CoA/N-19

R12.769a

CoA Note No. 19

THE COLLEGE OF AERONAUTICS
CRANFIELD

RATIONAL PRESENTATION OF HISTOGRAM DATA ON SAMPLES

by

J. A. C. WILLIAMS, B.Sc.

R12.769a

3 8006 10057 5847

NOTE NO. 19

JANUARY 1955

THE COLLEGE OF AERONAUTICS

CRANFIELD

Rational Presentation of Histogram Data on Samples

-by-

J. A. C. Williams, B.Sc., A.M.I.Mech.E., A.F.R.Ae.S.

SUMMARY

Misconceptions can be caused by the common method of presenting sample data by histograms. A rational method of treating and presenting histogram data is outlined which attempts to overcome misconceptions caused by the usual method and to make for consistent treatment.

Introduction

A noteworthy increase in the use of statistics has occurred in industry since 1939, and managements now rely on statistical analysis of data in considering action. Particular examples of such occur in:

1. engineering inspection with statistical quality control,
2. engineering design and maintenance with data on fatigue and life tests,
3. personnel management with psychological testing,
4. work study with design of time studies and in the ratio delay technique.

In all these cases data is obtained which must be presented by the technician to various levels of management and therefore clarity of presentation is required otherwise misconception may arise amongst those not trained in statistics. A common example of a form of presentation which may obscure data aften occurs with histograms prepared for samples taken from industrial processes. A typical method of drawing a histogram is shewn (with arbitrary units employed for abscissae) in Fig.1. where the abscissae represent a continuous variable. Amongst the misconceptions which may arise with those unused to statistical concepts are:

(a) the values presented are for a discontinuous variate, whereas a continuous variate is concerned,
(b) values at the extremes suggest that the sample contains individuals which are 'in error' or atypical,
(c) values more extreme than those shewn will not occur.

Such misconceptions are due partly to the method of presentation used, which depends upon a carryover from non-statistical ways of thought and are in no way inherent in the original data. A method is suggested here which is rational in form and which aims to overcome the objections against the usual method of presentation.

General Form of Sample Data

Sample data which has to be presented generally has the following properties:

1. values are obtained which are forced into discrete category values despite the fact that the values are really continuous, e.g. two pins which are measured have diameters .250101" and .250147" might be represented by two values at .2501",

2. the number of values obtained and presented in a histogram form only a sample which has to be representative of an infinite population and from which a hypothesis can be derived. Another sample of equal number from the same

population may give a histogram which would bear superficially no resemblance to the first sample. The sample data is approximate and any rigid categorisation is not appropriate,

3. the infinite population from which the sample is drawn is often assumed to be of the single mode symmetrial 'normal' type. The Sample is generally required to form an estimate of the population from which the sample is drawn,

4. in practice we can never ex hypothesi obtain an infinite population; for managerial purposes, in order to take earliest action possible, to minimise costs, and by the very nature of the data, small samples have to be treated. Typical sample sizes vary from 10 - 50 although samples of 4 - 5 are common. These characteristics will be taken into account in the suggested method of presentation.

Suggested Form of Treatment and Presentation

The form of presentation suggested here relies on a rationalisation from the binomial theorem. As shown in standard textbooks, if p% of a population has a certain property (say red colour) and q% of another property (say blue colour) then the probability of a sample from the population containing red and blue can be derived from the coefficients of terms of $(p + q)^m$ where m is the number in the sample. Thus, if p = q and the sample number be three, we should have the probabilities of picking the following combinations of red and blue from an infinite population:-

Combination	3 red	2 red 1 blue	1 red 2 blue	3 blue
Relative Probability	1	3	3	1

where the probability is given by the numerical coefficients of $(p + q)^3 = p^3 + 3p^2q + 3pq^2 + q^3$. The value of these coefficients are most easily remembered from Pascal's Triangle which in addition illustrates the symmetrical form of the coefficients (for p = q) and the increase in number of the coefficients with the index 'm'.

Index m:	Coefficients.	Histogram sample	
		n	Number
1	1 1	1	2
2	1 2 1	2	4
3	1 3 3 1	3	8
4	1 4 6 4 1	4	16
5	1 5 10 10 5 1	5	32
etc.	etc.		etc.

Each coefficient is derived as the sum of the two terms above it as shewn. In the limit the curve of the coefficients approaches the 'normal' curve.

For any line of the Pascal Triangle the term of the general form $p^r q^{m-r}$ is a discrete category (or class interval) value and the probability of obtaining a particular category value is given by the numbers shewn. Now the class intervals selected in preparing histograms for a continuous variable are arbitrary and determined principally by usage. When a sample is taken we do not know the 'best' class interval; we only know the rounded-off values of the individual members of the sample determined by custom. The method suggested here is to regard the extreme values as associated with the extreme coefficients from the Pascal Triangle from which we can obtain the best (for our purpose) class interval. In order to do this assume the sum of the numerical values of the coefficients in one line of the Pascal Triangle to be the number in the sample to be treated. This imposes a restriction on the sizes of sample that can be treated to powers of 2 but such a restriction can be overcome to some extent by scaling. Taking a sample of size 2^n the expected values should be distributed as shewn on the appropriate line of the Pascal Triangle above. The class intervals will then be $(n + 1)$ in number and the class interval $= \frac{\text{Sample Range}}{n}$

The derivation of these will be seen from Fig. 2. which has been shewn for a sample of 4. If the sample numbers other than a power of 2, values must be allocated in the manner of the nearest 2^n sample as fractional intervals cannot be allowed. The binomial distribution values with which to compare the sample obtained must then be scaled appropriately.

A simple example will be now given to illustrate the suggested method and compare it with that usually accepted.

A sample of five pins are measured and have the following diameters:

.250020, .250101, .249987, .249813, .250147.

It is suggested that the usual method of presenting this data would be to employ class intervals of .0001". Depending upon the selection of the mid point of the class intervals the following histogram values would result.

(a) Mid point.	No. of cases.	(b) Mid point.	No. of cases.
.2498	1	.24985	1
.2499	0	.24995	1
.2500	2	.25005	1
.2501	2	.25010	1
		.25015	1

Employing the method suggested and taking an allocation appropriate to a sample of 4.

$$\text{The class interval then} = \frac{.250147 - .249813}{2}$$

$$= .000167"$$

The class interval limits and the values occurring within those intervals are:

Class intervals	No. of cases.
.249730 - .249897	1
.249897 - .250064	2
.250064 - .250231	2

The three histograms obtained are shewn in Fig. 3.

It will be noted from both Fig.2 and 3. that the suggested method has an implicit correction for sample size to give an estimate of population range which overcorrects the sample data.

Conclusion

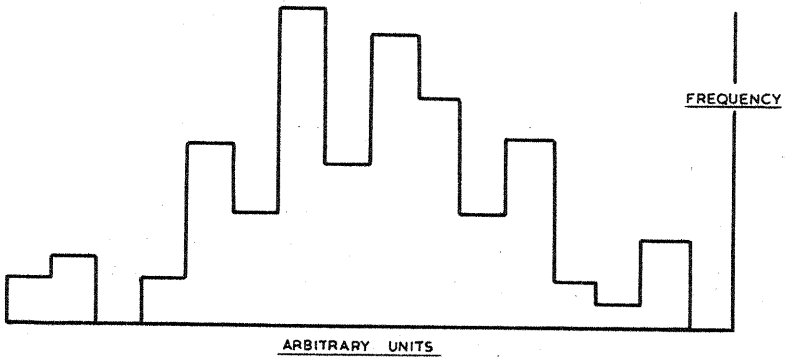
The method shewn can claim to have the following advantages over that normally employed:

1. it is rational and therefore leads to consistent treatment of data especially in treatment of class interval values when the variate is continuous.

2. it is less likely to shew discontinuity such as shewn in Fig.3. Method (a) when the data is in reality of a continuous form and reduces the large differences between adjacent class intervals as seen in Fig.1.

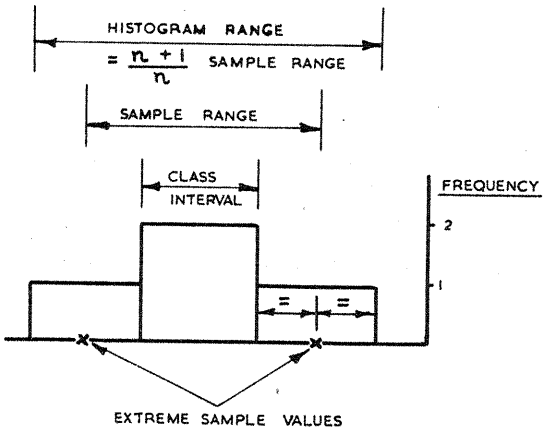
3. the generally increased size of the class intervals gives the suggestion of the approximate nature of data derived from samples which is lost in the normal method.

The main disadvantage would seem to be that the establishment of class intervals involves more arithmetic in the presention of the values especially when more than one sample is involved. The increased clarity of presentation should outweigh these points when data is intended for the use of management. The method is restricted to use with a continuous variate and cannot generally be applied to discontinuous variates.



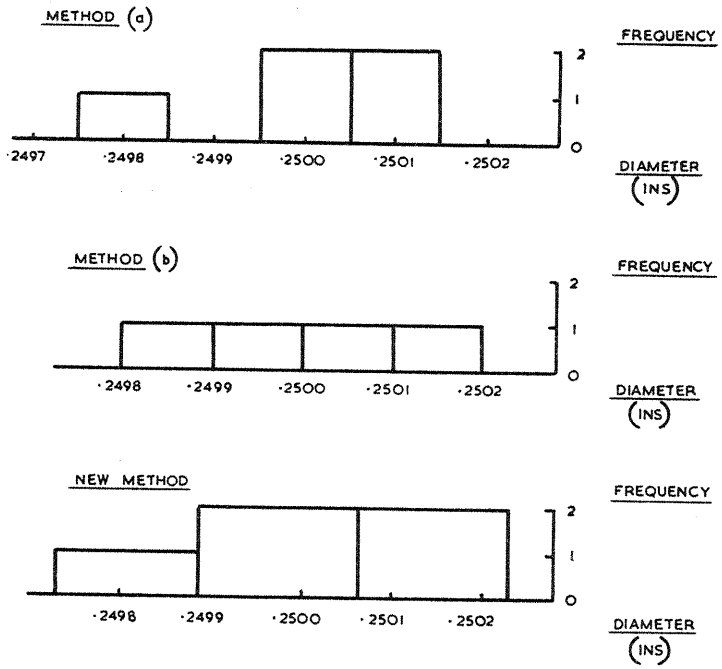
TYPICAL FORM OF HISTOGRAM

FIG. 1.



DERIVATION OF CLASS INTERVAL FOR SAMPLE OF 4.

FIG. 2.



VARIOUS FORMS OF PRESENTATION BY HISTOGRAM

FIG. 3.