1. 400 hundred draws are made at random with replacement from a box of numbered tickets containing 2 $\boxed{1}$s and 8 $\boxed{0}$s.

(a) (2 pts) What is the ***expected*** percentage of $\boxed{1}$s in the sample?

$$\textit{Expected percentage} = \textit{percentage of } \boxed{1} \textit{ s in box} = \frac{2}{10} \times 100\% = 20\%$$

(b) (2 pts) What is the ***standard error*** (SE) for the percentage of $\boxed{1}$s in the sample?

$$\textit{SE for the sample } \% = \frac{\sqrt{(\textit{proportion of } \boxed{1})(\textit{proportion of } \boxed{0})}}{\sqrt{\textit{sample size}}} \times 100\% = \frac{\sqrt{0.2 \cdot 0.8}}{\sqrt{400}} \times 100\% = 2\%$$

(c) (2 pts) What is the (approximate) probability that the sample percentage of $\boxed{1}$s is between 16% and 24%? Why?

Sample percentages approximately follow the normal curve when n *is big enough (which it is in this case), so*

$$P(16\% \leq \textit{sample } \% \leq 24\%) \approx 95\%$$

because this is the probability that the sample % falls in the range $EV(\%) \pm 2 \cdot SE(\%)$*, which is approximately equal to the area under the normal curve from* -2 *to* 2*.*

2. A market research firm surveyed a simple random sample of 3600 households from a large metropolitan area of more than 800,000 households.

(a) (4 pts) Of the sample households, 2160 owned two or more computers. Use this data to construct a 95%-confidence interval for the percentage of all households in the metropolitan area who own two or more computers. *Show your work.*

A 95%-confidence interval for the percentage of all households who own two or more computers is given by $(\textit{sample } \%) \pm 2SE(\%)$*. In this case, we have:*

$$\textit{sample } \% = \frac{2160}{3600} \times 100\% = 60\% \quad \textit{and} \quad SE(\%) \approx \frac{\sqrt{0.6 \times 0.4}}{\sqrt{3600}} \times 100\% \approx 0.816\%$$

so $2SE(\%) \approx 1.632\%$ *and the confidence interval is* $(60\% \pm 1.632\%)$*.*

(b) (2 pts) The 3600 households in the sample included 2500 children age 4 or younger. Of these 2500 children, 800 attended day care regularly.

True or false, and explain (briefly): A 95%-confidence interval for the percentage of all children age 4 or younger in the metropolitan area who attend day care regularly is $32\% \pm 1.87\%$.

False. *The sample % of children age* 4 *or younger who who attend day care regularly is* $(800/2500) \times 100\% = 32\%$ *and* $(\sqrt{0.32 \times 0.68}/\sqrt{2500}) \times 100\% \approx 0.933\%$*... So,* ***if this were a simple random sample*** *of children then the confidence interval would be correct. But it* ***isn't*** *a simple random sample of children — it is a* ***cluster sample****, so the calculation is incorrect.*

3. The average household income for all the households in the sample in the problem above was $\$52,000$ with an SD of $\$20,000$.

True or False, and justify your answer briefly:

(a) (2 pts) A 95%-confidence interval for the average household income in the metropolitan area is about $\$52,000 \pm \$667$.

True. *This is a simple random sample of households, the sample average is* $\$52,000$ *and the SE (for the average) is* $20000/\sqrt{3600} \approx \$333.33$, *so the given confidence interval is correct (because* $2 \times 333.33 = 666.66 \approx 667$*).*

(b) (2 pts) The chance is about 95% that the interval in (a) contains the average household income in the metropolitan area.

True. *This is precisely how a 95%-confidence interval for a population average is interpreted. More precisely, we expect that about 95% of all intervals constructed in this way will contain the population average and therefore any one of them (like the one in (a)) has a 95% chance of containing the population average.*

(c) (2 pts) The chance is about 95% that the interval in (a) contains the average household income in the sample.

False. *The chance is exactly 100% that the interval in (a) contains the average household income* ***in the sample****, because the sample average is the middle of the interval.*

(d) (2 pts) About 95% of the households in the metropolitan area have incomes between $\$51,333$ and $\$52,667$.

False. *The standard error (333.33 in this case) measures the variation between different sample averages,* ***not*** *between different household incomes. The variation in income between households is estimated by the sample SD = 20000. Moreover, there is no reason to believe that household income in this region has a normal distribution.*