MILITARY MEDICAL PRACTICE

BASIC NOTES ON STATISTICAL METHODS IN MEDICINE—1

Major D. A. MOORE, M.B., B.S., D.P.H., D.T.M.&H., R.A.M.C.

Introduction

THESE notes are not intended to supplant or augment textbooks of medical statistics. They are designed merely to give a simple, largely empirical explanation of those statistical tests in common use in clinical and epidemiological work together with some indication to their application by means of examples.

It is realised that readers will have had variable prior experience of medical statistics. Experts need read no further. The less confident are assured that the only mathematical skills required are those of the 3rd Form (beyond which standard the writer is ignorant).

It is important to note that there will be no exposition of the problems of the collection of data in the following pages. For an account of the pitfalls of sampling and a general justification of the need for statistical analysis the reader is referred to the clear account given in Chapters I to IV of the 9th edition of Sir A. Bradford-Hill's "Principles of Medical Statistics".

Frequency distributions

Variables

Data of any type and from any source will present as a series of variables descriptive of a particular characteristic. These may be Binary, Discrete or Continuous.

An example of a *binary* variable is that of sex: that is, it can only be male or female. Similarly a situation where the mere presence or absence of a particular characteristic is recorded will produce binary variables.

Fig. 1a. A Histogram. Waist measurements of 200 RQMS's aged 40 years and above.

Fig. 1b. The frequency polygon. Waist measurements of 200 RQMS's aged 40 years and above.

Waist in inches	Frequency	Totals
28	I	1
30	IIII	5
32	IIII II	12
34	IIII III	9
36	IIII IIII I	16
38	IIII IIII IIII	19
40	IIII IIII IIII IIII III	28
42	IIII IIII IIII IIII IIII II	27
44	IIII IIII IIII IIII IIII	25
46	IIII IIII IIII IIII I	21
48	IIII IIII IIII IIII	19
50	IIII IIII I	11
52	IIII II	7
54	III	3
56	II	2

Fig. 1c. The array. Waist measurements of 200 RQMS's aged 40 years and above.

Discrete variables are those that can have only whole number values, for example, the number of children in a family.

Continuous variables are those in which any value may be recorded and the variables form a continuous series. Continuous variables are by far the most commonly encountered in medical work.

Histograms and frequency polygons

The values of variables can be graphically represented against the frequency of occurrence of those values to form:

a. *A histogram*. Where rectangles are drawn above each class interval (see below) and the area of the rectangle represents the frequency of the particular value of the variables (Fig. 1a).

b. *A frequency polygon*. Here individual or grouped values are plotted against the frequency of their occurrence and the points so plotted joined by straight lines. The frequency polygon gives a general idea of the type of *distribution* involved (Fig. 1b).

Both the histogram and frequency polygon are constructed from an initial *array* (Fig. 1c).

Class intervals

In Figure 1c it will be noted that the precise measurements are not recorded but the observations are grouped together in *classes* forming *grouped frequencies*. This, in effect, converts a continuous series of variables into a discrete series. It is an enormously useful device to save time and effort. For reasons that need not detain us here the actual number of classes is best kept between 5 and 15. It is important to make the class intervals equal and quite unambiguous, for example:

<i>not</i>	10-15	<i>but</i>	10-14	<i>or</i>	10-
	15-20		15-19		15-
	20-25		20-24		20-
	25-30		25-29		25-
	30-35		30-34		30-

It is also wise to avoid open ended classes such as: 10-, 20-, 30-, 40-, 50-, 60 +

Dispersion and central tendency

Inspection of the frequency polygon will give some idea of the dispersion of frequency of values and whatever tendency to cluster there may be present. However more objective and precise measurements are needed. Those commonly used are:

a. *The range*. This is the difference between the largest and smallest values recorded. Its use as a measure of dispersion is self evident, but it is a "wasteful" measure as all intermediate values are disregarded.

b. *The mean*. This is the *arithmetic* mean; that is, the sum of all the values divided by the total number of values, or:

$$\bar{x} = \frac{\sum x}{n} \text{ where, by convention, } \bar{x} = \text{mean of } x$$

x = values

n = number of values

(N.B. For those unfamiliar with the notation " Σ "—this is merely the capital Greek letter *Sigma* and indicates "the sum of").

The mean obviously takes account of all values of x and is a measure of central tendency. It is also very useful in further statistical calculations. Its disadvantage is that where there are one or two very large or very small values of x then the mean may give a distorted impression of the true central tendency.

c. *The median.* If all the values in a series are arranged in ascending order of magnitude the median is taken to be the centre value. Plainly, with an even number of values there will be no single centre value. The convention here is to use the mean of the two central values as the median. The median is unaffected by exceptional outlying values and thus may be a useful indicator of whatever central tendency present. It is however of little value in further statistical calculation. Sometimes scatter about the median is expressed in percentiles, that is, the range divided into 100 equal parts; the median becoming the 50th percentile and the inter-quantile range being the distance between the 25th and 75th percentiles.

d. *The mode.* This is the most commonly occurring value in a series of values—the "fashionable" value. Like the median it is unaffected by extremes of range and very large or small outlying values.

e. *The standard deviation.* The above indices are very useful and should always be calculated for any series of values before more complex statistical techniques are applied. However, the best measure of dispersion is the standard deviation which will be fully dealt with below.

The normal distribution

Fig. 2. The frequency distribution of stature in the population.

The frequency distribution of values obtained from measurement in the biological context (and thus medicine) very commonly, although not invariably, falls into a fairly symmetrical shape, the *normal distribution*. In the perfect example this shape is completely symmetrical and the mean, median and mode coincide. This shape can be described and derived mathematically but a near perfect practical example is that of the frequency distribution of stature in the population (Fig. 2). The existence of this form of curve is of fundamental importance to most of the statistical theory behind the tests of significance that will be described later.

Skew distributions

Sometimes frequency distributions do not form a "typical" normal distribution but show a *skew* one way or the other (Fig. 3). The importance of this is that data giving rise to such distributions often needs mathematical manipulation before statistical techniques can be applied. Such methods of correction are beyond the scope of this paper and should be sought in the textbooks.

Fig. 3. Skew distribution.

The standard deviation

As has been seen there are several ways of measuring dispersion and central tendency or, comprehensively, *variation* within a given set of values. The standard deviation is a mathematical device to state this *variation or scatter about the mean* of any series of values as a single expression.

Derivation

Basically, the standard deviation is derived as the root mean square of the differences of individual values from their mean. Steps in the mathematical derivation are given below:

a. Calculate the mean: $\bar{x} = \frac{\sum x}{n}$

b. Calculate deviations by subtracting the mean from each value in turn:

$$x_1 - \bar{x}, x_2 - \bar{x}, x_3 - \bar{x} \dots \dots \dots x_n - \bar{x}$$

c. Square the deviations (which eliminates negative values):

$$(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, (x_3 - \bar{x})^2 \dots \dots \dots (x_n - \bar{x})^2.$$

d. Sum the squared deviations: $\sum (x - \bar{x})^2$ (Here x implies all values of x).

e. Divide the summed squared deviations by $n - 1$ (This, for a mathematical reason need not concern us here, gives a more accurate result than dividing merely by n where n has a relatively small value, say < 30):

$$\frac{\sum (x - \bar{x})^2}{n-1} \text{ This is the variance, } V, \text{ which will appear in later calculations.}$$

f. For the standard deviation, take the square root: $\sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$

Example. Having gone through this derivation at some length to illustrate the principle we can now illustrate the method of use by means of a simple example with a set of figures drawn from the population of Regimental Quarter Master Sergeants (RQMS's) already seen forming the array in Figure 1. In this example x represents individual waist measurements in inches. A random sample of 10 men was drawn (Table I).

Table I
A random sample of waist measurements of ten men

x	$(x - \bar{x})$	$(x - \bar{x})^2$	
37	-5	25	$SD = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$ $\therefore$ substituting we get: $SD = \sqrt{\frac{536}{10-1}}$ $= \sqrt{59.5}$ $= 7.71 \text{ inches}$
42	0	0	
30	-12	144	
48	6	36	
55	13	169	
40	-2	4	
40	-2	4	
45	3	9	
33	-9	81	
50	8	64	
420 = $\sum x$		536 = $\sum (x - \bar{x})^2$	
$n = 10$			
$\bar{x} = \frac{\sum x}{n} = \frac{420}{10} = 42 \text{ inches}$			

Therefore we have shown that in the random sample of 10 RQMS's the mean waist measurement is 42 inches with a standard deviation of 7.71 inches. The application of the calculated standard deviation will be explained later.

Alternative formula for calculation of the standard deviation

Now, having explained the process so far it must be said that the formula given above, while useful to illustrate the principle of the standard deviation, is laborious in practice. Instead a formula derived from it is used:

$$SD = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}} \quad \left(V = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} \right)$$

To show that this is no trick here is the same sample reworked (Table II).

Note that the standard deviation is expressed in the same units as the individual values.

Table II
Example as Table I reworked

$\bar{x}$	x^2	
37	1369	$SD = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}}$ <p>$\therefore$ Substituting we get:</p> $SD = \sqrt{\frac{18176 - \frac{(420)^2}{10}}{9}}$ $= \sqrt{\frac{536}{9}}$ $= \sqrt{59.5}$ $= 7.71 \text{ inches}$
42	1764	
30	900	
48	2304	
55	3025	
40	1600	
40	1600	
45	2025	
33	1089	
50	2500	
$420 = \Sigma x$	$18176 = \Sigma x^2$	

Probability

The normal distribution and the standard deviation

As has been seen in medicine we are making observations drawn from a "universe" of values (here "universe" implies all possible values of any given characteristic) which in most cases we know will be normally distributed. In the application of the standard deviation to such a normal distribution fundamental properties are found to exist. We are enabled to calculate theoretically how many of the observations will lie in the interval between the mean and the mean plus or minus any multiple of the standard deviation. Thus: approximately 68 per cent of the distribution will lie within the range mean ± 1 SD, 95 per cent will lie within the range mean ± 2 SD, 99 per cent will lie within the mean range ± 3 SD (Fig. 2).

Probability

The characteristics of the standard deviation vis-a-vis the normal distribution can now be seen to allow us to make some statement of probability. For example, a single value greater or smaller than the mean of all observations ± 2 SD (a range that will include 95 per cent of the distribution as we have seen above) must be drawn from 5 per cent of all observations that lie outside these limits. It therefore has a 1 in 20 *chance* of appearing. Similarly, an observation greater or smaller than the mean ± 3 SD would have roughly a 1 in 100 *chance* of appearing (as the range mean ± 3 SD includes about 99 per cent of the distribution). In other words, the *probability* (p) of obtaining by chance an observation outside the range mean ± 2 SD is about 1 in 20 or, expressed conventionally as a decimal of unity, 0.05 and for an observation outside the range mean ± 3 SD the probability is 1 in 100 or $p = 0.01$. Tables exist to allow ascertainment of probabilities from various multiples and decimals of the standard deviation.

Significance

Significance is a quality that really lies in the eye of the beholder. For example, in

some circumstances, such as the chance of you winning the football pools, $p = 0.5$ would be highly significant but in general scientific work a result that could occur as often as once in every two times by mere chance would be of little import. However, in this as in many other confusing situations in life, convention has come to the rescue and the level of significance is usually fixed, quite arbitrarily, at $p > 0.05$ or, another way, an observation falling within ± 2 SD of the mean is adjudged not significant. Nevertheless, despite this convention, quoted significance levels should be looked at critically and it should be remembered that with a large enough sample almost any event can be shown to satisfy the convention and be statistically significant. In other words, "significance" does not always imply "importance".

Working units

We have already seen grouped frequencies in the array (Fig. 1c). A combination of this trick with arbitrary *working units* can enormously lighten the arithmetical labour of statistical analysis of long series of variables. The method is best illustrated by an example where it will be used to calculate both the mean and the standard deviation.

Taking our population of 200 RQMS's we can calculate as follows from their recorded waist measurements (Table III).

Table III
Recorded waist measurements of 200 RQMS's

x (inches)	f (frequency of occurrence)	x (working units)	fx	fx^2
28—	1	—6	—6	36
30—	5	—5	—25	125
32—	12	—4	—48	192
34—	9	—3	—27	81
36—	16	—2	—32	64
38—	19	—1	—19	19
<hr/>				
40—	28	0	—157	
42—	27	1	27	27
44—	25	2	50	100
46—	21	3	63	189
48—	14	4	56	224
50—	11	5	55	275
52—	7	6	42	252
54—	3	7	21	147
56—	2	8	16	128
<hr/>				
$n = 200$			330	$\Sigma fx^2 = 1859$
			—157	
			$\Sigma fx = 173$	

To calculate the mean; $\bar{x} = \frac{\Sigma x}{n}$ or, in working units. $\frac{\Sigma fx}{n}$ Therefore substituting we get $\frac{173}{200} = 0.865$. What we really calculated at this stage is that the mean lies 0.865 of a working unit above (since it is positive) the mid-point of the class we have selected as), i.e. 40— or, to two places of decimals this implies 40.00 — 41.99; the mid-point being 40.995. But, we are working in double units (class interval is 2). Therefore we must multiply the working unit by a factor of 2 to convert it into inches. The true mean is therefore: $40.995 + (0.865 \times 2) = 42.73$ inches correct to 2 places of decimals.

The calculation of the standard deviation can be done so:

$$SD = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n - 1}}$$

or, in working units: $SD = \sqrt{\frac{\sum fx^2 - \frac{(\sum fx)^2}{n}}{n - 1}}$

therefore, substituting we get: $\sqrt{\frac{1859 - \frac{(173)^2}{200}}{199}} = \sqrt{\frac{1709}{199}} = 2.93 \text{ w.u.}$

As with the mean we must multiply this by the class interval to get the standard deviation in real units. therefore: $2.93 \times 2 = 5.86 \text{ inches.}$

Important points to remember in the use of this technique are:

- 0 represents the value of the mid-point of the class opposite which it is placed and it is immaterial which class is chosen to be so represented. Indeed calculations can be checked by going through the process twice with 0 opposite a different class each time.
- In resolving the working units the class interval or width must be taken into account (in our example above it was 2).
- The class interval should be equal.
- The values smaller in real units than those in the class represented by 0 should be given the negative values of working unit.

Coefficient of variation

The coefficient of variation is used in the comparison of the degree of variability around their respective means of two quite separate and independent distributions. This is done by expressing as a percentage the ratio of the standard deviation to the mean. Thus: Coefficient of variation = $\frac{\text{Standard deviation}}{\text{Mean}} \times 100$.

As, in comparing coefficients we are really comparing ratios, any difference in the units in which the individual distributions were recorded is immaterial.

Here is an example: In a large sample of mice the mean length of the rodents was 80 mm with a standard deviation of 4 mm. Similarly, calculations from a large sample of men showed a mean length (= height) of 68 inches with a standard deviation of 3.4 inches. The question is whether we are seeing, allowing for differences in absolute values and the units used, a similar variability of distribution about their respective means in these two samples of mice and men. Using our equation we get:

$$\text{Coefficient of variation (mice)} = \frac{4}{80} \times 100 = 5 \text{ per cent}$$

$$\text{Coefficient of variation (men)} = \frac{3.4}{6.8} \times 100 = 5 \text{ per cent}$$

So, we have shown that, despite the obvious difference in magnitude, between mice and men and the recording of the individual values in different units, we have been able to demonstrate that the distribution of their lengths about the mean length is similar for these particular samples.

(to be continued)

Mitchiner Medal

Lieutenant-Colonel R. Scott, M.A., B.M., B.Ch., F.R.C.S., D.T.M. & H., R.A.M.C., has been awarded the Mitchiner Medal for 1975 by the Royal College of Surgeons of England.

ACADEMIC ACHIEVEMENTS

F.R.C.O.G.

COLONEL J. H. ANDREW, M.B., ChB., M.R.C.O.G. late R.A.M.C.

M.R.C.P.

LIEUTENANT-COLONEL J. G. HOLMES-SMITH, M.B., B.S., M.R.C.S., L.R.C.P., D.obst. R.C.O.G., D.C.H., D.T.M. & H., R.A.M.C. MAJOR C. M. BATE, M.B., ChB., R.A.M.C. MAJOR P. LYNCH, M.B., ChB., R.A.M.C.

D.M.R.D.

MAJOR C. N. O'N. DIGGES, M.A., M.B., B.Ch., B.A.O., R.A.M.C.

D.C.H.

CAPTAIN J. HALLORAN, M.B., B.Ch., D.obst.R.C.O.G., R.A.M.C.