8 –Probability and Simulation

Problem set 8-0

1. In class you will be randomly divided into two groups, the “real coin” group and the “imaginary coin” group.

   Real Coin - The real coin group is to flip one coin 100 times and after each flip record an “H” if the coin turns up heads or “T” if the coin turns up tails in the table on the next page. It is important that you follow these instructions carefully. Use only one coin. Flip the coin once, record the result, repeat 100 times.

   Imaginary Coin - The imaginary coin group is to pretend they are flipping one coin 100 times and after each pretend flip record an “H” if you imagine the coin turns up heads or “T” if you imagine the coin turns up tails in the table on the next page. It is important that you follow these instructions carefully. Use only one imaginary coin. Imagine one flip, record the result, repeat 100 times.

   Write down your name, but not whether you are using a real coin or an imaginary coin.

Coin Flip Table

Name _____

Write an “H” or “T” in each box after flipping the real or imaginary coin.

1.	34.	67.
2.	35.	68.
3.	36.	69.
4.	37.	70.
5.	38.	71.
6.	39.	72.
7.	40.	73.
8.	41.	74.
9.	42.	75.
10.	43.	76.
11.	44.	77.
12.	45.	78.
13.	46.	79.
14.	47.	80.
15.	48.	81.
16.	49.	82.
17.	50.	83.
18.	51.	84.
19.	52.	85.
20.	53.	86.
21.	54.	87.
22.	55.	88.
23.	56.	89.
24.	57.	90.
25.	58.	91.
26.	59.	92.
27.	60.	93.
28.	61.	94.
29.	62.	95.
30.	63.	96.
31.	64.	97.
32.	65.	98.
33.	66.	99.
		100.

Problem Set 8-1

1. Give an example of an event with:

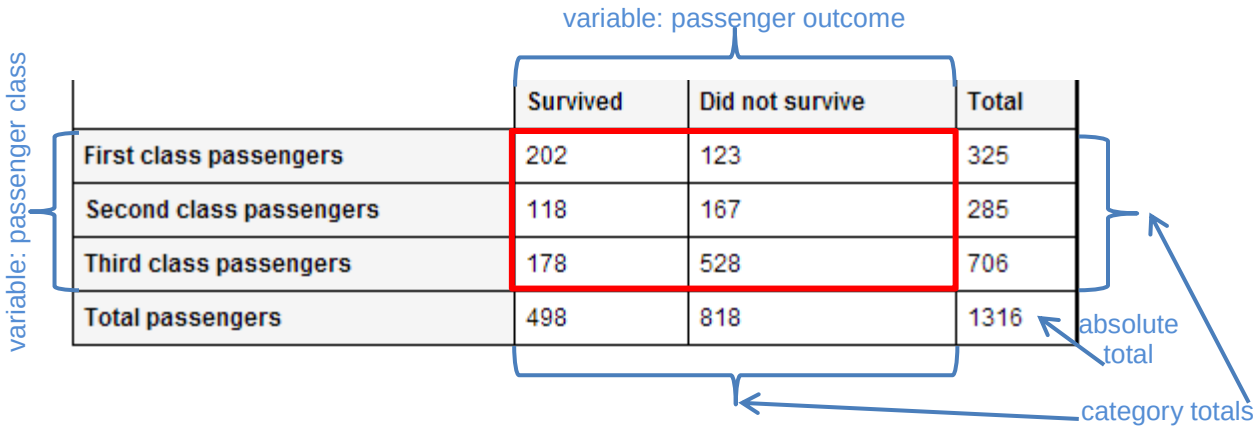
   a) probability = 0

   b) relative frequency = 1.

2. The data below describes passenger survival from the Titanic. (note, the data excludes crew members). One of the reasons there were so many fatalities was that for aesthetic reasons the ship did not carry enough lifeboats for its capacity. There was only room for a maximum of 52.9% of the boat's population in the lifeboats, but the survival rate was much less than even that.

   Source: *The Real Reason for the Tragedy of the Titanic*. The Wall Street Journal, n.d. Web. 2 Mar. 2013.

This data is organized in a **two-way table** – a table that classifies data based on possible categories for two different variables at the same time, one by rows and one by columns. It also includes the totals for each category and an absolute total.



	Survived	Did not survive	Total
First class passengers	202	123	325
Second class passengers	118	167	285
Third class passengers	178	528	706
Total passengers	498	818	1316

Data Table Source: *Common Core State Standards - Illustrations*. Illustrative Mathematics, n.d. Web. 27 Feb. 2013. <http://www.illustrativemathematics.org/illustrations/949>.

a) What is the sum of the numbers in the red box? Why?

b) What does the number 498 in the bottom row represent?

c) What is the relationship between the four numbers in the last column?

d) What's wrong with this question: "What is the probability that any passenger survived?"

e) Does this table give relative frequencies? Explain.

f) What was the relative frequency of survival?

g) What was the relative frequency of third class passengers on the ship?

h) What was the relative frequency of third class survivors?

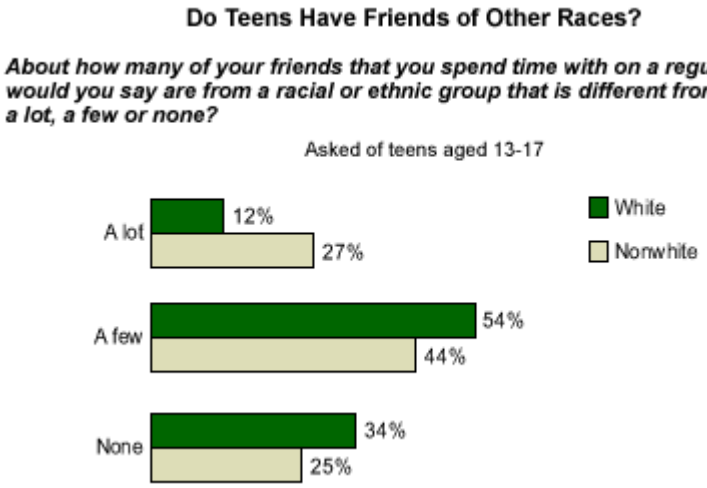
i) If we only consider survivors, what was the relative frequency of third class passengers?

j) Which is bigger between your answers for h and i and why?

k) Did class of passenger affect the likelihood of survival? Justify your answer with calculations, and if it did, provide a possible reason why that may have been the case.

l) Give one other example of other types of questions about relative frequency that can be answered with this table and answer your question. Show your work.

3. The following graph was published as part of a study on diversity involving teenagers done by Gallup in 2003.



Source: *Do Teens "Clique" With Diversity?* Gallup, n.d. Web. 27 Feb. 2013. <http://www.gallup.com/poll/10219/Teens-Clique-Diversity.aspx>.

a) If we asked the same questions of a random sample of 400 teenagers, 100 whites and 300 of other races and expected the same proportional responses that Gallup found, how many teens would fit each of the categories below? Use the graph to help you fill in the two-way frequency table. One example is given.

> A **two-way frequency table** represents frequency data broken down by possible outcomes in two different categories, or events. The last row and column give us totals for each outcome across the other category, and the bottom right cell gives us the total number of responses/trials.
> **Note**: a two-way frequency table is a specific type of two-way table.

	White	Nonwhite	**Total**
A lot		81	
A few			
None			
No answer			
Total			

b) Why was the row "No answer" necessary when it wasn't part of the graph? What observation about the graph makes that row necessary?

4. The tables you have seen in the previous problems were both two-way frequency tables. **A two-way relative frequency table** looks very similar, but the cells are filled with numbers between 0 and 1, representing the percentage of outcomes that fall in a certain category.

   Let's take the given frequency table of adult smokers (18 and older) in the US in 2011 and turn it into a relative frequency table:

	Women	Men	Total
Smoker	20,167,319	24,928,711	45,096,030
Non-smoker	102,058,854	90,481,988	192,540,843
Total	122,226,173	115,410,699	237,636,872

By the way, you might find it interesting that both the total number of men and total number of women in the US in this data set (provided by an affiliate of the Census Bureau) has a margin of error of ± 29,450 people…. So don't take it too literally!

Source: *Age and Sex - 2011 American Community Survey 1-Year Estimates*. US Census Bureau, n.d. Web. 28 Feb. 2013. <http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_1YR_S0101&prodType=table>.

Source: *Current Cigarette Smoking Among Adults — United States, 2011*. Center for Disease Control and Prevention, n.d. Web. 28 Feb. 2013. <http://www.cdc.gov/mmwr/preview/mmwrhtml/ mm6144a2.htm>.

a) In the frequency table, what does the number in the bottom right corner represent?

b) In a relative frequency table, what percentage would represent all responses?

To create a relative frequency table from a frequency table, divide the value in every cell by the absolute total.

c) Explain why the instruction above creates a relative frequency table. Then use it to fill in the relative frequency table below, rounding to the third decimal place. One cell is done for you.

	Women	Men	Total
Smoker			
Non-smoker			
Total			1

d) Do the total row and total column still add to the appropriate total in the bottom right cell? Would that always be the case when creating a relative frequency table from a frequency table?

e) What does the value 0.105 represent?

f) What does the value 0.514 represent?

Problem Set 8-2

1. Give an example of two dependent events:
   a) that have a large effect on each other.
   b) that have a small effect on each other.

2. Give a one sentence argument for each pair of events as to whether or not you think they are independent or dependent.
   a) first letter of your first name and last letter of your math teacher's last name
   b) country you live in and the number of syllables in your name
   c) eye color and hair color
   d) birthdate and annual salary
   e) gender and cancer

3. Let's take another look at the Titanic data from the last problem set and see how it relates to the ideas of conditional relative frequency and independent events.
   See reference in problem 8-1-2

	Survived	Did not survive	Total
First class passengers	202	123	325
Second class passengers	118	167	285
Third class passengers	178	528	706
Total passengers	498	818	1316

Image Source: *Common Core State Standards - Illustrations*. Illustrative Mathematics, n.d. Web. 27 Feb. 2013. <http://www.illustrativemathematics.org/illustrations/949>.

   a) Look back at the questions you answered about this table in problem set 8-1. Did any of those questions involve a condition? If so, which ones? How did you include the condition when solving the problem?
   b) What is the difference between these statements? Which of the statements is correct and why?

      "Most second-class passengers did not survive."

      "Most of the people who didn't survive were second class passengers."

   c) Are the events "passenger survived" and "passenger was in first class" independent events? Support your answer.

4. If two events are dependent, does that mean that one *causes* the other? Why or why not? Recall our conversations in Chapter 2 about causation and use examples to help you explain your answer.

Problem Set 8-3

Problem 1 is intended as an in-class pair/group activity. The rest of the problems are intended as homework.

1. Titanic Data Revisited

This problem is based on Illustrative Mathematics Common Core State Standards S-CP The Titanic 3.

We have looked at survival data from the Titanic twice, but only broken down by the class/deck of the passenger. Here is a more detailed data table that shows survival frequencies based on class and whether the passenger was a child, woman or man.

	Survived	Did not survive	Total
Children in first class	6	0	6
Women in first class	140	4	144
Men in first class	57	118	175
Children in second class	24	0	24
Women in second class	80	13	93
Men in second class	14	154	168
Children in third class	27	52	79
Women in third class	76	89	165
Men in third class	75	387	462
Total	498	818	**1316**

Source: British Parliamentary Papers, Shipping Casualties (Loss of the Steamship "Titanic") 1912, cmd 6352 *'Report of a Formal Investigation into the circumstances attending the foundering on the 15th April 1912, of the British Steamship "Titanic" of Liverpool, after striking ice in or near Latitude 41°46'N., Longitude 50°14'W., North Atlantic Ocean, whereby loss of life ensued.'* (London: His Majesty's Stationery Office, 1912), page 42

In Problem Set 8-2, you found that first class passengers were more likely to survive than second class, and second class more so than third. Some might believe that the rescue procedures were biased based on class. However, Victorian morality would have required the lifeboats be loaded with "women and children first".

Based on the data, who do you think was given priority in boarding the lifeboats? Write an argument that includes calculations to support your conjecture and be sure to analyze both aspects of the data: sex/age and class. Note, there are many ways to approach this problem, and you might want to make several different types of calculations before starting to write your analysis.

Extension: This table gives additional data on crew survival, which now allows us to address the entire ship population of 2,224.

	Survived	Did not survive	Total
Women in crew	20	3	23
Men in crew	192	693	885

Given this data and the previous table, which category (row) of people had the lowest relative frequency of survival? Why do you think that might be the case?

Use this additional data to amend your earlier calculations by including "crew" in the possible classes of passengers, and then including crew women with the rest of the women, and crew men with the rest of the men.

2. Out of an advisory of 10 students, some made varsity sports teams and some were chosen for honor roll. Here are the rosters:

Advisory Roster	Picked for Varsity	Picked for Honor Roll
Adam	Adam	Anna
Anna	Anna	Kim
Celeste	Celeste	Thomas
Kim	Kim	
Lucia	Miguel	
Miguel	Ming	
Ming		
Sebastian		
Sylvia		
Thomas		

a) If you were asked to find the relative frequency of advisory students on Varsity AND Honor Roll, which of the following two strategies would you prefer and why?

   i. Go down the Advisory roster, mark all the kids that made Varsity with a V. Then go down the list again, and mark all the kids that made Honor Roll with an H. Go down the list again, and count the number of kids that have VH next to their name, divide by the size of the advisory group.

   ii. Go down the Varsity list and mark the kids that are also on Honor Roll with an H. Count those students, and divide by the size of the advisory group.

b) What is the relative frequency of advisory students on Varsity AND Honor Roll?

3. Use the relative frequency table below to answer the following questions

	Ever Bullied		
Height	**Yes**	**No**	**Total**
Short	0.20	0.24	0.44
Not Short	0.14	0.42	0.56
Total	0.34	0.66	1

Data set provided by Floyd Bullard
Source: *Statistical Ideas and Methods,* Utts and Heckard, p. 166; England, Voss and Mulligan (2000)

a) Let's investigate the relative of frequency of kids that meet the description short OR bullied.

Consider the following statement:

$$RF(Short\ OR\ Bullied) \stackrel{?}{=} RF(Short) + RF\ (Bullied)$$

Is that statement true or not? Justify your answer.

b) Use your finger to trace the row and then the column that represent the kids you included to get your answer to *RF(Short) + RF (Bullied)*. What cell did you cross twice, meaning it was double counted? What group of kids does that represent?

c) Fix your answer in part a) by subtracting the value of the cell you double counted. That is RF(Short OR Bullied).

d) How would you describe the group of kids that does not satisfy either of the conditions short or bullied using RF notation? What is the RF of that group (hint: you can read it from the chart)

e) What is the relationship between the answers to c) and d)? Why?

Problem Set 8-4

1. Thunderstorms involve rain and lightning, but those events can happen without each other. (Technically dry lightning involves rain that evaporates before it hits the ground. It's also associated with causing wildfires. Source: *Dry Lightning*. Wikipedia, n.d. Web. 26 May 2013. <http://en.wikipedia.org/wiki/Dry_lightning>.)

   This problem is based on Illustrative Mathematics Common Core State Standards Illustration Source: *S-CP Rain and Lightning*. Illustrative Mathematics, n.d. Web. 26 May 2013. <http://www.illustrativemathematics.org/illustrations/1112>.

   a) If today's weather report states a 60% chance of rain, 15% chance of lightning, and 20% chance of lightning if it's raining, then what's the chance of rain AND lightning today?

   b) Given a 55% chance of rain, 20% chance of lightning, and 15% chance of lightning and rain, then what's the chance of rain OR lightning today? What's the chance of neither?

   c) Given a 50% chance of rain, 60% chance of rain or lightning, and 15% chance of rain and lightning, then what's the chance of lightning today?

2. *This problem is based on a similar problem from Floyd Bullard's handout "Some Short Probability Lessons"* Suppose a random sample of 1000 college students was polled on their magazine readership.

   a) The table below gives a possible breakdown for readership of Sports Illustrated. The numbers are given for both women and men.

	Reads SI	Does not read SI	Total
Women	110	590	700
Men	90	210	300
Total	200	800	1000

   Who is more likely to read SI: women or men? How do you know?

   b) During the same time period, those students bought an equal number of copies of National Geographic as copies of Sports Illustrated. However, unlike SI, men were just as likely as women to read National Geographic. Complete the table below with numbers that are consistent with that fact.

	Reads NG	Does not read NG	Total
Women			700
Men			300
Total	200	800	1000

c) With the data in these tables, can you calculate the relative frequency of women that read SI AND NG? If so, calculate it. If not, explain why you can't.

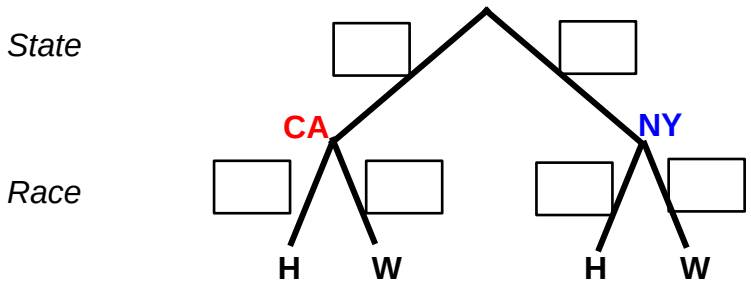
d) Answer part c) for RF (SI OR NG).

3. According to 2007 US Census data, approx. 13.1% of all babies born to residents of the US that year were born in CA and approx. 5.9% in NY. In CA, approx.. 52.5% of babies born were Hispanic, and approx. 27.5% were white. In NY, approx. 23.7% were Hispanic and approx. 49.6% were white.

Source: Births, Deaths, Marriages, and Divorces - Table 82. US Census Bureau, n.d. Web. 1 Mar. 2013. <http://www.census.gov/prod/2011pubs/11statab/vitstat.pdf>.

This type of data can be organized in a **tree diagram**. A relative frequency (or probability) tree diagram records relative frequencies (or probabilities) as decimals on branches for the possible responses for each category of information. (note: for the sake of calculation, it may be more convenient to write percentages as decimals between 0 and 1, inclusive.)

a) Fill in the relative frequencies from the information above on the following tree: (note: a complete tree would have a branch at the first level for each state, but this page isn't that big!)



b) If 0.059 represents RF(NY), what does 0.496 represent? Answer in notation.

c) If you were to trace the leftmost path down the tree and multiply the numbers 0.131 and 0.525, what have you calculated? Answer in notation.

d) Calculate the relative frequency of U.S. babies born in CA or NY.

e) Calculate the relative frequency of U.S. babies that are White New Yorkers.

f) Can we use this table to calculate the relative frequency of Hispanic babies? Why or why not?

Problem set 8-5

1. Create a simulation with Excel* to calculate the relative frequency of the number of times the sum of two dice is 7 in 10,000 rolls.
   a. If two fair dice are rolled, based on the simulation, give an estimate of the probability that the sum is 7?
   b. What is the smallest possible relative frequency and what would that have meant within the context of the simulation?
   c. If the relative frequency in your simulation would have been the smallest possible value you answered in (b), what would be the corresponding probability and what would that mean within the context of rolling two dice?
   d. What is the largest possible relative frequency and what would that have meant within the context of the simulation?
   e. If the relative frequency in your simulation would have been the largest possible value you answered in (d), what would be the corresponding probability and what would that mean within the context of rolling two dice?
   f. Give an event, E, such that the probability of the event occurring is 0, that is P(E)=0.
   g. Give an event, E, such that the probability of the event occurring is 1, that is P(E)=1.

   *Before you can do a simulation with Excel, you will need to open Excel → Tools → Add-ins…→ ☑ Analysis ToolPak and ☑ Analysis ToolPak – VBA → OK

Problem set 8-6

1. A *frequency distribution* is a table that shows the frequency of events; the events must be non-overlapping events. A ***relative*** *frequency distribution* is a table that shows the relative frequency of *disjoint* (non-overlapping) events. Supposed you asked 10 people to tell you their age and you got the following responses: 14, 14, 15, 15, 15, 16, 16, 17, 17, 18. The frequency and relative frequency distributions are shown below.

Frequency Distribution	
Age	Frequency
14	2
15	3
16	2
17	2
18	1

Relative Frequency Distribution	
Age	Relative Frequency
14	0.2 or 20%
15	0.3 or 30%
16	0.2 or 20%
17	0.2 or 20%
18	0.1 or 10%

   Go back to rolling two dice in which there are 11 possible sums: 2-12. Note these are disjoint events. Create a simulation in Excel with 10,000 trials that gives the relative frequency distribution of sums. The purpose of creating a relative frequency distribution is to be able to estimate the *probability* distribution of sums.

2. In a family with three children, what is the probability that the family will have 1 boy and 2 girls? Assume that there are no multiple births.*
   a. Create a simulation that gives an estimate of the probability above by finding the relative frequency of 1 boy and 2 girls in a family of three children.

   b. Create a simulation that gives the relative frequency distribution for the number of boys, that is for:
      0 boys and 3 girls
      1 boys and 2 girls
      2 boys and 1 girls
      3 boys and 0 girls

      *The assumption of no multiple births would not be appropriate if you wanted a very accurate answer to the question. "In the past two decades, the number of multiple births in the United States has jumped dramatically. Between 1980 and 2000, the number of twin births has increased 74 percent, and the number of higher order multiples (triplets or more) has increased fivefold, according to the National Center for Health Statistics. Today, about 3 percent of babies in this country are born in sets of two, three or more, and about 95 percent of these multiple births are twins". (1) "Twins means that there are 2 fetuses in the uterus. Seventy percent of twins are fraternal (not identical). This is when there are two eggs released by the mother. They are fertilized by 2 sperm. They are like siblings born at the same time. Except for being born at the same time, fraternal twins are no more alike than other brothers and sisters born to the same parents at a different time and place". (2)

Sources:
(1) *Multiples: Twins, Triplets and Beyond*. March of Dimes, n.d. Web. 6 Mar. 2003. <http://www.marchofdimes.com/681_4545.asp>.
(2) *Dichorionic Twins*. Dartmouth-Hitchcock Medical Center, n.d. Web. 6 Mar. 2003. <http://www.dartmouth.edu/~obgyn/mfm/PatientEd/twins_didi.html>.

Problem set 8-7

1. A company has 400 employees, 320 are men and 80 are women. Due to financial difficulties, the company needs to lay off 10 employees and they tell the employees that they will do so randomly. When the layoffs are announced, 5 of those laid off are men and 5 are women. The 5 women are considering filing a class action suit against the company and hope that they can demonstrate that it is highly unlikely that if the employees were fired at random that 5 women and 5 men would be fired.

   a. The women have asked for your help as they decide whether to sue the company. Create a simulation in Excel that picks 10 employees at random and identifies the number of women laid off. Run the simulation 100 times and with paper and pencil create a relative frequency distribution of the number of women laid off.

   b. If the women file the class action suit and ask you to appear as an expert witness, what information will you provide to support their claims?

   c. Now think of the situation from the perspective of a member of the jury. Given the information the expert witness provides, would you vote in favor of the plaintiff (the women) or the defendant (the company)?

Problem set 8-8

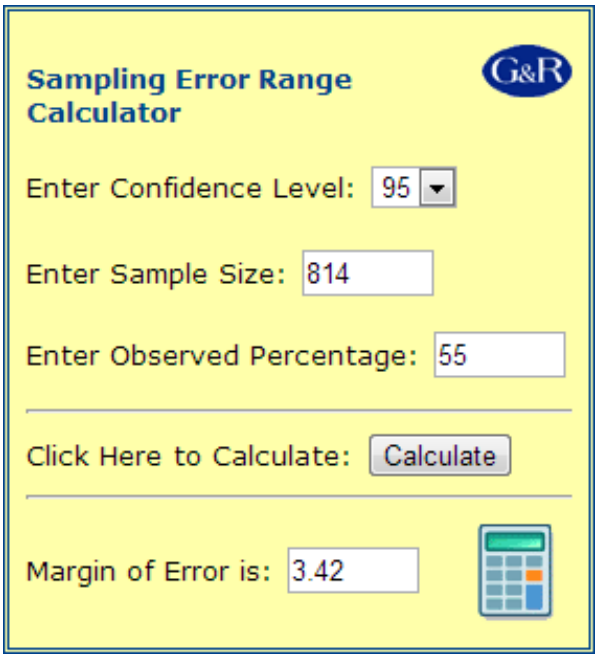
On January 14-15, 2013, CNN/Time/ORC conducted a survey asking Americans, "Do you favor or oppose stricter gun control laws"? According to the poll, "Fifty-five percent of Americans favor stricter gun laws". For this poll, like almost all professional polls, they reported four numbers:

55% of those surveyed were in favor of stricter gun control (the **sample proportion**)
814 was the **sample size** of adult Americans (usually about 1,000)
95% **confidence level** (this number is the same for almost all polls)
3.5% **margin of error**

The sample proportion +/- the margin of error gives the confidence interval. In this case 55% +/- 3.5% gives the confidence interval 51.5% to 58.5%. You can be 95% certain that the interval 51.5% to 58.5% captures the true population proportion.

Polling is expensive, costing about $2,000 for a typical poll according to Precision Polling; the smaller the sample size, the larger the margin of error. So, polling companies strike a balance between high expense and high margin of error.
Source: *How much will it cost to get 500 completes?* Precision Polling, n.d. Web. 28 Feb. 2013. <http://www.precisionpolling.com/about/faq#How_much_will_it_cost_to_get_500_completes>.

There are many "Margin of Error" web calculators. One such website can be found at http://gandrllc.com/setable.html. The image at right shows how the website was used to calculate the margin of error for the Gun Control Law poll. The website calculated a confidence interval of 3.42% and CNN rounded up (common practice) to 3.5%.



1. Use the website and trial and error to determine how big the sample size would need to be in order to lower the margin of error to 3.00%? What would the confidence interval be in that case and what does the confidence level say about this confidence interval?

   Source: Calculating Error Ranges. Gallup Robinson, n.d. Web. 5 Mar. 2013. <http://www.gallup-robinson.com/setable.html>.

2. If CNN wanted to reduce the margin of error of error to 1%, how big would the sample size need to be? Why do polling firms like CNN "settle" for margins of error like 3.5%?

3. What percent of Americans favor stricter gun control laws?

4. If a poll reports a confidence interval of 35.4% to 41.2%, what is the sample proportion and margin of error?

5. Our task at this point is to use simulation to “take a poll” over and over to see how the four numbers below relate to each other. Again, those numbers are:

   55% sample proportion
   814 sample size
   95% confidence level
   3.42% points margin of error (we will use the non-rounded margin of error)

   As we (hopefully) realized in problem 3 of this problem set, we do not know the true proportion of the population that is in favor of gun control, so instead we use the sample proportion to estimate the population proportion. For a moment, we are going to pretend we know that the population proportion is 55%. If we take repeated polls via simulation of sample size 814, then 95% of our confidence intervals should capture the assumed population proportion of 55%. In Excel, the simulation is accomplished with the following steps:

   a. Column A – Person # (The person being asked). Generate the numbers 1-814.

   b. Column B – Returns “F” for in Favor 55% of the time and “O” for Opposed 45% of the time. This can be accomplished with the formula =if(rand()<=.55,”F”,”O”). Note: We are pretending for a moment that we know the population proportion; of course we can never know this.

   c. Use the =countif function to find the frequency of “F”; =COUNTIF(B:B,"F").
      Use the =countif function to find the frequency of “O”; =COUNTIF(B:B,"O").

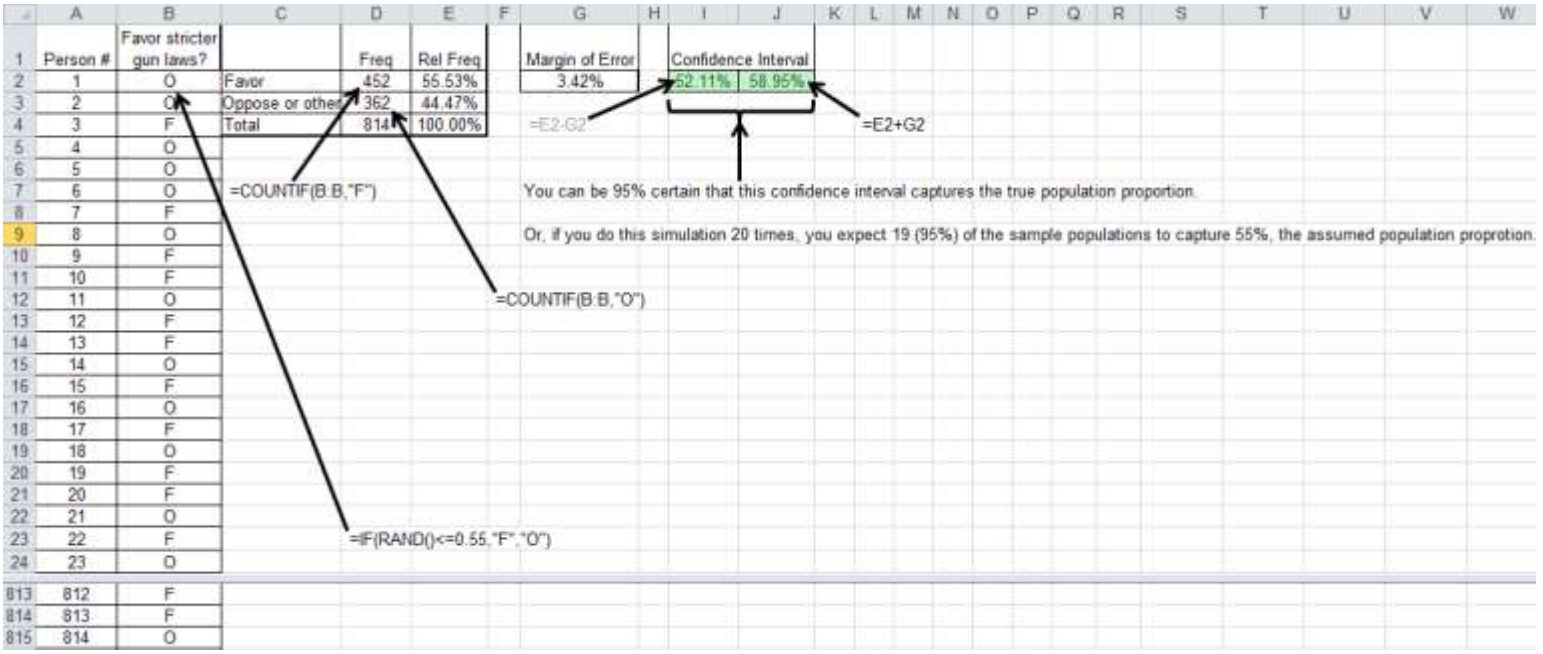
   d. Divide by the total frequency to find the relative frequency of “F” and “O” in column B. The relative frequency of “F”, being in favor of stricter gun laws, is the sample proportion.

   e. Create a cell for the margin of error which is 3.42%.

   f. Create the confidence interval by subtracting margin of error from the sample proportion and then adding the margin of error to the sample proportion.

   g. Run the simulation 20 times, and each time record the confidence interval and whether that interval captures the assumed population proportion of 55% (two examples: 52.97%-59.81%; yes and 55.06%-61.90%; no). The class will put the results of all students together in one large group of confidence intervals. We expect that 95% of the confidence intervals capture the assumed sample proportion of 55%.

The simulation should look as follows:



Problem set 8-9

1. During the Second World War, the German army produced tanks (Deutsche Panzer) and put a serial number (specifically the tire molds and tank gearboxes)[1] on each one. As the allied forces captured tanks they figured out fairly quickly that the Germans numbered the tank with serial numbers 1, 2, 3, and so on. This gave the allied forces a strategic advantage. Army statisticians told their commanding officers that they were confident they could estimate the *total* number of German tanks that had been produced by having the list of serial number of tank captured thus far. It turns out that the allied forces actually ended up having a better estimate of the number of tanks the German forces had, than the Germans themselves; the Germans, surprisingly, did not keep particularly accurate records of the tanks that were produced at various factories around the country.

   Your task now is the same task that the allied forces statisticians had then, namely to estimate the total number of tanks the Germans had based on the serial numbers of the tanks captured so far. What is so intriguing about this task is that there is no one right algorithm that best estimates the total number of tanks. Simulation is the perfect way to approach this problem.

   a. Let's do this first step in class as a class discussion or in small groups.
      If the first serial number is 50, what is your estimate of the total number of tanks? How did you arrive at you estimate?
      The next serial number is 30. Now what is your estimate?
      The next serial number is 60. Now what is your estimate?
      The next serial number is 70. Now what is your estimate?
      The next serial number is 5. Now what is your estimate?
      Describe the procedure or algorithm you used to arrive at these estimates.

   b. In order to test your procedure, we will first do the simulation as if we know the total number of tanks. Create a simulation that generates random serial numbers between 1 and 100 *without* repetition. Test the formula you or your classmates created.

   c. For the simulation you just ran, look what the two estimates are after all 100 tanks have been captured. Is there a refinement you can make to one or both of the formulas you used to create the estimates?

d. You not only need to be able to estimate how many tanks there are in total, but how confident you are in your answer. Look at the simulation you did in part b. How many serial numbers would you need before you would be willing to *confidently* give your estimate to a commanding officer?

Recreate the simulation knowing that there are 200 tanks. Now how many serial numbers would you need before you would be willing to *confidently* give an estimate to your commanding officer? Keep recreating the simulation so that you can fill in the table below. You may want to do this with a partner; each does half the work and then combine results in your table.

Total number of tanks	Number of tanks that need to be captured before you are "confident" of your estimate
100	
200	
300	
400	
500	
600	
700	
800	
900	
1,000	

e. Now comes the fun and creative part. So far we have used only two estimators, namely 2*mean and 2*median (with perhaps some refinements we discovered in part c). Can you come up with other estimators? If you are stuck, there are two hints on the next page…

(Part f continued)

Hint 1: Go back to assuming there are 100 tanks total. Can you come up with an order in which the first few tanks could be captured, such that you are *sure* there is a better estimate than the ones you have used so far?

Hint 2: If you search the web, you will find many other estimators. Try them out and see if you find one that you feel is better than the ones you have used so far. Write a sentence or two that describes why you switched to a new estimator or stuck with one of the two know estimators.

e. On the next page is a list of tank serial numbers in the order in which they were captured. (Note: you can copy them in the problem set below and then past them into Excel.) Two questions:
Having all of the serial numbers, how many tanks do you estimate there are?
After how many captured tanks would you be willing to report to your commanding officer that you have an estimate in which you are confident?

Sources:
1) *German Tank Problem.* Gary Smith, n.d. Web. 3 May 2008. <http://www.lhs.logan.k12.ut.us/~jsmart/tank.htm>.
2) *Unbiased and Minimum Variance Estimators*, Rowell , n.d. Web. 4 May 2008. <http://www.mathspace.com/NSF_ProbStat/Teaching_Materials/rowell/final/12_germantank_bcL7.doc>.
3) *German Tank Problem*, Wikipedia, n.d. Web. 3 May 2008. <http://en.wikipedia.org/wiki/German_tank_problem>.

Serial numbers for part e.

# of tanks captured	Serial #
1	46
2	408
3	258
4	425
5	261
6	459
7	249
8	412
9	456
10	15
11	52
12	228
13	40
14	159
15	295
16	126
17	36
18	406
19	186
20	294
21	24
22	39
23	181
24	244
25	423
26	310
27	33
28	445
29	200
30	164
31	94
32	124
33	410
34	18
35	282
36	160
37	180
38	116
39	138
40	90
41	10
42	309
43	436
44	477
45	80
46	260
47	165
48	62
49	147
50	214