COPING WITH STATISTICS -- A PRIMER FOR LIBRARIANS AND TRUSTEES

by John Borneman

WHY UNDERSTAND STATISTICS?

> *"... and although we may quote one to another with a chuckle the words of the Wise Statesman, 'Lies–damned lies–and statistics,' still there are some easy figures the simplest must understand, and the astutest cannot wriggle out of."*
> *Leonard Henry Courtney*

These are times of tighter budgets — for governments, industry, and individuals. Consequently, many people are questioning whether they are "getting their money's worth" from various tax supported entities, including libraries. Statistics are being increasingly used to attempt to prove or disprove the value of libraries.

A quick search on the Internet for news articles on libraries and statistics revealed these two good examples of this demand for data from libraries.

- *"...on February 15, commissioners asked library officials to provide statistics showing the amount of use the facilities get." (Lynch, 2006)*
- *"...These statistics will help the library determine 'how many of our resources do we put into books vs. technology...' " (Bournea, 2006)*

Therefore, instead of being a victim of other people's use of statistics, why not use them to help improve your own library or to help sell your library's effectiveness? The use of statistics for libraries may be placed into two categories. One, using data to make comparisons *between* libraries and two, using data to evaluate the performance or track record *within* one's library. However, be cautious. Much of the dislike of statistics comes from the misuse of data rather than its use. When creating or evaluating statistics, keep in mind the following points as described in The Internet Public Library:

- understand from where the numbers came (the "source");
- understand how the numbers were collected (sometimes given in a footnote);
- understand what date range the statistics cover (usually different than the date the statistics were published);
- understand who collected the data (how reliable is the agency or group who collected and analyzed the numbers in order to come up with the statistics).

MAKING COMPARISONS BETWEEN LIBRARIES

Distributions, Normal Data, and Averages

Any set of related data will generally fall into a distribution or pattern. Often data follows what is known as the "bell curve," or normal distribution. In the normal distribution, all the values of the data are equally centered about the average value of the data, and the most common values lie near that average. For example, if one were to measure the height of a randomly selected group of women, their heights would fall into a normal distribution.

Distributions of related data points may be described by defining three values: a measure of central tendency (in the case of the normal distribution, this is the average), a measure of the spread of the data (the standard deviation in a normal distribution), and an expression of the shape of the distribution (the bell curve in a normal distribution). (Note: instructions on how to calculate the standard deviation are outside the scope of this article.)

As can be seen in Figures 1 and 2, if the investigator calculates the average and the standard deviation, and if the data is normal, then a great deal of understanding of the entire data set is known. We can calculate the percentage of data over, under, or between any values we choose. And if we want to compare our library's data point with all the other libraries, we can calculate the percentage of libraries with higher or lower numbers than ours. However, many times a set of data is not evenly distributed about the average value and the distribution is skewed to one side or another. Additionally, some groups of data contain values that lie far outside the expected range of values. These are called

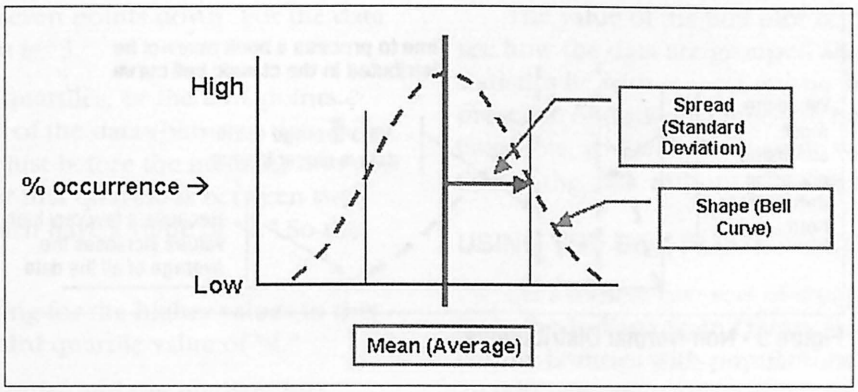

Figure 1 - The Normal Distribution

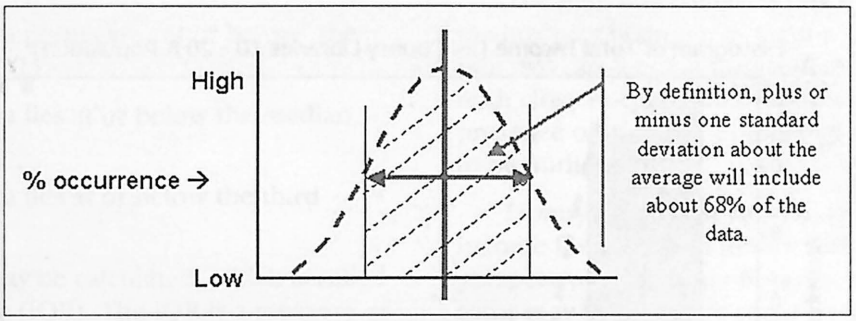

Figure 2 - Standard Deviation of a Normal Distribution

"outliers" and can point the investigator toward unique or special events and situations. In these cases, calculating the average and standard deviation is not always valid. (See Figure 3.)

Data may also be expressed in a histogram, which is one method of showing the actual distribution of our data. The chart in Figure 4 (Indiana State Library) shows that the distribution of the total income of county libraries (with populations between 10,000 and 20,000 persons). This data is fairly normal but does have some points that are potentially outliers.

Because some data is not normally distributed and may contain unique "outlying" data points, what is required is a set of statistics that are not dependent on how the data are distributed.

CALCULATING STATISTICS USING THE MEDIAN INSTEAD OF THE AVERAGE

Several years ago, a statistician named John Tukey promoted the use of a set of statistics other than the average and standard deviation. He felt that in many cases, investigators were assuming data were normal (and therefore that the average and standard deviation were valid statistics), when in fact the data were not normally distributed.

The math of these alternate statistics is simple to calculate and has the advantage of allowing the investigator to easily observe:

- the central tendency of the data
- the spread of the data about the center
- the total range of the data (highest point minus the lowest point)
- the shape of the distribution
- the presence (if any) of outliers or "odd" data points.

Tukey encouraged investigators to use the median of the data and what is called the "interquartile range" instead of the classic average and standard deviation — at least during initial observations and evaluations. The median, being the data point that lies in the middle of the data set, reduces the influence of unique or outlying data points, which artificially raise the average.

The first step in looking at data as Tukey proposed, is to calculate some basic parameters:

- the median (a measure of the central tendency of the data)
- the interquartile range (a measure of the spread of the data)
- the values of data that delineate any outliers.

As shown in the left set of data in Figure 5, the median is derived by first ordering the data in increasing value. Then by counting the total number of data points (in this case thirteen), we can find the median

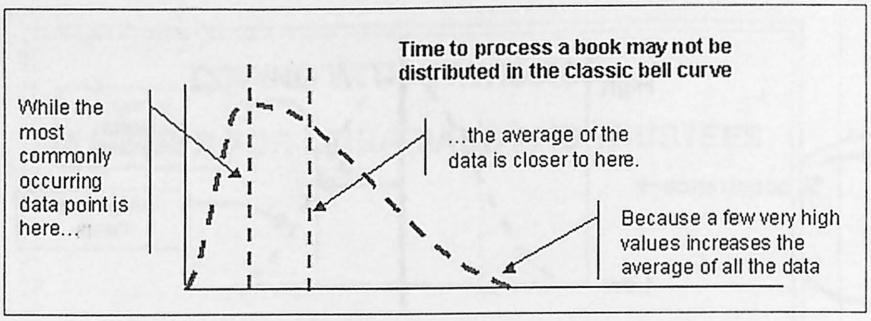

Figure 3 - Non-Normal Distributions

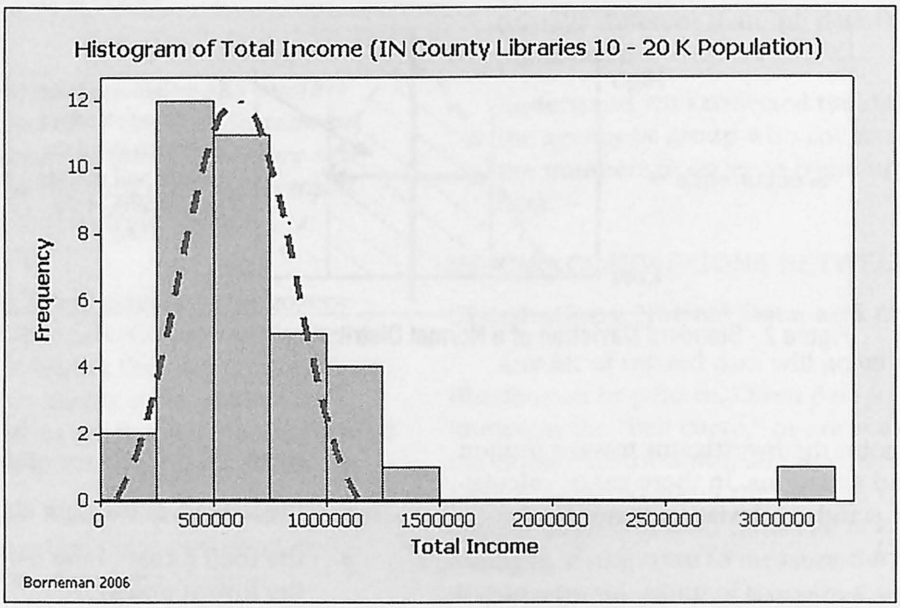

Figure 4 - Histogram (Distribution) of Total Income(2)

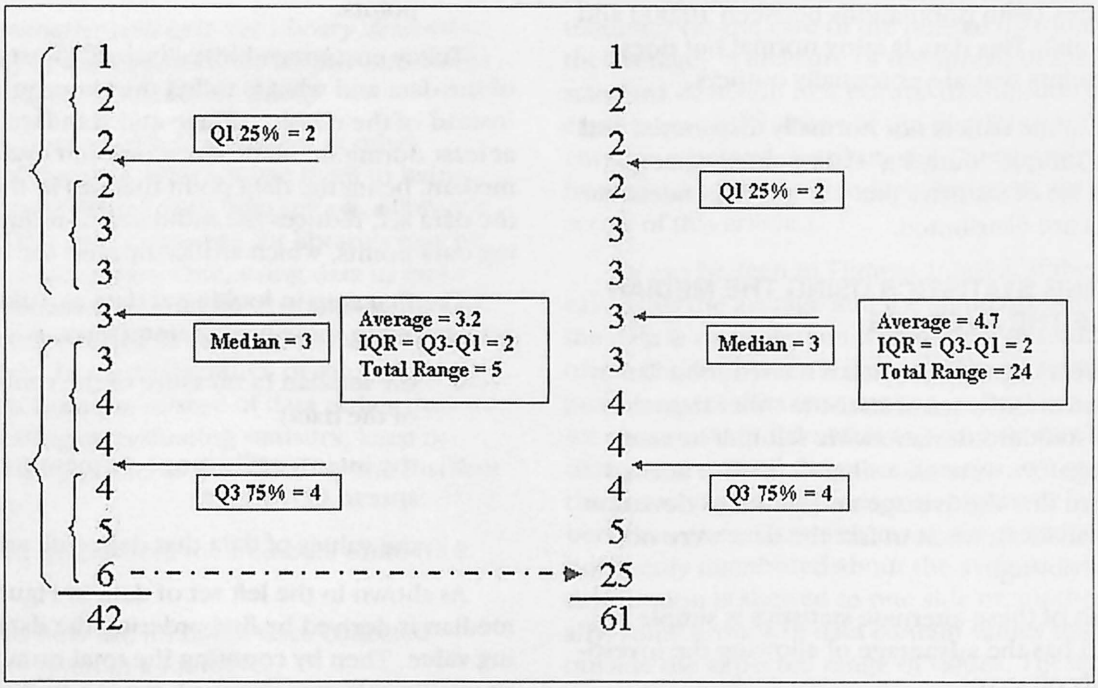

Figure 5 - Median And Quartile Calculations With A Comparison To The Average

(or middle) value lies seven points down. For the data in Figure 5, the median is "3."

Next calculate the quartiles, or the 25% points. Dividing the lower half of the data (between data point "1" and the data point just before the median) into two halves, we find that our first quartile is between two data points both of which have a value of "2." So the first quartile is "2."

Doing the same thing for the higher values in this set of data gives us a third quartile value of "4."

To understand quartiles and the median, just remember that:

- 25% of the data points are "at or below" the first quartile (Q1)
- 50% of the data lies at or below the median, and
- 75% of the data lies at or below the third quartile (Q3).

Another statistic may be calculated, which is called the Interquartile Range (IQR). The IQR is a measure of the spread of the data and is merely the value of the third quartile minus the value of the first quartile (Q3 – Q1 = IQR).

Finally, to calculate the limits for the outliers, just take the IQR (in this case "2") and multiply by 1.5. Then add this number onto the third quartile value or subtract it from the first quartile value.

For our very simple example, these upper and lower limits of "expected" data would be:

- Q3 + (2 x 1.5) = 4 + 2.5 = 6.5
- Q1 – (2 x 1.5) = 2 – 2.5 = 0
  (since we assume our data cannot be negative).

Therefore any data points greater than 6.5 are potentially outliers and may deserve special investigation. In the two data sets above, the first set of data does not have any outliers. All values of the data lie between 0 and 6.5. However, in the second set of data on the right side of the chart, the data incorporate an unusually high value ("25"). Observe that the key statistics are the same for this modified set of data, but the average has increased.

Additionally, since the upper limit for outliers is still calculated at 6.5, we can see that this data value of "25" is indeed an outlier. If this were a real set of data, we would begin to question where this data point came from. Remember, sometimes outliers are just transcription errors, so always check data for accuracy.

Tukey also created a graphical method of looking at the median and IQR data, which he called a "Box Plot." See Figure 6 for an annotated version of a box plot.

The value of the box plot is that, at a glance, we can see how the data are grouped and where the key statistics lie with respect to one another as well as the presence of outliers. Once you have studied enough box plots, you will also be able to assess the distribution of the data without plotting histograms.

USING THE BOX PLOTS TO COMPARE LIBRARIES

Let's look at two sets of comparative data taken from the Indiana State Library Statistics webpage for county libraries with populations between 10,000 and 20,000 persons.

The first data collected was for Total Income and the box plot is shown in Figure 7. Looking at this box plot, we can see that the median and the average are both close to each other in value. We also see the presence of two data points that have been calculated to be outliers.

Library officials or trustees can find the Total Income for their own library and easily place it in perspective to all other libraries. From that, an investigator may begin to ask some questions. For instance, if you know that your library's income is $700,000, it is easy to see that your library lies near the median. You now know that 50% of all libraries in your group have a total income that is equal to or less than your library's income.

The next box plot of data in Figure 8 presents a slightly different picture. These data (Total Operating Fund Expenditure) are skewed more toward the higher values (versus more toward the lower values in the "Total Income" box plot). There also seems to be a slightly wider spread between the median and the average, most likely due to the wider spread of data between the median and the third quartile. Look at the figure and compare the distance between the median and the third quartile with the distance between the median and the first quartile. The wider portion of the box between Q3 and the median indicates a wider spread of the data in this range.

Once again, you may compare where your library lies on the Box Plot and evaluate its position with the rest of the libraries.

In this case "your library" has an operating expenditure that is near the third quartile. In other words, "your library" spends as much or more than 75% of all the other libraries in its population group. But the real question is why? To answer this question will require deeper investigation. Look closely at all your expenditures. Ask yourself why these expenditures are being made. Consider traveling to some of the other libraries and comparing yourself to them. Pick two libraries that spend more than you and two that spend less. See how your library compares to them.

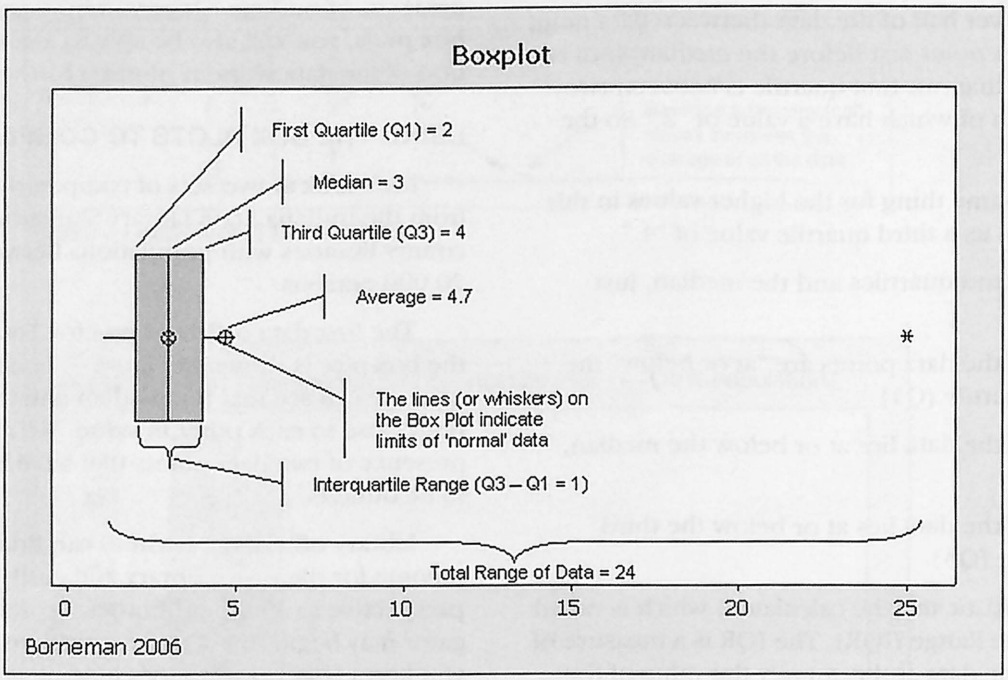

Figure 6 - Explanation of Box Plot

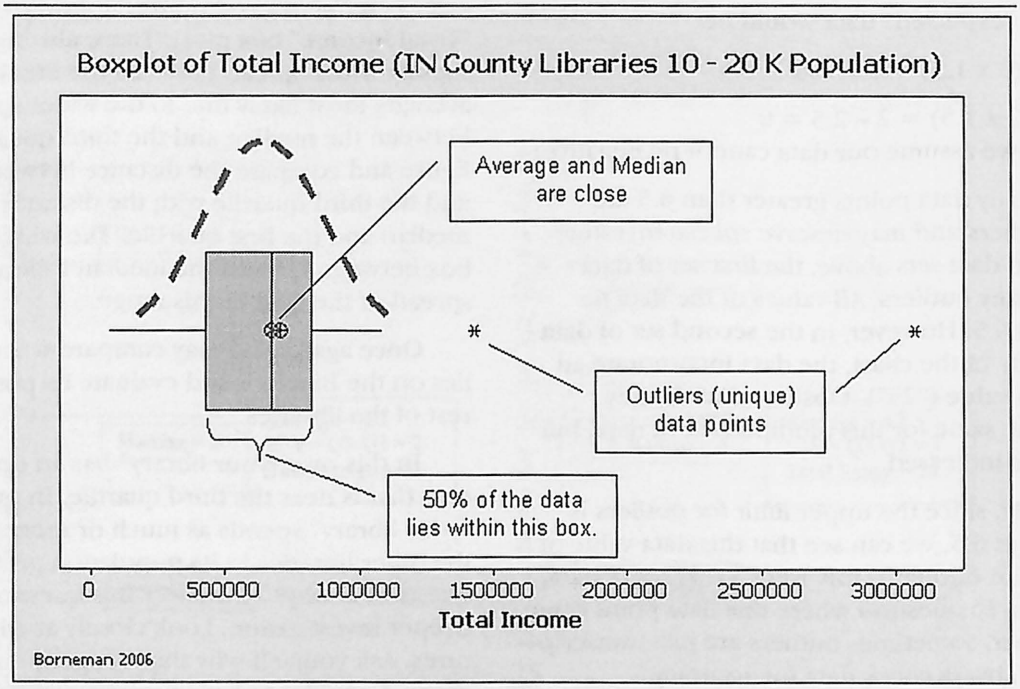

Figure 7 - Box Plot of County Library Income Data
(Compare this chart to the histogram above. Note that outliers are identified)

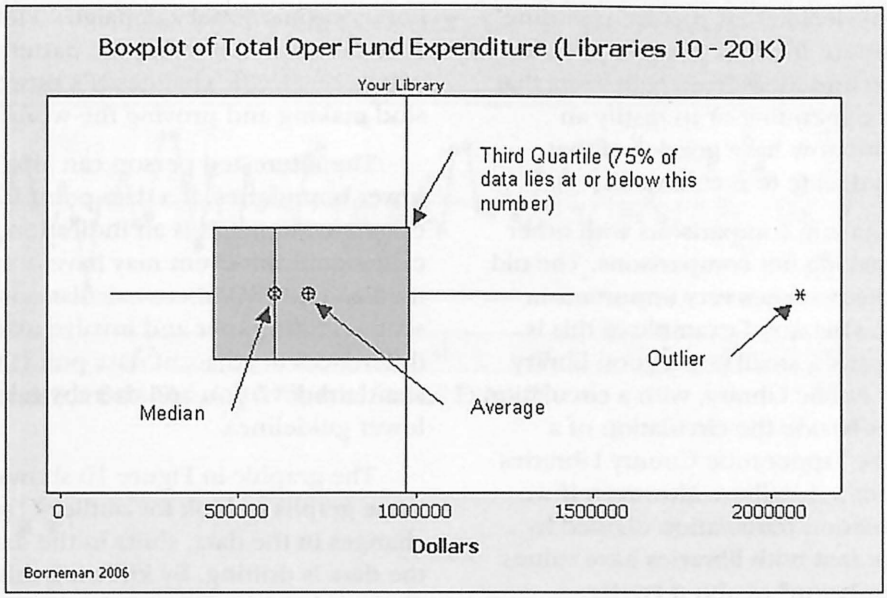

Figure 8 - Box Plot of Total Operating Expenditure of County Libraries (10-20 K)

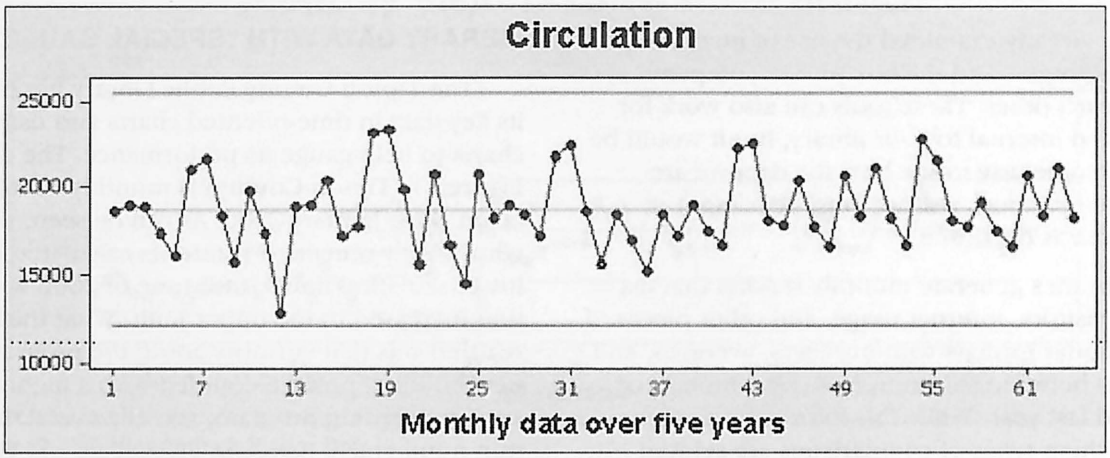

Figure 9 - Circulation Data Over Time

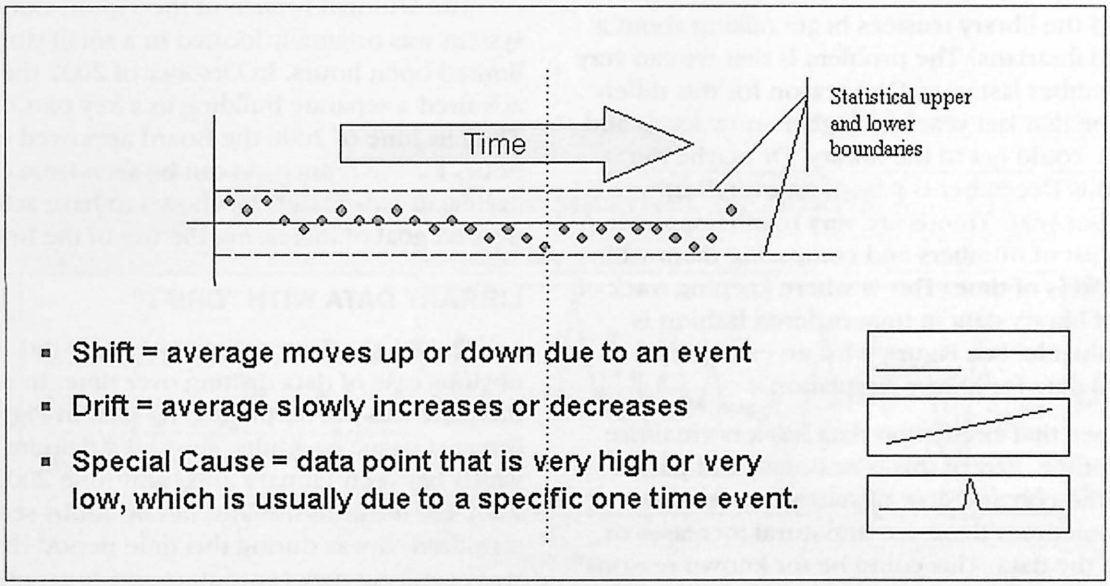

Figure 10 - Time Orientated Data

In the end, you may decide that you are spending just about the right amount for your library's specific conditions. Or you may find ideas from your visits that will allow you to reduce spending or to justify an increase. Either way, you now have good data that supports whatever you decide to recommend.

Remember when making comparisons with other libraries that you are making fair comparisons. The old "apples to apples" cliché becomes very important in statistical comparisons. One good example of this is circulation data. Obviously a small population library such as Tipton County Public Library, with a circulation of about 185,000, pales beside the circulation of a larger library such as the Tippecanoe County Libraries with circulation of about 1.4 million. However, if we adjust for county population (circulation divided by population) we can see that both libraries have values of "circulation per population" of about twelve (Tipton).

LOOKING WITHIN ONE'S OWN LIBRARY

We have already examined the use of medians, interquartile ranges, and the box plot to compare libraries to each other. These tools can also work for data generated internal to your library, but it would be nice to have some way to see how the data we are measuring (circulation, visitors, Internet usage) changes or reacts over time.

Many libraries generate monthly reports that list circulation, visitors, Internet usage, and other pieces of data in columnar formats with numbers, averages, and comparisons between this month and last month, or this year and last year. While this information can be interesting, these types of comparisons are seldom useful. For instance, what may we conclude if the circulation of December of this year is higher than December of one year ago? Is that a significant difference? Should the library trustees begin talking about a raise for the Librarians? The problem is that we can very seldom remember last year. One reason for this difference might be that last year had higher snow levels and fewer people could get to the library. Or maybe the circulation this December is part of an overall upwards trend since last year. Trends are very hard to spot when looking at a list of numbers and comparing them with previous periods of time. This is where keeping track of key pieces of library data in time-ordered fashion is extremely valuable. See Figure 9 for an example of time-ordered data for library circulation.

We can see that circulation data has a normal rise and fall over time. Part of this is seasonal, and part is just the natural ebb and flow of statistical variation. However, sometimes there are unnatural increases or decreases in the data. This could be for known reasons (such as a flood which closed the library or a new promotional campaign). Plotting information over time and observing the patterns of change and the effects of specific changes is a powerful tool for decision making and proving the worth of a change.

The interested person can also calculate upper and lower boundaries. If a data point falls outside these calculated limits, it is an indication that some statistically significant event may have occurred. Unlike the median and IRQ, these calculations are beyond the scope of this paper and involve using the average differences in adjacent data points to estimate the standard deviation and thereby calculate the upper and lower guidelines.

The graphic in Figure 10 shows how we can use these graphs to look for outliers, "special cause" changes in the data, shifts in the data, or for signs that the data is drifting. By knowing this information, the library administrators can begin to explore why these changes are happening and either celebrate the success or correct the problem.

LIBRARY DATA WITH "SPECIAL CAUSE"

The Tipton County Public Library has been tracking its key data in time-oriented charts and using these charts to help gauge its performance. The chart in Figure 11 (Tipton County) is monthly circulation data taken since January 2000. As can be seen, the data have consistently remained within its calculated upper and lower statistical limits until June of 2006 when circulation increased to the upper limit. What the librarians realized was that curiosity about the progress of their new building project, coupled with a highly effective summer reading program, actually raised the circulation numbers.

LIBRARY DATA WITH "SHIFT"

The Windfall Branch of the Tipton County Library system was originally located in a small storefront with limited open hours. In October of 2002 the Board acquired a separate building in a key part of town. Then in June of 2006 the Board approved increased hours for the branch. As can be seen from Figure 12, each action can easily be shown to have achieved its desired goal of increasing the use of the branch library.

LIBRARY DATA WITH "DRIFT"

Finally, the Tipton County Library did have a very obvious case of data drifting over time. In retrospect, this drift was not surprising. As seen in Figure 13, Internet usage by adults, showed a definite drift upwards between January 2000 and June 2002. After June 2002 the usage of the Internet by adults seems to have stabilized. It was during this time period that the library was acquiring new computers and re-arranging its computer area. Additionally, this time period was

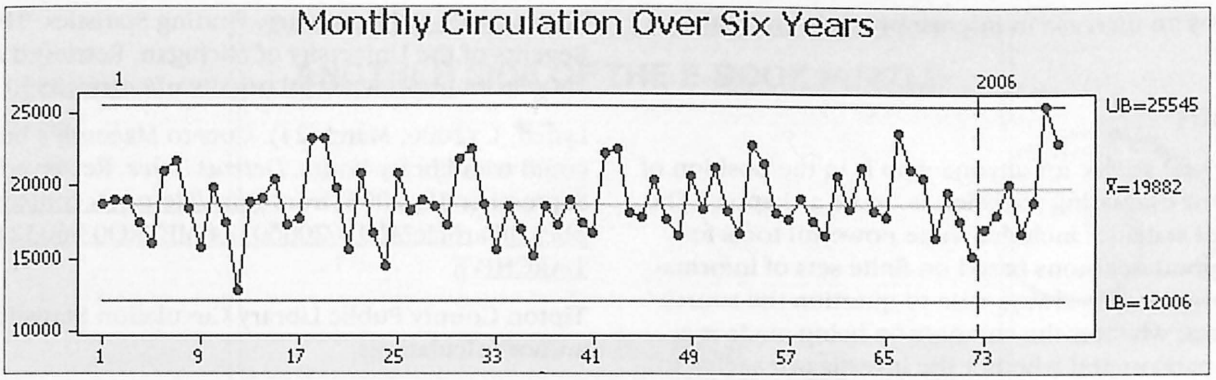

Figure 11 - Circulation Data Showing A "Special Cause" Event (3)

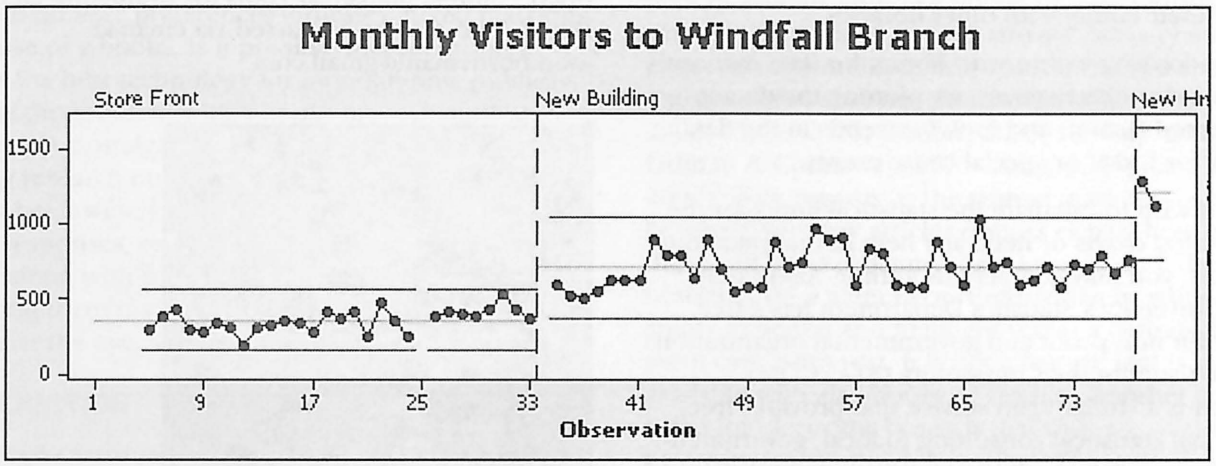

Figure 12 - Windfall Branch Visitor Data

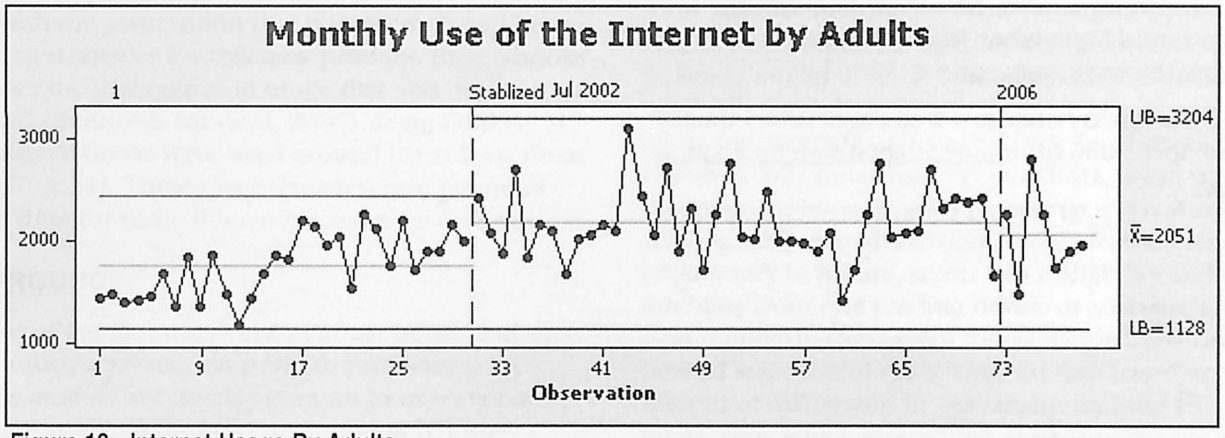

Figure 13 - Internet Usage By Adults

marked by an increase in interest by adults in "going online."

SUMMARY

The best advice for anyone who is in the position of creating or evaluating statistics, is "to be a skeptic." The science of statistics includes some powerful tools for making good decisions based on finite sets of information. However, it is always wise to question the source of the data, whether the comparison being made is a fair comparison and whether the investigator seems to have applied the appropriate types of statistics to the problem (Internet Public Library).

Remember that not all sets of data are normally distributed. Merely looking at the average value of a set of data can sometimes lead to incorrect decisions. The use of the median and interquartile range, along with the graphical box plot, can give investigators a great overview of the data and allow them to effectively compare their library with other libraries.

When looking within your library for data that can help you judge effectiveness, try plotting the data in time-oriented fashion and look for trends in the data. Also look for "odd" or special cause events.

If you wish to calculate the statistical limits for the time oriented charts or need any help in understanding these tools, you may contact this author. Alternately, Purdue University's Statistics Department has a free program for non-profit and governmental organizations called, Statistics in the Community (STATCOM). STATCOM is a student-run service that provides free, professional statistical consulting to local, governmental, and nonprofit organizations. Just visit www.stat.purdue.edu/external_relations/statcom.

Regardless of the exact type of statistics or calculations performed, use your data as an opportunity to improve.

SOURCES

Bournea, C. (2006, September 14). Survey to help library chart future course. *This Week Community News*. Retrieved September 14, 2006, from http://www.thisweeknews.com

Indiana State Library Statistics Web Page – 2005 data. Retrieved April 2006 from http://digital.statelib.lib.in.us/ils/toc.cfm

The Internet Public Library. Finding Statistics. The Regents of the University of Michigan. Retrieved April 2006 from http://www.ipl.org/div/pf/entry/48530

Lynch, J. (2006, March 24). Cuts to Macomb's budget could trim library hours. *Detroit News*. Retrieved September 19, 2006, from http://detnews.com/apps/pbcs.dll/article?AID=/20060324/METRO03/603240323/-1/ARCHIVE

Tipton County Public Library Circulation Statistics, author calculations.

ABOUT THE AUTHOR

John Borneman is an engineer and "statistical practioner" for a major electronics manufacturing company in Indiana. He also has been serving as a Trustee of the Tipton County Public Library Board for the past six years. Recently, John presented a paper at the 2006 Indiana Library Federation Annual Conference titled, "Simple Statistical Tools for Evaluating Library Data." John may be contacted via email at john.borneman@gmail.com.