Here's How to Incorporate "Big Data" into Your Statistics Class!

NCTM Annual Conference, Boston, MA
Thursday, April 16, 2015, 8-9:15 AM, Room 258 B (BCEC), Presentation #53
Complete lesson plans for these and other units are at
http://www.math.nmsu.edu/~breakingaway/ (Click on STATISTICS)

Patricia Baggett, baggett@nmsu.edu, and
Andrzej Ehrenfeucht, andrzej.ehrenfeucht@colorado.edu

Come and engage in classroom-tested activities involving three aspects of statistical investigations: Gather and analyze experimental data from hands-on-tasks; use TI-84 programs to simulate sets of "big data" for these tasks; and finally, investigate the mathematical models that explain the patterns observed in both real and simulated data.

Table of Contents

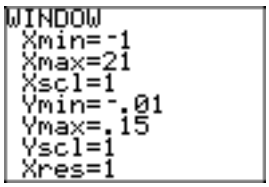
1. Tossing an icosahedral (20-sided) die

Part 1.
Props: Icosahedral dice, a cup, and a tally sheet.
Toss your icosahedral die a number of times and record the outcomes on the tally sheet. What do you notice? Do some numbers come up much more frequently than others?
Since an icosahedral die (an icosahedron) has 20 faces, there are 20 possible outcomes, and in a fair die, each one is equally likely. This means each one has a theoretical probability of .05 or 5%.
How close did you get to a constant .05 for each die? (To compute this, sum up the total number of rolls for the dice; call it N. Now for each outcome from one to twenty, use your tally sheet to get the number of times that particular outcome occurred, and divide that number by N. To change it to a percentage, multiply by 100. These 20 numbers are your actual, or empirical, frequencies.)
You may record your results here:

out-come	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
theor-etical probability	5 %	5 %	5 %	5 %	5 %	5 %	5 %	5 %	5 %	5 %	5 %	5 %	5 %	5 %	5 %	5 %	5 %	5 %	5 %
actual (empirical) frequency																			

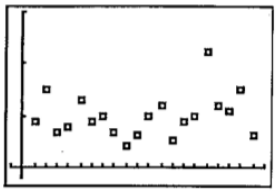
Do you think that, if you tossed your dice long enough, your actual frequencies would get closer and closer to 5%? We will try this in Part 2 using simulated icosahedral dice!

Part 2.
The program DISTRIB is in your calculator and on our website at https://www.math.nmsu.edu/~breakingaway/Statistics/Lessons/TID/TID.html.
Before running the program DISTRIB, set the window, format, and STATPLOT as follows:

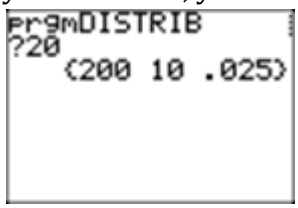


To run the simulation, start the program DISTRIB. At ?, enter 20 (for 20 sides):

You will see something like:



After every 200 throws, you will see something like:



The 10 means the largest difference between two outcomes was ten. And the .025 means that the difference between the current maximum empirical frequency and the theoretical probability is .025. (So here, since the theoretical frequency is 1/20 or .05, the current maximum empirical frequency is .075.) Press ENTER again to get more tosses. Run it as long as you want! To stop the program, press ON 1.

2. Statistics with M&Ms

This lesson has three parts. One is a hands-on activity with M&M's. The second is a calculator simulation involving collecting a lot of simulated data based on the hands-on activity, and the third is a theoretical interpretation of what is happening.

Part 1.

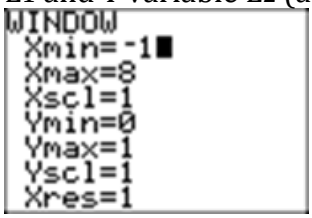
Each person needs a baggie with seven M&M's of the same color and a tally sheet. You also need an opaque container with a lid (a "Laughing Cow" cheese container works well).

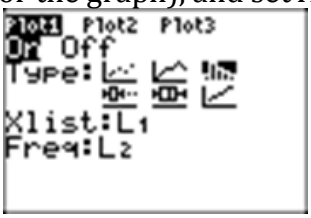
Place the seven M&M's in the container and put the lid on. Shake the container and then open it and count how many M&M's have the "M" side up. This number will be between 0 and 7.

Record the number on the tally sheet. (Above, the number is 3.)
Repeat this as long as you want, but at least 20 times, and then prepare a histogram. On the x-axis should be the numbers 0 through 7, labeled "Number of candies with the "M" side up". On the y-axis the numbers should be 0, 1, 2, 3, …and it should be labeled "frequency".

Look at the histograms that others in the class have made. Most likely they are all rather different. But this is not the final conclusion. In Part 2 we will see what happens when we have very large samples. Do the histograms remain different? Or do they begin to become more similar? (Usually they become more similar!)

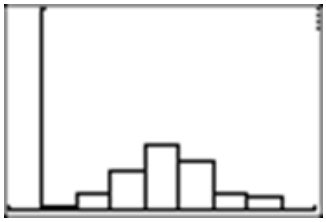
Part 2.
Simulating the activity in Part 1 with large samples
We will need the program MANDMS for our TI-83/84 calculators that simulates a large number of throws of 7 M&M's. The code is on your calculator, and also at https://www.math.nmsu.edu/~breakingaway/Statistics/Lessons/M&Ms/P2M&Ms/P2M&Ms.html. Before you run the program, you need to set the window, turn Plot1 on with X variable L1 and Y variable L2 (and type 3, a histogram for the graph), and set AXES ON in FORMAT:



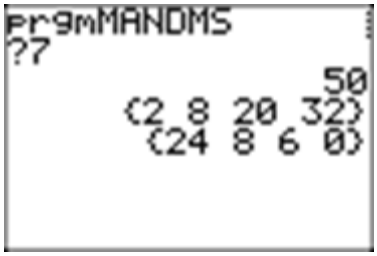


Now we're ready to run the program, MANDMS. At the ? we type 7, as we are tossing 7 M&Ms during each throw:

The program stops after 50 throws, and here is my first histogram. (Yours will most likely be different.)

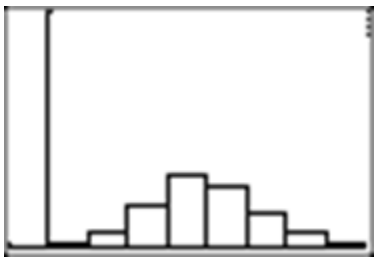


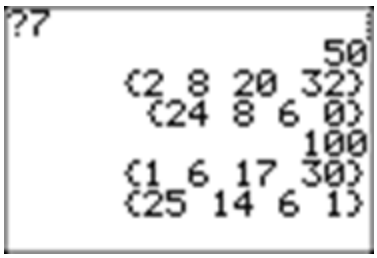
When I press ENTER, I get a screen of data:



It tells me there were 50 throws of M&Ms, and the percentages of each outcome were:

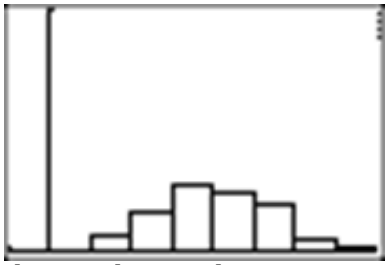
no Ms up	2%
1 M up	8%
2 Ms up	20%
3 Ms up	32%
4 Ms up	24%
5 Ms up	8%
6 Ms up	6%
7 Ms up	0%

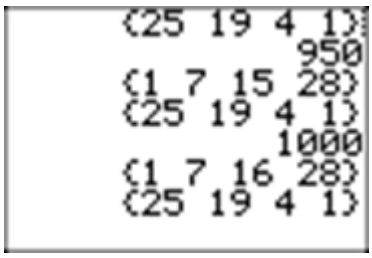




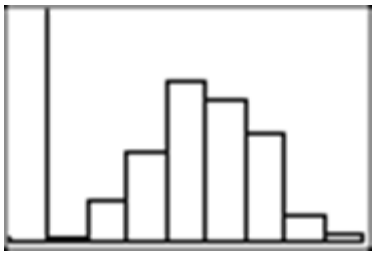
I press ENTER again, and I get another histogram, and ENTER again gives me more data:

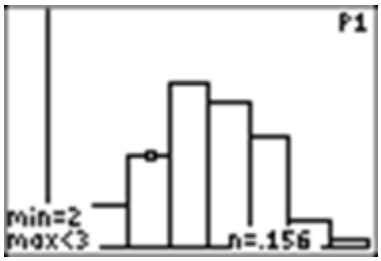
I'll stop at 1000 throws. (To stop the program, press ON1 or ON ENTER.)

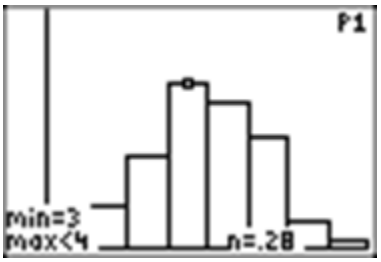




I change the window size, setting Ymax at .4, to stretch the graph vertically, and then I use TRACE to see the values:







I see that 2 M&Ms up appears 15.6% of the time, and 4 come up 28% of the time (rounded). (Your findings will probably be different!)
In Part 3, we give a theoretical interpretation of what is happening.

Part 3.

Here we show the theoretical probability that is used to explain the frequencies you observed working with M&Ms. It is Pascal's triangle.

1

1 1

1 2 1

1 3 3 1

1 4 6 4 1

1 5 10 10 5 1

1 6 15 20 15 6 1

1 7 21 35 35 21 7 1

1 8 28 56 70 56 28 8 1

1 9 36 84 126 126 84 36 9 1

1 10 45 120 210 252 210 120 45 10 1

(The triangle does not end here.)

The rows of the triangle are numbered starting at zero. We look at the 7th row, since we tossed 7 M&M's. We look at all $2^7 = 128$ possible combinations of M-up or M-down patterns of 7 M&M's. These eight numbers, 1, 7, 21, 35, 35, 21, 7, and 1, give us the number of combinations in which, out of 7 M&M's, 0, 1, 2, 3, 4, 5, 6, and 7 are M-up. The sum of these numbers is, of course, $128 = 2^7$. So the theoretical probability distributions based on counting cases is $1/2^7, 7/2^7, 21/2^7, \ldots, 1/2^7$, as shown in the table below.

0 M's up	1 M up	2 M's up	3 M's up	4 M's up	5 M's up	6 M's up	7 M's up
1/128	7/128	21/128	35/128	35/128	21/128	7/128	1/128
~.78%	~5.47%	~16.4%	~27.3%	~27.3%	~16.4%	~5.47%	~.78%

My data from the simulation above, in which the M&M's were "tossed" 1000 times:

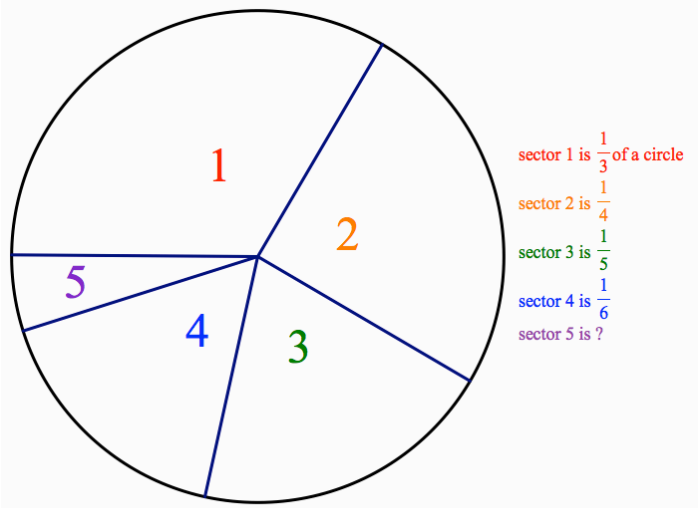
~.6%	~6.9%	~15.6%	~28%	~24.7%	~18.8%	~4.3%	~1.1%

Does Pascal's triangle really fit to the data? You will see that most of the time when the sample is quite big, the fit is pretty good.
For an explanation of the reasoning for why this is the case, see https://www.math.nmsu.edu/~breakingaway/Statistics/Lessons/M&Ms/P3M&Ms/P3M&Ms.html.

3. Spinners

Part 1. Designing a dial
Tools needed: A compass and protractor, paper, colored markers, and a calculator (optional)
Draw a large circle and divide it into the following 5 sectors:
1/3 of a circle, 1/4, 1/5, 1/6, and 1 - (1/3 + 1/4 + 1/5 + 1/6). Mark the sectors with the numbers 1 (for 1/3), 2 (for ¼), etc. Color the dial with five different colors, one for each sector.

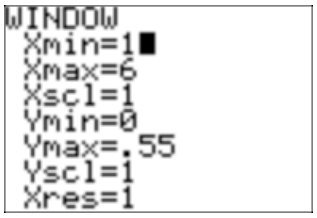
Can you figure out the number of degrees needed for each sector?
Here is a dial that we made with Geometer's Sketchpad:



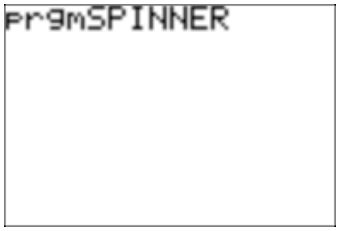
dial for a spinner

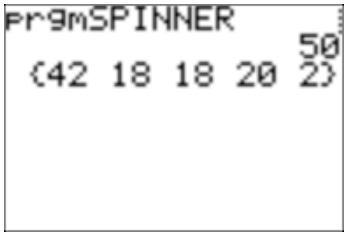
Part 2. Simulating a spinner
Your calculator contains a program, SPINNER. Its code can be found at https://www.math.nmsu.edu/~breakingaway/Statistics/Lessons/Spinners/P2Spinners/P2Spinners.html. It is used to simulate spins on a dial with sectors like those in Part 1. We can run it to collect a lot of simulated data. We want to compute the theoretical frequency of outcomes for the dial, and see if the simulated data, after many spins, approach the theoretical frequency.

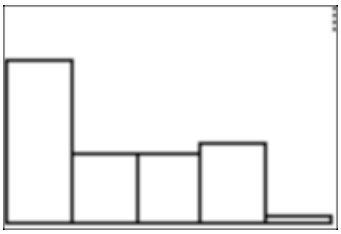
Before you run the program:
Set the window as below. And set STAT PLOT (note the new Type, a histogram):



Note that Xscl=1. Under FORMAT, Axes should be On
Now to run the program SPINNER, select it and press ENTER. You will see something like:

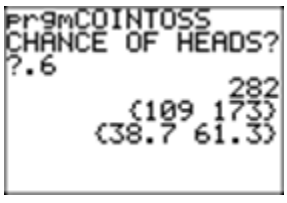






Here we see the *percentages* of outcomes landing on 1/3, 1/4, 1/5, 1/6, and 1/20 after 50 spins.
You may press ENTER for fifty more spins. I recommend spinning 500 times! After stopping the program (with ON 0 or ON ENTER), you can use TRACE and see the values for the five outcomes. The theoretical probabilities in whole percentages are 33, 25, 20, 17, and 5.

4. Tossing a (biased or unbiased) coin
A program in your calculator, COINTOSS, allows you to choose a bias towards heads for a coin toss, and then simulates tosses until you stop the program. The code for the program is at https://www.math.nmsu.edu/~breakingaway/Statistics/Lessons/TossingACoinUnits/tossing_coin_biased_unbiased.html
You may ask the program at any time, without stopping it, what the current outcomes of heads and of tails are, together with their frequencies. Here is an example, where the bias chosen is .6 towards heads. (You may choose any bias, including .5 for a fair coin):



When the program was stopped, there had been 282 tosses, with 109 showing tails and 173 heads. Thus tails came up 38.7% of the time, and heads 61.3% of the time.
Press enter again to see the new outcomes.

5. Betting on how many distinct prime factors a number has

In this unit you pay one dollar to play, and you generate a random number between one and 10 million on your calculator, randInt(1,E7). Your task is to guess how many distinct prime factors the number has. If you guess correctly, you get three dollars. If you guess incorrectly, you lose your dollar.

We use WolframAlpha to factor the number, http://www.wolframalpha.com.

The results may surprise you! Here are some examples:

Examples

4940664=2^3*3*41*5021	4 distinct factors
5566115=5*23*29*1669	4 distinct factors
9564142 = 2*7*29*23557	4 distinct factors
3858649=193*19993	2
9062410=2*5*7*37*3499	5
873619=873619	1 (it is prime!)
1472451=3*467*1051	3
2618824=2^3*13^3*149	3
6807953=29*181*1297	3
1702297=491*3467	2
925217=925217	1 (another prime!)
2068938=2*3^2*114941	3
8898764=2^2*7*19*43*289	5
8434026=2*3^2*468557	3
4999483=59*84737	2

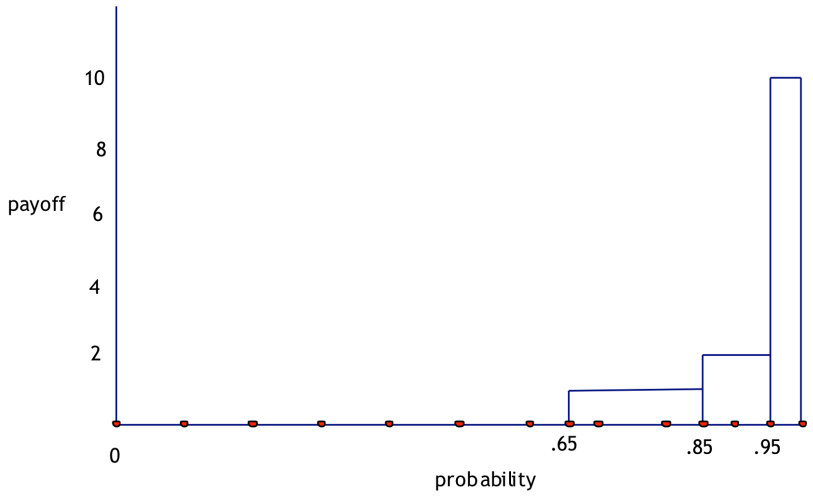
6. Mini Slot Machine

You may pretend that you are playing for pennies, or for dollars!

The probability of the payoff (for a 1¢ play) are:

payoff	0¢	1¢	2¢	10¢
probability	.65	.2	.1	.05

So the average return for 10¢ is 9¢. This means that the expected profit rate for the casino is 10% per play. In real casinos you sometimes know the rate of profit, but you are never told the actual distribution. This is how the payoff looks as a step function:



To run the program, select prgmMINISLOT. Here is my output for 34 tosses.
The first number is the trial number, the second is the payoff for that trial, and the third is the amount that I have won so far.

```
prgmMINISLOT
               {1 1 1}
               {2 2 3}
               {3 0 3}
               {4 1 4}
               {5 1 5}
               {6 0 5}
               {7 1 6}
               {8 0 6}
               {9 0 6}
              {10 0 6}
              {11 0 6}
              {12 0 6}
              {13 2 8}
              {14 1 9}
              {15 0 9}
             {16 10 19}
              {17 0 19}
              {18 0 19}
              {19 0 19}
              {20 0 19}
              {21 0 19}
              {22 1 20}
              {23 2 22}
              {24 2 24}
              {25 2 26}
              {26 0 26}
              {27 0 26}
              {28 0 26}
              {29 0 26}
              {30 2 28}
              {31 2 30}
              {32 2 32}
              {33 1 33}
              {34 1 34}
```

After 34 plays, I broke even!

7. Jumping Flea (A Random Walk)
This is a "story problem" and not the description of a real situation.
Imagine a flea that is jumping around in a random fashion. At each jump, it chooses a direction, and any direction has the same chance of being chosen as any other. Also it chooses the length of each jump, which can be any number between 0 and some maximal length. The flea starts its jumping in the center of a circle with a radius that is 10 times longer than its longest jump.

Can you guess how much ground it will cover before it leaves the circle by jumping randomly? This story introduces the concept of a "random walk", which is important in solving some problems (e.g. in physics).
The task is to write an animated simulation that will trace the jumps of an imaginary flea and compute the total distance the flea covers. The program FLEA is in your calculators, and the code is at
https://www.math.nmsu.edu/~breakingaway/Statistics/Lessons/flea/flea.html

Set the MODE: CLASSIC, DEGREE, PARAMETRIC
Set the FORMAT: CoordOff, Axes:Off
Use ZOOM to set window: ZStandard, ZSquare
Set the window: Tmin=0,
Tmax=360,Tstep=3
Define a circle of radius 10,
\X1T=10cos(T)
Y1T=10sin(T)

Here are some runs:

```
prgmFLEA
                89.83523308
                       Done
```

```
prgmFLEA
                89.83523308
                       Done
prgmFLEA
                100.3476116
                       Done
```

For all the units above and many more, go to https://www.math.nmsu.edu/~breakingaway/ and click on Statistics. For more information, contact Pat Baggett or Andrzej Ehrenfeucht, baggett@nmsu.edu, andrzej.ehrenfeucht@colorado.edu.