July, 1992

A Letter to the Editor

A Call for Help

I teach a biostatistics course at a four-year undergraduate university. The students taking this course are from the departments of nursing, occupational therapy, medical technology, health sciences and biology. The majority of these students are juniors and have very little mathematical background (often only Algebra I taken 5 to 10 years previously in high school). They are required to take this biostatistics course as a prerequisite for the research course in their own department. The departments want their students to be able to understand and critically evaluate current research, but not to originate statistical analyses. Most information I have received via this Section on Teaching Statistics in the Health Sciences pertains to graduate and/or medical school curricula. We are currently re-evaluating the contents of our course, and I would like to receive suggestions as to what people believe should be covered in a one-semester biostatistics at OUR level here at Towson State. I know that the diversity of departments will likely produce enough core material for a 10-credit course, but nevertheless, I would find these opinions very valuable in restructuring our course. The concerned departments here at Towson State will also be giving me their opinions, but a variety of external opinions is very useful.

Thank you,
Howard Kaplon
Towson State University

Feature Articles

MEDTREE:
An Introduction to Decision Analysis

Peter B. Imrey

Medical Information Science

University of Illinois

Formal decision analysis has become increasingly important in the analysis of clinical practices in the health professions including whether or not to intervene, choice of therapeutic modality, use of diagnostic tests, and screening strategies. This article describes an interactive microcomputer-based self-instructional program which has been used at the University of Illinois to introduce second-year medical students to the rudiments of decision analysis. MEDTREE is implemented on PC-compatibles, requires 261K disk space, and will be available shortly at no charge (see below).

MEDTREE uses a maximum expected utility approach to decision-making, and teaches this through a simple but general example: the decision of whether or not to employ a risky diagnostic test prior to making a therapeutic choice between two alternatives (one of which might be "do nothing"). It is assumed that the test has a small probability of a result with zero utility, e.g. test mortality, as might occur rarely with a coronary angiogram or a liver biopsy, and that expected two year survival proportion is the utility measure. Since no test is totally accurate, the decision analysis incorporates parameters of test accuracy into the calculation of expected utility, and thus

includes instructional material on predictive values of diagnostic tests.

MEDTREE leads the student through the construction of a decision tree which begins with the choice of whether or not to do the test. The tree branches depending on whether the patient survives the test, branches again on the outcome of the test for the survivors, once more on the therapeutic decision, and finally on the true disease state. Utilities for each combination of true disease state and therapeutic decision, as well as for test mortality, are then "folded back" towards the trunk of the tree, using the assumed pre-test probabilities of each disease state, of test-associated mortality, and the test's positive and negative predictive values (based on pre-test probabilities, sensitivity, and specificity).

Although we have used MEDTREE in conjunction with classroom lectures, it is a self-contained and quite flexible instructional module. Beginning with introductory material on the basic ideas of decision analysis (material which can be bypassed by a knowledgeable student), the program constructs a decision tree for the above problem step by step. At various points, students are prompted to interact with the program in this construction. As the tree is built, explanations corresponding to each node, branch, and utility are superimposed over portions of the tree, and removed as the tree grows. Once the tree is mature, the student is encouraged to alter any parameters of the problem, and watch how the utilities at each node, and possibly the appropriate decision, change immediately in response. The student can play with one parameter, or combinations of parameters, to study how sensitive the maximum-utility decision is to the diagnostic test's risk, the specificity and sensitivity of the diagnostic test, and to the pretest probabilities of each diagnosis.

To summarize the conditional relationship of the maximum-utility decision to any single parameter of the problem, with others fixed, the program will prepare a simultaneous graph of the utilities of each of three clinical strategies against varying values of the indicated parameter. The strategies are: i) skip the test, and use therapy 1; ii) skip the test, and use therapy 2; and iii) perform the test, accepting its risk, and choose the appropriate therapy depending on the outcome. The parameters which may be individually varied are the utilities of each disease state-treatment combination, the mortality, sensitivity and specificity of the diagnostic test, and the pretest probability of each disease state. Each plot is marked with vertical lines at values of the parameter where the maximum-utility decision changes, and each region on the plot is labeled with the maximum-utility decision for the corresponding range of parameter values. A student may customize both the decision tree and related graphs to a particular problem by making substitutions for the generic names for disease states and treatments which the program otherwise uses.

When working with a decision tree, the student may access extensive help facilities at any time. Pressing one key after highlighting any node of the tree makes available an explanation of how the utility of that node was obtained by the "folding back" process. The user may choose between a brief explanation, and a fully-detailed explanation including all appropriate formulae with commentary. Definitions are available for all fundamental terms used in the program, such as node, utility, sensitivity, specificity, and predictive values. A user may bypass all explanatory material when entering MEDTREE, and move directly to working with the full decision tree and associated graphs. However, the utility explanations, definitions, and all bypassed introductory and explanatory material are accessible directly from the tree through appropriate Help selections. MEDTREE can be used with specific "lab assignments", or students may just play with the program, to develop intuition about the effects of changing in combinations of parameters on the preference regions for various strategies. On leaving MEDTREE the user is given an opportunity to provide comments, which are stored in a file for later examination at the instructor's convenience.

Arrangements for routine distribution of MEDTREE by computer network are currently being developed, in conjunction with exploring the possibility of a general computer archive of materials that might prove useful to members of this Section. Members who wish to receive MEDTREE when arrangements for distribution are complete may notify me by letter (University of Illinois, 506 S. Mathews Ave., Urbana, IL 61801-3618), phone (217 333-2427), or e-mail (p-imrey@uiuc.edu or imrey@uiucvmd.bitnet). You

will receive either the software, or instructions for obtaining it, when the mechanism for electronic distribution has been worked out. A brief user's guide will be included, although use of the program is virtually self-explanatory. MEDTREE is not supported or warranted, but we have found the program useful, and hope others may also. While several general and powerful programs are available for decision analysis, its simplicity makes MEDTREE exceptionally friendly to the user and appropriate for elementary instruction.

Peter Imrey, Allan Levy, and John Mirowsky are responsible for the content and structure of MEDTREE, which was motivated by Fineberg, H. V., *Bull Cancer* (Paris) 67:395-404. MEDTREE was programmed by William Jockusch and Grant Jenkins in PASCAL. Development was partially supported by a Project Excel grant from the IBM Corporation and the Univ. of Illinois to the Dept. of Medical Information Science, College of Medicine at Urbana-Champaign.

See you at

the Section's

General Business Meeting

6:00 p.m.

Tuesday, August 11

Bayes' Theorem

Robert Elston
Bill Johnson
L.S.U. Medical Center, New Orleans

The following pictorial method of teaching Bayes' theorem is contributed by Robert Elston and Bill Johnson of the L.S.U. Medical Center, New Orleans. It is an extract from the forthcoming second edition of their book "Essentials of Biostatistics" to be published early next year by F.A. Davis, Philadelphia. It is reproduced here with permission of the publishers.

The Englishman Thomas Bayes wrote an essay on probability that he was hesitant to publish because he recognized the flaw in assuming, as he did in his essay, that all possible outcomes are equally likely. The essay was nevertheless published in 1763, after his death, by a friend. What is now called "Bayes' theorem" evolved from this essay and does not contain the original flaw. The theorem gives us a method of calculating new probabilities to take account of new information.

Suppose that twenty percent of a particular population has a certain disease, D. For example, the disease might be hypertension, defined as having an average diastolic blood pressure of 95 mm Hg or greater taken over a period of five days. In figure 4.2 we represent the whole population by a square whose sides are unity. The probability that a person has the disease, $P(D)$, and the probability that a person does not have the disease, $P(\overline{D}) = 1 - P(D)$, are indicated along the bottom axis. Thus the areas of the two rectangles are the same as these two probabilities.

Now suppose we have a test that picks up a particular symptom, S, associated with the disease. In our example, the test might be to take just one reading of the diastolic blood pressure, and S might be defined as this one pressure being 95 mm Hg or greater. Alternatively, we could say that the test result is positive if this one blood pressure is 95 mm Hg or greater, negative otherwise. Before being tested, a random person from the population has a 20% probability of having the disease. How does this probability change if it becomes known that the symptom is present?
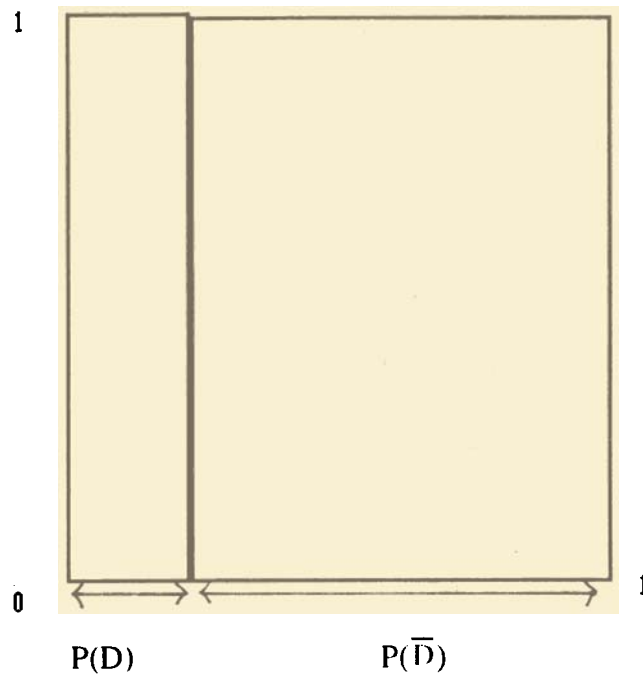
FIGURE 4.2 In the whole population, represented by a square whose sides are unity, the probability of having the disease is P(D), and of not having the disease is P($\overline{D}$), as indicated along the horizontal axis.
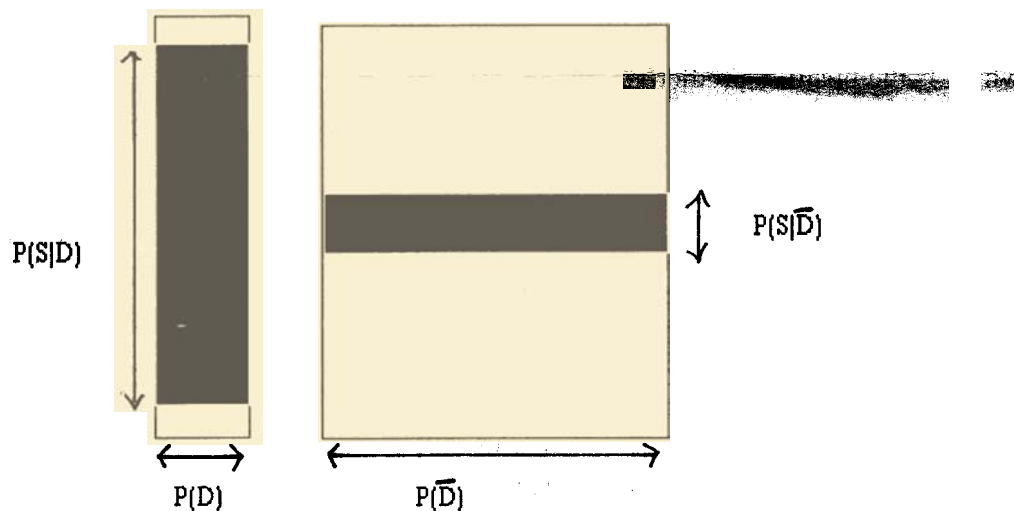

FIGURE 4.3 Within each subpopulation, D and $\overline{D}$, the conditional probability of having a positive test result or symptom, S, is indicated along the vertical axis. The dark rectangles represent the joint probabilities P(D) P(S|D) = P(S,D) and P($\overline{D}$) P(S|$\overline{D}$) = P(S,$\overline{D}$).

Assume that the symptom is present in 90% of all those with the disease but only 10% of all those without the disease, i.e. P(S|D) = 0.9 and P(S|$\overline{D}$) = 0.1. In other words, the sensitivity and the specificity of the test are both 0.9. These conditional probabilities are indicated along the vertical axis in Figure 4.3. The hatched rectangles represent the joint probabilities that the symptom is present and that the disease is present or not:

$$P(S \text{ and } D) = P(D) \, P(S|D) = (.2)\,(.9) = .18$$
$$P(S \text{ and } \overline{D}) = P(\overline{D}) \, P(S|\overline{D}) = (.8)\,(.1)\ .08.$$

If we know that the symptom is present, then we know that only the dark areas are relevant, i.e. we can write symbolically,

$$P(D|S) = P(D| \quad) =$$
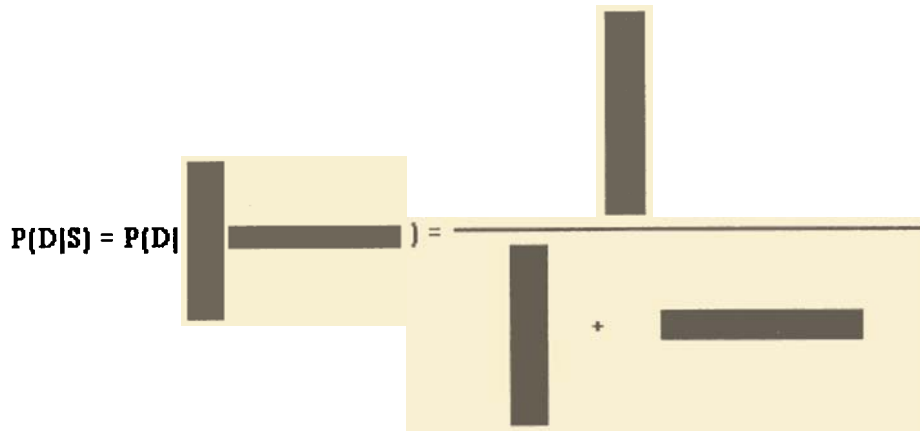
$$\frac{P(S \text{ and } D)}{P(S \text{ and } D) + P(S \text{ and } \overline{D})} = \frac{P(D)P(S|D)}{P(D)P(S|D) + P(\overline{D})P(S|\overline{D})}$$

$$= \frac{(0.2)(0.9)}{(0.2)(0.9) + (0.8)(0.1)} = 0.69 ,$$

which is the positive predictive value of the test.

This, in essence, is Bayes' theorem. We start with a **prior** probability of the disease, P(D), which is then converted into a **posterior** probability, P(D|S), given the new knowledge that symptom S is present.

More generally, we can give the theorem as follows. Let the new information that is available be that the event S occurred. Now suppose the event S can occur in any one of k distinct, mutually exclusive ways. Call these ways $D_1, D_2, \ldots, D_k$ (in the above example there were just two ways, the person either had the disease or did not have the disease; in general there may be k alternative diagnoses possible). Suppose that with no knowledge about S these have prior probabilities $P(D_1)$, $P(D_2), \ldots$, and $P(D_k)$, respectively. Then the theorem states that the **posterior** probability of a particular D, say D_j, conditional on S having occurred, is

$$P(D_j|S) = \frac{P(D_j)P(S|D_j)}{P(D_1)P(S|D_1) + P(D_2)P(S|D_2) + \ldots + P(D_k)P(S|D_k)}$$

$$\frac{P(D_j \text{ and } S)}{P(D_1 \text{ and } S) + P(D_2 \text{ and } S) + \ldots + P(D_k \text{ and } S)}$$

The theorem can thus be remembered as "the joint probability divided by the sum of the joint probabilities" (i.e. the posterior probability of a particular D, given that S has occurred, is equal to the joint probability of D and S occurring, divided by the sum of the joint probabilities of each of the D's and S occurring). This is illustrated in Figure 4.4.
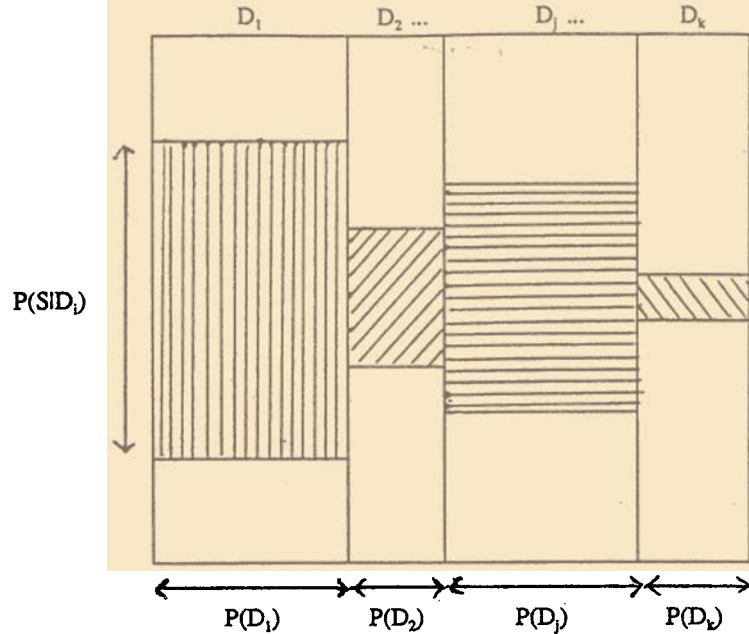
FIGURE 4.4. Bayes' theorem. The probability of various diagnoses, D_1, D_2. . . D_j,... D_k, are indicated on the horizontal axis and the conditional probability of a particular symptom, within each diagnostic class D_1, is indicated on the vertical axis. Thus each hatched rectangle is the joint probability of the symptom and each diagnostic class.

Steven Verhulst
Southern Illinois University
School of Medicine
Division of
Statistics and Research Consulting
P.O. Box 19230
Springfield, Illinois
62794-9230
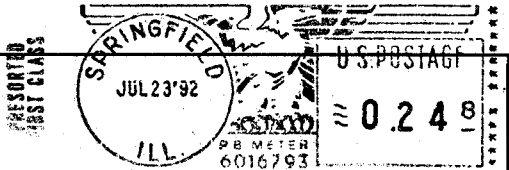
TEACHING OF STATISTICS IN THE HEALTH SCIENCES