Lesson 18: Connecting Graphical Representations and Numerical Summaries

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.

Classwork

Example 1: Summary Information from Graphs

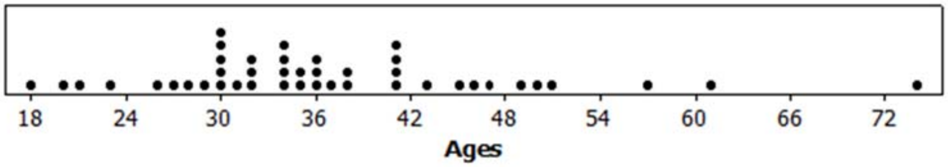
Recall that a *dot plot* includes a dot on a scale or number line for each observation in a data set. Dots are stacked on top of one another when there are multiple occurrences of a data value. Recall also that a *histogram* similarly uses a scale or number line to present the frequency or relative frequency 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.

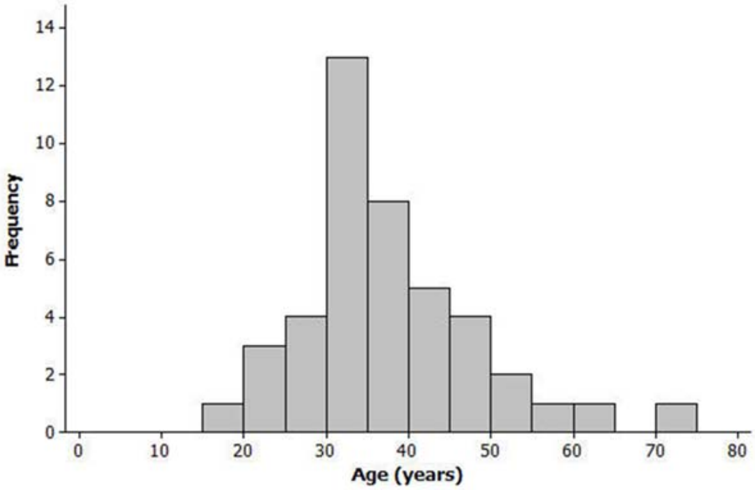
Here is a data set of the ages (in years) of 43 participants in a recent local 5-kilometer race.

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

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?

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

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

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

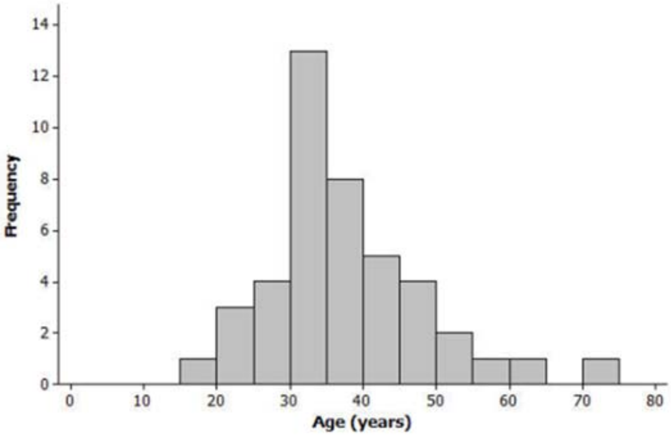
5. Do both the dot plot and the histogram convey information about the variability in the age distribution?

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?

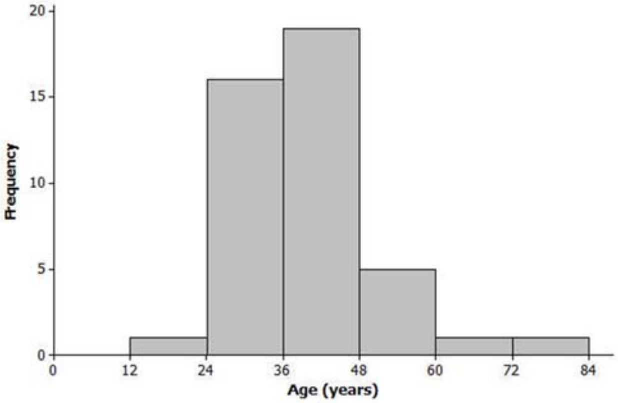
7. Explain why you would only be able to estimate the value of the median if you only had a histogram of the data.

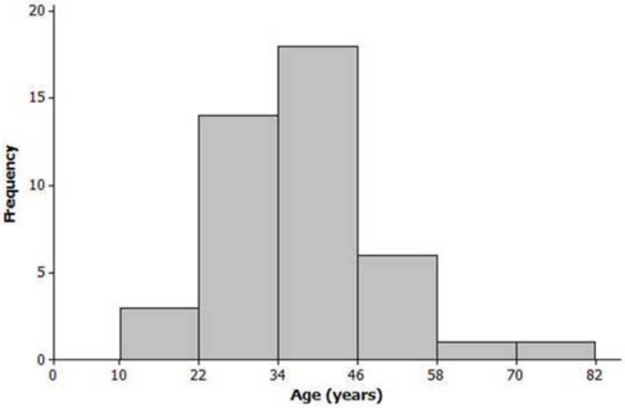
Exercises 8–13: Graphs and Numerical Summaries

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.

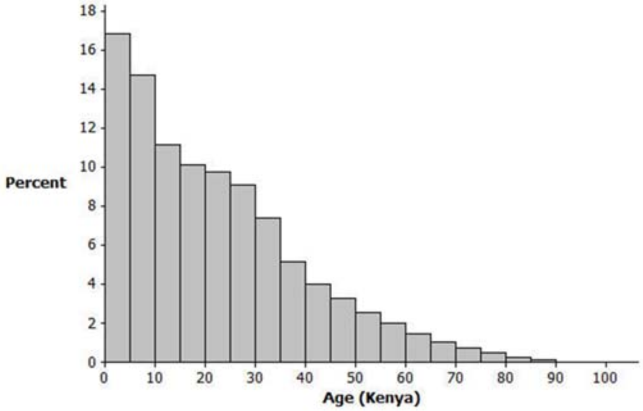


9. One of the histograms below is another valid histogram for the runners' ages. Select the correct histogram, and explain how you determined which graph is valid (and which one is incorrect) based on the summary measures and dot plot.





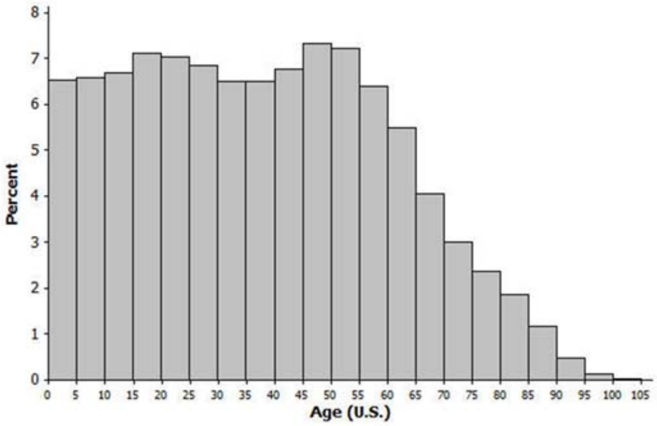
10. The histogram below represents the age distribution of the population of Kenya in 2010.



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?

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.

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?

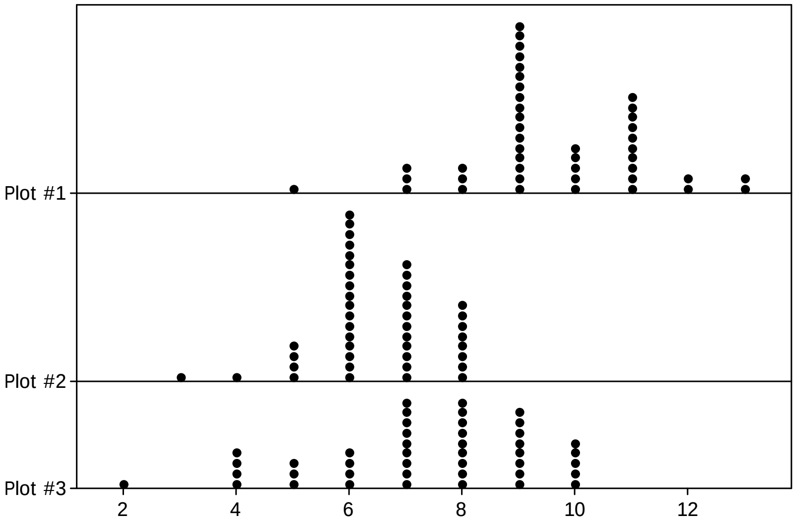


12. Use the histograms from Exercises 10 and 11 to answer the following:

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

    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?

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 # __________

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

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

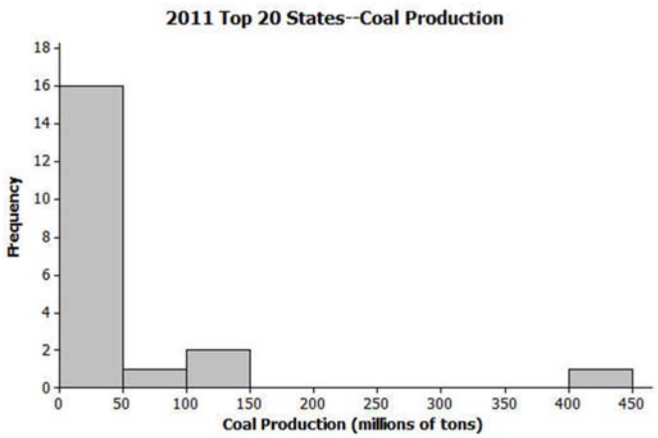
Lesson Summary

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.

Problem Set

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.



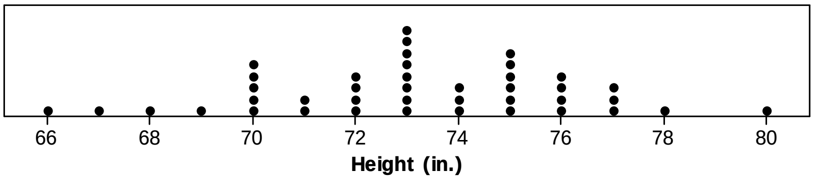
(U.S. Coal Production by State data as reported by the National Mining Association from http://www.nma.org/pdf/c_production_state_rank.pdf accessed May 5, 2013)

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

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.

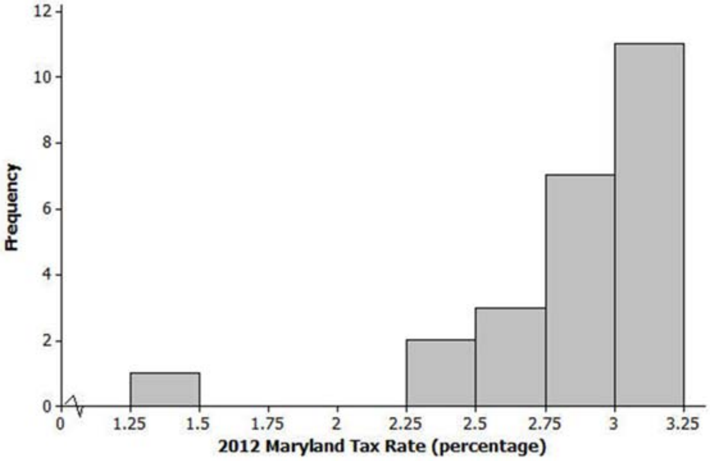


Data Source: http://www.texassports.com/sports/m-swim/mtt/tex-m-swim-mtt.html accessed April 30, 2013

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

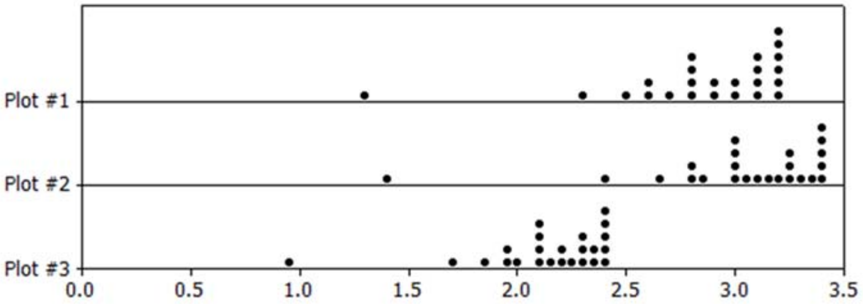
b. Based on this dot plot, make a histogram of the heights using the following classes: 66–< 68 inches, 68–< 70 inches, and so on.

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.



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