Spreadsheets in Education (eJSiE)

Volume 4 | Issue 3 Article 2

4-6-2011

Teaching Statistics in a Spreadsheet Environment Using Simulation

Graham D. Barr
University of Cape Town, gdi@iafrica.com

Leanne Scott
University of Cape Town, leanne.scott@uct.ac.za

Recommended Citation

Barr, Graham D. and Scott, Leanne (2011) Teaching Statistics in a Spreadsheet Environment Using Simulation, *Spreadsheets in Education (eJSiE)*: Vol. 4: Iss. 3, Article 2.
Available at: http://epublications.bond.edu.au/ejsie/vol4/iss3/2

Teaching Statistics in a Spreadsheet Environment Using Simulation

Abstract

The authors' experiences with a new teaching approach in an introductory statistics course involving some 1200 first year students in South Africa form the context for the development of the ideas in this paper. The paper focuses on the teaching of statistics within a spreadsheet environment whereby students are, inter alia, required to master the basics of MS Excel to perform statistical calculations. This approach has the advantages of developing the students' ability to work with data whilst also building their understanding of the algebraic relationships between elements embedded in the spreadsheet formulae which they use. The authors advocate a two-stage approach in which statistical understanding is built by initially empowering students to use simple spreadsheet operations, followed up by the use of more sophisticated simulation tools in a Visual Basic for Applications (VBA) programming environment. They demonstrate the use of a classroom experiment aimed at exploring the statistical distributions of a number of pieces of information generated by the students. Teaching sessions are then built around a suite of MS Excel VBA-based simulations which demonstrate the concept of random variation as well as show how statistical tools can be used to explore the concept of uncertainty.

Keywords

Simulation, Teaching, Statistics, Spreadsheets, Excel

TEACHING STATISTICS IN A SPREADSHEET ENVIRONMENT USING SIMULATION

Graham D. Barr and Leanne Scott
Department of Statistical Sciences, University of Cape Town

ABSTRACT

The authors' experiences with a new teaching approach in an introductory statistics course involving some 1200 first year students in South Africa form the context for the development of the ideas in this paper. The paper focuses on the teaching of statistics within a spreadsheet environment whereby students are, inter alia, required to master the basics of MS Excel to perform statistical calculations. This approach has the advantages of developing the students' ability to work with data whilst also building their understanding of the algebraic relationships between elements embedded in the spreadsheet formulae which they use. The authors advocate a two-stage approach in which statistical understanding is built by initially empowering students to use simple spreadsheet operations, followed up by the use of more sophisticated simulation tools in a Visual Basic for Applications (VBA) programming environment. They demonstrate the use of a classroom experiment aimed at exploring the statistical distributions of a number of pieces of information generated by the students. Teaching sessions are then built around a suite of MS Excel VBA-based simulations which demonstrate the concept of random variation as well as show how statistical tools can be used to explore the concept of uncertainty.

Keywords: Simulation; Teaching; Statistics; Spreadsheets; MSExcel

INTRODUCTION

Fundamental statistical concepts remain elusive to many students in their introductory course on statistics. Additional burdens such as language barriers can compound these difficulties and teachers continue to explore new ways of conveying foundational concepts. It is argued that teaching mathematical and statistical principles through a spreadsheet platform offers significant advantages. The structuring of a spreadsheet develops a general algebraic way of thinking as the process requires skills in expressing numerical relationships using algebraic notation. In the field of Statistics, the advantages of spreadsheets for teaching purposes are particularly marked as spreadsheets can simultaneously present an easily navigable yet extensive vista of numeric information stored in multiple rows and columns, along with the formulaic links between them, as well as an associated rich graphical depiction of the same data. However, the richest feature that a spreadsheet offers to the teacher of Statistics is its ability to show how one can mimic the process of repeated statistical experiments. By simulating statistical sampling one can reveal a range of subtle and often misunderstood ideas which are central to basic statistical knowledge, such as those of randomness and statistical distributions. Moreover, beyond the basic structure of the spreadsheet which lends itself so well to the teaching of statistical ideas, MSExcel, the most often used spreadsheet in academia and commerce, has an extremely powerful built-in programming language, Visual Basic for Applications (VBA). This allows teachers and students to enhance and leverage basic Excel power and functionality to a new level of flexibility and sophistication with click button automation and slickness.

In this paper we will relate how our experience with teaching first year Statistics courses at UCT has shown that the key concepts of randomness and distribution are most effectively taught through simulation in an MSExcel based spreadsheet environment using a 2-stage approach, first using a formula based spreadsheet and, subsequently with the enhancement of VBA. The focus of the paper is a carefully crafted teaching example which demonstrates to students how the random sampling of a set of distinct attributes associated with a set of individuals may reveal completely different distributions of each attribute. A key component of this teaching example is a set of associated customized Excel spreadsheets, firstly without and then with VBA enhancements, which elucidate these ideas and have been shown to be very didactically effective.

THE USE OF SPREADSHEETS IN MATHEMATICAL AND STATISTICAL EDUCATION

The usefulness of the spreadsheet for educational purposes has been widely recognised in the literature. The computer spreadsheet was invented by Dan

Bricklin and Bob Frankston who wrote Visicalc for the Apple II platform in 1979. Bricklin and Frankston sold the rights in VisiCalc to Lotus Development Corporation who developed the Lotus 1-2-3 package and as early as 1984, one year after the launch of Lotus 1-2-3, the spreadsheet had been noticed and recognized as a force in mathematics education (Arganbright, D, 1984)) and then one year later in statistics education (Soper and Lee, 1985)

In fact, the spreadsheet has become recognised across the educational spectrum, even for young learners in the early grades; see the work of Sutherland (2007) who states that "One way to help pupils move from a non-algebraic to an algebraic approach can be through work with spreadsheets." It is now universally recognised that the two-dimensional structure of spreadsheets along, with their associated graphical components, can facilitate the comprehension of a wide range of mathematical and statistical concepts by providing a supportive platform for conceptual reasoning; see Baker and Sugden (2003) for a comprehensive review of the application of spreadsheets in teaching across the mathematical, physical and economic sciences. Black (1999) was one of the first authors to recognise the usefulness of spreadsheets in a simulation context for teaching complex statistical concepts. The idea of using simulation, especially within VBA, was found by Barr and Scott (2008) to be particularly useful and effective for the teaching of first year statistics to large classes and they confirm the sentiments of Jones (2005) that statistical concepts and procedures taught within the context of a spreadsheet tend to be transparent to pupils, allowing them to look inside the "black box" of statistical techniques. A comprehensive survey of the use of simulation methods for teaching statistical ideas has been done by Mills (2002). It is seen that the literature supports the notion that spreadsheets lead students into a didactically rich and effectively open-ended line of inquiry with MSExcel as the de facto spreadsheet standard.

The core part of this paper is to showcase the teaching of the two foundational statistical concepts of randomness and underlying distribution through simulation using both simple spreadsheet functions and more sophisticated VBA programs. Our experience has lead us to believe that VBA-structured spreadsheets by themselves provide a difficulty for a large cohort of students; a leap into the dark to some extent. However, when properly scaffolded by a standard spreadsheet with formulae approach, it becomes a more effective learning tool. By itself, VBA simulation programs or simulation programs written in java on the web are neat and impressive but constitute too much of a black-box for students. By leading students through a formulaically structured approach on the spreadsheet first, and putting the appropriate building blocks in place, they find VBA programs, which leverage the first-tier analysis to a second, more accessible level.

THE NOTION OF RANDOM VARIATION : A MAJOR LEARNING HURDLE IN STATISTICS

One of the fundamental concepts of statistics, in fact one of the platforms on which we build the edifice that is the statistical discipline, is that of random variation. It is a notion that we as educators take for granted that people have an intuitive understanding of. It is, however, a subtle notion that apparently random and unpredictable events have underlying patterns which can be uncovered through (*inter alia*) long term observation.

An open ended invitation to describe their understanding of 'randomness' and how it affects our day to day lives was extended to a group of adult learners, all of whom were tertiary educators themselves. A brief discussion on the perceived need for an understanding of Statistics preceded this, touching on the fact that very little of what happens in life can be predicted with certainty, and indicating that statistics provides a mechanism to manage uncertainty associated with random variation. A variety of notions of randomness were articulated, from which some unexpected themes emerged, in particular a pervasive view of randomness as being a 'victimizing' force or a tool of malevolent authorities, associated with poor planning and discipline. In many cases, randomness was associated with chaos. All of the descriptions volunteered by the students were devoid of any notion of underlying pattern or distribution. Subsequent discussions confirmed that they believed the existence of an underlying pattern was in fact contradictory to the very idea of random variation.

From an educational point of view, it could be suggested that the consequence of (these) students' views of the nature of random variation is that there is an inflated view of the power of statistics to impose order on randomness or a jaded view of the discipline as a tool to disguise chaos and unfairness. It is suggested that beginning the statistics journey with a description of the world as containing innate patterns and order which are hidden from us through the random and unpredictable way in which individual outcomes are free to vary, may open up the power and interest of the discipline in a way that the traditional approach of teaching 'theory followed by its application' fails to do. We will show below, using an appropriate experiment, that the spreadsheet environment is an ideal canvas on which to sketch and unveil the ideas around random variation and underlying patterns.

TACKLING "RANDOM VARIATION" THROUGH THE CLASS EXPERIMENT: *RANDOM SELECTION; DIFFERENT PATTERNS!*

We begin the Class Experiment with a discussion with the students about different types of numbers, reflected both by the different measurement scales we choose to assign to them, and by the process that generates them. We ask them to consider the following experiment in which each student in the class will contribute four pieces of information, *viz*: (1) their first name; (2) their height (in cm); (3) an integer randomly selected within the range 1 to 50; and, finally, (4) their personal results of a (to be explained) experiment involving mice! Once we have generated this data we will be collecting it from everyone and constructing four separate histograms of each of the four number types. As part of this experiment we will be constructing ways to generate data for (3) by using, and exploring, the Excel random number generator. Data for (4) involves a mouse training experiment which tests the ability of 5 (simulated) randomly selected mice to navigate a simple maze, recording the number of successful mice. Students will be able to run the Excel models to record their own data for the class experiment. The focus of this class exercise is for students to answer the key question: *What shapes do we anticipate for these histograms?* We proceed by considering the randomly generated number.

THE RANDOM NUMBERS

Suppose we are interested in mimicking the National Lottery and (repeatedly) generating a random number which lies between 1 and 50. Each time we make a draw, this would be akin to drawing a number from a hat with the numbers 1 to 50 in it. If we want to keep drawing a number from this hat in such a way that all numbers are equally likely, we would have to also suppose that we have a very large hat that can hold such a large (and equal) quantity of each of the numbers that it doesn’t limit our thinking about the situation. What would a histogram of these numbers look like? Some discussion would probably lead us to conclude that we would expect all of the bars in the histogram to be of equal height. Now let’s see what sort of patterns we get when we randomly take numbers out of the hat. If we just pick one number it could pop up at any point in the specified interval (1; 50). The fact that the number is randomly selected means there is no way of telling (from the preceding numbers or, in fact, any other source of information) exactly what number is going to pop up next. In order to see a pattern of numbers we need to observe more than one randomly drawn number. Let’s see what happens when we generate 10 random numbers. Perhaps the pattern looks a little obscure… sometimes it looks very different from what we might have expected. What happens when we draw 100 numbers,…or 1000? It seems that as the pool of numbers that we are drawing grows, so the pattern of the numbers in the hat gets revealed. (see Figure 1 below) A small pool can give quite a misleading picture of the histogram of the numbers in the hat! However a big pool is less likely to do so. So how big a pool of numbers do we need to have access to in order to get a reliable picture of the numbers in the hat? Imagine that we hadn't known the shape of the histogram of the numbers in the hat.

The numbers might, for example, have had a different (other than flat/ rectangular) pattern of distribution? Let's use some spreadsheet commands to model our thinking of the above. We can easily simulate the drawing from our hat of a number (where all the numbers in the hat are integers that lie between 1 and 50 inclusive) by typing the formula:

```
= (RANDBETWEEN(1,50))
```

into our spreadsheet.

This computes (and rounds to the nearest integer) a single random number in the interval (1, 50). We can then resample this number by pressing the F9 key. We can also display this visually by plotting the number in a simple histogram. We should first set up some bin intervals, the simplest is the set of 10 intervals between 1 and 50 with interval width 5. We then use an array formula to compute the frequencies:

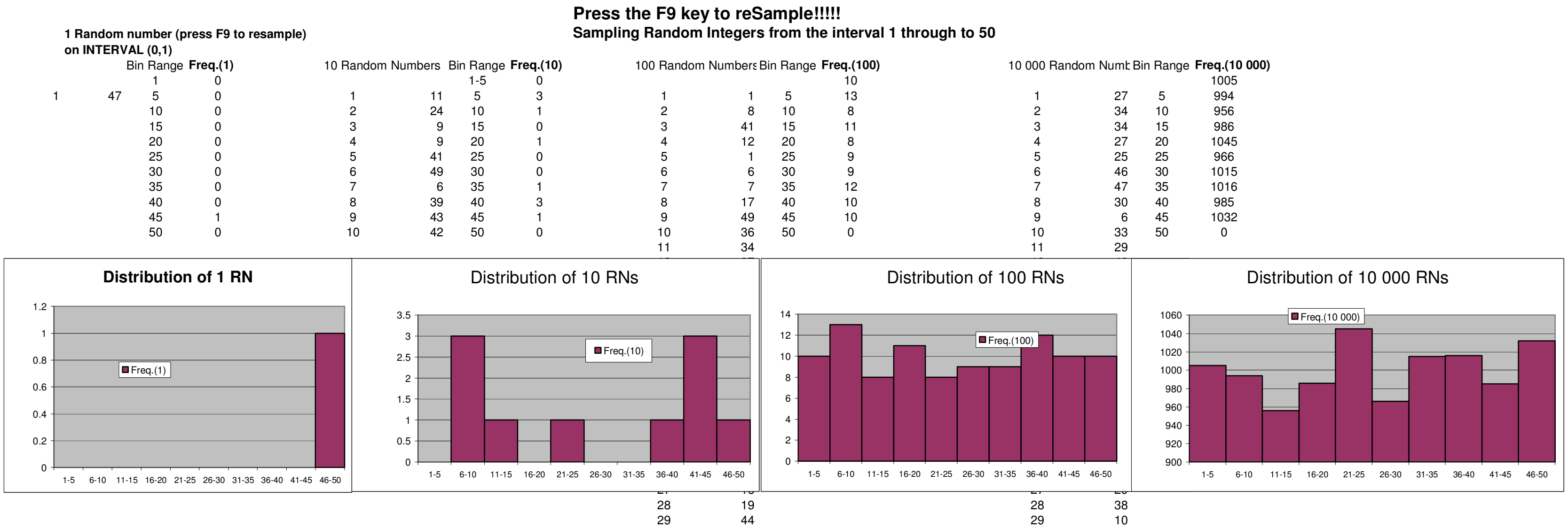
```
={FREQUENCY(data_range, bin_range)}
```

which can then be plotted in a histogram. We replicate this procedure for a sequence of sample sizes from 1 (single random number), through to 10 (10 random numbers), then 100 and finally 10 000.

By repeatedly pressing F9 (which simply recalculates the formulae and effectively re-samples) we get to replay the (random) selection of different sized pools of numbers from the hat. Each time we get a different selection of numbers. One feature becomes apparent. The pattern (of the numbers in the hat) becomes increasingly clearly revealed as we observe larger and larger pools of numbers from the hat. Although we cannot *at any stage* predict what the next number will be, by observing randomly selected pieces of information from the hat we can begin to piece together what the pattern of numbers in the hat must look like. In real life this is likely to mean we have to observe the pattern of (randomly revealed pieces of information) over time, before we can begin to understand something about the nature (or distribution!) of the numbers in the hat.

We can then use VBA to extend these ideas. The VBA version of the uniform (Figure 2 below) provides a neat and slick point and click demonstration of how simulated uniforms can be generated. A key feature of this is the ability to compare the theoretical (or expected) frequency graph with the empirical (that is the one generated by simulation).

FIGURE 1. RANDOM NUMBERS WITH INCREASING SAMPLE SIZE.



Press the F9 key to reSample!!!!!

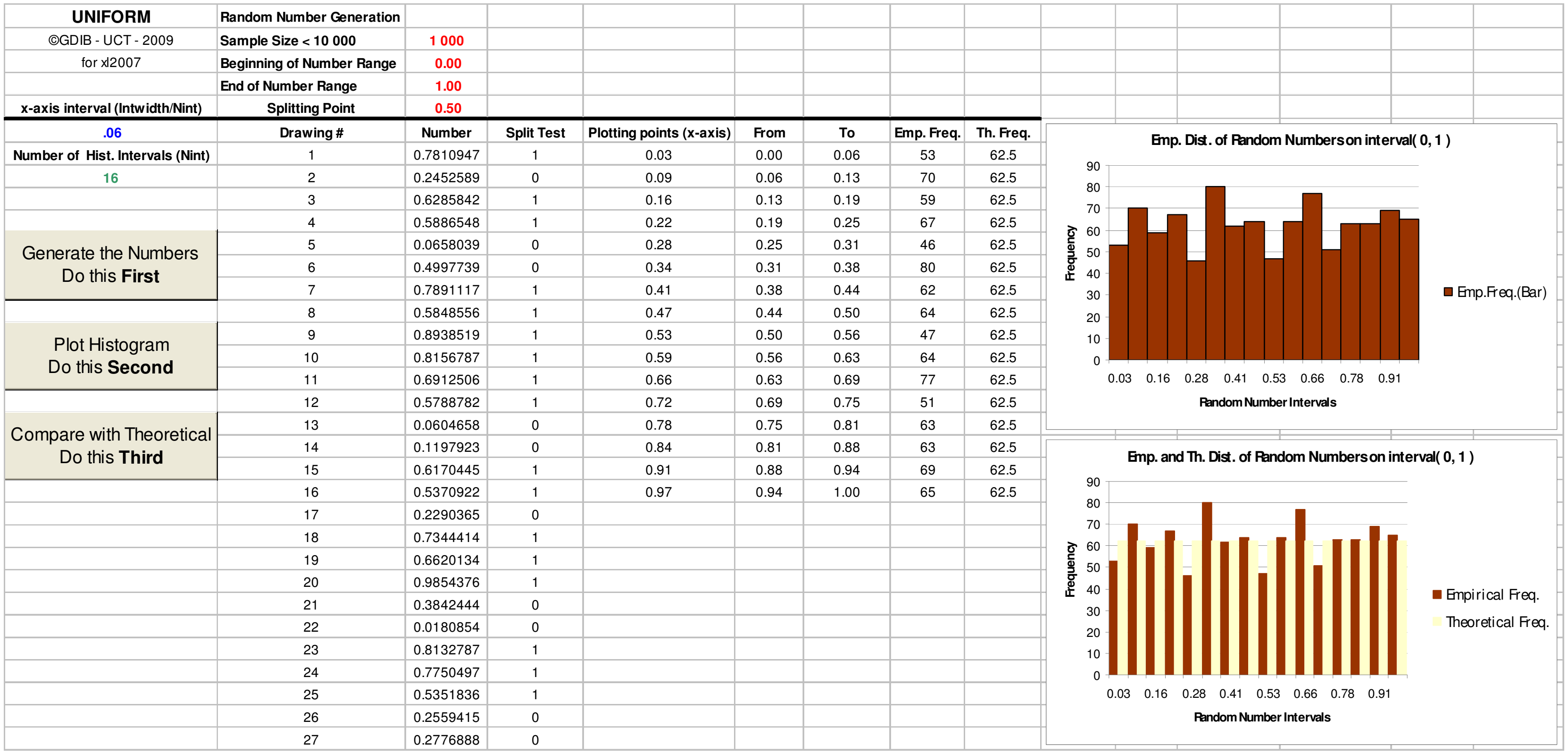
Sampling Random Integers from the interval 1 through to 50

1 Random number (press F9 to resample)
on INTERVAL (0,1)

		Bin Range	**Freq.(1)**	10 Random Numbers		Bin Range	**Freq.(10)**	100 Random Numbers		Bin Range	**Freq.(100)**	10 000 Random Numb		Bin Range	**Freq.(10 000)**
		1	0			1-5	0				10				1005
1	47	5	0	1	11	5	3	1	1	5	13	1	27	5	994
		10	0	2	24	10	1	2	8	10	8	2	34	10	956
		15	0	3	9	15	0	3	41	15	11	3	34	15	986
		20	0	4	9	20	1	4	12	20	8	4	27	20	1045
		25	0	5	41	25	0	5	1	25	9	5	25	25	966
		30	0	6	49	30	0	6	6	30	9	6	46	30	1015
		35	0	7	6	35	1	7	7	35	12	7	47	35	1016
		40	0	8	39	40	3	8	17	40	10	8	30	40	985
		45	1	9	43	45	1	9	49	45	10	9	6	45	1032
		50	0	10	42	50	0	10	36	50	0	10	33	50	0
								11	34			11	29		
								28	19			28	38		
								29	44			29	10		

FIGURE 2 USING VBA TO EXTEND THESE IDEAS.

The VBA version of the uniform provides a neat and slick point and click demonstration of how simulated uniforms can be generated. A key feature of this is the ability to compare the theoretical (or expected) frequency graph with the empirical (that is the one generated by simulation)



UNIFORM	**Random Number Generation**							
©GDIB - UCT - 2009	**Sample Size < 10 000**	1 000						
for xl2007	**Beginning of Number Range**	0.00						
	End of Number Range	1.00						
x-axis interval (Intwidth/Nint)	**Splitting Point**	0.50						
.06	**Drawing #**	**Number**	**Split Test**	**Plotting points (x-axis)**	**From**	**To**	**Emp. Freq.**	**Th. Freq.**
Number of Hist. Intervals (Nint)	1	0.7810947	1	0.03	0.00	0.06	53	62.5
16	2	0.2452589	0	0.09	0.06	0.13	70	62.5
	3	0.6285842	1	0.16	0.13	0.19	59	62.5
	4	0.5886548	1	0.22	0.19	0.25	67	62.5
Generate the Numbers Do this **First**	5	0.0658039	0	0.28	0.25	0.31	46	62.5
	6	0.4997739	0	0.34	0.31	0.38	80	62.5
	7	0.7891117	1	0.41	0.38	0.44	62	62.5
	8	0.5848556	1	0.47	0.44	0.50	64	62.5
Plot Histogram Do this **Second**	9	0.8938519	1	0.53	0.50	0.56	47	62.5
	10	0.8156787	1	0.59	0.56	0.63	64	62.5
	11	0.6912506	1	0.66	0.63	0.69	77	62.5
	12	0.5788782	1	0.72	0.69	0.75	51	62.5
Compare with Theoretical Do this **Third**	13	0.0604658	0	0.78	0.75	0.81	63	62.5
	14	0.1197923	0	0.84	0.81	0.88	63	62.5
	15	0.6170445	1	0.91	0.88	0.94	69	62.5
	16	0.5370922	1	0.97	0.94	1.00	65	62.5
	17	0.2290365	0					
	18	0.7344414	1					
	19	0.6620134	1					
	20	0.9854376	1					
	21	0.3842444	0					
	22	0.0180854	0					
	23	0.8132787	1					
	24	0.7750497	1					
	25	0.5351836	1					
	26	0.2559415	0					
	27	0.2776888	0					

THE MOUSE EXPERIMENT - GENERATING DATA FROM A BINOMIAL DISTRIBUTION

The uniform case above constitutes a starting point and necessary platform to consider more complex distributions. In particular, it leads naturally onto the Binomial. In the Binomial case we consider a fixed number of trials, where at each trial we have assigned a fixed probability of success or failure. In order to generate the fourth piece of information for our class experiment we consider the following scenario:

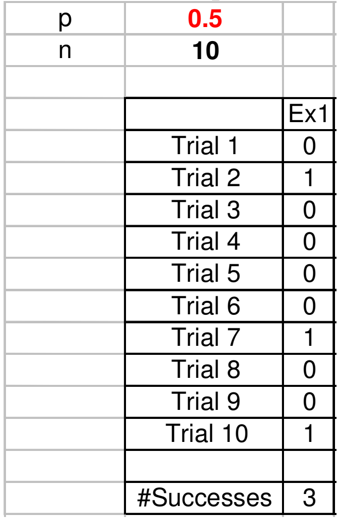
We, as statisticians, have been approached to mediate on a claim that an animal trainer has managed to train a group of ten mice to turn left at the end of a tunnel. As evidence of this feat he has cited the fact that in an observed demonstration, out of 10 trained mice, 9 of them turned left! In an attempt to pronounce on these results we try and build a model that mimics the behaviour of the mice. We start with the sceptical view of the situation and assume in fact that the mice have NOT been trained. Thus each mouse makes an arbitrary choice of left or right at the end of the tunnel which means that the probability of them turning left is 0.5 We are interested now in what different bunches of 10 hypothetical mice might do in terms of 'left turning' behaviour.

Let's assume we have a large number of mice at our disposal, which are very similar and which we put through the identical training regime. To run the experiment we select 10 mice, put them through the tunnel and record the number of successful mice, i.e. the number out of the 10 who turned left. We could then select another 10 randomly and perform the experiment again, in fact, we could repeat this as many times as we like. Our selection assumes the mice are always 'fresh', referred to as sampling 'without replacement' (i.e. our mice don't get tired or complacent or difficult!).

We use the spreadsheet to help us produce some simulated results for batches of 10 mice (see Figure 3). The mechanics of this are as follows: We select a random number between 0 and 1. If the number is less than or equal to 0.5 we assume the mouse went left, if not we assume it went right. Then we do this for 10 mice, labelling the results Trial 1, through to Trial 10. This comprises one experiment. We can see how easy this is on the spreadsheet. In cell C3 we put the value of p, in this case 0.5. The following formula:

=IF(RAND()<C3,1,0)

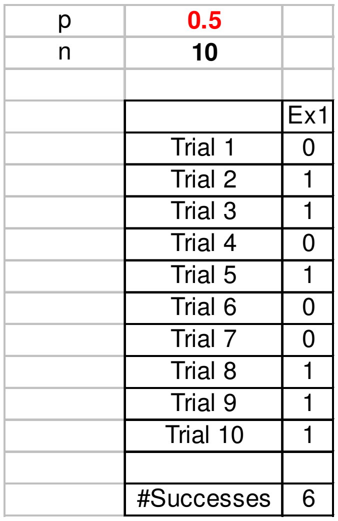
results in a 1 if the random number calculated is less than 1 and a 0 if it is not (that means it is greater than 1). We then copy this formula down a further nine cells to get the model results for one experiment.



p	0.5	
n	**10**	
		Ex1
	Trial 1	0
	Trial 2	1
	Trial 3	0
	Trial 4	0
	Trial 5	0
	Trial 6	0
	Trial 7	1
	Trial 8	0
	Trial 9	0
	Trial 10	1
	#Successes	3

In this case, Trials 2, 7 and 10 were a “success” (mouse turned left). Trials 3 to 6 and 8 and 9 were failures (turned right).

Let’s repeat this experiment under the same theoretical conditions (i.e. the same ineffective training program which yields p=0.5 and with the mice being selected randomly). Excel allows us to repeat the experiment just by pressing the F9 (recalculate) key.



p	0.5	
n	**10**	
		Ex1
	Trial 1	0
	Trial 2	1
	Trial 3	1
	Trial 4	0
	Trial 5	1
	Trial 6	0
	Trial 7	0
	Trial 8	1
	Trial 9	1
	Trial 10	1
	#Successes	6

This time we get 6 successes for the same value of p. If we keep pressing F9 we see we can a series of results, sometimes the same, sometimes different. These simulated results seem to form some type of pattern for a particular p. For example out of 10 mice we often get 4, 5 or 6 who are successful. This makes us curious to examine the pattern further. Let’s repeat this experiment 50 times and keep the results. We could write down the results but it’s a lot easier to copy the set of formulae across a number of different columns so that we can store the results of many random experiments. Let’s do it 50 times. The results are shown on the following page.

We can use the spreadsheet to calculate and display the pattern of results over the 50 experiments, each of which comprised 10 trials. That is, we have repeated the 10 trial experiment 50 times. Note the interesting pattern which we have plotted in a histogram. There are relatively high frequencies for 4, 5 and 6 successes, a number for 2, 3, 7 and 8 successes, BUT none for 0, 9 and 10 successes. In fact the histogram in Figure 3 which depicts the frequency of observed successes, infers that it is pretty unusual that nine out of the animal trainer’s ten mice would successfully turn left when they hadn’t been trained! What are the chances that nine out of ten mice turned left *purely by chance* on this occasion? We could keep repeating the experiment and see if we *ever* get an occasion when we get 9 out of 10 successes. However, perhaps we should consider that the training was effective! How would we model effectively trained mice? Well, its unlikely to have been 100% effective (after all most educational programs have less than 100% pass rates) so perhaps it is effective to the tune of say 80%. This would equate to a p parameter of 0.8. We don’t know what the true p value is but it does seem that the observed data are not consistent with a p of 0.5.

FIGURE 3 THE MICE EXPERIMENT

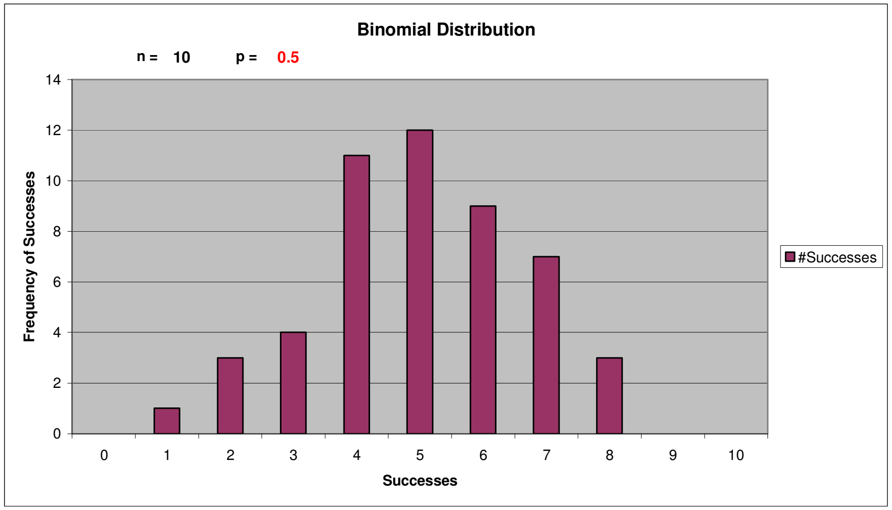
p	0.5
n	10

	Ex1	Ex2	Ex3	Ex4	Ex5	Ex6	Ex7	Ex8	Ex9	Ex10	Ex11	Ex12	Ex13	Ex14	Ex15	Ex16	Ex17
Trial 1	0	1	1	0	1	1	1	0	1	1	0	1	1	1	1	0	1
Trial 2	1	0	1	0	0	0	1	0	1	0	0	1	1	1	1	1	1
Trial 3	1	0	0	0	1	0	1	1	0	1	1	1	1	1	1	0	1
Trial 4	0	0	1	1	1	1	1	1	1	0	1	1	1	1	0	0	1
Trial 5	1	0	1	0	0	1	0	0	0	1	1	1	1	0	1	0	1
Trial 6	0	0	0	0	0	1	1	1	1	0	0	0	1	0	1	0	0
Trial 7	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
Trial 8	1	0	0	0	1	0	0	0	0	0	1	1	1	1	1	0	0
Trial 9	1	0	0	1	0	1	0	1	1	0	0	1	0	1	0	0	1
Trial 10	1	0	0	1	0	1	0	0	1	1	1	1	0	0	1	0	1
#Successes	6	2	4	4	5	7	5	4	6	4	5	8	7	6	7	1	7

	Ex18	Ex19	Ex20	Ex21	Ex22	Ex23	Ex24	Ex25	Ex26	Ex27	Ex28	Ex29	Ex30	Ex31	Ex32	Ex33	Ex34
Trial 1	0	0	1	0	1	0	1	0	0	1	1	0	0	1	0	1	1
Trial 2	0	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0
Trial 3	0	1	1	0	0	0	0	0	1	0	1	0	1	1	1	1	0
Trial 4	1	0	0	0	1	0	0	1	1	1	1	0	0	1	1	1	1
Trial 5	1	0	0	1	1	1	1	1	0	1	0	1	0	1	0	1	1
Trial 6	1	1	0	1	0	1	1	0	0	1	0	0	1	0	1	1	1
Trial 7	1	1	0	1	1	0	1	1	0	0	0	0	1	1	0	1	1
Trial 8	1	0	0	0	0	1	1	0	1	1	1	1	1	1	0	1	0
Trial 9	0	0	0	0	1	1	1	0	1	1	0	1	1	0	1	1	1
Trial 10	0	1	0	1	0	0	1	1	1	0	1	0	0	1	0	0	0
#Successes	5	4	2	4	6	5	7	4	5	7	6	3	5	7	4	8	6

	Ex35	Ex36	Ex37	Ex38	Ex39	Ex40	Ex41	Ex42	Ex43	Ex44	Ex45	Ex46	Ex47	Ex48	Ex49	Ex50
Trial 1	0	1	1	1	0	1	1	0	1	1	0	1	0	1	0	0
Trial 2	1	1	1	0	1	1	0	0	0	1	0	1	1	0	0	1
Trial 3	1	0	0	0	1	1	1	0	1	0	0	1	0	1	1	1
Trial 4	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0
Trial 5	1	1	0	1	0	1	0	0	1	1	1	0	1	1	0	1
Trial 6	0	1	0	1	0	0	0	1	1	0	0	0	0	1	1	0
Trial 7	1	1	1	0	0	0	0	0	0	1	1	0	1	0	1	1
Trial 8	1	1	1	0	1	1	0	0	1	0	0	0	0	0	1	0
Trial 9	0	1	1	0	1	0	0	0	0	1	1	0	0	0	0	0
Trial 10	1	1	0	0	0	0	1	1	1	0	1	0	1	0	1	0
#Successes	6	8	6	3	4	5	3	2	6	5	5	3	4	5	5	4

# Successes	Frequency for 50 Experiments
0	0
1	1
2	3
3	4
4	11
5	12
6	9
7	7
8	3
9	0
10	0



USING VBA TO EXTEND AND STREAMLINE THE SOLUTION

Our teaching approach over the last few years has been to focus on teaching things like the binomial distribution primarily through the means of a VBA simulation. Similar simulations exist on the Web written in Java like the excellent suite of programs produced under the *Visualization of and Experimentation with STAtistical Concepts* (VESTAC) Project at the Katholieke Universiteit Leuven. On reflection, however, we have come to the conclusion that jumping straight to a statistical simulation is too 'black-boxy', that is, it is graphically impressive and illuminating to a significant extent but, because no (spreadsheet) building blocks have been shown to students, it is impenetrable and intimidating for the average student. The approach we recommend here is a 2-stage approach where the ideas in stage-1 are explained through Excel sheet formulae up to a certain level, as expounded upon above. Then, when the basic ideas have been properly internalised, the leap is made to showcasing the click-button-and-fancy-graphics world of VBA. For engaged students, this transition can also provide an entrée into the power and flexibility of VBA programming. Barr and Scott (2008) give an account of using a suite of simulations written in VBA (including the mouse example) as teaching tools.

THE MOUSE EXAMPLE CONTINUED USING VBA

The use of VBA helps elucidate the example. For the VBA application we set n = 5 which makes the example more compact and allows us to more easily look at comparing the simulations of mouse behaviour under different p, but critically follow this by a comparison against what would be **expected** under an exact binomial distribution. In this stage 2 example the crucial distinction is then between what we call empirical distributions, that is those histograms simulated under assumed randomness with fixed p and those histograms derived directly from the frequencies obtained from the exact binomial probabilities assuming fixed p. To obtain the expected (theoretical distribution) we use the standard binomial distribution result $B(X, p, n)$

$$B(X = x\,;\, p\,;\, n) = \binom{n}{x} p^{x}\,(1-p)^{n-x}\ ;\ x = 1, \ldots, n \qquad (1)$$

In the stage 2 example, we can change p as before but, importantly, also change the number of experiments run. This allows us to demonstrate a key result, namely that as the number of simulations increases, the empirical distribution gets closer and closer to the theoretical distribution. The point and click buttons allow a staged and slick presentation by the teacher; the process should be:

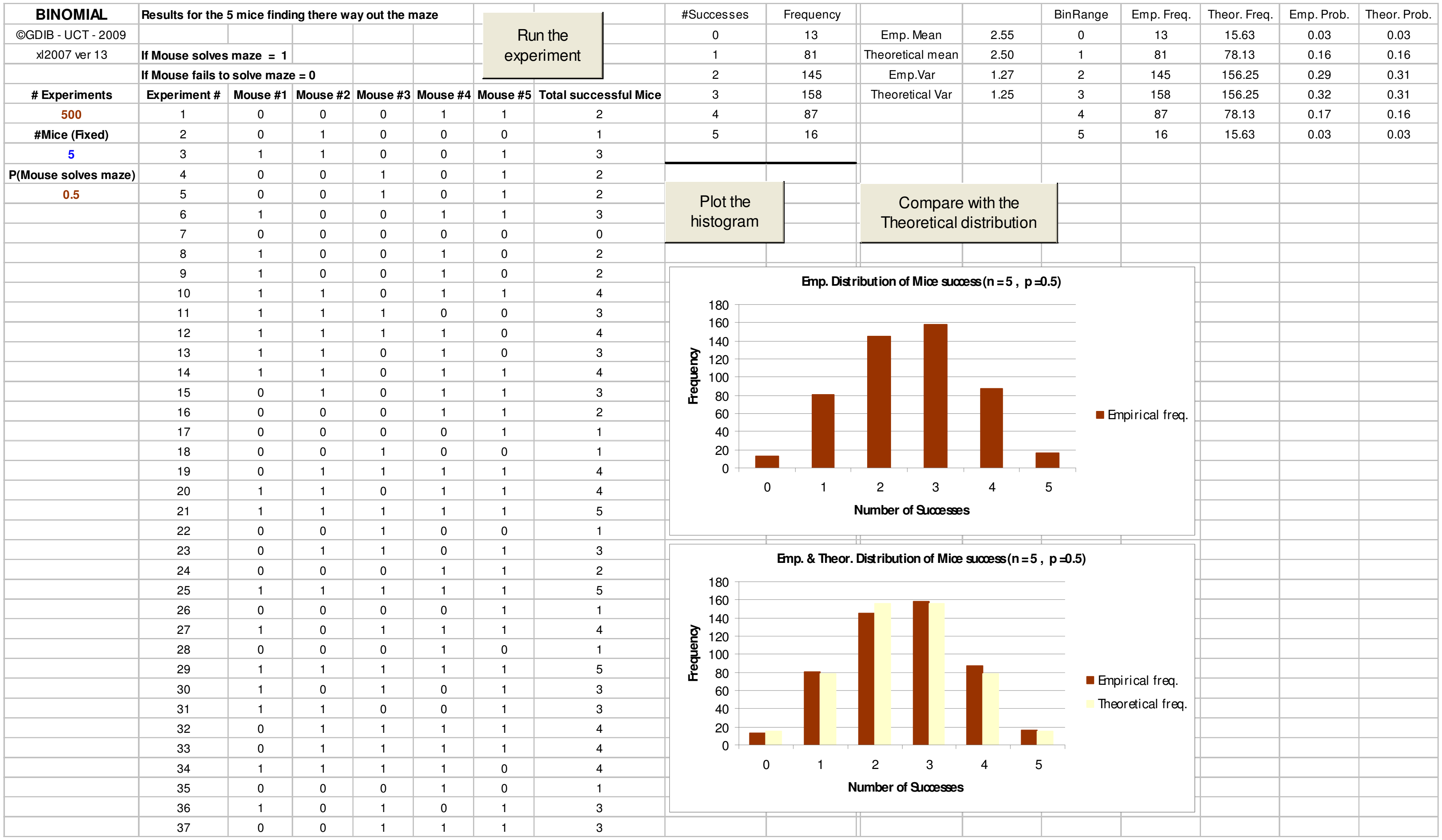
Step 1: Run the simulation with as many experiments deemed necessary and some value for p

Step 2 Plot the empirical distribution

Step 3 Plot the empirical and theoretical together

Then, by changing the number of experiments, we can demonstrate that the larger the number of experiments the closer the empirical distribution will get to the theoretical. We can also show that by changing p, the histograms become asymmetrical but the convergence criterion works as before. We can also in a parallel way discuss the notion of theoretical and empirical probability and frequency **(see Figure 4)**.

FIGURE 4 THE MICE EXPERIMENT WITH VBA



BINOMIAL	Results for the 5 mice finding there way out the maze						
©GDIB - UCT - 2009							
xl2007 ver 13	If Mouse solves maze = 1						
	If Mouse fails to solve maze = 0						
# Experiments	Experiment #	Mouse #1	Mouse #2	Mouse #3	Mouse #4	Mouse #5	Total successful Mice
500	1	0	0	0	1	1	2
#Mice (Fixed)	2	0	1	0	0	0	1
5	3	1	1	0	0	1	3
P(Mouse solves maze)	4	0	0	1	0	1	2
0.5	5	0	0	1	0	1	2
	6	1	0	0	1	1	3
	7	0	0	0	0	0	0
	8	1	0	0	1	0	2
	9	1	0	0	1	0	2
	10	1	1	0	1	1	4
	11	1	1	1	0	0	3
	12	1	1	1	1	0	4
	13	1	1	0	1	0	3
	14	1	1	0	1	1	4
	15	0	1	0	1	1	3
	16	0	0	0	1	1	2
	17	0	0	0	0	1	1
	18	0	0	1	0	0	1
	19	0	1	1	1	1	4
	20	1	1	0	1	1	4
	21	1	1	1	1	1	5
	22	0	0	1	0	0	1
	23	0	1	1	0	1	3
	24	0	0	0	1	1	2
	25	1	1	1	1	1	5
	26	0	0	0	0	1	1
	27	1	0	1	1	1	4
	28	0	0	0	1	0	1
	29	1	1	1	1	1	5
	30	1	0	1	0	1	3
	31	1	1	0	0	1	3
	32	0	1	1	1	1	4
	33	0	1	1	1	1	4
	34	1	1	1	1	0	4
	35	0	0	0	1	0	1
	36	1	0	1	0	1	3
	37	0	0	1	1	1	3

#Successes	Frequency			BinRange	Emp. Freq.	Theor. Freq.	Emp. Prob.	Theor. Prob.
0	13	Emp. Mean	2.55	0	13	15.63	0.03	0.03
1	81	Theoretical mean	2.50	1	81	78.13	0.16	0.16
2	145	Emp.Var	1.27	2	145	156.25	0.29	0.31
3	158	Theoretical Var	1.25	3	158	156.25	0.32	0.31
4	87			4	87	78.13	0.17	0.16
5	16			5	16	15.63	0.03	0.03

PUTTING ALL THIS TOGETHER: MAKING SENSE OF THE CO-EXISTENCE OF PATTERN AND RANDOMNESS

The interesting feature of the results of our 'mouse training' model is the pattern of results it revealed. Each time we repeated the 50 experiments, each consisting of 10 trials, we observed a different set of numbers but they appeared to keep a number of common features. This pattern became clearer the more times we repeated the experiment. The pattern was different from that of the numbers we drew out of the hat. What does this tell us about randomness? What causes the patterns? Remember we said the hat was our mechanism to ensure random selection, in other words to mimic the way data might present itself to us in an unpredictable way. All the numbers were mixed up in the hat and we drew them out in a way that meant no particular numbers were favoured or prejudiced.

What if we put different pieces of information in the hat (as distinct from numbers) into the hat, shuffle them and draw samples of them out the hat? Will the patterns related to different pieces of information all be the same? These pieces of information could be the results of the mice experiments that members of the class conducted, names of people or weights of people – there are a host of possibilities. Do all pieces of information associated with each person and which we could randomly sample from the hat have the same shaped histograms?

Well let's conduct our class experiment and find out: each student in the class should use the spreadsheet programs we have developed together to generate a (random) number between 1 and 50 and to run the mouse experiment and record the number of successful mice. They should write down on a piece of paper the required four pieces of information. We'll collect all these slips of paper in a large hat (not an electronic one this time), shuffle them thoroughly and then draw them out and construct histograms of the four different pieces of information. What shapes do we anticipate for each of these histograms?

We might not be surprised to observe that the random numbers have a flat, rectangular distribution. We also see that the 'successful mice observed' have the same shaped distribution as the one we saw repeatedly with our electronic mice running model. The heights may show one bell shaped histogram, or may have a hint of two humps of data, with the males being taller than the females. The 'names' may well show a few modes, depending on popular names and prevalence of language groups. Our 'randomising' hat has had the effect of giving us the data in random and unpredictable order, but the distinct patterns associated with each different piece of information have been preserved, and are revealed as we have access to more and more data.

We might not be surprised to observe that the random numbers have a flat, rectangular distribution. We also see that the 'successful mice observed' have the same shaped distribution as the one we saw repeatedly with our electronic mice running model. The heights may show one bell shaped histogram, or may have a hint of two humps of data, with the males being taller than the females. The 'names' may well show a few modes, depending on popular names and prevalence of language groups. Our 'randomising' hat has had the effect of giving us the data in random and unpredictable order, but the distinct patterns associated with each different piece of information have been preserved, and are revealed as we have access to more and more data. In fact, our reconstructing of the data into histograms reinforces two facets of random variation. On the one hand, it is reflected in the unpredictable way we frequently encounter (information in) life (stocks vary on the stock exchange, increments of growth of children, number of cars on highway at a particular time, etc). But, perhaps paradoxically, randomisation also provides the best mechanism to uncover the true pattern of an unknown measurable (eg household income). Selecting data (sampling) randomly ensures we have the best chance to see as broad a spectrum of the unknown pattern of data as quickly (efficiently) as possible!

CONCLUSION AND FURTHER WORK

The spreadsheet has many powerful didactic facets. It provides a flowing, dynamic model which links cells in a transparent way. At the first level it provides a two dimensional, visible, matrix-like calculating machine where at the press of a button the whole matrix may be recalculated. This can be used to simulate samples with different underlying distributions. Randomly sampled bits of information have statistical distributions which reflect how these bits of information are generated and what attributes of life they reflect. While randomly selected numbers will reflect a uniform distribution, heights of people will reflect a normal distribution, the results of a simple binary experiment a binomial distribution, and names across different cultural communities often a multi-modal distribution.

A key insight for learners is that although each attribute is selected from the hat of our experiment *randomly* the *patterns* or *distributions* of the attributes can be quite different. While an individual randomly generated/ encountered outcome is not predictable, the *pattern* of many outcomes *is* often predictable. This experiment thus constitutes a very powerful mechanism for learner differentiation between the concepts of randomness and the underlying pattern of the attribute itself. Our experience with the extensive use of Excel for first year statistics courses has led us to conclude that a pure VBA approach may be somewhat intimidating for learners and teachers alike so we have adopted a 2-stage approach, where Excel is used firstly in a straightforward formulaic approach, where the formulae are first explained to students in a flowing logical sequence and thereafter the jump to a VBA point-and-click automated approach is made. VBA based spreadsheets allow more flexibility and sophistication and in particular allow automated comparisons showing students how empirical distributions converge on theoretical distributions. We believe this approach is a step forward in the teaching of foundational concepts in first year Statistics, such as randomness and distribution, which are often poorly understood, even by statistics graduates.

The approach presented in this paper is both experiential and visual in that the learners generate their own data and build up understanding of concepts by seeing them unfold through experiments involving this data. This approach is in contrast to one which relies on learners digesting tracts of theory and then applying this theory to ‘static’ data sets. As such it is thought to be potentially of benefit to learners for whom the learning experience is made more onerous by language difficulties (e.g. non-mother tongue learners). The authors plan to explore ways of evaluating the approach outlined in this paper, focussing particular interest on this group of learners.

REFERENCES

Arganbright, D. (1984). The Electronic Spreadsheet and Mathematical Algorithms. *The College Mathematical Journal*, Vol 15, (pp 148—157).

Baker, J. E. & Sugden, S. J. (2003). Spreadsheets in education. The first 25 years. *eJournal of Spreadsheets in Education*, vol *1*(1), (pp 18-43). {Available at http://www.sie.bond.edu.au/, visited 25 August 2010}.

Barr G.D.I. & Scott L. (2008). A new Approach to Teaching Fundamental Statistical Concepts and an Evaluation of its Application at UCT, *SA Statist. J*., 42, 143–170.

Black, T. R. (1999). Simulations on Spreadsheets for Complex Concepts: Teaching Statistical Power as an Example, *International Journal of Mathematical Education in Science and Technology*, vol 30(4): (pp 473—81).

Jones, K. (2005). Using Spreadsheets in the Teaching and Learning of Mathematics: a research bibliography, *MicroMath*, vol 21(1), (pp 30-31).

Mills J. D. (2002). Using Computer Simulation Methods to Teach Statistics: A Review of the Literature, *Journal of Statistics Education*, vol 10 (1).

Soper, J.B. & Lee, M.P. (1985). Spreadsheets in Teaching Statistics, *The Statistician*, vol 34, (pp317-321).

Sutherland, R. (2007). A Dramatic Shift of Attention: from arithmetic to algebraic thinking. In; J. Kaput, D. Carraher, & M. Blanton (Eds.), *Algebra in the Early Grades*. Routledge.

Visualization of and Experimentation with STAtistical Concepts (VESTAC) Project at the Katholieke Universiteit Leuven.