THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

CHE384, From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Lecture 55
Robust Estimation

Chris A. Mack
Adjunct Associate Professor

http://www.lithoguru.com/scientist/statistics/

Robust Statistics

- Robustness: The ability of a statistical procedure to handle a variety of (non-normal) distributions, including outliers
  - Breakdown point: the fraction of contaminated data in a data set that can be tolerated by the statistical procedure
- Many statistics are not robust – even one bad data point can ruin the statistic
- Robust estimators are less efficient

P. Huber, *Robust Statistics*, New York: John Wiley and Sons (1981).

Contaminated Data

- Besides robustness against non-normal parametric distributions (e.g., Gamma distribution), we also seek robustness against contamination (a fraction of the data comes from a different distribution
- There are two models for contamination
  - Mean shift: some of the data comes from a distribution with a different mean
  - Variance shift: some of the data comes from a distribution with a larger variance

Location

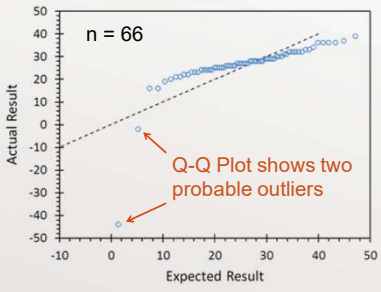
- The standard location statistic: mean $\bar{x}$
  - $SE(\bar{x}) = s/\sqrt{n}$ (assuming normal distribution)
- Some robust alternatives
  - Median, $SE = 1.253\, s/\sqrt{n}$ (b.p. = 0.5)
  - k-trimmed mean, $SE \approx \left(1 + \frac{2k}{n}\right)\frac{s}{\sqrt{n}}$ (b.p. = k/n)
  - k-Winsorized mean, $SE \approx \left(1 + \frac{2k}{n}\right)\frac{s}{\sqrt{n}}$

*Note: Definitions vary. Here, k points from top and k points from bottom are deleted or Winsorized.

Scale

- The scale statistic: standard deviation s
  - $SE(s) \approx s/\sqrt{2(n-1)}$ (normal distribution)
- Some robust alternatives
  - Median Absolute Deviation (MAD): $\text{MAD} = 1.4826\, Median(|x_i - Median(x)|)$, $SE(MAD) \approx 1.67\, s/\sqrt{2n}$, (b.p. = 0.5)
  - Interquartile Range, $SE(IQR) \approx 2.23\, s/\sqrt{2n}$, (b.p. = 0.25); for normal dist. IQR = 1.349σ

Example: Newcomb Speed of Light Data



Full data set:

mean =	26.21
median =	27.00
10% trimmed mean =	27.09
Sample StdDev =	10.75
0.7413*IQR =	5.00
MAD =	4.45

With outliers removed:

mean =	27.75
median =	27.50
10% trimmed mean =	27.74
Sample StdDev =	5.08
0.7413*IQR =	4.63
MAD =	5.19

Lecture 55: What have we learned?

- Explain robustness and the breakdown point
- What are some common robust location estimators?
- What are some common robust scale estimators?
- What is the main disadvantage of using robust estimators?