

## **NUMB3RS Activity: Outliers**

### **Episode: "Take Out"**

**Topic:** Statistics, Outliers

**Grade Level:** 9 - 12

**Objective:** Identify outliers with box-and-whisker plots and scatter plots

**Time:** 30 minutes

**Materials:** TI-83 Plus/TI-84 Plus graphing calculator

#### **Introduction**

In "Take Out," Amita and Charlie help the FBI track financial transfers from the US to Mexico. However, because there is so much traffic, it is difficult to find a specific transfer. Charlie explains that it is "a matter of using a target-specific optimization model. Something called Outlier Detection. We have been doing a brute-force search in which you throw a net across a river and catch everything. Outlier Detection, though, is target-specific. It's like fly-fishing the data stream, choosing your spot by the spawning behavior, selecting the right bait, a method by which we are able to cast our rod exactly where we wanted, and catch exactly the type of fish we needed."

An outlier is a piece of data that is significantly different from the other data in the set. A point that is far away may not be representative and may skew the results. The challenge, however, is to quantify what it means to be "significantly different." The goal of this activity is for students to gain an intuitive sense of how to identify a possible outlier. Analysis of a box-and-whisker plot and a scatter plot are included. Though the activity only shows the scatter plot, in the extensions, students are asked to find the least squares regression line and confidence level.

#### **Discuss with Students**

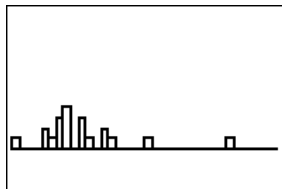
Familiarity with data displays is assumed for this activity, but a brief overview is given here. To enter the data into the calculator, press **[STAT]**, select **1:Edit...**, and enter the values into one of the lists. To create a graph, turn a plot on by pressing **[2nd] [STAT PLOT]** and choose the appropriate type of graph and the appropriate list.

This activity asks for four different types of displays: a histogram (**[2nd] [DRAW]**), a box-and-whisker plot (**[2nd] [DRAW]**), a modified box-and-whisker plot (**[2nd] [DRAW]**), and a scatter plot (**[2nd] [DRAW]**). The box-and-whisker plot does not show outliers, whereas the modified box-and-whisker plot does. In this activity, students calculate outliers with a formula and use the modified box-and-whisker plot to check their answer. One goal is to give students insight into how the calculator determines an outlier. For all of the data displays, to get an appropriate viewing window, press **[ZOOM]** and select **9:ZoomStat**.

Students will perform a linear regression and find the correlation coefficient. To set your calculator to display the correlation coefficient, press  $\boxed{2\text{nd}}$  [CATALOG] and choose **Diagnostics On**.

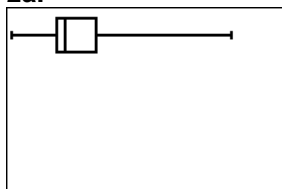
**Student Page Answers:**

**1a.**



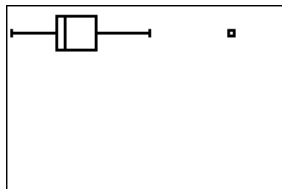
**1b.** Answers will vary, but the two highest values and the lowest value are possible candidates.

**2a.**

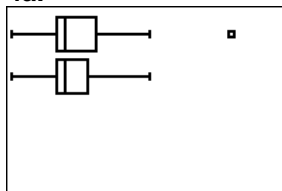


**2b.**  $Q1 = 1,275$  **2c.**  $Q3 = 2,300$  **2d.**  $IQR = 1,025$  **3a.**  $-262.5$  and  $3,837.5$  **3b.** Only  $\$6,000$  is an outlier.

**3c.**

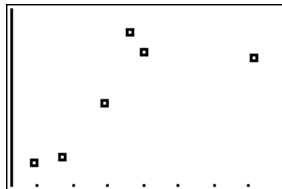


**4a.**



**4b.** original data: mean = 1,928, median = 1,500, mode = 1,500,  $IQR = 1,025$ ; adjusted data: mean = 1,714, median = 1,500, mode = 1,500,  $IQR = 850$  **4c.** The data sets have the same median and mode, but the mean and the  $IQR$  are both smaller for the data with the outlier removed.

**5a.**



Kobe Bryant's data point appears outside of the other data. **5b.** 0.746 **5c.** Kobe Bryant's data is the only point that greatly affects the correlation coefficient.

Name \_\_\_\_\_ Date \_\_\_\_\_


## NUMB3RS Activity: Outliers

In "Take Out," Amita and Charlie help the FBI track financial transfers from the US to Mexico. However, because there is so much traffic, it is difficult to find the specific transfer. Charlie explains that it is "a matter of using a target-specific optimization model. Something called Outlier Detection. We have been doing a brute-force search in which you throw a net across a river and catch everything. Outlier Detection, though, is target-specific. It's like fly-fishing the data stream, choosing your spot by the spawning behavior, selecting the right bait, a method by which we are able to cast our rod exactly where we wanted, and catch exactly the type of fish we needed."


An outlier is a piece of data that is significantly different from the other data in the set. A point that is far away may not be representative and may skew the results. The challenge, however, is to quantify what it means to be "significantly different." Definitions and explorations are given below for two types of analysis: a box-and-whisker plot and a scatter plot.

### Box-and-Whisker Plots

Consider the following data that lists the amounts of twenty different wire transfers: \$50, \$900, \$1,000, \$1,210, \$1,250, \$1,300, \$1,300, \$1,500, \$1,500, \$1,500, \$1,500, \$1,850, \$2,000, \$2,000, \$2,100, \$2,500, \$2,500, \$2,800, \$3,800, \$6,000.

1.
  - a. Use a graphing calculator to construct a histogram for the data. For a good window, press **ZOOM** and select **9:ZoomStat**. To set a good width of the bars for this data, press **WINDOW** and set **Xscl** to a value near 200.
  - b. An outlier is a point that seems significantly far away from other points. Do any points appear to be outliers?
2.
  - a. Change the plot to a box-and-whisker plot (select the  icon on the **STAT PLOTS** editor screen).
  - b. What is the value of Q1? (Use **TRACE** to find the value of Q1.)
  - c. What is the value of Q3?
  - d. Calculate the *interquartile range* (IWR), which is defined as  $Q3 - Q1$ .

A data point that is  $1.5 \times (\text{IQR})$  units above Q3 or  $1.5 \times (\text{IQR})$  units below Q1 is considered an *outlier*. The two values  $Q1 - 1.5(\text{IQR})$  and  $Q3 + 1.5(\text{IQR})$  are called *fences*.

3.
  - a. Find the two fences for the data set.
  - b. Which point(s) are outliers for the data set?
  - c. Verify your answer to 3b by plotting the modified box-and-whisker plot (select the  icon on the **STAT PLOTS** editor screen).

Examine the data with the outlier removed. Copy the data into a new list and remove the outlier.

4. a. Create a modified box-and-whisker plot with the adjusted data set. View both plots on the same screen.
- b. Determine the mean, median, mode and IQR for each data set. These values are all *measures of central tendency*.
- c. How do the two sets of data compare? Which measures of central tendency are the same, and which are different?

### **Scatter Plots**

The data in the chart below shows the numbers of points scored and minutes played by six members of the 2004–05 Los Angeles Lakers. (The data was obtained from the Illuminations website at <http://illuminations.nctm.org/ActivityDetail.aspx?ID=146>.)

Player	Points scored	Minutes played
Kobe Bryant	1,819	2,689
Caron Butler	1,195	2,746
Chucky Atkins	1,115	2,903
Lamar Odom	975	2,320
Chris Mihm	735	1,870
Jumaine Jones	577	1,830

5. a. Enter the data into a graphing calculator and generate a scatter plot. Do any data points appear to be outliers?

Perform a linear regression on the data. Press **[STAT]**, and in the **CALC** menu, select **4:LinReg(ax+b)**. Input (the name of the first list, the name of the second list), then **[ENTER]**. A regression gives the equation of a line that approximates the data. The  $r$  value, known as the *correlation coefficient*, indicates how well the regression equation models the data. The closer  $|r|$  is to 1, the better the fit.

- b. What is the correlation coefficient for the data?

Remove one data point and find the new correlation coefficient. Add that data point back in, remove a different point, and recalculate. Repeat for each data point. A point that affects the correlation coefficient significantly is considered an outlier.

- c. After testing all of the points in the data, which point affects the correlation coefficient the most? Is this the same point you chose in question 5a?

*The goal of this activity is to give your students a short and simple snapshot into a very extensive math topic. TI and NCTM encourage you and your students to learn more about this topic using the extensions provided below and through your own independent research.*

## Extensions

### For the Student

- Removing an outlier affects the value of statistical measurements. When an outlier is identified, a statistician must analyze the circumstances to determine if removing the point from the data set is reasonable. Consider the following activity at <http://illuminations.nctm.org/ActivityDetail.aspx?ID=151>.

Choose "Population Density" from the Data Sets menu, and notice that the map is mostly white, indicating that all of the states are similarly low in value. The District of Columbia has a population density of 8,370 people per square mile, which is significantly higher than any of the 50 states. In this case, it makes sense to remove DC as an outlier, because DC is a city and not a state. After DC is removed, it then appears that New Jersey is an outlier, but is it valid to remove it? Although New Jersey is "city-like" in that its southern area is a suburb of Philadelphia and its northern area is a suburb of New York City, it is still a state and should probably be left in the data set under consideration.

- While many mathematical formulas are universally accepted, there is not agreement on the method to calculate Q1 and Q3 for a box-and-whisker plot. All accepted methods yield similar results, but they can vary slightly for the same data set. Find books, calculators, or programs that use different methods and compare the results for the same data set. Can you create a data set such that a data point is an outlier using one method, but not a different method?
- Repeat the linear regression activity with data from the Detroit Pistons from the 2004–05 season. Go to <http://www.nba.com/pistons/stats/2004/index.html> to find player information. Investigate Ben Wallace as an outlier and consider how his results are different from Kobe Bryant's results.
- There are many different types of regressions and correlations. The linear regression calculated by the TI-83 Plus/TI-84 Plus graphing calculator is the *least squares line*, the equation that minimizes the sum of the square of the distances from each data point to the line. The correlation is the Pearson Product Moment Correlation. For another NUMB3RS activity exploring this correlation, go to <http://education.ti.com/exchange> and search for "7723."

### Additional Resources

- Another example of outliers in data sets can be found at <http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm>.
- To view the effects of adding data points to a scatter plot's regression line, view the interactive applet at <http://www.stat.sc.edu/~west/javahtml/Regression.html>.