

Outliers and Influential Points

by Sophia



WHAT'S COVERED

This tutorial is going to teach you about outliers and influential points. Our discussion breaks down as follows:

1. Outliers
2. Influential Points

1. Outliers

You may recall the term "outliers" when talking about univariate data. However, in bivariate data, **outliers** are a little bit different.

An outlier is any point that deviates substantially from the overall form of the remainder of the data points.

↪ **EXAMPLE** Let's take a look at these two data sets. One thing that you might realize is that the ones on the left seem quite random, whereas in the ones on the right, all the x's except one are 8, which might be a clue to something.

Table 1

x	y
10	746
8	677
13	1274
9	711
11	781
14	884
6	608
4	539

Table 2

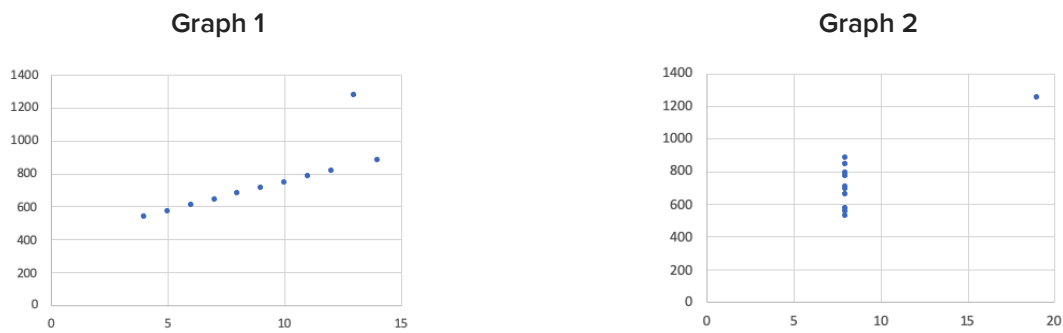
x	y
8	658
8	576
8	771
8	884
8	847
8	704
8	525
19	1,250

12	815
7	642
5	573

8	556
8	791
8	689

However, if you calculate the mean and standard deviation, you will find that they have the same mean for the x's, the same mean for the y's, the same standard deviation for the x's and the same standard deviation for the y's. Also, their correlations are the same at 0.816 in a positive direction.

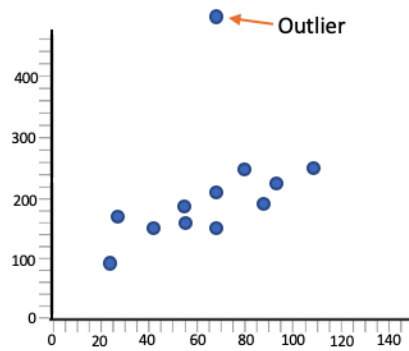
Based on that information, one would think that these two graphs will look fairly similar. Let's take a look:



Both graphs have an outlier that does not follow the overall trend of the graph. Depending on the pattern, the outlier could be an extreme x-value, an extreme y-value, extreme for both the x- and y-values, or neither.

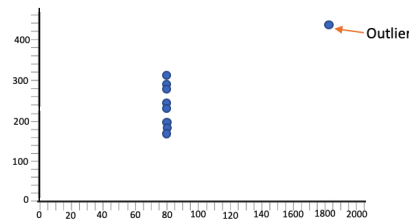
Types of Outliers	Example
Extreme x-values	This is an outlier in the x-direction because it's so much further to the right of the other pack of points but not in the y-direction. If you look horizontally, it's sort of in the middle lower part of the y-direction. It's an outlier in the x-direction but not the y-direction.

Extreme y-values



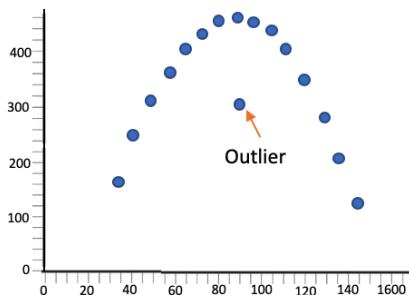
This is an outlier in the y-direction because it's so much higher than the other y-direction, but not the x-direction.

Extreme x- and y-values



This is an outlier in both the x- and y-direction because it's so much further to the right and also higher than the rest of the points.

Neither extreme x- or y-values



Even though it is not extreme in either the x- or y-direction, it doesn't fit the overall trend established by the rest of the data.



TERM TO KNOW

Outlier

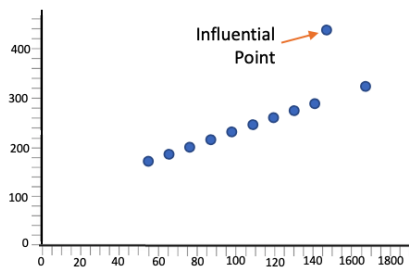
Points that deviate substantially from the overall form of the remainder of the data points.

2. Influential Points

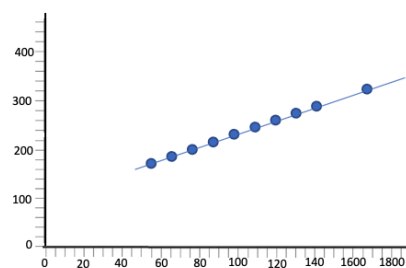
Influential points are points that, if removed, significantly changes a statistical measure. Usually, the measure that we're talking about changing is correlation, but it could also affect other measurements such as the mean of x or y and the standard deviation of x or y.

Some outliers are influential, and some are not.

⇒ **EXAMPLE** When the scatterplot on the left includes the outlier, the correlation coefficient is 0.816. However, when we remove the outlier, the correlation coefficient changes to 1. Since this dramatically changes the correlation, this outlier would be considered an influential point.

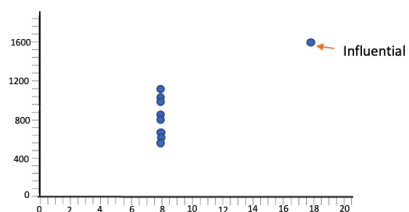


With outlier:
 $r = 0.816$

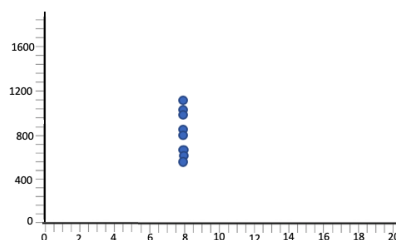


Without outlier:
 $r = 1$

⇒ **EXAMPLE** When the scatterplot below includes the outlier, the mean of x is 9, the standard deviation of x is 3.3, and the correlation is 0.816. However, when we remove the outlier, the mean becomes 8 because now all the x -values are 8, the standard deviation is 0 because they never deviate from 8, and the correlation is 0. Therefore, it changes all of these measures very substantially by being there. That outlier is certainly influential.



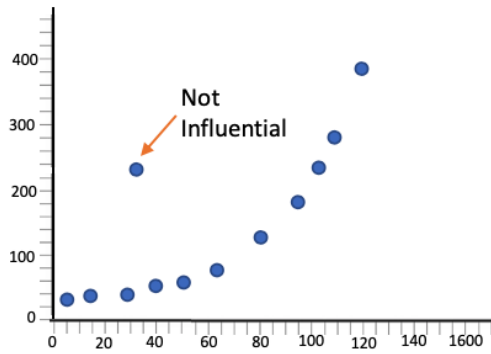
With outlier:
mean = 9
standard deviation = 3.3
 $r = 0.816$



Without outlier:
mean = 8
standard deviation = 0
 $r = 0$

⇒ **EXAMPLE** The outlier in the scatterplot below is not going to have a great effect on the correlation or the least squares regression line that these data sets create. In this case, a line is an inappropriate model, but if you did make a line, having this point versus removing this point wouldn't affect that line or the

correlation very much.



TERM TO KNOW

Influential Points

An observation that, if removed, significantly changes a statistical measure



SUMMARY

Important points on a scatterplot are influential points and outliers. Influential points substantially change at least one statistical measure. Outliers simply are points that deviate from the overall form of the rest of the points. They may be outliers in the x- or y-direction, but don't have to be, according to this definition. Be aware that different people use different definitions of outliers for scatterplots, so there's not one hard-and-fast definition.

Good luck!

Source: THIS TUTORIAL WAS AUTHORED BY JONATHAN OSTERS FOR SOPHIA LEARNING. PLEASE SEE OUR [TERMS OF USE](#).



TERMS TO KNOW

Influential Point

An observation that, if removed, significantly changes a statistical measure.

Outlier

In a scatter plot, an outlier is an observation that has an extreme x value, an extreme y value, both an extreme x and y, or is well away from the main trend of points.