Lesson 18: Connecting Graphical Representations and Numerical Summaries

Student Outcomes

- Students match the graphical representations and numerical summaries of a distribution. Matches involve dot plots, histograms, and summary statistics.

Lesson Notes

It can be difficult to understand a data set by just looking at raw data. Imagine that Joaquin's project is on finding the typical weight of bears. He finds an article about bears that provides an unsorted list of the weights of 250 bears with no numerical summaries or graphs of these data. It would be very difficult to quickly draw some conclusions from this data. Joaquin decides to design his project using this data.

Next Joaquin finds an article that states, "The median weight for the bears studied was 305 pounds." This is useful, but sometimes even numerical summaries alone cannot completely convey interesting aspects of a data distribution. Often, readers like Joaquin want to have a concise and useful summary of the information that is both numerical *and* visual.

In the next couple of lessons, students will begin to take the graphical representations and numerical summaries they learned and apply them to different situations. While working through these lessons, students should keep in mind their own statistical question. They should think about which graphs will best showcase their data and which numerical summaries will represent the data they are collecting.

Classwork

It can be difficult to understand a data set by just looking at raw data. Imagine that Joaquin's project is on finding the typical weight of bears. He found an article about bears that provided an unsorted list of the weights of 250 bears with no numerical summaries or graphs of these data. It would be very difficult to quickly draw some conclusions. Joaquin decided to design his project using this data.

Now consider the case where the article provides you with a statement, "the median weight for the bears studied was 305 pounds." This is useful, but sometimes even numerical summaries alone cannot completely convey interesting aspects of a data distribution. Often, readers want to have a concise and useful summary of the information that is both numerical and visual.

In the next couple of lessons, you will begin to take the graphical representations and numerical summaries you learned and apply them to different situations. While working through these lessons, keep in mind your own statistical question. Think about which graphs will best showcase your data and which numerical summaries will represent the data you are collecting.

Example 1 (3 minutes): Summary Information from Graphs

Review dot plots and histograms. Important points are as follows:

- Each picture reveals a great deal about what is occurring.
- Some pictures may look different from one another and highlight different aspects of the same data set, but graphs of the same data set can still have several similarities and impart similar information.
- Summary measures can be obtained or estimated from dot plots and histograms, and these measures complement the graphical information.

MP.4

Example 1: Summary Information from Graphs

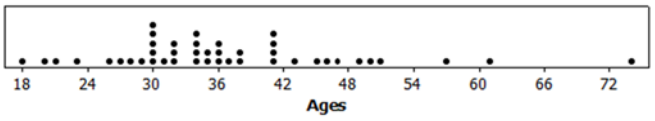
Recall that a *dot plot* includes a dot on a scale or number line for each observation in a data set. Dots are stacked on top of one another when there are multiple occurrences of a data value. Recall also that a *histogram* similarly uses a scale or number line to present the frequency or relative frequency of groups of data based on intervals of equal width. For each interval, the height of the bar is proportional to the number of observations in the interval; the taller the bar, the greater the number of observations in that interval. This means that when both graphs are generated for a given data set, the two graphs will display some similarities.

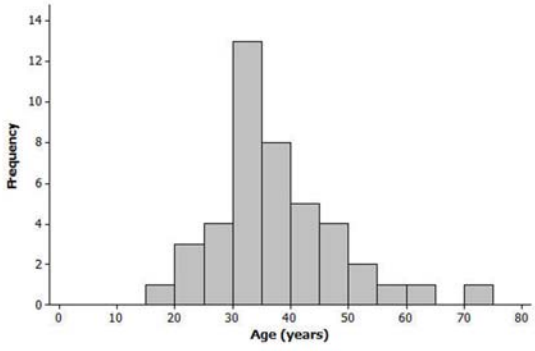
Here is a data set of the ages (in years) of 43 participants in a recent local 5-kilometer race.

20	30	30	35	36	34	38	46
45	18	43	23	47	27	21	30
32	32	31	32	36	74	41	41
51	61	50	34	34	34	35	28
57	26	29	49	41	36	37	41
38	30	30					

Here are some summary statistics, a histogram, and a dot plot for the data:

Minimum = 18, Q1 = 30, Median = 35, Q3 = 41, Maximum = 74; Mean = 36.81, MAD = 8.1





Exercises 1–7 (7–10 minutes)

Pose these questions to students one at a time.

Exercises 1–7

1. **Based on the histogram, would you describe the shape of the data distribution as approximately symmetric or as skewed? Would you have reached this same conclusion looking at the dot plot?**

   Both graphs show a slightly skewed right data distribution.

2. **Is it easier to see the shape of the data distribution from the histogram or the dot plot?**

   Generally, it is easier to see the shape of the data distribution from a histogram. In this case, the clustering and high frequency of ages in the 30s is more evident in the histogram.

3. **What is something you can see in the dot plot that is not as easy to see in the histogram?**

   When using the histogram, we cannot determine the exact minimum or maximum age—for example, we only know that the minimum age is between 15 and 19 years of age. Also, we can only approximate the median (we generally cannot figure out the exact median value from a histogram).

   Since the dot plot provides us with a dot for each observation, we can obtain specific (or sometimes rounded) values for a 5-number summary—and the entire data set—from the dot plot. With the dot plot, we see that the minimum is specifically 18. The median is the 22^{nd} observation (since there are 43 observations) and the 22^{nd} dot counting from left to right is 35 (we cannot be that precise with the histogram). The oldest runner (74) also appears to be a more extreme departure from the rest of the data in the dot plot as compared to the histogram.

4. **Do the dot plot and the histogram seem to be centered in about the same place?**

   Yes, as both graphs are based on the same data, they should generally communicate the same information regarding center.

5. **Do both the dot plot and the histogram convey information about the variability in the age distribution?**

   Yes, as both graphs are based on the same data, they should generally communicate the same information regarding variability. However, as mentioned earlier, the oldest runner (74) appears to be a more extreme departure from the rest of the data in the dot plot as compared to the histogram.

6. **If you did not have the original data set and only had the dot plot and the histogram, would you be able to find the value of the median age from the dot plot?**

   Yes, see response to Exercise 3.

7. **Explain why you would only be able to estimate the value of the median if you only had a histogram of the data.**

   The median is the 22^{nd} ordered observation in this data set since there are 43 observations. Counting from left to right, we know that the first 21 observations are in the first 4 classes: 15–19 (1 value), 20–24 (3 more values), 25–29 (4 more values), and 30–34 (13 more values). Cumulatively, we have encountered the lowest 21 observations by the time we are finished with the 30–34 class. So, the 22^{nd} value must be in the next class, which is 35–39 years of age. We just cannot determine the exact value from the histogram.

MP.6

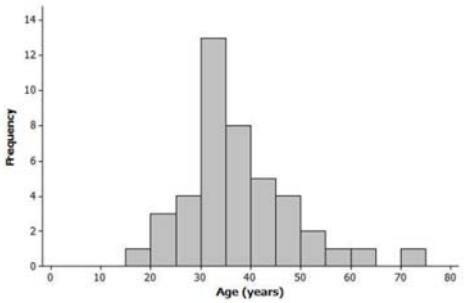
Exercises 8–13 (25 minutes): Graphs and Numerical Summaries

Pose the questions to students one at a time. Allow for more than one student to offer an answer for each question encouraging a brief (2 minute) discussion.

Note: In some cases, the questions have multiple and/or inexact answers.

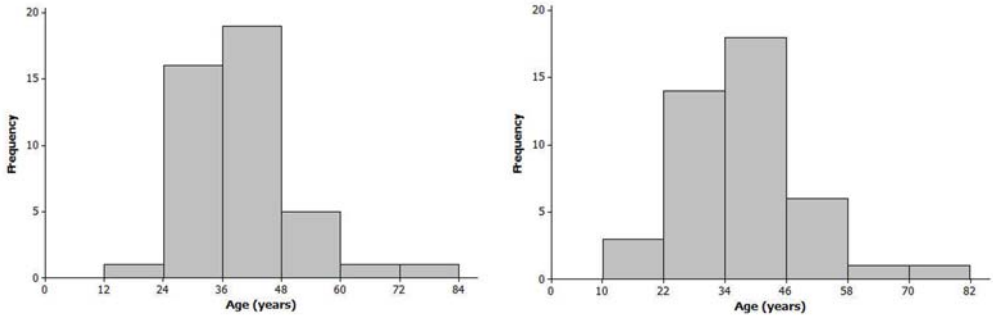
Exercises 8–13: Graphs and Numerical Summaries

8. **Suppose that a newspaper article was written about the race and the histogram of the ages from Example 1 was shown in the article. The writer stated, "The race attracted many older runners this year; the median age was 45." Explain how we would know that this is an incorrect statement based on just the histogram.**



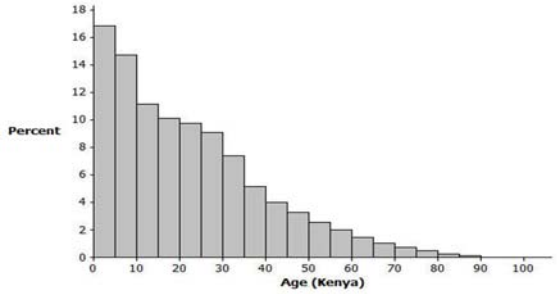
Several answers are possible, but students should concentrate on the shaded area (distribution) shown and the definition of a median. Specifically, at 45, it appears that less than half of the data are (or area is) at or above that value (or alternatively, more than half of the data are at or below that value). Another approach would be to state that the value at which the data (area) might be split, 50% below/50% above appears to be at a lower value than 45 (or in the interval 35–39).

9. **One of the histograms below is another valid histogram for the runners' ages. Select the correct histogram, and explain how you determined which graph is valid (and which one is incorrect) based on the summary measures and dot plot.**



One of the objectives is to reinforce that there is more than one way to draw a proper histogram for a distribution. This question is especially detail-oriented because students need to carefully reconcile components of the histogram with the data set (either as shown in raw form or in the dot plot). The histogram on the right is the correct graph because it is consistent with the dot pot/data. Most notably, the histogram on the right correctly shows there are 3 runners in the 10–21 age group, while the left histogram shows only 1 runner in the 12–23 age group (and there are actually 4 runners in that class). Other classes in the left histogram do not match the dot plot/data (e.g., the 48–59 group), so several answers are possible.

10. The histogram below represents the age distribution of the population of Kenya in 2010.



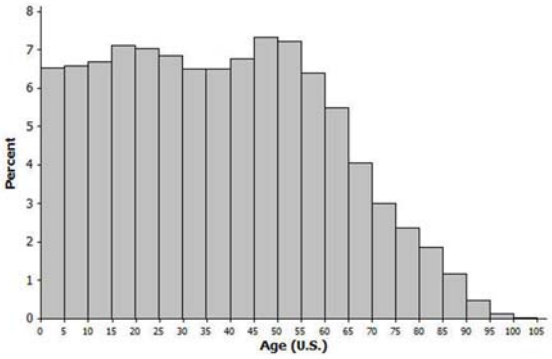
a. How do we know from the graph above that the first quartile (Q1) of this age distribution is between 5 and 9 years of age?

Since a histogram should display information that is consistent with summary measures, we are seeking a data value such that 25% of the distribution is at or below that value. While the 0–4 age group represents the lowest 17% approximately, the next group (age 5–9) appears to account for the next approximately 15% of the distribution. This means that cumulatively this second group (ages 5–9) roughly represents the lowest 17%–32%, thus the first quartile would be in that group.

b. Someone believes that the median age is near 30. Explain how the graph supports this belief, OR explain why the graph does not support this belief.

The median does NOT appear to be 30 years of age. See answers for Exercises 1 and 3 for guidance in determining this. Specifically, the 50^{th} percentile estimated by adding approximate percentages (and/or visually assessing the point at which the area seems split evenly) appears to be in the 15–19 age group.

11. The histogram below represents the age distribution of the population of the United States in 2010. Based on the histogram, which of the following ranges do you think includes the median age for the United States: 20–29, 30–39, or 40–49? Why?



Using similar arguments as described in the response to Exercise 10 part (b), the median appears to be in the 30–39 age group, most likely in the 35–39 class.

12. Use the histograms from Exercises 10 and 11 to answer the following:

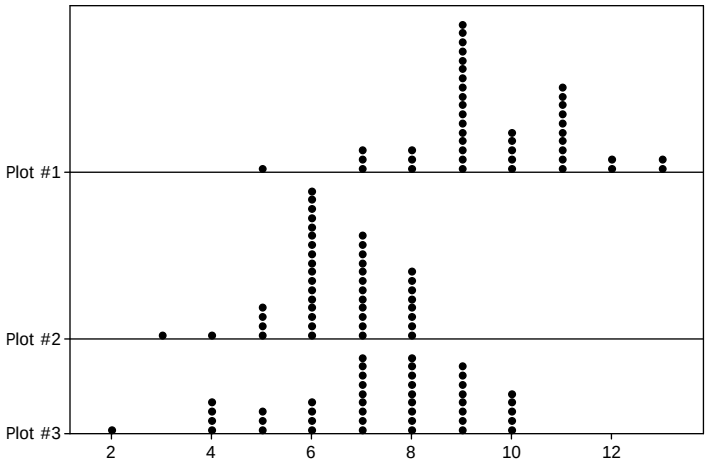
a. Which country's age distribution (Kenya or United States) has a third quartile in the 50s? How did you decide?

The third quartile of the U.S. is in the 50s, and this can be determined using methods of area visualization or cumulative percentage counting as described above. Note also that for the value of 50 in the Kenya distribution, there appears to be far less than 25% of the distribution shown above that value.

b. If someone believed that the average age of a person living in the United States was greater than the average age of a person living in Kenya, how could you support that claim by comparing the histograms?

There are a few ways to support this remark. First, there is a considerably higher percentage of high ages in the U.S. distribution. Secondly, as skewed right distributions tend to have a mean that is higher than the median, the fact that the U.S. has a higher median age than Kenya (Exercises 4 and 5) would support the idea that the U.S. would have a higher mean age than Kenya. Lastly, using a balance point argument, the balance point for the U.S. would be much further up the number line than the balance point for Kenya.

13. Match the following sets of summary measures with the corresponding dot plot. Only ONE dot plot matches each group of summary measures. Explain why you selected the dot plot or why the other dot plots would not represent the summary measures. Note: the same scale is used in each dot plot.



a. Median = 8 and IQR = 3 Plot # _____

Plot #3 – It is the only distribution visually centered near 8 and one can tell that the 22^{nd} ordered observation (the median in this case) is 8. Also, plot #3 is the only distribution with an IQR of 3.

b. Mean = 9.6 and MAD = 1.28 Plot # _____

Plot #1 – This plot is the only plot for which one could assume a mean value as high as 9.6. It is the only plot with values of 11, 12, and 13 – and there are several of these values.

c. Median = 6 and Range = 5 Plot # _____

Plot #2 – It is the only plot which appears to have a central value of 6. It is also the only plot with a range of 5 (each of the other plots has a range of 8).

Closing (5 minutes)

Consider posing the following questions; allow a few student responses for each:

- What kinds of information about a quantitative data distribution might not be presented well if we only use summary measures?
  - *Clustering, aspects of shape, extremeness of certain values, etc.*
- If dot plots can provide us with a way of figuring out exact (or nearly exact) observation values, why don't we always use dot plots to show a data distribution? What are some cases where a histogram might provide a better visual summary of the distribution or a dot plot might not work well?
  - *Clustering and gaps might be more easily shown in a histogram. A dot plot may be cumbersome for large data sets -- like the population distribution of an entire country!*

Lesson Summary

Generally, we can compute or approximate many values in a numerical summary for a data set by looking at a histogram or a dot plot for the data set. Thus, we can generally match a histogram or a dot plot to summary measures provided.

When making a histogram and a dot plot for the same data set, the two graphs will have similarities. However, some information may be more easily communicated by one graph as opposed to the other.

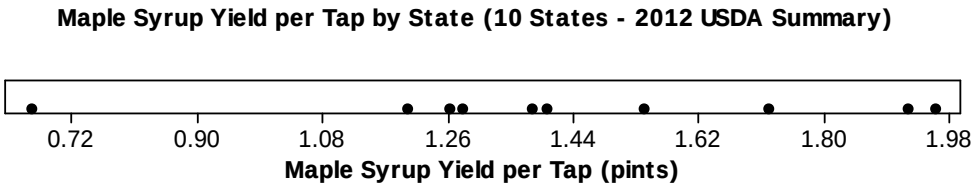
Exit Ticket (5–8 minutes)

Name _____ Date_____

Lesson 18: Connecting Graphical Representations and Numerical Summaries

Exit Ticket

1. Many states produce maple syrup, which requires tapping sap from a maple tree. However, some states produce more pints of maple syrup per tap than other states. The following dot plot shows the pints of maple syrup yielded per tap in each of the 10 maple syrup producing states as listed in the *US Department of Agriculture's 2012 Crop Production Summary*. For the dot plot, which ONE of the three sets of summary measures could match the graph? For each choice that you eliminate, list at least one reason for eliminating the choice.



(From the *United States Department of Agriculture National Agricultural Statistics Service Crop Production 2012 Summary*, ISSN: 1057-7823, p. 75, accessed May 5, 2013 available at http://usda01.library.cornell.edu/usda/current/CropProdSu/CropProdSu-01-11-2013.pdf.)

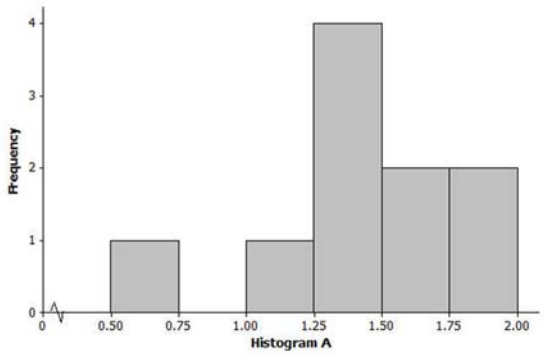
a. Minimum = 0.66, Q1 = 1.26, Median = 1.385, Q3 = 1.71, Maximum = 1.95, Range = 2.4; Mean = 1.95, MAD = 0.28

b. Minimum = 0.66, Q1 = 1.26, Median = 1.71, Q3 = 1.92, Maximum = 1.95, Range = 1.29; Mean = 1.43, MAD = 2.27

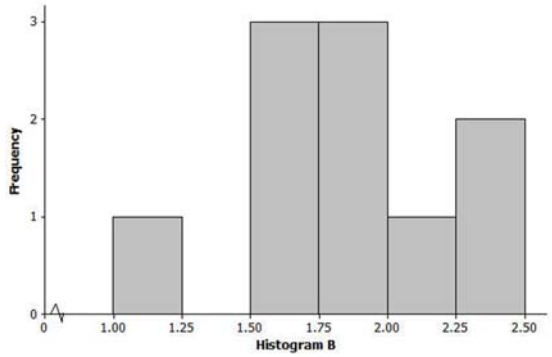
c. Minimum = 0.66, Q1 = 1.26, Median = 1.385, Q3 = 1.71, Maximum = 1.95, Range = 1.29; Mean = 1.43, MAD = 0.28

2. For the dot plot in problem 1, which ONE of the three histograms below could be a match? For each choice that you eliminate, list at least one reason for eliminating the choice.

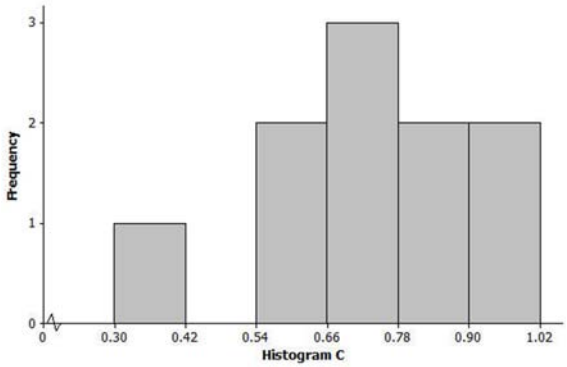
a.



b.



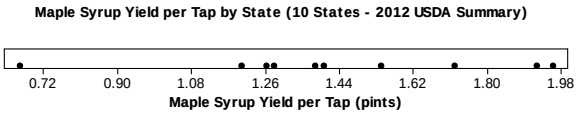
c.



Exit Ticket Sample Solutions

Note: Students do not need to list all elimination reasons listed below. The instructions asked, "For each choice that you eliminate, list *at least one* reason for eliminating the choice."

1. **Many states produce maple syrup, which requires tapping sap from a maple tree. However, some states produce more pints of maple syrup per tap than other states. The following dot plot shows the pints of maple syrup yielded per tap in each of the 10 maple syrup producing states as listed in the *US Department of Agriculture's 2012 Crop Production Summary*. For the dot plot, which ONE of the three sets of summary measures could match the graph? For each choice that you eliminate, list at least one reason for eliminating the choice.**



(From the *United States Department of Agriculture National Agricultural Statistics Service Crop Production 2012 Summary*, ISSN: 1057-7823, p. 75, accessed May 5, 2013 available at http://usda01.library.cornell.edu/usda/current/CropProdSu/CropProdSu-01-11-2013.pdf.)

   a. **Minimum = 0.66, Q1 = 1.26, Median = 1.385, Q3 = 1.71, Maximum = 1.95, Range = 2.4; Mean = 1.95, MAD = 0.28**

   b. **Minimum = 0.66, Q1 = 1.26, Median = 1.71, Q3 = 1.92, Maximum = 1.95, Range = 1.29; Mean = 1.43, MAD = 2.27**

   c. **Minimum = 0.66, Q1 = 1.26, Median = 1.385, Q3 = 1.71, Maximum = 1.95, Range = 1.29; Mean = 1.43, MAD = 0.28**

   The correct answer is (c).

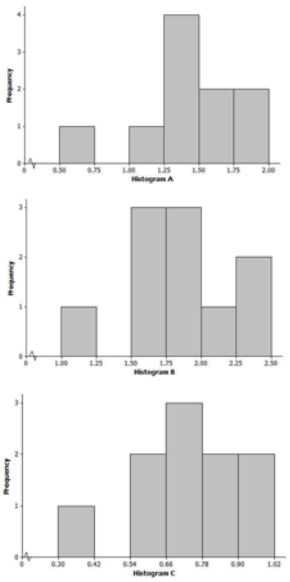
   Choice (a) would not work because the range is too large as the difference between maximum and minimum is only 1.29 pints. Also, the mean would not be that close to (or the same as) the maximum value in this case.

   Choice (b) would not work because a median value of 1.71 would be too high. Estimating the dot values, the 5^{th} and 6^{th} ordered observations (the median for a dataset of 10 items) are near 1.4. Also, the MAD is much too large as the range of the data is only 1.29 pints.

2. **For the dot plot in problem 1, which ONE of the three histograms below could be a match? For each choice that you eliminate, list at least one reason for eliminating the choice.**

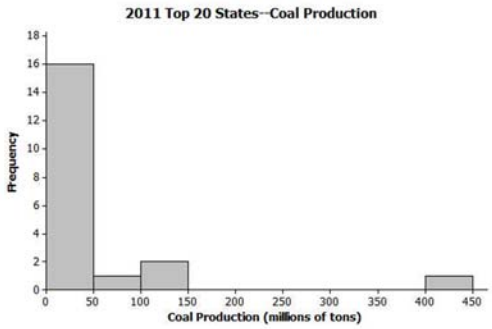
   The correct answer is (a).

   Graphs (b) and (c) are developed from data sets with similar shape features to the correct graph (graph (a)), but the range and distribution of values do not match. For example, graph (b) would not be valid as it is based on 3 observation values of 2 pints or more and there are no values that large in the original dot plot. Also, the smallest value in graph (b) is at least 1 pint, and the actual data set contains a value less than 1 pint. Graph (c) is based on values that are smaller than many of those presented in the dot plot; in fact all of the values in graph (c) are less than 1.02 pints, and nearly all of the 10 observations in the actual data set are greater than 1.02.



Problem Set Sample Solutions

1. The following histogram shows the amount of coal produced (by state) for the 20 largest coal producing states in 2011. Many of these states produced less than 50 million tons of coal, but one state produced over 400 million tons (Wyoming). For the histogram, which ONE of the three sets of summary measures could match the graph? For each choice that you eliminate, list at least one reason for eliminating the choice.



(U.S. Coal Production by State data as reported by the National Mining Association from http://www.nma.org/pdf/c_production_state_rank.pdf accessed May 5, 2013)

a. Minimum = 1, Q1 = 12, Median = 36, Q3 = 57, Maximum = 410; Mean = 33, MAD = 2.76

b. Minimum = 2, Q1 = 13.5, Median = 27.5, Q3 = 44, Maximum = 439; Mean = 54.6, MAD = 52.36

c. Minimum = 10, Q1 = 37.5, Median = 62, Q3 = 105, Maximum = 439; Mean = 54.6, MAD = 52.36

The correct answer is (b).

Choice (a) would not work because Q3 (the average of the 15th and 16th ordered observations) must be less than 50 since both the 15^{th} and 16^{th} ordered observations are less than 50. The mean is most likely greater than (not less than) the median given the skewed right nature of the distribution and the large outlier. The MAD value is most likely much larger than 2.76 given the presence of the outlier and its distance from the cluster of remaining observations.

Choice (c) would not work because since there are 20 observations, the median (the average of the 10^{th} and 11^{th} ordered observations) must be less than 50 since both the 10^{th} and 11^{th} ordered observations are less than 50. Likewise, the Q3 (the average of the 15^{th} and 16^{th} ordered observations) must be less than 50 since both the 15^{th} and 16^{th} observations are less than 50. The mean is most likely greater than (not less than) the median given the skewed right nature of the distribution and the large outlier.

2. The heights (rounded to the nearest inch) of the 41 members of the 2012–2013 University of Texas Men's Swimming and Diving Team are shown in the dot plot below.

Data Source: http://www.texassports.com/sports/m-swim/mtt/tex-m-swim-mtt.html accessed April 30, 2013

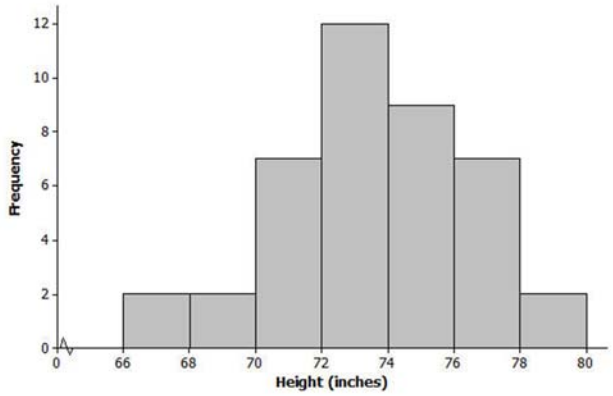
a. Use the dot plot to determine the 5-number summary (minimum, lower quartile, median, upper quartile, and maximum) for the data set.

The 5-number summary values for an ordered data set of 41 observations would be:

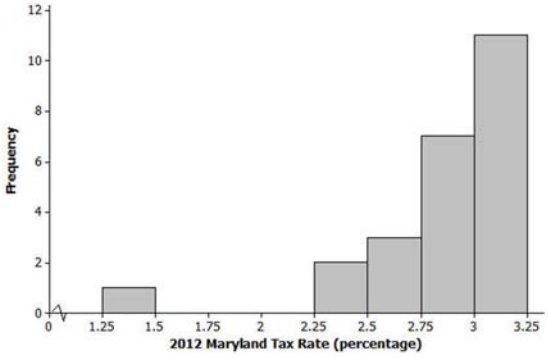
Min = 1st, Q1 = Average of 10^{th} and 11^{th}, Median = 21st, Q3 = Average of 31^{st} and 32^{nd}, Max = 41^{st}

Summary: 66, 71, 73, 75, 80

b. Based on this dot plot, make a histogram of the heights using the following classes: $66-<68$ inches, $68-<70$ inches, and so on.

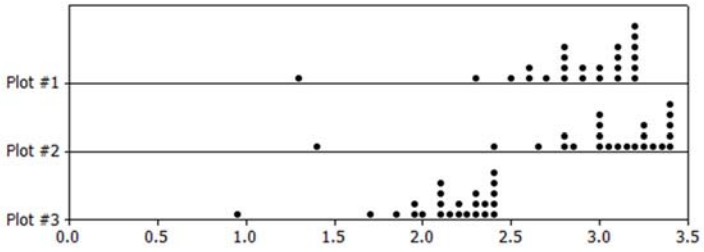


3. According to the website of the Comptroller of Maryland, "Maryland's 23 counties and Baltimore City levy a local income tax … Local officials set the rates, which range between 1.25% and 3.20% for the current tax year (2012)." A histogram of the 24 tax rates (in percentages) appears below.



Data Source: http://taxes.marylandtaxes.com accessed May 5, 2013

Which ONE of the three dot plots below matches the "2012 Maryland Tax Rates" histogram above? Explain how you determined the correct dot plot.



The correct response is "Plot #1." Plot #3 is eliminated as both its minimum and maximum values are too low (along with other reasons). Plot #2 is eliminated because several of its large values exceed the histogram's maximum possible value.

4. For each of the following five sets of summary measures, indicate if the set of summary measures could match the "2012 Maryland Tax Rates" histogram above. For each set of summary measures that you eliminate, explain why you eliminated that choice.

   a. Mean = 1.01, MAD = 5.4

   b. Median = 2.93, IQR = 0.45

   c. Mean = 3.5, MAD = 1.1

   d. Median = 3.10, IQR = 2.15

   e. Minimum = 1.25, Maximum = 3.20

Options (b) and (e) could match the picture (and option (e) matches the text introducing the context). Options (a) and (c) are eliminated as the mean values of 1.01 and 3.5 are not supported by the histogram (these values are more extreme, respectively, than the minimum and maximum values shown in the histogram). Option (d) is eliminated since 13 of the observations (that's more than half) are less than 3.0, so a median of 3.1 would be too large. (The IQR in option (d) is also too large.)