Curriculum Units by Fellows of the Yale-New Haven Teachers Institute
1985 Volume VIII: The Measurement of Adolescents

An Introduction to Statistical Thinking

Curriculum Unit 85.08.07
by Beverly Stern

Introduction

Information is all around us. It affects every aspect of our lives. To gather and work with numerical information using statistical concepts is the overall purpose of this unit. I am planning to use “An Introduction to Statistical Thinking” in technical math classes grades 9-12. The content is basic enough to be used with data from different disciplines. The structure is flexible and may be used as a whole or in part. There are three sections. Section I: Data provides for the experience of gathering, organizing and analyzing classroom data attendance, grades and temperature. The primary objective of this section is to take a set of data and be able to see both how the values are centering by determining the mean, median and mode averages and how they are varying by determining the range and making a Box plot diagram.

In the first section we work with only the full set of values being considered. The full set of values is called the population. Since it is not always possible or practical to study an entire population, we must use samples. Samples are subsets of the population. Selecting samples is one of the most important tasks of statistics because from the samples chosen inferences are made about the entire population. The reliability of the inferences depends in large part on the reliability of the sample.

Not all samples are useful. If we want reliable information about a population, the sample we choose must be representative of the entire population. The sample we want is called a random sample which means that every member of the population has an equal chance at being picked. The primary objective of Section II: Random Sample is to introduce the idea of randomness and to solve random sample problems.

Section III: Probability has as its objective to use the traditional definition of probability in finding probabilities and comparing experimental and theoretical probabilities. No matter how much information we acquire, we almost never know everything there is to know about any given situation. However, since decisions must be made, we guess-sometimes using only intuition and sometimes partial information. The more information, the more reliable the guess. Probability helps us put a numerical measure on the uncertainty of an event, on the risk we’re taking.

I consider Section I to be the most important part because it allows students to not only begin at the beginning of statistical work by gathering data, but it allows them also to be the generators of it. I am looking forward to the discussions in class as we evaluate the four week records.

Section I: Data

The primary objective of this section is to take a set of data and be able to see both how the values are centering by determining the mean, median and mode averages and how they are varying by determining the range and making a box plot diagram. The basic strategy for this is to have students gather classroom information and then work with it repeatedly finding the indicators in which we are interested. The structure I plan to use in gathering classroom material for this work has three levels. The first level is to have each student keep a daily record of class attendance and his or her own grades. Form A is a possible example of an individual form to use. The space near the bottom can be used for the students individual grades. Level two of classroom data gathering is the daily maintenance of a wall chart of which will be recorded the percents of class attendance, an average of class grades when grades are given and the room temperature. See Forms B, C and D as possible examples. The third level of data gathering is more flexible and consists of the assignment and test grades obtained from students' work in sections II and III.

The basic time frame is to have students keep the individual forms and the wall chart for a period of about four weeks. After the students understand what is expected of them, keeping the individual forms and the wall chart will take only a few minutes each day. During that four week period classroom time mostly will be spent on statistical topics such as those given in sections II and III. I think of this part as statistically cultivating the classroom soil.

For some students keeping the individual forms will be a fairly easy task: for others it will be difficult to even keep track of the forms. Both for the individual forms and the wall chart, the percent of attendance must be calculated. To do this we need a count of who is present and who is absent. How do we count a student excused, say to attend a student government meeting? Is that student absent or present? It is a decision that needs to be discussed and made. During the process of keeping these daily records many such questions may come up. Discussing these questions and coming to an agreement as a class will help students develop a better understanding of what it means to gather data. There is leeway. How solid are the numbers we read in the newspapers, magazines and books?

But what does this have to do with the primary objective of this section to develop skill in finding measures of central tendency and variance for a given set of data? It is generating the sets of data we are going to use in finding the mean, median, mode and range and from which we will make box plots. What sets of data? Each set of classroom grades from assignments and tests and the final set of percent of attendance for each student as well as the set of temperature readings is a set of data the students understand well and for each we can find the measures we want. First we find them together in class, then students take sets of data and find measures on their own either in class or for homework.

Students can bring in their own sets from the areas in which they are interested or sets of data can come from the *World Almanac* or any area of concern . The number of sets used depends on the class situation, but for each set used the values are ordered, the mean, median, mode, and range are determined and a box plot is made. Let's look at the statistical concepts involved.

The mean is the sum of the values of a set divided by the number of values and is the average with which most students are familiar.

Example: {200, 30, 125,92}

Mean = 200 + 30 + 125 + 92 = 111.75

4

The median may be defined as the number in the middle after the values have been put in order. There are always the same number of values above it as below it.

Example: {2,4,6,8,10}

Median = 6

If the number of values is even, there isn't a middle value, so you take the mean average of the two middle numbers.

Example: {2,4,6,8,10,12}

6 + 8

Median = 6 + 8 = 7

2

The mode is the value found most frequently. None of the three sets of data given above have a mode since each value was used only once.

Example: {2,2,4,8,7,7,3,2}

Mode = 2

Example: {7,9,3,7,5,9,1}

Mode = 7 and 9

The last example has two modes and is called bimodal.

The average to choose is the one that best serves your need. A shopkeeper would be interested in knowing the mode of the sizes of shirts he sold so he would know how to order. If you were considering buying s house in a particular neighborhood, knowing the median income for the families who live there might be the most helpful average to know. If you were interested in baseball you would watch the batting averages of your favorite players. Here the mean is used. Each average gives different information. It gives a different view of the data.

Consider the information in the table below on the XYZ Plant Incomes. Find the mean, median and mode.

Mean = 183,000 = $20,333 0

9

Median = $12,000

Mode = $12,000

Incomes for the XYZ Plant

owner	$60,000
manager	40,000
worker	15,000
worker	15,000
worker	12,000
worker	12,000
worker	12,000
helper	9,000
helper	8,000

If you were the owner and wanted to show how well you paid you would say your plant paid an average salary of $20,333. If you were a worker who wanted an increase you would say that the average wage was $12,000.

To look at the variation in this set of data we want to find the range and make s box plot. The range of a set of data is the difference between the largest value and the smallest. Using the data from the XYZ Plant above we have the range equals $60,000 8,000 = $52,000.

Range = largest value smallest value

The range gives us an indication of how far the data is spread. It is an indicator of variance, but it does not give us any information about how the individual values are distributed or how they vary. For this a box plot can be helpful.

(figure available in print form)
The box plot is a quick way of seeing how the data is distributed. There is a lot of information presented. To construct a box plot make a line and impose a scale on it that will include your lowest and highest values. Next plot your values by putting an x above the proper number. If there are more than one of the same value, stack them.

Next draw a light dotted line down indicating the median value. For our data it is $12,000. Next put a dot indicating the lowest and highest values. There are only two more numbers we have to find, the upper and lower quantities. The upper quartile is the median of the data above the set median. The lower quartile is the median of the data below the set median. For our example, the set median is $12,000. There are four values above it and four below. The median of the upper four values is 40,000 + 15,000 = 27,000 this is called the upper quartile, Doing similarly for the lower quartile we get 12,000 + 9,000 = 10,500. Mark lines indicating the quantities and draw a box going from the upper quartile to the lower quartile as shown. Finally draw a line going from the upper quartile end of the box to the highest value and similarly for a line going from the lower quartile end to the lowest value. These end lines are called the whiskers of the box plot. To do this is quick and easy once you have learn ed the pattern .

Taking a look at some of the information presented in our box plot, notice that the line in the center area of the box is the median so that tells us half the values are greater than or equal to it and half the values are less than or equal to it. The upper end of the box divides in half the frequencies of the upper-valued data and similarly for the lower end of the box for the lower valued data. The shape of the box will change as the set of data changes. It is a model that makes it easy to talk about distributions. One can easily say and understand

things like, “There are 2 values on the lower whisker of this box whereas the last one we did had 6.”

Part II: The Random Sample

The primary objective of the second section is to introduce the idea of randomness and to solve random sample problems. To get a real feeling for what randomness is, have the students generate their own random number tables. One way to do this is to put ten wooden squares, each of which has one of the digits 0 9 inclusive on it, in a container and shake. Have a student draw a digit and write it on the board. Making it obvious that the sample has been returned, shake and repeat for maybe twenty numbers. Since each time each digit has an equal chance at being picked, this is called a random selection of digits, and what you are beginning to do is to generate a random digit table on the board. After doing this much together on the board, have students generate their own. Use Form E, #1. They can work in pairs and save time, increase interest and use less sets of squares for generating their numbers.

When the tables are done, use one of them to randomly pick a committee from the class. One way to do this is as follows.

1. List all students’ names on the board. Have students write the list on back of their worksheets.
2. Number each name. Say it goes from 01 to 20. Since we have 20 students, we need to read in groups of two digits and so write our numbers 01,02, 03, . . . 20. They will see the reason for this as soon as they start reading the table to select a committee.
3. Have a student point arbitrarily to a spot on the table you are using. Explain that we will start there and read the numbers from the table. However, before we start to read the numbers, we have to decide if we want to read horizontally, vertically or diagonally. It doesn’t matter, but once we pick a way we should stay with it until the task is done. An overhead projector might be helpful here.

Consider the table below. If you were starting at the third row second digit, we could read 78, 85, 53, 32, 12, etc. In doing so we would be reading across and in groups of two digits but moving one digit at a time. If we had a larger table we might prefer to move two digits at a time and so starting from the same place we could read 78, 53, 21, 22, 21, 17, 73, 70, 23, 25, 33 etc.

Sample Random *Table*

row:

1 14073
2 43318
3 77853
4 21222
5 11773
6 70232

7 52333
8 90012
9 86746
10 64337

However we read it, the first number that is from 01 to 20 inclusive gives us the first member of the committee, and we keep going discarding numbers that do not have meaning for our task. If we are selecting a committee of three, we keep going until we get three numbers from 01 to 20 inclusive, and the names that correspond to these numbers are the ones on our committee. If we use the table above and keep counting the way we originally began we would read the numbers 78,85,53,32,21,12,22, 22,21,11,17. That would give us 12, 11 and 17. The names corresponding to 12, 11 and 17 would make up our class's randomly selected committee. The task is now completed.

By the time the class has selected a committee of three, then each student selects his or her own committee, #2 on Form E, most students probably will be able to do the random sample problem #3 on Form E.

#3 Form E. A batch of 200 new cars has just been completed. Your job is to randomly select 15 of the cars for a special safety check.

a. Describe how to do this.
b. Select the 15 cars. Use random number table on handout.
c. List the 15 numbers selected.

For this problem you will want a larger random digit table than the ones generated. Form F. A classroom set of copies of a random table is needed. Solution:

a. Number all the new cars 000 to 200 inclusive. Arbitrarily select a place to start on the larger table, decide if to read across, down or diagonally and begin reading in groups of three digits. Any three digit number 000 to 200 inclusive we keep, and any others we discard. Continue until we have 15 useful numbers. The cars with these numbers will be used for the special check.

Two possible extensions might be to use this method to take a survey or to do a simulation problem. To take a survey of the student body requires that several decisions be made. One decision is what question or questions do you want to ask? Since this unit deals with numerical values, you'll want numerical data back so you can evaluate it using the techniques from Section I. Possible questions could be "How much soda do you drink in a week?" or "What do you expect your annual income to be ten years from now?"

Another decision is how large a sample do you want? What is an adequate size? Too small and it may not be valid. Too large a sample may be too much work to do. Thirty seems to be a good size with which to work. Once you have the size of your sample, how will you go about getting a random sample, gathering the data, analyzing the results? Can you publish the results in the school newspaper?

The second extension could be this simulation problem from *Understandable Statistics* , Brase/Brase, p13.

A single pollen grain floating on the surface of water will move randomly from the impact of the water molecules. The task is to chart the course of a pollen grain as it moves on a drop of water for seven position changes. A problem, however, is that the pollen grain is so small and its movements are so fast that you would need to use a microscope and slow motion camera to see the changes. Since you do not have this equipment, you will have to use a random number table to simulate the observed direction of the pollen grain for seven position changes. Instructions. Allow that for each position change, the pollen grain is in the center of a circle marked in degrees as shown below. 0 degrees indicates east, 90 degrees indicates north, 180 degrees indicates west, and 270 degrees indicates south. The arrow points to the direction of change.

(figure available in print form)
Solution. Using a random number table, arbitrarily decide where to begin and in which direction you will read. Then, since there are 359 possible positions, begin reading in groups of three digits. Keep the numbers that are between 000 and 359 inclusive and discard those that are not. When you have seven such numbers, chart the position changes according to the instructions above. A possible looking solution might be as follows.
(figure available in print form)

III: Probability

The primary objective of this section is to use the basic definition of probability in finding theoretical and experimental probabilities. The basic definition of probability is the number of successful outcomes

$$P = \frac{\text{number of possible outcomes}}{\text{total number of possibilities}}$$

This can be written as P(E) = n(E) where P(E) means n(S) the probability of event E happening, n(E) means the number of times E could happen and n(S) means the number in the sample set which is the total number of possible outcomes.

Consider a die. It has six surfaces, and each surface has a set of 1, 2, 3, 4, 5, or 6 dots on it. If I roll a die, the only possible outcomes are 1, 2, 3,4 , 5 or 6. These six elements make up the sample set for our event the rolling of the die.

If I roll a die, I can ask for the probability of different events happening. What is the probability of the following.

a. P(1) = ___

b. P(even number) = ___

c. P(8) = ___

d. P(n > 5) = ___ where n means the number on the die

e. P(odd number) = ___

f. P(n 7) = ___

Since each of these is answered by P(E) = n(E)/n(S), the n(S) answers are as follows.

(figure available in print form)
Notice O means no possibility the event will happen. 1 means it will always happen. The probability of an event will always be between 0 and 1 or equal to one of them.
(figure available in print form)
If I roll a die, P(2) = 1/6. This could be written 1/6 as 1 or as 0.166. P(n 4) = 1/2 or 0.5. This may be an interesting way to review students' basic skills in fractions and decimals. We'll use both below.

I want to roll a die 12 times to see if the probability of getting 4 really is 1/6 as indicated by the definition.

Theoretical probability is what we have been talking about up to this point. Now we want to move out of the theoretical into the real world and try out that probability with a real die. I'll now roll a die 12 times.

results: theoretical probability P(4) = 1/6

experimental probability EP(4) = 4/12 = 1/3

In class one student could roll the die, another could tally it on the board. If we're lucky there will be a discrepancy to point out the difference between theoretical and experimental probability with a small sample of 12.

At this point, letting students roll dice and get how many times 4 comes up for each of them could be organized as follows.

Times Roll Die	Number of 4's	P(fraction)	P(decimal)
12	___	___	___
20	___	___	___
30	___	___	___

For each of the above three experiments, have students calculate the experimental probabilities they find using both fractional and decimal forms.

Notice that by using the basic definition of probability we can find simple probabilities, both theoretical, the probabilities that you might expect, and the experimental, the probabilities you get in the real world by doing experiments like roll a die, flip a coin, or draw a card from a deck. Further, by using small samples the experimental probability might be quite different from the theoretical, but as we increase the number of tries, that is as the number in our sample increases, the experimental probability moves closer and closer to the theoretical. How large a sample is needed? Again, 30 is usually considered to be fairly reliable sample.

Other easy activities done in the same or similar way are to ask how the theoretical and experimental probabilities compare for P(n > 1) in rolling a die, or for P(T) the probability of getting tails when flipping a coin, or P(2) the probability of drawing a two from a pack of cards.

In summary, with simple probability problems we can use the basic definition of probability to experiment with the difference between theoretical and experimental probabilities. The "simple" here means problems where it is easy to count the numbers you need as opposed to more difficult probability problems where the basic idea is the same but the counting of needed numbers becomes more difficult.

Lesson Plan Guide

(figure available in print form)

Materials

1. sets of 10 squares each square has on it one of the digits 0, 1, 2, . . . 9
2. thermometer to do readings for part of wall chart data
3. forms like or similar to the ones given in this unit
4. random number table and a classroom set of copies of it
5. wall chart like or similar to forms given in unit
6. coins, dice and/or cards as fit the classroom situation

Bibliography

1. Brase, C.H. and Brase, C.P. *Understandable Statistics* 2nd Ed. Lexington, MA: D.C. Heath and Co., 1983. A good book for independent study of statistics. It is readable and offers a Serious study of statistics without using calculus.

2. Huff, D. *How to Lie With Statistics* . New York: W.W. Norton and Co., 19-54. A brief book, 142 pages, that clearly shows you how to lie with statistics.

3. Jacobs, H.R. *Mathematics A Human Endeavor* . San Francisco: W.H Freeman and Co., 1970. An excellent reference book for many kinds of math topics. Chapters 7,8 and 9 on counting, probability and statistics respectively present clear theory and many good problems.

4. Markley, N. *Introduction to Probability* , Revised Ed. Lexington, MA: (Ginn Press, 1985. A good book for independent study of probability.

5. Tanur, J.M. *Statistics: A Guide to the Unknown* 2nd Ed. Oakland, CA: Holden-Day, 1978. Presents essays using statistics in various areas. Excellent. The four major categories used are “Our Biological World,” “Our Political World,” “Our Social World,” and “Our Physical World.”

(figure available in print form)
Form A

(figure available in print form)

Form B

(figure available in print form)

Form C

(figure available in print form)

Form D

Form E

Random Number Worksheet

1. Create a random number table 5 digits across and 10 rows deep.

row

1 __ __ __ __ __
2 __ __ __ __ __
3 __ __ __ __ __
4 __ __ __ __ __
5 __ __ __ __ __
6 __ __ __ __ __
7 __ __ __ __ __
8 __ __ __ __ __
9 __ __ __ __ __
10 __ __ __ __ __

2. Using the random digit table you have just generated, randomly select a committee of three students from this class.
    a. assign each name a number
    b. randomly select three names
    c. list names _____ _____ _____

3. A batch of 200 new cars has just been completed. Your job is to randomly select 15 of them for a special safety check.
    a. describe how to do this
    b. select the 15 cars using the large random number table handed out to you
    c. list the 15 numbers selected

Form F

(figure available in print form)