V. ANALYZING THE RESULTS

1. Coding the Responses

Data processing normally includes the following sub-processes:

- Questionnaire editing the answers to the questionnaires where the focus is on seemingly inconsistent responses (For instance, we may find someone who is 85 years old and still studying. Upon verification, it turns out that the age has been erroneously recorded and that the correct age is 8.5 years old.)
- Coding the answers or assigning numbers to the responses. This saves computer space and allows for more systematic analysis.
- Encoding is the process of entering the coded responses into a database system.
- Data editing is the process of verifying that encoded data is correct.

Coding the responses entails assigning numbers to the responses. Some researchers assign letters (called character) but this is not advisable since it is not possible to perform mathematical operations on letters. Remember that a space or blank is considered a character. For this reason, we assign a special number to nonresponse as well as to “not applicable” answers.

The simplest type of code is the listing code. As the name implies, the responses are simply listed out, there is no attempt to arrange them in whatever manner, and the codes will simply run from 1 to the number of different responses.

Advantage: The original responses are preserved
Disadvantage: Data analysis is difficult. For instance, the frequency distribution will yield a frequency of 1 for every answer.

Another type of code is the group code. Responses are sorted into meaningful groups or themes and a number is assigned to each group.

Example. Types of Irrigation Systems

1 Large-scale systems
- National Irrigation Systems
- Communal Irrigation Systems

2 Small-scale systems
- Small water impounding project
- Small farm reservoir
- Shallow-tube wells
- Low-lift pumps

3 None or rainfed

Advantage: Data analysis is easier and the results can be more meaningful. It also requires less computer space. Of course, over time, this latter reason is becoming irrelevant with advances in IT.

Disadvantage: We lose information on the original responses. For instance, if in the future we will be asked how many farmers use shallow tube wells, we will not be able to answer this outright since the response was lumped under code "2". We will need to go back to the questionnaires.

.

A better alternative is to use system code. This type of code combines some of the advantages of the listing code and the group code. Using the same example, the system code can look this way:

1 Large-scale systems
  11 National Irrigation Systems
  12 Communal Irrigation Systems
2 Small-scale systems
  21 Small water impounding project
  22 Small farm reservoir
  23 Shallow-tube wells
  24 Low-lift pumps
3 None or rainfed

The codes that will be encoded into the computer are the 2-digit codes. Note that if we only want to know the number of large-scale systems, we can just truncate the data to isolate the first digit.

Advantage: Responses are classified into meaningful groups but individual responses are preserved.

Disadvantage: Data will require more computer space.

Note that in system codes, it is not necessary to complete the number series (that is, from 1 to 9). It is also important to assign codes that facilitate recall. Take the following example:

Highest Educational Attainment

00	No formal schooling
11	Grade 1
12	Grade 2
13	Grade 3
14	Grade 4
15	Grade 5
16	Grade 6
17	Grade 7
21	First year HS
22	Second year HS
23	Third year HS
24	Fourth year HS
27	First year, post-secondary
28	Second year, post-secondary
29	Third year, post-secondary
31	First year, college
32	Second year, college
33	Third year, college
34	Fourth year, college
40	Post graduate

The postal code is a very popular example of a system code. Actually, There are a number of coding systems already developed within the Philippine Statistical System. It is always wise to subscribe to this standard coding system for at least two reasons. First, it saves us the time and energy simply thinking and developing the codes. Second, it facilitates comparison with official data. Following are some of the more common variables with a standard coding system.

Variable	Standard Coding System
Location of residence	Philippine Standard Geographical Codes
Occupation	Philippine Standard Occupational Codes
Industry or Sector of Employment	Philippine Standard Industry Code
Goods and Services	Standard International Trade Classification

We can also refer to NSO questionnaires and apply the same codes for similar variables like: source of water, type of housing, material of roof, material of walls, type of toilet, class of worker, etc.

2. Describing the Data Profile

a. Levels of Measurement

First, we need to distinguish between four levels of measurement, according to increasing degree of correspondence to the real number system.

- Nominal: The numbers assigned to objects are numerical but do not have a number meaning. They do not strictly correspond to numbers which means that they can neither be ordered nor added up. E.g., when you assign numbers to categories of sex, civil status, college degree, place of residence, etc. Note that it is not possible to have a mean sex, mean college degree, mean place of residence, etc.

- Ordinal: The objects of a set can be rank-ordered on an operationally defined characteristic or property.

Example 1: when educational attainment is rank-ordered on the basis of the number of years of schooling:

0 - no formal schooling
1 - less than elementary
2 - elementary graduate
3 - less than HS
4 - HS graduate
5 - less than college
6 - college graduate

Example 2: when cities and municipalities are ranked according to class:

1 - First class
2 - Second class
3 - Third class
4 - Fourth class
5 - Fifth class

Note that the assignation rule still does not strictly correspond to the real number system. A mean educational attainment of 2.6 does not have meaning. And how do you treat an individual with vocational training - similar to or higher than someone with less than college education? Similarly, how do you interpret a mean municipality class of 3.3?

- Interval. Numerically equal distances on interval scales represent equal distances in the property being measured. Intervals can be added or subtracted.

- Ratio. In addition to possessing the characteristics of nominal, ordinal and interval scales, the ratio scale has an absolute or natural zero that has empirical meaning. A zero income would mean no income, a person with income of 100 pesos has 100 times more than someone with an income of 1 peso.

b. Descriptive Statistics

After data has been collected, the next step is to explore characteristics of the data. A systematic approach would be to arrange them according to a certain property. When data is grouped according to magnitude, the resulting series is called a frequency distribution; if it is grouped according to time of occurrence, then it is called a time series; if according to geographic location, the resulting series is called a spatial distribution.

i. Frequency Distribution

A frequency distribution is an arrangement of numerical data according to size or magnitude.

Example : 1, 2, 1, 1, 2, 2, 1, 1, 1, 2

Re-arranged series: 1, 1, 1, 1, 1, 1, 2, 2, 2, 2

Sex	Freq
1(M)	6
2(F)	4

Construction of a Frequency Distribution

Following are the steps for constructing a frequency distribution:

1. Using the range of data as a guide, the data are divided into a number of conveniently sized groups.
2. The groups are then placed in a column with the lowest interval at the top and the rest follows according to size.
3. The data are then scored.

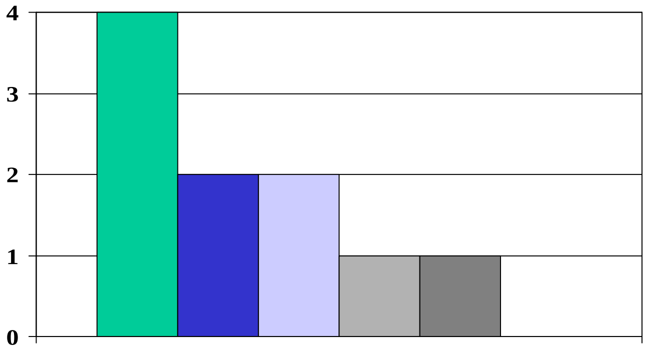
Example : 10, 12, 51, 45, 32, 15, 15, 20, 25, 31

Re-arranged series: 10, 12, 15, 15, 20, 25, 31, 32, 45, 51

Age	Freq
10 - 19	4
20 - 29	2
30 - 39	2
40 - 49	1
50 - 59	1

Types of Frequency Distribution

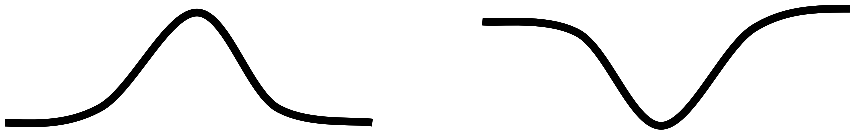
A frequency distribution may be graphically represented as a histogram. The width of the bars is equivalent to the size of the class interval (class width) while the length is equivalent to the frequency of observations falling within the interval.



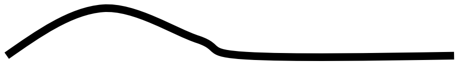
A frequency distribution may be characterized according to degree of symmetry, peakedness and modalities.

Degree of Symmetry

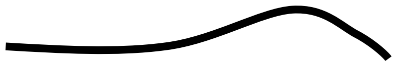
A frequency distribution may be symmetrical about a center. This means that observations to the right of the center are distributed in the same way as observations to the left.



A distribution may also be skewed. A positively skewed distribution is caused by a few observations with extremely high values

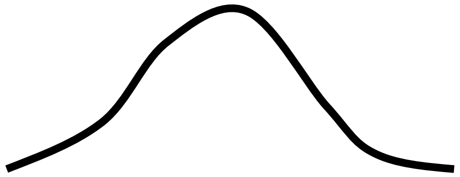


while a negative skewed distribution is caused by a few observations with extremely low values.

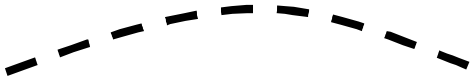


Peakedness

A distribution characterized by a high degree of peakedness, or a marked concentration around a certain value, is called leptokurtic.

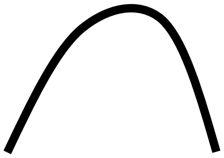


Otherwise, it is called platykurtic.

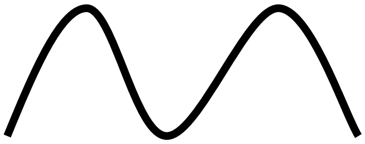


Modalities

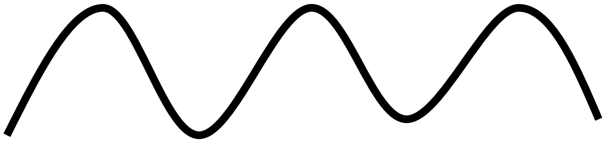
A distribution characterized by a concentration about a single value is called unimodal.



If concentration occurs around two values, then it is called bimodal.



Some distributions may have no or more than two modalities.



ii. *Summary Statistics*

It would be more convenient if the data being analyzed can be described in only one or two numbers. The most common summary statistics being used are the measures of central tendency and dispersion. Under these are different indices that correspond to different functions and come with different sets of advantages and disadvantages.

Measures of Central Tendency

These measures answer the question: ‘What is the typical value that has been observed?’

Arithmetic Mean

The arithmetic mean is a calculated average whose value is determined by every observation. It can easily pulled up or down by the presence of extreme values. In addition, it has the following characteristics:

- the sum of deviations about the mean is zero
- the sum of squares of the deviations about the mean is less than those computed about any other point
- its standard error is less than that of the median
- it has a determinate value

By formula, the arithmetic mean is the sum of the observations divided by the number of observations. E.g., given farmers 1, 2, 3, 4 each with 0.8, 1.2, 2.0, and 1.8 ha. of farmland, respectively, then the mean farm area is 1.45 ha.

Among the advantages of the arithmetic mean are the following:

- it is most commonly used
- it is easily understood
- it is generally recognized
- its computation is simple, and
- it may be treated algebraically

However, its major disadvantage is that its value may be distorted by the presence of extreme values and therefore may not be typical. If instead, we have farmers A, B, C and D with farms measuring 1, 1, 1, 13 ha., respectively. Then the mean farm area is 4 has. which gives a misleading picture of the distribution.

Median

The median is the value of the middle observation when the data items are arranged according to size. It is an average of position. Other characteristics are:

- it is affected by the number of items, not by the size of the extreme values
- the sum of the deviations about the median, signs ignored, is less than those computed about any other point

- it is the most typical value when the central values of the series are closely grouped
- a value selected at random is just as likely to be located above the median as below

If the number of observations is even, then the median is the average of the two middle observations. If odd, then the median is simply the middle observation.

The advantages of using the median as measure of central tendency include:

- it can easily be calculated
- it is not distorted in value by unusual items
- it can be considered as more typical of the series because of its independence from unusual values
- it can be calculated even when the distribution is open-ended

On the other hand, the disadvantages are the following:

- it is not so generally familiar as the arithmetic mean
- items must first be arranged according to size before it can be computed; however with the improvements in technology, this ceases to be an issue
- it has a larger standard error than the arithmetic mean
- algebraic manipulation involving the median is difficult

Mode

The mode is the most frequent or most common value, provided that a sufficiently large number of items are available to give a smooth distribution. Like the median, it is an average of position and is independent of extreme values.

The advantages of using the mode as measure of central tendency are:

- it is most typical and therefore the most descriptive
- it can be approximated by inspection where there only a small number of observations
- it is not necessary to arrange the values or know them if they are few in number.

The disadvantages are the following:

- if there is only a limited amount of data available, then its value can only be approximated
- its significance is limited when a large number of values is not available
- in a small number of items, the mode may not even exist for none of the values may be repeated

Geometric Mean

The geometric mean is the nth root of the product of n items. It is a calculated average and is dependent upon the size of all the values. However, unlike the mean, it is less affected by extreme items. In general, it is always smaller than the arithmetic mean.

Using the same example as before, the geometric mean farm area is 1.36 has.

The advantages are:

- it is a more typical average since it is less affected by extremes
- it may be manipulated algebraically
- it is useful in the computation of index numbers

The disadvantages are:

- it is not widely known
- it is difficult to compute, but then again, with electronic processing, this is no longer an issue
- it may be indeterminate if there are negative values in the series or where one of the items is zero

Harmonic Mean

The harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of the values. It is commonly used in averaging rates.

iii. Measures of Dispersion

Consider two groups of farmers. Group 1 consists of farmers 1, 2, 3, 4 with farm lands measuring 4, 4, 4, and 4 has, respectively. Group 2 consists of farmers A, B, C, and D with farm lands measuring 1, 1, 1, and 13 has., respectively. Both groups have a calculated average of 4 has. but quite obviously differ with respect to distributions. Reporting only the measure of central tendency will be a misrepresentation. Just imagine the consequences of such a practice were the numbers given above actually correspond to systematic measurements of depth, in feet, of two river systems.

There has to be another summary statistic that gives out information on the dispersion of observations.

Range

The range is the difference between the maximum and the minimum values. Sometimes, the difference is no longer computed, instead the maximum and minimum values are simply stated.

Mean Deviation

Also called absolute deviation, it is the average of the deviations of the items from either the arithmetic mean or the median, signs ignored.

Standard Deviation

The standard deviation is a special form of the mean deviation. It is computed by taking the square root of the quadratic mean of the deviations from the arithmetic mean.

Quartile Deviation

Also called the semi-interquartile range, it is computed as one-half the distance between the first and third quartiles.

iv. Relative Measure of Dispersion

At times, we are interested in comparing the dispersion of two or more distributions having different units of measurement, say feet and meters. Quite obviously, we can just convert the observations so that they would be in the same unit of measurement. A simpler way is to compare their coefficients of variation which is the ratio of the standard deviation to the mean, commonly expressed in percentage form.

c. Graphic Presentation

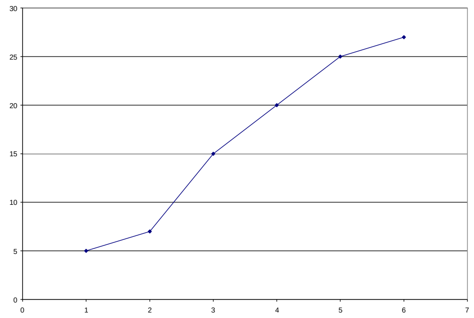
Data may also be graphically presented as graphs, charts, area diagrams or statistical maps.

The general rules for constructing graphs/charts are the following:

- Every graph must have a title which indicates the nature of the data, geographical area and time period covered.
- Data source should be indicated just under the graph and to the left; while footnotes, if any, are shown under and to the right.
- Each scale must have a scale caption indicating units used. Indicate zero point.
- Keep gridlines to a minimum. Tick marks may be used to indicate gradations on the scales although it is not necessary to show the fine gradations.

Line or Curve Graphs

Observations are represented as points and plotted according to their respective values on the X and Y scales. These points are connected by straight lines. The scales may follow either arithmetic, logarithmic or semi-logarithmic rulings.



Arithmetic Rulings

Equal amounts are assigned equal distances. Equal changes indicate identical absolute differences.

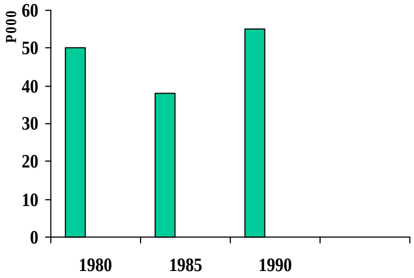
Logarithmic and Semi-logarithmic Rulings

Equal proportional changes are assigned equal distances. In this example, we see that the differences between 3 and 6 and between 100 and 200 both come out as a difference of 0.30103 representing the same 100% increase.

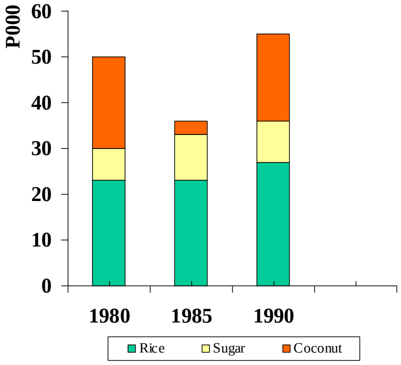
Usually, only one of the axes follow logarithmic ruling, e.g., X axis is a time dimension. This is referred to as semi-logarithmic ruling.

Bar Chart

Variables are compared using bars of varying length but uniform width. Here, we give examples of different types of bar charts: absolute simple, absolute subdivided, percent simple, and percent subdivided. Absolute charts use the raw observations as scale while percent charts use percentage scales.

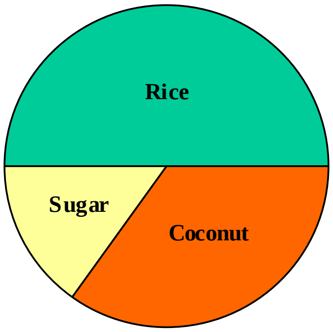


Subdivided charts are meant to present the composition of aggregate figures.



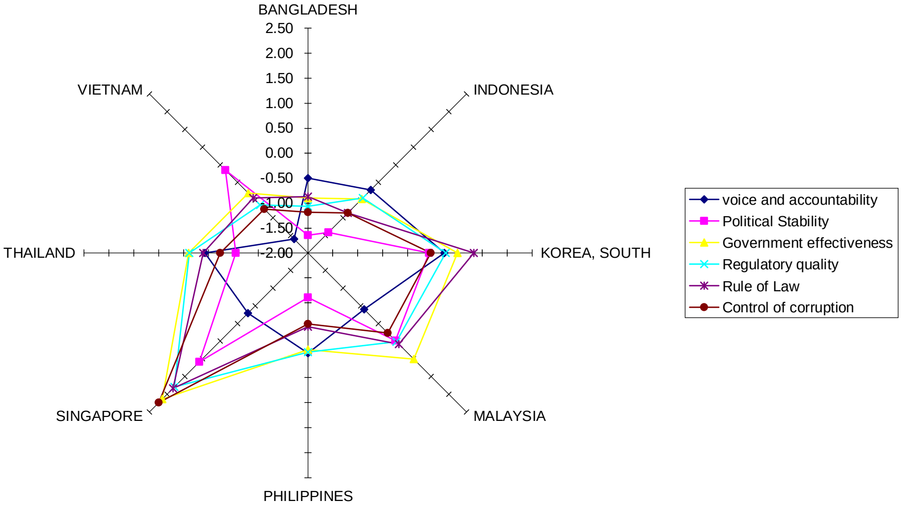
Pie Chart

A chart having a circular shape with subdivisions indicating the proportion of the component to the whole. Note that each percent is equivalent to 3.6 degrees.

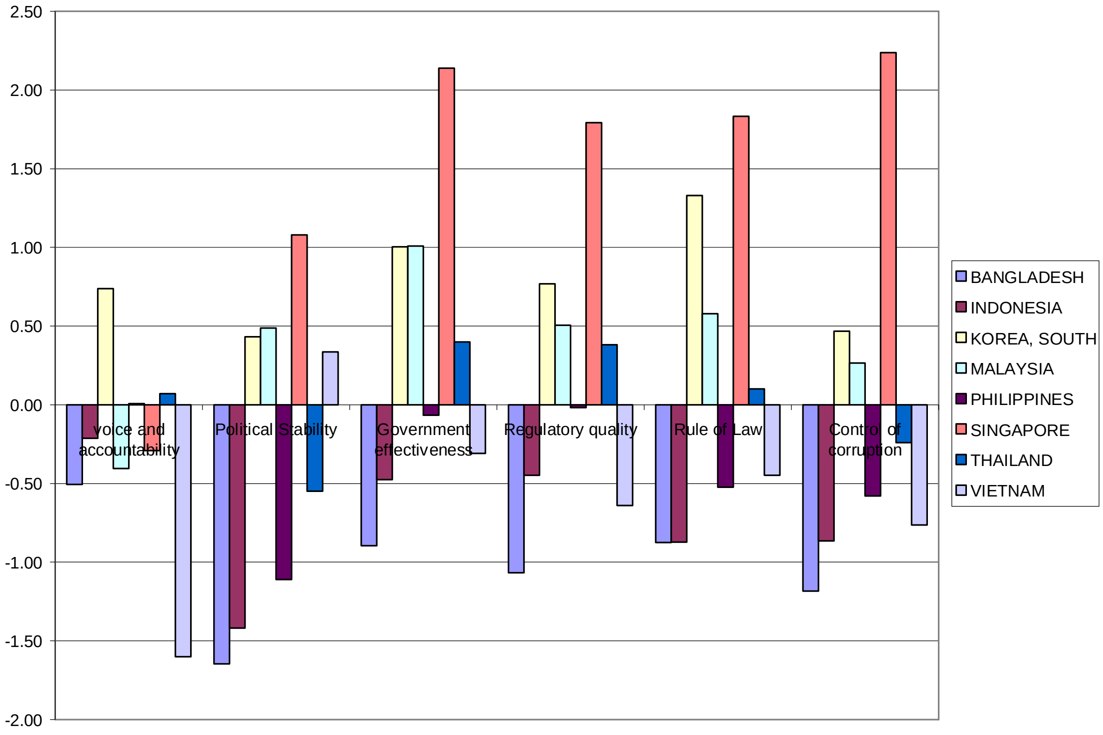


Web Chart

Increasingly becoming popular is the use of a web chart in illustrating data. This is equivalent to a 3-D graph as in the following where governance ratings of different countries on different dimensions ratings are plotted.



There is, however, the inherent difficulty of the naked eye to discern 3-D qualities on a two-dimensional medium like the paper. Consider the greater ease presented by the following representation of the same data:



3. Relationships Between Variables

Types of Relationship

Very often in practice, a relationship is found to exist between two (or more) variables. A mathematical form of the relationship can be very useful especially when we want to influence the behavior of one. However, we need to distinguish between three types of relationships before attempting to model the relationship:

- symmetrical -neither variables affect the other
- reciprocal - both variables affect one another
- asymmetrical - one the variables (independent variable) affect the other (dependent variable)

Two variables are said to have a symmetrical relationship if they arise from a common cause. For instance, the increase in the number of moviegoers and the increase in demand for dentists.

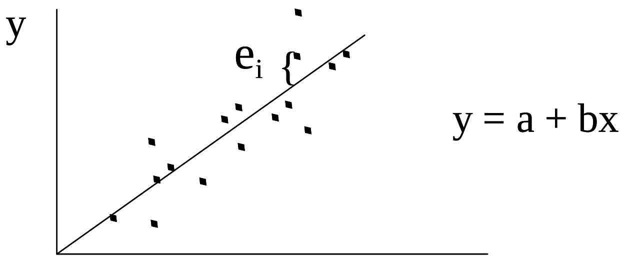
When the variables are interacting or reciprocal, it is not possible to identify the "cause" and the "effect". E.g., the relationship between literacy rate and growth.

An asymmetrical relationship assumes that one variable affects the other.

Regression Analysis

Regression analysis is the commonly used technique to model relationships between variables. Simple linear regression analysis models the linear relationship between one dependent and one independent variable. Multiple linear regression analysis, on the other hand, models the linear relationship between one dependent variable and a set of independent variables. There are other variants of the technique, including those that model nonlinear relationships, or those that analyze the relationships between two sets of variables (i.e., where there is more than one dependent variable).

Simple linear regression analysis finds a linear equation to explain the linear relationship between two variables, say x and y, such that the sum of the squared deviations of each observation from the respective predicted value is minimum.

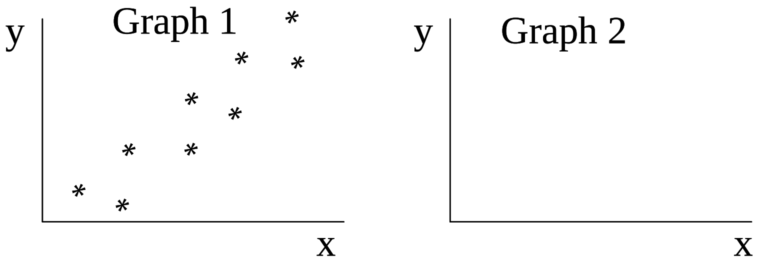


The usual form of the estimated linear regression model is given below:

$$y = a + b_1x_1 + b_2x_2 + ... + b_px_p$$

where a, the intercept is interpreted as the value of y when the independent variables $x_1, ..., x_p$ are set to zero; b_j is the expected change in y that will result from a unit change in x_j, all other x's remaining the same.

Whether or not a linear relationship exists between y and the x's is verified by the F value. The null hypothesis being tested is that the variability in y is purely random and the alternative is that it depends on the relationship between x and y. Consider the following graphs. Note that in graph 1, the values of y seem to differ according to the values of x. In contrast, in graph 2, the values of y appear to be random whatever is the value of x. We would then expect the F values to be very high (p very low) and very low (p very high) corresponding to graphs 1 and 2, respectively.



The strength of the linear relationship is given by the multiple correlation coefficient, R^2. This value gives the extent of variability in y that can be explained by the variability in x. It can happen that the R^2 is very low even when the F is sufficiently high. This means that while there exists a linear relationship, a substantial amount of the variability in y is still left unexplained by the variables considered.

Actually, the value of the R^2 increases with the number of independent variables. Therefore, the more appropriate statistic to consider is the adjusted or corrected R^2, which is simply the conventional R^2, corrected for the number of independent variables in the equation. If the additional independent variable has only a flimsy relationship with y, then we will actually observe the adjusted R^2 to decrease.

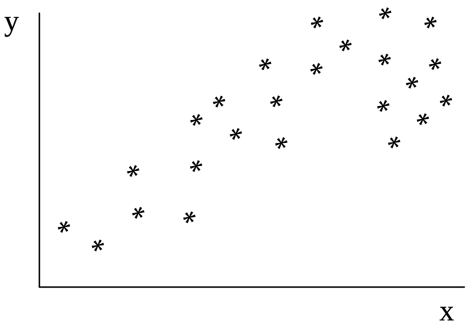
A Word of Caution

Be sure to verify the assumptions before utilizing the results of the regression analysis. The assumptions are that the error terms are independent of the x's; they are distributed normally about 0 and they have a common variance. The best way to do this is by exploring the scatterplots between y and the x's; between the error terms and the x's and the distribution of the error terms.

Following are some common anomalies. There are techniques for dealing with each one but they will not be discussed here.

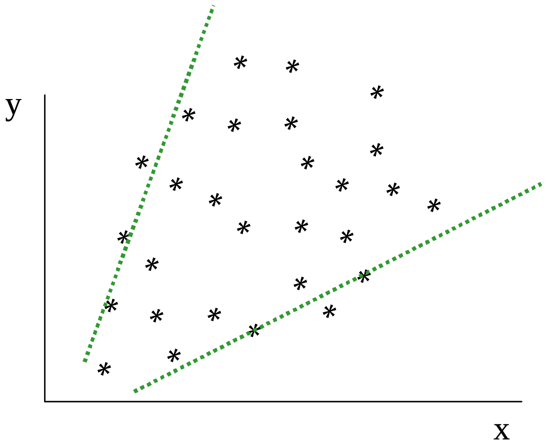
Nonlinear Relationship

In the graph shown, the x and y are related but in a nonlinear fashion. The regression result may show a very low, even insignificant F value. Hence, the linear regression analysis fails as a technique for modeling the relationship.



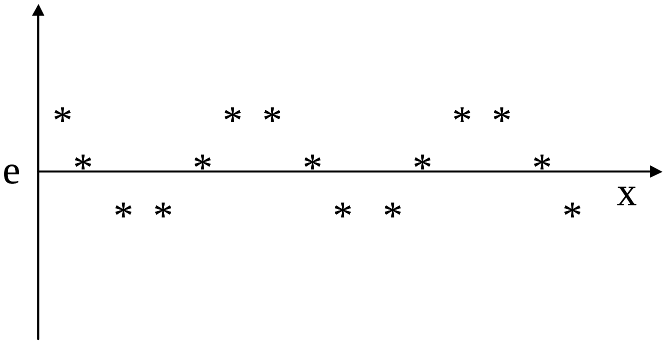
Heteroskedasticity

In the graph shown, we see that the variability of the y values depends on the x. As x becomes large, the range of values that y can assume becomes much wider.



Dependence Structure between X and the Error Term

In the graph below, we see that the values of the error term seem to be distributed in a systematic fashion depending on the value of x.



Nonnormality of Error Terms

If we plot the residuals, we should get an approximation of the normal distribution. If the deviation from the normal curve appears serious, then the proper adjustments need to be made.