INT. J. MATH. EDUC. SCI. TECHNOL., 1986, VOL. 17, NO. 1, 95–102

Changes needed in style and content of teaching statistics to medical undergraduates

by A. INDRAYAN†

Department of Statistics, The Ohio State University,
Columbus, Ohio, U.S.A.

(*Received 22 May 1984*)

The negative attitude of medical undergraduates towards statistics could be ascribed to adherance to traditional statistical topics, foreign to the medical sciences. The need is to teach in a medical language. This communication describes the change needed in style and content of teaching to implement this strategy. The medical topics are identified and the details of teaching specified. A discussion on normal values, their clinical significance, diagnostic criteria, and choice of diagnosis provides opportunity to teach measures of location and dispersion, distributions, type I and II errors, and some probability. Interpretation of laboratory results and of medical literature brings out essentials of sampling, some aspects of design and meaning of *P*-values.

1. Introduction

Variability and uncertainty are important characteristics of all living beings but none concerns us more than human beings. Medical sciences face such variabilities every day, and make decisions important to our life and health. From this angle, the science of minimizing and measuring uncertainties—statistics—looks a *sine qua non* to medicine. Ironically, the role is not fully appreciated at grass-roots level, though increased appreciation has been observed later in the career [1]. In a list of 18 subjects ranked by British medical students according to interest, usefulness and easiness, statistics got almost the last rank [2]. More than three-fourths of statistics instructors in American and Canadian medical schools, responding to a mail survey, perceived that students merely tolerate, dislike, even abhor their courses [3]. This negative attitude could be ascribed to two basic reasons, both relating to the clinical relevance of the subject. First, not so important, is the clubbing of teaching of biostatistics with epidemiology in many medical schools. Statistical methods do have a special significance for epidemiology, and sometimes they cannot be separated, so that the association should continue, but the emphasis on clinical issues should not be allowed to be diluted. Secondly, far more important, is trying to impart statistical education even to medical students through traditional statistical topics like presentation of data, measures of location and dispersion, and tests of hypotheses. These topics are foreign to medical undergraduates and alienate him. Even some of the recent series in medical journals [4, 5], though entitled 'statistics for clinicians', fail to recognize that statistics, as any other medical subject, need to be discussed in medical language. Some other series [6, 7] start with clinical issues but focus primarily on statistics as it appears in medical research rather than in everyday clinical practice.

† Present address: University College of Medical Sciences, New Delhi, India.

It has lately been recognized that the fault is not with the subject itself but with the style and content of teaching. The need to make statistics teaching more convincing [3] to the medical students and clinically relevant [8] has been emphasized. Such changes as shifting to seminar format [8], using local material and defining precise objectives [9], and of making it laboratory-problem oriented [10] have been tried. Much more needs to be done to convince the students that statistics is almost as essential for clinical practice as any other basic science conventionally incorporated in the medical curriculum.

We have been concerned with the aspect of the integration of statistics with medical sciences ever since we took up this challenge 15 years ago. The most important realization has been that quantitative concepts could more effectively be conveyed by orientation of teaching to medical issues. The following are some suggestions on how this could be done. It is necessary to relate such a change to the objective of teaching statistics at undergraduate level. The objective is not to apprise them of the statistical methodology, but to inculcate a quantitative appreciation and objectivity in medical practice and perhaps to develop a little more exactitude in the thinking of future physicians. Thus, the learning is supposed to be mostly affective rather than psychomotor. The format given below could easily achieve these objectives. In line with the tone of our suggestions, we refrain from using any formula in this communication. Also, in place of research examples, everyday clinical problems faced by physicians are used for greater educational impact. The description is divided in two major parts: the core topics comprising issues directly arising out of clinical questions, and the ancillary topics arising out of peripheral issues.

2. Core topics

2.1. *Normal values and their clinical use*

Statistics teaching customarily begins with measures of location. This could indeed be an excellent topic to introduce a medical student to a quantitative area. Realizing that the basic purpose of calculating the mean, median and mode is to get a representative central value, generally called normal in medical parlance when based on normal subjects, the teaching could very well start with the normals—their meaning and significance in a particular case. Thus the starting point is the normal values and not the mean, median or mode. Once the style is thus changed, the student may immediately develop a positive attitude towards learning, as this now becomes an area of interest—one of the medical sciences. Indicate that the concept of normals is necessary to make clinical decisions in the face of variations and uncertainties. Emphasize that the term 'normal' may have varied meanings in different contexts, but generally represents the usual level present in normal subjects. It is a kind of representative value generally obtained in terms of the mean, median or mode. A normal body temperature of 37 °C (98·6 °F) is the mean, while the normal haemoglobin level of 15 g/dl is an example of the mode. Normal height–weight in child growth assessment is a median. This also brings out the fact that the normal values, since they are central in nature, are only indicative and not definitive. Variations around them are perfectly normal and expected. Also, the normals could be different for various age-groups, sex, ethnicity, heredity, etc. They may depend upon the time of the day, exercise and posture. Blood pressure (b.p.) is an example which is affected by all these factors. In assessing the clinical significance of a measurement for a particular case, besides all the above intrinsic factors the possible

effect of extraneous sources such as errors in measurement, variations in laboratory techniques and instruments, and inter-observer differences also need to be considered.

The question, which of the three—mean, median or mode—is to be used in a particular setting needs to be properly addressed. The mean is generally preferred because it is universally understood and easily calculated. In some cases, e.g. haemoglobin, lower values are more commonly seen than the higher values in normal subjects, and the clinical interest is better served by mode rather than mean. Often, as in case of pulse rate and heart rate, the mean and mode do not differ in value. The answer thus depends on the pattern of variability of the observations. This is studied in terms of shape of the distributions such as symmetric (as of body temperature), positively skewed (as of blood pressure), and negatively skewed (as of haemoglobin level).

It is only after such convincing arguments as the above that one can hope to attract the attention of the students to the method of computing the mean, median and mode. Keep in mind that medical students, in general, are repelled by mathematical work. Therefore, the calculations have to be restricted to simple cases. For example, nothing on the calculation for grouped data should be taught at this stage, nor need it be at the undergraduate level. The purpose of teaching calculations is that they should better understand the meaning of mean, median and mode, and hence of normals.

In the case of the growth of a child, the interest sometimes is to know the percentage of children that have height or weight worse (or better) than the measurement for a particular child. Thus, the concept of percentiles, measuring locations other than the central, is introduced. If b.p. of an otherwise normal person is obtained as 160/95 mmHg, it helps to know, for example, that only 1% of normal people have a b.p. as extreme as this and thus the value has some clinical significance.

Single normal values like 37 °C (98·6 °F) for body temperature, rather than a range, are good for clinical use when the values in normal subjects do not differ much from each other. Blood glucose, blood pressure and serum cholesterol differ widely from person to person, as also within a person, so that there is a whole range of normal values for these variables.

2.2. *Normal range and the associated risks*

The foregoing discussion brings out the fact that some medical measurements are more variable than the others. The range of normality depends on the amount of variability seen in a normal population. Thus, there is a need to measure variability. The minimum to maximum ever seen in normal subjects, could be one such measure. It presents a highly distorted picture if these extremes happen to be unusual values. The most widely acceptable and frequently used is the standard deviation (s.d.). Giving the formula straight for computing s.d. may switch off the students and an explanation such as the following could help. Variability is the difference in the values from one another. In place of finding the difference from one another, it is convenient to get differences from a central value—the mean. Some of these deviations from the mean would be negative and the others positive so that their average is always zero. To get rid of negative signs, one can think of the absolute values and their average—leading to the mean deviation. It is mathematically simpler though to get rid of negatives by squaring each deviation. The average of these squared deviations is called variance. This is in squared units, e.g., for blood

cholesterol it could be 1600 $(mg/dl)^2$. To get back to the original units, take the square root and obtain the s.d.

In normal populations, many medical variables, but not all, have a distribution characterized by highest frequency around a single value, and an initially gradual and then fast tapering off of frequencies on either side of that central value. Body temperature, pulse rate, heart rate, blood glucose level, intra-ocular pressure, serum creatinine level, are all examples of such a variable. One of the properties of many such symmetric unimodal distributions is that mean ± 2 s.d. covers measurements of approximately 95% of the normal people. The percentage is exactly 95 in the case of what is known as a Gaussian distribution. The name Gaussian is preferable as normal has a different meaning in medical context. Make it absolutely clear that the range of normality so obtained has the inherent drawback of excluding 5% of the normal persons who happen to have high or low values at the extremes. But then, no practical limits can perhaps ever be constructed free from this risk. Till such time that a better method is evolved to delineate the normal range, such limits are accepted as a guide by which to make clinical decisions and have proved to be useful. The limitations of such a range are that (i) mean -2 s.d. may sometimes be an impossible negative as obtained by O'Brien and Shampo [4] for serum urea, (ii) the variable may not have a normal distribution, in which case the normal range could be obtained by using the same principle but by the method of percentiles [11], and (iii) a reading beyond such normal does not necessarily imply an adverse prognosis. The last is a serious objection, and can be overcome only by prospective study of people with readings at the extremes. The objective of such teaching at undergraduate level is not to provide skill in being able to find the normal range, but only to provide knowledge for the judicious use of the limits already available in the literature.

No matter how the limits are arrived at, there is always some risk of wrongly calling a healthy person sick, and then unnecessarily instituting treatment. This is a serious error and must be kept at a low level. This can be achieved by extending the range of normality. What unfortunately happens, for example in extending the normal b.p. limits to include, say, 165/95 mmHg, is that the risk of calling a sick person healthy increases. This is the second type of error of misclassification. When one error is reduced, the other goes up. The cut-off points are chosen to balance and minimize these risks. No details of the distribution of the variable in sick cases and its implications for the range of normality could be given at undergraduate level, but the discussion above could make students aware of both types of error, and of the fact that the errors can only be minimized, not eliminated.

The second kind of error—of wrongly classifying a sick person as healthy—is rare in cases where diagnosis and treatment is based primarily on the value of a particular measurement. Glaucoma, based on intra-ocular pressure, diabetes on glucose tolerance, and essential hypertension on b.p. levels, are such examples. In most of the other cases, as in enzymes for liver diseases, the overlap in the values for normal and sick people is huge and it takes on simultaneous consideration of several variables to come to a reasonable conclusion. Sometimes, it is vital to decide which one of the several variables is better in discriminating between the normal and the sick, or between various degrees of sickness. This is done as follows.

2.3. *Choice of diagnostic criteria*

In the case of liver diseases, for example, the diagnosis could be based on the signs and symptoms, or on the clinical chemical tests, or on morphological criteria. In

practice, there will possibly be a combination of these, but in case of conflicting evidence, there is the question of one criterion being better than another. The text provides information on the signs–symptoms–measurement generally found in a sick case. The actual problem faced by a clinician is the reverse—to establish the diagnosis on the basis of signs–symptoms–measurements. The percentage of times a particular test or criterian is correctly positive in established cases is called the sensitivity of that test. The percentage of times a particular test yields a negative result in known healthy subjects is called its specificity. Sensitivity and specificity are easily calculated by a retrospective look at the data on known sick and known healthy people. What is really needed for diagnosis is predictivity—the chance that a person has a particular problem when the test says he has (positive predictivity), or the chance that a person does not have that problem when the test says he has not (negative predictivity). Prospective study of cases with positive and negative tests is needed to get the right answer. The other method of calculating predictivity on the basis of sensitivity and specificity is to use Bayes's theorem. No mathematical details can be given at this stage except to emphasize that predictivity also depends on the rate of prevalence of that disease. A positive result on a test which is 95% sensitive and 92% specific provides almost no clue to the presence of disease if the disease is rare [12]. A negative skin test for tuberculosis has good negative predictivity to exclude the disease but a positive test indicates almost nothing. On the other hand, a positive histology of a biopsy tissue has good positive predictivity, but a negative histology does not necessarily exclude the disease. Thus, the choice of diagnostic criteria depends on the predictivity needed to confirm or exclude a diagnosis.

The opportunity could also be exploited to explain the concept of probability without going into details. Define it as the long-term chances or the long-term relative frequency. Notation, if at all, should be used with caution. One everyday instance of its use, as discussed above, is in evaluating the chances that a person has a particular problem given certain signs–symptoms–measurement. The final diagnosis and the type of treatment depends on such probabilities. Faced with profound variations and uncertainties, medicine has to depend on probabilities. These may not be formally calculated on paper, but they are invariably done in the mind. As more and more information through laboratory or other tests is made available, the probabilities change and the diagnosis becomes more and more sharp.

The teaching as above covers most of what is traditionally covered in a statistics course for medical undergraduates—the measures of central tendency, quantiles, measures of dispersion, normal and other distributions, type I and II errors, and some aspects of probability. We have intentionally excluded the epidemiologic part comprising measures of health, morbidity, mortality, fertility and the like. Our concern in this communication is primarily with the topics relevant to clinical practice. They have been presented in the way they arise from a medical context. Thus the style is changed, contents modified and the emphasis altered. Such a change could arouse much-needed interest in the students and demonstrate that statistical methods are also one of their own tools. If only this much could be taught with success, a major part of the objective of developing quantitative thinking and objectivity could be achieved.

The topics that follow are not essential, but are desirable. They do not arise directly from patient handling but have relevance to practice of medicine and could be covered in a more detailed course.

3. Ancillary topics

3.1. *Interpretation of laboratory results*

Subjective observations like signs–symptoms continue to play a vital role, and rightly so, in diagnosis and treatment. Nevertheless, medicine is becoming increasingly dependent on objective traits. These may be simple investigations like pulse rate, blood pressure, haemoglobin level, and blood group, or more complex like biopsies, electrocardiogram, and liver enzymes. In all these cases, the most important aspect is sampling. In the case of a blood specimen, one sample is generally considered enough since blood is thoroughly mixed internally and one sample is likely to yield the same result as another. Note that sampling is the only way to examine blood or, for that matter, urine, sputum and tissues. If the material is not homogeneous, as in the case of biopsy tissues, repeated investigations are sometimes needed to confirm or exclude a diagnosis. Sampling is thus an integral part of medical investigations. In fact, b.p. measurement is also a sample, and sometimes a series of measurements is required on the same subject. Sometimes the sample is obtained from various parts, organs, or systems of the body. The idea again is, how much confidence can we place on an investigation result? This, in turn, depends on the variability or the homogeneity of the material being sampled. Without going into exact quantitative aspects, the above discussion brings out the importance of sampling and its limitations.

3.2. *Critical appraisal of the medical literature*

During practice after graduation, someone in medicine is supposed to keep abreast with current development by going through the published reports. He sometimes finds himself flooded with literature distributed by pharmaceutical companies about the utility and effectiveness of their product. It is essential to be able to critically examine the evidence provided by such reports before the claim is accepted and put to practice.

Most of the pharmaceutical literature relates to the potency or effectiveness of a preparation measured in terms of percentage of a particular type of patient getting relief, the amount and duration of relief, and so on. Fortunately, it now is the rule rather than an exception that these claims are reported in quantitative terms, enabling the reader to make his own judgement in his own context. A statement of the type that effectiveness is 'at least 70% ($P<0{\cdot}10$)' generally accompanies the report. The intriguing quantity is the P-value given as less than 0·10 in the above example. This is the probability of the claim being wrong. The smaller the P-value, the greater is the confidence in the claim. Generally, a risk of 5% of erroneously accepting a claim is considered tolerable. Emphasize to the students that this is only half the story. The decision regarding acceptance or rejection of a regimen really depends on the severity of the condition and on the possible consequences of doing the wrong thing. If the prognosis for a disease in the absence of medication is severe, as in rabies, but not much harm is to be expected by unnecessary medication, it may be desirable to start the treatment even at a risk of more than 5% of 'treating' a normal person. On the other hand, if the therapy involves something like open-heart surgery, this could be done only after being almost certain that the person has the problem requiring the surgery.

The students should be made aware of spurious claims in some cases. The variable under report may really not be valid for the relief claimed. For example, can a preparation demonstrated to reduce the cholesterol level be considered effective in

providing relief from angina pectoris? Or, is a claim of a drug to provide relief from the common cold based actually on the common cold or on its surrogates?

The regimen of a drug is generally decided on the basis of bioassays and clinical trials. These experiments require elaborate planning and a sophisticated statistical analysis. Details of such methods can not be given at the undergraduate level but a mere reference serves to indicate what lies ahead if they plan to get involved in any such activity. Research involvement has several other statistical prerequisites which could be described only briefly at the undergraduate level, as follows.

3.3 *Data handling in medical research*

Medical research could be done either in the clinic, in the laboratory, or in the field. Special techniques of planning and conducting an investigation and of analysis are needed for each setting. The thrust in planning is to keep an eye on the sources of variability—intrinsic in the subjects due to age, sex, ethnicity, pregnancy, socio-economic conditions, etc., and extraneous due to observers, laboratory, instruments, etc. The latter should not be allowed to conceal the former. Various designs like double-blind, cross-over and factorial for clinic-based investigations, like randomized block, nested, and factorial for laboratory-based, and prospective, retrospective and cross-sectional for field studies, are available. Only the names, the meaning, and perhaps one or two relevant examples for each type of design can be given at the undergraduate level. The analysis should be simple so as to include only the description of data in terms of graphs, diagrams, frequency distribution, mean and s.d. for each group, etc. or may be deep to go into the type and strength of relationship, cause and effect, and the like. The latter could be based on any or all of such advanced methods as regression analysis, classification, stochastic processes, time series analysis, survival analysis and pharmacokinetics. Again, because of the mathematical nature of these techniques, no details can be given. The objective is to make the students aware of the existence of such methods and the complexities involved, so that in case the need should arise in his future endeavours, he is ready to seek professional statistical assistance. He then knows when to seek advice and what a statistician may be able to do for him.

4. Conclusion

As originally proposed, the suggestions on the style and content of teaching statistics to medical undergraduates are primarily based on the clinical needs as perceived at the undergraduate level. Besides the epidemiological aspects already mentioned, some other traditional topics like graphical representation of data have been excluded. In place of having a separate identity, the proposed teaching could well be merged with a course on decision-making [13], if such a course exists.

The proposal also excludes the topics needed primarily by the graduate students. The main disciplines requiring special statistical knowledge are pharmacology and preventive medicine/epidemiology. The former needs a detailed discussion of clinical trials, bioassays, bioavailability and pharmacokinetics, and the latter such topics as relative risk, the odds ratio, epidemic models, and life-table methods. We propose to give details of such teaching in a separate communication.

References

[1] KRALL, J. M., HALL, D. S., GARLAND, B. K., and PEARSON, R. J. C., 1983, *J. med. Educ.*, **58,** 815.
[2] Report of the Royal Commission on Medical Education: 1965–68, 1968 (London: Her Majesty's Stationery Office).
[3] COLTON, T., 1975, *J. med. Educ.*, **50,** 596.
[4] O'BRIEN, P. C., and SHAMPO, M. A., 1981, *Mayo Clinic Proc.*, **56,** 47.
[5] RAMACHANDRAN, K., 1982, *Ind. J. Pediat.*, **49,** 383.
[6] FEINSTEIN, A. R., 1970, *Clin. Pharmac. Ther.*, **11,** 135.
[7] GORE, S. M., 1981, *Br. med. J.*, **282,** 1605.
[8] MULVIHILL, M. H., WALLMAN, G., and BLUM, G., 1980, *J. med. Educ.*, **55,** 457.
[9] CLARKE, M., CLAYTON, D. G., and DONALDSON, L. J., 1980, *Int. J. Epidemiol.*, **9,** 179.
[10] MURPHY, J. R., 1980, *J. med. Educ.*, **55,** 27.
[11] FEINSTEIN, A. R., 1974, *Clin. Pharmac. Ther.*, **15,** 528.
[12] HILL, J. G., 1981, *Can. Med. Assoc. J.*, **125,** 819.
[13] MARGOLIS, C. Z., BARNOON, S., and BARAK, N., 1982, *J. med. Educ.*, **57,** 184.