Some Outlier Tests: Part Two

Tests with fixed overall alpha levels

Donald J. Wheeler

In part one we found the baseline portion of an *XmR* chart to be the best technique for identifying potential outliers among four tests with variable overall alpha levels. In this part we will look at tests which maintain a fixed overall alpha level regardless of how many values are being examined for outliers.

Before we get lost in the mechanics of detecting outliers it is important to think about the big picture. The objective of data analysis is not to compute the "right" numbers, but rather to gain the insight needed to understand what the data reveal about the underlying process. If the outliers are simply anomalies created by the measurement process, then they should be deleted before proceeding with the computations. But if the outliers are signals of actual changes in the underlying process represented by the data, then they are worth their weight in gold because unexpected changes in the underlying process suggest that some important variables have been overlooked. Here the deletion of the outliers will not result in insight. Instead, insight can only come from identifying why the unexpected changes happened. Nevertheless, whether we delete the outliers and proceed with our statistical computations, or stop to learn why the outliers happened, the first step is still the detection of the outliers.

TESTS WITH FIXED OVERALL ALPHA LEVELS

The tests covered in part one are completely determined by the number of values in the data set being tested. No choices on the part of the user were required. The tests considered here will depend upon both the number of values in the data set and the user's choice for the overall risk of a false alarm. So how do you make this choice?

Option One: If you think that one or more outliers are likely to be present in your data, then you will want to use an overall alpha of 10%. Such tests will detect more potential outliers than those with smaller alpha levels. About one time in ten these tests will give you a single false alarm, and about nine times out of ten they will have no false alarms. These tests will generally have a positive predictive value (*PPV*) value in the neighborhood of 88%. That is, the potential outliers identified will have about an 88% chance of being real outliers. So, when you are skeptical about the quality of the data, use an alpha level of 10%.

Option Two: If you do not know whether your data may or may not have any outliers, then use the traditional overall alpha level of 5%. You may find fewer potential outliers, but the larger outliers will still show up. Only about one test in twenty will have a single false alarm, and the general *PPV* values for potential outliers will be around 92%.

Option Three: If you are virtually certain that your data contain no outliers, then you may use an overall alpha level of 1%. Here you can almost completely avoid false alarms while still checking for the presence of large outliers. Typical *PPV* values here are around 97%. Use this option only when finding outliers is a low priority.

THE *ANOX* TEST FOR OUTLIERS

The analysis of individual values (*ANOX*) was developed by this author and James Beagle in 2017 as an extension of the *XmR* chart test for outliers [1]. Here the limits provide for a fixed risk of a false alarm. The *ANOX* test for outliers uses limits of:

$$Average \pm ANOX_{\alpha} * Average\ Moving\ Range$$

Where the *ANOX* scaling factor depends upon both the number of values, n, and the user's choice of alpha-level. These values are tabled below. (More extensive tables are available in [1].)

The risk that the single most extreme value in a set of n data will fall outside these limits by chance is defined by the stated alpha-level. For this reason, any and all points that fall outside the *ANOX* limits may be reasonably interpreted as outliers.

Figure 1 gives the *ANOX* scaling factors for an alpha level of 10%. This table should be used when you suspect that outliers may be present since it will identify more potential outliers than the other tables while holding the risk of a single false alarm to only 10%.

n	$ANOX_{.10}$	n	$ANOX_{.10}$	n	$ANOX_{.10}$	n	$ANOX_{.10}$	n	$ANOX_{.10}$
8	2.058	27	2.535	46	2.694	65	2.790	84	2.857
9	2.118	28	2.546	47	2.700	66	2.794	85	2.860
10	2.167	29	2.558	48	2.706	67	2.798	86	2.863
11	2.209	30	2.569	49	2.712	68	2.801	87	2.867
12	2.246	31	2.579	50	2.718	69	2.805	88	2.870
13	2.279	32	2.589	51	2.724	70	2.809	89	2.873
14	2.308	33	2.599	52	2.729	71	2.812	90	2.876
15	2.334	34	2.609	53	2.735	72	2.816	91	2.879
16	2.358	35	2.618	54	2.740	73	2.820	92	2.882
17	2.381	36	2.626	55	2.746	74	2.823	93	2.885
18	2.401	37	2.633	56	2.750	75	2.827	94	2.888
19	2.420	38	2.640	57	2.755	76	2.830	95	2.891
20	2.437	39	2.648	58	2.760	77	2.833	96	2.893
21	2.454	40	2.655	59	2.764	78	2.837	97	2.896
22	2.469	41	2.662	60	2.769	79	2.840	98	2.899
23	2.485	42	2.668	61	2.773	80	2.843	99	2.901
24	2.499	43	2.675	62	2.777	81	2.847	100	2.904
25	2.512	44	2.682	63	2.782	82	2.850	110	2.929
26	2.524	45	2.688	64	2.786	83	2.854	120	2.951

Figure 1: 10% *ANOX* Scaling Factors

Figure 2 gives the *ANOX* scaling factors for an alpha level of 5%. This table holds the risk of a single false alarm to only 5%, and may be used when you suspect outliers to be less likely.

n	$ANOX_{.05}$	n	$ANOX_{.05}$	n	$ANOX_{.05}$	n	$ANOX_{.05}$	n	$ANOX_{.05}$
8	2.279	27	2.741	46	2.890	65	2.978	84	3.042
9	2.343	28	2.751	47	2.896	66	2.981	85	3.044
10	2.389	29	2.762	48	2.901	67	2.985	86	3.047
11	2.432	30	2.772	49	2.906	68	2.989	87	3.051
12	2.468	31	2.782	50	2.911	69	2.993	88	3.054
13	2.498	32	2.791	51	2.916	70	2.997	89	3.057
14	2.526	33	2.801	52	2.922	71	3.000	90	3.060
15	2.550	34	2.810	53	2.927	72	3.004	91	3.062
16	2.575	35	2.820	54	2.932	73	3.007	92	3.065
17	2.595	36	2.826	55	2.937	74	3.010	93	3.067
18	2.614	37	2.833	56	2.941	75	3.013	94	3.070
19	2.632	38	2.840	57	2.946	76	3.017	95	3.072
20	2.648	39	2.847	58	2.950	77	3.020	96	3.075
21	2.665	40	2.854	59	2.954	78	3.023	97	3.077
22	2.681	41	2.860	60	2.959	79	3.027	98	3.080
23	2.695	42	2.866	61	2.963	80	3.030	99	3.082
24	2.708	43	2.873	62	2.966	81	3.033	100	3.085
25	2.720	44	2.879	63	2.970	82	3.036	110	3.105
26	2.730	45	2.885	64	2.974	83	3.039	120	3.126

Figure 2: 5% *ANOX* Scaming Factors

Figure 3 gives the *ANOX* scaling factors for an alpha level of 1%. This table holds the risk of a single false alarm to only 1%. It is biased against finding any outliers in favor of including all the data as good data. It should be used only when you are highly confident that there are no outliers in your data.

n	$ANOX_{.01}$	n	$ANOX_{.01}$	n	$ANOX_{.01}$	n	$ANOX_{.01}$	n	$ANOX_{.01}$
8	2.827	27	3.176	46	3.303	65	3.375	84	3.425
9	2.863	28	3.186	47	3.307	66	3.378	85	3.427
10	2.897	29	3.195	48	3.312	67	3.381	86	3.429
11	2.928	30	3.204	49	3.316	68	3.384	87	3.431
12	2.958	31	3.212	50	3.320	69	3.388	88	3.434
13	2.985	32	3.220	51	3.325	70	3.391	89	3.436
14	3.008	33	3.229	52	3.329	71	3.393	90	3.439
15	3.029	34	3.237	53	3.333	72	3.396	91	3.440
16	3.048	35	3.245	54	3.337	73	3.398	92	3.442
17	3.064	36	3.251	55	3.341	74	3.400	93	3.444
18	3.077	37	3.257	56	3.344	75	3.403	94	3.445
19	3.090	38	3.262	57	3.348	76	3.406	95	3.447
20	3.103	39	3.268	58	3.351	77	3.408	96	3.449
21	3.115	40	3.274	59	3.355	78	3.411	97	3.451
22	3.127	41	3.279	60	3.358	79	3.414	98	3.453
23	3.138	42	3.284	61	3.362	80	3.417	99	3.455
24	3.148	43	3.288	62	3.365	81	3.419	100	3.457
25	3.158	44	3.293	63	3.368	82	3.421	110	3.473
26	3.167	45	3.298	64	3.372	83	3.423	120	3.486

Figure 3: 1% *ANOX* Scaling Factors

The *PPV* curves for the *ANOX* tests for outliers are shown in Figure 4 along with the *PPV* curve for the *XmR* chart from part one.

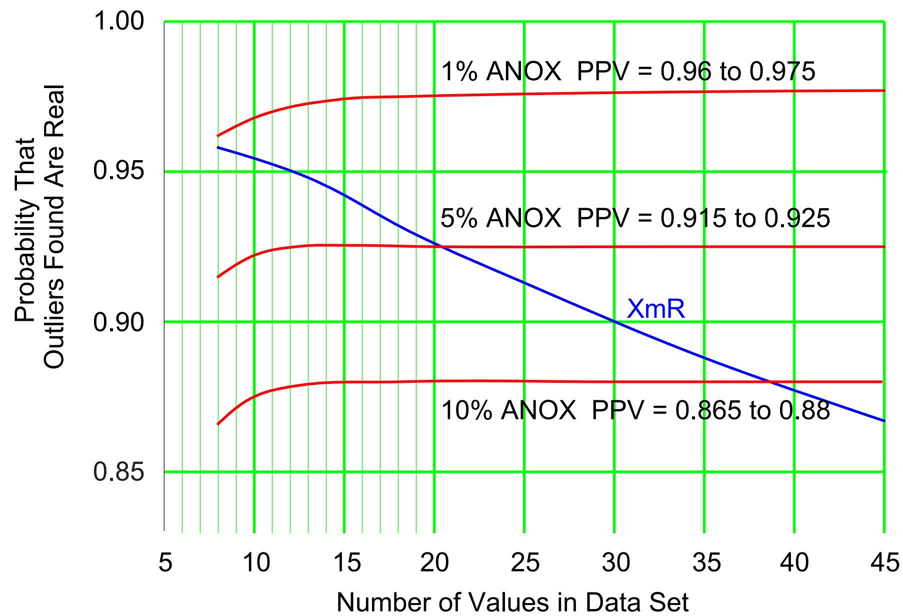

Figure 4: *PPV* Curves for ANOX Test for Outliers

So, whenever you use a 10% *ANOX* to test for outliers, the points you identify as potential outliers have an 88% chance of being real outliers that do not belong with the rest of your data.

When you use a 5% *ANOX* test, you may not find as many potential outliers as with a 10% *ANOX*, but the potential outliers you do find will have a 92% chance of being real outliers.

When you use a 1% *ANOX*, you will find fewer potential outliers than with a 5% *ANOX*, but those you do identify will have a 97% chance of being real outliers.

In contrast to these *ANOX* tests with their fixed alpha levels, the *XmR* chart will behave something like a 1% *ANOX* test when $n < 12$, similar to a 5% *ANOX* test when $13 < n < 30$, and something ike a 10% *ANOX* test when $n \geq 30$. Thus, the 5% *ANOX* and 10% *ANOX* tests will be more sensitive to outliers than the *XmR* test when n is small.

Since, like the *XmR* chart, the *ANOX* test uses the average moving range it also cannot be used on data that have been arranged in ascending or descending order. While the time order for the data is the preferred ordering, any arbitrary ordering that is independent of the values for the data may be used with the *ANOX* test.

One other limitation exists for the *ANOX* test, and this is the limitation imposed by chunky data. As long as the average moving range is greater than 0.9 measurement increments the *ANOX* test will work as advertised. When the average moving range drops below 0.9 measurement increments the chunkiness of the data will begin to cause the limits to shrink due to round-off effects. This shrinkage will increase the number of false alarms.

So *ANOX* combines the simplicity of the *XmR* chart with the advantage of being able to choose in advance a fixed overall alpha level for your test.

GRUBBS' TEST FOR OUTLIERS

In 1950 Frank E. Grubbs published a test for identifying outliers [2]. His test is equivalent to the following: Given n data ($n \geq 4$) compute the average and the global standard deviation statistic. Let $G(n,\alpha)$ be defined by:

$$G(n,\alpha) = \sqrt{\frac{(n-1)^2\ t^2}{n\,(n-2+t^2)}}$$

where the symbol t denotes the critical value from a Student's t-distribution with $[n–2]$ degrees of freedom which cuts off an upper tail area equal to $[\alpha/2n]$.

Grubbs' test uses the interval:

$$Average \ \pm\ G(n,\alpha)\ *\ Standard\ Deviation\ Statistic$$

Any and all values outside this interval are designated as outliers. Figure 5 gives selected values of $G(n,\alpha)$.

$G(n,\alpha)$	$n = 5$	$n = 10$	$n = 15$	$n = 20$	$n = 25$	$n = 30$
$\alpha = 0.10$	1.671	2.176	2.409	2.557	2.663	2.745
$\alpha = 0.05$	1.715	2.290	2.548	2.708	2.822	2.908
$\alpha = 0.01$	1.764	2.482	2.806	3.001	3.135	3.236

Figure 5: Selected $G(n,\alpha)$ Values for Grubbs' Test for Outliers

Figure 6 lists the *PPV* values for the Grubbs cut-offs given in Figure 5. Figure 7 shows the *PPV* curves for Grubbs' test.

PPV	$n = 5$	$n = 10$	$n = 15$	$n = 20$	$n = 25$	$n = 30$
$\alpha = 0.10$	0.816	0.870	0.876	0.877	0.879	0.881
$\alpha = 0.05$	0.845	0.913	0.923	0.925	0.926	0.927
$\alpha = 0.01$	0.865	0.956	0.972	0.973	0.974	0.975

Figure 6: *PPV* Values for Grubbs' Test for Outliers

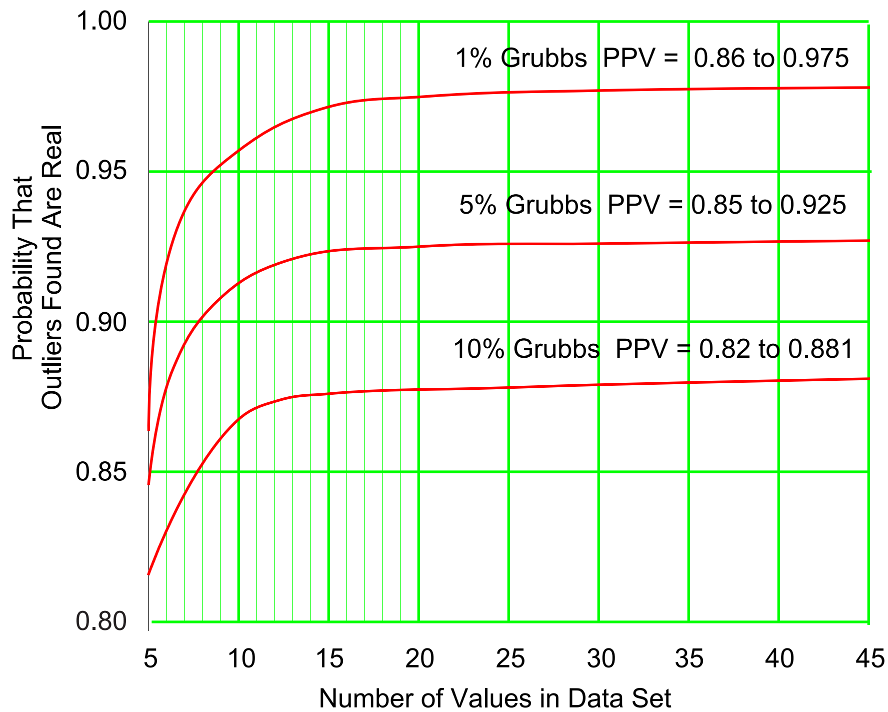

Figure 7: *PPV* Curves for Grubbs' Test for Outliers

These *PPV* curves are very close to those found for the *ANOX* test, which suggests fairly equivalent performance. Thus, both the *ANOX* test and Grubbs' test will allow you to test for outliers using a fixed overall alpha level that will result in a reasonable degree of belief that the outliers identified are indeed real.

LIMITATIONS

Most tables of the Grubbs' critical values include values for $n = 3$. These critical values are theoretical values derived under the assumption that the measurements are observations drawn from a continuum. Yet in practice, all data display some level of chunkiness, and this chunkiness places some limitations on Grubbs' test.

We start by recalling from Shiffler [3] that there is a maximum standardized value for a set of n data of:

$$\frac{|\,Observation - Average\,|}{Standard\ Deviation} \leq \frac{n-1}{\sqrt{n}}$$

This means that it will be *impossible* for any observation to ever fall outside the interval:

$$Average \pm \frac{n-1}{\sqrt{n}}\ Standard\ Deviations$$

For $n = 3$ this upper bound for a standardized value is 1.1547. Coincidentally, the 1% critical value for Grubbs' test for $n = 3$ is $G(3,\ 0.01) = 1.1547$. Thus, with three observations, it is impossible to ever get a value that will exceed the 1% Grubbs' cut-off. Hence, *for n = 3 Grubbs' test with alpha = 0.01 will never detect an outlier!*

For alpha = 0.05 and n = 3 the Grubbs' critical value is *G(3,0.05)* = 1.1543. In order to get one standardized value in between 1.1543 and 1.1547, a difference of 0.0004, the standard deviation will have to allow increments of 0.0002 in the standardized valules. When we invert this number we discover that the standard deviation will have to exceed 5000 measurement increments! Unless the standard deviation is greater than 5000 measurement increments it will be impossible to compute a standardized value in between the critical value of 1.1543 and the upper bound of 1.1547. And if we cannot compute a value that falls in this interval, the test will never detect an outlier.

For alpha = 0.10 and n = 3 the Grubbs' critical value is *G(3, 0.10)* = 1.1531. To allow a standardized value to fall half-way in between this critical value and the upper bound of 1.1547, the standard deviation will have to allow increments of 0.0008 in the standardized values. Inverting this we find that the standard deviation will have to exceed 1250 measurement increments before we can begin to use Grubbs' test for n = 3 and alpha = 0.10. Since it is rare to find data recorded using measurement increments that are 1250 times smaller than the standard deviation statistic, it is extremely unlikely that Grubbs' test for n = 3 and alpha = 0.10 will ever detect an outlier.

Since a test that only allows one outcome to occur is not a true test, you need to avoid using Grubbs' test with n = 3. Continuing in the same way for other values of n we end up with Figure 8 which lists the minimum number of measurement increments needed within the standard deviation statistic in order to use Grubbs' test for outliers.

	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 7$	$n = 8$	$n = 9$	$n = 10$
Grubbs' 0.01	—	533	79	29	16	10	7	5
Grubbs' 0.05	5054	107	27	13	8	6	4	4
Grubbs' 0.10	1264	53	17	9	6	5	4	3

Figure 8: Number of Measurement Increments in Std. Dev. Needed to Use Grubbs' Test

Thus, in general, in order to use Grubbs' test for n = 5, 6, or 7, you will need measurement increments that are at least one to two orders of magnitude smaller than the standard deviation statistic.

For example, if a set of n = 5 data had a standard deviation statistic of 4.56 units, and if the data were all recorded to the nearest 0.05 units, then the measurement increment would be 0.05 units and the standard deviation would be:

$$s = \frac{4.56 \text{ units}}{0.05 \text{ units per increment}} = 91.2 \text{ increments}$$

and we could use Grubbs' test at the 0.01 level with these data.

DIXON'S TEST FOR GAPS

Two other tests that also strike a balance between finding outliers and preserving good data are Dixon's test and the *W*-ratio test. These tests were discussed in earlier articles [4] [5] [6]. They differ from the tests above in that they are designed to find a gap in ordered data sets rather than looking for any and all outliers. Figure 9 shows the *PPV* curves for Dixon's test. Since the *W*-ratio test has power curves that are very similar to those of Dixon's test we expect the *PPV* curves for the *W*-ratio to be very close to those shown in Figure 9.

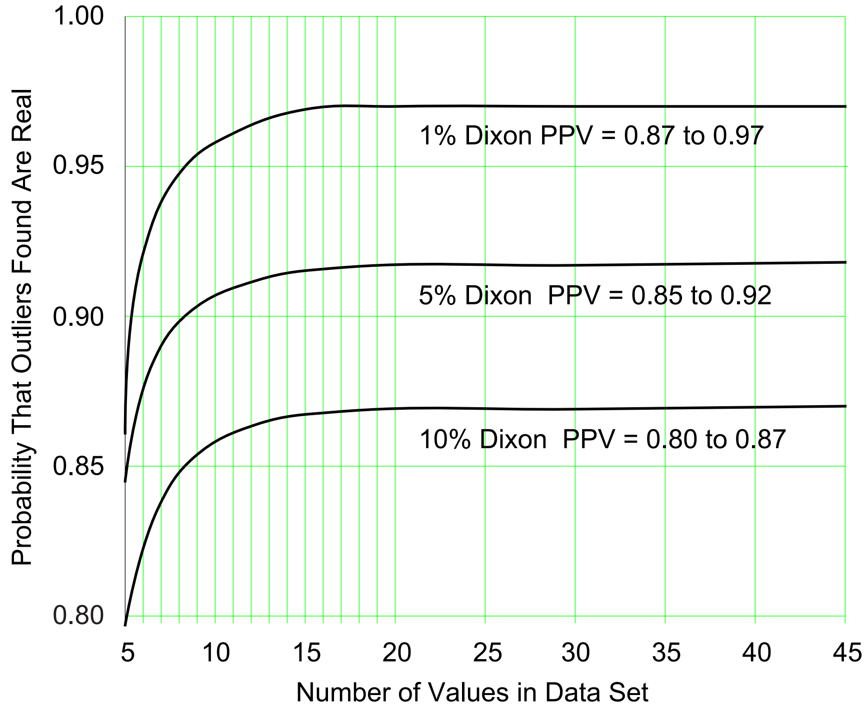

Figure 9: *PPV* Curves for Dixon'sTest for Outliers

Just as the measurement increment can create round-off issues with Grubbs' test, the same thing happens with Dixon's test and the *W*-ratio test. Both of these tests use the global range statistic for the set of *n* data.

$$\textit{Global Range} = \textit{Maximum of the n data} - \textit{Minimum of the n data}$$

In order for Dixon's test and the *W*-ratio test to have an overall alpha level that is close to the specified alpha level the global range will have to be greater than the number of measurement increments specified in Figure 10 [4].

Theoretical alpha-level	***k = 3***	***k = 4***	***k = 5***	***k = 6***	***k = 7***	***k = 8***	***k = 9***	***k = 10***
0.01	500	56	46	40	48	45	46	45
0.05	77	30	32	33	31	39	29	33
0.10	56	31	32	33	23	35	33	35

Figure 10: Minimum Number of Measurement Increments in Global Range for Robustness

To use Dixon's test or the *W*-ratio test at the alpha = 0.01 level you need a range statistic that exceeds roughly 50 measurement increments for $n \geq 4$. For larger alpha levels you need a range statistic that exceeds roughly 30 measurement increments for $n \geq 4$.

COMPARING THE OUTLIER TESTS

Figure 11 shows the *PPV* curves for all of the tests considered in both parts of this survey.

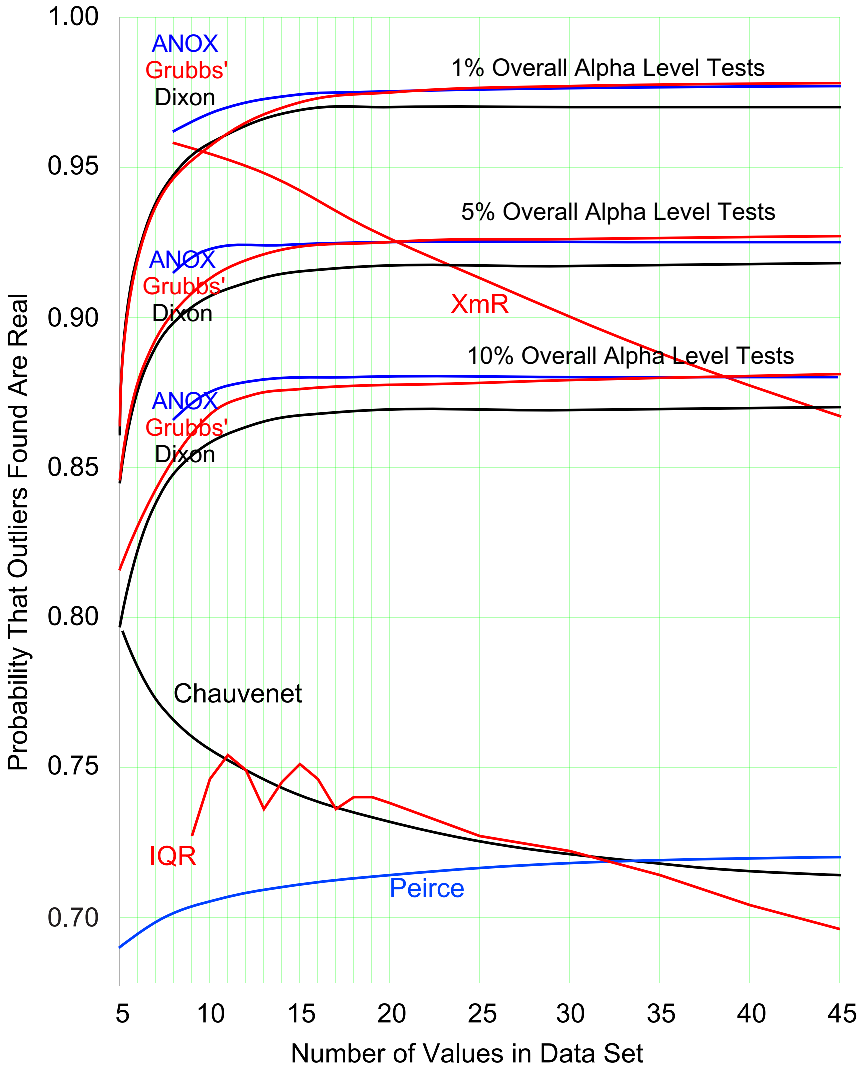

Figure 11: *PPV* Curves

The convergence between the curves for the *ANOX* and Grubbs' tests tells us that they are going to perform an equivalent job in practice. Dixon's curves (and those for the *W*-ratio which are not shown) are slightly below the other curves because they look for gaps instead of extreme values. However, the similarity of all these curves implies that these fixed overall alpha level tests all operate close to the absolute limit of what can be extracted from the data.

This means that other tests for detecting outliers simply cannot do any better job than these three approaches. While other tests may be dressed in different formulas, they will ultimately be either *equivalent* to ANOX, Grubbs', and Dixon, or they will be *inferior* to ANOX, Grubbs, and Dixon.

For example the modified Thompson's tau test is simply Grubbs' test performed with an overall alpha level that is *n times larger* than Grubbs' overall alpha level. As may be seen in

Figure 11, the effect of increasing the overall alpha level is to lower the *PPV* curve. Thus, the modified Thompson's tau test is going to *always be inferior* to Grubbs' test. It may find more potential outliers, but it will also have an excessive number of false alarms, undermining your faith in the reality of the potential outliers while removing good data. Such is the quid pro quo required of all such tests.

SUMMARY

Trying to identify all of the outliers is an unrealistic goal. Likewise, trying to avoid all false alarms is also an unrealistic goal. The trick is to strike a balance between these two goals: Identify those outliers that have a large effect while avoiding false alarms as much as possible.

Procedures that try to capture all of the outliers will go overboard and include good data in the dragnet along with the outliers. As shown in part one this is what happens with Peirce's test, Chauvenet's test, and the *IQR* test.

Procedures that keep the overall alpha level reasonably small will still find the major outliers without an undue increase in risk of false alarms. As shown, the *XmR* test, *ANOX*, Grubbs' test, and the Dixon and *W*-ratio tests all fall into this category.

Statistical inference is built on the assumption of homogeneous data. Outliers create a lack of homogeneity. In the rush to use their computerized computations people are going to continue to be interested in deleting the outliers.

The problem with deleting the outliers to obtain a homogeneous data set is that the resulting data set will no longer belong to *this* world. If the analysis of a *purified* data set ignores the assignable causes that lurk behind most outliers the results will not apply to the underlying process that produces the data. The real question about outliers is not how to get them out of the data, but why do they exist in the first place.

In this author's 50 years of experience in helping people analyze data, the more profound question has always been "Why are there outliers?" rather than "What do we find when we delete the outliers?"

There are many more tests for outliers, some with sopisticated mathematical theory behind them. Undoubtedly more tests will be created in the future. Many of these will follow Peirce and Chauvenet down the rabbit hole of trying to find *all* of the outliers so as to obtain a *purified* data set for their analysis. However, information theory places an upper bound on how much can be extracted from a given data set, and adding more tests will not change this upper bound. *ANOX*, Grubbs', Dixon, and the *W*-ratio all approach this upper bound. Other tests can do no better.

So, rather than arguing over which outlier test to use, it is better to find fewer outliers and to discover what happened to create those outliers than it is to find more outliers and delete them in order to analyze data that no longer describe reality.

REFERENCES

1. Donald J. Wheeler and James Beagle III, "ANOX: The Analysis of Individual Values," ***Quality Digest,*** September 4, 2017.

2. Frank E. Grubbs, "Sample Criteria for Testing Outlying Observations," ***Annals of Mathematical Statistics, v.21(1),*** pp. 27-58, 1950.

3. R.E. Shiffler, "Maximum z-Scores and Outliers," ***American Statistician, v. 42,*** pp. 79-80, 1988.

4. Donald J. Wheeler, "A Problem with Outlier Tests," ***Quality Digest Daily,*** September 1, 2014.

5. Donald J. Wheeler, "Analysis Using Few Data," ***Quality Digest Daily,*** June 6, 2012.

6. Donald J. Wheeler, "Analysis Using Few Data: Part Two" ***Quality Digest Daily***, Nov. 5, 2012.