So You Want to Use a *p*-Chart?

Not all count-based data will qualify

Donald J. Wheeler

First, let it be known that all charts for count-based data are charts for individual values. Regardless of whether we are working with a count or a rate, we obtain one value per time period and want to plot a point every time we get a value. This need to plot the current data is why the specialty charts for count-based data were developed before a general approach for charting individual values was discovered. The question addressed in this column is when to use the specialty charts with your count-based data.

The first of these specialty charts, the *p*-chart, was created by Walter Shewhart in 1924. At that time the idea of using the two-point moving range to measure the dispersion of a set of individual values had not yet reached the professional literature. (John von Neumann would introduce the use of successive differences to the mathematical world in 1941 and W. J. Jennett would have the idea of an XmR chart in 1942.)

So the problem Shewhart faced was how to create a process behavior chart for individual values based on counts. While he could plot the data in a running record, and while he could use an average value as the central line for this running record, the obstacle was how to compute the limits. With the average and range chart Shewhart had used the within-subgroup variation, but this approach did not work with subgroups of size one. However, if the counts could be said to follow either a binomial probability model or a Poisson probability model, then the dispersion could be estimated from the average. So, lacking any alternative, Shewhart decided to use theoretical limits based on a probability model for his count-based data. And that is the origin of the specialty charts.

Data Characterized by Binomial Model Area of Opportunity = **n**		Data Characterized by Poisson Model Area of Opportunity = ***a***	
Constant **n**	Variable **n**	Constant ***a***	Variable ***a***
np-Chart for Counts	***p***-Chart for Proportions	***c***-Chart for Counts	***u***-Chart for Rates

Figure 1: Specialty Charts for Count-Based Data

Both probability models in Figure 1 impose certain homogeneity conditions upon the data. Before a count of items can be said to be a binomial count each of the **n** items in any given time period *must* have the same probability **p** of possessing the attribute. Otherwise the sum of the **n** Bernoulli counts will not be binomial with parameters **n** and **p**. Before a count of events within some finite area of space or time or product, **a**, can be said to be a Poisson count, the events must logically be independent of each other and the events must be uniformly spread throughout the area of opportunity.

These models are fully specified by their mean value. This allows us to use the average to characterize both location and dispersion. Thus, *p*-charts, *np*-charts, *c*-charts, and *u*-charts all have limits that are based upon a theoretical relationship between the mean of a probability model and its dispersion. Hence, these specialty charts all use theoretical limits. If the counts can be reasonably modeled by either a binomial distribution or a Poisson distribution, then one of these specialty charts will provide appropriate limits for the data.

Over the years many textbooks and standards have forgotten that the assumption of a binomial model or a Poisson model is a prerequisite for the use of these specialty charts. This is a problem because there are many types of count-based data that cannot be characterized by either a binomial or a Poisson distribution. When such data are placed on a *p*-chart, *np*-chart, *c*-chart or *u*-chart the theoretical limits obtained will be wrong.

So what are we to do? The problem with the theoretical limits lies in the assumption that we know the exact relationship between the central line and the three-sigma distance. The solution is to obtain a separate estimate of dispersion, which is what the *XmR* Chart does: While the average will characterize the location and serve as the central line for the *X* Chart, the average moving range will characterize the actual dispersion in the data and serve as the basis for computing the three-sigma distance for the *X* Chart.

Thus, the major difference between the specialty charts and the *XmR* Chart is the way in which the three-sigma distance is computed. The *p*-chart, *np*-chart, *c*-chart, and *u*-chart will have the same running record, and essentially the same central lines, as the *X* Chart. But when it comes to computing the three-sigma limits the specialty charts use an *assumed* theoretical relationship to compute theoretical values while the *XmR* Chart actually measures the variation present in the data and constructs *empirical* limits.

To compare the specialty charts with the *XmR* Chart we shall use three examples. The first of these will use the data of Figure 2. These values come from an accounting department which keeps track of how many of their monthly closings of departmental accounts are finished "on-time." The counts shown are the monthly numbers of closings, out of 35 closings, that are completed on time. The limits are based on Years One and Two.

The Number of On-Time Closings for 35 Departments

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
One	32	30	32	33	32	28	30	31	32	32	32	33
Two	29	31	32	33	31	31	34	30	33	28	33	34
Three	33	33	31	29	33	30	26					

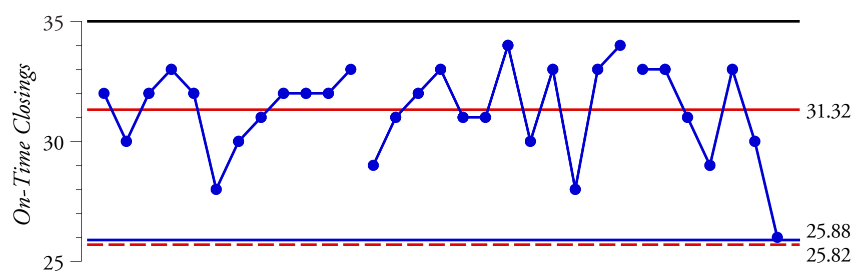

Figure 2: The *X* Chart and *np*-Chart for the On-Time Closing Data

Here both the *np*-chart and the *X* Chart computations give essentially the same limits. (The upper limit value of 36.8 is not shown since it exceeds the maximum value of 35 on-time closings.) Here the two approaches are essentially identical because these counts seem to be appropriately modeled by a Binomial distribution. If you are sophisticated enough to determine when this happens, then you will know when the *np*-chart will work and can use it successfully. On the other hand, if you do not know when a Binomial model is appropriate, then you can still use an *XmR* Chart. As may be seen here, when the *np*-chart would have worked, the empirical limits of the *X* Chart will mimic the theoretical limits of the *np*-chart, and you will not have lost anything by using the *XmR* Chart instead of the *np*-chart.

Our next example will use the on-time shipments for a plant. The data are shown in Figure 3 along with both the *X* Chart and the *p*-chart for these data. The limits are based on all 24 values.

The Proportion of On-Time Shipments

Month	Year	Total No.	No. On-Time	On-Time %	Month	Year	Total No.	No. On-Time	On-Time %
January	01	191	176	92.1	January	02	170	155	91.2
February	01	203	186	91.6	February	02	270	246	91.1
March	01	220	202	91.8	March	02	167	151	90.4
April	01	200	183	91.5	April	02	216	196	90.7
May	01	236	215	91.1	May	02	227	206	90.7
June	01	213	194	91.1	June	02	149	136	91.3
July	01	212	191	90.1	July	02	182	167	91.8
August	01	241	215	89.2	August	02	224	206	92.0
September	01	159	143	89.9	September	02	246	225	91.5
October	01	217	197	90.8	October	02	185	170	91.9
November	01	181	165	91.2	November	02	261	239	91.6
December	01	113	103	91.2	December	02	140	128	91.4

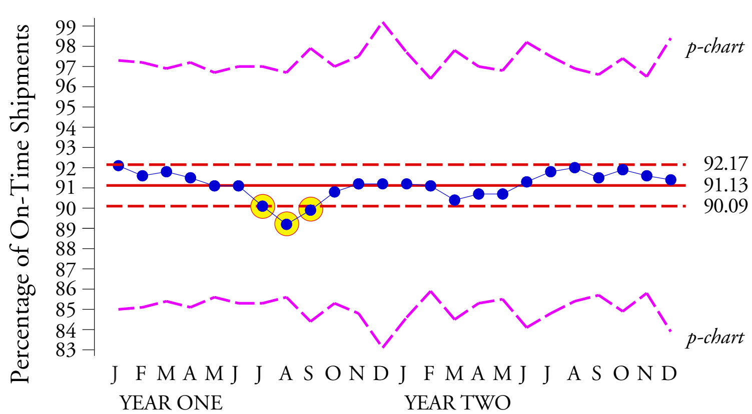

Figure 3: The *X* Chart and *p*-Chart for the On-Time Shipments

The *X* Chart shows a process with three points at or below the lower limit. The variable-width *p*-chart limits are five times wider than the limits found using the moving ranges. No points fall outside these limits. This discrepancy between the two sets of limits is an indication that the data of Figure 3 do not satisfy the Binomial conditions. (Even if we had nothing but the binomial limits we would know these data are not binomial data due to the way the running record hugs the central line.)

Here the probability of a shipment being on time is *not the same* for all of the shipments in any given month. This invalidates the Binomial model and makes the theoretical *p*-chart limits incorrect. However, the empirical limits of the *XmR* Chart, which do not depend upon the appropriateness of a particular probability model, are correct.

Our final comparison will use the data of Figure 4. There we have the percentage of incoming shipments for one electronics assembly plant that were shipped using air freight. The limits are based on all 8 values. Two points fall outside the variable width *p*-chart limits while no points fall outside the *X* Chart limits.

The Premium Freight Data

Month	Year	Total No. Shipments	No. Shipped Air Freight	Percentage Air Freight
May	01	6144	374	6.09
June	01	3792	227	5.99
July	01	4792	278	5.80
August	01	7226	346	4.79
September	01	4440	161	3.63
October	01	4896	232	4.74
November	01	6019	352	5.85
December	01	4101	277	6.75

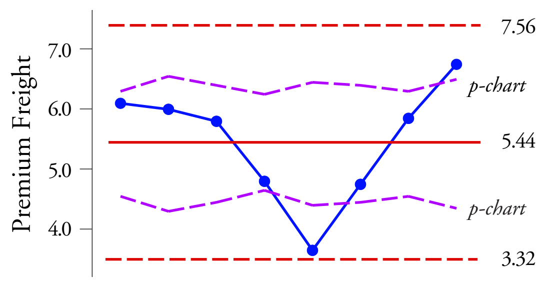

Figure 4: The X Chart and p-Chart for the Premium Freight Data

Figure 4 is typical of what happens when the area of opportunity for a count of items gets excessively large. The Binomial model *requires* that all of the items in any given time period will have the same chance of possessing the attribute being counted. Here this requirement is not satisfied. With thousands of shipments each month, the probability of a shipment being shipped by air is not the same for all of the shipments. Thus, the Binomial model is inappropriate, and the theoretical p-chart limits which depend upon the Binomial model are incorrect. The X Chart limits, which here are twice as wide as the p-chart limits, properly characterize both the location and dispersion of these data and are the correct limits to use.

Thus, the difficulty with using a p-chart, np-chart, c-chart, or u-chart is the difficulty of determining whether the Binomial or Poisson models are appropriate for the data. As seen in Figures 3 and 4, if you overlook the prerequisites for a specialty chart you will risk making a serious mistake in practice. This is why you should avoid using the specialty charts if you do not know how to evaluate the appropriateness of these probability models.

In contrast to this use of theoretical models which may or may not be correct, the *XmR* Chart provides us with empirical limits that are actually based upon the variation present in the data. This means that you can use an *XmR* Chart with count based data anytime you wish. Since the *p*-chart, the *np*-chart, the *c*-chart, and the *u*-chart are all special cases of the chart for individual values, the *XmR* chart will mimic these specialty charts when they are appropriate and will differ from them when they are wrong.

When the specialty charts with variable-width limits are appropriate, the fixed-width limits of an *XmR* Chart will approximate limits based on the average-sized area of opportunity.

All Count-Based Data	
Areas of Opportunity Relatively Constant	Areas of Opportunity Vary Considerably
XmR Chart for Counts	***XmR*** Chart for Rates

Figure 5: An Assumption-Free Approach for Count-Based Data

Thus, when you are confident that your counts within each time period satisfy the requirements for either a binomial probability model or a Poisson probability model, you may safely use an *np*-chart, *p*-chart, *c*-chart, or *u*-chart. If the theory is appropriate, the theoretical limits will be correct. If you are wrong about the theoretical model for your count-based data, then the theoretical limits will be incorrect.

Of course, you can guarantee that you have appropriate limits for your count-based data by simply using the *XmR* chart to begin with. The empirical approach will always be right.