**Econ 300 Spring 2020**

**Problem set 1 Suggested Solutions**

**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 1/27/2019.**

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?

**Naïve comparison: compare the academic performance, e.g. test scores, in big classes with the academic performance of the students in small classes**

**It’s misleading because other factors affect class size as well as academic performances. For example, private schools would have smaller classes than public schools. At the same time, the students in private schools are more likely to have parents or teachers with college degree. It’s hard to say that the difference of academic performance between small classes and big classes comes from the class size only.**

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

**Compare the academic performance of the students when they are in big class with the performance of the same students when they are in small class at the same time. No, because we can’t travel back in time.**

**Or, compare the same group of students when they’re in small class with when they’re in big class. It’s feasible if we randomly assign students**.

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?

**Cov(world ranking, # of winners)=-5.51.**

**It’s negatively correlated and it meets my expectation because the teams with a higher rankings will win the world cup more.**

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

**The Central Limit Theorem states that with sufficiently large sample size, the sampling distribution is approximately Normal.**

**It’s important because it allows us to use the normal distribution for statistical inference in situations where the underlying distribution is not normal.**

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

1. 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.
2. 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.

**95% confidence interval = 58 – 1.96(8/10), 58 + 1.96 (8/10)= (56.4, 59.6)**

1. **90% CI = (62 – 1.645( 11/(200)^0.5), 62 + 1.645( 11/(200)^0.5)**

**= (60.7, 63.3)**

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)

**Solution**

**summ X Y**

**cor X Y**

* **The sample means of X and Y (736, 276)**
* **The standard deviations of X and Y (364.4, 132.4)**
* **The correlation coefficient, r, between X and Y (0.926)**

6. 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

**The probability that the average attention span for the 60 children is less than 10 minutes is higher than the other probability.**

**The average attention span for 60 children distributes on a sampling distribution with a mean as 8 mins and a standard error as about 1.00(). So the probability of more than 10 mins is very small (about 5%) because 10 mins is more than 2 std. error away from the mean. In other words, the probability that the average attention span for 60 children with less than 10 mins is at least 95%.**

**However, an individual baby’s attention span has a distribution with a mean and a std. deviation of about 8 mins. 10 mins lays within 1 std. deviation around the mean. So its probability is less than the probability that the average attention span for the 60 children is less than 10 mins.**

7.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 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:

1. E(W) and E(V)

**E(X) = (0 \* 0.3) + (1 \*0.7) = 0.7**

**E(Y)= (0.22 \* 0) + (0.78 \* 1) = 0.78**

* **E(W)= 3 + 6 E(X)= 7.2**
* **E(V)= 20 - 7 E(Y)= 14.54**

1. Var(W) and Var(V)

**Var (X)= E(X^2) – (E(X))^2= ( 0 \* 0.3 + 1\* 0.7)- (0.7)^2= 0.7 – 0.49= 0.21**

**Var (Y)= 0.78 – 0.78^2= 0.1716**

**Var(W)= 6^2 Var (X) = 36 Var (X) = 7.56**

**Var (V)= 49 Var (Y)= 8.41**