Outlying Replicates

How to Identify and Deal with Unusual Data Points

Precision Scientific Software Inc.

Contents

- Normal Variation and Summary Statistics
- Conventional and Robust Statistical Parameters
- Calculation of Median and Interquartile Range
- Influential Data Points
- Types of Outliers
- Consequences of having no Outlier Strategy
- Subjective Handling of Outliers
- Statistical Outlier Detection Methods
- Dealing with Outlying Data Points
- Functionality of the Excel Template

Normal Variation and Summary Statistics

- Normal variation
  - There is variation in all replicate data
  - Variation is inherent in:
    - The object being tested
    - The people / instruments being used for testing
  - In outliers analysis we are looking for and handling unusual amounts of variation

- Summary statistics
  - Statistical parameters describe the important characteristics of the data
    - Central tendency (mean or median)
    - Dispersion (standard deviation or other measures of variability)
  - Statistical parameters can be described as either
    - Conventional – responsive to every data point
    - Robust – resistant to changes in individual data points

Conventional and Robust Statistical Parameters

- Conventional measures of scale
    - Mean and standard deviation are the most commonly used measures of central tendency and dispersion
    - Every data point affects their value so they are very sensitive to extreme points

- Robust measures of scale
    - Median is the value of the “middle” data value in a data set
        - If there are 25 replicates in the data set, the median is the value of the 13th largest replicate
            - If there are an even number of replicates, the median is the average of the two “middle” replicates
        - Even if a few of the largest values are much larger than all the other values, this will not affect the size of the middle value so the median will not change
    - Interquartile range is the difference between the 1st and 3rd quartile values
        - This is used as a robust measure of dispersion

Calculation of Median and Interquartile Range

Summary Statistics - All Dow Stocks

Dow Stock Symbol	Share Vol. (millions)	Median	Quartiles
TRV	1.5		
AXP	3.2		
GS	3.3		
MMM	3.4		
AMGN	3.5		
HON	3.7		
CAT	4.5		
MCD	5.3		5.3
UNH	5.4		
DOW	6.4		
NKE	6.7		
HD	7.9		
IBM	9.1		
WBA	9.4		
V	11.7	12.1	
BA	12.5		
JNJ	13.2		
CVX	13.7		
WMT	13.7		
DIS	15.8		
MRK	15.9		
PG	16.4		
CRM	19.3		19.3
JPM	19.8		
CSCO	22.1		
KO	22.5		
VZ	23.9		
INTC	35.1		
MSFT	37.8		
AAPL	**164.6**		
Mean	**17.7**	**Median**	**12.1**
Std Dev	**29.2**	**IQR**	**14.0**

Summary Statistics - Double AAPL Volume

Dow Stock Symbol	Share Vol. (millions)	Median	Quartiles
TRV	1.5		
AXP	3.2		
GS	3.3		
MMM	3.4		
AMGN	3.5		
HON	3.7		
CAT	4.5		
MCD	5.3		5.3
UNH	5.4		
DOW	6.4		
NKE	6.7		
HD	7.9		
IBM	9.1		
WBA	9.4		
V	11.7	12.1	
BA	12.5		
JNJ	13.2		
CVX	13.7		
WMT	13.7		
DIS	15.8		
MRK	15.9		
PG	16.4		
CRM	19.3		19.3
JPM	19.8		
CSCO	22.1		
KO	22.5		
VZ	23.9		
INTC	35.1		
MSFT	37.8		
AAPL	**329.1**		
Mean	**23.2**	**Median**	**12.1**
Std Dev	**58.5**	**IQR**	**14.0**

Summary Statistics - Exclude AAPL Volume

Dow Stock Symbol	Share Vol. (millions)	Median	Quartiles
TRV	1.5		
AXP	3.2		
GS	3.3		
MMM	3.4		
AMGN	3.5		
HON	3.7		
CAT	4.5		4.9
MCD	5.3		
UNH	5.4		
DOW	6.4		
NKE	6.7		
HD	7.9		
IBM	9.1		
WBA	9.4		
V	11.7	11.7	
BA	12.5		
JNJ	13.2		
CVX	13.7		
WMT	13.7		
DIS	15.8		
MRK	15.9		
PG	16.4		17.9
CRM	19.3		
JPM	19.8		
CSCO	22.1		
KO	22.5		
VZ	23.9		
INTC	35.1		
MSFT	37.8		
AAPL			
Mean	**12.6**	**Median**	**11.7**
Std Dev	**9.3**	**IQR**	**13.0**

- The Median and Interquartile Range are much less sensitive to the presence of extreme values than the Mean and Standard Deviation

- Excel formulas:

=MEDIAN(range)

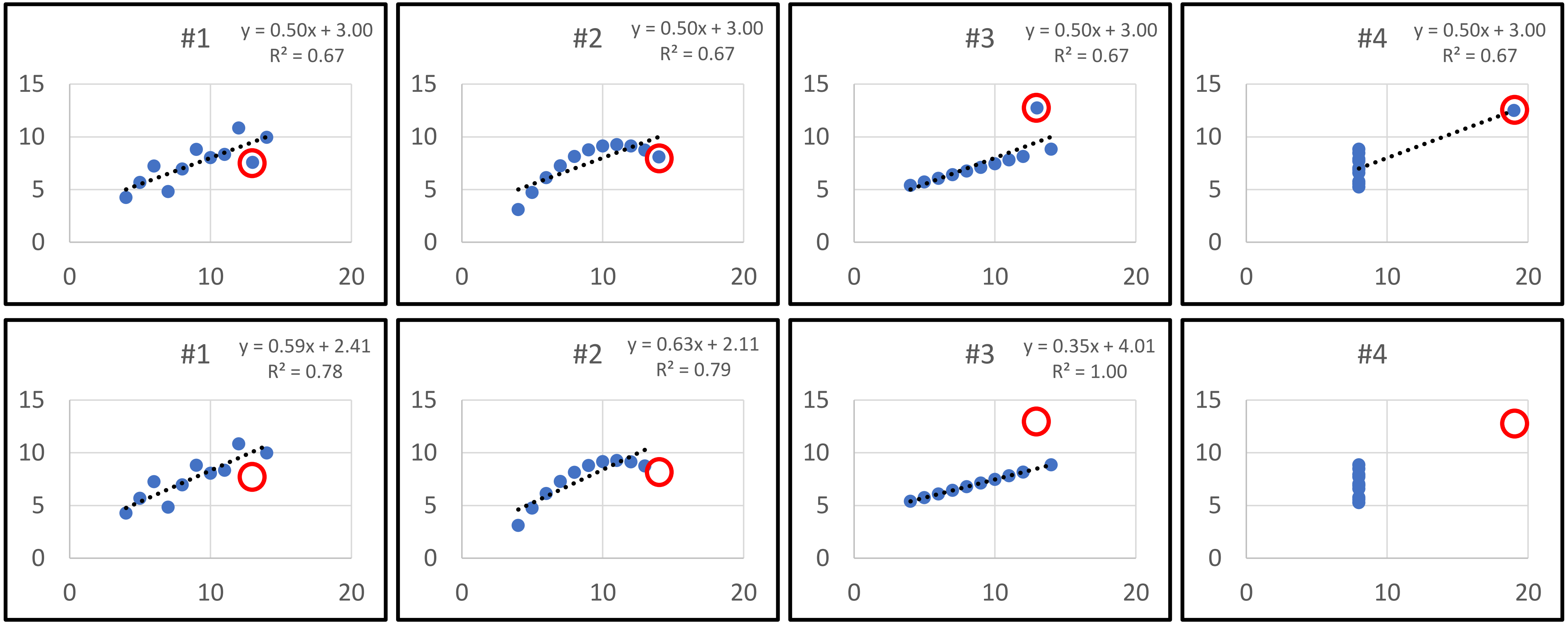
=QUARTILE.INC(range, 1)
=QUARTILE.INC(range, 3)

(the data does NOT have to be in order for these Excel formulas to work)

Precision Scientific Software Inc.

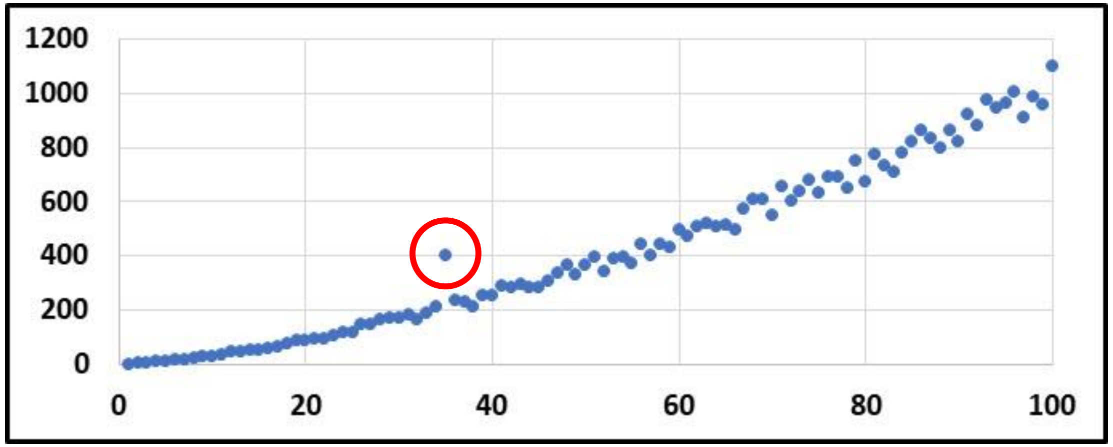
Influential Data Points

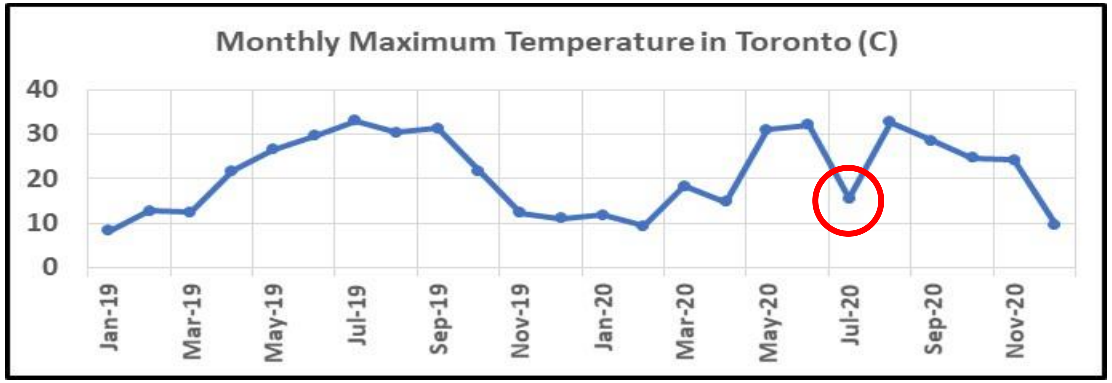
- Look for a single point with a large influence on the result of the analysis
  - These graphs have the same mean, standard deviation, and linear regression lines
  - Graphs 3 and 4 include single points with large influences on the statistical results

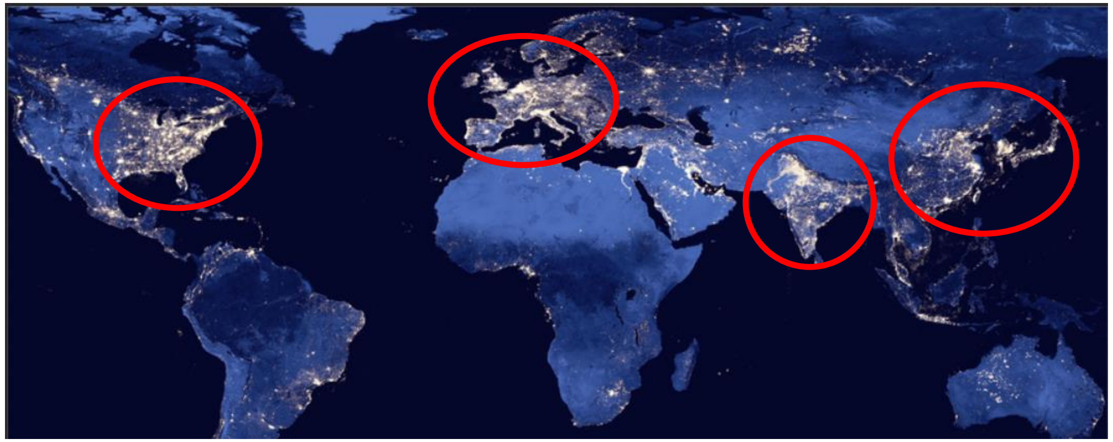


Types of Outliers

- Outliers are classified as
  - Global (or Point) outlier
    - A single point that is far removed from the other data points
  - Contextual outlier
    - A point that is unusual within the context in which it is found
  - Collective outlier
    - A group of points whose collective behaviour is unusual







(NASA)

Consequences of having no Outliers Strategy

- Outlier strategy
  - Recognize that outliers may exist in the data set
  - Use statistical tests to identify outliers
  - Determine the appropriate action to take based on clear, consistent, and defensible criteria
  - Document your actions and rationale

- Failure to have a strategy could result in
  - Being unaware that your results contain suspect data
  - Making business decisions based on faulty data
  - Taking inconsistent and unreproducible action
  - Having no record of previous actions
  - Having no audit trail to demonstrate your reasoning or justify your actions

Subjective Handling of Outliers

- Subjective approach to handling outliers
  - Designate data values as being outliers if the value doesn’t “look right” compared to either the other replicates or the expected result
  - Discard replicates if they “seem” distant from the other replicates

- Leads to several types of bias in data handling
  - Bias towards expectation
    - A data value further from the expected result is more likely to be identified as an outlier than a value closer to the expected result – even if both are equally far from the dataset average
      - Example, if the specification for a property is 85 and the dataset has a measured average of 92 then an individual data value of 97 is more likely to be identified as an outlier than an individual data value of 87
    - Increases the likelihood of missing problems in the production process

Subjective Handling of Outliers

- Biases in data handling
  - Bias of order of magnitude
    - Data values on the "other side" of a multiple of 10 appear further from each other than those on the same side
      - Example, if a dataset has a measured average of 980 then an individual data value of 1010 seems further away than an individual data value of 940
      - The extra digit makes the value much more noticeable in a visual scan of the data
  - Bias of "convenience"
    - Tendency to avoid the inconvenience of having to replace borderline values
  - Interpersonal biases
    - Different people will have different subjective criteria for identifying and dealing with possible outliers

Statistical Outlier Detection Methods

- A value is a statistical outlier if the frequency of its occurrence is greater that what would be predicted based on random variation
  - Example, a value more than 3 σ from the mean of a normally distributed population should only occur 0.27% of the time
    (0.135% on each of the high and low side of the mean)
  - A value 3 σ from the mean would only be expected to occur once in 370 replicates (1 / 0.0027 ≈ 370)
    - If it occurs in a dataset of 20 replicates then it is a statistical outlier

- There are many statistically rigorous tests to identify outliers
  - Some tests are very sophisticated and difficult to use and interpret
  - Two practical tests are
    - Tukey's Interquartile Range Fences (robust – based on the median)
    - Modified Thompson Tau Test (conventional – based on the mean)

Statistical Outlier Detection Methods

- Tukey's Interquartile Range Fences
  - Determine the upper and lower quartiles of the data (Q3 and Q1)
  - Interquartile Range: IQR = Q3 – Q1
  - An Outlier is a value: Less than Q1 – 1.5 x IQR or Greater than Q3 + 1.5 x IQR
  - An Extreme value is: Less than Q1 – 3 x IQR or Greater than Q3 + 3 x IQR
  - This method identifies all outliers "on the first pass" (it is not applied iteratively)

- Since this is a "robust" test (ie, based on the median and quartile values) it is reliable as long as the number of outliers does not encompass either of the upper or lower quartiles

Statistical Outlier Detection Methods

- Modified Thompson Tau Test
  - Calculate the Rejection Region using the formula:

$$\mu \pm \frac{t_{(\alpha/2\,,\,\mathrm{n}-2)} \times (n-1)}{\sqrt{n} \times \sqrt{(n-2) + (t_{(\alpha/2\,,\,\mathrm{n}-2)})^2}}$$

    - An Outlier is a value outside the rejection region based on α = 95%
    - An Extreme value is outside the rejection region based on α = 97.5%
  - If any value(s) are outside the rejection region, remove the value furthest outside
  - Recalculate and continue to remove the one value furthest out until no values remain outside the rejection region

- Since this is a “conventional” test (ie, based on the mean) the result is affected by the presence of any outliers
  - This is why the test is applied iteratively, rejecting one outlier at a time

Dealing with Outlying Data Points

- An outlying data point should be investigated to determine whether there is an "assignable cause" for its value
    - Experimental error
    - Instrument malfunction
    - Transcription error
    - Bad or mis-identified sample

- If an assignable cause is identified the data value should be rejected
    - The data value and its cause should be logged for future audit and continuous improvement purposes

- Sometimes the outlier is what we are searching for
    - New learning – especially at the boundaries of normal investigation
    - Fraud detection – for instance anomalous credit card activity

Dealing with Outlying Data Points

- If an assignable cause cannot be identified then a decision must be made

  - Keep the replicate
    - Not all outliers should be automatically discarded
    - Often a statistical outlier is simply the result of normal variation and is therefore a legitimate result
    - If it is not “extreme” or “influential” and does not significantly change the overall result

  - Discard the replicate
    - If it cannot be right (ie, if the value is outside the range of normal experience)
    - If it is practical to measure another replicate to replace the discarded value
    - You have a lot of data and discarding the replicate will not affect your results
    - If it is an influential point that “creates” a relationship in the data that otherwise wouldn’t exist

Dealing with Outlying Data Points

- If an assignable cause cannot be identified then a decision must be made

  - Investigate the replicate

    - If even a small change in the result is critical
      - Example: the temperature at failure of an aircraft engine part
    - If it is an extreme or influential point
      - Investigate whether unexpected factors are affecting the data
      - Run the analysis with and without the outlying data to see whether it changes the conclusion
    - If there is a large number of outliers
      - By definition outliers should be a rare event
      - A large number may indicate that there are two populations within the data
      - Run the analysis separately on each population to see whether they behave differently

Dealing with Outlying Data Points

- If an assignable cause cannot be identified then a decision must be made
  - Modify the value of the replicate
    - If it is an extreme value, rather than being discarding it can be modified ("Winsored" – named after Charles Winsor)
    - Replace the measured value with a value just inside the extreme threshold
    - This is preferrable to discarding the value because it does not reduce the standard deviation as much as discarding or replacing would
  - Whatever action is taken (even keeping the value) should be logged for future audit and continuous improvement purposes
    - If the data value is kept but the decision and rationale are not logged then an auditor may demand an explanation why this outlying value was kept but some other outlying value was discarded

Functionality of the Excel Template

- Worksheets implementing both the IQR and Tau methods
  - Distinguish between Outliers and Extreme values
  - Summary Statistics for all replicates and for non-outlying replicates
- Ability to set “meaningful” Tolerance limits to avoid unnecessarily stringent outlier identification
- Identify outlying replicates in a single characteristic
- Identify outlying replicates in a group of related characteristics
- Ability to apply a linear regression model to data to find outlying values based on the predicted relationship between two characteristics