CoA NOTE M & P No. 19

THE COLLEGE OF AERONAUTICS
CRANFIELD

SAMPLING TECHNIQUES AND THEIR APPLICATION TO MANAGERIAL PROBLEMS

by

H. C. Wiltshire

CoA Note M and P No. 19

October, 1968

THE COLLEGE OF AERONAUTICS

DEPARTMENT OF PRODUCTION ENGINEERING

Sampling techniques and their application to managerial problems

- by -

H.C. Wiltshire, M.Sc., C.Eng., F.I.Mech.E., F.I.E.E., M.B.I.M.

Summary

In future managers will depend more on numerical information in making decisions, and this information will be more readily available with the extension of the use of computers.

Some of the decisions made by management are of necessity based on information which, although expressed numerically is usually doubtful to a greater or lesser extent. The theory of probability provides a means of quantifying the degree of doubt in mathematical terms.

The basic concepts of sampling techniques are described and although a manager does not necessarily have to be thoroughly acquainted with the mathematical and statistical content, he should at least know what techniques are available and how they should be used. The measurements of central tendency and dispersion are described, together with the Normal and Negative exponential distributions.

The measurements of the standard error of the means for both small and large samples are given and exemplified. A method of determining the number of observations required in a Time Study is introduced. The general theory of significance tests is described and illustrated with examples.

t and χ^2 tests are described and the corresponding tables are appended.

The use of single and bivariate proportionate sampling techniques is illustrated, and the modifications required when the costs of the respective strata samples are incorporated.

The significance of symmetrical and asymmetrical confidence regions is shown. Examples are given of the application of statistical quality control and also of Sequential Analysis.

Acknowledgements

I am indebted to the Literary Executor of the late Sir Ronald A. Fisher, F R.S.; to Dr. Frank Yates, F.R.S., and to Oliver and Boyd Ltd., Edinburgh, for permission to reprint tables of t and χ^2 from their book Statistical Tables for Biological, Agricultural and Medical Research.

I am also indebted to the London University Press for permission to reproduce the tables of probability integrals and the z and x relationships from their publication Introduction to the Computation of Statistics by Shepherd Dawson.

Contents

1. Introduction

Some of the decisions made by management are of necessity based on information which, although expressed numerically is usually doubtful to a greater or lesser extent. The theory of probability provides a means of quantifying the degree of doubt in mathematical terms. The rigorous development of probability theory requires logical and mathematical difficulties of high order but it is assumed that the reader has an elementary knowledge of mathematics and statistics. However, the main concepts will be explained and their applications illustrated with examples.

A knowledge of sampling techniques is of paramount importance in market research, and it is also useful for scientific research, production and work study. It is considered essential that management should at least know what techniques are available and what information they can provide to assist in making decisions.

2. Probability and frequency distribution

The information on which management decisions are based is sometimes quite trustworthy and may take the form of direct relationship between the variables concerned making it possible to provide precise numerical information.

It often happens, however, that the information available, though expressed in the form of numbers, the numbers and the decisions based on them are usually doubtful to a greater or lesser extent. The theory of probability furnishes a means of expressing the degree of doubt in mathematical terms.

Probability and the binomial expansion

As shown in texts on statistics, the measure of the probability of an event occurring or not occurring is given by the terms of the expansion

$$(p+q)^n = p^n + np^{n-1}q + \frac{n(n-1)}{2!}p^{n-2}q^2 + \frac{n(n-1)(n-2)}{3!}p^{n-3}q^3 \ldots + q^n$$

where p represents the probability that an event will occur, and q the probability that it will not occur, p + q being equal to unity.

For example, if p represents the probability that a man aged 20 will die before he is 40 years of age, and q the probability that he will live, and n the number of men under consideration, then the probability that all will die before they are 40 is p^n and the following probabilities follow:-

n - 1 will die and 1 live is $np^{n-1}q$

n - 2 will die and 2 live is $\frac{n(n-1)}{2!}p^{n-2}q^2$

All will live to be 40 is q^n

n - 1 will live and 1 die is npq^{n-1}

n - 2 will live and 2 die is $\frac{n(n-1)}{2!} p^2 q^{n-2}$

If p is the probability of an event and n is the number of cases in a sample then the number of times the event happens in the sample is np. In N samples it will happen npN times and the mean of the sample will be np.

Frequency distributions

Management is often confronted with observed results of a variable the analysis of which requires the use of the observed results in the form of a frequency distribution; in graphical form this means the plotting of the ordinates against the number of times such ordinates occur for each of the varying abscissae.

In practice this is usually effected by grouping the observed data into a small number of frequency classes or class intervals, 10 to 20 being the usual practice. This is done not only to avoid the excessive paper work involved in recording each observed value of the variable, but also to obtain a better understanding of the general distribution of the values. An example of a frequency distribution is given in Fig. 1, which shows the variation in operating time to failure of incandescent lamps. The latter curve is seen to be symmetrical about the x = 1080 axis, but in many cases the curve may be skewed as shown in Fig. 2 in which is shown the numbers of Australian marriages arranged according to the ages of the bridegroom in 3 years groups, and in which is seen a positive skew, the longer tail of which is in the positive direction of x.

Another form of distribution is that known as the negative exponential or Poisson distribution which is based on the relationship

$$f(x) = \frac{1}{\alpha} e^{-x/\alpha}.$$

This curve is illustrated in Fig. 3, and is used in solving problems in queueing.

Moments of a frequency distribution

Frequency distribution may be described and differentiated according to characteristics of their own, and one of the most commonly employed are moments. The moment of a frequency distribution is its descriptive constants and measures its average value, selective scatter of the observations, its symmetry, and other characteristics. In practice the first four moments

$$\sum \frac{fy}{N}, \sum \frac{fy^2}{N}, \sum \frac{fy^3}{N}, \sum \frac{fy^4}{N}$$

usually provide an adequate description of a frequency distribution.

As defined above, the moments of distributions would be computed about the point $Y = 0$, but they may be computed about any arbitrary origin, in practice it is generally most convenient to compute the moments about the most frequently occurring value (as class interval of Y) or about the average value of the distribution ($\bar{Y}$) if the latter is readily available.

Moments are not the only means of describing a frequency distribution, as a large number of auxiliary procedures and formulae exists for the same purpose.

Measures of Central Tendency

The arithmetic mean of a frequency distribution is its first moment viz - $\sum \frac{fy}{N}$ where N is the total number of frequencies. If the frequency distribution is in the class interval unity, as is usually the case, the value of Y for any class interval is taken as the average of all the frequencies in that particular class interval.

When no other information is available the averages of the higher and lower limits in the class intervals are used.

When the origin is translated to the mean, the first moment becomes zero, that is $\sum \frac{Y - \bar{Y}}{N} = 0$ and this is one of the most useful properties of the mean and permits many valuable computational simplifications to be made in statistical analysis. Another useful property of the mean is that the sum of the squares of the deviation from the mean does not exceed the sum of the squares of the deviation from any other value, a property which is used to derive many statistical formulae.

Although the arithmetic mean is the nost frequently used measure of central tendency, it has its limitations as it is strongly influenced by extremely high values and consequently may yield an abnormally high mean for the entire distribution. Furthermore the mean cannot be computed if the distribution contains open-ended intervals unless reasonably accurate estimates of the mid-points of such intervals are possible.

The mode is another measure of central tendency and is the value in a series of observations occurring with the greatest frequency. It is the most meaningful measure in the cases of strongly skewed distributions but it is not used very frequently in practice except for extremely skewed or multi-model distributions.

Measure of Dispersion

The measurement of dispersion, i.e. the manner in which the observations are dispersed throughout the distribution is important because it enables the evaluation of the reliability of a measure of central tending to be made as a true measure of concentration.

Widely used are two measures of dispersion:- standard deviation; and coefficient of variation.

Variance is expressed as $\frac{X^2}{N}$ where $X = (x-\bar{x})$ and where $\bar{x}$ represents the mean, and the standard deviation is the square root of this quantity.

$$\text{Standard deviation} = \sigma = \sqrt{\frac{(x-\bar{x})^2}{N}} \qquad (1)$$

It should be noted that it is expressed in terms of the original units and can be defined as the square root of the mean of the squares of the deviation from the mean.

The coefficient of variation is a measure by which various distributions can be compared, and is expressed as:-

Coefficient of variation = $\frac{\sigma}{\text{mean}}$ and so σ and the mean are expressed in the same units the coefficient of variation becomes an absolute measure, i.e. it has no dimensions.

Measure of Skewness

The absolute measurement of skewness is the third moment, but cannot be used to compare the skewness of different distributions, is usually referred to as α_3:

$$\alpha_3 = \frac{\text{3rd moment about the mean}}{\sigma^3}$$

In general distributions are not considered to be very skewed unless the absolute value of α_3 is at least 2.

Types of Frequency Distribution

If in the binomial expansion

$$(p+q)^n = p^n + np^{n-1} + \frac{n(n-1)}{2!} p^{n-2}q^2 + \ldots.$$

p be given values between 0 and 1 and p+q = 1 the numerical coefficients when plotted give a series of frequency polygons of different shapes and one of them has the equation of the form

$$y = y_o e^{-x^2/c}.$$

This form will be discussed in some detail as it is the one usually described as the Normal curve, and it is the one on which many of the solutions to sampling problems are based.

Distributions of natural characteristics, i.g. stature are normal or nearly so and when the nature of a distribution is unknown and some hypothesis must be made it is usual to assume that it is normal; consequently a knowledge of its properties is extremely useful.

The Normal Curve

Apart from the straight line and the conic sections, no regular mathematical function occurs so widely as the Normal curve, the usual designation nowadays for which is also known as the Normal Probability Density function; the Maxwell Distribution, the Gaussian. On the Continent it was once evocatively termed la courbe de chapeau de gendarme. Its graceful symmetry comforts theoretical physicists and emboldens sociologists.

Francis Galton described it nearly 80 years ago:-

'I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the 'Law of the Frequency of Error'. The law would have been personified and diefied by the Greeks if they had known it. It reigns with serenity and in complete self-effacement effacement amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway ...'

'Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of magnitudes an unexpected and most beautiful form of regularity proves to have been latent all along'.

The historic mainstay of the curve has been its claim to be a distribution of errors, yet the link with errors was forged many years after the discovery of the curve. The Normal curve was first described in a brochure dated 12th November, 1733 by the 66 year-old De Moivre, an outstanding Anglo-French mathematician and a specialist in the theory of probability.

The curve is seen fitted to data concerning the life of electric lamps in Fig. 1 and the curve is also illustrated in Fig. 4. By its use tables giving in numerical terms the probability of the distribution of errors in random samples taken from populations which are normally distributed and such a table is given in Fig. 6.

In modern notation the equation of the Normal curve is

$$y = \frac{N}{\sigma\sqrt{2\pi}} \exp\left(\frac{-x^2}{2\sigma^2}\right) \qquad (2)$$

where N is the total frequency and σ the standard deviation. The curve is continuous and symmetrical about the ordinate y_o, the area between the curve and the axis of abscissae is N; y_o which is the mode ordinate

$= \frac{N}{\sigma\sqrt{2\pi}}$. If the total frequency be expressed as unity (i.e. N = 1) the equation becomes:-

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2} \qquad (3)$$

Probability Integrals

Referring to Fig. 5 let $\frac{1-\alpha}{2}$ be the area included by y_1 the axis of x and the smaller part (tail) of the curve, then $\frac{1+\alpha}{2}$ is the area of the rest, and the area from the ordinate y_0 to the ordinate y_1 is $\alpha/2$.

For example:

when $x = \sigma$ $\frac{1-\alpha}{2} = 0.159$ $\alpha = 0.622$ $y = 0.2420$

when $x = 2\sigma$ $\frac{1-\alpha}{2} = 0.0228$ $\alpha = 0.9544$ $y = 0.0540$

when $x = 3\sigma$ $\frac{1-\alpha}{2} = 0.0015$ $\alpha = 0.9974$ $y = .0044$

The probability integral table given in Fig. 6 is based on the equation (3), and in the case of a total frequency of N the actual value of y given in the table must be multiplied by $\frac{N}{\sigma}$.

Therefore if $N = 1000$ and $\sigma = 10$ then

when $x = \sigma$ $y = .242 \times \frac{1000}{10} = 24.9$

when $x = 3\sigma$ $y = 0.4$

and when $x = 0$, $y = 39.89$.

As an example if the mean stature of 1000 men be 67.5 inches and their standard deviation 2.6 in. and the distribution normal.

The area of the tail $\frac{1-\alpha}{2}$ is 0.159 when $x = \sigma$ and consequently there will be 0.159 x 1000 that is 159 who are taller than 67.5 + 2.6, that is 70.1 ins. and there will be 159 who are shorter than 67.5 - 2.6 that is, than 64.9: there will be 682 between 64.9 and 70.1 in. Since $\frac{1-\alpha}{2} = .023$ when $x = 2\sigma$ there will be 23 above 67.5 + 2(2.6), that is above 72.7 inches and 23 below 62.3 ins.

It will be noted from the table Fig. 6 that the area of the curve enclosed between two ordinates y_1 and y_2 can be easily determined, and that this area gives the probability that an observation taken at random will be between x_1 and x_2. For example, when $x_1 = \sigma$ $\frac{1+\alpha}{2} = .841$; and when $x_2 = 1.5\sigma$ $\frac{1+\alpha}{2} = .933$, therefore in the foregoing example where $\sigma = 2.6$ in. and $1.5\sigma = 3.9$ in., the number of men whose stature is between 67.5 + 2.6 and 67.5 + 3.9 in., that is between 70.1 and 71.4 in. will be 1000 (.933 - .841) that is 92.

The Significance of a Mean

Let the mean stature of all men in the U.K. be $\bar{x}$ and the mean of a large sample drawn at random be $\bar{x}_1$; then $\bar{x}_1$ will be equal to or very nearly equal to $\bar{x}$. The difference $\bar{x} - \bar{x}_1$ is the sampling error. If a large number of other samples of the same frequency be taken each independent of each other, that is, the first sample is put back before the second is taken and so on, all being drawn at random; then the sampling errors are distributed in a curve which is normal or nearly so, even if the statures of the whole population is somewhat skewed.

The standard deviation of the curve is $\frac{\sigma}{\sqrt{N}}$ where σ is the standard deviation of the whole population and N is the number in the sample that is:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} \qquad (4)$$

where $\sigma_{\bar{x}}$ is the standard deviation of the means of the samples.

For example the stature of the adult population in the U.K. is 67.5 inches and its standard deviation about 2.6 inches.

The standard deviation (standard error) of samples of 400 is $\frac{2.6}{\sqrt{400}} = 0.15$ ins.

In practical problems it is not possible to know the mean of the standard deviation of the whole population from the direct measurement of each individual. These values have to be estimated from samples by reasoning backwards from the above conclusions.

Suppose in a random sample of 400, it is required to know the probability of the mean being within certain ranges.

The standard deviation of the mean of the sample $= \frac{\sigma}{\sqrt{400}} = \frac{\sigma}{20}$ and the probability that the mean lies between $+3\sigma_m$ and $-3\sigma_m$, as shown in the table Fig. 6. It is practically certain that the true mean lies within $3\sigma_m$.

Again from the probability integral table in Fig. 6 it can be seen that if $\pm 2\sigma_m$ is taken as the range then the probability is about 1 in 20.

Hence from the mean of the sample and its standard deviation, an estimate can be made of the standard error of the mean and the odds that the true mean will be within any specified distance from the mean of the sample. For example if the average height of 8,500 men in the U.K. is 67.46 inches and the standard deviation of this sample was 2.57 inches, the standard deviation of the mean is $\frac{2.57}{\sqrt{8,500}}$ or .0279.

When $\frac{x}{\sigma} = 1.0$ $\alpha = 0.6826$ (from table Fig. 6) and therefore the chances are 683 in 1000 that the true mean lies between 67.46 ± 0.0279 that is between 67.488 and 67.432.

When $\frac{x}{\sigma} = 3.0$ $\alpha = .997$ (from Fig. 6) and therefore the chances are 997 in 1000 that the true mean lies between 67.46 ± 3 (.0279) i.e. between 67.544 and 67.376 inches.

Standard Errors and Confidence Intervals

Standard error is the estimated value of the standard deviation of the dample mean in the population, i.e. estimated from the sample data. For example if one sample had been taken from data having a reasonably normal distribution there would be a 68% probability of being correct if it were stated that a true mean lies within an interval of 1 standard error of the mean. This interval is known as the confidence interval and the associated probability known as the confidence coefficient expressed either as a decimal or as a percentage e.g. (.68 or 68%).

A Reasonably Normal Distribution

Since statistical analyses are based on reasonably normal distributions some attempt must be made to define what is meant by reasonably normal. Some points are now evident, it must be unimodel and the percentages of observations falling within 1 standard error of the mean might be stipulated to be at least 60 to 75%; the corresponding percentages being 90 to 99 of the observations fall within 2 standard deviations. Another stipulation would be that the absolute value of α_3 (the measure of skewness) is less than 2.

Parameters Defining a Complete Normal Distribution

Whenever a variable may be assumed to have a normal distribution, a knowledge of only two parameters, the mean and the standard deviation is sufficient to specify the entire distribution.

When the distribution is known it is then possible to make probability statements about the degree to which observed statistics approximate to the true unknown parameters. For example if it can be assumed that the age of motor cars can be normally distributed and the average age of a sample of cars was found to be 4 years with a standard deviation of 2.0 years, then it is possible to determine the precentage of cars outstanding between any two age limits.

3. Basic Concept and Rules for Random Sampling (Unrestricted Sampling)

The basic concept of random sampling is that the analysis of a small carefully selected segment of a population will yield information about that population almost as accurate as if the whole population had been studied; what is needed is a representative sample and not necessarily a large sample.

In random sampling each unit correctly included in the population sample should have exactly the same chance of being included in the final sample as any other unit in the population.

It can be said that the basis of sampling rests on two quite simple statistical rules.

(i) The rule of statistical regularity states that any group of objects taken from a much larger group will tend to have the same characteristics as that of the larger group.

(ii) The rule of large numbers states that large groups are statistically more stable than small groups.

The standard error expressed as a percentage gives the extent to which any data estimated from the sample will vary from the true value of the population from which the sample has been drawn; for example:-

68% of the units will be within ± 1 standard error of the mean
95% of the units will be within ± 2 standard errors of the mean
99% of the units will be within ± 3 standard errors of the mean.

Random (Unrestricted Sampling) (or simple sampling)

The two most common measures in business research are the mean and the percentage; the mean where variables are being investigated and the percentage when attributes are under observation.

$$\left.\begin{array}{l}\text{The variance of the mean of}\\ \text{an unrestricted sample}\end{array}\right\} = \frac{\text{Variance of the population}}{N}$$

where N is the number of observations in the sample.

In practice the square root of the variance of the mean, that is the standard error of the mean (denoted by $\sigma_{\bar{x}}$) is employed as the quantitative measure of dispersion $\sigma_{\bar{x}}$ is the estimate from the sample of the standard error of the mean.

$$\left.\begin{array}{l}\text{The standard error of the mean}\\ \text{of an unrestricted sample}\end{array}\right\} = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

As seen from the table in Fig. 6, there are 68 chances in a hundred that a range of sample mean 1 standard error will contain the population value.

The standard deviation can be conveniently calculated by using the relationship

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - d^2}$$

where d is the difference between the true and the arbitrary mean.

Let X be the variable measured from the arbitrary mean; and x its value when measured from the true mean.

Then $X = x + d$

and $X^2 = x^2 + d^2 + 2xd.$

$$\therefore \quad \Sigma(X^2) = \Sigma x^2 + \Sigma d^2 + 2d\Sigma x$$

$$= \Sigma x^2 + nd^2$$

for $\Sigma(x) = 0$ being the sum of deviations from the mean

$$\therefore \quad \frac{\Sigma X^2}{N} = \frac{\Sigma x^2}{N} + d^2 = \sigma^2 + d^2$$

$$\sigma^2 \equiv \frac{\Sigma X^2}{N} - d^2$$

It is often convenient to take the arbitrary origin at zero (or to assume that the mean is zero) and in this case

ΣX^2 = sum of the squares of the variables

d = the mean of the variables

and the formula $\sigma = \sqrt{\frac{\Sigma X^2}{N} - d^2}$ can be stated as

The square of the standard deviation is equal to the mean of the squares of the variables minus the square of their mean,

or can be expressed as:

$$\sigma = \sqrt{\frac{\overline{x^2}}{N}} = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma \bar{X}}{N}\right)^2} \qquad (5)$$

and the standard error of the mean $= \frac{\sigma}{\sqrt{N}} = \sqrt{\frac{1}{N}\left[\frac{\Sigma X^2}{N} - \left(\frac{\Sigma \bar{X}}{N}\right)^2\right]}$

The variance of an unrestricted sample in the proportional method is given by:-

$$\sigma^2 = \frac{pq}{N} \qquad (6)$$

where p is the proportion of the samples possessing given attributes, and q is the proportion not having the given attributees. For example in a survey among 971 English school-children revealed that 24% expected to take up teaching. Assuming this to be an unrestricted representative sample of all English school-children what is the true percentage if a confidence coefficient of 0.95 is desired?

The standard deviation of the population = $\sqrt{\frac{pq}{N}}$ or $\sqrt{\frac{(.24)(.76)}{971}}$

= .0137 or 1.37%.

From the table (Fig. 6) it is seen that .0228 of the area lies between the mean and $\pm$ 2.0σ, hence with a confidence coefficient of 0.95 the range would be 24% $\pm$ 2 x 1.37 or between 26.74 and 21.26%.

When the sample size is small (less than 30) the standard error formula must be modified to correct for the natural tendency of the standard deviation of a small sample to underestimate the true standard deviation. The necessary correction factor is the substitution of N-1 for N.

Then $$\sigma_{\bar{x}} = \sqrt{\frac{1}{N-1}\left[\frac{\Sigma X^2}{N} - \left(\frac{\Sigma \bar{X}}{N}\right)^2\right]} \qquad (7)$$

and in the proportional form

$$\sigma_{\bar{x}} = \sqrt{\frac{pq}{N-1}} \qquad (8)$$

<u>The Correction Factor when the Sample is Large in Relation to the Population</u>

The standard deviation of any statistic is affected not only by the absolute size of the sample but by the relative size of the sample to the population. Thus for a large sample the correct expression is:-

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{N}\left(1 - \frac{N}{P}\right) \text{ or } \sigma_{x} = \frac{\sigma}{\sqrt{N}}\sqrt{1 - \frac{N}{P}}$$

where σ^2 is the estimated variance of the population

N is the size of the sample

P is the size of the population

the adjustment factor is $\sqrt{\left(1 - \frac{N}{P}\right)}$ which however may be neglected, in general, if the sample constitutes less than 4% of the population. If it is willing to accept an error as high as 5% of the standard error estimate, the adjustment term may be neglected providing the sample is less than 10%, but where the sample comprises 10% or more of the population the adjustment term should certainly be employed.

For example if the variance of a population is known to be 70 the standard error of a random sample of 280 families would be:

$$\sigma_{\bar{x}} = \sqrt{\frac{70}{280}} = \sqrt{.25} = .5$$

As the sample comprised 25% of the population then the true standard error would be:

$$\sigma_{\bar{x}} = \sqrt{.25(.75)} = 0.433$$

a difference of 13.4% in the standard error computations.

An example of the use of the foregoing relationships and the use of the Normal distribution curve will now be given.

Twenty observations on the heights in inches of boys taken at random from a given age group are:

63.0	66.4	60.3	68.7	61.0
65.1	63.8	65.6	61.2	70.3
69.3	60.1	63.9	63.6	62.5
61.6	64.9	67.7	70.1	64.9

If the estimate of the average height of all boys in the age group is to have a 95% chance of being correct to $\pm\frac{1}{2}''$ it is required to determine how many observations will be necessary. The first step is to determine $\ddot{x}$ the average of the sample

$$\bar{x} = \frac{\Sigma x}{N} = \frac{1294}{26} = 64.7$$

The next step is to compute an unbiased estimate of the variance from the relationship

$$\sigma(x) = \frac{\Sigma(x-\bar{x})^2}{N-1}$$

The value of $\Sigma(x-\ddot{x})^2$ is computed as shown in the following table:-

x	$x-\bar{x}$	$(x-\bar{x})^2$
63.0	-1.7	2.89
65.1	+0.4	0.16
69.3	+4.6	21.16
61.6	-3.1	9.61
66.4	+1.7	2.89
63.8	-0.9	0.81
60.1	-4.6	21.16
64.9	+0.2	0.04
60.3	-4.4	19.36
65.6	+0.9	0.81
63.9	-0.8	0.64
67.7	+3.0	9.00
68.7	+4.0	16.00
61.2	-3.5	12.25
63.6	-1.1	1.21
70.1	+5.4	29.16
61.0	-3.7	13.69
70.3	+5.6	31.36
62.5	-2.2	4.84
64.9	+0.2	0.04
TOTALS 1294.0		197.08

The unbiased estimate of the variance

$$\sigma^2 = \frac{\Sigma(x-\bar{x})^2}{N-1} = \frac{197.08}{19}$$

$$= 10.37$$

Estimated standard deviation = $\sqrt{10.37}$ = 3.22

It is now assumed that the heights of all the boys in the age group are distributed with a standard deviation of 3.22 inches but with an unknown mean. From the table given in Fig. 6 it is seen that 95% of a normal distribution lies within ± twice the standard deviation of the mean.

Since it is required in this example that there is to be a 95% chance of the sample mean being within $\frac{1}{2}''$ of the population mean, the value of N can now be determined

$$\frac{2\sigma}{\sqrt{N}} = \text{permissible error}$$

$$\frac{2 \times 3.22}{\sqrt{N}} = \frac{1}{2}'' \qquad \therefore \; N = 166 \text{ observations.}$$

4. The Determination of the Number of Observations required in a Time Study

A similar method could be used to establish the number of observations required in a Time Study for a specific permissible error from the recorded actual times; the greater the variation in the observed times the greater will be the number of observations required to maintain a given order of accuracy.

Another more rapid method is based on the range, i.e. the difference between the highest and lowest value recorded; the relationship between the values of σ and the range w are given by the expression:

$$\sigma = kw \qquad (9)$$

and the various values of k for different numbers of observations in the sample are given below.

N Number of observations in sample.	k	N	k
2	.886	9.	.337
3	.591	10.	.325
4	.486	12.	.307
5	.430	14.	.294
6	.395	16.	.283
7	.370	18.	.275
8	.351	20.	.268

The estimate of σ from the range is more efficient when $N < 8$ and the method is shown in graphical form in Fig. 7 for samples of 4 units.

This provides a rapid means for the determination of the approximate number of observations required for a given range and accuracy required.

For example the four successive actual times taken in a Time Study are:-

13 secs.
10 secs.
11 secs.
14 secs.

48 secs.

From which $\bar{x} = \frac{48}{4} = 12$ secs.

Range = 14 - 10 = 4 secs.

$$\frac{\text{Range}}{\bar{x}} = \frac{4}{12} = .33$$

From the graph in Fig. 7 for a permissible error of 5% the number of observations required would be approximately 45.

Calculation of σ and $\sigma_{\bar{x}}$

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{N-1}}$$

The use of N-1 in the denomination, as explained before, is to offset the bias due to small samples. $\sigma_{\bar{x}}$ is the standard deviation of the means of sample of size N drawn from the population which is estimated to have a standard deviation of σ, the two quantities being related as follows:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

This relationship is based on an important theorem of statistics that for any population, regardless of distribution, if it has a finite variance σ^2 the variance of the means of samples of size N drawn from the population will be equal to the population variance divided by the sample size or:-

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{N}$$

The proof of this theorem can be found in texts on the theory of statistics. A second important theorem is the central-limit theorem which states that if a population has a finite variance σ^2 and a mean μ then the distribution of means of samples of size N drawn from the population approaches the normal distribution with a mean and a variance of

$$\frac{\sigma^2}{N}.$$

These two theorems permit the calculation of the estimated standard deviation of the mean from the estimated standard deviation of the sample, and permit the application of the t distribution to the sample mean regardless of the distributions of the populations from which the sample was drawn.

The Testing of Hypothesis

Sampling surveys may be designed to estimate a population characteristic, but also may be used in conjunction with a device to test the validity of some supposition or hypothesis concerning the population. Testing a statistical hypothesis comprises the evaluation of the significance of one or more related values prescribed by some hypothesis or theory. The problem in its original form is usually non-statistical in character but its solution depends on statistical analysis.

For example the question whether an 8% difference in average consumer purchases of a certain product between two cities as revealed by a sample survey indicates a real difference between the average purchases of the product or is due to sampling variations. In order to answer this question the problem must be reformulated in more statistical terms with the hypothesis stated as follows:

the 8% difference is due entirely to the chance element in the sample selected.

In this null, or negative hypothesis the original problem is restricted in the form of an hypothesis alleging that any observed or apparent differences are not significant but are due solely to random sampling fluctuations. The appropriate test of significance is then applied to yield results that will either support or contradict the hypothesis, and it is on this basis of these results that the hypothesis is either accepted or rejected. It should be noted that the results of a significance test do not definitely prove or disprove an hypothesis, they just lend support or cast doubt upon it.

5. The General Theory of Significance Tests

It is frequently necessary to determine the significance of a sample relative to itself or with respect to a value drawn from another sample. Also it is often desired to know whether the results of a sample survey do or do not contradict the theory that the true value is some specified value other than the sample value, or whether a sample could conceivably been drawn from a certain population as indicated by the significance of the difference between certain representative values and the population values. One of the most recurrent problems in market analysis is the significance of the difference between two sample surveys, taken either at the same time, or at different times.

In order to determine the significance of any difference it must be calculated what part of that difference is attributable to random sampling fluctuations, the universal measure of sampling fluctuations being the standard error.

Thus, given a mean value of a normally distributed population it is known that the mean value of random samples drawn from this population will differ from the true value, (due to random sampling operations) by more than 1 standard error approximately 32 times out of 100, or by 2 standard errors of the mean 45 times out of 1,000. The table in Fig. 6 indicates that roughly 68% of the area of the normal curve lies between ± 1 standard deviation of the mean, and it follows that 32% of the area will be outside this range.

If a sample statistic were found to be, say, 3.5 standard errors, then the membership in this population would be seriously questioned because the extreme rarity of such an occurrence (1 time in 2,000) makes it unlikely that the sample could have been drawn from that population.

It is because of this slim possibility, incidentally, that a statistical hypothesis cannot be confirmed <u>definitely</u> or denied.

The significance of a sample difference can readily be evaluated by first calculating the difference between the sample value and the other (population or sample) value in terms of standard errors. This is effected by dividing the difference between the two statistics by the standard error of the difference between the two statistics as estimated from the sample. The probability of the difference is then determined from an appropriate distribution table.

If the probability is high it indicates that there is little relative difference between the two samples and the difference consequently might easily be due to random sampling, the difference being taken as being non-significant.

The preset level of probability is known as the significance level, and for most business purposes it is usually set at 0.05. The hypothesis device is used for the normal distribution, the t, and X^2 distributions subsequently described.

The t Test

t can be defined as the difference between the mean of a sample and the true mean of the population from which it was drawn, divided by the estimated standard deviation of the mean.

Thus if μ represents the mean of the population

$$t = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \qquad (10)$$

The t distribution is not a normal distribution, although it approaches the normal under certain conditions.

It is an independent mathematical function which depends on the number of measurements involved in the computation of $\sigma_{\bar{x}}$.

If N is the number in the sample, ranging from 2 to ∞, t has a different distribution for each value of N, becoming equivalent to a normal distribution when N becomes infinite.

n, the number of degrees of freedom in the t test is taken as N - 1. For example, referring to table Fig. 8 which lists the values of t for various values of n, with 5 degrees of freedom (n = 5) there is only a 0.05 probability that t will be larger than 2.57, and if a t value is observed to be 2.90 then with a .05 chance of being wrong it would be considered that the null hypothesis is false; since an occurrence of a t value of 2.9 is an event which has less than .05 chance of happening if the hypothesis were true.

The level at which probability becomes significant

A probability level of 0.05 (confidence coefficient of .95) is used by many as the dividing line, and means that with a confidence coefficient of 0.95, in 95 cases out of a 100 the mean of a large sample will fall within 1.96 standard errors of the true mean when the sample is drawn from the same population.

If the difference could have occurred 1 time in 20 it is concluded that the difference is too large to be attributed to chance variations and is therefore significant. The probability level of 0.01 (confidence coefficient 0.99) is sometimes used by others who claim that if this difference could have arisen 1 time in 100 then it could be considered that factors other than chance could have caused this difference.

Standard Errors and Small-Sized Samples

The standard error of a statistic is modified by the size of the sample, because if the sample is relatively small there will be insufficient members to counteract the play of erratic sampling conditions, resulting in a sampling distribution somewhat different from the normal distribution.

The means of small-sized samples are distributed according to the t distribution which is shown tabulated in Fig. 8. As the size of the sample decreases it is seen that the t distribution diverges more and more from the normal distribution with increasing dispersion.

Thus although there is a 0.95 confidence coefficient that the sample mean $\pm$ 1.96 standard errors will contain the true population value for very large values of N, the allowable error is increased to 2.06 standard errors when the size of the sample is reduced to 25, and for samples of 3 the limits become the sample mean $\pm$ 3.182 standard errors.

In general, if the sample size exceeds 30 the t distribution approaches the normal distribution, but for less than 30 members the t distribution should be used, the degree of freedom (n) is taken as one less than the size of the sample, i.e. N - 1.

Degrees of Freedom

The degree of freedom in the t test is the number of indepedent measurements that are available for the calculation of the estimated standard deviation. If the standard deviation is known or assumed to be known from large amounts of previous data then there is no restriction of the standard deviation computation and consequently t values based on degrees of freedom of ∞ may be employed.

In the usual test when the standard deviation is estimated from N values, it is necessary to calculate the deviation from the sample mean where n the number of degrees of freedom is equal to N - 1.

The values of t for different degrees of freedom are given in the table Fig. 8; for example, with seven degrees of freedom there is only a 0.05 probability of observing t greater than 2.345 or smaller than - 2.365.

Estimate of True Mean

Suppose several measurements were made $x_1, x_2, \ldots, x_n$ and their mean $\bar{x}$ is calculated, $\bar{x}$ being the estimate of the true population mean μ.

For example, if in the analysis of a batch of material a mean of 17.14% nickel was found while the specification calls for 17.0% nickel, does the material meet the specification? The null hypothesis for a test of this nature is that the true mean μ is equal to the specified value μ_s, or

$$\text{Hyp. } \mu = \mu_s$$

To test the hypothesis it is necessary to estimate both μ and $\frac{\sigma}{\sqrt{N}}$, the best estimates of these two quantities are $\bar{x}$ and $\sigma_{\bar{x}}$ respectively.

The difference between two means

If $\bar{x}_1$ and $\bar{x}_2$ are two separate means and the numbers in the samples are respectively N_1 and N_2 the hypothesis is stated as $\mu_1 = \mu_2$.

There are available two estimates for the standard deviation.

$$\sigma_{\bar{x}_1} = \sqrt{\Sigma(x_1-\bar{x}_1)^2/N_1 - 1}$$

and

$$\sigma_{\bar{x}_2} = \sqrt{\Sigma(x_2-\bar{x}_2)^2/N_2 - 1}$$

These estimates are of the same thing and therefore can be pooled to give a better estimate.

The pooled estimate of the standard deviation is calculated from the following equation

$$\sigma_{p\bar{x}} = \sqrt{\frac{\Sigma(x_1-\bar{x}_1)^2 + (x_2-\bar{x}_2)^2}{N_1+N_2-2}}$$

This relationship is equivalent to weighting each estimated variance by the number of degrees of freedom available for its calculation. The expression to test the difference between two means is:-

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{p\bar{x}} \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \qquad (11)$$

where $\sigma_{p\bar{x}}$ is the pooled estimate of the standard deviation from both sets of data.

The last term in the denomination in equation $\sigma_{p\bar{x}} \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$ is to adjust the value of the pooled estimated standard deviation to be the estimated standard deviation of the difference between two means being equivalent to

$$\frac{\sigma_{p\bar{x}}}{\sqrt{N_1}} + \frac{\sigma_{p\bar{x}}}{\sqrt{N_2}}$$

Two numerical examples will now be given illustrating the use of the t test.

Suppose the corresponding values of two characteristics obtained from small samples are as shown in the following tables; and it is required to know if the difference between the means is significant at the 0.05 level.

	Group A			Group B		
	x_1	$x_1 - \bar{x}_1$	$(x_1 - \bar{x}_1)^2$	x_2	$(x_2 - \bar{x}_2)$	$(x_2 - \bar{x}_2)^2$
	12	1	1	9	0.6	0.36
	13	2	4	8	-0.4	0.16
	11	0	0	8	-0.4	0.16
	9	-2	4	9	0.6	0.36
	10	-1	1	8	-0.4	0.16
Σ	55		10	42		1.20

In group 1 $\bar{x}_1 = \frac{55}{5} = 11;$ $\Sigma(x_1-\bar{x}_1)^2 = 10$ $N_1 = 5.$

In group 2 $\bar{x}_2 = \frac{42}{5} = 8.4$ $\Sigma(x_2-\bar{x}_1)^2 = 1.2$ $N_2 = 5.$

The pooled value of $\sigma_{p\bar{x}} = \sqrt{\frac{\Sigma(x_1-\bar{x}_1)^2 + \Sigma(x_2-\bar{x}_2)^2}{N_1 + N_2 - 2}}$

$$= \sqrt{\frac{10 + 1.2}{5 + 5 - 2}} = \sqrt{\frac{11.2}{8}} = 1.183$$

$$\frac{\text{Difference of means}}{\text{its standard deviation}} = \frac{11 - 8.4}{\sigma_{p\bar{x}}\sqrt{\frac{N_1+N_2}{N_1N_2}}} = \frac{2.6}{1.183 \times \sqrt{\frac{10}{25}}} = \frac{2.6}{.747} = 3.48$$

From the table Fig. 8 the tabulated value of t for n = 8 at the 0.05 level is 2.306 and as the calculated value is 3.48, the difference of the means is indicated as significant.

In the second example, a plant is supposed to yield a petroleum product having an octane number of 90.0. The data given in the following table are nine daily analysis with a mean of 89.5; and it is required to know whether this could arise from a population with a mean of 90.0 at a 0.05 probability level.

The null hypothesis is set as $\mu = 90.0$ and the significant level at 0.05. To determine t it is necessary to estimate μ and $\frac{\sigma}{\sqrt{N}}$ and for these $\bar{x}$, and the estimated standard deviation of the means are used.

	x	$x - \bar{x}$	$(x - \bar{x})^2$
	91.1	1.6	2.56
	88.9	-0.6	.36
	89.7	+0.2	.04
	90.9	+1.4	1.96
	88.6	-0.9	.81
	89.3	-0.2	.04
	88.6	-0.9	.81
	89.5	0.0	0
	88.9	-0.6	.36
TOTALS	805.5		6.94

$N = 9 \quad \Sigma x = 805.5 \quad \bar{x} = \frac{805.5}{9} = 89.5$

$$\sigma_x^2 = \frac{\Sigma(x-\bar{x})^2}{N-1} = \frac{6.94}{8} = 0.8675$$

$$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{N} = \frac{.8675}{9} = .0964$$

$$\sigma_{\bar{x}} = 0.3105$$

Assuming the hypothesis to be true, i.e. that $\mu = \mu_s$

$$t = \frac{\bar{x} - \mu_s}{\sigma_{\bar{x}}} = \frac{89.5 - 90.0}{.3105} = -1.61$$

Since there are 9 observations, there are 8 degrees of freedom and the corresponding values of t at the .05 level and .10 levels are respectively from the table in Fig. 8, 2.306 and 1.86. The calculated value of t does not exceed the tabulated value, so that pending further data the plant is providing petrol with an octane number of 90.0 although the mean of the analysis is 89.5.

6. One-sided Tests (Asymmetrical Confidence Regions)

The acceptance limit so far has been taken as 0.05 probability level to include 0.025 of the area of the normal curve above the mean value and .025 of the area of the normal curve below the mean value constituting the range of a statistic ± 1.96 times its standard error. (For small sized samples where the t distribution is applicable the multiple will be 2.20 for samples of 12 and 2.06 for samples of 25 as given in the table Fig. 8. However, depending on the character of the problem a given probability level, or confidence coefficient need not necessarily be symmetrical.

When an asymmetrical confidence region is selected the significance or non-significance depends on the direction of the differences as well as on its magnitude. Thus if a 1% above and a 5% below normal probability is used, a sample statistic would have to exceed the population value by more than 2.33 standard errors before the difference would be taken as significant, but it would only need 1.65 standard errors less than the population value for the same decision to be reached as can be seen from the table Fig. 6.

As an example suppose that it is known from past records that the readership of a particular journal was 20% of the families in a given locality, but after a vigorous sales campaign conducted in the locality it was found in an unrestricted sample of 195 families approximately 26% were now reading the journal.

A decision must be reached as to whether the facts indicate the expenditure of funds to conduct the same sales campaign on a nation wide scale; and consequently the publishers are anxious to evaluate the effectiveness of the sales campaign.

The standard error is

$$\sigma = \sqrt{\frac{pq}{N}} = \sqrt{\frac{(.26)(.74)}{195}} = .0314$$

Using the symmetrical confidence region of 0.05 probability the region of acceptance is computed as 20% ± 1.96 x 1.34 or an interval between 32.14 and 19.86.

Since the value of 20% is within this region of acceptance the observed difference is concluded to be due to sampling fluctuations and is not significant with the inference that the promotional campaign was not successful. But in this problem the exclusive interest is in the significance of the sample values above that of the population or the symmetrical confidence region of 0.05 probability stipulated in this problem includes .025 above and .025 below the mean.

It is of no interest whether the samples are less than 20% because then it would be obvious that the promotional campaign had failed to increase the readership above the previous level; the entire interest in this problem is to avoid an erroneous decision.

From the probability integral table Fig. 6 it is seen that the area corresponding to + .05 probability is given by $\frac{x}{\sigma}$ Of 1.645, so that the asymmetrical confidence coefficient of .05 results in the acceptance interval 26% ± 1.645 x 3.14 or 31.16 and 20.84. The observed percentage is now outside the range of acceptance thereby inferring that the publisher would be advised to extend the sales compaign.

In common with the normal distribution the t distribution is symmetrical and the t values like those of the normal distribution measure the positive or negative values. A t value of 2.228 for 0.05 probability for 10 degrees of freedom means that there is a 0.025 probability of observing a t value greater than 2.228 and a 0.025 probability of observing a t value less than -2.228 where $\bar{x}$ is drawn from a population with a mean μ and t is calculated from the relationship in equation (10). When a t test is arranged to check whether a measured mean can be equal to some theoretical mean it is considered that there is a significant difference between the two if the measured mean is either greater or less than the critical value as computed by the equation (11).

There are many occasions however, when the interest is in establishing whether a measured value is either greater or less, but not both, than some critical value. For example, the purchaser of a substance which is specified to be, say, 99.3% pure might be interested in determining whether the mean value of a sample is significantly less than this purity, but not particularly interested if the purity is higher.

The hypothesis now becomes that the true mean is not less than a specified value, or:-

$$\text{Hyp.} = \mu > \mu_s$$

The corresponding relationship for t becoming

$$t = \frac{\mu_s - \bar{x}}{\sigma_{\bar{x}}}$$

If the alternative hypothesis is required, that the true mean is not greater than a specified value, the null hypothesis may be written:

hyp. $\mu \leq \mu_5$ and the corresponding t equation

$$t = \frac{\bar{x} - \mu_s}{\sigma_{\bar{x}}}$$

The critical values of t for these one-sided tests are those at the probability levels corresponding to twice the significant levels present. If a 0.05 significant level has been present the hypothesis is to be rejected if the calculated t value exceeds the 0.10 level because the 0.05 level corresponds to the probability for deviations of equal magnitude on either side of the mean.

7. The Test for Goodness of Fit (χ^2) (chi-squared)

The goodness of fit between observed frequencies and the frequencies in a normal distribution can be measured by a quantity χ^2 (pronounced kye squared), computed from the relationship:-

$$\chi^2 = \Sigma \frac{(f - f_t)^2}{f_t} \qquad (12)$$

where f is the observed frequency and f_t the theoretical or expected frequency; χ^2 measures the difference between the total actual and the total theoretical frequencies and tables are available which show the probability that any given value of χ^2 will occur by chance; thus giving the probability of a relationship (or difference) between two sets of variation.

In using the χ^2 test certain cautions are necessary; the total frequency must be the same in both the theoretical and the observed distributions; the number of groups should be the same and no group should contain less than ten individuals, all the individual members of the samp,e must be independent. Pearson the eminent mathematician recommends that small frequencies such as occur at the ends of a distribution should be grouped together.

As an example of the use of the χ^2 test consider a company using a large number of small transformers 30% of type A and 70% of type B, both types of the same rating. After a certain period there has been 452 burned out, 74 transformers of type A and 378 of type B.

Is there any evidence of a difference between the two types?

If there is no difference between the too types the expected or theoretical number burnt out would be in the same proportion as the numbers used.

Type A $\quad .3 \times 452 = 136$

Type B $\quad .7 \times 452 = \underline{316}$

$\underline{452}$

$$\chi^2 = \Sigma \frac{(f - f_c)^2}{f_c} = \frac{(74 - 136)^2}{136} + \frac{(378 - 316)^2}{316}$$

χ^2 being an additive function.

$$\chi^2 = \frac{(-62)^2}{136} + \frac{(62)^2}{316} = 40.5$$

Since there are only two categories comprising the total there is only one degree of freedom $(n = 1)$ and from the table Fig. 9 there is only a .0001 probability that χ^2 is larger than 10.827 so that the hypothesis that the two types are equal as regards burning out must therefore be rejected.

Degrees of Freedom and χ^2 evaluation

When an observed distribution is compared with a theoretical distribution, the sum total of the discrepancies between them will depend partly on the number of sub-groups.

If in forming the theoretical distribution it has been made to conform partially with the observations by using constants of the latter, then the freedom of comparison has been to some extent limited and allowance must be made for this.

If there are N groups in each distribution then to find the number of degrees of freedom from n must be subtracted the number of degrees used in the observed data in finding the theoretical distribution. If totals have been made equal it is N-1; if the means also have been made equal it is N-2.

8. Other Methods of Sampling

So far only the simple unrestricted random sampling has been considered, but in many cases other methods are preferable, and these other methods can be classified as i) single proportionate sample, ii) Bivariate proportionate sampling.

i) Single proportionate sampling

Often a better and cheaper method of taking samples is to divide the population into sections or strata and to take samples from each stratum. These stratified samples are of such a size that the relative size of each sample stratum is proportional to the relative size of the corresponding stratum in the population. In the case of the single proportionate sampling this is the only condition and the relationship can be expressed as

$$\frac{N_1}{P_1} = \frac{N_2}{P_2} \ldots\ldots \frac{N}{P} \qquad (13)$$

where N_1 N_2 are the numbers in the sample strata and P_1 P_2 are the corresponding proportions in the population. This single proportionate method is often used in practice but it is not theoretically valid unless the respective standard deviation in the various strata are equal, i.e. that the strata are homogeneous. To take heterogeneity into account, the proportions are modified by the respective standard deviations to produce a bi-variate proportional sampling method.

ii) Bi-variate proportionate sampling

The bi-variate proportionate sampling method can be expressed as

$$\frac{N_1}{P_1\sigma_1} = \frac{N_2}{P_2\sigma_2} \ldots\ldots = \frac{N}{P\sigma_s} \qquad (14)$$

The sampling variance of the mean in any single stratum is $\left(\frac{P_i}{P}\right)^2 \cdot \frac{\sigma_i^2}{N_i}$ therefore the sampling variance of the entire samples is a weighted average of individual variances, the weight of each stratum being the square of its relative size in the population. The relationship can be expressed as

$$\left.\begin{array}{l}\text{Sampling variance}\\ \text{of the bi-proportionate}\\ \text{sample}\end{array}\right\} = \left(\frac{P_1}{P}\right)^2 \frac{\sigma_1^2}{N_1} + \left(\frac{P_2}{P}\right)^2 \frac{\sigma_2^2}{N_2} + \left(\frac{P_3}{P}\right)^2 \frac{\sigma_3^2}{N_3}$$

$$= \sum_{i=1}^{s} \left[\left(\frac{P_1}{P}\right)^2 \frac{\sigma_i^2}{N_i}\right]$$

$$= \sum_{i=1}^{s} W_i^2 \frac{\sigma_i^2}{N_i} \qquad (15)$$

where P_i = size of each respective stratum in the population, there being s strata.

P = size of total population

W_i = relative size of each stratum in the population

$$= \frac{P_1}{P}$$

N_i = number of sample members in each stratum

σ_i^2 = variance of each stratum.

In the case of the mean, the formula becomes:-

$$\sigma_{\bar{x}}^2 = \sum_{i=1}^{s} W_i^2 \frac{\sigma_i^2}{N_i} = \sum_{i=1}^{s} \left[\frac{W_i^2 \sum_{i=1}^{N_i} (X_{ij} - x_i)^2}{N^2} \right] \qquad (16)$$

and in the case of the sample percentage

$$\sigma_p^2 = \sum_{i=1}^{s} W_i \frac{\sigma_i^2}{N_i} = \sum_{i=1}^{s} \left[W_i^2 \frac{p_1 q_1}{N_i} \right]$$

In a biproportionate sample

$$\sigma_{\bar{x}_1} - \sigma_{\bar{x}_2} = \sqrt{\sum_{i=1}^{s} \frac{W_{1i}^2 \sigma_i^2}{N_i} + \sum_{i=1}^{s} \frac{W_{2i}^2 \sigma_{2i}^2}{N_2}} \qquad (17)$$

where W_{1i} and W_{2i} are the proportions of the population in each of the strata and σ_{1i}^2 and σ_{2i} are the variances in the various strata. The expression within the square root sign is the sum of the different sample variance corresponding to the unrestricted sample formula.

For instance, if the bivariate sample formula given above has not been stratified, the relationship would be as for an unrestricted sample viz:-

$$\sigma_{\bar{x}_1} - \sigma_{\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$$

9. Cost Considerations and Optimum Allocation

The sample formulae given previously assume that a sample member from any stratum costs the same, but it frequently happens that the cost of selecting sample members varies between strata.

Much less expense is incurred in selecting members from, say, large-sized cities than from rural areas; in such cases optimum allocation can be obtained by taking differential sampling costs into account. The expressions used in such cases are just modifications of equation (14); if $c_1, c_2, c_3, \ldots c_n$ denote the cost of selecting members from the first, second, third, etc. strata respectively, the number drawn from each stratum should be such as to satisfy the following expression:-

$$\frac{N_1}{P_1\sigma_1/\sqrt{c_1}} = \frac{N_2}{P_2\sigma_2/\sqrt{c_2}} = \frac{N_3}{P_3\sigma_3/\sqrt{c_3}} = \frac{N_5}{P_5\sigma_5\ \sqrt{c_5}} \qquad (18)$$

Use of tables of Random numbers

A simple and rapid method of choosing a sample from a series of data is to number the data serially then select those particular numbers which are given in published tables of random numbers.

One such publication is the 'Statistical Tables for Biological, Agricultural, and Medical Research' by Fisher and Yates published by Oliver and Boyd Ltd.

10. Statistical Quality Control

Use of Sampling Methods for Control Purposes

Sampling methods are very frequently used to control a characteristic of a process. Such a characteristic may often be the physical dimensions of a machined component, or the hardness of a grinding wheel, etc.

The main principle of the control is to fix the required mean values and to determine the ranges, positive or negative which can arise in the particular process due to random fluctuations. Sampling then continues at appropriate intervals and if the values of such samples lie inbetween the two determined ranges, no action is necessary. When, however the values of the samples exceed the ranges set, the existence of an assigned cause is indicated, and the process stopped to re-adjust the process so that subsequent sample values lie between the two ranges. This method of sampling for control purposes when applied to dimensions in production is also known as statistical quality control and will be described.

The spread of values for the control limits for sample average is $\frac{1}{\sqrt{N}}$ times the production characteristic between the stability limit $X \pm 3\sigma$ where N is the sample size. Fig. 10(a) shows the method of control for a component

$3/8'' \begin{smallmatrix} +0.00 \\ -0.20 \end{smallmatrix}$ the drawing limit being represented by thick lines, and the control limits are shown by dotted lines.

From samples taken the average of the mean values $\bar{\bar{x}}$ is taken to be the average of the population, $\bar{x}_1$ and the average of the observed ranges (i.e. the difference between the highest and lowest values in the samples) is expressed as $\bar{w}$. The control limits are set from these averages in the from $\bar{\bar{x}} \pm Aw$, where A is a constant depending on the size of the sample.

Fig. 10(a) shows a side view of the normal distribution and the relationship between the control limits and the drawing limits.

Fig. 10(b) shows an actual control chart of a component where dimensions are $3/8'' \begin{smallmatrix} +0.00 \\ -0.20 \end{smallmatrix}$ and the averages of 4 samples in 1 hour intervals are shown as points on the chart. The chart indicates that at the 30th sample average the values commenced to rise beyond the upper control limit but a re-adjustment of the process restored the values well within the control limits. Charts are often prepared for the separate control of the deviations from the average, and the deviations of the ranges. This use of sampling control is quite widely used in industry, and in some cases has actually replaced 100% inspection of the products.

11. Sequential Analysis

Most of the statistical material so far discussed has been concerned with the total data available from which conclusions are drawn. It can be compared with a football match or a tennis tournament where the winner is decided after the whole game has been played.

In spite of one team being far ahead at the beginning the winner is not decided until all the data are available at the end.

However, there is another type of test in statistics for specific cases in which each item of data is considered as received and judgment is made on the receipt of each new observation. As soon as a significant result is obtained, no more data are needed and no more observations are taken.

This method is termed sequential analysis, and it is based on the establishment of a hypothesis about a population from which samples are taken. A statistic is calculated from the sample and on this statistic the hypothesis is either accepted or rejected. On commencement of the data being accumulated each new observation is checked against the hypothesis which gives rise to one of three possibilities; the hypothesis is rejected with a preset chance α of error, the hypothesis is accepted with a preset chance β of error, or no decision is made and another observation is made.

The advantage of this method is that usually an average of less than a half to a third as much data as are required for similar results by other methods. In most engineering work problems do not readily lend themselves to sequential analysis and the method is more applicable to inspection sampling.

The procedure is to plot some cumulative function of the data as against the number of observations as abscissae as shown in Fig. 11. Two parallel lines are drawn depending on the particular test being conducted and the probability levels preset.

The two parallel lines divide the area into three sections, one for rejection of the hypothesis, one for acceptance of the hypothesis, and the portion between the lines for no decision and continued testing. When the plot of the cumulative function of the data crosses either of the parallel limit lines, the indicated decision is made. As long as the plot stays within the parallel lines, no decision is indicated without a greater probability of error than was originally preset.

It is not essential to actually plot the data as the values of the two ordinates corresponding to the limit lines can be calculated for each consecutive onservation and these values can be tabulated together with the actual cumulated value from the data. When the observed data exceeds one or becomes less than the other, the indicated decision is made.

As an example of a single test-binomial distribution, in a system where a variable can have one of two possible values (a sample is either good or bad) a student either passes or fails the probability of an event occurring can be denoted by p and the alternative of the event not occurring is 1 - p.

It is necessary to establish the value of p for the population being sampled within some tolerable limits and with an acceptable chance of error. Suppose in a simple sampling problem it is wished to accept material with 1% defectives or less and that a 5% limit to the chance of rejecting material that meets this standard. Suppose also it is desired to reject material that has 5% or more defectives and a 1% limit to the chance of accepting material with this many defectives. The one hypothesis is that $p \leq 0.01$ and set an α value of 0.05 for a false rejection of this hypothesis. The alternative hypothesis is that $p < .05$ and a value of β of 0.01 for a false acceptance of this hypothesis. In general an acceptable limit to $p \leq p_1$ with an α maximum risk of false rejection is set and an unacceptable limit $p \geq p_2$ with a β maximum of false acceptance.

For the sequential test take one item at a time and plot the number of defectives as ordinate against the number of items inspected as abscissae. Two control limit lines are drawn on the inspection chart: the upper one:-

$$L_u = l_u + mN \qquad (19)$$

and the lower one

$$L_L = l_L + nN \tag{20}$$

where $$l_u = \frac{\log_{10} \frac{1-\beta}{\alpha}}{\log_{10} \left[\frac{p_2(1-p_1)}{p_1(1-p_2)}\right]}$$

$$lL = \frac{-\log_{10} \frac{1-\alpha}{\beta}}{\log_{10} \left[\frac{p_2(1-p_1)}{p_1(1-p_2)}\right]}$$

$$m = \frac{\log_{10} \frac{(1-p_1)}{(1-p_2)}}{\log_{10} \left[\frac{p_2(1-p_1)}{p_1(1-p_2)}\right]}$$

N = number of observations

As an example of the use of this method, suppose that in an inspection programme the desired quality is 2% defective and a maximum of 5% rejecting material that has 2% or less defectives.

The poorest quality that can be tolerated is 4% defective and a maximum of 2% chance of shipping material with 4% or more defectives.

$p_1 = .02$ $\qquad\qquad$ $\alpha = .05$

$p_2 = .04$ $\qquad\qquad$ $\beta = .02$

$$l_u = \frac{\log_{10} \frac{1-\beta}{\alpha}}{\log_{10} \frac{p_2}{p_1} \frac{(1-p_1)}{1-p_2}} = \frac{\log_{10} \frac{.98}{.05}}{\log_{10} 2 \times \frac{.98}{.96}} = \frac{\log_{10} 19.6}{\log \quad 2.04} = \frac{1.2923}{.3096} = 4.18$$

$$l_L = \frac{-\log_{10} \frac{1-\alpha}{\beta}}{\log_{10} \frac{p_2}{p_1} \frac{(1-p_1)}{(1-p_2}} = \frac{-\log_{10} \frac{.95}{.02}}{\log_{10} 2 \times \frac{.98}{.96}} = \frac{-\log_{10} 47.5}{\log_{10} \quad 2.04} = \frac{-1.6767}{.2096} = -5.42$$

$$m = \frac{\log_{10} \frac{1-p_1}{1-p_2}}{\log_{10} \frac{p_2}{p_1} \frac{(1-p_1)}{(1-p_2)}} = \frac{\log_{10} \frac{.98}{.96}}{\log 2 \times \frac{.98}{.96}} = \frac{\log 1.02}{\log_{10} 2.04} = \frac{.0086}{.3096} = .028$$

The two limit lines corresponding to this example are shown in Fig. 12.

List of illustrations

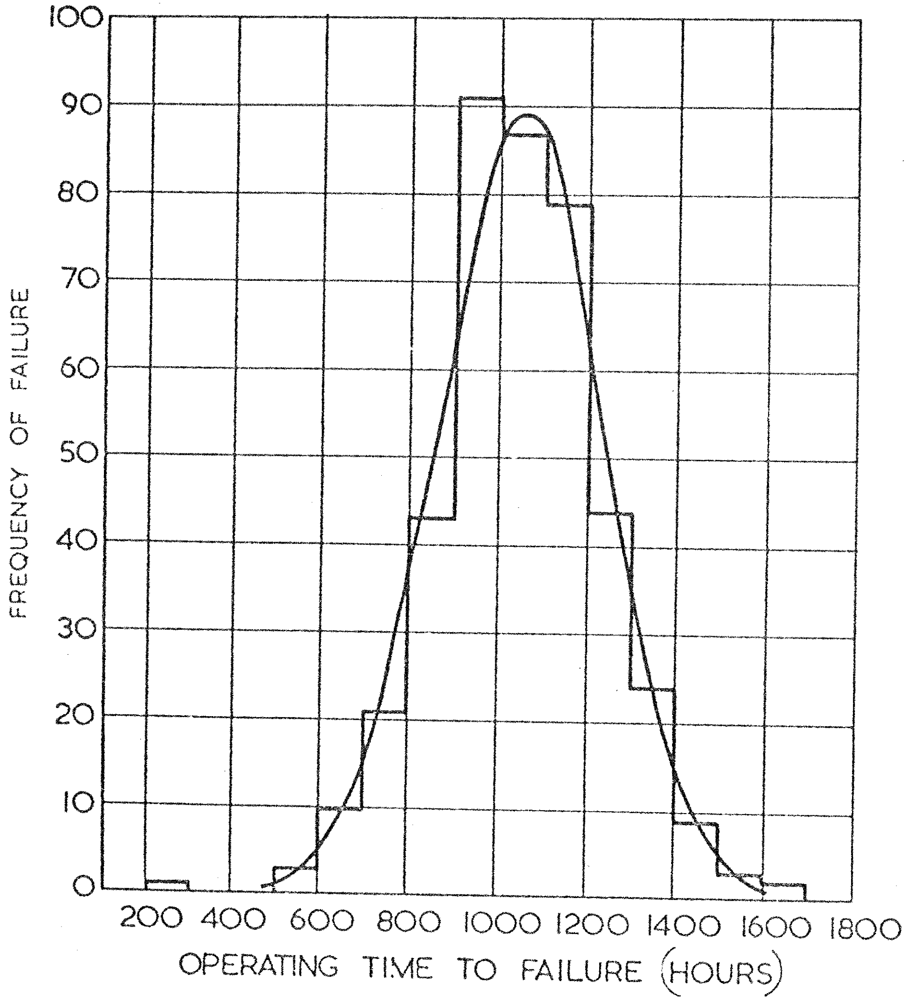


FIG.1.

NO. OF AUSTRALIAN MARRIAGES, ARRANGED ACCORDING TO THE AGE OF BRIDEGROOM IN 3 YEAR GROUPS.

AGE OF BRIDEGROOM	NO. OF MARRIAGES	AGE OF BRIDEGROOM	
16·5	294	55·5	1.655
19.5	10.995	58·5	1.100
22·5	61.001	61·5	810
25·5	73.054	64·5	649
28·5	56.501	67·5	487
31·5	33·478	70·5	326
34·5	20.569	73·5	211
37·5	14.581	76·5	119
40·5	9.320	79·5	73
43·5	6.236	82·5	27
46·5	4.770	85·5	14
49·5	3.620	88.5	5
52·5	2.190		
			301.785

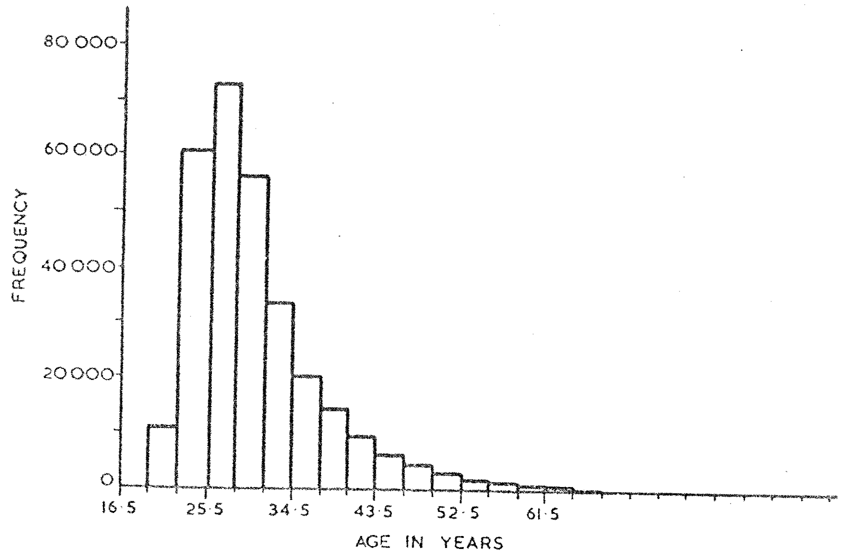


FIG. 2.

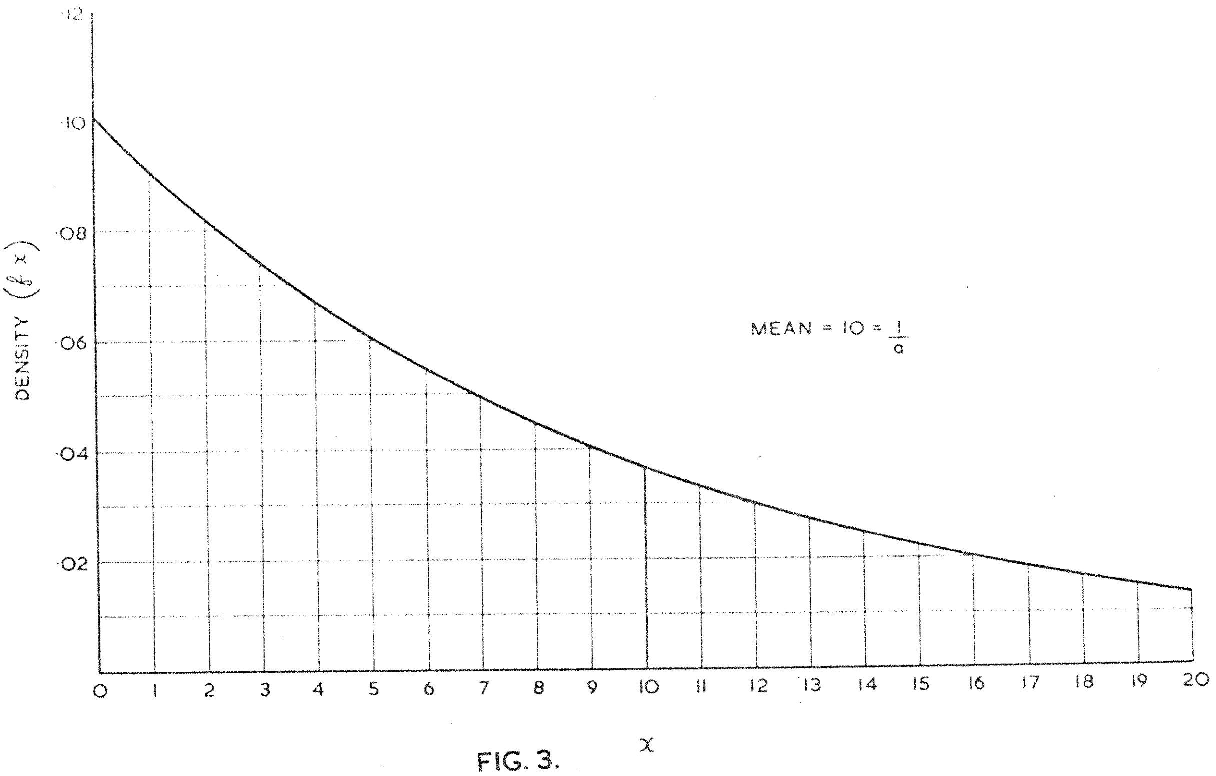


FIG. 3.

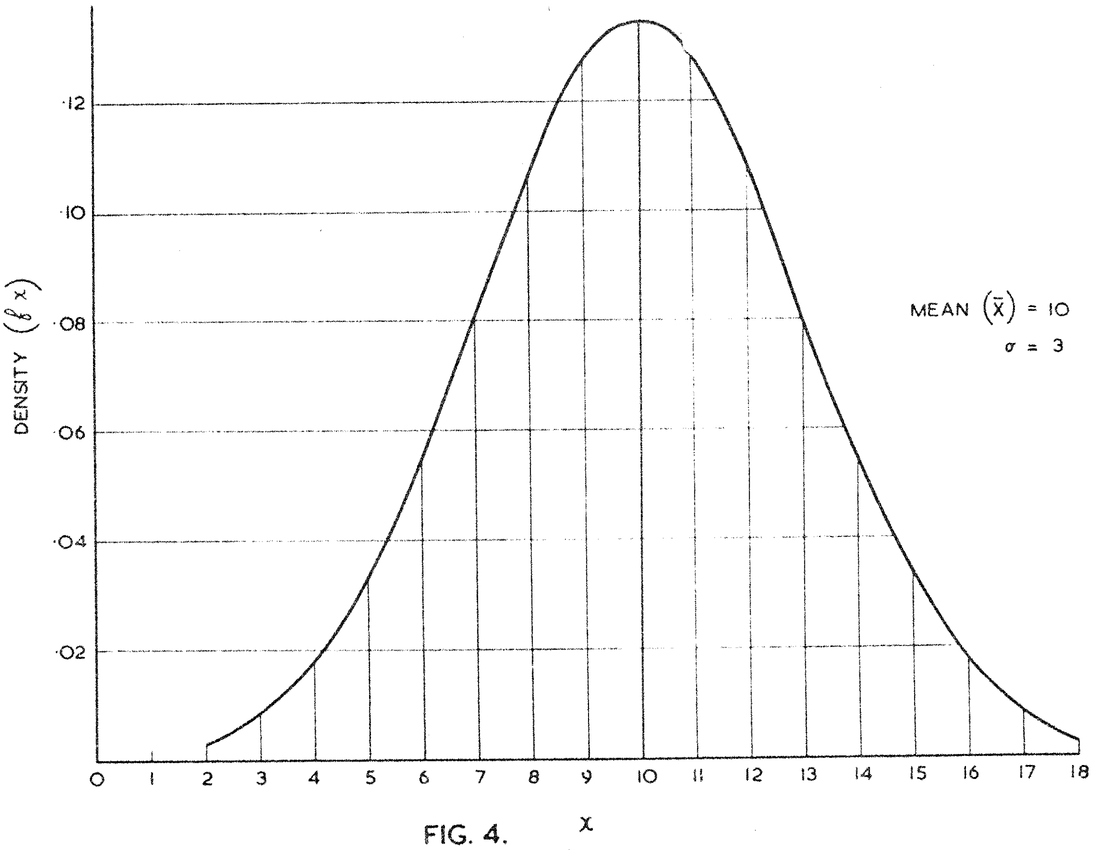


FIG. 4.

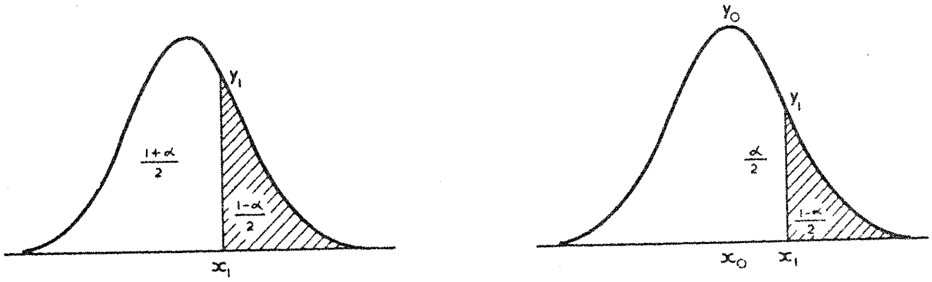


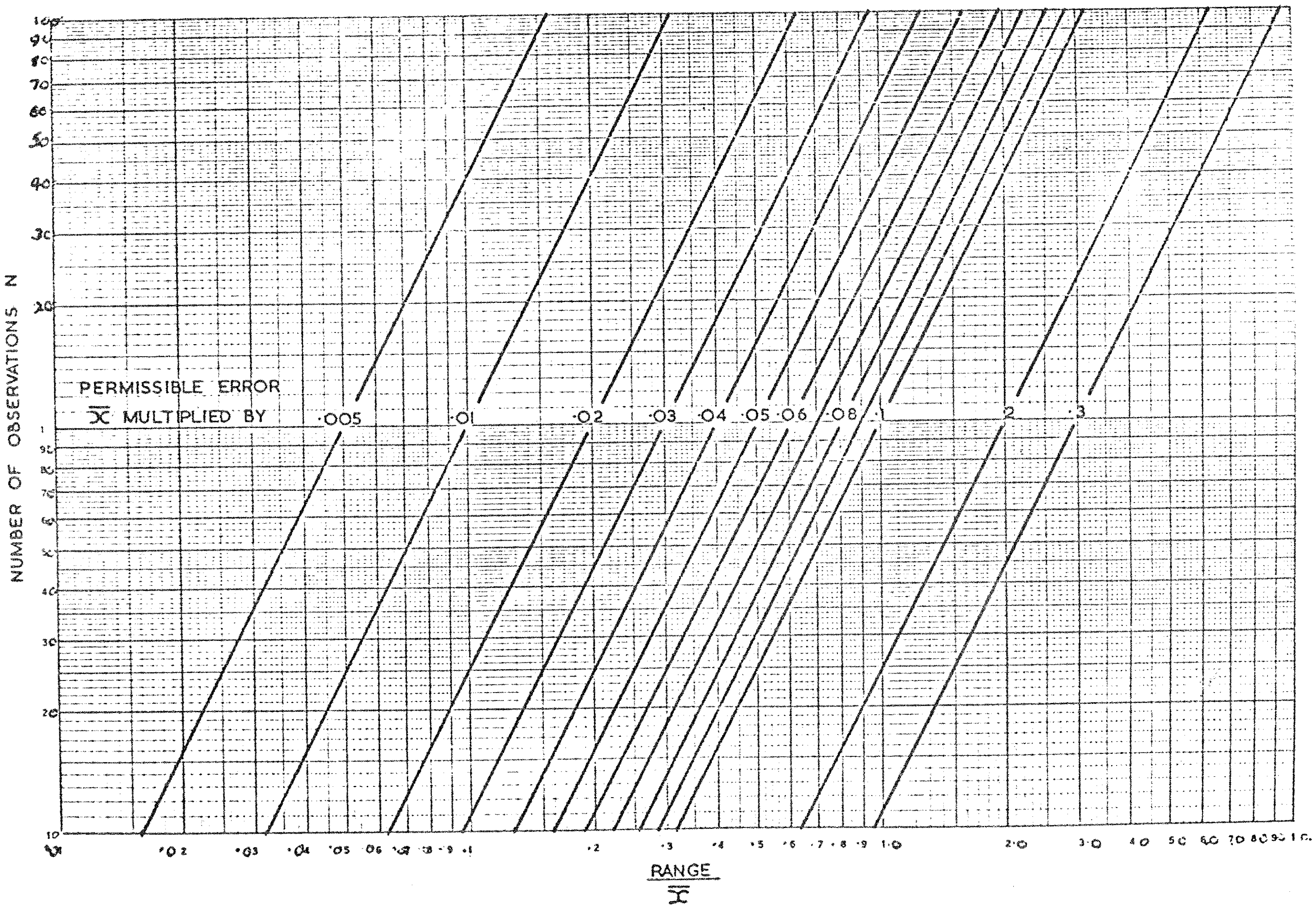
FIG. 5. PROBABILITY INTEGRALS

Table of Probability Integrals

$\frac{x}{\sigma}$	$\frac{1-\alpha}{2}$	$\frac{1+\alpha}{2}$	$\frac{1-\alpha}{1+\alpha}$	$\frac{1-\alpha}{\alpha}$	y
0.0	.5	.50	1.0	0	.3989
.1	.4602	.5398	.8524	11.5540	.3970
.2	.4207	.5793	.7263	5.3083	.3910
.3	.3821	.6179	.6184	3.2405	.3814
.4	.3446	.6554	.5257	2.2171	.3683
.5	.3085	.6915	.4462	1.6114	.3521
.6	.2743	.7257	.3779	1.2149	.3332
.7	.2420	.7580	.3192	.9377	.3122
.8	.2119	.7881	.2688	.7352	.2897
.9	.1841	.8159	.2256	.5826	.2661
1.0	.1587	.8413	.1886	.4648	.2420
1.1	.1357	.8643	.1570	.3724	.2179
1.2	.1151	.8849	.1300	.2989	.1942
1.3	.0968	.9032	.1072	.2401	.1714
1.4	.0808	.9192	.0879	.1926	.1497
1.5	.0668	.9332	.0716	.1542	.1295
1.6	.0548	.9452	.0580	.1231	.1109
1.7	.0446	.9554	.0466	.0979	.0940
1.8	.0359	.9641	.0373	.0774	.0790
1.9	.0287	.9713	.0296	.0609	.0656
2.0	.0228	.9772	.0233	.0477	.0540
2.1	.0179	.9821	.0182	.0371	.0440
2.2	.0139	.9961	.0141	.0286	.0355
2.3	.0107	.9893	.0108	.0219	.0285
2.4	.0082	.9918	.0083	.0167	.0224
2.5	.0062	.9938	.0062	.0126	.0175
2.6	.0047	.9953	.0047	.0094	.0136
2.7	.0035	.9965	.0035	.0070	.0104
2.8	.0026	.9974	.0026	.0051	.0079
2.9	.0019	.9981	.0019	.0038	.0060
3.0	.0013	.9987	.0014	.0027	.0044
3.1	.0010	.9990	.0010	.0019	.0033
3.2	.0007	.9993	.0007	.0014	.0024
3.3	.0005	.9995	.0005	.0010	.0017
3.4	.0003	.9997	.0003	.0007	.0012
3.5	.0002	.9998	.0002	.0005	.0009

Fig. 6

Reproduced by kind permission of the publishers University of London Press Ltd. from the 'Introduction to the Computation of Statistics' by Shepherd Dawson.



PERMISSIBLE ERROR
$\overline{x}$ MULTIPLIED BY
·005
·01
·02
·03
·04
·05
·06
·08
·1
·2
·3
NUMBER OF OBSERVATIONS N
RANGE / $\overline{x}$

FIG. 7.

DISTRIBUTION OF t

n	.9	.8	.7	.6	.5	.4	.3	.2	.1	.05	.02	.01	.001
							Probability						
1	.158	.325	.510	.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	.142	.289	.445	.617	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	.137	.277	.424	.584	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841	12.941
4	.134	.271	.414	.569	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	.132	.267	.408	.559	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032	6.859
6	.131	.265	.404	.553	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	.130	.263	.402	.549	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499	5.405
8	.130	.262	.399	.546	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	.129	.261	.398	.543	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	.129	.260	.397	.542	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	.129	.260	.396	.540	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	.128	.259	.395	.539	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	.128	.259	.394	.538	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	.128	.258	.393	.537	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	.128	.258	.393	.536	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	.128	.258	.392	.535	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	.128	.257	.392	.534	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	.127	.257	.392	.534	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	.127	.257	.391	.533	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	.127	.257	.391	.533	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	.127	.257	.391	.532	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	.127	.256	.390	.532	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	.127	.256	.390	.532	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	.127	.256	.390	.531	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	.127	.256	.390	.531	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	.127	.256	.390	.531	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	.127	.256	.389	.531	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	.127	.256	.389	.530	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	.127	.256	.389	.530	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	.127	.256	.389	.530	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	.126	.255	.388	.529	.681	.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	.126	.254	.387	.527	.679	.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	.126	.254	.386	.526	.677	.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	.126	.253	.385	.524	.674	.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

FIG. 8.

DISTRIBUTION OF χ^2

Probability

n	.99	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01	.001
1	$.0^3157$	$.0^3628$	.00393	.0158	.0642	.148	.455	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2	.0201	.0404	.103	.211	.446	.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210	13.815
3	.115	.185	.352	.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.345	16.268
4	.297	.429	.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277	18.465
5	.554	.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086	20.517
6	.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812	22.457
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475	24.322
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090	26.125
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666	27.877
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209	29.588
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725	31.264
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217	32.909
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688	34.528
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141	36.123
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578	37.697
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000	39.252
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409	40.790
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805	42.312
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191	43.820
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566	45.315
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932	46.797
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289	48.268
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638	49.728
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980	51.179
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314	52.620
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642	54.052
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963	55.476
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278	56.893
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588	58.302
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892	59.703

For larger values of n, the expression $\sqrt{2\chi^2} - \sqrt{2n-1}$ may be used as a normal deviate with unit variance, remembering that the probability for χ^2 corresponds with that of a single tail of the normal curve.

FIG. 9.

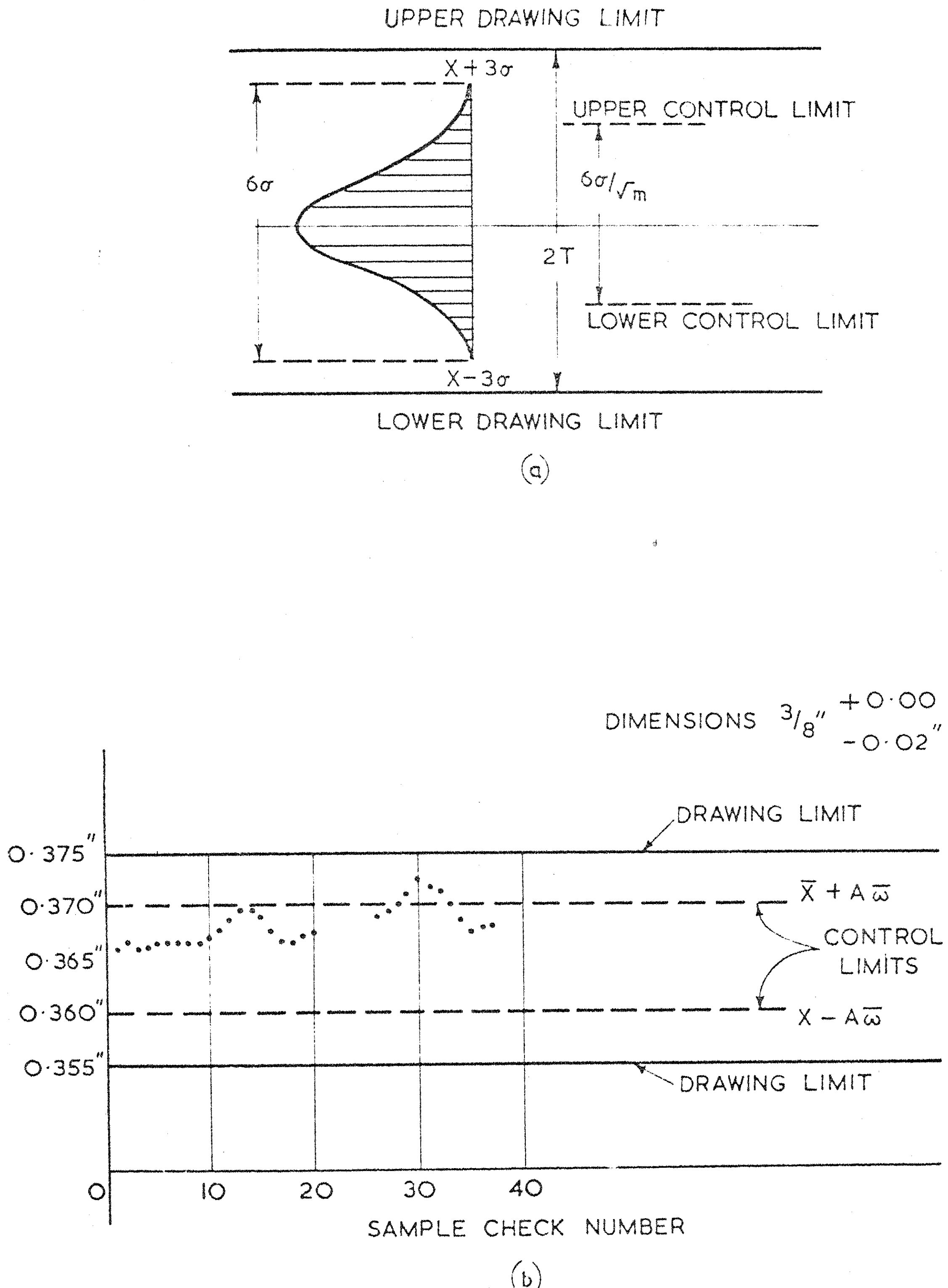

UPPER DRAWING LIMIT
X + 3σ
UPPER CONTROL LIMIT
6σ
6σ/√m
2T
LOWER CONTROL LIMIT
X − 3σ
LOWER DRAWING LIMIT
(a)
DIMENSIONS 3/8" +0·00 −0·02"
DRAWING LIMIT
0·375"
0·370"
0·365"
0·360"
0·355"
X̄ + A ω̄
CONTROL LIMITS
X − A ω̄
DRAWING LIMIT
0
10
20
30
40
SAMPLE CHECK NUMBER
(b)

FIG. 10.

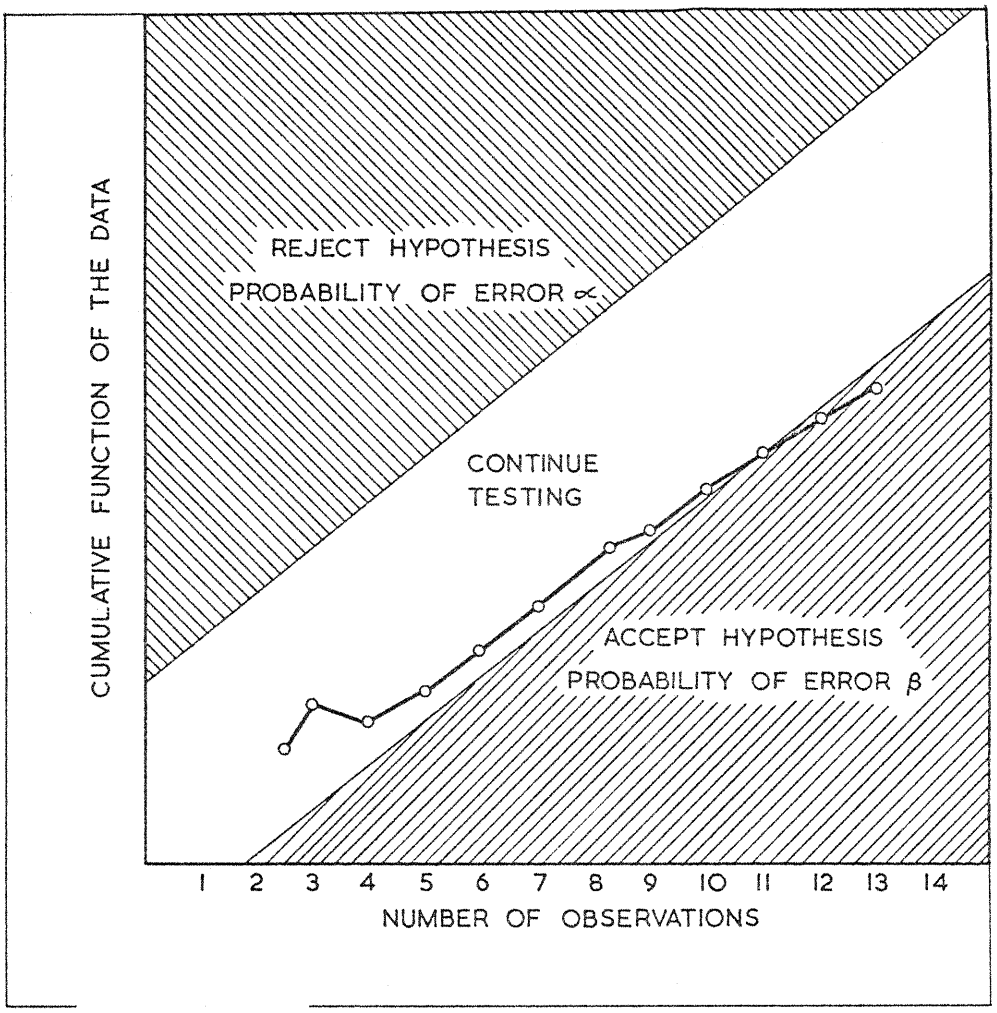

REJECT HYPOTHESIS
PROBABILITY OF ERROR ∝
CONTINUE
TESTING
ACCEPT HYPOTHESIS
PROBABILITY OF ERROR β
CUMULATIVE FUNCTION OF THE DATA
1 2 3 4 5 6 7 8 9 10 11 12 13 14
NUMBER OF OBSERVATIONS

FIG. 11.

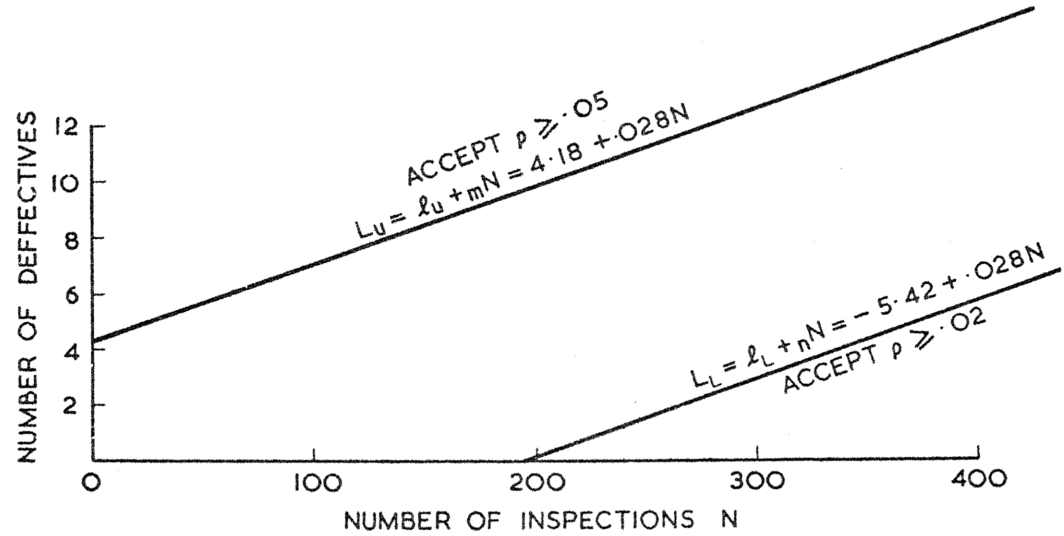

ACCEPT p ≥ ·05
Lu = ℓu + mN = 4·18 + ·028N
LL = ℓL + nN = − 5·42 + ·028N
ACCEPT p ≥ ·02
NUMBER OF DEFFECTIVES
2 4 6 8 10 12
0 100 200 300 400
NUMBER OF INSPECTIONS N

FIG. 12.