Midterm Test, ECO220Y, June 26, 2013 **SOLUTIONS**

(1) (a) $P(Unemployed \mid HS) = 0.088$. This is a conditional probability.

(b) $P(X = 1|n = 5, p = 0.074) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} = \frac{5!}{1!(4)!}0.074^1(0.926)^4 = 0.2720$

(2) (a) $V[Total] = V\left[\sum_{i=1}^{50} X_i\right] = \sum_{i=1}^{50} V[X_i] = \sum_{i=1}^{50} 160^2 = 50 * 25600 = 1{,}280{,}000$ liters-squared

$SD[Total] = \sqrt{1{,}280{,}000} = 1131.37$ liters

(b) The most important reason to check the scatter diagram is to make sure that there is in fact a linear relationship between the variables that we plan to summarize with a line: it appears linear in this case. A related reason is to check for outliers: there are none in this case. A final reason is to check for heteroskedasticity: there appears to be equal spread in this case. Because these are experimental data, the interpretation should be clearly causal: when the hotel raises the price of broadband usage by 1 cent per minute this reduces the minutes a guest uses by 4.8 minutes on average. The intercept does have an interpretation in this case: on average customers will use 119.1 minutes if the hotel does not charge for broadband usage. The R^2 means that 28% of the variation across guests in broadband usage is explained by variation in the price they are charged.

(3) (a) For a highly positively skewed distribution such as amount of money people spend on a wedding, the sample median is a better measure of what couples typically spend. The median marks the half-way point: half of couples spend more and half spend less. Instead the mean will be highly influenced (increased) by a handful of people who spend extremely large amounts of money and hence will be higher than what most people spend.

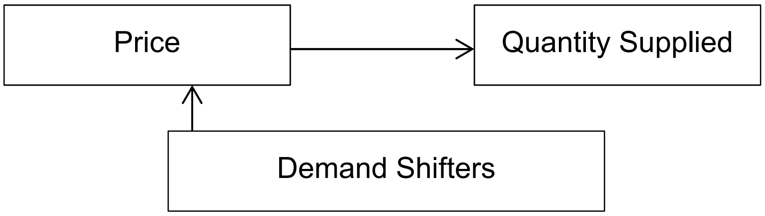
(b) The amount of money people spend on a wedding is a continuous random variable so the vertical axis must be labeled density and the graph should show a density function. The density function should show strong positive skew and not assign any density below $0. The horizontal axis should be labeled "Spending on a wedding in 2012 in dollars" or something similar and should show reasonable numeric tic marks. It should be consistent with a median of about $18,000 (i.e. it should look like about half the density is below that point and half above).

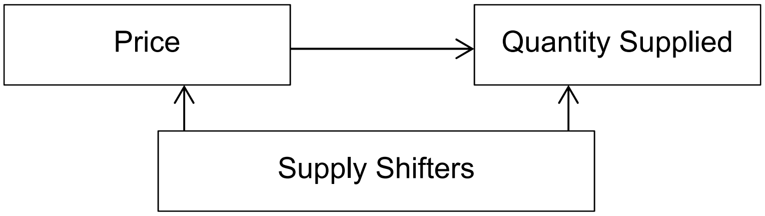
(c) This number cannot be calculated because we'd need to have the density function and we do not (just know that it is strongly positively skewed). Using the Normal distribution is highly inappropriate.

(d) No, the sample mean (a statistic) of $27,427 of spending on a wedding is a highly biased estimator because of non-sampling errors. As mentioned in the article excerpt, the sampling was done not from the entire population of people having a wedding but only from "gung-ho" (more extreme) brides that are actively participating in online wedding planning sites, which creates a serious selection bias. Hence the sample mean would be an upwardly biased estimator of the population mean.

(e) We find the standard deviation of the sample mean: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{\$18{,}000}{\sqrt{20{,}000}} = \$127.28$. This is a very tiny standard deviation compared to the population mean ($28,000), which means that the sample mean is hardly affected by sampling error. The huge sample size of 20,000 couples is indeed impressive. We can interpret it by using the Empirical Rule because according to the Central Limit Theorem (CLT) the sampling distribution of the sample mean will be Bell shaped (Normally distributed) as a sample size of 20,000 is surely sufficiently large. Hence 95.4 percent of the time the sample mean would be between \$27,745.44 and \$28,254.56, which is plus and minus two standard deviations, and is an extremely narrow range. (Note: This is NOT a confidence interval.)

(4) The development of improved corn plants is a good example of an unobserved (i.e. a lurking or confounding) variable in the estimation of the supply curve because it is a supply shifter. Supply shifters directly affect both the market quantity supplied and the market price. (Note: Following an increase in supply – a right-ward shift of the supply curve – the market price will go down while the market quantity supplied will go up. One needed to remember ECO100Y as well as homework from our course that reviewed those concepts to answer this question properly.) As the supply curve shifts, the observed time-series of price and quantities would be tracing out the DEMAND curve, not the supply curve. Of course demand would also be shifting over time, which would tend to trace out the supply curve. The combined effect of both curves shifting over time is that the scatter plot and OLS line above estimates neither the supply curve nor the demand curve. Another specific example of a choice that could have appeared instead of **(E)** and still been correct would be another supply shifter such as weather conditions, prices of farm land (an input), prices of fertilizers (an input), blights (insects, fungus, etc. that affects corn), another technological advance such as improved harvesting methods, etc. **(A)** – **(D)** are all demand shifters. Demand shifters do NOT SEPARATELY affect both the quantity supplied and the market price: instead they simply cause a movement along a given supply curve. Of course, a shift in demand results in a change in both the market price and the market quantity supplied (remember your graphs from ECO100): many students incorrectly wrote that demand can shift and that there would be no change in either the market price or the quantity supplied. In the first diagram below you can see that because price is affected quantity supplied is affected (i.e. movement along the supply curve so that a higher price translates into a higher quantity supplied). In contrast, in the second diagram the supply curve itself moves: even if the price stayed the same the quantity supplied would change. Hence supply shifters *separately* affect price and quantity.





(5) $P\left(\hat{P} \geq 0.02 \middle| n = 200, p = 0.01\right) = 1 - P(X = 0) - P(X = 1) - P(X = 2) - P(X = 3)$

$$P(X = x) = \frac{n!}{x!\,(n - x)!} p^x (1 - p)^{n-x}$$

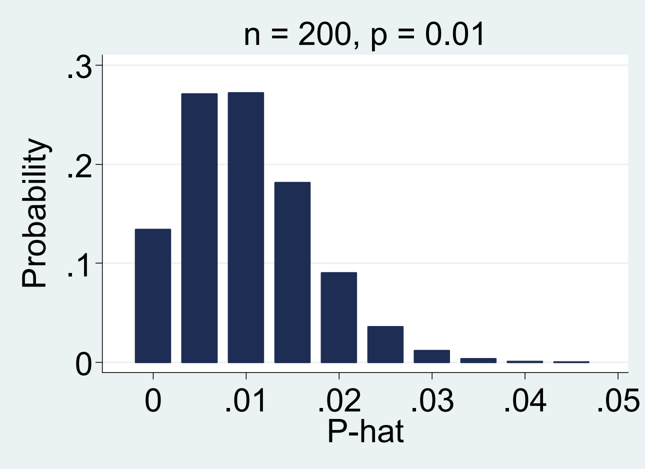
$P(X=0)=\frac{200!}{0!(200)!}0.01^0(0.99)^{200}=0.1340$

$P(X=1)=\frac{200!}{1!(199)!}0.01^1(0.99)^{199}=0.2707$

$P(X=2)=\frac{200!}{2!(198)!}0.01^2(0.99)^{198}=0.2720$

$P(X=3)=\frac{200!}{3!(197)!}0.01^3(0.99)^{197}=0.1814$

$$P(\hat{P}\geq 0.02|n=200,p=0.01) = 1-0.1340-0.2707-0.2720 - 0.1814=0.1419$$



Yes, sampling error is a plausible explanation: there is a 14.19% chance that even if the firm's claim were perfectly true that we could, by chance, observe a sample of 200 e-mails like the one we got.

(6) (a) It says that the margin of error is 2 percentage points for a confidence level of 95% (i.e. a significance level of 5%). It mentions three different point estimates so we can be conservative and put $\hat{P}$ = 0.5 in.

$ME=z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}=1.96\sqrt{\frac{0.5(1-0.5)}{1525}}=0.025$ (The article should have rounded up to 3%: as we'll see below the people who wrote/edited acted unethically in that they misrepresented the accuracy of the results. It does not matter whether it was deliberate or the result of ignorance (it is also unethical to report statistics you do not really understand).)

(b) When they divided the sample up to break out the results by geographic area this means that the sample size available for those inferences was far less than 1,525. For example, the ME for the 55% reported would be better estimated as $ME=z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}=1.96\sqrt{\frac{0.55(1-0.55)}{103}}=0.096$ because only 6.78% of the Canadian population lives in the Atlantic provinces and hence roughly only that fraction of the sample is used when finding the sample proportion of 0.55. Hence a margin of error of TEN percentage points NOT two percentage points is what we'd be dealing with. Alternatively, we could have illustrated this point with the proportion for Ontario where a much better estimate of the margin of error would be $ME=z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}=1.96\sqrt{\frac{0.41(1-0.59)}{590}}=0.040$ and hence a margin of error of FOUR percentage points NOT two percentage points.

(c)

$$\hat{P}\pm z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}=0.33\pm 1.96\sqrt{\frac{0.33(1-0.33)}{1525}}=0.33\pm 0.024$$

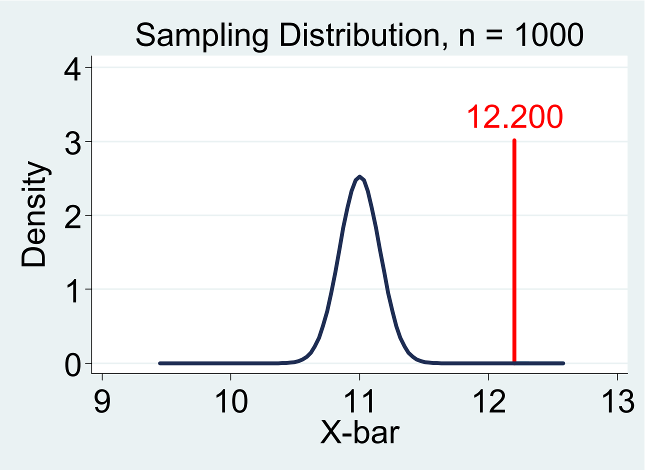
We are 95% confident that the interval from 30.6% and 35.4% would include the percent of all Canadians (adults) that believe that the scandal involving Senator Mike Duffy is the most embarrassing scandal of the three scandals asked about.

(7) (a)

$$P(\bar{X} > 12.2 \mid \mu = 11, \sigma = 5, n = 100)$$

$$= P\left(Z > \frac{12.2 - 11}{\frac{5}{\sqrt{1000}}}\right)$$

$$= P(Z > 7.6) \ \approx 0$$



(b) The three possible explanations for why a sample statistic may differ from the associated population parameter are: sampling error, non-sampling errors, or that the parameter is not what it is claimed to be. In part (a) we've determined that sampling error is a rather unlikely explanation for this discrepancy (less than a 1% chance). Hence the discrepancy is either because of non-sampling errors made when collecting the random sample of 100 taxi rides (e.g. selection bias or non-response bias if applicable) or because the taxi association is understating the average tip parameter.