1. A box model contains eight tickets marked 1, twenty marked 6, and twelve marked 8. Suppose you draw 150 tickets with replacement. Answer the questions below.

   (a) True or False. The population of values is normally distributed. *FALSE; the population only has 3 distinct values while the normal distribution is continuous (takes all values in a range)*

   (b) True or False. The population of values will be more normally distributed if we take a large sample from it. *FALSE; the population of values will always only have 3 distinct values - the population's distribution does not change with a change in sample size.*

   (c) True or False. The sampling distribution of the sample mean and the sampling distribution of the sample sum will be close to normally distributed if we take a large enough sample size ($n > 30$ will probably be large enough) *TRUE, this is what the Central Limit Theorem tells us: for a reasonably large number of draws with replacement from a box (i.e., a large sample,) the probability histogram of the sum and mean of those draws will follow a normal distribution, even if the content of the box does not.*

   (d) What is the expected value of the sum of the 150 tickets drawn? What is the expected value of the average of the 150 tickets drawn? *The expected value of the mean is the same as the average of the box. So* $EV_{mean} = \frac{8*1+20*6+12*8}{8+20+12} = \frac{224}{40} = 5.6$. *The expected value of the sum is* $EV_{sum} = number of draws * EV_{mean} = 150 * 5.6 = 840$.

   (e) What is the SE of the sum of the 150 tickets drawn? What is the SE of the average of the 150 tickets drawn? *The SD of the box is* $\sqrt{\frac{8*(1-5.6)^2+20*(6-5.6)^2+12*(8-5.6)^2}{40}} = \sqrt{\frac{241.6}{40}} = 2.46$. *Thus the SE of the sum of 150 draws is* $SE_{sum} = \sqrt{number\ of\ draws} * SD(box) = \sqrt{150} * 2.46 = 30.13$, *and the SE of the average is* $SE_{mean} = \frac{SD(box)}{\sqrt{number\ of\ draws}} = \frac{2.46}{\sqrt{150}} = 0.20$.

   (f) What is the approximate chance that the average of the 150 tickets drawn is less than 6.1 (Draw and label a picture to help you answer this question)? *The number of draws is large, so the CLT applies and we can use the normal approximation. This normal distribution will have* $EV_{mean} = 5.6$ *as the mean and* $SE_{mean} = 0.2$ *as the standard deviation. Converting to a z-score, we want the chance of being to the left of* $\frac{6.1-5.6}{0.2} = 2.5$, *which from the table, is* 0.9938, *or* 99.38%.

   (g) Suppose you now drew 4 tickets with replacement instead of 150. Without doing any additional calculations, will the SE of the *sum* of the 4 tickets be larger or smaller than in part (b)? Explain. *It will be smaller with 4 draws. The square root law* $SE_{sum} = SD(box) * \sqrt{n}$ *says the give or take number for the sum gets larger as the number of trials increases. (For 4 draws, the SE of the sum is* $2.46 * \sqrt{4} = 4.92$.*)*

(h) Again suppose you drew 4 tickets with replacement instead of 150. Without doing any additional calculations, will the SE of the *average* of the 4 tickets be larger or smaller than in part (b)? Explain.
It will be larger with 4 draws. The law of averages $SE_{mean} = \frac{SD(box)}{\sqrt{n}}$ *says the give or take number for the average gets smaller as the number of trials increases. (For 4 draws, the SE of the average is* $\frac{2.46}{\sqrt{4}} = 1.23$.*)*

(i) Again suppose you drew 4 tickets. Can you compute the approximate chance of getting an average less than 6.1? If you can, do so. If you can't, explain why.
You can't compute this. The population is not normally distributed. So we need the number of draws to be large enough to apply CLT. Recall the rule of thumb is that n of at least 30 will work for most boxes. Considering the distribution is asymmetrical, it's better to have at least 100 draws. Although you can calculate the exact probability by summing the probability of all situations satisfying this condition.

Summary:

i. *The formulae of* $EV_{mean}, EV_{sum}, SE_{mean}$ *and* SE_{sum} *have no requirement of n.*
ii. *CLT can only be used to approximate the distribution of the mean and the sum.*
iii. *Before applying CLT, always check if n is large enough first, unless the original distribution is normal. In most case,* $n \geq 30$ *is enough. If the data is very skewed, it's better to have* $n \geq 100$.

(j) Again suppose you drew 150 *tickets. Compute the approximate chance of getting at least 45% 8s, if possible.* Since 8's are the success - we can redraw the box as 12 successes (1's) and 28 failures (0's). So the $EV_{\%} = 12/40 * 100 = 30\%$. The $SD(box) = (1-0)\sqrt{\frac{12}{40} * \frac{28}{40}} = 0.458$. The $SE_{\%} = \frac{0.458}{\sqrt{150}} * 100 = 3.74\%$. Since our sample size is $n = 150$, we are confident that the CLT will kick in to make the sampling distribution of the sample proportion approximately normally distributed with 30% as the mean and 3.74% as the SD. $z - score = \frac{45-30}{3.74} = 4.01$ Thus $P(Percent\, of\, 8's \geq 45\% \approx P(Z \geq 4.01) < 0.0002$.