Mathematics in Medicine

Statistical ritual in clinical journals: is there a cure?—II

DONALD MAINLAND

Variability

STANDARD DEVIATION

The shambles mentioned at the beginning of Part I referred to the confusion in the use of standard deviations and standard errors of means,¹ with the latter apparently more popular, perhaps because they are so much smaller than their parent standard deviations and thus suggest higher precision. The author defined the standard deviation as representing the variation among the individual measurements,² but this does not help us unless we know what multiple of standard deviation would enclose a certain percentage of the observations. The formula commonly used is mean ± 1.96 SD (or 2 SD), on the assumption that we have individual measurements that, if increased in number, would form approximately a Gaussian distribution—an assumption for which most medical data provide scanty evidence. Bell shaped is not synonymous with Gaussian; and a trick such as the logarithmic transformation cannot be guaranteed to make the approximation safe.

We should remember that the Gaussian curve started as the "normal curve of error," a mathematical invention that was believed to represent the scatter of the variable (chance) errors in physical measurements, but thorough investigation³ failed to substantiate the belief.

Perhaps, however, some authors producing a standard deviation do not visualise any distribution. If they did, I would expect to see comments when mean *minus* 2 SD (or even *minus* 1 SD) leads to a ridiculous negative value in data such as: sodium concentration in saliva before and after a certain drug treatment; days off work after minor surgery for haemorrhoids; activity of an enzyme in ascitic fluid in hepatic cirrhosis; serum concentration of amyloid in rheumatoid arthritis; duration of labour; number of vaginal examinations during labour.

The widespread use of the Gaussian distribution to express intersubject variation has been apparently due, firstly, to the illegitimate application of the mathematical term normal (Gaussian) as if synonymous with healthy,⁴⁻⁵ and, secondly, to the use of standard deviation to produce standard error of the mean in the Gaussian based *t* test.

STANDARD ERROR OF MEAN

Standard error of the mean is also a standard deviation, an estimate of the variation among means of equal random samples from a population of means. It is a measure of the precision of

a mean,² but this is a risky definition, for sometimes standard error of the mean appears to be interpreted as an absolute limit of error, as in instrument calibration and in the statement: "The patients were seated in a controlled ambient temperature of $22^{\circ} \pm 1^{\circ}\text{C}$."

PERCENTILES

The safest way to present individual variation is to show the actual measurements, but if an overall picture is also desirable there is a simple method which entails no assumptions but has been long neglected by those who have been led to trust the Gaussian curve standard deviation, as was the reviewer of a pathology textbook who in 1982 wished that the author had given normal organ weights and their standard deviations. About 50 years previously the difficult problem of selecting "normal" organs had been discussed, and when material that met the stipulated requirements was collected the Gaussian curve was clearly unsuitable to represent the variation; so the weights were published in percentile form in tables which were later reproduced for anatomists.⁶⁻⁹ The percentile method is versatile¹⁰⁻¹¹ and there seems to be a revival of interest in it.^{2 12 13}

An imaginary world; populations and samples

The name of Gauss, inventor of the curve, should remind us that all statistical tests and estimates come from an abstract (imaginary) mathematical world in which populations—for example, a Gaussian distribution of measurements—can be completely defined and a distribution of truly random samples can be obtained. In the inspection of a consignment of lamp bulbs for the percentage defective the inspector can, like the mathematician, point to his population, and random sampling can be applied. In contrast, however minutely we define a population of patients with primary hypertension or patients with mammary cancer we can never be sure of including all the characteristics relevant to the purpose of our research. From an incompletely defined population we cannot take a strictly random sample and in ordinary medical research our samples are handy (available) samples. If we use on them tests and estimates derived from the imaginary world, therefore, our statements ought, strictly speaking, to begin with "If" and their verbs should be in the subjunctive—"If this were a strictly random sample from a population completely defined in all relevant respects the inference would be. . . ." Thus, as in everyday life, all that a handy sample can tell us definitely about its parent population is itself. True, we take a sample to learn about material of like kind, and if the other material is very much like the sample—say, repeatedly reliable products from a certain shop, or similar patients and circumstances in the same hospital the year after a clinical trial—we feel considerable confidence;

Kent Hills, Kent, Connecticut 06757, USA

DONALD MAINLAND, MB, DSC, retired professor of anatomy and medical statistics

Correspondence to: Kent Hills, Apt 3B1, Kent, CT 06757, USA.

but if we express the strength of our feelings numerically as 10 to 1, 50 to 1, or 99%, we are indulging in a kind of pseudo-quantification.

Gauss in group comparisons

Although it is an undependable guide to individual variability, the Gaussian curve is used annually in myriads of significance tests by χ^2 and the t test which are substitutes for randomisation trials.¹ The results are widely trusted because, even when a parent population is very non-Gaussian, samples of means and of binomial frequencies very often, with increase in sample size, approach the Gaussian form. An experimenter, however, who wishes to apply a test after a randomised controlled trial might ask how well founded is the almost universal faith.

χ^2 AND FISHER'S EXACT TEST

The "exact probability" test of 2×2 frequency tables with fixed marginal totals is equivalent to an infinitely large randomisation trial. Numerous comparisons showed that, with Yates's "correction for continuity," χ^2 commonly gave a close approximation to the exact test,¹⁰ but nowadays I use the latter because the formula (Armitage, section 4.8) entails only the addition and subtraction of logarithms of factorials on a small pocket calculator.¹⁴

THE t TEST AND ANALYSIS OF VARIANCE

The inquiring experimenter will find two requirements for the validity of the t test and analysis of variance: firstly, Gaussian distribution of observations and, secondly, equality of variation in the groups under comparison¹³—both often very unrealistic for groups under different treatments. He may sense that conformity with the two requirements is often judged by the eye, but he may read that it is not always possible to say how far the raw data can deviate from the ideal before the results become invalid,¹⁵ that the problem is somewhat subjective, and that expert help is recommended.^{15 16} Trying to see what an expert might offer, he could find in Armitage's advanced textbook (section 4.6) an informative statement.¹⁴ Even without understanding the mathematical formulae he would learn that in some cases only "approximations" to exact probability can be obtained, that the choice of methods often depends on the statistician's approach to the logic of statistical inference (three theories are alluded to), and therefore the choice may be contentious.

Longing for the time when randomisation trials¹⁷ by computer will become common, the experimenter would meanwhile probably favour non-parametric (distribution free) tests, such as Wilcoxon's rank order tests.

Multiple comparisons

Initially the concept of a 5% error in significance testing seems fairly simple, but its complexities are becoming appreciated, as was illustrated by a cardiologist who emphasised the danger after a clinical trial of producing too many false verdicts of "significant" by comparing subgroups, either formed retrospectively (data dredging) or by prerandomisation stratification.¹⁸ The extent of error and the proposed cures vary with the design of the study,^{14 19} but it is easy to see what may happen in clinical trials.²⁰

If throughout a lifetime we used the 5% ($P=0.05$) level of significance and if at the end we could discover all the verdicts that we had pronounced when in fact the randomisation was responsible, we would find approximately 5% significant (S) and 95% not significant (NS). If now we took the verdicts at

random, two at a time, as in testing a treatment difference in two subclasses, we would find mostly two NS's together, sometimes one S and one NS, and occasionally two S's. To find numbers for these expressions we could take several thousand uniform discs, 95% NS's and 5% S's, shuffle them thoroughly, and take samples of two at a time; but we can get the answers much more rapidly and accurately by expanding the binomial expression $(p+q)^N$, where $p=0.95$, $q=0.05$, and N is sample size. We need only the first term of the expansion, $0.95^2=0.9025$; that is, 90.25% of the two subsample tests would produce verdicts with two NS's and the other 9.75% would contain at least one S verdict. So, thinking that we had only a 5% risk of falsely rejecting the effect of randomisation, we would have almost a 10% risk. For comparisons of treatment in three subclasses $0.95^3=0.8574$ —that is, our error risk is about 14%.

More play with binomial expansions shows that in these simple cases if we wish to retain a risk of real error of 5% ($P=0.05$) but make two comparisons in the same experiment we should divide 0.05 by two and work with 0.025 as the cut off point. For three comparisons division of 0.05 by three would set our cut off point at 0.0167; and so on.

Confidence intervals

When a sample value (measurement or frequency) has been found an important question is: What may be the true value which would be approached by taking more and more of the same material under the same conditions? Statisticians now commonly answer by saying: "You can have 95% confidence that the true value lies in the interval between LL (lower limit) and UL (upper limit)." They supply the numerical values for measurement data, and often for binomial frequency data, from the Gaussian curve. It is not difficult, therefore, to find out what, in the imaginary world, "confidence" (a probability statement) implies. In the real world even expert statisticians find the same problems as in significance testing (Armitage, section 4.6).¹⁴ In medical research our problem is much greater, for we have handy samples only, and therefore interval estimates can merely indicate how great our ignorance might be, even if we had strictly random samples of an adequately defined population. We should note also whether the estimated limits are clinically or biologically unrealistic.

Correlation and regression

Some of the numerous pitfalls in correlation have recently been discussed.^{2 21-23}

Regression, perhaps more lucidly called "average trend," is much more useful in medicine than is correlation. It was, as usual, discussed with reference to the prediction of "expected" values, significance tests, confidence intervals, linearity, and the possibility of non-linearity.^{2 21-23} The specified requirements (Gaussian distributions and equal intersubject variation of Y values at all values of X) lead to problems that entail assumptions and frequently subjective decisions.

Perhaps some day an investigator who believes that scientific knowledge can be acquired without such entanglements will construct, for comparison with other independent observations, a regression line by the convenient "least squares" technique, without worrying about the meaning of "best fit." Then he may record the magnitude of the deviations of the Y values from the line, and look for differences in scatter at different parts of the line.

Future possibilities

I may have been overoptimistic in imagining that medical investigators would be willing to use a rather old fashioned, apparently tentative but intellectually demanding approach to

data analysis, instead of ritual that imparts an aura of mathematical proof. If I have been mistaken it is probably because my recent scrutiny of reports on small scale clinical research has given me the impression that, though often astray regarding statistical arithmetic, the authors were commonly anxious to perform a careful, unbiased study, and to weigh the results thoughtfully.

The extensive replacement of ritual by realistic and critical thinking would not be speedy, but it could be hastened by study of the writings of a clinical investigator who is well versed in statistics.^{24 25}

I thank Mrs R N Collins and Mrs R W Moore of the Sharon Hospital Health Sciences Library for their generous and invaluable help in obtaining published reports used in preparing these articles and Miss Maggie Wells for her helpful cooperation including her expert typing.

References

- 1 Mainland D. Statistical ritual in clinical journals: is there a cure? *I. Br Med J* 1984;288:841-3.
- 2 Altman DG. Statistics and ethics in medical research. *Br Med J* 1980;281:1542-4.
- 3 Topping J. *Errors of observation and their treatment*. London: Institute of Physics, 1955.
- 4 Elveback LR, Guillier CL, Keating FR. Health, normality and the ghost of Gauss. *JAMA* 1970;211:69-75.
- 5 Feinstein AR. Clinical biostatistics XII: on exorcising the ghost of Gauss and the curse of Kelvin. *Clin Pharmacol Ther* 1971;12:1003-16.
- 6 Boyd E. Method of establishing probable limits of normal variation in weights of organs. *Anat Rec* 1935;62:1-6.
- 7 Boyd E. Normal variability in weight of adult human liver and spleen. *Arch Pathol Lab Med* 1933;16:350-72.
- 8 Wald H. Weight of normal adult human kidneys and its variability. *Arch Pathol Lab Med* 1937;23:493-500.
- 9 Mainland D. *Anatomy as a basis for medical and dental practice*. New York and London: Paul B Hoeber, 1945.
- 10 Mainland D. *Elementary medical statistics*, 2nd ed. Philadelphia: WB Saunders Company, 1963. (Reprinted (Miller MC, ed) Ann Arbor: Biometry Imprint Series Press, 1978.)
- 11 Mainland D. SI units and acidity. *Br Med J* 1977;iii:1219-20.
- 12 Feinstein AR. Clinical biostatistics XXVII: the derangements of the "range of normal." *Clin Pharmacol Ther* 1974;15:528-40.
- 13 Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. *Br Med J* 1983;286:1489-93.
- 14 Armitage P. *Statistical methods in medical research*. Oxford: Blackwell, 1971.
- 15 Altman DG. Statistics and ethics in medical research. *Br Med J* 1980;281:1399-1401.
- 16 Gardner MJ, Altman DG, Jones DR, Machin D. Is the statistical assessment of papers submitted to the "British Medical Journal" effective? *Br Med J* 1983;286:1485-8.
- 17 Mainland D. Medical statistics—suggestions for the evaluation of introductory textbooks. *J Chronic Dis* 1983;36:345-51.
- 18 Hampton JR. Should every survivor of a heart attack be given a beta-blocker? Part I. Evidence from clinical trials. *Br Med J* 1982;285:33-6.
- 19 Wallenstein S, Zucker CL, Fleiss JL. Some statistical methods useful in circulation research. *Circ Res* 1980;47:1-9.
- 20 Tukey JW. Some thoughts on clinical trials, especially problems of multiplicity. *Science* 1977;198:679-84.
- 21 Altman DG. Statistics and ethics in medical research, V—analysing data. *Br Med J* 1980;281:1473-5.
- 22 Gore SM. Statistics in question. Assessment methods—recognising linearity. *Br Med J* 1981;283:711-3.
- 23 Gore SM. Statistics in question. Assessing methods—a feel for other things. *Br Med J* 1981;283:775-7.
- 24 Feinstein AR. *Clinical judgment*. Baltimore: Williams and Wilkins, 1967.
- 25 Feinstein AR. *Clinical biostatistics*. Saint Louis: C V Mosby, 1977.

(Accepted 29 November 1983)

Recently it was suggested that intercourse in young girls predisposed to uterine cancer. How valid is this claim?

This claim refers to only one type of uterine cancer—squamous carcinoma of the cervix. A connection between intercourse and cervical cancer was apparently first suggested in 1842, and it is now well documented that the disease is rare in nuns and common in prostitutes.¹ The two main coital factors seem to be early start of intercourse and multiple partners, but because these are interrelated it has been difficult to tell which is the more important. Evidence comes from two sources—epidemiology and colposcopy. In some epidemiological studies² the risk of cervical cancer was more closely associated with early age at first coitus than with other indicators of promiscuity, but a more recent study³ of women with cervical dysplasia or carcinoma in situ suggested that the total number of sexual partners is a more important factor than age at first coitus. Colposcopy⁴ has shown that compared with virgins, the cervixes of promiscuous teenage girls have smaller transformation zones with more frequent atypical transformation: in one series⁴ minor histological lesions tended to occur in girls who started coitus before age 15, and major histological abnormalities tended to occur in women who were promiscuous early in their sexual life. To explain all these observations, various theories have been proposed. It has been suggested that the cervix during adolescence may be particularly susceptible to some agent transmitted during intercourse, but the nature of the agent remains unclear. It may be a virus (such as herpes or papilloma), or even the DNA of the spermatozoa themselves. A male factor seems important: possibly some men produce particularly carcinogenic sperm, or perhaps promiscuous men¹ are more likely to acquire a viral agent and transfer it to their partners. Barrier contraception may offer some protection against the carcinogen, and cervical smearing should begin within a few years of first intercourse.⁵—JAMES OWEN DRIFE, senior lecturer in obstetrics and gynaecology, Leicester.

¹ Skegg DCG, Corwin PA, Paul C, Doll R. Importance of the male factor in cancer of the cervix. *Lancet* 1982;ii:581-3.

² Rotkin ID, Cameron JR. Clusters of variables influencing risk of cervical cancer. *Cancer* 1968;21:663-71.

³ Harris RWC, Brinton LA, Cowdell RH, et al. Characteristics of women with dysplasia or carcinoma in situ of the cervix uteri. *Br J Cancer* 1980;42:359-69.

⁴ Singer A. The uterine cervix from adolescence to the menopause. *Br J Obstet Gynaecol* 1975;82:81-99.

⁵ Wright VC, Ropelle MA. Age at time of first intercourse v chronologic age as a basis for Pap smear screening. *Can Med Assoc J* 1982;127:127-31.

In performing a perfusion scan of the lungs I understand that macroaggregates of a radioactive material, injected intravenously, impact in lung capillaries. From these sites they emit radiation, which then provides an indication of the number and patency of small vessels leading to the capillaries. What is the radiation dose in the cells of the capillary wall at the site of impaction of an average macroaggregate? Is it hazardous?

In most cases perfusion lung scans are carried out using Tc^{99m} labelled macroaggregated human albumin. About 600 000 particles are injected and the activity of Tc^{99m} is about 2 mCi (74 MBq). Roughly 90% of the particles are within the range 10-40 μ , and during the first passage through the pulmonary circulation after intravenous injection 90% of the particles are trapped within the capillaries of the lungs. They are subsequently cleared from the lungs, with an effective half life (biological plus physical decay) of about four hours. Most of the radiation dose to the lung tissue arises from the 140 keV gamma ray emission of Tc^{99m}. Estimates of the mean radiation dose to lung tissue for a standard 70 kg adult, assuming a uniform distribution of activity within the lung tissue, vary from 112 to 280 millirad per millicurie administered (0.03-0.075 mGy/MBq). Tc^{99m} has a complex decay scheme, however, and also emits various low energy x rays and electrons. These will not be very penetrating and will give rise to considerably higher doses (probably many thousands of rads) to the very small amounts of tissue actually in contact with the particles. The high dose region will extend only a small distance (< 10 μ) beyond the particle. If there are 600 000 particles the total mass of lung tissue receiving a very high dose is about 20 mg. The total mass of adult lungs is about 1000 g. There is very little difference between lung scan and any other nuclear medicine technique in which radioactive tracer is taken up in the cells of an organ of interest. Inevitably, since there is a small amount of low energy radiation, the radiation dose around the site of the radioactive atoms will be high, but since the volume of tissue affected is a small fraction of the whole this is unlikely to have serious biological consequences.—S EDWARDS, principal physicist, and M N MAISEY, consulting physician in nuclear medicine, London.

Technical product data. *Technescan MAA*. St Louis, Missouri: Mallinkrodt Inc. Robbins PJ, Feller PA, Nishiyama H. Evaluation and dosimetry of a Tc^{99m}-Sn-MAA lung imaging agent in humans. *Health Physics* 1976;130:173-8. Berger MJ. Distribution of absorbed dose around point sources of electrons and beta particles in water and their media. *J Nucl Med* 1971;12:suppl 5.