Analiza datelor experimentale

Elementele principale de reținut pentru analiza grafică a datelor experimentale sunt următoarele:

1. DATELE EXPERIMENTALE (x_i, y_i) obținute din experiment, conțin:
   variabila independentă "x" (se impune, de exemplu temperatura) şi
   variabila dependentă "y" (se măsoară, de exemplu rezistența electrică).
2. PLOT. Se reprezită grafic datele experimentale.
   alege abscisa (axa x) =mărimea reprezentată, valoare minimă/maximă
   alege ordonata (axa y) =mărimea reprezentată, valoare minimă/maximă
   alege unitatea de măsură pentru axe 1cm (x) =10°C, 1mm (y)=1 Ohm
3. IPOTEZA => $y = f(x)$
   "Funcția f modelează (descrie) felul cum se modifică variabila dependentă y în funcție de variabila independentă x."
   Ai funcția de fitare, ştii parametrii care trebuiesc determinați.
4. FITAREA (potrivirea)
   grafic este *curba care trece cel mai aproape de punctele experimentale*.
   găsirea parametrilor din funcția $f(x)$ care *minimizează suma pătratelor deviațiilor* (reziduurilor "$r_i=f_i-y_i$" $f_i = f(x_i)$) "hi-pătrat" $\chi^2=\sum(f_i-y_i)^2$
   [metoda celor mai mici pătrate] Karl Frie. Gauss 1809, Adrien Marie Legendre 1805
   [treaba asta o face un program soft, de exemplu ORIGIN, EXCEL]
5. CALITATEA FITĂRII, *goodness of fit*.
   Reziduurile (y_i-f_i) ca funcție de x, au o *distribuție aleatoare*. Dacă nu, atunci funcția propusă nu cuprinde tot comportamentul mărimii "y" şi trebuie modificată.

Utilizarea concretă a metodei celor mai mici pătrate pentru funcțiea liniară e descris mai jos (*opțional*).

Fitarea cu o funcție liniară a datelor experimentale (x_i, y_i):

$y = f(x) = ax+b$ $\qquad$ $f_i = ax_i+b$ $\qquad$ [a şi b sunt necunoscutele]

Funcția care trebuie minimizată:

$$\chi^2 = \sum_{i=1}^{N} (f_i - y_i)^2 = \sum_{i=1}^{N} (ax_i + b - y_i)^2$$

sau dezvoltat:

$$\chi^2 = a^2\sum_{i=1}^{N} x_i^2 + 2ab\sum_{i=1}^{N} x_i - 2a\sum_{i=1}^{N} x_i y_i + Nb^2 - 2b\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} y_i^2$$

unde "N" este numărul de măsurători. Aici "a" şi "b" sunt necunoscutele, iar sumele sunt de fapt nişte simple numere.

Ca să fie un minim trebuie ca prima derivată (după "a" şi după "b") să se anuleze:

$$\frac{\partial \chi^2}{\partial a} = 0 = 2a\sum_{i=1}^{N} x_i^2 + 2b\sum_{i=1}^{N} x_i - 2\sum_{i=1}^{N} x_i y_i \qquad \frac{\partial \chi^2}{\partial b} = 0 = 2a\sum_{i=1}^{N} x_i + 2Nb - 2\sum_{i=1}^{N} y_i$$

Cele 2 condiții duc la un sistem liniar de 2 ecuații cu 2 necunoscute:

$$a\sum_{i=1}^{N} x_i^2 + b\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i y_i$$

$$a\sum_{i=1}^{N} x_i + Nb = \sum_{i=1}^{N} y_i$$

Soluția sistemului este:

$$a = \frac{N\sum_{i=1}^{N} x_i y_i - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} \qquad b = \frac{N\left(\sum_{i=1}^{N} x_i^2\right)\left(\sum_{i=1}^{N} y_i\right) - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} x_i y_i\right)}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}$$

sau mai simplu: $$a = \frac{\sum_{i-1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i-1}^{N} (x_i - \bar{x})^2} \qquad b = \bar{y} - a\bar{x}$$

unde $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$ $\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N}$ sunt valorile medii ale lui "x" şi "y", şi am folosit identităţile:

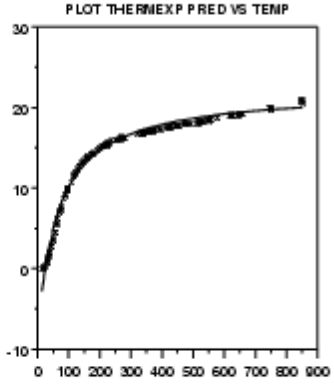
$$\bar{y}\sum_{i=1}^{N} x_i = \bar{x}\sum_{i=1}^{N} y_i = N\bar{x}\cdot\bar{y} \qquad \bar{x}\sum_{i=1}^{N} x_i = N\bar{x}^2$$

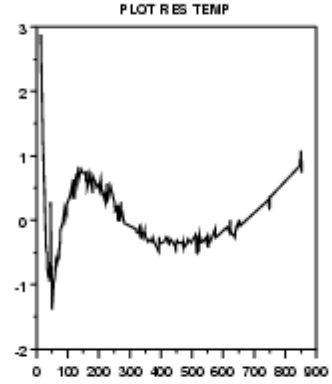
EXEMPLU: a = −80.2 / 22.8 = −3.5175 b = −4.8 − (−3.5175)(3.2) = 6.456 din:

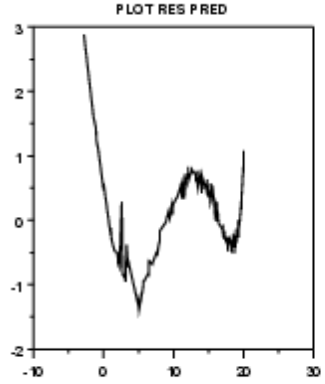
n=5	x	y	$x-\underline{x}$.	$y-\underline{y}$.	$(x-\underline{x})^2$.	$(x-\underline{x})(y-\underline{y})$
	0	6	–3.2	10.8	10.24	–34.56
	2	–1	–1.2	3.8	1.44	–4.56
	3	–3	–0.2	1.8	0.04	–0.36
	5	–10	1.8	–5.2	3.24	–9.36
	6	–16	2.8	–11.2	7.84	–31.36
∑	16	–24	0	0	22.8	–80.2
media	3.2	–4.8				

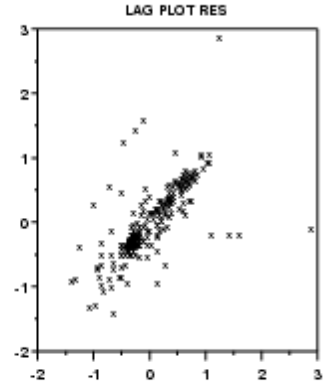
Exemplu NIST=> Thermal Expansion of Copper Case Study

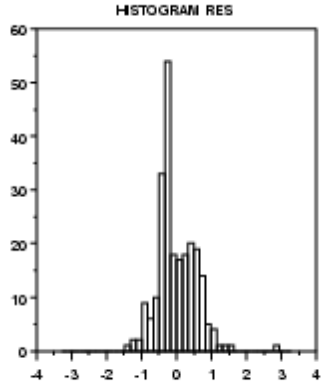
Aproximează coeficientul de dilatare printr-un raport de polinoame de gradul doi (pătratice)
Eroarea (graficul din mijloc sus) nu este aleatoare

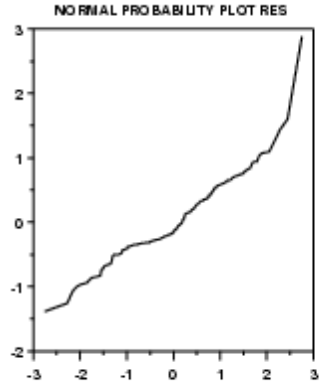




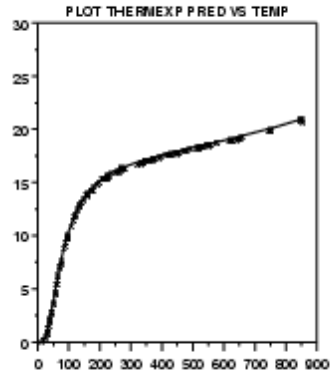


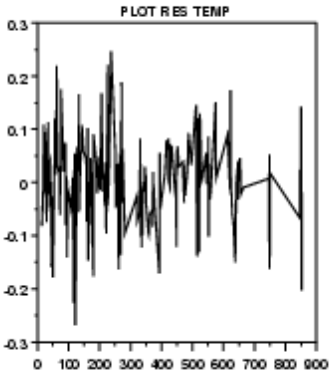


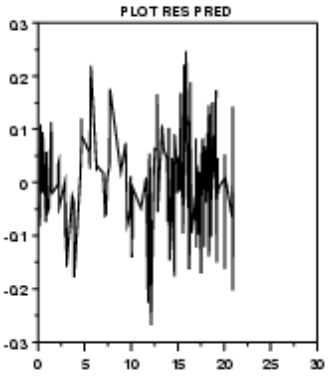


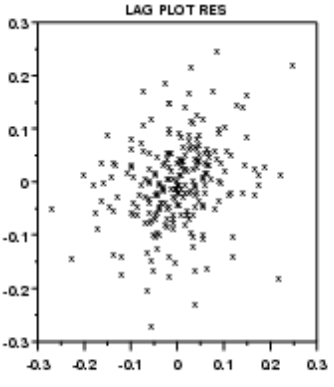


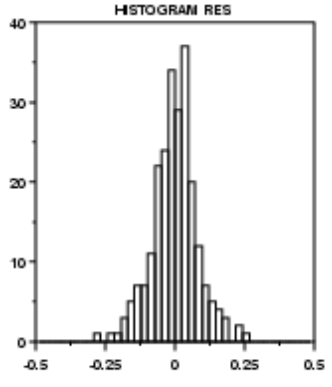
Aproximează coeficientul de dilatare printr-un raport de polinoame de gradul trei (cubice)
Eroarea (graficul din mijloc sus) este aleatoare şi mult mai mică

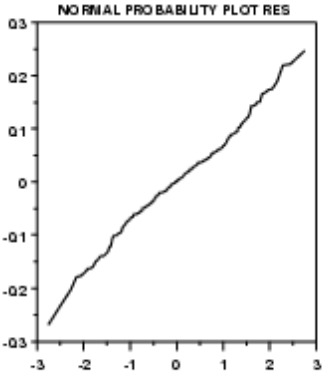












Surse

"LEAST SQUARES FITTING OF EXPERIMENTAL DATA"
http://35.9.69.219/home/modules/pdf_modules/m359.pdf
de la http://physnet2.pa.msu.edu/index.html

Cum se lucrează foarte corect => Thermal Expansion of Copper Case Study
http://www.itl.nist.gov/div898/handbook/pmd/section6/pmd64.htm

de la NIST, National Institute of Standards and Technology (USA)
Engineering Statistics Handbook
http://www.itl.nist.gov/div898/handbook/index.htm
NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

Gallery of Quantitative Techniques from the Handbook
http://www.itl.nist.gov/div898/handbook/quantgal.htm

DATAPLOT Summary [program de prelucrarea datelor]
http://www.itl.nist.gov/div898/software/dataplot/summary.htm
program for performing scientific, engineering, statistical, mathematical, and graphical analysis.

Adăugate fiindcă merită atenție (sunt doar în engleză!)

http://www.itl.nist.gov/div898/handbook/eda/section3/eda34.htm
1.3.4. Graphical Techniques: By Problem Category

http://www.itl.nist.gov/div898/handbook/eda/section3/lagplot.htm
1.3.3.15. Lag Plot

http://www.itl.nist.gov/div898/handbook/eda/section3/normprpl.htm
1.3.3.21. Normal Probability Plot

http://www.itl.nist.gov/div898/handbook/eda/section3/eda362.htm#CDF
1.3.6.2. Related Distributions

http://www.itl.nist.gov/div898/handbook/eda/section3/eda36.htm
1.3.6. Probability Distributions

http://www.itl.nist.gov/div898/handbook/eda/section3/eda364.htm
1.3.6.4. Location and Scale Parameters

http://www.itl.nist.gov/div898/handbook/eda/section3/6plot.htm
1.3.3.33. 6-Plot

http://www.itl.nist.gov/div898/handbook/eda/section3/eda365.htm
1.3.6.5. Estimating the Parameters of a Distribution

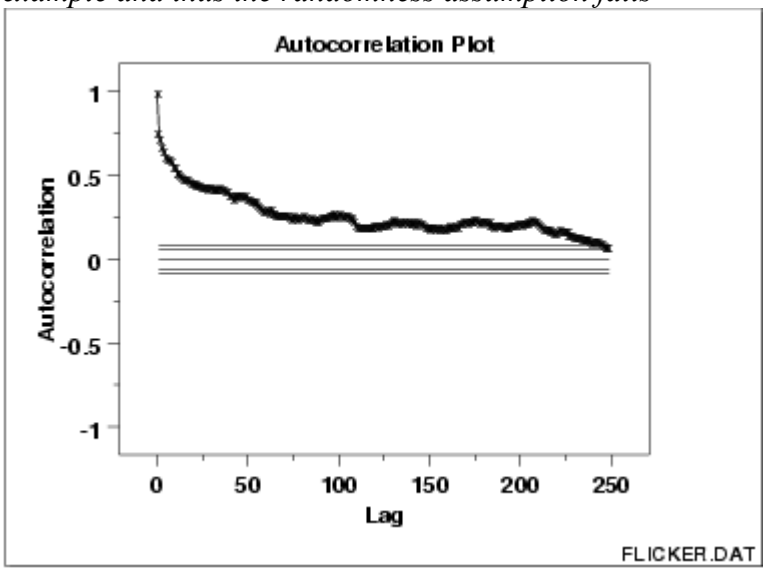
http://www.itl.nist.gov/div898/handbook/eda/eda.htm
1. Exploratory Data Analysis

1.3.3.1. Autocorrelation Plot

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

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

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



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

Definition: r(h) versus h
Autocorrelation plots are formed by

1. **Vertical axis**: Autocorrelation coefficient
   $R_h = C_h / C_0$
   where C_h is the autocovariance function
   $$C_h = \frac{1}{N} \sum_{t=1}^{N-h} (Y_t - \bar{Y})(Y_{t+h} - \bar{Y})$$
   and C_0 is the variance function
   $$C_0 = \frac{\sum_{t=1}^{N} (Y_t - \bar{Y})^2}{N}$$
   Note--R_h is between -1 and +1.
   Note--Some sources may use the following formula for the autocovariance function
   $$C_h = \frac{1}{N-h} \sum_{t=1}^{N-h} (Y_t - \bar{Y})(Y_{t+h} - \bar{Y})$$
   Although this definition has less bias, the (1/*N*) formulation has some desirable statistical properties and is the form most commonly used in the statistics literature. See pages 20 and 49-50 in Chatfield for details.
2. **Horizontal axis**: Time lag *h* (*h* = 1, 2, 3, ...)
3. The above line also contains several horizontal reference lines. The middle line is at zero. The other four lines are 95% and 99% confidence bands. Note that there are two distinct formulas for generating the confidence bands.

1. If the autocorrelation plot is being used to test for randomness (i.e., there is no time dependence in the data), the following formula is recommended:

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

where N is the sample size, z is the percent point function of the standard normal distribution and α is the. significance level. In this case, the confidence bands have fixed width that depends on the sample size. This is the formula that was used to generate the confidence bands in the above plot.

2. Autocorrelation plots are also used in the model identification stage for fitting ARIMA models. In this case, a moving average model is assumed for the data and the following confidence bands should be generated:

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

where k is the lag, N is the sample size, z is the percent point function of the standard normal distribution and α is. the significance level. In this case, the confidence bands increase as the lag increases.

Questions

The autocorrelation plot can provide answers to the following questions:

1. Are the data random?
2. Is an observation related to an adjacent observation?
3. Is an observation related to an observation twice-removed? (etc.)
4. Is the observed time series white noise?
5. Is the observed time series sinusoidal?
6. Is the observed time series autoregressive?
7. What is an appropriate model for the observed time series?
8. Is the model
Y = constant + error
valid and sufficient?
9. Is the formula $s_{\bar{Y}} = s/\sqrt{N}$ valid?

Importance: Ensure validity of engineering conclusions

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

1. Most standard statistical tests depend on randomness. The validity of the test conclusions is directly linked to the validity of the randomness assumption.
2. Many commonly-used statistical formulae depend on the randomness assumption, the most common formula being the formula for determining the standard deviation of the sample mean:
$s_{\bar{Y}} = s/\sqrt{N}$
where s is the standard deviation of the data. Although heavily used, the results from using this formula are of no value unless the randomness assumption holds.
3. For univariate data, the default model is
Y = constant + error
If the data are not random, this model is incorrect and invalid, and the estimates for the parameters (such as the constant) become nonsensical and invalid.

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

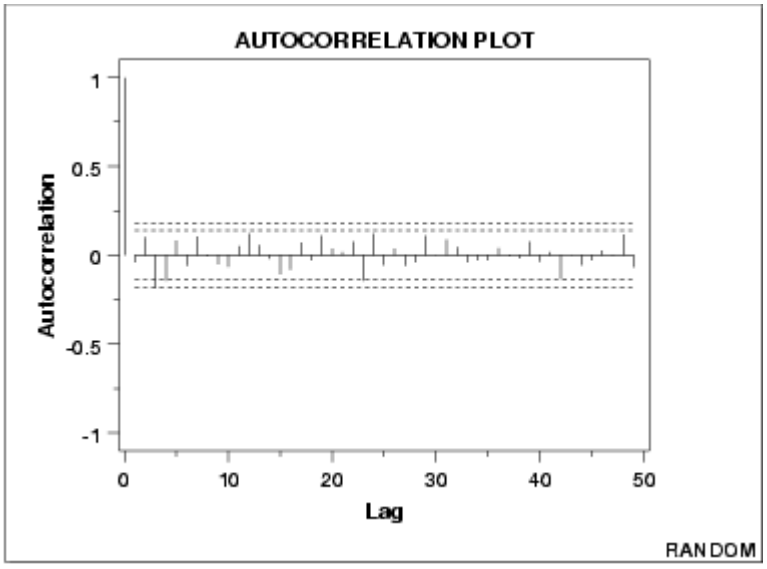
Examples
Examples of the autocorrelation plot for several common situations are given in the following pages.
1. Random (= White Noise)
2. Weak autocorrelation
3. Strong autocorrelation and autoregressive model
4 Sinusoidal model

Related Techniques
Partial Autocorrelation Plot
Lag Plot
Spectral Plot
Seasonal Subseries Plot
Case Study
The autocorrelation plot is demonstrated in the beam deflection data case study.

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

Autocorrelation Plot: Random Data

Autocorrelation Plot The following is a sample autocorrelation plot.



Conclusions We can make the following conclusions from this plot.

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

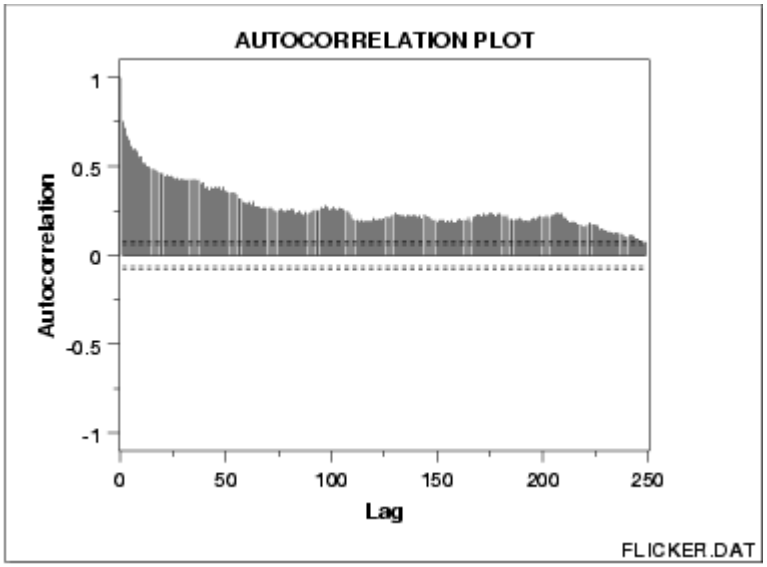
Discussion
Note that with the exception of lag 0, which is always 1 by definition, almost all of the autocorrelations fall within the 95% confidence limits. In addition, there is no apparent pattern (such as the first twenty-five being positive and the second twenty-five being negative). This is the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

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

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

Moderate Autocorrelation

Autocorrelation Plot The following is a sample autocorrelation plot.



Conclusions
We can make the following conclusions from this plot.

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

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

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

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

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

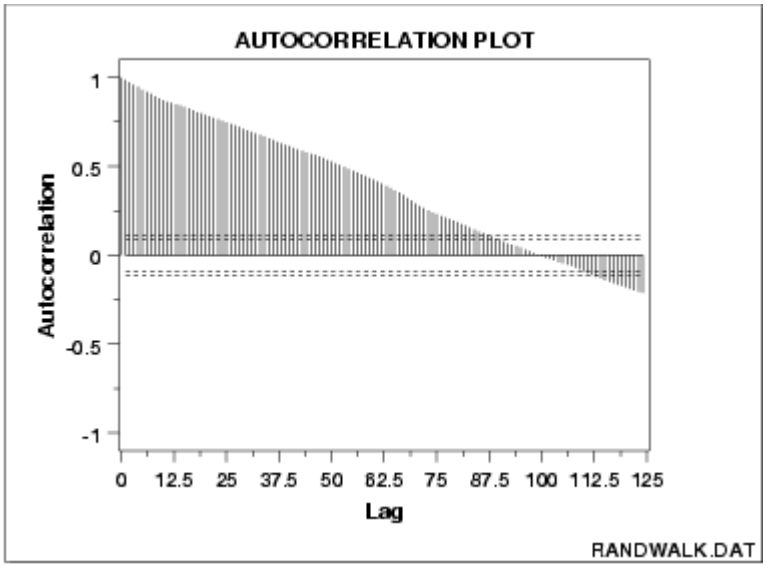
The randomness assumption for least squares fitting applies to the residuals of the model. That is, even though the original data exhibit randomness, the residuals after fitting $\mathbf{Y}_i$ against $\mathbf{Y}_{i-1}$ should result in random residuals. Assessing whether or not the proposed model in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

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

$$Y_i = A_0 + E_i$$

Strong Autocorrelation and Autoregressive Model

Autocorrelation Plot for Strong Autocorrelation The following is a sample autocorrelation plot.



Conclusions
We can make the following conclusions from the above plot.
The data come from an underlying autoregressive model with strong positive autocorrelation.

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

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

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

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

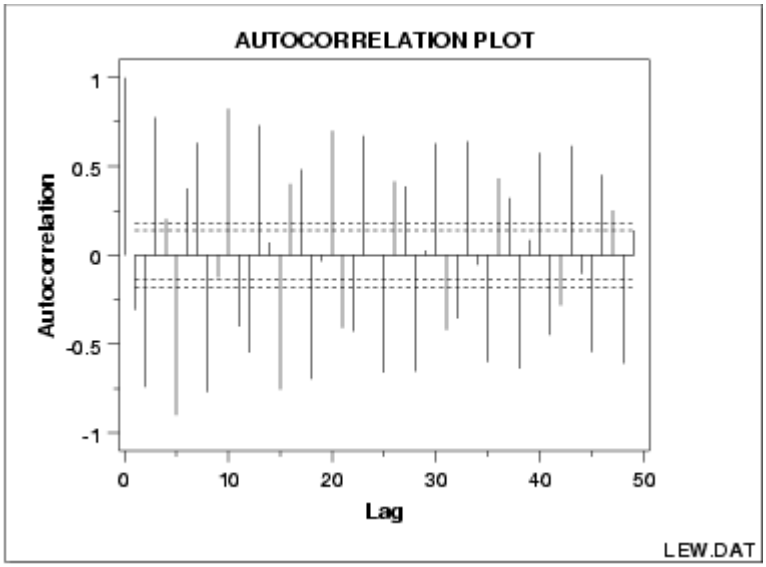
The randomness assumption for least squares fitting applies to the residuals of the model. That is, even though the original data exhibit randomness, the residuals after fitting $\mathbf{Y}_i$ against $\mathbf{Y}_{i-1}$ should result in random residuals. Assessing whether or not the proposed model in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

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

$$Y_i = A_0 + E_i$$

Autocorrelation Plot: Sinusoidal Model

Autocorrelation Plot for Sinusoidal Model The following is a sample autocorrelation plot.



Conclusions
We can make the following conclusions from the above plot.
The data come from an underlying sinusoidal model.

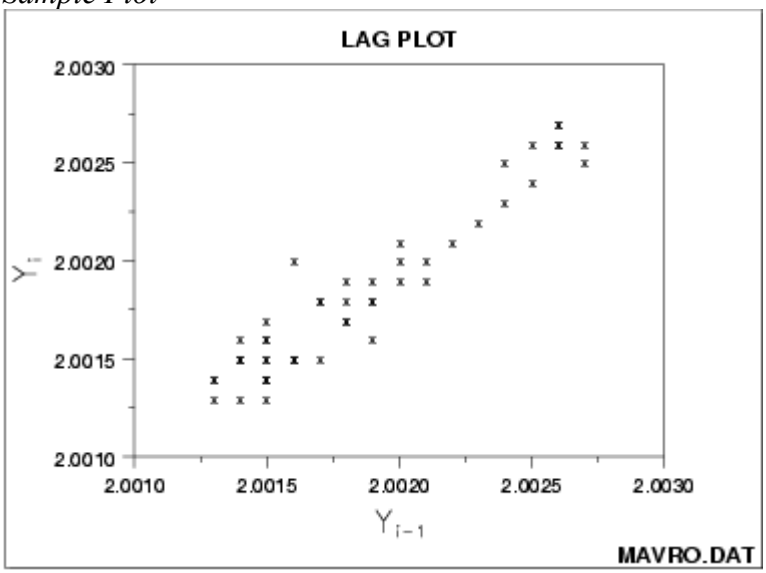
Discussion
The plot exhibits an alternating sequence of positive and negative spikes. These spikes are not decaying to zero. Such a pattern is the autocorrelation plot signature of a sinusoidal model.

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

1.3.3.15. Lag Plot

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

Sample Plot



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

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

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

Vertical axis: Y_i for all i

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

Questions

Lag plots can provide answers to the following questions:

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

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

Examples

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

Related Techniques

Autocorrelation Plot
Spectrum
Runs Test

Case Study

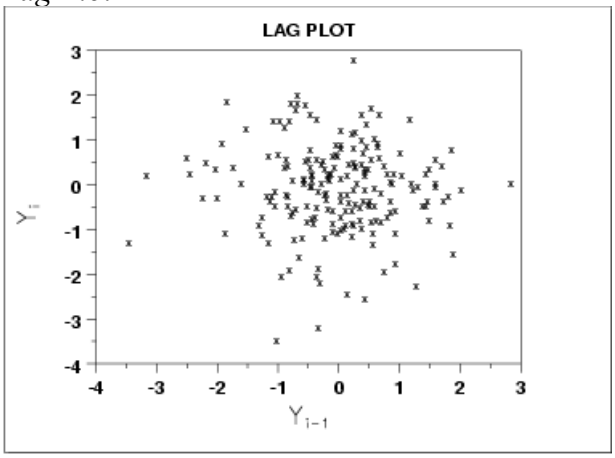
The lag plot is demonstrated in the beam deflection data case study.

Software

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

1.=Lag Plot: Random Data

Lag Plot



Conclusions

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

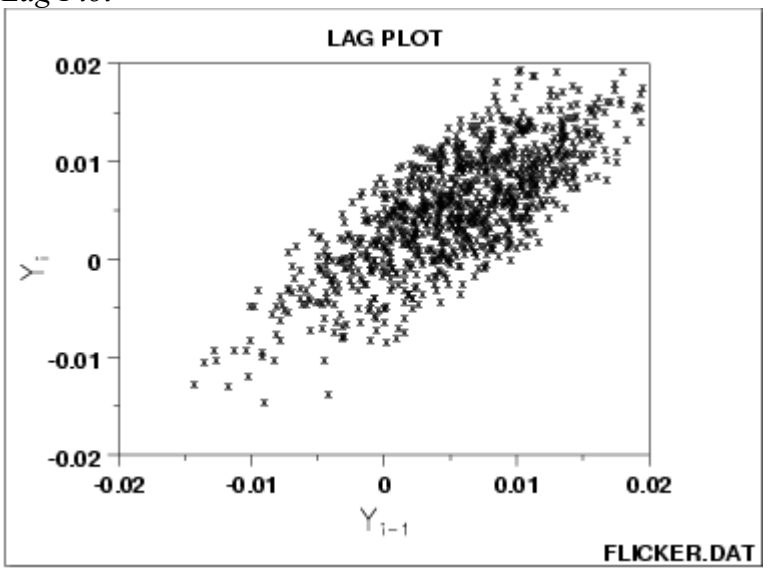
1. The data are random.
2. The data exhibit no autocorrelation.

3. The data contain no outliers.

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

2.=Lag Plot: Moderate Autocorrelation

Lag Plot



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

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

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

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

Recommended Next Step
Estimate the parameters for the autoregressive model:

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

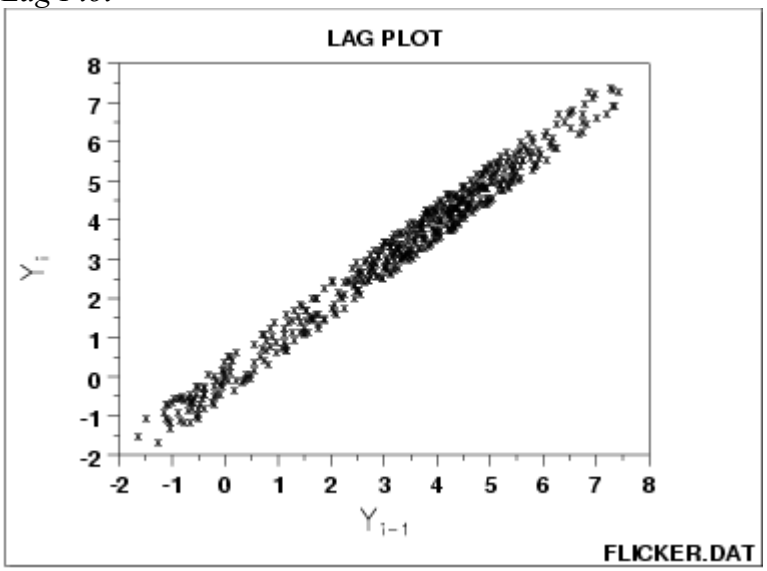
Since Y_i and Y_{i-1} are precisely the axes of the lag plot, such estimation is a linear regression straight from the lag plot.

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

$$Y_i = A_0 + E_i$$

3.=Lag Plot: Strong Autocorrelation and Autoregressive Model

Lag Plot



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

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

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

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

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

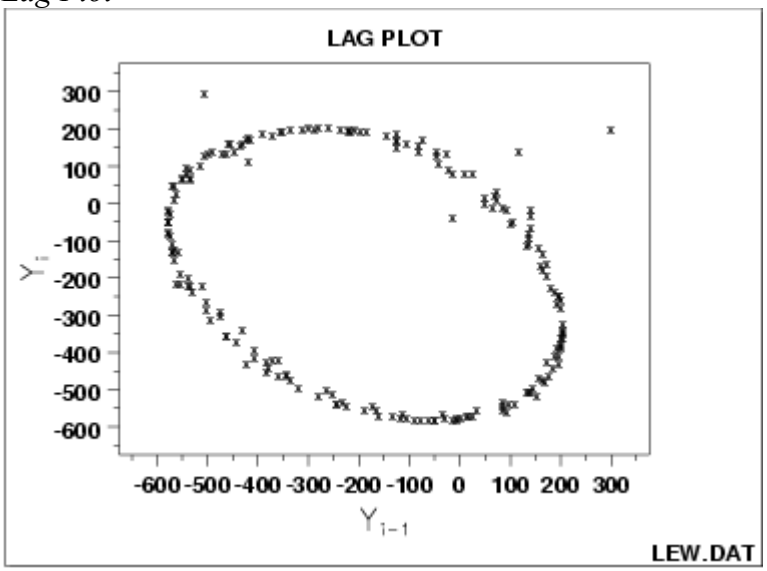
1. Extimate the parameters for the autoregressive model:
   $Y_i = A_0 + A_1 * Y_{i-1} + E_i$
   Since Y_i and Y_{i-1} are precisely the axes of the lag plot, such estimation is a linear regression straight from the lag plot.

   The residual standard deviation for this autoregressive model will be much smaller than the residual standard deviation for the default model
   $Y_i = A_0 + E_i$
2. Reexamine the system to arrive at an explanation for the strong autocorrelation. Is it due to the
   1. phenomenon under study; or
   2. drifting in the environment; or
   3. contamination from the data acquisition system?

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

4.=Lag Plot: Sinusoidal Models and Outliers

Lag Plot



Conclusions

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

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

Discussion

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

Consequences of Ignoring Cyclical Pattern

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

$$Y_i = A_0 + E_i$$

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

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

The proper model

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

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

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

Unexpected Value of EDA

Frequently a technique (e.g., the lag plot) is constructed to check one aspect (e.g., randomness) which

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

Recommended Next Step

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

1. Do a spectral plot to obtain an initial estimate of the frequency of the underlying cycle. This will be helpful as a starting value for the subsequent non-linear fitting.
2. Omit the outliers.
3. Carry out a non-linear fit of the model to the 197 points.
   $Y_i = C + \alpha \sin\left(2\pi\omega t_i + \phi\right) + E_i$

1.3.6. Probability Distributions

Probability Distributions
Probability distributions are a fundamental concept in statistics. They are used both on a theoretical level and a practical level.

Some practical uses of probability distributions are:

1. To calculate confidence intervals for parameters and to calculate critical regions for hypothesis tests.
2. For univariate data, it is often useful to determine a reasonable distributional model for the data.
3. Statistical intervals and hypothesis tests are often based on specific distributional assumptions. Before computing an interval or test based on a distributional assumption, we need to verify that the assumption is justified for the given data set. In this case, the distribution does not need to be the best-fitting distribution for the data, but an adequate enough model so that the statistical technique yields valid conclusions.
4. Simulation studies with random numbers generated from using a specific probability distribution are often needed.

Table of Contents

1. 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.

1. The probability that x can take a specific value is p(x). That is
   $P[X = x] = p(x) = p_x$
2. p(x) is non-negative for all real x.
3. The sum of p(x) over all possible values of x is 1, that is
   $\sum_j p_j = 1$
   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. $0 \le p(x) \le 1$

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 is
$$p[a \le x \le b] = \int_a^b f(x)dx$$
2. It is non-negative for all real x.
3. The integral of the probability function is one, that is
$$\int_{-\infty}^{\infty} f(x)dx = 1$$

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 covers 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.

+2. Related Distributions

Probability distributions are typically defined in terms of the probability density function. However, there are a number of probability functions used in applications.

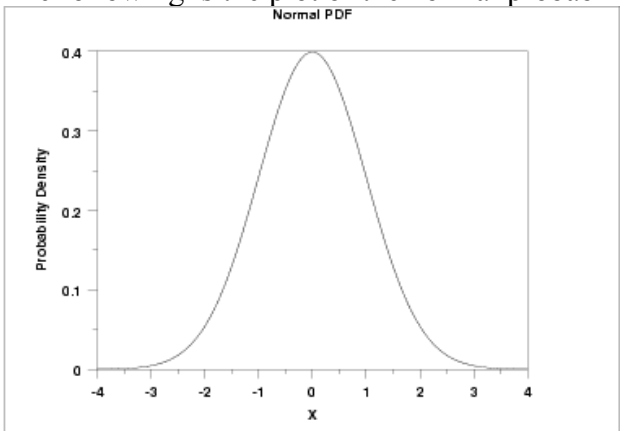
Probability Density Function
For a continuous function, the probability density function (pdf) is the probability that the variate has the value x. Since for continuous distributions the probability at a single point is zero, this is often expressed in terms of an integral between two points.

$$\int_a^b f(x)dx = Pr[a \le X \le b]$$

For a discrete distribution, the pdf is the probability that the variate takes the value x.

$$f(x) = Pr[X = x]$$

The following is the plot of the normal probability density function.



Cumulative Distribution Function
The cumulative distribution function (cdf) is the probability that the variable takes a value less than or equal to x. That is

$$F(x) = Pr[X \le x] = \alpha$$

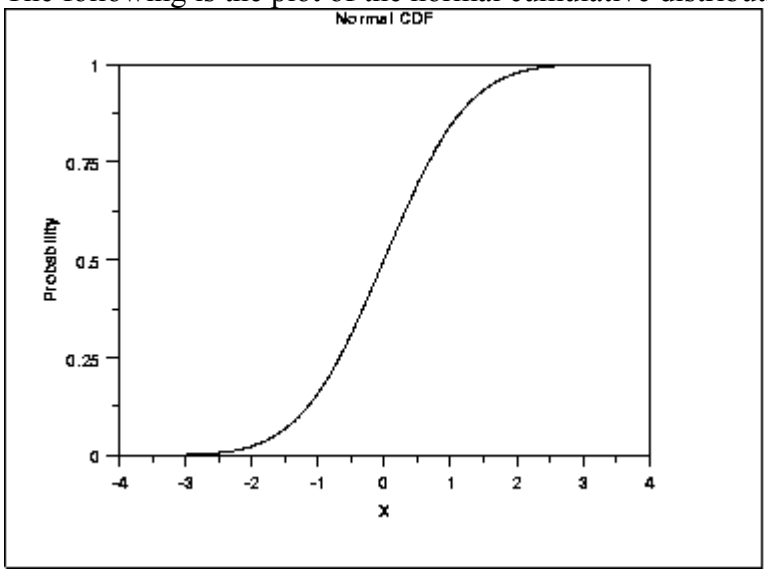
For a continuous distribution, this can be expressed mathematically as

$$F(x) = \int_{-\infty}^{x} f(\mu) d\mu$$

For a discrete distribution, the cdf can be expressed as

$$F(x) = \sum_{i=0}^{x} f(i)$$

The following is the plot of the normal cumulative distribution function.



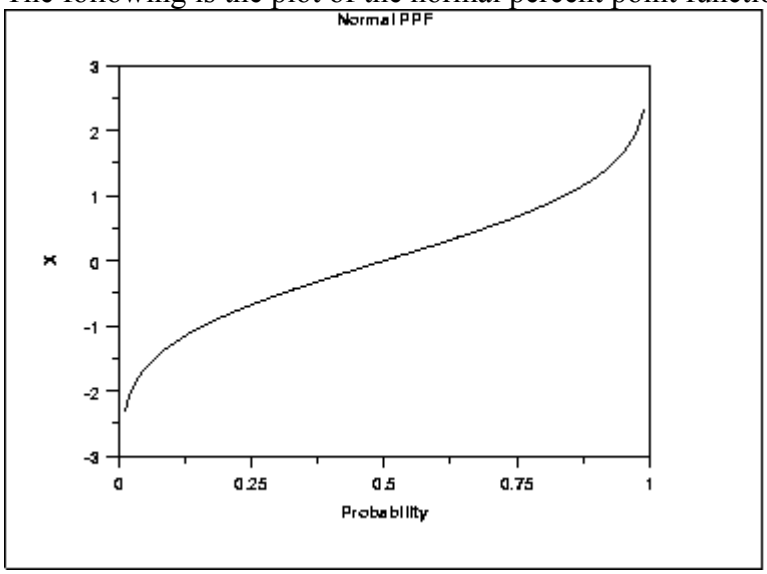
The horizontal axis is the allowable domain for the given probability function. Since the vertical axis is a probability, it must fall between zero and one. It increases from zero to one as we go from left to right on the horizontal axis.

Percent Point Function

The percent point function (ppf) is the inverse of the cumulative distribution function. For this reason, the percent point function is also commonly referred to as the inverse distribution function. That is, for a distribution function we calculate the probability that the variable is less than or equal to *x* for a given *x*. For the percent point function, we start with the probability and compute the corresponding *x* for the cumulative distribution. Mathematically, this can be expressed as $Pr[X \leq G(\alpha)] = \alpha$

or alternatively $x = G(\alpha) = G(F(x))$

The following is the plot of the normal percent point function.



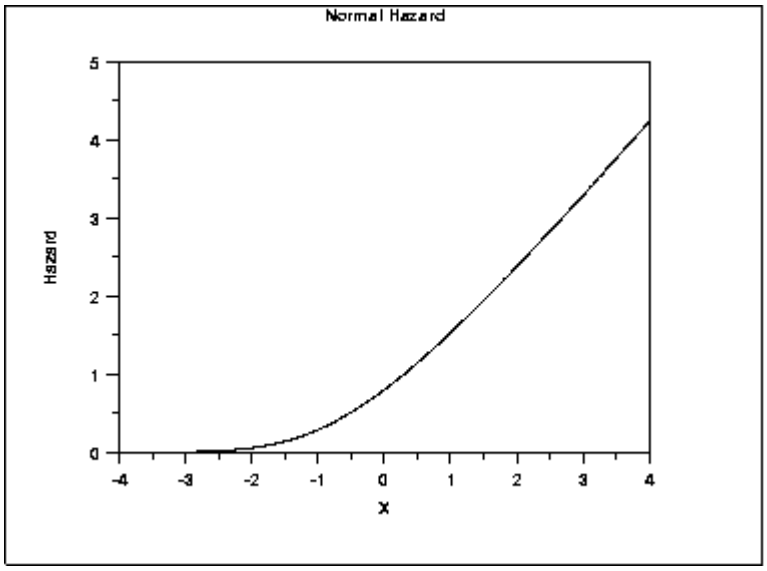
Since the horizontal axis is a probability, it goes from zero to one. The vertical axis goes from the smallest to the largest value of the cumulative distribution function.

Hazard Function

The hazard function is the ratio of the probability density function to the survival function, *S*(x).

$$h(x) = \frac{f(x)}{S(x)} = \frac{f(x)}{1 - F(x)}$$

The following is the plot of the normal distribution hazard function.



Hazard plots are most commonly used in reliability applications. Note that Johnson, Kotz, and Balakrishnan refer to this as the conditional failure density function rather than the hazard function.

Cumulative Hazard Function

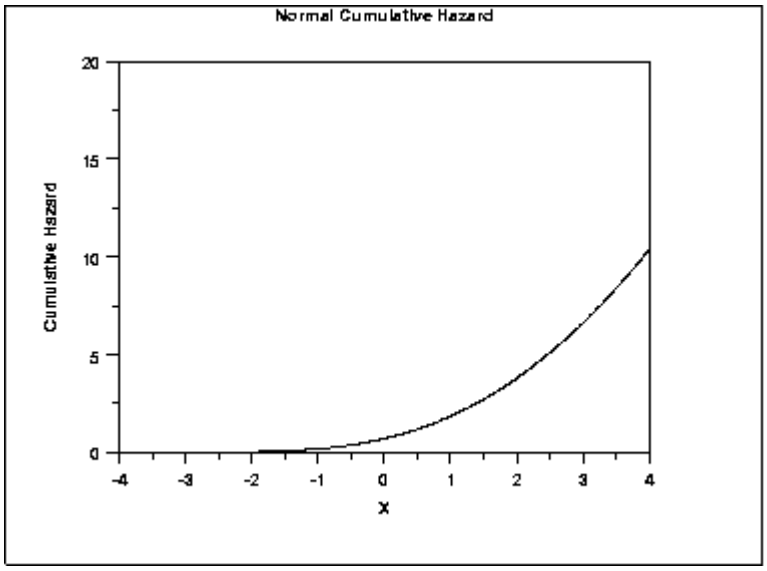
The cumulative hazard function is the integral of the hazard function. It can be interpreted as the probability of failure at time x given survival until time x.

$$H(x) = \int_{-\infty}^{x} h(\mu) d\mu$$

This can alternatively be expressed as

$$H(x) = -\ln(1 - F(x))$$

The following is the plot of the normal cumulative hazard function.



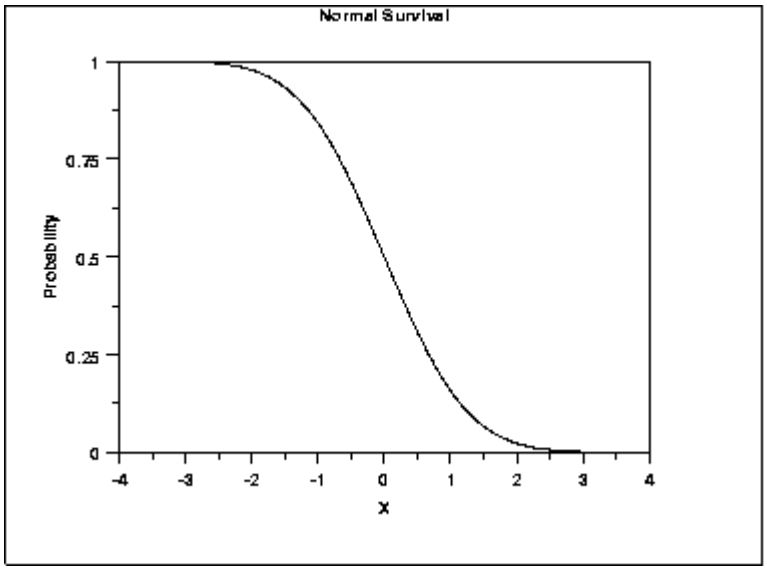
Cumulative hazard plots are most commonly used in reliability applications. Note that Johnson, Kotz, and Balakrishnan refer to this as the hazard function rather than the cumulative hazard function.

Survival Function

Survival functions are most often used in reliability and related fields. The survival function is the probability that the variate takes a value greater than x.

$$S(x) = Pr[X > x] = 1 - F(x)$$

The following is the plot of the normal distribution survival function.



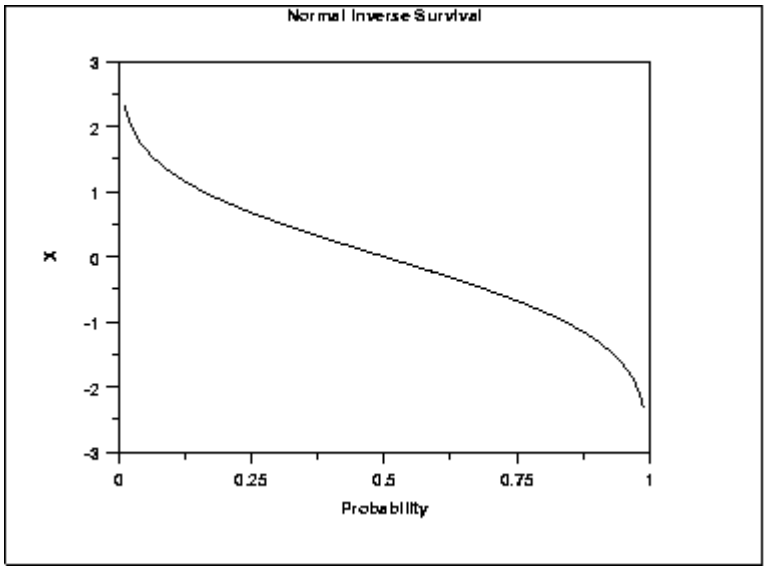
For a survival function, the y value on the graph starts at 1 and monotonically decreases to zero. The survival function should be compared to the cumulative distribution function.

Inverse Survival Function

Just as the percent point function is the inverse of the cumulative distribution function, the survival function also has an inverse function. The inverse survival function can be defined in terms of the percent point function.

$$Z(\alpha) = G(1 - \alpha)$$

The following is the plot of the normal distribution inverse survival function.



As with the percent point function, the horizontal axis is a probability. Therefore the horizontal axis goes from 0 to 1 regardless of the particular distribution. The appearance is similar to the percent point function. However, instead of going from the smallest to the largest value on the vertical axis, it goes from the largest to the smallest value.