

Did I Trap the Median?

Sarah Parks
California State University
Sacramento
Dept. of Mathematics and Statistics

Mathew Steinwachs
University of California
Davis
iAMSTEM Hub



Rafael Diaz
California State University
Sacramento
Dept. of Mathematics and Statistics
rdiaz@csus.edu

Marco Molinaro
University of California
Davis

Published: November 2013

Overview of Lesson

In the first part of this activity, students collect sample data from their class to estimate the median foot size of the population of students in the entire class. Each student obtains a point estimate and constructs an interval estimate for the median foot size in the entire class by using a simple random sample of 10 students in the class. Numeric summaries (mean and five-number summary) and box plots are used to obtain point and interval estimates, respectively, for the median foot size of the entire class. The teacher records the foot size of all students in class in order to obtain the population median foot size of the entire class. The confidence level, or reliability, for the interval estimates computed by the students is estimated by obtaining the proportion of students' sample interval estimates that trap the population median foot size of the entire class. In the second part of this activity, students use online statistical software to explore relationships between both different sample sizes and different shapes of a population distribution, and the variability in bounds of interval estimates and the intervals' reliability to estimate the population median (level of confidence or proportion of sample interval estimates that trap the population median). Conclusions are drawn about these relationships through computer simulations using the freely available statistical software **SeeIt**.

GAISE Components

This investigation follows the four components of statistical problem solving put forth in the Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report. The four components are: formulate a question, design and implement a plan to collect data, analyze the data by measures and graphs, and interpret the results in the context of the original question. This is a GAISE Level B activity.

Common Core State Standards for Mathematical Practice

1. Make sense of problems and persevere in solving them.
2. Reason abstractly and quantitatively.
3. Construct viable arguments and critique the reasoning of others.
4. Model with mathematics.
5. Use appropriate tools strategically.

Common Core State Standard Grade Level Content (High School)

S-ID. 1. Represent data with plots on the real number line (dot plots, histograms, and box plots).

S-ID. 2. Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, standard deviation) of two or more different data sets.

S-ID. 3. Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).

S-IC. 1. Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

NCTM Principles and Standards for School Mathematics

Data Analysis and Probability Standards for Grades 9-12

Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them:

- know the characteristics of well-designed studies, including the role of randomization in surveys and experiments;
- understand the meaning of measurement data and categorical data, of univariate and bivariate data, and of the term variable;
- understand histograms, parallel box plots, and scatterplots and use them to display data;
- compute basic statistics and understand the distinction between a statistic and a parameter.

Select and use appropriate statistical methods to analyze data:

- for univariate measurement data, be able to display the distribution, describe its shape, and select and calculate summary statistics.

Develop and evaluate inferences and predictions that are based on data:

- use simulations to explore the variability of sample statistics from a known population and to construct sampling distributions;
- understand how sample statistics reflect the values of population parameters and use sampling distributions as the basis for informal inference.

Prerequisites

Students should have the ability to make measurements (in cm) using a ruler and to organize data in a table. Students should know the difference between a population and a sample. Students should know the definition of percentiles. Students should have knowledge of how to use computer spreadsheets (e.g. Excel, Google Spreadsheets).

Learning Targets

Students will have a basic understanding of how to make inferences on a population median by drawing a simple random sample from a population and constructing interval estimates for the population median. Students will construct these interval estimates by calculating numeric summaries from the sample and by constructing a boxplot. Students will also be able to determine outliers in the sample data by using some of the features of the boxplot. Finally, students will be able to draw conclusions about the relationships between both sample size and the shape of a population distribution, and the variability in bounds of interval estimates as well as the reliability (level of confidence) of these intervals.

Time Required

Two 50-minute class periods.

Materials Required

- Ruler and Pencil
- Calculators
- Computer with Internet access (Firefox or Chrome only)
- Microsoft Office Excel or another spreadsheet program (e.g. Google Spreadsheets)
- Activity sheet for data collection (see page 21)

Instructional Lesson Plan for Day 1

The GAISE Statistical Problem-Solving Procedure

I. Formulate Question(s)

Begin the lesson by explaining that this will be a two-day activity. On the first day of the activity, students will be estimating the median of a small population (their class); and on the second day, students will be estimating the population median for larger data sets from the Census at School Web Site (<http://www.amstat.org/censusatschool/>).

On the first day, students will each measure their own foot size, and then they will collect a random sample of ten of their classmates' foot sizes in order to obtain point and interval estimates for the median foot size in the entire class. Make sure you emphasize that they are to take a *random* sample of size ten (so that data is not biased due to potential confounders such as gender).

Ask students to hypothesize the answers of some questions they will be investigating:

1. What are the possible ways to find out what the median foot size of the entire class is?
2. What are the advantages of collecting a sample of only ten foot sizes and not those of the entire class to access knowledge about the median foot size in the entire class?
3. What are the advantages and disadvantages of using the sample median to estimate the population median?
4. Is there any advantage in constructing an interval estimate as opposed to a point estimate (the sample median) for the population median?
5. Is it possible to ascribe a reliability value to the interval estimate (ascribe a probability that the interval contains the median)?
6. What are the factors that may affect the length and the reliability of an interval estimate?

II. Design and Implement a Plan to Collect the Data

In the data collection phase, ask students what they can do to make their measurements consistent (measure in the same units, measure from heel to big toe, measure the same foot, etc.). Make sure that each student measures his/her own right foot first. Tell students to round measurements to the nearest $\frac{1}{2}$ centimeter. After measuring their own right foot, have students collect foot sizes from ten random students. It is imperative that students collect a random

sample of 10 of their classmates. One possible method to ensure that each student collects a random sample of ten foot sizes from the class in a streamlined fashion is as follows:

1. Number off the students from one up to the total number of students in class. For example, if there are thirty students start with one student and instruct him/her to say “one,” and then instruct the second student to say “two,” etc. Follow this pattern until the thirtieth student is reached.
2. Tell students to generate ten random numbers from one to the total number of students in class using an online random number generator such as the one at <http://www.math.usu.edu/~schneit/CTIS/RNG/>. If using this online random number generator and, assuming again that there are thirty students in the class, each student can generate ten random integers without replacement by entering the numbers 10, 1 and 30 in the first three blanks in this Web site (see Image 1 below), and selecting a bullet that says “without replacement” before clicking “Go.”

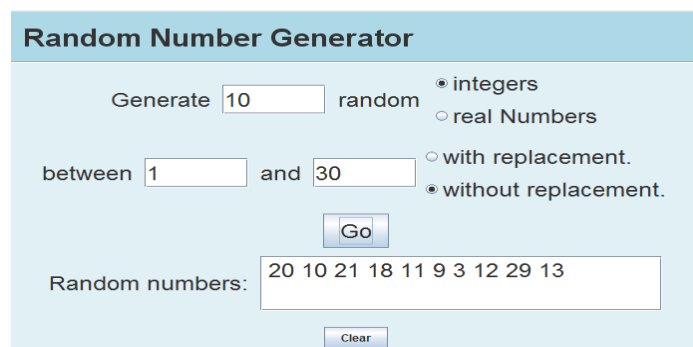


Image 1. Random Number Generator.

3. Have students write down the generated numbers from least to greatest on their data table. Explain that each number corresponds to a classmate. Next, have students in the entire class count off again, except this time, not only will they say their number but they will also say their name and their foot size. During this count off each student will be writing down only the foot sizes for each randomly generated number that corresponds to a fellow classmate. For example, Alice would say “I’m number one, Alice, size 21 cm,” and the next student will follow suit. If anyone needs foot size repeats, have students raise their hands and say which numbers they need and then have the students with the corresponding numbers say their foot sizes again. During this count off the instructor should write down the foot sizes of the entire class. This way each student gets his/her random sample and the instructor obtains the data for the whole population during one single count off.

A sample student data set is shown in the table below. A blank data table is provided on the Activity Sheet.

Name	Right Foot Size (cm)
1	24
2	24
3	25.5
4	21

5	23.5
6	24
7	24.5
8	22.5
9	24.5
10	26.5

Table 1. Example student data sample.

Record the right foot sizes (in cm) for the entire class in a data table. An example class data set is shown in the table below.

Name	Right Foot Size (cm)
1	24
2	24
3	25.5
4	21
5	23.5
6	24
7	24.5
8	22.5
9	24.5
10	26.5
11	23.5
12	23
13	25
14	22
15	21
16	25.5
17	22.5
18	25
19	22
20	23

Table 2. Example class data.

III. Analyze the Data

A. Computing and Displaying Numerical Summaries

Different statistical tools are used for estimating numerical values in a population. For example, when drawing a random sample one can calculate the sample mean (or average), or the median (50th percentile) to obtain an estimate of a “mid-value” (measure of center) of a population. Also the range, the interquartile range (difference from the 25th percentile or first quartile, to the 75th percentile or third quartile) as well as the standard deviation from sample data can be computed to estimate the spread of the values in a population (measures of variation).

A boxplot is a graphical summary for data that uses the 5-number summary (minimum, 25th percentile, median, 75th percentile, and maximum) to graphically display the distribution of a data set while highlighting measures of center (mean and median), measures of position (25th, 50th and 75th percentiles), and measures of variation (range, interquartile range). In addition to

this, a boxplot can reveal whether a data set contains any extreme points (outliers). In this lesson plan, some features of the boxplot of a sample data set will be used to construct interval estimates for the median of the population.

To construct a boxplot, students need to compute the 5-number summary of their sample data. To do this, start by asking students to order the numbers in their sample from smallest to largest. Now students can readily identify the minimum and maximum values in their sample data, and proceed to compute the quartiles. The median or second quartile (Q_2) is found by locating the midpoint of the entire ordered sample data set. Since we have an even number of data points in the example used in this lesson plan, we have two middle values so we find the median by averaging these two values. The 25th percentile or first quartile (Q_1) is found by calculating the median of the lower half of the sample data (first five numbers). For the sample data Q_1 is the sole value in the middle position (third data point) of the first five numbers. The 75th percentile or third quartile (Q_3) is found by calculating the median of the upper half of the sample data (last five numbers). In this case, this value is in the eighth position.

The steps to draw the boxplot using the sample data to construct an interval estimate for a population median can be better described by means of an example. This is done in section B below using the sample data in Table 1. In addition, the teacher should construct a boxplot for the data of the entire class for a later discussion.

B. Drawing the Boxplot and Identifying Outliers

The median right foot size for the entire class in the example of this lesson plan (Table 2) is 23.75, while for the student data sample (Table 1), it is 24.0 cm. Note that the sample median of 24 can be used as a point estimate of the population median of 23.75 cm. Point estimates are obtained with the hope that they are close to the population value that they are meant to estimate. However, point estimates without a confidence interval tend to be considered not much better than educated guesses. Confidence intervals, on the other hand, have the extra advantage that one can be quite confident of their accuracy; i.e., that the exact population value that is being estimated (the population median in this case) is captured or “trapped” by an interval constructed using sample data.

To place an interval estimate for the population median using the features of a box plot, start by having each student obtain the 5-number summary of his/her sample data as described in section A above. Notice that the smallest foot size for the sample data in Table 1 is 21.0 cm, and the largest foot size is 26.5 cm. The median value (Q_2) of 24.0 cm indicates that about half of the students in the data set have right foot sizes less than or equal to 24 cm, and that about half of the students have right foot sizes greater than or equal to 24 cm. The first quartile of the student data sample is 23.5 cm, and the third quartile is 24.5 cm. These values for Q_1 and Q_3 indicate that about 25% of the students in this sample have right foot sizes less than or equal to 23.5 cm, and about 25% of students have right foot sizes greater than or equal to 24.5 cm. These values also indicate about 50% or half of the students in the sample have right foot sizes between 23.5 and 24.5 cm.

To construct a boxplot follow these steps:

1. Mark the values of $Q_1 = 23.5$ cm, $Q_2 = 24.0$ cm, and $Q_3 = 24.5$ on a horizontal scale that spans across all the values in the sample data. Then construct a box above the scaled line using these values as indicated in Figure 1.

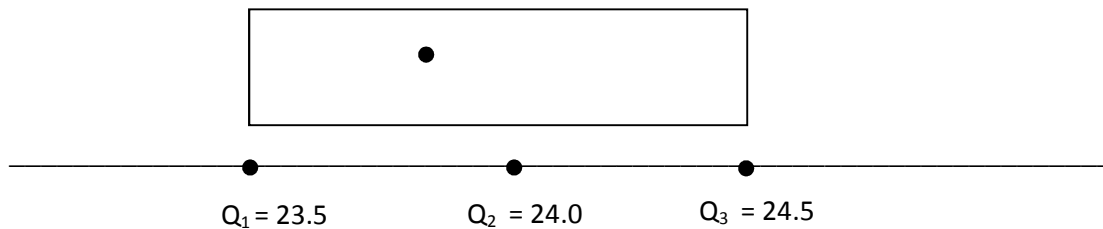


Figure 1. Boxplot: Step 1.

2. To find if there are any outliers or extreme values in a data set, compute the interquartile range (IQR), which is the difference between the third and first quartiles. Any data point beyond what are called the *lower outlier bound*, $Q_1 - 1.5(\text{IQR})$, or the *upper outlier bound*, $Q_3 + 1.5(\text{IQR})$, is considered to be an outlier. In this case $\text{IQR} = 24.5 \text{ cm} - 23.5 \text{ cm} = 1$; therefore any sample foot size smaller than $Q_1 - 1.5(\text{IQR}) = 23.5 - (1.5)(1) = 22 \text{ cm}$, or larger than $Q_3 + 1.5(\text{IQR}) = 24.5 + 1.5(1) = 26.0 \text{ cm}$ is an outlier. There are two outliers in this data set, 21.0 cm and 26.5 cm. These outliers are indicated by drawing stars above the scaled line at about half the height of the box as shown in Figure 2 below.

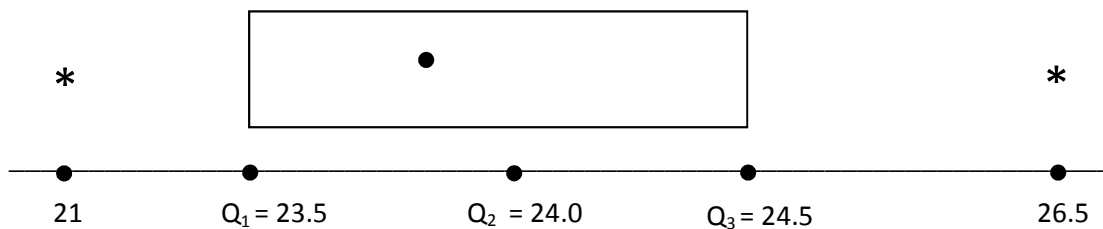


Figure 2. Boxplot: Step 2.

3. Finally, find the minimum value that is not an outlier and the maximum value that is not an outlier. Here, the minimum value that is not an outlier is 22.5, and the maximum value that is not an outlier is 25.5. Then add what are called the whiskers to the box by drawing horizontal lines at about half the height of the box, first from Q_1 down to the minimum value that is not an outlier, and second from Q_3 up to the maximum value that is not an outlier as indicated in Figure 3 below. Only when there are no outliers would the whiskers go as far as the minimum and maximum values in the data set. To avoid drawing the whiskers incorrectly, make sure to draw them after the outliers (if any) have been added to the graph.

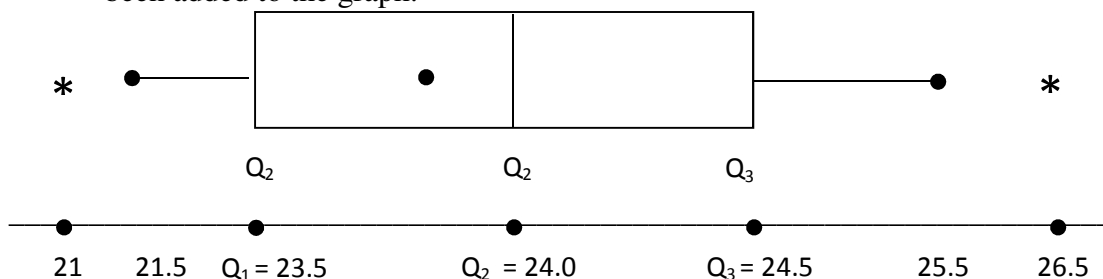


Figure 3. Boxplot: Step 3.

Notice that the boxplot for the student sample data is quite symmetric. The distributions of random sample data tend to reflect the distribution of the population. At this point you can write on the board the box-plot you obtained for the data of the entire class, and ask students if their sample data box plots resemble that of the population. There may be a small proportion of students whose boxplot may be quite different from the boxplot of the data of the entire class. This is due to random variation in the samples. However, most of the students should have a boxplot that resembles that of the population.

C. Constructing an Interval Estimate

Ask students to discuss how much their sample median differs from the population median. In the above example, the sample median of 24.0 cm is off by 0.25 cm from the population median of 23.75. Students should note the wide variability in estimation error when using their sample median as an estimate of the population median. Now ask students if they would consider it reasonable to provide an interval estimate that has a high probability of capturing or trapping the exact median of the population. If they could provide an interval that captures or traps the population median by using their own sample data, what would this interval be? One suggestion might be to use the endpoints of the whiskers of their boxplot as an interval that has a high probability of trapping the population median. However, students may also realize that this interval is too wide to help hone in on the value of the population median (that is, that this interval has a large margin of error). Then ask students whether the shorter interval from Q_1 to Q_3 (endpoints of the box instead of endpoints of the whiskers) would be more reasonable to estimate the location of the population median.

Now you can ask students how confident they are that each time they obtain a random sample of 10 students and obtain the first and third quartiles of this sample, the interval (Q_1, Q_3) captures or traps the population median. It would not be surprising to have students in class whose intervals (Q_1, Q_3) did not capture the population median. If so, this would prevent students from saying that they are 100% confident that each time they take a sample of 10 students and obtain the first and third quartiles of their sample, the interval (Q_1, Q_3) will trap the population median. So what is the level of confidence that students have for capturing the population median with the interval (Q_1, Q_3) from a random sample of 10 students? To answer this question students can obtain the reliability, or level of confidence, of using (Q_1, Q_3) from their sample of 10 students as an interval estimate for the population median: simply obtain the proportion of students in class whose interval estimate trapped the population median (class median). For example, if 15 of the 20 students (75%) in the class obtained an interval (Q_1, Q_3) that trapped the class median of 23.75 cm, then this means that each time someone takes a sample of 10 students from the class, we expect 75% of the intervals (Q_1, Q_3) will trap the population median right foot size.

IV. Interpret the Results

By now students should have an idea of the advantages of using interval estimates, which, once their level of reliability is known, are called confidence intervals. However, students may agree that a sample interval (Q_1, Q_3) is still too wide (that is, the interval has a large margin of error) as a predictor of the location of the median. Ask students questions pertaining to possible refinements for these confidence intervals such as the following, which will be explored on the second day of this lesson plan:

1. What do you think would happen to the sample interval (Q_1, Q_3) if the sample size increased from 10 to 15?
2. What do you think would happen to the sample interval (Q_1, Q_3) if the population distribution is not a symmetric distribution?
3. Do you have any idea of how to construct interval estimates that are shorter than the interval (Q_1, Q_3) ? Would a shorter interval necessarily change the level of reliability?

During the second day of this lesson plan students will see how increasing the sample size produces tighter intervals (Q_1, Q_3) . Also, they will explore how symmetric distributions produce interval estimates with smaller reliability (lower level of confidence) than non-symmetric distributions.

Assessment

A class of 25 students is selected and their IQ scores are recorded. A random sample of 10 students is taken from the class of 25 students. The data is shown in the tables below.

Class Data Table

Student	IQ Scores
1	90
2	101
3	106
4	108
5	125
6	130
7	115
8	91
9	112
10	107
11	76
12	103
13	69
14	94
15	106
16	78
17	121
18	80
19	85
20	80
21	99
22	76
23	92
24	89
25	121

Sample Data Table

Student	IQ Scores
1	96
2	101
3	106
4	108
5	125
6	130
7	115
8	93
9	112
10	107

Using the above tables answer the following questions:

- a) Calculate the 5-number summary for the **sample** data table.

- b) Determine the lower and upper outlier bounds. Are there any outliers?

- c) What are the minimum and maximum values that are not outliers? Note: if there are no outliers below (above) the lower (upper) outlier bound, then the minimum (maximum) value that is not an outlier matches the minimum (maximum) value of the data set.

- d) Construct a boxplot for foot sizes for the **sample** data table.

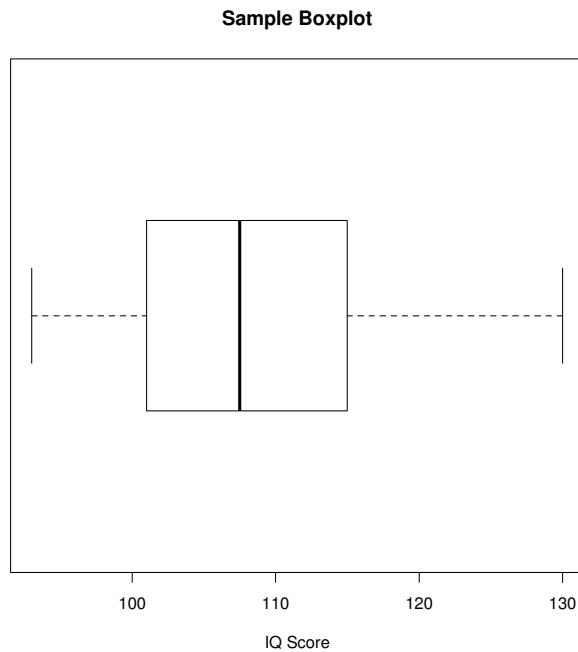
- e) Is the distribution of the data set symmetric or asymmetric?

- f) Compute the population median (median of the entire class of 25 students).

- g) Does the interval (Q_1 , Q_3) trap the median of the class data?

Answers

- a) 5-number summary: minimum = 93, first quartile, $Q_1 = 101$, median = 107.5, third quartile, $Q_3 = 115$, maximum = 130.
- b) $Q_1 = 101$, $Q_3 = 115$, $IQR = Q_3 - Q_1 = 115 - 101 = 14$;
Lower outlier bound: $Q_1 - 1.5(IQR) = 101 - 1.5(14) = 80$;
Upper outlier bound: $Q_3 + 1.5(IQR) = 115 + 1.5(14) = 136$;
No outliers (no points located beyond the outlier bounds).
- c) Minimum that is not an outlier = Minimum of the sample data (no outliers below the lower outlier bound) = 93
Maximum that is not an outlier = Maximum of the sample data (no outliers above the upper outlier bound) = 130
- d) See boxplot below:



- e) The box plot indicates that the distribution of the sample data is asymmetric due to a longer upper whisker, and a larger spread for the values between the third quartile and the median. (In the second day of this lesson plan, students will learn that when a box-plot shows an asymmetry in this direction the distribution of the data is said to be skewed right or positively skewed).
- f) The median of the entire class is 99.
- g) No, the sample boxplot does not trap the class median of 99. The class median does not fall between 101 (first quartile) and 115 (third quartile).

Instructional Lesson Plan for Day 2

The GAISE Statistical Problem-Solving Procedure

I. Formulate Question(s)

Begin day two's lesson by explaining to students that they will continue exploring properties of confidence intervals, but now they will be looking at the effect that different sample sizes and population distribution shapes have on the reliability and variation in width of confidence intervals. This exploration will require sampling a larger number of data sets (and therefore constructing a larger number of confidence intervals) than there are students in the classroom. Also the populations from which these data sets will be drawn will be much larger than the population of students in a classroom. Therefore, an online statistical software program will automate the process of drawing sample data sets and constructing confidence intervals. This statistical software can very quickly simulate the construction of a large number of confidence intervals, and it can also use these intervals to estimate their level of confidence or reliability.

The samples that students will be using here won't be samples of only size $n = 10$ as in Day 1, but they will vary for each section of this exploration: $n = 15, 30$ and 45 . The populations from which these samples will be drawn will have different distribution shapes: skewed right, skewed left, and bell-shaped (also known as a normal distribution). These population data sets were downloaded from the **U.S. Census at School** Web site (<http://www.amstat.org/censusatschool/>). Census at School is an international classroom project that engages students in grades 4-12 worldwide in statistical problem solving where students can also use data sets about themselves to perform data analysis. Students fill out a brief anonymous questionnaire about things such as their right foot size, their favorite subject, and the amount of time that it takes them to get to school. Subsequently, teachers can immediately access the data submitted by their students (as well as random samples of data submitted by students nationwide, and even worldwide) to teach students statistical data analysis.

The online statistical software program that students will be using is called **SeeIt**, which is a free, intuitive, Web-based statistical analysis tool with interactive graphical visualization features. **SeeIt** was created as part of the project "How Sure Are You?: Science, Biostatistics and Cancer Education" at the University of California Davis. In this project three Web-based teaching modules were created for high school teachers to teach students statistical analysis techniques using data-rich cancer education resources. All features of this program can be accessed freely at <http://centerforbiophotonics.github.io/SeeIt3/>.

After providing students with an introduction using the aforementioned information, hand out the *Did I Trap the Median? Simulation Worksheet: Day 2* located at the end of this lesson plan. Ask them to hypothesize what might be the answers to the following questions, and why?

1. What happens to the width of the intervals when the sample size increases? Do the bounds of the intervals vary more?
2. What happens to the level of confidence (reliability or percentage of sample intervals that trap the population median) of the interval estimate when the sample size increases?
3. What happens to the width of the interval estimate when the population distribution shape changes? Do the bounds of the intervals vary more?
4. What happens to the level of confidence (reliability or percentage of sample intervals that trap the population median) when the population distribution shape changes?

II. Design and Implement a Plan to Collect the Data

Instead of manually collecting data from a population in order to subsequently take samples and construct confidence intervals for the population median as was done on Day 1, students will be using **SeeIt** to automate the generation of a large number of sample data sets (1,000) from three large population data sets. Subsequently, they will use these sample data sets to produce a large number of confidence intervals for the population median. **SeeIt** will also compute the reliability of this large number of confidence intervals. The three population data sets that students will be using, which have different distribution shapes and were downloaded from the U.S. Census at School website as mentioned above, have already been pre-loaded into **SeeIt** expressly for this lesson plan. During this second day of the lesson plan, students will also be using different sample sizes ($n = 15, 30, \text{ and } 45$) to explore the effect that different sample sizes have on the variability of the width of confidence intervals, and on the reliability of confidence intervals.

The three population data sets are actually large samples of data that resulted from three questions answered by male and female 12th graders in 2013 from all states in the United States. These large samples of data can be treated as populations for the purposes of this exercise. These “population” data sets were carefully chosen for this lesson plan so that students analyze three different distribution shapes: skewed left, skewed right, and bell-shaped. These distributions shapes are illustrated in the histograms of Figure 4.

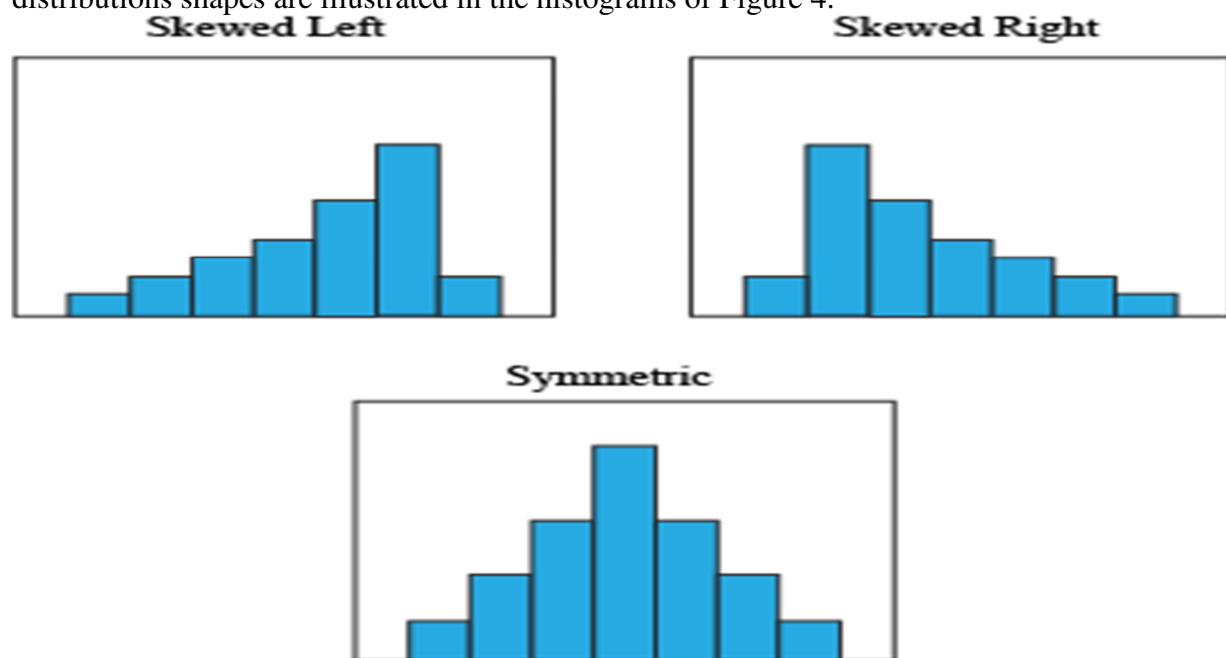


Figure 4: Population distribution shapes.

The bell-shaped data set comes from 451 students who answered the question “How many hours of sleep per night do you usually get when you have school the next day?” The skewed right data comes from 445 students who answered the question “Estimate how many hours a week you usually spend playing sports/outdoor games or activities.” Finally, the skewed left data comes from 428 students who answered the question “How important is it to you to have Internet access,” by using a scale from 0 (not important) to 1000 (very important).

After discussing the nature of the population data sets that students will be using, ask them to follow the instructions in the *Did I Trap the Median? Simulation Worksheet: Day 2* in order to open **SeeIt** directly into the link that contains the uploaded population data sets of this lesson plan. This link also contains the features of **SeeIt** that perform simulations with confidence intervals.

III. Analyze the Data

By following the instructions in the worksheet, students will summarize their results from **SeeIt** as shown in Tables 3, 4 and 5 below. Table 3 presents an example of the pattern that students should observe in the confidence levels (reliability) of the intervals (Q_1 , Q_3) when varying the sample size ($n = 15, 30$ and 45) and the shape of the distribution of the populations (skewed right, skewed left, and bell-shaped). The pattern that students should observe is that for the three population distributions, the reliability of the confidence intervals gets larger as the sample size increases. Also, the reliability is always smaller for the bell-shaped population distribution in comparison to the skewed distributions for each of the three sample sizes. Students should also note in **SeeIt** that the bounds of the confidence intervals vary more (have wider ranges) for the skewed distributions than for the bell-shaped distribution. **SeeIt** keeps track of the lower bounds (red vertical lines) and upper bounds (blue vertical lines) of the simulated confidence intervals so that students can note the variation in the ranges of the confidence interval bounds (see instructions in the *Did I Trap the Median? Simulation Worksheet: Day 2*). Tables 4 and 5, below, show an example of the pattern in ranges that students should also observe in this analysis.

Note: No two students will get the exact same answer since **SeeIt** generates different random samples for each student. However, most of the students, if not all, should get the same pattern of reliability estimates just described and presented in Table 3.

Population Distribution Shapes			
Sample Size	Bell-Shaped	Skewed Right	Skewed Left
$n = 15$	0.581	0.914	0.958
$n = 30$	0.770	0.968	0.991
$n = 45$	0.802	0.996	0.998

Table 3. Example of levels of confidence or reliability of the intervals estimates (Q_1 , Q_3) when varying sample sizes and the shape of the population distributions.

Range of Lower Bounds Q_1 's of the Intervals (Q_1, Q_3)			
Sample Size	Bell-Shaped	Skewed Right	Skewed Left
$n = 15$	From 4.0 to 6.5	From 0.0 to 5.8	From 200.0 to 800.0
$n = 30$	From 5.0 to 6.5	From 0.0 to 4.8	From 240.0 to 860.0
$n = 45$	From 5.0 to 6.8	From 0.0 to 5.0	From 260.0 to 800.0

Table 4. Example of ranges of lower bounds Q_1 's of the intervals (Q_1, Q_3) when varying sample sizes and the shape of the population distributions.

Range of Upper Bounds Q_3 's of the Intervals (Q_1, Q_3)			
Sample Size	Bell-Shaped	Skewed Right	Skewed Left
$n = 15$	From 7.0 to 9.0	From 0.6 to 30.0	From 800.0 to 1000.0
$n = 30$	From 7.0 to 8.8	From 4.8 to 22.0	From 860.0 to 1000.0
$n = 45$	From 7.0 to 8.5	From 5.0 to 16.0	From 860.0 to 1000.0

Table 5. Example of ranges of upper bounds Q_3 's of the intervals (Q_1, Q_3) when varying sample sizes and the shape of the population distributions.

IV. Interpret the Results

Finalize the lesson plan by explaining to students that there are two reasons for the pattern of confidence levels they obtained from **SeeIt** and summarized in a table (with values similar to those of Table 3):

- The level of confidence, regardless of the shape of the population distributions, gets larger as the sample size gets larger since the larger the sample size, the closer that the *sample* quartiles Q_1 and Q_3 get to *population* quartiles Q_1 and Q_3 . Since the population quartiles Q_1 and Q_3 do include the population median, sample quartiles Q_1 and Q_3 , being close to their population equivalents, will also tend to include the population median.
- Skewed distributions tend to produce sample data sets with extreme values to the left of most of the elements in the sample (when the population is skewed left), or to the right of most of the elements in the sample (when the population is skewed right). These extreme values tend to produce extreme values of Q_1 (when skewed left) or Q_3 (when skewed right) that in turn produce wider confidence intervals (Q_1, Q_3) than those of bell-shaped distributions. Wide confidence intervals have a higher probability of trapping the population median

Assessment

Open the online statistical software **SeeIt** using the link created for this assessment:

<http://centerforbiophotonics.github.io/SeeIt3/distributions-stew.html>. Click the sideways triangle next to “% Smokers” on **SeeIt**’s homepage. Using the data sets found under “% Smokers,” answer the following questions:

1. For the data set “Female% Smokers 2010,” answer the following questions:
 - a) What is the population distribution shape? Skewed right, skewed left, or bell-shaped?
 - b) Change the sample size to **10**, and find the level of confidence using the interval formula “Q1 to Q3.”
 - c) What is the range of the upper bounds of the confidence intervals for $n = 10$?
 - d) Change the sample size to **20** and find the level of confidence.
2. For the data set “Asian/Pacific Islander% Smokers 2010,” answer the following questions:
 - a) What is the population distribution shape? Skewed right, skewed left, or bell-shaped?
 - b) Change the sample size to **10**, and find the level of confidence using the interval formula “Q1 to Q3.”
 - c) What is the range of the upper bounds of the confidence intervals for $n = 10$?
 - d) Change the sample size to **20** and find the level of confidence.
3. What happens to the level of confidence as the sample size gets larger?
4. Which data set has the wider confidence interval range? Why?

Answers

(Note: Answers will not be exact, but they should be in the same range as the answers provided below.)

1. Female% Smokers 2010
 - a) Bell-shape
 - b) 0.929
 - c) From 16.0 to 25.5
 - d) 0.986
2. Asian/Pacific Islander% Smokers 2010
 - a) Skewed Right
 - b) .945
 - c) From 6.0 to 35.5
 - d) 0.981
3. Regardless of the shape of the population distribution, the level of confidence will increase as the sample size increases.
4. The data set "Asian/Pacific Islander% Smokers 2010" has the larger confidence interval range because it is skewed right. Populations that are either skewed right or skewed left contain extreme values, which produce wider confidence intervals.

Possible Extensions

A possible extension to this activity is to explore an adjustment to the fact that confidence intervals tend to get narrower as the sample size increases. **SeeIt** provides another formula to construct confidence intervals that shrink proportionately to the sample size: $Q_1 - 1.5(IQR)/\sqrt{n}$, $Q_3 + 1.5(IQR)/\sqrt{n}$. Students could use this alternate formula to produce interval estimates that are more accurate (narrower), and still have a good level of reliability.

References

1. Agresti, Alan, and Franklin, C. *Statistics: The Art and Science of Learning from Data*. Upper Saddle River, NJ: Pearson Prentice Hall, 2007. Print.
2. "The R Project for Statistical Computing." *The R Project for Statistical Computing*. N.p., n.d. Web. 15 July 2013. <<http://www.r-project.org/>>.
3. **SeeIt** [Software] (2011). Davis California: University of California Davis, iAMSTEM Hub. n.d. Web. 15 July 2013. <<https://sites.google.com/a/cbst.ucdavis.edu/sbcepublic/SeeIt>>.
4. Schneiter, Kady. "Random Number Generator." *Random Number Generator*. Utah State University, n.d. Web. 11 Aug. 2013. <<http://www.math.usu.edu/~schneit/CTIS/RNG/>>.

Acknowledgments

We thank Mary Richardson, Professor of Statistics at Grand Valley State University, for her valuable suggestions for the manuscript. We also thank Amy Smith, Project Coordinator of the NIH Science Education Partnership Award (SEPA) "How Sure Are You? Science, Biostatistics and Cancer Education" (PI; Marco Molinaro, UC Davis), for her comments in the improvement of the assessment sections of this lesson plan.

Did I Trap the Median? Boxplot Activity Sheet: Day 1

1. Describe the data collection process that will be used.
2. Record the length of your right foot in centimeters: _____
3. Record the right foot length of ten randomly chosen students in your class.

Name	Length of Right Foot (cm)

4. Arrange the foot lengths from smallest to largest.

5. Complete the table below showing numeric summaries for foot length for your ten randomly chosen classmates.

Mean	Minimum	First Quartile (Q1)	Median	Third Quartile (Q3)	Maximum

6. Determine what values would be considered to be outliers for your ten randomly chosen classmates. Are there any outliers?

7. Construct a horizontal boxplot for your ten randomly chosen classmates. In the event of having outliers for your data set, do not use outliers for the minimum or maximum values. For the minimum and maximum values, plot the minimum value that is not an outlier and the maximum value that is not an outlier.

Right Foot Length



8. What is the class median foot size? Does your Q1 to Q3 interval estimate trap the median for the entire class?

9. Based on the median of the entire class given by your instructor and the median of your particular ten randomly chosen classmates, calculate what proportion (percent) of box plots trap the median for the entire class. This is the reliability (confidence level) of using interval estimates from Q1 to Q3.

10. Think about what would happen if the sample size increased. Would the proportion of box plots that would trap the median increase or decrease? Why?

Did I Trap the Median? Simulation Worksheet: Day 2

1. Hypothesize what the answers to the following questions might be and state why.
 - a) What happens to the width of the confidence intervals when the sample size increases? Do the bounds of the intervals vary more? Why?

 - b) What happens to the level of confidence (reliability or percentage of sample intervals that trap the population median) of the interval estimate when the sample size increases? Why?

 - c) What happens to the width of the interval estimate when the population distribution shape changes? Do the bounds of the intervals vary more? Why?

 - d) What happens to the level of confidence (reliability or percentage of sample intervals that trap the population median) when the population distribution shape changes? Why?

2. **SeeIt** Setup:

- a) Open up **SeeIt** on your computer with either Google Chrome or Firefox and use the following link created for this lesson plan:

<http://centerforbiophotonics.github.io/SeeIt3/distributions-stew.html>

- b) Click on the triangle next to “Population Data.” The names of the population data sets needed for this lesson plan will appear as though they were documents in a folder called “Population Data.”

3. You’re going to analyze the set of data titled “Schoolnight Sleep Hours” first.

- a) Click and drag the title “Schoolnight Sleep Hours” into the space to the right that has the wrench icon. This data set can be thought of as a population of $N = 451$ students whose average amount of sleeping time was recorded. A histogram of this population will appear in the space with the wrench icon, and a histogram of a random sample (of size $n = 15$ by default) from this population will appear below this space.

- What is the shape of the distribution of this population? Skewed right, skewed left, or bell-shaped?

- Does the histogram of the random sample reflect the distribution of the population?

- b) Create a boxplot for the population data and the random sample by following these steps:

Note: If you click in the wrong place, and you don’t see things as they are described here, you can always reload the **SeeIt** Web page and start over again.

- i. Click the wrench icon. A pop-up box will appear.
- ii. In this pop-up box click on the bullets next to “Box Plot” (to change the histograms into box plots), “Advanced” (to display box plots that highlight outliers), and “Median” (to place a pin that “pins down” the median on the x-axis scale), so that the bullets have a check mark.
- iii. Close the pop-up box by clicking “X” on the upper right corner. The box plots will highlight the mean or average with a star, outlier(s) (if any) with “+” signs, and the population median with a round headed pin in the box plot of the population, and a square headed pin in the box plot of the sample.
- iv. You can find out the exact value of the population and sample medians by placing the cursor on each of the heads of the pins so that a small pop-up rectangle will display these values. You can gauge the rest of the values of the five-number summary and the mean from the scale located below the box plot.

- Does the boxplot provide the same information about the distribution of the population as the histogram? Are the whiskers of the box plot symmetric or is one of them longer than the other?
- What are the values of the five-number summary, and the mean of the population? Write these numbers in the following table:

Population ($N = 451$)	Mean	Minimum	First Quartile (Q_1)	Median	Third Quartile (Q_3)	Maximum
Schoolnight Sleep Hours.						

- Are there any outliers?

- What are the values of the five-number summary and the mean of this sample?

Sample ($n = 15$)	Mean	Minimum	First Quartile (Q_1)	Median	Third Quartile (Q_1)	Maximum
Schoolnight Sleep Hours.						

- Are there any outliers?
- Does the *sample* interval (Q_1, Q_3) contain or trap the *population* median?

c) Now change the default sample size of $n = 15$ to $n = 30$ in the space that has the sample box plot and click the button that says “sample.”

- What are the values of Q_1 and Q_3 of this new sample?
- Is the width of the interval (Q_1, Q_3) produced by this sample size narrower than that produced by the previous sample of size $n = 15$?

Change the sample size once again, as you did above, this time to $n = 45$.

- What are the values of Q_1 and Q_3 of this new sample?
 - Is the width of the interval (Q_1, Q_3) produced by this sample size narrower than that of that produced by the previous sample of sizes of $n = 15$ and $n = 30$?
- d) Change the sample size back to $n = 15$, but this time click the button “Sample” several times to obtain several samples of size 15 and their box plots. Note that **SeeIt** keeps track of the bounds of these confidence intervals in the space below the sample box plots: the lower bounds Q_1 are recorded with red vertical lines, and the upper bounds Q_3 are recorded with a blue vertical line. When the samples produce repeated values of Q_1 and Q_3 , **SeeIt** keeps track of these values by making the red and blue vertical lines darker. **SeeIt** also keeps track of the number of these confidence intervals that trap the population median: look under “Confidence Intervals” for the sentence that reads “The proportion of intervals that trap the population median after taking x samples is:”

You are now ready to make **SeeIt** automatically repeat the process of clicking the “Sample” button 1000 times. To do this, click the button “Sample All” in the space that keeps track of the bounds of the confidence intervals. Look again at the end of the sentence “The proportion of intervals that trap the population median after taking 1000 samples is:...” This is a much more accurate estimate of the level of confidence or reliability of the interval estimates than that obtained using only a few samples. Record this number in the cell that corresponds to “ $n = 15$ ” and “Skewed Left” of the Summary Table in question 6 of this worksheet. Also record the range of the 1000 lower bounds (red vertical lines) and upper bounds (blue vertical lines) in the second and third summary tables in question 6 of this worksheet. Now repeat this process for samples of sizes 30 and 45.

- e) Finally, click on the “X” on the upper-right corner in the space that contains the population box plot to erase this analysis (“Schoolnight Sleep Hours”) in order to prepare for the analysis of the next data set.

4. Now you're going to analyze the data set "Outdoor Activities" by repeating steps "a" through "e" from question "3," above, clicking now on "Outdoor Activities."
 5. Lastly, you're going to analyze the set of data called "Internet Access" following these same steps once again.
-
6. Write the levels of confidence and ranges of bounds you obtained for the intervals (Q_1 , Q_3) in questions 3, 4 and 5 above.

Level of Confidence or Reliability of the Intervals (Q_1, Q_3)			
	Bell-Shaped	Skewed Right	Skewed Left
$n = 15$			
$n = 30$			
$n = 45$			

Range of Lower Bounds Q_1's of the Intervals (Q_1, Q_3)			
	Bell-Shaped	Skewed Right	Skewed Left
$n = 15$	From to	From to	From to
$n = 30$	From to	From to	From to
$n = 45$	From to	From to	From to

Range of Upper Bounds Q_3's of the Intervals (Q_1, Q_3)			
	Bell-Shaped	Skewed Right	Skewed Left
$n = 15$	From to	From to	From to
$n = 30$	From to	From to	From to
$n = 45$	From to	From to	From to

- a) Within the bell-shaped population, which sample size had the largest level of confidence? Smallest level of confidence? Why?

- b) Within the skewed right population, which sample size had the largest level of confidence? Smallest level of confidence?

- c) Within the skewed left population, which sample size had the largest level of confidence? Smallest level of confidence?
- d) Now look at each of the population shape distributions as a whole. Overall, which shape traps the median the most? Which shape traps the median the least? Why do you think that is?
- e) Within the bell-shaped population, which sample size had the largest level of variability (range) in the lower and upper bounds of the confidence intervals? The smallest? Why?
- f) Within the skewed right population, which sample size had the largest level of variability (range) in the lower and upper bounds of the confidence intervals? The smallest? Why?
- g) Within the skewed left population, which sample size had the largest level of variability (range) in the lower and upper bounds of the confidence intervals? The smallest? Why?
- h) Now look at each of the population shape distributions as a whole. Overall, which shape has the largest variability in the lower and upper bounds of the confidence intervals? Why?

SeeIt Guide

To add data to an Excel Sheet:

- 1) Open Excel.
- 2) Label first column as “Label.”
- 3) Type in numbers “1” through “10” (or label until you reach your last data point) in the first column.
- 4) Label second column as “Data.”
- 5) Type in your data in second column.

To add a set of data from an Excel sheet:

- 1) Click the “+” next to “Add a Worksheet.”
- 2) Name your Worksheet in box next to “Title:” if you labeled your columns.
- 3) Click box next to “First Column is Label:” if you labeled your columns.
- 4) In your Excel worksheet, highlight all data in both columns using **Shift** key.
- 5) **Ctrl + Copy** data in Excel worksheet.
- 6) **Ctrl + Paste** data from Excel into **SeeIt** textbox.
- 7) Click “Load Worksheet From Form.”