Bayesian Analysis: A New Statistical Paradigm for New Technology

Gary L. Grunkemeier, PhD, and Nicola Payne, MPhil

Providence Health System, Portland, Oregon

Full Bayesian analysis is an alternative statistical paradigm, as opposed to traditionally used methods, usually called frequentist statistics. Bayesian analysis is controversial because it requires assuming a prior distribution, which can be arbitrarily chosen; thus there is a subjective element, which is considered to be a major weakness. However, this could also be considered a strength since it provides a formal way of incorporating prior knowledge. Since it is flexible and permits repeated looks at evolving data, Bayesian analysis is particularly well suited to the evaluation of new medical technology. Bayesian analysis can refer to a range of things: from a simple, noncontroversial formula for inverting probabilities to an alternative approach to the philosophy of science. Its advantages include: (1) providing direct probability statements—which are what most people wrongly assume they are getting from conventional statistics; (2) formally incorporating previous information in statistical inference of a data set, a natural approach which we follow in everyday reasoning; and (3) flexible, adaptive research designs allowing multiple looks at accumulating study data. Its primary disadvantage is the element of subjectivity which some think is not scientific. We discuss and compare frequentist and Bayesian approaches and provide three examples of Bayesian analysis: (1) EKG interpretation, (2) a coin-tossing experiment, and (3) assessing the thromboembolic risk of a new mechanical heart valve.

(Ann Thorac Surg 2002;74:1901–8)

"The Bayesian paradigm is conceptually simple, intuitively plausible, and probabilistically elegant" [1].

Readers of this journal are aware of the importance of statistics in the description and interpretation of surgical results. Standard statistical methods, and the philosophy of statistics that supports them, are variously referred to as classic, conventional, traditional, frequentist, or sampling theory statistics—when it is necessary to distinguish them from the less well-known alternative called Bayesian statistics. This competing philosophy and

See also page 2165

set of analysis tools is named after a man and his theorem that consider probability subjective and allow preexisting external information to formally contribute to the interpretation of data. The first use of the word "new" in the title of this paper is a bit ironic, since Bayes' paper was published (posthumously) in 1763 [2], while the "classic" statistics arose from the works of Fisher [3] and Neyman and Pearson [4] in the 1920's.

In both the traditional (frequentist) and the Bayesian statistical paradigms, we are concerned with estimating values called parameters. They are unknown, and our research studies are designed to collect data to help estimate them. Simple examples, considered in more detail below, are the probability that a patient has coronary artery disease (CAD); the probability a tossed coin will fall heads; and the thromboembolism rate with a new mechanical heart valve. In traditional statistics, these parameters are fixed, constant values, but in Bayesian analysis they are random variables. Unlike a fixed parameter, which can take only a single value, a random variable is characterized by a probability distribution over the range of values the parameter can take. This seemingly simple distinction leads to quite divergent theories of analysis.

EKG Interpretation

Bayes' theorem is simply a formula that inverts conditional probability statements (Appendix 1). For example, if we know the probability that a person with CAD will have a positive electrocardiogram (EKG), then what is the probability that a person who tests positive on an EKG has CAD? They are not the same, and Bayes' theorem is used to derive the latter from the former. Cardiology diagnosticians have been using this approach for many years [5, 6].

Electrocardiogram tests are not perfect, so we acknowledge that when a patient tests positive, there is a probability that the patient does (A) or does not (B) have CAD. Similarly, when a patient tests negative, there is another set of probabilities that the patient does (C) or does not

Address reprint requests to Dr Grunkemeier, East Pavilion MOB, Suite 33, 9155 SW Barnes, Portland, OR 97225; e-mail: ggrunkemeier@providence.org.

Published by Elsevier Science Inc

0003-4975/02/$22.00
PII S0003-4975(02)04535-6

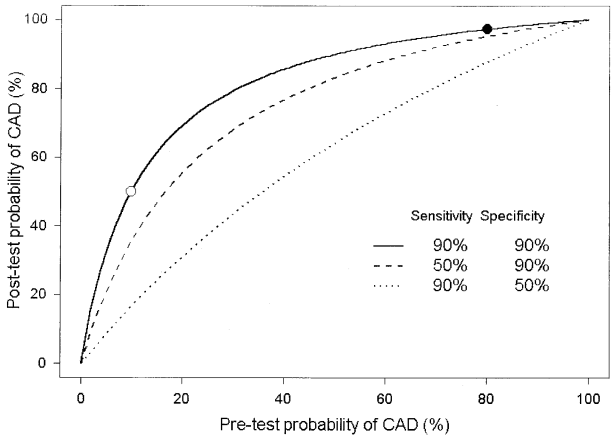

Fig 1. The relationship between post-test (vertical axis) *and pretest* (horizontal axis) *probabilities of having coronary artery disease* (CAD), *after a positive EKG test. Bayes' theorem is used to compute the posttest probabilities for different values of test sensitivity and specificity. The two symbols represent a typical high-risk patient* (solid circle) *and low-risk patient* (open circle), *both of whom tested positive.*

(D) have CAD. These four situations exhaust the possibilities. Sensitivity of the test is the percentage of CAD patients who test positive: (A/(A+C)); specificity of the test is the percentage of non-CAD patients who test negative: (D/(B+D)). Given these values, when a certain patient tests positive, what is the probability that the patient has CAD? This post-test probability can be computed using Bayes theorem (Appendix 1), and depends on the patient's pretest probability of CAD, that is, the prevalence of CAD in such patients. Figure 1 shows posttest probabilities for the entire range of pretest probabilities, for three combinations of specificity and sensitivity.

Suppose the sensitivity and specificity for a test are both 90% (the top curve in Fig 1). Now suppose 2 patients both test positive for CAD on EKG: (1) a 75-year-old male smoker who presented with chest pain and (2) a 35-year-old nonsmoking female found positive while undergoing a mandatory commercial airline pilot exam. If the pretest probability of CAD is 80% for the first patient and 10% for the second, then their probabilities of having CAD after testing positive on the EKG (Appendix 1) are 97% and 50%, respectively (Fig 1).

In this example, we used Bayes' theorem to combine external information (the pretest or prior probability of CAD) with the observed data (positive EKG), to estimate posttest or posterior probabilities of CAD. This did not illustrate any subjective or controversial aspects. Next we consider an example where the prior probabilities, the analog of the fixed, pretest probabilities of CAD, are unknown and subject to interpretation. This is the essence of Bayesian analysis.

Coin-Tossing Experiment

Probability theory originated to solve questions related to gambling, and games of chance still provide instructive examples. Suppose we just acquired a newly-minted Oregon quarter and want to know, before using it for gambling purposes, if it is "fair" (50% of its tosses will fall heads) or unfair (biased). So we toss this coin 10 times, and get 8 heads.

Frequentist Approach

Typically we are interested in showing a change or difference, so we pick a null hypothesis—in this case, that the coin is fair—which we hope to reject on the basis of our study data. The binomial distribution gives the probabilities of getting various numbers of heads. If the true probability is 50%, then the probability of getting 8 heads is only 4.4%. You might think that we reject the null hypothesis (at the 5% level) on this basis, but we must include the probability of results that are more extreme. The probability of 9 heads is 1.0% and of 10 heads is 0.1%, so we get a cumulative p-value of 5.5%. Moreover, we usually add in the other, equally extreme, end of the distribution—that is, the probabilities for getting 0, 1, or 2 heads—giving 11.0% for the final "p-value". So our coin is NOT (shown to be) biased, based on this exercise, which used scenarios that could have occurred, but did not.

However, if we had made 100 tosses and gotten 80 heads (same coin, still 80% heads), then the coin IS biased: the difference from 50% is "highly significant" ($p<0.000000001$). What is the chance that with only 10 tosses we would reach significance if the coin were biased? That is called the power of the test, and for 10 tosses it is only 38%, even if the true probability of heads were 80%. So, even though we cannot reject the null hypothesis, we should not accept it, either. This poses a dilemma, and because of this issue, it is currently recommended to emphasize confidence intervals that intrinsically incorporate the sample sizes, rather than hypothesis tests.

In this experiment, the single estimate of 80% is a point estimate. Based on only 10 tosses, this point estimate has low precision. We quantify this imprecision by using an interval estimate or confidence interval (CI). Many methods have been proposed to compute CI's for binomial percentages (Appendix 2). Four choices are shown in Table 1. (One reason for including several CI methods that yield different answers is to show another element of subjectivity in classic analysis.) When the confidence limit does not include 50.0%, a hypothesis test would reject the null hypothesis (the true probability of heads is 50%) and declare the coin biased. This is the case with two of the CI's, but not so with the other two. The "exact" test uses the binomial distribution, which corresponds to the p-value approach used above. A confidence interval contains the point estimate and gets narrower as the number of tosses grows. In a way, it contains the range of values that are consistent with the data we have observed.

Bayesian Approach

We do not really believe that 80% is our best, final estimate of the probability of getting heads. We do not

Table 1. Frequentist Confidence Intervals and Bayesian Probability Intervals for the Probability of Heads in the Coin-Tossing Experiment

		Lower	Mean	Upper
Frequentist	**Interval Method**			
	Normal	55	80	105
	Exact	44	80	98
	Wilson	49	80	94
	Likelihood ratio	50	80	96
Bayesian	**Prior Distribution**			
	Concentrated	45	54	63
	Diffuse	47	65	81
	Uniform	53	75	92

have any other data with this particular coin, but our experience with other coins tells us that it would be very unusual to have an 80% probability of getting heads. Can better use be made of these data?

Bayes' methodology provides a formal way of incorporating prior experience or opinions into the analysis. That is a definite advantage, but the price paid is that we must quantify this prior knowledge, and that can be a very subjective exercise. That is the criticism of the theory. I might pick one distribution, you another, both with good and justifiable (to ourselves) reasons. Two steps are involved: (1) select a 'prior' distribution for the quantity of interest, then (2) use the observed data to update this distribution, converting it to a 'posterior' distribution. This is equivalent to using the finding of a positive EKG to convert a patient's pre-test probability of CAD to a post-test probability.

DETERMINE THE PRIOR DISTRIBUTION. We recognize that all coins have a certain similarity and feel compelled to formally incorporate this knowledge into our estimating. How? By quantifying this knowledge, in the form of a prior distribution. Three possibilities are shown in Figure 2, as probability density functions from the beta distribution (Appendix 3).

Uniform. A very conservative choice: the probability of heads could be anything between 0% and 100% with equal probability. This is called the uniform distribution because the probability is evenly distributed along all possible outcomes. This opinion is reflected by the thinnest line in Figure 2; it has 90% of the probability of heads between 5% and 95% (shown by vertical lines).

Diffuse. We think the true probability is 50% and are "pretty sure" that it is between 25% and 75%. This opinion is shown by the thicker curve in Figure 2, which has 90% of the probability between 25% and 75%.

Concentrated. We think the probability is 50%, and would be surprised if it was outside 40%–60%. This opinion is shown by the thickest curve in Figure 2, which has 90% of the prior probability between 40% and 60%.

COMBINE THE OBSERVED DATA WITH THE PRIOR DISTRIBUTION. Bayes' theorem is invoked to combine the observed data (8 heads in 10 tosses) and the prior probability distributions to produce posterior probability distributions (Appendix 3). Figure 3 shows the posterior distributions for each of the priors in Figure 2. The mean of each of these distributions (vertical lines) is between the mean of the priors (all 50%) and the observed data (80%). All three of these estimates are more intuitively appealing than the frequentist estimate of 80%. The less concentrated the prior distribution, the more influence the observed data have on the posterior distribution. The most concentrated prior distribution (arguably the most reasonable) with a posterior mean of 54% is influenced least by the data (Fig 3; Table 1).

Even the least concentrated (uniform) prior distribution reduces the point estimate, from 80% to 75% (Fig 3; Table 1). For the uniform prior case, the Bayes' estimate is particularly easy to compute. For any combination of H heads in N tosses, it is given by the ratio (H+1)/(N+2), which in our example of H=8 and N=10 equals 9/12 = 75%. This seems counterintuitive; we observe 8 heads in 10 tosses, yet choose the point estimate 9/12 instead of

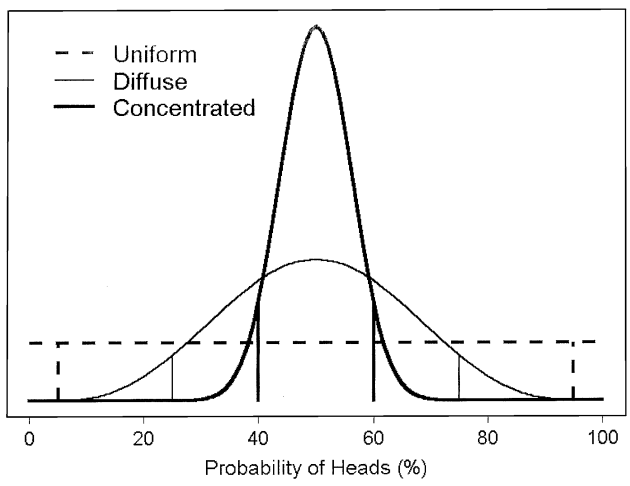

Fig 2. Three possible distributions (density functions) for the probability of getting heads in a coin-tossing experiment. The distributions represent varying degrees of belief about this probability before the experiment starts and are called "prior distributions." The vertical lines *indicate the 5% and 95% quantiles of each distribution.*

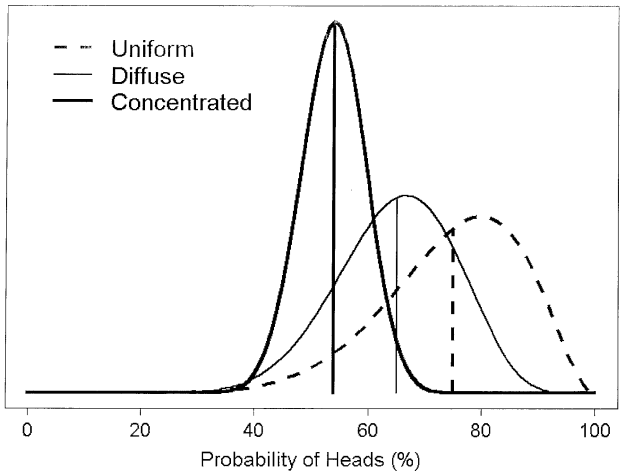

Fig 3. The posterior probabilities of heads in the coin-tossing experiment given 8 heads were observed in 10 tosses. The 3 distributions correspond to the 3 prior distributions plotted in Figure 2. The vertical lines *indicate the mean of each distribution: all 3 means are less than 80%, which is the traditional estimate.*

the usual 8/10? We can see the consequences of this choice by a computerized simulation of the coin tossing experiment. (1) We randomly generated 1000 different values of P uniformly distributed between 0% and 100%. (2) For each P we simulated 10 tosses of a coin with the probability P of getting heads. (3) Using the number of heads H generated in step 2, we computed two estimates of P: the frequentist estimate H/10 and the Bayesian estimate: (H+1)/12. Figure 4 compares these two estimates to the true probability P. It can be seen that the Bayes estimates on the right are in general closer to the diagonal line (of identity) than the frequentist estimates on the left.

As the observed data increases in amount, it dominates the prior. If we had 100 tosses with 80 heads, the posterior distributions in Figure 3 would be closer to each other, as well as closer to 80%.

Confidence Intervals Are Not Credible

We discussed the point estimates from the Bayes' approach, but what about the interval estimates? Table 1 also contains these Bayes estimates, called "credible intervals" or "probability intervals". Most physicians think that a 95% confidence interval has a meaning that can be described by this (direct) statement (A): "The probability the true mean value is in the interval is 95%." But, the true meaning is more convoluted. Frequentists assume that the true mean value is fixed, not random, so a probability statement about it has no meaning. Instead, they can only claim (indirectly) that (B): "If this experiment were repeated many times, 95% of the confidence intervals would contain the mean value."

Bayes' methods, on the other hand, consider that such parameters are random variables, so probability statement A is exactly what Bayesians can claim about their intervals. It could be argued that since most physicians use statement A to describe "confidence" intervals, what they really want are "probability" intervals. Since to get them they must use Bayesian methods, then they are really Bayesians at heart!

Evaluating New Technology

Analysis of new medical technology presents an ideal opportunity for Bayesian analysis, as summarized in a recent review article on its use in health technology assessment [7], and it is being actively promoted by the FDA. For the past several years, FDA has hosted meetings and workshops to discuss Bayesian approaches to new marketing approvals, given presentations at external meetings [8, 9], and supported the Bayesian research work of others [10].

The Bayesian approach to statistical analysis is more adaptive and less rigid than the traditional frequentist approach. The difference is quite apparent in the approach to analysis of clinical studies in which information is accumulated in an ongoing fashion. The traditional approach is to pick a sample size using estimates of anticipated performance; conduct the study on that number of patients; analyze the data at the end of the study; and either accept or reject the null hypothesis based on the p-value of the pivotal test statistic. Some variations of this allow for interim analyses of the data and the possibility of prematurely rejecting the null hypothesis. But these interim looks must be built into the original design, and require additional patients to be entered to allow for the extra hypothesis tests. It is important not to violate this rigid study protocol in order to protect the value of the p-value, to keep it from exceeding its nominal size. This is said to be an objective design, and

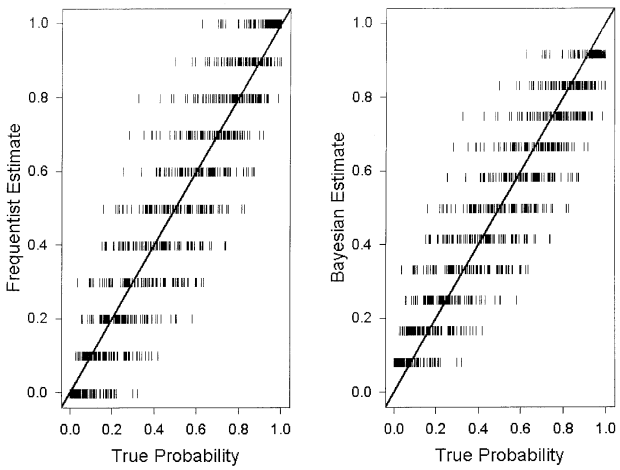

Fig 4. Scatter diagrams of the true probabilities (horizontal axes) *and the estimates of those probabilities* (vertical axes) *using two methods of estimation. On the left are the usual (frequentist) estimates (H/10) for H heads in 10 tosses. On the right are the Bayesian estimates ((H+1)/12) based on a uniform prior distribution. The* diagonal lines *represent perfect agreement between the true probability and its statistical estimate. The Bayesian method is seen to have better agreement.*

that only the data collected in the study is allowed to influence the conclusion of the study. But in reality, information outside of the study is utilized, since we need to supply guesses about the mean and variance of the treatment[s] being studied in order to estimate the required sample size.

The FDA often requires a clinical study to have an independent data monitoring board, comprised of persons not directly connected with the study. Their job is to review the data periodically, and determine whether the study should be stopped because of safety issues. It is ethically necessary to stop a study if it is revealed to be using a harmful therapy. Another reason for stopping a study prematurely could be that the new therapy has already declared itself to be so superior to the control that the remainder of the study patients could not possibly overturn that conclusion. A final reason for premature stopping could be that there is no difference between the treatments and that continuing the study until the prefixed number of patients have been studied could not possibly change this conclusion. The remainder of the study would thus be using resources and patients to provide no useful information. For the monitoring board to fulfill its obligation, they must have access to the data and be able to assess it statistically. But these periodic peeks at the data invalidate the formal study design, since they are not protected by provisions for interim analyses. Thus there is a fundamental incongruence between protecting the p-value and protecting the patients.

How to reconcile this issue? Bayesian analysis is perfectly suited to repeated, interim analyses. It does not require fixed prior sample sizes and one-time assessment, but can utilize accumulating information that becomes available as the study progresses, as the following example demonstrates.

New Mechanical Heart Valve

Assume we want to clinically study a new heart valve, and take one endpoint, eg, thromboembolism (TE). The Poisson distribution is used to describe and summarize event rates (Appendix 3). We are interested in estimating the event rate with the new valve and comparing it to an established standard. Just as with coin-tossing, we have previous experience with prosthetic heart valves. Thus, for example, although the TE rate with a new heart valve could theoretically be anything from 1% per year to 100% per year (or even higher), we know that it is more likely to be the former than the latter. We do not think a coin would have 25% or 75% heads, nor that a heart valve have a TE rate greater than, say, 10% per year.

Prior Distribution for the Thromboembolism Rate

The current FDA guidance document for heart valves defines objective performance criteria (OPC) for valve-related complications [11]. The OPC represents the performance of approved, currently available valves, and a new device must demonstrate that its performance is comparable. Comparability is defined as "significantly better than twice as bad" as the OPC; thus the null hypothesis (which we hope to reject) is that the valve's true mean rate is twice as high as the OPC. Using a one-sided test with a power of 0.80 and size of 0.05, this resulted in a required sample size of 324 patient years for TE with mechanical valves, whose OPC is 3.0%/year [12]. (For OPC of 1.2% per year, the rates for leak and endocarditis, 800 patient years are needed; this is the minimum sample size recommended for new approval studies.)

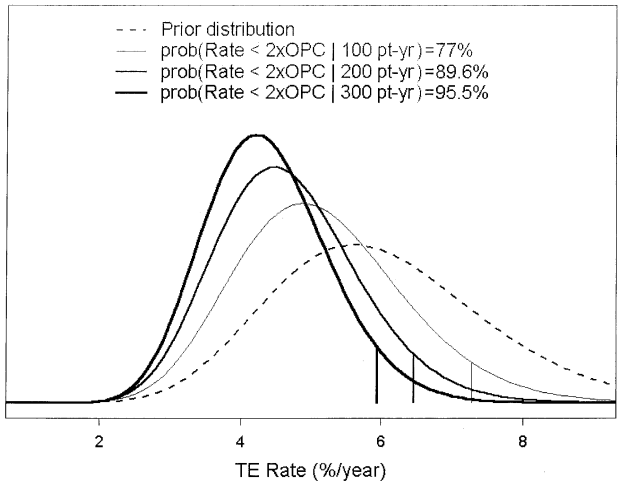

*Fig 5. The prior distribution (*dashed line*) and three posterior distributions (*increasingly thicker solid lines*, computed after increasing numbers of patient-years [*pt-yr*]), for the thromboembolism (*TE*) rate of a new mechanical heart valve, which has an observed rate of 3% per year. The* vertical lines *indicate the 95th percentiles of the posterior distributions; the probability (*prob*) that the TE rate is less than this value is 95%. When this line is less than 6% per year (two times the objective performance criteria [*OPC*]), the study can be stopped successfully.*

To be somewhat synchronous with the standard hypothesis testing design, we take our prior distribution to be the one used to represent the null hypothesis in the sample size estimation. This was a gamma distribution with a mean of twice the OPC (6% per year in this case), and the 5% quantile (the size of the hypothesis test) at a critical value (about 3.7% per year in this case) [12]. This prior distribution is indicated in Figure 5 by a dashed line. This might be called a conservative or "skeptical" [13, 14] prior, and would be suitable for one who was pessimistic about a new product and wanted to require a high burden of proof from the data. This is consistent with the null hypothesis, a dire situation which we hope the observed data is good enough to reject.

Sequential Analyses

Now, since there is no penalty for repeated analyses, we do so as each 100 patient years are accrued, an idea suggested by an FDA presentation (T. Z. Irony, FDA Issues and the Bayesian Framework). We stop the study and accept the new valve as soon as the probability is more than 95% that its TE rate is less than twice the OPC

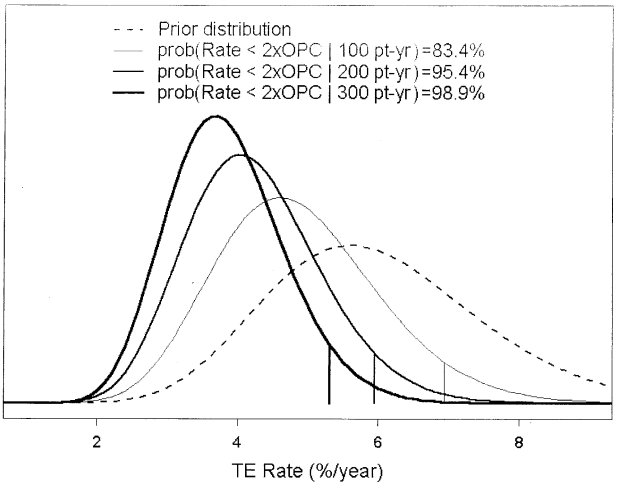

*Fig 6. The prior distribution and three posterior distributions for the thromboembolism (*TE*) rate of a new mechanical heart valve which has an observed rate of 2% per year. (*OPC = *objective performance criteria;* prob = *probability;* pt-yr = *patient-year.)*

(6% per year). The solid lines in Figure 5 are the resulting posterior distributions, assuming that the new valve has a TE rate of 3.0% per year. Even with this pessimistic prior, 300 patient years are enough to show a 95% probability that the true rate is less than 6% per year. This is slightly more efficient than the frequentist design, which required 324 patient years. And if the valve performs better, the study can be stopped sooner. Figure 6 shows that if the valve has a 2% per year TE rate, the study can be stopped after only 200 patient years.

Comments

Bayesian analysis requires an estimate of the distribution of the parameter(s) of interest to be made before the study starts. The analysis uses this estimate directly, and the result will vary depending on what it was. Thus two different analysts may get two different answers. However, one can disclose what the prior was and why, and let the consumer of the information see if they agree; or give a range of prior distributions, including one for the skeptic and one for the enthusiast [13]. There are also elements of subjectivity in classical analyses. Sample size estimation requires guesses, based on prior information. A Bayesian simply incorporates this information formally into the analysis rather than using it to fix the sample size beforehand. A predetermined sample size can be disadvantageous. If it is too small, the study will not have a high probability of reaching its appropriate conclusion. If the sample size is too large, the study will go on longer than it would have to, using more resources and patients than necessary. Bayesian analysis readily permits sequential assessments of the data. A bonus of using this method is that the resulting analysis has a more direct and satisfying interpretation, permitting, for example, true probability intervals to be given.

Our examples were quite simple compared to the wide range of problems for which Bayesian analysis can be used. We used two simple probability distribution, the binomial and the Poisson, and their conjugate priors (Appendix 3). Complex problems can also be solved by Bayes' methods, using software such as the BUGS system [15], for which an excellent website exists (http://www.mrc-bsu.cam.ac.uk/bugs/). Another interesting website, devoted to Bayesian analysis, has a wide range of information (http://www.bayesian.org/). And an easy to read review of medical statistics [16] is also available on the web (http://hesweb1.med.virginia.edu/biostat/teaching/bayes.short.course.pdf). Bayesian analysis is applicable in many other typical statistical situations, and is naturally suited for decision analysis [17] and for meta-analysis [18, 19], among other applications. It is also ideally suited for the evaluation of clinical studies of new medical technology, and FDA is now actively promoting this usage.

References

1. Press SJ. Bayesian Statistics: Principles, Models, and Applications. New York: John Wiley & Sons, 1989.
2. Bayes T. An essay towards solving a problem in the doctrine of chances. [Reprint of the original article which appeared in Philos Trans Roy Soc London. 1763: 53;370-418.]. Biometrika 1958;45:295–315.
3. Fisher RA. On the mathematical foundations of theoretical statistics. Phil Trans Roy Soc Series A 1922;222:309.
4. Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference, Part I. Biometrika 1928;20A:175.
5. Diamond GA, Forrester JS. Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease. N Engl J Med 1979;300:1350–8.
6. Staniloff HM, Diamond GA, Freeman MR, Berman DS, Forrester JS. Simplified application of Bayesian analysis to multiple cardiologic tests. Clin Cardiol 1982;5:630–6.
7. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Bayesian methods in health technology assessment: a review. Health Technology Assessment 2000;4:1–142.
8. Malec D, Campbell G. Towards a Bayesian Paradigm at the FDA's Center for Devices and Radiological Health. Case Studies in Bayesian Statistics Workshop 4. Pittsburgh, PA: Carnegie Mellon University, 1997.
9. Irony TZ, Pennello G, Campbell G. Bayesian Methodology at the Center for Devices and Radiological Health—Past, Present and Perspectives for the Future. Case Studies in Bayesian Statistics Workshop 6; Carnegie Mellon University. Pittsburgh, PA; 2001.
10. Berry D.A. Using a Bayesian Approach in Medical Device Development. A report supported by a contract from the Center for Devices and Radiological Health, US FDA; 1997.
11. Division of Cardiovascular R, and Neurological Devices, Center for Devices and Radiological Health, Food and Drug Administration. Draft Replacement Heart Valve Guidance, Version 4.1. 1994.
12. Grunkemeier GL, Johnson DM, Naftel DC. Sample size requirements for evaluating heart valves with constant risk events. J Heart Valve Dis 1994;3:53–8.
13. Spiegelhalter DJ, Freedman LS. Bayesian approaches to randomized trials. J. R. Statist. Soc. A. 1994;157:357–416.
14. Fayers PM, Ashby D, Parmar MK. Tutorial in biostatistics Bayesian data monitoring in clinical trials. Stat Med 1997;16:1413–30.
15. Gilks WR, Thomas A, Spiegelhalter DJ. A language and program for complex Bayesian modelling. Statistician 1994;43:169–78.

16. Harrell Jr. FE. Practical Bayesian Data Analysis from a Former Frequentist. From: Mastering Statistical Issues in Drug Development 2000;1–120.
17. Ashby D, Smith AF. Evidence-based medicine as Bayesian decision-making. Stat Med 2000;19:3291–305.
18. Larose DT, Dey DK. Grouped random effects models for Bayesian meta-analysis. Stat Med 1997;16:1817–29.
19. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. Stat Med 1995;14:2685–99.
20. Edwards FH, Peterson RF, Bridges C, Ceithaml EL. 1988: Use of a Bayesian statistical model for risk assessment in coronary artery surgery. Updated in 1995. Ann of Thorac Surg 1995;59:1611–2.
21. Vollset SE. Confidence intervals for a binomial proportion. Stat Med 1993;12:809–824.
22. Sahai, Kurshid A. Confidence intervals for the mean of a Poisson distribution: a review. Biomed J 1993;35:857–67.

Appendix 1

Bayes' Theorem

Conditional probability means the "probability of an event or condition (say, A) given that another condition (B) exists." This statement is written in mathematical shorthand as P(A|B), where the "|" symbol stands for "given", and is defined by the formula:

$$P(A|B) = P(AB)/P(B),$$

where P(AB) is the joint probability that both A and B occur, and P(B) is the unconditional or marginal probability of B.

In its simplest form, Bayes' Theorem can be thought of as a tool to invert the conditional probability: if we know P(A|B), we can use it to find P(B|A). By the definition above,

$$P(B|A) = P(AB) \times P(A),$$

and substituting the expression for the joint probability from above gives Bayes' theorem:

$$P(B|A) = P(A|B) \times P(B)/P(A).$$

The denominator of the right-hand side can be expanded using the theorem of total probability as

$$P(A) = P(A|B) \times P(B) + P(A|\textit{not}\ B) \times P(\textit{not}\ B).$$

Cardiac surgeons may be familiar with Bayesian risk models for operative mortality [20]. They are based on the above equation, with B being operative death. But, instead of just one prior condition, A, there are several, one for each risk factor in the model.

Rain, Go Away

Here is a nonmedical example to give the inclined reader an opportunity to apply the formula. We often hear it said "it always rains in Oregon on the weekend." Suppose condition A = RAIN and condition B = WEEKEND, and that this allegation is partly true: the probability of rain on a weekend day, P(A|B), is 50% and the probability of rain on a weekday, P(A|*not* B), is only 25%. If it is now raining outside, what is P(B|A), the probability that it is now the weekend?

The prior probability of B (weekend), without knowledge of the meteorological conditions, is 2/7, the fraction of weekend days in a week. Given, as above, that P(A|B)= 0.50 and P(A|*not* B) = 0.25, we can use Bayes' theorem to find the solution:

$$P(B|A) = 0.50(2/7)/(0.50(2/7) + 0.25(5/7)) \cong 0.444 = 44.4\%.$$

Thus, if it is raining, it is more likely to be a weekday (55.6%) than to be a weekend. This is because, even though rain is twice as likely on a weekend, there are more than twice as many weekdays.

EKG Interpretation

This same concept can be used in more practical situations, eg, the EKG interpretation example in the text. We know the sensitivity and specificity of the EKG test and the pretest probability of CAD for the patient. We want to know the post-test probability of CAD, that is, the probability of CAD given a positive EKG (+EKG). We use Bayes' formula, above, to solve this by substituting: P(CAD) is the pretest (unconditional) probability of CAD; P(+EKG|CAD) = sensitivity; P(+EKG|*not* CAD) = 1− Specificity.

For the first patient in the example, $P(CAD|+EKG) = 0.9 \times 0.8/(0.9 \times 0.8 + (1-0.9)) = 97\%$, shown by the filled circle in Figure 1. For the second patient in the example, $P(CAD|+EKG) = 0.9 \times 0.1/(0.9 \times 0.1 + (1-0.9) \times 0.9) = 50\%$, shown by the open circle in Figure 1.

Appendix 2

Confidence Intervals

Two primary aspects of conventional statistical analysis are P values ("Is the difference significant?") and confidence intervals ("What is the probable range of the estimate?"). Because of idiosyncrasies of hypothesis testing, it is often recommended to emphasize confidence intervals instead. This is especially important in the case of negative results (nonsignificant p values).

There are many options for computing a confidence interval (CI), and they give somewhat different results. For the binomial distribution used in the coin toss example, 12 methods have been reviewed [21]. For the Poisson distribution used in the heart valve example, 13 methods have been reviewed [22].

Table 1 contains four CI's:

(1) The usual one based on normal approximation to the binomial. This is simply the point estimate of the proportion plus and minus twice the standard error of that estimate. This one is symmetric, which is not ideal except for p=50%, and can be less than 0% or greater than 100%. In fact, in Table 1 it gives

105%, which would ordinarily be truncated to 100%.

(2) An "exact" CI based directly on the binomial distribution. You would think this would be the best, but it turns out that, for technical reasons, it is on average too wide.
(3) An improved method (sometimes called the Wilson method) using the normal approximation, based on inverting the usual one so as not to have to use the point estimate itself in computing the standard error.
(4) A method that uses the distribution of the likelihood ratio.

Appendix 3

Probability Distributions

Many different probability distributions are used to describe random variables. Best known is the normal distribution, the common bell-shaped curve, which describes many natural phenomena. The binomial distribution is used to describe binary (dichotomous) data and the Poisson distribution is used to describe count data.

Bayes' formula was given in Appendix 1 for converting simple pretest, or prior probabilities, into post-test, or posterior probabilities. An analogous formula can be used for entire probability distributions, f and g, and unknown parameter(s) B:

$$f(B|data) = g(data|B) \times f(B)/g(data).$$

Since the denominator does not contain the parameters, it is often omitted and the above is written as:

$$f(B|data) \sim g(data|B) \times f(B),$$

where $\sim$ means proportional to. The first term on the right-hand side is the likelihood of the data, so the above can be written more generally as,

posterior distribution $\sim$ likelihood $\times$ prior distribution.

When the prior and posterior distributions are members of the same family, this distribution is called conjugate to the likelihood. The beta distribution has two parameters, a and b, and is a conjugate prior to the binomial distribution. If the prior is $\beta(a, b)$ and the observed data is binomial (s success in n tries), then the posterior is $\beta(a+s, b+n-s)$. This was used for the coin toss example (Figs 2 and 3).

The gamma distribution has two parameters, the shape k and the (inverse) scale r, and is a conjugate prior to the Poisson distribution. Thus if the prior is $\gamma(k, r)$ and the observed data is Poisson (e events in t patient-years) then the posterior is $\gamma(k+e, r+t)$. This was used for the heart valve example (Figs 5 and 6). Both of these families of distributions, beta and gamma, are very flexible and are widely used for prior distributions.