GRAPHICAL STATISTICAL METHODS

By W. R. Hinton, A.M.I.E.E., M.I.R.E.

SUMMARY. Two things are attempted in this paper. First, to provide the beginner with powerful statistical tools and techniques for examining data graphically, and secondly, to direct the attention of those already using Probability Paper to the quality of fit required before an assumption of normality is justified. Graphical methods are described which require no theoretical knowledge of statistics at all, and the natural development of a conception of a 'Normal Population' and the evolution of Probability Paper, is described in most simple terms; however, sufficient information is given to prevent errors due to incorrect plotting, which have been seen in some previously published works. The examples have been chosen deliberately to be diverse, to emphasize the wide range of problems to which statistical methods can be applied, and the main purpose of the article is to encourage interest in this fascinating and most useful subject.

Introduction

BROADLY speaking, we are often concerned with examining the properties of similar things, and with forming an idea of what could be classified as a typical object. A typical object could be defined as one which is like the majority of objects in a fairly large sample. No less important is some idea of the way individuals differ from the typical one, and what proportion of the total have reasonably close properties to the typical one.

Generally these ideas are formed intuitively, and comprise the 'experience' of the observer, and quite often such ideas are difficult to define and communicate to anyone else.

Statistics is a subject which is primarily concerned with classifying, grouping and examining data, and offers a language by which the above ideas can be conveyed. For example, the most probable value of an observed quantity (i.e., the typical value) is called the 'Mode,' and a constant which is descriptive of the way in which individuals cluster round the typical value is called the 'Standard Deviation.' Likewise there are recognized graphical methods of presenting and analysing data, such as the Histogram and Ogive, which have been designed to convey particular ideas and allow certain conclusions to be drawn with a minimum of labour.

Ideas have been developed of 'populations' of individuals and of the distribution of individuals in the population, of samples of individuals from a population and of the distribution of individuals in the sample. It is not surprising to find, therefore, that an ideal parent population has been conceived, ideal, that is, in the amount of information which can be inferred from the way individuals are distributed in the population, and in the simplicity with which it can be described and defined. This ideal distribution is known as the 'Normal Distribution' or the 'Gaussian Distribution,' and is found to approximate closely to many distributions of observable quantities in nature.

The main point of testing data to see whether individuals follow a Normal Distribution, is to know whether the extensive predictions, proper to the ideal population, can be applied with any confidence to the problem in hand. This simply means that if the data is normally distributed, much labour is saved because the various predictions have been tabulated and published by theoretical workers in the field. If the data is not normal, one has to make one's own calculations and predictions.

The correct attitude is to decide what information is required from the data, and whether a graphical solution is adequate; if not, whether one has sufficient data to justify a normality test or a curve fitting analysis. Often one finds that there are so few individuals in the sample at one's disposal, that a curve fitting analysis would be rather absurd, and obviously, if this is so, any labour-saving system of curve fitting (like probability paper) is equally unreliable. For this reason probability paper is a snare for the unwary; as the temptation to use it for very small samples from unknown populations is very great. Of course, it is true that, if the data follows a normal distribution, the graph on the probability paper will be a straight line, but it may not be generally realized how little is the deviation from linearity which can be tolerated in a practical case for an assumption of normality to be justified, as will be shown later.

The novelty and fascination of probability paper tends to eclipse the usefulness of the common Ogive (from which it is derived), and an imperfect understanding of the evolution of probability paper can lead to errors due to incorrect plotting, or to a wrong interpretation of the curve. Some attempt is made, therefore, in the following notes, to provide a background of simple, graphical, statistical methods, and to demonstrate some inherent limitations of probability paper.

The Histogram

In surveying a mass of numerical results one would intuitively group identical results together, and perhaps arrange the resulting groups in ascending order of magnitude. A logical extension of this idea would be to group results which fell between definite boundaries, and to arrange these groups in ascending order of magnitude. In this way the vast amount of detail would be made more comprehensible, and any significant difference between one group and another would become apparent. For example, suppose that a Government Department was required to requisition suits for demobilized armed forces, and it was essential to conserve raw materials and labour. The problem would be to discover how many sizes of suits would be required, and how many of each should be manufactured.

The first step might be to take a sample of men at random, and measure their individual heights to the nearest inch, as shown in columns (1) and (2) in Table I. (Note, therefore, that the height of a man recorded as 58 in may actually be anywhere between 57½ in and 58½ in.)

TABLE I

(1) Height to the nearest Inch	(2) Number of Men	(3) Number of Men in Group	(4) % of Total Men in Group	(5) Cumulative % Men
58	1			
	4	5	0.8	
60	3			0.8
61	5			
62	13	52	8.4	
63	32			9.2
64	52			
65	74	214	34.6	
66	88			43.8
67	92			
68	80	256	41.3	
69	73			85.1
70	47			
71	25	84	13.6	
72	12			98.7
73	5			
74	2	8	1.3	
75	1			100
Totals	619	619	100	

Suppose that the data is grouped into broader classes, for example, into men with heights between 57½ in and 60½ in and between 60½ in and 63½ in, and so on, as shown in column (3). This can be plotted as shown in Fig. 1 (a), where each pillar represents a group of data by its height, being proportional to the number of men within the group, and the edges of the pillar define the group boundaries. Such a figure is known as a Histogram, and is valuable for showing the 'Mode' or most frequent value, which in this case is about 68 in, and the dispersion, or scatter, either side of the Mode. In a word, it shows the distribution of individuals in the sample. [In some problems it may be more convenient to plot the height of Histogram pillars proportional to the percentage of total observations, as given in Table I, column (4) and shown in Fig. 3.]

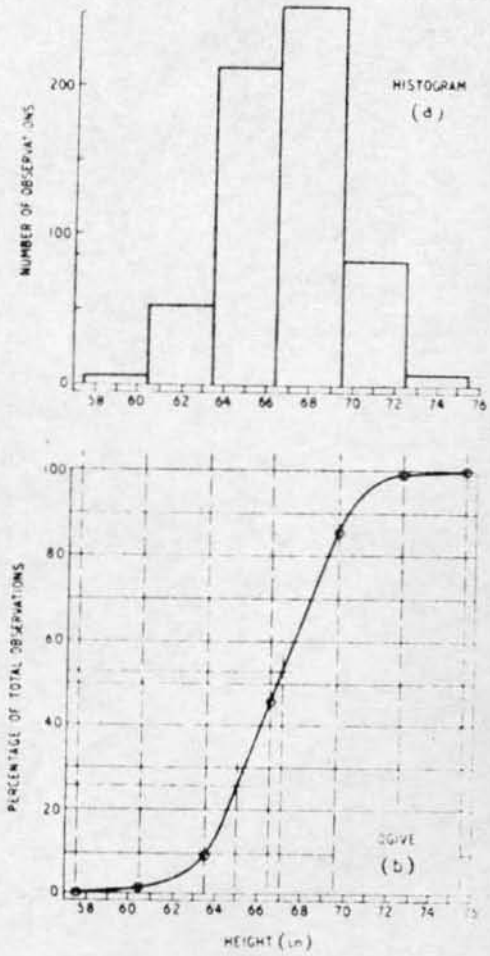

Fig. 1. *Histogram and Ogive showing the distribution of heights of 619 men, and the method of determining the proportion of men with heights between, say 65 in and 67 in.*

The Ogive

In practice, the group boundaries of the Histogram are chosen arbitrarily so as to make a presentable figure, and whereas it is possible to

see what proportion of men have heights between, say, 66½ in and 69½ in, it is not easy to see what proportion would have heights between 65 in and 67 in (say), or what proportion exceed 70 in in height.

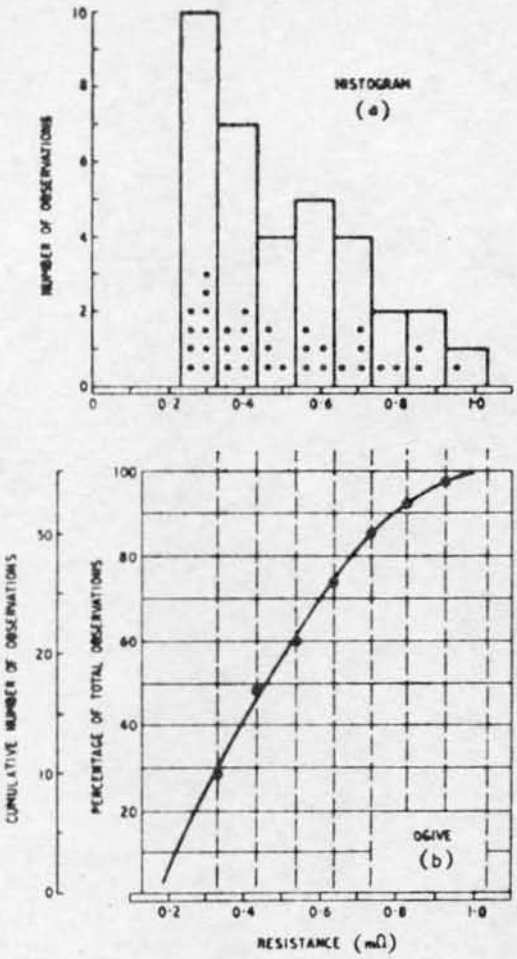

Fig. 2. *Dot Diagram, Histogram, and Ogive of the measured resistance of a heavy-current switch under full-load conditions, showing extrapolation of the extremes.*

To solve this kind of problem it is better to compute the cumulative percentage of men, as shown in Table 1, column (5). For example, 0.8% of the men had heights between 57½ in and 60½ in and 9.2% between 57½ in and 63½ in, and so on. Notice that no man is recorded with height less than 57½ in, and that 0.8% of the men were encountered over the range 57½ in to 60½ in, therefore the cumulative percentages must be plotted at the edges of the class intervals, as shown in Fig. 1 (b). This curve is an "Ogive," and provides different information from the Histogram. For example, the proportion of men having heights less than 67 in is 52% and that less than 65 in is 26%, therefore the proportion of men between heights 65 in and 67 in is the difference:

$$52 - 26 = 26\%.$$

This means that if one standard design of suit is intended to fit men between 65 in and 67 in in height, the number to be requisitioned should be 26% of the total, and so on. Likewise it seems hardly worth while making suits in bulk for statures less than about 60 in or over 73 in and these men could be fitted individually.

One valuable property of the Ogive is that it is fairly insensitive to the choice of class intervals, (histogram groups) and therefore smooths out the data which was coarsened by grouping for the histogram.

This example has shown what valuable information can be obtained from the application of very elementary statistical tools and a little common sense, and there has been no need to test the data for normality, or confuse oneself with predictions based on the Normal Distribution.

The Ogive is particularly useful for analysing data whose typical value (Mode) tends to be close to an extreme value. For example, in inspecting electrical switches, the switch resistance can never be less than the resistance of the conductors forming the switch parts, but always more, depending upon the condition of the contact surfaces. In such a case the Histogram is unsymmetrical or "skew" (Fig. 2 (a)) and the Ogive shows the minimum resistance quite clearly, which is of course, a useful parameter for judging the quality of the design.

If a technique is employed so that the extreme points of zero and 100% are not plotted, the Ogive gives an extrapolated value for the minimum resistance which is insensitive to the choice of histogram class intervals, (pillar widths).

This is a reasonable technique to adopt, because otherwise the arbitrary choice of class interval would arbitrarily fix the point of zero cumulative observations at the edge of the first pillar; whereas it is much better to let the entire data weight the choice of this point by extrapolation. Similar remarks apply to the 100% point. It is very instructive to experiment with grouping data differently, and observe how slightly the Ogive is affected.

Graphical Grouping of Data

It saves a great deal of time and labour to record each observation directly on the graph paper to be used for the Histogram, by making a bold dot opposite the appropriate point on the

bscissa, thus forming rudimentary columns of ots as observations are repeated. This method as the great advantage that one can see when ufficient data has been collected, as the distribution of dots begins to suggest a definite orm. It is then easy to mark the dots off nto suitable groups and construct a Histogram ased on the number of dots falling into each roup, as shown in Fig. 2. (Where a dot falls n a group boundary, one half of an observation hould count in each group.) This is called Dot Diagram, and the intermediate step of rawing the Histogram is not necessary for onstructing an Ogive, as the total number of lots encountered up to a given boundary can e seen directly. Likewise, by observing quanities to the nearest convenient unit, the dots orm clean columns and have the appearance f histogram pillars, which therefore, need not e drawn.

Another advantage of collecting experimental lata by dot diagrams, is that it shows up any rift in performance due to warming up of quipment, etc. This becomes apparent when ne finds the columns of dots not being filled n a random manner, but that ones hand is gradually moving across the page, as time goes on. This is most noticeable in the resistance measurements mentioned earlier, and the solution is to make several independent experiments, noting the time from the instant of switching on at which the observation is made. (Obviously the switch must be given time to cool down between experiments and strictly speaking, the observations should be made at about the same rate.) The data is recorded as bold dots, or other marks, located at the appropriate time, as shown in Fig. 3, and constitutes a Scatter Diagram which is intended to show whether one variable seems to depend on another. Clearly, in Fig. 3, the value of switch resistance observed, depended, to a great extent on the relative time at which the observation was made.

The method of treating the data is to divide the diagram into suitable vertical strips so that the dots enclosed can be considered to have occurred at roughly the same relative time. Then each strip of data can be treated as a normal dot diagram to form Histograms or Ogives as required. Fig. 3 thus gives a very clear picture of the performance of the switch. For example, about five minutes after switching on, the odds would be even that the switch resistance would not be greater than 0.32 mΩ because the Ogive shows that out of a large number of observations, one can expect 50% to lie below this value and 50% above. Likewise, the chance that the resistance would exceed 0.47 mΩ is about one in twenty, because the Ogive shows that about 95% of a large number of observations could be expected to be below this value.

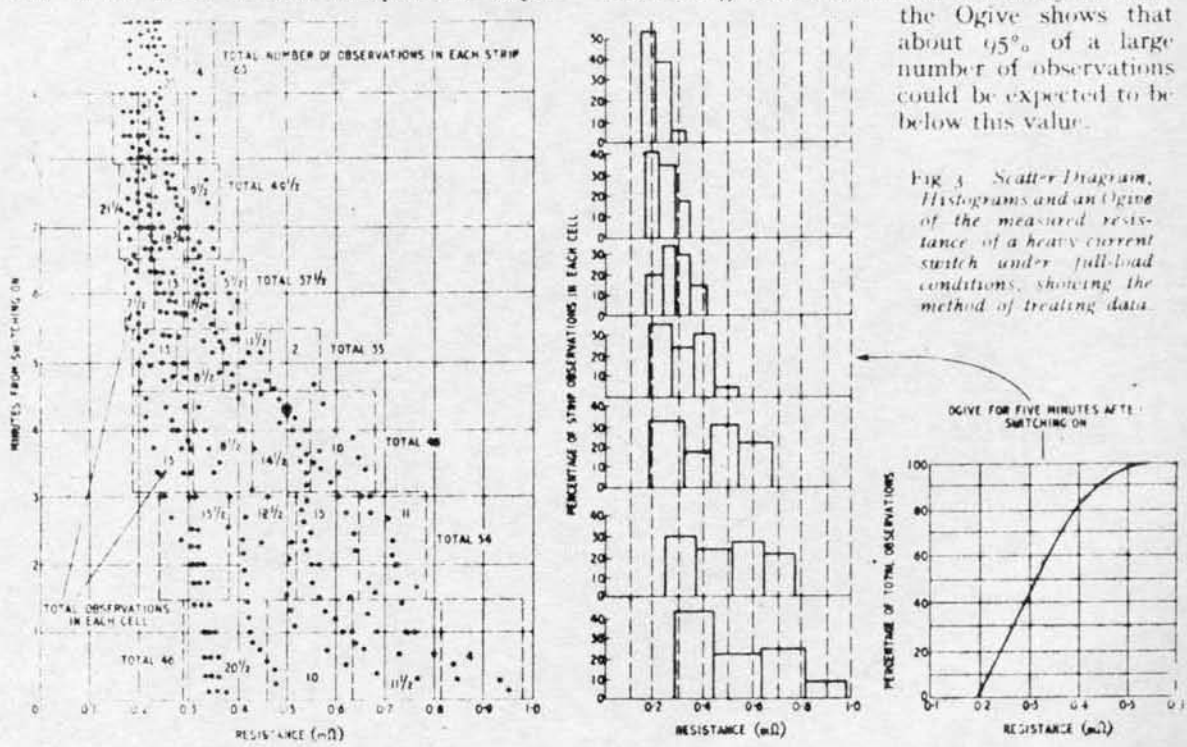

Fig. 3. *Scatter Diagram, Histograms and an Ogive of the measured resistance of a heavy current switch under full-load conditions, showing the method of treating data.*

Before leaving the subject of Dot and Scatter Diagrams, there is one disadvantage that should be mentioned, namely that because one can see the diagram taking shape, one is inclined to cheat, or shall we say, be biased in one's judgment, and cast the dots into where one thinks they should go. A great deal of self-discipline has to be exercised.

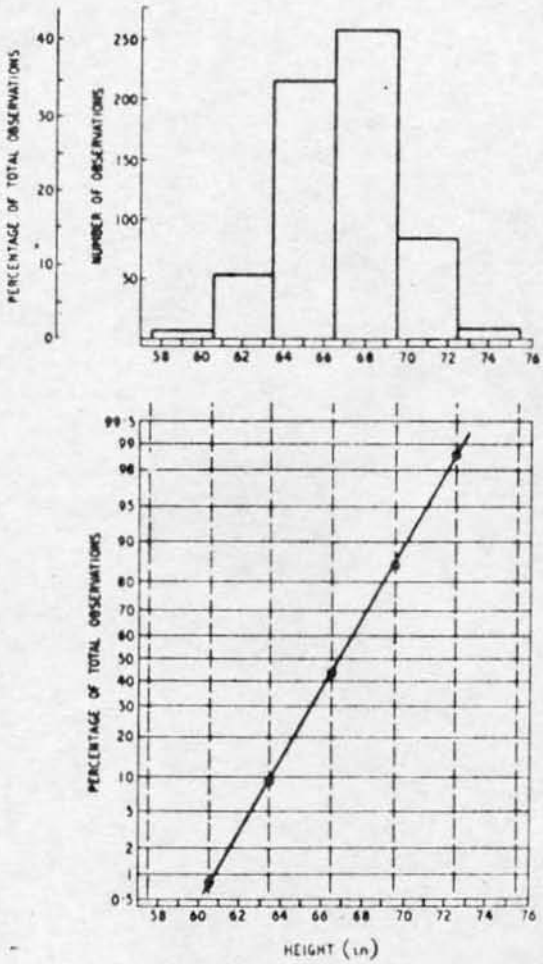

Fig. 4. Histogram with alternative scales and the Probability Paper Ogive of the distribution of heights of 619 men (as for Fig. 1), showing the correct method of plotting points at the edges of the class intervals.

Arithmetical Probability Paper

Probability Paper is simply a special graph paper for constructing Ogives, in which the vertical scale for 'Percentage of Total Observations' is stretched, at the extremities in a particular way. This scale is called the Probability Scale because the Ogive represents a population of a large number of individuals from which one can deduce the odds of an observation falling above or below a particular value. Let it be emphasized at once that it is not essential to use special probability paper to do this, indeed we have seen in the previous section that the common Ogive can be used.

The prefix, 'Arithmetical' means that the scale for the variable is linear, while with Logarithmic Probability Paper, the scale for the variable is logarithmic; the probability scale being the same as for the arithmetical paper. Obviously, one could use a logarithmic scale for the variable, in the construction of the common Ogive, if such a scale offered a clearer picture of the data.

Most Ogives are S-shaped curves, and become difficult to read at the extremes. If the scale is expanded over these sections, the readability of the graph is increased, but this does not mean that the reliability of the data is increased. In exactly the same way, inspecting the position of an ammeter pointer with a powerful magnifying lens does not increase the precision of a current measurement, because the calibration may not be reliable. This means that when Ogives are plotted on probability paper, one must be extremely cautious about making use of extrapolated information at the extremes of probability scales, say, outside the range of 5% to 95%.

At this stage it is convenient to emphasize the correct method of plotting data on probability paper. Since the curve is really an Ogive, the points corresponding to cumulative grouped data must be plotted at the boundaries of the Histogram groups, as shown in Fig. 4, and not at the central values of the groups which, by the way, is a common error. Since there is no zero or 100% on the probability scale, data for these points must necessarily be omitted.

If the plotted points fall exactly on a straight line, the data is normally distributed, because the probability scale has been specially stretched to make this so. Naturally one does not expect experimental points to fit a straight line perfectly, but the limited extent of deviation allowable may not be appreciated, and one should have at least twenty points to plot before drawing any serious conclusions about normality.

This is clearly demonstrated in Fig. 5, which shows data plotted on probability paper, from rectangular and triangular populations. Most people would feel justified in drawing a straight line through the points given by either of the 8 cell Histograms and might, therefore, be led erroneously to believe that both populations were normally distributed.

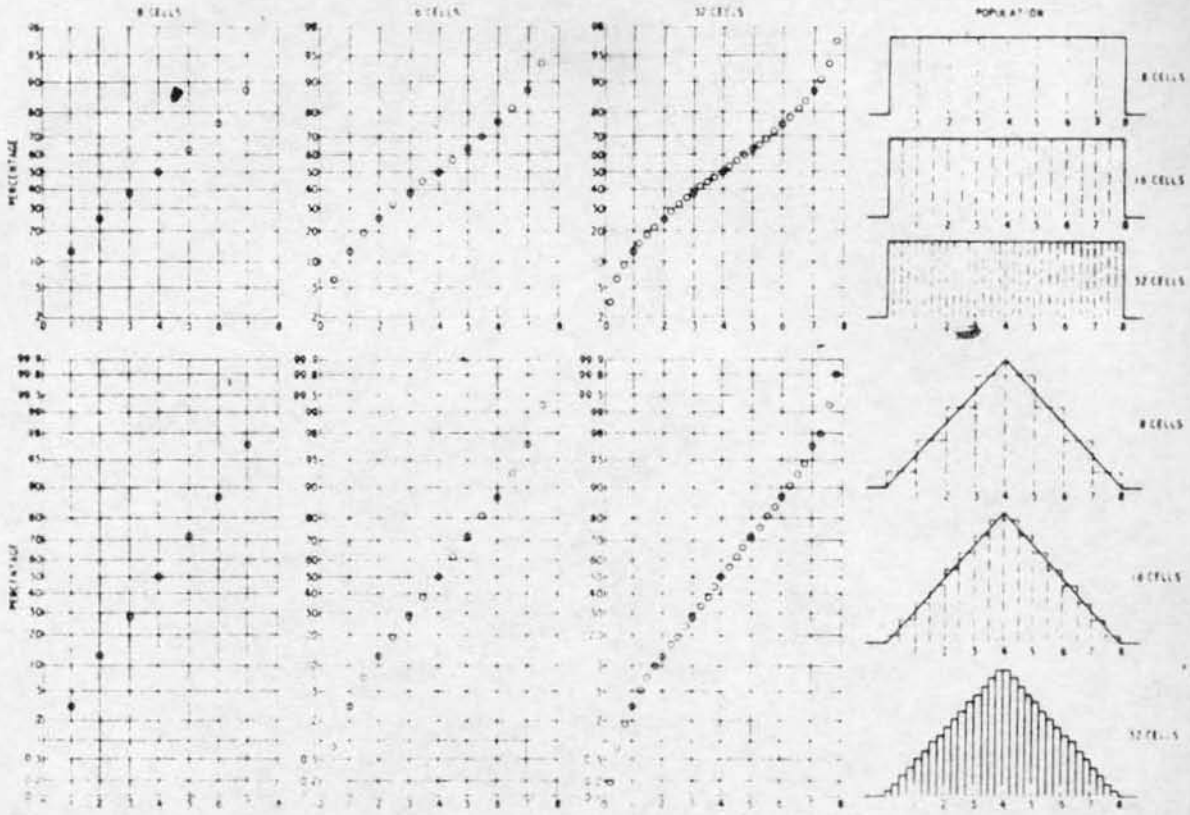

Fig. 5. *Ogives, plotted on Probability Paper, for 8-, 16- and 32-cell Histograms of data arising from rectangular and triangular parent populations, showing the small departure from linearity when few points are available*

The data for 16 Cells begins to show the curving of the extremities, which is clearly confirmed with 32 Cells, so that in practice it is advisable to have at least twenty points when testing for normality. Even then, the recommended attitude to be adopted, when an approximately straight line is obtained, is that such data *could* arise from a normal population, and not that it *does*.

The examples shown may be considered rather wide deviations from a normal population, and indicate the limited sensitivity of probability paper to differentiate between distributions when relatively few points are available. One feature worth noting is the relatively high probability of the first point for the rectangular distribution compared with the more normal triangular one; for example, 12.5% for the rectangular, against 3.1% for the triangular. This reveals that the former points are coming from a rectangular type of distribution because quite a high proportion of the results are encountered for quite a small invasion into the edge of the data.

Ungrouped Data

The popularity of probability paper is due, in no small measure, to the ease with which random individual observations can be handled, although the common Ogive can be used in exactly the same way. (This knowledge may save a lot of time and labour when supplies of the special graph paper are not available.)

Let us consider the case of our first example, the heights of men to be fitted with suits. The Histogram of Fig. 1 has analysed the results of 619 individual observations, and the question is whether a sample of say ten individuals could provide the same information. Obviously not: but a fair idea of the distribution can be obtained if certain assumptions are justified.

The first assumption to be made is that each observation has the same weight. In other words that each of the ten observations represents about the same number of men in the parent population of 619. This means that each individual observation in the sample is assumed to represent 61.9 men or 10% of the total men in the parent population.

The second assumption is that each sample individual is typical of the 10% of the parent population it is assumed to represent; that is, that the majority of the 61·9 heights of men can be considered to be clustered round the sample value.

A third assumption (that, from previous experience or other considerations, the data can be expected to be normally distributed) is valuable, but not essential, as it means that using probability paper, the best straight line can be drawn through the points, and so allows fairly small samples to be used.

The second assumption is the key to the method of plotting, because it means that if the sample individuals are arranged in ascending order of magnitude, each in turn corresponds to the mean value of successive groups of population. In other words, each sample individual is assumed to fall near the centre of the group and therefore in the example chosen, has 5% of the population below it to one boundary of its group, and 5% of the population above it to the other boundary of the group. In approaching the first sample individual then, 5% of the total parent population would be encountered, and in reaching the second, 15% would have been passed, made up of the first group (10%) plus half of the next group (5%), and so on. Thus for ten individuals in the sample, they should be plotted on the Ogive, or on probability paper, at the following percentages of total observations:

5, 15, 25, 35, 45, 55, 65, 75, 85, 95%.

In general, if there are n observations in a sample, they should be spaced at $100/n$% and the first observation should occur at $100/2n$%. Example: Suppose the heights of ten men, taken at random, were as shown in Table II, and that it was known that a normal distribution could be expected. The proportion of men with heights between 65 in and 67 in is required.

TABLE II

	Inches									
Heights of ten men taken at random, and placed in ascending order	63	64	65	66	66	67	68	69	70	71
Plot on Probability Scale at the following percentages	5	15	25	35	45	55	65	75	85	95

The results are shown plotted on probability paper in Fig. 6, and from the best straight line, we have:

Proportion of men having heights less than 67 in = 52%

Proportion of men having heights less than 65 in = 25%

Proportion of men having heights between 67 in and 65 in = 27%

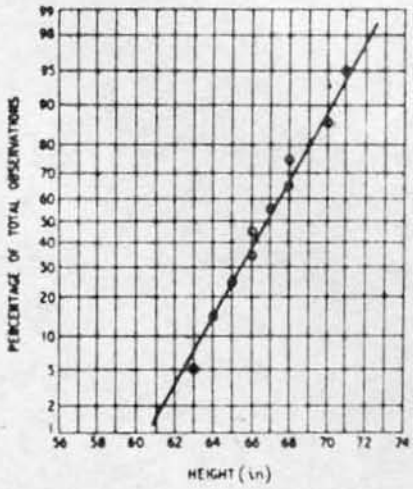

Fig. 6. Ogive, plotted on Probability Paper, of a small sample of ten individual observations taken at random from the population of 6119 heights of men shown in Figs. 1 and 4

The Normal Distribution

So much has been said about testing for normality, and whether or not individuals can be expected to follow a normal distribution, that a short discussion of its properties is appropriate at this stage.

As mentioned previously, the Normal Distribution is a conception of an ideal population of individuals, from the point of view of mathematical analysis. It is based on the assumption that deviations of individuals, from the mean value, obey three laws:

(1) The probability of small deviations from the mean value is greater than the probability of large deviations.

(2) The probability of a certain deviation above the mean value is equal to the probability of an equal deviation below the mean value.

(3) The probability of a "huge" deviation is very small indeed.

If we imagine a "Normal" Histogram, the first law means that the central pillars tend to peak near the mean; the second law means that the histogram is symmetrical about the mean; and the third, that the columns vanish fairly quickly as the distance from the mean increases. If we

now imagine the width of the histogram pillars to be made extremely small, and consequently the number of pillars extremely large, the tops of the pillars would follow a bell-shaped curve, as shown in Fig. 7. The mathematical law of this curve has been deduced from the postulations above, and a knowledge of this allows a great deal to be predicted about the probability of an individual falling between specified boundaries; i.e., into any specified Histogram group.

One of the most important deductions is that two parameters are sufficient to define and describe the distribution; and both parameters can be calculated without drawing the distribution, or conversely, obtained directly from the Histogram or Ogive, without calculation. These parameters are the 'Mean' and the 'Standard Deviation.'

Because the distribution is symmetrical, the Mean corresponds to the peak of the bell-shaped curve. The Standard Deviation is taken to be the distance from the Mean to the point of inflection of the curve, shown as sigma in Fig. 7, and both these points are easy to see on the Histogram.

As far as the Ogive is concerned, it has been calculated that, for a Normal Distribution, 34.2% of the total population can be expected between the mean and the standard deviation and, by symmetry, 50% of the population exists each side of the mean. This means that the standard deviation can be found by taking the difference in abscissa corresponding to ordinates of 50% and 15.8% (34.2% from the mean), or ordinates 50% and 84.2% as shown in Fig. 7.

The Standard Deviation is used as a yardstick for specifying the distribution of individuals in a normal population, and tables are published giving the proportions of population encountered between the mean and various deviations expressed as fractions or multiples of a standard deviation. Table III is a simplified version of one.

TABLE III

Deviation from Mean	Percentage of Total Population		
	Between Mean and Deviation (%)	Between ± Deviation (%)	Outside range of ± Deviation (%)
$\frac{1}{2}\sigma$ (approx.)	25	50	50
σ	34.2	68.3	31.7
2σ	47.7	95.4	4.6
3σ	49.85	99.7	0.3

For example, one can see that the chance of an individual observation falling outside the range $\pm 2\sigma$, is approximately one in twenty (4.6%). Likewise, the odds are about even that an observation would be within the range $\pm \frac{2}{3}\sigma$. (One should be very confident that the data is normally distributed, before venturing to predict for deviations greater than 3σ.)

Perhaps the most important of all deductions made by theoretical workers, concerning Normal Distributions, is that whatever the distribution of individuals in a parent population (rectangular, triangular, double humped, or any other shape), the distribution of the means of random samples drawn from that population, tends to be Normal, provided, of course, the parent population remains stable.

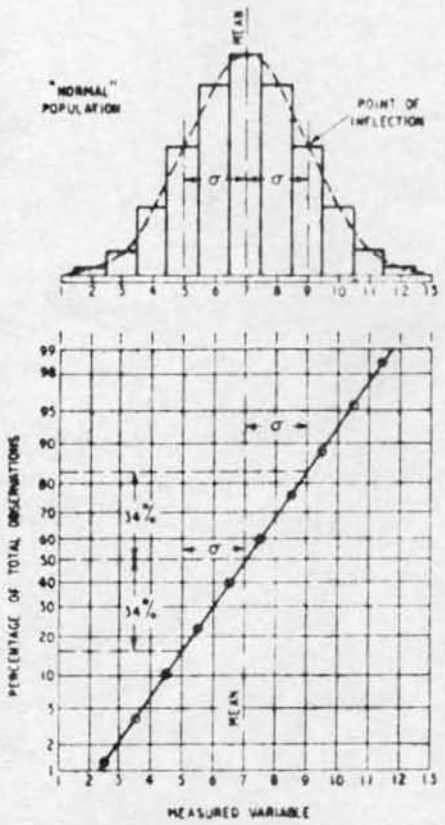

Fig. 7. *Histogram, and Probability Paper Ogive, of a 'Normal Distribution,' showing the method of recognizing the parameters 'Mean' and 'Standard Deviation'.*

This last comment is the key to one important application of Probability Paper, namely to test the stability of populations. Successive batches of twenty or more means of samples, should show roughly the same mean and standard deviation if the population is stable. This is the basis of Quality Control, where instability

in the parent population of objects means that rejects will occur unless the drift is arrested. When a sample is obtained with a mean value deviating from the grand mean by a certain amount, one is justified in using Normal Distribution Tables to compute the probability that such a deviation could be expected by pure chance, and act accordingly. Notice, however, that it is not absolutely essential to understand Normal Distributions to operate a Quality Control system; one could conduct a 100% inspection at the beginning of production, when the process was operating satisfactorily, and construct an Ogive from the results. This could then be used to define the limits within which subsequent sample individuals should fall.

Remember that if one intends to use the means of samples of five individuals, the Ogive must be constructed from the means of groups of five individuals. In short, decide on the size of sample, and commence with 100% inspection. This simply means that the samples are successive groups of individuals. Calculate the mean of each sample and plot as a dot diagram. When the dot diagram takes on a definite shape, plot the Ogive. Decide now how many false alarms can be tolerated, say one in ten, and select limit values from the Ogive accordingly. In the example chosen, one is prepared to be given a false alarm once in ten times, in other words 10% of the total population will be allowed to fall outside the Ogive limit lines so that about once in ten times an observed mean will exceed the limits by pure chance. This means that the limits must be chosen to allow 5% of the population to occur below the lower limit, and 5% above the upper limit. Thus draw lines on the Ogive at 5% and 95% and read off the corresponding limits of the variable.

Take samples at reasonable intervals, and investigate as soon as any sample mean exceeds either of the limits. (If the sample means are plotted as a scatter diagram in time, and the limit lines are marked on the scatter diagram, it is easy to see when there is a drift in the process which will eventually cause rejects.)

This elementary system of Quality Control is the basis of the more refined methods, which are designed to economize in labour by taking advantage of other theoretical deductions and properties associated with Normal Distributions.

Conclusions

The author hopes that the reader will agree that the first conclusion is that a great deal of useful work can be done graphically, with Histograms, Ogives and Scatter Diagrams, without any theoretical knowledge of Normal Distributions, and the like. This is important because it encourages the use of efficient statistical methods among those who might be put off by a more theoretical approach, and provides a good background for the appreciation of the more advanced ideas, later on. The danger of a limited theoretical knowledge lies in the temptation to use apparently simple tools, which have been developed from very advanced theory, because then they are imperfectly understood, and erroneous conclusions may be drawn from the results.

Probability Paper is directly concerned with Normal Distributions and this implies that some knowledge of the theory of these distributions is desirable for its correct use. Generally the use of it will be justified only when some application of the properties of normal distributions is sought, for example, when used as a labour-saving device to avoid calculations, and economize in size of sample, when the data is known to be normally distributed. Or again, when testing the stability of populations, making use of the expected normal distribution of sample means.

The actual testing of data for normality should be regarded cautiously, and not seriously attempted with less than twenty points, preferably more. One should also have a clear idea whether a test for normality is justified. For example, in the problem of providing suits for men of various heights, it is absolute nonsense to make elaborate tests and extensive predictions concerning the normality of the heights, in order to avoid waste of materials, when the selection is ultimately influenced by personal preferences in colours, designs and cuts. An Ogive analysis would be adequate to get the proportions approximately right.

Likewise, as the author knows to his cost, it is advisable to check the data for stability of parent population before attempting any serious curve fitting.

The final conclusion drawn, is that Probability Paper is rather an insensitive tool for discriminating between distributions, and a large number of points are required before pronouncing judgment on the probable Parent population.

Acknowledgments

The author wishes to express his thanks to the Chief Scientist of the Ministry of Supply for permission to publish this article, and to those friends and colleagues who first introduced the writer to this most fascinating and useful subject.