Summative Quiz 2 (Continuous Data Measures) Solutions:

1. Which of the following is true about an observational prospective cohort study being designed to study the association between cigarette smoking (smoking ≥ 1 pack of cigarettes per day, smoking < 1 pack of cigarettes per day, or no smoking) to be assessed on 3/1/19, and catching a cold between 3/2/19 and 3/31/19?

   Answer: Subjects self-select to be in the exposure groups, so potential confounders will need to be considered when assessing the outcome/exposure relationship.

2. Which of the following statements defines the 97.5th percentile value for a data sample, regardless of the sample data distribution?

   Answer: The value such that 2.5% of the sample observations are larger than this value, and 97.5% of the sample observations are less than (or equal to) this value.

3. 500 randomly selected university students were selected for a study on smoking status and stress levels. At the time of the study, students were asked to rate their stress levels on a scale of 0 – 100 with 0 indicating no stress. Students were also asked to disclose whether they were a non-smoker, a light smoker, or a heavy smoker.

   Of the 500 students, 75 identified themselves as heavy smokers, 150 identified themselves as light smokers, and 275 identified themselves as non-smokers. The average stress level reported for heavy smokers was 60.4. The average stress level reported for light smokers was 45.7. The average stress level reported by non-smokers was 56.6. Based on these results, which of the following statements is true?

   Answer: On average, heavy smokers have stress values of 3.8 greater than non-smokers

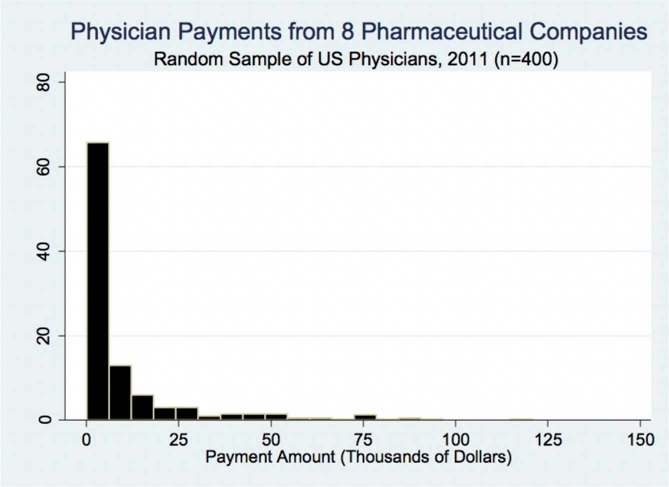
   Reasoning: 60.4 – 56.6 = 3.8.

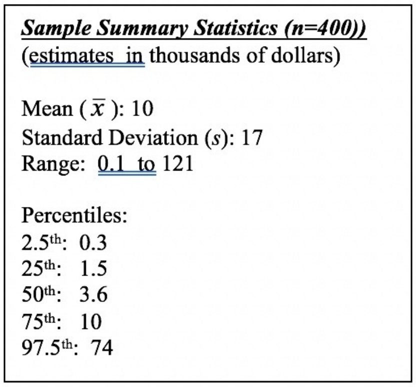
4. Which of the following sample statistics is least sensitive to the influence of outliers for a sample continuous data measures?

   Answer: The sample median (50th percentile)

   Reasoning: The sample median is the “middle” value in the data set, ie: the value that is greater than or equal to 50% of the data values, and less than 50% of the data values. Outliers, small or large, do not affect the position of the median (or most other percentiles) among all the data points, but will affect quantities that depend on the actual data values and not just the order of the data point.

5. This graph shows the total payment amounts in 2011 (in thousands of US$) for a **random sample** of 400 US physicians who received payments from any of 8 major pharmaceutical companies in 2011. The vertical axis represents the percent of physicians. (Source: Propublica "Dollars for Docs" Online Database)





Sample Summary Statistics (n=400))
(estimates in thousands of dollars)

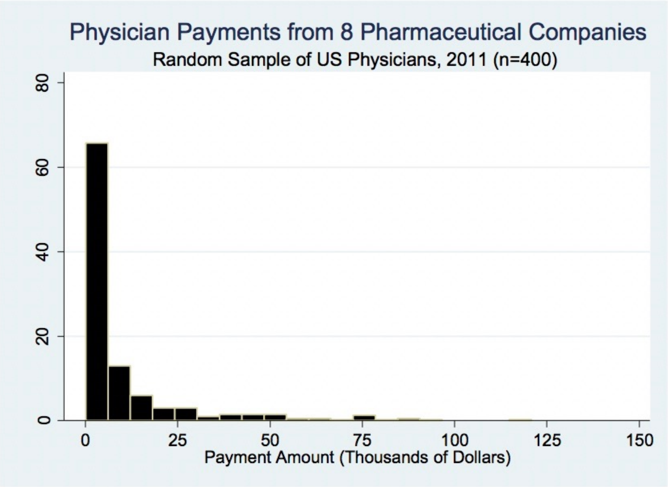
Mean ($\bar{x}$): 10
Standard Deviation (s): 17
Range: 0.1 to 121

Percentiles:
2.5th: 0.3
25th: 1.5
50th: 3.6
75th: 10
97.5th: 74

Which of the following statements best tells the story of these data?

Answer: The majority of physicians received relatively "small" payments (<$25,000) and a smaller percentage received payments greater than this majority.

6. This graph shows the payment amounts (in thousands of US$) for a **random sample** of 400 US physicians who received payments from any of 8 major pharmaceutical companies in 2011. The vertical axis represents the percent of physicians. (Source: Propublica “Dollars for Docs” Online Database)



Sample Summary Statistics (n=400))
(estimates in thousands of dollars)

Mean ($\bar{x}$): 10
Standard Deviation (s): 17
Range: 0.1 to 121

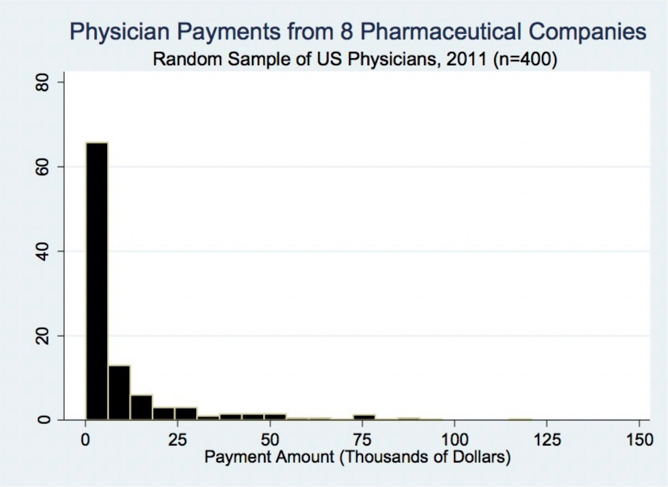
Percentiles:
2.5th: 0.3
25th: 1.5
50th: 3.6
75th: 10
97.5th: 74

Given the distribution of the sample data, what is the most likely distribution of the payment data in the population of all US physicians in 2011?

Answer: right (positively) skewed

Reasoning: The distribution of values in a sample from a larger population set should mimic (imperfectly) the distribution of values in the population. These 400 sample values have a right-skewed distribution.

7. This graph shows the payment amounts (in thousands of US$) for a **random sample** of 400 US physicians who received payments from any of 8 major pharmaceutical companies in 2011. The vertical axis represents the percent of physicians. (Source: Propublica “Dollars for Docs” Online Database)

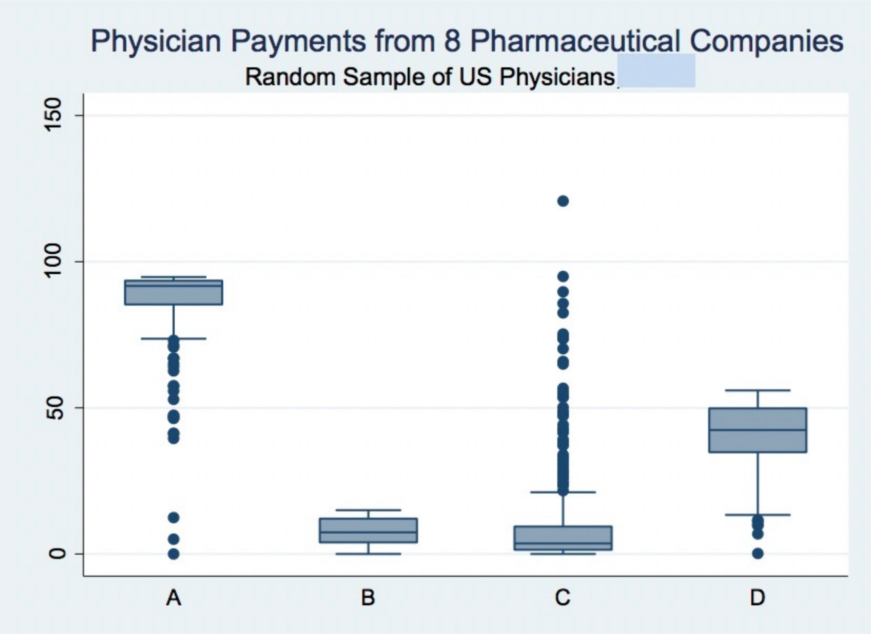


Sample Summary Statistics (n=400))
(estimates in thousands of dollars)

Mean ($\bar{x}$): 10
Standard Deviation (s): 17
Range: 0.1 to 121

Percentiles:
2.5th: 0.3
25th: 1.5
50th: 3.6
75th: 10
97.5th: 74

Which of the following boxplots corresponds to the physician payment distribution in this sample?



Answer: Boxplot C

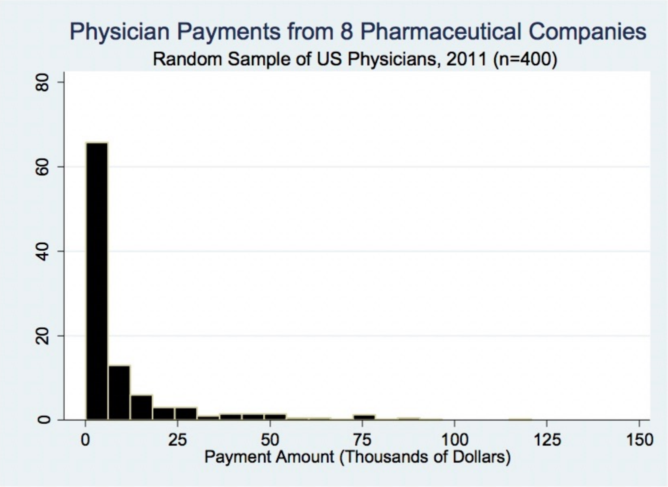
Reasoning: Both boxplots A and C evidence of a left skew, and B shows evidence of symmetry in the data distribution. Only boxplot C shows a right-skewed distribution, which is consistent with histogram. (Also, you can note that the empirical percentiles given in the table accompanying the histogram correspond to those show in boxplot C).

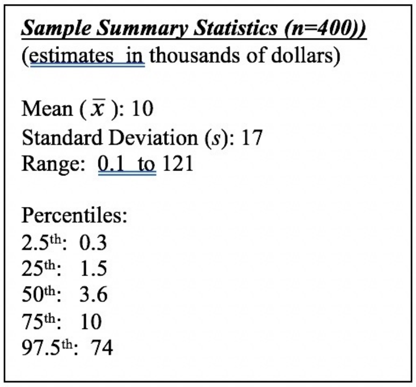
8. Suppose another researcher takes a random sample of 1000 subjects from the same population (US physicians who received payments from any of 8 major pharmaceutical companies in 2011). You have not yet seen these data. Likely, how will the sample standard deviation of these 1000 values (s_{1000}) compare to the sample standard deviation of individual physician payments in the sample of 400 (s_{400} =17)

   Answer: s_{1000} will likely be similar in value to 17 (s_{400}), but the exact relationship between these two sample standard deviation values cannot be predicted based on the information given.

   Reasoning: There is no systematic link between the expected change in the value of a sample standard deviation and the size of the sample on which the standard deviation is computed. While these estimates should be similar in value, there is no way to predict how the two estimates will exactly compare.

9. This graph shows the payment amounts (in thousands of US$), and selected sample summary statistics, for a **random sample** of 400 US physicians who received payments from any of 8 major pharmaceutical companies in 2011. The vertical axis represents the percent of physicians. (Source: Propublica "Dollars for Docs" Online Database)





Sample Summary Statistics (n=400))
(estimates in thousands of dollars)

Mean ($\bar{x}$): 10
Standard Deviation (s): 17
Range: 0.1 to 121

Percentiles:
2.5th: 0.3
25th: 1.5
50th: 3.6
75th: 10
97.5th: 74

Using these data from this sample of 400 physicians, estimate the percentage of US physicians who received payments of greater than $10,000 in 2011. This estimated percentage is:

Answer: 25

Reasoning: The 75th percentile estimated from these 400 observations is 10 ($10,000). The 75th percentile is the value that is greater than or equal to 75% of the sample observations and less than the remaining 25%. The 75th percentile estimated from this sample is the best estimate of the corresponding 75th percentile in the entire population of US physicians receiving payments from the 8 pharmaceutical companies in 2011

10. Which of the following is true about x̄, the sample mean?

Answer: The sample mean, x̄, is the best estimate of the underlying population mean, based on the sample data.