

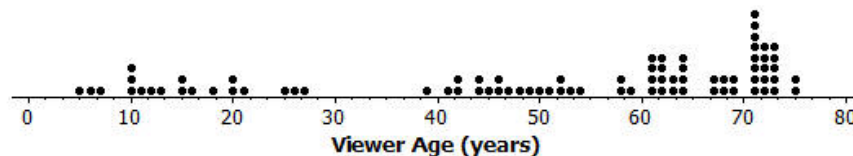
Lesson 7: Measuring Variability for Skewed Distributions (Interquartile Range)

Classwork

Exploratory Challenge 1/Exercises 1–3: Skewed Data and its Measure of Center

Consider the following scenario. A television game show, *Fact or Fiction*, was canceled after nine shows. Many people watched the nine shows and were rather upset when it was taken off the air. A random sample of eighty viewers of the show was selected. Viewers in the sample responded to several questions. The dot plot below shows the distribution of ages of these eighty viewers.

Dot Plot of Viewer Age



1. Approximately where would you locate the mean (balance point) in the above distribution?
2. How does the direction of the tail affect the location of the mean age compared to the median age?

3. The mean age of the above sample is approximately 50. Do you think this age describes the typical viewer of this show? Explain your answer.

Exploratory Challenge 2/Exercises 4–8: Constructing and Interpreting the Box Plot

4. Using the above dot plot, construct a box plot over the dot plot by completing the following steps:
- Locate the middle 40 observations, and draw a box around these values.
 - Calculate the median, and then draw a line in the box at the location of the median.
 - Draw a line that extends from the upper end of the box to the largest observation in the data set.
 - Draw a line that extends from the lower edge of the box to the minimum value in the data set.
5. Recall that the 5 values used to construct the dot plot make up the 5-number summary. What is the 5-number summary for this data set of ages?

Minimum age: _____

Lower quartile or Q1: _____

Median Age: _____

Upper quartile or Q3: _____

Maximum age: _____

6. What percent of the data does the box part of the box plot capture?
7. What percent of the data falls between the minimum value and Q1?
8. What percent of the data falls between Q3 and the maximum value?

Exercises 9–14

An advertising agency researched the ages of viewers most interested in various types of television ads. Consider the following summaries:

Ages	Target Products or Services
30–45	Electronics, home goods, cars
46–55	Financial services, appliances, furniture
56–72	Retirement planning, cruises, health care services

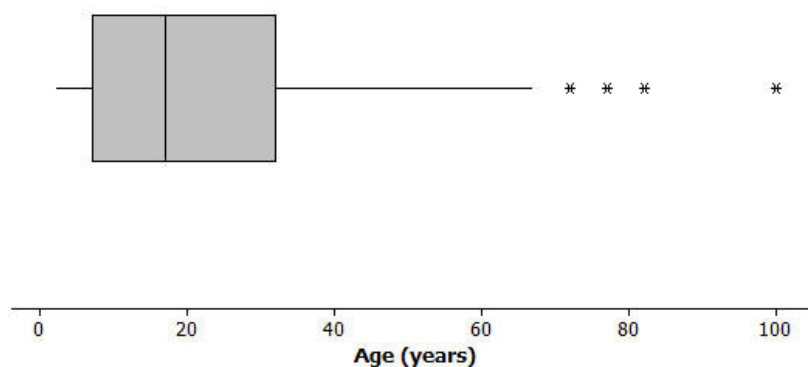
9. The mean age of the people surveyed is approximately 50 years old. As a result, the producers of the show decided to obtain advertisers for a typical viewer of 50 years old. According to the table, what products or services do you think the producers will target? Based on the sample, what percent of the people surveyed about the *Fact or Fiction* show would have been interested in these commercials if the advertising table is accurate?
10. The show failed to generate the interest the advertisers hoped. As a result, they stopped advertising on the show, and the show was cancelled. Kristin made the argument that a better age to describe the typical viewer is the median age. What is the median age of the sample? What products or services does the advertising table suggest for viewers if the median age is considered as a description of the typical viewer?
11. What percent of the people surveyed would be interested in the products or services suggested by the advertising table if the median age were used to describe a typical viewer?
12. What percent of the viewers have ages between Q1 and Q3? The difference between Q3 and Q1, or $Q3 - Q1$, is called the **interquartile range**, or **IQR**. What is the interquartile range (IQR) for this data distribution?

13. The IQR provides a summary of the variability for a skewed data distribution. The IQR is a number that specifies the length of the interval that contains the middle half of the ages of viewers. Do you think producers of the show would prefer a show that has a small or large interquartile range? Explain your answer.
14. Do you agree with Kristin's argument that the median age provides a better description of a typical viewer? Explain your answer.

Exploratory Challenge 3/Exercises 15–20: Outliers

Students at Waldo High School are involved in a special project that involves communicating with people in Kenya. Consider a box plot of the ages of 200 randomly selected people from Kenya.

Box Plot of Ages for Kenya



A data distribution may contain extreme data (specific data values that are unusually large or unusually small relative to the median and the interquartile range). A box plot can be used to display extreme data values that are identified as **outliers**.

Each "*" in the box plot represents the ages of four people from this sample. Based on the sample, these four ages were considered outliers.

15. Estimate the values of the four ages represented by an *.

An outlier is defined to be any data value that is more than $1.5 \times (IQR)$ away from the nearest quartile.

16. What is the median age of the sample of ages from Kenya? What are the approximate values of $Q1$ and $Q3$? What is the approximate IQR of this sample?

17. Multiply the IQR by 1.5. What value do you get?

18. Add $1.5 \times (IQR)$ to the 3rd quartile age ($Q3$). What do you notice about the four ages identified by an *?

19. Are there any age values that are less than $Q1 - 1.5 \times (IQR)$? If so, these ages would also be considered outliers.

20. Explain why there is no * on the low side of the box plot for ages of the people in the sample from Kenya.

Lesson Summary

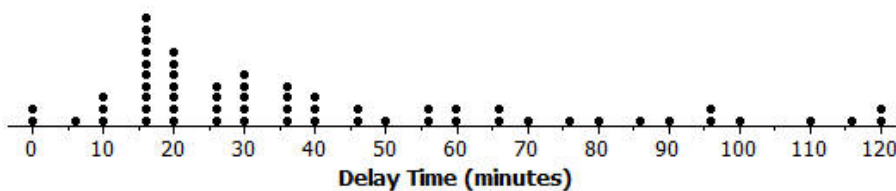
- Non-symmetrical data distributions are referred to as skewed.
- Left-skewed or skewed to the left means the data spreads out longer (like a tail) on the left side.
- Right-skewed or skewed to the right means the data spreads out longer (like a tail) on the right side.
- The center of a skewed data distribution is described by the median.
- Variability of a skewed data distribution is described by the interquartile range (IQR).
- The IQR describes variability by specifying the length of the interval that contains the middle 50% of the data values.
- Outliers in a data set are defined as those values more than $1.5(\text{IQR})$ from the nearest quartile. Outliers are usually identified by an “*” or a “•” in a box plot.

Problem Set

Consider the following scenario. Transportation officials collect data on flight delays (the number of minutes a flight takes off after its scheduled time).

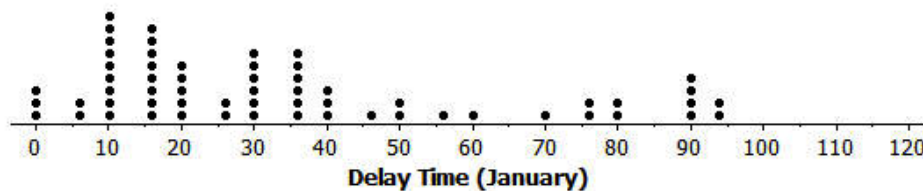
Consider the dot plot of the delay times in minutes for 60 BigAir flights during December 2012:

Dot Plot of December Delay Times



1. How many flights left more than 60 minutes late?
2. Why is this data distribution considered skewed?
3. Is the tail of this data distribution to the right or to the left? How would you describe several of the delay times in the tail?
4. Draw a box plot over the dot plot of the flights for December.

5. What is the interquartile range, or IQR, of this data set?
6. The mean of the 60 flight delays is approximately 42 minutes. Do you think that 42 minutes is typical of the number of minutes a BigAir flight was delayed? Why or why not?
7. Based on the December data, write a brief description of the BigAir flight distribution for December.
8. Calculate the percentage of flights with delays of more than 1 hour. Were there many flight delays of more than 1 hour?
9. BigAir later indicated that there was a flight delay that was not included in the data. The flight not reported was delayed for 48 hours. If you had included that flight delay in the box plot, how would you have represented it? Explain your answer.
10. Consider a dot plot and the box plot of the delay times in minutes for 60 BigAir flights during January 2013. How is the January flight delay distribution different from the one summarizing the December flight delays? In terms of flight delays in January, did BigAir improve, stay the same, or do worse compared to December? Explain your answer.



Box Plot of January Delay Times

