Anomalies

Leland Wilkinson
VP, Statistics
Tableau
lwilkinson@tableau.com

Data Meets Viz
Augsburg, Germany
June 28-30, 2015
Outliers

“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)

All the existing methods in statistics and machine learning packages for detecting outliers based on the mean and standard deviation of a distribution are wrong.

That is because, as $n$ increases, critical value of alpha must change in order to prevent false positives.

But picking alpha for a given $n$ makes detection of outliers circular.

Multivariate outlier detection problem is even harder.

Curse of dimensionality means interpoint distances tend toward a constant as $n$ held constant and $p$ heads toward infinity.

Many graphical methods aren’t much better.
Outliers

Don’t bother to Google
You’ll get this…

Outliers

The Story of Success

MALCOLM GLADWELL

#1 bestselling author of The Tipping Point and Blink
Outliers

What you will find if you persist

There are two popular tests
Both depend on a normal distribution
Both fail to offer protection for large samples

Grubbs (1950)

For the two-sided test, the hypothesis of no outliers is rejected if
\[ G > \frac{(N-1)}{\sqrt{N}} \left( \frac{t^2_{\alpha/2(N-2)}}{N-2 + t^2_{\alpha/2(N-2)}} \right) \]

with \( t_{\alpha/2(N-2)} \) denoting the critical value of the t-distribution with \( (N-2) \) degrees of freedom and a significance level of \( \alpha/2(N) \).

For the one-sided tests, we use a significance level of \( \alpha/N \).

In the above formulas for the critical regions, the Handbook follows the convention that \( t^{+} \) is the upper critical value from the t-distribution and \( t^{-} \) is the lower critical value from the t-distribution. Note that this is the opposite of what is used in some texts and software programs. In particular, Dataplot uses the opposite convention.

Tukey (1977)

The IQR and Outliers

- The IQR is short for “Interquartile Range”
- To calculate IQR, IQR = \( Q_3 - Q_1 \)
- Outliers are calculated using the IQR.
- The rule for outliers is that if a value is outside 1.5(IQR) then it is an outlier.
- So, if a value is more than \( Q_3 + 1.5(IQR) \) or less than \( Q_1 - 1.5(IQR) \) then it is an outlier.
Outliers

Why distance from location (mean, median, ...) is wrong

Remember Hawkins’ definition

“…arouse suspicions that it was generated by a different mechanism”

Wouldn’t you be inclined to say the one on the left is an outlier but not the right?

The two samples have the same mean and standard deviation.

So, the problem boils down to gaps, not distance from center

Dixon (1951)

\[ Q = \frac{\text{gap}}{\text{range}} \]

Tukey-Wainer-Schacht (1978)

\[ z_i = \frac{\sqrt{w_i g_i}}{\text{midmean}(y)} , \text{ where} \]

\[ w_i = i(n - i) \]
Outliers

Graphical methods

Probability plot is one of the best, IF you know the distribution
Outliers

Graphical methods

Box plots depend on normal distribution – useless for large $n$
See how many box plot outliers there are for $n = 100,000$?
Letter value box plots (Hofmann, Kafadar, Wickham, 2006) better

$n = 100,000$

$n = 100$
Outliers

Transformations affect outlier detection
For skewed batches, need to do transformations before testing for outliers
Other Anomalies

Skewness and Kurtosis
Use L-moments (based on weighted sums)
More robust (no third or fourth powers)

Spikes
Use dot plots
Check for stacks
Signal for Zero Inflated Poisson (ZIP) or other models

Multimodality
Smooth with a kernel
Do bump hunting by computing slope of tangent
Outliers

Multivariate Outliers

Mahalanobis Distance is most popular method
OK if you know distribution is multivariate normal
But estimate of covariance matrix can be unreliable when $p$ is large
Outliers

Multivariate Outliers

Minimum Spanning Tree

Compute MST and look for nodes having extremely long edges
Outliers

Multivariate Outliers

Clustering (k-means)

1. Choose very large $k$
2. Initialize $k$ centroids
3. Assign every point $y$ to nearest centroid (squared Euclidean distance)
4. Compute $SSW$
5. Repeat 3 and 4 until $SSW$ does not improve (get smaller)

- On each iteration, use outlier algorithm to decide if a distance to a centroid is beyond cutoff
- If so, leave point out of centroid
- Omitted points are outliers
Outliers

Multivariate Outliers

K-nearest neighbors (KNN)

Choose $k$

Compute KNN graph

For each vertex (point), compute distance to $k$th nearest neighbor

Use unidimensional outlier algorithm on these distances
Outliers

Multivariate Outliers

Parallel Coordinates

Useless for outlier detection

Notice, outlier profile at bottom (computed with MST) is mostly in the middle
Outliers

Multivariate Outliers


Original Data Matrix

Scagnostics Transform

Scagnostics Matrix

For each pair of columns in $X$, we compute 9 measures

$q = \frac{p(p-1)}{2}$
Outliers

Multivariate Outliers

Detecting outlying scatterplots by cluster analyzing scagnostics matrix
Compute scagnostics matrix and then cluster it
Use cluster outlier method to detect outlying scatterplots

Notice the plot in the upper left is an outlier even though it looks bivariate normal
Outliers

Multivariate Outliers

Scagnostics

Ladder of powers transformations reveal different scagnostics under different transformations
Inliers

Hidden rounding to integers

Histograms hide details

Stem-and-leaf and dot plots do not

In this batch, someone rounded some heights of baseball players to nearest inch
Inliers

Duplicate or near-duplicate cases
Multivariate sort
Pick delta profile distance (Euclidean or other distance metric)
Flag cases closer than delta
Duplicate cases found in some Iris datasets with this method
## Fraud

![Graph showing number of correct answers for different conditions](image)

**Smeeters & Liu (2011) JEXP**  
Thanks to Uri Simonsohn and Richard Gill
Fraud

Contrast black vs. white bars over each level

If $F$ value is very small, get suspicious

Because between-groups variation is too small relative to within

The *lower* tail of the $F$ distribution yields our $p$ value

For the Smeeters and Liu article, $F$ was so small that it revealed fraud

Smeeters was forced to resign from the university
Aggregation

Here is a plot of Birth Rate by Death Rate
Aggregated over 80 countries by Leader type
Aggregation

But when we disaggregate by Leader, the picture changes

This anomaly is called Simpson’s Paradox
Aggregation

Let’s see how these methods work real data...
Anomalies

Thank You!

Leland Wilkinson
VP, Statistics
Tableau
lwilkinson@tableau.com