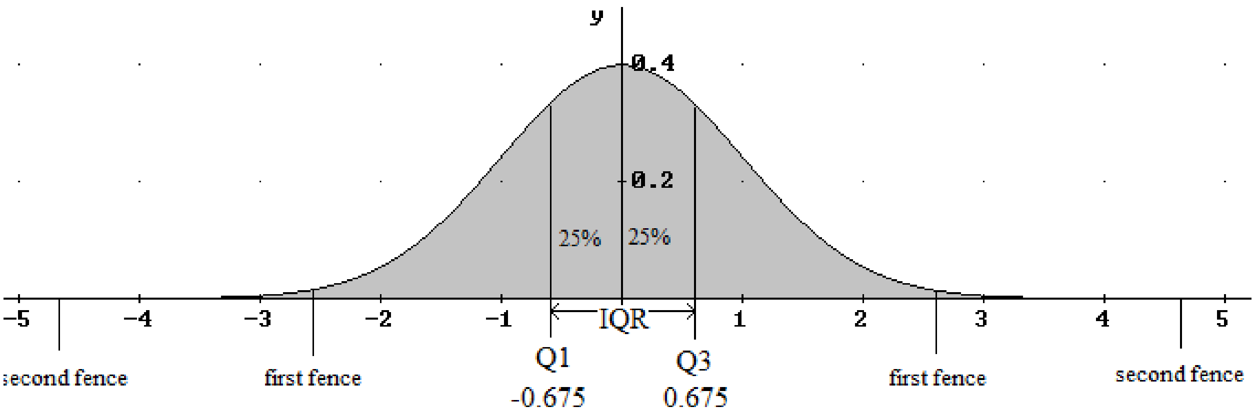
Where are the Outliers? – Prof. Richard B. Goldstein

John Tukey's Fences

Using the inter-quarterly range of IQR = $Q_3 - Q_1$ Prof. John Tukey's *Exploratory Data Analysis* set inner fences at $Q_1 - 1.5*IQR$ and $Q_3 + 1.5*IQR$ to represent the acceptable values. Values below $Q_1 - 1.5*IQR$ or above $Q_3 + 1.5*IQR$ are known as outliers. A second set of fences at $Q_1 - 3*IQR$ and $Q_3 + 3*IQR$ separate the extreme outliers.



For a standard Gaussian Normal Distribution,

$Q_1 = -0.67449$ $\quad Q_3 = 0.67449$ $\quad IQR = 1.34898$

$Q_1 - 3*IQR = -4.721428$ $\quad Q_1 - 1.5*IQR = -2.69796$

$Q_3 + 3*IQR = 4.721428$ $\quad Q_3 + 1.5*IQR = 2.69796$

How likely is an outlier for a normal distribution? The area between the first and second fences is 0.3488% on each side or 0.6976% totally. Therefore, roughly 1 out of 140 values is an outlier.

How likely is an extreme outlier? The area beyond the second set of fences is 1.17×10^{-6} on each side or 2.34×10^{-6} totally. Therefore an extreme outlier is roughly 1 out of 430,000.

What if the data is not normally distributed?

Consider n = 120 values of LDH (lactate dehydrogenase) taken in a laboratory. For the table shown on the following page

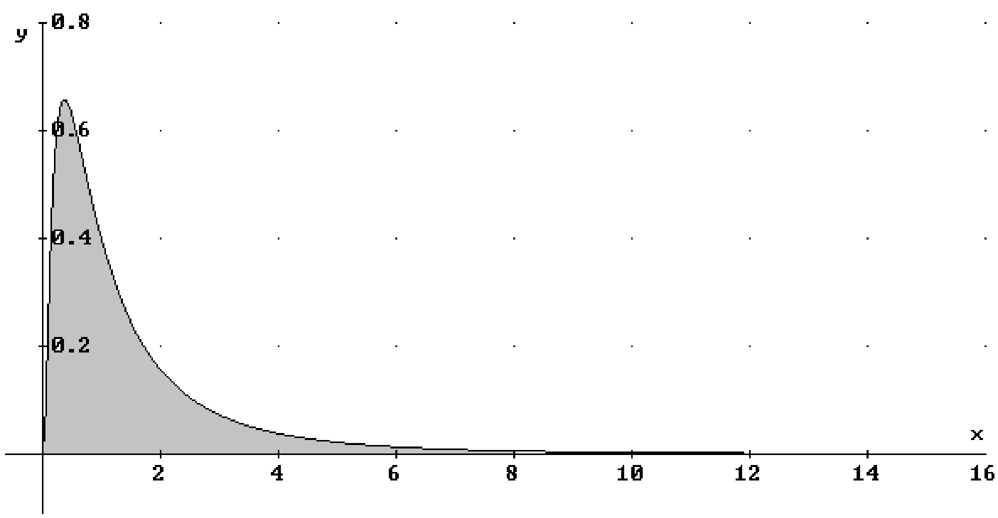
$Q_1 = x_{(30.75)} = 493 + 0.75(500 - 493) = 498.25$ $\quad Q_3 = x_{(90.25)} = 814 + 0.25(814 - 814) = 814$

IQR = 814 – 498.25 = 315.75

The fences on the right are 814 + 1.5(315.75) = 1287.625 and 814 + 3(315.75) = 1761.25. There are 9 values above the lower, inner fence of which 4 are above the outer fence.

#	LDH	#	LDH	#	LDH	#	LDH	#	LDH	#	LDH
1	321	21	469	41	537	61	609	81	717	101	900
2	324	22	472	42	538	62	622	82	720	102	934
3	357	23	472	43	544	63	635	83	723	103	939
4	377	24	478	44	547	64	642	84	729	104	958
5	387	25	480	45	549	65	644	85	739	105	983
6	403	26	481	46	550	66	650	86	762	106	1023
7	423	27	483	47	552	67	651	87	766	107	1077
8	428	28	490	48	553	68	653	88	792	108	1082
9	431	29	492	49	555	69	663	89	797	109	1130
10	434	30	493	50	564	70	672	90	814	110	1144
11	436	31	500	51	569	71	673	91	814	111	1168
12	442	32	509	52	572	72	674	92	819	112	1333
13	447	33	512	53	573	73	684	93	825	113	1368
14	448	34	513	54	575	74	687	94	828	114	1383
15	448	35	519	55	576	75	691	95	830	115	1385
16	452	36	531	56	576	76	694	96	838	116	1404
17	457	37	532	57	581	77	698	97	838	117	2327
18	466	38	533	58	590	78	713	98	845	118	2614
19	467	39	534	59	603	79	716	99	853	119	4537
20	469	40	536	60	608	80	717	100	864	120	66592

Could we have 9 outliers of which 4 are extreme outliers? What if the distribution is a highly skewed distribution like the log-normal distribution?



The normal distribution has the density function of $f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$ and the log-normal distribution has a density function of $f(x) = \frac{e^{-[\ln(x)]^2/2}}{x\sqrt{2\pi}}$.

Here, $Q_1 = e^{-0.67449} = 0.50942$, $Q_3 = e^{0.67449} = 1.96303$, IQR = 1.45361 and the fences on the right are at 4.14345 and 6.32875. Now, 7.758% of the values are to the right of the inner fence and 3.257% are to the right of the outer fence. Data falling into this distributional shape would seem to be outliers when they are just in the upper tail. Since most tests for outliers such as Tukey's fences or Grubb's test (see articles: http://en.wikipedia.org/wiki/Grubbs'_test_for_outliers and http://www.graphpad.com/quickcalcs/GrubbsHowTo.cfm) assume that we have a normal distribution, what should the analyst / statistician do?

Transforming the Data

There are transformations such as the Box-Cox transformation $y = \frac{x^{\lambda} - 1}{\lambda}$ that can make the distribution visually appear normally shaped and more importantly pass tests for normality (see Testing for Normality: http://www.providence.edu/mcs/rbg/stat/Testing_for_Normality.pdf). This was done on the LDH data with all 120 values. Both the skewness and kurtosis were unacceptable by D'Agostino's test. With 119 values the kurtosis was acceptable but the skewness was just barely acceptable. With 118 values both skewness and kurtosis were acceptable. The transformation used was $y = \frac{(\text{LDH})^{-0.7} - 1}{-0.7}$.

Letting the two highest values of LDH (4537 and 66592) be outliers and using the transformed data on the remaining 118 values the first fence will be at 2873 so that only the two highest values #119 and #120 would be declared outliers. It is somewhat circular whether one should declare outliers first and then do the transform or do the transform and then look for outliers. The number of outliers should be small in number typically at most one or two. A statistical definition states that an outlier should be "*a point in a sample widely separated from the main cluster of points in the sample.*" They should be errors in measurement or extremely unlikely events. Assuming that the LDH data is normally distributed, which of course it isn't by any test of normality, would produce an unacceptable 9 outliers.