FOREWORD

This is the third of a series of bulletins issued primarily for the use of members of the Inspection Engineering Department. Each of these bulletins treats of a particular phase of the general subject, "Control of Quality of Manufactured Product". An attempt has been made to make the discussion in each bulletin as nearly as possible a complete and independent unit so that the material contained therein may be used independently of that contained in other bulletins of the series. On the other hand, however, it is hoped that when all the bulletins in the series have been issued, they will constitute a unified treatment of the above subject, divided into the following Parts:

I - Introduction.

II - Presentation of Data by Means of Simple Statistics.

III - Basis of Quality Control.

IV - Detection of Quality Variations which Should not Be Left to Chance.

V - Measurement of Quality.

VI - Quality Standards for Raw Materials.

VII - Economic Control of Quality through Inspection.

VIII - Economic Control of Quality through Design.

IX - Tables and Nomograms with a Discussion of Nomographic Treatment of Data.

The order of presentation of these Parts has been governed by the immediate needs of the department. For example, I.E.B. 1 and I.E.B. 2 constitute as it were Parts IV and V of the completed story. The present bulletin, I.E.B. 3, constitutes Part II and treats of the subject,"Presentation of Data by Means of Simple Statistics".

The natural starting point in any engineering or scientific investigation is the collection of data and the presentation of information contained therein. Obviously this constitutes a major problem for the inspection engineer but, since the problem itself is perfectly general, the discussion has not been limited to what are customarily accepted as inspection engineering problems, so as to increase the usefulness of the results in other phases of engineering and scientific work.

Two concepts have been introduced, namely, that of total as contrasted with essential information. It is true that any scientific or engineering investigation customarily starts with the collection of data for the purpose of answering one or more specific questions. The information contained in such data and useful for answering a specific question is essential for that question. However, the information contained in such data and useful for answering all possible questions, constitutes the total information. Obviously, therefore, we cannot present the essential information contained in a series of observations unless we are given a specific question to be answered. We can, however, consider the problem of presenting the total information contained in a series of observations without setting up any specific question and since the essential information in any case cannot exceed the total, ways and means of presenting as much as possible of the total information contained in a series of observations by means of simple statistics have been considered.

It is obvious that such information must be presented by means of symmetric functions of the data for otherwise the conclusions would depend upon the order in which the data were taken. Three simple functions or statistics of this type, namely, the arithmetic mean, standard deviation and correlation coefficient are shown to be satisfactory for the presentation of a large amount of the information contained in any set of data, particularly where the number of observations is small. In fact, these simple statistics are shown to constitute, as it were, an almost universal language for the presentation of the information contained in sets of observations on one or more variables. One important novel feature of the discussion is that the presentation of information is treated independently of any sampling theory. We consider the significance of the average, standard deviation, correlation coefficient and similar simple statistics derived from observed data only to the extent that they present information contained in the data.

TABLE OF CONTENTS

CHAPTER I

Introduction

1. Why We Take Data

You go to your tailor for a suit of clothes and the first thing that he does is to make some measurements; you go to your physician because you are ill and the first thing that he does is to make some measurements. The objects of making measurements in these two cases are different. They typify the two general objects of making measurements now to be considered. They are:

a. To obtain quantitative information.

b. To obtain a causal explanation of observed phenomena.

Measurement to attain the first object enters into our everyday life because everything that we buy or sell is by the yard, the pound or some quantitative unit of measure. Such measurements are used at every step in the fabrication of commercial products from raw materials to the finished article. Particularly in the inspection of quality of product does this first function of measurement play an important rôle.

The second object for taking data is however of even greater importance than the first in the field of research and development because here we are in search of physical principles to explain the observed phenomena so that we may predict the future in terms of the past. For example, the savage of old observed an eclipse of the sun as we do today but he could not foretell as we can the time of the next eclipse. In the control of quality of manufactured product it is one thing to measure the quality to see that it meets certain standards and it is quite another thing to make use of these measurements to predict and control the quality in the future.

For example, three typical situations call for causal interpretation. They are:

A. We note differences in the qualities of a number of the same kind of things constituting a group even though so far as we know the qualities have been produced under the same essential conditions. We ask ourselves: "Why do these differences occur and is there a major controlling influence?" As a specific case we observe that the apples on a tree differ in size. Why? Must such differences be left to chance, or may they be explained in terms of causes which can be found and controlled?

B. Two series of observations of some quality have been taken under what may or may not have been the same essential conditions. From an analysis of the data can we determine whether or not the two sets of conditions were essentially the same? To be specific, two apple trees of the same kind are treated with different fertilizers. Do the differences between the sizes of the apples on one tree and those on the other indicate that the fertilizers were assignably different in their influence upon the size of the apples?

C. We obtain a series of pairs of quality characteristics on a number of the same kind of thing and from these we are to determine whether or not there is any underlying causal relationship between these two quality characteristics. To carry through our apple illustration, we assume that one hundred trees had been treated with fertilizers which are the same except for the nitrogen content. We wish to determine whether or not there is a causal relationship between the nitrogen content and the size of an apple.

These are typical questions, to answer which we need to use observed data.

2. Object of Analysis of Data

Of course we take data as just stated to answer certain questions of either one of the two types. The kind of data that the tailor takes in the example cited above did not call for much analysis in the technical sense of the term. Table 2.1 however presents a typical set of raw data which does require analysis. Here we have a series of 1370 pairs of observations on as many telephone poles. One of the series of observations represents the depth of sapwood X and the other the depth of penetration Y of the creosote into the sapwood.

X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
4.60	1.90	3.00	1.20	3.20	1.50	3.70	2.00	2.20	2.15	3.00	1.80	3.25	2.20	2.60	1.50	2.20	1.10	1.35	0.70	4.40	1.60	4.50	2.40	3.90	1.40	3.10	2.15
2.15	1.30	2.90	1.80	2.30	1.75	2.70	0.80	1.90	1.50	1.60	1.00	3.20	1.85	2.10	1.20	2.20	1.20	3.30	1.05	3.40	2.40	2.80	1.30	2.70	1.20	4.10	3.40
3.70	1.60	3.55	1.50	4.10	2.00	4.40	1.90	4.10	2.90	4.20	1.05	2.75	2.20	3.50	2.15	2.70	1.10	3.10	1.40	2.05	1.60	2.10	1.70	2.05	1.00	1.30	0.80
2.70	2.20	3.30	2.00	3.50	1.45	2.50	1.20	2.80	2.45	2.90	2.30	3.50	1.60	3.60	1.70	3.15	0.80	3.80	1.20	2.90	1.10	1.15	0.90	3.30	1.00	4.00	1.80
3.20	1.50	3.40	1.85	1.60	0.90	3.60	2.05	2.15	1.90	3.00	0.95	2.35	1.10	3.60	1.45	2.25	0.90	2.70	0.80	3.30	2.20	3.30	1.60	2.80	1.60	2.40	1.20
2.40	1.70	3.80	2.40	3.70	1.90	3.70	1.80	1.90	1.30	2.85	0.85	2.50	1.20	3.20	0.85	2.40	1.10	1.55	1.00	2.80	1.20	1.75	0.95	1.15	0.80	2.60	1.30
2.15	1.15	1.70	1.00	1.75	1.10	1.50	0.85	1.80	1.15	1.95	0.80	2.70	2.50	2.90	2.10	1.60	1.45	2.30	1.00	3.60	2.30	2.10	1.50	1.80	1.40	2.60	1.10
1.90	1.50	3.10	1.80	2.50	1.50	2.60	1.60	3.60	2.20	2.90	1.40	2.00	1.05	2.35	1.00	3.05	1.40	2.60	1.90	3.40	1.70	3.20	2.00	2.80	1.80	2.50	1.40
3.50	1.50	2.50	1.90	2.30	1.55	2.00	1.20	1.75	0.90	2.60	1.10	2.45	1.35	3.30	1.15	3.65	2.00	3.35	1.85	1.65	1.15	3.70	1.00	2.00	1.30	2.10	1.00
3.50	2.70	2.95	1.35	3.40	1.50	3.60	2.15	3.90	2.55	1.90	1.50	3.05	1.25	3.20	2.65	1.95	1.80	2.65	1.20	2.60	2.05	2.25	1.65	1.95	1.15	3.25	2.00
2.10	1.05	3.00	2.00	2.85	1.25	1.90	1.20	1.90	1.60	2.85	1.25	3.00	2.30	2.35	1.80	4.40	0.90	2.65	1.10	2.45	1.15	3.20	1.45	3.70	1.20	3.40	1.30
2.45	1.60	2.75	1.20	2.60	1.50	3.65	2.00	2.25	1.60	2.80	1.40	2.05	1.50	3.70	1.50	2.40	1.40	3.05	0.70	3.55	1.65	4.00	3.60	1.45	1.10	3.20	1.50
3.30	1.45	2.65	1.65	3.50	1.40	3.00	1.10	2.30	0.90	2.20	0.90	2.60	1.00	2.60	1.15	2.50	0.75	1.50	0.80	1.40	0.75	1.90	1.30	2.70	2.40	3.40	1.10
2.90	1.65	2.25	1.35	3.70	1.35	3.05	1.50	2.10	0.95	3.10	1.50	1.90	0.90	1.60	1.05	3.40	1.20	1.60	1.45	3.05	1.20	3.60	1.45	3.70	1.40	3.00	2.20
3.10	2.10	3.90	3.30	3.90	1.30	4.00	2.85	2.30	1.75	2.60	0.80	4.20	2.10	3.85	1.85	3.75	1.95	2.10	1.15	3.80	2.75	3.00	1.05	4.00	2.65	3.30	2.20
3.50	1.60	1.70	1.55	2.05	1.60	2.70	2.00	3.35	2.85	3.70	1.10	3.25	1.95	1.95	1.50	2.65	1.00	3.25	2.75	1.85	1.05	1.75	1.10	3.40	1.60	2.80	1.60
1.50	0.85	2.70	1.25	3.15	1.75	3.10	1.50	3.25	2.30	2.90	0.90	3.25	1.25	2.45	1.55	2.20	1.40	2.85	1.60	1.55	0.85	2.30	1.70	2.70	1.40	2.45	1.35
2.95	1.55	3.10	1.35	3.35	1.90	3.20	1.75	3.30	1.65	2.80	2.55	2.80	0.90	2.45	1.00	2.90	1.45	3.80	2.00	3.50	2.40	2.75	1.30	3.25	2.20	2.35	1.35
2.90	0.90	4.10	1.10	4.05	1.30	2.40	1.50	2.15	1.20	3.05	1.90	2.60	0.90	2.25	1.60	2.70	1.00	2.70	2.40	4.30	1.30	4.20	1.05	3.30	1.70	2.20	1.20
1.70	0.85	2.55	1.40	2.10	1.40	2.80	1.40	2.30	1.50	2.40	1.45	3.40	0.95	1.95	1.35	2.20	1.10	3.40	1.90	2.85	1.90	2.80	1.90	2.70	1.30	2.25	2.00
1.10	1.00	2.40	1.30	3.05	1.05	3.40	2.10	1.65	0.95	3.20	1.70	3.00	2.50	2.50	1.80	2.50	1.00	2.90	1.35	2.70	1.20	2.60	1.70	2.80	0.75	2.70	1.60
2.70	1.20	2.80	1.60	2.20	0.90	1.65	0.75	3.05	1.30	1.85	1.60	2.50	1.80	2.90	1.40	1.95	1.20	3.00	2.00	3.70	2.10	2.10	1.20	3.00	0.60	4.40	1.85
3.10	0.85	4.70	1.50	2.40	0.90	4.50	1.90	3.90	0.70	5.00	2.60	4.90	1.70	4.00	2.50	3.10	0.85	3.40	2.90	2.70	2.10	3.05	1.95	3.15	1.60	3.10	1.20
3.30	1.80	4.15	3.00	4.25	2.65	3.40	1.40	3.60	1.95	2.70	1.55	3.05	1.35	4.20	3.10	2.25	1.10	3.00	0.70	2.40	0.40	4.00	1.60	4.30	2.05	3.20	1.30
2.80	1.15	2.75	0.80	3.20	2.10	1.30	0.60	2.85	0.90	4.00	2.40	2.85	1.30	3.60	1.20	3.20	2.20	2.30	1.65	2.25	0.85	3.50	2.30	2.55	1.05	3.00	1.15
3.20	2.55	2.80	1.90	2.65	1.15	2.10	1.10	2.05	1.50	2.15	1.65	2.65	1.20	3.65	2.40	2.05	1.25	2.90	2.60	2.70	1.80	3.10	0.80	1.90	1.85	5.80	4.70
3.50	1.80	3.70	1.70	4.80	2.70	4.80	2.90	4.00	1.80	3.70	1.70	2.70	2.05	3.20	2.05	2.60	0.95	2.70	1.75	1.15	0.55	1.75	1.60	1.40	1.05	3.20	2.45
1.70	1.00	1.90	1.50	1.95	1.00	3.75	3.35	2.85	2.30	3.20	1.25	3.00	1.60	2.70	2.00	1.90	1.00	3.00	2.85	2.05	1.30	2.40	0.90	2.70	2.60	2.55	1.40
2.50	0.80	3.45	2.90	2.15	1.50	2.05	1.10	2.20	1.40	2.40	1.40	3.45	2.20	1.20	0.80	2.00	1.05	3.45	0.85	1.60	0.80	3.10	2.15	4.00	2.70	4.40	1.30
4.10	2.60	3.10	1.50	3.30	0.80	3.05	2.00	4.55	3.15	4.05	2.50	3.10	1.00	3.75	2.50	3.30	0.80	2.35	1.05	2.00	1.60	1.85	0.75	2.40	1.10	4.15	1.75
2.70	1.05	3.90	1.25	4.00	1.25	3.90	3.20	3.70	1.80	3.40	2.30	3.40	2.40	2.50	1.10	2.30	1.40	2.90	1.70	1.30	1.00	2.00	1.70	2.50	1.10	3.90	3.30
2.70	1.20	1.60	0.90	3.60	0.90	2.50	1.00	1.60	0.90	2.70	1.50	3.10	1.20	1.90	0.70	3.30	1.40	2.00	1.20	1.80	1.00	2.30	1.30	3.30	1.10	1.70	1.60
2.60	1.90	4.70	1.20	3.50	1.70	1.60	0.90	2.80	2.10	2.10	0.70	2.10	0.90	2.80	1.00	1.90	0.80	4.00	1.20	4.30	2.10	4.20	1.50	2.60	1.40	3.90	1.80
3.40	1.80	2.60	1.50	4.10	2.70	3.10	1.50	2.40	0.70	1.95	0.70	1.90	1.25	2.20	1.25	2.50	1.30	2.35	1.25	2.00	0.75	3.20	1.80	2.15	1.10	4.15	2.80
4.50	3.55	5.00	4.15	3.35	1.05	2.20	1.25	2.50	0.60	2.55	1.00	1.60	0.60	2.40	1.20	1.70	1.15	3.15	1.40	2.60	1.10	4.10	2.60	3.00	1.00	3.35	2.40
2.95	1.20	2.95	2.40	2.40	1.50	3.40	1.40	3.30	1.55	3.80	1.90	3.80	2.30	3.40	0.90	4.00	2.65	3.40	1.45	3.80	1.05	4.20	3.50	2.05	1.35	2.80	2.35
3.15	1.25	4.20	2.10	4.20	2.20	3.40	1.25	2.15	0.70	2.95	2.10	3.00	2.10	3.50	1.80	3.25	1.15	4.30	2.00	2.30	0.95	3.60	2.10	3.30	1.20	4.60	2.75
2.30	0.80	2.30	1.80	3.50	1.55	3.30	1.25	2.30	1.20	1.60	1.05	3.50	1.80	2.90	1.90	2.85	1.55	3.60	2.60	1.65	1.30	1.25	0.90	2.95	1.40	3.10	1.20
1.80	1.20	2.35	1.35	1.65	0.75	1.40	0.80	1.55	0.75	3.15	2.90	3.25	1.35	2.70	1.10	1.65	1.05	3.25	1.10	2.85	1.50	2.35	1.15	2.15	1.85	2.65	0.90
1.80	0.95	3.00	1.80	1.70	0.70	3.25	2.30	2.80	1.95	3.45	1.60	2.00	1.00	2.75	2.20	1.75	1.25	2.85	1.00	2.85	1.00	2.15	1.30	2.85	1.60	3.00	1.10
1.70	1.05	2.80	2.00	3.50	1.70	2.90	1.50	3.70	2.00	3.60	1.00	3.60	1.60	2.45	1.50	3.60	2.20	3.95	2.15	2.90	1.50	1.95	1.60	3.85	1.35	2.90	1.35
1.55	1.40	3.50	1.85	2.50	1.45	4.10	2.15	2.80	1.85	3.50	1.30	3.20	1.30	2.85	2.10	2.75	2.25	3.10	1.20	4.00	2.40	3.60	1.10	3.60	1.60	3.70	2.60
2.05	0.80	3.40	1.60	2.55	0.90	2.55	1.60	2.90	1.85	3.80	2.80	1.75	1.35	2.10	1.00	2.35	1.05	3.65	2.25	3.95	0.85	2.05	1.20	2.80	1.85	2.40	1.20
2.80	1.15	2.35	1.70	3.10	1.40	4.00	2.10	2.10	1.30	2.20	1.30	4.10	2.30	2.00	0.90	4.40	1.20	3.20	1.60	3.20	0.80	3.75	3.40	2.40	1.95	4.10	1.20
4.20	2.65	3.50	2.90	3.50	1.50	2.25	0.90	3.30	1.65	2.50	1.70	3.60	1.75	2.55	2.20	3.80	1.30	3.40	2.80	3.15	2.95	2.70	1.30	2.20	1.70	2.60	1.70
2.05	0.85	2.50	2.05	2.90	1.10	3.05	1.55	3.20	1.05	3.20	1.50	2.05	1.30	1.75	1.10	2.85	1.40	2.00	0.90	1.60	1.05	2.65	2.35	3.70	1.20	2.80	0.80
3.20	1.20	1.70	0.60	1.70	0.50	3.10	1.60	2.10	0.80	3.70	1.65	3.10	2.60	3.60	2.00	2.60	1.05	2.00	1.15	3.30	2.30	3.70	2.10	2.00	0.75	2.65	1.90
2.90	1.50	2.50	1.10	3.70	2.05	3.30	1.10	3.00	1.60	3.00	1.20	4.00	1.20	3.90	1.50	3.50	2.20	2.35	0.70	3.80	1.40	3.90	2.40	4.60	1.20	3.90	1.60
2.30	1.25	3.85	1.80	3.90	1.60	2.80	1.00	3.30	1.50	3.10	1.80	3.00	1.30	2.85	1.20	2.90	1.40	2.90	1.60	2.70	1.50	1.80	1.20	1.65	1.00	2.20	1.70
4.50	1.50	2.30	1.50	1.90	1.05	2.30	1.10	3.80	1.00	3.60	3.00	2.80	1.10	2.90	1.60	2.70	1.55	4.00	0.80	2.30	1.40	2.60	1.30	2.00	1.30	3.05	1.40
3.05	2.30	2.25	1.45	3.30	2.20	2.70	1.60	2.00	0.85	1.95	0.90	2.90	0.75	1.20	0.30	2.55	1.20	4.80	1.50	4.70	1.40	2.45	1.15	1.95	1.00	1.90	1.00
2.15	1.30	3.10	1.15	2.80	1.65	2.90	1.25	1.85	1.05	2.20	0.80	2.35	1.15	2.25	1.20	2.10	0.70	2.45	1.80	2.70	0.90	1.55	1.10	2.05	0.70	2.90	1.20
2.50	1.75	2.80	1.85	3.00	0.80	2.70	1.30	1.70	1.20	3.60	1.60	3.80	1.40	3.90	2.90	3.70	2.20	4.60	2.90	3.60	2.20	2.50	0.70	2.50	1.10	2.45	1.80
2.05	1.65	2.65	0.95	2.60	1.75	3.00	1.20	2.45	0.90	4.50	2.15	2.55	2.45	3.30	1.90	2.60	0.95	2.85	1.50	3.10	2.00	4.00	2.10	3.30	1.90	1.40	1.15
2.20	1.45	3.00	1.70	1.50	0.85	1.60	1.35	2.75	1.85	2.50	1.30	3.70	2.10	2.30	1.90	2.35	1.20	1.90	1.35	1.95	0.95	2.00	1.10	3.50	3.20	5.00	1.70
3.50	1.10	1.80	1.05	3.00	1.60	3.70	1.70	3.50	1.40	3.40	1.75	2.00	1.30	2.00	0.60	3.10	1.50	1.80	1.50	2.60	1.20	2.65	1.70	1.80	1.10	1.85	0.80
2.50	1.25	1.80	1.20	3.30	1.30	2.30	1.70	2.95	1.40	2.70	2.40	3.10	1.30	1.65	0.75	1.20	0.95	4.35	2.50	2.50	1.00	3.35	1.75	3.65	1.05	1.90	1.00
1.85	1.05	1.50	0.80	4.35	3.00	2.00	1.15	2.35	1.00	3.80	1.40	3.60	2.80	2.85	2.65	2.90	2.70	3.30	1.30	3.70	1.30	3.60	1.20	1.45	1.00	2.55	1.35
2.15	1.00	2.40	1.10	2.80	1.00	3.50	1.00	5.00	3.70	2.80	1.80	2.30	1.10	2.40	2.20	3.20	2.00	3.40	2.10	3.10	1.00	2.10	1.30	3.70	2.10	2.80	1.30
2.00	0.80	4.00	1.15	2.30	1.70	2.10	1.50	3.90	2.30	2.70	1.30	2.55	1.10	2.60	1.70	1.55	0.95	5.20	4.00	2.65	1.00	1.75	0.95	5.10	3.00	3.60	2.45
2.00	1.25	2.75	1.80	3.20	1.55	2.20	0.80	3.95	3.25	2.60	2.05	2.95	0.70	2.85	1.45	3.60	3.10	3.05	1.50	2.85	2.00	3.65	2.30	3.60	2.35	2.95	1.70
2.75	0.80	3.15	2.15	3.15	1.90	2.55	1.70	3.20	1.90	2.90	1.00	2.35	1.25	2.00	0.95	1.95	0.90	2.85	1.40	2.60	0.90	2.85	1.60	2.45	1.10	2.20	1.15
3.40	1.00	3.60	1.20	2.90	1.05	3.70	1.55	4.10	2.00	3.70	2.20	3.40	3.25	2.50	0.80	3.80	1.65	2.90	0.95	1.50	0.85	3.70	2.00	3.10	1.65	3.00	1.20
4.65	3.90	3.90	2.85	4.40	2.60	3.60	2.50	4.60	4.00	4.25	3.30	4.20	3.00	3.10	1.10	3.40	1.80	2.50	1.05	2.05	1.20	4.40	2.00	1.60	0.80	2.40	1.00
4.80	3.80	2.00	1.30	3.20	2.10	3.15	1.70	4.40	2.40	2.50	1.80	3.05	1.20	3.20	1.50	2.70	2.60	2.80	0.80	2.35	1.10	2.60	1.35	2.80	1.20	3.85	2.20
3.20	1.80	3.50	2.25	4.40	3.50	3.70	3.10	3.75	3.50	1.90	1.60	3.70	2.30	2.40	1.20	3.15	1.60	2.25	1.60	4.80	2.40	2.65	1.50	4.20	2.10	3.95	1.20
2.55	1.05	4.15	1.60	2.85	1.20	2.80	1.10	3.20	2.10	2.40	1.15	3.70	2.20	2.10	1.00	4.10	1.15	3.15	0.90	2.15	0.90	3.00	1.40	2.20	1.30	1.90	0.80
2.30	0.90	1.45	0.90	2.35	1.20	1.40	0.80	4.00	1.50	2.00	1.05	1.95	1.25	2.55	1.80	4.10	2.10	4.30	3.60	3.50	2.20	3.00	1.80	3.00	1.20	4.00	2.70
2.60	1.90	1.90	1.80	2.50	1.05	2.70	1.30	1.70	0.90	2.50	1.60	3.35	1.25	2.15	1.10	1.80	1.60	2.30	2.20	2.65	1.10	3.15	1.00	3.50	1.80	2.60	1.85
3.90	2.10	3.70	2.05	3.65	1.80	3.00	0.95	3.30	1.55	2.80	1.50	3.40	2.35	2.80	1.60	3.00	1.45	3.10	1.20	3.50	1.60	3.05	1.80	2.55	1.10	3.00	1.45
3.05	1.80	3.60	1.40	1.85	1.15	2.50	1.30	1.50	1.30	3.25	2.45	2.10	0.50	1.90	1.30	1.85	0.90	3.60	3.25	3.20	1.20	2.50	1.60	3.30	0.90	2.90	1.85
4.40	1.90	2.50	1.20	3.30	1.30	2.70	0.95	1.95	1.35	1.95	1.30	2.50	1.00	2.90	1.00	4.30	1.10	2.00	1.00	3.50	2.60	2.55	1.30	3.15	1.40	2.75	1.45
2.75	1.35	4.50	1.90	3.60	1.75	3.20	1.10	3.05	0.95	4.80	1.90	2.65	1.15	2.35	1.55	3.95	2.30	3.60	1.60	2.00	1.20	3.50	2.65	1.45	1.10	3.90	2.90
1.35	0.60	2.20	1.70	2.30	1.35	1.95	0.60	3.95	2.20	3.00	2.10	2.85	1.05	3.50	1.40	3.30	1.75	2.90	1.50	2.95	2.70	3.20	1.90	3.30	1.40	2.90	2.80
3.30	1.40	4.00	2.35	2.05	0.85	2.20	1.30	3.60	1.85	2.30	1.30	2.00	0.85	1.70	0.90	2.50	1.15	2.70	1.80	3.05	0.65	2.10	1.15	2.50	0.80	1.30	1.10
2.10	0.85	2.95	2.30	2.65	1.60	2.25	1.50	1.70	0.85	2.65	0.90	3.20	2.10	1.40	0.75	2.55	1.10	2.75	1.00	3.35	1.80	4.20	1.95	2.10	1.10	2.75	1.80
2.60	0.80	2.20	0.95	1.90	1.10	3.00	2.00	2.00	1.00	2.90	2.30	2.00	0.75	3.95	2.95	2.50	0.90	3.30	2.30	2.70	1.00	1.35	0.85	2.55	2.00	2.80	1.00
2.30	1.20	2.60	1.00	2.30	1.40	2.45	0.90	3.10	1.00	1.95	1.00	2.40	1.20	2.90	2.60	3.50	3.10	3.65	1.50	4.10	1.70	3.65	2.60	3.00	1.80	3.00	0.70
4.05	2.15	4.70	0.80	2.55	1.60	2.35	1.30	4.20	2.10	3.50	2.10	3.70	0.85	3.70	2.40	4.70	3.00	4.90	3.15	2.60	1.25	4.40	3.00	4.50	3.35	4.40	2.40
4.10	1.70	4.40	2.00	4.00	1.35	4.50	2.65	2.55	0.85	1.65	1.20	4.35	1.20	4.00	2.00	3.50	2.45	3.30	2.20	2.00	0.55	4.60	1.15	4.60	2.20	2.40	1.30
4.30	1.40	4.50	3.00	2.40	0.80	2.80	1.70	3.05	1.90	3.60	2.30	2.60	2.15	2.75	1.50	2.40	2.00	2.20	1.30	2.70	1.40	3.50	1.70	3.45	1.40	4.05	2.40
2.05	1.10	2.40	1.40	2.15	1.20	2.15	1.10	3.40	2.50	3.90	2.50	2.50	1.30	3.00	1.40	3.40	1.90	2.65	1.35	2.50	1.20	3.00	1.90	1.85	0.90	3.20	1.40
2.20	0.65	3.10	0.65	1.80	1.15	1.55	1.25	3.15	1.70	2.00	1.25	3.00	2.00	3.15	2.20	3.15	2.10	4.55	1.65	1.85	1.35	1.75	1.00	4.15	1.70	3.15	1.85
3.40	1.70	2.20	1.50	1.60	1.00	2.55	0.70	3.10	1.60	2.55	1.55	3.80	1.20	3.45	2.45	3.35	1.05	3.15	2.15	3.20	1.90	2.15	0.95	2.70	0.90	2.85	1.70
2.70	1.30	4.35	2.80	3.00	1.40	2.15	1.40	2.30	0.70	3.60	1.65	4.25	2.50	4.00	3.30	1.20	1.00	3.20	2.50	1.65	1.50	4.30	2.20	1.70	1.45	3.10	1.80
4.10	2.70	2.10	2.00	2.70	1.60	3.00	1.50	4.30	2.35	1.85	1.40	2.40	1.70	2.30	1.40	1.15	0.80	2.50	1.50	2.50	1.55	3.90	2.25	3.20	2.45	4.20	2.25
2.80	2.20	3.80	2.15	4.50	3.50	2.70	1.40	2.90	1.30	1.20	0.95	2.45	1.85	2.55	1.95	2.25	1.35	3.30	2.80	3.50	2.90	3.15	2.60	2.60	1.30	2.80	2.10
3.50	3.25	2.70	2.00	3.80	3.00	2.60	1.10	2.60	0.90	2.70	1.55	2.15	1.50	3.00	0.95	2.80	1.45	2.95	2.05	2.65	1.20	3.90	2.20	3.20	2.10	2.45	1.30
3.40	2.00	2.95	1.60	2.05	1.70	2.25	1.45	2.90	1.65	2.80	2.55	3.30	2.20	3.05	1.90	2.40	1.35	2.75	1.30	3.35	2.10	3.05	2.00	3.00	1.95	2.35	1.10
5.10	3.80	3.35	1.60	2.05	1.60	2.45	1.30	3.70	1.90	3.60	2.00	3.40	2.75	0.90	0.65	2.65	1.20	3.30	1.25	3.60	0.80	3.90	2.90	4.00	2.15	2.30	1.00
1.95	0.90	3.40	2.45	3.50	2.10	3.00	2.40	1.40	1.00	2.35	1.00	2.15	1.10	3.00	1.50	3.10	1.15	3.15	1.90	3.60	2.40	5.20	1.00	3.00	1.60	3.40	2.05
2.45	2.00	2.05	1.60	3.80	1.70	3.30	1.30	2.10	1.40	5.00	0.85	4.65	1.45	4.35	1.60	1.65	0.80	2.80	1.20	2.70	1.35	3.60	2.70	1.05	1.05	2.80	2.80
4.00	2.65	1.70	1.00	2.40	1.70	3.45	2.20	2.20	1.50	2.45	1.50	2.45	1.30	3.80	2.65	3.50	0.70	2.05	1.20	2.30	1.35	3.60	1.75	3.00	2.25	3.90	1.80
3.00	1.05	2.05	0.85	2.95	1.80	3.60	1.95	2.70	1.65	3.00	1.30	2.25	1.10	1.35	0.90	2.30	1.75	2.25	1.30	3.90	2.60	3.15	1.15	1.75	0.75	2.80	1.80
3.05	1.20	3.00	2.00	3.05	1.10	2.50	1.40	2.50	1.40	4.00	1.50	3.40	2.10	3.10	1.60	3.80	2.10	2.70	1.20	2.85	1.05	2.05	1.00	3.30	1.30	1.90	1.15
1.30	0.75	2.50	0.90	2.50	0.85	2.40	1.10	3.05	2.70	2.25	1.90	1.70	1.35	3.00	2.30	3.00	2.65	2.20	1.60	2.30	1.70	2.90	1.50	2.95	1.30	3.30	2.45
3.70	2.30	2.35	1.40	3.40	2.00	2.70	2.30	2.35	1.50	3.00	1.50	3.00	1.50	2.30	1.70	2.65	1.50	3.50	2.30	2.75	2.10	3.70	2.30	3.50	2.40	2.15	1.30
2.00	1.30	2.20	1.15	2.55	1.90	3.50	2.10	2.80	1.50	2.70	1.70	3.60	2.10	2.10	1.40	2.30	1.80	2.90	1.30	2.90	2.30	3.65	2.80				

TABLE 2.1

Our object is to extract from the raw data all of the essential information contained therein for the answer to questions which may be put in attaining the object for which the data were taken. For example, we might ask the following questions: How does the depth of sapwood vary from pole to pole? What is the range of variation? What is the most frequently observed depth of sapwood, the next most frequently observed depth and so on? These answers could be answered directly from the raw data given in Table 2.1 although to do so would be quite laborious. Such a vast array of figures in this form cannot readily be used.

Now we might ask another question. Does the depth of penetration depend upon the depth of sapwood? The unordered array of figures of Table 2.1 would be quite incomprehensible to the average individual when it comes to answering this question. Hence we must consider the available methods of analysis for reducing such a series of observations to some form in which they may be used more effectively to serve the purpose for which they were taken. This necessity is particularly marked when we face the problem of comparing several sets of observations such as the one presented in Table 2.1 for the purpose of determining whether or not these sets are significantly different one from another.

3. Methods of Analysis

There are certain more or less definite methods of analysis which are available for interpreting an original series of raw data. We shall consider briefly methods for presenting such data in both tabular and graphical forms which assist materially in helping us to grasp the significance of the original series of observations. We shall find, however, that the results obtained in this way are for the most part qualitative and in this particular, they do not serve effectively for the comparison of sets of data. To secure quantitative reduction of data we must therefore introduce methods for summarizing a series of values of the quality X by means of a few simple functions which express quantitatively such things as central tendency, dispersion and skewness of a distribution of values such as that of either the depth of sapwood or the depth of penetration presented in Table 2.1. In particular we need quantitative measures for the correlation between two or more qualities such as the two given in Table 2.1. We shall find moreover, that there are many ways for carrying out the details of such analyses and that there are many functions which may be used to measure such characteristics as central tendency and skewness, although some of these are far more effective than others.

4. Essential Information Defined

The fact that there is more than one function which would measure the characteristic central tendency of a distribution of data leads us to consider a basis for deciding upon the particular function to be used in a given case. We shall try to indicate ways and means of choosing that function which in a given case contains the essential information inherent in the original set of data for the purpose of obtaining the object for which the data were taken. We shall understand that the essential information, to the best of our knowledge in the light of available methods of analysis, answers the questions for which the data were taken so that whatever further analysis is made it will not add information sufficient to change to a practical extent the conclusions derived from the study of the data.

Obviously an analysis of the set of data should provide the essential information in the most useful form. These ideas in mind, we proceed to a consideration of the methods of analyzing data.

CHAPTER II

Presentation of Data by Tables and Graphs

1. General Problem

There are only a few forms in which the raw data with which we have to deal usually occur. For example, we may have a series of n observations of the quality of a single thing, such as n observations of the length of a rod, the resistance of a relay or the capacity of a condenser. In a similar way, we may have a series of n observations representing single observations of some quality characteristic on n different things, such, for example, as the 1370 observations of the depth of sapwood previously given in Table 2.1. We have, let us say, n values,

$$X_1, X_2, \ldots X_i, \ldots X_n, \qquad (2.1)$$

representing, in the one case, n measurements of the same quality on a single thing and, in the other case, single measurements of the same quality on n things.

In a similar way, we may have a series of observations representing n successively observed values of a group of m quality characteristics on some one thing or observed values of say m different qualities on each of let us say n things. In either case we have a series of observations, such as,

$$\begin{array}{l} X_{11}, X_{12}, \ldots X_{1i}, \ldots X_{1n} \\ X_{21}, X_{22}, \ldots X_{2i}, \ldots X_{2n} \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ X_{j1}, X_{j2}, \ldots X_{ji}, \ldots X_{jn} \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ X_{m1}, X_{m2}, \ldots X_{mi}, \ldots X_{mn}\,. \end{array} \qquad (2.2)$$

Naturally we always have a certain purpose in accumulating such a series of data and the object of tabular and graphical presentation is to assist in the interpretation of the raw data in terms of the object for which they were taken. As already noted, the distribution of values of depth of sapwood and also that of the depth of penetration as given in Table 2.1 illustrate the first form in which raw data may occur. Similarly, the two distributions taken together illustrate the second form in which raw data may occur. Particularly when we have such a large number of observations, it is very difficult to grasp the significance of the original data. This difficulty may be partially overcome by tabular and graphical presentation.

In general, perhaps the most useful form of table for presenting a single series of observations is that in which the original raw data are arranged or permuted in ascending order of magnitude. In a similar way, a set of observations measuring qualities on several things may be arranged in tabular form by permuting the series of observations in ascending order of magnitude in respect to one of the m quality characteristics and then tabulating the values of the associated characteristics accordingly.

Graphical presentation consists of the representation of the permutation of a single series of n observations arranged in ascending order of magnitude or of the permuted series of values in respect to one characteristic arranged in ascending order of magnitude together with the associated values for the other characteristics.

We can best make this point clear by considering several examples. We shall find in so doing that it is practical to present the original data in tabular or graphical form only when the number of observations is small. We shall also find that the data presented in tabular or graphical form lead for the most part only to qualitative conclusions.

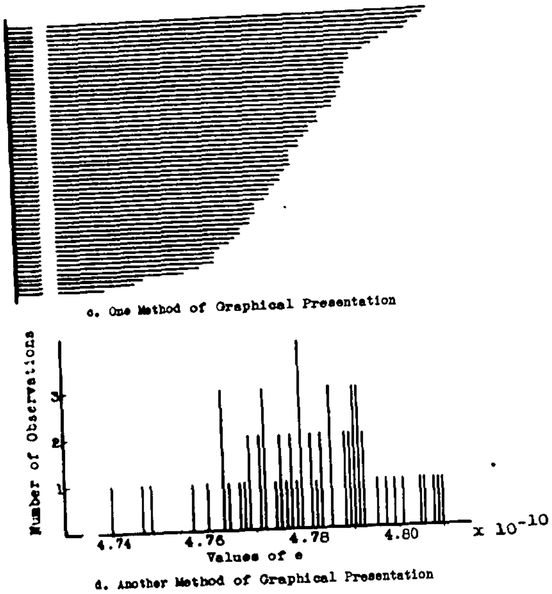
2. A Simple Illustration

In Part I we referred to the important problem of correcting data for errors of measurement and made use of Millikan's measurements of the charge on an electron as a typical set of data showing the effects of uncontrolled causes of variation in the observed data. In Fig. 2.1a we have the original set of raw data as presented by Millikan.[1] In "b" of this figure we have this original series of data arranged in ascending order of magnitude. We see at once how much easier it is to picture such things as the range of observed variation and the central tendency of the observed set of data by means of the permuted series of data than it is by the original series. In "c" and "d" of this same figure we

[1] The original data were given in terms of $e^{2/3}$ where e is the charge on an electron whereas this table of data has been given in terms of e.

have two of an indefinitely large number of possible graphical presentations of the observed results. In the first of these the lengthsof the lines are proportional to the observed charge on an electron. With but little difficulty we get from this figure a fairly good picture of the range in variation in size of the observed values of this charge but again we do not get any definite basis for quantitatively summarizing the results. Perhaps the more usual form of representing such permuted series of values is that given in Fig. 2.1d.

a. Observed Data	b. Permuted Data
4.781	4.740
4.795	4.747
4.769	4.749
4.792	4.758
4.779	4.761
4.775	4.764
4.772	4.764
4.791	4.764
4.782	4.765
4.767	4.767
4.764	4.768
4.776	4.769
4.771	4.769
4.789	4.771
4.772	4.771
4.789	4.772
4.764	4.772
4.774	4.772
4.776	4.774
4.791	4.775
4.777	4.775
4.765	4.776
4.785	4.777
4.805	4.777
4.788	4.778
4.801	4.779
4.785	4.779
4.782	4.779
4.808	4.779
4.771	4.781
4.809	4.781
4.790	4.782
4.779	4.783
4.788	4.783
4.778	4.785
4.791	4.785
4.768	4.785
4.783	4.788
4.740	4.788
4.775	4.789
4.761	4.789
4.792	4.790
4.758	4.790
4.764	4.790
4.810	4.791
4.799	4.791
4.779	4.791
4.797	4.792
4.790	4.792
4.747	4.795
4.769	4.797
4.806	4.799
4.779	4.801
4.785	4.805
4.790	4.806
4.777	4.808
4.749	4.809
4.781	4.810



c. One Method of Graphical Presentation

d. Another Method of Graphical Presentation

FIG. 2.1 - MILLIKAN'S DATA FOR THE CHARGE ON AN ELECTRON

3. Measurements of Relationship between Qualities

Let us consider two simple types of data representing relationships between quality characteristics. One is illustrated by the series of data presented in Table 2.2, showing the observed current I in amperes through a carbon contact at different voltages E. At a glance we see that the current increases with voltage although we cannot see so easily whether or not the rate of increase is constant. Now we may represent this series of pairs of observations graphically as in Fig. 2.2 and thereby show that there is a possible parabolic functional relationship between the current through

Voltage E in Volts	Current I in Amperes
3	.03
6	.07
9	.11
12	.15
15	.19
18	.24
21	.29
24	.34
27	.39
30	.45
33	.50
36	.55
29	.62
32	.69
45	.76
48	.86
51	.93

TABLE 2.2

the carbon contact and the voltage across it over the range of values given in the figure and hence that the rate of increase in current with voltage is not constant. Here the graphical presentation has an advantage over the tabular.

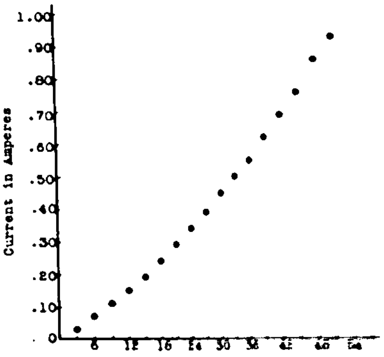


FIG. 2.2

If, however, we were to take another carbon contact and perform the same experiment, we would not in general obtain a series of points which would fit in with the series given above. As an illustration of this we introduce Fig. 2.3 which shows the same series of points as that presented in Fig. 2.2 together with a similar series of points observed for another carbon contact. That the relationship between these two characteristics for one carbon contact is different from that for another is obvious from the figure but how much these relationships differ is not so easily expressible either by tables of the observed values or by their graphical representation.

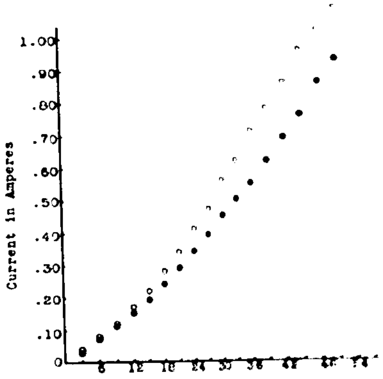


FIG. 2.3

Now let us introduce a problem of the second type. Again we shall consider two quality characteristics of granular carbon. Table 2.3 gives the measurements of the volumes of the pores and the surface areas of twenty-three different samples of carbon, the volumes being arranged in ascending order of magnitude. Do the associated values bear any definite relationship? This is equivalent to choosing the volume X as the independent variable. One form of graphical presentation of these results is given in Fig. 2.4.

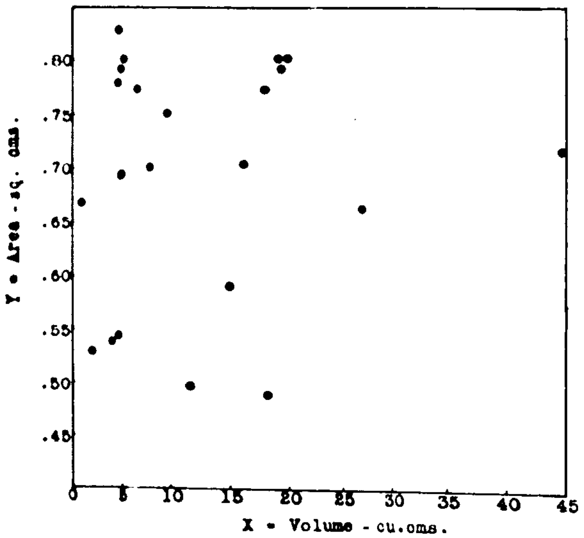
Certainly we cannot arrive at any definite conclusion from the tabular or graphical representation of these data as to whether or not there is a definite relationship between the two series of observations. We need a more refined method of analysis than that given by tables and graphs to measure this

Volume cu.cms. X	Area sq.cms. Y
.9	.667
1.9	.528
3.9	.538
4.5	.778
4.6	.827
4.6	.543
4.8	.792
4.9	.694
4.9	.694
5.1	.804
6.6	.772
7.8	.706
9.6	.750
11.7	.496
14.9	.591
16.2	.716
17.9	.771
18.2	.489
19.0	.811
19.2	.792
19.8	.803
26.8	.664
44.8	.718

TABLE 2.3

relationship. A similar series of pairs of values is given in Table 2.1 but to present this series in either tabular or graphical form such as that just introduced would be prohibitive for several obvious reasons.

It is evident that this same general procedure may be followed in presenting more complicated sets of data where the independent variable may be related to more than one other variable. In any such case, however, one of the first steps is to permute the independent variable and to observe the change in the dependent variable with this given permutation. The graphical presentation of the results involves the presentation of such a permutation and the series of observations associated with this permutation. The simple type of problem illustrated by the current vs. voltage across the contact is but a limiting case of the more complicated one illustrated by the relationship between the volume and area of granular carbon.



Regression Y on X

FIG. 2.4

4. <u>Frequency Distribution</u>

As previously stated, a simple way of picturing either of the series of 1370 observations presented in Table 2.1 is to arrange them in ascending order of magnitude. Such an arrangement or permutation is termed a <u>frequency distribution</u>. From such a tabular arrangement we can obtain with but little effort such characteristics of the distribution as <u>observed range</u> and <u>modal</u> or most frequently occurring value. However, this form of tabular presentation requires the same amount of space as does that for the same data in Table 2.1 and hence is not desirable.

To reduce the amount of space required we can represent this set of

1370 observations by a series of as many points on a straight line but some of the points would lie one on top of another and others would lie so close together as not to be easily distinguishable. Graphically, it is not feasible, therefore, to represent such a set of values in this way. Obviously, neither this simple tabular arrangement nor its graphical representation is satisfactory. We must look further.

To avoid these difficulties we customarily divide the range covered by the observations into something like thirteen to twenty equal intervals or cells, the boundaries of the intervals being so chosen that no observed point coincides therewith,thus avoiding uncertainty as to which cell a given value of X belongs. The number of things (in our case telephone poles) having a quality X lying within a given cell is termed the <u>frequency</u> for that cell, and, in a similar way, the ratio of the frequency of occurrence of a given value of X to the total number n of observations is a <u>relative frequency</u>. The series of relative frequencies constitutes a <u>relative frequency distribution</u>. The frequency distribution of depth of sapwood can in this way be reduced to the tabular form shown in Table 2.4. By this simple device of grouping the original observations into cells, we secure a tabular representation which is much simpler than that originally presented in Table 2.1 but at the same time we have slightly modified the original data. For example, we can no longer determine exactly from the table the observed range. We can, however, get a better picture of the clustering of the observed values about a central value somewhere near the cell whose midpoint is 2.8". Even though Table 2.4 does not present all of the original results in detail, let us for the moment be satisfied in seeing what we have gained not alone, as already indicated, by reducing the amount of space required and by indicating more clearly the nature of the distribution but also by making it possible to present the results more readily in graphical form.

Cell Midpoints in inches	Freq.
1.0	2
1.3	29
1.6	62
1.9	106
2.2	153
2.5	186
2.8	193
3.1	188
3.4	151
3.7	123
4.0	82
4.3	48
4.6	27
4.9	14
5.2	5
5.5	1

TABLE 2.4
Distribution of
Depth of Sapwood

Some of the forms of presentation are shown in Fig. 2.5. In the first of these the black dots represent ordinates proportional to the cell frequency, the ordinate for a given cell being placed at the midpoint of that cell. If we

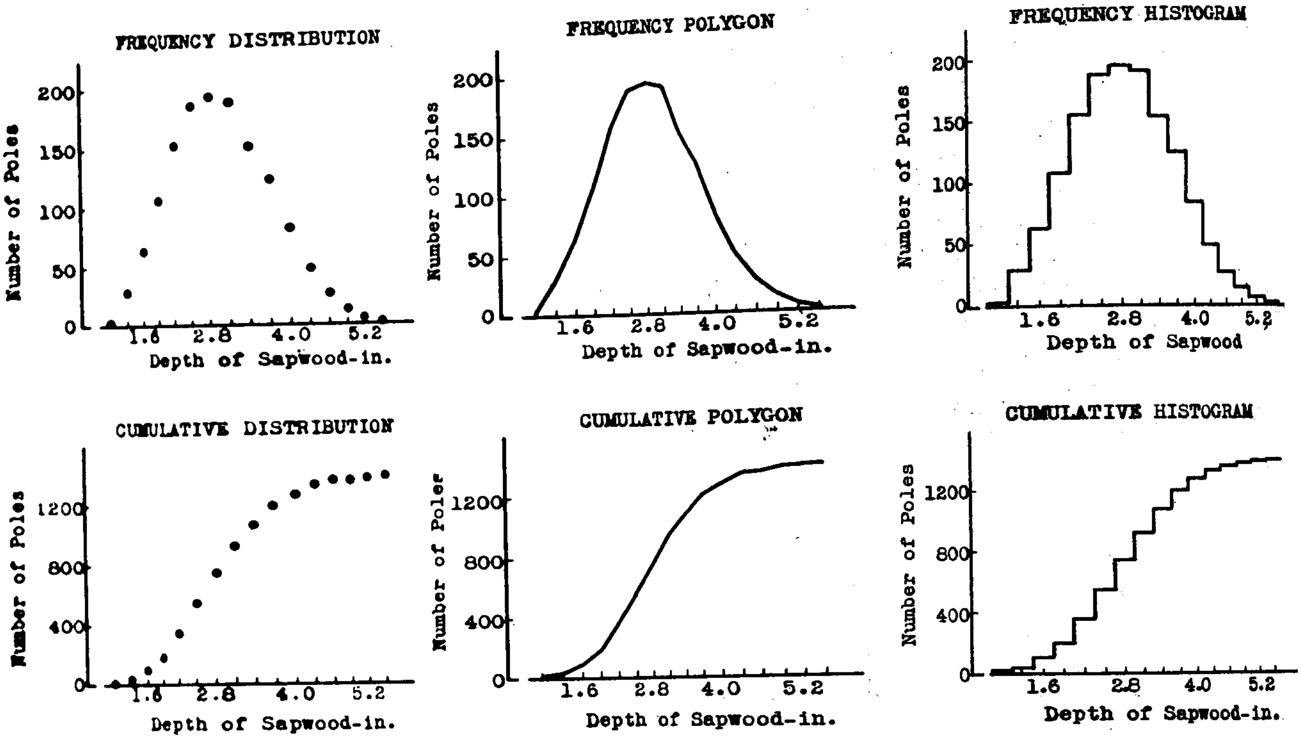


FIG. 2.5 - GRAPHICAL REPRESENTATION OF FREQUENCY DISTRIBUTION OF DEPTH OF SAPWOOD OF TELEPHONE POLES

join these points by a broken line, we obtain the frequency polygon. The method of obtaining the frequency histogram is clearly indicated by the figure itself. An ordinate in such graphical representations is termed a "frequency", meaning thereby the frequency of occurrence in the associated cell. However, this term often leads to confusion particularly in physics and engineering where it is so often used in another sense. Hence we shall adopt the practice of calling the ordinates "Number of _____" such as, in this case, "Number of Poles".

Instead of plotting the number in a given cell as an ordinate we may plot as the ordinate at a given value of abscissa the total number of observations having a value equal to or less than that of the given value of abscissa. In this way we get the cumulative distribution, cumulative polygon and cumulative histogram also shown in Fig. 2.5. These are often termed ogives. The suggested form for the title of such a cumulative chart is "Number of _____ having Quality X less than a Given Value". It is perhaps a matter of personal judgment depending upon the situation in hand as to whether the tabular or the graphical presentation of the frequency distribution of Table 2.4 is the more desirable.

Let us next try to present the data of Table 2.1 in such a way as to indicate whether or not there is any relationship between the two quality characteristics, depth of sapwood X and depth of penetration Y. In general, applying the same methods as those used above to obtain the reduced frequency distribution, we can obtain the correlation table or scatter diagram shown in Fig. 2.6. The number of poles having the depth of sapwood and depth of penetration represented by the midpoint of the rectangle is printed in the rectangle.

Y = Depth of Penetration in Inches

Σ	2	29	62	106	153	186	193	188	151	123	82	48	27	14	5	1	1250
4.6															1		1
4.3														1			1
4.0													2		1		3
3.7											1	1	1	2	1		6
3.4									2	4	5	3	2				16
3.1								1	2	4	2	5	3	1	1		19
2.8							2	7	12	7	14	4	3	2			51
2.5						2	12	6	11	11	12	7	1	2		1	65
2.2					2	7	19	24	27	29	15	6	3	1			133
1.9				2	10	19	22	36	22	44	9	7	3	1			183
1.6			5	14	39	34	45	42	29	21	10	7	3	3			252
1.3		1	11	36	48	59	51	40	29	13	10	6	3				304
1.0	1	12	33	41	42	50	37	22	15	10	2	4		1	1		271
.7	1	15	11	13	11	14	6	10	3	1	2		1				88
.4		1	2		1	1											5
	1.0	1.3	1.6	1.9	2.2	2.5	2.8	3.1	3.4	3.7	4.0	4.3	4.6	4.9	5.2	5.5	Σ

X = Depth of Sapwood in Inches

FIG. 2.6

If we were to erect a parallelpipedon from each rectangle as the base and with a height proportional to the number in this same rectangle the resulting figure would be a surface histogram, examples of which will be shown later in the discussion of the errors of an average. We might also construct surface polygons in a manner analogous to that used in constructing the frequency polygons.

What does the table or chart shown in Fig. 2.6 tell us about the relationship between the two variables therein considered? One thing is certain, - the distribution of values of penetration in a given column corresponding to a given depth of sapwood depend upon the depth of sapwood. In other words, knowing the depth of sapwood, we have some information about the depth of penetration. We shall be content, therefore, to say for the present that these two qualities appear to be correlated and that in general the depth of penetration appears to be greater, the greater the depth of sapwood. The table or chart of Fig. 2.6 does tell us something but what it tells is qualitative and not quantitative. For example, it does not tell us how close a relationship exists between the two qualities.

5. Choice of Cell Boundaries

Why do we suggest the use of from thirteen to twenty cells? This choice is to a large extent empirical. Experience has shown that when grouped in this way we appear to maintain most of the essential information contained in the original set of data. To take a larger number of cells often confuses the picture and, in particular, emphasizes sampling fluctuations, the significance of which we cannot consider until Part III. In general other things being equal, the outline of the frequency distribution will be more regular the smaller the number of cells. This is illustrated by the two frequency distributions of the data of Table 2.1 shown in Fig. 2.7.

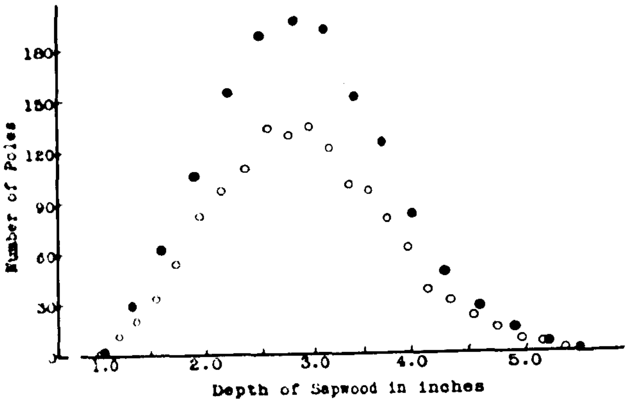


FIG. 2.7 - The Frequency Distribution of Table 2.1 Plotted with Different Cell Intervals to Show Effect of Classification on Graphical Representation

6. Further Illustrations

In Part I we introduced certain problems some of which we wish to carry through in all detail from chapter to chapter so as to illustrate all of the steps involved in arriving at practical solutions. In this chapter we shall present some of the data in tabular form which we presented in graphical form in Part I. For example, Table 2.5 gives the twelve frequency distributions for the quality characteristic, efficiency, previously shown in the polygons of Fig. 12 of Part I. In a similar way, Table 2.6 presents the original data shown in Fig. 13 of Part I.

Cell	Frequency											
Midpoints	July	Aug.	Sept.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June
-5.5		1	2	1	1		1	2	4	1		
-5.0			1			1	1	1	3	2	1	
-4.5								1	7			
-4.0				1	2			5	12	7	5	9
-3.5			3	1	1		5	10	24	19	15	24
-3.0	55	10	1	12	59	49	48	52	167	130	16	116
-2.5	141	90	119	99	157	152	137	125	221	168	146	206
-2.0	168	238	238	171	179	249	177	195	239	157	171	215
-1.5	249	265	213	312	302	359	320	330	281	241	237	322
-1.0	305	335	332	366	327	414	285	309	254	215	243	318
-0.5	231	313	238	234	161	117	162	140	134	132	150	153
0	64	46	3	3	11	9	13	27	49	100	106	79
0.5	26	2					1	2	4	24	10	8
1.0	9							1	1	4		
1.5	2											
Σ	1250	1300	1150	1200	1200	1350	1150	1200	1400	1200	1200	1450

TABLE 2.5 - Frequency Distributions for Data of Twelve Polygons of Fig. 12.

Already we have seen that the graphical presentation of these data did not make possible a quantitative comparison of the sets of observations. Now we see that the tabular presentation also fails to provide a basis for a direct and comprehensive quantitative comparison. For example, we cannot say from simply looking at the twelve frequency distributions in Table 2.5 how much these differ in respect to central tendencies and even in dispersions except for observed ranges.

In the next chapter we shall reduce these as well as other data to certain simple functions which later will be shown to contain the essential information in a more usable form than either the tabular or the graphical presentation.

n = number produced pn = number returned

Date	Company 1		Company 2		Company 3		Company 4		Company 5		Company 6		Company 7		Company 8		Company 9		Company 10	
	n	pn	n.	pn	n	pn	n	pn	n	pn	n	pn	n	pn	n	pn	n	pn	n	pn
July 18	22,291	-	24,309	1,092	24,245	744	42,592	-	5,033	-	7,104	-	14,225	1,890	43,119	2,807	20,661	1,440	23,404	1,044
21	22,715	1,780	26,037	1,007	23,804	766	44,864	1,742	5,030	103	6,961	451	13,839	1,223	44,698	1,395	20,612	1,103	21,552	932
Aug. 1	21,978	1,392	25,265	1,288	23,357	862	43,759	2,390	4,957	99	6,773	538	14,254	1,072	43,360	1,335	20,306	1,300	20,974	924
8	21,795	1,634	25,122	1,224	23,440	642	44,323	1,818	4,965	95	6,501	361	13,767	1,205	43,380	1,440	21,128	1,110	20,575	964
15	16,437	1,723	24,306	1,813	22,451	1,011	48,802	3,050	4,804	174	6,182	505	13,714	2,565	41,942	1,990	19,837	1,610	19,681	548
22	20,510	1,619	24,701	1,836	22,107	970	49,188	2,193	4,556	150	6,399	424	13,720	1,582	41,897	1,686	20,020	1,554	21,646	[illegible]
29	21,638	988	25,590	1,119	23,082	719	54,228	1,910	4,388	74	6,272	361	14,468	1,446	43,198	1,290	21,276	908	21,119	706
Sept. 5	22,710	1,184	25,284	1,209	23,080	806	52,247	2,383	7,270	41	6,590	325	14,312	1,598	43,460	1,227	21,201	1,104	20,391	843
12	20,805	1,113	25,076	1,094	22,972	959	51,304	2,164	4,102	40	6,352	417	14,459	1,842	47,176	1,430	21,102	962	21,429	734
19	20,628	1,917	23,340	1,828	22,430	151	45,750	2,938	4,559	48	6,141	586	14,367	2,305	48,416	1,256	21,550	1,008	20,752	976
26	21,962	867	5,123	941	24,642	756	49,647	1,693	4,643	95	6,476	315	13,654	1,673	54,373	1,005	23,553	843	22,067	630
Oct. 3	22,238	1,392	24,302	1,224	24,128	997	47,562	2,309	4,420	158	6,434	396	14,894	1,982	51,840	1,349	22,992	924	21,052	930
10	22,914	1,895	23,673	1,651	24,407	1,249	46,308	2,857	4,041	172	6,087	415	14,386	1,982	51,504	587	23,110	1,000	20,630	948
17	21,635	1,377	23,071	1,196	24,328	1,092	48,580	2,522	3,959	33	6,269	265	15,156	1,787	52,199	1,126	23,381	922	21,144	787
24	20,292	1,665	22,787	2,003	24,146	1,525	46,797	3,599	3,864	101	5,940	461	14,095	2,429	51,504	1,472	23,160	1,280	20,470	1,133
31	20,609	1,469	21,724	1,477	24,316	959	47,435	2,596	3,756	75	5,976	403	14,529	2,336	53,904	2,567	23,441	900	20,794	784
Nov. 7	20,810	1,601	21,616	1,315	24,270	1,015	47,225	2,259	3,794	68	6,141	333	14,017	2,394	52,609	1,726	23,005	1,060	21,076	768
14	19,996	2,071	21,281	1,507	23,887	1,144	46,464	2,634	3,773	0	5,769	421	12,817	2,055	52,842	2,163	23,169	1,088	20,801	1,097
21	19,500	1,462	21,290	1,374	24,008	1,057	47,364	2,187	3,721	63	5,844	377	13,487	2,026	53,486	1,017	22,476	1,067	20,710	838
28	18,654	822	22,471	1,310	23,900	501	48,318	-	3,584	50	5,664	300	12,538	1,002	52,552	1,421	22,333	692	20,479	609
Dec. 5	19,632	1,544	21,756	1,476	25,105	944	49,070	2,560	3,447	76	5,916	332	14,152	1,196	56,011	1,382	24,464	845	21,906	910
12	18,542	1,800	21,014	1,603	23,705	1,087	47,665	2,928	3,534	79	5,561	528	13,230	2,233	52,692	3,774	22,950	1,656	20,488	1,139
19	18,897	1,185	19,808	1,299	24,550	949	48,749	2,652	3,496	78	5,665	391	11,766	1,681	54,980	1,322	23,201	1,104	21,006	1,038
26	17,343	490	19,303	655	27,986	352	46,184	1,385	3,292	74	5,085	68	11,563	806	48,129	-	20,798	476	20,118	479
Jan. 2	17,674	688	19,333	862	23,049	563	47,220	1,862	3,015	11	5,070	292	11,084	97	48,599	1,973	20,940	1,171	19,343	817
9	19,423	970	21,940	679	24,943	694	50,001	2,211	3,311	26	5,464	331	12,953	1,531	53,515	2,657	23,746	952	21,290	630
16	19,200	1,260	21,300	1,132	25,062	831	58,816	2,632	3,278	77	5,408	421	12,114	1,512	55,955	3,388	23,257	1,160	21,069	981
23	18,748	1,501	19,897	1,323	24,125	1,107	47,123	2,531	3,189	125	5,177	421	11,675	1,482	52,009	1,480	22,707	1,148	20,160	1,008
30	18,655	1,325	21,171	1,620	24,599	1,047	48,313	2,300	3,043	94	5,241	342	11,467	1,396	55,987	1,480	22,732	1,319	19,490	1,119
Feb. 6	19,849	848	21,401	1,189	25,522	780	52,517	1,885	2,091	106	5,347	277	11,814	832	54,523	1,923	23,576	820	21,530	1,376
13	21,045	897	21,923	1,181	25,412	788	51,927	2,194	3,749	40	5,174	192	11,905	863	54,308	3,380	32,839	1,009	21,985	457
20	18,280	1,693	20,136	1,482	24,367	1,111	47,925	3,677	2,993	11	5,185	249	11,166	958	52,372	3,555	16,345	965	20,698	1,611
27	18,369	742	18,836	1,554	24,208	929	49,595	2,520	2,991	74	5,275	196	11,388	678	52,134	1,970	21,990	749	21,830	1,518
Mar. 6	17,628	1,301	19,569	1,414	23,834	1,170	48,328	634	3,067	69	5,075	333	11,149	1,638	51,726	4,329	21,708	1,379	20,141	1,917
13	17,836	1,172	19,654	1,466	24,005	1,204	48,714	583	3,033	49	4,995	285	11,036	1,104	52,797	2,575	21,776	1,076	20,568	1,379
20	18,055	1,013	19,862	1,525	25,286	1,247	48,576	535	3,040	54	4,997	290	13,222	1,336	54,976	1,720	15,090	1,144	19,547	1,244
27	17,741	992	19,663	1,500	25,408	1,337	47,385	546	2,934	62	4,905	306	16,522	1,441	50,559	2,078	5,621	375	5,382	[illegible]

TABLE 2.6 - Original Data of Fig. 13

CHAPTER III

Presentation of Data by Means of Simple Functions

1. Simple Functions to be Used

We carry over from the previous chapter two specific problems:

a. To reduce a series of n observed values,

$$X_1, X_2, \ldots X_i, \ldots X_n,$$

of some quality X to a few simple functions containing the essential information given by the original data.

b. To reduce a series of observed values,

$$\begin{array}{l} X_{11}, X_{12}, \ldots X_{1i}, \ldots X_{1n}, \\ X_{21}, X_{22}, \ldots X_{2i}, \ldots X_{2n} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ X_{j1}, X_{j2}, \ldots X_{ji}, \ldots X_{jn} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ X_{m1}, X_{m2}, \ldots X_{mi}, \ldots X_{mn}, \end{array}$$

representing n observations of m different quality characteristics to a few simple functions containing the essential information given in the original data and including measures of causal relationship between the quality characteristics. Table 2.7 presents for ready reference a list of those functions which we shall consider. Those marked by an asterisk are the more important when considered from the viewpoint of the theory of quality control and the analytical interpretation of all kinds of data. Hence they are taken up first in the discussion. Many of these functions have a graphical interpretation which cannot be fully understood until we have considered frequency curves and surfaces but we shall try to give in the present chapter an indication of the geometrical meaning of these terms adequate to give the reader a better initial picture of what the functions really mean.

2. Fraction p Defective or Non-Conforming

This simple measure of quality was described in Chapter II of Part I. In general p represents the fraction of the total number of observations lying between two specified limits. For example, the percent of stale bread returned

Fraction within certain limits	Averages for measurement of central tendency	Measures of dispersion	Measures of lopsidedness or skewness	Measures of flatness or kurtosis	Measures of relationship or correlation
*Fraction defective p	*Arithmetic Mean $\bar{X}$	*Standard deviation σ	*Skewness k	*Kurtosis β_2	*Correlation coefficient r
	$\frac{\text{Maximum+Minimum}}{2}$	Variance σ^2 .			Correlation ratio
	Median	Mean deviation			Partial correlation coefficient
	Mode	Observed range			Indefinitely large number of other measures
	Harmonic Mean	Percentiles			
	Geometric Mean	Symmetric ranges			
	Indefinitely large number of other averages	Asymmetric ranges			
		Indefinitely large number of other measures			

TABLE 2.7

is such a measure of quality.

3. Arithmetic Mean as a Measure of Central Tendency

By definition the arithmetic mean $\bar{X}$ of n real numbers X_1, X_2, ... X_i, ... X_n, is

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_i + \ldots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad (2.3)$$

Carrying through this computation for the 1370 observed values of depth of sapwood, we get 2.90" as the arithmetic mean depth. This particular measure of central tendency, as thus rigorously defined, depends upon all values of X and is easily calculated,- characteristics which do not apply to many of the other means in Table 2.7.

To calculate[1] the mean with the aid of Equation (2.3) is somewhat laborious of course and we can usually make use of an approximate method which will give a sufficient degree of accuracy. This method makes use of the grouped data as presented in the third column of Table 2.8. Obviously the mean calculated from the grouped data will not in general be equal to that given by equation 2.3 Table 2.8 illustrates the details of the method of calculating the arithmetic mean from the grouped data and in this way we secure the mean value of 2.914" instead of 2.900".

Mid-Cell Values in inches	Deviation in Cells from $\bar{o}$ X	Observed Frequency y	yX
1.0	0	2	0
1.3	1	29	29
1.6	2	62	124
1.9	3	106	318
2.2	4	153	612
2.5	5	186	930
2.8	6	193	1158
3.1	7	188	1316
3.4	8	151	1208
3.7	9	123	1107
4.0	10	82	820
4.3	11	48	528
4.6	12	27	324
4.9	13	14	182
5.2	14	5	70
5.5	15	1	15
Σ		1370	8741

$$_1\mu_1 = \frac{\Sigma\, yX}{\Sigma\, y} = \frac{8741}{1370} = 6.380292$$

m = units per cell = .3 inches

Arithmetic mean $\bar{X} = \bar{o} + m_1\mu_1 = 1.0 + 1.914088 = \underline{2.914088}$ inches

TABLE 2.8 - Illustration of Method of Estimating Arithmetic Mean.

4. The Standard Deviation as a Measure of Dispersion

Given a set of n real numbers, X_1, X_2, ... X_i, ... X_n, the standard deviation σ of this set about its mean value $\bar{X}$ is by definition

[1] It will be noted that in the calculation of the functions under consideration the numerical work is carried out to more places than will be used when it comes to interpretation. The reason for doing this will, however appear later. In general we shall find that our use of these functions in the mathematical calculations is based upon the assumption that the grouped observations are considered technically as a distribution of numbers. To do otherwise might lead to ridiculous and inaccurate relationships between the functions, particularly when it comes to the calculation of correlation coefficients and certain functions of the moments of the distributions soon to be defined.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}}$$

$$= \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n} - 2\,\frac{\bar{X}\sum_{i=1}^{n} X_i}{n} + \bar{X}^2} \qquad (2.4)$$

$$= \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n} - \bar{X}^2}$$

The exact value of σ can easily be obtained with the aid of equation 2.4 although to make the calculation in this way when the size n of the sample is large introduces a prohibitive amount of work. For this reason, as in the case of the average, we make use of the grouped data and calculate σ as indicated in Table 2.9.

Mid-Cell Values in inches	Deviation in Cells from ō X	Observed Frequency y	Xy	X^2y
1.0	0	2	0	0
1.3	1	29	29	29
1.6	2	62	124	248
1.9	3	106	318	954
2.2	4	153	612	2448
2.5	5	186	930	4650
2.8	6	193	1158	6948
3.1	7	188	1316	9212
3.4	8	151	1208	9664
3.7	9	123	1107	9963
4.0	10	82	820	8200
4.3	11	48	528	5808
4.6	12	27	324	3888
4.9	13	14	182	2366
5.2	14	5	70	980
5.5	15	1	15	225
Σ		1370	8741	65583

m = units per cell = .3 inches

$${}_1\mu_1 = \frac{\Sigma yX}{\Sigma y} = \frac{8741}{1370} = 6.380292$$

$${}_1\mu_2 = \frac{\Sigma yX^2}{\Sigma y} = \frac{65583}{1370} = 47.870803$$

$$\mu_2 = {}_1\mu_2 - {}_1\mu_1^2 = 7.162677$$

$$\sigma = m\,\mu_2^{1/2} = \underline{.798211} \text{ inches}$$

TABLE 2.9 - Illustration of the Method of Calculating the Standard Deviation.

In general, the further the set of values is spread out about the average, the larger becomes the standard deviation. For example, a small standard deviation indicates that the observed set of numbers is closely clustered about the arithmetic mean whereas a large value of standard deviation indicates that the numbers are spread out widely about the arithmetic mean. In the next chapter we shall find out how to interpret this measure of dispersion quantitatively as indicating the way in which the set of numbers is spread out about the arithmetic mean. For the time being it must suffice, however, for us to picture the significance of this measure as indicated graphically in Fig. 2.8. This figure shows two distributions differing only in standard deviation.

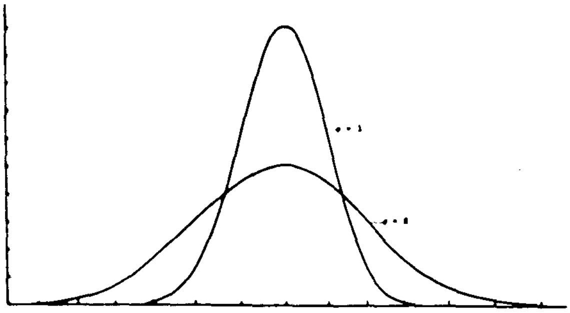


FIG. 2.8 - HOW THE STANDARD DEVIATION σ INDICATES DISPERSION. TWO DISTRIBUTIONS DIFFERING ONLY IN STANDARD DEVIATION.

5. Lopsidedness or Skewness

Given a series of numbers, we need some measure of the lack of symmetry of the distributions of the numbers of this series about the arithmetic mean. The particular function which we shall use most extensively in our study of quality control is designated by the letter k and defined by the expression

$$k = \frac{\dfrac{\sum\limits_{i=1}^{n} (X_i - \bar{X})^3}{n}}{\left[\dfrac{\sum\limits_{i=1}^{n} (X_i - \bar{X})^2}{n}\right]^{3/2}} = \frac{\dfrac{\sum\limits_{i=1}^{n} X_i^3}{n} - \dfrac{3\bar{X}\sum\limits_{i=1}^{n} X_i^2}{n} + 2\bar{X}^3}{\sigma^3} \qquad (2.5)$$

Obviously k is zero if the distribution is symmetrical. Of course, k may be either positive or negative and Fig. 2.9 shows schematically three distributions differing only in skewness

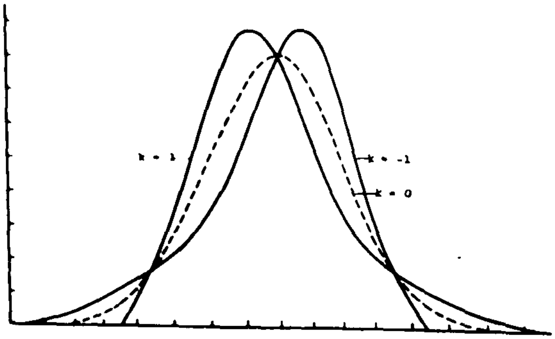


FIG. 2.9 - ILLUSTRATION OF SIGNIFICANCE OF k AS A MEASURE OF LOPSIDEDNESS OR SKEWNESS

6. Flatness or Kurtosis

The reader considering this subject for the first time may begin to wonder how many strange functions are to be introduced in the attempt to condense the essential information to usable form. Thus far we have introduced functions which measure the central tendency, skewness and dispersion of the distribution. Obviously we need to have some measure of the degree of flatness of the distribution. One such measure, β_2, technically termed kurtosis, is defined by the following relationship and its significance for the time being is illustrated in Fig. 2.10.

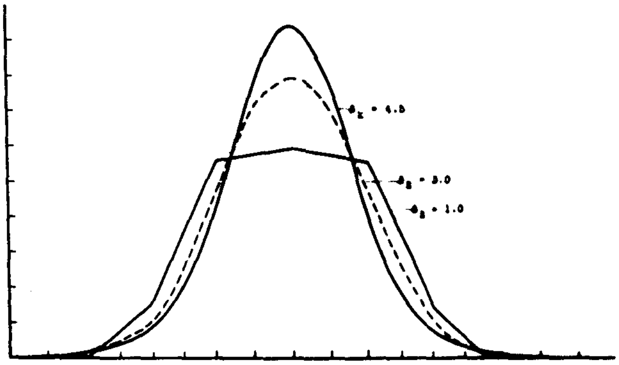


FIG. 2.10 - ILLUSTRATING USE OF β_2 AS MEASURE OF FLATNESS OF DISTRIBUTION

$$\beta_2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^4}{n} \left(\frac{n}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right)^2$$

$$= \frac{\frac{\sum_{i=1}^{n} X_i^4}{n} - 4\bar{X}\frac{\sum_{i=1}^{n} X_i^3}{n} + 6\bar{X}^2\frac{\sum_{i=1}^{n} X_i^2}{n} - 3\bar{X}^4}{\sigma^4}, \qquad (2.6)$$

7. Calculation of Functions

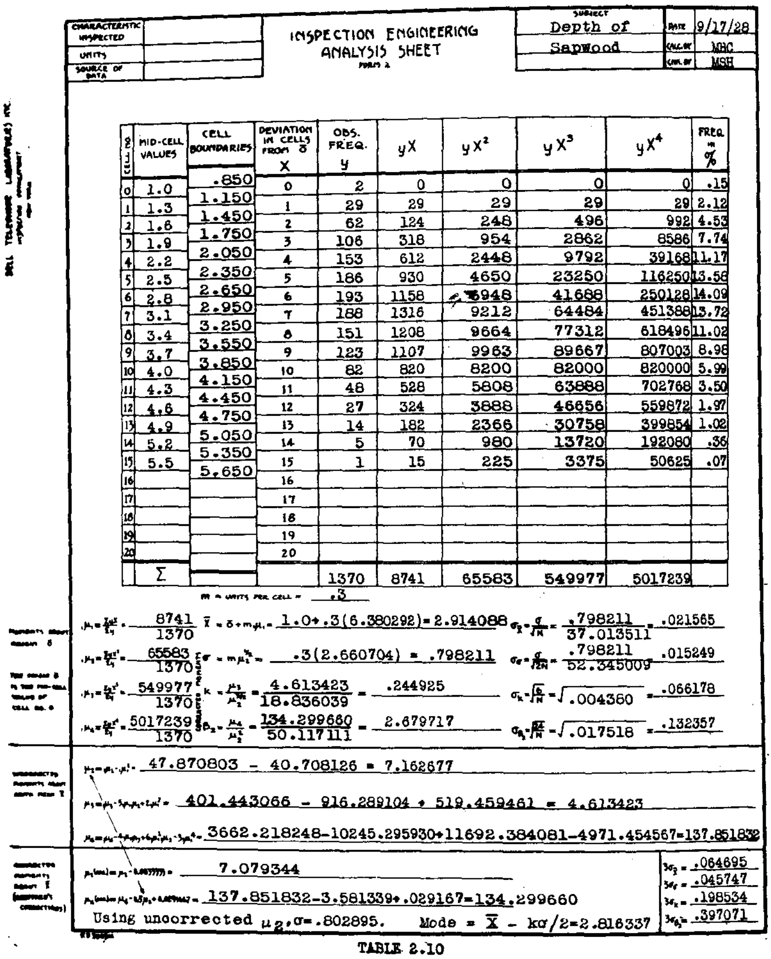
Let us see how simply the calculation of the above functions can be carried out. Again we shall use for illustrative purposes the distribution of depth of sapwood. For convenience we introduce a new term, namely, the moment of the distribution. By definition the jth moment, ${}_1\mu_j$, of a set of n values about the given origin is

$${}_1\mu_j = \frac{\sum_{i=1}^{n} X_i^j}{n} \qquad (2.7)$$

and, similarly, the jth moment of this same set of data about the arithmetic mean $\bar{X}$ is

$$\mu_j = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^j}{n}. \qquad (2.8)$$

By means of Equation (2.8) the formulas for the standard deviation, skewness and kurtosis reduce to the forms shown in the data sheet of Table 2.10. The details of the analysis involved are presented here so clearly that we need make no further comment.



CHARACTERISTIC INSPECTED		INSPECTION ENGINEERING ANALYSIS SHEET FORM 2	SUBJECT	Depth of Sapwood	DATE	9/17/28
UNITS					CALC. BY	MHC
SOURCE OF DATA					CHK. BY	MSH

CELL No.	MID-CELL VALUES	CELL BOUNDARIES	DEVIATION IN CELLS FROM δ X	OBS. FREQ. y	yX	yX^2	yX^3	yX^4	FREQ. IN %
		.850							
0	1.0	1.150	0	2	0	0	0	0	.15
1	1.3	1.450	1	29	29	29	29	29	2.12
2	1.6	1.750	2	62	124	248	496	992	4.53
3	1.9	2.050	3	106	318	954	2862	8586	7.74
4	2.2	2.350	4	153	612	2448	9792	39168	11.17
5	2.5	2.650	5	186	930	4650	23250	116250	13.58
6	2.8	2.950	6	193	1158	6948	41688	250128	14.09
7	3.1	3.250	7	188	1316	9212	64484	451388	13.72
8	3.4	3.550	8	151	1208	9664	77312	618496	11.02
9	3.7	3.850	9	123	1107	9963	89667	807003	8.98
10	4.0	4.150	10	82	820	8200	82000	820000	5.99
11	4.3	4.450	11	48	528	5808	63888	702768	3.50
12	4.6	4.750	12	27	324	3888	46656	559872	1.97
13	4.9	5.050	13	14	182	2366	30758	399854	1.02
14	5.2	5.350	14	5	70	980	13720	192080	.36
15	5.5	5.650	15	1	15	225	3375	50625	.07
16			16						
17			17						
18			18						
19			19						
20			20						
	Σ			1370	8741	65583	549977	5017239	

m = UNITS PER CELL = .3

$\mu_1' = \frac{\Sigma yX}{\Sigma y} = \frac{8741}{1370}$ $\quad \bar{X} = \delta + m\mu_1' = 1.0 + .3(6.380292) = 2.914088$ $\quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} = \frac{.798211}{37.013511} = .021565$

$\mu_2' = \frac{\Sigma yX^2}{\Sigma y} = \frac{65583}{1370}$ $\quad \sigma = m\mu_2^{1/2} = .3(2.660704) = .798211$ $\quad \sigma_\sigma = \frac{\sigma}{\sqrt{2N}} = \frac{.798211}{52.345009} = .015249$

$\mu_3' = \frac{\Sigma yX^3}{\Sigma y} = \frac{549977}{1370}$ $\quad k = \frac{\mu_3}{\mu_2^{3/2}} = \frac{4.613423}{18.836039} = .244925$ $\quad \sigma_k = \sqrt{\frac{6}{N}} = \sqrt{.004360} = .066178$

$\mu_4' = \frac{\Sigma yX^4}{\Sigma y} = \frac{5017239}{1370}$ $\quad \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{134.299660}{50.117111} = 2.679717$ $\quad \sigma_{\beta_2} = \sqrt{\frac{24}{N}} = \sqrt{.017518} = .132357$

$\mu_2 = \mu_2' - \mu_1'^2 = 47.870803 - 40.708126 = 7.162677$

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = 401.443066 - 916.289104 + 519.459461 = 4.613423$

$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 = 3662.218248 - 10245.295930 + 11692.384081 - 4971.454567 = 137.851832$

$\mu_2(\text{corr.}) = \mu_2 - .083333 = 7.079344$

$\mu_4(\text{corr.}) = \mu_4 - \tfrac{1}{2}\mu_2 + .029167 = 137.851832 - 3.581339 + .029167 = 134.299660$

Using uncorrected μ_2, $\sigma = .802895$. Mode $= \bar{X} - k\sigma/2 = 2.816337$

$3\sigma_{\bar{X}} = .064695$
$3\sigma_\sigma = .045747$
$3\sigma_k = .198534$
$3\sigma_{\beta_2} = .397071$

TABLE 2.10

8. Measure of Relationship

As engineers and scientists we are accustomed to think of two things being related when we can express one as a mathematical function of the other. For example, if a physical quality Y is related to another quality X we think of the relationship being expressible in the simple form

$$Y = f(X), \qquad (2.9)$$

meaning thereby that for every value of the independent variable X, Y is determined by the simple function f. If we again direct our attention, however, to the scatter diagram representing the observed values of depth of penetration Y and depth of sapwood X, we see that for a given value of X there are in general several values of Y. In fact, as we have already noted, there appears to be a general relationship between these two characteristics but we must discover some method of measuring this relationship.

If, for example, the depth of penetration is related to the depth of sapwood, the knowledge of the depth of sapwood should give us some information as to the depth of penetration. We must leave for later discussion a consideration of general measures of relationships of this nature. For the time being be

X = Depth of Sapwood

Y = Depth of Penetration

(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
X	Y	n_1	n_1XY	X	Y	n_1	n_1XY	X	Y	n_1	n_1XY
1.0	.7	1	.70	3.1	.7	10	21.70	4.0	3.4	5	68.00
	1.0	1	1.00		1.0	22	68.20		3.7	1	14.80
1.3	.4	1	.52		1.3	40	161.20	4.3	1.0	4	17.20
	.7	15	13.65		1.6	42	208.32		1.3	4	22.36
	1.0	12	15.60		1.9	36	212.04		1.6	7	48.16
	1.3	1	1.69		2.2	24	163.68		1.9	7	57.19
1.6	.4	2	1.28		2.5	6	46.50		2.2	6	56.76
	.7	11	12.32		2.8	7	60.76		2.5	7	75.25
	1.0	33	52.80		3.1	1	9.61		2.8	4	48.16
	1.3	11	22.88	3.4	.7	3	7.14		3.1	5	66.65
	1.6	5	12.80		1.0	15	51.00		3.4	3	43.86
1.9	.7	13	17.29		1.3	29	128.18		3.7	1	15.91
	1.0	41	77.90		1.6	28	152.32	4.6	.7	1	3.22
	1.3	36	88.92		1.9	22	142.12		1.3	3	17.94
	1.6	14	42.56		2.2	27	201.96		1.6	5	36.80
	1.9	2	7.22		2.5	11	93.50		1.9	3	26.22
2.2	.4	1	.88		2.8	12	114.24		2.2	3	30.36
	.7	11	16.94		3.1	2	21.08		2.5	1	11.50
	1.0	42	92.40		3.4	2	23.12		2.8	3	38.64
	1.3	48	137.28	3.7	.7	1	2.59		3.1	3	42.78
	1.6	39	137.28		1.0	10	37.00		3.4	2	31.28
	1.9	10	41.80		1.3	13	62.53		3.7	1	17.02
	2.2	2	9.68		1.6	21	124.32		4.0	2	36.80
2.5	.4	1	1.00		1.9	24	168.72	4.9	1.0	1	4.90
	.7	14	24.50		2.2	28	227.92		1.6	3	23.52
	1.0	50	125.00		2.5	11	101.75		1.9	1	9.31
	1.3	59	191.75		2.8	7	72.52		2.2	1	10.78
	1.6	34	136.00		3.1	4	45.88		2.5	2	24.50
	1.9	19	90.25		3.4	4	50.32		2.8	2	27.44
	2.2	7	38.50	4.0	.7	2	5.60		3.1	1	15.19
	2.5	2	12.50		1.0	2	8.00		3.7	2	36.26
2.8	.7	6	11.76		1.3	10	52.00		4.3	1	21.07
	1.0	37	103.60		1.6	10	64.00	5.2	1.0	1	5.20
	1.3	51	185.64		1.9	9	68.40		3.1	1	16.12
	1.6	45	201.60		2.2	15	132.00		3.7	1	19.24
	1.9	22	117.04		2.5	12	120.00		4.0	1	20.80
	2.2	18	110.88		2.8	14	156.80		4.6	1	23.92
	2.5	12	84.00		3.1	2	24.80	5.5	2.5	1	13.75
	2.8	2	15.68								

$n = 1370$

$\Sigma XYn_1 = 6765.77$

$\bar{X}\,\bar{Y} = 4.637654$

$\frac{\Sigma XYn_1}{n} = 4.938518$

$\sigma_X \sigma_Y = .498779$

$$r = \frac{\frac{\Sigma XYn_1}{n} - \bar{X}\,\bar{Y}}{\sigma_X \sigma_Y} = \frac{4.938518 - 4.637654}{.498779} = .603201$$

TABLE 2.11 – CALCULATION OF CORRELATION COEFFICIENT

content to accept the so-called correlation coefficient r defined in the following way.

$$r_{XY} = \frac{\frac{\sum_{i=1}^{n} X_i Y_i}{n} - \bar{X}\,\bar{Y}}{\sigma_X \sigma_Y} \qquad (2.10)$$

The method of calculating this correlation coefficient is illustrated in Table 2.11. We shall see later that the value of r must lie between +1 and -1.

9. Some Applications

Thus far we have introduced and defined certain simple functions which, when calculated for a given series of observations, will be shown to contain the essential information. If the series of observations consists of n observed values on each of m different quality characteristics, four simple functions are required to express the essential information contained in the distribution of each characteristic provided the sample size n is large and only two simple functions provided the sample size n is small in addition to $\frac{\underline{|m}}{\underline{|m-2}\;\underline{|2}}$ correlation coefficients. In each of these cases, it is understood that the sample size n is known.

We may make use of these functions in expressing the essential information contained in data previously introduced. For example, Table 2.12 presents the essential information contained in some of the previously given series of observations.

10. Other Measures of Central Tendency

In general the average of a series of values, X_1, X_2, ... X_i, X_n, is defined as a number greater than the least and less than the greatest when all of the values of X are not equal and equal to the common value of X when all of the n values of X are equal. In this case the arithmetic, harmonic and geometric means are simple examples. Obviously there are an indefinitely large number of averages or, in other words, functions satisfying the above conditions. Any one of these averages measures in a way the central tendency of a group of observations. Three averages other than the arithmetic mean are often used in engineering work. They are the median, $\frac{\text{Maximum } X + \text{Minimum } X}{2}$, and mode. If we arrange the series of values of X in their order of magnitude, the value of X below which

	July	Aug.	Sept.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June
n	1250	1300	1150	1200	1200	1350	1150	1200	1400	1200	1200	1450
$\bar{X}$	-1.298	-1.250	-1.368	-1.325	-1.504	-1.512	-1.490	-1.505	-1.765	-1.550	-1.501	-1.577
σ	.829	.672	.673	.623	.713	.638	.710	.754	.923	.985	.921	.862
k	-.009	-.439	-.785	-.770	-.541	-.490	-.573	-.717	-.353	-.093	-.191	-.192
β_2	2.729	3.287	4.854	4.208	3.143	3.025	3.673	4.566	3.331	2.637	2.390	2.519

a - Twelve Frequency Distributions of Table 2.5

n = 58

$\bar{X} = 4.780 \times 10^{-10}$ e.s.u.

$\sigma = .01497 \times 10^{-10}$ e.s.u.

b - Data Shown in Fig. 2.1

	Voltage	Current
n	17	17
$\bar{X}$	27	.42
σ	14	.27

r = .993

c - Data of Table 2.2

	Volume	Area
n	23	23
$\bar{X}$	11.9	.693
σ	9.9	.108

r = .099

d - Data of Table 2.3

	Sapwood	Penetration
n	1370	1370
$\bar{X}$	2.91	1.59
σ	.80	.62
k	.24	1.03
β_2	2.68	4.18

r = .603

e - Data of Table 2.1

	Company									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
n	19854	21701	24169	48570	3825	5786	13214	50906	21722	21390
$\bar{X}$	.0663	.0616	.0381	.0452	.0199	.0620	.1170	.0378	.0490	.0482
σ	.00177	.00163	.00123	.00094	.00226	.00317	.00280	.00085	.00147	.00150

$\bar{X} = \Sigma pn/\Sigma n$ where $p = \bar{X}$

$\sigma = \sqrt{pq/n}$ $q = 1-p$

$\bar{n}$ = average n

f - Data of Table 2.6

TABLE 2.12 - ESSENTIAL INFORMATION CONTAINED IN SOME OF THE PREVIOUSLY GIVEN SERIES OF OBSERVATIONS

there are as many values in the series as there are above, is by definition the median. In either grouped distributions or in the case where n is even the median must be determined by interpolation. The modal value of X is naturally that value of X which occurs most frequently. These three averages calculated for the series of fifty-eight observed values of the charge on an electron are

$$\text{Median charge} = 4.785 \times 10^{-10} \text{ e.s.u.}$$

$$\frac{\text{Max. Charge + Minimum Charge}}{2} = 4.775 \times 10^{-10} \text{ e.s.u.}$$

$$\text{Modal charge} = 4.779 \times 10^{-10} \text{ e.s.u.}$$

In the same way we might define and calculate any number of other mean values. Under certain conditions several of these mean values calculated for a given distribution may be identical. For example, if the distribution is unimodal and symmetrical, the arithmetic mean, median and $\frac{\text{Max. X + Min. X}}{2}$ are all equal. In general, however, as in the case just stated, these values are not the same. Therefore, it is natural that an engineer should want to know which of the mean values should be used in a given case.

Naturally the labor involved in calculating one mean value may be quite different from that in calculating another. For example, the $\frac{\text{Max. X + Min. X}}{2}$ can readily be determined almost by observation even though the number n values of X is large whereas this is not true for the arithmetic mean. The modal value of X cannot, however, so easily be determined. If one turns to any one of a number of excellent elementary treatises on the theory of statistics, he finds an extended discussion of the relative advantages of the different mean values. Such information is illuminating indeed but as engineers we are perhaps more concerned with the fundamental question as to what mean will best serve our purpose in the majority of cases arising in the problems of quality control. In succeeding chapters we must justify our acceptance of the arithmetic mean as being the most useful measure of the central tendency of a series of observations particularly when that measure is supposed to contain the essential information required in solving the problems of quality control.

11. Other Measures of Dispersion

The most important measure of dispersion other than the standard deviation is the mean deviation μ defined for the case of n values of X by the expression

$$\mu = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n} \qquad (2.11)$$

where as usual $|\quad|$ represents the absolute value of a quantity. In fact, the use of this measure is often suggested because it is usually assumed to be easier to calculate.

All values in the following table are multiplied by 10^{-10}

$\bar{X} = 4.7804655$

X	$\lvert X-\bar{X}\rvert$	X	$\lvert X-\bar{X}\rvert$	X	$\lvert X-\bar{X}\rvert$
4.740	.0404655	4.775	.0054655	4.789	.0085345
4.747	.0334655	4.776	.0044655	4.790	.0095345
4.749	.0314655	4.777	.0034655	4.790	.0095345
4.758	.0224655	4.777	.0034655	4.790	.0095345
4.761	.0194655	4.778	.0024655	4.791	.0105345
4.764	.0164655	4.779	.0014655	4.791	.0105345
4.764	.0164655	4.779	.0014655	4.791	.0105345
4.764	.0164655	4.779	.0014655	4.792	.0115345
4.765	.0154655	4.779	.0014655	4.792	.0115345
4.767	.0134655	4.781	.0005345	4.795	.0145345
4.768	.0124655	4.781	.0005345	4.797	.0165345
4.769	.0114655	4.782	.0015345	4.799	.0185345
4.769	.0114655	4.783	.0025345	4.801	.0205345
4.771	.0094655	4.783	.0025345	4.805	.0245345
4.771	.0094655	4.785	.0045345	4.806	.0255345
4.772	.0084655	4.785	.0045345	4.808	.0275345
4.772	.0084655	4.785	.0045345	4.809	.0285345
4.772	.0084655	4.788	.0075345	4.810	.0295345
4.774	.0064655	4.788	.0075345		
4.775	.0054655	4.789	.0085345	Σ	.6850000

$$\text{Mean Deviation } \mu = \frac{\Sigma |X-\bar{X}|}{n} = \frac{.68500}{58} = .01181 \text{ e.s.u.}$$

TABLE 2.13 Illustration of the Method of Calculating the Mean Deviation; Data of Figure 2.1

Table 2.13 shows the method of calculating the mean deviation for a series of observations, in this case the charge on an electron.[1] We shall see as we proceed, however, that this measure does not deserve the prominence that has already been attributed to it in this paragraph except for the fact that it is a measure often used in the theory of errors and, therefore, familiar to engineers. For this reason it is important that we show later just why this particular measure should not be used because of its inefficiency in presenting the essential information contained in the series of observations.

[1] This will be given in I.E.B. 4 constituting Part III of the complete story, "Control of Quality of Manufactured Product".

By dividing the standard deviation σ by the average $\bar{X}$ we get the coefficient of variation which expressed in percent is $\frac{100\sigma}{\bar{X}}$. This measure is often used because it indicates in a seemingly practical way the relative importance of the dispersion expressed in terms of the average. Another measure is the observed range, namely, Maximum X - Minimum X. For example, this measure is used extensively in some books[1] and often in scientific journals. It has the advantage of being easily calculated and easily understood, although we find that little else can be said for it.

Having arranged a series of n values of X in ascending order of magnitude and determined the value X_L such that the percentage S of the n values of X are less than X_L, then X_L is termed a percentile. When S = .25, S = .50 and S = .75, the corresponding values of X_L are termed the first quartile, second quartile or median and third quartile respectively. The range between any two percentiles is often used as a measure of dispersion and in particular the semi-interquartile range.

Obviously the sum of the absolute values of any given power of the deviations of the observed series of observations from the arithmetic mean can be taken as a measure of dispersion. Of this group of measures there is obviously an indefinitely large number and all of them are zero when the dispersion is zero.

In a similar way we might have an indefinitely large number of measures of dispersion in terms of either symmetrical or asymmetrical ranges.

What measure of dispersion shall we use in trying to record the essential information contained in a series of observations? Why are we justified in using the standard deviation previously suggested in this chapter? These questions must be answered as we proceed along with similar questions regarding our choice of each and every function proposed for use in the analysis of data.

For convenience in future reference we tabulate below some of the possible indefinitely large number of measures of dispersion for the series of fifty-eight observed values of the charge on an electron.

[1] See for example, "Timber, Its Strength, Seasoning and Grading", by H. S. Betts, published by McGraw-Hill, 1919.

Standard deviation = $.01497 \times 10^{-10}$ e.s.u.
Mean deviation = $.01181 \times 10^{-10}$ e.s.u.

Observed range, Max. X-Min. X = $.07000 \times 10^{-10}$ e.s.u.
Semi-interquartile range = $.01850 \times 10^{-10}$ e.s.u.

These will be used in later discussions.

CHAPTER IV

Presentation of Data by Frequency Curves

1. The Problem

Once more let us recall that there are two types of data to be presented, one of which is in the form of a series of n observed values, X_1, X_2, ... X_i, ... X_n, of some quality X and the other is a number of sets of n observations on each of m quality characteristics. For the moment let us recall that there are two purposes of taking data,- (1) The presentation of facts, and (2) the interpretation of the observed data in terms of causation. Obviously the presentation of facts requires the presentation of the original observations or else the presentation of a few simple functions which in turn may be used to reproduce the original observations. The total information contained in a series of observations or a number of series of observations is obviously that expressed in the observations themselves. The essential information, on the other hand, may or may not equal in amount the total information.

In most problems of quality control we are interested in extracting from a series of observations the essential information contained therein. It must be remembered, of course, that before we can decide in a specific case whether or not a given group of simple functions containsthe essential information we must have before us a clear cut question to be answered.

A simple illustration may serve to make this point clear. Going back to our old illustration of the tailor, if he wanted to make suits of clothes for a certain group of ten individuals, it would be necessary for him to know the measurements of each of the ten individuals. On the other hand, if he merely wanted to purchase the goods from which the suits were to be made, he would need only the average amount of cloth per suit. In the first case the essential information is the total information; in the second case, it is only that part of the total information contained in the simple function, the average. It is at once apparent, therefore, that we cannot set down any absolute criterion by which to determine whether or not a given set of functions contains the essential information, for what is essential information in the answer to one question need not be the essential information in the answer to another. We can, however, get a kind of upper limit to the amount of information contained in the simple functions derived from the set of observations by seeing how much of the total information is contained in the given group of functions. In what follows, therefore, we shall try to see just how far we can go in expressing the total information in terms of a few simple functions, it being evident of course that in most cases the essential information required is actually much less than the total information.

The method of attack will be somewhat as follows: We shall introduce the concept of frequency curve or function f(X) assumed to be such that

$$\int_a^b f(X, \lambda_1, \lambda_2, \ldots, \lambda_i, \ldots, \lambda_m)dX \qquad (2.12)$$

gives approximately the number of observed values of X lying within the range X = a to X = b, where the λ's are certain parameters to be determined from the observed data. It is obvious that these parameters must be symmetric functions of the observed set of data because the order in the original set should in no way affect the function. Now it is a well-known fact in algebra that symmetric functions can all be expressed in terms of sum functions, as they are called, where the sum functions are defined as

$$\begin{aligned} S_1 &= X_1 + X_2 + \ldots + X_i + \ldots + X_n \\ S_2 &= X_1^2 + X_2^2 + \ldots + X_i^2 + \ldots + X_n^2 \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ S_j &= X_1^j + X_2^j + \ldots + X_i^j + \ldots + X_n^j \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{aligned}$$

These particular sum functions, when divided by the number n have been given a particular name, moments. For example, $\frac{S_1}{n}$ is the average or first moment of a distribution about the origin. Similarly $\frac{S_2}{n}$ is the second moment about the origin. In this way we are led in a rational way to the particular choice of function in which we have already tried to express the information contained in a series of observations. We shall find that in general no frequency function f(X) can be found to satisfy the integral (2.12). In fact we cannot even express an upper limit to the error (other than 100%) with which any known function can be used in expressing the total number of observations lying within a given specified interval a to b.

We shall come to see that in general we cannot express the total information by means of a frequency function involving the estimates of the parameters derived from the observed data. In such cases, if we wish to express the total information contained in the series of observations, the only thing that we can do is to present the original series *in toto*.

On the other hand, we shall get acquainted with a very remarkable relationship. If we write the integral (2.12) in the form

$$\int_{\bar{X}-b}^{\bar{X}+b} f(X, \lambda_1, \lambda_2, \ldots, \lambda_i, \ldots, \lambda_m)dX,$$

indicating that the integral is to be extended over a symmetric range about the average $\bar{X}$, we find that we may express an upper bound to the error of approximation to this integral which *is absolutely independent of the original series of observations*. *In fact it is this very theorem which will form the background for many of the most important methods to be derived in the discussion of control of quality in future chapters*.

Returning now to the integral (2.12), a little consideration reveals certain difficulties lying in the way of finding the very important kind of function contained therein. Common sense dictates that the functional form representing one series of observations will, in general, be different from that representing another, unless possibly a very general type of function could be found. The importance and practical significance of this problem of searching for some almost miraculous function having the property of satisfying (2.12) has inspired numerous researches during the last three centuries and out of these extensive labors come a few very important and far-reaching results. Some very general forms of frequency function have been found which, as we shall see, come quite close to, even if they possibly do not attain the goal of expressing frequency distributions in terms of a single function involving a specified small number of parameters. On the other hand, however, these functions do not satisfy our present requirement of simplicity.

Before we venture far afield in the way of generalizations, let us acquire a few simple working principles which more than likely will help us to reduce most frequency distributions coming to our attention to simple functional forms that can be used in answering most of our practical problems.

Technically speaking, our problem of representing the information contained in a series of n observed values of a quality X is divisible into two parts,- (a) determination of the form of the function f(X), and (b) calculation of the estimates θ's of the parameters λ's in terms of the observed data.

We shall first consider two simple forms of function, one of which involves only two parameters and the other three parameters. It will be found in the one case that the estimates of the parameters will be taken as the average and standard deviation of the original set of observations and in the other case, the parameters will be expressed in terms of the average, standard deviation and skewness of the observed set of data.

2. The Normal Law Function

Let us get acquainted with perhaps the simplest useful function now on the market, as it were. Simple though it is, it helps us to grasp the significance of the average and standard deviation of a given series of observations in the way of expressing the information contained in the observations. This function is often referred to as the Normal Law although it is sometimes known as the Gaussian or as the LaPlacian Law. Formally it is expressed by the relation

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-\bar{X})^2}{2\sigma^2}}, \qquad (2.13)$$

where $\bar{X}$ and σ are the two parameters. Right now we are not going to worry over some of the proposed reasons why this normal law happens to be so useful. These may be more appropriately discussed in Part III (I.E.B.4).

Obviously this function is symmetrical about the parameter $\bar{X}$ so that $\bar{X}$ is the arithmetic mean of the normal distribution of X. Furthermore,

$$\int_{-\infty}^{+\infty} (X - \bar{X})^2 f(X) dX = \sigma$$

and hence the parameter σ is the standard deviation of the normal distribution of X.

Suppose now that having a set of n observed values of X, we make use of the average and standard deviation of this set of values by substituting them in the normal function to see how closely the integral of this function between any given limits can be depended upon to give us the approximate number of observed values of X within these limits. Of course, we may have a few doubts as to what may actually happen when we take such a bold step but first let us do it and then reason later why it so often gives so close an approximation to the truth.

As a specific illustration, let us make use of the 1370 observed values of depth of sapwood given above in Table 2.4. To obtain the values of the integral of the normal law over any given range, we may make use of Table 2.14.[1] For example, if we let $x = X - \bar{X}$, the normal law function may be written in the form

$$\Phi(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad \text{where } z = \frac{x}{\sigma} .$$

Table of Values of $F(z) = \frac{1}{\sqrt{2\pi}}\int_0^z e^{-\frac{1}{2}z^2} dz$

z	F(z)	z	F(z)	z	F(z)	z	F(z)	z	F(z)	z	F(z)	. z	F(z)	z	F(z)	z	F(z)	z	F(z)
.00	.0000	.30	.1179	.60	.2258	.90	.3160	1.20	.3850	1.50	.4332	1.80	.4641	2.10	.4822	2.40	.4918	2.70	.4966
.01	.0040	.31	.1217	.61	.2291	.91	.3186	1.21	.3869	1.51	.4345	1.81	.4649	2.11	.4826	2.41	.4920	2.71	.4967
.02	.0080	.32	.1255	.62	.2324	.92	.3212	1.22	.3888	1.52	.4358	1.82	.4656	2.12	.4830	2.42	.4923	2.72	.4968
.03	.0120	.33	.1293	.63	.2357	.93	.3238	1.23	.3907	1.53	.4370	1.83	.4664	2.13	.4834	2.43	.4925	2.73	.4969
.04	.0160	.34	.1331	.64	.2389	.94	.3264	1.24	.3925	1.54	.4382	1.84	.4671	2.14	.4838	2.44	.4927	2.74	.4970
.05	.0200	.35	.1369	.65	.2422	.95	.3290	1.25	.3944	1.55	.4395	1.85	.4679	2.15	.4842	2.45	.4929	2.75	.4970
.06	.0239	.36	.1406	.66	.2454	.96	.3315	1.26	.3962	1.56	.4406	1.86	.4686	2.16	.4846	2.46	.4931	2.76	.4971
.07	.0279	.37	.1443	.67	.2486	.97	.3340	1.27	.3980	1.57	.4418	1.87	.4693	2.17	.4850	2.47	.4933	2.77	.4972
.08	.0319	.38	.1481	.68	.2518	.98	.3365	1.28	.3997	1.58	.4430	1.88	.4700	2.18	.4854	2.48	.4935	2.78	.4973
.09	.0359	.39	.1518	.69	.2549	.99	.3389	1.29	.4015	1.59	.4441	1.89	.4706	2.19	.4858	2.49	.4936	2.79	.4974
.10	.0399	.40	.1554	.70	.2581	1.00	.3414	1.30	.4032	1.60	.4452	1.90	.4713	2.20	.4861	2.50	.4938	2.80	.4975
.11	.0438	.41	.1591	.71	.2612	1.01	.3438	1.31	.4049	1.61	.4463	1.91	.4720	2.21	.4865	2.51	.4840	2.81	.4975
.12	.0478	.42	.1628	.72	.2643	1.02	.3462	1.32	.4066	1.62	.4474	1.92	.4726	2.22	.4868	2.52	.4942	2.82	.4976
.13	.0517	.43	.1664	.73	.2673	1.03	.3485	1.33	.4083	1.63	.4485	1.93	.4732	2.23	.4872	2.53	.4943	2.83	.4977
.14	.0557	.44	.1701	.74	.2704	1.04	.3508	1.34	.4099	1.64	.4495	1.94	.4738	2.24	.4875	2.54	.4945	2.84	.4978
.15	.0596	.45	.1737	.75	.2734	1.05	.3532	1.35	.4115	1.65	.4506	1.95	.4744	2.25	.4878	2.55	.4946	2.85	.4978
.16	.0636	.46	.1773	.76	.2764	1.06	.3555	1.36	.4131	1.66	.4516	1.96	.4750	2.26	.4881	2.56	.4948	2.86	.4979
.17	.0675	.47	.1808	.77	.2794	1.07	.3577	1.37	.4147	1.67	.4526	1.97	.4756	2.27	.4884	2.57	.4949	2.87	.4980
.18	.0714	.48	.1844	.78	.2823	1.08	.3599	1.38	.4162	1.68	.4535	1.98	.4762	2.28	.4887	2.58	.4951	2.88	.4980
.19	.0754	.49	.1880	.79	.2853	1.09	.3622	1.39	.4178	1.69	.4545	1.99	.4768	2.29	.4890	2.59	.4952	2.89	.4981
.20	.0793	.50	.1915	.80	.2882	1.10	.3644	1.40	.4193	1.70	.4555	2.00	.4773	2.30	.4893	2.60	.4954	2.90	.4982
.21	.0832	.51	.1950	.81	.2911	1.11	.3665	1.41	.4208	1.71	.4564	2.01	.4778	2.31	.4896	2.61	.4955	2.91	.4982
.22	.0871	.52	.1985	.82	.2939	1.12	.3687	1.42	.4222	1.72	.4573	2.02	.4783	2.32	.4899	2.62	.4956	2.92	.4983
.23	.0910	.53	.2020	.83	.2968	1.13	.3708	1.43	.4237	1.73	.4582	2.03	.4788	2.33	.4901	2.63	.4958	2.93	.4983
.24	.0949	.54	.2054	.84	.2996	1.14	.3729	1.44	.4251	1.74	.4591	2.04	.4793	2.34	.4904	2.64	.4959	2.94	.4984
.25	.0987	.55	.2089	.85	.3024	1.15	.3749	1.45	.4265	1.75	.4599	2.05	.4798	2.35	.4906	2.65	.4960	2.95	.4984
.26	.1026	.56	.2123	.86	.3051	1.16	.3770	1.46	.4279	1.76	.4608	2.06	.4803	2.36	.4909	2.66	.4961	2.96	.4985
.27	.1064	.57	.2157	.87	.3079	1.17	.3790	1.47	.4292	1.77	.4617	2.07	.4808	2.37	.4911	2.67	.4962	2.97	.4985
.28	.1103	.58	.2191	.88	.3106	1.18	.3810	1.48	.4306	1.78	.4626	2.08	.4813	2.38	.4914	2.68	.4963	2.98	.4986
.29	.1141	.59	.2224	.89	.3133	1.19	.3830	1.49	.4319	1.79	.4633	2.09	.4817	2.39	.4916	2.69	.4965	2.99	.4986
								3.00	.4987	3.60	.4999								
								3.10	.4991	3.70	.4999								
								3.20	.4993	3.80	.5000								
								3.30	.4995	3.90	.5000								
								3.40	.4997	4.00	.5000								
								3.50	.4998										

TABLE 2.14

[1] Calculated from tables given by T. C. Fry, "Probability and Its Engineering Uses", with permission of the author.

The method of obtaining the empirical distribution is adequately illustrated in the data sheet of Table 2.15. Comparing the observed distribution, Column 9, with the theoretical one, Column 8, we see that there is a very close correspondence. Now there is something quite striking in the fact that a knowledge of the average $\bar{X}$ and the standard deviation σ of the observed set of 1370 values of depth of sapwood when used in this particular way in the normal function makes it possible for us to estimate so closely the number of observed values of X lying within any given limits. In other words, we see that the statistics, average and standard deviation, introduced in the previous chapter appear to contain a large amount of the information presented in the 1370 observed values of depth of sapwood in the sense that they make possible a close approximation to the observed distribution. This is so important that we tabulate in Table 2.16 observed and theoretical frequencies corresponding to certain ranges often used in analytical work.

0	1	2	3	4	5	6	7	8	9
Cell No.	Sapwood (a) in inches (b)	Cell Bound.	Deviations from $\bar{X}$ x	Z $(\frac{x}{\sigma})$	F(z)	Diff.	Freq	Approx Freq.	Obs. Freq.
		.25	2.6641	3.3376	.4995				
0	.4					.0010	1.4	1	
		.55	2.3641	2.9618	.4985				
1	.7					.0033	4.5	5	
		.85	2.0641	2.5859	.4952				
2	1.0					.0087	11.9	12	2
		1.15	1.7641	2.2101	.4865				
3	1.3					.0198	27.1	27	29
		1.45	1.4641	1.8342	.4667				
4	1.6					.0390	53.4	53	62
		1.75	1.1641	1.4584	.4277				
5	1.9					.0672	92.1	92	106
		2.05	.8641	1.0825	.3605				
6	2.2					.1004	137.5	138	153
		2.35	.5641	.7067	.2601				
7	2.5					.1305	178.8	179	186
		2.65	.2641	-.3309	.1296				
8	2.8					.1476	202.2	202	193
		2.95	.0359	+.0450	.0180				
9	3.1					.1451	198.8	199	188
		3.25	.3359	.4208	.1631				
10	3.4					.1240	169.9	170	151
		3.55	.6359	.7967	.2871				
11	3.7					.0924	126.6	127	123
		3.85	.9359	1.1725	.3795				
12	4.0					.0597	81.8	82	82
		4.15	1.2359	1.5483	.4392				
13	4.3					.0337	46.2	46	48
		4.45	1.5359	1.9242	.4729				
14	4.6					.0164	22.5	23	27
		4.75	1.8359	2.3000	.4893				
15	4.9					.0070	9.6	10	14
		5.05	2.1359	2.6759	.4963				
16	5.2					.0025	3.4	3	5
		5.35	2.4359	3.0517	.4988				
17	5.5					.0009	1.2	1	1
		5.65	2.7359	3.4275	.4997				
	Σ					.9992	1368.9	1370	1370

TABLE 2.15

Range

	$\bar{X} \pm .6745\sigma$	$\bar{X} \pm \sigma$	$\bar{X} \pm 2\sigma$	$\bar{X} \pm 3\sigma$
Theoretical	50.00%	68.27%	95.45%	99.73%
Observed	47.45%	66.57%	95.91%	99.93%
Difference	2.55%	1.70%	.46%	.20%

TABLE 2.16 - How Closely the Average $\bar{X}$ and Standard Deviation σ for the 1370 Observations of Depth of Sapwood Contain the Total Information. Note that theoretically almost all the observations should lie within the range $\bar{X} \pm 3\sigma$.

Of course we cannot expect such a close agreement between theoretical and observed frequency distributions if the observed distribution is not approximately symmetrical and at least unimodal. This suggests that we use some frequency function sensitive to skewness or lopsidedness of the observed distribution and quite naturally we take one involving the measure of skewness k introduced in the previous chapter.

3. The Second Approximation

The second approximation will be defined as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{x^2}{2\sigma^2}} \left[1 - \frac{k}{2}\left(\frac{x}{\sigma} - \frac{x^3}{3\sigma^3}\right)\right] \qquad (2.14)$$

where $x = X - \bar{X}$.

Or writing it in terms of $z = \frac{x}{\sigma}$, the corresponding expression is

$$\varphi(z) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{z^2}{2}} \left[1 - \frac{k}{2}\left(z - \frac{z^3}{3}\right)\right].$$

From this expression we readily obtain

$$\int_0^x f(x)\,dx = \sigma\int_0^z \varphi(z)\,dz = \int_0^z \frac{1}{\sqrt{2\pi}}\, e^{-\frac{z^2}{2}}\,dz - k\,\frac{1}{6\sqrt{2\pi}}\left[1 - (1-z^2)\,e^{-\frac{z^2}{2}}\right]$$

$$= F(z) - k\,f(z)$$

where F(z) and f(z) are given in Tables 2.14 and 2.17.

Table 2.18 gives the details of fitting the second approximation to the observed distribution of depth of sapwood. A comparison of the observed and

Table of Values of $f(z) = \frac{1}{6\sqrt{2\pi}}\left[1-(1-z^2)e^{-\frac{1}{2}z^2}\right]$

z	f(z)	z	f(z)	z	f(z)	z	f(z)	z	f(z)	z	f(z)	z	f(z)	z	f(z)	z	f(z)	z	f(z)
.00	.00000	.30	.00865	.60	.03095	.90	.05806	1.20	.08073	1.50	.09347	1.80	.09597	2.10	.09149	2.40	.08426	2.70	.07742
.01	.00001	.31	.00921	.61	.03183	.91	.05894	1.21	.08133	1.51	.09371	1.81	.09590	2.11	.09127	2.41	.08401	2.71	.07722
.02	.00004	.32	.00979	.62	.03272	.92	.05980	1.22	.08192	1.52	.09394	1.82	.09584	2.12	.09105	2.42	.08376	2.72	.07702
.03	.00009	.33	.01038	.63	.03361	.93	.06066	1.23	.08250	1.53	.09415	1.83	.09576	2.13	.09082	2.43	.08352	2.73	.07682
.04	.00016	.34	.01099	.64	.03450	.94	.06152	1.24	.08306	1.54	.09435	1.84	.09568	2.14	.09060	2.44	.08327	2.74	.07663
.05	.00025	.35	.01161	.65	.03540	.95	.06236	1.25	.08361	1.55	.09454	1.85	.09559	2.15	.09037	2.45	.08303	2.75	.07644
.06	.00036	.36	.01225	.66	.03631	.96	.06320	1.26	.08416	1.56	.09472	1.86	.09549	2.16	.09014	2.46	.08279	2.76	.07625
.07	.00049	.37	.01290	.67	.03721	.97	.06404	1.27	.08468	1.57	.09489	1.87	.09539	2.17	.08991	2.47	.08255	2.77	.07606
.08	.00064	.38	.01356	.68	.03812	.98	.06486	1.28	.08520	1.58	.09505	1.88	.09527	2.18	.08967	2.48	.08231	2.78	.07588
.09	.00081	.39	.01424	.69	.02904	.99	.06568	1.29	.08571	1.59	.09519	1.89	.09516	2.19	.08943	2.49	.08207	2.79	.07569
.10	.00099	.40	.01493	.70	.03995	1.00	.06649	1.30	.08620	1.60	.09533	1.90	.09503	2.20	.08919	2.50	.08183	2.80	.07551
.11	.00120	.41	.01564	.71	.04086	1.01	.06729	1.31	.08668	1.61	.09546	1.91	.09490	2.21	.08895	2.51	.08159	2.81	.07534
.12	.00143	.42	.01636	.72	.04178	1.02	.06809	1.32	.08715	1.62	.09557	1.92	.09477	2.22	.08871	2.52	.08136	2.82	.07516
.13	.00167	.43	.01708	.73	.04270	1.03	.06887	1.33	.08760	1.63	.09567	1.93	.09463	2.23	.08847	2.53	.08112	2.83	.07499
.14	.00194	.44	.01782	.74	.04362	1.04	.06965	1.34	.08805	1.64	.09577	1.94	.09448	2.24	.08823	2.54	.08089	2.84	.07482
.15	.00222	.45	.01857	.75	.04453	1.05	.07042	1.35	.08848	1.65	.09585	1.95	.09433	2.25	.08798	2.55	.08066	2.85	.07465
.16	.00253	.46	.01933	.76	.04545	1.06	.07118	1.36	.08890	1.66	.09592	1.96	.09417	2.26	.08774	2.56	.08043	2.86	.07448
.17	.00286	.47	.02011	.77	.04637	1.07	.07193	1.37	.08930	1.67	.09599	1.97	.09401	2.27	.08749	2.57	.08020	2.87	.07432
.18	.00319	.48	.02089	.78	.04728	1.08	.07267	1.38	.08970	1.68	.09604	1.98	.09384	2.28	.08724	2.58	.07998	2.88	.07416
.19	.00356	.49	.02168	.79	.04820	1.09	.07340	1.39	.09008	1.69	.09608	1.99	.09366	2.29	.08699	2.59	.07975	2.89	.07400
.20	.00392	.50	.02248	.80	.04911	1.10	.07412	1.40	.09045	1.70	.09612	2.00	.09349	2.30	.08674	2.60	.07953	2.90	.07384
.21	.00432	.51	.02329	.81	.05002	1.11	.07483	1.41	.09080	1.71	.09614	2.01	.09330	2.31	.08650	2.61	.07931	2.91	.07369
.22	.00473	.52	.02411	.82	.05093	1.12	.07552	1.42	.09115	1.72	.09616	2.02	.09312	2.32	.08625	2.62	.07909	2.92	.07354
.23	.00516	.53	.02494	.83	.05183	1.13	.07621	1.43	.09148	1.73	.09616	2.03	.09293	2.33	.08600	2.63	.07888	2.93	.07339
.24	.00561	.54	.02578	.84	.05274	1.14	.07689	1.44	.09180	1.74	.09616	2.04	.09273	2.34	.08575	2.64	.07866	2.94	.07324
.25	.00607	.55	.02662	.85	.05363	1.15	.07756	1.45	.09211	1.75	.09615	2.05	.09253	2.35	.08550	2.65	.07845	2.95	.07309
.26	.00654	.56	.02748	.86	.05453	1.16	.07822	1.46	.09241	1.76	.09613	2.06	.09233	2.36	.08525	2.66	.07824	2.96	.07295
.27	.00706	.57	.02833	.87	.05542	1.17	.07886	1.47	.09269	1.77	.09610	2.07	.09213	2.37	.08500	2.67	.07803	2.97	.07281
.28	.00757	.58	.02920	.88	.05631	1.18	.07950	1.48	.09296	1.78	.09606	2.08	.09192	2.38	.08475	2.68	.07782	2.98	.07267
.29	.00810	.59	.03007	.89	.05719	1.19	.08012	1.49	.09322	1.79	.09602	2.09	.09170	2.39	.08450	2.69	.07762	2.99	.07254
								3.00	.07240	3.60	.06771								
								3.10	.07118	3.70	.06739								
								3.20	.07016	3.80	.06714								
								3.30	.06933	3.90	.06696								
								3.40	.06866	4.00	.06683								
								3.50	.06813										

0	1	2	3	4	5	6	7	8	9	10	11	0
CELL NO.	Sapwood (a) in inches (b)	CELL BOUND.	DEV. FROM $\bar{X}$ x	z $\left(\frac{x}{\sigma}\right)$	$F(z)$	$\pm f(z)$	$\pm kf(z)$	$F(z)\pm kf(z)$	DIFF.	FREQ.	Obs. Freq.	CELL NO.
		.85	2.0641	2.5859	.4952	.0799	.0196	.5148				
0	1.0								.0065	9	2	0
		1.15	1.7641	2.2101	.4865	.0889	.0218	.5083				
1	1.3								.0181	25	29	1
		1.45	1.4641	1.8342	.4667	.0958	.0235	.4902				
2	1.6								.0399	56	62	2
		1.75	1.1641	1.4584	.4277	.0924	.0226	.4503				
3	1.9								.0719	99	106	3
		2.05	.8641	1.0825	.3605	.0729	.0179	.3784				
4	2.2								.1084	149	153	4
		2.35	.5641	.7067	.2601	.0406	.0099	.2700				
5	2.5								.1378	189	186	5
		2.65	.2641	.3309	.1296	.0105	.0026	.1322				
6	2.8								.1501	207	193	6
		2.95	.0359	.0450	.0180	.0003	.0001	.0179				
7	3.1								.1412	193	188	7
		3.25	.3359	.4258	.1631	.0164	.0040	.1591				
8	3.4								.1160	159	151	8
		3.55	.6359	.7967	.2871	.0488	.0120	.2751				
9	3.7								.0850	116	125	9
		3.85	.9359	1.1725	.3795	.0791	.0194	.3601				
10	4.0								.0560	77	82	10
		4.15	1.2359	1.5483	.4392	.0945	.0231	.4161				
11	4.3								.0336	46	48	11
		4.45	1.5359	1.9242	.4729	.0947	.0232	.4497				
12	4.6								.0184	25	27	12
		4.75	1.8359	2.3000	.4893	.0867	.0212	.4681				
13	4.9								.0091	13	14	13
		5.05	2.1359	2.6759	.4963	.0779	.0191	.4772				
14	5.2								.0040	6	5	14
		5.35	2.4359	3.0517	.4988	.0718	.0176	.4812				
15	5.5								.0017	2	1	15
		5.65	2.7359	3.4275	.4997	.0684	.0168	.4829				
	Σ								.9977	1370	1370	

TABLE 2.18

theoretical distributions in this table show that there is very close correspondence between the two. Suppose then that we plot the distribution of the observed set of 1370 values of depth of sapwood and then see how closely the curves derived from the two theoretical distributions used above actually fit the observed points. This is done in Fig. 2.11. Obviously the second approximation makes possible a little closer approximation to the original distribution than that obtained through the use of the normal law. It should be noted, of course, that a table calculated for the second approximation similar to Table 2.16 would be identical with it because the integral of the normal law between the limits $-z_1$ and $+z_1$ is identical with the integral of the second approximation between the same limits.

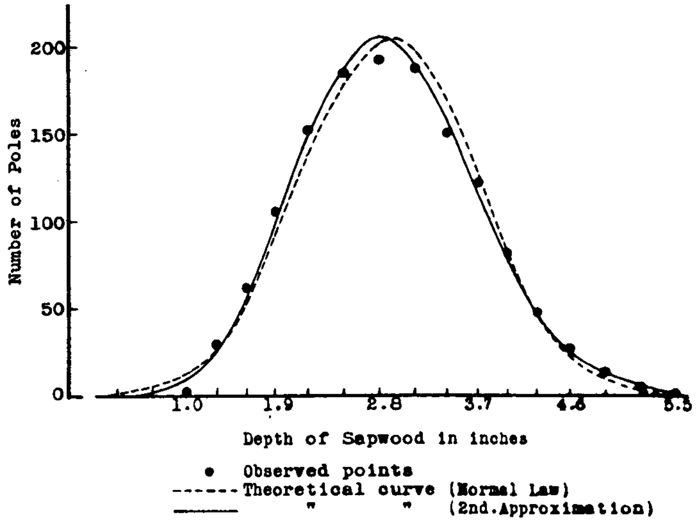


FIG. 2.11

4. Why the Arithmetic Mean $\bar{X}$ and Standard Deviation σ Always Contain a Large Part of the Essential Information

Given the arithmetic mean $\bar{X}$ and standard deviation σ of a set of observed values of some quality X, we have seen how to use the normal law integral to secure an approximate estimate of the number of values of X within any specified range $\bar{X} \pm z\sigma$. In the illustrative problem that we chose we found this estimate to be close indeed, but there are obvious reasons why the degree of approximation may not be so good in another case. For example, the application of the normal law integral implies the assumption that the observed values are distributed in accordance with the normal law function with an average $\bar{X}$ and standard deviation σ equal to those of a given set of n observed values of X.

In no case is this assumption rigorously satisfied,- the theory applies to a continuous distribution whereas the observed one is necessarily discontinuous and the theory applies to a special symmetrical unimodal distribution whereas the observed distribution is seldom symmetrical even though it be unimodal. To what extent then are we justified in assuming that the average $\bar{X}$ and standard deviation σ contain much of the total information given in the original series of observations in respect to the number of values of X within any symmetrical range $\bar{X} \pm z\sigma$? The ingenious work of the Russian statistician Tchebycheff comes to our aid in answering this question. He gives us a beautiful and very general theorem the proof of which can be framed in the simplest kind of elementary mathematics as we shall now see.

Given any set of n observed values expressible in the frequency distribution

$$\begin{array}{l} X_1 , X_2 , \ldots, X_i , \ldots, X_m \\ p_1n, p_2n, \ldots, p_in, \ldots p_mn \end{array}$$

where p_in represents the number of values of X_i, then

$$\bar{X} = \frac{\sum_{i=1}^{m} p_in X_i}{\sum_{i=1}^{m} p_in} = \sum_{i=1}^{m} p_iX_i,$$

and

$$\sigma^2 = \frac{\sum_{i=1}^{m} p_in(X_i - \bar{X})^2}{\sum_{i=1}^{m} p_in} .$$

Let P_zn denote the number of values of X such that $x = (X - \bar{X})$ does not exceed numerically $z\sigma$ where $z > 1$, and $n - P_zn$ denote the number of values of x that do exceed $z\sigma$.

Now we may write

$$\sigma^2 = \Sigma_1 p_i x_i^2 + \Sigma_2 p_i x_i^2$$

where Σ_1 denotes summation for all values of x_i which do not exceed $z\sigma$ and Σ_2 denotes summation for all values of x_i which do exceed $z\sigma$. Since all values of $p_i x_i^2$ are either positive or zero,

$$\sigma^2 \geqq \Sigma_2 p_i x_i^2 .$$

Obviously, therefore,

$$\sigma^2 \geqq \Sigma_2 p_i z^2 \sigma^2$$

since all values of x_i included in the summation Σ_2 are greater than $z\sigma$.

But $$\Sigma_2 p_i = 1 - P_z,$$

hence $$\sigma^2 \geqq (1 - P_z) z^2 \sigma^2,$$

or $$1 \geqq (1 - P_z) z^2,$$

$$(1 - P_z) \leqq \frac{1}{z^2} , \qquad (2.15)$$

and $$P_z \geqq 1 - \frac{1}{z^2} . \qquad (2.16)$$

Now P_z is just the thing we want to get from a knowledge of the average $\overline{X}$ and standard deviation σ of the n observed values of X. We see that no matter what set of observed values we may have, the number of these values $P_z n$ lying on or within the range $\overline{X} \pm z\sigma$ is equal to or greater than $(1 - \frac{1}{z^2})n$ whereas the number $(1 - P_z)n$ lying without this range cannot be greater than $\frac{1}{z^2}$ n.

We are now in a place to see just how accurately a knowledge of the average and standard deviation of a series of observations gives us the total information presented by the series itself. Let us recall the two purposes of taking data, namely, that of obtaining a series of observations as quantitative information and that of obtaining some causal interpretation of the data. The essential information in the first case is obviously completely given by the frequency distribution itself whereas in the second case, it will be found later that the essential information does not necessarily require the knowledge of the exact frequency distribution.

If a distribution is quite asymmetrical and cannot be represented by the second approximation, we have not, as far as we have gone, any general method for representing the observed frequency distribution from a knowledge of the average, standard deviation and skewness of that distribution. We can, however, with the aid of Tchebycheff's theorem, show just how close we can come to estimating the number of observed values within any given range $\bar{X} \pm z\sigma$ from a knowledge of the average and standard deviation alone where z is greater than unity. For example, in Fig. 2.12 the ordinate represents the fraction of

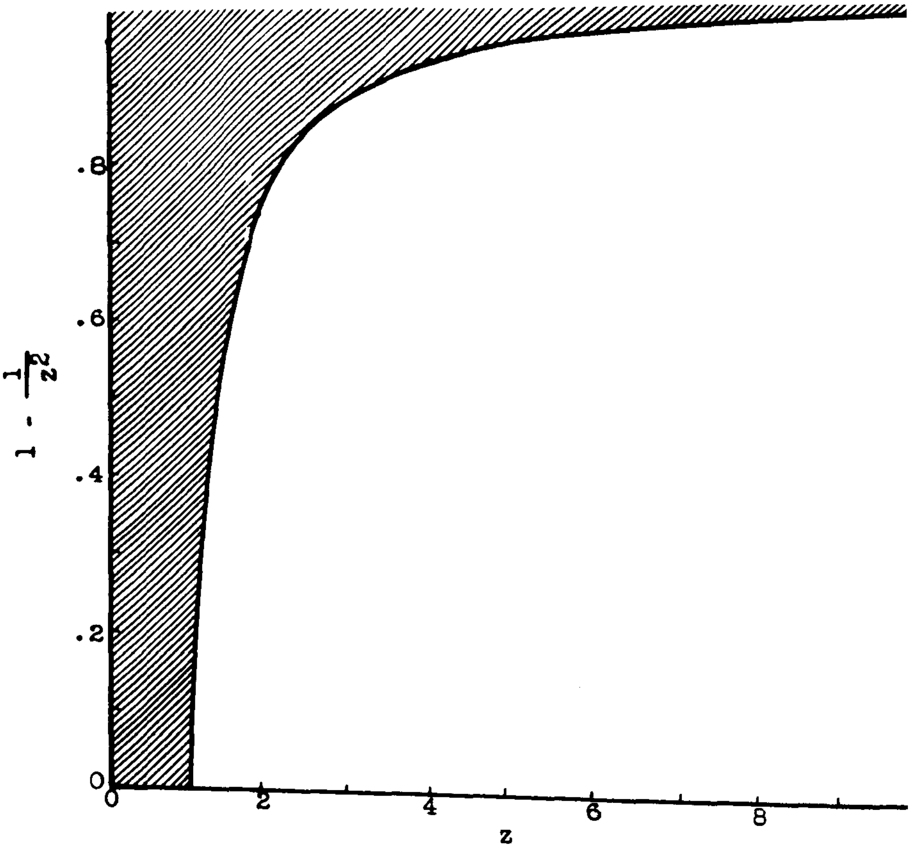


FIG. 2.12

observed values within the range $\bar{X} \pm z\sigma$. If we know, for example, the average and standard deviation of a series of numbers, no matter what the series is, we can say that not less than $1 - \frac{1}{z^2}$ of these numbers have values on or within the range $\bar{X} \pm z\sigma$. The function $1 - \frac{1}{z^2}$ is given in Fig. 2.12 and here we see that for z = 2 the number of observed values on or within the corresponding range is not less than .7500 n. In a similar way for z = 3 the number on or within the range $\bar{X} \pm 3\sigma$ is not less than .8889 n and for the case of $z = 6$, the number on or within the corresponding range is not less than .9722 n.

A little study of Fig. 2.12 reveals some rather startling information. We see that we cannot make a mistake of more than 25% in estimating the proportion of observations that will fall on or within the range of the average plus or minus 2σ. Similarly we cannot make a mistake of more than about 11% in making a corresponding estimate for the range $\bar{X} \pm 3\sigma$, and so on. The depth of the shaded area for a given value of z indicates the upper limit to the proportional amount of total information given by the original series and not included in the average and standard deviation, in respect to a symmetrical range about the average. Hence if we are interested in a range corresponding to a value of z equal to or greater than about three, we see that the average and standard deviation quite accurately give the required information. So long then as the information to be derived from the set of data involves a symmetrical range in the sense of the present discussion, we see that the average and standard deviation contain a large part of the total information. We shall find later that these same two parameters of a given set of data contain a large amount of the essential information for the causal interpretation of the set of data. When, however, the essential information involves the total frequency associated with any specified asymmetrical range, the average and standard deviation cannot be relied upon so implicitly. In such cases we must make use of more functions of the original data and use these functions as parameters in more complicated forms of frequency curves.

Before entering upon the discussion of these complicated frequency curves, however, let us recall here certain empirical information which indicates that the actual difference between the information contained in the average and standard deviation in respect to the frequency corresponding to a symmetrical range is not as great as that indicated in Fig. 2.12. Of the thousands of problems coming to our attention we have never found a frequency distribution where more than 2.5 percent fall outside the observed average plus or minus three times the observed standard deviation. This you see is only about 1/3 of the maximum possible error that can be made in estimating this percentage as indicated by Fig. 2.12.

5. Presentation of Data by Means of Estimates of Parameters in Either of Two General Frequency Functions

We recall that our object, as stated at the beginning of this Chapter, is to present at least in a practical sense, by means of a frequency function, the total information contained in the set of data, or in other words, to find a function f(X) which is such that

$$\int_a^b f(X)\,dX = \text{number of observed values of } X \text{ lying within the corresponding interval.}$$

We have shown that under certain very special conditions, we may express the total information contained in the set of data by means of the normal law frequency function. Also, we have seen that under more general conditions, we may attain the same end by the use of a somewhat more complicated function known as the Second Approximation. Under more general conditions than those yet considered, recourse must be had to still more complicated frequency functions if we are to present the total information given by the set of data.

We shall present now two general methods of solution, which, for a rather wide variety of types of observed distributions, serve the purpose already stated.

Inasmuch as the use of these two frequency functions now to be considered, have considerable in common, we may best compare by a diagram their mathematical expressions and the outline of the method of procedure followed in each case.

GRAM CHARLIER SERIES

Given an arbitrary frequency function f(x) (x is the deviation from the mean of the distribution) continuous and finite in the interval $-\infty$ to $+\infty$ and which vanishes at $x = \pm\infty$, to determine the coefficients a_0, $a_1, \ldots, a_n, \ldots$ in such a way that

$$a_0\varphi_0(x) + a_1\varphi_1(x) + \ldots + a_n\varphi_n(x) + \ldots$$

gives the best approximation to f(x) in the sense of least squares, where

$$\varphi_0(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

and $\varphi_n(x)$ is the nth derivative of $\varphi_0(x)$.

PEARSON SYSTEM OF CURVES

Assume that Y = f(X) is a continuous unimodal function of X which vanishes together with its derivative at the limits of the frequency range.

Such assumptions suggest that Y be defined by a differential equation of the form

$$\frac{dY}{dX} = \frac{Y(X+d)}{b_0 + b_1X + b_2X^2 + \ldots}$$

The constants d, b_0, b_1, $b_2, \ldots$ are to be determined from the moments of the given frequency function Y.

The disadvantage of this system for our purpose is that the

Under these conditions, it can be shown that

$$a_n = \frac{(-1)^n \sigma^{2n}}{\underline{|n}} \int_{-\infty}^{+\infty} f(x) H_n(x)\,dx$$

where

$$H_n(x) = \frac{x^n}{\sigma^{2n}} - \frac{n(n-1)}{2} \frac{x^{n-2}}{\sigma^{2n-2}} + \ldots$$

$$= (-1)^n \frac{\varphi_n(x)}{\varphi_0(x)}$$

From the definition of a_n, it is seen that the coefficients a_n, $n = 0, 1, 2, \ldots$ are determined in terms of the moments of the given function.

The great advantage for our purpose of this way of expressing f(x) is that the $\int_a^b f(x)\,dx$ can readily be obtained.

$\int_a^b Y\,dX$ can seldom be obtained and recourse must be had to quadrature formulas.

In what follows we shall give, in connection with the mathematical details involved in the use of these two systems, only those which seem to us essential to a clear understanding of the method of using either one of the two forms of representation. The tables given near the end of this section indicate how well either of these two functions represent, in a particular example, the total information contained in the data.

Use of Gram Charlier Series

There are several ways of determining the coefficients a_n of this series, and we shall choose to illustrate that one which makes use of the biorthogonal property of the functions $H_n(x)$ and $\varphi_m(x)$ introduced above.

Assume that the function f(x) can be expanded in a series

$$f(x) = a_0\varphi_0(x) + a_1\varphi_1(x) + a_2\varphi_2(x) + \ldots + a_n\varphi_n(x) + \ldots \qquad (2.17)$$

where $\varphi_0(x)$ and $\varphi_n(x)$ are defined as above.

Then, if the series (2.17) converges uniformally in the interval $-\infty$ to $+\infty$,

$$\int_{-\infty}^{+\infty} f(x)H_n(x)\,dx = a_0 \int_{-\infty}^{+\infty} \varphi_0(x)H_n(x)\,dx + a_1 \int_{-\infty}^{+\infty} \varphi_1(x)H_n(x)\,dx +$$

$$\ldots . + a_m \int_{-\infty}^{+\infty} \varphi_m(x)H_n(x)\,dx + \ldots + a_n \int_{-\infty}^{+\infty} \varphi_n(x)H_n(x)\,dx + \ldots \qquad (2.18)$$

where

$$H_n(x) = \frac{x^n}{\sigma^{2n}} - \frac{n(n-1)}{2}\frac{x^{n-2}}{\sigma^{2n-2}} + \frac{n(n-1)(n-2)(n-3)}{2.4}\frac{x^{n-4}}{\sigma^{2n-4}}$$

$$- \frac{n(n-1)(n-2)(n-3)(n-4)(n-5)}{2.4.6}\frac{x^{n-6}}{\sigma^{2n-6}} + \ldots . = (-1)^n \frac{\varphi_n(x)}{\varphi_0(x)} .$$

We shall show that

$$\int_{-\infty}^{+\infty} \varphi_m(x)H_n(x)\,dx = 0, \qquad (m \neq n),$$

$$\int_{-\infty}^{+\infty} \varphi_n(x)H_n(x)\,dx = \frac{(-1)^n \underline{|n}}{\sigma^{2n}}, \qquad (m = n).$$

Hence all the terms on the right of (2.18) vanish except the one in a_n and this term has the value

$$\frac{a_n(-1)^n \underline{|n}}{\sigma^{2n}}$$

Therefore, solving for the coefficient, we have

$$a_n = \frac{(-1)^n \sigma^{2n}}{\underline{|n}} \int_{-\infty}^{+\infty} f(x)H_n(x)\,dx \qquad n = 0,1,2,\ldots$$

Consider first $m > n$:

Integrating by parts, we have

$$\int_{-\infty}^{+\infty} \varphi_m(x)H_n(x)\,dx = \int_{-\infty}^{+\infty} H_n(x)\,d\,[\varphi_{m-1}(x)] = H_n(x)\ \varphi_{m-1}(x) \Big[_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} \varphi_{m-1}(x)H_n'(x)\,dx$$

$$= -\int_{-\infty}^{+\infty} \varphi_{m-1}(x)H_n'(x)\,dx,$$

since $H_n(x)\varphi_{m-1}(x)$ is a polynomial of degree $(n+m-1)$ in x multiplied by $\varphi_0(x)$. Such a product always vanishes at $x = \pm\infty$ independent of the degree of the polynomial.

Repeating the process, we find

$$-\int_{-\infty}^{+\infty}\varphi_{m-1}(x)H_n'(x)\,dx = -\int_{-\infty}^{+\infty}H_n'(x)\,d\,[\varphi_{m-2}(x)]$$

$$= -H_n'(x)\,\varphi_{m-2}(x)\Big|_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty}\varphi_{m-2}(x)H_n''(x)\,dx = \int_{-\infty}^{+\infty}\varphi_{m-2}(x)H_n''(x)\,dx.$$

Continuing the process, we would get finally

$$\int_{-\infty}^{+\infty}\varphi_m(x)H_n(x)\,dx = (-1)^{n+1}\int_{-\infty}^{+\infty}\varphi_{m-n-1}(x)H_n^{(n+1)}(x)\,dx = 0,$$

since $H_n(x)$ is a polynomial in x of degree n.

If m = n, we find

$$\int_{-\infty}^{+\infty}\varphi_n(x)H_n(x)\,dx = (-1)^n\int_{-\infty}^{+\infty}\varphi_0(x)H_n^n(x)\,dx.$$

Now

$$H_n^n(x) = \frac{\underline{|n}}{\sigma^{2n}} \text{ and } \int_{-\infty}^{+\infty}\varphi_0(x)\,dx = 1\,.$$

Therefore

$$\int_{-\infty}^{+\infty}\varphi_n(x)H_n(x)\,dx = (-1)^n\,\frac{\underline{|n}}{\sigma^{2n}}\,.$$

If $m < n$, we may proceed as follows:

By definition

$$\int_{-\infty}^{+\infty}\varphi_m(x)H_n(x)\,dx = (-1)^m\int_{-\infty}^{+\infty}H_m(x)H_n(x)\,\varphi_0(x)\,dx = (-1)^{m+n}\int_{-\infty}^{+\infty}\varphi_n(x)H_m(x)\,dx.$$

By repeated integration by parts, this last integral reduces to

$$(-1)^{2m+n+1}\int_{-\infty}^{+\infty}\varphi_{n-m-1}(x)H_m^{(m+1)}(x)\,dx$$

which has the value zero, since $H_m(x)$ is a polynomial of degree m in x.

Formally then, it is now merely a matter of grinding out as many coefficients a_n as we desire and substituting their values in Equation (2.17).

For example,

$$a_0 = \int_{-\infty}^{+\infty} f(x)H_0(x)\,dx = \int_{-\infty}^{+\infty} f(x)\,dx = 1$$

where $f(x)$ is taken to be a relative frequency distribution.

$$a_1 = -\int_{-\infty}^{+\infty} f(x)\,x\,dx = 0,$$

since the mean of the distribution is chosen as the origin.

$$a_2 = \frac{\sigma^4}{2}\int_{-\infty}^{+\infty} f(x)\left(\frac{x^2}{\sigma^4} - \frac{1}{\sigma^2}\right) dx = 0,$$

$$a_3 = -\frac{\sigma^6}{6}\int_{-\infty}^{+\infty} f(x)\left(\frac{x^3}{\sigma^6} - \frac{3x}{\sigma^4}\right) dx = -\frac{\mu_3}{6} = -\frac{k\sigma^3}{6},$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots ,$$

$$a_n = \frac{(-1)^n\,\sigma^{2n}}{\underline{\lvert n}}\int_{-\infty}^{+\infty} f(x)\left[\frac{x^n}{\sigma^{2n}} - \frac{n(n-1)}{2}\,\frac{x^{n-2}}{\sigma^{2n-2}} + \frac{n(n-1)(n-2)(n-3)}{2.4}\,\frac{x^{n-4}}{\sigma^{2n-4}} + \dots\right] dx.$$

Continuing in this way, we shall find that

$$f(x) = \varphi_0(x)\left[1 - \frac{k\sigma^3}{\underline{\lvert 3}}\left(\frac{3x}{\sigma^4} - \frac{x^3}{\sigma^6}\right) + \frac{\sigma^4}{\underline{\lvert 4}}(\beta_2 - 3)\left(\frac{3}{\sigma^4} - \frac{6x^2}{\sigma^6} + \frac{x^4}{\sigma^8}\right) + \frac{\sigma^5}{\underline{\lvert 5}}\left(10k - \frac{\mu_5}{\sigma^5}\right)\left(-\frac{15x}{\sigma^6} + \frac{10x^3}{\sigma^8} - \frac{x^5}{\sigma^{10}}\right)\right.$$

$$\left. + \frac{\sigma^6}{\underline{\lvert 6}}\left(30 - 15\frac{\mu_4}{\sigma^4} + \frac{\mu_6}{\sigma^6}\right)\left(\frac{x^6}{\sigma^{12}} - \frac{15x^4}{\sigma^{10}} + \frac{45x^2}{\sigma^8} - \frac{15}{\sigma^6}\right) + \dots\right], \qquad (2.19)$$

where k, σ, β_2, μ_4, μ_5, etc., are moments or functions of moments of the given distribution $f(x)$.

Before passing on, it is of interest to note that, if we cut off the series (2.19) at the term in $\varphi_3(x)$, we have

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{x^2}{2\sigma^2}}\left[1 - \frac{k}{2}\left(\frac{x}{\sigma} - \frac{x^3}{3\sigma^3}\right)\right]$$

the Second Approximation (2.14) already discussed; and cutting off the series at $\varphi_0(x)$ we have the Normal Law function.

Thus we see that the First and Second Approximations are really approximations to a series definition of f(x).

Returning now to (2.19) we may considerably simplify this expression by measuring x in units of σ, i.e., $x = z\sigma$. We then have

$$f(z) = \frac{1}{\sigma}\varphi_0(z)\left[1 - \frac{k}{\underline{\lfloor 3}}(3z - z^3) + \frac{1}{\underline{\lfloor 4}}(\beta_2 - 3)(3 - 6z^2 + z^4)\right.$$

$$+ \frac{1}{\underline{\lfloor 5}}\left\{10k - \frac{\mu_5}{\sigma^5}\right\}(-15z + 10z^3 - z^5)$$

$$\left. + \frac{1}{\underline{\lfloor 6}}\left\{30 - 15\frac{\mu_4}{\sigma^4} + \frac{\mu_6}{\sigma^6}\right\}(-15 + 45z^2 - 15z^4 + z^6) + \ldots\ldots\right] \qquad (2.20)$$

$$= \frac{1}{\sigma}\varphi_0(z) - \frac{k}{\sigma\underline{\lfloor 3}}\varphi_3(z) + \frac{1}{\sigma\underline{\lfloor 4}}(\beta_2 - 3)\,\varphi_4(z)$$

$$+ \frac{1}{\sigma\underline{\lfloor 5}}\left\{10k - \frac{\mu_5}{\sigma^5}\right\}\varphi_5(z) + \frac{1}{\sigma\underline{\lfloor 6}}\left\{30 - 15\frac{\mu_4}{\sigma^4} + \frac{\mu_6}{\sigma^6}\right\}\varphi_6(z) + \ldots \qquad (2.21)$$

where $\varphi_0(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$, and $\varphi_n(z)$ = the n^{th} derivative of $\varphi_0(z)$. This gives the ordinates of the function f(z) in the most convenient form for computation since the values of $\varphi_0(z)$, $\varphi_3(z)$, etc. have been tabulated[1] over a wide range of values of z.

However, there are two things that must be taken into consideration in the use of such a frequency function to represent an observed set of data. The first is that the computation of terms on the right of (2.21) even as far as $\varphi_6(z)$ becomes very laborious, and second, if we decide on a given degree of accuracy, we must insure that the terms neglected are those which do not affect the accuracy of the desired result. For example, it has been shown that as a First Approximation, we may use the term in $\varphi_0(z)$ alone. As a Second Approximation, we may use the terms in $\varphi_0(z)$ and $\varphi_3(z)$. As a Third Approximation, we should use the terms in $\varphi_0(z)$, $\varphi_3(z)$, $\varphi_4(z)$ and $\varphi_6(z)$ and so on.[2]

- -

[1] James W. Glover, "Tables of Applied Mathematics in Finance, Insurance, Statistics".

[2] Further discussion of this matter is given in a recent book "Probability and its Engineering Uses" by T. C. Fry.

Of course, these approximations are just those given before in terms of x only now the variable has been changed to z. We shall use just such combinations of terms in the example, to be given later.

Let us recall now that we started out to obtain a function which, when integrated over any given range, would give us the number of observations lying within the corresponding range, or in short, give us the total information given by the set of data. To this end, we return to equation (2.17) and integrate term by term, i.e.

$$\int_a^b f(x)dx = a_0 \int_a^b \varphi_0(x)dx + a_1 \int_a^b \varphi_1(x)dx + \dots + a_n \int_a^b \varphi_n(x)dx + \dots\dots$$

The expressions on the right can be integrated at once in terms of the original variable x, since by definition

$$\int_a^b \varphi_n(x)dx = \varphi_{n-1}(x)\Big|_a^b .$$

However, as in the case of the ordinates of f(x), it will be found convenient from the standpoint of computation to express this integral in terms of the variable $z = \frac{x}{\sigma}$.

Making the necessary substitutions, we find,

$$\int_0^x f(x)dx = \sigma \int_0^z f(z)dz = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-\frac{z^2}{2}} dz - \frac{k}{6}\left[\frac{1}{\sqrt{2\pi}} - (1-z^2)\ \varphi_0(z)\right]$$

$$+ \frac{1}{\lfloor\underline{4}} (\beta_2-3)(3z-z^3)\ \varphi_0(z) + \frac{1}{\lfloor\underline{5}} \left\{10k - \frac{\mu_5}{\sigma^5}\right\} \left[(3-6z^2 + z^4)\ \varphi_0(z) - \frac{3}{\sqrt{2\pi}}\right]$$

$$+ \frac{1}{\lfloor\underline{6}} \left\{30-15\ \frac{\mu_4}{\sigma^4} + \frac{\mu_6}{\sigma^6}\right\}(-15z + 10z^3 - z^5)\ \varphi_0(z)$$

or if an approximation for the ordinates of f(z) involves only terms of the order of $\varphi_4(z)$ and $\varphi_6(z)$

$$\sigma\int_0^z f(z)\,dz = \frac{1}{\sqrt{2\pi}}\int_0^z e^{-\frac{z^2}{2}}\,dz - \frac{k}{6}\left(\frac{1}{\sqrt{2\pi}} + \varphi_2(z)\right)$$

$$+ \frac{1}{\underline{|4}}\,(\beta_2-3)\;\varphi_3(z) + \frac{1}{\underline{|6}}\left(30 - 15\,\frac{\mu_4}{\sigma^4} + \frac{\mu_6}{\sigma^6}\right)\varphi_5(z) + \dots \quad (2.22)$$

Also since

$$\int_{z_1}^{z_2} f(z)\,dz = \int_0^{z_2} f(z)\,dz - \int_0^{z_1} f(z)\,dz$$

we need only the value of integrals of the form (2.22).

Example - Application of Series to Point Binomial $(.9 + .1)^{100}$

The given function f(z) now becomes the ordinates of the Point Binomial

$$(q+p)^n = (.9 + .1)^{100}$$

and the σ, k, β_2, etc., appearing in Equation (2.21) are the same constants found from the given Point Binomial.[1] Hence in using (2.21) to compute approximate ordinates for the Binomial, we should first calculate the coefficients of $\varphi_0(z)$, $\varphi_3(z)$, etc., which are independent of z.

Making use of the general expression[2] for the moments of the Point Binomial about the mean value pn, we have at once

$$pn = 10 = \text{mean value of Binomial distribution},$$

$$\sigma^2 = pqn = 9,$$

$$k = \frac{q-p}{\sigma} = .2666667,$$

$$\frac{-k}{\sigma\,\underline{|3}} = -\,.01481481,$$

$$\beta_2 = 3.0511111, \qquad \frac{\beta_2-3}{\sigma\,\underline{|4}} = .0007098765$$

[1] We choose this distribution because the results obtained not only illustrate the method of presenting in terms of parameters for a Gram Charlier series the information given by a distribution but also because the results will be made use of in Part III.

[2] To be developed in Part III.

$$\frac{1}{\sigma\lfloor 6}\left(30 - 15\frac{\mu_4}{\sigma^4} + \frac{\mu_6}{\sigma^6}\right) = .000325057$$

Equation (2.21) now takes the simpler form

$$f(z) = .3333333\ \varphi_0(z) - .01481481\ \varphi_3(z) + .0007098765\ \varphi_4(z)$$

$$+ .000325057\ \varphi_6(z) + \ldots \qquad (2.23)$$

Since the functions $\varphi_0(z)$, $\varphi_3(z)$, etc. have been tabulated by Glover, Equation (2.23) presents a fairly easy means of computing the approximate ordinates of the given Point Binomial.

Table 2.19 is arranged in such a way that the degree of approximation obtained by using only the terms in $\varphi_0(z)$; those in $\varphi_0(z)$ and $\varphi_3(z)$; those in $\varphi_0(z)$, $\varphi_3(z)$, $\varphi_4(z)$ and $\varphi_6(z)$ of (2.23) may readily be seen by comparing each with the actual ordinates of the Point Binomial given in column (11).

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
X	X-pn=x	$z = \frac{x}{\sigma}$	$\frac{1}{\sigma}\varphi_0(z)$	$\frac{k}{\sigma\lfloor 3}\varphi_3(z)$	$\frac{\alpha_2 - 3}{\sigma\lfloor 4}\varphi_4(z)$	$\frac{1}{\sigma\lfloor 6}\left(30-15\frac{\mu_4}{\sigma^4}+\frac{\mu_6}{\sigma^6}\right)\varphi_6(z)$	First Approximation (4)	Second Approximation (4)+(5)	Third Approximation (4)+(5) +(6)+(7)	$(.9+.1)^{100}$
0	-10	-3.3333	.00051	-.00062	.00007	.00000	.00051	-.00011	-.00004	.0000
1	- 9	-3.0000	.00148	-.00118	.00009	-.00014	.00148	.00050	.00025	.0003
2	- 8	-2.6667	.00380	-.00186	.00009	-.00035	.00380	.00195	.00169	.0016
3	- 7	-2.3333	.00874	-.00222	-.00000	-.00045	.00874	.00652	.00607	.0059
4	- 6	-2.0000	.01800	-.00160	-.00019	-.00019	.01800	.01640	.01602	.0159
5	- 5	-1.6667	.03316	.00055	-.00042	.00051	.03316	.03371	.03380	.0339
6	- 4	-1.3333	.05467	.00396	-.00052	.00124	.05467	.05863	.05935	.0596
7	- 3	-1.0000	.08066	.00717	-.00034	.00126	.08066	.08783	.00875	.0889
8	- 2	- .6667	.10648	.00806	.00012	.00022	.10648	.11454	.11488	.1148
9	- 1	- .3333	.12579	.00538	.00063	-.00125	.12579	.13117	.13055	.1304
10	0	0	.13298	.00000	.00085	-.00195	.13298	.13298	.13188	.1319
11	1	.3333	.12579	-.00538	.00063	-.00125	.12579	.12041	.11979	.1199
12	2	.6667	.10648	-.00806	.00012	.00022	.10648	.09842	.09876	.0988
13	3	1.0000	.08066	-.00717	-.00034	.00126	.08066	.07349	.07441	.0743
14	4	1.3333	.05467	-.00396	-.00052	.00124	.05467	.05071	.05143	.0513
15	5	1.6667	.03316	-.00055	-.00042	.00051	.03316	.03261	.03270	.0327
16	6	2.0000	.01800	.00160	-.00019	-.00019	.01800	.01960	.01922	.0193
17	7	2.3333	.00874	.00222	-.00000	-.00045	.00874	.01096	.01051	.0106
18	8	2.6667	.00380	.00185	.00009	-.00035	.00380	.00565	.00539	.0054
19	9	3.0000	.00148	.00118	.00009	-.00014	.00148	.00266	.00261	.0026
20	10	3.3333	.00051	.00062	.00007	.00000	.00051	.00115	.00120	.0012
21	11	3.6667	.00016	.00027	.00004	.00005	.00016	.00043	.00052	.0005
22	12	4.0000	.00004	.00010	.00002	.00004	.00004	.00014	.00020	.0002
23	13	4.3333	.00001	.00003	.00001	.00002	.00001	.00004	.00007	.0001
24	14	4.6667	.00000	.00001	.00000	.00001	.00000	.00001	.00001	.0000

TABLE 2.19

In a similar way, we may use Equation (2.22) after putting in the values of the constants to calculate the approximate sum of the ordinates of the Point Binomial between any given limits. Thus,

$$\sigma\int_0^z f(z)\,dz = \frac{1}{\sqrt{2\pi}}\int_0^z e^{-\frac{z^2}{2}}\,dz - .04444443\ (.3989423 + \varphi_2(z))$$

$$+ .002129623\ \varphi_3(z) + .000975171\ \varphi_5(z) + \ldots\ldots \qquad (2.24)$$

Table 2.20 indicates how accurately the integral of the function f(z) gives the proportion of the observations of a Point Binomial distribution lying within a given range. The meaning of column (4) will be explained later.

(1)		(2)	(3)	(4)
Range Limits		$\sigma\int_a^b f(z)\,dz$	$\sum_a^b (.9+.1)^{100}$	$\int_a^b f(x)dx$
X	x			Pearson Type
A to B	a to b	Gram Charlier Series		Curve(Quadrature)
10 - 13	0 - 1σ	.41520	.4249	.4489
10 - 16	0 - 2σ	.52790	.5282	.5732
10 - 19	0 - 3σ	.54670	.5468	.5993
7 - 13	-σ - +σ	.74814	.7590	.7544
4 - 16	-2σ - +2σ	.97076	.9717	.9654
1 - 19	-3σ - +3σ	.99797	.9981	.9966
7 - 19	-σ - +3σ	.87964	.8809	.9040

TABLE 2.20

The values in column (2) were obtained by linear interpolation with first differences in tables of $\varphi_0(z)$, $\varphi_3(z)$, etc.

The values in column (4) were obtained by the Trapezoidal Rule except in the case of the third value where Simpson's Rule was used.

Use of Pearson Type Curve

It can easily be shown[1] that the constants in the Pearson System of curves as defined in the above diagram can be determined in terms of the moments of the givendistribution Y = f(X). The resulting differential equation can then be integrated to give the equation of the given frequency function.

However, for the particular object in hand - that of fitting a Pearson type curve to a Binomial distribution - it will be found much more convenient first to integrate the differential equation and then to determine the constants in this equation by the "method of moments".[2]

It will be shown later that if we take as the differential equation of the given function

$$\frac{dY}{dX} = \frac{Y(X+d)}{b_0 + b_1 X} \qquad (2.25)$$

[1] See next chapter.

[2] To be discussed in the next chapter.

the type of curve resulting from integrating this equation may be expected to fit the Point Binomial distribution.

First transfer the origin to the point (-d,0) by setting

$$x = X + d.$$

Then, the differential equation becomes

$$\frac{dy}{dx} = \frac{xy}{b_0 + b_1(x-d)} = \frac{xy}{b_0' + b_1 x}.$$

Integrating this equation, we find

$$\log y = \frac{x}{b_1} - \frac{b_0'}{b_1^2} \log (b_1 x + b_0') + \log c$$

$$= \frac{x}{b_1} + \log \left\{1 + \frac{b_1}{b_0'} x\right\}^{-\frac{b_0'}{b_1^2}} + \log y_0,$$

where y_0 is a new constant of integration,

$$y = y_0 e^{\frac{x}{b_1}} \left\{1 + \frac{b_1}{b_0'} x\right\}^{-\frac{b_0'}{b_1^2}}$$

Now set[1] $\frac{1}{b_1} = -b$ and $\frac{b_0'}{b_1} = d$. Then the equation of our frequency function becomes

$$y = y_0 e^{-bx} (1 + \frac{x}{d})^{bd}$$

in which the origin of x is at (-d,0), the mode of the frequency distribution $Y = f(X)$.

It is now merely a matter of mathematical detail to determine the constants y_0, b and d. To determine y_0, we have the condition that

$$I = \int_{-d}^{\infty} y_0 e^{-bx} (1 + \frac{x}{d})^{bd} dx = N$$

where N equals the total frequency.

To evaluate I, set

$$1 + \frac{x}{d} = x_1.$$

[1] To avoid introducing a new letter, "d" is used again although its value is not necessarily the same as above.

Then, the integral becomes

$$I = d\, e^{bd}\, y_0 \int_0^{\infty} e^{-bd\, x_1}\, x_1^{bd}\, dx_1$$

and after setting bd $x_1 = x_2$, we have

$$I = \frac{dy_0\, e^{bd}}{(bd)^{bd+1}} \int_0^{\infty} e^{-x_2}\, x_2^{bd}\, dx_2 \,.$$

From Pierce's tables #481, we find

$$I = dy_0 \frac{e^{bd}}{(bd)^{bd+1}} \Gamma(bd+1) = N$$

from which

$$y_0 = \frac{N(bd)^{bd+1}}{d\, e^{bd}\, \Gamma(bd+1)} \,.$$

We shall apply the method of moments to find the remaining constants b and d and to do this, it will be found convenient to obtain first a general expression for the moments of this frequency curve about the start of the curve and then transform these moments, by means of well-known relationships given in the next Chapter, to moments about the mean value of the frequency function.

Denoting by ${}_1\mu_i$ the i^{th} moment of the frequency function about the start of the curve $(-d,0)$, we have

$${}_1\mu_i = \frac{y_0}{N} \int_{-d}^{\infty} e^{-bx} \left(1 + \frac{x}{d}\right)^{bd} (x+d)^i \, dx,$$

which in terms of the variable x_1 becomes

$$\frac{y_0}{N} e^{bd}\, d^{i+1} \int_0^{\infty} e^{-dbx_1}\, x_1^{bd}\, x_1^i\, dx_1,$$

or in terms of the variable x_2 equals

$$\frac{y_0}{N} e^{bd} \frac{d^{i+1}}{(bd)^{bd+i+1}} \int_0^{\infty} e^{-x_2}\, x_2^{bd+i}\, dx_2 \,.$$

Applying again the same formula from Pierce's tables and putting in the value of y_0 already found, we have

$${}_1\mu_i = \frac{\Gamma(bd+i+1)}{b^i\, \Gamma(bd+1)} \,.$$

Hence, the first three moments about the start of this curve are

$$ {}_1\mu_1 = \frac{1}{b}\frac{\Gamma(bd+2)}{\Gamma(bd+1)} = \frac{bd+1}{b}, $$

$$ {}_1\mu_2 = \frac{1}{b^2}\frac{\Gamma(bd+3)}{\Gamma(bd+1)} = \frac{1}{b^2}(bd+2)(bd+1), $$

$$ {}_1\mu_3 = \frac{1}{b^3}\frac{\Gamma(bd+4)}{\Gamma(bd+1)} = \frac{1}{b^3}(bd+3)(bd+2)(bd+1). $$

Denoting by μ_i the moments of the frequency curve about the mean of the distribution, we have the well-known relationships to be developed later,

$$ \mu_1 = 0, $$

$$ \mu_2 = {}_1\mu_2 - {}_1\mu_1{}^2, $$

$$ \mu_3 = {}_1\mu_3 - 3\,{}_1\mu_1\,{}_1\mu_2 + 2\,{}_1\mu_1{}^3. $$

From these last equations, we have at once

$$ \mu_2 = \frac{bd+1}{b^2} \text{ and} $$

$$ \mu_3 = \frac{2(bd+1)}{b^3}. $$

Hence, the simultaneous equations which determine b and d are

$$ b^2\,\mu_2 - bd-1 = 0, $$

$$ b^3\,\mu_3 - 2bd-2 = 0, $$

from which

$$ b = \frac{2u_2}{\mu_3} \text{ and} $$

$$ d = \frac{2u_2{}^2}{\mu_3} - \frac{\mu_3}{2u_2}. $$

To fit this Pearson curve to the Point Binomial, we must substitute for the moments of the frequency function, the like moments of the Point Binomial $(q+p)^n$, and these moments are given by

$$ \sigma^2 = \mu_2 = pqn, $$

$$ \mu_3 = pqn\,(q-p) = k\sigma^3, $$

and therefore

$$ b = \frac{2}{k\sigma} \text{ and} $$

$$ d = \frac{2\sigma}{k} - \frac{k\sigma}{2}. $$

Hence, we have for the final equation of the fitted curve the following

$$y = \frac{N\left(\frac{4}{k^2}-1\right)^{\frac{4}{k^2}} e^{-\frac{4}{k^2}+1}\, 2k}{\sigma\,(4-k^2)\,\Gamma\left(\frac{4}{k^2}\right)} \left[e^{-\frac{2}{k\sigma}x}\left(1+\frac{x}{\frac{2\sigma}{k}-\frac{k\sigma}{2}}\right)^{\frac{4}{k^2}-1}\right].$$

To fit this distribution to the Point Binomial $(.9+.1)^{100}$ we recall that in the differential equation form of the Pearson Curve with which we started, the mode of the curve was at

$$X = -d$$

or at $x = 0$ in the x coordinate.

Now the modal value and also the mean value in this Binomial distribution is

$$pn = 10$$

and this value will therefore correspond to $x = 0$ in the fitted curve. With this fact in mind, Table 2.21 shows a convenient form for computing the ordinates of the fitted curve and for comparison with the actual ordinates of the Point Binomial.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
X (Of Pinomial)	x (Of Frequency Curve)	$1+\frac{x}{\frac{2\sigma}{k}-\frac{k\sigma}{2}}$	log(3)	$\left(\frac{4}{k^2}-1\right)(4)$	$\frac{-2}{k\sigma}x$	(6)$\log_{10}e$	log y = (log y_0) + (5) + (7)	y	$(.9+.1)^{100}$
0	-10	.54751	$\overline{1}.7383921$	$\overline{15}.5461635$	25.0	10.8573625	$\overline{5}.5305549$	.000034	.0000
1	- 9	.59276	$\overline{1}.7728789$	$\overline{13}.4515592$	22.5	9.7716263	$\overline{4}.3602144$	.000229	.0003
2	- 8	.63801	$\overline{1}.8048275$	$\overline{11}.2167194$	20.0	8.6858900	$\overline{3}.0296383$	.001071	.0016
3	- 7	.68326	$\overline{1}.8345860$	$\overline{10}.8608765$	17.5	7.6001538	$\overline{3}.5880592$	.003873	.0059
4	- 6	.72851	$\overline{1}.8624355$	$\overline{8}.3995614$	15.0	6.5144175	$\overline{2}.0410078$	.010990	.0159
5	- 5	.77376	$\overline{1}.8886063$	$\overline{7}.8454981$	12.5	5.4286813	$\overline{2}.4012083$	.025189	.0339
6	- 4	.81900	$\overline{1}.9132839$	$\overline{5}.2089355$	10.0	4.3429450	$\overline{2}.6789094$	.047743	.0596
7	- 3	.86425	$\overline{1}.9366394$	$\overline{4}.4993269$	7.5	3.2572088	$\overline{2}.8835646$	.076483	.0889
8	- 2	.90950	$\overline{1}.9588027$	$\overline{3}.7238492$	5.0	2.1714725	$\overline{1}.0223506$	.105281	.1148
9	- 1	.95475	$\overline{1}.9798897$	$\overline{2}.8889059$	2.5	1.0857363	$\overline{1}.1016711$	.126378	.1304
10	0	1.00000	0.0000000	0.0000000	0	0.0000000	$\overline{1}.1270289$	.133977	.1319
11	1	1.04525	0.0192202	1.0619161	-2.5	-1.0857363	$\overline{1}.1032087$	.126826	.1199
12	2	1.09050	0.0376257	2.0788199	-5.0	-2.1714725	$\overline{1}.0343763$	.108237	.0988
13	3	1.13575	0.0552828	3.0543747	-7.5	-3.2572088	$\overline{2}.9241948$	.083984	.0743
14	4	1.18100	0.0722499	3.9918070	-10.0	-4.3429450	$\overline{2}.7758909$	.059689	.0513
15	5	1.22624	0.0885755	4.8937964	-12.5	-5.4286813	$\overline{2}.5921440$	.039097	.0327
16	6	1.27149	0.1043130	5.7632933	-15.0	-6.5144175	$\overline{2}.3759047$	.023763	.0193
17	7	1.31674	0.1195000	6.6023750	-17.5	-7.6001538	$\overline{2}.1292501$	.013466	.0106
18	8	1.36199	0.1341739	7.4131080	-20.0	-8.6858900	$\overline{3}.8542469$	.007149	.0054
19	9	1.40724	0.1483682	8.1973431	-22.5	-9.7716263	$\overline{3}.5527457$	.003571	.0026
20	10	1.45249	0.1621131	8.9567488	-25.0	-10.8573625	$\overline{3}.2264152$	.001684	.0012
21	11	1.49774	0.1754364	9.6928611	-27.5	-11.9430988	$\overline{4}.8767912$	.000753	.0005
22	12	1.54299	0.1883631	10.4070613	-30.0	-13.0288350	$\overline{4}.5052552$	.000320	.0002
23	13	1.58824	0.2009161	11.1006145	-32.5	-14.1145713	$\overline{4}.1130721$	.000130	.0001
24	14	1.63348	0.2131138	11.7745375	-35.0	-15.2003075	$\overline{5}.7012589$	.000050	.0000
25	15	1.67873	0.2249809	12.4301947	-37.5	-16.2860438	$\overline{5}.2711798$	.000019	.0000
26	16	1.72398	0.2365323	13.0684096	-40.0	-17.3717800	$\overline{6}.8236585$	.000007	.0000
27	17	1.76923	0.2477843	13.6900826	-42.5	-18.4575163	$\overline{6}.3595952$	.000002	.0000
28	18	1.81448	0.2587522	14.2960591	-45.0	-19.5432525	$\overline{7}.8798355$	.000001	.0000
29	19	1.85973	0.2694499	14.8871070	-47.5	-20.6289888	$\overline{7}.3851471$	.000000	.0000
30	20	1.90498	0.2798904	15.4639446	-50.0	-21.7147250	$\overline{8}.8762485$	.000000	.0000

$p = .1, \quad q = .9$ $\qquad$ $n = 100, \quad \sigma = 3$ $\qquad$ $k = .266667$

$\frac{2\sigma}{k} - \frac{k\sigma}{2} = 22.1$ $\qquad$ $\frac{4}{k^2} - 1 = 55.25$ $\qquad$ $-\frac{2}{k\sigma} = -2.5$

$y_0 = .133977$ $\qquad$ $\log y_0 = \overline{1}.1270289$ $\qquad$ $N = 1$

TABLE 2.21

To obtain the integral of this Pearson curve between the limits already considered, in the case of the Gram Charlier Series, we must make use of quadrature formulas, since no ready means of integrating this function between finite limits is available.

Applying the suitable quadrature formula to be used in each case, we find approximate expressions for the integral of this function taken between corresponding ranges as given in column (4) of Table 2.20.

6. Presentation of Information in Terms of Higher Moments

To one covering this ground for the first time, it must come as somewhat of a surprise to see how much information is actually tied up in the two simple functions or statistics, the average and the standard deviation of a distribution of numbers. To the same individual it must come as somewhat of a shock to see that the use of higher moments does not in all cases give us a proportional increase in the amount of information. This same individual now begins to appreciate the significance of the statement made in the previous chapter that the average and standard deviation are perhaps the two most useful statistics by which to present the information contained in a series of numbers. We have gone far enough to see, however, that under certain conditions practically the entire amount of information contained in a series of data can be presented by the first four moments, although in such cases it is necessary to know the functional form of the frequency distribution in which these four moments are to be used as estimates of certain parameters occurring in that distribution.

Quite naturally it is impossible for us to do more than briefly indicate in outline, as we have done, the more important methods for making use of frequency curves which in turn fix certain parameters to be estimated from the data so as to contain the greater part of the total information in the original set of data. To give a detailed treatment of the manner of using either one of the two general methods of presenting the information contained in a set of data through the use of moments interpretable in a given frequency curve would require a good sized book. Enough has been said, however, to show definitely the rather remarkable power of the method in certain examples.

Thus we have seen how closely the information contained in a Point Binomial distribution can be represented in terms of the first four moments for the particular case $(.9+.1)^{100}$. Similarly, it will be interesting to observe the frequency curves presented in Fig. 2.13, showing how accurately the total

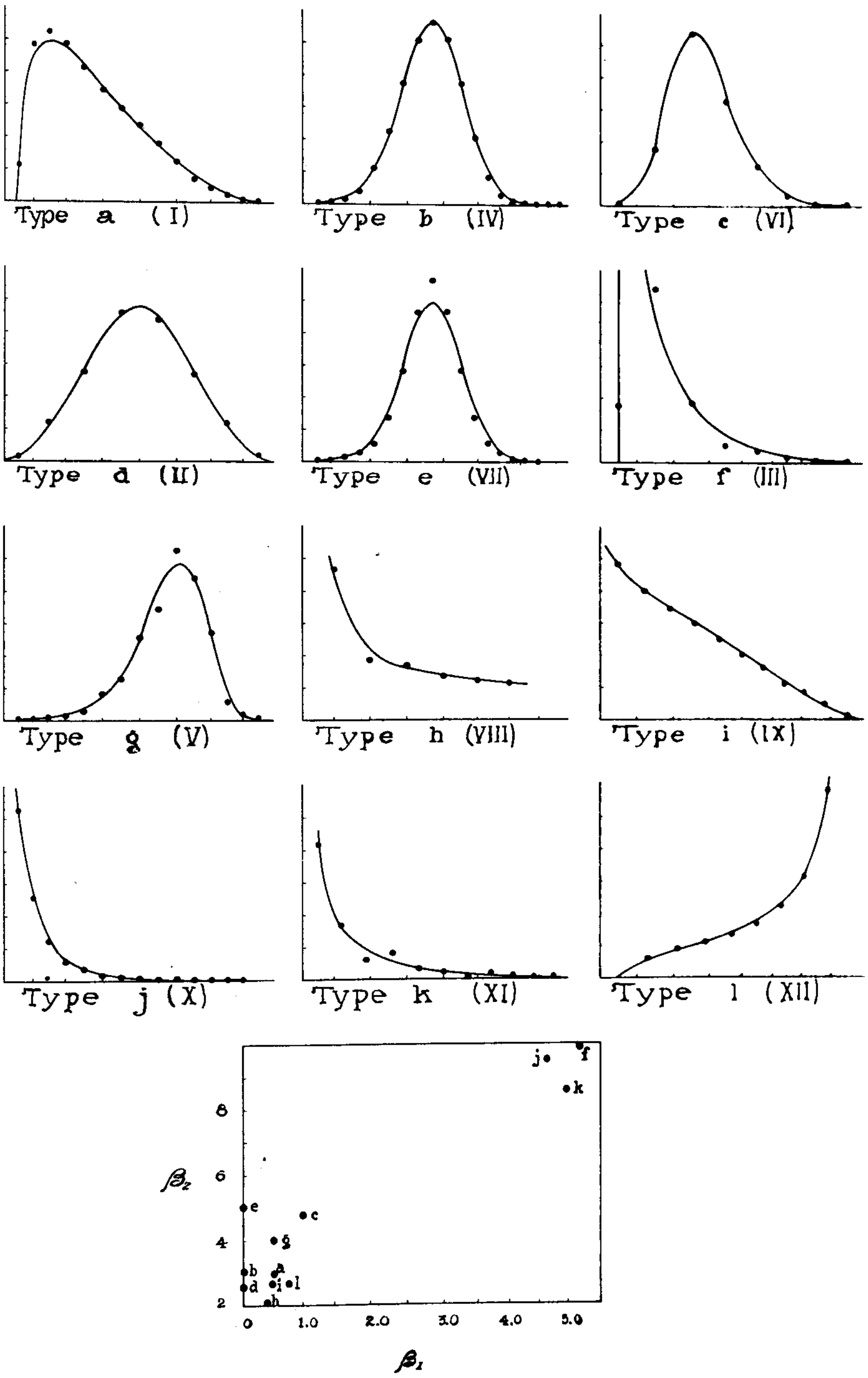

Type a (I)
Type b (IV)
Type c (VI)
Type d (II)
Type e (VII)
Type f (III)
Type g (V)
Type h (VIII)
Type i (IX)
Type j (X)
Type k (XI)
Type l (XII)
β_2
8
6
4
2
0
1.0
2.0
3.0
4.0
5.0
β_1
j
f
k
e
c
g
b
a
d
i
l
h

FIG. 2.13

information is presented by the first four moments of an observed set of data and the appropriate Pearson type curve for the particular distribution there presented. The solid lines in this figure show the Pearson type curve fitted to the observed points, represented by the black dots in the figure.

These curves were drawn from data presented by Elderton in his book, "Frequency Curves and Correlation". In the lower part of the figure are shown the points on the $\beta_1\beta_2$ plane (where β_1 is the square of the skewness k). In other words, knowing only β_1 and β_2 we would be led by Pearson's method to the choice of the frequency curve appropriate for representing the total information contained in the set of data to the degree of approximation shown graphically in this figure.

It is no small trick, however, to go through the necessary technical details required in fitting one of these frequency curves. Furthermore, past experience indicates that an engineer who attempts to follow what may appear to be such a universal method may meet with some disappointment when he finds that many distributions observed in engineering work have not been successfully fitted by means of any one of the given types of frequency curves. What then do these results give us definitely in respect to our present problem of presenting engineering information in terms of simple functions or statistics calculated from the set of n observed values of some quality X?

First of all we must keep in mind the significance of the average and standard deviation derivable through the use of Tchebycheff's theorem. If we are to go beyond this point in the presentation of facts through the use of simple statistics, it is necessary that we give some kind of a frequency curve involving these statistics as parameters provided we are interested in using these statistics for the purpose of presenting as much as possible of the total information contained in the original series of observations. The practice of giving the third and fourth moments in addition to the first and second without any indication of the frequency distribution which should be used therewith is perhaps of little value so far as the presentation of total information is concerned. It would be better in such cases to present the original observed distribution.

It may not be out of place to mention here that in Part III we shall find that, in general, k and β_2 are of comparatively little value when the sample size is small. On the other hand, we shall find there that k and β_2 may contain a comparatively large amount of the essential information even though no information is available as to the form of frequency function in which these statistics are used as parameters.

CHAPTER V

Presentation of Data to Show Relationship

1. Nature of the Problem

Up to the present time we have been concerned primarily with the presentation of the information contained in an observed series of data representing some quality characteristic X. In many problems in engineering work, however, we are interested not only in this phase of the subject but also in the study of the relationship between different observed characteristics, and it is this problem which we shall now consider.

2. The Concept of Mathematical Functional Relationship

Before entering upon the discussion of the problem involved in studying the relationships between quality characteristics, it is perhaps well for us to review briefly some of the fundamental concepts underlying our picture of mathematical relationship. For example, we say that Y is a function of X when for every value of X within a given domain for which the function is defined the value or values of Y are fixed.

In general a functional relationship involves certain constants or parameters and therefore may be written symbolically

$$Y = f(X, \lambda_1, \lambda_2, \ . \ . \ . \ . \ , \lambda_m),$$

the λ's being the parameters. In fact, as we have already seen and as will be further amplified, the information contained in the observed data is to be presented in the form of estimates of parameters which occur in the type of functional relationship to be used.

As a very simple case, we may take the relationship

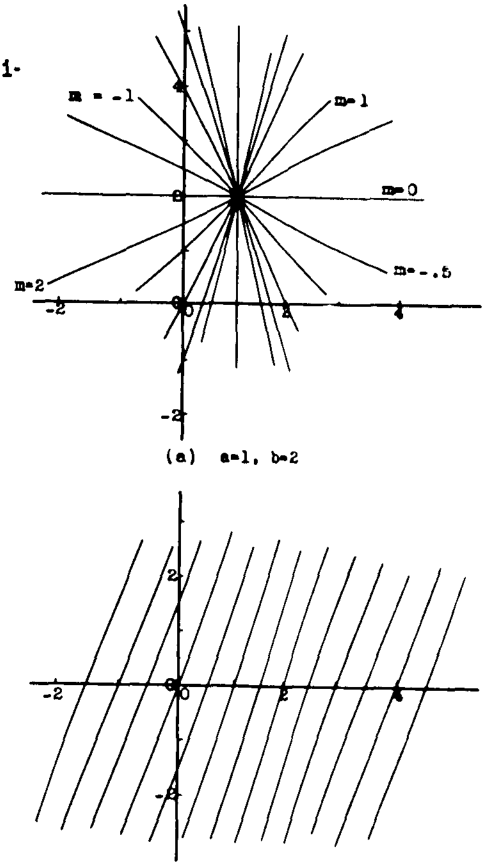
$$Y - b = m(X - a).$$

Obviously the graph of this relationship is a straight line passing through the point X = a, Y = b. If we take as a special case a = 1, b = 2 and give m all possible values, we get the pencil of lines through the point (1,2) illustrated in Fig. 2.14.

In a similar way if we put the equation of the straight line into the form

$$Y = mX + (b - ma)$$

and give m the same constant value, while the point (a,b) is assigned arbitrarily, we obtain a family of parallel lines through the point (a,b). As a special case, if we let m = 3, we get the set of parallel lines shown also in Fig. 2.14



(a) a=1, b=2

(b) m=3

FIG. 2.14 - GRAPH OF RELATIONSHIP
$Y - b = m(X - a)$

If we pass now to a slightly more complicated case, namely, the second degree parabola expressed by a functional relationship

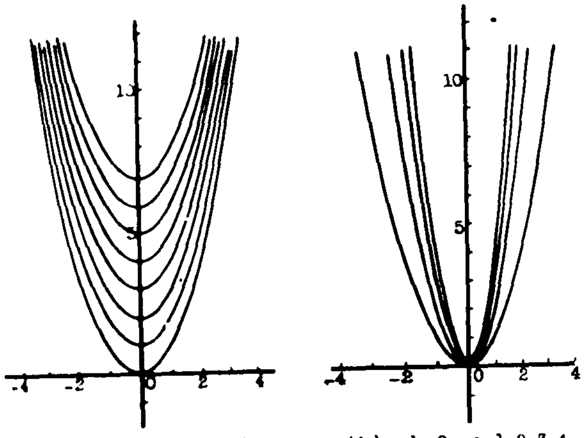
$$Y - b = aX^2,$$

we can also illustrate the significance of the parameters in this equation. For example, if we fix a and assign to b different values, we get a set of parabolas such as shown at the left, Fig. 2.15. In a similar way if we fix the value of b, and for example let it be 0 and change a, we get the set or family of parabolas shown at the right of the figure.

Enough has been said to illustrate the well known fact that for every functional form f there are, in general, a very large number of curves constituting, as we say, a family of curves, a single one of which is specified when the parameters involved in the functional form are fixed. On the other hand, all of the curves belonging to a given family have certain characteristics which are common to all and this commonness among the curves is defined by the function f.

From our present viewpoint we see then that in order to express a relationship between two or more variables, we must consider two things:

1. The form of the functional relationship.
2. The specific values of the parameters in that functional relationship.



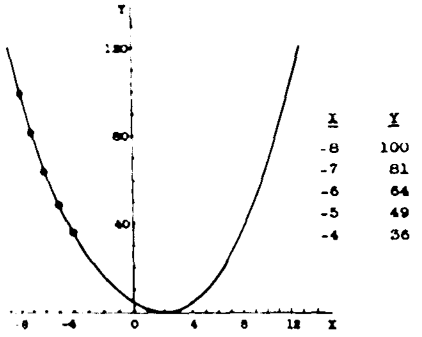
(a) a=1, b=0,1,...7 (b) b=0, a=1,2,3,4

FIG. 2.15 - GRAPH OF RELATIONSHIP $Y-b=aX^2$

Whenever one is given a mathematical relationship between two variables, he can of course, as already stated, calculate the value of one when the other is given. For example, if we take the parabola

$$Y = 4 - 4X + X^2,$$

we can calculate the value of Y corresponding to any given value of X and then plot these results to give the graphical picture of the relationship between Y and X. This familar process is illustrated below.



X	Y
-8	100
-7	81
-6	64
-5	49
-4	36

FIG. 2.16

What has been said in respect to the relationship between two variables can easily be extended to three or more. For example, if

$$Z = f(X,Y)$$

we say, in general, that for given values of X and Y, Z is determined.

3. Law of Relationship

Engineers in their everyday work frequently make use of laws or relationships between variable quality characteristics. In fact, one of the main objects of physical and engineering science is to discover and make use of such relationships.

A very simple *a priori* law is

$$s = \frac{1}{2}at^2, \tag{2.26}$$

where s is the space covered in the time t by a body falling freely from a position of rest. Almost everyone who has taken elementary physics has had the experience of estimating the value of a from simultaneously observed pairs of values of s and t. Equation (2.26) is the simple law of falling bodies.

In a similar way we have the law relating the force F acting on a mass m and the acceleration a produced in the velocity of the mass,

$$F = ma. \tag{2.27}$$

Underlying these conceptions of laws we have the idea of a functional relationship previously described. Such laws as those given by (2.26) and (2.27) assume that the relationship is a mathematically continuous one over certain possible ranges, even though it is obviously not possible to check experimentally such an assumption by measurement. In other words, we have here the introduction of the concept of a theoretical law which exists even though we can never prove by experiment that it does exist. In practice we can never do more than obtain an estimate of the parameter, in this case a, included in the law.

Over and against this theoretical law we have the empirical law, one of the simplest examples being that expressing the relationship between the length L_0 of a bar at an initial temperature t_0 and the length L_1 at a higher temperature t_1. Customarily this relationship is assumed to be given by the equation

$$L_1 = L_0(1 + \alpha t), \qquad (2.28)$$

where

$$t = t_1 - t_0.$$

In this case, however, it is necessary to note that the value of α is sufficiently accurate only for a small range of temperature t and holds for a temperature in the neighborhood[1] of 20° C.

To cover wider ranges of temperature, we need other formulae. Sometimes, for example, we find the expression

$$L_1 = L_0(1 + \alpha_1 t + \alpha_2 t^2 + \cdot \cdot \cdot)$$

4. Information Given by a Set of Data

First of all let us note the rôle played by the parameters in a given empirical law. In the one just described, α is proportional to the slope of the straight line (2.28) and in general this parameter is different for the different materials. For example, Kaye and Laby[2] give the following series of values of α.

- -

[1] In general, the coefficient α increases with the temperature but extrapolation of α may be unsound since some substances expand irregularly. Interpolation of α from constituent metals must be employed with caution in the case of alloys.

[2] G.W.C.Kaye and T.H.Laby, "Physical and Chemical Constants and Some Mathematical Functions".

Element	$\frac{\alpha}{x\ 10^{-6}}$	Element	$\frac{\alpha}{x\ 10^{-6}}$	Element	$\frac{\alpha}{x\ 10^{6}}$
Aluminium	25.5	Gold	13.9	Potassium	83
Antimony	12	Iridium	6.5	Selenium	36.8
Bismuth	15.7	Iron (cast)	10.2	Silver	18.8
C. (diamond)	1.2	" (wrought)	11.9	Sodium	75
" (gas carbon)	5.4	Steel, 10.5 to	11.6	Sulphur	c. 70
" (graphite)	7.9	Lead	27.6	Thallium, 40°	30.2
Cadmium	28.8	Magnesium	25.4	Tin	21.4
Cobalt	12.3	Nickel	12.8	Tungsten, 27°	4.44
Copper	16.7	Palladium	11.7	" 2027°	7.26
		Platinum	8.9	Zinc, 25.8 to	26.3

TABLE 2.22 - ILLUSTRATIVE TABLE:

How Useful Physical Information is Presented in Forms of Simple Functions Used as Estimates of Parameters in Assumed Relationships

Here we see that the linear function expresses a relationship which is approximately true for all the elements, whereas the parameters differ from one to the other.

From what precedes we see that if the law of functional relationship is known a priori, the only thing that experimental work is supposed to give is the numerical values of the parameters. For example, if it were true that the length of a specimen of material at any given temperature varied in a linear manner with temperature as indicated by (2.28), then all the information that an observed set of data consisting of pairs of values of lengths and temperatures could give would be tied up in the parameter α, technically termed the coefficient of linear expansion.

If the law of relationship is unknown, experimental work must give us an empirical relationship f and also estimates of the parameters involved in this relationship. In such a case all the information is not tied up in the estimates of the parameters unless these estimates are so chosen in terms of simple functions of the data that they may be used equally well in other possible relationships. This point will be stressed in a later paragraph, where it will be shown that the simple statistics $\bar{X}$, $\bar{Y}$, σ_x, σ_y and r are of much more value in this respect than many of those ordinarily tabulated. Before we can take up this point, however, we must consider some of the available ways and means of estimating the parameters.

5. Finding Estimates of Parameters to Express Information

In general, let us write the law of relationship between two variables X and Y in the form

$$Y = f(X, \lambda_1, \lambda_2, \ldots, \lambda_m), \qquad (2.29)$$

where the λ's are the parameters to be estimated from n observed pairs of values of X and Y. We shall consider four methods for estimating these parameters.

(A) Direct Calculation

If the number n of different pairs of observations is equal to or greater than the number m of parameters, we may use any set of m different pairs of values, say, X_1, Y_1; X_2, Y_2; . . . ; X_m, Y_m; and form the set of simultaneous equations

$$\begin{aligned} Y_1 &= f(X_1, \lambda_1, \lambda_2, \ldots, \lambda_m) \\ Y_2 &= f(X_2, \lambda_1, \lambda_2, \ldots, \lambda_m) \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ Y_m &= f(X_m, \lambda_1, \lambda_2, \ldots, \lambda_m), \end{aligned}$$

the solution of which will give one set of estimates $\theta_{11}, \theta_{12}, \ldots, \theta_{1m}$ of the m parameters. Now we may choose any other set of m pairs of different values of X and Y and solve in a similar way for a second set of estimates of the parameters, and so on. If all of the pairs of values happen to satisfy a single one of the family of curves represented by the assumed function, then the different sets of estimates will be identical. It would be, however, a very rare thing to have an observed set of n points that satisfy a single equation involving m parameters where $m < n$.

Obviously we would get, by such a method, $n_1 = \frac{\underline{|n}}{\underline{|n-m}\;\underline{|m}}$ sets of m parameters

$$\begin{aligned} &\theta_{11}, \theta_{12}, \cdots \theta_{1i}, \cdots \theta_{1m} \\ &\theta_{21} \;\; \theta_{22}, \cdots \theta_{2i}, \cdots \theta_{2m} \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ &\theta_{n_1 1}, \theta_{n_1 2}, \cdot\cdot\ \theta_{n_1 i}, \cdot\cdot\ \theta_{n_1 m}, \end{aligned}$$

where, in general, all of the sets will be different.

This method does not provide any criterion by which to choose the best set of statistics or estimates of the parameters from among the possible n_1 different sets, and even if it did, it would still have the disadvantage that any particular set makes use of the information contained in only m of the n pairs of observed values.

Let us illustrate these points by a simple example, taking from Table 2.2 the first six pairs of values of voltage and current: 3, .03; 6, .07; 9, .11; 12, .15; 15, .19; 18, .24.

Let us assume that the relationship between Y and X is given by the expression

$$Y = a_0 + a_1X + a_2X^2,$$

where in a particular case we have replaced the generalized λ's by the a's. To obtain estimates of the a's, we need only three pairs of values of X and Y. We may take, for example,

$$.03 = a_0 + 3a_1 + 9a_2$$
$$.07 = a_0 + 6a_1 + 36a_2$$
$$.11 = a_0 + 9a_1 + 81a_2 \ .$$

From these three equations we get

$$a_0 = -.01000$$
$$a_1 = +.01333$$
$$a_2 = +.00000 \ .$$

But in a similar way we get from the next three pairs of values

$$a_0 = +.09000$$
$$a_1 = -.00167$$
$$a_2 = +.00056 \ .$$

Similarly, we would get estimates of these three a's for every set of three values of X and Y, in all $\frac{\underline{|17}}{\underline{|14}\ \underline{|3}}$ or 680 sets of estimates! We have no way of deciding upon any one of the 680 sets of estimates of a_0, a_1 and a_2 in preference to any other. To give all of the sets we would have to calculate and then tabulate a total of 680 x 3 = 2040 estimates of the a's. Manifestly it would be much better to give merely the original 17 pairs of observed values of X and Y.

It is obvious, therefore, that this method of attack is not very desirable and is clearly of little value. We pass, therefore, to a consideration of other methods of finding estimates of the parameters.

(B) Graphical Method

In what follows we shall assume various forms for the functional relationship between X and Y and indicate in each case how to obtain graphical estimates of the parameters.

(a) $$Y = a_0 + a_1X .$$

In this case draw through the observed points the straight line that appears to fit them best. This process, of course, determines both the slope a_1 of the line and the Y-intercept a_0.

(b) $$Y = a_0 + a_1X + a_2X^2 .$$

The derivative of this function is given by

$$\frac{dY}{dX} = a_1 + 2a_2X$$

and, therefore, the ratio of ΔY to ΔX should plot against ΔX as an approximately straight line.

Hence, putting in the line that appears to fit this set of points best, we determine as above in (a) the parameters a_1 and a_2. To obtain an estimate of a_0 it is customary practice to find a value of a_0 corresponding to each of the n pairs of values of X and Y and then to take the average of these n estimates as the "best" estimate of a_0.

(c) $$Y = \frac{a_0 + a_1X}{X} .$$

In this case Y plotted against $\frac{1}{X}$ should give a straight line, since, if we let $X = \frac{1}{X_1}$, Y = Y, we get the form

$$Y = a_1 + a_0X_1$$

which represents a straight line.

(d) $$Y = \frac{a_0X}{1 + a_1X} .$$

Setting $Y_1 = \frac{X}{Y}$, X = X, we are led again to a linear relationship between Y_1 and X.

(e) $$Y = \frac{a_0X^2}{1 + a_1X^2} .$$

The plot of Y to $X_1 = \frac{Y}{X^2}$ is a straight line.

In cases (c), (d) and (e) the method of finding the two parameters in the resulting linear equation is the same as that used in (a).

(f) $$Y = a_0c^{a_1X} ,$$

where c is a known constant. Here

$$\log Y = \log a_0 + a_1X \log c$$

and hence the graph of Y against X on semi-logarithmic paper is a straight line from which a_0 and a_1 may be determined as before.

(g) $$Y = a_0X^{a_1} .$$

Here

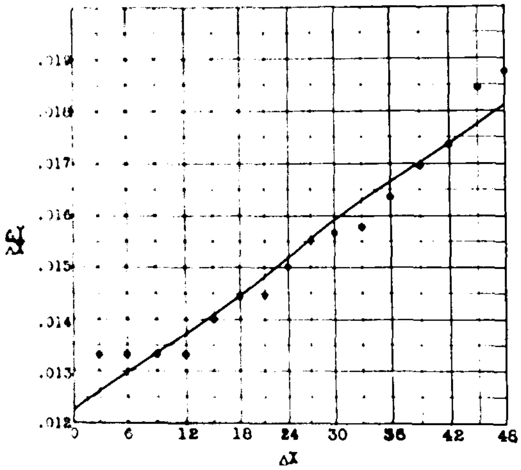
$$\log Y = \log a_0 + a_1 \log X$$

and the graph of Y against X on double logarithmic paper is a straight line, the parameters of which can be found as above.

Of course, these are only a few of the simplest ways in which the unknown parameters may be estimated. We shall now consider a numerical illustration.

Again we shall take the data relating the current Y in amperes to the voltage X in volts. The data sheet of Fig. 2.17 shows the necessary steps in the calculation of the parameters a_0, a_1 and a_2 in the assumed relationship

$$Y = a_0 + a_1X + a_2X^2 .$$



Voltage X	Current Y	ΔX	ΔY	$\frac{\Delta Y}{\Delta X}$	$.01225X$ $.0001222X^2$	$a_0 = Y - .01225X$ $- .0001222X^2$	Y
3	.03	0	0	------	.037850	-.007850	.013386
6	.07	3	.04	.013333	.077899	-.007899	.053435
9	.11	6	.08	.013333	.120148	-.010148	.095684
12	.15	9	.12	.013333	.164597	-.014597	.140133
15	.19	12	.16	.013333	.211245	-.021245	.186781
18	.24	15	.21	.014000	.260093	-.020093	.235629
21	.29	18	.26	.014444	.311140	-.021140	.286676
24	.34	21	.31	.014762	.364387	-.024387	.339923
27	.39	24	.36	.015000	.419834	-.029834	.395370
30	.45	27	.42	.015555	.477480	-.027480	.453016
33	.50	30	.47	.015666	.537326	-.037326	.512862
36	.55	33	.52	.015758	.599371	-.049371	.574907
39	.62	36	.59	.016389	.663616	-.043616	.639152
42	.69	39	.66	.016923	.730061	-.040061	.705597
45	.76	42	.73	.017381	.798705	-.038705	.774241
48	.86	45	.83	.018444	.869549	-.009549	.845085
51	.93	48	.90	.018750	.942592	-.012592	.918128

$$\Sigma = -.415893$$
$$\div 17 = -.024464 = a_0$$

$$Y = -.024464 + .01225X + .0001222X^2$$

FIG. 2.17

In using this method the values of a_0, a_1 and a_2 derived from the data depend upon the straight line drawn so as to "best" fit the points plotted and in the absence of any criterion by which to determine this "best" fit, there probably would result almost as many different assumed lines as there are independent trials at drawing the line. This is obviously a serious objection to this method of estimating the parameters.

We shall consider next the method of least squares used extensively by engineers and students of the physical sciences in estimating the parameters in a given equation.

(C) The Method of Least Squares

Suppose as before that we have a set of n pairs of values X and Y, that is, X_1, Y_1; X_2, Y_2; ... ; X_n, Y_n and that these pairs considered as coordinates of points in a plane lie approximately on the curvè

$$Y = f(X_1, \lambda_1, \lambda_2, \ . \ . \ . \ . \ , \lambda_m) \qquad (2.30)$$

The method of least squares, as usually applied, consists in taking those estimates of the parameters in (2.30) which make the sum of the squares $(\sum_{j=1}^{m} v_j^2)$ a minimum, where v_j^2 is defined by the equation

$$v_j^2 = [f(X_j, \lambda_1, \lambda_2, \ . \ . \ . \ . \ , \lambda_m) - Y_j]^2.$$

Let $v^2 = \sum_{j=1}^{n} v_j^2$, then the formal conditions for a minimum are

$$\frac{\partial(v^2)}{\partial\lambda_i} = 0, \ (i = 1, \ 2, \ .. \ , \ m). \qquad (2.31)$$

This system of m equations theoretically may be solved for the m unknown values of the parameters, thus securing a set of m statistics representing the n pairs of values of X and Y. Practically, however, it may not be feasible to solve the set (2.31) of m equations in that form. In such cases an approximation to a series expansion of f(X) may tend to lessen the difficulties.

If $f(X, \lambda_1, \ .. \ , \lambda_m)$ can be expanded in a Taylor's Series about some suitably chosen value $X = X_0$, we may write

$$f(X, \lambda_1, \ .... \ , \lambda_m) = f(X_0, \lambda_1, \ .... \ , \lambda_m) + \left(\frac{df}{dX}\right)_0 (X - X_0)$$

$$+ \frac{1}{\underline{|2}} \left(\frac{d^2f}{dX^2}\right)_0 (X - X_0)^2 + \ ... \qquad (2.32)$$

where $\left(\frac{d^if}{dX^i}\right)_0$ means the value of the i^{th} derivative of f for $X = X_0$. If we think of $f(X, \lambda_1, \ ..., \lambda_m)$ as some unknown but nevertheless definite function which represents the data, then the λ's have fixed values and the various derivatives of f have therefore definite values. This simply means that under rather general conditions we may to a first approximation represent the unknown function by a polynomial

$$Y = a_0 + a_1 X + a_2 X^2 + \cdot\cdot + a_m X^m \text{ .}$$

Assuming thererore as the simplest case that all terms of higher power than tne first may be neglected and measuring X and Y as before from their mean values, we have

$$v^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$

$$= \sum_{i=1}^{n} (y_i^2 + a_0^2 + a_1^2 x_i^2 - 2a_0 y_i - 2a_1 x_i y_i + 2a_0 a_1 x_i)$$

$$= n\sigma_y^2 + na_0^2 + na_1^2\sigma_x^2 - 2a_1 rn\sigma_x\sigma_y$$

where σ_x and σ_y are the functions or statistics introduced in Chapter II-3.

Hence to determine a_0 and a_1, we have

$$\frac{\partial(v^2)}{\partial a_0} = 2a_0 n = 0,$$

$$\frac{\partial(v^2)}{\partial a_1} = 2a_1 n\sigma_x^2 - 2rn\sigma_x\sigma_y = 0,$$

from which

$$a_0 = 0,$$

$$a_1 = r\frac{\sigma_Y}{\sigma_x},$$

and we see that the value of a_1 comes out in terms of our familiar friends σ_x, σ_y and r previously introduced.

Following the same procedure for the case of

$$y = a_0 + a_1 x + a_2 x^2,$$

we get, where the Σ stands for summation from 1 to n,

$$\Sigma a_0 + a_1\Sigma x + a_2\Sigma x^2 = \Sigma y,$$
$$a_0\Sigma x + a_1\Sigma x^2 + a_2\Sigma x^3 = \Sigma xy,$$
$$a_0\Sigma x^2 + a_1\Sigma x^3 + a_2\Sigma x^4 = \Sigma x^2 y.$$

or since $\Sigma x = \Sigma y = 0$, we have

$$na_0 + 0 + na_2\sigma_x^2 = 0$$
$$0 + na_1\sigma_x^2 + na_2\mu_3 = nr\sigma_x\sigma_y,$$
$$na_0\sigma_x^2 + na_1\mu_3 + na_2\mu_4 = \Sigma x^2 y,$$

where μ_i = the ith moment of X about the mean $\bar{X}$.

From the above set of equations, we find

$$a_0 = \frac{\begin{vmatrix} 0 & 0 & n\sigma_x^2 \\ nr\sigma_x\sigma_y & n\sigma_x^2 & n\mu_3 \\ \Sigma x^2 y & n\mu_3 & n\mu_4 \end{vmatrix}}{\Delta},$$

$$a_1 = \frac{\begin{vmatrix} n & 0 & n\sigma_x^2 \\ 0 & nr\sigma_x\sigma_y & n\mu_3 \\ n\sigma_x^2 & \Sigma x^2 y & n\mu_4 \end{vmatrix}}{\Delta},$$

$$a_2 = \frac{\begin{vmatrix} n & 0 & 0 \\ 0 & n\sigma_x^2 & nr\sigma_x\sigma_y \\ n\sigma_x^2 & n\mu_3 & \Sigma x^2 y \end{vmatrix}}{\Delta},$$

where

$$\Delta = \begin{vmatrix} n & 0 & n\sigma_x^2 \\ 0 & n\sigma_x^2 & n\mu_3 \\ n\sigma_x^2 & n\mu_3 & n\mu_4 \end{vmatrix}.$$

First of all we should note that all together eight statistics are required to be calculated from the original set of n pairs of values of X and Y. First we have to calculate the means $\bar{X}$ and $\bar{Y}$ which give the new origin of coordinates. In addition, we must evaluate σ_x, σ_y, r, μ_3, μ_4 and $\Sigma x^2 y$.

Let us apply this method to the determination of the parabola

$$y = a_0 + a_1 x + a_2 x^2$$

for the current versus voltage in the carbon contact example given above. The necessary details of the calculations are given below.

$\bar{X} = 27$

$\bar{Y} = .42176$

$\sigma_x = 14.6969$

$\sigma_y = .272186$

$\mu_3 = 0$

$\mu_4 = 83579.$

$r = .99301$

$n = 17$

$\Sigma x^2 y = 101.43$

$a_0 = -.0348916$

$a_1 = +.0183905$

$a_2 = +.00016153$

$$Y - \bar{Y} = a_0 + a_1(X - \bar{X}) + a_2(X - \bar{X})^2$$

Substituting

$$Y = .008091 + .009667X + .0001615\ X^2$$

(D) <u>Method of Moments</u>

Given a series of n pairs of values of X and Y, the method of moments, as well as the other three methods just described, gives a way of calculating certain simple functions or statistics of the observed data such that they contain much of the information presented in the original series of observations. Briefly the method is based upon the assumption that the quality characteristics X and Y are functionally related. Hence it follows that this function should be satisfied by the observed pairs of values except as these observed values may be influenced by errors of measurement.

We shall consider the application of the method to the simple case where the ordinates are given at discrete points located at equal intervals on the X-axis, such, for example, as we have in the series of observed pairs of values of current through and voltage across a carbon contact discussed in previous chapters. The method of moments may also be applied to find estimates of statistics to represent the information contained in a series of observations of some quality characteristic X. In this case the observed data are supposed to have been grouped into a frequency distribution so that the ordinate Y_i represents the observed number of values of X falling within the ith interval. We shall consider as a special case the representation of the information where the area or observed frequency Y_i is given for each interval of width h, the ordinate Y_i being situated at the midpoint of the interval. The details of carrying out the method for this case differ from those for the problem just preceding. In what follows, therefore, we shall refer to these two problems as Types I and II.

Specifically the method assumes the existence of a continuous function $f_1(X)$ such that, for data of Type I

$$f_1(X_i) = Y_i$$

and for data of Type II

$$\int_{X_i - \frac{h}{2}}^{X_i + \frac{h}{2}} f_1(X)dX = Y_i ,$$

for all given intervals. Moments of the curve which is chosen to represent or "fit" the data are then equated to estimates of like moments found from the hypothetically correct function $f_1(X)$.

A little consideration will show that in (I) there are at least two possible ways of equating moments and that in general, these two do not lead to the same estimates of the unknown parameters.

(a) We may equate moments of the discrete ordinates of the fitted curve $Y = f(X, \lambda_1, \ldots, \lambda_m)$ to like moments of the observed ordinates or

(b) We may equate moments of the fitted curve $Y = f(X, \lambda_1, \ldots, \lambda_m)$ found by integration to estimates of like moments found from the data.

<u>TYPE I(a)</u>

Given a series of equally spaced ordinates

$$Y_1, Y_2, \ldots, Y_n$$

at

$$X_1, X_2, \ldots, X_n$$

to fit a curve $Y = f(X, \lambda_1, \ldots, \lambda_m)$ to these ordinates.

Denoting by N_0, ${}_1v_1$, ${}_1v_2, \ldots, {}_1v_m$, the sum and first m moments of these ordinates about an arbitrary origin, we have

$$\sum_{i=1}^{n} Y_i = N_0,$$

$$\sum_{i=1}^{n} X_iY_i = N_0 \; {}_1v_1,$$

$$\ldots\ldots\ldots\ldots\ldots ,$$

$$\sum_{i=1}^{n} X_i^mY_i = N_0 \; {}_1v_m .$$

Denote by N, ${}_1\mu_1$, ${}_1\mu_2,\ldots,{}_1\mu_m$ the area and first m moments of the curve $f(X, \lambda_1,\ldots \lambda_m)$ to be fitted to the data.

Then
$$\sum_{i=1}^{n} f(X_i, \lambda_1,\ldots,\lambda_m) = N$$

$$\sum_{i=1}^{n} X_i f(X_i,\lambda_1,\ldots,\lambda_m) = N \; {}_1\mu_1$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$.\sum_{i=1}^{n} X_i^m f(X_i,\lambda_1,\ldots,\lambda_m) = N \; {}_1\mu_m$$

The method of determining the parameters $\lambda_1, \lambda_2,\ldots,\lambda_m$ now consists in equating

$$N = N_0,$$

$${}_1\mu_1 = {}_1v_1,$$

$${}_1\mu_2 = {}_1v_2,$$

$$\ldots\ldots\ldots.$$

$${}_1\mu_m = {}_1v_m .$$

From this system of equations we may solve for the λ's in two ways, i.e., by using the m moment equations together with $N = N_0$ or by using the (m-1) moment equations and $N = N_0$ where N is now expressed as a function of the parameters $\lambda_1, \lambda_2,\ldots,\lambda_m$.

If we choose the first of these ways, we have the following equations to solve for the λ's:

$$\left.\begin{array}{l}
\sum_{i=1}^{n} X_i \, f(X_i, \lambda_1,\ldots,\lambda_m) = \sum_{i=1}^{n} X_i \; Y_i \\
\sum_{i=1}^{n} X_i^2 \, f(X_i, \lambda_1,\ldots,\lambda_m) = \sum_{i=1}^{n} X_i^2 \; Y_i \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
\sum_{i=1}^{n} X_i^m \, f(X_i, \lambda_1,\ldots,\lambda_m) = \sum_{i=1}^{n} X_i^m \; Y_i
\end{array}\right\} \qquad (2.33)$$

A simple illustration will serve to make this formal outline clear.

Example: Given the series of data shown in Table 2.23,

X	Y
1	1
2	0
3	1
4	4
5	9
6	16
7	25
8	36

TABLE 2.23

to fit a continuous curve $Y = f(X, \lambda_1, \ldots, \lambda_m)$ to these ordinates in such a way that $f(X_i)$ is a good approximation to Y_i observed.

On plotting the series of values, it will be seen that they appear to follow a curve of parabolic type. Hence, we may choose as our function f to be fitted

$$Y = a_0 + a_1X + a_2X^2$$

where in a specific case, we have as before replaced the λ's by the a's.

Inasmuch as we must evaluate functions of the form

$$\sum_{i=1}^{n} X_i^m (a_0 + a_1X_i + a_2X_i^2)$$

we shall do well to tabulate its value for the n given values of X as shown in Table 2.24.

X	$X(a_0+a_1X+a_2X^2)$	$X^2(a_0+a_1X+a_2X^2)$	$X^3(a_0+a_1X+a_2X^2)$
1	$1(a_0+a_1+a_2)$	$1^2(a_0+a_1+a_2)$	$1^3(a_0+a_1+a_2)$
2	$2(a_0+2a_1+2^2a_2)$	$2^2(a_0+2a_1+2^2a_2)$	$2^3(a_0+2a_1+2^2a_2)$
3	....	....	....
4	....	....	....
5	....	....	....
6	....	....	....
7	....	....	....
8	$8(a_0+8a_1+8^2a_2)$	$8^2(a_0+8a_1+8^2a_2)$	$8^3(a_0+8a_1+8^2a_2)$
Σ	$36a_0+204a_1+1296a_2$	$204a_0+1296a_1+8772a_2$	$1296a_0+8772a_1+61776a_2$

TABLE 2.24

A little study of the formation of this table will show that the sums tabulated at the bottom are simply the sum of the first powers, second powers,...., fifth powers of the integers 1 to 8, multipled by a certain parameter.

In a similar way, we may display the calculation for the right side of Equations (2.33).

X	X^2	X^3	Y	XY	X^2Y	X^3Y
1	1	1	1	1	1	1
2	4	8	0	0	0	0
3	9	27	1	3	9	27
4	16	64	4	16	64	256
5	25	125	9	45	225	1125
6	36	216	16	96	576	3456
7	49	343	25	175	1225	8575
8	64	512	36	288	2304	18432
Σ				624	4404	31872

TABLE 2.25

Hence the Equations (2.33) become in this particular case

$$36a_0 + 204a_1 + 1296a_2 = 624$$

$$204a_0 + 1926a_1 + 8772a_2 = 4404$$

$$1296a_0 + 8772a_1 + 61776a_2 = 31872$$

which on solving will be found to give

$a_0 = 4$, $a_1 = -4$, $a_2 = 1$, and the function sought is therefore

$$f(X) = 4 - 4X + X^2 .$$

On calculating the values of $f(X)$ at $X = 1,2,\ldots,8$ they will be found to correspond exactly with the given Y values. We should hasten to add however, that in general, such a result would have been merely accidental, but in this case was to be expected, since the above values of X and Y were actually computed from the function

$$Y = 4 - 4X + X^2$$

By so doing we may compare below the virtues of the various methods of fitting curves to data since this can only be done in a given instance when the actual law of relationship between X and Y is known.

It may appear from the preceding example that in general, this method of curve fitting is a very good one, simple to apply, and involves no moment corrections of any kind. A little consideration however, will show that the application of this method may be next to impossible in certain instances such for example as would arise if we were to take for our function $f(X, \lambda_1, \ldots, \lambda_m)$ the familiar but simple normal law function

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(X-\bar{X})^2}{2\sigma^2}}$$

involving just two parameters $\bar{X}$ and σ.

Proceeding in the above fashion, we would come out with the two following equations:

$$\frac{X_1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(X_1-\bar{X})^2}{2\sigma^2}} + \frac{X_2}{\sigma\sqrt{2\pi}}\, e^{-\frac{(X_2-\bar{X})^2}{2\sigma^2}} +\ldots+ \frac{X_n}{\sigma\sqrt{2\pi}}\, e^{-\frac{(X_n-\bar{X})^2}{2\sigma^2}} = \sum_{i=1}^{n} X_iY_i$$

and

$$\frac{X_1^2}{\sigma\sqrt{2\pi}}\, e^{-\frac{(X_1-\bar{X})^2}{2\sigma^2}} + \frac{X_2^2}{\sigma\sqrt{2\pi}}\, e^{-\frac{(X_2-\bar{X})^2}{2\sigma^2}} +\ldots+ \frac{X_n^2}{\sigma\sqrt{2\pi}}\, e^{-\frac{(X_n-\bar{X})^2}{2\sigma^2}} = \sum_{i=1}^{n} X_i^2Y_i\,.$$

Now the right hand members of these equations are known constants and the $X_1, X_2, \ldots, X_n$; $X_1^2, X_2^2, \ldots, X_n^2$ are also given values, but where both unknowns occur in the exponents of e, the solution would obviously be at best a difficult task. Such difficulties, of course, will not be encountered by this method if we choose for f(X) as we did above a polynomial

$$Y = a_0 + a_1X + a_2X^2 + \ldots + a_mX^m\,.$$

TYPE I(b)

The function $f(X, \lambda_1, \ldots, \lambda_m)$ to be fitted to the data is to have its moments found by direct integration i.e.,

$$N\ {}_1\mu_m = \int X^m f(X, \lambda_1, \ldots, \lambda_m)dX$$

This means that before we can equate moments, we must estimate the value of moments of the ${}_1\mu_m$ type from the given set of ordinates. In other words, we now bring into play the function $f_1(X)$ representing a continuous curve which we assume passes through the tops of the given ordinates. It is the moments of this curve, defined by

$$N_0\ {}_1\nu_m = \int X^m f_1(X)dX$$

which we must then estimate from a knowledge of X^mY at certain points and equate these to the moments ${}_1\mu_m$ obtained from the function $f(X, \lambda_1, \ldots, \lambda_m)$ to be fitted.

Such estimation requires the use of a suitable quadrature formula, the purpose of such formula being, in general, to obtain an expression for the integral of a given function over a certain range when values of the function are given only at certain isolated points.

The general mode of procedure is still the same as in I(a) except that the summation signs in Equations (2.33) are now replaced by integrals whose limits must be appropriately assigned. Thus,

$$\left.\begin{array}{l} \int_{X_1-\frac{h}{2}}^{X_n+\frac{h}{2}} X\, f(X, \lambda_1,\ldots,\lambda_m)\,dX = \int_{X_1-\frac{h}{2}}^{X_n+\frac{h}{2}} X\, f_1(X)\,dX \\ \int_{X_1-\frac{h}{2}}^{X_n+\frac{h}{2}} X^2 f(X, \lambda_1,\ldots,\lambda_m)\,dX = \int_{X_1-\frac{h}{2}}^{X_n+\frac{h}{2}} X^2 f_1(X)\,dX \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \int_{X_1-\frac{h}{2}}^{X_n+\frac{h}{2}} X^m f(X, \lambda_1,\ldots,\lambda_m)\,dX = \int_{X_1-\frac{h}{2}}^{X_n+\frac{h}{2}} X^m f_1(X)\,dX \end{array}\right) \qquad (2.34)$$

Applying this theory to the set of data given in Table 2.23 and using for f(X), the function to be fitted, the same expression

$$f(X) = a_0 + a_1X + a_2X^2$$

the left side of Equations (2.34) become in this case, (h = 1)

$$\left.\begin{array}{l} \int_{.5}^{8.5} X(a_0 + a_1X + a_2X^2)\,dX = 36a_0 + \frac{614}{3}\,a_1 + 1305a_2 \\ \int_{.5}^{8.5} X^2(a_0 + a_1X + a_2X^2)\,dX = \frac{614}{3}\,a_0 + 1305a_1 + 8874.1a_2 \\ \int_{.5}^{8.5} X^3(a_0 + a_1X + a_2X^2)\,dX = 1305a_0 + 8874.1a_1 + 62858.25a_2 \end{array}\right) \qquad (2.35)$$

On the right of Equations (2.34) we require to estimate

$\int_{.5}^{8.5} X\, f_1(X)\,dX$, $\int_{.5}^{8.5} X^2\, f_1(X)\,dX$, and $\int_{.5}^{8.5} X^3\, f_1(X)\,dX$, when we are given the values of X_iY_i, $X_i^2Y_i$, and $X_i^3Y_i$ for $i = 1,2,3,..8$ as given in Table 2.25.

Applying a suitable quadrature formula[1], we estimate these three integrals to be

630.333333, 4472.766661, and 32581.84996

Hence, the three equations to be solved for a_0, a_1 and a_2 are

$$\begin{aligned} 108a_0 + 614a_1 + 3915a_2 &= 1890.999999 \\ 614a_0 + 3915a_1 + 26622.3a_2 &= 13418.29998 \\ 1305a_0 + 8874.1a_1 + 62858.25a_2 &= 32581.84996 \end{aligned}$$

from which the values of a_0, a_1, a_2 to six decimal places are

$$\begin{aligned} a_0 &= +4.000001 \\ a_1 &= -4.000000 \\ a_2 &= +1.000000 \end{aligned}$$

which agree very closely with the true values.

Before passing on to a discussion of Type II, we shall give a practical example of the kind just considered, that of fitting a smooth curve

$$Y = a_0 + a_1X + a_2X^2$$

to the data of Table 2.2. An outline of the details is shown below.

X Voltage "E"	Y Current "I"	XY	X^2Y	X^3Y
3	.03	.09	.27	.81
6	.07	.42	2.52	15.12
9	.11	.99	8.91	80.19
12	.15	1.80	21.60	259.20
15	.19	2.85	42.75	641.25
18	.24	4.32	77.76	1399.68
21	.29	6.09	127.89	2685.69
24	.34	8.16	195.84	4700.16
27	.39	10.53	284.31	7676.37
30	.45	13.50	405.00	12150.00
33	.50	16.50	544.50	17968.50
36	.55	19.80	712.80	25660.80
39	.62	24.18	943.02	36777.78
42	.69	28.98	1217.16	51120.72
45	.76	34.20	1539.00	69255.00
48	.86	41.28	1981.44	95109.12
51	.93	47.43	2418.93	123365.43
Σ		261.12	10523.70	448865.82

TABLE 2.26

[1] Tracts for Computers No. X by J. O. Irwin, Cambridge University Press, 1923, Page 7.

$$\int_{1.5}^{52.5} X\, f_1(X)\,dX = 783.3558916$$

$$\int_{1.5}^{52.5} X^2\, f_1(X)\,dX = 31595.4238754$$

$$\int_{1.5}^{52.5} X^3\, f_1(X)\,dX = 1349107.022$$

The like moments calculated from the fitted curves are

$$\int_{1.5}^{52.5} X(a_0 + a_1X + a_2X^2)\,dX = 1377a_0 + 48233.25a_1 + \frac{7596909}{4}\,a_2$$

$$\int_{1.5}^{52.5} X^2(a_0 + a_1X + a_2X^2)\,dX = 48233.25a_0 + \frac{7596909}{4}\,a_1 + 79767596.1375a_2$$

$$\int_{1.5}^{52.5} X^3(a_0 + a_1X + a_2X^2)\,dX = \frac{7596909}{4}\,a_0 + 79767596.1375a_1 + 3489832395.5625a_2$$

Hence the equations that determine a_0, a_1 and a_2 are

$$\begin{aligned}
5508a_0 &+ 192933a_1 &+ 7596909a_2 &= 3133.423566\\
192933a_0 &+ 7596909a_1 &+ 319070384.55a_2 &= 126381.6955\\
7596909a_0 &+ 319070384.55a_1 &+ 13959329582.25a_2 &= 5396428.088
\end{aligned}$$

from which we find

$$a_0 = .026486,$$
$$a_1 = .008306,$$
$$a_2 = .000182.$$

Let us pause long enough to consider the significance of the values of a_0, a_1 and a_2 just found. In terms of these we can express relationship between the current Y through and the voltage X across a carbon contact in the following way:

$$Y = .026486 + .008306X + .000182X^2.$$

By applying the method of least squares we previously obtained this relationship in the form

$$Y = .008091 + .009667X + .0001615X^2.$$

Starting with the same set of data in the two cases we thus arrive at two different functional relationships. Both cannot be the correct expression in the sense that they represent the hypothetically true relationship about which we shall hear more in Part III. In fact the two expressions above simply make use of the information contained in the original set of data when presented in terms of certain symmetric functions of the data.

In our previous notation we may write the above two relationships in the form

$$Y = \theta_{11} + \theta_{21} X + \theta_{31} X^2$$

and

$$Y = \theta_{12} + \theta_{22} X + \theta_{32} X^2$$

Now the analyst making use of the method of moments might present the information contained in the original data in terms of the estimates θ_{11}, θ_{21} and θ_{31} whereas the analyst making use of the method of least squares might in a corresponding way make use of θ_{12}, θ_{22} and θ_{32}. The point to be emphasized again is that when information is presented in this way, it contains that information which the analyst <u>puts in</u> by way of method of analysis as well as that which the data give. Unless the analyst can justify his method against all others, the presentation of the information in terms of a particular set of θ's is therefore open to criticism.

It is a simple matter to show that both sets of statistics involve n, $\bar{X}$, $\bar{Y}$, r, σ_x, σ_y together with certain other simple symmetric functions of the data. In so far, therefore, as we use these simple functions for presenting information we are making use, as it were, of a universal language the value of which is independent of the analyst.

TYPE II

In accordance with the general assumption about the function $f_1(X)$ made at the beginning, what we actually observe are

$$Y_1 = \int_{X_1-\frac{h}{2}}^{X_1+\frac{h}{2}} f_1(X)\,dX, \qquad Y_2 = \int_{X_2-\frac{h}{2}}^{X_2+\frac{h}{2}} f_1(X)\,dX, \ldots, \qquad Y_n = \int_{X_n-\frac{h}{2}}^{X_n+\frac{h}{2}} f_1(X)\,dX .$$

Now let[1]

$$Z = \int_{X - \frac{h}{2}}^{X_n + \frac{h}{2}} f_1(X)\, dX,$$

i.e., Z is the sum of the observed frequencies from some arbitrary value X to the last value X_n inclusive.

Thus,

$$Z_{X_1 - \frac{h}{2}} = \int_{X_1 - \frac{h}{2}}^{X_n + \frac{h}{2}} f_1(X) dX, \; Z_{X_2 - \frac{h}{2}} = \int_{X_2 - \frac{h}{2}}^{X_n + \frac{h}{2}} f_1(X) dX, \ldots, Z_{X_n - \frac{h}{2}} = \int_{X_n - \frac{h}{2}}^{X_n + \frac{h}{2}} f_1(X) dX, \; Z_{X_n + \frac{h}{2}} = \int_{X_n + \frac{h}{2}}^{X_n + \frac{h}{2}} f_1(X) dX$$

have respectively the values

$$Y_1 + Y_2 + \ldots + Y_n, \quad Y_2 + \ldots + Y_n, \quad \ldots\ldots\ldots, \quad Y_n, \quad 0.$$

Now what we require is

$$N_0 \; {}_1v_m = \int_{X_1 - \frac{h}{2}}^{X_n + \frac{h}{2}} X^m f_1(X)\, dX, \text{ where}$$

$$N_0 = \int_{X_1 - \frac{h}{2}}^{X_n + \frac{h}{2}} f_1(X)\, dX = Y_1 + Y_2 + \ldots + Y_n$$

by definition.

The purpose, as before, is then to equate the value of these integrals estimated from the observed data to like integrals of the function $f(X, \lambda_1, \ldots, \lambda_m)$ to be fitted.

To get an expression for $N_0 \; {}_1v_m$, we may proceed as follows:
By definition,

$$\frac{dZ}{dX} = - f_1(X).$$

Therefore,

$$N_0 \; {}_1v_m = - \int_{X_1 - \frac{h}{2}}^{X_n + \frac{h}{2}} X^m \frac{dZ}{dX}\, dX = - \int_{X_1 - \frac{h}{2}}^{X_n + \frac{h}{2}} X^m\, d(Z).$$

Integrating by parts, this becomes

[1] The method of treatment is that given by Pearson in Biometrika Volume I, pp. 282-283.

$$N_0\,{}_1v_m = -X^m Z\Big|_{X_1-\frac{h}{2}}^{X_n+\frac{h}{2}} + m\int_{X_1-\frac{h}{2}}^{X_n+\frac{h}{2}} Z\,X^{m-1}\,dX,$$

$$= (X_1-\tfrac{h}{2})^m\,N_0 + m\int_{X_1-\frac{h}{2}}^{X_n+\frac{h}{2}} Z\,X^{m-1}\,dX.$$

To evaluate the integral appearing in this last expression, we have given the values of $Z\,X^{m-1}$ at intervals of h, from $X_1-\frac{h}{2}$ to $X_n+\frac{h}{2}$ and it is now only necessary to apply again a suitable quadrature formula[1]. The moments of the fitted curve $f(X, \lambda_1, \ldots, \lambda_m)$ are of course obtained by direct integration.

A simple illustration may help to clarify the formal details just given.

Example: Let us take again the set of data given in Table 2.23 and as before let the equation of the curve to be fitted be

$$Y = a_0 + a_1X + a_2X^2.$$

For purposes of illustration of this process, we shall now assume that the values of Y given in Table 2.23 actually represent the frequency or area of observations lying within the unit interval of which the corresponding value of X is the midpoint.

The necessary details involved in the computation of ${}_1v_m$ up to the point of substitution into the quadrature formula are given in Table 2.27.

Y	X	$X-\frac{h}{2}$	$(X-\frac{h}{2})^2$	$Z_{X-\frac{h}{2}}$	$(ZX)_{X-\frac{h}{2}}$	$(ZX^2)_{X-\frac{h}{2}}$
		1/2	1/4	92	46.00	23.00
1	1					
		3/2	9/4	91	136.50	204.75
0	2					
		5/2	25/4	91	227.50	568.75
1	3					
		7/2	49/4	90	315.00	1102.50
4	4					
		9/2	81/4	86	387.00	1741.50
9	5					
		11/2	121/4	77	423.50	2329.25
16	6					
		13/2	169/4	61	396.50	2577.25
25	7					
		15/2	225/4	36	270.00	2025.00
36	8					
		17/2	289/4	0	000.00	0000.00

TABLE 2.27

[1] Tracts for Computers Loc. cit. page 6.

The specific expressions for the moments ${}_1v_1$, ${}_1v_2$, and ${}_1v_3$ are

$$N_{0\,1}v_1 = .5(92) + \int_{.5}^{8.5} Z\,dX$$

$$N_{0\,1}v_2 = (.5)^2(92) + 2\int_{.5}^{8.5} Z\,X\,dX$$

$$N_{0\,1}v_3 = (.5)^3(92) + 3\int_{.5}^{8.5} Z\,X^2\,dX,$$

and the values of Z, ZX and ZX^2 are given from .5 to 8.5 in Table 2.27.

Hence, applying the quadrature formula we find

$$N_{0\,1}v_1 = 627.333372$$

$$N_{0\,1}v_2 = 4455.711156$$

$$N_{0\,1}v_3 = 32473.099976$$

The like moments calculated from the curve $Y = a_0 + a_1X + a_2X^2$ are, of course, those given in Equations (2.35), and therefore the equations to determine a_0, a_1 and a_2 are

$$\begin{aligned} 108a_0 &+ 614a_1 + 3915a_2 = 1882.000116 \\ 614a_0 &+ 3915a_1 + 26622.3a_2 = 13367.13347 \\ 1305a_0 &+ 8874.1a_1 + 62858.25a_2 = 32473.09998 \end{aligned}$$

Solving these, we find

$$a_0 = +\ 3.916677$$

$$a_1 = -\ 4.000008$$

$$a_2 = +\ 1.000001$$

and the function f(X) sought is

$$Y = 3.916677 - 4.000008\ X + 1.000001\ X^2\ .$$

We recall now that the specific requirement placed on the fitted function f(X) is that

$$\int_{.5}^{X} f(X)dX = \text{observed frequency within the corresponding range.}$$

To see how well the fitted function serves this purpose, we find

$$\int_{.5}^{X} f(X)dX = \int_{.5}^{X} (3.916677 - 4.000008\, X + 1.000001\, X^2)dX$$

$$= - 1.500005 + 3.916677\,X - 2.000004\,X^2 + .333334\,X^3.$$

If in this last expression we set successively X = 1.5, 2.5,...,8.5, we shall obtain by subtraction the calculated frequency within each unit interval as shown in Table 2.28.

Mid-Value of Interval	Upper Limit of Integral X	$\int_{.5}^{X} f(X)dX$	Computed Frequency Within Interval	Actual Frequency Within Interval
	.5	.000000		
1			1.000004	1
	1.5	1.000004		
2			0.000003	0
	2.5	1.000007		
3			1.000004	1
	3.5	2.000011		
4			4.000011	4
	4.5	6.000022		
5			9.000020	9
	5.5	15.000042		
6			16.000035	16
	6.5	31.000077		
7			25.000052	25
	7.5	56.000129		
8			36.000075	36
	8.5	92.000204		

TABLE 2.28

The process outlined above for problems of Types I and II is quite general in its application and can be used in the representation of many forms of relationships such as arise in the study of physical and engineering problems involved in the control of quality of manufactured product. Enough has been said to indicate how the observed relationship between two variables may be expressed in terms of simple functions of the data by means of the method of moments. The necessary mathematical details are in general somewhat laborious although they become much less so if the assumed function $f_1(X)$ has high contact with the X-axis and vanishes at the extremes of the range.

Thus it can be shown[1] for Type I(b) that, if the continuous function $f_1(X)$ representing a series of discrete ordinates is such that it has high contact with the X-axis at the extremes of the range, the area and moments of

[1] Jones, C.D. "A First Course in Statistics", p. 200.

$f_1(X)$ may be found directly from the observations without introducing any corrections. A case of this kind arose in Chapter IV where the Pearson type curve was fitted to a Point Binomial distribution. No correction was applied to the moments obtained from the discrete ordinates of the binomial before equating them to the corresponding moments found from the Pearson curve by integration.

For the problems of Type II, even though $f_1(X)$ has high contact with the ends of the range, some adjustment is still necessary to allow for the fact that in calculating the moments from the observations we have assumed the frequency within a given interval to be centered at the middle of the interval, although we know that this is actually not the case. These adjustments are known in the literature as Sheppard's corrections.

Particularly in problems where we are interested in presenting the information contained in a series of observations of some particular quality characteristic, the information contained in the series of observations is often presented in the form of moments to which Sheppard's corrections have been applied. This practice needs some further consideration because here again we find that a so-called corrected moment has in it something which the analyst has put there in addition to what the original data contain and therefore should be used with due caution.

<u>Use of Sheppard's Corrections in the Presentation of Data</u>

Let us go through some of the detailed mathematical steps in arriving at corrections to be applied for the problems of Type II when the above condition as to high contact is satisfied. We shall find that, based upon the assumptions that are made, the moments calculated from the observed data must be corrected before they can be used either in an equation of relationship or in a theoretical frequency curve.

Assuming as before that the function $f_1(X)$ represents the data, the moments calculated from the observations are actually defined by

$$N \; {}_1\nu_m = \sum_{i=-\infty}^{+\infty} X_i^m \int_{x=-\frac{h}{2}}^{x=+\frac{h}{2}} f_1(X_i + x)\,dx \; .$$

If $f_1(X)$ can be expanded in a Taylor's Series about $X = X_i$, then

$$f_1(X_i + x) = f_1(X_i) + x f_1^{(1)}(X_i) + \frac{x^2}{\underline{|2}} f_1^{(2)}(X_i) + \frac{x^3}{\underline{|3}} f_1^{(3)}(X_i) + \cdots$$

and

$$\int_{-\frac{h}{2}}^{+\frac{h}{2}} f_1(X_i + x)\,dx = h\,f_1(X_i) + \frac{h^3}{2^2\,\underline{|3}}\, f_1^{(2)}(X_i) + \frac{h^5}{2^4\,\underline{|5}}\, f_1^{(4)}(X_i) + \cdot\;\cdot\;\cdot$$

Hence

$$N\;{}_1v_m = h \sum_{i=-\infty}^{+\infty} X_i^m\, f_1(X_i) + \frac{h^3}{2^2\,\underline{|3}} \sum_{i=-\infty}^{+\infty} X_i^m\, f_1^{(2)}(X_i) + \cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot$$

Now if $X^m\ f_1^{(j)}(X)$ vanishes together with all its derivatives at $X = \pm\,\infty$, for all integral values of j, we have, on applying the Euler MacLaurin Sum Formula

$$N\;{}_1v_m = \int_{-\infty}^{+\infty} X^m f_1(X)\,dX + \frac{h^2}{2^2}\,\frac{1}{\underline{|3}} \int_{-\infty}^{+\infty} X^m f_1^{(2)}(X)\,dX + \frac{h^4}{2^4\,\underline{|5}} \int_{-\infty}^{+\infty} X^m f_1^{(4)}(X)\,dX + \cdots \qquad (2.36)$$

The next step is to reduce terms like

$$\int_{-\infty}^{+\infty} X^m f_1^{(2j)}(X)\,dX$$

to moments of $f_1(X)$, $(j = 0,\ 1,\ 2,\ ..)$.

Integrating by parts we have

$$\int_{-\infty}^{+\infty} X^m f_1^{(2j)}(X)\,dX = \int_{-\infty}^{+\infty} X^m d\,[f_1^{(2j-1)}(X)] = -\,m \int_{-\infty}^{+\infty} X^{m-1}\ f_1^{(2j-1)}(X)\,dX.$$

Continuing we find

$$\int_{-\infty}^{+\infty} X^m f_1^{(2j)}(X)\,dX = (m)\ (m-1) \int_{-\infty}^{+\infty} X^{m-2} f_1^{(2j-2)}(X)\,dX = m(m-1)(m-2)\ldots(m-2j+1) \int_{-\infty}^{+\infty} X^{(m-2j)}\, f_1(X)\,dX,$$

by repeated application of the integration by parts.

Or

$$\int_{-\infty}^{+\infty} X^m f_1^{(2j)}(X)\,dX = N\,m(m-1)(m-2)\ \ ..\ \ (m-2j+1)\ \ {}_1\mu_{(m-2j)},$$

where ${}_1\mu_{m-2j}$ is the $(m-2j)^{th}$ moment of $f_1(X)$ about an arbitrary origin.

Substituting the value of these integrals into (2.36), we find

$$\begin{aligned} N\,{}_1\nu_m = N\,{}_1\mu_m &+ \frac{h^2}{2^2}\frac{N}{\underline{|3}}\,m(m-1)\;{}_1\mu_{m-2} \\ &+ \frac{h^4}{2^4\,\underline{|5}}\,Nm(m-1)(m-2)(m-3)\;{}_1\mu_{m-4} \\ &+ \frac{h^6}{2^6\,\underline{|7}}\,Nm(m-1)(m-2)(m-3)(m-4)(m-5)\;{}_1\mu_{m-6} + \cdots, \end{aligned}$$

which is the fundamental formula from which moments of the curve representing the data may be found in terms of the rough moments found from the grouped observations.

If moments are taken about the mean as origin then ${}_1\nu_m$ and ${}_1\mu_m$ become moments about the mean and the fundamental formula becomes

$$\nu_m = \mu_m + \frac{h^2}{2^2\,\underline{|3}}\,m(m-1)\;\mu_{m-2} + \frac{h^4}{2^4\,\underline{|5}}\,m(m-1)(m-2)(m-3)\;\mu_{m-4} + \cdots,$$

from which we obtain

$$\begin{aligned} &\mu_1 = \nu_1 = 0, \\ &\mu_2 = \nu_2 - \frac{h^2}{12}, \\ &\mu_3 = \nu_3, \\ &\mu_4 = \nu_4 - \frac{h^2}{2}\,\nu_2 + \frac{7}{240}\,h^4, \\ &\text{etc.} \ldots \end{aligned}$$

These equations express the corrected moments (μ's) in terms of those (ν's) for the original data. If the hypothetical curve $f_1(X)$ actually does satisfy the conditions in respect to contact required as a basis for Sheppard's corrections, the above relationshipsbetween the μ's and the ν's hold good. If, however, the assumptions involved in these corrections are not justified, the corrected moments may distort the information given by the original series of observations. The importance of this fact will become apparent in Part III.

It is felt that, in many problems arising in the control of quality, it is wise to make use of the original moments in presenting the information contained in the series of observations in such a way that they can be used in an almost universal manner as previously indicated, whereas the corrected moments naturally do not enjoy this generality. Of course, the observed moments may be easily obtained from the corrected ones by means of the above equations provided it is known that the moments have been corrected.

In closing this discussion on the method of moments, we shall indicate briefly how it leads to the use of the previously proposed simple moment functions of a distribution of values to express the information contained therein.

Application of Method of Moments to the Expression of the Information Contained in a Frequency Distribution

In the development of a Pearson Curve to represent the Point Binomial distribution, it was found desirable to first integrate the differential equation to obtain the expression of the frequency function and then to determine the parameters in this function by equating its moments to like moments of the Point Binomial.

However, for the purpose of showing how the various Pearson Curves arise as special cases of the same differential equation, we proceed in a somewhat different way. Starting with the general equation

$$\frac{dY}{dX} = \frac{Y(X + d)}{b_0 + b_1X + b_2X^2}$$

given in Part IV, we may determine the parameters b_0, b_1, b_2 and d as follows:

Writing the differential equation in the form

$$(b_0 + b_1X + b_2X^2)\ \frac{dY}{dX} = Y(X + d)\ ,$$

and multiplying both sides of this equation by X^m, we have

$$\int_{L_1}^{L_2} X^m(b_0 + b_1X + b_2X^2)\ \frac{dY}{dX}\ dX = \int_{L_1}^{L_2} X^mY(X + d)\,dX\ ,$$

where L_1 and L_2 denote the limits of the frequency range.

Integrating by parts, we find

$$X^m(b_0 + b_1X + b_2X^2)Y\ \Big|_{L_1}^{L_2} - \int_{L_1}^{L_2} Y\,[m\,b_0X^{m-1} + (m+1)b_1X^m + (m+2)b_2X^{m+1}]dX$$

$$= \int_{L_1}^{L_2} X^{m+1}Y\ dX\ +\ d\int_{L_1}^{L_2} X^m\,Y\,dX\ .$$

If Y vanishes at the ends of the frequency range, then the preceding equation gives

$$-m\,b_0\;{}_1\mu_{m-1} - (m+1)b_1\;{}_1\mu_m - (m+2)b_2\;{}_1\mu_{m+1} = {}_1\mu_{m+1} + d\;{}_1\mu_m \qquad (2.37)$$

where as before ${}_1\mu_i$ denotes the ith moment of the frequency function Y about an arbitrary origin.

Putting m = 0, 1, 2 and 3 successively in (2.37) we obtain four equations expressing the parameters in terms of moments of the given frequency function Y. However, by taking the origin at the mean of the distribution, we get a new differential equationsin the variables

$$\begin{aligned} x &= X - \bar{X} \\ y &= Y, \end{aligned}$$

as indicated below, but since this new equation has the same form as the old one, we shall for the sake of simplicity think of b_0, b_1, b_2 and d as the constants in this new equation.

Under these conditions, the equation which determined the constants are

$$\begin{array}{ccccccccc} 0 & + & b_1 & + & 0 & + & d & = & 0, \\ b_0 & + & 0 & + & 3b_2\mu_2 & + & 0 & = & -\mu_2, \\ 0 & + & 3b_1\mu_2 & + & 4b_2\mu_3 & + & d\mu_2 & = & -\mu_3, \\ 3b_0\mu_2 & + & 4b_1\mu_3 & + & 5b_2\mu_4 & + & d\mu_3 & = & -\mu_4. \end{array}$$

Solving these equations we find

$$b_0 = \frac{\begin{vmatrix} 0 & 1 & 0 & 1 \\ -\mu_2 & 0 & 3\mu_2 & 0 \\ -\mu_3 & 3\mu_2 & 4\mu_3 & \mu_2 \\ -\mu_4 & 4\mu_3 & 5\mu_4 & \mu_3 \end{vmatrix}}{\Delta}$$

$$b_1 = \frac{\begin{vmatrix} 0 & 0 & 0 & 1 \\ 1 & -\mu_2 & 3\mu_2 & 0 \\ 0 & -\mu_3 & 4\mu_3 & \mu_2 \\ 3\mu_2 & -\mu_4 & 5\mu_4 & \mu_3 \end{vmatrix}}{\Delta}$$

$$b_2 = \frac{\begin{vmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & -\mu_2 & 0 \\ 0 & 3\mu_2 & -\mu_3 & \mu_2 \\ 3\mu_2 & 4\mu_3 & -\mu_4 & \mu_3 \end{vmatrix}}{\Delta}$$

$$d = \frac{\begin{vmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 3\mu_2 & -\mu_2 \\ 0 & 3\mu_2 & 4\mu_3 & -\mu_3 \\ 3\mu_2 & 4\mu_3 & 5\mu_4 & -\mu_4 \end{vmatrix}}{\Delta}$$

where

$$\Delta = \begin{vmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 3\mu_2 & 0 \\ 0 & 3\mu_2 & 4\mu_3 & \mu_2 \\ 3\mu_2 & 4\mu_3 & 5\mu_4 & \mu_3 \end{vmatrix} .$$

Expanding these determinants we have

$$b_0 = - \frac{\mu_2(4\mu_2\mu_4 - 3\mu_3^2)}{10\mu_2\mu_4 - 12\mu_3^2 - 18\mu_2^3} ,$$

$$b_1 = - \frac{\mu_3(3\mu_2^2 + \mu_4)}{10\mu_2\mu_4 - 12\mu_3^2 - 18\mu_2^3} ,$$

$$b_2 = - \frac{2\mu_2\mu_4 - 3\mu_3^2 - 6\mu_2^3}{10\mu_2\mu_4 - 12\mu_3^2 - 18\mu_2^3} ,$$

$$d = \frac{\mu_3(\mu_4 + 3\mu_2^2)}{10\mu_2\mu_4 - 12\mu_3^2 - 18\mu_2^3} .$$

Before proceeding further, we should note that these four parameters are given in terms of the moments of the frequency function Y. Now this function is unknown and in practice we substitute for the moments of Y the corrected moments obtained from the data which are to be fitted. If the latter is such that the Pearson Curve representing them has high contact at the ends of the range, we should apply Sheppard's Corrections to the moments ν_m calculated from the observations.

To continue then, we assume that the above moments represent the corrected moments of the data so that we may now write the differential equation in the form

$$-\frac{1}{y}\frac{dy}{dx} = \frac{x + \dfrac{\mu_3(\mu_4 + 3\mu_2^2)}{10\mu_2\mu_4 - 12\mu_3^2 - 18\mu_2^3}}{\dfrac{\mu_2(4\mu_2\mu_4 - 3\mu_3^2) + \mu_3(3\mu_2^2 + \mu_4)x + (2\mu_2\mu_4 - 3\mu_3^2 - 6\mu_2^3)x^2}{10\mu_2\mu_4 - 12\mu_3^2 - 18\mu_2^3}}$$

If in this last equation we substitute

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}, \quad \beta_2 = \frac{\mu_4}{\mu_2^2}, \quad \sigma^2 = \mu_2,$$

we have the differential equation in the form

$$-\frac{1}{y}\frac{dy}{dx} = \frac{x + \dfrac{\sigma\sqrt{\beta_1}\,(\beta_2 + 3)}{10\beta_2 - 12\beta_1 - 18}}{\dfrac{\sigma^2(4\beta_2 - 3\beta_1)}{10\beta_2 - 12\beta_1 - 18} + \dfrac{\sigma\sqrt{\beta_1}\,(\beta_2 + 3)}{10\beta_2 - 12\beta_1 - 18}x + \dfrac{2\beta_2 - 3\beta_1 - 6}{10\beta_2 - 12\beta_1 - 18}x^2} \qquad (2.38)$$

The mode of this curve is the value of x for which $\frac{dy}{dx} = 0$ other than $y = 0$, that is,

$$x = -\frac{\sigma\sqrt{\beta_1}(\beta_2 + 3)}{10\beta_2 - 12\beta_1 - 18},$$

which gives the distance of the mode of the frequency function from the mean.

The various forms of frequency curves that may be obtained as an integral of (2.38) depend upon the nature of the roots of the equation

$$b_0 + b_1x + b_2x^2 = 0,$$

that is, they depend upon whether the "discriminant"

$$b_1^2 - 4b_0\, b_2 \gtreqless 0,$$

or upon whether

$$\frac{b_1^2}{4b_0\, b_2} \gtreqless 1.$$

This condition in terms of β_1 and β_2 becomes

$$\frac{\beta_1(\beta_2 + 3)^2}{4(4\beta_2 - 3\beta_1)(2\beta_2 - 3\beta_1 - 6)} \gtreqless 1.$$

Hence, given the relationship between β_1 and β_2 of a Pearson Curve, the nature of the curve is thus determined and this inequality is therefore called the "criterion"[1].

6. Stochastic Versus Functional Relationship

For several paragraphs, we have been considering the problem of presenting the information contained in an observed series of n pairs of two quality characteristics X and Y and indicating the relationship between them. For the most part however, we have talked about functional relationships.

Returning now to the 1370 pairs of observations of depth of penetration Y and depth of sapwood X of a certain group of 1370 telephone poles discussed in the previous chapter, we see at once that the customary concept of functional relationship does not appear to express our conception of relationship in this case. True it is that, in general, the deeper the sapwood, the deeper the penetration, but we are not able to say definitely for a given depth of sapwood what should be the depth of penetration.

To make this point clear, suppose we calculate the average depth $\bar{Y}$ of penetration of the poles in each of the columns of the scatter diagram of Fig. 2.6 above. Doing this, we get the following results:

Depth of Sapwood in inches X	Average Depth of Penetration in inches $\bar{Y}$	Number of Poles n
1.0	.85	2
1.3	.83	29
1.6	1.03	62
1.9	1.16	106
2.2	1.30	153
2.5	1.33	186
2.8	1.54	193
3.1	1.63	188
3.4	1.82	151
3.7	1.96	123
4.0	2.18	82
4.3	2.19	48
4.6	2.36	27
4.9	2.52	14
5.2	3.16	5
5.5	2.50	1

TABLE 2.29

[1] For a complete discussion of the various forms of Pearson Curves arising from different values of the criterion see W. Palin Elderton, "Frequency Curves and Correlation".

This table bears out the previous statement that the average depth of penetration increases with the depth of sapwood. Plotting the data of Table 2.29 as in Fig. 2.18 we see that the averages $\bar{Y}$ appear to lie approximately along a straight line or in other words, that

$$\bar{Y} = a_0 + a_1 X. \qquad (2.39)$$

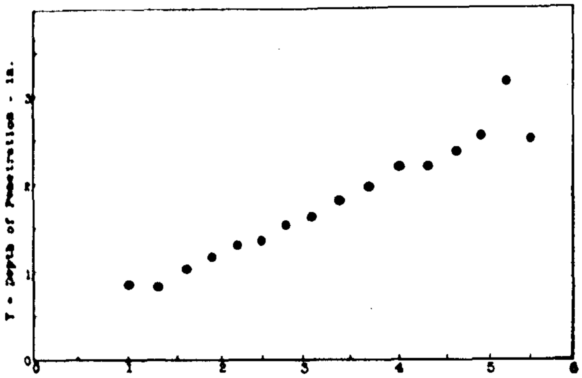


FIG. 2.18

Now we might proceed as before to find estimates of the parameters a_0 and a_1 by one of the four methods previously outlined, but we shall do it only for the Method of Least Squares, weighting each squared difference

$$v_i^2 = [\bar{Y}_i - (a_0 + a_1 X_i)]^2$$

by the number n_{Xi} of values X in the column designated by X_i.

In this way we get

$$v^2 = \Sigma\, n_{Xi}\, (\bar{Y}_i - a_0 - a_1 X_i)^2,$$

where the Σ is extended to the number of arrays of Y's.

The conditions for a minimum of v^2 are

$$\frac{\partial(v^2)}{\partial a_0} = 2\Sigma\, n_{Xi}\, (\bar{Y}_i - a_0 - a_1 X_i) = 0,$$

and

$$\frac{\partial(v^2)}{\partial a_1} = 2\Sigma n_{Xi} X_i (\bar{Y}_i - a_0 - a_1 X_i) = 0.$$

On examining these equations remembering that $n_{Xi}\ \bar{Y}_i$ is equal to the sum of all Y's in an X array of Y's, we see that, except for errors of grouping which vanish as dX—> 0, they are equivalent to those obtained by fitting the best straight line by the same method of least squares to the whole set of 1370 individual points. Hence, the slope of the best straight line is to a high degree of approximation

$$r \frac{\sigma_Y}{\sigma_X}$$

and the Equation (2.39) becomes when $(\bar{X}, \bar{Y})$ is chosen as the origin of coordinates

$$\bar{y}_x = r \frac{\sigma_y}{\sigma_x} x \qquad (2.40)$$

or putting in the values of r, σ_y and σ_x obtained from the data, we get $\bar{y}_x = .472209x$. The line (2.40) is shown in Fig. 2.19 and is technically termed the line of regression of y on x. It gives approximately the mean value of y associated with a given value of x. Furthermore, we have just seen that (2.40) may also be looked at as giving to a high order of approximation the best line through the individual points determined by the same method.

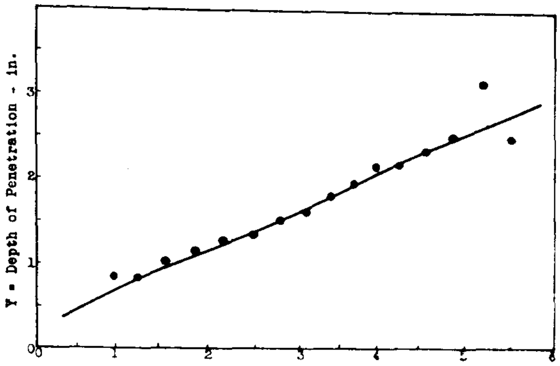


FIG. 2.19

Now we are ready to consider further the significance of the straight line obtained in either one of these two ways. On the one hand, had we proceeded according to the Method of Least Squares using the individual observations to find the relationship between the length L of a bar at some temperature t starting from some initial temperature t_0, we would have pronounced our findings as the empirical functional relationship between the two magnitudes. On the other hand, proceeding as we did in the present case and following customary practice, we would most likely have pronounced our findings as the empirical line of regression. Just how should we proceed in any physical case or what is the line of differentiation between the two methods of procedure? Asked such a question, an engineer might answer somewhat as follows:

As for the expansion of the rod, the length will vary in some definite functional way with temperature, that is, the length at a given temperature will always be the same except for small errors of measurement. In the case of the variation of depth of penetration with depth of sapwood, however, there are several factors supposed to influence the depth of penetration other than that of depth of sapwood so that even though the depth of sapwood remains constant, the variation in the other factors influencing penetration will in general, give rise to a whole series of values of the latter variable. In other words, it might be argued that there exists merely an apparent or stochastic relationship between depth of penetration and depth of sapwood.

A little closer consideration reveals however, that the so-called empirical functional relationship in the one case is the same fundamentally as the stochastic relationship in the other.

7. Use of Stochastic and Functional Relations

Of course we may use such relationships for estimating one variable in terms of another. Thus, from the expression

$$L = L_0 (1 + \alpha t)$$

introduced above, where $t = t_1 - t_0$ we may calculate L for any temperature difference t when α is known. Perhaps the engineer, in general, thinks of L as being the best estimate of the true length at temperature t. So far as the method of attaining it is concerned, we see that L is merely a point on some line of best fit although the observed values of length at the different temperatures used in obtaining this relationship did not, in general, fall on this line. In the same way from (2.40) we can estimate the mean depth of penetration corresponding to a given depth of sapwood.

In either case, however, we may be interested in knowing how the observed points used in determining an empirical relationship were distributed about the line of best fit. For example, we might wish to know how the observed depthsof penetration of telephone poles having a given sapwood thickness, say 4.3", were actually distributed about the average value given by the line of regression. Such information could be obtained from the scatter diagram. In general, however, we cannot afford to publish data in the form of scatter diagrams and hence, we need some method of presenting as much as possible of such information by means of simple functions or statistics.

Now in Chapter 3 of this part, we suggest the tabulation of the five statistics $\bar{X}$, $\bar{Y}$, σ_x, σ_y, and r. Let us next see if these statistics give us any information about the way the observed points are distributed about the linear relationship given in (2.40).

Suppose we calculate the standard deviation s_y of the n points in a scatter diagram about the line $y = r \frac{\sigma_y}{\sigma_x} x$ where y and x are measured from their respective mean values. This can easily be done and we have in fact

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - r\frac{\sigma_y}{\sigma_x}x_i)^2}{n},$$

$$= \frac{n\,\sigma_y^2 - 2nr^2\,\sigma_y^2 + nr^2\,\sigma_y^2}{n},$$

$$= \sigma_y^2\,(1 - r^2),$$

or

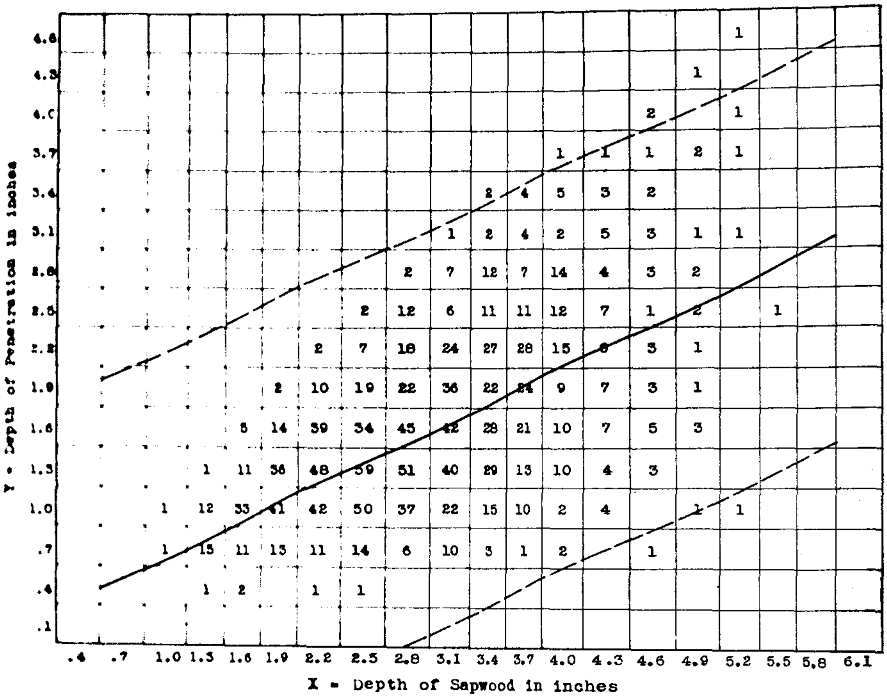
$$s_y = \sigma_y\sqrt{1 - r^2}\,. \qquad (2.41)$$

We see that nothing more is required in the way of statistics calculated from the raw data than those already mentioned in Chapter 3, Part II. Now let us see how we can make use of (2.41).

In what follows we shall limit our consideration to the case of linear regression or in other words, to the case where the straight line of best fit to the entire set of points is also to a high degree of approximation the line of best fit to the means of the columns.

Now, if the standard deviation of the y's in one column measured about the line of regression is practically the same as the standard deviation of the y's in any other column, then obviously this value of standard deviation is given by s_y and the distribution is said to be homo-schedastic. Under these conditions, if we draw two limit lines, one on each side of and parallel to the line of regression such that each line thus drawn is at a vertical distance $t\,s_y$ from the line of regression, we can make use of Tchebycheff's Theorem and state that not less than $1 - \frac{1}{t^2}$ of the points in the scatter diagram will fall inside the limit lines. Of course, if the points in the columns are normally distributed about the line of regression, then the total number of points in the scatter diagram falling within the limits thus set is given by the normal law integral, as would also be true if the points were distributed according to the second approximation discussed in the preceding chapter.

As an illustration, let us again consider the 1370 pairs of observed values of depth of penetration versus depth of sapwood. We have seen that the line of best fit to all the points is also approximately the line of regression which best fits the means of arrays of y's. Limits corresponding to $t\,s_y = 3s_y$ are given in Fig. 2.20. If all the conditions mentioned above had been met, then not less than $(1 - \frac{1}{9})$ 1370 or 1218 of the points should fall within the limits. By actual count, 1358 points fall within these limits. If we look further, however, we find that the condition of uniform variation about the line of regression (the condition of homoschedasticity) is not met. This is indicated by the results given in Table 2.30 below.



Y \ X	.4	.7	1.0	1.3	1.6	1.9	2.2	2.5	2.8	3.1	3.4	3.7	4.0	4.3	4.6	4.9	5.2	5.5	5.8	6.1
4.6																	1			
4.3																1				
4.0															2		1			
3.7													1	1	1	2	1			
3.4											2	4	5	3	2					
3.1										1	2	4	2	5	3	1	1			
2.8									2	7	12	7	14	4	3	2				
2.5								2	12	6	11	11	12	7	1	2		1		
2.2							2	7	18	24	27	28	15	6	3	1				
1.9						2	10	19	22	36	22	24	9	7	3	1				
1.6					5	14	39	34	45	42	28	21	10	7	5	3				
1.3				1	11	36	48	59	51	40	29	13	10	4	3					
1.0			1	12	33	41	42	50	37	22	15	10	2	4		1	1			
.7			1	15	11	13	11	14	6	10	3	1	2		1					
.4				1	2		1	1												
.1																				

FIG. 2.20

Depth of Sapwood in inches X	Standard Deviation of Penetration in inches σ	Number of Poles n
1.0	.15	2
1.3	.19	29
1.6	.27	62
1.9	.28	106
2.2	.34	153
2.5	.39	186
2.8	.47	193
3.1	.51	188
3.4	.59	151
3.7	.59	123
4.0	.67	82
4.3	.72	48
4.6	.89	27
4.9	.92	14
5.2	1.52	5
5.5	0	1

TABLE 2.30

This example illustrates a condition often found in practice, namely, that if the regression is practically linear, limits established as indicated above can be interpreted by means of Tchebycheff's Theorem as a rough approximation in the absence of further information although rigorously this Theorem does not apply except under the conditions stated.

Let us now review these facts from the standpoint of presentation of data. We see that the five simple statistics presented in Chapter 3, Part II serve to define an interpretable linear relationship provided certain conditions are satisfied by the data. Of course, they always serve to determine the line of best fit to the observed data where best fit means that the sum of the squares of the vertical deviations from this line is less than that from any other line. However, as we have seen, this relationship really becomes of use only when the line of best fit is also a good fit to the means of the column arrays and this is particularly true when the regression is linear.

We also have seen that these statistics serve to show how the points are distributed about the line of regression, provided the distribution is homo-schedastic. It is very important to note, however, that these statistics do not in themselves indicate whether or not they may be interpreted as above to indicate relationship unless we have the additional information that the regression is linear and that the distribution is homo-schedastic about the line of regression. We shall consider later another statistic to indicate whether or not the regression is linear.

8. Further Consideration of Presentation of Data Through Parameters in a Functional Relationship

In getting the line of best fit to the points in the scatter diagram of the 1370 pairs of values, we took that line such that the sum

$$\sum_{i=1}^{1370} (Y_i - a_0 - a_1X_i)^2$$

was a minimum.

Of course, we could have taken that line which would make

$$\sum_{i=1}^{1370} (X_i - b_0 - b_1Y_i)^2$$

a minimum.

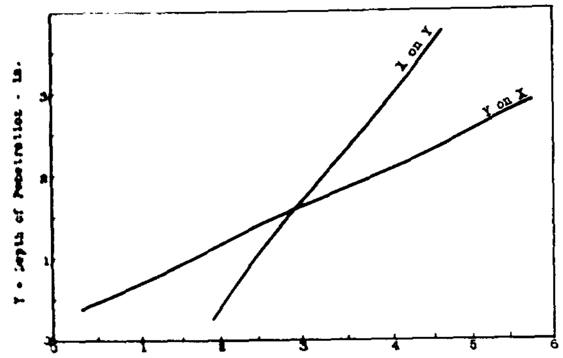
Following the same procedure as before, we would find that

$$x = r \frac{\sigma_x}{\sigma_y} y \tag{2.42}$$

where as before x and y are measured from the means $\bar{X}$ and $\bar{Y}$. Similarly, we would find the best line fitting the means of the arrays of X's in the scatter diagram to be

$$\bar{x}_y = r \frac{\sigma_x}{\sigma_y} y \tag{2.43}$$

to a high degree of approximation. Equation (2.43) is called the line of regression of x on y in contra-distinction to the line of regression of y on x in Equation (2.40) previously considered. The two lines of regression, y on x and x on y, for the 1370 pairs of numbers are shown in Fig. 2.21. They may be interpreted as giving approximately the mean value of y associated with a given value of x and the mean value of x associated with a given value of y respectively.



——Lines of Regression

FIG. 2.21

Quite naturally an engineer follows this simple distinction quite easily, but if he has not done so previously, he will most likely ask "What happened to our idea of functional relationship between y and x?" In the process of manipulation we really come out with two functional relations, both linear but nevertheless differing in their parameters, except in the special case where r = 1 in which the two become identical.

The significance of the above question becomes even more marked when we consider a set of data representing the relationship between two physical quantities such as the current Y, voltage X, relation for a carbon contact as previously considered. It will be recalled that by minimizing the sum

$$\sum_{i=1}^{n}(Y_i - a_0 - a_1X_i - a_2X_i^2)^2$$

we obtain the equation

$$Y = .008091 + .009667\,X + .0001615\,X^2 ,$$

whereas it follows that had we fitted the curve

$$X = b_0 + b_1Y + b_2Y^2$$

by minimizing the sum

$$\sum_{i=1}^{n}(X_i - b_0 - b_1Y_i - b_2Y_i^2)^2$$

we would have gotten

$$X = .8576 + 77.02\ Y - 25.181\ Y^2$$

Fig. 2.22 shows the graphs of these two equations. It will be observed that both fit very well indeed over the range of the observed data. Now which equation should be used? That perhaps is a matter for engineering judgment. At present we are not so much interested in this question as we are in considering the problem of presenting data by statistics. Let us then look over what we have done, keeping in mind our object, the presentation of information.

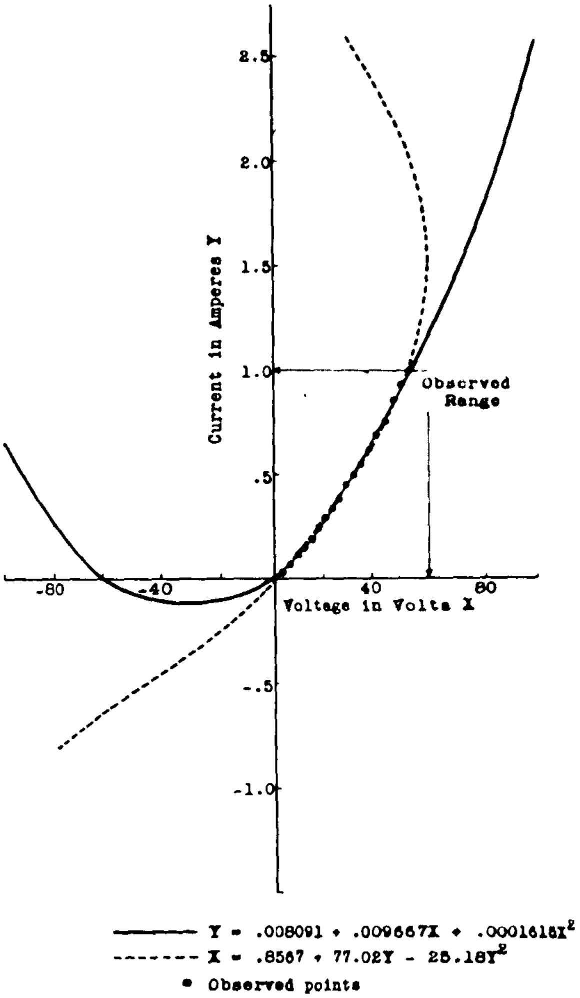


FIG. 2.28

Starting with the linear relationship, we have seen that in general, we get one line if we minimize the squares of the Y deviations of the points in a scatter diagram and get another line if we minimize the squares of the corresponding X deviations. In either case, we start with the assumption of a linear relationship between Y and X of the form

$$y = a_0 + a_1x \quad \text{or}$$

$$x = b_0 + b_1y$$

where x and y are measured from $\bar{X}$ and $\bar{Y}$ as before. It turns out that

$$a_0 = b_0 = 0$$

but in general,

$$a_1 \neq b_1.$$

Given a set of data then we could tabulate $\bar{X}$, $\bar{Y}$, a_1 and b_1, as indicating the relationship between Y and X. Since however,

$$a_1 = r\frac{\sigma_y}{\sigma_x} \quad \text{and}$$

$$b_1 = r\frac{\sigma_x}{\sigma_y},$$

we can get the same information from $\bar{X}$, $\bar{Y}$, σ_x, σ_y and r. Further, σ_x and σ_y give an indication of how X and Y are distributed. In other words, these five statistics give more than the information included in the two lines of best fit derived as above. We have also noted these five statistics can be interpreted to good advantage because they are involved in the estimates of certain parameters in an assumed <u>non-linear</u> relationship. In other words, we see that we may make use of the information thus expressed in a number of different ways, and now we shall go much further in the way of showing how very general the use of these five statistics really is.

9. Further Consideration of the Usefulness of the Information Given by $\overline{X}$, $\overline{Y}$, σ_x, σ_y and r

Fig. 2.23 shows the two lines of best fit already considered. Line A is obtained by minimizing $\sum_{i=1}^{n} v_{iy}^2$ and line B by minimizing $\sum_{i=1}^{n} v_{ix}^2$.

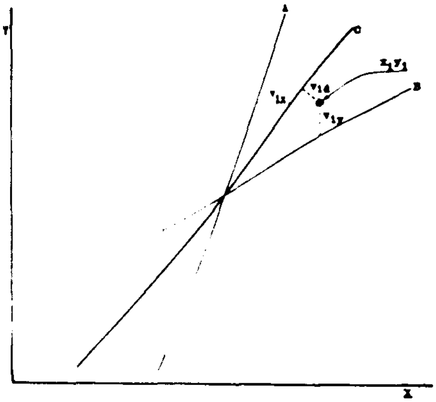


FIG. 2.23

Why should we take either of these lines, however, in preference to the one C so chosen that the sum of the squares $\sum_{i=1}^{n} v_{id}^2$ of the perpendicular distances of the points to this line is a minimum?

In fact in Part III we shall see that there may be a real object in choosing such a line as C. Therefore let us see how these same five statistics serve to determine the line C for a given set of data.

The equation to the line C may be written in the form

$$ax + by + c = 0$$

Where x and y are measured from the mean values $\overline{X}$ and $\overline{Y}$. Now of course, we are to find a, b and c such that $\sum_{i=1}^{n} v_{id}^2$ is a minimum. We shall now find a, b and c in terms of the five statistics already introduced.

We have

$$\sum_{i=1}^{n} v_{id}^2 = \left[\frac{ax_i + by_i + c}{\sqrt{a^2 + b^2}}\right]^2$$

$$= \frac{1}{a^2 + b^2} \sum_{i=1}^{n} \left[a^2 x_i^2 + 2ab\, x_i y_i + b^2 y_i^2 + 2a\, cx_i + 2b\, cy_i + c^2\right].$$

Since $\Sigma x_i = \Sigma y_i = 0$, the above expression reduces to

$$\sum_{i=1}^{n} v_{id}^2 = \frac{1}{a^2 + b^2} \left[a^2\, \Sigma x_i^2 + 2a\,b\, \Sigma x_i y_i + b^2\, \Sigma y_i^2 + n c^2\right].$$

Differentiating successively with respect to c, a and b and setting the resulting expressions equal to zero gives us the three normal equations from which we may determine a, b and c. We get:

$$\frac{2nc}{a^2 + b^2} = 0,$$

from which we see that c = 0.

$$\frac{1}{a^2 + b^2}\left[a\Sigma x_i^2 + b\Sigma x_i y_i\right] + \left[a^2\Sigma x_i^2 + 2ab\Sigma x_i y_i + b^2\Sigma y_i^2\right]\left[\frac{-a}{(a^2 + b^2)^2}\right] = 0$$

and

$$\frac{1}{a^2 + b^2}\left[a\Sigma x_i y_i + b\Sigma y_i^2\right] + \left[a^2\Sigma x_i^2 + 2ab\Sigma x_i y_i + b^2\Sigma y_i^2\right]\left[\frac{-b}{(a^2 + b^2)^2}\right] = 0$$

making use of the fact that c = 0.

The last two equations become, after simplifying:

$$(b^3 - a^2 b)\Sigma x_i y_i + ab^2\,(\Sigma x_i^2 - \Sigma y_i^2) = 0$$

and

$$(a^3 - ab^2)\Sigma x_i y_i + a^2 b\,(\Sigma y_i^2 - \Sigma x_i^2) = 0.$$

Dividing the first of these by b and the other by a, assuming that neither a nor b is zero, we obtain the single equation:

$$a^2\,(\Sigma x_i y_i) + ab\,(\Sigma y_i^2 - \Sigma x_i^2) - b^2\Sigma x_i y_i = 0,$$

from which we can determine the slope of the "least square" line. We may write the last equation as follows:

$$\left(\frac{a^2}{b^2}\right) + \frac{a}{b}\left(\frac{\sigma_y^2 - \sigma_x^2}{r\sigma_x\sigma_y}\right) - 1 = 0,$$

from which

$$\frac{a}{b} = \frac{1}{2r\sigma_x\sigma_y}\left[(\sigma_x^2 - \sigma_y^2) \pm \sqrt{(\sigma_x^2 - \sigma_y^2)^2 + 4r^2\sigma_x^2\sigma_y^2}\right]$$

One value of $\frac{a}{b}$ corresponds to that line for which Σv_{id}^2 is a maximum and the other value to that line for which Σv_{id}^2 is a minimum, since both values are obtained as a solution of the three normal equations.

The slopes of these two lines are

$$-\left(\frac{a}{b}\right)_1 = -\frac{1}{2r\sigma_x\sigma_y}\left[(\sigma_x^2 - \sigma_y^2) + \sqrt{(\sigma_x^2 - \sigma_y^2)^2 + 4r^2\sigma_x^2\sigma_y^2}\right]$$

and

$$-\left(\frac{a}{b}\right)_2 = -\frac{1}{2r\sigma_x\sigma_y}\left[(\sigma_x^2-\sigma_y^2)-\sqrt{(\sigma_x^2-\sigma_y^2)^2+4r^2\sigma_x^2\sigma_y^2}\right].$$

It follows, therefore, that the product of these slopes is -1, showing that the two lines are perpendicular.

Since c = 0, the line of best fit C, determined in this way, must go through the point $(\bar{X}, \bar{Y})$. We also see that the slope of this line is given in terms of σ_x, σ_y and r. Hence a knowledge of the five statistics $\bar{X}$, $\bar{Y}$, σ_x, σ_y and r give us all three lines A, B and C.

Suppose now that we consider the possibility of finding some function $f(X, Y, \lambda_1, \lambda_2, \ldots, \lambda_m)$ such that

$$\int_{Y_1}^{Y_2}\int_{X_1}^{X_2} f(X, Y, \lambda_1, \lambda_2, \ldots, \lambda_m)\,dX\,dY \qquad (2.44)$$

gives us approximately the number of observed pairs of values of X and Y within the corresponding limits. The function f in this case is the frequency function for two variables analogous to the corresponding frequency function for one variable treated in the previous chapter. Of course $Z = f(X,Y)$ is the equation of a surface and the integral (2.44) represents a certain section of the volume under this surface. What we shall do is to make use of a simple form of function for the surface as we did before for the frequency curve.

The equation of the surface which we shall use is

$$z = \frac{1}{2\pi\,\sigma_x\,\sigma_y\sqrt{1-r^2}}\,e^{-\frac{1}{2(1-r^2)}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}-2r\frac{xy}{\sigma_x\sigma_y}\right)} \qquad (2.45)$$

known as the normal law correlation surface. Here x and y are measured from their mean values $\bar{X}$ and $\bar{Y}$.

When the exponent of e in (2.45) is held constant, the ordinate z of the surface is constant. Thus, all the values of x and y that satisfy

$$\frac{1}{1-r^2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}-\frac{2rxy}{\sigma_x\sigma_y}\right) = \chi^2 \qquad (2.46)$$

will yield the value

$$z = \frac{1}{2\pi\,\sigma_x\sigma_y\sqrt{1-r^2}}\,e^{-\frac{1}{2}\chi^2}$$

for the ordinate. All such values of x and y lie on the ellipse given by (2.46) and hence this ellipse cuts the surface at a distance z above the xy-plane.

By changing to new rectangular coordinates x_1 and y_1, referred to axes which make a certain angle α with the old axes, we can simplify the equation of this ellipse by removing the xy term.

The relationship between the old and the new coordinates is given by

$$x = x_1 \cos \alpha - y_1 \sin \alpha,$$
$$y = x_1 \sin \alpha + y_1 \cos \alpha.$$

If we choose α so as to satisfy

$$\tan 2\alpha = \frac{2r\,\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2},$$

we shall obtain as the equation of the ellipse (2.46) in the new coordinates

$$a\,x_1^2 + b\,y_1^2 = \chi^2 \qquad (2.47)$$

where[1] a and b are defined by the equations

$$a + b = \frac{1}{(1-r^2)}\left(\frac{1}{\sigma_x^2} + \frac{1}{\sigma_y^2}\right)$$

$$(a - b) = -\frac{1}{(1-r^2)}\sqrt{\left(\frac{1}{\sigma_x^2} - \frac{1}{\sigma_y^2}\right)^2 + \frac{4r^2}{\sigma_x^2\,\sigma_y^2}}$$

so that

$$a = \frac{1}{2(1-r^2)}\left[\left(\frac{1}{\sigma_x^2} + \frac{1}{\sigma_y^2}\right) - \sqrt{\left(\frac{1}{\sigma_x^2} - \frac{1}{\sigma_y^2}\right)^2 + \frac{4r^2}{\sigma_x^2\,\sigma_y^2}}\right],$$

$$b = \frac{1}{2(1-r^2)}\left[\left(\frac{1}{\sigma_x^2} + \frac{1}{\sigma_y^2}\right) + \sqrt{\left(\frac{1}{\sigma_x^2} - \frac{1}{\sigma_y^2}\right)^2 + \frac{4r^2}{\sigma_x^2\,\sigma_y^2}}\right].$$

Also, from Equation (2.47) we see that the semi-axes of this ellipse are

$$\frac{\chi}{\sqrt{a}} \text{ and } \frac{\chi}{\sqrt{b}}.$$

We are now in a position to find the frequency or volume under the surface (2.45) bounded by the ellipse χ^2. The process of doing this is illustrated graphically in Fig. 2.24.

[1] See for example Osgood and Graustein "Analytic Geometry" pp. 242-243.

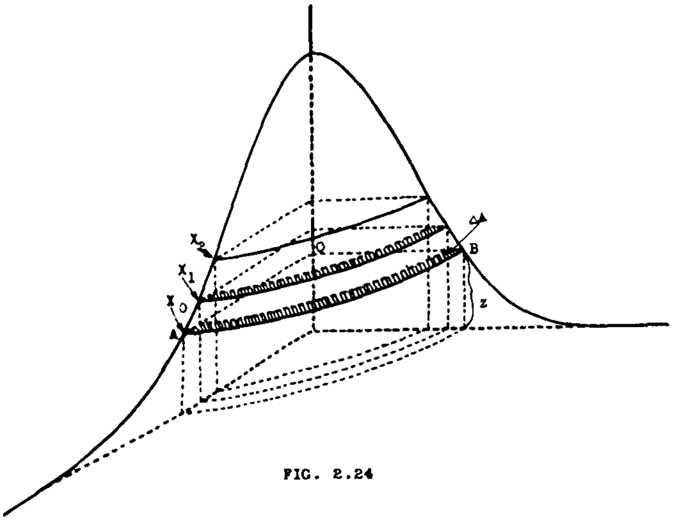


FIG. 2.24

AOB represents one quadrant of the ellipse χ_0, obtained by setting the right hand member of (2.46) equal to χ_0. Giving to χ a series of values starting from some fixed value χ_0 and progressing towards 0, the corresponding ellipses progress toward the apex of the surface, each ellipse being smaller than its predecessor which means, of course, that ΔA, the change in the area from one ellipse to the following one is negative.

If now we project the ellipse χ_1 upon the ellipse χ_0, we shall have a small ring of area $(\Delta A)_0$ lying between this projection and the ellipse χ_0. Furthermore, the volume of the surface lying under this ring of area is a thin disc of elliptical form whose volume is clearly $z_0(\Delta A)_0$.

If we were to carry out this process for succeeding ellipses $\chi_2, \chi_3, \ldots \chi_j, \ldots$ and form for each the corresponding volume $z_j\ (\Delta A)_j$, we would get thereby a series of elliptical discs each lying within its predecessor such that the sum of their volumes is approximately the volume of the surface lying within the ellipse χ_0.

Thus, approximately the volume within the ellipse χ_0 is

$$z_0\ (\Delta A)_0 + z_1\ (\Delta A)_1 + \ldots + z_j\ (\Delta A)_j + \ldots + z_n\ (\Delta A)_n$$

and the actual volume is the

$$\lim_{\substack{n \to \infty \\ \Delta A_j \to 0}} \sum_{j=0}^{n} z_j\ (\Delta A)_j = \int_{z_0}^{z} z\, dA = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} \int_{\chi_0}^{0} e^{-\frac{1}{2}\chi^2}\, dA.$$

The next step in the process is to substitute for dA its value in terms of χ. To do this, we recall that

$$A = \text{area of ellipse } \chi$$
$$= \frac{\pi\chi^2}{\sqrt{ab}}\,.$$

Now from the above values of a and b, we find

$$ab = \frac{1}{\sigma_x{}^2\ \sigma_y{}^2\,(1-r^2)}\ .$$

Hence

$$A = \chi^2 \pi\, \sigma_x \sigma_y\ \sqrt{1-r^2}$$

and the volume or relative frequency P of observations lying within the ellipse χ_0 is

$$P = -\int_{\chi_0}^{0} e^{-\frac{1}{2}\chi^2}\, \chi\, d\chi = 1 - e^{-\frac{1}{2}\chi_0^2}.$$

From the table of values of $e^{-\frac{1}{2}\chi^2}$ and $1 - e^{-\frac{1}{2}\chi^2}$ given below, we can read off the volume P under the frequency surface bounded by any ellipse χ and the xy-plane. This volume is our approximation to the number of points observed in the xy-plane within the same ellipse.

χ^2	Fraction outside $e^{-\frac{1}{2}\chi^2}$	Fraction inside $1-e^{-\frac{1}{2}\chi^2}$
.1	.951229	.048771
.2	.904837	.095163
.3	.860708	.139292
.4	.818731	.181269
.5	.778801	.221199
.6	.740818	.259182
.7	.704688	.295312
.8	.670320	.329680
.9	.637628	.362372
1.0	.606531	.393469
2.0	.367879	.632121
3.0	.223130	.776870
4.0	.135335	.864665
5.0	.082085	.917915
6.0	.049787	.950213
7.0	.030197	.969803
8.0	.018316	.981684
9.0	.011109	.988891
10.0	.006738	.993262
11.0	.004087	.995913
12.0	.002479	.997521
13.0	.001503	.998497
14.0	.000912	.999088
15.0	.000553	.999447
16.0	.000335	.999665
17.0	.000203	.999797
18.0	.000123	.999877
19.0	.000075	.999925
20.0	.000045	.999955

Fraction outside $e^{-\frac{1}{2}\chi^2}$	Fraction inside $1-e^{-\frac{1}{2}\chi^2}$	χ^2
.9000	.1000	.2107
.8000	.2000	.4463
.7500	.2500	.5754
.7000	.3000	.7134
.6000	.4000	1.0217
.5000	.5000	1.3863
.4000	.6000	1.9326
.3000	.7000	2.4080
.2500	.7500	2.7726
.2000	.8000	3.2198
.1000	.9000	4.6052
.0500	.9500	5.9915
.0100	.9900	9.2104
.0030	.9970	11.8194
.0027	.9973	11.8290

TABLE 2.31

For example, we see from this table that the χ^2 corresponding to a fraction .5000 inside the ellipse is 1.3863 and for a fraction .9973 is 11.8290. The data sheet in Fig. 2.25 shows the details of the computation necessary to obtain these two ellipses for the 1370 pairs of values of depth of sapwood and depth of penetration.

Actual observation shows that 683 and 1358 points fall within the theoretical ellipses. Note this close approximation obtained by means of a simple frequency function involving the five statistics $\overline{X}$, $\overline{Y}$, σ_x, σ_y and r, even though the regression is not strictly linear, and the data are far from homoscedastic.

Presentation of Relationship Between Several Variables in Terms of the Same Five Statistics

[illegible]	Tensile Strength in psi	Hardness in Rockwells "E"	Density in gms/cm^3
1	29314	53.0	2.666
2	34860	70.2	2.708
3	34018	64.3	2.665
4	30120	56.3	2.627
5	34020	78.5	2.581
6	30824	63.5	2.635
7	36386	71.4	2.671
8	31280	53.4	2.650
9	32184	82.5	2.717
10	33484	67.3	2.614
11	37694	69.5	2.584
12	34876	73.0	2.741
13	34640	56.7	2.619
14	34760	66.8	2.755
15	38080	96.4	2.846
16	28640	51.1	2.575
17	28810	74.4	2.561
18	28440	64.1	2.593
19	28070	77.8	2.639
20	24640	52.4	2.611
21	28770	69.1	2.696
22	28690	53.5	2.606
23	28850	64.3	2.616
24	32880	82.7	2.748
25	28810	56.7	2.518
26	34002	70.5	2.726
27	34470	87.5	2.675
28	29840	50.7	2.586
29	29710	72.5	2.547
30	29880	59.5	2.606
31	29880	71.3	2.648
32	27992	58.7	2.400
33	31862	76.5	2.692
34	27648	63.7	2.669
35	31640	69.2	2.628
36	30844	69.2	2.696
37	31888	61.4	2.648
38	36640	63.7	2.775
39	41878	94.7	2.874
40	30496	70.2	2.700
41	29848	60.4	2.583
42	32822	76.7	2.668
43	32822	62.9	2.679
44	30880	56.0	2.609
45	30880	63.2	2.721
46	28808	62.6	2.678
47	29190	78.0	2.610
48	30626	64.6	2.728
49	34332	64.0	2.709
50	34750	75.3	2.660
51	40678	84.8	2.949
52	28800	49.4	2.669
53	34646	74.2	2.624
54	31844	59.8	2.705
55	33802	75.2	2.736
56	34860	57.7	2.701
57	36690	79.3	2.776
58	38344	67.6	2.754
59	34440	77.0	2.660
60	34650	74.8	2.819

TABLE 2.32

PROPERTIES OF ALUMINUM DIE CASTINGS

Let us consider the data given in Table 2.32 These measurements were obtained by Committee XV-B2 of the American Society for Testing Materials in connection with a program calling for approximately 80,000 measurements to determine the physical properties of aluminum die castings.

Following the suggestions of Chapter 3, we would record the results in the form given below.

		T.S. X_1 in psi	Hardness X_2 in Rockwells	Density X_3 in gms./cc
Arithmetic Mean	$\overline{X}$	31869.4	69.825	2.6785
Std. Deviation	σ	3962.9	11.773	1.0986
No. of Measurements	n	60	60	60

Correlation coefficient r_{12} between T.S. and Hardness = .683

Correlation coefficient r_{13} between T.S. and Density = .657

Correlation coefficient r_{23} between Hardness and Density = .616

TABLE 2.33: Presentation of Information in Table 2.32 by Means of Simple Statistics

Naturally, we may use the information of Table 2.33 pertaining to any one of the pairs of quality characteristics in just the same manner as we have done in the previous

X = Depth of Sapwood in inches Y = Depth of Penetration in inches

$n = 1370$

$\bar{X} = 2.914088$ $\sigma_X = .798211$

$\bar{Y} = 1.591460$ $\sigma_Y = .624872$

$r = .603201$

$$\tan 2\alpha = \frac{2r\sigma_X\sigma_Y}{\sigma_X^2 \quad \sigma_Y^2} = 2.439350$$

$$2\alpha = 67° \; 42' \; 32"$$

$$\alpha = \underline{33° \; 51' \; 16"}$$

$$a = \frac{1}{2(1-r^2)}\left[\frac{1}{\sigma_X^2} + \frac{1}{\sigma_Y^2} - \sqrt{\left(\frac{1}{\sigma_X^2} - \frac{1}{\sigma_Y^2}\right)^2 + \frac{4r^2}{\sigma_X^2\sigma_Y^2}}\right] = \underline{1.191941}$$

$$b = \frac{1}{2(1-r^2)}\left[\frac{1}{\sigma_X^2} + \frac{1}{\sigma_Y^2} + \sqrt{\left(\frac{1}{\sigma_X^2} - \frac{1}{\sigma_Y^2}\right)^2 + \frac{4r^2}{\sigma_X^2\sigma_Y^2}}\right] = \underline{5.301130}$$

$$\underline{aX_1^2 + bY_1^2 = \chi^2}$$

$\underline{\text{Let } \chi^2 = 1.3863 \text{ or } 1 - e^{-\frac{1}{2}\chi^2} = .5000}$

$\chi = 1.1774$

$\frac{\chi}{\sqrt{a}} = 1.0784$ $\frac{\chi}{\sqrt{b}} = .5114$

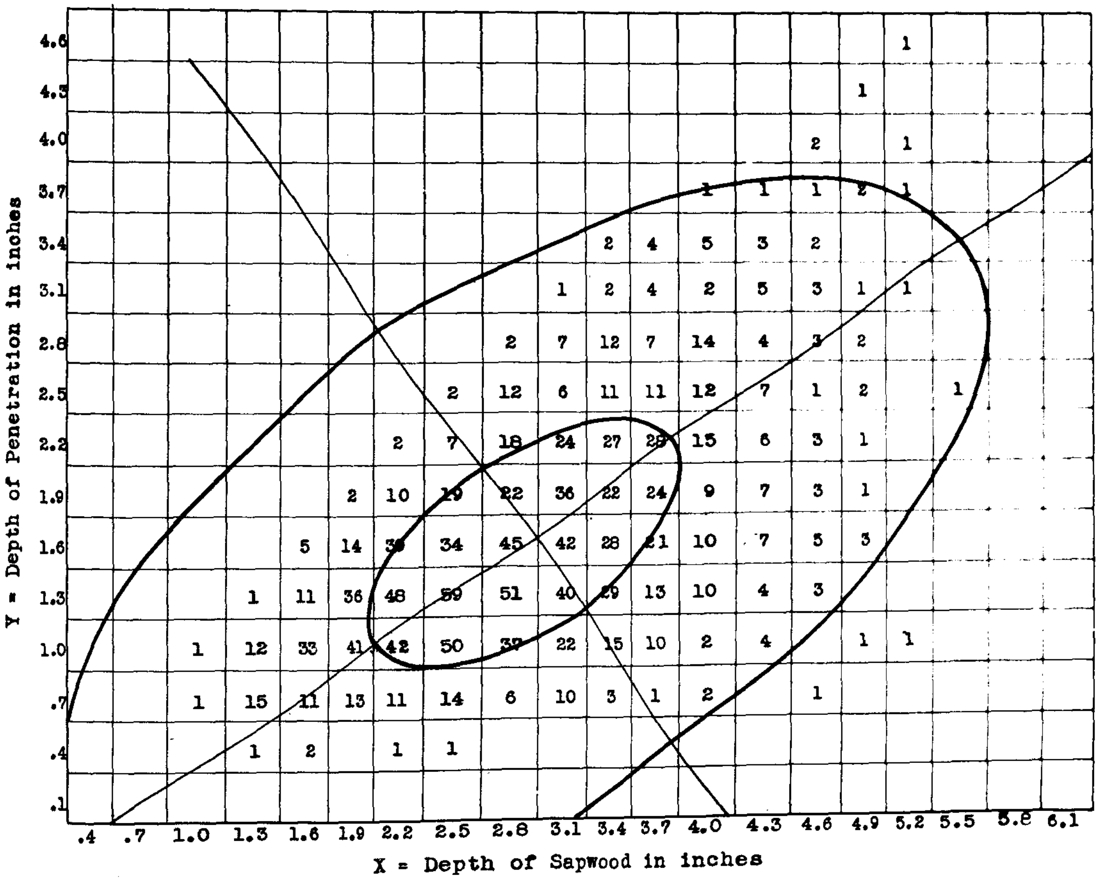
$1.191941X_1^2 + 5.301130Y_1^2 = 1.3863$

$\underline{\text{Let } \chi^2 = 11.8290 \text{ or } 1 - e^{-\frac{1}{2}\chi^2} = .9973}$

$\chi = 3.4393$

$\frac{\chi}{\sqrt{a}} = 3.1502$ $\frac{\chi}{\sqrt{b}} = 1.4938$

$1.191941X_1^2 + 5.301130Y_1^2 = 11.8290$



	.4	.7	1.0	1.3	1.6	1.9	2.2	2.5	2.8	3.1	3.4	3.7	4.0	4.3	4.6	4.9	5.2	5.5	5.8	6.1
4.6																	1			
4.3																1				
4.0															2		1			
3.7													1	1	1	2	1			
3.4											2	4	5	3	2					
3.1										1	2	4	2	5	3	1	1			
2.8									2	7	12	7	14	4	3	2				
2.5								2	12	6	11	11	12	7	1	2		1		
2.2							2	7	18	24	27	28	15	6	3	1				
1.9						2	10	19	22	36	22	24	9	7	3	1				
1.6					5	14	39	34	45	42	28	21	10	7	5	3				
1.3				1	11	36	48	59	51	40	29	13	10	4	3					
1.0			1	12	33	41	42	50	37	22	15	10	2	4		1	1			
.7			1	15	11	13	11	14	6	10	3	1	2		1					
.4				1	2		1	1												
.1																				

FIG. 2.25 - CALCULATION AND FIGURE FOR 50% AND

7 ELLIPSE [illegible] OF T[illegible]LE 2.1

paragraphs. Thus Fig. 2.26 shows such results for the pair of characteristics, tensile strength and hardness.

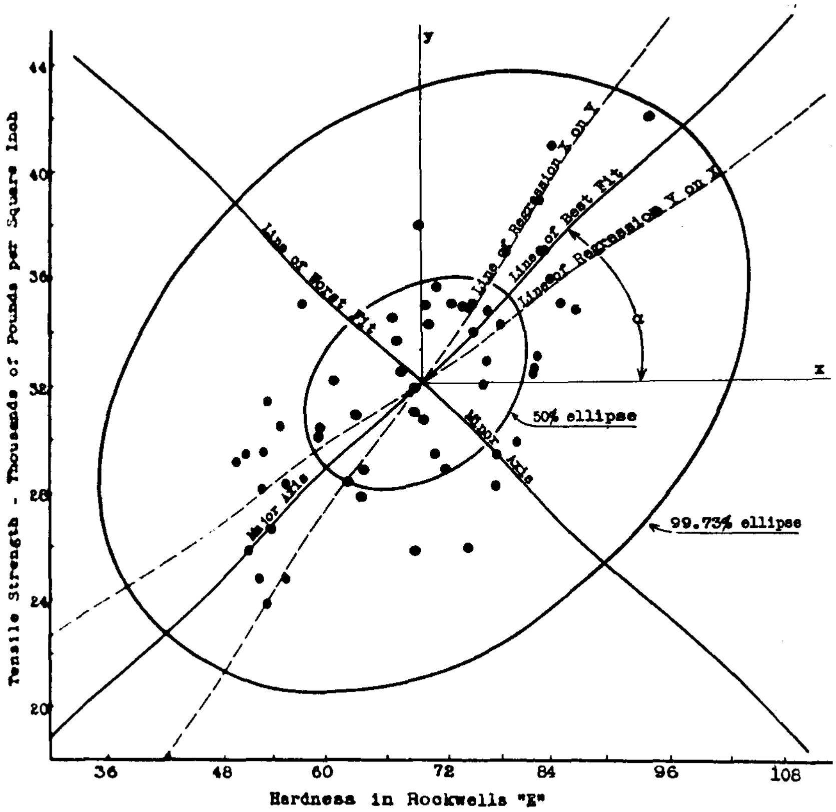


FIG. 2.26

It is natural, however, for us to wish to picture all of the results given in Table 2.33 in a way to indicate the nature of the relationship between the three quality characteristics. Suppose then that we find the plane of best fit whose equation is of the form

$$x_1 = a + bx_2 + cx_3$$

making use of the method of least squares where, of course, x_1, x_2, and x_3 are measured from their means $\bar{X}_1$, $\bar{X}_2$, and $\bar{X}_3$.

The quantity to be minimized is

$$v^2 = \sum_{i=1}^{n} \left[x_{1i} - (a + b\,x_{2i} + c x_{3i})\right]^2.$$

Hence the equations that determine a, b and c are:

$$2\sum_{i=1}^{n} \left[x_{1i} - (a + b\,x_{2i} + c\,x_{3i})\right] = 0,$$

$$2\sum_{i=1}^{n} \left[x_{1i} - (a + b\,x_{2i} + c\,x_{3i})\right] x_{2i} = 0,$$

$$2\sum_{i=1}^{n} \left[x_{1i} - (a + b\,x_{2i} + c\,x_{3i})\right] x_{3i} = 0.$$

From the first of these equations, we have, since $\Sigma x_{1i} = \Sigma x_{2i} = \Sigma x_{3i} = 0$,

$$a = 0.$$

The remaining two equations then give

$$(n\sigma_2^2)b + (n r_{23}\sigma_2\sigma_3)c = n r_{12}\sigma_1\sigma_2 ,$$

$$(n r_{23}\sigma_2\sigma_3)b + (n\sigma_3^2)c = n r_{13}\sigma_1\sigma_3 ,$$

where r_{ij} is the correlation coefficient between x_i and x_j and σ_j is the standard deviation of the j^{th} variable. In this way we get

$$a = 0$$

$$b = \frac{\sigma_1}{\sigma_2}\ \frac{(r_{12} - r_{13}\, r_{23})}{1 - r_{23}^2},$$

$$c = \frac{\sigma_1}{\sigma_3}\ \frac{(r_{13} - r_{12}\, r_{23})}{1 - r_{23}^2}.$$

as expressions of the parameters in this plane in terms of the simple statistics of Table 2.33. Substituting the numerical values given in Table 2.33 in the equation of the plane, we get

$$x_1 = 150.9888\ x_2 + 15310.348\ x_3$$

as the plane of regression of tensile strength on hardness and density.

Fig. 2.27 gives two pictures of this plane, indicating its relationship to the observed points. The standard deviation $\sigma_{1.23}$ of the points from this plane is given by

$$\sigma_{1.23} = \sigma_1 \frac{\begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{vmatrix}^{\frac{1}{2}}}{(1 - r_{23}^2)^{1/2}} = 1384 \text{ psi}$$

Furthermore, if the points are distributed homo-scedastically about this plane, then not less than $1 - \frac{1}{t^2}$ of the points will lie within the volume contained between the two planes, equally spaced on either side of and parallel to the plane of regression at a vertical distance of $t\,\sigma_{1.23}$ from this plane.

In a similar way, we could find planes of best fit by minimizing either the squares of deviations in x_2 or x_3. Also, there is one plane such that the

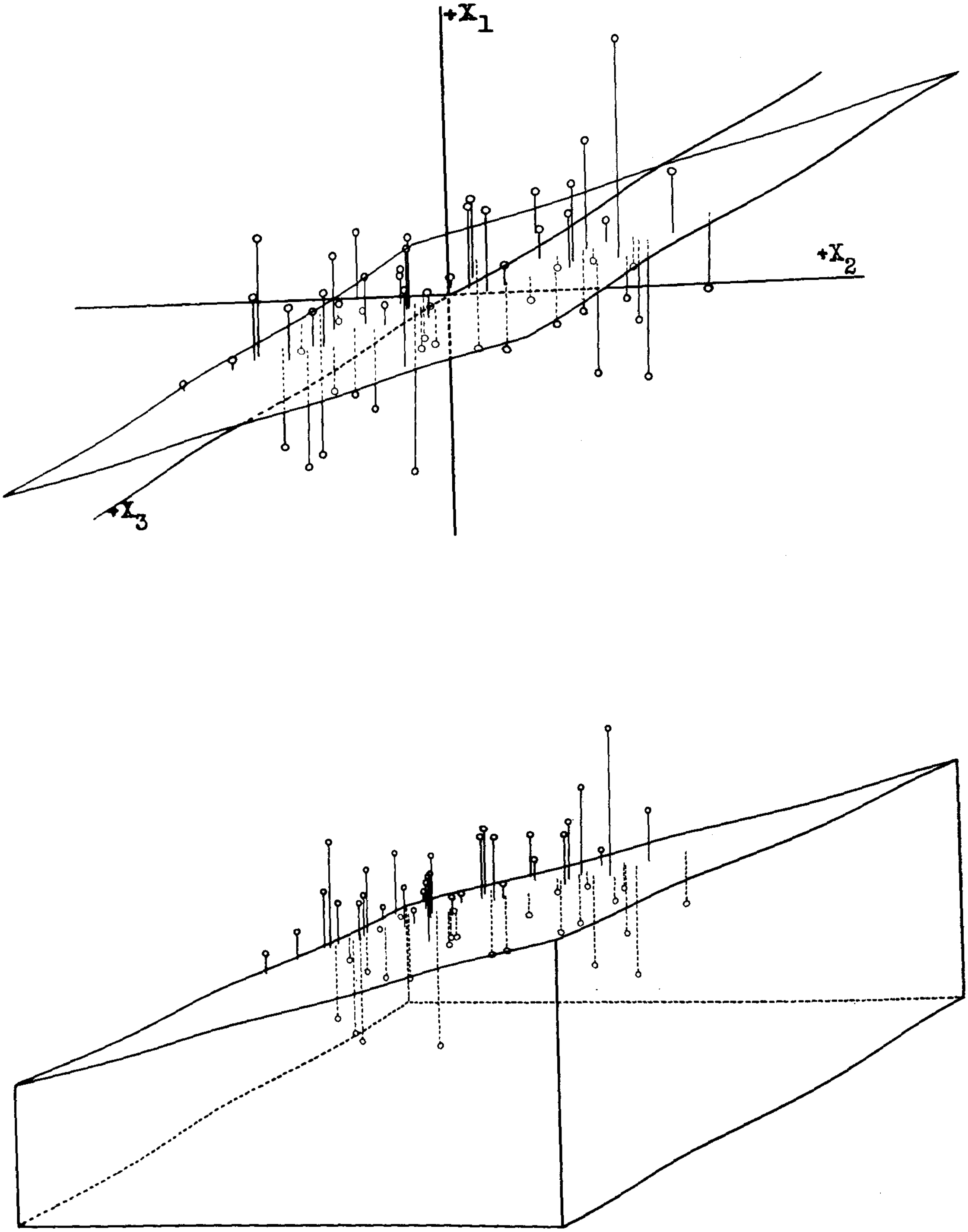


FIG. 2.27

sum of the squares of the perpendicular distances of the points to this plane is a minimum. In each case, we would find that the parameters of these planes may be expressed in terms of the statistics given in Table 2.33.

This same reasoning can be extended to the treatment of m different variables. Thus, if we are given n values of each of m variables we may, by extending the geometrical analogue of the case of three variables, represent them as n points in space of m dimensions, getting thereby a scatter of points in this hyperspace. Just as in the case of three variables, we may seek to determine a mathematical relationship between the variable x_1, say, and the remaining variables $x_2, x_3, \dots, x_m$. In the simplest case we may assume this relationship to be linear and proceed as before to find by the method of least squares estimates of the parameters in this assumed relationship in terms of simple functions of the data.

Hence, denoting this relationship by

$$x_1 = a_1 + a_2 x_2 + \dots + a_m x_m$$

where the x's are deviations from their respective mean values, the method involves making

$$v^2 = \sum_{i=1}^{n} \left[x_{1i} - (a_1 + a_2 x_{2i} + \dots + a_m x_{mi})\right]^2$$

a minimum.

Differentiating this expression successively with respect to $a_1, a_2, \dots, a_m$ and setting each derivative equal to zero, we may solve for the a's and substitute these in the equation of the plane. In this way we get

$$x_1 = -\sigma_1 \sum_{j=2}^{m} \frac{R_{1j}}{R_{11}} \frac{x_j}{\sigma_j},$$

where R_{ij} is the co-factor of the element standing in the i^{th} row and the j^{th} column of the determinant

$$R = \begin{vmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1m} \\ r_{21} & 1 & r_{23} & \cdots & r_{2m} \\ r_{31} & r_{32} & 1 & \cdots & r_{3m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{m1} & r_{m2} & r_{m3} & \cdots & 1 \end{vmatrix}$$

and σ_j is the standard deviation of the j^{th} variable. Thus we can see that, even in the general case of m variables where we assume a linear relationship, the simple functions, the average, standard deviation and correlation coefficient are sufficient to determine estimates of the parameters occurring in this relationship.

The standard deviation of the n points from this hyperplane of best fit can be shown to be

$$\sigma_{1.23\ldots m} = \sigma_1\sqrt{\frac{R}{R_{11}}},$$

where R and R_{11} are defined as above.

If the regression of x_1 on the remaining $(m-1)$ variables is linear it can also be shown that to a first degree of approximation the above plane of best fit also becomes the plane of regression. Hence provided the distribution about this plane is homo-scedastic, we may use Tchebycheff's Theorem to show that not less than $n(1-\frac{1}{t^2})$ of the n observed points in the m-hyperspace lie inside the planes

$$\bar{x}_1 \pm t\sigma_{1.23\ldots m}\,.$$

Thus far, of course, we have been considering the use of the statistics of Table 2.33 as estimates of parameters in assumed linear relationships between the variables considered. As in the case of two variables, we found that more than five simple statistics were required to estimate parameters in assumed parabolic relationships, so we would find in this case that more functions than given in Table 2.33 are required to estimate the parameters in an assumed non-planer relationship.

Suppose now that we consider the problem of finding a function f involving ℓ parameters

$$f(X_1, X_2, \ldots, X_m, \lambda_1, \lambda_2, \ldots, \lambda_\ell)$$

such that the

$$\int\int\int\cdots\int f(X_1, X_2, \ldots, X_m, \lambda_1, \lambda_2, \ldots, \lambda_\ell)\, dX_1 \ldots dX_m$$

taken over a given volume in m-space will give the proportion of the observed points lying within this space.

Let us take as a simple function in the m variables

$$z = z_0\, e^{-\frac{1}{2}\chi^2}$$

where

$$\chi^2 = \frac{1}{R}\left\{R_{11}\frac{x_1^{\,2}}{\sigma_1^{\,2}} + R_{22}\frac{x_2^{\,2}}{\sigma_2^{\,2}} + \ldots + 2R_{12}\frac{x_1 x_2}{\sigma_1\sigma_2} + \ldots\right\} \qquad (2.48)$$

where the R's are defined as above, and z_0 is a constant depending on the determinant R and the m standard deviations.

Analogous to the case of the normal correlation surface of two variables, we see that all values of the m variables which satisfy the Equation (2.48) when χ has the value χ_0 lie on the m-dimensional ellipse

$$\frac{1}{R}\left\{R_{11}\frac{x_1^2}{\sigma_1^2} + R_{22}\frac{x_2^2}{\sigma_2^2} + \ldots + 2R_{12}\frac{x_1x_2}{\sigma_1\sigma_2} + \ldots\right\} = \chi_0^2$$

and for all such values of the m variables

$$z = z_0\, e^{-\frac{1}{2}\chi_0^2} .$$

Hence, to find the proportion P of the n observed points lying outside any m-dimensional ellipse χ, we must calculate

$$P = \frac{\int_\chi^\infty z_0\, e^{-\frac{1}{2}\chi^2}\, dV}{\int_0^\infty z_0\, e^{-\frac{1}{2}\chi^2}\, dV} ,$$

where V is the volume of the m space ellipse χ.

By referring the ellipse to principal axes and then squeezing it into an m-dimensional sphere, it can be shown that the volume of this sphere is proportional to χ^m and therefore

$$dV = C\, m\, \chi^{m-1}\, d\chi$$

where C is a factor of proportionality. Hence the fraction or relative frequency of sets of observations lying outside the ellipse χ is

$$P = \frac{\int_\chi^\infty e^{-\frac{1}{2}\chi^2}\, \chi^{m-1}\, d\chi}{\int_0^\infty e^{-\frac{1}{2}\chi^2}\, \chi^{m-1}\, d\chi} .$$

A table of values of χ^2 for a large range of values of P and m is given by R. A. Fisher[1].

[1] Statistical Methods for Research Workers.

10. Concluding Remarks on the Presentation of Relationship

In this chapter we have barely touched upon the very important problem of presenting the information contained in sets of data so as to show relationship. It is believed, however, that we have covered in some detail those points which will be found most helpful in the study of control of quality of manufactured product. Starting with the consideration of the real meaning of functional relationship, we soon found that it is not feasible in general to represent observed relationships by mathematical functions. It would be perhaps more to the point to say that in the customary case one set of observed values cannot be expressed as a function of another in the mathematical sense. On the other hand, such functional relationships may serve to express approximately certain truths about the observed relationship between the sets of data themselves.

In general, the study of relationship between measurements of two or more quality characteristics involves some kind of an assumption which often can be expressed in mathematical form. The assumed relationship involves certain parameters which in turn must be estimated from the observed data. In this chapter we have considered four different methods of obtaining such estimates. We have considered in detail several ways of obtaining these estimates by the method of least squares. In general, however, it has been seen that the values of the parameters even in a given form of mathematical function are actually different depending upon the method used in estimating them. In the general case we have no definite criterion to guide us in the choice of method of estimating the parameters; hence two or more individuals, given a set of data, might reasonably choose to find different groups of parameters, which in turn most likely will differ one from another. Therefore, to be able to interpret the meaning of an estimate of a given parameter, we must have some knowledge of the method used in estimating it. This fact has been illustrated in considerable detail in respect to the very simple linear relationship where it is shown that by the method of least squares we may obtain several such relationships having different estimates of parameters even when calculated from the same set of data.

The very important thing in this connection is to have seen that all of the commonly derived estimates of parameters in a simple functional form can be expressed in terms of the five statistics for paired variables, namely, the averages and standard deviations together with the correlation coefficient between the variables. This is a very important fact because it shows how universal is the use of these five statistics. If a given set of data, consisting of n observations on two or more characteristics, is reduced to the simple statistics, namely, the averages, standard deviations and correlation coefficients, it has been seen that these may be used in numerous different ways in the interpretation of the data, particularly when relationships between quality characteristics are linear. On the other hand, it is seen that these simple statistics are required in the estimates of parameters in more complicated assumed functional relationships. What all of this means is that the method of presenting data by the five statistics suggested in Chapter III of the present part constitutes a type of universal tool for the study of interpretation of data.

We have seen, however, that there are some very definite limitations even to the use of these functions for the presentation of data. For example, in the case of two variables it is necessary to know whether or not the line of best fit for the entire series of points on the scatter diagram is also the line of best fit for the means of the appropriate set of arrays and to know whether or not the distribution is homo-scedastic in respect to the line of regression. In other words, an engineer having taken a number of pairs of observations on two quality characteristics, and desiring to tabulate the results of this experiment in terms of simple functions to be used by future generations, must give some idea as to whether or not the conditions mentioned above have been realized in respect to the given set of observations.

As a concrete illustration the tabulation of the averages, standard deviations and correlation coefficient for the case of 1370 observations of depth of penetration and depth of sapwood are not sufficient in themselves to give all of the necessary information in the interpretation of the data. For example, it should also be stated that the regression is approximately linear and that the standard deviation in the column arrays increases in almost linear fashion with the number of the array proceeding from left to right of the scatter diagram.

Naturally it is not to be expected that, for an observed distribution of pairs of values, the line of best fit obtained by minimizing the deviations with respect to one of the variables should at the same time be exactly the line of regression. This will only be true if the regression is strictly linear. Hence to measure this lack of linearity another statistic is often introduced, namely, the correlation ratio of one variable on another. The nature of this measure may be seen as follows: Let n_{xi} be the number of observed points in the i^{th} column array and $\bar{y}_i$ be the observed mean of the y's in this same array. Furthermore let

$$e_i = r\frac{\sigma_y}{\sigma_x}x_i - \bar{y}_i$$

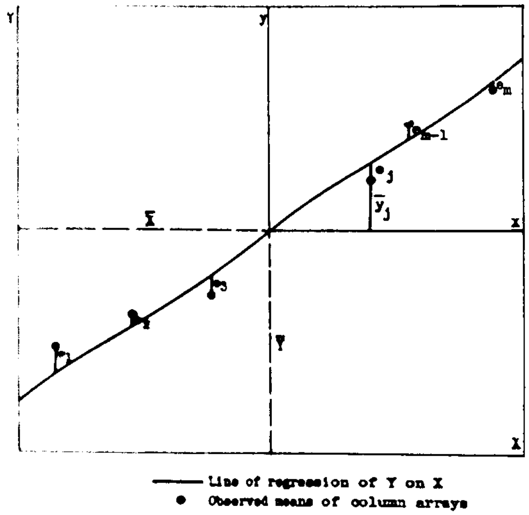


FIG. 2.28

as indicated in Fig. 2.28 then

$$\frac{1}{n}\sum_{i=1}^{m} n_{xi}e_i^2 = \frac{1}{n}\sum_{i=1}^{m} n_{xi}\left(r\frac{\sigma_y}{\sigma_x}x_i - \bar{y}_i\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{m} n_{xi}\left(r^2\frac{\sigma_y^2}{\sigma_x^2}x_i^2 - 2r\frac{\sigma_y}{\sigma_x}x_i\bar{y}_i + \bar{y}_i^2\right)$$

$$= \frac{1}{n}\left[n r^2\frac{\sigma_y^2}{\sigma_x^2}\sigma_x^2 - 2r\frac{\sigma_y}{\sigma_x}nr\sigma_y\sigma_x + n\sigma_{\bar{y}}^2\right]$$

$$= \sigma_{\bar{y}}^2 - r^2\sigma_y^2$$

$$= \sigma_y^2\left(\eta_{yx}^2 - r^2\right),$$

where m is the number of column arrays, $n\sigma_{\bar{y}}^2$ is the sum of weighted squared means of arrays and $\frac{\sigma_{\bar{y}}}{\sigma_y}$ is replaced by the symbol η_{yx}, termed the correlation ratio of y on x.

Similarly we define the correlation ratio of x on y by

$$\eta_{xy} = \frac{\sigma_{\bar{x}}}{\sigma_x}.$$

From what has preceded, we see that, when

$$\eta_{yx}^2 - r^2 = 0,$$

the means of the column arrays of Fig. 2.28 lie on the line of best fit to all the points and the regression is therefore linear.

When

$$\eta_{yx}^2 - r^2 > 0,$$

we see that the means of the column arrays do not coincide with the line of best fit and the regression is strictly not linear. Thus $\eta^2_{yx} - r^2$ is a kind of measure of non-linearity of regression.

Now obviously $\sigma_{\bar{y}} \leq \sigma_y$, so that by definition

$$\eta_{yx} \leq 1,$$

and this of course is also true of η_{xy}. Hence, to sum up, we have

$$r^2 \leq \eta^2_{yx} \leq 1.$$

Previously we have seen that the standard deviation s_y of the points about the line of regression is

$$s_y = \sigma_y \sqrt{1 - r^2} ,$$

and since s_y is always equal to or greater than zero, we have

$$-1 \leq r \leq + 1.$$

It follows from what has just preceded that the analyst should wherever possible give the correlation ratio as a measure of the lack of linearity. Of course, this factor cannot be given in a large number of instances found in engineering work because it must be calculated from grouped data whereas in practice the number of observations is often too small to justify any grouping. In this case perhaps as good a practice as any is merely to record the values of the five statistics, assuming no definite knowledge is available in respect to the nature of the relationship between the two or more variables under consideration.