What is EDA?

Approach

Exploratory Data Analysis (EDA) is an approach or 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
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 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.

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 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.

How Does Exploratory Data Analysis Differ from Classical Data Analysis?

Data Analysis Approaches

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)

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

Method of dealing with the underlying model for the data distinguishes the 2 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.

In the real world, data analysts freely mix elements of the above approaches (and other approaches).

Model

Classical

The classical approach imposes models (both deterministic[1] and probabilistic[2]) 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

[1] An inevitable consequence of antecedent sufficient causes
[2] Based on, or affected by probability, randomness, or chance

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.

Model Comparison

1. Focus
    a. 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.
    b. Exploratory - For exploratory data analysis, the focus is on the data--its structure, outliers, and models suggested by the data.
2. Techniques
    a. Classical - Classical techniques are generally quantitative in nature. They include ANOVA, t tests, chi-squared tests, and F tests.
    b. Exploratory - EDA techniques are generally graphical. They include scatter plots, character plots, box plots, histograms, bi-histograms, probability plots, residual plots, and mean plots.
3. Rigor
    a. 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".

b. 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.

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.

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 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.

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 and 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.

What are the EDA Goals?

Primary 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[3] 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;

[3] Excessively sparing or frugal

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.

The Role of Graphics

Quantitative/Graphical

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

1. Quantitative
2. Graphical

Quantitative

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

1. hypothesis testing
2. analysis of variance
3. point estimates and confidence intervals
4. 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:

1. scatter plots
2. histograms
3. probability plots
4. residual plots
5. box plots
6. 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:

1. testing assumptions
2. model selection
3. model validation
4. estimator selection
5. relationship identification
6. factor effect determination
7. 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.

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.

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.

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
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 follows the “response = deterministic component + random component” which boils down to “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:

1. the data are uncorrelated with one another
2. the random component has a fixed distribution
3. the deterministic component consists of only a constant
4. 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 behave like:

1. random drawings
2. from a fixed distribution
3. with fixed location (namely, 0 in this case)
4. 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.

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.

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

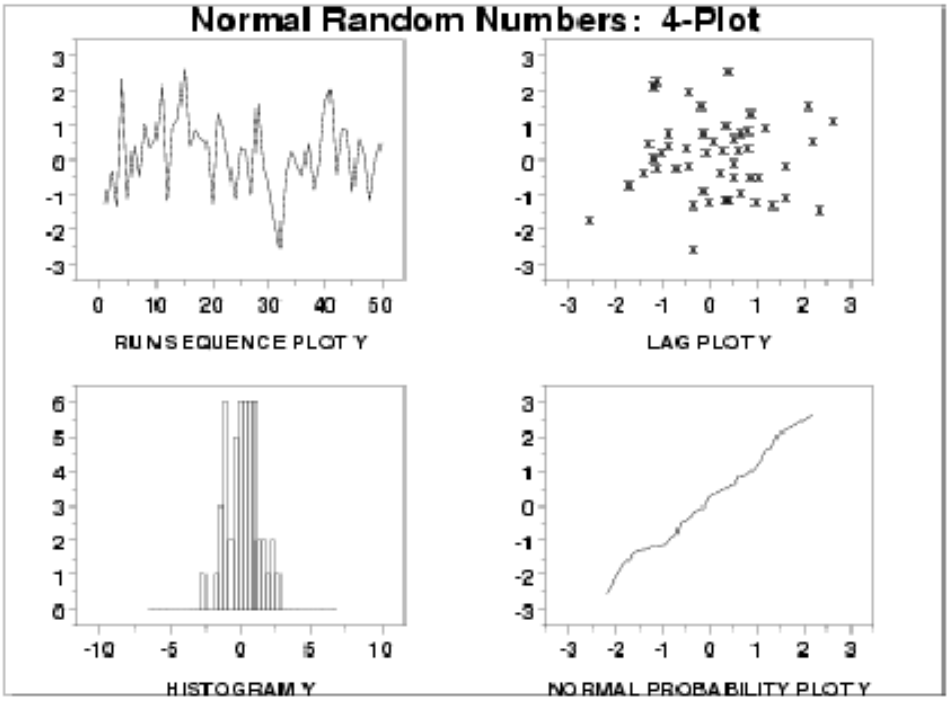
The following EDA techniques are simple, efficient, and powerful for the routine testing of underlying assumptions:

1. run sequence plot (Y_i 1. 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



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

Consequences of Non-Randomness

Randomness Assumption

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

Consequences 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.

Consequences of Non-Fixed Location Parameter

Location Estimate

The usual estimate of location is the mean 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:

a. may be invalid and the numerical value optimistically small
b. the location estimate may be poor
c. the location estimate may be biased

Consequences of Non-Fixed Variation Parameter

Variation Estimate

The usual estimate of variation is the standard deviation from N measurements Y_1, Y_2, ... , 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

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 mid-mean, 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.

Other consequences that flow from problems with distributional assumptions are:

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

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.

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 quantitative 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.

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.

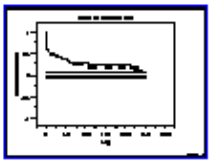
EDA techniques are tied in with each of the questions. There are some EDA techniques (e.g., the scatter plot) that are broad-brushed and apply almost universally. On the other hand, there are a large number of EDA techniques that are specific and whose specificity is tied in with one of the above questions. Clearly if one chooses not to explicitly identify relevant questions, then one cannot take advantage of these question-specific EDA techniques.

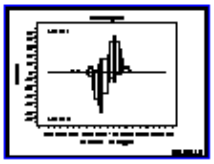
EDA Approach Emphasizes Graphics

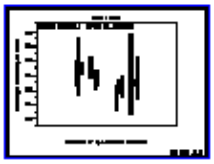
Most of these questions can be addressed by techniques discussed in this chapter. The process modeling and process improvement chapters also address many of the questions above. These questions are also relevant for the classical approach to statistics. What distinguishes the EDA approach is an emphasis 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.

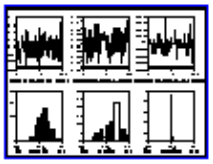
Graphical Techniques: Alphabetic

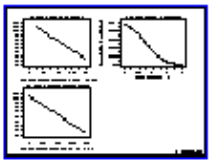
This section provides a gallery of some useful graphical techniques. The techniques are ordered alphabetically, so this section is not intended to be read in a sequential fashion. The use of most of these graphical techniques is demonstrated in the case studies in this chapter. A few of these graphical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

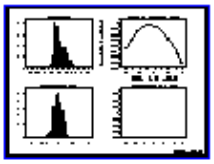

Autocorrelation Plot: 1.3.3.1

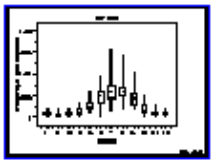

Bihistogram: 1.3.3.2

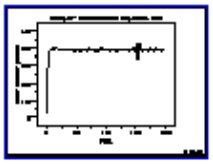

Block Plot: 1.3.3.3

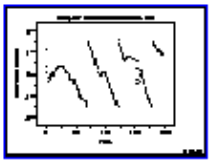

Bootstrap Plot: 1.3.3.4

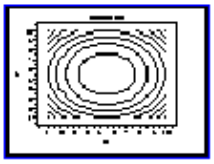

Box-Cox Linearity Plot: 1.3.3.5

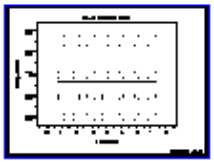

Box-Cox Normality Plot: 1.3.3.6

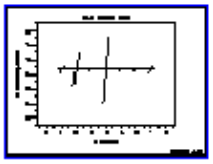

Box Plot: 1.3.3.7

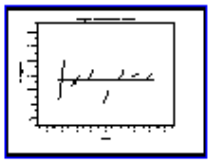

Complex Demodulation Amplitude Plot: 1.3.3.8

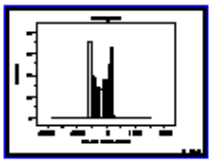

Complex Demodulation Phase Plot: 1.3.3.9

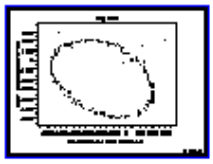

Contour Plot: 1.3.3.10

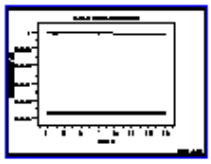

DEX Scatter Plot: 1.3.3.11

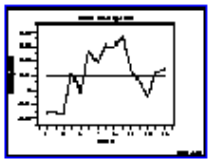

DEX Mean Plot: 1.3.3.12

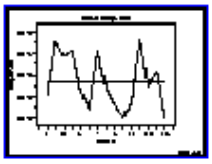

DEX Standard Deviation Plot: 1.3.3.13

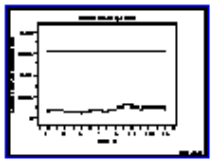

Histogram: 1.3.3.14

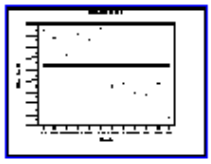

Lag Plot: 1.3.3.15

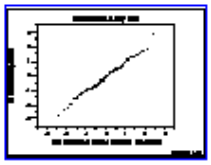

Linear Correlation Plot: 1.3.3.16

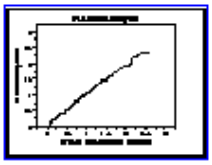

Linear Intercept Plot: 1.3.3.17

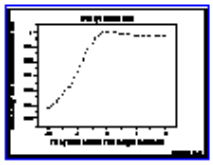

Linear Slope Plot: 1.3.3.18

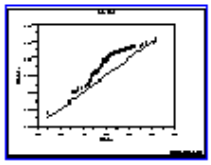

Linear Residual Standard Deviation Plot: 1.3.3.19

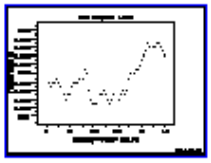

Mean Plot: 1.3.3.20

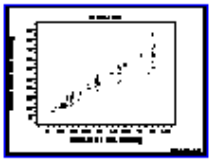

Normal Probability Plot: 1.3.3.21

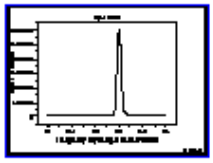

Probability Plot: 1.3.3.22

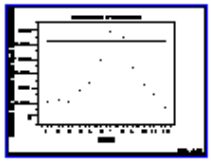

Probability Plot Correlation Coefficient Plot: 1.3.3.23


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

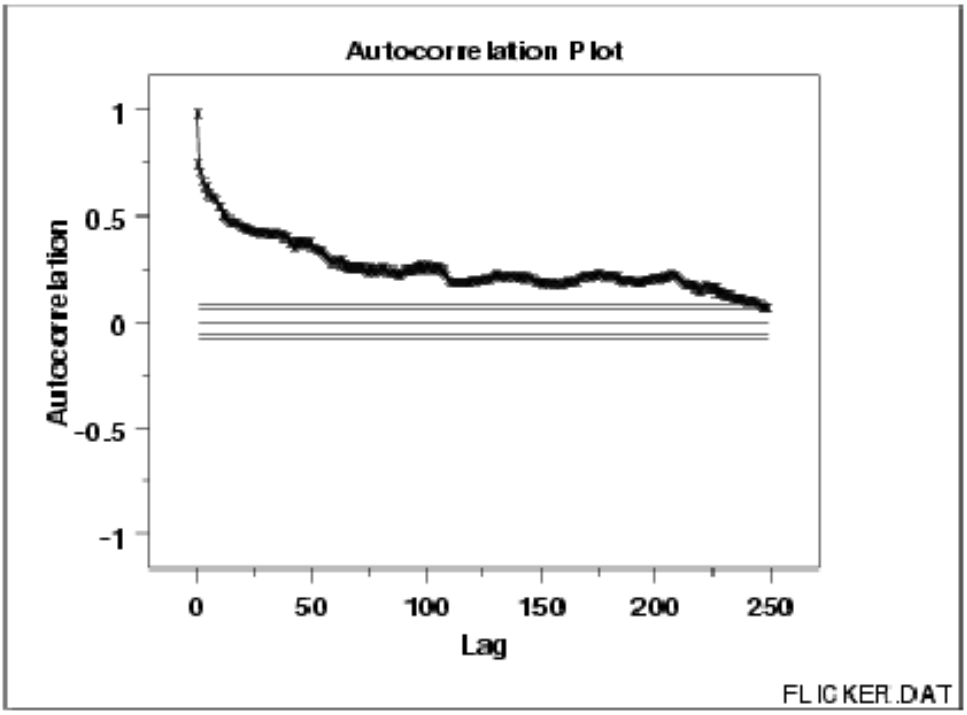
Autocorrelation Plot

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.

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?
4. Is the observed time series white noise?
5. Is the observed time series sinusoidal?
1. Is the observed time series autoregressive?
2. What is an appropriate model for the observed time series?
3. Is the model *Y* = constant + error valid and sufficient?
4. Is the formula 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 means: where *S* is the standard deviation of the data.

3. Although heavily used, the results from using this formula are of no value unless the randomness assumption holds.
4. 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.

Conclusions

We can make the following conclusions from this plot.

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

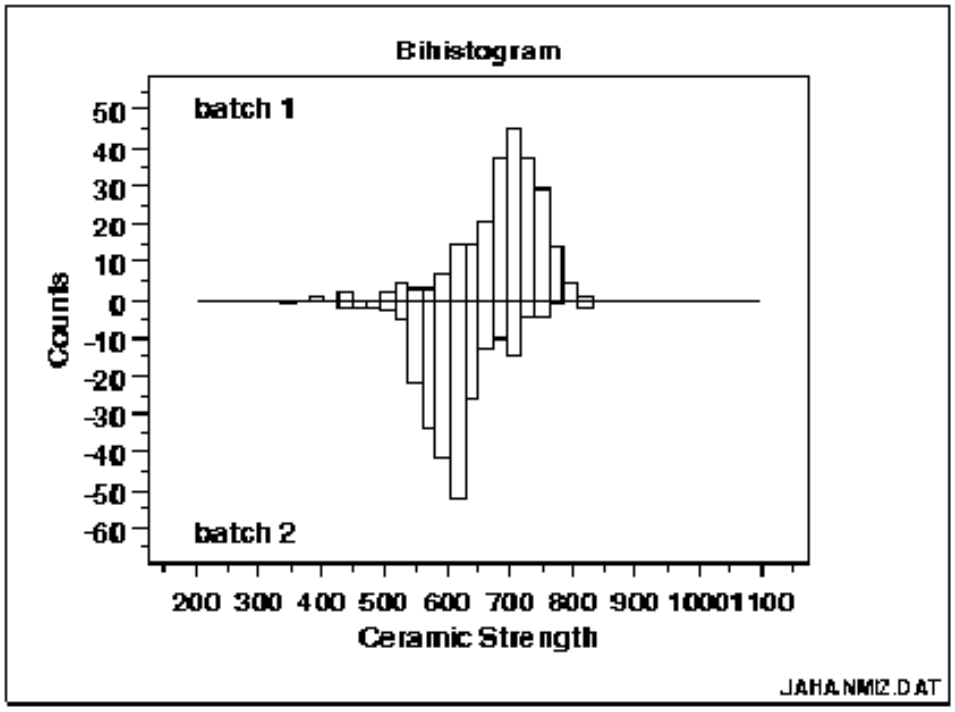
Bi-histogram

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

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

1. location
2. variation
3. distribution

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



Sample Plot: This bi-histogram reveals that there is a significant difference in ceramic breaking strength between batch 1(above) and batch 2 (below) from the above bi-histogram, 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 bi-histogram 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).

Questions

The bi-histogram 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 bi-histogram is an important EDA tool for determining if a factor "has an effect". Since the bi-histogram 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 bi-histogram 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.

Block Plot

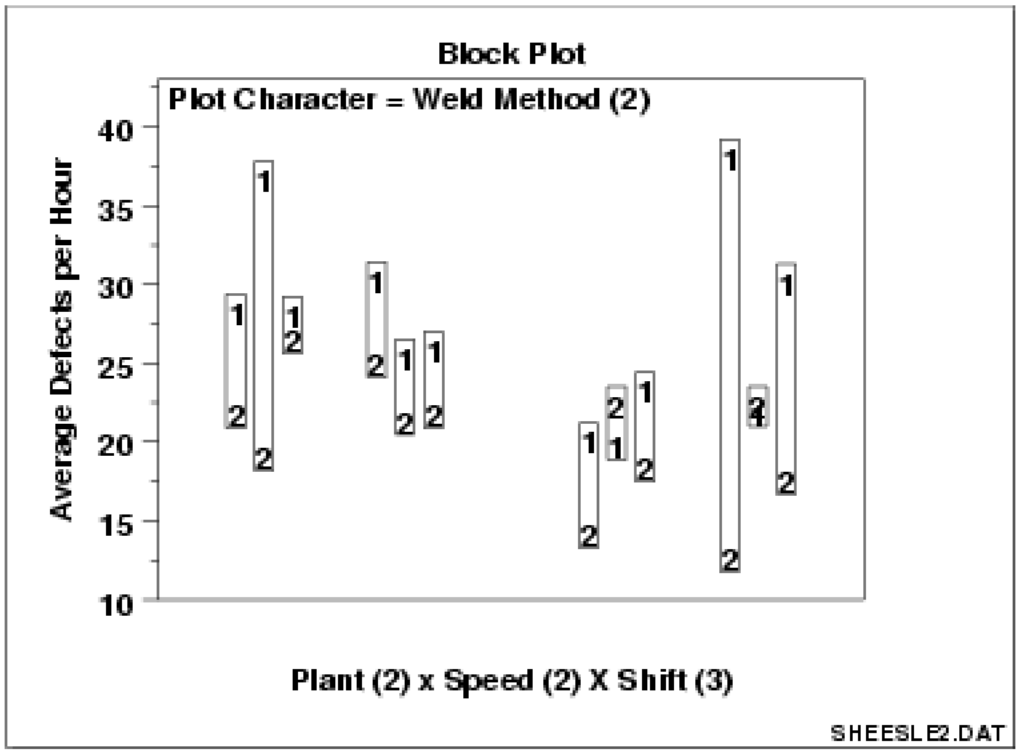
Purpose: Check to determine if a factor of interest has an effect robust over all other factors

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.



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: (shown in the plot above)

1. weld strength (2 levels)

2. plant (2 levels)
3. speed (2 levels)
4. shift (3 levels)

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:

1. The left 6 bars are from plant 1 and the right 6 bars are from plant 2.
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.
3. 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:

1. A quantitative procedure (analysis of variance) is replaced by a graphical procedure.
2. 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.

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

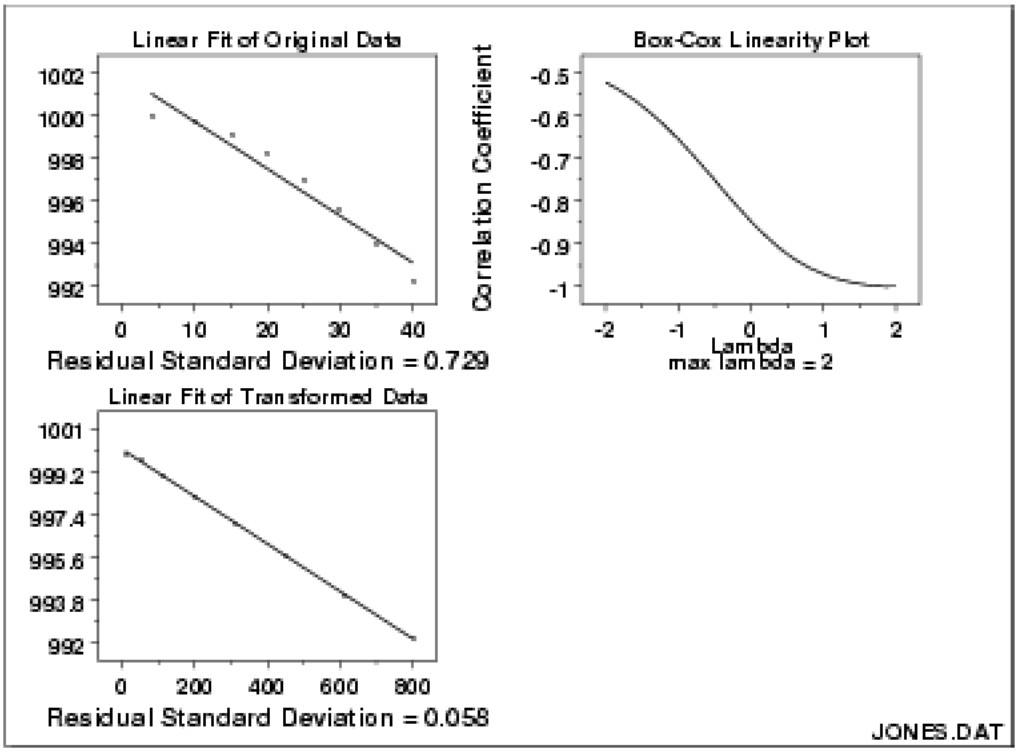
The Box-Cox linearity plot is a plot of the correlation between Y and the transformed X for given values of. That is, is the coordinate for the horizontal axis variable and the value of the correlation between Y

and the transformed X is the coordinate for the vertical axis of the plot. The value of corresponding to the maximum correlation (or minimum for negative correlation) on the plot is then the optimal choice for.

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.

The plot of the original data with the predicted values from a linear fit indicates that a quadratic fit might be preferable. The Box-Cox linearity plot shows a value of = 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).

Sample Plot



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

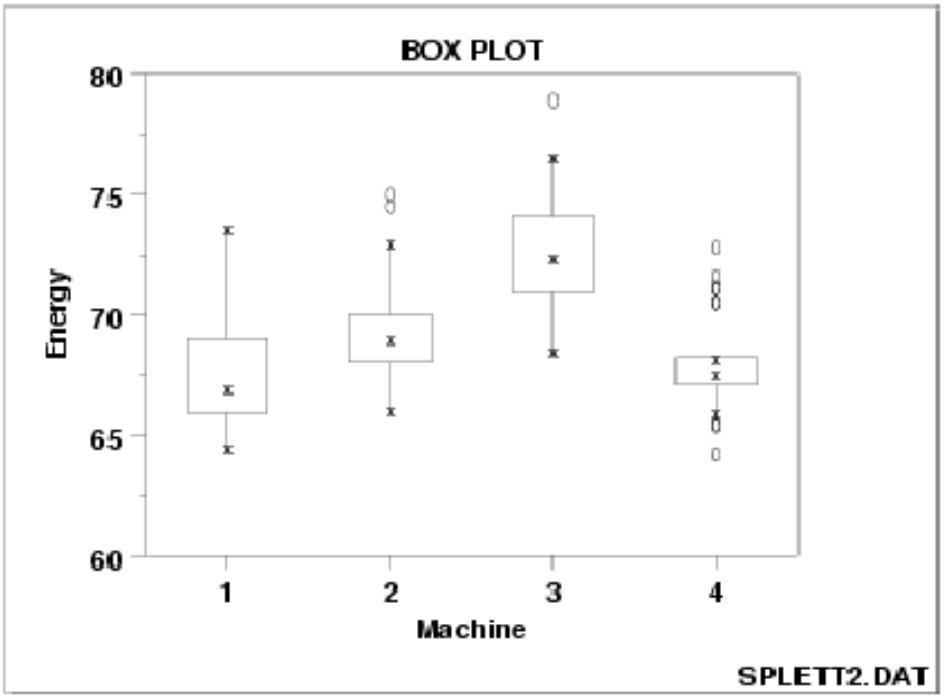
Transformations can often significantly improve a fit. The Box-Cox linearity plot provides a convenient way to find a suitable transformation without engaging in a lot of trial and error fitting.

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.

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).

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

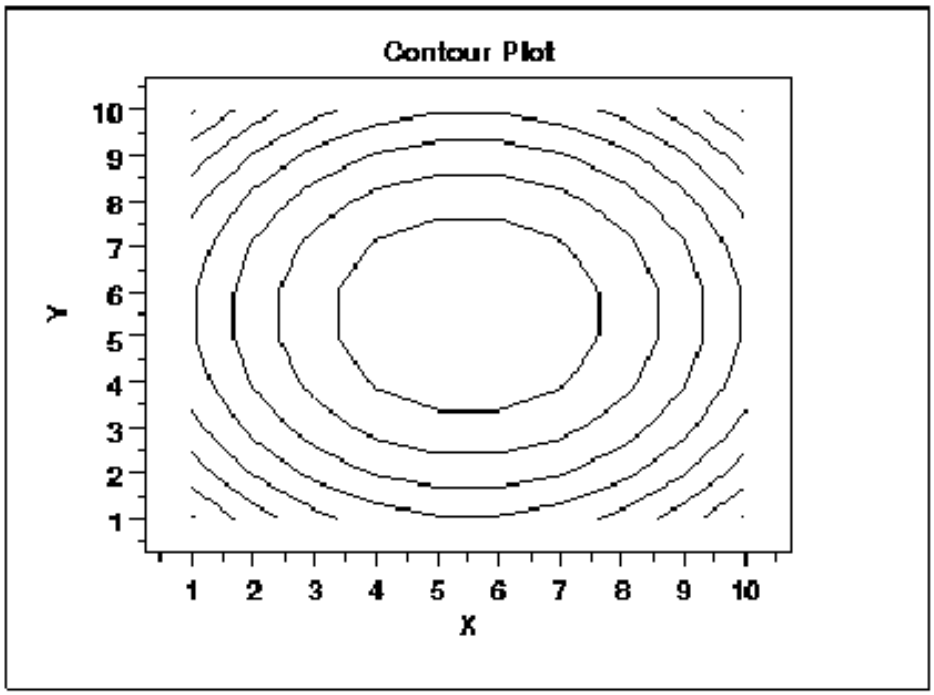
The box plot is an important EDA tool for determining if a factor has a significant effect on the response with respect to either location or variation. The box plot is also an effective tool for summarizing large quantities of information.

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.

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

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:

center (i.e., the location) of the data

spread (i.e., the scale) of the data

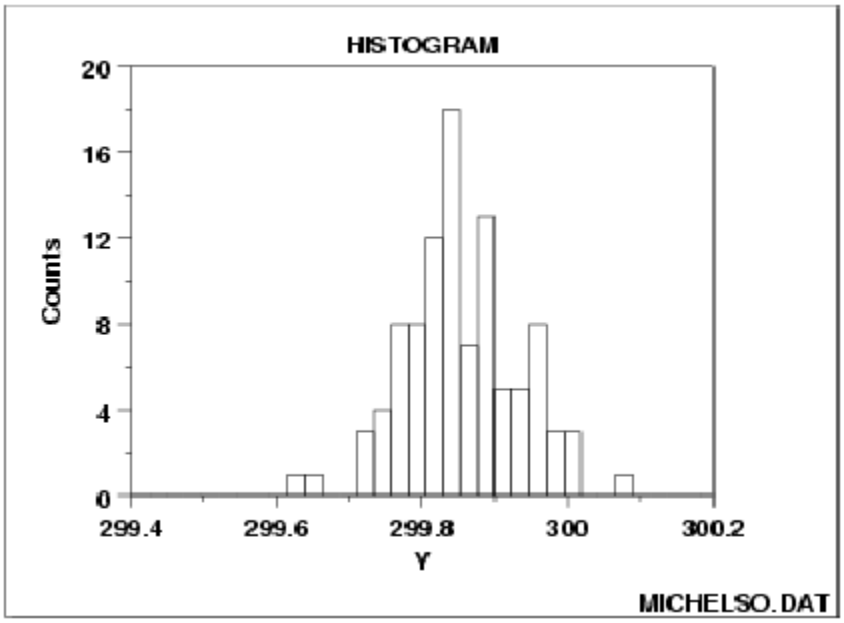
skewness of the data

presence of outliers

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



Questions

The histogram can be used to answer the following questions:

What kind of population distribution do the data come from?

Where are the data located?

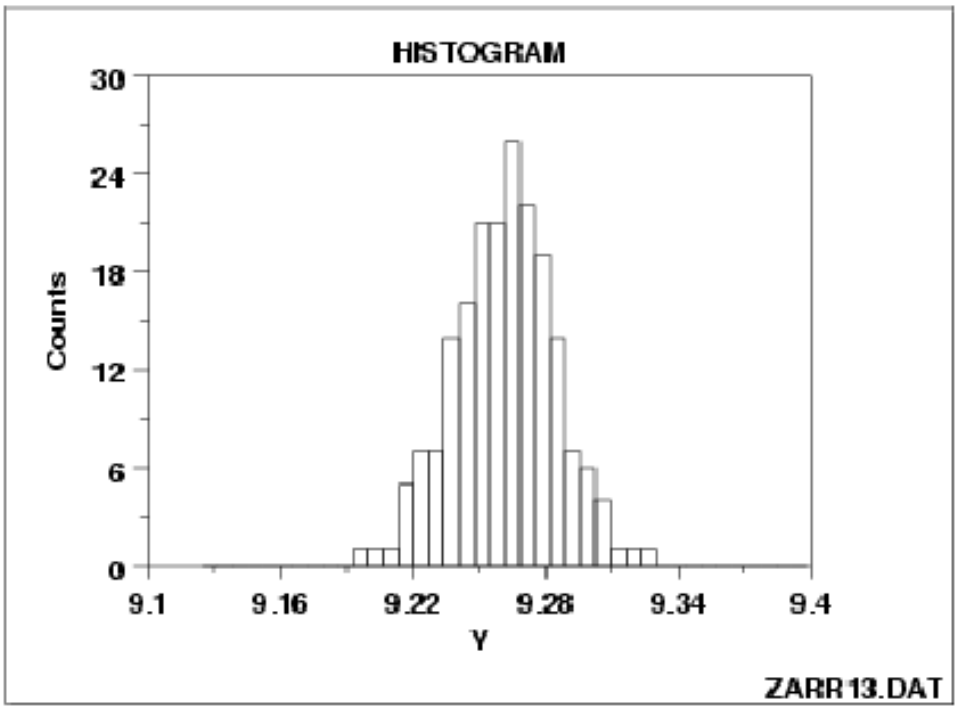
How spread out are the data?

Are the data symmetric or skewed?

Are there outliers in the data?

Histogram Interpretation: Normal

Symmetric, Moderate- Tailed Histogram



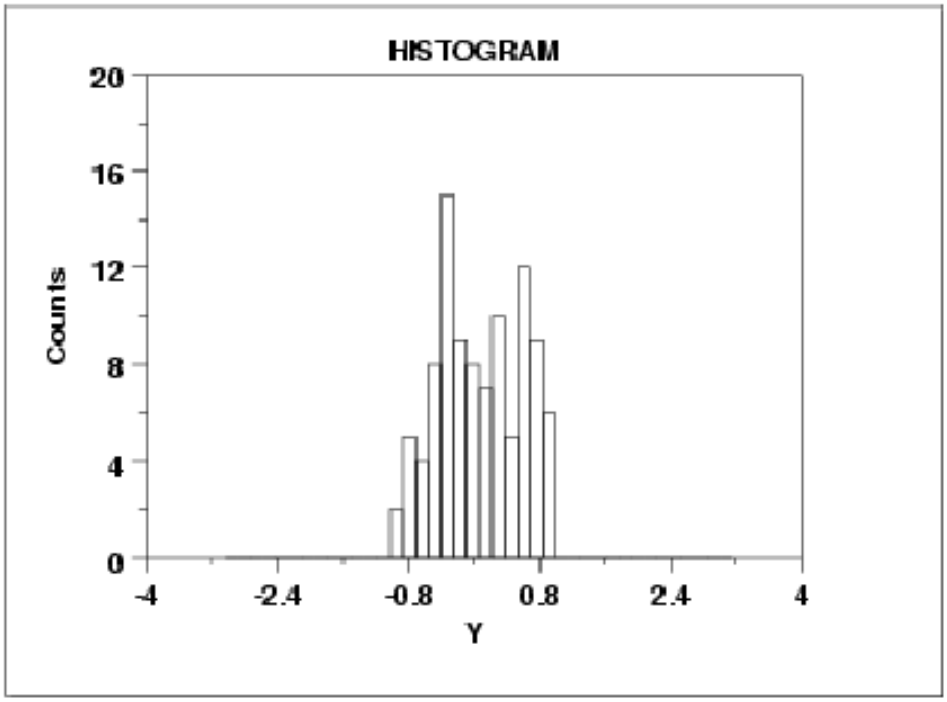
Note the classical bell-shaped, symmetric histogram with most of the frequency counts bunched in the middle and with the counts dying off out in the tails. From a physical science/engineering point of view, the normal distribution is that distribution which occurs most often in nature (due in part to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

Histogram Interpretation: Symmetric, Non-Normal, and 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 = (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.

Mean Plot

Purpose: Detect changes in location between groups

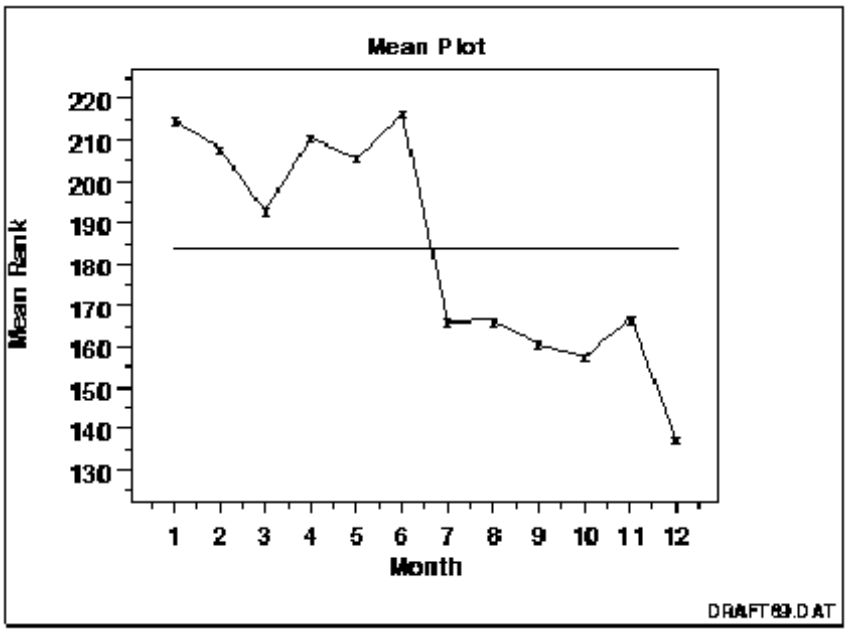
Mean plots are used to see if the mean varies between different groups of the data. The grouping is determined by the analyst. In most cases, the data set contains a specific grouping variable.

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.

Sample Plot



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.

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.

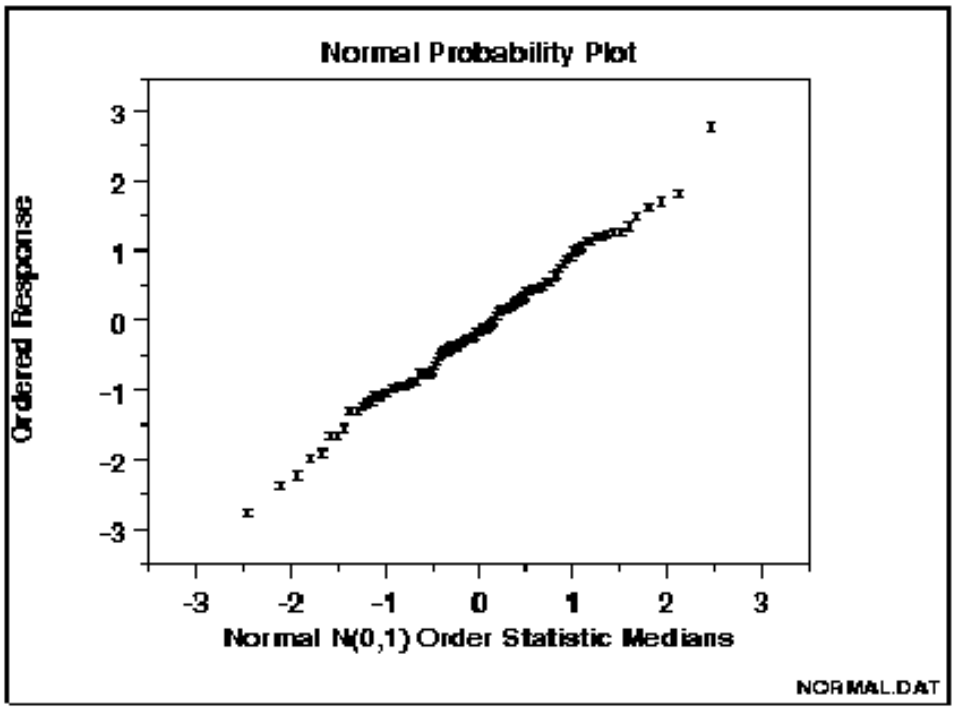
Normal Probability Plot

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.

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.

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

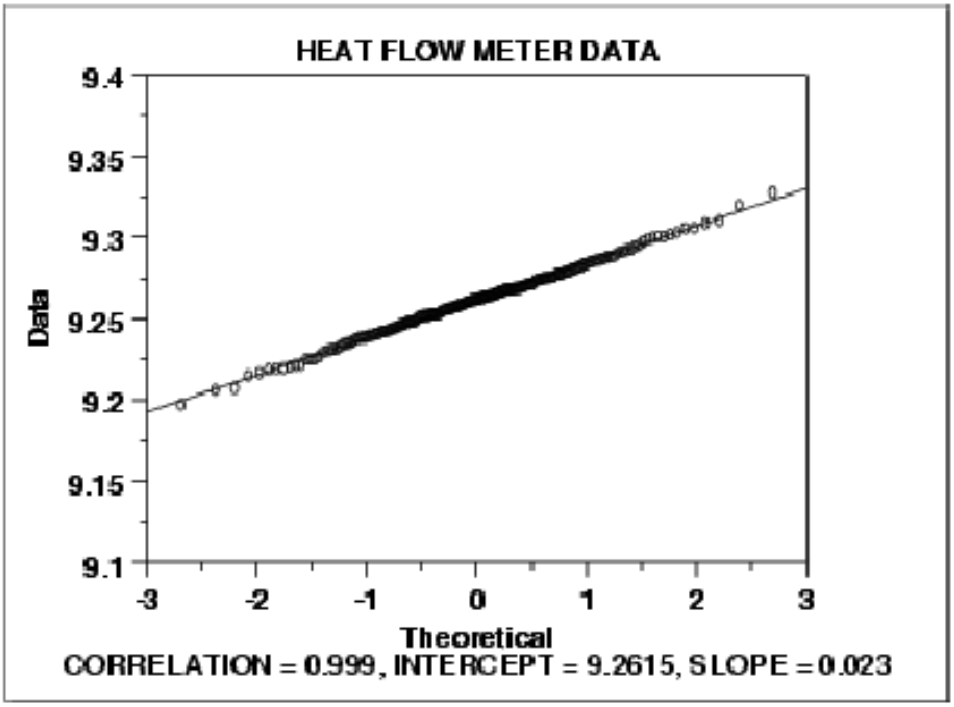
Probability plots are used to assess the assumption of a fixed distribution. In particular, most statistical models are of the form:

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.

Normal Probability Plot: Normally Distributed Data

Normal Probability Plot



Conclusions

We can make the following conclusions from the above plot. The normal probability plot shows a strongly linear pattern.

1. 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.

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:

random drawings

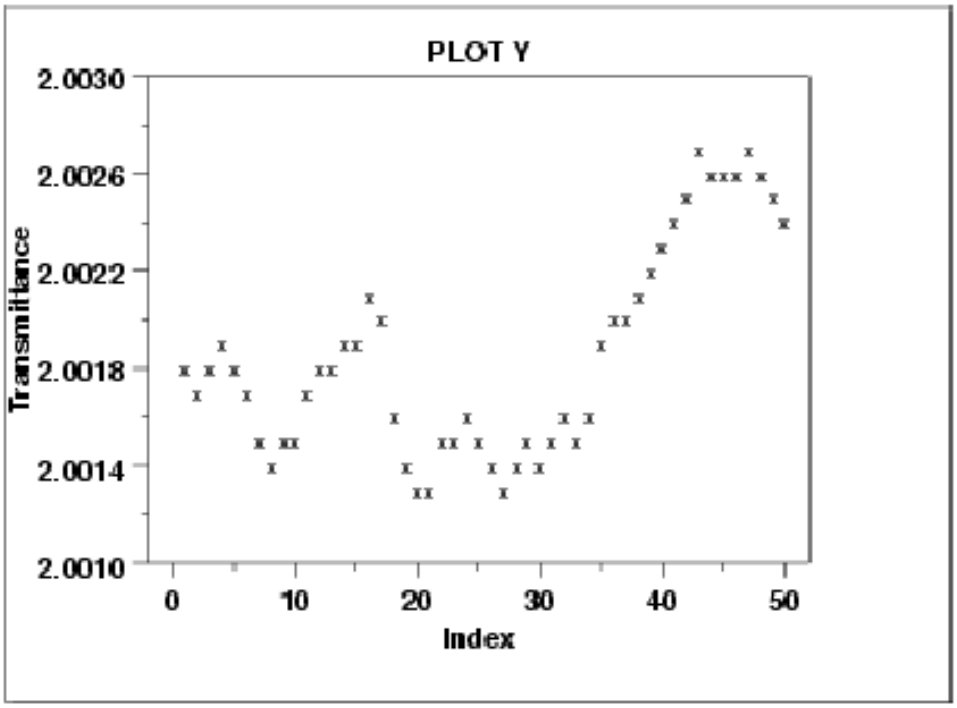
from a fixed distribution

with a common location

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.

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 = constant + error

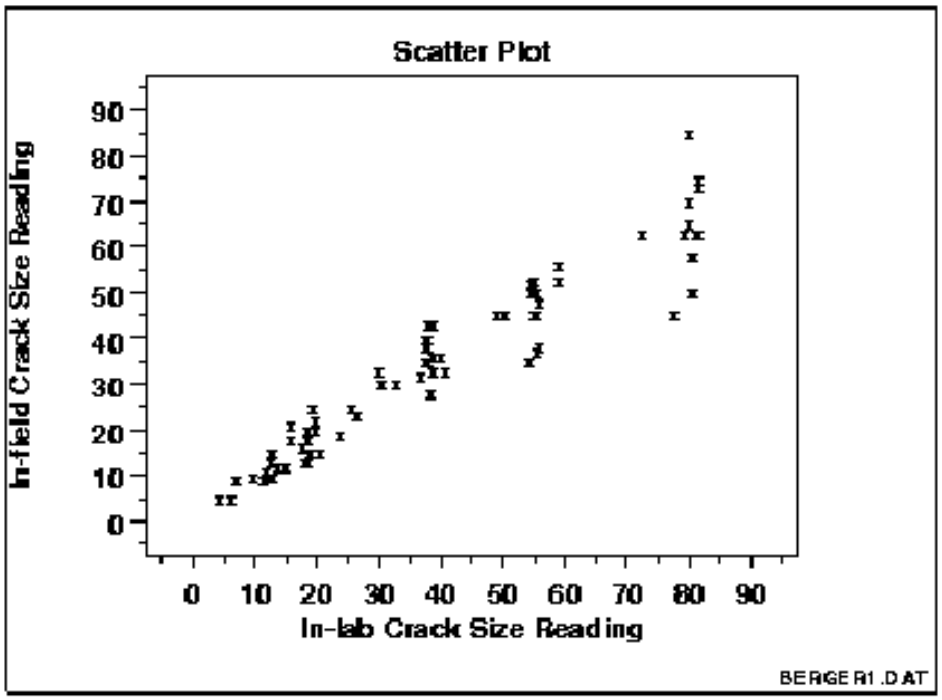
where the error is assumed to be random, from a fixed distribution, and with constant location and scale. The validity of this model depends on the validity of these assumptions. The run sequence plot is useful for checking for constant location and scale. Even for more complex models, the assumptions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

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.

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 scatter plot 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

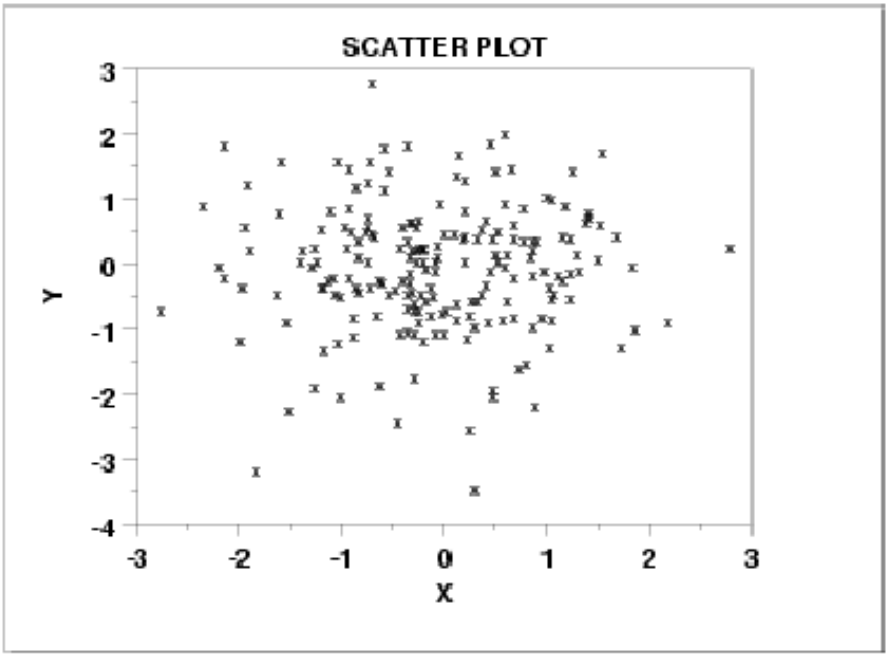
The scatter plot uncovers relationships in data. "Relationships" means that there is some structured association (linear, quadratic, etc.) between X and Y. Note, however, that even though causality implies association association does NOT imply causality.

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.

Scatter Plot: No Relationship

Scatter Plot with No Relationship



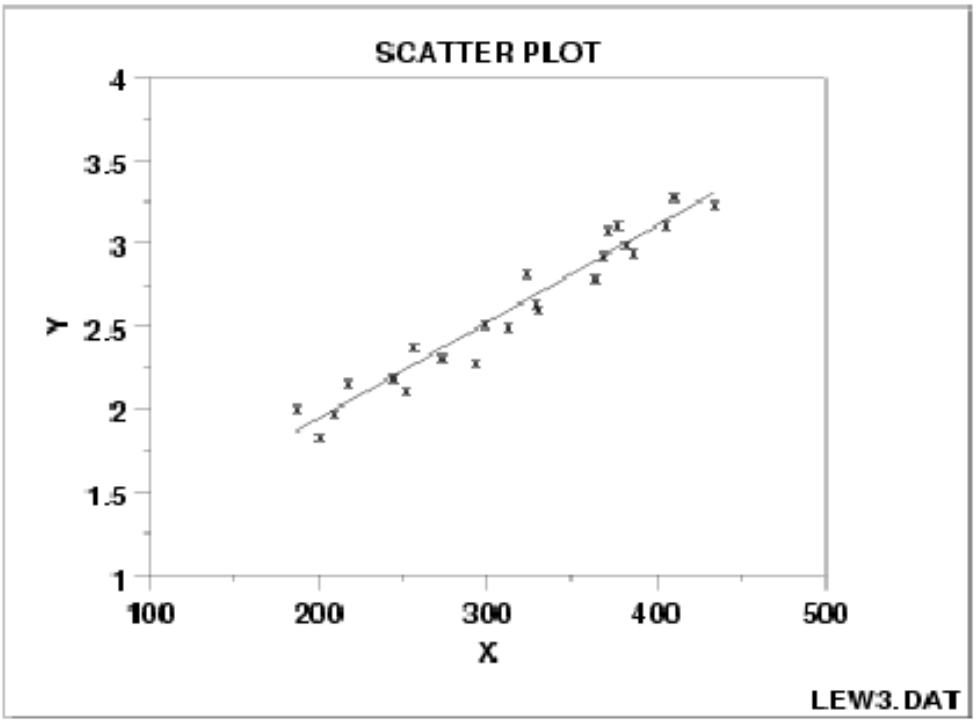
Discussion

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 predictability 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.

Scatter Plot: Strong Linear (positive correlation) Relationship

Scatter Plot Showing Strong Positive Linear Correlation

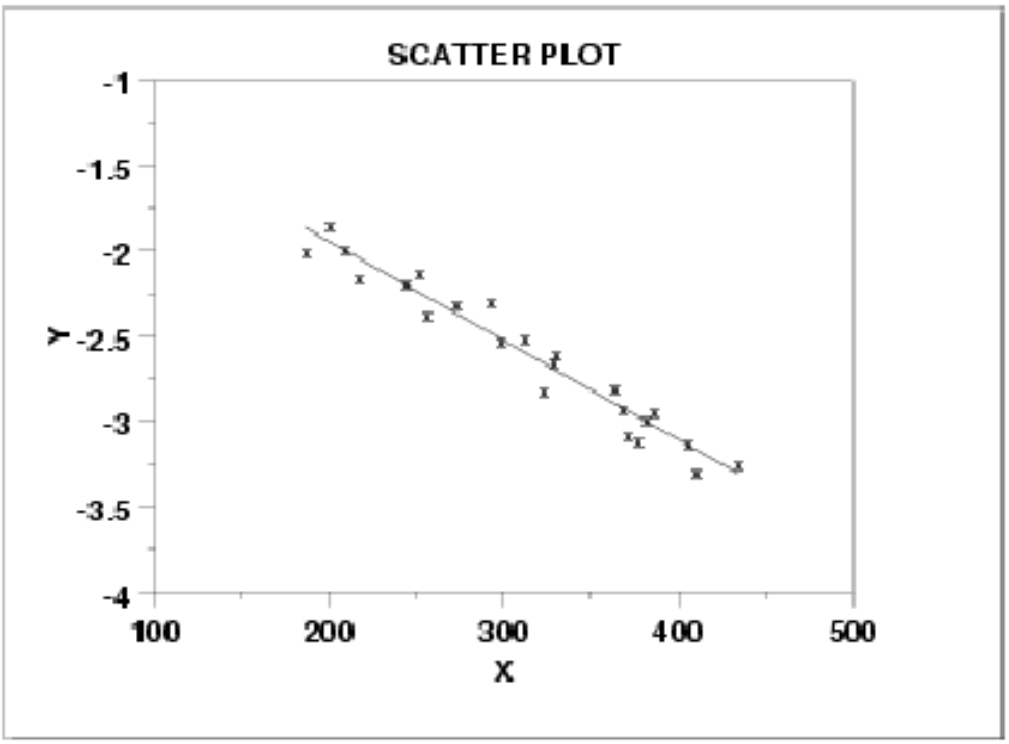


Discussion

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*.

Scatter Plot: Strong Linear (negative correlation) Relationship

Scatter Plot Showing a Strong Negative Correlation

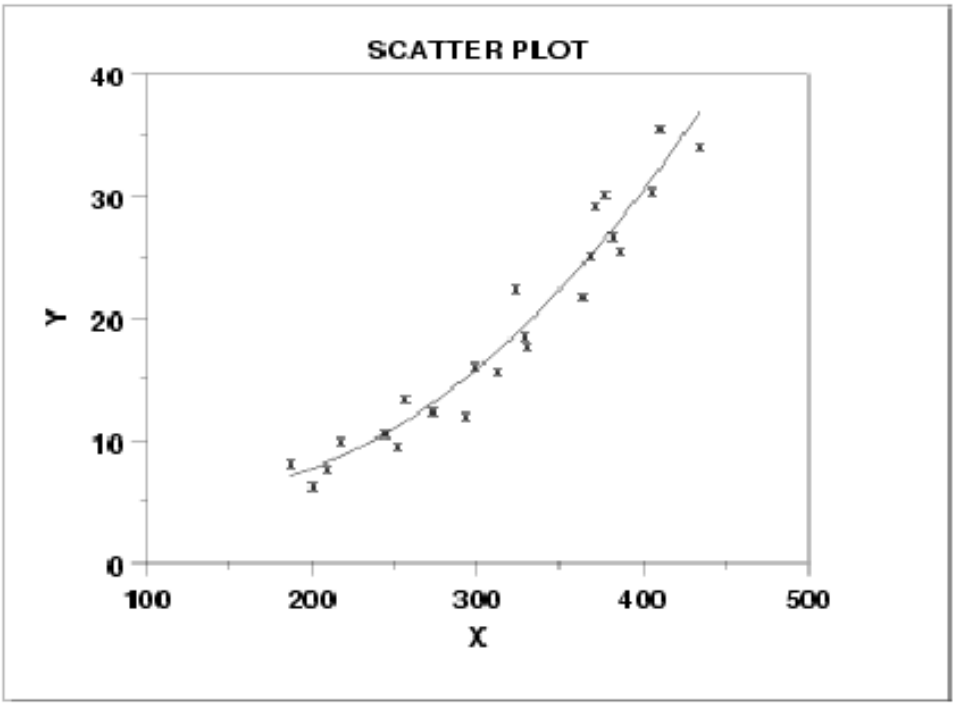


Discussion

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*.

Scatter Plot: Quadratic Relationship

Scatter Plot Showing Quadratic Relationship

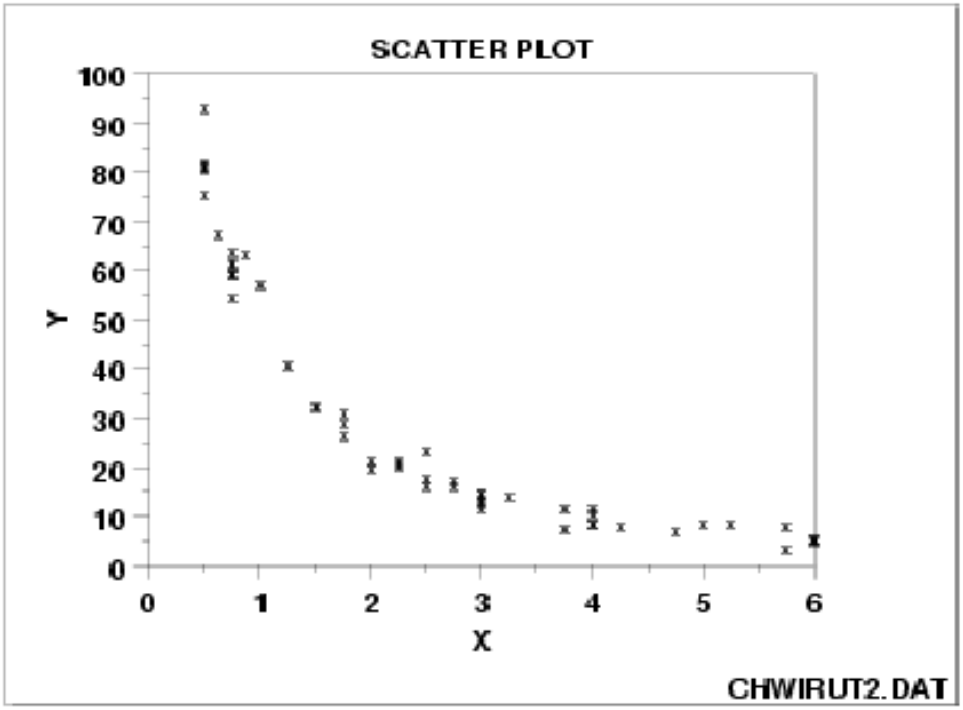


Discussion

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 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.

Scatter Plot: Exponential Relationship

Scatter Plot Showing Exponential Relationship

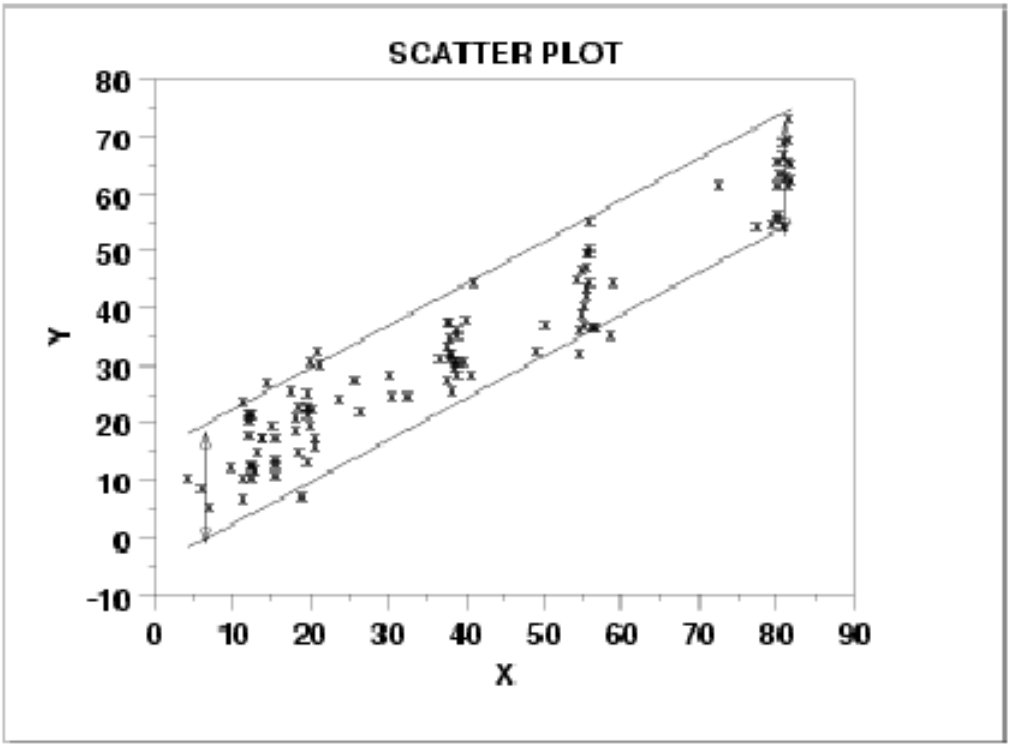


Discussion

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 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.

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

Scatter Plot Showing Homoscedastic Variability



Discussion

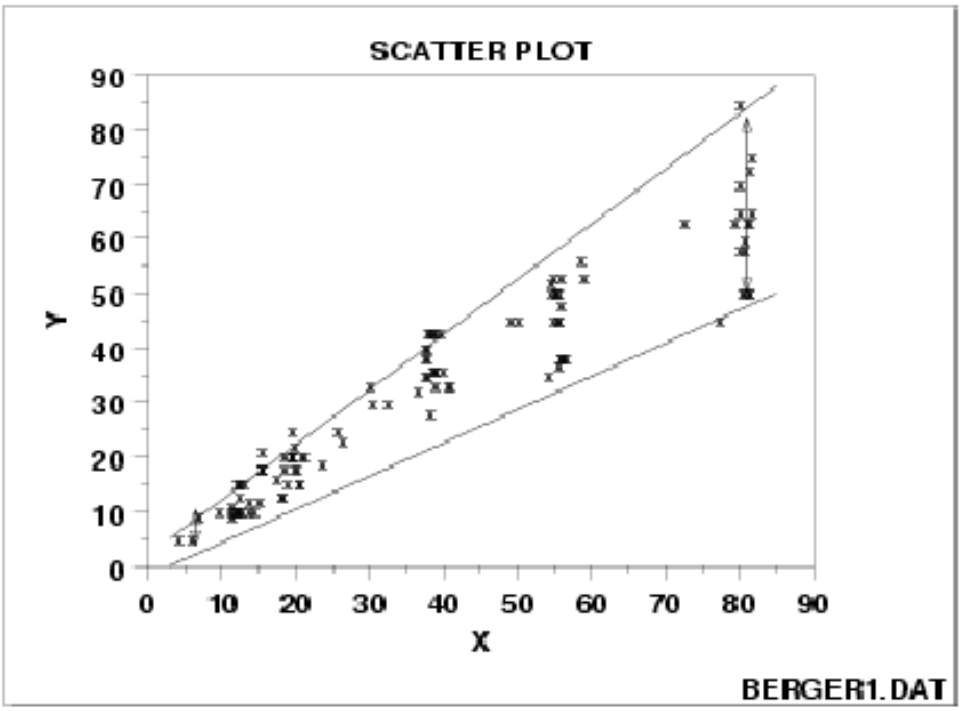
This scatter plot reveals a linear relationship between *X* and *Y*: for a given value of *X*, the predicted value of *Y* will fall on a line. The plot further reveals that the variation in *Y* about the predicted value is about the same (+- 10 units), regardless of the value of *X*.

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.

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 heteroscedastic (that is, non-constant 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

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, un-weighted 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.

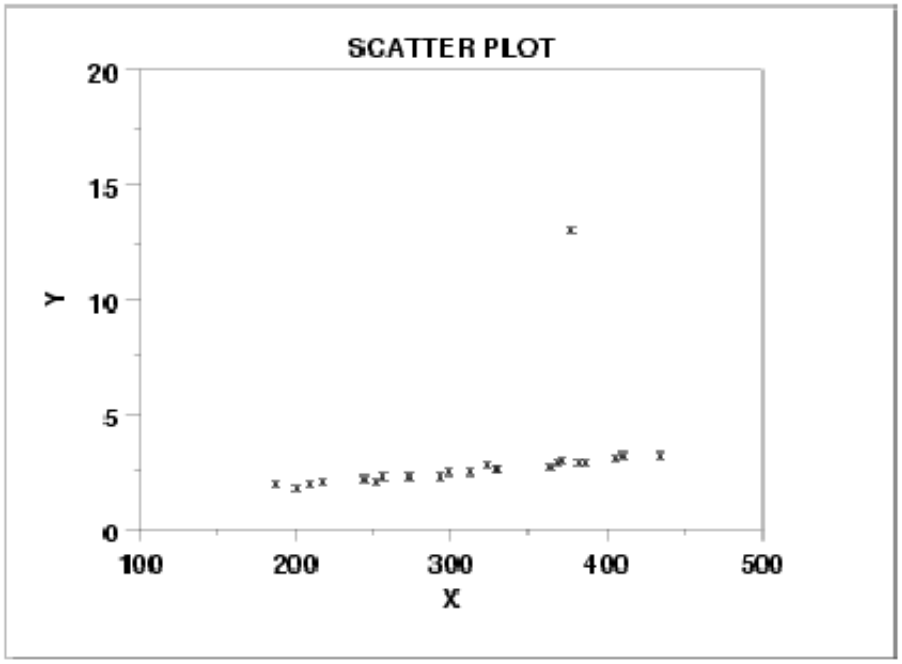
Note further that if heteroscedasticity does exist, it is frequently useful to plot and model the local variation as a function of *X*, as in. This modeling has two advantages:

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

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

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
2. 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).

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 complementary 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 statements 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:

Interval estimation

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 values of an interval which will, with a given level of confidence (i.e., probability), contain the population parameter.

Hypothesis Tests

Hypothesis tests also address the uncertainty of the sample estimate. However, instead of providing an interval, a hypothesis test attempts to refute a specific claim about a population parameter based on the sample data.

For example, the hypothesis might be one of the following:

the population mean is equal to 10

the population standard deviation is equal to 5

the means from two populations are equal

the standard deviations from 5 populations are equal

To reject a hypothesis is to conclude that it is false. However, to accept a hypothesis does not mean that it is true, only that we do not have evidence to believe otherwise. Thus hypothesis tests are usually stated in terms of both a condition that is doubted (null hypothesis) and a condition that is believed (alternative hypothesis).

A common format for a hypothesis test is:

1. H_0: A statement of the null hypothesis, e.g., two population means are equal.
2. H_a: A statement of the alternative hypothesis, e.g., two population means are not equal.

Test Statistic: The test statistic is based on the specific hypothesis test.

Significance Level: The significance level, defines the sensitivity of the test. A value of = 0.05 means that we inadvertently reject the null hypothesis 5% of the time when it is in fact true. This is also called

the type I error. The choice of is somewhat arbitrary, although in practice values of 0.1, 0.05, and 0.01 are commonly used.

The probability of rejecting the null hypothesis when it is in fact false is called the power of the test and is denoted by 1 - . Its complement, the probability of accepting the null hypothesis when the alternative hypothesis is, in fact, true (type II error), is called and can only be computed for a specific alternative hypothesis.

Critical Region: The critical region encompasses those values of the test statistic that lead to a rejection of the null hypothesis. Based on the distribution of the test statistic and the significance level, a cut-off value for the test statistic is computed. Values either above or below or both (depending on the direction of the test) this cut-off define the critical region.

Practical Versus Statistical Significance

It is important to distinguish between statistical significance and practical significance. Statistical significance simply means that we reject the null hypothesis. The ability of the test to detect differences that lead to rejection of the null hypothesis depends on the sample size.

For example, for a particularly large sample, the test may reject the null hypothesis that two process means are equivalent. However, in practice the difference between the two means may be relatively small to the point of having no real engineering significance. Similarly, if the sample size is small, a difference that is large in engineering terms may not lead to rejection of the null hypothesis. The analyst should not just blindly apply the tests, but should combine engineering judgment with statistical analysis.

Confidence Limits for the Mean

Purpose: Interval Estimate for Mean

Confidence limits for the mean (Snedecor and Cochran, 1989) are an interval estimate for the mean. Interval estimates are often desirable because the estimate of the mean varies from sample to sample. Instead of a single estimate for the mean, a confidence interval generates a lower and upper limit for the mean. The interval estimate gives an indication of how much uncertainty there is in our estimate of the true mean. The narrower the interval, the more precise is our estimate.

Confidence limits are expressed in terms of a confidence coefficient. Although the choice of confidence coefficient is somewhat arbitrary, in practice 90%, 95%, and 99% intervals are often used, with 95% being the most commonly used. As a technical note, a 95% confidence interval does **not** mean that there is a 95% probability that the interval contains the true mean. The interval computed from a given sample either contains the true mean or it does not. Instead, the level of confidence is associated with the method of calculating the interval.

The confidence coefficient is simply the proportion of samples of a given size that may be expected to contain the true mean. That is, for a 95% confidence interval, if many samples are collected and the confidence interval computed, in the long run about 95% of these intervals would contain the true mean.

Questions

Confidence limits for the mean can be used to answer the following questions:

1. What is a reasonable estimate for the mean?
2. How much variability is there in the estimate of the mean?
3. Does a given target value fall within the confidence limits?

Two-Sample *t*-Test for Equal Means

Purpose: Test if two population means are equal

The two-sample *t*-test (Snedecor and Cochran, 1989) is used to determine if two population means are equal. A common application of this is to test if a new process or treatment is superior to a current process or treatment. There are several variations on this test. The data may either be paired or not paired. By paired, we mean that there is a one-to-one correspondence between the values in the two samples. That is, if X_1, X_2, ..., X_n and Y_1, Y_2, ... , Y_n are the two samples, then X_i corresponds to Y_i. For paired samples, the difference X_i - Y_i is usually calculated. For unpaired samples, the sample sizes for the two samples may or may not be equal. The formulas for paired data are somewhat simpler than the formulas for unpaired data.

In some applications, you may want to adopt a new process or treatment only if it exceeds the current treatment by some threshold. In this case, we can state the null hypothesis in the form that the difference between the two populations means is equal to some constant ( ) where the constant is the desired threshold.

Questions

Two-sample *t*-tests can be used to answer the following questions:

1. Is process 1 equivalent to process 2?
2. Is the new process better than the current process?
1. Is the new process better than the current process by at least some pre-determined threshold amount?

One-Factor ANOVA

Purpose: Test for Equal Means Across Groups

One factor analysis of variance (Snedecor and Cochran, 1989) is a special case of analysis of variance (ANOVA), for one factor of interest, and a generalization of the two-sample *t*-test. The two-sample *t*-test is used to decide whether two groups (levels) of a factor have the same mean. One-way analysis of variance generalizes this to levels where *k*, the number of levels, is greater than or equal to 2.

For example, data collected on, say, five instruments have one factor (instruments) at five levels. The ANOVA tests whether instruments have a significant effect on the results.

Definition

The Product and Process Comparisons chapter (chapter 7) contains a more extensive discussion of 1-factor ANOVA, including the details for the mathematical computations of one-way analysis of variance. The model for the analysis of variance can be stated in two mathematically equivalent ways. In the following discussion, each level of each factor is called a cell. For the one-way case, a cell and a level are equivalent since there is only one factor. In the following, the subscript ***i*** refers to the level and the subscript ***j*** refers to the observation within a level. For example, Y_{23} refers to the third observation in the second level.

The first model decomposes the response into a mean for each cell and an error term. The analysis of variance provides estimates for each cell mean. These estimated cell means are the predicted values of the model and the differences between the response variable and the estimated cell means are the residuals.

The second model is This model decomposes the response into an overall (grand) mean, the effect of the *i*th factor level, and an error term. The analysis of variance provides estimates of the grand mean and the effect of the *i*th factor level. The predicted values and the residuals of the model are:

1. The distinction between these models is that the second model divides the cell mean into an overall mean and the effect of the *i*th factor level.
2. This second model makes the factor effect more explicit, so we will emphasize this approach.

Model Validation

Note that the ANOVA model assumes that the error term, E_{ij}, should follow the assumptions for a univariate measurement process. That is, after performing an analysis of variance, the model should be validated by analyzing the residuals.

Question

The analysis of variance can be used to answer the following question Are means the same across groups in the data?

Multi-factor Analysis of Variance

Purpose: Detect significant factors

The analysis of variance (ANOVA) (Neter, Wasserman, and Kunter, 1990) is used to detect significant factors in a multi-factor model. In the multi-factor model, there is a response (dependent) variable and one or more factor (independent) variables. This is a common model in designed experiments where the experimenter sets the values for each of the factor variables and then measures the response variable.

Each factor can take on a certain number of values. These are referred to as the levels of a factor. The number of levels can vary between factors. For designed experiments, the number of levels for a given factor tends to be small. Each factor and level combination is a cell. Balanced designs are those in which the cells have an equal number of observations and unbalanced designs are those in which the number of observations varies among cells. It is customary to use balanced designs in designed experiments.

Definition

The Product and Process Comparisons chapter (chapter 7) contains a more extensive discussion of 2-factor ANOVA, including the details for the mathematical computations. The model for the analysis of variance can be stated in two mathematically equivalent ways. We explain the model for a two-way ANOVA (the concepts are the same for additional factors). In the following discussion, each combination of factors and levels is called a cell. In the following, the subscript ***i*** refers to the level of factor 1, ***j*** refers to the level of factor 2, and the subscript ***k*** refers to the *k*th observation within the (*i*,*j*)th cell. For example, Y_{235} refers to the fifth observation in the second level of factor 1 and the third level of factor 2.

The first model decomposes the response into a mean for each cell and an error term. The analysis of variance provides estimates for each cell mean. These cell means are the predicted values of the model and the differences between the response variable and the estimated cell means are the residuals.

The second model decomposes the response into an overall (grand) mean, factor effects (and represent the effects of the *i*th level of the first factor and the *j*th level of the second factor, respectively), and an error term. The analysis of variance provides estimates of the grand mean and the factor effects.

The distinction between these models is that the second model divides the cell mean into an overall mean and factor effects. This second model makes the factor effect more explicit, so we will emphasize this approach.

Model Validation

Note that the ANOVA model assumes that the error term, ***Eijk***, should follow the assumptions for a univariate measurement process. That is, after performing an analysis of variance, the model should be validated by analyzing the residuals.

Measures of Scale

Scale, Variability, or Spread

A fundamental task in many statistical analyses is to characterize the *spread*, or variability, of a data set. Measures of scale are simply attempts to estimate this variability.

When assessing the variability of a data set, there are two key components:

How spread out are the data values near the center?

How spread out are the tails?

Different numerical summaries will give different weight to these two elements. The choice of scale estimator is often driven by which of these components you want to emphasize. The histogram is an effective graphical technique for showing both of these components of the spread.

Definitions of Variability

For univariate data, there are several common numerical measures of the spread: variance - the variance is defined as where is the mean of the data. The variance is roughly the arithmetic average of the squared distance from the mean. Squaring the distance from the mean has the effect of giving greater weight to values that are further from the mean. For example, a point 2 units from the mean adds 4 to

the above sum while a point 10 units from the mean adds 100 to the sum.

Although the variance is intended to be an overall measure of spread, it can be greatly affected by the tail behavior.

The standard deviation restores the units of the spread to the original data units (the variance squares the units).

range - the range is the largest value minus the smallest value in a data set. Note that this measure is based only on the lowest and highest extreme values in the sample. The spread near the center of the data is not captured at all.

average absolute deviation - the average absolute deviation (AAD) is defined as where is the mean of the data and **|*Y*|** is the absolute value of ***Y***. This measure does not square the distance from the mean, so it is less affected by extreme observations than are the variance and standard deviation.

median absolute deviation - the median absolute deviation (MAD) is defined as where is the median of the data and **|*Y*|** is the absolute value of ***Y***. This is a variation of the average absolute deviation that is even less affected by extremes in the tail because the data in the tails have less influence on the calculation of the median than they do on the mean.

interquartile range - this is the value of the 75th percentile minus the value of the 25th percentile. This measure of scale attempts to measure the variability of points near the center.

In summary, the variance, standard deviation, average absolute deviation, and median absolute deviation measure both aspects of the variability; that is, the variability near the center and the variability in the tails. They differ in that the average absolute deviation and median absolute deviation do not give undue weight to the tail behavior. On the other hand, the range only uses the two most extreme points and the interquartile range only uses the middle portion of the data.

Autocorrelation

Purpose: Detect Non-Randomness, Time Series Modeling

The autocorrelation (Box and Jenkins, 1976) function can be used for the following two purposes:

1. To detect non-randomness in data.
2. To identify an appropriate time series model if the data are not random.

Definition

Given measurements, Y_1, Y_2, ..., Y_N at time X_1, X_2, ..., X_N, the lag k autocorrelation function is defined as Although the time variable, X, is not used in the formula for autocorrelation, the assumption is that the observations are equi-spaced.

Autocorrelation is a correlation coefficient. However, instead of correlation between two different variables, the correlation is between two values of the same variable at times $\boldsymbol{X_i}$ and $\boldsymbol{X_{i+k}}$. When the autocorrelation is used to detect non-randomness, it is usually only the first (lag 1) autocorrelation that is of interest. When the autocorrelation is used to identify an appropriate time series model, the autocorrelations are usually plotted for many lags.

Questions

The autocorrelation function can be used to answer the following questions

1. Was this sample data set generated from a random process?
2. Would a non-linear or time series model be a more appropriate model for these data than a simple constant plus error model?

Importance

Randomness is one of the key assumptions in determining if a univariate statistical process is in control. If the assumptions of

constant location and scale, randomness, and fixed distribution are reasonable, then the univariate process can be modeled as: where E_i is an error term.

If the randomness assumption is not valid, then a different model needs to be used. This will typically be either a time series model or a non-linear model (with time as the independent variable).

What is a Probability Distribution?

Discrete Distributions

The mathematical definition of a discrete probability function, p(x), is a function that satisfies the following properties.

The probability that x can take a specific value is p(x). That is p(x) is non-negative for all real x.

The sum of p(x) over all possible values of x is 1, that is where j represents all possible values that x can have and p_j is the probability at x_j.

One consequence of properties 2 and 3 is that 0 <= p(x) <= 1.

What does this actually mean? A discrete probability function is a function that can take a discrete number of values (not necessarily finite). This is most often the non-negative integers or some subset of the non-negative integers. There is no mathematical restriction that discrete probability functions only be defined at integers, but in practice this is usually what makes sense.

For example, if you toss a coin 6 times, you can get 2 heads or 3 heads but not 2 1/2 heads. Each of the discrete values has a certain probability of occurrence that is between zero and one. That is, a discrete function that allows negative values or values greater than one is not a probability function. The condition that the probabilities sum to one means that at least one of the values has to occur.

Continuous Distributions

The mathematical definition of a continuous probability function, f(x), is a function that satisfies the following properties.

1. The probability that x is between two points a and b
2. It is non-negative for all real x
3. The integral of the probability function is one

What does this actually mean? Since continuous probability functions are defined for an infinite number of points over a continuous interval, the probability at a single point is always zero. Probabilities are measured over intervals, not single points. That is, the area under the curve between two distinct points defines the probability for that interval. This means that the height of the probability function can in fact be greater than one. The property that the integral must equal one is equivalent to the property for discrete distributions that the sum of all the probabilities must equal one.

Probability Mass Functions versus Probability Density Functions

Discrete probability functions are referred to as probability mass functions and continuous probability functions are referred to as probability density functions. The term probability functions cover both discrete and continuous distributions. When we are referring to probability functions in generic terms, we may use the term probability density functions to mean both discrete and continuous probability functions.