Identifying Outliers

Mr. Merrick · September 29, 2025

Introduction

Outliers are unusually extreme values in a dataset that do not seem to fit with the general pattern. Detecting outliers is important because they can distort measures of center and spread, and sometimes reveal important real-world phenomena. In statistics, two widely used methods for detecting outliers are:

1. Tukey's 1.5(IQR) rule (based on the Interquartile Range).
2. The two standard deviation rule (based on the normal distribution).

1. Tukey's 1.5(IQR) Rule

John Tukey (1915–2000), a pioneering statistician, introduced the boxplot in 1977. Along with it came a systematic rule for identifying outliers using the Interquartile Range (IQR).

$$\text{IQR} = Q_3 - Q_1$$

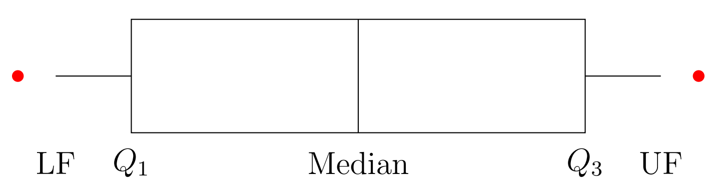
where Q_1 is the first quartile (25th percentile) and Q_3 is the third quartile (75th percentile).

Tukey's Fences

$$\text{Lower Fence} = Q_1 - 1.5 \times IQR \qquad \text{Upper Fence} = Q_3 + 1.5 \times IQR$$

Any data point below the lower fence or above the upper fence is considered an outlier.

Visualization



Here the red dots represent outliers beyond the fences.

2. Two Standard Deviation Rule

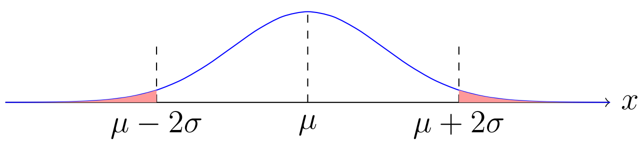
This rule relies on the properties of the normal distribution. It is often used for bell-shaped data distributions.

If μ is the mean and σ the standard deviation, then most of the data lies within two standard deviations:

$$[\mu - 2\sigma,\ \mu + 2\sigma]$$

Any data point outside this range is flagged as a potential outlier.

Visualization



The shaded red regions show potential outliers, since only about 5% of data lie outside $\mu \pm 2\sigma$ for a normal distribution.

Comparison

- Tukey's method is **non-parametric** (does not assume normality) and works well for skewed data. **Use this method for AP Statistics unless otherwise stated**.
- The two standard deviation rule assumes data is **approximately normal**.

Outliers should not be discarded automatically; rather, they should be studied carefully. They might represent data-entry errors, or they might reveal interesting phenomena worth deeper investigation.

Practice: Outliers

1. **Compute fences from raw data (Tukey).**
The data set (sorted) is:

$$5,\ 7,\ 8,\ 9,\ 10,\ 12,\ 13,\ 13,\ 14,\ 18,\ 35.$$

Find Q_1, Q_3, IQR, the lower/upper fences, and list any outliers by Tukey's 1.5(IQR) rule.

Solution.

There are $n = 11$ values, so the median is the 6th value: 12. Lower half $= \{5, 7, 8, 9, 10\}$ so $Q_1 = 8$. Upper half $= \{13, 13, 14, 18, 35\}$ so $Q_3 = 14$. IQR $= 14 - 8 = 6$. Fences: LF $= 8 - 1.5(6) = 8 - 9 = -1$, UF $= 14 + 9 = 23$. Outliers are > 23 or < -1, so 35 only.

2. **Use a five-number summary (Tukey).**
A distribution has five-number summary: min $= 5$, $Q_1 = 22$, Median $= 29$, $Q_3 = 35$, max $= 62$. Determine the IQR, fences, and which (if any) of min or max are outliers.

Solution.

IQR $= 35 - 22 = 13$. Fences: LF $= 22 - 1.5(13) = 22 - 19.5 = 2.5$, UF $= 35 + 19.5 = 54.5$. Since $5 > 2.5$, min is *not* an outlier. Since $62 > 54.5$, max is an outlier.

3. **Apply the Two-Standard-Deviation rule.**
In a (roughly normal) class score distribution with $\mu = 72$ and $\sigma = 9$, use the 2σ rule to flag potential outliers. Classify each value: 45, 54, 90, 96.

Solution.

Two-sigma interval: $[72 - 18, 72 + 18] = [54, 90]$. Values outside are flagged: 45 (outlier) and 96 (outlier). Endpoints 54 and 90 are *not* outliers.

4. **Method choice on skewed data.**
A right-skewed data set of daily website hits includes a single very large day. Explain why Tukey's IQR method is generally preferred to the 2σ rule for skewed distributions. In one sentence, state what assumption the 2σ rule is relying on.

Solution.

Tukey's method is based on quartiles and IQR, which are resistant to extreme values and do not assume any particular shape. The 2σ rule relies on the distribution being approximately normal (symmetric, light tails), which is violated under strong right skew.

5. **z-score version of the 2σ rule.**
Show that the 2σ rule is equivalent to flagging any observation with $|z| > 2$, where $z = (x - \mu)/\sigma$. Then, for $\mu = 50$, $\sigma = 8$, write the non-outlier interval and classify $x = 33$ and $x = 66$.

Solution.

$|z| > 2 \iff |x - \mu| > 2\sigma \iff x < \mu - 2\sigma$ or $x > \mu + 2\sigma$. With $\mu = 50, \sigma = 8$, interval is $[34, 66]$. So 33 is an outlier; 66 is on the boundary and *not* an outlier.

6. **Edge case at the fence (Tukey).**
True or false: If a data point lies *exactly* on a Tukey fence, it is an outlier. Justify briefly.

Solution.

False. Tukey outliers are typically defined as points *beyond* the fences. A point exactly on a fence is not considered an outlier.

7. **How transformations affect outlier rules.**
Suppose every value in a data set is transformed by (a) adding c, or (b) multiplying by $k > 0$. Describe how each method's outlier thresholds change.

   (a) Tukey 1.5(IQR) fences.

   (b) Two-sigma bounds.

Solution.

(a) **Tukey fences.**

- *Add c:* Quartiles Q_1, Q_3 each increase by c, so $\text{IQR} = Q_3 - Q_1$ is unchanged. Fences translate by c:

$$\text{LF}' = (Q_1 + c) - 1.5\,\text{IQR}, \qquad \text{UF}' = (Q_3 + c) + 1.5\,\text{IQR}.$$

  The keep-region has the same width; only its location shifts.

- *Multiply by $k > 0$:* Q_1, Q_3 and IQR all scale by k, so

$$\text{LF}' = k(Q_1 - 1.5\,\text{IQR}), \qquad \text{UF}' = k(Q_3 + 1.5\,\text{IQR}).$$

  The entire interval stretches if $k > 1$ and contracts if $0 < k < 1$.

(b) **Two-sigma bounds.**

- *Add c:* $\mu' = \mu + c$, $\sigma' = \sigma$. Bounds translate by c:

$$[\mu + c - 2\sigma,\ \mu + c + 2\sigma].$$

  Length unchanged.

- *Multiply by $k > 0$:* $\mu' = k\mu$, $\sigma' = k\sigma$. Bounds scale by k:

$$[k\mu - 2k\sigma,\ k\mu + 2k\sigma] = k[\mu - 2\sigma,\ \mu + 2\sigma].$$

Invariance of outlier count (for $k > 0$). Both transformations preserve the set of outliers when the thresholds are recomputed from the transformed data

8. **Compare methods on the same summary.**
A normal-approximate variable has mean $\mu = 40$ and standard deviation $\sigma = 6$. A different sample from the same process has $Q_1 = 36$ and $Q_3 = 44$. Compare the two sets of thresholds:

$$\text{Tukey fences: } Q_1 \pm 1.5\text{IQR}, \quad \text{and} \quad 2\sigma \text{ bounds: } \mu \pm 2\sigma.$$

Which method is tighter here?

Solution.

$\text{IQR} = 44 - 36 = 8$, so Tukey fences are $36 - 1.5(8) = 24$ and $44 + 12 = 56$. Two-sigma bounds are $40 \pm 12 \Rightarrow [28, 52]$. The 2σ bounds are tighter (narrower) in this case.