SEIS 631

Dan Ward

Assignment 4

**Question 1 [MULTIPLE CHOICE] Which of the following is false?**

(e) The smallest house is 534 square feet and the largest is 3,642 square feet.

**Question 2: Describe the distribution of this sample? How does it compare to the distribution of the population?**

The distribution of the sample has a min of 520 and a max of 2524, with a median of 1384, mean of 1433. The extremes of the sample are not as high on the high end or low on the low end. The overall trend of the histogram is similar, it is unimodal with the peak at the same point.

**Question 3 [MULTIPLE CHOICE] Suppose we took two more samples, one of size 100 and one of size 1000. Which would you think would provide a more accurate estimate of the population mean?**

(c) Sample size of 1000

**Question 4: Describe the sampling distribution (the distribution of the sample means that you just created), and be sure to specifically note its mean. Which theorem describes the behavior of the sampling distribution?**

The distribution looks very similar to the normal distribution. The mean is 1499 which is very close to the mean of the total population. At the same time the min and max values are much less extreme than that of the population.

The theorem that describes this behavior is the Central Limit Theorem.

**Question 5: [MULTIPLE CHOICE] Which of the following is true about the elements in the sampling distributions you created?**

(a) Each element represents a mean square footage from a simple random sample of 50 houses.

**Question 6: It makes intuitive sense that as the sample size increases, the center of the sampling distribution becomes a more reliable estimate for the true population mean. Also as the sample size increases, what can you say about the variability of the sampling distribution? Explain your answer.**

As the sample size increases the mean of the sample trends towards the mean of the population, thus reducing the overall variability. The more random data points we have from the population the more accurately it will represent the population mean.

**Question 7: Take a random sample of size 50 from price. Using this sample, what is your best point estimate of the population mean?** From this sample the mean price is 174,929

**Question 8: Since you have access to the population, simulate the sampling distribution of price by taking 5000 samples from the population of size 50 and computing 5000 sample means. Store these means in a vector called sample\_means50. Plot the data, then describe the shape of this sampling distribution. Based on this sampling distribution, what would you guess the mean home price of the population to be?** The distribution looks like the normal distribution. Based on this data it looks like the mean price of a home is around $180,000 to $185,000

**Question 9: Change your sample size from 50 to 150, then compute the sampling distribution using the same method as above, and store these means in a new vector called sample means150. Describe the shape of this sampling distribution, and compare it to the sampling distribution for a sample size of 50. Based on this sampling distribution, what would you guess to be the mean sale price of homes in Ames?**

The sampling of the 150 sample set is a tighter normal distribution and more symmetric than the 50 sample set. The mean of the 150 sample set, appears to be tighter at the $180,000 point.

**Question 10: [MULTIPLE CHOICE] Which of the following is false?**

1. The variability of the sampling distribution with the smaller sample size (sample means50) is smaller than the variability of the sampling distribution with the larger sample size (sample means150).

**Question 11: Compare the distribution of the original population price to the sampling distribution (distribution of sample means). If they appear different, explain why.**

All three plots appear to be a normal distribution with the population data differing from the other two sample means, such that the population data is right skewed and it has a much greater variability. This is due to the fact that the other two data sets are means of a randomly selected group so they will remove the extreme variability.