Q Manage Health Care
Vol. 13, No. 4, pp. 216–221

Tukey's Control Chart

Farrokh Alemi, PhD

This article presents the idea of Tukey's Control Chart, a method of analyzing data based on the concepts developed by John Tukey for calculation of confidence intervals for medians. The procedure is simple to implement (no need to calculate averages or standard deviations); it does not assume any distribution of the data; it can be applied to small data sets; and it is robust and not affected by occasional unusual observations (outliers). The article provides examples of the application of Tukey's Control Chart to both patients' lifestyle management and business process improvement.

Key words: *control chart, John Tukey, outliers, process improvement, quality improvement, small data sets, statistical process control*

There are numerous control charts available[1]—each suitable for one type of data. For a single observation per time period, a commonly used method of charting data is the Moving Range Chart, sometimes referred to as the XmR chart.[2,3] The XmR chart has the advantage that it does not assume a Normal distribution in the data. But because it is based on averages of consecutive differences, an XmR chart can be easily affected by an extreme value (eg, an outlier). Such extreme values spread the distance between the control limits and reduce the ability of the chart to identify changes in the data. When one is examining a large number of data points, the extreme observation is averaged with many other observations and the impact of the outliers on the control limits is mitigated. On the other hand, when one is dealing with a small number of data points (eg, <15), the outlier can have a large effect on the tightness of control limits. An alternative that would not be sensitive to an occasional extreme observation, even when working with a small data set, is needed.

One such alternative can be seen in the work of John Tukey—a statistician famous for his work with medians and exploratory data analysis.[4] In his classic book on robust data analysis, Tukey suggested confidence intervals for medians.[5] Since medians are not affected by outliers, they are ideal for analysis of small data sets. We used the procedures described by John Tukey for calculating confidence intervals for medians to construct a control chart. We call this type of chart "Tukey's Control Chart."

From the College of Nursing and Health Science, George Mason University, Fairfax, Va.

Corresponding author: Farrokh Alemi, PhD, College of Nursing and Health Science, George Mason University, 4400 University Dr, Fairfax, VA 22030 (e-mail: falemi@gmu.edu).

Tukey's Control Chart has several advantages. First, Tukey's Control Chart does not assume any distribution of the data. It does not assume that the observations are Normal in distribution, or that they follow any other prespecified distribution. As such, Tukey's Control Chart can be applied to any interval-based data.

Second, Tukey's Control Chart can be applied to small data sets. The more data one has, the more precision one can have in constructing the upper control limit (UCL) and the lower control limit (LCL). However, practically speaking, most improvement teams do not have access to large databases. One advantage of Tukey's Control Chart is that it can be used with small databases; in fact, one can construct a Tukey's Control Chart with as low as 7 data points in the preintervention period. This is an absolute minimum for the data needed to construct the chart and not a recommended number. The actual number of data points depends on the consequence of waiting and collecting more data versus using too little data and making an error in judgment.

Third, Tukey's Control Chart is not affected by unusual data points (eg, an outlier). In his seminal work, Tukey showed that confidence intervals based on medians are more robust and less likely to be affected by an outlier than confidence intervals based on means. Tukey's Control Chart is based on analysis of medians. In contrast, the XmR chart is based on analysis of means (in particular, the mean difference of consecutive data points). Thus, one would expect Tukey's Control Chart to be more accurate than an XmR chart—especially in small data sets (where an unusual observation can significantly affect the spread between the limits).

Finally, Tukey's Control Chart does not require standard deviations or averages to be calculated. Since medians can be calculated by examining the ordered data, Tukey's Control Chart is remarkably easy to construct, even for the people having no background in statistics. It is conceivable that a manager or a clinician listening to a presentation can construct a control chart without using a calculator or a computer.

CALCULATION OF CONTROL LIMITS FOR TUKEY'S CONTROL CHART

We will use Tukey's suggested limits for calculation of confidence intervals. The procedure calculates control limits from the difference of the Upper "Fourth" and the Lower Fourth of data, a concept that Tukey named Fourth Spread. Most readers are familiar with median, a value where half the data are below and half the data are above it. A Lower Fourth is similar to 25% quartile and is the median of the first half of the data; 25% of the data are below this value. An Upper Fourth is similar to 75% quartile and is the median of the upper half of the data; 75% of the data are below this value. The difference between the two Fourths is referred to as Fourth Spread. The UCL is the sum of the Upper Fourth and 1.5 times the Fourth

Table 1

DATA ON EXERCISE TIMES FOR A HYPOTHETICAL PATIENT

		Sorted in order of length of exercise		
Day of observation	**Minutes of exercise**	**Rank**	**Day of observation**	**Minutes of exercise**
1	30	1	2	0
2	0	2	3	25
3	25	3	1	30
4	30	4	4	30
5	32	5	5	32
6	35	6	6	35
7	50	7	7	50
8	45			
9	31			
10	20			
11	40			
12	60			
13	45			
14	60			
15	45			
16	32			
17	50			
18	60			

Spread. The LCL is Lower Fourth minus 1.5 times the Fourth Spread.

Here is the procedure for calculating Tukey's control limits:

1. List values from smallest to largest.
2. Calculate median. If the number of observations is odd, take the middle value. If the number of observations is even, take the average of the 2 middle-ranked numbers.
3. Divide the data set into 2 halves using the median. Include the median in both the halves if it is one of the observed data points.
4. Calculate Lower Fourth—the median of the lowest 50% of the data, data from the smallest number till (or including) the median.
5. Calculate Upper Fourth—the median of the top 50% of the data, data from (or including) the median of the full data set to the highest value.
6. Calculate Fourth Spread as the difference between the two Fourths.
7. Calculate UCL and LCL using the following 2 formulas:
   LCL = Lower Fourth − 1.5 × Fourth Spread
   UCL = Upper Fourth + 1.5 × Fourth Spread

APPLICATIONS OF TUKEY'S CONTROL CHART

Examples in management of patients' lifestyles

Jane collected data regarding her exercise times (Table 1). She planned to exercise 3 times a week, and each time she exercised she recorded the time in minutes. When she planned to but did not exercise, she recorded a 0 for the minutes of exercise. Data were collected for 7 days prior to and 9 days post the intervention. After this period, she and her spouse joined a mixed-group volleyball team. The question she wanted to know was whether joining the team had made a difference in her exercise time.

The first step is to sort the preintervention data in order of length of exercise. This is shown in the last column of Table 1. Next, we calculate the median; this is the value where half of the data ($7 \times 0.5 = 3.5$ ~3 points) are below it and half of the data (3 points) are above it. The median is 30, the fourth data point; 3 values are below it.

Since the median is an actual data point, we include this point in the lower data set. To calculate the Lower Fourth, we calculate the halfway point for the first half of the data. When we include the median, we have 4 points in the lower data set. The 25% quartile is halfway between the second and third data points; in other words, between 25 and 30, which is 27.5.

To calculate the Upper Fourth, we calculate the halfway point for the upper half of the data. Again, because the median is an actual data point, we include this point in the upper data set. With the median, we have 4 data points from the median to the highest value. The Upper Fourth is between the fifth and sixth data points, and therefore its value is 33.5.

The Fourth Spread is the difference between the Upper and Lower Fourths, which is $33.5 - 27.5 = 6$. The UCL is $33.5 + 1.5 \times 6 = 42.5$. The LCL is $27.5 - 1.5 \times 6 = 18.5$. A chart of the data is provided in Figure 1.

Examination of the chart shows that in the first 7 days, there was one very low point of no exercise, a statistical abnormality (perhaps an outlier), and one high point. The control limits are calculated from 7 data points in the preintervention period. On several occasions, the exercise time exceeded the UCL. On these days, there was a real increase in exercise time compared to the first 7 days. Since these days corresponded to joining the volleyball team, the intervention seems to have worked in changing the patient's lifestyle.

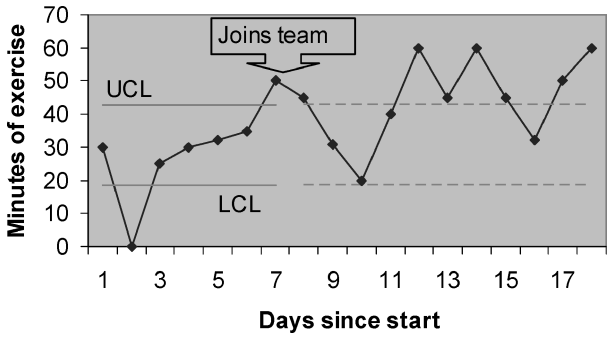

Figure 1. Tukey's Control Chart for the data shown in Table 1.

Table 2

RECORDED WEIGHT VALUES

		Sorted values	
Week	**Pounds over ideal weight**	**Rank**	**Pounds over ideal weight**
1	10	1	3
2	11	2	5
3	7	3	7
4	5	4	7
5	9	5	8
6	7	6	9
7	3	7	10
8	8	8	11
9	6		
10	6		
11	3		
12	0		
13	4		
14	−1		
15	−5		
16	−2		

Let us look at another example, this time on weight loss. A 48-year-old man measured his weight for 8 weeks. Then he and his spouse changed their food-shopping habits. They excluded all sweets from their shopping (they stopped buying pops, sweetened cereals, and chocolates for the house). The data for this person is provided in Table 2. Weight was recorded once a week.

As before, the first step is to sort the preintervention data, from least amount of pounds overweight to the highest value. This is shown in the last column of Table 2. Next, we calculate the median; this is the value where half of the data ($8 \times 0.5 = 4$ points) are below it and half of the data (4 points) are above it. The value should be between the fourth and fifth data points, or between 7 and 8, and so the median is 7.5.

Since the median is not an actual data point, we do not include this point in the calculations of Fourths. To calculate the Lower Fourth, we pick the halfway point for the first half of the data. We have 4 points in the lower data set. The Lower Fourth is halfway between the second and third points; in other words, between 5 and 7, that is, 6.

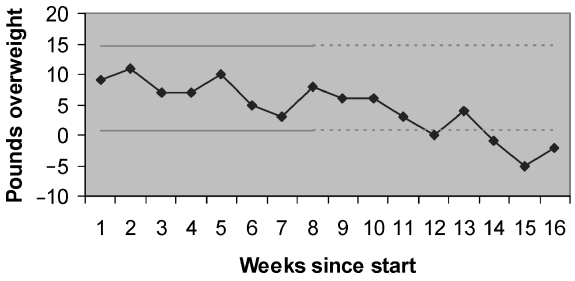

Figure 2. Tukey's Control Chart for the data shown in Table 2.

To calculate the Upper Fourth, we calculate the halfway point for the upper half of the data. Again, because the median was not an actual data point, we do not include this point in the upper data set. We have 4 data points for the highest values. The Upper Fourth is between the sixth and seventh data points (between 9 and 10), that is, 9.5.

The Fourth Spread is the difference between the Upper and Lower Fourths, which is $9.5 - 6 = 3.5$. The UCL is $9.5 + 1.5 \times 3.5 = 14.75$. The LCL is $6 - 1.5 \times 3.5 = 0.75$. A chart of the data is provided in Figure 2.

Examination of the chart shows that in the first 8 weeks, all data points were within the limit. No

Table 3

BUDGET DEVIATIONS IN 12 MONTHS

Month	**Budget deviation in 1000s**
1	23
2	−5
3	−70
4	−7
5	−8
6	9
7	12
8	30
9	24
10	25
11	−4
12	−2

Table 4

SORTED DATA FOR BUDGET DEVIATIONS

Rank	Budget deviation in 1000s
1	−70
2	−8
3	−7
4	−5
5	−4
6	−2
7	9
8	12
9	23
10	24
11	25
12	30

weight was lost in the preintervention period, even though there were considerable fluctuations. Over the remaining 8 weeks and compared to the first 8 weeks, on 4 occasions, the weight was lower than the LCL. Therefore, there was a real decrease in weight in the postintervention period.

Example in business management

Suppose we are looking at 12 month of data regarding our clinic's budget. The question is whether the expenditures for any particular month are higher than the general pattern of expenditure across the 12 months. Table 3 shows the budget deviation (expenditure − budget amount) for each of the months in thousands of dollars.

The first step is to sort the data, which is shown in the second coloumn in Table 4.

There are 12 data points, and so the median is halfway between the sixth- and seventh-ranked data points. The median is not included in the lower and upper data sets because it is not an actual value in the data. The Lower Fourth is halfway between the 6 data points with the lowest ranks, that is, between the third and the fourth data points, and has the value −6. The upper data set is the points ranked 7 through 12. Median of this data set is halfway between the 9th- and 10th-ranked data items. It is 23.5. The Fourth Spread is 29.5.

The UCL is $23.5 + 1.5 \times 29$ and the LCL is $-6 - 1.5 \times 29.5$. Figure 3 shows the control chart.

The chart shows that all months are within control limits except for March, when there was a large deviation from the budgeted amount.

For additional examples of the application of Tukey's Control Chart, please see http://cqi.gmu.edu/frTukey.asp.

DISCUSSION

This article presented Tukey's Control Chart, a method of analyzing data based on the calculation of

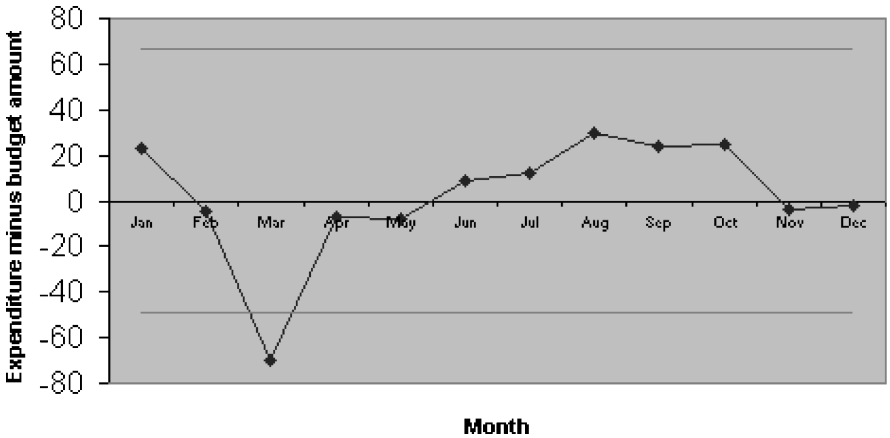

Figure 3. Tukey's Control Chart for budget deviation (in 1000s).

confidence interval for medians. The procedure has several advantages:

1. It is simple to implement (no need to calculate averages or standard deviations)
2. It does not assume any distribution of the data
3. It can be applied to small data sets
4. It is robust and not affected by occasional unusual observations (outliers)

Our personal experience with improvement efforts within the health care industry shows that most teams rely on small data sets, typically less than 15 data points. In these small data sets, Tukey's Control Chart may be a reasonable robust method of analysis.

While this article laid out the procedures for constructing Tukey's Control Chart, it did not provide independent data regarding the relative accuracy of Tukey's and XmR charts. John Tukey's own research, of course, has already established that confidence intervals for medians are robust and preferred to confidence intervals for means, especially in data streams that contain an outlier. It is logical to assume that these findings extend to control charts. Nevertheless, the proof of the relative accuracy of Tukey's Control Chart over XmR chart remains to be established in direct studies involving these charts. We encourage additional research to see if the approach proposed here is more accurate and easier to implement than that for XmR charts in typical health care improvement efforts.

REFERENCES

1. Amin SG. Control charts 101: a guide to health care applications. *Qual Manag Health Care*. 2001;9(3):1–27.
2. Shewhart W. *Statistical Method from the Viewpoint of Quality Control*. Mineola, NY: Dover Publications; 1986. ISBN: 0486652327.
3. Berwick DM. Controlling variation in health care: a consultation from Walter Shewhart. *Med Care*. 1991;29(12):1212–1225.
4. Tukey JW, Jones LV. *The Collected Works of John W. Tukey: Philosophy and Principles of Data Analysis 1949–1964*. Vol 3. New York: Chapman & Hall/CRC; 1986. ISBN: 0412742500.
5. Hoaglin DC, Mosteller F, Tukey JW, eds, *Understanding Robust and Exploratory Data Analysis*. New York: Wiley-Interscience; 2000. ISBN: 0471384917.