Outlier Analyses
Step-by-Step Guide

Authors: Danielle Crain, Chris Lysy

January 2020
The IDEA Data Center (IDC) created this publication under U.S. Department of Education, Office of Special Education Programs grant number H373Y190001. Richelle Davis and Rebecca Smith serve 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 IDC’s work and its partners, see www.ideadata.org.

Suggested Citation:
Outlier Analyses: Step-by-Step Guide

Purpose and Intended Audience

Outlier analysis provides an important tool for examining data to identify observations [among local education agencies (LEAs), schools, students] with data that deviate from an established norm so that states can investigate the observations as possible data errors. This guide introduces the principles of outlier analysis and includes six tutorials on completing an outlier analysis. It is a companion for the IDEA Data Quality: Outlier Analyses Tool, an Excel-based tool states can use to identify outliers using the interquartile range approach described in the step-by-step tutorials. IDC designed these two products—the outlier analyses tool and this step-by-step guide—for state personnel responsible for the IDEA 618 and/or 616 data to use. IDEA Part B state staff working with LEAs also can use the products to analyze their local data. Any state staff with the ability to examine and analyze IDEA 618 and/or 616 data also can benefit from these technical assistance (TA) products. Such staff include data managers, Information Technology (IT) personnel, coordinators, and directors.

This guide provides an overview of outlier analysis. It is organized around four questions:

1. **What is an outlier?**
2. **Why is outlier analysis important for data validity and reliability?**
3. **What action should states take after conducting an outlier analysis?**
4. **How can states conduct and display an outlier analysis?**

**Question 1: What Is an Outlier?**

“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). Outlier analysis also includes investigating whether the data are valid or invalid. In the fields of statistics and data mining, outliers may be referred to as anomalies, abnormalities, deviants, and discordant data. When conducting an outlier analysis, states define what value or combination of values are outside the expected norm. These parameters can help specify the LEAs that have data outside of the “normal” parameters the state sets. There may be times when states see differences in LEAs...
that the state considers normal. It is up to the states to determine what constitutes a “sufficient” anomaly.

Outlier analysis may identify valid as well as invalid data. Invalid outliers are the target of outlier analysis, as they represent errors in the data. On the other hand, valid outliers may appear to be outside the norm, but investigation demonstrates that the data are not in error. Valid outliers may occur due to random variation, which occurs due to chance and is inherent in a system.

**Question 2: Why Is Outlier Analysis Important for Data Validity and Reliability?**

Outlier analysis is primarily important because it helps to identify errors in the data, which, when investigated, may reveal systematic errors in data collection, coding, or entry. Invalid outliers should be corrected, and the processes that resulted in such errors should be fixed.

Outlier analysis also can be important because it may identify LEAs that are performing better or worse than the norm. Identifying these high or low performers provides opportunities for understanding the factors behind high performance or providing targeted TA where it is needed.

**Question 3: What Action Should States Take After Conducting an Outlier Analysis?**

After conducting an outlier analysis, states should investigate any identified outliers to understand why the data are so different from the norm. If the data are outside the parameters states set for valid outliers, then states should follow up with the LEAs to determine the root cause of the outlying data. Questions to focus outlier investigations follow.

1. **Are the outliers found in just one LEA?**

Knowing that all outliers are found in just one LEA can help state personnel focus their investigation into the cause of outliers.

2. **Are the same LEAs identified with outliers in more than one data submission**

State personnel may want to review LEAs that have outlier data in more than one data submission. The outliers may indicate a need for the LEA to review the data entry or coding policies. They also may indicate that the LEA lacks understanding of the data that the data submissions require.

3. **Are multiple outliers commonly identified in the same LEAs?**

If an outlier analysis reveals multiple outliers commonly identified in the same LEAs, state personnel may want to review the similarities in demographics or data collection practices in these LEAs or both.

---

1 The Office of Elementary and Secondary Education (OESE) compiled a list of possible causes of data quality problems related to the *Elementary and Secondary Education Act* (ESEA) and other data reporting. States can review the list to help determine areas of data quality that the outlier analysis affects (U.S. Department of Education 2006).
4. Are the LEAs with outliers using non-standard data collection definitions?

State personnel may want to review the definitions LEAs use to ensure that, within the state context, the LEAs understand and use the definitions the Office of Special Education Programs (OSEP) provides for the IDEA 618 data collections. For example, outliers in the discipline data could be due to an LEA’s interpretation of the terms “suspension” and “expulsion.”

5. Are the LEAs with outliers using non-standard methods for aggregating the data?

States that collect aggregated data from LEAs may want to review the methods LEAs use to aggregate student-level data to create totals. Inconsistencies in how LEAs aggregate data could lead to outliers. For example, the state education agency (SEA) may want to review the methods LEAs use to aggregate the number of students with disabilities by race/ethnicity to ensure that the LEAs appropriately count each child in the seven categories OSEP requires.

6. Are the LEAs with outliers using non-standard methods to collect the data?

State personnel may want to review whether LEAs are using similar, standard policies and procedures for collecting the data. For example, this can include ensuring that all LEAs use the U.S. Department of Education’s race/ethnicity guidelines. It also could include ensuring all locals define suspension and expulsion consistently.

7. Did the small n-size affect the analysis?

States can analyze n-size in two different ways. The n-size can skew the results of the analysis. If an LEA has a small population, then it may lead to outliers because of the proportion of the population to the rest of the state’s LEAs. To investigate if this scenario has occurred, users would disaggregate the data and determine if they are sensible based on the local small populations. The second scenario can lead to outliers when reporting IDEA data. States can review the n-size used to calculate the Annual Performance Report (APR) for Part B Indicator B4 around discipline and significant discrepancy to determine if outliers have skewed the results of the analyses.

For more information on examining root cause, states can review Equity, Inclusion, and Opportunity: Addressing Success Gaps, White Paper.

Question 4: How Can States Conduct and Display an Outlier Analysis?

States can use several possible approaches to conduct an outlier analysis. The IDEA Data Quality: Outlier Analyses Tool provides assistance with automatic calculation of an outlier analysis. State staff who want to calculate their outlier analysis using their databases and/or programs can use this step-by-step guide instead.

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

This guide provides six different tutorials covering different methods states can use to identify and visualize outliers. Five of the six tutorials explain approaches using Microsoft Excel. The sixth approach uses Tableau to visualize outliers. Staff should pick the method, or methods, that make the most sense to them.

Identifying an Outlier

Identifying an outlier starts with a simple question, “What Is Normal?” The six tutorials in this guide can help state staff define what is normal and use data visualization to make identified outliers more noticeable.
Outliers are the numbers outside of a range of data considered normal. The following three 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 can help support outlier analyses by making identified outliers more noticeable. The following three 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
- Tutorial 6: Dot Plots in Tableau

A Range of Normal

As stated previously, 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 LEAs 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)."

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."
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. \[ \text{Lower Fence} = \text{Quartile 1} - (1.5 \times \text{IQR}) \]

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

In this example, the formula for the Lower Fence would be “=D951-(1.5*D953),” and the formula for 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. Staff would consider the value of 15.38 percent 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 they should look further at districts if the districts had any less than 50 percent for Indicator B5A. Staff should just go ahead and identify those cases.
Outlier Analyses: Step-by-Step Guide

Tutorial 3: Simply Sorting

Sometimes the easiest way to do an analysis would be simply to 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.

![Excel screenshot showing Step 1 of Simply Sorting]

Step 2.

Staff should sort by the measure.

![Excel screenshot showing Step 2 of Simply Sorting]
**Step 3.**

Staff should look more deeply 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 analyses.

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 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 also can adjust the format (color/font/additional symbols).

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

![Image of Excel sheet with highlighted cells](image1.png)

**Step 3.**

If staff determined the normal range using a qualitative approach, they should follow the same instructions as in Step 2. They should 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.

![Image of Excel sheet with highlighted cells](image2.png)
Alternative Quick Heat Maps

A quicker approach to using 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 identify possible outliers quickly.
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 always will remain in line with the data. Excel’s standard chart functions also can assist in identifying outliers but, ultimately, the standard chart functions will 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 are instructing Excel to put in a series of repeating blanks based on the data and then add 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!

![Spreadsheet image](image1.png)

**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.

![Spreadsheet image](image2.png)
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 visualize hundreds of rows and multiple measures quickly and easily. 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 should stop here.

![Diagram showing step 2](image)

**Step 3.**

Staff should click on the *Show Me* tab and select the bar chart.

![Diagram showing step 3](image)
**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 also may 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 introduces the principles of outlier analysis and 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 www.ideadata.org to get in touch with their IDC State Liaison if they have any questions about these approaches.

References
