Introduction to Bayesian methods I: measuring the strength of evidence

Steven N Goodman

Clin Trials 2005 2: 282
DOI: 10.1191/1740774505cn098oa

CLINICAL TRIALS *PRESENTATION* *Clinical Trials* 2005; **2**: 282–290

Introduction to Bayesian methods I: measuring the strength of evidence

Steven N Goodman

Bayesian inference is a formal method to combine evidence external to a study, represented by a prior probability curve, with the evidence generated by the study, represented by a likelihood function. Because Bayes theorem provides a proper way to measure and to combine study evidence, Bayesian methods can be viewed as a calculus of evidence, not just belief. In this introduction, we explore the properties and consequences of using the Bayesian measure of evidence, the Bayes factor (in its simplest form, the likelihood ratio). The Bayes factor compares the relative support given to two hypotheses by the data, in contrast to the *P*-value, which is calculated with reference only to the null hypothesis. This comparative property of the Bayes factor, combined with the need to explicitly predefine the alternative hypothesis, produces a different assessment of the strength of evidence against the null hypothesis than does the *P*-value, and it gives Bayesian procedures attractive frequency properties. However, the most important contribution of Bayesian methods is the way in which they affect both who participates in a scientific dialogue, and what is discussed. With the emphasis moved from "error rates" to evidence, content experts have an opportunity for their input to be meaningfully incorporated, making it easier for regulatory decisions to be made correctly. *Clinical Trials* 2005; **2**: 282–290. www.SCTjournal.com

Introduction

The world does not need another introduction to Bayesian methods; there are many introductions to Bayesian methods in medical journals [1–3], and now we have an excellent book as well [4]. In a brief talk one cannot teach how to perform a Bayesian analysis. However, we have a unique opportunity to have a constructive conversation among statisticians, physicians, regulators, policy makers and industry about how the Bayesian approach to data analysis might lead to better decision making in the regulatory arena. My goal will be to help ensure that we are talking about the right things.

There is a lot of mythology both about Bayesian and frequentist approaches to inference that gets in the way of meaningful dialogue. When people from different religions or cultures meet, it often takes time before they recognize their shared goals. My aim will therefore be to dispel certain myths, and hopefully to convince you that we are all share a common goal; to decide correctly. It is not about the many buzzwords that are tossed around – objectivity, subjectivity, coherence, bias, efficiency, or significance – it is about getting it right.

It is a daunting task to give a statistics lecture to such a mixed audience, an audience that includes some from whom I have learned and others who may have expressly avoided learning statistics. With such a heterogeneous group, the best way to figure out how to deliver your message is... with a test. OK, a self-assessment. Here it is:

> A well done randomized controlled trial is reported for a new electrical stimulator for pain control. The authors state that it has turned out somewhat surprisingly (i.e., they thought that it would have perhaps no more than a 25% chance of being true before the experiment) that those who had the stimulator treatment had a 15% absolute reduction in the chance of migraine (95% CI: 0–30%, $P = 0.05$). What is the probability that this association is real?
>
> A) Less than 75%
> B) 75% to 94%
> C) ≥95%

Johns Hopkins School of Medicine
Address for correspondence: Department of Oncology, Division of Biostatistics, Johns Hopkins School of Medicine, 550 N. Broadway, Suite 1103, Baltimore, MD 21205, USA. E-mail: sgoodman@jhmi.edu

10.1191/1740774505cn098oa

Based on the show of hands, about 25% felt it was "A", a little more for "B", and most for "C". The answer is "A". This audience actually did well. When I teach physicians, who are typically reluctant to answer most of my questions, this is one of the few questions that they are not scared to answer. Almost everyone raises their hands, confidently, for the wrongest answer of the lot, $\geq$95%.

The important thing is not that answer *per se*, but what it reveals. It shows that we are not taught how to use the number "25%" – the likelihood of this claim being true before the experiment. And most actually do not even understand the question "What is the probability that this association is real?" because they have never been asked that before. This talk will be partly about what that question means and how that "25%" is relevant.

I will start in a nonproselytizing mode, saying the things I *won't* say. I won't say that if we turn to Bayesian methods all of our problems will go away. Some may change, and a few may go away, but they won't all go away. I won't say that the only right thinkers in the statistics world are Bayesian. We've made quite a lot of progress with standard methods, and they embody certain sensible principles that even Bayesians must adhere to. And I won't say that the Bayesian approach doesn't have its own difficulties.

What I *will* say is that if we turn to Bayesian methods, difficult issues will be discussed in the right way by the right people. I will say that some of the dilemmas that FDA decision makers face are artifacts of a particular statistical philosophy, whose characteristics are enshrined in regulations, and not due to the requirements of science. And I will say that the Bayesian perspective provides the best way to think about evidence. In fact, that is going to be my focus today, the Bayesian and likelihood perspective on evidence.

It is sometimes implied that since standard, "objective" methods have served us well, we shouldn't fix what ain't broke. Let us remind ourselves what they have delivered into our laps. Here is a list of things that have been identified as cancer risks: electric razors; broken arms (but only in women); fluorescent lights; allergies; breeding reindeer; being a waiter; owning a pet bird; being short; being tall; and hot dogs. And, in case anyone is feeling safe – having a *refrigerator* [5]. We are apparently all at risk. These results were not produced by Bayesian methods.

Consider a study reported in the *New York Times*. The headline read "Magnets lessen the foot pain of diabetics". In the article we find these quotes: "A finding that runs counter to many studies", "We have no idea how or why the magnets work, but it is a real breakthrough", "The study must be regarded as preliminary but the early results are clear and the treatment ought to be put to use immediately" [6]. A Bayesian perspective would help prevent silliness like that.

Bayes theorem

Most are probably familiar with the procedures of medical inference. I learned them the hard way. My first two years of medical school were spent studying textbooks listing the features of various diseases. I learned about Chagas disease, malaria and hepatitis, I learned if you had hepatitis you had these symptoms, and if you had pneumonia you had those symptoms. A medical student in the pre-clinical years becomes a walking textbook. Then I got onto the wards. My first rotation was surgery, and they didn't say the woman in room 3 had hepatitis with cirrhosis. They said she had a cough, a rash and splenomegaly, and that I should go figure out her diagnosis. I felt rather foolish because my ability to go from the signs and symptoms to the diagnosis was quite limited. I had learned the science of medicine – the deductive process of going from disease to symptoms, but not the art, the more difficult but more useful inductive process of diagnosing disease from signs and symptoms.

We have exactly the same process in statistical inference. We start with some unknown underlying truth, like a difference in cure rates, and we learn in statistics class how to turn the probability crank and calculate how likely we are to observe each possible outcome; that collection of probabilities is called a probability distribution. What we don't learn, if we are not taught Bayesian methods, is how to go in the reverse direction; how to calculate the probability of underlying truths given what we observe in the data, the inductive direction. There is only one, formal, coherent calculus of inductive statistical inference: Bayes theorem. What we call "standard" frequentist methods are a collection of conventions, principles and procedures to control errors over the long run. But these methods do not tell us, in a particular study, what everyone wants to know: how likely our claims are to be true.

Figure 1 shows a simple way to state Bayes theorem. Before seeing the data, we have the relative odds that a hypothesis is true. We multiply this by the Bayes factor, which uses the data from a particular study, and from that we get a final or posterior odds that the hypothesis is true.

Note that if the alternative hypothesis is the complement of the null hypothesis, that is, $\Pr(H_0) + \Pr(H_A) = 1$, then $\Pr(H_0)/\Pr(H_A)$ becomes $\Pr(H_0)/(1 - \Pr(H_0))$, the prior odds of the null hypothesis. The prior odds should not come out of the air; it should be based on real evidence external to a study. The likelihood ratio, or Bayes factor, implicitly says

$$\underbrace{\frac{\Pr(H_0 \mid Data)}{\Pr(H_A \mid Data)}}_{\substack{Post\text{-}study \\ Odds}} = \underbrace{\frac{\Pr(H_0)}{\Pr(H_A)}}_{\substack{Pre\text{-}study \\ Odds}} \times \underbrace{\frac{\Pr(Data \mid H_0)}{\Pr(Data \mid H_A)}}_{\substack{Likelihood\ ratio \\ or\ Bayes\ Factor}}$$

Figure 1 The "odds" form of Bayes theorem (H_0 = null hypothesis, H_A = alternative hypothesis).

that the strength of the current evidence is determined by how well each competing explanation – otherwise known as the null and alternative hypotheses – predicts the data we observed. This equation offers a formal calculus not only of external plus current evidence, but it also enables us to measure evidence properly within an experiment.

Bayes theorem as a calculus of evidence

There are two views of Bayes theorem. One can look at it as a calculus of belief; that the theorem tells us what we should believe after an experiment based on the data and what we believed before. This perspective tends to engender distrust, or wariness at the very least, since the entire regulatory system is based on the use of empirical evidence to protect us from biased "beliefs". A system that incorporates beliefs at its foundation would seem to be anathema to the regulator. But an alternative perspective is that Bayesian methods offer us a calculus of evidence. That is my perspective, and will be the focus of the remainder of this presentation.

Problems with *P*-values

It is impossible to talk about a measure of statistical evidence without first discussing the *P*-value, the traditional measure of evidence, which I think is baked into every brick of regulatory buildings. Its definition is easy to state, but hard to interpret. The *P*-value is the probability of getting a result as or more extreme than the observed result if the null hypothesis (of chance) were true. A much harder question is what that means, and how it should be used. I have conducted an informal survey of textbooks, and it is extremely difficult, even if one finds the *P*-value definition, to find a subsequent understandable explanation of why we use them the way we do. There is a remarkable degree of variation in treatment of the *P*-value, with some books barely mentioning it, some devoting whole chapters to it, some skipping over the definition and going straight to the methods (hoping that you won't notice), and some saying things that are flat-out wrong.

Everyone wants the *P*-value to be the probability of the null hypothesis. That is, when $P = 0.05$, we want to say that there is only a 5% chance that the hypothesis of no effect is true. That would correspond to the "$\geq$95%" answer on the earlier quiz for the question of whether a non-null hypothesis is true. But the *P*-value cannot measure the probability of the null hypothesis, because its calculation assumes the null hypothesis is true. You can't have it both ways. You can't have a measure assuming something to be true while simultaneously measuring how likely the same thing is to be false.

There are several other things the *P*-value is not. It is not the probability that you will make a Type I error if you reject the null hypothesis. It is not the probability that the observed data occurred by chance, a circumlocution that actually means the probability of the null hypothesis. For example, if we define "chance" as being a fair coin flip, the probability that I will get 10 consecutive heads with a normal coin is less than 1/1000, but the probability that those 10 heads "occurred by chance" is 100%. The *P*-value is not the probability of the observed data under the null (chance) hypothesis, because the *P*-value includes the probability of more extreme data. In fact, the *P*-value is almost nothing sensible you can think of. I tell students to give up trying.

Many of the worst problems with the *P*-value are caused not by these misinterpretations, but by its combination with hypothesis testing, which I will not discuss here, as there has been extensive discussion in the literature [7–10]. It is critical to understand that what is at stake here is not the ability to answer questions on tests. These issues profoundly affect how we talk about real problems, how we solve (or don't solve) them, and who in the room gets to participate in discussions. I will illustrate this with the transcript of an FDA advisory panel's discussion of carvedilol. Carvedilol was a cardiovascular drug in which the studies submitted as the basis for FDA approval showed a small effect on heart failure outcomes, where investigators thought they would see a big effect, and a big effect on mortality where the investigators thought they would see the smallest effect. The problem was that the primary endpoint was designated as the less important heart failure related outcomes. Here is an excerpt from that discussion [7,8]:

> **L. Moyé, MD, PhD**: What we have to wrestle with is how to interpret *P*-values for secondary endpoints in a trial which frankly was negative for the primary. . . . In a trial with a positive endpoint. . . you

> haven't spent all of the alpha on that primary endpoint, and so you have some alpha to spend on secondary endpoints... In a trial with a negative finding for the primary endpoint, you have no more alpha to spend for the secondary endpoints.
>
> **Dr Lipicky**: What are the *P*-values needed for the secondary endpoints?... Certainly we're not talking 0.05 anymore... You're out of this 0.05 stuff and I would have liked to have seen what you thought was significant and at what level... What *P*-value tells you that it's there study after study?
>
> **Dr Konstam**: ... what kind of statistical correction would you have to do that survival data given the fact that it's not a specified endpoint? I have no idea how to do that from a mathematical viewpoint.

By the way, we are talking about cardiology here. I feel sorry for these poor panel members; they have been sucked into a whirlpool of statistical gibberish. This kind of dialogue disenfranchises the very people you want to provide the most substantive input, the people with the insight into the biological mechanism and the clinical reality. You do not want doctors talking about how to "spend their alpha". You want him or her talking about biologic plausibility, supporting evidence, the relationship between endpoints, the other clinical parameters, and so on. That may well have happened in other parts of this discussion, but there is little doubt that many felt it all came down to "alpha". The reason these panelists are struggling is that conventional statistics provides no language, no way to coherently relate their biologic or clinical knowledge to the statistical procedures used to make this decision. This panel ultimately recommended disapproval of this drug, although it was approved subsequently by panel with the same chair (Dr Moyé).

Understanding statistical likelihood

To understand Bayesian methods, we must start by understanding statistical likelihood. "Likelihood" is the statistical support provided to a hypothesis by the observed data. It is calculated from the data, but it becomes a property of the hypothesis. We therefore must distinguish likelihood in this technical sense from its informal probability interpretation. For now, we must resist talking about how "likely" something is; that concept will be reserved for words like chance and probability. When we talk about the "likelihood of a hypothesis" we will mean how strongly it is supported by the data [9,10]. This distinction between probability and likelihood is made particularly important, and particularly difficult, because we measure likelihood by using probability:

$$\textbf{Likelihood}\ (\textbf{Hypothesis} \mid \text{Data}) = \mathbf{c} \times \textbf{Probability}\ (\text{Data} \mid \textbf{Hypothesis})$$

where $\mathbf{c}$ = arbitrary constant.

This definition says that the better the hypothesis predicts the data, the stronger the evidential support for that hypothesis. The proportionality constant is quite important, because it signals that evidence is not an absolute measure. The only way to make the constant go away, and to make the evidence unique, is to take ratios. Hence, the ratio of probabilities in Bayes theorem (Figure 1) is called a "likelihood ratio".

Measuring evidence in the form of a ratio forces one, as standard methods do not, to be explicit about the alternative hypothesis. To measure evidence for or against the null hypothesis, we must answer the question, "Compared to what?" This explicit specification of an alternative, an explicitness that frequentist methods lack, produces many of the attractive properties of Bayesian procedures, for example, insensitivity to stopping rules. We will look shortly at a simple numerical example to see exactly how this works.

P-values and Bayes factors

Before proceeding to the example, let us qualitatively contrast the Bayes factor with the *P*-value (Table 1). The *P*-value is noncomparative, calculated in relation to only the null hypothesis. The Bayes factor is comparative. The *P*-value uses both observed plus hypothetical, more extreme data. The "more extreme" data depends critically on the design of the experiment. If we are looking multiple times, that "more extreme" data will be different than if we are looking once because the possible outcomes are different. The *P*-value thus depends on the design of the experiment and what we imagine will be done in the future. When the "subjectivity

Table 1 Properties of Bayes factors and *P*-values as measures of evidence

P-value	Bayes factor
Noncomparative	Comparative
Observed + hypothetical data	Only observed data
Evidence only negative	Evidence negative or positive
Alternative hypothesis implicit, data-dependent	Alternative hypothesis explicit, data-independent
Sensitive to stopping rules and study design	Insensitive to stopping rules and study design
No formal justification or interpretation	Formal justification and interpretation

of priors" is brought up in this workshop, it will be important to keep that frequentist exercise of imagination in mind.

The Bayes factor uses an explicit and predefined alternative hypothesis, something that the most hardened objectivist would like. It is often taught that you can't have evidence for the null hypothesis, only against it. That is untrue; it is an artifact of using a measure of evidence that involves only one hypothesis. The Bayes factor can be negative or positive. You can have evidence *for* the null hypothesis if you compare it to a hypothesis that is less well supported. The *P*-value calculation is affected by stopping rules, but the Bayes factor is not. Finally, the *P*-value has no formal justification or interpretation; its use is dictated by convention, conventions that in fact vary widely among scientific disciplines, even among different specialties of medicine [11]. The Bayes factor has a formal justification through Bayes theorem that gives it a consistent interpretation across all settings.

Interpreting Bayes factors

Table 2 shows how Bayes factors of various strengths combine with the prior probability to produce the final probabilities reported in the table. The Bayes factors (BF) in this table have been stated in two equivalent ways. The BF is first defined as the evidence for the alternative hypothesis versus the null hypothesis, that is, with the likelihood for the alternative in the numerator, and the null in the denominator. Therefore, BFs over 1.0 indicate stronger support for the alternative than the null, or more evidence against the null. This is a natural way to think about evidence in the Bayesian context: as positive support for one hypothesis versus another. However, because of the *P*-value, we have become accustomed to thinking about evidence in the negative sense, that is, that the smaller the evidential number is, the more evidence against the null. Table 2 therefore also shows the reciprocal of the first measure, which should be interpreted as the evidence for the null hypothesis. As with the *P*-value, the smaller this number, the more evidence against the null.

The table shows that a Bayes factor of 5 is enough evidence to move from a starting probability of the alternative hypothesis from 25% to a final probability of 62%. Similarly, it would move a starting probability of 50% to a final probability of 83%, and from 75% to 94%. For a Bayes factor of 10, which is called "moderately strong" evidence, it moves one from 25% to 77%, from 50% to 91%, and from 75% to 97%.

To preview something I will explain shortly, a *P*-value of 0.05 has a maximum evidential strength of a Bayes factor of about 7, so it falls in the category of at most moderate strength against the null. That will bring you from a prior probability of even odds on the alternative to about 87% meaning that a 50:50 hypothesis still has at least a 153 chance of being wrong after observing a *P*-value of 0.05.

The next step needed to understand Bayes theorem is to be clear on the concept of a statistical hypothesis. A statistical hypothesis is a specification of the underlying probability distribution that governs what we observe. If we were investigating cure rates, a statistical hypothesis could be that the true cure rate was 50%, sufficient information to generate a probability distribution using the binomial formula. This kind of specification is called a "simple" hypothesis, in that the underlying distribution can be fully described by specifying a single quantity, in this case the true cure rate. A statement that the cure rate was greater than 15% would constitute a "composite" hypothesis, as it includes all cure rates over 15%. That hypothesis is the composite of all the simple hypotheses that it comprises. The hypothesis that "the treatment is beneficial" is also composite, including all treatment differences above zero. This distinction between simple and composite hypotheses is necessary to understand the difference between likelihood ratios and Bayes factors; likelihood ratios involve only simple hypotheses, whereas Bayes factors encompass any kind of null or alternative. Here both terms will be used.

Table 2 Calibrating evidential strength: the effect of Bayes factors of various strengths on probabilities of the alternative hypothesis

			Final probability when prior probability is		
Strength of evidence	BF	1/BF	25%	50%	75%
Zero	1	1	25	50	75
Moderate	5	0.2	62	83	94
Moderate/strong	10	0.1	77	91	97
Strong	20	0.05	83	95	98
Very strong	40	0.025	93	97.5	99
Very strong	80	0.0125	96	99	99.6

Constructing a likelihood curve

Showing how a likelihood curve is constructed will show clearly how different likelihood is from probability. The example we will explore will be a study in which we estimate a cure rate from 15 patients. Figure 2a shows the probability density functions associated with several possible true rates, that is, how often we would observe each possible outcome (1 cure, 2 cures, and so on) under hypothetical cure rates (π) of 20%, 33% and 60%.

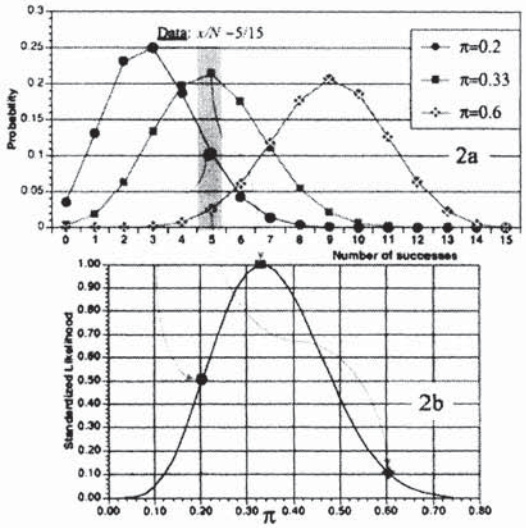

Figure 2 (a) Binomial probability distribution functions for three underlying cure rates, with $N = 15$. The data of five cures in the 15 patients is highlighted. (b) Likelihood curve constructed from that data. Arrows indicate how the curve is constructed from the probability distributions.

The curves peak at the expected number of outcomes. With a 20% cure rate ($\pi = 0.20$) you expect to see three successes out of 15, where that curve peaks. Similarly, when $\pi = 0.33$, the curve peaks at 5 and when $\pi = 0.6$ it peaks at 9. For each of these probability density functions, we keep the underlying hypothesis fixed and the possible outcomes vary.

To begin the construction of a likelihood curve, we must specify the data. This is the central concept of likelihood; the data remains fixed, and the hypotheses vary, in contrast to probability density functions, where the hypothesis is fixed, and the possible data vary. In this example, since the hypotheses are on a continuous scale, and the data are discrete, this translates into probability density functions that are discrete and a likelihood that is continuous, emphasizing the qualitative difference between the two. This qualitative difference is not as evident when dealing with Gaussian data, as both the probability functions and the likelihood are Gaussian.

Figure 2b shows the likelihood curve for the observed data of five successes out of 15 subjects. This curve is constructed from a cross-section of all of the probability density functions at the point of five successes. The arrows on the figure show how three such points translate from the probability density realm to likelihood. The likelihood function allows us to immediately see and calculate the relative support that these data give to any two hypotheses. The BF is the ratio of the heights of the likelihood at the two points corresponding to those hypotheses. In this case, the evidence favors the $\pi = 0.33$ hypothesis over the $\pi = 0.2$ hypothesis by a factor of $1/0.48 = 2.1$, and the same hypothesis over the $\pi = 0.6$ hypothesis by a likelihood ratio of $1/0.11 = 8.8$. However, we are not limited to measuring the evidence for the maximally supported hypothesis; we can choose any pair. Contrasting the likelihoods for $\pi = 0.20$ with $\pi = 0.60$ tells us that the $\pi = 0.20$ is favored by a factor of $0.48/0.11 = 4.2$.

Identical *P*-value ≠ identical evidence

The next example will use Gaussian likelihood, where it is possible to calculate the maximum possible evidence against the null hypothesis for any given *Z*-score (or *P*-value). The hypothesis for which there is maximum evidence is the hypothesis equal to the effect we observe – the "maximum likelihood estimate". In the Gaussian setting, this maximum evidence has a simple mathematical relationship to the *P*-value.

Our example will use two trials, both with $P = 0.05$. One is a large trial, showing a 5% difference in cure rates with a 95% CI of 0–10%, and the second is a small study with an observed difference of 20%, 95% CI: 0–40%. We will use likelihood to show how the evidence against the null hypothesis differs in these two cases, even though they have identical *P*-values. Figure 3 shows the likelihood curves from the two studies; the narrow curve is from the large one, and the wider curve from the small RCT. For both curves, the likelihood ratio for the maximally supported difference versus the null is the same; it is a ratio of the peak of each curve to its height at the null

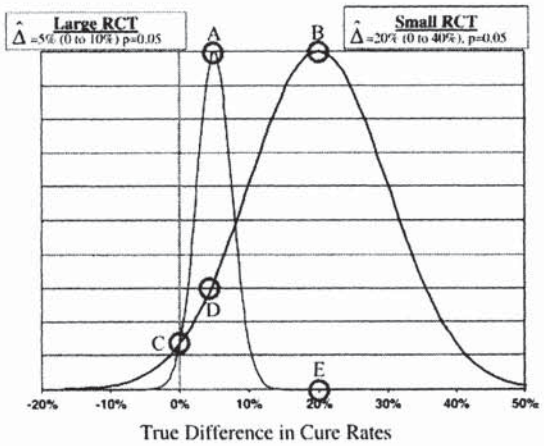

Figure 3 Likelihood curves from two hypothetical studies.

hypothesis (Points A and B, divided by C). Simple Gaussian mathematics tells us that this ratio is $e^{Z^2/2}$[12]. Conversely, the evidence for the null versus the maximally supported alternative is the reciprocal, $e^{-Z^2/2}$. For. $Z = 1.96$, corresponding to a two-sided $P = 0.05$, this likelihood ratio is 0.15 (=1/6.8).

The fact that we get the same likelihood ratio for the two maximally supported hypotheses does not mean the evidence supplied by the two studies against the null hypothesis is the same. These likelihood ratios are actually not comparable because they were calculated against two different alternative hypotheses; a 5% difference in the large trial and a 20% difference in the small trial. Let us see what happens if we ask the same evidential question of both sets of data.

In the large trial, the evidence for the null versus the alternative hypothesis of 5% was 0.15. The evidence for that same pair of hypotheses in the small trial is the ratio of point C in the figure to point D, which is 0.40. This means that the small trial supports the null hypothesis 40% as strongly as the 5% hypothesis; fairly weak evidence. This reflects what we knew intuitively from the confidence interval; the small trial does not distinguish well between a null and a 5% effect. So, if we ask the same evidential question of these different data, and measure the evidence properly, we get different answers; 0.15 for the big trial and 0.40 from the small trial.

The contrast between the strength of evidence in the two situations is much more dramatic when we examine the null versus a 20% effect. In the small trial, with an observed effect of 20%, that equaled C/B = 0.15. In the big trial, it equals C/E, greater than a million to one *in favor of the null hypothesis*. This evidential calculation formalizes what we knew intuitively; that the big trial "rules out" the 20% difference far more strongly than it does a null difference. This is expressed in a measure of evidence that actually supports the null, a concept that cannot be recaptured with a *P*-value. Table 3 summarizes these four results.

Table 4 shows how to reinterpret *P*-values in an evidential fashion. A *P*-value of 0.05 corresponds to a number 0.15 or greater, not 0.05. That Bayes factor brings you from a prior probability on the null hypothesis from 75% down to 31% or from 50% down to 13%. If you are going to have only 5% confidence in the null hypothesis after a *P*-value of 0.05, you have to have started out with a prior probability on the null of only 26% to begin with. You have to be at least three-quarters convinced of a non-null relationship before an experiment in order to claim 95% confidence after the study. For a *P*-value of 0.01, you must be no less than 40% convinced of the alternative hypothesis to have 95% percent probability after you are done.

Table 3 Bayes factors for the null hypothesis versus different alternative hypotheses for two trials in Figure 3 (see text).

H_A	Data ($P = 0.05$)	BF (H_0 vs. H_A| Data)
$\Delta = 5\%$	Big trial ($\hat{\Delta} = 5\%$)	0.15
$\Delta = 5\%$	Small trial ($\hat{\Delta} = 20\%$)	0.4
$\Delta = 20\%$	Small trial ($\hat{\Delta} = 20\%$)	0.15
$\Delta = 20\%$	Big trial ($\hat{\Delta} = 5\%$)	$>10^6$

Simple versus composite hypotheses

The hypotheses considered thus far are simple hypotheses, but more typical is a composite alternative, like "the treatment is effective". To measure the evidence for a composite hypothesis requires averaging the likelihood curve with some weighting across its components, and comparing that weighted average to the likelihood of the null hypothesis. The optimal weighting function is a Bayesian prior. This is shown graphically in Figure 4.

This makes it clear why focusing on the evidence for the maximum likelihood estimate, which in the Gaussian case is a function of the *Z*-score and directly related to the fixed sample size *P*-value, is a bit like data dredging. Choosing just that hypothesis among all components of the composite is like looking across all patient subgroups, finding the one with the largest effect, and citing this as the summary of the evidence across all subgroups. The previous example demonstrated how summarizing a likelihood curve always at its peak fails to

Table 4 Correspondence of two-sided, fixed sample size *P*-value under Gaussian distribution with the minimum Bayes factor, that is, the strongest against the null hypothesis

			Decrease in probability of the null hypothesis, %	
P-value (*Z*-score)	Minimum Bayes factor	Strength of evidence	From	To no less than
0.10	0.26	Weak	75	44
(1.64)	(1/3.8)		50	21
			17	5
0.05	0.15	Moderate	75	31
(1.96)	(1/6.8)		50	13
			26	5
0.03	0.1	Moderate	75	22
(2.17)	(1/10.5)		50	9
			33	5
0.01	0.04	Moderate	75	10
(2.58)	(1/28)	to strong	50	3.5
			60	5
0.001	0.005	Strong to	75	1
(3.28)	(1/217)	very strong	50	0.5
			92	5

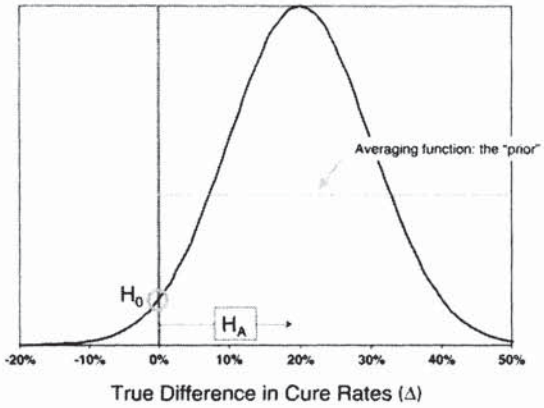

Figure 4 Demonstration of Bayes factor calculation for the null hypothesis ($\Delta = 0$) versus the composite alternative ($\Delta > 0$). The Bayes factor is equal to the height of the likelihood curve at the null hypothesis (circled), divided by the average height of the likelihood curve, averaging with the Bayesian prior probability function. The prior shown here is flat or "uninformative".

distinguish between situations of clearly different evidential import, and overstates the evidence against the null. An additional consequence is that it disconnects evidential interpretations from frequency consequences, which has serious consequences for data monitoring. In the next section, we will see how evidence and error are linked if the evidence is measured properly with the Bayes factor.

Stopping rules

It is often taught that if you look frequently at the data you are likely to come up with a spurious result, requiring adjustments to the *P*-value. The purported need for such adjustments show most clearly how the *P*-value is expected to play two roles; that of evidential measure, and an indirect measure of the frequency of Type I error. The adjustment is based on the principle that the probability of misleading evidence approaches 100% as the number of looks becomes large. But a widely underappreciated mathematical result is that when we measure evidence with the Bayes factor, the probability of misleading evidence is bounded. We can write that mathematically like this:

$$\text{As number of looks} \rightarrow \infty$$
$$\Pr(P < \alpha | H_0) \rightarrow 1$$
$$\Pr(\text{BF}(H_0 \text{vs.} H_A) < \alpha | H_0) \leq \alpha$$

This theorem states that when the null hypothesis is true, if we look repeatedly for a Bayes factor of less than 5%, strong evidence against the null, this will not occur more than 5% of the time, no matter how many times we look [9,13,14] (In most common situations, this upper bound is even lower [15].) There is thus a direct relationship between the strength of evidence and the maximum probability of misleading evidence, but only if we measure the evidence properly.

This basic result is often surprising to those schooled in the frequentist paradigm and it is worth spending a moment to examine how it occurs. I will use a fairly heuristic argument. The lack of an upper bound in the *P*-value situation occurs for reasons related to the "data dredge" analogy; if we are looking for any random high, we are sure to find it. Another way to frame that is that the *P*-value is an evidential answer to a data-derived alternative hypothesis, equal to the observed effect. We state our alternative hypothesis as composite, but we in fact measure the evidence only for a single member of that composite. In contrast, in the Bayesian approach, if a large difference is observed, the evidence for that difference is averaged in a prespecified manner with the lesser evidence for other treatment differences, tempering the evidence and putting limits on how often misleading evidence is obtained.

So the Type I error rate has a relationship to the evidential strength when evidence is defined appropriately, based on a prespecification of how the evidence across the composite alternative will be calculated. The importance of prespecifying hypotheses is well appreciated in the regulatory setting, but it is not recognized that the principle is violated by the use of the *P*-value. Viewed through this prism, the prior probability distribution is not a source of unwelcome arbitrariness, but is a valuable prior constraint, serving the same purpose as prior specification of a primary outcome measure.

This Bayesian discipline permits later freedom with design and interpretation without high rates of misleading evidence. Because the frequentist approach lacks this feature, the frequency of misleading evidence must be controlled instead through rigid constraints on the design and analysis that often do not make evidential sense.

Conclusions

What does the FDA need to know about Bayesian statistics?

1) Bayes theorem has a separable data and belief component and can be viewed as a calculus of evidence, not just belief.
2) Likelihood-based evidential measures have very attractive frequentist properties as well as sound theoretical foundations and intuitive interpretations.

3) Standard inferential methods represent evidence inappropriately, typically overstating the evidence against the null, and produce unnecessary rigidity in design and interpretation.
4) The use of Bayesian methods can have an impact far beyond the numbers they produce, even when they are similar to those produced by standard methods. Bayesian approaches affect how we talk about evidence, what evidence is seen as relevant, and who participates in that dialogue.

For standard approaches to produce similar procedures or conclusions to Bayesian ones, they can require us to "bend the rules" so the frequentist constraints make evidential (i.e., Bayesian) sense. Bending rules is awkward, particularly in a regulatory context; it is hard to know exactly when or how they can be stretched, such bending can seem arbitrary, and it takes both experience and expertise to figure out when we are going too far. Bayes theorem offers a coherent framework that allows us to know when we are being reasonable and when we are not. From the Bayesian perspective, many things that require a "bent rule" under the standard paradigm reflect scientific common sense that hews quite closely to Bayesian dictates.

The final word will be given to AWF Edwards, a student of RA Fisher, in a well-known passage that reflects on a source of trouble that I think has generated this meeting:

> What used to be called judgment is now called prejudice. What used to be called prejudice is now called the null hypothesis. It is dangerous nonsense dressed up as a scientific method and will cause much trouble before it is widely appreciated as such [11].

References

1. **Brophy JM, Joseph L.** Placing trials in context using Bayesian analysis. GUSTO revisited by Reverend Bayes [see comments]. *JAMA* 1995; **273**: 871–75.
2. **Goodman SN.** Towards evidence-based medical statistics, I: the P-value fallacy. *Annals of Internal Medicine* 1999; **130**: 995–1004.
3. **Lilford RJ, Braunholtz D.** For debate: The statistical basis of public policy: a paradigm shift is overdue. *BMJ* 1996; **313**: 603–607.
4. **Spiegelhalter DJ, Abrams KR, Myles JP.** *Bayesian approaches to clinical trials and health-care evaluation.* Chicester, UK: Wiley, 2004.
5. **Simon R, Altman DG.** Statistical aspects of prognostic factor studies in oncology [editorial]. *Br J Cancer* 1994; **69**: 979–85.
6. **Noble H.** Magnets lessen foot pain of diabetics, a study finds. *New York Times* 9 January 1999, A16.
7. **Fisher LD.** Carvedilol and the Food and Drug Administration (FDA) approval process: the FDA paradigm and reflections on hypothesis testing. *Control Clin Trials* 1999; **20**: 16–39.
8. **Fisher LD, Moyé LA.** Carvedilol and the Food and Drug Administration (FDA) approval process: an introduction. *Control Clin Trials* 1999; **20**: 1–15.
9. **Royall R.** *Statistical evidence: a likelihood paradigm.* London: Chapman and Hall, 1997.
10. **Berger JO, Wolpert RL.** The likelihood principle. *IMS Series* 1988; **6**.
11. **Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Kruger L.** *The empire of chance.* Cambridge: Cambridge University Press, 1989.
12. **Edwards W, Lindman H, Savage LJ.** Bayesian statistical inference for psychological research. *Psych Rev* 1963; **70**: 193–242.
13. **Wald A.** *Sequential analysis.* New York: Wiley, 1947.
14. **Joseph B, Kadane MJS, Seidenfeld T, Skyrms B.** *Rethinking the foundations of statistics.* Cambridge: Cambridge University Press, 1999.
15. **Blume JD.** Likelihood methods for measuring statistical evidence. *Stat Med* 2002; **21**: 2563–99.