2

MASTER

SYSTEM CONCEPT PAPER
FOR
EDIT AND AUDIT ENHANCEMENTS
TO THE FRS

DISCLAIMER

This book was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

leg

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

System Concept Paper for Edit & Audit Enhancements to the FRS

1. General Information

1.1 Summary

The purpose of the Financial Reporting System (FRS) is to establish a system of uniform segmented reporting for firms having different functional and management structures. As a result, patterns of activities and results can be considered and analyzed in an industry rather than firm-specific context.

In order to assure the quality and integrity of data collected from the respondent companies, various validation schemes may be employed in the data acceptance. Inter Systems, Inc. (ISI), under Task 1 of the contract, is required to review the current editing/auditing techniques utilized by EIA on the FRS, prepare an overall system concept paper for enhancing the automated review process, and develop specifications and output report layouts for special edit modules, ratios, and range checks, and predictive models that will be applied to the incoming submissions in support of the auditing effort.

1.2 Environment

System Sponsor: U.S. Department of Energy
Energy Information Administration

Developer: Office of the Financial Reporting System
12th Street & Pennsylvania Ave., NW
Washington, DC 20461

Computer Center: DOE/Forrestal Computer Facility
1000 Independence Ave., SW
Washington, DC 20585

Software: COBOL, INQUIRE, TSO, SUPERWYLBUR,
SAS, CCGS, PL1, FORTRAN

Hardware: IBM 3033 general purpose computer

1.3

References

- a) DOE/EIA Contract #DE-AC01-80EI10745
- b) Department of Energy
Energy Information Administration
Reporting System Clearance
1-12-79
Book 2
- c) Performance Profiles of Major Energy Producers 1977
DOE/EIA-0206 (77)
- d) DOE/EIA-28 Review Program 8/8/79 (Audit Handbook)
- e) Form EIA-28 Processing Manual by Cexec, Inc.
- f) Discriminate Function (DIF) Handbook Doc 6588 (Rev. 11-79)
- g) 'Probes of the Technological Future'
by North & Pike
Published by Harvard Business Review 1969
- h) 'Business & Economic Forecasting'
by Spencer, Clark & Hoguet
Published by Richard D. Irwin 1961
- i) 'Introduction to Business Statistics'
by Hadley
Published by Holden-Day, Inc.
- j) 'Less Risk in Inventory'
by Brown
Published by Harvard Business Review 1959
- k) 'Time Series Analysis, Forecasting & Control'
by Box-Jenkins
Published by Holden-Day, Inc.

- l) 'Time Series Forecasting'
by McLaughlin & Boyle
Published by Marketing Research Technique
- m) 'Basic Statistics with Business Applications'
by Clelland, deCani, Brown, Bush & Murry
Published by John Wiley & Sons, Inc.
- n) 'Macro-economic Activity: Theory, Forecasting & Control'
by Evans
Published by Harper & Row Publishers, Inc.
- o) 'Input-Output Economics'
by Leontief
Published by Oxford University Press
- p) 'Discussion Paper #138'
by Evans & Preston
Wharton School of Finance & Commerce
The University of Pennsylvania
- q) 'Harvard Business Review on Management'
Published by Harper & Row

2. Overview

2.1 Background

The requirement for FRS is based on Section 205 of Public Law 95-91 which established the Department and the EIA. The FRS collects data on revenues, capital expenditures, and operations of energy companies. A major emphasis of the FRS program is to assure the quality and integrity of the data collected. This is done in three different ways. First, the form (EIA-28) on which FRS data are reported is designed so that certain key elements correspond to publicly reported and audited numbers. The disaggregated data called for by the FRS EIA-28 forms are carefully checked to see that they do indeed relate back to key aggregated items already reported publicly by the respondents. Second, the FRS forms which are quite complex, are designed with numerous interrelated schedules to permit

an array of automated cross-checks and reviews for internal consistency. Finally, individual company data are manually compared with those of other companies, groups of companies or with related data from other sources. Significant deviations are then subject to special review.

During the initial phase, the FRS program collected a wide array of financial and operating data from 27 major energy companies for the year 1977. The group was chosen from the top 50 publicly owned domestic crude oil producers in 1976. The producers had at least one percent of domestic production or reserves of oil, natural gas, coal, or uranium, or at least one percent of refining capacity or petroleum product sales. Data collection for 1978 and for the years 1974-1976 is now essentially completed.

The Office of Management and Budget has approved renewal of the FRS reporting forms to permit collection of data for 1979 and 1980 and data for 1979 began to arrive in August 1980. In subsequent phases of the program, data will be collected from an additional 150-200 coal producers, uranium producers, refiners, marketers, and other energy companies as necessary to produce an accurate industry profile.

The FRS now contains 40 separate data bases which reside on an IBM system located at the EIA Computer Facility. From these data bases, over 250 batch output reports are generated which summarize the financial and operational characteristics of the respondents. The data bases have been loaded into the INQUIRE information retrieval and reporting system from which the FRS analysts develop ad-hoc reports.

2.2 Objectives

Review the current 40 schedules and 140 pages of incoming submissions to insure that the existing computerized edit and audit checks are appropriate. On the basis of that review, a system concept paper and general specifications will be developed for special edit modules that process data from prior year submissions of the respondent group to establish range tests and ratio tests that are statistically based and that can be adjusted via the introduction of a parameter. Additionally, specifications will be developed for a set of programs that will select and store values for approximately 300 ratios or range checks that can be applied to submissions and development of specifications for the associated output reports. Deliverables from this task will be a system concept paper and specifications for edit and audit enhancements.

2.3 Existing Methods and Procedures

Conceptual Edit/Audit Software

When the completed EIA-28 schedules are returned to the Department of Energy, the Office of Data Collections and Survey Operations reviews the forms for general consistency and completeness. After the returned forms have been logged in, Data Collection personnel screen the schedules for coding and format errors, and then send the schedules to be keypunched. Data entry is performed directly from the EIA-28 forms. The records are then processed through an edit program. Among the edits performed on the data are code validations, range checks, cross-footings, and down-footings.

Data with keypunch or minor format errors are corrected and reprocessed through the edit program. The edited data are processed through an audit program which checks the data on an inter-schedule basis, performs field-relational validations, and validates the data against some external sources. Valid data is then loaded onto INQUIRE data bases from which it is retrievable by the FRS office.

The edit and audit modules generate files of valid data, files of erroneous data, and an edit/audit report. Numerous analytical reports and dumps have been generated from the INQUIRE data bases.

FRS Review Program

The FRS Review Program has been developed to provide the Energy Information Administration (EIA) with a method of evaluating the reliability of the information collected by Form EIA-28. The review procedures are segregated into two phases: 1) desk review procedures, which include a series of analytical and mechanical functions to be performed at EIA, and; 2) on-site review procedure, a review and discussion of desk review findings and other appropriate interviews. As of August 1980, no on-site reviews have been conducted.

The Form EIA-28 is designed to depict a respondent's operations as separate lines of business as defined by the instructions. Each segment is intended to be treated as a separate entity entering into transactions with other segments and third parties.

The final product of the desk review audit is a "Review Control Sheet".

An example of some of the data review audit procedures are as follows: Note: See the Audit Handbook, p.4.

- . Disaggregated data elements are used to calculate totals which are then checked against aggregated data items reported by the companies.

For example:

#7 Review schedule 5100 for:

- a) Officer certification.
- b) Revisions in address, names, IRS ID number, etc.
Update mailing files.
- c) Completeness.

#8 Agree all amounts in column (A) of schedules 5110, 5120, and 5130 to the consolidated financial statements in the 10-K. Exhibit B should provide reconciliations between the 10-K statement line items and the EIA-28 line items on each schedule.

- . As the numerous reporting schedules are interrelated, a series of cross-checks are performed to ensure of the integrity and consistency of the data across various schedules. Note: See the Audit Handbook, p. 19.

For example:

On determination of Income Taxes on Schedule 5115 -

- a) Where ITC is generated on 5115, determine that capital expenditures are noted on schedule 5130/D.
  - b) Column A of schedule 5115 should be compared to Schedule 5246 (petroleum), 5341 (coal) and 5441 nuclear to ascertain there was current year production when excess depletion is noted on Column A and vice versa.
  - c) Compare column D to the segmental income statements (schedules 5210, 5310, 5410, 5510, and 5610) to ascertain that foreign operations exist where tax credits have been generated and vice versa.
- . The current years opening balance is compared against prior-year closing balance for the same company to ensure year to year consistency in reporting.
  - . Data concerning financial activities of public companies are also available through other sources and are used to verify the information reported on the FRS schedules.

For example:

- . On determination of Petroleum - Property, Plant, and Equipment on Schedule 5222 -
  - 1. Schedule 5222 reconciles changes in net property and other capitalized cost balances in a format similar to the SEC (Securities and Exchange Commission) Form 10-K Schedules V and VI.

- a) Relate balances in column A and H to the net beginning and ending balances of Schedules V and VI.
- b) Relate additions, per column B, to Schedule V additions.
- c) Relate disposals, per column D, to the net dispositions per V and VI.

Audit Support Program

The Audit Support Program is a new addition to the FRS software and replaces or supplements many of the ratio-developed-limit-tests, which had been used by the FRS Review Program. The ratios which are developed by this new software program are used for company level limit testing.

2.4 Proposed Methods and Procedures

This segment of the document presents the system concepts to enhance the edit/audit modules and to ensure the integrity of the data contained within the 40 separate data bases. This requires reviewing the current 40 schedules and 140 pages of incoming submissions.

a) Trend Projections

The data obtained by the submissions consists of those variables to which trend analysis may be easily applied. Trend projection is a technique which fits a trend line to a mathematical equation.

There are several math formulas within the trend projection technique of forecasting, which can fit the DOE/EIA requirements for FRS ratio development. An example of these are slope-characteristic, method of least squares, polynomials, logarithms, and so on.

A visual representation of the use of the method of least squares is shown in Figure 1.

Example of Trend Projection Using the Method of Least Squares

FIGURE 1

This math formula is a deviation of the straight line equation which is written as

$$y = \beta_0 + \beta_1 x$$

where β_0 is the y-intercept and β_1 is the slope of the line. The y-intercept of β_0 is the value of y when $X=0$. The slope β_1 is the change in y for a 1-unit change in X. Thus the calculation for the straight line deviations (denoted as $\hat{y}$ rather than y) is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ represent sample estimates of β_0 and β_1 . The Least Square Estimators formulas of

β_0 and β_1 are

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \quad \text{and} \quad \hat{\beta}_0 = \hat{y} - \hat{\beta}_1 \bar{x}$$

with SS_{xy} = the cross product summation.

This group of formulas will minimize the deviations of the measurements and allow the line to be "best-fitted".

The straight line that is created is better known as the regression line and may be used as the input vehicle to an extrapolation calculation in order to project that line into the future.

A second method that might be used in the development of the ratios may be the use of a scattergram. The primary purpose of a scattergram is to provide a descriptive picture of the sample when two variables are measured.

Reference Figure 2.

FIGURE 2

The popular numerical measure of the strength of the relationship between two variables is the correlation coefficient. The correlation coefficient provides a measure of the linear (straight-line) relationship between two variables. That is, when two variables have a positive relationship (y tends to increase as X increases) or a negative relationship (y tends to decrease as X increases), the correlation coefficient will reflect this.

It is computed by

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

where

$$SS_{xy} = \sum (x - \bar{x})(y - \bar{y}) = \frac{\sum xy - (\sum x)(\sum y)}{n}$$

$$SS_{xx} = \sum (x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$SS_{yy} = \sum (y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n}$$

This formula assures us of always obtaining a number which is between -1 and +1 no matter what the scale of the variables.

The accuracy of forecasting using trend projections is very good for short term (0 to 3 months) forecast and good for medium term (3 months to 2 years) forecast, per Harvard Business Review on Management.

The complexity of the various mathematic equations varies from little expertise required to moderately complex when using the Box-Jenkins technique as a reference point.

b) Ranges

Ranges are developed through an analysis of the ratios frequency of distribution. In order to determine the frequency of distribution a number of chronological steps must be taken.

A histogram is built and smoothed forming the bell-shaped distribution (also known as a normal distribution curve). Following this, identifying the mode (i.e., the single value that best characterizes the entire distribution of measurements) and identifying the quartiles (i.e., the upper and lower limits of the range). To determine the percentage of the population that falls within the quartiles a standard deviation calculation must be done. This formula is

$$S = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}}$$

where

y = the measurement.

n = number of measurements.

The results of this calculation will allow us to state a rule that describes accurately the variability of a bell-shaped curve distribution. This rule is called the Empirical Rule. The Empirical Rule states that for a distribution of measurements that is approximately bell-shaped, the interval

1. $(\bar{y} \pm S)$ will contain 68% of the measurements.
2. $(\bar{y} \pm 2S)$ will contain 95% of the measurements.
3. $(\bar{y} \pm 3S)$ will contain all or almost all of the measurements.

See figure 3 for an example of frequency of distribution as shown on an normal distribution curve.

Example of Industry Measurements for Domestic Refining
Average Price Per Barrel of Crude

FIGURE 3

c) Limit Tests

Limit tests are very straight forward. Data variables subject to limit testing should contain, as part of their software definitions, the upper and lower limits against which they will be compared. Variable values outside these limits should be flagged as potential errors.

Note that reasonableness checks should be applied to data elements and not to data variable groups. Group checking loses some of its accuracy when off-setting changes occur. In addition, since each data element is subject to limit tests, discrepancies will be noted individual items and, therefore, the components of a group will be checked.

Note also that for trend checking using either standard average values or weighted average values, the computed percent change should be compared in absolute value to the specified percent related to the data variable.

Frequent exceptions for the same data variables may suggest a change in the specified limits. These changes should be researched and approved before they are made.

d) Forecasting Methods

Due to the erratic and sometimes volatile changes in the petroleum industry an accurate method of determining both short and long term standards must be developed. To handle the increasing complexity of the petroleum industry many forecasting techniques have been developed.

The selection of a method depends on many factors - the context of the analysis, the relevance and availability of historical data, the degree of accuracy desirable, the period to be forecast, the cost/benefit (or value) of the forecast, and the time available for making the analysis.

There are three basis types of analysis - qualitative techniques, time series analysis and projection (progression/regression), and causal models. The first uses qualitative data (expert opinion, for example) and information about special events of the kind already mentioned, and may or may not take the past into consideration.

The second, on the other hand, focuses entirely on patterns and pattern changes, and thus relies entirely on historical data.

The third uses highly refined and specific information about relationships between system elements, and is powerful enough to take special events formally into account. As with time series analysis and projection techniques, the past is important to causal models.

Qualitative Techniques

The data on the FRS data bases, for the years 1974 through 1976 is only partial when compared to that of the more recently collected years, i.e., 1977. A qualitative technique may be applicable in part for use in the development of the FRS standards.

Qualitative techniques are used when data are scarce, for example when only partial historical data are available. They use human judgment and rating schemes to turn qualitative information into quantitative estimates.

The objective here is to bring together in a logical, unbiased, and systematic way all information and judgments which relate to the factors being estimated. Such techniques are frequently used in new-technology areas, or areas that have little or no historical data. A sample of these techniques is as follows:

1) Delphi Method:

A panel of experts is interrogated by a sequence of questionnaires in which the responses to one questionnaire are used to produce the next questionnaire. Any set of information available to some experts and not others is thus passed on to others, enabling all experts to have access to the information for forecasting. This technique eliminates the bandwagon effect of majority opinion.

2) Panel Consensus:

This technique is based on the assumption that several experts can arrive at a better forecast than one person. There is no secrecy, and communication is encouraged. The forecasts are sometimes influenced by social factors, and may not reflect a true consensus.

3) Visionary Forecast:

A prophecy that uses personal insights, judgment, and, when possible, facts about different scenarios of the future. It is characterized by subjective

guesswork and imagination; in general, the methods used are non-scientific.

Any of these three methods might be used by DOE/EIA personnel in the determination of specific standards and the ranges for those standards though the accuracy would be questionable.

Time Series Analysis

In general all statistical techniques are based on the assumption that existing patterns will continue into the future. This assumption is more likely to be correct over the short term than it is over the long term, and for this reason these techniques provide us with reasonably accurate forecasts for the immediate future but do quite poorly further into the future (unless the data patterns are extraordinarily stable). A sample of these techniques are as follows:

1) Moving Average:

Each point of a moving average of a time series is the arithmetic or weighted average of a number of consecutive points of the series, where the number of data points is chosen so that the effects of irregularity is eliminated.

This method is very similar to trend analysis but its accuracy of forecasting is rated poor for both short (0 - 3 months) and medium (3 months - 2 years) term forecasting.

2) Exponential Smoothing:

This technique is similar to the moving average, except that more recent data points are given more weight. Descriptively, the new forecast is equal to the old one plus some proportion of the past forecasting error. Adaptive forecasting is somewhat the same except that seasonals are also computed. There are many variations of exponential smoothing: some are more versatile than others and may be applicable to the FRS standards requirements.

3) Box-Jenkins:

Exponential smoothing is a special case of the Box-Jenkins technique. The time series is fitted with a mathematical model that is optional in the sense that it assigns smaller errors to history than any other model. This is apparently the most accurate statistical routine presently available.

4) X-11:

Developed by Julius Shiskin of the Census Bureau, this technique decomposes a time series into seasonals, trend cycles, and irregular elements. Primarily used for detailed time series analysis (including estimating seasonals); but he has extended its uses to forecasting and tracking and warning by incorporating other analytical

methods. Used with special knowledge, it is perhaps the most effective technique for medium range forecasting (3 months - 2 years).

Causal Models

A causal (cause/effect) model is the most sophisticated kind of forecasting tool. It expresses mathematically the relevant causal relationships, and may include pipeline considerations (i.e., inventories) and market survey information. It may also directly incorporate the results of a time series analysis. A sample of the techniques is as follows:

1) Regression Model:

This technique functionally relates sales to other economic, competitive, or internal variables and estimates an equation using the least-squares technique. Relationships are primarily analyzed statistically, although any relationship should be selected for testing on a rational ground.

2) Econometric Model:

An econometric model is a system of interdependent regression equations that describes some sector of economic sales or profit activity. The parameters of the regression equations are usually estimated simultaneously. Due to the system equations inherent in such models, they will better express the causalities involved than an ordinary regression equation and hence will predict turning points more accurately.

3) Input-Output Model:

A method of analysis concerned with the interindustry or interdepartmental flow of goods or services in the economy or a company and its markets. It shows what flows of inputs must occur to obtain certain outputs. Considerable effort must be expended to use these models properly and additional detail, not normally available, must be obtained if they are to be applied to specific business.

4) Economic Input-Output Model:

Econometric models and input-output models are sometimes combined for forecasting. The input-output model is used to provide long-term trends for the econometric model; it also stabilizes the econometric model.

5) Discriminant Function Model:

The discriminant function technique developed and used by the Internal Revenue Service, may be best described as a scoring system. The objective of the system is to maximize the differentiation between two sub-populations, i.e., "need-to-examine" sub-population of returns (those with significant tax change) and the "no-need-to-examine" sub-populations of returns.

(DIF formulas are usually developed from reported line item data collected during a TCMP (Taxpayer Compliance Measurement Program) survey and certain data from the Master File).

There are three prerequisites that must be met before discriminant function systems can be developed, namely: (1) availability of experienced technical personnel in the operations research or the mathematical statistical area; (2) existence of a body of scientifically collected, unbiased and detailed data on reported and corrected (after examination) FRS submission information (IRS requires a sample of approximately 5,000 returns in an examination class for DIF development); and (3) computerization of returns processing. If these three basic conditions can be met, a minimum of six months of intensive research work is needed to develop reliable discriminant function formulas. Then additional time is required for objective and detailed evaluation of the formulas, and for the design of procedures and computer programming for system installation. Once these prerequisites have been met, the essential steps in this process is as follows:

- a) For each examination class, define the two sub-populations of returns to be differentiated.
- b) For each examination class, breakdown each of the reported line items into intervals of value or attributes developed through distribution analysis of the line items.
- c) Develop a DIF formula by examination class from those return characteristics that exhibit the most discrimination capability.
- d) Develop optimal weights for the intervals of each of the return characteristics that exhibit the most formula.
- e) Calculate a DIF score for each return. This score is the algebraic sum of the assigned weights.
- f) Measure the effectiveness of each of the DIF formulas by its productivity.

By using such models DOE/EIA could identify company data which are significantly at variance with established averages. The models will also be able to show, through analysis of historical data, that the variance is a result of significantly changed operational factors, of reporting errors, or purely random variation.

Conclusion

To put in proper perspective the requirements that must be met in the development of the System Specification Paper for Edit and Audit Enhancements, the following must be done in chronological order:

- 1) Properly identify approximately 300 key data element fields. These fields will be used in the development of the ratios. These key fields may be those which are utilized in the company's fiscal report or may be the same selected information requested by DOE/EIA for the years 1974, 1975 and 1976.
- 2) Determine the methodology to be used in the development of the standards (i.e., trend projections, exponential smoothing, etc.). This is the most important segment of the entire task. If the wrong methodology is chosen, it will be reflected in that it must be constantly modified to reflect the volatile changes in the petroleum industry.
- 3) Develop the ratios using the approved method.
- 4) Establish the ranges for the ratios using a standard deviation formula.
- 5) Implement the ratios and ranges into the FRS (system).

Realizing that this is looked upon as a pilot project and when the DOE/EIA personnel have gained sufficient expertise in the handling of the ratios and ranges, we recommend either the Box-Jenkins or X-11 techniques. Either of these products will answer the long term requirements of FRS.

In order to answer to the near term requirements as defined in Task I, using the limited data that is available (i.e., partial data on years 1974, 1975 and 1976), while adhering to DOE/EIA request for non-complexity and ease of implementation, we recommend the use of a group of trend projection techniques to identify the ratios and build the ranges for those ratios. Reference Figure 6.

These programs will be parameter driven, allowing the user to increase or decrease the number of ratios that the user desires. This will also allow the user the flexibility of expanding or reducing the ranges, thereby fine tuning the system.

FLOWCHART WORKSHEET

SYSTEM _____ DATE _____
PROGRAM _____ PROGRAMMER _____ PAGE _____ OF _____

Figure 6

PRODUCT NO. 5910

GLOSSARY OF COMMON STATISTICAL TERMS

Analysis of variance: Analysis of data based on a partitioning of the sum of squares of deviations of the y-values (response measurements) about their mean.

Biased estimator: Estimator whose probability distribution has a mean value that is not equal to the value of the estimated parameter.

Central Limit Theorem: Theorem states that the probability distribution of the sample mean (or sum) will be approximately normal when certain conditions are satisfied.

Chi-square: Test statistic used to test the null hypothesis of independence for the two classifications of a contingency table.

Class boundary: Dividing point between two cells in a frequency histogram.

Classes: Cells of a frequency histogram.

Class frequency: Number of observations falling in a class (referring to a frequency histogram).

Coefficient of linear correlation: Measure of linear dependence between two random variables.

Completely randomized design: Consider an experiment that involves k different ways of treating a set of experimental units so that k treatments can be compared. A completely randomized design assigns the experimental units to the treatments in such a way that any one allocation of experimental units to the treatments is just as probable as any other.

Confidence coefficient: Probability that an interval estimate (a confidence interval) will enclose the parameter of interest.

Confidence interval: Concerned with interval estimation. Two numbers, computed from sample data, that form an interval estimate for some parameter.

Contingency table: Two-way table constructed for classifying count data. The entries in the table show the number of observations falling in the cells. The objective of an analysis is to determine whether the two directions of classification are dependent (contingent) upon one another.

Degrees of freedom: Parameter of a measurements t and the chi-square probability distributions. Degress of freedom measure the quantity of information available in normally distributed data for estimating the population variance σ^2

Design of an experiment: Plan for collecting experimental data. The objective is to acquire a specified quantity of information (pertinent to a population parameter) at minimum cost.

Deviation from the mean: Distance between a sample observation and the sample mean $\bar{y}$.

Discrete random variable: Random variable that can assume only a finite number or a countable infinity of values.

Empirical Rule: Rule that describes the variability of data that possess a mound-shaped frequency distribution.

Error of estimation: Distance between an estimate and the true value of the parameter estimated.

Estimate: Number computed from sample data used to approximate a population parameter.

Estimator: Rule that tells how to compute an estimate based on information contained in a sample. An estimator is usually given as a mathematical formula.

Event: Outcome of an experiment.

Experiment: Process of making an observation.

Experimental units: Objects upon which measurements are taken.

Factorial experiment: Experiment that includes all combinations of the levels of the factor.

Factors: Independent experimental variables that are related to a dependent response variable.

Frequency: Number of observations falling in a cell or classification category.

F statistic: Test statistic used to compare variances from two normal populations. Used in the analysis of variance.

Histogram: Graphical method for describing a set of data.

Interaction (factor interaction): When the effect of one factor on a response depends on the level(s) of one (or more) other factors.

Interval estimate: Two numbers computed from the sample data. The interval formed by the numbers should enclose some parameter of interest. An interval estimate is usually called a confidence interval.

Interval estimator: A rule that explains how to calculate from sample data the two numbers required for an interval estimate.

Least squares: Method of curve fitting that selects as the best-fitting curve the one that minimizes the sum of the squares of deviations of the data points from the fitted curve.

Level of factor: Intensity setting.

Level of significance: Refers to the outcome of a specific statistical test for an hypothesis. The level of significance the test is the probability of drawing a value of the test statistic that is as contradictory, or more contradictory, to the null hypothesis as the value observed.

Linear correlation: Dependence between two random variables.

Lower confidence limit: Smaller of the two numbers that form a confidence interval.

Matched pairs test: Statistical test for the comparison of two population means. The test is based on paired observations, one from each of the two populations.

Mathematical model: Mathematical equation that characterizes (or models) the behavior of one or more variables associated with a natural phenomenon.

Mean: Average of a set of measurements. The symbols $\bar{y}$ and μ denote the means of a sample and a population, respectively.

Median: Middle measurement when a set is ordered according to numerical value.

Multinomial experiment: Experiment that consists of n identical independent trials, where each trial can result in one of k (a fixed number) outcomes.

Nonparametric methods: Usually refers to statistical tests of hypotheses about population probability distributions - but not about specific parameters of the distributions.

Normal distribution: Bell-shaped probability distribution. The curve possesses a specific mathematical formula.

Null hypothesis: Hypothesis under test in a statistical test of an hypothesis.

Parameter of a population: Numerical descriptive measure of a population.

Parametric methods: Statistical methods for estimating parameters or testing hypotheses about population parameters.

Point estimate: See Estimate.

Population: Set of measurements, actual or conceptual, that is of interest to the experimenter. Samples are selected from the population.

Probability: As a practical matter, we think of the probability of an event as a measure of one's belief that the event will occur when the experiment is conducted once. The exact definition, giving a quantitative measure of this belief, is subject to debate. The relative frequency concept is most widely accepted.

Probability distribution, continuous: Smooth curve that gives the theoretical frequency distribution for the continuous random variable. An area under the curve over an interval is proportional to the probability that the random variable will fall in the interval.

Probability distribution, discrete: Listing, mathematical formula, or histogram that gives the probability associated with each value of the random variable.

Qualitative factor: Factors that are not quantitative (see quantitative factor).

Quantitative factor: One that can take values corresponding to points on a real line.

Randomized block design (containing p treatments and b blocks of p experimental units each): The treatments are randomly assigned to the units in each block, with each treatment appearing exactly once in every block.

Random sample: Sample of n measurements selected in such a way that every different sample of n elements in the population has an equal probability of being selected.

Random variable: Random variable is associated with an experiment. Its values are numerical events that cannot be predicted with certainty.

Range of a set of measurements: Difference between the largest and smallest members of the set.

Regression line: Line fit to data points using the method of least squares.

Rejection region: Set of values of a test statistic that indicates rejection of the null hypothesis.

Relative frequency: Class frequency divided by the total number of measurements.

Research hypothesis: Hypothesis to be accepted if null hypothesis is rejected; sometimes called the alternative hypothesis.

Sample: Subset of measurements selected from a population.

Significance level: See Level of Significance.

Sign test: Nonparametric statistical test used to compare two populations.

Standard deviation: Measure of data variation.

Standardized normal distribution: Normal distribution with mean and standard deviation equal to 0 and 1, respectively. The standardized normal variable is denoted by the symbol z .

Test statistic: Function of the sample measurements, used as a decision maker in a test of an hypothesis.

Treatment: Specific combination of factor levels.

Type 1 error: Rejecting the null hypothesis when it is true.

Unbiased estimator: Estimator that has a probability distribution with mean equal to the estimated parameter.

Upper confidence limit: Larger of the two numbers that form a confidence interval.

Variance: Measure of data variation.

z statistic: Standardized normal random variable that is frequently used as a test statistic.