Introductory Econometrics
Lecture 1: Introduction
Suggested Solution

by Hieu Nguyen

Fall 2024

1.

A researcher is analyzing data on the financial wealth of 100 professors at a small liberal arts college. The values of their wealth range from \$400 to \$400,000, with a mean of \$40,000, and a median of \$25,000. However, when entering these data into a statistics software package, the researcher mistakenly enters \$4,000,000 for the person with \$400,000 wealth. How much does this error affect the mean and median?

Solution: We employ a formula from the lecture #1 slides for the sample mean:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

A difference between the sample mean for the mistaken sample ($\tilde{x}_n$), $n = 100$, and the sample mean for the correct sample (x_n) can be computed as:

$$\tilde{x}_n - x_n = \frac{1}{n}\tilde{x}_{100} - \frac{1}{n}x_{100} = \frac{1}{100}(\$4,000,000 - \$400,000) = \$36,000,$$

because only the last observation changes. Still, the remaining parts of both formulas ($i = 1, 2, \ldots, 99$) are identical and thus cancel each other out. We can then express:

$$\tilde{x}_n = \$36,000 + x_n = \$76,000 > x_n = \$40,000.$$

We see that the sample mean is largely affected (it almost doubles). On the other hand, the median is not affected as only the last observation changes but the 'middle' observation remains the same.

2.

Which has a higher expected value and which has a higher standard deviation: a standard six-sided die (D6) or a four-sided die (D4) with the numbers 1 through

4 printed on the sides? Explain your reasoning without doing any calculations, then verify doing the math.

Solution: We employ a formula from the lecture #1 slides for the expected value of a discrete variable:

$$E[X] = \sum_{i=1} x_i P(X = x_i),$$

and because we assume 'fair' dice, we can directly compute $E[D6] = 3.5$ and $E[D4] = 2.5$.

Next, we employ formulas for the variance and the standard deviation:

$$\text{Var}[X] = E\left[(X - E[X])^2\right] = E[X^2] - (E[X])^2,$$

$$\sigma_X = \sqrt{\text{Var}[X]},$$

and obtain:

$$E[X^2_{D6}] = \frac{91}{6}; \quad \text{Var}[X_{D6}] = \frac{35}{12}; \quad \sigma_{X_{D6}} \approx 1.7,$$

$$E[X^2_{D4}] = \frac{30}{4}; \quad \text{Var}[X_{D4}] = \frac{5}{4}; \quad \sigma_{X_{D4}} \approx 1.12.$$

Thus the D6 has a larger standard deviation as could have been intuitively reasoned because the possibilities are more spread out on the D6.

3.

The heights of U.S. females between age 25 and 34 are approximately normally distributed with a mean of 66 inches and a standard deviation of 2.5 inches. What fraction of U.S. female population in this age interval is taller than 70 inches (the height of average adult U.S. male of this age)?

Solution: We employ formulas from the lecture #1 slides for the probability computational rule and standardization of a random variable:

$$P(X > x) = 1 - P(X \leq x),$$

$$X \sim N(\mu, \sigma^2) \rightarrow Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

We then compute:

$$P(X \geq 70) = 1 - P(X \leq 70) = 1 - P\left(\frac{X - 66}{2.5} \leq \frac{70 - 66}{2.5}\right) = 1 - P(Z \leq 1.6) = 1 - 0.9452 = 5.5\%.$$

The last equality (=) is based on a search in statistical tables for the standard normal distribution (e.g., Wooldridge, 2016, pg. 743-744).

4.

A woman claims that she had been pregnant for 310 days before giving birth. Completed pregnancies are normally distributed with a mean of 266 days and a standard deviation of 16 days. Use statistical tables to determine the probability that a completed pregnancy lasts i) at least 270 days, ii) at least 310 days.

Solution: We employ the same formulas as in the previous exercise (it does not matter whether we use the probability computational rule first or we standardize first, here we standardize first) and compute:

$$P(X \geq 270) = P(Z \geq 0.25) = 1 - P(Z \leq 0.25) = 1 - 0.5987 = 40\%,$$

$$P(X \geq 310) = P(Z \geq 2.75) = 1 - P(Z \leq 2.75) = 1 - 0.9970 = 0.3\%.$$