ACTIVITY 4: Worksheet Answers

The purpose of this activity is to give you an understanding of the sampling distribution of a statistic and to examine how sample size impacts the variability of the sampling distribution.

The sampling distribution of the sample mean, which is a statistic, is found by obtaining all possible samples of size *n*, and calculating the mean for each of those samples. The distribution of those sample means form the sampling distribution.

In the previous activity, you calculated three estimates for the mean life expectancy for all countries by computing the sample mean for sample sizes n=5, n=15, and n=30. Each of those sample means has a sampling distribution.

Suppose we wanted to find the sampling distribution for the sample mean that was calculated from a simple random sample of size n=5. We would have to find all possible samples of size five from our population of 195 countries. That would be 2,231,243,664 samples! Once we selected all the samples, we would have to calculate the sample mean for all of the samples and graph those 2,231,243,664 sample means using a boxplot.

Don't worry! You do not need to find the sample means for 2,231,243,664 samples; that is not reasonable! Instead, we will use the means that you calculated for the various sample size groups from the previous activity and the instructor will provide an additional 29 sample means for each sample size group. You will then create a side-by-side boxplot that will illustrate the change in variability of the mean from one sample to the next as the sample size increases.

Let's set up the Excel spreadsheet so that you can analyze your data.

- Your instructor will provide an Excel spreadsheet with three columns that contains an additional 29 sample means for each sample size group that you are going to consider.

- Input the sample means your group computed for the life expectancy in Activity 3 in the Excel spreadsheet on Line 33 under the appropriate column (n=5, n=15, and n=30).

  Note: After you include your data, there will be a total of 30 sample means in each column.

Before you get started, answer the following questions based on what you have read and completed so far.

1. Describe how the numbers on the Excel spreadsheet were obtained. Revisit Activities 2 and 3 if needed.

   A simple random sample of five, fifteen or thirty countries was selected, their life expectancies were recorded, and a mean was calculated for all the life expectancies in the sample.

2. The side-by-side boxplots you create will serve as estimates for what type of distribution? Circle the correct response and justify.

   Sampling Distribution / Population Distribution

   The data that will be used to create the boxplots are the means found for 30 simple random samples of 30 countries. Since the means are being graphed and not the values for single countries, the boxplots serve as estimates of the sampling distributions.

You are now ready to represent your data by creating boxplots. To do so, follow the directions below.

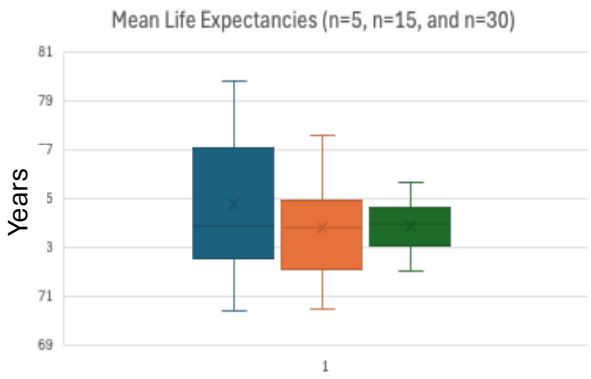
Steps for Creating Side-by-Side Boxplots:

1. Highlight all three columns of data.
2. Select "Insert" in the top menu.
3. To get the boxplot, click the icon that looks like a histogram (all bars are blue) and select the box and whisker option below the histogram.
4. A graph with three boxplots will pop up.
5. Click on the box that says "Chart Title" and create an appropriate title.
6. Change the y-axis to better show the boxplots by following the steps below.
   A. Click on the y-axis numbers to pull up "Format Axis"
   B. Click on the icon that looks like a histogram
   C. Select the "Axis Options" dropdown menu and change the bounds to 69 to 81.

Boxplots of Sampling Distributions:

Sketch your side-by-side boxplots in the space below.

Below is an example of what the students' boxplots should look like. Note: They will need to write in the units for the y-axis.



Answer the following questions based on the boxplots.

1. The side-by-side boxplots are estimates of the sampling distribution for each sample size group. They give you an idea of the variability of the sample mean from one sample to the next. Based on the boxplots, what do you notice concerning the variability as the sample size increases?

   As the sample size gets larger, the variability of the box (middle 50% of the data) decreases showing that the means become clustered more closely together.

2. From what you have observed, why is it important for a study to have a larger sample size?

   With a larger sample size, the sample mean will vary less from one sample to the next. When there is less variability in the sample mean from one sample to the next, you will obtain a good estimate of the population mean for each possible sample.

Takeaways from the Activities thus far:

In the first four activities, we looked at the concepts of a representative sample and sampling variability. When statisticians want to find an estimate for a population parameter, they take a sample and calculate a statistic to use as an estimate. Their estimate must have low variability and low bias for it to be accurate.

1. When reading a study, what should you look for to determine whether they attempted to minimize the bias?

   It would be important to determine how the sample was obtained, consider how the questions were phrased, and know the rate of non-response.

2. When reading a study, what should you look for to determine whether they attempted to reduce variability?

   It is important to look for the sample size that was used.