The Creative World of Statistics and Probability

Leslie Austin

INTRODUCTION

Robert E. Lee High School is a comprehensive urban high school established in 1962 in the southwest side of downtown Houston and is part of the Houston Independent School District (HISD). Lee serves 2092 9^{th} through 12^{th} grade young adults. Sixty-four percent of Lee's students are economically disadvantaged and 43% are at-risk of dropping out, according to standards set by the Texas Education Agency. Seventy-four percent, (1557) of Lee students carried the limited English proficient (LEP) label at one time; of these, 51% (796 students) still receive services. This 2000-2001 academic year, 1007 students indicated that they were born in a country other than the U.S.; Lee High School serves students from 60 different countries who speak 33 different languages. Many of these students did not attend school in their home countries or attended sporadically due to challenging circumstances.

I teach science in a school where many students find it difficult to connect their academic learning to real-life experiences. In addition, my students are LEP; therefore I often use visual aids to help increase their English by having them perform hands-on experiments. I want to expose students to some of the many uses of probability and statistics. They see statistics almost everyday although informal, but they do not have an understanding of why it is used. I want to empower them with the understanding that they can decide for themselves whether to believe information that the media tries to impress upon them. Knowledge is power and if I can help them understand the uses and misuses of statistics, it will help them later on in life.

My goal is to do a variety of mini projects with students. After I introduce and show my students how to use fundamental concepts in statistics and probability, I want to demonstrate the uses of them through games (dice, coin toss, gambling). I anticipate students will learn which games are suitable for making profits, and that gambling is designed to help individuals lose money.

Next, I want them to apply statistical concepts through classroom projects. I intend to use my classes as a sample of the population, the entire student body. For one of the projects I will create a few survey forms for my classes and we will compile and analyze the statistical results. The purpose of these activities will be to show students that questions can be asked in different ways to solicit certain responses. In other words some of the statistics we see generalized in the public, are results of a pre-meditated questioning style designed to make people respond a certain way. When they have grasped this concept I intend to have students create surveys for the entire student body. They will then decide which classes will be the sample group. This exercise will demonstrate the effectiveness of using a sample size to represent the entire population. These students need to learn that when the media gives statistics on a subject, the information was collected on a select number of people, and then extrapolated to represent the entire population.

Using statistics for health-related issues is another project I would like to incorporate. These students are at the experimental age and I would like them to find the probability of contracting certain sexually transmitted diseases. This project will demonstrate another use of statistics for actual life experiences.

It is my vision that these mini projects will be successful in linking statistical science to real life situations. Hopefully, students will be able to interpret and analyze statistics on their own, and they will feel empowered through knowledge of subject matter to not take future statistics at face value. Ultimately I hope an interest is sparked in students to want to pursue mathematics and/or sciences in higher education.

OVERVIEW OF UNIT

This curriculum unit will focus on two areas: using data to interpret statistics, and using data to interpret probability. Students will need to have a fundamental understanding of these topics in order to execute the projects effectively.

Data, the information collected from experiments can be of two types, quantitative or qualitative. Quantitative data are those that represent the amount (or quantity) of something measured on a numerical scale. For example, the time (measured in seconds) it takes for a solute to dissolve in a solvent during a chemistry experiment. Time is a quantitative variable. Qualitative data, sometimes referred to as categorical data, can only be classified. For example, a list of all HISD students and their ethnic backgrounds would be a set of qualitative data.

The ability to analyze and interpret data is a critical problem-solving tool. Today students are inundated with information; computers are able to store, sort and analyze massive amounts of data. Students should learn to apply important concepts and skills across all subject areas. They need to make the connection from use of data in their academic environment to their social environment.

In every day life, people make decisions through their own informal collection of data. For example, consumers often base their choice of current and future purchases on past experiences or information gathered from acquaintances. Students enter the classroom with some knowledge, although informal, about how to use data. Classroom experiences need to build on students' natural experiences and show them how to interpret and analyze the data presented to them. Students also need to learn how to effectively reason and make decisions when a large amount of data is given. "When data has been gathered, organized, represented, and analyzed by someone else, students need to rely on their own experiences to question data gathering processes, sampling biases, conclusions, and so on" (Zawojewski).

Statistics is the science of data. It involves collecting, classifying, summarizing, organizing, analyzing, and interpreting data. Statistics is commonly applied to two types of problems: 1) summarizing, or describing data, or 2) using data samples to make inferences about the data set from which the sample was selected. The branch of statistics used for the organization, summarization, and description of data sets is called *descriptive statistics*. The branch of statistics concerned with using sample data to make an inference about a population is called *inferential statistics*.

The data set that characterizes a topic of interest to us, i.e. the data we want to know more about is referred to as a population. A sample is a subset of data selected from a population. For this curriculum unit, the words population and sample are used to represent the objects from which the measurements are taken.

Probability is how society measures the uncertainty associated with events. For example, it may be reported that the probability of snow on a given day is 30%. This statement acknowledges that it is uncertain whether it will snow on the given. It indicates that weather forecasters measure the likelihood of its occurrence as being 30%. Probability is also used in decision-making processes to help us make forecasts.

There are two concepts of probability; 1) the relative frequency with which an event occurs in the long run (*statistical probability*), and 2) the degree of belief, which it is reasonable to place in a proposition on given evidence (*inductive probability*). One example of statistical probability is coin tossing. For example, when a penny is tossed no one can tell which way it will fall; but it is expected that the proportions of heads and tails after a large number of tosses will be equal. "A statistical probability is the limiting value of the relative frequency with which some event occurs" (Source: Utts).

Statistical probability is an empirical concept. It deals with what actually happens in the real world. It can only be proved or disproved by observation and measurement. The only way to find whether the probability of tails is ½ when a coin is tossed is by tossing the coin a large number of times. Confirmation or disproof can only come from performing actual experiments. With statistical probability, no experiment can go on forever; either the experimenter or apparatus will eventually wear out.

Inductive probability is the concept that the degree of belief which it is rational to place in a hypothesis or proposition on given evidence. Juries effectively illustrate this concept in deliberations. "What the jury does is to decide, after hearing the evidence, what, as reasonable men, they ought to believe and with what strength they should hold that belief" (Mendenhall & Sincich). If they consider the probability very high, the verdict is guilty; otherwise, the verdict is not guilty. Inductive probability is a logical concept. It refers to the logical relationship between a hypothesis in which we believe and the evidence on which that belief is based. Data from inductive probability differs from statistical probability. In jury assessments each trial is unique and can never be considered as one of a large number of similar trials.

Using Data to Interpret Statistics

There are two main aims to statistical analysis: to reach a decision, and to summarize information collected in the data. How does one make sense out of data collected? Data sets can be described using either graphical or numerical methods in conjunction with a variety of tools.

Qualitative data sets are usually described graphically through bar graphs or pie charts. Bar graphs show the frequency corresponding to each category. Pie charts (or circle graphs) divide a circle into slices, thus represents where each slice corresponds to a category. The size of the slice is proportional to the category frequency. Summary tables describe qualitative data numerically. The summary table gives the frequency and relative frequency for the data collected. Examples of these graphical and numerical methods are shown in Table 1 and Figures 1 and 2.

Table 1 contains the data used to create the graphs for Figures 1 and 2. Both graphs show that from the 77 students sampled, more students earned a letter grade of C than any other grade.

Table 1: Number of Students and Grades Earned in Course

		Grade	# Students		
		A	9		
		B	20		
		C	34		
		D	10		
		F	4		

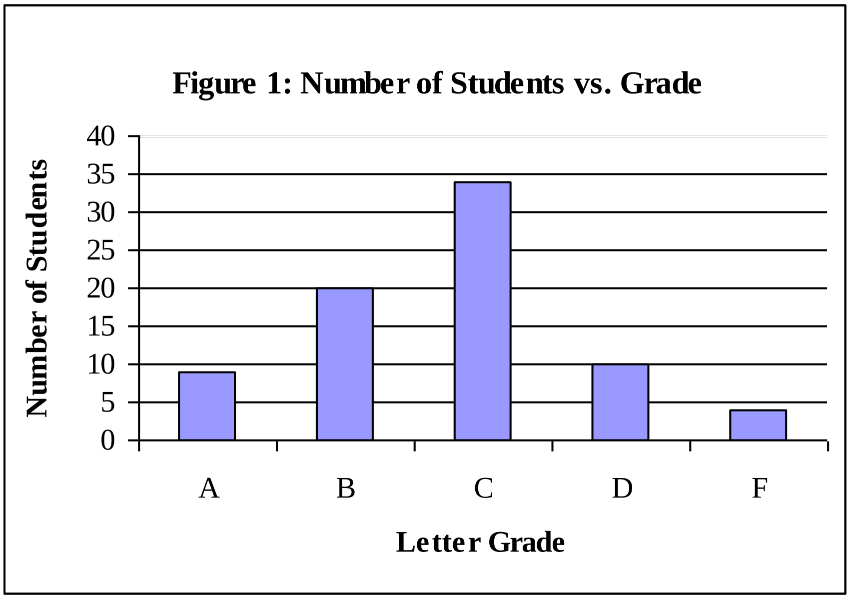

Figure 1 depicts this result by graphing the number of students related to each corresponding letter grade. From the graph one can see that the largest bar corresponds to a grade of C, 34 students.

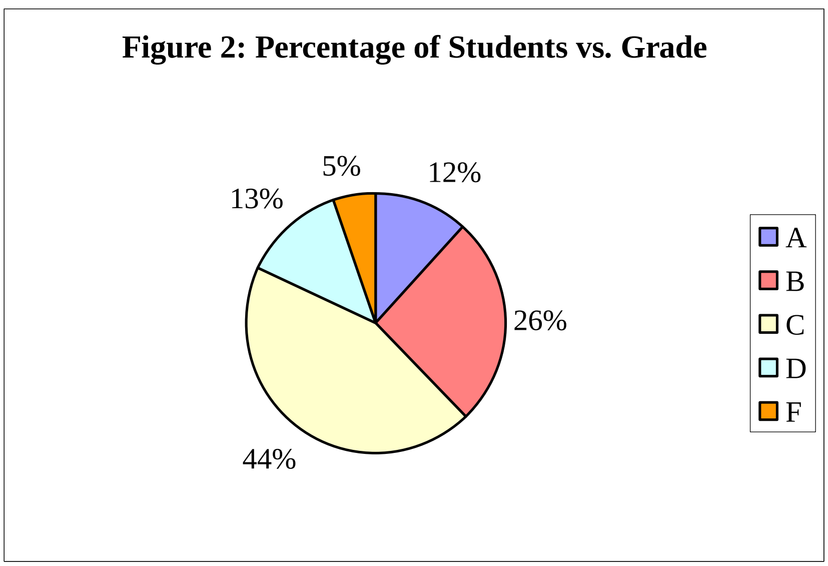

Figure 2 shows the percentage of students that earned each letter grade. It also shows that grade C is the largest section, or biggest slice of the pie, 44%.

In addition, students will be able to describe quantitative data sets using numerical methods. They will be able to locate the center of the relative frequency distribution, measure variation in data, and describe the position of an observation within the data set (Source: Mendenhall and Sincich). The most common measure of data toward central tendency is the arithmetic mean (generally referred to as the mean). The mean is the average of the measurements taken, and the sample mean is normally denoted by $\bar{y}$ and it is calculated by the following equation:

$$\text{Arithmetic Mean: } \frac{\sum_{i=1}^{n} y_i}{n}$$

For example, let's sample a class where the following shoe sizes were measured:

7, 9, 11, 12, 11, 15, 17, 11, 13

where y represents each size. We would solve the mean by adding each shoe size and dividing by the total number of shoes sampled, eight. The mean is (93/8) or 11.6.

Another measure toward central tendency is to calculate the median. The median is the middle number when measurements are arranged in ascending or descending order. If an odd number of measurements are taken, half the measurements would fall below the median, and half the measurements would fall above the median. The median can be calculated using the following equations depending on whether the number of measurements (n) is odd or even.

$$\text{Median: } \begin{cases} y_{[(n=1)/2]} & \text{if n is odd} \\ \dfrac{y_{(n+2)} + y_{(n/2+1)}}{2} & \text{if n is even} \end{cases}$$

The median in the above example is 11.

The mode is another common measurement toward central tendency. It is the value of y that occurs with the greatest frequency. Graphically it is the peak of the relative frequency distribution. In this example the mode is 11 because that shoe size occurs three times whereas the other shoe sizes occur once.

One of the objectives of this unit is to teach students to measure data variation. The most common measures of this are the range, variance, and standard deviation. The range is the difference between largest and smallest measurements in a data set. In this example the range would be 10, the result of 7 (smallest size) subtracted from 17 (largest size).

Standard deviation refers to the way measurements can vary. Students will also learn how to describe variation data using the Empirical Rule:

1. *Approximately 68% of the measurements will fall within 1 standard deviation of their mean.*
2. *Approximately 95% of the measurements will fall within 2 standard deviations of their mean.*
3. *Almost all measurements fall within 3 standard deviations of their mean.*

Highly unusual values from data sets, i.e. outliers will be discussed. Students will learn to construct box plots and calculate percentiles to measure the relative standing of a value within a data set.

Using Data to Determine Probability

How do we measure the uncertainty associated with events? Data for determining probability is collected by observation or measurement. The process of obtaining an observation or taking a measurement is called an experiment. When experiments are conducted only once, each observation is referred to as a simple event; it is a basic outcome of an experiment and cannot be broken down into simpler outcomes. For example if you roll a die once, the six possible outcomes or simple events to this experiment are:

1. Observe a 1
2. Observe a 2
3. Observe a 3
4. Observe a 4
5. Observe a 5
6. Observe a 6

The collection of all the simple events of an experiment is called the sample space. For example, the sample space for tossing a coin is

Sample Space

Observe a head

S: {H, T}

Observe a tail

H represents observing a head, and T represents observing a tail.

Venn diagrams are used to depict sample space and simple events pictorially. The following pictures are examples of Venn diagrams.

Throwing one die

* 1	* 4
* 2	* 5
* 3	* 6

Tossing one coin

* H	* T

Often data is collected for several simple events, and referred to as an event. The probability of an event is computed by summing the probabilities of the simple events that comprise it. Students will learn how to calculate probabilities of events using a 5-step approach (Source: Mendenhall and Sincich).

1. Define the experiment, the type of observation and how it will be made.
2. List the simple events.
3. Select the simple events to be contained in the event.
4. Sum the simple event probabilities to calculate the event probability.

Random variables

The random observation from a population may result from any possible number of outcomes. The random variable is the numerical quantity whose value depends on the outcome. Its value is determined by the outcome. Thus a specific characteristic of a population is determined if the characteristic of the variable is identified (Source: Choi).

Reliability

Reliability refers to something that can be depended upon time after time. A reliable measurement will give you or anyone else approximately the same results time after time, when taken on the same object or individual. The most reliable measurements are physical ones taken with the proper instrument.

Variability

Variability refers to changes without a consistent pattern. Most measurements have some degree of variability in that they include unpredictable errors or discrepancies that are not readily explained (Source: Utts). For example taking height measurements of males and females, if there were no variability within each group, it would be easy to detect a difference between males and females. The more variability there is within each group, the more difficult it is to detect a difference between groups.

LESSON PLANS

Lesson Plan Number One

Measurement Variation –Time Allotment: 90 minutes

The purpose of this lesson is to have students use statistics to calculate various measurements. Students will use the entire classes shoe sizes to collect data.

Objectives

- Students will be able to observe variation in measurement through collection of data.
- Students will be able use technology to compute statistical results

Materials – metric ruler, calculator (TI-83 plus)

***Pre-Lesson* –** Students need to have prior knowledge of using TI-83 plus calculator.

Hook – The teacher will ask the class who has the smallest and largest shoe sizes?

Directed Lesson Sequence – The teacher will use this exercise as an introduction to statistics, in particular the calculation of range. Students will become familiar with statistical terms and the entire class will perform use given data to perform calculations.

Independent Practice – Students will be divided into groups of three. Each group will measure the heights of five males and five females. After the measurements are taken they will compute the mean, median and mode for the group intermixed, then for males and females separately. Next they will answer the following questions.

1. Explain exactly how you measured the heights
2. Are your measures valid?
3. Are your measures reliable?
4. How does variability in the measurements within each group compare to the difference between the two groups?

Next students will enter data into their calculators to find statistical results. They will compare the answers.

Closure – The class will compare the data results to see if groups arrived at similar values.

Lesson Number Two

Surveys -Time Allotment: 270 minutes

Objective

- Students will be able to determine bias and unbiased in survey questioning

Materials – Working brain ☺

Hook- Teacher will ask questions soliciting student response. Teacher will rephrase the questions to solicit an opposite response from students.

Directed Lesson Sequence – Teacher will explain the purpose of surveys, how they are used in society and what can be learned from them through the following activities:

1. Activity 1 – Students will complete a professionally prepared survey.
2. Students will answer survey questions without bias
3. Students will write well-defined and clear surveys
4. Students will improve first draft of survey data

Guided Practice – Instructor will create a survey of three questions to measure attitudes toward something of interest to her. She will design a revised version by changing a few words in each question to make it deliberately biased. Half of the students will answer the first survey and the remaining half will answer the second survey. The two groups will compare responses and discuss what happened.

Independent Practice – Students will work in groups of four to select a topic of interest to them. Once the class decides on a topic they will construct a survey of no more than six questions to be asked to a class of their choice. Each group will go around to the class and administer the survey. Half the class will get the original version; half of the class will get the biased version.

Closure – Students will construct a mini report on the results from data and what they learned.

Lesson Number Three

Is This Game Fair? By Hope Martin

Time Allotment: 90 minutes

Objective(s)

- Students will be able to experience the concept of chance and likelihood.
- Students will be introduced to probability

Materials

- 2 six sided number cubes
- Calculators
- Blackline masters

Preparation – Photocopy a classroom set of blackline master 1. Make overhead transparencies of blackline masters 1 and 2.

Directed Lesson Sequence – The teacher will explain the rules of the game. The game is designed for two students, each of whom starts with ten points. One student is the "player" and the other, the "opponent". Only the player rolls the cubes. Each time the player rolls a sum of 7; the opponent must give up, or transfer three points to the player. When the play rolls a sum other than 7, he or she must give up, or transfer one point to the opponent. The winner is the individual with the most points at the end of ten rolls. If a participant runs out of points before ten rolls, he or she automatically loses.

Activity 1

After students are familiar with the rules, play one game and place results on the blackline master 1 overhead. The teacher is the player and the students are the opponents. The teacher will ask the students to predict in writing whether or not the game is fair.

Activity 2

Put students in teams of two. Give each team two number cubes and one copy of the blackline master 1 to keep track of the points in the game. Each pair of students should play ten rounds, recording their results after each toss.

Activity 3

Students will note trends by collecting all the class data on an overhead transparency of blackline master 2. As each team reports its winner, fill in the cells on the transparency. Determine the percentage of games won by the opponents and won by the players. Then pose the following questions for discussion:

1. Do you feel more strongly about your prediction? Why?

2. Do you feel less convinced about your prediction? Why?

3. Do you want to modify your prediction? Why?

If the data collected is not convincing, have students play another round and combine the new information with the old. Repeat the game until all students are convinced that the opponent has an advantage. Have students reflect on the above questions and modify their predictions after each round.

Activity 4

Once students are convinced that the opponent has a distinct advantage in this game, analyze the game mathematically. Use Table A to investigate the various ways that sums can be made on two number cubes.

Once the class has helped the teacher complete the table, discuss the probability of rolling a sum of 7, 6 out of 36, or 1.6. Consider this, the odds in favor of rolling a sum of 7 (1 out of 5, or 1:5). Given this information, ask students how the game could be changed to make it a fair game. The opponent should give up more than 3 points to the player for each sum of 7.

Table A: Sums of Seven

+	1	2	3	4	5	6
1						
2						
3						
4						
5						
6						

ANNOTATED BIBLIOGRAPHY

Teacher Sources

Choi, Sung C., *Introductory Applied Statistics in Science*, Prentice Hall, 1978.

This text was helpful in explaining the theory behind statistics and probability. It explained the calculations behind the concepts (or calculations) students will be learning.

Figure This Math Challenge for Families

http://figurethis.org/challenges/c14/challenge.htm An interactive Internet source that provides a wonderful opportunity for students to take challenges by answering historical statistics and probability questions.

Mendenhall, W. & Sincich, Terry, *Statistics for Engineering and the Sciences,* 4^{th} ed., Prentice Hall, 1995.

This book explains the concepts of statistics and probability in a concise manner. It filters out the extraneous information most mathematical textbooks include to get to the "root" of the issue, defining the terminology in "laymen's" terms. The text uses concrete examples to explain certain phenomena, such as the probability of certain weather patterns occurring or other events taking place. The authors did an excellent job of breaking down the major operations used to explain data collected from experiments, which is critical to the activities that will be performed in the classroom.

Probability of Same Birthday

http://www.stat.psu.edu/~resources/ClassNotes/rho_01.htm This Internet source provided a variety of interactive games students can play that will make the concept of probability easy to understand.

Utts, Jessica M., *Seeing Through Statistics*, California, Brooks/Cole Publishing Company, 1999.

Utts' book provides important case studies involving statistical use. She shows the application of statistics in many ever day, "real life" situations, through the benefits and risks of using statistics. She explains how to effectively conduct experiments and make observations. The language used in the book is easy for teachers with non-mathematical backgrounds to comprehend.

Zawojewski, Judith S., *Dealing with Data and Chance*, National Council of Teachers in Mathematics, Inc., 1995

This workbook was phenomenal in the types of activities it had for student use. These activities were geared toward getting students to think on a higher level. Each activity is highly participatory, but more importantly, students are able to make connections (link classroom to real world) while having fun. The activities in this workbook also foster integration of technology into the classroom.

Wetherill, G. Barrie, Elementary *Statistical Methods*, Redwood Burn Limited, 1982.
Wetherill's text explains the fundamentals of statistics and probability. He provides examples of how they can be incorporated into mathematical and science courses.

Student Sources

Figure This Math Challenge for Families
http://figurethis.org/challenges/c14/challenge.htm

Probability of Same Birthday
http://www.stat.psu.edu/~resources/ClassNotes/rho_01.htm

Zawojewski, Judith S., *Dealing with Data and Chance*, National Council of Teachers in Mathematics, Inc., 1995