Econ 300 Spring 2020
Problem set 1
Total Points: 50

Instructions: Please write/type your answers and hand in a hard copy at the beginning of the class. The due date of the problem set is 1/27/2020.

1. Suppose you are studying the research question whether bigger class size increases students' academic performance. To make it simple, let's divide classes into two groups: big and small. "big" is defined as number of students greater than 40, and "small" is below or equal to 40.

a) What is the naïve comparison? Why can it be misleading?

b) What is the ideal comparison? Is it ever feasible to make the ideal comparison?

2. Soccer ranking and # of world-cup winner (Men)

Team	Current world ranking	# of world-cup winners
Brazil	9	5
Italy	10	4
Germany	4	4
Argentina	1	2
Uruguay	8	2
France	7	1
England	13	1
Spain	8	1
Netherlands	26	0

Calculate the covariance between the two variables. Are they positively correlated or negatively correlated? Does the result meet your expectation?

3. Explain what central limit theorem is and why it is important in statistical inference?

4. Suppose a new standardized test is given to 100 randomly selected third-grade students in New Jersey. The sample average score $\bar{X}$ on the test is 58 points, and the sample standard deviation, s, is 8 points.

a. The authors plan to administer the test to all third-grade students in New Jersey. Construct a 95% confidence interval for the mean score of all New Jersey third graders.

b. Suppose the same test is given to 200 randomly selected third graders from Iowa, producing a sample average of 62 points and sample standard deviation of 11 points. Construct a 90% confidence interval for the mean score of all Iowa third graders.

5 [STATA] The table given below shows the relationship between cigarettes consumed and lung cancer.

Observation #	Country	Cigarettes consumed per capita in 1930 (X)	Lung cancer deaths per million people in 1950 (Y)
1	Switzerland	530	250
2	Finland	1115	350
3	Great Britain	1145	465
4	Canada	510	150
5	Denmark	380	165

Calculate the below statistics using STATA.

- The sample means of X and Y
- The standard deviations of X and Y
- The correlation coefficient, r, between X and Y

STATA HINTS: First load STATA and type "edit," which brings up something that looks like a spreadsheet. Enter the smoking and cancer values in the first two columns. Double-click the column headers to enter variable names (e.g. "smoke", "death"). Close the editor window when you are done. The following commands will be useful:

list: lists the data (to be sure you typed it in correctly)

summarize: computes sample means and standard deviations (the option ",detail" gives additional statistics, including the sample variance)

correlate: produces correlation coefficient (with the option ",covariance" this command produces covariances)

6 (OPTIONAL: BONUS QUESTION worth 5 points). The attention span of a two-year old is normally distributed with a mean and standard deviation both about 8 minutes. Suppose we randomly survey 60 two-year olds.
Which do you think is higher? Explain.

- The probability that an individual baby's attention span is less than 10 minutes
- The probability that the average attention span for the 60 children is less than 10 minutes

7 (OPTIONAL: BONUS QUESTION worth 5 points).

	Rain (X=0)	No Rain (X=1)	Total
Long Commute (Y=0)	0.15	0.07	0.22
Short Commute (Y=1)	0.15	0.63	0.78
Total	0.30	0.70	1.00

Using the random variables X and Y from the table given above, consider two new random variables W=3+6X and V=20-7Y. Compute:

a) E(W) and E(V)

b) Var(W) and Var(V)