TAKE IT TO THE MAT

A NEWSLETTER ADDRESSING THE FINER POINTS OF MATHEMATICS INSTRUCTION

Southern Nevada Regional Professional Development Program

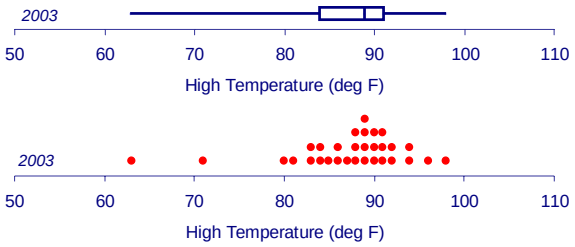
December 2003 — High School Edition

rpdp.ccsd.net

In the last two issues of *Take It to the MAT*, we looked at the process for constructing box-and-whisker plots and comparing distributions using parallel boxplots. In this edition, we will expand our construction of box-and-whisker plots to include *outliers*, those values that are outside the general pattern of the data.

Las Vegas High Temps 10/03 (in ºF)
98
96
94
94
92
92
91
91
91
90
90
90
89
89
89
89
88
88
88
87
86
86
85
84
84
83
83
81
80
71
63

We'll continue to use the daily high temperatures in Las Vegas during October 2003. The data in the table have been sorted to facilitate analysis. The box-and-whisker plot below has a very long whisker between the minimum value (63º) and the first quartile (84º). Indeed, a line plot (sometimes called a dot plot) of the temperatures shows that both 63º and 71º appear to be set apart from the rest of the data.



Are those two values *outliers*? They appear to be, but are they really? Is whether or not an observation is an outlier in the eye of the beholder? To some degree, yes. But, statisticians have a way to measure whether an observation is an outlier. Just like we can quantify center by calculating the mean or median, or quantify spread by computing range or standard deviation, we can quantify whether an observation is an outlier.

The method we will use is called Tukey's Rule. It states that an observation is an outlier if:

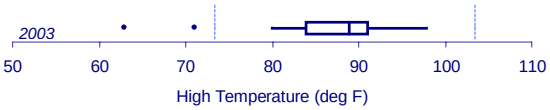
—it lies more than 1.5 × interquartile range below the first quartile (Q1 – 1.5·IQR)

—it lies more than 1.5 × interquartile range above the third quartile (Q3 + 1.5·IQR)

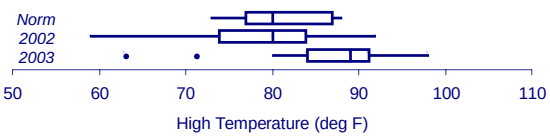
The *interquartile range* is the distance between the first and third quartile. It is a measure of spread that gives the range of the middle half of the data. For the October 2003 data, the five number summary is {63, 84, 89, 91, 98}. The interquartile range is IQR = Q3 – Q1 = 91 – 84 = 7. By Tukey's Rule, an observation is considered to be an outlier if it is smaller than Q1 – 1.5·IQR = 84 – (1.5)(7) = **73.5**, or larger than Q3 + 1.5·IQR = 91 + (1.5)(7) = **101.5**. Thus, both 63º and 71º are outliers in this data set.

The values of 73.5 and 101.5 computed above are often called *fences*. Any observation beyond the fences, when looking from the center, is an outlier. To create a boxplot showing outliers, we extend the whisker from the first quartile to the smallest observation inside the fence. We similarly extend the whisker from the third quartile to the largest observation inside the fence. Any observations considered outliers are marked with single points.

There are no high outliers. The two outliers, 63º and 71º, are marked with dots and the lower whisker is extended from the first quartile (84º) to the smallest value that is not an outlier, which is 80º.



Rather than looking like an extremely skewed distribution as the first boxplot suggests, the distribution now looks more symmetric with a couple of outliers. We would draw a similar conclusion from examining the dotplot.



Remember, the power of the boxplot is in comparing distributions. In the last issue, we compared temperatures in October 2003 to those in October 2002 and the average since 1937. (See the November 2003 issue for raw data.) There were no outliers in October 2002 and we wouldn't expect any for the normal temperatures because they are averages. Now that we've separated October 2003's outliers, what new conclusions can you draw?