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

Lecture 18
Testing for Outliers, part 2

Chris A. Mack
Adjunct Associate Professor

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

Outliers

- Outlier: an observation so different from the others that one suspects it was generated by a different mechanism
- Two outlier mechanisms:
  - The true distribution has heavy tails
  - The data are "contaminated" by a second distribution with either a significantly different mean or a significantly larger variance

Studentized Data

- Most outlier procedures begin by taking the absolute value of the "studentized" data

How many standard deviations away from the mean is this data point?

$$T_i = \frac{|x_i - \bar{x}|}{s_x}$$

(named for the student's t distribution)

- For the case of model-fit residuals, some extra work is required to studentize them
  - More on this later

Robustness and Outliers

- Consider one "bad" data point, $x_{outlier}$

All the data points except $x_{outlier}$ are iid

$$E[\bar{x}] = \mu + \frac{x_{outlier} - \mu}{n}$$

$$E[s^2] = \sigma^2 + \frac{(x_{outlier} - \mu)^2}{n}$$

- For $x_{outlier} \gg n\mu, n\sigma$: $\bar{x} \approx \frac{x_{outlier}}{n}, s \approx \frac{|x_{outlier}|}{\sqrt{n}}$

$$T = \frac{|x_{outlier} - \bar{x}|}{s} \approx \frac{n-1}{\sqrt{n}}$$

The maximum possible value of T

Grubbs' Test

- The most common outlier rejection technique is the Grubbs' test
  - The test assumes the basic data come from a normal population
  - First, identify the number of outliers that you want to test
    - Examples: one upper tail outlier; one upper and one lower tail outlier; two lower tail outliers, etc.
    - For two outliers, we use a different critical value depending on whether the outliers are in the same or different tails

Frank E. Grubbs, "Sample Criteria for Testing Outlying Observations", *Ann. Math. Statist.*, **21**(1), 27-58 (1950).

Grubbs' Test

- Calculate the Grubb's statistic (equation below), the ratio of sum of squares errors

$$Grubbs'\ ratio = \frac{SSE\ with\ outliers\ removed}{SSE\ of\ full\ data\ set}$$

$$SSE \quad \sum (x_i - \bar{x})^2$$

- Is it less than the critical value from the table?

Grubbs' Test Alternate Formulation

- For a single outlier (extreme max or extreme min point),

$$T = \frac{|x_{extreme} - \bar{x}|}{s_x} \qquad s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Note: $Grubbs'\ ratio = 1 - \frac{T^2}{n-1}$

- Critical value for T:

$$T_{crit} = \frac{n-1}{\sqrt{n}}\sqrt{\frac{\left(t_{\frac{\alpha}{2n},n-2}\right)^2}{n-2+\left(t_{\frac{\alpha}{2n},n-2}\right)^2}}$$

$t_{\frac{\alpha}{2n},n-2}$ = critical t-value with DF = n–2 and a significance level of $\alpha/(2n)$. For one-sided tests, use a significance level of α/n.

Iterated Tests

- To find an unknown number (k) of outliers in a given sample, apply the Grubbs T-statistic test iteratively
  - Called the Extreme Studentized Deviate (ESD)
  - If an outlier is found, remove it and repeat the test for the remaining data
  - The critical T-statistic value depends on k (the current iteration number)
- Critical values can be found in
  - R.B. Jain, "Percentage points of many-outlier detection procedures", *Technometrics*, **23**, 71-76 (1981).

Peirce's Criterion

- Developed by Benjamin Peirce in 1852 (the first statistical outlier removal procedure)
  - Compare the probability of the data with the outliers to the probability of the data without the outliers
  - Assumes a normal distribution
  - Can remove multiple outliers in iterative approach
  - Uses T with lookup tables for critical value; not as common as the Grubbs' test

Stephen M. Ross, "Peirce's criterion for the elimination of suspect experimental data", *Journal of Engineering Technology*, **20**(2), 38-41 (2003).

Lecture 18: What have we learned?

- What is a studentized outlier?
- What is the maximum possible value for a studentized outlier?
- Be able to perform the Grubbs' test, using the tables provided.