MAY I INTRODUCE STATISTICS?

BY

Major G. FRASER ANDERSON, M.B., Ch.B., D.P.H.

Royal Army Medical Corps

PART I

INTRODUCTION

THE purpose of this paper is to explain the terms commonly used in statistical analysis and to indicate the application of the simpler statistical methods.

Many impolite remarks have been passed, are being passed, and will, no doubt, continue to be passed regarding statistics and statisticians. Some of them are true and a few of them are trite. The expression most frequently used is probably to the effect that "figures can be made to prove anything."

Figures can prove many things; for example, if you are endeavouring to complete a journey at the average speed of 30 miles per hour and you accomplish the first half of it with an average of 15 m.p.h., nothing less than a magic carpet will be necessary for the rest of the journey. That may appear incredible at first sight; that it is true is due to the fact that you have used for the first half of the trip the time you had allowed for the whole journey. I can remember, also, when two members of a cricket club were running neck and neck to top the bowling averages for the season. With one match to go both had the same average; the fast bowler had already taken 100 wickets for 1,000 runs and the slow bowler 10 for 100; in the last match the fast bowler took five for five and the slow bowler five for twenty. Consternation is the mildest term for the emotion which was experienced by the former when he discovered that his labours had resulted in his having a final average of 9·57 whereas his rival ended up with an average of 8. The injustice is only apparent, however, as, although both had improved their averages, the slow bowler had done so to a greater extent relatively to the number of wickets and runs he had already taken and conceded. Had the last day's performances been spread over the whole season, not even an apparent injustice would have been suggested. Batting and bowling averages are not always what they seem! What I like to believe is that the expression "figures can be made to prove anything" originated at about the same time as "the exception proves the rule"; at the time, in fact, when the word "prove" had not lost its relationship with the original "probe." I think we

can agree that the exception probes the rule and that figures can probe anything.

Those who sheer away from the subject of statistical methods (those, in fact, who make the rude remarks) are usually frightened either by the mathematics involved, the technical expressions encountered, or by the Greek symbols employed. In practice the mathematics is not difficult, the expressions used are mostly self-explanatory, and the symbolism easily understood where it cannot be entirely eliminated. The faint-hearted, however, may prefer to omit, on first reading it at least, that part of the paper which is printed in small type. This can be done without disturbing the sequence of the remainder.

One fact which must be understood at the outset is that, no matter what figures can be made to do, *statistical methods can prove nothing*. They are merely tools in the hands of the research worker by means of which he is enabled to describe, relate and assess the value of his observations. They can be used and they can be abused. What is of importance is that those who read original work in which figures are used by the writer should be able to judge whether he has in fact used them or abused them.

The term " statistical significance " is not an absolute one. It is based on probability. If there are three horses in a race, each of which is considered by the bookmaker to have an equal chance of winning he will (if he is not interested in profit !) open his book by offering two to one against each of the runners. By two to one he means that he considers a horse to have one chance in three of winning; this may be expressed as a probability of $\frac{1}{3}$; each horse has an equal chance in this case, and $\frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 1$. It is always the case that probabilities expressed in this way do add up to 1 provided that all probabilities are included. If you throw a die, your chance of throwing a 6 is 1/6, and your chance of throwing something else is 5/6; 1/6 and 5/6 again making 1. Your chance of throwing a 5 or a 6 is $1/6 + 1/6 = \frac{1}{3}$; your chance of throwing anything else is 4/6 or 2/3 and again $\frac{1}{3} + 2/3 = 1$. Here has been used the constant rule that the probability of *either* one thing happening *or* another is found by *adding* together the probabilities of each. On the other hand your chance of throwing two sixes with two dice is $1/6 \times 1/6 = 1/36$. In other words the odds are 35 to 1 against. Here the rule is that the probability of one thing happening *as well* as another is found by *multiplying* the individual probabilities. To extend this, the odds against your throwing three sixes with three dice are 215 to 1 ($1/6 \times 1/6 \times 1/6 = 1/216$). In other words the chance is so small that one might consider the throw to be significant. If you performed the feat several times it would be so significant that you might be accused of using loaded dice. Nevertheless proof would be lacking, because the " probability," no matter how small, would still exist. Similarly with what is called statistical significance ; when the result of an experiment is said to be statistically significant all that is meant is that the probability of such a result having been obtained by chance is extremely small ; how small depends upon the rigour of the standard which has been adopted. One of the reasons why statistical signifi-

cance can never be regarded as proof is that one is nearly always working, not with the population as a whole, but with a sample of that population. Obviously the value placed upon the results of an experiment depends largely on the size of the sample and on the extent to which it is representative of the population. If you took ten women at random and found that all the blondes were blue-eyed and all the brunettes brown-eyed you might be interested but not significantly impressed. If you increased your sample tenfold and found the same result you might begin to think there was a definite relationship between fair hair and blue eyes on the one hand and dark hair and brown eyes on the other. No matter how big your sample, however, you could never be *absolutely sure* that the next blonde you came across would have blue eyes.

Averages

The best known and most useful form of average is the *arithmetic mean*, usually called the " mean " or " average," so that when the word " average " is used, the arithmetic mean is usually intended. The mean of a number of observations is, of course, normally found by adding all the observations and dividing the total by the number of observations. When the number of observations is large, however, this would be a laborious operation, and the usual method is simplified by the construction of a *frequency distribution table* (Table I).

Table I.—Frequency Distribution Table Showing the Numbers of Colds Suffered by Each of 814 Individuals in the Course of One Year

No. of colds	*No. of individuals*
0	14
1	173
2	247
3	233
4	112
5	23
6	12
7+	0
Total	814

Instead of adding together 814 separate figures the same result is obtained by multiplying each figure in the first column of the table by its corresponding figure in the second column and adding the results. Thus the total number of colds is :

$(0 \times 14) + (1 \times 173) + (2 \times 247) + (3 \times 233) + (4 \times 112) \times (5 \times 23) + (6 \times 12) = 0 + 173 + 494 + 699 + 448 + 115 + 72 = 2001$, and the mean number of colds suffered by the 814 individuals of course is $2001/814 = 2{\cdot}46$.

In the above example the values of one's observations were anything from 0 to 6. This method would, however, still be laborious if there were many possible values of one's observations. In such a case one's findings are usually

tabulated in the form of a *grouped frequency distribution table* and the mean can be determined by a modification of the method used above (Table II).

TABLE II.—GROUPED FREQUENCY DISTRIBUTION TABLE SHOWING THE SYSTOLIC BLOOD PRESSURE OF 222 HEALTHY ADULTS AND THE METHOD OF CALCULATING THE MEAN

I	*II*	*III*	*IV*
Systolic B.P.	*Frequency*	*Group No.*	*II x III*
95–99	2	−7	−14
100–104	4	−6	−24
105–109	7	−5	−35
110–114	12	−4	−48
115–119	19	−3	−57
120–124	24	−2	−48
125–129	27	−1	−27
			−253
130–134	35	0	
135–139	26	1	26
140–144	21	2	42
145–149	18	3	54
150–154	13	4	52
155–159	6	5	30
160–164	5	6	30
165–169	2	7	14
170–174	1	8	8
	222		256
			−253
			3

The first two columns form an example of a grouped frequency distribution table. To calculate the mean, an arbitrary origin, or starting point, is selected, usually the centre point of one of the groups towards the middle of the table. In the example 132 has been chosen as the starting point (all observations between 129·5 and 134·5 will be included in the 130–134 group). A third column is then added, numbering the 130–134 group (i.e. that including the starting point) zero, the higher groups 1, 2, 3, etc., and the lower groups −1, −2, −3, etc. Column 1 can now be ignored, and a fourth column is formed by multiplying the figures in columns II and III. When the figures in column IV are added, making due allowance for + and − signs, the total, in this case 3, divided by the number of observations, 222, gives the discrepancy in "working units" of the arbitrary origin or starting point from the true mean. The "working unit" is the number of units included in each group, in this case 5. Thus :

$$\text{Discrepancy in working units} = \frac{3}{222} = 0{\cdot}0135$$

$$\text{Size of working unit} = 5$$

$$\text{Arbitrary origin or starting point} = 132$$

$$\text{Therefore, the true arithmetic mean} = 132 + 0{\cdot}0135 \times 5$$

$$= 132 + 0{\cdot}0675$$

$$= 132{\cdot}0675$$

(N.B.—If the sum of column IV had been negative, the discrepancy would have been subtracted from the starting point instead of having being added to it ; e.g. if the total of column IV had been − 9, the other values remaining the same, then :

$$\text{Discrepancy in working units} = -\frac{9}{222} = -0{\cdot}0405$$

$$\text{Size of working units} = 5$$

$$\text{Arbitrary origin or starting point} = 132$$

$$\text{Therefore, the true arithmetic mean} = 132 - 0{\cdot}0405 \times 5$$
$$= 132 - 0{\cdot}2025$$
$$= 131{\cdot}79751$$

It will be realized that the above method is completely accurate only if the mean of all the readings in each group is coincident with the central point of the group. This is an acceptable assumption which becomes more and more nearly true as the number of observations increases.

Another form of average which is sometimes used is called the *median*. This, as its name implies, is the central observation when all the observations are arranged in order of magnitude. Thus, if there are twenty-one men on parade and they are " sized " in one rank in order of height, then the height of the middle man, i.e. the eleventh from either end, will be the median height of all those on parade. The essential point about the median is that there are an equal number of observations on each side of it. If there are an even number of observations then the median is taken as the mean of the two central observations. Thus, if there had been twenty-two men on parade, the eleventh and twelfth men would have been equally central and the mean of their two heights would have given the median. It will be seen that there would still have been the same number of observations, viz. 11, on either side of the median.

To find the median in a grouped frequency distribution requires but a modicum of common sense and the acceptance of the assumption already suggested, that the observations in any one group are evenly distributed throughout that group. Reverting to Table II, it is obvious that the median observation of the 222 is the mean of the one hundred and eleventh and the one hundred and twelfth observations. (There would then be 111 observations on each side.) There are 95 observations in the first seven groups and 35 in the eighth group (130–134). It is therefore obvious that the hundred and eleventh and the hundred and twelfth observations must lie in that group. The question is where ? The answer is that the two required observations will be the sixteenth and seventeenth in that group (111–95 = 16, and 112–95 = 17). If we may assume that the observations in that group are evenly distributed then the required observations will be 129·5 + 16/35 of 5 and 129·5 + 17/35 of 5. (In other words, the lower limit of the group plus 16/35 and 17/35 of the size of the group, there being 35 observations in the group.) Accordingly the median is the mean of 129·5 + 80/35 and 129·5 + 85/35, i.e. the mean of 131·79 and 131·91, i.e. 131·85.

A form of average which is occasionally used is the *mode*. This, as its name suggests, is the most fashionable or most frequently occurring, observation. In Table I, for example, the mode is 2, as more individuals suffered two colds than any other number. The mode, however, can rarely be determined accurately by the study of numerical data as errors in sampling or the small size of the sample may easily result in an observation occurring most frequently which differs considerably from that of the population as a whole.

If a graph of the frequency of any variable for the total population were drawn, the highest point of the curve would represent the mode, since the maximum frequency would occur at that point. It has been found that, in curves

which are not greatly asymmetrical, it is better not to rely on the most frequently occurring observation but to determine the mode by applying the following formula :

$$\text{Mode} = \text{mean} - 3\,(\text{mean} - \text{median})$$

In distributions which are perfectly symmetrical the mean and median are equal ; the mode will then equal the mean $- 3 \times 0$; in other words the mean, median and mode will all be equal.

The median and mode are less valuable as averages than the mean and are much less commonly used. Their most obvious use is where the mean would give a false picture by reason of there being one or two exceptional outlying observations in a series which is otherwise fairly uniform.

E.g. the inhabitants of a village might consist of twenty cottagers earning £300 each year, and a squire with an income of £3,000. To say that the average income of the inhabitants was in the neighbourhood of £430 would obviously be a distortion of the true state of affairs (although the mean in fact is over £428 10s.). On the other hand, the median and mode in this case would both be £300, a much more accurate figure.

The Measurement of Scatter or Dispersion

Averages, which we have been considering so far, may be described as measurements of position. They fix the position around which the variables occur. They rarely, however, give a sufficiently adequate description of those variables. Suppose that the twenty-one soldiers whom we had on parade in the last section had a mean height of 5 ft. 8 in. ; they might nevertheless vary from 5 ft. 3 in. to 6 ft. 3 in. ; or, going towards the other extreme, they might all be between 5 ft. 7 in. and 5 ft. 9 in. To complete the picture, we must have some measures which will describe the *scatter*, as it is commonly called, of one's observations round the mean.

The range is the simplest measure of scatter or dispersion. Thus in the parade above-mentioned one could say the mean height of the soldiers was 5 ft. 8 in. with a range of 5 ft. 3 in. to 6 ft. 3 in., or, on the other hand, 5 ft. 8 in. with a range of 5 ft. 7 in. to 5 ft. 9 in. The range, however, is not a good measure of dispersion as it depends on only two of the observations, and these may easily mislead, one or both of them, for example, being exceptionally divorced from the rest. (Thus, in the parade above-mentioned the range might be 5 ft. 3 in. to 6 ft. 3 in. and yet all but the two extreme observations be within the range of 5 ft. 7 in. and 5 ft. 9 in.)

The *inter-quartile range*, as its name indicates, involves the division of the observations into quarters. This is accomplished by taking first the median of all observations, thus dividing them into two numerically equal halves, and then taking the median of those two resulting halves. The three resulting " medians " are termed " quartiles " and the inter-quartile range is the range between the upper and lower quartiles.

In its simplest form it can be applied to seven observations. If these are

arranged in order of magnitude, the median observation is, of course, the fourth. (There would then be three observations on each side.) Of each of these groups of three observations the median is the second, leaving one on each side, and the inter-quartile range the range between those two final medians, the upper and lower quartiles. Thus, one might apply the inter-quartile range to a series of seven observations of height in a sample, say, of children who are being investigated in a nutritional experiment. If their heights in inches were 42, 46, 47, 49, 50, 52 and 57, then the median observation would be 49, and the upper and lower quartiles 46 and 52; the inter-quartile range would be 6 inches. (N.B. that this expression of scatter avoids the exaggeration which would otherwise be suggested by the lowest and highest observations being far removed from their neighbour.)

In practice the inter-quartile range is usually halved, and quoted as the *semi-inter-quartile range*. In the last example this would, of course, be 3 inches. Although this measure is an improvement on the total range it is not amenable to further statistical treatment and is still liable to be anomalous.

A measure which uses all the observations is the *mean deviation*. This, as its name implies, is the mean amount by which all the observations deviate from their average. The median rather than the arithmetic mean is usually taken as the average for this purpose. In the last example the observations in inches were 42, 46, 47, 49, 50, 52 and 57. Of these the median observation is 49. The deviations of each observation from 49 are 7, 3, 2, 0, 1, 3 and 8. The mean deviation is, then, the sum of these deviations, 24, divided by their number, 7. The mean deviation is, accordingly, $24/7 = 3{\cdot}43$.

The Standard Deviation.—This is the most useful of all measures of scatter and is, in words, the square root of the mean of all squared deviations of observations from their mean. The symbol usually employed to represent the standard deviation is the Greek letter σ (sigma). The abbreviation S.D. may be used instead. In our last example the observations were 42, 46, 47, 49, 50, 52 and 57.

$$
\begin{aligned}
\text{Mean} &= \text{sum of observations divided by their number} \\
&= (42 + 46 + 47 + 49 + 50 + 52 + 57)/7 \\
&= 343/7 \\
&= 49 \\
\text{Deviations from the mean} &= -7, -3, -2, 0 + 1 + 3 \text{ and } + 8 \\
\text{Squared deviations} &= 49, 9, 4, 0, 1, 9 \text{ and } 64 \\
\text{Mean of squared deviations} &= (49 + 9 + 4 + 0 + 1 + 9 + 64)/7 \\
&= 136/7 \\
&= 19{\cdot}43 \\
\text{Standard deviation} &= \sqrt{19{\cdot}43} \\
&= \sqrt{4{\cdot}41}
\end{aligned}
$$

That method of calculating the S.D. is simple when the mean, as in the above example, is a whole number. This, naturally, is not usually the case. As a rule, therefore, the deviations are not whole numbers and squaring them would be laborious. However, it can be shown mathematically that the S.D. can be obtained equally well by finding the

square root of (mean of the squared observations minus the square of their mean). To demonstrate the truth of this, let us apply it to the same example (Table III).

TABLE III

Observations	*Squared observations*
42	1764
46	2116
47	2209
49	2401
50	2500
52	2704
57	3249
343	16943

Then :

Mean of squared observations = their sum divided by their number
= 16943/7
= 2420·43

Mean of observations = 343/7
= 49

Squared mean of observations = 2401

Therefore, S.D. $= \sqrt{2420{\cdot}43 - 2401}$
$= \sqrt{19{\cdot}43}$
= 4·41

The same result as was obtained before.

The above calculation can be expressed in the shape of a formula :

$$\text{S.D.} = \sqrt{\frac{S(X^2)}{N} - \overline{X}^2}$$

where $S(X^2)$ means the sum of the squared observations,
N is the number of observations,
and $\overline{X}$ is the mean of the observations.

In order to calculate the S.D. from a frequency distribution table, the formula is modified thus :

$$\text{S.D.} = \sqrt{\frac{S(fX^2)}{N} - \left(\frac{S(fX)}{N}\right)^2}$$

Where f is the frequency with which each value of X occurs.

The use of this formula is shown in the following example, in which the standard deviation is found of the observations in Table I.

No. of colds $\times$	*Frequency* f	fX $(f \times X)$	fX^2 $(fX \times X)$
0	14	0	0
1	173	173	173
2	247	494	988
3	233	699	2,097
4	112	448	1,792
5	23	115	575
6	12	72	432
7+	0	0	0
	814	2,001	6,057

$$\text{Then S.D.} = \sqrt{\frac{S(fX^2)}{N} - \left(\frac{S(fX)}{N}\right)^2}$$

$$= \sqrt{\frac{6057}{814} - \left(\frac{2001}{814}\right)^2}$$

$$= \sqrt{7{\cdot}441 - 6{\cdot}043}$$

$$= \sqrt{1{\cdot}398}$$

$$= 1{\cdot}182$$

In a grouped frequency distribution table the standard deviation can be found from the modified formula. As an example, Table II is reproduced with one additional column (Table IV).

TABLE IV

I Systolic B.P.	*II* Frequency *f*	*III* Group No. *X*	*IV* fX $(f \times X)$	*V* fX^2 $(fX \times X)$
95–99	2	−7	−14	98
100–104	4	−6	−24	144
105–109	7	−5	−35	175
110–114	12	−4	−48	192
115–119	19	−3	−57	171
120–124	24	−2	−48	96
125–129	27	−1	−27	27
130–134	35	0	−253	0
135–139	26	1	26	26
140–144	21	2	42	84
145–149	18	3	54	162
150–154	13	4	52	208
155–159	6	5	30	150
160–164	5	6	30	180
165–169	2	7	14	98
170–174	1	8	8	64
	222		256	1,875
			− 253	
			3	

Then, by the modified formula :

$$\text{S.D. in "working units"} = \sqrt{\frac{S(fX^2)}{N} - \left(\frac{S(fX)}{N}\right)^2}$$

$$= \sqrt{\frac{1875}{222} - \left(\frac{3}{222}\right)^2}$$

$$= \sqrt{8{\cdot}4457}$$

$$= 2{\cdot}906$$

and since the working unit in this case is a group of 5 actual units, then S.D. in actual units is $2{\cdot}906 \times 5 = 14{\cdot}53$.

The Coefficient of Variation

It should be noted that the scatter of one set of variables cannot be directly compared with the scatter of another set unless the mean of each is the same. Thus, the length of needles and of ninepins may be equally variable, but that would not be apparent by comparing their standard deviations expressed, say, in centimetres. Scatter can be compared, however, by expressing the standard deviation as a percentage ratio of the mean, thus :

$$\text{V, (the coefficient of variation)}, = \frac{\text{S.D.} \times 100}{\text{mean}}$$

In this way the units of measurement are eliminated, as they occur both in the numerator and the denominator, and V becomes a ratio independent of these units.

For example let us compare the relative variability of the data in Tables II and III.

For Table II :

$$\text{Mean} = 132{\cdot}0675 \text{ (}\textit{see}\text{ p. 268)}$$
$$\text{S.D.} = 14{\cdot}53 \quad \text{(}\textit{see}\text{ p. 271)}$$
$$\text{and V} = \frac{14{\cdot}53 \times 100}{132{\cdot}0675}$$
$$= 11$$

For Table III :

$$\text{Mean} = 49 \text{ (}\textit{see}\text{ p. 268)}$$
$$\text{S.D.} = 4{\cdot}41 \text{ (}\textit{see}\text{ p. 271)}$$
$$\text{and V} = \frac{4{\cdot}41 \times 100}{49}$$
$$= 9$$

In effect, this means that the observations in Table II are about 1·2 times as scattered around their mean as are the observations in Table III.

The Normal Distribution

Much statistical theory and method depend upon the fact that variables subjected to the many uncontrollable causes which are collectively called chance form a frequency distribution which can be depicted graphically in the form of a curve and that many biological observations have what is known as a " normal " distribution and may be represented by a " normal " or Gaussian symmetrical curve. This is most easily explained by an example. Accordingly, in fig. 1 there is represented a histogram of the grouped frequency distribution of the systolic blood pressures we have already considered in Table II.

If the size of each group had been made infinitely small and the number of observations sufficiently large, the histogram would have resolved itself into a symmetrical curve as shown in fig. 2.

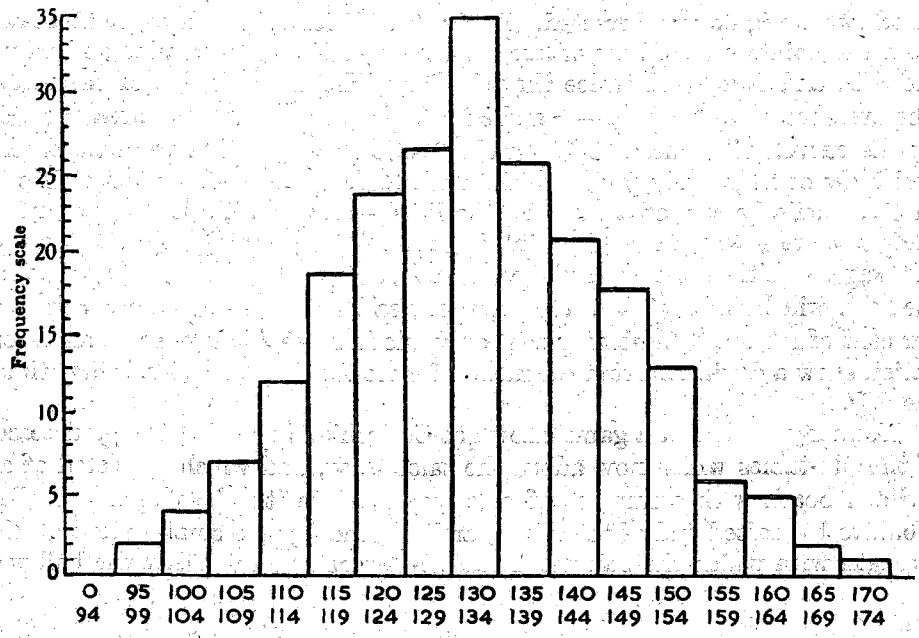

FIG. 1.—Histogram of frequency distribution of systolic blood pressure.

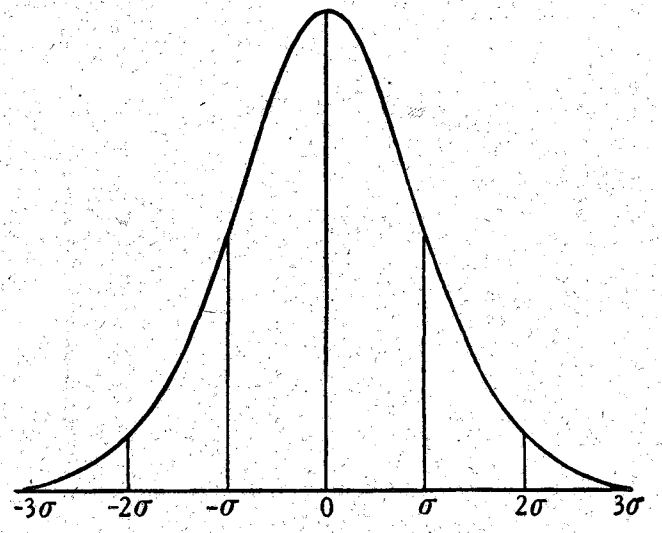

FIG. 2.

In a symmetrical distribution, as we have seen, the mean, mode and median are coincident and they are represented in the figure by the central line which divides the area contained by the curve into two equal parts. The area contained by any two uprights from the base to the curve represents the proportion of the total observations which fall between the two points on the base from which

these two uprights were erected. In fig. 2 additional lines have been drawn from the points on the base where, in a normal distribution, plus and minus the S.D. and twice and thrice the S.D. would lie on each side of the mean. The area between the lines $-\sigma$ and σ is 68 per cent of the total area enclosed by the curve. This means that an observation which differs from the mean, positively or negatively, by more than the S.D. will occur about one time in three. Similarly, the area between the lines -2σ and 2σ is 95 per cent of the total area, which means that the probability of any single observation showing a positive or negative deviation from the mean of more than twice the S.D. will be about 1/20. Again, the area between -3σ and 3σ is 99·7 per cent of the total, indicating that only one in some 370 times will an observation show a deviation from the mean of more than plus or minus three times the S.D.

You may remember a game known as Corinthian bagatelle, the predecessor of the pin-tables which now adorn the saloon bar, and which consisted of an inclined board with a number of pins inserted into it; the player propelled a ball to the raised end of the board, and, depending to a small extent on the strength with which the ball was struck, but mostly on whether the ball was

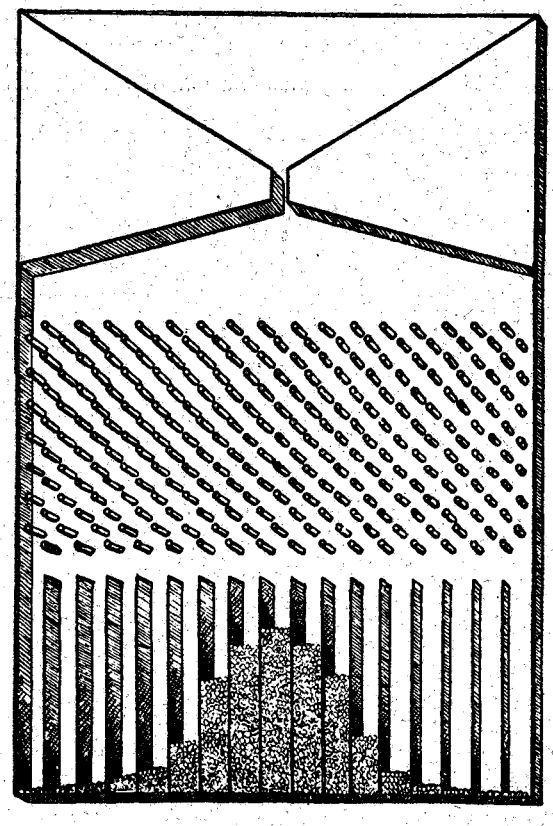

FIG. 3.

diverted to right or left by each successive pin that impeded its progress, the ball eventually came to rest in a particular part of the board marked 5, 10, 20, or 50, and so on. (We will discount for the moment the skill with which the more experienced player could influence the ball's destiny by a judicious nudge of the table when his opponent wasn't looking.) *Galton's Quincunx* is an apparatus of a similar nature, devised, however, to eliminate the influence of the player's skill, and in which the pins are symmetrically placed, each making equilateral triangles with its nearest neighbours (fig. 3).

It will be seen that—

"The ball no question makes of ayes or noes
But right or left . . . . . goes"

depending entirely upon how it is diverted by the successive pins which it strikes. In other words, there is no influence here other than chance.

If a sufficient number of balls were used the final result would represent a histogram of a normal distribution. If the number of balls were infinitely large and their size and the size of the bays infinitely small the final result would be a representation of the normal or Gaussian curve.

The conception of *statistical significance* is based on the foregoing. As has already been pointed out the term significance is used to indicate that the odds are heavy against a certain estimate, difference, or coefficient deviating so much from its expected value by chance. In most cases the odds of 19 to 1 against are usually accepted as the lower limit of signiciance, i.e. a probability of 1/20. This corresponds, as was seen above, to the chance of getting a deviation from the mean of a normal distribution greater than plus or minus twice the standard deviation.

(*To be concluded*)