Outlier Detection

Outlier detection is both easy and difficult.

- It is easy since there are several relatively straightforward tests for the presence of outliers.
- It is difficult since there are no firm rules as to when outlier removal is appropriate.

Outliers may be due to:

- Chance.
- Measurement error.
- Experimental error.

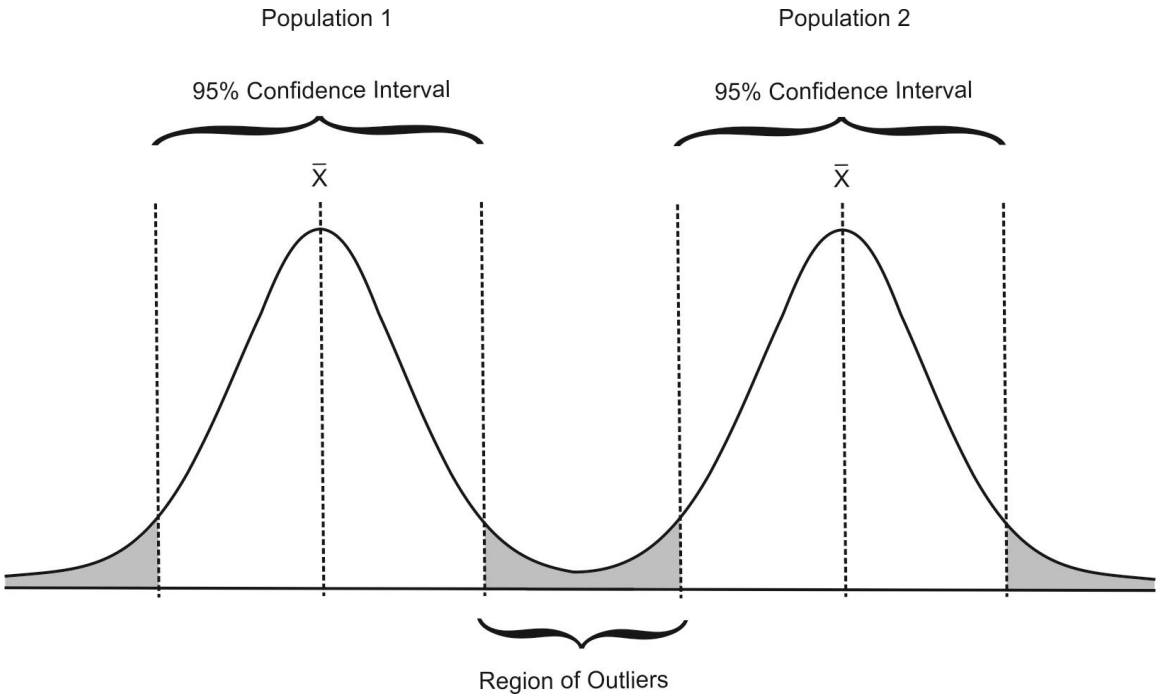
Outliers may or may not be a problem, depending on many factors:

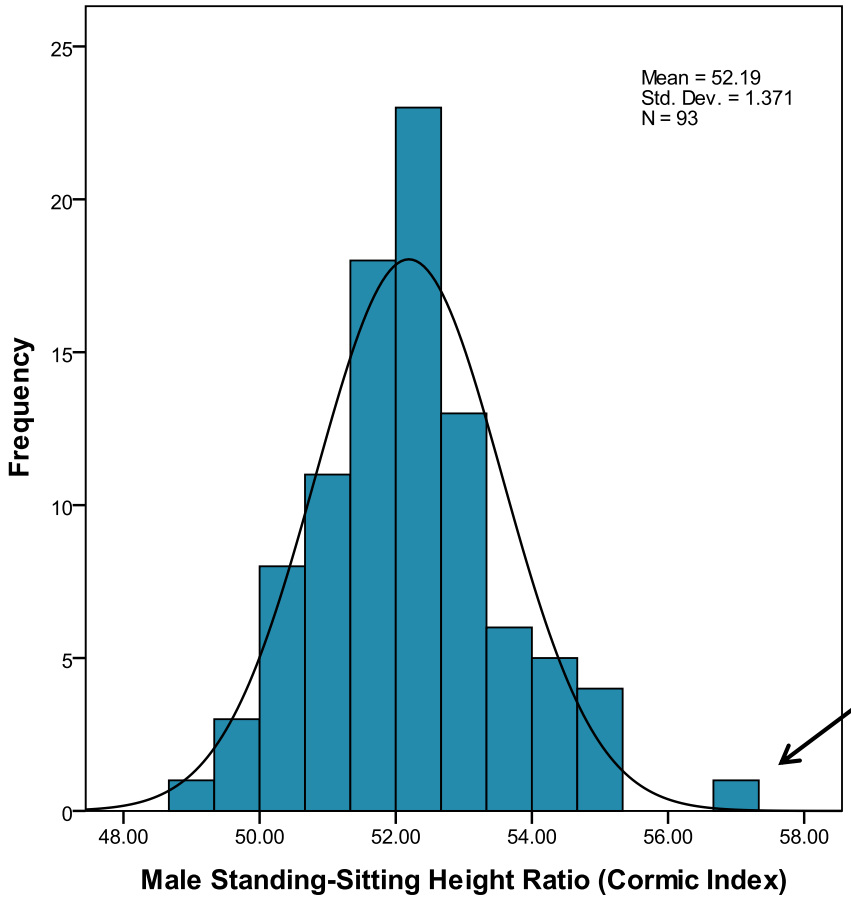
- Some statistical tests are robust and can accommodate outliers, others may be severely influenced by outliers.
  - Parametric test can unduly influenced.
  - Non-parametric tests rarely are.

- Some data types will naturally contain extreme values.
  - Radiation levels often have extreme values (spikes).

- The presence of outliers may, in fact, be of interest.
  - Again, radiation spikes.

The outlier(s) may fall in a region of population overlap. This type of outlier must be removed from the data set.





Is this observation (57.00) an outlier?

In some cases a single outlier may influence normality, however, in this case the data are normal even with this observation.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Male Standing-Sitting Height Ratio (Cormic Index)	.065	93	.200*	.986	93	.304

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Should this observation be examined further at this point?

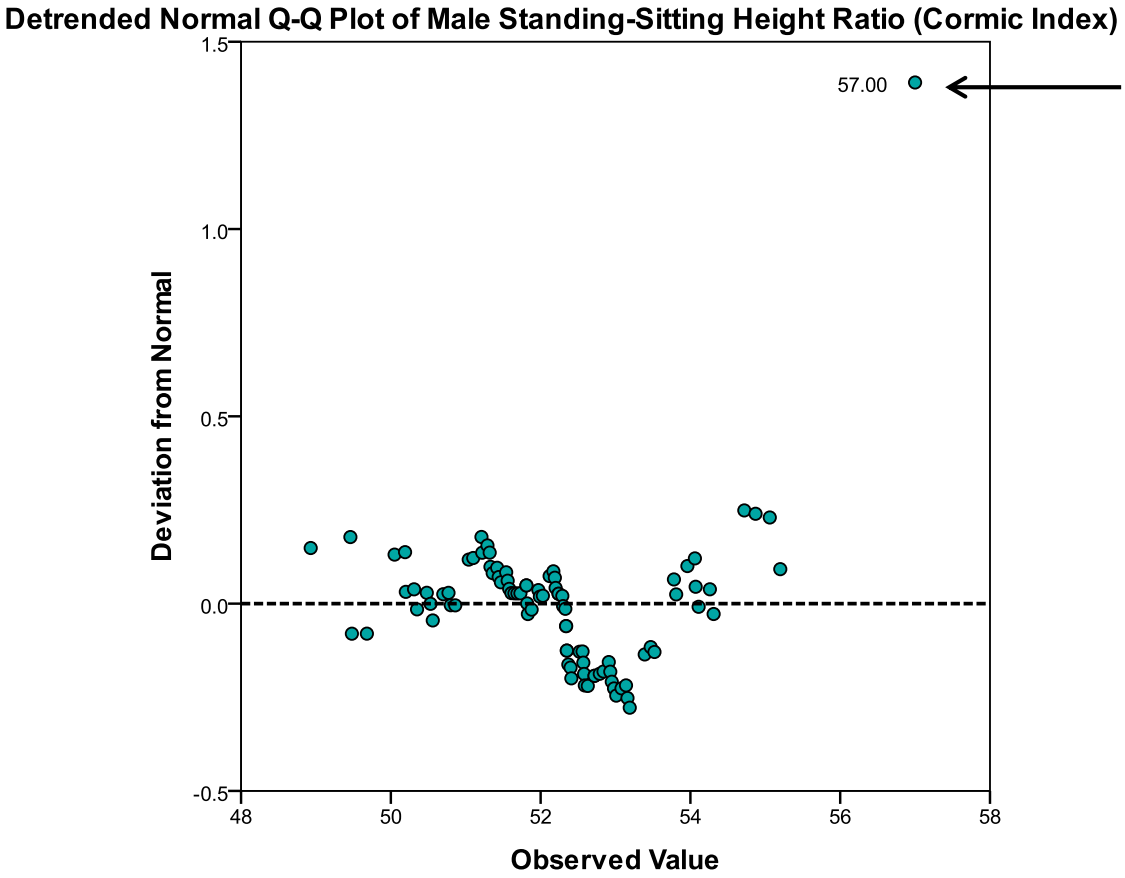
Male Standing-Sitting Height Ratio (Cormic Index)
Stem-and-Leaf Plot

```
Frequency    Stem &  Leaf

    1.00 Extremes    (=<48.9)
    3.00       49 .  446
   12.00       50 .  012334557788
   25.00       51 .  0122233344455566678888899
   31.00       52 .  0111222233333334455555677789999
   11.00       53 .  00111345789
    7.00       54 .  0012378
    2.00       55 .  02
    1.00 Extremes    (>=57.0)

Stem width:      1.00
Each leaf:       1 case(s)
```

The observation 57.0 is considered to be an extreme value in the stem and leaf plot.

When examining potential outliers, the detrended normal Q-Q plot is useful.

- Observations are transformed to z-scores and plotted as standard deviations from the mean.



This observation is nearly 1.5 standard deviations from the mean.

The best method of determining if an observation is an outlier is to use an outlier test.

- The test gives the probability that an observation is from a different population.
- It is defensible.
- It DOES NOT tell you whether or not to remove the extreme observation(s)…

Grubbs Outlier Test

$$G_{\max} = \frac{x_{\max} - \bar{x}}{s} \quad or \quad G_{\min} = \frac{\bar{x} - x_{\min}}{s}$$

where G_{max} is used if the observation is greater than the mean and G_{min} is used if it is less than the mean, and where x_{max} or x_{min} is the extreme observation value.

Ho: The observation is not different than the sample population.
Ha: The observation is different than the sample population.

$$n = 93$$

$$\bar{x} = 52.2$$

$$s = 1.38$$

$$G_{\max} = \frac{57.0 - 52.2}{1.38} = 3.49$$

From the G table at n=93 and α=0.05 the critical value is 3.18.

Since 3.49 > 3.18, reject H_o.

The observation is from a different population ($G_{3.49}$, $p < 0.025$).

Critical Values of Grubb's Outlier (G) Test

Taken from Grubb 1969, Table 1

N	α=0.05	α=0.025	α=0.01
3	1.15	1.15	1.15
4	1.46	1.48	1.49
5	1.67	1.71	1.75
6	1.82	1.89	1.94
7	1.94	2.02	2.10
8	2.03	2.13	2.22
9	2.11	2.21	2.32
10	2.18	2.29	2.41
11	2.23	2.36	2.48
12	2.29	2.41	2.55
13	2.33	2.46	2.61
14	2.37	2.51	2.66
15	2.41	2.55	2.71
16	2.44	2.59	2.75
17	2.47	2.62	2.79
18	2.50	2.65	2.82
19	2.53	2.68	2.85
20	2.56	2.71	2.88
21	2.58	2.73	2.91
22	2.60	2.76	2.94
23	2.62	2.78	2.96
24	2.64	2.80	2.99
25	2.66	2.82	3.01
30	2.75	2.91	
35	2.82	2.98	
40	2.87	3.04	
45	2.92	3.09	
50	2.96	3.13	
60	3.03	3.20	
70	3.09	3.26	
80	3.14	3.31	
90	3.18	3.35	
100	3.21	3.38	

Calculated value falls about here.

Dixon Outlier (Q) Test

$$Q = \frac{x_n - x_{n-1}}{x_n - x_1}$$

Where x_n is the suspected outlier, x_{n-1} is the next ranked observation, and x_1 is the last ranked observation.

Note that the data have to be ranked, with the suspected outlier as the first observation.

In SPSS Analyze > Descriptive Statistics > Explore, then choose the *Statistics* button and *Outliers*.

Extreme Values

			Case Number	Value
Male Standing-Sitting Height Ratio (Cormic Index)	Highest	1	1	57.00
		2	2	55.20
		3	3	55.06
		4	4	54.87
		5	5	54.72
	Lowest	1	93	48.93
		2	92	49.46
		3	91	49.48
		4	90	49.68
		5	89	50.05

This gives the upper and lower extremes AND the next several observations, very useful when using the Dixon test.

Extreme Values

			Case Number	Value	
Male Standing-Sitting Height Ratio (Cormic Index)	Highest	1	1	57.00	←
		2	2	55.20	←
		3	3	55.06	
		4	4	54.87	
		5	5	54.72	
	Lowest	1	93	48.93	←
		2	92	49.46	
		3	91	49.48	
		4	90	49.68	
		5	89	50.05	

H_o: The observation is not different than the sample population.

H_a: The observation is different than the sample population.

$$n = 93$$

$$Q = \frac{57.00 - 55.20}{57.00 - 48.93} = \frac{1.8}{8.07} = 0.223 \quad Q_{Critical} = 0.1881 \quad 0.05 > p > 0.02$$

The observation is from a different population ($Q_{0.223}$, $0.05 > p > 0.02$).

n	CL	70%	80%	90%	**95%**	98%	**99%**	99.5%
	SL	30%	20%	10%	**5%**	2%	**1%**	0.5%
	α	0.30	0.20	0.10	**0.05**	0.02	**0.01**	0.005
51		0.1079	0.1374	0.1819	0.2206	0.2651	0.2941	0.3204
52		0.1071	0.1365	0.1808	0.2191	0.2632	0.2927	0.3191
53		0.1067	0.1357	0.1797	0.2182	0.2620	0.2920	0.3177
54		0.1060	0.1349	0.1788	0.2169	0.2606	0.2899	0.3163
55		0.1052	0.1340	0.1777	0.2160	0.2595	0.2880	0.3140
56		0.1047	0.1334	0.1768	0.2145	0.2582	0.2873	0.3136
57		0.1041	0.1326	0.1759	0.2135	0.2570	0.2859	0.3118
58		0.1036	0.1320	0.1752	0.2126	0.2555	0.2845	0.3098
59		0.1030	0.1312	0.1741	0.2116	0.2545	0.2828	0.3089
60		0.1024	0.1304	0.1733	0.2106	0.2531	0.2816	0.3075
61		0.1019	0.1299	0.1726	0.2095	0.2522	0.2812	0.3071
62		0.1014	0.1294	0.1717	0.2085	0.2510	0.2792	0.3061
63		0.1009	0.1286	0.1707	0.2075	0.2500	0.2784	0.3041
64		0.1004	0.1281	0.1703	0.2070	0.2493	0.2775	0.3031
65		0.1000	0.1275	0.1694	0.2057	0.2480	0.2766	0.3025
66		0.0997	0.1272	0.1689	0.2053	0.2472	0.2754	0.3006
67		0.0991	0.1264	0.1679	0.2045	0.2466	0.2742	0.2996
68		0.0987	0.1260	0.1674	0.2037	0.2457	0.2735	0.2990
69		0.0982	0.1254	0.1667	0.2030	0.2445	0.2724	0.2983
70		0.0979	0.1249	0.1660	0.2020	0.2436	0.2714	0.2968
71		0.0974	0.1243	0.1652	0.2013	0.2429	0.2709	0.2959
72		0.0970	0.1238	0.1648	0.2005	0.2420	0.2696	0.2946
73		0.0967	0.1234	0.1641	0.1996	0.2409	0.2682	0.2934
74		0.0961	0.1228	0.1635	0.1990	0.2402	0.2677	0.2932
75		0.0960	0.1225	0.1631	0.1984	0.2398	0.2667	0.2922
76		0.0955	0.1221	0.1626	0.1980	0.2387	0.2662	0.2912
77		0.0952	0.1217	0.1620	0.1973	0.2382	0.2656	0.2905
78		0.0948	0.1212	0.1613	0.1964	0.2372	0.2646	0.2897
79		0.0943	0.1205	0.1605	0.1955	0.2365	0.2637	0.2885
80		0.0939	0.1201	0.1601	0.1950	0.2360	0.2633	0.2876
81		0.0937	0.1198	0.1596	0.1943	0.2349	0.2621	0.2870
82		0.0935	0.1195	0.1594	0.1940	0.2345	0.2614	0.2859
83		0.0930	0.1189	0.1586	0.1934	0.2337	0.2608	0.2852
84		0.0928	0.1187	0.1583	0.1927	0.2330	0.2599	0.2844
85		0.0925	0.1182	0.1576	0.1922	0.2322	0.2588	0.2836
86		0.0921	0.1178	0.1573	0.1918	0.2319	0.2584	0.2832
87		0.0918	0.1174	0.1567	0.1909	0.2309	0.2573	0.2818
88		0.0915	0.1171	0.1563	0.1906	0.2304	0.2568	0.2811
89		0.0913	0.1167	0.1557	0.1899	0.2298	0.2566	0.2808
90		0.0910	0.1165	0.1554	0.1896	0.2294	0.2558	0.2798
91		0.0906	0.1160	0.1547	0.1887	0.2285	0.2548	0.2790
92		0.0903	0.1156	0.1544	0.1885	0.2279	0.2543	0.2788
93		0.0902	0.1154	0.1540	0.1881	0.2272	0.2539	0.2784
94		0.0899	0.1151	0.1537	0.1876	0.2272	0.2535	0.2775
95		0.0896	0.1147	0.1532	0.1869	0.2259	0.2524	0.2766
96		0.0894	0.1144	0.1528	0.1865	0.2257	0.2521	0.2764
97		0.0892	0.1141	0.1524	0.1860	0.2251	0.2512	0.2755
98		0.0890	0.1138	0.1521	0.1856	0.2247	0.2513	0.2751
99		0.0887	0.1134	0.1516	0.1851	0.2240	0.2499	0.2738
100		0.0885	0.1131	0.1512	0.1846	0.2234	0.2498	0.2737

CL: Confidence level (%); SL: Significance level (%); α: Significance level. Headers for commonly used CL or SL or α are given in bold face (*e.g.*, for RM applications). The mean values of the standard error of the mean ($\overline{x}_{se}$) for these critical values ($\overline{x}$) are (respective % errors are also reported in parentheses): ~0.00011 (for $\alpha = 0.30$, 0.09%); ~0.00011 (for $\alpha = 0.20$, 0.07%); ~0.00009 (for $\alpha = 0.10$, 0.041%); ~0.00008 (for $\alpha = 0.05$, 0.029%); ~0.00007 (for $\alpha = 0.02$, 0.020%); ~0.000043 (for $\alpha = 0.01$, 0.012%); and ~0.000028 (for $\alpha = 0.005$, 0.007%).

Characteristics of the Dixon and Grubbs Tests

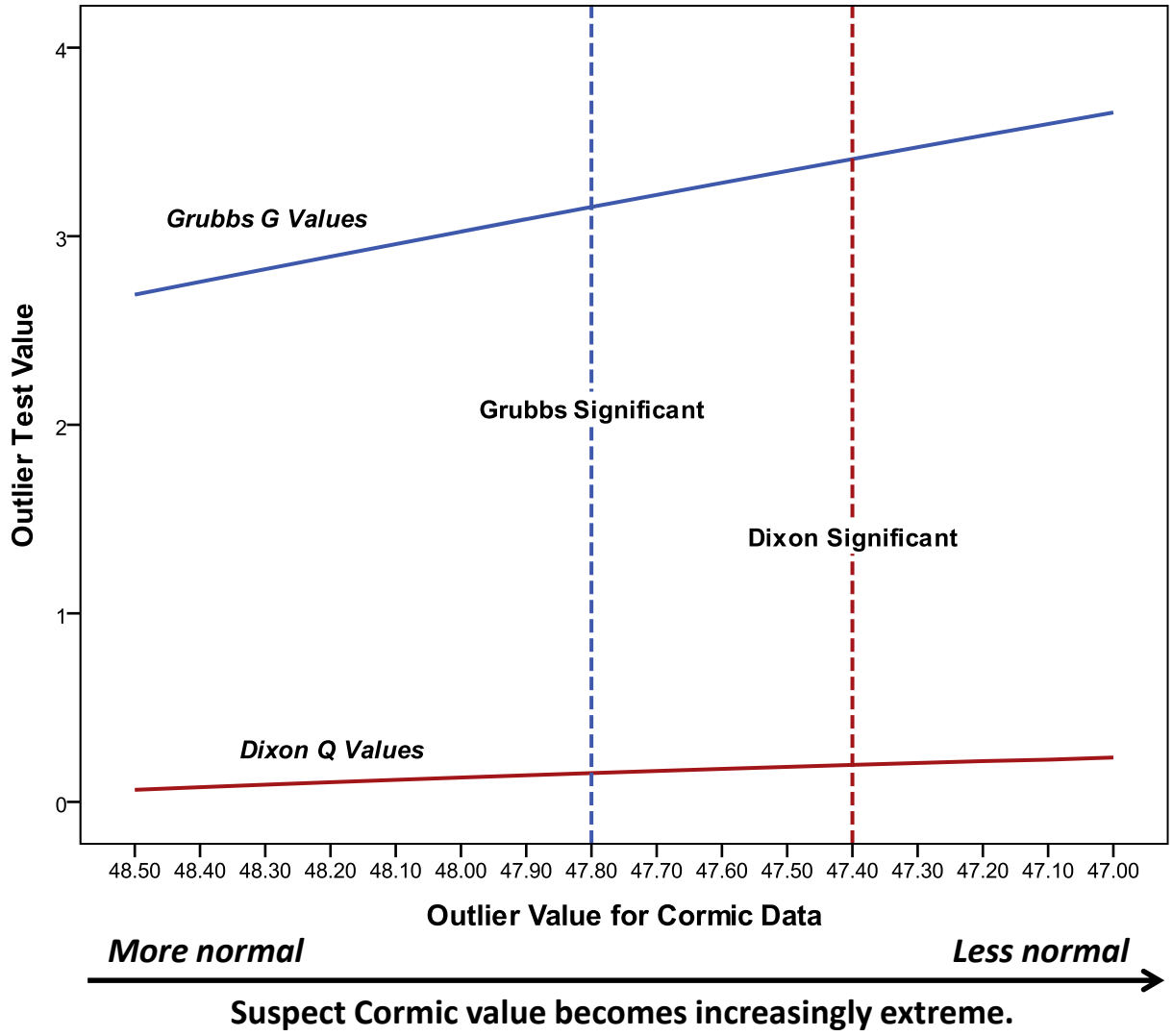
Dixon Q:

- Is the ratio of the ‘outlier gap’ to the data range.
- Similar to the w/s (range) normality test.

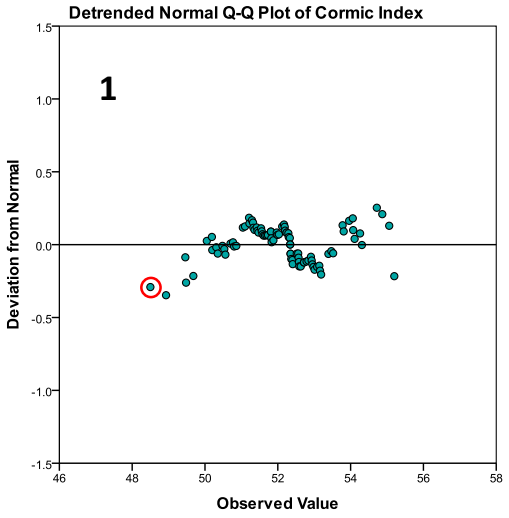
Grubbs G:

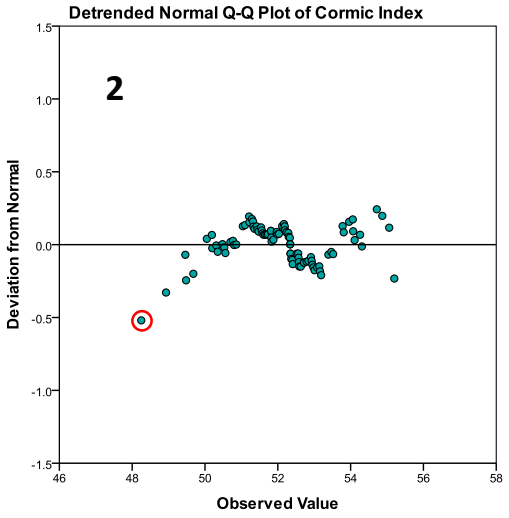
- Is essentially a z score that references a modified t table.
- Very similar to a one-sample t test.

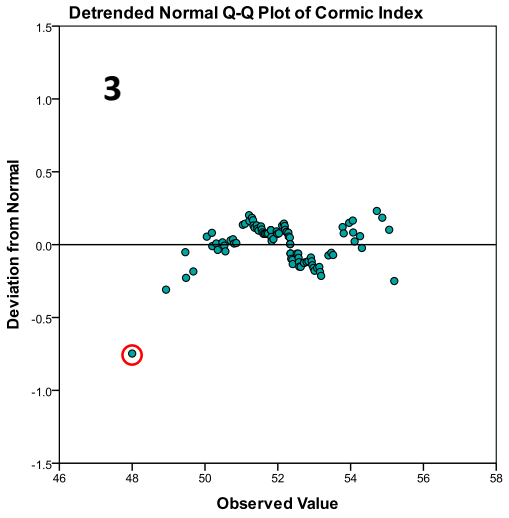
The Grubbs test picks up extreme values earlier than the Dixon test, so choose the test that is most appropriate based on your knowledge of the data

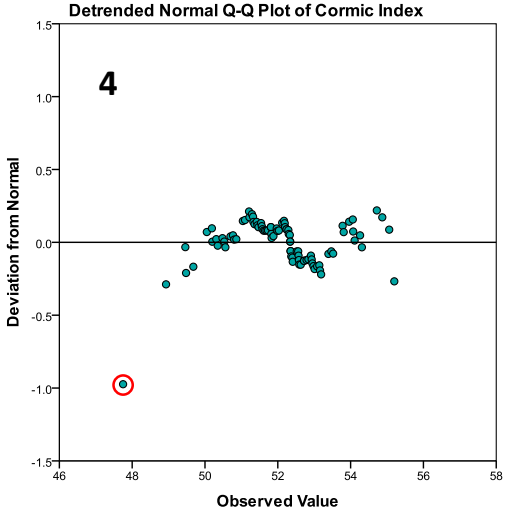


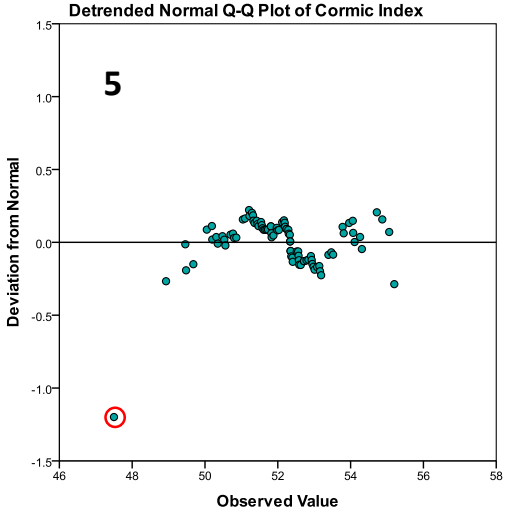
The same data used the generate the previous graph, displayed as a detrended Q-Q plot.

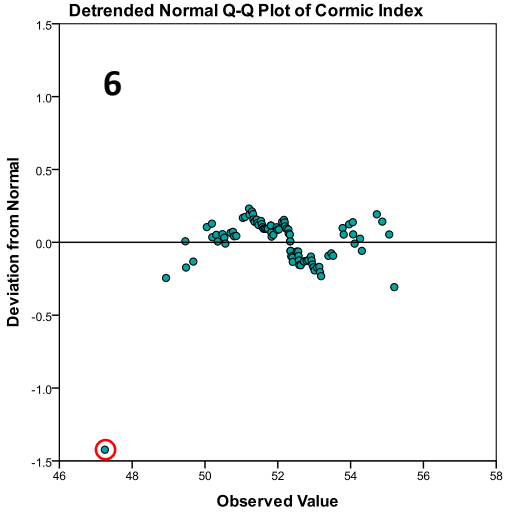












Final notes:

Outlier tests are an iterative process.

1. Check most extreme value for being an outlier.
2. If it is, remove it.
3. Check for the next extreme value using the new, smaller sample.
    - It is smaller because the first outlier was removed.
4. Repeat the process.

Once all outlier are removed the sample can be analyzed.