Chapter 1: Introduction to Data Analysis

- Preview:

  - o Data and its science, statistics
  - o Stationary and nonstationary processes; displaying data from each.
  - o Assessing causes of variation.

- **What's the IDEA?**

  - o Data have variation.
  - o The variation has a pattern (data distribution).
  - o By analyzing the pattern, we can tell something about the process or population the data came from.

- Data=Facts Which Convey Information (for example, the PREZHITE data set (found in SASDATA.PREZHITE):

YEAR	WINNER	WINPARTY	WINHITE	LOSER	LOSPARTY	LOSEHITE
1868	Grant	(R)	68.5	Seymour	(D)	71.5
1872	Grant	(R)	68.5	Greeley	(D)	70.0
1876	Hayes	(R)	68.5	Tilden	(D)	.
1880	Garfield	(R)	72.0	Hancock	(D)	74.0
1884	Clevelan	(D)	71.0	Blaine	(R)	.
1888	Harrison	(R)	66.0	Clevelan	(D)	71.0
1892	Clevelan	(D)	71.0	Harrison	(R)	66.0
1896	McKinley	(R)	67.0	Bryan	(D)	72.0
1900	McKinley	(R)	67.0	Bryan	(D)	72.0
1904	T.Roosev	(R)	70.0	Parker	(D)	72.0
1908	Taft	(R)	72.0	Bryan	(D)	72.0
1912	Wilson	(D)	71.0	T.Roosev	(P)	70.0

1916	Wilson	(D)	71.0	Hughes	(R)	71.0
1920	Harding	(R)	72.0	Cox	(D)	.
1924	Coolidge	(R)	70.0	Davis	(D)	72.0
1928	Hoover	(R)	71.0	Smith	(D)	.
1932	F.Roosev	(D)	74.0	Hoover	(R)	71.0
1936	F.Roosev	(D)	74.0	Landon	(R)	68.0
1940	F.Roosev	(D)	74.0	Wilkie	(R)	73.0
1944	F.Roosev	(D)	74.0	Dewey	(R)	68.0
1948	Truman	(D)	69.0	Dewey	(R)	68.0
1952	Eisenhow	(R)	70.5	Stevenso	(D)	70.0
1956	Eisenhow	(R)	70.5	Stevenso	(D)	70.0
1960	Kennedy	(D)	72.0	Nixon	(R)	71.5
1964	Johnson	(D)	75.0	Goldwate	(R)	72.0
1968	Nixon	(R)	71.5	Humphrey	(D)	71.0
1972	Nixon	(R)	71.5	McGovern	(D)	73.0
1976	Carter	(D)	69.5	Ford	(R)	72.0
1980	Regan	(R)	73.0	Carter	(D)	69.5
1984	Regan	(R)	73.0	Mondale	(D)	70.0
1988	Bush	(R)	74.0	Dukakis	(D)	68.0
1992	Clinton	(D)	74.0	Bush	(R)	74.0
1996	Clinton	(D)	74.0	Dole	(R)	73.0
2000	Bush	(R)	71.0	Gore	(D)	73.0

- Statistics: The Science of Data

- Displaying Data

  - Frequency Histogram: shows **Data Distribution**: static pattern of variation. A histogram of the election winner's heights from the prezhite data set is shown in Figure 1.

  - Time Series Plot (or Line Plot): shows pattern of variation evolving over time. Figure 2 displays a time series plot of election winner's heights from the prezhite data set.

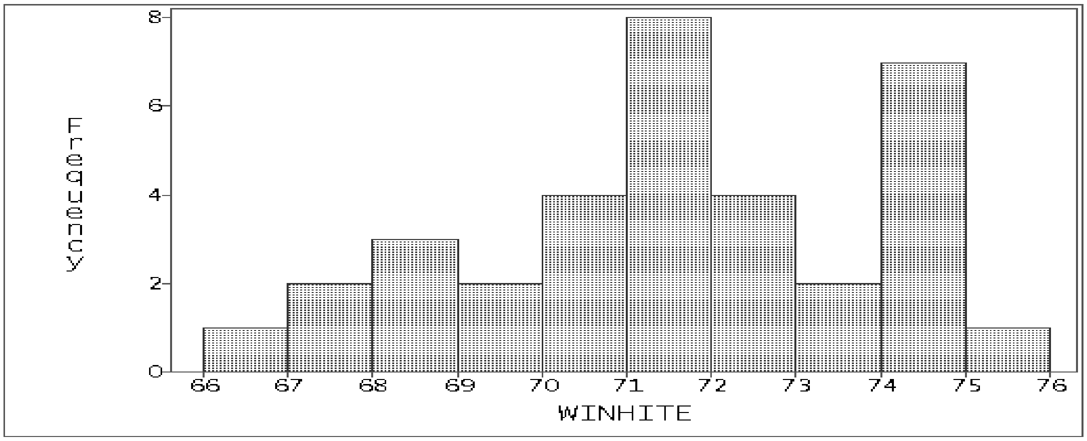

Figure 1: *Histogram of election winner's heights.*

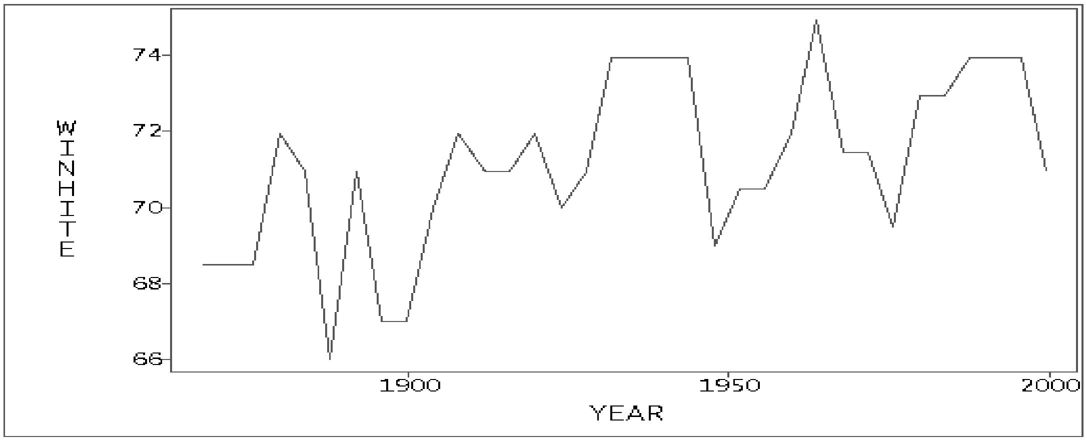

Figure 2: *Time series plot of election winner's heights.*

What's the IDEA?

The first step in analyzing data should **ALWAYS** be to plot it. **BUT...** be sure to use appropriate plots.

- Stationary Processes
  - o Pattern of variation does not change as more data are taken.

o To assess stationarity data must be plotted versus time. Figure 3 shows a histogram of some simulated data:

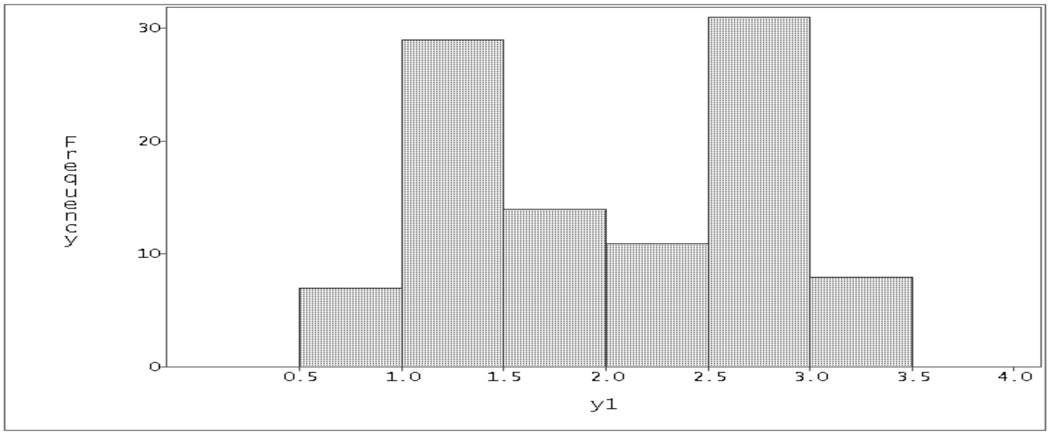

Figure 3: *Histogram of some simulated data.*

The histogram in Figure 3 will always look the same regardless of the order in which the data were taken. Figure 4 shows three possible plots of these same data, obtained by rearranging the order in which the data are plotted. These three plots tell very different stories about the process that produced the data. This example shows that to properly analyze data taken over time, it is necessary to plot the data versus time.

Figure 5 displays a histogram of monthly sales of cars and car parts for the years 1971-1991 (Data found in SASHELP.USECON). From it, we can only see a pattern of variation for all months in the period.

In order to see if or how the monthly sales change over time, we must construct a time series or line plot, as shown in Figure 6.

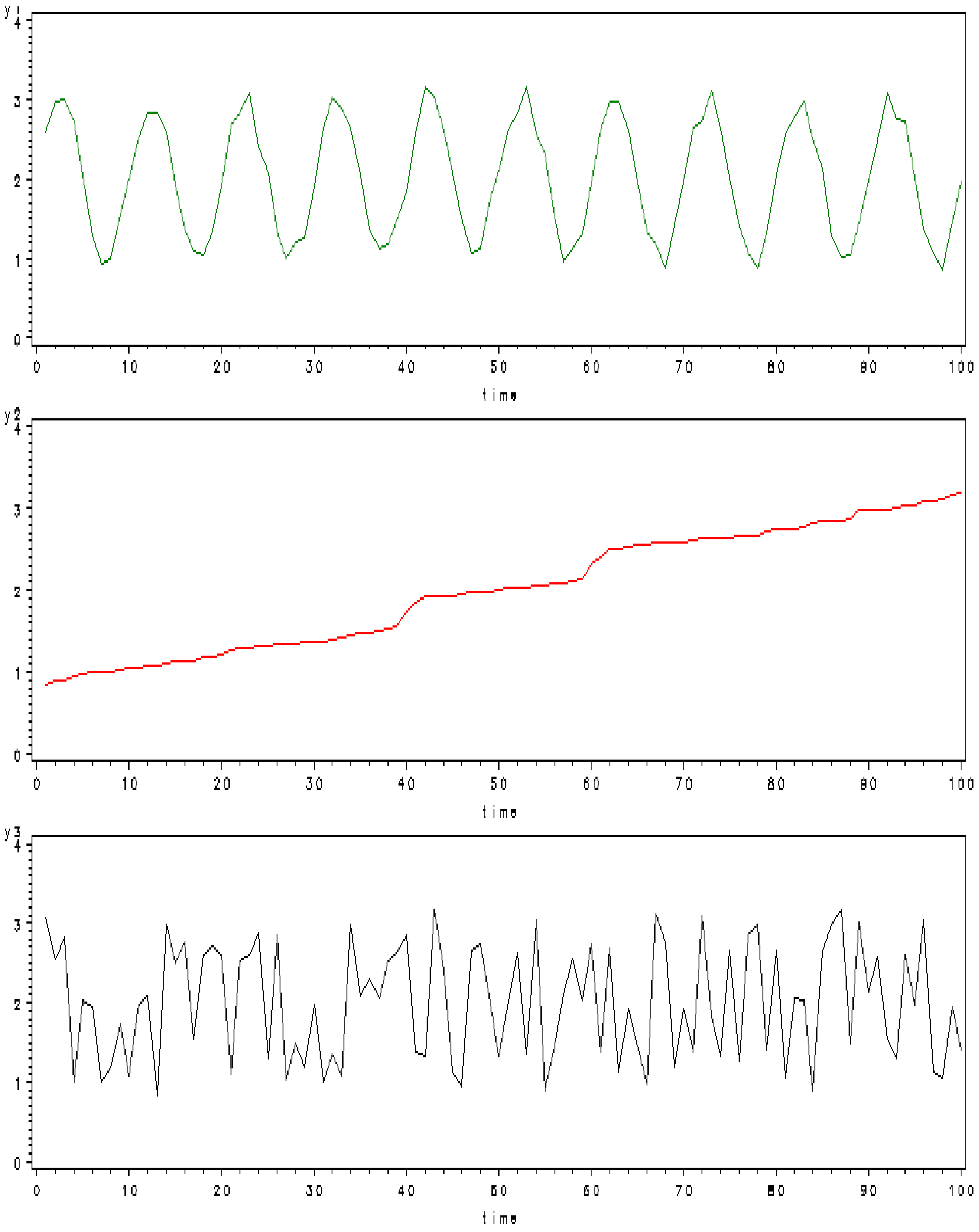

Figure 4: *Three possible time series plots for the simulated data.*

This plot shows a clear increase in level and spread over time.

o An l-term Moving Average replaces the data value at time t with the average of itself and the

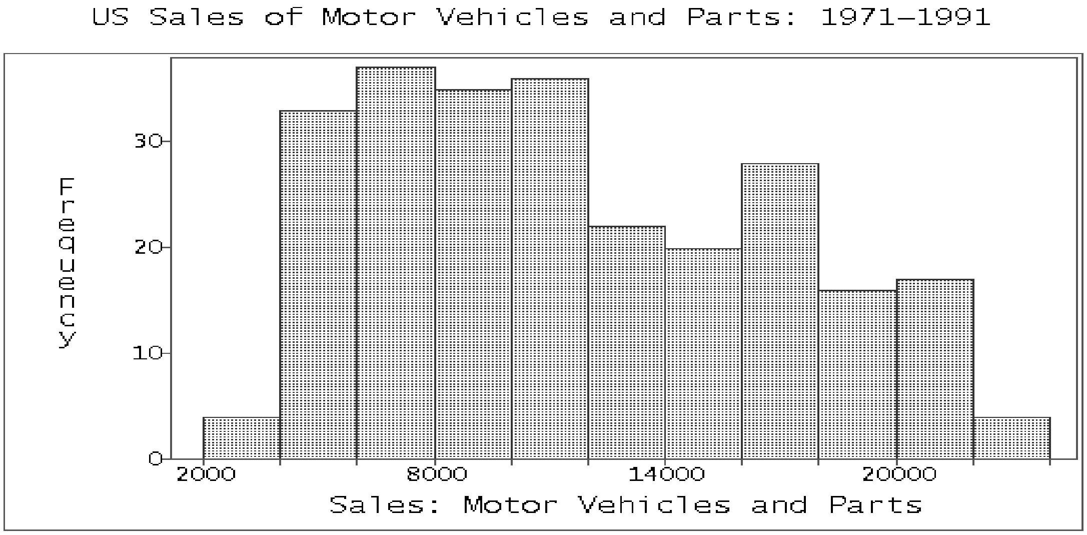

Figure 5: *Histogram of monthly sales of cars and car parts for the years 1971-1991.*

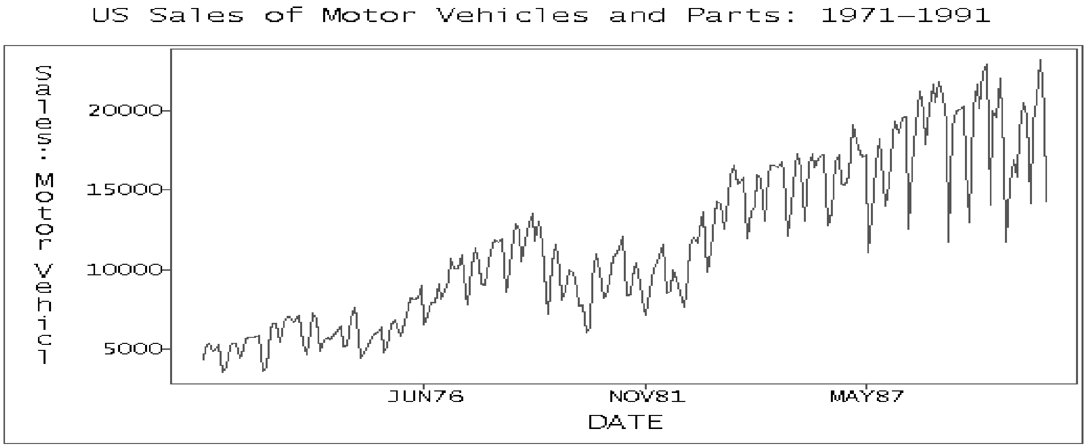

Figure 6: *Time series plot of monthly sales of cars and car parts for the years 1971-1991.*

$l - 1$ previous data values. Moving averages

- $*$ Remove extraneous variation.
- $*$ Help show trends: changes in average value over time.

The variation in the car sales data seems to be a result of both an increasing trend and yearly cyclic behavior. The 12 term moving average, shown in Figure 7, helps remove the yearly cycles so that we can focus on the trend.

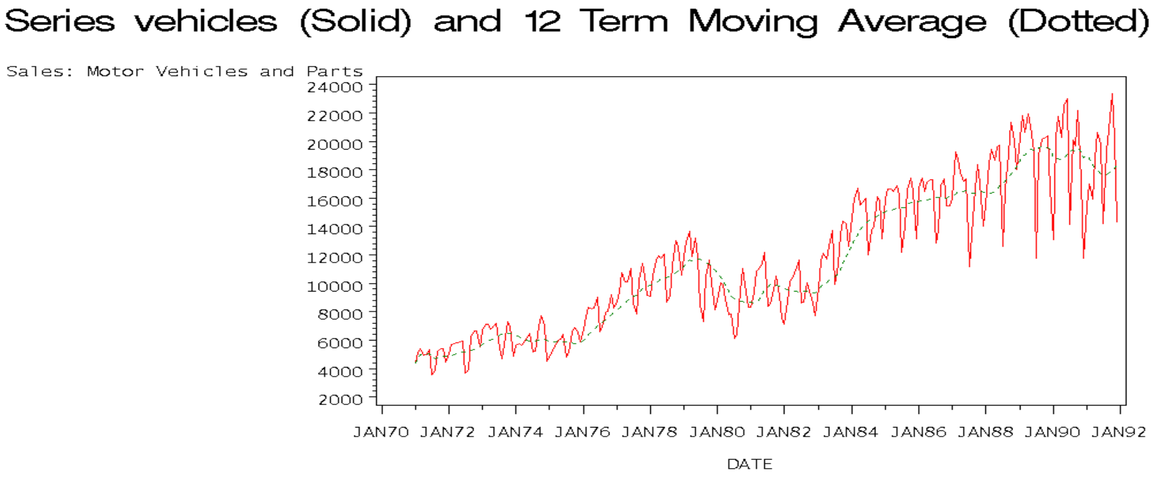

Figure 7: *Monthly sales of cars and car parts with 12 term moving average.*

Figure 8 is a time series plot of the weights of 150 bags of 1/2 ounce bags of potato chips taken from the production line every two minutes during a continuous five hour period. The plot shows a stationary pattern. The 10 term moving average supports the conclusion that the process is stationary.

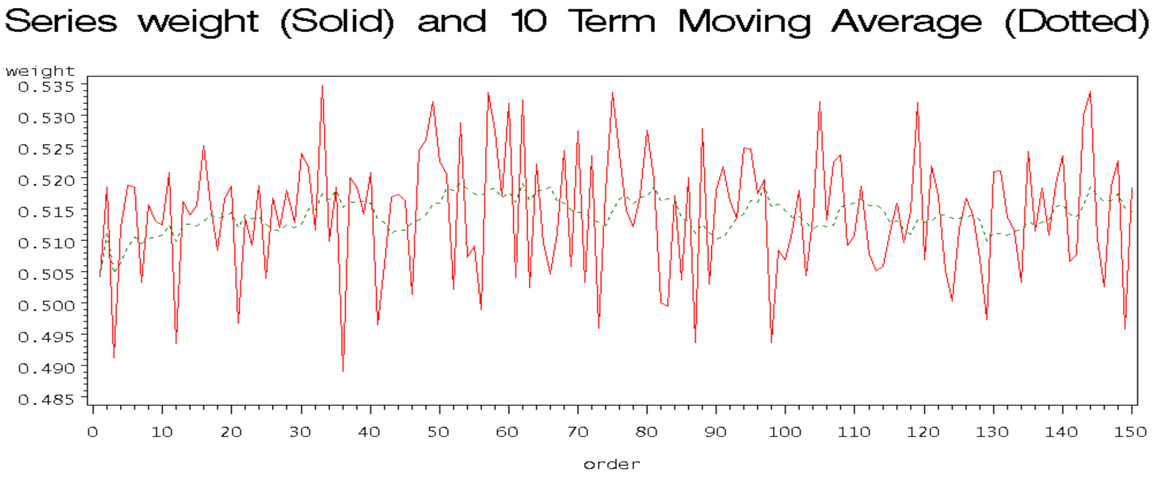

Figure 8: *Weights of half ounce bags of potato chips with 10 term moving average.*

What's the IDEA?

If data are taken over time, **ALWAYS** check stationarity by plotting versus time. If the pattern is nonstationary, displays and measures designed for data from a stationary process are probably inappropriate. A moving average can help detect nonstationary trends.

- Causes of Variation

  - o Common Causes: Systemic

  - o Special Causes: Periodic and unpredictable

The graph of the potato chip weights in Figure 8 shows only common causes of variation, since the pattern of variation remains the same throughout. The same is not true of the process graphed in Figure 9, which is a plot of the viscosity of 30 consecutive batches of paint primer. The large jump at batch 20 indicates a special cause of variation.

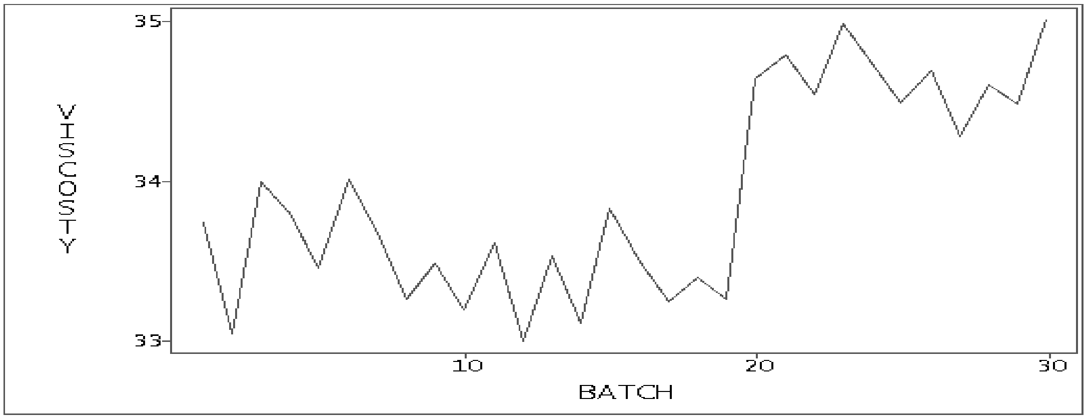

Figure 9: *Viscosity of 30 batches of paint primer.*

• Graphical Tools for Identifying Causes of Variation

- o Line or Time Series Plots are used to detect causes of variation through changes in the pattern of variation over time. Can be stratified to detect differences in different subgroups (machines, operators, etc)
- o Stratified Plots are used to detect causes of variation for stationary processes by breaking data into different subgroups.
- o Process Flow Diagrams show process steps and flow of information and material.
- o Ishikawa Diagrams organize ideas about causes of variation into categories.

The book has examples of each of these. The next application demonstrates one use of time series plots and stratified plots.

• Application: Gage R&R

- o Repeatability: consistency of gage (measuring tool) in repeated measurements by same operator on same part.
- o Reproducibility: variation between different operators.

In a gage R&R study of a laser ranging device, four operators each took fifteen measurements of the same distance. The data are in SASDATA.LASER and SASDATA.LASER0.

Figure 10 shows four time series plots, one for each operator, on the same graph. We say the plots are **stratified** by operator. None of these plots shows evidence of nonstationarity, which means we can analyze the pattern of variation without regard to order.

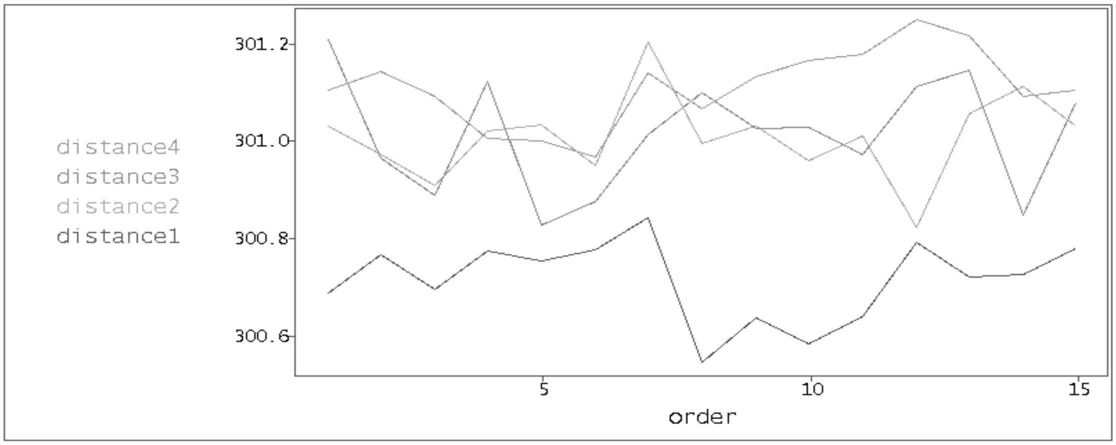

Figure 10: *Line plot of laser distance measurements stratified by four operators.*

Figure 11 is a plot of the measurements stratified by operator but ignoring order. We call this a **stratified plot**. The two most important things to look for in a stratified plot are differences in the spreads of the data for different strata (called **within variation**) and differences in the locations or centers of the data in the different strata (called **between variation**).

For a gage R&R study,

- The within variation measures the **repeatability** of the gage: the consistency of the gage in repeated measurements.
- The between variation measures the **reproducibility** of the measurement process: the consistency of measurements taken by different operators.

Consideration of the stratified plot in Figure 11, shows that all operators exhibit roughly the same repeatability, since the spreads of the measurements are roughly the same. It also shows that while the measurements of operators 2-4 are centered at roughly the same value, those of operator 1 are centered about a substantially lower value than the other three. This means there is a reproducibility problem with the measuring process.

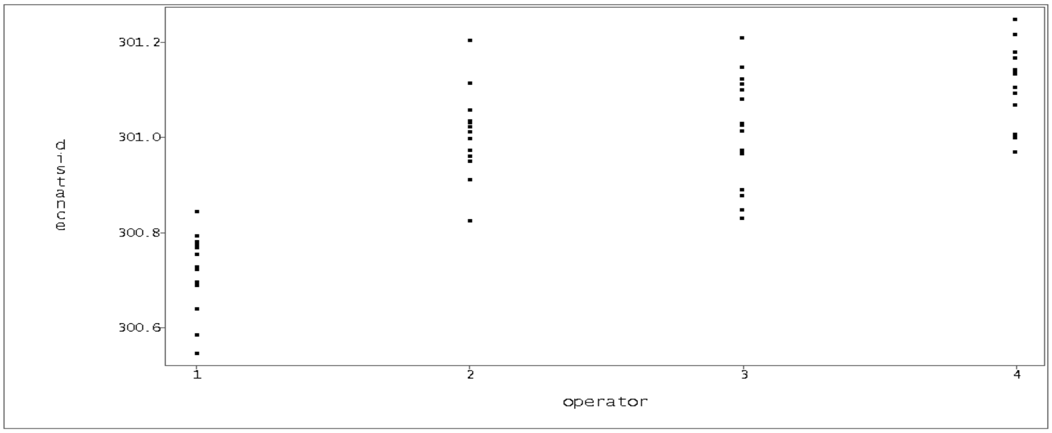

Figure 11: *Stratified plot of laser distance measurements by four operators.*

What's the IDEA?

Variation can sometimes be broken into pieces which identify different sources of the variation and the amount of variation each source contributes. One example is breaking variation into between and within components. Plots of data stratified by these sources can be helpful in analyzing the structure of variation.