OUTLIERS

QUESTION:

I would be interested in what people tell students they are supposed to do with such data points identified as "outliers".

ANSWERS:

1) I generally tell my students that they should conduct a separate analysis of the data without the outliers, and then mention the outliers separately. For example, at this early stage of the course they might say something like "Excluding the two very unusual students who have well over 100 CDs in their music collections, the mean number of CDs among the other students in this class was 42."

2) Go back and verify the data is correct/real, if that is possible. Make sure it isn't the result of a data entry error. Also, for a survey check to see if it might be the result of someone giving false answers:

Ex: How much do you weigh? 6502 lbs

If it's a data entry problem, fix it. If you're sure it's a false answer, perform the analysis with and without the false answer, but draw your conclusions from the analysis without the outlier and explain why to your reader.

If you have no obvious reason to declare the point(s) invalid, then perform the analysis with and without the outlier(s). If your conclusions don't change, no further action is probably needed. If your conclusions do change, then you could repeat the evaluation with a larger sample. If the outlier(s) occur again, it might be that your population may have an unusual distribution (non-normal at least). This might indicate alternate analysis methods are necessary to make conclusions. If you have a "hunch" as to what might have caused the outlier and sufficient time & money, design a new study/experiment to evaluate that hunch.

3) These are good suggestions. I find one has to deal with outliers and data quality issues all the time in real life. I remember long ago when two of my students did an internship at the Mayo Clinic and spent most of their time cleaning up the data! Data entry errors can often introduce multiple outliers of which you only detect one but then clean up the rest. I used to use this example in class: What would you think if you saw an age of 3264? This was in Plymouth, NH, where the ZIP was 03264. So maybe the respondent put their ZIP where their age went on the form, but also maybe the data entry clerk left out the age and so other fields are off by one position -- ZIP in the age field, weight in the ZIP field, etc. Once a bunch of twos in a SEX variable coded 0-1 revealed the entire data deck (yes, punched cards!-) was scrambled so most of the data was wrong!

Unfortunately, in a textbook/classroom situation, we are rarely able to go back and check these things out. Usually someone else has already cleaned up the data. Often we do not even SEE the data -- just summary statistics.

Of added relevance to AP Stats.:

Lots of outliers on one side of a boxplot may be a sign of skewness rather than an outlier problem.

Lots on both sides may indicate heavy tails.

Bimodality can mask outleirs in a boxplot -- AND a boxplot is not very good at showing bimodality. So for a single variable, I would not make a boxplot my first (and certainly not my ONLY) display.

A realistic source of apparent bimodality: more than one unit used to measure the same thing, so the same height might be 6.0, 72 or 183 depending on who measured it.