Students' Understanding of Statistical Inference: Implications for Teaching

by

Robyn Reaburn, B.App.Sci.(Medical Technology),

BA, Dip.Teach.,Grad.Dip.Sci.,MSc.

Submitted in fulfilment of the

requirements for the Degree of Doctor of

Philosophy

University of Tasmania, October, 2011.

This thesis contains no material which has been accepted for a degree or diploma by the University or any other institution, except by way of background information and duly acknowledged in the thesis, and to the best of my knowledge and belief no material previously published or written by another person except where due acknowledgement is made in the text of the thesis, nor does the thesis contain any material that infringes copyright.

The research associated with this thesis abides by the international and Australian codes on human and animal experimentation, the guidelines by the Australian Government's Office of the Gene Technology Regulator and the rulings of the Safety, Ethics and Institutional Biosafety Committees of the University.

Abstract

It was of concern to the researcher that students were successfully completing introductory tertiary statistics units (if success is measured by grades received), without having the ability to explain the principles behind statistical inference. In other words, students were applying procedural knowledge (surface learning) without concurrent conceptual knowledge.

This study had the aim of investigating if alternative teaching strategies could assist students in gaining the ability to explain the principles behind two tools of statistical inference: *P*-values and confidence intervals for the population mean. Computer simulations were used to introduce students to statistical concepts. Students were also introduced to alternative representations of hypothesis tests, and were encouraged to give written explanations of their reasoning. Time for reflection, writing and discussion was also introduced into the lectures.

It was the contention of the researcher that students are unfamiliar with the hypothetical, probabilistic reasoning that statistical inference requires. Therefore students were introduced to this form of reasoning gradually throughout the teaching semester, starting with simple examples that the students could understand. It was hoped that by the use of these examples students could make connections that would form the basis of further understanding.

It was found that in general, students' understanding of *P*-values, as demonstrated by the reasoning used in their written explanations, did improve over the four semesters of the study. Students' understanding of confidence intervals also improved over the time of the study. However for confidence intervals, where sim-

ple examples were more difficult to find, student understanding did not improve to the extent that it did for *P*-values.

It is recommended that statistics instructors need to appreciate that tertiary students, even those with pre-tertiary mathematics, may not have a good appreciation of probabilistic processes. Students will also be unfamiliar with hypothetical, probabilistic reasoning, and will find this difficult. Statistics instructors, therefore, need to find connections that students can make to more familiar contexts, use alternative representations of statistical processes, and give students time to reflect and write on their work.

Acknowledgements

I would like to thank my supervisors, Professor Jane Watson and Associate Professor Kim Beswick, and my research supervisor, Associate Professor Rosemary Callingham, for their help and support over the time of my candidature. They have been consistently helpful, cheerful, and encouraging. It has been my good fortune to work with them. Thanks also to Dr. Des Fitzgerald, coordinator of the statistics unit described in this study, for his advice.

I wish to thank my parents, Kenneth and Phyllis Duck. They did all they could so that their daughters could have the education that they could never have. My father died during the time this thesis was written; he would have been thrilled to know that it has been finished.

Thanks to my daughters, who have managed well with a mother who has been tired and occasionally very absent minded. Most of all thanks must go to my husband Nicholas, for his consistent love and support. Much has happened in our lives over the past year – this thesis would never have been finished without him.

.

1. Introduction

1.1 Why do this research?

I have long felt that many students, although they can successfully follow the procedure to carry out a hypothesis test, do not understand the reasoning behind this process. In particular, it appears that students find difficulty in explaining the reasoning behind the *P*-values in hypothesis testing and in understanding that confidence intervals are used to estimate population parameters.

I, myself, was a successful undergraduate statistics student in that I received high grades. Looking back, however, I realise that although I was successfully following the process I would not have been able to explain the reasoning behind confidence intervals and hypothesis tests.

I am now a lecturer of a first year statistics unit at a tertiary institution and I have found that when students are asked questions that require conceptual understanding (in contrast to procedure) they often demonstrate a lack of understanding. For example, in one assignment students are asked to calculate the confidence interval for a mean, and then in a separate question they are asked to calculate the interval where 95% of the individuals are expected to lie. These questions cause intense angst and confusion. It is apparent from their answers that the students often do not appreciate the difference between the questions.

In another assignment students may be asked to explain the meaning of the phrase “significant difference” when testing for differences in population means. Since the sample means are not identical, why can it be concluded from the given *P*-

value that the population means may not be different from each other? It is apparent from the student answers that the students can successfully complete the process and conclude that the null hypothesis should be accepted. Many of them cannot, however, explain the role of sampling variation and what the *P*-value is in conceptual terms. It would appear that the students are using procedural knowledge only.

The literature indicates that my suspicion, that many students do not understand hypothesis testing, is also of concern to others. For example, Garfield (2002, p. 3) has found "that students can often do well in a statistics course, earning good grades on homework, exams and projects, yet still perform poorly on a measure of statistical reasoning such as the Statistical Reasoning Assessment." Garfield and Ahlgren (1988) also report that students use procedural knowledge without understanding the concepts behind what they are doing:

> The experience of most college faculty members in education and the social sciences is that a large proportion of university students in introductory statistics courses do not understand many of the concepts they are studying … Students often tend to respond to problems involving mathematics in general by falling into "number crunching" mode, plugging quantities into a computational formula or procedure without forming an internal representation of that problem. (p. 46)

My experiences as both a student and lecturer confirm that this "number crunching mode" can lead to success in a statistics course if success is only measured by the grade received. The purpose of this study was to find alternative methods of instruction that would enhance the students' gaining of conceptual knowledge of hypothesis testing and the estimation of population parameters.

1.2 The Research Questions

Before this research commenced the Data Handling and Statistics unit was taught in a didactic style and little attempt was made to discover the students' conceptions before and during the unit. The unit was taught in four modules. The first module consisted of the summarisation of data and data collection methods. The second module introduced probability and probability distributions. The third module introduced confidence intervals and hypothesis testing. The final module introduced the Analysis of Variance (ANOVA) and simple and multiple linear regression. One consequence of the way the material was presented was that students had to grapple with the formal hypothesis testing procedures at the same time as they were introduced to hypothetical and probabilistic reasoning. They had little or no time to gather experience with drawing conclusions using probability and to become familiar with hypothetical reasoning before the use of formal procedures.

With these factors in mind the research questions were:

- What are students' understandings of probability and stochastic processes on entering university? Are there any differences in understandings between those students who have studied statistics in their previous mathematics courses and those who have not?
- What are students' understandings of *P*-values at the end of their first tertiary statistics unit? How did these understandings change over the time of the study?

- What are students' understandings of confidence intervals at the end of their first tertiary statistics unit? How did these understandings change over the time of the study?

1.3 A note on the terminology

Students come into any learning environment with their own views of the world which they have gained over their life experience. These beliefs may be inconsistent with formal knowledge, that is, "inconsistent with commonly accepted and well-validated explanations of phenomena or events" (Ormrod, 2008, p. 245). In the educational literature views that are not consistent with formal knowledge, are referred to as "misconceptions," "misunderstandings" or, in the more recent literature, "alternative conceptions" (Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007). In the literature pertaining to tertiary statistics education, however, the term "misconceptions" is generally retained. Therefore this term is used for this study.

As stated in Section 1.2, the study aims to investigate students' understanding. This is based on the reasoning used in their answers to specified questions. What is meant by "understanding" and "reasoning" in mathematics and statistics education has resulted in considerable debate (for example, see Ben-Zvi & Garfield, 2004). Therefore it is important to define what is meant by these terms in this study. In this study students are considered to have "understanding" if they can make connections among related concepts, can represent concepts in different ways, and have conceptual and not just procedural knowledge (Australian Curriculum, Assessment and Reporting Authority (ACARA), 2011). "Reasoning" is shown when students justify the strategies they use and justify their conclusions (ACARA, 2011).

2. Literature Review – part I: Statistical Reasoning

2.1 What is statistics?

“Statistics is the science of collecting, organising, analysing, interpreting, and presenting data” (Doane & Seward, 2007, p. 3). In practice, statistics involves the use of numbers within a context and involves data collection, summarising these data in some way and making interpretations and decisions.

There are two general areas of statistics, descriptive statistics and inferential statistics. With descriptive statistics, data are summarised with graphs, tables and numbers such as means and standard deviations. Inferential statistics involves the making of conclusions about entire populations from samples (Doane & Seward, 2007). This latter field involves the use of probability and hypothetical reasoning. Because variation is universal, and no two samples are alike, no sample is likely to be exactly representative of the population from which it was drawn. The use of samples, therefore, always results in uncertainty concerning the accuracy of the conclusions inferred from samples.

Statistical analyses require some mathematical skills. The use of computers, however, has greatly reduced the time taken in computation, so that analyses are now much easier to do, although not necessarily easier to understand. Although statistics requires the use of numbers, from the students’ viewpoint there are important differences between statistics and other branches of mathematics with which they may be more familiar. In statistics the numbers always have a context, there is a need for correct data collection, and there is always uncertainty about the answers to questions posed about populations when samples are used. For students who are used to working towards a single “correct” answer in other branches of mathematics the need to address these differences can be unexpected and disconcerting.

These differences have led some writers to look at statistics as being not a branch of mathematics at all. For example, Shaughnessy (2006, p. 78) states, “Statisticians are quite insistent that those of us who teach mathematics realise that statistics is not mathematics, nor is it even a branch of mathematics.”

The successful use of statistics, however, does require skills that are usually regarded as mathematical. Not only does the discipline of statistics require the summary and interpretation of data with graphs and numbers such as the mean and median, it also requires hypothetical reasoning that in turn uses the mathematics of probability. Common statistical procedures are based on what Cobb and Moore (1997, p. 803) refer to as “elaborate mathematical theories [and] the study of these theories is part of the training of statisticians.”

2.2 Statistical reasoning

In practice, statistical reasoning involves being able to assess how well data are collected, describe the data, draw conclusions from the data, and allow for the uncertainty that results from the use of a sample. Students, therefore, need to understand how sampling is influenced by the variation that is present in every process (Wild & Pfannkuch, 1999). They need to grapple with the question, if variation is omnipresent and sampling variation is also omnipresent, what can be said about a population, when there is only a sample available? (Moore, 1990). Students need to recognise that a sample gives some information about a population, and the sample puts limits on the estimated value of a characteristic of the population. That is, students need to be able to cope with the conflicting ideas that samples do not exactly represent a population but are in some way still representative of that population (Rubin, Hammerman, & Konold, 2006).

A result of the tension between the representativeness and variability of samples is that statistical inference leads to the formation of conclusions based on a hypothetical reasoning process (hypothesis testing), and which are stated in probabilistic terms. The result of the presence of variation leads the user of statistics to answer the following question, "Is the observed effect larger than can be reasonably attributed to chance alone?" (Moore, 1990).

2.3 Hypothesis testing and confidence intervals

2.3.1 Hypothesis testing

What distinguishes science from other fields of knowledge? It was in the search for the answer to this question that Popper (1963) proposed the criterion of "falsifiability, or refutability, or testability" (p. 37). By this criterion, "statements or systems of statements, in order to be ranked as scientific, must be capable of conflicting with possible, or conceivable observations" (p. 39). A "theory that is not refutable by any conceivable event is non-scientific" (p. 36).

This proposal, that scientific statements must be capable of being falsified, is sometimes introduced to students with reasoning similar to this. A statement is made such as:

All swans are white.

It is not possible to prove this statement true. No matter how many white swans are observed, there is always the possibility that the next swan observed may not be white. In contrast, it is possible to disprove this statement by the observation of only one swan of another colour. Therefore, according to Popper's criterion of falsification, because the statement about swans is capable of being disproved, it is scientific.

Similar reasoning is used in statistical hypothesis testing. A proposition (the "null hypothesis", designated H_0) is made about a parameter (for example, the mean) of an entire population. This is written in such a way so that the sample data may be used to find evidence against it. The sample data are then collected and the appropriate sample statistic calculated. If the sample statistic is one that could be reasonably expected from a population with the proposed characteristic, then the hypothesis is accepted. If, however, the sample statistic is not one that would be expected from a population with the proposed characteristic the hypothesis is rejected. A complication is added by the omnipresence of sampling variation. Since it is known that if another sample were taken it would be different from the first, and that any one sample may or may not be representative of the population from which it was drawn, the decision to reject or accept the hypothesis is always made with uncertainty. Thus the hypothesis is never definitively proved or disproved.

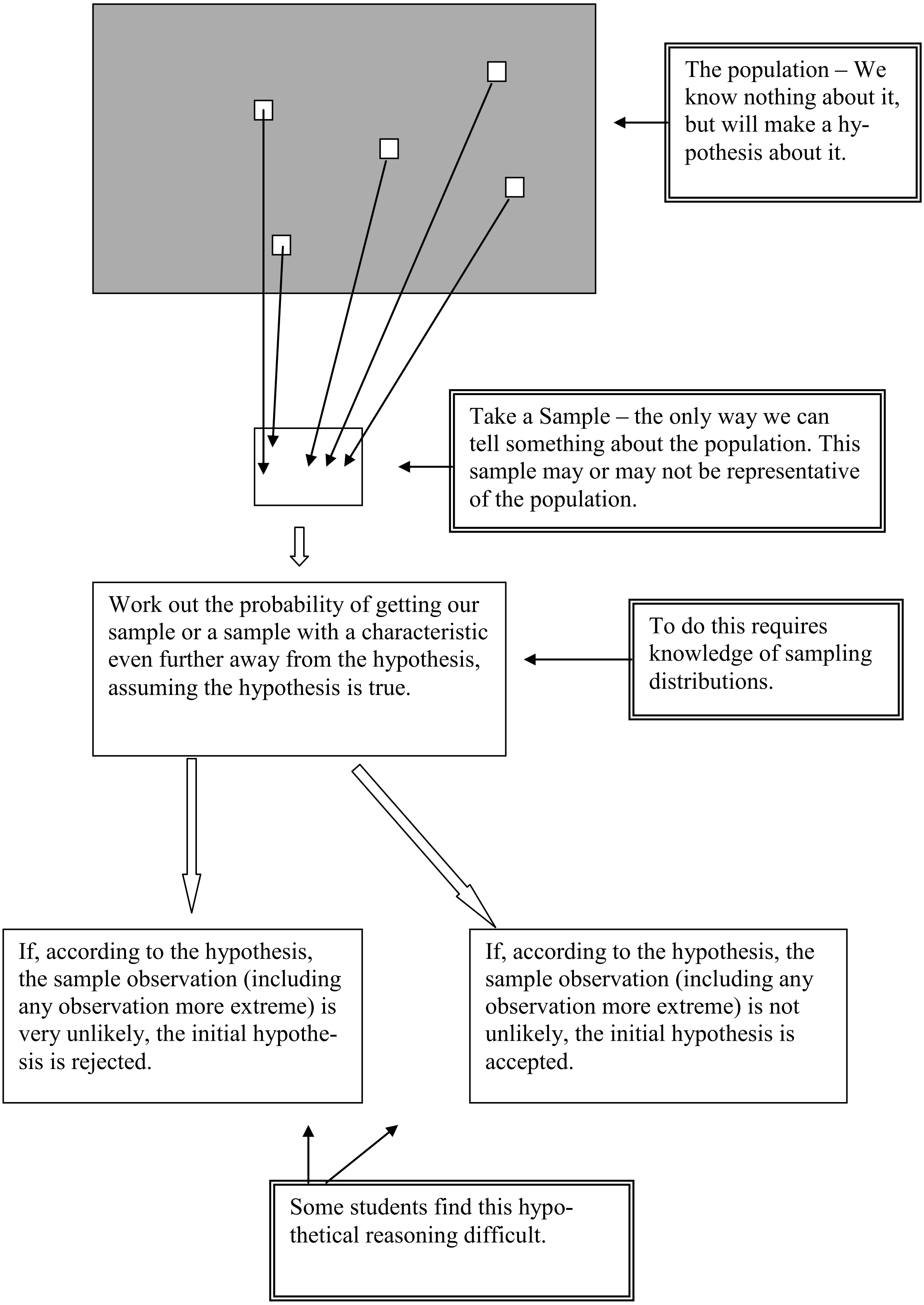
To manage this uncertainty the mathematics of probability is used to assist in making the decision to reject or accept the hypothesis. A probability is calculated using the following reasoning: If the null hypothesis about the population is true, how likely is the sample statistic or a statistic that is even less likely? For example, if the hypothesis is about the value of a population mean, the probability would be expressed in mathematical terms as:

$$P((|\bar{x}-\mu|\geq 0)\,|\,H_o=\mu)$$

where $\bar{x}$ is the mean of the sample, and μ is the mean of the population. If this probability is found to be very low, then it is concluded that evidence has been found against the null hypothesis and it is rejected. If this probability is not very low, then the hypothesis is accepted.

Lipson, Kokonis and Francis (2003) have summarised the reasoning involved in hypothesis testing as a stepwise process. The first step involves the recognition that no two samples are alike, even if they had been drawn from the same population. The second step involves comparing the sample result with that expected from the hypothesised population. To do this, a knowledge of sampling distributions (the pattern into which the sample statistics from the hypothesised population would fall) is required. If the hypothesis should be rejected, then the next step involves recognition that there is an inconsistency between the sample and the hypothesised population, and that the sample may not belong to that of the hypothesised population.

A diagrammatic model of a hypothesis test is in Figure 2.3.1.1.



The population – We know nothing about it, but will make a hypothesis about it.

Take a Sample – the only way we can tell something about the population. This sample may or may not be representative of the population.

Work out the probability of getting our sample or a sample with a characteristic even further away from the hypothesis, assuming the hypothesis is true.

To do this requires knowledge of sampling distributions.

If, according to the hypothesis, the sample observation (including any observation more extreme) is very unlikely, the initial hypothesis is rejected.

If, according to the hypothesis, the sample observation (including any observation more extreme) is not unlikely, the initial hypothesis is accepted.

Some students find this hypothetical reasoning difficult.

Figure 2.3.1.1. Model of a hypothesis test.

In summary, successful hypothesis testing requires:

- An understanding of randomness and probability.
- An understanding of data collection and the recognition that samples may not be representative of the parent population.
- An understanding of what summary statistics such as the mean and standard deviation represent, that is, an understanding that is more than just how these numbers are calculated.
- An understanding of how sample statistics such as the mean relate to the equivalent statistics in the population (in populations these statistics are known as *parameters*).
- An understanding that variation is omnipresent, and of the extent of variation to be expected in the data.
- An understanding of the legitimate interpretations of hypothesis tests, including the setting up and correct interpretations of the null and alternative hypotheses, and correct interpretations of *P*-values and levels of significance.

These areas are discussed in turn in Sections 2.4 and 2.5.

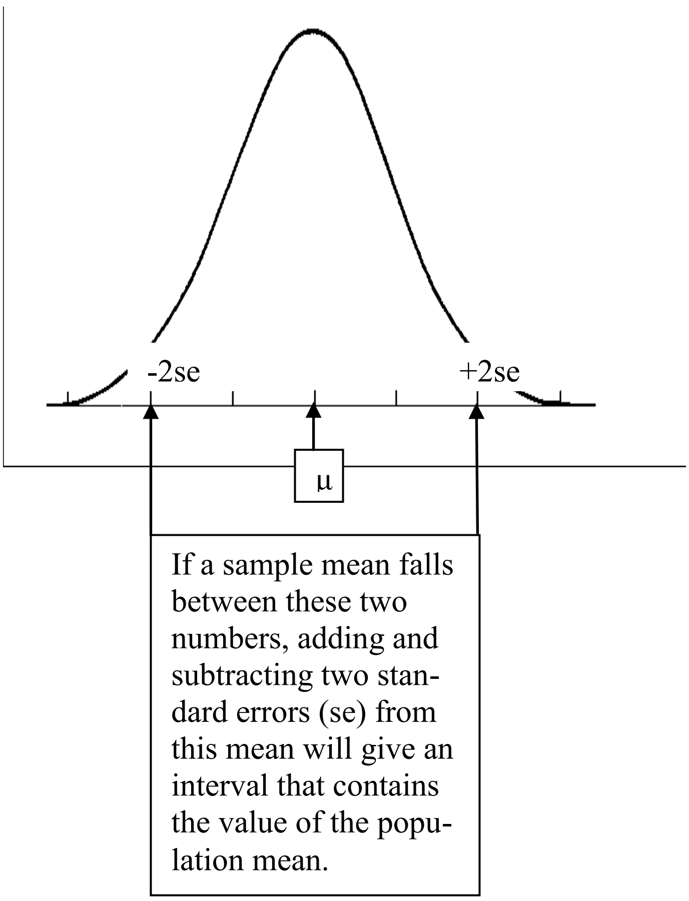
2.3.2 Confidence intervals

Confidence intervals estimate a population parameter based on a sample. They give a range in which it is considered likely the value of the population parameter will lie. In the Data Handling and Statistics unit at the University of Tasmania, students are required to estimate and interpret confidence intervals for the mean.

To understand the process, students need to know that approximately 95% of data that belongs to a Normal distribution will be within two standard deviations of the mean. They also need to know that if an infinite number of samples of the same size were taken, and the sample means calculated for each one, these sample means in turn would form a Normal distribution. What follows is a statement known as the Central Limit Theorem. If the sample size is large enough (a rule of thumb is 20 or more) then the distribution formed by the sample means is a Normal distribution, *regardless of the distribution of the original population*. This Normal distribution has the same mean as the original population, and the standard deviation of the sample means (known, rather confusedly, as *the standard error of the mean*) is equal to the standard deviation of the original population divided by the square root of the sample size. Therefore, a larger sample size will result in a smaller standard error.

If sample means have a Normal distribution, then the same rule applies to this distribution as any other Normal distribution. Approximately 95% of the sample means are then found within two standard errors of the population mean. This indicates that most sample means are within a "reasonable" distance of the population mean. The direct consequence of this knowledge is illustrated in Figure 2.3.2.1.

Students then have to contend with the idea that the process used to make the estimate of the population mean will be "true" 95% of the time, as 5% of the time a sample mean will be found that is outside of the two standard error interval. Section 2.5.4.6 examines the literature describing misconceptions students have about the process of finding and interpreting confidence intervals.



If a sample mean falls between these two numbers, adding and subtracting two standard errors (se) from this mean will give an interval that contains the value of the population mean.

Figure 2.3.2.1. The relationship between the distribution of sample means and the process of finding a confidence interval to estimate the value of the population mean.

2.4 Misconceptions with probabilistic reasoning

2.4.1 Introduction

According to constructivist theories of learning all students come into any learning environment with their own preconceptions that may or may not be correct. Students combine concepts into a *schema* – a mental representation of an associated set of perceptions, ideas or actions. If new knowledge is understood, this means that a student has successfully assimilated the new information into an appropriate schema. If the student has an inappropriate or non-existent schema then assimilating later ideas can become difficult, if not impossible (Krause, Bochner, & Duchesne, 2007).

If students have pre-existing inappropriate or non-existent schemas about probability they will not be able to understand statistical inference. Statistical inference relies on the mathematics of probability from the selection of the sample to the drawing of the final conclusions. The literature shows, however, that a person's intuitive views of probability are often inappropriate or incomplete. These inappropriate or incomplete views may be difficult to detect because probability questions, using examples such as coin tosses, can be simple to answer. As people get older their intuitive perceptions of statistical phenomena, even though inappropriate, may get stronger, and formal instruction may not correct these intuitive views (Moore, 1990). Furthermore, it has been found that students may use the formal views inside the classroom, but revert to their own intuitive views, whether correct or not, outside the classroom (Chance, delMas, & Garfield, 2004).

Over the last three decades a body of research has been produced on probabilistic reasoning. This research has identified several errors in intuitive reasoning that are described in the following sections.

2.4.2 The contribution of Tversky and Kahneman

In their groundbreaking work, Tversky and Kahneman (1982b) provided extensive research that described the judgement heuristics (the "rules of thumb") that are used in probabilistic reasoning. One of these judgement heuristics is known as the “representative heuristic,” in which probabilities are evaluated by the degree to which A resembles B. If A is very similar to B, the probability that A originates from B is judged to be high. If A is not similar to B, the probability that A originates from B is judged to be low (Tversky & Kahneman, 1982b). This heuristic manifests itself in several forms. The first of these involves the neglect of base rate frequencies and is illustrated the following example. The participants in Tversky and Kahneman’s study had to assess the probability that “Steve” was a farmer, salesman, airline pilot or librarian. Steve was described as being meek and tidy with a passion for detail. The participants tended to suggest that Steve was a librarian, as he fitted the stereotype of a librarian. In using this reasoning, however, the participants ignored the base rate frequencies, that is, the number of each occupation in the population.

Using this representative heuristic also led participants not to realise that deviations from the expected value generated by a random process are more likely in samples of small size. The participants, therefore, expected that 50% of coin tosses would be heads even with a small number of tosses. This expectation was reinforced by a belief in the “fairness” of the laws of chance. Therefore in a series of coin tosses the independence of each event was ignored so that if a series of Tails had been tossed, a Head was regarded as more likely than a Tail. This latter misconception is known as the “Gambler’s fallacy” (see also Fischbein & Schnark, 1997).

This heuristic also led the participants to predict that two samples drawn from a population would be more like each other than is the case in reality. The participants who used the representativeness heuristic also expected that one outlier in data (say a very high result) would be cancelled out by further data (say a very low result) instead of being merely diluted by other data (Tversky & Kahneman, 1982b).

Tversky and Kahneman (1982b) described another judgement heuristic that they referred to as “availability” (p. 11). This led the participants to judge the likelihood of an event by the ease with which instances of it could be brought to mind. By this heuristic a class that can easily be retrieved will appear more numerous than a class of equal or higher frequency that cannot be so easily retrieved. Therefore the subjects considered words in the form of „_ _ _ _ ing’ to be more common than words in the form „_ _ _ _ _ n _’, even though the first example is a subset of the second (Tversky & Kahneman, 1983). This heuristic also led to the participants assuming correlation between variables when, in fact, this correlation does not exist (Tversky & Kahneman, 1982b, pp. 11-13).

Tversky and Kahneman (1982b, pp. 15-16) also described the phenomenon of “anchoring” where a person will make an estimate from an initial value and then make adjustments to suit the situation. Unfortunately, the adjustments are usually insufficient. For example, in a study where the subjects had to judge the probability of getting seven consecutive red marbles where 90% of the marbles are red (a “conjunctive” event) the participants started with the probability of the initial draw (0.90) and then estimated down for seven in a row but then did not make the adjustment large enough (the final probability is 0.48). When they had to judge the probability of getting at least one red marble in seven tries where 10% of the marbles were red, the subjects started with the probability of the initial draw

(0.10) and then estimated up to allow for the seven tries but again did not make a large enough adjustment (the final probability is 0.52).

2.4.3 Other misconceptions about probability

Since the work of Tversky and Kahneman other researchers have examined people's understanding of probability. Konold (1989) added another judgement heuristic to the list, the "outcome approach." With this heuristic it was found that participants in his study made errors because they had a desire to predict the outcome of a single trial in a probabilistic process, and would judge their predictions as correct or not on this single trial. Lecoutre (1992) added the "equiprobability bias," where two outcomes of different probabilities are judged to be equally probable even when this is not the case. For example, participants in her study judged that if two dice were rolled then a combination of a „5' and a „6' was equally likely as two sixes. The participants failed to realise that there were twice as many ways of achieving a „5' and a „6' rather than two sixes.

Fischbein and Schnark (1997) have described what they call the "time-axis fallacy" where they found that people can predict the probability of events in the future, but do not use later knowledge to predict the probability of events in the past. This fallacy can be illustrated with an example.

> You have an urn with two white marbles and two black marbles. You take out a white marble. Without replacing it, you take out a second marble. What is the probability that this second marble is also white?
>
> You put all the marbles back and then take out a marble and put it aside without looking at it. You take out a second marble that is white. What is the probability that the first marble was white? (Adapted from Fischbein and Schnark, 1997)

It was found that in answering the second part of this question the participants correctly realised that the second draw could not influence the first draw, but they failed to realise that knowledge of the second draw could be used to determine the probability of the first draw (Falk, 1986).

If a student has poor intuitive ideas on how probabilistic processes work, it can be expected that this student will have difficulty understanding statistical inference and hypothesis testing, which are based on these probabilistic processes. In addition, Garfield and Ahlgren (1988) have found that these problems are exacerbated if the student has poor basic mathematical skills, especially in the use of proportions, fractions, decimals and percents.

2.4.4 Misconceptions about conditional probability

Because hypothesis testing involves conditional probabilities it is important that students can interpret conditional statements.

Formally, a conditional probability is presented as

$$\mathrm{P(A|B)} = \frac{\mathrm{P(A \cap B)}}{\mathrm{P(B)}}$$

The sample space is therefore restricted to points in an event B; that is, only points in A that are in set B are of interest. For example, if the probability of getting a king, given that the requirement that the card is a spade, the formula would read

$$\mathrm{P(King|Spade)} = \frac{\mathrm{P(King\ and\ Spade)}}{\mathrm{P(Spade)}}$$

In this case, the only card of interest is the one card that is both a king and a spade, out of the total number of spades. Cards of other suits are not of interest.

Watson and Kelly (2009) found that students in elementary school can use informal conditional reasoning with probabilities, and can understand that sampling without replacement affects the probabilities of outcomes of subsequent selections. Students can also calculate conditional probabilities accurately when the data are in the form of a frequency table. However, when conditional statements are put into social contexts, the students' background knowledge may interfere with the calculations of these probabilities. The "time-axis fallacy" may also play a part; that is, students can predict the probability of events in the future given past events, but cannot give the probability of events in the past given later knowledge (Fischbein & Schnarch, 1997).

2.5 Other misconceptions about statistical reasoning

2.5.1 Misconceptions about randomness

Randomness has a different mean-ing in everyday speech compared to its meaning in mathematics. In everyday speech randomness can refer to any event that cannot be predicted, is haphazard, or is without a definite purpose. In mathematics, a process is random if the occurrence of an event follows a probability distribution. Therefore, if an event has a 60% chance of occurring (for example selecting a red marble from a bag with 60 red marbles and 40 blue marbles), although the result of each individual selection cannot be predicted, it is known that in the long term a red marble will be drawn approximately 60% of the time.

The varying meanings that students apply to the term randomness have been investigated by LeCoutre, Rovira, LeCoutre and Poiteviniau (2006). Some participants in their study believed that randomness applied for any occurrence where the cause is unknown. If a cause is found, these participants then believed that the occurrence was no longer random. The authors also found that some participants believed that randomness applied to any situation where the probability was easy to compute. Of interest was the finding that a background in probability study had little effect on the accuracy of the participants' beliefs. This illustrates how students' intuitive knowledge, while inappropriate, may be difficult to correct (Krause et al., 2007).

2.5.2 Misconceptions about sampling

In general, senior high school students and students in undergraduate statistics courses have little experience with sampling (Rubin, Bruce, & Tenney, 1991). Students, therefore, have no knowledge of the extent to which samples may or may not be representative of populations, and how much one sample may vary from the next. As a result, students generally expect samples to be much more representative of the population than they really are, and if a sample is found to be representative, they regard it as "accurate" (Rubin et al., 1991). It has also been found that some students ignore the effect of sample size on the characteristics of a sample. They will concentrate on the possible increase of the within sample variation as a sample size increases, without realising the effect that increasing sample size has on the standard error (Finch, 1998). This latter finding has also been noted by delMas, Garfield, Ooms and Chance (2007).

2.5.3 Misconceptions about measures of central tendency

For students to perform hypothesis tests successfully, they need to understand concepts related to the mean and other measures of central tendency. In particular, it needs to be understood that the arithmetic mean is in some way representative of a group. It is because of this representativeness that many common hypothesis tests are about the mean of a population, or the difference in means of two populations. The calculation of the arithmetic mean, the mode and median are simple, yet students from primary school to the tertiary level have been found to have difficulties with using and understanding these statistics.

For some school students the mean is defined as the algorithm to calculate it, and there is no conceptual understanding of what the answer might represent. In a study of students from Grades 4 to 8, Mokros and Russell (1995) found that some students did not see the arithmetic mean as representative of the data, but saw the mode as the most representative number instead. Mokros and Russell did, however, find that reasoning improved as the students became older. In a study by Strauss and Bichler (1988), some school students (aged 8 to 14) did not see the mean as a number representative of a data set, but were more likely to do so as the students became older. Some of the students in Strauss and Bichler's study also had difficulties understanding the result when the mean was a fraction that had no counterpart in reality, such as one third of a soccer ball per person. As the students in this study became older they approached the mean in a more conceptual way, in that they either saw the mean as a "reasonable" number, as a "midpoint" or as a "balance" point, where the sum of distances of the data points above the mean to the mean is equal to the sum of the distances of the data points below the mean to the mean (p. 36).

Once students see the mean as a representative number for a data set, they should also be able to see that this representativeness allows for comparisons of data sets. This understanding should later be extended to making inferences involved in comparing populations (Gal, Rothschild, & Wagner, 1990).

Watson and Moritz (1999) gave students from Grades 6 to 9 a problem that required them to compare two data sets that were presented in the form of graphs. Many of the students did not use the mean in their conclusions, and of those who did (10% of the Grade 6 students and 54% of Grade 9 students), did not always do so successfully. Similarly, for students in Grades 3 to 9, Gal, et al. (1990) found that some students did not use the mean in their comparisons even though they had demonstrated familiarity with the algorithm. In a study of students from Grades 5 to 8, Hancock, Kaput and Goldsmith (1992) also found that some students did not use means to make comparisons between groups of unequal size, many using totals instead. Hancock et al. (1992) also found that students tended to give individual cases in a group more importance than is desirable. In addition, they would produce graphs if required for their assessments, but then would ignore them when drawing their conclusions, indicating that they did not see the representative nature of these forms of data representation either.

Because understanding generally improves as the students become older it would be expected that late secondary students and tertiary students would have a more sophisticated and accurate understanding of the mean and other measures of central tendency. These students should also be more likely to know which measure of central tendency is most appropriate for a given data set. However, many still have difficulty with these ideas. Pollatsek, Lima and Well (1981) found that when given a situation in which a weighted mean was required, a “surprisingly large proportion” of college students could not calculate it, did not understand the concept, and did not recognise that the ordinary mean was not appropriate. Many of these students did not comprehend that an error had been made, even when given follow up questions that were designed to prompt them to recognise their errors.

Groth and Bergner (2006) examined the understanding of Grade 12 students and preservice teachers related to the mean, median and mode. They found that most students chose to use the mean of data when it was more appropriate to use the median. Only 7% of these students could discuss which measure was most suitable for each example they were given. Although some students could explain that the mean and median measured the centre of the data or were in some way representative of the data, there were still students who could only explain these measures in terms of the algorithms used to calculate them.

2.5.4 Misconceptions about statistical inference

Garfield and Ahlgren (1988) have described how students in tertiary institutions may not understand the concepts of inferential statistics and therefore fall into “„number crunching’ mode, plugging quantities into a computational formula or procedure without forming an internal representation of that problem” (p. 46). The following sections describe the literature regarding misconceptions in statistical inference, particularly those regarding hypothesis testing and the generation of confidence intervals.

2.5.4.1 Misconceptions about variability

Because variation is omnipresent, Reid and Reading (2005) suggested that the success of students in statistics depends on how well they can develop an understanding of variation in different contexts. The ability to look at a data set, with its variation, as a whole comes with experience. Garfield, delMas and Chance (2007) found that at the beginning of an introductory university statistics course their students tended to focus on individual points, and the range, but not to focus on where most of the data were. They found that extensive practice was needed for students to see the data as a whole. In a study of pre-service teachers, Leavy (2006) found that these teachers concentrated on the summary statistics and ignored the variation, rather than focussing on the data as a whole.

Liu and Thompson (2005) studied in detail tertiary students' understanding of the standard deviation. The students were required to arrange the bars in a histogram of discrete data so that the highest standard deviation would be produced. Initially the students spread the bars evenly across the *x*-axis, illustrating they were not considering the standard deviation in terms of total deviation from the mean. Liu and Thompson concluded that few students had a conception of the standard deviation that combined both frequency and deviation from the mean.

2.5.4.2 Misconceptions about sampling distributions and the Central Limit Theorem

The understanding of sampling distributions requires the integration of several concepts – sampling and variability, the Normal distribution, the distribution of sample means, and standard errors (Batanero, 2008; Chance, delMas, & Garfield, 2004). Therefore if students have misconceptions about any of these topics, or have not or cannot put them together correctly, they are not going to have a full conceptual understanding of how statistical inference takes place. DelMas, Garfield, Ooms, and Chance (2007) found that indeed there are students who cannot use these concepts simultaneously. In particular, they cannot deal with concurrent use of the mean of the sample, the mean of the population and the mean of the sampling distribution.

It has also been found that students may expect sets of data to be normally distributed when this is not the case. Bower (2003) described how some students have a tendency to believe that the larger the sample size, the closer the distribution of any statistic will approximate a Normal distribution, even when there is an obvious lower bound such as zero. These students will believe, therefore, that something is wrong if a non-Normal distribution is found. DelMas et al. (2007) discovered that there are some students who, at the end of a tertiary statistics unit, showed a more fundamental lack in understanding, in that they could not draw a Normal distribution when given its parameters.

The law of large numbers states that as the number of trials increases, the observed proportions of a probabilistic process will converge to the theoretical proportion. For example, the larger the number of coin tosses, the closer the proportion of heads will be to fifty percent. The Central Limit Theorem states that if the sample size is large enough, the means of samples of a fixed size will form a Normal distribution, even if the original population is not normally distributed. Thompson, Liu and Saldanha (2007) describe how some students mix up the law of large numbers with the Central Limit Theorem. These students believe that as the sample size increases, the distribution of the sample means will not necessarily be Normal, but look more like the distribution of the original population. As a consequence, they do not understand that variability among samples means is less than that of the individuals of the population, and that variability among sample means will decrease with an increase in sample size.

Lipson (2002) studied students' understanding of the sampling distribution of a statistic (for example, that sample means form a Normal distribution) in contrast to the distribution of the sample (the pattern formed by the individuals in the sample). In her study, 43% of the participants could state that the sample statistic was determined from the sample, and that the variability of this sample statistic is described by the sampling distribution. Twenty two percent of the participants, however, stated that the distribution of the sample (the individuals) was the same as the distribution of the sampling statistic. This left 35% percent of the students who did not clearly indicate their reasoning. Further evidence of difficulties with the distribution of samples was supplied in a study of teachers by Liu and Thompson (2005) who found that the participants did not incorporate the idea of the distribution of sample statistics at all, and therefore did not consider whether or not a statistic was unexpected when deciding on the outcome of inferential reasoning. Lipson (2002, p. 1) concluded that the concept of the sampling distribution is "multifaceted and complex" and as a result students find this difficult.

If students have misconceptions about the distribution of the sample, the distribution of sampling statistics and the Central Limit Theorem, it is not surprising that they may try to cope by looking for rules to apply when carrying out statistical inference, and not use conceptual understanding. Unfortunately, learning by applying rules may lead to error. If understanding is not present, a student may apply the rules inconsistently or even misremember them (Chance et al., 2004).

2.5.4.3 Misconceptions about null and alternative hypotheses and their interpretation

Part of the procedure for hypothesis testing is to propose a hypothesis about the population of interest. This hypothesis should be falsifiable (see Section 2.3.1) and has to be stated correctly, so that when the hypothesis test is completed an incorrect statement will not be accepted or rejected. Unfortunately, some students do not even realise that the null hypothesis refers to a population, but instead think that a null hypothesis can refer to both a population and a sample (Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007).

A further problem is that the steps can be carried out correctly but the meaning of the results can be misinterpreted as students may have an inappropriate understanding of what accepting or rejecting a null hypothesis really means. This latter problem was investigated by Haller and Krauss (2002). Of particular interest is that many of their subjects were instructors in undergraduate statistics courses. The researchers surveyed 113 staff and students (including statistics instructors) from psychology departments at six universities. The participants were given the example of an independent samples *t*-test to determine if there was a significant difference in the means of control and experimental groups. In their example the *P*-value was .01. The participants were asked to agree or disagree with each of the following statements.

1. You have absolutely disproved the null hypothesis (that is, there is no difference between the population means).
2. You have found the probability of the null hypothesis being true.
3. You have absolutely proved your experimental hypothesis (that there is a difference between the population means).

4. You can deduce the probability of the experimental hypothesis being true.
5. You know, if you decide to reject the hypothesis, the probability that you are making the wrong decision.
6. You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.

Eighty percent of the participants who were statistics instructors marked at least of one of these statements as true, while all the psychology students marked at least one of these statements as true. When analysed by statement, Statement 5 had the highest percentage incorrect (average 74%). For this statement it appears that the caveat, "if H_0 is true", was not known, forgotten or ignored. Statements 4 and 6 were the next highest in percentage incorrect (average 47%); while Statements 1 to 3 were the most often correct (average incorrect 18%).

Further investigation by Haller and Krauss (2002) showed that for some of these students and instructors the results of a hypothesis test was regarded in the same light as a mathematical proof. Thompson et al. (2007) also found in their research this tendency to believe that the results of a hypothesis test indicate proof. Subjects, believed that to reject a hypothesis test means that the hypothesis has been proved wrong, and that further evidence in unnecessary.

It cannot be expected that students will have a good understanding of hypothesis testing if their instructors have misconceptions as well. It would also appear that that students of statistics wish to be able to make definite conclusions about their data, which is not possible.

2.5.4.4 Misconceptions about the interpretation of the *P*-value and the level of significance (α)

In the hypothesis testing procedure, the level of significance (denoted by α) indicates the level of probability at which the null hypothesis will go from being accepted to rejected. It also indicates the maximum chance of rejecting the null hypothesis if this hypothesis *is actually true*. The *P*-value is the probability that given a particular hypothesis about a population parameter, the sample statistic or one even more unlikely, is observed.

Several studies have found a combination of the following misconceptions about the level of significance to be held by both students and researchers. The level of significance may be regarded as the probability that one of the hypotheses is true. The level of significance may also be regarded as the probability of being wrong, or just as the probability of making a mistake (Batanero, 2008; Nickerson, 2000).

Similar misconceptions apply to the meaning of the *P*-value. The *P*-value is believed by some to be the probability that the null hypothesis is true. The *P*-value is also interpreted as the probability that the event of interest could happen, given that the null hypothesis is true (Gliner, Leech, & Morgan, 2002). Students may also believe that any one *P*-value will be replicated if the experiment is replicated. They do not realise that a particular sample statistic is unlikely to be replicated, and, furthermore, do not realise that the probability of having a sample statistic that exactly replicates the null hypothesis is extremely small (Cumming, 2006; Mittag & Thompson, 2000; Nickerson, 2000).

Mittag and Thompson (2000) found that students may believe that the *P*-value indicates the strength of a relationship. Therefore they will believe that statistical significance also implies practical significance (Gliner et al., 2002). A study of researchers, however, found that these subjects were generally aware that this was not the case (Mittag & Thompson, 2000).

2.5.4.5 General misconceptions about hypothesis tests

Students may carry out hypothesis tests with correct procedures and conclusions, but then give these conclusions unwarranted meanings. For example, if a null hypothesis is rejected, students may conclude that the theory behind the experiment is true. If the null hypothesis is accepted, then the experiment might be regarded as a failure (Nickerson, 2000).

Another problem is that students may carry out a hypothesis test, and then look for non-statistical reasons for their conclusions. For example, Lipson, Kokonis, and Francis (2003) found that even when the students in their study verbalised that the likelihood of the sample coming from the hypothesised population was small, they tended to look for a practical explanation (deliberate tampering with the sample, for example) rather than dealing with the statistical solution. Kaplan (2009) found that students rated the results of statistical inference on the strength of their own beliefs. Therefore if they did not believe a conclusion, the students would look at problems in design and ask for further information. They did not do this, however, if they believed the results. The students did not realise that it was necessary to discuss the strengths and weaknesses of any experiment, and not just when they were surprised by the result.

2.5.4.6 Misconceptions about the interpretation of confidence intervals

Confidence intervals are relatively simple to calculate but appear to cause problems in interpretation. The purpose of confidence intervals is to estimate a population parameter from the sample, with an indication of the uncertainty due to chance variation (Moore, 1990). If, for example, it is stated that the 95% confidence interval for the mean is between 20 cm and 25 cm, the value of the population mean is believed to be between these two numbers (including the end points). To the educated reader, the 95% indicates that the process used will give a correct estimation 95% of the time, but that for any one interval it is not known whether the answer given is correct. That is, the accuracy of the individual result is unknown, but the overall level of uncertainty is known.

In a study of undergraduate students, delMas et al. (2007) found that about one third of students believed that a confidence level indicated the percentage of population values that lie between the confidence limits. These students also had a tendency to believe that the confidence level represents the percentage of sample values within the confidence limits. In addition, the majority of the subjects in delMas et al.'s study indicated that the level of confidence denoted the percentage of all sample means that lie between the confidence limits. These students did not understand that their knowledge of Normal distributions led to a *process* that allows estimation of the population mean. As a result, they were not able to take the step from the knowledge that 95% of all sample means are within two standard errors of the population mean, to knowledge that this leads to a process that enables inference about the population mean.

In a study of researchers, Cumming (2006) found that there was a common misconception that if a sample were taken and the sample mean and confidence interval were calculated, then the level of confidence would give the percentage of means that would fall within the original confidence interval if replicate samples were taken.

2.6 The persistence of preconceived views

2.6.1 Introduction

Students can hold multiple and contradictory views on probability as well as on other topics in mathematics, and on many other physical phenomena (Konold, 1995). In the classroom, students may use the formal methods, but resort to their own intuitive knowledge and methods outside the classroom. Because intuitive beliefs appear self evident and obvious, students are reluctant to change them (Fischbein & Schnarch, 1997). When presented with information that conflicts with their previous views, students may rather look for and find evidence for their misconceptions than use a new theory (Dunbar, Fugelsang, & Stein, 2004; Shaughnessy, 1992). If the new theories do not fit into the structure provided by their intuitive theories, these new theories are then integrated into their schema with difficulty (Dunbar et al., 2004). It is for this reason that instruction may not correct previously held misconceptions.

2.6.2 How students change their previous conceptions

In his book "The Structure of Scientific Revolutions," Kuhn (1996) suggests that scientists work with traditions that provide coherent models that define rules and standards for accepted scientific practice, referred to as paradigms. The observation of new phenomena may influence these paradigms in three ways. Firstly, the observations might fit in well with the existing paradigms. Secondly, a new theory may be developed that articulates well with the existing paradigms. Thirdly, a new paradigm may result that replaces the old paradigm. This will happen, however, only if attempts at articulation fail. Kuhn states: “Only when these attempts at articulation fail do scientists encounter the third type of phenomena, the recognised anomalies whose characteristic feature is their stubborn refusal to be assimilated to existing paradigms” (p. 97). Strike and Posner (1985) made a similar claim for individuals. Students will not make a major change in a former conception, or form a new conception, until they have found that less radical changes will not work. This change will not take place, however, unless the new/adjusted conception is at least minimally understood, appears initially plausible and has explanatory power. These ideas are further developed in the discussion on theories of learning in Chapter 3.

2.7 Implications of the literature for teaching statistics

For those who teach applied statistics courses, the message from the literature is twofold. Firstly, because it appears that students can successfully complete statistics courses using procedural knowledge only, if conceptual understanding is considered to be important, then conceptual understanding will need to be assessed (Kelly, Sloane, & Whittaker, 1997).

Secondly, it has to be appreciated that the concepts such as randomness, distributions of sample statistics and probabilistic conclusions are abstract and complex. Understanding these concepts is complicated by the lack of background in mathematics that many students may have. Students may also have difficulty in working with the non-deterministic view of the world that the discipline of statistics requires (Yilmaz, 1996).

With these factors in mind, Cobb and McClain (2004) recommended that the teaching of statistics should focus on developing statistical ideas, use real data and classroom activities that promote statistical understanding, use appropriate technological tools that promote statistical reasoning, and promote classroom discourse. In addition, assessment should be used to monitor students' conceptual development. Furthermore, it has to be recognised that it is easy to underestimate the difficulty students have in understanding basic concepts of probability and statistics, and to overestimate how well students understand these concepts.

3. Literature Review part II: The nature of learning

3.1 Introduction - What is learning?

If someone says they have "learnt" something, what do they mean? For some people learning has taken place if they can reproduce a series of facts, whereas for others, learning means something more comprehensive. For example, on giving students a text to study, Martin and Säljö (1976) found that some students aimed just to reproduce the contents of the text ("surface learning"), whereas other students tried to understand the intention of the author of the text ("deep learning"). In a more general context, deep learning, in contrast to learning that merely reproduces content, can be considered to have occurred when the student can understand the material, relate parts to a whole, integrate it with existing knowledge and apply it in real world situations (Boulton-Lewis, 1995).

Boulton-Lewis (1995) described the acquisition of knowledge as a series of steps where students begin with declarative knowledge and progress to procedural knowledge, conditional knowledge, theoretical knowledge and then to metatheoretical knowledge. In this system, declarative knowledge consists of factual knowledge. Procedural knowledge allows the manipulation of declarative knowledge to undertake a task, to solve a problem, and to make decisions. These can be compared with surface learning. Conditional knowledge allows a person to know when to use certain procedures for different purposes. Theoretical knowledge (deep learning) involves being able to make abstracted or generalised statements going beyond particular instances. Metatheoretical knowledge is knowledge about the process of abstraction and theory building (Krause et al., 2007; Mason & Spence, 1999).

With this in mind it can be said that students who can complete their statistics courses but cannot explain the reasoning behind what they do have achieved surface learning or procedural knowledge, and have not achieved theoretical knowledge or deep learning.

3.2 How learning occurs

3.2.1 Introduction

One aim of this research was to discover students' understanding about probability and hypothesis testing so that the teaching of statistics to first year university students could be adapted to make deep learning more likely to be achieved. From reading students' work in the past it was clear that some of these students had ideas about statistical inference and hypothesis testing that were not in their lecture notes, not in their text, and were never presented to them in their lectures or tutorials. These students were not merely reproducing what they had learnt or read, but were somehow processing this information for themselves.

It is proposed that examining the theories of how students learn may enhance teaching practice so that deep learning, without misconceptions, may become more likely. Because this research is focussed on students' understanding, the following discussion concentrates on cognitive models of learning. Cognitive learning theories concentrate on internal mental processes and on how learners manipulate both new and familiar information.

3.2.2 Information processing theory

Information processing theory states that human memory does not simply retain information but is an active system. This system actively selects the sensory data that are to be processed, transforms the data into meaningful information, and stores much of the information for later use. Learning comes about when information from the environment is transformed into cognitive structures.

Several models have been proposed to describe how human memory works. The first to be described here is the “multistore” model (Krause et al., 2007). According to this model, information is noted by the senses and if attention is paid to the information, it will be transferred to the short term memory. If the information is considered important enough it will be transferred to the long term memory, if not, it will be forgotten. This long term memory is made up of episodic memory, which holds memories of personal events, semantic memory, which holds language and knowledge of how the world works, and procedural memory, which holds knowledge of procedures for performing the skills we need (Krause et al., 2007).

An alternative model is called the “levels of processing model” where attention is paid to the level of information processing. According to this model “deep processing” occurs when information is analysed and enriched by making connections with existing knowledge. Information that is analysed more deeply will be remembered (Krause et al., 2007).

These information theories are useful when designing a teaching program because they alert instructors to the possibility that since information is not just merely taken in without further processing, learners may gain ideas that were not intended by the lesson. They also can help instructors to understand why students may not retain information they have been given. Learners may not retain information because they have failed to pay adequate attention, they may be not motivated to remember, they might have inadequate memory skills, or they might not have the right cue to recall the information. These problems may be exacerbated if students are taught in a didactic way (Perkins & Simmons, 1988). Learners may also not retain the information because the short term memory is limited (Krause et al., 2007). Consequently, if too much new information is given at once, students will not be able to process this information effectively (Wieman & Perkins, 2005).

3.2.3 Constructivist theories of learning

Constructivist theories, like information processing theories, suggest that learners are not passive recipients of knowledge, but play an active part in their own learning. Rather than learners passively receiving knowledge as it is given, learners actively construct new knowledge by linking it to prior knowledge and understanding. It is argued by some (radical constructivists) that knowledge, no matter how it be defined, is in the heads of persons, and that as a result learners have no alternative but to construct what they know on the basis of their experience (von Glaserfield, 1995).

There are variants of this theory. Psychological constructivism focuses on individual learners and how they construct their own knowledge. Social constructivists focus on social interaction as a key component to learning, and indeed, may argue that this interaction is essential for learning (Krause et al., 2007). It is argued that as learners live in their own particular social and cultural environments, particular meanings are given to the events and objects that are encountered, and therefore all learning is socially mediated (Tobin, Tippins, & Gallard, 1994).

According to constructivism, learners operate with a *schema.* A schema is a cluster of ideas about a particular object or experience that is used by the person to organise existing knowledge in a way that makes sense. When learners come across a new situation, they may be able to *assimilate* the new knowledge into a pre-existing schema without modification. If an inconsistency arises between new information and a current schema, learners experience *disequilibrium*, and then have to modify the existing schema by the process of *accommodation* (Krause et al., 2007).

Using constructivist theory, Perkins and Simmons (1988) described how errors may occur as students acquire new information.

- Students may have naïve, underdifferentiated, and malprioritised concepts that may rival and override those of the new topic.
- Students may have difficulty accessing freshly acquired knowledge, especially if it was given in a didactic fashion.
- Students may mix up the new knowledge in various ways, and the result is a garbled version of what was intended.
- Observations that do not fit into their previous intuitions may be ignored, and only observations that fit into these prior intuitions will be acknowledged (see Section 2.6.1).
- Students may not be aware of the need for coherence in their knowledge, and therefore may prefer to keep to their previous intuitions than to seek internal coherence.

If Perkins and Simmons are correct, then students will resist the process of accommodation; that is, they will prefer to hold to a view of the world that is internally inconsistent rather than go through the process of modifying an existing schema. Students will only modify an existing schema if something easier will not work, and if the new schema is in some way plausible.

3.2.4 Implications of the cognitive models for teaching

If we accept that learners construct knowledge and understandings based on what they already know and believe, then it would seem important that instructors should discover what knowledge their students have, and what their problems are likely to be. "It is easier to orient students towards a particular area of conceptual construction if one has some idea of the conceptual structures they are using at present" (von Glaserfield, 1995, p. 185).

If students have prior conceptions that are not consistent with what is being taught, instead of undergoing the process of accommodation, they may build further misconceptions, or use what is being taught inside the learning environment but use their own ideas outside it (Bransford, Brown, & Cocking, 2000). Because these students are using what is taught inside the learning environment they may appear to have made an accommodation when, in fact, they have not done so. Strike and Posner (1985) stated:

> Typically, students will attempt various strategies to escape the full implication of a new conception or to reconcile it with existing beliefs. Accommodation may, thus, have to wait until some unfruitful attempts at assimilation are worked through. It rarely seems characterised by a flash of insight, in which old ideas fall away to be replaced by new visions, or as a steady logical progression from one commitment to another. Rather it involves much fumbling about and many false starts and mistakes. (pp. 221-222)

The cognitive models of learning therefore suggest that instructors, being aware of likely problems, should give their students opportunities to work through new material and should give them the time needed to come to terms with it. These models also suggest that the instructors need to monitor their students' understanding to check that an appropriate assimilation or accommodation has taken place.

Information processing models suggest that instruction should be designed to assist students to store new information in the memory by attracting the learners' attention to the relevant information and organising this information in a way that makes it easy to assimilate (Krause et al., 2007). They also suggest that only a certain amount of new material should be given at one time.

3.3 Affective factors

It would be expected that instructors who are keen for their students to gain conceptual understanding, and, it must be said, achieve good grades, will usually try to develop the best program possible for their students. However, the aims of the instructors may be thwarted by the beliefs and aims of their students, which may not coincide with theirs. Pintrich, Marx and Boyle (1993) stated:

> The assumption that students approach their classroom learning with a rational goal of making sense of the information and coordinating it with their prior conceptions may not be accurate. Students may have many social goals in the classroom context besides learning – such as making friends, finding a boyfriend or girlfriend, or impressing their peers – which can short circuit any in-depth intellectual engagement. (p. 173)

Even if the students have a desire to achieve a high grade, they may have different approaches to their learning:

> Even if the focus is on academic achievement, students may adopt different goals for or orientations to their learning. For example, it appears that a focus on mastery or learning goals can result in deeper cognitive processing on academic tasks than a focus on the self (ego-involved) or a focus on performance (grades, besting others), which seems to result in more surface processing and less overall cognitive engagement (Pintrich et al., 1993, p. 173).

Students' beliefs about intelligence also influence the way they approach their learning. For example, if students believe that intelligence is fixed they are more likely to display helpless behaviour when faced with a difficult task. If they believe that knowledge is fixed and certain, they are more likely to acquire surface learning instead of deep learning. They will also tend to regard tentative knowledge as absolute. If students believe that learning is quickly achieved, they may oversimplify information, perform poorly on tests, and be overconfident about their understanding of information. If they believe knowledge is composed of isolated facts they will have difficulties in understanding (Schommer, 1993).

In the university environment, where it is not compulsory to attend all classes, differing motivations and attitudes to learning may be reflected in attendance. Those students who are motivated only to gain the minimum requirements to pass may have irregular attendance. Some students are reluctant to be involved in whole class discussion, either because they prefer to let others make the effort of contributing, are shy, or because they have language difficulties. Some of these latter problems may be overcome when the students work in small groups. In contrast, there are students who are motivated to understand their subjects as much as possible, or who develop an interest as the semester continues.

3.4 The use of the SOLO taxonomy in assessing learning

Piaget suggested that as people grow older they have increasingly more complex ways of reasoning available to them (Krause et al., 2007). At first children learn using concrete modes, and as they grow older they are able to use increasingly abstract modes of reasoning. Biggs and Collis (1982) suggested that there are natural stages in the growth of learning of any complex material that are analogous to the developmental stages of reasoning described by Piaget. As a result Biggs and Collis developed a system for assessing learning (The SOLO Taxonomy – Structure of Observed Learning Outcomes) that takes into account the demonstration of increasingly abstract knowledge and increasing complexity of the learner's reasoning. In this system, an answer that does not address the elements of a task is considered to be "Prestructural." An answer that employs a single element of the task only is considered to be "Unistructural," whereas an answer that employs several elements in a task is considered to be "Multistructural." Those answers that create connections among the elements of a task to form an integrated whole are considered to be "Relational" (Watson & Callingham, 2003). Using the principles of the SOLO model, a form of assessment can be developed where students are not only assessed on right/wrong answers, but can be assessed on the level of sophistication of their reasoning.

With these ideas in mind, the items in the questionnaires used in this study were designed so that a hierarchical structure could be used in grading the responses to them. Therefore answers that showed a higher level of statistical reasoning were assessed with a higher score than those that showed a lower level of reasoning. For example, one of the questions asked the students to determine the probability of a coin coming up with a Head after four Tails had come up in a row, and to give an explanation for their answer. Those students who did not address the task were considered to give a Prestructural response and received a code of "0." Those students who answered that as there were "only two outcomes" the probability is 0.5 were considered to give a Unistructural response and received a code of "1." Those students who added that as the toss of coin is "not affected by the outcomes of previous tosses" the probability is 0.5 were considered to give a Multistructural response, as this response indicates that the students were considering more aspects of the overall picture. These students received a code of "2." If a student had considered the pros and cons of the competing explanations would have been considered to give a Relational response (no student did this).

Further details of the assessment models used in this study are explained in Chapter 4.

4. Literature review part III: Measurement in the social sciences

4.1 Scales used in measurement

Measurement can be defined as the assignment of numbers to objects or events according to rules (Stevens, 1946). The attributes of the things being measured determine the scale that applies the numbers, and in turn the scale determines the mathematical properties of the measurements and the statistical operations that can be applied to the measurements.

The simplest form of measurement uses a nominal scale where numerals are used only as labels. For example, a “1” may be used for a person enrolled in a Bachelor of Arts, and a “2” may be used for a Bachelor of Science. With a nominal scale numerals are used only as labels and no order is implied with these numerals. In general, the only statistic available for nominal data is the number of cases in each category. Yet even though nominal data have no implied order, nominal data such as race or gender may still be important explanatory variables (Wright & Linacre, 1989).

At the next level of complexity the labels are ordered into successive categories, which increase or decrease in status along some intended underlying variable (an “ordinal” scale). For example, a person’s level of understanding may be described as “none”, “average” or “extensive.” These ordered categories can be considered to be related to each other by a series of steps. Therefore “none” would be considered as step zero, “some” could be considered to be one step up, and “extensive” to be the next step up. This form of ordering, however, does not reflect the distances each step may require (Wright & Linacre, 1989). For example, it might be harder to go from "average" to "extensive" than from "none" to "average".

The next level of complexity involves the use of an interval scale where the intervals are equal (Stevens, 1946). Such scales are often used in the physical world and are often simple to understand. For example to measure length a particular item of interest is matched against another length that has been standardised in some way, for example, marked in centimetres. The measurement of length also involves a "ratio" scale. By this it is meant that a zero length is a true zero, as it is not possible to have a length that is less than zero, and 20 cm, for example, is twice as long as 10 cm.

The measurement of temperature also involves an interval scale but the process of measurement is indirect. In a common mercury thermometer temperature (in Celsius or Fahrenheit) is measured by the expansion of mercury in a calibrated glass tube, unlike the measurement of length where a length is lined up against a standard length. There is another important difference between these measuring scales. Whereas temperature measured on the Celsius or Fahrenheit scales uses interval scales these are not ratio scales. This is because it is possible to have a temperature below zero, and 20^{o}, for example, is not twice as hot as 10^{o} on either scale.

In the social sciences much measurement is even more indirect than that of temperature and the choice of scale is more problematic. Attributes such as intelligence cannot be measured by a simple measuring instrument, but instead are measured by a person's responses to some form of test or observation of behaviour. The data then, consist of counts of observations. This problem is further complicated by a disagreement as to what the nature of intelligence is and what it is that intelligence tests measure (Gould, 1981). Because the measurement may differ with the process or the test used to take the measurement, it is important that the test used, the scores of others being measured and the context of the test be known to the reader (Willmott & Fowles, 1974). This is in contrast to the measurement of length where if a reading of 1.2 m is given, for example, nothing else needs to be known about the process by which the measurement was made – the measure is “invariant.”

With these factors in mind the question arises as to how attributes and behaviour, where the data are in the form of counts can be measured and be placed on a scale so that comparisons between individuals or changes in individuals can be quantified. Before this issue can be addressed, however, the process of measurement itself needs to be examined.

4.2 Measurement Theory

In the 1920s, Campbell listed the logical requirements of measurement that became the basis for what is now known as the Theory of Fundamental Measurement. According to this theory, any property being scaled must be described by a quantitative variable, where a quantitative variable is one whose values are defined by a set of ordinal and additive relations (Barrett, 2003). For an ordinal scale to exist, the suitable relations must be defined within it, and the relationships on the scale must be transitive and strongly connected (Michell, 2002).

Formally, this can be described in the following manner (Barrett, 2003). In the real number system, if X, Y and Z are three values of a variable Q, then Q is ordinal if and only if:

- $X \geq Y$ and $Y \geq Z$ then $X \geq Z$ (transitive property)
- both $X \geq Y$ and $Y \geq X$ then $X = Y$ (antisymmetric property)
- either $X \geq Y$ or $Y \geq X$ (are "strongly connected" – that is, every pair can be compared)

Once it is established that the variable Q is ordinal, then to establish additivity, the following axioms must apply:

For any ordinal variable Q ($X + Y = Z$):

- $X + (Y + Z) = (X + Y) + Z$ (associative property)
- $X + Y = Y + X$ (commutative property)
- $X \geq Y$ if and only if $X + Z \geq Y + Z$ (monotonic property)
- If $X > Y$ there exists a value of Z such that $X = Y + Z$ (that is, the relation is solvable)
- $X + Y > X$ (positivity)
- There exists a natural number n such that $nX \geq Y$ (that is, there are no infinite elements or non-zero infinitesimals - The Archimedean condition)

If these apply the relation is additive and Q is a quantitative variable. The practical consequence of these axioms is that the measuring scale is capable of being ordered, and it is possible to designate less than, more than, and equality.

How can these axioms be applied to the social sciences, where physical objects are not usually measured, and many of the attributes measured are variables that are latent or unobservable? Mental attributes such as intellectual abilities, personality traits and social attitudes are often judged on answers to test items on a questionnaire or similar. For example, a person's spelling ability may be judged by the number of spelling errors in a test, but this is not "measuring" spelling (Wright, 1997). If care is not taken, a measurement made by placing subjects on a scale may just be a monotonic transformation of the observed test scores (Barrett, 2003). As Michell (2001) states:

> Psychological tests are thought to provide clues about processes underlying intellectual performance. However, just because performance on such tests possess quantitative features (e.g., test scores), it does not follow that these features reflect the workings of exclusively quantitative causes. (p. 213)

In particular, Wright (1997) points out the importance of noting that raw scores (for example test marks) are not measures. This is due to the inequality of the units and the non-linearity of the raw scores. If the test score is plotted against an ability measure, the increase in value for one more right answer is steepest where the items are most dense, towards the 50% correct point. At the extremes (towards 0% and 100%) the increase in the curve is flatter. In other words, this means it is much more difficult for a person to move from a score of 90% to 95% than from 50% to 55%. The result of this non-linearity is that any statistical method (including linear regression, factor analysis and ANOVA) that uses raw scores (including those generated by Likert Scales) will be affected by bias (Wright, 1997).

In 1964, Luce and Tukey published a set of axioms that showed that if two or more variables (for example *A* and *X*) are non-interactively related to a third variable (*P*) then differences between the elements in *A* and *X* may be equated by their effects on *P*, a process known as "conjoint additivity" (Michell, 1994). By this process latent or unobservable variables such as intelligence can be measured on an interval scale (Barrett, 2003).

For these variables to show conjoint additivity they must show the condition of solvability; that is, for any a in A and x,y in X there must exist b in A such that $(a,x) = (b,y)$, where (a,x) and (b,y) denote the elements in P determined by the conjoined elements in A and X. Similarly, for any a and b in A and x in X there must exist y in X such that $(a,x) = (b,y)$ (Michell, 1994).

Second, there must be no differences on either A or X that are infinitesimally small or infinitely large relative to the others (the Archimedean condition), and third, the sums of differences on one factor must retain whatever parity the part differences have with differences on the other factor. That is, for any a_1,a_2 and a_3 in A, x_1, x_2 and x_3 in X,

if $(a_2,x_1) \geq (a_1,x_2)$ and $(a_3,x_2) \geq (a_2,x_3)$ then $(a_3,x_1) \geq (a_1,x_3)$ (Michell, 1994).

Any measure that is to be used in a statistical method needs to be additive (or conjoint additive) and linear. The measure should also be invariant, that is, the values attributed to any variables in the measurement system should be independent of the measurement instrument used (Bond & Fox, 2007). The measurement should also be subject to item free and sample free calibration and be consistent, in that if a person answers a more difficult question correctly (or endorses a more extreme statement), then all of the less difficult items (or less extreme statements) should be answered correctly (Wright, 1997). Finally, the test used should be unidimensional, in that it describes only one attribute of the object measured. One of the processes by which this can be achieved is described in Sections 4.4 and 4.5.

4.3 Should the social sciences use measurement?

"Our main purpose is briefly stated in the subtitle of the new journal, *Psychometrika*, namely, to encourage the development of psychology as a quantitative rational science." Thurstone (1937)

Given that much of what is studied in the social sciences is latent, and therefore not directly measurable, it is reasonable to ask why mathematical procedures should be used at all – why the processes described in the previous section should even matter. It could be argued that human attributes such as attitude and behaviour should remain in the qualitative realm.

One motivation for the quantification of human attributes came about from the view that true, scientific knowledge must be measurable and amenable to mathematical analysis. It was also felt that quantification led to a more thorough, rigorous knowledge such as that found in the physical sciences. One such argument was put by Thorndike (as cited in Barrett, 2003, p. 426), "Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quality as well as its quantity."

Another reason for using measurement in the social sciences is that in everyday discourse values and attitudes are readily quantified. Thurstone (1928) stated: "The main argument so far has been to show that since in ordinary conversation we readily and understandably describe individuals as more and less pacifistic or more and less militaristic in attitude, we may frankly represent this linearity in a unidimensional scale" (p. 538).

One motivating factor in the search for quantification of human attributes and behaviour was that if these could be measured, then it would be possible to quantify changes in individuals across time and to make comparisons between groups. An early example of such a quantitative comparison is found in the work of Trabue who studied the development of language ability (measured by the "completion-test") from Grade 2 students to college graduates (Thurstone, 1925). Intelligence tests, first introduced by Binet to identify students who needed learning support, also provided a stimulus to quantify human attributes (Gould, 1981).

Finally, and possibly most importantly, quantification was considered desirable because it enables statistical inference (Wright, 1997). This desire for inference has, unfortunately, led to the use of statistical tests on data unsuitable for these tests. Such misuse may be ignored or even unknown by the readers of the research, and their conclusions may give the impression of legitimacy just because the tests were carried out (Merbitz, Morris, & Grip, 1989).

If the social sciences are to use valid quantification, forms of measurement need to be used that comply with the theory of fundamental measurement. The search for such methods has been continuing over the last century. One response to this need has been the development of item response theory, which is discussed in the next section.

4.4 Item Response Theory

When giving tests to measure human attributes (or in fact in the making of any measurement), experimental error is unavoidable. Unless adequate control is provided for the "error," or "nuisance," variables, valid inferences cannot be made (van der Linden & Hambleton, 1997).

In general, there are three ways of coping with this experimental error. These are matching or standardisation, randomisation, and statistical adjustment. If the conditions are matched or standardised, then the subjects operate under the same levels of error or nuisance variables, and thus the effects of these variables cannot explain any difference in experimental outcomes. Unfortunately this technique has restricted generalisability as the experimental results obtained will hold only in similar conditions.

Randomisation is based on the principle that if error variables cannot be manipulated to create the same effects, random selection of conditions guarantees that these effects, on average, can be expected to be the same, and will therefore not have a systematic influence of the responses.

In contrast, statistical adjustment is a technique of post hoc control. Mathematical models are used to make adjustments to test (raw) scores by processes based on item response theory (IRT). Much of the work in statistical adjustment can be related back to Thorndike, who was one of the early pioneers in the search for methodology to transform the counting of observations to measurement. Being aware of the non-linearity of raw scores, he searched for a way of obtaining a scale that could be applied to compare ability over different age and grade groups. Thurstone (1928) has described Thorndike's method in a study on completion tests of language scales. In this study, Thorndike calculated the proportion of correct answers for each grade, and then graded the difficulty of each item using the deviation from the mean for all the correct answers for that grade. He then devised a common educational scale across all grades by adding to each scale value the mean value for the grade. This process had the disadvantage, however, that it assumed that the dispersion of abilities to be normally distributed and having equal dispersion across all grades. Using the same data, Thurstone found that in fact the dispersion increased as the students became older.

Thurstone (1925), therefore, devised a method in which it was still assumed that the distributions of ability in each age/grade group followed a Normal distribution, but that allowed for variability in the dispersion around the mean for each group. For each age group he placed each test item on the Normal curve such that the percentage to the right of this point represented the percentage of students who answered the question correctly. From these placements on the scale Thurstone then presented a formula to calculate the standard deviation. This method allowed for the item difficulties to be described in two ways, by age/grade level using the *z*-score, or by finding the 50% mark for each age/grade level, so comparison between the different age/grade levels could be made. Thus Thurstone's methodology gave a rating for difficulty for each item that allowed each group to have its own level of dispersion. However it still had the disadvantage of assuming a Normal distribution for the pattern of response, even when the data suggested otherwise.

Modern IRT methods are similar to Thurstone's work in that the models place item locations and respondent values as points on a scale of the quantitative variable (Bock, 1997). However, they differ in that the continuous variable controlling the probability of a correct response is assumed to be latent variable such as ability, and not an independently measured variable such as chronological age (Bock, 1997).

In the 1940s the Normal ogive model (based on the work of Thurstone) was introduced. In this model, the probability of detecting a stimulus was written as a function of the strength of the stimulus. If the probability of success of a person with ability *θ* on item *i* is represented by $P_i(\theta)$, this model has the form:

$$P_i = \int_{-\infty}^{a_i(\theta-b_i)} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$$

The difficulty parameter b_i is the point on an ability scale where the person has a probability of success on the item of 0.50. The value of a_i is proportional to the slope of the tangent to the response function at this 0.50 point. As a_i increases, the ability of the item to discriminate between persons with varying abilities also increases (van der Linden & Hambleton, 1997). With this model the ability variable is assumed to be normally distributed (Bock, 1997).

IRT was developed further by Lazarsfield in the 1950s who was studying the adjustment patterns of returned soldiers. He administered multiple item attitude questionnaires and then used dichotomously classified responses (0,1) to develop probabilistic models that were dependent on the conditional probability of the response pattern. Of importance was his introduction of the idea of “local independence,” by which he meant the statistical independence of one person’s response from another person’s response having the same value of the underlying latent variable. Although he used his model to assign respondents to classes, he did not use the model to estimate values of the parameter for individuals (Bock, 1997).

Further developments in item response theory were made in the 1980s by Birnbaum and Lord who substituted the ogive model with a logistic item response model. This logistic model is in the form:

$$p = \frac{1}{1 + \exp(-z)}$$

In this model, *z* has the form $z = a(\theta\text{-}b),$ where θ is the latent ability variable.

One advantage of this model was that Maximum Likelihood Estimates (MLE) of a person's scale score became possible, thus allowing a judgement as to the precision of the MLE estimates. This model, however, still gave no general solution for how to estimate the parameters of the item response models using samples of test data (Bock, 1997).

A great step forward occurred in 1960 when Rasch presented a new method of analysis that gave independent measures of the difficulty of the item and the ability of the person. Rasch recognised that as long as the test items belong to a calibrated set of items that define the variable under study, then no matter who responds to the test, each test or rating scale item will have a constant level of difficulty, and that each person will have a constant level of ability regardless of which particular test items are encountered (Wright & Linacre, 1989). Therefore:

> A person having a greater ability than another person should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another means that for any person the probability of solving the second item is the greater one (Rasch, 1960, as cited in Bond & Fox, 2007, p.10).

Rasch also recognised that the interaction between a person and an item cannot be fully predetermined and will also involve an additional, unpredictable component. With this knowledge Rasch determined a mathematical model that converts observed counts, on an ordinal scale, into a unidimensional, linear scale. In so doing, the ordinal scale was transformed into an interval where the size of each step could be quantified and compared to the size of the other steps. A table of expected probabilities is produced that determines the likelihood of each person with a specified ability correctly answering an item with a specified level of difficulty. For items of a level of difficulty above a person's level of ability, this probability of success will decrease as the difference between the ability and difficulty increases. For items of a level of difficulty below a person's level of ability, this probability of success will increase as the difference between the ability and difficulty increases (Bond & Fox, 2007).

An advantage of the Rasch model is that sufficient statistics are used to estimate the parameters. With these sufficient statistics, consistent maximum likelihood estimates of the parameters can be made that are independent of other parameters, thus resulting in an invariant measure. Another advantage is that since this model uses individuals, rather than groups, no assumptions about the distribution of abilities need to be made (Wright, 1997). In addition, the measures that result from the Rasch model show conjoint additivity (Bond & Fox, 2007, see Section 4.2).

Because for each Rasch analysis the mean of the item difficulty is calibrated as zero, Rasch measurements of item difficulty and person ability are at the interval, rather than ratio, level of measurement. In addition, if longitudinal measures are made, it is possible to ‚anchor' the scale by the inclusion of items of the same difficulty (Bond & Fox, 2007).

In summary, Rasch analysis provides the social sciences a method of measurement of latent variables that fulfil the requirements of fundamental measurement theory, including conjoint additivity. The mathematics behind this model is described in the next section.

4.5 The Mathematics of the Rasch Model

The Rasch model will be introduced with the dichotomous model which is used when there are only two responses. From this model the Partial Credit Model will be developed, which is used when responses are graded onto a scale which takes into account partially correct answers.

4.5.1 The Dichotomous model

The dichotomous model is used when the answers are classified either as correct or incorrect only. Usually a correct answer is coded as "1" and an incorrect answer is coded as "0." To perform the Rasch analysis, firstly the percentage of correct answers is calculated for each person, and then for each item. These raw score totals are sufficient statistics for estimating the person ability *(θ_j)* and the item difficulty *(δ_i)*.

To estimate the person ability θ_j, the odds of success are first computed by calculating the ratio of each person's percentage correct *(p)* over the percentage incorrect *(1-p)*, and then the natural log of these odds is calculated. A similar process is used for the items. These log odds ratios, known as logits, are then placed on a scale of intervals of one logit. In Rasch analysis, the average logit is arbitrarily set at 0, so that positive logits indicate a higher than average probability and negative logits indicate a lower than average probability.

In the calculations the person estimates are firstly constrained while the item estimates are calculated. These first estimates are then used in an iterative process to produce an internally consistent set of item and person parameters. The iteration process continues until the maximum difference in item and person values during successive iterations converges to a preset value. When this is complete, the difference in person ability and item difficulty values will produce the Rasch probabilities of success, and as a result, the ordinal level data are transformed into interval level data that is suitable for inference (Masters & Wright, 1997).

Given θ_j and δ_i the function (*f*) expressing the probability of a successful response can be expressed as:

$$\frac{\mathrm{P_{ij1}}}{\mathrm{P_{ij0}} + \mathrm{P_{ij1}}} = \frac{\exp\,(\theta_j - \delta_i)}{1 + \exp\,(\mathrm{\theta_j} - \mathrm{\delta_i})},$$

where P_{ij1} is the probability of person *j* scoring 1 on item *i*, P_{ij0} is the probability of person *j* scoring 0, θ_j is the ability of person *j*, and δ_i is the difficulty of item *i* . This difficulty defines the location on the measurement variable at which a score of 1 on item *i* is as likely as a score of 0.

From this equation it follows that when item difficulties are greater than person abilities (negative $\theta_j - \delta_i$ values) persons have a lower than 50% chance of correctly answering the item. When item difficulties are lower than person abilities (positive $\theta_j - \delta_i$ values), persons have a higher than 50% probability of correctly answering the item (Bond & Fox, 2007).

Using conditional MLE the parameters for the items can be estimated without knowing the parameters for the persons, and vice versa. That is, person free measures and item free calibrations are obtained. Therefore abstract measures that transcend specific persons' responses to specific items at a specific time are achieved (Masters & Wright, 1997).

4.5.2 The Partial Credit Model

The Partial Credit Model (Masters, 1982) is developed from the dichotomous model (van der Linden & Hambleton, 1997). It can be used in any situation in which performances on an item or an assessment criterion are recorded in two or more ordered categories, and where there is an aim of obtaining measures on a latent unidimensional variable (Masters & Wright, 1997). These ordered categories can be considered as a series of steps. In this study the Partial Credit Model was used because students' answers were graded on a scale according to the level of understanding shown by their answers, rather than simply graded as right or wrong.

In the Partial Credit Model, the probability of a student responding in the x^{th} category as opposed to the $x\text{-}1^{th}$ category is dependent on the difficulty of the x^{th} level. The Partial Credit Model considers the number of steps a person has made beyond the lowest level of performance (Callingham & Watson, 2005). The derivation of the mathematical model follows.

For an item with more than two response categories, and when there is increasing ability of the person being tested, there comes a point where a lower score becomes less likely because the next higher score becomes more likely. That is, as ability increases, there is a point where a person is more likely to score a '1' than a „0', and then a point where a person is more likely to score a '2' instead of a „1'. It follows from the intended order $0<1<2, \ldots, <m_i$ of a set of categories that the conditional probability of scoring x rather than $x\text{-}1$ on an item should increase monotonically throughout the ability range. This is modelled as:

$$\frac{P_{ijx}}{P_{ijx-1}+P_{ijx}}=\frac{\exp\,(\theta_j-\delta_{ix})}{1+exp(\theta_j-\delta_{ix})},x=1,2\ldots,m_i$$

Where P_{ijx} is the probability of person *j* scoring *x* on item *i*, P_{ijx-1} is the probability of person j scoring *x-1,* θ_j is the ability of person *j*, and δ_{ix} is an item parameter governing the probability of scoring *x* rather than *x-1* on item *i.* These parameters are estimated by an iterative, conditional or joint maximum likelihood procedure (Masters, 1982).

How well the model fits the data, and fits on a unidimensional scale, can be assessed by the "goodness of fit." Goodness of fit in a Rasch model is applied to item fit, person fit, and global fit, where the latter indicates the overall fit of a set of data to the Partial Credit Model.

Item and person fit are assessed with the weighted ("infit") and the unweighted ("outfit") mean square procedures. If a person *j* with ability θ_j responds to item *i*, the person's response x_{ij} takes a value between 0 and the maximum possible score on item *i,* m_i. If the probability of person *i* scoring *h* on item *i* is denoted by P_{ijh}, then:

The expected value of x_{ij} is:

$$\mathrm{E_{ij}}=\sum_{k=0}^{m_j}\mathrm{kP_{ijk}}$$

The variance is:

$$\mathrm{W_{ij}}=\sum_{k=0}^{m_j}(\mathrm{k}-\mathrm{E_{ij}})^2\mathrm{P_{ijk}}$$

The standardised difference between person *j*'s observed and expected response to item *i* is:

$$\mathrm{z_{ij} = (x_{ij} - E_{ij}) / \sqrt{W_{ij}}}$$

The outfit mean square for each item *i* is then:

$$\mathrm{u_i = \sum_{j}^{N} z_{ij}^2 / N}$$

For each person *j*:

$$\mathrm{u_j = \sum_{i}^{N} z_{ij}^2 / N}$$

These outfit statistics can be oversensitive to outliers. Therefore the "infit" is also used, which is weighted.

The weighted mean square for each item *i* is then:

$$\mathrm{v_i = \sum_{j}^{N} W_{ij}\, z_{ij}^2 / \sum_{j}^{N} W_{ij}}$$

For each person *j*:

$$\mathrm{v_j = \sum_{i}^{N} W_{ij} z_{ij}^2 / \sum_{i}^{N} W_{ij}}$$

(Masters & Wright, 1997).

The expected value of these statistics is +1. An infit or outfit mean square value of (1 + *x*) indicates 100*x*% more variation between the observed and the model predicted response than if the data and the model were perfectly compatible. An outfit square of less than one indicates $100x\%$ less variation than predicted by the model (Bond & Fox, 2007).

4.5.3 How Rasch analysis was used in this study

In this study, each participant was asked to answer two questionnaires, one on entry to the unit, and the other at the end of the unit. The participants were also asked to allow the answers to one of the tests used for formal assessment to be included in the study. The questionnaires were analysed quantitatively using the Rasch partial credit model. Using this model, independent assessments of the difficulty of the items and the students' ability were gained. It was planned that a similar analysis would be carried out on responses to the test items, but it was found that these did not fall on a unidimensional scale and were therefore unsuitable for Rasch analysis. The partial credit model was used because each answer in the questionnaires was not merely graded as correct or incorrect, but was rated according to the level of sophistication of reasoning used in the answer.

A judgement needs to be made as to which items fit on the unidimensional continuum. To decide whether or not an item fits on this continuum the "fit" statistics are used. These fit statistics are the infit mean square (infit MNSQ) and the outfit mean square (outfit MNSQ), where the mean square is the mean value for the squared residuals for an item. The infit MNSQ gives more weight to the performances of persons with abilities close to the item's difficulty. If this statistic is satisfactory, then it can be concluded that a single construct is being measured. The outfit MNSQ is more sensitive to the influence of outlying scores (Bond & Fox, 2007). Both of these statistics, the infit and outfit MNSQ, can be transformed to a *t* statistic known as the standardised *z* value (ZSTD) that follows approximately the *t* or standardised Normal distribution when the items fit the model's expectations (Hsueh, Wang, Sheu, & Hsieh, 2004). Items with infit and outfit ZSTD values of greater than ±2 are considered to be poor fitting and are examined further. They may be removed from the analysis or modified. Winsteps (www.winsteps.com.htm), a program dedicated to Rasch analysis, displays these standardised values in "bubble chart" by which the fit of the items and persons can be visualised. An example can be seen in Figure 4.5.3.1.

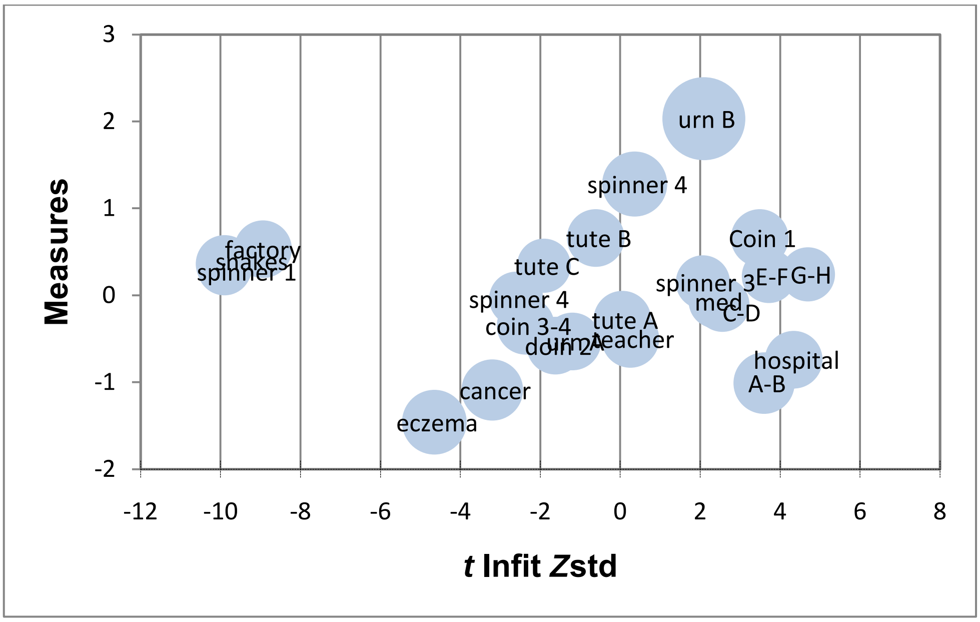


Figure 4.5.3.1. An example of a "bubble chart" for a Rasch analysis.

Once the items that were on the unidimensional scale were selected, they were placed in order of difficulty and then grouped according to this level of difficulty. Each group was then analysed by the cognitive demands were made by the questions in this group, and then a set of criteria were determined that characterised the qualitative changes demanded for successful answers for each group (Callingham & Watson, 2005).

The students were then divided into groups according to their ability, using the groupings obtained from the item analysis. For example, as the highest group of items in the first questionnaire had ratings from 0.7 logits and above, the students with an ability rating of 0.7 logits and above were placed in the highest ability group. These groupings were then used to qualitatively compare the answers across groupings. The abilities of the students were also compared between the first and second questionnaires, to see if there was any improvement overall, and to see if there were any significant differences in the ability between the semesters of the study for each student cohort. This comparison was made possible because if a number of test items are common to both tests, the Rasch analysis enables these common items to be anchored to first analysis' difficulty ratings. A correlation was also made between the initial ability of the students and their final scores for the Data Handling and Statistics unit, to see if it was possible to predict a student's success from their initial ability ratings.

One of the statistics generated by Rasch analysis via the Partial Credit Model is the Rasch-Thurstone threshold. This threshold gives the item difficulty on a scale at a point at which the lower and higher categories of an answer have equal probability of being observed. An example is shown in Figure 4.5.3.2 – an item person map. In this figure the items are listed with their difficulty ratings, and the persons, indicated by an „X' are indicated by their ability ratings. There are two persons who have the same rating as the item "tute A .1". Both have a 50% chance of gaining a score of zero, and a 50% chance of gaining a score of one on this item. Those persons who have a higher ability score, and are therefore above these two on the chart, have a greater than 50% chance of scoring one on this item. Persons on the chart have a greater than 50% probability of obtaining the indicated score for any item below them on the scale (the lower the item the more likely they are to gain the indicated score) and a less than 50% probability of obtaining the indicated score for any item above them on the scale (the higher the item the less likely they are to gain the indicated score).

```
                    |
                    |
                    |   urnB    .2
                    |
2     X             |
                    |
                    |
      X             |   spinner4 .2
                    |
                    |   urnB    .1
      XX            |
                    |
      XXXXX         |
1                   |   tuteB .2      coin1   .2
      XXXX          |
                    |
      XXXXX         |   spinner4  .1  tuteC   .2
      XX            |
                    |   spinner3 .2
      XXXXXXXXXXX   |
      XXXXXX        |   med     .2
                    |
      XXXXXXX       |   coin1   .1    tuteA   .2    tuteB  .1
0     XXXXXXXXXX    |   coin3-4  .2
                    |
      XXXXXXX       |   tuteC   .1    coin2   .2    teacher  .2
                    |
      XXX           |   urnA    .2
                    |   spinner3 .1
      XXX           |
      XX            |   med     .1
                    |
      XX            |   tuteA   .1
      X             |   coin3-4 .1
-1                  |   teacher .1    cancer  .2
      X             |   coin2   .1    urnA  .1
                    |
      X             |
                    |
                    |
                    |
                    |
                    |
                    |   cancer  .1
-2                  |
                    |
                    |
      X             |
```

Figure 4.5.3.2 . An example of an Item and person map.

5. The use of computer technology in statistics education

5.1 Introduction

Computers are widely used in statistics education to perform routine data analysis tasks and thus relieve the student from tedious calculations. In this role the computer is just an advanced calculator. Although this is helpful, the use of computers in this way does not enhance learning of abstract concepts (Mills, 2002).

In contrast, the use of computers for simulations provides a model of a system or process that demonstrates scientific or other principles. Computer simulations have been used in a variety of contexts such as economics, decision theory, Newtonian mechanics, physics, and mathematics (de Jong et al., 1999). In statistics, computer simulations allow students to do what they cannot do in reality, that is, repeat a study many times (Burrill, 2002). Therefore students can observe for themselves the extent and effect of the variability that is inherent in, and complicates, any statistical analysis.

5.2 Discovery Learning and Simulation

Constructivist theories of learning are based on the assumption that students do not merely take in information, but actively build knowledge for themselves in the light of their previous knowledge. This understanding has led some authors to conclude that knowledge students construct on their own is more valuable than knowledge that has been shown, demonstrated or explained by a teacher (Klahr & Nigam, 2004).

It is argued that one way students can be prompted to construct their own knowledge is to use some form of discovery learning, where the students are actively engaged in activities that will lead them to discover a principle for themselves, instead of just being told by the instructor. In the sciences students can be given questions in which they need to form testable hypotheses, and then plan and execute an experiment (de Jong & van Jooligen, 1998). In statistics, simulation provides one of the means by which an experiment can be carried out. For example, using a computer enables students to take hundreds of random samples, and examine the population of the means of these samples.

The literature demonstrates, however, that discovery learning without guidance is not necessarily successful. Students are not always proficient in noticing regularities in the data and may also be poor at interpreting graphs. Because students may be resistant to conceptual change, they may not be able to adapt hypotheses even when the data are contradictory (de Jong & van Jooligen, 1998). In statistics education, it has also been found that students can apparently be successful in their use of the computer simulation software yet still demonstrate a lack of understanding and hold misconceptions afterwards (delMas, Garfield, & Chance, 1999). Without careful guidance, students may even continue to build on previous misconceptions (Mills, 2002).

For conceptual change to take place, or for new knowledge to be assimilated, the literature suggests that guided discovery learning or a combination of discovery learning and direct instruction is needed (delMas et al., 1999; Lane & Peres, 2006). Guided discovery learning involves giving the students a structure, a series of questions that guide the learner to a predetermined goal (Lane & Peres, 2006).

Lane and Peres suggest that knowledge acquisition is improved if students are asked to make predictions as to what will happen before a simulation: the “query first – answer later” method. By this means students are forced to confront discrepancies between what they expect and what actually occurs and therefore are encouraged to make accommodations of their current schemas (Garfield & Ahlgren, 1988; Hardiman, Pollatsek, & Well, 1986; Mills, 2002; Posner, Strike, Hewson, & Gertzog, 1982). If the students are not put in a position where they are made to face these discrepancies, they may look for, and find evidence for, their own previous knowledge, which may not be accurate (Shaughnessy, 1992). In this study for example, students drew a histogram of the means of the random samples they had taken from a normally distributed population. They were then asked to predict the shape of the distribution of sample means taken from a Uniform, then binomial population, and then take random samples and draw histograms of the means from these populations. Most students thought that these distributions would be the same as the parent populations and were therefore surprised by their results.

5.3 Simulation in statistics

Statistical inference is based on probability. One of the ways to look at a probability is in terms of a long term frequency. For example, the probability of getting a “3” on a throw of a die is one in six. This does not mean that one “3” will turn up once in the next six throws, but if the die were thrown a very large number of times approximately one sixth of the numbers will be a “3”. According to Chance and Rossman (2006) the application of the long term frequency idea of probability to inferential statistics, that is, to consider what would happen if a random process were repeated many times, is difficult for many people to understand. The advantage of computer simulation is that many repeated trials can be carried out quickly.

Another advantage of computer simulation is that it is possible to carry out trials of problems that may be impossible to do in reality. For example, a simple simulation can determine what would happen to the ratio of male and female births if China's "one child" policy was replaced with a "stop when a son in born" policy. Other advantages include that the feedback is immediate and it is possible to re-do the simulation quickly. In *Microsoft Excel*, for example, any simulation based on random numbers can be repeated with the press of one button. It is also possible to link graphical representations to numerical representations (Snir, Smith, & Grosslight, 1995) and in *Microsoft Excel* the graphical representations update automatically as the data are changed.

Computer simulations can also be used to demonstrate hypothesis tests. For example, Erickson (2006), described a procedure whereby the hypothesis test for the equality of two population means can be demonstrated. He gave an example where data had been collected to test the effect of a new fertiliser on plant growth. The null hypothesis was that the fertiliser has no effect on growth. If this was the case, then group membership would have no influence on height. Using simulation, membership of the two groups can be repeatedly randomly shuffled and the difference in mean heights calculated. These differences can then be compared with that of the original grouping. If the observed difference in mean height of the original grouping is greater than that obtained when the plants are randomly shuffled, then it can be concluded that there was something special about the original grouping; that is, the fertiliser does make a difference. An example of the results of such a process is shown in Figure 5.3.1.

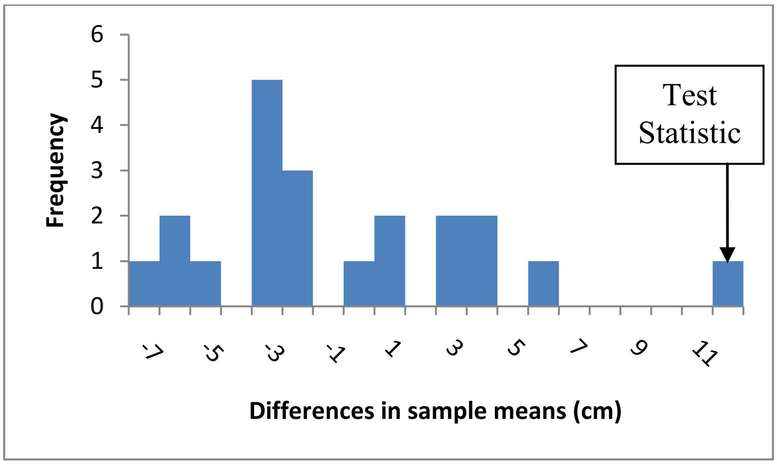


Figure 5.3.1. Example of the results obtained when data are shuffled at random, compared with the observed test statistic.

With this process the distribution of statistics and the comparison of the test statistic with the distribution are modelled visually. Students can, therefore, draw a subjective conclusion as to the likelihood of the test statistic, if the null hypothesis were true, before calculating the numerical *P*-value (Erickson, 2006).

Simulation, however, does not automatically produce understanding or produce sound statistical reasoning. For example, Lipson (2002) found that the sampling software used in her study helped in the understanding of sampling distributions but then the students failed to link the sampling distribution to hypothesis testing and estimation. Lipson, Kokonis and Francis (2003) found that the ways students interacted with the software were complex and often difficult to understand by the instructors with statistical expertise.

Simulation, then, is one tool to help students understand the role of variation in statistical decision making, and can be used to challenge students' assumptions and preconceived notions. However, simulation does not guarantee good understanding.

5.4 How computers were used in this study

Computers in this study were used in three ways. The first was to provide students with a relatively quick way of performing the often tedious calculations involved in statistical analysis. Secondly, computers were used to demonstrate simple principles such as the effect of errors and extreme values on the values of the mean and median. Thirdly, computers were used for simulations to provide guided discovery learning. Neither of the last two practices had been used in the teaching of the unit before this study. Each simulation had a predict-test-evaluate format. That is, students were given a scenario and asked for a prediction. This was followed by a simulation of the scenario after which students were asked to re-evaluate their previous prediction. The details of the computer demonstrations and simulations are found in Appendix D.

6. The Study Design

6.1 Introduction

In the physical and biological sciences, it is possible to design experiments in which confounding factors are controlled, treatments are randomly allocated and given in known amounts, and a control (objects of study without the treatment) is included. These experiments are carried out in a positivist environment, in that the only phenomena of interest are those that can be observed and measured. The data are subject to numerical analysis, and causal relationships are determined by altering the variables one by one. These experiments are also replicable, in that they can be repeated in identical circumstances. In education, however, experiments with these conditions are not possible. Each educational research setting, even with the same materials and lesson plan, is a unique interaction between the researchers, the instructors and the students. As a consequence, even if random allocation is used to allocate students to different learning environments, it is impossible to control for all the variables. Therefore educational experiments are not replicable, and if judged by the standards used in the physical and biological sciences, are not scientific. It is necessary, therefore, for other methods to be found for carrying out educational research other than by a traditional scientific experiment. The following sections describe some current research designs in education and their potential for providing scientific knowledge. This is then followed by a description of the design of this study, the instruments used in the study, and the details of the analyses used to assess these instruments.

6.2 Research Designs in Education

Educational researchers have to make decisions about the nature of the research design, and, if researching in the classroom, the relationship between the researcher and the instructor. The research could be performed by a researcher who is not part of the institution where the study is occurring and is thus separate from the usual instructional environment. For example, a researcher might enter an institution to carry out work with the students separate from their usual instruction. Because such a researcher would not have preconceptions of the students, this form of research can have the advantage of objectivity. Other advantages include that the researcher may also be able to study concepts that a busy instructor may not have time to examine, and may not be able to investigate because of constraints in the curriculum. If care is not taken, however, this form of research may be irrelevant to the usual instructor's needs or even invalid owing to lack of knowledge or understanding of the social situation (Hammersley, 1993).

An alternative is for the research to be performed by the researcher in collaboration with the usual instructor. The researcher might enter the usual classroom environment and participate in the students' instruction. This allows for objectivity from the researcher as well as input from the usual instructor who has knowledge of the setting and the participants that an outsider cannot have (Hammersley, 1993). This collaborative setting is the form of research recommended by the National Research Council (NRC) (2002).

A third option is for the research to be performed by the instructor alone, as occurs in action research. This has the advantage that the researcher has the knowledge base of the setting and participants, but can have the disadvantage of a lack of objectivity resulting from not being able to see the phenomenon in its wider context (Hammersley, 1993).

The study design itself then needs to be addressed. Three research paradigms that are in use in educational research are the qualitative research paradigm, the scientific research paradigm and the critical theoretic research paradigm (Ernst, 1998).

The qualitative research paradigm is concerned with human understanding and interpretation and uses methods such as ethnographic studies and case studies. These studies are interested in the nature of social phenomena, and tend to produce unstructured data and often describe a small number of cases in detail (Atkinson & Hammersley, 1998). The critical theoretic research paradigm aims to understand a phenomenon with the aim of promoting social and institutional change. The current study was not carried out within either of these paradigms, and therefore these are not examined further.

The scientific research paradigm uses the positivist approach and is concerned with prediction, objectivity, replicability and the discovery of scientific generalisations (Ernest, 1998). Therefore the data are structured and quantitative. However, it has been argued (in Section 6.1) that replicability is not possible in a classroom context. How then, can scientific research be carried out in education? This issue is discussed in the next section.

6.2.1 Scientific Research in Education

In the previous sections it was claimed that educational research cannot always be performed in controlled conditions, or use randomised treatments, and cannot be replicated. This lack of replicability can lead to difficulty in assigning causal effects (Posner et al., 1982). To complicate matters further, the investigator is often a participant in the research, and the data may not be quantitative. As a consequence, the level of certainty that applies to research in the social sciences is lower than in the physical sciences (NRC, 2002).

If the positivist view is taken, that only quantitative measures, within strictly controlled conditions, contribute to true knowledge (NRC, 2002), it will then be assumed that other forms of knowledge are not genuine, or at the very least, inferior. Consequently any education research, with its lack of strictly controlled conditions, will be considered to be inferior.

Since the beginning of research in education, the validity of the methodologies of the social sciences compared with those used in the physical sciences have been constantly questioned (NRC, 2002). Despite the continuing debate as to its validity, educational research continued to develop in tandem with the development of models of human behaviour (NRC, 2002).

It has been argued that if certain principles are followed, educational research can indeed provide genuine and even scientific knowledge. This view has resulted in the development of the “design study,” or the “design experiment.” Design experiments are created with the aim of “seek[ing] to trace the evolution of learning in complex, messy classrooms and schools, test and build theories of teaching and learning, and produce instructional tools that survive the challenges of everyday practice” (Shavelson, Phillips, Towne, & Feuer, 2003, p. 25).

The question is then raised concerning how these aims can be achieved using research that is in some way “scientific.” According to the Committee on Scientific Principles for Educational Research of the NRC (2002) a scientific study should first pose significant questions that reflect understanding of the relevant theoretical, methodological and empirical work that has come before, and use empirical means that are based on observation (NRC, 2002). The study should also allow direct investigation of the questions of interest with methods that are appropriate and effective, and provide a "logical chain of reasoning from evidence to theory and back again that is coherent, shareable, and persuasive to the sceptical reader" (NRC, 2002, p. 4). The findings should also be generalisable across studies, and be available for professional scrutiny and critique. The authors claim that science progresses both by proposal of new theories and by elimination of theories that have been refuted by newly acquired evidence.[1] It is therefore proposed that scientific research in education should propose hypotheses, or conjectures that are stated in "clear, unambiguous, and empirically testable terms" (NRC, 2002, p. 18).

The methodology of design research is further described by Cobb, Confrey, diSessa, Lehrer, and Schauble (2003). Researchers create the conditions for developing new theories while allowing for the possibility that these theories may be refuted (cf. with Karl Popper, Section 2.3.1) by modifying classroom settings, procedures, and instructional artefacts (Shavelson et al., 2003). As a result, design experiments are prospective, in that designs are implemented with a hypothesised learning process in mind that can be exposed to scrutiny. They are also reflective, in that during the experiment conjectures can be confirmed, refined or refuted. In the latter case alternatives may be generated. This feature of design experiments

[1] That science progresses by this "neat" fashion is a matter of debate. Scientific discovery can be made by flashes of insight and other non-conventional means. Scientific theories may be around for years before they are accepted by the scientific community, as was tectonic plate theory.

leads to them being inherently cyclical in nature. As hypotheses or conjectures are refined or refuted, the design is adjusted so that the experiment may recommence; that is, the research is carried out with cycles of "design, enactment, and analysis" (Baumgartner et al., 2003, p. 6). Design experiments are also collaborative, in that practitioners (instructors for example) work together with researchers. In this way both the researchers' aims and the instructors' knowledge of the local context may be considered.

In summary, in a design experiment, before each teaching episode a set of hypotheses are proposed and a sequence of situations planned to test these hypotheses. These hypotheses are then confirmed, refined, or refuted (Steffe, Thompson, & Von Glaserfield, 2001). It is this feature of generating testable hypotheses that leads to the claim that design based research can generate causal explanations and therefore result in "scientific" knowledge such as obtained in the more traditional scientific experiments (Baumgartner et al., 2003).

6.2.2 How "Scientific" Does Knowledge Have To Be?

From the discussion in the previous sections it is apparent that what constitutes "scientific" knowledge varies according to a person's point of view. If one is a positivist, all information that is not measureable is not scientific and consequently not valid. It can be argued that by aiming for "scientific" knowledge the educational researcher is, by implication, subscribing to the view that other forms of knowledge are inferior, and may be asking too much of the type of data that results. It can be argued that if the educational research provides knowledge so that students' educational experiences might be improved, or produces "Instructional tools that survive the challenges of everyday practice" (Shavelson et al., 2003, p. 25), then the research is valid. Any research should be appreciated for the value of the knowledge it produces, whether “scientific” or not.

It was stated earlier that design experiments are collaborative. Another form of research frequently used in education, with similarities to the design research methodology, but is not collaborative, is known as Action Research. This is described in the next section.

6.2.3 The Action Research Method of educational research

Action research occurs when practitioners carry out research in their own environments. In the education context, “Action research is any systematic enquiry conducted by teacher researchers, principals, school counsellors, or any other stakeholders in the teaching/learning environment to gather information about how their particular schools operate, how they teach, and how well their students learn” (Mills, 2007, p. 5). Because the research is practitioner based it is insider research; that is, the researcher will influence what is happening (McNiff, Lomax, & Whitehead, 2003).

Action research has had a varied history that is reflected in the various philosophical/ethical views in today’s literature on the topic. In the USA, an early proponent of action research was Kurt Lewin, who “focussed on understanding and changing human actions, often around issues of reducing prejudice and increasing democratic behaviours” (Noffke, 1994, p. 10). As a result, much of action research in the USA has been aligned with the progressive education movement. In Australia, action research arose during a time when there was a shift towards collaborative curriculum planning and action research was seen as part of the development of a more participatory education system (Noffke, 1994). In Britain, action research emerged in the 1960s and 1970s as part of a shift away from a central curriculum development that predetermined what is to be learned by students, toward curricula that emphasised the students' own search for meaning (Hammersley, 1993; Noffke, 1994). One of the major influences on action research was Stenhouse, who first, advocated inquiry learning, and second, and most important in this context, emphasised the role of the teacher as a reflective practitioner. A reflective practitioner is one who has made a commitment to the systematic inquiry into his or her own teaching and the testing of theory, and who is ready to allow others to observe his/her work and is prepared to discuss it (Hammersley, 1993).

The varied antecedents of action research have led to differing emphases on the nature and purpose of modern action research. Depending on the researcher, action research can be performed to modify classroom practice, to do something good in the world through direct social action, to encourage emancipation of a disadvantaged group in society or to enhance the professional development of teachers (Feldman & Minstrell, 2000). Whatever the reason for which the action research is carried out, like the design experiment, action research it is inherently cyclical. Practitioners review their own practice, identify what they want to improve, try it out, and review what happens (McNiff et al., 2003).

6.2.4 How this study fitted the research paradigms

This study had the following features of a design experiment and the action research method.

- It was cyclical – each intervention was subjected to review and retrial.
- The aim was to generate knowledge of teaching and learning so that improvements in these areas could be made.
- Testing took place through practice (here the practice of teaching).
- Validity was enhanced through triangulation.
- The findings are to be made public and therefore subject to review.
- This study was practitioner based.

This study was not collaborative as the researcher was also the practitioner, and therefore did not have this characteristic of a design experiment.

6.3 The study – aims, participants, tasks, interventions

6.3.1 The aims of the study

The first of the aims of this study was to gain knowledge of the students' understandings of statistical processes on entering university. In the first week of the semester the students were given a questionnaire that included items on stochastic processes, conditional probability, the use of proportional reasoning, and the determination of differences between sets of data. Details of this questionnaire are described in Section 6.3.3.3. The other aims were to gain knowledge of students' beliefs and difficulties in understanding *P*-values and confidence intervals, and to use this knowledge to develop teaching programs so that student understanding of these concepts could be enhanced. Knowledge of students' understandings of *P*-values and confidence intervals was gained from the reasoning used in responses to a second questionnaire given at the end of the unit and by the reasoning used in responses to selected questions in the second of two tests that were part of the students' formal assessment. The second questionnaire and test items are described in Sections 6.3.3.4 and 6.3.3.5.

6.3.2 The participants in the study

This study was carried out over four teaching semesters at the University of Tasmania. The subjects of the study for each of the four semesters (pre-intervention and the three cycles of the intervention) were volunteers from the Data Handling and Statistics unit at the University of Tasmania, which is a first year one-semester unit. As described in the previous section, the participants were asked to complete two questionnaires and to make available their responses to the second of their formal tests. At the beginning of the study the first questionnaire was given to students on both campuses of the University where the unit is taught. The responses of all these students were used in the analysis. Owing to circumstances beyond the researcher's control, only students at one campus were available after this time for the study, and the interventions were applied only to these students.

6.3.3 The sources and analysis of the questionnaires and the test items

6.3.3.1 Introduction

Because participants in the study were volunteers, it was important that the questionnaires were not too time consuming so that the participants would be more likely to complete them. The challenge, therefore, was to use or design questions that would be quick to answer, but still allow an assessment of the students' statistical understanding. Therefore multiple choice and short answer questions made up the major portion of these questionnaires. Interviews could not be used owing to ethical considerations. As the researcher was the lecturer, the lecturer could not know who had agreed to participate, and only the student numbers of the participants were given to the researcher after the unit results were published.

If written well, multiple choice questions can assess different levels of understanding from surface to deep learning. They can also be written so students must discriminate among options that vary in degree of correctness, thus allowing for a hierarchical system of grading. They also have the advantage of minimising writing so that a substantial amount of material can be covered in a relatively short time. Multiple choice items also reduce the guessing students carry out with a true-false test (Kubiszyn & Borich, 2003). However a disadvantage is that is has been found that students may use surface strategies and game playing strategies in choosing their answers (Biggs, 1999).

Restricted-response items, that is, short answer questions, are often used to measure comprehension, application and analysis (Kubiszyn & Borich, 2003). These also allow for grading on a hierarchical scale. In the questionnaires used for this study, students were usually asked to give reasons for their answers. The importance of this process was demonstrated in the trial of the first questionnaire. The respondents were given a situation where the probability of a success was a half. On being asked what number of successes out of 50 trials would be surprising, one respondent answered "25." At first this would appear that this respondent could not understand the concept of a simple expected value. However the explanation showed that the respondent was thinking of variability, and therefore did not expect to get “exactly 25,” but a number near it. Therefore, whereas the answer of “25” was incorrect a higher order of reasoning was being used than the answer initially indicated.

6.3.3.2 The sources of the questionnaire items

Several of the questionnaire items were based on The Statistical Reasoning Assessment (Garfield & Ahlgren, 1988). The Statistical Reasoning Assessment (SRA) was developed and validated as part of the ChancePlus Projects in the United States of America to evaluate the effectiveness of a new statistics curriculum for high school students. The aim of the SRA was to assess students’ statistical understanding in contrast to assessing computational accuracy and procedural knowledge. The items were designed to assess understanding about statistical measures, uncertainty, samples and association. The SRA was also designed to assess common misconceptions. These included misconceptions about sample size, the representativeness of a sample, and the equiprobability bias.

After distributing the test to experts for content validation, the authors of the SRA then administered the questions to students in the form of open ended questions. Based on these answers selected responses were used to construct a multiple choice format. Criterion related validity was then attempted by correlating the student scores on the SRA with the student marks obtained from the formal assessments. This resulted in a low correlation. The authors suggested that this finding shows that the use of successful statistical reasoning and the presence of misconceptions are unrelated to students’ performance in a statistics unit.

The authors of the initial SRA graded their questionnaire by placing questions into categories depending on whether the reasoning was correct or identified the presence of misconceptions. Watson and Callingham (2003) used many of the items of the Statistical Reasoning Assessment, however, they used a hierarchical grading system based on the SOLO Taxonomy. A hierarchical grading procedure was used to grade the questions in the questionnaires and tests in this study. Therefore the student who might answer "25" in the question described earlier would receive a higher score than a student who gave another incorrect answer because of poor statistical reasoning. This hierarchical system of grading has two practical advantages. One is that it allows for Rasch analysis (details are in Chapter 4) and the other is that students who are guessing are more likely to be detected.

6.3.3.3 The first questionnaire

The first questionnaire was designed to assess the presence of some of the misconceptions identified by Fischbein and Schnarch (1997), Kahneman and Tversky (1982), Tversky and Kahneman (1982a), and Garfield and Ahlgren (1988) and whether or not certain correct reasoning was present. These misconceptions included the representativeness heuristic (Tversky & Kahneman, 1982a), the time-axis fallacy (Fischbein & Schnarch, 1997), and the over or underestimating of sampling variability (Garfield & Ahlgren, 1988). Examples of successful reasoning included the calculation and interpretation of simple and conditional probabilities, being able to recognise independence in a simple context, and the ability to interpret a simple two-way table (Garfield & Ahlgren, 1988).

Because hypothesis testing involves the use of conditional probabilities it was considered important to assess whether or not students could answer simple conditional probability questions. If they could not do this, then it was expected that they would find the more sophisticated conditional reasoning involved in hypothesis testing to be difficult. There were three questions regarding conditional probability included in the first questionnaire. One required the use of a frequency table. The second involved forward and backward conditional statements, and the third involved conditional reasoning without the use of numbers.

A final question was included to see if students could make judgements about differences between two data sets in a simple context. This also involved judgements based on equal and unequal sample sizes (Watson & Moritz, 1999). If students could not use informal processes to determine differences in a simple context, it was expected that they would have difficulty in the determination of differences with hypothesis testing using any other reasoning than that provided by procedural knowledge. A summary of the statistical reasoning tested for, whether correct or incorrect, and the questions that relate to each form of reasoning tested for is given in Table 6.3.3.3.1. The questionnaire is provided in Appendix B1 and the coding rubric is in Appendix C1, with explanations in Chapters 7 and 8.

Table 6.3.3.3.1

Statistical reasoning assessed in the first questionnaire

Reasoning	Questions
Correct	
Correctly interprets probabilities	B1, B2, B3d[*#]
Calculates simple probabilities	C1[#], D2[#]
Understands sampling variability	D6c[#], D1b*, D3[#], D4[#], D6a[#]
Understands independence in simple contexts	C2[#], C3[#], C4[#]
Correctly interprets conditional probabilities in two way tables	E2[●], F1, Ad*
Calculates simple conditional probabilities	E3a[#]
Makes simple inferences when the group sizes are equal	F2abc[▲]
Misconceptions	
Holds the outcome orientation misconception	B3e*[#]
Holds the law of small numbers	D1c*
Believes that previous outcomes influence independent events	C2[#], C3[#], C4[#]
Over or underestimates sampling variability	C2[#], C3[#], C4[#], D5[#], D6b[▲]
Takes the time factor is taken into account with conditional probability	E3b[●]
Makes incorrect simple inferences when the group numbers are not the same size	F2d[▲]

* SRA

[#] Watson and Callingham (2003)

[●] Watson and Kelly (2007)

[▲] Watson and Moritz (1999)

6.3.3.4 The second questionnaire

Three of the questions in the second questionnaire were repeated from the first questionnaire as “anchors” for the Rasch analysis (see Section 4.5.3). These anchors provided a means by which the students’ ability ratings could be compared between the beginning and the end of the unit. The other questions were included to assess students’ knowledge about statistical significance, their ability to judge the likelihood of outcomes using standard errors, and their ability to explain the use of random processes and the importance of random sampling. There were also questions that were included to judge their ability to use hypothetical and conditional probability reasoning, with and without labelling the probabilities involved as *P*-values. A summary of the statistical reasoning tested for and the questions that relate to these can be found in Table 6.3.3.4.1. The questionnaire is provided in Appendix B2 and the coding protocol is in Appendix C2.

Table 6.3.3.4.1

Statistical reasoning assessed in the second questionnaire

Reasoning	Second questionnaire
Correct	
Understands sampling variability	8d*
Understands importance of large samples	6 part 2b*
Makes simple probability judgements when problem expressed in words	2b♠, 5ab
Calculates simple expected values	2a
Describes randomness	1#
Draws conclusions from a conditional probability	2c♠
Interprets a *P*-value	4♦
Draws conclusion from a *P*-value	4♦
Takes variability into account using a standard error	5ab
Explains the importance of chance processes in sample selection and allocation of treatments	5d, 9♥, 8b
Misconceptions	
Over/under estimates sampling variability	5c

* SRA
#Watson and Callingham (2003)
♠Brightman and Schneider (1992)
♦Written by researcher and used in Australian Maritime College Exams
♥From the assessment used in the Graduate Diploma in Science (statistics) at the University of Tasmania.

6.3.3.5 The test

Part of the formal assessment of the Data Handling and Statistics unit involved two 50-minute tests. These tests were written by the senior lecturer who was in charge of the unit and were part of the formal assessment and could not be changed for this study. The questions from the second test (given in the last week of the unit) that were included in this study assessed students' interpretation of *P*-values, interpretation of confidence intervals, and tested for the presence of the misconception that high *P*-values indicate that the null hypothesis is true. The statistical reasoning tested for and the questions involved are described in Table 6.3.3.5.1. The test items are found in Appendix B3 and the coding protocol is found in Appendix C3.

Table 6.3.3.5.1

Statistical reasoning assessed in the selected test items

Reasoning	Test item
Correct	
Interprets a *P*-value	3
Draws conclusion from a *P*-value	3
Interprets a confidence interval in terms of an estimate of the population mean	5a, 6d
Compares groups based on their averages	8bf
Misconceptions	
States that the "95%'" in a confidence interval represents where the samples will be	5b
Believes a high *P*-value indicates the null hypothesis is true	1

6.3.3.6 The analysis of the questionnaires and the test

The questionnaires were analysed quantitatively using the Rasch partial credit model. Using this model, independent assessments of the difficulty of the items and the students' ability were gained. It was planned that a similar analysis would be carried out on responses to the test items, but it was found that these did not fall on a unidimensional scale and were therefore unsuitable for Rasch analysis.

After the completion of the Rasch analysis, the items in each questionnaire were then divided into groups, in decreasing order of difficulty, depending on the cognitive demands of the items. The students were then divided into groups according to their ability, using the groupings obtained from the item analysis. For example, as the highest group of items in the first questionnaire had ratings from 0.7 logits and above, the students with an ability rating of 0.7 logits and above were placed in the highest ability group.

The responses to the first questionnaire for all the semesters involved in the study were combined into one analysis. This was reasonable because this questionnaire was administered at the beginning of each semester before any teaching took place. The ability ratings were then compared among semesters using the Kruskal-Wallis H procedure, to determine if there were significant differences in mean ability among the students from each of the four semesters involved in the study. This was of interest as the cohorts of students were different from the first semester to the second semester in each calendar year. In general, the students in the first semester of each calendar year were enrolled in courses that required a higher Tertiary Entrance Score than those in the second semester. The items in the first questionnaire were also tested using the Mann-Whitney U procedure, to see if there were any differences between the responses of the students who did and did not have previous statistical experience. The responses of the second questionnaire were also combined into one analysis, even though there were different teaching programs for each semester. This was partly because only a small number of students agreed to participate in the study for the second cycle of the intervention. Even without this problem this combination was reasonable because some of the items from the first questionnaire were repeated and used to anchor the item difficulties of the second questionnaire. As a consequence, the item and person difficulties would be placed on the same place of the logit scale no matter if the analyses were separated or not.

The final quantitative analysis of the questionnaires compared the gain (or otherwise) in students' ability from the first to the second questionnaire by the means of a paired *t*-test. Correlations were then calculated and interpreted between the ability ratings from the first and second questionnaires, the first questionnaire and the students' final result of their formal assessments, and the second questionnaire and the students' final result of their formal assessments.

A qualitative analysis of the students' responses for each questionnaire then followed. In this analysis, the responses of the students for each item were compared over the ability groups, to see how the form of reasoning used by the students in the higher ability groups differed from those in the lower ability groups.

6.4 The design for this study

6.4.1 Introduction

The aims of this study were to gain knowledge of students' beliefs and difficulties in reasoning about *P*-values and confidence intervals, and to use this knowledge to develop teaching programs that would enable students to explain how these statistical methods are derived and used in inferential statistics. The study was cyclical, in that each the nature of each intervention was based on the results and experiences of the previous semester. The following sections describe the teaching strategies used for the pre-intervention semester, and the three cycles of the intervention.

6.4.2 The pre-intervention semester

The first semester of the study, the pre-intervention semester, was used to gain knowledge of students' beliefs before the teaching program was altered. Traditionally, the unit had been taught with the use of two lectures, one tutorial session and one "practical" session per week, where the statistical calculations they required to carry out the formal assessment were demonstrated and practiced with the use of *Microsoft Excel*. The lectures were presented in didactic form, that is, they were used to pass on, and explain the information that the students were required to know. The students were expected to take notes and although they were encouraged to ask questions if required, very little questioning actually took place. The tutorial sessions were used to give the instructions for their formal assessments, and to answer questions from the formal notes that they were given in place of a text book.

The material of the unit was, and continued to be, divided into four modules. At the end of each module the students were assessed by a project, and they were given a test at the end of modules two and four. The content of the unit is shown in Table 6.4.2.1.

Table 6.4.2.1

A description of the Data Handling and Statistics unit for the pre-intervention semester

Module 1		Introduction to the discipline of statistics, sampling and experimental design
	Introduction	The discipline of statistics, its uses, and the branches of statistics
	Types of data	Measurement, categorical, response and explanatory variables
	Displaying data	Histograms, scatterplots and ogives
	Summary statistics	Mean, median, mode, variance, standard deviation, range and IQR
	Sampling methods	Random sampling Systematic sampling Cluster sampling Stratified sampling
	Survey and experimental design	Survey design and sources of error in surveys Experimental design
	Statistical independence	Contingency tables
Module 2		Probability
	Introduction to probability	Classical Empirical Subjective
	Discrete distributions	Binomial Poisson
	Continuous distributions	Normal Standard Normal *t*-distribution
Module 3		Statistical inference
	Introduction	Sampling distribution of the mean Confidence intervals
	Hypothesis testing	One sample *t*-test Two sample *t*-test Chi-squared tests for independence and goodness of fit
Module 4		Statistical applications
	Linear regression	Simple linear regression Multiple regression
	ANOVA	One factor ANOVA
	Setting the level of significance	Type I and Type II errors Considerations for setting the level of significance

Evidence of students' understandings of *P*-values and confidence intervals in the pre-intervention semester was collected using the responses to the second of the questionnaires and from responses to selected items in the second test that was used as part of students' formal assessment.

These responses were coded based on the SOLO taxonomy; the coding rubrics are in Appendix C. The responses were then subjected to a Rasch analysis (using the Partial Credit Model) to give ratings of the item difficulty and of the students' ability at the beginning of the semester. The results of the quantitative and qualitative analyses, described in Section 6.3.3.6, are in Chapters 7 and 8. Descriptive analyses of the progression of students' reasoning about *P*-values and confidence intervals are in Chapters 9 and 10.

6.4.3 The first cycle of the intervention

The responses from the students at the end of the pre-intervention semester showed that most students were able to carry out hypothesis testing and to calculate confidence intervals successfully, in that the procedures were followed and the answers were numerically accurate. In contrast, the responses to questions that required the students to show their reasoning showed that for many of these students understanding was not present. This was evident from the fact that no student attempted to define the meaning of a *P*-value in the first of the questions where this would have helped to a answer the problem, and only one quarter of the students attempted to define the meaning of the *P*-value in the second question where this would have helped explain the problem. Only one quarter could successfully answer questions that asked for an explanation of the meaning of a confidence interval.

As described in Chapter 2, Garfield and Ahlgren (1988) and Yilmaz (1996) point out that in general, early tertiary students are not familiar with hypothetical, probabilistic reasoning and find it difficult, and these difficulties can be made worse if instruction is given in a didactic fashion. In addition, Rubin, Bruce and Tenney (1991) point out that in general, tertiary students have not been exposed to sampling and do not have a realistic understanding of the relationship between samples and the original population and the variation between samples of the same population. This intervention was designed, therefore, to give students exposure to sampling and to give students exposure to the hypothetical, probabilistic process.

The first strategy that was used in order to enhance learning was the introduction of guided discovery learning via the use of computer simulation. For each simulation, the students were given a scenario, asked to make a prediction of the outcome, and then use the simulation to test their prediction (delMas et al., 1999; Garfield & Ahlgren, 1988; Mills, 2002). Each simulation was carried out before the relevant material was introduced formally in a lecture or tutorial. For example, to introduce the Central Limit Theorem, students were given some data that were normally distributed. They were then informed that they would be taking samples from these data and calculating the sample means. They were asked to predict the shape of the distribution that these sample means would have. They then used *Microsoft Excel* to take 500 samples, calculate the mean of these samples, and draw a histogram of these means. These means formed a Normal distribution. They were then asked to undergo a similar process for data that were uniformly distributed. As they had been led to believe that the sample means would have the same distribution as the population distribution, most of the students predicted that the sample means would have a Uniform distribution as well, and seemed to be surprised when a Normal distribution resulted. They then repeated the process with a Binomial distribution, one with a small sample size (n = 5) and then with a larger sample size (n = 25). It was only after this practical exercise was completed that the Central Limit Theorem was introduced formally in a lecture. A simpler example was used to compare the characteristics of the mean and median. The students entered a set of given numbers, with a range from 5 cm to 10 cm, into an *Excel* spreadsheet and calculated the mean and median. They were then asked to replace one of the numbers with a very large number of their choice. They immediately saw that the mean changed dramatically, but the median remained unchanged. The carrying out of simulations such as this is one of the benefits of using computers to enhance the learning of statistics, as repeated sampling is quick, and since visual effects such as graphs are connected to the data, any changes to the data gives an immediate visual demonstration. The details of all the simulations the students carried out are found in Appendix D.

The second strategy was to introduce the reasoning behind hypothesis testing early in the unit (Week 2) using examples such as the Chinese Birth problem, where the students used coins to simulate what would happen to the ratio of girls to boys if the "One child policy" in China was replaced by a "Have children until a boy is born" policy. After this introduction, each time a hypothesis test was introduced the students were asked a series of questions.

- What is the assumption about the population?

- Given this assumption, were our sample results very unlikely, unlikely, likely, or very likely?
- What conclusion can we make about our assumption about the population?

Because it was believed that this hypothetical thinking was unfamiliar to the students, the formal language of hypothesis testing (for example, null hypothesis and *P*-value) was not introduced until half way through the semester. This was so that students could become accustomed to hypothetical thinking without simultaneously having to learn the formal language.

The third strategy used in the semester was to familiarise students with the process of confidence intervals by having them physically draw random samples and looking at the characteristics of their samples. First, students were asked to take random samples from a population. The population consisted of 100 workers at a shipping port, whose blood lead concentrations were written onto equally sized squares of paper and these were placed in a paper bag. For each sample, the students calculated the mean, and these values were placed on a number line. By these means the students were introduced to the ideas that sample means vary from sample to sample, and that sample means can be used to make an estimate of the value of the population mean. It was after this that the students were introduced to the Central Limit Theorem with a computer simulation, and this was followed by a formal introduction to confidence intervals.

A detailed description of the traditional teaching plan, and the additional material introduced for the first cycle of the intervention is given in Appendix A. The evidence used to judge the success or otherwise of this cycle of the intervention was the same as in the previous semester; that is, the responses to the second questionnaire and the second test. The results of the quantitative and qualitative analyses, described in Section 6.3.3.6, are in Chapters 7 and 8. Descriptive analyses of the progression of students' understanding about *P*-values and confidence intervals are in Chapters 9 and 10.

6.4.4 The second cycle of the intervention

The analysis of the student responses (given in detail in Chapters 9 and 10) after the first cycle of the intervention showed that unlike the pre-intervention semester, where no students had made an attempt to define a *P*-value for either of the questions about *P*-values, 22% of the students attempted to explain the meaning of the *P*-value for one question, and 30% of the made this attempt for the other question where this was useful. Most of their explanations, however, showed misconceptions. There was no apparent improvement in the understanding of confidence intervals. Therefore some additional teaching strategies were introduced into the second cycle.

Constructivist theories of learning propose that student learning is enhanced if connections can be made between students' previous knowledge and the new knowledge being introduced. However students are not, in general, familiar with the probabilistic process used in hypothesis testing and the calculation of *P*-values. An example, therefore, needed to be found that would introduce this process and be easily understood. After some searching, an example was found that was believed could take this role. This is the "It is hot outside" problem (Shaughnessy & Chance, 2005) represented in table form in Table 6.4.4.1. With this problem the proposition was made that the weather is hot, but when the data were collected (looking out of the window), it was observed that everyone was wearing winter clothes. Because it is unlikely that people would be wearing winter clothes on a hot day, the observation was incompatible with the proposition and the proposition was rejected. This problem was considered to be easy to understand and was used as the template for all further hypothesis testing.

Table 6.4.4.1

An example of the probabilistic hypothetical process

My hypothesis	It is hot outside today
Data	When we look out of the window, everyone we see is wearing winter clothes (woolly hats, gloves and coats).
What is the probability of seeing people wearing winter clothes if it is hot outside?	Very, very low.
What do you conclude about my hypothesis?	It is incorrect.

Morgan (2001) and Pugalee (2001) state that by writing down their reasoning, students become aware of gaps in their knowledge and their understanding is increased as they strive to fill in these gaps. Students, therefore, were also encouraged to write down the meaning of the appropriate *P*-value for every hypothesis test they were introduced to in the unit.

To help increase students' understanding of confidence intervals, students were introduced to sampling as in the previous cycle. In this cycle, however, the introduction to sampling started in the first week of the semester, and more opportunities were given for students to take samples and to make estimations of the value of the population mean. The introduction to confidence intervals took place in stages. First of all, the sampling exercise was used not only to give a demonstration of the variation among sample, but also to introduce the idea that "most" sample means were “nearby” to the population mean. As the semester progressed, the students were then introduced to the idea that the proportion of means that are “nearby” is known. After the introduction to the Central Limit Theorem, they were introduced to the idea that because approximately 95% of sample means are within two standard errors of the population mean, 95% of the time the interval that consists of the sample mean ± two standard errors would contain the value of the population mean.

One result of the introduction of these strategies was that there was an increased amount of interaction among the students and between the students and lecturer. The consequence was that as the formal lectures became more conversational and the difference between the tutorials and the formal lectures became increasingly blurred.

The evidence for the success or otherwise from this cycle of the intervention was the same as in the previous semesters; that is, the answers to the second questionnaire and selected items of the second test. The results of the quantitative and qualitative analyses, described in Section 6.3.3.6, are in Chapters 7 and 8. Descriptive analyses of the progression of students’ understanding about *P*-values and confidence intervals are in Chapters 9 and 10.

During this cycle notes were also taken by the researcher of the teaching strategies used to teach *P*-values and confidence intervals and the students’ reactions to these strategies. This was so that more knowledge about students’ problems in these areas could be gained to assist in the planning of the third cycle of the intervention. Owing to ethical considerations the researcher could not know who was participating in the study, therefore no identifying information was kept, and the notes of the students' work were written in a descriptive format only.

6.4.5 The third cycle of the intervention

It was difficult to make conclusions about students' understanding for the second cycle of the intervention because only six students agreed to participate in the data collection. The data that were available, however, showed that some students were confused about the nature of the *P*-value as their answers were inconsistent. For example, one student stated that a *P*-value of .01 indicated that a new drug worked better than the previous drug 1% of the time, but recommended that the new drug be used. Misconceptions were also revealed in their understanding of confidence intervals. However, two out of the six students answered both parts of the confidence interval question correctly.

Confrey (1990) and Tobin, Tippins and Gallard (1994) stated that learning and understanding is improved if students are encouraged to explain their reasoning; the act of searching for the words to explain their meaning clarifies their ideas. Therefore the strategy used in the previous cycle, namely the encouragement of students to write and share their ideas, was extended to confidence intervals as well as the work on *P*-values.

Moreno and Duran (2004) and Ozgun-Koca (1998) found that using multiple representations can improve students' learning and understanding. In this cycle of the intervention alternative representations were used for each hypothesis test so that students could gain a visual representation of the likelihood of each test statistic, if the null hypothesis were true. For example, if the test statistic, $t = 3.6$, were calculated, a diagram such as that in Figure 6.4.5.1 would be drawn.

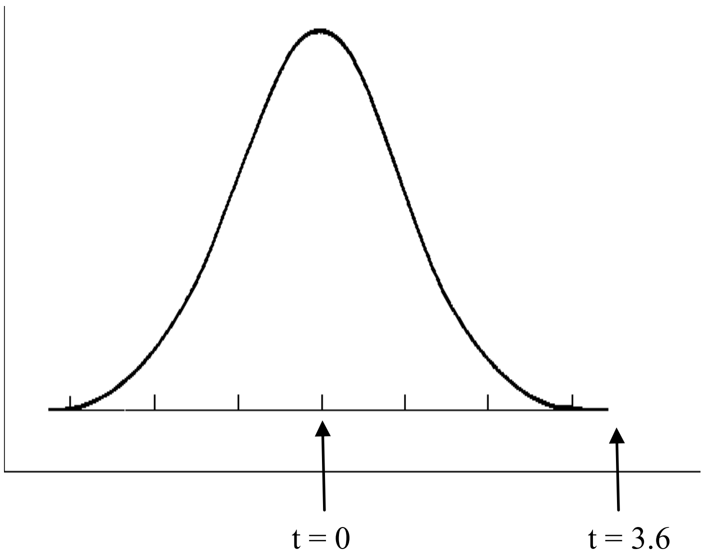


Figure 6.4.5.1. Example of a t-distribution used to give a visual demonstration of the likelihood of the test statistic.

The students were then asked to consider how likely the given statistic or one of a more extreme value would be if the null hypothesis were true using words (for example “unlikely”, “likely”). They then were asked to consider if the given sample statistic belonged to the proposed distribution or a distribution centered on another parameter.

From discussions in the previous cycle of the intervention, it became apparent that some students were confused about why statistical hypotheses are written in the way they are, that is, as the hypothesis of no difference. In this cycle, therefore, a link was made between the writing of null hypotheses and the scientific method as proposed by Popper (1963). According to Popper, it is never possible to prove a statement true. As a result scientific statements are those that make conjectures that can be falsified. Similarly, null hypotheses are also statements that cannot be proven true, but are written in a form so that evidence can be found against them.

In the earlier cycles of the intervention, the students were introduced to the hypothetical, probabilistic process early on in the semester, however the formal terminology (in particular null hypothesis and *P*-value) was not introduced to later in the semester. In the third cycle this terminology was introduced with the first example ("It is hot outside") so that students had more time to become familiar with these terms. Therefore the presentation of the problem in Table 6.4.4.1 was altered to that in Table 6.4.5.1.

Table 6.4.5.1

The "It is hot outside" problem in its new format

Null Hypothesis	It is hot outside today
Data	When we look out of the window, everyone we see is wearing winter clothes (woolly hats, gloves and coats).
P-value	Q: What is the probability of seeing people wearing winter clothes if it is hot outside? A: Very, very low.
What do you conclude about my hypothesis?	It is incorrect.

The evidence for the success or otherwise from this cycle of the intervention was the same as in the previous semesters; that is, the answers to the second questionnaire and selected items of the second test. Qualitative and quantitative analyses were carried out as described in Section 6.3.3.6, and the results of these analyses are found in Chapters 7 to 10.

6.5 A summary of the study design

Aims

- To gain knowledge of students' understandings of statistical processes on entering university.
- To gain knowledge of students' beliefs and difficulties in understanding *P*-values and confidence intervals.
- To use this knowledge to develop teaching programs so that students' understanding of *P*-values and confidence intervals can be enhanced.

Tasks requested from participants

Participants were asked to complete:

- One questionnaire on entry to the unit
- One questionnaire at the end of the unit, and to give
- Permission to use results of the second test that was used in the formal assessment.

The Pre-intervention semester (See Section 6.4.2)

The unit was taught as per previous practice.

The total number of participants was 26, 13 of these were from the Hobart campus. The Hobart students completed the first questionnaire only. Thirteen participants were from the Launceston campus, nine of whom completed the second questionnaire, and 12 of whom completed the test.

Cycle one of the intervention (See Section 6.4.3)

The changes that were introduced were:

- The use of guided discovery learning with the use of computer simulations in a “predict, test, re-think” format.
- The introduction of the hypothetical reasoning used in hypothesis testing early on in the semester. No formal language (e.g., the term “*P*-value”) was used.
- The demonstration of sampling variation by taking samples from a population and calculating and comparing the sample means.

The total number of participants was 26 out of a possible 29, 20 of whom completed the second questionnaire, and 23 of whom completed the test.

Cycle two of the intervention (See Section 6.4.4)

The further changes that were introduced were:

- The introduction of the hypothetical, probabilistic process by the “It is hot outside” problem.
- The encouragement of students to write down the meaning of each *P*-value in words.
- The staged introduction of confidence intervals.

The total number of participants was 7 out of a possible 27, four of whom completed the second questionnaire, and six of whom completed the test.

Cycle three of the intervention (See Section 6.4.5)

The further changes that were introduced were:

- The encouragement of students to write about confidence intervals as well as *P*-values.
- The use of alternative representations for *P*-values.
- The introduction of Popper's work on falsifiable statements.
- The introduction of formal statistical terminology (e.g., the term "*P*-value") from the beginning of the semester.

The total number of participants was 16 out of a possible 26, 12 of whom completed the test. The second questionnaire was not given.

6.6 Constraints on the research

The participants in the study were student volunteers. Therefore the numbers were restricted not only by the number who enrolled in the unit but also by those who chose to participate by allowing their data to be used. Because the lecturer was also the researcher, it was important that no student could be penalised, or believe to be penalised, if he/she chose not to participate. Therefore the researcher did not know who had agreed to participate and was only given the student enrolment numbers (so that the test results could be used) after the students' unit results were published. This restricted the data that could be collected during the second cycle of the intervention when teaching notes were made. Only general summaries and impressions could be recorded. This also prevented the use of follow up interviews.

Complications were also added by the nature of the students involved. In the first semester of each year the students were mainly from the biomedical science programs, whereas in the second semester the students were mainly from the aquaculture and environmental science programs. The score required for university entrance to the biomedical science programs is higher than that to enter the other programs. In addition, there is a requirement that the students in the biomedical science programs should have successfully completed a pre-tertiary mathematics unit. The mathematical competence of the biomedical science students, therefore, tended to be higher than the other students.

Because the participants were working towards various degrees, the course content, as set by the School of Mathematics and Physics at the university where this study took place, had also to be covered and could not be changed by the researcher.

Chapters 7 and 8 describe the results of the Rasch analysis of the questionnaires. The responses of the students are also analysed to see what forms of reasoning the students used, and how this reasoning varied from students with high and low ability, where the ability was determined by the Rasch analysis. These are followed by Chapters 9 and 10 in which descriptive analyses are given of the progression of students' understanding about *P*-values and confidence intervals over the time of the study.

7. Results of the Quantitative and Qualitative Analysis of the First Questionnaire

7.1 Introduction

This chapter begins with the results of the Rasch analysis of the first questionnaire and is followed by a qualitative analysis of students' responses grouped according to their ability levels. All of the Rasch analyses were carried out using Winsteps version 3.70.1.1 (www.winsteps.com.htm). Each analysis used the Partial Credit Model. Details of the theory underpinning this model can be found in Section 4.5.2.

Students' responses to each item in the questionnaire were assigned a score depending on the statistical reasoning shown in the response. For each item, comparisons were made among the scores of the four semesters (the pre-intervention semester and the three cycles of the intervention) and among the scores of students with and without previous statistical experience. Because the data were made up of discrete numbers (from zero to three), the assumptions for 2-sample *t*-tests and ANOVA, that is, normality and homogeneity of variance, could not be met. Instead of 2-sample *t*-tests, therefore, testing for differences between two groups (previous statistical experience or not) was carried out by using the Mann-Whitney U test, a non-parametric procedure. When testing for differences among three or more groups (among the four semesters), the Kruskal-Wallis H was used instead of the Analysis of Variance (ANOVA).

For the Mann-Whitney U test and the Kruskal-Wallis H tests, the ranks of the measurements were used instead of the actual measurements. These "ranks" were obtained by ordering the data from highest to lowest. For data that are appropriate for *t*-tests or ANOVA, these non-parametric tests are approximately 95% as powerful as their parametric counterparts. However, if the assumptions of *t*-tests or ANOVA are “seriously” violated, then the non parametric tests are more powerful[2] (Zar, 1974). The term “powerful” in this context reflects the ability of a test to detect significant differences if they exist.

For these non parametric tests the null and alternative hypotheses are of the form:

H_0: The groups have the same measure

H_1: The groups do not have the same measure

Similarly to ANOVA, the null hypothesis for the Kruskal-Wallis test is rejected if at least two groups are found to be significantly different from each other; rejecting the null hypothesis does not imply that all the groups are significantly different.

On the completion of the Rasch analysis, the items were placed on a scale in order of difficulty, and then grouped according to this level of difficulty. Each group was then analysed by the cognitive demands that were made by the questions in this group, and then a set of criteria were determined that characterised the qualitative changes demanded for successful responses for each group (Callingham & Watson, 2005). A qualitative analysis of the students’ responses for each questionnaire then followed. The students were then grouped by level of ability, where the groupings were based on the item groupings described above. The responses of the students for each item were then compared over the ability groups, to see

[2] As it happened, when the tests were repeated using ANOVA or 2-sample *t*-tests, all the *P*-values were within 2% of the figures obtained by these tests.

how the forms of reasoning used by the students in the higher ability groups differed from those used by the students in the lower ability groups.

7.2 Rasch analysis of the First Questionnaire

7.2.1 Introduction

The first questionnaire was designed to answer the following questions:

- What understandings and misconcpetions were held by students before undertaking the unit?
- Are there any differences in initial understanding between the students in each semester?

The second question was of interest because the students were from different cohorts. In general, the students in the first semester of each calendar year were enrolled in courses that required a higher Tertiary Entrance Score than those in the second semester.

7.2.2 Items in the First Questionnaire

There were 23 items in the first questionnaire. These included questions that required students to calculate and interpret simple and conditional probabilities. Other questions required students to demonstrate knowledge and understanding of variation and to make judgements about differences between samples using informal estimates of the centre of the data and the extent of the variation in the data. They were also required to make a judgement in a situation where the events were independent, but there were different numbers of trials. A summary of the items with their labels used in the Rasch analysis is found in Table 7.2.2.1.

Table 7.2.2.1

A description of the items used in the Rasch analysis with the labels used in this analysis

Question	Label	Description of knowledge shown
B1	Snakes	Can indicate that a probability of 1/6 does not indicate that the event will occur exactly 1 in 6 times.
B2	Cancer	Can accurately interpret the statement “14% of women will develop breast cancer sometime during their lifetime.”
B3	Eczema	Can accurately interpret the statement “For application to skin areas there is a 15% chance of developing a rash.”
C1	Coin1	Can calculate the probability of 4 tails in a row when tossing a fair coin.
C2	Coin2	Can indicate that after 4 tails in a row, the next toss will still be independent.
C3,C4	Coin3-4	Can continue with reasoning by independence for another coin toss.
D1	Hospital	Can determine that for events that are equally likely and independent (boy or girl being born) that the sample with the lowest sample size is more likely to deviate from the expected number of 50%.
D2	Spinner 1	Can calculate the chance of an even number for this situation.
D3	Spinner 2	With the knowledge from *Spinner 1*, can calculate the expected number of even numbers for 50 trials.
D4	Spinner 3	Can indicate that if 50 more trials in *Spinner 2* were carried out, the same number of even numbers would not be expected.
D5	Spinner 4	Can give an indication of the number of even numbers that would not be expected in 50 spins of the spinner in *Spinner 1*.
D6a	Tute A	Can determine that the data from 50 spins of the spinner (when presented in graphical form), are not likely to be real because the graph is perfectly symmetrical.
D6b	Tute B	Can determine that the data from 50 spins of the spinner (when presented in graphical form) are not likely to be real because the graph presents several unlikely events.
D6c	Tute C	Can determine that the data from 50 spins of the spinner (when presented in graphical form) are likely to be real because the graph is centred on the expected value and a reasonable amount of variation is present.
E1	Teacher	Can determine the most likely out of two conditional probabilities presented in verbal form.
E2	Factory	Can determine a conditional probability when the data are in a simple 2 x 2 table.

Table 7.2.2.1 (Continued)

	Label	Description of knowledge shown
E3a	Urn A	Can calculate a simple forward probability.
E3b	Urn B	Can calculate the probability of an event in the past given later information.
F1	Med	Can make a judgement of the effectiveness of a treatment using proportional reasoning.
F2a	A-B	Can compare the scores of two sets of data of equal size, presented in graphical form, where all of one group have higher individual scores than the other, without using formal algorithms.
F2b	C-D	Can compare the scores of two sets of data of equal size, presented in graphical form, where one group has a higher mean score than the other.
F2c	E-F	Can compare the scores of two sets of data of equal size, where the mean, median and mode are equal, but one group has a larger range than the other.
F2d	G-H	Can compare the scores of two sets of data of *un*equal size, where one group has a higher mean than the other.

7.2.3 The Rasch Analysis of the Items (Partial Credit Model)

In the process of the Rasch analysis it was found that some items fitted poorly to the model and these were removed and the analysis re-run. This was a step-wise process. At first only the items with the more extreme *t*-statistics, less than negative three or greater than plus three, were removed. This process continued until all the items remaining in the analysis had a satisfactory *t*-statistic within the range of negative two to plus two. By the end of this process, the items that were removed were *Snakes*, *Hospital, Spinner 1, Spinner 2, Coin 1, Factory*, and *A-B, C-D, E-F* and *G-H.* The fitting of the items in the original analysis before these items were removed can be seen in the bubble chart in Figure 7.2.3.1. Because of the way the items' arrangement changed as the items with very large *t*-statistics were removed, some items that appeared to have a poor fit in the first analysis remained in the final analysis.

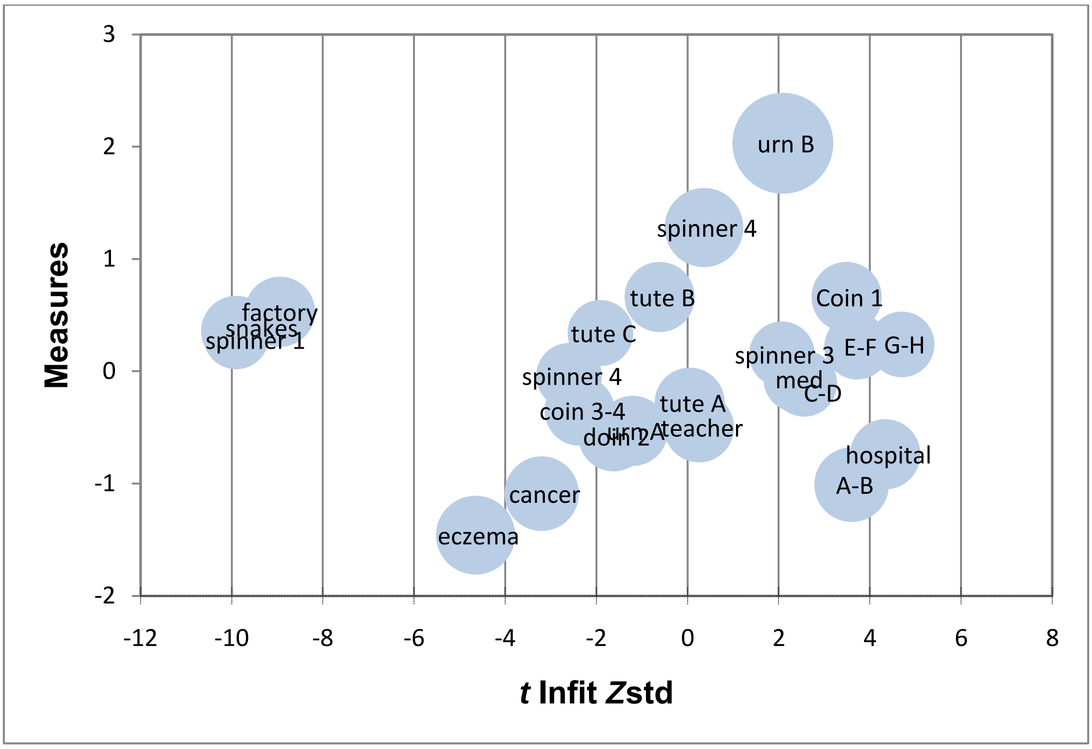


Figure 7.2.3.1. Bubble chart for item fit in the first attempt at the Rasch analysis.

The variables *Snakes, Spinner1* and *Factory* appeared close together in the bubble chart, as did the items *Hospital, Spinner 3, Coin 1, A-B, C-D, E-F, G-H* and *Med.* This raised the question of whether or not factor or cluster analysis would be suitable for these data. When a factor analysis was attempted, the Kaiser-Meyer-Olkin statistic was 0.529, which indicates that these data are probably not suitable for factor analysis. An examination of the factors produced confirmed this supposition. Only one factor was produced with a small number of items, and these items had no relation either to the pattern in Figure 7.2.3.1 or to the cognitive demands of the items. The cluster analysis was also unsatisfactory, as it did not show linkages that relate to the items either by place on the chart, or by cognitive demand.

The list of items used in the main analysis with their mean difficulty measures are shown in Table 7.2.3.1. This list is in order of difficulty, and is measured in logits, the logarithm of the odds of success. This analysis had an item reliability score of 0.95 (maximum obtainable is 1.0, see Appendix E1). This indicates that the replicability of item placements along the scale if these same items were given to another equal sized sample of similar participants, is predicted to be high (Bond & Fox, 2007).

Table 7.2.3.1

Items in decreasing order of mean difficulty for the Rasch analysis

NAME	MEASURE
Urn B	1.80
Spinner 4	1.13
Coin 1	0.55
Tute B	0.55
Tute C	0.25
Spinner3	0.06
Med	-0.15
Tute A	-0.37
Coin3-4	-0.45
Teacher	-0.61
Coin 2	-0.64
Urn A	-0.70
Cancer	-1.43

More detail of the items is shown in Figure 7.2.3.2, which shows the item map with the Rasch-Thurstone thresholds for each category of each item. There are five people (indicated by “X”) at the same level as items *Spinner4.1* and *Tute C.2*. This indicates that these people have a 50% probability of scoring a zero or one for *Spinner 4*, and a 50% probability for scoring a one or two for *Tute C*. Further details of the Rasch Thustone thresholds are described in Section 4.5.3. The spread of the participants and the questionnaire items were well matched, in that in general, they both cover the same range.

```
         |                    |
         |                    |
         |                    |    urnB    .2
         |                    |
     2   |   X                |
         |                    |
         |                    |
         |   X                |    spinner4 .2
         |                    |
         |                    |    urnB    .1
         |   XX               |
         |                    |
         |   XXXXX            |
     1   |                    |    tuteB .2        coin1   .2
         |   XXXX             |
         |                    |
         |   XXXXX            |    spinner4  .1    tuteC   .2
         |   XX               |
         |                    |    spinner3 .2
         |   XXXXXXXXXXX      |
         |   XXXXXX           |    med     .2
         |                    |
         |   XXXXXXX          |    coin1   .1      tuteA   .2     tuteB  .1
Logits 0 |   XXXXXXXXXX       |    coin3-4  .2
         |                    |
         |   XXXXXXX          |    tuteC   .1      coin2   .2     teacher  .2
         |                    |
         |   XXX              |    urnA    .2
         |                    |    spinner3 .1
         |   XXX              |
         |   XX               |    med     .1
         |                    |
         |   XX               |    tuteA   .1
         |   X                |    coin3-4 .1
    -1   |                    |    teacher .1      cancer  .2
         |   X                |    coin2   .1      urnA  .1
         |                    |
         |   X                |
         |                    |
         |                    |
         |                    |
         |                    |
         |                    |
         |                    |    cancer  .1
    -2   |                    |
         |                    |
         |                    |
         |   X                |
```

Figure 7.2.3.2. Item and person map of the Rasch analysis of the first questionnaire, showing the Rasch-Thurstone thresholds. Each "X" represents one person.

The items were then divided according to increasing complexity of the reasoning needed to respond to the questions successfully. Examining the variable map gave tentative groupings where there was an apparent jump in difficulty. The decision concerning where to separate these groupings was assisted by importing the item difficulties into a spreadsheet and using these to produce a bar graph of the item difficulties in order of increasing difficulty. This bar graph is shown in Figure 7.2.3.3.

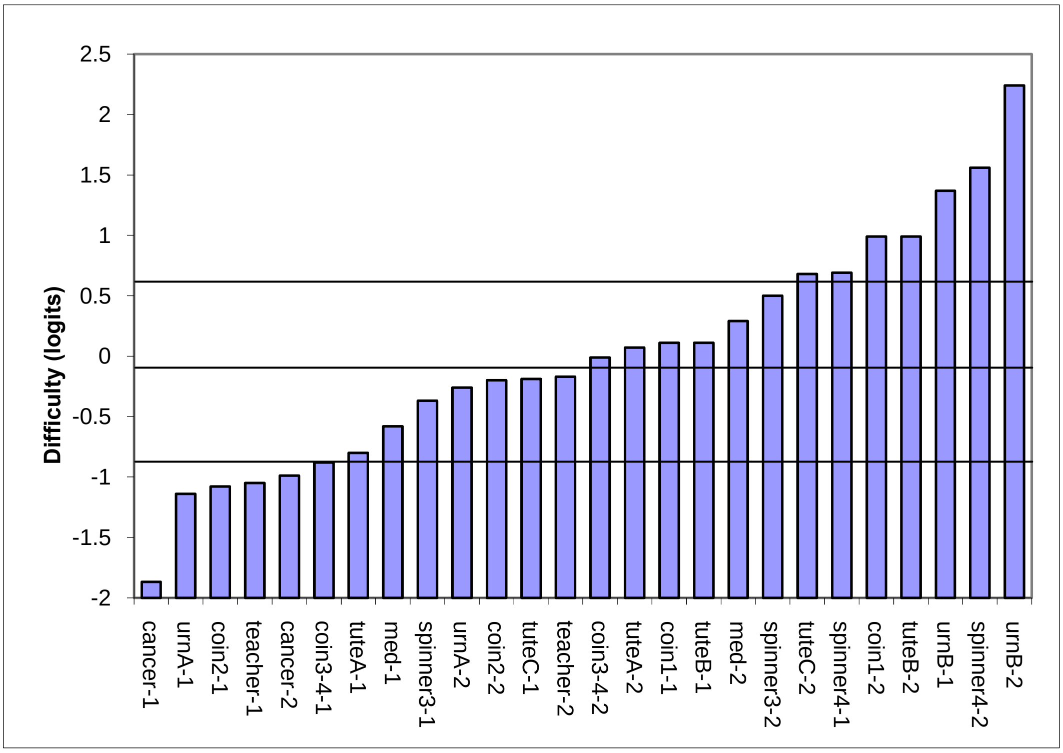


Figure 7.2.3.3. Items in order of difficulty from lowest to highest for the first questionnaire

The items were then assessed on the type of reasoning required by the students to respond to each question at each level successfully. As a result of this process, the items were divided according to the lines in Figure 7.2.3.3. From the bar graph, it would appear that *Tute A* should be placed into the lower grouping. To gain a category one response for *Tute A*, however, the student needed to have an understanding that probabilistic processes show variation, which requires a higher cognitive level than what was needed to answer *Coin 3-4.1*. *Coin 3-4.1* was placed in the lower grouping as a response at this level only required the interpretation of a simple verbal and numeric statement, as did the other items in this grouping. The groupings, together with the cognitive demands made on the students in each level of difficulty are summarised in Table 7.2.3.2.

Table 7.2.3.2

Summary of the cognitive demands made on the students for each group of items

Description of Item	Measure of difficulty	Scoring Rubric	Reasoning used
Urn B .2	2.24	Correctly calculated the probability of a past event given later knowledge	The presence of variation in probabilistic processes was recognised and the prediction of the degree of this variation included a “greater than” and a “less than” response. Independent probabilistic processes were calculated correctly. The probability of an earlier event given later knowledge was calculated correctly.
Spinner 4 .2	1.56	Prediction of what would be unusual was correct and included both “greater than” and “less than” response	
Urn B .1	1.37	Conditional probability not recognised therefore joint probability was calculated	
Tute B .2	0.99	Explanation was based on two or more reasons	
Coin1 .2	0.99	Probabilities of multiple independent events were successfully calculated	
Spinner 4 .1	0.69	Prediction of what would be unusual was correct but included “greater than” or “less than” only	The presence of variation in probabilistic processes was recognised but the degree of this variation was not predicted completely. Independent probabilistic processes were recognised, and proportional reasoning was used. Explanations were complete.
Tute C .2	0.68	Explanation was based on two or more reasons	
Spinner 3 .2	0.50	Explanation was based on expectation of variation	
Med .2	0.29	Explanation of comparison between two groups was based on proportional reasoning	
Tute B .1	0.11	Explanation was based on one reason only	
Coin 1 .1	0.11	Multiple independent events were not successfully calculated, but independence was correctly identified	
Tute A .2	0.07	Prediction was based on expectation of variation in real situation	
Coin3-4 .2	-0.01	Identified independence of each outcome	

Table 7.2.3.2 (Continued)

Description of Item	Measure of difficulty	Scoring Rubric	Reasoning used
Teacher .2	-0.17	Recognised the difference in two conditional statements and explanation was based on unequal probability	The presence of variation in probabilistic processes was not recognised. Simplified explanations were given.
Tute C .1	-0.19	Explanation was based on one reason only	
Coin 2 .2	-0.20	Recognised independence of coin tosses	
Urn A .2	-0.26	Correctly calculated probability of future event given past knowledge	
Spinner 3 .1	-0.37	Simplified explanation, for example, “It's chance only”	
Med .1	-0.58	Used proportional reasoning but only described one group	
Tute A .1	-0.80	Identified situation as made up but no explanation given	
Coin 3-4 .1	-0.88	Simplified explanation, for example, “1 in 2 chance”	Simple probabilistic statements in verbal and numeric form were interpreted correctly.
Cancer .2	-0.99	Verbal probability statement interpreted correctly	
Teacher .1	-1.05	Explanation based on personal experience, for example, “More teachers in my school are women”	
Coin 2 .1	-1.08	Simplified explanation given, for example, “It's just chance”	
UrnA .1	-1.14	Calculated probability of future event given past knowledge as a joint probability	
Cancer .1	-1.87	Interpreted verbal probabilistic statement in a way that underestimated probability/did not recognise difference between incidence and mortality	

7.2.4 Rasch analysis of persons

The person analysis had a person reliability score of 0.60 (maximum score available is 1.0). This gives a measure of the replicability of person ordering that would be expected if a test of similar items were given to the same people. The lower score compared to the item reliability score indicates that this reliability is not as high as the item reliability.

A histogram of these data is shown in Figure 7.2.4.1. The histogram shows that most of the students had an ability measure between -1.5 and +2.0 logits, with one student having an ability measure below -2.5. The results of the Kruskal-Wallis test (*P* = .424, α = .05) indicated that there was no significant difference in person ability scores on the first questionnaire among the four semesters. The details of the analysis are in Appendix E1.

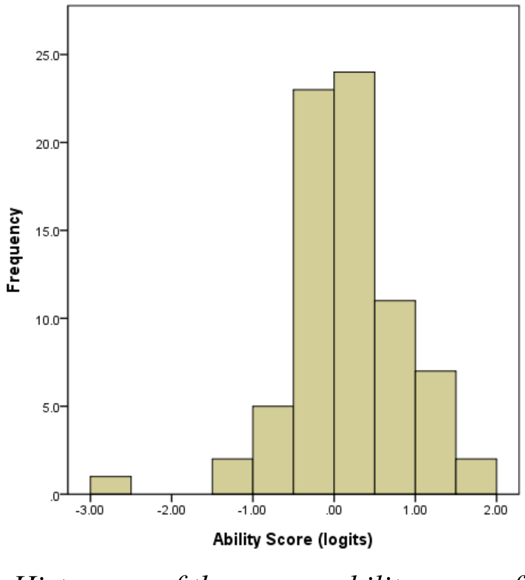


Figure 7.2.4.1. Histogram of the person ability scores for the Rasch analysis of the first questionnaire.

The students were then divided into four groups, according to where they could achieve in relation to the item difficulty levels found in Table 7.2.3.2. These groups were labelled, from highest to lowest, as A, B, C and D. This is illustrated in Figure 7.2.4.2. Students in Group D were able to interpret simple probabilistic statements, those is Group C were able to answer more difficult items but only using simple explanations, those in Group B were able to recognise the presence of variation in probabilistic processes, and to use proportional reasoning, whereas those in Group A were able to show a more accurate understanding of the degree of variation shown by probabilistic processes.

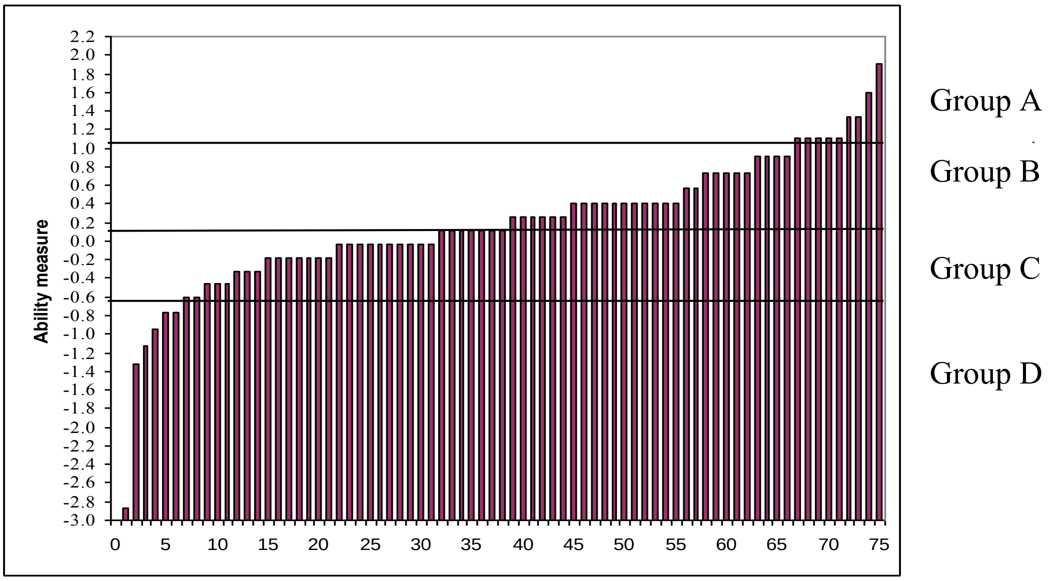


Figure 7.2.4.2. Ability levels of students divided to correspond with table 7.2.3.2.Each column represents one student.

Table 7.2.4.1 shows the scores obtained by the students for each item in the Rasch analysis, with their corresponding frequencies. Unless indicated, there was no significant difference in the scores between the semesters, and no significant difference in the scores according to whether the students had previous statistical experience or not (Appendix E1). A qualitative analysis of the students’ responses is found in Section 7.3.

Table 7.2.4.1

Frequency of scores received by the 75 students for each of the items of the first questionnaire

	Score			
Item	3	2	1	0
Cancer	N/A	59	12	4
Eczema	N/A	68	6	1
Coin 1	N/A	28	4	43
Coin 2	N/A	43	23	9
Coin 3-4	N/A	33	36	6
Hospital **	22	12	34	7
Spinner 1	N/A	N/A	73	2
Spinner 2	N/A	24	41	10
Spinner 3	N/A	33	15	27
Spinner 4	N/A	10	18	10
Tute A	N/A	43	13	19
Tute B	N/A	15	30	30
Tute C	N/A	19	35	21
Med	N/A	40	10	25
A-B	20	21	25	9
C-D	9	19	27	20
E-F	6	25	10	34
G-H	10	19	9	37
Factory	N/A	N/A	65	10
Teacher	N/A	48	12	15
Urn A	N/A	48	15	12
Urn B	N/A	6	8	61
Snakes	N/A	N/A	72	3

N/A – not applicable – this score was not available for this item.

** There is a significant difference (*P* < .01) between students who claimed previous statistical experience than those who did not. The mean rank for the group without previous statistical experience (2.06) is higher than for those students with previous statistical experience (1.40).

7.3 Qualitative analysis of the first questionnaire

In this section the students' responses are compared across the ability groups, to see what types of reasoning the students used and how these varied between the ability groups. Because it was desired that the knowledge gained would be as complete as possible, all the items are included, regardless of whether or not they fitted the unidimensional scale in the Rasch analysis. The items are divided according to the form of required reasoning: the interpretation of verbal probabilistic statements, the interpretation and calculation of statistically independent events, the interpretation and calculation of conditional events, and the judgement of differences between groups. The students were also required to answer items that would indicate whether or not they expected variation in stochastic processes.

7.3.1 Questions requiring interpretation of verbal probabilistic statements

In this section the questions that required the students' ability to interpret verbal probabilistic statements are examined.

"Cancer"

The students were asked to choose the best interpretation, out of a set of options, for the following.

> It is estimated that 14% of women will develop breast cancer sometime during their lifetime.

The responses according to ability are found in Table 7.3.1.1. The majority of the students in groups A to C gave the response: "Not many women get breast cancer, but it is not that uncommon." This response was given a score of "2." The responses "Not many die from breast cancer, but it is not that uncommon," and "It is not likely that a woman will get breast cancer" resulted in a score of "1." All other responses received a code of "0."

Table 7.3.1.1

Responses to "Cancer" according to the students' ability

Ability Group	A	B	C	D
Response	(n = 13)	(n = 31)	(n = 25)	(n = 6)
Not many women get breast cancer, but it is not that uncommon	10	28	19	2
Not many women die from breast cancer, but it is not that uncommon	1		1	1
It is not likely that a woman will get breast cancer	2	2	4	
It is not likely that a woman will die of breast cancer		1		1
More women than not will get breast cancer			1	1
It is very likely that a woman will get breast cancer				1

Of all the items that remained in the final Rasch analysis, this one was the least difficult for the students. The level of difficulty was 1.43 logits below the mean of zero. Seventy-nine percent of the students received the highest score of "2," and a further 16% received the score of "1."

"Eczema"

The students were given a series of options to answer the following question.

> The following message is printed on a bottle of prescription medication:
>
> > Warning: For application to skin areas there is a 15% chance of developing a rash. If a rash develops, consult your physician.
>
> Which is the best interpretation of this warning?

The responses, according to ability, are found in Table 7.3.1.2. Very few students did not give the response, “About 15 out of 100 people who use this medication develop a rash.” This response resulted in the highest score of “2.” The responses “Don’t use the medication on your skin; there is a good chance of developing a rash” and “There is hardly any chance of getting a rash in using this medication” resulted in a score of “1.”

Table 7.3.1.2

Responses to “Eczema” according to the students’ ability

Ability Group	A	B	C	D
Response	(n = 13)	(n = 31)	(n = 25)	(n = 6)
About 15 out of 100 people who use this medication develop a rash	13	28	21	6
If a rash develops, it will probably involve only 15% of the skin		1		
There is hardly any chance of getting a rash in using this medication		2	3	
Don’t use the medication on your skin; there is a good chance of developing a rash			1	

Ninety-one percent of the students received a score of “2,”,with a further 8% of the students receiving a score of “1.” Because the students did so well on this item, it had a very low *t*-score (Figure 7.2.3.1) and did not fit on the unidimensional scale.

“Snakes”

The students were required to imagine they were playing a game of snakes and ladders where a “6” was required for a player to commence the game. They were asked to choose from the five statements.

After four rounds no-one has started. Which of the following statements best matches your conclusion?

a. Since a „6’ hasn’t come up yet, it will come up in the next round.
b. Since the chance of getting a ‘6’ is 1 in 6, the die should have come up with a „6’ four times by now, so something is wrong with it.
c. Throwing a die is a chance event, so it just happens like this sometimes.
d. If a „6’ doesn’t come up soon, there must be something wrong with the die.
e. There must be something wrong with the die.

All but three students responded to this question correctly (option (c)). This high number of correct responses resulted in a very low *t*-score and the item did not fit on the unidimensional scale for the Rasch analysis; this is illustrated in Figure 7.2.3.1.

The responses to these items suggested that the students found it easy to interpret verbal probabilistic statements. The responses to *Snakes* indicated that they could interpret a probability in terms of a long term frequency, and the responses to *Eczema* indicated they could also do this when the probability was given in terms of a percentage. The responses to *Cancer* indicated that most students thought an incidence of 14% was “not that uncommon,” or “not likely,” but some of the responses indicated that there was an association in their minds of cancer and death, with the result they did not realise the question only referred to incidence.

7.3.2 Questions requiring an understanding of statistical independence

This section examines the questions that needed an understanding of statistical independence to be successfully answered. Formally, events A and B are independent if the conditional probability of A given B is the same as the unconditional probability of A, that is,

$$P(A|B) = P(A).$$

Informally, understanding independence can be thought as recognising that events governed by randomness have no memory, so that the previous outcomes have no influence on any future outcomes. For example, the probability of a head on the toss of a fair coin will remain at 50%, even if several heads in a row have resulted from previous tosses.

The coin questions

The first coin question, *Coin 1*, required the students to calculate the probability of getting four tails in a row. The second question asked the students to choose from a series of options to indicate whether a head or a tail would be more probable for the fifth toss, or whether they are equally likely. The third and fourth questions asked them to calculate the probability of getting each of a head and a tail on the fifth toss.

"Coin 1"

This item, with a difficulty score of 0.55 logits, was the hardest out of the coin questions. The results (Table 7.2.4.1) show that 57% of the students could not correctly calculate the probability of four heads in four coin tosses and thus did not achieve a score of "2." Neither could they state that the probability was 50%

each and every time the coin was tossed (which would result in a score of "1"). The numbers of each response given, as shown by ability group, are in Table 7.3.2.1.

Table 7.3.2.1

Responses given to the probability of tossing four tails in a row by ability group

	Response						
Ability group	1 in 16 (correct)	1 in 2	1 in 4	1 in 8	Other	No response	Total
Group A	9	1	1	1	0	1	13
Group B	12	8	3	1	6	1	31
Group C	8	11	3	0	2	1	25
Group D	0	5	1	0	0	0	6

It would appear that one of the "1 in 4" and one of the "1 in 8" responses came from inappropriate arithmetic. For example, the reasoning given by two students were

"50% each time therefore 1/4"

"½ * ½ * ½ * ½ = 1/8"

There may have been other students with correct reasoning but faulty arithmetic but this cannot be determined as most of the students did not explain how they got their answers. It is apparent that the majority of the students in Group A could answer this question successfully, and that the proportion of students who could do this decreased down the groups until Group D, where none of the students could do so.

"Coin 2"

The students were then asked to choose from a series of options to determine which choice someone should make for the fifth toss. Should a person choose

Heads, Tails, or doesn't it matter? Table 7.2.4.1 indicates that approximately 57% of the students received a score of "2." This indicates that these students gave the correct response ("it does not matter") with a suitable justification. Approximately 31% of the students gave the correct response without a full justification, or a justification that relied on the fact that there were only two outcomes but did not mention the independence of each toss (score "1"). Table 7.3.2.2 shows that a higher proportion of students of lower ability groups provided responses that scored "1." This question had an item difficulty of 0.64 logits below the mean, 1.19 logits below *Coin 1*, and was the easiest of the *Coin* questions.

Table 7.3.2.2

Responses to "Coin 2" according to the ability of the students.

Ability Group	A	B	C	D
Response	(n = 13)	(n = 31)	(n = 25)	(n = 6)
Response directly states that each toss is independent/Does not depend on what happened before	8	9	8	
Response implies independence, for example, the probability is "still" 50%	3	13	6	2
Response is correct (1/2) but does not use independence in argument – only two outcomes	2	5	9	3
Response is incorrect		4	2	1

"Coin 3-4"

The students were then asked to calculate the probability of first, a head for the 5th toss, and then the probability of a tail for the 5th toss. Because most students combined their responses these two questions were coded together. Table 7.2.4.1 shows that 92% of the students either calculated the probability correctly with a full explanation based on independence (score "2") or with a response that implied independence or based on the argument that there were only two outcomes (score "1"). This item, with an item difficulty of 0.45 logits below the mean, was the second easiest of the coin questions. The responses, according to ability, are found in Table 7.3.2.3. This table shows that Group A students were more likely to use the independence of the coin tosses in their reasoning. Six students (from Groups B and C) showed inconsistency, in that they stated that heads or tails would be more likely for the 5th toss and then went on to say that the probabilities of heads or tails were 50% for this toss.

Table 7.3.2.3.

Responses to "Coin 3-4" according to the ability of the students.

Ability Group	A	B	C	D
Response	(n = 13)	(n = 31)	(n = 25)	(n = 6)
Response directly states that each toss is independent/Does not depend on what happened before	5	10	5	
Response implies independence, for example, the probability is "still" 50%	3	10	6	1
Response is correct (1/2) but does not use independence in argument – only two outcomes	5	10	12	4
Response is incorrect		1	2	1

"Hospital"

The students were asked to answer the following question.

> Half of all newborns are girls and half are boys. Hospital A records an average of 50 births per day. Hospital B records an average of 10 births per day. On a particular day, which hospital is more likely to record 80% or more of female births?

To respond to the Hospital Problem successfully, it is necessary to look beyond the probability of each single birth, and to appreciate the effect of sample size. If the sample size is ignored, it will be concluded that as the two events are described by the same statistic, and as each birth is independent, the two hospitals will be equally likely to record 80% or more girls. As a sample increases in size, however, the sampling statistic (here the proportion of girls born) is more likely to approach the theoretical value for the entire population (Fischbein & Schnarch, 1997). Using the Binomial distribution the probabilities are:

$$P(x \geq 8 | n = 10, p = 0.5) = 0.0547$$

$$P(x \geq 40 | n = 50, p = 0.5) = 0.0000$$

It was not expected that students would have been able to carry out these calculations. What was required was that they should use their intuition to answer the question: Is it more likely to get eight or more girls out of 10 births, or 40 or more girls out of 50 births?

The most common response (given by 49% of the students) was that the two hospitals were equally likely to record 80% or more of female births in a day as each birth was an independent event. Even though this response is incorrect, it resulted in a score of “1,” as it indicated a higher level of statistical understanding than lead to a response of Hospital A. Those students who stated that Hospital B was more likely to record 80% of more female births, either because the sample size was smaller for Hospital B, or larger for Hospital A, received a score of “3.” Those students who said that it was “easier” for B to get 80% or more female births received a score of “2.” Twenty-three (40%) of the students gave the correct response of Hospital B.

For this question, the Mann-Whitney U test was significant according to whether or not the students claimed to have experience with statistics in previous mathematics courses (*P* = .016, see Appendix E1). Those students who did not claim previous statistical experience had a higher mean rank (48.03) than those who did (34.83). This is possibly due to the students with previous statistical experience being distracted by their knowledge of statistical independence. In Table 7.3.2.4 students responses are divided by the category of previous experience as well as by ability.

Table 7.3.2.4

Students' responses to the "Hospital" question according to previous statistical experience and ability level

	Previous statistics experience				No previous statistics experience			
Ability Group	A	B	C	D	A	B	C	D
Response	(n = 8)	(n = 25)	(n = 19)	(n = 5)	(n = 5)	(n = 6)	(n = 6)	(n = 1)
More likely for B as the sample size is smaller/Less likely for A as the sample size is larger	1	7	4	1	1	2		
"Easier" for B		5	3	2	4		4	
Equally likely – the probability of 50% for each birth is constant	7	8	10	2		4	2	1
A – more births, more girls		4	2					
No response		1						

A high proportion of students from all ability groups gave the equally likely option, and as a result, this item did not fit on the unidimensional scale for the Rasch analysis. In Figure 7.2.3.1 it can be seen the *Hospital* item had a high *t*-score.

7.3.3 Students' Awareness of Variation in Stochastic Processes

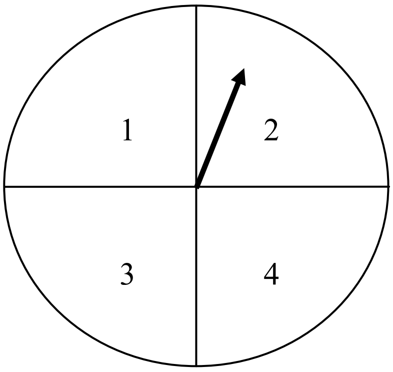
The *Spinner* and the *Tutorial* questions described in this section were included to test students' awareness of the presence of variation in stochastic processes. The correct responses required awareness that if repeated, any process that has a random component will be unlikely always to give identical results. Even though a large number of trials will converge to the expected value determined by the conditions of the process, a small number of trials are likely to show variation in the outcomes.

The spinner questions

The first question was to set the scene for the questions after it. Once students determine that out of 50 spins the expected value of the number of spins that land on an even number is 25, they are asked a series of questions to determine if they are aware that this result will not be achieved for every set of 50 spins.

“Spinner 1”

A tutorial group used this spinner. If you were to spin it once, what is the chance it will land on an even number?



All but two students gave the correct response of ½ or 1 in 2 or 50%. Because such a high number students answered this question correctly, the question had a very low *t*-score, and did not fit into the unidimensional model of the Rasch analysis.

“Spinner 2”

Out of the next 50 spins, how many times do you think the spinner will land on an even number?
Why do you think this is?

If the students answered they would expect the spinner to land on an even number 25 times, and indicated that they expected that the answer would not necessarily be exactly 25 they received a code of “2.” Table 7.2.4.1 shows that 32% of the students achieved this score. If the students indicated that they would expect exactly 25 even numbers then the students received a code of “1.” This was achieved by 55% of the students. The remaining students could not make this calculation. With such a high number of students receiving a score of "1" or "2," this item had a low *t*-score and did not fit on the unidimensional scale of the Rasch analysis. The responses according to ability are in Table 7.3.3.1.

Table 7.3.3.1

Responses to "Spinner 2" by ability

Ability Group	A	B	C	D
Response	(n = 13)	(n = 31)	(n = 25)	(n = 6)
Approximately 25	6	14	6	
25	6	15	17	4
Can't tell as it is random	1	2	2	2

"Spinner 3"

If you were to spin it 50 times again, would you expect to get the same number out of 50 to land on an even number? Why do you think this?

The *Spinner 3* item was included to prompt the students to think about variation. For this question 76% of the students gave a response that showed that they would expect some variation between groups of 50 spins to occur, with varying degrees of sophistication in their responses. Those responses that stated that the process was random or similar received a score of "2," while those that stated that it was "just chance" or "anything can happen" received a score of "1."

The results of questions *Spinner 2* and *Spinner 3* are combined in Table 7.3.3.2 to show how many students in each ability group indicated that they would expect variation in the number of even spins that land on an even number for each set of 50 spins. For groups A and B it is apparent that most of the students were aware that variation is to be expected. Students in groups C and D were much less likely to be aware of this variation.

The item difficulty rating of *Spinner 3*, 0.06 logits above the mean, illustrates that it was not too difficult for most of the students to answer successfully. It is of concern, however, that nearly a quarter of the students did not consider that the results of different trials would vary, even when prompted to do so.

Table 7.3.3.2

The number of students in each ability group who were aware/not aware that variation is to be expected in the spinner questions. Some students did not answer, therefore the number of responses does not sum to 75.

	Variation is expected	Variation is not expected
Ability Group A	12	1
Ability Group B	23	3
Ability Group C	15	9
Ability Group D	2	3

"Spinner 4"

How many times out of 50 spins, landing on an even number, would surprise you? Why do you think this is?

Using the Binomial distribution the probability distribution for the number of times the spinner lands on an even number out of 50 spins is illustrated in Figure 7.3.3.1. From this distribution it is apparent that the most likely outcome is for the spinner to land on an even number 25 times out of 50 spins. The probability of getting below 13 or above 37 is extremely low.

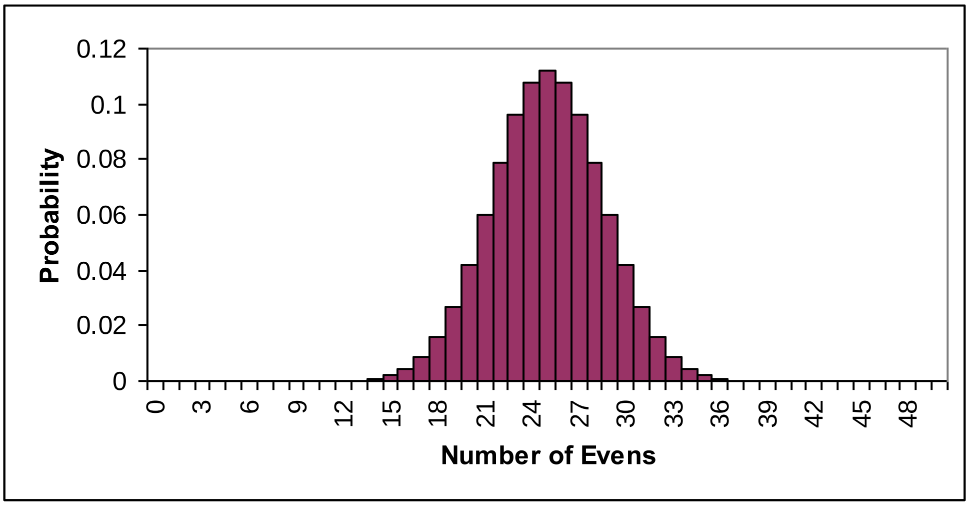


Figure 7.3.3.1. The probability distribution for the number of times a spinner lands on an even number out of 50 trials when the probability of landing on an even number is 50%.

The students were not expected to be familiar with the Binomial distribution, nor were they expected to be able to carry out the calculations that resulted in Figure 7.3.3.1. Therefore some leeway was given for their responses. Those who gave responses of above 35 (or a number close to 35), and below 15 (or a number close to 15), received the highest score of "2." Most of the students, however, gave responses that were one sided, in that they stated above a certain number, or below a certain number, but not both, and these were given a score of "1." This omission contributed to this question having the second highest difficulty rating, 1.13 logits above zero, 2.56 logits above the easiest item. Another difficulty was the inability of 31 of the students to think in terms of a range of numbers. For example, one student gave a response of "40," instead of 40 and above. The responses according to ability are found in Table 7.3.3.3.

Table 7.3.3.3

Reponses to "Spinner 4" by ability

Ability Group	A	B	C	D
Response	(n = 13)	(n = 31)	(n = 25)	(n = 6)
Above 35 (approximately) AND below 15 (approximately)	3	4	2	
Above 35 (approximately) OR below 15 (approximately) but not both	8	9	4	
Zero only		2	1	1
50 only		5	7	2
25 only				2
Other single number stated	1	2	3	
Zero and 50 only	1	2	1	1
No number is unexpected		2	2	
No response			4	
Idiosyncratic*		5	1	

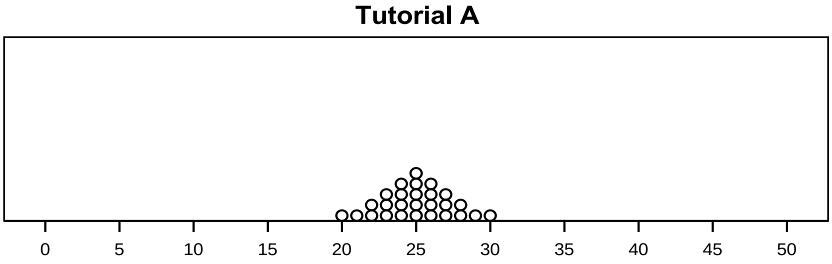
* Idiosyncratic – these responses were either incomprehensible, or did not address the question

The Tutorial Questions

In these questions students were given graphical representations of the supposed results for three tutorial groups that spun the spinner (shown in *Spinner 1*) 50 times and recorded the number of times an even number was obtained. These questions not only tested for students' awareness of variation in stochastic processes but further tested their ideas concerning what reasonable outcomes might be. The following introduction to the question was given.

> The members of three statistics tutorial groups did 50 spins and graphed the number of times the spinner landed on an even number. Each circle represents one person in the tutorial group. In some cases, the results were just made up without actually doing the experiment.

“Tute A”



a. Do you think tutorial A’s results are made up or really from the experiment?
    i. Made up
    ii. Real from experiment

   Explain your answer.

The student responses, according to their ability grouping, are found in Table 7.3.3.4. It is apparent that 75% of the students realised the results were likely to have been made up. Those responses that stated that the results were too symmetrical or perfect received a score of “2.” Those responses that stated the results were likely to have been made up without an explanation received a code of “1.” This item had a difficulty rating of 0.37 logits below zero and was the easiest of the *Tute* questions. Approximately one quarter of the students, however, were not concerned about the symmetry of the problem, giving further evidence that some students were not aware of the likelihood of variation in the outcomes of a stochastic process.

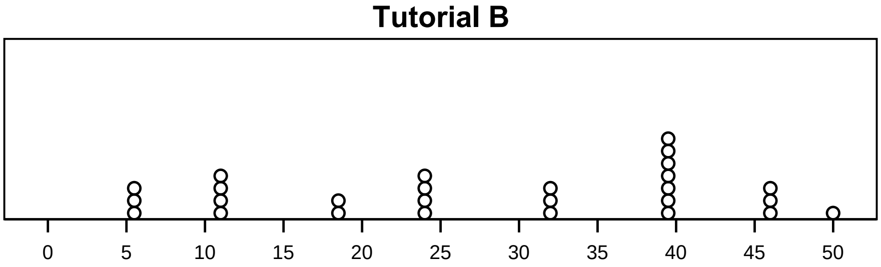
Table 7.3.3.4

Responses to "Tute A" according to ability group

Ability Group	A	B	C	D
Response	(n = 13)	(n = 31)	(n = 25)	(n = 6)
Made up				
Too perfect/neat/uniform	12	27	13	
Real				
Follows bell shaped curve	1		1	
Got expected number		3	5	3
No response			2	2
Idiosyncratic*		1	4	1

* Idiosyncratic – these responses were either incomprehensible, or did not address the question.

"Tute B"



b. Do you think tutorial B's results are made up or really from the experiment?
   i. Made up
   ii. Real from experiment
   Explain your answer.

This item gave results which at first glance, being more varied than for *Tute A,* appear more likely to be real. When the graph is examined closely, however, several problems are revealed that indicate that the results were likely to have been made up. There are no results for 25, the number with the highest probability; there are multiple results on each of some values, with gaps in-between; and the answers of 46, 50 and 6 even numbers are so unlikely as to be virtually impossible. From Figure 7.3.3.1 it is also apparent that the results of 11 and 39 even numbers are also extremely unlikely.

Some students, however, did state that the range was too wide, without being specific about the probability of each result. Those students who gave more than one reason for saying the results were made up were given a score of "2," and those students who gave one reason only were given a score of "1." Table 7.2.4.1 indicates that 60% of the students realised the results were likely to have been made up. The responses grouped according to ability are found in Table 7.3.3.5.

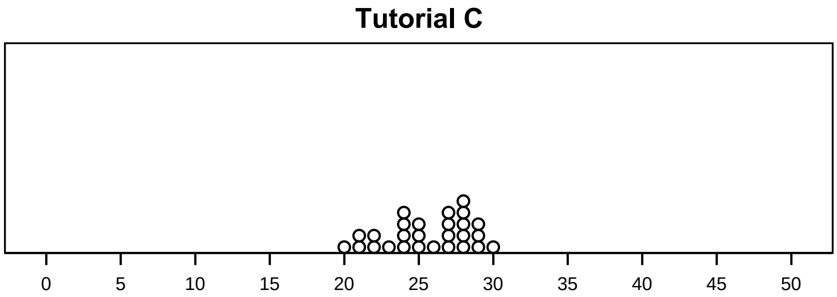
Table 7.3.3.5

Responses to "Tute B" by ability group. Because some students gave more than one response, the total number of responses is greater than the number of students.

Ability Group	A	B	C	D
Response	(n = 13)	(n = 31)	(n = 25)	(n = 6)
Made up				
Too many answers on too few values	2	6	1	
Would not get zero and/or 50 evens out of 50	3	6		
No 25s/not clustered around 25	4	3	1	
The range is too wide	11	8	1	
Real				
More random than Tutorial A		11	11	
No response			5	3

Approximately 35% of the students in Group B and 44% of the students in Group C looked at the increased variation compared with tutorial A and therefore considered the results to be real. Half of the Group D students did not respond. According to the Rasch analysis, this question, with an item difficulty of 0.55 logits above the zero, was the hardest of the *Tute* questions. For some of the students who stated that the data were likely to be real, it is unclear whether or not they were just comparing the pattern with the pattern in *Tute A* or actually looking at the numbers on the horizontal axis.

"Tute C"



c. Do you think tutorial C's results are made up or really from the experiment?
   i. Made up
   ii. Real from experiment
   Explain your answer.

The results in the plot are clustered around 25 with some variation, and the range is from 20 to 30. Therefore these results could be real. Students who gave more than one reason for saying "real" were given a code of "2," students who gave one reason only were given a code of "1" and students who said the results were made up received a code of "0." Table 7.2.4.1 indicates that 72% of the students recognised the results as likely to be real. According to the Rasch analysis, this item with an item difficulty of 0.25 logits above zero, was the second most difficult of the *Tute* questions. Table 7.3.3.6 gives the students' responses according to their ability.

Table 7.3.3.6

Responses to Tutorial C according to ability group. Because some students gave more than one response, the total number of responses is greater than the number of students.

Ability Group	A	B	C	D
Response	(n = 13)	(n = 31)	(n = 25)	(n = 6)
Made up				
Range is too small		4	3	1
Real				
Grouped around 25 with some variation	11	10	4	2
The range is reasonable	3	3	3	
More varied than tutorial A		4	4	
Shows randomness		1	2	
No response		5	8	2
Idiosyncratic*			3	1

* Idiosyncratic – these responses were either incomprehensible, or did not address the question.

7.3.4 Questions requiring judgements of differences between groups

For the items in this section the students were required to make judgements of differences between groups without the use of formal statistical procedures. For the first item (*Med*) the students were required to compare the number of people whose eczema improved for those who were treated and the control group. As there were unequal numbers of people in the two groups proportional reasoning was required to answer the item successfully.

For the second series of items the students were required to compare the results of a test between two groups, where the results were shown in graphical form. For the first pair, the two groups had equal numbers and all of one group did better than the other (*A-B*). For the second pair the two groups again had equal numbers and there is some overlap in the results, but one group clearly had a higher mean score than the other (*C-D*). In the third pair there were equal numbers in each group and whereas the means, medians, and modes were equal, one group had a wider spread than the other group (*E-F*). For the last pair (*G-H*) the group numbers were not identical. It was expected that students would have to make a judgement of the value of either the mean or median to make a decision as to which group performed "better." Students were not expected to be familiar with ideas of statistically significant differences.

"Med"

A new medication is being tested to determine its effectiveness in the treatment of eczema, an inflammatory condition of the skin. Thirty patients with eczema were selected to participate in the study. The patients were randomly divided into two groups. Twenty patients in an experimental group received the medication, while ten patients in a control group received no medication. The results after two months are shown below.

	Experimental Group (Medication)	Control Group (No Medication)
Improved	8	2
No improvement	12	8

Based on this data, you think the medication was:

A. Somewhat effective B. Basically ineffective

If you chose option A, select the one explanation below that best describes your reasoning.	If you chose option B, select the one explanation below that best describes your reasoning.
a. 40% of the people (8/20) in the experimental group improved	a. In the control group, 2 people improved even without the medication.
b. 8 people improved in the experimental group while only 2 improved in the control group	b. In the experimental group, more people didn't get better than did (12 vs. 8).
c. In the experimental group, the number of people who improved is only 4 less than the number who didn't improve (12-8) while in the control group the difference is 6 (8-2).	c. the difference between the numbers who improved and didn't improve is about the same in each group (4 vs. 6).
d. 40% of patients in the experimental group improved (8/20), while only 20% improved in the control group (2/10)	d. In the experimental group, only 40% of the patients improved (8/20).

Students who compared the groups to each other using proportional reasoning (Ad) received the highest score of “2.” Table 7.2.4.1 indicates that 53% of the students were able to do this. Students who used proportional reasoning but only mentioned the results of one group (Aa, Bd), or used the raw numbers and not the proportions (Ab) received a score of “1” (13%). All other students received a score of “0.” Table 7.3.4.1 shows the responses of the students according to ability group. This item had an item difficulty of 0.15 logits below the mean, indicating that the item was of average difficulty.

Table 7.3.4.1

Responses to “Med” according to ability group

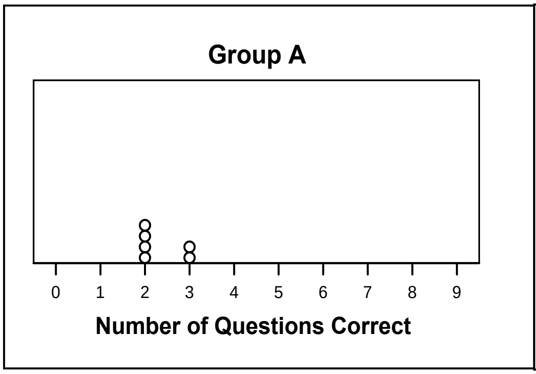
Ability Group	A	B	C	D
Responses	(n = 13)	(n = 31)	(n = 25)	(n = 6)
Somewhat effective				
Option Ad	10	18	11	2
Option Aa		1	3	
Option Ab		1		1
Basically ineffective				
Option Bd	3	2	3	1
Option Ba		2	1	
Option Bb		1	2	
Option Bc		3	4	
No response		3	1	2

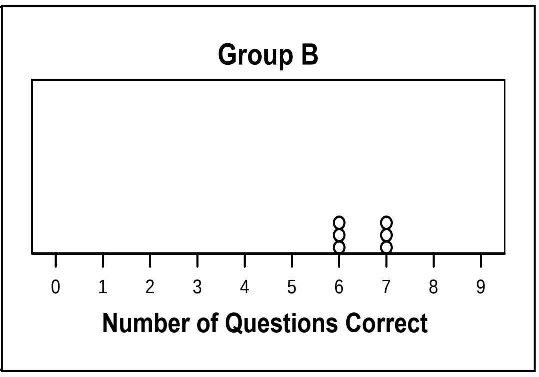
Comparisons between groups

“A-B”

A tertiary institution is comparing the scores of some tutorial groups on a test of basic statistics facts. The test had nine questions.
The scores for two of these tutorial groups are shown in the charts below. Each circle represents one person. Therefore for Group A four people answered two questions correctly, and two people answered three questions correctly.

Did the two groups perform equally well, or did one group perform better? Please give reasons for your answer.





There was a wide variety of responses. Those students who received the highest score of “3” either stated that all of Group B had a higher results than all of the other group, estimated the means but did not calculate them, or used totals while also indicating that the group sizes were equal. Scores of “2” were given for responses that referred to Group B’s results without referring to the results of the other group. Scores of “2” were also given for responses where the means had been calculated fully (a less efficient strategy for this item) or for responses that referred to the totals of each group but did not make the equal group sizes explicit. A score of “1” was given to those responses that stated Group B had “more questions correct" or similar without further explanation. Table 7.2.4.1 shows that 88 % of the students stated that Group B had a better performance than Group A. The responses, according to ability group, are in Table 7.3.4.2. It is apparent that the students in the ability Groups A and B were much more likely to give complete explanations than those in the other two groups.

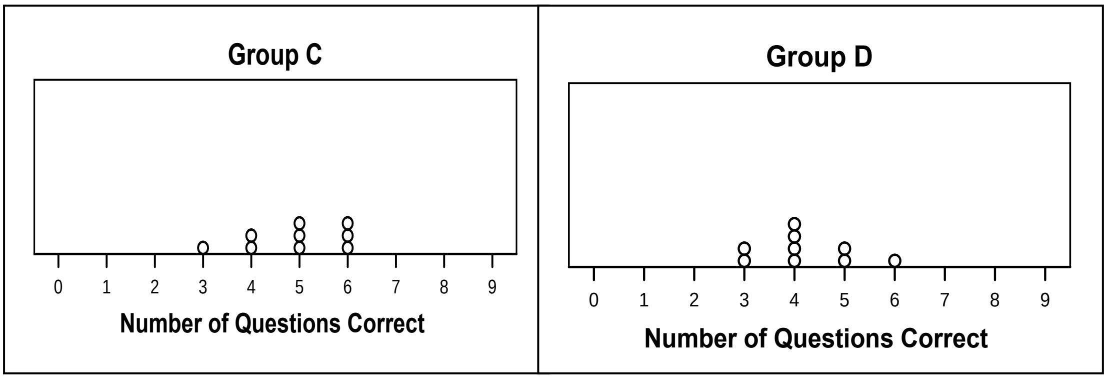
Table 7.3.4.2

Responses to "A-B" by ability group

Ability Group	A	B	C	D
Response	(n = 13)	(n = 31)	(n = 25)	(n = 6)
Entire group in B did better than the entire group in A	6	6	2	
Mean and/or median used in answer	2	7	1	
Group B had "more" correct answers	3	8	10	1
Totals used		1		
Scores of one group compared with the other		5	3	
Scores of one group mentioned only		2	4	1
"B did better" – no further explanation			2	
Group A did better	1	1	2	
No response	1	1	1	2
Idiosyncratic*				2

*Idiosyncratic – indicates that the response did not have any relationship with the question or was unintelligible

“C-D”



> Did the two groups perform equally well, or did one group perform better? Please give reasons for your answer.
>
> (For this question group C has a mean of 4.9, and group D has a mean of 4.2. This information was not given to the students.)

Again a wide variety of responses was given. A score of “3” was given to those students who calculated/estimated the means, used the totals and stated that the number of people in each group was equal, or used the frequency of the number of people with each score. A score of “2” was given to the students who stated the scores of one group and not the other, and a score of “1” was given to those students who stated that one group did “better” or similar, without further explanation. Table 7.2.4.1 indicates that 73% of the students correctly chose Group C as having the better performance. Table 7.3.4.3 shows the responses by ability. This shows that in Ability Groups B and C the most common reason given was that “more” people got “more” correct, without giving a full explanation.

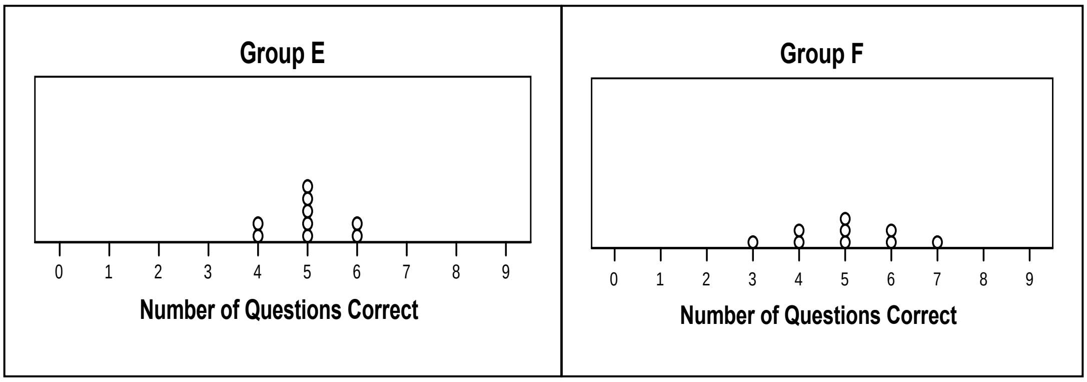
7.3.4.3

Responses to "C-D" by ability grouping

Ability Group	A	B	C	D
Response	(n = 13)	(n = 31)	(n = 25)	(n = 6)
Mean and/or median used in answer	4	3	1	
Totals used	1	1	1	
Scores of one group compared with the other	3	6	6	
C "performed better" – no further explanation		2	1	
"More people in group C got "more" correct	4	12	10	2
The groups were equal		4	3	
Group D did better	1	1		1
No response		1	3	2
Idiosyncratic*		1		1

*Idiosyncratic – indicates that the response did not have any relationship with the question or was unintelligible

"E-F"



Did the two groups perform equally well, or did one group perform better? Please give reasons for your answer.

This question confused some students who tried to say that one group had "more" higher or lower results than the other. It is apparent that these students did not regard the mean or the median as a balancing point, as in this example it is clear that every lower score is balanced by a higher score. It is also apparent from Table 7.3.4.4 that some students made their judgements according to the range, saying that this one group was "more consistent" than the other. Overall, 55% stated that the two groups performed equally well.

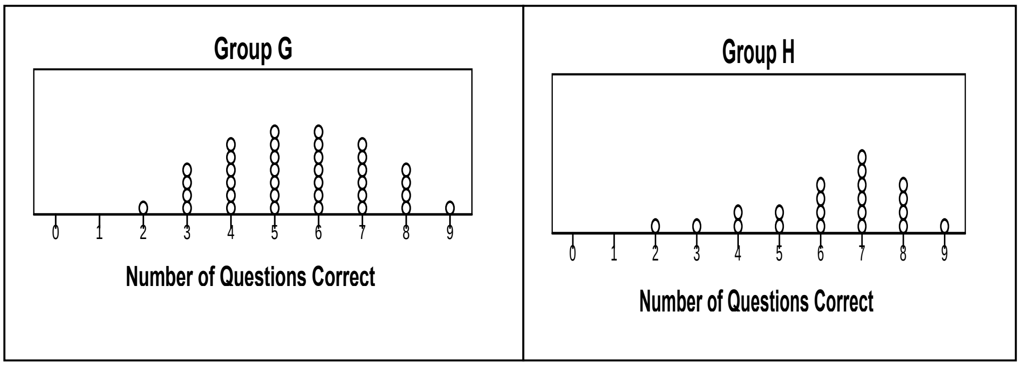
Table 7.3.4.4

Responses to "E-F" by ability grouping

Ability Group	A	B	C	D
Response	(n = 13)	(n = 31)	(n = 25)	(n = 6)
Equal				
Means and/or median used in response	7	8	4	
Totals used	2	4	3	1
No explanation	1	2	4	
Group E				
More consistent		3		
More got higher scores	1	6	6	
Averages used			1	1
No explanation		3	3	
Group F				
More got higher scores	1	1	1	
No explanation	1		1	1
No response		4	2	2
Idiosyncratic*				1

*Idiosyncratic – indicates that the response was unintelligible

“G-H”



Did the two groups perform equally well, or did one group perform better? Please give reasons for your answer.

(The average of group G is 5.5 and the average group H is 6.2. This information was not given to the students.)

Students who estimated or calculated the mean or median, or used proportional reasoning received the highest score of “3.” Those students who stated that group H had a higher mean or stated that most of Group H had “higher results” with no further explanation got a score of “2,” while those who just stated that Group H got a “better” results with no further explanation received a score of “1.” From Table 7.2.4.1 it is apparent that only 51% of students stated that Group H had the superior performance.

It was expected that the students would use some sort of proportional reasoning (a higher proportion of the group H group had higher scores) or would use a measure of central tendency such as the mean or median to answer this last question. Table 7.3.4.5, however, shows that many students chose Group G as having the best performance because there were more people in Group G, or because they thought there were “more” people in the higher range or because they thought

Group G was more “balanced.” Some of the students said the problem could not be solved, or was not fair, and there were some who gave no explanation at all. It is expected that all students attending university would have carried out the algorithm for the mean during their previous education. Therefore these incorrect responses suggest that there is a poor understanding of the reasons for which means are calculated. This example also demonstrates the proposition of Gal, Rothschild and Wagner (1990) that students, having learned a statistical skill, may not actually use the skill when it is needed outside of the classroom environment in which it was originally learned.

Table 7.3.4.5

Responses to “G-H” by ability group

Ability Group	A	B	C	D
Response	(n = 13)	(n = 31)	(n = 25)	(n = 6)
Group H				
Group H has a higher mean/median	10	8	2	
Proportional reasoning – a higher proportion within the group has higher scores	2	3	7	1
More got higher scores			1	
No explanation		3	2	
Too hard/cannot be done	1	2	1	
Not fair			1	
Group G				
More people therefore performed better		3	4	
More in higher range		2	2	
More balanced		1	1	
No explanation		3	2	2
Equal		4	2	
No response		2		3

7.3.5 Conditional probability questions

The conditional probability questions were in three forms. The first question required the students to calculate a simple conditional probability where the data were presented in the form of a table. According to Watson and Kelly (2007), students tend to find conditional probability questions easier when in the form of a table than in verbal form. The second question asked the students to distinguish between two conditional probability statements and to determine which probability was higher. The third question required the students to calculate probabilities in a simple context. However, whereas the first probability was in the "forward" direction (what is the probability of the second event if the first was ...?), the second probability was in the "backward" direction (what is the probability of the first event if the second event was ...?). Research by Fischbein and Schnark (1997) shows that many people do not realise that although the second event does not affect the first, the first probability can be revised as a result of knowledge of the second event (the "time-axis fallacy" – see Section 2.4.3).

“Factory”

The table below shows the number of defective TV’s produced every week at two factories by the day shifts and by the night shifts.

	Factory A	Factory B
Day	40	30
Night	40	60

a. How many defective TV’s are produced at Factory B every week?

b. How many defective TV’s are produced by a night shift every week?

c. If you were told that a defective TV was produced at Factory A, what is the probability it was produced by a day shift?

From Table 7.2.4.1 it is apparent that 87% of the students gave the correct response of (c), and thus received a score of “1.” With this high number of correct responses, the item had a *t*-score that was very low, so that the item did not fit on the unidimensional scale. The number of incorrect responses, by ability group, is shown in Table 7.3.5.1. It is evident that students in Group D were considerably more likely then students in other ability groups to have made an error.

Table 7.3.5.1

Number and percentage of incorrect responses to “Factory” by ability group

Ability Group	A	B	C	D
Response	(n = 13)	(n = 31)	(n = 25)	(n = 6)
Number of incorrect responses	2	2	2	4
Percentage of group incorrect	15	6	8	67

“Teacher”

Which probability do you think is bigger?

a. The probability that a woman is a schoolteacher
OR
b. The probability that a schoolteacher is a woman.
OR
c. Both (a) and (b) are equally likely.

Please explain your answer.

The highest score of “2” was given to those students who chose “b” and could explain their thinking; Table 7.2.4.1 shows that 64% of students obtained this score. A score of “1” was given to students who stated “b” but used personal experience for their justification; this score was given to 16% of the responses. The item difficulty was 0.61 logits below zero, and therefore it was one of the easier items in the questionnaire. The responses, by ability group, are shown in Table 7.3.5.2.

Table 7.3.5.2

Responses to “Teacher” by ability grouping

Ability Group	A	B	C	D
Response	(n = 13)	(n = 31)	(n = 25)	(n = 6)
There are two choices for a teacher, male or female, but there are many occupations for a woman to choose	5	15	11	1
The probability of a schoolteacher is approximately 50%, the probability of a woman being a schoolteacher is much less	7	5	1	
More women than schoolteachers		2	6	1
Personal experience		1	1	
Both (a) and (b) mean the same thing	1	6	2	1
No response		2	2	2
Idiosyncratic*			2	1

*Idiosyncratic – indicates that the response did not have any relationship with the question or was unintelligible

“Urn A” and “Urn B”

An urn has 2 white balls and 2 black balls in it. Two balls are drawn out without replacing the first ball.

a. What is the probability that the second ball is white, given that the first ball was white? Please explain your answer (Urn A).

b. What is the probability that the first ball was white, given that the second ball was white? Please explain your answer (Urn B).

The answer to both questions is 1/3. For both questions a score of “2” was given for correct responses with an explanation. A score of “1” was given for correct responses where the probability only was stated, or for those responses where the joint probability (one in six) was calculated. Table 7.2.4.1 shows that whereas 84% of the students were able to get a score of “2” or “1” for *Urn A*, only 19% could do so for *Urn B*. *Urn A* was one of the easiest items on the questionnaire; *Urn B*, however, was the hardest item of all. The difference in item difficulty scores was 2.5 logits. The responses, according to student ability groups, are found in Tables 7.3.5.3 and 7.3.5.4.

Table 7.3.5.3

Responses to "Urn A" by ability grouping.

Ability Group	A	B	C	D
Response	(n = 13)	(n = 31)	(n = 25)	(n = 6)
1 in 3 – There are three balls left, one of which is white	12	26	16	3
1 in 6 – calculated a joint probability	1	4	3	
Other numerical answer, excluding 1 in 6			5	2
No response		1	1	1

Table 7.3.5.4

Responses to "Urn B" by ability grouping

Ability Group	A	B	C	D
Response	(n = 13)	(n = 31)	(n = 25)	(n = 6)
1 in 3	2	4	3	
1 in 6 – calculated a joint probability	2	1		
1 in 2 – result of second draw cannot affect first draw	6	12	11	1
1 in 4	1	4	3	2
No response	2	10	8	3

Almost all of the students who answered correctly for *Urn A* stated that since there were originally four balls, and one had been removed, the remaining probability was 1 in 3. Almost all of the 44% of the students who answered *Urn B* incorrectly stated that the result of the second draw could not influence the first draw. This confirms the presence for some students of the time-axis fallacy described by Fischbein and Schnark (1999).

Of the nine students (12% of the total) who gave the correct response for *Urn B*, only three of them gave a full explanation. Another drew a diagram and two others gave partial explanations, stating that they were not confident that they were correct. The other three students gave no explanation at all.

7.4 Summary and discussion

Almost all of the students could successfully interpret simple probability statements as shown by their responses to *Snakes*, *Cancer* and *Eczema*. However, the results show that although most (65%) of the students were aware that the outcome of a coin toss is independent of the outcomes before it, only 39% of the students could successfully calculate the probability of getting four heads in a row.

The responses to *Hospital* indicate that although many of the students were aware that each birth was independent, they were not aware that the sample size had an influence on the likelihood that a statistic (here the number of girls born in a day) would deviate from the expected value. For this sample of students, those who had no previous statistical experience were more likely to give a correct response. It would be interesting to see if this result were repeated with another study.

The *Spinner* questions showed that 76% could state that they would not expect the same outcome in repeated random experiments, which leaves almost one quarter of the students not able to do so. This has worrying implications for students of statistics, where the whole subject is based on the idea that individual and sampling variation is universal. It is also apparent that the students did not think through which outcomes would be unexpected in the *Spinner* scenario. Only 12% of the students, when asked which outcomes would be unlikely, gave responses in terms of greater than AND less than given numbers. For the answer to *Tute A*, where the given outcomes were completely symmetrical, 75% of the students indicated the results were made up. For the answer to *Tute B*, where there were several "unreal" outcomes, only 60% of the students stated that the results were made up, and for *Tute C*, where the results were reasonably to be expected in this situation, 72% of the students stated that the results were real. In each case there was a substantial minority of students who did not give the correct response. Overall, only 37% of the students responded correctly for all three *Tute* questions.

For the conditional probability questions, 87% of the students could calculate the conditional probability when the data were presented in a table. For the school-teacher problem, 80% of the students could distinguish appropriately between the two statements. The simple forward conditional problem (*Urn A*) was answered successfully by 84% of the students, whereas only 19% could successfully answer the "backward" probability (*Urn B*). In general, the students thought that because the choice of the second ball would not influence the choice of the first ball they did not appreciate that probabilities in the past could be reassessed in the light of later information.

The responses to the questions *Med*, and *A-B*, *C-D*, *E-F* and *G-H* show that some of the students did not use proportional reasoning when required. In addition, only 55% of the students correctly answered *E-F* where the mean, median, and modes were equal, but the ranges were different. The idea of the mean as a balancing point or a representative number did not seem to be present. Their responses confirmed the findings of Groth and Bergner (2006) that some students come to university without any more knowledge of the mean apart from the algorithm used to calculate it.

The questionnaire was designed so that no formal statistical experience would be necessary to answer the questions, and it was hoped that the questions would be within the capabilities of students who were comfortable with numbers. The results of the first questionnaire confirm that students entering university, even with pre-tertiary mathematics qualifications, may have some misconceptions about stochastic processes. They may expect short runs in a repetitive trial to be more like the long term expected value than will happen in reality, and also have unrealistic views about what the likely outcomes are for a stochastic process (Tversky & Kahneman, 1982b). In addition, some students expected that the outcomes of stochastic trials to be repeated without variation.

The responses to the first questionnaire also demonstrate that students may find conditional probabilities easier when the data are in tabular from than when the data are presented in a verbal form (Watson & Kelly, 2007). Some students demonstrated internally inconsistent views on probability. For example, they may have stated that a head is more likely after four tails in a row, but then said the probability of a tail on the next toss is “still” 50%.

Finally, the responses of some students, even though they all had post Grade 10 mathematics experience, suggested that these students were not familiar with the mean as a representative number that could be used to make comparisons between sets of data (Groth & Bergner, 2006). In addition, some of these students did not use proportional reasoning when it would have assisted in making comparisons between sets of data.

This chapter described students’ responses to a questionnaire that was given at the beginning of their first unit in statistics on entering university. The next chapter describes their responses to a questionnaire that was given to them at the end of the unit, and tests for their understanding of randomness, random allocation, and significant differences.

8. Results of the Quantitative and Qualitative Analysis of the Second Questionnaire

8.1 Introduction

This chapter begins with the results of the Rasch analysis of the second questionnaire. This is then followed by a qualitative analysis of the responses given for each item by the students at varying levels of ability. The relationship between the ability of the students, as judged by the first questionnaire, and the ability of the students as judged by the second questionnaire is then investigated. This is followed by an investigation of the relationship of the ability score from the first questionnaire and the students' final mark from their formal assessment. An investigation into the relationship between the students' ability scores from the second questionnaire and the students' final mark from their formal assessment is also included.

All of the Rasch analyses were carried out using Winsteps version 3.70.1.1 (www.winsteps.com.htm). Each analysis used the Partial Credit Model. Details of the theory of this model can be found in Section 4.5.2. As explained in Section 7.1, in testing for differences in means among three or more groups (over the three semesters), the Kruskal-Wallis H was used instead of the Analysis of Variance (ANOVA).

Thirty-three students completed the second questionnaire. The reduction in numbers is because some students dropped out of the unit, some were absent on the day the second questionnaire was administered, and, owing to circumstances beyond the researcher's control, the second questionnaire was not administered to the students in the third cycle of the intervention. In the pre-intervention semester 9 students answered the second questionnaire, 20 did so in the first cycle of the intervention, as did 4 in the second cycle of the intervention.

8.2. Rasch analysis of the second questionnaire

8.2.1 Introduction

The second questionnaire was administered at the end of the Data Handling and Statistics unit. It was designed to test students' ability to explain informally the processes behind *P*-values and hypothesis testing and to explain why randomness in inferential statistics is so important.

8.2.2 Items in the Second Questionnaire

There were 17 items in the second questionnaire. The students were first required to give an example of a random event and explain why it was random. They were also asked to judge the likelihood of two sample means based on the number of standard errors these means were away from the proposed population mean, and to give an explanation. For some of the items the students were asked to explain why random sampling or allocation was important. They were also required to explain the meaning of "significant difference" for a scenario where one sample mean was higher than the other, and to make an inference based on graphical data.

Four questions were repeated from the first questionnaire (*Hospital, UrnA, UrnB* and *Med);* these were used to anchor the analysis so that a comparison could be made between the ability level of the students at the first and second questionnaires. A paired *t*-test was then used to see if there was a significant change in ability. In the second questionnaire, the question *Med* (renamed *Med A*) had an addition (named *Med B*) requiring the students to choose from a list of comments about the validity of the experiment that produced the *Med* data. A summary of the items on the second questionnaire with the labels used in the Rasch analysis is found in Table 8.2.2.1.

Table 8.2.2.1

A description of the items used in the Rasch analysis.

Label	Description of knowledge shown
Random	Can give an example of a random event and explain why it is random.
CB 1	Can calculate a simple expected value.
CB 2	Can judge the probability of an event in words – very likely, fairly likely, possible, unlikely or very unlikely.
CB 3	Can use hypothetical reasoning based on the probability to make judgements.
Fish	Can explain the meaning of a "significant difference," and the meaning of a *P*-value.
Cereal A	Can make judgements on the likelihood of an event based on the number of standard errors the event is from the proposed population value.
Cereal B	Can make judgements on the likelihood of an event based on the number of standard errors the event is from the proposed population value.
Cereal C	Can make an estimate of a value that is unlikely given a proposed value.
Cereal D	Can explain the importance of random sampling.
Med A	Can make a judgement on the effectiveness of a treatment using proportional reasoning.
Med B	Can make judgements on the experimental design used in *Med A*.
Urn A	Can calculate a simple forward probability.
Urn B	Can calculate an event in the past given later information.
Pacific A	Can make a judgement on the difference between two groups using both the centre and variation of the results.
Pacific B	Can explain the reasons for random allocation.
Hospital	Can determine that for events that are equally likely and independent (boy or girl being born) the sample with the lowest sample size is more likely to deviate from the expected number of 50%.
Fred	Can explain the importance of random allocation.

8.2.3 The Rasch Analysis of the Items (Partial Credit Model)

In the process of the Rasch analysis it was found that some items fitted poorly to the model and these were removed and the analysis re-run. This was a step-wise process. At first only the items with the more extreme *t*-statistics, less than negative three or greater than plus three, were removed. This process continued until all the items remaining in the analysis had a satisfactory *t*-statistic within the range of negative two to plus two. By the end of this process, the items that were removed were *Hospital, Cereal A, Cereal B* and *Cereal C*. The fitting of the items in the original analysis before these items were removed can be seen in the bubble chart in Figure 8.2.3.1. Because of the way the items' arrangement changed as the items with very large *t*-statistics were removed, some items that appeared to have a poor fit in the first analysis remained in the final analysis. This final analysis had an item reliability score of 0.90. Table 8.2.3.1 shows the list of items retained in the final analysis with their mean levels of difficulty. This shows that the hardest question was *Med B.* This is the question where the students were asked to make judgements about the experimental design used in *Med A*. The variable map in Figure 8.2.3.2 shows the items with their respective Rasch-Thurstone thresholds.

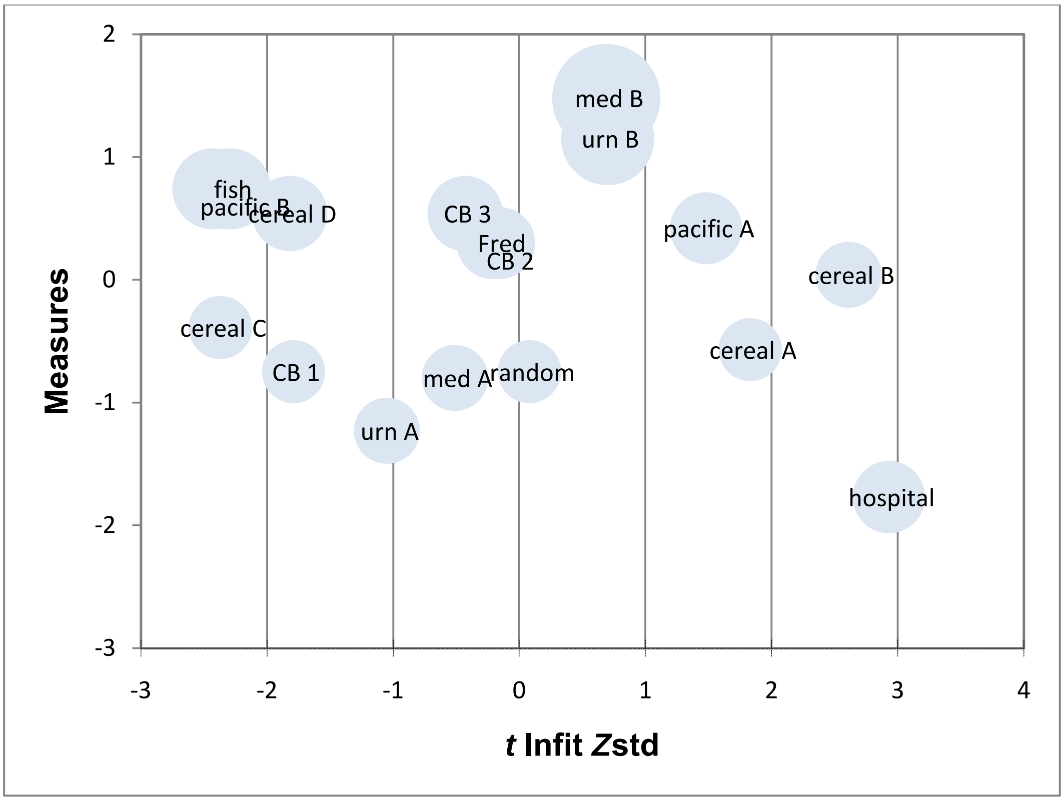


Figure 8.2.3.1. Bubble chart of the items of the second questionnaire.

Table 8.2.3.1

Items of the second questionaire in order of mean difficulty

NAME	MEASURE
Med B	2.61
Urn B	2.08
Pacific B	1.83
Fish	1.83
Cereal D	1.61
CB 3	1.61
Pacific A	1.48
CB 2	1.35
Fred	1.35
CB 1	0.80
Random	0.80
Med A	0.02
Urn A	-0.55

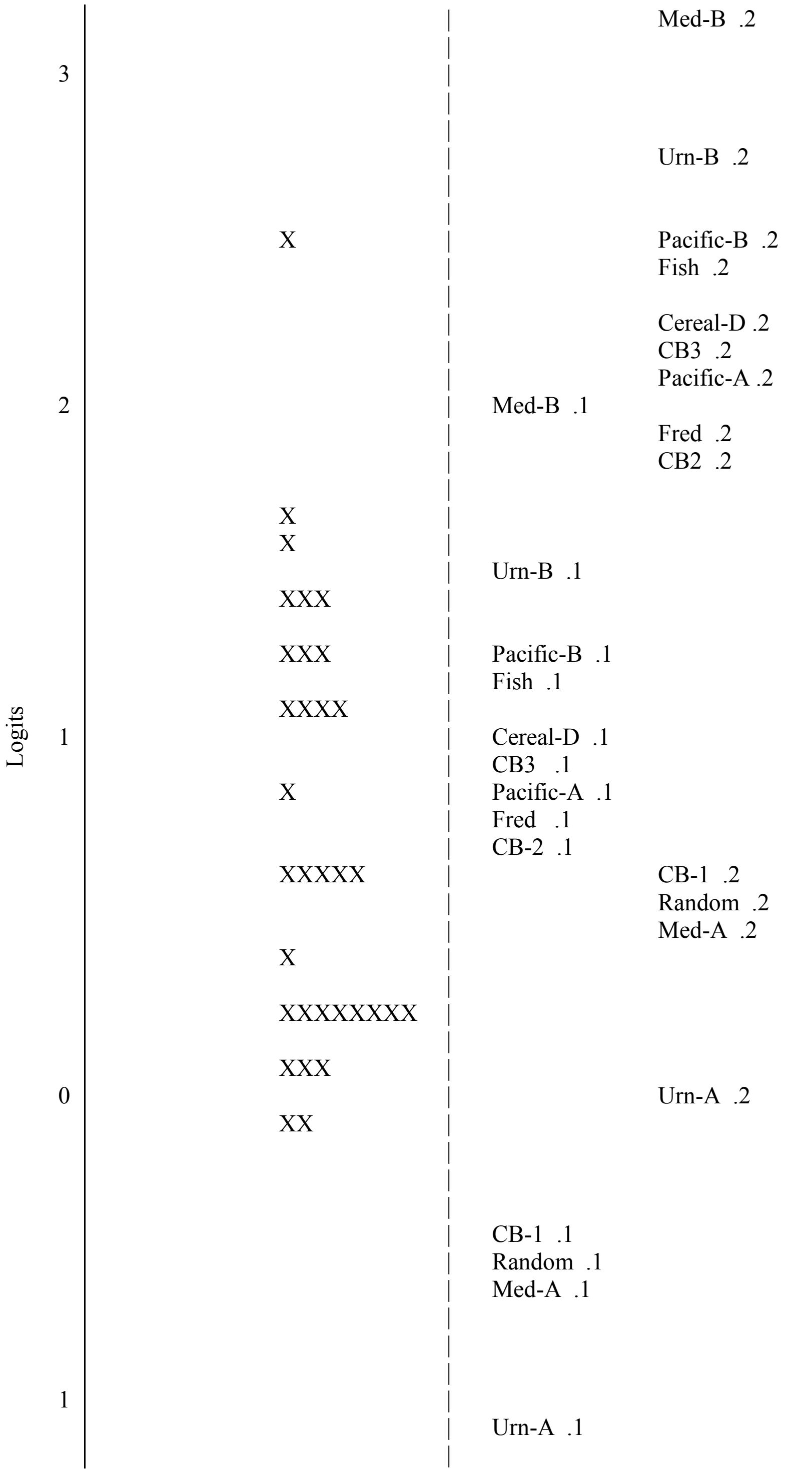


Figure 8.2.3.2. Item and person map of the Rasch analysis of the second questionnaire, showing the Rasch-Thurstone thresholds. Each "X" represents one person.

Figure 8.2.3.2 shows that, according to the analysis, all the students have a less than 50% chance of getting the highest score for *Med B* and *Urn B*. It also shows that there were ten questions that were found to be difficult by most of the students.

The items were then divided according to increasing complexity of the reasoning needed to answer the questions successfully. Examining the variable map gave tentative clusters where there were apparent jumps in difficulty. The decision about where to separate these clusters was assisted by importing the item difficulties into a spreadsheet and using these to produce a bar graph of the item difficulties in order of increasing difficulty. This bar graph is shown in Figure 8.2.3.3.

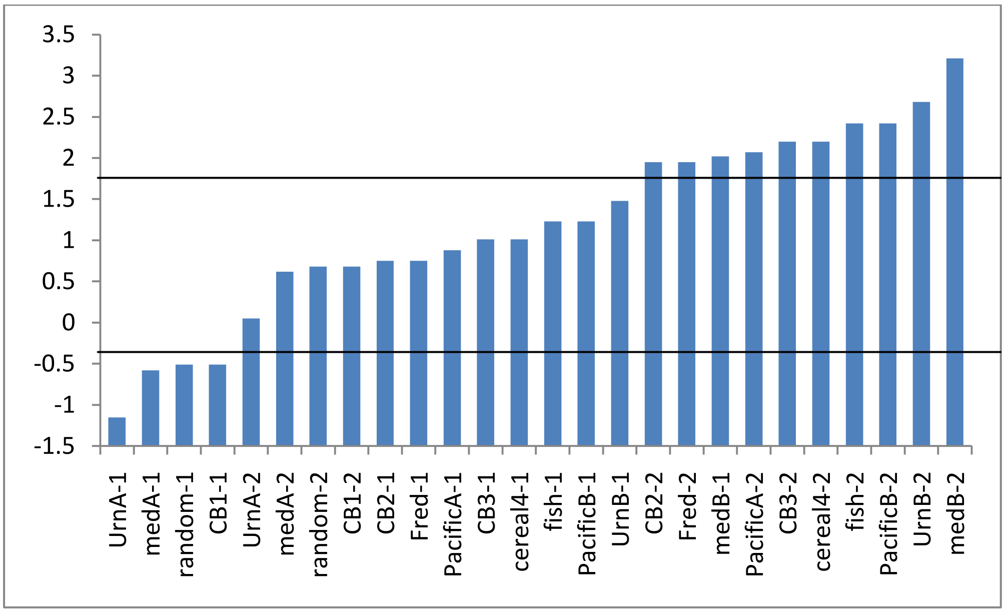


Figure 8.2.3.3. Items in order of difficulty according to their Rasch-Thurstone thresholds, showing the division of items according to the cognitive demands made by these items.

The items were then assessed on the type of resoning required by the students to answer each question successsfully at each level. As a result of this process, the items were divided according to the lines in Figure 8.2.3.3. "*Urn A* .2" was placed above the lowest group because this score for *Urn A* required the students to give an explanation.

The cognitive demands made by these items are summarised in Table 8.2.3.3. This table demonstrates that as student ability decreased, they were less likely to be able to explain the meaning of a "significant difference," to be able to explain the importance of random allocation, and less likely to use proportional reasoning when required.

Table 8.2.3.3

Summary of the cognitive demands made on the students in each cluster of items.

Description of Item	50%PRB	Scoring Rubric	Reasoning used
Med B .2	3.21	Recognises that the data given are sufficient for a conclusion to be drawn	Probabilities given in numerical form are described verbally. The importance of random allocation is recognised. The term “significant difference” is understood, and the *P*-value is defined. Both the centre and variation of the data are used when making comparisons between two data sets.
Urn B .2	2.68	Calculates a reverse conditional probability and explains	
Pacific B .2	2.42	Recognises and describes the importance of random allocation	
Fish .2	2.42	Describes the difference between observed and significant differences, defines *P*-value	
Cereal D .2	2.2	Recognises and describes the importance of random allocation	
CB 3 .2	2.2	Uses a probability in a hypothetical reasoning process to make a conclusion	
Pacific A .2	2.07	Uses both the centre and variation of two data sets to make a comparison	
Med B .1	2.02	Does not recognise that the data given are sufficient for a conclusion to be drawn	
Fred .2	1.95	Recognises and describes the importance of random allocation	
CB 2 .2	1.95	Correctly describes the level of a given numerical probability in words	

Table 8.2.3.3 (Continued)

Description of Item	50%PRB	Scoring Rubric	Reasoning used
Urn B .1	1.48	Calculates a reverse conditional probability does not explain	The importance of random allocation is recognised but explained in superficial terms. The importance of sampling variation in determining differences in data is recognised, but the centre of the data is not used as well. Proportional reasoning is used when necessary.
Pacific B .1	1.23	Recognises that random allocation of is necessary but only a superficial explanation is given (e.g., “bias introduced”)	
Fish .1	1.23	Describes that the observed differences are due to sampling variation, but does not define the *P*-value	
Cereal D .1	1.01	Recognises that random allocation of is necessary but only a superficial explanation is given (e.g., “bias introduced”)	
CB 3 .1	1.01	Bases a conclusion of a hypothetical reasoning process on the presence of possible sampling variation only	
Pacific A .1	0.88	Uses the variation of two data sets to make a comparison, but not the centre	
Fred .1	0.75	Recognises that random allocation of is necessary but only a superficial explanation is given (e.g., “bias introduced”)	
CB 2 .1	0.75	Over or understates the level of a given probability in words	
CB 1 .2	0.68	Correctly calculates the value of a simple expected value	
Random .2	0.68	Correctly identifies a random process and gives suitable reasons for answer	
Med A .2	0.62	Uses proportional reasoning when comparing two data sets where the results are given as the percentage of success	
Urn A .2	0.05	Calculates a forward conditional probability and explains answer	

Table 8.2.3.3 (Continued)

Description of Item	50%PRB	Scoring Rubric	Reasoning used
CB 1 ".1"	-0.51	Correctly calculates the value of a simple expected value but inappropriately rounds up the answer	The necessity for proportional reasoning when necessary is not recognised. Can calculate simple probabilities in simple contexts and identify a random process. Explanations are not usually given, and when given, are inaccurate.
Random ".1"	-0.51	Correctly identifies a random process but reasons for answer are not given	
Med A ".1"	-0.58	Does not use proportional reasoning when comparing two data sets where the results are given as the percentage of success	
Urn A ".1"	-1.15	Calculates a forward conditional probability but does not explains answer	

8.2.4 Rasch analysis of persons

The analysis of persons had a person reliability score of 0.32 (see Appendix E2). This shows that the level of replicability of person ordering that would be expected if a test of similar items were given to the same people is not high. A histogram of the person ability scores for the second questionnaire is found in Figure 8.2.4.1. The histogram shows that the students who completed the second questionnaire had an ability score between -0.5 and 2.5 logits. In the previous questionnaire, the majority of the students had ability scores between -1.5 and 2.0 logits. There is no significant difference in students' ability among the three semesters where the second questionnaire was given (see Appendix E3).

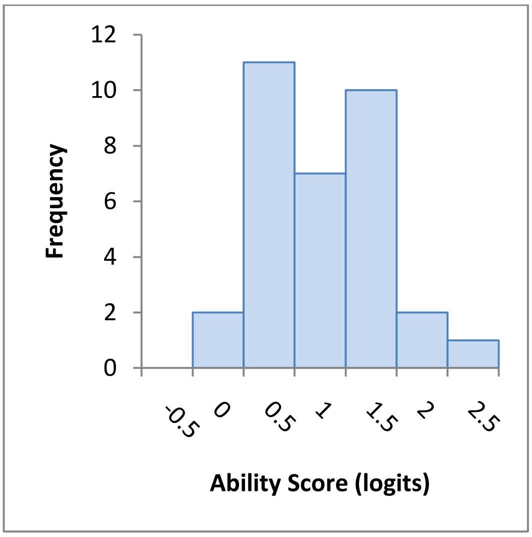


Figure 8.2.4.1. Histogram of person ability scores for the Rasch analysis of the second questionnaire.

The students were then divided into three groups, according to the levels in the items of difficulty reported in Table 8.2.3.3. These groups were labelled, from highest to lowest as Group A, Group B and Group C. Figure 8.2.3.2 demonstrates that there was only one student who had a better than 50% chance of answering some of the more difficult items correctly, and that overall the students found the second questionnaire difficult. As a result there were very few students in group A, and most of the students were in group B. As the qualitative analysis described in Section 8.3 progressed it became apparent that the students at the higher level of Group B were giving answers at greater depth, and thus receiving higher scores, than the students at the lower level of Group B. As a result, Group B was divided into two (Groups B_1 and B_2) according to the dotted line in Figure 8.2.4.2. This resulted in three students in Group A, 16 students in Group B_1, 12 students in Group B_2 and two students in Group C.

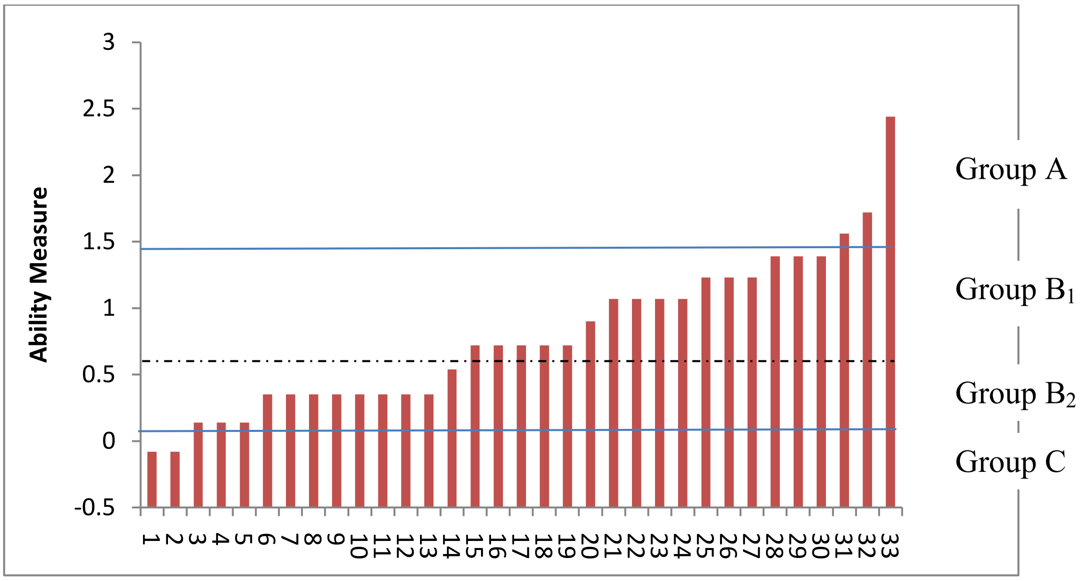


Figure 8.2.4.2. Ability levels of students.

Table 8.2.4.1 shows the frequency of the scores the students received for the questions in the second questionnaire. There were no significant differences in the scores among the three semesters for any of these items (see Appendix E2).

Table 8.2.4.1

Frequency of scores received by the 33 students for each of the items on the second questionnaire

	Score			
Item	3	2	1	0
CB 1				
Pre-intervention	N/A*	5	4	0
First cycle of intervention	N/A	11	5	4
Second cycle of intervention	N/A	1	2	1
Combined	N/A	17	11	5
CB 2				
Pre-intervention	N/A	0	5	4
First cycle of intervention	N/A	4	6	10
Second cycle of intervention	N/A	1	2	1
Combined	N/A	5	13	15
CB 3				
Pre-intervention	N/A	1	5	3
First cycle of intervention	N/A	3	5	12
Second cycle of intervention	N/A	0	1	3
Combined	N/A	4	11	18
Fish				
Pre-intervention	N/A	0	6	3
First cycle of intervention	N/A	0	11	9
Second cycle of intervention	N/A	0	2	2
Combined	N/A	0	19	14
Fred				
Pre-intervention	N/A	1	4	4
First cycle of intervention	N/A	6	4	10
Second cycle of intervention	N/A	0	1	3
Combined	N/A	7	9	17
Cereal A				
Pre-intervention	4	1	3	1
First cycle of intervention	7	8	4	1
Second cycle of intervention	1	0	2	1
Combined	12	9	9	3
Cereal B				
Pre-intervention	2	1	0	6
First cycle of intervention	0	5	5	10
Second cycle of intervention	1	1	0	2
Combined	3	7	5	18
Cereal C				
Pre-intervention	N/A	4	4	1
First cycle of intervention	N/A	5	10	5
Second cycle of intervention	N/A	1	3	0
Combined	N/A	10	17	6

Table 8.2.4.1 (Continued)

	Score			
Item	3	2	1	0
Cereal D				
Pre-intervention	N/A	0	4	5
First cycle of intervention	N/A	0	13	7
Second cycle of intervention	N/A	0	2	2
Combined	N/A	0	19	14
Random				
Pre-intervention	N/A	4	1	4
First cycle of intervention	N/A	14	4	2
Second cycle of intervention	N/A	2	0	2
Combined	N/A	20	5	8
Urn A				
Pre-intervention	N/A	7	0	2
First cycle of intervention	N/A	16	2	2
Second cycle of intervention	N/A	3	1	0
Combined	N/A	26	3	4
Urn B				
Pre-intervention	N/A	1	1	7
First cycle of intervention	N/A	2	3	15
Second cycle of intervention	N/A	0	0	4
Combined	N/A	3	4	26
Med A				
Pre-intervention	N/A	4	3	2
First cycle of intervention	N/A	12	5	3
Second cycle of intervention	N/A	3	0	1
Combined	N/A	19	8	6
Med B				
Pre-intervention	N/A	0	1	8
First cycle of intervention	N/A	2	2	16
Second cycle of intervention	N/A	0	1	3
Combined	N/A	2	4	27
Hospital				
Pre-intervention	5	3	1	0
First cycle of intervention	3	7	9	1
Second cycle of intervention	2	0	1	1
Combined	14	6	11	2
Pacific A				
Pre-intervention	N/A	2	1	6
First cycle of intervention	N/A	12	2	6
Second cycle of intervention	N/A	1	0	3
Combined	N/A	15	3	15
Pacific B				
Pre-intervention	N/A	0	4	5
First cycle of intervention	N/A	0	8	12
Second cycle of intervention	N/A	0	4	0
Combined	N/A	0	16	17

*N/A- not applicable – this score was not available for these questions

8.3 Qualitative analysis of the second questionnaire

In this section the students' responses are compared across the ability groups, to see what types of reasoning the students used and how these varied between the higher and lower ability groups. Because it was desired that the knowledge gained would be as complete as possible, all the items are included, regardless of whether or not they fitted the unidimensional scale in the Rasch analysis.

8.3.1 The circuit breaker questions

These questions required the students to make a judgement of the likelihood of getting 3 defective circuit breakers in a box of 25 if the underlying rate of defectives for all the circuit breakers was 5%. The students were required to use a probabilistic hypothetical process. In essence, they were required to carry out a hypothesis test but without the formal procedures.

You work for a manufacturer of circuit breakers. Owing to the difficulty of the process, it is expected that 5% of these will be defective. The occurrence of the defective breakers occurs randomly. The breakers are sold in boxes of 25.
One of your customers buys a box with three defective breakers. This is 12% of the contents of the box. Your customer is furious. You are told that your underlying rate is 12%, not 5% and they will take their custom elsewhere.

a. *If* 5% are defective overall, then *on average* how many defective breakers would you expect to find per box? **("CB1")**

b. It can be calculated that *If* the underlying rate is 5%, the probability of getting 3 or more defectives in a box is 13/100. Based on these figures, getting three or more defective circuit breakers in a box is:
   i. Very likely
   ii. Fairly likely
   iii. Possible
   iv. Unlikely
   v. Very unlikely **("CB2")**

c. Does this box provide sufficient evidence that the underlying rate of defectives for all the circuit breakers is greater than 5% as the customer claims? Explain your answer. **("CB3")**

"CB 1"

The highest score of "2" was given for the answer 1.25. The integer answer (1) resulted in a score of "1." Table 8.2.4.1 indicates that all except five of the students gave one of these answers. Table 8.3.1.1 shows the answers by student ability group. It shows that as the ability decreased the students were more likely to give the integer answer. The Rasch analysis indicates that this was the easiest of the three circuit breaker questions that remained in the analysis.

Table 8.3.1.1

Answers to "CB 1" by ability group

Ability Group	A	B_1	B_2	C
Answer	(n = 3)	(n = 16)	(n =12)	(n = 2)
1.25	3	10	5	
1		4	6	1
Other		1	1	
No answer		1		1

"CB 2"

In this question the students were required to interpret the numerical probability (13%) in words. The score "2" was given for the response "fairly likely," and a score of "1" was given for the response "very likely" or "possible." Table 8.2.4.1 indicates that 55% of the students received one of these scores. Table 8.3.1.2 shows that the other options were more likely to be given as the student ability decreased. This question was found to be harder than *CB 1*, having an item difficulty of 1.35 logits above zero, 0.55 logits above *CB 1*.

Table 8.3.1.2.

Answers to "CB 2" by ability group

Ability Group	A	B_1	B_2	C
Answer	(n = 3)	(n = 16)	(n =12)	(n = 2)
Fairly likely	2	2	1	
Very likely/possible	1	8	4	
Other		6	7	2

"CB 3"

For this question students were required to use the given probability (13%) to decide whether or not the observed box was an unusual event given the underlying rate. This probability indicates that if the underlying rate of defectives is still 5% a box with three defectives is not an unusual event. Reponses that used the probabiltiy of 13% resulted in a score of "2." Those students who stated that because randomness is always present, there will be boxes with more or less than the observed value, or carried out a formal hypothesis test without further explanation received a score of "1." Table 8.2.4.1 shows that only 45% of the students used one of these forms of reasoning. As a result, this item had a difficulty rating of 1.61 logits above zero, 0.26 logits above *CB 2*, and 0.81 logits above *CB 1*. Table 8.3.1.3 gives the answers according to the students' ability group.

Table 8.3.1.3

Answers to "CB 3" according to ability group.

Ability Group	A	B_1	B_2	C
Answer	(n = 3)	(n = 16)	(n =12)	(n = 2)
With a probability of 13%, a box with three defectives is not unusual, even if overall 5% are defective	1			
Used a formal hypothesis test without further explanation	1		4	
Variation is present, other boxes will have different numbers of defectives		4		
This box does not necessarily reflect the whole population		1		1
Cannot decide from one box		5	5	
This box is just due to random chance	1	5		
No answer		1	1	1
Idiosyncratic*			2	

* Answers were unitelligilbe

8.3.2 Explaining the meaning of "significant difference"

In the *Fish* question the students were asked to explain why, given that one treatment had a higher sample mean than the other treatment, the conclusion was that there was no significant difference in means? To explain this, the students needed to appreciate the role of sampling variation, and to think of the *sample* means that are to be expected if they both samples have come from two populations with equal means.

You are looking at the effects of supplementing trout fish feed with vitamin E. Some of the fish are given the standard commercial feed, and others are given the same feed with double the level of vitamin E. After a suitable time, you measure the weights of the fish. The results are in the table below:

	Standard feed	Extra E
Mean weight (g)	256.4	263.1
Standard deviation (g)	12.3	11.2

You perform the two sample *t*-test and the *P*-value you receive is 0.45. Therefore you tell your supervisor that the extra vitamin E has not made any difference to the mean weight of the fish.
Your supervisor says that the mean weight of the fish given „extra E’ ***is*** higher than the mean weight of those who were given the standard feed. Explain to your supervisor why you say that even though the „extra E’ feed has a higher mean weight, the „extra E’ feed has not made a *significant* difference.

For the highest score of “2” students needed to explain specifically the meaning of the *P*-value. Table 8.2.4.1, however, indicates that no student did so. The majority of the students gave an explanation in terms of the usual variation that is to be expected between samples. The answers, according to student ability, are shown in Table 8.3.2.1. This was found to be one of the harder questions, with a difficulty rating of 1.83 above zero, 2.38 logits above the easiest question (Urn A).

Table 8.3.2.1

Answers to "Fish" according to ability group

Ability Group	A	B_1	B_2	C
Answer	(n = 3)	(n = 16)	(n =12)	(n = 2)
The two means are within one standard deviation of each other, the means are close together	1	2	1	
Results show that the differences could be due to – sampling variation	1	3	1	
A "significant difference" means that the means are too far apart to be explained by normal sampling variation		2		
There is a high chance that these sample results could be recorded even if there is no difference in the overall population			1	
Results are not significant – no explanation		4	2	
P-value less than 0.05 but no further explanation given				
Hypothesis test carried out with no further explanation		1		1
P-value defined in answer but not accurately defined			2	
The difference in means is significant		1		
P-value is too high to prove H_o is incorrect				1
No answer	1	2	4	
Idiosyncratic*		1	1	

*Answers were self-contradictory

8.3.3 Judgement as to the likelihood of sample means, given a population mean

For the *Cereal* questions the students were required to judge how likely a sample mean would be given the value of the population mean. They were also asked to consider what values of the sample mean would be unlikely given the population mean and the standard error of the mean. They then had to consider the effect of non-random sampling on their conclusions.

"Cereal A"

You are working for a consumer organisation. As part of your duties, you select 49 boxes of "Get up and Go" cereal at *random* and weigh the boxes. On the label of the boxes you read that the minimum weight of the box is 800g.
The standard deviation of the weight of the boxes is 14g. Therefore the standard error of the mean (the standard deviation of all possible sample means) is estimated to be 2g.

a) Assume the manufacturer's claim that the minimum weight of 800g is correct. For your 49 boxes, is a sample mean of 799g
   i. Very likely
   ii. Likely
   iii. Possible
   iv. Unlikely
   v. Very unlikely?

Give reasons for your answer.

The highest score ("3") was given to those students who answered "very likely" and used the standard error in their reasoning. A score of "2" was given to those who chose "likely" or "possible" and used the standard error in their reasoning. Students who used the standard error appropriately, but stated that they had used the standard deviation, received a score of "1." All other answers, including those that used the standard deviation, received a score of "0." Table 8.3.3.1 gives the answers of the students according to their ability. As similar reasoning was used across all the ability levels, the item did not fit into the unidimensional scale on the Rasch analysis.

Table 8.3.3.1

Answers to "Cereal A" according to ability group

Ability Group	A	B_1	B_2	C
Answer	(n = 3)	(n = 16)	(n =12)	(n = 2)
Used reasoning based on the standard deviation, not the standard error	1			
Used reasoning that did not refer to the standard error or the standard deviation	2	4	1	
Used reasoning based on the fact that the sample mean was within one standard error of the proposed population mean		7	7	1
Used reasoning based on the fact that the sample mean was within two standard errors of the proposed population mean		2	1	
Used reasoning based on the standard error, but stated the standard deviation was used		3		
Used reasoning based on the standard deviation, not the standard error			3	
No answer				1

“Cereal B”

Cereal B followed directly from *Cereal A*. This time the sample mean is exactly two standard errors lower than the proposed population mean.

> b) Again assuming the manufacturer’s claim to be correct, for your 49 boxes, is a sample mean of 796g
>
> i. Very likely?
> ii. Likely?
> iii. Possible?
> iv. Unlikely
> v. Very unlikely?
>
> Give reasons for your answer.

Those students who answered “likely” with an appropriate explanation using standard errors received the highest score of “3.” Those responses that indicated “very likely” or “possible” with an appropriate explanation using the standard error received a score of “2.” Those students who used the standard error appropriately but stated that they had used the standard deviation received a score of “1.” Table 8.3.3.2 shows that 30% of the students received either a “3” or a “2,” with 15% of the students receiving a score of “1.” As for to *Cereal A*, similar forms of reasoning were used by the students in all the ability groups, and therefore this item did not fit onto the unidimensional scale of the Rasch analysis.

Table 8.3.3.2

Answers to "Cereal B" by ability group

Ability Group	A	B_1	B_2	C
Answer	(n = 3)	(n = 16)	(n =12)	(n = 2)
Used reasoning based on the fact that the sample mean is two standard errors from the population mean	1	10	4	1
Used reasoning that did not refer to the standard error or the standard deviation	2	2	1	
Used reasoning based on the standard deviation, not the standard error		1	4	
No answer			3	1
Idiosyncratic*		3		

*Idiosyncratic – indicates the answer did not relate to the question or was unintelligible

"Cereal C"

Here the students were required to make a prediction of the sample means that would be unlikely if the population mean was 800g. They were expected to take the standard errors into account.

c) At what value of a sample mean below 800g would you start to suspect the manufacturer's claim to be untrue? Give reasons for your answer.

Responses that indicated that sample means of three standard errors or more below 800g would be unexpected scored "2." Responses that indicated that sample means of two standard errors or more below would be unexpected resulted in a score of "1." A score of "1" was also given for those responses that used the exact number, that is, a response of "794 g" instead of "< 794 g." Those responses that used the standard error appropriately, but stated that the standard deviation had been used, also received a score of “1.” Those responses that used the standard deviation of 14 g resulted in a score of “0.” The answers, according to ability group, are in Table 8.3.3.3.

The responses to the Cereal questions indicate that there is some confusion in distinguishing between the standard error and standard deviation. Some students used the standard error in their reasoning but stated that they had used the standard deviation, and some other students used the standard deviation even though in this context it was not appropriate to do so. In addition, some students used reasoning in *Cereal A* and *Cereal B* that did not use any measure of spread of the data. For example, “There is a lot of variation.” Owing to similar forms of reasoning being used across all ability groups, the item did not fit onto the unidimensional scale of the Rasch analysis.

Table 8.3.3.3

Answers to "Cereal C" by ability group

Ability Group	A	B_1	B_2	C
Answer	(n = 3)	(n = 16)	(n =12)	(n = 2)
Less than 796 g, more than two standard errors from 800 g	1	4	2	1
Less than 794 g, more than three standard errors from 800 g	2	3	1	
Less than 796 g, more than two standard deviations from 800 g			1	
Less than 796 g, no explanation			1	
796 g, two standard errors from 800 g		2		
794 g, more than two standard errors from 800 g		3		
One standard deviation of 14 g above or below				1
Other number		4	4	
Idiosyncratic*			3	

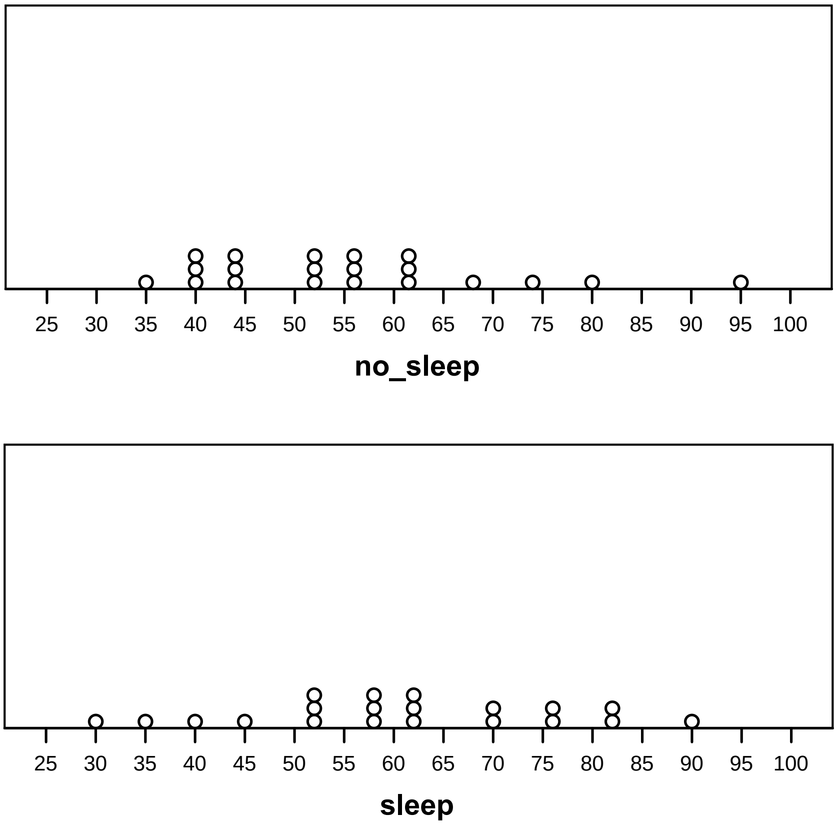
* Idiosyncratic – indicates the answer did not relate to the question or was unintelligible

8.3.4 The Use of Informal Inference

“Pacific A”

Forty students from the Pacific University participated in a study of the effect of sleep on test scores. Using random allocation, 20 of the students were required to stay up all night studying the night before the test (no-sleep group). The other 20 students (the control „slεep’ group) were required to be in bed by 11:00pm on the evening before the test. The test scores for each group are shown in the graphs below. Each dot on the graph represents a particular student’s score. For example, the 3 dots above the 40 in the top graph indicated that 3 students in the no-sleep group scored 40 on the test.

(Sample statistics: sleep: $\bar{x}$ = 60.6, s = 16.1, no-sleep: $\bar{x}$ = 55.7, s = 15.1. These results were not given to the students)



A. Examine the two graphs carefully. Then circle the conclusion from the 6 possible conclusions listed below the ONE you MOST agree with.

a. The no-sleep group did better because none of these students scored below 35 and the highest score was achieved by a student in this group.
b. The no-sleep group did better because its average appears to be a little higher than the average of the sleep group.
c. There is no difference between the two groups because there is considerable overlap in the scores of the two groups.
d. There is no difference between the two groups because the difference between their averages is small compared to the amount of variation in the scores.
e. The sleep group did better because more students in this group scored 75 and above.
f. The sleep group did better because its average appears to be a little higher than the average of the no-sleep group.

For the highest code of "2" students were required to make the judgement of a possible difference considering both the difference in means and the variation in the two groups (answer d). Students who considered the amount of overlap only (answer c) received a code of "1." Table 8.2.4.1 shows that only 55% of the students were able to make these judgements successfully. Frequencies of the answers, according to ability groups, are shown in Table 8.3.4.1. *Pacific A* had an item difficulty of 1.48 logits, 2.03 logits above the easiest item in the Rasch analysis.

Table 8.3.4.1

Answers to "Pacific A" according to ability group

Ability Group	A	B_1	B_2	C
Answer	(n = 3)	(n = 16)	(n =12)	(n = 2)
d	2	8	1	
c		4	1	2
e		2	1	
f	1	2	9	

8.3.5 Questions that deal with randomness – what is random, and why randomise?

The first of the questions that dealt with randomness (*Random*) asked the students to give an example of "something that happens in a random way" and to explain why their chosen example was random. Statistically, a random event is one where an individual outcome cannot be predicted but the long term pattern can be predicted. By the end of the unit students should have had an understanding of this. Therefore a code of "2" was given to students who gave a suitable example with an appropriate explanation and a code of "1" was given to students who used a suitable example but without an accompanying explanation. Frequencies of the answers, according to ability groups are in Table 8.3.5.1. This table shows that most students were able to give a suitable answer with an explanation, but as the level of ability decreased the students were more likely not to answer or to give an idiosyncratic explanation. This was one of the easier questions, with an item difficulty level of 0.80 logits above zero.

Table 8.3.5.1

Answers to "Random" by ability group

Ability Group	A	B_1	B_2	C
	(n = 3)	(n = 16)	(n =12)	(n = 2)
Coin/die/Tattslotto or similar with explanation	3	12	7	
Coin/die/Tattslotto or similar with no explanation		1	1	
No answer				1
Idiosyncratic*		3	4	1

* Idiosyncratic – indicates the answer did not relate to the question or was unintelligible

"Cereal D"

This is the last of the *Cereal* questions (see Section 8.3.3).

> You are working for a consumer organisation. As part of your duties, you select 49 boxes of "Get up and Go" cereal at *random* and weigh the boxes. On the label of the boxes you read that the minimum weight of the box is 800g.
> The standard deviation of the weight of the boxes is 14g. Therefore the standard error of the mean (the standard deviation of all possible sample means) is estimated to be 2g.
>
> (d) If you didn't use random sampling, how would this affect your previous answers in this question?

For the highest score of "2" the students were required to explain that for the statistical analyses to be valid, each possible sample had to be equally likely. Responses that indicated that by not using random sampling bias may be introduced in some way, or that the conclusions will be invalid, received a score of "1." Responses that indicated that the boxes from one place may have been from the same batch and could be similar also received a score of "1." All other answers received a score of "0." The answers, according to ability group, are found in Table 8.3.5.2. This item had a difficulty level of 1.61 logits above zero, and was in the middle range of difficulty for this questionnaire.

Table 8.3.5.2

Answers to “Cereal D” according to ability group

Ability Group	A	B_1	B_2	C
	(n = 3)	(n = 16)	(n =12)	(n = 2)
The sample could be biased and the results will be invalid	1	6		
Not every sample will be equally likely		1		
The results will be invalid, sampling must be random		2		1
The boxes manufactured at the same time could have similar errors		1	2	
The results could be biased/skewed			5	
The results won’t represent the population		1		
The results will not fit a Normal distribution		2		
No answer	2	1	1	
Idiosyncratic*		2	4	1

*Idiosyncratic – indicates that the answer did not relate to the questions or was unintelligible

“Pacific B”

Pacific B followed on from *Pacific A* (see Section 8.3.4).

> Forty students from the Pacific University participated in a study of the effect of sleep on test scores. Using random allocation, 20 of the students were required to stay up all night studying the night before the test (no-sleep group). The other 20 students (the control „sleep’ group) were required to be in bed by 11:00pm on the evening before the test.
>
> Atlantic University repeated the same study but allowed the students to *choose* which of the groups (sleep or no-sleep) they could go into. The Pacific University claims that allowing the students to choose could bias the results. Atlantic University claims that this does not matter.
>
> Which University do you think is correct? Give reasons for your answer.

The coding was similar to that of *Cereal D*, in that those who stated that there was not an equal chance of getting each allocation received the highest score of "2," and those who stated that bias could be introduced received a code of "1." Table 8.2.4.1 shows that no student received the highest available score, and that 48% of the students gave answers that indicated that bias would be introduced if random allocation did not take place. The answers, according to ability, are found in Table 8.3.5.3. This question had an item difficulty rating of 1.83, 0.35 logits above *Pacific A*.

Table 8.3.5.3

Answers to "Pacific B" according to ability group

Group	A	B_1	B_2	C
	(n = 3)	(n = 16)	(n =12)	(n = 2)
Letting students self-select will bias results and invalidate the statistical analysis	1			
Letting students self-select will bias results	1	9	6	
Pacific is correct – no further explanation	1	1		
Letting students self-select will mean the allocation is not random		2	2	
Since people prefer to sleep, there will not be enough students in the non-sleep group		1		
It is better to let students self-select		2	3	1
No answer		1	1	
Idiosyncratic*				1

*Idiosyncratic – answer does not address question

"Fred"

In this question "Fred" had a paper rejected because he had not randomly allocated his treatments. He then did so, and found that his allocations were just as before. The students were required to think what they would say to him to convince him that the random allocation was required. The reasoning is that the hypothesis tests, which look at the probabilities of sample outcomes if a hypothesis about the population is true, would be invalid if every allocation were not equally likely. The question of bias is also important; random allocation decreases the effects of unknown confounding factors that may be present. The instructions to the students were the following.

Fred is a plant geneticist and sent the results of his research to a scientific journal but his paper was rejected. He has sought help from a statistician. Here is their conversation:

Fred: I've just had my paper containing some important results rejected because I didn't use random allocation of my treatments. Now I have to repeat the whole experiment!
Statistician: Tell me what you did.
Fred: I had a bench with eight pots sitting next to each other along the bench. In the first four pots I put my new wonder species and in the next four pots I put the standard species. As I expected, my wonder species produced much higher growth.
Statistician: OK. Of course, there may have been some other factor varying along the bench which is responsible for the difference.
Fred: I'm not that stupid! The temperature, light and everything else is controlled in this glasshouse. If I thought there was another effect, I would have allocated the treatments to take account of the fact.
Statistician: In that case, our task is simple. We will produce an allocation plan by generating random numbers in the computer.
They do this, and find that the wonder variety is allocated to the first four pots, and the standard variety to the other four pots, just as before.
Fred: Great! You have just proved my results were valid because they were obtained under the layout recommended by random allocation.
Statistician: No! The editor rejected the WAY you obtained the layout, not the layout itself.
Fred: !!!!!

Can you explain to Fred why the randomisation was so important? See if you can provide an argument for randomisation that will overcome Fred's problem.

The coding was similar to the previous questions in this section, in that the highest scores were given to students who could state that each possible allocation should be equally likely. Table 8.2.4.1 indicates that 21% of the students received this score. Twenty-seven percent of the students gave answers that mentioned that unintended bias could be a problem if the treatments were not randomly allocated. These students, and those who stated that allocation should be random without further explanation received a score of "1." Table 8.3.5.4 shows the student answers according to ability. This question had an item difficulty rating of 1.35 logits above zero, which is 1.90 logits above the easiest question.

Table 8.3.5.4

Answers to "Fred" according to ability group

Ability Group	A	B_1	B_2	C
	(n = 3)	(n = 16)	(n =12)	(n = 2)
Each layout needs to be equally likely	2	2	1	1
Allocation should be random	1	1	3	
Random allocation makes the statistical calculations legitimate		2	1	
Random allocation removes bias		8	2	
It is only chance it happened that way				1
No answer		3	3	
Idiosyncratic*			2	

*Idiosyncratic –indicates the answer either does not answer the question or is unintelligible

8.3.6 Repeated questions from the first questionnaire

“Hospital”

Further details for the coding of the *Hospital* problem can be found in Section 7.3.2.

> Half of all newborns are girls and half are boys. Hospital A records an average of 50 births per day. Hospital B records an average of 10 births a day. On a particular day, which hospital is more likely to record 80% or more of female births?
>
> a. Hospital A (with 50 births a day)
> b. Hospital B (with 10 births a day)
> c. The two hospitals are equally likely to record such an event.
>
> Please explain your answer.

Table 7.3.2.4 demonstrates that 45% of the 75 students chose hospital B when answering the first questionnaire, whereas in the second questionnaire 61% of the 33 students chose hospital B (Table 8.2.4.1). Because the same students answered both questionnaires, the samples are not independent and a test for the difference in proportions was not carried out. In the first questionnaire 45% of the students chose the equally likely option, whereas in the second questionnaire 27% of the students did so. Further analysis of the *Hospital* questions indicates that out of the 33 students who answered both questionnaires, 7 received the highest score of “3” in both questionnaires. Eleven students changed their answers, with 10 receiving a higher score than for the first questionnaire, and one student receiving less than for the first questionnaire. The answers, according to ability, are found in Table 8.3.6.1.

Table 8.3.6.1

Answers to "Hospital" for the second questionnaire, by ability

Ability Group	A	B_1	B_2	C
	(n = 3)	(n = 16)	(n =12)	(n = 2)
Hospital B – More likely to deviate from average with smaller sample size/more likely to be close to average with larger sample size	1	5	6	
Hospital B – More likely to get 8 of 10 then 40 out of 50		6		2
Equal –Probability of 50% boy or girl for every birth	2	4	5	
Hospital A – More likely to get more girls with more births		1	1	

"Urn A and Urn B"

Further details of the coding for the items *Urn A* and *Urn B* can be found in Section 7.3.5.

> An urn has 2 white balls and 2 black balls in it. Two balls are drawn out without replacing the first ball.
>
> a. What is the probability that the second ball is white, given that the first ball was white? Please explain your answer
>
> b. What is the probability that the first ball was white, given that the second ball was white? Please explain your answer.

Table 7.3.5.3 indicates that in the first questionnaire 76% of the 75 students gave the correct answer of one in three for *Urn A*. For the second questionnaire, 70% of the 33 gave the correct answer for *Urn A*. Because the same students answered both questionnaires, the samples are not independent and a test for the difference in proportions was not carried out. Table 8.3.6.2 indicates that most students used the argument that since there were two black balls and one white ball left, the probability of getting a white ball on the second draw was one in three.

Table 8.3.6.2

Answers to "Urn A" for the second questionnaire, by ability group

Ability Group	A	B_1	B_2	C
Answer	(n = 3)	(n = 16)	(n = 12)	(n = 2)
1 in 3 – There are three balls left, one of which is white	2	14	5	2
1 in 3 – no explanation		2	1	
1 in 6 – calculated a joint probability	1			
Other numerical answer			2	
No answer			1	
Idiosyncratic*			3	

* Idiosyncratic – indicates the answer either does not answer the question or is unintelligible

The answers to *Urn B* by ability group are found in Table 8.3.6.3. This shows that 21% of the students gave the correct answer of one in three in the second questionnaire, compared to 12% in the first questionnaire (Table 7.3.5.4).

Table 8.3.6.3

Answers to "Urn B" for the second questionnaire, by ability group

Ability Group	A	B_1	B_2	C
Answer	(n = 3)	(n = 16)	(n = 12)	(n = 2)
1 in 3 – There are three balls in question, one of which is white	2			
1 in 3 – no explanation	1	3	1	
1 in 2– because at the beginning there were four balls, two of which were white		9	7	2
One in four – because at the beginning there were four balls		3	1	
Other numerical answer			1	
No answer		1	2	

"Med A" (called "Med" in the first questionnaire)

The instructions were.

A new medication is being tested to determine its effectiveness in the treatment of eczema, an inflammatory condition of the skin. Thirty patients with eczema were selected to participate in the study. The patients were randomly divided into two groups. Twenty patients in an experimental group received the medication, while ten patients in a control group received no medication. The results after two months are shown below.

	Experimental Group (Medication)	Control Group (No Medication)
Improved	8	2
No improvement	12	8

Based on this data, you think the medication was:

A. Somewhat effective B. Basically ineffective

<u>If you chose option A</u>, select the one explanation below that best describes your reasoning.	<u>If you chose option B</u>, select the one explanation below that best describes your reasoning.
a. 40% of the people (8/20) in the experimental group improved	a. In the control group, 2 people improved even without the medication.
b. 8 people improved in the experimental group while only 2 improved in the control group	b. In the experimental group, more people didn't get better than did (12 vs. 8).
c. In the experimental group, the number of people who improved is only 4 less than the number who didn't improve (12-8) while in the control group the difference is 6 (8-2).	c. the difference between the numbers who improved and didn't improve is about the same in each group (4 vs. 6).
d. 40% of patients in the experimental group improved (8/20), while only 20% improved in the control group (2/10)	d. In the experimental group, only 40% of the patients improved (8/20).

Table 7.2.4.1 indicates that 53% of the 75 students who completed the first questionnaire used proportional reasoning to answer this question. For the second

questionnaire, 58% of the 33 students used proportional reasoning. The answers, according to ability, are found in Table 8.3.6.4.

Table 8.3.6.4

Answers to "Med A" for the second questionnaire by ability group

Ability Group	A	B_1	B_2	C
Answer	(n = 3)	(n = 16)	(n = 12)	(n = 2)
Somewhat effective				
Option Ad	2	12	6	1
Option Aa		1		
Option Ac		1	1	
Basically ineffective				
Option Bd	1	1	2	
Option Bb				1
Option Bc		1	1	
No answer			2	

"Med B"

Med B was a new question that followed from *Med A*. It asked the students to consider the experimental design used in Med A.

> Listed below are several possible reasons one might question the results of the experiment described above. Please circle the letter for EVERY reason you agree with.
>
> a. It's not legitimate to compare the two groups because there are different numbers of patients in each group.
> b. The sample of 30 is too small to permit drawing conclusions.
> c. The patients should not have been randomly put into groups, because the most severe cases may have just by chance ended up in one of the groups
> d. I'm not given enough information about how doctors decided whether or not the patients improved. Doctors may have been biased in their judgements.
> e. I don't agree with any of these statements.

Students who answered “e,” received a score of “2.” Those who answered “d” received a code of “1.” Those who answered both “d” and “e” received a score of “0” as these answers are inconsistent with each other. Table 8.2.4.1 indicates that 18% of the students received a score of “1” or “2.” The frequency of the answers, according to ability groups are found in Table 8.3.6.5. *Med B* had the highest item difficulty rating on the Rasch analysis (2.61 logits above zero). This is in contrast to *Med A* with an item difficulty rating of 0.02 logits above zero. It can be concluded that the students found comparing the numbers in *Med A* simpler than making judgements on the experimental design.

Table 8.3.6.5

Answers to “Med B” by ability group

Ability Group	A	B_1	B_2	C
Answer	(n = 3)	(n = 16)	(n = 12)	(n = 2)
(e)	1	1		
(b) and (d)	2	6	4	
(a) and (b) and (c) and (d)			1	
(b) and (c) and (d)		3	1	
(a) and (c) and (d)			1	
(a) and (b) and (d)				1
(a) and (b)				1
(a) and (d)		2	2	
(b) and (d)				
(b) and (c)			2	
(a)		1		
(c)		1		
(d)		1	1	
No answer		1		

8.4 Relationships among ability measures and scores from formal assessments

A paired *t*-test was carried to see if there was a significant change in the students' ability measures from the first to second questionnaire. The results indicate that there was a significant gain in the mean ability between questionnaires one and two. The mean ability measure for the first questionnaire for these 33 students was 0.311, and the mean ability measure for the second questionnaire was 0.787, a significant difference ($P < .001$, see Appendix E3). To investigate differences among semesters, each student's ability measure from the first questionnaire was subtracted from the ability measure of the second. A single factor ANOVA was then carried out on these differences. The results indicate that there was no significant difference between the means of these differences among the semesters ($P = .133$, see Appendix E3). This indicates that although the students were from different cohorts (see Section 6.6) the overall performance on these questionnaires did not differ.

The ability measures from the first questionnaire were then correlated with the ability measures of the second questionnaire. This relationship is demonstrated in Figure 8.4.1. The correlation coefficient was .178, and the relationship was non-significant ($P = .374$, see Appendix E3).

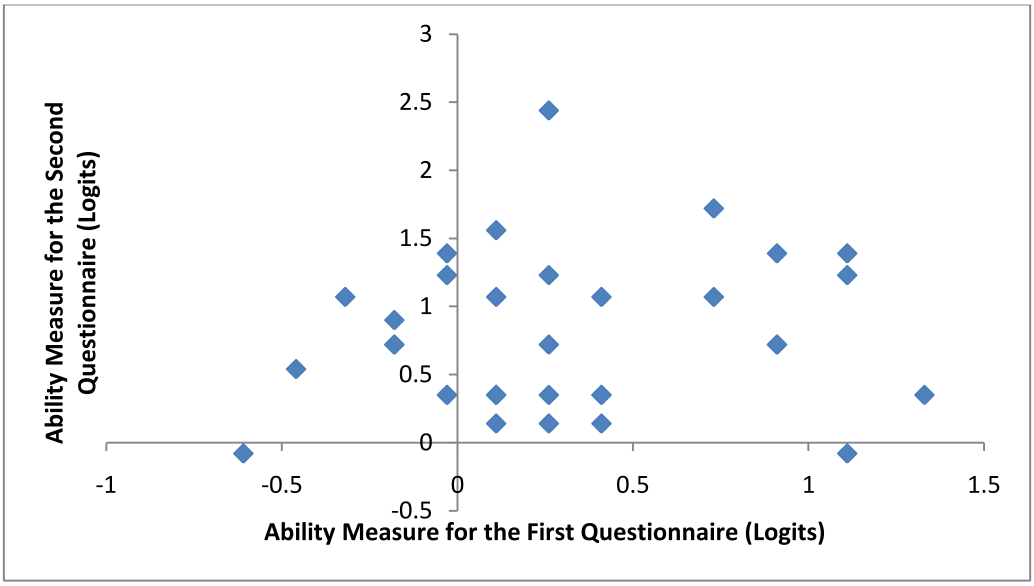


Figure 8.4.1. Scatterplot of the relationship of the students' ability measures between the first and second questionnaires.

The relationship between the students' ability measures from the first questionnaire and their scores from their formal assessment was then investigated, and is demonstrated in Figure 8.4.2. The score from the formal assessment was a result of a combination of the students' results from four projects and two in-class tests. Even though scores are not "measures" (see Section 4.2) it would be expected that if there was a relationship between the ability level and the score from the formal assessment there would be some trend in Figure 8.4.2; however, there was no such trend. The correlation coefficient was 0.017 and the relationship was non-significant (*P* = .935, see Appendix E3).

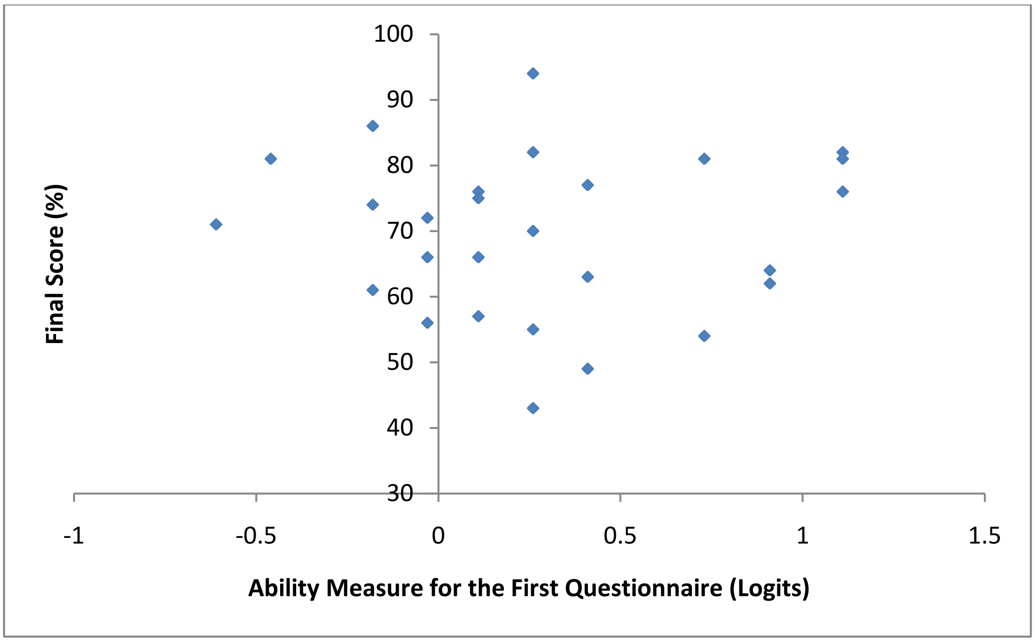


Figure 8.4.2. Relationship between the students' final scores of the formal assessment and the ability measures from the first questionnaire.

Similarly, the relationship between the students' ability measures from the second questionnaire and their scores from their formal assessment was also investigated, and is demonstrated in Figure 8.4.3. The correlation coefficient was 0.354, and the relationship was significant at the 10% level ($P = 0.070$). However, the point (2.44,94), indicated by the star, is influential. When this point was removed from the analysis the relationship was non-significant ($P = .387$), with a correlation coefficient of 0.177.

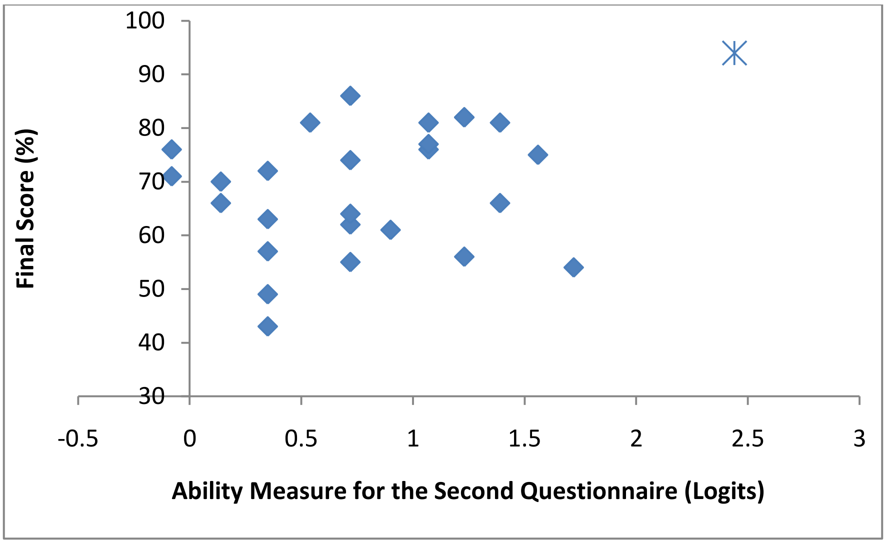


Figure 8.4.3. Relationship between the students' final scores from the formal assessment and the ability measures from the second questionnaire.

The lack of correlation between the ability score from the first questionnaire and the final score from the formal assessment indicates that this ability measure does not predict the outcome of an individual student. It is possible that students with poor prior understanding of statistical concepts (including probability) can make up for this lack of understanding as the semester progresses. The results of the ability measure from the second questionnaire also do not predict the final outcome of an individual student on formal assessment. The items in the second questionnaire tested for the ability to explain inferential processes in an informal manner, while in general, the formal assessments were based on formal statistical procedures.

The two questionnaires were also assessing different forms of reasoning, which possibly explains the lack of correlation between the ability scores from the analysis of these questionnaires. As described, the second questionnaire tested for informal reasoning used in inferential processes, whereas the first questionnaire examined students' prior knowledge of statistical concepts.

8.5 Summary and Discussion

The responses to the second questionnaire demonstrated five main problems in students' understanding when undertaking inferential processes. First, the Cereal questions showed that some students confused the terms "standard deviation" and "standard error." Some of the students used the standard deviation when the standard error should have been used, while other students stated they had used the standard deviation but had actually correctly used the standard error.

Second, most of the students did not use both the means and standard deviations when making a comparison between two data sets in the *Pacific A* question. They were more likely to use the standard deviations alone. An increase of 1.2 logits in ability was required for the students to make this upward transition in reasoning, a substantial gain in a questionnaire that encompassed a range of a little of 4 logits.

There was also a tendency for students not to use the given probability in making their conclusions, but to use other arguments. In the *Circuit Breaker* questions, the students preferred to say that one box did not give strong enough evidence to claim that the underlying rate of defectives had changed, because other boxes could have more or less defectives within them. Again an increase of 1.2 logits in ability was required for the students to make this upward transition in reasoning, a substantial gain in a questionnaire that encompassed a range of a little of 4 logits..

When answering the question that asked the students to explain the nature of a "significant difference," none of the students attempted to explain the meaning of the given *P*-value. Some of the other responses, however, did indicate that some of the students showed they understood the effect of sampling variation when making a judgement of difference between two samples. These students either stated that the means were not far apart when the standard deviations were taken into account, or stated that the given difference in means could occur just from sampling variation.

In addition, few students could explicitly state that for calculations based on probability, all possible combinations of treatment allocation or selection need to be equally likely. Most of the students, however, were aware that a lack of random selection and random allocation could introduce some sort of "bias" or somehow skew the results.

Finally, it was observed that there was no significant correlation between the rated ability of the students from the first questionnaire to the rated ability of the students from the second questionnaire; nor was there a significant correlation between these ability measures and the students' final score from their formal assessments. This suggests that two questionnaires and the formal assessments tested for dissimilar forms of statistical reasoning.

The next two chapters describe the development of students' understanding of *P*-values and confidence intervals as the intervention progressed. They also contain a description of the teaching strategies used for the semester of the second cycle of the intervention during which there was a focus on improving students' understanding of confidence intervals and *P*-values.

9. An Analysis of Students' Understanding of *P*-values.

9.1. Introduction

This chapter describes the progression of students' understanding of *P*-values as the intervention progressed. It describes the work through four semesters, the pre-intervention semester, and the three cycles of the intervention. In the first semester of this study (the second teaching semester of 2007) the unit was taught as in previous years. This was followed by the first and second cycles of the intervention (the first and second teaching semesters of 2008) and then by the third cycle of the intervention (in the first teaching semester of 2009). In addition, the teaching program of second cycle of the intervention is described. During this cycle notes were taken of the students' responses to certain problems, and their reactions to the different strategies. This was so that more knowledge about students' problems in understanding *P*-values could be gained to assist in the planning of the third cycle of the intervention.

A *P*-value answers the following question: If the proposition (hypothesis) about the population is true, how likely is the sample, or an even more unlikely sample? An understanding of the procedure for calculating a *P*-value requires consideration of what samples might be likely if the proposed population characteristic is true. For example, if the proposition is that the mean of a population is 200 g, then it would be expected that most of the sample means would be "nearby" to 200 g. If the sample mean were 205 g, and the researcher wanted to know if the population mean had increased, then the P-value would be determined by the calculation:

$$P(\bar{x} \geq 205 | \mu = 200)$$

What constitutes “nearby” depends on the standard deviation of the population, and therefore the standard error of the mean for each case.

Students’ understanding of P-values was assessed by examining the responses to the following questions in the test that was held during the last week of each semester, with the same wording each time.

1. A P-value of 0.98 indicates that the null hypothesis is almost certainly true. Is this statement correct? Give reasons for your answer.

2. In the test of a null hypothesis that a new drug produces the same expected benefit as the standard drug, versus the alternative hypothesis that the new drug produces a higher expected benefit than does the standard drug, a P-value of 0.01 is obtained. Explain what this result means to a patient who has read the result on the web but has no statistical training. Avoid all statistical jargon.

An ideal response to Question 1 would indicate that the P-value is the probability of the observed sample statistic or one more extreme if the null hypothesis were true. As a result this P-value would indicate that the observed sample statistic is *very likely* if the null hypothesis were true, but does not prove the null hypothesis true. Responses similar to this would receive a code of “3.” Other acceptable responses include that the true situation in the population could be close to that of the null hypothesis or that hypothesis tests only find evidence against the null hypotheses but cannot prove them true (code “2”). An acceptable response that would receive a lower score (“1”) is the general statement that nothing is ever proved in inferential statistics.

An ideal response to Question 2 would include a definition of the *P*-value in this context. If it were true that the new drug has the same expected benefit as the standard drug, then the probability of the results shown by the new drug would be only 1%. That is, the new drug’s results were unlikely if it did not work better than the standard drug. A response similar to this would receive a code of “3.” Students who performed a standard hypothesis test without further explanation received a code of “2,” whereas those who stated that the new drug “works better” without further explanation received a code of “1.”

9.2 Results of the pre-intervention semester (Semester 2 – 2007)

9.2.1 Teaching strategies

The pre-intervention semester was taught according to the previous practice where the students were first introduced to descriptive statistics, probability theory, and then the formal hypothesis testing procedure. With this format, the formal hypothesis testing procedure and its terminology were introduced in one lecture. As a result the terms “null and alternative hypothesis,” “level of significance,” “*P*-value,” “rejecting/accepting the null hypothesis” were all introduced at the same time as the probabilistic hypothetical reasoning procedure. Subsequent lectures consisted of describing hypothesis tests in various contexts.

Each week of the semester consisted of two traditional lectures where the students were given the information they were required to know with some explanation.

Each week the students were also required to attend one tutorial where they were given exercises to work on, for example writing null and alternative hypotheses for a given set of problems. They also attended one computer, “practical” session, where they were given the instructions to carry out the required statistical procedures in *Microsoft Excel*.

9.2.2 Student answers to the *P*-value items in the test

Out of the 13 students that agreed to take part in the study, 12 completed the test. In their answers to question one, no student attempted to explain the meaning of the *P*-value. Most of the answers indicated that the students were aware that inferential statistics and the process of hypothesis testing do not result in certainty but no further reasoning was given. Two students indicated the statement was true.

For the second question, all students apart from two gave an answer that contained a mention of the null and alternative hypotheses and the *P*-value, even though the instructions stated to avoid statistical jargon. Only four of these attempted an explanation of the meaning of the *P*-value, but none did so successfully. Their answers indicated that these students believed either that the *P*-value is the probability of being incorrect (also reported by Gliner, Leech & Morgan, 2002), that it indicates the rate of replication of the conclusion, or that the *P*-value gives the probability of seeing a difference.

In summary, the evidence indicated that most of the students were not confident enough to attempt to explain the meaning of the *P*-value, and those who did make this attempt held misconceptions as to its meaning. Most of the students, however, were aware that the practice of inferential statistics does not result in certainty.

9.3 Results of the first cycle of the intervention (Semester 1 – 2008)

9.3.1 Teaching strategies

In the first cycle of the intervention extensive guided discovery learning via computer simulation was introduced. For each idea introduced via simulation, for example the distribution of sample means, the material was not introduced into the lecture or tutorial until after the simulation had been carried out. In addition, the hypothetical probabilistic process via the use of the "Chinese Birth Problem" was introduced early in the semester. For this problem the students were asked to investigate the influence on the ratio of boys to girls in the Chinese population if the "One child policy" were replaced with a "Have children until a boy is born" policy. The students were each required to investigate 10 families by using a coin (details are in Appendix D). There were 26 students in this cycle of the intervention, 23 of whom completed the test.

9.3.2 Student answers to the *P*-value items in the test

This semester was the first where students attempted to explain the meaning of the *P*-value in their responses to Question 1. Five students attempted to explain the *P*-value, but only one was correct. Sixteen students indicated that the statement was false because accepting a null hypothesis only indicates there is insufficient evidence to reject it. For the second question, seven students attempted to explain the meaning of the *P*-value. Their answers indicated that the students believed either that the *P*-value gives the probability of a difference between the treatments, or the *P*-value gives the rate at which the two treatments will give the same result, or the *P*-value gives the probability that the null hypothesis is correct. One student, however, gave an answer that although not being entirely correct, would give the reader a partial understanding of the meaning of the *P*-value. This student said: “Since the likelihood of obtaining our test value is very unlikely assuming the same benefit as the standard drug, we conclude that the new drug gives a greater expected benefit than the standard drug.”

Of the other fourteen students, nine used the expressions *P*-value and/or null and alternative hypothesis without further explanation and five students stated that the new drug works “better” with no other explanation. The students who stated that the new drug works better without further explanation may have taken the instruction not to use statistical jargon to a greater extent than was intended by the writer of the test, and it is difficult to know if understanding was present or not. Those students who used the null and alternative hypothesis process without explanation may have done this because they did not have conceptual understanding and therefore resorted to procedural knowledge, but again, it is difficult to know if this was the case.

Of interest is the change in the explanations given for the answers to Question 1. For the pre-intervention semester, some of the students just stated that hypothesis tests do not result in certainty. In this semester, however, most of the students stated that accepting a null hypothesis test means that there is insufficient evidence to reject it. This shows a more sophisticated level of reasoning of the nature of hypothesis tests than just a general “statistics does not give proof” statement.

9.4 Results of the second cycle of the intervention (Semester 2 – 2008)

9.4.1 Teaching strategies

From the responses given at the end of the pre-intervention semester and at the end of the first cycle of the intervention it was apparent that the students were finding the hypothetical, probabilistic reasoning used in hypothesis tests difficult. Some of the students could explain that hypothesis testing does not result in definitive proof, or that accepting a null hypothesis just means that there is not enough evidence to reject it, but did not explain their reasoning any further.

It was felt by the researcher that a suitable example that would be easily understood was needed to introduce the hypothetical, probabilistic process. If a suitable example could be found, it was intended that it could act as a template for further hypothesis tests. After some searching, the following example, the "It is hot outside" problem, where a statement about the weather was assessed according to the clothes people were observed wearing, was found (Shaughnessy & Chance, 2005, see Section 6.4.4). The example was placed in a table as shown in Table 9.4.1.1.

Table 9.4.1.1

The first example of the hypothetical probabilistic process used in the second cycle of the intervention - The "It is hot outside" problem.

My hypothesis	It is hot outside today.
Data	When we look out of the window, everyone we see is wearing winter clothes (woolly hats, gloves and coats).
What is the probability of seeing people wearing winter clothes if it is hot outside?	Very, very low.
What do you conclude about my hypothesis?	It is incorrect.

This was followed by another example where the process was believed to be more complicated. It is based on a Sherlock Holmes story in which the theft of a race horse was being investigated ("Silver Blaze," also described in Shaughnessy & Chance, 2005). It is described in Table 9.4.1.2.

Table 9.4.1.2

An example of the probabilistic hypothetical process based on the Sherlock Holmes story, Silver Blaze.

The hypothesis	The horse was stolen by a stranger.
Data	The guard dog did not bark.
What is the probability of the guard dog not barking, if there was a stranger on the premises?	Very, very low.
What do you conclude about the hypothesis?	It is incorrect.
Conclusion	The horse was stolen by someone the dog knew – it was an inside job.

To introduce the use of numerical probabilities into this process, it was then repeated with the "Chinese Birth Problem," which had been used to introduce the process of simulation to the students in the previous week (Section 9.3.1). When the results of all the students were collated, it was found that the simulation predicted there would be 116 girls and 140 boys born. The resulting table, based on the examples in Tables 9.4.1.1 and 9.4.1.2, is shown in Table 9.4.1.3.

Table 9.4.1.3

The "Chinese Birth Problem," using the probabilistic hypothetical process based on table 9.4.1.1

My hypothesis	The ratio of girls to boys will remain unchanged.
Data	116 girls, 140 boys. (The results of the in-class simulation)
What is the probability of getting 116 girls or fewer out of 256 births if the hypothesis is correct?	0.08 (8%)
What do you conclude about my hypothesis?	Some agreed with it, and some did not. This led to questioning by the students – how low is "too low"? In answer, they were informed that how low is "too low" is determined by a convention, and this convention used a value of 5% (0.05).

This table format was then used as a basis for all other hypothesis tests used throughout the semester. For every hypothesis test, the students also were required to write out the meaning of the *P*-value for the specific context. For example, if the test was for the difference in two population means, they were required to write something similar to:

> The *P*-value is the probability of getting two sample means this far or further apart if the populations have equal means.

The students were encouraged to discuss their ideas with each other and then with the rest of the class. As they became more familiar with this process, time spent in interaction between the students and the lecturer increased, and the lectures became less formal. In this semester, seven students agreed to take part in the study, six of whom completed the test.

9.4.2 Student answers

In answer to Question 1, four of the students agreed with the proposition that a high *P*-value would indicate that the null hypothesis was very likely to be true.

One stated that there was not enough evidence to prove the statement true, and the other answer was incomprehensible. The answers indicated that the students either believed that the *P*-value gives the probability that the null hypothesis is correct or that the *P*-value gives the rate at which the null hypothesis will occur. For the second question, the students showed that they believed either that the *P*-value gives the rate at which the new treatment will work better, the *P*-value gives the probability of the observation (partly correct) or a low *P*-value indicates that the alternative hypothesis is true.

Some of the responses to Question 1 and 2 were inconsistent, suggesting that these students were confused about the nature of the *P*-value. Further evidence of this confusion was shown by internal inconsistencies within some answers. For example, one student stated that the new drug works better 1% of the time, but then stated that the new drug worked better.

This semester was disappointing for two reasons. One reason was the apparent lack of improvement demonstrated in understanding. The second reason was the low number of students who agreed to take part in the study making it difficult to draw firm conclusions. Further details of students' progression throughout this semester are found in Section 9.6.

9.5 Results of the third cycle of the intervention (Semester 1 – 2009)

9.5.1 Teaching strategies

This was the final semester of the intervention where data were collected for this research. The students were again encouraged to ask questions and to reflect on each problem. This time the students were also shown a visual representation of each hypothesis test, and encouraged to create one for themselves. It was found that as these changes were introduced the time students spent interacting with each other and the lecturer increased more than in the second cycle of the intervention, and the difference between tutorials and lectures became increasingly blurred.

For each hypothesis test, the students were encouraged to write out the meaning of each *P*-value, and to share their ideas with each other before sharing them with the rest of the class. It was also decided to introduce some of the formal terminology of hypothesis testing from the first week of the unit, so that this could become familiar to the students over time. The revised table for the "It is hot outside" problem is shown in Table 9.5.1.1.

Table 9.5.1.1

The new grid for the "It is hot outside"problem

The null hypothesis about the population	It is hot outside today.
Data	When we look out of the window, everyone we see is wearing winter clothes (woolly hats, gloves and coats).
P-value	What is the probability of seeing people wearing winter clothes if it is hot outside?
Decision about hypothesis	Reject/ Accept
Conclusion	It is not hot outside today.

In addition, a visual representation of each hypothesis test that was based on the appropriate distribution was introduced. To give a visual representation of each hypothesis test, each test statistic (for example the value of "*t*") was located on the sampling distribution and located on a sketch graph. For example, if the null hypothesis was that there was no difference in means, and the *t*-statistic was 3.6, a diagram such as shown in Figure 9.5.1.1 was drawn.

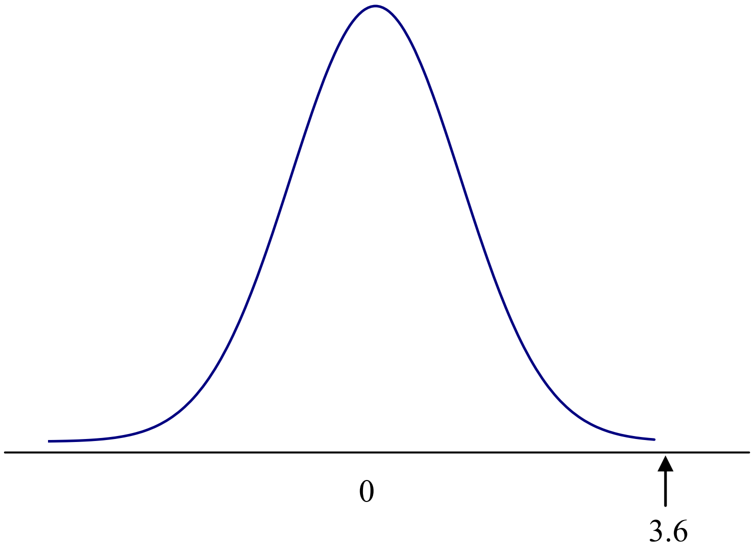


Figure 9.5.1.1. Diagram showing position of a proposed t-statistic.

Following the diagrammatic representation, questions similar to these were asked.

- Where do you think 3.6 will be (before adding the appropriate number and arrow)?
- If the population distribution is really centered on our proposed value, do you think our sample data belongs to this distribution, or another distribution?
- How likely do you think our sample data are (or values even further away), if H_o is true? (In verbal terms – likely/unlikely).

The responses were then compared to the numerical probability. This was calculated using the relevant probability distribution (for example, using the Binomial distribution), or by reading the *P*-value from the computer calculations.

A further change was made by introducing students to Popper's work on falsifiable propositions in science (see Section 2.3.1). This was introduced to help students to be able to put statistical hypotheses in the wider scientific context. Specifically, Popper's ideas were introduced to help students realise why the null hypothesis is the one of no difference, because it is easier to find evidence against an equality than to find evidence for a difference. If a greater understanding could be achieved, it was believed that it would be less likely that students would learn the convention for writing null hypotheses by rote and then make mistakes because understanding was not present.

9.5.2 Student answers to the *P*-value items in the test

Sixteen students agreed to take part in the study in this semester, twelve of whom completed the test. All of these students indicated that the proposition in Question 1 was false. Eight of the students not only attempted to explain the meaning of the *P*-value, but also did so correctly. One student defined the *P*-value as being the probability that the null hypothesis is true. The other students did not explain the meaning of the *P*-value in their answers, but merely stated that hypothesis tests do not give proof.

For Question 1, the biggest difference in results between this and the previous semesters was the proportion of those who correctly explained the meaning of the *P*-value. This difference is illustrated by the codes given to the students for this question, which are shown in Figure 9.5.2.1. This increase in score for this cycle of the intervention is confirmed by the results of the Kruskal-Wallis test that showed the difference in scores for the four semesters was significant with the third cycle of the intervention having the highest mean rank ($P < .001$). The details of this analysis are found in Appendix E4 and the mean ranks are displayed in Table 9.5.2.1.

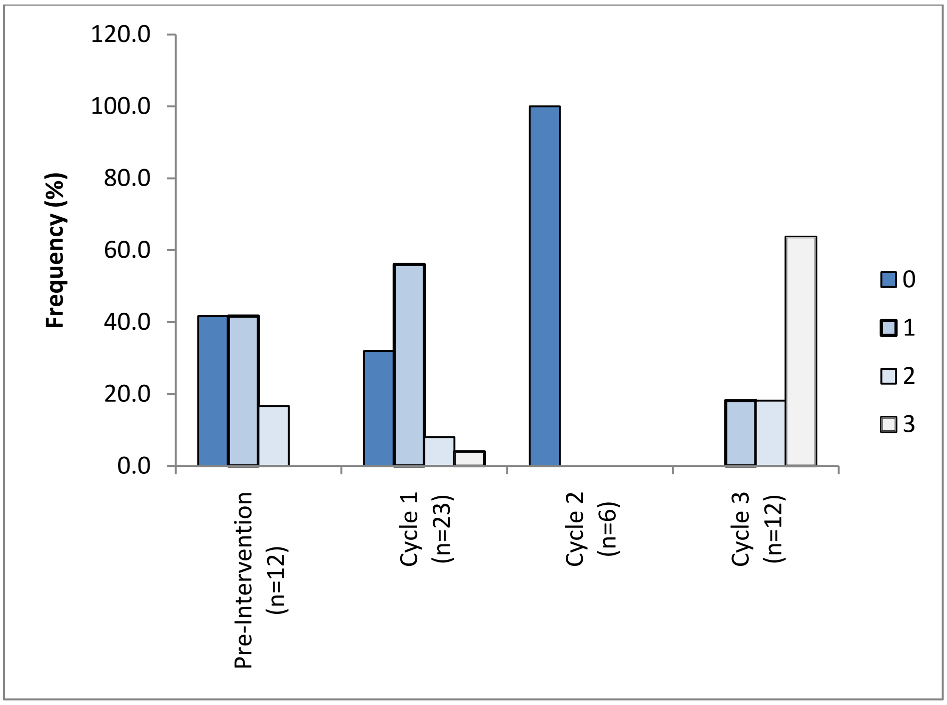


Figure 9.5.2.1. Percentage of students in each semester who received the indicated codes for Question 1 of the test.

For the Question 2, three of the students not only explained the meaning of the *P*-value in their answers, but also did this correctly. Six gave correct answers to the question itself, but did not explain the *P*-value. Two students used the null and alternative testing procedure without further explanation. The comparison between the scores from the four semesters is shown in Figure 9.5.2.2. In the first three semesters no students received the highest score of "3," whereas in the last semester no student received a score of "0." This increase in score for this cycle of the intervention is confirmed by the results of the Kruskal-Wallis test that showed the difference in scores for the four semesters was significant with the highest mean rank being for the third cycle of the intervention (*P* = .015). The details of this analysis are found in Appendix E4, and the mean ranks are shown in Table 9.5.2.1.

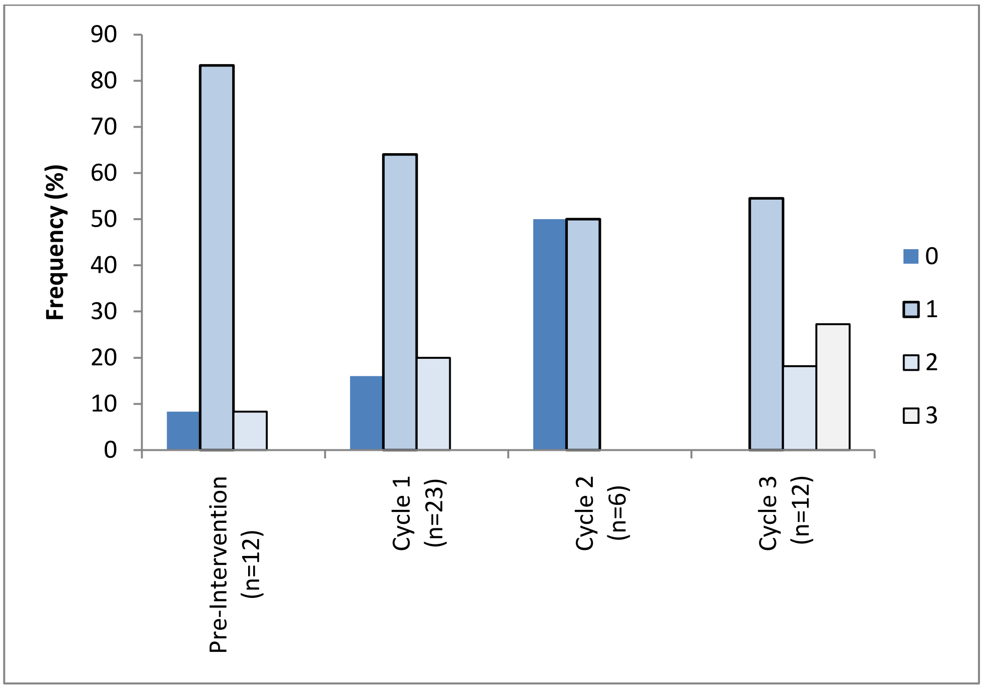


Figure 9.5.2.2. Percentage of students who received the indicated score for Question 2 of the test.

Table 9.5.2.1

Mean ranked scores for the P-value questions

	Mean Rank	
Semester	**Question 1**	**Question 2**
Pre-intervention	22.14	26.00
Cycle 1	26.04	26.86
Cycle 2	13.33	15.25
Cycle 3	42.54	36.33

9.6 A description of the teaching strategies used in the teaching of *P*-values in second cycle of the intervention

This section gives a description of the strategies used to teach *P*-values in the second semester of the intervention. Notes were kept throughout the semester of not only the students' answers but also their reactions to different strategies. The intention of keeping these notes was to gain further information into students' difficulties to assist in the planning of the third cycle of the intervention. Owing to ethical considerations the researcher could not know who was participating in the study, therefore no identifying information was kept, and the notes of the students' work were written in a descriptive format only.

9.6.1 Stage 1 – Introduction to probabilistic reasoning

The initial step was to introduce the students to simulation and to introduce the idea that the sample may not be exactly representative of a population via the use of the "Chinese birth problem" described in Section 9.4.6. The students were then asked to consider if a sample would have exactly 50% boys and 50% girls even if the population had this ratio. The students answered confidently that this would not be expected. Because null hypotheses are usually framed in terms of no difference, the following proposition was then posed: "My proposition is that the population ratio of boys to girls is unchanged. How far away from the one to one ratio would the sample have to be before you would think I was incorrect?" Some students suggested that they would disagree with the proposition if the proportion of girls was less than 85% of the boys, while others disagreed and stated that they would disagree with the proposition if the proportion of girls to boys was greater than 40:60/60:40.

The actual probabilities of getting these proportions were not calculated. The aim of this exercise had been to make sure that students were familiar with the principles of simulation, that they had become aware of sampling variation, and that they were aware that samples do not exactly reflect the population from which they were drawn.

The next step was to introduce numerical probability into the “Chinese Birth Problem.” At this stage the actual calculation procedures were not introduced, so that the students could think about the problem without being distracted by these procedures. The students had to use the probability of getting 116 girls or less out of 256 births, 8%, to make a judgement of the likelihood of this sample, assuming the ratio in the populating was still 1:1. They were asked to choose between the options of very “unlikely,” “unlikely,” “likely” and “very likely,” and then decide if the sample indicated that the population ratio had changed or not.

The students wrote down their answers and these were collected and were returned at the next session. It was apparent from the students’ answers that they found the reasoning difficult, and some stated this explicitly. This difficulty was reflected in inconsistency. For example, some students stated that the sample was unlikely but the population ratio would remain the same as before, and others stated that the sample was likely but the population ratio would change.

Students who were consistent in their reasoning were divided between those who stated that the sample was unlikely if the hypothesis were true and that the population ratio had changed, and those who stated that the sample was likely if the hypothesis were true so the population ratio had not changed. Some of the latter specifically stated that the observed ratio was not sufficiently different from the proposed population ratio to state that the population ratio would change. It was also apparent from their answers that some of the students did not appreciate how varied samples from a population could be, and that they expected a sample ratio to be closer to the population ratio than would be expected in reality.

Because it had become evident that this form of hypothetical reasoning was difficult for some of the students, an example was sought that could act as a model for their reasoning, an example that the students could connect to easily. After some searching the "It is hot outside" and the "Silver Blaze" examples were found and it was at this point they were first used (Section 9.4.1). The reasoning used for these examples was then repeated with the Chinese Birth Problem with the addition of the actual *P*-value. Table 9.6.1 shows the Chinese Birth Problem as it was presented to the students. The students did not agree with each other as to whether or not the given probability was too low to accept the hypothesis. They were told that the level of the "cut-off point" was a matter of convention, that it was 0.05 (5%) and that later on they would work on why this level might be changed to a higher or lower level.

Table 9.6.1

The Chinese Birth Problem with the associated P-value as given to the students

My hypothesis	The ratio of girls to boys will remain unchanged.
Data	116 girls, 140 boys.
What is the probability of getting 116 girls or less out of 256 births if the ratio if the hypothesis is correct?	0.08 (8%)
What do you conclude about my hypothesis?	Some agreed with it, and some did not. The students were told of the convention of 0.05.

9.6.2 Stage 2 – Consolidation – hypothetical probabilistic reasoning in another context

Before the students proceeded further with hypothesis testing, they were introduced to the mathematics of probability and the concept of a probability distribution. In this unit, the students were introduced to the Poisson, Binomial, Normal and Student's *t*- distributions. They were then given an example where they were asked to use the Poisson distribution to make a decision about the effect of temperature on the hatching rate of a species of butterfly. In this example, the number of butterfly eggs that hatched on a leaf after the temperature in the greenhouse had been raised by 3^o Celsius was compared to the average number that had hatched per leaf before the temperature change. With the given data, the *P*-value in this example was 0.04 (4%).

The student responses were collected and returned at the next session. These indicated that a small number of students were so unhappy that only one leaf was taken into account they would not discuss the problem any further. Some of the responses were inconsistent. These stated that the given outcome was very unlikely if the temperature change had made no difference, but then concluded that the temperature had not altered the hatching rate, or used the reverse reasoning.

Most of the students, however, argued that the outcome was so unlikely if the temperature had not made a difference that the hatching rate really had been altered by the change. Most of the students also explicitly used the probability in their answers.

9.6.3 Stage 3 – Simulation of *P*-values

After the students had been introduced to the hypothetical reasoning used in hypothesis tests they carried out a series of simulations in their practical sessions to further their understanding of *P*-values. One of the benefits of the *Excel* program is that the spreadsheet capability can be used to collect a large number of "samples" easily, and if these samples are based on the random number generator the process can be repeated quickly.

With these simulations, the *Excel* program was used to generate the samples that would be expected in each case if the null hypothesis were true, and, if the data were graphed, to give a visual representation of the variation among samples. The sample statistics that were generated from these random samples were then compared with the value of the appropriate test statistic.

The first such exercise was based on an idea from Ericksen (2006). In this situation the null hypothesis (that the population of voters was evenly divided for or against a proposition to enable city dwellers to keep koalas as pets) was simulated and compared to the result of a pre-election poll. The result of this pre-election poll was that out of the 50 voters surveyed; only 19 stated that they were in favour of the proposition. In the simulation, 500 samples (n = 50) were constructed. The proportion of samples that had 19 or less "yeses" (an estimate of the *P*-value) was then calculated. The students' simulations gave estimates of the *P*-value between .045 and .070. The calculated *P*-value, $P(X \leq 19|p = 0.5)$ is 0.059. The results of one such simulation are demonstrated in Figure 9.6.3.1. The columns representing the samples with 19 or less voters in favour of the proposition are left unshaded.

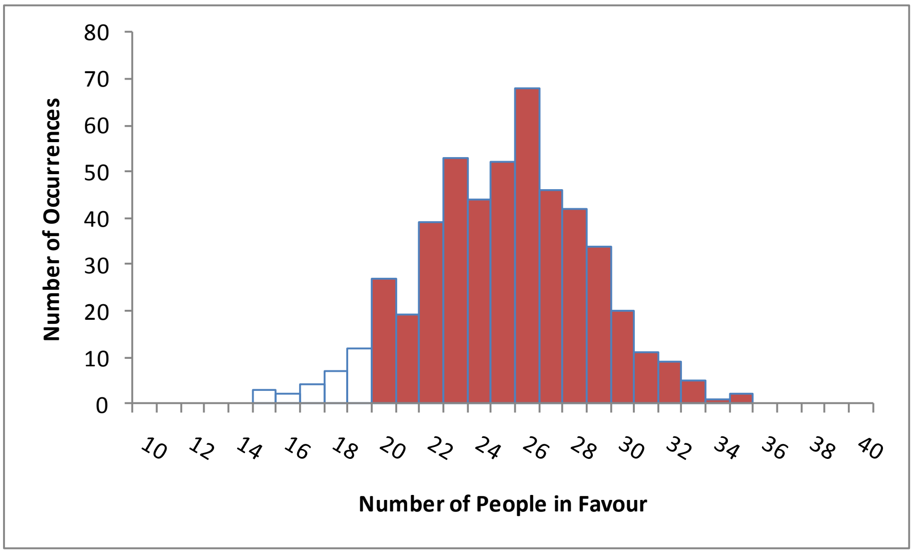


Figure 9.6.3.1. Results of a simulation showing the number of samples with 19 or less voters in favour of a proposition, assuming the population overall is evenly divided. The P-value from this simulation is estimated to be 0.062.

The next time the students met after the simulation, a comparison was made between the Koala exercise and the "It is hot outside" problem; this comparison is shown in Table 9.6.3.1.

Table 9.6.3.1

Comparison of the "Koala" and "It is hot outside" problems.

	It's hot outside	Koala
Hypothesis	It's hot outside.	The proportion of all the voters who agree with the proposition is 50%.
Data	Everyone we see is wearing winter clothes.	19 out of 50 people agree with the proposition.
Probability of our outcome, if hypothesis is true.	Very, very low.	From practical: 4.5%, 6.2%, 5.1% (Student results) Calculation in Excel: =binomdist(19,50,0.5,1) = 0.059
What we think about our hypothesis	It is not hot outside.	Two answers were given – one for and one against the hypothesis. The students were reminded of the convention of the "cut-off" of 0.05.

The next simulation was again based on the work of Erickson (2006). The data were obtained from the Census at School website[3] and consisted of a randomly drawn sample of the heights of 38 Grade 12 students. From these data an *Excel* spreadsheet was designed so that the difference between the mean of the first 16 Grade 12 students in the sample (all female) and the mean of the last 18 Grade 12 students (all male) was automatically calculated. The difference in mean heights was 11.62 cm. The students then used the spreadsheet to allocate the 34 students in the data set randomly between the two groups of the same size as before. Each time the difference in means was automatically recalculated. Each student performed the reallocation 20 times, and then placed his or her results on a tally on the whiteboard. Out of 360 results, a difference of 11.62 cm or more was observed only 3 times. This gave an approximate *P*-value of 0.01. An example of one such simulation is found in Figure 9.6.3.2.

[3] http://www.abs.gov.au/censusatschool

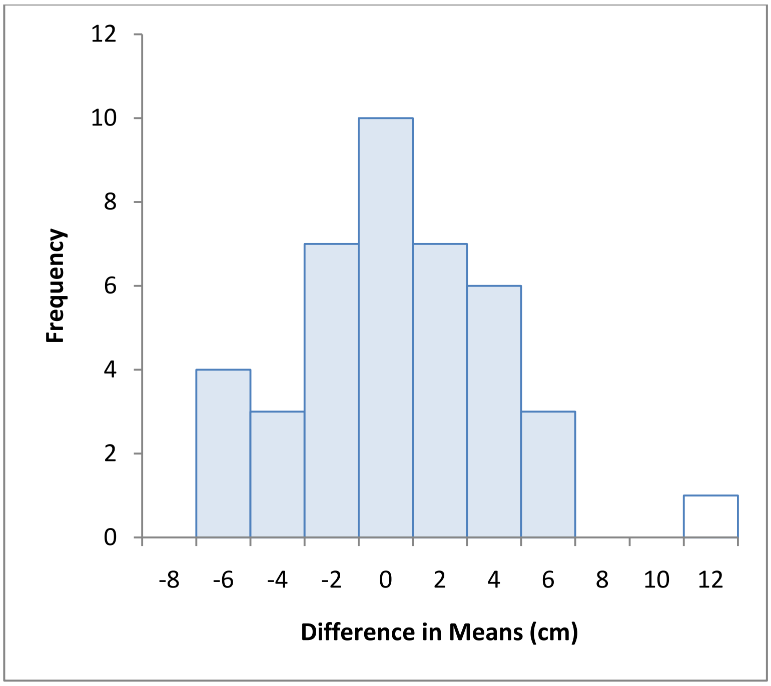


Table 9.6.3.2. Results of a simulation comparing the difference in means when the data are randomly allocated compared to the test statistic (unshaded bar).

9.6.4 Stage 4 – Introduction to the formal hypothesis testing procedure

In the lecture following the Grade 12 heights simulation students were introduced to the formal hypothesis testing procedure. The students worked on the handout shown in Figure 9.6.4.1 by filling in the gaps and answering the questions. It was at this point that the terms "*P*-value" and "null hypothesis" were first introduced. They were then introduced to the formal calculations associated with the 2-sample *t*-test for the difference in means. After this they were given the handout presented in Figure 9.6.4.2. This handout lists common problems in the understanding of *P*-values. It was intended that if known misconceptions were made explicit to the students, then these misconceptions could be avoided (Garfield & Ahlgren, 1988).

Our question about the **whole population**	Do adult males have a higher mean height than adult females?
The hypothesis – known as the “null hypothesis”. This is also about the population.	H_o: $\mu_M = \mu_F$ The null hypothesis is usually the one of NO DIFFERENCE. Statisticians put up a statement and then look for evidence against it. You will see that this usually gives a benchmark by which the appropriate probabilities can be calculated.
The data from our **sample**	
P-value	This is the probability of getting a sample with a difference of 11.62cm or more, given that the null hypothesis is true. That is, P(difference in means is 11.62cm or more|No difference). From our practical we estimate this to be ______. Formal calculations show the *P*-value to be 0.002.
Decision about the hypothesis	Choose one of these statements: A. This probability is so low that we will **reject** the null hypothesis. B. This probability is not low enough for us to say we have strong enough evidence against the null hypothesis to reject it, and we will **accept** the null hypothesis. Remember: Our cut-off point for going from B to A is 0.05 (5%).
Conclusion in plain English	

Figure 9.6.4.1. The handout used to introduce the formal hypothesis testing procedure for the difference in two population means.

The *P*-value tells you how likely it is to get the sample you got (or a more extreme sample) if the null hypothesis is true.

Many people are confused about the *P*-value and try to read too much into it. In an experiment, you get a certain set of results, like a sample mean. The hypothesis test asks whether random chance can account for those sample results if the null hypothesis is true.

The *P*-value is the likelihood, if H_o is actually true, that random chance could give you the results you got. It is a **conditional probability**:

P-value = P(this sample | H_o is true)

When you write it in symbols like that, you can see right away that the *P*-value is not the probability that the hypothesis is true or false:

* The *P*-value is not the probability that H_o is true.

* The *P*-value is not the probability that H_o is false.

* The *P*-value is not the probability that your results are due to random chance.

* The *P*-value is not the probability that your results are not due to random chance.

Once again, the P-value is just a measure of how likely your results would be if H_o is true and random chance the only factor in selecting the sample.

There's one other thing: the P-value is not a measure of the size or importance of an effect. A small P-value means you can be pretty confident in rejecting H_o. But it doesn't tell you by how much you're rejecting H_o (the effect size), or whether that rejection has any practical consequences.

(Reference: http://www.tc3.edu/instruct/sbrown/stat/pvalue.htm)

Figure 9.6.4.2. The handout given to explain the meaning of the P-value and to describe some common misconceptions.

The next time the class met was in a tutorial session. In this tutorial the students were given a hypothesis test to work on (The Fish Problem) where it was stated that in the previous year the population mean weight for the fish in a holding pond before release was 300 g. This year, a sample only was taken when the fish were at the same stage of development, and the sample mean was 280 g. Did this sample give evidence to suggest that the mean weight was less than the previous year? The students were given a handout, which is illustrated in Table 9.6.4.2, and were asked to fill in the gaps.

Table 9.6.4.2

The handout used for the Fish Problem.

Our question about the **whole population**	
The hypothesis – known as the "null hypothesis". This is also about the population.	H_o:
The data from our **sample**	
P-value	The *P*-value is 0.16. Write the meaning of this *P*-value is words, remembering it is a conditional probability.
Decision about the hypothesis	Choose one of these statements: A. This probability is so low that we will **reject** the null hypothesis. B. This probability is not low enough for us to say we have strong enough evidence against the null hypothesis to reject it, and we will **accept** the null hypothesis. Remember: Our cut-off point for going from B to A is 0.05 (5%).
Conclusion is plain English	

The students' comments while they were working on this problem indicated that most of the students were now comfortable with the hypothetical reasoning involved in hypothesis testing. These students used arguments suggesting that the sample mean was likely in a population with a mean of 300 g. Although explanations such as this do not give a formal explanation of the meaning of a *P*-value they do show evidence of understanding of the process of the hypothesis test. A minority stated that they still found the reasoning difficult – some stating that they did not "get it."

The students were then asked to give an example of an experiment they had worked on in their own disciplines. As it happened, most of the students were aquaculture students, and they had been working on fish nutrition. As a result, another fish problem was proposed. In this problem the mean weight of fish that had been given extra Vitamin C was compared to the mean weight of those who had been given the standard feed (the difference in means was reported as 0.7 kg). They were asked to write out the null hypothesis (H_o), what the P-value would be in words (it had a numerical value of 0.27), and their conclusions. The answers were collected and returned in the next session.

Most of the students showed an understanding that was very close to being correct. These students stated that the *P*-value was the chance of seeing a difference of 0.7 kg in the means if H_o was true. Some students had an understanding that was close to the formal definition of the *P*-value, stating that the *P*-value was the probability of seeing a difference of 0.7 kg. A smaller number of students gave evidence of misconceptions of the meaning of the *P*-value, for example, that the *P*-value was the probability that H_o was correct.

Overall, the students were starting to gain an understanding of formal hypothesis testing. It was apparent, however, that a small number of students still had problems in understanding the meaning of the *P*-value. For some of these with problems the consequence was that they drew the incorrect conclusion for the hypothesis test. Instead of accepting the null hypothesis, it was rejected and the conclusion was made that the extra level of Vitamin C had increased the mean weight of the fish.

In later sessions the students were given more examples of hypothesis tests that they worked on together in groups. For each hypothesis test they were asked to explain the meaning of the *P*-value in context, and draw a conclusion from the value of this *P*-value, and then share their conclusions with the rest of the students. During the discussions it was apparent that a small minority of students were still struggling with the process of the hypothetical reasoning based on the *P*-value.

9.6.5 Stage 5 – *P*-values in other contexts – chi-squared tests for independence, the analysis of variance, and linear regression

Toward the end of the semester the students were introduced to some applications of hypothesis testing, chi-squared tests for independence, the analysis of variance (ANOVA) and simple and multiple linear regression. The first application to be introduced was the chi-squared test for independence. It was introduced with the means of a simulation in a practical session, which was based on work taken from Burrill (2002). The aim was for the students to compare the rate of children suffering from Haemolytic Uraemic Syndrome (HUS) who had been given a certain antibiotic to the rate for those who had not. In the sample, five children who had been given antibiotics suffered from HUS. First of all, the students had to calculate the expected number of children in the sample who had been given antibiotics who would suffer from HUS if the disease occurred at random; this number was 1.7. The simulation then randomly allocated the children to the groups who had been given antibiotics and those who had not, and the number of children with HUS who had been given antibiotics was recorded. Because the exercise was based on the random number generator in *Excel*, the simulation was easily repeated. Each student repeated the simulation 20 times and the class data were tallied. The results of this simulation are found in Table 9.6.5.1.

Table 9.6.5.1

Results of the simulation with the expected numbers of children affected with HUS if the disease occurs at random

Number of children with HUS	Count
0	50
1	63
2	46
3	15
4	16
5 (Observed value)	0
Total	190

Because none of the 190 simulations predicted that 5 children would be affected, the *P*-value was estimated to be less than 0.01 (1/190). The conclusion was that treatment with antibiotics did increase the chance of a child contracting HUS. Figure 9.6.5.1 shows the bar graph of the data in Table 9.6.5.1. The test statistic, the number of children who contracted HUS who received antibiotics, is designated with the solid fill. This figure gives a visual illustration of the comparison of the test statistic with the predicted values if the disease occurred at random. It suggests that the administration of the antibiotic to children did increase the chances of the children to suffer from HUS.

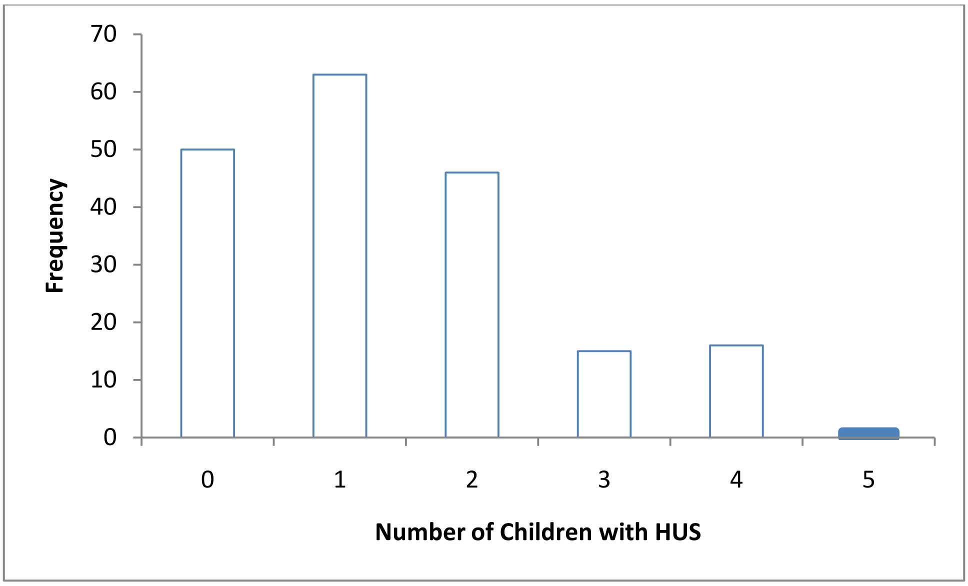


Figure 9.6.5.1. Results of a simulation comparing the actual number of times five children were observed with HUS compared to the number if the disease occurred at random.

One of the applications of hypothesis testing with which the students were expected to become familiar with was the single factor Analysis of Variance (ANOVA). This process was introduced in an informal way. The students were given data that consisted of the time to answer calls for an airline reservation system for five different shifts during a week. The students were asked to plot the data (Figure 9.6.5.2) and to make a judgement as to which groups were significantly different from each other based on an estimate of the mean and spread of the data in each group.

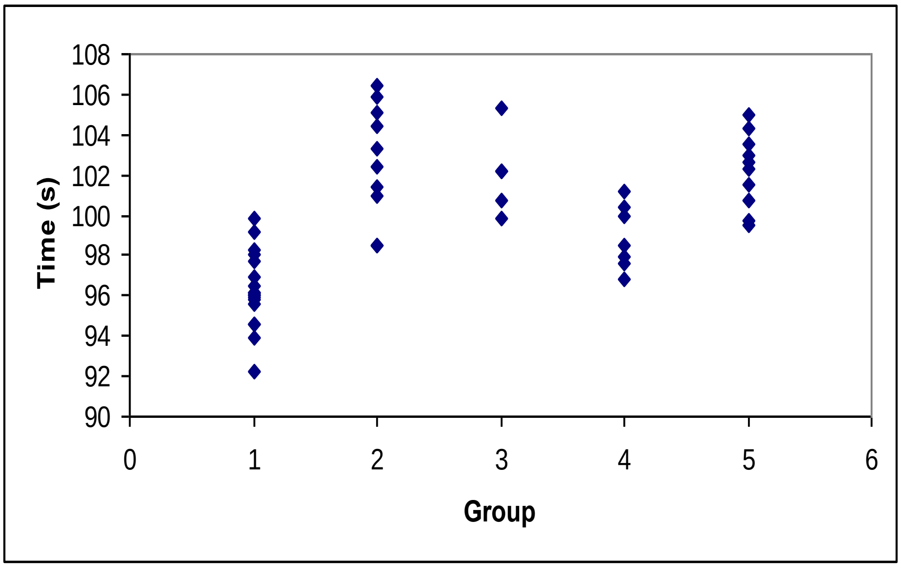


Figure 9.6.5.2. Times to answer calls for an airline reservation system for five shifts.

At the next lecture the students' predictions were written on the whiteboard before the formal procedure for ANOVA was introduced. All the students agreed that groups 1 and 2, 1 and 3, and 1 and 5 had means that were significantly different from each other. They did not agree on whether or not groups 2 and 4, and 2 and 5 had means that were significantly different from each other, and they all agreed that the other combinations of groups did not have means that were significantly different. The students were then asked to explain the meaning of the *P*-value in this context, that is, the probability of the observed differences, or those more extreme, if all the population means were equal. The formal procedures indicated that significant differences were found between the means of groups 1and 2, 1 and 3, 1 and 5, 2 and 4, and 2 and 5. This corresponded closely with the students' predictions.

The final of the simulations the students carried out in a practical session involved observing what happens to the estimation of the equation of the line of best fit when measurements are taken. The process of measurement always results in error, therefore the equation derived from measurement data will be only an estimation of the situation for the entire population. As for any sample data, it is expected that variation will exist from sample to sample.

The simulation asked the students to pretend that they were the “God of Algebra” so that they know what the true equation was for their data, in this case $y = 3x + 2$. A random component was then added to each “y” value. The original data and the data with the random component were graphed and their equations compared. For any data in *Excel* that are based on the random number generator the random numbers are easily recalculated and so the simulation can be quickly repeated. One example is shown in Figure 9.6.5.3. The complete line is that for the “true” situation, $y = 3x + 2$, and the dotted line is that for the data with the random component.

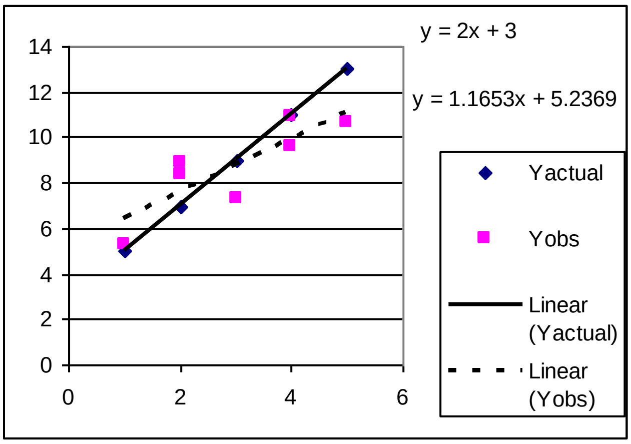


Figure 9.6.5.3. Comparison of the line for the “true” situation, and the line of best fit for the “sample” data.

A similar process was then carried out where the true situation was one where there was no influence of the independent variable on the dependent variable; in this case the true situation was given the equation $y = 3$. A random component was added as before and the two lines were drawn again, and their equations compared. One example of this process is shown in Figure 9.6.5.4. This figure illustrates how a person may be misled into believing that there is a significant relationship between two variables when in fact, this is not the case. The *P*-value in this context is the probability of the sample gradient or greater if the population gradient is zero (that is, there is not a significant relationship between the two variables).

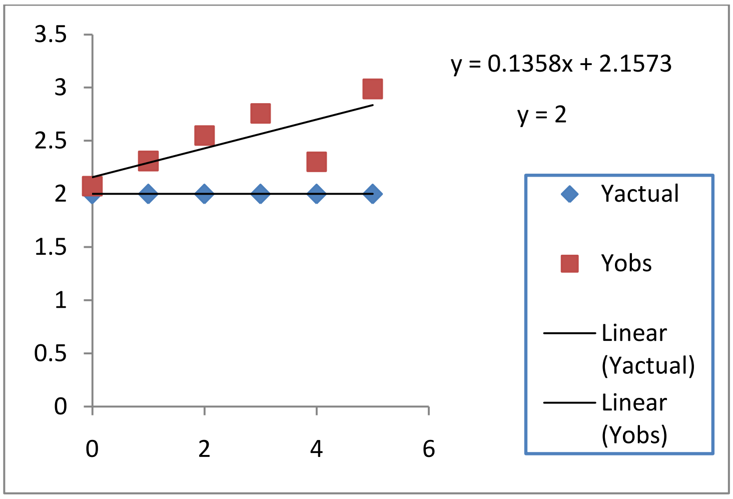


Figure 9.6.5.4. Comparison of the line for the "true" situation where there is no relationship between the dependent and independent variable, and the line of best fit for the "sample" data.

9.6.6 Stage 6 - revision

A week before the final lecture, the students were given a practice test and were asked to bring their answers to this final lecture. The question in this practice test that related to *P*-values was the following.

> A *P*-value of 0.9 means that it is almost certain that the null hypothesis is true.
>
> True or false? Give reasons for your answer.

As the students shared their answers it became apparent that some students were convinced that the statement was true. They were asked “What is a *P*-value?” As the discussion progressed it also came apparent that although some students were confident in correctly describing a *P*-value, they still thought that the high *P*-value would mean the null hypothesis would be *likely* to be true. An example was given to them where the “true” value of the parameter was very close to that proposed in the null hypothesis, and they were asked to consider what the *P*-values might be like for sample statistics close to the null hypothesis. The *P*-values would be high, even though the hypothesis did not identify the correct parameter. The semester ended with a test that contributed to the formal assessment, part of which was used for this study. A discussion of the student answers to the selected questions from this test is found in Section 9.4.2.

9.7 Summary of the misconceptions identified during the study in comparison to the literature

The following summarises the misconceptions shown by the students in their understanding of *P*-values during the four semesters of the study. These misconceptions were demonstrated in the student responses to the test items that were part of the second test used in the formal assessments. The study showed that the following misconceptions about *P*-values were held by some of the students.

First, some of the students stated that if the *P*-value is below 0.05, then the null hypothesis is accepted. Those students who had this view would appear not to have understood the meaning of *P*-value in any way, but have merely attempted to learn a rule which they then misapplied. This tendency has also been observed by Chance, delMas and Garfield (2004). If the student had an idea of the true nature of the *P*-value, it would seem unlikely that this mistake should be made.

Second, some of the students stated that the *P*-value is the probability that the null hypothesis is true. This misconception has also been observed by Gliner, Leech and Morgan (2002). This is a simpler interpretation than the correct definition of the *P*-value, and would appear to make sense to the students. Therefore, not only was it the most commonly held misconception about *P*-values, but also it was tenaciously held. Telling the students that this was not the meaning of the *P*-value did not appear to make any difference.

Other misconceptions included that the *P*-value is the probability of being incorrect (also observed by Gliner, Leech and Morgan, 2002), that the *P*-value indicates the rate of replication of the conclusion, that the *P*-value gives the probability of seeing a difference and the *P*-value gives the rate at which a specific treatment will show a difference. In addition, there was one other misconception that was partly correct, that the *P*-value gives the probability of the observation. This, too is a simpler interpretation than the correct definition of the *P*-value, and also appears to make sense to the students.

9.8 How students' understanding changed over the intervention

At the end of the pre-intervention semester no students attempted to explain the meaning of a *P*-value in their responses to the test questions, even though this could have improved the quality of their explanations. As the intervention progressed, the proportion of students who used the meaning of the *P*-value to explain their answers, increased. The largest improvement was noted at the end of the third intervention.

The third intervention used a combination of strategies, the introduction of hypothetical, probabilistic process early on in the semester, simulations, students writing down their reasoning, students comparing their work with others, alternative representations, and relating the process of hypothesis tests to the general scientific method as described by Popper (1963).

Some of these strategies enabled the students to make connections between the unfamiliar and the familiar. It was apparent from the students' responses at the time of their entry into the statistics unit that in general, they had not been exposed to hypothetical, probabilistic reasoning. The introduction of the "It is hot outside" problem, however, enabled students to understand the process in a simple context. Once this was completed, they were able to connect more difficult problems to the initial problem. The representation of the hypothesis tests in visual form enabled students to make connections between the probability distribution, what sample statistics would be likely given the population parameter, the likelihood of the test statistic given the parameter, and the numerical result for the *P*-value. Some of the simulations also led the students to have a visual representation of what sample statistics are likely given a population parameter, and they were able to compare their test statistics with these "likely" sample statistics.

Other simulations, for example the one used to introduce the Central Limit Theorem (see Section 6.4.3), allowed the students to "discover" some statistical principles for themselves. The element of surprise that resulted from the predict, test, re-evaluate format appeared to increase students' interest. Even the simple comparison between means and medians, where a data point was replaced by a large number, seemed to startle the students, and some of them appeared to enjoy putting in the largest number they could think of to see how much they could make the mean change, while the median remained constant. These simulations are all possible because of a spreadsheet's ability to instantaneously update the calculations and to "resample."

The connection of the writing of null hypotheses with Popper's ideas on falsification was of interest to the students. It was apparent that some of them had not been exposed to ideas of what makes a scientific proof, and that repeated observations in confirmation of a theory do not give definitive proof. This process enabled some students to make a more intelligent decision about the null hypothesis, instead of just learning that the null hypothesis is the one of "no difference."

It is not contended that one strategy alone led to the increase in understanding shown by the students. It is believed that it was the combination of strategies that led to the observed improvement. This combination helped to gain the students' interest and enabled them to make connections between the unfamiliar to the familiar, and connections between visual and written representations.

9.9 Implications for teaching

This study confirms the proposition that students find the reasoning used in inferential statistics unfamiliar and difficult (Garfield & Ahlgren, 1988; Yilmaz, 1996). As a result, students often make the error that a *P*-value is the probability that the null hypothesis is true, as this is a simpler concept to understand.

One of the implications of constructivist theories of learning is that instructors need to be aware of the misconceptions that students are likely to have or to develop and take steps to avoid these. This study demonstrated, however, that merely pointing out the misconceptions does not necessarily mean that the students will avoid them or correct them. If students have an idea that makes sense to them, it will persist. Students will not change a conception unless they see some benefit (Posner et al., 1982), and so the instructor needs to take steps that will encourage the students to make sense of the conception the instructor wishes the students to have.

As a result of this study, the concept of a *P*-value is now taught with the following combination of strategies in the Data and Handling Statistics unit. First, the *P*-value is introduced with a simple example (the “It is hot outside problem”) and the terms “*P*-value” and “null hypothesis” are introduced in the first week of semester so that the students have more time to become familiar with the terminology.

Second, for each hypothesis the students are encouraged to write out the meaning of the *P*-value in that particular context and the students are encouraged to help each other in this process. The students are not rushed and individual questions are answered. In addition, diagrams are used as well as verbal explanations. In these diagrams the test statistic is placed on a diagram of the hypothesised distribution and questions are asked to assist the students to make a link between the likelihood of the value of the test statistic, and the numerical probability of the *P*-value.

By these means the students are encouraged to develop connections between the unfamiliar hypothetical, probabilistic reasoning used in hypothesis testing and something familiar, so that they can find a means of developing a concept of the *P*-value that is personally meaningful.

The next chapter describes in a similar fashion the students' understanding of confidence intervals over the period of the study.

10. An Analysis of Students' Understanding of Confidence Intervals

10.1 Introduction

This chapter describes the progression of students' understanding of confidence intervals as the intervention progressed. It describes the work through four semesters, the pre-intervention semester, and the three cycles of the intervention. In the first semester of this study (the second teaching semester of 2007) the unit was taught as in previous years. This was followed by the first and second cycles of the intervention (the first and second teaching semesters of 2008) and then by the third cycle of the intervention (in the first teaching semester of 2009). In addition, the teaching program of second cycle of the intervention is described. During this semester notes were taken of the students' responses to certain problems and their reactions to the different teaching strategies. This was so that more knowledge about students' problems in understanding confidence intervals could be gained to assist in the planning of the third cycle of the intervention.

At the end of this introductory statistics unit, students were expected to be able to perform simple hypothesis tests, and be able to estimate population means from given samples. A population mean can be estimated with the value of the sample mean (a "point" estimate). Alternatively, a range of values can be calculated in which it is believed the value of the population mean is likely to be. This process, finding the 95% "confidence interval for the mean," takes into account the standard error (and hence the precision) of the estimate. The calculations for confidence intervals are relatively simple; all that are needed is the value of the sample mean, the value of the standard deviation of the sample (as the population standard deviation is not usually known), and the "*t*" value for the appropriate level of confidence. In this unit, the practice is to use approximate 95% confidence intervals and therefore the value of "*t*" is "2." Using this value the formulae to determine the limits of the 95% confidence interval are:

$$\text{Lower limit } = \bar{x} - 2 * \frac{s}{\sqrt{n}}$$

$$\text{Upper limit } = \bar{x} + 2 * \frac{s}{\sqrt{n}}$$

where $\bar{x}$ refers to the sample mean, s refers to the sample standard deviation, and n refers to the size of the sample. It is then likely that the value of the population mean is in the range from the lower to upper limit. The likelihood that this has occurred is indicated by the level of confidence. The "95%," therefore, indicates that a process has been used that will give a range that includes the actual value of the population mean 95% of the time the process is used. For any single interval, however, it cannot be known if the value of the population mean is included or not.

It appears that students find the reasoning behind the use of these formulae difficult to understand. To understand the process the following knowledge needs to be brought together.

1. Ninety-five percent of individuals in a normally distributed population are within approximately two standard deviations of the population mean.
2. If the sample size is large enough, sample means form a Normal distribution (The Central Limit Theorem). This distribution is centred on the

population mean and has the standard error as its standard deviation. The standard error ($\sigma_{\bar{x}}$) is linked to the population standard deviation (σ) by the formula $\sigma_{\bar{x}} = \sigma / \sqrt{n}$.

3. As a result of (1) and (2), 95% of the possible sample means with a given sample size are within two standard errors of the population mean.
4. Since a sample mean will be within two standard errors of the population mean 95% of the time, the interval calculated by the formulae above will include the value of the population mean 95% of the time.

From the student answers, it appeared that they found steps one to three to be fairly straightforward, but had difficulty in grasping the reasoning from step three to step four.

To assess students' understanding of confidence intervals the students were required to answer the following questions in a test held in the final week of each semester, with the same wording used each time.

> The 95% confidence interval for the expected number of visits by Tasmanians to a doctor during 1998 is 7 to 11.
>
> a. In completely non technical words, explain what this statement means.
> b. What does the 95% refer to?

An ideal response to part (a) would state that the mean number (or expected number) of visits by Tasmanians to a doctor in 1998 was estimated to be between 7 and 11. This response would result in a code of "2." A response that stated the above, but added "95% of the time," or stated that on average, Tasmanians visited a doctor between 7 and 11 times would receive a code of "1."

An ideal response to part (b) would state that the given interval has a 95% chance of including the value of the population mean. Alternatively, it could be stated that the process used would include the value of the population mean 95% of the time it was used. These responses would receive a code of "2." A response that stated that 95% of sample means will be within two standard errors of the population mean, but did not explain how these related to the confidence interval would receive a code of "1."

10.2 Results of the pre-intervention semester (Semester 2 – 2007)

10.2.1 Teaching strategies

The pre-intervention semester was taught according to previous practice, where the students were introduced to the Central Limit Theorem, and then to confidence intervals in a lecture. Each week of the semester consisted of two traditional lectures where the students were given, with some explanation, the material they were required to know. The students were also required to attend one tutorial per week where they were given exercises to work on, for example calculating the confidence interval from given sample statistics. The students also attended one computer, "practical" session per week, where they were given the instructions to carry out the required statistical procedures in *Microsoft Excel*. There were no computer exercises that related to confidence intervals, except how to calculate statistics such as the mean and standard deviation in *Excel*. There were 14 students who provided data for this part of the study, 12 of whom completed the second test.

10.2.2 Student answers to the confidence interval questions

In their answers to part (a) of the question, one quarter of students answered correctly. The other students stated either that between seven and eleven Tasmanians visited a doctor, that 95% of Tasmanians went to the doctor 7 to 11 times, or that the average falls between 7 and 11, 95% of the time.

For part (b), one quarter of the students answered correctly, but apart from one student, these were not the students who answered part (a) correctly. The other students either stated that 95% refers to the number of Tasmanians visiting a doctor, 95% refers to the proportion of means that fall within 7 and 11, or that the mean for 95% of the population is between 7 and 11.

10.3 Results of the first cycle of the intervention (Semester 1 – 2008)

10.3.1 Teaching strategies

In the first cycle of the intervention guided discovery learning via computer simulation was introduced. For each idea introduced via simulation, the appropriate material was not introduced into the lecture or tutorial until after the simulation had been carried out. The first of the simulations that applied to confidence intervals introduced the Central Limit Theorem. Thompson, Liu and Saldanha (2007) have reported that it is common for students to think that the distribution of sample means will have the same distribution of the parent population. This simulation was designed to encourage students to gain this belief but then be confronted with conflicting evidence. In this simulation, students were given some data that were normally distributed. They were then informed that they would be taking samples from these data and calculating the sample means. They were asked to predict the shape of the distribution that these sample means would have. They then used *Microsoft Excel* to take 500 samples, calculated the mean of these samples, and draw a histogram of these means. These means formed a Normal distribution. They were then asked to undertake a similar process for data that were uniformly distributed. Because they had been led to believe that the sample means would have the same distribution as the population distribution, most of the students predicted that the sample means would have a Uniform distribution as well, and seemed to be surprised when a Normal distribution resulted. They then repeated the process with a Binomial distribution, one with a small sample size (n = 5) and then with a larger sample size (n = 25).

In the tutorial that followed the practical, the students were given a handout to encourage them to consider the consequences that result from sample means being normally distributed (see Figures 10.6.2.1 and 10.6.2.2). For example, approximately 67% of sample means will be within one standard error of the population mean. In the next practical class, the students carried out a simulation with the aim that they would see the proportion of intervals that would include the value of the population mean. For this simulation, students constructed 100 random samples drawn from a population with a given mean. They calculated the mean, standard deviation and standard error for each sample, and then calculated the interval constructed by adding and subtracting one, two and then three standard errors from the sample mean. Using the "IF" function in *Excel* they determined the number of intervals out of the 100 that had the value of the population mean within them for each number of standard errors. It was only after these simulations were carried out that the students were given a formal lecture on confidence intervals.

10.3.2 Student answers to the confidence interval questions

For this semester 23 students participated in the study. Approximately one third of these gave a correct answer to part (a). For example, "There is a 95% chance that the mean number of visits is between 7 and 11." The other students indicated they thought that 95% of Tasmanians visited a doctor between 7 and 11 times, or that the mean for 95% of the population is between 7 and 11, or that 95% of the sample means fall within 7 and 11.

For part (b), approximately one quarter of the students gave a correct answer, however only three students answered both parts correctly. The misconceptions shown were different from those in 2007, and indicated an improvement in understanding. The students in this semester stated that 95% of sample means fall within two standard deviations of population mean, or that 95% of data falls between 7 and 11 or that 95% refers to the probability that the number of visits is between 7 and 11.

It is apparent from the answers to parts (a) and (b) that several students believed the confidence interval to be about the number of visits to the doctor, rather than an estimate of the value of the population mean.

10.4 Results of the second cycle of the intervention (Semester 2 – 2008)

10.4.1 Teaching strategies

As a result of the simulation to introduce students to the Central Limit Theorem, it was apparent that students were more confident in their knowledge that sample means form a Normal distribution if the sample size is large enough. It was also apparent, however, that some students had not connected the fact that approximately 95% of the sample means were within two standard errors of the population mean to the process of calculating confidence intervals. The responses from some of the students also indicated that they thought that confidence intervals refer to the number of occurrences and not to the population mean.

As a result, for this semester the students were required to carry out two additional simulations where they would use a sample mean to estimate the value of the population mean. These were both carried out by hand so that there would be no distractions from following the computer instructions.

In the second week of the semester the students were given small paper bags containing 100 squares of paper on which were written the blood lead levels for workers at a sea port. Working in groups, samples of size 10 were collected and the means of these samples calculated. To ensure random sampling, the samples were taken with replacement, and the bags shaken each time. A number line was drawn on the whiteboard and the sample mean values were place on it. This gave a range of values for estimates of the population mean. The students were asked to agree on the range in which they thought the population mean value might be. The results for one tutorial group are shown in Figure 10.4.1.1. This tutorial group decided that they would estimate the value of the population mean for the blood lead levels to be between 17.5 µg/l and 22.5 µg/l.

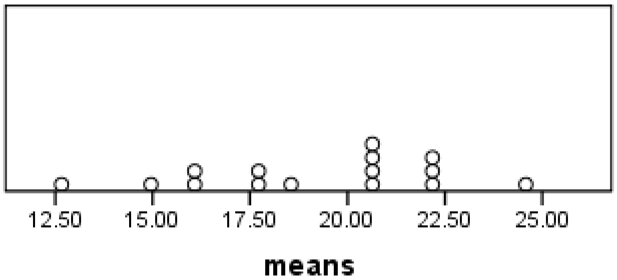
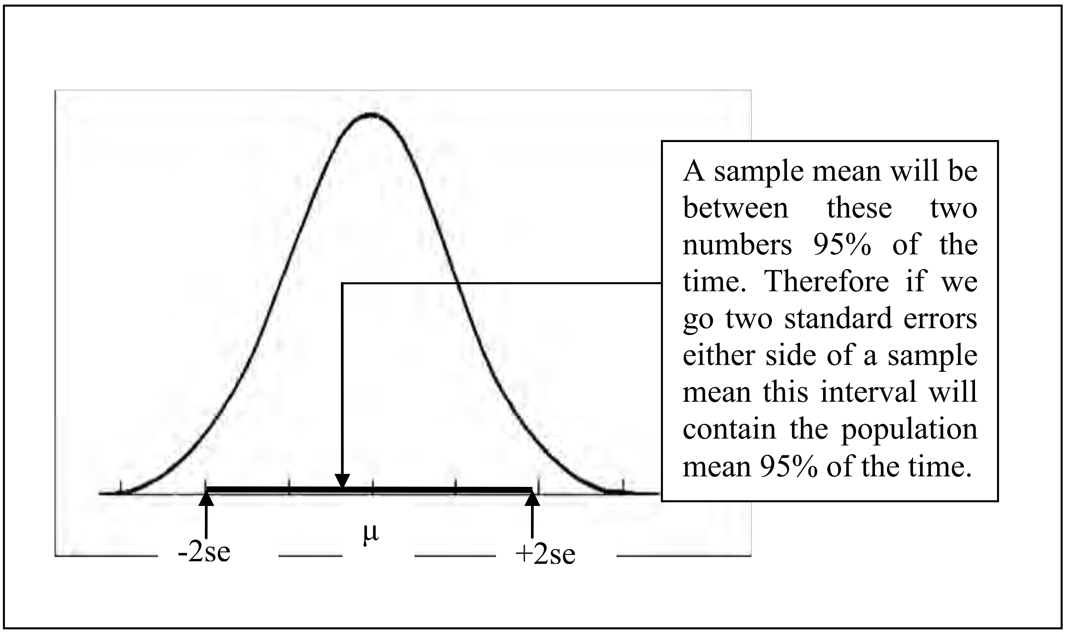


Figure 10.4.1.1. Values of sample means calculated from one tutorial group.

This exercise was designed so that students would become aware of sampling variation and that the value of a sample mean will not be exactly the same as the value of the population mean. They were also made aware that "most" sample means are "near" the population mean although what constitutes "most" and "near" was not clarified at this point.

Later on in the semester, the students were introduced to the Central Limit Theorem via simulation, as described in Section 9.3.1, and then the formal theory of confidence intervals. In the tutorial after this formal introduction, the students were asked to carry out a second simulation to determine the proportion of bumblebees out of the total number of bees in a national park (bumblebees were in the news at that time, as they were displacing native bees in the Tasmanian environment). The students were given large jars of counters of different colours, one of which represented the bumblebees, and asked to take a sample of 30 "bees." From this the students calculated the confidence interval for the proportion of bumblebees.

Another strategy was to use diagrams so that students could visualise the derivation of confidence intervals. One such diagram is illustrated in Figure 10.4.1.2.



A sample mean will be between these two numbers 95% of the time. Therefore if we go two standard errors either side of a sample mean this interval will contain the population mean 95% of the time.

Figure 10.4.1.2. Example of a diagram used to illustrate the derivation of confidence intervals.

10.4.2 Student answers to the confidence interval questions

The responses of six students were available for the study in this semester. In answer to part (a), half gave correct answers, for example, “The mean number of visits to doctors is between 7 and 11.” The other students demonstrated misconceptions, either stating that the mean will fall between 7 and 11, 95% of the time, or that 95% of Tasmanians visited a doctor between 7 and 11 times.

For part (b), two of the students gave correct answers, and these students gave correct answers to part (a) as well. The other students stated that 95% of the time the sample means would be between 7 and 11, that the population mean will between these two numbers 95% of the time, or that the “95%” refers to two standard deviations either side of the population mean. From these responses it was apparent that some students did not understand that they are using a process of inference to find the value of the population mean, and did not appreciate that the population mean is a single figure.

The number of students who agreed to participate in the study for this semester was small, making it difficult to come to firm conclusions. Further details of the students' progression throughout the semester are found in Section 10.6.

10.5 Results of the third cycle of the intervention (Semester 1 – 2009)

10.5.1 Teaching strategies

This was the final semester of the intervention in which data were collected for this research. In this semester the students carried out the same simulations, all those that were carried out by hand and by computer, from the previous two cycles. Throughout the semester the students were also asked to write down what confidence intervals were for and the principles behind their derivation. It was intended that by discussing their answers with each other and with the lecturer, students would become aware of the gaps in their knowledge and try to fill in these gaps (Morgan, 2001; Pugalee, 2001). Because some students appeared to find this process difficult, care was taken not to rush the students. This part of the intervention, combined with the intervention to increase understanding of *P*-values (described in Sections 9.4.1 and 9.5.1) led to an increased interaction between the students and between the students and the lecturer. One result was that the lectures became more informal and on some occasions there was no difference between the lectures and tutorials.

10.5.2 Student answers to the confidence interval questions

Sixteen students agreed to provide data for the study in this semester, twelve of whom completed the test. Eight of these students answered part (a) correctly, in that they stated that the value of the population mean was between 7 and 11. The other students stated either that 95% of Tasmanians visited a doctor between 7 and 11 times, or that there is a 95% chance that the number of people visiting a doctor is between 7 and 11, or that there is a 95% chance that a Tasmanian visited a doctor 7 to 11 times.

For part (a), the biggest difference between this and the previous semesters was that no student received a code of "0." Figure 10.5.2.1 shows the distribution of scores given to the students over the period of the study. It also shows that in the third cycle of the intervention, a higher percentage of students received a score of "2." The differences between the scores of the semester, however, were not significant as determined by the Kruskal-Wallis test (*P* = .397). Details of the analysis can be found in Appendix E5.

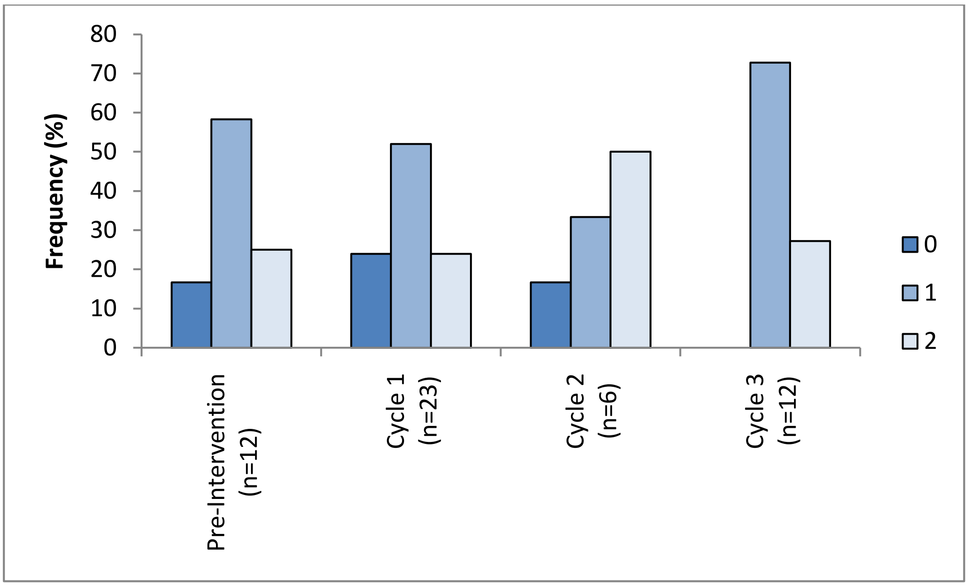


Figure 10.5.2.1. Percentage of each code given to the students' answers to part (a) of the confidence interval question in the test.

For part (b), six students in the third cycle answered correctly, but only two students answered both parts correctly. The other students either stated that the "95%" refers to two standard deviations, that there is a 95% chance that the results (not specified) will fall between 7 and 11, or that 95% refers to the range of population means that fall within two standard deviations.

The biggest difference between the scores received by the students in the final semester and the previous semesters in the study is the higher proportion of students who received a code of "2." Figure 10.5.2.2 gives the distribution of codes given to the students' answers for part (b) of the confidence interval question during the four semesters of the study. The Kruskal-Wallis test indicates that the differences among the semesters were significant, with the third cycle of the intervention having the highest mean rank ($P = 0.009$, see Appendix E5). The mean ranks for both the confidence interval questions for each semester are displayed in Table 10.5.2.1.

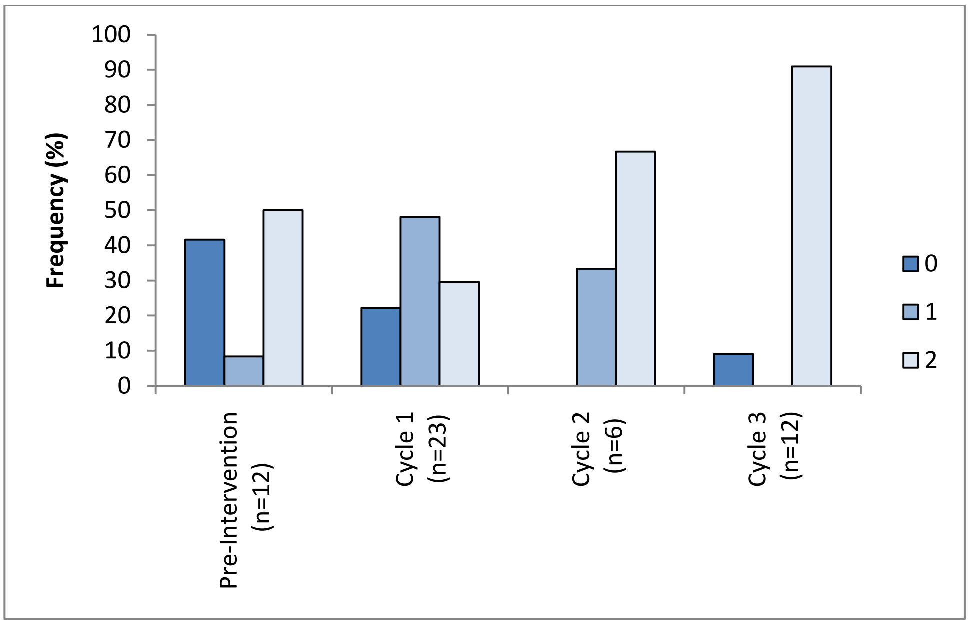


Figure 10.5.2.2. Percentage of each code given to the students to answer part (b) of the confidence interval question in the test.

Table 10.5.2.1

Mean ranked scores for the confidence interval questions

	Mean Rank	
Semester	**Part (a)**	**Part (b)**
Pre-intervention	25.05	26.86
Cycle 1	25.22	21.62
Cycle 2	32.50	31.00
Cycle 3	32.00	38.58

10.6 A description of the teaching strategies used in the teaching of confidence intervals in the second cycle of the intervention

This section gives a description of the strategies used to teach confidence intervals in the second cycle of the intervention. Notes were kept throughout the semester of not only the students' answers but also their reactions to different strategies. This was so that more knowledge about students' problems in understanding confidence intervals could be gained to assist in the planning of the third cycle of the intervention. Owing to ethical considerations the researcher could not know who was participating in the study, therefore no identifying information was kept, and the notes of the students' work were written in a descriptive format only.

10.6.1 Stage 1 – Introduction to the distribution of sample means

The initial step in the teaching of confidence intervals had the aim of giving students experience of the presence of variation and its extent among samples in a simple context. It was also intended to give students the idea that although a sample mean will not have the exact value of the population mean, the sample mean can be used to estimate the value of the population mean.

In the introduction to this exercise, the students were instructed that they were going to carry out some inferential statistics, in that they were going to draw a conclusion about a population from a sample. They were then given envelopes that contained pieces of paper that had written on them on the blood lead levels of 100 people. They were asked to take multiple samples of size 10, with replacement after each individual draw, and calculate the mean of each sample. There were two tutorial groups. The sample means for each group were placed on a number line and these are reproduced in Figures 10.6.1.1 and 10.6.1.2.

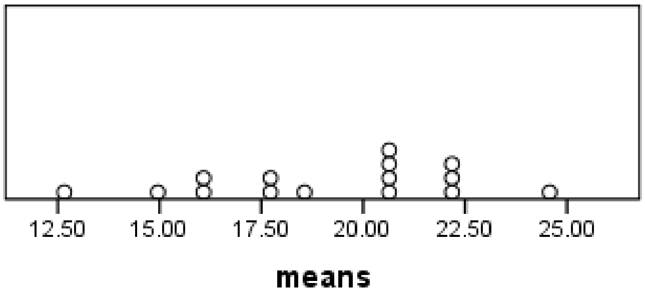


Figure 10.6.1.1. Sample means obtained by the first tutorial group.

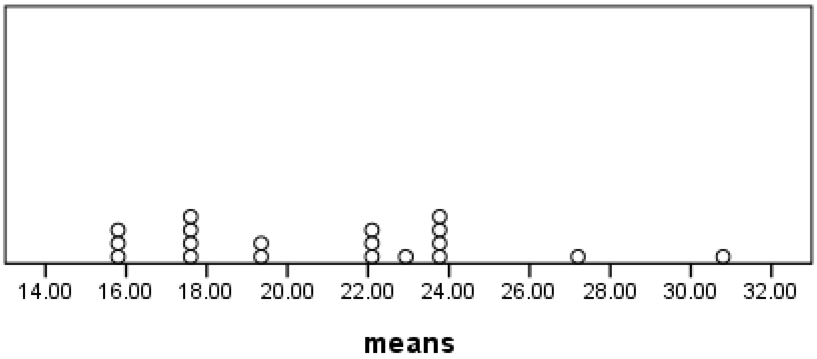


Figure 10.6.1.2. Sample means obtained from the second tutorial group.

In each tutorial group the students were asked to come to an agreement on which results they would regard as outliers, and which results they would include in the range that would estimate the value of the population mean. The students in the first tutorial came to the conclusion that the population mean was a value between 17.5 µg/l and 22.5 µg/l. The students in the second tutorial came to the conclusion that the population mean was a value between 15.9 µg/l and 23.9 µg/l. For this example the value of the population mean (20.0 µg/l) was known, and this value was marked on the diagrams and the conclusion was written, “Most sample means are close to the population mean, but a small number are far away.”

At the end of the lecture that was held after the tutorial described in the previous paragraph, the students were asked to answer the question, “What is the relationship between sample means and the population mean?” The answers were collected to be returned at the next lecture. Most of the students wrote down the conclusion from the tutorial exactly as it had been written on the whiteboard, but some tried to explain the conclusion in their own words. In general, these students stated that sample means give an idea of the population mean, but are not exactly the same as the population mean.

At the beginning of the next lecture, after their responses to the question described in the previous paragraph had been returned, a plot of the sample means from both tutorial groups, shown in Figure 10.6.1.3, was displayed to the students. The mean of the sample means, marked with the triangle and reference line, is 20.0 µg/l, which is equal to the population mean, and the students were asked to add the following statement to their conclusion from the tutorial, “The mean of the sample means we collected is the same as the population mean.” The students were then told it is possible to know the proportion of sample means that will be “near” and “far away” from the population mean and this would be dealt with later in the unit. The next section describes the practical session that was used to introduce the Central Limit Theorem and the formal introduction to confidence intervals.

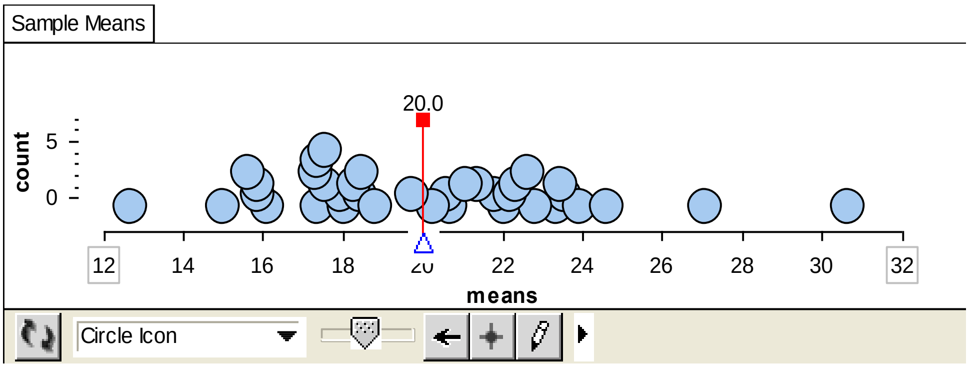


Figure 10.6.1.3. Plot of the sample means calculated from data collected in the tutorials. The mean of the sample means is shown by the reference line (plot produced in Tinkerplots (Konold & Miller, 2005)).

10.6.2 Stage 2 – Putting the information into a mathematical format

After being introduced to the idea that "most" sample means were "near" the population mean, and that sample means could be used to estimate the value of the population mean, the students were then introduced to the Central Limit Theorem via computer simulation. This theorem states that if a sample size is large enough, sample means calculated from samples of the same size form a Normal distribution. Using *Excel*, the students constructed four distributions: A Normal distribution, a Uniform distribution, and two Binomial distributions, one with a sample size of five and other with a sample size of 25. Samples were then drawn and the means of these samples calculated. The students were then required to plot histograms of these sample means.

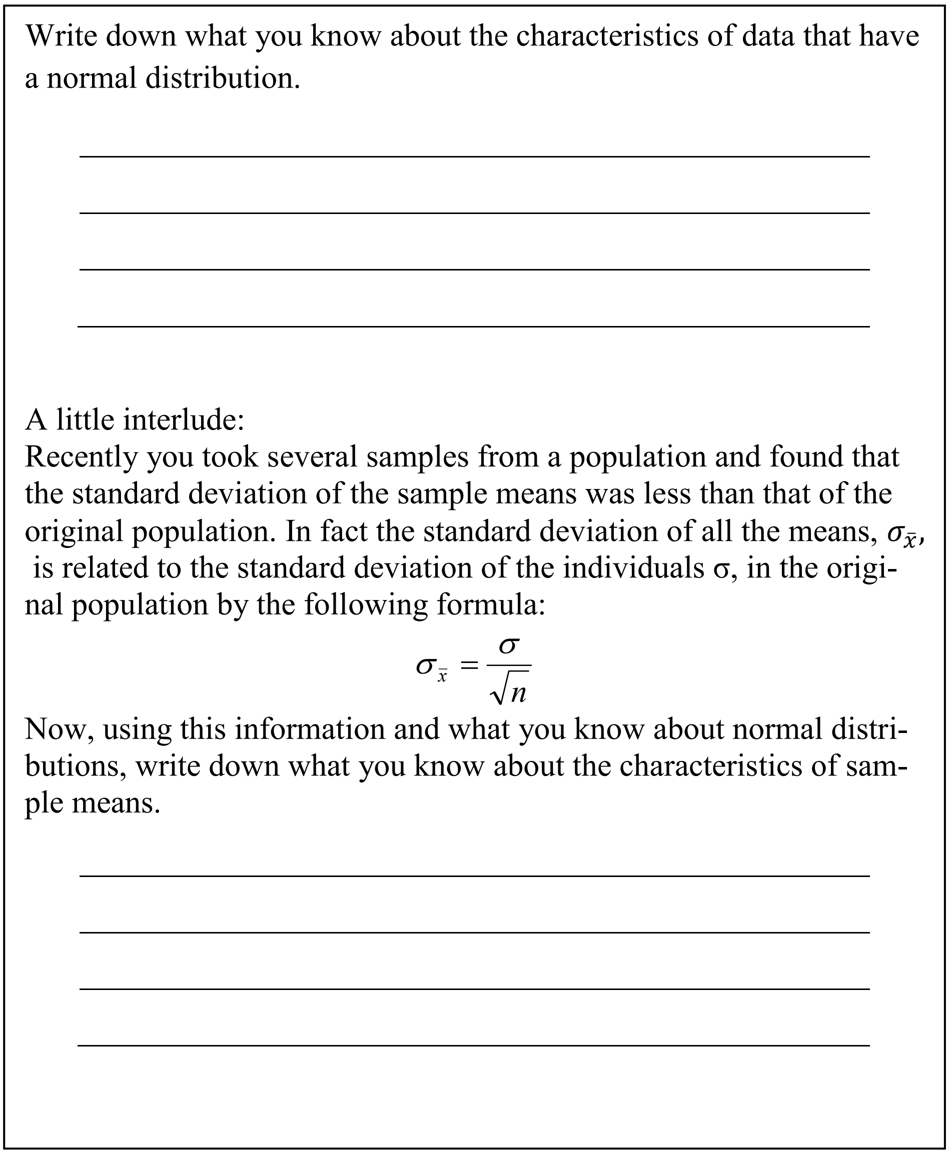
Before the distributions were created, the students were asked to predict what sort of distribution the sample means would have. For the Normal distribution they generally predicted that the sample means would be normally distributed, and this was the result. Most of the students then predicted that the sample means from the Uniform distribution would also be Uniform, and were therefore surprised at the resulting Normal distribution. By the time they had progressed to the Binomial distributions most of the students were suggesting that a Normal distribution might result. As the instructor walked around the room, if a histogram was on a student's screen, he/she was asked the following series of questions.

- What shape does the histogram have?
- What distribution do the sample means have when the original population is Normally/Uniformly/Binomially distributed (as appropriate)?
- Were you surprised?

The idea of the Central Limit Theorem was reinforced whenever students finished the work and left the room. As they left they were asked to explain the principle they had observed that session. If they did not give an answer that in some way described the Central Limit Theorem they were asked a series of questions.

- What distribution did the sample means from the Normal distribution have?
- What distribution did the sample means from the Uniform distribution have?
- What distribution did the sample means from the Binomial distribution with the larger sample size have?
- What is your conclusion?

The next step was for the students to put together the knowledge of Normal distributions gained from a previously held formal lecture, and the knowledge that sample means form a Normal distribution. In the previous simulation, they had also found that the standard deviation of the sample means was less than that of the standard deviation of the original population. They were given a handout (Figure 10.6.2.1) in a tutorial and asked to work in groups to give the answers. In this handout they were expected to be able to state that approximately 67% of sample means would be within one standard error of the population mean, approximately 95% of sample means would be within two standard errors of the population mean, and that approximately 99.7% of sample means would be within three standard errors of the population mean.



Write down what you know about the characteristics of data that have a normal distribution.

____

____

____

____

A little interlude:
Recently you took several samples from a population and found that the standard deviation of the sample means was less than that of the original population. In fact the standard deviation of all the means, $\sigma_{\bar{x}}$, is related to the standard deviation of the individuals σ, in the original population by the following formula:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Now, using this information and what you know about normal distributions, write down what you know about the characteristics of sample means.

____

____

____

____

Figure 10.6.2.1. The first part of the worksheet given to students to make the connection between Normal distributions and the distributions of sample means.

After the answers had been discussed, the students were then encouraged to make the link between the distribution of sample means and statistical inference for the value of the population mean. Because sample means form a Normal distribution, when an individual sample mean is calculated it will be within two standard errors of the value of the population mean 95% of the time (approximately). The second part of the students' handout is shown in Figure 10.6.2.2.

Implications for inference

A sample is taken so that the value of the population mean can be estimated. The mean of this sample is calculated.

How often will this sample mean be within one standard error of the population mean? ____________________
How often will this sample mean be within two standard errors of the population mean? ____________________
How often will this sample mean be within three standard errors of the population mean? ____________________
How often will this sample mean be outside of three standard errors of the population mean?

____

In conclusion, is it more likely that the value of the sample mean is close to the value of the population mean, or far away?

Figure 10.6.2.2. The second part of the worksheet showing the implications for statistical inference for the value of the population mean that result from sample means belonging to a Normal distribution.

The next time the students met was for a practical session where they were given a simulation that gave a practical demonstration of the proportion of sample means that fall within one, two or three standard errors of the population mean. Although the students calculated confidence intervals, this term was not introduced at this stage. The students generated 100 samples with a known population mean using *Excel*. For each sample, the mean, standard deviation, and standard error were calculated. The required intervals were then determined (sample mean ± one standard error, sample mean ± two standard errors, and ± three standard errors). Using the "IF" function in Excel, the students then added up the number of confidence intervals that included the value of the population mean in each case. For one standard error, the student results were in the range of 60% to 74%. For two standard errors, the students results were in the range of 92% to 98%, and for three standard errors, the results were in the range of 99% to 100%.

10.6.3 Stage 3 - Formal introduction to confidence intervals for the mean

The formal introduction to confidence intervals took place in the next lecture that was scheduled after the practical sessions described in the previous section. The purpose of this lecture was to demonstrate the importance of knowing the standard error as well as the sample mean, and to demonstrate how confidence intervals are derived.

The example given was of a peanut butter manufacturer who had two factories, where one of the factories was newer than the other so that the jars were being filled with greater precision. For both factories the jars were to be filled to 500 g. For the older factory the standard deviation of the fill was 5 g, and for the newer factory the standard deviation of the fill was 2 g. In both factories a sample of 25 jars was taken and the sample mean for both factories was 498.7 g. The students were required to calculate the standard errors and work out how many standard errors the sample mean was below the required population mean for each factory. Would they be concerned about this sample mean if they were the quality control manager? For the older factory, the sample mean was 1.3 standard errors below the proposed population mean, and the students indicated that they would not be concerned about this sample value. For the newer factory, however, the sample mean was 3.25 standard errors below the proposed population mean, and the sample indicated that the jars were being under filled.

This example illustrated how a point estimate (reporting only the value of the sample mean) was not adequate, as the same statistic could have different consequences in different contexts. Therefore in many scientific journals the standard error is reported. The students were told about an alternative method, the confidence interval. They had been given lecture notes based on the peanut butter factory examples, and a diagram from these notes is reproduced in Figure 10.4.1.2. The students were then asked to consider the previous computer session when they had observed the proportion of intervals that included the value of the population mean for each number of standard errors either side of the sample mean. They then calculated the approximate confidence interval for the peanut butter factory. For the older factory, the confidence interval for the mean was from 496.7 g to 500.7 g, and for the newer factory, the confidence interval for the mean was from 498.3 g to 499.1 g. The confidence interval for the mean in the newer factory did not include the required value for the peanut butter jars.

In the next lecture the students were given a handout that was designed to consolidate their newly acquired knowledge about the connection between a population and its sample means, and confidence intervals. The work was based on Rowntree (1981). The first section of the handout asked the students to consider what they could say about a population of gnomes if one gnome should come and visit them. What could they say about the height of the population of gnomes? They then had to imagine what further information they could gain if some more gnomes came to join their original visitor, which would let them gain an estimate of the population standard deviation. They were also asked to write down what they knew about their sample mean and the relationship this sample would have to the population of gnomes.

The final section of the handout (also based on Rowntree, 1981) included the diagram shown in Figure 10.6.3.2. The values of three supposed sample means were marked by A, B and C. The students had to draw the length of two standard errors either side of each sample mean and indicate which of these intervals would include the value of the population mean.

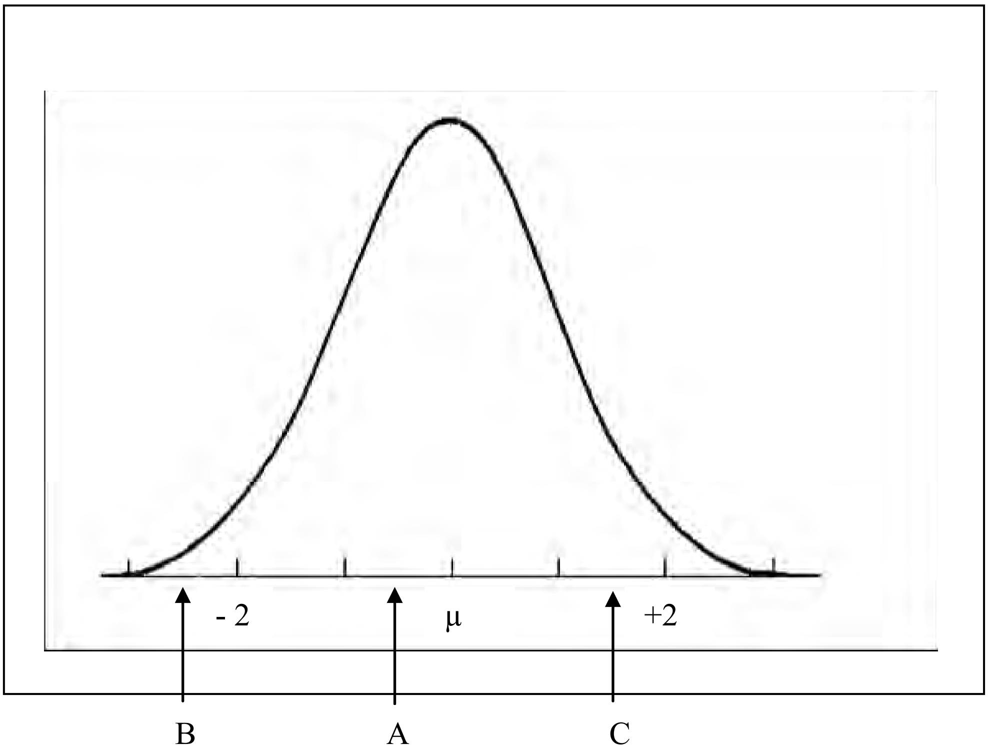


Figure 10.6.3.2. Part of the exercise to consolidate knowledge of confidence intervals. The x-axis on the normal curve is marked in standard errors.

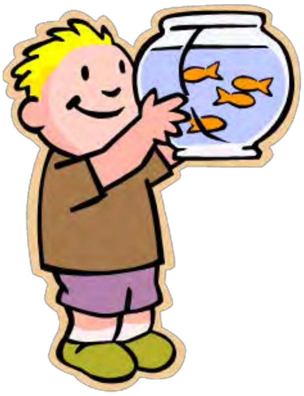
10.6.4 Stage 4 – Practice and consolidation

Towards the end of the semester, the students were given the paper bags with the blood lead data they had used at the beginning of the semester (as described in

Section 10.6.1). From these data, the students were asked to draw out a sample of 25 and to calculate the 95% confidence interval for the population mean. One group found the range to be very wide. On querying their results one of the group realised that the standard deviation, not the standard error had been used.

The students were then asked, “What does the „95%’ refer to?” They were asked to write down their answers and these were returned the next time the students met. The answers to the question indicated that although approximately half of the students could explain the meaning of the “95%” some of the students believed that the “95%” indicated that 95% of the individuals would be within the calculated range. The other students stated that 95% of the sample means would be within the calculated range. It is apparent from these answers that some students did not realise that they were making an inference about the population mean. In addition, whereas they realised that approximately 95% of a normally distributed population is within two standard deviations of the population mean, they had not made the connection from this fact to how confidence intervals are derived.

Because there had been previous exposure to confidence intervals it had been hoped that more of the students would have been able to explain the meaning of a confidence interval. A new handout was devised in an attempt to aid in the visual representation of the principle behind the calculation of confidence intervals. The introduction to this handout is shown in Figure 10.6.4.1.

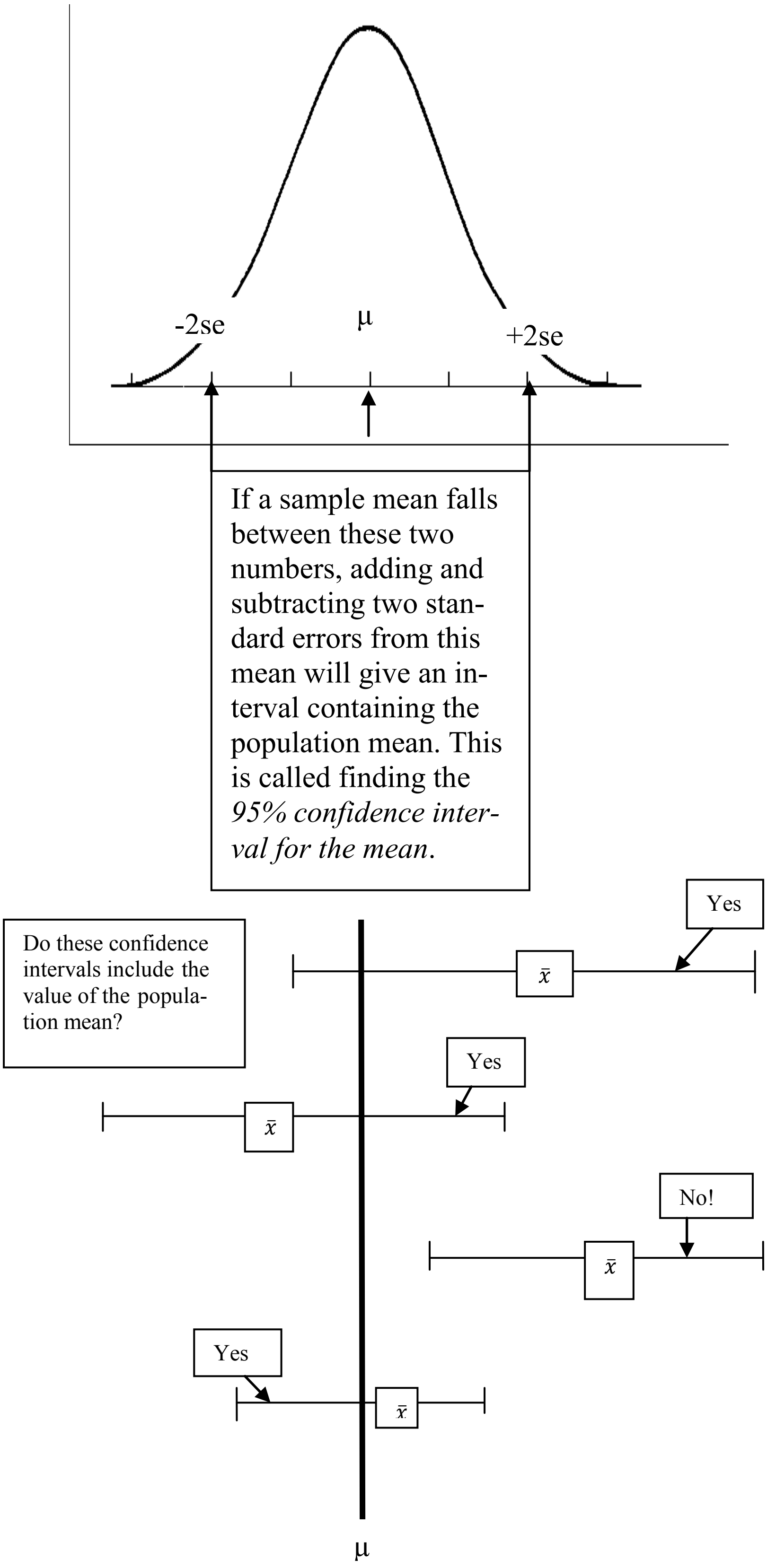


You have several hundred fish in a holding pond. You want to know the mean weight of the fish. Measuring the weight of ALL the fish is not practical. Therefore you take a SAMPLE of 25 fish. The mean of this sample is 300g. The standard deviation of this sample is 20g.

What is your estimate of the value of the mean weight of all the fish?

Figure 10.6.4.1. The introduction to the new handout on confidence intervals.

The handout continued with a series of questions for the students to answer about the distribution of sample means, including the proportion of sample means that would be within one, two or three standard errors of the population mean. The handout continued with the diagram shown in Figure 10.6.4.2. At the conclusion of the handout, the students were required to calculate the confidence interval and state what it meant for this situation.



If a sample mean falls between these two numbers, adding and subtracting two standard errors from this mean will give an interval containing the population mean. This is called finding the *95% confidence interval for the mean.*

Do these confidence intervals include the value of the population mean?

Figure 10.6.4.2. The final part of the new handout on confidence intervals.

10.6.5 Stage 5 – Responses to questions on part of the formal assessment

During the unit the students had to complete four projects as part of their formal assessment. The data for the third project partly consisted of the catch in tonnes of the fishing fleet of Tasmania for four fishing trips. The skippers of the boats were divided into two groups, “experienced” and “inexperienced.” The students were asked to calculate the 95% confidence interval for the mean fishing catch for boats with experienced skippers, and then for inexperienced skippers. They then had to use the data to estimate the range in which the tonnages of 95% of the individual catches for all the boats would fall and explain why this second interval was wider than either of the confidence intervals.

The calculations for these two questions are almost identical but the interpretations of the answers are very different. The answers to the confidence intervals are found by subtracting and adding two standard errors to the sample means. The answers give an estimate of the value of the population mean. Approximately half of the students could successfully give this explanation. To calculate the range where 95% of the individual tonnages will be requires subtracting and adding two standard deviations to the sample mean. Because the latter answer refers to the range of individuals, and not to the mean, this range is much wider than for the confidence interval. Approximately a third of the students gave an explanation along these lines. Some students stated that the range was wider because the standard deviation was used, but did not explain why this was. Some of the students claimed that the interval for the second analysis was wider because the whole data set was used.

At the first lecture after the projects were returned the differences between the two questions were explained. The idea of trying to find a mysterious number "out there" was introduced by the statement, "Only the god of mathematics knows what the value of the population mean is, and he will not tell us!" It was then suggested that an educated guess can be made, that the sample can give us an idea of what the value of the population mean might be.

10.6.6 Stage 6 – revision

In the week before the final lecture, the students were given a practice test and were asked to bring their answers to this final lecture. Before the students shared their answers to this test, they were given a list of some of the questions that had been asked through the semester, and they were asked how they would be answered. The first of these questions was, "What is the value of the population mean?" Initially there was no response. After a short time they were asked to think about how the boy found out the mean weight of the fish in the holding pond. A student then answered that the boy took a sample and calculated the confidence interval. The questions on this practice test that directly related to confidence intervals are presented in Figure 10.6.6.1.

A new bus service has been trialled. To be profitable, in the long term, the expected amount taken in as fares per trip must be at least $100.
Over the trial period of 64 days (*n*=64), the average amount taken in fares per trip is $97 ($\bar{x}$ =97) and the standard deviation of the amounts is $16 (*s*=16).

a. The standard error of the sample mean is $2. Present the formula which is used to compute the standard error and list the values taken by each element in the formula.
b. The sampling distribution of the average amount taken per trip is well approximated by a Normal distribution. Why is it reasonable to make this statement?
c. Present a formula for computing an approximate 95% confidence interval for the expected amount taken per trip, and establish that it produces limits of $93 and $101.
d. Based on the 95% confidence interval, advise the operator on how this finding relates to his requirement that, in the long term, the expected amount taken per trip must be at least $100.

Figure 10.6.6.1. The question on the practice test relating to confidence intervals.

During the discussion on this question it became apparent that some students still had not realised that a confidence interval makes an estimate of the value of a parameter. Disturbingly, there were a small number of students who did not realise that the population mean is one, fixed number, but stated that the confidence interval gives a range where the population mean would be 95% of the time. In addition, there were a small number of students who believed that the confidence interval represented where 95% of the sample means will fall. This statement, while not correct, does connect with the fact that 95% of sample means are within two standard errors of the population mean, but here the connection to the derivation of confidence intervals has not been made. There were no more lectures or tutorials until the day of the test. The results of the relevant test questions for this semester were presented in Section 10.4.

10.7 Summary of the misconceptions identified during the study in comparison to the literature

This study shows that the following misconceptions about confidence intervals were held by some of the students. These misconceptions were demonstrated in the student responses to the test items as part of the formal assessments.

When answering the confidence interval questions, some students stated that ninety-five percent of the sample means will be within the calculated interval; this confirms the findings of delMas, Garfield, Ooms and Chance (2007). Other students stated that ninety-five percent of the population will be within the calculated interval, and this is also confirms the findings by delMas et al. (2007). There was one additional misconception not found in the literature, this was that ninety-five percent of the population means will be within the stated interval.

The last misconception is disturbing, as it suggests that these students, fortunately a very small number, may not even be aware of what the difference is between a sample and a population. Owing to the ethical considerations brought to bear by the researcher being the instructor, it was not possible to carry out follow up interviews to find out if this really was the case, or whether this statement was made because there was only a lack of understanding of the confidence interval, and the students felt that some answer was required. The other misconceptions do relate closely to true statements, that 95% of a population is within two standard deviations of the population mean, and that 95% of the sample means are within two standard errors of the population mean.

10.8 A summary of students' understanding over the intervention

The student responses to the test items about confidence intervals at the end of the pre-intervention cycle (Section 10.2.2) showed that only one quarter of the students could give correct answers. This rose to three quarters by the end of the third cycle of the intervention, but for all semesters in the study most of the students who gave correct responses did not do so for both parts of the question.

The answer that showed the least understanding was the answer that "between 7 and 11 people visited a doctor" for the year. This answer occurred twice in the pre-intervention semester, and once in the third cycle of the intervention. This response is particularly worrying because it makes no sense at any level of understanding. It is not conceivable that only 7 to 11 people visited a doctor in Tasmania over a whole year.

Another response that showed little understanding was that the confidence interval gives the proportion of population means between the two numbers. This is also concerning because statistical inference is based on the relationship between samples and populations, and this response suggests that the students involved did not understand the meaning of the terms "population" and "sample." This needs to made clear at the beginning of the unit.

The other misconceptions, and the more common misconceptions shown by the students, were that 95% of the population visited a doctor between 7 and 11 times, that 95% of the sample means were between these numbers, that the mean (not stated whether population or sample mean) is between these numbers 95% of the time, or that the mean for 95% of the population is between these numbers.

These responses, although not correct in this context, do have the virtue of making some sense, unlike the responses discussed in the previous paragraphs.

Constructivist theory states that if something does make sense to students then they will not change their understanding unless they can see some benefit to do so. It would seem that the students with these understandings find it easier to keep these beliefs than to make the effort to change. What is also of interest in these responses is that they show a relationship to the true situation. It is true that approximately 95% of sample means are within two standard errors of the population mean, and these students seem to have gained this knowledge but failed to apply it to the correct derivation of confidence intervals.

It is the belief of this researcher that no one teaching strategy was of significant benefit in helping students understand the principle behind confidence intervals. Instead, there was a combination of strategies that led to an improvement in students' understanding. The simulations, both by computer and by hand demonstrated first, that no two sample means will be identical, and second, the principle of the Central Limit Theorem. The addition of the diagrams helped give a visual connection to the theory, and the writing about their understanding helped students realise where they had gaps in their knowledge and motivated them to search for the understanding they needed.

10.9 Implications for teaching

It is evident that in each semester there were students who did not understand what confidence intervals are for, that is, to find an estimate of the value of the population mean. The proportion of students with correct answers did not increase over the study to the same extent as for the questions relating to *P*-values. This could be because an example was not found that could play the role of the “It is hot outside” problem that was used in the teaching of *P*-values.

This study confirms a tenet of constructivist theory, that repetition alone, although it may help students to remember, will not help understanding. The example of the boy with the goldfish bowl had the aim of putting a confidence interval into a context that the students could relate to, but this was not entirely successful.

As a result of this study, the concept of confidence intervals is now taught with the following combination of strategies in this unit. First, the students are introduced to the ideas that a sample mean can be used to estimate the population mean and that “most” samples are “near” the population mean by physically drawing samples. This exercise also allows the students to observe the extent to which samples can vary. Second, the Central Limit Theorem is introduced by computer simulation. After the students have come to appreciate that sample means form Normal distributions, they are then encouraged to consider the consequences of this fact. That is, that sample means have the characteristics of all normally distributed data. Students are also asked to write their reasoning of the purpose and process of determining confidence intervals at regular intervals.

One of the demonstrations used in this study is now not always used in the teaching of the unit. Unless the student cohort appears to be of above average computer competence, the demonstration “How confidence intervals work” (see Appendix D8) is not used, as students tend to become caught up in the instructions without thinking about what the results mean.

Over recent semesters the students are told that confidence intervals make an "educated guess" as to the value of the population mean. Students seem to be comfortable with this terminology. However, more research is needed in this area.

11. Discussion with Implications for Teaching

This chapter summarises the findings of the study – the students' understanding of probability and stochastic processes on entering university and the students' understandings of confidence intervals and *P*-values at the end of the unit. The implications of these findings for instructors of statistics are also discussed. Some areas for future research are also identified. This study asked the following questions.

- What are students' understandings of probability and stochastic processes on entering university? Are there any differences in understandings between those students who have studied statistics in their previous mathematics courses and those who have not?
- What are students' understandings of *P*-values at the end of their first tertiary statistics unit? How did these understandings change over the time of the study?
- What are students' understandings of confidence intervals at the end of their first tertiary statistics unit? How did these understandings change over the time of the study?

What are students' understandings of probability and stochastic processes on entering university? Are there any differences in understandings between those students who have studied statistics in their previous mathematics courses and those who have not?

The discipline of inferential statistics would not exist if variation were not universal. Without variation, all samples would be representative of the population and all samples would be alike. As a result of variation, the mathematics of probability is used to make inferences about populations when only samples are available. The answers to the first questionnaire indicated that a substantial proportion of the students had unrealistic views about probability and variation on entering the university. Approximately one quarter of the students did not expect variation in outcomes for a stochastic process; a similar proportion believed that the outcomes of an experiment, represented in graphical form, would be real, even though the outcomes presented were exactly symmetrical around the expected value.

According to Rubin, Bruce, and Tenney (1991), in general, students at senior high school and undergraduate level have little experience in sampling, and therefore do not appreciate how representative (or not) samples may be of their parent populations, and how much one sample may vary from another. This assertion was confirmed by the present study.

In a study by Tversky and Kahnemann (1982b), the authors found evidence to suggest that the participants did not have realistic ideas of the outcomes of stochastic processes. In this study, when students were asked about the outcomes of a process where the two outcomes (odd or even numbers) were equally likely, 40% of the students did not realise that out of 50 trials, getting 6, 46 or 50 even numbers would be in any way unusual. This item was presented in graphical form and the previous item consisted of a graph that was exactly symmetrical. The responses from some of these students suggest that they might have only been comparing the pattern of data with the previous item, and not reading the numbers on the horizontal axis. However, some students specifically stated that any quantity of even numbers would be expected out of 50 trials; these students clearly did not have a realistic view of the possible variation in outcomes for a process such as this.

In general, two out of the three conditional probability questions, which were in a written and graphical form, were relatively easy for the students. For the questions where the students were required to calculate the numerical probability, 84% of the students answered successfully when the question was in a simple "forward" form (You have two white and two black balls, what is the probability of getting a white ball on the second draw if the first ball was white?). When the question was asked in a "backward" form, however, only 19% of the students could answer the problem successfully (If the second ball was white, what is the probability of getting a white ball on the first draw?). Most of the students argued that a later event could not affect an earlier event, not realising that later information could shed light on an earlier event. This gives further evidence to the presence of the "Time Axis Fallacy," described by Fischbein and Schnarch (1997). There may be implications for the understanding of Bayes' theorem, where earlier probabilities are revised in the light of later knowledge. Students will not understand the use of Bayes' theorem without the knowledge that later information can change earlier estimates of the probability of an event. It may even be that familiarity with Bayes' theorem would improve students' performance on this question. Students' understanding of Bayes' theorem could well be a topic of further research.

Tversky and Kahneman (1982b) found that the participants in their study had a tendency not to realise that deviations from the expected value of a random process are more likely in samples of small size. The findings of the present study suggest that students who have previous formal knowledge of statistical independence may be more likely to have this misconception than those who do not have this previous knowledge, but confirmation is needed from a study involving more students without previous statistical experience than were available in this study. This was the only item in the first questionnaire where a significant difference was found between these two groups of students (Appendix E1). The students with previous statistical experience gave more importance to the independence of each individual outcome than to the sample size. It is possible that this knowledge led them not to ask the question, “Is it more likely to get 8 girls born out of 10 births, or 40 girls born out of 50 births?” In addition, some of the students showed evidence of holding the Gambler’s fallacy (Fischbein &Schnark, 1997), that is, the belief that after several Tails had been obtained when tossing a coin that a Head would be more likely.

Of particular interest to the researcher were the responses to the questions in which the students were asked to compare the scores for two groups where the information was given in graphical form. Watson and Moritz (1999) gave the same questions to students in Grades 3 to 9. Twenty-nine percent of the Grade 9 students in their study used strategies that involved using either multiple calculations or proportional reasoning. It would be expected that because the students in the present study were older, a higher proportion would be using such strategies. This did occur, but to a lesser extent than was expected by the researcher. When the students were required to compare the scores for two groups that had the same mean, median and mode, but where one had a wider range than the other, only 55% of them answered this question successfully by stating that the two groups performed equally well. Others answered by saying that one group was more "consistent" than the other, or that one group had more members with "higher" or "lower" scores. It would appear that these students did not see the arithmetic mean as a balancing point or as a representative number of a group. The latter problem was also illustrated by the number of students, who, after performing three comparisons where the group sizes were equal, could not perform a fourth comparison when the group sizes were not equal. For the unequally sized groups, only 51% of the students correctly identified the group with the superior performance. Of these, 39% used proportional reasoning, whereas the others used an estimate of the mean and median, or gave no explanation. Thirty-five percent of the students stated that the group with the inferior performance was better because there were either more people in the group overall, the scores were more "balanced" or there were "more people in the higher range." Eight per cent of the students either said the problem could not be done at all, or was "not fair." Groth and Bergner (2006) have suggested that some students come to university without any knowledge of the arithmetic mean apart from the algorithm used to calculate it and this study confirms their proposition. It is extremely unlikely that there are any tertiary students who have not calculated the arithmetic mean at some time in their previous education. Yet, it appears that some of these students do not have an understanding of what means are used for. As a result of the answers to these questions, the researcher has inserted questions about the purpose of the arithmetic mean into a tutorial session of the first year tertiary statistics unit, and has found that the students usually have difficulty in answering.

There were two ways of approaching the question where there are different numbers of people in each group. One way was to estimate the mean or median, and the other was to use proportional reasoning. Proportional reasoning was required to answer an item where the benefit (or otherwise) of a trial medication was to be assessed by examining the number of people in the treatment and control groups who improved. Fifty-three percent of the students explicitly used proportional reasoning, comparing the proportions in both the treatment and control group. Sixteen percent used proportional reasoning but only mentioned the proportions in one group, while the other students used reasoning that did not involve proportions. Garfield and Ahlgren (1988) found that any misconceptions students may have about probabilistic processes are exacerbated if students have poor mathematical skills, citing proportional reasoning as one of those skills that is most important.

Students' will also have problems if they lack other basic mathematical skills. For example, it was shown in this study that only 57% of the students could correctly calculate the probability of getting four Tails in a row when tossing a coin. Thirty-three percent gave an answer of 1 in 2, and 11% gave the answer of 1 in 4. For a small number of students, it was evident that it was their basic arithmetic, not their reasoning that was at fault (for example, " $1/2 * 1/2 * 1/2 * 1/2 = 1/8$").

It was evident that students were generally unfamiliar with sampling, and in general did not have realistic views on the variation that arises from stochastic processes. It is recommended therefore, that statistics' instructors do not to assume this knowledge exists and should give their students exposure to these processes. This exposure can be provided by the modern computer where large numbers of samples can be generated very quickly and easily using programs such as *Microsoft Excel*. Possible alternatives to *Excel* include *Fathom* software (Konold & Miller, 2005) and samplers available on the World Wide Web. In the unit, as now taught by the researcher, students are asked to predict the range of sample means they expect to get from a given population and then to compare their prediction with their actual values.

This study has also provided further evidence that students may have very little idea of the purpose of the arithmetic mean or the mean and may not see these numbers as representative numbers that can be used for comparisons. For these students hypothesis tests for the comparison of two means will have little meaning. As a result, instructors need to provide opportunities for the students to consider the purposes of these statistics, for example, by providing data sets and asking the students to make comparisons between them.

It was of concern to the researcher that because some of the students studying the Data Handling and Statistics unit were enrolled in a course where there was a higher Tertiary Entrance Score than other units, and where there was a requirement to have completed a pre-tertiary mathematics unit, that these students would be more mathematically competent than the other students. These students were concentrated in the first semester of each calendar year. It was found however, that there was no significant difference in mean ability among the semesters for the students who completed the first questionnaire. This suggests that the students' previous mathematical experience did not influence their ability to complete the first questionnaire. The details of this analysis are in Appendix E1.

What are students' understandings of *P*-values at the end of their first tertiary statistics unit? How did these understandings change over the time of the study?

The results of this study indicated that, whereas a minority of students explained the *P*-value fully in their final formal assessments before the initial intervention, the proportion that did so increased over the time of the study. This was confirmed by the Kruskal-Wallis test where the highest mean rank was found at the end of the third cycle of the intervention (see Appendix E4). Apart from the second cycle of the intervention, where only six students contributed data, there was an increase in mean ranked scores for both test questions that applied to *P*-values. Students who described the *P*-value correctly indicated that the *P*-value was the probability of a sample observation or a more extreme observation, if the null hypothesis were correct. Some of the students showed a partial understanding, in that they stated that the *P*-value is the probability of the observation if the null hypothesis were true. Although this reasoning is incomplete, it does indicate that these students realise that hypothesis tests look at the likelihood of observations given a proposition about a population. This partial understanding of a *P*-value is simpler to understand than the correct definition, and would appear to make sense to the students.

The most common misconception in defining *P*-values was found to be the belief that *P*-values give the probability that the null hypothesis is true. This misconception has also been reported by Gliner et al. (2002). It was found in this study that this misconception was difficult to correct. Telling the students in lectures and tutorials, and including an explanation in the students' lecture notes, was not effective. Garfield and Ahlgren (1988) suggested that known misconceptions should be pointed out to students so that they may be avoided, but this study suggests that if a belief is firmly held, this may not be enough. Strike and Posner (1985) suggested that students will not change their conceptions unless presented with a good reason to do so. One way to do this is to confront students with their misconceptions by the use of the predict-test-evaluate format (Garfield & Ahlgren, 1988; Hardiman et al., 1986; Mills, 2002; Posner et al., 1982). Finding a way to use this format for abstract ideas as the *P*-value, however, is difficult.

Some text books try to avoid this misconception, that the *P*-value gives the probability that the null hypothesis is true, by using the term “not reject the null hypothesis” in place of “accept the null hypothesis” when the *P*-value is higher than the level of significance (For example Lind, Marchal & Mason, 2001, and Croucher, 2002). During the time of this study the practice used by the instructors in the unit, “accept the null hypothesis was used.” Since the time this study was completed the practice has been changed to “not reject the null hypothesis.” It is observed, however, that some students are still holding this misconception. More research needs to be done in this area.

The responses to the first questionnaire indicated that over 80% of the students could answer simple conditional probability questions, both in tabular and verbal form. This study showed, however, that the more complex process of the hypothetical, probabilistic reasoning used in hypothesis testing was difficult, and confirmed that students were unfamiliar with this process (Garfield & Ahlgren, 1988). This may partly explain why students have difficulty in making the change from believing that the *P*-value is the probability of the null hypothesis being true, to the correct definition. It seems likely that this misconception seems more intuitive than the formal definition of a *P*-value, and makes sense to the students. Strike and Posner (1985) stated that for a concept to be changed, the student must find the new interpretation makes sense, and that some benefit is to be gained by making the change. If this misconception works well for students trying to make sense of the process of hypothesis testing, they will be reluctant to make the extra effort to change their understanding.

Other misconceptions shown by this study included that the *P*-value is the probability of being incorrect (also identified by Gliner et al., 2002), the rate of replication of the conclusion, the probability of seeing a difference, or the rate at which the treatment will show a difference.

Cobb and McClain (2004, cited in Garfield and Ben-Zvi, 2008, p. 48) suggest that statistics instructors should not underestimate how difficult the process of hypothetical, probabilistic reasoning is for new students of statistics. In this study this was demonstrated not only by students' difficulties in understanding *P*-values, but also by the fact that only one student confidently used the given probability in the circuit breaker questions in the second questionnaire, without resorting to a formal hypothesis test. The ideal answer to this item would have stated that even if the underlying rate of defective circuit breakers is only 5%, the probability of 13% of getting three defectives in a box of 25 breakers suggests that this event is not so rare as to suggest the underlying rate is greater than specified. Most students avoided using these numbers in their reasoning, being content to state that since variation exists, other boxes will have more or less defectives. This suggests that students found the latter reasoning easier to do than to use the hypothetical reasoning involving the numerical values.

A small number of students stated that if a *P*-value were below .05, then the null hypothesis should be accepted. All of these students made no attempt to otherwise define the *P*-value, suggesting that these students do not have an understanding that the *P*-value is a probability that in some way relates to the sample observation, but instead have tried to learn a rule, which they then have misapplied. The tendency of students to try to compensate for a lack of understanding by learning rules has also been noted by Chance et al. (2004).

Some considerable time was spent by the researcher (including consultation with colleagues) to find an example that used probabilistic reasoning and to which students could easily relate. If the right example were found, it was hoped it could be used to help the unfamiliar become familiar. Eventually an example was found in Shaughnessy and Chance (2005) and became the basis for the "It is hot outside" problem where the students were asked to judge the likelihood of observing people walking about in winter clothing if the weather were hot. In the last two cycles of the intervention this problem was introduced in the first week of the semester and became the basis of all further work in hypothesis testing. With this process the students became familiar with the terminology, procedure and reasoning of hypothesis tests over a period of time. Although there is no formal evidence, from the researcher's point of view the students appeared to accept the process of hypothesis testing more readily than before the intervention, where all of the elements of hypothesis testing had been introduced in one lecture.

In this study it was also found that the use of diagrams in the process of hypothesis testing and calculating *P*-values was an aid to understanding for some students. In these diagrams, the distribution of the sample statistic centred on the value of the null hypothesis was drawn and the test statistic placed on the distribution. On the basis of the placement of the test statistic, a visual judgement could be made as to the likelihood of this statistic or a more extreme statistic given the null hypothesis and related to the numerical *P*-value.

Morgan (2001) and Pugalee (2001) suggested that if students are encouraged to explain their reasoning, their ideas are clarified as they work to explain their ideas. Consistent with this, another strategy was to ask students to write down the definition of the *P*-value for each context in words. Each time this was done the students were encouraged to talk to each other before sharing their answers to the whole group. It was also found important to allow the students time to do their writing in a relaxed atmosphere.

From this study, it is apparent that students' understanding of *P*-values can be enhanced, but one strategy will not work for all students, but rather a mixture of strategies is required. It is of note that it was in the third cycle of the intervention, where all strategies were combined, that the highest mean ranks were obtained for the *P*-value questions. The results suggest that instructors need to appreciate how unfamiliar students are with the hypothetical, probabilistic reasoning used in making decisions with *P*-values. Consequently, students will have problems in explaining a *P*-value, and if they resort to learning a rule, they are likely to make mistakes in its use. It is therefore recommended that instructors use examples from a variety of sources, and use a variety of techniques in aiding students to make sense of their work.

What are students’ understandings of confidence intervals at the end of their first tertiary statistics unit? How did these understandings change over the time of the study?

The Central Limit Theorem states that if the sample size is large enough, the sample mean belongs to a Normal distribution, regardless of the distribution of the original population. Because the sample means are normally distributed, then it follows that approximately 95% of the sample means will be within two standard deviations of the population mean. The standard deviation of the sample means is less than that of the population, and is called the “standard error of the mean.” It follows that if 95% of sample means are within two standard errors of the population mean, then 95% of the time the sample mean ± two standard errors will include the value of the population mean.

This study has indicated that students find such reasoning difficult. They seem to find the information that sample means belong to a Normal distribution fairly easy to comprehend, but then have immense difficulty in making the leap to how this leads to a process for estimating the population mean. The teaching of the topic was taken in steps, the first of which introduced students to the idea that sample means vary, and that “most” sample means are “near” the population mean. The students performed a simulation as an introduction to the Central Limit Theorem, and were then introduced to the idea that the facts about any Normal distribution can be applied to sample means. It is because sample means are normally distributed that the process of estimating a population mean by a confidence interval is possible.

The difficulty that students have in understanding confidence intervals is illustrated by their misconceptions: that the confidence interval contains 95% of the sample means (the most common misconception), that 95% of the population is within the interval (also reported by delMas et al., 2007), or that the interval contains the mean for 95% of the population. From students' responses it was also evident that some students did not realise the purpose of confidence intervals, that is, to use the sample to make an estimate of the value of the population mean. This misconception persisted despite the fact that students were given practice in taking samples and using them to estimate the population mean.

The belief that 95% of the sample means (approximately) are within the confidence interval is only a small difference from the knowledge that 95% of the sample means are within two standard errors of the population mean. It is therefore understandable that students should hold such a misconception. It is a statement that is easily understood, although factually incorrect, and this is possibly why the misconception is hard to correct. If the students believe that this misconception helps their understanding and it makes sense, then they will not be motivated to change it. The challenge for instructors is to find ways that will cause the students to want to change this belief.

It was also apparent that some students were extremely confused by the distinction between the "standard deviation" and the "standard error." This was illustrated by the responses to questions both in the test and the second questionnaire. Some students stated they had used the standard deviation in their responses but had actually used the standard error as required. Others stated they had used the standard error in their responses but had actually used the standard deviation, and there was a small number who used the standard deviation and seemed oblivious to the existence of the standard error. These problems are possibly not helped by the terminology, and this is another occasion when directly telling the students about a problem did not necessarily prevent it.

The strategy that had some positive effect was coming back to the concept of the confidence interval several times, not just restating the problem, but giving the students time to write about their understanding. The students were also encouraged to ask questions of each other and of the researcher. Writing about a topic helps students recognise gaps in their knowledge and understanding, and helps them to fill in these gaps (Morgan, 2001; Pugalee, 2001). Strike and Posner (1985) also pointed out that students, when working to achieve accommodation of new ideas, may attempt many strategies and make many mistakes before this accommodation is achieved. It is of note that it was when this strategy was used, in the third cycle of the intervention, the mean rank for the second part of the confidence interval question was at its highest (see Appendix E5). Whereas there was not a significant difference in mean ranks among the semesters for the first part of the confidence interval question, it was of note that no student received the lowest score in the third cycle of the intervention (see Figure 10.5.2.1), which was a change from previous semesters.

It is recommended, therefore, that statistics instructors, appreciating students' difficulties in this area, give students time to reflect upon and write about their understanding, and to create an environment in which students feel comfortable to share their work with others.

Constraints on the research - suggestions for further research

In this study, the researcher and the instructor in the statistics unit were the same person. It was imperative, therefore, that students were not constrained in any way to take part in the research, and did not feel that they could be penalised if it was not their wish not to take part. One consequence was that the number of students who agreed to take part varied considerably, for no apparent reason. The participation rate varied from 20% to 85% of the students taking the unit.

More importantly, because the researcher was only given the student numbers of those who participated after their unit results had been formalised and published (all their assessed work was labelled by student number only) it was not possible to carry out follow up interviews. It is believed that in particular, the area of student understanding of confidence intervals requires more research. The literature is not extensive in this area, and this study shows that students are having difficulties that are not easily corrected. More research is required into the influence of the different terminology (standard deviation, standard error) on student understanding. More research is also required into the difficulties students have in making the connection between understanding that sample means belong to a Normal distribution, that 95% (approximately) of sample means are within two standard errors of the population mean, and the consequence that 95% of the intervals calculated by the sample mean ± two standard errors will contain the value of the population mean. The last step appears to be the problem, and if more knowledge could be gained about this problem, it is likely that alternative strategies could be developed to aid students in making this step. More research is also required into students' understanding so that instructors may find ways to change the belief that a confidence interval contains 95% of the sample means.

Because the participants were working towards various degrees, the course content, as set by the School of Mathematics and Physics at the university where this study took place, had to be covered and could not be changed by the researcher.

12. A personal reflection

This study originated from frustration – frustration that the students I taught were not "getting it." In discussions with colleagues and on reading the literature it became apparent that my concern was of concern to others as well. It also originated from the realisation of my own lack of understanding as an undergraduate. Despite my high grades, I cannot say that I "got it" before doing post-graduate work in statistics. I now realise that I had not really appreciated the significance of the Central Limit Theorem, particularly that it described a connection between samples and populations. I also remember the frustration of trying to learn what the *P*-value represented by rote and failing at this. As conceptual understanding was achieved, the *P*-value became easy to use. I was determined, therefore, to find alternative ways to make it possible that my students would understand the concepts behind their work – to try and reduce the number of "cook book" statisticians.

During the study I have found that it is possible to help students to develop a greater understanding than is achieved by didactic teaching. This was achieved by a combination of teaching methods, increased interaction with the students, and by the connecting the strange world of inferential statistics to things they already know. A world where nothing is certain is unfamiliar to many, and some of the students who said they were "good at maths" openly stated that they found this world frustrating and uncomfortable. These students were used to working in an environment where there were right answers and enjoyed finding these answers.

Although there is no empirical evidence, it is my belief that with the new teaching program the students are much more engaged in their work. Some of them say that they have to think about the numbers, that the scientific journals they read make more sense, and that they now listen to reports in the media with more scepticism. I also find that the increased interaction I have with the students during instruction time has made my job more interesting and enjoyable. It is now not unusual for students to tell me about how their work in statistics has helped in their other subject areas. It is also not unusual for a student to stop me as I go about the campus to tell me about something he or she has read where statistical analysis was important and that he or she thought interesting.

From my point of view as instructor, the introduction of the simulations has smoothed the teaching path in three areas where previously I felt the students were having particular difficulties. First of all, although not all the students make the final link between the Central Limit Theorem and confidence intervals, they are much more likely than before to remember that sample means form a Normal distribution. In the previous teaching format the students did not often remember this, and again repetition was not a successful strategy. I have also found that students' understanding of the principles behind the analysis of variance has improved. Asking the students to make a visual assessment of where the significant differences are before the formal lecture has led the students to appear more interested when the variation "among" and "within" groups is discussed. They also seem to enjoy seeing how their predictions compare with the results of the formal calculations. The third area where the students now seem to have much less difficulty is testing for a significant relationship between the dependent and independent variables in linear regression. Because the students have seen how the gradient of the lines may vary even when there is no relationship, they seem comfortable with the idea that a non-zero gradient may occur in this situation.

I have also found that since the students have been introduced to visual representations of *P*-values, the last lecture of the semester, on Type I and Type II errors, goes much more smoothly. Because they are already familiar with sketches of probability distributions, they appear to find the diagrams used to describe how the distribution of the "true" situation may overlap with the distribution of the proposed situation (the null hypothesis) easier to understand. The questions on the test that ask students to explain the consequence of a Type I error in a particular context are usually done well.

A constant frustration throughout the study has been students' understanding of confidence intervals. A colleague recently asked, "How do we get across that there is this number „out there' we are trying to find?" There has been an improvement in understanding over the time of the study, but not to the same extent as for *P*-values. Since the time of the study I have been stating that the sample mean allows us to make an educated guess of the value of the population mean, and this seems to work well. Here more research is needed.

Another frustration is the materials supplied by the School of Mathematics and Physics, which the students were given in place of a text book. The students frequently complain (verbally and in their student evaluations of the unit) about these materials, stating that they are difficult to comprehend. There is hope for the future, however, as it has been suggested that these materials be rewritten and that I should be involved. It is gratifying that my research should be of use in this way.

Finally, it has been pleasing to me to read comments such as, “The lecturer makes a boring subject interesting,” in the student evaluations. It is my aim that one day the students will say, “This subject is interesting.”

REFERENCES

Atkinson, P., & Hammersley, M. (1998). Ethnography and participant observation. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research* (pp. 248-261). London: Sage Publications.

Australian Curicculum, Assessment and Reporting Authority. (2011). *The Australian Curriculum: Mathematics.* Version 1.2, 8 March, 2011. Sydney: ACARA.

Barrett, P. (2003). Beyond psychometrics: Measurement, non-quantitative structure, and applied numerics. *Journal of Managerial Psychology, 18*(5), 421-439.

Batanero, C. (2008). Statistics education as a field for research and practice. In M. Niss (Ed.), *Proceedings of the 10th International Congress on Mathematical Education, 4-11 July, 2004.* [CDRom] Copenhagen: IMFUFA, Department of Science, Systems and Models, Roskilde University.

Baumgartner, E., Bell, P., Brophy, S., Hoadley, C., Hsi, S., Joseph, D., et al. (2003). Design based research: An emerging paradigm for educational enquiry. *Educational Researcher, 32*(1), 5-8.

Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning and thinking: Goals, definitions, and challenges. In D. Ben-Zvi & J. Garfield (Eds), *The challenge of developing statistical literacey, reasoning and thinking* (pp. 3-15). Dordrecht, The Netherlands: Kluwer Academic Press.

Biggs, J. (1999). *Teaching for quality learning at university*. Buckingham: Society for Research into Higher Education and Open University Press.

Biggs, J., & Collis, K. (1982). *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome)*. New York: Academic Press.

Bock, R. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice, 16*, 21-33.

Bond, T., & Fox, C. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. New Jersey: Lawrence Erlbaum Associates.

Boulton-Lewis, G. (1995). The SOLO taxonomy as a means of shaping and assessing learning in higher education. *Higher Education Research and Development, 14*(2), 143-154.

Bower, K. (2003, May). *Some misconceptions about the normal distribution.* Paper presented at the Six Sigma Forum for the American Society of Quality, Milwaukee, WI.

Bransford, J., Brown, A., & Cocking, R. (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Brightman, H., & Schneider, H. (1992). *Statistics for business problem solving*. Cincinnati, OH: South-Western Publishing Company.

Burrill, G. (2002). Simulation as a tool to develop statistical understanding. In B. Phillips (Ed), *Developing a statistically literate society* (Proceedings of the 6th International Conference on Teaching Statistics, Cape Town, South Africa) [CDRom]. Voorburg, The Netherlands: International Statistics Institute.

Callingham, R., & Watson, J. (2005). Measuring statistical literacy. *Journal of Applied Measurement, 6*(1), 2-31.

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenges of developing statistical literacy, reasoning and thinking* (pp. 295-323). Dordrecht, The Netherlands: Kluwer Academic Press.

Chance, B., & Rossman, A. (2006). Using simulation to teach and learn statistics. In A. Rossman & B. Chance (Eds.), *Developing a statistically literate society* (Proceedings of the 7th International Conference on Teaching Statistics, Salvador, Brazil) [CDRom]. Voorburg, The Netherlands: International Statistics Institute.

Clarke, R. (1946). An application of the Poisson distribution. *Journal of the Institute of Actuaries, 72*, 481.

Cobb, P., & McClain, K. (2004). Principles of instructional design for supporting the development of students' statistical reasoning. In D. Ben-Zvi & J. Garfield (eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 375-395). Dodrecht: Kluwer Academic Publishers.

Cobb, G., & Moore, D. (1997). Mathematics, statistics and teaching. *The American Mathematical Monthly, 104*(9), 801-823.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9-13.

Confrey, J. (1990). What constructivism implies for teaching. In R. Davis, C. Maher & N. Noddings (Eds.), *Constructivist views on the teaching and learning of mathematics* (pp. 107-210). Reston,VA: National Council of Teachers of Mathematics.

Croucher, J. (2002). *Statistics: Making business decisions.* Sydney: McGraw-Hill Irwin.

Cumming, G. (2006). Understanding replication: Confidence intervals, *p*-values, and what's likely to happen next time. In B. Phillips (Ed.), *Developing a statistically literate society* (Proceedings of the 7th International Conference on Teaching Statistics, Salvador, Brazil) [CDRom]. Voorburg, The Netherlands: International Statistics Institute.

de Jong, T., Martin, E., Zamarro, J.-M., Esquembre, F., Swaak, J., & van Joolingen, W. (1999). The integration of computer simulation and learning support: An example from the physics domain of collisions. *Journal of Research in Science Teaching, 36*(5), 597-615.

de Jong, T., & van Jooligen, W. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research, 68*(2), 179-201.

delMas, R., Garfield, J., & Chance, B. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education, 7*(3). Retrieved from http://www.amstat.org/publications/jse

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal, 6*(2), 25-58.

Doane, D., & Seward, L. (2007). *Applied statistics in business and economics*. Boston, MA: McGraw-Hill Irwin.

Dunbar, K., Fugelsang, J., & Stein, C. (2004). In M. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 193-205). New York: Laurence Erlbaum and Associates.

Erickson, T. (2006). Using simulation to learn about inference. In B. Phillips (Ed.), *Developing a statistically literate society* (Proceedings of the 7th International Conference on Teaching Statistics, Salvador, Brazil) [CDRom]. Voorburg, The Netherlands: International Statistics Institute.

Ernest, P. (1998). The epistemological basis of qualitative research in mathematics education: A postmodern perspective. In A. R. Teppo (Ed.), *Qualitative research methods in mathematics education* (pp. 22-39). Reston, VA: National Council of Teachers of Mathematics.

Falk, R. (1986). Conditional probabilitites: Insights and difficulties. In R. Davidson & J. Swift (Eds.), *Proceedings of the Second International Conference on Teaching Statistics* (pp. 714-716). Victoria, BC: University of Victoria.

Feldman, A., & Minstrell, J. (2000). Action research as a research methodology for the study of teaching and learning of science. In A. Kelly & R. Lesh (Eds.), *Handbook of Design in Mathematics and Science Education* (pp. 429-456). Mahwah, NJ: Lawrence Erlbaum and Associates.

Finch, S. (1998). Explaining the law of large numbers. In L. Pereira-Mendoza, L. Seu Kea, T. Wee Kee, & W. K. Wong (Eds.), *Statistical education – Expanding the network* (Proceedings of the 5th International Conference on Teaching Statistics, Vol.2, pp. 731-736). Voorburg, The Netherlands: International Statistical Institute.

Fischbein, E., & Schnarch, D. (1997). The evolution with age of probabilistic, intuitively based misconceptions. *Journal for Research in Mathematics Education, 28*(1), 96-105.

Franklin, L. (1992). Using simulation to study linear regression. *The College Mathematics Journal, 23*(4), 290-295.

Gal, I., Rothschild, K., & Wagner, D. (1990, April). *Statistical concepts and statistical reasoning in school children: Convergence or divergence?* Paper presented at the Meeting of the American Educational Research Association, Boston, MA.

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education, 10*(3). Retrieved from http:/www.amstat.org/publications/jse

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education, 19*(1), 44-63.

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice.* Dordrecht, The Netherlands: Springer.

Garfield, J., delMas, R., & Chance, B. (2007). Using students' informal notions of variability to develop an understanding of formal measures of variability. In M. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 117-148.). Mahwah, NJ: Lawrence Erlbaum.

Gliner, J., Leech, N., & Morgan, G. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *The Journal of Experimental Education, 71*(1), 83-92.

Gould, S. (1981). *The mismeasure of man*. New York: WW Norton & Company.

Groth, R., & Bergner, J. (2006). Preservice elementary teachers' conceptual and procedural knowledge of the mean, median and mode. *Mathematical Thinking and Learning, 8*(1), 37-63.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7*(1), 1-20.

Hammersley, M. (1993). On the teacher as researcher. *Educational Action Research, 1*(3), 425-445.

Hancock, C., Kaput, J., & Goldsmith, L. (1992). Authentic enquiry with data: Critical barriers to classroom implementation. *Educational Psychologist, 27*(3), 337-364.

Hardiman, P., Pollatsek, A., & Well, A. (1986). Learning to understand the balance beam. *Cognition and Instruction, 3*(1), 63-86.

Hsueh, I., Wang, W., Sheu, C., & Hsieh, C. (2004). Rasch analysis of combining two indices to assess comprehensive ADL function in stroke patients. *Stroke, 35*(3), 721–726.

Kahneman, D., & Tversky, A. (1982). Subjective probability: A judgement of representativeness. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 32-47). New York: Cambridge University Press.

Kaplan, J. (2009). Effect of belief bias on the development of undergraduate students' reasoning about inference. *Journal of Statistics Education, 17*(1). Retrieved from http:/www.amstat.org/publications/jse

Kelly, A., Sloane, F., & Whittaker, A. (1997). Simple approaches to assessing underlying understanding of statistical concepts. In I. Gal & J. Garfield

(Eds.), *The assessment challenge in statistics education* (pp. 85-90). Amsterdam: IOS Press.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction. *Psychological Science, 15*(10), 661-667.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction, 6*(1), 59-98.

Konold, C. (1995). Issues in assessing conceptual understanding in probability and statistics. *Journal of Statistics Education, 3*(1). Retrieved from http:/www.amstat.org/publications/jse

Konold, C., & Miller, C. (2005). *Tinkerplots* (Version 1.0). [Computer Software]. Emeryville, CA: Key Curriculum Press.

Krause, K., Bochner, S., & Duchesne, S. (2007). *Educational psychology for learning and teaching*. South Melbourne, VIC: Thomson.

Kubiszyn, T., & Borich, G. (2003). *Educational testing and measurement: Classroom application and practice.* (7th ed.). New York: John Wiley and Sons.

Kuhn, T. (1996). *The structure of scientific revolutions* (3rd ed.). Chicago: University of Chicago Press.

Lane, D., & Peres, S. (2006). Interactive simulations in the teaching of statistics: Promise and pitfalls. In B. Phillips (Ed.), *Developing a statistically literate society* (Proceedings of the 7th International Conference on Teaching Statistics, Salvador, Brazil) [CDRom]. Voorburg, The Netherlands: International Statistics Institute.

Leavy, A. (2006). Using data comparison to support a focus on distribution: Examining preservice teachers' understandings of distrubtion when engaged in statistical enquiry. *Statistics Education Research Journal, 5*(2), 89-114.

LeCoutre, M. (1992). Cognitive models and problem spaces in "purely random" situations. *Educational Studies in Mathematics, 23*(6), 557-568.

LeCoutre, M., Rovira, K., LeCoutre, B., & Poitevineau, J. (2006). People's intuitions about randomness and probability: An empirical study. *Statistics Education Research Journal, 5*(1), 20-35.

Lind, D., Marchal, W. & Mason, R. (2001). *Statistical techniques in business and economics* (11th ed). Boston, MA: McGraw-Hill Irwin.

Lipson, K. (2002). The role of computer based technology in developing understanding of the concept of sampling distribution. In B. Phillips (Ed.), *Developing a statistically literate society* (Proceedings of the 6th International Conference on Teaching Statistics, Cape Town, South Africa). [CDRom]. Voorburg, The Netherlands: International Statistical Institute.

Lipson, K., Kokonis, S. & Francis, G. (2003). Investigation of students' experiences with a web-based computer simulation. *Proceedings of the International Association for Statistical Education Satellite Conference on Statistics and the Internet, Berlin, August, 2003*. [CDRom]. Voorburg, The Netherlands: International Statistical Institute.

Liu, Y, & Thompson, P. W. (2005). Teachers' understanding of hypothesis testing. In S. Wilson (Ed.), *Proceedings of the 27th Annual Meeting of the International Group for the Psychology of Mathematics Education,* Vicksburg, VA: Virginia Tech.

Martin, F., & Säljö, R. (1976). On qualitative differences in learning. 1: Outcome and process. *British Journal of Educational Psychology, 46*(1), 4-11.

Mason, J., & Spence, M. (1999). Beyond mere knowledge of mathematics: The importance of knowing-to act in the moment. *Educational Studies in Mathematics, 38*(1-3), 135-161.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

Masters, G., & Wright, B. (1997). The partial credit model. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101-122). New York: Springer.

McNiff, J., Lomax, P., & Whitehead, J. (2003). *You and your action research project* (2nd ed.). London: Routledge Falmer.

Merbitz, C., Morris, J., & Grip, J. (1989). Ordinal scales and foundations of misinference. *Archives of Physical Medicine and Rehabilitation, 70*(4), 306-312.

Michell, J. (1994). Numbers as quantitative relations and the traditional theory of measurement. *British Journal of the Philosophy of Science, 45*(2), 389-406.

Michell, J. (2001). Teaching and misteaching measurement in psychology. *Australian Psychologist, 36*(3), 211-218.

Michell, J. (2002). Steven's theory of scales of measurement and its place in modern psychology. *Australian Journal of Psychology, 54*(2), 99-104.

Mills, G. (2007). *Action research: A guide for the teacher researcher*. Upper Saddle River NJ: Merrill Prentice Hall.

Mills, J. (2002). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education, 10*(1). Retrieved from http://www.amstat.org/publications/jse

Mittag, K., & Thompson, B. (2000). A national survey of American Educational Research Association members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher, 29*(4), 14-20.

Mokros, J., & Russell, S. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education, 26*(1), 20-39.

Moore, D. (1990). Uncertainty. In L. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95-137). Washington, DC: National Academy Press.

Moreno, R., & Duran, R. (2004). Do multiple representations need explanations? The role of verbal guidance and individual differences in multimedia mathematics learning. *Journal of Educational Psychology, 96*(3), 492-503.

Morgan, C. (2001). The place of pupil writing in learning, teaching and assessing mathematics. In P. Gates (Ed.), *Issues in Mathematics Teaching* (pp. 232-244). New York: Routledge Falmer.

National Research Council. (2002). *Scientific research in education.* Washington DC: National Academy Press.

Nickerson, R. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241-301.

Noffke, S. (1994). Action research: Towards the next generation. *Educational Action Research, 2*(1), 9-21.

Ormrod, J. (2008). *Educational psychology*. Columbus, OH: Pearson.

Ozgun-Koca, S. (1998, October–November). *Students' use of representations in mathematics education.* Paper presented at the Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, Raleigh, NC.

Perkins, D., & Simmons, R. (1988). Patterns of misunderstanding: An integrative model for science, math and programming. *Review of Educational Research, 58*(3), 303-326.

Pintrich, P., Marx, R., & Boyle, R. (1993). Beyond conceptual change: The role of motivational beliefs and classroom contextual factors in the process of conceptual change. *Review of Educational Research, 63*(2), 167-199.

Pollatsek, S., Lima, S., & Well, A. (1981). Concept or computation: Students' understanding of the mean. *Educational Studies in Mathematics, 12*(2), 191-204.

Popper, K. (1963). *Conjectures and refutations*. London: Routledge and Kegan Paul.

Posner, G., Strike, K., Hewson, P., & Gertzog, W. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education, 66*(2), 211-227.

Pugalee, D. (2001). Writing, mathematics, and metacognition: Looking for connections through students' work in mathematical problem solving. *School Science and Mathematics, 101*(5), 236-245.

Reid, J., & Reading, C. (2005). Developing consideration of variation: Case studies from a tertiary introductory service statistics course. In *Proceedings of the 55th Session of the International Statistical Institute, Sydney, Australia*. [CDRom] Voorburg, The Netherlands: International Statistics Institute.

Rowntree, D. (1981). *Statistics without tears: A primer for non-mathematicians*. Harmondsworth, England: Penguin Science.

Rubin, A., Bruce, B., & Tenney, Y. (1991). Learning about sampling: Trouble at the core of statistics. In D. Vere-Jones (Ed.), *Proceedings of the 3rd International Conference on Teaching Statistics* (Vol. 1, pp. 314-319.). Voorburg, The Netherlands: International Statistical Institute.

Rubin, A., Hammerman, J., & Konold, C. (2006). Exploring informal inference with interactive visualisation software. In A. Rossman & B. Chance (Eds.), *Developing a statistically literate society* (Proceedings of the 7th International Conference on Teaching Statistics, Salvador, Brazil) [CDRom]. Voorburg, The Netherlands: International Statistics Institute.

Shaughnessy, J. (1992). Research in probability and statistics: Reflections and directions. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 465-494). New York: Macmillan.

Shaughnessy, J. (2006). Research on students' understanding of some big ideas in statistics. In Gail Burrill, (Ed.), *Data and Chance. 2006 Yearbook of the National Council of Teachers of Mathematics* (pp. 77-98). Reston, VA: National Council of Teachers of Mathematics.

Shaughnessy, J., & Chance, B. (2005). *Statistical questions from the classroom.* Reston, VA: NCTM.

Shavelson, R., Phillips, D., Towne, L., & Feuer, M. (2003). On the science of education design studies. *Educational Researcher, 32*(1), 25-28.

Snir, J., Smith, C., & Grosslight, L. (1995). Conceptually enhanced simulations: A computer tool for science teaching. In D. Perkins, J. Schwartz, M. West, & M. Wiske (Eds.), *Software goes to school: Teaching for understanding with new technologies* (pp. 106-129). New York: Oxford University Press.

Sotos, A., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2*(2), 98-113.

Steffe, L., Thompson, P., & Von Glaserfield, E. (2001). Teaching experiment methodology: Underlying principles and essential elements. In S. Carver & D. Klahr (Eds.), *Cognition and instruction: Twenty-five years of progress* (pp. 205-226). Mahwah, NJ: Laurence Erlbaum and Associates.

Stevens, S. (1946). On the theory of scales of measurement. *Science, 103* (2684), 677-680.

Strauss, S., & Bichler, E. (1988). The development of children's concept of the arithemetic average. *Journal for Research in Mathematics Education, 19*(1), 64-80.

Strike, K., & Posner, G. (1985). A conceptual change view of learning and understanding. In L. West & A. Pines (Eds.), *Cognitive structure and conceptual change* (pp. 259–266). New York: Academic Press.

Thompson, P., Liu, Y., & Saldanha, L. (2007). The intricacies of statistical inference. In M. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 207-231.). Mahwah NJ: Erlbaum Thomson.

Thurstone, L. (1925). A method of scaling psychological and educational tests. *The Journal of Educational Psychology, 16*(7), 433-451.

Thurstone, L. (1928). Attitudes can be measured. *The American Journal of Sociology, 33*, 529-554.

Thurstone, L. (1937). Psychology as a quantitative rational science. *Science,85,* 227-232.

Tobin, K., Tippins, D., & Gallard, J. (1994). Research on instructional strategies for teaching science. In D. Gabel (Ed.), *Handbook of research on science teaching and learning* (pp. 45-93). New York: Macmillan Publishing Company.

Tversky, A., & Kahneman, D. (1982a). Evidential impact of base rates. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 153-162). New York: Cambrige University Press.

Tversky, A., & Kahneman, D. (1982b). Judgement under uncertainty: Heuristics and biases. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 3-22). New York: Cambridge University Press.

Tversky, A., & Kahneman, D., (1983). Extension versus intuitive understanding: The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293–315.

van der Linden, W., & Hambleton, R. (1997). Item response theory: Brief history, common models, and extensions. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 1-28). New York: Springer.

von Glaserfield, E. (1995). *Radical Contructivism: A way of knowing and learning.* London: The Falmer Press.

Watson, J., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal, 2*(2), 3-46.

Watson, J., & Kelly, B. (2007). The development of conditional probability reasoning. *International Journal of Mathematics Education in Science and Technology, 38*(2), 213-235.

Watson, J., & Kelly, B. (2009). Development of student understanding of outcomes involving two or more dice. *International Journal of Science and Mathematics Education, 7*(1), 25-54.

Watson, J., & Moritz, J. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics, 37*(2), 145-168.

Wieman, C., & Perkins, K., (2005). Transforming physics education. *Physics Today*, *58*(11), 36-41.

Wild, C., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*, 223-248.

Willmott, A., & Fowles, D. (1974). *The objective interpretation of test performance: The Rasch model applied*. Windsor: NFER Publishing Company Ltd.

Wright, B. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16*(4), 33-45.

Wright, B., & Linacre, J. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation, 70*, 857-860.

Yilmaz, M. (1996). The challenge of teaching statistics to non-specialists. *Journal of Statistics Education, 4*(1). Retrieved from http://www.amstat.org/publications/jse.

Zar, J. (1974). *Biostatistical Analysis.* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Appendix A: Details of the Data Handling and Statistics Unit - The traditional teaching program with the additions for the first cycle of the intervention

	Teaching plan – pre-intervention	Additional material and/or alterations for the first cycle of intervention
WEEK 1		
Lectures	Introduction to the unit • Introduction to the discipline of statistics • Why do we use samples? • Classifying data	Introduction to simulation The Chinese birth problem. If the one child policy was replaced with a "keep having children until a boy is born" policy, what would be the effect on the ratio of girls to boys born?
Practical	Introduction to *Microsoft Excel*, including pivot tables and graphing	
Tutorial	• The unusual event problem (A table of data was presented that gives the number of people present and the number who died for each gender and cabin class on the Titanic. Without being told the source of the data, students were required to give suggestions that would explain the pattern of data.) • Introduction to the first project	

	Teaching plan – pre-intervention	Additional material and/or alterations for the first cycle of intervention
WEEK 2		
Lectures	Further introduction to unit and how to obtain data • Displaying data • Measures of central tendency and spread • Sampling • Survey design • Experimental design	Comparison of the mean and median Measures of central tendency and spread were not covered until after the practical session, where the effects of extreme values on the values of the mean and median had been demonstrated The results of the Chinese problem were used as an introduction to hypothetical, probabilistic reasoning. From the sample, would they conclude that the ratio of boys to girls will change?
Practical	Work on project	The effects of extreme values or errors on the mean and median • The students entered a small data set into Excel, calculated the mean, median and standard deviation, and then introduced a very large number into the data.[†] • The button (clustering problem). What is "random"?
Tutorial	The effect of bias on reported statistics An example of a double blind experiment from *The Australian*, and an article from *New Scientist* (*Hype and Herceptin*) showing how the influence of financial and other interests can influence the statistics chosen in reporting results.	Students were asked the following questions Means versus medians • Why does the Real Estate Institute of Tasmania report quarterly median house prices? Randomness • What does "random" mean?

	Teaching plan – pre-intervention	Additional material and/or alterations for the first cycle of intervention
WEEK 3		
Lectures	Introduction to statistical independence • The statistical definition of independence • Independence in a contingency table • Introduction to probability	The use of probability to make decisions Problems that need probability reasoning in their solution The effect of sample size on the outcomes of a Binomial experiment
Practical	Work on project	Variation among samples Students were instructed that they were going to take samples from a population with a given mean and asked to predict the lowest and highest sample means they would expect to get. After the samples were taken they compared their results to their predictions. They also compared the standard deviation of the means to the standard deviation of the original data
Tutorial	Questions about project 1 – due week 4	

	Teaching plan – pre-intervention	Additional material and/or alterations for the first cycle of intervention
WEEK 4		
Lectures	Introduction to probability distributions • Introduction to probability distributions • The Binomial distribution with example • The Poisson distribution with example	Introduction to probability distributions • The Binomial distribution was introduced by students using a probability tree to calculate Binomial probabilities. The resulting pattern was linked to the Binomial formula • The introduction to the Poisson distribution was followed by an example where it was proposed that a change in experimental conditions did not affect the outcome. The Poisson distribution was used to calculate the probability of the outcome assuming the change in conditions made no difference. The students were then required to make a conclusion about the effect of the change in conditions
Practical	Introduction to calculating probability in *Excel*	A real application of the use of the Poisson distribution The V1 rockets problem – this was based on the work of Clarke (1946). The students were asked to compare the frequencies predicted by the Poisson distribution with the observed frequencies and decide whether the data were Poisson distributed. This example was again used to demonstrate Chi-squared goodness-of-fit tests in week 9
Tutorial	Project 2	The consequences of the data being Poisson distributed were discussed

	Teaching plan – pre-intervention	Additional material and/or alterations for the first cycle of intervention
WEEK 5		
Lectures	Continuous probability distributions • The Normal distribution • The Standard Normal distribution • The *t*-distribution – traditional lecture	
Practical	Work on project	Introduction to the Central Limit Theorem via simulation Students simulated populations with Normal, Uniform and Binomial distributions, took sample means and observed the distribution of the sample means
Tutorial	Probability problems using the Normal, Binomial and Poisson distributions	Applications of the Central Limit Theorem If sample means are normally distributed, then sample means belong to a distribution that has the characteristics of any Normal distribution
WEEK 6		
Lectures	The sampling distribution of the mean and confidence intervals	The students were asked the following questions: • What do we mean by the word "sample" in statistics? • Does random sampling guarantee representativeness? • What can we say about the population from a sample?
Practical	Work on assignment	Introduction to hypothesis testing by simulation The survey problem. How likely it is that 19 or less people are found in favour of a proposition in a sample of 50, if the population is evenly split? How confidence intervals work.
Tutorial	Questions on project 2	

	Teaching plan – pre-intervention	Additional material and/or alterations for the first cycle of intervention
WEEK 7		
Lectures	• Practice test 1 • Introduction to hypothesis testing	Introduction to hypothesis testing (no formal hypothesis testing language used) How likely is it that 19 or less people are found in favour of a proposition in a sample of 50, if the population is evenly split? The answer to this question was estimated from the results of the practical and was compared with the probability calculated using the Binomial distribution
Practical	Calculation of 2-sample *t*-tests in *Excel*	Introduction to 2-sample *t*-tests via simulation The Grade 12 heights problem* - How likely is it that two sample means this far apart will occur if the population means are equal?
Tutorial	Consolidation of confidence intervals - The Gnome's visit (see Appendix C for details)	
WEEK 8		
Lectures	Test 1	Introduction to 2-sample *t*-tests (based on the Grade 12 heights problem) How likely is it that two sample means this far apart will occur if the population means are equal? The answer to this question was estimated from the results of the practical and was compared with the probability calculated from the formal calculations. Formal hypothesis testing language was not used

	Teaching plan – pre-intervention	Additional material and/or alterations for the first cycle of intervention
Practical	Work on project	Introduction to chi-squared tests of independence via simulation The haemolytic uraemic syndrome (HUS) example -How likely is it that this number of children treated with antibiotics will get HUS if antibiotic treatment makes no difference? (REF)
WEEK 9		
	Introduction to chi-squared tests for independence and goodness-of-fit	Introduction to chi-squared tests for independence and goodness-of-fit • Test for independence - How likely is it that this number of children treated with antibiotics will get HUS if antibiotic treatment makes no difference? The answer to this question was estimated from the results of the practical and was compared with the probability calculated from the formal calculations • Goodness-of-fit - The V1 rocket problem – are the data Poisson distributed? The formal language for hypothesis testing used in this lecture - the null hypothesis, the alternative hypothesis, cut-off point – level of significance, probability – *P*-value
Practical	Calculation of chi-squared tests in *Excel*	Introduction to hypothesis tests on the line of best fit via simulation Example of fitting line of best fit to a problem where there are errors in the measurement.*
Tutorial	Project questions	Project questions

	Teaching plan – pre-intervention	Additional material and/or alterations for the first cycle of intervention
WEEK 10		
Lectures	Introduction to regression analysis • Simple regression analysis • Multiple regression analysis	Introduction to regression analysis The determination of the significance of the line of best fit was linked to the work in the last practical session.
Practical	Regression analysis in *Excel*	Regression analysis and an introduction to the Analysis of Variance (ANOVA) • Regression analysis in *Excel* • Demonstration – the Analysis of Variance – a visual assessment of whether or not two or more groups may have the same mean.†
Tutorial	Introduction to assignment 4	
WEEK 11		
Lectures	Introduction to ANOVA, and Type I and Type II errors • ANOVA • Type I and Type II errors	ANOVA Students used the work from the practical to estimate which groups had significantly different means before the lecture. These estimations were compared with the results of the formal calculations
Practical	Calculation of ANOVAs in *Excel*	ANOVA in excel
Tutorial	Revision	
WEEKS 12 &13	Revision and Test 2	

Appendix B: The Questionnaires and test questions

B1 The first questionnaire

Please answer questions in either the spaces provided or, when applicable, by circling the most appropriate answer.

Please note that these questions are designed to gather information on students’ understanding of probability as they enter a statistics unit and are of varying difficulty. If you find any of these difficult, this does **not** in any way reflect your ability to complete the Data Handling and Statistics unit successfully.

Student Number:…………………………………………….

Section A:

For questions one to five please circle the answer that is applicable to you.

1	Campus:	Launceston	Hobart			
2	Gender:	Male	Female			
3	Semester:	1	2			
4	Course:	Biomedical Science	Aquaculture	Pharmacy	Medicine	Other
5	Highest level at which you studied mathematics		……………			
6	Was there any statistics in your last mathematics course?		Yes	No		

Section B:

1. You are playing the game snakes and ladders with four other people. This game requires you to get a „6’ on a die before you can start. After four rounds no-one has started. Which of the following statements best matches your conclusion?

a. Since a „6’ hasn’t come up yet, it will come up in the next round.
b. Since the chance of getting a ‘6’ is 1 in 6, the die should have come up with a „6’ four times by now, so something is wrong with it.
c. Throwing a die is a chance event, so it just happens like this sometimes.

d. If a „6’ doesn’t come up soon, there must be something wrong with the die.
e. There must be something wrong with the die.

2. It is estimated that 14% of women will develop breast cancer sometime during their lifetime. What is the best interpretation of this statement?

a. It is very likely that a woman will die of breast cancer.
b. Not many women die from breast cancer, but it is not that uncommon.
c. Not many women get breast cancer, but it is not that uncommon.
d. It is not likely that a woman will die of breast cancer.
e. More women than not will get breast cancer.
f. It is very likely that a woman will get breast cancer.
g. It is not likely that a woman will get breast cancer.
h. More women than not will die of breast cancer.

3. The following message is printed on a bottle of prescription medication:

> **Warning: For application to skin areas there is a 15% chance of developing a rash. If a rash develops, consult your physician.**

Which is the best interpretation of this warning?

a. Don’t use the medication on your skin; there is a good chance of developing a rash.
b. For applications to the skin, apply only 15% of the dose.
c. If a rash develops, it will probably involve only 15% of the skin.
d. About 15 out of 100 people who use this medication develop a rash.
e. There is hardly any chance of getting a rash in using this medication.

Section C.

1. Imagine that you are the captain of your local cricket team. The next season is coming up and you know that it is important that you win the toss so you can choose whether or not you will bat first.

   You decide that you will choose heads for the entire season. What is the chance that the coin will come up tails (and you losing the tosses) 4 times out of 4?

2. Suppose tails did come up 4 times out of 4. For the 5th toss, should you choose
   a. Heads
   b. Tails
   c. Doesn't matter

Please explain your answer.

3. What is the probability of getting heads on this next toss? Explain your answer.

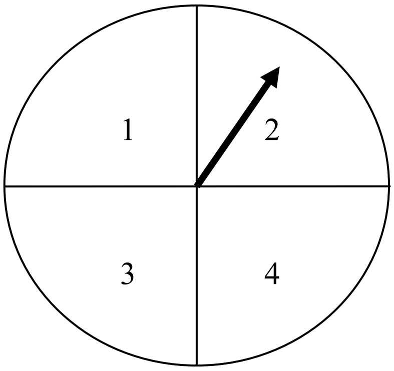
4. What is the probability of getting tails on this next toss? Explain your answer

Section D

1. Half of all newborns are girls and half are boys. Hospital A records an average of 50 births per day. Hospital B records an average of 10 births a day. On a particular day, which hospital is more likely to record 80% or more of female births?
   a. Hospital A (with 50 births a day)
   b. Hospital B (with 10 births a day)
   c. The two hospitals are equally likely to record such an event.

Please explain your answer.

2. A tutorial group used this spinner. If you were to spin it once, what is the chance it will land on an even number?

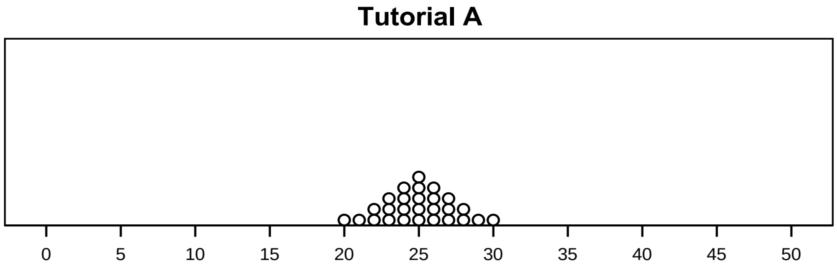


3. Out of the next 50 spins, how many times do you think the spinner will land on an even number? Why do you think this is?

4. If you were to spin it 50 times again, would you expect to get the same number out of 50 to land on an even number? Why do you think this?

5. How many times out of 50 spins, landing on an even number, would surprise you? Why do you think this is?

6. The members of three statistics tutorial groups did 50 spins and graphed the number of times the spinner landed on an even number. Each circle represents one person in the tutorial group. In some cases, the results were just made up without actually doing the experiment.



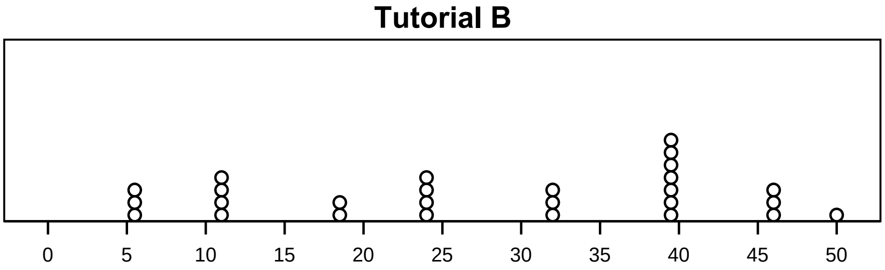
Tutorial A

a. Do you think tutorial A’s results are made up or really from the experiment?

i. Made up

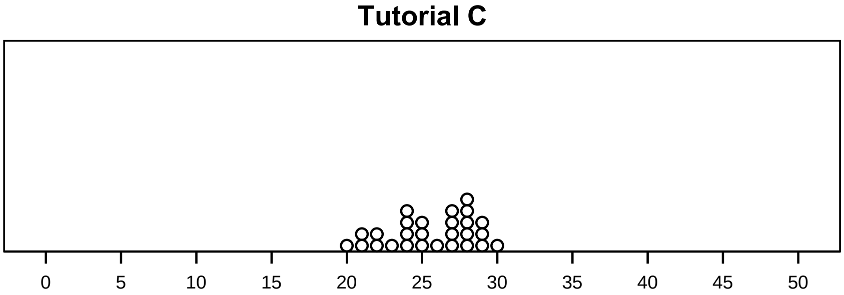
ii. Real from experiment

Explain why you think this.



b. Do you think tutorial B's results are made up or really from the experiment?
   i. Made up
   ii. Real from experiment

Explain why you think this is



c. Do you think tutorial C's results are made up or really from the experiment?
   i. Made up
   ii. Real from experiment

Explain why you think this is

Section E

1. Which probability do you think is bigger?

a. The probability that a woman is a schoolteacher
OR

b. The probability that a schoolteacher is a woman.
c. Both (a) and (b) are equally likely.

Please explain your answer.

2. The table below shows the number of defective TV's produced every week at two factories by the day shifts and by the night shifts.

	Factory A	Factory B
Day	40	30
Night	40	60

a. How many defective TV's are produced at Factory B every week?

b. How many defective TV's are produced by a night shift every week?

c. If you were told that a defective TV was produced at Factory A, what is the probability it was produced by a day shift?

3. An urn has 2 white balls and 2 black balls in it. Two balls are drawn out without replacing the first ball.

a. What is the probability that the second ball is white, given that the first ball was white? Please explain your answer

b. What is the probability that the first ball was white, given that the second ball was white? Please explain your answer.

Section F

1. A new medication is being tested to determine its effectiveness in the treatment of eczema, an inflammatory condition of the skin. Thirty patients with eczema were selected to participate in the study. The patients were randomly divided into two groups. Twenty patients in an experimental group received the medication, while ten patients in a control group received no medication. The results after two months are shown below.

	Experimental Group (Medication)	Control Group (No Medication)
Improved	8	2
No improvement	12	8

Based on this data, you think the medication was:

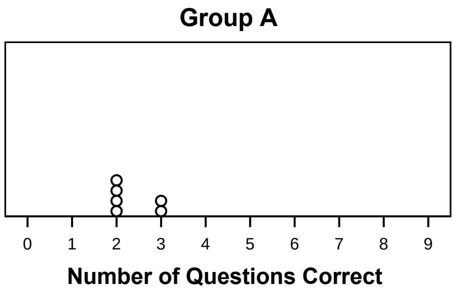
A. Somewhat effective

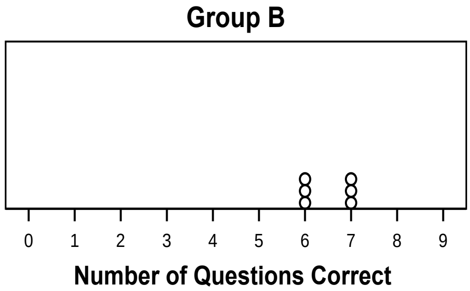
B. Basically ineffective

If you chose option A, select the one explanation below that best describes your reasoning.	If you chose option B, select the one explanation below that best describes your reasoning.
a. 40% of the people (8/20) in the experimental group improved	a. In the control group, 2 people improved even without the medication.
b. 8 people improved in the experimental group while only 2 improved in the control group	b. In the experimental group, more people didn't get better than did (12 vs. 8).
c. In the experimental group, the number of people who improved is only 4 less than the number who didn't improve (12-8) while in the control group the difference is 6 (8-2).	c. the difference between the numbers who improved and didn't improve is about the same in each group (4 vs. 6).
d. 40% of patients in the experimental group improved (8/20), while only 20% improved in the control group (2/10)	d. In the experimental group, only 40% of the patients improved (8/20).

2. A tertiary institution is comparing the scores of some tutorial groups on a test of basic statistics facts. The test had nine questions.

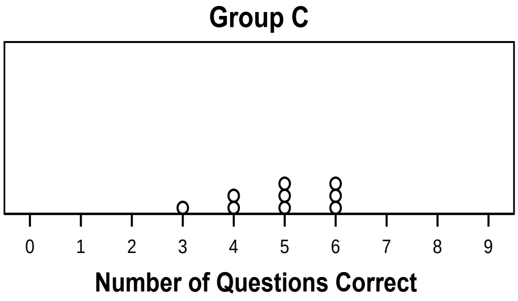
a. The scores for two of these tutorial groups are shown in the charts below. Each circle represents one person. Therefore for Group A four people answered two questions correctly, and two people answered three questions correctly.

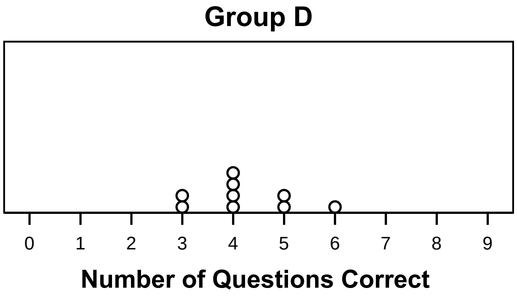




Did the two groups perform equally well, or did one group perform better? Please give reasons for your answer.

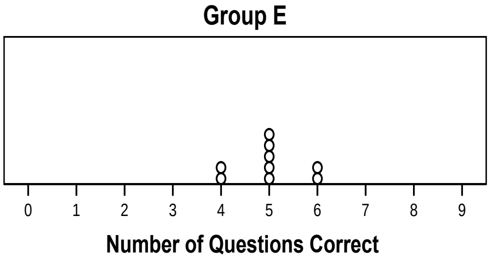
b. Now compare tutorial groups C and D.

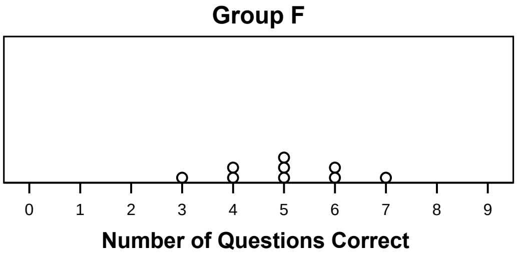




Did the two groups perform equally well, or did one group perform better? Please give reasons for your answer.

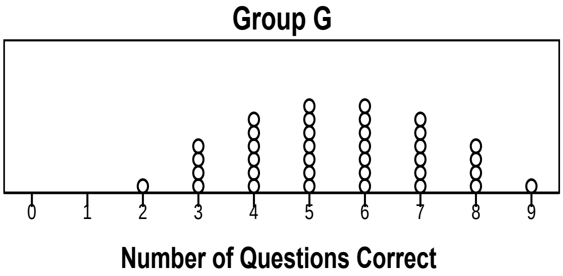
c. Now compare groups E and F

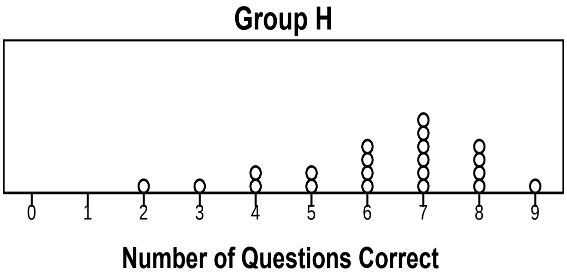




Did the two groups perform equally well, or did one group perform better?
Please give reasons for your answer.

d. Now compare groups G and H





Did the two groups perform equally well, or did one group perform better?
Please give reasons for your answer.

Thank you very much for your time.

B2 The Second Questionnaire

Student number: ……………………………………………..

Please note that this questionnaire is investigating forms of statistical thinking different to that in your formal assessments. How hard or easy you find these questions does NOT reflect your ability to complete successfully the final test.

Question 1.

Give an example of something that happens in a random way. Explain why you think this is an example of randomness.

Question 2.

You work for a manufacturer of circuit breakers. Owing to the difficulty of the process, it is expected that 5% of these will be defective. The occurrence of the defective breakers occurs randomly. The breakers are sold in boxes of 25.
One of your customers buys a box with three defective breakers. This is 12% of the contents of the box. Your customer is furious. You are told that your underlying rate is 12%, not 5% and they will take their custom elsewhere.

a. *If* 5% are defective overall, then *on average* how many defective breakers would you expect to find per box?

b. It can be calculated that *If* the underlying rate is 5%, the probability of getting 3 or more defectives in a box is 13/100. Based on these figures, getting three or more defective circuit breakers in a box is:

   i. Very likely
   ii. Fairly likely
   iii. Possible
   iv. Unlikely
   v. Very unlikely

Continued next page

c. Does this box provide sufficient evidence that the underlying rate of defectives for all the circuit breakers is greater than 5% as the customer claims? Explain your answer.

Question 3.

Half of all newborns are girls and half are boys. Hospital A records an average of 50 births per day. Hospital B records an average of 10 births a day. On a particular day, which hospital is more likely to record 80% or more of female births?

d. Hospital A (with 50 births a day)
e. Hospital B (with 10 births a day)
f. The two hospitals are equally likely to record such an event.

Please explain your answer.

Question 4.

You are looking at the effects of supplementing trout fish feed with vitamin E. Some of the fish are given the standard commercial feed, and others are given the same feed with double the level of vitamin E. After a suitable time, you measure the weights of the fish. The results are in the table below:

	Standard feed	Extra E
Mean weight (g)	256.4	263.1
Standard deviation (g)	12.3	11.2

You perform the two sample t-test and the p-value you receive is 0.45. Therefore you tell your supervisor that the extra vitamin E has not made any difference to the mean weight of the fish.

Your supervisor says that the mean weight of the fish given „extra E' ***is*** higher than the mean weight of those who were given the standard feed. Explain to your supervisor why you say that even though the „extra E' feed has a higher mean weight, the „extra E' feed has not made a *significant* difference.

Question 5.

You are working for a consumer organisation. As part of your duties, you select 49 boxes of "Get up and Go" cereal at *random* and weigh the boxes. On the label of the boxes you read that the minimum weight of the box is 800g.
The standard deviation of the weight of the boxes is 14g. Therefore the standard error of the mean (the standard deviation of all possible sample means) is estimated to be 2g.

a. Assume the manufacturer's claim that the minimum weight of 800g is correct. For your 49 boxes, is a sample mean of 799g

    (i) Very likely
    (ii) Likely
    (iii) Possible
    (iv) Unlikely
    (v) Very unlikely?

    Give reasons for your answer.

b. Again assuming the manufacturer’s claim to be correct, for your 49 boxes, is a sample mean of 796g

    (i) Very likely
    (ii) Likely
    (iii) Possible
    (iv) Unlikely
    (v) Very unlikely?

Give reasons for your answer.

c. At what value of a sample mean below 800g would you start to suspect the manufacturer’s claim to be untrue? Give reasons for your answer.

d. If you didn’t use random sampling, how would this affect your previous answers in this question?

Question 6.

A new medication is being tested to determine its effectiveness in the treatment of eczema, an inflammatory condition of the skin. Thirty patients with eczema were selected to participate in the study. The patients were randomly divided into two groups. Twenty patients in an experimental group received the medication, while ten patients in a control group received no medication. The results after two months are shown below.

	Experimental Group (Medication)	Control Group (No Medication)
Improved	8	2
No improvement	12	8

Based on this data, you think the medication was:

A. Somewhat effective	B. Basically ineffective
If you chose option A, select the one explanation below that best describes your reasoning.	If you chose option B, select the one explanation below that best describes your reasoning.
a. 40% of the people (8/20) in the experimental group improved	a. In the control group, 2 people improved even without the medication.
b. 8 people improved in the experimental group while only 2 improved in the control group	b. In the experimental group, more people didn't get better than did (12 vs. 8).
c. In the experimental group, the number of people who improved is only 4 less than the number who didn't improve (12-8) while in the control group the difference is 6 (8-2).	c. the difference between the numbers who improved and didn't improve is about the same in each group (4 vs. 6).
d. 40% of patients in the experimental group improved (8/20), while only 20% improved in the control group (2/10)	d. In the experimental group, only 40% of the patients improved (8/20).

Continued next page

Listed below are several possible reasons one might question the results of the experiment described above. Please circle the letter for EVERY reason you agree with.

a. It’s not legitimate to compare the two groups because there are different numbers of patients in each group.
b. The sample of 30 is too small to permit drawing conclusions.
c. The patients should not have been randomly put into groups, because the most severe cases my have just by chance ended up in one of the groups
d. I’m not given enough information about how doctors decided whether or not the patients improved. Doctors may have been biased in their judgements.
e. I don’t agree with any of these statements.

Question 7

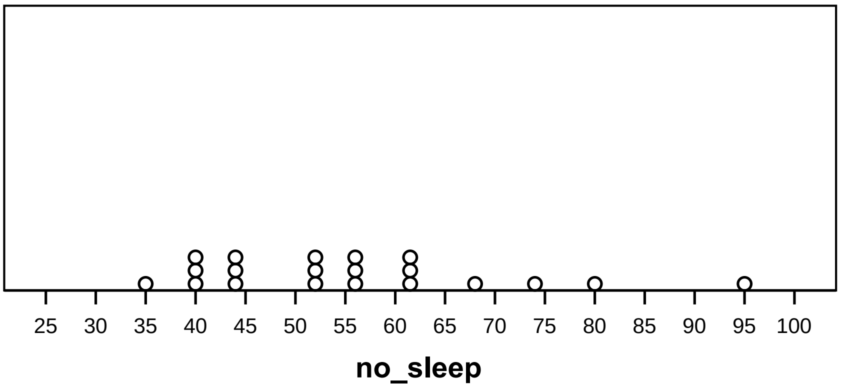
An urn has 2 white balls and 2 black balls in it. Two balls are drawn out without replacing the first ball.

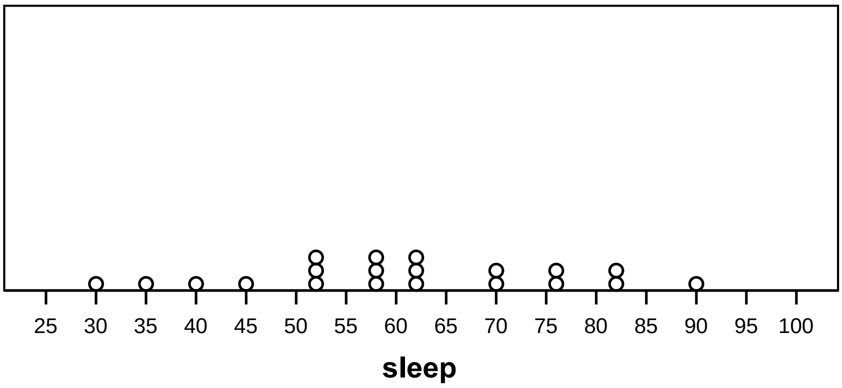
c. What is the probability that the second ball is white, given that the first ball was white? Please explain your answer

d. What is the probability that the first ball was white, given that the second ball was white? Please explain your answer.

Question 8

Forty students from the Pacific University participated in a study of the effect of sleep on test scores. Using random allocation, twenty of the students were required to stay up all night studying the night before the test (no-sleep group). The other 20 students (the control „slœp' group) were required to be in bed by 11:00pm on the evening before the test. The test scores for each group are shown in the graphs below. Each dot on the graph represents a particular student's score. For example, the 3 dots above the 40 in the top graph indicated that 3 students in the no-sleep group scored 40 on the test.





A. Examine the two graphs carefully. Then circle the conclusion from the 6 possible conclusions listed below the ONE you MOST agree with.

a. The no-sleep group did better because none of these students scored below 35 and the highest score was achieved by a student in this group.
b. The no-sleep group did better because its average appears to be a little higher than the average of the sleep group.
c. There is no difference between the two groups because there is considerable overlap in the scores of the two groups.
d. There is no difference between the two groups because the difference between their averages is small compared to the amount of variation in the scores.
e. The sleep group did better because more students in this group scored 75 and above.

f. The sleep group did better because its average appears to be a little higher than the average of the no-sleep group.

B. Atlantic University repeated the same study but allowed the students to *choose* which of the groups (sleep or no-sleep) they could go into. The Pacific University claims that allowing the students to choose could bias the results. Atlantic University claims that this does not matter. Which University do you think is correct? Give reasons for your answer.

Question 9

Fred is a plant geneticist and sent the results of his research to a scientific journal but his paper was rejected. He has sought help from a statistician. Here is their conversation:

Fred: I've just had my paper containing some important results rejected because I didn't use random allocation of my treatments. Now I have to repeat the whole experiment!

Statistician: Tell me what you did.

Fred: I had a bench with eight pots sitting next to each other along the bench. In the first four pots I put my new wonder species and in the next four pots I put the standard species. As I expected, my wonder species produced much higher growth.

Statistician: OK. Of course, there may have been some other factor varying along the bench which is responsible for the difference.

Fred: I'm not that stupid! The temperature, light and everything else is controlled in this glasshouse. If I thought there was another effect, I would have allocated the treatments to take account of the fact.

Statistician: In that case, our task is simple. We will produce an allocation plan by generating random numbers in the computer.

They do this, and find that the wonder variety is allocated to the first four pots, and the standard variety to the other four pots, just as before.

Fred: Great! You have just proved my results were valid because they were obtained under the layout recommended by random allocation.

Statistician: No! The editor rejected the WAY you obtained the layout, not the layout itself.

Fred: !!!!!

Can you explain to Fred why the randomisation was so important? See if you can provide an argument for randomisation that will overcome Fred's problem.

Thank you for your time

B3 The Test Questions used in this study

1. A *P*-value of 0.98 indicates that the null hypothesis is almost certainly true. Is this statement correct? Give reasons for your answer.

2. In the test of a null hypothesis that a new drug produces the same expected benefit as the standard drug, versus the alternative hypothesis that the new drug produces a higher expected benefit than does the standard drug, a p-value of 0.01 is obtained. Explain what this result means to a patient who has read the result on the web but has no statistical training**.**

3. A large number of Tasmanian residents were asked for their estimate of the number of times they visited a GP in 2009. From the data, 95% confidence limits were calculated for the mean number of visits by Tasmanians to a GP during 2009. The confidence interval was reported as 7 to 11.

   (a) In completely non-technical words, explain what this reported statement means.

   (b) What does the "95%" refer to?

Appendix C: The coding protocols

C1 Coding protocol for the first questionnaire

Question	Score	Explanation
B1 “Snakes”	1	c
	0	All others
B2 “Cancer”	2	c
	1	g, b
	0	NR, all others
B3 “Eczema”	2	d
	1	e, a
	0	b, c, NR, multiple selections
C1 “Coin 1”	2	1 in 16 or similar
	1	50%, 50-50 every time the coin is tossed
	0	Other number, NR
C2	2	=, ½, each value ½ with reasoning that implies independence
“Coin 2”	1	=, ½, each value ½ with reasoning that does not imply independence, for example “there are only two outcomes”
	0	NR, a, b, illogical reasoning
C3, C4	2	=, ½, each value ½ with reasoning that implies independence
“Coin 3-4”	1	=, ½, each value ½ with reasoning that does not imply independence, for example “there are only two outcomes”
	0	NR, a, b, illogical reasoning
D1 “Hospital”	3	b- reasoning suggests that probability for Hospital B is higher
	2	b- “easier” or similar for Hospital B
	1	c – uses argument of independence of individual births
	0	Hospital A, NR, illogical reasoning
D2 “Spinner 1”	1	50%, ½, 50-50, half
	0	NR, other
D3 “Spinner 2”	2	Possibility of variation in outcomes indicated
	1	Strict probability used
	0	NR, not reasonable number
D4 “Spinner 3”	2	Recognition of the possibility of variation
	1	Anything can happen
	0	Yes, NR

Coding protocol for the first questionnaire (continued)

Question	Score	Explanation
D5 “Spinner 4”	2	>35, <15 or similar
	1	>35 or similar only, larger extremes
	0	0 and/or 50 only, other single number, NR
D6a “Tute A”	2	Made up – too perfect/symmetrical/neat
	1	Made up – other reasoning
	0	Real, NR
D6b “Tute B”	2	Made up with reasoning that includes two or more of: • Too many results on two few numbers • Would not get 0 and/or 50 evens • No 25s • Not clustered around 25 • Range is too wide
	1	Made up – one of above reasons
	0	No reason for choice, NR, Real
D6c “Tute C”	2	Real with reasoning that includes two or more of: • Grouped around 25 • Some variation present/evidence of randomness • Range is reasonable
	1	Real – with one of above reasons
	0	Made up, NR
E1 “Teacher”	2	b, Reasoning which takes all women’s occupations, proportion of teachers into account
	1	b, personal experience that large numbers of teachers are women
	0	a, NR, c same question
E2 “Factory”	1	1 in 2/50%/ ½
	0	Other answer
E3a “Urn A”	2	1/3 with reasoning explained
	1	1/3 without reasoning explained/ joint probability calculated
	0	NR, other answer
E3b “Urn B”	2	1/3 with reasoning explained
	1	1/3 without reasoning explained, joint probability calculated
	0	NR, other answer
F1 “Med”	2	Ad – Uses proportional reasoning comparing one group to the other
	1	Aa, Ab, Bd – Uses the results from one group without mentioning results of other group/uses raw scores for both groups
	0	All others, no response

Coding protocol for the first questionnaire (continued)

Question	Score	Explanation
F2 “A-B”	3	Group B – All of Group B has a higher score than all of Group A, estimates means without full calculations, uses totals and explains that numbers in each group are equal
	2	Group B – uses Group B’s results without reference to Group A, fully calculates means, uses totals without stating that group sizes are equal
	1	Group B – “more” correct or similar
	0	Group A, NR
F2 “C-D”	3	Group C – calculates/estimates mean, uses totals and explains that numbers in each group are equal, calculates frequency of each score
	2	Group C – uses Group C’s results without reference to group D, uses totals without stating that group sizes are equal
	1	Group C – “more” correct or similar
F2 “E-F”	3	Equal – mean or median used, uses totals and explains that numbers in each group are equal
	2	Equal – uses totals without stating that group sizes are equal
	1	Equal – no explanation
	0	One group performed better than the other
F2 “G-H”	3	Group H – mean, median used/proportional reasoning
	2	Group H – uses results of one group only
	1	Group H – “more” correct or similar
	0	Cannot be done, not fair, NR

C2 Coding protocol for the second questionnaire

Question	Coding	Explanation
1	2	Coin, Die or similar with appropriate explanation
"Random"	1	Coin, Die or similar, no explanation
	0	NR, Can't think of anything, Car breakdown or similar.
2a	2	1.25
"CB 1"	1	1
	0	NR, other answers
2b	2	(ii)
"CB 2"	1	(i), (iii)
	0	(iv), (v), NR
2c	2	No, used probability of 13% in reasoning
"CB 3"	1	Variation exists, other boxes may have less/Used formal hypothesis test with no further explanation
	0	NR, Yes
3 "Hospital"	3	b- reasoning suggests that probability for Hospital B is higher
	2	b- "easier" or similar for Hospital B
	1	c – uses argument of independence of individual births
	0	Hospital A, NR, illogical reasoning
4	2	Explains *P*-value in reasoning
"Fish"	1	Difference comes about by normal variation between samples
	0	NR, other answer
5a	3	(i) - Appropriate use of standard errors in answer
"Cereal A"	2	(ii), (iii) - Appropriate use of standard errors in answer
	1	Appropriate use of standard error in answer but students states that standard deviation was used
	0	NR, other answer
5b	3	(ii) - Appropriate use of standard errors in answer
"Cereal B"	2	(iii) - Appropriate use of standard errors in answer
	1	Appropriate use of standard errors in answer but student states that standard deviation was used
	0	NR, other answer
5c	2	Three or more standard errors below 800 g
"Cereal C"	1	Two or more standard errors below 800 g, 794 g instead of <794 g, used standard errors but student states that standard deviation was used
	0	NR, other answer
5d	2	All possible samples need to be equally likely
"Cereal D"	1	Bias introduced/conclusions invalid
	0	NR, Other

Coding for second questionnaire (Continued)

Question	Coding	Explanation
6a “Med A”	2	Ad – Uses proportional reasoning comparing one group to the other
	1	Aa, Ab, Bd – Uses the results from one group without mentioning results of other group
	0	All others, no response
6b “Med B”	2	e
	1	d
	0	NR, any other answer, or combination of answers
7a “Urn A”	2	1/3 with reasoning explained
	1	1/3 without reasoning explained, joint probability calculated
	0	NR, other answer
7b “Urn B”	2	1/3 with reasoning explained
	1	1/3 without reasoning explained, joint probability calculated
	0	NR, other answer
8a “Pacific A”	2	d – uses both mean and standard deviation
	1	c – uses standard deviation only
	0	NR, others
8b “Pacific B”	2	Pacific university - all possible allocations of treatments should be equally likely
	1	Pacific University – bias may be introduced
	0	NR, Atlantic
9 “Fred”	2	All possible combinations must be equally likely
	1	Bias may be introduced/random allocation must be used but no explanation given
	0	NR, other answers

C3 Coding protocols for the test items

Question	Coding	Explanation
1	3	False – P-value is probability of sample value or more extreme if H_o true, so sample likely if H_o true
	2	False – Can only find evidence against H_o/True situation could be a value close to H_o
	1	False - cannot prove true or untrue in inferential statistics
	0	True
2	3	If it were true that the new drug has the same expected benefit as the standard drug, then the probability of the results shown by the new drug is only 1%.
	2	Hypothesis test performed with no further explanation
	1	New drug will do "better" or similar
	0	Inappropriate use of *P*-value
3a	2	Average number of visits to a doctor was between 7 and 11
	1	Average number of visits to a doctor was between 7 and 11, 95% of the time/on average Tasmanian visited a doctor between 7 and 11 times
	0	7 to 11 Tasmanians visit a doctor, other answer
3b	2	Process used is correct 95% of the time/Chance parameter will be included is 95%
	1	95% of sample means will be within two standard errors of the population mean
	0	True for 95% of population, 95% of time population mean is between 7 and 11, other answer

Appendix D: The simulations and demonstrations

D1 Introduction

This appendix contains the instructions for the simulations and demonstrations as they were given to the students. At the time of writing the students had access to *Microsoft Excel 95*. The instructions need altering if they are to be used in *Microsoft 97* or a later version.

D2 Introduction to simulation – the Chinese birth problem

The aims of this exercise were:

- To introduce students to the process of simulation.
- To introduce students to the idea that samples will not replicate exactly the populations from whence they came.

Student instructions

Using simulation, an experiment or process in real life can be mimicked to see what would happen if the conditions were changed. In statistics, simulation is useful because we can repeat a process many times that could not be repeated in reality.

Example: The Republic of China has had a one child policy for many years. In rural areas this has caused distress to those who want a boy to carry on a farm. What would happen to the ratio of male to female births if the policy was changed so that a couple could keep having children until a baby boy was born? (Assume that no prenatal testing takes place)

Choose one of the options below:

- There would be an increase in the number of boys born compared to girls born.
- There would be an increase in the number of girls born compared to boys born.
- The ratio of boys born to girls born would remain equal.

This is not a question that could be answered easily in reality. However, with a coin it will take just a few minutes. For this simulation a „head' will represent a boy, and a „tail' will represent a girl.

Now do 10 trials. In each trial toss a coin and keep tossing until a head is obtained. This will represent the number of children born until a boy is born. Each trial will represent the births for one family, therefore you will be trialling the births for a total of 10 families. Record your answers in the table below.

Trial number	H or T	B or G
Example (do not include this in your final tally)	TTH	GGB
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
Total number of children born:		
Number of boys:		

Now, answer the question again.

What would happen to the ratio of male to female births if the policy was changed so that a couple could keep having children until a baby boy was born?

Choose one of the options below:

- There would be an increase in the number of boys born compared to girls born.
- There would be an increase in the number of girls born compared to boys born.
- The ratio of boys born to girls born would remain equal.

Is this answer different to your first answer? If so, what was incorrect with your reasoning? What is the correct explanation?

D3 Demonstration – Means vs. Medians

The aims of this exercise were:

- To introduce students to the formulae for calculating summary statistics in *Microsoft Excel*.
- For students to observe the effects of an error in the data on the mean and median of this data.

Student instructions

1. Open up the excel file butterfly.xls. This contains the data for the butterfly wing lengths for 20 specimens.
2. Go to cell A17 and type *mean*=. In cell B17 type =average(B2:B16). Remember you can highlight the data in the brackets instead of typing in the cell range. Press enter. Write the answer here:___________
3. Go to cell A18 and type median=. In cell B18 type =median(B2:B16). Press enter. Write the answer here:____________
4. Go to cell A19 and type std.dev.=. In cell B19 type =stdev(B2:B16). Press enter. Write the answer here: ________
5. Now change one of the numbers to a very high value, say 1000. Before you do this, answer the following question:
    a. After the value is changed, the new mean will be:
       Higher Lower The same

    b. The new median will be:
       Higher Lower The same

    c. The standard deviation (a measure of the spread of the data will be:
       Larger Smaller The same

6. Write down the new values:
    a. The new mean is: _________
    b. The new median is: ___________
    c. The new standard deviation is: _________

7. Experiment with a few more changes, then circle the correct answer in the following statements:
    a. If there is an error in the data, the mean/median is the LEAST affected by this error.
    b. If there is an extreme value (an outlier) in the data, the mean/median is LEAST affected by this extreme value.

8. The Real Estate Institute of Tasmania reports quarterly median house prices. This is because:

____

____

____

D4 Simulation – What is “random”?

The aim of this exercise was to stimulate students’ thinking of the meaning of randomness in the statistical sense. The exercise was in the “predict, test, re-evaluate” format (see Section 5.4).

Student instructions

Section A

One problem that occurs in epidemiology is how deal with the situation when several people in a small area come down with the same disease. Is there a common cause for this event, or is this a chance event? For example, if five children in a town of 3,000 people get the same form of leukaemia, does this mean there is something wrong in the town’s environment, or is this just a chance happening?

You might remember that in 2006 the ABC studios in Brisbane were relocated at great expense after there were several cases of breast cancer in the women who worked there. A Statistical analysis concluded that this was a cluster, and not a chance happening, even though the cause could not be identified.

(For more information go to www.abc.net.au/rn/healthreport and go to the transcript for February 5th, 2007).

Here is a 5 X 5 grid. You have 50 buttons to distribute. If the buttons are distributed at RANDOM, how many buttons do you expect to see in each cell? As you think about this consider the following questions:

1. Do you expect that a random process will give a perfectly even spread, that is, exactly two buttons per cell?
2. If you answered no to question 1, how much variation is allowable before you will say the pattern is not random, but clustered?
3. Look at grid A which gives the number of buttons in each cell. Were the buttons distributed randomly, or are they clustered? Why do you think this?
4. Look at grid A which gives the number of buttons in each cell. Were the buttons distributed randomly, or are they clustered? Why do you think this

Grid A

3	0	3	4	1
3	3	3	3	5
1	1	3	0	3
2	1	2	1	2
0	2	2	1	1

5. Now look at Grid B. Were the buttons distributed randomly, or are they clustered? Why do you think this?

Grid B

2	1	3	0	3
1	3	1	2	1
5	3	1	2	1
2	2	2	1	2
3	1	1	6	1

Section B

Now you are going to simulate this for yourself.

1. Go to an *Excel* worksheet and in cell A1 type in *X* and in cell B1 type in *Y*.
2. In cell A2 type *=ROUNDUP(5*RAND(),0)*. Be careful to match up all the parentheses. Press enter. This will give you a number between one and five, produced by a random process. You will notice that the random

number generator recalculates every time you press enter, do not worry about this.

3. Copy across to cell B2 and then down until row 51. You now have 50 pairs of numbers.
4. Highlight the two columns and then make a pivot table; put X in the rows section, and Y in the columns section. Pick up Y or X and place in the table to give the counts. Make sure your data are in the form of **counts**, and not sums.
5. You have now simulated distributing 50 buttons into a 5 X 5 grid. Look at this pattern. Is it more (less) clustered than you expected? If you have time, do another pivot table and see the results.
6. Now go back to your answers for questions 3 and 4 in Section A. Do you agree with your answers from before? Give reasons.

Section C.

Now consider, what makes something random? This will be discussed in your tutorial and lectures. You will also be given the answers to questions 3 and 4.

D5 Demonstration – the V1 rocket problem

The aims of this exercise were

- To give students a real example of a Poisson process where the assumption of independence was extremely important.
- To introduce students to a goodness-of-fit problem in an informal manner.

Student instructions (most of the following is directly quoted from Clarke (1946)

A Real and Interesting Problem for You to Work On*

An Application of the Poisson Distribution

By R.D.Clarke, F.I.A.

Of the Prudential Assurance Company, Ltd.

Readers of Lidstone's *Notes on the Poisson frequency distribution (J.I.A.* Vol LXXI, p.284) may be interested in an application of this distribution which I recently had occasion to make in the course of a practical investigation.

During the flying-bomb attack on London, frequent assertions were made that the points of impact of the bombs tended to be grouped in clusters. It was accordingly decided to apply a statistical test to discover whether any support could be found for this allegation.

An area was selected comprising 144 square kilometres of south London over which the basic probability function of the distribution was very nearly constant, i.e. the theoretical mean density was not subject to material variation anywhere within the area examined. The selected area was divided into 576 squares of ¼ square kilometre each, and a count was made of the numbers of squares containing 0,1,2,3,…,etc. flying bombs. Over the period considered the total number of bombs within the area involved was 537. The expected number of squares corresponding to the actual numbers yielded by the count were then calculated by the Poisson formula.

The result provided an example of **conformity/non conformity** to the Poisson law.

The actual results were as follows - complete the table:

No. of flying bombs per square	Probability of no. of bombs per square (Poisson)	Expected no. of squares (Poisson)	Actual no. of squares
0			229
1			211
2			93
3			35
4			7
5 and over			1
			576

The closeness of fit which in fact appears lends **support/no support** to the clustering hypothesis.

*From the Journal of the Institute of Actuaries, Vol. 72 (1946), p.481.

One of the assumptions of the Poisson distribution is that each unit of time/volume/space is independent from the others. Because these data fitted a Poisson distribution, it was concluded that these rockets were falling randomly.

D6 Simulation – the sampling distribution of the mean

The aim of this exercise was for students to discover that if the sample size is large enough, the means of these samples means form a Normal distribution. .

The exercise was in the "predict, test, re-evaluate" format (see Section 5.4).

Student instructions

The Distribution of Sample Means

Today we are going to examine the "sampling distribution of the mean". What is meant by this?

By the distribution, we mean the pattern into which the data fall. Figure 1 shows the histogram of the heights of 200 grade 5 children. You can see that the shape of this histogram is very close to that of a Normal distribution. The data are left skewed, that it, there are a small number of children who are shorter than the rest. The result is that the shape is not exactly symmetrical.

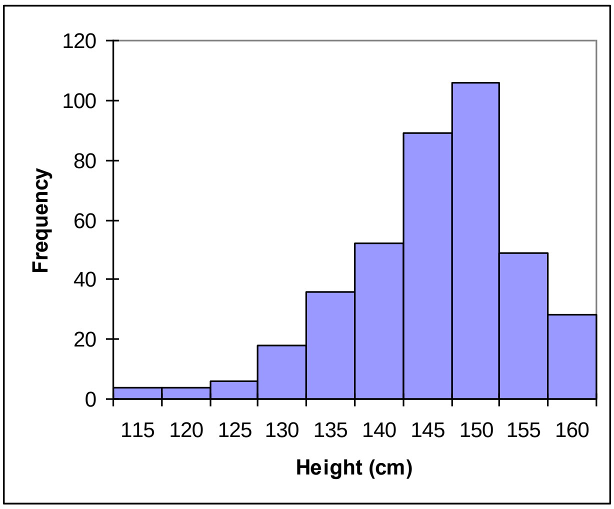


Figure 1. Distribution of heights of 200 grade 5 children.

Now, if you were to take a large number of samples from these data, and then calculate the sample mean for all of these samples, would a histogram of these means have the same shape as the original data? Complete the following sentence:

I think that the histogram of the means will have the ***same shape/different shape*** to that of the original data.

It would be possible, but very time consuming, to take a large number of samples from these data and then calculate and graph the means. So we will go another route using simulation.

Scenario 1 – the original population forms a Normal distribution

1. A random observation from a Normal population with mean μ = 100 and standard deviation σ = 20 can be produced in Excel by:

   =NORMINV(RAND(),100,20)

   Type this in cell A2 of a new sheet. Copy this across for 5 columns (A2 to E2). These numbers consist of a random sample of size 5 from the Normal distribution.
2. In cell F1 type in the word MEAN, and in cell F2 calculate the mean of this sample.
3. Now, highlight the cells from A2 to F2 and copy these cells down to row 1001. Now you have 1000 samples of size 5 from a normally distributed population which has a mean of 100 and a standard deviation of 20. In column F the means of all these samples has been calculated. (It might take a moment or two for Excel to do the calculations, wait!)
4. What shape will the histogram of the means have? To answer this question, in cells H2:H14 type the numbers 70, 75, 80 … up to 130. From here on you must follow these directions VERY CAREFULLY! In particular, do NOT press enter until you have fully read the instructions. Highlight all the cells from I2:I14. Type:

   =FREQUENCY(F2:F1001,H2:H14)

   Check that you have not made a mistake and then ***PRESS CTRL, SHIFT and ENTER.*** This is an array formula and excel will add braces to the outside of the formula.
5. Now highlight the cells from H2 to I14 and use the chart tool to draw a column chart. What shape does histogram have? Sketch it here.
6. You can press F9 and see what happens each time you draw ANOTHER 1000 samples of size 5.
7. What is your conclusion? When a population has a Normal distribution, the sample means form ________________________________.

8. Calculate the mean of all the means, and the standard deviation of the means.
   Mean = ________________, Standard deviation = ___________________

THIS WORK TAKES UP CONSIDERABLE MEMORY. TO STOP THE EXERCISES GETTING TOO SLOW, CLEAR THE SHEET BEFORE DOING THE NEXT EXERCISE.

Scenario 2. The Population Has A Uniform Distribution

A Uniform distribution is one where every number has an equal chance of being selected. Therefore the random number generator in excel produces a Uniform distribution because every number between 0 and 1 has the same chance of being selected.

In the previous section, we saw that if the original POPULATION is normally distributed, the SAMPLE means form a _______________ distribution.

The population in this exercise is uniformly distributed; the histogram of a Uniform distribution is in Figure 2. Note that every bar is roughly the same height.

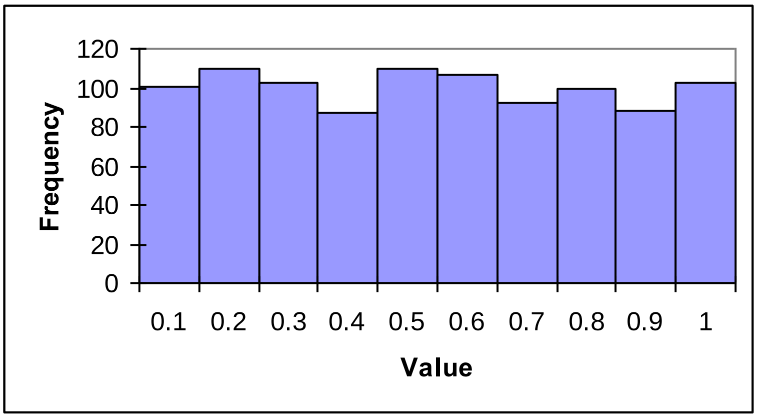


Figure 2. Example of a Uniform distribution.

What shape do you think the histogram of the SAMPLE MEANS will have?

____

____

Now to find out.

1. Go to a new sheet in Excel. In cell A2 type =rand(). Copy this formula across to cell E2. These five numbers represent a random sample of size five from the uniformly distributed population.
2. Go to cell F1 and type in the word MEAN. Go to cell F2 and calculate the average of the five cells.
3. Highlight all the cells from A2 to F2 and copy these down to row 1001. You now have 1000 samples of size 5 from a population which has a Uniform distribution. The averages of all these samples are in column F.
4. In cells H2 to H11 type in the numbers 0.1, 0.2, 0.3 … up to 1. Highlight the cells I2 to I11 and use the frequency formula as before, that is, type:

   =FREQUENCY(F2:F1001,H2:H11)

   And then press CTRL, SHIFT and ENTER.

5. Highlight the cells H2 to I11 and produce a column chart. What distribution does this histogram have? _________________ Are you surprised? Sketch it here.

Scenario 3 – A Binomial distribution

In this section you will use the random numbers to construct samples from a Binomial distribution where the probability of a success is 0.75. To do this, all the random numbers that are less than 0.75 will be labelled „1', and all those with random numbers 0.75 or less will be labelled '0'.

1. Go to a new sheet in Excel. In cell A2 type =rand(). Copy this formula across to cell E2. Then, making sure these five cells are highlighted, copy these cells down to row 1001.
2. Now go to cell G2 and type: =IF(A2<0.75,1,0). Press enter and then copy this formula across to cell K2.
3. In cell L1 type: mean. In cell L2 type: =average(G2:K2). Press enter. Then highlight the cells from G2 to L2 and copy them down to row 1001. The L column represents the mean of 1000 samples of size 5, drawn from Binomial distribution where the probability of a success is 0.75.
4. What pattern will the histogram of the averages have? To find out, draw a histogram. To do this go to cell N2 and type in 0.1, 0.2, … up to 1 in cell N11.

Then go to Tools, Data Analysis and use the histogram tool. The L column goes into the Data Range box (leave out the label), and the cells in the N column go into the bin range. Don't forget to select chart output.

5. What does the pattern of the means look like? Sketch it here.

6. Now repeat steps one to five, but this time you will select samples of size 25. Therefore the random numbers will go in columns A to Y, and the 1's and 0's will go in columns AA to AY, and the means will go into column AZ. Draw the histogram. Make the bin from 0 to 1, in steps of 0.1. What pattern does the histogram of the means have? Sketch it here.

Now go back to the original problem of the heights of children in grade 5. What distribution will the sample means have from this population?

____

____

____

NOW FOR THE OVERALL CONCLUSION.

FOR ANY POPULATION:

IF THE SAMPLE SIZE IS LARGE ENOUGH, THE SAMPLE MEANS FORM A ______________ DISTRIBUTION. FOR THIS TO BE TRUE, THE ORIGINAL POPULATION ***MUST BE/DOES NOT HAVE TO BE*** A NORMAL DISTRIBTUTION.

(THIS IS TRUE EXCEPT FOR SOME RARE EXTREMELY SKEWED POPULATIONS.)

D7 Simulation – hypothesis testing

The purpose of this simulation was to introduce students to hypothesis testing in an informal and visual way. The assumption in this exercise is that the population is evenly divided between those who are and are not in favour of a local council proposition to keep koalas as pets. If a formal hypothesis test was to be carried out this assumption would constitute the null hypothesis. This simulation also gives students a practical demonstration of the sampling variation that can result from taking repeated samples from the same population.

This simulation is based on an article by Erickson (2006). The exercise was in the "predict, test, re-evaluate" format (see Section 5.4).

Student instructions

A survey

There is a proposition on the ballot in the upcoming council election which, if passed, will make it legal to keep koalas as pets. As president of the local Koala Foundation, you hope that this proposition will fail. In a poll of 50 voters, only 19 (38%) will say they will vote yes.

You are now very happy. Since only 38% of your sample vote yes, you are sure the proposition will fail. However, your 50 voters are only a ***sample from the total population*** of voters. If you should survey another 50 voters, do you expect to get exactly 19 in favour again? Why or why not?

Is this proportion of 38% likely to be true for the ***whole*** population of voters?

50% is the threshold, the borderline between success and failure. Let us imagine, then, that the population as a whole is evenly split, 50% for and 50% against.

Now ask, if the situation for the ***whole population*** is ***50% in favour*** of the proposition, how likely is it that we get a ***sample result*** as low as 38%? Or, if 50% of **all** the voters are in favour of the proposition, how likely is it that we will get a **sample** of 19 voters out of 50 in favour? Let us use the computer to answer this question.

Step 1: Go to an *Excel* workbook. In cell A1, type: =rand().

(The random function gives a number where all the numbers between 0 and 1 are equally likely. You will notice that each time you do something in the worksheet these numbers change. Don't worry about this.)

Copy this cell across to column AX, and then keeping the top row highlighted, copy down to row 500. You should now have 500 rows of 50 numbers each.

Step 2: Now go to cell A503 and type: =if(A1>0.5,1,0). Again copy this cell across to column AX and then down to row 1002.

Each row represents a sample of 50 people. Each cell represents a single person. Owing to the way it is set up a „1' represents a person who said „yes, they are in favour', and a „0' represents a person who said „no', they are not in favour.

Step 3: How many of your samples of 50 (represented by each row) have 19 or less people in favour? To find out, go to cell AY503 and type: =sum(A503:AX503). Press enter and copy down for the 500 rows.

Have a quick look at the range of values you have. What is the lowest number out of 50 people in favour of the proposition (out of your 500 simulations)? (Don't spend too much time on this)_______________

What is the highest number out of 50 people in favour (out of your 500 simulations)? __________________

Now for the tricky bit. Make sure you follow the instructions EXACTLY.

When you get to step 4b, do NOT press enter without reading to the end!

Step 4a: Go to cell BA503, and type: 12. Continue down the column with the numbers 13, 14, etc until you get to 40. The numbers should be in cells BA503 to BA531.

Step 4b: Highlight the cells BB503 to BB531 and type:
=frequency(AY503:AY1002,BA503:BA531) Do NOT press enter!

Check that you have typed in the formula correctly, and then press control, shift and enter at the same time.

Step 5: Highlight the cells BA503 to BB531 and draw a column graph. What is the shape of this graph?

How many times, out of 500 samples, do you get a sample where there are 19 people or less in favour of the koala proposition? To answer this, go to step 6.

Step 6: Go to cell BA533 and type: 19 or less=
In cell BB533 type: =SUM(BB503:BB510). This gives you the number of samples with 19 or less people in favour.

Step 7: To make this number a proportion, in cell BA534 type: proportion=
In cell BB534 type: =BB533/500. You can make this into a percentage in the cell below (*100) if you wish.

Now, in how many samples out of 500 are 19 people or less in favour of the proposition, if the population is evenly split? What proportion of the total number of samples is this? Remember you can press F9 and automatically repeat the whole process with a new set of 500 samples.

Conclusion: Therefore, IF the POPULATION is evenly split, getting a SAMPLE with 19 people or less out of 50 in favour is:

a. Almost impossible
b. Unlikely but still possible
c. Very likely
d. Almost certain

From your sample of 19 people out of 50 in favour, could you conclude that the proposition might still pass when ALL of the ratepayers vote?

Conclusion: Therefore, IF the POPULATION is evenly split, getting a SAMPLE with 19 people or less out of 50 in favour is:

a) Almost impossible
b) Unlikely but still possible
c) Very likely
d) Almost certain

From your sample of 19 people out of 50 in favour, could you conclude that the proposition might still pass when ALL of the ratepayers vote?

At the next lecture the hypothesis test was set up formally, but without the alternative hypothesis. The *P*-value was calculated using the Binomial distribution:

$P(x \leq 19|n = 50, p = 0.5) = 0.0595$.

The formula in excel, with which the students were familiar is:

=BINOMDIST(19,50,0.5,1)

D8 Demonstration/simulation – how confidence intervals work

The aims of this demonstration were:

- For students to see that some confidence intervals include the value of the population mean, while others do not.
- For students to see that the proportion of the confidence intervals that include the value of the population mean depends on the number of standard errors used in the calculation.

Student Instructions

Aim: To use computer simulation to demonstrate how confidence intervals work.

1. Open up *Excel*. In cell A2 type „=12+rand()' and then copy across for 30 columns, i.e. to column AD. Then copy down all these columns for 100 rows, i.e. to row 101. You should now have numbers in 100 rows for 30 columns. Each **row** will be treated as a separate random sample of size n = 30. Notice that each time you press *enter* the random function will recalculate the values in the cells; don't worry about this.

   Note: Since these numbers belong to a Uniform distribution the average for the whole population of numbers will be 12.5.

2. Go to cell AE1 and type „mean'. In cell AF1 type „std.error'. In cell AG1 type „LL' (for lower limit) and in cell AH1 type „UL' (for upper limit).
3. Now to calculate the mean for each of our samples of size 30. Go to cell AE2 and type „=average(A2:AD2)'. Press enter and copy this formula down for the 100 rows.
4. Now to calculate the standard error for each of the samples. Go to cell AF2 and type =stdev(A2:AD2)/sqrt(30)'. Again press enter and copy down for the 100 rows.
5. Now you have 100 samples of size 30. You have also calculated the standard error for each of these samples. Remember that it is usual to not be sure of the population standard deviation and to have to **estimate** it from the sample.

   Approximately how many of these **sample** means do you expect will be within one standard error of the **population** mean? ___________

6. Now calculate the one standard error **below** the sample mean. Go to call AG2 and type =AE2-1*AF2. Press enter and copy down.
7. Now calculate the one standard error **above** the sample mean. Go to call AH2 and type =AE2+1*AF2. Press enter and copy down.
8. You now have the upper and lower limits for an interval which **estimates** where the population mean lies. Does this interval actually contain the **population** mean? Go to cell AI2, and type (be careful!) =IF(AND(12.5>AG2,12.5<AH2),1,0). Copy down. If your interval does contain the population mean this formula will return „1', if the interval does not contain the population mean this formula will return „0'.

   Look at some of the rows with a „0' and you will see that the numbers between the upper and lower limit do NOT include 12.5.

9. How many of these intervals contain the population mean? To determine this, click in cell AI102 and then go to the tool bar and press Σ, the Auto-

Sum button, and press enter. What is the number? __________ is it close to what you estimated in step five?

10. Press F9 and you will see that the values in the worksheet are recalculated. Repeat this and watch the pattern to see how many of your intervals contain the population mean.
11. Now repeat the process for two standard errors above and below the mean. First of all, how many **sample** means do you expect will be within two standard errors of the **population** mean? _________ Now all you need to do is go to cells AG2 and AH2 and change „1*' to '2*' and copy the formulas down. How many of these sample means are within two standard errors of the population mean? ______ Is this what you expected?
12. Now repeat the process for three standard errors. First of all make your prediction (______), and then compare it with your actual value (_______).

D9 Simulation – The Grade 12 Heights problem

This exercise was based on the work of Erickson (2006). The aim of this simulation was to simulate the null hypothesis, in this case that there is no difference in mean heights between year 12 male and female students. Each time the simulation was carried out the difference in means was recorded and compared with the difference in means when the data were divided by gender. The results were pooled so that an informal *P*-value could be calculated. This simulation was followed by a lecture that introduced the formal 2-sample *t*-test procedure.

Student Instructions

1. Open up the file G12heights.xls. This file contains the heights and gender of a random sample of 34 grade 12 students who participated in the Census at Schools program. These data were obtained from the random sampler available from the Australian Bureau of Statistics web site: www.abs.gov.au.
2. Are adult males taller than females? You are most likely to say „yes', but it is likely that all the females doing this unit know some males shorter than they, and all the males will know some females taller than they. So how can we say males are taller than females? This is where a summary statistic such as the mean becomes useful. A more precise question would be: Is the mean height for males taller than the mean height for females?

3. Have a look at the data in the G12 file. The data have been sorted into males and females, and the mean heights for each gender have been calculated. Have a look in cell H2, you will notice that in **this sample** the male mean height is 11.62cm higher than the mean height for females. What are we likely to find for the whole **population**? Answer the following question:
4. If you took another random sample from this Census at School data for grade 12 students, will you get the same difference between the mean male and mean female height? Explain your answer:
5. We will call the observed difference of 11.62cm the *test statistic*. Now, imagine that in reality, the mean heights for males and females are equal. If this is the case, gender will have NO INFLUENCE on height. Could this be true? How can we test this? One way of testing this follows:
6. Go to the worksheet and in cell B2, type in =rand() and copy this formula down to cell B35.
7. Highlight the data in columns B and C. and go to Data, Sort, sort by RANDOM, and press OK.
8. Now the results are distributed randomly, what is the difference in means between the first and second group? Is it close to 11.62? Write your difference down in the table below.

Test statistic = 11.62		
Randomised Results:	7:	14:
1:	8:	15:
2:	9:	16:
3:	10:	17:
4:	11:	18:
5:	12:	19:
6:	13:	20:

9. Repeat step 2 until you have 20 differences in means recorded. Note that this process assumes that being a male or female doesn't make a difference to the overall mean height.
10. Place your results on the tally in the whiteboard. These will be discussed in the next lecture. Answer the following.
11. Does the observed difference of 11.62cm belong to the distribution of differences when the data are scattered at random? Yes/No
12. If there is really no difference in male and female mean height, then the observed difference in the sample of 11.62cm is:
    Common Likely Unlikely Very Unlikely
13. From these data, do you think that males really do have a mean height that is higher than that for females? Give reasons for your answer.

D10. Simulation – the chi-squared test for independence.

This simulation is based on a paper by Burrill (2002). The purpose of this exercise was to introduce students to the chi-squared test for independence in an informal way. The students were required to compare the observed results with those obtained if the null hypothesis were true, that is, that the two categorical variables were independent. The students were introduced to the formal procedure for the chi-squared test for independence in the lecture after this exercise.

Student Instructions

Headline in newspaper: Antibiotics can worsen E-coli infections.

According to a study from the University of Washington School of Medicine (Wong, Jelacic, Habeeb, Watkins, & Tarr, 2000)*, children who may be infected with the bacteria E-coli 0157:H7 should not be treated with antibiotics because they raise the risk of a potentially deadly complication called haemolytic uraemic syndrome (HUS). Researchers looked at 71 children with E-coli poisoning, nine of whom were treated with antibiotics. Of the nine, five developed HUS. Among the remaining 62, five developed HUS. Do the data support the headline?
The original data are shown in Table 1.
Table 1. Data showing incidence of HUS with treated and untreated children

	Antibiotics	No Antibiotics	Total
HUS	5	5	10
No HUS	4	57	61
Total	9	62	71

At this time, do you think that the evidence suggests that the researchers are correct, that treating with antibiotics increases the chance of a child getting HUS?
Now the totals in each row and column have to remain constant, that is the total number getting HUS remains at 10, and the total number given antibiotics remains at 9 and so on. Fill in the next table, assuming that the antibiotics have a STRONG POSITIVE effect on contracting HUS, i.e. everyone receiving antibiotics will get HUS.

Table 2. Strong positive evidence relating antibiotics and HUS

	Antibiotics	No Antibiotics	Total
HUS			10
No HUS			61
Total	9	62	71

The other extreme is that antibiotics have a STRONG NEGATIVE effect on getting HUS, i.e. everyone receiving antibiotics will NOT contract HUS. Assuming this is the scenario, fill in Table 3.

Table 3. Strong negative evidence relating antibiotics and HUS

	Antibiotics	No Antibiotics	Total
HUS			10
No HUS			61
Total	9	62	71

Now assume that the HUS affects children with E-coli 0157:H7 completely at random. If this is the case, the proportion of all the children with E-coli 0157:H7 who contract HUS, will be the same as the proportion of those taking antibiotics and contacting HUS, and the same as the proportion of those who did not take antibiotics yet contracted HUS. (For example, if ¼ of all the children with E-coli 0157:H7 contract HUS, then ¼ of the children who are not on antibiotics will get HUS, and ¼ of those not taking antibiotics will get HUS). Complete the following table. Go to one decimal place.

Table 4. Numbers of children contracting HUS if disease occurs at random.

	Antibiotics	No Antibiotics	Total
HUS			10
No HUS			61
Total	9	62	71

Compare this with Table 1, which contains the actual numbers contracting HUS, so do you think there is a relationship between contracting HUS and whether or not a child receives antibiotics? Give reasons.

The random number generator was then used in *Excel* to simulate the number of children who contracted HUS who were on antibiotics. The students were asked to repeat the simulation 10 times and to tally their results on the whiteboard. The students were then asked:

Now answer the question again. Do you think there is a relationship between contracting HUS and whether or not a child receives antibiotics? Give reasons. Is this answer different from your last answer?

*Wong, C., Jelacic, S., Habeeb, R., Watkins, S., & Tarr, P. (2000). The risk of hemolytic-uremic syndrome after antibiotic treatment of escherichia coli O157:H7 infections. *The New England Journal of Medicine, 342*, 1930-1936.

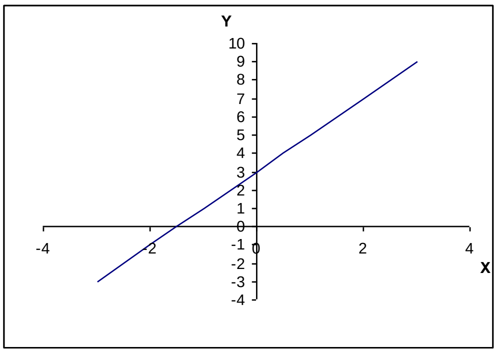
D11 Simulation – Fitting a line of best fit to data with measurement error

This exercise was based on the work by Franklin (1992). The aim of this exercise was to show students that as measurements involve error, the underlying regression population parameters (β_0 and β_1) are different from the estimated coefficients obtained from the sample, *a* and *b*. The exercise gives a visual representation of the way a sample line may vary around the "true" line. The other aim of the exercise is for the students to have a visual demonstration of regression lines with gradients that are not significantly different from zero.

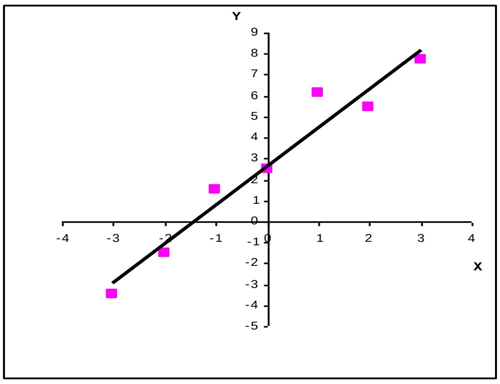
Student instructions

Finding the Line of Best Fit for Real Data

Most of you will have found the equations of lines in school. For example, given the points (2,7) and (4,11) you would have found that the equation of the line that goes through these two points is: y = 2x + 3. The „2' here refers to how steep the line is, the *gradient*, and the „3' here refers to where the graph intercepts the Y-axis, the *intercept*. The graph looks like this:

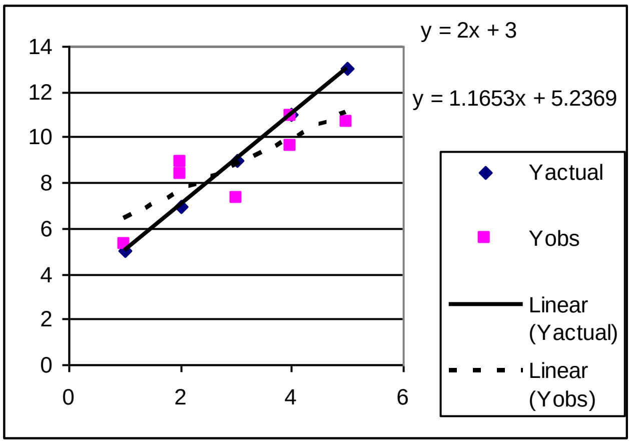


You may also have found the "line of best fit" for experimental data, since for real data you do not get all the points falling exactly on the line. Therefore, the graph may look like this:



No matter how sophisticated your equipment, ALL MEASUREMENTS ARE MADE WITH ERROR. Also, your data are only a sample and may not represent the entire population. If you were to repeat the experiment, you will not get exactly the same values. How does the inevitable experimental error affect your estimates of the gradient and intercept? Imagine that you are the deity of linear algebra, so that you alone know the true gradient and intercept of an experiment a lowly mortal is about to perform. This will mean that you alone know that the true relationship between X and Y is: Y = 2X +3.

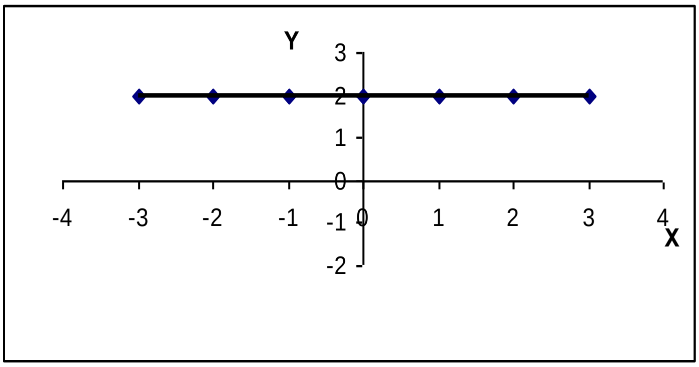
1. Go to a new sheet in excel. In cell A1 type X, and in the A column type in the whole numbers from 1 to 5.
2. In cell B1 type Yactual, and in cell B2 type: =2*A2+3. Copy this formula down. These are the values that experimenters will get if their equipment is perfect, so that no errors are made in the measurements.
3. Now to model the experimental error, go to cell F1 and type error, and then go to cell F2 and type, =NORMINV(rand(),0,1). Copy the formula down. This mimics an error from a Normal distribution with mean zero and standard deviation 1.
4. Now go to cell C1 and type the heading Yobserved, and in cell C2 type =B2+F2. Copy this formula down. This column mimics the readings you get in a real experimental situation, where there are errors in the measurement.
5. Now highlight the cells A1:C6, and plot a scatterplot, making sure the cells in the A column are in the X-values cell in the window.
6. How do the real results compare with the actual situation? First of all, put the mouse over one of the Yactual data points. Right click, go to add trendline, options, display equation on graph, and click OK. Do the same for the Yobs data, but also right click on the trendline and format it so that it is a dashed line. The result should look something like this:



7. See how far the estimate of the gradient and intercept of the line of best fit for the experimental data are from the true situation. By pressing F9 you can repeat the whole experiment. Try this several times, and see how the situation changes each time you repeat your experiment.

Remember, in the real experimental world you do not know what the true situation is, you only estimate it from your sample data. From the example shown here you would say that the line follows the equation : y = 1.2x + 5.2, and would not know that it is really y = 2x +3.

One of the problems with experimental data is that we might think that we have a relationship between x and y when none really exists. That is, a person changes the x values thinking that this will change the value on y, and measure these y values. Again you are the deity of linear algebra and know that the x values have no effect on y. The values of y are equal to 2 no matter what else changes. Graphically, it looks like this:



As we know, in our experiment the measurements will be made with an error. What effect will this have on our line?

1. Go to column A and type in the heading and numbers as before.
2. Go to column B and put in the heading Yactual and type in „2’ for every number.
3. Go to column F and type in the heading and the same formula to mimic the measurement errors made in the experiment.
4. Go to column C and type in the heading and formula as the previous section. Then draw the charts and add the trendlines, and format the observed trendline as before.
5. Note your observations. Does the experimental data follow a horizontal line, or is it sloped? Press F9 again to mimic the repeating of your experiment. Write down the equation of two of your experimental lines.

From an idea in: Franklin, L (1992), Using simulation to study linear regression, *The College Mathematics Journal,* 23, pp. 290-295.

D12 Demonstration – The Analysis of Variance

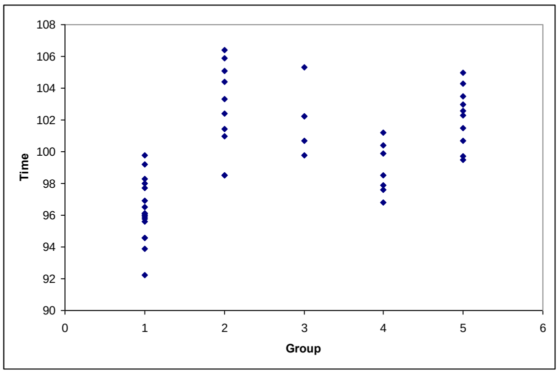
The aim of this demonstration was for students to gain an informal understanding of the process of the Analysis of Variance. The students were required to plot some data and make a judgement of which groups were significantly different from each other. The formal procedure for the Analysis of Variance was introduced in the lecture following this exercise.

Student Instructions

Introduction to the Analysis of Variance

In the chapter of your notes, “Analysis of Variance for Designed Experiments”, an example is given where an airline examines the time spent per call by telephone operators who answered callers wishing to make airline reservations. The data are in the file „Airline.xls’, and represent the times for calls (in seconds) to reservation operators in an airline reservation system. The calls were grouped according to the time of day or day of week. Group 1 represents calls made on weekdays between 8am and 4pm. Group 2 represents calls made between 4pm and midnight, group 3 represents calls made between midnight and 8am, group 4 represents calls made on a Saturday, and group 5 represents calls made on Sunday.

Open up this file. Using the data in columns A and B, draw a scatterplot showing the spread of results for each group. It should look something like this:



Calculate the overall mean, and the mean for each group. How does each group mean compare with the overall mean?

Group 1 has a mean that is **lower/higher** than the overall mean.

Group 2 has a mean that is **lower/higher** than the overall mean.

Group 3 has a mean that is **lower/higher** than the overall mean.

Group 4 has a mean that is **lower/higher** than the overall mean.

Group 5 has a mean that is **lower/higher** than the overall mean.

Are there **significant** differences between the average times between the groups? You already know how to compare the means for two groups (the 2-sample t-test), but what about 5 groups at once? It is not valid to compare the groups by doing a series of 2-sample t-tests; the reason for this will be explained in the lecture.

Now print out the graph (at a fairly large size). Mark the position of the mean for each group. Now compare how far each group mean is from the other. To decide whether or not there is a significant difference in means you will need to consider the position of each mean, the difference of means between each group, and the overlap between the spread in each group. Now answer the following:

I think group 1 **is/is not** significantly different to group 2.

I think group 1 **is/is not** significantly different to group 3.

I think group 1 **is/is not** significantly different to group 4.

I think group 1 **is/is not** significantly different to group 5.

I think group 2 **is/is not** significantly different to group 3.

I think group 2 **is/is not** significantly different to group 4.

I think group 2 **is/is not** significantly different to group 5.

I think group 3 **is/is not** significantly different to group 4.

I think group 3 **is/is not** significantly different to group 5.

I think group 4 **is/is not** significantly different to group 5.

Please bring these instructions to the next lecture

Appendix E: Statistical analyses

All analyses were completed using PASW Statistics 18.0.2 (SPSS inc., http://www.spss.com), unless otherwise stated.

E1 Analyses of the first questionnaire

Table E1.1

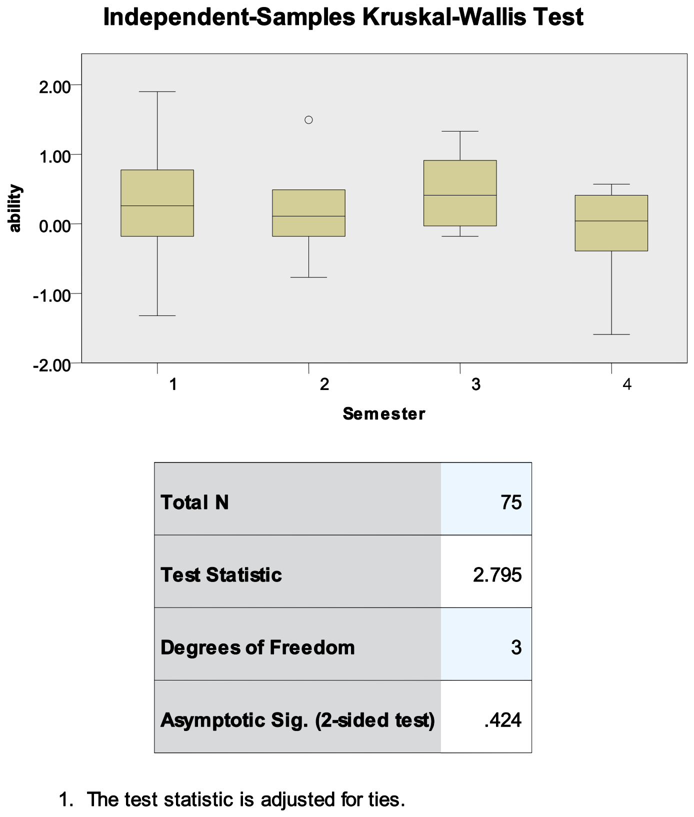
Summary of the Rasch item analysis for the first questionnaire (Winsteps)

```
     SUMMARY OF 13 MEASURED Items
+-----------------------------------------------------------------------------+
|           RAW                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      83.2      75.0         .00     .17      1.01    -.1   1.02    -.1 |
| S.D.      30.5        .0         .83     .02       .25    1.9    .27    1.4 |
| MAX.     130.0      75.0        1.80     .23      1.48    2.7   1.64    1.8 |
| MIN.      20.0      75.0       -1.43     .15       .62   -3.9    .60   -3.1 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .18  ADJ.SD     .81  SEPARATION  4.39  Item   RELIABILITY  .95 |
|MODEL RMSE    .17  ADJ.SD     .81  SEPARATION  4.67  Item   RELIABILITY  .96 |
| S.E. OF Item MEAN = .24                                                     |
+-----------------------------------------------------------------------------+
UMEAN=.000 USCALE=1.000
Item RAW SCORE-TO-MEASURE CORRELATION = -1.00
```

Table E1.2

Summary of the Rasch person analysis for the first questionnaire (Winsteps)

```
SUMMARY OF 75 MEASURED Persons
+-----------------------------------------------------------------------------+
|           RAW                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      14.4      13.0         .19     .41        .99     .0   1.02     .1 |
| S.D.       4.1        .0         .71     .07        .37    1.1    .52    1.0 |
| MAX.      23.0      13.0        1.90     .97       2.29    2.8   2.73    3.0 |
| MIN.       1.0      13.0       -2.87     .38        .44   -1.9    .27   -1.5 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .44  ADJ.SD      .55  SEPARATION  1.24  Person RELIABILITY  .60 |
|MODEL RMSE    .42  ADJ.SD      .57  SEPARATION  1.36  Person RELIABILITY  .65 |
| S.E. OF Person MEAN = .08                                                   |
+-----------------------------------------------------------------------------+
Person RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) Person RAW SCORE RELIABILITY = .59
```



Total N	75
Test Statistic	2.795
Degrees of Freedom	3
Asymptotic Sig. (2-sided test)	.424

1. The test statistic is adjusted for ties.

Figure E1.1. Results of Kruskal-Wallis test for differences in ability among the four semesters of the study.

Table E1.2

Items and P-values for the Kruskal-Wallis test for differences in scores among semesters for each item in the first questionnaire.

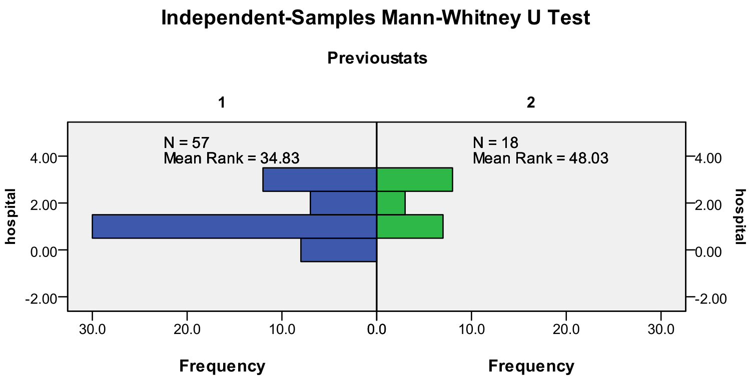
Item	*P*-value	Item	*P*-value
Snakes	.074	Tute B	.920
Cancer	.471	Tute C	.841
Eczema.	.787	Teacher	.743
Coin 1	.128	Factory	.635
Coin 2	.416	Urn A	.116
Coin 3-4	.547	Urn B	.361
Hospital	.016	Med	.428
Spinner 1	.574	A-B	.270
Spinner 2	.371	C-D	.543
Spinner 3	.422	E-F	.952
Spinner 4	.374	G-H	.487
Tute A	.577		

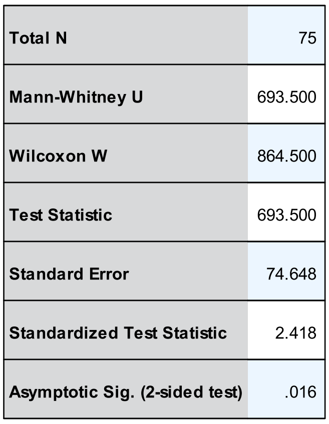
Table E1.3

Items and P-values for the Mann Whitney U tests for differences in scores for the students with and without previous statistical experience.

Item	*P*-value	Item	*P*-value
Snakes	.074	Tute B	.920
Cancer	.471	Tute C	.841
Eczema.	.787	Teacher	.743
Coin 1	.128	Factory	.635
Coin 2	.416	Urn A	.116
Coin 3-4	.547	Urn B	.361
Hospital	.016*	Med	.428
Spinner 1	.574	A-B	.270
Spinner 2	.371	C-D	.543
Spinner 3	.422	E-F	.952
Spinner 4	.374	G-H	.487
Tute A	.577		

* Significant at $\alpha = .05$





Total N	75
Mann-Whitney U	693.500
Wilcoxon W	864.500
Test Statistic	693.500
Standard Error	74.648
Standardized Test Statistic	2.418
Asymptotic Sig. (2-sided test)	.016

Figure E1.2. Back to back histograms and statistics for the “Hospital” question.

E2 Analyses of the second questionnaire

Table E2.1

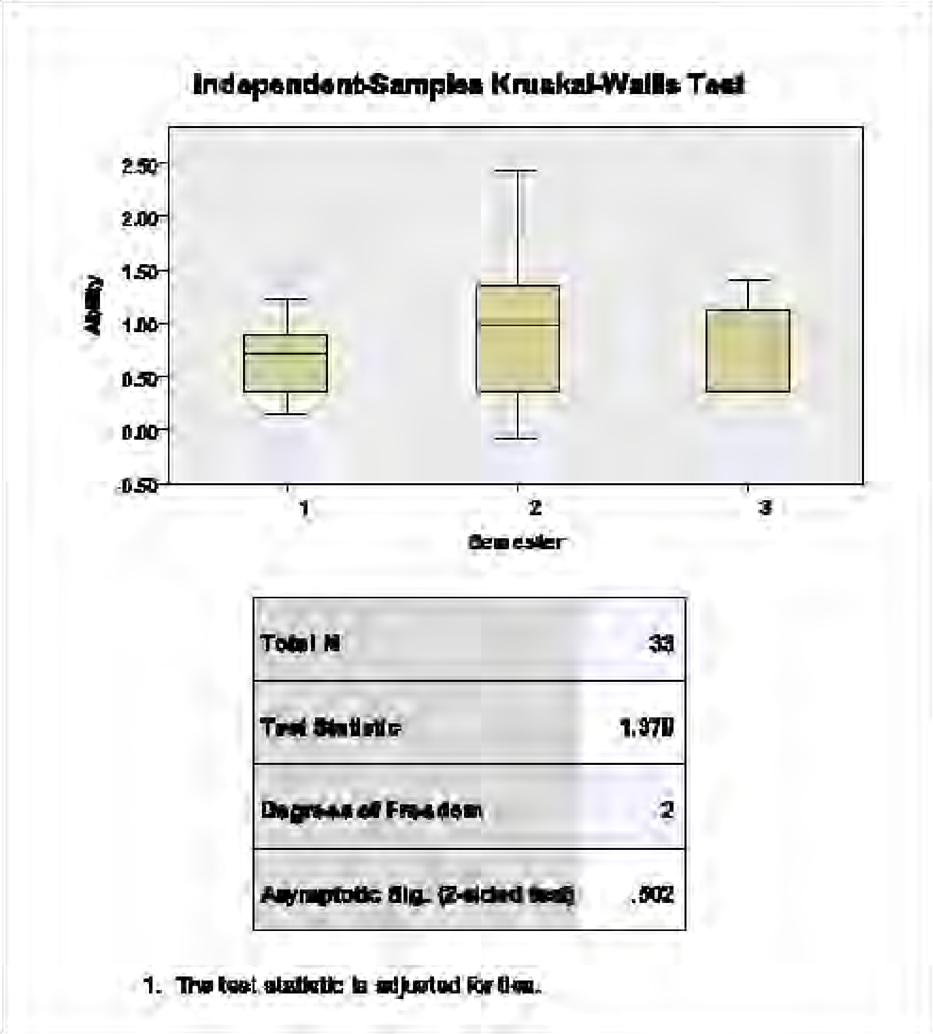
Summary of the Rasch item analysis for the second questionnaire (Winsteps)

```
      SUMMARY OF 13 MEASURED Items
+-----------------------------------------------------------------------------+
|           RAW                               MODEL         INFIT        OUTFIT    |
|          SCORE      COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       26.7       33.0         .00      .28      1.02      .0   1.00      .0 |
| S.D.       14.8         .0         .96      .03       .28     1.2    .24      .9 |
| MAX.       55.0       33.0        1.44      .37      1.41     1.5   1.35     1.1 |
| MIN.        8.0       33.0       -1.89      .25       .55    -2.2    .60    -1.5 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .30  ADJ.SD     .92  SEPARATION  3.06  Item   RELIABILITY  .90 |
|MODEL RMSE    .28  ADJ.SD     .92  SEPARATION  3.30  Item   RELIABILITY  .92 |
| S.E. OF Item MEAN = .28                                                     |
+-----------------------------------------------------------------------------+
UMEAN=.000 USCALE=1.000
Item RAW SCORE-TO-MEASURE CORRELATION = -1.00
429 DATA POINTS. APPROXIMATE LOG-LIKELIHOOD CHI-SQUARE: 720.10
```

Table E2.2

Summary of the Rasch person analysis for the second questionnaire (Winsteps)

```
     SUMMARY OF 33 MEASURED Persons
+-----------------------------------------------------------------------------+
|           RAW                               MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      10.5      13.0        -.40      .44      1.02     .0   1.00     .0 |
| S.D.       3.2        .0         .58      .02       .44    1.2    .49    1.1 |
| MAX.      20.0      13.0        1.29      .49      1.90    2.3   2.13    2.4 |
| MIN.       6.0      13.0       -1.29      .41       .44   -1.8    .43   -1.6 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .48  ADJ.SD     .33  SEPARATION   .69  Person RELIABILITY  .32 |
|MODEL RMSE    .44  ADJ.SD     .39  SEPARATION   .88  Person RELIABILITY  .44 |
| S.E. OF Person MEAN = .10                                                   |
+-----------------------------------------------------------------------------+
Person RAW SCORE-TO-MEASURE CORRELATION = 1.00
CRONBACH ALPHA (KR-20) Person RAW SCORE RELIABILITY = .43
```



Independent-Samples Kruskal-Wallis Test

Total N	33
Test Statistic	1.379
Degrees of Freedom	2
Asymptotic Sig. (2-sided test)	.502

1. The test statistic is adjusted for ties.

Figure E2.1. Box plots and statistics for the Kruskal-Wallis test for the differences in students' ability among semesters for the second questionnaire.

Table E2.3

Items and P-values for the Kruskal-Wallis test for difference in scores among semesters for each item in the second questionnaire.

Item	P-value
Pacific B	.094
Fish	.567
Cereal D	.564
Fred	.505
Random	.207
CB 1	.491
CB 2	.639
CB 3	.372
Med A	.675
Med B	.780
Urn A	.963
Urn B	.464
Pacific A	.861

E3 Comparisons between ability scores and final scores from formal assessment

Table E3.1

Results of Paired t-test comparing students' ability from the first and second questionnaire (Microsoft Excel).

t-Test: Paired Two Sample for Means		
	First Questionnaire	*Second Questionnaire*
Mean	0.311	0.787
Variance	0.240	0.331
Observations	33	33
Pearson Correlation	0.130	
Hypothesized Mean Difference	0	
df	32	
t Stat	-3.870	
P(T<=t) one-tail	0.000	
t Critical one-tail	1.694	
P(T<=t) two-tail	0.001	
t Critical two-tail	2.037	

Table E3.2

Results of ANOVA on differences of ability score among semesters (Microsoft Excel).

ANOVA: Single Factor

SUMMARY

Semester	*Count*	*Sum*	*Average*	*Variance*
Pre-intervention	9	2.22	0.247	0.108
Cycle 1	20	13.33	0.667	0.613
Cycle 2	4	0.12	0.030	0.463

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*
Between Semesters	1.99	2	0.997	2.151	0.134
Within Groups	13.907	30	0.464		
Total	15.902	32			

Table E3.3

Results of correlation analysis between the ability scores from the first questionnaire (Q1), the second questionnaire (Q2), and the final score (final), with all data included.

Correlations

		Q1	Q2	Final
Q1	Pearson Correlation	1	.178	.017
	Sig. (2-tailed)		.374	.935
Q2	Pearson Correlation	.178	1	.354
	Sig. (2-tailed)	.374		.070*
Final	Pearson Correlation	.017	.354	1
	Sig. (2-tailed)	.935	.070	

* Significant at α = .10

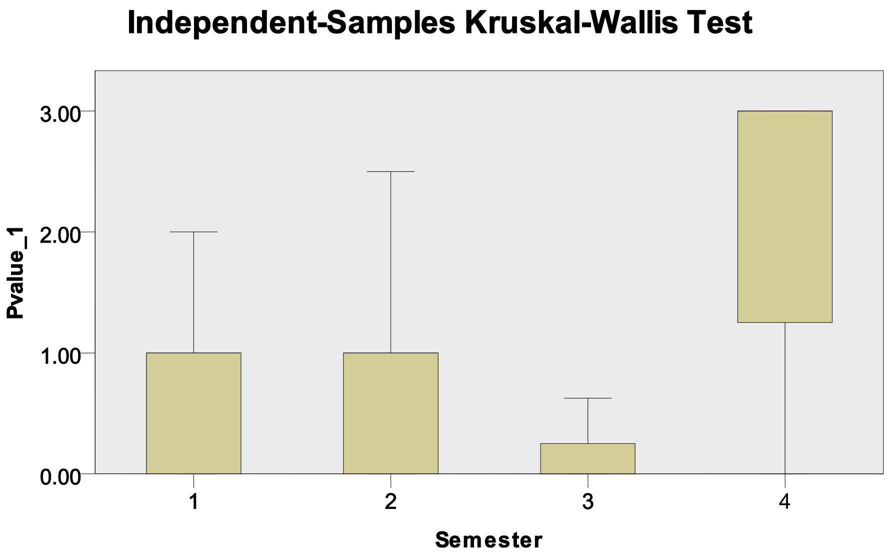
Table E3.4

Results of correlation analysis between the ability scores from the second questionnaire (Q2), and the final score (final), with the influential point excluded.

Correlations

		Q2	Final
Q2	Pearson Correlation	1	.177
	Sig. (2-tailed)		.387
Final	Pearson Correlation	.177	1
	Sig. (2-tailed)	.387	

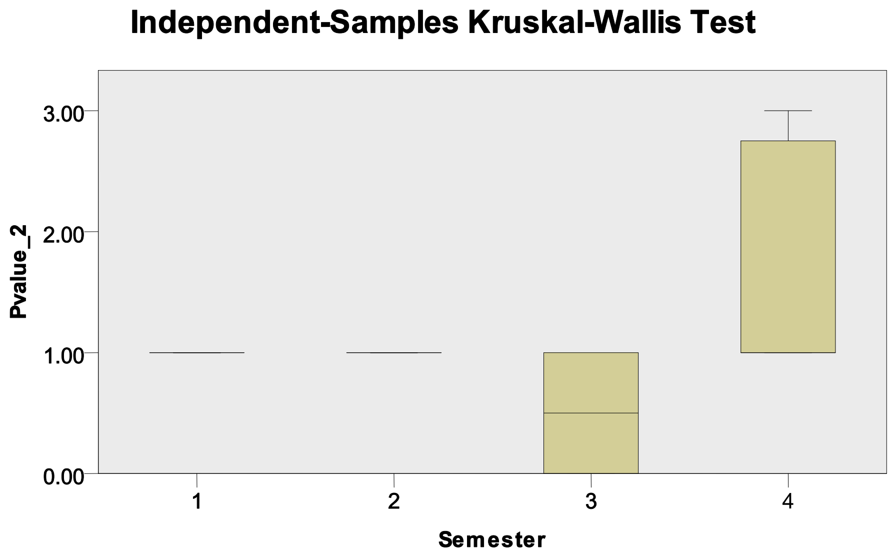
E4. Comparison among semesters for the *P*-value and confidence interval questions on the test


Independent-Samples Kruskal-Wallis Test

Total N	54
Test Statistic	19.399
Degrees of Freedom	3
Asymptotic Sig. (2-sided test)	.000

1. The test statistic is adjusted for ties.

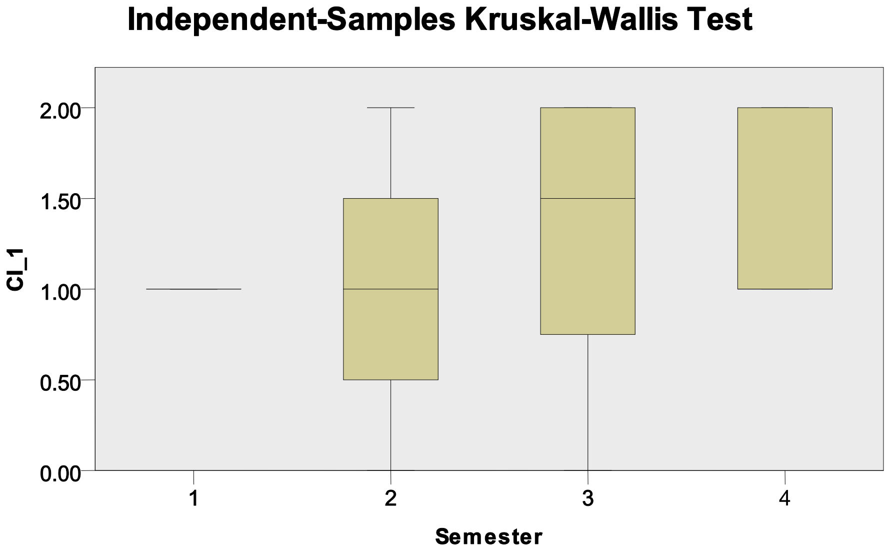
Figure E4.1. Box plots and statistics for the Kruskal-Wallis test for difference in scores among semesters for the first P-value question.



Total N	54
Test Statistic	10.485
Degrees of Freedom	3
Asymptotic Sig. (2-sided test)	.015

1. The test statistic is adjusted for ties.

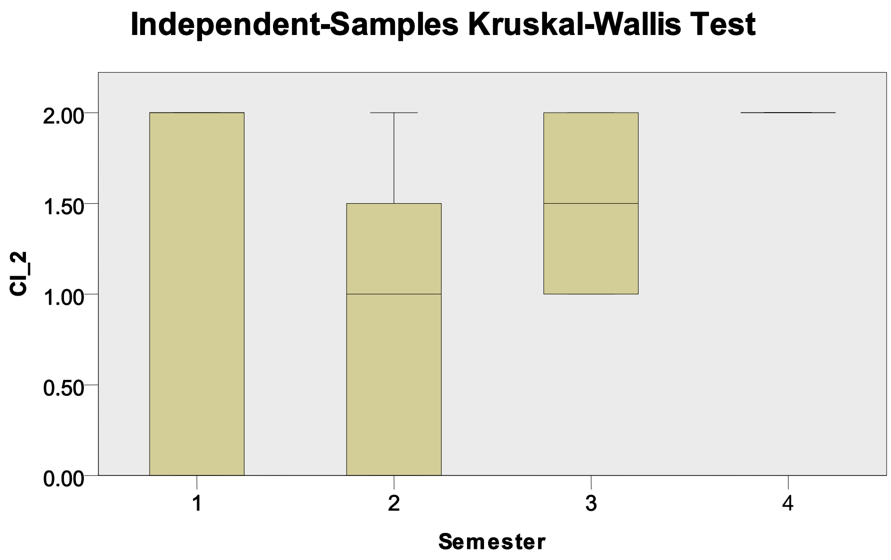
Figure E4.2. *Box plots and statistics for the Kruskal-Wallis test for difference in scores among semesters for the second P-value question.*



Total N	54
Test Statistic	2.966
Degrees of Freedom	3
Asymptotic Sig. (2-sided test)	.397

1. The test statistic is adjusted for ties.

Figure E4.3. Box plots and statistics for the Kruskal-Wallis test for difference in scores among semesters for the first confidence interval question.



Total N	54
Test Statistic	11.463
Degrees of Freedom	3
Asymptotic Sig. (2-sided test)	.009

1. The test statistic is adjusted for ties.

Figure E4.4. Box plots and statistics for the Kruskal-Wallis test for difference in scores among semesters for the second confidence interval question.

Table E4.1

Mean ranked scores for the confidence interval questions

	Mean Rank	
Semester	**Part (a)**	**Part (b)**
Pre-intervention	25.05	26.86
Cycle 1	25.22	21.62
Cycle 2	32.50	31.00
Cycle 3	32.00	38.58

Table E4.2

Mean ranked scores for the P-value questions

	Mean Rank	
Semester	**Question 1**	**Question 2**
Pre-intervention	22.14	26.00
Cycle 1	26.04	26.86
Cycle 2	13.33	15.25
Cycle 3	42.54	36.33