CHAPTER 3: SIMPLE EXTENSIONS OF CONTROL CHARTS

1. Introduction: Counting Data and Other Complications

The simple ideas of statistical control, normal distribution, control charts, and data analysis introduced in Chapter 2 give a good start on learning to analyze data -- and hence ultimately to improve quality. But as you try to apply these ideas, you will sometimes encounter challenging complications that you won't quite know how to deal with. In Chapter 3, we consider several common complications and show how slight modifications or extensions of the ideas of Chapter 2 can be helpful in dealing with them. As usual, we shall rely heavily on examples.

So long as the data approximately conform to randomness and normality, as in all but one example of Chapter 2, we have a system both for data analysis and quality improvement. To see whether the data do conform, we have these tools:

- The run chart gives a quick visual indication of whether or not the process is in statistical control.
- The runs count gives a numerical check on the visual impression from the run chart.
- If the process is in control, we can check to see whether the histogram appears to be roughly what we would expect from an underlying normal distribution.
- If we decide that the process is in control and approximately normally distributed, then a control chart, histogram, and summary statistical measure tell us what we need from the data:
- Upper and lower control limits on the range of variation of future observations.
- If future observations fall outside those control limits -- are "outliers" -- we can search for possible special causes impacting the process.
- The histogram and the sample mean and standard deviation give an idea of the process capability.

With these tools alone, we can deal with many applications, as illustrated by the examples of Chapter 2. But these tools are not a complete system of data analysis that can be applied without modification in all applications:

- The data may appear to be in a state of statistical control -- no sequential patterns -- but the normal distribution is not a good "model" for the histogram: the data may not appear to be compatible with an underlying normal curve that would emerge if we could get a lot more data from the same process.

- The data may **not** be in a state of statistical control. The **run chart** (control chart minus control limits) may then be a useful guide to our thinking about root causes, but the control limits -- computed under the assumption that the process **is** in control -- can be misleading. For example, they may suggest that the process is in control because all points are within the computed control limits, but visual examination and/or the runs count may show clearly that the process is **out** of control.

Statistical Control but Nonnormal Histogram

In Chapter 2 where the data take the form of numerical measurements -- for example, deviations from target or a dimensional reading -- a first choice of "model" for all applications was the normal distribution. That is, the sample histogram, though ragged, suggested that a smooth, "bell-shaped" curve would emerge if we could get a lot more data from the process.

To deal with applications where the process is in control but the normal distribution is not a good model for the histogram, we shall consider these tools:

- For data that reflect the results of counting -- such as counts of defects on a personal quality checklist -- a first choice of possible model is not the normal distribution but the **Poisson distribution** (approximate pronunciation, "pwa-sohn"). A special control chart that is appropriate to the Poisson distribution is a **c-chart**, implemented by an option under the ***SPSS*** sequence **Graphs/Control…**. (We will explain and illustrate the exact commands in this introductory section.)

- For measurements of cycle times in repetitions of a process, the normal distribution may sometimes be useful but often a better choice can be based on the **exponential distribution**. ***SPSS*** does not have a special control chart for the exponential distribution, but we will improvise one: by a preliminary data transformation we obtain approximate normality, so we then can use the same control chart that was introduced in the previous chapter. (We will present this approach in Section 2.)

- For percentage data -- for example, percent of defects to total trials -- a promising first choice of model is the **binomial distribution**. A special control chart that is appropriate to the binomial distribution is a **p-chart**, also implemented via the ***SPSS*** sequence **Graphs/Control…** (We will present this approach in Section 3.)

Not in Statistical Control

A more challenging problem arises when the data are not even in a state of statistical control to begin with. For example, there may be a trend: the process is getting systematically better or worse through time. If the departures from statistical control are only slight, analyses based on control charts will still give a guide. If the departures are substantial, however, we need statistical tools that go beyond simple control charts. In Sections 4 through 6 of this chapter we consider simple examples that illustrate:

- Section 4: "**random-walk**" behavior, with stock market example.

- Section 5: **trend**.

- Section 6: **periodic effects**, such as **seasonality**.

On a first reading of this chapter, you could skip directly to Section 4 at this point, which begins the treatment of out-of-control data.

Counting Data

For the first of the new tools, we consider data arising from **counts**, as opposed to **measurements**. Several applications in Chapter 2 entailed measurements of elapsed time. With sufficient accuracy in our measuring instrument -- a digital watch -- and in our reading of it, we could in principle obtain readings to any desired number of decimal places. Time measurements are especially important in quality management because process cycle times are crucial, as we mentioned in the first section of Chapter 2.

But other measurements are also important. For example, there are dimensional measurements such as diameters of manufactured parts (recall the bushing diameters in Chapter 2); weights; volumes; temperatures; chemical concentrations; etc.

Counts, by contrast, are necessarily integers, that is, "whole numbers". In a production run of manufactured parts, for example, you could have 3 defects on a given day, or zero defects, or 4 defects, but never 2.37 or 0.21. Counts -- especially counts of defects or errors -- are important in quality management. Here are typical quality applications of counting data:

- Customer arrivals in a queue by five minute intervals.

- Completed assemblies by shift.

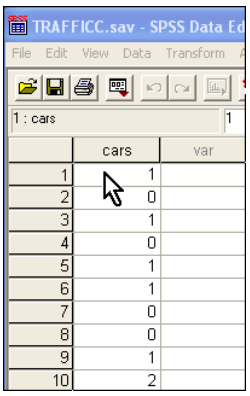
- Accidents per month.

- Customer complaints by week.

For our first illustration, however, we shall use a lightning data set of the kind that anyone can easily collect in a short time: vehicular traffic counts. Below we show a simple ***SPSS*** analysis of traffic counts. As usual, you should follow the output closely, hands on, to be sure that you understand all steps. Here is the observer's brief note about the data contained in the file TRAFFICC.sav:

> Traffic study of vehicles traveling west on US 64 from the West
> edge of Gilman, Wisconsin, 1 September 1993, at about 9:30 AM.
> Variable named **cars** is number of vehicles in 36 successive one minute intervals

Clearly the variable of interest, **cars**, is a count, similar to the examples given above. The first few rows of the **Data Editor** look like this:



	cars
1	1
2	0
3	1
4	0
5	1
6	1
7	0
8	0
9	1
10	2

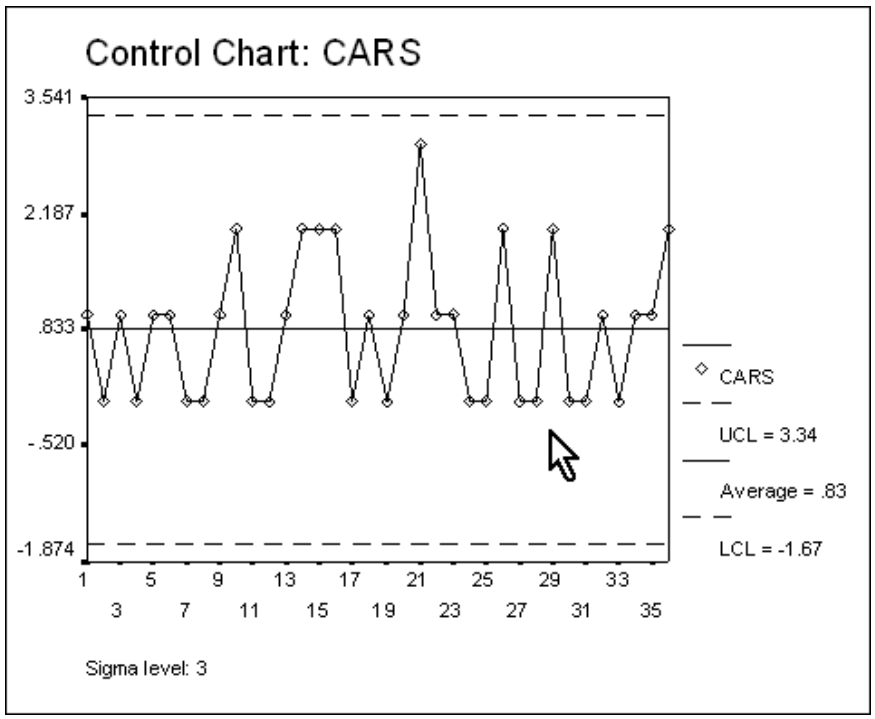
Here are the descriptive statistics for **cars**:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
CARS	36	0	3	.83	.845
Valid N (listwise)	36				

We see that the range of values is quite limited—from 0 through 3.

Next, we execute **Graphs/ Control…**, choosing the option **Individuals, Moving Range**, as we have been doing before.



Visual analysis suggests nothing out-of-control in the sequential behavior of the data points, and this is confirmed by the runs test:

Runs Test

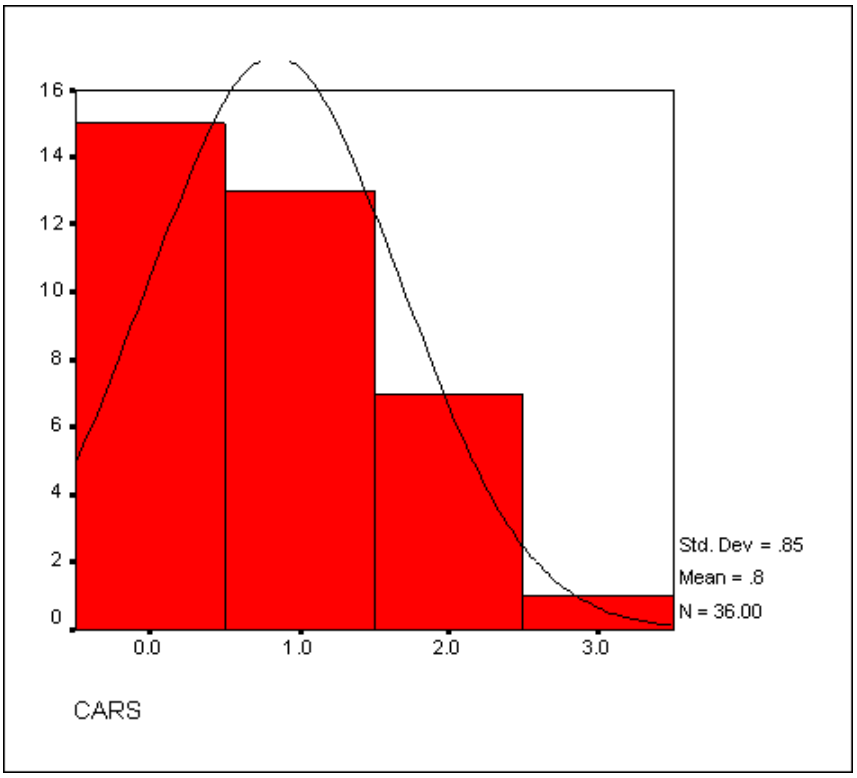
	CARS
Test Value[a]	.83
Cases < Test Value	15
Cases >= Test Value	21
Total Cases	36
Number of Runs	21
Z	.696
Asymp. Sig. (2-tailed)	.486

a. Mean

Although the sequential behavior of the data is compatible with the assumption of statistical control, you can see even without a histogram or box-and-whisker plot that the points are not symmetrical about the mean of 0.83. There is a single row of points at 0, just below the mean, then a lot of white space extending down to the LCL (Lower control limit), which is -1.67, an impossible value for **cars**.

Further, more than half the points are above the mean, and they extend almost up to the UCL (Upper control limit). Neither the LCL nor the UCL appears to be properly placed: the control limits seem unrealistic relative to the pattern of the data points.

This happens because the control limits are based on the model of the normal distribution, and the normal distribution is **not a good model** for these data, as we see when we apply **Graphs/ Histogram…** to **cars**:

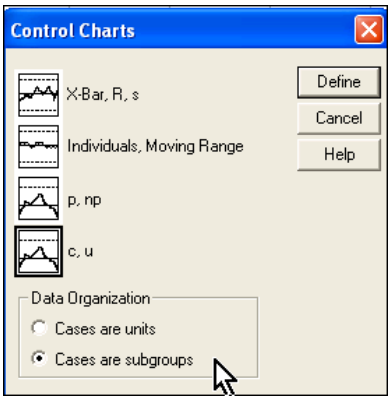


The time sequence of the points may be consistent with statistical control, but the histogram is not consistent with the model of a normal distribution: it is clearly "skewed" to the right, not symmetrical, with its right tail longer than the left. Also, the data are clearly discrete, that is, only whole numbers 0, 1, 2, and 3 are observed.

As we have noted in Chapter 2 and in the first part of the current section, the normal distribution is often a good approximate model for data, as in most of the applications of Chapter 2. As anticipated above, however, for data consisting of **counts**, another distribution is often appropriate: **the Poisson distribution**. If so, a control chart called "**c-chart**" plays the same role as does the chart that we have been using for approximately normal data.

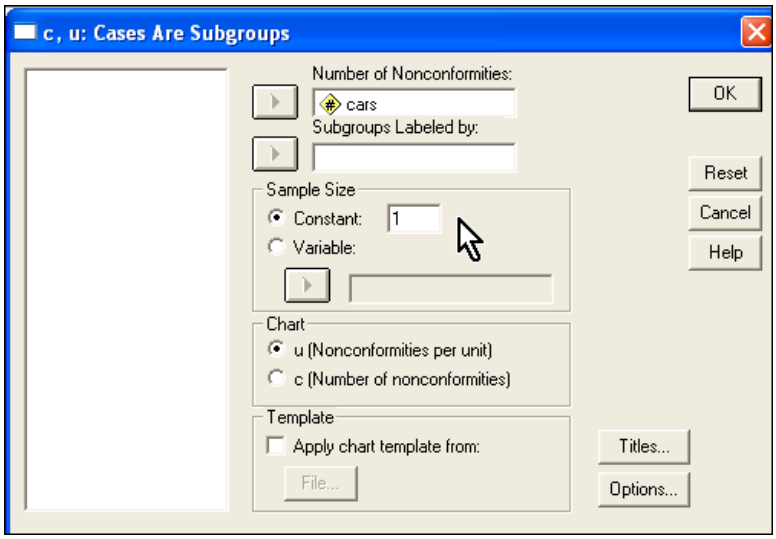
We say that the Poisson distribution is often "a good model" for counting data.[1] That is, histograms of actual data tend to resemble, aside from the raggedness to be expected in small samples, the theoretical pattern specified by the Poisson distribution.

We'll show how **c-chart** works, then discuss more about the Poisson distribution and how it can be interpreted. The procedure may seem a little confusing at first because the terminology used in the ***SPSS*** tools for this purpose follows the traditional usage in the special field called **"Statistical Process Control".** At the first step we execute **Graphs/ Control…**



Then in the window at left that appears we highlight the choice labeled **c,u** and we mark the little circle for **Cases are subgroups.**

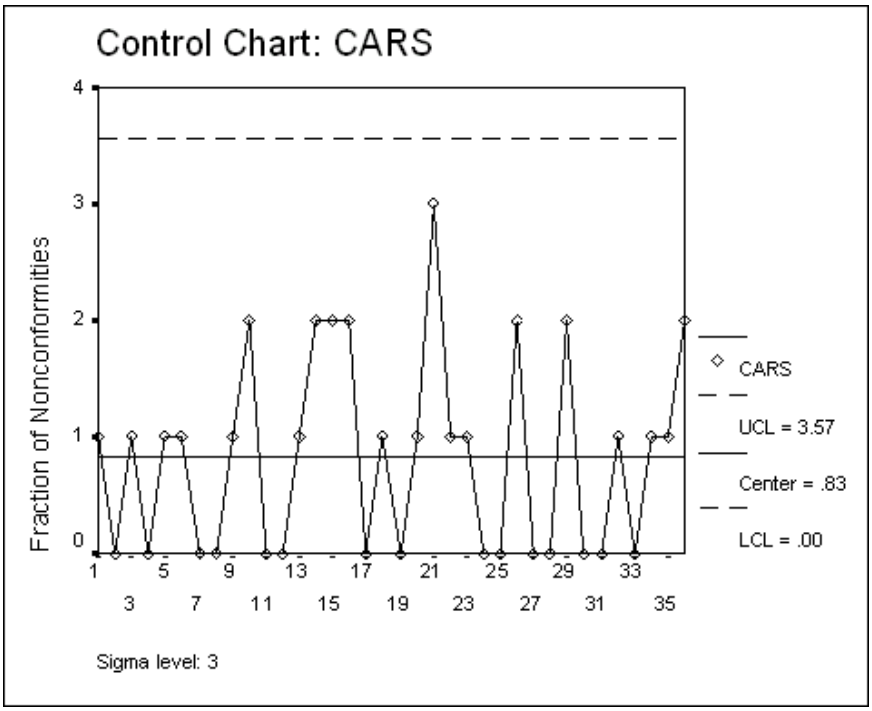
Here is the window that opens when we click on the **Define** button:



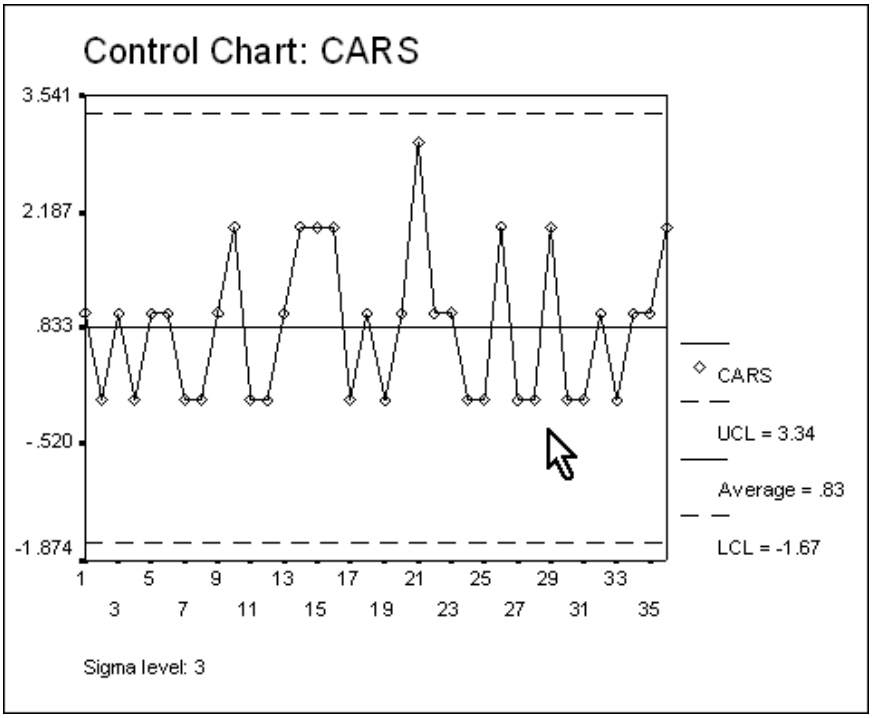
After moving **cars** into the top box we indicate under **Chart** that we want to plot **u(Nonconformities per unit)** and we enter a constant sample size of **1** (representing the one-minute interval of observation).

This is the plot that results after clicking **OK**:

[1] If the data are **not** the results of counting, put the Poisson distribution out of your mind when thinking of possible models!



The points are exactly the same as on the first control chart, but the control limits are placed differently. For your convenience in making the comparison of control limits, we display previous chart again:



The center line is at the mean 0.83 on both charts, but compare the lower control limits:

- The c-chart shows an LCL of zero, the minimum possible count, which is **less** than three standard deviations below the mean. (In practice, in this example, with a mean count less than 1, even a zero count would not arouse substantial suspicion of a special cause.)

- On the earlier chart, by contrast, the LCL is three standard deviations below the mean, -1.67; yet a count can't be less than zero.

- Similarly, the c-chart gives 3.57 for the upper control limit, which is **more** than three standard deviations above the mean. Thus, in practice, a count of 4 or more would be needed to arouse substantial suspicion of a special cause, and there are no counts that large in this data set.

The difference between the two sets of control limits is that the chart that we plotted first assumes a normal distribution as a model for the data, while the c-chart assumes a Poisson distribution as a model.

If you compare the two charts, you will see that the Poisson assumption is more realistic for this data set. It implies, in particular, that if the mean count is less than 1, as here, the sample histogram will be skewed with the "long tail" pointing to the right, that is, toward the larger values; the "left tail" will be blunt.

There is no law of statistics that says that the Poisson distribution will give a good description of the histogram for counting data, just as there is no law that says that the normal distribution will give a good description of the histogram for measurement data. Both, however, are good first thoughts for appropriate models in appropriate circumstances: the normal for measurement data, the Poisson for counting data.

In doing data analysis, however, we have to decide whether the actual data of a particular application do or do not conform approximately to what would be expected under reasonable models. If the time sequence of the data suggests a state of statistical control and **if we are dealing with counting data**, an easy way to do this is to use the **c-chart** to see whether the Poisson control limits appear to be realistically placed relative to the data points.

Another check is provided by the following. If the Poisson assumption is appropriate, the sample mean should be close to the square of the sample standard deviation (the square of the standard deviation is called the variance, not to be confused with an accounting "variance"). Here Mean = 0.83 and Std. Deviation = 0.845 and Std. Deviation squared = 0.7144, so we are at least in the right ball park.

Conditions Leading to the Poisson Distribution

What kind of process is described by the Poisson distribution? The intuitive idea is this: if occurrences of a particular phenomenon, such as a vehicle passing a particular checkpoint, are in statistical control and occur "at random through time", then the histogram of traffic counts for successive intervals of time will be approximately described by a Poisson distribution. In the present application, and in many others (including counts of errors and defects) this assumption often works surprisingly well. One interesting application is the occurrence of accidents.

We have observed that there is no law that any particular set of data **must** conform to the Poisson distribution. For example, if we counted vehicles per second on a congested expressway in a big

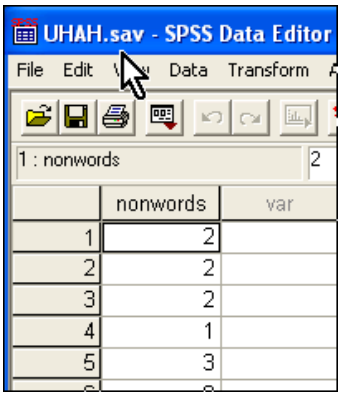
city at rush hour, we would **not** expect either that the process would be in statistical control or that the Poisson distribution would describe the histogram. Traffic jams do not lead to "random occurrences through time": your arrival at a particular point depends on the car ahead of you!

Whatever we may anticipate in advance, therefore, we must always look at the data to see what is going on in the current application!

When, in looking at the data, we see that simple assumptions fail, we will have to learn what further steps are necessary in the data analysis. Most of the balance of treatment of data analysis in this book, **STM**, is an elaboration of these further steps.

Another Application: "Non-Words" in Public Speaking

The following lightning data set provides another example of data for which the Poisson model is often roughly realistic. It also suggests possibilities for personal quality improvement projects aimed at improvement of oral presentation skills. Most of us use "non-words" such as "uh" and "ah" rather frequently in our everyday speech, so much so that we hardly notice it. However, non-words are regarded by speech experts as undesirable, because they make the oral delivery choppy and disjointed. Toastmasters' Clubs actually count the frequency of non-words in talks given by their members, who are urged to work hard to reduce their frequency. A visit to a Toastmasters' breakfast meeting, where members practice giving short talks on a variety of subjects, led to the idea for the following data set from a file called UHAH.sav:



The variable called **nonwords** is the count of the use of expressions such as “uh”, “ah”, “er”, etc. in successive 30 second intervals during a talk on quality management at a Toastmasters’ meeting in 1994. The file consists of 28 cases in all.

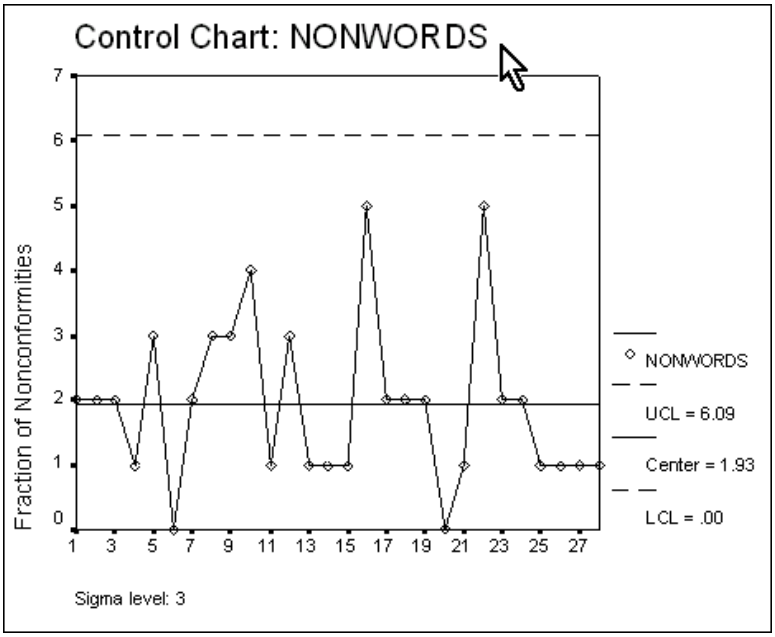
Here are the desciptive statistics:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
NONWORDS	28	0	5	1.93	1.274
Valid N (listwise)	28				

Note that the square of the standard deviation, 1.274, is 1.6244, roughly comparable to the mean, 1.93, and thus roughly consistent with the assumption of a Poisson distribution.

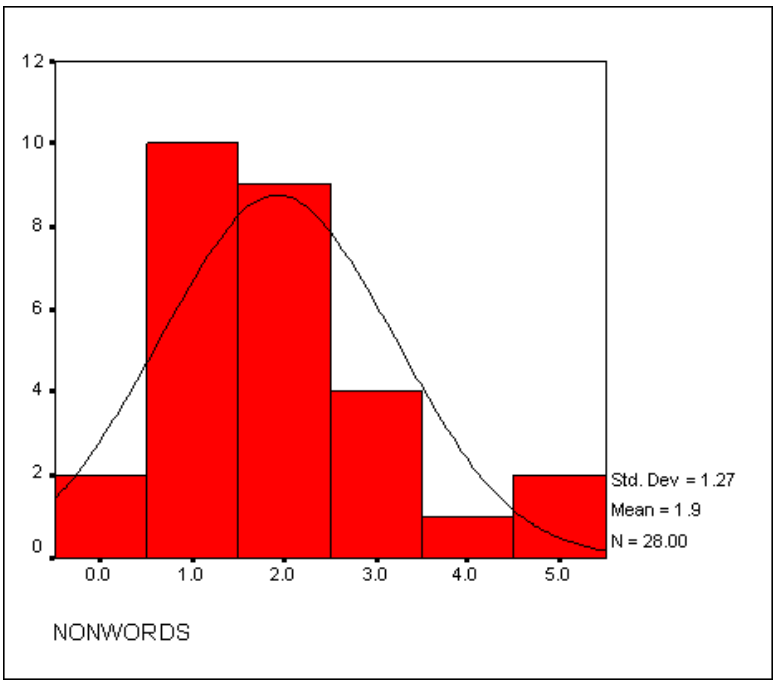
Next we do a c-chart, a runs test, and examine the histogram for **nonwords**:



Runs Test

	NONWORDS
Test Value[a]	1.93
Cases < Test Value	12
Cases >= Test Value	16
Total Cases	28
Number of Runs	12
Z	-.871
Asymp. Sig. (2-tailed)	.384

a. Mean



The data points in the c-chart are well below the upper control limit, and the runs test shows the number of runs to be close to the expected value. As for the histogram above, the normal curve fits a little better than in the case of TRAFFICC.sav, but the distribution is still skewed to the right, with a large buildup to the left of the peak of the normal curve, and a smaller, but noticeable bar in the extreme right tail. (Try the normal probability plot yourself to confirm that the fit is not good.)

We must admit that the skewness is less pronounced than for the traffic counts. The mean of **nonwords** is 1.93, whereas the mean for the traffic count, **cars,** was only 0.83. Thus the higher mean count is associated with a less skewed histogram. This particular comparison illustrates a general tendency: the larger the mean of a Poisson distribution, the less the skewness and the more that the Poisson resembles the normal distribution. It follows that if the mean is large, the normal distribution can be used to approximate the Poisson, and we can use either the individuals chart or the c-chart equally well to check on randomness. In the present case, however, the mean is not large enough for the normal fit to be accepted.

Data on non-words are very easy to collect, and they lend themselves not only to self-improvement studies but to all kinds of interesting statistical comparisons. For example, the frequency of use of non-words may depend on the degree of preparation of the talk, the possibility of interruption, or the nervousness of the speaker. Also, there is great variation from one person to another in the frequency of use of non-words in everyday conversation.

2. Data Arising as Time Intervals: the Exponential Distribution

This section deals with a very common complication that arises when time intervals between occurrences (for example, cycle times -- the time required for one unit to pass through a process) are not well approximated by a normal distribution, so that a control chart for individuals is not very helpful. Coping with this complication entails a subtlety that some readers may wish to bypass on first reading; **they can safely skip to Section 3.**

When events occur successively through time, we can view the same process from two perspectives:

- occurrences during successive time periods, as in traffic counts per minute in Section 1;
- the time intervals between occurrences.

Thus for accident data, we can tabulate accidents per month or we can count working days between accidents.

This idea can be illustrated by reexamination of the traffic data set that we studied in Section 1. We not only counted vehicles per minute for 36 minutes, but we timed the intervals between the passage of the same 30 vehicles. (Actually, we did the latter and reconstructed the former.) The interval data are contained in a separate file named TRAFFICI.sav. Here we display the observer's remarks without showing the actual data:

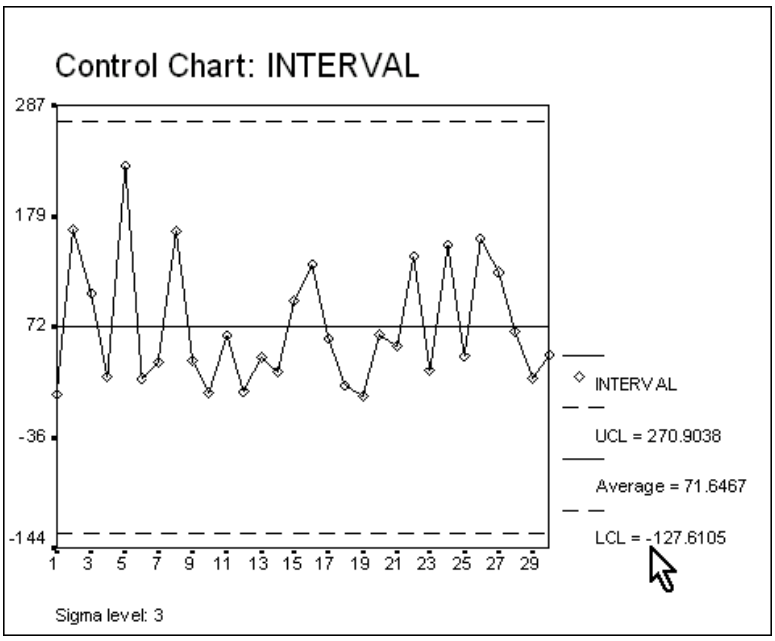
> Traffic study of vehicles traveling west on US 64 from the west
> edge of Gilman, Wisconsin, 1 September 1993, at about 9:30 AM.
> **interval**: time in seconds since last vehicle passed fixed point.
> timing was started when the first vehicle passed the point; the
> recorded 30 numbers therefore reflect times for the subsequent
> 30 vehicles after the first. The 30 vehicles reflected here
> the same as those counted by one-minute intervals in TRAFFICC.sav
> Mean 71.65 total time 35:49.40
> **interval** refers to the time intervals between the thirty
> vehicles, while **cars** refers to the number of vehicles in each
> of the 36 minutes (the count adds to 30). (There were no
> additional vehicles in time 35:49.40 to 36:00.)

The following descriptive statistics are obtained for the new variable, **interval.** (This time we are careful to check the little box labeled **Save standardized values as variables**.) :

Descriptive Statistics

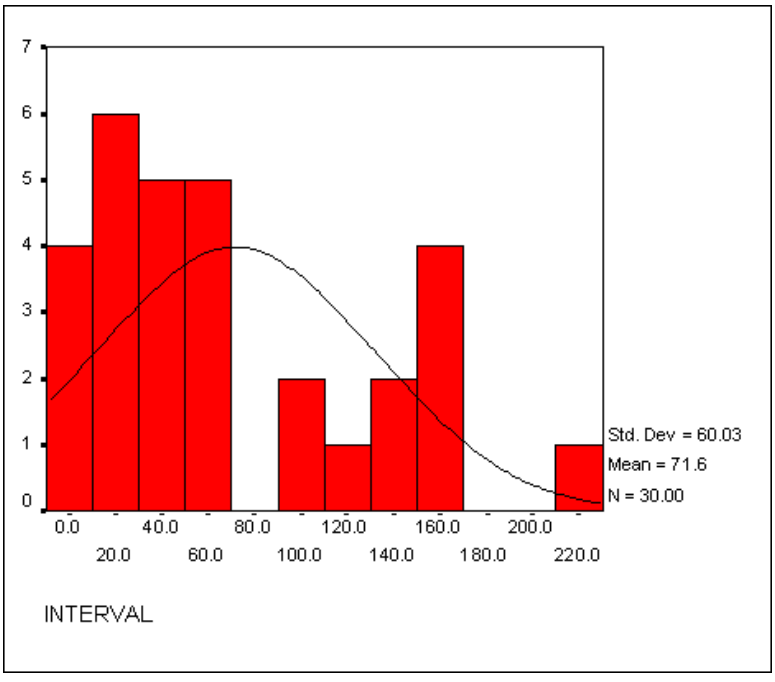
	N	Minimum	Maximum	Mean	Std. Deviation
INTERVAL	30	5.21	227.09	71.6467	60.02984
Valid N (listwise)	30				

Next we plot the data on a control chart for individuals:



There are no indications of unusual values that might indicate "special causes", but we can see that, as in the traffic counts, there is a greater range of values above the mean than below. (The LCL is negative--an impossible value for time intervals!) We shall not show the runs test. (You should be sure to do it yourselves.) It shows no indication that the process is out of control.

The histogram clearly indicates that the normal probability model is not appropriate for these data:



In dealing realistically with this application, we can draw on theory. In the same circumstances for which the Poisson distribution is a good model for a histogram of counts of occurrences during each

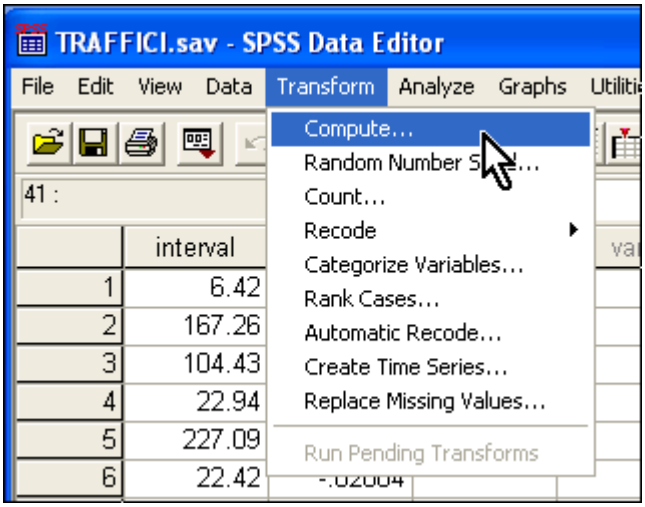
of a series of time intervals, a distribution called the **exponential distribution** is a good model for the histogram of the times elapsed between successive occurrences.

The exponential distribution is strongly skewed to the right, suggesting frequent short times between occurrences but a scattering of quite long times, which is what we're seeing here.

SPSS does not offer a special control chart for the exponential distribution, but we can adapt the **individuals chart** to this purpose by a special computation called a **data transformation**. In particular, for purposes of data analysis we can look at, not the times themselves as above, but the **cube root of the times**.

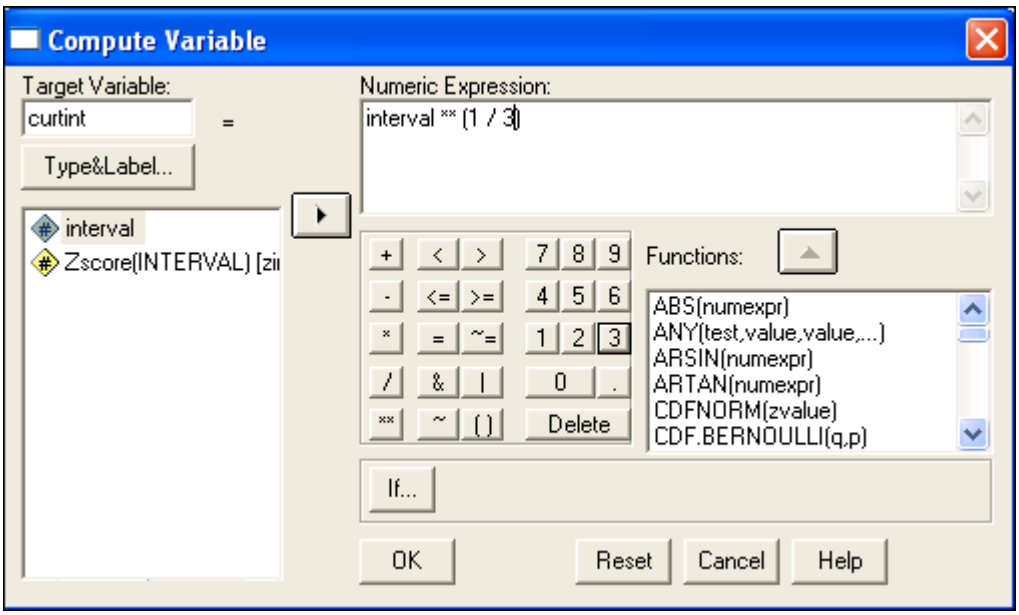
Very roughly, **if the exponential assumption holds, the transformed data will be approximately normal**. Hence instead of making a new kind of chart, we can make a chart that is already available serve the need.[2] The transformation itself is simply an interim convenience to facilitate our examination of the data, since the transformed data can be viewed from the familiar perspective of the normal distribution. (In any application of our results, we can always go back to the original times in seconds if we desire to do so.)

In the **SPSS** data editor, on the menu bar just to the left of **Analyze**, there is a choice named **Transform**. After clicking there we follow by clicking on the first choice in the submenu, **Compute…**



This action, in turn, opens the following dialog window:

[2]Another way to accomplish the same result would be to use the original time intervals, without transformation, then to set the control limits as follows: LCL = 0 and UCL equals six times the standard deviation, in this application UCL = 6*60.0 = 360.0. If you superimpose these limits on the I-Chart above, you will find them realistic. In particular, no points are close to UCL = 360, and there is no hint of special causes.



When the window first appears the white boxes under **Target Variable:** and **Numeric Expression:** are blank. Our aim is to transform the variable **interval** to a new variable that is the **cube root of interval**. Thus we arbitrarily type in the name **curtint** (abbreviation for "cube root of interval") in the **Target Variable:** box. The numeric expression shown above was created by highlighting **interval**, clicking on the insertion button, and then using the mouse to insert the necessary symbols from the keyboard; but it also could have been typed in directly. Thus, we are ready to perform the transformation

curtint= interval(1/3)**

Clicking on **OK** inserts the new variable into **Data Editor**.[3]

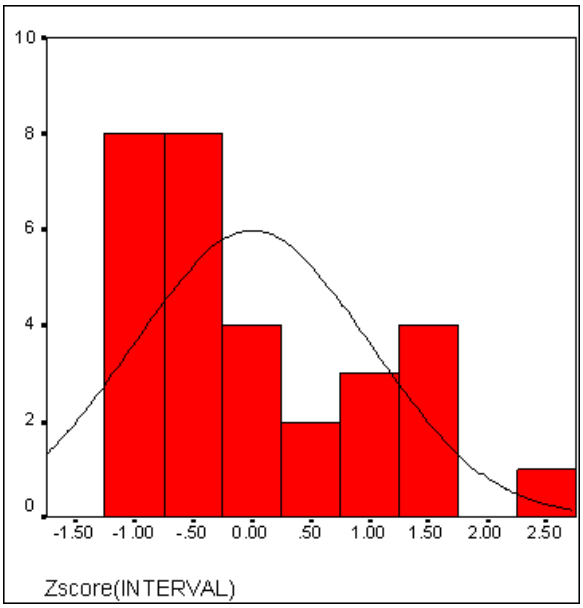
Descriptive Statistics

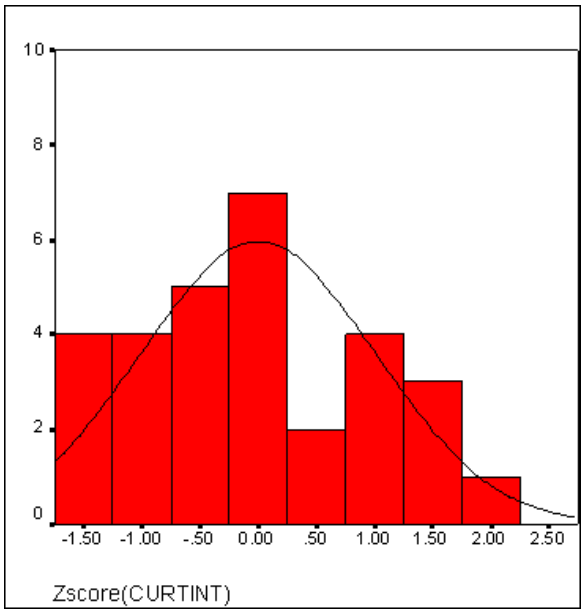
	N	Minimum	Maximum	Mean	Std.	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
INTERVAL	30	5.21	227.09	71.6467	60.02984	.922	.427
CURTINT	30	1.73	6.10	3.8062	1.21806	.078	.427
Valid N (listwise)	30						

In the display above we compare the descriptives for **interval** and its cube root, **curtint**. Clearly the new variable is more compactly distributed, i.e., its standard deviation is smaller relative to its mean value, but that may not tell us much since the scale is drastically changed. We are more interested in whether the new variable looks more normally distributed, so this time we chose as an option the **skewness** statistic. We have a rough rule of thumb that says that any value of **skewness** that is outside of the range **plus or**

[3] The "**" symbol actually means "raise to the power". Recall that taking the cube root of a number is the same as raising it to the power 1/3.

minus 0.5 indicates non-normality. The values in the table indicate that **curtint** may pass the test. We also made sure that **curtint** was standardized so that we can compare the histograms for the two sets of **z-values** using the same intervals on the axes, as shown below:





Although not perfect (after all, the sample size is only 36), the distribution for **curtint** does look more normal.

There is another rule of thumb for checking for the appropriateness of the exponential distribution. Recall that if the Poisson distribution is applicable to counting data, the sample mean should be approximately equal to the square of the sample standard deviation, that is, the variance. **For the exponential distribution, there is a similar numerical guideline: the sample mean should be approximately equal to the standard deviation (not to the square of the standard deviation, which applies to the Poisson)**.

This rule of thumb is satisfied in the current application: the mean of the original time intervals is 71.65 and the standard deviation is 60.03. (The rule of thumb "mean equals standard deviation" applies only to the original time intervals, not to the transformed time intervals.)

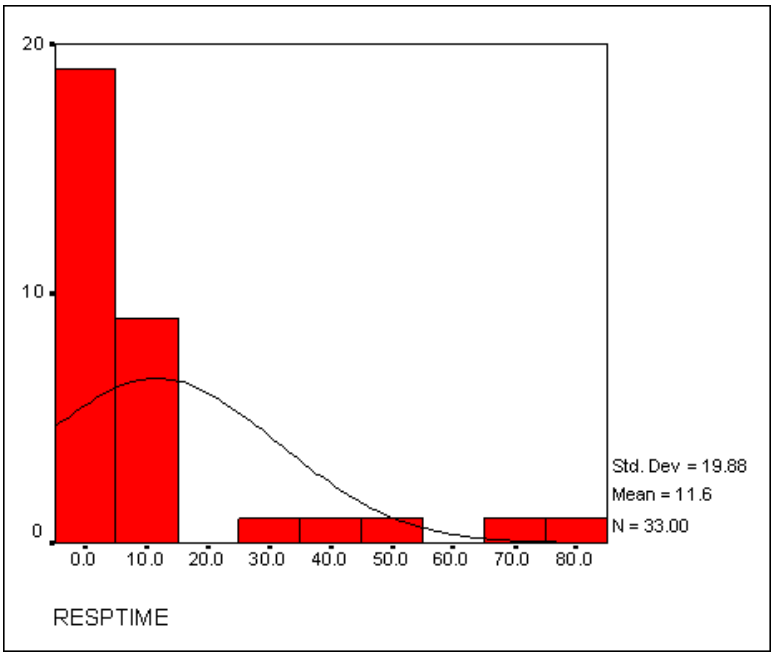
Not all time-interval data will be approximately exponential. In some instances, indeed, they may be more nearly normal. You have to look at each data set on its own merits. But when you do encounter time interval data that have lots of short intervals and a scattering of long ones, it's worth trying the cube root transformation just demonstrated. (Or, at the least, use the original, untransformed data with the special control limits suggested in the footnote above: set LCL = 0 and UCL equal to six times the sample standard deviation.)

A Quality Management Application of Cycle Times

Next we look at a quality management application of the analysis developed above. The data are in the file, OCONNOR.sav. Here is the brief descriptive note:

> Time measured in total elapsed time during business hours (to the nearest half-hour) for president of a manufacturer of industrial fasteners to provide price quotations on prospective orders that had been requested by sales persons. These data refer to the first 33 consecutive orders starting 21 January 1991, before making a change in the quotation process.

The variable is named **resptime**, for "response time", and the histogram below is called for:



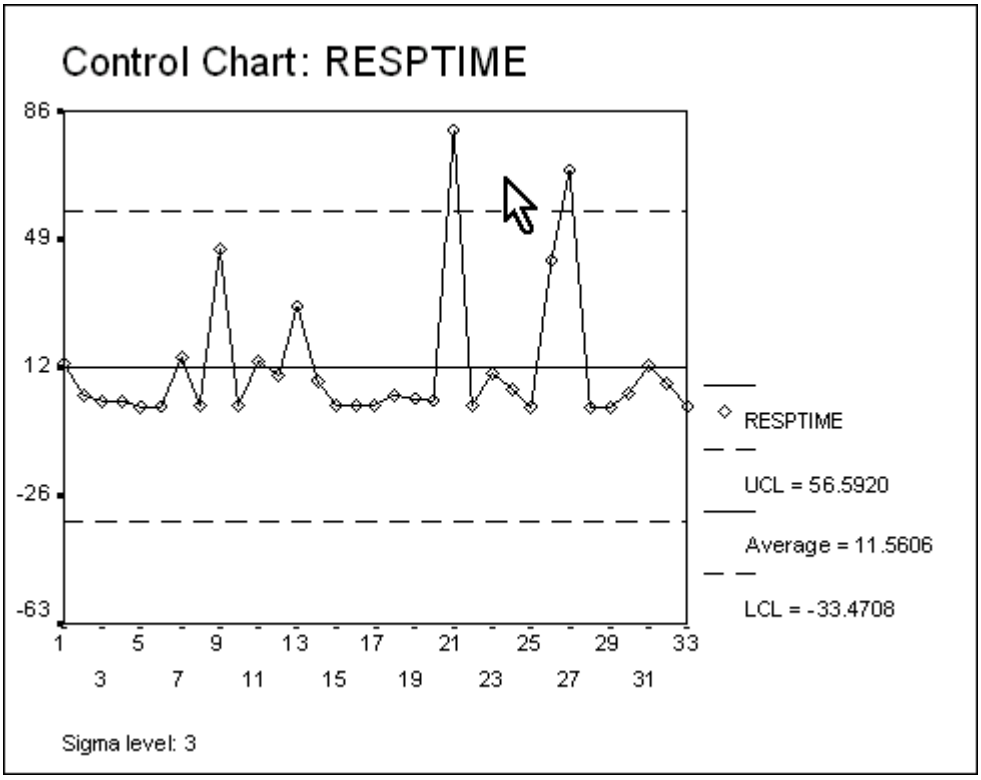
We see the classical **backward J-shape** of the exponential distribution-- most of the price quotations were provided very quickly, usually within a single day, but a few responses took several days. As we see in the following descriptive statistics, the maximum response time in the sample is 80.5 hours. Note also the very large positive skewness statistic.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
RESPTIME	33	.00	80.50	11.5606	19.87569	2.396	.409
Valid N (listwise)	33						

The standard deviation, 19.876, is substantially larger than the mean, 11.561, so the exponential model may not be strictly appropriate. Nonetheless, it may turn out that the cube

root transformation will be helpful for rough analysis. Let's continue with some checks to see if the process is in control.



This time we have two data points that fall above the upper control limit. The labels on the horizontal axis tell us that the first exception occurred at point 21 (80.5 hours). The second violation was at point 27 (68.5 hours).

The practical question is, "Do points 21and 27 really signal a good chance of finding a special cause?" The upper control limit was computed under the assumption of a normal distribution, and it is apparent from the plot that that assumption is wide of the mark: most observations are very close to the mean level, but there are five observations scattering substantially above it, and point 21 happens to be the most extreme of these. (Note also that 80.5 hours is less than 6 times the sample standard deviation of 19.88, that is, 119.28; this was the alternative approach to control charting suggested in a footnote above.)

Further, the LCL computed under the normality assumption is an impossible, negative, -33.47 hours.

Qualitatively at least, this is the kind of result we would expect under the exponential distribution. Therefore, let's follow through with our general strategy of data analysis. First, the runs count is consistent with statistical control for the time sequence of the data:

Runs Test

	RESPTIME
Test Value[a]	11.5606
Cases < Test Value	24
Cases >= Test Value	9
Total Cases	33
Number of Runs	16
Z	.634
Asymp. Sig. (2-tailed)	.526

a. Mean

Next, let's call up the **Transform/Compute…** dialog box and create the new variable **curtrstm= resptime**(1/3)**, the cube root transformation:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
CURTRSTM	33	.00	4.32	1.5863	1.20242	.566	.409
Valid N (listwise)	33						

Note that the skewness statistic for **curtrstm** is considerable less than that for the original variable, **resptime**. It is just a hair outside of the "rule-of-thumb" range, -0.5 to +0.5, that we have given above for normal data. Thus we are hopeful that the deviation from normality will not be too serious.

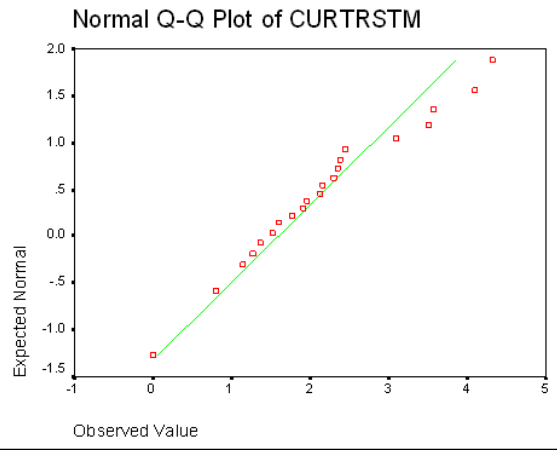
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
CURTRSTM	.109	33	.200*	.938	33	.058

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

CURTRSTM

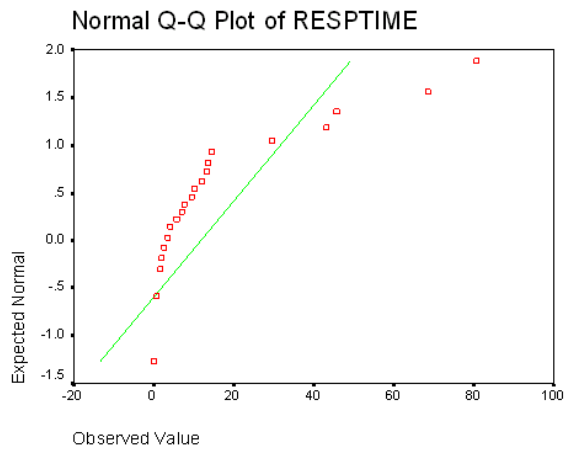


Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
RESPTIME	.290	33	.000	.623	33	.000

a. Lilliefors Significance Correction

RESPTIME



We have shown the **Normal Q-Q plots** for **curtrstm** and **resptime** side by side to emphasize the improvement that the cube root transformation has achieved. It is not perfect, but the plot for **curtrstm** is much closer to the desired straight line, and the significance level for the Shapiro-Wilk statistic is greater than 0.05, although just by a hair.[4]

You can verify, by running a control chart for **curtrstm**, that points 21 and 27 no longer fall outside the control limits. Instead of outliers, they should be seen as part of a group of five rather high observations, for each of which the price quotation took several business days.

A reasonable practical implication might be this: **it is desirable to explore for root causes of delay common to all the multi-day response times rather than to look only for some special cause at observations 21 and 27.** Remember the outlying observation of the apparently misplaced mile marker in the cruise-control example of Chapter 2: the first priority was clearly to figure out what went wrong with mile marker 18 **in particular**. That marker was grossly out of line with the rest of the data. In the current application, observations 21 and 27 are not isolated extremes but just two of a number of rather long response times.

3. Analysis of Defect Rates

An important category of quality measurements concerns defect rates, error rates, or rates of occurrence of any undesirable outcome. Such data can be expressed as proportions or ratios, that is, the number of defective outcomes divided by the number of opportunities for defective outcomes to occur.

If the defects follow a simple chance mechanism -- analogous to heads and tails in independent tosses of a possibly loaded coin -- a standard statistical distribution, the **binomial distribution**, is applicable as a statistical model for defect rates.

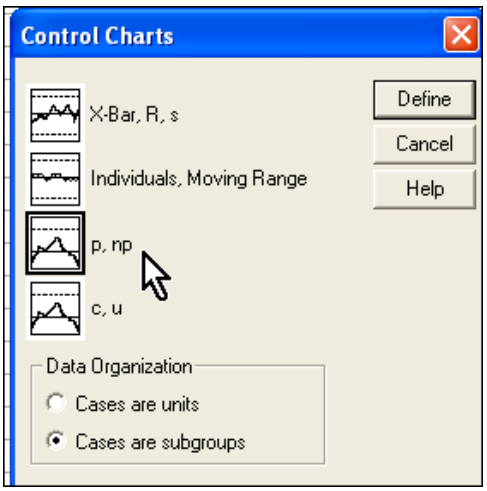
Again, there is no law that requires defect rates to conform to the binomial distribution, but, just as counts often conform approximately to the Poisson distribution and numerical measurements to a normal distribution, defect rates often conform approximately to the binomial distribution.

Just as we used the **c-chart** when we suspected that the Poisson distribution might be a good model, so we use the **p-chart** when we suspect that the binomial distribution might be a good model. If our guess turns out badly, then we have further work to do, and this will be taken up later.

For illustration we shall use data on mortality rates in the intensive care unit of a hospital. The data are contained in the file INTCARE.sav. The hospital has kept track of mortality over time in 60 successive groups of 20 patients each. The variable **mort** indicates the number of patients who expired within six months of admission to the ICU. The second column contains **n**,

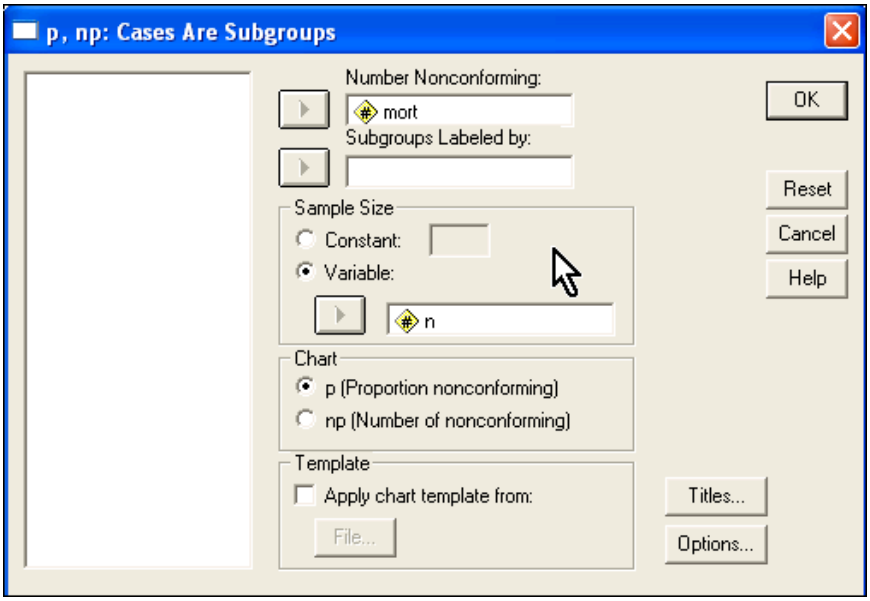
[4] Make sure that you remember what the significance level (p-value) is telling you. Should you reject the hypothesis that the data for **curtrstm** are normal or should you not reject? Explain your answer.

the number in each group. Although in this example the size of each group is 20, it need not be constant for the **p-chart** to be applicable.



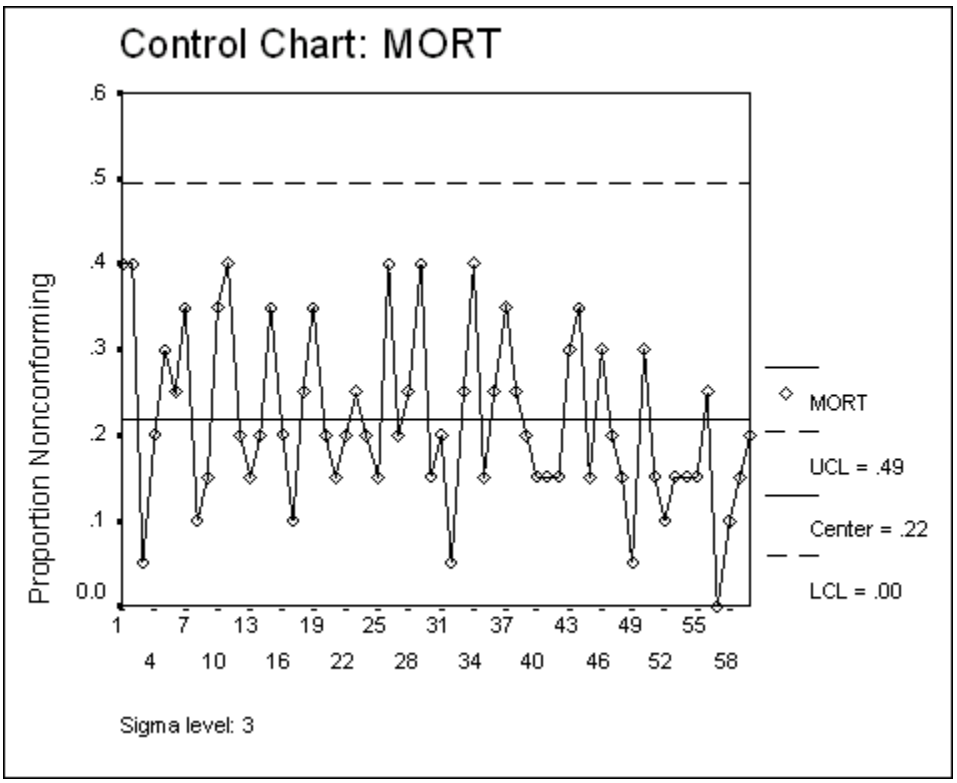
In the **Control Charts** window at left we highlight the icon for **p,np** and make sure that the little circle at the bottom for **Cases are subgroups** is marked.

Then we set up the next window like this:



The variable name **mort** has been moved into the top space and the variable **n** is moved to indicate the sample size. Alternatively in this example, we could have marked the **Constant** circle for **Sample Size** and typed in the number 20. Finally, under **Chart** we indicate that we want to show the **Proportion nonconforming, p**, where in this case "nonconforming" is an unfortunate euphemism.

Here is the resulting **p-chart**:



We see that on the average about 4.4 patients (0.22 times 20) died within six months after admission to the ICU.

Note that the control limits in the **p-chart** above are **not** symmetrical: the distance from the lower control limit of 0 to the center line at 0.22 is less than the distance from the center line to the upper control limit at 0.49. Overall, the binomial control limits seem well placed, although in Section 5 we shall point out that there is a hint of downward trend in mortality. (A trend would mean that the process is not strictly in statistical control.)

With respect to statistical control of the sequence of data, there are no symptoms of special causes and the plot is consistent with a state of randomness.

Note that the points plotted in the **p-chart** are the ratio of **mort** to **n**, that is, the proportion of deaths in each successive group of patients. To analyze this ratio further we must first perform the ***SPSS*** transformation **mortrate= mort/n**. You should verify by examining a histogram with a superimposed normal curve and by performing the normal probability plot that the proportion, **mortrate,** does not pass the tests for normality.

The histogram does not look very normal, with the buildup to the left of the peak of the normal curve and its box-like shape. The binomial distribution is one of many different non-normal distributions that are well approximated by the normal distribution if certain parameters (in this case, the group sample size) are large enough, but for these data **n**=20 is not large enough**.** The **p-chart**, however, is especially designed for plotting the time sequence of binomial

observations. One reason that we need it is that the usual chart for individuals could give us a negative lower control limit for **mortrate**, and we know that the variable cannot be less than zero.

The "Hot Hand"

Since we have introduced the binomial distribution, we can show an interesting application of it. If data conform to the binomial distribution, the sequential pattern of "successes" and "failures" (above, "deaths" and "survival") should resemble the patterns encountered in independent coin tossing, where the coin may possibly be biased towards heads or tails.

We can therefore check out data on successes and failures in various processes to see if this holds. In sports, as we have seen in Chapter 2, it is widely believed that successes tend to cluster and failures are **not** like coin tosses. Sports fans often think that they see "hot streaks" and "cold streaks", "the hot hand", and "shift of the momentum of the game". If true, the binomial model would be far from realistic.

The runs count can check the reasonableness of the binomial model. If the model is a good approximation, runs of successes and failures should be compatible with what would be expected under the hypothesis of statistical control.

	shotmade
1	0
2	1
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	1
11	0
12	0
13	0
14	1
15	0
16	1
17	1
18	1
19	0
20	1
21	0
22	1
23	1
24	1
25	1
26	0
27	0
28	1
29	1
30	0

We can illustrate very simply by a lightning data set, BULLCELT.sav. The contents are described as follows:

Consecutive shots made (1) and missed (0) by both teams (excluding freethrows) in NBA basketball game between the Chicago Bulls and the Boston Celtics, starting late in first quarter and continuing for the entire second quarter. Game played in Boston Garden, 4 April 1993; finally won by Bulls, 101-89.

The variable of interest is called **shotmade**. It consists only of ones or zeros, but if it is not random we cannot say that it follows the binomial distribution. We can make our point by showing only the first 30 cases below although we have recorded 61 in total:

You can see for yourselves that the data are one long string of 0's and 1's--- 29 successes and 32 failures, to be exact. Next, do the runs test. (We do not show it here, letting you do it on your own. Be sure to check the statement that follows with your own ***SPSS*** output.)

It would appear that there is no evidence here to support the theory of the Hot Hand: observed and expected runs are almost the same, and "significance" is 0.91. This result is typical of what is found when actual performance records are kept and analyzed.

In this application there is a strong tension between statistical analysis and our natural emotional reactions. For example, in the first 15 shots recorded above, only three baskets were made. If you had been watching, your reaction would have been that both teams were "cold".

What, in fact, is happening is that about half the shots were made, but chance fluctuations -- binomial variation -- are such that there are short stretches of the game in which many more or many fewer than half of the shots are made. The danger is that we will regard these short stretches as indicators of special causes. For example, if Michael Jordan had made only 3 of his last 15 attempts (and he occasionally had apparent "cold streaks" like that), we might have been tempted to think that he should stop shooting and pass off to team mates. (Fortunately neither Jordan nor Bulls' coach Phil Jackson succumbed to that temptation!)

4. Checking Stock Market Information

In this section we give an example in which the process is badly out of statistical control, so that Chapter 2 does not apply at all, but in which a simple extension of Chapter 2 provides a good analysis.

The application has intrinsic interest of its own. Often in quality management, it is desired to measure company performance against external benchmarks such as the performance of the company's common stock. Presumably the good effects of the company's quality management program should somehow be reflected in the stock market. Also, stock market performance might give clues as to external special causes that might be impacting the company, favorably or unfavorably.

In practice, top managements often have great difficulty in figuring out possible explanations of why their company stock price has been doing whatever it is doing. One source of the difficulty is that there is normally a great deal of unexplained variation in stock prices from one time period to the next, yet it is tempting to look for special causes when nothing but the normal market churning is going on. This is closely related to tampering, which was discussed in Section 11 of Chapter 2.

Control charts on stock prices are useless and misleading for analyzing stock prices. However, a simple adaptation, to be shown in this section, can give insight. We could use almost any stock market price series through time -- whether the reported prices be daily closings, weekly closings, or monthly closings -- but for illustration, we shall use a stock market series that almost everyone is familiar with -- the Dow Jones Industrial Index -- rather than an individual company stock. The statistical behavior of company stock prices would be quite similar. We shall work with monthly closings for a period of four years, contained in the file STOCKS.sav.

The cases run from September, 1977 through August, 1981. We call the variable **dj** (for Dow-Jones) and show the first few observations below:

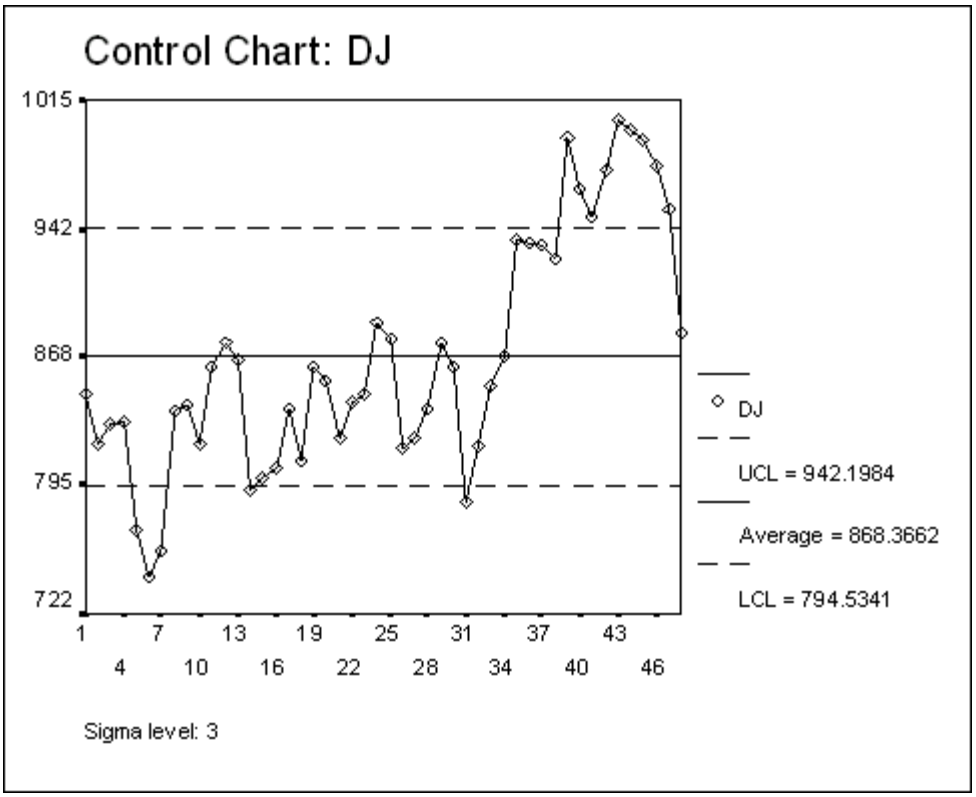
	dj
1	847.11
2	818.35
3	829.70
4	831.17
5	769.92
6	742.12
7	757.36
8	837.32
9	840.61

Here are the descriptive statistics:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
DJ	48	742.12	1003.87	868.3663	67.49627
Valid N (listwise)	48				

Now we'll start out in our usual way, but it will turn out to be a disaster because **dj is not even remotely in statistical control.** Don't panic: we'll show how to get out of the mess and obtain a simple analysis that tells us just what we want to know!



It should be obvious from the control chart above that **dj** is not random. For now the important point is that **dj** falls outside the control limits in many places and in **no way** can be considered to be in statistical control. The points are **not** varying unpredictably around a constant level with constant variance through the time sequence; they are extremely **nonrandom**. They start out below the mean, and gradually meander upwards, with consecutive points usually close together. This meandering behavior is characteristic of data on stock prices, as well as prices on other organized exchanges, including commodity prices and exchange rates. To reinforce the point, here is the runs count, with only 8 runs as against 23.5 expected:

Runs Test

	DJ
Test Value[a]	868.3662
Cases < Test Value	30
Cases >= Test Value	18
Total Cases	48
Number of Runs	8
Z	-4.676
Asymp. Sig. (2-tailed)	.000

a. Mean

←----This should say < .0005 to be exact.

An Alternative, and Appropriate, Analysis

The control chart for successive **levels of dj** is not a useful approach to analysis. There is, however, a simple trick that makes it possible to apply the control chart concept effectively.

This trick is based on the fact that the key to the information conveyed by stock prices are not the closing prices themselves but the **changes** in closing prices from month to month. These changes reflect the new information that has become available in the month since the previous closing.

We can use an ***SPSS*** transformation that simply computes the changes from one month to the next -- the **difference** between this month's price and last month's price. 48 months of prices give 47 months of price changes. Under **Transform/Compute…** we enter the following in the **Compute Variable** window:

diffdj = dj – LAG(dj),

where **LAG(variable)** is a function that returns the immediately previous value of its argument, in this case the last value of **dj**. In the common notation of time series data we could alternatively define

$$\mathbf{diffdj = dj_t - dj_{t-1}}$$

We compare **dj** with the new variable **diffdj** to see just how **LAG( )** works:

	dj	diffdj
1	847.11	.
2	818.35	-28.76
3	829.70	11.35
4	831.17	1.47
5	769.92	-61.25
6	742.12	-27.80
7	757.36	15.24
8	837.32	79.96
9	840.61	3.29
10	818.95	-21.66
11	862.27	43.32

←---The dot indicates missing. There is no value to subtract.
←-- -28.76 = 818.35 – 847.11
←-- 11.35 = 829.70 – 818.35
etc.

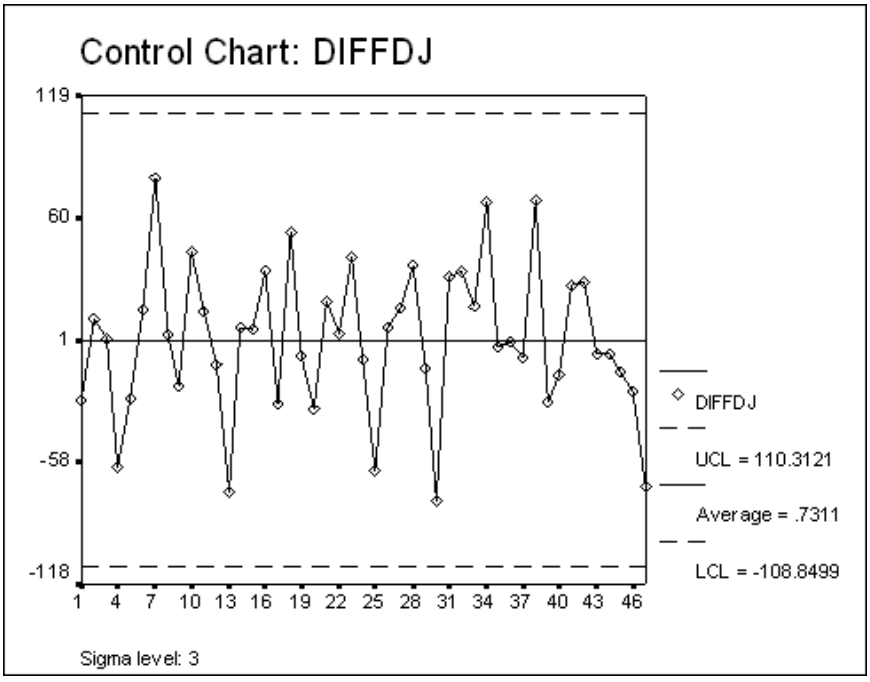
Compare the descriptive statistics for the two variables, **dj** and **diffdj**:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
DJ	48	742.12	1003.87	868.3663	67.49627
DIFFDJ	47	-77.39	79.96	.7311	36.32932
Valid N (listwise)	47				

The mean of **diffdj** is 0.7311, suggesting that on average, **dj** itself has been drifting up a little less than one point a month over this four year period. **Further, the standard deviation of the changes diffdj is 36.329, which is much smaller than the standard deviation of 67.496 in the original dj series. (In this sense, the process of differencing "explains" much of the variation of dj.)**

The statistical behavior of **diffdj** is much different, and more easily interpreted, than that of **dj** itself. The output below suggests that the changes **diffdj** are in a state of statistical control and approximately normally distributed! This means that there are **no** apparent special causes impinging on the **dj** index during this period, and that the control chart analysis can be based on **diffdj**, even though control charts were completely inappropriate for **dj** itself. If you think about it for a moment, you can see that by knowing the values of **diffdj** and just one value of **dj** you can reconstruct the whole time series for **dj**.



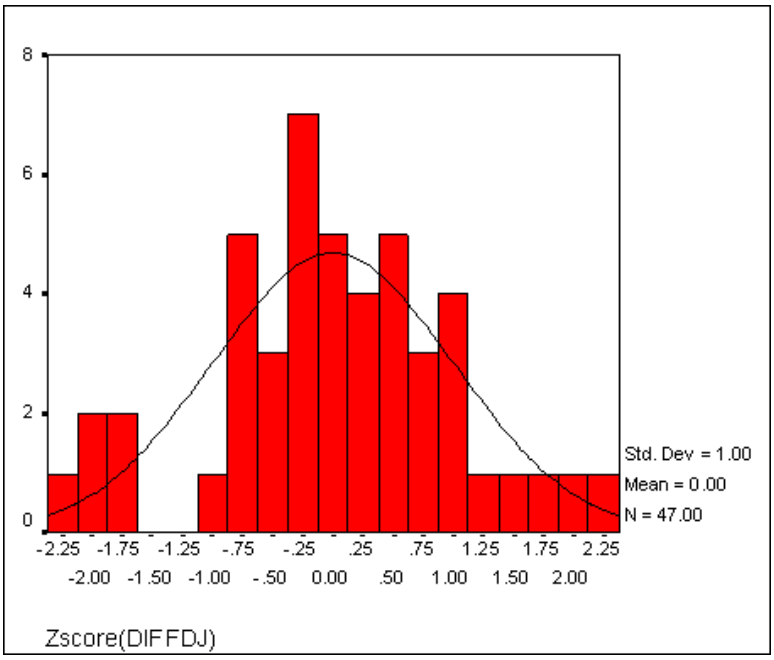
We'll look next at the runs test results which confirm the visual impression of randomness:

Runs Test

	DIFFDJ
Test Value[a]	.7311
Cases < Test Value	23
Cases >= Test Value	24
Total Cases	47
Number of Runs	21
Z	-.882
Asymp. Sig. (2-tailed)	.378

a. Mean

In the histogram below we display the distribution of the standardized values of **diffdj:**



And here is the **Shapiro-Wilk Q-Q** plot:

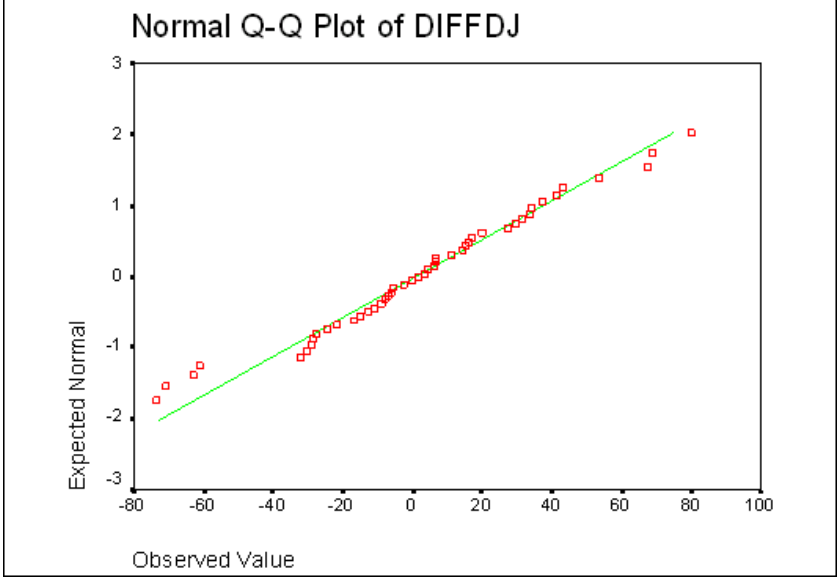
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
DIFFDJ	.073	47	.200*	.978	47	.531

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

DIFFDJ



The high p-value supports the hypothesis that the data are normally distributed.

The changes in **dj** -- **diffdj** -- thus can be analyzed by standard control charts, and other statistical tools as well can be brought to bear even though the original **dj** prices were badly out of control. **dj** is a good example of a series for which the **Random Walk Model** is a good approximation. That is, the original series itself is not random, but first-differencing results in a transformed series **diffdj** that appears to be random. Be sure to read the appendix at the end of this chapter which discusses this subject further— especially after we discuss **autoregression** in Chapter 6.

In modern finance, percentage changes -- called **returns** -- are usually used instead of simple changes in statistical analyses of stock market prices, as well as of prices in other organized markets such as bonds, commodities, and foreign exchange. If the overall price level does not change substantially, either percentage changes or simple changes are usually approximately in statistical control. If there are substantial changes in level, the percentage changes or returns are preferable for analysis.

5. A New Wrinkle: Trend

We have approached data analysis of time-ordered data from the perspective of control charts. That's a good way to begin, and, as we have seen, it often permits us to understand how the process is behaving without need for further analysis.

But sometimes the control charts are only the first step, and further analysis is needed. For the balance of this chapter, we shall consider applications that ultimately will require more than control charts for satisfactory analysis.

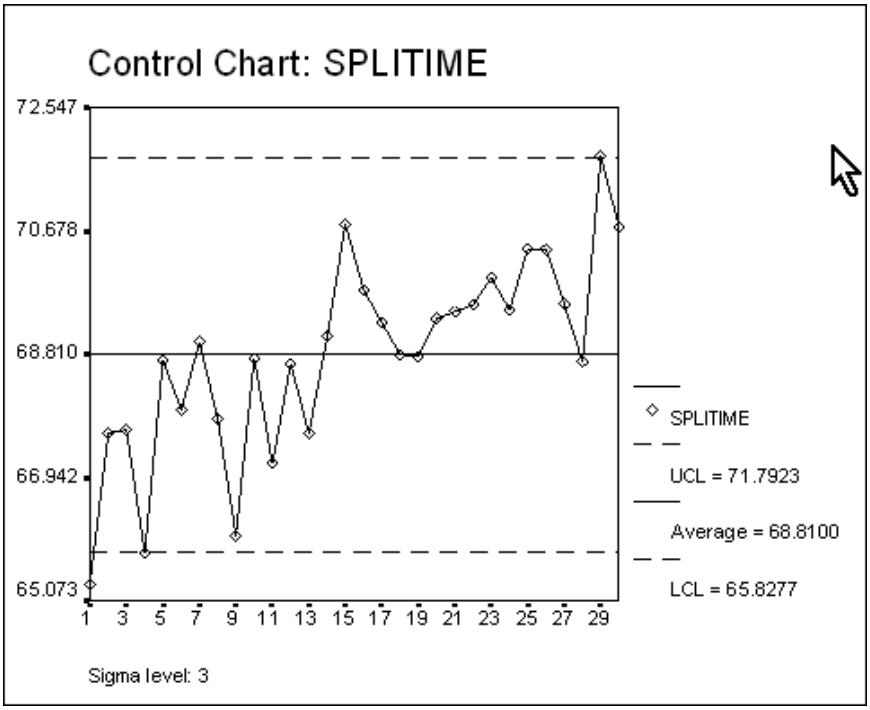
For the first application, we'll look at another lightning data set, this one based on timings of the splits (times at intermediate stages) in a running workout. The ***SPSS*** file is LAPSPLIT.sav

> Data from a running workout around a block that is approximately 3/8 miles, divided by markers into approximately three 1/8 mile segments. Timed with a Casio Lap Memory 30 digital watch to obtain timings for each of 30 1/8 mile segments (3 3/4 miles in total). Attempted to run at constant (and easy) perceived effort, without looking at the splits at each 1/8 mile marker.
> Afternoon of 19 September 93. Data are in seconds.

The variable of interest is named **splitime.** First we look at the descriptive statistics and a control chart:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
SPLITIME	30	65.35	71.82	68.8100	1.48369
Valid N (listwise)	30				



Visual examination of the graph above should show you that the data are wildly out of control. There is an obvious **trend: keeping effort constant, the runner was tending to slow down.** The plot shows this clearly: most of the early points are faster than the mean (below the center line) and most of the later points are slower than the mean (above the center line). The runs count confirms the visual analysis; there are too few runs:

Runs Test

	splitime
Test Value[a]	68.8100
Cases < Test Value	14
Cases >= Test Value	16
Total Cases	30
Number of Runs	8
Z	-2.775
Asymp. Sig. (2-tailed)	.006

a. Mean

In particular, the process is **not** varying around a constant level.

Trends are a common and important symptom of departure from statistical control. They can tell us, for example, that a process is steadily getting better or getting worse, thus suggesting the need to look for possible root causes. The root causes, in turn, may suggest appropriate changes in the process.

- If the trend is towards improvement, we want to maintain whatever it is that is causing the trend.

- If the trend is towards deterioration, we want to find the root cause and remove it.

A trend does not of itself tell us what the root cause is. We have to study the process directly, using all available information and knowledge, and seeking more data if necessary. The discovery of a trend just lets us know that we should be facing up to a problem of which we might otherwise have been unaware. More than just control limits, we need careful visual analysis of the behavior of the points and runs counts, which we have been using regularly. We also need other tools to be introduced later. For example, in Chapter 4 we shall see how to fit a linear trend to these data and check for its statistical significance.

We shall return to this example in Chapter 4 for a discussion of trend fitting using simple linear regression. For now, the main lesson is the importance of visual analysis and the "interocular traumatic test".

The Mortality Data Reconsidered

To reinforce that lesson, look back to the intensive care mortality data in Section 3. If you look carefully -- it's not nearly as obvious as in the running example -- there is a hint of a trend, this time a **downward** trend. The earlier points tend to be above the center line; the later points tend to be below it.

If this is not just a chance aberration, it is of great importance: the performance of the intensive care unit appears to have been **improving** systematically during this time. The practical questions would have been, "Does this improvement reflect something good we're doing?" "If so, what is it? We'd like to be sure that we keep doing it."

6. Another New Wrinkle: Periodic Effects

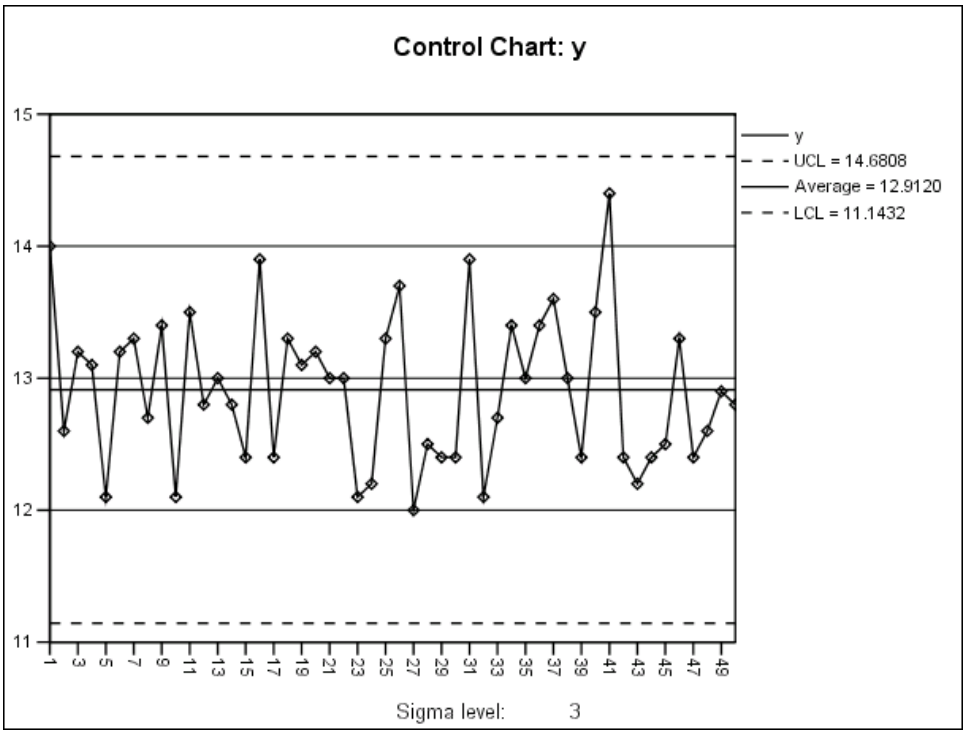
Consider next a manufacturing data set, taken from a standard text on quality control. Specific units of measurement are not given, but presumably it is some kind of dimensional or yield measurement. The file is ISHIKAWA.sav:

> First 50 observations from Ishikawa, *Guide to Quality Control*, Table 7.2, page 66. Each successive five observations represent measures taken at successive times on the same data: 6:00, 10:00, 14:00, 18:00, and 22:00. Thus the 50 observations represent 10 days, which we shall assume to be consecutive working days. The original data set has 75 further observations for another 25 working days; to save space we will not look at these at this time.

Since we do not know exactly what kind of measurements we are dealing with we have simply named the variable "**y**".

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
y	50	12.00	14.40	12.9120	.57237
Valid N (listwise)	50				



Runs Test

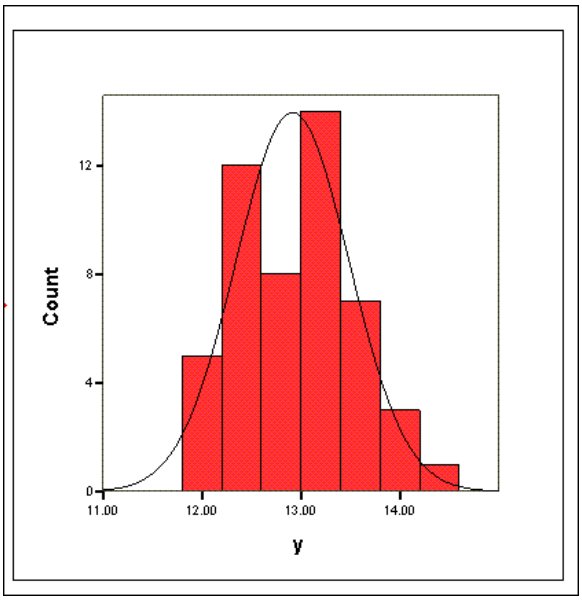
	y
Test Value[a]	12.9120
Cases < Test Value	25
Cases >= Test Value	25
Total Cases	50
Number of Runs	26
Z	.000
Asymp. Sig. (2-tailed)	1.000

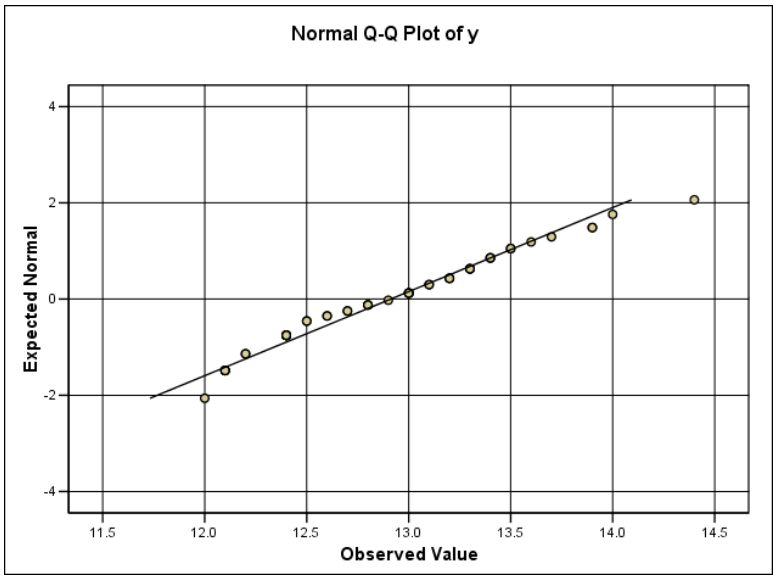
a. Mean

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
y	.114	50	.099	.964	50	.130

a. Lilliefors Significance Correction





It would appear that we have an example of a process in statistical control, and with a nearly normal distribution, although we do see a hint of skewness to the right in the histogram. That is, there is a tendency for the larger observations to straggle out.

But wait! Recall that each set of five measurements was taken at one of five times during the day: 6:00, 10:00, 14:00, 18:00, and 22:00. It is therefore possible that there is some kind of systematic within-day pattern that we are missing, a **periodic** or **seasonal** effect.

To bring out this possible periodic effect visually we need to create a new variable using the ***SPSS*** sequence **Data/Insert Variable** after highlighting the second column in **Data Editor**. At first the new column looks like this:

	y	VAR00001
1	14.00	.
2	12.60	.
3	13.20	.
4	13.10	.
5	12.10	.

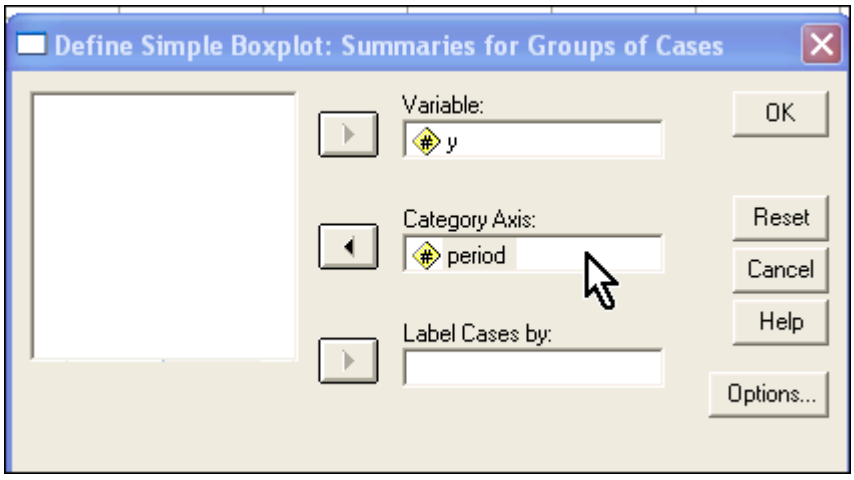
After clicking on the **Variable View** tab in **Data Editor** we rename the new variable **period** and change the **Decimals** entry to 0 and the **Width** to 1. That section of the display should look like this:

	Name	Type	Width	Decimals
1	y	Numeric	4	2
2	period	Numeric	1	0

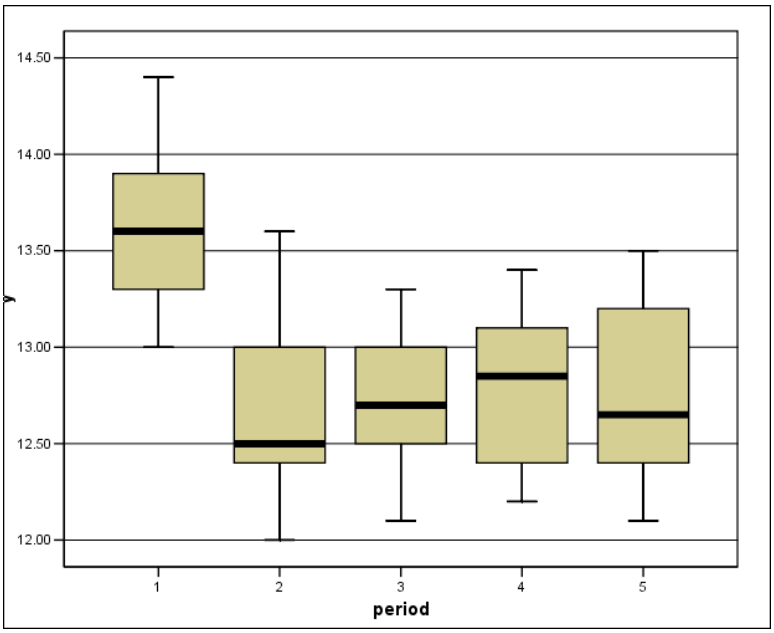
We now have to insert some values for the new variable, **period**. Since the data set consists of only 50 cases the easiest way to do this is to type in successively the numbers 1,2,3,4,5,1,2,3,4,5, etc. (There are other methods to achieve the same results if the number of cases is large but they are rather complicated and we will not discuss them here.) We see that period is merely an indicator of the time of day at which the process measurement was observed. The **Data Editor** now looks like this:

	y	period
13	13.00	3
14	12.80	4
15	12.40	5
16	13.90	1
17	12.40	2
18	13.30	3
19	13.10	4
20	13.20	5
21	13.00	1
22	13.00	2
23	12.10	3
24	12.20	4
25	13.30	5
26	13.70	1
27	12.00	2
28	12.50	3
29	12.40	4
30	12.40	5
31	13.90	1

The next step is execute the sequence **Graphs/Boxplot…** and to set up the resulting dialog window like this:



Note that the new variable, **period**, will determine the categories. Here is the plot:

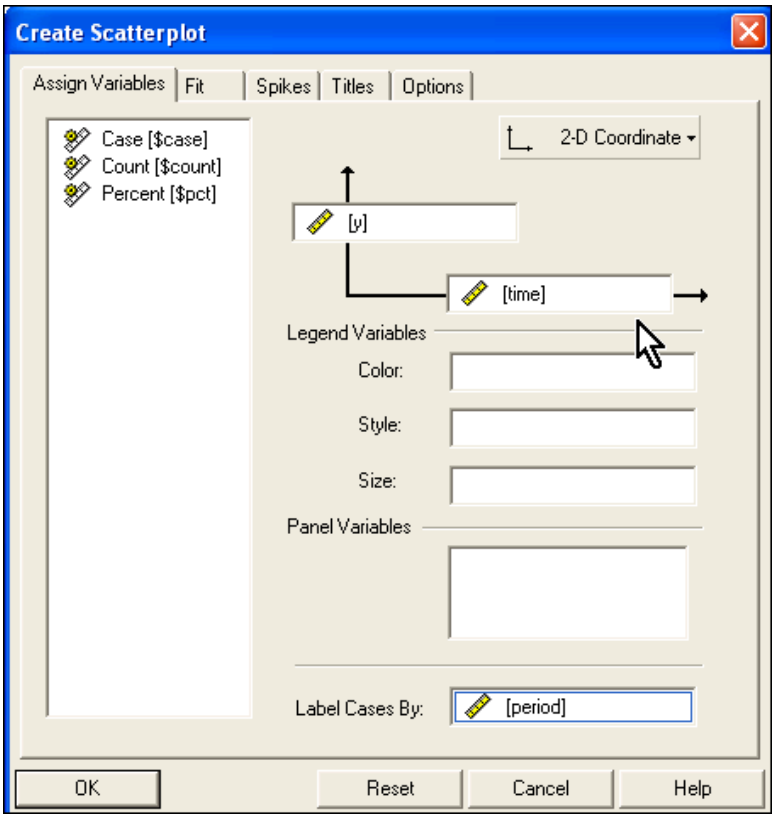


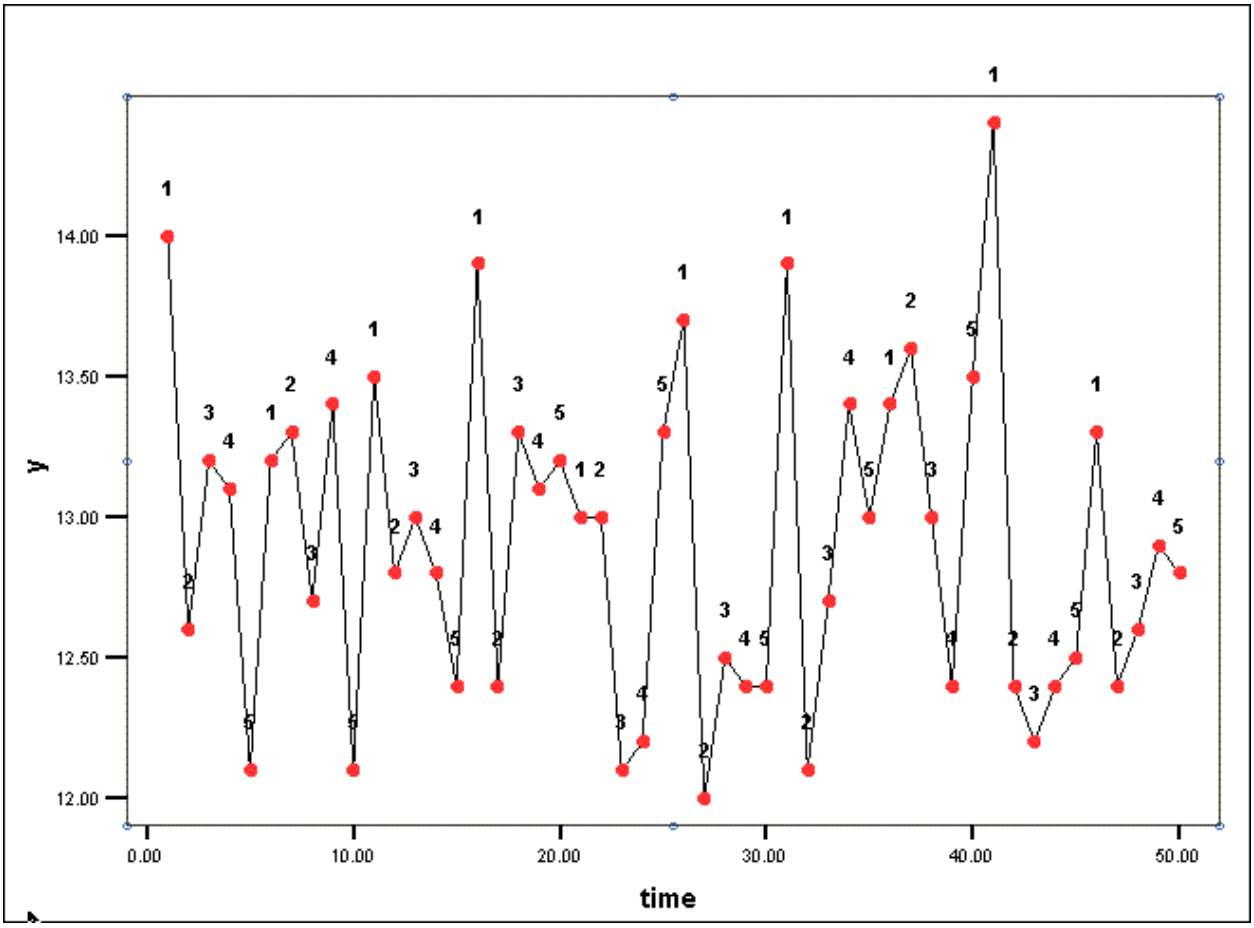
Observe that the median value for period = 1, that is, the first measurement of each day, is about one full unit higher than the remaining measurements for the same day.

We have detected an apparent "startup effect", an effect that is totally incompatible with statistical control. We don't know whether high values are good or bad, but the visual analysis sends a clear signal to study the startup process itself to see why the early morning readings tend to be higher than those taken later in the day.

It is possible to make allowance for the startup effect in fitting the data by linear regression, and we shall show how to do this in Chapter 4. In the meantime, we have seen a triumph for the visual analysis that yields information not given by the standard control chart.

For a final display we show the result of using the ***SPSS*** sequence **Graphs/Interactive/Scatterplot…** followed by a considerable degree of "fiddling around" and trial and error with the **Chart Editor**, a feature that is available for advanced users. If you want to try it, fine, but do not take valuable time away from your mastering the fundamentals that we have covered thus far, and unless you have a lot of previous experience with ***SPSS*** you may waste several hours before you achieve the display that we show below. (For your information the new variable, **time**, shown in the following dialog window, was created by the transformation **time=$casenum**, where **$casenum** is a built-in function that is not even mentioned in the *Brief Guide*.)





It is apparent visually that the **1**'s tend to be at the top of the plot; on average, they are substantially higher than the other points.

7. A Final New Wrinkle: Intervention Effects

In the next data set, quality improvement -- of putting in golf -- was the major objective. The project was undertaken by a student who was an excellent golfer but who had never been satisfied with his putting game.

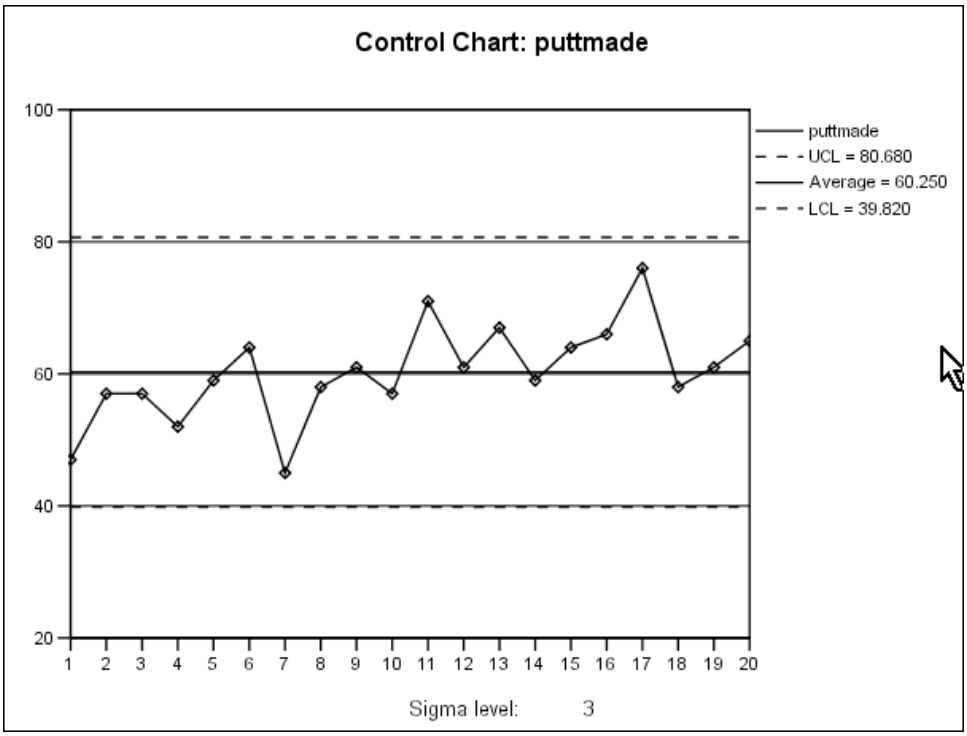
PUTTING.sav:

47 57 57 52 59 64 45 58 61 57
71 61 67 59 64 66 76 58 61 65
Indoor putting experiment. Putts sunk per 100 trials from fixed distance. At the end of the first 10 groups of 100, it was noticed that 136 of 443 misses were left misses and 307 were right misses. It was reasoned that the position of the ball relative to the putting stance was a problem. "I concluded that the ball was too far "back" (too much in the middle) of my putting stance. I moved the ball several inches forward in my stance, keeping just inside my left toe." The final 10 observations were made with the modified stance.

The data in the file PUTTING.sav are opened in ***SPSS*** and the variable is named **puttmade**. The successive numbers are the numbers of successful putts out of each group of 100, thus they could be called "percentages of putts made." The descriptive statistics and Control Chart are as follows:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
puttmade	20	45.0	76.0	60.250	7.3117
Valid N (listwise)	20				



Again, all points are within control limits, and the runs check (not shown) was satisfactory, but the process is clearly not in control.

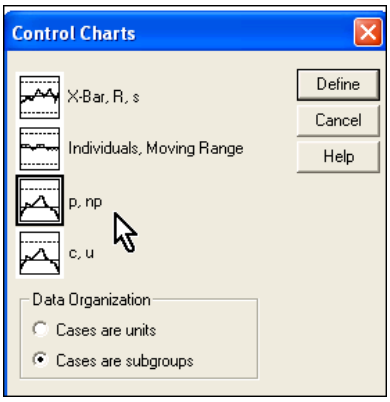
Your first thought may be that this is an upward trend due to the practice the student was getting as he did the successive series of 100 putts. Another interpretation is that there was a **level shift** upward for the last 10 observations as the student changed his stance based on his interpretation of the pattern of left and right misses for the first 10 observations. When we come back to this application in Chapter 4, we shall show that there is good evidence that it was the stance change, not the effect of practice.

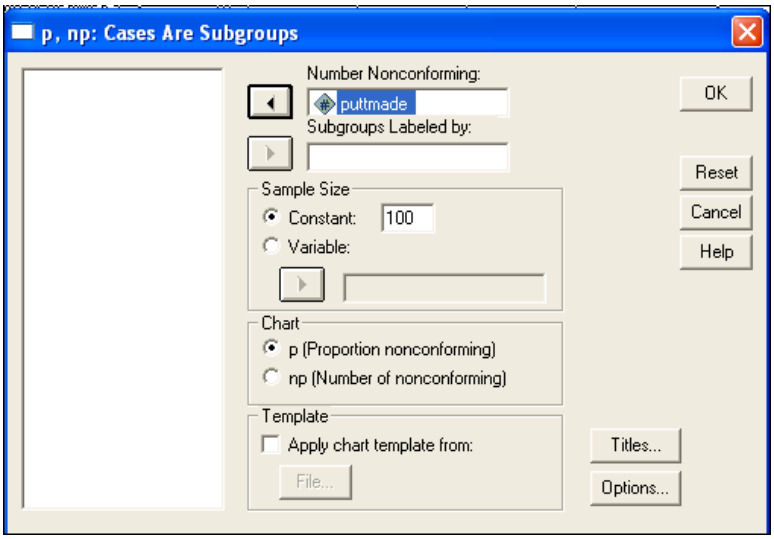
We have here an example of a good quality improvement project: preliminary data suggested the modification of stance; the subsequent data suggested that the modification of stance led to a substantial improvement, an estimated 16 percent. This amounts to several strokes per round.

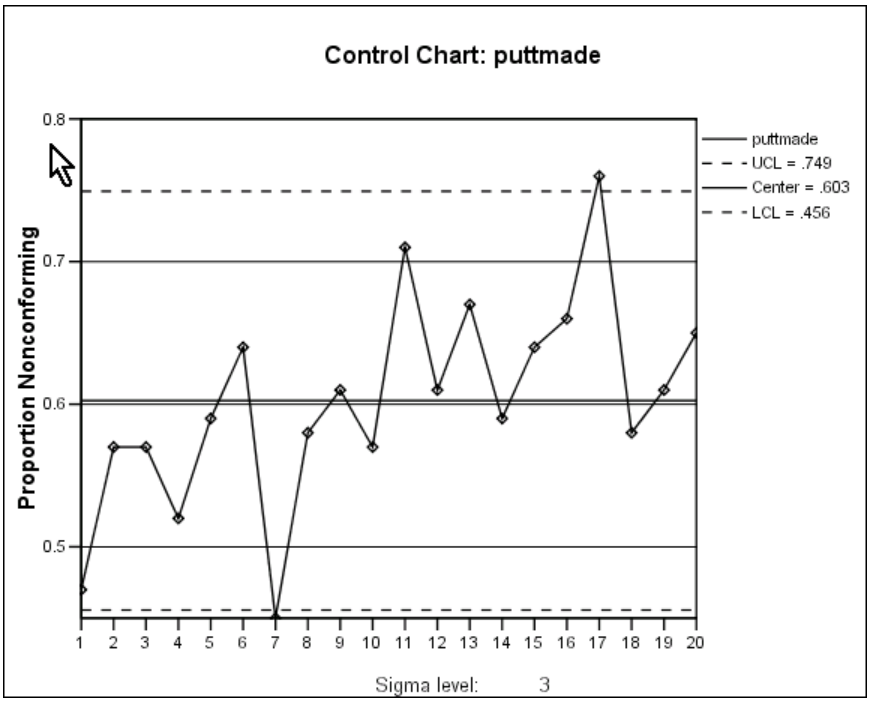
Central to the golfer's success was keeping data on right misses versus left misses, noticing the surprising pattern in these right and left misses, and formulating a shrewd hypothesis that suggested a process change that might lead to improvement. This illustrates the essence of quality improvement:

- Learning from past data how to do things differently and, it is hoped, better;
- then checking statistically whether the change had the desired effect.

Since it is possible that the binomial distribution might describe the golfer's percentage of successes for any one series of 100 putts, we could alternatively have looked at these data from the perspective of a **p-chart.** The images below show the set up that is required**:**







The **p-chart** detects two seemingly out-of-control individual points, one on the low side at point 7 and one on the high side at point 17. It is clear from our discussion, however, that looking for special causes at those two particular points would be fruitless. They are simply a reflection of

the golfer's success in improving his game by improving his stance: Point 7 occurred before the change of stance, while Point 17 occurred afterwards. (Or, if the trend were due to improvement with practice, this interpretation of the apparently "out-of-control" points would apply.) What is important is not observations 7 and 17, but the mean of observations 11-20 compared to the mean of observations 1-10; or in the alternative interpretation, the upward trend.

Appendix: RANDOM WALKS

Let $\ldots, X_{t-1}, X_t, X_{t+1}, \ldots$ be a time series, where X_t means the value of the series observed at time t. Some series are called "random walks". The characteristics of these random walks are that they meander in an unpredictable manner, but they are not themselves random because their mean and standard deviation are not constant. They also exhibit a great deal of persistence -- that is, successive values tend to be much closer to immediately preceding values than to values that are far away. Thus random walks show too few runs to pass the runs test. Later on when we learn about autocorrelation, we will see that random walks are very highly autocorrelated-- just another way of saying that successive values are closely associated.

The best way of defining a random walk is to say that it is a time series for which the **first differences** are random. First differencing is a transformation that creates $D_t = X_t - X_{t-1}$. In other words, the first difference at time t is the variable X at time t minus X observed at the immediately preceding time period, $t-1$. So, remember this: **A random walk is not random, but its first differences are random.**

Dow Jones' series of industrials is a good example of a time series that fits the model of a random walk very well. Many economic time series seem to be well described by the random walk model, but certainly not all.

CAUTION: Never transform a series that is already random to first differences. The resulting first differences will not be random.

Look at the following after we have studied the autoregression model (Chapter 6):

The simplest autoregression model is

$$X_t = \beta_0 + \beta_1 X_{t-1} + \varepsilon_t$$

where β_0 is a constant, and ε_t is a random error term with mean 0 and constant standard deviation. Now to say that the first differences of a time series are random is to say that

$$D_t = X_t - X_{t-1} = \beta_0 + \varepsilon_t$$

In other words the first difference is equal to a constant (its mean) plus a random error term with mean 0 and constant standard deviation. Note, however, that by moving X_t to the right-hand-side of the equation, we can write

$$X_t = \beta_0 + X_{t-1} + \varepsilon_t ,$$

of the same form as the autoregression model, but with the slope coefficient, β_1, equal to one. Thus we see that a random walk is just a special case of autoregression, where the slope is one.