Key to exam 1, eco 311, spring 2014

Q1: the standard error of $\bar{y}$ is

$$se = \frac{5}{\sqrt{n}} = \frac{0.9993776}{\sqrt{200}} = 0.07$$

Q2: rigorously speaking, the unbiasedness of $\bar{y}$ only requires equal means of all observations: $E(y_i) = \mu$ for all $i$. A random (i.i.d) sample of course satisfies this requirement.

Q3: the histogram is not symmetric because the sample skewness is -.1126, less than zero. We don’t know if this is significantly different from zero, though. The histogram is skewed to the left. That means there are some extremely small values in this sample, and those small outliers cause the sample mean to be less than the sample median.

Q4:

$$t = \frac{\bar{y} - c}{se} = \frac{2.984904 - 2}{0.07} = 14.07 > 1.96$$

so we reject the null hypothesis at 0.05 level. Here we use the critical value of normal distribution because the sample size is large enough so that the central limit theorem implies that $\bar{y}$ follows a normal distribution. If you use the critical value of the t distribution, then the degree of freedom is so large that the t critical value is identical to the normal critical value.

Q5: the proper critical value should be 1.645 other than 1.28 because

$$P(-1.645 < Z < 1.645) = P(Z < 1.645) - P(Z < -1.645) = 0.95 - 0.05 = 0.9$$

$$\Rightarrow 90\% \text{ confidence interval is } [\bar{y} - 1.645se, \bar{y} + 1.645se]$$

$$= [2.984904 - 1.645 \times 0.07, 2.984904 + 1.645 \times 0.07] = [2.87, 3.10]$$

Loosely speaking, this interval will contain the true value of $\mu_y$ with 90% probability.

Q6: In theory, this is the most difficult question. Now the sample size is only 50. So the central limit theorem cannot be applied. The distribution of $\bar{y}$ can be normal, or not. Most people forgot discussing the normality. If $y$ is normally distributed, then $\bar{y}$ must also be normally distributed (not because of central limit theorem, but because a linear combination of normal is still normal). In that case we only need to recalculate the standard error using the smaller sample size $se = \frac{0.9993776}{\sqrt{50}}$ and use the new se to obtain the new confidence interval.
On the other hand, if normality does not hold, then we need to get many sample means using many different samples. Then the percentiles of those sample means can define the confidence interval for $\mu_y$. If you want to know more about this method, google “bootstrap”

Q7: In reality, this turns out to be the most difficult question. The proper critical value should be -1.645 because

$$P(Z > -1.645) = 1 - P(Z < -1.645) = 1 - 0.05 = 0.95$$

Then

$$0.95 = P(y > x) = P\left(\frac{y - \mu_y}{\sigma} > \frac{x - \mu_y}{\sigma}\right) = P(Z > -1.645)$$

Thus

$$\frac{x - 2.984904}{0.9993776} = -1.645 \Rightarrow x = 2.984904 - 1.645 \times 0.9993776 = 1.34$$

In class, we know how to define the “big value” as the one that is greater than, say, 95% observations. This problem is concerned with “small value”, which is less than 95% observations. In short, statistically speaking, “big” or “small” value corresponds to certain percentile.

Q8: i.i.d means independently and identically distributed. The three assumptions are

$$E(y_i) = \mu, \var(y_i) = \sigma^2, \cov(y_i, y_j) = 0, \forall i, \forall j \neq i$$

Q9: the key is to choose the sample using random numbers such as student campus ID numbers. For example, one random sample may be 100 Miami students whose ID numbers are odd, or end with digit 7. Notice that the random number is out of control of people, so is exogenous. In contrast, somebody suggests that we can get the sample by doing the survey in the Armstrong student center. This is problematic if rich students tend to go there more often than poor students, because the food there can be pricey. The sample average family income using this biased sample can be greater than the true value. This is an example of selection bias, which occurs when people choose endogenously to be in the sample.

Q10:

$$E(\bar{y}) = E\left(\frac{y_1 + y_2 + y_3}{3}\right) = \frac{\mu + \mu + \mu}{3} = \mu$$
This shows that the sample mean can still be an unbiased estimator for the population mean even if the sample is not i.i.d. This non i.i.d sample, however, satisfies the key condition of equal mean: \( E_y = \mu, \forall i \).

Q11:

\[
\text{var}(\bar{y}) = \frac{\text{var}(y_1) + \text{var}(y_2) + \text{var}(y_3)}{9} = \frac{1 + 2 + 3}{9} = 0.67
\]

Many people forgot squaring 3.

Q12:

\[
E(\bar{y}^2) = E(\bar{y}) + E(\bar{e}) = \mu_y + E(\bar{e}) \neq \mu_y
\]
because \( E(\bar{e}) > 0 \).

Q13:

\[
\text{var}(\text{Bernoulli } y) = p(1-p) = 0.24
\]

Q14: This statement is false because the focus is on \( \bar{y} \), which can be continuous, and therefore can be normally distributed, even if \( y \) is discrete (taking only two values)

Q15:

\[
\sum(y_i - \bar{y}) = \sum y_i - n\bar{y} = \sum y_i - n\frac{\sum y_i}{n} = \sum y_i - \sum y_i = 0
\]

Q16: a one-unit increase in \( x \) is associated with 0.4549946 increase in \( y \).

Q17: the key assumption is \( \text{cov}(x, u) = 0 \), which means \( x \) is exogenous and uncorrelated with the error term \( u \), which represents all other factors. Consider a counterexample. Suppose \( y \) is student’s gpa; \( x \) is student’s attendance, and \( u \) is student’s motivation. Because \( x \) and \( u \) are correlated here, the coefficient of \( x \) in the simple regression has no causal interpretation. The coefficient captures the effect of not only \( x \), but also \( u \), so is a biased estimator of true causal effect of \( x \) on \( y \).

You cannot “wing it” for eco 311. You have to study hard, on a regular basis. Seek help from me whenever questions arise. Don’t wait.