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

Lecture 19 Some Final Thoughts on Outliers

Chris A. Mack

Adjunct Associate Professor

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

A Digression on Statistical Tests

- There are two important characteristics of a statistical test:
  1. The significance level, α = the probability of rejecting the null hypothesis when it is true (type I error)
  2. The power of the test = 1 – β, β = the probability of failing to reject the null hypothesis when it is false (type II error)
- We want high power (small β) and small α
  - Different tests have different powers for the same α
  - Higher sample size gives higher power

Comparing Outlier Tests

- Consider a data set where the Dixon Q-test fails to identify the extreme data point as an outlier, but the Grubbs' test does (for the same α)
  - We don't think of one test as being "right" and the other "wrong"
  - If we reject the null hypothesis (call the data point an outlier), then we know that our type I error rate is < α (which is set by us)
  - If we fail to reject the null hypothesis, it could be because our test has insufficient power (we don't set the value of β directly)
- The Grubbs' test has more power (given its assumptions are true), which is why we prefer it

What to Do with an Outlier?

- Can you identify the cause of the outlier?
  - Yes. Can the data be corrected?
    - Yes. Correct the data and include it in the analysis.
    - No. Remove the data but document everything.
  - No. Perform analysis with and without outlier. Does it affect your conclusions?
    - No. Don't worry, but of course document everything.
    - Yes. Ouch. Your conclusions may be suspect. You may need to take more data, or abandon the assumption of normality.

What to Do with an Outlier?

- When repeating the data analysis, there are many options of what to do with the outlier
  - Delete the outlier
  - Truncate (delete both the min and max data points)
  - Winsorize the outlier (set its value equal to its closest neighbor)
  - Replace the outlier with its expected value (from the Q-Q plot)
- Whether we delete, truncate, Winsorize, or replace the data depends on whether we identify the cause
  - We always delete spurious data
- In any case, document exactly what you did

Three Types of Outlier Causation

- Case 1: You notice the problem before you detect the outlier
  - E.g., a measurement tool breaks and must be repaired, you suspect calibration will be off
- Case 2: You investigate after the outlier is observed and identify a cause
  - Beware of just-so stories
- Case 3: You never find a cause
- Question: when do you report the existence of outliers in your data?

Is the Cause Important?

- Whether an outlier is important depends on the decision you are trying to make
  - Testing the accuracy of missiles, a few go way off course because of a software bug
  - Developing a measurement procedure, you are supplied with a degraded sample
- Spurious data vs. outlier depends on what is important to you

An Alternative

- An alternative to outlier rejection is robust estimation
  - Robust statistics have good performance for data drawn from a wide range of probability distributions, especially for distributions that are not normally distributed
  - More on robust estimation later
  - Bayesian approaches are also available
- In any case, it is always good to identify outliers for the lessons that can be learned
  - Further, outlier rejection can be thought of as a cheap version of robust estimation

Conclusions

- Tests for outliers and normality are related, since outliers result in non-normality
  - Most such tests are only useful when n > 20
- Typical testing sequence:
  - Graph the data as a histogram, boxplot, and Q-Q plot
  - Perform moment tests (skewness, then kurtosis)
  - If non-normality is detected, check for outliers (assuming a normal distribution can be justified)
  - If outliers are removed or adjusted, recheck for normality
  - If a non-normal distribution is suspected, use the empirical CDF to identify candidate distributions

Outlier Test Summary

- Testing for Normality (recommended)
  - Q-Q plots: is the normality assumption justified?
  - Skewness, kurtosis, etc.: Good for detecting the presence of outliers, but doesn't identify which data are outliers
- Multiple of IQR (recommended)
  - Robust; useful for identifying potential outliers to test or investigate
- Chauvenet's criterion
  - Simple; assumes normal distribution; arbitrary cut-off; not rigorous

Outlier Test Summary (2)

- Dixon Q-test
  - Most useful for small data sets; masking occurs with two or more outliers on same side of median
- Grubbs' test (recommended)
  - Most common and rigorous for data assumed normal; simple for one or two outliers, but can be used iteratively to identify more
- Peirce's criterion
  - Assumes normal distribution; can be used for any number of outliers; not rigorously studied for power and effectiveness

More Reading on Outliers

- Vic Barnett and Toby Lewis, *Outliers in Statistical Data*, John Wiley & Sons (1st edition from 1978, 3rd edition from 1994).
- F. J. Anscombe and Irwin Guttman, "Rejection of Outliers", *Technometrics*, Vol. 2, No. 2, pp. 123-147 (May, 1960).
- Frank E. Grubbs, "Procedures for Detecting Outlying Observations in Samples", *Technometrics*, Vol. 11, No. 1, pp. 1-21 (1969).

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

Lecture 19: What have we learned?

- Why should one focus on identifying the cause of an outlier?
- Name the four options for what to do with an outlier that can't be ignored
- What is an important alternative to outlier testing and rejection?
- Describe the recommended testing sequence for outliers

