SOLUTIONS

(1) We check the rule of thumb that shows the Normal approximation for the sampling distribution of the sample proportion is appropriate. We can check that $n\hat{P} = 14 > 10$ and $n(1-\hat{P}) = 28 > 10$ OR we can check that $\hat{P} \pm 3\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} = 0.3333 \pm 3\sqrt{\frac{0.3333(1-0.3333)}{42}} = (0.12, 0.55)$ is in the valid range for a proportion (0,1). We believe this is a random sample and certainly 42 people are far less than 10% of the relevant population of people.

$$\hat{P} \pm z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$$

$$0.3333 \pm 1.96\sqrt{\frac{0.3333(1-0.3333)}{42}}$$

$$0.3333 \pm 1.96 * 0.0727$$

LCL = 0.19; UCL = 0.48; The 95% confidence interval estimate is (0.19, 0.48).

Interpretation: We are 95% confident that the fraction of individuals who would not notice an obvious fight when focused on something else will be captured by the interval from 0.19 to 0.48. In other words, the margin of error on the reported point estimate of 0.33 is 0.14, which is considerable.

(Note: The interpretation must make it clear that the random element is the interval and not the unknown population parameter.)

(Note: The question did not demand a certain confidence level. Students could have picked 90%, 99%, or some other number. Here are the 90% and 99% CI estimators, respectively.

$0.3333 \pm 1.645\sqrt{\frac{0.3333(1-0.3333)}{42}} = (0.21, 0.45)$ $0.3333 \pm 2.576\sqrt{\frac{0.3333(1-0.3333)}{42}} = (0.15, 0.52))$

(2) (a) True. The scatter diagram clearly shows scatter, which means that there are residuals and hence they will have a positive standard deviation. The assumption of homoscedasticity – equal variance of the residuals – appears to be met by the scatter diagram as the amount of scatter is roughly constant all along the line.

(b) True. We can calculate the coefficient of correlation as 0.75 ($= \sqrt{0.5625}$) and one way to interpret it is in term of standard deviations as given in the statement. (Also, the statement does not imply causality and merely describes how we observe X and Y moving with respect to each other.)

(c) False. The phenomenon of regression to the mean occurs when the y-variable is at least partially subject to randomness: that is, there is some scatter (not a perfect positive or negative relationship). Clearly we have scatter in this case and we discussed this in part (a). Further, the correct statement in part (b) shows regression to the mean: a one standard deviation increase in X is associated with less than a one standard deviation increase in Y. (The line is in fact steeper than 45 degrees – i.e. slope is greater than one – but the link between a slope being less than one and there being regression to the mean only holds for standardized data.)

(3) X = number of women in sample
p = 0.60
n = 100
P(X ≤ 30 | p = 0.60, n = 100) = ?

X ~ Binomial but check if we can use the Normal Approximation to find the probability. Students may use the rule of thumb from the textbook that n*p = 60 > 10 and n*q = 40 > 10 OR they may use the one we emphasized in class: $np \pm 3*(np(1-p))^{0.5} = 60 \pm 4.9$ which is well within the interval from 0 to 100. Once we've checked the rule of thumb we can use the Normal Approximation:
$P(X \leq 30) = P(Z \leq (30 - 60)/(100*0.6*0.4)^{0.5}) = P(Z \leq -6.12) \approx 0$

(Note: Some students may have been even more precise in their use of a continuous approximation to a discrete distribution and found P(X ≤ 30.5).)
(Note: Some students may have approached this in terms of proportions ($\hat{P} = \frac{X}{n}$), which would yield the exact same answer and is perfectly reasonable.)
(Note: Some students may have defined X to be the number of men rather than the number of women. P(X ≥ 70 | p = 0.4, n = 100) = P(Z ≥ 6.12) ≈ 0. This yields the same final answer.)

Sentence to fill in: "The probability of obtaining a sample as extreme as that one is for all practical purposes zero: $P(X \leq 30) = P(Z \leq (30 - 60)/(100*0.6*0.4)^{0.5}) = P(Z \leq -6.12) \approx 0$."

(4) (a) Choice (B) is wrong because if there is no variation at all in prices then we will never be able to estimate the effect of a change in price on quantity demanded.

Choice (C) is wrong because it would not address the fundamental problem with observational data that firms – no matter what the nature of competition – will react to market demand conditions in choosing prices and that will make price endogenous. (It is also ridiculous: how will we control the nature of competition in an international commodity market?)

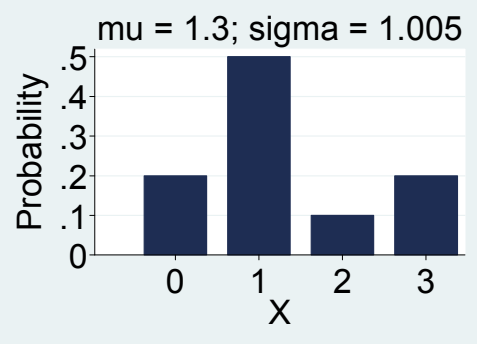
Choice (D) is wrong for two reasons. It is both impossible – how will you make sure that a person's tastes do not affect their choice about how much coffee to drink?! – and unnecessary because the key to experimental data is ensuring the X variable (prices) is randomly set. Once you do that you do not have to worry that some other factors may be systematically affecting both your X and your Y variable (since nothing can affect something that is random).

(b) No we should definitely not conclude that the positive association between coffee prices and coffee sold indicates an upward sloping demand for Arabica coffee and this is not what the article excerpts suggest. The excerpts merely describe equilibrium prices and quantities (quantity demanded=quantity sold) in this market. (We would not want to repeat the embarrassing historical mistake of Henry Moore who thought he found an exception to the Law of Demand in pig iron.) Our research question is about how price affects quantity demanded: elasticity is the percent change in quantity demanded divided by the percent change in price. The Law of Demand (downward sloping demand) says the elasticity of demand is negative: as prices go up, quantity demanded goes down, other things equal. The excerpts summarize observational data where the price of coffee is endogenous: it is affected by demand shifters such as the taste for coffee and rising income in developing countries which also affect the quantity demanded. Hence the high equilibrium price is partially caused by high demand (rightward shift in demand) as well as supply shocks like bad weather (leftward shift in supply) as opposed to the erroneous idea that high quantity demanded is being caused by high prices, which is what a positive elasticity of demand would mean.

(5) (a) $E[X] = \mu = 0 * 0.2 + 1 * 0.5 + 2 * 0.1 + 3 * 0.2 = 1.3$

$V[X] = \sigma^2 = (0 - 1.3)^2 * 0.2 + (1 - 1.3)^2 * 0.5 + (2 - 1.3)^2 * 0.1 + (3 - 1.3)^2 * 0.2 = 1.01$

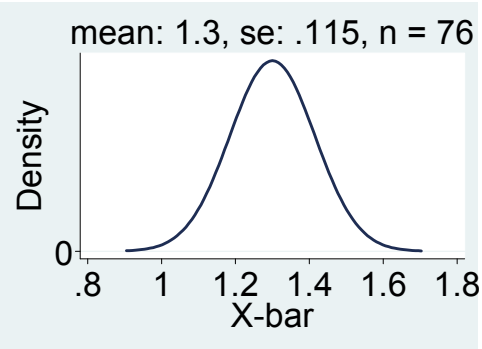
$SD[X] = \sigma = \sqrt{1.01} = 1.005$



(b) $E[\bar{X}] = \mu = 1.3$

$SD[\bar{X}] = standard\ error = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1.005}{\sqrt{76}} = 0.115$

The shape should be Normal (Bell shaped) because of the Central Limit Theorem (CLT) and we have a sample size of 76 which is sufficiently large given the above population.



(c) $P(X > 1) = P(X = 2) + P(X = 3) = 0.3 \quad [or\ \ P(X > 1) = 1 - P(X = 0) - P(X = 1) = 0.3]$

$P(\bar{X} > 1) = P\left(Z > \frac{1-1.3}{0.115}\right) = P(Z > -2.61) = 0.5 + 0.4955 = 0.9955$

[The sample mean is much less variable around the population mean (μ = 1.3) than the population is, which means there is a much greater chance that the sample mean will be bigger than one (a value below the mean).]

(6) (a) $E[\bar{X}] = \mu = \frac{a+b}{2} = \frac{0+1}{2} = 0.5$

$$V[\bar{X}] = \frac{\sigma^2}{n} = \frac{\frac{(b-a)^2}{12}}{n} = \frac{\frac{(1-0)^2}{12}}{10} = \frac{1}{120} = 0.0083333$$

$$SD[\bar{X}] = \sqrt{\frac{1}{120}} = 0.0912871$$

The results in the table above are not exactly equal to these theoretical results because of simulation error. Because we did 100,000 simulation draws the amount of simulation error is very small.

(b) These simulation results do suggest that the sample median is an unbiased estimator of the population mean because on average the sample median is coming out very close to the population mean (which is 0.5). These simulation results also suggest that the sample median is consistent because we see that the standard deviation of the sample median (its standard error) is getting smaller and smaller as the sample size increases – it goes from 0.14 to 0.09 to 0.05 as the sample size rises from 10 to 30 to 100 – and that it is unbiased. The sample median is converging towards the correct answer as the sample size grows. However, the sample median is NOT relatively efficient because the sample mean is also an unbiased estimator of the population mean and the sample mean has a smaller standard deviation (is less subject to sampling error) for all of the reported sample sizes: for example, for a sample size of 100 the s.e. of the sample mean is only 0.03 whereas it is 0.05 for the sample median.

(c) Yes. The Law of Large numbers says that a sample should be highly representative of the population when the sample size becomes large. The STATA summary shows a sample with a large sample size (n = 100,000) that aligns very closely with the population it is drawn from: Uniform[0, 1]. We see the sample mean and the sample median are very close to the population mean and median (0.5). Further the sample standard deviation (s = 0.2884576) is very close to the population standard deviation ($\sigma = (1 - 0)/12^{0.5} = 0.2886751$). In fact all of the reported sample statistics (including the range, IQR, percentiles, etc.) are very close to the population parameters because there is very little sampling error with a sample size of 100,000.

(d) This STATA summary describes a simple random sample whereas the results in Table 6.1 describe *sampling distributions* for two measures of central tendency (the sample mean and the sample median) for sample sizes ranging from 10 to 100. The sample size for the random sample in this STATA summary is 100,000. There is no relationship between it and the simulation results in Table 6.1. Further, it is no surprise that the smallest 1 percent of our sample is MUCH smaller than the smallest sample means and medians we observed as the sample mean and median will be much less variable than observations in a sample (or the population). In other words, we expect observations in a sample to be much more variable than a sample mean (or median) would be. Hence we should expect that the 1st percentile of the sample will be much smaller (more extreme) than the 1st percentile of a sampling distribution (of the mean or median), which is in fact what we see.