MAD + Outlier Detection

- **Mean, Median, Mode**: Geometric interpretation
  - Optimizes L_0, L_1, L_2 distance metric.
  - Use some constant s, minimize $\|x - \vec{s}\|_p$.
  - Higher norm metric susceptible to outliers

- **Z-score**: A "measure" of how many standard deviations away from the mean a point is.

$$Z_i = \frac{x_i - \bar{x}}{s}$$

Modified Z-Score use median and median absolute deviation (MAD) to compute $M_i = \frac{0.6745(x_i - \tilde{x})}{MAD}$; Heuristic: $M_i > 3.5 \Rightarrow$ Outlier.

- **Median Absolute Deviation**: L_1 equivalent to the standard deviation

$$MAD := \text{median}(|x_i - \tilde{x}|)$$; Measure of spread

Formal Outlier Tests:

- ▲ **Grubb's Test**: Test for a single outlier.

  H_A: There is one outlier in the dataset

  $$G := \max \frac{|Y_i - \bar{Y}|}{s}$$; Test if G > Critical Value

  $$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{(t_{\alpha/2N, N-2})^2}{N-2+(t_{\alpha,2N,N-2})^2}}$$

- ▲ **Generalized ESD**: Test for up to r outliers
  - Iterate up to r times, Compute Grubb's statistic
  - At each iteration, compute λ_i, critical region for i^{th} iteration.
  - Flag as outlier is $G_i > \lambda_i$; Remove point at end of iteration.