Problem Set (Week 1)

Econ 103

Lecture 1 and 2

1. For each variable indicate whether it is nominal, ordinal, or numeric.

   (a) Grade of meat: prime, choice, good.

   > **Solution:** ordinal

   (b) Type of house: split-level, ranch, colonial, other.

   > **Solution:** nominal

   (c) Income

   > **Solution:** numeric

2. A drive-time radio show frequently holds call-in polls during the evening rush hour. Explain in no more than two sentences why such polls are likely to be biased.

   > **Solution:** People who are listening to the radio during rush hour are disproportionately likely to be commuters driving home from work. People who are employed and drive to work are not representative of the population at large.

3. Which of these studies are based on experimental data? Which are based on observational data?

   (a) A biologist examines fish in a river to determine the proportion that show signs of disease due to pollutants poured into the river upstream.

   > **Solution:** Observational

(b) In a pilot phase of a fund-raising campaign, a university randomly contacts half of a group of alumni by phone and the other half by a personal letter to determine which method results in higher contributions.

Solution: Experimental

(c) To analyze possible problems from the by-products of gas combustion, people with with respiratory problems are matched by age and sex to people without respiratory problems and then asked whether or not they cook on a gas stove.

Solution: Observational

(d) An industrial pump manufacturer monitors warranty claims and surveys customers to assess the failure rate of its pumps.

Solution: Observational

4. An emergency room institutes a new screening procedure to identify people suffering from life-threatening heart problems so that treatment can be initiated quickly. The procedure is credited with saving lives because in the first year after its initiation, there is a lower death rate due to heart failure compared to the previous year among patients seen in the emergency room. Do you agree? Explain.

Solution: No. There could be many other reasons why death rates decreased, including improved medical technology in other areas. It could also be that the patients who came into the ER in the second year happened to be less sick, on average. In other words, there are many possible confounders.

5. Suppose that x_i is measured in centimeters and y_i is measured in feet. What are the units of the following quantities?

(a) Interquartile Range of x

Solution: centimeters

(b) Covariance between x and y

Solution: centimeters $\times$ feet

(c) Correlation between x and y

Solution: unitless

(d) Skewness of x

Solution: unitless

(e) Variance of y

Solution: feet^2

Lecture 3

6. The *mean deviation* is a measure of dispersion that we did not cover in class. It is defined as follows:

$$MD = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

(a) Explain why this formula averages the absolute value of deviations from the mean rather than the deviations themselves.

Solution: As we showed in class, the average deviation from the sample mean is zero regardless of the dataset. Taking the absolute value is similar to squaring the deviations: it makes sure that the positive ones don't cancel out the negative ones.

(b) Which would you expect to be more sensitive to outliers: the mean deviation or the variance? Explain.

Solution: The variance is calculated from squared deviations. When x is far from zero, x^2 is much larger than $|x|$ so large deviations "count more" when calculating the variance. Thus, the variance will be more sensitive to outliers.

7. Consider a dataset $x_1, \ldots, x_n$. Suppose I multiply each observation by a constant d and then add another constant c, so that x_i is replaced by $c + dx_i$.

(a) How does this change the sample mean? Prove your answer.

Solution:

$$\frac{1}{n}\sum_{i=1}^{n}(c+dx_i) \quad = \quad \frac{1}{n}\sum_{i=1}^{n}c+d\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)=c+d\bar{x}$$

(b) How does this change the sample variance? Prove your answer.

Solution:

$$\frac{1}{n-1}\sum_{i=1}^{n}[(c+dx_i)-(c+d\bar{x})]^2=\frac{1}{n-1}\sum_{i=1}^{n}[d(x_i-\bar{x})]^2=d^2s_x^2$$

(c) How does this change the sample standard deviation? Prove your answer.

Solution: The new standard deviation is $|d|s_x$, the positive square root of the variance.

(d) How does this change the sample z-scores? Prove your answer.

Solution: They are unchanged as long as d is positive, but the sign will flip if d is negative:

$$\frac{(c+dx_i)-(c+d\bar{x})}{ds_x}=\frac{d(x_i-\bar{x})}{ds_x}=\frac{x_i-\bar{x}}{s_x}$$