ISSN 2250-3153

Teaching Computational Thinking in Probability Using Spread Sheet Simulation

Anand R*, Manju M*, Anju M Kaimal*, Veenaa Deeve NV*, Chithra R*

* Center for Computational Engineering and Networking, Amrita School of Engineering, India

Abstract- Probability is an effective mathematical tool to calculate the likelihood at various instances rising as a effective research area. The programmers are continuously trying to accomplish full automation system in real world applications which can be made possible only through the effective calculation of probability in various circumstances. Though probability is very much important, the students and teachers are not obtaining the full experience of theory because of the lack of proper tools for exploring it. This paper has concentrated on bringing up some basic probability terms and implementation in Microsoft Excel which could be understood by all from the scratch.

Index Terms- Spread sheet simulation, Probability Distribution, Higher Order Moments, Birthday Problem, and Data Analysis

I. INTRODUCTION

The Probability and Statistics is playing a main role in various fields like major business problems. So the need of understanding the probability should be given higher priority than marks. But the students are not getting the chance to actually taste the real probability theory, so this paper has tried to bring a limelight on the various basic probability terms by making them understand the probability concepts easily. It will be good to start with the lines of Arthur Benjamin from his speech titled "Teach statistics before calculus!" in the year 2008. Being a mathematician he states that only very few people actually use calculus in the conscious and meaningful way. On the other hand, probability plays a more important role in real life. The mathematics behind probability can be explained as a decision at risk, a reward, and randomness which helps in understanding data that helps the individuals to live the life with lots of fun, as the probability is the basics for the games and gambling when it is taught properly. He also states that understanding probability will avoid the economic mess in nations and can also help in analyzing the trends for predicting the future.

In every day's life in some way or other the people have to encounter probability. In the fascinating world of statistics, the individual who can explore the chance for occurrence of anything in their life prior itself can achieve his goal in a much easier manner than expected. Each individual have to face lots of situations which are totally uncertain. Probability is basically used for predicting this uncertainty and fit 'some' certainty to it[11]. This kind of prediction of uncertainty will be useful when a person is put forth in a decision making situation. But, the current education system has made students to consider the probability and statistics as a kind of mathematical formula and have failed to make them experience the need for it in the real life. The subject of probability will be more useful to the students if it is taught in a right manner.

Understanding the fact that the students are not taught probability properly, John Garfield did an extensive research on this field and wrote an article named" How Students Learn Statistics" .After his research, he concluded that there are large group of students who lack a lucid understanding about the subject .This proves that there is a massive gap between the theoretical and practical understanding of the subject. He also points that the students who are answering most of the question asked in an examination fails to use the same in the real time situations [1]. This proves that the traditional teaching system is not actually effective enough to help the students to test the real understanding or to solve any real life problems, which seems to be very pathetic.

It is not known to many teachers that the process of listening to lectures and solving problems in the probability will not help the students to absorb the essence of the subject. It is because even if the probability teachers are asked about their expectation from the students after one year of class will not be the one to compute a standard deviation by hand or the ability of converting normal variables to standard normal variables or to perform the computation of expected values by looking up their probabilities on the z table. But the actual understanding of the students regarding the basic statistical concepts and ideas is not helping to become statistical thinkers or to evaluate quantitative information [1].

In general, students learn by fitting the restructuring process of the new information provided to them in the class with their own cognitive framework. According to the thesis on Seventh Grade curriculum in probability by David Nganga Njenga, the students are actually facing several barriers in understanding the probability and the foremost barrier among them is the limited experience with uncertainty. When an individual is put in amid situation relating to chance and uncertainty of future events, a difficulty is faced in assessing the risk, especially when the situations get more complex [6]. For example, the sentence which we come actually come across each and every day in the media about the weather report say that there is a sixty percentage chance of rain on the next day needs some sophistication in order to articulate its meaning.

The belief in determinism and a tendency to look for a cause is an important misconception in student's life [6]. For instance, while playing a game of chance where a coin is being tossed, children often think that, if they get a tail in one event, then the result in the next toss is expected to be head. This is not randomness because while tossing a coin anything is possible, including getting all heads when a coin is tossed ten times.

One more challenge which makes the students to fail in understanding the probability is that the probability cannot be measured as easy as distance, weight or time. When the die is tossed there is nothing that can indicate with it except the fact that the probability of a particular face is $\frac{1}{6}$. Even after the die is tossed, there is no way we can measure the probability of what just happened. These factors post a serious challenge in understanding the probability.

Lots of mathematicians and psychologists have tried to address the issue of the method of teaching statistics effectively to the students. The International conference on teaching statistics was started in 1982 and it helped to form an educational committee of diverse choice of location which brings together statistician all over the globe from different fields to share their ideas. Thus various activities have been framed to make the students to get the feel of what really probability means. Even though lots of money was spent for it, it incorporates considerable knowledge for the students. People also thought of interactive game based software that helps students to learn the subject with fun. Currently, we have lot of free Java applet based games which helps the students to visualize the situation.

Even after all these continuous effort, student are still facing problem in understanding the real life situations. This is primarily due to the gap between theoretical probability and the experimental probability as said earlier. The theoretical probability provides clear insight of different section on probability with the aid of explanations and formulas. The practical activities such as games help to get the feel of probability. But still there is a big gap between the numbers obtained with the aid of theoretical knowledge and these software game based activities. For instance, almost all the students can calculate moments of a distribution using formula but very few can actually explain what that number signifies. Most of the game based activities were not able to address this issue.

It is found that the simulation of the outcomes will be one of the best ways proposed to make students get the feel of the different numerical value they obtain by hand calculation; also by building mathematical models that approximate observation in the real world, students are able to investigate chance events and develop their understanding of probability concepts.

The above simulation model can be used to solve the issue of Monte Carlo Simulation. The time independency and repeated sampling of the input data will provide the output in the form of the probability distribution. There are more number of commercial simulation models like SLAM, Arena, Extend, SIMSCRIPT, or GPSS are available which needs lots of experienced to work with it and hence is not suitable for the students [2].

Spread sheets, on the other hand, offer many pedagogical advantages for learning simulation. Taking into consideration most of the students have expertise in using Spread Sheet, Students find easy to understand the simulation. The students can work on Excel without any guidance as it is a user-friendly learning system. The Excel can serve as such a valuable and versatile tool that helps the students to maximize the potential of their learning skills. Added to that, Spread sheets allow users to quickly and easily develop visualizations of data, and gain useful insights that typical simulation output does not provide [2]. Also with the aid of dynamic updating facility, new results can be immediately obtained as data are changed. This feature is of particular interest from a teaching perspective as Excel provides dynamic updating of values so that the results for various values can be viewed. Similarly the Graphs and charts are used so that new results are obtained. Excel does not expect the user to know any difficult programming language also to work with it. Since clicking on any cell in Excel provides the formula used for simulation, students find it very easy to understand Excel simulation. So Excel is an ideal tool for teaching statistics to students.

Motivated by all these literature reviews [1-6], the work is done for dozens of simulations in excel which are primarily intended to understand primary probability concepts. These experiments are primarily intended to minimize the misconceptions that both teachers and students may have while dealing with probability. Also, we have attempted to explain it through the simulated outcomes from popular probability distribution; calculated important parameters of the distribution without using any traditional formulas and have explained the numerical result obtained in a logical way. Thus we have tried to reduce the gap between theoretical and practical probability. The various examples and simulation suggested in this thesis hopefully will increase the quality of instruction in probability and ultimately motivate the students in this important field of mathematics.

The first step taken in this work is to understand which area students really lack in understanding. Hence a preliminary survey was taken in which 65 students were assessed by asking simple statistics questions to them. Section II elaborates about the survey and the inference from it. From their answers it is found that most of the students lack a clear picture about randomness. It was also observed that most of the students find it difficult to understand the physical meaning of moments. An attempt has been made to address these issues. Section III focus on simulating coin toss problem in Excel and brings a limelight on all the basic statistics concepts. The next two sections explain data analysis; the primary focus is on understanding discrete and continuous Probability distribution with the aid of excel. Birthday problem is discussed and analyzed in Section VI. Finally, the work is discussed and concluded in last two Sections.

II. PRELIMINARY SURVEY

Before starting the survey the need of the areas in the probabilities in which the students expect to be much clear in

their understanding criteria. So a case study among a group of students is conducted by asking a series of question that comprised some topics which they have studied in their university syllabus. This study includes the process of testing the understanding of the students on the concepts of statistics and is carried out at the department of Centre for Excellence in Computational Engineering and Networking (CEN), Amrita University in 2012. 70 students volunteered to participate in the test out of 260 students; the test was designed to examine student's practical understanding rather than procedural abilities. The probability course includes basics of both probability and statistics. Although the course manual is filled with formal definitions and theory, and many references to N.P.Bali, B.S.Grewal "contemporary Statistics" textbook, it is designed in such a way that the students are able to pass the course simply by knowing the routine processes, and not necessarily understanding the theory.

In order to access the skill of each student, 15 questions are framed. The questions are framed in such a manner that the only students who have good idea about uncertainty of outcomes, proper physical understanding of mean, variance and moments of the given distribution without aid of formulas can only answer them properly.

A. QUESTION PAPER
The survey Question paper is shown below:

Assessment
Preliminary survey

Name:_____
Date: _____

1. I am tossing a fair coin 10 times. Which outcome is more likely to occur?
    a. I get exactly 5 heads
    b. I get somewhere between 3 to 7 heads
    c. I get no heads

Reason:_____

2. I am tossing a coin 10 times. Which of these will never occur?
    a. I get 0 heads
    b. I get 10 heads
    c. Both a and b
    d. None of these

Reason:_____

3. There is a small rural hospital R and a big urban hospital U. I am interested in percentage of male birth on a particular day in both these hospitals. Which hospital is more likely to have 70% male birth on a particular day?
    a. R
    b. U

Reason:_____

4. I am tossing a coin 1000 times. Can I say for sure that I will get at least one head.
    a. Yes
    b. No

Specify Reason for the answer:_____

5. How will you generate 1000 random numbers for the given probability distribution ?

X	1	2	3	4
P(X=x)	0.2	0.3	0.4	0.1

_____

6. If the probability distribution is not given, is there any other alternative ways to generate random numbers from that distribution

_____

7. Assume that there are 35 people in a party. What is the probability that two of them have birthday on the same date?(Approximate range)
    a. 0.01 - 0.1
    b. 0.1 - 0.3
    c. 0.3 - 0.7
    d. 0.7 and above.

8. I have a Random variable X from binomial distribution, with n=3 and p=0.5. Now by formula,

$$E(X)=np=3\times\frac{1}{2}=1.5$$

i. What is the physical meaning of 1.5?

_____

ii. Without using formula, is there any other way to find $E(X)$ if you are given 1000 outcomes of X.
    a. Yes. Method:_____
    b. Cannot be determined

9. I have a Random variable X from uniform distribution. Given a=0,b=1. I am providing the following information

$$E(X)=\int_a^b xf(x)dx=\left[\frac{x^2}{2}\right]_a^b=\frac{a+b}{2}=\frac{0+1}{2}=0.5$$

$$E(X^2)=\int_a^b x^2 f(x)dx=\left[\frac{x^3}{3}\right]_a^b=\frac{a^2+b^2+ab}{3}=\frac{0+1+0}{3}=0.33$$

$$E(X^3)=\int_a^b x^3 f(x)dx=\left[\frac{x^4}{4}\right]_a^b=\frac{(a^2+b^2)(a+b)}{4}=\frac{1(1)}{4}=0.25$$

i. In your own words, explain what the value of 0.5,0.5 and 0.25 for the moments signify?

_____

_____

ii. Without using formula, is there any other way to find $E(X), E(X^2), E(X^3)$ if you are given 1000 outcomes of X.

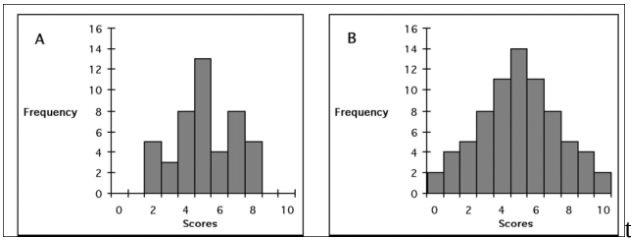
a. Yes. Method:_____

_____

b. Cannot be determined

10. Which of the following distribution shows MORE variability?

a. A
b. B



the

Circle the reason which made you to select answer

a) Because it's bumpier
b) Because it's more spread out
c) Because it has a larger number of different scores
d) Because the values differ more from the center
e) Other (please explain) : _____

_____

11. Assume that I have 100 data of height of a class which follows normal distribution. I am interested in analyzing two cases. In case 1, am considering 5 people height as a single unit and finding local average of each unit and finding the distribution. In case 2, I am making 10 people height as a single unit and finding local average and plotting the distribution. Which case will have minimum spread?

a. Distribution in case 1
b. Distribution in case 2

Reason for the answer:_____

_____

12. I have two independent random variables X1 and X2. Let Y1 = X1+X2 and Y2 = X1-X2. Then which is true?

a. Variance of Y1> Variance of Y2
b. Variance of Y1< Variance of Y2
c. Variance of Y1=Variance of Y2

13. Is there any relation between exponential distribution and Poisson distribution?

_____

_____

14. In your statistics class, you learn a lot about distributions like uniform, normal etc and their properties. What is the use of learning about these distribution(apart from getting marks)

_____

_____

15. List some of the topics in probability which you find difficult to understand?

_____

_____

III. RESULTS

The current probability textbooks explain clearly about randomness and uncertainty but it does not give clear picture of it. This lack in the visualization of uncertainty leads to lot of misconception about basic ideas and concepts. According to the observations made on the answers of the 70 students it is observed that 14 students are unable to answer the question at all. Most of the student answered question 1 correctly by choosing option b which implies that students have fair idea that small samples are more likely to deviate from the population than the large samples. But only 35 of them have answered the question 3 correctly and out of which only 10 of them are able to give correct explanation on the reason for giving that answer. Thus from the observations made in which the students have answered very low number of questions, the fact that the students cannot transfer their understanding to a similar real-world context is clearly shown.

The survey clearly gives the kind of understanding the students are having on the probability and statistics. The students use a model of probability that lead them to make yes or no decisions about single events rather than looking at the series of events. Students have a misconception that if the chance of occurrence of an event is 70 %, students misinterpret that the event will happen for sure. Also that 30% chance of occurrence of an event is interpreted that the event will not happen at all. The 50% chance is a tricky case where it is interpreted as one cannot say the either way [1]. 95% of the students who underwent the test went wrong in third and fourth questions,

which suggested that the student have biggest misconception in understanding the result obtained from probability model. The first 4 questions are designed to see if the students have clear visualization of the uncertainty situation.

The next two questions were asked to know how much students are exposed to experimental probability. The student's knowledge should get restricted to just some formulas. A lot of knowledge has to be obtained by 'doing' statistics, not by just 'knowing' statistics. .But the current education system doesn't give importance to practical lab sessions on probability. So almost everyone were clueless while answering Question 5 and 6.The first step to practice probability is to generate the random numbers the random numbers for the given distribution. This idea is very essential to carry out probability experiments using computers. The next session clearly explains how random numbers can be generated for the given distribution. Also, there may be lot of cases where the probability distribution is unknown. For that case, random numbers can be generated using cumulative distribution. In section V-C, the exponential distribution is simulated using this method and it will give the readers how to generate random numbers for any distribution.

Mean variance and higher order moments are some of the very important parameters which can characterize a distribution. The current curriculum and the examination board have trained students to compute these parameters for almost any popular distribution. But students don't have any insight of what the numerical result means. Student may have answered items correctly on a test because they know what the expected answer is, but still have incorrect ideas. Using Questions 5 and 6, we tried to access what the students have understood about these parameters.

The answers suggest that none of them have clear insight of the method to compute these parameters without aid of formula. 5 of the students have written that average of all outcomes gives $E(X)$. But they too don't have any clue how to calculate higher order moments. The pathetic situation is that none of the students had idea why some probability distribution like binomial ,Poisson, normal are taught to them in the curriculum. This gave a motivation to focus primarily on explaining these concepts in this paper.

Variance is an important parameter in the concepts of the probability which gives a clear idea about how the data are spread out. Hence for analyzing any distribution in a better way the sound knowledge in this parameter is required by all the students. Question 8 and 9 aimed at finding how comfortable students are in visualizing the variance. Although most of them answered the question correctly, the reason given by them for choosing those answer is not convincing.

Question 7 and 13 were asked basically to test how much understanding they had about theoretical probability. Even though those questions were straight forward, only 4 students were able to attempt that question, which proves that students learn only by 'doing' probability. The feel of the technology can be obtained only by doing things practically in lab. Particularly for question 7, which is also popularly known as 'birthday problem', most students choose first or second option as the answer. They think that probability of getting two persons birthday on same day for 35 people is very less. But the actual answer is option c. Its probability value is quite high. Majority of the people will think that high probability can be obtained only when no of people is more than 300. But, in reality, if n(number of people in the room) increases to even 100, probability of getting 2 person getting birthday on the same day is almost 99%. The tricky question was framed to understand the mindset of the students. Even though four of the students approached the problem correctly, they were unable to end up in exact numeric answer because of the heavy computation involved. In excel, this problem can be approached in a much simpler way and a detailed picture is provided in Section VI.
None of the students are aware of the reason why they learn distribution like Uniform, Normal was taught in the curriculum. This is really pathetic which proves that the current education scenarios haven't provided any insight to the practical problems to the student's .Towards the end of next section, this issue is addressed in detail. Also from Question 15, we got a feedback from students to get an idea about topics student find tiresome. Most of the student found it difficult to identify which distribution to use for a given problem. Also they find it difficult to deal with continuous distribution

IV. DISCUSSION

The study helped in understanding the area where more focus is needed. So here a small work is done that can help the students to unlearn the concepts which they have studied already by removing the misconception in the basic probability. The need for making the students to feel what randomness or uncertainty means is very important; also to make them understand that the probability is just to fit some number to uncertainty but it never leads to certainty; Probability is a prediction; Statisticians are not magicians or God.

The only way to make the students to visualize probability is by simulating the outcomes of the experiment. It could be well understood that the curriculum have made most of the students strong in theoretical probability. Majority of students possesses good understanding about Random variable, probability distribution, Analytical calculation of moments etc. But they lacked the idea of generating random numbers for a given distribution and obtaining the parameters. A simple coin tossing experiment is taken and has tried to explain basic probability concepts in a simpler way. The implementation part also includes some commonly used probability distribution and explained concepts which are not normally focused in textbooks. All the results of the experiments are presented graphically and have made interactive so as to make students learn topics with fun.

Instead of viewing the problem in a traditional way such as giving a formal definition and formulas, the focus is given fully in making the students to familiarize themselves with the terms like Random variable, probability distribution, mean, variance and moments. There are lots of resources which are actually already giving some idea about these topics and hence this work

will not concentrate more on the theoretical part. Also, in Excel based experiments, the process of highlighting the algorithm which is used to implement a distribution in excel is explained very well in a step by step manner; hence it is not discussed here. But towards the end of the discussion part, the link which has the excel file containing our work has been given. Even the step by step explanation are given clearly in the excel sheet itself which can help the individuals to get a clear understanding about the topics discussed.

V. COIN TOSSES EXPERIMENT

Let's take an illustration on Coin toss experiment in Excel which will help the students to take a tour on all the basic parameters defining the probability of a particular event. Let us consider an unbiased coin where {H,T} be the possible outcome which constitutes the sample space, where H denotes the occurrence of Head and T denotes the occurrence of tail. Consider the occurrence of head as the favorable event. The Random variable X denotes number of heads.

Let us simulate N=2000 outcomes of coin toss experiment in Excel. In each of the simulated outcome, X can take either 0 or 1, where both the outcomes are equi-probable.

The above situation can be simulated with the aid of random number generation. The random variables are generated using the predefined function "RAND()" which will actually generate the uniform random variables between 0 and 1.Since that there is a need for generation of the values in such a way that, 50% of values generated should be 0,and the rest 50% percentage of values generated should be 1.Now, if we consider values between 0 and 0.5 (given by RAND() function) it will surely constitute 50% of the total random numbers generated. Similarly values between 0.5 and 1 form **5**0% of the values**.** This comes from the property of uniform probability distribution.

The generation of 2000 random numbers is considered to be the first step. There is a need for generating the random numbers between 0 and 1 in such a way that to generate the numbers 0 and 1 with probabilities 0.5 and 0.5 respectively. This can be accomplished as follows: ={=IF(J3<=0.5,0,1))},where J3 is the excel cell location having the random value.
After simulating the outcomes, the first task is to find the number of times the head has occurred. In other words, it can be said as the probability distribution of X. In Excel, the ÇOUNTIF () function actually helps in counting the number of cells in a range, that meets a given criteria. Using the formula

{=ÇOUNTIF($K$3:$K$2002,0)},

it will be easier to count the number of outcomes of X which takes the value 0 can be found considering the simulated outcomes present in the range K3 and K2002. In the same manner, the number of occurrence of X=1 can also be found using the formula {=ÇOUNTIF ($K$3:$K$2002,1)}.

Since we are generating outcomes with the aid of random numbers, the outcome changes each time the excel sheet is refreshed and as a result we get different set of 2000 outcomes each time. In one of such outcomes, occurrence of 0 is 1008 times and occurrence of 1 is 992 times. So the probability of X taking value zero is obtained as 0.504(Not exactly 0.5) and probability of X taking value 1 is 0.496. A bar chat in which X axis represents the value X takes (0 and 1) and Y axis represents their probabilities. This bar graph represents the probability distribution of X. The simulation has been made interactive with the help of scroll bar option available in Excel. Using scroll bar, the value of n can be changed, where n denote the number of outcomes. Fine tuning of n with the aid of scroll bar changes the number of random numbers simulated. In the work, n can be changed from 10 to 2000. It can be observed that when n=10, there are chance that X=0 occurs just 2 or 3 times. There is less chance for equi-probable output as the value of n taken is as low as 10. This can be very well understood from the bar chart. Both the bin will have unequal height. But when we increase n, we can observe that both the bins approach to same height. So instead of giving lot of theoretical explanation, using a single bar chart, we have tried to effectively transfer the information to the students.

The next step is to give a clear physical meaning of E(X) and also the higher order moments. In general, by formula it is assumed that X takes discrete values and its higher order moments are given as

$$E(X)=\sum_x x.p(X=x)$$

$$E(X^2)=\sum_x x^2.p(X=x)$$

$$E(X^3)=\sum_x x^3.p(X=x)$$

$$\text{M}$$

$$\text{M}$$

The students are already familiar with these moments for a given distribution, but when they are asked to deal with a real time data they are feeling difficulty to do so without formulas which could be easily addressed in the following experiment.

Expectation of X is defined as the likely value that the random variable X can take which actually means the average of all the outcomes. So if the average of all 2000 simulated outcomes of the Random variable X, the expected value of X, $E(X)$ can be found easily. In excel, there is an inbuilt command called AVERAGE() which can be used for the purpose. In one of the simulation sets, obtained the value of $E(X)$ is 0.4985. Theoretically, using formula, the calculated expected value of X, $E(X)$ will be 0.5, but in real time, one can never reach this value. When value of n is increased more, better precision which will be near to 0.5 can be found.

A similar way has been adopted in order to find $E(X^2),E(X^3),\text{L}$ *etc.* In order to find $E(X^2)$, the value of X is squared and then the average is calculated. In this experiment, since X takes only 2 values 0 and 1 and squaring

them leads to same number, $E(X^2)$ will have the same value that of $E(X)$. But this is not the case with other distributions and $E(X^2)$ will have different value when compared to $E(X)$. A similar way need to be adopted to obtain $E(X^3)$. The cube of the value of X is calculated and then finds average is obtained for that. Thus one can easily give a clear insight of what these moments actually conveys in a distribution.

Readers will have a doubt what is the use of knowing how to calculate $E(X)$ and higher order moments without formula. But this lucid understanding is the first step to apply any statistical concept in real time data. In data analysis, the first and foremost area of interest to the statisticians is where the data value is concentrated and what the spread of data values is. The formal question can be answered by calculating $E(X)$ and the latter question can be answered by calculating variance of X say $V(X)$, for which the value of $E(X^2)$ is needed. Since $V(X) = E(X^2) - [E(X)]^2$. In those cases, we will not have readymade probability distribution table to calculate $E(X), E(X^2)$ So having the knowledge of what $E(X)$ and $E(X^2)$ values conveys and how to calculate them without formula is very essential in real time data analysis.

In the statistics curriculum, the students would have read about distribution like Uniform, Poisson, Normal, etc. What is the use of learning these distribution and their properties in real time? There is one more important step in real time data analysis, that is, data fitting. When the students are provided with raw data, it will difficult to analyze the data from the scratch as it is tedious and time consuming. So what is normally done is to fit the raw data into any of the known distribution. The advantage is that, since the properties of that distribution is already known, it becomes very easy for analyzing the data.

When analyzing data, in most cases, just plotting the histogram of the data doesn't help to decide which distribution has to be fitted. The higher order moments is widely used to fit a distribution to real time data. For most of the real time data, just by finding 3 higher order moments, say $E(X), E(X^2), E(X^3)$ the data can be fitted into any of the known distribution. How the data fitting is done with the aid of moments is out of context for this paper. What the work is focusing is that given a raw data, students must build the ability to calculate $E(X), E(X^2), E(X^3)$ without formula. Understanding this makes the student to open themselves to the various difficulties that they were facing so far and hence it is believed that this knowledge helps the student to apply statistics to at least simple problem they encounter in day today life.

In data fitting, the trial will be fully focusing on the steps to fit the unknown data to any of the known distribution. So the next two sections are dedicated for the simulation of some common known distribution in Excel and for analyzing the parameters of the distribution which will help students to get comfortable while handling real time problems.

VI. DISCRETE PROBABILITY DISTRIBUTION

The distribution of a random variable X which is discrete is termed as discrete probability distribution. Some of the popular discrete probability distributions taught in the curriculum are Binomial distribution, Geometric distribution, Poisson distribution. The work has concentrated more on the steps of how to simulate these distributions in Exceland plot the histogram for the same.

A. BINOMIAL DISTRIBUTION

Binomial distribution is a kind of discrete distribution in which there will be a series of fixed number of independent trials (n) leading to only two outcomes like success and failure. In Binomial distribution, the probability of success is considered as p and the probability of failure is given as q which is equal to 1-p.

In excel implementation; there is a need to generate M number of outcomes say 2000 from binomial distribution. For every event in M outcomes the sub task is to find the number of successes for the given n number of trials. This is accomplished through *what if analysis* in excel which is generally defined as the process in which changing the values in a single cell affects the actual formula in our worksheet. Here we are considering a small sequence of trials(say 10) and based on the outcomes we are calculating the number of successes which in turn can be extended to make changes in the actual random variable sequence of M. Thus the value of X can be calculated based on the outcome of our sub-process of finding the number of success.

When n=10, X takes values from 0 to 10. The probability of occurrence of the each of the 11 discrete values of X can be obtained with the aid of ÇOUNTIF () function in excel, which was explained in detail in previous section. The bar chart of the distribution is drawn in excel. Also $E(X), E(X^2), E(X^3)$ is computed from outcomes as explained in Section III. It can be seen that for the given n, the parameters of the binomial distribution depends on p. If p value is large, say 0.8, there is more chance of getting success in n trials and so X takes value between 7 and 10. So the probability distribution is skewed to the left. Also, since X take large values, $E(X), E(X^2), E(X^3)$ will also have large values. On the other hand, if p has small value say 0.2, there will be lesser number of successes and so the distribution skews towards the right and also $E(X), E(X^2), E(X^3)$ will have low values. If p is taken between 0.4 and 0.6, we get approximately symmetric Gaussian curve.

The excel sheet is simulated with binomial samples and is made so interactive that the students can change the value of p and observe the change in distribution parameters. Also using scroll bar provision, the students can change the number of outcomes

(N) that is needed to be simulated. If N is small, say 10, we will not get perfect shape for the distribution but if it is increased as much as 1000 or 2000, the distribution takes any of the three mentioned shapes based on the value of p.

B. POISSON DISTRIBUTION

The Poisson distribution is one of the discrete distributions where individual event occur at random in a given interval where the interval can be either time or space. The rate of occurrence of the event is given by the value of λ (lambda). The Poisson distribution is different from other kind of distributions because it does not involve a series of trials. It is unique as it models the number of occurrence in a particular interval. The random variable X is defined as the number of success in a given interval.

The Binomial distribution which is discussed in section IV-A can be said as a special case of Poisson distribution. If $X \in B(n, p)$ and if n is very large and p is very small then X can be approximated to $X : P_o(np)$ where $\lambda = np$. For example n taken as 100 and p as 0.02, then binomial distribution gets approximated to Poisson distribution. Using this idea in mind it can be extended from the previous implementation of Binomial distribution in Excel to the Poisson distribution by changing the n as large number and p with a small value.

The probability distribution table is obtained with the aid of ÇOUNTIF() command and the distribution is plotted. In Poisson distribution, since the probability of success p is very low, the distribution always is right skewed and so the distribution is always a decaying curve. How sharp is the decay is decided by the value of λ which is directly proportional to n and p. The distribution is plotted in excel and is made interactive where the user has provision to decide the number of outcome to be simulated (N)

C GEOMETRIC DISTRIBUTION

In this discrete distribution we have series of independent trials and we get only two outcomes say either success or failure and we are interested in number of trials until we get our first success. If X is defined as a random variable from Geometric distribution we can denote X as $X \in Geo(p)$ where p is the probability of success.

The distribution is generated in Excel using almost the same procedure followed for binomial distribution. There are some slight changes that need to be incorporated. In binomial distribution, the numbers of trails are fixed, say 10. But in geometric distribution the experiment has to be repeated till first success is encountered. So theoretically, for the worst case, infinite trials are required. But in Excel implementation, the trial is limited to 100. Here in the sub process, instead of counting the number of success, the numbers of trial are counted till the first success is obtained. It is not so simple to obtain the 100 trials for each simulation and hence the '*what if analysis* tool in Excel is used to address the problem. The Random variable is simulated following almost the same method used for binomial distribution.

The distribution is plotted in the bar chart. The scenario follows exactly the opposite when compared to binomial distribution. If p is large, then within less number of trials success can be achieved and so X takes small values, which makes the curve skewed towards the right. Added to that, the value of $E(X), E(X^2), E(X^3)$ are also low. The case becomes vice versa when p is small and so we get a left skewed curve. The Excel implementation of the distribution has been made interactive like previous two distributions.

It is needed to take care in the process of choosing p in this excel simulation. Since the upper limit of number of trials is set as 100, if p is very small, we will not obtain success even after 100 trials and so there may be difficulty in finding probability distribution. In order to avoid this problem, p is taken larger than $\frac{1}{50} = 0.02$ such that we will be mostly able to achieve success mostly within 100 trials.

VII. CONTINUOUS DISTRIBUTION

In continuous distribution the random variable X takes continuous values. From the students feedback, it is understood that they find it difficult to understand continuous distribution. This is because probability is not defined for a particular value but a range of values. Also they find it difficult to visualize formulas owing to the fact that it involves integration. In this section we perform a detailed study of three popular continuous distributions like Uniform, Exponential and Normal distributions.

A. UNIFORM DISTRIBUTION

The uniform distribution is also called as rectangular distribution which has constant probability throughout the given interval [a, b]. Since we have "RAND ()" function in Excel which helps in generating the uniform values from 0 to 1, and it can be modified to range from a to b byusing the formula {=a+(b-a)*RAND()}.
Using the above formula, the simulation of 2000 outcomes of uniform distribution can be done easily. In order to draw histogram, the number of bins that will be used in the distribution must be decided priory. In this work 10 bins are used and so the range $[a, b]$ is equally divided into 10 sub ranges. Using ÇOUNTIF () command, the number of values of X in each of the sub range is counted. Now the 10 bins are plotted as a histogram with range on x axis and count on Y axis. The values will be of almost same count for all the sub ranges and so it is expected approximately to get a rectangular plot.
One of the challenges students face is visualizing $E(X)$ and higher order moments. This is primarily because of the integration term involved in the formula

$$E(X) = \int x.f(x)dx$$
$$E(X^2) = \int x^2.f(x)dx$$
$$E(X^3) = \int x^3.f(x)dx$$
$$\text{M}$$
$$\text{M}$$

But without using formula, finding $E(X)$ and higher order moments is same like that of discrete distribution. For example, to find $E(X)$,just find average of all the values of X. Initially readers will find it difficult to associate integration with summation of the outcomes of X. The students normally find it difficult to visualize integration formula than summation formula. So by this excel simulation, a better physical meaning is obtained for $E(X)$. Similarly $E(X^2)$ is found by squaring all the X terms and finding the average.
The sheet has been made interactive and N can be varied from 10 to 2000. For low values of n, the curve is not smooth but as n gets increased, a smooth rectangular curve is obtained.

B. EXPONENTIAL DISTRIBUTION

The exponential distribution is commonly called as negative exponential distribution which could also be said as a continuous analogue of geometric distribution. For a random variable X from exponential distribution, the cumulative distribution function, F(x) is given as:

$$F(x) = 1 - e^{-\lambda x}$$
$$e^{-\lambda x} = 1 - F(x)$$
$$-\lambda x = \ln(1 - F(x))$$
$$x = -1/\lambda.\ln(1 - F(x))$$

Since 1-F(x) is a random number between 0 and 1, a simplified formula can be used as $x = -1/\lambda.\ln(F(x))$, where, F(x) is the random number, which can be generated in excel using the RAND() function. The equation $x = -1/\lambda.\ln(F(x))$ helps in generating the random variable X.

Using the similar procedures followed in V-A, histogram is plotted for the distribution and the moments are calculated without the aid of formulas. The histogram shows that the distribution is a decaying one and the rate at which it decays is given by λ. The distribution has been implemented in excel and it has been made interactive. The user has the provision to change n value from 10 to 2000. Also provision has been provided to change λ. It can be seen that when the value of λ is increased, the rate of decaying prolongs and the distribution gets spread out.
There is a relationship between the Poisson distribution, which was discussed in section IV-C and the exponential distribution. If Poisson provides an approximate description of the number of occurrences per interval of time, then the exponential will provide the description of the length of time between occurrence[9] .This relationship can be proved with an aid of an excel experiment.

The value of λ is fixed as 6, which means that we will encounter on an average 6 occurrence/hour. 2000 random numbers has been generated for the exponential distribution for the above value of λ. One set of those random numbers is pasted in a separate column. These values represent the length of time between each occurrence. By cumulating the values of the random variable, the actual occurrence time can be obtained. Now by defining Y as the number of occurrence for unit time. it can be observed that Y follows Poisson distribution. This relationship between the two distribution has been brought to the limelight.

C. NORMAL DISTRIBUTION

Normal distribution is one of the widely known distributions. Most of the real time data fall in this distribution. The distribution is bell shaped.

In Excel, the presence of the "NORMINV()" function will directly help to get the normal distributed random numbers if the mean and variance is specified properly. Alternatively, box Millar formula can be used to simulate the distribution [12]. But in the work, "NORMINV()" function has been used. Thus simulation of the distribution becomes straight forward and thus the study of the distribution by plotting the histogram will become much easier.

Normal distribution is one of the widely used distributions. More than 90% of the data we encounter in real life like height, weight and other continuous data comes from this distribution. To get a better picture of the spread of the distribution and how it changes due to clubbing of data, a simple experiment has been performed. 2000 random outcomes have been simulated from normal distribution with mean 170 and variance 5. Let this outcome denotes height values of 2000 students in a class. The distribution is plotted and the variance is found using VAR() function in excel. After this step, two cases are analyzed. In the first case, 5 students are considered a single unit and find local average. This can be done in excel by splitting the 2000 data present in a single column to 5 separate column, each containing 400 data and then finding average of each row. Now the distribution is plotted and its variance is obtained. Similarly, in other case the 2000 data are split into group of 10. Variance is obtained for this case too. It can be observed that the variance in the second case is less than that of first. Thus, we get a better idea of how clubbing of data influences the spread of the distribution by practically simulating the distribution in Excel.
Also, in order to make students understand how the variance changes while adding or subtracting two different data type, another experiment is performed. Two random variables, say X1 and X2 are generated from normal distribution. Here X1's mean and variance are 10 and 2 respectively and X2 have mean as 5 and variance as 3. The Random variable Y1 is calculated using the sum of X1 and X2. Similarly, Y2 is calculated with the difference between X1 and X2.It is found that Y1 and Y2 have almost same variance. These provide a clear insight that if two distributions is added or subtracted, the variance of the resulting distributions remains unaltered.

VIII. BIRTHDAY PROBLEM

Excel can serve as a good tool for solving probability problems which are computationally expensive. For example, the problem of finding the individuals having same birthday out of 32 persons can be taken at random. Theoretically, the probability of this problem can be done using the formula $1-\left(\frac{{}^{365}P_{32}}{365^{32}}\right)$. Finding the computation theoretically or manually is very difficult even by using the calculator. This can be easily solved using simulation in excel. In simulation, 32 persons birthday are randomly generated and check whether two persons having same birthday or not. This process is repeated for large number of times. Then the total number of iteration gives the real probability.

IX. DISCUSSIONS

In order to provide clear insight to the students, the entire work has been made available in a single Zip file. The Excel file can be downloaded from the link: *nlp.amrita.edu:8080/sisp/Excelsimulation/probability.zip*. Each experiment was made as much interactive as possible so that students learn the information with fun. Also, in the left side of each sheet, step by step procedure has been explained which helps the students to follow the simulation in a much easier manner.

The concentration of the work is given on very small number of distribution only. But there is scope to simulate other distribution in excel. One way of doing it is by knowing its cumulative distribution. Once Cumulative distribution is known, the probability distribution can be generated by following the same procedure as mentioned in section V-B.

X. CONCLUSION

We have tried to explain the various probability definitions and distributions like discrete and continuous using the basic mathematical tool so that the students can really experience the taste of every concept of Probability and Statistics. Some of the distinct real world problems like Birthday problems that are computationally expensive can be solved in a much simpler manner using Spreadsheet.

Future work is to explore the multidimensional probability and central limit theorem using the spreadsheet.

ACKNOWLEDGMENT

Thank you to Dr. K.P Soman who is the primary motivator of this work and the participants of the survey.

REFERENCES

[1] Joan Garfield.(1995).How Students Learn Statistics.InternufionnlSkUisfical Review,63,1,25-34.

[2] James, R. Evans.*Spreadsheets as a Tool for Teaching Simulation*.INFORMS Transaction on Education. 1:1, 27-37.

[3] Giuseppe Cicchitelli. (2006). Demonstrations In Probability And Statistics Using Excel.ICOTS.7.

[4] David, H. Moen., John, E. Powell. (2005).*Illustrating the Central Limit Theorem Through Spread SheetSimulations* .College Teaching Methods & Styles Journal.Volume 1, Number 2.

[5] Shuqin Yang.(July 2006).*Applications of Excel in teaching Statistics*. The China Papers.

[6] David NgangaNjenga.Seventh-Grade Curriculum In Probability (A Guide For Teachers).A Thesis Submitted to the Graduate Faculty of the Louisiana State University and Agricultural and Mechanical College in partial fulfillment of the requirements for the degree of Master of Natural Sciences.

[7] Seal, Kala Chand.Przasnyski, Zbigniew H. (2005) *Illustrating Spreadsheets in Education (*eJSiE*)*.Vol. 2.Iss.1, Article 4.

[8] Thomas J. Pfaff.Aaron Weinberg. (2009).*DoHands-On Activities Increase Student Understanding?*: A Case Study.Journal of Statistics Education.Volume 17, Number 3.

[9] Arnab Bhattacharyya.(2003) The Poisson Distribution.MIT Junior Lab.

[10] John C.B. Cooper.(2005).The Poisson and Exponential Distribution, Applied probability trust.

[11] Dawn Griffiths, (2009), *Head First Statistics*, O'Reilly Media, Inc.

[12] Jonathan Goodman, (2005), *Chapter 2: Simple Sampling of Gaussians*, *Lecture notes on Monte Carlo Methods*, Courant Institute of Mathematical Science, NYU.

AUTHORS

First Author – Anand R, student, Amrita School Of Engineering,anandbhuvanam@gmail.com.
Second Author – Manju M, student, Amrita School Of Engineering,manjummurali@gmail.com.
Third Author – Anju M Kaimal, student, Amrita School Of Engineering,anjumoh.
Fourth Author – Veenaa Deeve NV, student, Amrita School Of Engineering,veenaa8190@gmail.nam.kaimal@gmail.com
Fifth Author – Chithra R, student, Amrita School Of Engineering,rchithra89@gmail.com.

Correspondence Author – Anand R, anandbhuvanam@gmail.com, anand.ece.104@gmail.com, +91-9791129552