Question 1.1. At least _____% of the people are between 25 and 65 years old. **(6 Points)**

Type your answer here, replacing this text.

SOLUTION: 93.75; Since we cannot make further assumptions about the *distribution* of ages within the community (i.e. we cannot assume them to be normally distributed), we must use Chebyshev's inequality to create a bound on the variability. Since we know that the mean of the ages is 45 with an standard deviation of 5, we see that 25 is -4 SDs from the mean and 65 is +4 SDs from the mean. Chebyshev's tells us that +/-4 SDs from the mean covers at least 93.75% of the people in the distribution.

Question 1.2. At most _____% of the people have ages that are not in the range 25 years to 65 years. **(6 Points)**

Type your answer here, replacing this text.

SOLUTION: 6.25; Since the total percentage of people within the community is 100%, and since we know that at least 93.75% of the people are between 25 and 65 years old, we can substract 100% - 93.75% = 6.25% to get the maximum number of people that are not in the range 25 to 65 years old. We know that this translates to there being *at most* 6.25% of people outside of the range [25,65] because Chebyshev tells us that *at least* 93.75% within the range (i.e. the percentage could be higher than this 93.75%).

Question 1.3. At most _____% of the people are more than 65 years old. **(6 Points)**

Hint: If you're stuck, try thinking about what the distribution may look like in this case.

Type your answer here, replacing this text.

SOLUTION: 6.25; Question 1.2 tells us that "At most 6.25% of the people have ages that are not in the range 25 years to 65 years". We don't know anything about the distributions of this 6.25% of people, but we do know that we would *maximize* the number of people that were more than 65 years old if we put ALL 6.25% of people that weren't in the range [25 year, 65 years]. In this case, there would be no one that was less than 20 years. Thus, this extreme tells us that at most 6.25% of the people are more than 65 years old.

Question 2.2. Suppose the data science class decides to construct a 90% confidence interval instead of a 95% confidence interval, but they still require that the width of the interval is no more than 6% from left end to right end. Will they need the same sample size as in 2.1? Pick the right answer and explain further without calculation. **(6 Points)**

1. Yes, they must use the same sample size.
2. No, a smaller sample size will work.
3. No, they will need a bigger sample.

Type your answer here, replacing this text.

SOLUTION: 2; In terms of standard deviations, the width of a 90% confidence interval is smaller than a 95% confidence interval. We know that we need to look +/- 2 SDs from the mean for our 95% CI. Thus, the number of deviations we have to look left and right will be less for our 90% CI.

However, this question requires that both intervals have a width at most 0.06. To achieve this, we would have to take a larger sample size for our 95% CI. Thus, for our 90% CI, we could take a smaller sample size than what we needed for our 95% CI.

Algebraically, we can think about it in the following way:

$$\text{Desired Width} \geq 4 \cdot \frac{\text{Pop SD}}{\sqrt{\text{Sample Size}}}$$

The above equation works for constructing a 95% CI. We want a width at most 0.06. Let's rearrange the above equation to solve for sample size:

$$\text{Sample Size} \geq (4 \cdot \frac{\text{Pop SD}}{0.06})^2$$

The 4 in the above equation is unique to a 95% CI. As we mentioned before, if we chose to construct a 90% CI instead, we would need less than four SDs as our width. Clearly, if we swapped the 4 on the righthand side of the above equation with something less than 4, say 3, the sample size needed would decrease.

Question 2.4. This shows that the percentage in a normal distribution that is at most 1.65 SDs above average is about **95%**. Explain why 1.65 is the right number of SDs to use when constructing a **90%** confidence interval. **(6 Points)**

Type your answer here, replacing this text.

SOLUTION: The level of confidence is a central area. On the left hand side, the confidence interval will stop at 1.65 SDs below the center. That is, it will stop at -1.65 in standard units. Since the area to the right of 1.65 is 5%, the area below -1.65 is also 5%, so the area between -1.65 and 1.65 is 90%.

Question 3.2. Why does the Central Limit Theorem (CLT) apply in this situation, and how does it explain the distribution we see above? **(6 points)**

Type your answer here, replacing this text.

SOLUTION: If we think of Yes votes as 1 and No votes as 0, then the sample is a collection of numbers. A resample is sampled with replacement from that collection, so a resample mean is the mean of a sample with replacement from some collection of numbers. The CLT therefore applies to the resample means: across resamples, they will have an approximately normal distribution. That's why the histogram above is bell-shaped.