Why 1.5*IQR?

Origins of EDA

- John W. Tukey, often considered the father of EDA, published "Exploratory Data Analysis" at a time when computer-aided visualization was still nascent.
- He introduced plots such as the stem-leaf plot and the five-point boxplot.
- He came up with the 1.5 factor in the IQR method of finding outliers.

John W. Tukey

EXPLORATORY DATA ANALYSIS

Objectives of EDA :

- *Suggest hypotheses* about the causes of observed phenomena
- *Assess assumptions* on which statistical inference will be based
- Support the selection of *appropriate tools and techniques*
- Provide a basis *for further data collection* through surveys or experiments

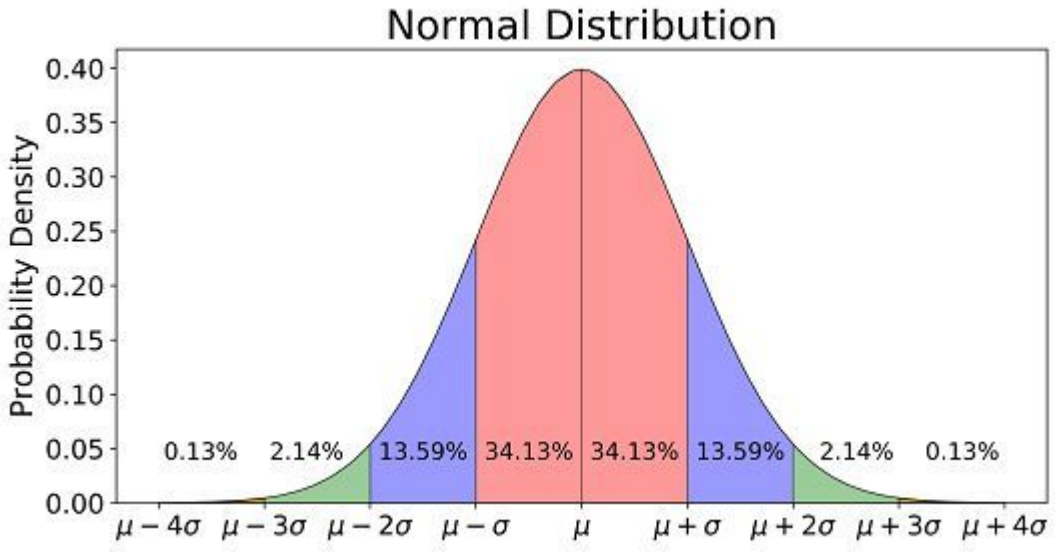
John Tukey,
Exploratory Data Analysis
(New York: Pearson, 1977)

IQR Method of Outlier Detection

- To detect the outliers using this method, we define a decision range.
- Any data point lying outside this range is considered an outlier.
- The range is as given below:
  - Lower Bound: (Q1 - 1.5 * IQR)
  - Upper Bound: (Q3 + 1.5 * IQR)

Why only 1.5 times the IQR?

- For example, let's say our data follows Gaussian distribution.
- The first and the third quartiles, Q1 and Q3, lies at -0.675σ and +0.675σ from the mean, respectively.



SCALE 1 IN IQR OUTLIER DETECTION

Lower Bound:

$= Q1 - 1 * IQR$

$= Q1 - 1 * (Q3 - Q1)$

$= -0.675\sigma - 1 * (0.675 - [-0.675])\sigma$

$= -0.675\sigma - 1 * 1.35\sigma$

$= -2.025\sigma$

Upper Bound:

$= Q3 + 1 * IQR$

$= Q3 + 1 * (Q3 - Q1)$

$= 0.675\sigma + 1 * (0.675 - [-0.675])\sigma$

$= 0.675\sigma + 1 * 1.35\sigma$

$= 2.025\sigma$

SCALE 1 IN IQR OUTLIER DETECTION

- So, when scale is taken as 1, then according to the IQR method, any data which lies beyond 2.025σ from the mean (μ), on either side, shall be considered as outlier.
- But as we know, the data is useful up to 3σ on either side of the μ.
- So we can't take scale = 1, because this makes the decision range too exclusive resulting in too many outliers.
- In other words, the decision range gets so small (compared to 3σ) that it considers some data points as outliers, which is not desirable.

SCALE 2 IN IQR OUTLIER DETECTION

Lower Bound:

$= Q1 - 2 * IQR$

$= Q1 - 2 * (Q3 - Q1)$

$= -0.675\sigma - 2 * (0.675 - [-0.675])\sigma$

$= -0.675\sigma - 2 * 1.35\sigma$

$= -3.375\sigma$

Upper Bound:

$= Q3 + 2 * IQR$

$= Q3 + 2 * (Q3 - Q1)$

$= 0.675\sigma + 2 * (0.675 - [-0.675])\sigma$

$= 0.675\sigma + 2 * 1.35\sigma$

$= 3.375\sigma$

Conclusion

- When scale is taken as 1.5, then according to the IQR method any data that lies beyond 2.7σ from the mean (μ), on either side, shall be considered an outlier.
- This decision range is the closest to what Gaussian Distribution tells us, i.e., 3σ.
- In other words, this makes the decision rule closest to what Gaussian distribution considers for outlier detection, and this is exactly what we wanted.
- To get exactly 3σ, we'd have to take the scale = 1.7.