

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.

Answers may vary, but an event with probability 0 is one that could NEVER happen and an event with probability 1 is one that MUST happen.

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?

The numbers in the red box must sum to 1316 because they represent the breakdown of all of the passengers into particular categories.

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

The number 498 represents the total number of passengers that survived.

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

The first three numbers in the last column represent the total numbers of first, second and third class passengers. Those three numbers sum to the fourth number in the column, representing the total number of passengers.

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

The term "probability" refers to the likelihood of a future event. Since the Titanic has already sunk, we should use the term "relative frequency".

e) Does this table give relative frequencies? Explain.

No. The table gives us frequencies, meaning the number of times a particular event like "first class passenger survived" occurred. The relative frequency of that event would require comparing it to (dividing it by) the total number of passengers.

f) What was the relative frequency of survival?

The relative frequency of survival is calculated by:

$$\frac{\text{number of passengers that survived}}{\text{total number of passengers}} = \frac{498}{1316} \approx 0.378 \text{ or } 37.8\%$$

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

$$\frac{706}{1316} \approx 0.536 \text{ or } 53.6\%$$

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

$$\frac{178}{1316} \approx 0.135 \text{ or } 13.5\%$$

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

$$\frac{178}{498} \approx 0.357 \text{ or } 35.7\%$$

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

The answer to part i is bigger because we are counting third class survivors out of a relatively small group, namely only survivors. Part h is smaller because we are counting third class survivors out of a bigger group, namely all passengers, so they represent a smaller percentage.

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.

Answers may vary.

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.

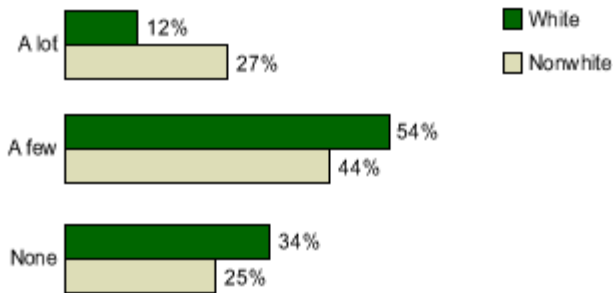
Answers may vary.

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

Do Teens Have Friends of Other Races?

About how many of your friends that you spend time with on a regular basis would you say are from a racial or ethnic group that is different from yours – a lot, a few or none?

Asked of teens aged 13-17



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	12	81	93
A few	54	132	186
None	34	75	109
No answer	0	12	12
Total	100	300	400

- 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?

If all students answered the question, then the sum of the relative frequencies for each type of student should be 100%. $12\%+54\%+34\%=100\%$, which tells us all white students answered the question with one of the three given answers. However, $27\%+44\%+25\%=96\%$, which tells us that 4% of Nonwhite students chose not to answer the question (or gave an answer that didn’t fit one of the given categories).

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 $\pm 29,450$ 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?
The total number of data points in the set, or in this case, people in the US.
- b) In a relative frequency table, what percentage would represent all responses?
1 as a decimal, or 100%

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.

A relative frequency is defined as the ratio of the number of times a particular event occurs compared to the total number of trials/data points, which gives us the percentage of the total that fits a particular category.

	Women	Men	Total
Smoker	0.085	0.105	0.190
Non-smoker	0.429	0.381	0.810
Total	0.514	0.486	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?

In this case, the total row and total column do still add to the overall total of 100%. That should always be the case when creating a relative frequency table from a frequency table if you account for all decimal places. If you're rounding the relative frequencies your totals may be slightly off.

- e) What does the value 0.105 represent?

0.105 tells us that in 2011, approximately 10.5% of the US population were men and smokers.

- f) What does the value 0.514 represent?

0.514 tells us that in 2011, approximately 51.4% of people in the US were women (bonus: why not 50%?)

Problem Set 8-2

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

Answers may vary.

- b) that have a small effect on each other.

Answers may vary.

2. Give a one sentence argument for each pair of events as to whether or not you think they are independent or dependent.

Note that you aren't asked to say if the events are definitely dependent or independent. You would have to have all the data in the world in order to prove empirically that two events are completely dependent or independent; hence the use of the term "Likely" in the solutions.

- a) first letter of your first name and last letter of your math teacher's last name

Likely Independent. America is a country of immigrants with quite a variety of names represented. In this situation it is quite unlikely that the first letter of your first name has anything to do with your math teacher's name.

- b) country you live in and the number of syllables in your name

Likely Dependent due to differences in languages/cultures. For example, people in Thailand tend to have long names while people in Korea tend to have short names.

- c) eye color and hair color

Likely Dependent. [the rest of the supporting statements are left for you to do]

- d) birthdate and annual salary

Likely Independent.

- e) gender and cancer

Answers may vary.

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?

Part i most certainly involves a condition. It even uses the term “given”. You incorporated the condition by changing the possible outcomes to be just the column of survivors, and the “total” outcomes to be the total number of survivors.

Part j asked you to explore how having a condition changed the relative frequency of a given event.

Your argument in part k likely involved conditions as well.

- 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.”

The first statement is correct because it is saying that when you look at the set of 285 second class passengers, more of them did not survive (167) than did (118), which is true. Note: Using the word “most” in this statement spins it to make you think vastly more did not survive than did, which isn’t really the case, so one could argue as to the degree of its “correctness”.

The second statement is saying that when you look at the set of 818 people that did not survive, that more of them were second class than first or third. This statement is incorrect, because there were far more third class passengers that didn’t survive (528) than second class (167).

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

Let S represent “passenger survived” and 1 represent “passenger was in first class”

$$\mathbf{RF(S)} = \frac{498}{1316} \approx 0.378 \text{ or } 37.8\%$$

$$\mathbf{RF(S|1)} = \frac{202}{325} \approx 0.622 \text{ or } 62.2\%$$

RF(S) \neq RF(S|1) therefore S and 1 are not independent, they are dependent.

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.

Saying two events are dependent does not mean that one causes the other. For example, the likelihood you wear a jacket depends on the weather. However, wearing a jacket does not cause the weather to be cold, and cold weather cannot force you to wear a jacket.

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.

This problem is intentionally open-ended; answers will vary.

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?
- 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.
 - 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.

While both strategies are valid and will yield the same answer, the second strategy is more efficient. Rather than always looking at the entire set of students, it takes advantage of the fact that we can first narrow down to the students that fit at least one of our criteria, and then look at which of those students satisfy both. It simply requires looking at fewer names.

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

Anna and Kim are the only two students that satisfy both criteria, out of ten advisees, so the relative frequency of students on both Varsity AND Honor Roll is

$$\frac{2}{10} = 0.2 \text{ or } 20\%$$

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:

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

Is that statement true or not? Justify your answer.

The category of kids that were Short OR Bullied would have to include all the Short kids, who had a relative frequency of 44% and the kids that were Not Short, but still Bullied, which was 14% for a total of 58%

The relative frequency of kids that were Short was 44% and the relative frequency of kids who were Bullied was 34%. $44+34 = 78\%$

$0.58 \neq 0.78$, therefore

$RF(\text{Short OR Bullied})$ DOES NOT EQUAL $RF(\text{Short}) + RF(\text{Bullied})$

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

The cell containing 0.20 was crossed twice, which means it was double counted since it was included in the totals of 0.44 and 0.34

That cell represents the kids that are Short AND Bullied

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

$$0.58 = 0.44 + 0.34 - 0.20$$

$$RF(\text{Short OR Bullied}) = RF(\text{Short}) + RF(\text{Bullied}) - RF(\text{Short AND Bullied})$$

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)

RF(Not Short AND Not Bullied) = 0.42

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

0.58 + 0.42 = 1, because the conditions in c) and d) together cover all possible types of outcomes

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?

Take how likely it is that it rains, 0.6 and then multiply it by the chance that there's lightning given that it's already raining: $0.6 \cdot 0.2 = 0.12$, so 12%

In notation: $P(R \cap L) = P(R) \cdot P(L|R) = 0.6 \cdot 0.2 = 0.12$, so 12%

- 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?

If we add the chance of rain and the chance of lightning, we get 75%, but we've double counted the chance of rain and lightning, 15%, so when we take away the double counting we get 60%

In notation: $P(R \cup L) = P(R) + P(L) - P(R \cap L) = 0.55 + 0.2 - 0.15 = 0.60$, so 60%

Having neither rain nor lightning is the complement of having rain OR lightning, so the chance of neither is 40%

In notation: $P(\sim(R \cup L)) = 1 - P(R \cup L) = 1 - 0.6 = 0.4$, so 40%

- 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?

**Since we know the chance of rain OR lightning, we know that encompasses the chance of rain and the chance of lightning but with the chance of both subtracted from it. So to get just the chance of lightning, we'd want to take the chance of rain OR lightning, subtract the chance of rain and add the chance of both:
 $0.6 - 0.5 + 0.15 = 0.25$ or 25%**

**In notation: since $P(R \cup L) = P(R) + P(L) - P(R \cap L)$; by algebraic manipulation
 $P(L) = P(R \cup L) - P(R) + P(R \cap L) = 0.6 - 0.5 + 0.15 = 0.25$ or 25%**

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?

$\frac{110}{700} \approx 15.7\%$ of the women read SI and $\frac{90}{300} = 30\%$ of the men read SI, so for this particular sampling, the men are almost twice as likely to read SI.

- 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

There is more than one way to solve this problem, but a system of equations is possible. Consider the four unknown numbers a, b, c and d.

	Reads NG	Does not read NG	Total
Women	a	b	700
Men	c	d	300
Total	200	800	1000

The table tells us that $a + c = 200$, and the problem tells us that $\frac{a}{700} = \frac{c}{300}$ so we

can solve one equation for one of the variables, $a = \frac{7c}{3}$, and substitute into the

other equation: $\frac{7c}{3} + c = 200$ so $\frac{10c}{3} = 200$ and $c = 60$. You could solve it using b and d instead if you like. Using the row and column totals to help, we get:

	Reads NG	Does not read NG	Total
Women	140	560	700
Men	60	240	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.

No, we cannot calculate the relative frequency of women that read SI AND NG because while we know $RF(SI)$ and $RF(NG)$, we have no idea what affect reading one magazine has the likelihood of reading the other. It could be that the women that read one magazine just like magazines and therefore are highly likely to read the other, or it could be two completely different audiences given the difference in subject matter.

- d) Answer part c) for $RF(SI \text{ OR } NG)$.

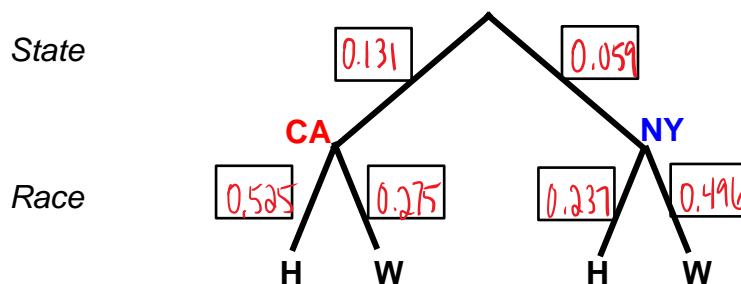
No, we cannot calculate the relative frequency of women that read SI OR NG because again, while we know $RF(SI)$ and $RF(NG)$, we do not know what the overlap is between the two, if there even is one. There could be a large overlap or none at all, we don't know.

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.

$0.496 = RF(W | NY)$, or what percentage of babies born in NY were White.

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.

$0.131 \cdot 0.525 = RF(CA) \cdot RF(H | CA) = RF(CA \cap H)$

In other words, we've calculated the relative frequency that any given baby born in 2007 was a Hispanic Californian

(which was $0.131 \cdot 0.525 = 0.0687$ or 6.87% of the total national births that year – do you think that's a lot? What kind of implications does that number have?)

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

$RF(CA \cup NY) = RF(CA) + RF(NY) - RF(CA \cap NY)$

In this case, the RF of babies born in CA or NY is just the sum of the RFs of CA and NY because there is no possible overlap that could be double counted.

$RF(CA \cup NY) = 0.131 + 0.059 = 0.19$ or 19%

(So almost 20% or $\frac{1}{5}$ th of births in the entire country in 2007 happened in NY or CA... it would be interesting to know the size of NY + CA relative to the entire country and compare)

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

$RF(NY \cap W) = RF(NY) \cdot RF(W | NY) = 0.059 \cdot 0.496 = 0.0293$ or 2.92%

(Can't you just feel the political ramifications of this data?!?)

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

No, we cannot, because "the relative frequency of Hispanic babies" implies in the US, since this data is coming from a national survey. We only have information about 2 states, so we cannot.

(In case you were wondering, it was almost 25%)

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? **Solution posted.**
 - b. What is the smallest possible relative frequency and what would that have meant within the context of the simulation? **0%; Sum of 7 did not happen.**
 - 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? **0%; Sum of 7 will never happen.**
 - d. What is the largest possible relative frequency and what would that have meant within the context of the simulation? **100%; Sum of 7 happened every time.**
 - 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? **100%; Sum of 7 will happen every time.**
 - f. Give an event, E, such that the probability of the event occurring is 0, that is $P(E)=0$.
Sample answer: Roll a die. P(8)
 - g. Give an event, E, such that the probability of the event occurring is 1, that is $P(E)=1$.
Sample answer: Flip a coin. P(H or T)

*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.*
- 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. **Solution posted.**
 - 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**

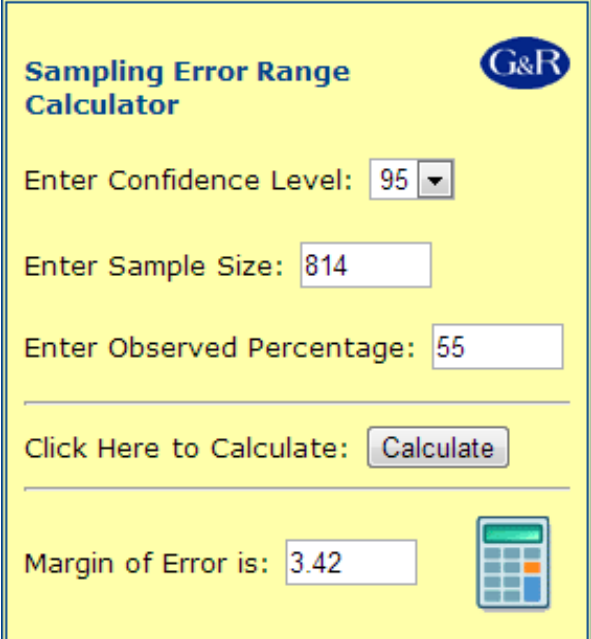
The sample proportion +/- the margin of error gives the confidence interval. In this case 55% +/- 3.5% gives the confidence interval 51.5% to 58.5%. You can be 95% certain that the interval 51.5% to 58.5% captures the true population proportion.

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%.



Sampling Error Range Calculator

Enter Confidence Level: 95

Enter Sample Size: 814

Enter Observed Percentage: 55

Click Here to Calculate: Calculate

Margin of Error is: 3.42

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?

They would need to sample about 1,056 people to reduce the margin of error to 3.00%. The confidence interval would be 52% to 58%. You can be 95% confident that the confidence interval, 52% to 58%, captures the true population proportion. (Note: the exact wording of the last sentence is very important!)

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%?

**CNN would need to sample about 9,507 people to reduce the margin of error to 1%.
Poling firms “settle” for 3.5% because polling is expensive and takes time.**

3. What percent of Americans favor stricter gun control laws?
We don't know; we will never know. In order to know the answer to that question, we would need to ask every single person in the US and that would be impossible.
4. If a poll reports a confidence interval of 35.4% to 41.2%, what is the sample proportion and margin of error?
The sample proportion is 38.3% and the margin of error is 2.9%.
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:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	Person #	Favor stricter gun laws?		Freq	Rel Freq		Margin of Error		Confidence Interval														
2	1	O	Favor	452	55.53%		3.42%		52.11%	58.95%													
3	2	O	Oppose or other	362	44.47%																		
4	3	F	Total	814	100.00%																		
5	4	O																					
6	5	O																					
7	6	O	=COUNTIF(B:B,"F")																				
8	7	F																					
9	8	O																					
10	9	F																					
11	10	F																					
12	11	O																					
13	12	F																					
14	13	F																					
15	14	O																					
16	15	F																					
17	16	O																					
18	17	F																					
19	18	O																					
20	19	F																					
21	20	F																					
22	21	O																					
23	22	F																					
24	23	O																					
813	812	F																					
814	813	F																					
815	814	O																					

You can be 95% certain that this confidence interval captures the true population proportion.
 Or, if you do this simulation 20 times, you expect 19 (95%) of the sample populations to capture 55%, the assumed population proportion.

=E2-G2 =E2+G2

=COUNTIF(B:B,"F")
 =COUNTIF(B:B,"O")
 =IF(RAND()<=0.55,"F","O")

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 $2 \times \text{mean}$ and $2 \times \text{median}$ (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 known 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 paste 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