1. Exploratory Data Analysis

This chapter presents the assumptions, principles, and techniques necessary to gain insight into data via EDA--exploratory data analysis.

1. Exploratory Data Analysis - Detailed Table of Contents [1.]

This chapter presents the assumptions, principles, and techniques necessary to gain insight into data via EDA--exploratory data analysis.

1.1. EDA Introduction

Summary

What is exploratory data analysis? How did it begin? How and where did it originate? How is it differentiated from other data analysis approaches, such as classical and Bayesian? Is EDA the same as statistical graphics? What role does statistical graphics play in EDA? Is statistical graphics identical to EDA?

These questions and related questions are dealt with in this section. This section answers these questions and provides the necessary frame of reference for EDA assumptions, principles, and techniques.

Table of Contents for Section 1

1.1.1. What is EDA?

Approach

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

1. maximize insight into a data set;
2. uncover underlying structure;
3. extract important variables;
4. detect outliers and anomalies;
5. test underlying assumptions;
6. develop parsimonious models; and
7. determine optimal factor settings.

Focus

The EDA approach is precisely that--an approach--not a set of techniques, but an attitude/philosophy about how a data analysis should be carried out.

Philosophy

EDA is not identical to statistical graphics although the two terms are used almost interchangeably. Statistical graphics is a collection of techniques--all graphically based and all focusing on one data characterization aspect. EDA encompasses a larger venue; EDA is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow with the more direct approach of allowing the data itself to reveal its underlying structure and model. EDA is not a mere collection of techniques; EDA is a philosophy as to how we dissect a data set; what we look for; how we look; and how we interpret. It is true that EDA heavily uses the collection of techniques that we call "statistical graphics", but it is not identical to statistical graphics per se.

History

The seminal work in EDA is Exploratory Data Analysis, Tukey, (1977). Over the years it has benefitted from other noteworthy publications such as Data Analysis and Regression, Mosteller and Tukey (1977), Interactive Data Analysis, Hoaglin (1977), The ABC's of EDA, Velleman and Hoaglin (1981) and has gained a large following as "the" way to analyze a data set.

Techniques

Most EDA techniques are graphical in nature with a few quantitative techniques. The reason for the heavy reliance on graphics is that by its very nature the main role of EDA is to open-mindedly explore, and graphics gives the analysts unparalleled power to do so, enticing the data to reveal its structural secrets, and being always ready to gain some new, often unsuspected, insight into the data. In combination with the natural pattern-recognition capabilities that we all possess, graphics provides, of course, unparalleled power to carry this out.

The particular graphical techniques employed in EDA are often quite simple, consisting of various techniques of:

1. Plotting the raw data (such as data traces, histograms, bihistograms, probability plots, lag plots, block plots, and Youden plots.
2. Plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.
3. Positioning such plots so as to maximize our natural pattern-recognition abilities, such as using multiple plots per page.

1.1.2. How Does Exploratory Data Analysis differ from Classical Data Analysis?

Data Analysis Approaches

EDA is a data analysis approach. What other data analysis approaches exist and how does EDA differ from these other approaches? Three popular data analysis approaches are:

1. Classical
2. Exploratory (EDA)
3. Bayesian

Paradigms for Analysis Techniques

These three approaches are similar in that they all start with a general science/engineering problem and all yield science/engineering conclusions. The difference is the sequence and focus of the intermediate steps.

For classical analysis, the sequence is

Problem => Data => Model => Analysis => Conclusions

For EDA, the sequence is

Problem => Data => Analysis => Model => Conclusions

For Bayesian, the sequence is

Problem => Data => Model => Prior Distribution => Analysis => Conclusions

Method of dealing with underlying model for the data distinguishes the 3 approaches

Thus for classical analysis, the data collection is followed by the imposition of a model (normality, linearity, etc.) and the analysis, estimation, and testing that follows are focused on the parameters of that model. For EDA, the data collection is not followed by a model imposition; rather it is followed immediately by analysis with a goal of inferring what model would be appropriate. Finally, for a Bayesian analysis, the analyst attempts to incorporate scientific/engineering knowledge/expertise into the analysis by imposing a data-independent distribution on the parameters of the selected model; the analysis thus consists of formally combining both the prior distribution on the parameters and the collected data to jointly make inferences and/or test assumptions about the model parameters.

In the real world, data analysts freely mix elements of all of the above three approaches (and other approaches). The above distinctions were made to emphasize the major differences among the three approaches.

Further discussion of the distinction between the classical and EDA approaches

Focusing on EDA versus classical, these two approaches differ as follows:

1. Models
2. Focus
3. Techniques
4. Rigor
5. Data Treatment
6. Assumptions

1. Exploratory Data Analysis
1.1. EDA Introduction
1.1.2. How Does Exploratory Data Analysis differ from Classical Data Analysis?

1.1.2.1. Model

Classical	The classical approach imposes models (both deterministic and probabilistic) on the data. Deterministic models include, for example, regression models and analysis of variance (ANOVA) models. The most common probabilistic model assumes that the errors about the deterministic model are normally distributed--this assumption affects the validity of the ANOVA F tests.
Exploratory	The Exploratory Data Analysis approach does not impose deterministic or probabilistic models on the data. On the contrary, the EDA approach allows the data to suggest admissible models that best fit the data.

1. Exploratory Data Analysis
1.1. EDA Introduction
1.1.2. How Does Exploratory Data Analysis differ from Classical Data Analysis?

1.1.2.2. Focus

Classical	The two approaches differ substantially in focus. For classical analysis, the focus is on the model--estimating parameters of the model and generating predicted values from the model.
Exploratory	For exploratory data analysis, the focus is on the data--its structure, outliers, and models suggested by the data.

1. Exploratory Data Analysis
1.1. EDA Introduction
1.1.2. How Does Exploratory Data Analysis differ from Classical Data Analysis?

1.1.2.3. Techniques

Classical — Classical techniques are generally quantitative in nature. They include ANOVA, t tests, chi-squared tests, and F tests.

Exploratory — EDA techniques are generally graphical. They include scatter plots, character plots, box plots, histograms, bihistograms, probability plots, residual plots, and mean plots.

1.1.2.4. Rigor

Classical	Classical techniques serve as the probabilistic foundation of science and engineering; the most important characteristic of classical techniques is that they are rigorous, formal, and "objective".
Exploratory	EDA techniques do not share in that rigor or formality. EDA techniques make up for that lack of rigor by being very suggestive, indicative, and insightful about what the appropriate model should be. EDA techniques are subjective and depend on interpretation which may differ from analyst to analyst, although experienced analysts commonly arrive at identical conclusions.

1. Exploratory Data Analysis
1.1. EDA Introduction
1.1.2. How Does Exploratory Data Analysis differ from Classical Data Analysis?

1.1.2.5. Data Treatment

Classical — Classical estimation techniques have the characteristic of taking all of the data and mapping the data into a few numbers ("estimates"). This is both a virtue and a vice. The virtue is that these few numbers focus on important characteristics (location, variation, etc.) of the population. The vice is that concentrating on these few characteristics can filter out other characteristics (skewness, tail length, autocorrelation, etc.) of the same population. In this sense there is a loss of information due to this "filtering" process.

Exploratory — The EDA approach, on the other hand, often makes use of (and shows) all of the available data. In this sense there is no corresponding loss of information.

1.1.2.6. Assumptions

Classical The "good news" of the classical approach is that tests based on classical techniques are usually very sensitive--that is, if a true shift in location, say, has occurred, such tests frequently have the power to detect such a shift and to conclude that such a shift is "statistically significant". The "bad news" is that classical tests depend on underlying assumptions (e.g., normality), and hence the validity of the test conclusions becomes dependent on the validity of the underlying assumptions. Worse yet, the exact underlying assumptions may be unknown to the analyst, or if known, untested. Thus the validity of the scientific conclusions becomes intrinsically linked to the validity of the underlying assumptions. In practice, if such assumptions are unknown or untested, the validity of the scientific conclusions becomes suspect.

Exploratory Many EDA techniques make little or no assumptions--they present and show the data--all of the data--as is, with fewer encumbering assumptions.

1. Exploratory Data Analysis
1.1. EDA Introduction

1.1.3. How Does Exploratory Data Analysis Differ from Summary Analysis?

Summary A summary analysis is simply a numeric reduction of a historical data set. It is quite passive. Its focus is in the past. Quite commonly, its purpose is to simply arrive at a few key statistics (for example, mean and standard deviation) which may then either replace the data set or be added to the data set in the form of a summary table.

Exploratory In contrast, EDA has as its broadest goal the desire to gain insight into the engineering/scientific process behind the data. Whereas summary statistics are passive and historical, EDA is active and futuristic. In an attempt to "understand" the process and improve it in the future, EDA uses the data as a "window" to peer into the heart of the process that generated the data. There is an archival role in the research and manufacturing world for summary statistics, but there is an enormously larger role for the EDA approach.

1. Exploratory Data Analysis
1.1. EDA Introduction

1.1.4. What are the EDA Goals?

Primary and Secondary Goals

The primary goal of EDA is to maximize the analyst's insight into a data set and into the underlying structure of a data set, while providing all of the specific items that an analyst would want to extract from a data set, such as:

1. a good-fitting, parsimonious model
2. a list of outliers
3. a sense of robustness of conclusions
4. estimates for parameters
5. uncertainties for those estimates
6. a ranked list of important factors
7. conclusions as to whether individual factors are statistically significant
8. optimal settings

Insight into the Data

Insight implies detecting and uncovering underlying structure in the data. Such underlying structure may not be encapsulated in the list of items above; such items serve as the specific targets of an analysis, but the real insight and "feel" for a data set comes as the analyst judiciously probes and explores the various subtleties of the data. The "feel" for the data comes almost exclusively from the application of various graphical techniques, the collection of which serves as the window into the essence of the data. Graphics are irreplaceable--there are no quantitative analogues that will give the same insight as well-chosen graphics.

To get a "feel" for the data, it is not enough for the analyst to know what is in the data; the analyst also must know what is not in the data, and the only way to do that is to draw on our own human pattern-recognition and comparative abilities in the context of a series of judicious graphical techniques applied to the data.

1.1.5. The Role of Graphics

Quantitative/ Graphical

Statistics and data analysis procedures can broadly be split into two parts:

- quantitative
- graphical

Quantitative

Quantitative techniques are the set of statistical procedures that yield numeric or tabular output. Examples of quantitative techniques include:

- hypothesis testing
- analysis of variance
- point estimates and confidence intervals
- least squares regression

These and similar techniques are all valuable and are mainstream in terms of classical analysis.

Graphical

On the other hand, there is a large collection of statistical tools that we generally refer to as graphical techniques. These include:

- scatter plots
- histograms
- probability plots
- residual plots
- box plots
- block plots

EDA Approach Relies Heavily on Graphical Techniques

The EDA approach relies heavily on these and similar graphical techniques. Graphical procedures are not just tools that we could use in an EDA context, they are tools that we must use. Such graphical tools are the shortest path to gaining insight into a data set in terms of

- testing assumptions
- model selection
- model validation
- estimator selection
- relationship identification
- factor effect determination
- outlier detection

If one is not using statistical graphics, then one is forfeiting insight into one or more aspects of the underlying structure of the data.

1.1.6. An EDA/Graphics Example

Anscombe Example

A simple, classic (Anscombe) example of the central role that graphics play in terms of providing insight into a data set starts with the following data set:

Data

X	Y
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.84
7.00	4.82
5.00	5.68

Summary Statistics

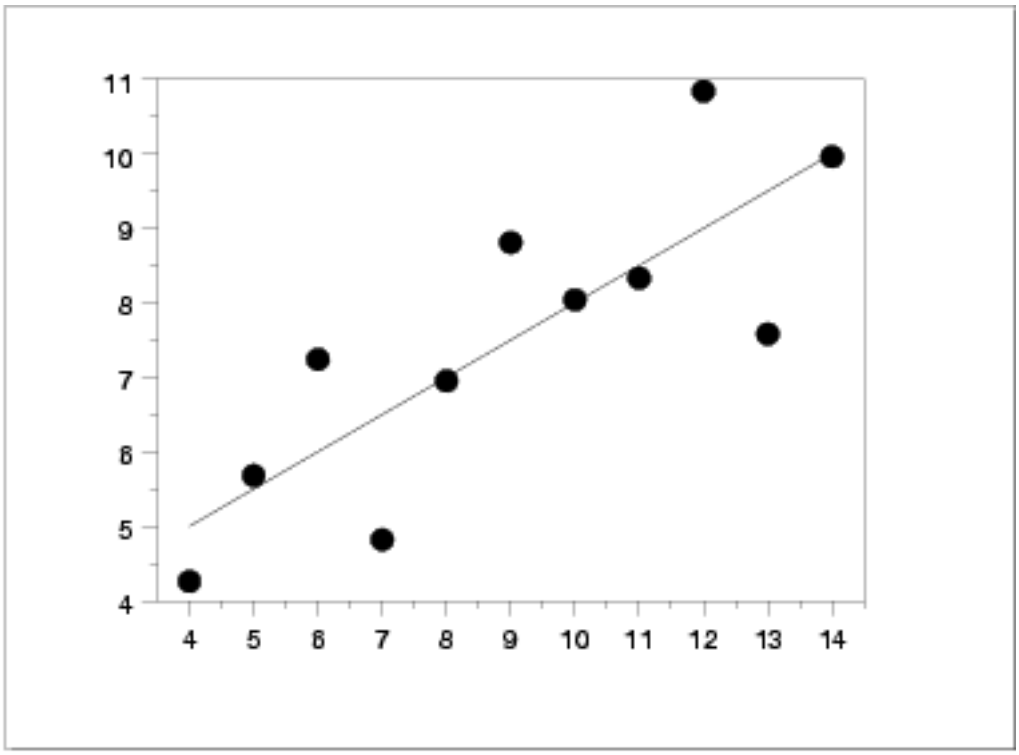
If the goal of the analysis is to compute summary statistics plus determine the best linear fit for Y as a function of X, the results might be given as:

N = 11
Mean of X = 9.0
Mean of Y = 7.5
Intercept = 3
Slope = 0.5
Residual standard deviation = 1.237
Correlation = 0.816

The above quantitative analysis, although valuable, gives us only limited insight into the data.

Scatter Plot

In contrast, the following simple scatter plot of the data



suggests the following:

1. The data set "behaves like" a linear curve with some scatter;
2. there is no justification for a more complicated model (e.g., quadratic);
3. there are no outliers;
4. the vertical spread of the data appears to be of equal height irrespective of the *X*-value; this indicates that the data are equally-precise throughout and so a "regular" (that is, equi-weighted) fit is appropriate.

Three Additional Data Sets

This kind of characterization for the data serves as the core for getting insight/feel for the data. Such insight/feel does not come from the quantitative statistics; on the contrary, calculations of quantitative statistics such as intercept and slope should be subsequent to the characterization and will make sense only if the characterization is true. To illustrate the loss of information that results when the graphics insight step is skipped, consider the following three data sets [Anscombe data sets 2, 3, and 4]:

X2	Y2	X3	Y3	X4	Y4
10.00	9.14	10.00	7.46	8.00	6.58
8.00	8.14	8.00	6.77	8.00	5.76
13.00	8.74	13.00	12.74	8.00	7.71

9.00	8.77	9.00	7.11	8.00	8.84
11.00	9.26	11.00	7.81	8.00	8.47
14.00	8.10	14.00	8.84	8.00	7.04
6.00	6.13	6.00	6.08	8.00	5.25
4.00	3.10	4.00	5.39	19.00	12.50
12.00	9.13	12.00	8.15	8.00	5.56
7.00	7.26	7.00	6.42	8.00	7.91
5.00	4.74	5.00	5.73	8.00	6.89

Quantitative Statistics for Data Set 2

A quantitative analysis on data set 2 yields

N = 11
Mean of X = 9.0
Mean of Y = 7.5
Intercept = 3
Slope = 0.5
Residual standard deviation = 1.237
Correlation = 0.816

which is identical to the analysis for data set 1. One might naively assume that the two data sets are "equivalent" since that is what the statistics tell us; but what do the statistics not tell us?

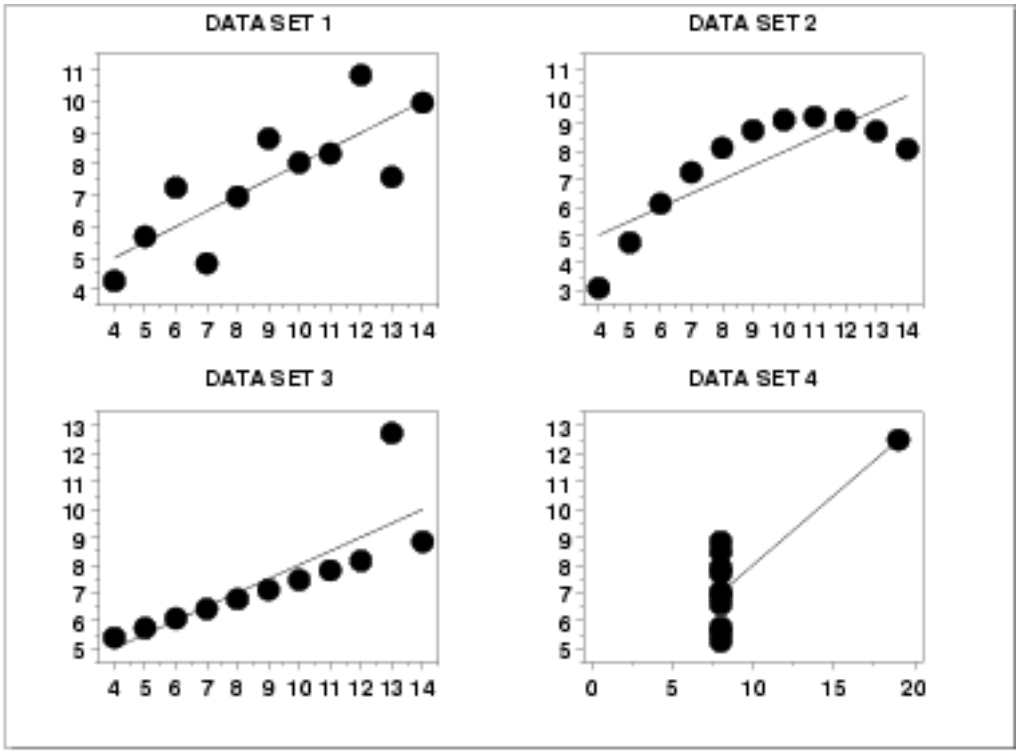
Quantitative Statistics for Data Sets 3 and 4

Remarkably, a quantitative analysis on data sets 3 and 4 also yields

N = 11
Mean of X = 9.0
Mean of Y = 7.5
Intercept = 3
Slope = 0.5
Residual standard deviation = 1.236
Correlation = 0.816 (0.817 for data set 4)

which implies that in some quantitative sense, all four of the data sets are "equivalent". In fact, the four data sets are far from "equivalent" and a scatter plot of each data set, which would be step 1 of any EDA approach, would tell us that immediately.

Scatter Plots



Interpretation of Scatter Plots

Conclusions from the scatter plots are:

1. data set 1 is clearly linear with some scatter.
2. data set 2 is clearly quadratic.
3. data set 3 clearly has an outlier.
4. data set 4 is obviously the victim of a poor experimental design with a single point far removed from the bulk of the data "wagging the dog".

Importance of Exploratory Analysis

These points are exactly the substance that provide and define "insight" and "feel" for a data set. They are the goals and the fruits of an open exploratory data analysis (EDA) approach to the data. Quantitative statistics are not wrong per se, but they are incomplete. They are incomplete because they are numeric **summaries** which in the summarization operation do a good job of focusing on a particular aspect of the data (e.g., location, intercept, slope, degree of relatedness, etc.) by judiciously reducing the data to a few numbers. Doing so also **filters** the data, necessarily omitting and screening out other sometimes crucial information in the focusing operation. Quantitative statistics focus but also filter; and filtering is exactly what makes the quantitative approach incomplete at best and misleading at worst.

The estimated intercepts (= 3) and slopes (= 0.5) for data sets 2, 3, and 4 are misleading because the estimation is done in the context of an assumed linear model and that linearity assumption is the fatal flaw in this analysis.

The EDA approach of deliberately postponing the model selection until further along in the analysis has many rewards, not the least of which is the ultimate convergence to a much-improved model and the formulation of valid and supportable scientific and engineering conclusions.

1.1.7. General Problem Categories

Problem Classification

The following table is a convenient way to classify EDA problems.

Univariate and Control

UNIVARIATE	CONTROL
Data: A single column of numbers, *Y*. Model: *y* = constant + error Output: 1. A number (the estimated constant in the model). 2. An estimate of uncertainty for the constant. 3. An estimate of the distribution for the error. Techniques: • 4-Plot • Probability Plot • PPCC Plot	Data: A single column of numbers, *Y*. Model: *y* = constant + error Output: A "yes" or "no" to the question "Is the system out of control?". Techniques: • Control Charts

Comparative and Screening

COMPARATIVE	SCREENING
Data: A single response variable and k independent variables (Y, X_1, X_2, ... , X_k), primary focus is on *one* (the primary factor) of these independent variables. Model: $y = f(x_1, x_2, ..., x_k) + error$ Output: A "yes" or "no" to the question "Is the primary factor significant?". Techniques: ● Block Plot ● Scatter Plot ● Box Plot	Data: A single response variable and k independent variables (Y, X_1, X_2, ... , X_k). Model: $y = f(x_1, x_2, ..., x_k) + error$ Output: 1. A ranked list (from most important to least important) of factors. 2. Best settings for the factors. 3. A good model/prediction equation relating Y to the factors. Techniques: ● Block Plot ● Probability Plot ● Bihistogram

Optimization and Regression

OPTIMIZATION	REGRESSION
Data: A single response variable and k independent variables (Y, X_1, X_2, ... , X_k). Model: $y = f(x_1, x_2, ..., x_k) + error$ Output: Best settings for the factor variables. Techniques: ● Block Plot	Data: A single response variable and k independent variables (Y, X_1, X_2, ... , X_k). The independent variables can be continuous. Model: $y = f(x_1, x_2, ..., x_k) + error$ Output: A good model/prediction equation relating Y to the factors.

• Least Squares Fitting • Contour Plot	Techniques: • Least Squares Fitting • Scatter Plot • 6-Plot

Time Series and Multivariate

TIME SERIES	MULTIVARIATE
Data: A column of time dependent numbers, Y. In addition, time is an indpendent variable. The time variable can be either explicit or implied. If the data are not equi-spaced, the time variable should be explicitly provided. Model: $y_t = f(t)$ + error The model can be either a time domain based or frequency domain based. Output: A good model/prediction equation relating Y to previous values of Y. Techniques: • Autocorrelation Plot • Spectrum • Complex Demodulation Amplitude Plot • Complex Demodulation Phase Plot • ARIMA Models	Data: k factor variables (X_1, X_2, ... , X_k). Model: The model is not explicit. Output: Identify underlying correlation structure in the data. Techniques: • Star Plot • Scatter Plot Matrix • Conditioning Plot • Profile Plot • Principal Components • Clustering • Discrimination/Classification Note that multivarate analysis is only covered lightly in this Handbook.

1. Exploratory Data Analysis

1.2. EDA Assumptions

Summary

The gamut of scientific and engineering experimentation is virtually limitless. In this sea of diversity is there any common basis that allows the analyst to systematically and validly arrive at supportable, repeatable research conclusions?

Fortunately, there is such a basis and it is rooted in the fact that every measurement process, however complicated, has certain underlying assumptions. This section deals with what those assumptions are, why they are important, how to go about testing them, and what the consequences are if the assumptions do not hold.

Table of Contents for Section 2

1.2.1. Underlying Assumptions

Assumptions Underlying a Measurement Process

There are four assumptions that typically underlie all measurement processes; namely, that the data from the process at hand "behave like":

1. random drawings;
2. from a fixed distribution;
3. with the distribution having fixed location; and
4. with the distribution having fixed variation.

Univariate or Single Response Variable

The "fixed location" referred to in item 3 above differs for different problem types. The simplest problem type is univariate; that is, a single variable. For the univariate problem, the general model

response = deterministic component + random component

becomes

response = constant + error

Assumptions for Univariate Model

For this case, the "fixed location" is simply the unknown constant. We can thus imagine the process at hand to be operating under constant conditions that produce a single column of data with the properties that

- the data are uncorrelated with one another;
- the random component has a fixed distribution;
- the deterministic component consists of only a constant; and
- the random component has fixed variation.

Extrapolation to a Function of Many Variables

The universal power and importance of the univariate model is that it can easily be extended to the more general case where the deterministic component is not just a constant, but is in fact a function of many variables, and the engineering objective is to characterize and model the function.

Residuals Will Behave According to Univariate Assumptions

The key point is that regardless of how many factors there are, and regardless of how complicated the function is, if the engineer succeeds in choosing a good model, then the differences (residuals) between the raw response data and the predicted values from the fitted model should themselves behave like a univariate process. Furthermore, the residuals from this univariate process fit will behave like:

- random drawings;
- from a fixed distribution;
- with fixed location (namely, 0 in this case); and
- with fixed variation.

Validation of Model

Thus if the residuals from the fitted model do in fact behave like the ideal, then testing of underlying assumptions becomes a tool for the validation and quality of fit of the chosen model. On the other hand, if the residuals from the chosen fitted model violate one or more of the above univariate assumptions, then the chosen fitted model is inadequate and an opportunity exists for arriving at an improved model.

1.2.2. Importance

Predictability and Statistical Control	Predictability is an all-important goal in science and engineering. If the four underlying assumptions hold, then we have achieved probabilistic predictability--the ability to make probability statements not only about the process in the past, but also about the process in the future. In short, such processes are said to be "in statistical control".
Validity of Engineering Conclusions	Moreover, if the four assumptions are valid, then the process is amenable to the generation of valid scientific and engineering conclusions. If the four assumptions are not valid, then the process is drifting (with respect to location, variation, or distribution), unpredictable, and out of control. A simple characterization of such processes by a location estimate, a variation estimate, or a distribution "estimate" inevitably leads to engineering conclusions that are not valid, are not supportable (scientifically or legally), and which are not repeatable in the laboratory.

1.2.3. Techniques for Testing Assumptions

Testing Underlying Assumptions Helps Assure the Validity of Scientific and Engineering Conclusions

Because the validity of the final scientific/engineering conclusions is inextricably linked to the validity of the underlying univariate assumptions, it naturally follows that there is a real necessity that each and every one of the above four assumptions be routinely tested.

Four Techniques to Test Underlying Assumptions

The following EDA techniques are simple, efficient, and powerful for the routine testing of underlying assumptions:

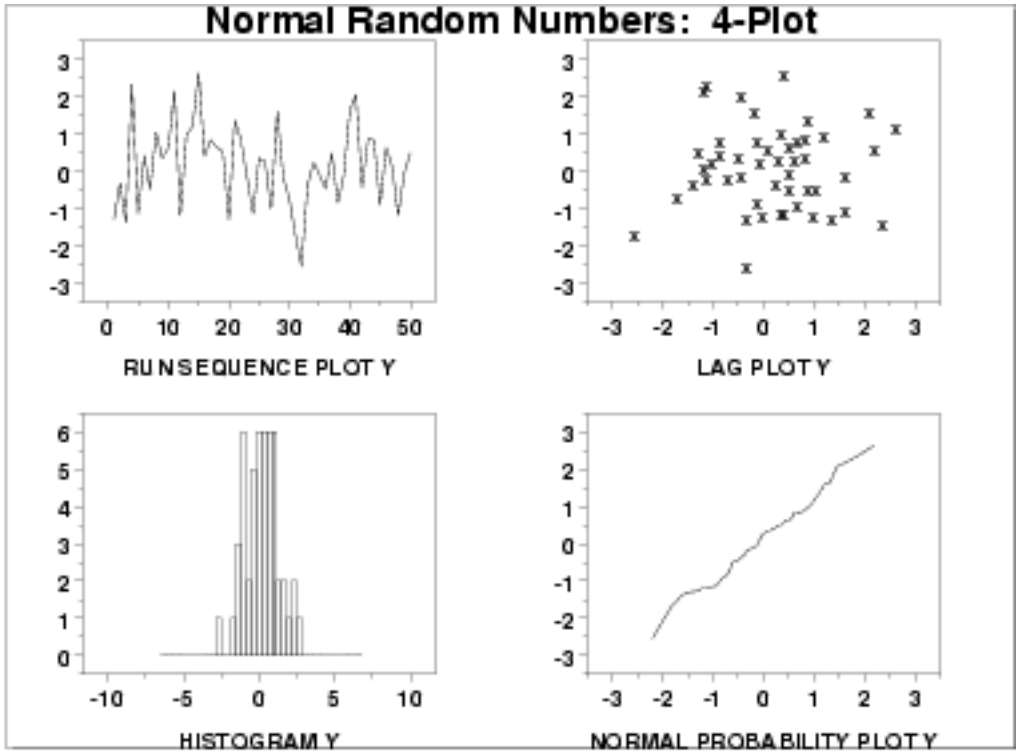
1. run sequence plot (Y_i versus i)
2. lag plot (Y_i versus Y_{i-1})
3. histogram (counts versus subgroups of Y)
4. normal probability plot (ordered Y versus theoretical ordered Y)

Plot on a Single Page for a Quick Characterization of the Data

The four EDA plots can be juxtaposed for a quick look at the characteristics of the data. The plots below are ordered as follows:

1. Run sequence plot - upper left
2. Lag plot - upper right
3. Histogram - lower left
4. Normal probability plot - lower right

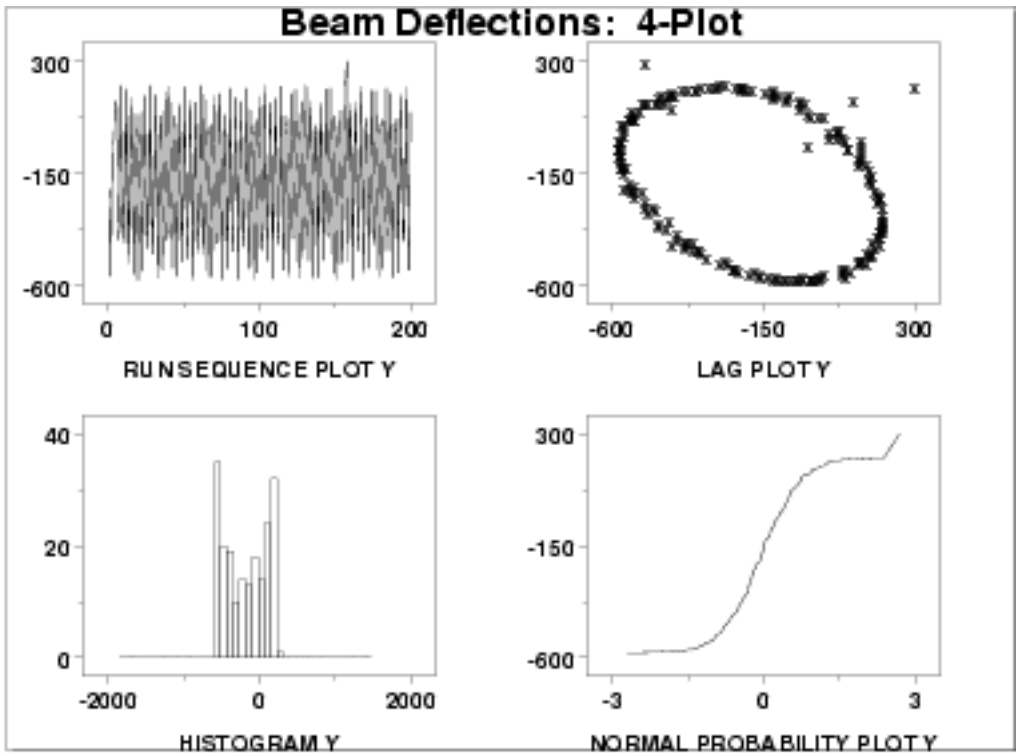
Sample Plot: Assumptions Hold



This 4-plot reveals a process that has fixed location, fixed variation, is random, apparently has a fixed approximately normal distribution, and has no outliers.

Sample Plot: Assumptions Do Not Hold

If one or more of the four underlying assumptions do not hold, then it will show up in the various plots as demonstrated in the following example.



This 4-plot reveals a process that has fixed location, fixed variation, is non-random (oscillatory), has a non-normal, U-shaped distribution, and has several outliers.

1.2.4. Interpretation of 4-Plot

Interpretation of EDA Plots: Flat and Equi-Banded, Random, Bell-Shaped, and Linear

The four EDA plots discussed on the previous page are used to test the underlying assumptions:

1. **Fixed Location:**
   If the fixed location assumption holds, then the run sequence plot will be flat and non-drifting.
2. **Fixed Variation:**
   If the fixed variation assumption holds, then the vertical spread in the run sequence plot will be the approximately the same over the entire horizontal axis.
3. **Randomness:**
   If the randomness assumption holds, then the lag plot will be structureless and random.
4. **Fixed Distribution:**
   If the fixed distribution assumption holds, in particular if the fixed normal distribution holds, then
   1. the histogram will be bell-shaped, and
   2. the normal probability plot will be linear.

Plots Utilized to Test the Assumptions

Conversely, the underlying assumptions are tested using the EDA plots:

- **Run Sequence Plot:**
  If the run sequence plot is flat and non-drifting, the fixed-location assumption holds. If the run sequence plot has a vertical spread that is about the same over the entire plot, then the fixed-variation assumption holds.
- **Lag Plot:**
  If the lag plot is structureless, then the randomness assumption holds.
- **Histogram:**
  If the histogram is bell-shaped, the underlying distribution is symmetric and perhaps approximately normal.
- **Normal Probability Plot:**

If the normal probability plot is linear, the underlying distribution is approximately normal.

If all four of the assumptions hold, then the process is said definitionally to be "in statistical control".

1.2.5. Consequences

What If Assumptions Do Not Hold?

If some of the underlying assumptions do not hold, what can be done about it? What corrective actions can be taken? The positive way of approaching this is to view the testing of underlying assumptions as a framework for learning about the process. Assumption-testing promotes insight into important aspects of the process that may not have surfaced otherwise.

Primary Goal is Correct and Valid Scientific Conclusions

The primary goal is to have correct, validated, and complete scientific/engineering conclusions flowing from the analysis. This usually includes intermediate goals such as the derivation of a good-fitting model and the computation of realistic parameter estimates. It should always include the ultimate goal of an understanding and a "feel" for "what makes the process tick". There is no more powerful catalyst for discovery than the bringing together of an experienced/expert scientist/engineer and a data set ripe with intriguing "anomalies" and characteristics.

Consequences of Invalid Assumptions

The following sections discuss in more detail the consequences of invalid assumptions:

1. Consequences of non-randomness
2. Consequences of non-fixed location parameter
3. Consequences of non-fixed variation
4. Consequences related to distributional assumptions

1.2.5.1. Consequences of Non-Randomness

Randomness Assumption

There are four underlying assumptions:

1. randomness;
2. fixed location;
3. fixed variation; and
4. fixed distribution.

The randomness assumption is the most critical but the least tested.

Consequeces of Non-Randomness

If the randomness assumption does not hold, then

1. All of the usual statistical tests are invalid.
2. The calculated uncertainties for commonly used statistics become meaningless.
3. The calculated minimal sample size required for a pre-specified tolerance becomes meaningless.
4. The simple model: y = constant + error becomes invalid.
5. The parameter estimates become suspect and non-supportable.

Non-Randomness Due to Autocorrelation

One specific and common type of non-randomness is autocorrelation. Autocorrelation is the correlation between Y_t and Y_{t-k}, where k is an integer that defines the lag for the autocorrelation. That is, autocorrelation is a time dependent non-randomness. This means that the value of the current point is highly dependent on the previous point if $k = 1$ (or k points ago if k is not 1). Autocorrelation is typically detected via an autocorrelation plot or a lag plot.

If the data are not random due to autocorrelation, then

1. Adjacent data values may be related.
2. There may not be n independent snapshots of the phenomenon under study.

3. There may be undetected "junk"-outliers.
4. There may be undetected "information-rich"-outliers.

1.2.5.2. Consequences of Non-Fixed Location Parameter

Location Estimate

The usual estimate of location is the mean

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

from N measurements Y_1, Y_2, ... , Y_N.

Consequences of Non-Fixed Location

If the run sequence plot does not support the assumption of fixed location, then

1. The location may be drifting.
2. The single location estimate may be meaningless (if the process is drifting).
3. The choice of location estimator (e.g., the sample mean) may be sub-optimal.
4. The usual formula for the uncertainty of the mean:

$$s(\bar{Y}) = \frac{1}{\sqrt{N(N-1)}}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

   may be invalid and the numerical value optimistically small.
5. The location estimate may be poor.
6. The location estimate may be biased.

1.2.5.3. Consequences of Non-Fixed Variation Parameter

Variation Estimate

The usual estimate of variation is the standard deviation

$$s_Y = \frac{1}{\sqrt{(N-1)}}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

from N measurements $Y_1, Y_2, \ldots, Y_N$.

Consequences of Non-Fixed Variation

If the run sequence plot does not support the assumption of fixed variation, then

1. The variation may be drifting.
2. The single variation estimate may be meaningless (if the process variation is drifting).
3. The variation estimate may be poor.
4. The variation estimate may be biased.

1.2.5.4. Consequences Related to Distributional Assumptions

Distributional Analysis

Scientists and engineers routinely use the mean (average) to estimate the "middle" of a distribution. It is not so well known that the variability and the noisiness of the mean as a location estimator are intrinsically linked with the underlying distribution of the data. For certain distributions, the mean is a poor choice. For any given distribution, there exists an optimal choice-- that is, the estimator with minimum variability/noisiness. This optimal choice may be, for example, the median, the midrange, the midmean, the mean, or something else. The implication of this is to "estimate" the distribution first, and then--based on the distribution--choose the optimal estimator. The resulting engineering parameter estimators will have less variability than if this approach is not followed.

Case Studies

The airplane glass failure case study gives an example of determining an appropriate distribution and estimating the parameters of that distribution. The uniform random numbers case study gives an example of determining a more appropriate centrality parameter for a non-normal distribution.

Other consequences that flow from problems with distributional assumptions are:

Distribution

1. The distribution may be changing.
2. The single distribution estimate may be meaningless (if the process distribution is changing).
3. The distribution may be markedly non-normal.
4. The distribution may be unknown.
5. The true probability distribution for the error may remain unknown.

Model

1. The model may be changing.
2. The single model estimate may be meaningless.
3. The default model

   $$Y = \text{constant} + \text{error}$$

   may be invalid.
4. If the default model is insufficient, information about a better model may remain undetected.
5. A poor deterministic model may be fit.
6. Information about an improved model may go undetected.

Process

1. The process may be out-of-control.
2. The process may be unpredictable.
3. The process may be un-modelable.

1. Exploratory Data Analysis

1.3. EDA Techniques

Summary

After you have collected a set of data, how do you do an exploratory data analysis? What techniques do you employ? What do the various techniques focus on? What conclusions can you expect to reach?

This section provides answers to these kinds of questions via a gallery of EDA techniques and a detailed description of each technique. The techniques are divided into graphical and quantitative techniques. For exploratory data analysis, the emphasis is primarily on the graphical techniques.

Table of Contents for Section 3

1. Exploratory Data Analysis
1.3. EDA Techniques

1.3.1. Introduction

Graphical and Quantitative Techniques

This section describes many techniques that are commonly used in exploratory and classical data analysis. This list is by no means meant to be exhaustive. Additional techniques (both graphical and quantitative) are discussed in the other chapters. Specifically, the product comparisons chapter has a much more detailed description of many classical statistical techniques.

EDA emphasizes graphical techniques while classical techniques emphasize quantitative techniques. In practice, an analyst typically uses a mixture of graphical and quantitative techniques. In this section, we have divided the descriptions into graphical and quantitative techniques. This is for organizational clarity and is not meant to discourage the use of both graphical and quantitiative techniques when analyzing data.

Use of Techniques Shown in Case Studies

This section emphasizes the techniques themselves; how the graph or test is defined, published references, and sample output. The use of the techniques to answer engineering questions is demonstrated in the case studies section. The case studies do not demonstrate all of the techniques.

Availability in Software

The sample plots and output in this section were generated with the Dataplot software program. Other general purpose statistical data analysis programs can generate most of the plots, intervals, and tests discussed here, or macros can be written to acheive the same result.

1.3.2. Analysis Questions

EDA Questions

Some common questions that exploratory data analysis is used to answer are:

1. What is a typical value?
2. What is the uncertainty for a typical value?
3. What is a good distributional fit for a set of numbers?
4. What is a percentile?
5. Does an engineering modification have an effect?
6. Does a factor have an effect?
7. What are the most important factors?
8. Are measurements coming from different laboratories equivalent?
9. What is the best function for relating a response variable to a set of factor variables?
10. What are the best settings for factors?
11. Can we separate signal from noise in time dependent data?
12. Can we extract any structure from multivariate data?
13. Does the data have outliers?

Analyst Should Identify Relevant Questions for his Engineering Problem

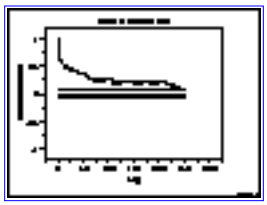
A critical early step in any analysis is to identify (for the engineering problem at hand) which of the above questions are relevant. That is, we need to identify which questions we want answered and which questions have no bearing on the problem at hand. After collecting such a set of questions, an equally important step, which is invaluable for maintaining focus, is to prioritize those questions in decreasing order of importance. EDA techniques are tied in with each of the questions. There are some EDA techniques (e.g., the scatter plot) that are broad-brushed and apply almost universally. On the other hand, there are a large number of EDA techniques that are specific and whose specificity is tied in with one of the above questions. Clearly if one chooses not to explicitly identify relevant questions, then one cannot take advantage of these question-specific EDA technqiues.

EDA Approach Emphasizes Graphics

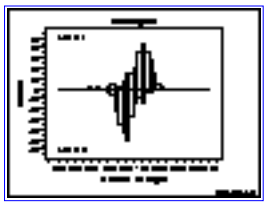
Most of these questions can be addressed by techniques discussed in this chapter. The process modeling and process improvement chapters also address many of the questions above. These questions are also relevant for the classical approach to statistics. What distinguishes the EDA approach is an emphasis on graphical techniques to gain insight as opposed to the classical approach of quantitative tests. Most data analysts will use a mix of graphical and classical quantitative techniques to address these problems.

1.3.3. Graphical Techniques: Alphabetic

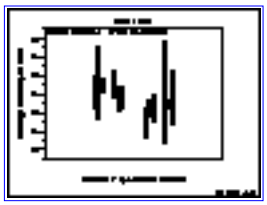
This section provides a gallery of some useful graphical techniques. The techniques are ordered alphabetically, so this section is not intended to be read in a sequential fashion. The use of most of these graphical techniques is demonstrated in the case studies in this chapter. A few of these graphical techniques are demonstrated in later chapters.



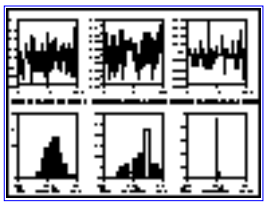
Autocorrelation Plot: 1.3.3.1



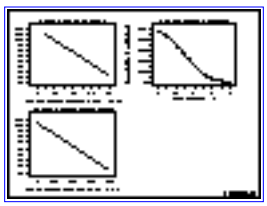
Bihistogram: 1.3.3.2



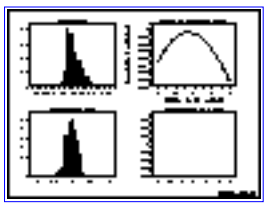
Block Plot: 1.3.3.3



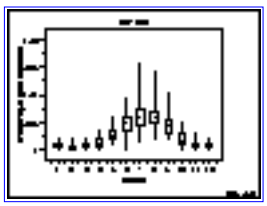
Bootstrap Plot: 1.3.3.4



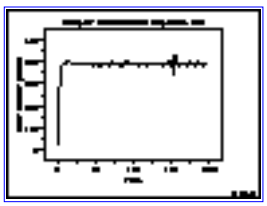
Box-Cox Linearity Plot: 1.3.3.5



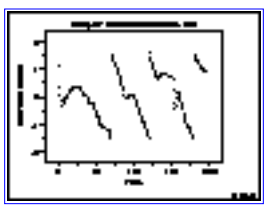
Box-Cox Normality Plot: 1.3.3.6



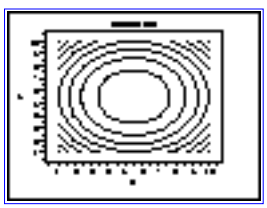
Box Plot: 1.3.3.7



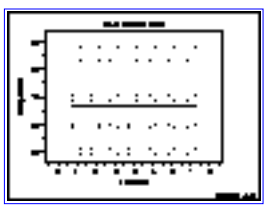
Complex Demodulation Amplitude Plot: 1.3.3.8



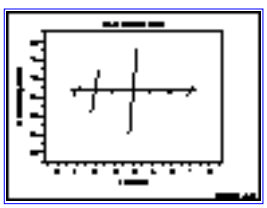
Complex Demodulation Phase Plot: 1.3.3.9



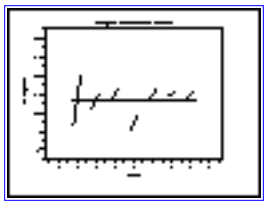
Contour Plot: 1.3.3.10



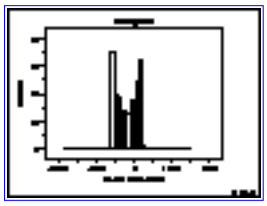
DEX Scatter Plot: 1.3.3.11



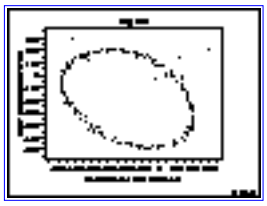
DEX Mean Plot: 1.3.3.12



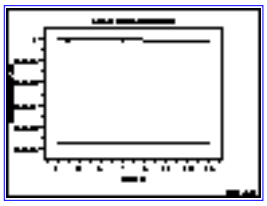
DEX Standard Deviation Plot: 1.3.3.13



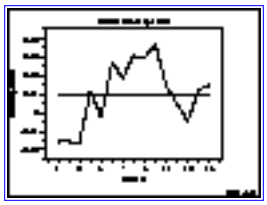
Histogram: 1.3.3.14



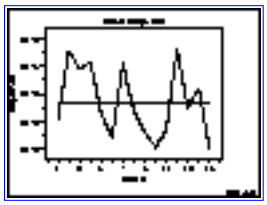
Lag Plot: 1.3.3.15



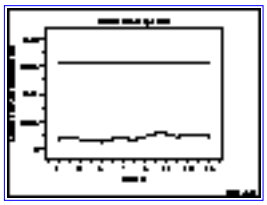
Linear Correlation Plot: 1.3.3.16



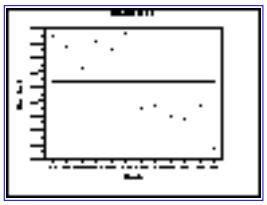
Linear Intercept Plot: 1.3.3.17



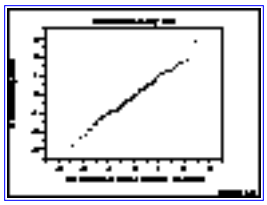
Linear Slope Plot: 1.3.3.18



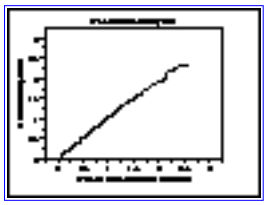
Linear Residual Standard Deviation Plot: 1.3.3.19



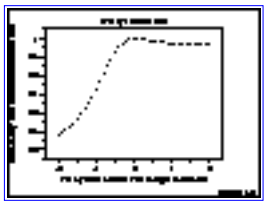
Mean Plot: 1.3.3.20



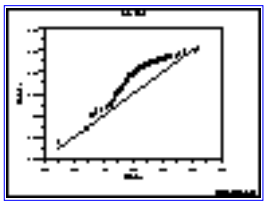
Normal Probability Plot: 1.3.3.21



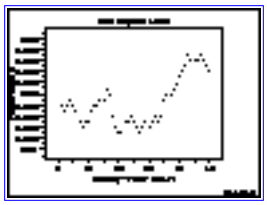
Probability Plot: 1.3.3.22



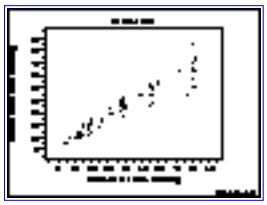
Probability Plot Correlation Coefficient Plot: 1.3.3.23



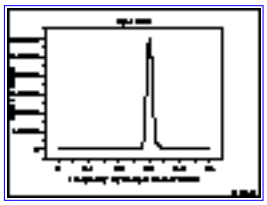
Quantile-Quantile Plot: 1.3.3.24



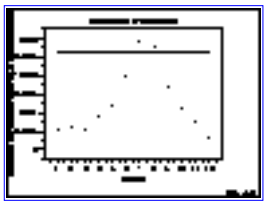
Run Sequence Plot: 1.3.3.25



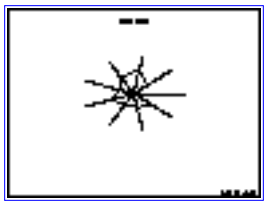
Scatter Plot: 1.3.3.26



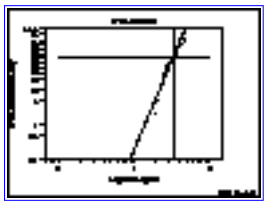
Spectrum: 1.3.3.27



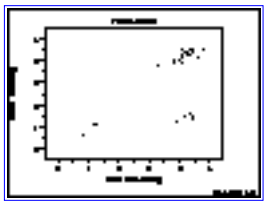
Standard Deviation Plot: 1.3.3.28



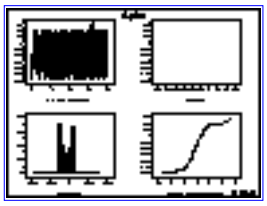
Star Plot: 1.3.3.29



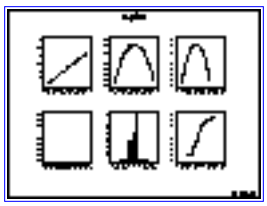
Weibull Plot: 1.3.3.30



Youden Plot: 1.3.3.31



4-Plot: 1.3.3.32



6-Plot: 1.3.3.33

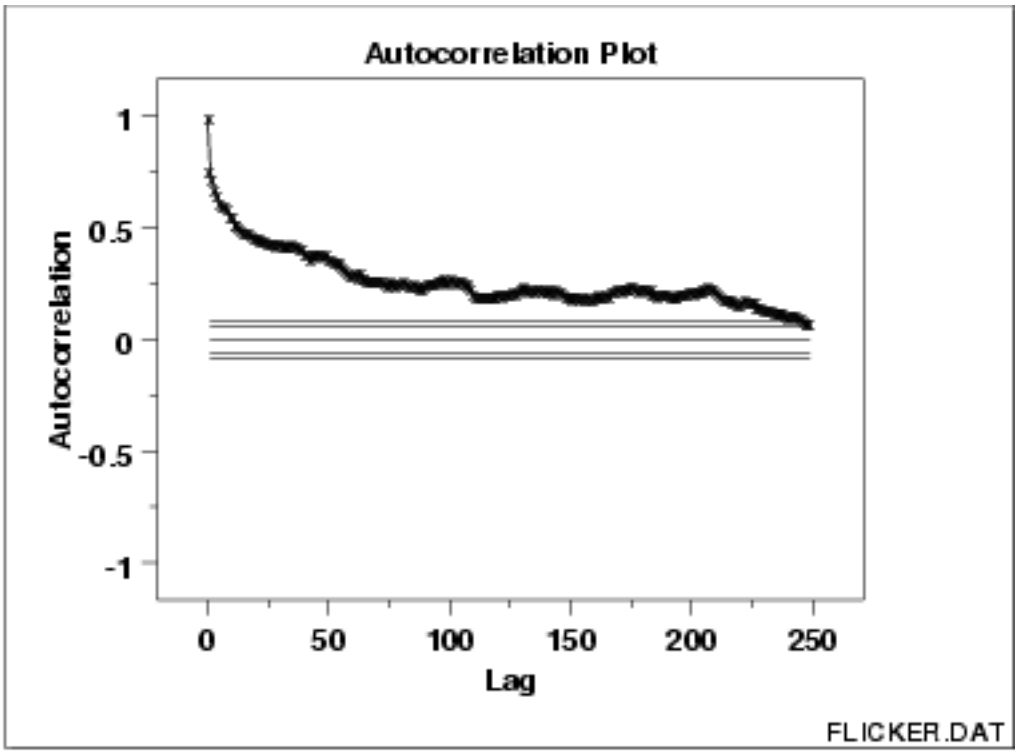
1.3.3.1. Autocorrelation Plot

Purpose: Check Randomness

Autocorrelation plots (Box and Jenkins, pp. 28-32) are a commonly-used tool for checking randomness in a data set. This randomness is ascertained by computing autocorrelations for data values at varying time lags. If random, such autocorrelations should be near zero for any and all time-lag separations. If non-random, then one or more of the autocorrelations will be significantly non-zero.

In addition, autocorrelation plots are used in the model identification stage for Box-Jenkins autoregressive, moving average time series models.

Sample Plot: Autocorrelations should be near-zero for randomness. Such is not the case in this example and thus the randomness assumption fails



This sample autocorrelation plot shows that the time series is not random, but rather has a high degree of autocorrelation between adjacent and near-adjacent observations.

Definition: r(h) versus h

Autocorrelation plots are formed by

- Vertical axis: Autocorrelation coefficient

$$R_h = C_h/C_0$$

where C_h is the autocovariance function

$$C_h = (1/N) \sum_{t=1}^{N-h} (Y_t - \bar{Y})(Y_{t+h} - \bar{Y})$$

and C_0 is the variance function

$$C_0 = \frac{\sum_{t=1}^{N} (Y_t - \bar{Y})^2}{N}$$

Note--R_h is between -1 and +1.

- Horizontal axis: Time lag *h* (*h* = 1, 2, 3, ...)
- The above line also contains several horizontal reference lines. The middle line is at zero. The other four lines are 95% and 99% confidence bands. Note that there are two distinct formulas for generating the confidence bands.
  1. If the autocorrelation plot is being used to test for randomness (i.e., there is no time dependence in the data), the following formula is recommended:

$$\pm \frac{z_{1-\alpha/2}}{\sqrt{N}}$$

     where *N* is the sample size, *z* is the percent point function of the standard normal distribution and α is the. significance level. In this case, the confidence bands have fixed width that depends on the sample size. This is the formula that was used to generate the confidence bands in the above plot.
  2. Autocorrelation plots are also used in the model identification stage for fitting ARIMA models. In this case, a moving average model is assumed for the data and the following confidence bands should be generated:

$$\pm z_{1-\alpha/2} \sqrt{\frac{1}{N}(1 + 2\sum_{i=1}^{k} y_i^2)}$$

     where *k* is the lag, *N* is the sample size, *z* is the percent

point function of the standard normal distribution and α is. the significance level. In this case, the confidence bands increase as the lag increases.

Questions

The autocorrelation plot can provide answers to the following questions:

1. Are the data random?
2. Is an observation related to an adjacent observation?
3. Is an observation related to an observation twice-removed? (etc.)
4. Is the observed time series white noise?
5. Is the observed time series sinusoidal?
6. Is the observed time series autoregressive?
7. What is an appropriate model for the observed time series?
8. Is the model

   Y = constant + error

   valid and sufficient?
9. Is the formula $s_{\bar{Y}} = s/\sqrt{N}$ valid?

Importance: Ensure validity of engineering conclusions

Randomness (along with fixed model, fixed variation, and fixed distribution) is one of the four assumptions that typically underlie all measurement processes. The randomness assumption is critically important for the following three reasons:

1. Most standard statistical tests depend on randomness. The validity of the test conclusions is directly linked to the validity of the randomness assumption.
2. Many commonly-used statistical formulae depend on the randomness assumption, the most common formula being the formula for determining the standard deviation of the sample mean:

   $$s_{\bar{Y}} = s/\sqrt{N}$$

   where s is the standard deviation of the data. Although heavily used, the results from using this formula are of no value unless the randomness assumption holds.
3. For univariate data, the default model is

   Y = constant + error

   If the data are not random, this model is incorrect and invalid, and the estimates for the parameters (such as the constant) become nonsensical and invalid.

In short, if the analyst does not check for randomness, then the validity of many of the statistical conclusions becomes suspect. The autocorrelation plot is an excellent way of checking for such randomness.

Examples

Examples of the autocorrelation plot for several common situations are given in the following pages.

1. Random (= White Noise)
2. Weak autocorrelation
3. Strong autocorrelation and autoregressive model
4. Sinusoidal model

Related Techniques

Partial Autocorrelation Plot
Lag Plot
Spectral Plot
Seasonal Subseries Plot

Case Study

The autocorrelation plot is demonstrated in the beam deflection data case study.

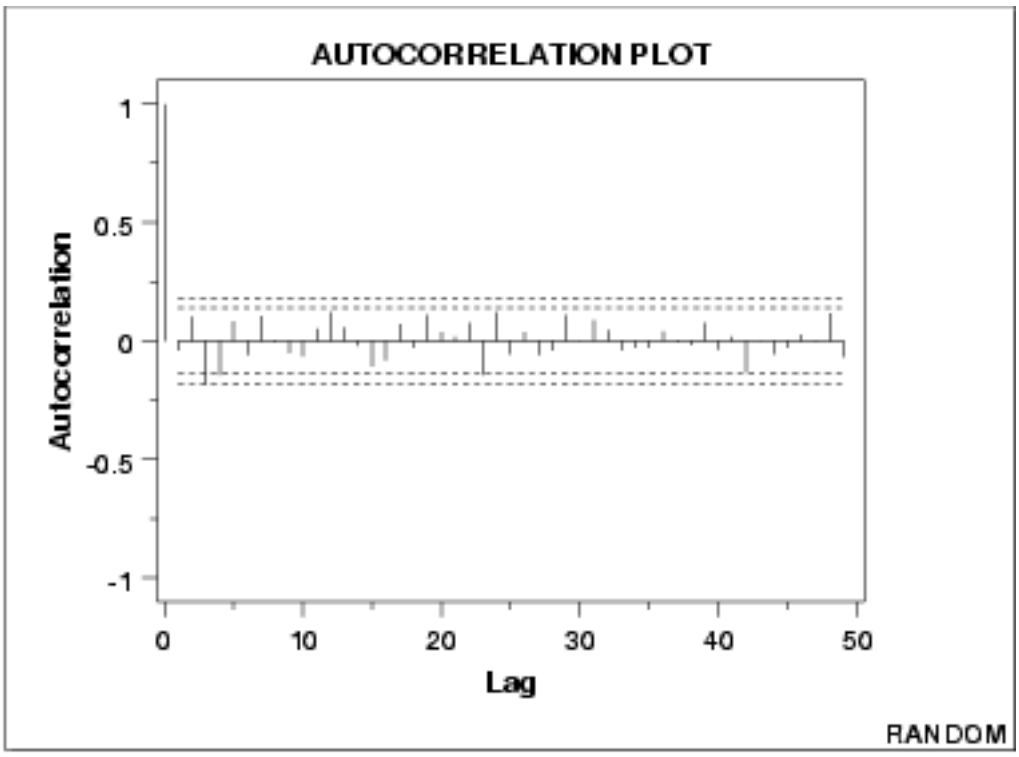
Software

Autocorrelation plots are available in most general purpose statistical software programs including Dataplot.

1.3.3.1.1. Autocorrelation Plot: Random Data

Autocorrelation Plot

The following is a sample autocorrelation plot.



Conclusions

We can make the following conclusions from this plot.

1. There are no significant autocorrelations.
2. The data are random.

Discussion

Note that with the exception of lag 0, which is always 1 by definition, almost all of the autocorrelations fall within the 95% confidence limits. In addition, there is no apparent pattern (such as the first twenty-five being positive and the second twenty-five being negative). This is the abscence of a pattern we expect to see if the data are in fact random.

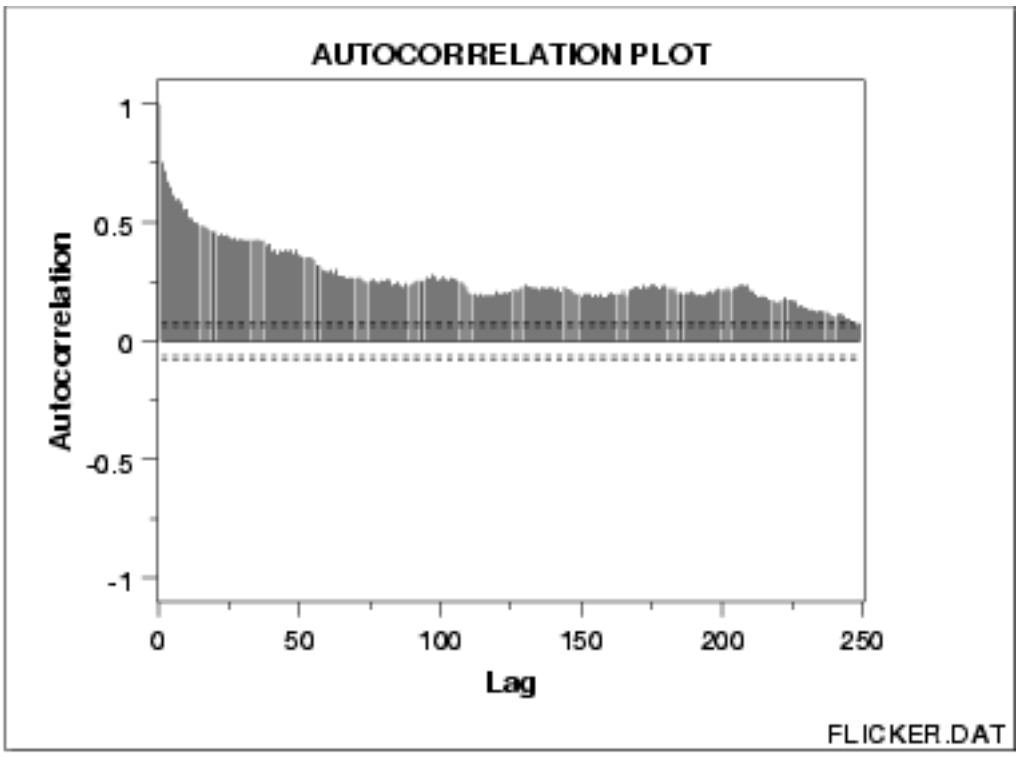
A few lags slightly outside the 95% and 99% confidence limits do not neccessarily indicate non-randomness. For a 95% confidence interval, we might expect about one out of twenty lags to be statistically significant due to random fluctuations.

There is no associative ability to infer from a current value Y_i as to what the next value Y_{i+1} will be. Such non-association is the essense of randomness. In short, adjacent observations do not "co-relate", so we call this the "no autocorrelation" case.

1.3.3.1.2. Autocorrelation Plot: Moderate Autocorrelation

Autocorrelation Plot

The following is a sample autocorrelation plot.



Conclusions

We can make the following conclusions from this plot.

1. The data come from an underlying autoregressive model with moderate positive autocorrelation.

Discussion

The plot starts with a moderately high autocorrelation at lag 1 (approximately 0.75) that gradually decreases. The decreasing autocorrelation is generally linear, but with significant noise. Such a pattern is the autocorrelation plot signature of "moderate autocorrelation", which in turn provides moderate predictability if modeled properly.

Recommended Next Step

The next step would be to estimate the parameters for the autoregressive model:

$$Y_i = A_0 + A_1 * Y_{i-1} + E_i$$

Such estimation can be performed by using least squares linear regression or by fitting a Box-Jenkins autoregressive (AR) model.

The randomness assumption for least squares fitting applies to the residuals of the model. That is, even though the original data exhibit randomness, the residuals after fitting $\mathbf{Y_i}$ against $\mathbf{Y_{i-1}}$ should result in random residuals. Assessing whether or not the proposed model in fact sufficiently removed the randomness is discussed in detail in the Process Modeling chapter.

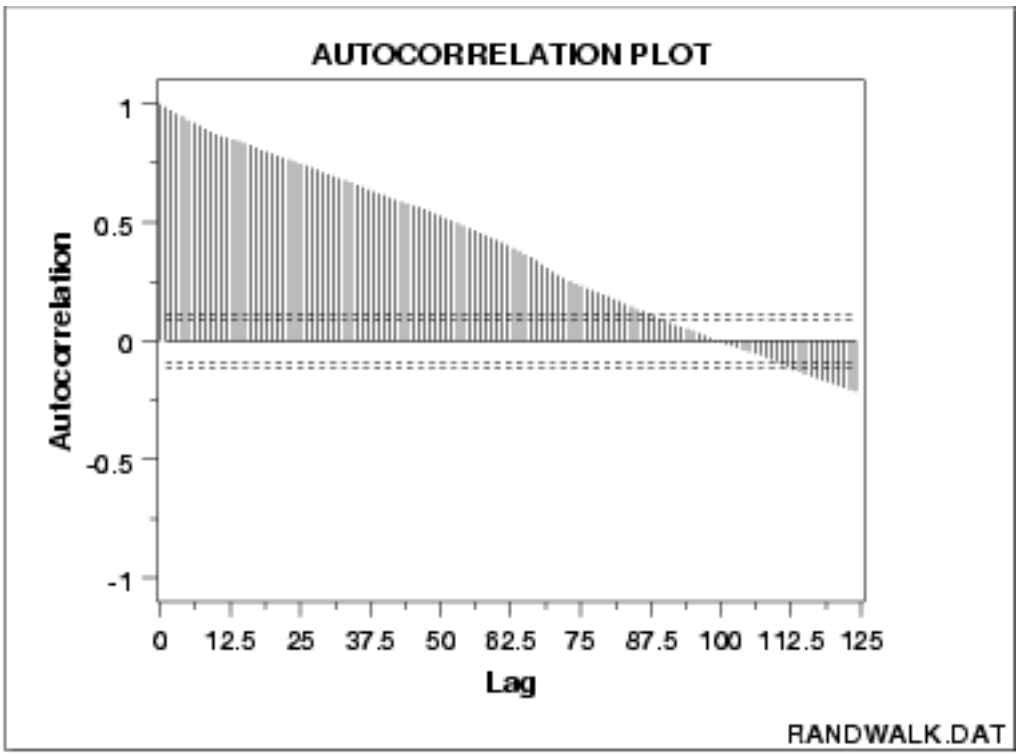
The residual standard deviation for this autoregressive model will be much smaller than the residual standard deviation for the default model

$$Y_i = A_0 + E_i$$

1.3.3.1.3. Autocorrelation Plot: Strong Autocorrelation and Autoregressive Model

Autocorrelation Plot for Strong Autocorrelation

The following is a sample autocorrelation plot.



Conclusions

We can make the following conclusions from the above plot.

1. The data come from an underlying autoregressive model with strong positive autocorrelation.

Discussion

The plot starts with a high autocorrelation at lag 1 (only slightly less than 1) that slowly declines. It continues decreasing until it becomes negative and starts showing an incresing negative autocorrelation. The decreasing autocorrelation is generally linear with little noise. Such a pattern is the autocorrelation plot signature of "strong autocorrelation", which in turn provides high predictability if modeled properly.

Recommended Next Step

The next step would be to estimate the parameters for the autoregressive model:

$$Y_i = A_0 + A_1 * Y_{i-1} + E_i$$

Such estimation can be performed by using least squares linear regression or by fitting a Box-Jenkins autoregressive (AR) model.

The randomness assumption for least squares fitting applies to the residuals of the model. That is, even though the original data exhibit randomness, the residuals after fitting $\mathbf{Y_i}$ against $\mathbf{Y_{i-1}}$ should result in random residuals. Assessing whether or not the proposed model in fact sufficiently removed the randomness is discussed in detail in the Process Modeling chapter.

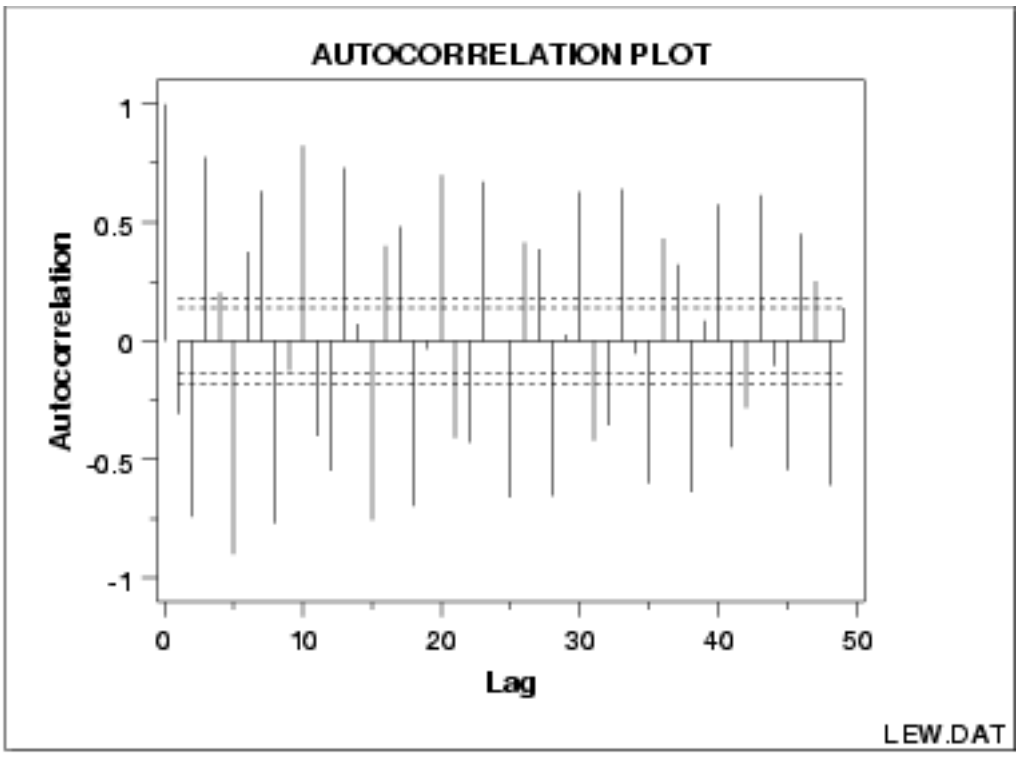
The residual standard deviation for this autoregressive model will be much smaller than the residual standard deviation for the default model

$$Y_i = A_0 + E_i$$

1.3.3.1.4. Autocorrelation Plot: Sinusoidal Model

Autocorrelation Plot for Sinusoidal Model

The following is a sample autocorrelation plot.



Conclusions

We can make the following conclusions from the above plot.

1. The data come from an underlying sinusoidal model.

Discussion

The plot exhibits an alternating sequence of positive and negative spikes. These spikes are not decaying to zero. Such a pattern is the autocorrelation plot signature of a sinusoidal model.

Recommended Next Step

The beam deflection case study gives an example of modeling a sinusoidal model.

1.3.3.2. Bihistogram

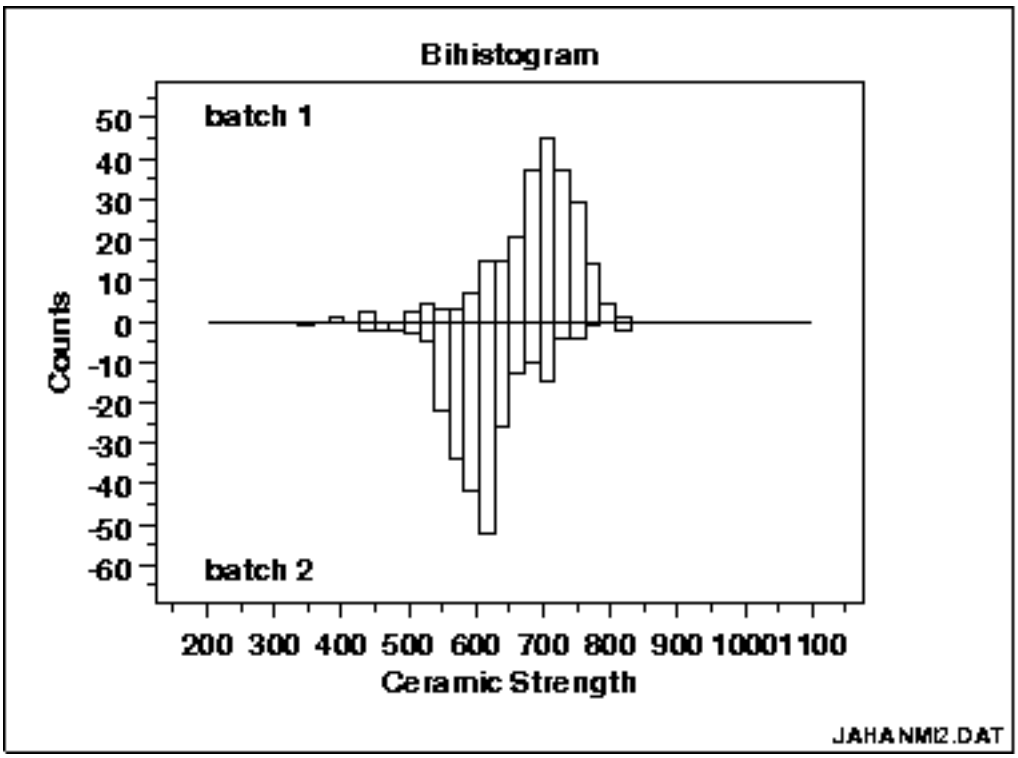
Purpose: Check for a change in location, variation, or distribution

The bihistogram is an EDA tool for assessing whether a before-versus-after engineering modification has caused a change in

- location;
- variation; or
- distribution.

It is a graphical alternative to the two-sample t-test. The bihistogram can be more powerful than the t-test in that all of the distributional features (location, scale, skewness, outliers) are evident on a single plot. It is also based on the common and well-understood histogram.

Sample Plot: This bihistogram reveals that there is a significant difference in ceramic breaking strength between batch 1 (above) and batch 2 (below)



From the above bihistogram, we can see that batch 1 is centered at a ceramic strength value of approximately 725 while batch 2 is centered at a ceramic strength value of approximately 625. That indicates that these batches are displaced by about 100 strength units. Thus the batch

factor has a significant effect on the location (typical value) for strength and hence batch is said to be "significant" or to "have an effect". We thus see graphically and convincingly what a t-test or analysis of variance would indicate quantitatively.

With respect to variation, note that the spread (variation) of the above-axis batch 1 histogram does not appear to be that much different from the below-axis batch 2 histogram. With respect to distributional shape, note that the batch 1 histogram is skewed left while the batch 2 histogram is more symmetric with even a hint of a slight skewness to the right.

Thus the bihistogram reveals that there is a clear difference between the batches with respect to location and distribution, but not in regard to variation. Comparing batch 1 and batch 2, we also note that batch 1 is the "better batch" due to its 100-unit higher average strength (around 725).

Definition: Two adjoined histograms

Bihistograms are formed by vertically juxtaposing two histograms:

- Above the axis: Histogram of the response variable for condition 1
- Below the axis: Histogram of the response variable for condition 2

Questions

The bihistogram can provide answers to the following questions:

1. Is a (2-level) factor significant?
2. Does a (2-level) factor have an effect?
3. Does the location change between the 2 subgroups?
4. Does the variation change between the 2 subgroups?
5. Does the distributional shape change between subgroups?
6. Are there any outliers?

Importance: Checks 3 out of the 4 underlying assumptions of a measurement process

The bihistogram is an important EDA tool for determining if a factor "has an effect". Since the bihistogram provides insight into the validity of three (location, variation, and distribution) out of the four (missing only randomness) underlying assumptions in a measurement process, it is an especially valuable tool. Because of the dual (above/below) nature of the plot, the bihistogram is restricted to assessing factors that have only two levels. However, this is very common in the before-versus-after character of many scientific and engineering experiments.

Related Techniques	t test (for shift in location) F test (for shift in variation) Kolmogorov-Smirnov test (for shift in distribution) Quantile-quantile plot (for shift in location and distribution)
Case Study	The bihistogram is demonstrated in the ceramic strength data case study.
Software	The bihistogram is not widely available in general purpose statistical software programs. Bihistograms can be generated using Dataplot

1. Exploratory Data Analysis
1.3. EDA Techniques
1.3.3. Graphical Techniques: Alphabetic

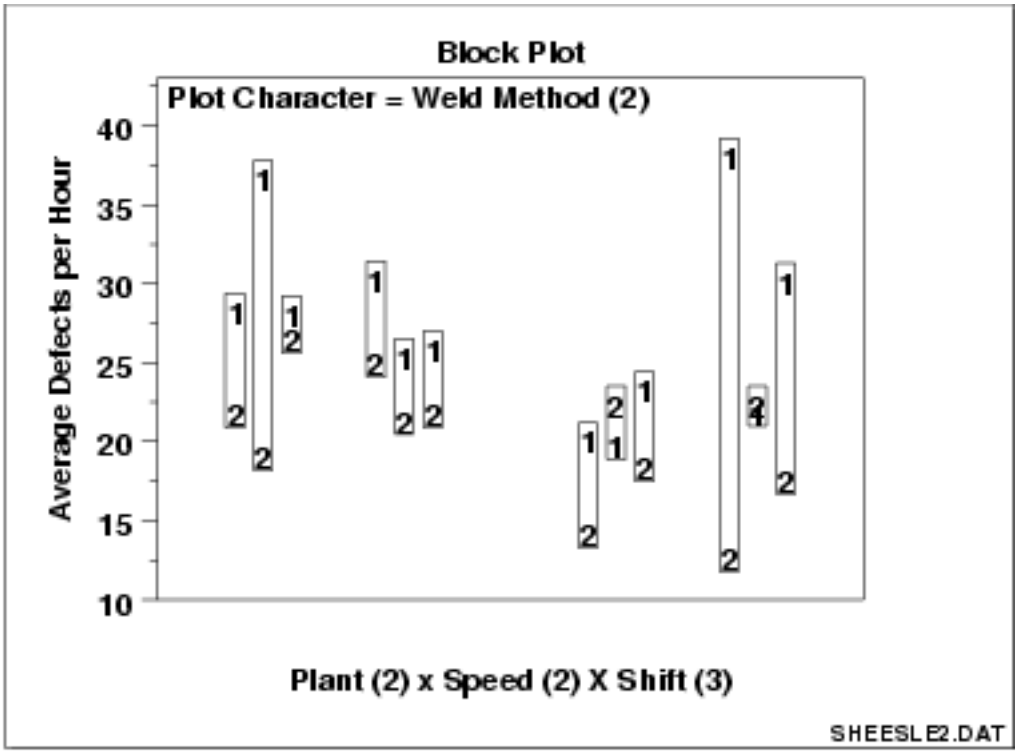
1.3.3.3. Block Plot

Purpose: Check to determine if a factor of interest has an effect robust over all other factors

The block plot (Filliben 1993) is an EDA tool for assessing whether the factor of interest (the primary factor) has a statistically significant effect on the response, and whether that conclusion about the primary factor effect is valid robustly over all other nuisance or secondary factors in the experiment.

It replaces the analysis of variance test with a less assumption-dependent binomial test and should be routinely used whenever we are trying to robustly decide whether a primary factor has an effect.

Sample Plot: Weld method 2 is lower (better) than weld method 1 in 10 of 12 cases



This block plot reveals that in 10 of the 12 cases (bars), weld method 2 is lower (better) than weld method 1. From a binomial point of view, weld method is statistically significant.

Definition

Block Plots are formed as follows:

- Vertical axis: Response variable Y
- Horizontal axis: All combinations of all levels of all nuisance (secondary) factors X1, X2, ...
- Plot Character: Levels of the primary factor XP

Discussion: Primary factor is denoted by plot character: within-bar plot character.

Average number of defective lead wires per hour from a study with four factors,

1. weld strength (2 levels)
2. plant (2 levels)
3. speed (2 levels)
4. shift (3 levels)

are shown in the plot above. Weld strength is the primary factor and the other three factors are nuisance factors. The 12 distinct positions along the horizontal axis correspond to all possible combinations of the three nuisance factors, i.e., 12 = 2 plants x 2 speeds x 3 shifts. These 12 conditions provide the framework for assessing whether any conclusions about the 2 levels of the primary factor (weld method) can truly be called "general conclusions". If we find that one weld method setting does better (smaller average defects per hour) than the other weld method setting for all or most of these 12 nuisance factor combinations, then the conclusion is in fact general and robust.

Ordering along the horizontal axis

In the above chart, the ordering along the horizontal axis is as follows:

- The left 6 bars are from plant 1 and the right 6 bars are from plant 2.
- The first 3 bars are from speed 1, the next 3 bars are from speed 2, the next 3 bars are from speed 1, and the last 3 bars are from speed 2.
- Bars 1, 4, 7, and 10 are from the first shift, bars 2, 5, 8, and 11 are from the second shift, and bars 3, 6, 9, and 12 are from the third shift.

Setting 2 is better than setting 1 in 10 out of 12 cases

In the block plot for the first bar (plant 1, speed 1, shift 1), weld method 1 yields about 28 defects per hour while weld method 2 yields about 22 defects per hour--hence the difference for this combination is about 6 defects per hour and weld method 2 is seen to be better (smaller number of defects per hour).

Is "weld method 2 is better than weld method 1" a general conclusion?

For the second bar (plant 1, speed 1, shift 2), weld method 1 is about 37 while weld method 2 is only about 18. Thus weld method 2 is again seen to be better than weld method 1. Similarly for bar 3 (plant 1, speed 1, shift 3), we see weld method 2 is smaller than weld method 1. Scanning over all of the 12 bars, we see that weld method 2 is smaller than weld method 1 in 10 of the 12 cases, which is highly suggestive of a robust weld method effect.

An event with chance probability of only 2%

What is the chance of 10 out of 12 happening by chance? This is probabilistically equivalent to testing whether a coin is fair by flipping it and getting 10 heads in 12 tosses. The chance (from the binomial distribution) of getting 10 (or more extreme: 11, 12) heads in 12 flips of a fair coin is about 2%. Such low-probability events are usually rejected as untenable and in practice we would conclude that there is a difference in weld methods.

Advantage: Graphical and binomial

The advantages of the block plot are as follows:

- A quantitative procedure (analysis of variance) is replaced by a graphical procedure.
- An F-test (analysis of variance) is replaced with a binomial test, which requires fewer assumptions.

Questions

The block plot can provide answers to the following questions:

1. Is the factor of interest significant?
2. Does the factor of interest have an effect?
3. Does the location change between levels of the primary factor?
4. Has the process improved?
5. What is the best setting (= level) of the primary factor?
6. How much of an average improvement can we expect with this best setting of the primary factor?
7. Is there an interaction between the primary factor and one or more nuisance factors?
8. Does the effect of the primary factor change depending on the setting of some nuisance factor?

9. Are there any outliers?

Importance: Robustly checks the significance of the factor of interest	The block plot is a graphical technique that pointedly focuses on whether or not the primary factor conclusions are in fact robustly general. This question is fundamentally different from the generic multi-factor experiment question where the analyst asks, "What factors are important and what factors are not" (a screening problem)? Global data analysis techniques, such as analysis of variance, can potentially be improved by local, focused data analysis techniques that take advantage of this difference.
Related Techniques	t test (for shift in location for exactly 2 levels) ANOVA (for shift in location for 2 or more levels) Bihistogram (for shift in location, variation, and distribution for exactly 2 levels).
Case Study	The block plot is demonstrated in the ceramic strength data case study.
Software	Block plots can be generated with the Dataplot software program. They are not currently available in other statistical software programs.

1.3.3.4. Bootstrap Plot

Purpose: Estimate uncertainty

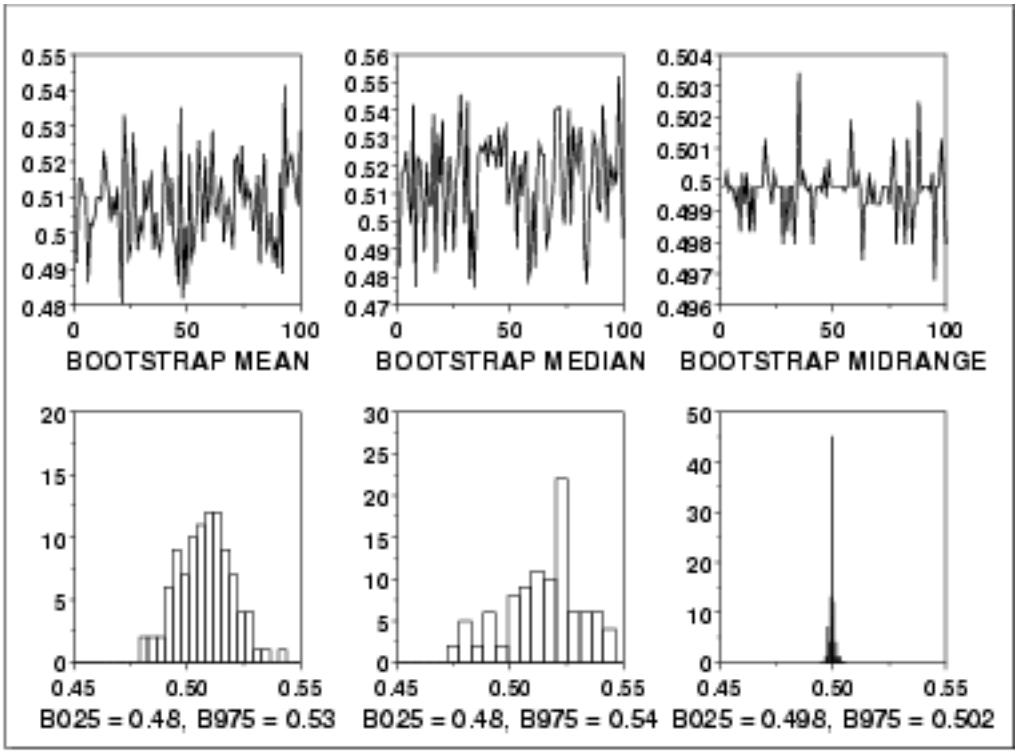
The bootstrap (Efron and Gong) plot is used to estimate the uncertainty of a statistic.

Generate subsamples with replacement

To generate a bootstrap uncertainty estimate for a given statistic from a set of data, a subsample of a size less than or equal to the size of the data set is generated from the data, and the statistic is calculated. This subsample is generated *with replacement* so that any data point can be sampled multiple times or not sampled at all. This process is repeated for many subsamples, typically between 500 and 1000. The computed values for the statistic form an estimate of the sampling distribution of the statistic.

For example, to estimate the uncertainty of the median from a dataset with 50 elements, we generate a subsample of 50 elements and calculate the median. This is repeated at least 500 times so that we have at least 500 values for the median. Although the number of bootstrap samples to use is somewhat arbitrary, 500 subsamples is usually sufficient. To calculate a 90% confidence interval for the median, the sample medians are sorted into ascending order and the value of the 25th median (assuming exactly 500 subsamples were taken) is the lower confidence limit while the value of the 475th median (assuming exactly 500 subsamples were taken) is the upper confidence limit.

Sample Plot:



This bootstrap plot was generated from 500 uniform random numbers. Bootstrap plots and corresponding histograms were generated for the mean, median, and mid-range. The histograms for the corresponding statistics clearly show that for uniform random numbers the mid-range has the smallest variance and is, therefore, a superior location estimator to the mean or the median.

Definition

The bootstrap plot is formed by:

- Vertical axis: Computed value of the desired statistic for a given subsample.
- Horizontal axis: Subsample number.

The bootstrap plot is simply the computed value of the statistic versus the subsample number. That is, the bootstrap plot generates the values for the desired statistic. This is usually immediately followed by a histogram or some other distributional plot to show the location and variation of the sampling distribution of the statistic.

Questions

The bootstrap plot is used to answer the following questions:

- What does the sampling distribution for the statistic look like?
- What is a 95% confidence interval for the statistic?
- Which statistic has a sampling distribution with the smallest variance? That is, which statistic generates the narrowest confidence interval?

Importance

The most common uncertainty calculation is generating a confidence interval for the mean. In this case, the uncertainty formula can be derived mathematically. However, there are many situations in which the uncertainty formulas are mathematically intractable. The bootstrap provides a method for calculating the uncertainty in these cases.

Cautuion on use of the bootstrap

The bootstrap is not appropriate for all distributions and statistics (Efron and Tibrashani). For example, because of the shape of the uniform distribution, the bootstrap is not appropriate for estimating the distribution of statistics that are heavily dependent on the tails, such as the range.

Related Techniques

Histogram
Jackknife

The jacknife is a technique that is closely related to the bootstrap. The jackknife is beyond the scope of this handbook. See the Efron and Gong article for a discussion of the jackknife.

Case Study

The bootstrap plot is demonstrated in the uniform random numbers case study.

Software

The bootstrap is becoming more common in general purpose statistical software programs. However, it is still not supported in many of these programs. Dataplot supports a bootstrap capability.

1.3.3.5. Box-Cox Linearity Plot

Purpose: Find the transformation of the X variable that maximizes the correlation between a Y and an X variable

When performing a linear fit of Y against X, an appropriate transformation of X can often significantly improve the fit. The Box-Cox transformation (Box and Cox, 1964) is a particularly useful family of transformations. It is defined as:

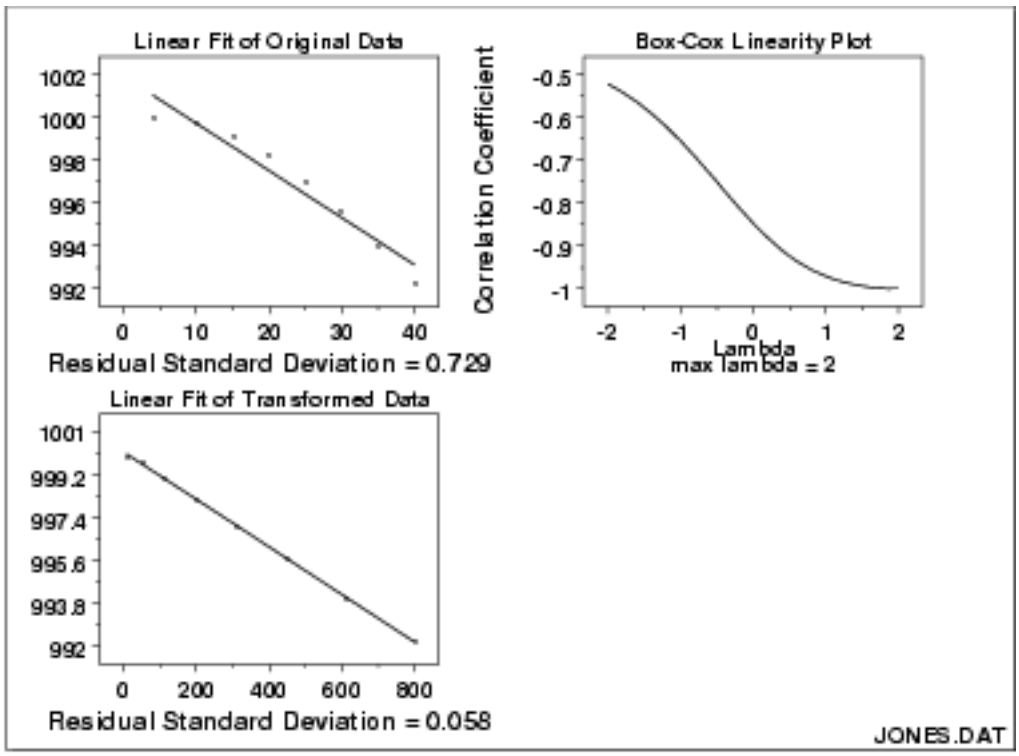
$$T(X) = (X^{\lambda} - 1)/\lambda$$

where X is the variable being transformed and λ is the transformation parameter. For $\lambda = 0$, the natural log of the data is taken instead of using the above formula.

The Box-Cox linearity plot is a plot of the correlation between Y and the transformed X for given values of λ. That is, λ is the coordinate for the horizontal axis variable and the value of the correlation between Y and the transformed X is the coordinate for the vertical axis of the plot. The value of λ corresponding to the maximum correlation (or minimum for negative correlation) on the plot is then the optimal choice for λ.

Transforming X is used to improve the fit. The Box-Cox transformation applied to Y can be used as the basis for meeting the error assumptions. That case is not covered here. See page 225 of (Draper and Smith, 1981) or page 77 of (Ryan, 1997) for a discussion of this case.

Sample Plot



The plot of the original data with the predicted values from a linear fit indicate that a quadratic fit might be preferable. The Box-Cox linearity plot shows a value of $\lambda = 2.0$. The plot of the transformed data with the predicted values from a linear fit with the transformed data shows a better fit (verified by the significant reduction in the residual standard deviation).

Definition

Box-Cox linearity plots are formed by

- Vertical axis: Correlation coefficient from the transformed X and Y
- Horizontal axis: Value for λ

Questions

The Box-Cox linearity plot can provide answers to the following questions:

1. Would a suitable transformation improve my fit?
2. What is the optimal value of the transformation parameter?

Importance: Find a suitable transformation

Transformations can often significantly improve a fit. The Box-Cox linearity plot provides a convenient way to find a suitable transformation without engaging in a lot of trial and error fitting.

Related Techniques

Linear Regression
Box-Cox Normality Plot

Case Study The Box-Cox linearity plot is demonstrated in the Alaska pipeline data case study.

Software Box-Cox linearity plots are not a standard part of most general purpose statistical software programs. However, the underlying technique is based on a transformation and computing a correlation coefficient. So if a statistical program supports these capabilities, writing a macro for a Box-Cox linearity plot should be feasible. Dataplot supports a Box-Cox linearity plot directly.

1.3.3.6. Box-Cox Normality Plot

Purpose: Find transformation to normalize data

Many statistical tests and intervals are based on the assumption of normality. The assumption of normality often leads to tests that are simple, mathematically tractable, and powerful compared to tests that do not make the normality assumption. Unfortunately, many real data sets are in fact not approximately normal. However, an appropriate transformation of a data set can often yield a data set that does follow approximately a normal distribution. This increases the applicability and usefulness of statistical techniques based on the normality assumption.

The Box-Cox transformation is a particulary useful family of transformations. It is defined as:

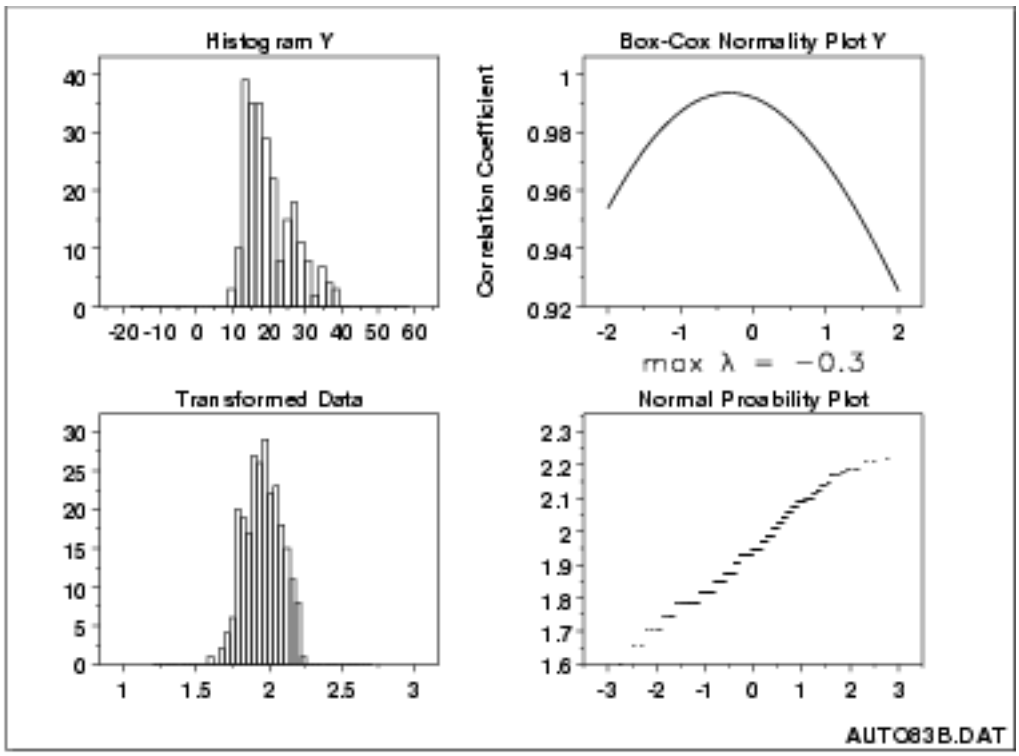
$$T(Y) = (Y^{\lambda} - 1)/\lambda$$

where Y is the response variable and λ is the transformation parameter. For $\lambda = 0$, the natural log of the data is taken instead of using the above formula.

Given a particular transformation such as the Box-Cox transformation defined above, it is helpful to define a measure of the normality of the resulting transformation. One measure is to compute the correlation coefficient of a normal probability plot. The correlation is computed between the vertical and horizontal axis variables of the probability plot and is a convenient measure of the linearity of the probability plot (the more linear the probability plot, the better a normal distribution fits the data).

The Box-Cox normality plot is a plot of these correlation coefficients for various values of the λ parameter. The value of λ corresponding to the maximum correlation on the plot is then the optimal choice for λ.

Sample Plot



The histogram in the upper left-hand corner shows a data set that has significant right skewness (and so does not follow a normal distribution). The Box-Cox normality plot shows that the maximum value of the correlation coefficient is at λ = -0.3. The histogram of the data after applying the Box-Cox transformation with λ = -0.3 shows a data set for which the normality assumption is reasonable. This is verified with a normal probability plot of the transformed data.

Definition

Box-Cox normality plots are formed by:

- Vertical axis: Correlation coefficient from the normal probability plot after applying Box-Cox transformation
- Horizontal axis: Value for λ

Questions

The Box-Cox normality plot can provide answers to the following questions:

1. Is there a transformation that will normalize my data?
2. What is the optimal value of the transformation parameter?

Importance: Normalization Improves Validity of Tests

Normality assumptions are critical for many univariate intervals and hypothesis tests. It is important to test the normality assumption. If the data are in fact clearly not normal, the Box-Cox normality plot can often be used to find a transformation that will approximately normalize the data.

Related Techniques	Normal Probability Plot Box-Cox Linearity Plot
Software	Box-Cox normality plots are not a standard part of most general purpose statistical software programs. However, the underlying technique is based on a normal probability plot and computing a correlation coefficient. So if a statistical program supports these capabilities, writing a macro for a Box-Cox normality plot should be feasible. Dataplot supports a Box-Cox normality plot directly.

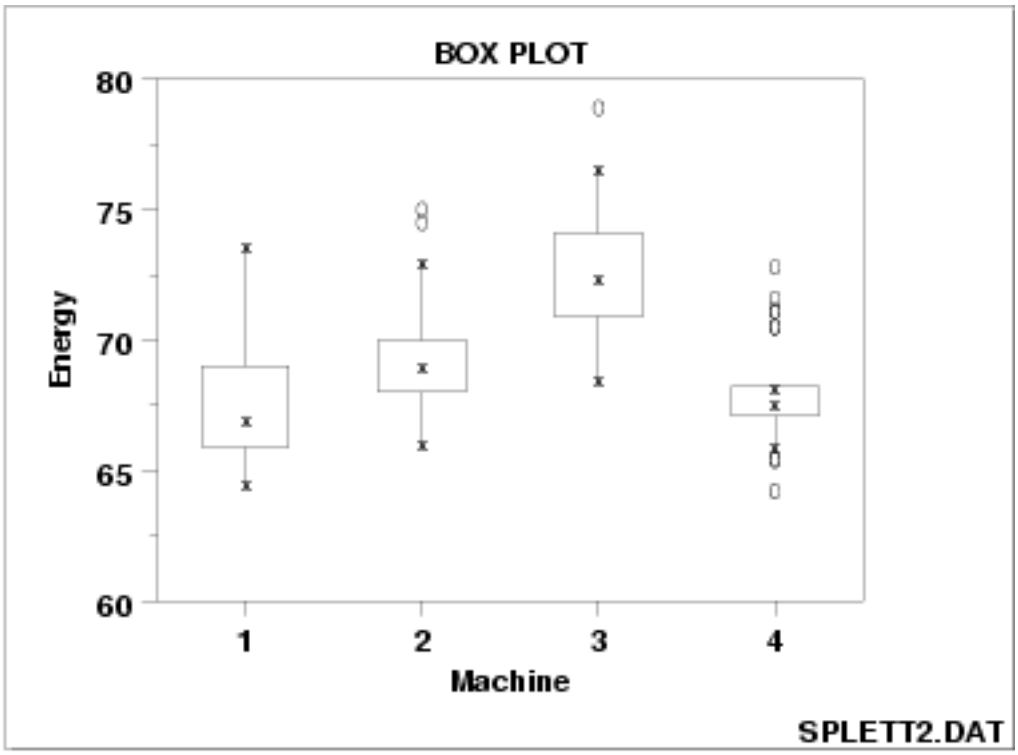
1. Exploratory Data Analysis
1.3. EDA Techniques
1.3.3. Graphical Techniques: Alphabetic

1.3.3.7. Box Plot

Purpose: Check location and variation shifts

Box plots (Chambers 1983) are an excellent tool for conveying location and variation information in data sets, particularly for detecting and illustrating location and variation changes between different groups of data.

Sample Plot: This box plot reveals that machine has a significant effect on energy with respect to location and possibly variation



This box plot, comparing four machines for energy output, shows that machine has a significant effect on energy with respect to both location and variation. Machine 3 has the highest energy response (about 72.5); machine 4 has the least variable energy response with about 50% of its readings being within 1 energy unit.

Definition

Box plots are formed by

Vertical axis: Response variable
Horizontal axis: The factor of interest

More specifically, we

1. Calculate the median and the quartiles (the lower quartile is the 25th percentile and the upper quartile is the 75th percentile).
2. Plot a symbol at the median (or draw a line) and draw a box (hence the name--box plot) between the lower and upper quartiles; this box represents the middle 50% of the data--the "body" of the data.
3. Draw a line from the lower quartile to the minimum point and another line from the upper quartile to the maximum point. Typically a symbol is drawn at these minimum and maximum points, although this is optional.

Thus the box plot identifies the middle 50% of the data, the median, and the extreme points.

Single or multiple box plots can be drawn

A single box plot can be drawn for one batch of data with no distinct groups. Alternatively, multiple box plots can be drawn together to compare multiple data sets or to compare groups in a single data set. For a single box plot, the width of the box is arbitrary. For multiple box plots, the width of the box plot can be set proportional to the number of points in the given group or sample (some software implementations of the box plot simply set all the boxes to the same width).

Box plots with fences

There is a useful variation of the box plot that more specifically identifies outliers. To create this variation:

1. Calculate the median and the lower and upper quartiles.
2. Plot a symbol at the median and draw a box between the lower and upper quartiles.
3. Calculate the interquartile range (the difference between the upper and lower quartile) and call it IQ.
4. Calculate the following points:

   L1 = lower quartile - 1.5*IQ
   L2 = lower quartile - 3.0*IQ
   U1 = upper quartile + 1.5*IQ
   U2 = upper quartile + 3.0*IQ

5. The line from the lower quartile to the minimum is now drawn from the lower quartile to the smallest point that is greater than L1. Likewise, the line from the upper quartile to the maximum is now drawn to the largest point smaller than U1.

6. Points between L1 and L2 or between U1 and U2 are drawn as small circles. Points less than L2 or greater than U2 are drawn as large circles.

Questions

The box plot can provide answers to the following questions:

1. Is a factor significant?
2. Does the location differ between subgroups?
3. Does the variation differ between subgroups?
4. Are there any outliers?

Importance: Check the significance of a factor

The box plot is an important EDA tool for determining if a factor has a significant effect on the response with respect to either location or variation.

The box plot is also an effective tool for summarizing large quantities of information.

Related Techniques

Mean Plot
Analysis of Variance

Case Study

The box plot is demonstrated in the ceramic strength data case study.

Software

Box plots are available in most general purpose statistical software programs, including Dataplot.

1.3.3.8. Complex Demodulation Amplitude Plot

Purpose: Detect Changing Amplitude in Sinusoidal Models

In the frequency analysis of time series models, a common model is the sinusoidal model:

$$Y_i = C + \alpha \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

In this equation, α is the amplitude, ϕ is the phase shift, and ω is the dominant frequency. In the above model, α and ϕ are constant, that is they do not vary with time, t_i.

The complex demodulation amplitude plot (Granger, 1964) is used to determine if the assumption of constant amplitude is justifiable. If the slope of the complex demodulation amplitude plot is zero, then the above model is typically replaced with the model:

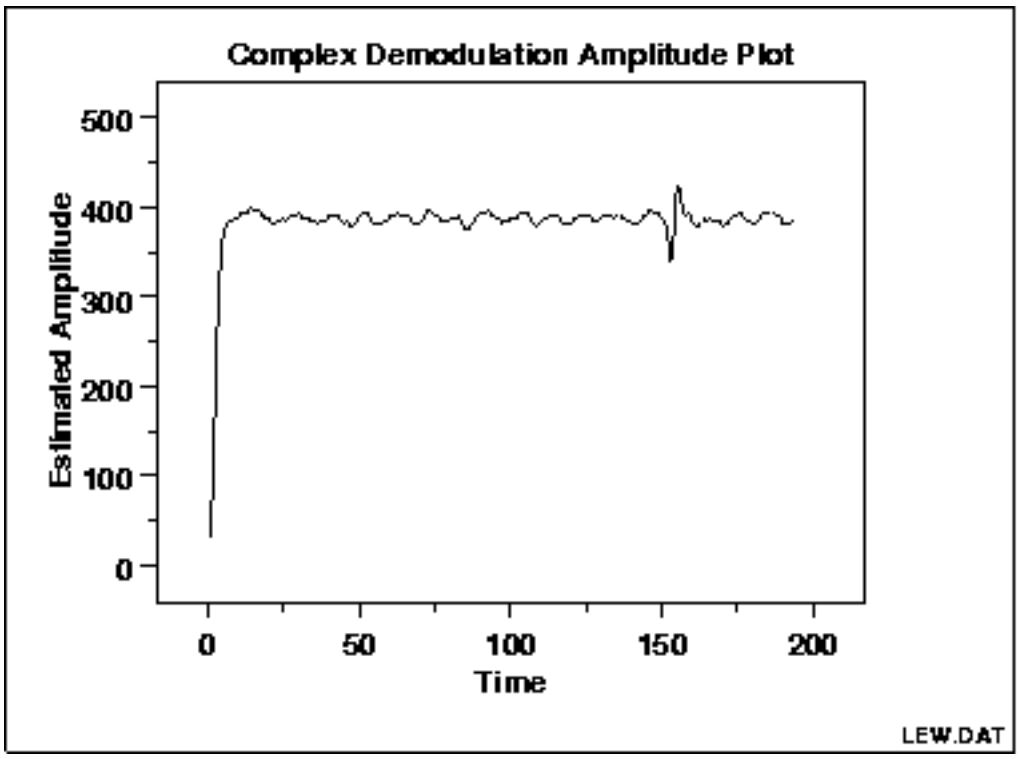
$$Y_i = C + \alpha_i \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

where $\hat{\alpha}_i$ is some type of linear model fit with standard least squares. The most common case is a linear fit, that is the model becomes

$$Y_i = C + \left(B_0 + B_1 * t_i\right) \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

Quadratic models are sometimes used. Higher order models are relatively rare.

Sample Plot:



This complex demodulation amplitude plot shows that:

- the amplitude is fixed at approximately 390;
- there is a start-up effect; and
- there is a change in amplitude at around $x = 160$ that should be investigated for an outlier.

Definition: The complex demodulation amplitude plot is formed by:

- Vertical axis: Amplitude
- Horizontal axis: Time

The mathematical computations for determining the amplitude are beyond the scope of the Handbook. Consult Granger (Granger, 1964) for details.

Questions The complex demodulation amplitude plot answers the following questions:

1. Does the amplitude change over time?
2. Are there any outliers that need to be investigated?
3. Is the amplitude different at the beginning of the series (i.e., is there a start-up effect)?

Importance: Assumption Checking

As stated previously, in the frequency analysis of time series models, a common model is the sinusoidal model:

$$Y_i = C + \alpha \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

In this equation, α is assumed to be constant, that is it does not vary with time. It is important to check whether or not this assumption is reasonable.

The complex demodulation amplitude plot can be used to verify this assumption. If the slope of this plot is essentially zero, then the assumption of constant amplitude is justified. If it is not, α should be replaced with some type of time-varying model. The most common cases are linear ($B_0 + B_1*t$) and quadratic ($B_0 + B_1*t + B_2*t^2$).

Related Techniques

Spectral Plot
Complex Demodulation Phase Plot
Non-Linear Fitting

Case Study

The complex demodulation amplitude plot is demonstrated in the beam deflection data case study.

Software

Complex demodulation amplitude plots are available in some, but not most, general purpose statistical software programs. Dataplot supports complex demodulation amplitude plots.

1.3.3.9. Complex Demodulation Phase Plot

Purpose: Improve the estimate of frequency in sinusoidal time series models

As stated previously, in the frequency analysis of time series models, a common model is the sinusoidal model:

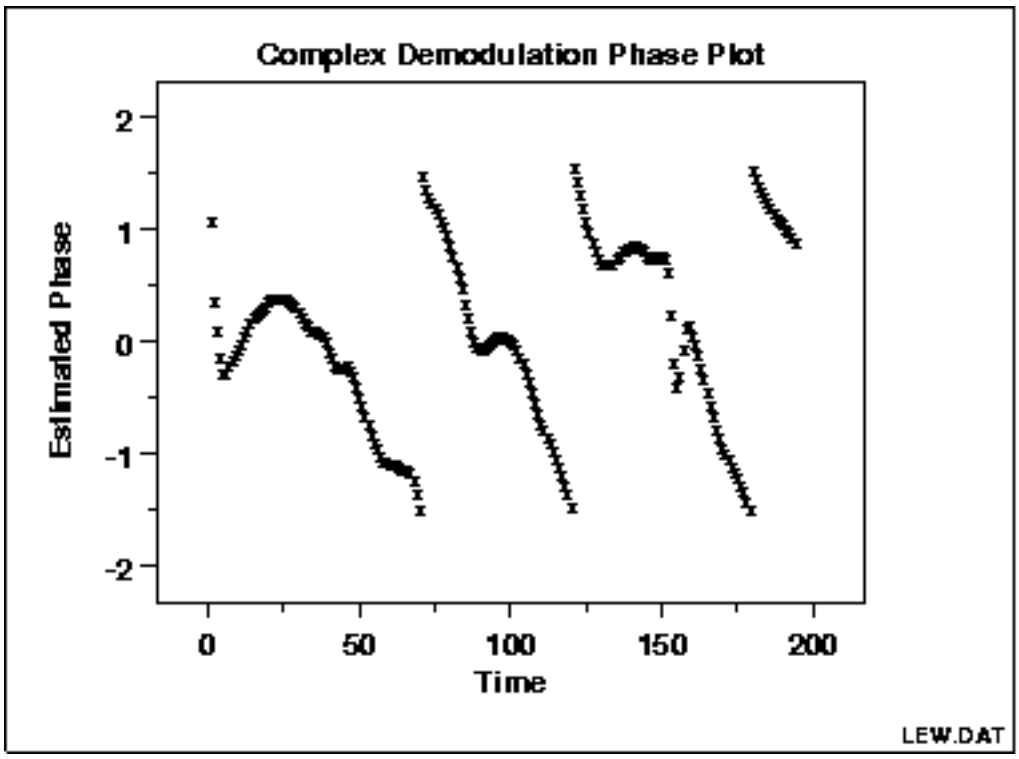
$$Y_i = C + \alpha \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

In this equation, α is the amplitude, ϕ is the phase shift, and ω is the dominant frequency. In the above model, α and ϕ are constant, that is they do not vary with time t_i.

The complex demodulation phase plot (Granger, 1964) is used to improve the estimate of the frequency (i.e., ω) in this model.

If the complex demodulation phase plot shows lines sloping from left to right, then the estimate of the frequency should be increased. If it shows lines sloping right to left, then the frequency should be decreased. If there is essentially zero slope, then the frequency estimate does not need to be modified.

Sample Plot:



This complex demodulation phase plot shows that:

- the specified demodulation frequency is incorrect;
- the demodulation frequency should be increased.

Definition

The complex demodulation phase plot is formed by:

- Vertical axis: Phase
- Horizontal axis: Time

The mathematical computations for the phase plot are beyond the scope of the Handbook. Consult Granger (Granger, 1964) for details.

Questions

The complex demodulation phase plot answers the following question:

Is the specified demodulation frequency correct?

Importance of a Good Initial Estimate for the Frequency

The non-linear fitting for the sinusoidal model:

$$Y_i = C + \alpha \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

is usually quite sensitive to the choice of good starting values. The initial estimate of the frequency, ω, is obtained from a spectral plot. The complex demodulation phase plot is used to assess whether this estimate is adequate, and if it is not, whether it should be increased or decreased. Using the complex demodulation phase plot with the spectral plot can significantly improve the quality of the non-linear fits obtained.

Related Techniques	Spectral Plot Complex Demodulation Phase Plot Non-Linear Fitting
Case Study	The complex demodulation amplitude plot is demonstrated in the beam deflection data case study.
Software	Complex demodulation phase plots are available in some, but not most, general purpose statistical software programs. Dataplot supports complex demodulation phase plots.

1. Exploratory Data Analysis
1.3. EDA Techniques
1.3.3. Graphical Techniques: Alphabetic

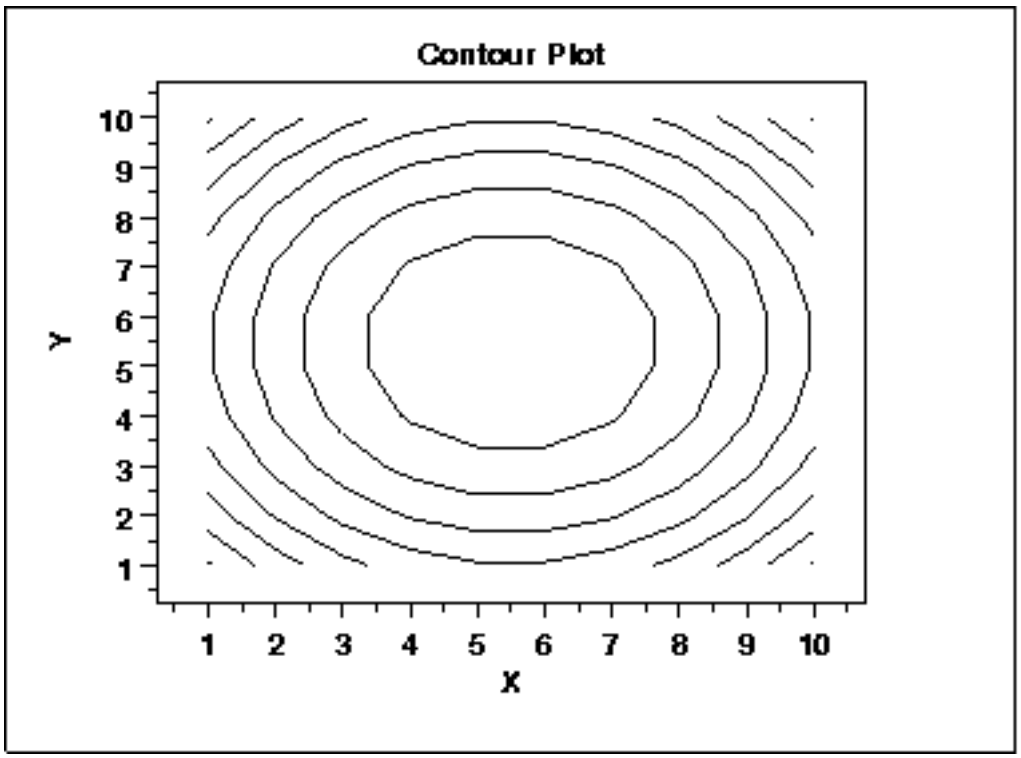
1.3.3.10. Contour Plot

Purpose: Display 3-d surface on 2-d plot

A contour plot is a graphical technique for representing a 3-dimensional surface by plotting constant *z* slices, called contours, on a 2-dimensional format. That is, given a value for *z*, lines are drawn for connecting the (x,y) coordinates where that *z* value occurs.

The contour plot is an alternative to a 3-D surface plot.

Sample Plot:



This contour plot shows that the surface is symmetric and peaks in the center.

Definition

The contour plot is formed by:

- Vertical axis: Independent variable 2
- Horizontal axis: Independent variable 1
- Lines: iso-response values

The independent variables are usually restricted to a regular grid. The actual techniques for determining the correct iso-response values are rather complex and are almost always computer generated.

An additional variable may be required to specify the Z values for drawing the iso-lines. Some software packages require explicit values. Other software packages will determine them automatically.

If the data (or function) do not form a regular grid, you typically need to perform a 2-D interpolation to form a regular grid.

Questions

The contour plot is used to answer the question

How does Z change as a function of X and Y?

Importance: Visualizing 3-dimensional data

For univariate data, a run sequence plot and a histogram are considered necessary first steps in understanding the data. For 2-dimensional data, a scatter plot is a necessary first step in understanding the data.

In a similar manner, 3-dimensional data should be plotted. Small data sets, such as result from designed experiments, can typically be represented by block plots, dex mean plots, and the like (here, "DEX" stands for "Design of Experiments"). For large data sets, a contour plot or a 3-D surface plot should be considered a necessary first step in understanding the data.

DEX Contour Plot

The dex contour plot is a specialized contour plot used in the design of experiments. In particular, it is useful for full and fractional designs.

Related Techniques

3-D Plot

Software

Contour plots are available in most general purpose statistical software programs. They are also available in many general purpose graphics and mathematics programs. These programs vary widely in the capabilities for the contour plots they generate. Many provide just a basic contour plot over a rectangular grid while others permit color filled or shaded contours. Dataplot supports a fairly basic contour plot.

Most statistical software programs that support design of experiments will provide a dex contour plot capability.

1.3.3.10.1. DEX Contour Plot

DEX Contour Plot: Introduction

The dex contour plot is a specialized contour plot used in the analysis of full and fractional experimental designs. These designs often have a low level, coded as "-1" or "-", and a high level, coded as "+1" or "+" for each factor. In addition, there can optionally be one or more center points. Center points are at the mid-point between the low and high level for each factor and are coded as "0".

The dex contour plot is generated for two factors. Typically, this would be the two most important factors as determined by previous analyses (e.g., through the use of the dex mean plots and a Yates analysis). If more than two factors are important, you may want to generate a series of dex contour plots, each of which is drawn for two of these factors. You can also generate a matrix of all pairwise dex contour plots for a number of important factors (similar to the scatter plot matrix for scatter plots).

The typical application of the dex contour plot is in determining settings that will maximize (or minimize) the response variable. It can also be helpful in determining settings that result in the response variable hitting a pre-determined target value. The dex contour plot plays a useful role in determining the settings for the next iteration of the experiment. That is, the initial experiment is typically a fractional factorial design with a fairly large number of factors. After the most important factors are determined, the dex contour plot can be used to help define settings for a full factorial or response surface design based on a smaller number of factors.

Construction of DEX Contour Plot

The following are the primary steps in the construction of the dex contour plot.

1. The *x* and *y* axes of the plot represent the values of the first and second factor (independent) variables.
2. The four vertex points are drawn. The vertex points are (-1,-1), (-1,1), (1,1), (1,-1). At each vertex point, the average of all the response values at that vertex point is printed.
3. Similarly, if there are center points, a point is drawn at (0,0) and the average of the response values at the center points is printed.
4. The **linear** dex contour plot assumes the model:

$$Y = \mu + 0.5(\beta_1 * U_1 + \beta_2 * U_2 + \beta_{12} * U_1 * U_2)$$

where μ is the overall mean of the response variable. The values of β_1, β_2, β_{12}, and μ are estimated from the vertex points using a Yates analysis (the Yates analysis utilizes the special structure of the 2-level full and fractional factorial designs to simplify the computation of these parameter estimates). Note that for the dex contour plot, a full Yates analysis does not need to performed, simply the calculations for generating the parameter estimates.

In order to generate a single contour line, we need a value for *Y*, say Y_0. Next, we solve for U_2 in terms of U_1 and, after doing the algebra, we have the equation:

$$U_2 = \frac{2(Y_0 - \mu) - \beta_1 * U_1}{\beta_2 + \beta_{12} * U_1}$$

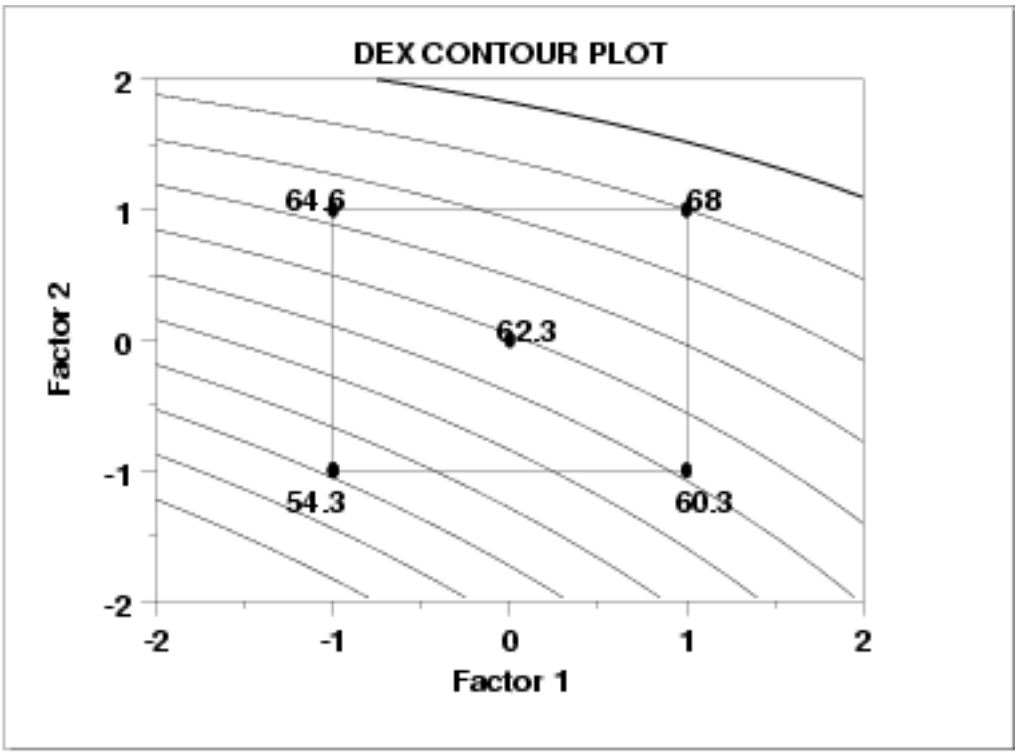
We generate a sequence of points for U_1 in the range -2 to 2 and compute the corresponding values of U_2. These points constitute a single contour line corresponding to $Y = Y_0$.

The user specifies the target values for which contour lines will be generated.

The above algorithm assumes a linear model for the design. Dex contour plots can also be generated for the case in which we assume a quadratic model for the design. The algebra for solving for U_2 in terms of U_1 becomes more complicated, but the fundamental idea is the same. Quadratic models are needed for the case when the average for the center points does not fall in the range defined by the vertex point (i.e., there is curvature).

Sample DEX Contour Plot

The following is a dex contour plot for the data used in the Eddy current case study. The analysis in that case study demonstrated that X1 and X2 were the most important factors.



Interpretation of the Sample DEX Contour Plot

From the above dex contour plot we can derive the following information.

1. Interaction significance;
2. Best (data) setting for these 2 dominant factors;

Interaction Significance

Note the appearance of the contour plot. If the contour curves are linear, then that implies that the interaction term is not significant; if the contour curves have considerable curvature, then that implies that the interaction term is large and important. In our case, the contour curves do not have considerable curvature, and so we conclude that the X1*X2 term is not significant.

Best Settings

To determine the best factor settings for the already-run experiment, we first must define what "best" means. For the Eddy current data set used to generate this dex contour plot, "best" means to **maximize** (rather than minimize or hit a target) the response. Hence from the contour plot we determine the best settings for the two dominant factors by simply scanning the four vertices and choosing the vertex with the **largest** value (= average response). In this case, it is (X1 = +1, X2 = +1).

As for factor X3, the contour plot provides no best setting information, and so we would resort to other tools: the main effects plot, the interaction effects matrix, or the ordered data to determine optimal X3 settings.

Case Study

The Eddy current case study demonstrates the use of the dex contour plot in the context of the analysis of a full factorial design.

Software

DEX contour plots are available in many statistical software programs that analyze data from designed experiments. Dataplot supports a linear dex contour plot and it provides a macro for generating a quadratic dex contour plot.

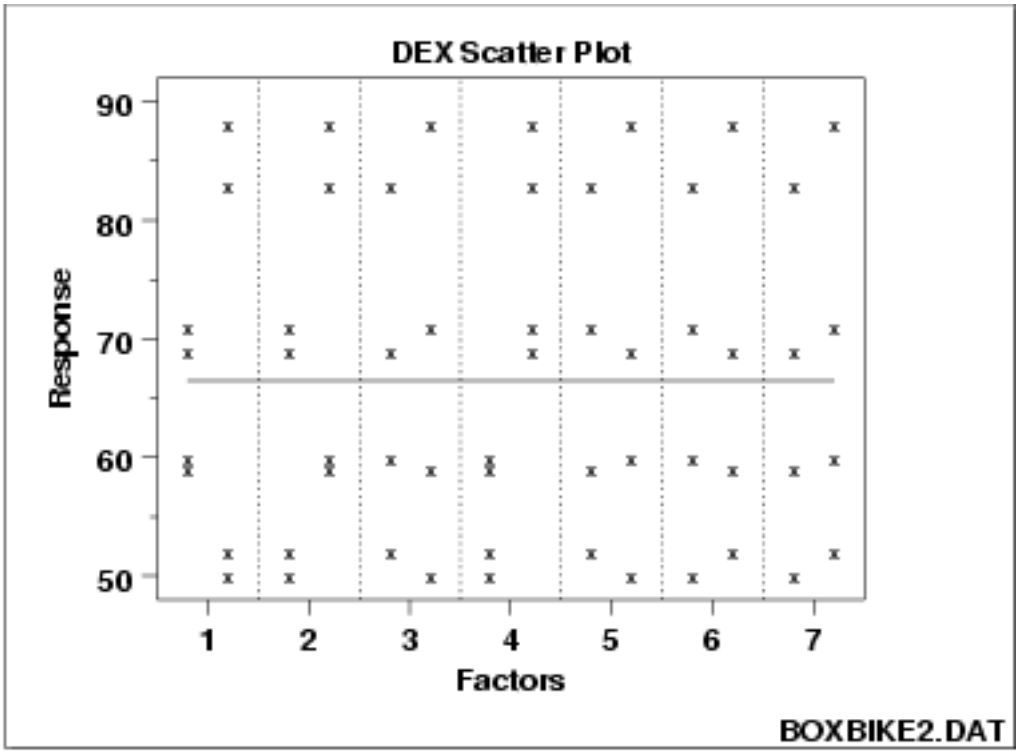
1.3.3.11. DEX Scatter Plot

Purpose: Determine Important Factors with Respect to Location and Scale

The dex scatter plot shows the response values for each level of each factor (i.e., independent) variable. This graphically shows how the location and scale vary for both within a factor variable and between different factor variables. This graphically shows which are the important factors and can help provide a ranked list of important factors from a designed experiment. The dex scatter plot is a complement to the traditional analyis of variance of designed experiments.

Dex scatter plots are typically used in conjunction with the dex mean plot and the dex standard deviation plot. The dex mean plot replaces the raw response values with mean response values while the dex standard deviation plot replaces the raw response values with the standard deviation of the response values. There is value in generating all 3 of these plots. The dex mean and standard deviation plots are useful in that the summary measures of location and spread stand out (they can sometimes get lost with the raw plot). However, the raw data points can reveal subtleties, such as the presence of outliers, that might get lost with the summary statistics.

Sample Plot: Factors 4, 2, 3, and 7 are the Important Factors.



Description of the Plot

For this sample plot, there are seven factors and each factor has two levels. For each factor, we define a distinct *x* coordinate for each level of the factor. For example, for factor 1, level 1 is coded as 0.8 and level 2 is coded as 1.2. The *y* coordinate is simply the value of the response variable. The solid horizontal line is drawn at the overall mean of the response variable. The vertical dotted lines are added for clarity.

Although the plot can be drawn with an arbitrary number of levels for a factor, it is really only useful when there are two or three levels for a factor.

Conclusions

This sample dex scatter plot shows that:

1. there does not appear to be any outliers;
2. the levels of factors 2 and 4 show distinct location differences; and
3. the levels of factor 1 show distinct scale differences.

Definition: Response Values Versus Factor Variables

Dex scatter plots are formed by:

- Vertical axis: Value of the response variable
- Horizontal axis: Factor variable (with each level of the factor coded with a slightly offset *x* coordinate)

Questions

The dex scatter plot can be used to answer the following questions:

1. Which factors are important with respect to location and scale?
2. Are there outliers?

Importance: Identify Important Factors with Respect to Location and Scale

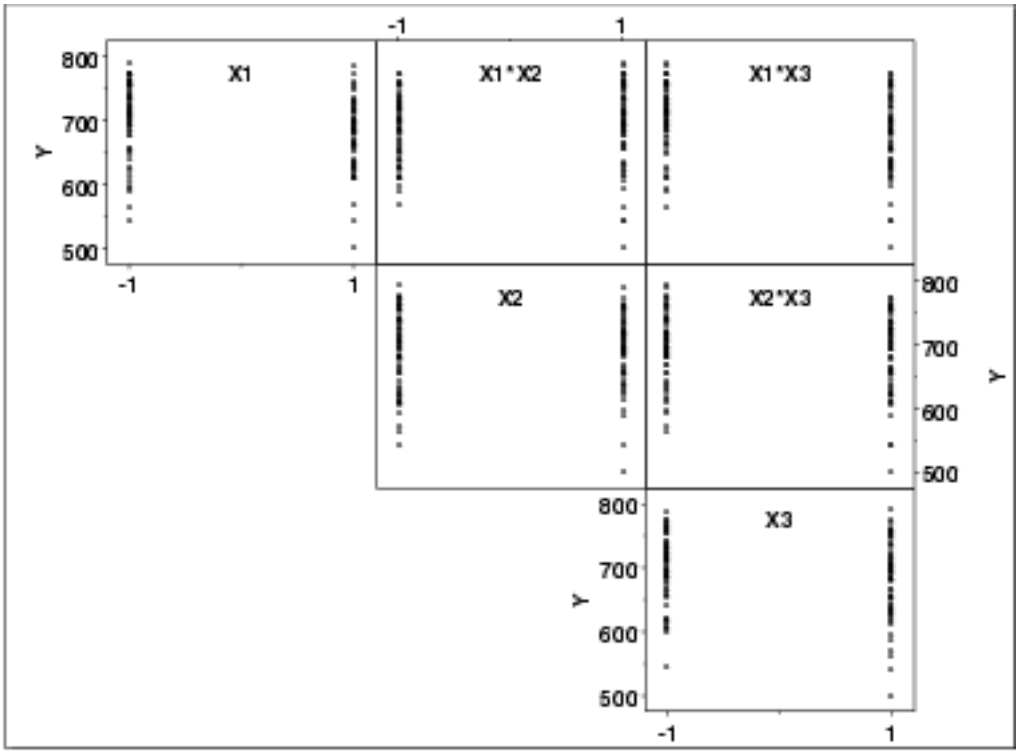
The goal of many designed experiments is to determine which factors are important with respect to location and scale. A ranked list of the important factors is also often of interest. Dex scatter, mean, and standard deviation plots show this graphically. The dex scatter plot additionally shows if outliers may potentially be distorting the results.

Dex scatter plots were designed primarily for analyzing designed experiments. However, they are useful for any type of multi-factor data (i.e., a response variable with 2 or more factor variables having a small number of distinct levels) whether or not the data were generated from a designed experiment.

Extension for Interaction Effects

Using the concept of the scatterplot matrix, the dex scatter plot can be extended to display first order interaction effects.

Specifically, if there are k factors, we create a matrix of plots with k rows and k columns. On the diagonal, the plot is simply a dex scatter plot with a single factor. For the off-diagonal plots, we multiply the values of $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$. For the common 2-level designs (i.e., each factor has two levels) the values are typically coded as -1 and 1, so the multiplied values are also -1 and 1. We then generate a dex scatter plot for this interaction variable. This plot is called a dex interaction effects plot and an example is shown below.



Interpretation of the Dex Interaction Effects Plot

We can first examine the diagonal elements for the main effects. These diagonal plots show a great deal of overlap between the levels for all three factors. This indicates that location and scale effects will be relatively small.

We can then examine the off-diagonal plots for the first order interaction effects. For example, the plot in the first row and second column is the interaction between factors X1 and X2. As with the main effect plots, no clear patterns are evident.

Related Techniques

Dex mean plot
Dex standard deviation plot
Block plot
Box plot
Analysis of variance

Case Study

The dex scatter plot is demonstrated in the ceramic strength data case study.

Software

Dex scatter plots are available in some general purpose statistical software programs, although the format may vary somewhat between these programs. They are essentially just scatter plots with the X variable defined in a particular way, so it should be feasible to write macros for dex scatter plots in most statistical software programs. Dataplot supports a dex scatter plot.

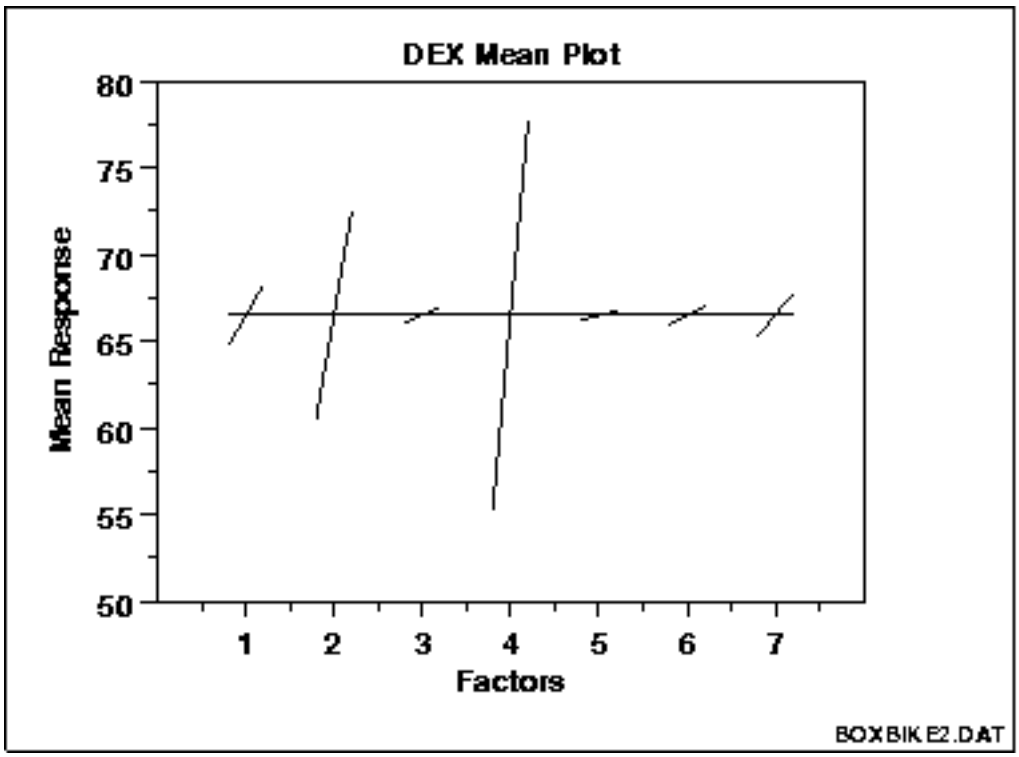
1.3.3.12. DEX Mean Plot

Purpose: Detect Important Factors with Respect to Location

The dex mean plot is appropriate for analyzing data from a designed experiment, with respect to important factors, where the factors are at two or more levels. The plot shows mean values for the two or more levels of each factor plotted by factor. The means for a single factor are connected by a straight line. The dex mean plot is a complement to the traditional analysis of variance of designed experiments.

This plot is typically generated for the mean. However, it can be generated for other location statistics such as the median.

Sample Plot: Factors 4, 2, and 1 are the Most Important Factors



This sample dex mean plot shows that:

1. factor 4 is the most important;
2. factor 2 is the second most important;
3. factor 1 is the third most important;

4. factor 7 is the fourth most important;
5. factor 6 is the fifth most important;
6. factors 3 and 5 are relatively unimportant.

In summary, factors 4, 2, and 1 seem to be clearly important, factors 3 and 5 seem to be clearly unimportant, and factors 6 and 7 are borderline factors whose inclusion in any subsequent models will be determined by further analyses.

Definition: Mean Response Versus Factor Variables

Dex mean plots are formed by:

- Vertical axis: Mean of the response variable for each level of the factor
- Horizontal axis: Factor variable

Questions

The dex mean plot can be used to answer the following questions:

1. Which factors are important? The dex mean plot does not provide a definitive answer to this question, but it does help categorize factors as "clearly important", "clearly not important", and "borderline importance".
2. What is the ranking list of the important factors?

Importance: Determine Significant Factors

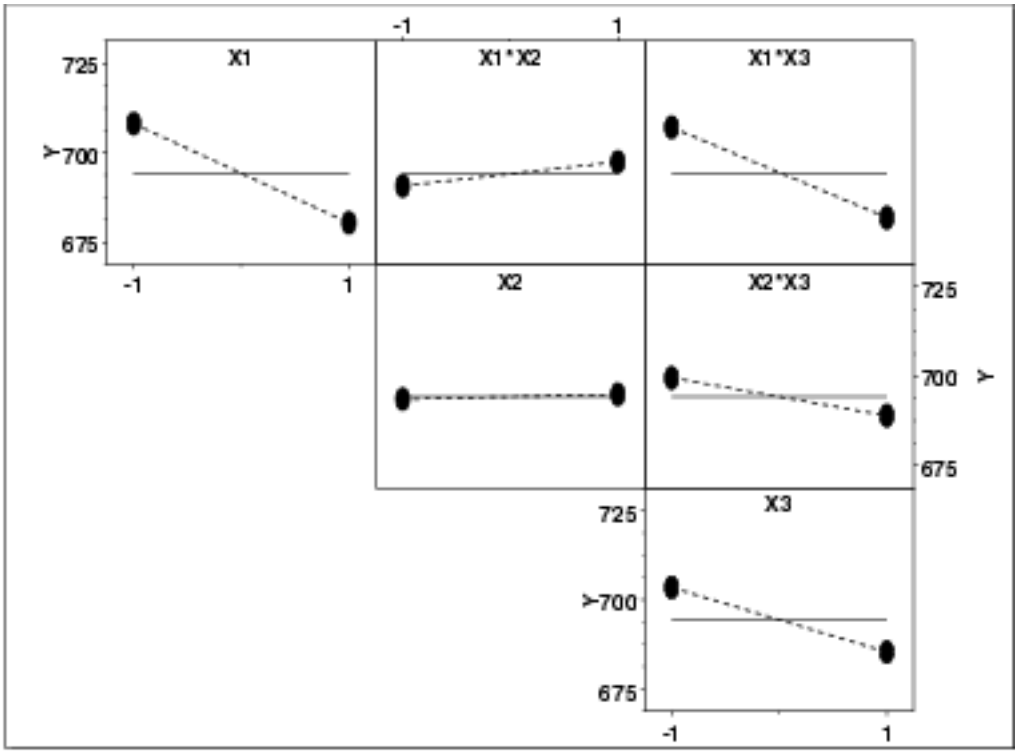
The goal of many designed experiments is to determine which factors are significant. A ranked order listing of the important factors is also often of interest. The dex mean plot is ideally suited for answering these types of questions and we recommend its routine use in analyzing designed experiments.

Extension for Interaction Effects

Using the concept of the scatter plot matrix, the dex mean plot can be extended to display first-order interaction effects.

Specifically, if there are k factors, we create a matrix of plots with k rows and k columns. On the diagonal, the plot is simply a dex mean plot with a single factor. For the off-diagonal plots, measurements at each level of the interaction are plotted versus level, where level is X_i times X_j and X_i is the code for the ith main effect level and X_j is the code for the jth main effect. For the common 2-level designs (i.e., each factor has two levels) the values are typically coded as -1 and 1, so the multiplied values are also -1 and 1. We then generate a dex mean plot for this interaction variable. This plot is called a dex interaction effects plot and an example is shown below.

DEX Interaction Effects Plot



This plot shows that the most significant factor is X1 and the most significant interaction is between X1 and X3.

Related Techniques

Dex scatter plot
Dex standard deviation plot
Block plot
Box plot
Analysis of variance

Case Study

The dex mean plot and the dex interaction effects plot are demonstrated in the ceramic strength data case study.

Software

Dex mean plots are available in some general purpose statistical software programs, although the format may vary somewhat between these programs. It may be feasible to write macros for dex mean plots in some statistical software programs that do not support this plot directly. Dataplot supports both a dex mean plot and a dex interaction effects plot.

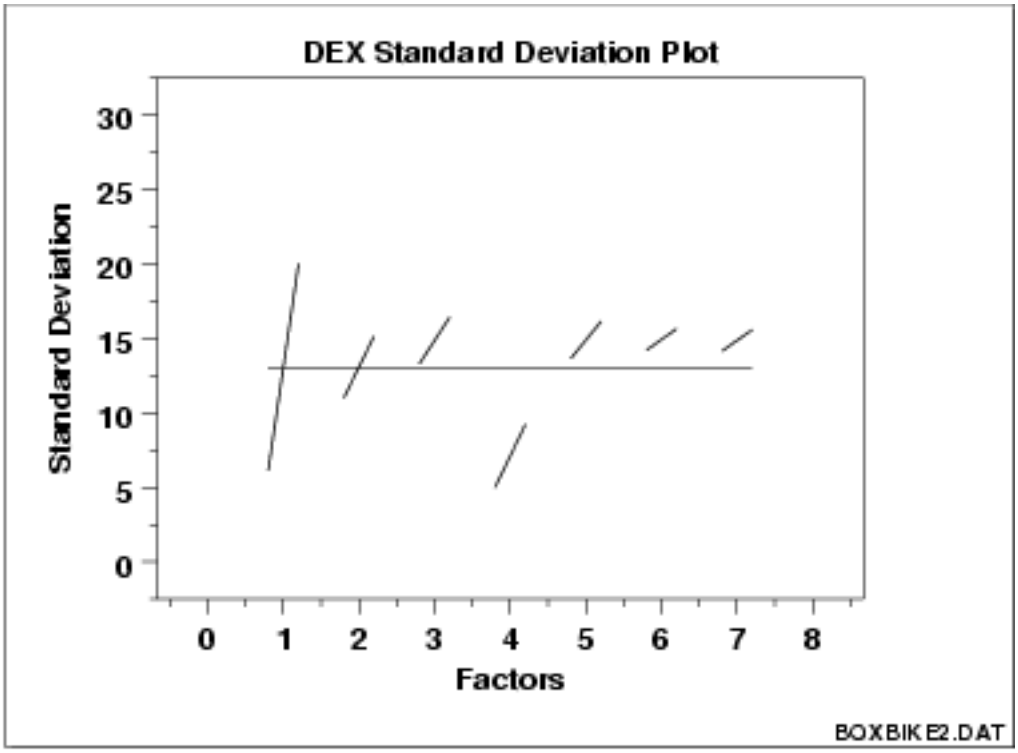
1.3.3.13. DEX Standard Deviation Plot

Purpose: Detect Important Factors with Respect to Scale

The dex standard deviation plot is appropriate for analyzing data from a designed experiment, with respect to important factors, where the factors are at two or more levels and there are repeated values at each level. The plot shows standard deviation values for the two or more levels of each factor plotted by factor. The standard deviations for a single factor are connected by a straight line. The dex standard deviation plot is a complement to the traditional analysis of variance of designed experiments.

This plot is typically generated for the standard deviation. However, it can also be generated for other scale statistics such as the range, the median absolute deviation, or the average absolute deviation.

Sample Plot



This sample dex standard deviation plot shows that:

1. factor 1 has the greatest difference in standard deviations between factor levels;
2. factor 4 has a significantly lower average standard deviation than the average standard deviations of other factors (but the level 1 standard deviation for factor 1 is about the same as the level 1 standard deviation for factor 4);
3. for all factors, the level 1 standard deviation is smaller than the level 2 standard deviation.

Definition: Response Standard Deviations Versus Factor Variables

Dex standard deviation plots are formed by:

- Vertical axis: Standard deviation of the response variable for each level of the factor
- Horizontal axis: Factor variable

Questions

The dex standard deviation plot can be used to answer the following questions:

1. How do the standard deviations vary across factors?
2. How do the standard deviations vary within a factor?
3. Which are the most important factors with respect to scale?
4. What is the ranked list of the important factors with respect to scale?

Importance: Assess Variability

The goal with many designed experiments is to determine which factors are significant. This is usually determined from the means of the factor levels (which can be conveniently shown with a dex mean plot). A secondary goal is to assess the variability of the responses both within a factor and between factors. The dex standard deviation plot is a convenient way to do this.

Related Techniques

Dex scatter plot
Dex mean plot
Block plot
Box plot
Analysis of variance

Case Study

The dex standard deviation plot is demonstrated in the ceramic strength data case study.

Software

Dex standard deviation plots are not available in most general purpose statistical software programs. It may be feasible to write macros for dex standard deviation plots in some statistical software programs that do not support them directly. Dataplot supports a dex standard deviation plot.

1.3.3.14. Histogram

Purpose: Summarize a Univariate Data Set

The purpose of a histogram (Chambers) is to graphically summarize the distribution of a univariate data set.

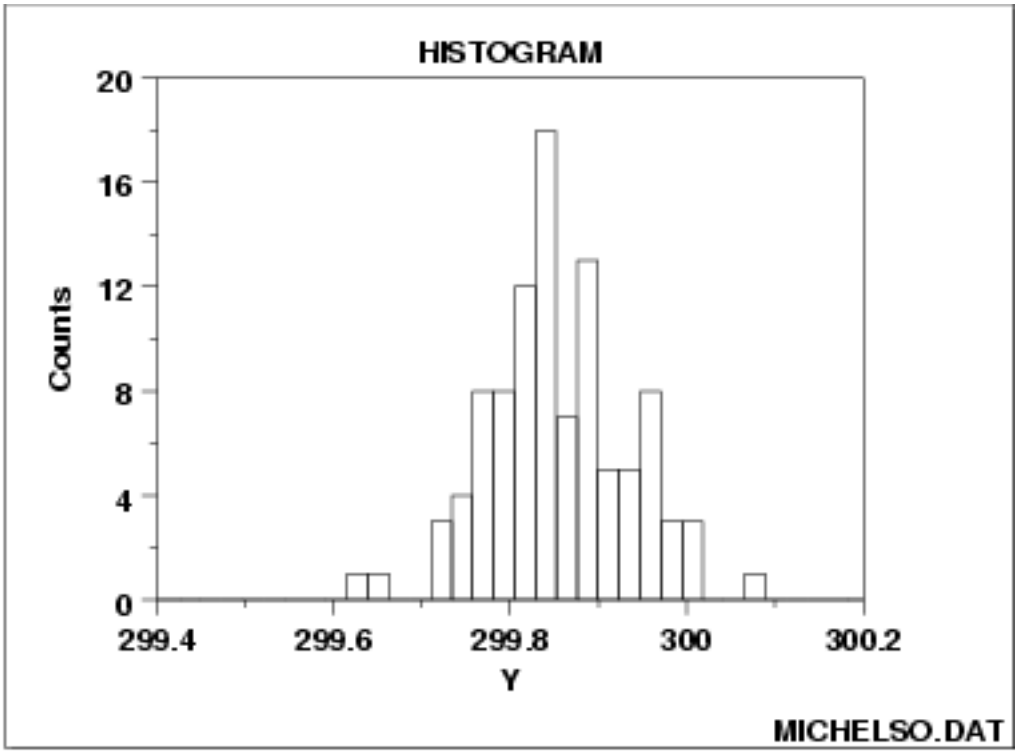
The histogram graphically shows the following:

1. center (i.e., the location) of the data;
2. spread (i.e., the scale) of the data;
3. skewness of the data;
4. presence of outliers; and
5. presence of multiple modes in the data.

These features provide strong indications of the proper distributional model for the data. The probability plot or a goodness-of-fit test can be used to verify the distributional model.

The examples section shows the appearance of a number of common features revealed by histograms.

Sample Plot



Definition

The most common form of the histogram is obtained by splitting the range of the data into equal-sized bins (called classes). Then for each bin, the number of points from the data set that fall into each bin are counted. That is

- Vertical axis: Frequency (i.e., counts for each bin)
- Horizontal axis: Response variable

The classes can either be defined arbitrarily by the user or via some systematic rule. A number of theoretically derived rules have been proposed by Scott (Scott 1992).

The cumulative histogram is a variation of the histogram in which the vertical axis gives not just the counts for a single bin, but rather gives the counts for that bin plus all bins for smaller values of the response variable.

Both the histogram and cumulative histogram have an additional variant whereby the counts are replaced by the normalized counts. The names for these variants are the relative histogram and the relative cumulative histogram.

There are two common ways to normalize the counts.

1. The normalized count is the count in a class divided by the total number of observations. In this case the relative counts are normalized to sum to one (or 100 if a percentage scale is used). This is the intuitive case where the height of the histogram bar represents the proportion of the data in each class.
2. The normalized count is the count in the class divided by the

number of observations times the class width. For this normalization, the area (or integral) under the histogram is equal to one. From a probabilistic point of view, this normalization results in a relative histogram that is most akin to the probability density function and a relative cumulative histogram that is most akin to the cumulative distribution function. If you want to overlay a probability density or cumulative distribution function on top of the histogram, use this normalization. Although this normalization is less intuitive (relative frequencies greater than 1 are quite permissible), it is the appropriate normalization if you are using the histogram to model a probability density function.

Questions

The histogram can be used to answer the following questions:

1. What kind of population distribution do the data come from?
2. Where are the data located?
3. How spread out are the data?
4. Are the data symmetric or skewed?
5. Are there outliers in the data?

Examples

1. Normal
2. Symmetric, Non-Normal, Short-Tailed
3. Symmetric, Non-Normal, Long-Tailed
4. Symmetric and Bimodal
5. Bimodal Mixture of 2 Normals
6. Skewed (Non-Symmetric) Right
7. Skewed (Non-Symmetric) Left
8. Symmetric with Outlier

Related Techniques

Box plot
Probability plot

The techniques below are not discussed in the Handbook. However, they are similar in purpose to the histogram. Additional information on them is contained in the Chambers and Scott references.

Frequency Plot
Stem and Leaf Plot
Density Trace

Case Study

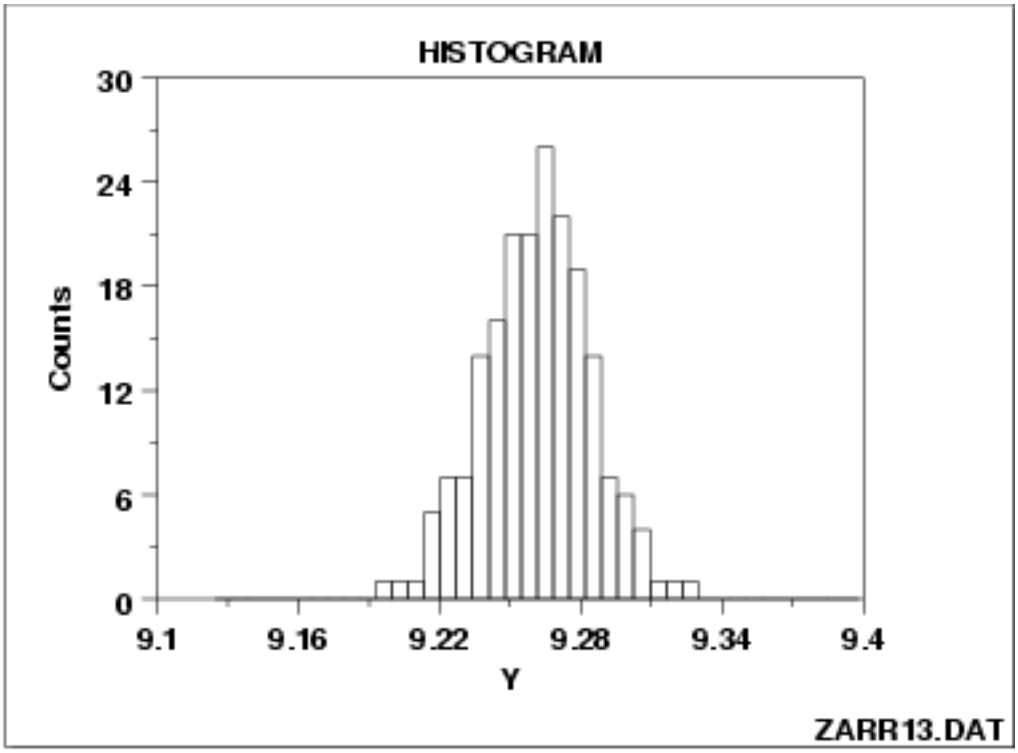
The histogram is demonstrated in the heat flow meter data case study.

Software

Histograms are available in most general purpose statistical software programs. They are also supported in most general purpose charting, spreadsheet, and business graphics programs. Dataplot supports histograms.

1.3.3.14.1. Histogram Interpretation: Normal

Symmetric, Moderate-Tailed Histogram



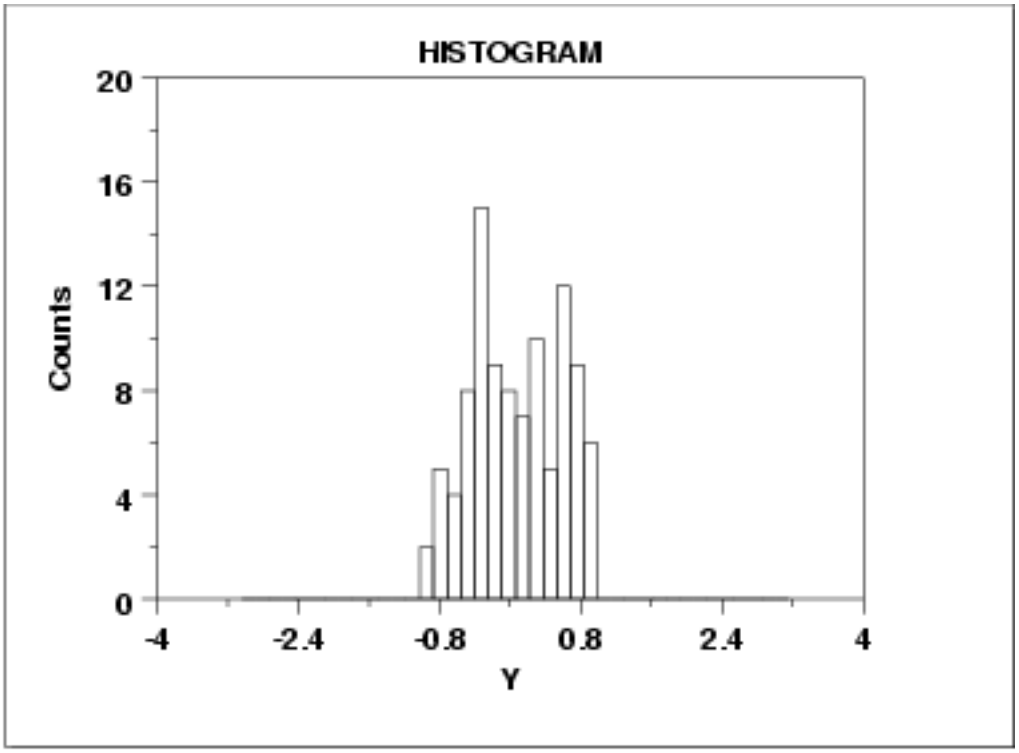
Note the classical bell-shaped, symmetric histogram with most of the frequency counts bunched in the middle and with the counts dying off out in the tails. From a physical science/engineering point of view, the normal distribution is that distribution which occurs most often in nature (due in part to the central limit theorem).

Recommended Next Step

If the histogram indicates a symmetric, moderate tailed distribution, then the recommended next step is to do a normal probability plot to confirm approximate normality. If the normal probability plot is linear, then the normal distribution is a good model for the data.

1.3.3.14.2. Histogram Interpretation: Symmetric, Non-Normal, Short-Tailed

Symmetric, Short-Tailed Histogram



Description of What Short-Tailed Means

For a symmetric distribution, the "body" of a distribution refers to the "center" of the distribution--commonly that region of the distribution where most of the probability resides--the "fat" part of the distribution. The "tail" of a distribution refers to the extreme regions of the distribution--both left and right. The "tail length" of a distribution is a term that indicates how fast these extremes approach zero.

For a short-tailed distribution, the tails approach zero very fast. Such distributions commonly have a truncated ("sawed-off") look. The classical short-tailed distribution is the uniform (rectangular) distribution in which the probability is constant over a given range and then drops to zero everywhere else--we would speak of this as having no tails, or extremely short tails.

For a moderate-tailed distribution, the tails decline to zero in a moderate fashion. The classical moderate-tailed distribution is the normal (Gaussian) distribution.

For a long-tailed distribution, the tails decline to zero very slowly--and hence one is apt to see probability a long way from the body of the distribution. The classical long-tailed distribution is the Cauchy distribution.

In terms of tail length, the histogram shown above would be characteristic of a "short-tailed" distribution.

The optimal (unbiased and most precise) estimator for location for the center of a distribution is heavily dependent on the tail length of the distribution. The common choice of taking N observations and using the calculated sample mean as the best estimate for the center of the distribution is a good choice for the normal distribution (moderate tailed), a poor choice for the uniform distribution (short tailed), and a horrible choice for the Cauchy distribution (long tailed). Although for the normal distribution the sample mean is as precise an estimator as we can get, for the uniform and Cauchy distributions, the sample mean is not the best estimator.

For the uniform distribution, the midrange

midrange = (smallest + largest) / 2

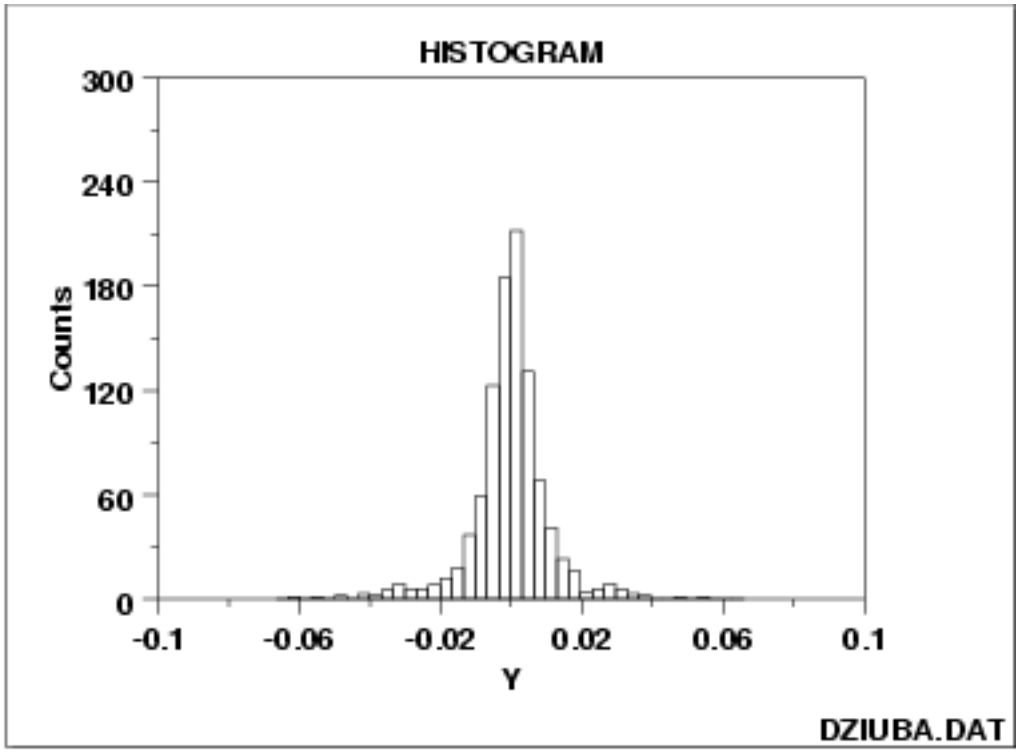
is the best estimator of location. For a Cauchy distribution, the median is the best estimator of location.

Recommended Next Step

If the histogram indicates a symmetric, short-tailed distribution, the recommended next step is to generate a uniform probability plot. If the uniform probability plot is linear, then the uniform distribution is an appropriate model for the data.

1.3.3.14.3. Histogram Interpretation: Symmetric, Non-Normal, Long-Tailed

Symmetric, Long-Tailed Histogram



Description of Long-Tailed

The previous example contains a discussion of the distinction between short-tailed, moderate-tailed, and long-tailed distributions.

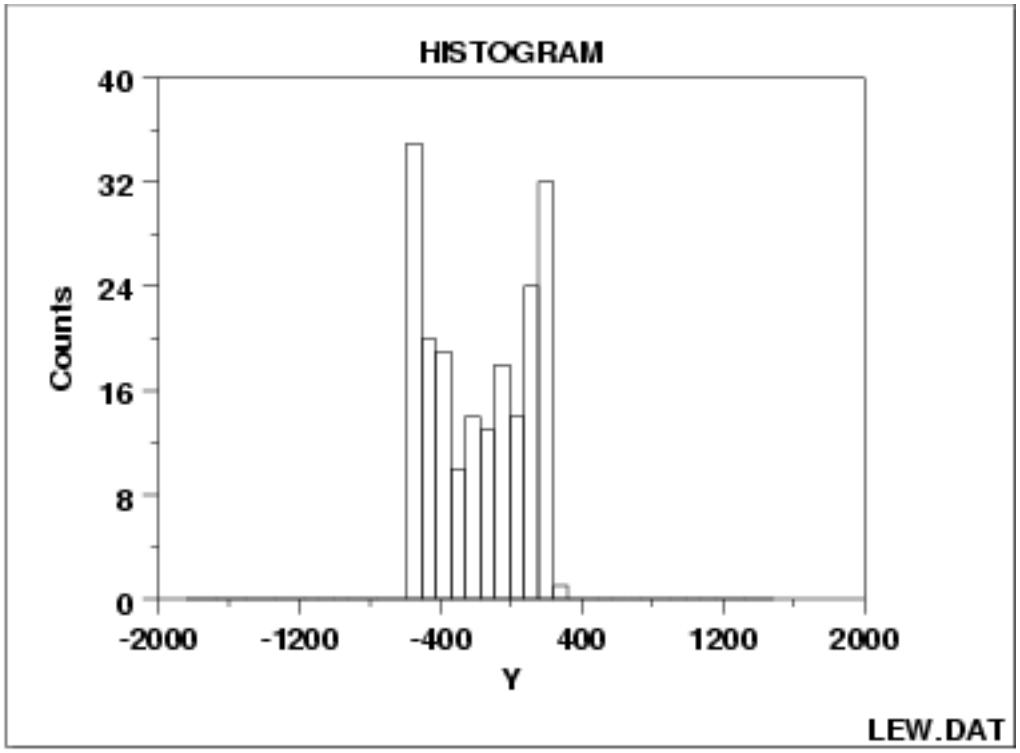
In terms of tail length, the histogram shown above would be characteristic of a "long-tailed" distribution.

Recommended Next Step

If the histogram indicates a symmetric, long tailed distribution, the recommended next step is to do a Cauchy probability plot. If the Cauchy probability plot is linear, then the Cauchy distribution is an appropriate model for the data. Alternatively, a Tukey Lambda PPCC plot may provide insight into a suitable distributional model for the data.

1.3.3.14.4. Histogram Interpretation: Symmetric and Bimodal

Symmetric, Bimodal Histogram



Description of Bimodal

The mode of a distribution is that value which is most frequently occurring or has the largest probability of occurrence. The sample mode occurs at the peak of the histogram.

For many phenomena, it is quite common for the distribution of the response values to cluster around a single mode (unimodal) and then distribute themselves with lesser frequency out into the tails. The normal distribution is the classic example of a unimodal distribution.

The histogram shown above illustrates data from a bimodal (2 peak) distribution. The histogram serves as a tool for diagnosing problems such as bimodality. Questioning the underlying reason for distributional non-unimodality frequently leads to greater insight and

improved deterministic modeling of the phenomenon under study. For example, for the data presented above, the bimodal histogram is caused by sinusoidality in the data.

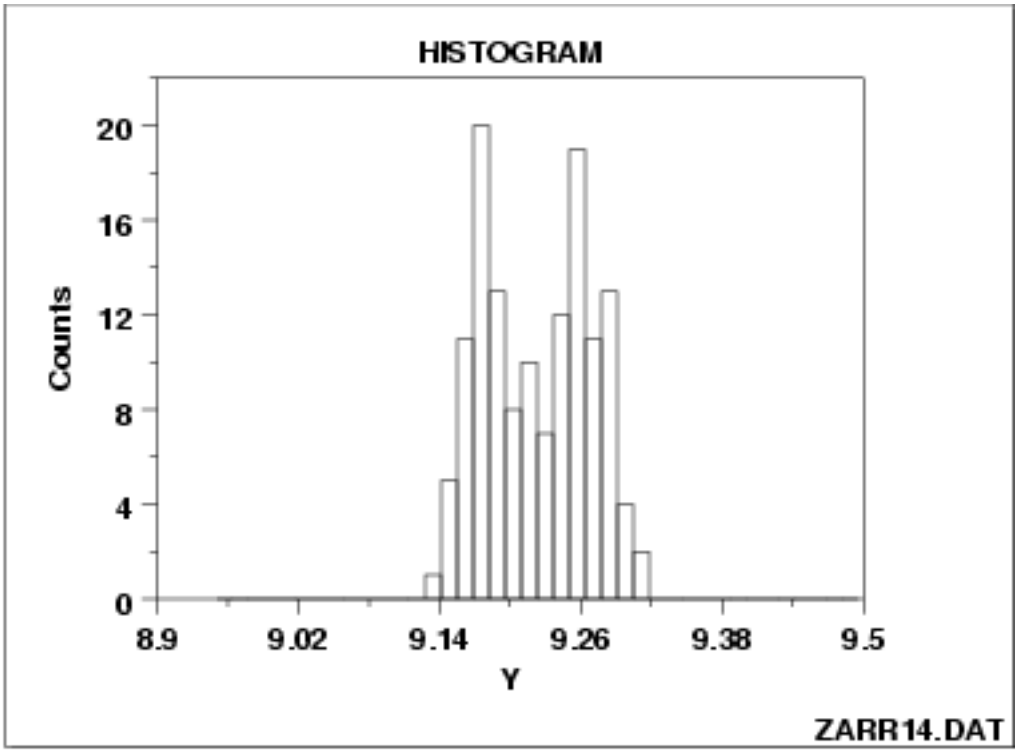
Recommended Next Step

If the histogram indicates a symmetric, bimodal distribution, the recommended next steps are to:

1. Do a run sequence plot or a scatter plot to check for sinusoidality.
2. Do a lag plot to check for sinusoidality. If the lag plot is elliptical, then the data are sinusoidal.
3. If the data are sinusoidal, then a spectral plot is used to graphically estimate the underlying sinusoidal frequency.
4. If the data are not sinusoidal, then a Tukey Lambda PPCC plot may determine the best-fit symmetric distribution for the data.
5. The data may be fit with a mixture of two distributions. A common approach to this case is to fit a mixture of 2 normal or lognormal distributions. Further discussion of fitting mixtures of distributions is beyond the scope of this Handbook.

1.3.3.14.5. Histogram Interpretation: Bimodal Mixture of 2 Normals

Histogram from Mixture of 2 Normal Distributions



Discussion of Unimodal and Bimodal

The histogram shown above illustrates data from a bimodal (2 peak) distribution.

In contrast to the previous example, this example illustrates bimodality due not to an underlying deterministic model, but bimodality due to a mixture of probability models. In this case, each of the modes appears to have a rough bell-shaped component. One could easily imagine the above histogram being generated by a process consisting of two normal distributions with the same standard deviation but with two different locations (one centered at approximately 9.17 and the other centered at approximately 9.26). If this is the case, then the research challenge is to determine physically why there are two similar but separate sub-processes.

Recommended Next Steps

If the histogram indicates that the data might be appropriately fit with a mixture of two normal distributions, the recommended next step is:

Fit the normal mixture model using either least squares or maximum likelihood. The general normal mixing model is

$$M = p\phi_1 + (1-p)\phi_2$$

where *p* is the mixing proportion (between 0 and 1) and ϕ_1 and ϕ_2 are normal probability density functions with location and scale parameters μ_1, σ_1, μ_2, and σ_2, respectively. That is, there are 5 parameters to estimate in the fit.

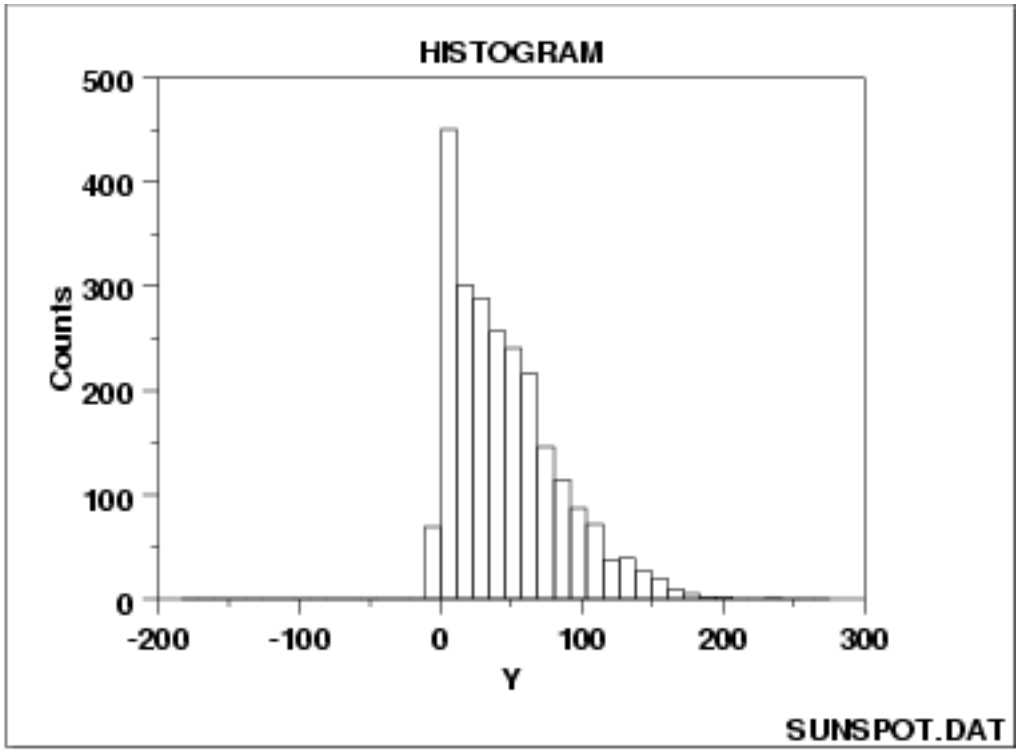
Whether maximum likelihood or least squares is used, the quality of the fit is sensitive to good starting values. For the mixture of two normals, the histogram can be used to provide initial estimates for the location and scale parameters of the two normal distributions.

Dataplot can generate a least squares fit of the mixture of two normals with the following sequence of commands:

```
RELATIVE HISTOGRAM Y
LET Y2 = YPLOT
LET X2 = XPLOT
RETAIN Y2 X2 SUBSET TAGPLOT = 1
LET U1 = <estimated value from histogram>
LET SD1 = <estimated value from histogram>
LET U2 = <estimated value from histogram>
LET SD2 = <estimated value from histogram>
LET P = 0.5
FIT Y2 = NORMXPDF(X2,U1,S1,U2,S2,P)
```

1.3.3.14.6. Histogram Interpretation: Skewed (Non-Normal) Right

Right-Skewed Histogram



Discussion of Skewness

A symmetric distribution is one in which the 2 "halves" of the histogram appear as mirror-images of one another. A skewed (non-symmetric) distribution is a distribution in which there is no such mirror-imaging.

For skewed distributions, it is quite common to have one tail of the distribution considerably longer or drawn out relative to the other tail. A "skewed right" distribution is one in which the tail is on the right side. A "skewed left" distribution is one in which the tail is on the left side. The above histogram is for a distribution that is skewed right.

Skewed distributions bring a certain philosophical complexity to the very process of estimating a "typical value" for the distribution. To be

specific, suppose that the analyst has a collection of 100 values randomly drawn from a distribution, and wishes to summarize these 100 observations by a "typical value". What does typical value mean? If the distribution is symmetric, the typical value is unambiguous-- it is a well-defined center of the distribution. For example, for a bell-shaped symmetric distribution, a center point is identical to that value at the peak of the distribution.

For a skewed distribution, however, there is no "center" in the usual sense of the word. Be that as it may, several "typical value" metrics are often used for skewed distributions. The first metric is the mode of the distribution. Unfortunately, for severely-skewed distributions, the mode may be at or near the left or right tail of the data and so it seems not to be a good representative of the center of the distribution. As a second choice, one could conceptually argue that the mean (the point on the horizontal axis where the distributiuon would balance) would serve well as the typical value. As a third choice, others may argue that the median (that value on the horizontal axis which has exactly 50% of the data to the left (and also to the right) would serve as a good typical value.

For symmetric distributions, the conceptual problem disappears because at the population level the mode, mean, and median are identical. For skewed distributions, however, these 3 metrics are markedly different. In practice, for skewed distributions the most commonly reported typical value is the mean; the next most common is the median; the least common is the mode. Because each of these 3 metrics reflects a different aspect of "centerness", it is recommended that the analyst report at least 2 (mean and median), and preferably all 3 (mean, median, and mode) in summarizing and characterizing a data set.

Some Causes for Skewed Data

Skewed data often occur due to lower or upper bounds on the data. That is, data that have a lower bound are often skewed right while data that have an upper bound are often skewed left. Skewness can also result from start-up effects. For example, in reliability applications some processes may have a large number of initial failures that could cause left skewness. On the other hand, a reliability process could have a long start-up period where failures are rare resulting in right-skewed data.

Data collected in scientific and engineering applications often have a lower bound of zero. For example, failure data must be non-negative. Many measurement processes generate only positive data. Time to occurence and size are common measurements that cannot be less than zero.

Recommended Next Steps

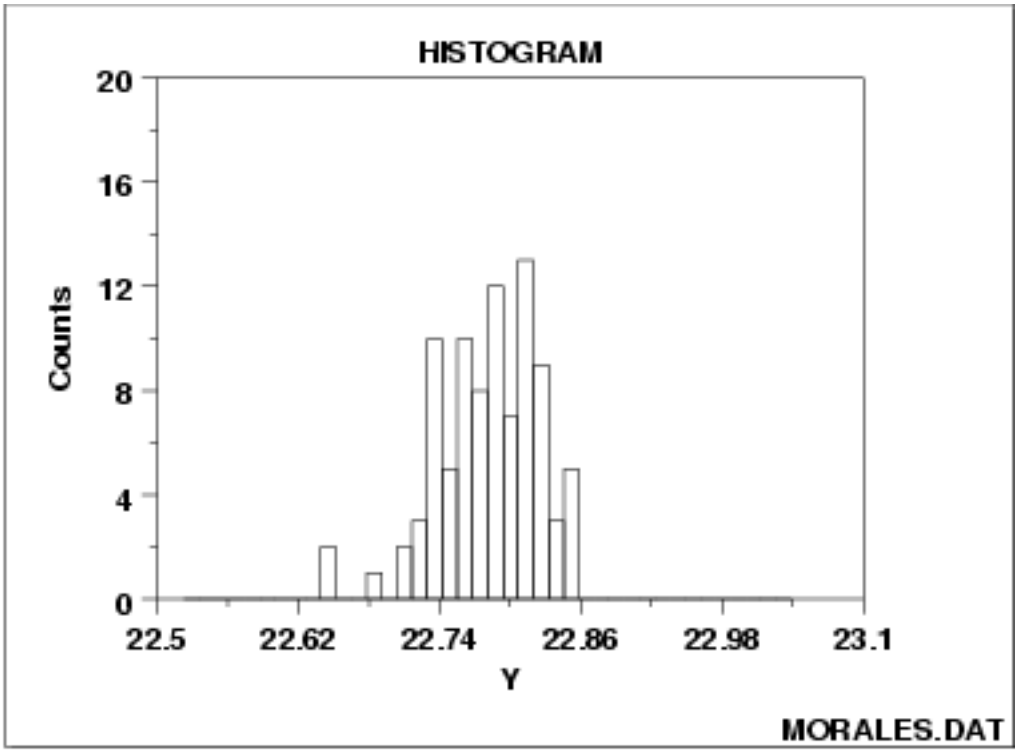
If the histogram indicates a right-skewed data set, the recommended next steps are to:

1. Quantitatively summarize the data by computing and reporting the sample mean, the sample median, and the sample mode.
2. Determine the best-fit distribution (skewed-right) from the
   - Weibull family (for the maximum)
   - Gamma family
   - Chi-square family
   - Lognormal family
   - Power lognormal family
3. Consider a normalizing transformation such as the Box-Cox transformation.

1. Exploratory Data Analysis
1.3. EDA Techniques
1.3.3. Graphical Techniques: Alphabetic
1.3.3.14. Histogram

1.3.3.14.7. Histogram Interpretation: Skewed (Non-Symmetric) Left

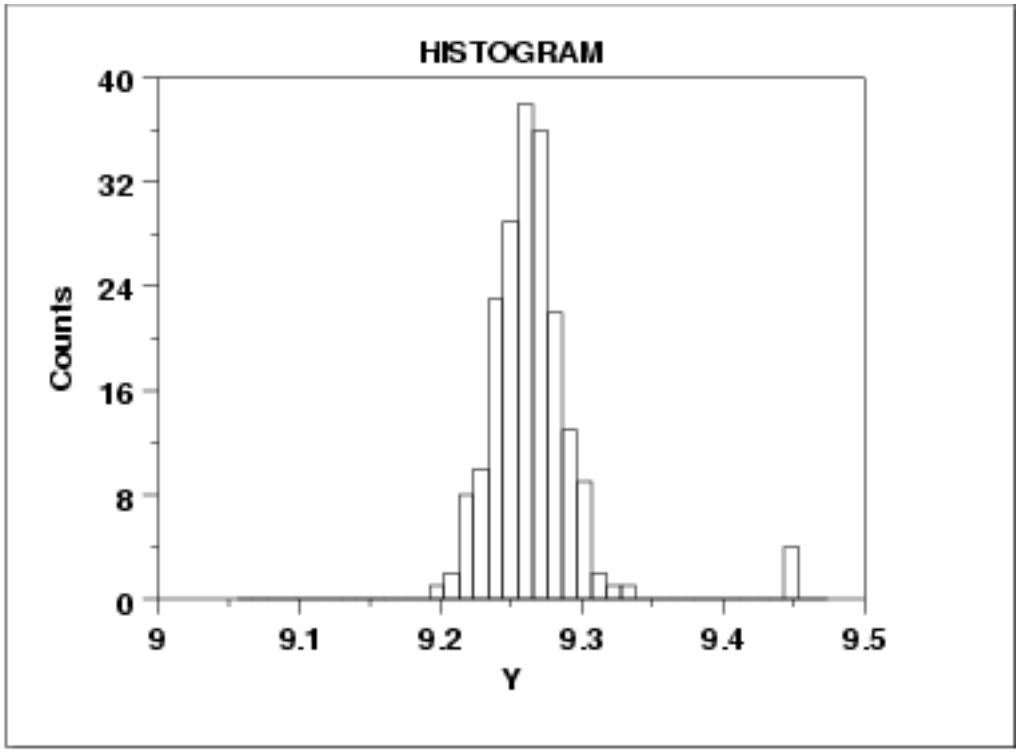
Skewed Left Histogram



The issues for skewed left data are similar to those for skewed right data.

1.3.3.14.8. Histogram Interpretation: Symmetric with Outlier

Symmetric Histogram with Outlier



Discussion of Outliers

A symmetric distribution is one in which the 2 "halves" of the histogram appear as mirror-images of one another. The above example is symmetric with the exception of outlying data near Y = 4.5.

An outlier is a data point that comes from a distribution different (in location, scale, or distributional form) from the bulk of the data. In the real world, outliers have a range of causes, from as simple as

1. operator blunders
2. equipment failures
3. day-to-day effects
4. batch-to-batch differences
5. anomalous input conditions

6. warm-up effects

to more subtle causes such as

1. A change in settings of factors that (knowingly or unknowingly) affect the response.
2. Nature is trying to tell us something.

Outliers Should be Investigated

All outliers should be taken seriously and should be investigated thoroughly for explanations. Automatic outlier-rejection schemes (such as throw out all data beyond 4 sample standard deviations from the sample mean) are particularly dangerous.

The classic case of automatic outlier rejection becoming automatic information rejection was the South Pole ozone depletion problem. Ozone depletion over the South Pole would have been detected years earlier except for the fact that the satellite data recording the low ozone readings had outlier-rejection code that automatically screened out the "outliers" (that is, the low ozone readings) before the analysis was conducted. Such inadvertent (and incorrect) purging went on for years. It was not until ground-based South Pole readings started detecting low ozone readings that someone decided to double-check as to why the satellite had not picked up this fact--it had, but it had gotten thrown out!

The best attitude is that outliers are our "friends", outliers are trying to tell us something, and we should not stop until we are comfortable in the explanation for each outlier.

Recommended Next Steps

If the histogram shows the presence of outliers, the recommended next steps are:

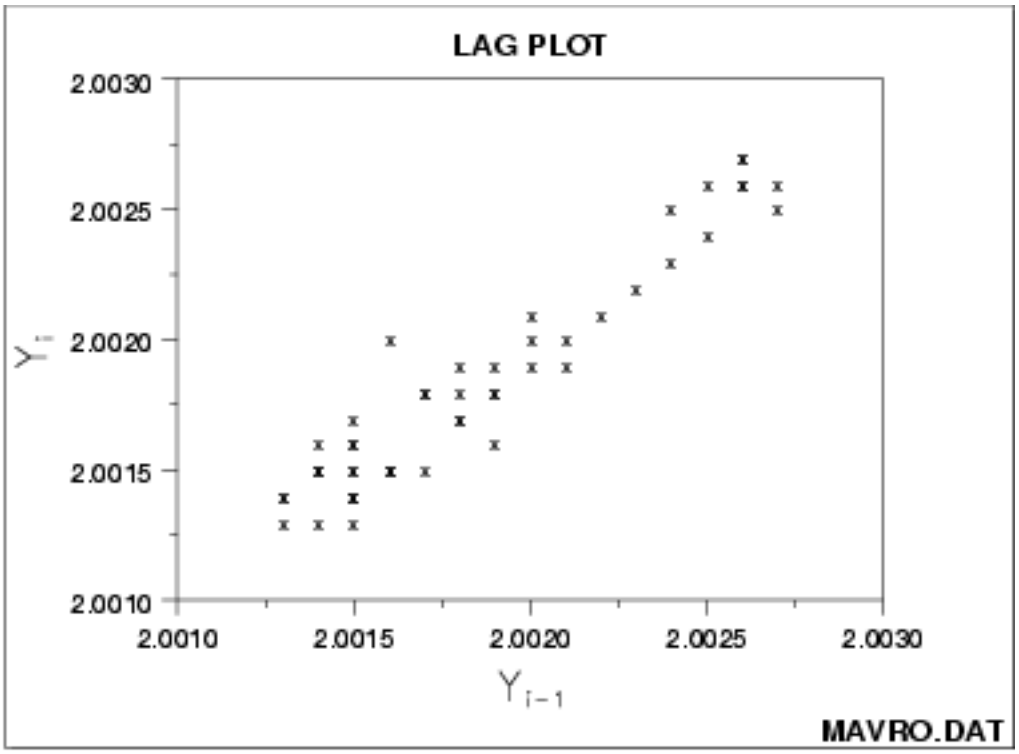
1. Graphically check for outliers (in the commonly encountered normal case) by generating a box plot. In general, box plots are a much better graphical tool for detecting outliers than are histograms.
2. Quantitatively check for outliers (in the commonly encountered normal case) by carrying out Grubbs test which indicates how many sample standard deviations away from the sample mean are the data in question. Large values indicate outliers.

1.3.3.15. Lag Plot

Purpose: Check for randomness

A lag plot checks whether a data set or time series is random or not. Random data should not exhibit any identifiable structure in the lag plot. Non-random structure in the lag plot indicates that the underlying data are not random. Several common patterns for lag plots are shown in the examples below.

Sample Plot



This sample lag plot exhibits a linear pattern. This shows that the data are strongly non-random and further suggests that an autoregressive model might be appropriate.

Definition

A lag is a fixed time displacement. For example, given a data set Y_1, Y_2 ..., Y_n, Y_2 and Y_7 have lag 5 since 7 - 2 = 5. Lag plots can be generated for any arbitrary lag, although the most commonly used lag is 1.

A plot of lag 1 is a plot of the values of Y_i versus Y_{i-1}

- Vertical axis: Y_i for all i
- Horizontal axis: Y_{i-1} for all i

Questions

Lag plots can provide answers to the following questions:

1. Are the data random?
2. Is there serial correlation in the data?
3. What is a suitable model for the data?
4. Are there outliers in the data?

Importance

Inasmuch as randomness is an underlying assumption for most statistical estimation and testing techniques, the lag plot should be a routine tool for researchers.

Examples

- Random (White Noise)
- Weak autocorrelation
- Strong autocorrelation and autoregressive model
- Sinusoidal model and outliers

Related Techniques

Autocorrelation Plot
Spectrum
Runs Test

Case Study

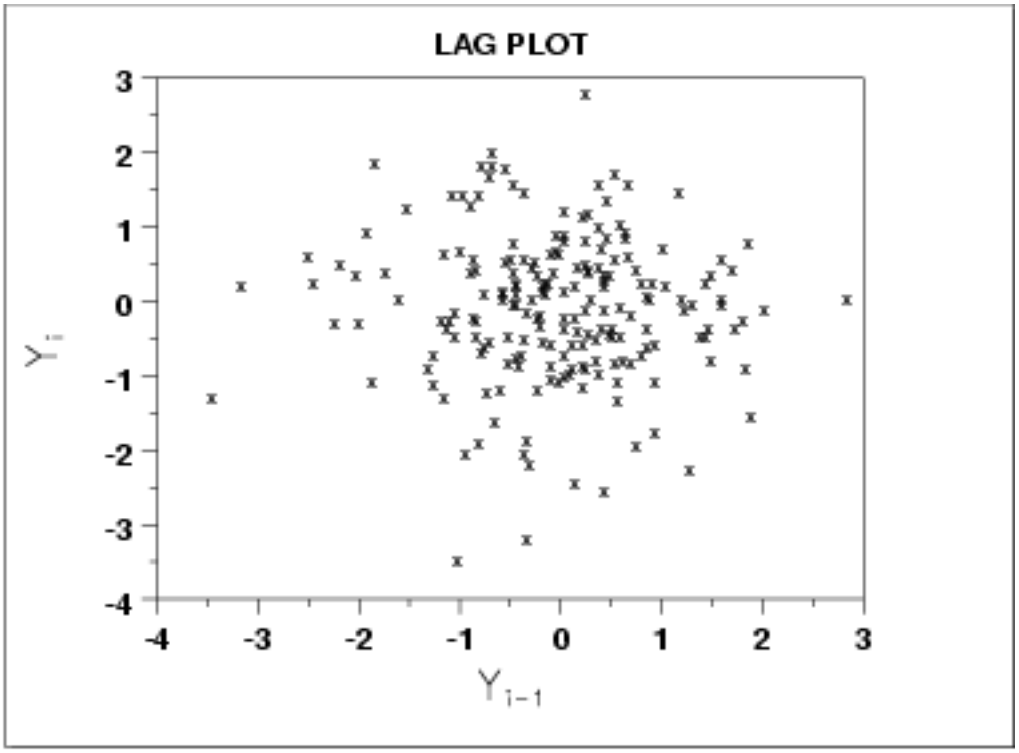
The lag plot is demonstrated in the beam deflection data case study.

Software

Lag plots are not directly available in most general purpose statistical software programs. Since the lag plot is essentially a scatter plot with the 2 variables properly lagged, it should be feasible to write a macro for the lag plot in most statistical programs. Dataplot supports a lag plot.

1.3.3.15.1. Lag Plot: Random Data

Lag Plot



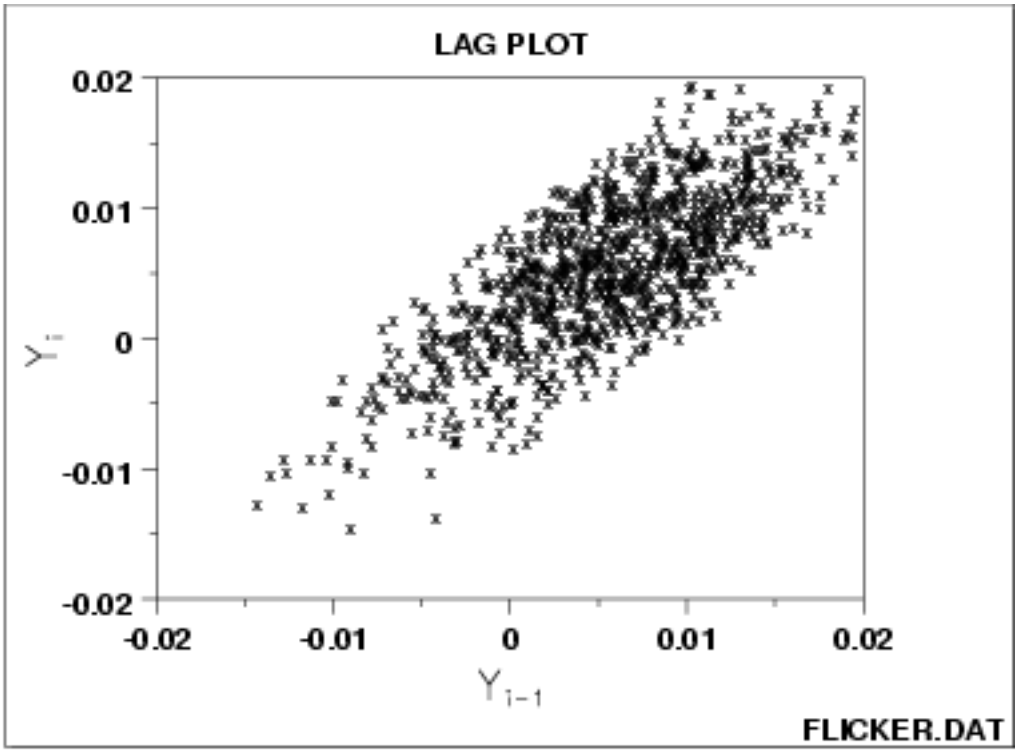
Conclusions We can make the following conclusions based on the above plot.

1. The data are random.
2. The data exhibit no autocorrelation.
3. The data contain no outliers.

Discussion The lag plot shown above is for lag = 1. Note the absence of structure. One cannot infer, from a current value Y_{i-1}, the next value Y_i. Thus for a known value Y_{i-1} on the horizontal axis (say, Y_{i-1} = +0.5), the Y_i-th value could be virtually anything (from Y_i = -2.5 to Y_i = +1.5). Such non-association is the essence of randomness.

1.3.3.15.2. Lag Plot: Moderate Autocorrelation

Lag Plot



Conclusions We can make the conclusions based on the above plot.

1. The data are from an underlying autoregressive model with moderate positive autocorrelation
2. The data contain no outliers.

Discussion

In the plot above for lag = 1, note how the points tend to cluster (albeit noisily) along the diagonal. Such clustering is the lag plot signature of moderate autocorrelation.

If the process were completely random, knowledge of a current observation (say $Y_{i-1} = 0$) would yield virtually no knowledge about the next observation Y_i. If the process has moderate autocorrelation, as above, and if $Y_{i-1} = 0$, then the range of possible values for Y_i is seen to be restricted to a smaller range (.01 to +.01). This suggests prediction is possible using an autoregressive model.

Recommended Next Step

Estimate the parameters for the autoregressive model:

$$Y_i = A_0 + A_1 * Y_{i-1} + E_i$$

Since Y_i and Y_{i-1} are precisely the axes of the lag plot, such estimation is a linear regression straight from the lag plot.

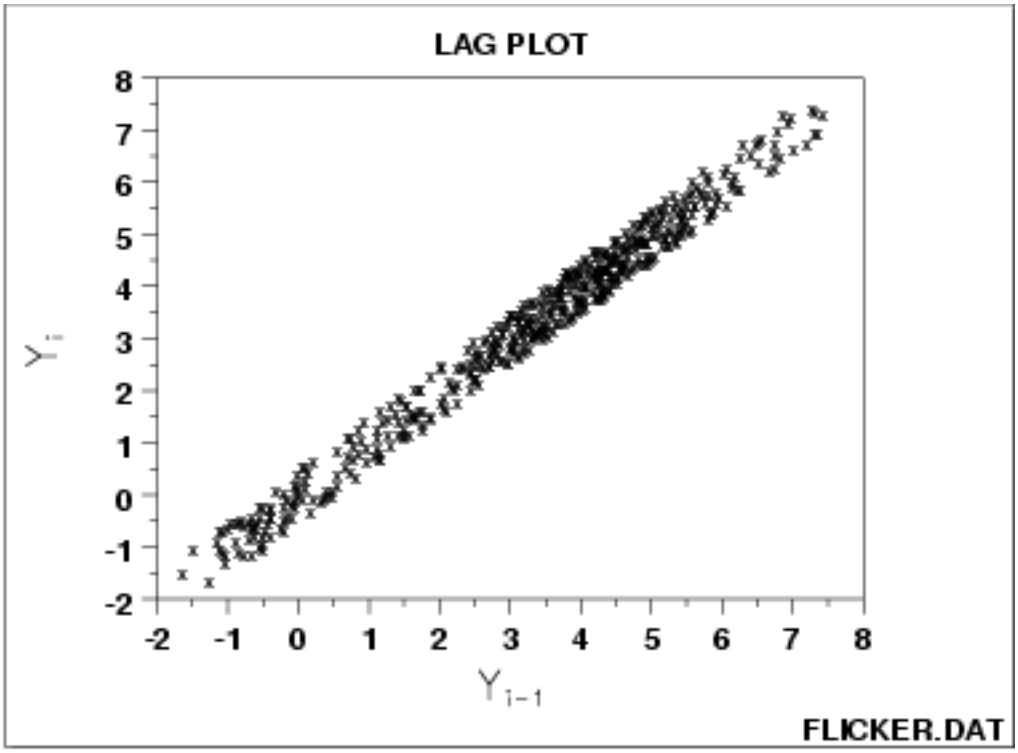
The residual standard deviation for the autoregressive model will be much smaller than the residual standard deviation for the default model

$$Y_i = A_0 + E_i$$

1. Exploratory Data Analysis
1.3. EDA Techniques
1.3.3. Graphical Techniques: Alphabetic
1.3.3.15. Lag Plot

1.3.3.15.3. Lag Plot: Strong Autocorrelation and Autoregressive Model

Lag Plot



Conclusions We can make the following conclusions based on the above plot.

1. The data come from an underlying autoregressive model with strong positive autocorrelation
2. The data contain no outliers.

Discussion

Note the tight clustering of points along the diagonal. This is the lag plot signature of a process with strong positive autocorrelation. Such processes are highly non-random--there is strong association between an observation and a succeeding observation. In short, if you know Y_{i-1} you can make a strong guess as to what Y_i will be.

If the above process were completely random, the plot would have a shotgun pattern, and knowledge of a current observation (say $Y_{i-1} = 3$) would yield virtually no knowledge about the next observation Y_i (it could here be anywhere from -2 to +8). On the other hand, if the process had strong autocorrelation, as seen above, and if $Y_{i-1} = 3$, then the range of possible values for Y_i is seen to be restricted to a smaller range (2 to 4)--still wide, but an improvement nonetheless (relative to -2 to +8) in predictive power.

Recommended Next Step

When the lag plot shows a strongly autoregressive pattern and only successive observations appear to be correlated, the next steps are to:

1. Extimate the parameters for the autoregressive model:

$$Y_i = A_0 + A_1 * Y_{i-1} + E_i$$

   Since Y_i and Y_{i-1} are precisely the axes of the lag plot, such estimation is a linear regression straight from the lag plot.

   The residual standard deviation for this autoregressive model will be much smaller than the residual standard deviation for the default model

$$Y_i = A_0 + E_i$$

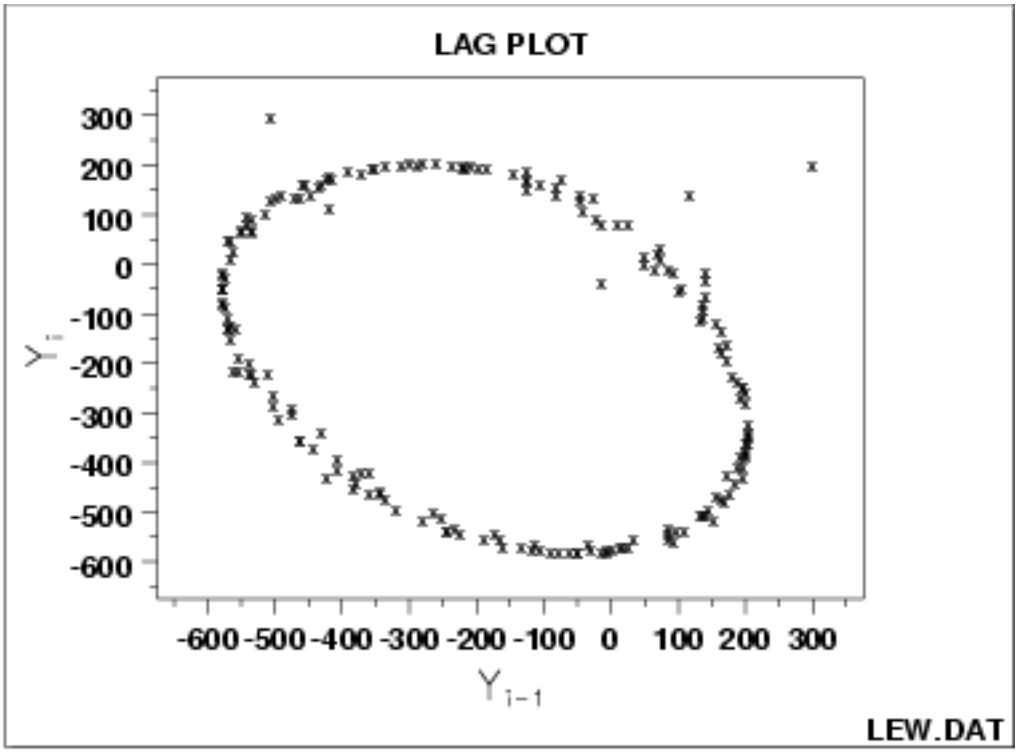
2. Reexamine the system to arrive at an explanation for the strong autocorrelation. Is it due to the
   1. phenomenon under study; or
   2. drifting in the environment; or
   3. contamination from the data acquisition system?

   Sometimes the source of the problem is contamination and carry-over from the data acquisition system where the system does not have time to electronically recover before collecting the next data point. If this is the case, then consider slowing down the sampling rate to achieve randomness.

1. Exploratory Data Analysis
1.3. EDA Techniques
1.3.3. Graphical Techniques: Alphabetic
1.3.3.15. Lag Plot

1.3.3.15.4. Lag Plot: Sinusoidal Models and Outliers

Lag Plot



Conclusions

We can make the following conclusions based on the above plot.

1. The data come from an underlying single-cycle sinusoidal model.
2. The data contain three outliers.

Discussion

In the plot above for lag = 1, note the tight elliptical clustering of points. Processes with a single-cycle sinusoidal model will have such elliptical lag plots.

Consequences of Ignoring Cyclical Pattern

If one were to naively assume that the above process came from the null model

$$Y_i = A_0 + E_i$$

and then estimate the constant by the sample mean, then the analysis would suffer because

1. the sample mean would be biased and meaningless;
2. the confidence limits would be meaningless and optimistically small.

The proper model

$$Y_i = C + \alpha \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

(where α is the amplitude, ω is the frequency--between 0 and .5 cycles per observation--, and ϕ is the phase) can be fit by standard non-linear least squares, to estimate the coefficients and their uncertainties.

The lag plot is also of value in outlier detection. Note in the above plot that there appears to be 4 points lying off the ellipse. However, in a lag plot, each point in the original data set Y shows up twice in the lag plot--once as Y_i and once as Y_{i-1}. Hence the outlier in the upper left at Y_i = 300 is the same raw data value that appears on the far right at Y_{i-1} = 300. Thus (-500,300) and (300,200) are due to the same outlier, namely the 158th data point: 300. The correct value for this 158th point should be approximately -300 and so it appears that a sign got dropped in the data collection. The other two points lying off the ellipse, at roughly (100,100) and at (0,-50), are caused by two faulty data values: the third data point of -15 should be about +125 and the fourth data point of +141 should be about -50, respectively. Hence the 4 apparent lag plot outliers are traceable to 3 actual outliers in the original run sequence: at points 4 (-15), 5 (141) and 158 (300). In retrospect, only one of these (point 158 (= 300)) is an obvious outlier in the run sequence plot.

Unexpected Value of EDA

Frequently a technique (e.g., the lag plot) is constructed to check one aspect (e.g., randomness) which it does well. Along the way, the technique also highlights some other anomaly of the data (namely, that there are 3 outliers). Such outlier identification and removal is extremely important for detecting irregularities in the data collection system, and also for arriving at a "purified" data set for modeling. The lag plot plays an important role in such outlier identification.

Recommended Next Step

When the lag plot indicates a sinusoidal model with possible outliers, the recommended next steps are:

1. Do a spectral plot to obtain an initial estimate of the frequency of the underlying cycle. This will be helpful as a starting value for the subsequent non-linear fitting.
2. Omit the outliers.
3. Carry out a non-linear fit of the model to the 197 points.

$$Y_i = C + \alpha \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

1.3.3.16. Linear Correlation Plot

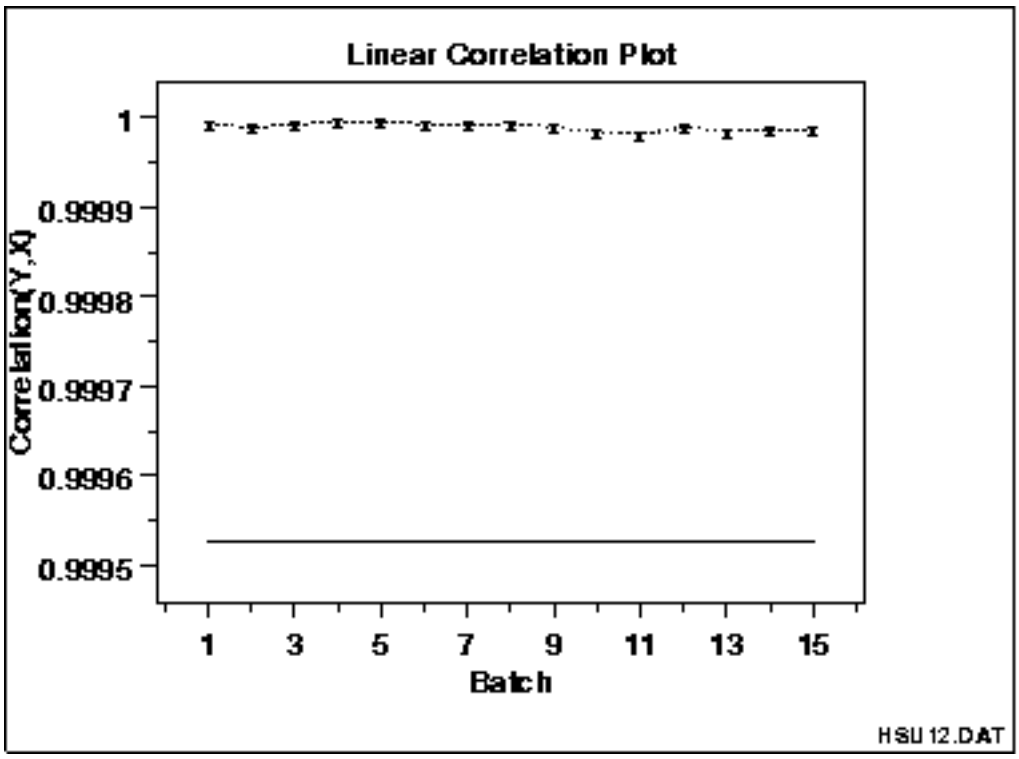
Purpose: Detect changes in correlation between groups

Linear correlation plots are used to assess whether or not correlations are consistent across groups. That is, if your data is in groups, you may want to know if a single correlation can be used across all the groups or whether separate correlations are required for each group.

Linear correlation plots are often used in conjunction with linear slope, linear intercept, and linear residual standard deviation plots. A linear correlation plot could be generated intially to see if linear fitting would be a fruitful direction. If the correlations are high, this implies it is worthwhile to continue with the linear slope, intercept, and residual standard deviation plots. If the correlations are weak, a different model needs to be pursued.

In some cases, you might not have groups. Instead you may have different data sets and you want to know if the same correlation can be adequately applied to each of the data sets. In this case, simply think of each distinct data set as a group and apply the linear slope plot as for groups.

Sample Plot



This linear correlation plot shows that the correlations are high for all groups. This implies that linear fits could provide a good model for each of these groups.

Definition: Group Correlations Versus Group ID

Linear correlation plots are formed by:

- Vertical axis: Group correlations
- Horizontal axis: Group identifier

A reference line is plotted at the correlation between the full data sets.

Questions

The linear correlation plot can be used to answer the following questions.

1. Are there linear relationships across groups?
2. Are the strength of the linear relationships relatively constant across the groups?

Importance: Checking Group Homogeneity

For grouped data, it may be important to know whether the different groups are homogeneous (i.e., similar) or heterogeneous (i.e., different). Linear correlation plots help answer this question in the context of linear fitting.

Related Techniques

Linear Intercept Plot
Linear Slope Plot
Linear Residual Standard Deviation Plot
Linear Fitting

Case Study The linear correlation plot is demonstrated in the Alaska pipeline data case study.

Software Most general purpose statistical software programs do not support a linear correlation plot. However, if the statistical program can generate correlations over a group, it should be feasible to write a macro to generate this plot. Dataplot supports a linear correlation plot.

1.3.3.17. Linear Intercept Plot

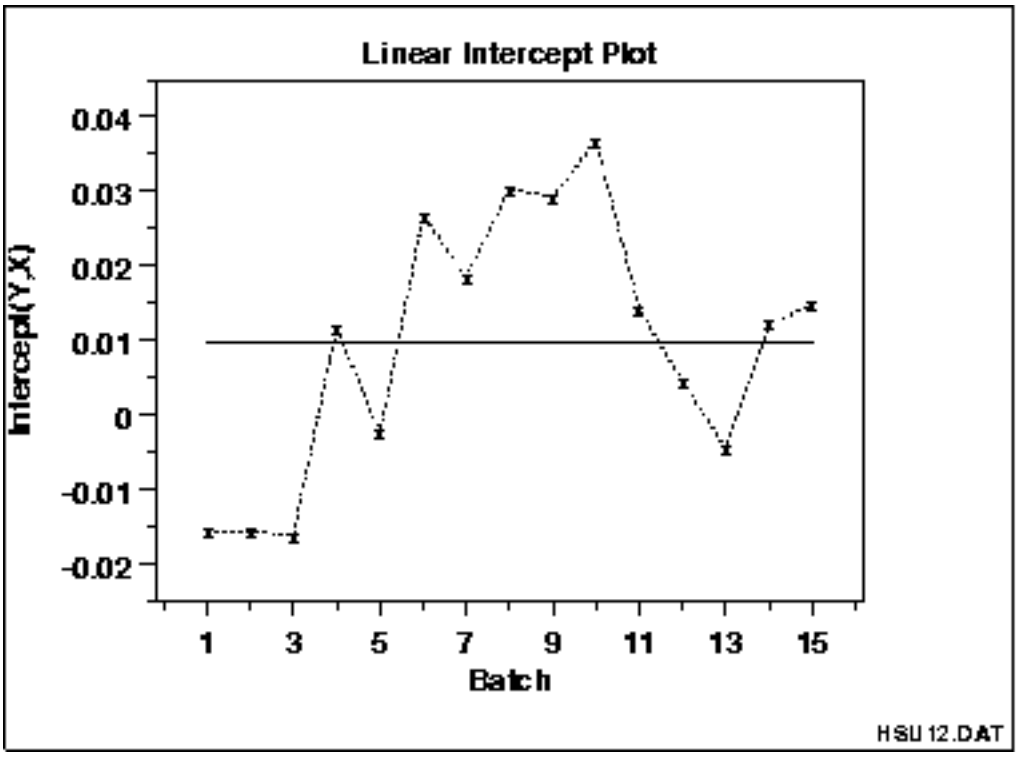
Purpose: Detect changes in linear intercepts between groups

Linear intercept plots are used to graphically assess whether or not linear fits are consistent across groups. That is, if your data have groups, you may want to know if a single fit can be used across all the groups or whether separate fits are required for each group.

Linear intercept plots are typically used in conjunction with linear slope and linear residual standard deviation plots.

In some cases you might not have groups. Instead, you have different data sets and you want to know if the same fit can be adequately applied to each of the data sets. In this case, simply think of each distinct data set as a group and apply the linear intercept plot as for groups.

Sample Plot



This linear intercept plot shows that there is a shift in intercepts. Specifically, the first three intercepts are lower than the intercepts for

the other groups. Note that these are small differences in the intercepts.

Definition: Group Intercepts Versus Group ID

Linear intercept plots are formed by:

- Vertical axis: Group intercepts from linear fits
- Horizontal axis: Group identifier

A reference line is plotted at the intercept from a linear fit using all the data.

Questions

The linear intercept plot can be used to answer the following questions.

1. Is the intercept from linear fits relatively constant across groups?
2. If the intercepts vary across groups, is there a discernible pattern?

Importance: Checking Group Homogeneity

For grouped data, it may be important to know whether the different groups are homogeneous (i.e., similar) or heterogeneous (i.e., different). Linear intercept plots help answer this question in the context of linear fitting.

Related Techniques

Linear Correlation Plot
Linear Slope Plot
Linear Residual Standard Deviation Plot
Linear Fitting

Case Study

The linear intercept plot is demonstrated in the Alaska pipeline data case study.

Software

Most general purpose statistical software programs do not support a linear intercept plot. However, if the statistical program can generate linear fits over a group, it should be feasible to write a macro to generate this plot. Dataplot supports a linear intercept plot.

1.3.3.18. Linear Slope Plot

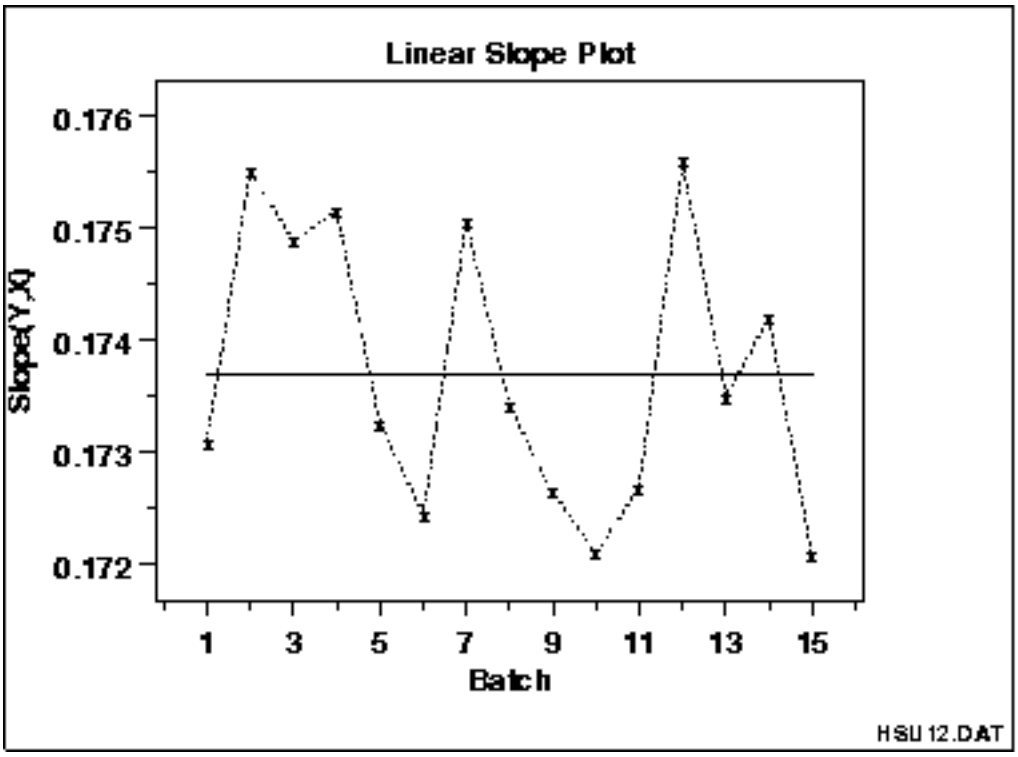
Purpose: Detect changes in linear slopes between groups

Linear slope plots are used to graphically assess whether or not linear fits are consistent across groups. That is, if your data have groups, you may want to know if a single fit can be used across all the groups or whether separate fits are required for each group.

Linear slope plots are typically used in conjunction with linear intercept and linear residual standard deviation plots.

In some cases you might not have groups. Instead, you have different data sets and you want to know if the same fit can be adequately applied to each of the data sets. In this case, simply think of each distinct data set as a group and apply the linear slope plot as for groups.

Sample Plot



This linear slope plot shows that the slopes are about 0.174 (plus or minus 0.002) for all groups. There does not appear to be a pattern in the variation of the slopes. This implies that a single fit may be adequate.

Definition: Group Slopes Versus Group ID

Linear slope plots are formed by:

- Vertical axis: Group slopes from linear fits
- Horizontal axis: Group identifier

A reference line is plotted at the slope from a linear fit using all the data.

Questions

The linear slope plot can be used to answer the following questions.

1. Do you get the same slope across groups for linear fits?
2. If the slopes differ, is there a discernible pattern in the slopes?

Importance: Checking Group Homogeneity

For grouped data, it may be important to know whether the different groups are homogeneous (i.e., similar) or heterogeneous (i.e., different). Linear slope plots help answer this question in the context of linear fitting.

Related Techniques

Linear Intercept Plot
Linear Correlation Plot
Linear Residual Standard Deviation Plot
Linear Fitting

Case Study

The linear slope plot is demonstrated in the Alaska pipeline data case study.

Software

Most general purpose statistical software programs do not support a linear slope plot. However, if the statistical program can generate linear fits over a group, it should be feasible to write a macro to generate this plot. Dataplot supports a linear slope plot.

1.3.3.19. Linear Residual Standard Deviation Plot

Purpose: Detect Changes in Linear Residual Standard Deviation Between Groups

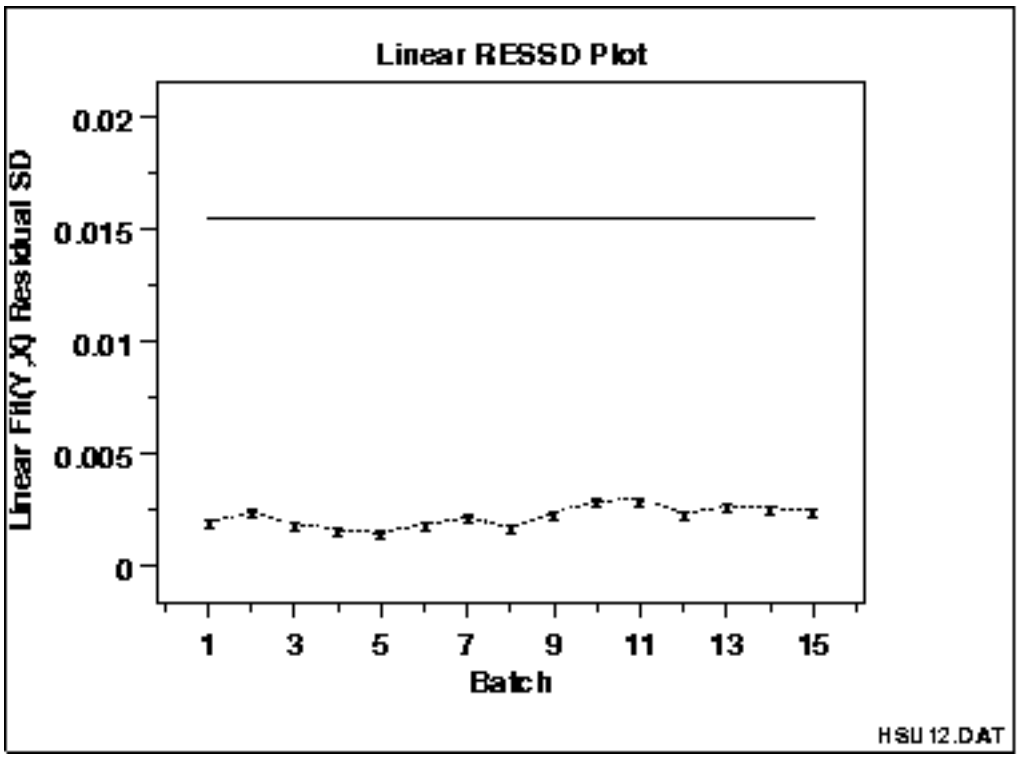
Linear residual standard deviation (RESSD) plots are used to graphically assess whether or not linear fits are consistent across groups. That is, if your data have groups, you may want to know if a single fit can be used across all the groups or whether separate fits are required for each group.

The residual standard deviation is a goodness-of-fit measure. That is, the smaller the residual standard deviation, the closer is the fit to the data.

Linear RESSD plots are typically used in conjunction with linear intercept and linear slope plots. The linear intercept and slope plots convey whether or not the fits are consistent across groups while the linear RESSD plot conveys whether the adequacy of the fit is consistent across groups.

In some cases you might not have groups. Instead, you have different data sets and you want to know if the same fit can be adequately applied to each of the data sets. In this case, simply think of each distinct data set as a group and apply the linear RESSD plot as for groups.

Sample Plot



This linear RESSD plot shows that the residual standard deviations from a linear fit are about 0.0025 for all the groups.

Definition: Group Residual Standard Deviation Versus Group ID

Linear RESSD plots are formed by:

- Vertical axis: Group residual standard deviations from linear fits
- Horizontal axis: Group identifier

A reference line is plotted at the residual standard deviation from a linear fit using all the data. This reference line will typically be much greater than any of the individual residual standard deviations.

Questions

The linear RESSD plot can be used to answer the following questions.

1. Is the residual standard deviation from a linear fit constant across groups?
2. If the residual standard deviations vary, is there a discernible pattern across the groups?

Importance: Checking Group Homogeneity

For grouped data, it may be important to know whether the different groups are homogeneous (i.e., similar) or heterogeneous (i.e., different). Linear RESSD plots help answer this question in the context of linear fitting.

Related Techniques	Linear Intercept Plot Linear Slope Plot Linear Correlation Plot Linear Fitting
Case Study	The linear residual standard deviation plot is demonstrated in the Alaska pipeline data case study.
Software	Most general purpose statistical software programs do not support a linear residual standard deviation plot. However, if the statistical program can generate linear fits over a group, it should be feasible to write a macro to generate this plot. Dataplot supports a linear residual standard deviation plot.

1.3.3.20. Mean Plot

Purpose: Detect changes in location between groups

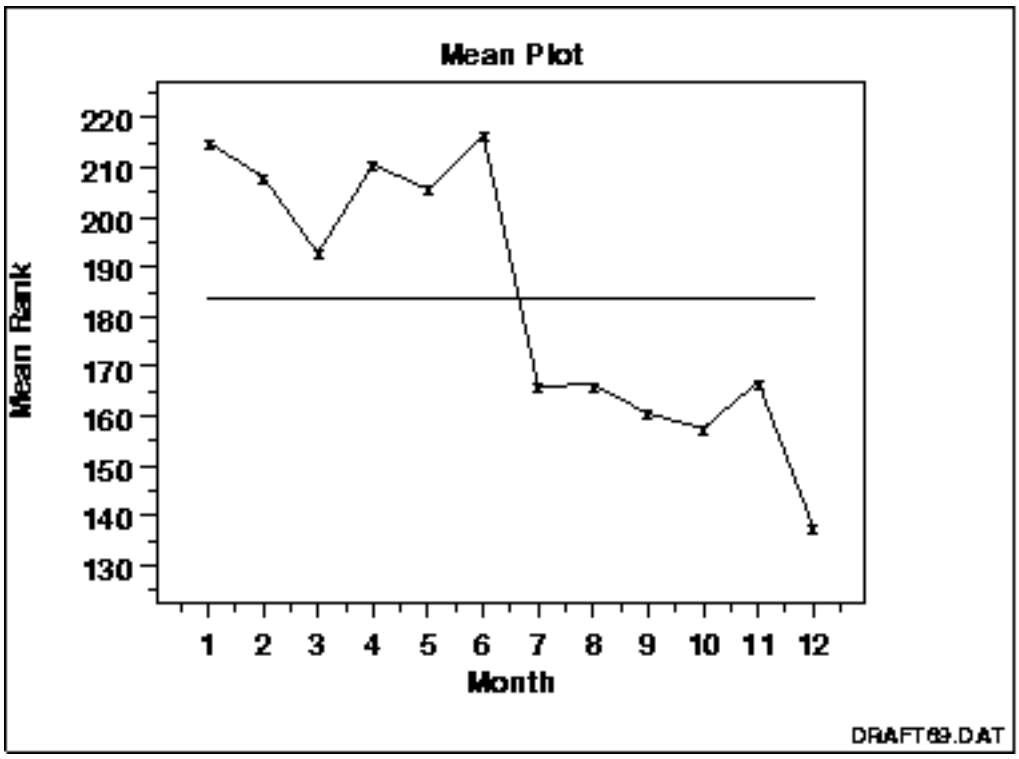
Mean plots are used to see if the mean varies between different groups of the data. The grouping is determined by the analyst. In most cases, the data set contains a specific grouping variable. For example, the groups may be the levels of a factor variable. In the sample plot below, the months of the year provide the grouping.

Mean plots can be used with ungrouped data to determine if the mean is changing over time. In this case, the data are split into an arbitrary number of equal-sized groups. For example, a data series with 400 points can be divided into 10 groups of 40 points each. A mean plot can then be generated with these groups to see if the mean is increasing or decreasing over time.

Although the mean is the most commonly used measure of location, the same concept applies to other measures of location. For example, instead of plotting the mean of each group, the median or the trimmed mean might be plotted instead. This might be done if there were significant outliers in the data and a more robust measure of location than the mean was desired.

Mean plots are typically used in conjunction with standard deviation plots. The mean plot checks for shifts in location while the standard deviation plot checks for shifts in scale.

Sample Plot



This sample mean plot shows a shift of location after the 6th month.

Definition: Group Means Versus Group ID

Mean plots are formed by:

- Vertical axis: Group mean
- Horizontal axis: Group identifier

A reference line is plotted at the overall mean.

Questions

The mean plot can be used to answer the following questions.

1. Are there any shifts in location?
2. What is the magnitude of the shifts in location?
3. Is there a distinct pattern in the shifts in location?

Importance: Checking Assumptions

A common assumption in 1-factor analyses is that of constant location. That is, the location is the same for different levels of the factor variable. The mean plot provides a graphical check for that assumption. A common assumption for univariate data is that the location is constant. By grouping the data into equal intervals, the mean plot can provide a graphical test of this assumption.

Related Techniques

Standard Deviation Plot
Dex Mean Plot
Box Plot

Software

Most general purpose statistical software programs do not support a mean plot. However, if the statistical program can generate the mean over a group, it should be feasible to write a macro to generate this plot. Dataplot supports a mean plot.

1.3.3.21. Normal Probability Plot

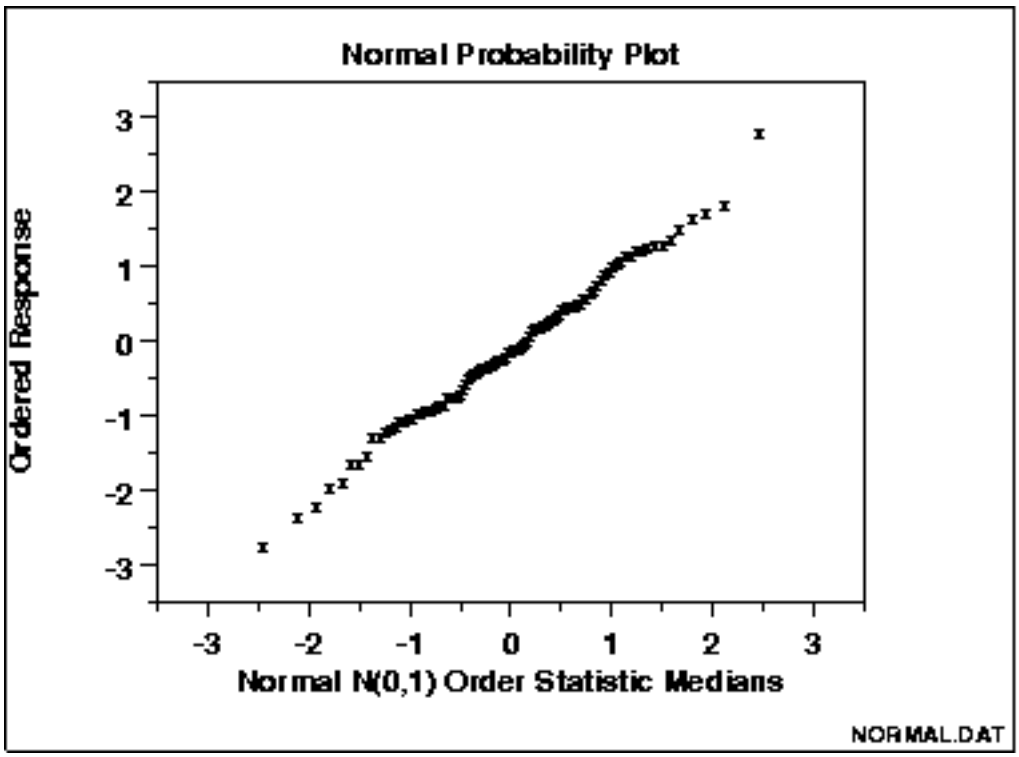
Purpose: Check If Data Are Approximately Normally Distributed

The normal probability plot (Chambers 1983) is a graphical technique for assessing whether or not a data set is approximately normally distributed.

The data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line. Departures from this straight line indicate departures from normality.

The normal probability plot is a special case of the probability plot. We cover the normal probability plot separately due to its importance in many applications.

Sample Plot



The points on this plot form a nearly linear pattern, which indicates that the normal distribution is a good model for this data set.

Definition: Ordered Response Values Versus Normal Order Statistic Medians

The normal probability plot is formed by:

- Vertical axis: Ordered response values
- Horizontal axis: Normal order statistic medians

The observations are plotted as a function of the corresponding normal order statistic medians which are defined as:

$$N(i) = G(U(i))$$

where U(i) are the uniform order statistic medians (defined below) and G is the percent point function of the normal distribution. The percent point function is the inverse of the cumulative distribution function (probability that x is less than or equal to some value). That is, given a probability, we want the corresponding x of the cumulative distribution function.

The uniform order statistic medians are defined as:

$$m(i) = 1 - m(n) \text{ for } i = 1$$
$$m(i) = (i - 0.3175)/(n + 0.365) \text{ for } i = 2, 3, ..., n-1$$
$$m(i) = 0.5^{(1/n)} \text{ for } i = n$$

In addition, a straight line can be fit to the points and added as a reference line. The further the points vary from this line, the greater the indication of departures from normality.

Probability plots for distributions other than the normal are computed in exactly the same way. The normal percent point function (the G) is simply replaced by the percent point function of the desired distribution. That is, a probability plot can easily be generated for any distribution for which you have the percent point function.

One advantage of this method of computing probability plots is that the intercept and slope estimates of the fitted line are in fact estimates for the location and scale parameters of the distribution. Although this is not too important for the normal distribution since the location and scale are estimated by the mean and standard deviation, respectively, it can be useful for many other distributions.

The correlation coefficient of the points on the normal probability plot can be compared to a table of critical values to provide a formal test of the hypothesis that the data come from a normal distribution.

Questions

The normal probability plot is used to answer the following questions.

1. Are the data normally distributed?
2. What is the nature of the departure from normality (data skewed, shorter than expected tails, longer than expected tails)?

Importance: Check Normality Assumption

The underlying assumptions for a measurement process are that the data should behave like:

1. random drawings;
2. from a fixed distribution;
3. with fixed location;
4. with fixed scale.

Probability plots are used to assess the assumption of a fixed distribution. In particular, most statistical models are of the form:

response = deterministic + random

where the deterministic part is the fit and the random part is error. This error component in most common statistical models is specifically assumed to be normally distributed with fixed location and scale. This is the most frequent application of normal probability plots. That is, a model is fit and a normal probability plot is generated for the residuals from the fitted model. If the residuals from the fitted model are not normally distributed, then one of the major assumptions of the model has been violated.

Examples

1. Data are normally distributed
2. Data have fat tails
3. Data have short tails
4. Data are skewed right

Related Techniques

Histogram
Probability plots for other distributions (e.g., Weibull)
Probability plot correlation coefficient plot (PPCC plot)
Anderson-Darling Goodness-of-Fit Test
Chi-Square Goodness-of-Fit Test
Kolmogorov-Smirnov Goodness-of-Fit Test

Case Study

The normal probability plot is demonstrated in the heat flow meter data case study.

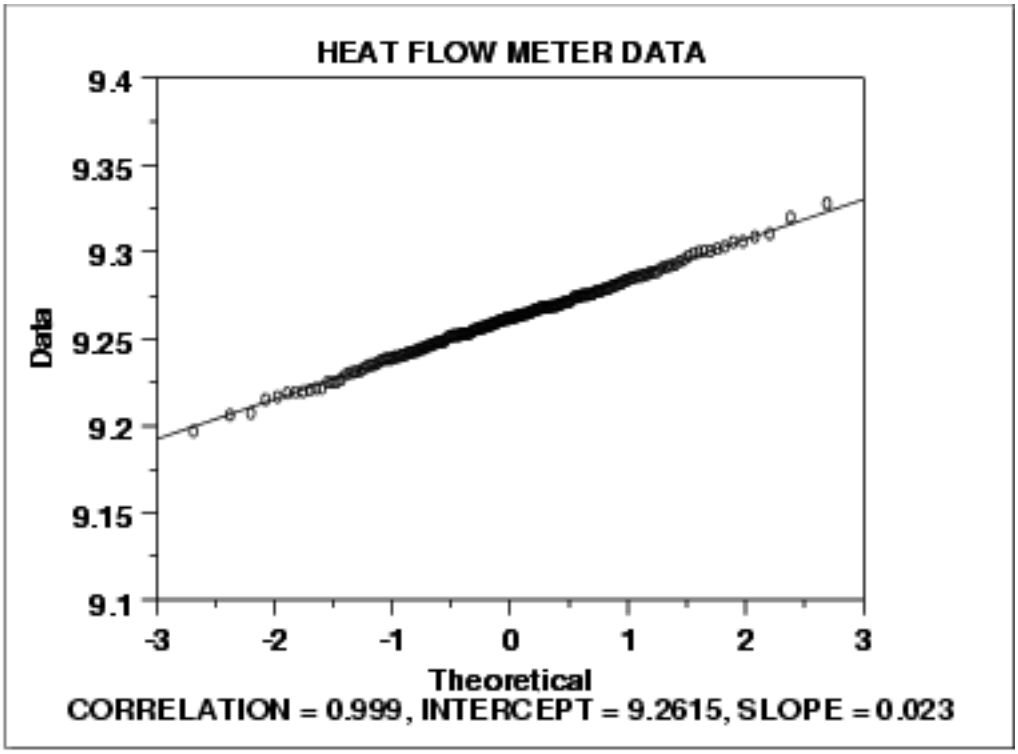
Software

Most general purpose statistical software programs can generate a normal probability plot. Dataplot supports a normal probability plot.

1.3.3.21.1. Normal Probability Plot: Normally Distributed Data

Normal Probability Plot

The following normal probability plot is from the heat flow meter data.



Conclusions

We can make the following conclusions from the above plot.

1. The normal probability plot shows a strongly linear pattern. There are only minor deviations from the line fit to the points on the probability plot.
2. The normal distribution appears to be a good model for these data.

Discussion

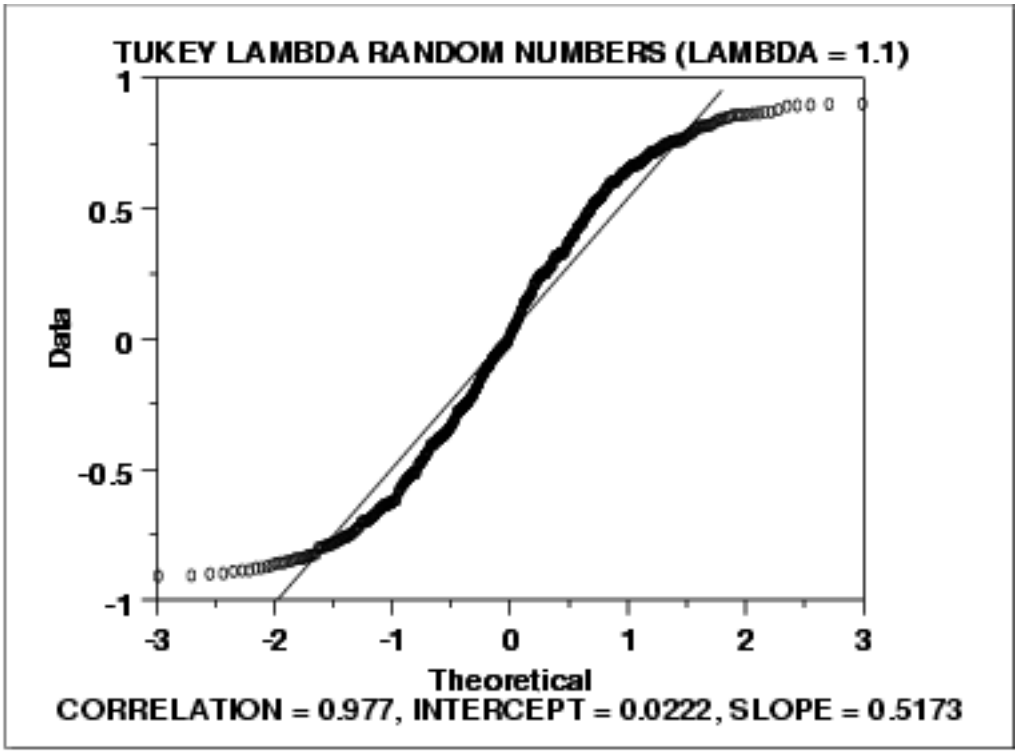
Visually, the probability plot shows a strongly linear pattern. This is verified by the correlation coefficient of 0.9989 of the line fit to the probability plot. The fact that the points in the lower and upper extremes of the plot do not deviate significantly from the straight-line pattern indicates that there are not any significant outliers (relative to a normal distribution).

In this case, we can quite reasonably conclude that the normal distribution provides an excellent model for the data. The intercept and slope of the fitted line give estimates of 9.26 and 0.023 for the location and scale parameters of the fitted normal distribution.

1.3.3.21.2. Normal Probability Plot: Data Have Short Tails

Normal Probability Plot for Data with Short Tails

The following is a normal probability plot for 500 random numbers generated from a Tukey-Lambda distribution with the λ parameter equal to 1.1.



Conclusions

We can make the following conclusions from the above plot.

1. The normal probability plot shows a non-linear pattern.
2. The normal distribution is not a good model for these data.

Discussion

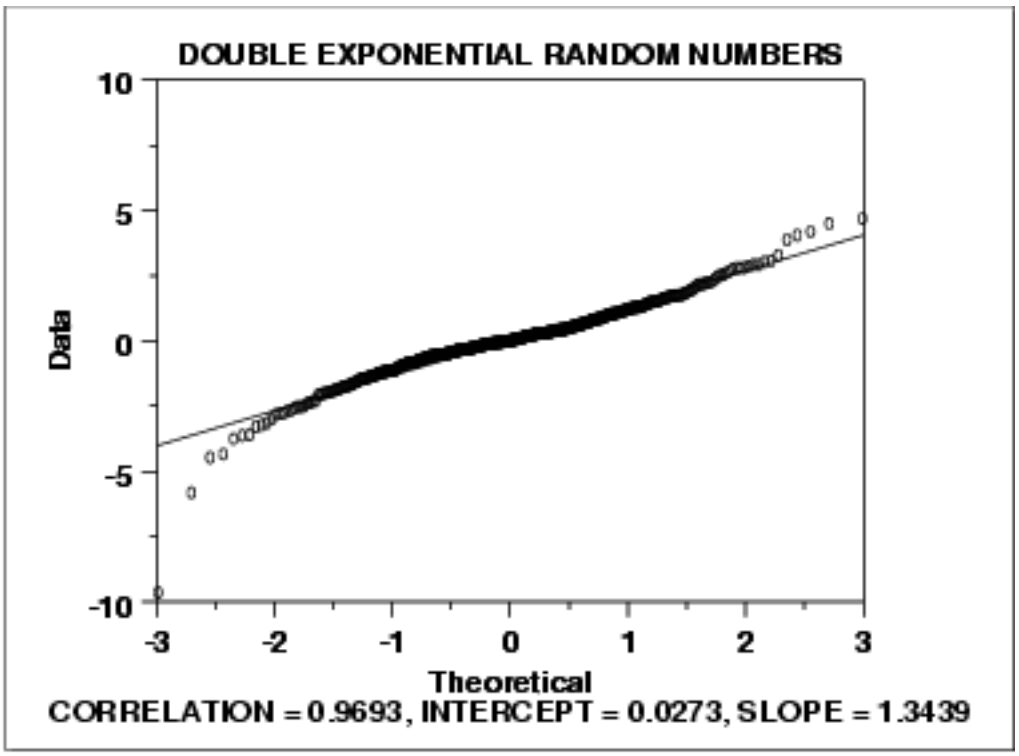
For data with short tails relative to the normal distribution, the non-linearity of the normal probability plot shows up in two ways. First, the middle of the data shows an S-like pattern. This is common for both short and long tails. Second, the first few and the last few points show a marked departure from the reference fitted line. In comparing this plot to the long tail example in the next section, the important difference is the direction of the departure from the fitted line for the first few and last few points. For short tails, the first few points show increasing departure from the fitted line *above* the line and last few points show increasing departure from the fitted line *below* the line. For long tails, this pattern is reversed.

In this case, we can reasonably conclude that the normal distribution does not provide an adequate fit for this data set. For probability plots that indicate short-tailed distributions, the next step might be to generate a Tukey Lambda PPCC plot. The Tukey Lambda PPCC plot can often be helpful in identifying an appropriate distributional family.

1.3.3.21.3. Normal Probability Plot: Data Have Long Tails

Normal Probability Plot for Data with Long Tails

The following is a normal probability plot of 500 numbers generated from a double exponential distribution. The double exponential distribution is symmetric, but relative to the normal it declines rapidly and has longer tails.



Conclusions

We can make the following conclusions from the above plot.

1. The normal probability plot shows a reasonably linear pattern in the center of the data. However, the tails, particularly the lower tail, show departures from the fitted line.
2. A distribution other than the normal distribution would be a good model for these data.

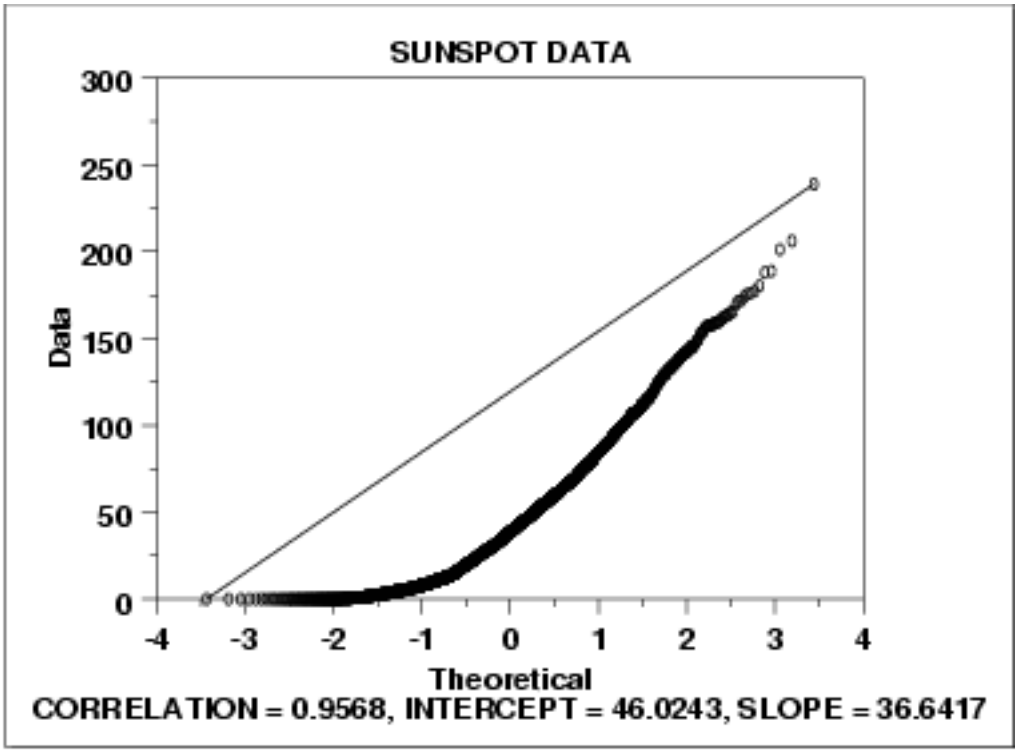
Discussion

For data with long tails relative to the normal distribution, the non-linearity of the normal probability plot can show up in two ways. First, the middle of the data may show an S-like pattern. This is common for both short and long tails. In this particular case, the S pattern in the middle is fairly mild. Second, the first few and the last few points show marked departure from the reference fitted line. In the plot above, this is most noticeable for the first few data points. In comparing this plot to the short-tail example in the previous section, the important difference is the direction of the departure from the fitted line for the first few and the last few points. For long tails, the first few points show increasing departure from the fitted line *below* the line and last few points show increasing departure from the fitted line *above* the line. For short tails, this pattern is reversed.

In this case we can reasonably conclude that the normal distribution can be improved upon as a model for these data. For probability plots that indicate long-tailed distributions, the next step might be to generate a Tukey Lambda PPCC plot. The Tukey Lambda PPCC plot can often be helpful in identifying an appropriate distributional family.

1.3.3.21.4. Normal Probability Plot: Data are Skewed Right

Normal Probability Plot for Data that are Skewed Right



Conclusions We can make the following conclusions from the above plot.

1. The normal probability plot shows a strongly non-linear pattern. Specifically, it shows a quadratic pattern in which all the points are below a reference line drawn between the first and last points.
2. The normal distribution is not a good model for these data.

Discussion

This quadratic pattern in the normal probability plot is the signature of a significantly right-skewed data set. Similarly, if all the points on the normal probability plot fell above the reference line connecting the first and last points, that would be the signature pattern for a significantly left-skewed data set.

In this case we can quite reasonably conclude that we need to model these data with a right skewed distribution such as the Weibull or lognormal.

1.3.3.22. Probability Plot

Purpose: Check If Data Follow a Given Distribution

The probability plot (Chambers 1983) is a graphical technique for assessing whether or not a data set follows a given distribution such as the normal or Weibull.

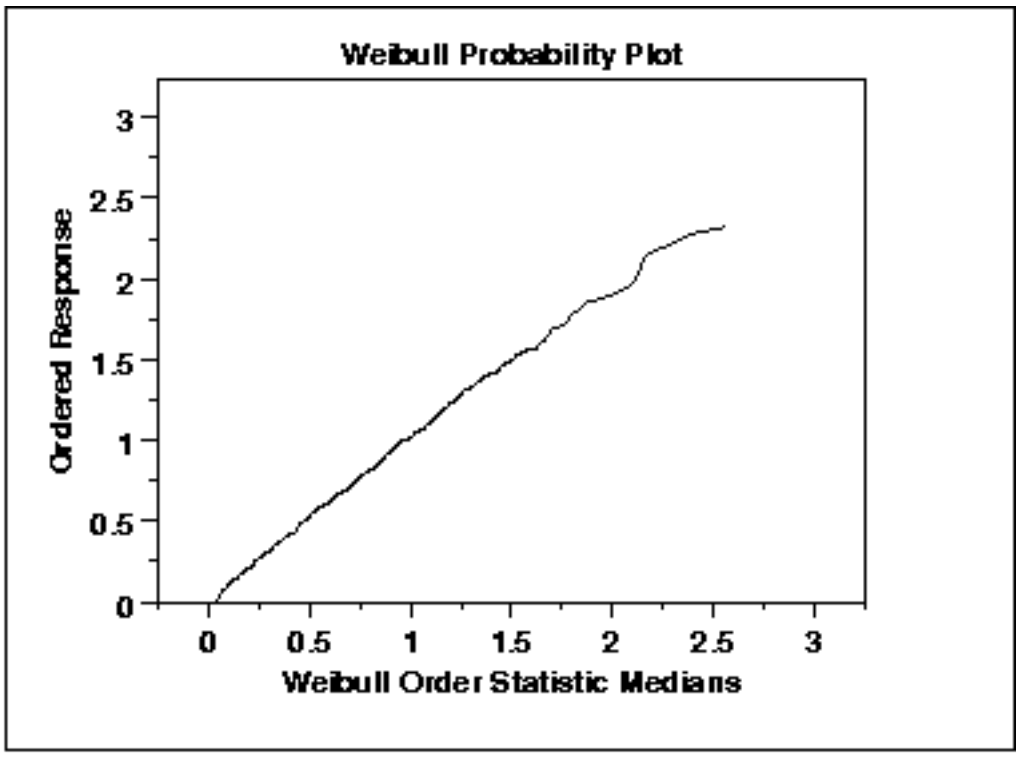
The data are plotted against a theoretical distribution in such a way that the points should form approximately a straight line. Departures from this straight line indicate departures from the specified distribution.

The correlation coefficient associated with the linear fit to the data in the probability plot is a measure of the goodness of the fit. Estimates of the location and scale parameters of the distribution are given by the intercept and slope. Probability plots can be generated for several competing distributions to see which provides the best fit, and the probability plot generating the highest correlation coefficient is the best choice since it generates the straightest probability plot.

For distributions with shape parameters (not counting location and scale parameters), the shape parameters must be known in order to generate the probability plot. For distributions with a single shape parameter, the probability plot correlation coefficient (PPCC) plot provides an excellent method for estimating the shape parameter.

We cover the special case of the normal probability plot separately due to its importance in many statistical applications.

Sample Plot



This data is a set of 500 Weibull random numbers with a shape parameter = 2, location parameter = 0, and scale parameter = 1. The Weibull probability plot indicates that the Weibull distribution does in fact fit these data well.

Definition: Ordered Response Values Versus Order Statistic Medians for the Given Distribution

The probability plot is formed by:

- Vertical axis: Ordered response values
- Horizontal axis: Order statistic medians for the given distribution

The order statistic medians are defined as:

N(i) = G(U(i))

where the U(i) are the uniform order statistic medians (defined below) and G is the percent point function for the desired distribution. The percent point function is the inverse of the cumulative distribution function (probability that *x* is less than or equal to some value). That is, given a probability, we want the corresponding *x* of the cumulative distribution function.

The uniform order statistic medians are defined as:

m(i) = 1 - m(n) for i = 1
m(i) = (i - 0.3175)/(n + 0.365) for i = 2, 3, ..., n-1
m(i) = 0.5**(1/n) for i = n

In addition, a straight line can be fit to the points and added as a reference line. The further the points vary from this line, the greater the

indication of a departure from the specified distribution.

This definition implies that a probability plot can be easily generated for any distribution for which the percent point function can be computed.

One advantage of this method of computing proability plots is that the intercept and slope estimates of the fitted line are in fact estimates for the location and scale parameters of the distribution. Although this is not too important for the normal distribution (the location and scale are estimated by the mean and standard deviation, respectively), it can be useful for many other distributions.

Questions

The probability plot is used to answer the following questions:

- Does a given distribution, such as the Weibull, provide a good fit to my data?
- What distribution best fits my data?
- What are good estimates for the location and scale parameters of the chosen distribution?

Importance: Check distributional assumption

The discussion for the normal probability plot covers the use of probability plots for checking the fixed distribution assumption.

Some statistical models assume data have come from a population with a specific type of distribution. For example, in reliability applications, the Weibull, lognormal, and exponential are commonly used distributional models. Probability plots can be useful for checking this distributional assumption.

Related Techniques

Histogram
Probability Plot Correlation Coefficient (PPCC) Plot
Hazard Plot
Quantile-Quantile Plot
Anderson-Darling Goodness of Fit
Chi-Square Goodness of Fit
Kolmogorov-Smirnov Goodness of Fit

Case Study

The probability plot is demonstrated in the airplane glass failure time data case study.

Software

Most general purpose statistical software programs support probability plots for at least a few common distributions. Dataplot supports probability plots for a large number of distributions.

1.3.3.23. Probability Plot Correlation Coefficient Plot

Purpose: Graphical Technique for Finding the Shape Parameter of a Distributional Family that Best Fits a Data Set

The probability plot correlation coefficient (PPCC) plot (Filliben 1975) is a graphical technique for identifying the shape parameter for a distributional family that best describes the data set. This technique is appropriate for families, such as the Weibull, that are defined by a single shape parameter and location and scale parameters, and it is not appropriate for distributions, such as the normal, that are defined only by location and scale parameters.

The PPCC plot is generated as follows. For a series of values for the shape parameter, the correlation coefficient is computed for the probability plot associated with a given value of the shape parameter. These correlation coefficients are plotted against their corresponding shape parameters. The maximum correlation coefficient corresponds to the optimal value of the shape parameter. For better precision, two iterations of the PPCC plot can be generated; the first is for finding the right neighborhood and the second is for fine tuning the estimate.

The PPCC plot is used first to find a good value of the shape parameter. The probability plot is then generated to find estimates of the location and scale parameters and in addition to provide a graphical assessment of the adequacy of the distributional fit.

Compare Distributions

In addition to finding a good choice for estimating the shape parameter of a given distribution, the PPCC plot can be useful in deciding which distributional family is most appropriate. For example, given a set of reliabilty data, you might generate PPCC plots for a Weibull, lognormal, gamma, and inverse Gaussian distributions, and possibly others, on a single page. This one page would show the best value for the shape parameter for several distributions and would additionally indicate which of these distributional families provides the best fit (as measured by the maximum probability plot correlation coefficient). That is, if the maximum PPCC value for the Weibull is 0.99 and only 0.94 for the lognormal, then we could reasonably conclude that the Weibull family is the better choice.

Tukey-Lambda PPCC Plot for Symmetric Distributions

The Tukey Lambda PPCC plot, with shape parameter λ, is particularly useful for symmetric distributions. It indicates whether a distribution is short or long tailed and it can further indicate several common distributions. Specifically,

1. λ = -1: distribution is approximately Cauchy
2. λ = 0: distribution is exactly logistic
3. λ = 0.14: distribution is approximately normal
4. λ = 0.5: distribution is U-shaped
5. λ = 1: distribution is exactly uniform

If the Tukey Lambda PPCC plot gives a maximum value near 0.14, we can reasonably conclude that the normal distribution is a good model for the data. If the maximum value is less than 0.14, a long-tailed distribution such as the double exponential or logistic would be a better choice. If the maximum value is near -1, this implies the selection of very long-tailed distribution, such as the Cauchy. If the maximum value is greater than 0.14, this implies a short-tailed distribution such as the Beta or uniform.

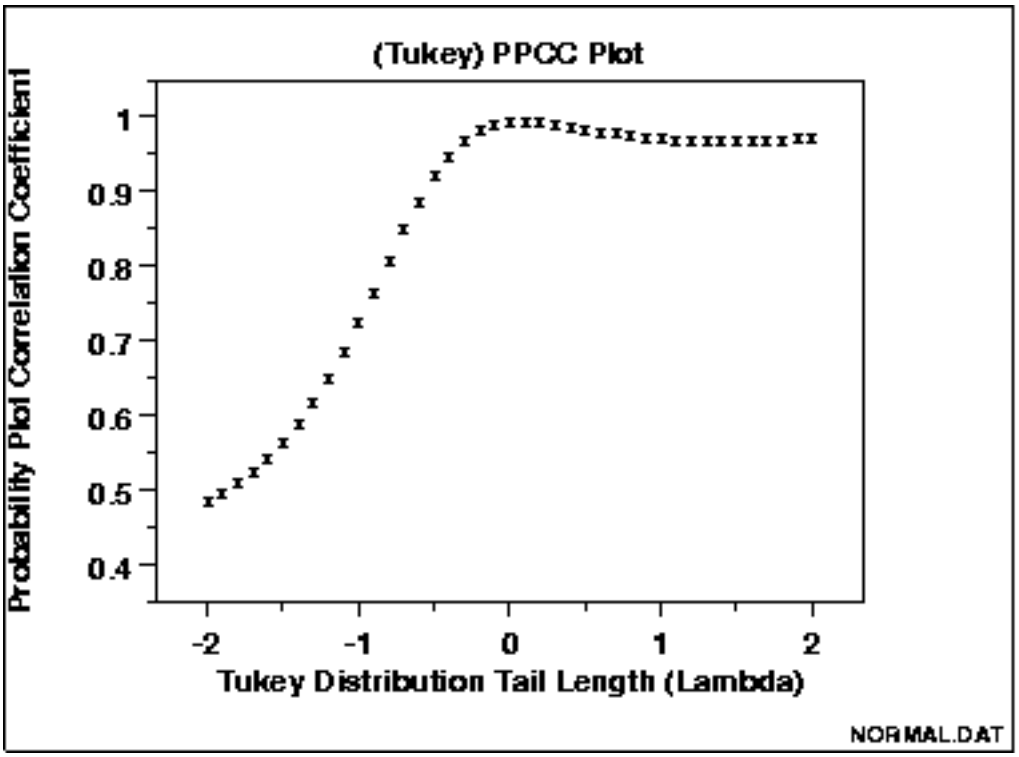
The Tukey-Lambda PPCC plot is used to suggest an appropriate distribution. You should follow-up with PPCC and probability plots of the appropriate alternatives.

Use Judgement When Selecting An Appropriate Distributional Family

When comparing distributional models, do not simply choose the one with the maximum PPCC value. In many cases, several distributional fits provide comparable PPCC values. For example, a lognormal and Weibull may both fit a given set of reliability data quite well. Typically, we would consider the complexity of the distribution. That is, a simpler distribution with a marginally smaller PPCC value may be preferred over a more complex distribution. Likewise, there may be theoretical justification in terms of the underlying scientific model for preferring a distribution with a marginally smaller PPCC value in some cases. In other cases, we may not need to know if the distributional model is optimal, only that it is adequate for our purposes. That is, we may be able to use techniques designed for normally distributed data even if other distributions fit the data somewhat better.

Sample Plot

The following is a PPCC plot of 100 normal random numbers. The maximum value of the correlation coefficient = 0.997 at λ = 0.099.



This PPCC plot shows that:

1. the best-fit symmetric distribution is nearly normal;
2. the data are not long tailed;
3. the sample mean would be an appropriate estimator of location.

We can follow-up this PPCC plot with a normal probability plot to verify the normality model for the data.

Definition: The PPCC plot is formed by:

- Vertical axis: Probability plot correlation coefficient;
- Horizontal axis: Value of shape parameter.

Questions The PPCC plot answers the following questions:

1. What is the best-fit member within a distributional family?
2. Does the best-fit member provide a good fit (in terms of generating a probability plot with a high correlation coefficient)?
3. Does this distributional family provide a good fit compared to other distributions?
4. How sensitive is the choice of the shape parameter?

Importance Many statistical analyses are based on distributional assumptions about the population from which the data have been obtained. However, distributional families can have radically different shapes depending on the value of the shape parameter. Therefore, finding a reasonable choice for the shape parameter is a necessary step in the analysis. In many analyses, finding a good distributional model for the data is the primary focus of the analysis. In both of these cases, the PPCC plot is a valuable tool.

Related Techniques
Probability Plot
Maximum Likelihood Estimation
Least Squares Estimation
Method of Moments Estimation

Case Study The PPCC plot is demonstrated in the airplane glass failure data case study.

Software PPCC plots are currently not available in most common general purpose statistical software programs. However, the underlying technique is based on probability plots and correlation coefficients, so it should be possible to write macros for PPCC plots in statistical programs that support these capabilities. Dataplot supports PPCC plots.

1.3.3.24. Quantile-Quantile Plot

Purpose: Check If Two Data Sets Can Be Fit With the Same Distribution

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

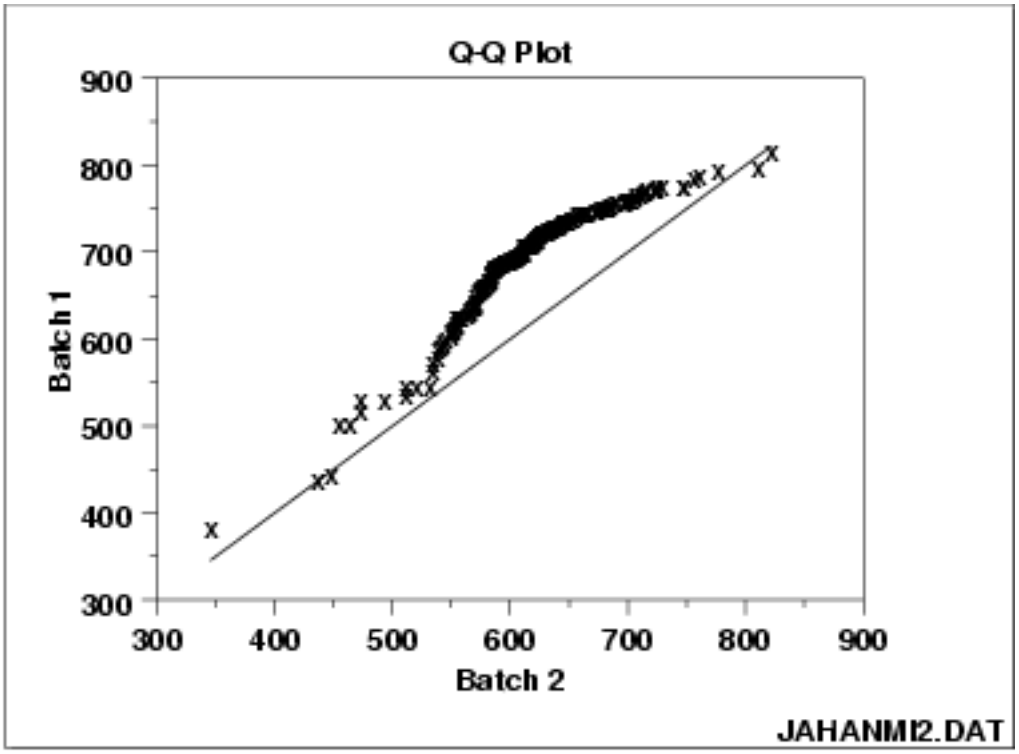
A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.

The q-q plot is similar to a probability plot. For a probability plot, the quantiles for one of the data samples are replaced with the quantiles of a theoretical distribution.

Sample Plot



This q-q plot shows that

1. These 2 batches do not appear to have come from populations with a common distribution.
2. The batch 1 values are significantly higher than the corresponding batch 2 values.
3. The differences are increasing from values 525 to 625. Then the values for the 2 batches get closer again.

Definition: Quantiles for Data Set 1 Versus Quantiles of Data Set 2

The q-q plot is formed by:

- Vertical axis: Estimated quantiles from data set 1
- Horizontal axis: Estimated quantiles from data set 2

Both axes are in units of their respective data sets. That is, the actual quantile level is not plotted. For a given point on the q-q plot, we know that the quantile level is the same for both points, but not what that quantile level actually is.

If the data sets have the same size, the q-q plot is essentially a plot of sorted data set 1 against sorted data set 2. If the data sets are not of equal size, the quantiles are usually picked to correspond to the sorted values from the smaller data set and then the quantiles for the larger data set are interpolated.

Questions

The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior?

Importance: Check for Common Distribution

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.

Related Techniques

Bihistogram
T Test
F Test
2-Sample Chi-Square Test
2-Sample Kolmogorov-Smirnov Test

Case Study

The quantile-quantile plot is demonstrated in the ceramic strength data case study.

Software

Q-Q plots are available in some general purpose statistical software programs, including Dataplot. If the number of data points in the two samples are equal, it should be relatively easy to write a macro in statistical programs that do not support the q-q plot. If the number of points are not equal, writing a macro for a q-q plot may be difficult.

1.3.3.25. Run-Sequence Plot

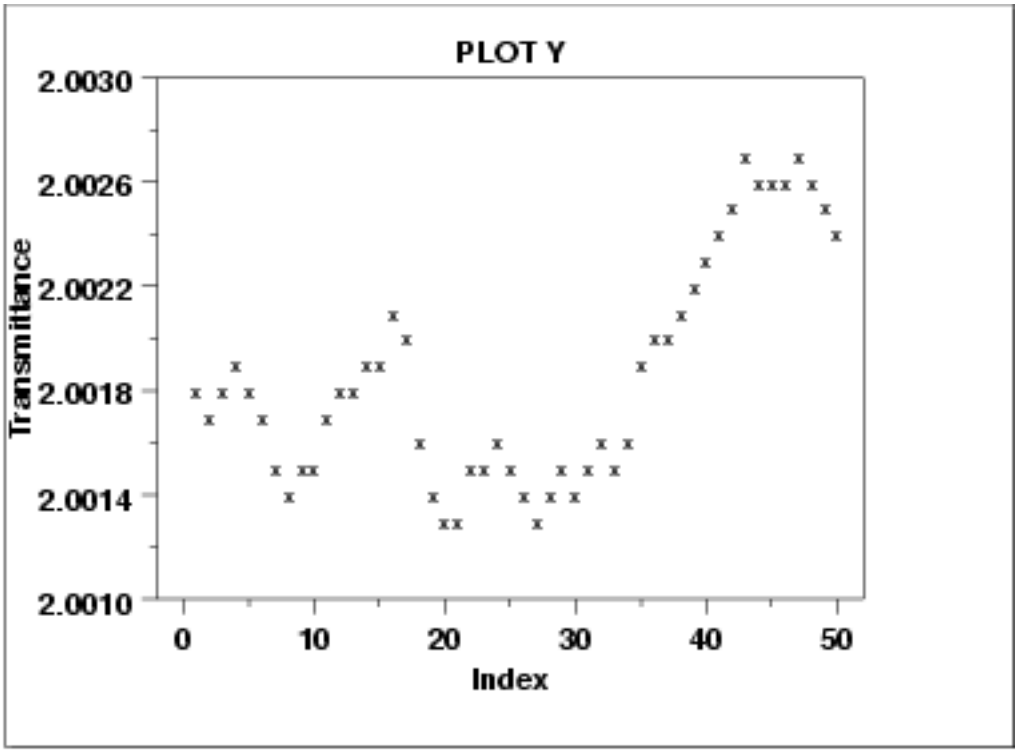
Purpose: Check for Shifts in Location and Scale and Outliers

Run sequence plots (Chambers 1983) are an easy way to graphically summarize a univariate data set. A common assumption of univariate data sets is that they behave like:

1. random drawings;
2. from a fixed distribution;
3. with a common location; and
4. with a common scale.

With run sequence plots, shifts in location and scale are typically quite evident. Also, outliers can easily be detected.

Sample Plot: Last Third of Data Shows a Shift of Location



This sample run sequence plot shows that the location shifts up for the last third of the data.

Definition: y(i) Versus i

Run sequence plots are formed by:

- Vertical axis: Response variable Y(i)
- Horizontal axis: Index i (i = 1, 2, 3, ...)

Questions

The run sequence plot can be used to answer the following questions

1. Are there any shifts in location?
2. Are there any shifts in variation?
3. Are there any outliers?

The run sequence plot can also give the analyst an excellent feel for the data.

Importance: Check Univariate Assumptions

For univariate data, the default model is

$$Y = \text{constant} + \text{error}$$

where the error is assumed to be random, from a fixed distribution, and with constant location and scale. The validity of this model depends on the validity of these assumptions. The run sequence plot is useful for checking for constant location and scale.

Even for more complex models, the assumptions on the error term are still often the same. That is, a run sequence plot of the residuals (even from very complex models) is still vital for checking for outliers and for detecting shifts in location and scale.

Related Techniques

Scatter Plot
Histogram
Autocorrelation Plot
Lag Plot

Case Study

The run sequence plot is demonstrated in the Filter transmittance data case study.

Software

Run sequence plots are available in most general purpose statistical software programs, including Dataplot.

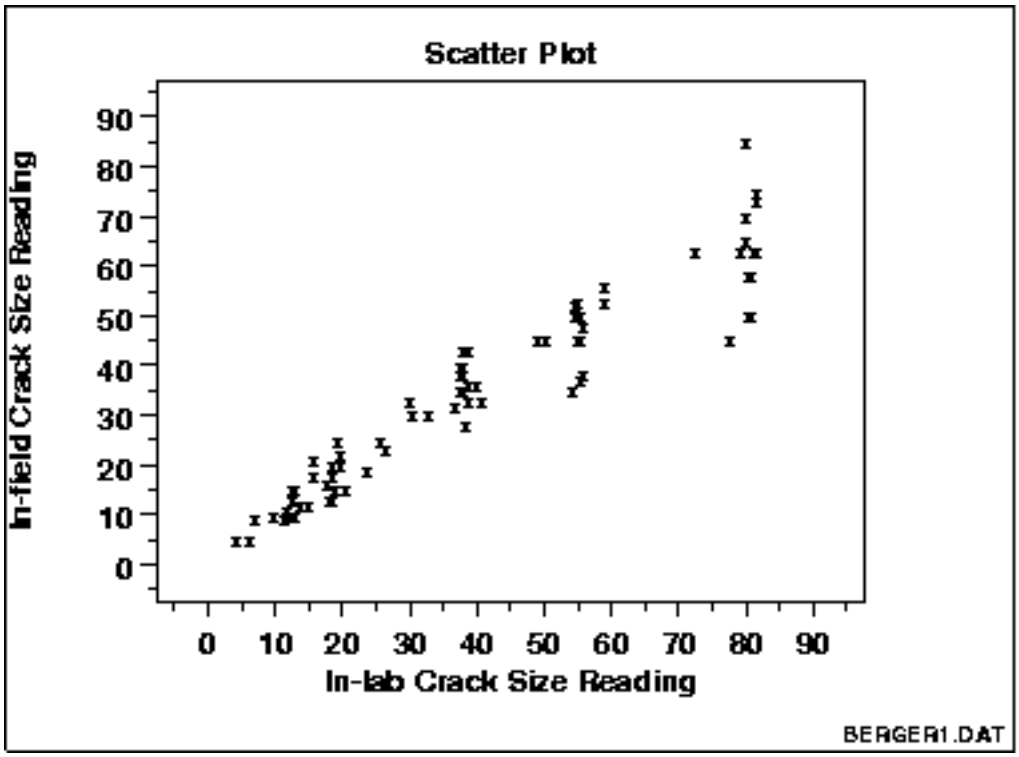
1. Exploratory Data Analysis
1.3. EDA Techniques
1.3.3. Graphical Techniques: Alphabetic

1.3.3.26. Scatter Plot

Purpose: Check for Relationship

A scatter plot (Chambers 1983) reveals relationships or association between two variables. Such relationships manifest themselves by any non-random structure in the plot. Various common types of patterns are demonstrated in the examples.

Sample Plot: Linear Relationship Between Variables Y and X



This sample plot reveals a linear relationship between the two variables indicating that a linear regression model might be appropriate.

Definition: Y Versus X

A scatter plot is a plot of the values of Y versus the corresponding values of X:

- Vertical axis: variable Y--usually the response variable
- Horizontal axis: variable X--usually some variable we suspect may ber related to the response

Questions

Scatter plots can provide answers to the following questions:

1. Are variables X and Y related?
2. Are variables X and Y linearly related?
3. Are variables X and Y non-linearly related?
4. Does the variation in Y change depending on X?
5. Are there outliers?

Examples

1. No relationship
2. Strong linear (positive correlation)
3. Strong linear (negative correlation)
4. Exact linear (positive correlation)
5. Quadratic relationship
6. Exponential relationship
7. Sinusoidal relationship (damped)
8. Variation of Y doesn't depend on X (homoscedastic)
9. Variation of Y does depend on X (heteroscedastic)
10. Outlier

Combining Scatter Plots

Scatter plots can also be combined in multiple plots per page to help understand higher-level structure in data sets with more than two variables.

The scatterplot matrix generates all pairwise scatter plots on a single page. The conditioning plot, also called a co-plot or subset plot, generates scatter plots of Y versus X dependent on the value of a third variable.

Causality Is Not Proved By Association

The scatter plot uncovers relationships in data. "Relationships" means that there is some structured association (linear, quadratic, etc.) between X and Y. Note, however, that even though

causality implies association

association does NOT imply causality.

Scatter plots are a useful diagnostic tool for determining association, but if such association exists, the plot may or may not suggest an underlying cause-and-effect mechanism. A scatter plot can never "prove" cause and effect--it is ultimately only the researcher (relying on the underlying science/engineering) who can conclude that causality actually exists.

Appearance

The most popular rendition of a scatter plot is

1. some plot character (e.g., X) at the data points, and
2. no line connecting data points.

Other scatter plot format variants include

1. an optional plot character (e.g, X) at the data points, but
2. a solid line connecting data points.

In both cases, the resulting plot is referred to as a scatter plot, although the former (discrete and disconnected) is the author's personal preference since nothing makes it onto the screen except the data--there are no interpolative artifacts to bias the interpretation.

Related Techniques

Run Sequence Plot
Box Plot
Block Plot

Case Study

The scatter plot is demonstrated in the load cell calibration data case study.

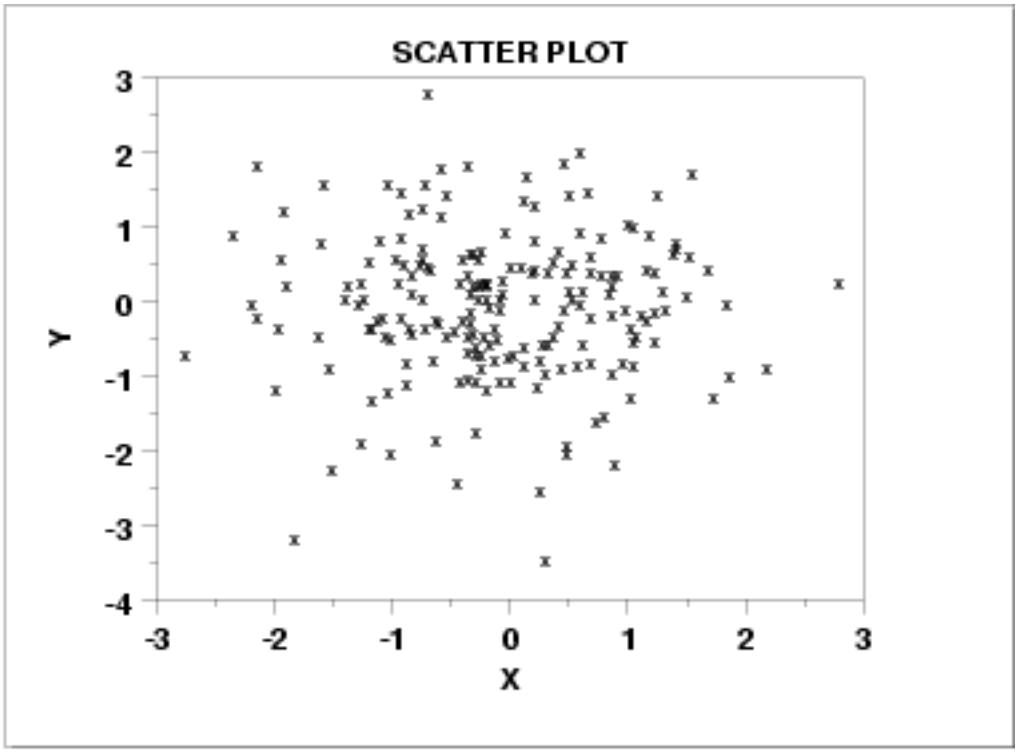
Software

Scatter plots are a fundamental technique that should be available in any general purpose statistical software program, including Dataplot. Scatter plots are also available in most graphics and spreadsheet programs as well.

1. Exploratory Data Analysis
1.3. EDA Techniques
1.3.3. Graphical Techniques: Alphabetic
1.3.3.26. Scatter Plot

1.3.3.26.1. Scatter Plot: No Relationship

Scatter Plot with No Relationship

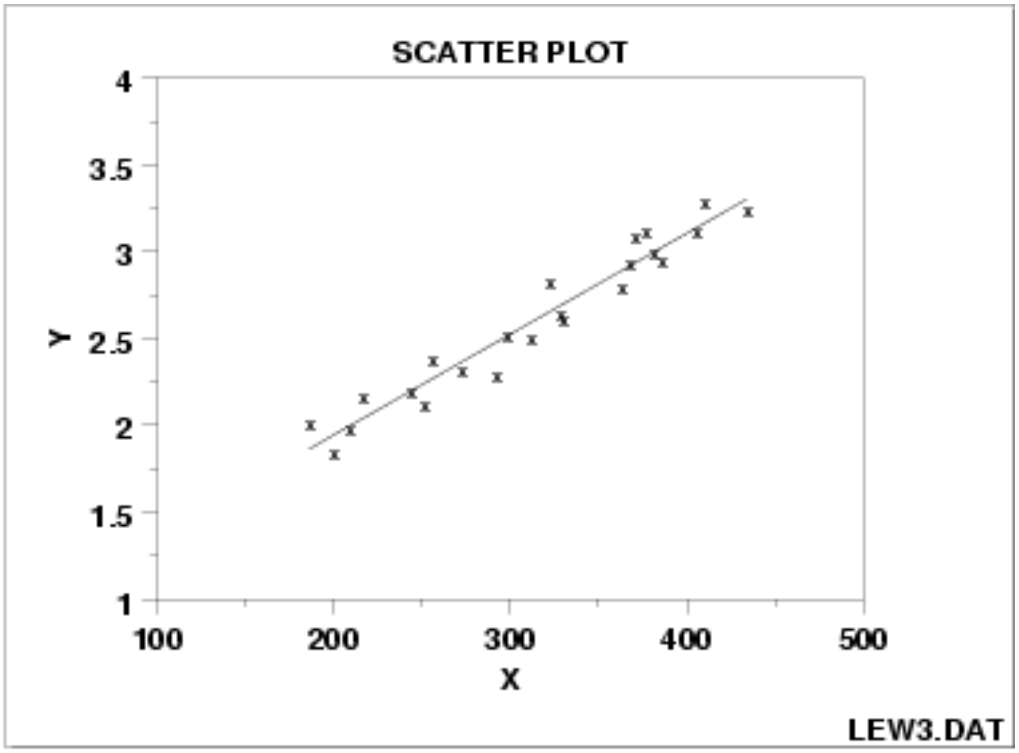


Discussion

Note in the plot above how for a given value of X (say $X = 0.5$), the corresponding values of Y range all over the place from $Y = -2$ to $Y = +2$. The same is true for other values of X. This lack of predictablility in determining Y from a given value of X, and the associated amorphous, non-structured appearance of the scatter plot leads to the summary conclusion: no relationship.

1.3.3.26.2. Scatter Plot: Strong Linear (positive correlation) Relationship

Scatter Plot Showing Strong Positive Linear Correlation

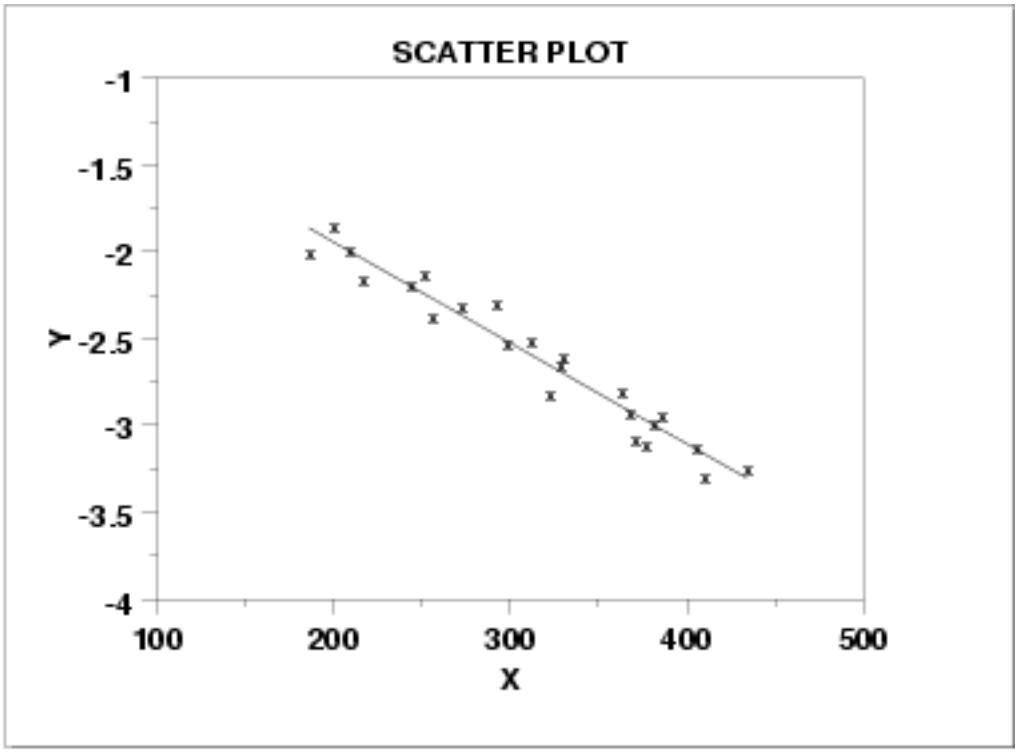


Discussion

Note in the plot above how a straight line comfortably fits through the data; hence a linear relationship exists. The scatter about the line is quite small, so there is a strong linear relationship. The slope of the line is positive (small values of *X* correspond to small values of *Y*; large values of *X* correspond to large values of *Y*), so there is a positive co-relation (that is, a positive correlation) between *X* and *Y*.

1.3.3.26.3. Scatter Plot: Strong Linear (negative correlation) Relationship

Scatter Plot Showing a Strong Negative Correlation

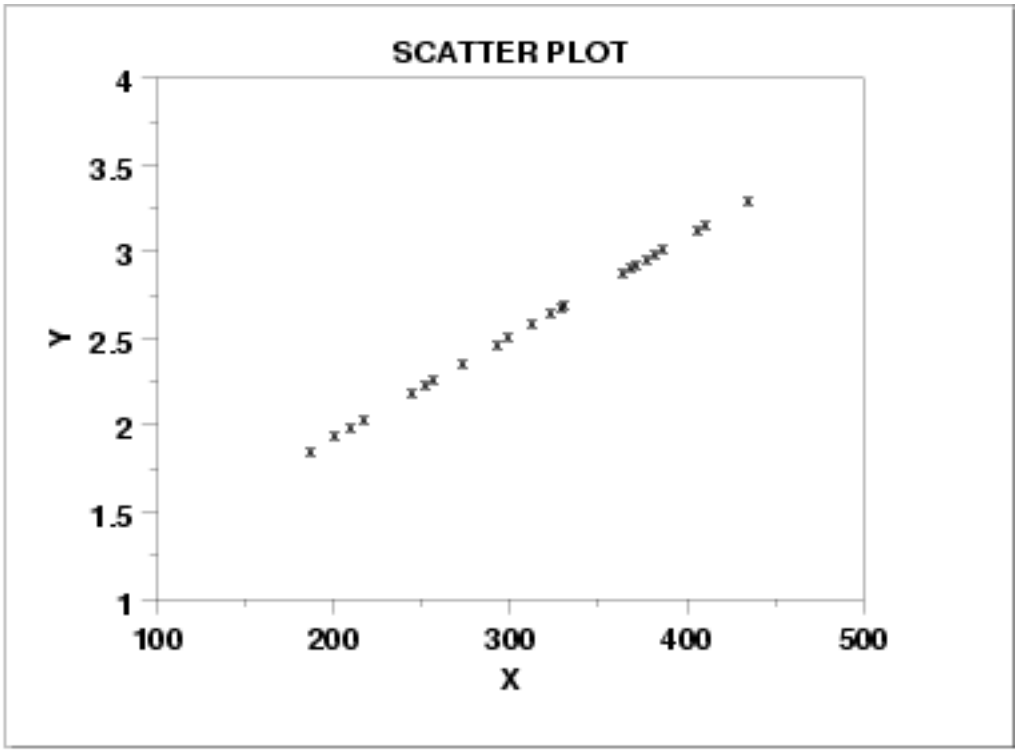


Discussion

Note in the plot above how a straight line comfortably fits through the data; hence there is a linear relationship. The scatter about the line is quite small, so there is a strong linear relationship. The slope of the line is negative (small values of X correspond to large values of Y; large values of X correspond to small values of Y), so there is a negative co-relation (that is, a negative correlation) between X and Y.

1.3.3.26.4. Scatter Plot: Exact Linear (positive correlation) Relationship

Scatter Plot Showing an Exact Linear Relationship



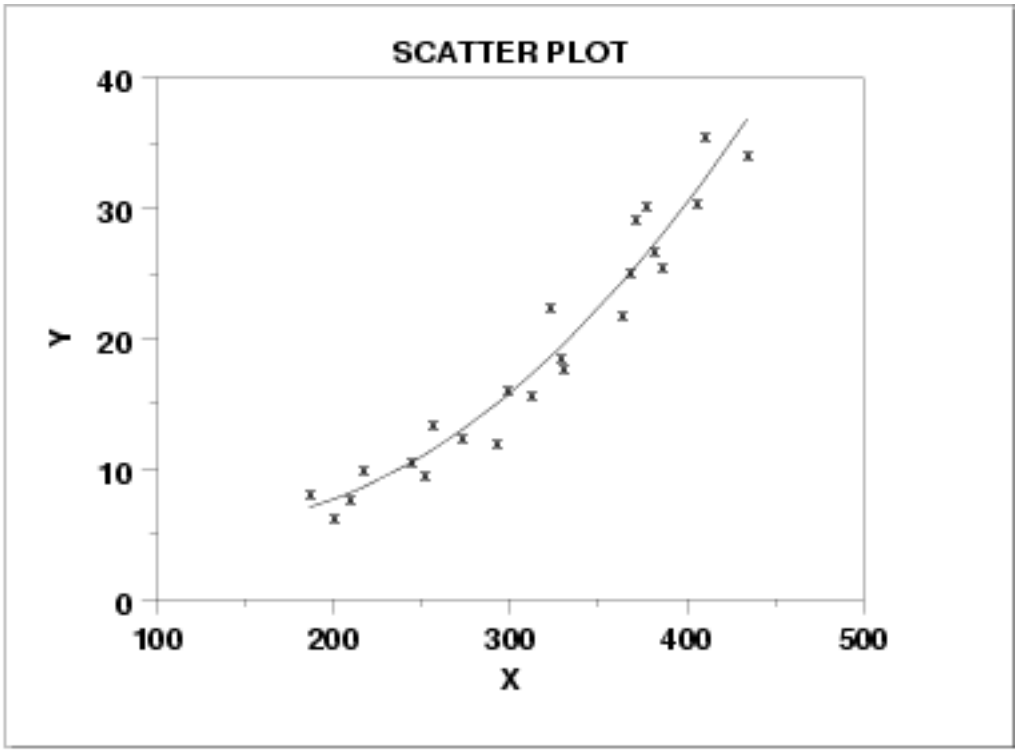
Discussion

Note in the plot above how a straight line comfortably fits through the data; hence there is a linear relationship. The scatter about the line is zero--there is perfect predictability between *X* and *Y*), so there is an exact linear relationship. The slope of the line is positive (small values of *X* correspond to small values of *Y*; large values of *X* correspond to large values of *Y*), so there is a positive co-relation (that is, a positive correlation) between *X* and *Y*.

1. Exploratory Data Analysis
1.3. EDA Techniques
1.3.3. Graphical Techniques: Alphabetic
1.3.3.26. Scatter Plot

1.3.3.26.5. Scatter Plot: Quadratic Relationship

Scatter Plot Showing Quadratic Relationship



Discussion

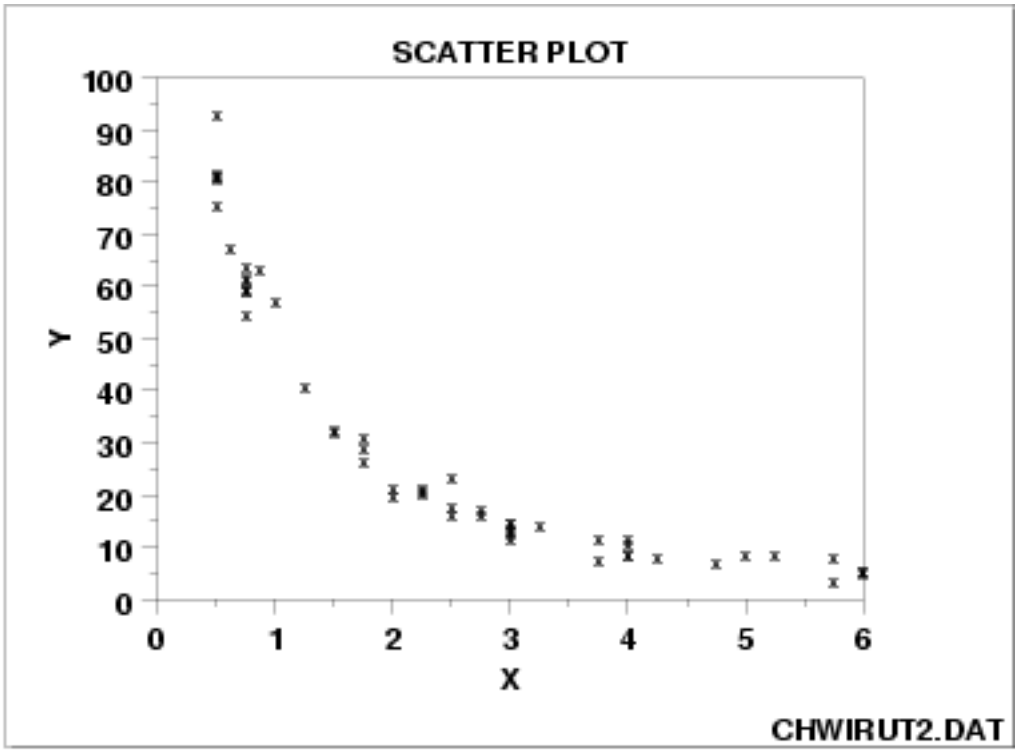
Note in the plot above how no imaginable simple straight line could ever adequately describe the relationship between *X* and *Y*--a curved (or curvilinear, or non-linear) function is needed. The simplest such curvilinear function is a quadratic model

$$Y_i = A + BX_i + CX_i^2 + E_i$$

for some A, B, and C. Many other curvilinear functions are possible, but the data analysis principle of parsimony suggests that we try fitting a quadratic function first.

1.3.3.26.6. Scatter Plot: Exponential Relationship

Scatter Plot Showing Exponential Relationship



Discussion

Note that a simple straight line is grossly inadequate in describing the relationship between *X* and *Y*. A quadratic model would prove lacking, especially for large values of *X*. In this example, the large values of *X* correspond to nearly constant values of *Y*, and so a non-linear function beyond the quadratic is needed. Among the many other non-linear functions available, one of the simpler ones is the exponential model

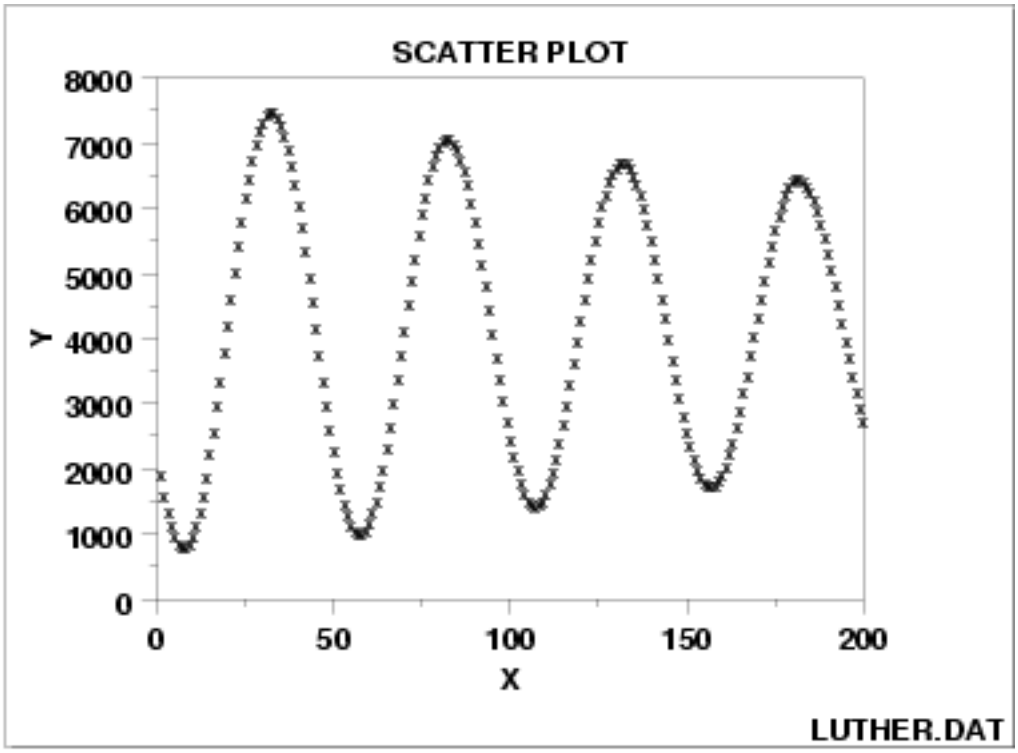
$$Y_i = A + Be^{CX_i} + E_i$$

for some A, B, and C. In this case, an exponential function would, in fact, fit well, and so one is led to the summary conclusion of an exponential relationship.

1. Exploratory Data Analysis
1.3. EDA Techniques
1.3.3. Graphical Techniques: Alphabetic
1.3.3.26. Scatter Plot

1.3.3.26.7. Scatter Plot: Sinusoidal Relationship (damped)

Scatter Plot Showing a Sinusoidal Relationship



Discussion

The complex relationship between *X* and *Y* appears to be basically oscillatory, and so one is naturally drawn to the trigonometric sinusoidal model:

$$Y_i = C + \alpha \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

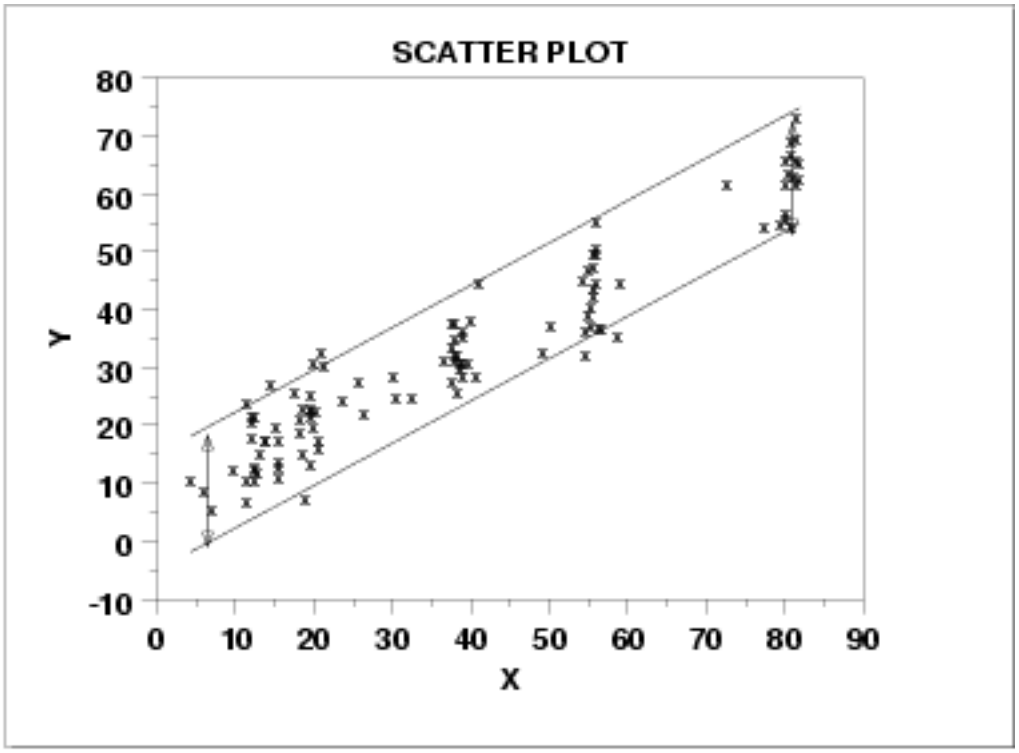
Closer inspection of the scatter plot reveals that the amount of swing (the amplitude α in the model) does not appear to be constant but rather is decreasing (damping) as *X* gets large. We thus would be led to the conclusion: damped sinusoidal relationship, with the simplest corresponding model being

$$Y_i = C + \left(B_0 + B_1 * t_i\right) \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

1. Exploratory Data Analysis
1.3. EDA Techniques
1.3.3. Graphical Techniques: Alphabetic
1.3.3.26. Scatter Plot

1.3.3.26.8. Scatter Plot: Variation of Y Does Not Depend on X (homoscedastic)

Scatter Plot Showing Homoscedastic Variability

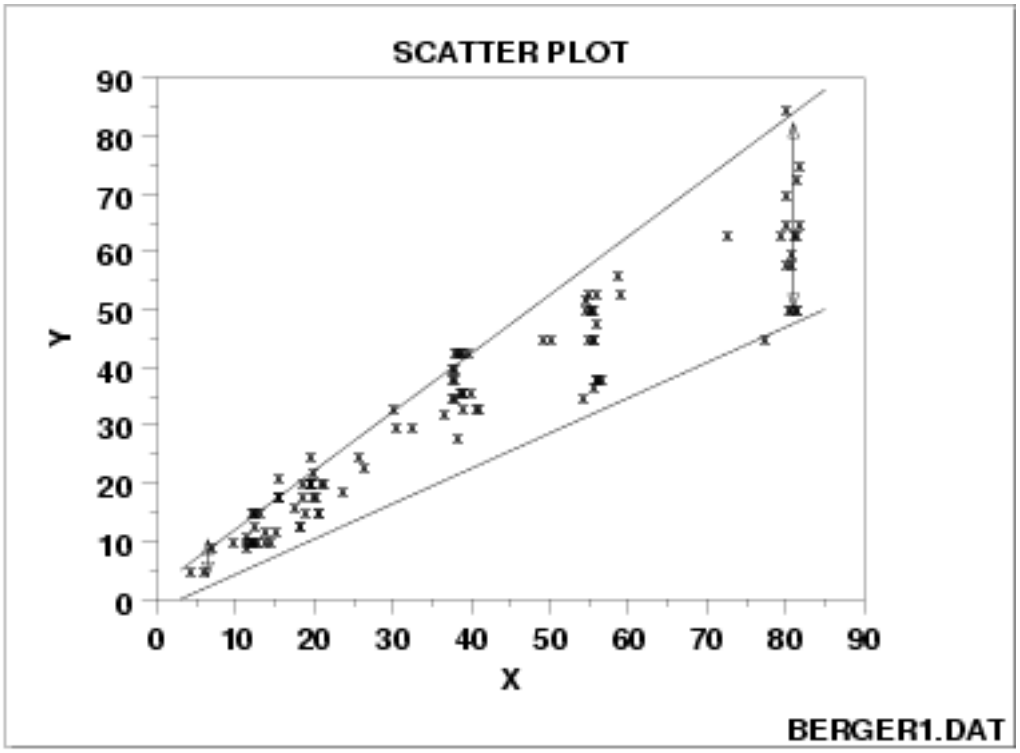


Discussion

This scatter plot reveals a linear relationship between *X* and *Y*: for a given value of *X*, the predicted value of *Y* will fall on a line. The plot further reveals that the variation in *Y* about the predicted value is about the same (+- 10 units), regardless of the value of *X*. Statistically, this is referred to as homoscedasticity. Such homoscedasticity is very important as it is an underlying assumption for regression, and its violation leads to parameter estimates with inflated variances. If the data are homoscedastic, then the usual regression estimates can be used. If the data are not homoscedastic, then the estimates can be improved using weighting procedures as shown in the next example.

1.3.3.26.9. Scatter Plot: Variation of Y Does Depend on X (heteroscedastic)

Scatter Plot Showing Heteroscedastic Variability



Discussion

This scatter plot reveals an approximate linear relationship between *X* and *Y*, but more importantly, it reveals a statistical condition referred to as heteroscedasticity (that is, nonconstant variation in *Y* over the values of *X*). For a heteroscedastic data set, the variation in *Y* differs depending on the value of *X*. In this example, small values of *X* yield small scatter in *Y* while large values of *X* result in large scatter in *Y*.

Heteroscedasticity complicates the analysis somewhat, but its effects can be overcome by:

1. proper weighting of the data with noisier data being weighted less, or by

2. performing a *Y* variable transformation to achieve homoscedasticity. The Box-Cox normality plot can help determine a suitable transformation.

Impact of Ignoring Unequal Variability in the Data

Fortunately, unweighted regression analyses on heteroscedastic data produce estimates of the coefficients that are unbiased. However, the coefficients will not be as precise as they would be with proper weighting.

Note further that if heteroscedasticity does exist, it is frequently useful to plot and model the local variation $var(Y_i|X_i)$ as a function of *X*, as in $var(Y_i|X_i) = g(X_i)$. This modeling has two advantages:

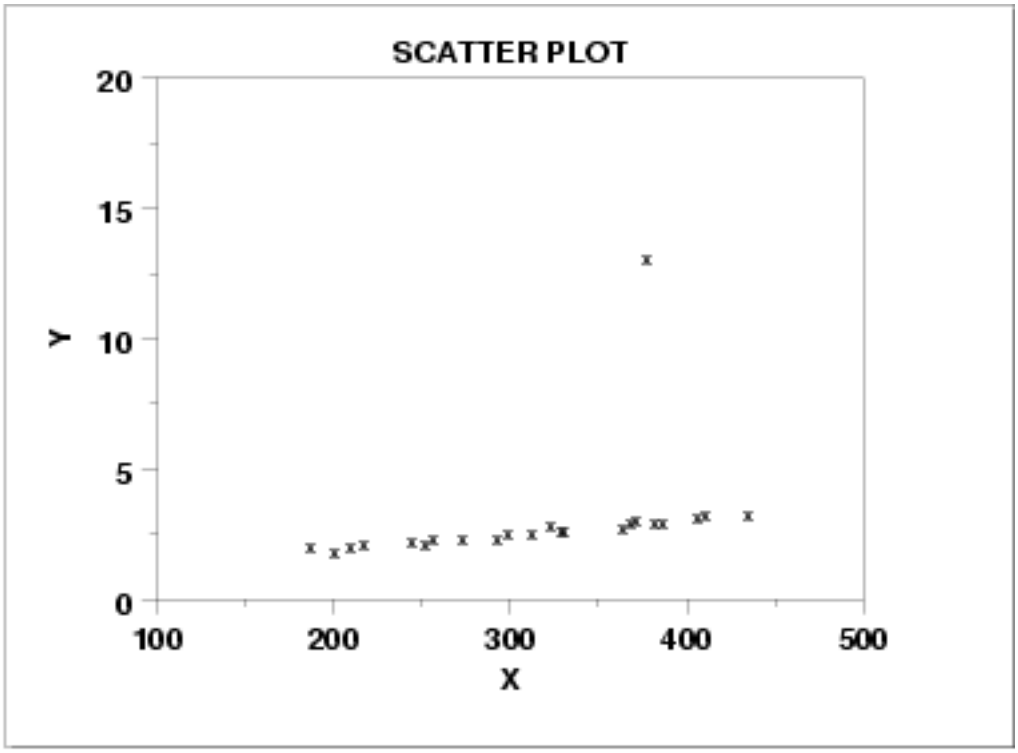
1. it provides additional insight and understanding as to how the response *Y* relates to *X*; and
2. it provides a convenient means of forming weights for a weighted regression by simply using

$$w_i = W(Y_i|X_i) = \frac{1}{Var(Y_i|X_i)} = \frac{1}{g(X_i)}$$

The topic of non-constant variation is discussed in some detail in the process modeling chapter.

1.3.3.26.10. Scatter Plot: Outlier

Scatter Plot Showing Outliers



Discussion

The scatter plot here reveals

1. a basic linear relationship between *X* and *Y* for most of the data, and
2. a single outlier (at *X* = 375).

An outlier is defined as a data point that emanates from a different model than do the rest of the data. The data here appear to come from a linear model with a given slope and variation except for the outlier which appears to have been generated from some other model.

Outlier detection is important for effective modeling. Outliers should be excluded from such model fitting. If all the data here are included in a linear regression, then the fitted model will be poor virtually everywhere. If the outlier is omitted from the fitting process, then the resulting fit will be excellent almost everywhere (for all points except

the outlying point).

1.3.3.26.11. Scatterplot Matrix

Purpose: Check Pairwise Relationships Between Variables

Given a set of variables $\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_k}$, the scatterplot matrix contains all the pairwise scatter plots of the variables on a single page in a matrix format. That is, if there are *k* variables, the scatterplot matrix will have *k* rows and *k* columns and the *i*th row and *j*th column of this matrix is a plot of $\boldsymbol{X_i}$ versus $\boldsymbol{X_j}$.

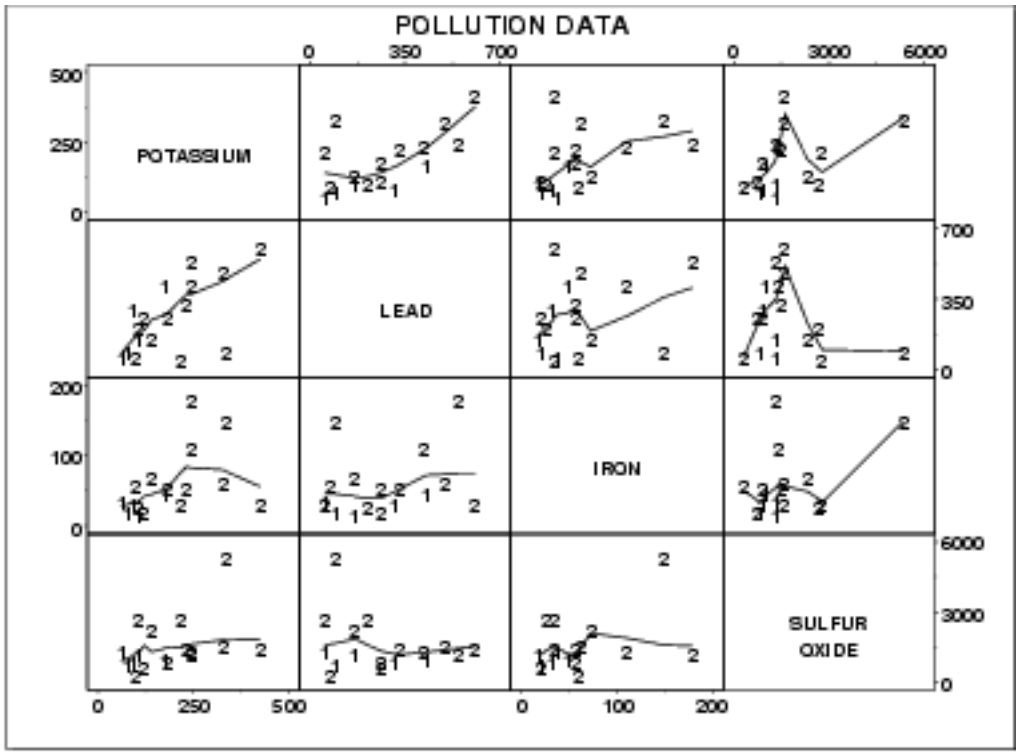
Although the basic concept of the scatterplot matrix is simple, there are numerous alternatives in the details of the plots.

1. The diagonal plot is simply a 45-degree line since we are plotting $\boldsymbol{X_i}$ versus $\boldsymbol{X_i}$. Although this has some usefulness in terms of showing the univariate distribution of the variable, other alternatives are common. Some users prefer to use the diagonal to print the variable label. Another alternative is to plot the univariate histogram on the diagonal. Alternatively, we could simply leave the diagonal blank.
2. Since $\boldsymbol{X_i}$ versus $\boldsymbol{X_j}$ is equivalent to $\boldsymbol{X_j}$ versus $\boldsymbol{X_i}$ with the axes reversed, some prefer to omit the plots below the diagonal.
3. It can be helpful to overlay some type of fitted curve on the scatter plot. Although a linear or quadratic fit can be used, the most common alternative is to overlay a lowess curve.
4. Due to the potentially large number of plots, it can be somewhat tricky to provide the axes labels in a way that is both informative and visually pleasing. One alternative that seems to work well is to provide axis labels on alternating rows and columns. That is, row one will have tic marks and axis labels on the left vertical axis for the first plot only while row two will have the tic marks and axis labels for the right vertical axis for the last plot in the row only. This alternating pattern continues for the remaining rows. A similar pattern is used for the columns and the horizontal axes labels. Another alternative is to put the minimum and maximum scale value in the diagonal plot with the variable

name.

5. Some analysts prefer to connect the scatter plots. Others prefer to leave a little gap between each plot.
6. Although this plot type is most commonly used for scatter plots, the basic concept is both simple and powerful and extends easily to other plot formats that involve pairwise plots such as the quantile-quantile plot and the bihistogram.

Sample Plot



This sample plot was generated from pollution data collected by NIST chemist Lloyd Currie.

There are a number of ways to view this plot. If we are primarily interested in a particular variable, we can scan the row and column for that variable. If we are interested in finding the strongest relationship, we can scan all the plots and then determine which variables are related.

Definition

Given *k* variables, scatter plot matrices are formed by creating *k* rows and *k* columns. Each row and column defines a single scatter plot

The individual plot for row *i* and column *j* is defined as

- Vertical axis: Variable $\boldsymbol{X_i}$
- Horizontal axis: Variable $\boldsymbol{X_j}$

Questions

The scatterplot matrix can provide answers to the following questions:

1. Are there pairwise relationships between the variables?
2. If there are relationships, what is the nature of these relationships?
3. Are there outliers in the data?
4. Is there clustering by groups in the data?

Linking and Brushing

The scatterplot matrix serves as the foundation for the concepts of linking and brushing.

By linking, we mean showing how a point, or set of points, behaves in each of the plots. This is accomplished by highlighting these points in some fashion. For example, the highlighted points could be drawn as a filled circle while the remaining points could be drawn as unfilled circles. A typical application of this would be to show how an outlier shows up in each of the individual pairwise plots. Brushing extends this concept a bit further. In brushing, the points to be highlighted are interactively selected by a mouse and the scatterplot matrix is dynamically updated (ideally in real time). That is, we can select a rectangular region of points in one plot and see how those points are reflected in the other plots. Brushing is discussed in detail by Becker, Cleveland, and Wilks in the paper *"Dynamic Graphics for Data Analysis"* (Cleveland and McGill, 1988).

Related Techniques

Star plot
Scatter plot
Conditioning plot
Locally weighted least squares

Software

Scatterplot matrices are becoming increasingly common in general purpose statistical software programs, including Dataplot. If a software program does not generate scatterplot matrices, but it does provide multiple plots per page and scatter plots, it should be possible to write a macro to generate a scatterplot matrix. Brushing is available in a few of the general purpose statistical software programs that emphasize graphical approaches.

1.3.3.26.12. Conditioning Plot

Purpose: Check pairwise relationship between two variables conditional on a third variable

A conditioning plot, also known as a coplot or subset plot, is a plot of two variables conditional on the value of a third variable (called the conditioning variable). The conditioning variable may be either a variable that takes on only a few discrete values or a continuous variable that is divided into a limited number of subsets.

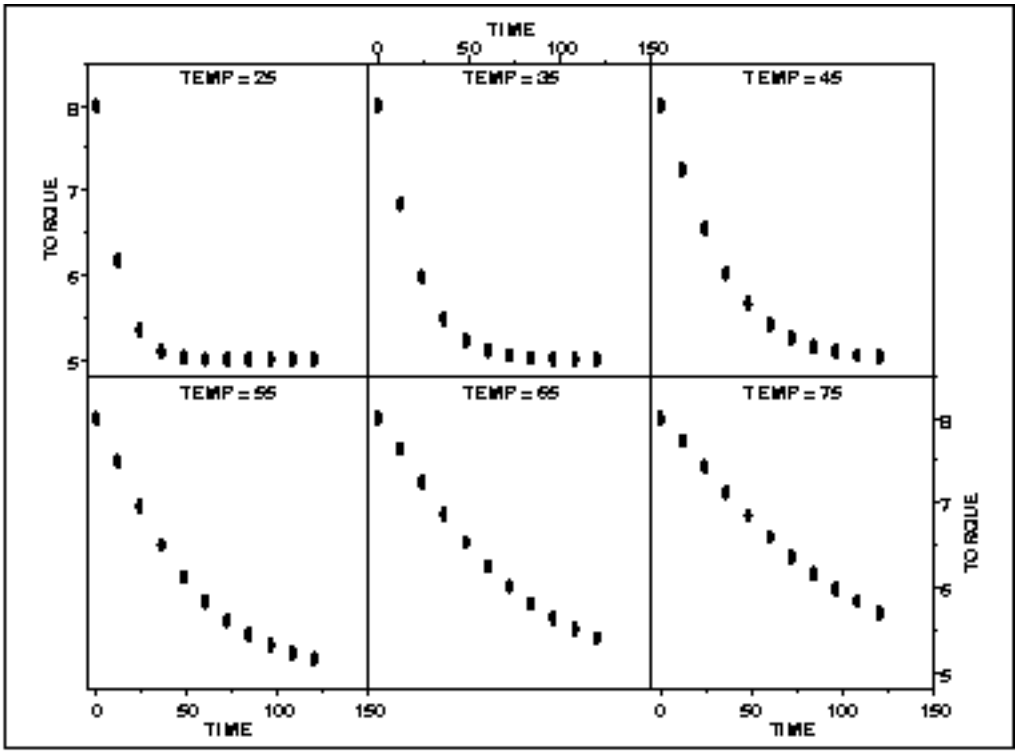
One limitation of the scatterplot matrix is that it cannot show interaction effects with another variable. This is the strength of the conditioning plot. It is also useful for displaying scatter plots for groups in the data. Although these groups can also be plotted on a single plot with different plot symbols, it can often be visually easier to distinguish the groups using the conditioning plot.

Although the basic concept of the conditioning plot matrix is simple, there are numerous alternatives in the details of the plots.

1. It can be helpful to overlay some type of fitted curve on the scatter plot. Although a linear or quadratic fit can be used, the most common alternative is to overlay a lowess curve.
2. Due to the potentially large number of plots, it can be somewhat tricky to provide the axis labels in a way that is both informative and visually pleasing. One alternative that seems to work well is to provide axis labels on alternating rows and columns. That is, row one will have tic marks and axis labels on the left vertical axis for the first plot only while row two will have the tic marks and axis labels for the right vertical axis for the last plot in the row only. This alternating pattern continues for the remaining rows. A similar pattern is used for the columns and the horizontal axis labels. Note that this approach only works if the axes limits are fixed to common values for all of the plots.
3. Some analysts prefer to connect the scatter plots. Others prefer to leave a little gap between each plot. Alternatively, each plot can have its own labeling with the plots not connected.

4. Although this plot type is most commonly used for scatter plots, the basic concept is both simple and powerful and extends easily to other plot formats.

Sample Plot



In this case, temperature has six distinct values. We plot torque versus time for each of these temperatures. This example is discussed in more detail in the process modeling chapter.

Definition

Given the variables ***X***, ***Y***, and ***Z***, the conditioning plot is formed by dividing the values of ***Z*** into *k* groups. There are several ways that these groups may be formed. There may be a natural grouping of the data, the data may be divided into several equal sized groups, the grouping may be determined by clusters in the data, and so on. The page will be divided into *n* rows and *c* columns where $nc \geq k$. Each row and column defines a single scatter plot.

The individual plot for row *i* and column *j* is defined as

- Vertical axis: Variable ***Y***
- Horizontal axis: Variable ***X***

where only the points in the group corresponding to the *i*th row and *j*th column are used.

Questions

The conditioning plot can provide answers to the following questions:

1. Is there a relationship between two variables?
2. If there is a relationship, does the nature of the relationship depend on the value of a third variable?
3. Are groups in the data similar?
4. Are there outliers in the data?

Related Techniques

Scatter plot
Scatterplot matrix
Locally weighted least squares

Software

Scatter plot matrices are becoming increasingly common in general purpose statistical software programs, including Dataplot. If a software program does not generate conditioning plots, but it does provide multiple plots per page and scatter plots, it should be possible to write a macro to generate a conditioning plot.

1.3.3.27. Spectral Plot

Purpose: Examine Cyclic Structure

A spectral plot (Jenkins and Watts 1968 or Bloomfield 1976) is a graphical technique for examining cyclic structure in the frequency domain. It is a smoothed Fourier transform of the autocovariance function.

The frequency is measured in cycles per unit time where unit time is defined to be the distance between 2 points. A frequency of 0 corresponds to an infinite cycle while a frequency of 0.5 corresponds to a cycle of 2 data points. Equi-spaced time series are inherently limited to detecting frequencies between 0 and 0.5.

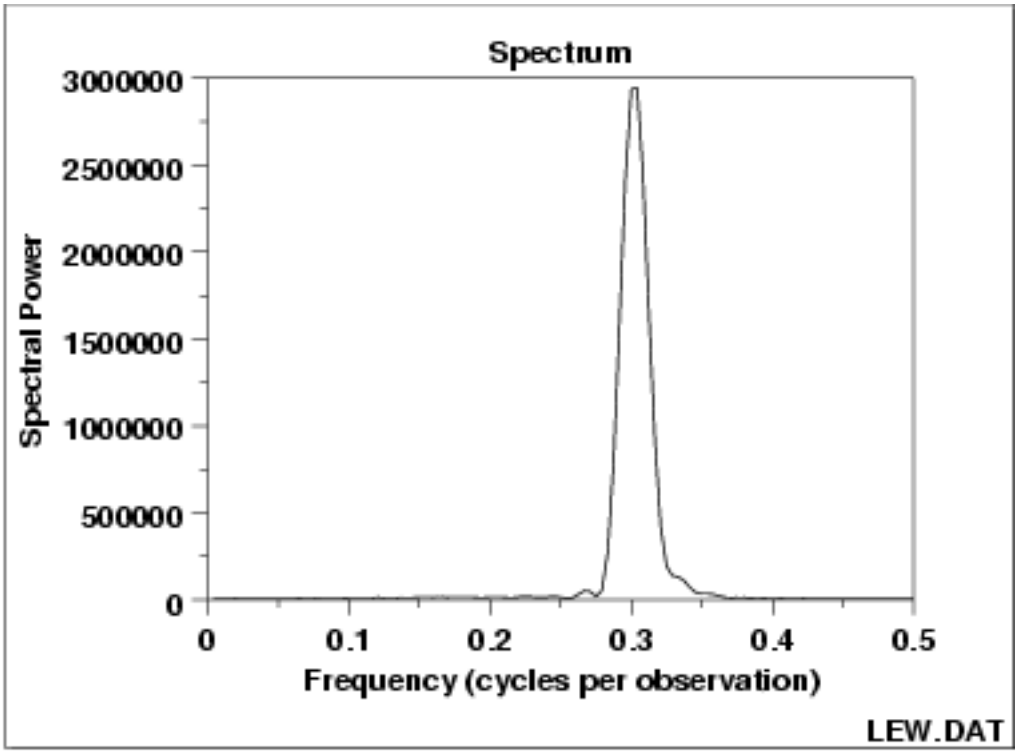
Trends should typically be removed from the time series before applying the spectral plot. Trends can be detected from a run sequence plot. Trends are typically removed by differencing the series or by fitting a straight line (or some other polynomial curve) and applying the spectral analysis to the residuals.

Spectral plots are often used to find a starting value for the frequency, ω, in the sinusoidal model

$$Y_i = C + \alpha \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

See the beam deflection case study for an example of this.

Sample Plot



This spectral plot shows one dominant frequency of approximately 0.3 cycles per observation.

Definition: Variance Versus Frequency

The spectral plot is formed by:

- Vertical axis: Smoothed variance (power)
- Horizontal axis: Frequency (cycles per observation)

The computations for generating the smoothed variances can be involved and are not discussed further here. The details can be found in the Jenkins and Bloomfield references and in most texts that discuss the frequency analysis of time series.

Questions

The spectral plot can be used to answer the following questions:

1. How many cyclic components are there?
2. Is there a dominant cyclic frequency?
3. If there is a dominant cyclic frequency, what is it?

Importance Check Cyclic Behavior of Time Series

The spectral plot is the primary technique for assessing the cyclic nature of univariate time series in the frequency domain. It is almost always the second plot (after a run sequence plot) generated in a frequency domain analysis of a time series.

Examples

1. Random (= White Noise)
2. Strong autocorrelation and autoregressive model
3. Sinusoidal model

Related Techniques

Autocorrelation Plot
Complex Demodulation Amplitude Plot
Complex Demodulation Phase Plot

Case Study

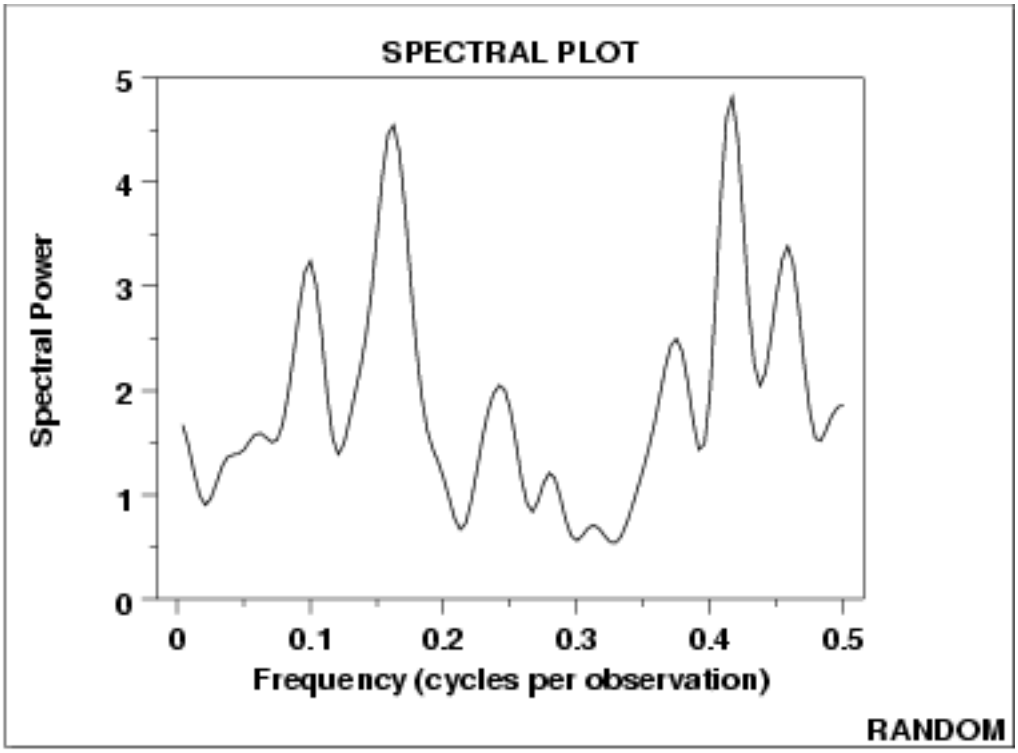
The spectral plot is demonstrated in the beam deflection data case study.

Software

Spectral plots are a fundamental technique in the frequency analysis of time series. They are available in many general purpose statistical software programs, including Dataplot.

1.3.3.27.1. Spectral Plot: Random Data

Spectral Plot of 200 Normal Random Numbers



Conclusions

We can make the following conclusions from the above plot.

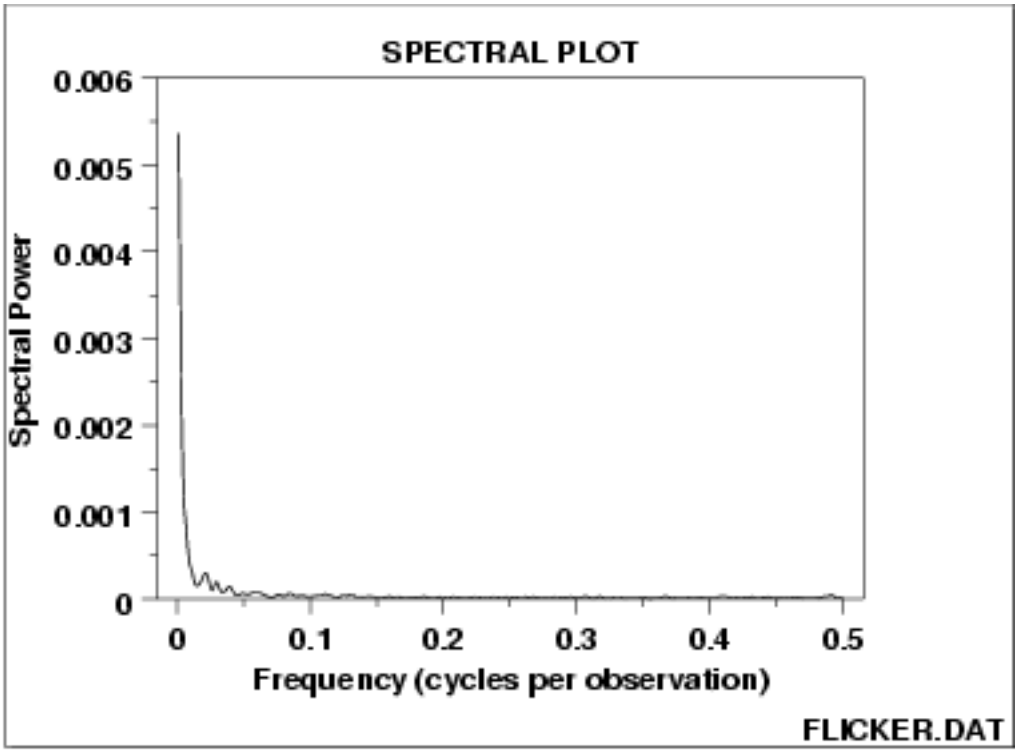
1. There are no dominant peaks.
2. There is no identifiable pattern in the spectrum.
3. The data are random.

Discussion

For random data, the spectral plot should show no dominant peaks or distinct pattern in the spectrum. For the sample plot above, there are no clearly dominant peaks and the peaks seem to fluctuate at random. This type of appearance of the spectral plot indicates that there are no significant cyclic patterns in the data.

1.3.3.27.2. Spectral Plot: Strong Autocorrelation and Autoregressive Model

Spectral Plot for Random Walk Data



Conclusions

We can make the following conclusions from the above plot.

1. Strong dominant peak near zero.
2. Peak decays rapidly towards zero.
3. An autoregressive model is an appropriate model.

Discussion

This spectral plot starts with a dominant peak near zero and rapidly decays to zero. This is the spectral plot signature of a process with strong positive autocorrelation. Such processes are highly non-random in that there is high association between an observation and a succeeding observation. In short, if you know Y_i you can make a strong guess as to what Y_{i+1} will be.

Recommended Next Step

The next step would be to determine the parameters for the autoregressive model:

$$Y_i = A_0 + A_1 * Y_{i-1} + E_i$$

Such estimation can be done by linear regression or by fitting a Box-Jenkins autoregressive (AR) model.

The residual standard deviation for this autoregressive model will be much smaller than the residual standard deviation for the default model

$$Y_i = A_0 + E_i$$

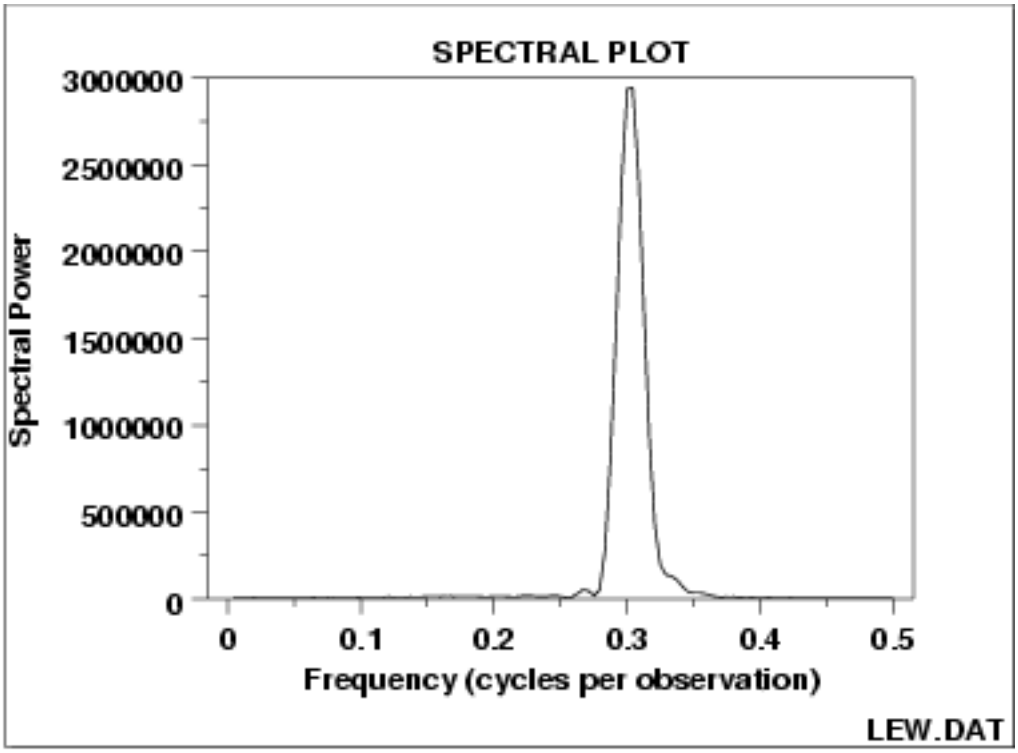
Then the system should be reexamined to find an explanation for the strong autocorrelation. Is it due to the

1. phenomenon under study; or
2. drifting in the environment; or
3. contamination from the data acquisition system (DAS)?

Oftentimes the source of the problem is item (3) above where contamination and carry-over from the data acquisition system result because the DAS does not have time to electronically recover before collecting the next data point. If this is the case, then consider slowing down the sampling rate to re-achieve randomness.

1.3.3.27.3. Spectral Plot: Sinusoidal Model

Spectral Plot for Sinusoidal Model



Conclusions

We can make the following conclusions from the above plot.

1. There is a single dominant peak at approximately 0.3.
2. There is an underlying single-cycle sinusoidal model.

Discussion

This spectral plot shows a single dominant frequency. This indicates that a single-cycle sinusoidal model might be appropriate.

If one were to naively assume that the data represented by the graph could be fit by the model

$$Y_i = A_0 + E_i$$

and then estimate the constant by the sample mean, the analysis would be incorrect because

- the sample mean is biased;
- the confidence interval for the mean, which is valid only for random data, is meaningless and too small.

On the other hand, the choice of the proper model

$$Y_i = C + \alpha \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

where α is the amplitude, ω is the frequency (between 0 and .5 cycles per observation), and ϕ is the phase can be fit by non-linear least squares. The beam deflection data case study demonstrates fitting this type of model.

Recommended Next Steps

The recommended next steps are to:

1. Estimate the frequency from the spectral plot. This will be helpful as a starting value for the subsequent non-linear fitting. A complex demodulation phase plot can be used to fine tune the estimate of the frequency before performing the non-linear fit.
2. Do a complex demodulation amplitude plot to obtain an initial estimate of the amplitude and to determine if a constant amplitude is justified.
3. Carry out a non-linear fit of the model

$$Y_i = C + \alpha \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

1.3.3.28. Standard Deviation Plot

Purpose: Detect Changes in Scale Between Groups

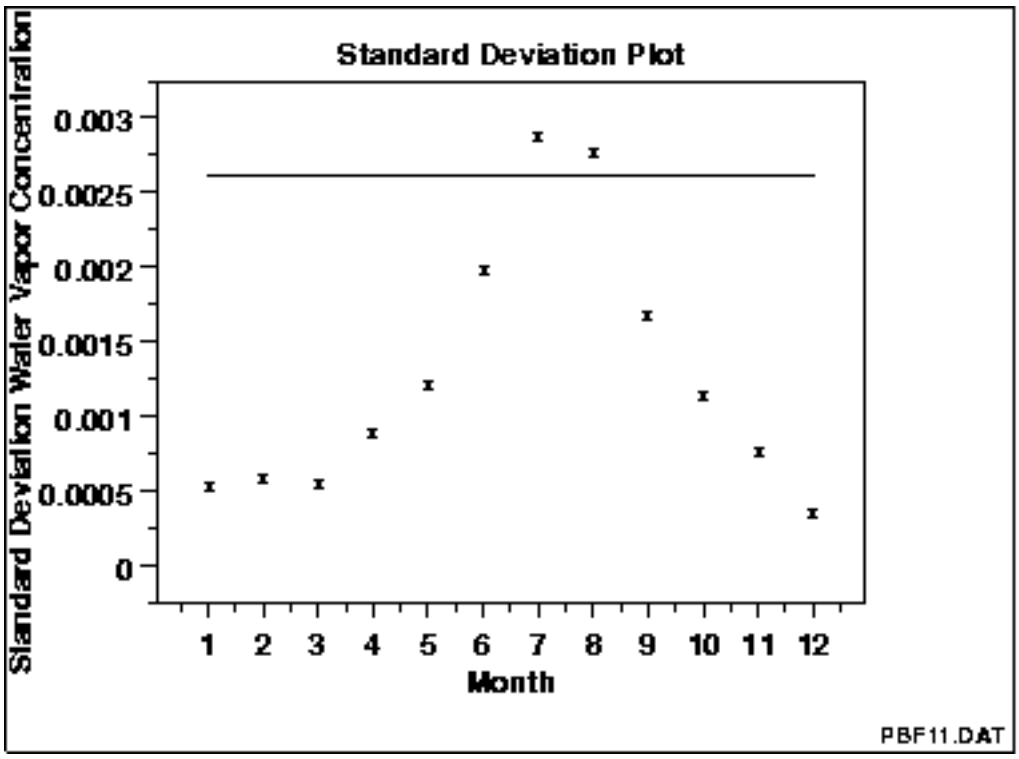
Standard deviation plots are used to see if the standard deviation varies between different groups of the data. The grouping is determined by the analyst. In most cases, the data provide a specific grouping variable. For example, the groups may be the levels of a factor variable. In the sample plot below, the months of the year provide the grouping.

Standard deviation plots can be used with ungrouped data to determine if the standard deviation is changing over time. In this case, the data are broken into an arbitrary number of equal-sized groups. For example, a data series with 400 points can be divided into 10 groups of 40 points each. A standard deviation plot can then be generated with these groups to see if the standard deviation is increasing or decreasing over time.

Although the standard deviation is the most commonly used measure of scale, the same concept applies to other measures of scale. For example, instead of plotting the standard deviation of each group, the median absolute deviation or the average absolute deviation might be plotted instead. This might be done if there were significant outliers in the data and a more robust measure of scale than the standard deviation was desired.

Standard deviation plots are typically used in conjunction with mean plots. The mean plot would be used to check for shifts in location while the standard deviation plot would be used to check for shifts in scale.

Sample Plot



This sample standard deviation plot shows

1. there is a shift in variation;
2. greatest variation is during the summer months.

Definition: Group Standard Deviations Versus Group ID

Standard deviation plots are formed by:

- Vertical axis: Group standard deviations
- Horizontal axis: Group identifier

A reference line is plotted at the overall standard deviation.

Questions

The standard deviation plot can be used to answer the following questions.

1. Are there any shifts in variation?
2. What is the magnitude of the shifts in variation?
3. Is there a distinct pattern in the shifts in variation?

Importance: Checking Assumptions

A common assumption in 1-factor analyses is that of equal variances. That is, the variance is the same for different levels of the factor variable. The standard deviation plot provides a graphical check for that assumption. A common assumption for univariate data is that the variance is constant. By grouping the data into equi-sized intervals, the standard deviation plot can provide a graphical test of this assumption.

Related Techniques	Mean Plot Dex Standard Deviation Plot
Software	Most general purpose statistical software programs do not support a standard deviation plot. However, if the statistical program can generate the standard deviation for a group, it should be feasible to write a macro to generate this plot. Dataplot supports a standard deviation plot.

1.3.3.29. Star Plot

Purpose: Display Multivariate Data

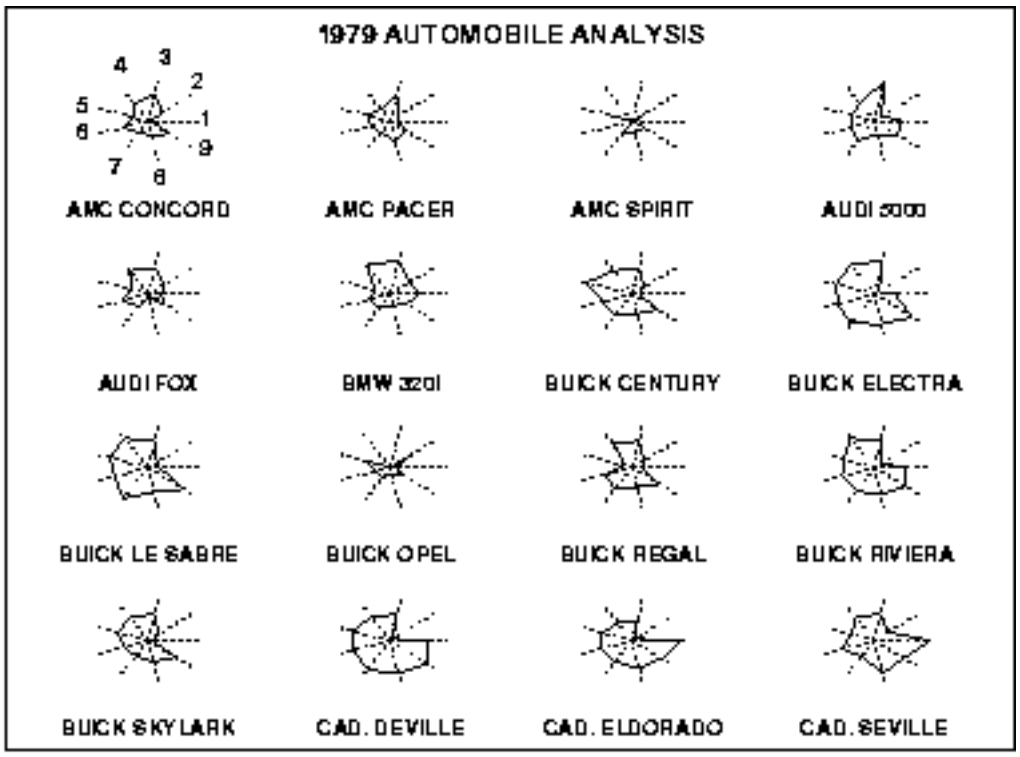
The star plot (Chambers 1983) is a method of displaying multivariate data. Each star represents a single observation. Typically, star plots are generated in a multi-plot format with many stars on each page and each star representing one observation.

Star plots are used to examine the relative values for a single data point (e.g., point 3 is large for variables 2 and 4, small for variables 1, 3, 5, and 6) and to locate similar points or dissimilar points.

Sample Plot

The plot below contains the star plots of 16 cars. The data file actually contains 74 cars, but we restrict the plot to what can reasonably be shown on one page. The variable list for the sample star plot is

1 Price
2 Mileage (MPG)
3 1978 Repair Record (1 = Worst, 5 = Best)
4 1977 Repair Record (1 = Worst, 5 = Best)
5 Headroom
6 Rear Seat Room
7 Trunk Space
8 Weight
9 Length



We can look at these plots individually or we can use them to identify clusters of cars with similar features. For example, we can look at the star plot of the Cadillac Seville and see that it is one of the most expensive cars, gets below average (but not among the worst) gas mileage, has an average repair record, and has average-to-above-average roominess and size. We can then compare the Cadillac models (the last three plots) with the AMC models (the first three plots). This comparison shows distinct patterns. The AMC models tend to be inexpensive, have below average gas mileage, and are small in both height and weight and in roominess. The Cadillac models are expensive, have poor gas mileage, and are large in both size and roominess.

Definition

The star plot consists of a sequence of equi-angular spokes, called radii, with each spoke representing one of the variables. The data length of a spoke is proportional to the magnitude of the variable for the data point relative to the maximum magnitude of the variable across all data points. A line is drawn connecting the data values for each spoke. This gives the plot a star-like appearance and the origin of the name of this plot.

Questions

The star plot can be used to answer the following questions:

1. What variables are dominant for a given observation?
2. Which observations are most similar, i.e., are there clusters of observations?
3. Are there outliers?

Weakness in Technique	Star plots are helpful for small-to-moderate-sized multivariate data sets. Their primary weakness is that their effectiveness is limited to data sets with less than a few hundred points. After that, they tend to be overwhelming. Graphical techniques suited for large data sets are discussed by Scott.
Related Techniques	Alternative ways to plot multivariate data are discussed in Chambers, du Toit, and Everitt.
Software	Star plots are available in some general purpose statistical software progams, including Dataplot.

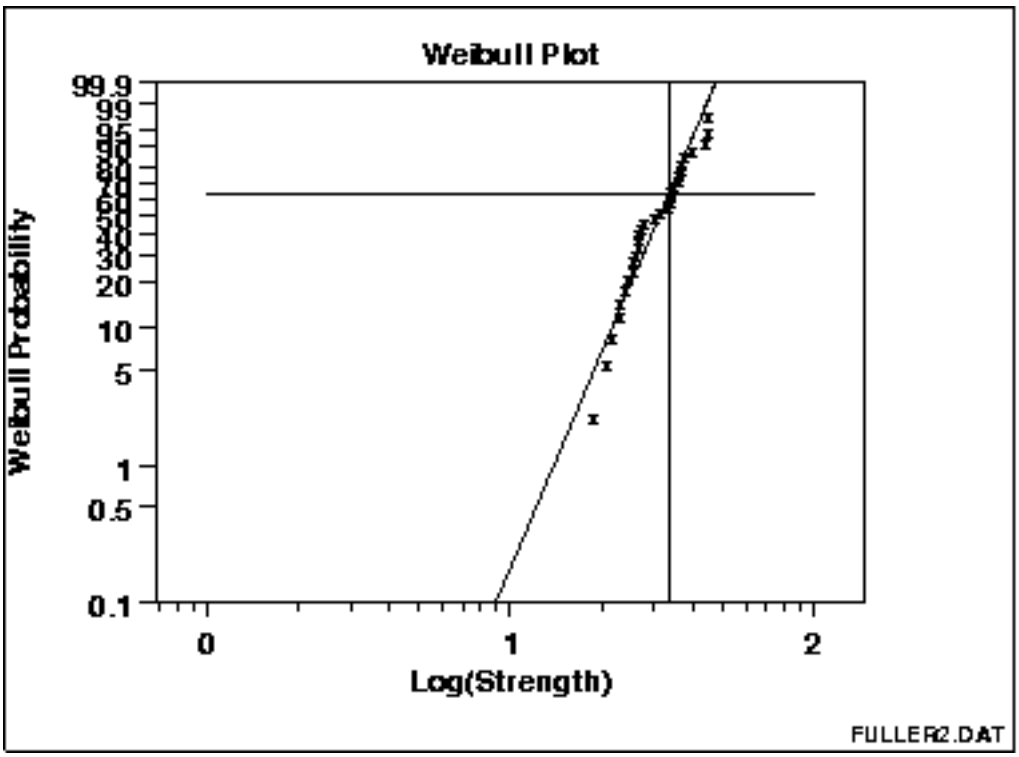
1.3.3.30. Weibull Plot

Purpose: Graphical Check To See If Data Come From a Population That Would Be Fit by a Weibull Distribution

The Weibull plot (Nelson 1982) is a graphical technique for determining if a data set comes from a population that would logically be fit by a 2-parameter Weibull distribution (the location is assumed to be zero).

The Weibull plot has special scales that are designed so that if the data do in fact follow a Weibull distribution, the points will be linear (or nearly linear). The least squares fit of this line yields estimates for the shape and scale parameters of the Weibull distribution. Weibull distribution (the location is assumed to be zero).

Sample Plot



This Weibull plot shows that:

1. the assumption of a Weibull distribution is reasonable;
2. the shape parameter estimate is computed to be 33.32;
3. the scale parameter estimate is computed to be 5.28; and

4. there are no outliers.

Definition: Weibull Cumulative Probability Versus LN(Ordered Response)

The Weibull plot is formed by:

- Vertical axis: Weibull cumulative probability expressed as a percentage
- Horizontal axis: LN of ordered response

The vertical scale is ln-ln(1-p) where p=(i-0.3)/(n+0.4) and *i* is the rank of the observation. This scale is chosen in order to linearize the resulting plot for Weibull data.

Questions

The Weibull plot can be used to answer the following questions:

1. Do the data follow a 2-parameter Weibull distribution?
2. What is the best estimate of the shape parameter for the 2-parameter Weibull distribution?
3. What is the best estimate of the scale (= variation) parameter for the 2-parameter Weibull distribution?

Importance: Check Distributional Assumptions

Many statistical analyses, particularly in the field of reliability, are based on the assumption that the data follow a Weibull distribution. If the analysis assumes the data follow a Weibull distribution, it is important to verify this assumption and, if verified, find good estimates of the Weibull parameters.

Related Techniques

Weibull Probability Plot
Weibull PPCC Plot
Weibull Hazard Plot

The Weibull probability plot (in conjunction with the Weibull PPCC plot), the Weibull hazard plot, and the Weibull plot are all similar techniques that can be used for assessing the adequacy of the Weibull distribution as a model for the data, and additionally providing estimation for the shape, scale, or location parameters.

The Weibull hazard plot and Weibull plot are designed to handle censored data (which the Weibull probability plot does not).

Case Study

The Weibull plot is demonstrated in the airplane glass failure data case study.

Software

Weibull plots are generally available in statistical software programs that are designed to analyze reliability data. Dataplot supports the Weibull plot.

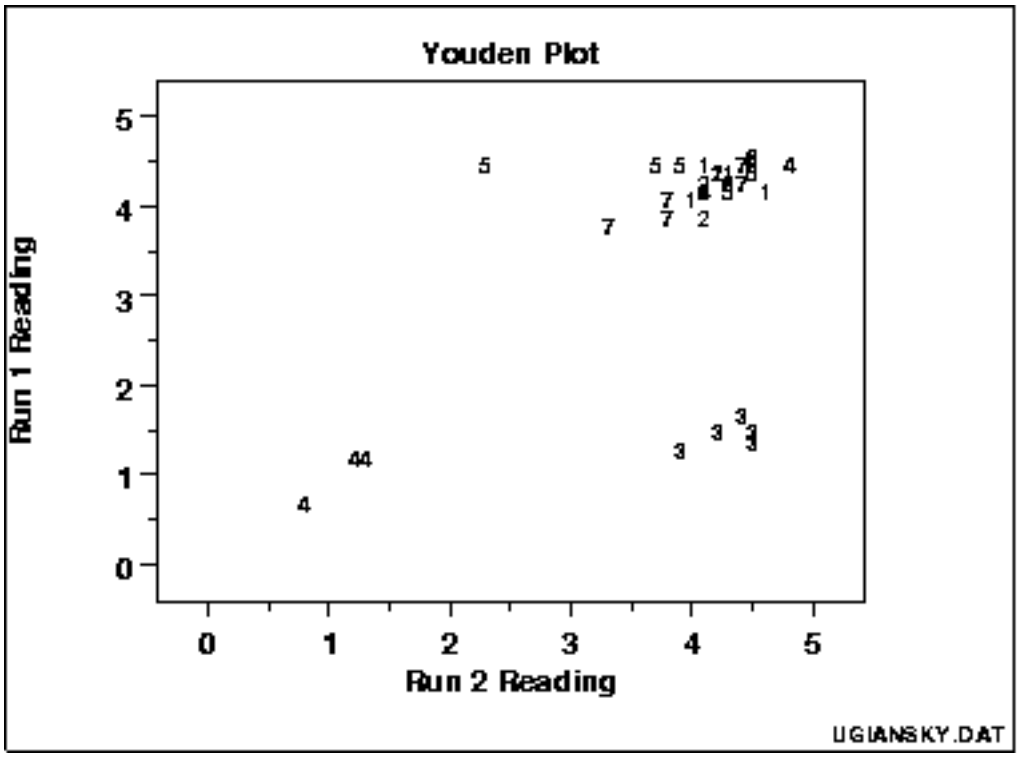
1.3.3.31. Youden Plot

Purpose: Interlab Comparisons

Youden plots are a graphical technique for analyzing interlab data when each lab has made two runs on the same product or one run on two different products.

The Youden plot is a simple but effective method for comparing both the within-laboratory variability and the between-laboratory variability.

Sample Plot



This plot shows:

1. Not all labs are equivalent.
2. Lab 4 is biased low.
3. Lab 3 has within-lab variability problems.
4. Lab 5 has an outlying run.

Definition: Response 1 Versus Response 2 Coded by Lab

Youden plots are formed by:

1. Vertical axis: Response variable 1 (i.e., run 1 or product 1 response value)
2. Horizontal axis: Response variable 2 (i.e., run 2 or product 2 response value)

In addition, the plot symbol is the lab id (typically an integer from 1 to *k* where *k* is the number of labs). Sometimes a 45-degree reference line is drawn. Ideally, a lab generating two runs of the same product should produce reasonably similar results. Departures from this reference line indicate inconsistency from the lab. If two different products are being tested, then a 45-degree line may not be appropriate. However, if the labs are consistent, the points should lie near some fitted straight line.

Questions

The Youden plot can be used to answer the following questions:

1. Are all labs equivalent?
2. What labs have between-lab problems (reproducibility)?
3. What labs have within-lab problems (repeatability)?
4. What labs are outliers?

Importance

In interlaboratory studies or in comparing two runs from the same lab, it is useful to know if consistent results are generated. Youden plots should be a routine plot for analyzing this type of data.

DEX Youden Plot

The dex Youden plot is a specialized Youden plot used in the design of experiments. In particular, it is useful for full and fractional designs.

Related Techniques

Scatter Plot

Software

The Youden plot is essentially a scatter plot, so it should be feasible to write a macro for a Youden plot in any general purpose statistical program that supports scatter plots. Dataplot supports a Youden plot.

1.3.3.31.1. DEX Youden Plot

DEX Youden Plot: Introduction

The dex (Design of Experiments) Youden plot is a specialized Youden plot used in the analysis of full and fractional experiment designs. In particular, it is used in support of a Yates analysis. These designs may have a low level, coded as "-1" or "-", and a high level, coded as "+1" or "+", for each factor. In addition, there can optionally be one or more center points. Center points are at the midpoint between the low and high levels for each factor and are coded as "0".

The Yates analysis and the the dex Youden plot only use the "-1" and "+1" points. The Yates analysis is used to estimate factor effects. The dex Youden plot can be used to help determine the approriate model to use from the Yates analysis.

Construction of DEX Youden Plot

The following are the primary steps in the construction of the dex Youden plot.

1. For a given factor or interaction term, compute the mean of the response variable for the low level of the factor and for the high level of the factor. Any center points are omitted from the computation.
2. Plot the point where the *y*-coordinate is the mean for the high level of the factor and the *x*-coordinate is the mean for the low level of the factor. The character used for the plot point should identify the factor or interaction term (e.g., "1" for factor 1, "13" for the interaction between factors 1 and 3).
3. Repeat steps 1 and 2 for each factor and interaction term of the data.

The high and low values of the interaction terms are obtained by multiplying the corresponding values of the main level factors. For example, the interaction term X_{13} is obtained by multiplying the values for X_1 with the corresponding values of X_3. Since the values for X_1 and X_3 are either "-1" or "+1", the resulting values for X_{13} are also either

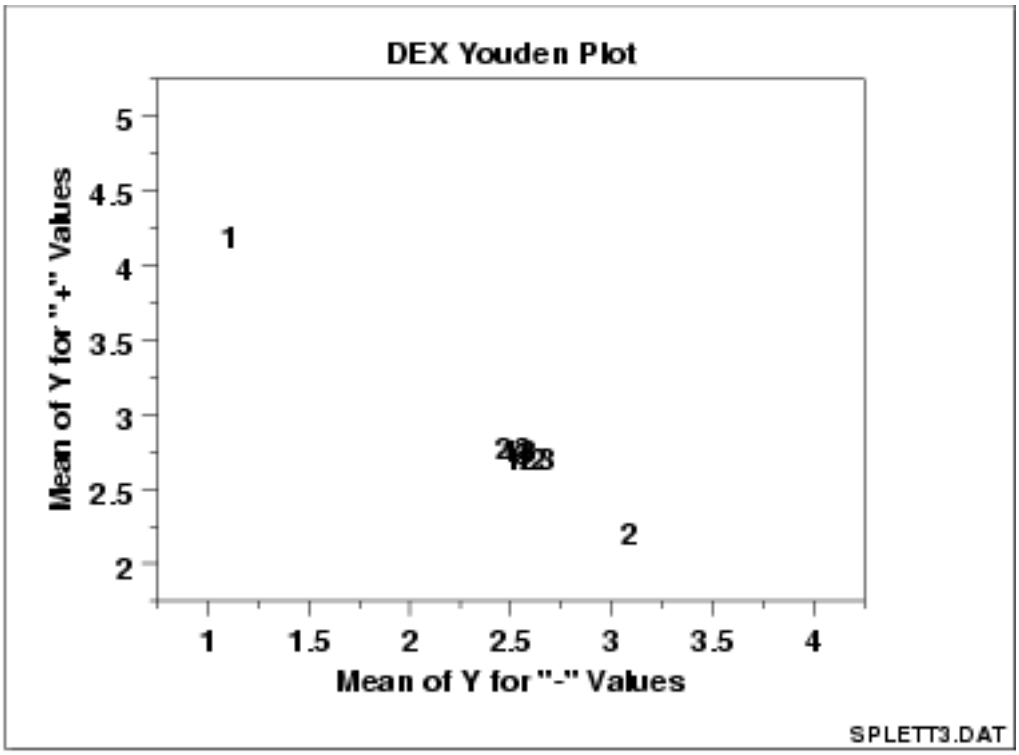
"-1" or "+1".

In summary, the dex Youden plot is a plot of the mean of the response variable for the high level of a factor or interaction term against the mean of the response variable for the low level of that factor or interaction term.

For unimportant factors and interaction terms, these mean values should be nearly the same. For important factors and interaction terms, these mean values should be quite different. So the interpretation of the plot is that unimportant factors should be clustered together near the grand mean. Points that stand apart from this cluster identify important factors that should be included in the model.

Sample DEX Youden Plot

The following is a dex Youden plot for the data used in the Eddy current case study. The analysis in that case study demonstrated that X1 and X2 were the most important factors.



Interpretation of the Sample DEX Youden Plot

From the above dex Youden plot, we see that factors 1 and 2 stand out from the others. That is, the mean response values for the low and high levels of factor 1 and factor 2 are quite different. For factor 3 and the 2 and 3-term interactions, the mean response values for the low and high levels are similar.

We would conclude from this plot that factors 1 and 2 are important and should be included in our final model while the remaining factors and interactions should be omitted from the final model.

Case Study The Eddy current case study demonstrates the use of the dex Youden plot in the context of the analysis of a full factorial design.

Software DEX Youden plots are not typically available as built-in plots in statistical software programs. However, it should be relatively straightforward to write a macro to generate this plot in most general purpose statistical software programs.

1.3.3.32. 4-Plot

Purpose: Check Underlying Statistical Assumptions

The 4-plot is a collection of 4 specific EDA graphical techniques whose purpose is to test the assumptions that underlie most measurement processes. A 4-plot consists of a

1. run sequence plot;
2. lag plot;
3. histogram;
4. normal probability plot.

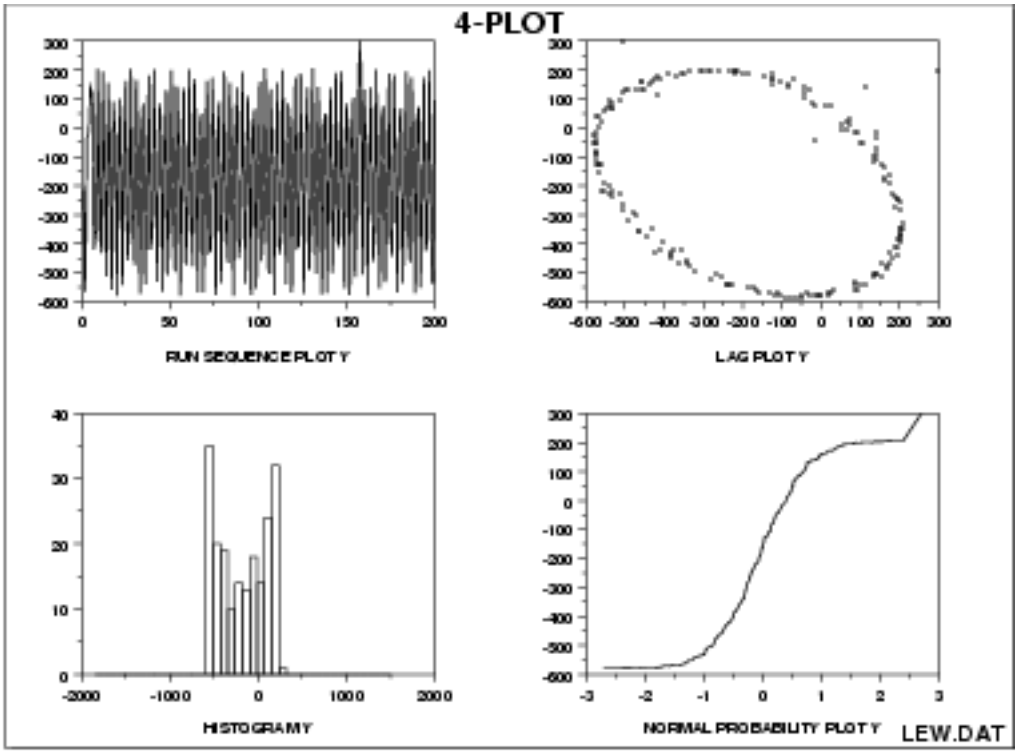
If the 4 underlying assumptions of a typical measurement process hold, then the above 4 plots will have a characteristic appearance (see the normal random numbers case study below); if any of the underlying assumptions fail to hold, then it will be revealed by an anomalous appearance in one or more of the plots. Several commonly encountered situations are demonstrated in the case studies below.

Although the 4-plot has an obvious use for univariate and time series data, its usefulness extends far beyond that. Many statistical models of the form

$$Y_i = f(X_1, ..., X_k) + E_i$$

have the same underlying assumptions for the error term. That is, no matter how complicated the functional fit, the assumptions on the underlying error term are still the same. The 4-plot can and should be routinely applied to the residuals when fitting models regardless of whether the model is simple or complicated.

Sample Plot: Process Has Fixed Location, Fixed Variation, Non-Random (Oscillatory), Non-Normal U-Shaped Distribution, and Has 3 Outliers.



This 4-plot reveals the following:

1. the fixed location assumption is justified as shown by the run sequence plot in the upper left corner.
2. the fixed variation assumption is justified as shown by the run sequence plot in the upper left corner.
3. the randomness assumption is violated as shown by the non-random (oscillatory) lag plot in the upper right corner.
4. the assumption of a common, normal distribution is violated as shown by the histogram in the lower left corner and the normal probability plot in the lower right corner. The distribution is non-normal and is a U-shaped distribution.
5. there are several outliers apparent in the lag plot in the upper right corner.

Definition: 1. Run Sequence Plot; 2. Lag Plot; 3. Histogram; 4. Normal Probability Plot

The 4-plot consists of the following:

1. Run sequence plot to test fixed location and variation.
   - Vertically: Y_i
   - Horizontally: i
2. Lag Plot to test randomness.
   - Vertically: Y_i
   - Horizontally: Y_{i-1}
3. Histogram to test (normal) distribution.
   - Vertically: Counts
   - Horizontally: Y
4. Normal probability plot to test normal distribution.
   - Vertically: Ordered Y_i
   - Horizontally: Theoretical values from a normal N(0,1) distribution for ordered Y_i

Questions

4-plots can provide answers to many questions:

1. Is the process in-control, stable, and predictable?
2. Is the process drifting with respect to location?
3. Is the process drifting with respect to variation?
4. Are the data random?
5. Is an observation related to an adjacent observation?
6. If the data are a time series, is is white noise?
7. If the data are a time series and not white noise, is it sinusoidal, autoregressive, etc.?
8. If the data are non-random, what is a better model?
9. Does the process follow a normal distribution?
10. If non-normal, what distribution does the process follow?
11. Is the model

    $$Y_i = A_0 + E_i$$

    valid and sufficient?
12. If the default model is insufficient, what is a better model?
13. Is the formula $s_{\bar{Y}} = s/\sqrt{N}$ valid?
14. Is the sample mean a good estimator of the process location?
15. If not, what would be a better estimator?
16. Are there any outliers?

Importance: Testing Underlying Assumptions Helps Ensure the Validity of the Final Scientific and Engineering Conclusions

There are 4 assumptions that typically underlie all measurement processes; namely, that the data from the process at hand "behave like":

1. random drawings;
2. from a fixed distribution;
3. with that distribution having a fixed location; and
4. with that distribution having fixed variation.

Predictability is an all-important goal in science and engineering. If the above 4 assumptions hold, then we have achieved probabilistic predictability--the ability to make probability statements not only about the process in the past, but also about the process in the future. In short, such processes are said to be "statistically in control". If the 4 assumptions do not hold, then we have a process that is drifting (with respect to location, variation, or distribution), is unpredictable, and is out of control. A simple characterization of such processes by a location estimate, a variation estimate, or a distribution "estimate" inevitably leads to optimistic and grossly invalid engineering conclusions.

Inasmuch as the validity of the final scientific and engineering conclusions is inextricably linked to the validity of these same 4 underlying assumptions, it naturally follows that there is a real necessity for all 4 assumptions to be routinely tested. The 4-plot (run sequence plot, lag plot, histogram, and normal probability plot) is seen as a simple, efficient, and powerful way of carrying out this routine checking.

Interpretation: Flat, Equi-Banded, Random, Bell-Shaped, and Linear

Of the 4 underlying assumptions:

1. If the fixed location assumption holds, then the run sequence plot will be flat and non-drifting.
2. If the fixed variation assumption holds, then the vertical spread in the run sequence plot will be approximately the same over the entire horizontal axis.
3. If the randomness assumption holds, then the lag plot will be structureless and random.
4. If the fixed distribution assumption holds (in particular, if the fixed normal distribution assumption holds), then the histogram will be bell-shaped and the normal probability plot will be approximatelylinear.

If all 4 of the assumptions hold, then the process is "statistically in control". In practice, many processes fall short of achieving this ideal.

Related Techniques

Run Sequence Plot
Lag Plot
Histogram
Normal Probability Plot

Autocorrelation Plot
Spectral Plot
PPCC Plot

Case Studies

The 4-plot is used in most of the case studies in this chapter:

1. Normal random numbers (the ideal)
2. Uniform random numbers
3. Random walk
4. Josephson junction cryothermometry
5. Beam deflections
6. Filter transmittance
7. Standard resistor
8. Heat flow meter 1

Software

It should be feasible to write a macro for the 4-plot in any general purpose statistical software program that supports the capability for multiple plots per page and supports the underlying plot techniques. Dataplot supports the 4-plot.

1.3.3.33. 6-Plot

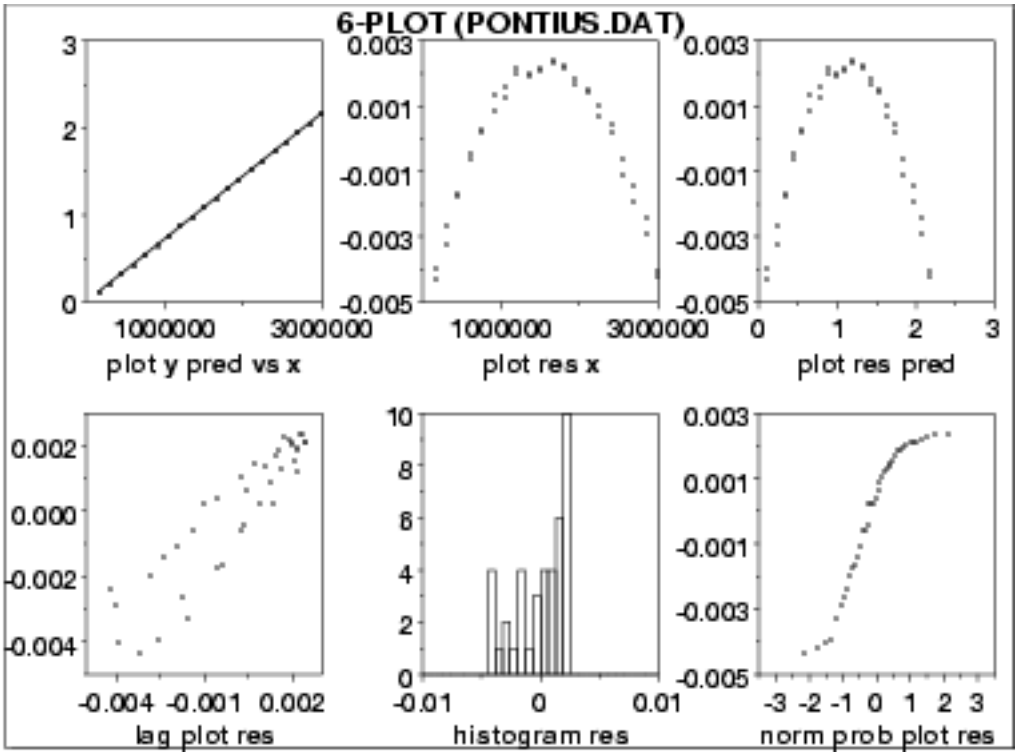
Purpose: Graphical Model Validation

The 6-plot is a collection of 6 specific graphical techniques whose purpose is to assess the validity of a Y versus X fit. The fit can be a linear fit, a non-linear fit, a LOWESS (locally weighted least squares) fit, a spline fit, or any other fit utilizing a single independent variable.

The 6 plots are:

1. Scatter plot of the response and predicted values versus the independent variable;
2. Scatter plot of the residuals versus the independent variable;
3. Scatter plot of the residuals versus the predicted values;
4. Lag plot of the residuals;
5. Histogram of the residuals;
6. Normal probability plot of the residuals.

Sample Plot



This 6-plot, which followed a linear fit, shows that the linear model is not adequate. It suggests that a quadratic model would be a better model.

Definition: 6 Component Plots

The 6-plot consists of the following:

1. Response and predicted values
    - Vertical axis: Response variable, predicted values
    - Horizontal axis: Independent variable
2. Residuals versus independent variable
    - Vertical axis: Residuals
    - Horizontal axis: Independent variable
3. Residuals versus predicted values
    - Vertical axis: Residuals
    - Horizontal axis: Predicted values
4. Lag plot of residuals
    - Vertical axis: RES(I)
    - Horizontal axis: RES(I-1)
5. Histogram of residuals
    - Vertical axis: Counts
    - Horizontal axis: Residual values
6. Normal probability plot of residuals
    - Vertical axis: Ordered residuals
    - Horizontal axis: Theoretical values from a normal N(0,1)

distribution for ordered residuals

Questions

The 6-plot can be used to answer the following questions:

1. Are the residuals approximately normally distributed with a fixed location and scale?
2. Are there outliers?
3. Is the fit adequate?
4. Do the residuals suggest a better fit?

Importance: Validating Model

A model involving a response variable and a single independent variable has the form:

$$Y_i = f(X_i) + E_i$$

where Y is the response variable, X is the independent variable, *f* is the linear or non-linear fit function, and E is the random component. For a good model, the error component should behave like:

1. random drawings (i.e., independent);
2. from a fixed distribution;
3. with fixed location; and
4. with fixed variation.

In addition, for fitting models it is usually further assumed that the fixed distribution is normal and the fixed location is zero. For a good model the fixed variation should be as small as possible. A necessary component of fitting models is to verify these assumptions for the error component and to assess whether the variation for the error component is sufficiently small. The histogram, lag plot, and normal probability plot are used to verify the fixed distribution, location, and variation assumptions on the error component. The plot of the response variable and the predicted values versus the independent variable is used to assess whether the variation is sufficiently small. The plots of the residuals versus the independent variable and the predicted values is used to assess the independence assumption.

Assessing the validity and quality of the fit in terms of the above assumptions is an absolutely vital part of the model-fitting process. No fit should be considered complete without an adequate model validation step.

Related Techniques

Linear Least Squares
Non-Linear Least Squares
Scatter Plot
Run Sequence Plot
Lag Plot
Normal Probability Plot
Histogram

Case Study

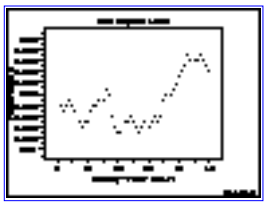
The 6-plot is used in the Alaska pipeline data case study.

Software

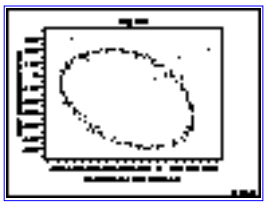
It should be feasible to write a macro for the 6-plot in any general purpose statistical software program that supports the capability for multiple plots per page and supports the underlying plot techniques. Dataplot supports the 6-plot.

1.3.4. Graphical Techniques: By Problem Category

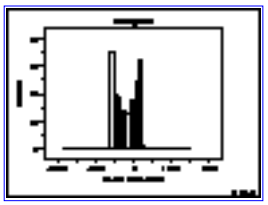
Univariate
y = c + e



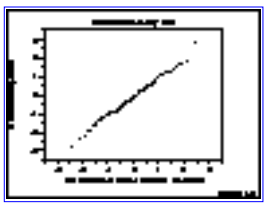
Run Sequence Plot: 1.3.3.25



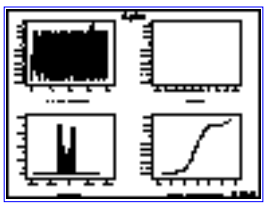
Lag Plot: 1.3.3.15



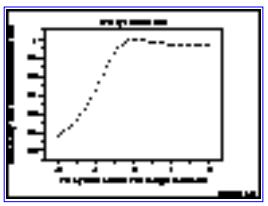
Histogram: 1.3.3.14



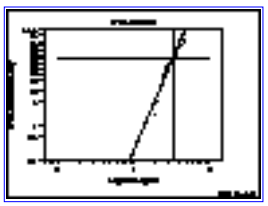
Normal Probability Plot: 1.3.3.21



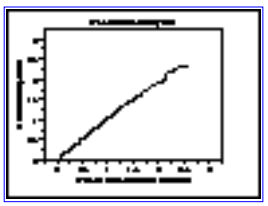
4-Plot: 1.3.3.32



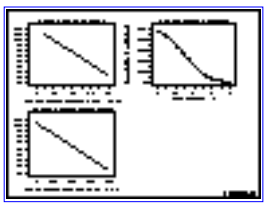
PPCC Plot: 1.3.3.23



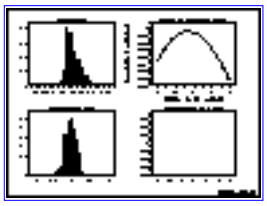
Weibull Plot: 1.3.3.30



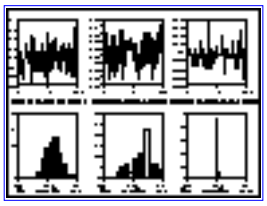
Probability Plot: 1.3.3.22



Box-Cox Linearity Plot: 1.3.3.5

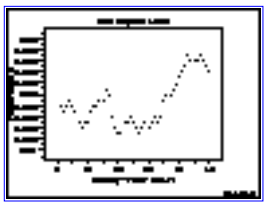


Box-Cox Normality Plot: 1.3.3.6

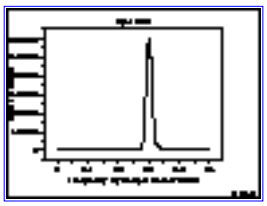


Bootstrap Plot: 1.3.3.4

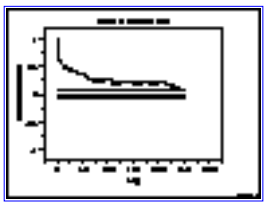
Time Series
y = f(t) + e



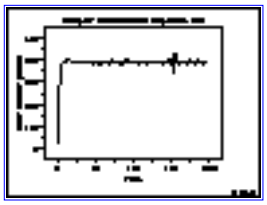
Run Sequence Plot: 1.3.3.25



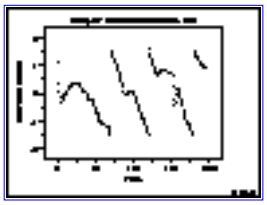
Spectral Plot: 1.3.3.27



Autocorrelation Plot: 1.3.3.1



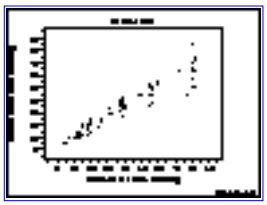
Complex Demodulation Amplitude Plot: 1.3.3.8



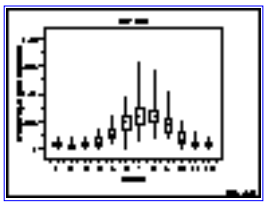
Complex Demodulation Phase Plot: 1.3.3.9

---

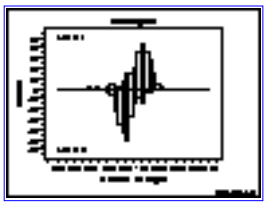
1 Factor
y = f(x) + e



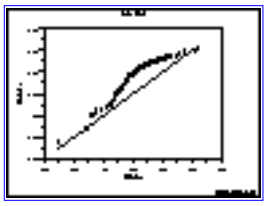
Scatter Plot: 1.3.3.26



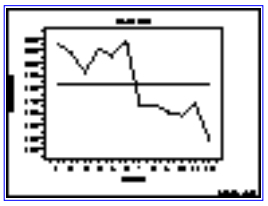
Box Plot: 1.3.3.7



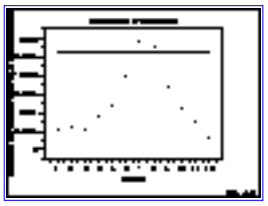
Bihistogram: 1.3.3.2



[Quantile-Quantile Plot: 1.3.3.24]

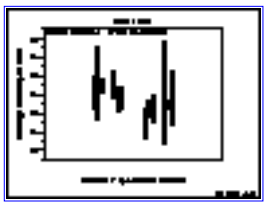


[Mean Plot: 1.3.3.20]



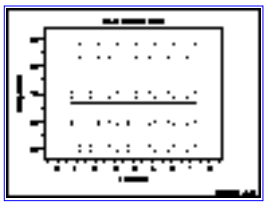
[Standard Deviation Plot: 1.3.3.28]

Multi-Factor/Comparative
y = f(xp, x1,x2,...,xk) + e

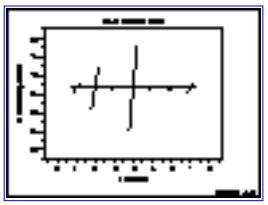


[Block Plot: 1.3.3.3]

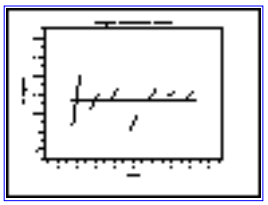
Multi-Factor/Screening
y = f(x1,x2,x3,...,xk) + e



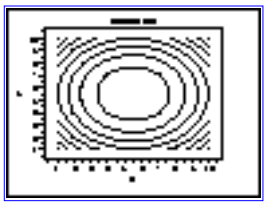
[DEX Scatter Plot: 1.3.3.11]



[DEX Mean Plot: 1.3.3.12]

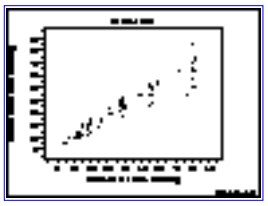


[DEX Standard Deviation Plot: 1.3.3.13]

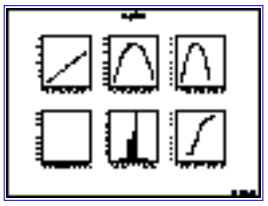


[Contour Plot: 1.3.3.10]

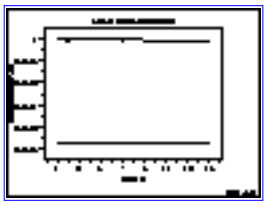
Regression
y = f(x1,x2,x3,...,xk) + e



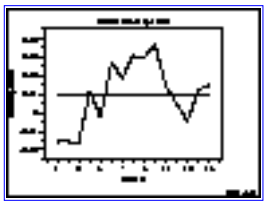
Scatter Plot: 1.3.3.26



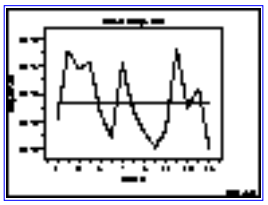
6-Plot: 1.3.3.33



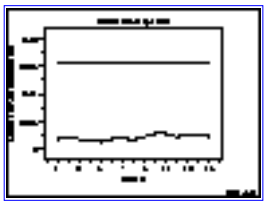
Linear Correlation Plot: 1.3.3.16



Linear Intercept Plot: 1.3.3.17

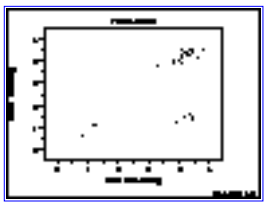


Linear Slope Plot: 1.3.3.18



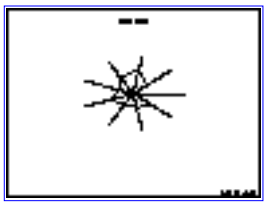
Linear Residual Standard Deviation Plot:1.3.3.19

Interlab
(y1,y2) = f(x) + e



Youden Plot: 1.3.3.31

Multivariate
(y1,y2,...,yp)



Star Plot: 1.3.3.29

1.3.5. Quantitative Techniques

Confirmatory Statistics

The techniques discussed in this section are classical statistical methods as opposed to EDA techniques. EDA and classical techniques are not mutually exclusive and can be used in a complamentary fashion. For example, the analysis can start with some simple graphical techniques such as the 4-plot followed by the classical confirmatory methods discussed herein to provide more rigorous statments about the conclusions. If the classical methods yield different conclusions than the graphical analysis, then some effort should be invested to explain why. Often this is an indication that some of the assumptions of the classical techniques are violated.

Many of the quantitative techniques fall into two broad categories:

1. Interval estimation
2. Hypothesis tests

Interval Estimates

It is common in statistics to estimate a parameter from a sample of data. The value of the parameter using all of the possible data, not just the sample data, is called the population parameter or true value of the parameter. An estimate of the true parameter value is made using the sample data. This is called a point estimate or a sample estimate.

For example, the most commonly used measure of location is the mean. The population, or true, mean is the sum of all the members of the given population divided by the number of members in the population. As it is typically impractical to measure every member of the population, a random sample is drawn from the population. The sample mean is calculated by summing the values in the sample and dividing by the number of values in the sample. This sample mean is then used as the point estimate of the population mean.

Interval estimates expand on point estimates by incorporating the uncertainty of the point estimate. In the example for the mean above, different samples from the same population will generate different values for the sample mean. An interval estimate quantifies this uncertainty in the sample estimate by computing lower and upper