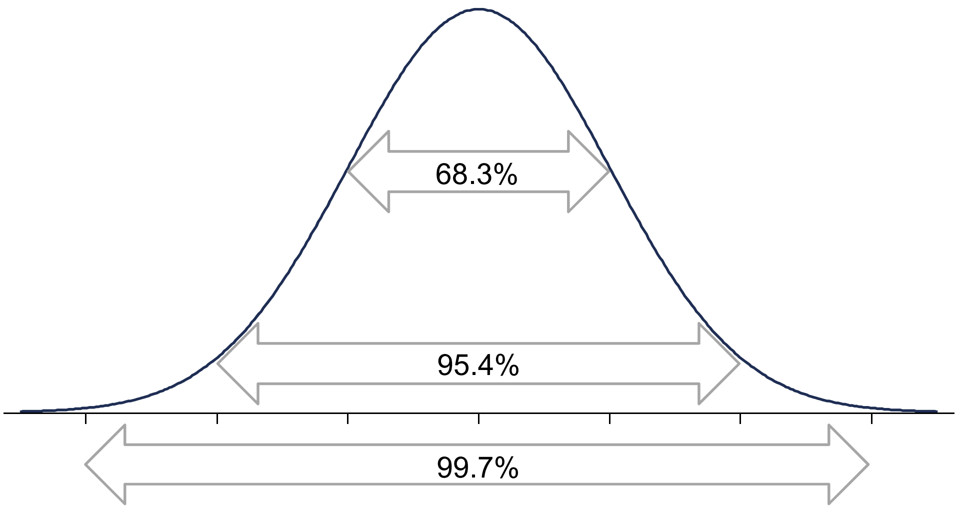
ECO220Y: Homework 11

Required Exercises: Chapter 10: 8 – 10, 11, 13, 33 – 35, 42, 44, 49, 50, 55, 56, 58, 61

(1) We can often use the Normal model for the sampling distribution of the sample proportion ($\hat{P}$) and the sampling distribution of the sample mean ($\bar{X}$). (Regarding *when* we can use the Normal model, for $\hat{P}$ we discussed rules of thumb based on the Normal approximation to the Binomial and for $\bar{X}$ we discussed the Central Limit Theorem (CLT). In both cases, the larger the sample size, the better the Normal model fits the sampling distribution.) With the Normal model we use the Empirical Rule to help label the horizontal axis values. Label each tick mark in a graph like that below and give an overall label to the x-axis. Starting from the left, the seven tick marks on the x-axis show 3 s.d.'s below the mean, 2 s.d.'s below the mean, 1 s.d. below the mean, the mean, 1 s.d. above the mean, 2 s.d.'s above the mean and 3 s.d.'s above the mean, respectively. Practice completing this graph in the parts below.



(a) Using the template above, sketch the sampling distribution of the sample proportion if the population proportion is 0.3 and the sample size is 1,000.

(b) Using the template above, sketch the sampling distribution of the sample mean if the population mean is 200, the population standard deviation is 100, and the sample size is 1,000. (Remember, you are sketching the *sampling distribution*, not the population distribution or the distribution of the sample. It is unlikely that the population or sample are Normal: these are very often skewed or otherwise not Normal.)

(2) Suppose the distance customers travel to a retail location is 8.29 km on average with a s.d. of 6.72 km.

(a) How do you know the distribution of travel distance is positively skewed and certainly not Normally distributed?

(b) Sketch a rough graph of what the distribution of travel distances may look like? Label it fully. Give a brief explanation of why you drew it the way you did.

(c) If you randomly select a customer, what is the chance s/he traveled between 8 and 9 km?

For Parts **(d) – (f)**, suppose that for a random sample of 54 customers the sample mean is only 7.91 km.

(d) Sketch a graph showing a plausible distribution your sample. Briefly explain why you drew it the way you did.

(e) Draw a fully labeled graph of the sampling distribution of $\bar{X}$. Briefly explain why you drew it the way you did.

(f) What is the chance sampling error explains why the sample mean is smaller than the presumed population mean? Is sampling error a plausible explanation for the discrepancy?

(3) In general, what are "standard errors"? Are they the same thing as standard deviations?

(a) What is the standard error (also written as s.e. or SE) of $\hat{P}$?

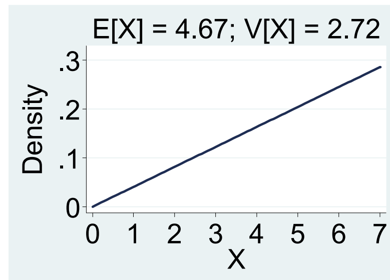
(b) What is the standard error (also written as s.e. or SE) of $\bar{X}$?

(c) What factors affect the size of the standard error of $\hat{P}$?

(d) What factors affect the size of the standard error of $\bar{X}$?

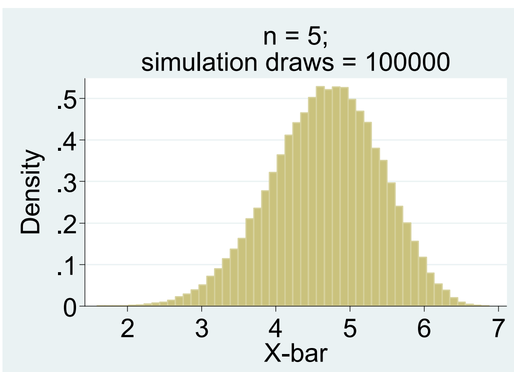
(e) Why are small standard errors preferable? What can a researcher do to get smaller standard errors?

(4) Consider the population distribution to the right. I'll refer to this a ramp-shaped.



(a) How large of a sample size is necessary such that the CLT applies and the sampling distribution of the sample mean is Normally distributed? How can you answer this question?

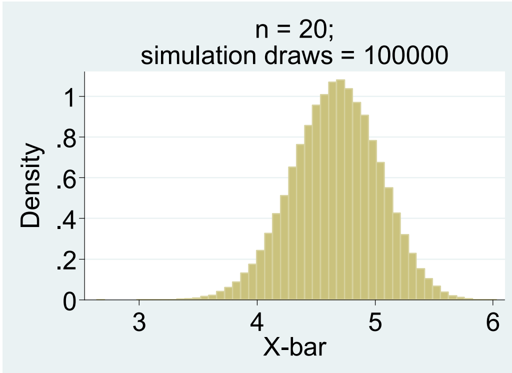
(b) Consider the following results of a Monte Carlo simulation. Review these results and assess how they inform the answer the question in Part (a).




```
                            X-bar
-------------------------------------------------------------
      Percentiles      Smallest
 1%     2.850204       1.587979
 5%     3.396197       1.673443
10%     3.690418       1.681243       Obs              100000
25%     4.176221       1.715443       Sum of Wgt.      100000

50%     4.697264                      Mean           4.665603
                        Largest       Std. Dev.      .7362233
75%     5.190035       6.771663
90%     5.599048       6.772544       Variance       .5420248
95%     5.824379       6.783857       Skewness      -.2448487
99%      6.19832       6.875974       Kurtosis       2.869772
```

(c) What does theory tell us the STATA summary should say about the mean and s.d.? How do these theoretical results compare to the actual values reported (4.665603 and .7362233)?

(d) Consider the following results of a Monte Carlo simulation. Review these results and assess how they inform the answer the question in Part (a).

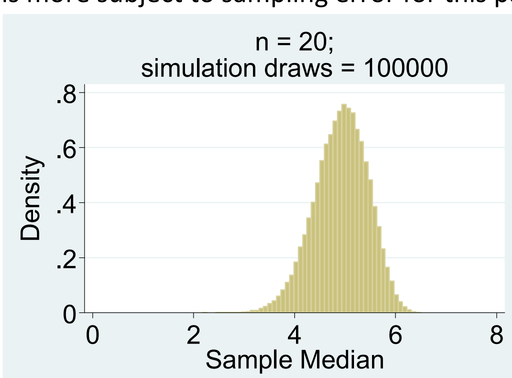



```
                          X-bar
-------------------------------------------------------------
      Percentiles      Smallest
 1%     3.774917       2.641003
 5%     4.047826       3.006247
10%     4.188137       3.071831       Obs              100000
25%     4.419997        3.08211       Sum of Wgt.      100000

50%     4.674486                      Mean           4.666657
                        Largest       Std. Dev.      .3690027
75%     4.922324       5.941281
90%     5.135538        5.95461       Variance        .136163
95%     5.258768       5.977927       Skewness      -.1330226
99%     5.484841       6.029417       Kurtosis       2.977807
```

(e) How would you expect the graph in the previous part to differ if instead of 100,000 simulation draws you did 1,000,000 simulations draws?

(f) Comparing the simulation results in Parts **(b)** and **(d)**, why are STATA summary results for the mean extremely similar (4.665603 versus 4.666657) but for the standard deviation very different (.7362233 versus .3690027)?

(g) Considering also the simulation results below for the sample median for a sample size of 20, which sample statistic is more subject to sampling error for this particular ramp population: sample mean or sample median?

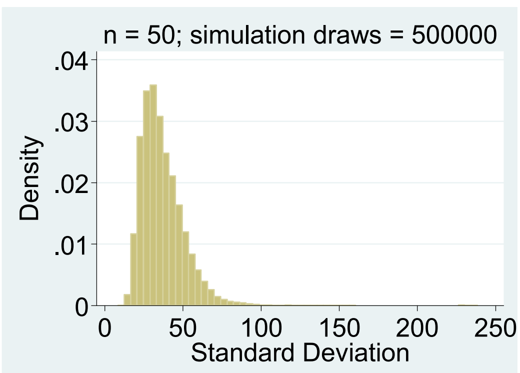



```
                         median
-------------------------------------------------------------
      Percentiles      Smallest
 1%     3.596848        2.17536
 5%     4.011773       2.487504
10%     4.223432       2.596495       Obs              100000
25%     4.573925       2.625426       Sum of Wgt.      100000

50%     4.946564                      Mean           4.919341
                        Largest       Std. Dev.       .526078
75%       5.2945        6.43844
90%     5.579855       6.448429       Variance        .276758
95%     5.737637       6.449238       Skewness      -.2969399
99%     6.000215       6.487709       Kurtosis       2.975978
```

(5) Recall the 2012 Ontario Public Sector Salary example from Lecture 10. The summary of the population of employees' salaries is reproduced below. Also, in class we used a Monte Carlo simulation to investigate the sampling distribution of the sample mean and the sample median. Simulation can be used to find the sampling distribution of any sample statistic. Below are also the results for the sample standard deviation.

(a) What does the graph show?

```
.  su salary, detail;
                            salary
-------------------------------------------------------------
      Percentiles      Smallest
 1%      100.168            100
 5%     100.9921            100
10%     102.0471            100       Obs               88545
25%     105.7447            100       Sum of Wgt.       88545

50%     115.3013                      Mean           127.5176
                        Largest       Std. Dev.      39.64454
75%     133.2821        843.095
90%     164.5416       935.2365       Variance        1571.69
95%      193.125        1036.74       Skewness       5.019101
99%     296.8753           1720       Kurtosis       64.99817
```




```
                       Standard deviation
-------------------------------------------------------------
      Percentiles      Smallest
 1%     16.97221       8.288662
 5%      20.3301       10.36123
10%     22.49249       10.43655       Obs              500000
25%      27.1103       10.57678       Sum of Wgt.      500000

50%     34.18233                      Mean           36.97279
                        Largest       Std. Dev.      14.27555
75%     44.02704       234.5967
90%     54.34834       235.4911       Variance       203.7912
95%     61.70084       236.8879       Skewness       2.561164
99%     81.39713       238.5874       Kurtosis       24.08033
```

(b) What does the value 36.97279 in the second STATA summary mean? Does it appear that the sample standard deviation is an unbiased estimator of the population standard deviation?

(c) What does the value 14.27555 in the second STATA summary mean? Can you use the Empirical Rule to interpret it?

(d) Suppose you collected a random sample of 50 public sector employees and the sample standard deviation came out to be 26.43. Is this higher or lower than the population standard deviation? Is sampling error a plausible explanation for the discrepancy?

(e) Suppose you collected a random sample of 50 public sector employees and the sample standard deviation came out to be 43.21. Is this higher or lower than the population standard deviation? Is sampling error a plausible explanation for the discrepancy?

(f) Suppose you collected a random sample of 50 public sector employees and the sample standard deviation came out to be 100.97. Is this higher or lower than the population standard deviation? Is sampling error a plausible explanation for the discrepancy?