Outlier Analyses: Step-by-Step Guide

Authors
Danielle Crain
Chris Lysy

January 2018
Version 1.0
The IDEA Data Center (IDC) created this document under U.S. Department of Education, Office of Special Education Programs Grant No. H373Y130002. Richelle Davis and Meredith Miceli serves as the project officers.

The views expressed herein do not necessarily represent the positions or policies of the U.S. Department of Education. No official endorsement by the U.S. Department of Education of any product, commodity, service, or enterprise mentioned in this publication is intended or should be inferred. This product is in the public domain. Authorization to reproduce it in whole or in part is granted.

For more information about the IDEA Data Center’s work and its partners, see www.ideadata.org.

Suggested Citation:

A Step-by-Step Guide for Completing an Outlier Analysis

“An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism” (Hawkins, 1980). Outliers are also referred to as anomalies, abnormalities, deviants, and discordant in the data mining and statistics fields. Often what constitutes a ‘sufficient’ anomaly is based on subjective judgment. Some anomalies may be embedded in a large amount of “noise” and may not be of interest. Most of the time, the significant deviations are the ones of interest.” See IDEA Data Quality: Outlier Analyses Brief for more information about the importance of outlier analyses and data quality.

The purpose of this document is to demonstrate three ways of completing outlier analyses. It also demonstrates three ways of displaying the analyses. State staff who want to calculate their outlier analyses using their databases and/or programs can use this guide instead of the IDEA Data Quality: Outlier Analysis Tool that provides assistance with automatic calculation of outlier analyses.

There is no single right way to do an outlier analysis. Staff need to choose an approach and be systematic.

The tutorials included in this guide each present different methods staff can use to identify and visualize outliers. Staff should pick the method, or methods, that make the most sense to them.  

Identifying an Outlier

Identifying an outlier that starts with a simple question, “What Is Normal?” The testimonials in this guide can help state staff define what is normal and use data visualization to make identified outliers more noticeable.

A Range of Normal

Outliers are the numbers outside of a range of data considered normal. The following tutorials present different ways of identifying what is normal:

- Tutorial 1: Systematically Determining What Is Normal Using the Interquartile Range
- Tutorial 2: Qualitatively Defining a Normal Range
- Tutorial 3: Simply Sorting

Data Visualization Support for Outlier Analyses

Data visualization can help support outlier analyses by making identified outliers more noticeable. The following tutorials present visual approaches that pair well with any of the first three approaches to calculating outlier analysis.

- Tutorial 4: Heat Maps in Excel
- Tutorial 5: Dot Plots in Excel

1 Sample Part B data were used to illustrate each tutorial, but these approaches will work with Part C data or any other data source.
A Range of Normal

As stated above, outliers are the numbers outside of a range of data that state staff identified as normal. Here are a few ways to define a normal range.

(Note: Within this section, the guide refers to the edges of the normal range as Upper and Lower Fences.)

**Tutorial 1: Systematically Determining What Is Normal Using the Interquartile Range**

Tutorial 1 is the typical statistical approach to defining what is normal. This approach uses an easy-to-calculate “interquartile range” to identify a normal range for a provided series of data. This series could be data from across all local education agencies (LEAs) and local lead agencies (LLAs) in a state for a single measure. Most likely, state staff will analyze a distribution within a single column in Excel.

**Step 1.**

To start, staff will need an Excel workbook with at least two columns of data. The following example uses district-level Indicator B5 data.
**Step 2.**
The first calculations are for the first and third quartiles, using the following formulas in Excel: “=percentile(<Cell Range>,0.25)” and “=percentile(<Cell Range>,0.75)”; staff should just replace <Cell Range> with the range of values, they are checking.

For this example, the formula for the Quartile 1 calculation would be “=PERCENTILE(D2:D949,0.25),” and the Quartile 3 calculation would be “=PERCENTILE(D2:D949,0.75).”

![Excel screenshot](image1)

**Step 3.**
Next, staff should calculate the interquartile range, which is simply Quartile 3 – Quartile 1.

In this example, the formula would be “=D952-D951.”

![Excel screenshot](image2)
**Step 4.**

Using the interquartile range (IQR), staff should come up with a “normal range” by setting up fences. The Lower Fence is the bottom of the range, and the Upper Fence is the top.

Calculate the Lower Fence by subtracting 1.5 times the interquartile range from Quartile 1. [Lower Fence = Quartile 1 – (1.5 * IQR)]

Calculate the Upper Fence by adding 1.5 times the interquartile range to Quartile 3. [Upper Fence = Quartile 3 + (1.5 * IQR)]

In this example, the formula for the Lower Fence would be “=D951-(1.5*D953),” and the Upper Fence would be “=D952+(1.5*D953).”

(Note: This is subjective. To create a larger range (fewer outliers), staff should multiply by a larger number than 1.5 (maybe 2 or 3). To create a smaller range, staff should not use a multiplier at all. If staff do not use a multiplier, the definition of the percentile indicates that staff will identify about half of their data points.)
Step 5.

Once the Lower and Upper Fences are known, state staff can identify outliers. An outlier would be any number falling below the Lower Fence or above the Upper Fence.

In this example, that would be any number less than 25.6 percent and any number over 122.6 percent. The value of 15.38 percent would be considered an outlier because it falls below 25.6 percent.

Note: With data that are widely dispersed, it is likely that the Lower Fence could be a negative number or the Upper Fence could fall outside of the possible range of values. Staff could interpret a negative Lower Fence as having no Lower Fence with no low outliers.
Tutorial 2: Qualitatively Defining a Normal Range

Another way to identify outliers would be to use qualitative information. State staff can identify the numbers that seem out of place given what they know about their state’s data, or they can ask colleagues about their own expectations based on experience. Staff should determine a range that seems to make sense, then look at the data.

Turning to the last example, let’s say that after speaking with colleagues, staff determined that districts should be looked at further if they had any less than 50 percent for Indicator B5A. Staff should just go ahead and identify those cases.
Tutorial 3: Simply Sorting

Sometimes the easiest way to do an analysis would be to simply sort the data. The goal of an outlier analysis is not to support statistical analyses but, rather, to identify potential data issues. There is no harm in singling out data that staff would usually consider part of a normal range.

Step 1.
Staff should highlight their range of data.

Step 2.
Staff should sort by the measure.
**Step 3.**

Staff should look deeper into the values at the top and bottom of the sort.
Data Visualization Support for Outlier Analyses

Data visualization can help support outlier analyses by making identified outliers more noticeable. Here are a few ways state staff can use data visualization to support their outlier analysis.

Tutorial 4: Heat Maps in Excel

By using conditional formatting in conjunction with the approaches shown previously, staff can make outliers stand out. This can be particularly important when there are multiple columns of data or many rows.

Step 1.

Staff should highlight the range of data and click on the Conditional Formatting button in the Home tab.
Step 2.
If staff determined the normal range $w$ by using the interquartile range, they should select **Highlight Cells Rules**, then select **Less Than**. In the box that asks for a number, they should enter the cell location of the Lower Fence value. They can also adjust the format (color/font/additional symbols).

Staff should do the same for the Upper Fence but use **Greater Than**.

Step 3.
If staff determined the normal range using a qualitative approach, they should follow the same instructions as in Step 2. They should either enter the number directly into the **LESS THAN** box or place the Lower Fence value into a cell in the worksheet and refer to that.
Alternative Quick Heat Maps

A quicker approach to heat maps in Excel is using the Color Scales feature found under the Conditional Formatting drop-down menu. This will automatically color a set of selected cells based on the range of values. Staff also can use data bars and icon sets to quickly identify possible outliers.
**Tutorial 5: Dot Plots in Excel**

Another approach to identifying outliers involves visualizing the data. An easy approach is using an in-cell formula that will create a simple dot plot next to the data. Once the formula creates the plot, staff simply look for data points that do not seem to fit with the others.

By using an Excel formula to create an in-cell chart, dot plots will always remain in line with the data. Excel’s standard chart functions also can assist in identifying outliers but the standard chart functions will ultimately be disconnected from the original data.

**Step 1.**

This example uses Excel’s REPT function to create this visual. Staff should start by selecting a cell next to the first data point. Basically, staff is instructing Excel to put in a series of repeating blanks based on the data and then adding some kind of character at the end.

In this example, the formula looks like this:

```
=REPT(“ “,B2*50)&“l”
```

In order to get the dot at the end, staff should use the Wingdings font. The *50 multiplier creates the spacing. If staff desire less spacing, they should use a smaller multiplier.
**Step 2.**

Staff then just need to copy the formula down once they are happy with the look of the first cell. That is all that it takes!

![Excel screenshot](image)

**Step 3.**

It is easy to change the font or change the character and get a different look. Here is an alternative using the same exact formula with the Arial Black font in 10 point instead of Wingdings. The font change also changes the spacing, so staff should set up cells within the same column using the same font.

![Excel screenshot](image)
Tutorial 6: Dot Plots in Tableau

Using an interactive visualization program like Tableau can be useful if there is a large amount of data. Such a program allows the user to quickly and easily visualize hundreds of rows and multiple measures. If visualizing public data, state staff can use Tableau Public for free. For private data, staff would need at least a personal version of the Tableau desktop license.

**Step 1.**
Staff should open the Excel data in Tableau. The measures staff use should be under the *Measures* space in Tableau, and descriptive variables, such as district, should be under *Dimensions*. If this is not the case, staff should right-click on the variable and convert.
**Step 2.**  
Staff should place the measure variable in the *Columns* box and the district variable in the *Rows* box. The default should be a bar graph. If satisfied with the bar chart, staff stop here.

**Step 3.**  
Click on the *Show Me* tab and select the bar chart.
Step 4.
To change to a dot plot, staff should just change the Marks from Automatic to Circle.

Step 5.
To create a line to identify the Lower and Upper Fences, staff should just right-click on the x axis and select Add Reference Line, Band, or Box. Then in the Value box, select Create New Parameter.
**Step 6.**
Staff should change the name of the parameter to “Fence” and set the current value as the Lower Fence Value and click **OK**. Then, staff should click **Apply** and **OK**.

**Step 7.**
To include both the Upper and Lower Fences, staff should just follow the same steps a second time.

Staff may also opt to use the **Parameters** box that the program created on the screen to shift the Fence based on what they are trying to see.
Conclusion

This guide includes a handful of approaches state agency staff can use to identify and visualize outliers. There are many others. Staff should find a set of approaches that works well for their state and their data and then apply the approaches systematically.

State staff should visit ideadata.org to get in touch with their IDC State Liaison if they have any questions about these approaches.