

PLCY 2455  
Problem Set #1  
Summer 2017  
Due MONDAY 6/19/2017, by 9am

Last Name:      Sarcone

First Name:      Krystal

Group members with whom you worked:

Remington Pontes (#'s 2 & 3)

Alex Barba (# 4)

Tynesia Fields (# 4)

\_\_\_\_\_

## QUESTION 1 – MEMO ON THE EVOLUTION OF PUBLIC DEBT SINCE THE GLOBAL FINANCIAL CRISIS

Your task is to write a one-page memo characterizing how public indebtedness (measured as a country's public debt as a share of its GDP) changed in different countries **between 2007 and 2015**, a period during which many advanced economies initially resorted to fiscal stimulus, and then some to austerity, as they attempted to cope with the aftermath of the global financial crisis. To do this, you will analyze data from the World Economic Outlook (found in an Excel workbook called WEO\_Tables.xls). We suggest that you conduct extensive data analysis by computing the descriptive statistics discussed in class and then decide which results are the most informative, relevant, and interesting to present in your memo.

A good memo will:

1. Make clear how the distribution of public debt changed between 2007 and 2015. You should comment not only on the change in average debt but also how other aspects of the distribution changed (e.g., 90th percentile, IQR, etc...).
2. Indicate any relevant patterns in changes in the distribution of public debt over time. For example, you might examine the differences across geographic regions, between advanced and non-advanced economies, or between economies that suffered a financial crisis and those that did not.
3. Use tables and/or graphs that convey the most salient findings.
4. Be written in a language that a policy maker, who is intelligent but not well versed in statistics, can understand.

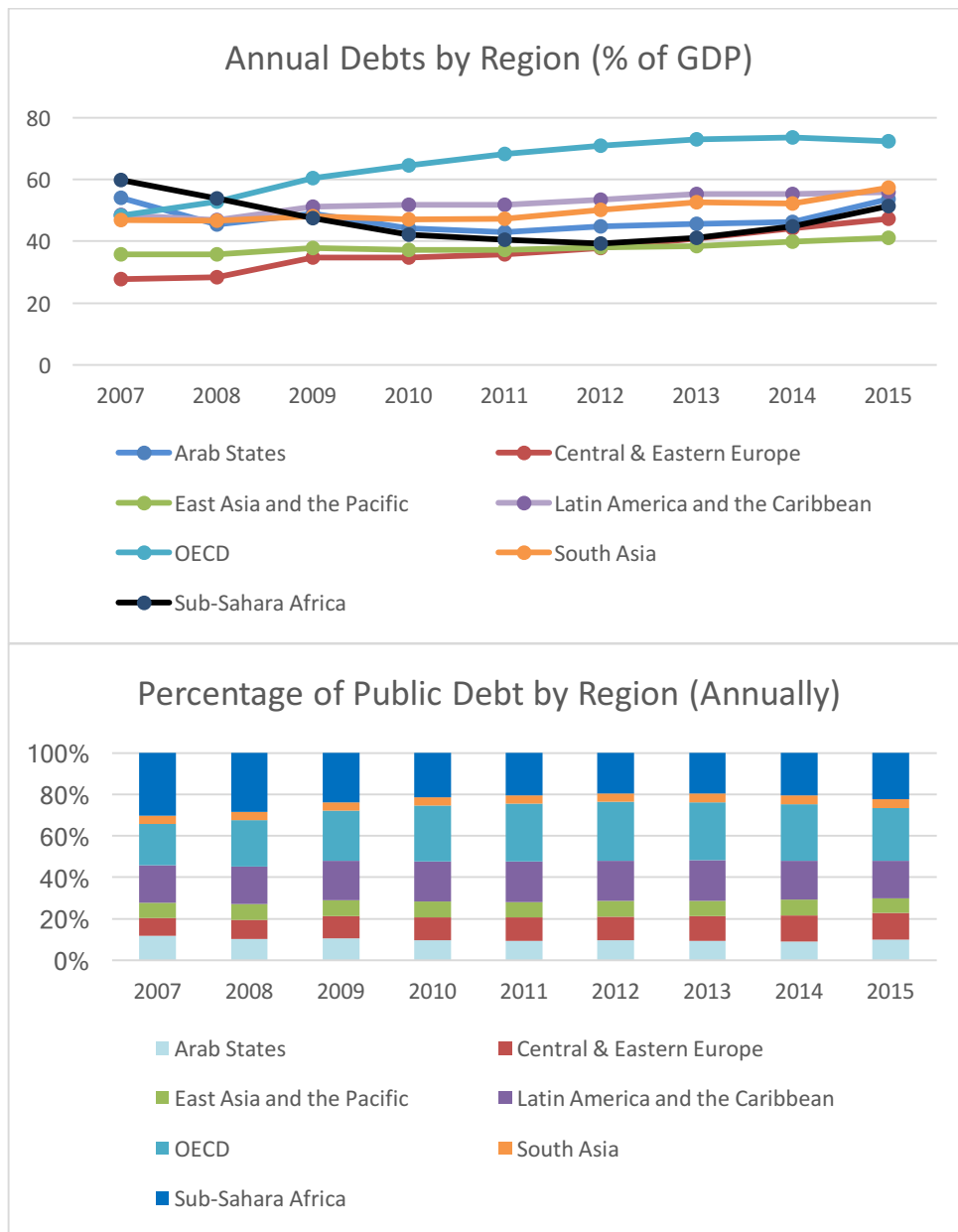
Other things to note:

- Memo format: One page (including any graphs and tables), single-spaced, with twelve-point font. Please be sure that any graphs and/or tables are large enough to read.
- The Excel worksheet "Debt" shows public debt as a percentage of GDP for all countries, annually, from 2000 to 2015.
- OECD membership can be used as an indicator for an advanced economy. The information on OECD membership is provided in column "Region" in the Excel worksheet "Debt."
- We have posted a guide to memo writing that may be of help to you. Notice that the memo we are asking you to write is not a decision or action memo (so some of the guidance in these documents may not be completely suitable), but the documents as a whole may prove useful, particularly to those of you with limited experience writing memos.

Note that this is the key question of this assignment, and one in which we expect you will spend a good portion of the time you devote to the assignment.

**Of the financial crisis that occurred, n=24, 79% of them occurred in OECD countries with only 4 occurring in Central and Eastern Europe and one occurring in Sub-Saharan Africa. On top of facing more of the financial crisis burden OECD countries also did the poorest with regards to the sum of change in debt between 2007-2015. OECD countries had a total sum of change in debt of 820 whereas Sub-Saharan Africa improved their debt situation during that time period with a sum of total change in debt of -359. Globally the average public debt rose during that period from 47.38 to 55.08**

**However, each region experienced the this time period differently both in annual changes in debt but also in their regional percent of total public debt**



**QUESTION 2 – COLLEGES AND INTERGENERATIONAL MOBILITY**

Data recently released from the [Equality of Opportunity Project](#) estimates the joint distribution of parents’ and kids’ incomes for all colleges in the United States. This question asks you to explore these data for a college of your choice. You may download these data in “College Mobility.xls”.

2(a) Choose a college to analyze, and report both the name and ID number (“super\_opeid”). NOTE: You may work with your group on this question (as others), but you must choose your own college that is different from others in your group.

**super\_opeid:** 3401  
**Name:** Brown University

2(b) Lay out a probability table that you could use to calculate various marginal, joint, and conditional probabilities in this setting. Note that you do not need to fill in the table.

	par_q1	par_q2	par_q3	par_q4	par_q5	
kq1						
kq2						
kq3						
kq4						
kq5						

2(c) Calculate the following probabilities for your chosen college:

2.C(1) P(Parents in Bottom 20%)

**par\_q1, provided in data = 2.9%**

2.C(2) P(Kids in Top 20% | Parents in Bottom 20%)

**kq5\_cond\_parq1, provided in data = 53%**

2.C(3) P(Kids in Top 20% AND Parents in Bottom 20%)

**Calculated = 1.5%**

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$P(0.53) = \frac{P(K5 \text{ and } P1)}{0.029}$$

$$P(K5|P1) = \frac{P(K5 \text{ and } P1)}{P(P1)}$$

$$P(K5 \text{ AND } P1) = 1.5\%$$

2.C(4) P(Kids in Top 20% AND Parents in Top 20%)

**Calculated = 43.87%**

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$P(0.596) = \frac{P(K5 \text{ and } P5)}{0.736}$$

$$P(K5|P5) = \frac{P(K5 \text{ and } P5)}{P(P5)}$$

$$P(K5 \text{ AND } P5) = 43.87\%$$

2.C.(5) P(Kids in Top 20%) = **58.9%**

#2.c.5. <sup>We're</sup> Looking for all Kids in Top 20%

We Need to find all joints for Kids in the top 20%

in 2.c.4. we found joint for

Kids top 20 AND Parents in top 20

using  $P(A|B) = \frac{P(A \cap B)}{P(B)}$  b/c we were given

$P(A|B)$  +  $P(B)$  data:

	Pg1	Pg2	Pg3	Pg4	Pg5	
Kg5	0.01539	0.02365	0.03871	0.07256	0.43912	= .589
top 20	0.029	0.043	0.071	0.121	0.736	1

$$P(K5|P1) = \frac{K5 \cap P1}{P1}$$

$$0.5307 = \frac{x}{0.029}$$

$$P(K5|P2) = \frac{K5 \cap P2}{P2}$$

$$0.5499 = \frac{x}{0.043} = 0.0236457$$

$$P(K5|P3) = \frac{K5 \cap P3}{P3}$$

$$0.5452 = \frac{x}{0.071} = 0.0387092$$

$$P(K5|P4) = \frac{K5 \cap P4}{P4}$$

$$0.5997 = \frac{x}{0.121} = 0.07256$$

$$P(K5|P5) = \frac{K5 \cap P5}{P5}$$

$$0.596 = \frac{x}{0.736} = .454$$

The paper analyzing these data focused on upward mobility from the bottom quintile to the top quintile, but there are many other potential definitions of mobility.

2(d) Calculate  $P(\text{Kids in Top 40\%} \mid \text{Parents in Bottom 40\%})$  for your chosen college. Be sure to show the formula you used for this calculation, in terms of the statistics provided in the dataset. = 72.5%

2.d. Solve for  $K4 \cap P4$  +  $K4 \cap P5$

using  

$$P(A|B) = \frac{A \cap B}{B}$$

$$K4 \mid P4 = \frac{K4 \cap P4}{P4}$$

$$K5 \mid P5 = \frac{K5 \cap P5}{P5}$$

$$0.149275 = \frac{K4 \cap P4}{0.121}$$

$$0.14127 = \frac{K5 \cap P5}{0.736}$$

$$K4 \cap P4 = 0.01806$$

$$K5 \cap P5 = 0.43901$$

$$K4 \cap P3$$

$$K4 \cap P2$$

$$K4 \cap P1$$

$$K4 \mid P3 = \frac{K4 \cap P3}{P3}$$

$$K4 \mid P2 = \frac{K4 \cap P2}{P2}$$

$$K4 \mid P1 = \frac{K4 \cap P3}{P1}$$

$$0.14982 = \frac{K4 \cap P3}{0.071}$$

$$0.17749 = \frac{K4 \cap P2}{0.043}$$

$$0.19129 = \frac{K4 \cap P3}{0.029}$$

$$K4 \cap P3 = 0.0106$$

$$K4 \cap P2 = 0.00763$$

$$K4 \cap P3 = 0.0054741$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Solve for  $P(A \cap B)$



$$P(K4|P1) = \frac{P(K4+P1)}{P(P1)} = 0.1913 = \frac{0.0055477}{0.029} = K4+P1$$

$$P(K5|P1) = \frac{P(K5+P1)}{P(P1)} = 0.5307 = \frac{0.0153903}{0.029} = K5+P1$$

$$P(K4|P2) = \frac{P(K4+P2)}{P(P2)} = 0.1775 = \frac{0.0076325}{0.043} = K4+P2$$

$$P(K5|P2) = \frac{P(K5+P2)}{P(P2)} = 0.5499 = \frac{0.0236457}{0.043} = K5+P2$$

$$P(\text{Kids Top 40} | \text{Parents Top 40}) = \frac{P(K4+P1)}{P1} + \frac{P(K5+P1)}{P1} + \frac{P(K4+P2)}{P2} + \frac{P(K5+P2)}{P2}$$

K4 | P1

K5 | P1

K4 | P2

K5 | P2

$$= \frac{0.0055477}{0.029} + \frac{0.0153903}{0.029} + \frac{0.0076325}{0.043} + \frac{0.0236457}{0.043}$$

0.19

0.5307

0.1775

0.5499

	par_q1 Bottom 20	par_q2 Bottom 40	par_q3 Top 60	par_q4 Top 40	par_q5 Top 20	
kq1 Bottom 20	0.00343078	0.0042251	0.00813958	0.00962404	0.0660155	0.091435
kq2 Bottom 40 40	0.00197589	0.00362098	0.00490465	0.0115012	0.0662525	0.08825526
kq3 Bottom 60	0.00270536	0.00381438	0.00858842	0.00923235	0.0607893	0.08512983
kq4 Top 40	0.00558193	0.00759063	0.01062693	0.01805338	0.1040034	0.14585629
kq5 Top 20	0.01548685	0.02351505	0.03866823	0.07252935	0.4391241	0.58932362
	0.02918081	0.04276613	0.07092781	0.12094032	0.7361849	1

*par\_q# = Parents Quartile*  
*kq# = Kids Quartile*

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		par_q1	par_q2	par_q3	par_q4	par_q5			par_top10pc	par_top5pc	par_top1pc	par_toppt1pc	
2		Bottom	Bottom	Top	Top	Top			0.622028128	0.48012759	0.188650885	0.031548193	
3		20	40	60	40	20							
4	kq1								kq1_cond_parq1	kq2_cond_parq1	kq3_cond_parq1	kq4_cond_parq1	kq5_cond_parq1
5	Bottom	0.0034308	0.0042251	0.0081396	0.009624	0.0660155	0.091435		0.117569735	0.067711874	0.092710403	0.191287548	0.53072044
6	20												
7	kq2								kq1_cond_parq2	kq2_cond_parq2	kq3_cond_parq2	kq4_cond_parq2	kq5_cond_parq2
8	Bottom 40	0.0019759	0.003621	0.0049047	0.0115012	0.0662525	0.0882553		0.098795499	0.084669302	0.089191569	0.177491527	0.549852103
9	40												
10	kq3								kq1_cond_parq3	kq2_cond_parq3	kq3_cond_parq3	kq4_cond_parq3	kq5_cond_parq3
11	Bottom	0.0027054	0.0038144	0.0085884	0.0092324	0.0607893	0.0851298		0.114758675	0.069149938	0.121086731	0.149827393	0.545177263
12	60												
13	kq4								kq1_cond_parq4	kq2_cond_parq4	kq3_cond_parq4	kq4_cond_parq4	kq5_cond_parq4
14	Top	0.0055819	0.0075906	0.0106269	0.0180534	0.1040034	0.1458563		0.079576787	0.095098143	0.076338086	0.149275106	0.599711877
15	40												
16	kq5								kq1_cond_parq5	kq2_cond_parq5	kq3_cond_parq5	kq4_cond_parq5	kq5_cond_parq5
17	Top	0.0154869	0.023515	0.0386682	0.0725293	0.4391241	0.5893236		0.089672436	0.089994427	0.082573428	0.141273516	0.596486193
18	20												
19		0.0291808	0.0427661	0.0709278	0.1209403	0.7361849	1		par_q1	par_q2	par_q3	par_q4	par_q5
20									0.029180808	0.042766131	0.070927811	0.12094032	0.736184929
21		[=SUM(E13:F18)]	0.0521744										
22		[=SUM(E19:F19)]	0.0719469										
23		[=D21/D22]	0.7251796										
24													

**Also See Attached Excel: “College Mobility SARCONe”**



2(e) Repeat your calculation in (d) for all colleges,

**See Attached Excel: “College Mobility SARCONE”**  
**Column I has the probability per college for Kids in the Top 40% GIVEN Their parents are in the Bottom 40%)**

and calculate the correlation between your new measure of upward mobility and P(Kids in Top 20% | Parents in Bottom 20%) across colleges.

**The Correlation =0.87**

	C	D	E	F	G	H	I	J
1								
2	par_q1	par_q2	kq4_cond_parq1	kq5_cond_parq1	kq4_cond_parq2	kq5_cond_parq2	Top 40   Bottom 40	[=CORREL(I:I,F:F)]
3	0.443575115	0.326790166	0.19232666	0.045164714	0.207874309	0.084689854	0.260853338	0.874670466
4	0.052441352	0.101031977	0.267327011	0.273902819	0.23309061	0.247651504	0.501410575	
5	0.154555182	0.187158613	0.186568282	0.096129661	0.228475947	0.142476096	0.331035222	
6	0.093524223	0.122917146	0.166350345	0.174900016	0.136387221	0.244054939	0.3635074	
7	0.129224392	0.185503884	0.263695825	0.14582912	0.289182793	0.162432578	0.434333448	
8	0.087048019	0.12853861	0.208177236	0.374334421	0.193888463	0.412743474	0.596892824	
9	0.114288185	0.177607137	0.243769699	0.086981264	0.249162607	0.152085685	0.373645891	
10	0.060760294	0.115355947	0.314623361	0.157415789	0.33797786	0.224236633	0.531103901	
11	0.025902086	0.111165487	0.118694457	0.778812955	0.076348212	0.037713813	0.262112139	
12	0.159531326	0.218920736	0.292089748	0.197068896	0.326453078	0.136637399	0.474079159	
13	0.12763433	0.215705448	0.194817465	0.296616386	0.330014723	0.291179177	0.572956445	
14	0.072003043	0.129967505	0.169420902	0.204771333	0.305753401	0.231689889	0.47924385	
15	0.201153378	0.191781065	0.167844657	0.085034882	0.19321453	0.120987702	0.282809548	
16	0.243561168	0.264324457	0.245893058	0.173234095	0.263938898	0.180948632	0.432533907	
17	0.293899452	0.229185474	0.140578238	0.104723209	0.232336527	0.091329655	0.279636328	
18	0.134097428	0.216393038	0.206855341	0.033365204	0.200724051	0.057951104	0.251614433	
19	0.217516425	0.241491306	0.228839692	0.112214294	0.240031436	0.130175696	0.356391922	
20	0.039403551	0.101879336	0.037291662	0.852088913	0.065597025	0.781799484	0.859105791	
21	0.233010812	0.281636744	0.298995202	0.110702773	0.285142185	0.153592261	0.425587951	
22	0.445306555	0.288225748	0.099749932	0.032280397	0.126022333	0.035961285	0.14379983	
23	0.129735245	0.163736763	0.253381074	0.425518694	0.381646527	0.198604248	0.62386056	
24	0.022652305	0.069142471	0.205165924	0.30077224	0.324696225	0.329935735	0.617938617	
25	0.051434021	0.126789504	0.324022705	0.266473381	0.282635996	0.310532439	0.592397214	

2(f) In a short paragraph (100 words), comment on the implications of your result in (e).

**A correlation of 0.87 reveals a relatively strong association between the likelihood of a child being in the top 20% given their parents are in the top 20%. This not only makes logical sense but is replicated in the data when you look at an institutional level. At Brown University the probability of a child being in the top 20% is 58.9%. Given that their parents are in the top 20% that probability remains essentially the same at 59.6% however if their parents are from the bottom 20% that probability drops to 2%. Thus it looks as though not only is there a strong correlation, but there’s much less mobility for children coming from parents in the bottom 20%.**

**QUESTION 3 - HIV INFECTION**

What can one learn from an HIV test? We examine this issue in the context of Rhode Island, where the prevalence of HIV is about 0.24 percent. Suppose that the test for detecting HIV identifies correctly 99.9 percent of those actually infected and 99.6 percent of those actually not infected. (These are the approximate rates over the past decade.)

- (a) Use Bayes’ Rule to calculate the probability that a person who tests positive is infected (i.e.  $P(\text{HIV}|+)$ ). This number is usually referred to as the *positive predictive value of the test*.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \begin{matrix} A = H \\ B = + \end{matrix}$$

$$3a) \quad P(H|+) = \frac{P(H \cap +)}{P(+)} = \frac{(0.0023976)}{(0.006388)} = 0.3753287$$

- (b) Use Bayes’ Rule to calculate the probability that a person who tests negative is not infected (i.e.  $P(\text{NO HIV}|-)$ ). This number is usually referred to as the *negative predictive value of the test*.

$$3b) \quad P(H^c|-) = \frac{P(H^c \cap -)}{P(-)} = \frac{(0.99361)}{(0.993612000)} = 0.9999979871$$

	HIV +	HIV -	
Test +	0.0023976	0.00399	0.006388
Test -	0.00000240000	0.99361	0.99361200000
	0.0024	0.9976	1

- $P(+ | \text{HIV}+) = 0.999$  Sensitivity – True Positive**
- $P(- | \text{HIV}-) = 0.996$  Specificity – True Negative**
- $P(+ | \text{HIV}-) = 0.004$  False Positive**
- $P(- | \text{HIV}+) = 0.001$  False Negative**

**Also see attached excel “HIV SARCONE”**

(c) Read the following passage:

*"Two weeks ago, a 3-year-old child in Winston Salem, North Carolina, was struck by a car and rushed to a nearby hospital. Because the child's skull had been broken and there was a blood spill, the hospital performed an HIV test. As the traumatized mother was sitting at her child's bedside, a doctor came in and told her the child was HIV-positive. Both parents are negative. The doctor told the mother that she needed to launch an investigation into her entire family and circle of friends because this child had been sexually abused. There was no other way, the doctor said, that the child could be positive. A few days later, the mother demanded a second test. It came back negative. The hospital held a press conference where a remarkable admission was made. In her effort to clear the hospital of any wrongdoing, a hospital spokesperson announced that 'these HIV tests are not reliable; a lot of factors can skew the tests, like fever or pregnancy. Everybody knows that.'"*

Celia Farber, Impression Magazine, June 21, 1999. Reported by Christine Maggiore: Is the "AIDS test" Accurate? (<http://healtoronto.com/testcm.html>)

Write a short paragraph (100-200 words) to the hospital commenting on whether the claims made by the doctor and the hospital spokesperson were sound. The letter should be written in language that the head of the hospital (who is intelligent and educated, but not well-versed in statistics) can understand.

**I find the recent actions and response of Hospital X quite disappointing and above all else unprofessional. It appears that the physicians in your practice could benefit from a refresher in protocol and bedside manners. The first glaring area of concern is that a second HIV test was not automatically and immediately performed upon receiving the positive results of the first, this is protocol in many health institutions. This is because we must always consider false positives, especially in this case where false positives can truly have a negative emotional impact on the individual or family. And if the child were indeed positive, an HIV counselor with proper training should discuss with the family the possible implications of such.**

**Secondly the statement made by the hospital representative in the press conference is ill-informed and only makes a bad situation worse, not better. HIV tests have very high sensitivity and specificity, meaning they're extremely reliable in detecting HIV when the person is HIV +, or providing a negative result with the person is HIV negative. However, they are not perfect and though much less common, false positives and false negatives do occur. I hope the hospital takes this as an opportunity to improve their protocols and communication etiquette with patients and the public.**

(d) In the article above, the three-year-old child first has a positive test result and then a negative result. Suppose that a different child's first test result comes back negative but then a second test comes back positive. After the two tests, is the probability that the second child (negative then positive) has HIV higher, lower, or the same as the first child (positive then negative)? Justify your answer using either equations or calculations, and then give a brief intuitive argument (100 words) supporting your answer.

Handwritten calculations and notes:

$$P(-|H) = \frac{P(-|H)}{H} = \frac{P(0.0000024)}{0.0024} = 0.001$$

$$P(+|H^c) = \frac{P(+|H^c)}{H^c} = \frac{0.00399}{0.9976}$$

$$P(-|H^c) = \frac{P(-|H^c)}{H^c} = \frac{0.99361}{0.9976}$$

Child 1: T1, T2  
 Child 2: T1, T2

$P(+|H^c) \cdot P(-|H^c) = P(0.0040) \cdot P(.9960) = 0.0039836022$   
 $P(-|H) \cdot P(+|H) = (0.001) \cdot (.999) = 0.000999$   
 Not the SAME!

• Assuming HIV status does not change between Test 1 & Test 2 + child 2 is truly not HIV positive.

But if child 2 is HIV (+)  $P(-|H) \cdot P(+|H) = (0.001) \cdot (.999) = 0.000999$

d) Assuming that child 1 had not had a blood transfusion prior to the accident nor had been sexually assaulted + truly does not have HIV.

**The outcome of the individual tests are independent events. Unlike removing cards from a deck and not putting them back in the deck each turn and calculating probability of selecting a 4-card, the HIV test events are more like flipping a weighted coin. The odds are not 50:50, but the probability of a + or – result remain the same each coin toss given your HIV status doesn't change.**

**Thus, if we are assuming both children are truly HIV negative, the probability of a + test followed by a – test or a – test followed by a + test, are the same. However, if one of the children is truly HIV+ and the other HIV- the probabilities are NOT the same.**

**QUESTION 4 – HURRICANE PREPARATION**

You are tasked by FEMA to estimate expected hurricane damage in the upcoming year.

The data file “stormData.xlsx” (Source: <http://www.icatdamageestimator.com/viewdata>) contains information on the 242 Atlantic hurricanes that have made landfall on the United States from 1900 – present. Note that some storms appear twice in the data if they made landfall in more than one state (e.g. Hurricane Andrew (1992) made landfall in Florida, re-entered the Gulf of Mexico and made a second landfall in Louisiana).

Variables in the dataset include:

- **BASE DAMAGE (\$):** Estimated damage from the storm in the United States in the year the storm made landfall.
- **CURRENT DAMAGE (\$ 2016) :** Estimated damage from the storm if the storm struck today. Current damage adjusts base damage for inflation, as well as changes in coastal population and the changes in the value of coastal property.
- **DAMAGE RANK:** Ordinal ranking based on current damage
- **CATEGORY AT LANDFALL:** Category on the Saffir-Simpson hurricane scale when the storm made landfall in the United States (<http://www.nhc.noaa.gov/aboutsshws.php>). Note: “TS” denotes a hurricane that made landfall as a tropical storm.
- **WINDS AT LANDFALL:** Maximum sustained wind speed when the storm made landfall in the United States.

(a) Major hurricanes are classified as Category 3, 4 or 5 hurricanes and are considered especially dangerous. How much damage should we expect if a major hurricane makes landfall in the U.S.? **\$18,258,161,764.71**

See Excel “stormDATA SARCONE” Tab 4.a.

66	Storm 4 in 1947	1	17,100,000,000
67	Storm 4 in 1947	0	2,420,000,000
68	Storm 5 in 1936	1	80,000,000
69	Storm 6 in 1916	1	2,130,000,000
70	Storm 7 in 1944	1	11,590,000,000
71	Storm 7 in 1944	0	9,020,000,000
72	Storm 7 in 1948	1	4,550,000,000
73	Storm 8 in 1906	1	2,130,000,000
74	Storm 9 in 1945	1	16,910,000,000
75	Tampa Bay	1	4,410,000,000
76	Valasco	1	2,000,000,000
77	Wilma	1	24,560,000,000
78	<b>Unique Storms</b>	<b>68</b>	<b>1,241,555,000,000</b>
79	<b>Expected Value</b>	<b>\$ 18,258,161,764.71</b>	<b>Total Damage</b>
80			

(b) If we use the hurricanes from 1900-present as a guide, how many hurricanes should we expect in the coming year?

**Average # of Major Hurricanes Per Year = 1.372**

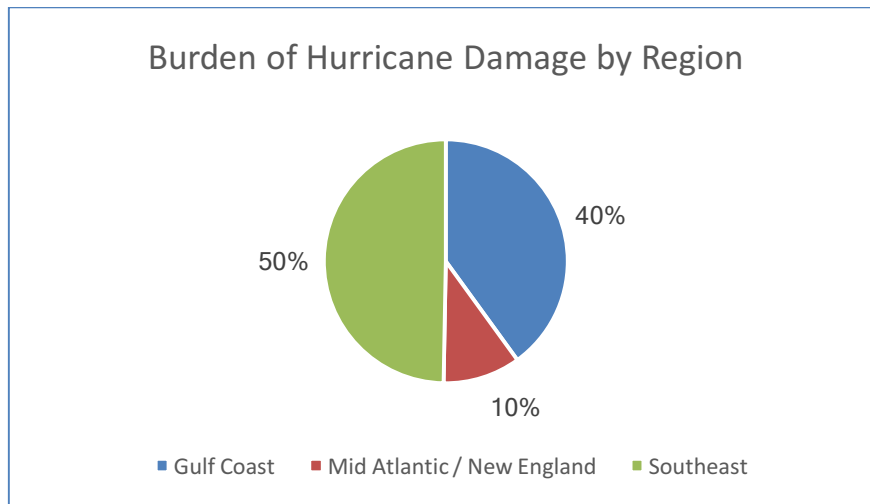
How much damage should we expect nationwide? (Hint: first create a table that sums the number of hurricanes and hurricane damage by year)

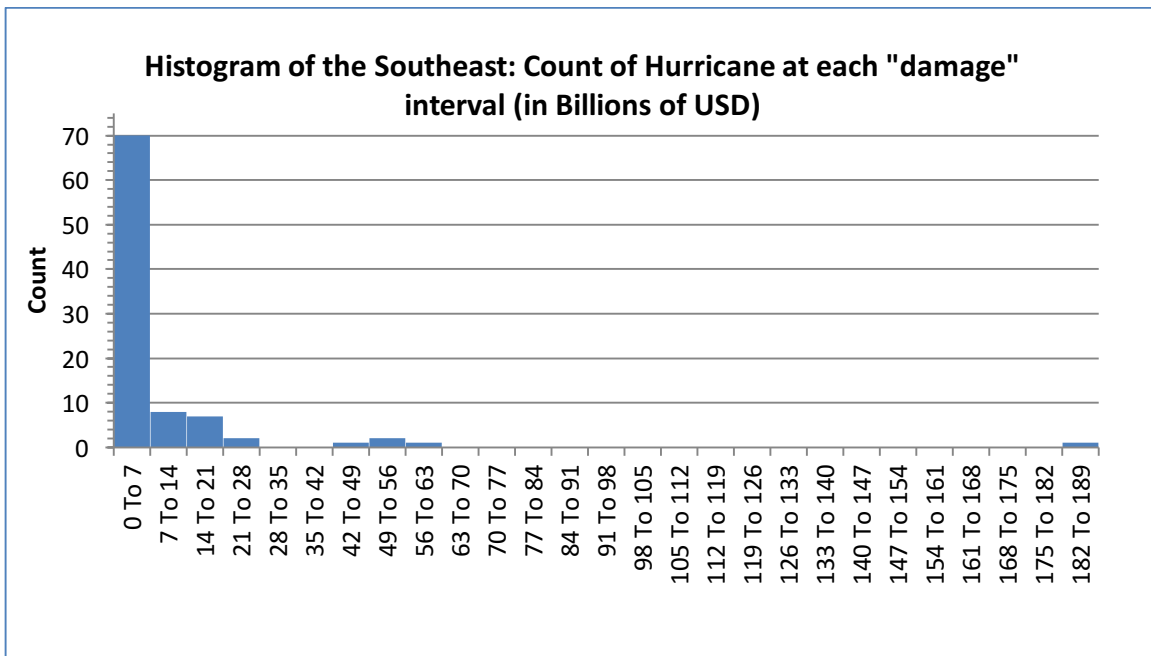
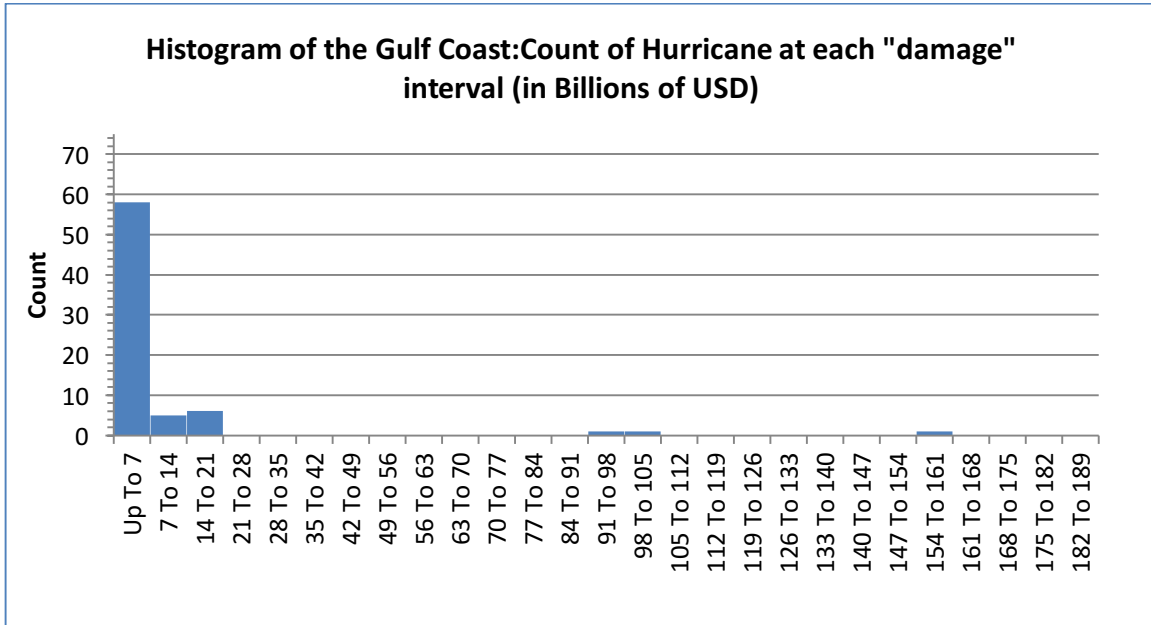
**Average Damage Per Year = \$18,116,178,571.43**

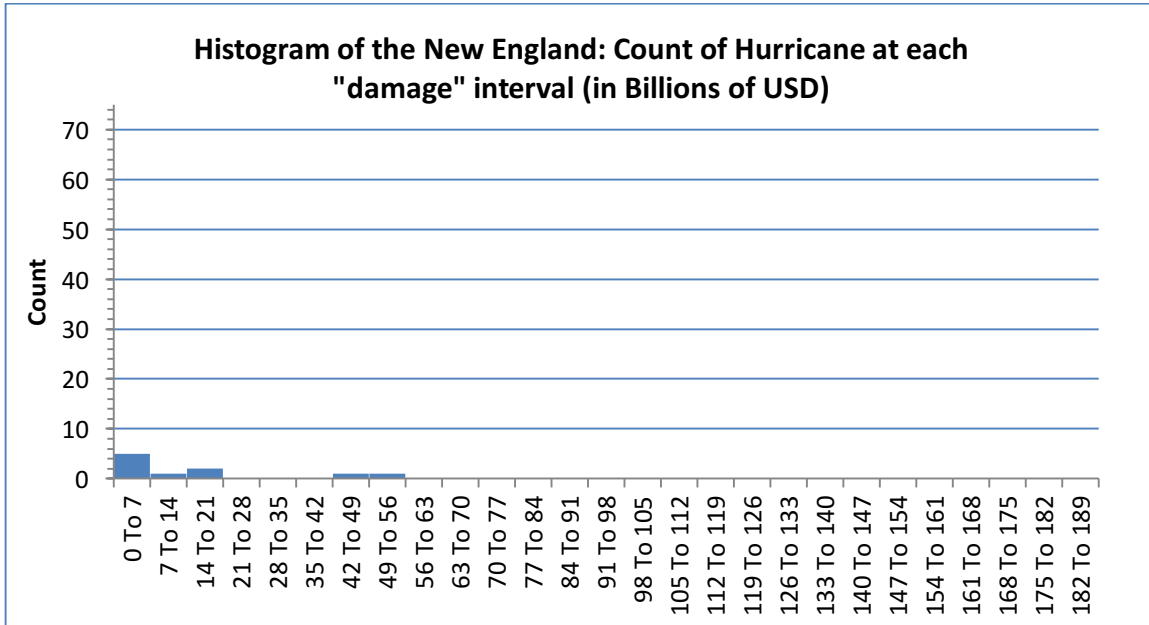
See Excel “stormDATA SARCONE” Tab 4.b.

1	6	30,000,000,000	2005	\$ 136,163,000,000.00
1		83,000,000		
1		11,920,000,000		
1		24,560,000,000		
0	0	-	2006	-
1	1	54,000,000	2007	\$ 54,000,000.00
1		1,110,000,000		
1	3	4,530,000,000	2008	\$ 25,960,000,000.00
1		20,320,000,000		
0	0	-	2009	-
0	0	-	2010	-
1	1	7,330,000,000	2011	\$ 7,330,000,000.00
1		2,410,000,000		
1	2	51,210,000,000	2012	\$ 53,620,000,000.00
155	1.371681416	\$ 1,521,759,000,000.00	Total Sum of Hurricane Damage	\$ 1,521,759,000,000.00
Total Unique Storms	Average Major Hurricanes Per Year	\$ 8,695,765,714.29	Total Number of Years	\$ 18,116,178,571.43
155		Average Damage Per Hurricane	113	Average Damage Per Year
	(X) = # of Storms/Year	Years	P(x)	(x) P(x)
	0	29	0.2566371681	0.00000
	1	43	0.3805309735	0.38053
	2	22	0.1946902655	0.38938
	3	14	0.1238938053	0.37168
	4	2	0.0176991150	0.07080
	5	0	0.0000000000	0.00000
	6	3	0.0265486726	0.15929
		113	1.00	1.37168

(c) Using the past as a guide, plot the distribution of hurricane damage that might hit in the upcoming seasons. You should do this separately for each of the three regions.







See Excel “stormDATA SARCONE” 4.c. Tabs

In a short paragraph (less than 100 words), how does the risk of hurricane damage vary by region? Provide at least one comparison based on a statistic other than the mean.

**I chose to keep my axis the same so visually we could clearly see that by region the number of hurricanes that cause damage on the lower end of the billion-dollar range but also in the higher less common, but costly end are much more common in the Gulf Coast and even more so in the Southeast. All three histograms are skewed right with extremely costly hurricanes in the hundreds of billions of dollars serving as outliers. I also graphed that proportionally the Southeast take on 50% of the financial burden of hurricane damage, followed by the Gulf Coast at 40% and the Mid-Atlantic only enduring 10% of all hurricane damage costs.**



## QUESTION 5 – GOAL-SCORING STRATEGIES IN SOCCER/FOOTBALL

Read the following passage:

Timothy Treep, a former Royal Air Force Wing Commander, tracked play-by-play data for matches and served as a quantitative consultant for Football League teams as early as the 1950s. He took it upon himself to attend every Swindon Town F.C. match — sometimes with a miner’s helmet on his head to better illuminate his notes — and meticulously scribble down play-by-play diagrams of how everything went down. More than 60 years before player-tracking cameras became all the rage in pro sports, Treep was mapping out primitive spatial data the old-fashioned way, by hand.

Poring over all the scraps of data he’d collected, Treep eventually came to a realization: Most goals in soccer come off of plays that were preceded by three passes or fewer. And in Treep’s mind, this basic truth of the game should dictate how teams play. The key to winning more matches seemed to be as simple as cutting down on your passing and possession time, and getting the ball downfield as quickly as possible instead. The long ball was Treep’s secret weapon.

“Not more than three passes,” Treep admonished during a 1993 interview with the BBC. “If a team tries to play football and keeps it down to not more than three passes, it will have a much higher chance of winning matches. Passing for the sake of passing can be disastrous.”

This was it: Maybe the first case in history of an actionable sports strategy derived from next-level data collection, such as it was. And Treep got more than a few important folks to listen to his ideas, too. It took him a few decades of preaching, but Treep’s recommended playing style was adopted to instant success by Wimbledon F.C. in the 1980s, and then reached the highest echelons of English soccer — channeled as it was through the combination of England manager Graham Taylor and Football Association coaching director Charles Hughes, each of whom believed in hoofing the ball up the pitch and chasing it down (and now seemed to have the data to back up their intuition). The long ball was suddenly England’s official footballing policy.

- (a) In a short paragraph (100 words), explain to your statistics professor why Treep was using the wrong probability to assess the relative efficiency of different scoring strategies, and what probability would be more appropriate. (Note: Please ignore the issue of correlation vs. causation.)

**Like the OJ example in class, in which case the wrong marginal variable was on the right of the | “given” symbol on which the first probability is conditional; In this soccer example, instead of considering the percentage of goals generated by sequences of different passes, Treep should have calculated the probability that given a set number of passes, what was the probability of scoring a goal.**

- (b) Write a short paragraph (100-200 words) to a soccer coach explaining your logic in (a). The letter should be written in language that the soccer coach (who is intelligent and educated, but not well-versed in statistics) can understand.

**As a coach of soccer you know that the game is naturally and fundamentally involves lots of turnovers and short possessions, and thus it would make sense that lots of goals are scored off of short possessions. But that doesn’t mean that making sure your passing game is short, and not more than three passes increases your likelihood of scoring a goal. Soccer is complex and so too are statistics. But what’s wrong with the findings of Treep is that short sequences are already more common in soccer, so it makes sense more goals come from short passes. Now with better technology and cameras and stat packages we can see that other factors play a role in scoring probability, such as possession.**

PLCY 2455

Problem Set #2

Summer 2016

Due MONDAY 6/26/2017, by 9am

Last Name: Sarcone \_\_\_\_\_

First Name: Krystal \_\_\_\_\_

Group members with whom you worked:

Cassie Taylor  
\_\_\_\_\_

Elisabeth Perez  
\_\_\_\_\_

Marisa Vang  
\_\_\_\_\_

\_\_\_\_\_

## QUESTION 1 – DEMAND FOR FLU SHOTS

As the leaves change color, and the air turns crisp, it can mean only one thing – influenza vaccination season! Suppose it is October 6, 2014, and you are working on a team at the U.S. Centers for Disease Control and Prevention (“CDC”) on flu-related issues. Concerned about a vaccine shortage – which happens from time to time<sup>1</sup> – you and your team make plans based on your best guess that 53.5 percent of American adults intend to get vaccinated this year.

Then, on November 16, 2016, Rasmussen, a polling company, publishes the results of a national poll of 1,000 American adults:<sup>2</sup>

---

## Most Intend to Get A Flu Shot This Year

in **Lifestyle**

 Facebook  Twitter  Email this  Share This

Wednesday, November 16, 2016


It's flu season again, and that's left most Americans running out to get their flu shots.

A new Rasmussen Reports telephone and online survey finds that 56% of American Adults intend to get a flu shot this year. That's down just two points from 58% in 2014 which was [the highest finding since we first started asking this question in 2006](#). Thirty-eight percent (38%) don't plan on getting a flu shot, up slightly from 35% two years ago. (To see survey question wording, [click here](#).)

### RELATED ARTICLES

- [Americans Report Paying More For Health Care Compared to Five Years Ago](#)
- [Americans Insist: No Vaccine, No School](#)
- [What America Thinks: Public Health Agencies Rate Our Trust](#) ▶

### Sign up for free daily updates

 Your e-mail here

OK

You focus specifically on the percentage of American adults that say they intend to get a flu shot this year. Assume that 560 of the 1,000 adults in the sample gave this response.

(a) The Basics. Define the following terms in this specific context:

1. Population: **Adults residing in the USA**
2. Sample: **Non-institutionalized reachable adults surveyed(n=1,000)**
3. Estimate: **The sample estimate is 56%(0.56) 560/1,000**

---

<sup>1</sup> See, e.g., <http://health.usnews.com/health-news/news/articles/2013/01/15/spot-shortages-of-fluvaccine-tamiflu-reported-fda-head-says>. For more, see <http://www.cdc.gov/flu/about/qa/vaxdistribution.htm>.

<sup>2</sup> [http://www.rasmussenreports.com/public\\_content/lifestyle/general\\_lifestyle/november\\_2016/most\\_intend\\_to\\_get\\_a\\_flu\\_shot\\_this\\_year](http://www.rasmussenreports.com/public_content/lifestyle/general_lifestyle/november_2016/most_intend_to_get_a_flu_shot_this_year)

(b) Using the Rasmussen survey results, test the hypothesis that 53.5 percent of American adults intend to get vaccinated this year. Follow the steps described in class:

1. State the **null hypothesis** ( $H_0$ ).

$$H_0 = q_0 = 0.535$$

2. Set a **significance level** ( $\alpha$ ).

$$\alpha = 0.05$$

3. Calculate the **estimate** from the sample.

$$\text{Given} = 56\% \text{ } 0.56 \text{ (560/1,000)}$$

4. Define the **sampling distribution** and use it to calculate the **p-value**.<sup>3</sup>

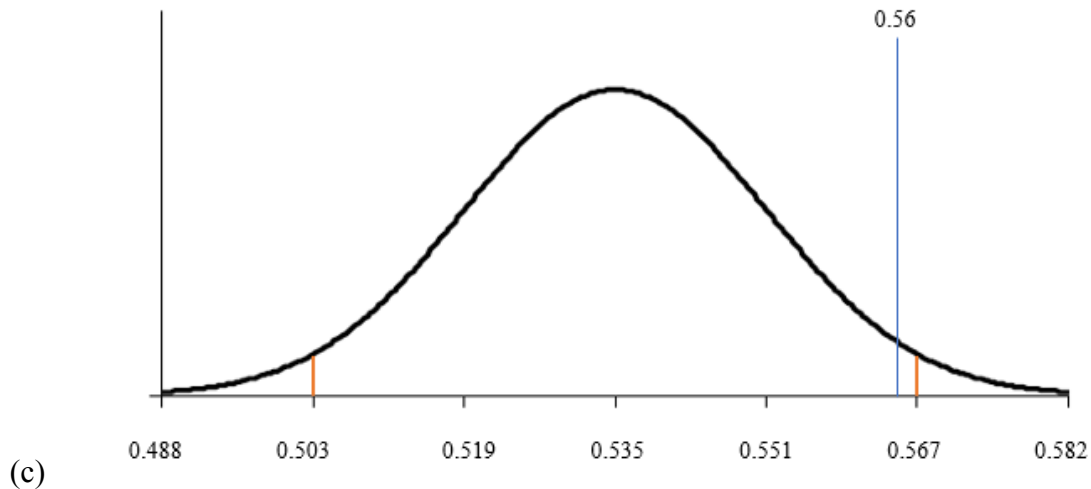
i **Shape: Normal**

ii **Mean: 0.535**

iii **Standard error:**  $\sqrt{\frac{q_0(1-q_0)}{n}} = \sqrt{\frac{0.535(1-0.535)}{1000}} = 0.0158$

**Z-Score:**  $0.025 / 0.0158 = 1.58$

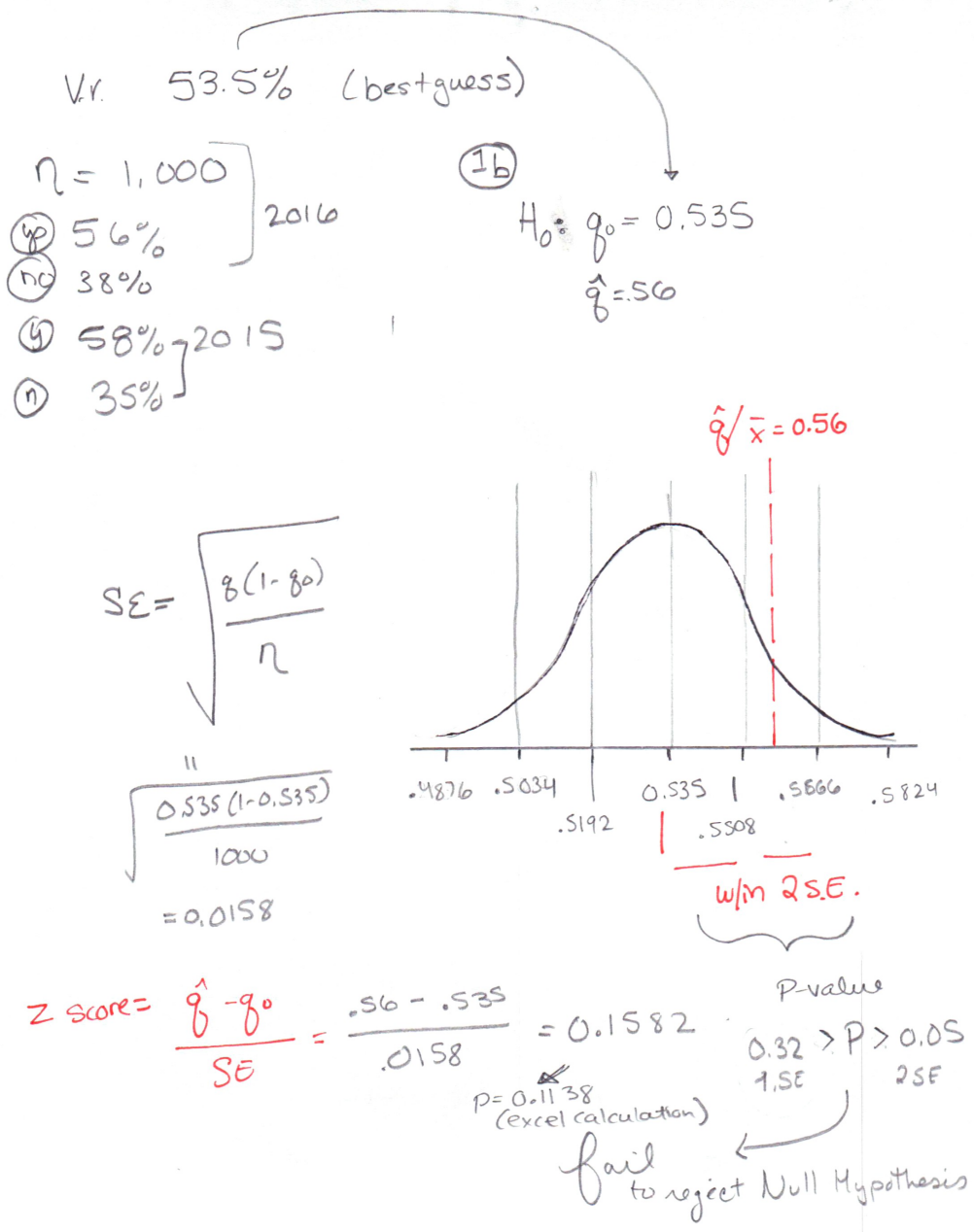
**P-Value:**  $=2*(1-\text{NORM.S. DIST}(\text{ABS}(1.58), \text{TRUE})) = 0.112$



1. **Reject or fail to reject**  $H_0$  based on whether  $p\text{-value} < \alpha$ .

**We fail to reject our null hypothesis  
because our p-value is greater than 0.05**

<sup>3</sup> Note that, assuming that the Central Limit Theorem applies (which you can check), you can calculate the p value using Excel. Suppose that you have a z-score  $Z$ . To calculate the two-tailed p-value, use the following Excel command:  $=2*(1-\text{NORM.S. DIST}(\text{ABS}(Z), \text{TRUE}))$ . For example, you could calculate the p-value of  $Z = -1.2$  using  $=2*(1-\text{NORM.S. DIST}(\text{ABS}(-1.2), \text{TRUE}))$ .



(d) What is the interpretation of the sampling distribution and the p-value in this context? Explain the result to your project leader, a statistical neophyte who is extremely concerned that the Rasmussen survey result (56 percent) is greater than the vaccination rate (53.5 percent) that you anticipated and used as the basis for your flu-season plans.

**The distribution of sampling, if we draw enough samples will be normally distributed and centered around the population proportion. Assuming our best guess of anticipating 53.5% of people getting a flu shot, and a standard**

**error of 0.0158 we know our estimate falls within two standard deviations/errors from the mean (between 0.5508 and 0.5666) and with a p-value of 0.1138 we believe that if our hypothesis were true is approximately an 11.38% chance of seeing a vaccination rate of 56% or as extreme.**

- (e) Calculate a 95% Confidence Interval for your estimate in (b), and explain this to your Statistics professor.

$$95\% \text{ CI} = \hat{q} \pm 1.96 * SE$$

$$95\% \text{ CI} = \widehat{56} \pm 1.96 * 0.0158 = (0.529, 0.591)$$

**If we were to repeat polling over and over again, 95% of all confidence intervals will contain the true average number of people who get flu shots next year.**

- (f) Suppose that you conduct a new poll to test the same null hypothesis, that 53.5 percent of American adults intend to get vaccinated this year. Suppose that exactly 56 percent of adults in your new sample intend to get vaccinated this year. What is the minimum number of individuals you would need in your new sample to be able to reject the null hypothesis at a 5 percent significance level?

We already know the z-score of 0.012755 at  $\alpha = 0.05$  thus we plug that into the equation and solve for  $n$ .

$$SE = \sqrt{\frac{q_0(1-q_0)}{n}} \qquad 0.012755 = \sqrt{\frac{0.535(1-0.535)}{n}}$$

$$0.012755^2 = \frac{0.248775}{n} \qquad n = \frac{0.248775}{0.012775^2} \qquad n = 1,524.35$$

**A minimum of 1,525 individuals are needed to reject  $H_0$  at  $\alpha = 0.05$**

## QUESTION 2 – WELFARE PROGRAMS IN INDIA

The Mahatma Gandhi National Rural Employment Guarantee Act (NREGA, <http://nrega.nic.in/netnrega/home.aspx>), one of the world's largest welfare programs enacted by the Government of India in 2005, guarantees 100 days of paid employment to adults in rural India willing to do unskilled manual work at the minimum statutory wage. The program was rolled out over a three-year period across districts in rural India. You are interested to know whether the average monthly per capita expenditure in the Indian state of Rajasthan has changed in the aftermath of the program's implementation. In 2006, you randomly sample 500 adults in Dungarpur district in Rajasthan, which was one of the 200 districts where NREGA was first implemented. You also have access to a random sample

of 500 adults from the same district in 2005, before the program was implemented there.<sup>4</sup> You can download these data (NREGA.xlsx) from the course website.

Use these data to conduct a hypothesis test at the 5 percent significance level to assess whether the monthly per capita expenditure has changed in the aftermath of the NREGA implementation. Use the 5 steps outlined in class:

1. State the **null hypothesis** ( $H_0$ ).

**The null hypothesis is that there is no difference in mean monthly per capita expenditure before and after the NREGA program intervention.**

$$H_0: \mu_1 - \mu_2 = 0 \quad \text{or} \quad H_0: \mu_1 = \mu_2$$

2. Set a **significance level** ( $\alpha$ ).  $\alpha = 0.05$

3. Calculate the **estimate** from the sample.

$$\bar{x}_1 - \bar{x}_2 = 453.2660 - 435.4411 = 17.8249$$

4. Define the **sampling distribution** and use it to calculate the **p-value**.<sup>5</sup> 5.

**Reject or fail to reject**  $H_0$  based on whether  $p\text{-value} < \alpha$ .

a **Shape = normal**

b **Mean = 0**

$$c \quad SE = \text{formula} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{148.87^2}{500_1} + \frac{175.16^2}{500_2}} = 10.28$$

**Standard Deviation data were obtained from the excel files**

$$S_1 = 148.87$$

$$S_2 = 175.16$$

**P-value from excel = 0.083 and thus**

**we fail to reject the  $H_0$  at  $\alpha = 0.05$**

---

<sup>4</sup> The numbers quoted in this example were created for educational purposes and do not reflect official estimates by the Government of India or other analytical work.

<sup>5</sup> Note that, assuming that the Central Limit Theorem applies (which you can check), you can calculate the p-value using Excel. Suppose that you have a z-score  $Z$ . To calculate the two-tailed p-value, use the following Excel command: `=2*(1-NORM.S.DIST(ABS(Z), TRUE))`. For example, you could calculate the p-value of  $Z = -1.2$  using `=2*(1-NORM.S.DIST(ABS(-1.2), TRUE))`.

**QUESTION 3 – THE BOUNDARY BETWEEN REJECT AND FAIL TO REJECT**

Suppose you draw a random sample from the population and conduct a hypothesis test based on the following four quantities:

$$H_0: q = 0.40$$

$$qq = 0.36 \quad n = 1000$$

$$\alpha = 0.05$$



(a) Do you reject or fail to reject the hypothesis that  $q = 0.40$ ?

$$H_0: q_0 = 0.40$$

$$\hat{q} = 0.36$$

$$SE = \sqrt{\frac{q(1-q)}{n}} = \sqrt{\frac{0.4(1-0.4)}{1000}} = 0.01549$$

$$Z\text{-Score: } (\hat{q} - q_0) / SE = -2.581988962$$

$$P\text{-Value: } = 2 * (1 - \text{NORM.S. DIST}(\text{ABS}(-2.581988962), \text{TRUE})) = 0.009823$$

$$CI = 0.36 \pm 1.96 * SE = (0.3296, 0.3904)$$

With a p-value of  $< 0.05$  we reject the null hypothesis that  $q = 0.40$

While holding the other three quantities constant, change each quantity in turn so that you are just on the boundary between rejecting and failing to reject the null hypothesis. For each case, report the range of values for which you would reject the hypothesis and the range of values for which you would fail to reject the hypothesis.<sup>6</sup> In addition, in each case describe intuitively why increasing/decreasing the quantity in question changes the pvalue.

- (b)  $H_0: q = 0.40$        $qq =$        $n = 1000$        $\alpha = 0.05$
- (c)  $H_0: q = 0.40$        $qq = 0.36$        $n =$        $\alpha = 0.05$
- (d)  $H_0: q = 0.40$        $qq = 0.36$        $n = 1000$        $\alpha =$

<sup>6</sup> For example, state, “I would reject the hypothesis for any value of      that is [greater than/less than/between]     , and I would fail to reject the hypothesis for any value of      that is [less than/greater than/between]     .”



b.)  $H_0: q = 0.40$        $0.37 < \hat{q} < 0.43$        $n = 1000$        $\alpha = 0.05$

Reject the null if  $\hat{q} \geq 0.43$  and if  $\hat{q} \leq 0.37$

Fail to reject if  $0.37 < \hat{q} < 0.43$

And increasing  $\hat{q}$  will increase the z-score and decrease the p-value.

(3b)

$$\left| \frac{\hat{q} - q_0}{SE} \right| = 1.96$$

$$\frac{\hat{q} - 0.40}{SE} = 1.96$$

$$\hat{q} - 0.40 = (1.96)(SE)$$

$$\hat{q} - 0.40 = (1.96)(0.015)$$

$$\hat{q} - 0.40 = 0.0294$$

$$\hat{q} = 0.4294$$

$H_0 = q = 0.40$   
 $\hat{q} = 0.43$  find  $\hat{q}$   
 $n = 1000$   
 $\alpha = 0.05$   
 $SE = \sqrt{\frac{0.4(1-0.4)}{n}} = 0.015$

c.)  $H_0: q = 0.40$        $\hat{q} = 0.36$        $n = 576.23$        $\alpha = 0.05$

Reject the null if  $n \geq 576.23$

Fail to reject if  $n < 576.23$

(c)

$$H_0 = q = 0.40$$

$$\hat{q} = 0.36$$

$$\alpha = 0.05$$

$$n = 576.24 \text{ find } n^*$$

$$|z\text{-score}| = 1.96$$

$$\left| \frac{0.36 - 0.4}{SE} \right| = 1.96$$

$$|-0.04| = (1.96)(SE)$$

$$\left( \frac{0.04}{1.96} \right)^2 = \left( \sqrt{\frac{0.24}{n}} \right)^2$$

$$0.00041649 = 0.24 \frac{(n)}{n}$$

$$n = 576.24$$

Decreasing the n will increase the standard error and decrease the z-score. This results in an increase in the p-value.

d.)  $H_0: q = 0.40$       $\hat{q} = 0.36$       $n = 1000$       $\alpha = 0.0098$

Reject the null if  $\alpha \geq 0.009$

Fail to reject if  $\alpha < 0.0098$

$H_0: q = 0.40$       $n = 1000$       $\alpha = 0.009$      find alpha  
 $\hat{q} = 0.36$

$\frac{\hat{q} - q_0}{SE} = Z\text{-score}$

$\sqrt{\frac{0.4(0.6)}{1000}} = 0.015$

$\left| \frac{0.36 - 0.40}{0.015} \right| = \frac{-0.04}{0.015}$

$Z\text{-score} = 2.6\bar{6}$       $\alpha = 0.009$

Increasing the significance level ( $\alpha$ ) will expand the ranges of where the truths can lie. This will increase the p-value.

**QUESTION 4: POLLUTION AND HEALTH IN LOS ANGELES**

Particulate matter (PM) pollution is generated by industrial activity and motor vehicles. Of different types of pollution, PM 2.5 poses the greatest health risks – the particles are sufficiently fine to penetrate deep into the lungs.

You are advising the current Mayor of Los Angeles, Eric Garcetti, on public health issues related to pollution. As part of your analysis, you run two regressions examining daily data. You use the number of cardiorespiratory and non-cardiorespiratory emergency room visits in the Los Angeles metropolitan area as the dependent variable and the average level of PM2.5 pollution on that day as the independent variable. Consider the following table of results.<sup>7</sup>

<sup>7</sup> These results are a simplified version of actual research that examines the relationship between pollution and health outcomes in [http://faculty.haas.berkeley.edu/rwalker/research/SchlenkerWalker\\_Airports\\_2012.pdf](http://faculty.haas.berkeley.edu/rwalker/research/SchlenkerWalker_Airports_2012.pdf)

**Table 1: Pollution and Emergency Room Visits, by Cause**

	Emergency Room Visits	
	Cardiorespiratory	Non-Cardiorespiratory
PM 2.5 pollution (micrograms / cubic meter, 24 hour average)	32.6 (15.2)	21.6 (65.7)
Constant	164.0 (49.6)	769.3 (276.5)
R-squared	0.115	0.002
N	5475	5475

Note: standard errors are in parentheses.

(a) State a reasonable null hypothesis (in non-technical language) for the slope of the regression where cardiorespiratory emergency room visits is the dependent variable. Calculate the approximate p-value.

**$H_0: \beta_1 = 0$  essentially if there is no change in cardiorespiratory ED visits**

**Estimate:  $\widehat{\beta}_1 = 32.6$**

**Standard Error: = 15.2**

**T-statistic:  $= \frac{\widehat{\beta}_1 - \beta_0}{SE} = 2.145$**

**P-value = 0.01601**

(b) Construct a 95% confidence interval for the slope of the regression where cardiorespiratory emergency room visits is the dependent variable.

**95 % CI =  $32.6 \pm 1.96 * SE$**

**95 % CI =  $32.6 \pm 1.96 (15.2)$**

**95 % CI =  $32.6 \pm 29.79$**

**95% CI = ( 2.808 , 62.392 )**

c.) Although Mayor Garcetti is a graduate of Columbia School of International and Public Affairs (MIA, 1993), it's been a while since he has taken statistics. Explain the main takeaways from the table using non-technical language in a short (100 words or less) paragraph.

**We can't necessarily say if there is causation with this data of 5,475 however there appears to be no real strong linear correlation with cardio respiratory ED visits and pollution of 2.5 PMs (with r-squared values of 0.115 and 0.002 respectively for cardiovascular vs. non-cardiovascular emergency room visits.**

# PLCY 2455

## ProblemSet #3 Summer 2017

### Due MONDAY 7/3/2017 by 9am

Last Name: Sarcone

First Name: Krystal

Group members with whom you worked:

---

---

---

#### **QUESTION 1 – OMITTED VARIABLE BIAS IN THE NEWS**

Use the concept of omitted variable bias you learned in class to read and critically assess study findings, as presented in NYT Op-ed “Making Tyrants Do Time” (<http://www.nytimes.com/2011/09/16/opinion/making-tyrants-do-time.html>) and answer the following questions. You do not need to read the full study by Prof. Sikkink.

- a) What is the key question the author is trying to answer? What is the outcome (dependent variable) of interest in her study? What independent variables did she include in her study?

**The question the author is trying to answer is whether I.C.C. trials encourage criminals to hide, leverage more authoritarian power, or undermine nation state stability or whether such trials have positive impacts on the societies at hand, i.e. fewer executions, forced disappearances, political imprisonment or less torture. The independent variables of interest are prosecutions/trials/extraditions.**

- b) What is the most important source of potential omitted variable bias that you can think of? In a short paragraph, describe that omitted variable bias. Be sure to discuss the direction of that bias.

**Omitted Variable Bias essentially explains the direction of bias that results from relevant confounding variables (a covariate having a relation with both exposure/independent/x and outcome/dependent/y variables) not being controlled for in a regression. In this particular research/article one could argue a variety of potential omitted variables. One particular omitted variable I would consider would be the severity of crimes committed by government officials. I wonder if controlling for the severity of crimes, i.e. a handful of political imprisonment vs. mass genocide or ethnic cleansing we'd see different findings. I worry about this omitted variable because it impacts both the possible likelihood of being tried (x/independent variable) for human rights violations and with potential societal outcomes of stability, and fewer human right violations (dependent variable of interests).**

**O = Severity of Crimes  
X = Prosecution/Trials  
Y = Resultant Societal Stability & Peace**

**sign(O, X) / sign (severity of crimes, prosecution) = + (positive)  
sign(O, Y) / sign (severity of crimes, resultant societal stability) = - (negative)  
sign of bias = - (negative)**

**I would argue that severity of crimes is positively correlated with likelihood or the extent of prosecution for human rights violations, resulting in a positive bias. And I would also argue that a society experiencing more sever crimes, will have a harder time recovering from such hardships and thus a negative correlation would result from this interaction, making the sign of the bias negative.**

**Subsequently, with a negative bias, for this particular OVB, we could potentially see that the findings, not reported in any statistical sense in this NY Times article, would be underestimated, when controlling for severity of crimes committed. This is because without a controlling for negative OVB, the coefficient are often underestimated.**

- c) You're hired as the Special Envoy for North Korean Human Rights. In a short paragraph with non-technical language, explain how to interpret this study's findings.

**In this article, the population of interest are nation states "in-transition," meaning they are either 1) an authoritarian government transitioning to a democratic one, or 2) countries experiencing war transitioning to peace and**

**North Korea is neither. Though we view North Korea as a rogue state that has political leverage via nuclear weapons and though they face international condemnation for some of the worst human rights violations in the world, North Korea does not fit the criteria by which nations were selected for analysis in Sikkink’s research and takes place 13 years after the time frame for her analysis.**

**Notwithstanding, Sikkink’s research “found that prosecutions tend not to exacerbate human rights violations, undermine democracy or lead to violence,” meaning that should I.C.C. pursue prosecution of DPRK (Democratic People’s Republic of Korea) leaders, such as Kim Jong-il, for human rights violations, which has been urged by South Korea earlier this year, if this research were indicative of the truth, we could rest assured that such trials would not make the situation worse on the ground for North Korean citizens. Additionally, the article claims the regional prosecutions may also impact neighboring countries to decrease level of repression or human rights violations. This sort of “spillover effect” possibly observed in South America with Argentinian and Chilean trials paving the way for decreased regional contemporary military coups is also something to consider in the case of North Korea. However, I would argue that North Korea may be an outlier, in that, similar to Syria being unlikely deterred by trials of fellow regional leaders such as Mr. Mubarak in Cairo.**

**QUESTION 2 – TRAFFIC FATALITIES AND SEATBELTS**

Traffic crashes are the leading cause of death for Americans between the ages of 5 and 32. Through various spending policies, the federal government has encouraged states to institute mandatory seat belt laws to improve safety. You are asked to investigate how effective these laws are in increasing seat belt use and reducing fatalities.

Panel of data from 50 U.S. states, plus the District of Columbia, for the years 1983-1997 are available.<sup>1</sup> The key variables are as follows:

<i>fatalityrate</i>	=	Number of fatalities per million traffic miles
<i>sb_usage</i>	=	Seatbelt usage rate
<i>speed65</i>	=	Dummy variable for 65 MPH state speed limit
<i>speed70</i>	=	Dummy variable for 70 MPH state speed limit
<i>ba08</i>	=	Dummy variable for state blood alcohol limit $\leq$ 0.08%
<i>drinkingage21</i>	=	Dummy variable for state drinking age of 21
<i>lninc</i>	=	Average per capita income (logarithmic scale)
<i>age</i>	=	Mean age of drivers in state

---

<sup>1</sup> Note: These data were provided by Professor Liran Einav of Stanford University and were used in his paper with Alma Cohen “The Effects of Mandatory Seat Belt Laws on Driving Behavior and Traffic Fatalities” in The Review of Economics and Statistics, 2003, Vol. 85, No. 4, pp 828-843.

You regress the fatality rate on a series of different variables:

Regressor	(1)	(2)	(3)
<i>sb_useage</i>	0.0041*** (0.0012)	-0.0057*** (0.0012)	-0.0037*** (0.0011)
<i>speed65</i>	0.00014 (0.00041)	-0.00040 (0.00030)	-0.00078* (0.00042)
<i>speed70</i>	0.0024*** (0.0005)	0.0012*** (0.0003)	0.0008** (0.0003)
<i>ba08</i>	-0.0019*** (0.0004)	-0.0014*** (0.0004)	-0.0008** (0.0004)
<i>drinking21</i>	0.00008 (0.00099)	0.00074 (0.00051)	-0.00110** (0.00050)
<i>lninc</i>	-0.0181*** (0.0011)	-0.0135*** (0.0014)	0.0062 (0.0039)
<i>age</i>	-0.00001 (0.00016)	0.00098** (0.00038)	0.00130*** (0.00040)
<i>State Effects</i>	No	Yes	Yes
<i>Year Effects</i>	No	No	Yes
<i>R-Squared</i>	0.544	0.874	0.897

Notes: Dependent variable for all regressions is the fatality rate per million traffic miles. All regressions include 765 state-year observations, corresponding to 50 states plus the District of Columbia between 1983 and 1997. Stars indicate statistical significance at the 10% (\*), 5% (\*\*), and 1% (\*\*\*) levels respectively.

- a) Does Column 1 suggest that increased seat belt use reduces fatalities? What other omitted variables might explain this relationship?

**The coefficient ( $\beta_0 = 0.0041$ ) of seatbelt usage indicates that yes, when controlling for top speeds, blood alcohol content, drinking age, average age, and average income, for seatbelt usage there is an increase of 0.0041 fatalities per million traffic miles.**

**Any omitted variables that might explain this relationship other than the**

**above listed ones controlled for include ones such as quality of roads, safety standards for cars, maybe gender, state laws or any difference between states and years.**

- b) Consider the following potential sources of bias. For each one, first indicate whether it would bias the coefficient of interest upwards or downwards relative to the causal effect of seatbelts on fatalities. Then, indicate in which columns the control variables and fixed effects are sufficient to remove the bias, and in which columns the bias would remain:

- i. States with higher speed limits tend to have a higher fatality rate and lower seatbelt usage.

**O = Higher Speed Limits**

**X = Seatbelt Usage**

**Y = Fatality Rates**

**sign(O, X) / sign (higher speed limits, seatbelt usage) = - (negative)**

**sign(O, Y) / sign (higher speed limits, fatality rates) = + (positive)**

**sign of bias = - (negative)**

**The omitted variable of interest is controlled for in columns 1, 2 and 3.**

- ii. Over time, as seat belt usage has increased in all states, airbags have also become standard equipment in cars and reduced fatality rates.

**O = Better Safety Standards, i.e. Airbags**

**X = Seatbelt Usage**

**Y = Fatality Rates**

**sign(O, X) / sign (Airbags, seatbelt usage) = + (positive)**

**sign(O, Y) / sign (Airbags, fatality rates) = - (negative)**

**sign of bias = - (negative)**

**The omitted variable of interest is controlled for in column 3, which accounts for differences over time. Columns 1 and 2 do not control for changes over time (year fixed effect).**

- iii. States that were more urbanized in 1983 tend to have higher fatality rates (because of more cars on the road) and higher seatbelt usage rates (due to higher education of its citizens).

**O = Urbanized**

**X = Seatbelt Usage**

**Y = Fatality Rates**

**sign(O, X) / sign (Urbanized, seatbelt usage) = + (positive)**



**sign(O, Y) / sign (Urbanized, fatality rates) = + (positive)**  
**sign of bias = + (positive)**

**Controlled for in columns 2 & 3, which accounts for differences between states. Column 1 does not control for differences between states or changes over time.**

- iv. Certain states have increased seatbelt usage over time especially quickly. and have also especially increased the quality of road safety barriers, as part of a broad program of improvements in highway safety.

**O = Highway safety**  
**X = Seatbelt Usage**  
**Y = Fatality Rates**

**sign(O, X) / sign (Highway safety, seatbelt usage) = + (positive)**  
**sign(O, Y) / sign (Highway safety, fatality rates) = - (negative)**  
**sign of bias = - (negative)**

**Controlled for in column 3, which accounts for differences between states and years.**

- c) Suppose that you are evaluating a law that is projected to increase seat belt usage from 52% to 90% on average. Suppose that the average number of vehicle miles traveled in a state is 13 billion. Based on the results in your column of choice, how many lives would be saved?

**Column of Choice: 3**  
**Average vehicle miles traveled: 13 billion**  
**Current Seat Belt Usage: 52%**  
**Goal Seat Belt Usage: 90%**  
**Estimated Lives Saved:**

**Estimated Lives Saved = (0.0037)(90-52) = 0.1406 / per million**

**13 billion / 1 million = 13,000**

**(0.1406)(13,000) = 1,827.8**

**Estimated Lives Saved = 1,828**