Quiz #2 Identifying Univariate Outliers / Influential Data Points

The purpose of any sample is to represent a particular population. "Extreme" scores in a sample are likely to influence the sample-based estimates of both center and the spread of the population scores. There are many sources of extreme scores, sampling a member not of that population, bad measurement or recording, errors in data entry, etc. For whatever reason they have come to exist, extreme data points will lessen the ability of the sample statistics to represent the population of interest. We will take a 4-fold approach to outliers. First, make every effort to obtain accurate measures of members of the sample of interest and to enter those values correctly. Second, employ a conventional technique for identifying value that are "extreme" and so will influence the mean and spread values for the sample. Three, carefully examine each identified outlier to determine its cause. If there has been an error, change the data values accordingly. If there is no identifiable error, consider what it might mean if the value is accurate, perhaps analyzing outlier data separately. Fourth, exclude outliers from the primary analyses, to improve the representativeness of the sample, and so of the statistical results and interpretations.

The procedure for identifying outliers is taken from the following reference. You should cite this text whenever you write a paper using this procedure. Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). Understanding Robust and Exploratory Data Analysis. New York: Wiley.

The process:

1. Order the values and note the depth of each (the rank from the nearest extreme value).

depth	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
value	32	47	53	59	77	77	81	90	96	118	120	120	131	135	143	151	162	174	187	189	195	205	210	220	248	281	300	309	337	475

2. Find the DEPTH of the median as: (N + 1) / 2 = 31 / 2 = 15.5

3. Find the DEPTH of the fourths (similar to quartiles) as: (median depth + 1)/2 = (15 + 1)/2 = 16/2 = 8

   NOTE: When computing fourths, always drop any fractional part of the depth of the median before adding the 1. In this case, the calculated median depth was 15.5. The fraction is dropped, changing the median depth to 15, which is then entered into the formula.

4. Find the VALUES of the fourths: the lower fourth has a value at depth of 8 is 90; the upper fourth is 210.

5. Find the FOURTH SPREAD as the difference between the values of the 1st and 3rd fourths:

   210 - 90 = 120

6. An outlier is defined as any score which is more than 1 1/2 fourth spreads beyond either fourth. If the data were normally distributed about 7/1000 cases would be identified as outliers.

   Lower outlier bound is = lower fourth value - 1.5(fourth spread) = 90 - 1.5(210 - 90) = -90 (no "too small" outliers)

   Upper outlier bound is = upper fourth value + 1.5(fourth spread) = 210 + 1.5(210 - 90) = 390 (475 is an outlier)

REMEMBER: Identifying and "tossing" outliers is not the end of the process! You should try to determine why they happened and if the data values can be "recovered" or perhaps the outliers provide you with some useful information.

Identifying Outliers: More about this interpolation stuff.....

Whenever the depth of a median or a fourth is a decimal (??.5), then you must interpolate. That is, you must find the value of the median or fourth by taking the average of the two values with adjacent depths.

Example #1 -- no interpolation necessary

depth ==> 1 2 3 2 1 remember mdn depth = (N + 1) / 2, 4th depth = (mdn depth + 1) / 2
value ==> 10 12 15 19 21

The depth of the median is (5 + 1) / 2 = 3, so the value of the median is 15

The depth of the fourths is (3 + 1) / 2 = 2 , so the values of the 1st and 3rd fourths are 12 and 19, respectively.

Example #2 -- interpolation necessary for the median and the fourths

depth ==> 1 2 3 4 4 3 2 1
value ==> 21 26 30 36 37 38 42 46

The depth of the median is (8 + 1) / 2 = 4.5, so the value of the median is interpolated as the average of those values with depth = 4, (36 + 37) / 2 = 36.5

The depth of the fourths is (4 + 1) / 2 = 2.5 , so the values of the 1st and 3rd fourths are interpolated as the average of those values with depths of 2 and 3, (26 + 30) / 2 = 28 for the 1st fourth
(38 + 42) / 2 = 40 for the 3rd fourth

NOTE: Since the median depth was a fraction (4.5), the median depth was truncated to a whole number (4) before being applied to formula to compute the depth of the fourths.

Example #3 -- interpolation necessary for median, but not for the fourths

See the example on the last page -- the median would be (143 + 151) / 2 = 147
Note: Again, the fractional median depth was truncated to compute the fourth depth.

Example #4 -- interpolation necessary for the fourths, but not for the median

depth ==> 1 2 3 4 3 2 1
value ==> 21 26 30 36 37 38 42

The depth of the median is (7 + 1) / 2 = 4, so the value of the median is 36

The depth of the fourths is (4 + 1) / 2 = 2.5 , so the values of the 1st and 3rd fourths are interpolated as the average of those values with depths of 2 and 3, (26 + 30) / 2 = 28 for the 1st fourth and
(37 + 38) / 2 = 37.5 for the 3rd fourth

When you will have to interpolate is a function of the number of data points, but there are only these four combinations.

Outlier Analysis via SPSS

NOTE#1: **This is a two "run" process!** The **first "run"** is used to obtain the Q1 & Q3 statistics needed to compute the "extremes" to define the outliers. The **second "run"** uses these values to select the "nonoutliers" and get the "screened" statistics.

NOTE#2: The by-hand version used fourths to determine the "limits", whereas the by-SPSS version uses quartiles. The difference between the results of these two procedures has been demonstrated to usually be negligible. Procedurally, the difference is that when computing the depth of a fourth, any factional part of the median depth is truncated, while "proper interpolation" is used when calculating quartiles.

1st Run -- getting the values to identify outliers

Analyze → Descriptives → Frequencies

- Select the desired variable(s)
- Select "Statistics" and request quartiles

Statistics

SCORE

N	Valid	30
	Missing	0
Mean		167.4000
Std. Deviation		100.1439
Percentiles	25	87.7500
	50	147.0000
	75	212.5000

Applying the formulas to lientify the upper and lower extremes:

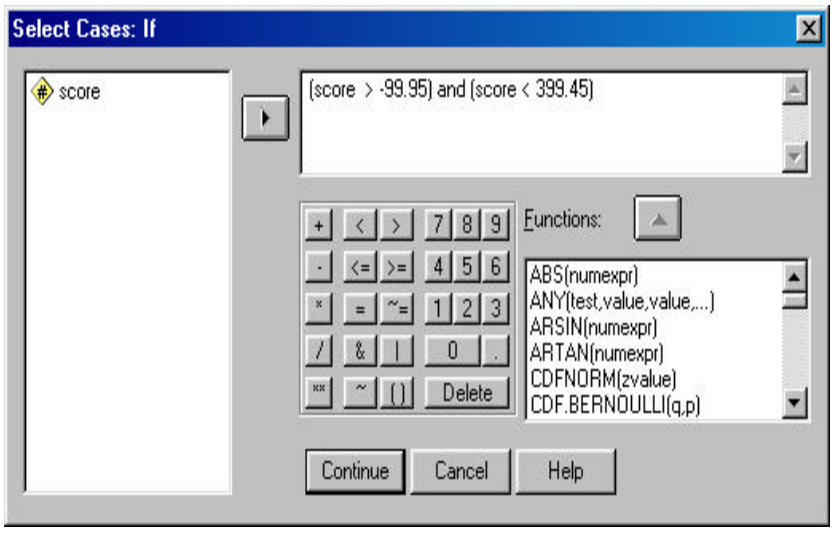
lower extreme = Q1 - 1.5(Q3 - Q1) = 87.75 - 1.5(212.5 - 87.75) = -99.95

upper extreme = Q3 + 1.5(Q3 - Q1) = 212.5 + 1.5(212.5 - 87.75) = 399.45

2nd Run -- getting the screened statistics

Data → Select Cases

- Click the "If condition is satisfied"
- Press the "If…" button
- specify "limits" of "non-outliers"



Then
Re-run the frequencies to get the screened univariate statistics

Statistics

SCORE

N	Valid	29
	Missing	0
Mean		156.7931
Std. Deviation		83.0135

These are referred to as the "screened" mean and Std. We believe this to be a better estimate of the pop values

REMEMBER, while screening can help eliminate the over-influence of extreme data points, identifying and "tossing" them is not the complete process. You should make every attempt to determine WHY the outlier values occurred!! If they were produced by a data-entry error, you can correct the file and not lose the data point. If the subject really does represent another population or a "special case" of the population of interest you will want to know this information.