IC 233

"If a man will begin with certainties, he shall end in doubts; but if he will be content to begin with doubts, he shall end in certainties."

Francis Bacon

GUIDE TO ENGINEERING STATISTICS

by

J. Visman and Jacqueline L. Picard

Western Regional Laboratory (Edmonton)
METALS REDUCTION AND ENERGY CENTRE

Information Circular IC 233

A Contribution to Scientific Knowledge in Canada

June 1970

FOREWORD

Most of the research investigations undertaken by scientists and engineers in the Mines Branch require the use of statistical analysis to reach valid conclusions from necessarily limited observed or experimental data. This practical handbook, written especially for the technical research user of statistics, gathers in a simplified form many of the most useful and powerful statistical techniques, usually found in a number of different textbooks, generally written from the statistician's viewpoint. Although the senior author, Dr. Jan Visman, is an internationally recognized authority on sampling statistics, this Guide was prepared from the non-specialist viewpoint and should be very useful not only for the Mines Branch research scientists but also for other engineers and research workers in Canada and elsewhere.

John Convey

John Convey,
Director

Ottawa, June 1970

AVANT-PROPOS

Pour la plupart des recherches entreprises par les ingénieurs et scientifiques à la Direction des mines, l'emploi d'analyses statistiques est requis pour atteindre des conclusions valables à partir de données expérimentales ou observées nécessairement limitées. Ce manuel pratique, écrit spécialement pour le chercheur technique utilisant la statistique, rassemble sous une forme simplifiée de nombreuses techniques statistiques parmi les plus utiles et puissantes, qu'on trouve ordinairement dans plusieurs livres différents, généralement écrits du point de vue du statisticien. Bien que le principal auteur, le Dr. Jan Visman, ait une réputation internationale en statistique d'échantillonnage, ce Guide a été préparé du point de vue du non-spécialiste, et devrait être très utile, non seulement aux chercheurs scientifiques de la Direction des mines, mais aussi à d'autres ingénieurs et chercheurs au Canada et ailleurs.

John Convey

John Convey,
Directeur

Ottawa, juin 1970

Mines Branch Information Circular IC 233

GUIDE TO ENGINEERING STATISTICS

by

J. Visman* and Jacqueline L. Picard**

ABSTRACT

This text provides guidelines for the selection and the application of statistical techniques that are commonly used in science and industry.

The emphasis is on how to solve statistical problems and, by conveying the basic concepts of variability, to prepare the reader for further self-study of textbooks in his or her particular field.

Instructions in the form of a Summary of Operations, presented in Section 1, are recommended to those readers for whom the application of statistical analysis is not a daily routine. A tabular listing of statistical problems and procedures provides a short-cut to the practical application of techniques. A general sampling theory for segregated populations is introduced, with condensed instructions that cover most of the variates.

Definitions of terms and symbols are presented in an appendix preceding the alphabetic register of subjects.

Many techniques in this guide can only be applied legitimately for calculating first-order estimates of a variance, a probability, a ratio, etc. For more critical situations where specific conditions - too complicated to be mentioned here - have to be satisfied, the reader is well advised to obtain the assistance of a professional statistician. Between this high level of perfection and that of the "educated guess" there is scope for a guide to statistics which it is hoped this volume will provide for its readers.

* Head and ** Technical Officer, Western Regional Laboratory, Metals Reduction and Energy Centre, Mines Branch, Department of Energy, Mines and Resources, Edmonton, Alberta.

Direction des mines

Circulaire d'information IC 233

GUIDE À LA STATISTIQUE TECHNOLOGIQUE

par

J. Visman* et Jacqueline L. Picard**

RÉSUMÉ

Ce texte a l'objet de servir de guide à la sélection et à la mise en oeuvre de techniques statistiques qui s'utilisent souvent dans les divers domaines des sciences et de l'industrie.

Il s'agit ici surtout de souligner la manière par laquelle se résolvent les problèmes statistiques. Par ailleurs, ce guide servira de préparatif à l'étude de la statistique dans le domaine particulier du lecteur en lui donnant des notions élémentaires de la variabilité.

Aux lecteurs pour lesquels l'utilisation de l'analyse statistique n'est pas une pratique journalière, la méthode est présentée sous forme de mode opératoire. Un résumé de divers problèmes et de procédés statistiques en forme de tableau sert de raccourci pour l'emploi de ces techniques. Une théorie générale de l'échantillonnage pour les populations ségrégées est présentée avec un précis de la technique qui traite de la plupart des variates. La définition des termes et caractères se trouve dans un appendice qui précède la table alphabétique des matières.

Les techniques dont on parle ne peuvent légitimement être utilisées que pour le calcul d'estimations de premier ordre, par exemple d'une variance, d'une probabilité, d'une proportion, etc. Lorsqu'il s'agit de situations plus difficiles où nous devons satisfaire à certaines conditions précises, trop compliquées pour être discutées ici, le lecteur devrait bien obtenir l'aide d'un statisticien. Entre ce niveau élevé de perfection et celui du jugement pratique, il y a de la place pour un guide à la statistique et c'est ce que nous espérons avoir ici fourni au lecteur.

* Chef et ** Agent technique, Laboratoire régional de l'ouest, Centre de l'énergie et de réduction des métaux, Direction des mines, ministère de l'Énergie, des Mines et des Ressources, Edmonton, Alberta.

CONTENTS

(CONTENTS, cont'd)

(CONTENTS, concluded)

TABLES*

* Note: Tables 1.3, 1.4, 4.1, 4.3, 4.6, 4.7, 4.8, 4.9, 4.11, 6.1, 6.2, 6.4 and 6.5, small and untitled, are interspersed in their respective sections.

FIGURES*

Appendix C - Charts of Statistical Functions

* Note: Figures 4.1, 4.2, 4.3, 6.1, 6.2, and 6.3, small and not captioned, are interspersed in their respective sections.

INTRODUCTION

For the professional worker engaged in research, development or investigational work, it is generally easy to recognize certain natural phenomena or processes as statistical problems. Experience shows, however, that the application of the large variety of statistical techniques and tests found in textbooks is commonly left to specialists. In view of the present scarcity of this high-priced skill, a large amount of statistical work is left undone where the need for it is recognized, because the investigator feels that he is inadequately trained to "think in probabilities" and lacks the time for sufficient study. The primary object of this Guide is to help bridge the gap. In the broad sense, it aims to assist in interpreting certain phenomena encountered in engineering research and industrial production as statistical problems, by first formulating the problem, then investigating it by experiment and, finally, analyzing the data and interpreting the results in a meaningful way.

As the need for interpretation of statistical data is becoming more general with the proliferation of computers in industry, so is the need of those involved with industrial problems for a guide in the application of engineering statistics.

The sequence of operations may be subdivided into five steps, starting with the definition of the problem. This requires, first of all, a detailed knowledge of the <u>natural laws</u> underlying the process or phenomenon, including its actual and theoretical behaviour, and secondly, a knowledge of the <u>quantitative</u> aspect,

i. e. the magnitude of the factors involved and their ranges. The first part is basically of a non-statistical nature. It establishes the fundamental relationship and its scientific correctness or conformity to truth, and is primarily a qualitative appraisal. The second part, on the other hand, is concerned with experiment and induction. It provides estimates of the observed factors or phenomena and of their significance in relation to the sum total of all the unobserved factors and chance variations that might occur.

The important point to remember here is that statistics is a formal logic only, incapable of proving or disproving the truth. Its value depends entirely on the correctness or reality of the premises which lie at the root of the relationship that is being tested by statistical means.

The most commonly made mistake is that statistical relationships are taken at face value as representing the "facts", whereas in reality there might be no true relationship (spurious correlation). For instance, a statistical relationship was found to exist between the number of storks flying over East Prussia during a certain period and the number of babies born there during the same period. This is obviously a spurious correlation, but it might appear to a child as factual confirmation of what it has accepted as the truth, namely, that storks bring babies. Similarly, various controversies (e. g., regarding fluoridation and the causes of lung cancer, etc.) could originate from a fallacious or incomplete set of assumptions which are mistakenly identified with the truth.

Statistics can be defined as the science of the collection and organization of quantitative data according to relative frequency of occurrence as a basis for drawing valid conclusions.

As such, it is a highly efficient tool for the analysis of those natural phenomena and industrial processes whose true natures are obscured or masked by large variations.

The scholars of the Middle Ages provided the necessary foundation for the laws of nature in certain assumed theories and expressions of the Scriptures. However, they rarely if ever tested the results by experiment, as it was their maxim that "Reason is the Sovereign of Nature" and that, therefore, truth of the natural world as well as of the spiritual world must be derived from reason and authority. Their thinking was "Aristotelian", that is, essentially philosophical and of a qualitative nature. Since the Renaissance, however, the need for the collection of facts by observation and experiment has been recognized, and this principle, which has been called the "Galilean" approach, forms the basis for modern advancement in science.

Problems of a statistical nature involving a number of variables are often dealt with by non-statisticians in the conventional way of studying the effect of one variable at a time while trying to keep the other variables constant. This method is not efficient, since it requires unnecessary repetition of tests and is limited in the sense that only operating variables can be studied. As a rule, also, behaviour observed under these conditions will differ from that when all variables are operating simultaneously, an effect due to interaction between the variables.

The old method therefore has limited application, and any conclusions drawn are subject to the "ceteris paribus" proviso ("all other things being equal"). With modern statistical methods, these restrictions are eliminated and a maximum amount of information can be obtained with a minimum amount of work.

This Guide sets forth a procedure for the collection and interpretation of data. Detailed instructions are provided for solving the principal types of statistical problems encounttered in the field of engineering practice and research. The required sequence of operations is given in condensed form in Table 1.1, where the statistical procedure has been subdivided into five steps. For most problems, each step will have to be considered to some degree. Even for simple problems where only one or two of the steps may be required, it is considered worthwhile to go through the entire procedure in order to see it in its proper perspective.

1. GENERAL PROCEDURE

Table 1.1 - Summary of Operations

1. Defining a statistical problem

Analyze the actual and theoretical aspects of the problem and list the nature of variables involved. Formulate the physical or chemical relationship between the dependent and independent variables.

2. Premises

Assign the main independent variables. Choose the estimators for same. Estimate their relative magnitude and reputed range. Evaluate the residual variations (secondary factors and chance deviations, or errors, combined).

3. Experimental procedure

Design the testing technique. Determine the number of observations. Determine the size of the sample and number of increments. Collect the data.

4. Analysis of data

Make preliminary estimate of standard deviation from range. Round off the observations. Eliminate "tramp" values. Estimate missing data. Normalize the relationship by transformation of variables. Apply analysis of variance and tests of significance. Choose significant factors for correlation and regression analysis.

5. Correlation

Choose the appropriate formula for the relationship between dependent and independent variables. Determine the regression coefficients and constant. Find the correlation coefficient.

1.1 Defining a Statistical Problem

A statistical problem may be said to arise when data variations caused by factors other than those accounted for are too large to be ignored. For instance, the problem of providing quick estimates of the heat value of a certain type of coal can be solved by using the experimental relationship between the heat value and the ash content. This relationship is not exact, however, because the heat value depends not only upon the percentage of ash but also upon the percentage of moisture and combustible matter, and the composition of the latter. The efficacy of the experimental formula will depend upon the precision that is required.

When the B.t.u. figures estimated in this way are sufficiently precise, there remains but one statistical problem: a "curve-of-best-fit" must be drawn through the points relating ash content (dry basis) and B.t.u. value, as found through analysis of a number of samples.

In certain cases where it is necessary to know the precision of the estimated B.t.u. value as well, for instance when coal is sold on a B.t.u. basis with terms involving a penalty clause, the problem falls into three parts. First of all, the curve-of-best-fit, or "regression curve", must be found. Since it cannot be drawn by eye accurately enough, its "most likely" location requires a calculation known as regression analysis. Secondly, the precision of a single B.t.u. figure obtained from the curve must be calculated, and thirdly, a certain minimum number of samples must be collected from the consignment for ashing, in order to ensure a B.t.u. figure of predetermined precision. The foregoing concerns only the quantitative aspect of the problem. The statistical treatment in itself does not necessarily prove whether the relationship between the two variables is real or spurious.

In the above example there need be no doubt about the reality of the relationship. This is not always so, however. In actual fact, the majority of statistical problems require the experience and judgment of a professional worker in the field, in order to analyze the intrinsic relationship between the variables.

For instance, it is often found that within a certain range the compressive strength of coal briquets is inversely proportional to the surface moisture content of the coal entering the briquetting press. Yet, the actual cause of deterioration of the compressive strength of the finished briquet is neither the initial nor the instantaneous moisture content of the briquet, but rather the porosity which results from the presence of moisture during formation of the briquet in the press.

The above example illustrates the solution of a very important aspect of the problem, not by statistics, but simply by a detailed knowledge of the physical or chemical mechanisms and the environmental influences. The main factors governing the outcome of a process under investigation and the main sources of error in sampling and analysis should be known. Care in conducting the experiment, and awareness of the concomitant factors that might interfere with the test, are of the essence.

"Defining the problem" thus signifies the essentially qualitative evaluation of all the facts involved, based on a detailed knowledge of the process or phenomenon. It is a major step, one that is indispensable to the operation that follows: establishing the premises which lead up to the statistical treatment.

Summary -

"Defining the problem" stands for the qualitative analysis of the physical and/or chemical relationships between the dependent and independent variables. This is the step which establishes the reality of the relationship. A functional expression of the relationship is often helpful:

$$z = f(x,y, \ldots)$$

where x, y, ... are the independent variables and z the dependent variable, i.e. the effect that is being studied. All the variables that could possibly affect z should be listed and their relationship formulated in accordance with the physical or chemical laws that apply.

1.2 Premises

In this section, two steps of a quantitative nature are taken. Certain assumptions which will be made here regarding the problem will require subsequent verification.

First, the independent variables are classified according to their expected importance and the most important assigned as the "main factors". The "residual factors" which remain should all be of approximately the same order of influence and should not include any that are of appreciably greater importance. Their estimated composite effect should be smaller than the effect of the smallest "main factor".

Secondly, the "main factors" are measured in one way or another. If a factor cannot be measured directly, some quantity must be found which is closely related to it insome manner and which can be measured with ease and sufficient precision. For instance, it is well known that the efficiency of a continuous blender depends not only upon the number of circulations of the material passing through it, but also upon the rate of feed and the volume of the blender. When determining the efficiency of the blender, the rate of feed and the volume can be measured directly, but the number of circulations made by the material is not so easily found. In this case, a related, measurable quantity ("estimator" or "parameter") must be used. This introduces another assumption, namely that the factor (number of circulations) and its parameter are related. In the example, the speed of the impeller used for circulating the material in the blender could be used as the parameter, even though the exact relationship may not be known.

Summary -

Assign the main factors and rank them in descending order of importance.

Evaluate the combined effect of the residual factors and variations as accurately as possible from previous experience. If this remainder is seen to be larger than the smallest main factor, one or more of the residual factors will have to be classified as main factors for the experiments and analyses that will follow. Assign parameters for the main factors that cannot be measured directly.

1.3 Experimental Procedure

Verification of the above "Premises" is generally attained through preliminary testing. This means an additional series of operations before the actual data are collected for the procedure called "Analysis of Variance". The preliminary test and the actual experiment are both considered to be part of the experimental procedure.

Experimental procedure covers the range of tests from simple ones, such as determining the precision of a burette reading, to complicated factorial tests for determining the optimum conditions of a metallurgical process, or for ascertaining the cause(s) of certain diseases. Where information is not available, either a test is performed or the data are obtained by direct observation of the process. For instance, in a comparison of the success rate of a new surgical method with that of the conventional one, results can only be gathered as they become available. The same applies to studies in the field of economics and other areas where experimentation may be impractical or physically impossible. In most cases, however, the experimenter is free, within certain limits, to conduct a true experiment. This raises the question of how the test should be designed.

To help answer this, a classification of statistical procedures is given in Table 1.2. Experimental procedure will generally consist of the following parts: 1) a preliminary test to verify the premises and to obtain estimates of the "variances" involved; 2) an estimate of sample size and of the number of increments needed to obtain a pre-assigned accuracy for the individual observations; and 3) an estimate of the minimum number of observations required to attain a certain accuracy of the end result.

For instance, if an experimenter wants to assess the qualitative relationship between the average length of Douglas-fir shoots and the average summer temperature, the experiment will be designed to provide data for a "regression analysis". It is assumed that measurements are made at various latitudes and altitudes across the country in order to introduce variation into the average summer temperature.

The question of how many fir shoots should be measured requires a preliminary survey of the variability of fir-shoot length for a given latitude and altitude. This introduces a sampling problem.

The above example illustrates that experimental procedure for statistical problems generally consists of several elemental operations which are carried out one after the other in order to ensure maximum efficiency in the ultimate test.

Summary -

Design the experimental procedure by placing in chronological order the elemental operations needed to determine the number of observations and to calculate the size of the sample and number of increments required to ensure a pre-assigned accuracy for each observation. Collect the data for the "Analysis of Variance".

Table 1.2 - CLASSIFICATION OF STATISTICAL PROCEDURES

OUTLINE OF STEPS REQUIRED IN THE TREATMENT OF STATISTICAL PROBLEMS

I. Determine the type of problem or procedure under investigation.			II. Determine the ESTIMATOR, i.e. the "yardstick", to be used in measuring the item, property, event or phenomenon.				III. Collect the observations. Calculate the STATISTIC, i.e. the quantity used for expressing the result of the calculation.	Ref.
Type of Problem		Typical Applications/Procedure						
1. RANKING: Judging the order of preference or merit of an item or property.		Comparison of the taste, odour, colour, mental ability, or other subjective qualities, of one or more objects or individuals, by one judge or by several judges.	The RANKING NUMBER given by the judge to qualify the order of preference or merit.	1 item - or property	1 or 2 judges		Spearman's Rank Correlation Coefficient	(1)
					>1 judge		Coefficient of Concordance	(2)
				>1 item or property	1 judge only		Coefficient of Consistency	(3)
					>1 judge		Coefficient of Agreement	(4)
2. CALCULATION OF PROBABILITIES: The expected relative frequency of occurrence or non-occurrence of an event, property, etc.		Essentially for randomly dispersed variates (see also under Type 5: Tests of Significance). Expected number of defectives; coin, dice, and card games; randomness of oscillatory series.	The ABSOLUTE FREQUENCY OF OCCURRENCE of the event, and for binomial distributions, the ABSOLUTE NON-OCCURRENCE of the event or phenomenon.				The relative frequency of occurrence of the event, from the population distribution curve (normal, binomial, Poisson); the number of permutations and combinations (binomial only).	(5), GES
3. CALCULATION OF LIKELIHOOD: The likelihood of occurrence or contingency of events or phenomena.		Comparison of an observed frequency distribution with the expected frequency distribution, e.g. the normal distribution, the binomial or Poisson distribution.	The DIFFERENCE between the observed absolute frequency (O) of the occurrence of an event or phenomenon and the expected value (E).				Chi-square test; degrees of freedom (d.f.); probability level (P).	(5), GES
4. SAMPLING: Procedures for estimating a variate (X) of a material lot, with accuracy (a), preassigned or known in advance. Samples either collected and measured individually (single samples), or collected by "increments" from all over the lot and combined into one "gross sample" before an analysis is made.	Variate of limited range. Coefficient of variation is small. Little or no segregation (random dispersion of X).	Specification and prediction of quality. Sampling for defectives; sampling heterogeneous products for determination of physical or chemical properties; testing randomness of the distribution of (X) in space or in time; control charts. Random selection of samples should be adhered to if possible - systematic sampling requires special precautions in the evaluation of sample data.	The probability of occurrence of ≥c or <c items (X) in population of size (N) when sampling for attributes. The MEAN and VARIANCE of (X) when sampling for a continuous variate (X).				For (simple) random sampling: The minimum size of sample required to attain an accuracy (a) at level P. For stratified (random) sampling: The variances within and between strata; the total accuracy (a) at level P.	(5), GES
	Variate has large range. Large coefficient of variation. Noticeable or high degree of segregation (X is dispersed non-randomly over the consignment - "spotty pattern").		When the mean ($\bar{x}$) and standard deviation (s) are related, $s = f(\bar{x})$, use transformed variate $x' = \int dx/f(x)$ as estimator to reduce or eliminate covariance. Examples: s proportional to $\bar{x}^2$ - take reciprocals of x, s proportional to $\bar{x}$ - take logarithms of x, s proportional to $\sqrt{\bar{x}}$ - take square roots of x.				For increment sampling (using gross samples): The variance component due to random variation; the variance component caused by segregation and the variances of sample preparation and analysis; the total variance; the overall accuracy at level P.	
5. TESTS OF SIGNIFICANCE: Testing the assumption that the observed variations of a property or phenomenon are caused by chance (testing the Null Hypothesis). (See also under Type 2: Calculation of Probabilities.)		This includes testing the statistical significance of one or more variable factors that are expected to contribute to a phenomenon or event; to determine a physical or chemical property of a material or article.	Direct or indirect observation of the variable.	1 variable	1 set of data	Outlying observation (tramp), Fiducial limit of a mean.	t-test (normal distribution only)	(5)
						Fiducial limit of s or s^2.		GES
					>1 set of data	Difference between 2 means.	t-test (normal distribution only)	GES
						Diff. between 2 variances.	F-test, Z-test	(5)
				>1 variable	Type: $Z = f(X, Y, ..)$	Variates are classified in rows, columns, blocks, and replicates.	Analysis of variance methods (factorial tests, randomized blocks, F-test, etc.)	(6), GES
6. CORRELATION: Finding the quantitative, experimental relationship between the independent variables (causes) and the dependent variable (effect); expressing the goodness-of-fit of this relationship in an experimental and/or index formula.		Correlation is applied where the relationship between cause and effect is masked by a large number of relatively small, random influences.	By direct or indirect observation, measure the main factors in descending order of significance if possible. The factors chosen should be substantially independent of one another. Transform the estimator if necessary (see under Type 4: Sampling).				Regression coefficient(s) and constant; correlation coefficient (r); covariance; error variance; level of significance.	GES

References:-(1) Moroney, M J., "Facts from Figures", Penguin Books, 2nd ed 1953, p. 334 ff; (2) Ibid., p.336 ff; (3) Ibid., p. 340 ff; (4) Ibid., p. 348 ff; (5) Cowden, D.J., "Statistical Methods in Quality Control", Prentice Hall, 1957;
(6) Mentzer, E.G., "Tests by the Analysis of Variance", Wright Air Development Centre Tech. Rep. 53-23, Jan. 1953.
GES = "Guide to Engineering Statistics"

1.4 Analysis of Data

Every set of statistical data has a pattern of its own. This pattern can be presented in the form of a "frequency distribution" showing the relative or absolute frequencies of the items or values obtained.

Experience has shown that successive sets of data of the same variable, collected under comparable conditions, show frequency distributions having approximately the same average, same range, and same shape. The single observations of any set will generally "crowd" around a mean value and will deviate from this mean by an amount which cannot be predicted individually. However, from a large number of observations it appears that small variations with respect to the mean are generally more frequent than large variations, and that all variations cluster around a mean value within a limited range. It is then possible to predict limits within which the variable will lie when the experiment is repeated under comparable conditions. Fundamentally, every statistical technique is a means of evaluating, either directly or indirectly, the frequency distribution represented by the data, from the average, the degree of dispersion, and the shape of the distribution.

Example -

If replicate determinations are made of the specific gravity of a material, the observations will be distributed around the true (unknown) specific gravity, according to a frequency curve which is not unlike the familiar bell-shaped curve of Gauss-Laplace, more generally known as the Normal Curve.

As a rule, the number of data obtained will not be sufficient to show this curve in great detail. If only three or four determinations are done, however, these will generally fall within the range of such a curve, as would be verified by repeating the determination several hundreds of times. The frequency distribution is found by subdividing the range of observed specific gravities into a number of classes of equal interval and counting the number of observations found within each of the classes. The diagram representing the number of observations per class interval is called a histogram. The outline of this histogram approximates the distribution that would result if the number of observations and the number of class intervals were to be increased to infinity.

In the above example, all the specific gravity observations, taken together, constitute the "parent distribution" ("population"). This distribution typifies both the material itself and the method by which the specific gravity was determined.

An estimate of the average value and of the range of such a population can be found from a limited number, say three or four observations. It is clear that the estimate will be affected by the errors or deviations in each observation with respect to the mean.

Under normal conditions, a specific gravity determination will produce a set of observations distributed according to the "Normal Curve", the parameters of which are easily found. It is clear also that, under these circumstances, the greater the number of observations the more stable the mean value becomes.

It often happens that one or two observations in a set appear to be different from the remainder. They may or may not be part of the parent distribution. This means in effect that, owing to some unforeseen cause, a "systematic error" larger than a "chance error" has crept into the data. The parent distribution of this set of observations may therefore be no longer of the Normal type, and it becomes necessary to check the figures for outlying data ("tramps").

The chances of non-normality increase as the experiment becomes more complicated. When designing tests, systematic shifts in the values of the variables may be introduced deliberately in order to simplify the experimental procedure. The resulting frequency distributions are then quite often non-normal and more complicated methods for the analysis of these distributions are required. See Section 2, "Analysis of Variance".

Summary -

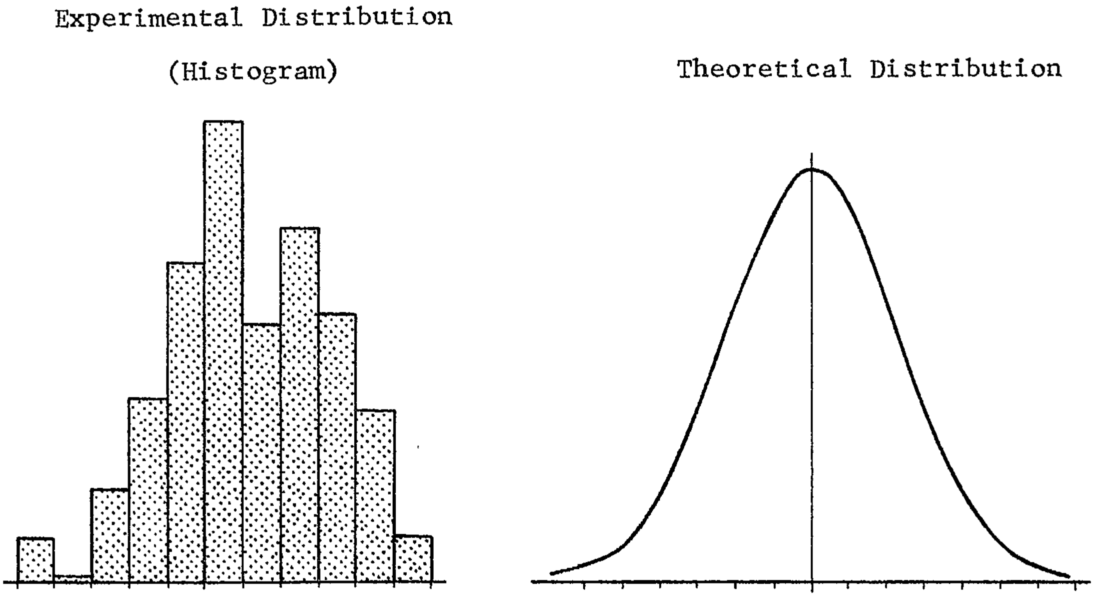
The analysis of experimental data is essentially the analysis of the frequency distribution(s) obtained from the data. Basically, it involves the determination and comparison of means, ranges and shapes of the distributions.

1.5 Frequency Distributions

This section deals with the frequency distributions that commonly occur in statistical experiments, the parameters used for describing these distributions, and the tests that are used for comparing the parameters.

Mathematical statistics deal with variables, i.e. physical or chemical properties, attributes or events which show a certain range of variability. It can be shown by experiment that most variables do not behave chaotically but, rather, conform to a certain pattern of behaviour. This can be illustrated by the experiment of Galton: a board covered with staggered rows of nails is used to scatter the course of a large number of beads which are introduced at the top of the board and are eventually trapped in a series of partitions at the base of the board. The end result is indicated by the distribution of the beads arrested between partitions.

The frequency distribution of the beads shows a marked orderliness, which is caused by the nails interfering with the gravitation of the beads. It shows that small deviations from the mean are more frequent than large deviations. The beads tend to crowd around the mean value. This phenomenon is known as the "central tendency".



Experimental Distribution

(Histogram)

Theoretical Distribution

Fig. 1.1 - Frequency Distribution

A collection of individuals (be it observations, items, or events), when related in this manner, is said to form a "universe" or "population". This is referred to as the Law of Large Numbers. Numerous experiments have shown that this law has a wide application in nature and in nearly every field of human endeavour.

In practice, all kinds of frequency distributions are found. These include symmetrical bell-shaped distributions with single tops (unimodal - see Fig. 1.1) like the one found in the

Galton experiment; positively- or negatively-skewed asymmetrical distribution; double-topped (bimodal) distributions (see Fig. 1.2); and others of seemingly irregular shape.

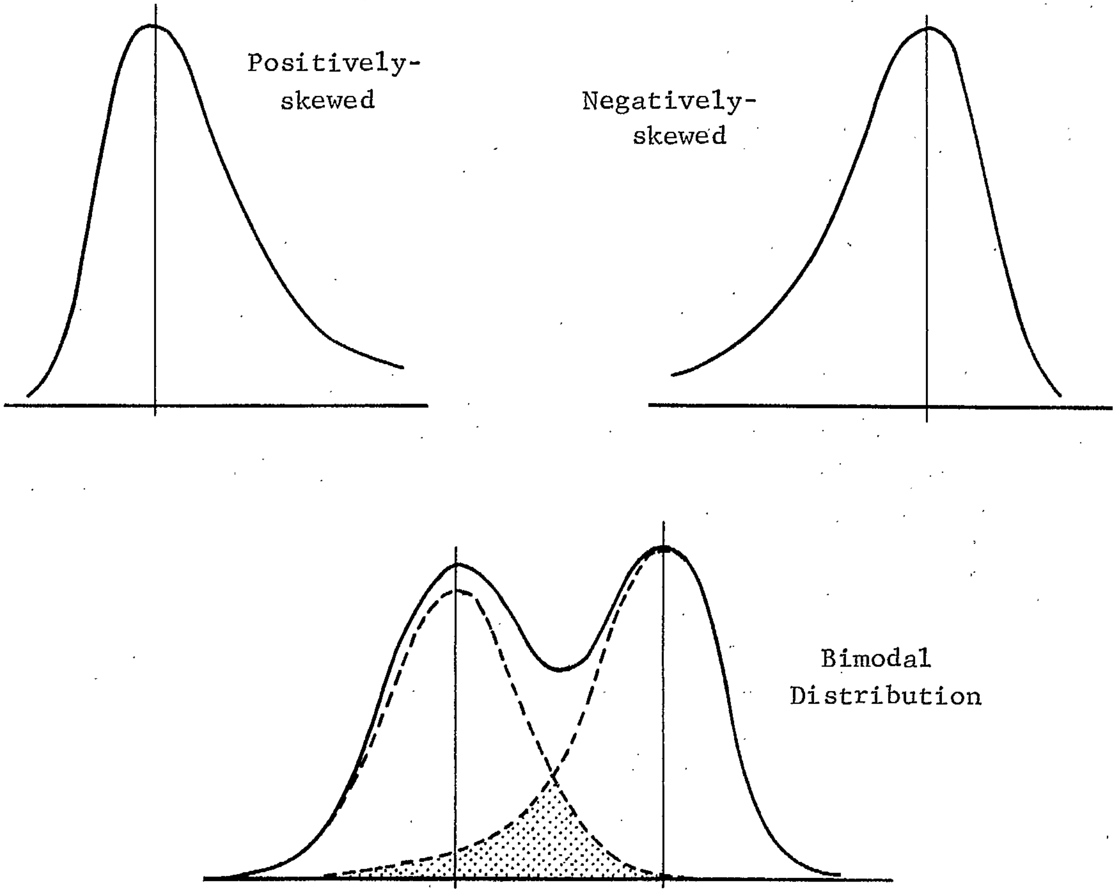


Fig. 1.2 - Asymmetrical Frequency Distributions

When dealing with the Normal curve, statistical interpretation boils down to finding the average value, measuring the scatter of the observations, and checking on the normality of distribution of the data. Although procedure is basically the same for non-normal distributions, emphasis here is shifted to the testing of differences between means and differences in scatter. A compound frequency distribution is looked upon as the sum of two or more single-top distributions, each of them caused by one main factor.

Summary -

The aim of the statistical procedure is to describe frequency distributions of observed data in terms of average and scatter as they relate to the normal distribution, for the purpose of estimating the true value and range of the variable or variables involved.

1.6 The Normal Curve

The formula for the normal curve expresses the relationship between the values of a variable (p) and their relative frequencies (y) for a total number of observations (n).

$$y = \frac{n}{\sigma\sqrt{2\pi}} . e^{-(p-\mu)^2/2\sigma^2} \qquad \ldots\ldots \text{(Eq. 1)}$$

The formula has two parameters: the true mean value (μ), and the standard deviation of the population (σ). The latter is a measure of scatter and will be discussed in the next section. The shape of the frequency curve of any normal population can thus be evaluated once the true mean (μ) and the "population standard deviation " (σ) are known. Geometrically, this "true" standard deviation (σ) represents the distance between the mean and the points of inflexion on the normal curve. It is the root-mean-square of the deviations with respect to the mean value μ.

$$\sigma^2 = \frac{\Sigma(p-\mu)^2}{n} \qquad \ldots\ldots \text{(Eq. 2)}$$

One of the properties of the normal curve is that the area under the curve between $\mu+\sigma$ and $\mu-\sigma$ (shaded area Fig. 1.3) is 68% of the total area. This means that 68% of the deviations ($p-\mu$) are smaller than (σ). Similarly, the area between limits

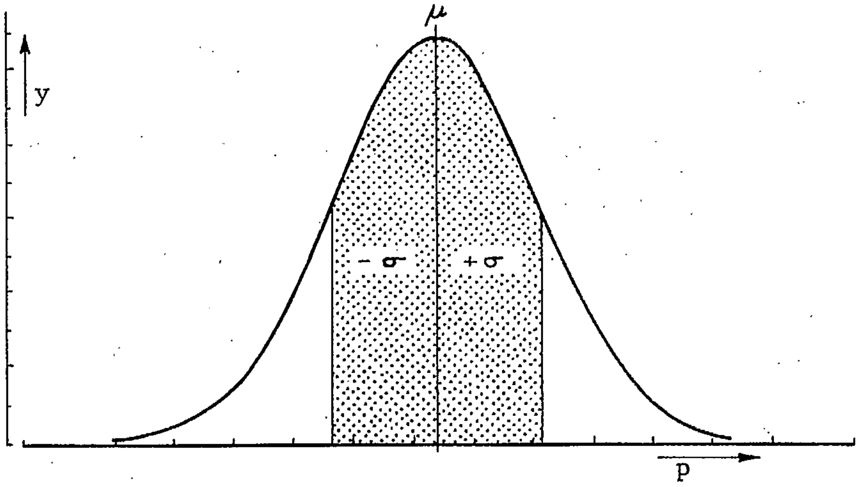


Fig. 1.3 - The Normal (Gauss) Curve

$+2\sigma$ and -2σ (theoretically 1.96s) is 95% of the total area, or nineteen out of every twenty observations; the area between $+3\sigma$ and -3σ is 99.7%. (The 2σ limit is commonly used as the measure of precision in engineering forecasts. See Section 5, examples 4, 8, 10 and 12.)

Thus, the standard deviation is not merely a kind of average, but is also a means of calculating the chance or the probability of occurrence of a certain error or deviation from the mean value. The essence of statistical procedure is in fact the calculation of probability of occurrence of any phenomenon which is subject to the Law of Large Numbers. In simple words, the standard deviation is used to describe the existing situation and, in addition, to predict future behaviour.

1.6.1 Estimate of standard deviation from the observations

The basic calculation consists of finding an estimate of the true mean, and of the true standard deviation (or true standard error), from a limited number of observations as a first step in calculating the observed quantity and its range of scatter. It is clear that only an infinite number of observations will produce a complete picture of the phenomenon. In practice, however, only an estimate of (σ) is obtained and with a limited precision which depends upon the number of observations.

This estimated standard deviation is designated by the symbol (s) and is based on deviations from the arithmetic mean ($\overline{P}$) instead of the true mean. If the true mean were known, a better estimate of the true standard deviation could be found. The following equation may be used for computing the most probable estimate of the true standard deviation from a finite number of observations:

$$s^2 = \frac{\Sigma x^2}{(n-1)} \qquad \text{..... (Eq. 3)}$$

where x = ($p-\overline{P}$). The standard deviation (s) and its square the variance (s^2) are the two "statistics" most commonly used for estimating the scatter of an infinite population based on a limited number of observations.

A second, derived equation may be used which facilitates the calculation of (s), particularly when dealing with a large number of observations:

$$s^2 = \frac{\Sigma p^2 - (\Sigma p)^2/n}{(n-1)} \qquad \text{..... (Eq. 4)}$$

In this form, the standard deviation or variance can be determined by using the observed values (p) directly without having to calculate their deviations from the mean, i.e. ($p - \overline{P}$).

Example -

Observation of the automobile accident rate in a certain town during five equal periods showed the following results: 6, 8, 3, 9, and 5. The variance, according to Equation 4, works out as follows:

$$\Sigma p^2 = 6^2 + 8^2 + 3^2 + 9^2 + 5^2 = 215$$

$$\Sigma p = 6 + 8 + 3 + 9 + 5 = 31$$

$$(\Sigma p)^2/n = 31^2/5 = 192.2$$

$$\text{variance, } s^2 = \frac{215 - 192.2}{5 - 1} = 5.70$$

standard deviation, s = 2.39

Average, $\overline{P} = \Sigma p/n = 6.2$

1.6.2 Quick method for estimating the standard deviation

The standard error (s) can be quickly determined from the range (w) of a series of (n) observations, using values given in Table 1.3. This method applies to a series of not more than 10 observations, and is to be used only when the observations are normally distributed after elimination of any tramp observations.

Table 1.3

n	s/w
2	0.89
3	0.59
4	0.49
5	0.43
6	0.40
7	0.37
8	0.35
9	0.34
10	0.32

Note: The terms standard deviation and standard error have been used interchangeably because there is no fundamental difference between the two. The term "standard error" is used when the observations differ mainly as a result of human or instrument errors. In all other cases the term "standard deviation" is employed and is generally to be preferred.

1.6.3 Other parameters for measuring scatter

Various quantities have been used in the literature to describe the range of scatter of a series of observations. Two of these are mentioned here.

The first one is the average deviation (g) which is found from the data by simply averaging the deviations:

$$g = \frac{\Sigma|(p-\overline{P})|}{n} \qquad \text{..... (Eq. 5)}$$

The second statistic used is the probable error (r) which indicates the limits (+ and -) on either side of the mean between which theoretically 50% of the observations are found.

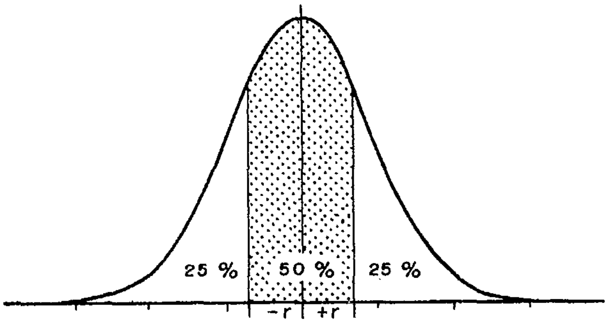


Fig. 1.4 - The Probable Error

In theory, there is a constant ratio between the standard deviation and both (g) and (r), but only in the Normal case. If this condition of Normality is met, factors given in Table 1.4 below can be used for converting from one to the other.

Table 1.4

$s = 1.252 \cdot g$	$g = 0.798 \cdot s$
$s = 1.484 \cdot r$	$r = 0.674 \cdot s$
$g = 1.184 \cdot r$	$r = 0.845 \cdot g$

2. ANALYSIS OF VARIANCE

Where there are two or more possible sources of variation in a set of data, a technique known as Analysis of Variance can be used to determine how much of the total variation for all the observations taken together can be attributed to the different causes. For example, a test produces three sets of three observations each. The mean values as well as the variances of the sets are found to differ. An answer as to whether or not the differences are significant is given by the Analysis of Variance, which provides the means of calculating the odds that the observed differences were caused by chance. Only a partial answer can be obtained, however: if the difference is larger than can be explained by chance variation, it is a significant difference. If the difference is small, on the other hand, it may be significant but the possibility cannot be proven. In other words, the hypothesis that no difference exists (the Null Hypothesis) can never be proven, but can only be disproved.

Each of the three variances in the above example contributes to the overall variance of the nine observations. This overall variance, which is a compound variance, generally results from tests which deal with a multiplicity of factors. The main value of the Analysis of Variance technique lies in its use as a means of finding estimates of the individual variance components that contribute to the overall variation of the data. In so doing, it reveals the relative influence of the individual factors.

The four examples which will follow in this section constitute a mental experiment to show in a simplified manner what type of factors generally enter into observations that are obtained from a test program. Essential points are summarized in Table 1.2, under problem Type 5.

2.1 General Procedure

For purposes of calculation, the data are arranged in tables of rows, columns, blocks and replicates as required. The data of each column should have a natural tie, i.e. a distinctive feature which characterizes this column as being distinct from other columns. The same applies to each row, cell, etc. The magnitude and significance of each of the separate variances can be found by following the procedure outlined below:

1) Arrange the data to be analyzed in Columns and Rows as shown in Examples 1 and 2 (single observations) and 3 and 4 (duplicate observations).
2) Using the formulas given in Tables 2.2 and 2.6, calculate the following:
   a) Sum of Squares: (S.S.)
      i) Between Columns (β)
      ii) Between Rows (γ)
      iii) Total S.S. (α)
   b) Interaction (ϵ) and/or Error (τ)
   c) Degrees of Freedom (d.f.)
   d) Mean Squares (M.S.)
3) Find "true variance" estimates.
4) Check statistical significance of the variance components by means of the F-test.

Example 1

The data used in this example may be taken to be the exact values representing a certain process or phenomenon being tested and are free from error. There are no replicates. The data are arranged in three rows (R = 3) and three columns (C = 3). Row sums A, B, D and column sums P, Q, T are found and entered in the data table, together with the overall sum, M:

Table 2.1 - Test Data (Example 1)

	C			
R	2	6	10	A_{18}
	4	8	12	B_{24}
	6	10	14	D_{30}
	P_{12}	Q_{24}	T_{36}	M_{72}

Using a Variance Table set up like the following, the variance components relating to the data may now be calculated:

Table 2.2 - Analysis of Variance (Example 1)

Source of Variation		Sum of Squares (S.S.)	Degrees of Freedom (d.f.)	Mean Square (M.S.)	Expected Mean Square
Between Columns	C	$\beta = \frac{P^2+Q^2+T^2}{R} - \frac{M^2}{CR} = 96$	(C-1) = 2	$V_\beta = \frac{\beta}{2} = 48$	$V_\beta = V_\tau + RV_C$
Between Rows	R	$\gamma = \frac{A^2+B^2+D^2}{C} - \frac{M^2}{CR}\ 24$	(R-1) = 2	$V_\gamma = \frac{\gamma}{2} = 12$	$V_\gamma = V_\tau + CV_R$
Residual (error)		$\tau = \alpha - (\beta + \gamma) = 0$	(C-1)·(R-1)=4	$V_\tau = \frac{\tau}{4} = 0$	$V_\tau = V_\tau$
Total		$\alpha = \Sigma p_i^2 - \frac{M^2}{CR} = 120$	(CR-1)=8	—	Compound Variance

"True Variance" estimates

1) $V_C = \frac{V_\beta - V_\tau}{R} = \frac{48 - 0}{3} = 16$

2) $V_R = \frac{V_\gamma - V_\tau}{C} = \frac{12 - 0}{3} = 4$

3) $V_\tau = 0$

Notes

1. The true variance V_C ($= s_c^2$) in the Table refers to observations in different columns but in the same row, and is called the

"variance between columns". The true variance V_R ($=s_R^2$) refers to observations in different rows but in the same column, and is called the "variance between rows". The true variance V_τ is called the "residual variance", "error variance", or "interaction variance" (if no error is involved).

2. Because V_τ is 0, the true variance estimates are exact. This may be checked by calculating the variances directly from the observations, e.g. s^2 (2, 6, 10) = 16; etc.

3. The difference between the row and column variances (V_γ and V_β respectively) is tested using the "F-test": the Null Hypothesis that no difference exists between two variances is tested at a significance level P. In this test, a ratio (F) is computed from s_1^2/s_2^2 with the larger variance always in the numerator so that F is always greater than 1. In the example, $F=V_\beta/V_\gamma=48/12=4.0$. Entering a table of F-values for degrees of freedom (d.f.) = 2 and 2 respectively, the following values are found: $F_1 = 99$; $F_5 = 19$; $F_{10} = 9.0$.

Subscripts of F denote the Probability Levels, i.e. the probability that a given difference is due to random or chance variation. If a computed F-value exceeds the theoretical one at P=0.01, this means that there is only 1 chance out of 100 that the difference is attributable to random variation. In other words, the difference is considered to be "highly significant". Similarly,

$F_5 < F < F_1$ - difference is "significant"

$F_{10} < F < F_5$ - difference is "possibly significant"

$F < F_{10}$ - difference is "not significant"

In this example, since the computed value of F is only 4, i.e., $F < F_{10}$ for 2 and 2 d.f. respectively, it can be concluded that the difference found from the data is probably not a significant one.

The significance levels (P) of variances M. S. (Mean Square) are commonly indicated by asterisks as follows:

$***$ $P < 0.01$ - "highly significant"

$**$ $P = 0.01$ to 0.05 - "significant"

$*$ $P = 0.05$ to 0.10 -"possibly significant"

- $P > 0.10$ - "not significant"

Though arbitrary, these levels are generally accepted for investigational and research work in many areas of technological inquiry.

Notes on variance formulas

As given in the preceding Variance Table, the Sum of Squares between Columns may be found from:

$$\beta = \frac{P^2 + Q^2 + T^2}{R} - \frac{M^2}{CR}$$

which is derived from

$$\beta = \Big[\underbrace{(P/R)^2 + (Q/R)^2 + (T/R)^2}_{\text{Crude S.S. of Column Averages}} - \underbrace{(M/R)^2/C}_{\text{Correction Term}}\Big] \cdot R$$

(Note the similarity of this formula with the one used for calculating the variance of a set of single observations.)

From this, it is seen that:

$\beta =$ R x Sum of Squares of the Column Averages

= R x Average estimate of the Sum of Squares of "deviations between columns" (i.e. single observations in the same row, in different columns, and which include V_τ).

Since this average estimate still contains V_τ, $V_\beta = R \cdot V_C + V_\tau$ from which

$$V_C = \frac{V_\beta - V_\tau}{R} .$$

Example 2

The data are the same as in Example 1 (the row sums and column sums are unchanged) except for the introduction of a reading error which has altered some of the individual observations.

Table 2.3 - Test Data (Example 2)

		C		
	3	5	10	A_{18}
R	4	9	11	B_{24}
	5	10	15	D_{30}
	P_{12}	Q_{24}	T_{36}	M_{72}

Table 2.4 - Analysis of Variance (Example 2)

Source of Variation		S. S.	d.f.	M.S.	Expected Mean Square
Between Columns	C	$\beta = 96$	2	*** $V_\beta = 48$	$V_\beta = V_\tau + RV_C$
Between Rows	R	$\gamma = 24$	2	** $V_\gamma = 12$	$V_\gamma = V_\tau + CV_R$
Error		$\tau = 6$	4	$V_\tau = 1.5$	$V_\tau = V_\tau$
Total		$\alpha = 126$	8	-	Compound Variance

"True Variance" estimates

1) $V_C = \dfrac{48 - 1.5}{3} = 15.5$ (16.0)

2) $V_R = \dfrac{12 - 1.5}{3} = 3.5$ (4.0)

3) $V_\tau = 1.5$

For calculation of β, γ, τ, and α, see Table 2.2, Example 1.

Notes

1. The "true variance" estimates are no longer exact, owing to reading errors (compare with the exact values in brackets

which were found in Example 1).

2. The F-test in this case disproves the hypothesis that the variation between columns was caused by reading (chance) errors. The value of the ratio, $F = V_\beta/V_\gamma = 48/1.5 = 32$, is highly significant for 2 and 4 d.f. respectively. The F-table gives a value of 18 for 2 and 4 d.f. at P = 1%, indicating that the difference between columns was not caused by chance (reading errors): the probability is less than 1% ($P<0.01$) that this conclusion is wrong.

3. The F-test for the difference between rows gives $F=12/1.5=8$ for d.f. 2 and 4, indicating that the variance between rows is significant at a level (P) between 1 and 5% ($F_1 = 18$, $F_5 = 6.94$).

Example 3

The data used are the same as before, except that they are now represented by duplicate instead of single readings (row and column sums are doubled). This will demonstrate the introduction of an analytical error.

Table 2.5 - Test Data (Example 3)

C

S_1 1.5-2.5	S_2 5.5-6.5	S_3 9.5-10.5	A 36
S_4 3.5-4.5	S_5 7.5-8.5	S_6 11.5-12.5	B 48
S_7 5.5-6.5	S_8 9.5-10.5	S_9 13.5-14.5	D 60
P 24	Q 48	T 72	M 144

Symbols

S = cell total (e.g. $S_1 = 4.0$; $S_2 = 12.0$),

H = number of replicates = 2,

P_i = individual observation,

C = 3, R = 3.

Table 2.6 - Analysis of Variance (Example 3)

Source of Variation		S. S.	d.f.	M. S.	Expected Mean Square
Between Columns	C	$\beta = \frac{P^2+Q^2+T^2}{RH} - \frac{M^2}{CRH}$ $= 192$	$(C-1)$ $= 2$	*** $V_\beta = 96$	$V_\beta = V_\tau + HV_{CR} + RHV_C$
Between Rows	R	$\gamma = \frac{A^2+B^2+D^2}{CH} - \frac{M^2}{CRH}$ $= 48$	$(R-1)$ $= 2$	*** $V_\gamma = 24$	$V_\gamma = V_\tau + HV_{CR} + CHV_R$
Interaction	CxR	$\epsilon = \delta - (\beta + \gamma) = 0$	$(C-1)(R-1)$ $= 4$	$V_\epsilon = 0$	$V_\epsilon = V_\tau + HV_{CR}$
Between Cells		$\delta = \frac{\Sigma S_i^2}{H} - \frac{M^2}{CRH}$ 240	$(CR-1)$ $= 8$	—	Compound Variance
Error		$\tau = (\alpha - \delta) = 4.5$	$CR(H-1)$ $= 9$	$V_\tau = 0.5$	$V_\tau = V_\tau$
Total		$\alpha = \Sigma p_i^2 - \frac{M^2}{CRH} = 244.5$	$(CRH-1)$ $= 17$	—	Compound Variance

"True Variance" estimates

1) $V_C = \frac{V_\beta - V_\tau - HV_{CR}}{RH} = 16.0 \quad (16.0)$

2) $V_R = \frac{V_\gamma - V_\tau - HV_{CR}}{CH} = 4.0 \quad (4.0)$

3) $V_{CR} = \frac{V_\epsilon - V_\tau}{H} \simeq 0$ (Variances are never negative)

4) $V_\tau = 0.5$

Notes

1. The sensitivity of the F-test has been increased by augmenting the number of degrees of freedom; i.e. by increasing the number of observations. The "between columns" and "between rows" variances are found to be highly significant (P<0.01).

2. Interaction would have been indicated if the variance "between columns" had shown an increase (or decrease) for each row, going from row 1 to row 2 to row 3. Since no such change is found in this example, the interaction variance $V_\epsilon = 0$.

Example 4

The column and row averages remain unchanged, but the reading and analytical errors (Examples 2 and 3) are now both included.

Table 2.7 - Test Data (Example 4)

C

	2.5-3.5	4.5-5.5	9.5-10.5	A 36
	3.5-4.5	8.5-9.5	10.5-11.5	B 48
R	4.5-5.5	9.5-10.5	14.5-15.5	D 60
	P 24	Q 48	T 72	M 144

Table 2.8 - Analysis of Variance (Example 4)

Source of Variation		S. S.	d. f.	M. S.	Expected Mean Square
Between Columns	C	$\beta = 192$	2	*** $V_{\beta}=96$	$V_{\beta}=V_{\tau}+HV_{CR}+RHV_{C}$
Between Rows	R	$\gamma = 48$	2	** $V_{\gamma}=24$	$V_{\gamma}=V_{\tau}+HV_{CR}+CHV_{R}$
Interaction	CxR	$\epsilon = 12$	4	** $V_{\epsilon}= 3$	$V_{\epsilon}=V_{\tau}+HV_{CR}$
Between Cells		$\delta = 252$	8	-	Compound Variance
Error		$\tau = 4.5$	9	$V = 0.5$	$V_{\tau}= V_{\tau}$
Total		$\alpha = 256.5$	17	-	Compound Variance

"True Variance" estimates

1) $V_C = (V_{\beta} - V_{\epsilon})/RH = 15.5$ (16.0)

2) $V_R = (V_{\gamma} - V_{\epsilon})/CH = 3.5$ (4.0)

3) $V_{CR} = (V_{\epsilon} - V_{\tau})/H = 1.25$ (1.0)

4) $V_{\tau} = 0.5$

For calculation of β, γ, ϵ, δ, τ, and α, see Table 2.6, Example 3.

Notes

1. The "true" variance estimates are not exact (see theoretical values in brackets).

2. It is clear that an interaction variance has been simulated here from the combined effects of the reading and analytical errors, which were known in advance to be true sources of variation in this example.

2.2 Discussion of Analysis of Variance Results

The four preceding examples were given to illustrate the mathematical procedure and physical background of the Analysis of Variance technique. A further example may more clearly show the meaning of the "natural tie". It will be assumed that 10 sets of triplicate observations are available with each set representing data which refer to a different "level" in a range of levels, e.g. 10 volatile matter (VM) determinations done in triplicate at 10 different temperatures. If the 10 sets are arranged in 10 rows of 3 columns each in the order obtained from the tests, the variance between rows (V_R) will reflect the influence of temperature on the VM readings because each row refers to a specific temperature. This is the "natural tie" for the figures within each row. The variance between columns (V_C), on the other hand, does not reflect any special factor, because the observations in each column have no common tie. This would occur, however, if the first column should happen to contain the lowest value of each set, the second column the middle value and the third column the highest value of each set; the between-columns variance (V_C) would then reflect the error variance.

The significance of the various mean squares has been tested by comparing them with the error variance V_T. This was the correct procedure for the examples given. Suppose, however, that the interaction variance in Example 4 had proved to be insignificant, as in fact it actually is; a new estimate of V_T would then have been obtained by combining the interaction and residual sums of squares and dividing by the sum of their respective degrees of freedom (d.f.), and the main variance re-tested on this new basis.

The paradoxical result for the interaction variance of Example 4 (see note 2) illustrates a problem that faces the experimenter in all statistical work. Statistics being a formal logic only, detailed knowledge of the process and the history of the experiment are generally required before the validity of the analysis of variance and subsequent correlation can be established.

The Analysis of Variance technique can be applied to data that might have been obtained previously for a different purpose, e.g. operational control of a plant. These data are often not complete enough nor suitable for such an analysis, and grave doubts may arise about the validity of the results. It is necessary in such a case to design a test beforehand. Where the nature of a physical or chemical process is known and the ranges of the factors involved are also known, it is possible to select a number of levels to be tested for each factor and the minimum number of replicates that will be required to establish significance. Tests of this kind when combined with Analysis of Variance are called Factorial Tests. (See para. 2.4 and Section 3.)

The nature of the data that were used in the above mental experiment is neither known nor is it important, since the Analysis of Variance operates independently of it. In experimental work, procedure is generally designed with the purpose of determining the true relationship that exists between two or more physical/chemical attributes. This purpose is achieved by means of measurable quantities called parameters. The resultant experimental equation will truly represent the relationship between the variables, provided that the parameters have no other elements in common that might mask or distort them. If the parameters have such an element in common (e.g. a common denominator, a common factor, or a constant in common), then the true relationship between the variable attributes may be either partly or entirely masked by spurious correlation. A common form of spurious correlation occurs when the parameters show significant correlation as expressed by the correlation coefficient (r), even though the variable properties are not related in any way. An example based on a publication by Karl Pearson (7) is given below, using three series of random numbers. Series 1 and 2 are the variables, series 4 and 5 their respective parameters.

Table 2.9 - Spurious Correlation

Series	1	2	3	4	5
	x_1	x_2	x_3	x_1/x_3	x_2/x_3
	54.9	76.1	18.3	3.00	4.16
	62.2	63.5	16.0	3.89	3.97
	57.0	78.2	21.3	2.68	3.67
	74.7	78.0	25.1	2.98	3.11
	65.4	84.2	13.9	4.70	6.06
Mean	62.84	76.00	18.92	3.45	4.19
Variance	61.19	58.08	19.49	0.69	1.25
Variation Coefficient	0.12	0.10	0.23	0.24	0.27
Correlation Coefficient	$r_{1,2} = 0.1637$			$r_{4,5} = 0.8507$	

This example demonstrates that great care is required in the choice of parameters, especially when using dimensionless ratios that are plotted against one another or against a single variable.

They may have an element in common that will invalidate subsequent correlation of variables, as described in Section 4.

2.3 Procedure for Experimental Design

The following general procedure serves as a guide to the experimenter who is faced with the problem of having to plan an experiment for testing the nature or behaviour of some phenomenon.

1. Examine the problem. Determine the "operating variables" and "other factors". Subdivide the latter category into "controlled" and "more or less uncontrolled" factors. List everything under these three headings.

Table 2.10 - Example: Briquetting of Coal in a Roll-Press

Operating variables (1)	Other factors	
	Controlled (2)	More or less uncontrolled (3)

% binder;
type of binder;
method of dispersing binder.
	Fluxer temp.;
	speed of press;
	rate of feed;
	temperature of coal entering the press.
		Moisture % of feed;
		particle size of feed:
		temperature of coal in mixer;
		cooling of coal between mixer and press;
		cooling of briquets;
		steam pressure and quality.

2. Choose the 2 to 4 most important factors for the test, bearing in mind that the combined effect of the factors that are omitted should be small enough not to interfere with the test. If necessary, the experiment can be split into two separate tests, e.g. for the above,

Test (1): effect of % asphalt binder and moisture content.

Test (2): effect of type of binder and method of dispersion.

3. Factors may be distributed either normally (as in Fig. 1.1) or non-normally (skew unimodal, bimodal, etc.; see Fig. 1.2). For example, moisture % is unimodal; method of dispersion is bimodal. Try in both cases, first of all, to set up the test as a bimodal one by taking the factors at two levels, one "high" and one "low". Both levels should be chosen so that the working range is covered, and so that reproducible results can be obtained.

For instance, tests are run with asphalt "high" (e.g. 5%) and asphalt "low" (e.g. 3.5%). Linear behaviour is anticipated here. If linearity is not expected, the range should be reduced, without, however, sacrificing or endangering the bimodal nature of the tests.

This type of experiment is known as one with "fixed constants". It is designed as a factorial test, using 2^n tests for (n) factors, and H replicates per test. If this factorial test is impossible because of the nature of the phenomenon, then use the type of design that employs more than two tests per factor and (H) replicates per test, with rows, columns, cells, sub-cells, etc. This type of test is necessary for "normally distributed" factors; that is, the distribution of the factor over the columns of the table, for instance, is unimodal and gradual, and the differences from column to column are small and of the same order of magnitude as the variations within the columns. Of course, with one "normally distributed" factor and one or two "fixed constants", the same type of variance analysis will have to be employed. Many possible types of this "mixed" nature are given in Mentzer's manual (6).

It is noted that both the factorial test and the other designs using more than 2 rows, columns, etc., for the analysis of fixed constants, are very powerful. This is because all of the partial variances can be tested with the error variance, the degrees of freedom of which can be determined in advance from the number of replicates. On the other hand, where the distribution of factors is unimodal, the partial variances cannot always be tested. Artifices need to be found, and generally the test is a less powerful one.

2.4 Design of Factorial Tests

The minimum required number of tests for any n-factorial design is fixed at (2^n) tests. The minimum required number of replicates (H) per test which can be found from Table 2.11 ensures sufficient accuracy of the (error) variance and therefore a sufficiently powerful test. The number of replicates is based on a predetermined precision of the (error) variance, $a_V \leq 52\%$. See derivation below.

Table 2.11 - Design of Factorial Tests

Number of Factors (n)	Minimum no. of Replicates (H)	Minimum no. of Observations (2^n H)
1	30	60
2	16	64
3	9	72
4	5	80
5	3	96

Derivation of % Precision of the Error Variance, $\%a_V$

The standard deviation of a variance is:

$$s_V = s^2\sqrt{2/f},$$

where s = standard error of a single observation and

f = degrees of freedom.

Expressed as a percent,

$$\%\ s_V = 100s_V/V = 100s_V/s^2$$

$\therefore \% \ s_V = (100s^2\sqrt{2/f})/s^2 = 200/\sqrt{2f}$,

and % precision of the variance at P=95% level,

$\% \ a_V = 2 \cdot s_V = 400/\sqrt{2f} = 52\%$ for 30 observations.

Note: $a_V \leq 52\%$ corresponds with the generally accepted minimum of 30 single observations per test.

3. FACTORIAL ANALYSIS

This section deals with the analysis of data obtained from a factorial test. The procedures are illustrated with an example of an actual test involving three factors:

The compressive strength of briquets in relation to % asphalt (A); moisture % of "Green Briquets" (B); and asphalt type - natural vs cracked (C).

Two cases will be considered:

1. No replicate observations available
2. Method of calculation when using replicates

3.1 No Replicates

Factorial test for 3 factors is based on 2^3 or 8 combinations of three factors, each one at two levels: a "high" and a "low". Levels are chosen to cover the working range, while avoiding extreme values.

The data in Table 3.1 can be used to find all the variances, including Main Effects (A, B, C) and Interactions (A x B; A x C; B x C; A x B x C), as follows:

When the sum of the 4 tests containing "A" (314 + 301 + 282 + 223) is compared to the sum of the 4 remaining tests containing "a" (304 + 134 + 264 + 158), the difference indicates the effect of changing the asphalt content from A to a only; the other factors, moisture and asphalt type, cancel out because both their high and low values are represented in each of the above two groups. The difference in the groups is an estimate of the effect of the % asphalt between level "A" and level "a".

The same data can be regrouped to produce the difference between the moisture levels B and b. In this case, the effects of % asphalt and asphalt type cancel out, and so on.

Table 3.1 - Test Data*

	A		a	
	B	b	B	b
C	ABC	AbC	aBC	abC
c	ABc	Abc	aBc	abc

Test No.	Compressive Strength, lb	Legend		
abc	304	Asphalt level:	High, 5.54%	A
aBc	134		Low, 4.37%	a
Abc	314	Moisture level:	High, 11.9%	B
ABc	301		Low, 8.1%	b
abC	264	Asphalt type:	Straight-run	C
aBC	158		Cracked	c
AbC	282			
ABC	223			

* Compressive-strength values are the averages of the replicates given in Table 3.3, and represent the average levels of asphalt and moisture.

The formulas and diagrams given below may be used to calculate the variances of the main effects and their interactions. This is a simplified method (8) which reduces the calculations for "Sums of Squares" to a minimum.

Main Effects:

$$\boxed{S_A = (\Sigma A - \Sigma a)^2 \;/2^n} \quad \text{(Eq.6)}$$

$$\boxed{S_B = (\Sigma B - \Sigma b)^2 \;/2^n} \quad \text{(Eq.7)}$$

$$\boxed{S_C = (\Sigma C - \Sigma c)^2 \;/2^n} \quad \text{(Eq.8)}$$

$S_{A,B,C}$ = Sum of Squares A,B,C;

$\Sigma A,B,C$ = Sum of obs. A,B,C;

n = no. of factors = 3

Interactions:

1st order

$$\boxed{S_{AB} = (\Sigma\square - \Sigma\boxdot)^2/2^n} \quad \text{(Eq.9)}$$

$\Sigma\square$ = Sum obs. indicated by blank squares.

$\Sigma\boxdot$ = Sum obs. indicated by shaded squares.

	A	a
B	▨	
b		▨

Make similar diagrams for the other 1st-order interactions A x C and B x C, and calculate S_{AC} and S_{BC}.

2nd order

$$\boxed{S_{ABC} = (\Sigma\square - \Sigma\boxdot)^2/2^n} \quad \text{(Eq.10)}$$

	A		a	
	B	b	B	b
C		▨	▨	
c	▨			▨

, The above formulas and diagrams provide the information required for the following table in which the variances are tested:

Table 3.2 - Analysis of Variance (Factorial Test)

Source of Variation	Sum of Sq. (S.S.)	d.f.	Variance $\frac{S.S.}{d.f.}$	Test Ratio (F)
Main Effects A	8450	1	8450	4.29*
B	15138	1	15138	4.96*
C	1985	1	1985	Not.Sig.
Interactions 1st order AB	5202	1	5202	" "
AC	1104	1	1104	" "
BC	40	1	40	" "
2nd order ABC	1513	1	1513	" "
Total	33,432	7	Compound Variance	

To test the variances, list them in descending order of magnitude and test the largest variance first:

$$F_B = \frac{15138}{\Sigma R/6} = \frac{15138}{3049} = 4.96,$$

where $\Sigma R/6$ = the sum of the remaining Sums of Squares (8450 + 1985 + 5202 + 1104 + 40 + 1513), divided by the sum of their d.f. (=6). F-Ratios (df_1 = 1; df_2 = 6) are 5.99 (5%) and 3.78 (10%). The test ratio F_B shows the variance to be "possibly significant, further evidence required".

The other variances are checked in the same way, leaving out those sums of squares that prove to be significant or possibly significant.

$$F_A = \frac{8450}{9844/5} = 4.29$$

F-ratios ($df_1 = 1$; $df_2 = 5$) are 6.61 (5%) and 4.06 (10%). The test ratio found indicates the variance to be "possibly significant, further evidence required".

$$F_{AB} = \frac{5202}{4642/4} = 4.48$$

F-ratios ($df_1 = 1$; $df_2 = 4$) are 7.71 (5%) and 4.51 (10%). This test ratio does not show significance. The same conclusion applies to the remaining variances in Table 3.2.

The test shows that only two variances proved possibly significant; further evidence is required to support this conclusion.

3.2 Replicates

More information can be obtained by repeating each test a number of times; in this case, 2 tests had originally been done for each combination of factors. The complete data and the Analysis of Variance are given in Tables 3.3 and 3.4.

Table 3.3 - Test Data (Duplicate Observations)

abc	aBc	Abc	ABc	abC	aBC	AbC	ABC
327	121	307	278	264	147	286	199
281	146	321	324	265	169	278	247
304	134	314	301	264	158	282	223

Table 3.4 - Analysis of Variance (Replicates)

Source of Variation		Sum of Squares (S.S.)	d.f.	Variance SS/d.f.	Test Ratio (F)
Main Effects	A	16,900	1	16,900	37.72 ***
	B	30,450	1	30,450	67.97 ***
	C	3,906	1	3,906	8.72 **
Interactions 1st Order	AB	10,506	1	10,506	23.45 **
	AC	2,256	1	2,256	5.04 *
	BC	81	1	81	--
2nd Order	ABC	3,025	1	3,025	6.75 **
Error		3,953	8	494	
Total		71,077	15	Compound Variance	-

$$\text{Test ratio} = \frac{\text{Variance tested}}{\text{Error variance}}$$

The equations used for calculating the Sums of Squares are the same as above (Eq. 1-5), except that, instead of 2^n, the denominator now reads $2^n \cdot H$ (H = number of replicates = 2).

The error variance is found from:

$$V_T = (\Sigma p_i^2 - \frac{\Sigma h_j^2}{2})/2^n(H-1) = 494,$$

where

p_i = individual observation (i),

h_j = sum of H replicates (j),

H = number of replicates.

Note that since the variance for the Interaction BC is smaller than the error variance, a new estimate of the error variance must be obtained: V_T = (81 + 3,953)/9 = 448. The test ratio F for the remaining variances is calculated on the basis of this new estimate:

$$F = \frac{\text{Variance tested}}{\text{Error variance}}$$

Theoretical F-ratios ($df_1 = 1$, $df_2 = 9$) are: 10.56 (1%), 5.12 (5%) and 3.36 (10%).

It appears that all but one of the variances now prove to be significant owing to the fact that duplicate observations were used. This latter test is said to be more "powerful" than the former one. Comparison of these two tests stresses the need for adequate test design to ensure meaningful results.

Conclusions

1) A drop in asphalt content from 5.54 to 4.37% (=1.17%) produces a reduction in compressive strength of 65 lb, or 56 lb per % asphalt. This is an average value for the two types of asphalt used. See conclusion 4.

2) A rise in initial moisture content from 8.1 to 11.9%(=3.8%) causes a reduction in compressive strength of 87 lb, or 23 lb per % moisture. In other words, for each percent more moisture in the coal, the asphalt content must be raised by 0.4% in order to maintain the same compressive strength for the briquets. This again is an average figure for the two types of asphalt. See conclusion 5.

3) The interaction variance AB shows that the effect of moisture is less detrimental at a high percentage of asphalt (9.5 lb per % moisture) than at a low percentage of asphalt (36 lb per % moisture).

	A	a
B	1048	583
b	1192	1137

4) Interaction AC shows that, possibly, natural asphalt increases compressive strength by 35 lb per % asphalt, whereas the cracked asphalt (type c) increases compressive strength by 76 lb per %

asphalt. Further evidence is required to confirm this conclusion, however.

	A	a
C	1010	845
c	1230	875

5) The 2nd-order interaction is significant and shows that the effect of initial moisture content differs somewhat for the two types of asphalt.

	A		a	
	B	b	B	b
C	446	564	316	529
c	602	628	267	608

For natural asphalt (type C), the compressive strength of the briquets decreases by 16 and 28 lb per % moisture (for high asphalt content and low asphalt content respectively). For cracked asphalt (type c), the compressive strength of the briquet decreases by 3 and 45 lb per % moisture. It appears that briquets made with cracked asphalt are more susceptible to moisture than those made with natural asphalt.

The system illustrated by the above examples can be applied to factorial tests with different numbers of factors. Condensed instructions given below allow the analysis of variance for up to seven factors.

3.3 Condensed Instructions for n - Factorial Analysis to a Maximum of 7 Factors

The sum of squares (S_x) is calculated as follows:

$$S_{\text{main effects}} = S_A = (\Sigma A - \Sigma a)^2\ /2^nH$$

$$S_B = (\Sigma B - \Sigma b)^2/2^nH, \text{ etc.}$$

$$S_{\text{Interaction}} = S_{\text{Int.}} = (\Sigma\square - \Sigma\boxdot)^2/2^nH$$

The tables below apply for each i^{th}-order interaction of any n-factorial test up to n = 7 factors.

$\Sigma\square$, $\boxdot$ = sum of all the individual observations indicated by the diagrams.

1st-order Interaction

	A	a
B	▨	
b		▨

2nd-order Interaction

	A		a	
	B	b	B	b
C		▨	▨	
c	▨			▨

3rd-order Interaction

		A		a	
		B	b	B	b
C	D	▨			▨
	d		▨	▨	
c	D		▨	▨	
	d	▨			▨

4th-order Interaction

		A				a			
		B		b		B		b	
		C	c	C	c	C	c	C	c
D	E		▨	▨		▨			▨
	e	▨			▨		▨	▨	
d	E	▨			▨		▨	▨	
	e		▨	▨		▨			▨

5th-order Interaction

			A				a			
			B		b		B		b	
			C	c	C	c	C	c	C	c
D	E	F	▨			▨		▨	▨	
		f		▨	▨		▨			▨
	e	F		▨	▨		▨			▨
		f	▨			▨		▨	▨	
d	E	F		▨	▨		▨			▨
		f	▨			▨		▨	▨	
	e	F	▨			▨		▨	▨	
		f		▨	▨		▨			▨

6th-order Interaction

			A								a							
			B				b				B				b			
			C		c		C		c		C		c		C		c	
			D	d	D	d	D	d	D	d	D	d	D	d	D	d	D	d
E	F	G		▨	▨		▨			▨	▨			▨		▨	▨	
		g	▨			▨		▨	▨			▨	▨		▨			▨
	f	G	▨			▨		▨	▨			▨	▨		▨			▨
		g		▨	▨		▨			▨	▨			▨		▨	▨	
e	F	G	▨			▨		▨	▨			▨	▨		▨			▨
		g		▨	▨		▨			▨	▨			▨		▨	▨	
	f	G		▨	▨		▨			▨	▨			▨		▨	▨	
		g	▨			▨		▨	▨			▨	▨		▨			▨

The <u>residual</u> <u>variance</u> is calculated by subtraction or as follows:

1) With replicates:

$$V_T = \frac{\Sigma p^2 - (\Sigma p)^2/2^n H - [\Sigma\square^2 + \Sigma\boxdot^2 - (\Sigma\square + \Sigma\boxdot)^2/2^n H]/H}{2^n(H-1)}$$

$\left.\begin{matrix}\square\\ \boxdot\end{matrix}\right\} =$ Sum of H replicates of 1 block. Total number of blocks = 2^n.

2) No replicates:

a) Find all variances and arrange them in descending order of magnitude.

b) Test largest variance with residual variance found from <u>all</u> the others. Continue with the 2nd largest variance, etc. A better way of testing the variances can be used if the variance of a single observation is known or can be found from a separate test. This error variance is then used in the denominator of the test-ratio, F.

<u>Note</u>

1) Interaction variances should preferably be smaller than the variances of the main effects. If any is larger, there is a possibility of improvement in the test procedure or with the choice of factors.

2) If any variance estimate turns out to be negative, there is an error in the calculation. Variances are never negative.

4. CORRELATION

Correlation is defined as the true relationship between two or more things, parts, etc. It is distinguished from spurious correlation, which has been discussed in an earlier section (par. 2.2) and which can be created unwittingly when choosing dependent parameters for these things, parts, etc.

It is assumed, in this section, that the credibility of the observed relationship has been established scientifically, and that the parameters used for measuring the variable things, parts, etc., do not have any element in common. It is wise to remember that many dimensionless ratios that are used as parameters and plotted against one another do have random elements in common (9).

The correlation of experimental data can be conveniently carried out in the form of graphs with two or more scales on which the data are plotted. As a rule, the resulting scatter diagram is first inspected visually. The experimenter then determines the basic equation of the "curve-of-best-fit". This is preferably a straight-line equation. Non-linear scatter should be linearized, if possible, by using a log-scale, a log-log scale, a probability-scale, combinations of such scales, or other forms of transformation of variables. Curvilinear correlation may also be used if preferred. A number of model equations in the form of Worksheets are presented in this section.

4.1 The Curve-of-Best-Fit

The condition for a "curve-of-best-fit" is that the sum of squares of the deviations of the observed points from the curve is a minimum. In other words,

$$s^2 \xrightarrow[\text{min}]{} \frac{\Sigma(p - P)^2}{n - 1}$$

and therefore

$$\Sigma(p - P)^2 \longrightarrow \text{minimum},$$

where

s = standard deviation,

p = observed value of point (on x- or y-axis), and

P = corresponding value on the curve.

The choice of axis for p depends upon the experimenter's decision regarding the source of errors involved:

I. Y only subject to error

II. X only subject to error

III. X and Y both subject to error

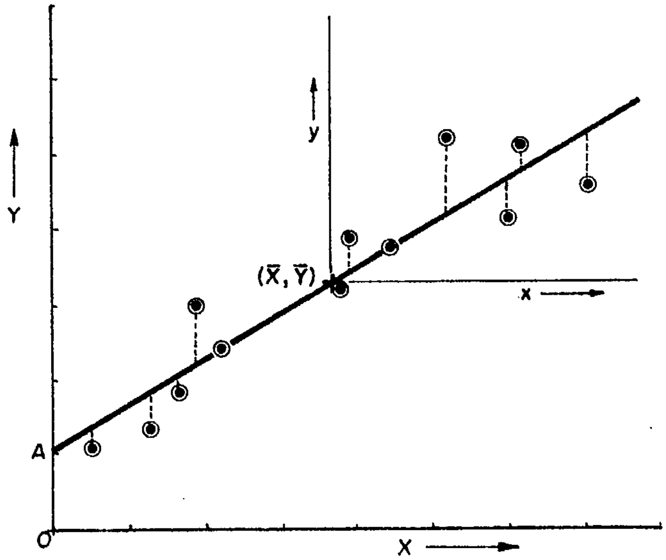


Fig. 4.1

4.1.1 Y only subject to error (Case I)

The condition to be fulfilled for the curve-of-best-fit is that the sum of squares of the vertical deviations (d_{yi}) is a minimum. Taking a curve of the form

$$Y = A + BX,$$

the origin can be shifted so that it coincides with the overall mean ($\overline{X}$, $\overline{Y}$), and a new system of coordinates y and x can be defined: $y = Y-\overline{Y}$, $x = X-\overline{X}$ (see Fig.4.1).

Equation of the curve then becomes:

$$y = Bx.$$

Since an observed point

$$y_i = y \pm d_{yi} = Bx_i \pm d_{yi}$$

then its vertical deviation from the curve,

$$\pm d_{yi} = Bx_i - y_i$$

and the sum of squares of the deviations for all points is

$$\Sigma(d_{yi})^2 = \Sigma(Bx_i - y_i)^2.$$

The condition that $\Sigma(d_{yi})^2$ is a minimum is fulfilled when

$$0 = 2\cdot\Sigma(Bx - y_i)x_i$$

and $$\Sigma Bx_i^2 - \Sigma x_i y_i = 0$$

Therefore, $$B_y = \frac{\Sigma xy}{\Sigma xx}$$

where xx stands for x^2. (B) is called the "regression coefficient" of the straight line through ($\overline{X}$, $\overline{Y}$). Geometrically, (B) is defined as the tangent of the angle between the regression curve and the horizontal axis.

Transforming back to the original system (X,Y), the constant (A) is then found from:

$$\boxed{A = \overline{Y} - B\overline{X}}$$

Constant (A) is the value of Y for X = 0, or, geometrically, the point at which the regression curve intercepts the vertical (Y) axis.

4.1.2 X only subject to error (Case II)

The condition to be fulfilled is that the sum of squares of the horizontal deviations (d_{xi}) is a minimum. (See Fig. 4.2.)

Cyclic replacement of x, y in the basic equation and proceeding as in the case I derivation,

$$1/B = \frac{\Sigma xy}{\Sigma yy}$$

and therefore,

$$\boxed{B_x = \tan\alpha = \frac{\Sigma yy}{\Sigma xy}}$$

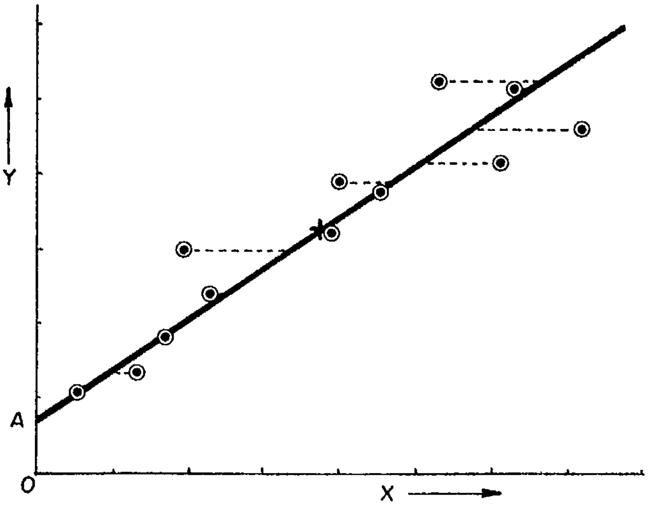


Fig. 4.2

4.1.3 Both X and Y subject to error (general case)

In this case, deviations from the line-of-best-fit are taken in some average direction which has been weighted according to the respective variabilities of X and Y (see Fig. 4.3). The regression coefficient,

$$B_{xy} = (kB_y + jB_x)/(k + j)$$

$$= (k \frac{\Sigma xy}{\Sigma x^2} + j \frac{\Sigma y^2}{\Sigma xy})/(k + j),$$

or

$$\boxed{B_{xy} = \frac{k(\Sigma xy)^2 + j(\Sigma x^2 \Sigma y^2)}{(k + j)\ (\Sigma x^2 \Sigma xy)}}$$

where k = variation coefficient of Y, = $s_y/\overline{Y}$,

j = variation coefficient of X, = $s_x/\overline{X}$.

The coefficients k and j are, in fact, weighting factors and can be determined from the reputed precisions ($s_{x,y}$) and average values of X and Y.

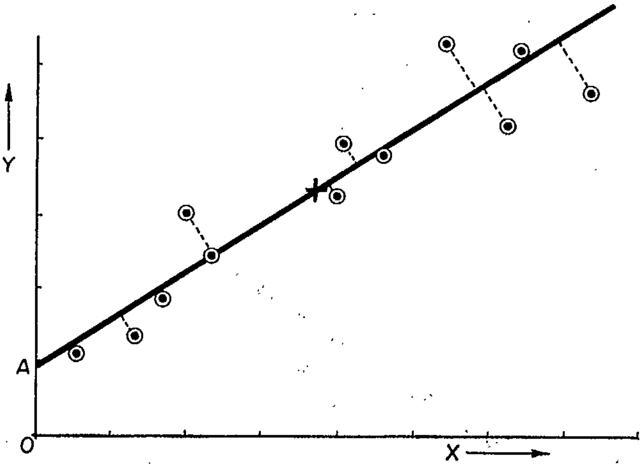


Fig. 4.3

Example:

The experimental relationship between the Btu value (Y) of a bituminous coal and its ash content (X) is found, from a series of 43 samples of this coal, for ash contents ranging from 14.7 to 20.6 percent. There are three possible expressions for this relationship:

I. Error in Y only: Y calc. = 13,337 - 132.33 X

II. Error in X only: Y calc. = 13,396 - 136.98 X

III. Error in both X and Y: Y calc. = 13,362 - 134.28 X.

In general terms, the most correct expression for a given relationship:

Assume Case I applies when $j < 1/2k$, i.e. when $B_y s_x < 1/2\ s_y$.

Assume Case II " " $k < 1/2j$, i.e. when $B_x s_x > 2\ s_y$.

Assume Case III " " $1/2(k,j) < (j,k) < 2(k,j)$,
i.e. when $1/2 s_y < B_{xy} s_x < 2\ s_y$.

4.2 Correlation Coefficient

The correlation coefficient (r) is a measure of the "goodness of fit" of the observations with respect to the regression curve. For perfect correlation, r = 1 and for complete lack of a relationship, r = 0.

The general equation for linear and non-linear regression for case I, error in Y only:

$$\boxed{r^2 = 1 - \frac{\Sigma d_y^2}{\Sigma y^2}}$$

The deviation of an observation from the overall mean is generally only partly explained by the regression curve. The residue (N) or % Not Explained, is found from:

$$\boxed{\%\ N = 100 \cdot \sqrt{1 - r^2}}$$

Table 4.1

r	% not explained (N)
1.00	0
0.99	13
0.98	19
0.95	31
0.90	44
0.80	60
0.60	80

For linear relationships, the following may be employed:

$$\boxed{r^2 = \frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma y^2}}$$

It is advisable to use linear correlation as much as possible by linearizing the relationship, i.e. by using a transformation variate.

4.3 Regression Analysis Worksheets

The worksheets presented in the sections that follow have been set up to deal with the Case I situation only, i. e., "Y only subject to error".

4.3.1 Linear correlation - one independent variable

Regression Formula:

$$\boxed{Y = A + BX}$$

X	Y	x	y	x^2	y^2	xy		Y'	d_y	d_y^2
		$(X-\overline{X})$	$(Y-\overline{Y})$			-	+	(=Y calc.)	(Y-Y') (2)-(9)	$(10)^2$
1	2	3	4	5	6	7	8	9	10	11

Coefficients:

$$B = \frac{\Sigma xy}{\Sigma x^2}$$

$$A = \overline{Y} - B\overline{X}$$

4.3.2 Linear correlation - one independent variable - weighted observations(grouped data)

Regression Formula: $\boxed{Y = A + BX}$

					x	y					nxy			d_y	
n	X	nX	Y	nY	$(X-\bar{X})$	$(Y-\bar{Y})$	x^2	y^2	nx^2	ny^2	-	+	Y'	$(Y-Y')$ (4)-(14)	nd_y^2
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Coefficients:

$$B = \frac{\Sigma nxy}{\Sigma nx^2}$$

$$A = \bar{Y} - B\bar{X}$$

4.3.3 Linear correlation - two independent variables

Regression Formula: $\boxed{Z = AY + BX + C}$

			x	y	z		zy		zx			yx				d_z	
X	Y	Z	$(X-\bar{X})$	$(Y-\bar{Y})$	$(Z-\bar{Z})$	z^2	-	+	-	+	y^2	-	+	x^2	Z'	(3)-(16)	d_z^2
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

Coefficients:

$$B = \frac{\Sigma zy\,\Sigma yx - \Sigma zx\Sigma y^2}{(\Sigma yx)^2 - \Sigma y^2\Sigma x^2}$$

$$A = \frac{\Sigma zy - B\,\Sigma yx}{\Sigma y^2}$$

$$C = \bar{Z} - A\bar{Y} - B\bar{X}$$

4.3.4 Linear correlation - two independent variables - weighted observations (grouped data)

Regression Formula : $\boxed{Z = AY + BX + C}$

n	X	nX	Y	nY	Z	nZ	x	y	z	nz^2	nzy		nzx		ny^2	nyx		nx^2	Z'	d_Z	nd_Z^2
							$(X-\overline{X})$	$(Y-\overline{Y})$	$(Z-\overline{Z})$		-	+	-	+		-	+			(6)-(20)	$n(21)^2$
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22

Coefficients:

$$B = \frac{\Sigma nzy \cdot \Sigma nyx \ - \ \Sigma nzx \cdot \Sigma ny^2}{(\Sigma nyx)^2 \ - \ \Sigma ny^2 \cdot \Sigma nx^2}$$

$$A = \frac{\Sigma nzy \ - \ B\Sigma nyx}{\Sigma ny^2}$$

$$C = \overline{Z} - A\overline{Y} - B\overline{X}$$

4.3.5 Linear correlation - three independent variables

Regression Formula: $\boxed{W = AY + BX + CZ + D}$

W	Y	X	Z	w	y	x	z	wy		y^2	yx		yz		wx		x^2	xz		wz		z^2	W'	d_W	d_W^2
				$(W-\overline{W})$	$(Y-\overline{Y})$	$(X-\overline{X})$	$(Z-\overline{Z})$	-	+		-	+	-	+	-	+		-	+	-	+			(1)-(24)	$(25)^2$
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26

Coefficients:

$$C = \frac{(\Sigma wy\Sigma yx - \Sigma wx\Sigma y^2)\cdot\overbrace{(\Sigma yx\Sigma yz - \Sigma y^2\Sigma xz)}^{Q} - \overbrace{(\Sigma wy\Sigma yz - \Sigma wz\Sigma y^2)}^{R}\cdot\left[(\Sigma yx)^2 - \Sigma y^2\Sigma x^2\right]}{\underbrace{(\Sigma yx\Sigma yz - \Sigma y^2\Sigma xz)}_{Q}{}^2 - \underbrace{\left[(\Sigma yz)^2 - \Sigma y^2\Sigma z^2\right]}_{S}\cdot\left[(\Sigma yx)^2 - \Sigma y^2\Sigma x^2\right]}$$

$$B = \frac{R - CS}{Q}$$

$$A = \frac{\Sigma wy - B\Sigma yx - C\Sigma yz}{\Sigma y^2}$$

$$D = \overline{W} - A\overline{Y} - B\overline{X} - C\overline{Z}$$

4.3.6 Linear correlation - three independent variables - weighted observations (grouped data)

Regression Formula: $\boxed{W = AY + BX + CZ + D}$

n	W	nW	Y	nY	X	nX	Z	nZ	w	y	x	z	nw^2	nwy		nwx		nwz	
									$(W-\overline{W})$	$(Y-\overline{Y})$	$(X-\overline{X})$	$(Z-\overline{Z})$		-	+	-	+	-	+
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

ny^2	nyx		nyz		nx^2	nxz		nz^2	W'	d_W	nd_W^2
	-	+	-	+		−	+			(2)-(30)	$n(31)^2$
21	22	23	24	25	26	27	28	29	30	31	32

Coefficients:

$$C = \frac{(\Sigma nwy\Sigma nyx - \Sigma nwx\Sigma ny^2)\cdot\overbrace{(\Sigma nyx\Sigma nyz - \Sigma ny^2\Sigma nxz)}^{Q'} - \overbrace{(\Sigma nwy\Sigma nyz - \Sigma nwz\Sigma ny^2)}^{R'}\cdot\left[(\Sigma nyx)^2 - \Sigma ny^2\Sigma nx^2\right]}{\underbrace{(\Sigma nyx\Sigma nyz - \Sigma ny^2\Sigma nxz)}_{Q'}{}^2 - \underbrace{\left[(\Sigma nyz)^2 - \Sigma ny^2\Sigma nz^2\right]}_{S'}\cdot\left[(\Sigma nyx)^2 - \Sigma ny^2\Sigma nx^2\right]}$$

$$B = \frac{R' - CS'}{Q'}$$

$$A = \frac{\Sigma nwy - B\Sigma nyx - C\Sigma nyz}{\Sigma ny^2}$$

$$D = \overline{W} - A\overline{Y} - B\overline{X} - C\overline{Z}$$

4.4 Simplified Calculations for Large Numbers of Observations

When a large number of observations are to be treated, and provided that the variables are equally spaced, calculation can be greatly simplified by grouping of the observations and rank-numbering of the groups.

This is illustrated in the table below, where frequencies are recorded for various class-intervals of both X and Y. The simple numbers 1, 2, 3, ... given as the values of X and Y in the table are the "rank-numbers" and replace, for ease of calculation, the true value (e.g., mid-point) of each group (class).

4.4.1 Linear correlation - one independent variable

Regression Formula: $\boxed{Y = A + BX}$

TABLE 4.2 Frequency Table of X and Y

		X									Total
		1	2	3	4	5	6	7	8	9	No.
	7	1	2	1							4
	6		1	3	3	1					8
	5		1	-	6	4	1				12
Y	4		1	-	5	10	10	5	1		32
	3			1	6	15	20	15	6	1	64
	2				1	4	11	9	6	-	31
	1						4	6	4	1	15
Total No.		1	5	5	21	34	46	35	17	2	166

(Halves move to the right)

Calculations for the Line-of-Best-Fit are:

1. Sum of squares of variable X about the Mean;
2. Sum of squares of variable Y about the Mean;
3. Sum of Products of X and Y about the Means.

1. Sum of Squares of Variable X about Mean

Sum X = (Column Totals) $\cdot$ X = ΣX=(1 x 1) + (5 x 2) + ...(2 x 9)= 955

Crude Sum Squares = ΣX^2= (1 x 1^2) + (5 x 2^2) + ...(2 x 9^2) = 5873

Correction due to Mean = $(\Sigma X)^2/n$ = $(955)^2/166$ = 5494

$\Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/n$ = (5873 - 5494) = 379

Variance (V) = $\Sigma x^2/(n-1)$ = 379/165 = 2.30

2. Sum of Squares of Variable Y about Mean

Sum Y = (Row Totals)$\cdot$Y = ΣY = (15 x 1) + (31 x 2) + ...(4 x 7)=533

Crude Sum Squares =ΣY^2=(15x1^2) + (31x2^2) + ...(4x7^2) = 2011

Correction term = $(\Sigma Y)^2/n$ = 1711

$\Sigma y^2 = \Sigma Y^2 - (\Sigma Y)^2/n$ = (2011 - 1711) = 300

Variance (V) = $\Sigma y^2/(n-1)$ = 300/165 = 1.82

3. Sum of Products of X and Y about Means

Sum of Products = $\Sigma f_i \cdot X_i \cdot Y_i$, where f_i = frequency of $(X_i Y_i)$.

As shown in Table 4.3 below, fill in for each value of X the sums obtained as follows:

Column 1 (X = 1); $\Sigma f_i \cdot Y_i$ = (1 x 7) = 7

X = 2; $\Sigma f_i \cdot Y_i$ = (1 x 4) + ... (2 x 7) = 29

⋮

X = 9; $\Sigma f_i \cdot Y_i$ = (1 x 1) + (1 x 3) = 4

TABLE 4.3

X	1	2	3	4	5	6	7	8	9	Total
$\Sigma f_i Y_i$	7	29	28	88	119	131	89	38	4	533
$X\Sigma f_i Y_i$	7	58	84	352	595	786	623	304	36	2845

Crude Sum of Products = $X_i \Sigma f_i Y_i$ = ΣXY = 2845

Correction term = $\Sigma X \Sigma Y/n$ = (955)$\cdot$(533)/166 = 3066

$\underline{\Sigma xy} = \Sigma XY - \Sigma X \Sigma Y/n = 2845 - 3066 = \underline{-221}$

For this example, the correlation coefficient r, found from $r = \sqrt{\frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma y^2}} = \sqrt{\frac{(-221)^2}{(379)\cdot(300)}} = 0.656$, indicates a high degree of correlation for 164 degrees of freedom. (See table, "Significance of Correlation Coefficient", in Appendix B.)

TABLE 4.4 Analysis of Variance of Regression

Source of Variation	Sum Squares	d.f.	Variance
Regression	$r^2 \Sigma y^2 = (\Sigma xy)^2 / \Sigma x^2 =$ 129	1	129
Remainder	$(1-r^2)\Sigma y^2 = \Sigma d_y^2 =$ 171*	164	1.04
Total	$\Sigma y^2 =$ 300	n-1=165	—

$$* \ \Sigma d_y^2 = \Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2}$$

The F-ratio, F = 129/1.04 = 124.04, is highly significant for 1 and 164 d.f. respectively.

Coefficients for the regression formula Y = A + BX:

$$B = \frac{\Sigma xy}{\Sigma x^2} = \underline{-0.58}$$

$$A = \frac{\Sigma Y}{n} - \frac{B \Sigma X}{n} = \frac{533}{166} - (-0.58) \cdot \frac{(955)}{166} = \underline{6.55}$$

Equation for the Line-of-Best-Fit:

$$\boxed{Y = 6.55 - 0.58X}$$

4.42 Linear correlation - two independent variables

Regression Formula: $\boxed{Z = AY + BX + C}$

TABLE 4.5 Frequency Table of X, Y, and Z

X	Y	Z						Total	
		1	2	3	4	5	6		
7	5								7
	4				2	-	1	3	
	3				2	1	1	4	
	2								
	1								
6	5								44
	4			-	17	9	6	32	
	3			2	6	4	-	12	
	2								
	1								
5	5		1	-	1	3	-	5	80
	4		4	12	14	21	5	56	
	3		-	8	9	2	-	19	
	2								
	1								
4	5	-	1	2	3	2		8	79
	4	3	18	20	11	6		58	
	3	-	3	8	-	-		11	
	2			-	-			-	
	1			1	1			2	
3	5	1	1	3	1			6	73
	4	10	30	13	7			60	
	3	2	2	2	1			7	
	2								
	1								
2	5	3	5	1	1			10	48
	4	8	21	4	2			35	
	3			3				3	
	2								
	1								
1	5		-	2				2	12
	4		4	4				8	
	3		1	1				2	
	2								
	1								
Total		27	91	86	78	48	13	343	

Coefficients:

$$B = \frac{\Sigma zy\Sigma x^2 - \Sigma zx\Sigma xy}{\Sigma y^2\Sigma x^2 - (\Sigma xy)^2}$$

$$A = \frac{\Sigma xz\Sigma y^2 - \Sigma yz\ \Sigma xy}{\Sigma x^2\Sigma y^2 - (\Sigma xy)^2}$$

$$C = \frac{\Sigma Z - B\Sigma Y - A\Sigma X}{n}$$

Correlation coefficient:

$$r^2 = \frac{B\Sigma yz + A\Sigma xz}{\Sigma z^2}$$

Calculations for the Line-of-Best-Fit are:

1. $\Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/n$
2. $\Sigma y^2 = \Sigma Y^2 - (\Sigma Y)^2/n$
3. $\Sigma z^2 = \Sigma Z^2 - (\Sigma Z)^2/n$
4. $\Sigma xy = \Sigma XY - \Sigma X\Sigma Y/n$
5. $\Sigma xz = \Sigma XZ - \Sigma X\Sigma Z/n$
6. $\Sigma yz = \Sigma YZ - \Sigma Y\Sigma Z/n$

1. Calculation of Σx^2:

Sum X = $\Sigma X = (12 \times 1) + (48 \times 2) + \ldots (7 \times 7) = \underline{1356}$

Crude sum squares = $\Sigma X^2 = (12 \times 1^2) + (48 \times 2^2) + \ldots (7 \times 7^2) = \underline{6052}$

Correction: $(\Sigma X)^2/n = \underline{5361}$

$$\boxed{\Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/n = 691}$$

2. Calculation of Σy^2:

Sum Y = $\Sigma Y = (2 \times 1) + (58 \times 3) + (252 \times 4) + (31 \times 5) = \underline{1339}$

Crude sum squares = $\Sigma Y^2 = (2 \times 1^2) + (58 \times 3^2) + \ldots (31 \times 5^2) = \underline{5331}$

Correction: $(\Sigma Y)^2/n = \underline{5227}$

$$\boxed{\Sigma y^2 = 104}$$

3. Calculation of Σz^2:

Sum Z = $\Sigma Z = (27 \times 1) + (91 \times 2) + \ldots (13 \times 6) = \underline{1097}$

Crude sum squares = $\Sigma Z^2 = (27 \times 1^2) + (91 \times 2^2) + \ldots (13 \times 6^2) = \underline{4081}$

Correction: $(\Sigma Z)^2/n = \underline{3508}$

$$\boxed{\Sigma z^2 = 573}$$

4. Calculation of Σxy:

TABLE 4.6

X	1	2	3	4	5	6	7	Total
ΣY	2x3= 6 4x8=32 2x5=10 48	3x3= 9 35x4=140 10x5= 50 199	7x3= 21 60x4=240 6x5= 30 291	2x1= 2 11x3=33 58x4=232 8x5=40 307	19x3= 57 56x4=224 5x5=25 306	12x3=36 32x4=128 164	4x3=12 3x4=12 24	
ΣXY	48	398	873	1228	1530	984	168	5229

Crude Sum = ΣXY = 5229

Correction = $\Sigma X\Sigma Y/n$ = 5294

$$\boxed{\Sigma xy = \Sigma XY - \Sigma X\Sigma Y/n = -65}$$

5. Calculation of Σxz:

TABLE 4.7

X	1	2	3	4	5	6	7	Total
Σ Z	5x2=10 7x3=21 31	- - 99	- - 169	- - 240	- - 326	- - 199	- - 33	
ΣXZ	31	198	507	960	1630	1194	231	4751

Crude Sum = ΣXZ = 4751

Correction = $\Sigma X\Sigma Z/n$ = 4337

$$\boxed{\Sigma xz = 414}$$

6. Calculation of Σyz:

TABLE 4.8

Z	1	2	3	4	5	6		Total
ΣY	2x3 = 6 21x4 =84 4x5 =20 110	- - - 366	- - - 325	- - - 297	- - - 190	- - - 51	- - - -	
ΣYZ	110	732	975	1188	950	306		4261

Crude Sum = ΣYZ = 4261

Correction = $\Sigma Y \Sigma Z/n$ = 4282

$$\boxed{\Sigma yz = -21}$$

Coefficients:

$$B = \frac{(-21 \times 691) - (414 \times -65)}{(104 \times 691) - (-65)^2} = \underline{0.1833}$$

$$A = \underline{0.6164}$$

$$C = \underline{0.0458}$$

For this example, equation of the Line-of-Best-Fit:

$$\boxed{Z = 0.0458 + 0.1833X + 0.6164Y}$$

The correlation coefficient

$$r^2_{z.xy} = \frac{0.1833 \times (-21) + (0.6164 \times 414)}{573} = 0.4386$$

$$r_{zxy} = \underline{0.6623}$$

indicates a high degree of correlation for 340 degrees of freedom.

4.5 Example of Regression Analysis with Grouped Data

Correlation of Summer Temperature vs. Length of Fir Shoots:

Assuming a linear relationship,

Regression Formula $\boxed{Y = A + BX}$

X = temperature, °C; Y = length of shoot, mm.;

The data have been weighted, i.e., X and Y have been independently grouped in classes of equal interval and the frequency (n) of each class recorded. In calculating correlation, the mid-point of each class is taken as the value corresponding to X and Y respectively.

TABLE 4.9

X Class-Interval	n_x	Y Class-Interval	n_y
6.45 - 6.95	1	35 - 45	1
6.95 - 7.45	1	45 - 55	3
7.45 - 7.95	1	55 - 65	5
7.95 - 8.45	1	65 - 75	11
8.45 - 8.95	6	75 - 85	8
8.95 - 9.45	8	85 - 95	6
9.45 - 9.95	6	95 - 105	2
9.95 - 10.45	8	105 - 115	1
10.45 - 10.95	1	115 - 125	1
10.95 - 11.45	2	125 - 135	1
11.45 - 11.95	1		
11.95 - 12.45	1		
12.45 - 12.95	2		
	39		39

In order to use the regular tables shown earlier (Tables 4.3.1 to 4.3.6) for calculating correlation, the number of classes of X and Y must be equal, and the frequency (n) of each class-interval of X must equal that of the corresponding class-interval of Y. Neither of these conditions can be met here without difficulty. It is simpler in such a case to construct a table of frequencies from the original data as shown on the following page.

Regression Formula : $\boxed{Y = A + BX}$

X = Average Summer Temperature, °C;

Y = Length of Fir Shoots, mm.

TABLE 4.10 Frequencies of X vs Y

		X = Class Mid-point													
		6.7	7.2	7.7	8.2	8.7	9.2	9.7	10.2	10.7	11.2	11.7	12.2	12.7	Total
Y = Class Mid-point	130													1	1
	120													1	1
	110												1		1
	100								1		1				2
	90						2		2		1	1			6
	80					1	2	1	3	1					8
	70						4	5	2						11
	60		1	1		3									5
	50				1	2									3
	40	1													1
Total		1	1	1	1	6	8	6	8	1	2	1	1	2	39

This table shows the true distribution and frequencies of the class-intervals of each variable with r espect to one another, and is simply another form of the scatter diagram.

Calculation of the Line-of-Best-Fit

For Y = A + BX

Coefficients:

I. $B_y = \Sigma xy/\Sigma x^2$;

II. $B_x = \Sigma y^2/\Sigma xy$;

III. $B_{xy} = (kB_y + jB_x)/(k + j)$

$$A = \overline{Y} - B\overline{X} = \frac{\Sigma Y}{n} - \frac{B\Sigma X}{n}$$

1. Calculation of Σx^2

$\Sigma X = (1 \times 6.7) + (1 \times 7.2) + \ldots (2 \times 12.7) = 377.8$

Crude Sum Squares= $\Sigma X^2 = (1 \times 6.7^2) + (1 x 7.2^2) + \ldots (2 x 12.7^2) = 3,725.06$

Correction = $(\Sigma X)^2/n = (377.8)^2/39 = 3,659.82$

$\Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/n = 3725.06 - 3659.82 = 65.24$

Variance = $s_x^2 = \Sigma x^2/(n-1) = 65.24/38 = 1.72$; standard error, $s_x = 1.31$

2. Calculation of Σy^2

$\Sigma Y = (1 \times 40) + (3 \times 50) + \ldots (1 \times 130) = 3,000.00$

Crude Sum Squares= $\Sigma Y^2 = (1 x 40^2) + (3 x 50^2) + \ldots (1 x 130^2) = 244,200$

Correction = $(\Sigma Y)^2/n = (3,000)^2/n = 230,769$

$\Sigma y^2 = \Sigma Y^2 - (\Sigma Y)^2/n = 244,200 - 230,769 = 13,431$

Variance = $s_y^2 = 13,431/38 = 353.45$; standard error, $s_y = 18.80$

3. Calculation of Σxy

Column 1 $(X = 6.7)$; $\Sigma Y = (1 \times 40) = 40$

$X = 7.2$; $\Sigma Y = (1 \times 60) = 60$

⋮

$X = 12.7$; $\Sigma Y = (1 \times 120) + (1 \times 130) = 250$

TABLE 4.11

X	6.7	7.2	7.7	8.2	8.7	9.2	9.7	10.2	10.7	11.2	11.7	12.2	12.7	Total
ΣY	40	60	60	50	360	620	430	660	80	190	90	110	250	3,000
XΣY	268	432	462	410	3132	5704	4171	6732	856	2128	1053	1342	3175	29,865

Crude Sum Products = $X\Sigma Y = 29,865$

Correction = $(\Sigma X \cdot \Sigma Y)/n = (377.8 \times 3,000)/39 = 29,062$

$\Sigma xy = 29,865 - 29,062 = 803$

It will be recalled that, depending upon which of the variables is most subject to error, one of three possible equations may be used for expressing the relationship (see Fig. 4.4). The three are given for illustration purposes:

I. Y subject to error only:

If variations in the data are assumed or are known to be largely due to measurement errors or to the variability of the length of fir shoots (Y):

$$B = \Sigma xy/\Sigma x^2 = 803/65.24 = \underline{12.31}$$

$$A = 3000/39 - B(377.8/39) = \underline{-42.33}$$

$$\boxed{Y \text{ calc.} = 12.31X-42.33}$$

II. X subject to error only:

If variations are largely due to measurement errors or to variability of temperature (X):

$$B = \Sigma y^2/\Sigma xy = \underline{16.73}$$

$$A = 76.9 - (16.73 \text{ x } 9.7) = \underline{-85.14}$$

$$\boxed{Y \text{ calc.} = 16.73X-85.14}$$

III. Both X and Y subject to error:

If variability or measurement errors are not largely confined to either X or Y:

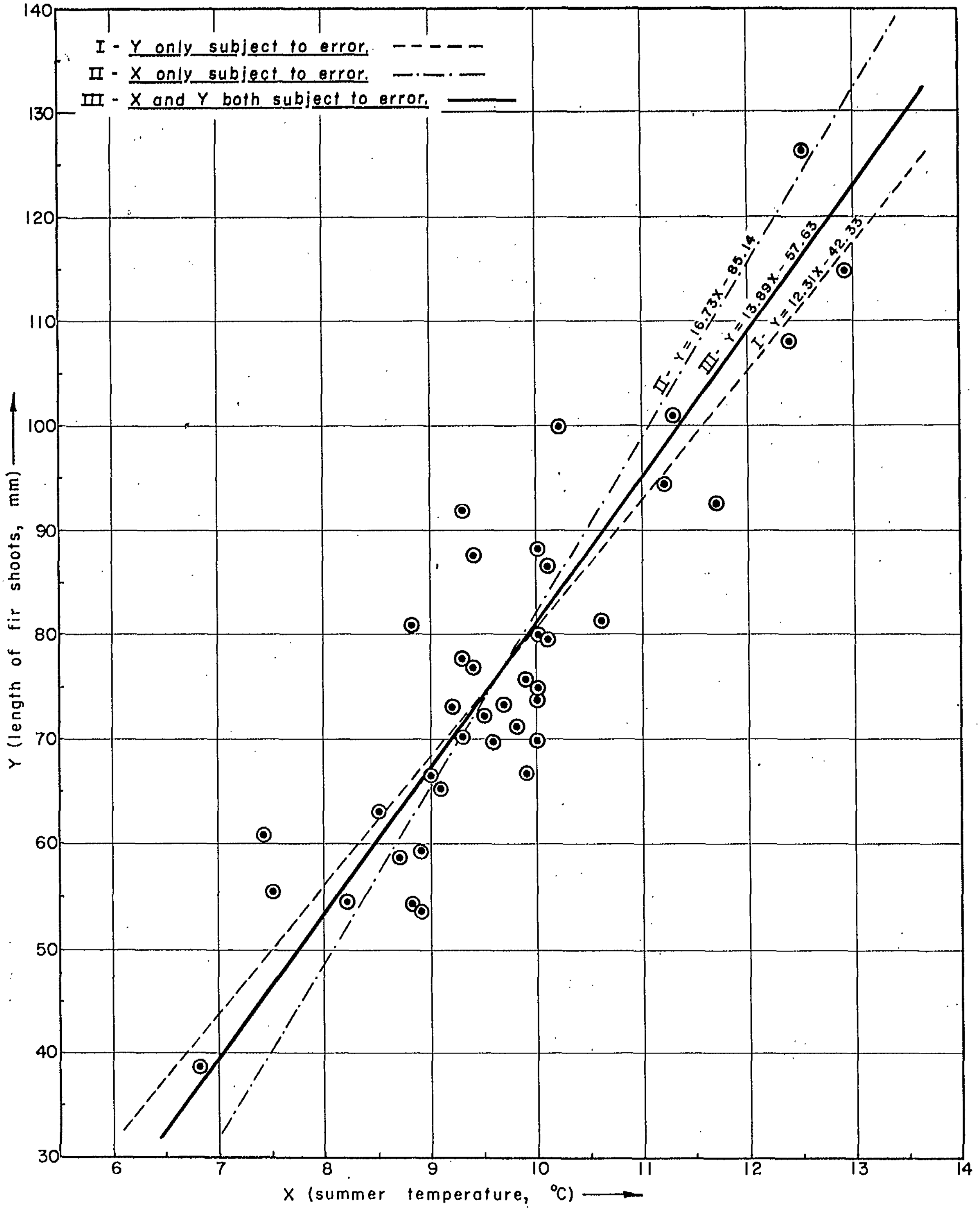
Variation Coefficients: $k = 0.244$, $j = 0.136$.

$$B = (kBy + jBx)/(k + j) = \underline{13.89}$$

$$A = 76.9 - (13.89)(9.7) = \underline{-57.63}$$

$$\boxed{Y \text{ calc.} = 13.89X - 57.63}$$

Fig. 4.4 - Correlation of Length of Fir Shoots vs. Summer Temperature



Referring back to 4.1.3, it is seen that for $k = 0.244$ and $j = 0.136$, Case III applies (both X and Y subject to error) since $1/2\ k < j < 2k$. In other words, the most accurate estimate of Y, length of fir shoot for a temperature X, will be given by equation III in this example, i.e. the weighted average curve of the two independent regressions.

Correlation coefficient :

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{803}{\sqrt{(65.24)(13{,}431)}} = \underline{0.857}$$

This value of (r) denotes a high degree of correlation for 37 degrees of freedom.

4.6 Worksheets - Non-Linear Regression

Worksheets that follow are for calculation of 2nd-, 3rd- and 4th-degree curves-of-best-fit for two variables.

4.6.1 Non-linear correlation - 2nd degree

Regression Formula : $\boxed{y = a + bx + cx^2}$

X	Y	x	y	x^2	y^2	xy		x^2y		x^3		x^4	Y'	d_y	d_y^2
		$(X-\bar{X})$	$(Y-\bar{Y})$			−	+	−	+	−	+			(2)-(14)	
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Regression coefficients:

$$c = \frac{\Sigma x^2 y\, \Sigma x^2 - \Sigma xy \Sigma x^3}{\Sigma x^4 \Sigma x^2 - (\Sigma x^2)^3/n - (\Sigma x^3)^2}$$

$$b = \frac{\Sigma xy - c\Sigma x^3}{\Sigma x^2}$$

$$a = -\frac{c\Sigma x^2}{n}$$

Correlation coefficient:

$$\boxed{r^2 = 1 - \frac{\Sigma d_y^2}{\Sigma y^2}}$$

To obtain equation on basis of the original units X, Y, substitute $Y - \bar{Y} = y$ and $X - \bar{X} = x$ in the regression formula for y and determine new coefficients:

$C = c$

$B = b - 2c\bar{X}$

$A = a - b\bar{X} + c\bar{X}^2 + \bar{Y}$

Regression Formula, original units:

$$\boxed{Y = A + BX + CX^2}$$

4.6.2 Non-linear correlation - 2nd degree - weighted observations (grouped data)

Regression Formula : $\boxed{y = a + bx + cx^2}$

n	X	nX	Y	nY	x $(X-\bar{X})$	y $(Y-\bar{Y})$	x^2	y^2	nx^2	ny^2	nxy		nx^2y		nx^3		nx^4	Y'	d_y (4)-(19)	d_y^2
											-	+	-	+	-	+				
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21

Regression coefficients :

$$c = \frac{\Sigma nx^2y\Sigma nx^2 - \Sigma nxy\Sigma nx^3}{\Sigma nx^4\Sigma nx^2 - (\Sigma nx^2)^3/n - (\Sigma nx^3)^2}$$

$$b = \frac{\Sigma nxy - c\Sigma nx^3}{\Sigma nx^2}$$

$$a = \frac{-c\Sigma nx^2}{n}$$

Original Units :

Coefficients:

$C = c$

$B = b - 2c\bar{X}$

$A = a - b\bar{X} + c\bar{X}^2 + \bar{Y}$

Regression Formula:

$Y = A + BX + CX^2$

4.6.3 Non-linear correlation - 3rd degree

Regression Formula : $\boxed{y = a + bx + cx^2 + dx^3}$

For sets of single observations, the following columns should be added to worksheet 4.6.1:

x^3y	x^5	x^6

Simultaneous equations for regression coefficients:

$$d = \frac{\Sigma x^3y - a\Sigma x^3 - b\Sigma x^4 - c\Sigma x^5}{\Sigma x^6}$$

$$c = \frac{\Sigma x^2y - a\Sigma x^2 - b\Sigma x^3 - d\Sigma x^5}{\Sigma x^4}$$

$$b = \frac{\Sigma xy - c\Sigma x^3 - d\Sigma x^4}{\Sigma x^2}$$

$$a = -\frac{c\Sigma x^2 - d\Sigma x^3}{n}$$

Convert to original units : $Y = A + BX + CX^2 + DX^3$

$D = d$

$C = c - 3d\bar{X}$

$B = b - 2c\bar{X} + 3d\bar{X}^2$

$A = a - b\bar{X} + C\bar{X}^2 - D\bar{X}^3 + \bar{Y}$

Similarly, for weighted observations, the following columns should be added to worksheet 4.6.2:

nx^3y	nx^5	nx^6

4.6.4 4th degree - non-linear regression

Regression Formula : $\boxed{y = a + bx + cx^2 + dx^3 + ex^4}$

Add the following columns to worksheet 4.6.1:

x^3y	x^5	x^6	x^4y	x^7	x^8

Simultaneous equations for regression coefficients :

$$e = \frac{\Sigma x^4y - a\Sigma x^4 - b\Sigma x^5 - c\Sigma x^6 - d\Sigma x^7}{\Sigma x^8}$$

$$d = \frac{\Sigma x^3y - a\Sigma x^3 - b\Sigma x^4 - c\Sigma x^5 - e\Sigma x^7}{\Sigma x^6}$$

$$c = \frac{\Sigma x^2y - a\Sigma x^2 - b\Sigma x^3 - d\Sigma x^5 - e\Sigma x^6}{\Sigma x^4}$$

$$b = \frac{\Sigma xy - c\Sigma x^3 - d\Sigma x^4 - e\Sigma x^5}{\Sigma x^2}$$

$$a = -\frac{c\Sigma x^2 - d\Sigma x^3 - e\Sigma x^4}{n}$$

Convert to original units : $Y = A + BX + CX^2 + DX^3 + EX^4$

$$E = e$$

$$D = d - 4e\bar{X}$$

$$C = c - 3d\bar{X} + 6e\bar{X}^2$$

$$B = b - 2c\bar{X} + 3d\bar{X}^2 - 4e\bar{X}^3$$

$$A = a - b\bar{X} + c\bar{X}^2 - d\bar{X}^3 + e\bar{X}^4 + \bar{Y}$$

Similarly, for weighted observations, the following columns should be added to worksheet 4.6.2:

nx^3y	nx^5	nx^6	nx^4y	nx^7	nx^8

5. SAMPLING

In this section a general theory(10) is presented for the sampling of materials and statistical populations that are encountered in the field of engineering and in other disciplines.

The basic concept of this theory is that the variability of a material consignment, or of any other population, can be expressed by two variance components that are constants and reflect statistical properties of the material in the same way that other material constants reflect certain physical and chemical properties of materials.

These variance components or "sampling constants" are used for designing sampling experiments and, more specifically, for determining in advance the precision of a sampling experiment as a function of sample size and the number of increments. Application of the method is illustrated with twelve examples that cover a large variety of materials and conditions. Condensed instructions are presented in a table.

5.1 Notes on the Problems of Sampling

When it is required to measure some property or attribute of a large volume of material or some other statistical population, a small representative portion is collected as a sample for testing. Sometimes, as in opinion polls, the information can be obtained without actually "collecting" the sample. This does not, however alter the procedure that follows.

The sample value will generally differ from the true, unknown value of the material consignment. This difference, called sampling error, has a frequency distribution with a mean value and a variance.

It is necessary to estimate this sampling error before the quality evaluation can be reported with any degree of assurance or precision. The sampling error depends upon the nature of the material and on the manner of sampling. These two main factors are expressed as variance components that contribute to the overall sampling variance, s^2.

Let the true unknown value of a variate (Q) of a material consignment be X; the sample value of this same variate will have a value x, and the difference (x - X) will be characterized by a frequency distribution with standard error, s. The sampling error can then be expressed as a function of s.

Materials and variates may vary over wide ranges and the circumstances under which the samples are collected can vary widely, but the causes of variation in sample value are limited. Two factors are inherent in the nature of the consignment, namely, "random variation" and "segregation". These are statistical properties in the nature of material constants, that will be termed "sampling constants" (A and B, respectively) of the population. A and B are variance components that can either be determined from a specially designed test or be estimated from prior knowledge if the material is known by composition and distribution.

The other factors influencing precision of the sample are the number (N) of increments collected from all parts of the lot, and the size (W) of the resultant gross sample. W and N are in the nature of operating variables that can, within certain limits, be regulated at will by the sampler. The equation $s^2 = A/W + B/N$ operates independently of the shape of the parent distribution of the variate。 It applies generally for first-order estimates of the upper limit of the sampling variance (s^2). The degree of segregation (z) is described by $z^2 = B/A$.

Every sampling operation consists essentially of either extracting one single sample from a given quantity of material or extracting from different parts of the lot a series of small portions or "increments" that are combined into one "gross sample". The latter method, known as "sampling by increments", will be considered here. The former method can be regarded as a special case of incremental sampling in which the number of increments equals one.

5.1。1 Comparison with other sampling theory

One theory for sampling materials that are non-randomly distributed is known as "stratified sampling" or "representative (random) sampling (of stratified populations)"。 In this theory, the precision of sampling is expressed as the sum of the variance "within-strata" and the variance "between-strata", the strata indicating parts of the material consignment whose mean values differ significantly from the overall mean value of the consignment. Sometimes, as in incremental sampling, these "strata" are imaginary, because they become identical to the portions represented by the individual increment. The "within" and "between" variance estimates are then a function of the size and number of increments。 It is common to identify the "between-strata" variance with the "trend variance", and the "within-strata" variance with the "random variance"。 Clearly, however, with different size and number of increments, the estimates of the between-strata variance and the within-strata variance will change. These variance estimates, therefore, cannot be regarded as constants and consequently cannot be used without certain corrections for calculating in advance the number and size of increments required to attain a projected over-all precision of sampling.

The meaning of "random sampling error" as used here goes back to a classical experiment where a number of black and white balls are mixed in a vase and a sample consisting of one or more balls is withdrawn at random. The random error occurs when the hand collecting the sample selects by chance a white ball instead of a black ball, or vice versa. The resulting variance is the "random variance", of which the "within-strata variance" used in representative sampling gives a biased estimate (depending upon the size of the samples used) when dealing with materials that are non-randomly distributed. This random variance is determined by the average composition of the material (in this case the relative amount of black or white balls) and by the size of the sample only. The same definition of random variance is adopted for variates with parent distributions that are not of the binomial type.

5.1.2 Definitions

In this section, the term "random variance" keeps its original meaning. "Trend variance" has been omitted because of its confusing nature; in its place a new term, "segregation variance", is introduced which denotes the variance caused solely by deviations resulting from the non-random distribution of a consignment. Its physical meaning is simple to explain: the deviation of any sample value from the true mean of the lot or consignment is the algebraic sum of its random error and a remaining error which results from the fact that the variate is non-randomly distributed over the lot. The latter is called the segregation error and its variance the segregation variance. It will be shown that the segregation variance component of single samples is independent of sample size; it depends only upon the degree of segregation of the consignment. It will also be shown that the maximum degree of segregation, as expressed by the variance of segregation, is directly related to the random variance. This relationship is utilized to estimate sampling precision.

5.2 General Sampling Theory

When a sample of given size is drawn from an infinite population, its theoretical variance is always larger than if a sample of the same size were drawn from a finite population having otherwise identical characteristics.

The fact that in practice all populations are finite does not necessarily invalidate the theoretical estimate of the variance, provided stipulation is made that it is an estimate of the maximum value that this variance will attain for an infinite population. The same problem is encountered when samples are drawn either systematically or at random from a stratified population. Samples that straddle the boundaries between two strata contribute less to the sampling variance estimate than those that are drawn wholly from individual strata. The latter variance estimate is consequently always larger than the former.

A model population is introduced to demonstrate the fundamental relationship and its general applicability. Variance values found from tests on this model are only maximum estimates, because the model represents the conditions that cause the largest possible variations. Sampling variances derived from the tests are accurate by first-order approximation only. Conditions other than those governing test results from the model will lead to variance estimates that are smaller, as for instance when the samples are very large or when the population is relatively small. Other conditions are discussed in the text.

The above limitations do not seriously interfere with the requirements of industry regarding the testing and safeguarding of quality.

5.2.1 Model population

The model population of "black" and "non-black" items, as illustrated by a "sampling board" (Figure 5.1), is used for analyzing variability of samples drawn from segregated consignments.

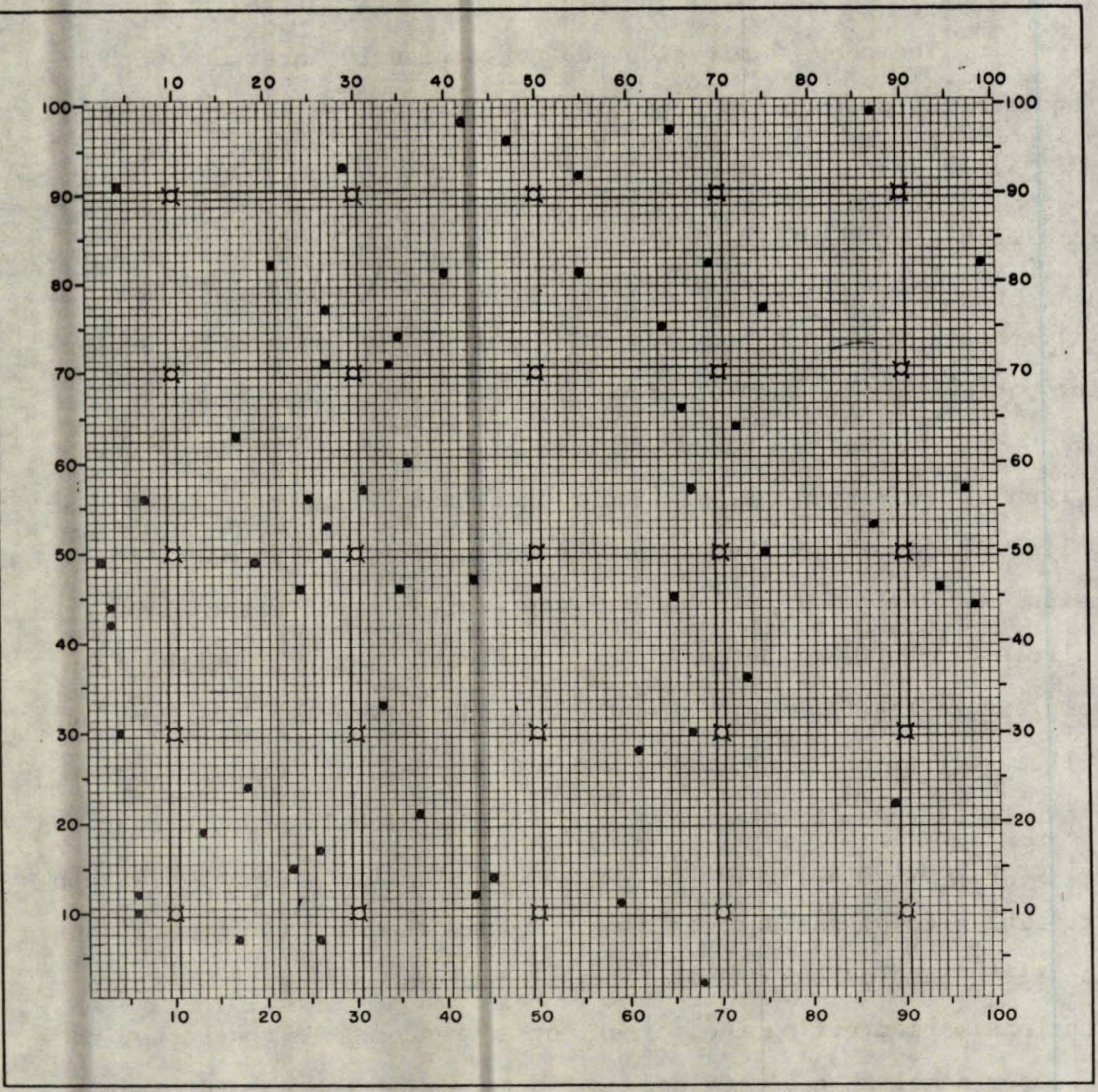
This sampling board consists of a piece of 10" x 10" wire screen with 10 openings per linear inch, and a supply of 5,000 lead pellets. The lead pellets can be used entirely, or in part, for making model populations that are segregated in different ways. The pellets can be distributed in any conceivable manner, ranging from complete segregation to near-perfect random mixtures. The samples collected from this population are not removed but merely counted. A sample is taken by placing a square frame with its centre over the selected station and counting the number of pellets enclosed within it. The size of the samples can thus be varied and the number chosen at will. Samples can be collected either systematically at fixed stations marked off on the screen, or at random. In the latter case, a random sampling table is used for determining the co-ordinates.

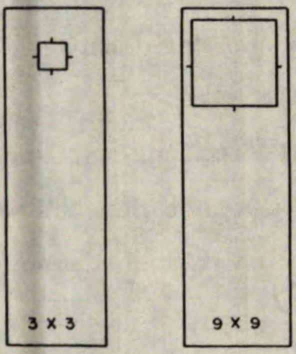
The method of analysis consists essentially of collecting samples of different size from a given population and determining the relationship between sample variance and sample size.

It will be shown (Eq. 13) that the total sampling variance (s^2) consists of a random variance component (s_p^2/w') that depends upon the size (w') of the sample, and a segregation variance component (s_s^2) that is independent of sample size.

The results of experiments carried out with the sampling

Fig. 5.1 - Sampling Board





Samplers

Supply of Lead Pellets

Legend

⌑	- SYSTEMATIC SAMPLING STATION
●	- LEAD PELLET

board are presented in the form of graphs which show the relationship between the variance of single samples and sample size, the latter being determined by the number of screen openings in a square frame. In the tests reported here, three different sample sizes are used:

$w_1' = 1$,

$w_2' = 9$ (located in the square of 3 x 3 openings),

and $w_3' = 81$ (9 x 9 openings).

The numbers of pellets (x) found within the square frames are recorded and the series thus obtained used for calculating variance estimates. A simple formula for calculating this measure of dispersion for a series of observations is presented in Table 5.4, where p can be taken as equal to $\frac{x}{w}$.

5.2.2 Relationship between the degree of segregation and the parent frequency distribution

Example 1

An example of complete segregation will be studied first by placing 2,500 beads in one corner of the sampling board (the lower left corner as shown by the inset on Figure 5.2). This corresponds to a binomial population designated by p = 0.25. Samples collected from this mixture will be either 100% black or 100% white, except for those that straddle the boundary between the black area and the white area. This latter restriction is of little consequence as long as the samples are small compared with the "patch" of 2,500 beads, as shown in Table 5.1, where three series of systematic samples and three series of random samples are presented having sizes 1, 9 and 81 respectively. Figure 5.2 shows that the six variance estimates found from these series do not deviate significantly from a straight horizontal line corresponding to the binomial variance $s^2 = p(1-p) = 0.1875$. The fiducial limits of the variance estimates correspond to variance ratios $F_{95} = 1.52$ (24 and ∞ d.f.) for variance estimates larger than

0.1875, and $F_{95} = 1.73$ (∞ and 24 d.f.) for variance estimates smaller than 0.1875. The results of this sampling experiment show that there is no significant difference between the samples drawn at random and the samples collected systematically. The same conclusion is found when the Chi-square (χ^2) test is applied (see Table 1.2 for references).

The experiments also show that, while the size-variance curve of a completely random mixture is defined by a straight line sloping down at an angle of 45° on a double-log scale, the sample variance never exceeds the theoretical value of 0.1875 in the case of complete segregation and remains substantially constant over the entire interval.

Patterns showing partial segregation may take many forms that are impossible to deal with in every aspect. The gradual transition from complete segregation into complete randomness can, however, be illustrated in an orderly fashion and the conclusions that can be drawn from this apply generally to any pattern of distribution.

To study the characteristics of partial segregation it will be assumed that mixing takes place in five equal steps, reducing the degree of segregation first from 1.0 to 0.8, then to 0.6, to 0.4, to 0.2, and finally to 0. When segregation is zero, the number of pellets within the black square should be 25% of the original number. The total reduction from 100% pellets to 25%, divided into five equal steps, is a reduction of 15% or 375 pellets for each step.

The following mental experiment can now be conducted: 375 pellets are selected at random from the black square containing 2,500 pellets (Figure 5.2), and are redistributed randomly over the remaining three quarters of the sampling board (the degree of segregation is reduced from 1 to 0.8).

A sample drawn from the black quarter of the sampling board

Fig. 5.2 - Size Variance Curve (complete segregation)

P = 0.25

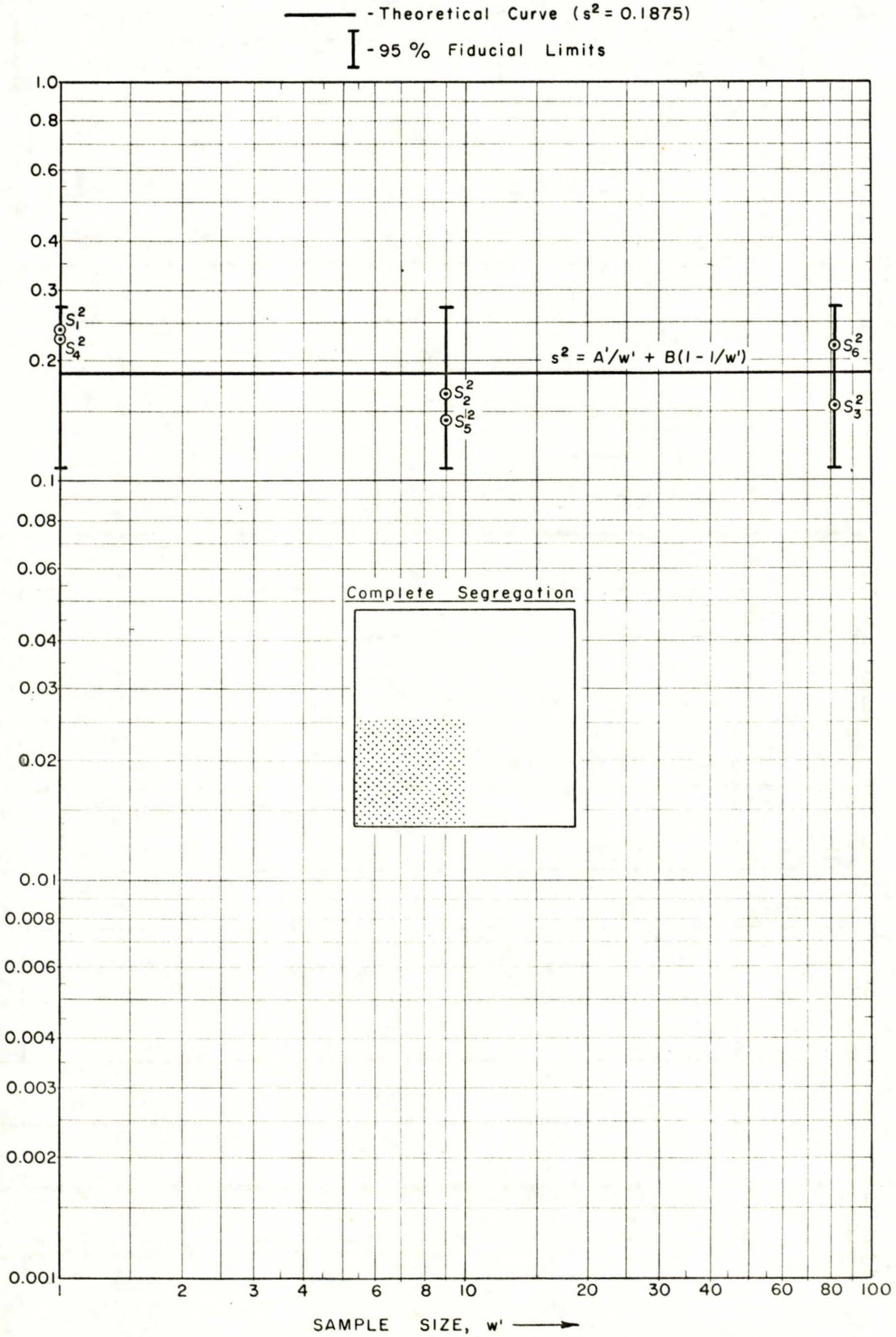


TABLE 5.1

Complete Segregation (Figure 5.2) p = 0.25

	Systematic Samples						Random Samples											
Sample Size	1		9		81		1				9				81			
Sample No.	x_1	x_1^2	x_2	x_2^2	x_3	x_3^2	coordinates		x_4	x_4^2	coordinates		x_5	x_5^2	coordinates		x_6	x_6^2
1							17	07	1	1	68	55			44	04	81	6,561
2							76	74			34	74			22	33	81	6,561
3							37	21	1	1	30	30	9	81	78	46		
4							13	19	1	1	13	77			84	09		
5							04	30	1	1	70	40			26	52	27	729
6							70	97			74	59			71	13		
7							33	77			57	29			91	58		
8							24	46	1	1	25	97			38	18	81	6,561
9							03	44	1	1	65	68			67	24		
10							54	80			76	60			54	76		
11	1	1	6	36	45	2,025	04	94			27	48	9	81	96	96		
12	1	1	6	36	45	2,025	43	77			42	55			57	46		
13	1	1	4	16	25	625	18	24	1	1	37	90			69	92		
14							66	21			86	65			36	42	81	6,561
15							79	90			53	72			10	45	81	6,561
16	1	1	9	81	81	6,561	12	99			00	66			77	10		
17	1	1	9	81	81	6,561	72	27			39	37	9	81	84	45		
18	1	1	6	36	45	2,025	07	72			68	32			57	65		
19							34	95			29	20	9	81	03	04	81	6,561
20							45	14	1	1	61	30			29	26	81	6,561
21	1	1	9	81	81	6,561	52	38			29	68			53	34	18	324
22	1	1	9	81	81	6,561	85	68			94	49			75	23		
23	1	1	6	36	45	2,025	66	88			98	69			91	20		
24							60	11			94	10			93	57		
25							44	80			24	82			30	27	81	6,561
Sum	9	9	64	484	529	34,969			8	8			36	324			693	53,541
s^2	0.2400		0.1647		0.1510				0.2267				0.1400				0.2180	

will have an expected value:

$$E(X)_{black} = (2500 - 375)/2500 = 0.85$$

Similarly, for samples drawn from the other three-quarters, we find expected sample values

$$E(X)_{white} = 375/7500 = 0.05.$$

The expected variance calculated from these figures for a degree of segregation 0.8 is,

$$E\ (\text{variance}) = E\left[[X - E(X)]^2\right] = 0.1200.$$

The total variance for a degree of segregation of 0.8 is 0.64 times the total variance for the entirely segregated mixture.

Continuing the experiment for lower degrees of segregation, the results presented in Table 5.2 are found when four samples are collected (one from each quarter of the sampling board) for each test.

TABLE 5.2

Effect of Segregation on Total Variance

Degree of Segregation (z)	Deviation from Mean Grade $X - E(X) = X - 0.25$ for Each Quarter	Total Expected Variance	
		$E(s^2)$	Fractional
1.0	0.75; 0.25; 0.25; 0.25	0.1875	1.00
0.8	0.60; 0.20; 0.20; 0.20	0.1200	0.64
0.6	0.45; 0.15; 0.15; 0.15	0.0675	0.36
0.4	0.30; 0.10; 0.10; 0.10	0.0300	0.16
0.2	0.15; 0.05; 0.05; 0.05	0.0075	0.04
0.0	0.00; 0.00; 0.00; 0.00	0.0000	0.00

This table shows that the degree of segregation (z) and the expected variance are related:

$$E\ (\text{variance}) = 0.1875\ z^2$$

A similar relationship holds for all ratios of "black" and "white" mixtures other than for 2,500 out of 10,000.

Practically speaking, the expected variance is the limit of the total variance as sample size increases. The expected variance is therefore identical to the segregation variance:

$$\text{E (variance)} = s_s^2.$$

Furthermore, the total variance for complete segregation appears to be identical with the parent variance, i.e. the variance of single items, which in this case follows from the binomial equation $s_p^2 = p(1-p)$.

From the foregoing equations it is seen that:

$$s_s = z \cdot s_p \qquad \text{..... (Eq. 11)}$$

Summarizing the conclusions from the above experiment:

1. The segregation variance has a maximum value equal to that of the parent variance of the population.
2. The segregation variance is, within the range of actual sampling practice, substantially independent of sample size. It never exceeds the parent variance.
3. The ratio between the segregation variance and the parent variance depends solely on the degree of segregation (z).
4. The total variance of samples consisting of a single unit equals the parent variance (s_p^2) regardless of the degree of segregation.

On the basis of experimental evidence, it is proposed that the expected variance of sampling satisfies the following relationship:

$$E(s^2) = s_p^2/w' + E(s_s^2)\ (1 - 1/w') \qquad \text{.... (Eq. 12)}$$

where s_p^2 = parent variance; variance of single units;

$E\ (s_s^2)$ = expected value of the segregation variance; and

w' = sample size, expressed in number of units.

This equation formulates the general relationship $s = f\ (s_p,\ s_s,\ w')$ for any degree of segregation (z = 0 to 1) and sample size $w' \geqq 1$ (compare Figures 5.2, 5.3, 5.4 and 5.5 as illustrations).

For samples consisting of two units the total variance becomes, by first approximation,

$$s^2 = 1/2\ s_p^2 + 1/2\ s_s^2.$$

For samples consisting of ten or more units, Equation 12 can be written by first approximation as:

$$s^2 = s_p^2/w' + s_s^2 \qquad \ldots\ldots\ (\text{Eq. } 13)$$

It is noted that the parent variance (s_p^2) is a constant which, according to the binomial equation, depends only upon the composition of the material. It is designated as "sampling constant A'".

The segregation variance (s_s^2) for one and the same material depends upon the degree of segregation (z) only, in accordance with Equation 11. It is known from experience that, while (z) may range from zero to 1, the stability of the segregation variance under otherwise normal conditions of handling, storage, and transportation is comparable to that of the parent variance. To illustrate with figures, it is known that noticeable blending occurs when a mixing device reduces the segregation variance of a product by a factor of 3 or more. Conversely, an increase of the segregation variance by a factor of 3 to 4 or more is equivalent to a distinct separating action. Therefore, while s_s^2 may change, its value for a given material consignment will be constant within limits normal for variance estimates (F-ratio), unless the consignment is noticeably mixed or segregated. Segregation variance s_s^2 is designated as "sampling constant B".

The practical value of the "sampling constants" is demonstrated by examples 2 and 3.

Example 2

General Equation 12 was tested by distributing 2,500 lead pellets non-randomly over the sampling board. The samples of different sizes were collected systematically and at random as in the first example. The results are presented in Figure 5.3 and Table 5.3.

Two variance estimates, s_1^2 and s_3^2 , obtained from the systematic samples were used to evaluate the sampling constants from Equation 12, which can now be written as:

$$s^2 = A'/w' + B(1 - 1/w') \quad \text{....... (Eq. 14)}$$

The two constants were found by substituting the observed values for s_1^2, s_3^2, w_1', and w_3' in the following equations:

$$s_1^2 = A'/w_1' + B(1 - 1/w_1') \quad \text{....... (Eq. 14a)}$$

$$s_3^2 = A'/w_3' + B(1 - 1/w_3') \quad \text{....... (Eq. 14b)}$$

It follows that A' = 0.1824 and B = 0.00761.

From these values the size-variance curve for Equation 14 was found; it is shown in Figure 5.3. This size-variance curve is approximately the algebraic sum of a straight line A'/w', sloping down at 45 degrees from a point $w' = 1$; $s^2 = 0.1824$, and a straight horizontal line, B = 0.00761. The former represents the random variance component and the latter, the segregation variance component. The degree of segregation is found from the equivalent of Equation 11.

$$z = \sqrt{B/A'} = 0.20$$

Here, the Chi-square test provides spot-checks for the "goodness-of-fit" of Equation 14, using the experimental variance estimates s_2^2, s_4^2, s_5^2 and s_6^2.

The size-variance curve calculated from s_1^2 and s_3^2 falls within the confidence interval defined by $(n-1)\cdot s_i^2/\chi^2$ of each of the above four variance estimates for probability levels P = 0.025 and 0.975. For example, the confidence interval of the variance estimate s_2^2, which was found from 25 (systematic) samples, is 0.026 — 0.088 at the 95% level. The calculated variance (Equation 14) for A = 0.1875; B = 0.00761; w' = 9, falls within this range at 0.028. Since s_2^2 shows the largest difference of all, the Chi-square test confirms the statistical identity of the calculated variance (Equation 14) and all four experimental variance estimates at the 95% level. In Figure 5.3 the confidence interval is shown only for the experimental variance s_2^2. It is noted that similar results were found when applying the F-test. The Chi-square test was preferred, however, it being the more rigorous of the two tests. Frequency distributions for samples larger than 1 unit (w' = 1) will generally show deviations from the binomial distribution when the material is segregated. When the samples contain only a small number of units, as they necessarily do in the experiments performed with the sampling board, these departures from the theoretical binomial frequency distribution cannot always be proven significant. When, however, the number of units contained in the sample becomes very large, such as in molecular binomial mixtures (fluids, pulps, etc.), the difference between the frequency curve of sample values as found from a test and the frequency curve of the sample values observed in the same material consignment when it is randomly mixed, will generally be significant, the more so when the degree of segregation is high.

Fig. 5.3 - Size Variance Curve (partial segregation)

P = 0.25

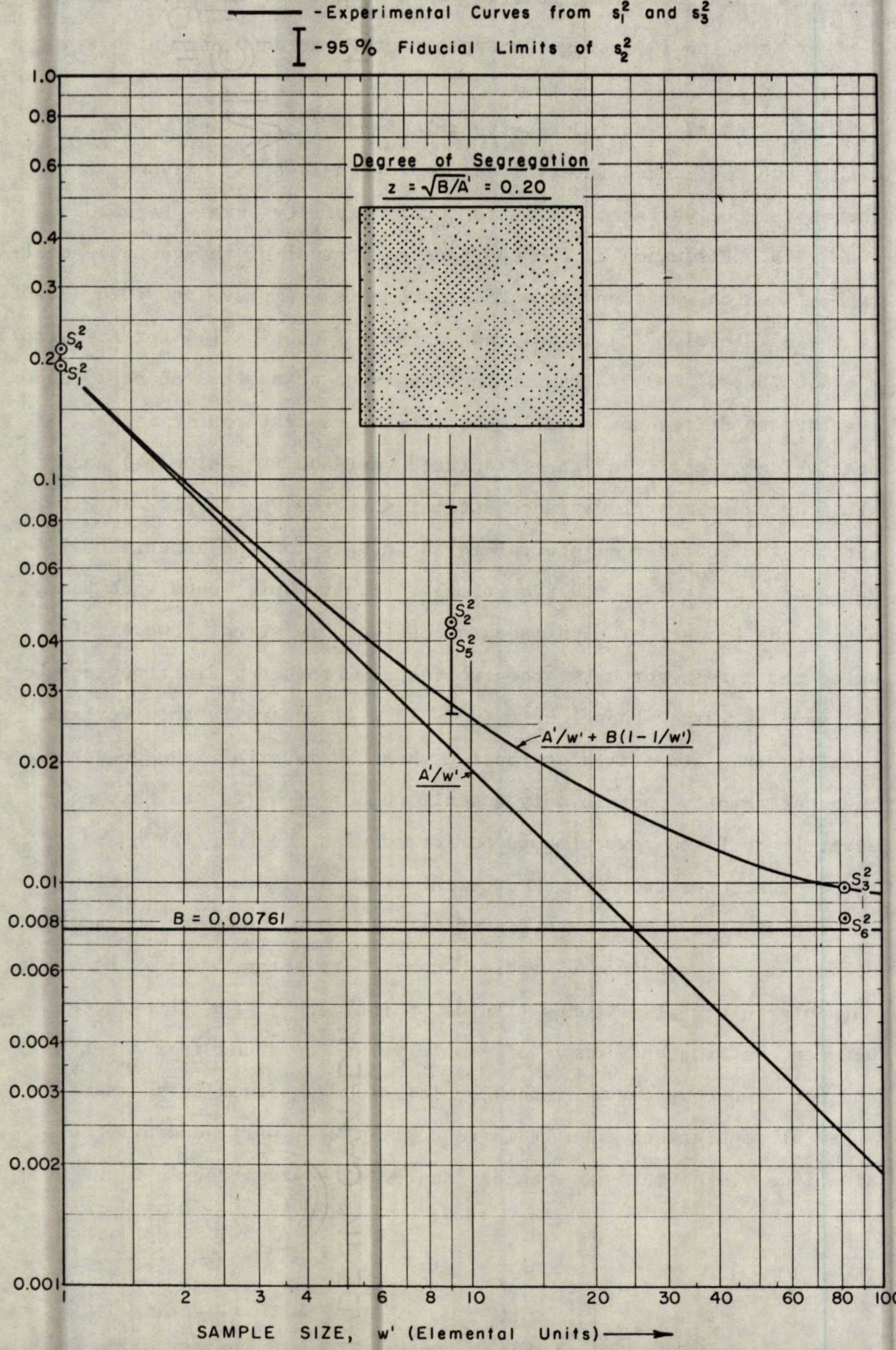


TABLE 5.3

Partial Segregation (Figure 5.3): p = 0.25

	Systematic Samples						Random Samples											
Sample Size	1		9		81		1				9				81			
Sample No.	x_1	x_1^2	x_2	x_2^2	x_3	x_3^2	coordinates		x_4	x_4^2	coordinates		x_5	x_5^2	coordinates		x_6	x_6^2
1			2	4	21	441	03	36			46	33			60	16	17	289
2			2	4	20	400	47	96	1	1	98	26			11	22	18	324
3			2	4	14	196	43	47	1	1	63	16	2	4	14	77	20	400
4			2	4	16	256	73	36	1	1	71	80	1	1	10	94	7	49
5			2	4	15	225	86	61			62	45	4	16	95	39	28	784
6			1	1	15	225	97	42			42	27	6	36	24	84	26	676
7			1	1	30	900	74	81			53	07	1	1	51	42	23	529
8			3	9	35	1,225	24	14			32	36	1	1	79	17	35	1,225
9					25	625	67	57	1	1	37	07	5	25	89	53	20	400
10			4	16	19	361	62	20			32	51	2	4	73	31	20	400
11			1	1	19	361	16	56			32	13			88	63	20	400
12			2	4	21	441	76	50	1	1	90	55			97	01	6	36
13	1	1	1	1	20	400	62	26			79	38	1	1	54	63	27	729
14			1	1	22	484	27	71	1	1	78	58	2	4	14	78	20	400
15	1	1	5	25	28	784	66	07			53	59	4	16	10	59	8	64
16	1	1	4	16	26	676	12	96			05	57			88	33	16	256
17	1	1	6	36	27	729	56	96			03	12			26	21	22	484
18			1	1	21	441	85	68			72	10	5	25	49	12	23	529
19	1	1	3	9	22	484	99	27			93	14	4	16	81	34	11	121
20					15	225	26	31			15	21	3	9	76	29	29	841
21			1	1	11	121	55	38			31	06	1	1	23	57	21	441
22			1	1	20	400	59	54			62	18	2	4	83	60	21	441
23			2	4	19	361	56	82			43	44	1	1	01	86	6	36
24	1	1	2	4	38	1,444	35	46	1	1	09	32	4	16	30	32	22	484
25			8	64	46	2,116	64	22			90	53	3	9	30	44	17	289
Sum	6	6	57	215	565	14,321			7	7			52	190			483	10,627
s^2	0.1980		0.0438		0.00986				0.2100				0.04210				0.00823	

In fact, although the frequency distribution of large samples from segregated mixtures can take on any shape, independently of the shape of the parent distribution, the variance of such large samples is directly related to the variance of the frequency distribution of the single units. This relationship is demonstrated for variates that can be expressed by parameters having a binomial parent distribution. It will be shown further on that the same concept applies to parent distributions of different type, including normal, poissonian, and irregular parent distributions (see under "Non-binomial Variates", 5.4.2).

Example 3

A test similar to the preceding ones was done, using 1,000 lead pellets distributed as evenly as possible over the sampling board. The curve for Equation 14 was based on variance estimates s_1^2 and s_3^2 (see Figure 5.4). All the other values which were determined independently appear to check within the limits of chance variation with the curve

$$s^2 = 0.1086/w' + 0.00137\ (1 - 1/w').$$

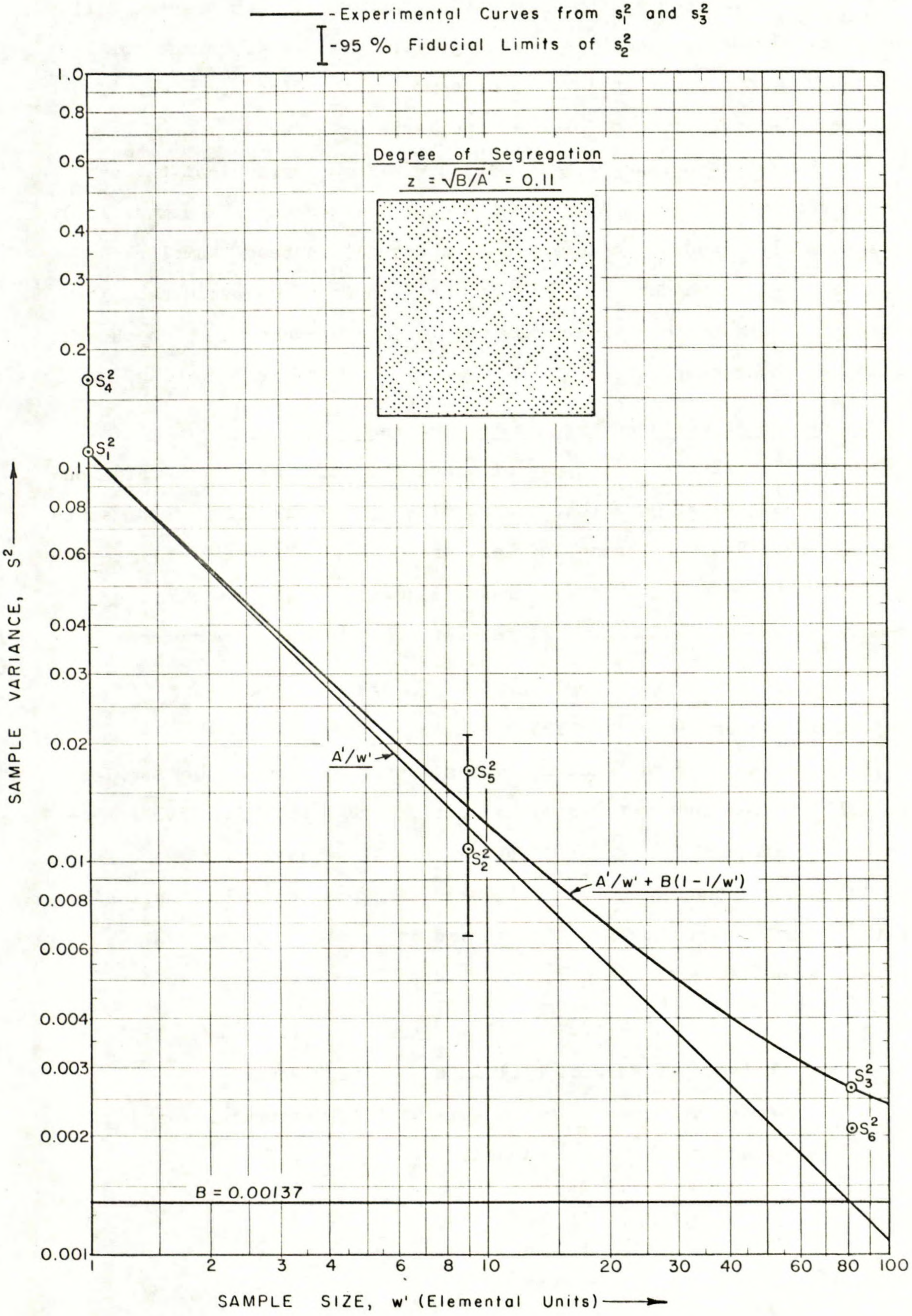
The degree of segregation found from $z = \sqrt{B/A'} = 0.11$.

These three examples confirm the correctness of the general Equation 14 for a range of conditions varying between complete segregation and near-random dispersion of the variate.

In practice, the use of samples which consist of only a few units is common in such fields as the microscopic analysis of particle mixtures and in sampling for defectives. In many instances, however, the samples collected consist of a very large number of units that cannot be counted. In these cases,sample size is expressed in some unit of measurement (1 gram, 1 pound, etc.); each unit of measurement may contain thousands or millions of elementary units of the

Fig. 5.4 - Size Variance Curve (minor segregation)

P = 0.10



binomial. As a result, the size-variance curve of such samples will be generally determined by the segregation variance component only. In other words, the actual range of sample sizes lies somewhere within the less steep section of the size-variance curve.

For this type of material it would be impractical to use the parent variance for sampling constant A', because the number (w') of binomial units is too large to be counted. Instead, sampling constant A' can be determined for a single unit of measurement. It is then necessary to indicate the unit of measurement to which this sampling constant refers.

5.2.3 Practical units and proximate equation

To illustrate the use of practical units and their relationship to the general equation, the results of another test are presented in Figure 5.5. One thousand lead pellets were distributed with a high degree of segregation (see inset Figure 5.5) and the sampling constants calculated from variance estimates s_1^2 and s_3^2 as before:

$$s^2 = 0.09923/w' + 0.01078\ (1 - 1/w')$$

degree of segregation, z = 0.33.

The other variance estimates (obtained from random samples as well as from systematic samples) correspond with this curve as before within the 95% fiducial limits. It will be assumed for the sake of convenience that the size of samples is expressed in a practical unit of measurement equal to ten elementary units. The general equation now becomes:

$$s^2 = A/w + B\ (1 - 1/10w) \quad \ldots\ldots\ (\text{Eq. } 15)$$

where A = variance of samples of 1 unit of measurement,

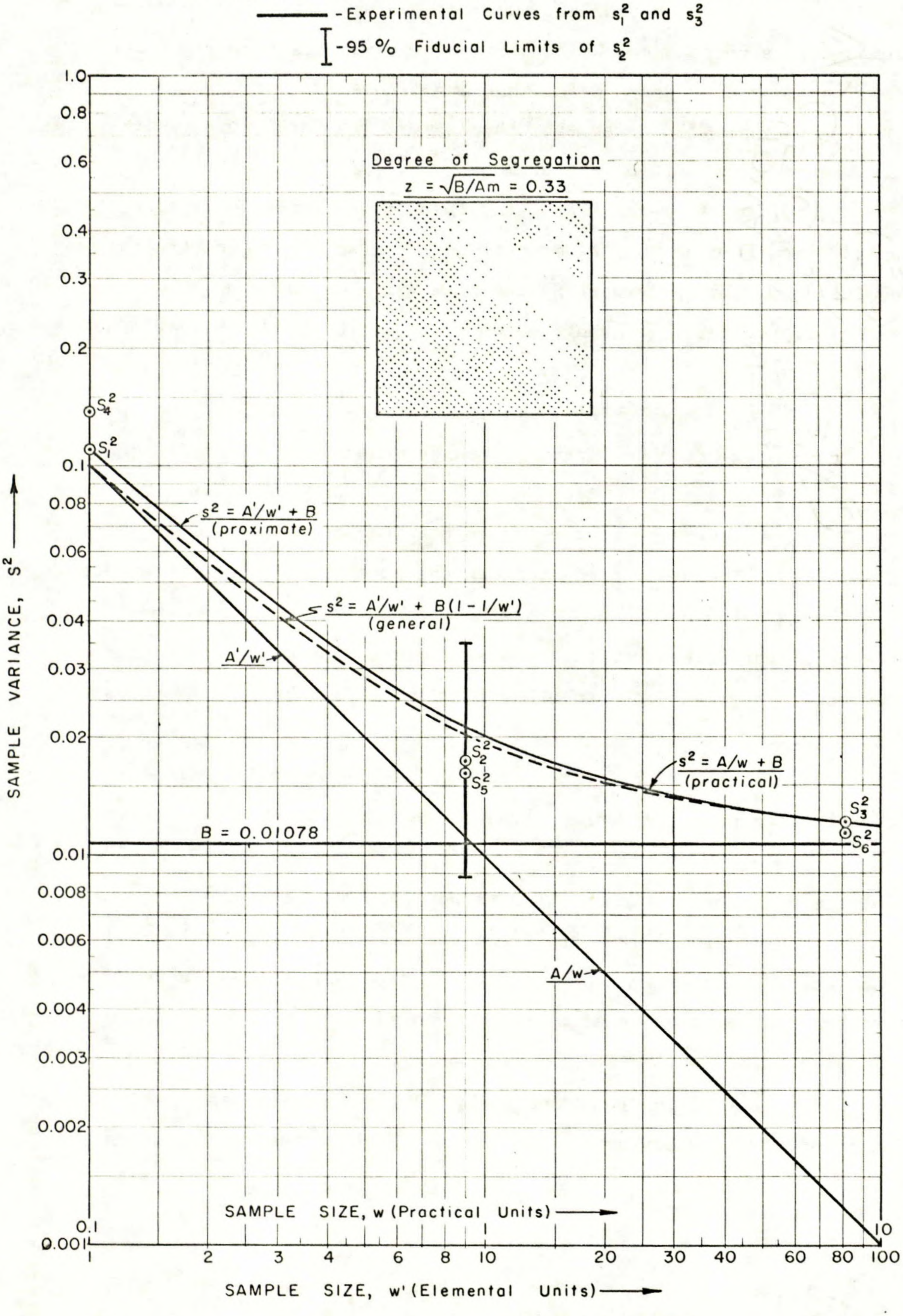
w = sample size expressed in same unit of measurement, and

A/w = random variance component.

Fig. 5.5 - Size Variance Curve (Practical Units)

P = 0.10

(This diagram illustrates the use of Elemental and Practical Units)



It is noted that the numerical value of the random variance component is not changed by this transformation, as shown in Figure 5.5, the only difference being that A = 1/10 A'.

It is also noted that the segregation variance B is independent of the unit of measurement.

In those cases where samples must be expressed in some unit of measurement that is many times the size of an elemental binomial unit, the upper part of the size variance curve as shown in Figure 5.5 is not used, and the general Equation 14 can be replaced by:

$$s^2 = A'/w' + B,$$

or, when using practical units of measurement,

$$s^2 = A/w + B \qquad \text{....... (Eq. 16)}$$

The curve corresponding to this equation is also shown in Figure 5.5. The discrepancy between the general curve and the practical curve turns out to be negligible for a first approximation of the total variance estimate. The same conclusion holds for higher degrees of segregation. Equation 16 will be used henceforth unless otherwise indicated.

The equation for the degree of segregation (z) likewise changes when practical units of measurement are used:

$$z = \sqrt{B/Am} \qquad \text{....... (Eq. 17)}$$

where m = number of elemental units per unit of measurement.

Equation 17 appears to be useful because (z) can often be estimated from available data on the average composition and distribution of a material consignment. Examples 4 and 5 (Section 5.3) illustrate the application of Equation 17.

It is noted that the product Am is dimensionless and can be estimated from any other unit for which the value of A is known.

In view of the above tests, it can be concluded that the variance of single samples drawn systematically or at random from segregated materials consignments can be expressed as a function of two constants which are determined by the composition of the material, by the degree of segregation of the consignment, and by the size of the sample.

When single samples are combined, as is done in incremental sampling, the total variance of a gross sample consisting of N increments has a maximum value equal to 1/N times the total variance of the single samples. Theoretically, this maximum value will be attained only when the "patches" caused by segregation of the consignment are themselves distributed at random. This condition may not prevail in actual practice, and the total variance as formulated for gross samples consisting of N increments,

$$s^2 = A/Nw + B/N,$$

is in fact an estimate of the upper limit of the gross sample variance. The estimate of the total variance obtained from this equation is therefore a safe estimate; the same equation can be written as follows:

$$s^2 = A/W + B/N \qquad \ldots\ldots\ (\text{Eq. } 18)$$

where W = Nw = the gross sample size.

This equation is the general expression of variability for gross samples drawn from material consignments that are not perfect mixtures.

5.3 Materials of Unknown Composition

Sampling constants (A and B) and the degree of segregation (z) for materials of unknown composition can be determined by means of the duplicate sampling method, using small and large samples. This test requires the collection of two series of single samples, from which an estimate of the total variance (s^2) is found. For the first series, relatively small samples (w_1) are chosen to ensure that the first term (A/w) in Equation 16 contributes more to the total variance than the second term. The estimate (s_1^2) therefore largely reflects the random sampling component (A/w). The second series of samples are of relatively large size (w_2) and, as a result, the variance found from this series is caused mainly by the segregation component B. The following equations derived from Equation 16 provide maximum estimates, by first-order approximation, of sampling constants A and B.

$$A = w_1 \cdot w_2 \cdot (s_1^2 - s_2^2)/(w_2 - w_1) \qquad \text{.......(Eq. 19)}$$

$$B = s_2^2 - A/w_2 \qquad \text{....... (Eq. 20)}$$

The error of reduction and analysis of individual samples has been ignored in these equations; the inflation caused in the estimates of A and B is generally of no consequence. The sample sizes (w_1, w_2) should generally be the smallest and largest sizes practically possible.

The degree of segregation (z) is expressed by Equation 17. In many materials that are mass-produced, the degree of segregation (z) does not change too much although the pattern of distribution may vary, and it is possible to estimate B without a test when A and (z) are known.

A condensed schedule of the calculations required for determining sampling constants A and B and the degree of segregation (z) is presented in Table 5.4.

TABLE 5.4

Calculation of (A,B) and (z) for Materials of Unknown Composition

Sample No.	Small Samples		Large Samples		Calculations
1	p_1	p_1^2	p_2	p_2^2	Determine the variance for each series, (s_1^2) and (s_2^2), with the equation: $s^2 = \dfrac{\text{sum } p^2 - (\text{sum } p)^2/n}{n - 1}$ Determine (A,B) from Equations 19 and 20. Find (z) from Equation 17.
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
n (see note)	.	.	.	.	
	sum p_1	sum p_1^2	sum p_2	sum p_2^2	NOTE: It is recommended to collect a minimum of 25 to 30 samples for each series.
Average size of samples	w_1		w_2		

Example 4

An untreated stove coal (1-1/2" x 2-3/8") was sampled by collecting 35 increments with an average weight of 185 grams each, and a second series of 35 increments with an average weight of 6,539 grams each. These samples were analyzed for ash content. The variance for the small samples (calculated from fractional ash content) was $s_1^2 = 0.0234$; the variance for the large samples was $s_2^2 = 0.00219$. Sampling constants found from Equations 19 and 20 are:

$$A = 4.04 \text{ for samples of 1 gram}$$

$$B = 0.00157$$

The weight of the gross sample and the number of increments can be found for any pre-assigned accuracy from Equation 18:

$$s^2 = 4.04/W + 0.00157/N$$

For instance, a sampling precision of 1% ash (corresponding with $s = \frac{1}{1.96}\,\% \simeq 0.005$) would be obtained 19 times out of 20 (see Section 1.6) when collecting 128 increments with a total weight of 320 kilograms. The average particle weight of the coal was found from a sieve analysis to be 29.6 grams. Consequently, the number of particles per gram of sample was $m = 1/29.6$, and the degree of segregation, as calculated from Equation 17, was found to be $z = 0.11$.

Example 5

The results of a general election were used in the following duplicate sampling test: the variance s_1^2 of the individual political support for a certain party (X) was compared with the variance of the average political vote for the same party in the ridings. The average number of votes per riding was $w_2 = 15{,}430$, while $w_1 = 1$. The variance s_1^2 was found to be 0.27; variance s_2^2 appeared to be 0.0045. The resulting variance formula is:

$$s^2 = 0.27/W + 0.0045/N$$

The number of investigators required for probing the political opinion of the same population at some future date and the number of interviews to be made by each investigator can be estimated in advance by using this equation. For instance, public opinion regarding the same party (X) could be determined to the nearest 1.5% by about 320 pollsters who would each interview 20 persons. The degree of segregation (z) for this population, with regard to its political support for party (X), follows from Equation 17: for $m = 1$, $z = 0.13$.

The following example demonstrates the application of Equations 16, 17 and 18 for materials that are characterized by a variate (X) but that do not consist of mixtures of identical units.

Example 6

Mixtures of particles of unequal size that are sampled for size analysis can be regarded as binomial mixtures by defining variate (X) as a particle size interval within two given size limits. The material consignment can then be regarded as consisting of two fractions, (X) and (non-X), as before. The precision of the weight percentage of particles (X) found from a sample is determined by Equations 16 and 17. Estimates of the sampling constants A and B can be found from a duplicate sampling test, as demonstrated previously, by collecting two series of samples, one series consisting of relatively small samples and the second series of relatively large samples.

The substance to be sampled might occur in the form of broken aggregate, solids in suspension, or droplets in an emulsion. When a material occurring in one of these forms is sampled, the chance error as expressed by the binomial variance is now caused by the accidental interchange of units of differing size and depends therefore on the size and relative abundance of the units.

When the particles are small and the number of particles per unit of weight is large, the value of the sampling constant A for samples of unit weight will generally be small compared to that of sampling constant B. Since the effect of segregation prevails over random variation, the frequency distribution of (X) will generally show an irregular form, depending upon the pattern of segregation and the number of particles contained in each sample used for the determination of (X).

Solid Aggregates

When the material consignment consists of a solid aggregate, random errors caused by the accidental interchange of units (X) and (non-X) are automatically precluded because no movement of these units relative to one another is possible. While this does not exclude all random variations, most of the variations are caused by segregation when the elemental units that are the carriers of the variate are very small in comparison with the sample.

In materials of this type, the variability of (X) is often of the binomial kind, as, for instance, when sampling ore in place for its metal content. The ore consists of a mixture of molecular units (X) and other constituents (non-X). All variability originates from this binomial mixture, but substantially in the form of segregation. The sampling constant B for molecular units can be calculated with the binomial equation or measured directly.

The practical value of the binomial theory lies in its application to materials of known composition and distribution, as will be demonstrated in the next section.

5.4 Materials of Known Composition and Distribution

When the main characteristics and distribution of a material consignment are known, its sampling constants can often be determined without a test. Sampling precision as expressed by the total variance of sampling can be determined from Equations 16, 17 and 18 for binomial variates when the average value of the variate and the degree of segregation (z) of the consignment are known.

5.4.1 Binomial variates

Sampling constant A is calculated from the binomial equation, which takes on different forms, depending upon the type of material and the variate. Sampling constant B is calculated from A, the degree of segration (z), and the ratio (m) which denotes the number of units of the material contained in the unit of measurement used for expressing variate (X).

The "materials" are subdivided into three main classes (see Table 5.5). The first class deals with materials consisting of discrete units, each one of which bears a characteristic quality (X) or (non-X). Variability in the values of samples drawn from a consignment of such a material results from the fact that these elementary units can move relative to one another; they can be either randomly mixed or can cause a certain degree of segregation in the consignment. It is generally easy to separate units (X) from units (non-X) in these substances by physical or chemical methods. Most gases, fluids, and mixtures of these with solids (amalgams, suspensions, pastes) belong to this class. Applications of the method can be found in the fields of microchemistry and assaying. Likewise, the sampling of mass-produced items and similar "discrete populations" also belongs in this first class.

The second class of substances includes materials in which variability is caused, as before, by the free movement of elemental units. In this case, however, the variate (X) is not restricted to certain units, but is spread in varying degrees over all the elemental units. Granular solids such as broken coal and ore, wheat and many other materials fall into this class. The units can be separated into two fractions characterized by "high-X" and low-X"; the variability caused by relative movement of the units of these two fractions is reflected in the variations of the sample drawn from such material.

A third class of materials is recognized where variability is caused by uneven dispersion of the variate (X) throughout the consignment. Essentially, these materials differ from the ones above only in that the elemental units (X) and (non-X), which may be real or imaginary, cannot move relative to one another; this reduces random variation. Many physical properties, such as the tensile strength

Table 5.5 - CALCULATION OF SAMPLING CONSTANTS FOR MATERIALS OF KNOWN COMPOSITION AND DISTRIBUTION

BINOMIAL VARIATES ONLY

Class of Material	I			II	III
	Material consisting of separate items characterized by (X) and (non-X) in gaseous, liquid, or solid form, or in mixtures of some (suspensions, emulsions, pulps, or pastes). Items (X) can be separated from items (non-X) by physical or chemical methods.			Material consisting of separate aggregates of (X) and (non-X). The aggregates are characterized by "high-X" and "low-X" and are separable.	Other materials: 1. Variate (X) is dispersed without being accumulated in separate physical units. 2. (X) occurs in units that cannot be identified or separated.
	Items are countable	The number of items in the sample is too large to be counted			
Material-Group No.	(1)	(2)	(3)	(4)	(5)
Method of evaluating average grade of consignment	The average grade is determined by counting the number of items (X) and (non-X) in the sample, either directly or after separating items (X) from (non-X).	The average grade is determined by separating the sample by suitable physical and/or chemical methods into two fractions, (X) and (non-X). Fractions are measured by a parameter, expressed in a suitable unit of measurement.		The average grade is determined directly, by suitable chemical and/or physical analytical methods.	
		Items (X) have same specific gravity as items (non-X).	Items (X) differ significantly in specific gravity from items (non-X).	Units may have different size and/or specific gravity.	Standard specimen of the material may be required for specific tests.
Parameter used for measuring average grade	Variate (X)	A dimension of the items: length (width, height, depth, diameter, thickness, etc.); surface area; volume.	Weight of fractions (X) and (non-X)	Weight of fractions "high-X" and "low-X"	A length (diameter, depth, exponsion, etc.); time; load (force), or other parameters used in the test.
Unit of measurement	Number	Unit of weight, volume, length, area; surface area per unit of weight, etc.	A unit of weight	A unit of weight	A unit of weight, force, time, length, surface area, suitable for measuring the parameter.
Examples	1. Sampling for public opinion. 2. Proportion of defectives (X) in the manufacturing of mass-produced goods.	1. Size analyses. 2. The fineness of hydraulic cement, by surface area (turbidimeter). 3. Sampling of textiles for wool content.	1. Lightweight pieces in aggregate. 2. Float-sink analysis of coal.	1. Ash content (X) of a consignment of broken coal. 2. Sampling of sands for heavy minerals.	1. Sampling of ores in place. 2. The abrasion of crushed gravel by weight loss. 3. Ductility of bitumen by elongation.
Sampling constants	$\boxed{A = p(1-p)}$ A = random unit variance. p = average fractional number of items (X) known by approximation. $\boxed{B = Az^2}$ B = segregation variance. z = degree of segregation (known).	$\boxed{A = p(1-p)/m}$ A = as in (1). p = average proportional amount of (X) fraction. m = average number of items per unit of measurement. $\boxed{B = Amz^2}$ B = as in (1). z = as in (1).	$\boxed{A = p(1-p)d/Dm}$ A = as in (1). p = as in (2), fractional weight. d = specific gravity of items (X) or (non-X). D = average specific gravity of material. m = as in (2). $\boxed{B = Amz^2}$ B = as in (1). z = as in (1).	$\boxed{A = p(1-p)(a_1-a_2)^2 d_1 d_2/D^2 m}$ A = as in (1). p = as in (2), fractional weight. $a_{1,2}$ = X-values of fractions (1,2). $d_{1,2}$ = specific gravity of fractions (1,2). D = specific gravity of material. m = as in (2). $\boxed{B = Amz^2}$ B = as in (1). z = as in (1).	1. (X) is separable chemically. $\boxed{B = p(1-p)dz^2/D}$ B = as in (1). p = average proportional amount of chemical constituent. d = as in (3). z = as in (1). D = as in (3). 2. (X) is not separable chemically. $\boxed{B = s^2}$ B = as in (1). s = standard deviation of (X) from available data.

of a wax or the abradability of a gravel, are in this category. Distribution of such a variate over the consignment can be attributed to segregation of elementary units characterized by either (X) or (non-X) which cannot be separated and often cannot even be identified.

All three classes are seen as binomial populations; samples collected from material consignments belonging to the third class have a variance that is substantially determined by segregation.

Five groups of materials are recognized under the main classifications; these will now be described in more detail, (see Table 5.5).

Group No. 1 deals with substances that occur in the form of separate units, each characterized by either (X) or (non-X). A feature of this group of materials is that the samples are analyzed by counting the individual units (X) and (non-X).

Groups No. 2 and No. 3 include materials consisting of separate units which are too numerous to be counted individually and which are consequently measured by some dimension of the items (length, surface area, volume or weight, etc.) expressed in a suitable unit of measurement (inch, square foot, gallon, pound, etc.).

Group No. 2 includes materials where the items characterized by variate (X) have the same specific gravity as items (non-X); for example, granular materials sampled for size analysis.

Group No. 3 deals with materials consisting of items (X) that differ significantly in specific gravity from items (non-X). These are materials that are sampled for specific gravity analysis (e.g. by float-sink analysis).

Groups No. 4 and No. 5 include materials in which the variate (X) is dispersed without being necessarily accumulated in separate physical units of the material.

Group No. 4 includes all materials consisting of separate aggregates that are characterized by either a high percentage of variate (X) or a low percentage of variate (X), the two components being separable.

Group No. 5 includes other materials where the variate (X) is either dispersed without being accumulated in separate physical units or occurs in units that cannot be identified or separated.

The examples that follow may serve to illustrate the use of Table 5.5:

Group 1

Example 7

A mass-produced item is known to contain about 4% defectives. Therefore, p = 0.04 and sampling constant A = 0.0384. It follows from Equation 14 that the effect of any segregation can be eliminated by collecting and testing sample items one by one (w' = 1). The number (N) of items required for determining the percentage of defectives to the nearest 1% nineteen times out of twenty is found from

$$N = A/s^2_{av}$$

where $s^2_{av} = (0.01/1.96)^2 = 26 \times 10^{-6}$ represents the variance of the average expressed as a fraction of N (see Section 1.6). Consequently, N = 1476.

Example 8

The results of a general election are used to determine the number of investigators to be employed in a poll to survey changes in political popularity, and the number of persons to be interviewed by each investigator. The party whose election returns were closest to 50% was party X, its vote amounting to 61% of the total returns; this figure is subject to the greatest variations and is used as a yardstick for evaluating sampling precision of the poll. Consequently p = 0.61

and the sampling constant A = 0.24. The degree of segregation for X is known to be z = 0.13. It follows that the sampling constant $B = Az^2 = 0.0041$. From the many possible combinations of (w) and (N), a value w = 16 is chosen as a reasonable figure for the number of persons that can be interviewed by one investigator in one day.

It is found from Equation 18 that by employing 155 investigators, the results of the poll will indicate political popularity with a precision of 2%, nineteen out of twenty times, $s^2 = \left(\frac{0.02}{1.96}\right)^2 = \left(\frac{0.24}{16} + 0.0041\right)/N$. The total number of persons interviewed would thus be: wN = 2480.

Group 2

Example 9

It is required, for operational control in an ore beneficiation plant, that a daily sample of minus 14 mesh sand be collected for sieve analysis. The precision of the sieve curve is important, especially with regard to the silt fraction which needs to be determined with a precision of 1% nineteen out of twenty times. The sand is segregated (z = 0.20) and the average amount of silt (minus 200 mesh material) is 3%.

The accidental interchange of silt particles with sand particles during sampling is determined by the size of the particle. Errors thus caused depend primarily upon the size and relative abundance of the coarse particles; i.e., the sand fraction. The weighted average particle weight of the sand fraction (14 x 200 mesh) of this ore is known to be 0.010 gram. Therefore, m = 100 when the sample weight is expressed in grams. It follows that:

$$A = p(1-p)/m = 0.0003$$

$$B = Amz^2 = 0.0012.$$

Samples in this plant are collected automatically by increments weighing 30 grams each. The minimum required number of increments found from Equation 18:

$$N = 47.$$

Group 3

Example 10

A non-uniform lightweight aggregate is tested by float-sink analysis for determining the percentage of lightweight pieces. The material is known to contain approximately 10% by weight of lightweight pieces floating on bromotrichloromethane (sp. gr. = 2.00); the average specific gravity of the floats is d = 1.6; the average specific gravity of the entire aggregate is D = 2.3, and the degree of segregation is known to be z = 0.3. The size of the lightweight aggregate is minus 1-1/2 inch; the weighted average particle weight is 15 grams; and hence, m = 1/15 = 0.067. The sampling constants A = 0.934 and B = 0.0056 are found from the equations given in Table 5.5 under Group No. 3. Increments are collected by an automatic sample cutter, each cut weighing approximately 400 grams. The minimum number of increments required to attain a sample precision of 1% is found from Equation 18:

$$N = 303.$$

The weighted average particle weight can be determined from a sieve analysis, using the following equation:

$$V = \Sigma k^3 q / \Sigma q,$$

where V = weighted average particle volume, cu. cm.,

q = weight of individual size fraction, and

k = central value of individual size fraction, cm.

Group 4

Example 11

A minus 1-1/2 inch mine-run slack coal having an average ash content of about 30% is sampled for ash by an automatic sampler collecting increments of 5 lb. This coal is known to contain approximately 64% (p = 0.64) floats at 1.60 sp. gr. with 5% ash (a_2 = 0.05), and 36% sinks with approximately 80% ash (a_1 = 0.80). The specific gravities of these two fractions are known to be d_2 = 1.30; d_1 = 2.35; the overall specific gravity D = 1.60.

The weighted average particle weight of this coal is 5.26 grams. Because the weight of sample is expressed in pounds (1 lb. = 454 grams), the ratio m = 454/5.26 = 86. The degree of segregation of the mine-run slack is known to be z = 0.13. From this the sampling constants (see Table 5.5, Group 4) are:

$$A = 0.00180$$

$$B = 0.002616$$

The minimum number of increments required to determine the ash content with a precision of 1% ash, nineteen out of twenty times, is N = 115. Gross sample weight is therefore 575 pounds.

Group 5

Materials in this group occur as a solid or fluid mass in which the variate (X) is either dispersed without being accumulated in separate physical units, or occurs in units that cannot be identified or separated and must be measured in some indirect manner.

In these circumstances there can be no accidental interchange of units (X) and (non-X) during sample collection except at the molecular level as in the sampling of fluids. Therefore, while sampling constant A may have a distinct value for molecular units or similar very small aggregates, its value for any practical unit of measurement becomes negligibly small as the ratio (m) approaches infinity. While the binomial distribution is inoperative with regard to chance variations that occur during sample collection, it is still the prime cause of all segregation.

For materials in this group where the variate (X) is a constituent that can be extracted by chemical means, sampling constant A can generally be calculated for molecular units and constant B can be estimated as before from the average composition of the material and its degree of segregation (z).

For other materials in this group, where (X) does not refer directly to units that can be determined or separated by chemical extraction (such as the compressive strength of briquets, the ductility of bitumen, etc.), sampling constant B can only be determined from available variance data.

Example 12

The sampling of ore in place will be used as an example to illustrate use of the equations given in Table 5.5 under Group 5.

Channel samples are collected from a zinc vein containing 10% metallic zinc in the form of smithsonite ($ZnCO_3$); the degree of segregation of the metal is known to be $z = 0.20$. Since the zinc occurs in the form of the carbonate, it follows that the proportional amount of this constituent is $p = 0.20$; the specific gravity of smithsonite is $d = 4.4$; the average specific gravity of the ore is $D = 2.8$. Sampling constant B,

$$B = p(1 - p)\cdot dz^2/D = 0.010,$$

and the total sample variance:

$$s^2 = 0.010/N.$$

This variance is independent of sample weight. The number of increments required to attain a sampling precision of 1% zinc is found to be:

$$N = 384.$$

5.4.2 Non-binomial variates

In actual sampling practice, many instances are found where the variate has a non-binomial parent distribution. For example, in sampling for the number of defectives, the variate has a parent distribution of the Poisson type. In many other cases the parent distribution is a normal curve although frequency curves of irregular shape are encountered as well.

While the parent frequency curves of variates may differ, they have one property in common: the difference between the true value of any sample and the true mean of the material lot from which a sample originates can be expressed as the algebraic sum of two deviations, one caused by random variation, the other by segregation. The usefulness of this distinction lies in the fact that it applies to any variate and to any material.

The law of propagation of errors applies (see derivation in Appendix A) provided that these two individual deviations are independent of each other for any sample or increment. It is impossible to prove, by mathematical analysis, the correctness of this assumption for all materials and all variates. From tests on the sampling board and results of field trials, however, it can be understood intuitively that the law of propagation of errors has a general application here, meaning that Equations19 and 20 apply, independently of the type of frequency distribution of the variate (X). It may be noted, also, that in cases where the mean value and the standard deviation of a variate are related, it is often possible to transform the variate by substitution of a variate whose mean (M) and standard deviation (s) are substantially independent of one another.

Generally, if (s) is even approximately a function of the mean (M) of (X), the transformations given below are appropriate for stabilizing (s).

When relationship s = f (M):	Use (s) calculated from:
(s) proportional to M^2	Reciprocals of observations
(s) proportional to M	Logarithms of observations
(s) proportional to $\sqrt{M}$	Square roots of observations

Such transformation variates can be used in extreme cases where the above conclusions would not apply.

6. SELECTED TECHNIQUES

The simple techniques presented in this section will find application in a wide variety of problems. Some will be especially useful for the preliminary organization and evaluation of data.

6.1 Rules for Rounding

To minimize errors which might result from rounding off of data in calculations:

1. The rounding interval for a series of observations should be no more than 0.6 time the standard deviation of a single observation.
2. When rounding figures that end with a 5, round to the nearest even number.

Examples:

1.385 ..... round off to 1.38

1.475 ..... round off to 1.48

When two or more decimals are to be eliminated, always round off in a single step.

3. When dealing with several series each containing (n) observations, data are rounded off according to the average range ($\bar{w}$) of each series as shown in Table 6.1 below. The average range ($\bar{w}$) is the average difference between the highest and lowest value of the data.

TABLE 6.1

No. of observations per series	Minimum no. of series	Round off to a maximum interval
2	5	0.6 $\overline{w}$
3	3	0.4 $\overline{w}$
4	2	0.3 $\overline{w}$
5 - 10	1	0.2 $\overline{w}$

6.2 Estimating Missing Data

If it has been found necessary in the course of an experiment to discard observations that were judged to be unreliable, or if some observations are simply missing, the set of data can be completed by estimating the missing values. For this purpose, the set can be regarded as consisting of rows and columns of data. As will be illustrated, blocks or replicates are re-arranged to form sub-rows (or sub-columns) and the sums are then determined in order to estimate the missing value or values.

6.2.1 One observation missing

Estimation of a missing value X_{AP} proceeds from the following:

The residual sum of squares for a complete set with no missing data is,

$$\tau = \alpha - (\beta + \gamma)$$

X_{AP}			A'
			B
			D
P'	Q	T	M'

or,

$$\tau = \left[x^2 - \frac{M^2}{CR}\right] - \left[\frac{P^2 + Q^2 + T^2}{R} - \frac{M^2}{CR} + \frac{A^2 + B^2 + D^2}{C} - \frac{M^2}{CR}\right]$$

$$= x^2 - \frac{P^2 + Q^2 + T^2}{R} - \frac{A^2 + B^2 + D^2}{C} + \frac{M^2}{CR} .$$

The condition $\boxed{\tau = \text{minimum}}$ is fulfilled when

$$\frac{\partial \tau}{\partial X_{AP}} = 0.$$

$$\boxed{0 = X_{AP} - P/R - A/C + M/CR} \quad \ldots\ldots \text{(Eq. 21)}$$

Substituting: $A' = A - X_{AP}$, $P' = P - X_{AP}$ and $M' = M - X_{AP}$ in Equation 21:

$$\boxed{X_{AP} = \frac{A'R + P'C - M'}{(C-1) \cdot (R-1)}} \quad \ldots\ldots \text{(Eq. 22)}$$

Note: There is no preference regarding rows or columns, for arranging the data, i.e. replicates may be either placed in the columns and tests in the rows, or vice versa.

Example:

Briquet strength for two types of asphalt (K,k) and for two methods of application (emulsifier m and atomizer M). Arranging the data as shown in the table below, the missing value is found from Equation 22:

$$X_{AP} = \frac{(406 \times 2) + (153 \times 4) - 964}{(4-1) \cdot (2-1)} = \underline{153}$$

C

			KM	Km	kM	km	Sum
R	Replicate	1	159	108	153	138	558
		2	151	106	?	149	406
	Sum		310	214	153	287	964

After filling in the missing value, the data are analyzed in the normal way, but with one exception, namely, the number of degrees of freedom for the residual variance V_τ is one less than for the normal case, owing to the fact that one value has been estimated, thereby reducing the number of independent observations by one.

6.2.2 Two observations missing

a) Missing observations in different rows and different columns:

$$\boxed{\begin{aligned} X_{AP} &= \frac{A'R + P'C - M' - X_{DT}}{(C - 1) \cdot (R - 1)} \\ X_{DT} &= \frac{D'R + T'C - M - X_{AP}}{(C - 1) \cdot (R - 1)} \end{aligned}}$$

X_{AP}			A'
			B
		X_{DT}	D'
P'	Q	T'	M'

giving 2 equations with 2 unknowns.

In the analysis of variance, subtract 2 extra degrees of freedom for calculating V_T.

b) Missing values in same row:

$$\boxed{\begin{aligned} X_{AP} &= \frac{A'R + P'C - M' + X_{AT} (R-1)}{(C - 1) \cdot (R - 1)} \\ X_{AT} &= \frac{A'R + T'C - M' + X_{AP} (R-1)}{(C - 1) \cdot (R - 1)} \end{aligned}}$$

X_{AP}		X_{AT}	A'
			B
			D
P'	Q	T'	M'

In the analysis of variance, subtract 2 extra degrees of freedom as in a) for calculating V_T.

c) Missing values in same column:

$$\boxed{\begin{aligned} X_{AP} &= \frac{A'R + P'C - M' + X_{DP} (C-1)}{(C - 1) \cdot (R - 1)} \\ X_{DP} &= \frac{D'R + P'C - M' + X_{AP} (C-1)}{(C - 1) \cdot (R - 1)} \end{aligned}}$$

X_{AP}			A'
			B
X_{DP}			D'
P'	Q	T	M'

In the analysis of variance, subtract 2 extra degrees of freedom for calculating V_T.

Formulas for more than 2 missing values can be found in the same way by partial differentiation of the formula for (τ) and substitution of $A' = A - X_{AP}$, etc. For calculating V_τ in the analysis of variance, subtract 1 extra degree of freedom for each missing value that has been estimated.

6.2.3 Correction of sum of squares between variates

Reduction of the degrees of freedom due to missing data inflates the sum of squares between variates. Supposing the variates are classified in columns, it follows generally, that,

$$\Delta\beta = \frac{C-1}{C} \cdot \left[(X_{AP} - \frac{A'}{C} - 1)^2 + (X_{DT} - \frac{D'}{C} - 1)^2 + \ldots\ldots \right],$$

and the corrected sum of squares between variates,

$$\beta' = \beta - \Delta\beta.$$

When variates are classified in rows instead of columns,

$$\Delta\gamma = \frac{R-1}{R} \cdot \left[(X_{AP} - \frac{P'}{R} - 1)^2 + (X_{DT} - \frac{T'}{R} - 1)^2 + \ldots\ldots \right].$$

6.3 Checking for "Tramp" Values

In frequency distributions, a "tramp" value is one whose deviation (x_i) from the mean ($\overline{P}$) is greater than can be attributed to chance variation. It is also therefore, a biased value.

The suspected value (observation) can be tested to determine the probability of bias in terms of the standard deviation (s) and the Normal Distribution as shown in the example given below.

Example:

In the set of data shown in Table 6.2 and Fig. 6.1, observation p_4 appears to be a "tramp". This value will be checked for bias:

$$s^2 = \frac{\Sigma p_i^2 - (\Sigma p_i)^2/n}{n-1} = \underline{8.5}$$

$$s = 2.92$$

$$\overline{P} = 4.0$$

where

n = total number of observations,

s = standard deviation, in class units,

p_i = individual observation, and

$\overline{P}$ = arithmetic mean of observations.

Deviation of the observation p_4 from the mean ($\overline{P}$),

$x_4 = (p_4 - \overline{P}) = 5 = \underline{1.72s}$.

TABLE 6.2

Observation No.	p_i	p_i^2
1	2	4
2	3	9
3	4	16
4	9	81
5	2	4
	$\Sigma p_i = 20$	$\Sigma p_i^2 = 114$

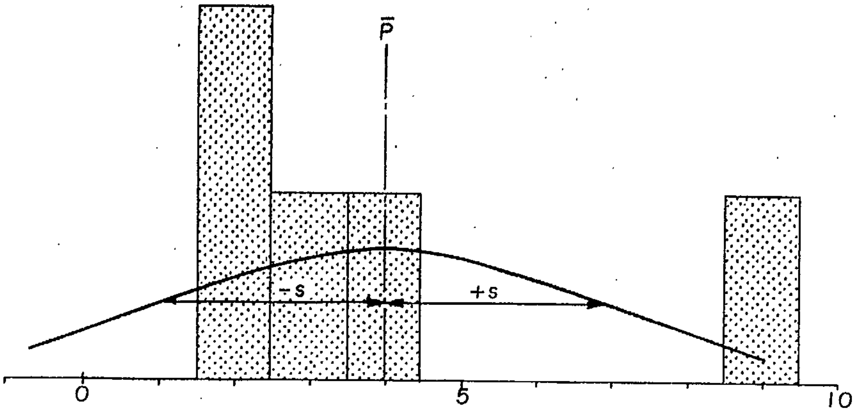


Fig. 6.1

From a Normal Distribution table it is seen that the chances of finding a random deviation of x_4= 1.72s are 4.27%, or less than 1 chance in 20 when the number of observations is very large.

The estimate of (s) in this case is based on only five observations, however, and the fiducial limit (a) with respect to the true(unknown) mean value $\overline{P}$ should therefore be corrected in accordance with the t-test for normal distributions ($a = t \cdot s$). The t-table for degrees of freedom (d.f) $= \infty$ shows that at the 5% level, a = 1.96s, which is identical to the Normal Distribution. For d.f. = 4, we find a = 2.78s at the 5% level; a = 2.13s at the 10% level; and a = 1.42s at the 25% level. Consequently, the level here is not 4.27% but approximately 19% or about 1 in 5. While p_4 cannot be rejected on this basis, it must be remembered that it strongly influences the estimate of (s) itself.

The above calculation is therefore repeated for 4 observations excluding p_4, and we find:

$$x_4' = \underline{6.53s'}$$

This difference indicates a systematic bias at a level of less than 1%, and on this basis, p_4 should be eliminated. The corrected parameters (see also Fig. 6.2) are:

$n' = 4$

$\overline{P}' = 2.75$

$s' = 0.957$

$x_4' = 6.53s'$

$\quad = \underline{6.25}$

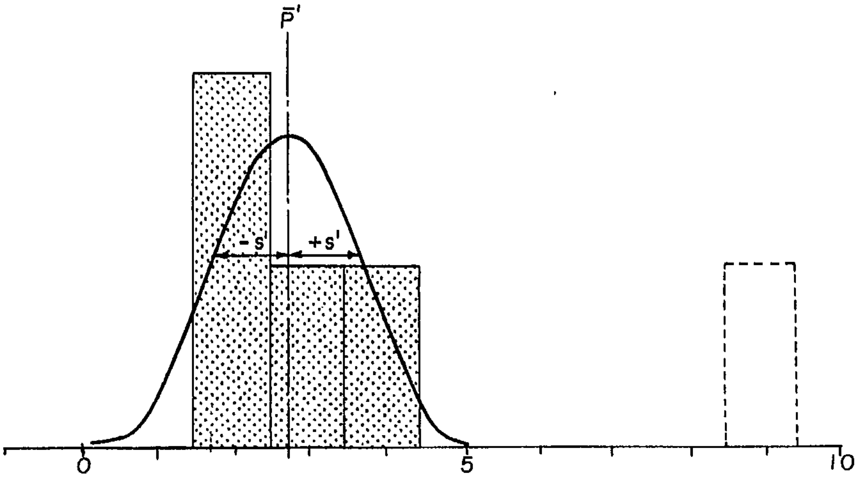


Fig. 6.2

6.4 Randomness of an Oscillatory Series

The following table can be used to test whether the distribution of (m) items A and (n) items B in a series of (m + n) items is random or non-random at the 5% level.

TABLE 6.3 - Value of g

Minimum number of groups - 5% level

Larger no. of items	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
20	3	4	5	6	7	8	9	10	10	11	12	12	13	13	14	14	15	15	16
19	"	"	"	"	"	"	"	9	"	"	11	"	"	"	"	"	14	"	
18	"	"	"	"	"	"	"	"	"	"	"	"	12	"	13	"	"		
17	"	"	"	"	"	"	8	"	"	10	"	11	"	12	"	13			
16	"	"	"	"	"	7	"	"	9	"	"	"	"	"	"				
15	"	"	"	"	"	"	"	"	"	"	10	"	11	"					
14	"	"	"	"	6	"	"	8	"	9	"	10	"						
13	"	"	"	5	"	"	7	"	"	"	"	"							
12	"	"	"	"	"	"	"	"	8	"	9								
11	"	"	4	"	"	6	"	7	"	8									
10	"	"	"	"	"	"	"	"	7										
9	"	3	"	"	5	"	6	"											
8	"	"	"	4	"	5	"												
7	2	"	"	"	"	"													
6	"	"	"	"	4														
5	"	"	3	"															
4	"	2	"																
3	"	"																	
2	"																		

Smaller no. of items

This table is symmetrical; (n) and (m) can be read on either scale. Use the vertical scale for the larger number.

The test applies to a series of numbers in which homogeneous groups having a property (A) alternate with homogeneous groups identifiable by characteristic (B). For instance, the series AA BBB A BB AAAA contains g = 5 homogeneous groups in a total of m = 7 items (A), and n = 5 times (B). The distribution of A and B over the series is random at the 5% level because the number of homogeneous groups found (g = 5) exceeds the required minimum of 4 shown in the g-table.

Example 1

A feeding test on calves is carried out using two different types of feed (A,B). Eight calves are tested for each type. When the average daily weight-increases for all calves are arranged in ascending order, it is found that there are 4 homogeneous groups of feed types A and B. Since the minimum number of groups for m = n = 8 is g = 6, this indicates a non-random distribution. It is thus concluded that either the two types of feed do not have the same effect, or the calf groups are not identical.

Note: This test gives a quick indication only; further study, e.g. by analysis of variance, is required to evaluate the data more fully.

Example 2

A regression curve is drawn through 29 points. There are 14 points (A) above the line and 15 points (B) beneath the line, distributed in 9 homogeneous groups. According to the g-table at (15,14), there should be at least 11 homogeneous groups; distribution of the points is therefore not random. The regression curve does not fit the data sufficiently well for values of $g < 11$.

6.4.1 The randomness of an oscillatory time series may be tested by ascertaining the number of turning points. In a random series of n terms, this number has a mean value of

$$\boxed{\bar{n}_t = 2/3\ (n - 2)}$$

and a variance

$$\boxed{s_t^2 = (16n - 29)/90} \quad .$$

A member U_t is said to be a "peak" if

$$(U_{t-1}) < U_t > (U_{t+1}),$$

and a "trough" if

$$(U_{t-1}) > U_t < (U_{t+1}) .$$

In either case it is a "turning point". The results of the above 2 formulas are independent of the parent distribution. The interval between turning points is called a "phase" (1/2 wave length). For large (n), the average number of points per unit interval or phase is 2/3 and the average phase is therefore 1.5 turning points. Hence, the average distance between peaks (i.e. the wave length) is 3 turning points, which is what we expect to find in a random series.

Example

For $n = 48$ terms;

Expected $\bar{n}_t = 2/3\ (48-2) = 30.67$; Observed $\bar{n}_t = 14$.

$s_t^2 = 8.21$; $s_t = 2.866$.

Since the difference between expected and observed mean values $\bar{n}_t = 16.67 > 5\ s_t$, distribution of the terms is non-random, i.e., the time series shows a trend.

Note: Borderline cases cannot be judged without repeating the experiment and applying the χ^2 test. In this case the χ^2 test is used as follows:

	$\bar{n}_t$	non-$\bar{n}_t$	Total
Observed	14	34	48
Expected	30.67	17.33	48
Difference	-16.67	16.67	0

$$\chi^2 = \frac{(-16.67)^2}{30.67} + \frac{(16.67)^2}{17.33} = \underline{25.1}$$

For d.f. = 1, conclusion is that the oscillations are not random, indicating segregation.

6.5 Estimating Sampling Bias or Analytical Bias

A simple test can be used for detecting bias in a series of duplicate observations obtained by some method of sampling or analysis. When a sampling device is to be tested, analysis is carried out on duplicate samples collected from different materials, using the device. For instance, the significance of bias of a sample splitter can be found by collecting one duplicate from the "save" side of the splitter, and the second duplicate from the "reject" side, repeating the operation (n) times. To test an analytical method, one duplicate obtained by the method may be compared to a second duplicate obtained by some standard method, or to the known true value or, if the bias of an individual analyst is in question, to the average of a large number of observations from different analysts.

Using the results calculated as indicated in Table 6.4, the F-test is applied to check for bias, as follows:

$$F = \frac{\text{Mean Square between columns}}{\text{Residual Mean Square}} = \frac{(n - 1)A^2}{nB - A^2} .$$

TABLE 6.4

Obs. No.	Duplicates		$(x_1 - x_2)$	$(x_1 - x_2)^2$
	x_1	x_2		
1	20.13	20.03	0.10	0.0100
2	20.39	20.07	0.32	0.1024
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
n	20.10	20.05	0.05	0.0025
Sum			A	B
Mean	$\overline{x}_1$	$\overline{x}_2$		

The F-table is entered at d.f. = 1 and d.f. = n-1 respectively.

The actual bias value for sampling devices is found from:

$$\text{Bias} = (\overline{x}_1 - \overline{x}_2)\cdot(1 - c),$$

where c = cutter ratio. The term "cutter" refers to a sampling device (usually an automatic sampler), where a receptacle or cutter selects a small portion from the main flow of the material. The ratio between the size of the sample secured by the cutter and the size of the entire material lot that was sampled is called the "cutter ratio".

When two analytical methods are compared, the actual bias value is found from:

$$\text{Bias} = (\overline{x}_1 - \overline{x}_2)/2.$$

6.5.1 Precision of a biased sample or biased method of analysis

Precision, expressed as the standard deviation of the difference between duplicates, can be formulated as the maximum permissible difference (P = 5%) between duplicates.

$$(x_1 - x_2)_{\max} = 2\sqrt{(B - A^2/n)/(n - 1)}$$

Note: For a very large number of duplicate observations, the coefficient ahead of the root sign is 1.96 instead of 2; for a smaller number of duplicate observations, the coefficient increases slightly (see t-table).

6.6 Construction of the Normal Distribution Curve

This procedure is used for determining the most probable parent distribution curve (Normal Distribution) for a set of (n) observations, whose mean ($\overline{P}$) and standard deviation (s) are known. The method is particularly useful for a large number of observations.

1. Classify the (n) observations into 8 - 12 classes of equal interval and draw a frequency histogram.
2. Calculate the standard deviation (s) from the observations and express it in class-units (the class interval of the histogram is used as the unit); e. g. when the standard deviation is 2% and the class intervals are 5%, then $s_c = 2/5 = 0.4$.
3. Calculate the mode (top of curve) y_o from

$$\boxed{y_o = n/(s_c\sqrt{2\pi}) = 0.3989\ n/s_c}$$

4. Using factors in Table 6.5, calculate ordinates y for five positive and five negative values of x, where x is the distance along the abscissa taken from the mean ($\overline{P}$). The most probable form of the normal curve for the observations can now be drawn through the points derived from the table.

TABLE 6.5

Abscissa $= \pm x$	Ordinate $= y$
0	y_o
$0.5s_c$	$0.8825y_o$
s_c	$0.6065y_o$
$1.5s_c$	$0.3248y_o$
$2s_c$	$0.1353y_o$
$3s_c$	$0.0111y_o$

Comparison of the curve with the histogram may show up discrepancies between the observed frequencies and the theoretical frequencies for the various classes. Large discrepancies indicate significant departures from the normal curve. This can be checked using the t-test and/or the χ^2 test.

Example

Given, n = 85 observations of the moisture content of a raw material, grouped in classes having an interval of 2% moisture, the mean moisture content ($\overline{P}$) is 18.6%, and the standard deviation (s) equals 1.3%. The shape of the normal frequency distribution curve which best fits these data is determined in the following manner:

The standard deviation expressed in class-units (s_c) is

$$s_c = \frac{s}{\text{class interval}} = \frac{1.3}{2} = \underline{0.65},$$

and the mode, y_o, is found from:

$$y_o = \frac{0.3989\ n}{s_c} = \frac{0.3989 \times 85}{0.65} = \underline{52.2}.$$

Using the values given in Table 6.5, the coordinates for the normal curve are the following:

Abscissa, $(=\overline{P} \pm x)$	Ordinate
x	y
0	52.2
0.325	46.1
0.650	31.7
0.975	17.0
1.300	7.1
1.950	0.6

6.7 Calculation of Index Formula

A useful modification of the regression formula is that obtained in an index formula: one which expresses relationship in comparative or relative terms. It becomes possible by this means to gauge the percentage change that will take place in the value of a variable by reason of change in values of the other variables.

It will be assumed that the regression formula is of the form:

$$\boxed{P = AX + BY + CZ + D}$$

Procedure:

1. Substitute the average values, $\overline{P}$, $\overline{X}$, $\overline{Y}$, $\overline{Z}$, for P, X, Y, Z, in the regression formula; thus, $\overline{P} = A\overline{X} + B\overline{Y} + C\overline{Z} + D$.

2. Multiply both sides of this equation by $\frac{100}{\overline{P}}$:

$$100 = \frac{100A\overline{X}}{\overline{P}} + \frac{100B\overline{Y}}{\overline{P}} + \frac{100C\overline{Z}}{\overline{P}} + \frac{100D}{\overline{P}}$$

3. Replace "100" in the first four terms by the index percentage numbers I_P, I_X, I_Y, I_Z. The index formula thus reads:

$$\boxed{I_P = \frac{A\overline{X}}{\overline{P}} I_X + \frac{B\overline{Y}}{\overline{P}} I_Y + \frac{C\overline{Z}}{\overline{P}} I_Z + \frac{100D}{\overline{P}}}$$

Note that the constant term has no index number but retains the value $\frac{100}{\overline{P}}$.

The index formula is used as follows: if the average value of variable X increases by 10%, then the index I_X increases by 10 points as well, and consequently the index value of the dependent variable, I_P, will increase by $\frac{A\overline{X}}{\overline{P}}$ x 10%. The effect of a change in the value of each of the independent variables (X,Y,Z) on the dependent variable (P) can thus be read off directly from the index formula.

6.8 Law of the Propagation of Errors

When a relationship between two or more variables (X,Y,Z) can be expressed mathematically Z = f(x,y), and estimates of the variances of the independent variables (X,Y) are available, the variance of the dependent variable (Z) is found from:

$$\left(\frac{\partial f}{\partial Z}\right)^2 s_Z^2 = \left(\frac{\partial f}{\partial X}\right)^2 s_X^2 + \left(\frac{\partial f}{\partial Y}\right)^2 s_Y^2$$

Example 1: A rectangular piece of land having sides X and Y has been measured with a noticeable error. The error in the surface area of the land z = xy will be:

$$s_z^2 = y^2 s_x^2 + x^2 s_y^2$$

This is depicted in the sketch, where ys_x and xs_y represent the areas of two strips alongside the piece of land. The small square, $s_x\ s_y$, in the upper right hand corner is ignored, it being of a lower order of magnitude.

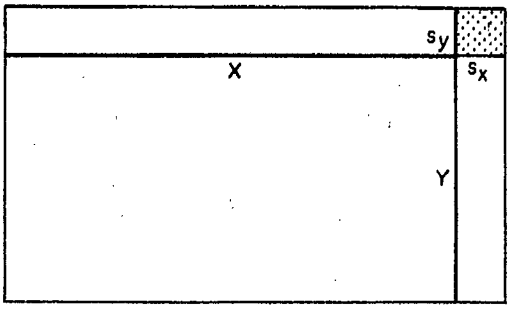


Fig. 6.3

Example 2: The formula for the variance of a mean is found by applying the Law of Propagation of Errors as follows:

$$P = \Sigma p_i/n = \frac{p_1 + p_2 + \ldots\ldots p_n}{n}$$

$$s_p^2 = \left(\frac{1}{n}\right)^2 s_1^2 + \left(\frac{1}{n}\right)^2 s_2^2 + \ \ldots \left(\frac{1}{n}\right)^2 s_n^2$$

where $s_1 = s_2 = s_3 = \ \ldots s_n = s.$

Therefore $s_p^2 = \frac{s^2}{n}.$

Example 3: The variance of the difference between two variates equals the sum of the variancesof the individual variates:

$$Z = X - Y$$

It follows directly from the formula for the propagation of errors that

$$s_z^2 = s_x^2 + s_y^2.$$

The Law of Propagation of Errors applies generally to any relationship, provided that the variables X, Y, ... are substantially independent of one another.

A further application of the Law of Propagation of Errors is illustrated in Appendix A.

ACKNOWLEDGEMENTS

Permission from the Director, Mines Branch, Department of Energy, Mines and Resources, for publishing this Guide is gratefully acknowledged.

The authors are indebted to Mr. F. A. Sorensen, U. S. Steel Corporation, and to Dr. R. P. Charbonnier, Mines Branch, Department of Energy, Mines and Resources, Ottawa, whose valuable comments and great patience in reviewing earlier drafts have made completion of this Guide possible.

The contents of Section 5, Sampling, were first published in 1962 as a Mines Branch Report, R-93, under the title "Towards a Common Basis for the Sampling of Materials". We derive a great deal of satisfaction from the fact that the views expressed at that time have now been accepted by the ASTM and incorporated in a general sampling theory (10). In this regard, thanks are due in large part to Prof. J. R. McGregor of the Department of Mathematics, University of Alberta, Edmonton, and again to Dr. R. P. Charbonnier, for reading the original manuscripts and providing useful observations.

The editorial assistance of Mr. P.E. Shannon is gratefully acknowledged.

LITERATURE REFERENCES

1. Moroney, M. J., "Facts from Figures", 2nd ed., Penguin Books Ltd., London, 1953, p. 334 ff.
2. Ibid., p. 336 ff.
3. Ibid., p. 340 ff.
4. Ibid., p. 398 ff.
5. Cowden, D. J., "Statistical Methods in Quality Control", Prentice Hall Inc., Englewood Cliffs, N. J., 1957.
6. Mentzer, E. G., "Tests by the Analysis of Variance", Wright Air Development Centre, Tech. Rept. 53-23, Jan. 1953.
7. Pearson, K., "On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurements of Organs", Proc. Royal Soc. (London), Ser. A, <u>60</u>, 1932, pp. 489-502.
8. Boyard, G., "Application du plan de travail factoriel à la recherche en flottation", Revue de l'Industrie Minérale, <u>35</u> (1954), No. 607, pp. 346-365.
9. Benson, M. A., "Spurious Correlation in Hydraulics and Hydrology", J. Hydraulics Div., Proc. Am. Soc. Civil Eng., July 1965, pp. 4393-99.
10. Visman, J., "A General Sampling Theory", ASTM Materials Research and Standards, Vol. 9, No. 11, November, 1969.

GLOSSARY OF TERMS

Synonymous or related terms used in the text of the definitions are underscored. Terms that are defined elsewhere in this Glossary are spatiated.

accuracy, a term generally used to indicate the closeness of agreement between an experimental result and the true value; it is affected by chance errors as well as by bias. The term accuracy is not used specifically as a measure of variability; see precision.

attribute, a quality of an item or component part that is either affirmed or denied (e.g. a machine part is either accepted, or rejected as defective).

average value, s. mean value.

bias, s. systematic error; significant bias.

biased sample, a sample whose composition is biased by contamination with foreign matter or by disproportionate inclusion or exclusion of certain true components of the shipment.

central value, the value half-way between the upper and lower limits of a class interval; see frequency distribution.

chance deviation, s. chance error.

chance error, error associated with a probability distribution and whose algebraic sum tends to zero; random error; random deviation; chance deviation.

component, s. random sampling variance, segregation variance.

composite sample, s. gross sample.

consignment, s. shipment.

<u>constant</u>, s. unit random variance; segregation variance.

<u>degree of segregation</u>, a numerical value on a scale ranging from zero, indicating a perfect mixture, to one indicating a state of complete segregation. See segregated, homogeneous.

<u>deviation</u>, s. error; chance error; standard error.

<u>distribution</u>, s. frequency distribution; parent distribution; probability distribution.

<u>duplicate samples</u>, two samples collected from the same population.

<u>error</u>, 1) a mistake; 2) the difference between the observed or estimated value and the mean value, or the true value, or some other standard value; <u>deviation</u>. The term deviation is commonly used when an involuntary error or discrepancy is indicated, while the term error infers that the difference can be controlled to some degree by a voluntary act.

<u>error</u>, s. chance error; standard error; systematic error.

<u>error variance</u>, the mean square of errors.

<u>frequency distribution</u>, graphical or tabular presentation of the quantitative relationship between the relative abundance of material units of a given size within a given range or class interval of a variate (ordinate), and the values of the variate representing the central values of these classes, in numerical order (abscissa).

<u>gross sample</u>, a sample consisting of a number of increments; <u>composite sample</u>.

<u>homogeneous</u>, 1) of the same nature or kind throughout; 2) the state of being perfectly blended; zero degree of segregation.

<u>increment</u>, a sample taken by one operation of a sampling device, for the purpose of combining it with other increments to form a gross sample. An increment is usually

not analyzed separately; and if so, is preferably termed a one-increment sample.

lot, s. shipment.

mean value, arithmetic average; average value.

parameter, 1) a quantity conveniently used for indirectly measuring a variate or variable property of a material or any other statistical population.
para (Gr.) = beside, near.

parent distribution, frequency distribution of a variate characteristic of a statistical population of units having a specified size.

population, s. statistical population.

precision, 1) a term used to indicate the capability of a person, an instrument or a method to obtain reproducible results;
2) a measure of the chance error as expressed by the variance, the standard deviation, or a multiple of the standard deviation (see ASTM Recommended Practice E177, Parts 27 and 30 (1968).

probability distribution, frequency distribution of any random variable, e.g. a chance error.

random deviation, s. chance error.

random error, s. chance error.

random sampling, collecting samples at random.

random sampling variance, variance of the parent distribution of a given sample; random variance component.

random, s. unit random variance; random variance component; random sampling variance.

random variance component, s. random sampling variance.

replicate samples, a series of more than two samples taken from the same population.

representative sample, a sample without bias; true sample; unbiased sample.

resample, 1) to sample again, for the purpose of replacing another sample collected previously; 2) a sample collected for the purpose of replacing another sample.

sample, 1) a quantity of material taken from a larger quantity for the purpose of estimating properties of the larger quantity; 2) to collect sample.

sampling constant A, s. unit random variance.

sampling constant B, s. segregation variance.

sampling, s. random sampling; random sampling variance; systematic sampling; total variance of sampling.

segregated, 1) not homogeneous; 2) the state of being imperfectly mixed.

segregation, degree of segregation.

segregation variance, 1) variance due to segregation; 2) difference between the total variance of sampling and the random sampling variance; segregation variance component; sampling constant B.

segregation variance component, s. segregation variance.

shipment, 1) a commercial or negotiable quantity of material that is transferred from seller to buyer; 2) a discrete quantity of material that is presented for inspection and acceptance; consignment; 3) a specified quantity of material from a common source; lot.

significant bias, bias that is of appreciable economic importance to the concerned parties.

standard deviation, s. standard error.

standard error, the root mean square of errors; standard deviation.

statistical collection, s. statistical population.

statistical population, 1) a collection of discrete items or units of a given size characterized by a common variate; universe; statistical collection; 2) all of the pieces, particles, items, persons or other component parts that constitute the whole content of that which is the subject of specific interest separately and individually and about which knowledge is to be inferred from one or more samples drawn from it.

subsample, 1) a sample taken from another sample; 2) to collect sample from another sample.

systematic error, error that is consistently positive, or consistently negative; 2) error associated with a probability distribution whose mean value does not tend to the true value; bias.

systematic sampling, collecting samples at regular intervals.

total variance of sampling, 1) the mean square of errors due to sampling; 2) the sum of the random sampling variance and the segregation variance.

true sample, s. representative sample.

true value, 1) mean value of a variate; 2) any standard value concurrently accepted by joint parties as a basis for negotiation.

unbiased sample, s. representative sample.

unit random variance, variance of the parent distribution of a sample of unit size (e.g., 1 item, 1 lb., 1 kg.) unit variance; sampling constant A.

unit variance, s. unit random variance.

universe, s. statistical population.

value, s. central value; mean value; true value.

variance, the root mean square of errors.

variance, s. error variance; random sampling variance; segregation variance; total variance of sampling; unit random variance.

variate, a quantity used to express and measure a variable property of a material or any other statistical population.

LIST OF IMPORTANT SYMBOLS

α Total sum of squares of all items in Analysis of Variance (ANOVA).

A Sampling Constant, denoting the parent variance of a sample of unit size; random unit variance.

a, a_{95} Precision at a given probability level (e.g. P = 95%).

a_V Precision of the error variance.

β Sum of squares between columns of an array (ANOVA).

B Sampling Constant, denoting the segregation variance of a consignment.

δ Sum of squares between cells of an array (ANOVA).

d_x Horizontal deviation of an observed value and corresponding value on the regression curve.

d_y Vertical deviation of an observed value and corresponding value on the regression curve.

d.f. Degrees of freedom.

ϵ Interaction sum of squares (ANOVA).

E(...) Expected value of a quantity.

F Ratio of two variance estimates; test-ratio; F-ratio.

F_A, F_B.. Test ratio of variance (subscripted) (ANOVA).

f Degrees of freedom.

γ Sum of squares between rows of an array (ANOVA).

g Average deviation in a series of observations; theoretical minimum number of homogeneous groups required to ascertain randomness of the groups in an oscillatory series.

H Number of replicate observations.

h, h_j... Sum of H replicates.

$I_p, I_x \ldots$	Index numbers relating to terms in a regression formula.
j	Variation coefficient of the independent variable, X.
k	Variation coefficient of the dependent variable, Y.
μ	True mean value of a population.
M.S.	Mean square: = Sum of squares/d.f. = Variance.
m	Average number of elemental units (items) per unit of measurement.
N	Residue or proportion of variation not explained by regression; number of increments (items) in a gross sample.
n	Frequency of observations in an interval of grouped observations; total number of observations; as exponent, refers to the number of factors in Factorial testing.
$\bar{n}_t$	Average number of "turning points" in an oscillatory time series containing n terms.
$\bar{P}$	Arithmetic mean of a number of observations p_i.
P	Probability (significance) level.
$p_1, p_2 \ldots$	Individual observations.
p	Binomial probability.
χ^2	Chi-square, a quantity representing the relative size of differences between observed and theoretical frequencies (Chi-square test).
r	Probable error; correlation coefficient.
σ	True standard deviation of a population: = E(s).
S.S.	Sum of squares.
$S_A, S_B \ldots$	Sum of squares of factor (subscripted) in Factorial Analysis.
S	Sum of observations within a cell (ANOVA).
s	Standard error (deviation) of a population based on a limited number of observations.

s_c	Standard deviation expressed in class-units.
s_s	Standard deviation of a standard error.
s_v	Standard deviation of a variance.
s^2	Variance; mean square of errors (deviations); total sampling variance.
s_p^2	Random unit variance; parent variance of a sample of unit size.
s_s^2	Segregation variance component of total sampling variance.
s_t^2	Variance of turning points in an oscillatory time series.
τ	Error/residual sum of squares (ANOVA).
t	Absolute value of the ratio of a variate to its standard error (t-test).
t_{i1}, t_{i2}	Random deviation (1) and deviation caused by segregation (2) of a sample (i) from the true average value of the population (lot, consignment).
U_t	Member of an oscillatory time series.
V	Variance estimate; variate; attribute of a variate.
V_β	Mean square between columns of an array (ANOVA).
V_ϵ	Interaction mean square of an array (ANOVA).
V_γ	Mean square between rows of an array (ANOVA).
V_τ	Error/residual mean square of an array (ANOVA).
V_C	True variance estimate between columns of an array (ANOVA).
V_{CR}	True variance estimate of interaction in an array (ANOVA).
V_R	True variance estimate between rows of an array (ANOVA).
W	Size of gross sample.
w	Range of values in a series of n observations; size of

	increments in sampling; fractional number of units in binomial sampling.
X	Variate of a lot, consignment or any other population.
$X_{AP}, \ldots$	A missing value in an array, to be estimated. Subscripts denote row and column coordinates.
$x_1, \ldots$	Observed value of a variate.
x	Deviation of a value from the arithmetic mean ($x = p - \bar{P}$).
$\bar{x}_1, \ldots$	Arithmetic mean of a series of observations.
y	Relative frequency of a variable.
y_o	Mode of frequency curve.
z	Quantity representing the degree of segregation of a population attribute.

-- ---

APPENDIX A

LAW OF PROPAGATION OF ERRORS

Application to Random and Segregation Variations

The true value (x) of a sample (i) collected from a segregated population having a true average value (y) can be written as follows:

$$x_i = y \pm t_{i1} \pm t_{i2}$$

where t_{i1} = random deviation, and

t_{i2} = deviation caused by segregation.

The total deviation for any sample (i) is, therefore:

$$x_i - y = t_i = \pm t_{i1} \pm t_{i2}.$$

For a large number of samples, it follows that

$$t_1^2 = t_{11}^2 + t_{12}^2 \pm 2t_{11} \cdot t_{12}$$

$$t_2^2 = t_{21}^2 + t_{22}^2 \pm 2t_{21} \cdot t_{22}$$

$$\cdot\cdot \quad \cdot\cdot \quad \cdot\cdot \quad \cdots$$

$$\cdot\cdot \quad \cdot\cdot \quad \cdot\cdot \quad \cdots$$

$$t_n^2 = t_{n1}^2 + t_{n2}^2 \pm 2t_{n1} \cdot t_{n2}.$$

The average:

$$\frac{\Sigma t_i^2}{n} = \frac{\Sigma t_{i1}^2}{n} + \frac{\Sigma t_{i2}^2}{n} + \frac{2\Sigma(\pm t_{i1}\, t_{i2})}{n}.$$

It follows, by first-order approximation, that:

$$s^2 = s_1^2 + s_2^2,$$

where s_1^2 = random variance, and

s_2^2 = segregation variance.

The mean value of the double products is of a lower order of magnitude owing to opposite signs, provided that there is no correlation between t_{i1} and t_{i2}.

This derivation applies to any type of parent distribution and supports the general validity of Equation 16.

APPENDIX B

Significance of Correlation Coefficient (r) for 5% and 1% Points

(d.f. = N - n_r)

% Points	d.f. (φ)	Number of Variables*, (n_r)				d.f. (φ)	Number of Variables, (n_r)			
		2	3	4	5		2	3	4	5
5	1	.997	.999	.999	.999	24	.388	.470	.523	.562
1		1.000	1.000	1.000	1.000		.496	.565	.609	.642
5	2	.950	.975	.983	.987	25	.381	.462	.514	.553
1		.990	.995	.997	.998		.487	.555	.600	.633
5	3	.878	.930	.950	.961	26	.374	.454	.506	.545
1		.959	.976	.983	.987		.478	.546	.590	.624
5	4	.811	.881	.912	.930	27	.367	.446	.498	.536
1		.917	.949	.962	.970		.470	.538	.582	.615
5	5	.754	.836	.874	.898	28	.361	.439	.490	.529
1		.874	.917	.937	.949		.463	.530	.573	.606
5	6	.707	.795	.839	.867	29	.355	.432	.482	.521
1		.834	.886	.911	.927		.456	.522	.565	.598
5	7	.666	.758	.807	.838	30	.349	.426	.476	.514
1		.798	.855	.885	.904		.449	.514	.558	.591
5	8	.632	.726	.777	.811	35	.325	.397	.445	.482
1		.765	.827	.860	.882		.418	.481	.523	.556
5	9	.602	.697	.750	.786	40	.304	.373	.419	.455
1		.735	.800	.836	.861		.393	.454	.494	.526
5	10	.675	.671	.726	.763	45	.288	.353	.397	.432
1		.708	.776	.814	.840		.372	.430	.470	.501
5	11	.553	.648	.703	.741	50	.273	.336	.379	.412
1		.684	.753	.793	.821		.354	.410	.449	.479
5	12	.532	.627	.683	.722	60	.250	.308	.348	.380
1		.661	.732	.773	.802		.325	.377	.414	.442
5	13	.514	.608	.664	.703	70	.232	.286	.324	.354
1		.614	.712	.755	.785		.302	.351	.386	.413
5	14	.497	.590	.646	.686	80	.217	.269	.304	.332
1		.623	.694	.737	.768		.283	.330	.362	.389
5	15	.482	.574	.630	.670	90	.205	.254	.288	.315
1		.606	.677	.721	.752		.267	.312	.343	.368
5	16	.468	.559	.615	.655	100	.195	.241	.274	.300
1		.590	.662	.706	.738		.254	.297	.327	.351
5	17	.456	.545	.601	.641	125	.174	.216	.246	.269
1		.575	.647	.691	.724		.228	.266	.294	.316
5	18	.444	.532	.587	.628	150	.159	.198	.225	.247
1		.561	.633	.678	.710		.208	.244	.270	.290
5	19	.433	.520	.575	.615	200	.138	.172	.196	.215
1		.549	.620	.665	.698		.181	.212	.234	.253
5	20	.423	.509	.563	.604	300	.113	.141	.160	.176
1		.537	.608	.652	.685		.148	.174	.192	.208
5	21	.413	.498	.552	.592	400	.098	.122	.139	.153
1		.526	.596	.641	.674		.128	.151	.167	.180
5	22	.404	.488	.542	.582	500	.088	.109	.124	.137
1		.515	.585	.630	.663		.115	.135	.150	.162
5	23	.396	.479	.532	.572	1000	.062	.077	.088	.097
1		.505	.574	.619	.652		.081	.096	.106	.115

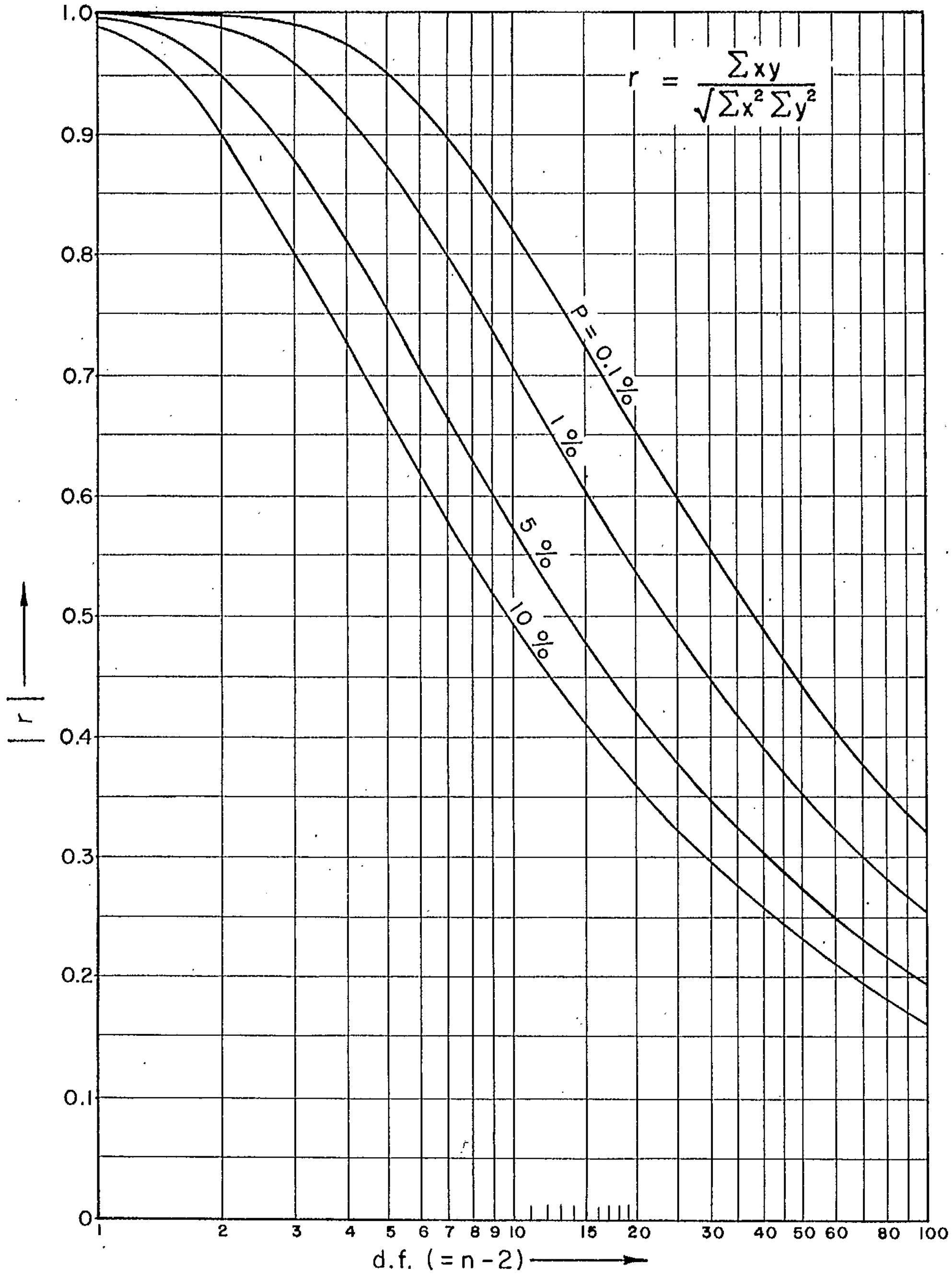
* Includes both dependent and independent variables

APPENDIX C

Charts of Statistical Functions

Chart 1

Correlation Coefficient r - 2 Variables*



*For moro than 2 variables, rofer to table of r-valuos, p. 147.

Chart 2

5 % Level of F-ratio

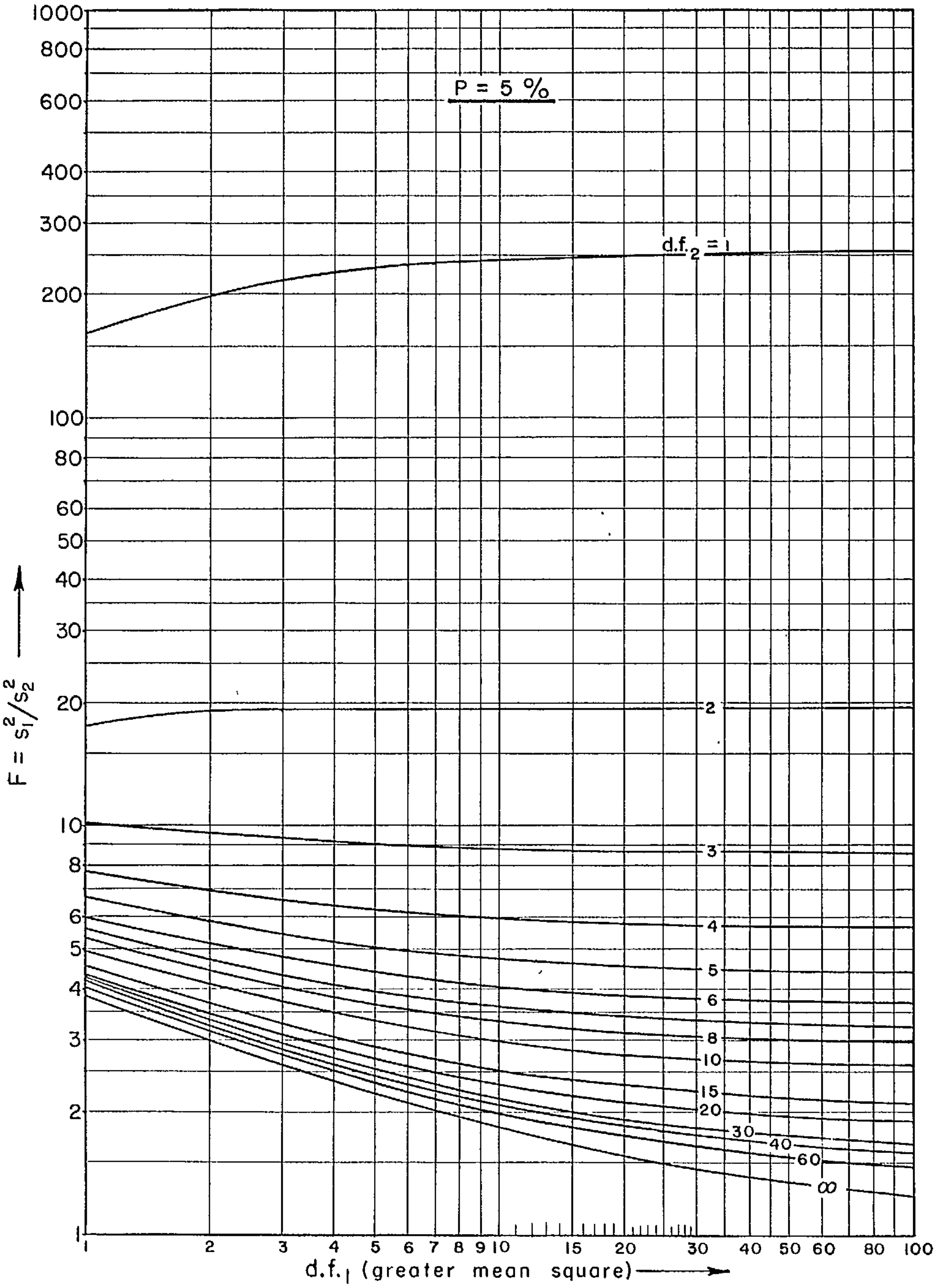


Chart 3

Significance Levels of t - Equal Tails

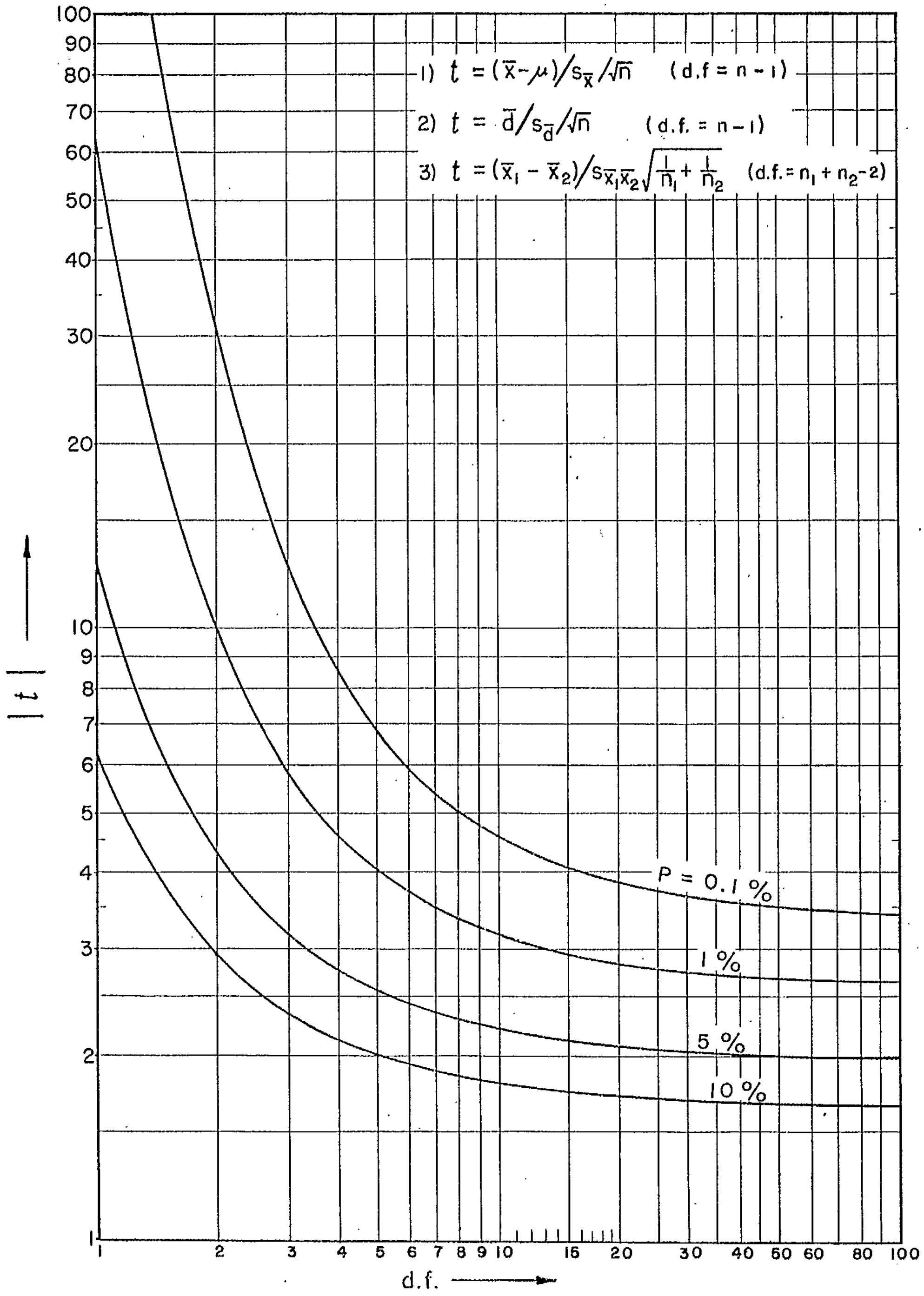
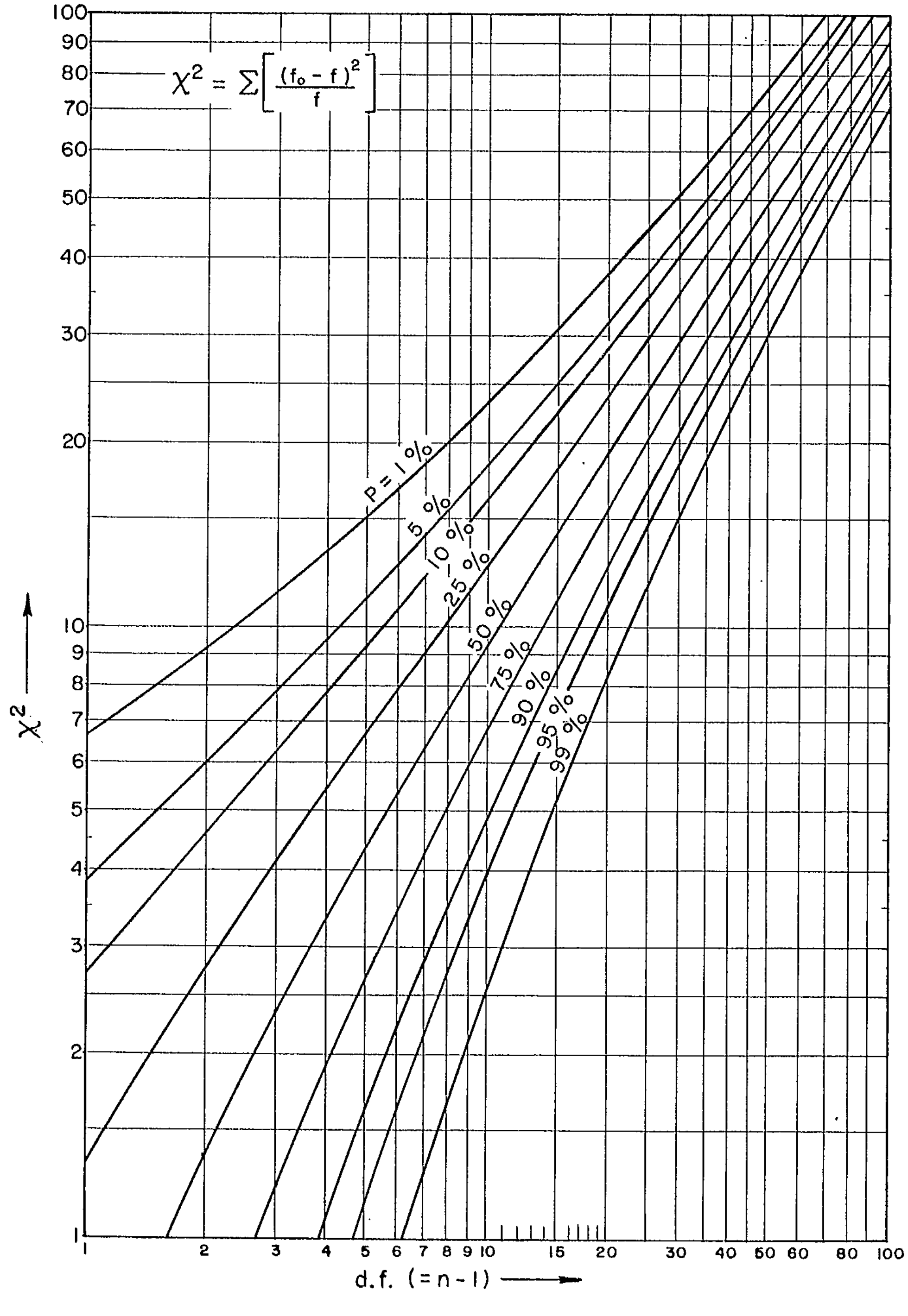


Chart 4

Significance Levels of χ^2



INDEX