Twitter Simulation

Domain: Statistics and Probability 6.SP.2 & 7.SP.2

Cluster: Develop understanding of statistical variability. (6.SP)

Standard: Understand that a set of data collected to answer a statistical question has a distribution which can be described by its center, spread, and overall shape.

Cluster: Use random sampling to draw inferences about a population. (7.SP)

Standard: Use data from a random sample to draw inferences about a population with an unknown characteristic of interest. Generate multiple samples (or simulated samples) of the same size to gauge the variation in the estimates or predictions.

Goal: The goal of this activity is for students to develop and run a simulation to create a sampling distribution to estimate an answer to a statistical question and to explore the effect of sample size on the variability within a sampling distribution.

Materials needed:

- paper bags
- bingo chips or counters of some kind of two colors
- recording sheet or paper/pencil or white boards to record sample data
- calculators
- Core Math Tools: General purpose tools, Statistics and Probability, Data Analysis (link provided here: http://www.nctm.org/standards/content.aspx?id=32706 (scroll to the bottom of the page.)

Context: We see proportions used on a daily basis in all sorts of media. For example, local television anchors may report that 60% of area voters are in favor of a school bond issue or that 45% of the American public favors a particular political candidate. All of these proportions are based on random sampling from the population.
In this activity we will use simulations to estimate the answer to the statistical question "What proportion of teachers have a Twitter account?"

Summary of Simulation: This activity simulates drawing a random sample from a population and estimating the true proportion of the population using both a hands-on activity and computer simulation. The secondary goal is for students to use a series of simulations changing sample sizes taken from a population to make a conjecture on the relationship between the sample size and the variability or spread of the sample data. Discussion should be fostered throughout the simulations especially those questions asking students to make conjectures and explain their reasoning or thinking behind their conjectures.

Set Up for Hands-On Simulation:
•Students are in pairs or small groups
•Each group is given a bag with 60 green chips and 40 yellow chips. They should NOT look in the bag.
•Students randomly draw 20 chips from the bag without replacement. They will plot the proportion of green chips on a line plot.
•Replace the chips in the bag, shake it up, and draw 20 again. Each group should repeat this process for a total of 10 trials plotting the proportion of green chips each time.

Questions:
•Based on your sampling distribution, what proportion of teachers have Twitter accounts?
•How does your estimated proportion compare with other groups' estimates or the rest of the class?
•What is the range of your sampling distribution?
•How does the range of your sampling distribution compare with other groups' estimates or the rest of the class?
•What do you think might happen to the range of your sampling distribution if the sample size for each trial was increased to 40? 80? 120?

Testing our conjectures about increasing the sample size.
Teacher Notes: One way to help students see the relationship between the sample size and the spread of the data is plot the sample size versus the range for 20, 40, 60, 80, 100.

Using the Core Math Tools:
•Open Core Math Tools and select Data Analysis
•Enter data into Column A (or whatever column you want). Use 1 to represent green chips and 0 to represent yellow.
•Under Statistics, select Distribution of Samples and select the column with your data.
•Under the Controls menu, select Reset.
•Under the Options menu, make sure to de-select with replacement (you could have the students re-run this activity using with replacement and compare the results)
•Set the sample size.
•Set the number of runs.
•Select Conduct.
•Once the sample runs are completed, under the Tool menu select Descriptive Statistics. (To find the range students will need the Minimum and Maximum values. The standard deviation and the variation are also given and could be used as well.)
•Select Reset again from the Controls menu, change the sample size and repeat the process.

Possible Questions:•Do you think the relationship between the sample size and the range of the sampling distribution is linear or non-linear? Why?

•After making a scatterplot for this data, what do you notice about the data? Is it linear or non-linear? Does it appear to be bounded/have asymptotes? Explain your reasoning.

Blood Donor Simulation

Domain: **Statistics and Probability 7.SP.8C**
Cluster: **Investigate chance processes and develop, use, and evaluate probability models.**
Standard: **8. Find probabilities of compound events using organized lists, tables, tree diagrams, and simulation.**
c. Design and use a simulation to generate frequencies for compound events. ***For example, use random digits as a simulation tool to approximate the answer to the question: If 40% of donors have type A blood, what is the probability that it will take at least 4 donors to find one with type A blood?***

Goal: The purpose of this activity is for students to create and run a simulation to approximate the probability of a specific event (at least four blood donors).
Materials Needed:

- Recording Sheet
- Core Math Tools, General Purpose Tools, Statistics and Probability, Simulation (link provided here: http://www.nctm.org/standards/content.aspx?id=32706 [Scroll to the bottom of the page.]
    - IF computer access is not available, similar simulations can be run with graphing calculators or 10-sided number cubes (or some other means of generating random numbers)
- Pencils
- Calculators

Summary of Simulation:
The goal is to represent a situation in which students can simulate a person with Type A blood being selected. Ideally you would use the software being utilized in this session; however, other tools can be used to achieve similar results. For example, out of any ten numbers, four of them (40%, which is the probability of a person having Blood Type A) represent selecting a person with Blood Type A. Selecting the remaining six digits indicates someone with blood type other than A. A nice context might be something along the lines of the following scenario: Suppose you are volunteering at a blood donation center and you are required to keep track of how many people come in before a person comes in who has Type A blood. Discussion about how to collect and tally this data should be encouraged. A frequency table collecting the number of persons entering the blood donation center until the nth person with Blood Type A enters will facilitate data collection. The frequency table will be constructed in this fashion:

Number of Persons Entering Until Finding a Person with Type A Blood	Tally	Frequency
One (1st person entering had Type A)		
Two (2nd person entering had Type A)		
Three (3rd person entering had Type A)		
Four (4th person entering had Type A)		
Five (5th person entering had Type A)		
Six (6th person entering had Type A)		

Example Data:

1st trial: A This outcomes indicates a Type A blood donor was the first person entering

2nd trial: A Again, this outcomes indicates a Type A blood donor entering next

3rd trial: B No, so go to the next person

B No, so go to the next person

O No, so go to the next person

A In this trial we have the fourth person entering as Type A so tally in the "four"

4th trial: O No, so go to the next person

A In this trial we have the second person entering as Type A so tally in the "two"

5th trial: O No, so go to the next person

AB No, so go to the next person

A In this trial we have the third person entering as Type A so tally in the "three"

6th trial: B No, so go to the next person

O No, so go to the next person

B No, so go to the next person

AB No, so go to the next person

B No, so go to the next person

A In this trial we have the sixth person entering as Type A so tally in the "six"

Based on these outcomes, our table might look something like this:

Number of Persons Entering Until Finding a Person with Type A Blood	Tally	Frequency
One (1st person entering had Type A)	//	
Two (2nd person entering had Type A)	/	
Three (3rd person entering had Type A)	/	
Four (4th person entering had Type A)	/	
Five (5th person entering had Type A)		
Six (6th person entering had Type A)	/	

In this example we had six trials and of those 6, 2 outcomes required four or more people before we found a donor with Type A blood. Based on this data, the approximate probability it will take at least 4 donors before we find one with Type A blood would be 2 out of 6 or approximately 33%. (This is a really small sample, so more data needs to be collected.)
To run the simulation from the Core Math Tools, use the Menu options to select:
Build/Other Simple Events/Blood Type
We will run the simulation conducting 1 trial at a time until we reach Blood Type A. At that point we will tally the appropriate outcome on our table.
Move to Overhead Frequency Distribution; discuss the finite number of outcomes listed vs. possibilities. Why is it not necessary to include individual outcomes after the "Five"? What could a new chart look like?
Run Simulation, collect data, discuss results

We are looking for an approximation for the probability that it will take at least four people before we find one with Type A blood. The grayed region/rows suggest the tallies we are interested in that represent four or more people to find on with Type A blood. Since the blood types are randomly generated, it is possible to have 10, 14, 21, etc. people enter before we found a donor with Type A blood. This suggests we really only need to extend the first column to have the following: one, two, three, four or more.

Investigating Compact Car Data

Domain: **Statistics and Probability 8.SP.1, 8.SP.2, 8.SP.3**
Cluster: **Investigate patterns of association in bivariate data.**
Standard: **1. Construct and interpret scatter plots for bivariate measurement data to investigate patterns of association between two quantities. Describe patterns such as clustering, outliers, positive or negative association, linear association, and non-linear association.**
2. Know that straight lines are widely used to model relationships between two quantitative variables. For scatter plots that suggest a linear association, informally fit a straight line, and informally assess the model fit by judging the closeness of the data points to the line.
3. Use the equation of a linear model to solve problems in the context of bivariate measurement data, interpreting the slope and intercept.

Goal: Students will investigate data on the weight and mileage of compact cars. Based on a scatter plot of the data, they will describe important features of the data set, consider whether an outlier should be included or omitted, fit a line to the data, and use the line to make predictions about cars with different weights.

Materials Needed:

- Core Math Tools, General Purpose Tools, Statistics and Probability, Data Analysis (link provided here: http://www.nctm.org/standards/content.aspx?id=32706 [Scroll to the bottom of the page.]
    - If computer access is not available, students could create a scatter plot on graph paper and obtain similar results.
- Pencils
- Calculators

Predicting Gas Mileage for Compact Cars

The table at the right contains data on the weight and mileage of 16 different compact car models (Source: Edmunds.com). If possible, we would like to use the weight of the car to make a reasonable prediction for the car's mileage.

(a) Make a scatter plot for the data. Why is a scatter plot a better display to consider than histograms for the two columns?

(b) Describe important features of the scatter plot (patterns in the data and/or points that don't fit the pattern).

(c) Should unusual points (if any) be eliminated from the data set in this case? Why or why not?

(d) Construct a line that fits the data reasonably well, and find an equation for this line. What (if anything) does the y-intercept mean in this situation?

MODEL	WEIGHT in 100 lbs.	MPG
Audi A4	34.50	32
Chevrolet Cobalt	32.16	32
Ford Focus	26.36	34
Honda Civic	26.90	40
Honda Civic Hybrid	28.75	51
Hyundai Accent	24.03	36
Kia Spectra	29.72	35
MAZDA3	28.11	34
Mercedes-Benz C280	34.60	28
Nissan Sentra	28.97	36
Saturn ION	28.05	32
Subaru Impreza	30.67	28
Suzuki Aerio	27.16	31
Toyota Corolla	25.95	38
Toyota Yaris	23.26	39
VW Rabbit	29.11	30

(e) If we plan to buy a compact car that weighs 3000 pounds, what kind of gas mileage should we expect? If we add a 100-lb luggage rack, approximately how much can we expect our mileage to drop?

Some Like It Hot

Domain: **Statistics and Probability 8.SP.4**
Cluster: **Investigate patterns of association in bivariate data.**
Standard: **4. Understand that patterns of association can also be seen in bivariate categorical data by displaying frequencies and relative frequencies in a two-way table. Construct and interpret a two-way table summarizing data on two categorical variables from the same subjects. Use relative frequencies calculated on rows or columns to describe possible association between the two variables.**

Goal: Students will collect data and investigate a possible relationship between gender and liking hot, spicy food.

Materials Needed:

- Core Math Tools, General Purpose Tools, Statistics and Probability, Spreadsheet (link provided here: http://www.nctm.org/standards/content.aspx?id=32706 [Scroll to the bottom of the page.]
    - If computer access is not available, students could make frequency and relative frequency tables on paper and obtain similar results.
- Pencils
- Calculators

Some Like It Hot

On the Food Network's *Heat Seekers,* "Spicy-food gurus Aarón Sanchez and Roger Mooking traveled across the country to find the most mouth-burning plates – and tasted every one of them" (Source: http://www.foodnetwork.com/heat-seekers/index.html). Is this more of a guy thing?

(a) Collect data and fill in the frequency table below.

	Number of males	Number of females	Total
Like hot food			
Don't like hot food			
Total			

(b) Based on the data, find the relative frequencies in the table below.

	Rel. Freq. among males	Rel. Freq. among females	Rel. Freq. within entire data set
Like hot food			
Don't like hot food			

(c) Does liking hot food appear to be a guy thing? Explain why or why not.

(d) Suppose you set up the table differently (see below). Could you answer the same question? Fill in the table and explain.

	Rel. Freq. of males	Rel. Freq. of females
Among people who like hot food		
Among people who don't like hot food		
Within entire data set		