Name_____ Class_____ Date_____

9-2

COMMON CORE

CC.9-12.S.ID.1*,
CC.9-12.S.ID.2*,
CC.9-12.S.ID.3*

Data Distributions and Outliers

Essential question: *Which statistics are most affected by outliers, and what shapes can data distributions have?*

1 EXAMPLE Using Line Plots to Display Data

Twelve employees at a small company make the following annual salaries (in thousands of dollars): 25, 30, 35, 35, 35, 40, 40, 40, 45, 45, 50, 60.

A Create a line plot of the data by putting an X above the number line to represent each data value. Stack the Xs for repeated data values.

B Complete the table. Round to the nearest hundredth, if necessary.

Mean	Median	Range	IQR	Standard deviation

REFLECT

1a. *Quantitative data* are numbers, such as counts or measurements. *Qualitative data* are categories, such as attributes or preferences. For example, employees' salaries are quantitative data while employees' positions within a company are qualitative data. Is it appropriate to use a line plot for displaying quantitative data, qualitative data, or both? Explain.

_____

_____

1b. The line plot allows you to see how the data are distributed. Describe the overall shape of the distribution of employees' salaries.

_____

1c. When you examine the line plot, do any data values appear to be different than the others? Explain.

_____

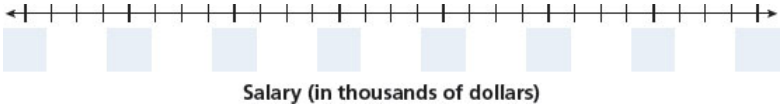
Outliers An **outlier** is a value in a data set that is relatively much greater or much less than most of the other values in the data set. Outliers are determined using either the IQR or the standard deviation. Below is one way to determine whether a data value is an outlier.

Determining Whether a Data Value Is an Outlier
A data value x is an outlier if $x < Q_1 - 1.5(IQR)$ or if $x > Q_3 + 1.5(IQR)$.

2 EXPLORE Investigating the Effect of an Outlier in a Data Set

Suppose the list of salaries in the previous example is expanded to include the owner's salary, which is $150,000. Now the list of salaries is: 25, 30, 35, 35, 35, 40, 40, 40, 45, 45, 50, 60, 150.

A Create a line plot for the revised data set. Choose an appropriate scale for the number line.



B Complete the table. Use a calculator and round to the nearest hundredth, if necessary.

Mean	Median	Range	IQR	Standard deviation

C Complete each sentence by stating whether the statistic increased, decreased, or stayed the same when the data value 150 was added to the original data set. If the statistic increased or decreased, say by what amount.

The mean _____.

The median _____.

The range _____.

The IQR _____.

The standard deviation _____.

REFLECT

2a. Show that the data value 150 is an outlier, but the data value 60 is not. Use the inequalities given at the top of the previous page to support your answer.

_____

_____

2b. What effect does the outlier have on the overall shape of the distribution?

_____

2c. For the original data set, you can conclude that the salary of a typical employee is $40,000 regardless of whether you used the mean or the median. For the revised data set, you could say that the salary of a typical employee is either $48,500 or $40,000 depending on whether you used the mean or the median. Which average salary is more reasonable for the revised data set? Explain your reasoning.

_____

_____

2d. Based on how the IQR and standard deviation are calculated, explain why the IQR was only slightly affected by the addition of the outlier while the standard deviation was dramatically changed.

_____

_____

_____

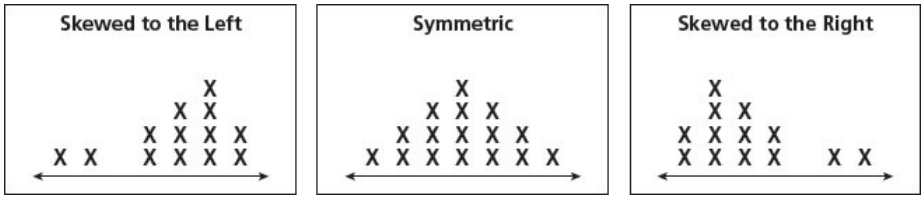
_____

2e. Because the median and the IQR are based on quartiles while the standard deviation is based on the mean, the center and spread of a data set are usually reported either as the median and IQR or as the mean and standard deviation. Which pair of statistics would you use for a data set that includes one or more outliers? Explain.

_____

_____

Shapes of Distributions A data distribution can be described as **symmetric**, **skewed to the left**, or **skewed to the right** depending on the general shape of the distribution in a line plot or other data display.



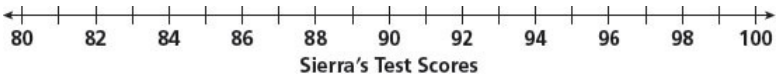
3 EXAMPLE Comparing Data Distributions

The tables list Sierra's and Jacey's scores on math tests in each quarter of the school year. Create a line plot for each student's scores and identify the distribution as symmetric, skewed to the left, or skewed to the right.

Sierra's Scores			
I	II	III	IV
88	86	92	88
94	90	87	91
91	95	94	91
92	91	88	93
90	94	96	89

Jacey's Scores			
I	II	III	IV
89	76	87	82
83	86	86	85
86	87	72	86
83	88	73	88
87	90	84	89

A Create and examine a line plot for Sierra's scores.



Sierra's Test Scores

The distribution is centered on one value (91) with the data values to the left of the center balanced with the data values to the right, so the distribution is symmetric.

B Create and examine a line plot for Jacey's scores.

Jacey's Test Scores

The data values cluster on the right with a few data values spread out to the left of the cluster, so the distribution is skewed to the left.

REFLECT

3a. Find the mean and median for Sierra's test scores. How do they compare?

3b. Will the mean and median in a symmetric distribution always be equal or approximately equal? Explain.

3c. Find the mean and median for Jacey's test scores. How do they compare?

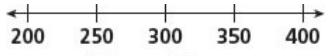
3d. Will the mean and median in a skewed distribution always be different? Explain.

PRACTICE

1. a. Rounded to the nearest $50,000, the values (in thousands of dollars) of homes sold by a realtor are listed below. Use the number line to create a line plot for the data set.

300 250 200 250 350

400 300 250 400 300



Values of Homes
(in thousands of dollars)

b. Suppose the realtor sells a home with a value of $650,000. Which statistics are affected when 650 is included in the data set?

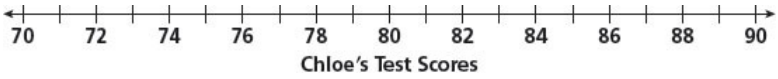
c. Would 650 be considered an outlier? Explain.

2. In Exercise 1, find the mean and median for the data set with and without the data value 650. Why might the realtor want to use the mean instead of the median when advertising the typical value of homes sold?

3. The table shows Chloe's scores on math tests in each quarter of the school year.

Chloe's Scores			
I	**II**	**III**	**IV**
74	77	79	74
78	75	76	77
82	80	74	76
76	75	77	78
85	77	87	85

a. Use the number line below to create a line plot for Chloe's scores.



Chloe's Test Scores

b. Complete the table below for the data set.

Mean	Median	Range	IQR	Standard deviation

c. Identify any outliers in the data set. Which of the statistics from the table above would change if the outliers were removed?

d. Describe the shape of the distribution.

e. Which measure of center and which measure of spread should be used to characterize the data? Explain.

4. Give an example of a data set with a symmetric distribution that also includes one or more outliers.

5. Suppose that a data set has an approximately symmetric distribution, with one outlier. What could you do if you wanted to use the mean and standard deviation to characterize the data?