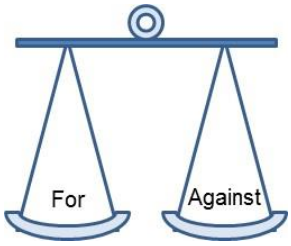
Outing the Outliers – Tails of the Unexpected

Cynics would say that we can prove anything we want with statistics as it is all down to interpretation and misinterpretation. To some extent this is true, but in truth it is all down to making a judgement call based on the balance of probabilities.



… a word (or two) from the Wise?

"The great tragedy of Science: the slaying of a beautiful hypothesis by an ugly fact"
Thomas Henry Huxley
(1825-1895)
British Biologist

The problem with using random samples in estimating is that for a small sample size, the values could be on the "extreme" side, relatively speaking – like throwing double six or double one with a pair of dice, and nothing else for three of four turns. The more random samples we have the less likely (statistically speaking) we are to have all extreme values. So, more is better if we can get it, but sometimes it is a question of "*We would if we could, but we can't so we don't!*" Now we could argue that Estimators don't use random values (*because it sounds like we're just guessing*); we base our estimates on the "actuals" we have collected for similar tasks or activities. However, in the context of estimating, any "actual" data is in effect random because the circumstances that created those "actuals" were all influenced by a myriad of random factors. Anyone who has ever looked at the "actuals" for a repetitive task will know that there are variations in those values. What we want to know is, is there a pattern to the variation, and therefore can we pick a value that suits our purpose, or better still three values[1]; generally speaking we will want to avoid the extreme values.

To help us identify whether a value is reasonably representative of all the others, and is not an outlier, or more generally, where a sample statistic falls in relation to the population to which it belongs, statisticians have developed a number of tests, some of which are known by the name of their "inventor" or who added significantly to the "body of knowledge", and others by a single latter (e.g. Z, t, F or U)

Before we explore these, we need to explore the idea of making an assumption (or a Hypothesis) and then how we might substantiate or repudiate that assumption or hypothesis.

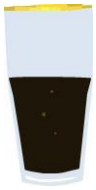
[1] Optimistic, Most Likely and Pessimistic perspectives, but not necessarily Minimum or Maximum in an absolute sense

1. Hypothesis Testing

Statistical Tests usually involve either Significance Testing or Hypothesis Testing, where the practitioner, estimator, planner etc tests the validity of an assumption or hypothesis against another. These hypotheses are usually referred to as:

- The Null Hypothesis
- The Alternative Hypothesis

As with everything statistical, there is always more than one interpretation of these things. We have to decide which perspective we are going to take: Something is assumed to be true until we prove it is false (Significance Testing), or something is assumed to be false unless we can prove it is true (Hypothesis Testing). This is akin to the legal perspective of "*innocent until proven guilty*"! Both type of test refer to Hypotheses, and sentence them based on the level of significance calculated; to some extent this is the difference between the optimistic and pessimistic perspectives – is this glass half-full or half-empty?



Definition

Null Hypothesis

A Null Hypothesis is that supposition that the difference between an observed value or effect and another observed or assumed value or effect, can be legitimately attributable to random sampling or experimental error. It is usually denoted as H_0.

Definition 1: Null Hypothesis

In experiments, or in using empirical results, the Null Hypothesis generally assumes that the implied relationship in the data is wrong (Field, 2005, p.739), and we have to test whether that assumption could be true. In the context of the justice system, the Null Hypothesis can be likened to "Not Guilty"; it is the prosecution's job to show that the evidence does not support that assumption, beyond reasonable doubt.

Definition

Alternative Hypothesis

An Alternative Hypothesis is that supposition that the difference between an observed value and another observed or assumed value or effect, cannot be legitimately attributable to random sampling or experimental error. It is usually denoted as H_1.

Definition 2: Alternative Hypothesis

The Alternative Hypothesis is called the Experimental Hypothesis by Field (2005, p.730). If the Null Hypothesis can be shown to be wrong then the Alternative Hypothesis is implied to be correct, i.e. that the relationship generated by the empirical results is valid. In the context of our judicial example, the defendant has just been found guilty.

As with any Justice System though, there is always the possibility of a miscarriage of justice, where the verdict of the court is inadvertently misplaced based on the evidence presented; the same can be said of Statistical Hypothesis Testing. We can classify any errors in the interpretation of Statistical Tests in two ways, for instance:

- **Type I Error:** False positive i.e. accepting a hypothesis we should have rejected e.g. reaching an innocent verdict for a guilty person
- **Type II Error:** False negative i.e. rejecting a hypothesis we should have accepted e.g. reaching a guilty verdict for an innocent person

The problem is that if we decrease the chance of one type of error, then generally speaking we will be increasing the other type of error. Consequently, as estimators, we have to decide which of the two errors are the lesser or greater of the two evils:

1 a) Be too optimistic – win the job and lose money … or

1 b) Be too pessimistic – avoid losing money by not winning the job

2 a) Be too optimistic – treat a disease with medication that doesn't actually work … or

2 b) Be too pessimistic – don't treat a disease with medication that would have worked had we used it

In terms of "Tails of the Unexpected", this is not a spelling mistake, but a reference to the low probability of values falling in the tails of a Probability Distribution. The "tails" being those little curly bits at either "end" of a Probability Distribution (either CDF or PDF/PMF) that go "flat" or "asymptotically flat" along the axis as illustrated in Figure 1. Values in these areas have low significance in the context of all others.

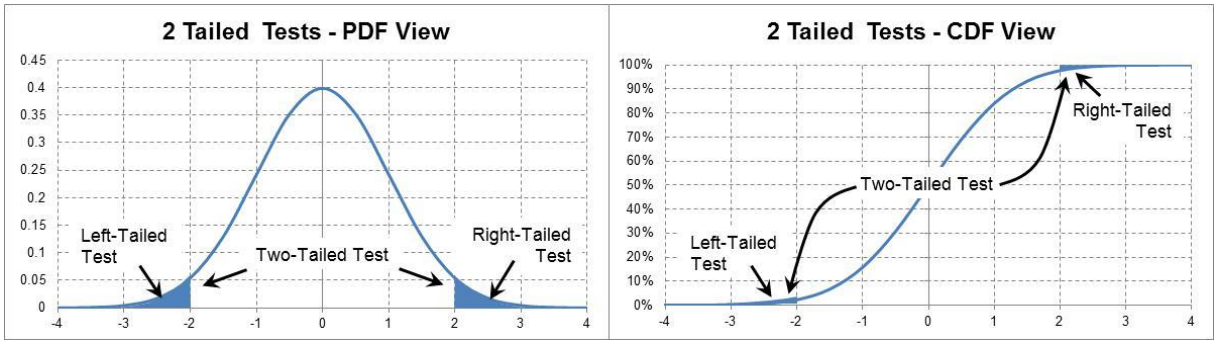


Figure 1: Hypothesis Testing - A Tale of Two Tails

The tests are all implying that the chances of getting a value in either tail, that far away from the assumed value of the Null Hypothesis, is remote.

Tests can have the following tails:

Left-tailed Test: The Alternative Hypothesis is that the true value is less than the value assumed in the Null Hypothesis – used to test for a negative difference.

Right-tailed Test: The Alternative Hypothesis is that the true value is greater than the value assumed in the Null Hypothesis – used to test for a positive difference

Two-tailed Test: The Alternative Hypothesis is that the value assumed in the Null Hypothesis is simply wrong – used to test that there is simply a difference between the values and it doesn't matter which way

When it comes to our interpretation of "beyond reasonable doubt", we have a choice over the level of probability that constitutes that "reasonable doubt"; we can be excused for thinking that it has to be outside the realms that we can physically determine on a graph, but in truth, rarely is it interpreted with such extreme vigour! Instead the acceptable level of probability or confidence in the result, is dependent on a degree of estimating judgement, or is a matter of custom and practice. (*However, we should challenge custom and practice if we do not think it is appropriate – that's all part of the reality of being an estimator!*)

Depending on the context of the analysis being performed, and the consequences of getting a false positive or a false negative test result, the significance levels chosen are often from, but not restricted to, the values in Table 1. To continue the earlier legal analogy it is recommended that the Significance Level is decided before the data is analysed to avoid "selection bias" in choosing a particular answer; in judicial terms we might be accused of "*leading the witness*".

Possible Context where the consequences of being wrong are …	***Confidence Level for Left or Right-Tailed Tests***	***Confidence Interval for Two-Tailed Tests***
Moderate – used in situations where a general purpose Pareto type of approach is acceptable or where the consequences of getting a Type I Error (False Positive) is highly undesirable	10% / 90%	80%
High – with possible financial losses or minor reputational damage,	5% / 95%	90%
High – with possible heavy financial losses or serious reputational damage	2.5% / 97.5%	95%
Very high – with potentially life-threatening implications e.g. medical research where we want to minimise the risk of a Type II Error (False Negative)	1% / 99%	98%
Unthinkable or untenable (truly beyond reasonable doubt)	0.1% / 99.9%	99.8%

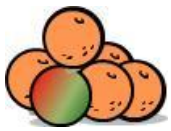
Table 1: Typical Confidence Levels (Significance Levels)

1.1. Mitigation of Type I and Type II Outlier Errors

When we run any statistical test we always run with the risk of being misguided by the data available, i.e. accepting a hypothesis we should have rejected (Type I Error) or rejecting a hypothesis we should have accepted (Type II Error). This is not because we are fundamentally inept, but because the data sample has led us to that conclusion; if we had had different data (possibly just one more data point) we may have reached a different conclusion.

Potentially both types of error may lead to our estimates being skewed one way or the other; leaving in data that we could have legitimately removed, and could lead to a skewed or even atypical answer. However, removing a legitimate "extreme value" outlier can be equally flawed, giving more confidence in a "central" value than is due in the wider reality of things.

For example, suppose we wanted to know the average weight of an orange. To do this we might weigh six pieces of fruit from a bowl, and having made the assumption that they were all oranges, divide by six to get the average weight of an orange. Unfortunately, we may have failed to notice that one of the alleged oranges was in fact an apple. (*Accept it; people do dumb things in life!*)



Total weight of 6 "alleged oranges" = 1200g

Estimated average weight of an orange = 200g

Actual weight of 1 rogue apple = 150g

Actual weight of 5 real oranges = 1050g

True average weight of an orange = 210g

The average weight of a real orange is 5% higher than our estimated weight.

In this example we can resolve the issue without the need for a statistical test by resorting to one of the fundamental principles of estimating: normalisation, or comparing like with like; in this case by eliminating all fruit that are not oranges!

We can always try to mitigate the effects of the outlier by factoring. Assuming that we knew as a "Rule of Thumb" that an orange weighed some 40% more than an apple of similar size, then we could normalise the quantity of oranges in our sample to reflect that we have $5^{5}/_{7}$ equivalent oranges[2].

When we are "certain" that our comparative data is equivalent and comparable, or has been normalised to an acceptable level of equivalence and comparison, then we must stand back and consider whether we still have any outliers.

[2] 1 orange = 1.4 apples = > 1 apple = $^{5}/_{7}$ orange. Note: this relationship is not always true; the data is based on bowl of fruit selected at random from the kitchen at home. *I knew I'd find a practical use for all that fruit!*

Caveat Augur

There is a view that any removal of outliers is inappropriate, especially if we are certain that the data is all drawn from the same population – if an atypical value has occurred once then who's to say it won't occur again? We have to have some sympathy with that argument, but Estimators also have to be pragmatic; there's no point in allowing for the improbable if all we want is the typical value or range of values.

Note: Taking account of the improbable may be better served in many cases by examining the risks, opportunities, or simply the range of uncertainty around the basic task.

Consider the example of some sequential data. If we have two possible outliers in close proximity (e.g. in a time or other natural sequence) where one value is high and the other is low, we need to assure ourselves that we have not uncovered a case of "data contamination" where values have been incorrectly recorded against one event or activity rather than another. A not uncommon example is where manual recording of cost or time is required. Can we really be sure that some of the costs have not "migrated" from one event or activity to the other to a greater extent than might be expected to happen in the normal course of operations as a consequence of human error? We can't be sure of that even when we don't see any obvious outliers – so we have to accept that that is part of the natural "noise" or scatter around the true relationship. However, if we do have two suspected outliers that have cross-contaminated values, we have three options:

1. Reject both the potential outliers (*with a loud "tutting" noise*)
2. Estimate (*or is that "guess"*) the degree of contamination and artificially adjust the "actuals" so that we can then use it along with the other data we have. (*Maybe just take the average of the two for each instead?*)
3. Make an assessment of the potential degree of contamination as above, then put the two outliers to one side. We can then perform the analysis without the two outliers and create our estimate using the remaining data. Finally, we can then use the two adjusted data points that we set aside to test the sensitivity or sensibilitity of our estimate, asking ourselves whether the adjusted data points fit the pattern?

Personally, I would always go with the last one, even though the second one is making some attempt at normalisation.

When we look at the data we should keep an open mind about the nature of the underlying relationship that we expect. For instance, in Figure 2 (left-hand plot) we might suspect that there is a potential outlier against an assumption of a linear relationship, but if we can convince ourselves that the relationship is non-linear (right-hand plot), then the case for a potential outlier diminishes. This does not mean that we should always assume a non-linear relationship just to accommodate an apparent outlier. We have to ask ourselves which relationship makes more sense in it context.

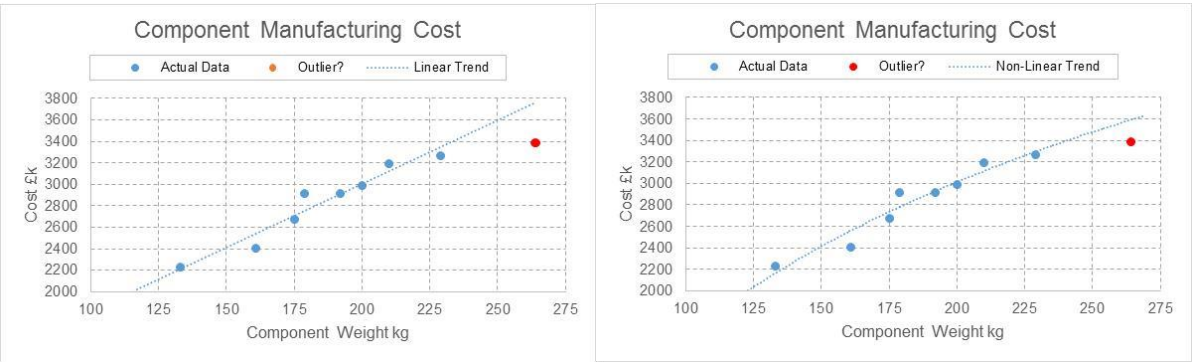


Figure 2: Example – A Linear Outlier may not be a Non-Linear Outlier

At the moment though all we have is conjecture about whether something is an outlier or not. There are a number of tests we can to use to aid us in that decision rather than leave it down to subjectivity (remember that two estimators' subjective opinions are likely to differ.) Before we consider some of the better known ones, they do all have one shortfall in common. Most of them assume that data is distributed Normally (broadly speaking) around some underlying pattern or relationship. This is probably only true for linear relationships. If we think that the relationship is non-linear instead, then other models of scatter are more appropriate. For instance, the Lognormal Distribution is more appropriate for the scatter of the data around a class of non-linear relationships called Power Functions.

If we have a non-linear relationship, then we should always consider whether we can transform it to a linear one before we apply these Outlier tests. If we are unhappy with assumption of normality, we can always try fitting the data scatter to some other non-Normal distribution. It is important where possible to consider the scatter around the assumed relationship.

Let's look at some of the techniques open to us to identify potential outliers.

2. Outing the Outliers: Detecting and Dealing with Outliers

A very important type of test that estimators and other analysts should perform, but one perhaps *(at the risk being accused of an over-generalisation*) that is not always performed quite as formally as it might, is in the detection of Outliers.

There is always the easy option, of which we have probably all been guilty at some time, of looking at some data and excluding a value or two that clearly don't match the pattern formed by the rest of the data. It doesn't mean we were or were not justified in making that judgement call, but we can hardly claim that it was TRACEable[3].

So, what is an "outlier"? The Oxford English Dictionary (Stevenson, 2011) gives us four alternatives:

i. A person or thing situated away or detached from the main body or system

ii. A person or thing differing from all other members of a particular group or set

iii. In Geology: A younger rock formation isolated among older rocks

[3] TRACE = Transparent, Repeatable, Appropriate, Credible and Experientially-based

iv. In Statistics: A data point on a graph or in a set of results that is very much bigger or smaller than the next nearest data point

We would probably accept the first definition, but in relation to the second option, good estimating practice would encourage us to avoid comparing like with unlike as part of our standard normalisation process.

The only link we can probably make to the third definition is that inclusion or exclusion of outliers is a contentious issue, and estimators may find themselves in-between a rock and a hard place whatever they decide to do!

The fourth option sounds quite promising, but fails if we have two similar points close to each other but distant from the rest; we might want to consider them both as potential outliers.

An outlier is sometimes referred to as an "extreme value", implying a very low or high value relative to all others, but this may not be the case as we will see shortly; it may be just "displaced" from a pattern. This "extreme value" view is somewhat one-dimensional, and estimating is often a multi-dimensional problem that we can only resolve by looking at the context within which the data was created.

Let's combine the first and last definition offered by the Oxford English Dictionary.

Definition

Outlier

An outlier is a value that falls substantially outside the pattern of other data. The outlier may be representative of unintended atypical factors or may simply be a value which has a very low probability of occurrence.

Definition 3: Outlier

There is no rigid quantification of what constitutes the degree of displacement of a potential outlier from the rest of the data pack, and each case should be examined on its own merits. Having said that all of the tests we will review have Confidence-based rules associated with them or implied by them.

Consider the three plots of data in each of Figures 3 and 4. We have highlighted one data value or point differently to the rest.

- In Figure 3, Example 1A the highlighted value appears somewhat displaced from the rest of the data. The intuitive response is usually to classify it as an Outlier. On the other hand, Example 1C would usually be considered to be an Inlier (i.e. not an Outlier) because its value is relatively close to the next smallest value. The difficulty arises with the middle one, Example 1B. If we moved it to the right we would probably say "Outlier", whereas if we moved it to the left, we would lean more towards including it as an Inlier.

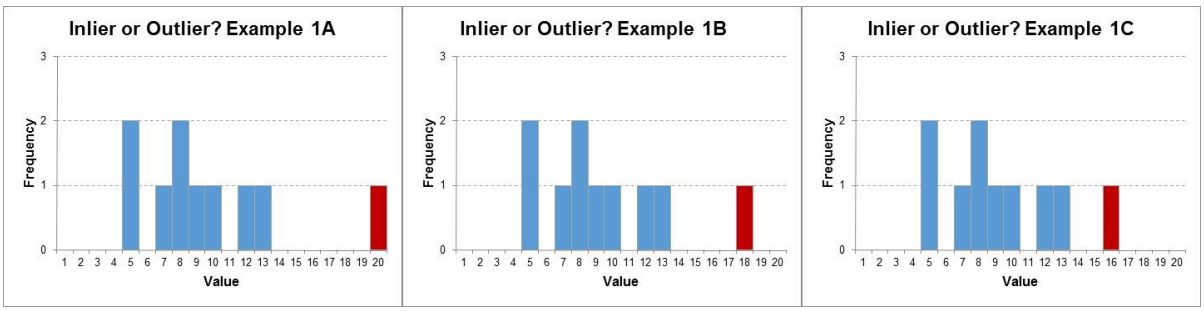


Figure 3: When does an Extreme Value Become an Outlier? Example 1

- The second set of examples in Figure 4 is for the same three datasets but this time we have added some context to each in that we have related the values to some other variable (for example, linking a cost of an item to its weight). Now, we would probably conclude that Example 2A is in fact an Inlier, not an Outlier, and that Example 2C is in fact an Outlier, not an Inlier. The jury is still out on the middle one, Example 2B

.

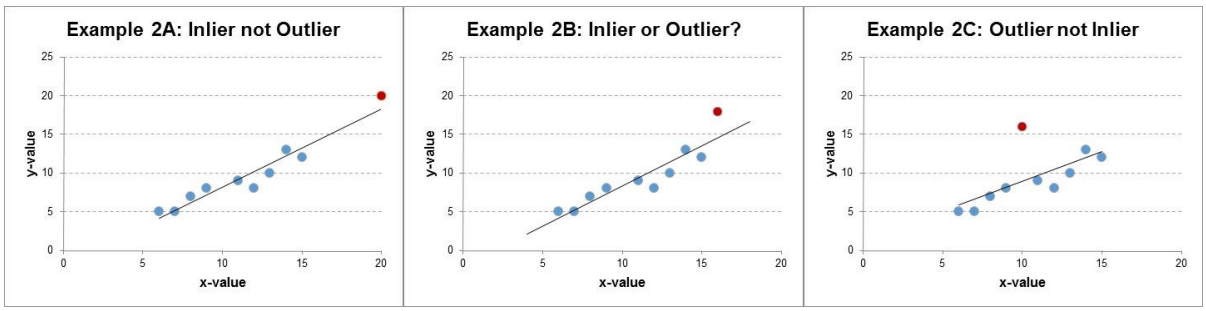


Figure 4: When does an Extreme Value Become an Outlier? Example 2

If we performed an internet search on the detection of outliers, we might conclude that there is almost a plethora of tests and techniques that we can apply to detect outliers; unfortunately, they don't always point us to the same conclusion. In short there is no simple sure-fire test that will say once and for all "*that is an outlier and that is not*", but before we jump straight into considering what's on offer, let's consider a few alternative strategies for dealing with outliers – think of it as a Quality Assurance Step: Preventative Action is better than Corrective Action.

2.1 Tukey Fences

Time for an honesty session: How many of us did a 'double take' here having initially misread the title as 'Turkey Fences'? Hmm, surprisingly many! However, I can assure you that it does read 'Tukey' without an 'r', after John Tukey (1977), who as an esteemed former Professor of Science could never be considered to be a Turkey!

Tukey's technique is very simple and rather elegant, relying on our understanding of Quartiles and Interquartile Ranges in order to define two inner and two outer fences:

1. Calculate the end of the first and third Quartiles of our data range. We can use Microsoft Excel's function[4] **QUARTILE.INC(*array, quart*)** where array is our data range, and quart is an integer referring to the Quartile we are interested in.
2. Calculate the Interquartile Range (IQR) of the data sample. This is simply the difference between Quartile 3 and Quartile 1.
3. The **Upper Inner Tukey Fence** is positioned at the value calculated by adding one and a half times the IQR to the value of the third Quartile
4. The **Lower Inner Tukey Fence** is positioned at the value calculated by subtracting one and a half times the IQR from the value of the first Quartile
5. The **Upper Outer Tukey Fence** is positioned at the value calculated by adding three times the IQR to the value of the third Quartile
6. The **Lower Outer Tukey Fence** is positioned at the value calculated by subtracting three times the IQR from the value of the first Quartile
7. Any data point falling between the Inner and Outer Fences (*on the same side obviously*) is categorised as a "potential outlier"
8. Any data point falling outside the Outer Fences (*on either side*) is deemed to be an "extreme outlier"

Now, the use of the IQR, and the choice of multipliers of one and a half and three, are not some random selection, but relate directly *(in an approximation sense of the word, estimators will be pleased to hear*) to the underlying assumption of Normality i.e. that the sample data is Normal Distributed. Figure 5 illustrates how Tukey Inner Fences are a very close approximation to a Standard Normal Distribution, being close to the Mean $\pm$ 3 Standard Deviations that give us a 99.73% Confidence Interval. (Although Figure 5 relates Tukey Fences to the Standard Normal Distribution, they can equally be mapped against any Normal Distribution.)

[4] In Microsoft Excel 2003 and earlier the function was **QUARTILE(*array, quart*)**

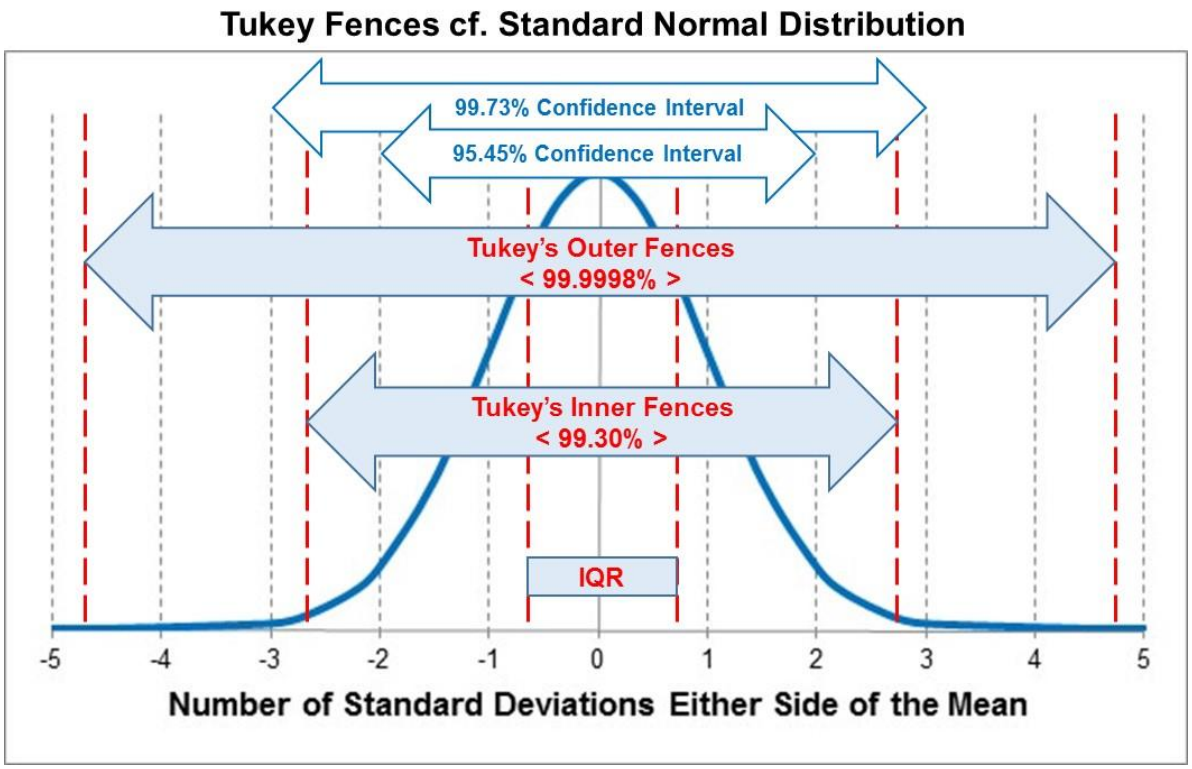


Figure 5: Principle Underpinning Tukey Fences

From a practical standpoint most of us would probably accept that any value beyond ± 3 Standard Deviations from the Mean (equivalent to the bounds of a 99.73% Confidence Interval) would be reasonable grounds for its classification as an outlier. The Inner Tukey Fences (equivalent to ± 2.7 Standard Deviations bounding a Confidence Interval of 99.3%) also sound like a reasonable basis for identifying potential Outliers. (See Table 2 for the supporting data.)

> Now we might wonder why Tukey stopped slightly short of the 3 sigma boundary, when if he had used a multiplier of 1.75 instead of 1.5, he would have been closer to that landmark boundary. *But does it really matter? How precisely inaccurate do we need to be?*
>
> We might also wonder why Tukey didn't try to equate his Inner Fence to being equivalent to a 95% Confidence Interval. In terms of rounded numbers, this would have been equivalent to an IQR Multiplier of one. Let's revisit that in Section 2.2.

If we go beyond the Tukey Inner Fences to the Outer Fences, then these are pitched just outside the four and a half times the standard deviation (4.5 sigma) distance from the Mean. (This nine sigma range forms the basis of the oxymoron we refer to as six-sigma process control.) Anything beyond these points is in the "one in fourteen million" category. *That by anyone's reckoning is an extreme outlier!*

	Tukey Fence Muliplier	Standard Normal Distribution		
		Z-Score	CDF	Comment on Relevance of Points
Lower Outer Tukey Fence	-3	-4.721	0.0001%	Q1 - 3 x IQR
		-4.5	0.0003%	Mean - 4.5 x Standard Deviations
		-3	0.135%	Mean - 3 x Standard Deviations
Lower Inner Tukey Fence	-1.5	-2.698	0.35%	Q1 - 1.5 x IQR
	-1	-2.023	2.15%	Q1 - IQR
		-2	2.28%	Mean - 2 x Standard Deviations
		-1	15.87%	Mean - 1 x Standard Deviations
IQR = Q3-Q1		-0.674	25%	Q1, End of First Quartile
		0.000	50%	Q2, Median
		0.674	75%	Q3, End of Third Quartile
		1	84.13%	Mean + 1 x Standard Deviations
		2	97.72%	Mean + 2 x Standard Deviations
	1	2.023	97.85%	Q3 + IQR
Upper Inner Tukey Fence	1.5	2.698	99.65%	Q3 + 1.5 x IQR
		3	99.865%	Mean + 3 x Standard Deviations
		4.5	99.9997%	Mean + 4.5 x Standard Deviations
Upper Outer Tukey Fence	3	4.721	99.9999%	Q3 + 3 x IQR

Table 2: Tukey Fences in the Context of a Standard Normal Distribution

However, where we have small sample sizes (SSS) it may be more appropriate to assume a Student t-Distribution for the sample, and examine where the Tukey Fences stand in that context.

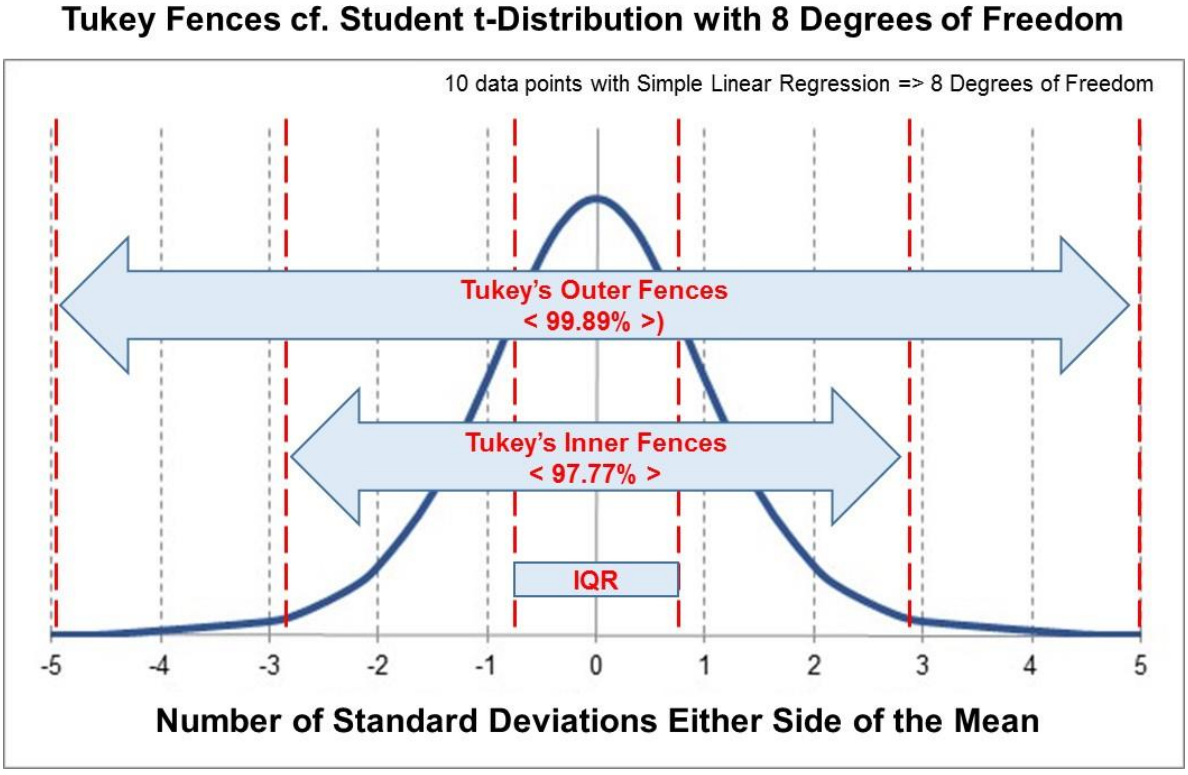


Figure 6: Principle Underpinning Tukey Fences Revisited with a Student t-Distribution

Let's consider the case of a small sample size of 10 data points scattered around a Linear Line of Best Fit, implying 8 degrees of freedom in a Student t-Distribution. We can redraw Figure 5 to get Figure 6. Similarly, if we assume that we have only have a sample size of 6, (and 4 degrees of freedom), to which we want to find the Line of Best Fit, then we would get the Tukey Fences shown in Figure 7. This latter diagram seems to fit with the chosen Outer and Inner Fences more logically around the usual Significance Levels that Statisticians often bandy around of 1% and 5% respectively.

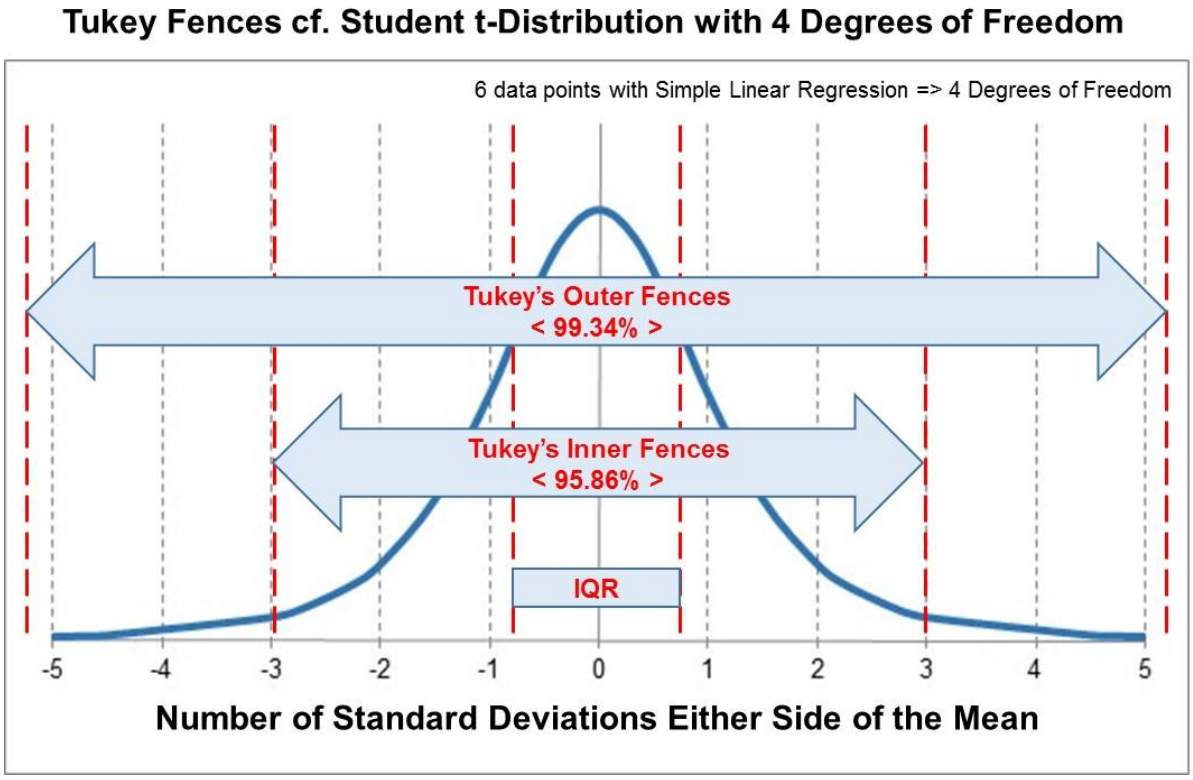


Figure 7: Principle Underpinning Tukey Fences Revisited with a Student t-Distribution

The use of Tukey Fences hinges on the use of Quartiles. (*Hmm, wouldn't that make them gates and not fences?*) However, by interpolation we can imply four quartiles from any two numbers, but from an estimating perspective it is hardly sensible and even less meaningful in the context of identifying outliers. There is a logical argument that the minimum sample size we should consider for quartiles is five, one to define each end of the four quartiles i.e. Minimum, Median, Maximum and the first and third Quartile end points, but even that is stretching the bounds of sensibility. There is perhaps an even stronger argument that there should be at least eight data points thus ensuring at least two data points fall in each quartile.

Let's look at an example in action. For this we will use the data from our earlier Example 2B in Figure 4. From a Confidence Interval perspective, this is equivalent to Figure 6 with 10 data points:

1. Firstly we need to calculate the provisional Line of Best Fit (LoBF, which is only provisional because it will change if we identify and exclude an Outlier.) We can use the **SLOPE(*y-range, x-range*)** and **INTERCEPT(*y-range, x-range*)** functions in Microsoft Excel
2. We can then determine how far each observed point deviates from the LoBF

3. Calculate the First and Third Quartile positions using **QUARTILE.INC(*quart*)** function, and calculate their difference as the Interquartile (IQR) Range
4. Finally, we can construct our Tukey Fences around first and third Quartiles.

The results are demonstrated in Table 3 and Figure 8.

x	y	Line of Best Fit	Deviation from LoBF		Absolute Deviation	Deviation Rank
6	5	4.18	0.82		0.82	5
7	5	5.23	-0.23		0.23	10
8	7	6.27	0.73		0.73	6
9	8	7.31	0.69		0.69	7
11	9	9.40	-0.40		0.40	9
12	8	10.44	-2.44		2.44	2
13	10	11.48	-1.48		1.48	4
14	13	12.52	0.48		0.48	8
15	12	13.56	-1.56		1.56	3
16	18	14.61	3.39	< Not an Outlier	3.39	1

						Fence Multiplier	Fence Position
Count	10						
Mean	11.1	9.5	9.50	0.00			
Std Dev	3.48	3.98	3.63	1.64			
Provisional Regression Slope			1.04				
Provisional Regression Intercept			-2.07				
					Lower Outer	-3	-7.00
			Quartile 0	-2.44	Lower Inner	-1.5	-4.10
		IQR	Quartile 1	-1.21			
			Quartile 2	0.13	1.93		
			Quartile 3	0.72			
			Quartile 4	3.39	Upper Inner	1.5	3.61
					Upper Outer	3	6.51

Table 3: Example of Tukey Fences Based on Line of Best Fit Deviations

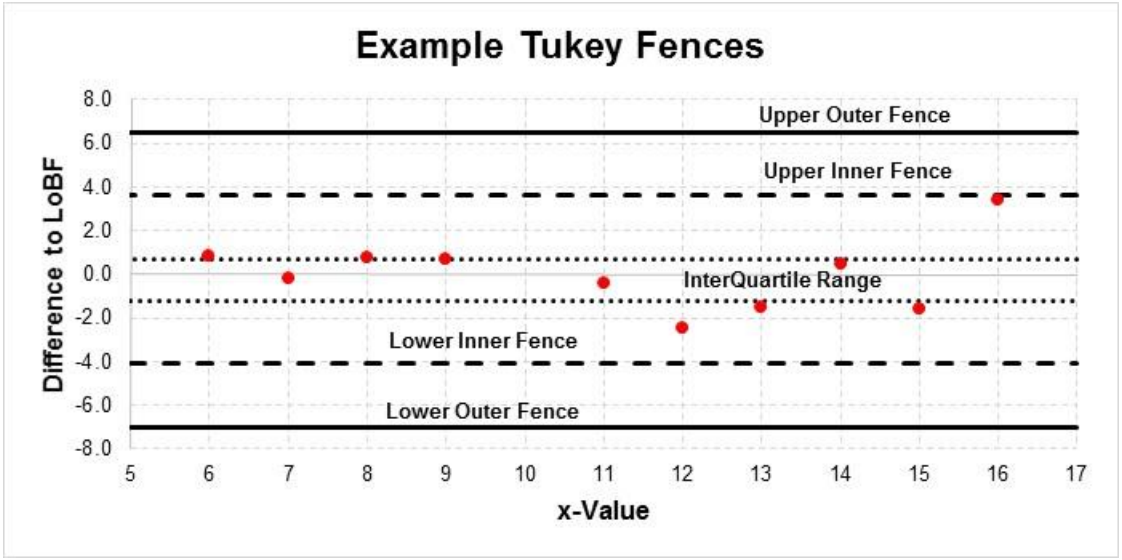


Figure 8: Example of Tukey Fences Based on Line of Best Fit Deviations

As we suspected in Example 2B earlier, this was going to be close. In this particular case the last data point (16,18) is not being flagged as a potential outlier by Tukey Inner or Outer Fences, lying just inside the Upper Inner Fence, which is at the 98.88% Confidence Level.

Despite (*or should that read 'because of'*) their simplicity and elegance, the integrity of Tukey Fences is not necessarily maintained if we add a potential outlier to the pot! For instance, suppose we find an extra value to add to our sample, Let's suppose that it is y=12 when x=10. We may find the results surprising (Table 4 and Figure 9.) All of a sudden we get two potential outliers … the new point, AND the one that we just decided wasn't an outlier!

x	y	Line of Best Fit	Deviation from LoBF		Absolute Deviation	Deviation Rank
6	5	4.68	0.32		0.32	8
7	5	5.69	-0.69		0.69	7
8	7	6.70	0.30		0.30	9
9	8	7.71	0.29		0.29	10
11	9	9.73	-0.73		0.73	6
12	8	10.74	-2.74		2.74	3
13	10	11.75	-1.75		1.75	5
14	13	12.75	0.25		0.25	11
15	12	13.76	-1.76		1.76	4
16	18	14.77	3.23	< Potential Outlier	3.23	2
10	12	8.72	3.28	< Potential Outlier	3.28	1

Count	11					
Mean	11	9.73	9.73	0.00		
Std Dev	3.32	3.85	3.35	1.90		
Provisional Regression Slope			1.01			
Provisional Regression Intercept			-1.37			
			Quartile 0	-2.74		
		IQR	Quartile 1	-1.24	1.55	
			Quartile 2	0.25		
			Quartile 3	0.31		
			Quartile 4	3.28		

	Fence Multiplier	Fence Position
Lower Outer	-3	-5.87
Lower Inner	-1.5	-3.55
Upper Inner	1.5	2.63
Upper Outer	3	4.95

Table 4: Impact of an Additional Data Point on Tukey Fences Based on Line of Best Fit Deviations

As its deviation from the Line of Best Fit is only marginally greater than that of the original point that was just inside the inner fence from (Table 4), then intuitively we may have expected that this would have been similarly positioned, or at worst would have just popped over the fence onto the Potential Outlier side … if anything, as the deviation is very similar to the original suspect point then we may have expected it to confirm that neither point was an Outlier, not drag the other one with it into "no man's land" between the Tukey Inner and Outer Fences. It is not unreasonable in some people's mind to expect this to be flagged as a Potential Outlier. Unfortunately, life as an Estimator is full of disappointments and unwanted surprises. As we will observe from Table 4 the addition of the extra point has moved the goalposts.

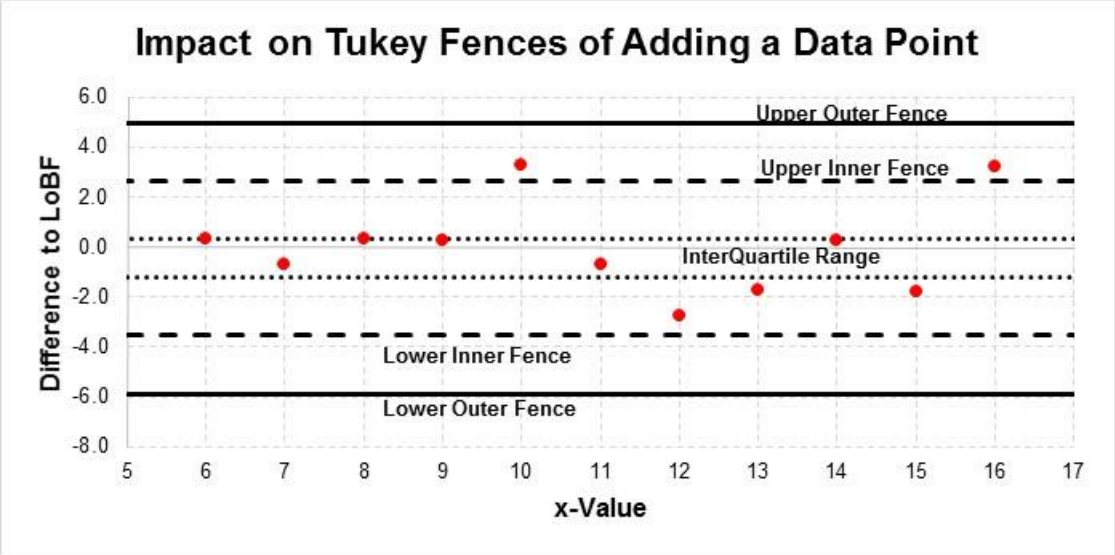


Figure 9: Example of Tukey Fences Based on Line of Best Fit Deviations

We might conclude from this that Tukey Fences are not particularly robust for small sample sizes. However, here we have a peculiar set of values. The Median is a more robust measure of Central Tendency than the common Average or Arithmetic Mean, as it is less susceptible to change by an Outlier. Tukey's technique doesn't directly use the Median, but the Interquartile range around the Median. The introduction of the extra point has reduced the Median by equivalent of half a Rank Position, but it has also changed the third Quartile quite significantly, creating a knock-on effect to the Tukey Fence Positions.

If we follow the advice and eliminate one outlier at a time then we would first remove the data point (10,12) as a potential Outlier, and then re-do the test with the remaining points, which as we saw previously suggests that this is not an Outlier. As the Deviations from the Line of Best Fit are so similar, we may feel a little uncomfortable rejecting one and not the other. The key to this is that word "potential". Perhaps we should try an alternative test?

… let's just reflect for a moment on Confidence Intervals

You may have spotted something of an inconsistency or double standards being applied when it comes to Outlier Detection with Tukey Fences and Hypothesis Testing. Surely the determination of potential outliers is a matter of hypothesizing on the existence of an outlier and then testing that hypothesis.

When it comes to thresholds or Critical Values for Hypothesis Testing we are frequently happy to accept a 95% Confidence Interval, and sometimes as low as 90%, yet here we are with Tukey Fences pushing the boundaries as it were out to 99.3%.

As we saw with Figures 5 to 7 the Confidence Interval associated with Tukey Inner Fences varies depending on the number of data points (and therefore Degrees of Freedom).

With this in mind, for larger Sample Sizes, we might want to consider what we will call Tukey Slimline Fences.

2.2 Tukey Slimline Fences – For Larger Samples

If we refer back to Figure 7, the Tukey Inner Fences with only 6 data points is equivalent to a Confidence Interval of some 96%. This is reasonably comparable with a Confidence Interval of some 95% for a Normal Distribution Range of the Mean ± 2 Standard Deviations.

If we have a large data sample (nominally greater than 30 data points) then there may be a case, depending on the criticality of the estimate being produced, to use Tukey Fences with an IQR multiplier of ± 1 for Tukey Inner Fences to identify Potential Outliers, and ± 2 for the Outer Fences.

However, we should not reject Potential Outliers determined in this manner without first performing another more rigorous test.

Finally, in Microsoft Excel 2010[5] and later we can calculate Q1 and Q3, using the inclusive Quartile function **QUARTILE.INC(*array, quart*)** where ***quart*** takes the parameter value 1 or 3 and ***array*** is the sample array. Consequently, we can derive Tukey Fences (*either the Full Fat or Slimline version*) based on an appropriate multiplier value, simply in relation to the first and third quartiles in a single step for each fence.

For the Formula-philes: Tukey Fences in One Step

Consider a range of values x_i to x_n in a sample. Denote the First and Third Quartile end points as *Q1* and *Q3*. Consider also, a positive constant, *m*, to be used as the Interquartile Range multiplier in determining Tukey Fences.

The Interquartile Range, *IQR*, is:	$IQR = Q3 - Q1$	(1)
Using (1) the Lower Tukey Fence, *LTF*, based on multiplier, *m* applied to the IQR, is:	$LTF = Q1 - m(Q3 - Q1)$	(2)
Similarly, the corresponding Upper Tukey Fence, UTF, can be expressed as:	$UTF = Q3 + m(Q3 - Q1)$	(3)
Re-arranging (2):	$LTF = (1 + m)Q1 - mQ3$	(4)
Re-arranging (3):	$UTF = (1 + m)Q3 - mQ1$	(5)

… typically the Inner Tukey Fences would use a value of m = 1.5, and the Outer Tukey Fences would use a value of m = 3, but any modified multiplier could be used

[5] In earlier versions of Excel, the function was simply **QUARTILE(*array, quart*)**.

2.3 Chauvenet's Criterion

Chauvenet's Criterion (Chauvenet, 1863) is also based on that assumption of Normality *(I agree; when was the life of an estimator ever normal?*) It calculates the number of data points we might reasonably expect to get from a known sample size based on the Cumulative Distribution Function (CDF) of a Normal Distribution; in effect it uses the Z-Score. The procedure is simple enough. (If we are looking at one-dimensional data rather than a scatter around a line of best fit, then this procedure must be adapted to reflect a standard Z-Score.)

1. Count the number of observations in the sample
2. Calculate the Line of Best Fit (LoBF) using Microsoft Excel **SLOPE(*y-range, x-range*)** and **INTERCEPT(*y-range, x-range*)** functions
3. Calculate the Deviation (difference) between each point and the Line of Best Fit
4. Calculate the Deviation Mean using **AVERAGE(*range*)** in Microsoft Excel. Note: this should always be zero. If it is not, then there is something wrong with the LoBF.
5. Calculate the Standard Deviation of the point Deviations using **STDEV.S(*range*)** in Excel
6. Calculate the absolute value for a Z-Score for each Deviation (we can use **ABS(*Z*)** in Excel), i.e. find the absolute value of the difference from Step 3 (ignoring any negative signs) between each point's deviation value divided by the Standard Deviation of the Deviations from Step 6.
    - Strictly speaking, we should deduct the Deviation Mean from each Point's Deviation, but as it is zero (Step 4), it is irrelevant here.
7. Determine the probability of getting a Z-Score larger than that calculated in Step 6 for each point (we can use complementary function **1-NORM.S.DIST(ABS(*Z*), *TRUE*)** in Excel for this.) We must multiple this by two to get a two-tailed probability.
8. Multiply the Z-Score probability for each observation by the total number of observations counted in Step 1
9. Round the answer from Step 8 to the nearest integer. This then represents the number of observations we would reasonably expect to get this far from the Deviation Mean of zero with the sample size in question
10. Any observation with a net score of zero can be deemed to be a potential outlier

If we identify any potential outliers, we can then set them aside and repeat the procedure until there are no additional potential outliers detected.

Applying Chauvenet's Criterion to the data in Example 2B (reproduced in the upper half of Figure 10) we get the results in Table 5 which tell us that we shouldn't expect any points to be as far away from the line of best fit as (16,18) appear to be. This blatantly contradicts the result indicated previously by Tukey Fences! (*Oh dear!*)

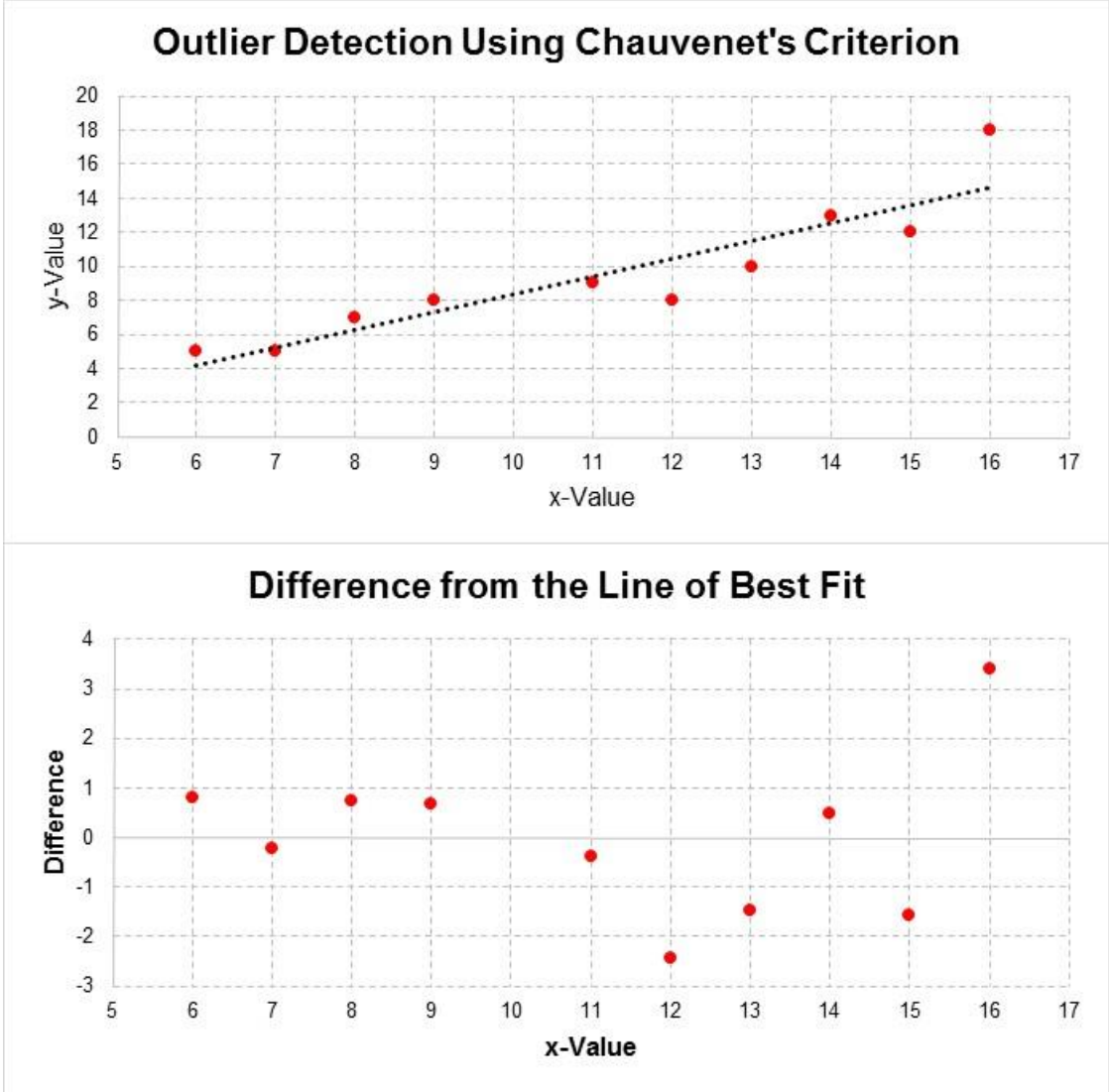


Figure 10: Example of Using Chauvenet's Criterion to Detect an Outlier

x	y	Line of Best Fit	Difference to LoBF	Absolute Z-Score	Prob > Z ~ N(0,1)	Expected # Points	Rounded # Points	
6	5	4.18	0.82	0.497	61.9%	6.19	6	
7	5	5.23	-0.23	0.138	89.0%	8.90	9	
8	7	6.27	0.73	0.446	65.6%	6.56	7	
9	8	7.31	0.69	0.420	67.5%	6.75	7	
11	9	9.40	-0.40	0.241	80.9%	8.09	8	
12	8	10.44	-2.44	1.487	13.7%	1.37	1	
13	10	11.48	-1.48	0.903	36.7%	3.67	4	
14	13	12.52	0.48	0.291	77.1%	7.71	8	
15	12	13.56	-1.56	0.954	34.0%	3.40	3	
16	18	14.61	3.39	2.069	3.9%	0.39	0	< Outlier

Count	10				
Mean	11.1	9.5	9.50	0.00	↑
Std Dev	3.48	3.98	3.63	1.64	↑
Provisional Regression Slope			1.04		
Provisional Regression Intercept			-2.07		

Table 5: Example Use of Chauvenet's Criterion to Detect Potential Outliers

Despite contradicting the conclusion indicated by Tukey Fences, the Chauvenet approach seems to be quite a reasonable one on the face of it for detecting a potential outlier. Whether we then choose to exclude the outlier from our analysis is a separate issue. However, as the number of observations or data points increases so too does the threshold or Critical Value of the Z-Score by which we calculate the number of observations that we might reasonably expect (i.e. equivalent to Step 8 in our procedure). Table 6 highlights the issue it then gives us:

- The advantage that this technique gives us is an objective measure with a repeatable procedure.
- Its shortcoming is that where the outlier is close to the Critical Value, then one more or one less point may pull it or push it back over the "wall".

Sample Size	Min Z-Score Probability to get at least one value	Potential Outlier when Absolute Z-Score Exceeds
4	12.5%	1.534
5	10.0%	1.645
6	8.3%	1.732
7	7.1%	1.803
8	6.3%	1.863
9	5.6%	1.915
10	5.0%	1.960
12	4.2%	2.037
15	3.3%	2.128
20	2.5%	2.241
25	2.0%	2.326
33	1.5%	2.429
50	1.0%	2.576
75	0.7%	2.713
100	0.5%	2.807

Table 6: Chauvenet's Criterion Critical Values

The principle that underpins Chauvenet's Criterion is that it assumes probabilistically that we have a greater chance of getting a more remote value with larger sample sizes. This then implies that the Critical Value of the Z-Score increases with the number of data points.

Some of us may find this disappointing as we might feel inclined to argue that a point is either an outlier or it is not! However, if we reflect on the Z-Score calculation then perhaps it is not so bad as there is a degree of compensation inherent in the calculation.

For the Formula-phobes: Z-Score Critical Values

Suppose we have a small sample with a known outlier of a high value. Suppose that we increase the size of the sample with other values that are typical of the main body of data i.e. no more outliers. The Mean of the original sample will be skewed to the right in comparison with the larger sample, as the contribution made by the outlier's inflated value will be diluted by dividing it by a larger sample size quantity. The deviation from the sample mean of the outlier is greater in the case of the large sample.

Similarly, the Standard Deviation of the larger sample will be smaller too, but the effect of the larger sample size quantity is reduced by the square root function used in its calculation.

Sample	Point Number	Value	Mean	Std Dev	Z-Score	
1	1	3	7	4.83	0.828	
	2	5			0.414	
	3	6			0.207	
	4	14			1.449	< Outlier
2	1	3	6	3.63	0.828	
	2	5			0.276	
	3	6			0.000	
	4	14			2.207	< Outlier
	5	5			0.276	
	6	4			0.552	
	7	8			0.552	
	8	3			0.828	

As with Tukey Fences, we should resist any temptation to reject multiple outliers in a single iteration, especially if:

- The points are at either end of the scale in relation to the main body of the data
- The two points are not physically close to each other

As Estimators, we are all prone to asking the question "what if". For example:

- In our example in Table 5, if the first point (6,5) was not available to us, then the value (16,18) would NOT be a potential outlier according to Chauvenet's Criterion. (*This would also have been the case with Tukey, by the way*.)
- The question many of us are probably thinking is "what if we added that extra data point as we did with the Tukey Fences example?" (*Now is that me being clairvoyant or what?*)

Let's do that. In Table 7 we have added the point (10,12) and re-run our calculations.

This has gone in the opposite direction to Tukey Fences! This test is saying that neither of the two suspect points are outliers. Now some of us may be thinking words that we cannot print but they boil down to "*Why does this happen?*" or perhaps even "*Statistics! I always said it was just all smoke and mirrors!*" However, this is not always the case, the two tests are often consistent with each other, but sometimes we can get an arrangement of values where the bizarre happens.

x	y	Line of Best Fit	Difference to LoBF	Absolute Z-Score	Prob > |Z| ~ N(0,1)	Expected # Points	Rounded # Points	
6	5	4.68	0.32	0.167	86.7%	9.54	10	
7	5	5.69	-0.69	0.363	71.6%	7.88	8	
8	7	6.70	0.30	0.158	87.5%	9.62	10	
9	8	7.71	0.29	0.153	87.8%	9.66	10	
11	9	9.73	-0.73	0.382	70.2%	7.72	8	
12	8	10.74	-2.74	1.439	15.0%	1.65	2	
13	10	11.75	-1.75	0.918	35.9%	3.95	4	
14	13	12.75	0.25	0.129	89.7%	9.87	10	
15	12	13.76	-1.76	0.927	35.4%	3.89	4	
16	18	14.77	3.23	1.697	9.0%	0.99	1	< Not an Outlier
10	12	8.72	3.28	1.726	8.4%	0.93	1	< Not an Outlier

Count	11				
Mean	11	9.73	9.73	0.00	↑
Std Dev	3.32	3.85	3.35	1.90	↑
Provisional Regression Slope			1.01		
Provisional Regression Intercept			-1.37		

Table 7: Impact of an Additional Data Point on Chauvenet's Criterion Based on LoBF Deviations

In the context of Chauvenet's Criterion why has a case of "one potential outlier" turned into a case of "no outliers"? If we think about it, it begins to make some sense:

- If we have a relatively small number of observations and we remove one from the "middle ground" then we will have less evidence to support the Measures of Central Tendency as being representative of the whole data set, so the distribution flattens and widens in effect.
- On the other hand, if we add another point in the region of the first potential outlier, we are in effect moving the Sample Mean towards that "distribution tail" and also widening the dispersion. The net result is a lowering of the absolute value of the Z-Score which in turn reduces the chance of an outlier.

With a small number of random observations in our sample we have a greater chance of having an unrepresentative distribution which means that we may identify a potential outlier at a relatively low Z-Score – probably not a good idea … but perhaps there is something that we can do about it?

2.4 Variation on Chauvenet's Criterion for Small Sample Sizes (SSS)

In Section 2.1 we introduced a discussion on the Student t-Distribution; we said that we consider a Student t-Distribution to be the Small Sample Size Equivalent of a Normal Distribution. The scatter of data points around a Line of Best Fit (LoBF) will be a Student t-Distribution with degrees of freedom of two less than the number of data points. It only approximates to a Normal Distribution for larger sample sizes (>30). Perhaps we should then look at our normalised deviation Z-statistic as a Student t-Distribution instead of as a Normal Distribution. In Tables 8 and 9 we have revisited our two Chauvenet's Criterion examples from Tables 5 and 7 but replaced the Probability calculation with a two-tailed t-Distribution. We can do this using the Microsoft Excel Function **1-T.DIST.2T(*x*,*deg_freed*)**.

In both cases, the revised test is suggesting that these points are not Outliers, as we can expect one value of each that far from the line of best fit, unlike the traditional Chauvenet's Criterion based on a Normal Distribution.

x	y	Line of Best Fit	Difference to LoBF	Absolute Z-Score	Prob > |Z| ~ t(0,n-2)	Expected # Points	Rounded # Points	
6	5	4.18	0.82	0.497	63.2%	6.324	6	
7	5	5.23	-0.23	0.138	89.3%	8.934	9	
8	7	6.27	0.73	0.446	66.8%	6.676	7	
9	8	7.31	0.69	0.420	68.6%	6.856	7	
11	9	9.40	-0.40	0.241	81.5%	8.154	8	
12	8	10.44	-2.44	1.487	17.5%	1.754	2	
13	10	11.48	-1.48	0.903	39.3%	3.931	4	
14	13	12.52	0.48	0.291	77.8%	7.783	8	
15	12	13.56	-1.56	0.954	36.8%	3.680	4	
16	18	14.61	3.39	2.069	7.2%	0.723	1	< Not an Outlier

Count	10				
Mean	11.1	9.5	9.50	0.00	↑
Std Dev	3.48	3.98	3.63	1.64	↑
Provisional Regression Slope			1.04		
Provisional Regression Intercept			-2.07		

Table 8: Example of Substituting a t-Distribution into Chauvenet's Criterion

x	y	Line of Best Fit	Difference to LoBF	Absolute Z-Score	Prob > |Z| ~ t(0,n-2)	Expected # Points	Rounded # Points	
6	5	4.68	0.32	0.167	87.1%	9.58	10	
7	5	5.69	-0.69	0.363	72.5%	7.97	8	
8	7	6.70	0.30	0.158	87.8%	9.66	10	
9	8	7.71	0.29	0.153	88.2%	9.70	10	
11	9	9.73	-0.73	0.382	71.1%	7.82	8	
12	8	10.74	-2.74	1.439	18.4%	2.02	2	
13	10	11.75	-1.75	0.918	38.3%	4.21	4	
14	13	12.75	0.25	0.129	90.0%	9.90	10	
15	12	13.76	-1.76	0.927	37.8%	4.16	4	
16	18	14.77	3.23	1.697	12.4%	1.36	1	< Not an Outlier
10	12	8.72	3.28	1.726	11.9%	1.30	1	< Not an Outlier

Count	11				
Mean	11	9.73	9.73	0.00	↑
Std Dev	3.32	3.85	3.35	1.90	↑
Provisional Regression Slope			1.01		
Provisional Regression Intercept			-1.37		

Table 9: Impact of an Additional Data Point on Chauvenet's Criterion Using a t-Distribution

Perhaps it may help us to understand what is going on here if we look at the Q-Q Plots for 9, 10 and 11 data points (Figure 11), in which the only differences are the two suspect data points.

If we reject both the suspected outliers we definitely get a better Q-Q Plot, closer to a true linear relationship (left hand plot). However, the slight mirrored S-Curve is suggestive that a Student t-Distribution may be a potentially better solution, based on our previous Figure 15.

When leave the first suspect data point in our Q-Q Plot we still have a reasonable straight line (centre plot), albeit not as good.

The addition of the extra data point in the right hand plot, makes a marginal improvement in the straight line Q-Q Plot. This is supported by the increase in probability associated with these two points in the revised Chauvenet's Criterion we calculated in Tables 8 and 9.

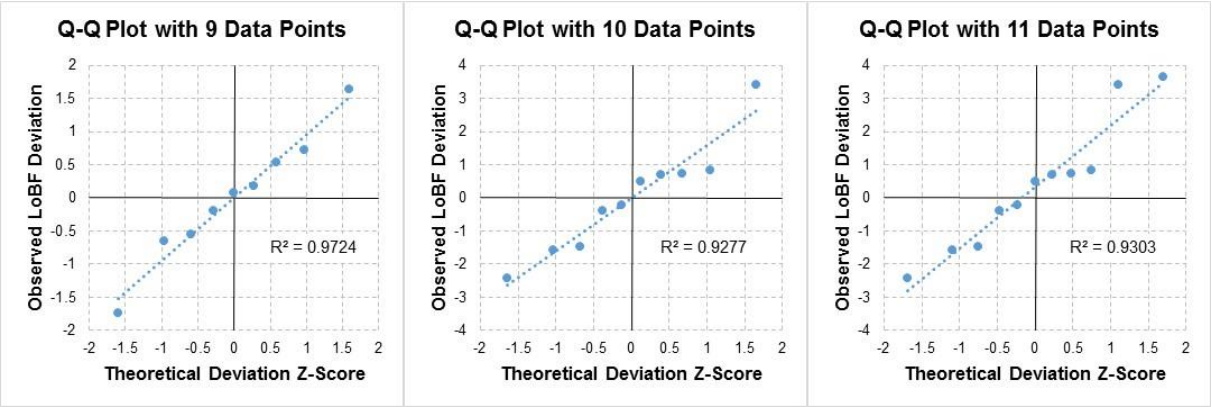


Figure 11: Q-Q Plot Comparison for Example Data

2.5 Peirce's Criterion

Peirce's Criterion (Peirce, 1852; Gould, 1855) pre-dates Chauvenet's Criterion by some eleven years and can be applied to cases of multiple potential outliers. To use it we require access to a set of tables for the "Maximum Allowable Deviation", R, for "Ratio" as Peirce called it (*not to be confused with Pearson's Linear Correlation Coefficient, R.*) Peirce's R is again based on an assumption of Normality but is less tolerant of outliers than Chauvenet's Criterion until we get a sample size of 35 or more (Table 10).

Chauvenet (1863) commented that Peirce's Criterion and associated procedure was statistically more robust than his proposal. Gould (1855) made a valiant effort to create a Table of Critical Values based on a rather convoluted-looking formula, and these are still used today, but there are also a few anomalies in comparison to Peirce's original work. The Ratio used in comparison with the Maximum Allowable Deviation, R, is the same as the Z-Score that we use for Chauvenet's Criterion, and as we will see, for some other tests as well.

The procedure for Peirce's Criterion is outlined succinctly by Ross (2003):

1. Calculate the Mean and Standard Deviation for the sample in question
2. Assume one potential outlier initially, and obtain R from the Table of Critical Values for the appropriate sample size (Table 10)
3. Calculate the Z-Score for each point
4. Compare the value from Step 2 with that from Step 3, and mark the data point as an outlier if the Max Z-Score (Step 3) is greater than the Critical Value, R (Step 2)
5. Now assume that we have two outliers (still keeping all the data points and the data calculated for the mean, standard deviations and Z-Scores)

Sample Size	Peirce's Criterion									Chauvenet's Criterion
	Number of Suspected Data Points									Min Z-Score Probability to get at least one value
	1	2	3	4	5	6	7	8	9	
3	1.196									1.383
4	1.383	1.078								1.534
5	1.509	1.200								1.645
6	1.610	1.299	1.099							1.732
7	1.693	1.382	1.187	1.022						1.803
8	1.763	1.453	1.261	1.109						1.863
9	1.824	1.515	1.324	1.178	1.045					1.915
10	1.878	1.570	1.380	1.237	1.114					1.960
11	1.925	1.619	1.430	1.289	1.172	1.059				2.000
12	1.969	1.663	1.475	1.336	1.221	1.118	1.009			2.037
13	2.007	1.704	1.516	1.379	1.266	1.167	1.070			2.070
14	2.043	1.741	1.554	1.417	1.307	1.210	1.120	1.026		2.100
15	2.076	1.775	1.589	1.453	1.344	1.249	1.164	1.078		2.128
16	2.106	1.807	1.622	1.486	1.378	1.285	1.202	1.122	1.039	2.154
17	2.134	1.836	1.652	1.517	1.409	1.318	1.237	1.161	1.084	2.178
18	2.161	1.864	1.680	1.546	1.438	1.348	1.268	1.195	1.123	2.200
19	2.185	1.890	1.707	1.573	1.466	1.377	1.298	1.226	1.158	2.222
20	2.209	1.914	1.732	1.599	1.492	1.404	1.326	1.255	1.190	2.241
21	2.230	1.938	1.756	1.623	1.517	1.429	1.352	1.282	1.218	2.260
22	2.251	1.960	1.779	1.646	1.540	1.452	1.376	1.308	1.245	2.278
23	2.271	1.981	1.800	1.668	1.563	1.475	1.399	1.332	1.270	2.295
24	2.290	2.000	1.821	1.689	1.584	1.497	1.421	1.354	1.293	2.311
25	2.307	2.019	1.840	1.709	1.604	1.517	1.442	1.375	1.315	2.326
26	2.324	2.037	1.859	1.728	1.624	1.537	1.462	1.396	1.336	2.341
27	2.341	2.055	1.877	1.746	1.642	1.556	1.481	1.415	1.356	2.355
28	2.356	2.071	1.894	1.764	1.660	1.574	1.500	1.434	1.375	2.369
29	2.371	2.088	1.911	1.781	1.677	1.591	1.517	1.452	1.393	2.382
30	2.385	2.103	1.927	1.797	1.694	1.608	1.534	1.469	1.411	2.394
31	2.399	2.118	1.942	1.812	1.710	1.624	1.550	1.486	1.428	2.406
32	2.412	2.132	1.957	1.828	1.725	1.640	1.567	1.502	1.444	2.418
33	2.425	2.146	1.971	1.842	1.740	1.655	1.582	1.517	1.459	2.429
34	2.438	2.159	1.985	1.856	1.754	1.669	1.597	1.532	1.475	2.440
35	2.450	2.172	1.998	1.870	1.768	1.683	1.611	1.547	1.489	2.450
36	2.461	2.184	2.011	1.883	1.782	1.697	1.624	1.561	1.504	2.460
37	2.472	2.196	2.024	1.896	1.795	1.711	1.638	1.574	1.517	2.470
38	2.483	2.208	2.036	1.909	1.807	1.723	1.651	1.587	1.531	2.479
39	2.494	2.219	2.047	1.921	1.820	1.736	1.664	1.600	1.544	2.489
40	2.504	2.230	2.059	1.932	1.832	1.748	1.676	1.613	1.556	2.498

Table 10: Peirce's R Values for Suspect Data Compared with Chauvenet's Criterion Critical Values

6. Look up the Critical Value of R from the Table 10 for two outliers for the appropriate sample size
7. If this results in two values exceeding the Maximum Allowable Deviation, then mark them both as outliers and continue to the next highest number of suspect points (and so on)

8. If only one data point falls inside the Critical Value of R, then stop and only reject the previous outliers.
9. We can now recalculate the sample mean and standard deviation based on the remaining data

Whilst in theory, as Table 10 implies, we can apply Peirce's Criterion to very small sample sizes, or moderately small ones, where we have a significant number of "suspect" data points, However, we should really be questioning whether we should be using any outlier test on such a high proportion of the data. For example, if we had a sample size of 9, would it be reasonable to classify (and potentially remove) 5 of them?

In Table 11 we apply Peirce's Criterion to the original example data we used for Tukey Fences and Chauvenet's Criterion, taking the Critical Value of Peirce's R from Table 10 based on a sample size of 10. In this case, first assuming one outlier, the test agrees with Chauvenet's Criterion that the point furthest from the Line of Best Fit is an outlier. We can then move to the next stage where we assume two outliers and re-test. This time Peirce's Criterion indicates that the furthest point from the Line of Best Fit is indeed an outlier, but that the next nearest is not. In conclusion, the furthest point from the Line of Best Fit is an Outlier.

x	y	Line of Best Fit	Difference to LoBF	Absolute Z-Score	Reverse Rank	Assume 1 Outlier	Assume 2 Outliers
6	5	4.18	0.82	0.497	5		
7	5	5.23	-0.23	0.138	10		
8	7	6.27	0.73	0.446	6		
9	8	7.31	0.69	0.420	7		
11	9	9.40	-0.40	0.241	9		
12	8	10.44	-2.44	1.487	2		< Not an Outlier
13	10	11.48	-1.48	0.903	4		
14	13	12.52	0.48	0.291	8		
15	12	13.56	-1.56	0.954	3		
16	18	14.61	3.39	2.069	1	< Outlier	< Outlier

Count	10				
Mean	11.1	9.5	9.50	0.00	↑
Std Dev	3.48	3.98	3.63	1.64	↑
Provisional Regression Slope			1.04		
Provisional Regression Intercept			-2.07		

Peirce's R Value	
1.878	1.570

Table 11: Example of the Application of Peirce's Criterion

In Table 12 we have re-run the test for the second example in which we added the second potential outlier. On the first pass of the test, on the assumption of one suspect value, the Z-Score is less than Peirce's Critical Value of R, and therefore we would not reject the most distant point from the Line of Best Fit. If we were to follow Peirce's procedure as described by Ross (2003) then we would not proceed to a second stage of assuming two outliers. However, if we started with the assumption of two outliers then this would lead us to the same conclusion that neither of the two suspect points are indeed outliers.

In both cases, Peirce's Criterion gives the same results as Chauvenet's Criterion. However, in different circumstances, i.e. alternative values, they could easily have given us conflicting answers.

x	y	Line of Best Fit	Difference to LoBF	Absolute Z-Score	Reverse Rank	Assume 1 Outlier	Assume 2 Outliers
6	5	4.68	0.32	0.167	8		
7	5	5.69	-0.69	0.363	7		
8	7	6.70	0.30	0.158	9		
9	8	7.71	0.29	0.153	10		
11	9	9.73	-0.73	0.382	6		
12	8	10.74	-2.74	1.439	3		
13	10	11.75	-1.75	0.918	5		
14	13	12.75	0.25	0.129	11		
15	12	13.76	-1.76	0.927	4		
16	18	14.77	3.23	1.697	2		*Test not Required*
10	12	8.72	3.28	1.726	1	< Not an Outlier	*Test not Required*

						Peirce's R Value	
Count	11						
Mean	11	9.73	9.73	0.00	↑	1.925	1.619
Std Dev	3.32	3.85	3.35	1.90	↑		
Provisional Regression Slope			1.01				
Provisional Regression Intercept			-1.37				

Table 12: Example of the Application of Peirce's Criterion with Additional Suspect Data Point

2.6 Iglewicz and Hoaglin's MAD Technique

Most of the Outlier Tests require the calculation of a value based on the Mean of the sample in question. Unfortunately, the Mean is not a robust statistic and is sensitive to changes in the constituent data, such as potential outliers. The Median of the other hand is more robust and will not vary as significantly if a potential outlier is present or added to the sample.

That's probably where Iglewicz and Hoaglin (1993) got their idea for a MAD Technique. It centres (*pun intended*) on a double Median … the Median of the absolute deviations from the Median, otherwise known as the Median Absolute Deviation or MAD for short. It is based on the Z-Score but uses the Median instead of the Mean and the Median Absolute Deviation instead of the Standard Deviation. Iglewicz and Hoaglin called this their Modified Z Score or M-Score.

Let's go through the procedure in Table 13 using the first of our two examples that we have been using throughout this section.

1. Calculate the Line of Best Fit (LoBF) using Microsoft Excel **SLOPE(*y-range, x-range*)** and **INTERCEPT(*y-range, x-range*)** functions
2. Calculate the Deviation (difference) between each point and the Line of Best Fit. (The average or Mean Deviation should be zero for reference.)
3. Calculate the Deviation Median using **MEDIAN(*range*)** in Excel.
4. Calculate the Absolute Deviation from the Deviation Median for each point using **ABS(*Point Deviation – Median Deviation*)** in Excel
5. Calculate the Median Absolute Deviation (MAD) by taking the Median of the individual Absolute Point Deviations from Step 4
6. Calculate the M-Score for each point by multiplying the point's Absolute Deviation from the Median (from Step 4) by the constant 0.6745 and dividing by the Median Absolute Deviation (from Step 5)
7. If a point is greater than a Critical Value of 3.5, then it can be classed as an outlier

Using this technique and critical value we can see that our suspect point (16,18) should not be classed as an Outlier as the M-Score is less than the recommended value of 3.5 (*but only just … as was the case with Tukey's Inner Fence*).

	x	y	Line of Best Fit	Difference to LoBF	Absolute Deviation from Median	Constant 0.6745 Absolute M-Score	
	6	5	4.18	0.82	0.69	0.718	
	7	5	5.23	-0.23	0.35	0.367	
	8	7	6.27	0.73	0.61	0.631	
	9	8	7.31	0.69	0.56	0.587	
	11	9	9.40	-0.40	0.52	0.543	
	12	8	10.44	-2.44	2.56	2.669	
	13	10	11.48	-1.48	1.61	1.672	
	14	13	12.52	0.48	0.35	0.367	
	15	12	13.56	-1.56	1.69	1.760	
	16	18	14.61	3.39	3.27	3.402	< Not an Outlier
Mean	11.1	9.5	9.5	0.0			
Median	11.5	8.5	9.92	0.13	0.65	< MAD	
Provisional Regression Slope			1.04				
Provisional Regression Intercept			-2.07				

Table 13: Example of the Application of Iglewicz and Hoaglin's MAD Technique

We can repeat the technique for our second example using the extra data point (Table 14), which confirms that neither suspect point is an outlier.

	x	y	Line of Best Fit	Difference to LoBF	Absolute Deviation from Median	Constant 0.6745 Absolute M-Score	
	6	5	4.68	0.32	0.07	0.050	
	7	5	5.69	-0.69	0.94	0.649	
	8	7	6.70	0.30	0.05	0.038	
	9	8	7.71	0.29	0.05	0.032	
	11	9	9.73	-0.73	0.97	0.675	
	12	8	10.74	-2.74	2.98	2.068	
	13	10	11.75	-1.75	1.99	1.381	
	14	13	12.75	0.25	0.00	0.000	
	15	12	13.76	-1.76	2.01	1.393	
	16	18	14.77	3.23	2.98	2.068	< Not an Outlier
	10	12	8.72	3.28	3.04	2.105	< Not an Outlier
Mean	11	9.7	9.7	0.0			
Median	11	9	9.73	0.25	0.97	< MAD	
Provisional Regression Slope			1.01				
Provisional Regression Intercept			-1.37				

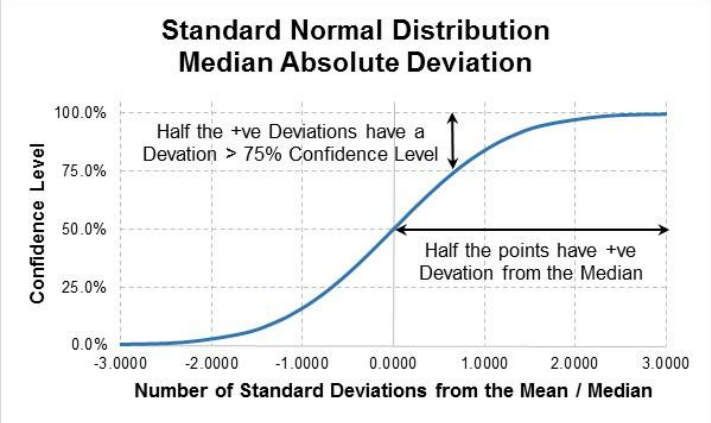
Table 14: Example of the Iglewicz and Hoaglin's MAD Technique with Additional Suspect Data Point

As Estimators we are probably wondering "*Why multiply by the constant of 0.6745?*" The value comes from the Reciprocal of the 75% Confidence Level of a Standard Normal Distribution. Iglewicz and Hoaglin observed that the expected value of the Median Absolute Deviation is approximately 67.45% of the Standard Deviation … hence the reason it is referred to as a Modified Z-Score.

For the Formula-phobes: Justifying the 75% Confidence Level as the M-Score Constant

Consider a Standard Normal Distribution. By definition it has a Standard Deviation of 1.

Consider all the points to the right of the Median or 50% Confidence Level. They all have a positive deviation from the Median. 50% of these positive points occur above the 75% Confidence Level or third Quartile, and 50% of them below it. The third quartile is therefore the Median of the upper half points.



Similarly, the first quartile is the median of the lower half points, all of which have a negative deviation from the population Median.

As the Standard Normal Distribution is symmetrical, the absolute value of the first and second quartile deviations equals the value of the third and fourth quartile deviations, so we can use the 75% Confidence Level as the Median of the Absolute Deviations from the Median … which by definition is the Median Absolute Deviation.

This argument is valid for any symmetrical distribution.

However, we can argue that the presence of this constant is largely redundant, and that we can delete it from the calculation and adjust the Critical Value for the Outlier determination to 5.2, or 5.19 (*or 5.1891 if we want to be unnecessarily precise*):

$$3.5 \div 0.6745 = 5.1891$$

However, the value of the constant should be regarded as a theoretical value only … and as such it is one that we should not take as being "sacrosanct" in practice. (*Wow, that's a bold statement!)* If we consider a random sample of 30 observations from a Normal Distribution, and calculate the sample's Median Absolute Deviation and its Standard Deviation, we can determine the ratio of the two for that sample. It is highly unlikely that the ratio will be exactly 0.6745. If we did it again, we'd probably get a different answer. We can use Monte Carlo Simulation to model the range of values we might get and the associated Confidence Levels for those values. In Figure 12 we get the simulation output based on 10,000 iterations of a Sample Size 30. In the left hand graph, we show the range and relative frequency that each ratio value occurs, and in the right hand graph, we show the confidence level that the ratio will

be a particular value or less. The graphs look to be fairly symmetrical and could be approximated by a Normal Distribution.

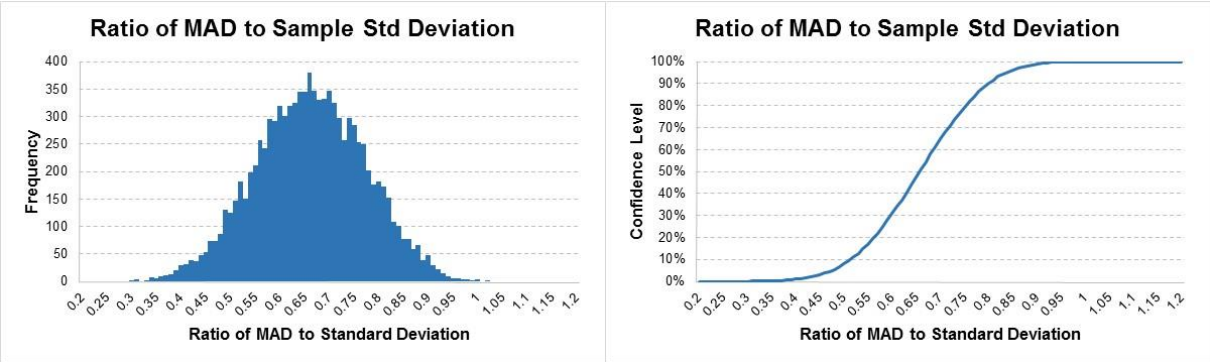


Figure 12: Range of Potential Ratio Values for Median Absolute Deviation cf. Standard Deviation (30)

Quite often in real life estimating we will not have a sample size as large as 30, so let's look at the equivalent Monte Carlo Simulation based on a smaller Sample Size of, say, 10 data points. The range of potential ratio values is shown in Figure 13. It is still broadly symmetrical but wider, and displaced slightly to the left compared with the Sample Size of 30.

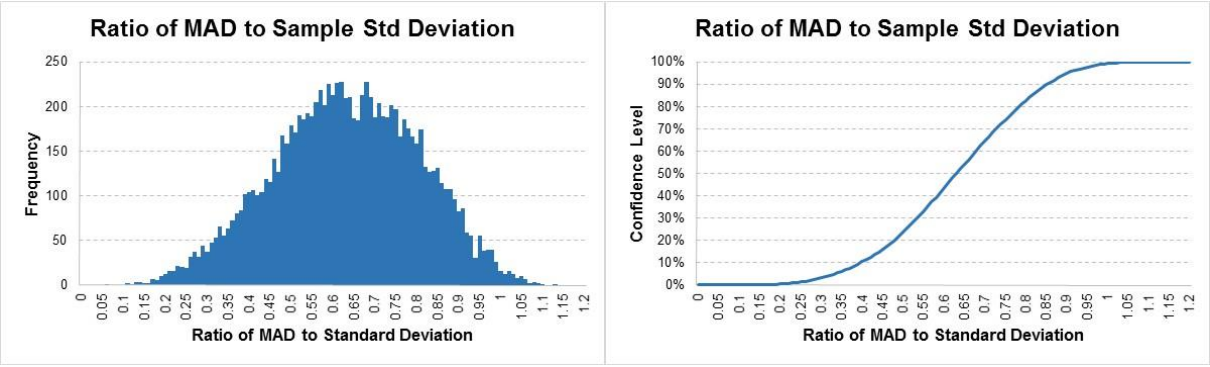


Figure 13: Range of Potential Ratio Values for Median Absolute Deviation cf. Standard Deviation (10)

These two simulations create some interesting statistics as shown in Table 15. (*Well, I found them interesting and we've already confirmed that I need to get out more!*)

	Sample Size	30	10
		Ratio	Ratio
Confidence Level	2.50%	0.43	0.28
	5%	0.47	0.33
	10%	0.51	0.39
	Median	0.660	0.62
	Mean	0.661	0.63
Confidence Level	90%	0.80	0.85
	95%	0.83	0.90
	97.50%	0.87	0.94

Table 15: Confidence Levels for M-Score Constant

Depending on the Sample Size, Iglewicz and Hoaglin M-Score Constant of 0.6745 occurs to the right of the Median for smaller sample sizes and to the left for larger samples.

By implication our confidence in the Critical Value of 3.5 is only around 50%, whereas if we had used a constant of around 0.9 our confidence in the Critical Value would increase to some 97.5%.

Using the same random samples, we can produce another pair of Monte Carlo Models that show that the Critical Value of 3.5 or less (using the 0.6745 constant) occurs with some 87.5% Confidence for a sample size of 10, and at the 91.5% Confidence Level for a sample size of 30, suggesting that we will get an outlier from random sampling around 10% of the time … which seems a little high. If we want to use a 5% Significance Level, then with the aid of Monte Carlo Simulation we can derive the following alternative Rule of Thumb for the Critical Values of the M-Score for probable outliers for varying Sample Sizes greater than 10 in Table 16; (the Rule of Thumb begins to breakdown for smaller Samples than this).

Rule of Thumb	Sample Size	M-Score Critical Value
The M-Score will vary depending on the Sample Size such that: M = 3.5 + 10 / (Sample Size)	10	4.5
	15	4.17
	20	4
	25	3.9
	30	3.83

Table 16: Suggested M-Score Critical Values for Varying Batch Sizes at the 5% Significance Level

These increased values will make it even less likely that either of our examples are outliers.

2.7 Grubbs' Test

Frank Grubbs (1969) proposed a test to detect a SINGLE outlier, again on the assumption that the data is Normally Distributed. It compares the largest Absolute Deviation from the Mean of the data points in a sample to their Standard Deviation. So, by default it only considers the Minimum or Maximum Value in a sample. (*As does any test for a single Outlier in reality*.)

Grubbs' Test assumes the Null Hypothesis that there is no outlier in the sample. The Alternative Hypothesis is that there is exactly one outlier in the data set. Grubbs' Test is sometimes referred to as the Maximum Normed Residual Test; (*yes, well, I think we'll stick to Grubbs' Test here.*)

By comparing the Deviation from the Sample Mean and dividing by the Sample Standard Deviation, it bears more than a passing resemblance to a Z-Score, but that's where the similarity ends. Here, the Critical Value of Grubbs' Test Statistic is derived from a Student t-Distribution, albeit a somewhat more complicated one than that which we contemplated for the SSS Chauvenet's Criterion in Section 2.4.

The really nice thing about Grubbs' test is that it allows the user to specific the Significance Level at which to apply the cut-off or Critical Value that determines whether a point is an outlier or not. (*It's all beginning to sound a bit more promising, isn't it*?)

For the Formula-philes: Grubbs' Test	
Consider a range of values x_1 to x_n in a sample with a Mean of $\bar{x}$ and a sample standard deviation s:	
Grubbs' test Statistic, G, is defined as:	$G = \frac{max(\lvert x_i - \bar{x} \rvert)}{s}$
If the Maximum Value is furthest from the Mean:	$G = \frac{x_{max} - \bar{x}}{s}$
If the Minimum Value is furthest from the Mean:	$G = \frac{\bar{x} - x_{min}}{s}$
At a Two-Tailed Significance Level of α, the point tested is an outlier if:	$G > \frac{(n-1)}{\sqrt{n}} \sqrt{\frac{\left(t_{\alpha/2n,n-2}\right)^2}{n-2+\left(t_{\alpha/2n,n-2}\right)^2}}$

Yes, it's OK to wince a little at this complicated expression for the Critical Value of G, especially when we realise that there is no single function in Microsoft Excel that will do this for us, although it is included in some other Commercial-off-the-Shelf software applications, but not Excel.

However, if we take it one step at a time, we can create a table of Critical Values for Grubbs' Statistic in Microsoft Excel. Table 17 derives the Critical Value in a number of small steps for a range of Sample Sizes and a Two-Tailed Significance Level of 10%. However, the good news is that we can download Tables of Critical Values from the Internet using commom Significance Levels (1%, 5%, 10% etc.)

From this table we will see that the Critical Values of Grubbs' G are much higher than the equivalent for Peirce's R (Section 2.5), and yet the calculation of Grubbs' G Test Statistic is fundamentally the same as Peirce's R Test Statistic! (*It's OK to ask "What's all that about?"*)

The basic difference is:

- Peirce's Criterion can be used to detect multiple suspect values or outliers
- Grubbs' Test looks to see if there is one, and only one, outlier so inherently the boundaries have to be much more stringent i.e. further away, and by default are more "forgiving" towards less extreme values.
- Grubbs' Test utilises the Student t-Distribution which was "discovered" until 1908 by Gosset

So, what if you try to use Grubbs' techniques iteratively? Well, assuming that it has identified an extreme value, we would be really unlucky to get a second observation that far away from the rest of the pack (*or we are looking at two distributions mixed up*), so in that sense Grubbs' technique is only expected to find a single outlier.

2 Tailed Significance Level, α			5%	Grubb's G, Critical Value		
Sample Size, n	Degrees of Freedom n-2	$\frac{(n-1)}{\sqrt{n}}$	$\frac{\alpha}{2n}$	t = T.INV(α/2n,n-2)	$t^2 / (n-2+t^2)$	Critical Value
A	*B*	*C*	*D*	*E*	*F*	*G*
				=T.INV(*D*,*B*)	$=E^2/(B+E^2)$	$=C\sqrt{F}$
4	2	1.5000	0.625%	-8.860	0.9752	1.481
5	3	1.7889	0.500%	-5.841	0.9192	1.715
6	4	2.0412	0.417%	-4.851	0.8547	1.887
7	5	2.2678	0.357%	-4.382	0.7934	2.020
8	6	2.4749	0.313%	-4.115	0.7384	2.127
9	7	2.6667	0.278%	-3.947	0.6899	2.215
10	8	2.8460	0.250%	-3.833	0.6474	2.290
11	9	3.0151	0.227%	-3.751	0.6099	2.355
12	10	3.1754	0.208%	-3.691	0.5768	2.412
13	11	3.3282	0.192%	-3.646	0.5472	2.462
14	12	3.4744	0.179%	-3.611	0.5208	2.507
15	13	3.6148	0.167%	-3.584	0.4970	2.548
20	18	4.2485	0.125%	-3.510	0.4063	2.708
25	23	4.8000	0.100%	-3.485	0.3456	2.822
30	28	5.2947	0.083%	-3.479	0.3018	2.908

Table 17: Generation of Critical Values for Grubbs' Statistic @ 5% Significance Level

			Critical Value @ 5% Level >	2.290	
x	y	Line of Best Fit	Difference to LoBF	G: |Abs Dev| (Std Dev)	
6	5	4.18	0.82	0.497	
7	5	5.23	-0.23	0.138	
8	7	6.27	0.73	0.446	
9	8	7.31	0.69	0.420	
11	9	9.40	-0.40	0.241	
12	8	10.44	-2.44	1.487	
13	10	11.48	-1.48	0.903	
14	13	12.52	0.48	0.291	
15	12	13.56	-1.56	0.954	
16	18	14.61	3.39	2.069	< Not an Outlier

Count	10			
Mean	11.1	9.5	9.50	0.00
Std Dev	3.48	3.98	3.63	1.64
Provisional Regression Slope			1.04	
Provisional Regression Intercept			-2.07	

Table 18: Example of Grubbs' Outlier Test @ 5% Significance Level

In Table 18 we revisit the example we have been using to compare with all the tests so far, we will find that Grubbs' G-Statistic for the point furthest from the Line of Best Fit i.e. the point (16,18) is less than the Critical Value at the 5% Level of Significance for a sample size of ten, which from Table 17, has the value 2.290.

If we examine the example with the additional suspect data point, we will find (Table 19) Grubbs' Test will also support the hypothesis that there are no outliers.

			Critical Value @ 5% Level >		2.355
x	y	Line of Best Fit	Difference to LoBF	G: |Abs Dev| (Std Dev)	
6	5	4.68	0.32	0.167	
7	5	5.69	-0.69	0.363	
8	7	6.70	0.30	0.158	
9	8	7.71	0.29	0.153	
11	9	9.73	-0.73	0.382	
12	8	10.74	-2.74	1.439	
13	10	11.75	-1.75	0.918	
14	13	12.75	0.25	0.129	
15	12	13.76	-1.76	0.927	
16	18	14.77	3.23	1.697	< Not an Outlier
10	12	8.72	3.28	1.726	< Not an Outlier
Count	11				
Mean	11.00	9.73	9.73	0.00	
Std Dev	3.32	3.85	3.35	1.90	
Provisional Regression Slope			1.01		
Provisional Regression Intercept			-1.37		

Table 19: Example of Grubbs' Outlier Test @ 5% Significance Level with Additional Data Point

We can say then that in both cases, using this particular test that there is insufficient evidence to support the rejection of the Null Hypothesis of there being no outliers (*yes, it's not a double but a treble negative*) in other words we can say that the values (16,18) and (10,12) are probably not outliers.

	Critical Values			
Sample Size, n	Grubbs' Test @ 5% Level	Grubbs' Test @ 10% Level	Chauvenet based on ~t(0,n-2)	Chauvenet based on ~N(0,1)
4	1.481	1.463	2.556	1.534
5	1.715	1.671	2.353	1.645
6	1.887	1.822	2.296	1.732
7	2.020	1.938	2.281	1.803
8	2.127	2.032	2.283	1.863
9	2.215	2.110	2.293	1.915
10	2.290	2.176	2.306	1.960
11	2.355	2.234	2.320	2.000
12	2.412	2.285	2.335	2.037
13	2.462	2.331	2.350	2.070
14	2.507	2.372	2.365	2.100
15	2.548	2.409	2.380	2.128
20	2.708	2.557	2.445	2.241
25	2.822	2.663	2.500	2.326
30	2.908	2.745	2.546	2.394

Table 20: Comparison of Critical Values

As the statistic we are testing in Grubbs' Test is the same one that we are testing in Chauvenet's Criterion (*the traditional and our revised SSS one using the t-Distribution*), it is really just a question of how we determine the Critical Value.

In Table 20 we compare the Critical Values for a range of Sample Sizes. This shoes that Grubbs' Test is less likely to reject a value as an Outlier than the traditional Chauvenet's Criterion, but more likely to do so that the proposed SSS Chauvenet technique using the Student t Distribution for Sample Sizes of ten or less. For very small samples we should be questioning the wisdom of rejecting any point at all.

Whilst there appears to be large differences between the Revised Chauvenet Criterion Test and the other three sets of Critical Values for very small sample sizes of 4, 5 or 6, in practice it is very difficult to create an example where all four tests would say "Reject". This is because the Line of Best Fit will compensate more for single displaced points more than in the case of large sample sizes.

2.8 Generalised Extreme Studentised Deviate (GESD)

Generalized Extreme Studentised Deviate (ESD), despite its name sounding like a radicalised student protest movement from the nineteen-sixties or seventies, is a more general purpose version of Grubbs' Test allowing multiple outliers rather than just the one.

If we have a number of potential outliers we can run the Grubbs' Test iteratively, but we must first identify how many outliers we think we have. In a practical sense it is easier to use with one-dimensional data (*we'll expand on that shortly*.) The procedure is:

1. Identify the suspect data points
2. Perform a Grubbs' Test on the full data set
3. Remove the most extreme of our suspect points, and perform a Grubbs' Test on the remaining data (irrespective of the result of the previous test)
4. Continue by removing the furthest most point until we have performed a Grubbs' Test on all our suspect data
5. If the last of these tests show that the outlier is significant, then we can reject all the previous suspect data points regardless of their individual tests. This compensates for any of the innermost outliers (*if that is not too much of an oxymoron*) distorting the mean values for the outermost outliers
6. If the innermost is not an outlier, we can look back at the last test that was significant and reject those outwards.

The difficulty we have for two or multiple dimensional data, unlike one dimensional data, is that as we set aside one suspect data point it fundamentally changes the Line of Best Fit through the remaining data, and this can then change our view of the potential number of suspect data points.

2.9 Dixon's Q-Test

Dixon's Q-Test is conceptually very simple but requires access to published Tables of Critical Values (Table 21) in order to determine the outcome. These tables are not available in Microsoft Excel but values can be sourced from the Internet, but forewarned is forearmed…

Caveat Augur

If you use this test and access Q-Tables from the internet, make sure they come from a reputable source. A simple trawl will highlight that there are conflicting values published, which is not helpful.

Critical Values of Q			
Obs, n	Two Tailed Confidence		
	99%	95%	90%
3	0.994	0.970	0.941
4	0.926	0.829	0.765
5	0.821	0.710	0.642
6	0.740	0.625	0.560
7	0.680	0.568	0.507
8	0.634	0.526	0.468
9	0.598	0.493	0.437
10	0.568	0.466	0.412
11	0.542	0.444	0.392
12	0.522	0.426	0.376
13	0.503	0.410	0.361
14	0.488	0.396	0.349

Table 21: Critical Values for Dixon's Q-Test

The premise of Dixon's Q-Test (Dixon, 1950) is that an outlier by definition is significantly distant from the rest of the data with which it is being considered. The Test compares the distance between the potential outlier and its nearest neighbour (i.e. the gap) in comparison to the overall range of the data (including the potential outlier). It is intended to be used to detect a single outlier only. Table 22 looks at this for our sample example.

It is not suitable for our second example where we have two suspect data points.

x	y	Line of Best Fit	Deviation from LoBF	Rank
6	5	4.18	0.82	9
7	5	5.23	-0.23	5
8	7	6.27	0.73	8
9	8	7.31	0.69	7
11	9	9.40	-0.40	4
12	8	10.44	-2.44	1
13	10	11.48	-1.48	3
14	13	12.52	0.48	6
15	12	13.56	-1.56	2
16	18	14.61	3.39	10

Sort Order	Deviation from LoBF	Gap	Q Statistic (Max Gap) / Range
1	-2.44	0.87	
2	-1.56		
3	-1.48		
4	-0.40		
5	-0.23		
6	0.48		
7	0.69		
8	0.73		
9	0.82		
10	3.39	2.58	

Count	10			
Mean	11.1	9.50	9.50	0.00
Std Dev	3.48	3.98	3.63	1.64
Provisional Regression Slope			1.04	
Provisional Regression Intercept			-2.07	

Obs, n	10		
Max Endpoint Gap >	2.58	>	0.442
Range	5.83	> >	
Critical Value @ 95% Level >			0.466
Critical Value @ 90% Level >			0.412

Table 22: Example of Dixon's Q-Test for Outliers

The Critical Value of Dixon's Q-Test statistic for a sample size of 10 is 0.466 at the 95% Significance Level and 0.412 at the 90% Confidence Level. Our example value is a deviation gap of 2.58 out of a range of 5.83, giving us a Q-Statistic of 0.442. This is significant at the 90% level but not at the 95% level. We have a decision to make … reject or keep!

2.10 Doing the JB Swing - Using Skewness and Excess Kurtosis to identify Outliers

In relation to Skewness and Excess Kurtosis, they have the useful property of quantifying whether our sample data is anything approaching a Normal Distribution:

- Skewness measures the degree to which the distribution "leans" to one side or the other with a Skewness Coefficient of zero being synonymous with a Symmetrical

Distribution; the Normal Distribution or Student t-Distribution assumption that most Outlier Tests have in common suggests that we would expect the sample's Skewness Coefficient to be reasonably close to zero in an ideal world (*which is usually not the case for Estimators*.)

- Excess Kurtosis measures the degree of peakedness or spikeyness of a distribution relative to its effective range in comparison with that of a Normal Distribution, which is baselined to have an Excess Kurtosis value of zero[6].
    - In the case of a Student t-Distribution, which might be considered to be an acceptable approximation to a Normal Distribution for small sample sizes, the Excess Kurtosis is 6/(ν-4) where ν is the number of Degrees of Freedom. If we are considering the scatter around a regression line then we can assume the Degrees of Freedom to be two less than the Sample Size, *n*. Therefore, the Excess Kurtosis for a Student t-Distribution representing the scatter around a Line of Best Fit would be 6/(*n*-4).
    - A Symmetrical Triangular Distribution can give us a reasonable impersonation of a Normal Distribution. The Excess Kurtosis in that case would be -0.6.

It does not seem to be an unreasonable assumption then that if our data sample has a Skewness Coefficient close to zero (i.e. slightly positive or negative) and an Excess Kurtosis between in the range of approximately -0.6 and 6/(n-4) then perhaps we can say that for the purposes of applying our Outlier Tests that "Normality reigns" (or near Normality at least). We may find, however, that if our data has an outlier, especially an extreme one, it will distort our measures of Skewness and, or Peakedness (Excess Kurtosis). Small sample sizes are more prone to statistical distortion than larger ones.

The Jarque-Bera Statistic (Jarque & Bera, 1987) combines the measures of Skewness and Excess Kurtosis, to test for Normality. Perhaps we can use it here as well to detect potential outliers, i.e. as values that disrupt our assumption of Normality. Let's explore that thought…

As an **indicator** only then, perhaps we can look at the "shape" of the sample data with and without our suspected outlier, i.e. does the removal of the suspect data point or points move the Skewness and/or the Excess Kurtosis significantly closer to zero? We can also measure what we might call the "JB Swing" (*No, that's not the name for a new Jazz Band*.)

Table 23 illustrates the JB Swing procedure and results using our benchmark sample … and nine other random samples (*just in case we think that the result in itself was an outlier … don't worry about being sceptical, it goes with the job*.) These extra samples are ostensibly Normally Distributed or Normalesque deviations around a regression line (e.g. Student t-Distributions).

1. Arrange the data in ascending order (*just so that we can easily eliminate a suspect data point*). We have highlighted the suspect point in bold font in each sample.
2. Calculate the Skewness Coefficient using Microsoft Excel's **SKEW(*range*)** function with and without the suspected outlier. Let us assume that the data is Normally Distributed, and therefore we expect a Skewness Coefficient of around zero.
3. Calculate the Excess Kurtosis using Microsoft Excel's **KURT(*range*)** function with and without the suspected outlier. If the data is Normally Distributed then we should expect an Excess Kurtosis of around zero, but for a Student t-Distribution with ten

[6] The Kurtosis of a Normal Distribution is 3. Excess Kurtosis is defined to be the Kurtosis minus 3

points, we would expect a value of 1; for nine points, this would increase to around an Excess Kurtosis of 1.2[7]

4. We can calculate the Jarque-Bera Statistic for the scatter around the regression line using the Skew and Excess Kurtosis

   JB = Sample Size x (Skew Squared + a quarter of Excess Kurtosis Squared) / six

   For a Sample Size of 10 with a t-Distribution, we would expect JB = 5/12 (or 0.4167)

   For a Sample Size of 9 with a t-Distribution, we would expect JB = 27/50 (or 0.54)

5. Typically, we would reject a sample as being non-Normal at the 5% Significance level if the JB-Statistic was greater than 6. This Test would be performed using a Chi-Squared Right Tailed Test using **CHISQ.TEST.RT(*JB*,2).** Here, we can look to improve on the Significance level of the JB Statistic by an amount that we specify. In this example we have chosen a 40% swing towards perfect Normality

Sample 1 Order	Ascending Rank Order	Sample Example Number									
		1	2	3	4	5	6	7	8	9	10
6	1	-2.44	-1.67	-1.83	**-2.27**	-1.52	-1.89	**-2.67**	**-3.14**	-3.04	-2.11
9	2	-1.56	-1.02	-1.66	-1.52	-1.49	-1.83	-1.81	-1.90	-1.50	-1.71
7	3	-1.48	-1.01	-1.57	-0.26	-1.23	-1.59	-1.57	-1.59	-1.20	-1.66
5	4	-0.40	-0.26	-1.31	-0.14	-0.35	-0.38	-0.60	-0.56	-0.96	-0.55
2	5	-0.23	-0.24	-0.28	-0.07	-0.13	-0.14	-0.20	0.22	-0.84	-0.46
8	6	0.48	-0.02	0.40	0.25	-0.11	-0.02	0.70	0.98	0.06	-0.08
4	7	0.69	-0.01	1.20	0.91	0.75	0.21	1.03	1.08	0.26	0.78
3	8	0.73	0.63	1.24	0.92	0.82	0.30	1.39	1.16	1.07	1.66
1	9	0.82	0.71	1.54	1.02	0.85	0.87	1.65	1.56	2.07	1.66
10	10	**3.39**	**2.88**	**2.27**	1.16	**2.39**	**4.48**	2.08	2.20	**4.08**	**2.47**
Skewness	With Remote Value	0.59	1.25	0.08	-1.04	0.48	1.63	-0.37	-0.63	0.73	0.18
	Without Remote Value	-0.65	-0.36	0.13	-1.01	-0.17	-0.43	-0.39	-0.57	0.05	0.21
Excess Kurtosis	With Remote Value	1.08	2.63	-1.76	0.36	-0.06	3.85	-1.25	-0.72	0.79	-1.34
	Without Remote Value	-1.04	-0.57	-2.06	0.96	-1.66	-1.43	-1.32	-0.94	0.22	-1.36
Jarque-Bera Statistic	With Remote Value	1.07	5.49	1.30	1.87	0.39	10.61	0.88	0.89	1.15	0.80
	Without Remote Value	1.15	0.35	1.80	2.09	1.19	1.17	0.99	0.91	0.02	0.85
Jarque-Bera Significance	With Remote Value	59%	6%	52%	39%	82%	0%	64%	64%	56%	67%
	Without Remote Value	56%	84%	41%	35%	55%	56%	61%	63%	99%	65%
JB Swing Indicator	Significance Swing	-2%	78%	-11%	-4%	-27%	55%	-3%	-1%	43%	-2%
	Swing > 40%		Outlier?				Outlier?			Outlier?	

Table 23: Doing the JB Swing

In our original Sample 1, the removal of the suspect point does nothing to improve the degree of Skewness, Excess Kurtosis, or the value of the Jarque-Bera Statistic for Normality, so we would conclude that the suspect data point is not an outlier.

[7] Excess Kurtosis for the Scatter around a Regression Line assuming a Student t-Distribution is 6/(n-4)

Sample 6, however, shows that the removal of the suspect point, 4.48 improves the Skewness and Excess Kurtosis towards nominal values close to zero. The JB statistic shows the most improvement, swinging from a totally unacceptable value over 10 to one just above 1.

Samples 2 and 9, show that both the Skewness and Excess Kurtosis improve towards zero is the suspect point is removed. Whilst the JB Statistic for the sample size of ten is not indicative of non-Normality in itself at the 5% Significance Level, there is still an order of magnitude improvement towards the nominal value of zero that we would expect for a Normal Distribution.

Note: If we are using this technique with Regression data scatter (as in our example) then we should re-test the scatter around the revised regression line after we have removed any outlier; the deviations will change and so might their Skewness and Excess Kurtosis

Incidentally, in case we were wondering, if we ran the JB Swing Indicator for our example with the additional suspect data point then we would find that the significance swing would not be triggered.

3. Outlier Tests: A Comparison

The proliferation of outlier tests tells us one important thing – it is not a clear-cut matter when it comes to deciding whether a point should be considered to be an outlier or not, and more importantly, whether we should be excluding it from our data. There is a definite component of subjective judgement that has to be applied, even where we try to use an objective measure because there is no single measure of correctness.

Caveat Augur

As we have already discussed, there is also a school of thought that we should never exclude an outlier.

In response to that, let's just say that perhaps it is better to "set it aside" and not include it in the initial analysis, and re-introduce it later as part of the estimate validation and sensitivity analysis stage of our process.

That way we can always raise a risk or opportunity to cover the possibility that we may get that result again.

As we have seen there are a number of tests to detect outliers, none of which are fool proof, hence the proliferation of them. (*Yes, I know, all we wanted was a simple reliable test. Soon you can go for a lie down in a darkened room.*) So which should we be using? It is worth considering whether as a matter of good practice all Estimators should only use Outlier Tests with which they feel comfortable and understand. The logic of some tests are easier to follow than others.

Using the data from Table 23, we have run all the tests for each of our ten sample examples and compared the results in Table 24 for:

- Chauvenet's Criterion (plus its SSS t-Distribution variation)
- Peirce's Criterion
- Grubbs' Test
- Tukey Fences (both Traditional and Slimline)
- Iglewicz and Hoaglin M-Score (the MAD Technique)
- Dixon's Q-Test
- JB Swing in Skewness and Excess Kurtosis

There is a lot of agreement between the various tests (*which may be comforting for the statistical cynics amongst us*) but there are also some differences *(which may be expected from those same statistical cynics amongst us.*)

- All the tests agree that Sample 6 contains an Outlier, but even then the Tukey Outer Fence does not classify it as an extreme outlier. However, had we included a significance the 99% Level for Grubbs' test, its Critical Value of 2.482 would have indicated that it was not an Extreme Outlier in the same sense as Tukey.
- In contrast, Peirce's Criterion appears to be overly strict and highlights 6 out of 10 of our random samples as having outliers … one more than our Slimline Tukey Fences that we have already indicated are better suited to large sample sizes.
- Chauvenet's Criterion is almost as intolerant as Peirce's Criterion, identifying potential outliers in half of our samples.
- Our revised SSS (Small Sample Size) Chauvenet's Criterion using a Student t-Distribution rather than a Normal Distribution, is more lenient and only identifies one sample in which we have an outlier. This is the consistent with Grubbs' Test and the Iglewicz-Hoaglin MAD Technique
- Dixon's Q-Test seems to be the outlier amongst outlier tests in the context of our samples. Whilst it agrees with other tests in some cases, it also produces some results that others do not. For example, it does not highlight Sample 9 as a potential outlier whereas the Grubbs' Test, SSS Chauvenet and JB Swing do. In contrast this highlights Sample 1 as containing an Outlier at the 90% level, but not at the 95% level, suggesting that the 90% Level is too stringent.
- Our JB-Swingometer (with a 40% Positive Confidence Swing) also flags up Samples 2, 5 and 9 in common when some of the other tests.

With larger sample sizes we may find that there is a greater (but not total) consistency between the various tests. With smaller sample sizes we are more likely to find an increase in conflicting results. We should exercise great caution when considering potential outliers with very small samples as the practice of rejecting outliers becomes even more questionable as the sample size reduces; what appears to be an extreme value in a small sample may not be the case, just the luck of the draw … it's a bit like the lottery in which all the numbers bar one are low; it's just a fluke of the random sampling used. In these cases, we are probably better erring towards the more tolerant tests.

				Sample 1 Order	Ascending Rank Order	Sample Example Number									
						1	2	3	4	5	6	7	8	9	10
				6	1	-2.44	-1.67	-1.83	-2.27	-1.52	-1.89	-2.67	-3.14	-3.04	-2.11
				9	2	-1.56	-1.02	-1.66	-1.52	-1.49	-1.83	-1.81	-1.90	-1.50	-1.71
				7	3	-1.48	-1.01	-1.57	-0.26	-1.23	-1.59	-1.57	-1.59	-1.20	-1.66
				5	4	-0.40	-0.26	-1.31	-0.14	-0.35	-0.38	-0.60	-0.56	-0.96	-0.55
				2	5	-0.23	-0.24	-0.28	-0.07	-0.13	-0.14	-0.20	0.22	-0.84	-0.46
				8	6	0.48	-0.02	0.40	0.25	-0.11	-0.02	0.70	0.98	0.06	-0.08
				4	7	0.69	-0.01	1.20	0.91	0.75	0.21	1.03	1.08	0.26	0.78
				3	8	0.73	0.63	1.24	0.92	0.82	0.30	1.39	1.16	1.07	1.66
				1	9	0.82	0.71	1.54	1.02	0.85	0.87	1.65	1.56	2.07	1.66
				10	10	3.39	2.88	2.27	1.16	2.39	4.48	2.08	2.20	4.08	2.47
			Sample Mean			0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			Sample Standard Deviation			1.640	1.255	1.528	1.141	1.242	1.844	1.626	1.736	2.021	1.595
			Z-Score for Potential Outlier			2.069	2.296	1.485	1.994	1.928	2.429	1.643	1.810	2.020	1.548
Peirce's Criterion	Critical Z-Score (10 pts)		1.878			Outlier?	Outlier?		Outlier?	Outlier?	Outlier?			Outlier?	
Grubbs' Test	Critical Z @	95%	2.29				Outlier?				Outlier?				
	Critical Z @	98%	2.410								Outlier?				
Chauvenet's Criterion	Based on Normal Dist	No of Values Expected		0	0	1	0	1	0	1	1	0	1		
		Traditional Test Result		Outlier?	Outlier?		Outlier?		Outlier?			Outlier?			
	Based on Student t	No of Values Expected		1	1	2	1	1	0	1	1	1	2		
		Revised SSS Test Result							Outlier?						
Iglewicz and Hoaglin M-Score	Max M-Score (calculation not shown)			3.402	2.533	1.045	1.943	1.777	4.640	1.558	2.372	3.146	1.304		
	Trad Critical M-Score		3.5						Outlier?						
	SSS Critical M-Score		4.5						Outlier?						
Tukey Fences	Deviation from Q3 divided by IQR >			1.39	1.86	0.38	0.21	0.87	2.68	0.30	0.43	1.60	0.37		
	Deviation from Q1 divided by IQR >			0.64	0.65	0.12	1.79	0.28	0.39	0.51	0.73	0.95	0.26		
	Traditional	Inner Fence	1.5 IQR		Outlier?		Outlier?		Outlier?			Outlier?			
		Outer Fence	3 IQR												
	Slimline	Inner Fence	1 IQR	Outlier?	Outlier?		Outlier?		Outlier?			Outlier?			
		Outer Fence	2 IQR						Outlier?						
Dixon's Q-Test	Max Gap / Range (calculation not shown)			0.442	0.477	0.178	0.219	0.395	0.567	0.181	0.232	0.282	0.176		
	Critical Q @	90%	0.412	Outlier?	Outlier?				Outlier?						
	Critical Q @	95%	0.466		Outlier?				Outlier?						
Jarque-Bari Swing Indicator	Skewness	With Remote Value		0.59	1.25	0.08	-1.04	0.48	1.63	-0.37	-0.63	0.73	0.18		
		Without Remote Value		-0.65	-0.36	0.13	-1.01	-0.17	-0.43	-0.39	-0.57	0.05	0.21		
	Excess Kurtosis	With Remote Value		1.08	2.63	-1.76	0.36	-0.06	3.85	-1.25	-0.72	0.79	-1.34		
		Without Remote Value		-1.04	-0.57	-2.06	0.96	-1.66	-1.43	-1.32	-0.94	0.22	-1.36		
	Jarque-Bera Statistic	With Remote Value		1.07	5.49	1.30	1.87	0.39	10.61	0.88	0.89	1.15	0.80		
		Without Remote Value		1.15	0.35	1.80	2.09	1.19	1.17	0.99	0.91	0.02	0.85		
	Jarque-Bera Significance	With Remote Value		59%	6%	52%	39%	82%	0%	64%	64%	56%	67%		
		Without Remote Value		56%	84%	41%	35%	55%	56%	61%	63%	99%	65%		
	JB Swing	Significance Swing		-2%	78%	-11%	-4%	-27%	55%	-3%	-1%	43%	-2%		
		Swing >	40%		Outlier?				Outlier?			Outlier?			

Note: SSS = Small Sample Size

Table 24: Comparison of Different Outlier Detection Techniques

In determining whether data should be considered to be an outlier, there is no definitive technique that works consistently in all cases. It very much depends on our "tolerance level" as indicated in Table 25.

Tests that are more tolerant of "Extreme Values"	Middle of the Road Tests	Tests that are less tolerant of "Extreme Values"
Grubbs' Test		Peirce's Criterion
Tukey Traditional Outer Fence	Tukey Traditional Inner Fence	Chauvenet's Citerion (Traditional)
SSS Adaptation of Chauvenet's Criterion	Tukey Slimline Outer Fence	Tukey Slimline Inner Fence
Iglewicz-Hoaglin M-Score (MAD Technique)	JB Swing	
	Dixon's Q-Test	

Table 25: Summary of Main Differences Between Outlier Detection Techniques

If we were to sum up Statistical Testing in one phrase it would be:

> **We can never say "Never" for certain … but with some degree of Confidence we might say "Hardly ever" … and that is a significantly more honest reflection of reality for an estimator.**

We wouldn't want our "Tails of the Unexpected" to turn out to be our personal horror story.

References

Chauvenet, W, (1863), *A Manual of Spherical and Practical Astronomy, Vol II: Theory and Use of Astronomical Instruments*, Philadelphia, J. B. Lippincott & Company, pp. 474 - 566

Dixon, WJ, (1950) "Analysis of Extreme Values", *The Annals of Mathematical Statistics*, 50, 4, 488-506

Field, A, (2005), *Discovering Statistics Using SPSS, 2ne Edition*, London, Sage, p.730 and p.739

Gosset WS (writing as "Student") (1908), "The probable error of a mean", *Biometrika*, March 6 (1) pp.1-25

Gould, BA, (1855), "On Peirce's criterion for the rejection of doubtful observations, with tables for facilitating its application," *Astronomical Journal IV*, 83, 81-87

Grubbs, F (1969), "Procedures for Detecting Outlying Observations in Samples", *Technometrics*, 11(1), February, pp. 1-21

Iglewicz, B & Hoaglin D (1993), "Volume 16: How to Detect and Handle Outliers", in *The ASQC Basic References in Quality Control: Statistical Techniques*, Mykytka, ED, (Ed), ASQC Quality Press

Jarque, CM, & Bera, AK, (1987). "A test for normality of observations and regression residuals", *International Statistical Review*, Vol. 55, No. 2, 1987, pp. 163–172

Peirce, B, (1852), "Criterion for the rejection of doubtful observations" *Astronomical Journal II*, 45,161- 163

Ross, SM (2003), "Peirce's criterion for the elimination of suspect experimental data", *Journal of Engineering Technology*, Fall

Stevenson, A (Ed), (2011), *Concise Oxford English Dictionary, 12th Edition*, Oxford, Oxford University Press

Tukey, J (1977), *Exploratory Data Analysis*, Addison-Wesley, Reading MA, pp. 39-43

