Lecture 17
Testing for Outliers, part 1

Chris A. Mack
Adjunct Associate Professor

© Chris Mack, 2016
Data to Decisions

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)

Are Outliers Bad?

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

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

Rare Events

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

<table>
<thead>
<tr>
<th>n = (n−2)/2</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student’s t, DF = 5</td>
<td>3.00E-02</td>
<td>1.78E-02</td>
<td>1.06E-02</td>
<td>6.44E-03</td>
<td>4.01E-03</td>
</tr>
<tr>
<td>Student’s t, DF = 10</td>
<td>1.41E-02</td>
<td>7.61E-03</td>
<td>2.61E-03</td>
<td>1.15E-03</td>
<td>5.46E-04</td>
</tr>
<tr>
<td>Student’s t, DF = 20</td>
<td>8.10E-03</td>
<td>4.36E-03</td>
<td>1.60E-03</td>
<td>6.96E-04</td>
<td>3.25E-04</td>
</tr>
<tr>
<td>Student’s t, DF = 40</td>
<td>5.16E-03</td>
<td>2.80E-03</td>
<td>1.20E-03</td>
<td>4.86E-04</td>
<td>2.20E-04</td>
</tr>
<tr>
<td>Student’s t, DF = 100</td>
<td>3.49E-03</td>
<td>2.02E-03</td>
<td>1.28E-03</td>
<td>5.27E-04</td>
<td>2.52E-04</td>
</tr>
<tr>
<td>Two-tailed Normal</td>
<td>2.77E-03</td>
<td>4.75E-04</td>
<td>6.31E-05</td>
<td>6.68E-06</td>
<td>5.78E-07</td>
</tr>
</tbody>
</table>

- 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

Outliers
Max (without outliers)
Upper quartile
Median
Lower quartile
Min (without outliers)

Dixon Q-test

- Identify one suspect (extreme) data point
- 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 Table

<table>
<thead>
<tr>
<th>n</th>
<th>Q_{critical}</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.885</td>
</tr>
<tr>
<td>3</td>
<td>0.679</td>
</tr>
<tr>
<td>4</td>
<td>0.558</td>
</tr>
<tr>
<td>5</td>
<td>0.484</td>
</tr>
<tr>
<td>6</td>
<td>0.434</td>
</tr>
<tr>
<td>7</td>
<td>0.398</td>
</tr>
<tr>
<td>8</td>
<td>0.370</td>
</tr>
<tr>
<td>9</td>
<td>0.349</td>
</tr>
<tr>
<td>10</td>
<td>0.331</td>
</tr>
</tbody>
</table>

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?