Now You SeeIt, Now You Don't: Using SeeIt to Compare Stacked Dotplots to Boxplots

STatistics Education Web
STEW
online journal of K-12 statistics lesson plans

Alberto Guzman-Alvarez
University of California Davis,
iAMSTEM Hub

Amy Falk Smith
University of California Davis,
iAMSTEM Hub

Marco Molinaro
University of California Davis,
iAMSTEM Hub

Rafael Diaz
California State University
Sacramento
rdiaz@csus.edu

Published: May 2014

Overview of Lesson

In the first part of the lesson students will collect a dataset set by measuring the height of their right-hand reach. Then students will learn how to enter the data online into the freely available statistical software *SeeIt* in order to visualize their data using a stacked dotplot (all *SeeIt* visualizations used in this lesson are available in Fathom as well; therefore, teachers already using Fathom can simply highlight and copy the datasets used in this lesson from the "*SeeIt* edit worksheet" and import them into Fathom).

For general help with *SeeIt*, follow this link to the *SeeIt* help page: http://centerforbiophotonics.github.io/SeeIt3/disthelp.html. Students will note that stacked dotplots allow them to see the individual elements of a dataset. Subsequently, students will learn to identify stacked dotplots of five common distributions: Normal or Bell-Shaped, Skewed Right, Skewed Left, Flat or Uniform, and Bi-Modal.

In the second part of the lesson, students will use *SeeIt* to organize the individual dots on the stacked dotplot of their right-hand reach, in a way that leads to the creation of a boxplot. Students will note that boxplots provide an aggregate-view of their data (individual elements in the dataset cannot be seen). More specifically, students will use a special feature of *SeeIt* to divide the dots in their stacked dotplot into four equal parts, and obtain the five-number summary (minimum, 25th percentile or first quartile (Q1), median or second quartile (Q2), 75th percentile or third quartile (Q3), and maximum) of their data. Then students will use *SeeIt* to superimpose a boxplot on their stacked dotplot. Students will repeat this gradual data visualization from stacked dotplots to boxplots using datasets that have the five distributions described in the first part of the lesson, and that have already been preloaded in *SeeIt*. This part of the lesson shows students how to recognize a boxplot from a stacked dotplot, and vice versa, for all five distributions. The lesson concludes by showing students that even though the individual elements of a dataset cannot be seen in a boxplot, this popular type of graph provides, in many instances, several advantages over stacked dotplots.

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 Standards Grade Level Content (High School)

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

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:

- 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.

Prerequisites

Students should have the ability to take measurements using a ruler as well as construct a dataset table. Students should have knowledge of how to use computer spreadsheets (e.g. Excel, Google Spreadsheets).

Learning Targets

Students will be able to enhance their individual-case view with an aggregate-view of data by using a free online visualization tool that allows them to gradually transition from stacked dotplots (individual-case view) to boxplots (aggregate-view). This gradual transition is based on the research of Bakker et al (2004), who studied the difficulties that young students have when recognizing patterns of datasets in the commonly used boxplots. Students will also be instructed on the advantages that, in many instances, boxplots have over stacked dotplots. This activity also intends to introduce students to a not-so-trivial, but important statistical analysis skill, which consists of recognizing statistical features of datasets as properties of their aggregated, and not of their individual, elements.

Time Required

One and a half 50-minute class periods.

Materials Required

- Activity Sheets (at the end of this lesson)
- Printable Meter Ruler
- Computer with Internet access (Firefox, Safari, or Chrome only)
- Microsoft Office Excel or another spreadsheet program (e.g. Google Spreadsheets)

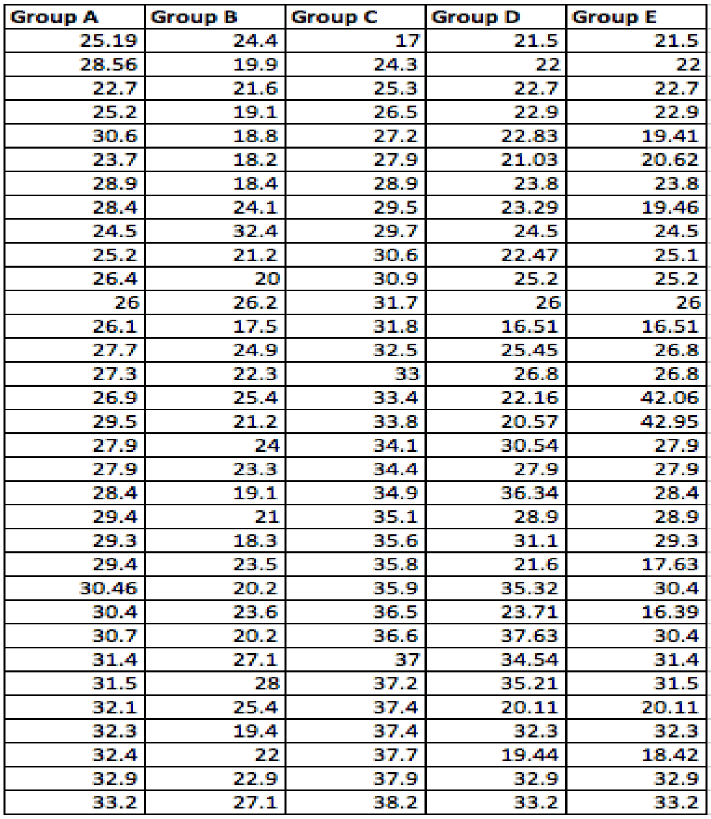
Instructional Lesson Plan for Day 1

The GAISE Statistical Problem-Solving Procedure

I. Formulate Question(s)

Begin the lesson by explaining that this will be a two-day activity. On day one, students will be learning about the importance of representing datasets in graphs in order to obtain more information than that obtained by simply looking at a list of the dataset. Students will learn that the distribution, or shape, of a stacked dotplot, can provide meaningful information about a dataset as they learn to identify five common distribution shapes: Bell-Shaped, Skewed Right, Skewed Left, Flat or Uniform, and Bi-Modal. Subsequently, students will create their own dataset by collecting the right-hand reach measurements of the entire class and entering these measurements into *SeeIt*. Finally, students will learn how to create a stacked dotplot in *SeeIt* in order obtain information from their data by highlighting features of its distribution.

Proceed to show Table 1 to students. This table includes the scores from a physical fitness test taken by five different hypothetical groups of middle school and high school students. The test was developed by the Marine Corps, and involved challenging activities including running an obstacle course, climbing a rope, sit-ups, and push-ups. The maximum possible score was 50 points.



Group A	Group B	Group C	Group D	Group E
25.19	24.4	17	21.5	21.5
28.56	19.9	24.3	22	22
22.7	21.6	25.3	22.7	22.7
25.2	19.1	26.5	22.9	22.9
30.6	18.8	27.2	22.83	19.41
23.7	18.2	27.9	21.03	20.62
28.9	18.4	28.9	23.8	23.8
28.4	24.1	29.5	23.29	19.46
24.5	32.4	29.7	24.5	24.5
25.2	21.2	30.6	22.47	25.1
26.4	20	30.9	25.2	25.2
26	26.2	31.7	26	26
26.1	17.5	31.8	16.51	16.51
27.7	24.9	32.5	25.45	26.8
27.3	22.3	33	26.8	26.8
26.9	25.4	33.4	22.16	42.06
29.5	21.2	33.8	20.57	42.95
27.9	24	34.1	30.54	27.9
27.9	23.3	34.4	27.9	27.9
28.4	19.1	34.9	36.34	28.4
29.4	21	35.1	28.9	28.9
29.3	18.3	35.6	31.1	29.3
29.4	23.5	35.8	21.6	17.63
30.46	20.2	35.9	35.32	30.4
30.4	23.6	36.5	23.71	16.39
30.7	20.2	36.6	37.63	30.4
31.4	27.1	37	34.54	31.4
31.5	28	37.2	35.21	31.5
32.1	25.4	37.4	20.11	20.11
32.3	19.4	37.4	32.3	32.3
32.4	22	37.7	19.44	18.42
32.9	22.9	37.9	32.9	32.9
33.2	27.1	38.2	33.2	33.2

Table 1. Physical fitness scores for five hypothetical group of students.

Ask students about what they can tell from the data table. Explain to students that looking at a long list of numbers may not be very useful in terms of extracting meaningful information about the data. Ask students to think of a way to visualize the data in this table in a more organized and informative fashion. The expected answer is by using *graphs*. Now ask students which graph(s) they can think of to organize the data from this table? Answers will vary, but continue to introduce a basic graph called a stacked dotplot using Figure 1.

When these data are graphed into stacked dotplots, the graphs of each group of students will look like those in Figure 1. Students should be told that each of these graphs represent the *distribution* of scores of each group of students, with each dot representing one individual score. The more dots that stack over a particular score, the more people that received that score. Now ask students about any additional information they gain when the data are graphed as a stacked dotplot and the distribution can be visualized.

Briefly discuss that one of the basic features of a distribution is its shape. Figure 1 displays five common shapes of distributions, and it is followed by a discussion of the information that these distributions may provide about data in general (students will then use the discussion below to answer questions in terms of the physical fitness scores).

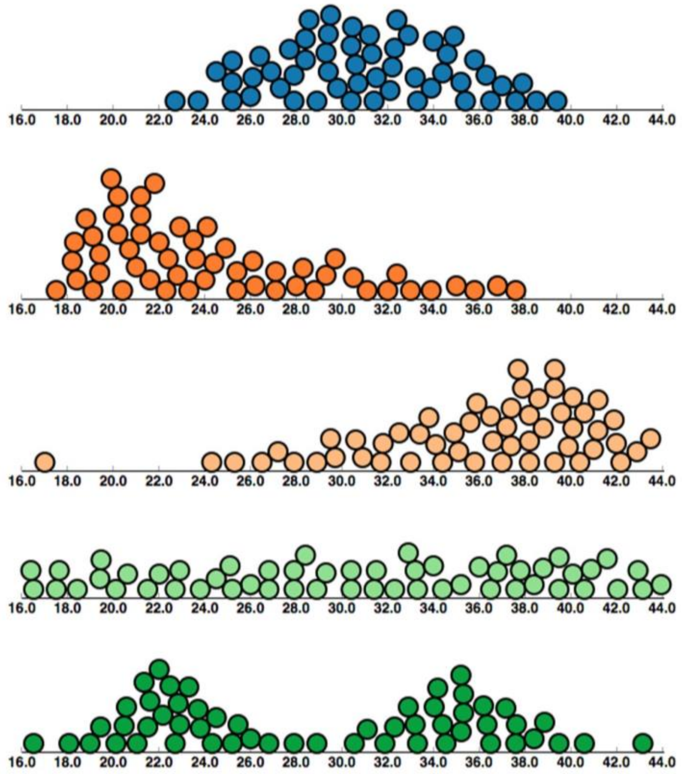


Figure 1. Stacked dotplots of common distribution shapes.

A. **Bell-Shaped or Normal:** symmetric data that are concentrated in the middle, with equal tails on either side. Many variables that occur naturally have bell-shaped distributions, e.g., height and weight of students in a high school.

B. **Skewed Right:** most of the data are concentrated on the *left*, with a tail trailing off to the *right*. Note that the word “right” on skewed right refers to the side of the **tail**, not to the side where the majority of the data are concentrated. Skewed right data can occur, for example, with test scores on a very difficult test, where most people do very poorly and a few high scores constitute the tail on the right of the distribution.

C. **Skewed Left:** most of the data are concentrated on the *right*, with a **tail** trailing to the *left*. A skewed left distribution can occur, for example, with test scores from an easy test, where most students receive a high score and a few low scores constitute the tail on the left of the distribution .

D. **Uniform or Flat:** the data are evenly spread out over the distribution. A uniform/flat distribution occurs when the proportion of elements in the dataset is the same over intervals with the same length. This distribution can occur when plotting the numbers that come up when tossing a fair die a large number of times.

E. **Bi-Modal:** the data is concentrated on both the right and the left, creating two distinct peaks. Bimodal distributions can indicate that a dataset contains two subgroups that may need to be graphed and analyzed separately. For example, if the age of the people that use a gym in a high school is collected, the graph of the stacked dotplot may produce two distinct peaks. One peak would belong to the distribution of the ages of the students, and the other peak would show the distribution of the ages of the teachers/adults.

Ask students to go to Question 1 of the Activity Sheet and interpret the shape of the distributions of the performance of each of the groups of students (Figure 1) based on the explanations provided in A-E above.

Next ask students to go to Question 2 of the Activity Sheet and write an advantage and a disadvantage of presenting the data in stacked dotplots as opposed to a table. Then make a list on the board with the different answers of the students. Some possible answers are provided below.

Advantages:

- Can see data in an organized fashion.
- Can see how data are distributed over the range of values.
- In the case of stacked dotplots, can easily identify minimum and maximum values.
- Can easily identify the presence of extreme values.
- Can get a sense of where most of the values are concentrated (hill) and the spread of the whole data (range).
- Can identify the shape of the distribution of the data: symmetry, skewedness, etc. and surmise possible reasons for the shape of the distribution.

Disadvantages:

- May not be easy to know exact value of all points.
- Not easy to identify summary values of the data, such as mean, median, percentiles, etc.

Explain to students that graphs offer a "big picture view." With graphs you often lose individual data point accuracy, but gain an idea of the overall shape and pattern of the data.

II. Design and Implement a Plan to Collect the Data

Tell students that they will practice representing and interpreting data using stacked dotplots by creating and graphing their own dataset. They will create this dataset by measuring the hand reach of all students in the class. For the data collection phase, students will step up to a paper ruler taped on the wall and measure the height of their right hand reach to the nearest cm.
You can download a 2-meter printable ruler here:
http://www.vendian.org/mncharity/dir3/paper_rulers/UnstableURL/meterstick_cm_in_2meters.pdf

As students step up to the ruler one by one, they can either write their right hand reach on a table on the board, or read it out loud for the students in class to record it. To streamline the creation of the dataset in *SeeIt*, students can enter the measurements into a worksheet in *SeeIt* as each student measures his/her reach. The steps of the data collection process may be as follows:

Tape to the wall a paper ruler that covers at least the possible range of hand reach of students in your class (the ruler does not need to have markings starting from the floor, but start, say, at 150 cm, and finish, say, at 250 cm).

Draw on the board a table with two columns with the titles “Name” and “Reach” where students will be writing their name and their right-hand reach.

III. Analyze the Data

Ask students to use their computer to go into *SeeIt* at:

http://goo.gl/lkU7Ek

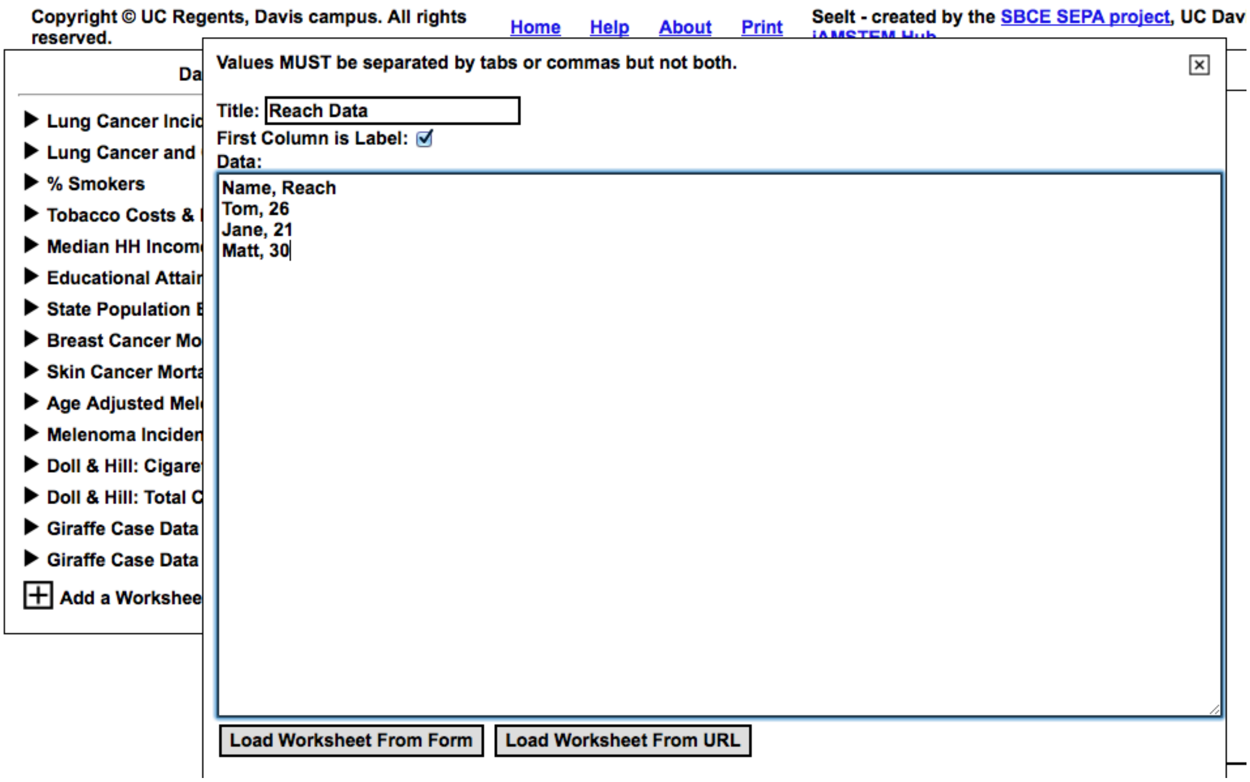
(The help page of SeeIt is located at: http://centerforbiophotonics.github.io/SeeIt3/disthelp.html)

Note: If students don’t have access to a computer in the classroom, the teacher can record the data in a spreadsheet in his/her computer, and provide it to the students to analyze in *SeeIt* once students have access to a computer.

Once students have access to a computer ask them to click on “Add a Worksheet” at the bottom of the datasets panel located on the left of the screen in order to display a *SeeIt* edit worksheet. Ask students to add a title to their dataset in the rectangle that says ***Enter Worksheet Title***, and click on the box to the right of “First Column is Label” at the top of the edit worksheet. Students must follow the following format when entering the dataset into the worksheet (see Figure 2):

First row must have the Labels of the two columns separated by a comma and followed by a space, e.g., Name, Reach.

All rows after the first one need to have the name of the student (one single word) followed by a comma and separated by a space and the right-hand reach of the student, to the nearest cm (see Figure 2 below).



Values MUST be separated by tabs or commas but not both.

Title: Reach Data

First Column is Label: ☑

Data:

```
Name, Reach
Tom, 26
Jane, 21
Matt, 30
```

Figure 2. Format to enter data in an edit worksheet in *SeeIt*.

Have students step to the ruler one by one and reach as high as they can with their right hand. Record name and reach on the table on the board (you can also enter the data in your own *SeeIt* worksheet while you project the screen of your computer for the rest of the class). Students in class will be entering the data into *SeeIt* as each student steps up to measure their reach. If students are working in a computer lab in which the cookies are cleared after logging off, have students keep a copy of the classroom data in an Excel/Google Spreadsheet (select the dataset in the *SeeIt* worksheet, copy it, and then paste it into an Excel/Google Documents spreadsheet), then save the spreadsheet on a USB, or email to themselves. Students can then copy the dataset from the Excel/Google Documents spreadsheet and paste into a *SeeIt* worksheet on day two of this activity.

Once all data have been entered, have students click on the rectangle that says “Load Worksheet from Form” that is located at the bottom of the edit worksheet.

The class reach data will now be loaded onto *SeeIt* and should appear under the title the dataset was given on the list of datasets located on the left of the screen in *SeeIt*. Ask students to click on the title of their dataset in order to display the name of the variable of their measurements (“Reach Data” in this case). Note: If the dataset needs editing, students can click on the pencil icon below the title of the dataset so that the *SeeIt* worksheet with the dataset appears again. Students should click on “Reach Data” and drag it over to the window to the right in *SeeIt*. *SeeIt* will produce a stacked dot-plot distribution of the classroom reach data similar to the graph shown in Figure 3. Ask students to click on “Display Options” and increase the dot size to better view the stacked dot plot.

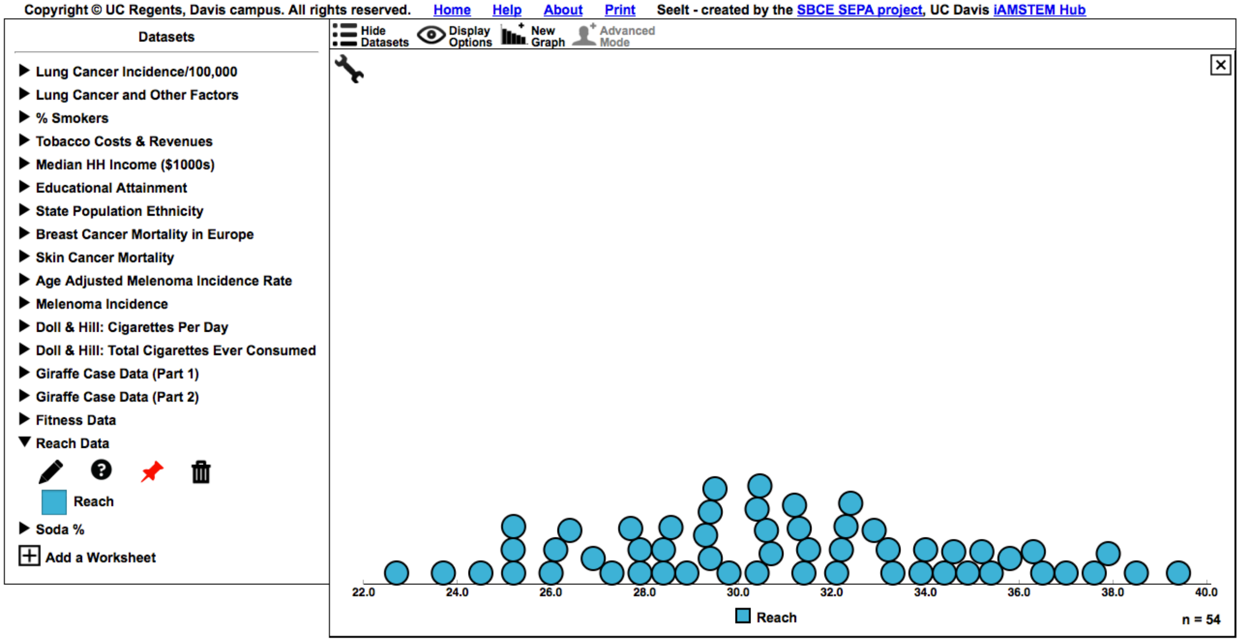

Copyright © UC Regents, Davis campus. All rights reserved.
Home
Help
About
Print
SeeIt - created by the SBCE SEPA project, UC Davis iAMSTEM Hub
Datasets
Hide Datasets
Display Options
New Graph
Advanced Mode
Lung Cancer Incidence/100,000
Lung Cancer and Other Factors
% Smokers
Tobacco Costs & Revenues
Median HH Income ($1000s)
Educational Attainment
State Population Ethnicity
Breast Cancer Mortality in Europe
Skin Cancer Mortality
Age Adjusted Melenoma Incidence Rate
Melenoma Incidence
Doll & Hill: Cigarettes Per Day
Doll & Hill: Total Cigarettes Ever Consumed
Giraffe Case Data (Part 1)
Giraffe Case Data (Part 2)
Fitness Data
Reach Data
Reach
Soda %
Add a Worksheet
22.0
24.0
26.0
28.0
30.0
32.0
34.0
36.0
38.0
40.0
Reach
n = 54

Figure 3. Example of a stacked dotplot in *SeeIt*.

Ask students to go to Question 3 on the Activity Sheet to answer questions about the information they obtain about their hand reaches when visualizing their measurements with a stacked dotplot.

Students can also continue practicing stacked dotplots in *SeeIt* by displaying the data for the various distributions described at the beginning of the lesson (Normal Distribution, Skewed Right, Skewed Left, Uniform, and Bi-Modal). To do this, students should click on the dataset with title “Fitness Data” in order to display the data for each individual group. To display multiple graphs (as in Figure 4) stacked on top of each other students can click on “New Graph” in the bar at the top of the graphing window.

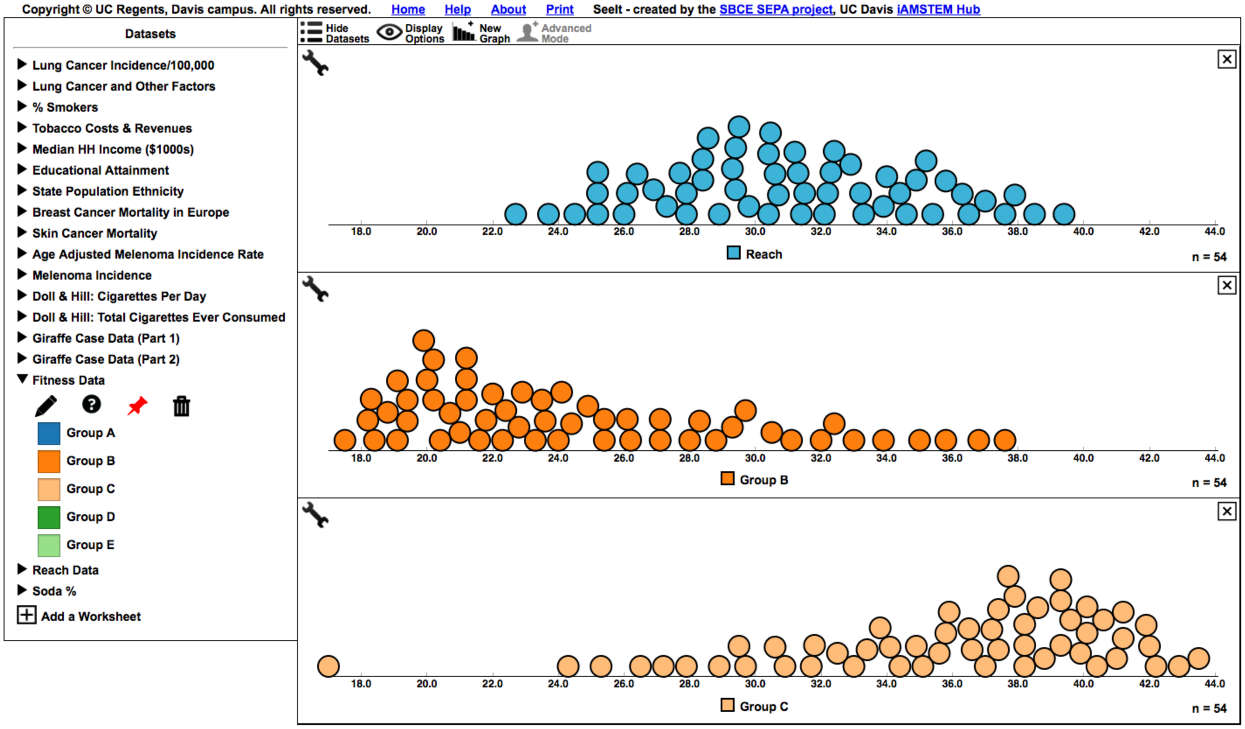

Copyright © UC Regents, Davis campus. All rights reserved.
Home
Help
About
Print
SeeIt - created by the SBCE SEPA project, UC Davis iAMSTEM Hub
Datasets
Lung Cancer Incidence/100,000
Lung Cancer and Other Factors
% Smokers
Tobacco Costs & Revenues
Median HH Income ($1000s)
Educational Attainment
State Population Ethnicity
Breast Cancer Mortality in Europe
Skin Cancer Mortality
Age Adjusted Melenoma Incidence Rate
Melenoma Incidence
Doll & Hill: Cigarettes Per Day
Doll & Hill: Total Cigarettes Ever Consumed
Giraffe Case Data (Part 1)
Giraffe Case Data (Part 2)
Fitness Data
Group A
Group B
Group C
Group D
Group E
Reach Data
Soda %
Add a Worksheet
Hide Datasets
Display Options
New Graph
Advanced Mode
18.0
20.0
22.0
24.0
26.0
28.0
30.0
32.0
34.0
36.0
38.0
40.0
42.0
44.0
Reach
n = 54
Group B
n = 54
Group C
n = 54

Figure 4. Multiple stacked dotplots in *SeeIt.*

IV. Interpret the Results

Ask students to interpret their hand reach data in terms of all the information that can be extracted from a stacked dotplot by answering Question 4 on the Activity Sheet.

Assessment

Cans of a particular soda are supposed to be automatically filled at 98.2% of their capacity. When you purchase 16 cans of this type of soda, and measure the percentage of liquid these cans contain, you obtain the following measurements:

Can Number	Percentage of Soda
1	98.16
2	98.32
3	97.83
4	98.26
5	98.24
6	97.96
7	98.21
8	98.16
9	98.20
10	98.31
11	98.30
12	98.26
13	98.08
14	98.11
15	98.26
16	98.02

If it is very important that the cans be filled quite accurately, answer the following questions based on the data collected.

1. Can you tell by looking at the table above if the machine that automatically fills these cans did a good job? Or, did the machine tend to overfill or under fill these cans?

2. If you obtained a stacked dotplot of this data, what type of distribution would you expect to see if the machine is doing a good job? i.e., the machine neither tended to overfill or under fill these cans?

3. Access *SeeIt* at:

http://goo.gl/lkU7Ek

Click on the title of the dataset called "Soda %" located in the list on the left of the screen. This will show the name of the only variable in this dataset called "Percentage of Soda"

Use the cursor to click on "Percentage of Soda" and drag this variable name into the graphing window on the left of the screen.

Click where it says "Display Options" on the top bar, then use the "+"/ "-" signs to increase/decrease the size of the data points in the plot for a better view of the stacked dotplot.

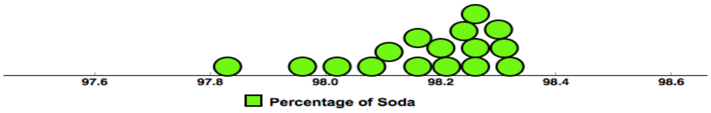
4. When looking at the stacked dotplot, do your answers about the performance of this machine change from what you answered in Part (a) above?

Answers

1. Difficult to determine the performance of the machine, since data are not organized. A graph would help.

2. Either a flat (uniform) or a bell-shaped distribution centered around 98.2.

3. From *SeeIt*:



4. Answer varies. However, students may tend to say that the machine is doing a good job when just looking at the table, since most of the numbers in the table start with 98. A look at the stacked dotplot indicates that the machine tended to underfill the cans.

Instructional Lesson Plan for Day 2

The GAISE Statistical Problem-Solving Procedure

I. Formulate Question(s)

On day two, students will learn how to organize the individual dots on the stacked dotplot of their data by using a special feature in *SeeIt* that leads to the creation of a boxplot. *SeeIt* allows students to evenly divide the data into four parts so students can visualize the density of the data between each quartile and obtain the five-number summary. Studies have shown that a main difficulty in learning and teaching boxplots is that boxplots reduce a dataset into just five key values, and the individual values can no longer be visualized (Bakker et al, 2004; Pfannkuch, 2006). Another issue with learning and teaching boxplots is that, in comparison to other graphical representations, such as histograms, boxplots do not display frequencies but rather densities (Bakker et al, 2004; Pfannkuch 2006). *SeeIt* in agreement with the conclusions from these studies, introduces boxplots to students by using a combination of dotplots and quartiles, and presents dotplots and boxplots at the same time. By the end of this lesson students should be able to look at a stacked dotplot and visualize its corresponding boxplot and vice versa.

To begin day two, ask students if there are features of a dataset they wish they could visualize when looking at a graph but that they cannot visualize when looking at a stacked dotplot. Answers that will lead to the next section of the lesson are: mean, median, quartiles, or more comprehensively, the five-number summary (minimum, 25th percentile or first quartile (Q1), median or second quartile (Q2), 75th percentile or third quartile (Q3), and maximum).

Tell students that now they will learn to highlight the five-number summary of a stacked dotplot in *SeeIt*. Tell them that they will do so in *SeeIt* by dividing the data points into four equal parts.

Ask student to use their computer to go into *SeeIt* at:
http://goo.gl/lkU7Ek

Students will have to upload their data back into *SeeIt* in the same manner as day one. This process can be facilitated by having students copy their reach data from *SeeIt*, and paste it in an Excel or Google Spreadsheet on day one of the lesson so that students can simply copy and paste the data back into *SeeIt* on day two.

The class reach data will now be loaded into *SeeIt* and should appear under the title the dataset was given on the list of datasets located on the left of the screen in *SeeIt.*

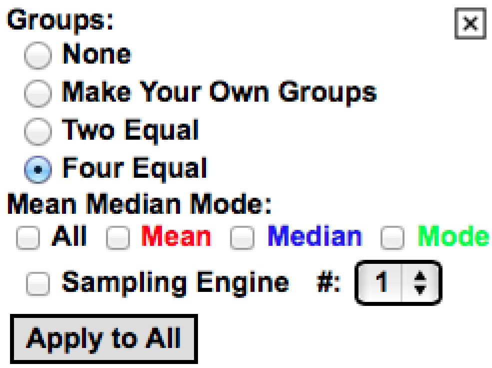
Students should click on "Reach" and drag it over the window to the right in *SeeIt*. *SeeIt* will produce a stacked dotplot distribution of the classroom's reach data. Ask students to click on "Display Options" and increase the dot size to better view the stacked dot plot, if needed.

II. Design and Implement a Plan to Collect the Data

Students will be using the data they collected in day one of this activity, and also datasets that have already been preloaded into *SeeIt.*

III. Analyze the Data

To divide the data into four equal parts, ask students to click on the wrench icon in the graphing widow of *SeeIt* to open a window where they need to click on the bullet next to “Four Equal” (see Figure 5).



Groups:
○ None
○ Make Your Own Groups
○ Two Equal
● Four Equal
Mean Median Mode:
☐ All ☐ Mean ☐ Median ☐ Mode
☐ Sampling Engine #: 1
Apply to All

Figure 5. Selecting the four equal groups option in *SeeIt*.

This will divide the data into four equal parts, each representing 25% of the data (if the number of data points is not divisible by 4, *SeeIt* uses an approximation to divide the dataset). The number in between the lines represents how many of the data points are in-between these lines. In Figure 6, for example, 13points fall into the part on the left. Students can determine the exact values of the five-number summary (minimum, 25th percentile or first quartile (Q1), median or second quartile (Q2), 75th percentile or third quartile (Q3), and maximum) by hovering the cursor anywhere on the vertical lines representing these values. The stacked dotplot should resemble that of Figure 6.

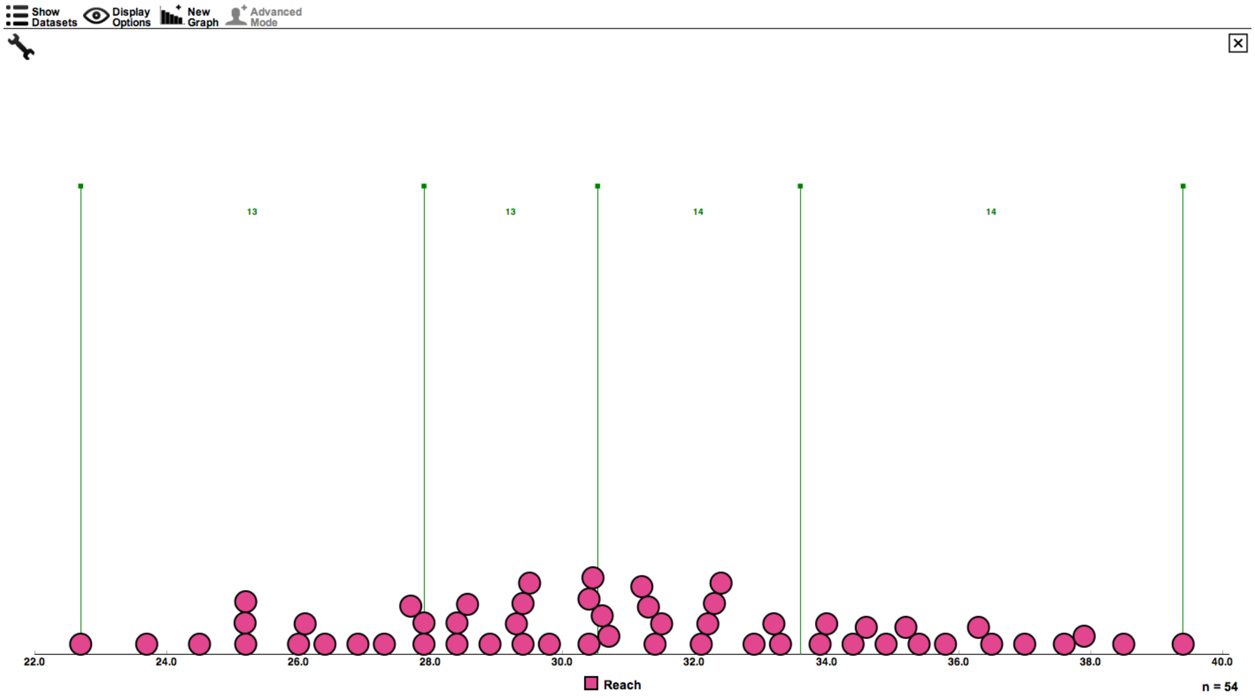


Figure 6. Stacked dotplot divided into four equal parts.

Ask students to go to Question 1 on the Activity Sheet and answer questions about the information that now can be obtained from the stacked dotplot once it is divided into four equal parts.

Proceed to tell students that there is another graphical representation, called a boxplot, that summarizes the data by highlighting the five-number summary and displaying the density of data points in-between these five numbers. Larger (shorter) distances in-between two of these numbers denotes a lower (higher) density of data points in the interval determined by these two points. Boxplots also allow, in most instances, for identification of the shape of the distribution of the data.

To create a boxplot in *SeeIt*, direct students to click on the "Advanced Mode" icon in the bar at the top of the *SeeIt* window. Students should now click on the wrench icon to display the window where they will click on the bullet next to "Box Plot".

SeeIt will produce a boxplot superimposed over the stacked dotplot that highlights the five-number summary. The rectangle in the middle of the graph is called "the box," and the lines on the sides of the rectangle are called "the whiskers." Additionally, *SeeIt* adds a star inside of the box to represent the location of the average or mean of the dataset. Students should get an image similar to that of Figure 7.

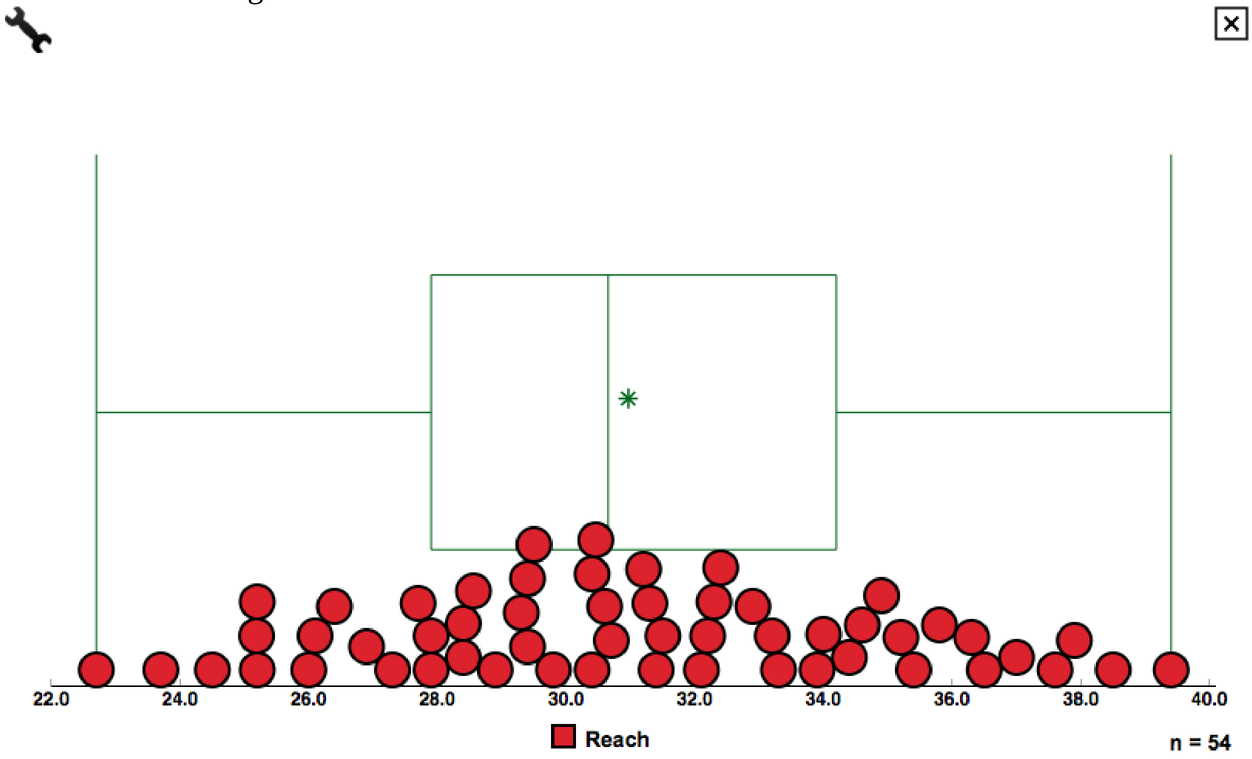


Figure 7. Stacked dotplot with boxplot.

Ask students to go to Question 2 of the Activity Sheet and identify the features of the boxplot that indicate the shape of the distribution of the dataset. For example, if the data set of their hand reaches is roughly symmetric and bell-shaped: the box representing the middle 50% of the data is divided by the median roughly in the middle, and the whiskers are also symmetric and longer than each of the roughly-equal halves of the box. These symmetric and longer whiskers indicate that the top and bottom 25% of the data are spread symmetrically over a longer interval (the spread of the tails) than each of the intervals that divide the middle 50% of the data in two parts (the halves of the bell).

Now, ask students to visualize in *SeeIt* the dataset with the boxplot only and not the individual data points. To do this, students need to click on the "Display Option" icon in the top bar to open a window where they need to click on the bullet next to "Hide." Then ask students to go to Question 3 of the Activity Sheet.

Now students will repeat this gradual data visualization from stacked dotplots to boxplots using datasets that have the five distributions described in day one: Normal or Bell-Shaped, Skewed Right, Skewed Left, Flat or Uniform, and Bi-Modal.

Have students go into *SeeIt* and click on the dataset under the name Fitness Data. This dataset contains the variables that have the shape of the five common distributions that were previously discussed. Group A shows a normal distribution, Group B shows a skewed right distribution, Group C shows a skewed left distribution, Group D shows a uniform distribution, and Group E shows a bi-modal distribution.

Ask students to click on the variable that says Group A (under the dataset titled Fitness Data), and drag it into the empty graph window of *SeeIt*. Students can click on the icon that says New Graph on the bar at the top of *SeeIt*, and then drag the variable of Group B into the space that will open below the graph of Group A.

This process can be repeated for all groups of students in the dataset. After the fourth graph the screen will start to scroll down. Ask students to click on "Display Options" and increase the dot size to better view the stacked dotplot if needed.

Ask students to divide the distributions into four parts as shown above and illustrated in Figure 8. For each different distribution ask students to identify the values of the five-number summary.

Note: Students can click on the "Apply to All" button located in the window that appears when clicking on the wrench icon (same icon to select the option to divide a plot into four equal parts) to expedite this process. Student should get an image similar to that of Figure 8.

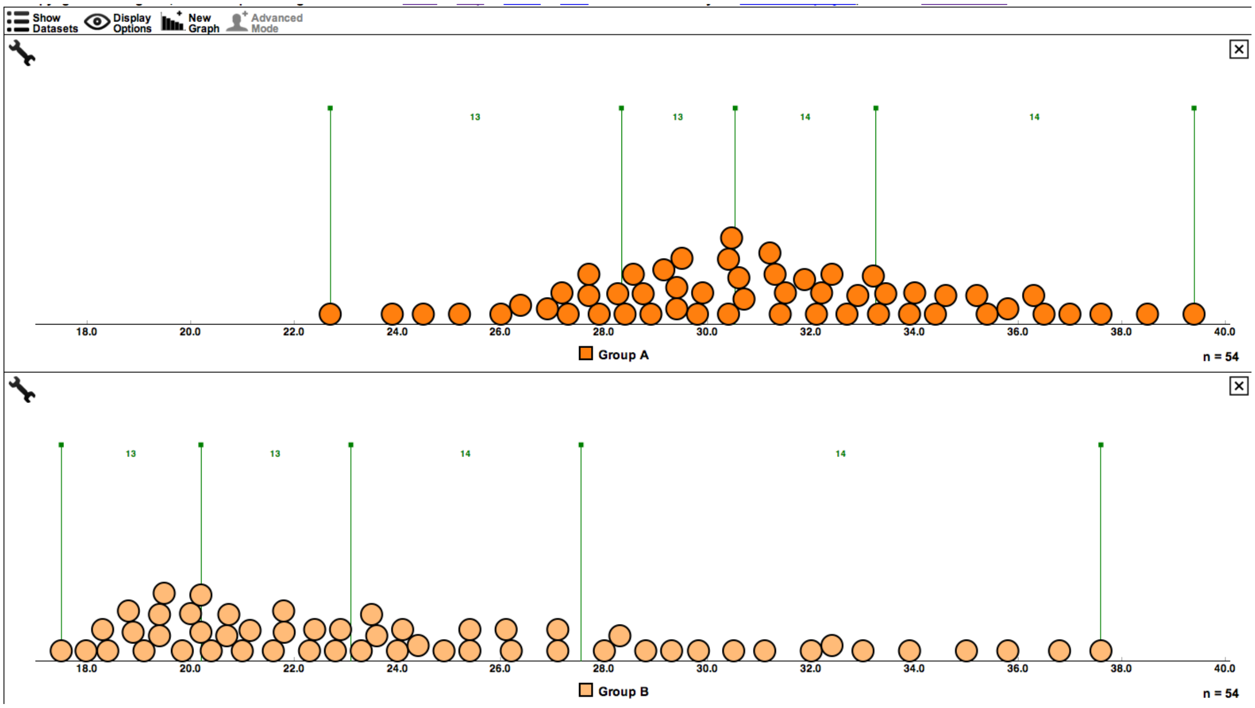


Figure 8. Stacked dotplots of the groups of students divided into four equal parts.

Finally, ask students to create a boxplot for each of the five distributions.

Note: Students can click on the button “Apply to All” to create boxplots for all stacked dotplots at once similarly to the way that all stacked dotplots can be divided into four equal parts at once (see previous note). The five distributions, along with their corresponding boxplots, are shown in Figure 9.

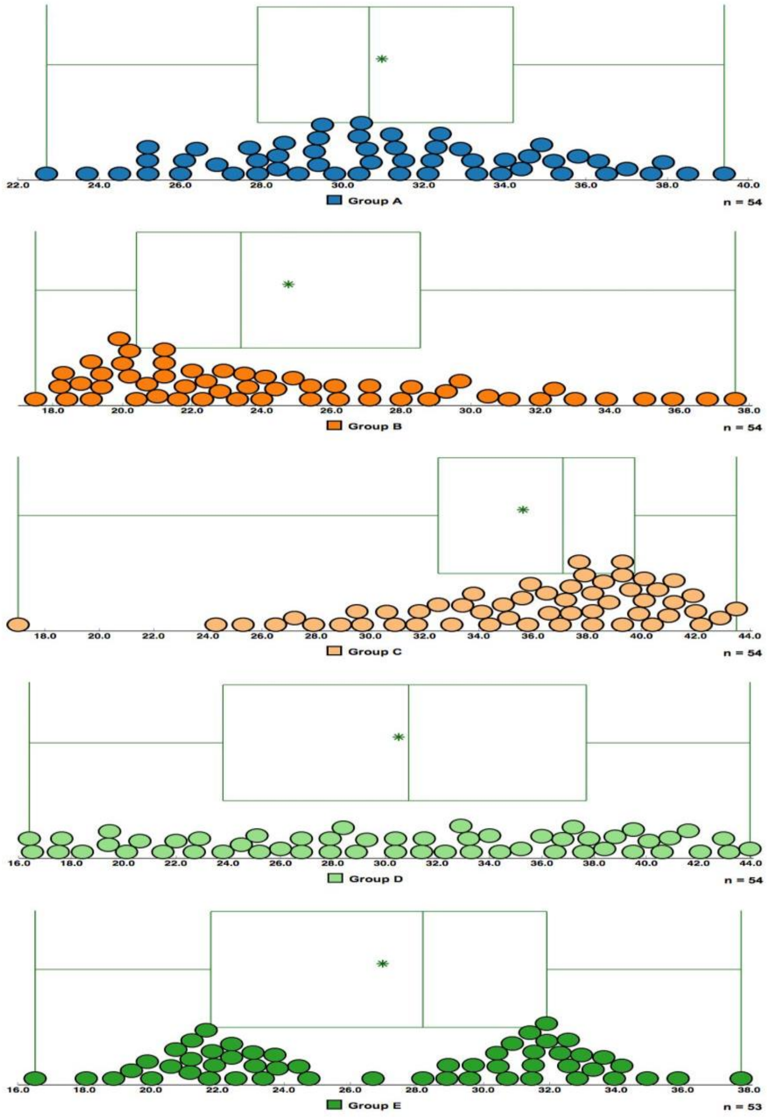


Figure 9. Boxplots superimposed of the stacked dotplots of all groups of students.

IV. Interpret the Results

Ask students to identify the features of the boxplots that correspond to the shapes of the distributions they identified in the stacked dotplots.

A. Bell-Shaped or Normal: symmetric data that is concentrated in the middle, with equal tails on either side. The boxplot will be fairly symmetric with Q1 and Q3 being centered around the peak of the distribution, while the median will be fairly centered between Q1 and Q3. The lower and upper whiskers will extend symmetrically in opposite directions, and their length will be longer than the roughly-equal halves of the box.

B. Skewed Right: most of the data are concentrated on the *left*, with a tail trailing off to the *right*. With skewed right distributions, most of the data being concentrated on the left will cause the first quartile to be closer to the median, and this will create a box divided by the median with a smaller left side. Also, the whisker on the right will be longer than the one of the left indicating a longer spread of the upper 25% of the data which corresponds to the long right tail in the stacked dotplot.

C. Skewed Left: most of the data are concentrated on the *right*, with a **tail** trailing to the *left*. With skewed left distributions, most of the data being concentrated on the right will cause the third quartile to be closer to the median, and will create a box divided by the median with a smaller right side. Also, the whisker on the left will be longer than the one of the right indicating a longer spread of the lower 25% of the data which corresponds to the long left tail in the stacked dotplot.

D. Uniform or Flat: the data are evenly spread out over the distribution. The four intervals into which the quartiles divide the range of the data have all the same length. Thus the intervals in between the five-number summary in a boxplot have all the same length creating whiskers with the same length as the equal-halves of the box.

E. Bi-Modal: the data is concentrated on both the right and the left, creating two distinct peaks. The boxplot for a *symmetric* bi-modal distribution can be a bit misleading when distinguishing it from the boxplots of a bell-shaped or a uniform distribution. The length of equal halves of the box, in the symmetric case, will either be longer or shorter than the length of the equally-long whiskers depending on the distance between the peaks. Peaks close to each other will make the equal halves of the box shorter than the equally-long whiskers (similar to a boxplot of a bell-shaped distribution), and vice versa when the peaks are separated by a longer distance. It is also possible that the boxplot of a bimodal distribution will have all quartiles equally separated making it look like the boxplot of a uniform distribution. This shows one of the limitations that boxplots can present to even experienced boxplot users.

Finish this activity by asking students to compare and contrast boxplots and stacked dot plots. Then ask them to answer Question 4 on the Activity Sheet.

Advantages:

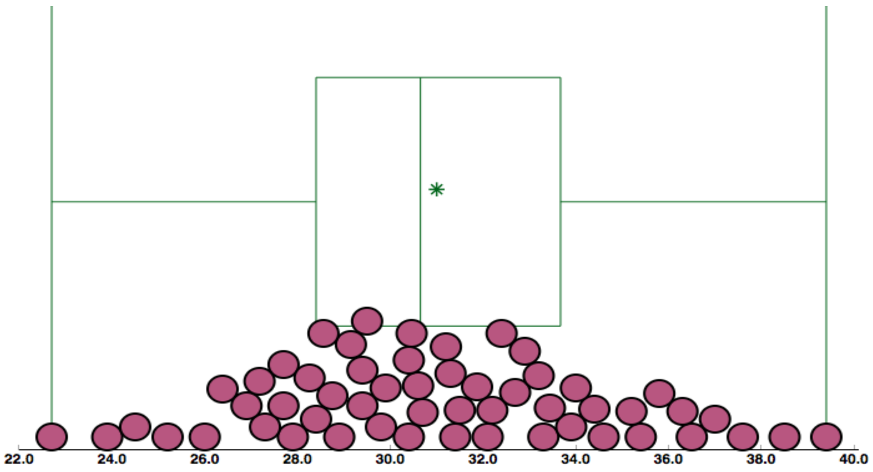
- Boxplots organize datasets with more information than stacked dotplots in terms of including the five-number summary (thus one can obtain a sense of a center and average spread of values from center).
- Boxplots can handle the display of very large datasets, while stacked dotplots can become messy.
- Boxplots can be displayed side-by-side or on top of each other for comparison purposes. Several stacked dotplots are difficult to display side-by-side or on top of each other.

Disadvantages:

- Boxplots cannot identify the individual elements of a dataset.
- Boxplots of symmetric data (bell-shape, uniform, bimodal, etc) may be difficult to differentiate.

Assessment

1. Use the graph below to answer the questions that follow.

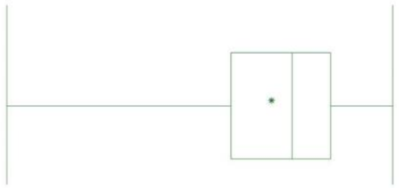


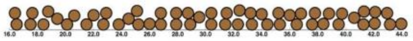
a) Approximate the value of the median.

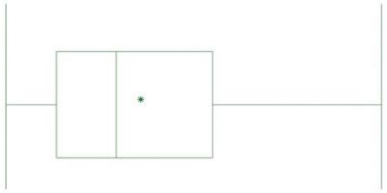
b) Approximate the values of the 25th percentile (or first quartile Q1), and the 75th percentile (or third quartile Q3).

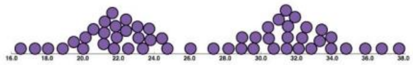
c) Approximate the minimum and maximum values of the data set.

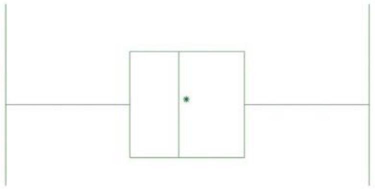
2. A series of boxplots and stacked dotplots are shown on the left column below. Please draw the corresponding dotplot/boxplot for each graph:

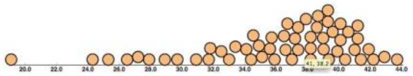


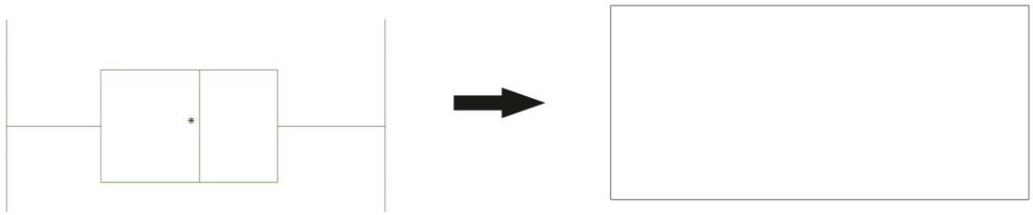


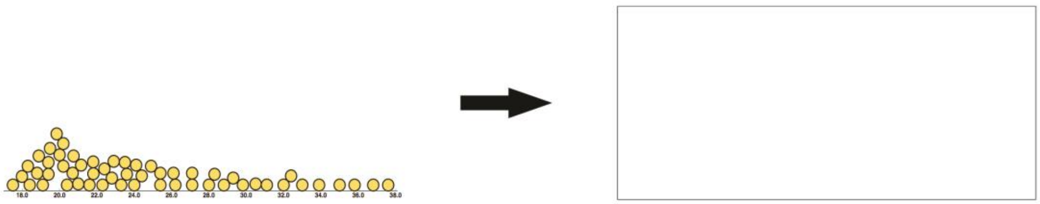


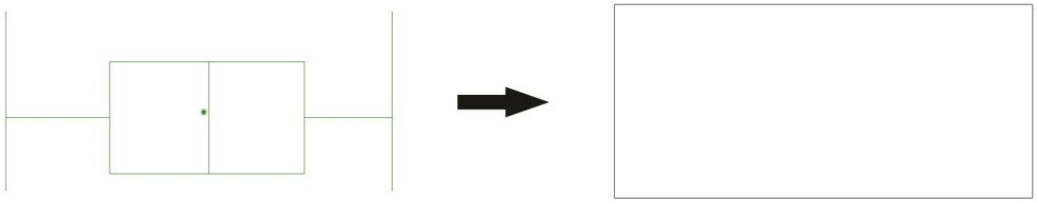


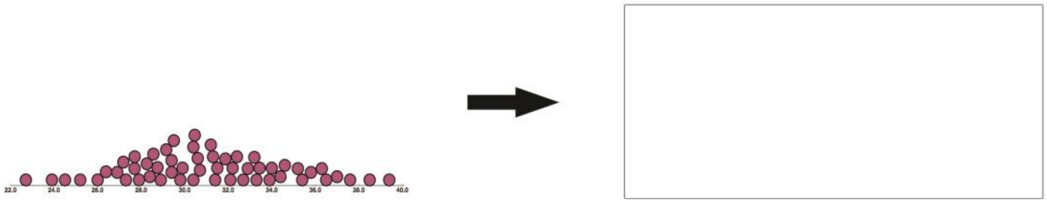







18.0
20.0
22.0
24.0
26.0
28.0
30.0
32.0
34.0
36.0
38.0




22.0
24.0
26.0
28.0
30.0
32.0
34.0
36.0
38.0
40.0

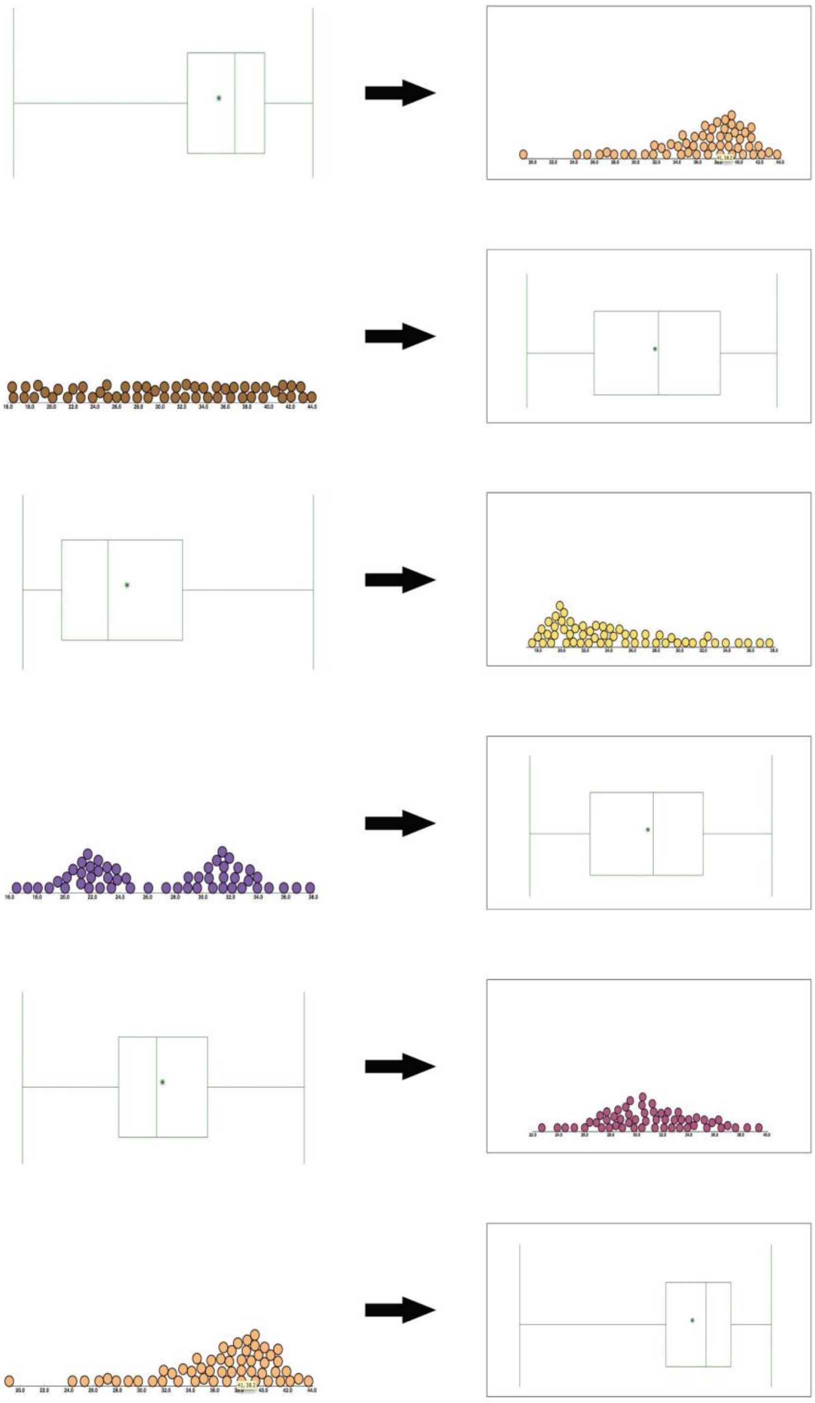
Answers

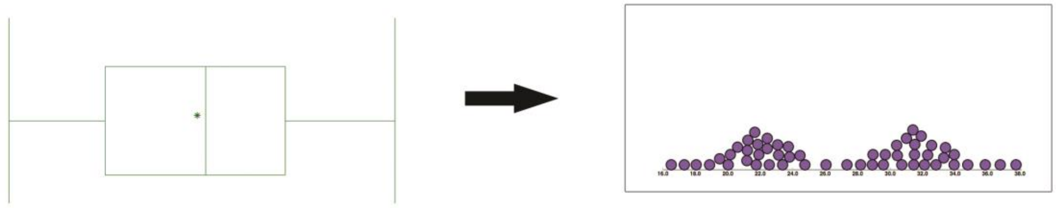
1. a) 31

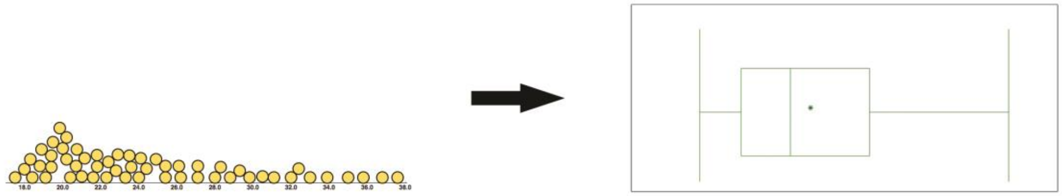
b) Q1 = 29 and Q3 = 33

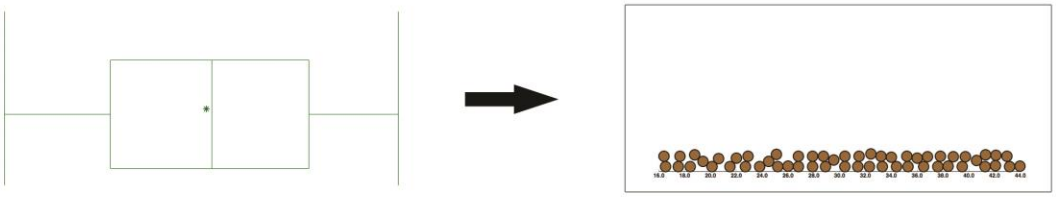
c) The minimum is around 22. The maximum is around 40.

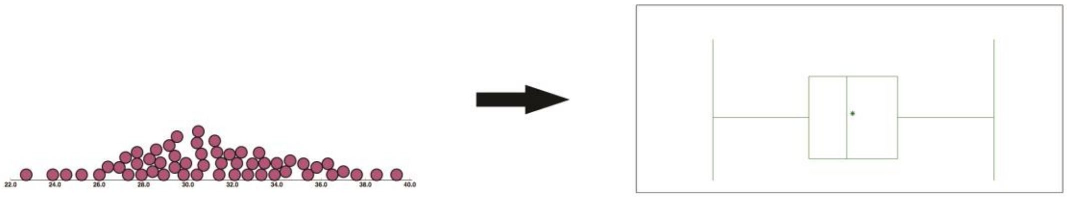
2.




16.0
18.0
20.0
22.0
24.0
26.0
28.0
30.0
32.0
34.0
36.0
38.0


18.0
20.0
22.0
24.0
26.0
28.0
30.0
32.0
34.0
36.0
38.0


16.0
18.0
20.0
22.0
24.0
26.0
28.0
30.0
32.0
34.0
36.0
38.0
40.0
42.0
44.0


22.0
24.0
26.0
28.0
30.0
32.0
34.0
36.0
38.0
40.0

Possible Extensions

SeeIt can also identify outliers when producing boxplots. You can extend the teaching of this lesson to include outliers by clicking the advanced bullet next to Boxplot when creating a boxplot. Also, since the mean of the data set is represented with a star in the boxplots of *SeeIt*, students can learn about skewness when comparing means and medians.

Phantom/TinkerPlot users: To download the *SeeIt* data for use in another program simply click on the title of the dataset in *SeeIt* you want to download. Then click on the pencil icon that will appear just below the title of the dataset. This will display the *SeeIt* worksheet window with the dataset in it. Highlight the data and copy. From here you are able to paste the data into an Excel or Google Spreadsheet for use in Phantom or TinkerPlots.

References

1. Bakker, A; Biehler, R; Konold, C (2004). Should Young Students Learn about Boxplots? Curricular Development in Statistics Education. Sweden.

2. Pfannkuch, M (2006). Comparing Box Plot Distributions: A Teacher's Reasoning. Statistics Education Research Journal, 5(2), 27-45.

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>.

Acknowledgments

We thank Matthew Steinwachs and Pearl Buniger for their skillful contributions with the technical aspects of the datasets in *SeeIt*, and the creation of some of the graphics used in this lesson plan. We also thank Alisa Lee, Education Manager at the UC Davis iAMSTEM Hub, for her valuable help in managing the project "How Sure Are You? Science, Biostatistics and Cancer Education," which funded the research involved in this lesson plan through an NIH Science Education Partnership Award (SEPA). Finally, we thank Tobias Spencer, a high school teacher who directed the pilot testing of the contents of this lesson plan in several high schools.

Now You SeeIt, Now You Don’t: Using SeeIt to Compare Stacked Dotplots to Boxplots
Activity Sheet Day 1

1. Answer these questions after learning about information that can be obtained from the shape of a distribution.

a. Which group of students is most likely to be made up of mostly athletes? Why? What is the shape of the distribution of this group?

b. Which group is most likely to be made up of students who are mostly not physically active? Why? What is the shape of the distribution of this group?

c. Which group is most likely to be made up of students with mostly an “average” level of fitness? Why? What is the shape of the distribution of this group?

d. Which group seems to be made up of a mixture of a subgroup of students who are physically active, and another subgroup of students who are not? Why? What is the shape of the distribution of this group?

e. Which group seems to be made up of students with an equal number of students from all levels of physical fitness? Why? What is the shape of the distribution of this group?

2. Write down an advantage and a disadvantage of visualizing the distribution of a dataset using a stacked dotplot as opposed to looking at the dataset listed on a table.

3. Once you are done graphing the stacked dotplot of the hand reaches in the class in *SeeIt analyze* the data by answering the following questions.

a. What is the shape of the distribution of hand reaches?

b. Can you readily get other information from the stacked dotplot of the hand reaches such as:

i. The maximum and minimum hand reach? If so, what are these values?

ii. The location of the hill(s) of the distribution? That is, the value around which the majority of the hand reaches is concentrated? If so, what is this value?

iii. The presence of any extreme values or outliers? If so, what are their values?

Note: Hovering the cursor on any data point in the stacked dotplot will display the name and the reach of the student that the point represents.

4. Once you are done graphing the stacked dotplot of the hand reaches in the class in *SeeIt interpret* the data by answer the following questions.

a. What range would you use to describe the most common hand reaches of the classroom? (Hint: use a range around the hill of the distribution).

b. What does the shape of the distribution tell you about your classroom? For example, since the dataset is made up of measurements of something natural, does the data appear to be bell-shaped? Is the data bimodal, and thus indicate that there are two distinct subgroups in the classroom? (e.g. males and females differing in their hand reach measurements).

c. Are there any extreme hand reaches? If so, why are these values away from the rest?

Now You SeeIt, Now You Don't: Using SeeIt to Compare Stacked Dotplots to Boxplots Activity Sheet Day 2

1. Answer these questions after you have produced the stacked dotplot of the hand reaches divided into four equal parts in *SeeIt*.

a. What is the median hand reach in the class?

b. What are the endpoints of the middle 50% of the data? Is your hand reach in this group? If it is in this group, can you identify the data point that corresponds to your hand reach?

c. What are the endpoints of the lower 25% of the data? Is your hand reach in this group? If it is in this group, can you identify the data point that corresponds to your hand reach?

d. What are the endpoints of the upper 25% of the data? Is your hand reach in this group? If it is in this group, can you identify the data point that corresponds to your hand reach?

e. Which of the four groups is concentrated over the shortest interval created by the five-number summary (minimum, 25th percentile or first quartile (Q1), median or second quartile (Q2), 75th percentile or third quartile (Q3), and maximum)? Over the next shortest? and so on.

f. Do shorter intervals created by the five-number summary correspond to groups that are more clustered or less clustered together?

g. Have you been able to obtain more information about the hand reaches by adding vertical lines that divide the stacked dotplot into four equal parts? Would it be helpful to have a graph that highlights the five-number summary of a dataset?

2. After obtaining the stacked dotplot of the hand reaches in day 1 of this lesson, you were asked to identify the shape of the distribution. What are the features of the boxplot that correspond to the shape of this distribution? E.g. If the upper whisker is longer than the lower one then you can conclude that the dataset is skewed right, etc.

3. Is the boxplot alone more (without the individual data points) informative of a dataset than a stacked dotplot alone (without the lines dividing it into four equal parts)? Why?

4. Answer these questions after comparing and contrasting boxplots and stacked dotplots.

a. What are the advantages and disadvantages of a boxplot in comparison to a stacked dotplot?

b. Would you prefer to use a stacked dotplot or a boxplot to obtain information about a dataset? Why?