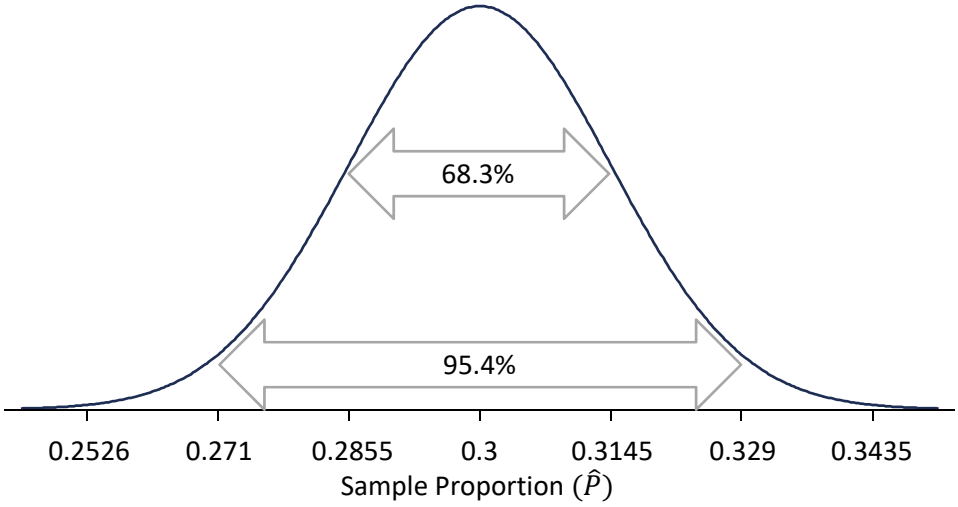
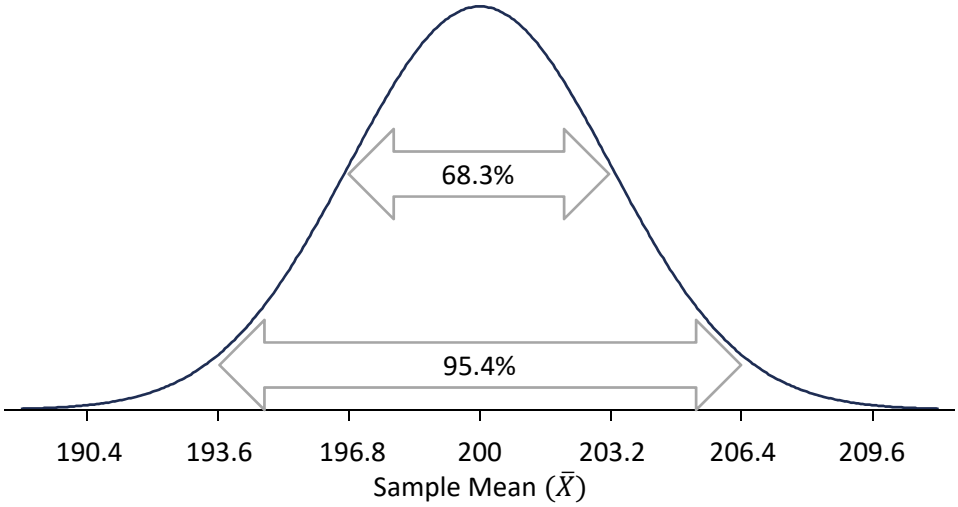
Homework 11: ECO220Y – SOLUTIONS

Required Problems:

(1) (a) $E[\hat{P}] = p = 0.3$ and $SD[\hat{P}] = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.3(1-0.3)}{1000}} = 0.0145$



(b) $E[\bar{X}] = \mu = 200$ and $SD[\bar{X}] = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{1000}} = 3.2$



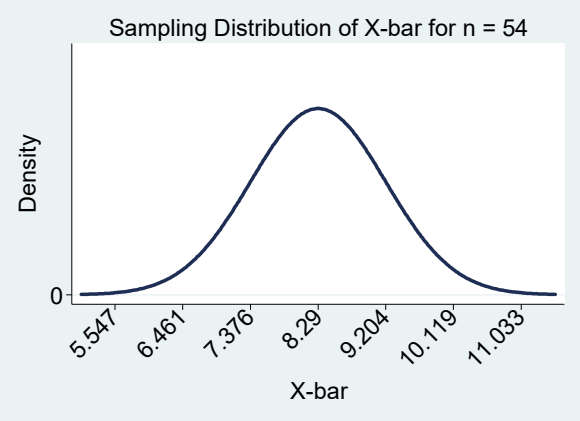
(2) (a) Because travel distances must be positive and there is not room to go even two standard deviations below the mean without heading into negative territory. Given the impossibility of a long tail below the mean (given the necessity of positive distances), the large s.d. can only be explained by a tail above the mean.

(b) You should draw a continuous density function that is clearly positively skewed and does not assign any density to negative distances.

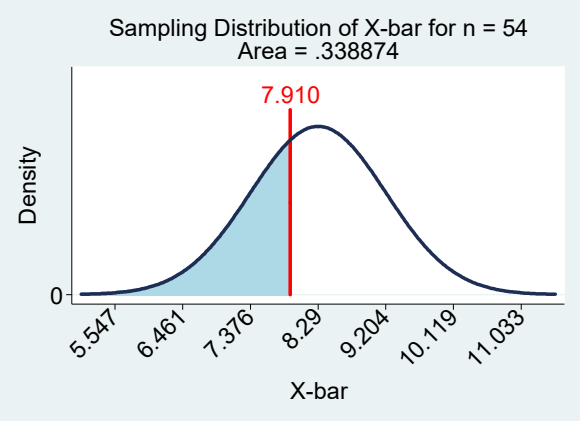
(c) It is impossible to find this probability with the given information: you certainly CANNOT use the Normal table as we've established that the distribution will certainly be right skewed and hence not Normal.

(d) You should draw a histogram that is clearly positively skewed (starting at zero).

(e) Use the CLT as a sample size of 54 will be sufficiently large such that the sampling distribution of the sample mean is Normal (Bell shaped) even though the population is definitely NOT Normal.



(f) Yes, sampling error is a plausible explanation for why our sample mean came out to be that much smaller than the population mean: $P(\bar{X} < 7.91 \mid \mu = 8.29, \sigma = 6.72, n = 54) = 0.34$.



(3) As explained in Section 10.5 of the textbook, standard errors are *estimates* of the standard deviation of sample statistics. They are estimates and hence not the same thing as the theoretical standard deviations of sample statistics derived in the textbook and in lectures. We need estimates because the actual standard deviation formulas involve *parameters* whose values we usually do not know.

(a) The standard error of $\hat{P}$ is simply the *estimate of* $SD[\hat{P}] = \sqrt{\frac{p(1-p)}{n}}$, which is $SE[\hat{P}] = \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$. The standard error is a practical necessity that addresses the inconvenient truth that we cannot get the exact value of $SD[\hat{P}]$ without knowing p, which is the (typically) unknown parameter we will need to make an inference about.

(b) The standard error of $\bar{X}$ is simply the *estimate of* $SD[\bar{X}] = \frac{\sigma}{\sqrt{n}}$, which is $SE[\bar{X}] = \frac{s}{\sqrt{n}}$. Again, we are very unlikely to know the value of the population standard deviation – the parameter σ – when we are trying to make an inference about an unknown population mean – the parameter μ. Hence, a practical compromise is to compute the standard error: it replaces the unknown σ with an estimate of it from our sample, which is the sample standard deviation s.

(c) As shown by the formula, $SE[\hat{P}] = \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$, the standard error of the sample proportion depends on two things: the sample size and the value of $\hat{P}$.

(d) As shown by the formula, $SE[\bar{X}] = \frac{s}{\sqrt{n}}$, the standard error of the sample mean depends on two things: the sample size and the value of the sample standard deviation.

(e) Standard errors measure the amount of sampling error: bigger values of the standard error mean more sampling error whereas smaller values of the standard error mean less sampling error. When we are trying to make inferences about unknown population parameters using a random sample and its statistics, we (ideally) do not want too much sampling error, which will mean less precise inferences. The lever the researcher uses to control sampling error is the sample size: bigger sample sizes mean less sampling error. Notice that both formulas for standard errors (for the statistics $\hat{P}$ and $\bar{X}$) have $\sqrt{n}$ in the denominator. Bigger sample sizes mean smaller standard errors (although there are diminishing returns given the shape of the square root function).

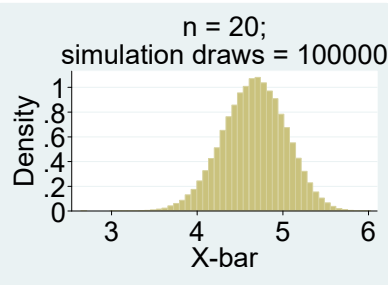
(4) (a) We cannot be sure. The population is skewed (although not heavily like salaries). We'd feel confident that a sample size like 100 would be sufficiently large but maybe unsure if the rough rule of thumb of 30 would hold given that the population is clearly skewed.

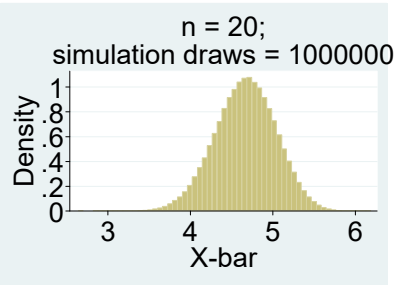
(b) A sample size of 5 has moved us towards a Normal shaped sampling distribution of the sample mean but there is still some negative skew.

(c) $E[\bar{X}] = \mu$ and $SD[\bar{X}] = \frac{\sigma}{\sqrt{n}}$. Hence $E[\bar{X}] = 4.67$ and $SD[\bar{X}] = \frac{\sqrt{2.72}}{\sqrt{5}} = 0.7376$. These are very similar to the STATA summary results from the Monte Carlo simulation as we would expect given the large number of simulation draws and small amount of simulation error.

(d) While there is still a tiny bit of negative skew this is close to Normal. It appears that a sample size of 20 is nearly sufficiently large (a bigger sample size would be better). Of course the Normal is always an approximation so it depends on the context as to how good of an approximation that we demand.

(e) You would expect no change (other than somewhat more bins). The amount of simulation error is already extremely small with 100,000 simulation draws. These show the results of those computer simulations:





(f) Because $E[\bar{X}] = \mu$ and $SD[\bar{X}] = \frac{\sigma}{\sqrt{n}}$ and the sample size for the first is 5 whereas the second is 20.

(g) The sample median is more subject to sampling error because it has a bigger standard deviation across samples, which is a quantitative measure of sampling error: 0.526078 versus 0.3690027.

(5) (a) The graph shows the simulated sampling distribution of the sample standard deviation of salary (measured in $1000's) for a random sample of 50 Ontario public sector employees in 2012.

(b) It means that on average the sample s.d. is 36.973. In other words, if you *imagine* repeatedly drawing drawing random samples of 50 employees on average the sample standard deviation will be 36.973 (where units are $1000s). We see from the first STATA summary of the population that the population standard deviation is actually 39.64454. The sample standard deviation is a biased measure of the population standard deviation. (If you're interested: in general the sample variance is an unbiased estimator of the population variance. However, remember that to get the s.d. you have to take a square root, which is a non-linear transformation. Hence just because the variance is unbiased does not mean that the standard deviation is unbiased.)

(c) It means that the standard deviation of the sampling distribution of the sample s.d. is 14.276. In other words, if you *imagine* repeatedly drawing random samples of 50 employees the standard deviation will vary across those samples and one measure of how much it will vary (i.e. of sampling error) is the standard deviation, which is 14.276 (where units are $1000s). (Just like we can think about a mean of a mean we can think about the s.d. of a s.d.) No, we cannot use the Empirical Rule to understand this better because we see that the shape of the sampling distribution is not Normal.

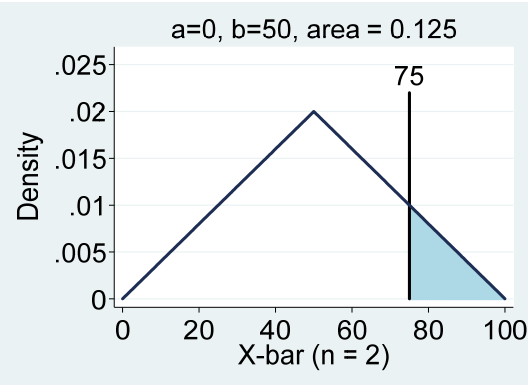
(d) If you collected a random sample of 50 public sector employees and the sample standard deviation came out to be 26.43 then this is lower than the population standard deviation, which is 39.64454 (see the first STATA summary in this question). Sampling error is a plausible explanation for the discrepancy because we can see from the STATA summary of the simulated sampling distribution of the sample standard deviation that the probability of such a low sample standard deviation is between 0.10 and 0.25 (quite a high chance).

(e) If you collected a random sample of 50 public sector employees and the sample standard deviation came out to be 43.21 then this is higher than the population standard deviation, which is 39.64454 (see the first STATA summary in this question). Sampling error is a plausible explanation for the discrepancy because we can see from the STATA summary of the simulated sampling distribution of the sample standard deviation that the probability of such a high sample standard deviation over 0.25 (a very high chance).

(f) If you collected a random sample of 50 public sector employees and the sample standard deviation came out to be 100.97 then this is higher than the population standard deviation, which is 39.64454 (see the first STATA summary in this question). Sampling error is NOT a plausible explanation for the discrepancy because we can see from the STATA summary of the simulated sampling distribution of the sample standard deviation that the probability of such a high sample standard deviation less than 0.01 (a low chance).

(6) The most surprising would be a sample median below $107K (i.e. choice **(C)**). From the Stata summary, we can see that P(Sample Median < 108.8) ≈ 0.01, which means that there must be less than a 1 percent chance of sample median below $107K: it is highly unlikely and hence would be quite surprising. In contrast, there is more than 1% chance of a median above $122K (choice **(A)**), a more than 25% chance of sample median between $113K and $116K (choice **(B)**), and a more than 4% chance of a sample median between $120K and $125K (choice **(D)**). All those arguments rely on the "Percentiles" part of the Stata output.

(7) (a) $P(\bar{X} > 75) = ?$ The shape of the population is Uniform from 0 to 100 because percentile scores are evenly spread out (i.e. the chance of being between the 40th and 50th percentile is 10% and the chance of being between the 90th and 100th percentile is also 10%). Hence, $X \sim U[0,100]$. A sample size of 2 is not sufficiently large for application of the CLT. However, if $X_1 \sim U[0,100]$ and $X_2 \sim U[0,100]$ and these are independent (which is surely true in this case if we've drawn two children at random: i.e. not done something like draw a sibling pair) then we know from Lecture 9 that $(X_1 + X_2) \sim T[0,200]$, which is the Triangle distribution. Further, $\bar{X} = \frac{(X_1+X_2)}{2}$, which is just a linear transformation of $(X_1 + X_2)$. Remember that linear transformations do *not* affect the shape. Hence, $\bar{X} \sim T[0,100]$ where $2a = 0$ and $2b = 100$, which means $a = 0$ and $b = 50$. Remembering that we find areas under the Triangle density function using A=0.5*base*height (no need for a statistical table), we find that $P(\bar{X} > 75) = 0.5 * 25 * 0.01 = 0.125$, which is illustrated in the graph above.



(b) Given that the population is Uniform, which is perfectly symmetric, a sample size of 20 is sufficiently large to apply the CLT. Further, for $X \sim U[0,100]$ we know $E[X] = \mu = 50$ and $V[X] = \sigma^2 = \frac{(100-0)^2}{12} = 833.33$. Hence, we standardize and use the Normal table: $P(\bar{X} > 60) = P\left(Z > \frac{60-50}{28.8675/\sqrt{20}}\right) = P(Z > 1.55) = 0.5 - 0.4394 = 0.06$.

(c) The top three quartiles means above the 25th percentile. Hence, $X \sim U[25,100]$. This affects the mean and standard deviation – we start with $\mu = 62.5$ and $\sigma = 21.6506$ – but the general approach in Part (b) still works.