Outlier

From Wikipedia, the free encyclopedia

In statistics, an **outlier** is an observation point that is distant from other observations.[1][2] An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set.[3]

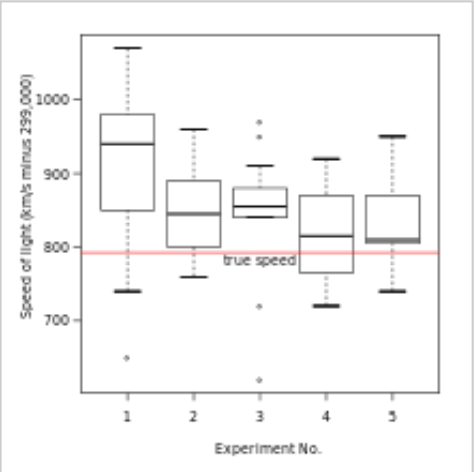

Figure 1. Box plot of data from the Michelson–Morley experiment displaying four outliers in the middle column, as well as one outlier in the first column.

Outliers can occur by chance in any distribution, but they often indicate either measurement error or that the population has a heavy-tailed distribution. In the former case one wishes to discard them or use statistics that are robust to outliers, while in the latter case they indicate that the distribution has high skewness and that one should be very cautious in using tools or intuitions that assume a normal distribution. A frequent cause of outliers is a mixture of two distributions, which may be two distinct sub-populations, or may indicate 'correct trial' versus 'measurement error'; this is modeled by a mixture model.

In most larger samplings of data, some data points will be further away from the sample mean than what is deemed reasonable. This can be due to incidental systematic error or flaws in the theory that generated an assumed family of probability distributions, or it may be that some observations are far from the center of the data. Outlier points can therefore indicate faulty data, erroneous procedures, or areas where a certain theory might not be valid. However, in large samples, a small number of outliers is to be expected (and not due to any anomalous condition).

Outliers, being the most extreme observations, may include the sample maximum or sample minimum, or both, depending on whether they are extremely high or low. However, the sample maximum and minimum are not always outliers because they may not be unusually far from other observations.

Naive interpretation of statistics derived from data sets that include outliers may be misleading. For example, if one is calculating the average temperature of 10 objects in a room, and nine of them are between 20 and 25 degrees Celsius, but an oven is at 175 °C, the median of the data will be between 20 and 25 °C but the mean temperature will be between 35.5 and 40 °C. In this case, the median better reflects the temperature of a randomly sampled object than the mean; naively interpreting the mean as "a typical sample", equivalent to the median, is incorrect. As illustrated in this case, outliers may indicate data points that belong to a different population than the rest of the sample set.

Estimators capable of coping with outliers are said to be robust: the median is a robust statistic, while the mean is not.[4]

Contents

Occurrence and causes

In the case of normally distributed data, the three sigma rule means that roughly 1 in 22 observations will differ by twice the standard deviation or more from the mean, and 1 in 370 will deviate by three times the standard deviation.[5] In a sample of 1000 observations, the presence of up to five observations deviating from the mean by more than three times the standard deviation is within the range of what can be expected, being less than twice the expected number and hence within 1 standard deviation of the expected number – see Poisson distribution – and not indicate an anomaly. If the sample size is only 100, however, just three such outliers are already reason for concern, being more than 11 times the expected number.

In general, if the nature of the population distribution is known a priori, it is possible to test if the number of outliers deviate significantly from what can be expected: for a given cutoff (so samples fall beyond the cutoff with probability p) of a given distribution, the number of outliers will follow a binomial distribution with parameter p, which can generally be well-approximated by the Poisson distribution with $\lambda = pn$. Thus if one takes a normal distribution with cutoff 3 standard deviations from the mean, p is approximately 0.3%, and thus for 1000 trials one can approximate the number of samples whose deviation exceeds 3 sigmas by a Poisson distribution with $\lambda = 3$.

Causes

Outliers can have many anomalous causes. A physical apparatus for taking measurements may have suffered a transient malfunction. There may have been an error in data transmission or transcription. Outliers arise due to changes in system behaviour, fraudulent behaviour, human error, instrument error or simply through natural deviations in populations. A sample may have been contaminated with elements from outside the population being examined. Alternatively, an outlier could be the result of a flaw in the assumed theory, calling for further investigation by the researcher. Additionally, the pathological appearance of outliers of a certain form appears in a variety of datasets, indicating that the causative mechanism for the data might differ at the extreme end (King effect).

Identifying outliers

There is no rigid mathematical definition of what constitutes an outlier; determining whether or not an observation is an outlier is ultimately a subjective exercise. There are various methods of outlier detection.[6][7][8] Some are graphical such as normal probability plots. Others are model-based. Box plots are a hybrid.

Model-based methods which are commonly used for identification assume that the data are from a normal distribution, and identify observations which are deemed "unlikely" based on mean and standard deviation:

- Chauvenet's criterion
- Grubbs' test for outliers
- MMS[9] - test for outliers in linear regression
- Peirce's criterion[10][11][12][13]

> It is proposed to determine in a series of m observations the limit of error, beyond which all observations involving so great an error may be rejected, provided there are as many as n such observations. The principle upon which it is proposed to solve this problem is, that the proposed observations should be rejected when the probability of the system of errors obtained by retaining them is less than that of the system of errors obtained by their rejection multiplied by the probability of making so many, and no more, abnormal observations. (Quoted in the editorial note on page 516 to Peirce (1982 edition) from *A Manual of Astronomy* 2:558 by Chauvenet.)

- Dixon's Q test
- ASTM E178 Standard Practice for Dealing With Outlying Observations
- Mahalanobis distance and leverage are often used to detect outliers, especially in the development of linear regression models.

Other methods flag observations based on measures such as the interquartile range. For example, if Q_1 and Q_3 are the lower and upper quartiles respectively, then one could define an outlier to be any observation outside the range:

$$[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)]$$

for some nonnegative constant k.

In the data mining task of anomaly detection, other approaches are distance-based[14][15] and density-based,[16] and most of them use the distance to the k-nearest neighbors to label observations as outliers or non-outliers.[17]

- Modified Thompson Tau test[18]

The modified Thompson Tau test is a method used to determine if outlier exists in a data set. The strength of this method lies in the fact that it takes into account a data set's standard deviation, average and provides a statistically determined rejection zone; thus providing an objective method to determine if a data point is an outlier.

How it works: First, a data set's average is determined. Next the absolute deviation between each data point and the average are determined. Thirdly, a rejection region is determined using the formula: $RejectionRegion = \frac{t_{\alpha/2}(n-1)}{\sqrt{n}\sqrt{n-2+t_{\alpha/2}^2}}$; where $t_{\alpha/2}$ is the critical value from the Student t distribution, n is the sample size, and s is the sample standard deviation. To determine if a value is an outlier: Calculate δ = |(X - mean(X)) / s|. If δ > Rejection Region, the data point is an outlier. If δ ≤ Rejection Region, the data point is not an outlier.

The modified Thompson Tau test is used to find one outlier at a time (largest value of δ is removed if it is an outlier). Meaning, if a data point is found to be an outlier, it is removed from the data set and the test is applied again with a new average and rejection region. This process is continued until no outliers

remain in a data set.

Some work has also examined outliers for nominal (or categorical) data. In the context of a set of examples (or instances) in a data set, instance hardness measures the probability that an instance will be misclassified ($1 - p(y|x)$ where y is the assigned class label and x represent the input attribute value for an instance in the training set t).[19] Ideally, instance hardness would be calculated by summing over the set of all possible hypotheses H:

$$\begin{aligned} IH(\langle x, y\rangle) &= \sum_{H} (1 - p(y, x, h)) p(h|t) \\ &= \sum_{H} p(h|t) - p(y, x, h) p(h|t) \\ &= 1 - \sum_{H} p(y, x, h) p(h|t). \end{aligned}$$

Practically, this formulation is unfeasible as H is potentially or infinite and calculating $p(h|t)$ is unknown for many algorithms. Thus, instance hardness can be approximated using a diverse subset $L \subset H$:

$$IH_L(\langle x, y\rangle) = 1 - \frac{1}{|L|} \sum_{j=1}^{|L|} p(y|x, g_j(t, \alpha)$$

where $g_j(t, \alpha)$ is the hypothesis induced by learning algorithm g_j trained on training set t with hyperparameters α. Instance hardness provides a continuous value for determining if an instance is an outlier instance.

Working with outliers

The choice of how to deal with an outlier should depend on the cause.

Retention

Even when a normal distribution model is appropriate to the data being analyzed, outliers are expected for large sample sizes and should not automatically be discarded if that is the case. The application should use a classification algorithm that is robust to outliers to model data with naturally occurring outlier points.

Exclusion

Deletion of outlier data is a controversial practice frowned upon by many scientists and science instructors; while mathematical criteria provide an objective and quantitative method for data rejection, they do not make the practice more scientifically or methodologically sound, especially in small sets or where a normal distribution cannot be assumed. Rejection of outliers is more acceptable in areas of practice where the underlying model of the process being measured and the usual distribution of measurement error are confidently known. An outlier resulting from an instrument reading error may be excluded but it is desirable that the reading is at least verified.

The two common approaches to exclude outliers are truncation (or trimming) and Winsorising. Trimming discards the outliers whereas Winsorising replaces the outliers with the nearest "nonsuspect" data.[20] Exclusion can also be a consequence of the measurement process, such as when an experiment is not entirely capable of measuring such extreme values, resulting in censored data.[21]

In regression problems, an alternative approach may be to only exclude points which exhibit a large degree of influence on the estimated coefficients, using a measure such as Cook's distance.[22]

If a data point (or points) is excluded from the data analysis, this should be clearly stated on any subsequent report.

Non-normal distributions

The possibility should be considered that the underlying distribution of the data is not approximately normal, having "fat tails". For instance, when sampling from a Cauchy distribution,[23] the sample variance increases with the sample size, the sample mean fails to converge as the sample size increases, and outliers are expected at far larger rates than for a normal distribution. Even a slight difference in the fatness of the tails can make a large difference in the expected number of extreme values.

Set-membership uncertainties

A set membership approach considers that the uncertainty corresponding to the *i*th measurement of an unknown random vector x is represented by a set X_i (instead of a probability density function). If no outliers occur, x should belong to the intersection of all X_i's. When outliers occur, this intersection could be empty, and we should relax a small number of the sets X_i (as small as possible) in order to avoid any inconsistency.[24] This can be done using the notion of q-relaxed intersection. As illustrated by the figure, the q-relaxed intersection corresponds to the set of all x which belong to all sets except q of them. Sets X_i that do not intersect the q-relaxed intersection could be suspected to be outliers.

Alternative models

In cases where the cause of the outliers is known, it may be possible to incorporate this effect into the model structure, for example by using a hierarchical Bayes model or a mixture model.[25][26]

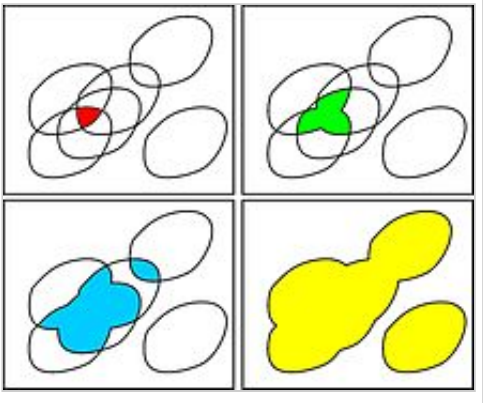

Figure 2. q-relaxed intersection of 6 sets for q=2 (red), q=3 (green), q= 4 (blue), q= 5 (yellow).

See also

- Anomaly time series
- Robust regression
- Studentized residual
- Data transformation (statistics)
- Local Outlier Factor

References

1. Grubbs, F. E. (February 1969). "Procedures for detecting outlying observations in samples". *Technometrics* **11** (1): 1–21. doi:10.1080/00401706.1969.10490657. "An outlying observation, or "outlier," is one that appears

to deviate markedly from other members of the sample in which it occurs."
2. Maddala, G. S. (1992). "Outliers". *Introduction to Econometrics* (2nd ed.). New York: MacMillan. pp. 88–96 [p. 89]. ISBN 0-02-374545-2. "An outlier is an observation that is far removed from the rest of the observations."
3. Grubbs 1969, p. 1 stating "An outlying observation may be merely an extreme manifestation of the random variability inherent in the data. ... On the other hand, an outlying observation may be the result of gross deviation from prescribed experimental procedure or an error in calculating or recording the numerical value."
4. Ripley, Brian D. 2004. Robust statistics (http://www.stats.ox.ac.uk/pub/StatMeth/Robust.pdf)
5. Ruan, Da; Chen, Guoqing; Kerre, Etienne (2005). Wets, G., ed. *Intelligent Data Mining: Techniques and Applications*. Studies in Computational Intelligence Vol. 5. Springer. p. 318. ISBN 978-3-540-26256-5.
6. Rousseeuw, P; Leroy, A. (1996), *Robust Regression and Outlier Detection* (3rd ed.), John Wiley & Sons
7. Hodge, Victoria J.; Austin, Jim, *A Survey of Outlier Detection Methodologies*, CiteSeerX: 10.1.1.109.1943
8. Barnett, Vic; Lewis, Toby (1994) [1978], *Outliers in Statistical Data* (3 ed.), Wiley, ISBN 0-471-93094-6
9. Adikaram, K.K.L.B.; Hussein, M.A.; Effenberger, M.; Becker, T. (2014). "Outlier Detection Method in Linear Regression Based on Sum of Arithmetic Progression". *The Scientific World Journal*. doi:10.1155/2014/821623.
10. Benjamin Peirce, "Criterion for the Rejection of Doubtful Observations" (http://articles.adsabs.harvard.edu/cgi-bin/nph-iarticle_query?1852AJ.....2..161P;data_type=PDF_HIGH), *Astronomical Journal* II 45 (1852) and Errata to the original paper (http://articles.adsabs.harvard.edu/cgi-bin/nph-iarticle_query?1852AJ.....2..176P;data_type=PDF_HIGH).
11. Peirce, Benjamin (May 1877 – May 1878). "On Peirce's criterion". *Proceedings of the American Academy of Arts and Sciences* **13**: 348–351. doi:10.2307/25138498. JSTOR 25138498.
12. Peirce, Charles Sanders (1873) [1870]. "Appendix No. 21. On the Theory of Errors of Observation". *Report of the Superintendent of the United States Coast Survey Showing the Progress of the Survey During the Year 1870*: 200–224.. NOAA PDF Eprint (http://docs.lib.noaa.gov/rescue/cgs/001_pdf/CSC-0019.PDF#page=215) (goes to Report p. 200, PDF's p. 215).
13. Peirce, Charles Sanders (1986) [1982]. "On the Theory of Errors of Observation [Appendix 21, according to the editorial note on page 515]". In Kloesel, Christian J. W., *et alia*. *Writings of Charles S. Peirce: A Chronological Edition*. Volume 3, 1872-1878. Bloomington, Indiana: Indiana University Press. pp. 140–160. ISBN 0-253-37201-1.
14. Knorr, E. M.; Ng, R. T.; Tucakov, V. (2000). "Distance-based outliers: Algorithms and applications". *The VLDB Journal the International Journal on Very Large Data Bases* **8** (3–4): 237. doi:10.1007/s007780050006.
15. Ramaswamy, S.; Rastogi, R.; Shim, K. (2000). *Efficient algorithms for mining outliers from large data sets*. Proceedings of the 2000 ACM SIGMOD international conference on Management of data - SIGMOD '00. p. 427. doi:10.1145/342009.335437. ISBN 1581132174.
16. Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; Sander, J. (2000). *LOF: Identifying Density-based Local Outliers* (PDF). *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. SIGMOD. pp. 93–104. doi:10.1145/335191.335388. ISBN 1-58113-217-4.
17. Schubert, E.; Zimek, A.; Kriegel, H. -P. (2012). "Local outlier detection reconsidered: A generalized view on locality with applications to spatial, video, and network outlier detection". *Data Mining and Knowledge Discovery*. doi:10.1007/s10618-012-0300-z.
18. John M. Cimbala (September 12, 2011). "Outliers" (PDF).
19. Smith, M.R.; Martinez, T.; Giraud-Carrier, C. (2014). "An Instance Level Analysis of Data Complexity (http://link.springer.com/article/10.1007%2Fs10994-013-5422-z)". Machine Learning, 95(2): 225-256.
20. Wike, Edward L. (2006). *Data Analysis: A Statistical Primer for Psychology Students*. pp. 24–25. ISBN 9780202365350.
21. Dixon, W. J. (June 1960). "Simplified estimation from censored normal samples". *The Annals of Mathematical Statistics* **31** (2): 385–391. doi:10.1214/aoms/1177705900.
22. Cook, R. Dennis (Feb 1977). "Detection of Influential Observations in Linear Regression". Technometrics (American Statistical Association) 19 (1): 15–18.
23. Weisstein, Eric W. Cauchy Distribution. From MathWorld--A Wolfram Web Resource (http://mathworld.wolfram.com/CauchyDistribution.html)
24. Jaulin, L. (2010). "Probabilistic set-membership approach for robust regression" (PDF). *Journal of Statistical Theory and Practice*.
25. Roberts, S. and Tarassenko, L.: 1995, A probabilistic resource allocating network for novelty detection. Neural Computation 6, 270–284.

26. Bishop, C. M. (August 1994). "Novelty detection and Neural Network validation". *Proceedings of the IEE Conference on Vision, Image and Signal Processing* **141** (4): 217–222. doi:10.1049/ip-vis:19941330

- ISO 16269-4, Statistical interpretation of data — Part 4: Detection and treatment of outliers
- Strutz, Tilo (2010). *Data Fitting and Uncertainty - A practical introduction to weighted least squares and beyond*. Vieweg+Teubner. ISBN 978-3-8348-1022-9.

External links

Wikimedia Commons has media related to ***Outliers***.

- Renze, John, "Outlier" (http://mathworld.wolfram.com/Outlier.html), *MathWorld*.
- Balakrishnan, N.; Childs, A. (2001), "Outlier", in Hazewinkel, Michiel, *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
- Grubbs test (http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm) described by NIST manual
- how to detect univariate outliers (http://www.psychwiki.com/wiki/Detecting_Outliers_-_Univariate), how to detect multivariate outliers (http://www.psychwiki.com/wiki/Detecting_Outliers_-_Multivariate) and how to deal with outliers (http://www.psychwiki.com/wiki/Dealing_with_Outliers)

Retrieved from "https://en.wikipedia.org/w/index.php?title=Outlier&oldid=706024124"

Categories: Statistical charts and diagrams | Statistical terminology | Data analysis | Robust statistics | Statistical outliers

---