FROM TESTING TO DECISION MAKING: CHANGING HOW WE TEACH STATISTICS TO HEALTH CARE PROFESSIONALS

Dalene Stangl
Duke University
USA

In their first and often only statistics course, health-care professionals are taught Bayes' theorem in the context of diagnostic testing. They learn the concepts of sensitivity/specificity and predictive value positive/negative and how Bayes' theorem can assist in diagnostic decision-making. Then the class moves on often spending weeks on tests of significance. This paper will argue for changing this practice, and instead focusing such courses on statistics for decision-making beyond diagnostic testing. It will argue that such changes will make our health-care professionals better consumers of statistical information and better decision makers.

INTRODUCTION

The statistical training of medical researchers usually begins with an introductory undergraduate statistics course, which is perhaps followed by a course or two while in nursing, graduate or medical school. Any other statistical education comes from 'on-the-job' training and continuing education courses. Laake (1998) and Phillips et al. (1998) discuss teaching statistics to professionals in health-related sciences. Typically courses cover a mixture of statistical analysis and research design. While nearly all courses spend weeks on hypothesis testing, few courses address the use of statistics directly in decision-making. The lucky student may be taught Bayes' theorem in the context of diagnostic testing. They learn the concepts of sensitivity/specificity and predictive value positive/negative and how Bayes' theorem can assist in making diagnoses. However, this is where using statistics to aid decision-making starts and ends in these classes. This paper will argue for changing this practice by focusing such courses on statistics for decision-making.

The work of physicians, nurses and other health-care professionals is a sequence of decision-making. I asked a group of health-care professionals to list decisions they make on a daily basis. Combining and generalizing their responses resulted in the following list. Does this patient have illness A or B? Should I order a diagnostic test to rule out illness C? How reliably is this patient able to give a medical history? If the patient has illness B, should I treat with medicine X or Y? What is the efficacy of new medicine Z relative to medicines X and Y? What side effects are prominent enough that the patient should be warned about the possibility of their appearance? Should a specialist see this patient? What are the chances this patient will survive? What is this patient's long-term prognosis if he/she does survive? When should this patient come back for a follow-up visit?

How many of these questions can be answered by testing a hypothesis? The answer is not many. So why is it that in teaching introductory statistics we persist spending 1/2 or more of our time teaching tests of hypothesis? In Rosner's *Fundamental's of Biostatistics* (2000), Pagano and Gauvreau's *Principles of Biostatistics* (2000), and Zar's *Biostatistical Analysis* (1999) the index entry for hypothesis testing shows the entries presented in Table 1.

Table 1
Entries under Hypothesis Testing in 3 Favorite Biostatistics Textbooks

Rosner	Pagano and Gauvreau	Zar
for clustered binary data	alternative hypothesis	See also specific hypotheses
multiple linear regression	comparison of two proportions	& tails
multiple logistic regression	concepts	
one-sample inference	confidence intervals and	
one-way analysis of variance	errors type I and II	
person-time data methods	mean	
relationship to confidence	normal approximation to	
intervals	binomial distribution	
sample problem involving	null hypothesis	
stratified person-time data	one-sided tests	
methods	power	
two-sample inference	proportion	
two-sample inference for	*p*-value	
incidence-rate data	significance level	
two-way analysis of variance	two-sided tests	
for zero and non-zero null		
correlations		

Each of these books has multiple chapters on hypothesis testing, and many other chapters with sections on hypothesis testing within particular methodologies such as ANOVA and regression. When we search these texts for the use of statistics for decision-making we see something quite different.

Table 2
Entries under Bayes' or Decision-Making in 3 Favorite Biostatistics Textbooks

Rosner	Pagano and Gauvreau	Zar
Bayes' theorem	Bayes' theorem	No entries
Decision rule for meta-analysis	in diagnostic testing	

In Rosner (2000) the index entry for decision anything is a single reference to "Decision rule for meta-analysis." If one looks-up this reference, one finds Rosner teaching the use of hypothesis testing for deciding between fixed and random-effect models. This is a commonly taught technique for meta-analysis. The books by Pagano and Gauvreau (2000) and by Zar (1999) don't have entries for decision anything, although Pagano and Gauvreau do teach the utility of Bayes' theorem in diagnostic testing.

The most important gap in current statistical education arises from the fact that while decision-making is the incentive for most clinical and research efforts, the decision process usually remains informal and ad hoc. The statistician's role has been to provide data summaries of empirical evidence, while determining how information is combined across sources and should be used is left to the clinician and researcher. This disjuncture between statistical synthesis and decision-making is an unnatural and undesirable one, because it undermines the impact of quantitative information. Adopting a Bayesian statistical perspective that coherently incorporates decision-theoretic methods provides a natural bridge for this gap.

But how does one teach Bayesian methods to persons with little mathematical training? There has been much discussion of didactical problems in teaching Bayesian inference to undergraduates. (Albert, 1997; Berry, 1997; Moore, 1997). Moore argues that it is, at best, premature to teach the ideas and methods of Bayesian inference in a first statistics course for general students. He argues that: 1) Bayesian techniques are little used, 2) Bayesians have not yet agreed on standard approaches to standard problem settings, 3) Bayesian reasoning requires a grasp of conditional probability, a concept confusing to beginners, and 4) an emphasis on

Bayesian inference might impede the trend toward experience with real data and a better balance among data analysis, data production, and inference in first statistics courses.

Similar arguments could be made for not teaching Bayesian methods to medical researchers; however, these arguments are as specious for the education of health-care personnel as for undergraduates. (Stangl, 1998) Teaching statistics in a way that undermines the impact of quantitative information is as irresponsible as failing to teach the fact that there is no universal solution to the problem of inductive inference.

A COMPARISON

When we do not teach students to link statistical methods and decision-making, students inappropriately use tests of significance for decision rules even when cautioned against doing so. They see no other option! Last semester I ran an experiment to compare how students use statistics for decision making after an introductory course. In Class 1, I taught only Frequentist inference using the popular textbook *Statistics*, by Freedman, Pisani and Purves (1997). In Class 2, I taught Frequentist inference using the same textbook, but I also taught Bayesian inference using my own supplemental material. During the Bayesian inference component students learned about prior, posterior, and predictive distributions, and how to incorporate predictive distributions into a decision analysis.

As part of the final exam, students were given the article, "Neurological Dysfunction in Children with Chronic Low-Level Lead Absorption," by Landrigan (1975). The article appeared in *The Lancet* in 1975. This article investigates the relationship between low-level lead absorption and neuropsychological function in 124 children living within 6.6 kilometers of a large lead-emitting smelter. The article was chosen because it cuts across natural, social and health sciences and it uses only statistical methods that are introduced in my introductory statistics course. Students were asked to read the article and then calculate some simple test statistics and relative risks based on information given in the article. Students were also asked to comment on age and gender as potential confounding variables and comment on changes in results depending on whether particular subgroups were included in the analysis. Finally students were asked whether the findings in the paper warranted a public-health warning on the hazards of living within 6.6 kilometers of a lead-emitting smelter and were asked to provide a rationale for their decision using statistical evidence from the paper. Because the policy question was only tangentially related to the research question and statistical results of the paper (both high and low lead level groups lived within 6.6 kilometers of the smelter) answers clearly distinguished between students that understood and could critically evaluate the paper's statistical content.

What this comparison showed was that students in the two classes interpreted both the question asked and the statistical results from this paper quite differently. The primary differences can be grouped as 1) perception of the question, 2) use of prior information, 3) use of statistics presented, and 4) critique of study design.

Students who had been exposed to the Bayesian paradigm perceived the question as a policy decision with many facets whereas students who were taught only the Frequentist paradigm perceived the question as one of statistical significance. The former group most frequently concluded that there was not enough relevant information provided by the Landrigan (1975) study to warrant a public warning, while the latter group most frequently concluded that the dozen or so statistically significant test results did warrant a public warning.

Students who had been exposed to the Bayesian paradigm more often wanted information from previous studies than students that were not exposed to the Bayesian paradigm. They saw this information, especially an article that used geographical proximity to the smelter rather than blood-lead levels, as relevant to the decision at hand. Meanwhile students exposed only to the Frequentist paradigm accepted without question Landrigan et al's (1975) claim that the design of previous studies were weaker. For Landrigan's research question, "*What was the impact of blood-lead level on neuropsychological dysfunction in children?*" the geographic proximity study was of less relevance, but for the policy question presented to students, the geographic proximity study was of more value than the Landrigan study.

Students who had been exposed to the Bayesian paradigm rarely used the p-values from tests of null hypotheses (no difference between the two groups of children) as decision rules,

whereas the students exposed only to Frequentist methods did so with abandon. The former students wanted posterior distributions for the difference between groups. They wanted predictive distributions for future observations. They wanted time to think about the costs and benefits of making one decision over another. Students exposed only to Frequentist methods sometimes requested confidence intervals but overwhelmingly concluded that the dozen or so statistically significant test results demonstrated that we need to protect our children and provide the public warning.

While both groups could find fault with the Landrigan (1975) study, those exposed to the Bayesian paradigm were far less likely to take claims by the authors as fact. They understood research as an exercise full of subjective judgments. Hence they took the results with far greater skepticism. They were far more likely to mention the problems with multiplicities in testing, the potential confounding of age and gender, and the need to carefully consider the costs and benefits of decisions in both directions. In the Frequentist only group, these issues were rarely mentioned. My hunch is that some of these issues do not really make sense to students until the Frequentist paradigm is juxtaposed to the Bayesian one. It is only then that sampling distributions and p-values can be understood as different from posterior and predictive probabilities. It is only then that students really grasp that a p-value is not the probability that the null hypothesis is true.

In summary, the students who had been exposed to the Bayesian paradigm were better consumers of the statistical analysis presented in the Landrigan (1975) paper. They read it more critically, they viewed it as subjective rather than factual empirical evidence, they were better able to focus on the decision question, and they better understood what tests-of-significance do and do not tell us.

A CASE STUDY FOR TEACHING

How does one teach Bayesian inference to health-care professionals? There is a good introductory textbook by Berry (1996), but for those who don't want to teach from primarily a Bayesian perspective, one is left to devise their own course materials. I use a supplement written by Michael Lavine and myself to accompany Freedman, Pisani & Purves (1997) introductory textbook, *Statistics*. The supplement teaches probability as subjective belief, parameters as random quantities, and that the goal of statistical analysis is to enhance one's decision-making. Students are taught about prior and posterior distributions for parameters and about predictive distributions for future observations. Students first work through the calculation for discrete parameter spaces and then are shown how Bayes' works in continuous-parameter spaces using beta-binomial and normal-normal models. In addition to this supplement, students are exposed to a case study.

The case study (Stangl, 2001) involves the GUSTO clinical trial, a trial comparing tissue plasminogen activator (t-PA) and streptokinase (SK) for the treatment of myocardial infarction. The results of the trial were first presented in the *New England Journal of Medicine*, (The Gusto Investigators, 1993) and were subsequently reanalyzed by Brophy and Joseph in the *Journal of the American Medical Association* (1995*)*. The statistical argument in the *NEJM* paper uses confidence intervals and tests of significance. Finding an increased survival of 1% and rejecting the null hypothesis of no difference between treatments, the GUSTO investigators conclude that t-PA is clinically superior. In the *JAMA* paper, Brophy and Joseph use Bayesian statistical arguments to argue that the jury is still out. They find that the posterior probability that survival on t-PA is greater than survival on SK by at least 1% ranges from 0% to 36% depending on how much weight is placed on previous trials. A third source for the case is an article, "The Mathematics of Making up Your Mind", by W. Hively (1996). The article appeared in the popular science magazine *Discover* in May 1996. It covers the differences between inferential paradigms and highlights the controversies that can arise between them. The article uses the GUSTO trial as their primary example.

After introducing and discussing the case, there are two student exercises both based on role-playing. One is a written exercise, the other a mock legal trial. Students are expected to use the information from the three articles. In the written role-playing exercise students are asked to role-play 3 individuals: 1) a government policy maker deciding whether Medicare will pay for t-PA, the more expensive treatment, 2) an insurance company deciding whether their company will

pay for the more expensive drug, and 3) a son/daughter who's parent was given the more expensive drug, and the insurance company is refusing to pay. They must present a written statistical argument (Bayesian or Frequentist) to defend each position.

In the second role-playing exercise, a mock legal trial, students are given roles of plaintiff, defendant, prosecuting attorney, defense attorney, or expert (statistical) witness (one for each side). Students must role-play a malpractice suit against a doctor who prescribes the cheaper drug (SK) and the patient dies.

The case study is an excellent way of teaching Bayesian methods to health-care professionals ranging from future nurses to old-hand pharmaceutical statisticians. It brings to light how differently questions are addressed and statistical results are interpreted between the Bayesian and Frequentist paradigms. A PowerPoint presentation of the case along with an audio explanation is available at www.stat.duke.edu/~dalene. Free software, Real Player, is required to view the presentation.

CONCLUSIONS

Health care professionals use empirical information via statistical summaries to make decisions. Current statistical education of these individuals falls short, because classroom time is overspent on hypothesis testing rather than statistics for decision-making. This strategy undermines the impact of quantitative information and can lead to incoherent decision-making. This paper proposes an alternative and provides resources for initiating change. It argues that health-care professionals should be taught Bayesian methods of inference and decision-making. It provides information on a supplement and case study for doing so.

REFERENCES

Albert, J. (1997). Teaching Bayes' rule: A data-oriented approach. *The American Statistician*, *51,* 247-253.

Berry, D.A. (1997). Teaching elementary Bayesian statistics with real applications in science. *The American Statistician*, *51*, 241-246.

Berry, D.A. (1996). *Statistics: A Bayesian perspective*. Duxbury Press.

Brophy J.M., & Joseph, L. (1995). Placing trials in context using Bayesian analysis: Gusto revisited by Reverend Bayes. *Journal of the American Medical Association, 273*, 871-875.

Freedman, D., Pisani, R., & Purves, R. (1997). *Statistics* (3rd edn.). Norton.

Hively, W. (1996). The mathematics of making up your mind. *Discover*, *May*, 90-97.

Landrigan, P.J., Whitworth, R.H., Baloh, R.W., Staehling, N.W., Barthel, W.F., & Rosenblum, B.F. (1975). Neurological dysfunction in children with chronic low-level lead absorption," *The Lancet*, March, 708-712.

Moore, D.S. (1997). Bayes' for beginners? Some reason to hesitate. *The American Statistician*, *51*, 254-261.

Pagano M., & Gauvreau K. (2000). *Principles of Biostatistics*. California: Brooks/Cole, Duxbury Thomson Learning.

Rosner, B. (2000). *Fundamentals of Biostatistics*. Kentucky: Duxbury Press.

Stangl, D., (1998). Classical and Bayesian paradigms: Can we teach both. In L. Pereira-Mendoza, L.S. Kea, T.W. Kee, and W. Wong (Eds.). *Proceedings of the 5th International Conference on Teaching Statistics (Vol 1)* (pp. 251-258). International Statistics Institute.

Stangl, D. (2001). A case study for teaching Bayesian methods. In *Proceedings of JSM 2001,* ASA Section on Education.

The GUSTO Investigators (1993). An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *New England Journal of Medicine, 329*, 673-682.

Zar, J.H. (1999). *Biostatistical Analysis*. New Jersey: Prentice Hall.