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

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

# Lecture 17
# Testing for Outliers, part 1

Chris A. Mack
*Adjunct Associate Professor*

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



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

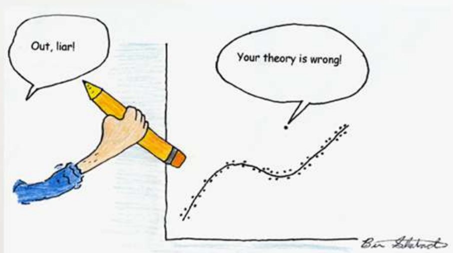
## Outliers

- Outlier: an observation so different from the others that one suspects it was generated by a different mechanism
  – A one-time, large systematic error (also called a data flyer, wild observation, maverick, etc.)
- Possible causes of outliers:
  – Error in recording the measurement
  – Failure of the measurement process/tool
  – One sample was fundamentally different from other samples being measured
  – Failure of the experimental process (e.g., sample did not receive the proper treatment)



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

## Are Outliers Bad?




Cartoon by Ben Shabad, http://davidmlane.com/ben/cartoons.html



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

## Outliers vs. Spurious Data

- Outlier: an observation so different from the others that one suspects it was generated by a different mechanism
- Spurious Data Point: a data value that has nothing to teach us about the subject matter of interest
  – We remove spurious data without guilt
  – Not all outliers are spurious



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

## Why Detect Outliers?

- For non-robust statistics, one bad data point can ruin the analysis
  – One outlier will violate the normal distribution assumption, for example
  – We detect in order to correct, adjust, or reject
- Detecting outliers is the first step to discovering the mechanism that caused the outlier
  – Sometimes we are more interested in the causes of outliers than in the analysis of the "good" data



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

## Rare Events

- Rare events that do not entail a different mechanism can happen, but are not true outliers

Table of two-sided p-values

| $(x-\bar{x})/s$ | 3 | 3.5 | 4 | 4.5 | 5 | 5.5 | 6 |
|---|---|---|---|---|---|---|---|
| Student's t, DF = 5 | 3.0E-02 | 1.7E-02 | 1.0E-02 | 6.4E-03 | 4.1E-03 | 2.7E-03 | 1.8E-03 |
| Student's t, DF = 10 | 1.3E-02 | 5.7E-03 | 2.5E-03 | 1.1E-03 | 5.4E-04 | 2.6E-04 | 1.3E-04 |
| Student's t, DF = 20 | 7.1E-03 | 2.3E-03 | 7.0E-04 | 2.2E-04 | 6.9E-05 | 2.2E-05 | 7.2E-06 |
| Student's t, DF = 40 | 4.6E-03 | 1.2E-03 | 2.7E-04 | 5.7E-05 | 1.2E-05 | 2.4E-06 | 4.7E-07 |
| Student's t, DF = 100 | 3.4E-03 | 7.0E-04 | 1.2E-04 | 1.8E-05 | 2.5E-06 | 2.9E-07 | 3.2E-08 |
| Two-tailed Normal | 2.7E-03 | 4.7E-04 | 6.3E-05 | 6.8E-06 | 5.7E-07 | 3.8E-08 | 2.0E-09 |

- Statistical tests calculate the probability of the suspect data occurring by chance (p-value)

## Using Rareness to Detect Outliers

- If the probability of getting the extreme data point is far smaller than 1/n (n = number of data points), we can consider the data point an "outlier"
  - P-value*n = probability of getting one data point (out of n) this unusual or more so due to random chance
  - Assumes we know the underlying distribution
  - Chauvenet's criterion: reject if p-value < 1/(2n)
- There are many other statistical tests for detecting outliers, and none of them are perfect
  - Multiple of IQR (for outlier labeling)
  - Dixon Q-test
  - Grubbs' Test
  - Peirce's Criterion





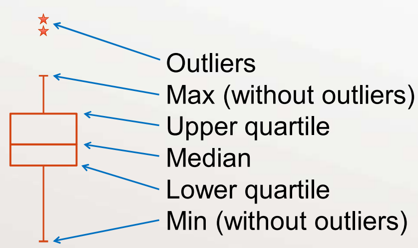
## Multiple of IQR test

- Use Interquartile range (IQR), which is insensitive to outliers (robust)
  - IQR = 75% quartile – 25% quartile
  - Upper limit = 75% quartile + 1.5*IQR
  - Lower limit = 25% quartile - 1.5*IQR
- Any data points outside of the upper/lower limits are labeled as outliers
  - For a normal population, about 1% of data points could be expected to be so labeled ($n$ dependent)
  - "Far" outliers use a 3*IQR criterion





## Box and Whisker Plot








## Dixon Q-test

- Identify one suspect (extreme) data point

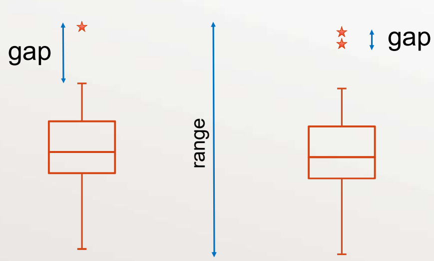
$$Q = \frac{|x_{suspect} - x_{closest}|}{x_{max} - x_{min}} \quad \text{(also called } r_{10}\text{)}$$

- Look up critical Q value from table
- Reject as an outlier if $Q > Q_{critical}$
- Mostly used when n is small (so calculating the standard deviation is suspect); e.g., n < 20
- Problem: masking (what if there are two outliers?)





## Dixon Q-test Masking








## Dixon Q-test Table

- Critical values assume a normal distribution
- It is common to use a smaller significance level for outlier rejection than for other statistical tests
  - E.g., use α = 0.01 for outlier rejection when using α = 0.05 for other tests with the same data
  - α = risk of rejecting good data

| n | α/2 = 0.1 | 0.05 | 0.02 | 0.01 | 0.005 |
|---|---|---|---|---|---|
| 3 | 0.885 | 0.941 | 0.976 | 0.988 | 0.994 |
| 4 | 0.679 | 0.765 | 0.846 | 0.889 | 0.920 |
| 5 | 0.558 | 0.642 | 0.729 | 0.782 | 0.823 |
| 6 | 0.484 | 0.562 | 0.646 | 0.699 | 0.744 |
| 7 | 0.434 | 0.508 | 0.586 | 0.637 | 0.681 |
| 8 | 0.398 | 0.467 | 0.543 | 0.591 | 0.634 |
| 9 | 0.370 | 0.436 | 0.509 | 0.555 | 0.595 |
| 10 | 0.349 | 0.412 | 0.481 | 0.526 | 0.566 |
| 11 | 0.331 | 0.392 | 0.459 | 0.503 | 0.542 |
| 12 | 0.317 | 0.376 | 0.441 | 0.483 | 0.521 |
| 13 | 0.305 | 0.362 | 0.425 | 0.466 | 0.503 |
| 14 | 0.294 | 0.350 | 0.412 | 0.452 | 0.487 |
| 15 | 0.285 | 0.339 | 0.399 | 0.439 | 0.474 |
| 20 | 0.251 | 0.301 | 0.356 | 0.392 | 0.425 |
| 25 | 0.230 | 0.276 | 0.329 | 0.363 | 0.394 |
| 30 | 0.216 | 0.259 | 0.309 | 0.343 | 0.372 |
| 40 | 0.196 | 0.237 | 0.284 | 0.314 | 0.342 |
| 50 | 0.183 | 0.222 | 0.266 | 0.296 | 0.323 |
| 60 | 0.173 | 0.211 | 0.253 | 0.282 | 0.308 |
| 70 | 0.166 | 0.202 | 0.244 | 0.271 | 0.297 |
| 80 | 0.160 | 0.195 | 0.236 | 0.263 | 0.288 |
| 90 | 0.155 | 0.190 | 0.229 | 0.256 | 0.280 |
| 100 | 0.151 | 0.185 | 0.223 | 0.250 | 0.274 |

Surendra P. Verma and Alfredo Quiroz-Ruiz, "Critical values for six Dixon tests for outliers in normal samples up to sizes 100, and applications in science and engineering", *Revista Mexicana de Ciencias Geológicas*, **23**(2), 133-161 (2006).

## Lecture 17: What have we learned?

- What is an outlier?
- What is the difference between an outlier and a spurious data point?
- How does the Box and Whisker plot identify outliers?
- Be able to perform the Dixon Q test.  What can go wrong with this test?