Why a Bayesian Approach to Drug Development and Evaluation?

Frank E Harrell Jr
Department of Biostatistics, Vanderbilt University School of Medicine
and
Lisa LaVange
Department of Biostatistics, University of North Carolina Chapel Hill

April 27, 2024

Contents

Abstract

P-values and the $p < 0.05$ rule of thumb came into use before the computing revolution. Assuming the null hypothesis is true greatly simplified the model, often requiring only manual calculations. But the traditional straw-man null hypothesis testing approach to establishing statistical evidence about efficacy or safety of a drug has a number of deficiencies, many of them caused by the indirectness of the approach, including the use of probabilities of events that already occurred conditional on facts that are unknown. *P*-values and type I assertion probability α are often assumed to provide the error probabilities regulators need, but the chance that an approved drug is ineffective given the data is actually the direct Bayesian posterior probability that efficacy falls below an acceptable level. Furthermore, the rules of logic supporting proof by contradiction, based on certainties and not probabilities, don't apply to the uncertainties of traditional null hypothesis testing. Just as in medical diagnosis, forward probabilistic thinking leads to optimum decisions. The Bayesian approach involves direct estimation of time-forward probabilities of clinical interest and does not need to concern itself with long-run operating characteristics such as the number of false positives from a large set of imagined exact replications of exactly null clinical trials. Instead, Bayesian methods aim to maximize the probability of making the best decision about drug efficacy and safety for the single problem at hand. The Bayesian approach applies to complex study designs, incorporates applicable prior information, results in cleaner interpretations on a clinical as opposed to a randomness scale, and provides a fully self-contained model-based approach to inference needing no after-the-fact adjustments for context/multiplicities. In some ways frequentist hypothesis testing involves modeling noise while Bayesian inference involves modeling signal.

The Bayesian approach is outlined, and demonstrated through relatively simple simulations. By focusing on an extreme example in which one analyzes the data up to 500 times for 500 subjects, the advantages of Bayes for evidence generation and saving sample size by earlier stopping are shown. A simulated parallel-group randomized clinical trial with two efficacy endpoints is used to demonstrate how Bayes is used to quantify evidence for efficacy with joint probabilities involving both outcomes, something impossible to calculate in the frequentist paradigm. Bayesian methods should be used for simple problems as well as for complex situations such as adaptive designs and use of prior data where a frequentist solution is not available. When sponsors and reviewers are comfortable applying and interpreting Bayesian methods in simple cases they will be more able to interpret results in complex situations. When relevant prior data are not available for incorporation into the prior distribution, a little skepticism goes a long way.

1 High-Level View

Would a regulator rather know
1. the chance of making an assertion of efficacy if a drug has no effect or
2. the chance that the drug is ineffective (either has no effect or harms the patient)?

Bayesian modeling provides the perfect point in the logic flow at which to inject context-specific skepticism, or relevant positive evidence from other studies

So the Bayesian and frequentist approaches are based on inverse measures: one deals with probabilities of hypotheses given the data and the other involves probabilities of data sets given hypotheses.

Berry [8]

A fundamental tenet of the Bayesian approach: data does not create beliefs; rather it modifies existing beliefs.

AFM Smith, Seminar in Bayesian Statistics, 1982

We need an evidence measure that ignores ignorable contexts and factors in contexts that matter. And instead of computing a measure of how surprising the data are if the null hypothesis is true, Bayes reacts to whatever unknown value of the parameter is thrown at it, focused by the prior.

What is the principal distinction between Bayesian and classical statistics? It is that Bayesian statistics is fundamentally boring. There is so little to do: just specify the model and the prior, and turn the Bayesian handle. There is no room for clever tricks or an alphabetic cornucopia of definitions and optimality criteria.

Dawid [21]

The difference between Bayesian and frequentist inference in a nutshell: With Bayes you start with a prior distribution for θ and given your data make an inference about the θ-driven process generating your data (whatever that process happened to be), to quantify evidence for every possible value of θ. With frequentism, you make assumptions about the process that generated your data and infinitely many replications of them, and try to build evidence for what θ is not.

Essence of Bayes: Bayesian statistics is a mechanism for rationally updating beliefs in the light of new data. It is tailored for decision making.

Compute probabilities of things you don't know assuming things you do.

The chance that an assertion is true is more actionable than the chance of making the assertion given it's false.

It is important to be able to compute probabilities of non-trivial effects and simultaneous probabilities about multiple endpoints.

Posterior probabilities are perfectly calibrated independent of the analysis frequency and stopping rule.

A Bayesian approach to the simplest study design can without modification handle complex sequential or adaptive designs.

The benefits of obtaining direct, simply stated evidence about effects of interest, formally incorporating extra-study data and handling complex designs, are worth the price of having a prior distribution to anchor probability calculations.

The process of developing a skeptical, optimistic, or uniformative prior distribution before a study begins results in better science, and enhances objectivity in judging evidence at study end.

Compute probabilities of things you don't know assuming things you do.

> p-values assume no drug effect and compute the probability of observing data more extreme than yours; the Bayesian posterior probability is the chance the effect is positive given the observed data.

The chance that an assertion is true is more actionable than the chance of making the assertion when it's false.

> The second is the type I assertion probability α. An example of the first is the chance of a positive drug effect of 0.98, which means that one has a 0.02 chance of being wrong, whether or not the drug is approved. A type I (long-run false positive chance) assertion probability of 0.05 may be useful when designing the study, but once results are available p-value=0.03 is just a measure of how surprising the data are if the null hypothesis is true, and gives no clue about our evidence for the null and only indirect evidence for the non-null.

It is important to be able to compute probabilities of non-trivial effects and simultaneous probabilities about multiple endpoints.

> Examples: the chance that blood pressure is lowered $\geq$ 5mmHg, the chance that at least 3 of 5 efficacy endpoints are improved by the drug.

Posterior probabilities are perfectly calibrated independent of the analysis frequency and stopping rule.

> The current chance the drug is effective has the same intrepretation whether or not this probability was also computed 100 subjects ago.

A Bayesian approach to the simplest study design can without modification handle complex sequential or adaptive designs.

> Frequentist methods must envision replications of the experiment and must incorporate the intended analysis schedule in computing probabilities of data extremes under the null, which can be very complicated. Forward Bayesian posterior probabilities are merely functions of the prior distribution and the current data and are computed the same whether in the context of a simple one-look data analysis or in an analysis conducted inside a complex adaptive or sequential strategy.

The benefits of obtaining direct, simply stated evidence about effects of interest, formally incorporating extra-study data and handling complex designs, are worth the price of having a prior distribution to anchor probability calculations.

> Example: it is easy to compute the chance of a positive drug effect on two clinical endpoints in children at the 10^{th} data look given a specific skeptical prior distribution for this effect, the posterior distribution from an adult trial, and given that the adult data are 0.7 relevant to children.

1.1 Drug Approval

FDA's approval of a drug for a medical indication is based on the totality of the evidence for efficacy and safety from clinical studies. Disapprovals can sometimes be revisited but once a drug is approved for market the decision is seldom revisited unless significant evidence for safety problems emerges. Consider late-phase studies, and for now pretend that efficacy and safety can be assessed independently. In reality, regulators consider benefit-risk tradeoffs whereby the amount of efficacy that must be demonstrated is increased when there is a safety problem, and the amount of harm that can be tolerated is increased when the clinical benefit is substantial. As seen in Section 7, the Bayesian approach can formalize the analysis of benefit-risk tradeoffs[17].

There are several types of errors that can be made in the regulatory process related to efficacy studies:

1. Disapproval for a drug that is actually safe and effective

2. Approval of a drug that is safe but whose effectiveness has been overestimated and is positive but clinically trivial

3. Approval of a drug that is not effective (or is not at least trivially effective) but is safe

4. Approval of a drug that is not effective and not safe

5. Approval of a drug that has negative effectiveness but is safe

6. Approval of a drug that has negative effectiveness and is unsafe

Here negative drug effectiveness is taken to mean a worsening of an efficacy outcome that is not a safety outcome, e.g., the drug shortens walking distance but dues not increase risk of a clinical event. For the moment assume that the drug having effectiveness in the wrong direction is unlikely to show statistical evidence for positive effectiveness so that the last two errors are uncommon.

Errors in decision making are often made because the evidence was misleading. For example, there could be insufficient evidence from an underpowered study, or the efficacy could be overestimated because of statistical variation and imprecision. Wrong decisions can also be made because the evidence was misinterpreted. For example, a regulator may discount a "statistically significant" result because she believes a multiplicity adjustment should be applied to it when from a purely evidentiary viewpoint the other statistical tests that were considered were irrelevant. On the other hand, a regulator may approve a drug for which "statistical significance" is clear but where the magnitude of the drug's effect is not consistent with noticeable benefit to patients.

There are different issues for non-inferiority studies including using the wrong non-inferiority margin or significance level, and problems with interpretations of confidence intervals.

Returning to efficacy studies, the following depicts a simplified version of decisions and outcomes.

	Statistical evidence that drug works	
	+	-
Drug works	correct	incorrect
Drug doesn't work	incorrect	correct

To feel confident in an approval decision, a regulator needs to know that the chance of the error in the lower left table entry is small. The probability of this error is the conditional probability that the drug doesn't work given that the statistical evidence is interpreted to mean that the drug does work. This probability is not available in the frequentist paradigm but is a natural product of the Bayesian paradigm. To date frequentist inference is more commonly chosen in drug development, so regulators attempt to mitigate the lack of a relevant probability of "regulator's regret[1]" by considering the more easily computed probability that the statistical evidence in the long run favors drug effectiveness (by at least that observed) when in fact the drug has no effect (and has no chance of a negative effect). But once the statistical evidence is in, this probability is irrelevant just as the probability of a positive diagnostic test is irrelevant once the test result is known. Bayes' rule is needed to turn the probability on its head, as described later. The probability of getting a false positive statistical result over many trial repetitions is **not** the probability of regulator's regret.

[1]The Oxford dictionary defines *regret* as "a feeling of sadness, repentance, or disappointment over an occurrence or something that one has done or failed to do." Here we typically mean sadness over approving a drug that in the final analysis is ineffective or actually has a negative effect. The probability of regret is not just the probability of no effect or harm given that the drug was approved but is the probability of no effect or harm given the data used in the determination. Of course, regulators can also experience regret in the other direction—failing to approve a drug that is actually effective and safe.

The probability that a drug doesn't work given that the evidence is interpreted to mean effectiveness is analogous to one minus the positive predicted value in a medical diagnostic test. The frequentist type I assertion probability, i.e., the probability that the evidence can be positive given the drug doesn't work, is analogous to one minus specificity, which does not allow a physician to make a diagnosis.

An excellent review of the use of p-values at the FDA is by Kennedy-Shaffer [38]. Ionan et al. [36] have an excellent review of Bayesian approaches used at the FDA.

1.2 Statistical Inference

Biostatistics is evolving to meet ever-increasing complexities of biomedical research. While traditional statistical inference can meet the needs of relatively simple experimental designs, there is a growing need for highly sequential or adaptive randomized trials, platform trials, basket trials, biomarker-guided therapeutic trials, and use of relevant historical information such as incorporating adult data into pediatric trials that are unable to enroll a sufficient number of children. Full frequentist solutions are not available in these settings, whereas the Bayesian approach provides elegant solutions that are relatively simple conceptually. Even in applications where full frequentist solutions are available, e.g., the two-sample equal-variance t-test, a Bayesian analysis is easier to interpret, more clinically relevant, and extremely simply handles multiple data looks in sequential testing.

For inferring causation or association or for making statistical estimates of quantities of interest and their uncertainties, one must choose one or more of the existing schools of statistical inference and estimation: frequentist, Bayesian, and likelihood. It is not possible to choose a statistical paradigm that is devoid of problems. Every paradigm has its own challenges and shortcomings. A growing number of statisticians recommend use of the Bayesian paradigm, not because it is perfect but because it has fewer problems than the frequentist approach, provides more clear and clinically relevant interpretation, and is able to solve complex problems such as obtaining valid inference in response-adaptive trials.

As discussed in Section 3.1, there are many drawbacks to the traditional frequentist approach, which is especially challenged by arbitrary multiplicity adjustments and by difficulties in analyzing adaptive randomized trials, incorporating outside information, quantifying clinical significance, inferring non-inferiority, drawing simultaneous inference about multiple clinical endpoints, performing sample size re-estimation, and obtaining exact inference in situations that do not involve a normal distribution. Bayesian inference has one primary difficulty: settling on a particular encapsulation of the state of prior knowledge (or lack thereof). Once the prior distribution is selected, the above problems largely fall away. In short, *the Bayesian paradigm replaces a lot of arguments by one big argument.* As will be described below, when known-to-be-highly-relevant prior information is not available, much of the problem can be solved by using somewhat skeptical priors or non-informative priors.

In a nutshell, the frequentist approach uses only objective data at the beginning, but the interpretation can involve endless subjective debates at the end. The Bayesian approach incorporates subjectivity in quantifying beliefs at the beginning, and at the end the result is a concise clinically relevant statement about the beliefs as updated objectively by the data. A subtle but important point is that frequentist interpretations are colored/biased by observing the data, including having regret about how type I assertion probability α was divided/spent after revealing p-values, and perhaps even more importantly, these interpretations cause confusion about totality of evidence from multiple clinical endpoints and risk/benefit trade-offs. The Bayesian approach allows pre-specification of how all final evidential quantities are to be calculated, not to be influenced by data. With Bayes, the prior distribution is fixed before data are available, and may not be changed in light of data from the experiment.

Excellent introductions to Bayesian analysis may be found in [52, 40, 42, 41, 7, 57, 3, 47, 44], and

nice introductions and comparisons with frequentist and likelihood approaches may be found here and here. See also Richard McElreath's online lectures.

Natanegara et al. [49] provides key historical references and the results of the best available survey of pharmaceutical industry, regulatory, and research institutions to determine hindrances to adoption of Bayesian approaches in medical product development. The following were the the most highly endorsed reasons and needs :

- insufficient knowledge about the Bayesian approach
- lack of clarity from regulatory authorities
- need for better Bayesian training
- need for fully worked-out case studies
- need for more information about trial design, sample size determination, success criteria, and interim decisions

Self-contained R code [61] that does all the calculations and graphics below is included to make things more concrete.

Notation

Symbol	Meaning
$P(A)$	probability of event A or of the truth of assertion A
$P(A\|B)$	conditional probability that assertion A is true given that assertion B is true
p-value	frequentist p-value under a null hypothesis
$p()$	probability distribution or density function
θ	an unknown parameter or vector of unknown parameters
PP	Bayesian posterior probability given the current data

2 Background

> Data are noisy and inferences are probabilistic
>
> Kruschke [41, p. 19]

2.1 Probabilistic Thinking and Decision Making

> Optimum decisions require good data, a statistical model, and a utility/loss/cost function. The optimum decision maximizes expected utility and is a function of the Bayesian posterior distribution and the utility function[3]. Utility functions are extremely difficult to specify in late-phase drug development. Even though we seldom provide this function we need to provide the inputs (posterior probability distribution) to decisions that will be based on an informal utility function.

Statistics is all about judgment and decision making under uncertainty. The key to understanding uncertainty is understanding probability. Probability theory is all that is needed for the Bayesian approach, and the advantages of using probability theory to make predictions or statements about evidence cannot be overstated. Chief among the advantages is the ability to assign probabilities to assertions of direct interest.

Bayesian inference when exercised in its fullest form optimizes a loss/cost/utility function to arrive at an optimal decision such as whether a drug should be marketed. In the majority of

cases the utility function is unknown, so calculation of probabilities and margins of error are used to make decisions. As so eloquently described in Silver [54], optimum decisions are made when the decision maker understands all the key uncertainties and understands probabilities. This is known to lead to optimal betting strategies and forecasts. For example, Silver points out that that winningest poker players are those who are well calibrated in estimating probabilities and who never act as though the probability were exactly 0 or 1.

A key to understanding probability is the notion of which events or assertions are having their probabilities computed, and which conditions are being assumed or which data are being utilized. All probabilities are conditional on *something*. Probabilities that are immediately useful are probabilities of something unknown given (conditioning on; assuming) something known. Such probabilities respect the forward flow of information and time.

A nice interactive demonstration of conditional probability is at http://setosa.io/conditional.

The following table provides side-by-side examples of backwards and forwards probabilities. The third line in the table illustrates the common unhelpful way of reporting disease associations through such language as "$\frac{1}{5}^{\text{th}}$ of diabetics are African American whereas African Americans make up only $\frac{1}{10}^{\text{th}}$ of the population."

Type	Backwards Probability	Forwards Probability
Forecast	P(current state|future event occurs)	P(future event|current state)
Diagnosis	P(positive test|disease) (sensitivity)	P(disease|positive test)
Disease Incidence	P(black|has diabetes)	P(diabetes|black)
General	P(data|assertion X)	P(assertion X true|data)
p-value vs. PP	P(data **in general** more extreme|no effect)	P(effect|**these** data)

Medical decisons related to disease diagnosis can be thought of as maximizing an expected utility [62]. One can readily see that the expected utility is a function of the forward probability of disease and not of sensitivity and specificity.

The order of conditioning is all-important. P(female | U.S. senator) means "of senators what is the proportion of females?" and is $\frac{21}{100}$ as of 2017. P(senator | female) means "of females what is the proportion of senators" and is $\frac{21}{165{,}000{,}000}$. For non-statisticians, the table below gives more examples of translation of tendencies (assertions) into symbolic probability statements. Below, Δ refers to the unknown treatment effect/difference.

Assertion	Probability Statement
50 year old has disease now	P(disease|age = 50)
disease-free 50 y.o. will get a disease within 5y	$P(T \leq 5 \mid \text{age} = 50)$
	T = time until disease
50 y.o. male has disease	P(disease|age = 50 and male)
A drug really lowers blood pressure[a]	$P(\theta < 0 \mid \text{data})$
	θ = unknown bp Δ, data=RCT data
Drug ↓ blood pressure or ↑ exercise time	$P(\theta_1 < 0 \text{ or } \theta_2 > 0 \mid \text{data})$
	θ_1 =unknown bp Δ, θ_2 =unknown ex. time Δ

[a] By "really lowers blood pressure" we mean that the process generating the data is such that blood pressure is lower for subjects on treatment B vs. treatment A, and a random error is added to the treatment-specific mean to reflect biologic variability. Thus there is a true tendency for subjects on treatment B to have lower blood pressure, with the tendency camouflaged by biologic variability.

Spiegelhalter's classic paper [56] demonstrates the power of forward-thinking probabilities in decision making about patient management and clinical trials.

As discussed later, probabilities arising from the frequentist paradigm are indirect and hard for non-statisticians (and some statisticians) to understand and to actualize. When a probability

being estimated is actually the probability of the event or assertion of interest (as opposed to the probability of the data given the assertion is true or false), many advantages follow. The forward probability is self-contained and defines its own error probability for decision making. In making the decision about licensing a drug for market, a Bayesian PP of 0.96 of efficacy means exactly that if the drug is licensed, the chance of the decision being incorrect, i.e., of the drug being ineffective or harmful, is 0.04. This is exactly the probability of regulator's regret that is needed. On the other hand, type I "error", though comforting to some reviewers, is just the probability of rejecting H_0 given the efficacy is exactly zero. This is **not at all** the same as the probability of harm or no benefit for a drug given the data. The needed probability is one minus the PP.

> Thus to claim that the null P value is the probability that chance alone produced the observed association is completely backwards: The P value is a probability computed assuming chance was operating alone. The absurdity of the common backwards interpretation might be appreciated by pondering how the P value, which is a probability deduced from a set of assumptions (the statistical model), can possibly refer to the probability of those assumptions.

Greenland et al. [30] 2016

Just as with rain forecasting, medical diagnosis, and assessing patient prognosis, the key ingredient to demonstrating the validity of probabilistic reasoning is the accuracy of the probabilities—what is called *calibration in the small*. When the relationship between the assessed probabilities and the true probabilities of the outcome is the line of identity, the probability assessment tool is perfectly calibrated and is fit for purpose. As has been done extensively in the diagnosis and prognosis literature, examples below demonstrate how one empirically checks the known mathematics by demonstrating perfect calibration of PPs. The *calibration curve* is the relationship between an estimated probability and the actual probability. If only a few probability estimates are made, one checks their calibration merely by comparing the probabilities to the proportion of occurrences of a condition over a large number of replications. When probability estimates vary continuously, there are various ways to estimate the true probability as a function of the estimated probabilities including logistic regression and the *loess* nonparametric smoother [5], which is similar to computing smoothed moving proportions.

2.2 Bayes' Theorem

Bayesian estimation and inference come from Bayes' theorem. Suppose that one is interested in whether or not a subject has condition A vs. not A (denoted $\overline{A}$) and whether or not the subject has condition B vs. not B (denoted $\overline{B}$). The theorem comes from the laws of conditional and total probability. $P(A|B)$ denotes the probability that condition A holds given that condition B holds, and similarly for $P(B|A)$. $P(A)$ is the probability that A holds regardless of B, and likewise for $P(B)$. Bayes' theorem is stated as

$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A)} \tag{1}$$

To avoid dealing with the marginal probability of A, Bayes re-expressed the last equation. Consider the case where B can take on only two values B and $\overline{B}$:

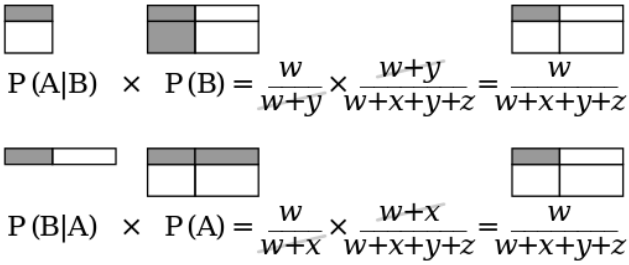
$$P(B|A) = \frac{P(A|B) \times P(B)}{(1 - P(B)) \times P(A|\overline{B}) + P(B) \times P(A|B)} \tag{2}$$

This can be written as posterior odds = prior odds × likelihood ratio since a probability equals $\frac{\text{odds}}{1+\text{odds}} = \frac{1}{\frac{1}{\text{odds}}+1}$:

$$\frac{P(B|A)}{1 - P(B|A)} = \frac{P(B)}{1 - P(B)} \times \frac{P(A|B)}{P(A|\overline{B})} \tag{3}$$

Bayes' rule is visualized in Figure 1.

Relative size	Case B	Case $\bar{B}$	Total
Condition A	w	x	$w+x$
Condition $\bar{A}$	y	z	$y+z$
Total	$w+y$	$x+z$	$w+x+y+z$



$$P(A|B) \times P(B) = \frac{w}{w+y} \times \frac{w+y}{w+x+y+z} = \frac{w}{w+x+y+z}$$

$$P(B|A) \times P(A) = \frac{w}{w+x} \times \frac{w+x}{w+x+y+z} = \frac{w}{w+x+y+z}$$

Figure 1: Contingency table illustrating Bayes' theorem, from Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Bayes_theorem_visualisation.svg. w x y and z represent frequency counts. Think of Condition A as representing a positive diagnostic test, Condition not A (Condition $\bar{A}$) as a negative test, Case B as "diseased" and Case $\bar{B}$ as "non-diseased".

For example, knowing the fraction of football game watchers who are male, the fraction of males, and the fraction of people who watch football, one can compute the fraction of males who watch football. Bayes' theorem is all about reversal of probabilities or conditions. Knowing the probability from backwards conditioning one can compute the probability under forwards conditioning, but only if the appropriate marginal probabilities are known. Without the marginal "anchors", it is impossible to compute an absolute probability. All this will be seen to be relevant when one wants to make a probability statement about a drug effect given the data where the data model (e.g., data are normal with unknown mean and variance) is a function of the unknown drug effect. We are interested in the probability that the drug effect exceeds a chosen level given the data, and the marginal probability (prior) for the unknown efficacy or safety parameter θ is absolutely necessary for calculation of this conditional probability.

An excellent video tutorial on Bayes basics may be found here.

> Bayes' theorem dictates that two individuals starting with the same beliefs (distribution) about an unknown parameter, who are given the same data, use the same data model, and agree not to redefine their prior beliefs after seeing the data, must have identical beliefs about the parameter (same conclusion about drug effectiveness) after analyzing the data.

2.3 What is a Posterior Probability?

Bayes theorem is a statistical technique that develops inference by incorporating baseline beliefs (prior) with observed data (trial results) to establish a posterior probability. This can, ideally should, be performed as an iterative process where each trial serves as the prior for subsequent trials. The exact effect (magnitude and direction) of a drug is never known with complete certainty but with the Bayes approach one can determine the probability. Moreover, Bayes offers the flexibility to ask multiple and a variety of very complex questions, such as what is the probability that the drug will have a specific effect size **and** avoid a specific adverse effect. The probability of a drug having a specific effect is determined through examination of the posterior distribution (AKA posterior density function) which is a curve akin to a histogram of the magnitude of the drug's effect (x-axis) and probability or relative degree of belief in that

effect (y-axis). This posterior distribution plot can be queried in many ways, such as asking what is the probability (area under the curve) that the drug's effect is between a and b or is the drug better than another drug (or placebo). It is, however, critical that the parameters of interest (potential questions) be pre-specified along with the prior distributions. A singular advantage of the Bayes approach over the frequentist approach is that it answers the question that clinicians (and regulators) are most interested in—"What is the probability that this drug does X?" rather than "What is the probability that I would have observed the same effect or something more extreme if the trial results were merely due to chance, i.e., if the drug does nothing?"

Probabilities in the frequentist world are long-run relative frequencies. They do not apply to one-time events, i.e., experiments that cannot be replicated, and subjectivity is not allowed. What exactly then is a PP? In the possibly biased coin example in the next section, the PP is a fully objective probability that has the usual long-run frequency interpretation because the prior distribution is not from an opinion but rather follows directly from the formulation of the problem. More generally we do not have an unassailable truth for the state of prior knowledge, and we are forced to quantify the anchor or starting point using degrees of belief. Before further discussing degrees of belief, note that one can think of a PP as a kind of conditional probability, e.g., **if** the probability that the drug actually improves blood pressure, before we knew the data from the new study, is $\frac{1}{2}$, the post-data probability of efficacy **must be** 0.93 because of Bayes' theorem.

Bayes' theorem uses new evidence (data) to translate a prior probability to a PP. If the prior probability is subjective, representing a degree of belief, one can say that the PP is a post-data degree of belief. Probability here is not a long-run relative frequency but is a metric that is between 0 and 1 and obeys certain basic axioms. As well described by Kruschke [41], "probabilities assign numbers to possibilities." All of the properties of belief probabilities needed for statistical inference are contained in Kolmogorov's axioms from 1956:

1. A probability value is ≥ 0.
2. The sum of all the probabilities across all the possibilities is 1.0.
3. If two events are mutually exclusive, the probability that either event occurs (i.e., the probability of the union of two events) is the probability of the first plus the probability of the second.

2.4 Example: Prior to Posterior

Consider a discretely-valued unknown parameter θ where it is easy to see how the data update the prior. A novelty coin maker makes a biased coin with the chance of coming up heads equal to 0.6. The coin maker randomly mixes in fair coins so that $\frac{3}{10}$ of the coins are actually fair. A coin is chosen at random from the mix and we wish to infer whether it is fair or not by making $n = 40$ tosses and observing the number of heads (y). The observed number was $y = 23$. Only two values of θ are possible; the PP of θ is zero if θ is not 0.5 or 0.6. The PP that $\theta = 0.5$ is proportional to $0.3 \times 0.5^y \times 0.5^{n-y}$, and the PP that $\theta = 0.6$ is proportional to $0.7 \times 0.6^y \times 0.4^{n-y}$. Summing these two provides the normalizing constant to get actual PPs that range from 0 to 1. These are shown as a function of y in Figure 2.

```
prior1 ← 0.3; prior2 ← 0.7
n ← 40
y ← 5:35
post1 ← prior1 * (0.5 ^ y) * (0.5 ^ (n - y))
post2 ← prior2 * (0.6 ^ y) * (0.4 ^ (n - y))
total ← post1 + post2
post1 ← post1 / total
post2 ← post2 / total
plot(y,  post1, type='n',
     xlab=expression(y), ylab='Posterior Probability')   # Fig. 2
```

```
abline(v=c(20, 23), col=gray(.9))
abline(h=prior2, col=gray(.9))
lines(y, post1)
lines(y, post2, col='blue')
text(23.5, 0.71, 'P(coin is 0.6 unfair)\nincreases with\n# heads',
     col='blue', adj=0)
text(23.5, 0.30, 'P(coin is fair)\ndecreases with\n# heads', adj=0)
op1 ← round(post1[y == 23], 2)
op2 ← round(post2[y == 23], 2)
```

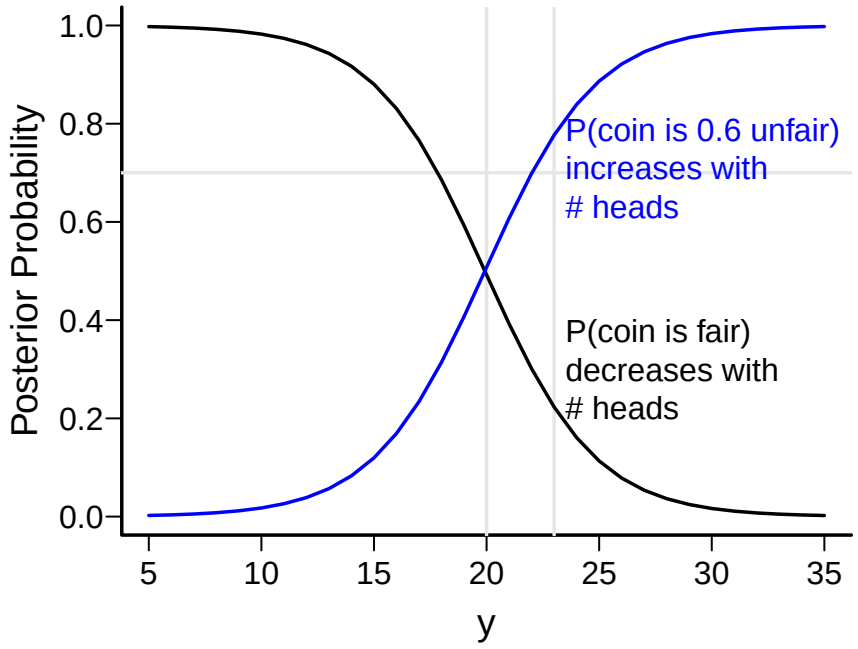


Figure 2: Posterior $P(\theta = 0.5|y)$ (black curve) and $P(\theta = 0.6|y)$ (blue curve) as a function of the observed number of heads y after 40 coin tosses. The prior probability that $\theta = 0.6$ is shown with a horizontal grayscale line. Vertical grayscale lines are shown at $y = 20$ and at the observed number of heads, $y = 23$.

The prior probabilities that the unknown probability of a head is respectively 0.5 and 0.6 were 0.3 and 0.7. With $y = 23$ the PPs are 0.22 and 0.78 so that evidence for fairness of the coin has lessened. One can read off PPs from Figure 2 had other numbers of heads been observed. The break-even value is $y = 20$ where the observed proportion of heads is $\frac{1}{2}$. The break-even point is at a proportion of $\frac{1}{2}$ and not to the right because the prior probability made it unlikely apriori for the coin to be fair. That is, observing $y = 20$ heads out of 40 tosses makes it equally likely for the coin to be fair vs. biased because of the predisposition to be biased. Had the prior probability for $\theta = 0.5$ been lower, the break-even point would be shifted to the left. The maximum posterior density estimate of θ is the value of θ that yields a PP > 0.5 in this example (where the blue curve is higher than the black), since there are only two alternative values of θ.

2.5 Bayesian Inference Model: General Case

The example above was for a discrete outcome with a finite number of possibilities for the unknown parameter. In general, we are interested in a continuous *parameter space*, e.g., in providing evidence about infinitely many possibilities for a parameter such as a mean, difference in means, or an odds ratio. When referring to a probability about the mean of a continuous variable, it must be the case that the probability that the mean equals a specific value is zero.

So to deal with continuous parameters one must deal with *probability density functions* instead of a discrete probability that exactly one value is true. For a continuous probability distribution, probability pertains only to a range of parameter values. The probability density function is related to the probability of being *near* a parameter value; it is the limit of the probability of being in a small interval divided by the width of that interval, as the width tends to zero.

The general expression used for all Bayesian inference is stated as the probability distribution function for the unknown parameters θ given the data equals the probability distribution function p for the data y given the parameter value(s) (called the *likelihood function*) times the unconditional distribution function for the parameters, divided by the marginal distribution of the data. Since the latter does not involve any unknown parameters, it eventually cancels out and does not effect inference so can be ignored, resulting in a statement of equality being replaced by a sufficient statement of proportionality. The formal expression is[4]

$$p(\theta|y) \propto p(y|\theta)p(\theta) \tag{4}$$

This can be expressed as

$$\text{prior belief about } \theta \xrightarrow{\text{data}} \text{ current belief about } \theta \tag{5}$$

This can also be expressed as "**if** you have degree of belief $p(\theta)$ about θ before seeing data from the experiment, you **must** have degree of belief $p(\theta|y)$ after unveiling the data."

The posterior density $p(\theta|y)$ is often summarized by posterior means, quantiles, cumulative probabilities, etc. To obtain quantities of interest one must integrate out θ from the above equation, requiring complex multi-dimensional integration that is usually done numerically. Simulation methods sample from $p(\theta|y)$ to estimate quantities of interest. It is often easier to sample from the posterior distribution than to derive the distribution's mathematical form.

A nice interactive demonstration for the two-sample Bayesian t-test (but with variance assumed to be known) may be found here. The standard deviation σ of the prior may be specified by the user, to indicate the amount of skepticism to use (here the prior is assumed to have a mean of zero). The observed effect (standardized difference in means) and sample size n can also be easily varied. It is easy to see that when n is large the prior has very little effect on the posterior. When n is large or the prior has a large σ, the posterior agrees with the likelihood function which summarizes the information in the data alone, ignoring the prior. Another good interactive demonstration is here.

2.6 What is Bayesian Inference Doing?

> Bayesian data analysis is a clear conceptual framework for learning

> The Bayesian approach is a common sense approach. It is simply a set of techniques for orderly expression and revision of your opinions with due regard for internal consistency among their various aspects and for the data.

Edwards, Lindman, Savage (1963) [23]

[4] A nice overview from Larry Wasserman is available at `www.stat.cmu.edu/~larry/=stat705/Lecture14.pdf`.

Type I error for smoke detector: probability of alarm given no fire

Bayesian posterior probability: probability there's a fire given current air data, whether or not an alarm is triggered

Smoke detector designed by a frequentist investigator who runs single-site RCTs: require a 0.05 false alarm probability, and require a probability (power) of 0.8 to detect an inferno

Actionable probabilities with a Bayesian smoke detector: set to sound an alarm if the probability of a fire exceeds 0.02 while you are at home or exceeds 0.01 while you are away

The Bayesian approach to statistical inference and decision making conditions on what is known to make probability statements about what is unknown. This is in contrast to the frequentist approach which conditions in reverse, i.e., conditions on the unknown status of H_0 to derive a probability statement about what is known—the data. This is akin to proof by contradiction, e.g., one assumes there is no treatment efficacy and then computes the probability of observing results as or more extreme than those observed, i.e., the p-value. A small p-value (typically $p < 0.05$) is taken to mean that what has been observed is surprising enough that one questions the premise (the null hypothesis). By having to reason while reversing the way that time and information actually flow, the frequentist approach runs into difficulties caused by (1) multiplicities due to sequential monitoring and multiple clinical endpoints and (2) adaptation of a clinical trial after it begins. There is also arbitrariness in how exactly the null hypothesis is conceived when one tries to bring evidence against a large effect as in a non-inferiority study. There is no unique prescriptive way that statisticians derive multiplicity adjustments when using frequentist methods, leaving the adjustment open to debate at every stage. A significant portion of the multiplicity problem arises from the use of null hypothesis testing (as opposed to examining evidence for non-trivial effects).

Frequentist approaches have no way of incorporating information outside a trial, and they cannot make evidential statements about totality of evidence. By contrast, a Bayesian PP that a drug does not raise the probability of death by more than 0.04 and that it either (1) lowers mortality by any amount, (2) improves exercise time by more than 5 minutes on average or (3) lowers blood pressure by more than 2 mmHg on average is very easy to compute by taking 50,000 draws from the posterior distributions of the three efficacy/safety parameters and computing the fraction of the 50,000 for which the above condition holds. See Sections 7 and 8.

Frequentist inference has the virtue and drawback of being multi-focal of having no single overarching principle of inference. From the user's point of view, having multiple principles (unbiasedness, asymptotic efficiency, coverage, etc.) gives more flexibility and, in some settings, more robustness with the downside being that application of the frequentist approach requires the user to choose a method as well as a model.

Gelman [27]

Bayesian methods are often characterized as 'subjective' because the user must choose a 'prior distribution', that is, a mathematical expression of prior information. The prior distribution requires information and user input, that's for sure, but I don't see this as being any more 'subjective' than other aspects of a statistical procedure, such as the choice of model for the data (for example, logistic regression) or the choice of which variables to include in a prediction, the choice of which coefficients should vary over time or across situations, the choice of statistical test, and so forth. Indeed, Bayesian methods can in many ways be more 'objective' than conventional approaches in that Bayesian inference, with its smoothing and partial pooling, is well adapted to including diverse sources of information.

Gelman, rwconnect.esomar.org/the-abcs-of-bayesian-basics

2.7 The Use of Prior Information . . . Or Not

Since Bayesian methods condition on what is known, it is important to define "known" aside from the observed data. Examples of what is considered "known" in a Bayesian analysis are:

- nothing (use of a flat or non-informative prior distribution)
- evidence from other studies of the same drug or device
- evidence from other studies of the same class of drug or device
- evidence from a previous clinical trial for which the current trial is a replication
- evidence about the same treatment in a different patient population
- historical evidence from observational studies
- skepticism that a treatment can have major efficacy because either
  - multiple previous clinical trials in the drug class were "negative"
  - the reviewer who needs to be convinced of an effect is a skeptic

A common misconception is that Bayesian methods will systematically require fewer or more subjects to be studied or will lower the evidence bar. Also, many non-statisticians assume that the use of Bayesian methods **requires** one to borrow information or to use expert opinion or historical data in the analysis. This is not the case; Bayesian methods are capable of only using the data at hand in conjunction with any prior distribution: non-informative, skeptical, or optimistic. The evidence bar is under complete control of reviewers according to the prior, what PP of efficacy is deemed convincing, and which effect cutoff, e.g., minimum mean difference between treatments, is inserted into the probability calculation. Most importantly, Bayes uses a better measure of evidence that increases the chance of individually approving and disapproving the *right* treatments, not just arriving at a certain false positive probability over multiple datasets in the long run.

Many Bayesian statisticians are happy to use flat (non-informative) prior distributions to let the data speak for themselves and to avoid injecting subjective bias in the the analysis. Andrew Gelman has written extensively about this issue and has pointed out that it often leads to unreliable inference, e.g., just as with frequentist methods there is a non-negligible probability of concluding efficacy when the treatment actually causes harm. Simulations presented below can easily be modified to demonstrate problems from a very practical standpoint. When for example the unknown mean μ in a single-arm study has a non-informative prior with no restrictions on its range, many of the simulations will have a value of $\mu > 1000$ when $\sigma = 1$. This will result in power indistinguishable from 1.0 and a needed sample size of 2 subjects. Another problem with flat priors is that if the study is stopped early for efficacy and the reviewer's prior does not allow for enormous treatment effects, the efficacy estimate will be miscalibrated (too high) to that reviewer (see Figure 13).

Recent excellent papers about forming priors are [66, 20], the latter delving into elicitation of priors from experts. Kopp-Schneider et al. [39] demonstrate the impossibility of preserving type I assertion probability if using external information.

> Prior beliefs are overt, explicitly debated, and founded on publicly accessible previous research. A Bayesian analyst might have personal priors that differ from what most people think, but if the analysis is supposed to convince an audience, then the analysis must use priors that the audience finds palatable. It is the job of the Bayesian analyst to make cogent arguments for the particular prior that is used. The research will not get published if the reviewers and editors think that the prior is untenable. Perhaps the researcher and the reviewers will have to disagree about the prior, but even in that case the prior is an explicit part of the argument, and the analysis should be conducted with both priors to assess the robustness of the posterior. Science is a cumulative process, and new research is presented always in the context of previous research. A Bayesian analysis can incorporate this obvious fact. . . . the priors are overt, public, cumulative, and overwhelmed as the amount of data increases. Bayesian analysis provides an intellectually coherent method for determining the degree to which beliefs should change.

Kruschke [41, p. 317]

> Any frequentist criticizing the Bayesian paradigm for requiring one to choose a prior distribution must recognize that she has a possibly more daunting task: to completely specify the experimental design, sampling scheme, and data generating process that were actually used and would be infinitely replicated to allow p-values and confidence limits to be computed.

> In the early years, many people had philosophical concerns about the status of the prior distribution, thinking that the prior was too nebulous and capricious for serious consideration. But many years of actual use and real-world application have allowed reality to overcome philosophical anxiety.
>
> . . . the practical results along with the rational coherence of the approach have trumped earlier concerns. The remaining resistance stems from having to displace deeply entrenched and institutionalized practices.

Kruschke and Liddell [42]

> The default conclusion from a noninformative prior analysis will almost invariably put too much probability on extreme values. A vague prior distribution assigns much of its probability on values that are never going to be plausible, and this disturbs the posterior probabilities more than we tend to expect—something that we probably do not think about enough in our routine applications of standard statistical methods.

Gelman [28]

> If your goal is to lie with statistics, you'd be a fool to do it with priors, because such a lie would be easily uncovered. Better to use the more opaque machinery of the likelihood. Or better yet—don't actually take this advice!—massage the data, drop some "outliers,", and otherwise engage in motivated data transformation.

McElreath [47]

2.7.1 Priors That Merely Exclude Impossible Values

In some situations a study's judge may not have any opinion about likely true values for a parameter of interest, but knows with virtual certainty that certain ranges of values are impossible. For example, one may exclude a standard deviation of systolic blood pressure that is above the mean because negative values of blood pressure are not possible. Another example is discounting a treatment effect that is larger than anything ever observed in the disease being

studied. Within an interval of possibility the prior may sometimes be taken as flat.

Consider estimation of the mean μ in a one-sample problem with known $\sigma = 1$, and suppose that the range of plausibility for μ is $[a, b]$. How does a flat prior in $[a, b]$ with zero probability outside that interval translate into a posterior distribution for a given observed sample mean? Let's take the observed mean to be 4 after a sample of size $n = 10$. Consider (1) a prior that has a wider interval of non-zero probability, (2) one that is narrower but still wide enough to include the observed mean, and (3) one that is still narrower and makes it appear that the observed mean of 4 is not to be trusted, hence should be discounted. The results are in Figure 3.

```
# Function to compute the posterior density with a truncated flat prior over x
      in [a,b]
# x=observed sample mean, s=population SD, n=sample size
pp ← function(x, a, b, s=1, n, xlim=c(0, 8)) {
    s ← s / sqrt(n)
    mu ← if(missing(a)) seq(xlim[1], xlim[2], length=150)
    else
      seq(a, b, length=150)
    eps ← 0.00001
    mus ← seq(min(mu), max(mu), by=eps)

    if(missing(a)) return(list(x=mu, y=dnorm(mu, mean=x, sd=s)))
    # http://math.stackexchange.com/questions/1787177
    y ← dnorm(x, mu, s) / (pnorm(b, mu, s) - pnorm(a, mu, s)) *
         (mu ≥ a & mu ≤ b)
    # Evaluate on a finer grid for numerical integration
    yf ← dnorm(x, mus, s) / (pnorm(b, mus, s) - pnorm(a, mus, s)) *
         (mus ≥ a & mus ≤ b)
    # Get area under the density function because we may only have the function
    # up to a constant of proportionality
    area ← eps * sum(yf[-1] + yf[-length(yf)]) / 2  # trapezoid rule
    y ← y / area; yf ← yf / area
    # Also compute posterior cumulative distribution function at selected mu's
    mus2 ← seq(2.5, 5.5, by=0.25)
    cdf ← approx(mus, eps * cumsum(yf), xout=mus2)
    list(x=c(xlim[1], a, mu, b, xlim[2]), y=c(0, 0, y, 0, 0), cdf=cdf)
}
plot(pp(x=4, n=10, a=0, b=6), type='l', ylim=c(0,10),
     xlab=expression(mu), ylab='Posterior Density')   # Fig. 3
lines(pp(x=4, n=10, a=0, b=4), col='blue')
lines(pp(x=4, n=10, a=0, b=3), col='red')
```

The prior with the widest interval of possibility for μ yields a posterior that is indistinguishable from that arising with a completely non-informative prior, and the posterior mode (most likely value for μ in the posterior distribution) is 4. When the interval of possibility is $[0, 4]$, the posterior mode is also the sample mean of 4. But when the interval is $[0, 3]$ the posterior mode moves the sample mean of 4 (that is thought to be an overestimate due to sampling variability) to the highest possible value for μ of 3.

The y-axis scale in Figure 3 is not important in absolute terms. It can be labeled *relative degree of belief* and is scaled so that the area under the posterior density is 1.0. Non-statisticians may be helped by seeing a histogram-like plot of probabilities that μ will be in intervals of width 0.25 for the black curve above, corresponding to a flat prior on $\mu \in [0, 6]$. The height of the bars sum to 1.0, and the height of each bar is the posterior probability that the unknown μ is in that bar's interval. The graph is in Figure 4.

```
cdf ← pp(x=4, n=10, a=0, b=6)$cdf
xs  ← cdf$x
interval.probs ← diff(cdf$y)
cat('Sum of interval probabilities:', sum(interval.probs), '\n')
```

```
Sum of interval probabilities: 0.9999978
```

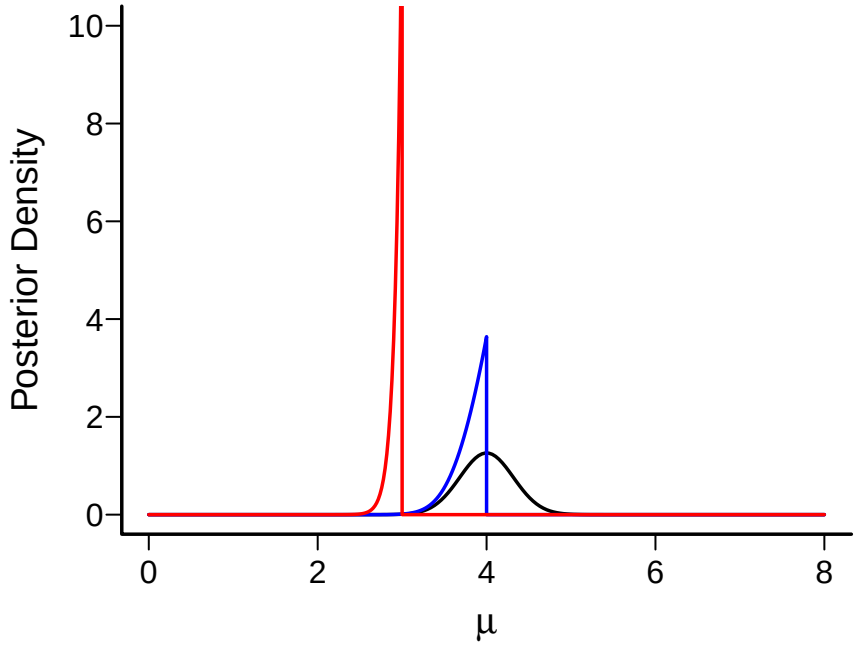


Figure 3: Posterior density functions under flat priors over different intervals. Black: interval of possibility for μ is $[0, 6]$; Blue: interval is $[0, 4]$; Red: interval is $[0, 3]$. The black posterior density function is the limit of heights of the bars in Figure 4 divided by the bar width, as the width approaches zero.

```
plot(0, 0, xlim=c(0,6), ylim=c(0, max(interval.probs)), type='n',
     xlab=expression(mu), ylab='Posterior Probability for Interval')
segments(xs[-1], 0, xs[-1], interval.probs, col=gray(.8))
lines(c(xs[1], rep(xs[-1], each=2)),    # Fig. 4
      c(rep(interval.probs, each=2), interval.probs[length(interval.probs)]))
```

General statistical guidance about selection of priors may be found at https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations.

2.8 Bayesian Inferences are Exact, To Within Simulation Error

A subtlety that is often not given enough attention by statisticians is that in the vast majority of statistical analyses, p-values and confidence intervals are approximate, and the adequacy of the approximations used is often not obvious. In the frequentist framework, only a handful of statistical tests provide exactly correct (and not conservative) p-values, e.g.,

- the linear model when the residuals truly have a normal distribution with equal variances
- the special case of the linear model that is the two-sample t-test with equal variances
- the one-sample t-test
- certain tests from simple exponential distributions
- the Wilcoxon and Wilcoxon signed-rank tests when there are no ties in the data

Other methods such as logistic models, Cox models, and mixed effects models use approximations. For nonlinear models such as logistic regression, normal approximations and even

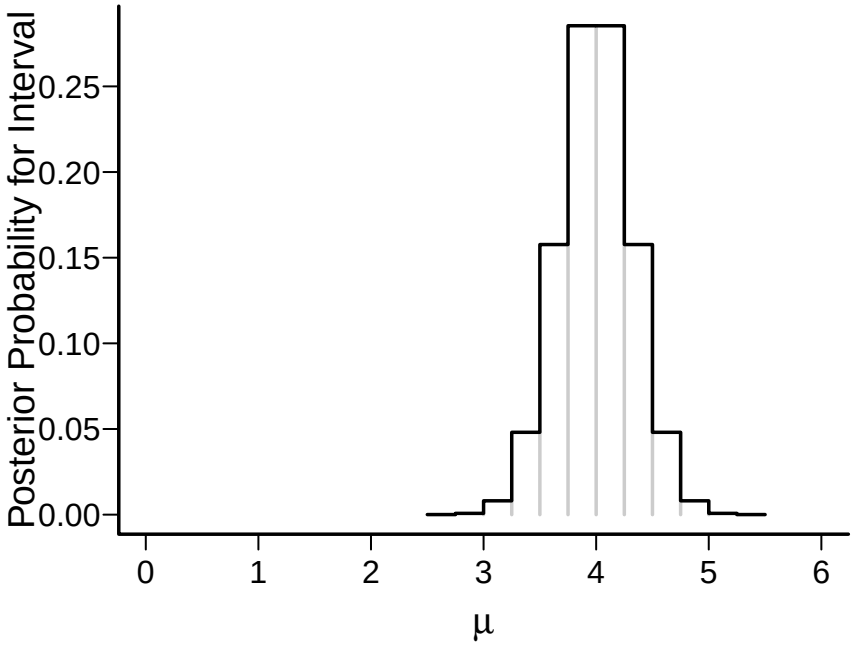


Figure 4: Posterior interval probabilities under a flat prior over the interval $\mu \in [0, 6]$

likelihood ratio tests can be inaccurate because of the non-quadratic (non-Gaussian) nature of the log-likelihood function.

By contrast, the Bayesian approach is exact if the data model is correctly specified. Occasionally exact analytic solutions are available, but more often a simulation method is used to handle the integrations involved. Once one has established that the simulation method used to obtain Bayesian posterior distributions has properly converged, i.e., the only error remaining is simulation error which can be cured by simulating tens of thousands of posterior draws, Bayesian inference requires no approximations given sufficiently many simulations[5]. Even in something as simple as a 2×2 contingency table, frequentists cannot agree on a p-value [15]. Once the Bayesian or reviewer chooses the prior distributions, inference for this situation is exact[6]. Even the "exact" confidence interval for a single proportion may not be accurate [2], whereas a Bayesian credible interval for the unknown probability is exact, given the prior.

3 Measures of Evidence

3.1 Frequentist

The p-value is the probability of obtaining data as or more impressive than the observed data given the null hypothesis H_0 is true. Here probability refers to long-run relative frequency. It is **not** the probability that the observations were produced by chance alone. H_0 is usually a test of zero effect/difference, and the test is called a null hypothesis significance test (NHST) by Bayesians. Even though in most settings outside of frequentist hypothesis testing a probability is used to describe uncertainty in an assertion given what is actually known, in NHST the p-value is a probability related to something already observed, given what cannot be known. Being

[5]For example in the worst case where the PP is 0.5, the margin of error in estimating it from 50,000 draws is 0.004.

[6]Note that Fisher's so-called "exact" test is anything but. Type I assertion probabilities can be substantially lower than claimed [18], hurting power.

backwards in terms of time/information flow is at the heart of the p-value's problems as a measure of evidence, especially when considering sequential tests or multiple endpoints. NHST has come under fire for lack of clinical relevance [45] as has over-reliance on p-values [65]. p-valuesare commonly misinterpreted by experienced scientists and even many statisticians [48]. Arbitrary multiplicity adjustments are needed because of the "what *could* have happened" approach used in NHST.

> One of the more important varieties of prejudince against the null hypothesis . . . comes about as a consequence of researchers much more identifying their own theoretical predictions with rejections (rather than with acceptances) of the null hypothesis. The consequence is an ego involvement with rejection of the null hypothesis that often leads researchers to interpret null hypothesis rejections as valid confirmations of their theoretical beliefs while interpreting nonrejections as uninformative and possibly the result of flawed mehods.
>
> Greenwald et al. [31]

A baseball analogy may help with the issue of time/information flow: a fan is interested in knowing the chance that a left-handed hitter will get a hit against a left-handed pitcher, so the probability of a hit given the handedness is relevant. Few fans (except those interested in sensitivity and specificity?) would be interested in the probability of being left handed given the batter just made a hit.

Another analogy is medical diagnostic testing. Physicians are taught that even with the availability of relevant prospective cohort data, sensitivity and specificity are the quantities that should be used to arrive at a diagnostic probability. Sensitivity and specificity (sens and spec) are in reverse time and information order, which presents major problems. Sens is the probability that a diagnostic test will be positive (even though we may already know it is negative for the patient at hand) given the disease is present, i.e. $P(T^{+}|D^{+})$. Spec is the probability that the test is negative given the disease is absent, i.e., $P(T^{-}|D^{-})$. Even though $P(D^{+}|T)$ is available as a simple proportion of cohort patients having the test outcome who are shown to have disease, clinicians relying on sens and spec must use Bayes' rule to reverse the conditioning: $P(D^{+}|T) = P(T|D^{+})P(D^{+})/P(T)$ where $P(D^{+})$ is the cohort disease prevalence and $P(T)$ is the prevalence of the particular test outcome. $P(D^{+}|T^{+}) = \frac{sens \times prevalence}{sens \times prevalence + (1-spec) \times (1-prevalence)}$. The use of time-backwards probabilities sens and spec creates a host of problems including:

1. Many physicians still do not appreciate that very high sens and spec can be ruined by low disease prevalence, just as p-values provide weak evidence about an effect.

2. It becomes natural to assume that sens and spec are constants, which is far from the truth [35] [7]

3. Because time has been reversed, what *could have happened that didn't* becomes important, just as with sequential frequentist tests and α-spending. Case in point is the need to make complex adjustments to sens and spec in the presence of verification/referral bias, e.g., when only a fraction of patients having T^{-} get the procedure that yields the final diagnosis[19]. Under the assumptions required for the usual adjustment, the adjustment cancels out once Bayes' rule is used, leading to the simple forward proportion of diseased patients in the cohort given the known test result.

For a physician, unlike recommending a biopsy when a high-specificity test is positive, choosing

[7] Time-backwards probabilities sensitivity and specificity would only be useful if they were unifying constants, but they vary strongly over types of patients. This is especially true when the disease is on a continuum of severity but has been dichotomized. Patient factors that are associated with extent of disease will significantly affect sensitivity because more severe disease is easier to detect. Diagnostic studies based on prospective cohorts should use forward probabilities (called "positive predictive values" and "negative predictive values" in the unusual case that the diagnostic test is all-or-nothing) that incorporate patient covariates to automatically model pre-test probabilities. See [34, Chapter 19] for much more information about diagnostic modeling and diagnostic test evaluation using forward probabilities.

biopsy when the probability of malignancy is 0.06 means that the chance that the physician is too aggressive is 0.94. Choosing to biopsy when the probability of cancer is 0.8 means the chance is 0.2 of the biopsy not being necessary.

Specificity and p-values are closely related and cause similar problems, due to conditioning on what is unknown to derive a probability statement about what is known (the data and the test outcome T).

The p-value has been called "the degree to which the data are embarrassed by the null hypothesis [46]." As such they can only provide evidence against something, **never** evidence in favor of something. In efficacy studies evidence against lack of efficacy can be provided, but never evidence in favor of no efficacy. Efficacy is inferred by having an abundance of evidence against "no efficacy."

In NHST one sets a type I (α) assertion probability, i.e., the probability of a false positive result. This has one well-appreciated and one little-appreciated consequence. Clinically trivial effects can be declared "significant" because given sufficiently large n signal can be detected in presence of noise, and the type I probability never gets below α even as the sample size goes to infinity. If $\alpha = 0.05$, an expected one in twenty studies will be false positive no matter how large their sample sizes. Bayesian and likelihood approaches can easily prevent this problem. A likelihood approach lets both type I and type II probabilities converge to zero [10].

Type I assertion probability may be useful at the study design stage [11]. Frequentists like type I "error" control, but after the study is completed, the only way to commit a type I error is to know with certainty the treatment has zero effect. But then the study would not have been necessary. Type I probability is a long-run operating characteristic for a sequence of hypothetical studies. Thinking of p-values that a sequence of hypothetical studies might provide, when the type I probability is α this means $P(p\text{-value} < \alpha|\text{zero effect}) = \alpha$. Neither a single p-value nor α is the probability of a decision error. They are "what if" probabilities, if the effect is zero. The p-value for a single study is merely the probability that data more extreme than ours would have been observed had the effect been exactly zero and the experiment was capable of being re-run infinitely often. It is nothing more than this. It is not a false positive probability for the experiment at hand. To compute the false positive probability one would need a prior distribution for the effect, and then one might as well be fully Bayesian and enjoy all the benefits.

> A basic difficulty for most students is the proper formulation of the alternatives H_0 and H_1 for any given problem and the consequent determination of the proper critical region (upper tail, lower tail, two-sided). . . .
>
> *Comment.* Small wonder that students have trouble. They may be trying to think. . . .
>
> *More on the teaching of statistics.* Little advancement in the teaching of statistics is possible, and little hope for statistical methods to be useful in the frightful problems that face man today, until the literature and classroom be rid of terms so deadening to scientific enquiry as null hypothesis, population (in place of frame), true value, level of significance for comparison of treatments, representative sample.
>
> Statistical significance of B over A thus conveys no knowledge, no basis for action.

Deming [22]

> . . . Another concern is that Bayesian methods do not control error rates as indicated by p values. . . . This concern is countered by repeated demonstrations that error rates are extremely difficult to pin down because they are based on sampling and testing intentions.

Kruschke and Liddell [42]

> If the design were unknown, then it is not possible to calculate a P value. ... Every practicing statistician must deal with data from experiments the designs of which have been compromised. For example, clinical trials are plagued with missing data, patients lost to follow-up, patients on the wrong dosing schedule, and so forth. Practicing statisticians cannot take the unconditional perspective too seriously or they cannot do statistics!
>
> Berry [8]

There are four other subtle but important problems with p-values. First, the use of "proof by contradiction" to make inference never applied:

> The following is almost but not quite the reasoning of null hypothesis rejection:
>
> If the null hypothesis is correct, then this datum (D) can not occur.
> It has, however, occurred.
> Therefore the null hypothesis is false.
>
> If this were the reasoning of H_0 testing, then it would be formally correct. ... But this is not the reasoning of NHST. Instead, it makes this reasoning probabilistic, as follows:
>
> If the null hypothesis is correct, then these data are highly unlikely.
> These data have occurred.
> Therefore, the null hypothesis is highly unlikely.
>
> By making it probabilistic, it becomes invalid. ... the syllogism becomes formally incorrect and leads to a conclusion that is not sensible:
>
> If a person is an American, then he is probably not a member of Congress. (TRUE, RIGHT?)
> This person is a member of Congress.
> Therefore, he is probably not an American. (Pollard & Richardson, 1987)
> ... The illusion of attaining improbability or the Bayesian Id's wishful thinking error ...
>
> Induction has long been a problem in the philosophy of science. Meehl (1990) attributed to the distinguished philosopher Morris Raphael Cohen the saying "All logic texts are divided into two parts. In the first part, on deductive logic, the fallacies are explained; in the second part, on inductive logic, they are committed."
>
> Cohen [16]

> A person is interested in a probability model. But guided by the philosophy of p-values, he asks no questions about this model, and instead asks what is the probability, given the data and some other model, which is not the model of interest, of seeing an ad hoc statistic larger than some value. (Any change in a model produces a different model.) Since there are an infinite number of models that are not the model of interest, and since there are an infinite number of statistics, the creation of p-values can go on forever. Yet none have anything to say about the model of interest.
>
> Why? Fisher (1970) said: "Belief in null hypothesis as an accurate representation of the population sampled is confronted by a logical disjunction: Either the null is false, or the p-value has attained by chance an exceptionally low value."
>
> Fisher's "logical disjunction" is evidently not one, since the either-or describes different propositions. A real disjunction can however be found: Either the null is false and we see a small p-value, or the null is true and we see a small p-value. Or just: Either the null is true or it is false and we see a small p-value. Since "Either the null is true or it is false" is a tautology, and is therefore necessarily true, we are left with, "We see a small p-value." The p-value casts no light on the truth or falsity of the null.
>
> Frequentist theory claims, assuming the truth of the null, we can equally likely see any p-value whatsoever. And since we always do (see any value), all p-values are logically evidence for the null and not against it. Yet practice insists small p-value is evidence the null is (likely) false. That is because people argue: For most small p-values I have seen in the past, the null has been false; I now see a new small p-value, therefore the null hypothesis in this new problem is likely false. That argument works, but it has no place in frequentist theory (which anyway has innumerable other difficulties).
>
> Any use of p-values in deciding model truth thus involves a fallacy or misunderstanding. This is formally proven by Briggs (2016, chap. 9), a work which I draw from to suggest a replacement for p-values, which is this. Clients ask, "What's the probability that if I know X, Y will be true?" Instead of telling them that, we give them p-values.

Briggs [13]

Second, the p-value **is not** the probability of achieving a result as impressive as that observed. That probability is zero when the distributions are continuous. The p-value is the probability of observing a result **more impressive** than that observed.

Third, the type I assertion probability is computed under the assumption that the treatment has no effect, and does not entertain the possibility that it is actually harmful.

Finally, it could be argued that the type I error is always zero, if "error" is taken to mean being incorrect in concluding a drug has nonzero effect, as all non-placebos have *some* effect.

> The nil hypothesis is always false. Tukey (1991) wrote that "It is foolish to ask 'Are the effects of A and B different?' They are always different—for some decimal place". Schmidt (1992) . . . reminded researchers that, given the fact that the nill hypothesis is always false, the rate of Type I errors is 0%, not 5%, and that only Type II errors can be made.

Cohen [16]

Why are p-values still used?

> Feinstein [25] believes their status ". . . is a lamentable demonstration of the credulity with which modern scientists will abandon biologic wisdom in favor of any quantitative ideology that offers the specious allure of a mathematical replacement for sensible thought."

> "It is incomparably more useful to have a plausible range for the value of a parameter than to know, with whatever degree of certitude, what single value is untenable." — Oakes [50]

Hypothesis testing usually entails fixing n; many studies stop with $p = 0.06$ when adding 20 more patients could have resulted in a "positive" study. And the frequentist approach to unblinded sample size re-estimation would require an adjustment for multiple comparisons that makes the final test (after adding 20) conservative, effectively ignoring many of the first wave of patients.

Confidence intervals go hand-in-hand with NHST, and the indirect reasoning that is central to the frequentist approach ensures that confidence intervals have complex interpretations. They have long-run operating characteristics rather than providing evidence directed solely at the study at hand. 0.95 confidence limits are numbers so constructed that if reconstructed afresh for 1000 studies one expects 950 of the confidence limits to contain the true unknown population parameter. Confidence intervals, because they are flat, also give the false impression that all values of the unknown parameter are equally likely. This should be contrasted with the Bayesian posterior distribution.

> I see that the 0.95 confidence interval for the mean blood pressure difference is $[2, 7]$. But I want to know the confidence I should have in it being in the interval $[0, 5]$ and you're telling me it can't be computed with frequentist confidence intervals?

see Wagenmakers et al. [63]

As an example of the typical statement of clinical trial results in the frequentist world, the difference in mean blood pressure between treatments A and B of 6 mmHg is associated with $p = 0.01$ and a 0.95 confidence interval of $[3, 9]$. An event (6 mmHg) of relatively low probability has just been witnessed if H_0 is true.

> The worry is that, when data are weak and there is strong prior information that is not being used, classical methods can give answers that are not just wrong—that's no dealbreaker, it's accepted in statistics that any method will occasionally give wrong answers—but clearly wrong; wrong not only just conditional on the unknown parameter but also conditional on the data. Scientifically inappropriate conclusions. That's the meaning of 'poor calibration.' Even this, in some sense, should not be a problem—after all, if a method gives you a conclusion that you know is wrong, you can just set it aside, right?—but, unfortunately, many users of statistics consider to take $p < 0.05$ or $p < 0.01$ comparisons as 'statistically significant' and to use these as a motivation to accept their favored alternative hypothesis. This has led to such farces, in recent claims, in leading psychology journals that various small experiments have demonstrated the existence of extra-sensory perception or huge correlations between menstrual cycle and voting, and so on.

Gelman [27]

> So what happened with the development of efficacy measures is we developed a whole new field called biostatistics. It had been sort of an orphan corner of mathematics until the Kefauver-Harris Amendments, and there had been extremely important advances in how do you study efficacy of drugs. Most of it devolves down to whether or not you're likely to see a benefit more than chance alone would predict. But *how likely* and *how much* benefit was left for some free-floating kind of notion by the FDA. So *any* benefit in essence, more than any toxicity in essence, would lead to licensure. That has led to what I call "small effectology."

Nortin Hadler, Interviewed by Tom Ashbrook *On Point*, WBUR radio, 2016-03-29, 15:26

> p-value: the chance that someone else's data are more extreme than mine if H_0 is true, not the chance that H_0 is true given my data

> Aside from ignoring applicable pre-study data, the p-value is at least monotonically related to what we need. But it is not calibrated to be on a scale meant for optimum decision making.

> The criterion of $p < .05$ says that we should be willing to tolerate a 5% false alarm rate in decisions to reject the null value. In general, frequentist decision rules are driven by a desire to limit the probability of false alarms. The probability of false alarm (i.e., the p value) is based on the set of all possible test results that might be obtained by sampling fictitious data from a particular null hypothesis in a particular way (such as with fixed sample size or for fixed duration) and examining a particular suite of tests (such as various contrasts among groups). Because of the focus on false alarm rates, frequentist practice is replete with methods for adjusting decision thresholds for different suites of intended tests. ...
>
> Bayesian decisions are not based on false alarm rates from counterfactual sampling distributions of hyopthetical data. Instead, Bayesian decisions are based on the posterior distribution from the actual data.

Kruschke and Liddell [42]

> ...Neyman and Pearson outline the price that must be paid to enjoy the purported benefits of objectivity: We must abandon our ability to measure evidence, or judge truth, in an individual experiment. ...Hypothesis tests are equivalent to a system of justice that is not concerned with which individual defendent is found guity or innocent (that is , "whether each separate hypothesis is true or false") but tries instead to control the overall number of incorrect verdicts (that is, "in the long run of experience, we shall not often be wrong"). Controlling mistakes in the long run is a laudable goal, but just as our sense of justice demands that individual persons be correctly judged, scientific intuition says that we should try to draw the proper conclusions from individual studies.

Goodman [29] .

3.1.1 Computing p-values Using Simulation

Simulation often exposes what is really going on without using math or distribution theory. The R code below shows how one can compute a p-value using basic ideas and no theory other than the fact that the mean is an optimal measure of central tendency when the data come from a normal distribution. As used in detailed Bayesian simulations in later sections, consider the one-sample problem where the data come from a normal distribution with unknown mean μ and known variance $\sigma^2 = 1$, where $\mu > 0$ indicates efficacy in a single arm study. The one-sided p-value is the probability upon repeated sampling of getting a sample mean at least as large as the observed mean given $\mu = 0$. We first draw a sample of size $n = 30$ from a true $\mu = 0.3$. Then we run 100,000 studies where $\mu = 0$ and save the estimated means. This can be done almost instantly if we use the knowledge that the sample means have a population mean of zero and a variance of $\frac{1}{30}$. We take the opportunity to show that the one-sided 0.95 confidence interval has the intended properties.

In addition we compute what some researchers hope that a p-value represents: the chance of getting a result **as** impressive as that observed. Because the sample mean has a continuous distribution, this probability is actually zero, so we relax the criterion and compute the probability of getting a sample mean at least as large as that observed but no more than 0.1 units ($\frac{\sigma}{10}$) larger than it. This is labeled "probability of approximately as impressive" below.

```
n ← 30
set.seed(1)
y     ← rnorm(n, 0.3, sd=1)             # generate data
ybar  ← mean(y)                         # observed mean
ucl   ← ybar + qnorm(0.95) / sqrt(30)   # upper C.L.
```

```
# Run 100,000 studies and compute their sample means:
repeated.ybar ← rnorm(100000, 0, sd=sqrt(1/30))
# TRUE/FALSE variables are converted to 1/0 when taking the mean
# This is an easy way to compute a proportion
p     ← mean(repeated.ybar ≥ ybar)
pa    ← mean(repeated.ybar ≥ ybar & repeated.ybar ≤ ybar + 0.1)
repeated.ucl ← repeated.ybar + qnorm(0.95) / sqrt(30)
cover ← mean(repeated.ucl ≥ 0)
cat('Observed mean           : ', round(ybar, 3), '\n',
    'Upper 0.95 1-sided CL   : ', round(ucl, 3), '\n',
    'One-sided p-value       : ', round(p, 4), '\n',
    'Exact p-value           : ', round(1 - pnorm(ybar, 0, sd=1/sqrt(30)),
                                    4), '\n',
    'Confidence coverage     : ', round(cover, 4), '\n',
    'P(Approx. as impressive): ', round(pa, 4), '\n',
    sep='')
```

```
Observed mean           : 0.382
Upper 0.95 1-sided CL   : 0.683
One-sided p-value       : 0.0186
Exact p-value           : 0.0181
Confidence coverage     : 0.949
P(Approx. as impressive): 0.0146
```

Next, modify the simulation so that two looks are taken at the data and a stopping rule is used. Using the cutoff on the mean of the first 15 subjects that yields a nominal type I probability of 0.05, stop the study at 15 subjects if the mean exceeds this cutoff. Otherwise use the mean of 30 subjects at the end and consider whether it exceeds the cutoff that preserves type I probability had a single analysis of 30 subjects been done. Compute the actual p-value under this decision rule after computing the two nominal p-values. The simulation code exposes some assumptions: the intended early look is actually carried out (e.g., the data monitoring committee did not cancel the meeting at the last minute) and is ignored if the nominal p-value > 0.05.

```
set.seed(1)
# Make first look
y1 ← rnorm(n / 2, 0.3, sd=1)
ybar1 ← mean(y1)
# Make second look
y2 ← rnorm(n / 2, 0.3, sd=1)
ybar2 ← mean(c(y1, y2))      # combine to get n=30
ybar.at.stop  ← ifelse(ybar1 * sqrt(15) ≥ qnorm(0.95), ybar1, ybar2)
ybar.at.stopb ← ifelse(ybar1 * sqrt(15) ≥ qnorm(0.975),ybar1, ybar2)
# Run 100,000 studies.  For each get mean with n=15 and 30 and apply the same
    stopping rule
repeated.ybar1 ← rnorm(100000, 0, sd=sqrt(1/15))
# Compute overall mean with n=30:
repeated.ybar2 ← (repeated.ybar1 + rnorm(100000, 0, sd=sqrt(1/15))) / 2
repeated.ybar  ← ifelse(repeated.ybar1 * sqrt(15) ≥ qnorm(0.95),
                        repeated.ybar1, repeated.ybar2)
repeated.ybarb ← ifelse(repeated.ybar1 * sqrt(15) ≥ qnorm(0.975),
                        repeated.ybar1, repeated.ybar2)
pval1          ← mean(repeated.ybar1 ≥ ybar1)
pval2          ← mean(repeated.ybar2 ≥ ybar2)
pval           ← mean(repeated.ybar  ≥ ybar.at.stop)
pvalb          ← mean(repeated.ybarb ≥ ybar.at.stopb)

cat('Sample mean at first look     :', round(ybar1, 3), '\n',
    'Sample mean at end            :', round(ybar2, 3), '\n',
    'Nominal p-value at first look :', round(pval1, 4), '\n',
    'Nominal p-value at end        :', round(pval2, 4), '\n',
    'p-value accounting for looks  :', round(pval,  4), '\n',
    'p-value " " with alpha=0.025  :', round(pvalb, 4), '\n', sep='')
```

```
Sample mean at first look      :0.401
Sample mean at end             :0.382
Nominal p-value at first look :0.0604
Nominal p-value at end         :0.018
p-value accounting for looks   :0.0585
p-value " " with alpha=0.025   :0.0367
```

When the null hypothesis is exactly true and one has two chances to declare efficacy, even though half of the subjects in the "second chance" are also in the first, the true p-value is much larger than the p-value that would be computed had the first look not been done, even though the first look was inconsequential[8]. As expected, if the initial look used a nominal $\alpha = 0.025$ for stopping, the true p-value is smaller, because this ignores the first look in a greater number of simulated trials.

The fix for the positive bias in the final mean chosen by the stopping rule is quite complicated, which translates to extreme difficulty in deriving confidence intervals[9]. The sampling distribution of the final mean from our stopping rule is given in Figure 5. The discontinuity is at the critical value of the sample mean from the first test with $n = 15$. Bayesian inference on the other hand does not concern itself with sampling distributions. Instead of considering probabilities of observing specific values of summary statistics over study replications and accounting for stopping rules, Bayesian analysis considers probabilities of specific values of the unknown efficacy parameter.

```
hist(repeated.ybar, nclass=100, xlab=expression(bar(Y)), main='')
abline(v=qnorm(0.95) / sqrt(15), col=gray(.85))   # Fig. 5
```

3.2 Bayesian

The Bayesian approach to statistical inference recognizes that there are no absolute truths, yet we seek the truth about an assertion such as a drug is effective. Bayesian evidence is couched in terms of degrees of belief (this being the Bayesian notion of probability), and two observers who started with the same knowledge base and biases, and given the same data and statistical model for the data, would necessarily arrive at the same conclusion about the assertion. Relative changes in evidence, e.g., likelihood ratios in the likelihood paradigm or the ratio of posterior to prior odds in the Bayesian paradigm, are functions only of the data at hand. But a final evidence measure for an effect can only be quantified on an absolute scale given a pre-data anchor or prior distribution. At the heart of Bayesian modeling is the movement of prior belief to current (posterior) belief.

> This form (probability of unknown given what is known) has enormous benefits. It is in plain language; specialized training is not needed to grasp model statements ... Everything is put in terms of observables. The model is also made prominent, in the sense that it is plain there is a specific probability model with definite assumptons in use, and thus it is clear that answers will be different if a different model or different assumptions about that model are used ...
>
> Briggs [13]

Because Bayesians use full conditioning on available information and do not condition on unknowable quantities such as the true treatment effect, probability statements operate forward in time and information flow and can be interpreted out of context. The Bayesian approach uses a direct forward probability model [33]. Multiple looks do not matter, and the stopping rule used

[8] This is because the first look *could have been* consequential. The simulations take the "could have" detours.

[9] In fact, the frequentist approach can result in stopping early for efficacy but having the final confidence interval include the null value.

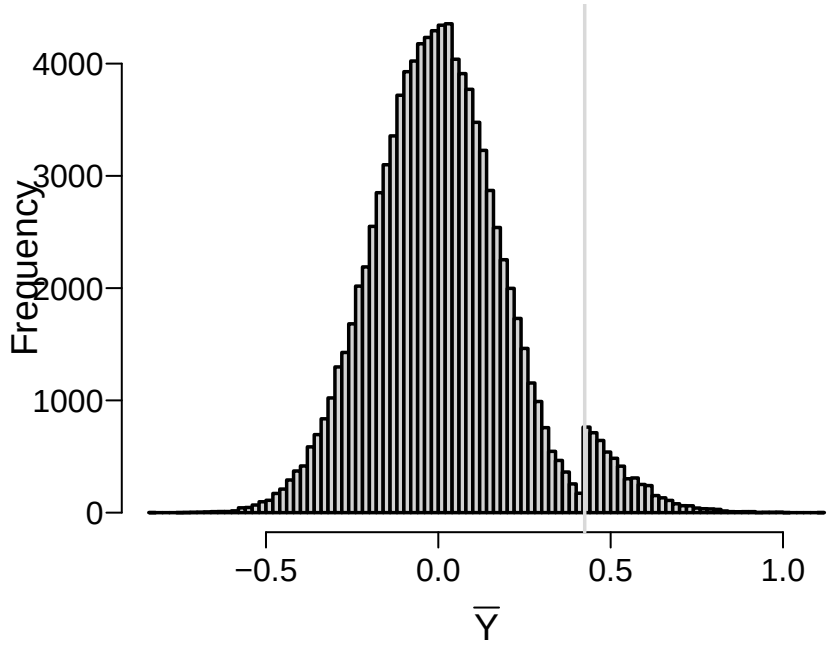


Figure 5: Sampling distribution of final estimate of the mean in a two-stage sequential single arm trial, under the null hypothesis

for a study is not relevant to the interpretation of the data. There are two ways for Bayesians to cheat: by changing the prior after seeing the data, or by hiding data. If the PP for efficacy is 0.95 and the study enrolls more subjects to refine the information but the new PP is 0.93, failure to condition on the new information and instead reporting the 0.95 is cheating.

A Bayesian analysis of the hypothetical blood pressure study mentioned above might be stated as follows: Using a normal prior distribution that assumed (1) the pre-study chance that the drug worsens blood pressure is $\frac{1}{2}$ and (2) the pre-study chance of a large ($\geq$ 10 mmHg) improvement in mean blood pressure is only $\frac{1}{10}$, the posterior mean blood pressure reduction was 5 mmHg with a 0.95 credible interval of $[2.5, 8]$. The probability of any reduction in blood pressure is 0.97, and the probability of at least a 2 mmHg reduction is 0.9. Note that the credible interval is what is sought by clinicians when they compute confidence limits. With probability 0.95 the unknown true mean blood pressure reduction is between 2.5 and 8 mmHg. The probability of a blood pressure reduction being in a "similarity zone" of $[-2, 2]$ mmHg could easily be computed. Contrast this with the frequentist result from Section 3.1:

- B-A sample mean blood pressure difference = 6 mmHg: not discounted by prior skepticism
- $p = 0.01$: chance of observing a mean difference > 6 mmHg in infinitely many repeats of the same experiment if the true mean difference is zero is 0.01
- 0.95 confidence interval $[3, 9]$: infinitely many repeats of the same experiment in which the 0.95 confidence limits were recalculated using new data would have the true unknown mean difference contained in the new interval 0.95 of the time

The overall result could be plotted as a posterior density function as in the coin tossing example below. The four probabilities just listed can be depicted by shaded regions of the density function. Envisioning replications of the study is not a part of the Bayesian interpretation.

When the posterior probability density function is plotted, it enhances cognition by virtue of not being flat but of showing regions where the unknown parameter value is more likely to be, i.e., regions where the function is high. And the progression of increasing information content in a study as it progresses can be readily seen. Consider a coin-flipping experiment in which apriori the coin is believed to be more fair than not, encapsulated in a prior distribution for the probability θ of heads being a beta distribution having shape parameters $\alpha = \beta = 10$. The mean of this distribution is 0.5, and the entire prior distribution is shown in the graph below. The number of heads from N tosses follows a binomial distribution with parameter θ. A random number generator is used to "toss the coin" 100 times, and the posterior distribution is shown after $N = 10, 20, \ldots, 100$ tosses. This distribution is a beta distribution with parameters $Y + \alpha$ and $N - Y + \beta$ where Y is the number of heads after N tosses. Prior and posterior distributions are shown in Figure 6.

```
x ← seq(0, 1, length=200)
set.seed(1)
alpha ← beta ← 10
Y ← 0
# Plot beta distribution density function
plot(x, dbeta(x, alpha, beta), type='l', ylim=c(0, 10),
     xlab=expression(theta), ylab='', col='blue', bty='l')
abline(v=0.5, col=gray(.9))    # Fig. 6
for(N in seq(10, 100, by=10)) {
  Y ← rbinom(1, 10, 0.5)  # 10 new tosses
  # Posterior distribution updated
  alpha ← alpha + Y
  beta  ← beta + 10 - Y
  lines(x, dbeta(x, alpha, beta),
        col=if(N < 100) gray(.95 - N / 120) else 'red',
        lwd=N * 2 / 100)
  }
```

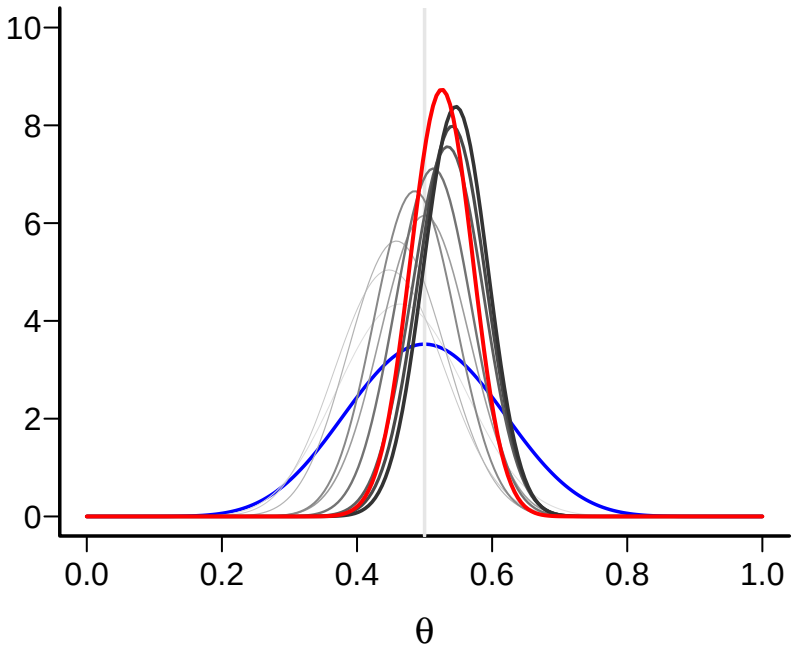


Figure 6: Prior distribution (blue) and posterior distributions as the trials progress (darkness of lines increases). The final posterior at $N = 100$ is in red.

The number of heads tossed by the end (100 tosses) was 53.

What if a study had been designed to stop when the PP of efficacy exceeds 0.95 and the statistician had taken 200 looks at the data? A PP of 0.97 stands and needs no re-interpretation in light of multiple looks. This is demonstrated in a simulation below.

Some like to think of one minus a PP as being akin to the p-value. This clouds thinking and is only appropriate in very special cases [28]. As analogies, sens and spec tell a physician very little about the probability of disease, and the probability that a batter is left handed (which does not immediately lead to knowing the probability of a hit) is not of interest because the viewer knows his handedness once he steps to the plate.

3.2.1 Alternative Take on the Prior

When comparing the Bayesian inference under a skeptical prior to Bayesian inference under a non-informative (flat) prior (or with frequentist inference using an unadjusted p-value), one can think of the effect of skepticism as being equivalent to discounting the results by effectively ignoring a certain number of observations [68]. Consider the case where one wishes to estimate an unknown mean μ from a sample of size n using a skeptical prior that has mean 0 and variance $\frac{1}{\tau}$ (τ is called the *precision*). The population standard deviation is taken to be 1.0. With the observed sample mean being $\overline{Y}$, the posterior density of $\mu|\overline{Y}$ may be shown to be normal with variance $\hat{\sigma}^2 = \frac{1}{\tau+n}$ and (discounted) mean $\hat{\mu} = n\overline{Y}\hat{\sigma}^2 = \frac{n}{n+\tau}\overline{Y}$. The PP $P(\mu > 0) = \Phi(\frac{\hat{\mu}}{\hat{\sigma}}) = \Phi(\frac{n\overline{Y}}{\sqrt{\tau+n}})$, where Φ is the cumulative distribution function of the standard normal distribution. Now compare this to the situation where $\tau = 0$, i.e., where the prior for μ is flat indicating no prior information. In that case $\hat{\mu} = \overline{Y}$ and $\hat{\sigma}^2 = \frac{1}{n}$ and $P(\mu > 0) = \Phi(\overline{Y}\sqrt{n})$. What is the effective skepticism in the estimate of μ with $\tau > 0$ compared to the undiscounted estimate? To answer this question, suppose that the discounted analysis had used a sample of size m. What is the value of m that would yield the same PP of $\mu > 0$ as an undiscounted analysis for a lower sample size n? Assume $\overline{Y}$ doesn't change and set $\overline{Y}\frac{m}{\sqrt{\tau+m}} = \overline{Y}\sqrt{n}$ so that $\frac{m}{\sqrt{\tau+m}} = \sqrt{n}$. Solving for m yields $2m = n + \sqrt{n^2 + 4n\tau}$ from which one obtains the increment in sample size $m - n$ needed to achieve the same level of evidence discounted by a skeptical prior as compared with the evidence from a flat prior. This is depicted in Figure 7 as a function of the prior precision τ ($\frac{1}{\sigma^2}$).

```
z ← list()
n ← seq(1, 100, by=2)
for(tau in c(.01, .25, 1, 2, 4, 10, 20))
  z[[paste0('tau=', tau)]] ←
        list(x=n, y=0.5 * (n + sqrt(n^2 + 4 * n * tau)) - n)

labcurve(z, pl=TRUE, xlab='Sample Size With No Skepticism',   # Fig. 7
         ylab='Extra Subjects Needed Due to Skepticism')
```

Moderate skepticism effectively discards around 5 subjects, so one can readily see that the Bayesian power is not ruined by skepticism once the actual sample size is moderate. The skepticism eventually wears off, with the data likelihood overwhelming the prior. Contrast this with frequentist multiplicity adjustments, which never wear off as $n \uparrow$.

An optimistic prior (e.g., derived from data on adults where efficacy was positive) can be thought of as providing additional observations to the analysis of children.

3.3 Contrasting Frequentist and Bayesian Evidence and Errors

Suppose that in a fixed sample size study a single endpoint is compared for treatment A vs. B. Let E denote the true unknown efficacy measure with $E = 0$ indicating exactly zero difference. The frequentist approach attempts to show that the data are implausible under the assumption that $E = 0$, and does not make any probability statement about E. E is either zero or not. The Bayesian approach makes probability statements about the unknown E by computing PPs. In the vast majority of Bayesian models E is continuous and the probability that E exactly

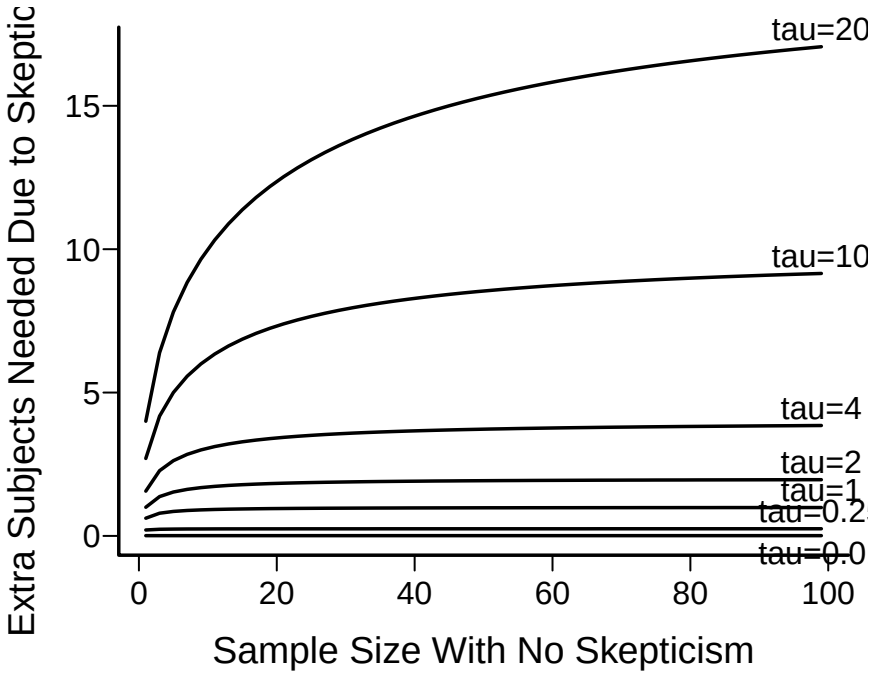


Figure 7: Effect of discounting by a skeptical prior with mean zero and precision τ: the increase needed in the sample size in order to achieve the same posterior probability of $\mu > 0$ as with the flat (non-informative) prior. $\tau = 20$ corresponds to a very skeptical prior.

equals any specific value e is zero. If $E > 0$ denotes benefit of treatment B, Bayesian PPs are often of the form $P(E > c|\text{data})$. If c is zero, the Bayesian inference provides evidence for *any* efficacy. If $c > 0$ then evidence is being quantified for efficacy greater than some, usually non-trivial, amount. Just as in forecasting the chance of rain tomorrow, there are no Bayesian "errors" per se; there are just small PPs for something that turned out to happen or larger probabilities for things that didn't. In probabilistic thinking one might say that the only real errors are assigning a probability of exactly zero to something that happened or a probability of exactly 1.0 to something that didn't. More errors are made by decision makers who are in the difficult position of having to act on the probabilities when their actions are constrained to be all–or–nothing. Decision making under uncertainty is best done using probabilistic thinking, unless a loss/utility/cost function is available for optimization using Bayesian decision analysis.

The following examples contrast evidential measures and errors for the two paradigms.

Design:

Frequentist: Design the study to have $\alpha = 0.05, \beta = 0.1$. Once data are available, these are not relevant because they are long-run operating characteristics about a sequence of trials and do not apply to the current trial. α depends on (often unknown) intentions while β depends on a single parameter value (efficacy). You can also use a frequentist design to yield a specified confidence interval width if the sample size is fixed.

Bayesian: Choose a prior and design the study to have a 0.95 credible interval of a specified width or smaller, or to have a proportion > 0.9 of simulated trials such that $P(E > c) > 0.95$ for a pre-specified c.

Type of errors:

Frequentist: Type I assertion probability α: prob. of declaring efficacy when $E = 0$
Type II error: prob. of failing to declare efficacy when $E = c$ for some particular $c > 0$
Prob. of asserting efficacy never drops no matter how $N \uparrow$ since we usually fix α

Bayesian: PP=$P(E > c|\text{data})$
If judge efficacious, chance of an error is 1-PP
If judge ineffective, chance of an error is PP

$p = 0.03$**:**

Frequentist: Conclude efficacy. This is either right or wrong; no probability is associated with the true unknown E.
Interpretation: If $E = 0$ and one ran a series of identical trials, one would see an observed *estimate* of E as large or larger than that observed 0.03 of the time.

Bayesian: PP is its own error probability

$p = 0.2$**:**

Frequentist: Can't conclude $E = 0$ but fail to have evidence for $E \neq 0$. No measure of $P(E = 0)$ is available.

Bayesian: Simple PP of no effect or harm: $P(E < 0)$

Clinical significance:

Frequentist: With large N, trivial effect can yield $p < 0.05$

Bayesian: Compute PP that the true effect is more than trivial

$p = 0.04$**, 5 other trials "negative":**

Frequentist: No way to take the other 5 trials into account other than using non-quantitative subjective arguments

Bayesian: Skepticism about efficacy in the current treatment setting would already be captured in the prior; otherwise the other trials could be used as a prior or a Bayesian hierarchical model could be used to borrow their information.

3.4 Problems Caused by Use of Arbitrary Thresholds

Much has been written about the problems of using arbitrary thresholds for "statistical significance" in frequentist NHST [31]. Though Bayesian posterior probabilities would improve inference in many ways, similar problems could arise were an arbitrary cutoff be placed on PPs. Science as well as regulatory actions have been damaged by thresholding. Once it is known whether or not an evidence measure exceeds the declared threshold, conclusions tend to be stated as if there is no uncertainty [29, 4]. Imagine how a more honest accounting of evidence could result in greater objectivity with less arbitrariness by considering a sentence that carries along a PPs in parentheses: Treatment B probably (0.94) resulted in lower blood pressure and was probably (0.78) safer in comparison with treatment A[10].

[10] A more radical idea would be to have the font size of "was better" proportional to the PP in "Treatment B was better than treatment A."

4 Multiplicity

Consternation from both statisticians and clinicians about how to handle multiplicities exposes weaknesses in the frequentist approach. As mentioned above, there are no statistical principles that lead to unique frequentist solutions for multiple comparison problems. When considering adaptive clinical trials or sample size re-estimation, the problems magnify.

4.1 Frequentist

It is well known that the more hypotheses that are all false are tested, the greater the chance of positive assertions of effects increases. The frame of reference for what constitutes "hypotheses" is not clear. Does it include hypotheses in other studies the investigator may happen to be involved in? Does it include all patient subgroups, endpoints, and study monitoring looks? The frequentist approach considers the *sample space* in inference, which must take into account hypotheses that *might have been tested* in addition to those that *were tested*, in violation of the *likelihood principle*, which states that under the chosen statistical model, all of the evidence in a sample relevant to model parameters is contained in the likelihood function [8]. [11]

Consider a 4-treatment study with treatments denoted by A B C D. A frequentist assessment of A vs. B frequently is discounted because C was compared to D. Next consider a clinical trial monitored using a group sequential α-spending method. An early look at the data is discounted because of planned future looks. Later looks are discounted for earlier inconsequential looks. In unblinded sample size re-estimation, the first wave of data must be discounted to preserve the overall α level at the end of the extended study. None of these multiplicity adjustments are satisfactory from a scientific viewpoint [9].

Frequentist multiplicity adjustments are always ad hoc.

4.2 Bayesian

Bayesian PPs are well calibrated no matter what type of or how many multiplicities are present. Skepticism about an effect is focused on the effect of interest, not other effects tested. The current posterior density is an accurate reflection of study evidence at any point in time. Bayesian inference obeys the likelihood principle. The data and not the context for the data are important for inference[12]. The benefits of not dwelling on the sample space of contemplated experiments but instead using the likelihood principle cannot be overstated. Frequentist significance testing deals with "what would have occurred following results that were not observed at analyses that were never performed" [24]. The probability of a test statistic as or more extreme than an observed value depends on all samples that *might have arisen*, whereas Bayes uses only the sample that *has arisen*. To limit the sample space (to for example limit α) there must be more planning and less flexibility.

In the A B C D treatment study, Bayesian inference for A vs. B is not discounted because C was compared to D. A vs. B is discounted only because of prior information for how A might compare to B. In a sequential trial, the current PP is self-contained, well calibrated, and meaningful when taken out of context of the number of data looks or the stopping rule. As mentioned earlier,

[11] The "paradox of two sponsors" illustrates how frequentist statistics violates this principle. Suppose that sponsor 1 has designed the study for one interim look, choosing an α cutoff of 0.047 at the second analysis to preserve the overall type I assertion probability at $\alpha = 0.05$. The sponsor conducted an inconsequential interim analysis and now comes with a final dataset with a p-value of 0.049 so does not receive approval for the treatment. Sponsor 2 comes with identical data but did not conduct an interim analysis so pays no α-spending penalty, resulting in $p = 0.049$, significance at $\alpha = 0.05$, and an approved treatment. This cannot make sense.

[12] For example, the likelihood principle asserts that the inference about the population probability θ of an event is identical whether one samples 20 patients and counts 5 events or one enrolls patients until 5 events have occurred and this happened to require 20 patients. The first situation involves the binomial distribution and the second the negative binomial distribution. In either case the likelihood of the data is $\theta^5(1-\theta)^{15}$ yet the frequentist approach would obtain two conflicting confidence intervals for θ.

cheating around multiple looks is only possible when a study is extended, less promising results are obtained, and the new data are suppressed.

To get a better sense of why repeated looks do not distort the meaning of PPs, consider a probabilistic pattern recognition system for identifying enemy targets in combat. Suppose the initial assessment when the target is distant is a probability of 0.3 of being an enemy vehicle. Upon coming closer the probability rises to 0.8. Finally the target is close enough (or the air clears) so that the pattern analyzer estimates a probability of 0.98. The fact that the probabilty was < 0.98 earlier is of no consequence as the gunner prepares to fire a canon. Even though the probability may actually decrease while the shell is in the air due to new information, the probability at the time of firing was completely valid based on then available information.

> In the frequentist world, multiplicity comes from the chances you give data to be extreme, not the chances you give true effects to exist.

5 Posterior Probabilities With Sequential Analysis

> (In a Bayesian analysis) It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience.
>
> Edwards, Lindman, Savage (1963) [23]

With sequential testing and early study termination, the current point estimate is promoted or pulled back by the prior, providing perfect calibration. Bayesian inference can only go wrong because of an incorrectly specified data model (which hurts frequentist inference alike) or because the prior distribution used in the analysis is in conflict with the prior used by the study's judge.

What if the efficacy of a treatment were assessed at will and the study terminated the first time that a PP of efficacy exceeded 0.95? The PP at this point will on average be above 0.95, but remains perfectly calibrated for a reviewer *as long as she does not substitute a different prior during the review than was used during the analysis*. In the following simple simulation example, a one-arm study has a maximum sample size of $N = 500$ subjects, and efficacy is assessed after each subject, resulting in 500 data looks. The efficacy measurement is assumed to have a normal distribution with standard deviation (SD) 1.0. Efficacy corresponds to the mean subject response μ being greater than zero.

5.1 Skeptical Prior

Though the choice of prior has no impact on the PP calibrations shown below, let's use a skeptical prior that favors no effect, allows for harm to be as likely as benefit, and places a low probability on a large effect. Specifically the prior is a 1:1 mixture of two normal distributions each having mean zero. The SD of the first distribution is chosen so that $P(\mu > 1) = 0.1$, and the SD of the second distribution is chosen so that $P(\mu > 0.25) = 0.05$. The prior density is shown in Figure 8. Mixtures of normals are one of many good approaches for bringing skepticism about large treatment effects into the final study interpretation [59][13].

```
sd1 ← 1    / qnorm(1 - 0.1)
sd2 ← 0.25 / qnorm(1 - 0.05)
```

```
wt  ← 0.5   # 1:1 mixture
pdensity ← function(x) wt * dnorm(x, 0, sd1) + (1 - wt) * dnorm(x, 0, sd2)
x ← seq(-3, 3, length=200)   # Figure 8
```

[13] Mixtures of distributions can provide a formal way to handle pertinent historical data. The prior could be a 3:1 mixture of a non-informative or somewhat skeptical distribution and the posterior distribution from a completed study if experts believed that the completed study was $\frac{1}{4}$ applicable to the new setting [53, 6].

```
plot(x, pdensity(x), type='l', xlab=expression(mu), ylab='Prior Density')
```

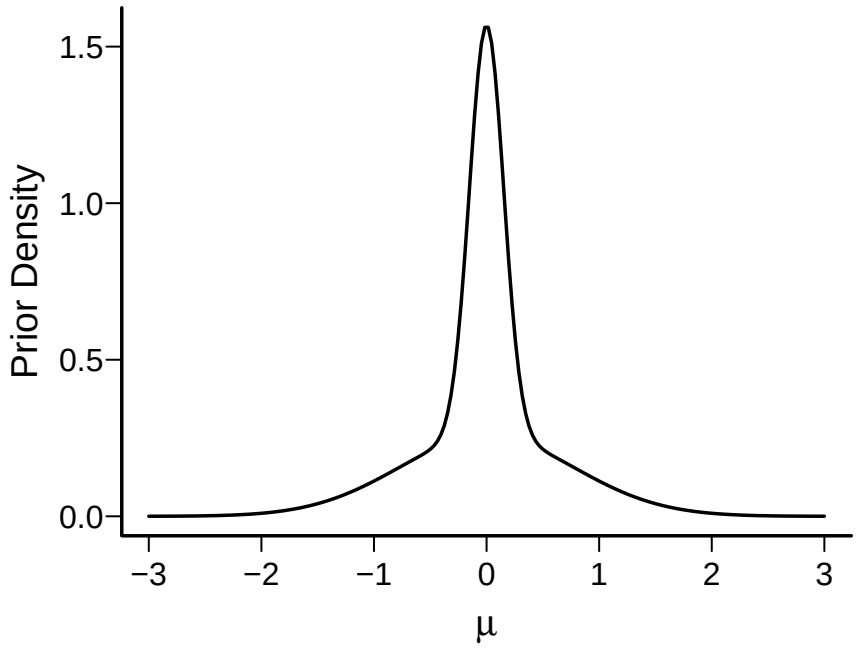


Figure 8: Skeptical prior distribution for the unknown mean in a single arm study. This is a 1:1 mixture of zero mean normals with SD=0.780 and 0.152 respectively.

Next simulate 10,000 studies to reflect the real-world case in which the true treatment effect is not known. For each study a single value of μ is sampled from the above prior distribution. Then 500 data values are simulated from a normal distribution having mean μ and SD=1. The 500 values are revealed one-at-a-time so that data look j has a sample size of j subjects.

```
simseq ← function(N, prior.mu=0, prior.sd, wt, mucut=0, mucutf=0.05,
                  postcut=0.95, postcutf=0.9,
                  ignore=20, nsim=1000) {
  prior.mu ← rep(prior.mu, length=2)
  prior.sd ← rep(prior.sd, length=2)
  sd1 ← prior.sd[1]; sd2 ← prior.sd[2]
  v1 ← sd1 ^ 2
  v2 ← sd2 ^ 2
  j ← 1 : N
  cmean ← Mu ← PostN ← Post ← Postf ← postfe ← postmean ← numeric(nsim)
  stopped ← stoppedi ← stoppedf ← stoppedfu ← stopfe ← status ←
    integer(nsim)
  notignored ← - (1 : ignore)

  # Derive function to compute posterior mean
  pmean ← gbayesMixPost(NA, NA, d0=prior.mu[1], d1=prior.mu[2],
                                v0=v1, v1=v2, mix=wt, what='postmean')

  for(i in 1 : nsim) {
    # See http://stats.stackexchange.com/questions/70855
    component ← if(wt == 1) 1 else sample(1 : 2, size=1, prob=c(wt, 1. - wt))
    mu ← prior.mu[component] + rnorm(1) * prior.sd[component]
    # mu ← rnorm(1, mean=prior.mu, sd=prior.sd) if only 1 component

    Mu[i] ← mu
    y  ← rnorm(N, mean=mu, sd=1)
```

```
    ybar ← cumsum(y) / j     # all N means for N sequential analyses
    pcdf ← gbayesMixPost(ybar, 1. / j,
                         d0=prior.mu[1], d1=prior.mu[2],
                         v0=v1, v1=v2, mix=wt, what='cdf')
    post  ← 1 - pcdf(mucut)
    PostN[i] ← post[N]
    postf ← pcdf(mucutf)
    s ← stopped[i] ←
      if(max(post) < postcut) N else min(which(post ≥ postcut))
    Post[i]  ← post[s]    # posterior at stopping
    cmean[i] ← ybar[s]    # observed mean at stopping
    # If want to compute posterior median at stopping:
    #    pcdfs ← pcdf(mseq, x=ybar[s], v=1. / s)
    #    postmed[i] ← approx(pcdfs, mseq, xout=0.5, rule=2)$y
    #    if(abs(postmed[i]) == max(mseq)) stop(paste('program error', i))
    postmean[i] ← pmean(x=ybar[s], v=1. / s)

    # Compute stopping time if ignore the first "ignore" looks
    stoppedi[i] ← if(max(post[notignored]) < postcut) N
    else
      ignore + min(which(post[notignored] ≥ postcut))

    # Compute stopping time if also allow to stop for futility:
    # posterior probability mu < 0.05 > 0.9
    stoppedf[i] ← if(max(post) < postcut & max(postf) < postcutf) N
    else
      min(which(post ≥ postcut | postf ≥ postcutf))

    # Compute stopping time for pure futility analysis
    s ← if(max(postf) < postcutf) N else min(which(postf ≥ postcutf))
    Postf[i] ← postf[s]
    stoppedfu[i] ← s

    ## Another way to do this: find first look that stopped for either
    ## efficacy or futility.  Record status: 0:not stopped early,
    ## 1:stopped early for futility, 2:stopped early for efficacy
    ## Stopping time: stopfe, post prob at stop: postfe

    stp ← post ≥ postcut | postf ≥ postcutf
    s ← stopfe[i] ← if(any(stp)) min(which(stp)) else N
    status[i] ← if(any(stp)) ifelse(postf[s] ≥ postcutf, 1, 2) else 0
    postfe[i] ← if(any(stp)) ifelse(status[i] == 2, post[s],
                                    postf[s]) else post[N]
  }
  list(mu=Mu, post=Post, postn=PostN, postf=Postf,
       stopped=stopped, stoppedi=stoppedi,
       stoppedf=stoppedf, stoppedfu=stoppedfu,
       cmean=cmean, postmean=postmean,
       postfe=postfe, status=status, stopfe=stopfe)
}
```

```
set.seed(3)
z ← simseq(500, prior.mu=0, prior.sd=c(sd1, sd2), wt=wt, postcut=0.95,
           postcutf=0.9, nsim=10000)
mu      ← z$mu
post    ← z$post
postn   ← z$postn
st      ← z$stopped
sti     ← z$stoppedi
stf     ← z$stoppedf
stfu    ← z$stoppedfu
cmean   ← z$cmean
postmean← z$postmean
postf   ← z$postf
rmean ← function(x) formatNP(mean(x), digits=3)
```

In Figure 9 is shown the relationship between the PP of efficacy at the conclusion of the study and the posterior computed at the time of stopping early for efficacy (or final posterior if no early stopping).

```
cuts ← c(1, 5, 10, 25, 50, 100, 250, 500, 1000, 3000, 4000)
ggfreqScatter(postn, post, xlab='Posterior at Study End', cuts=cuts,
              ylab='Posterior Upon Stopping')   # Fig. 9
```

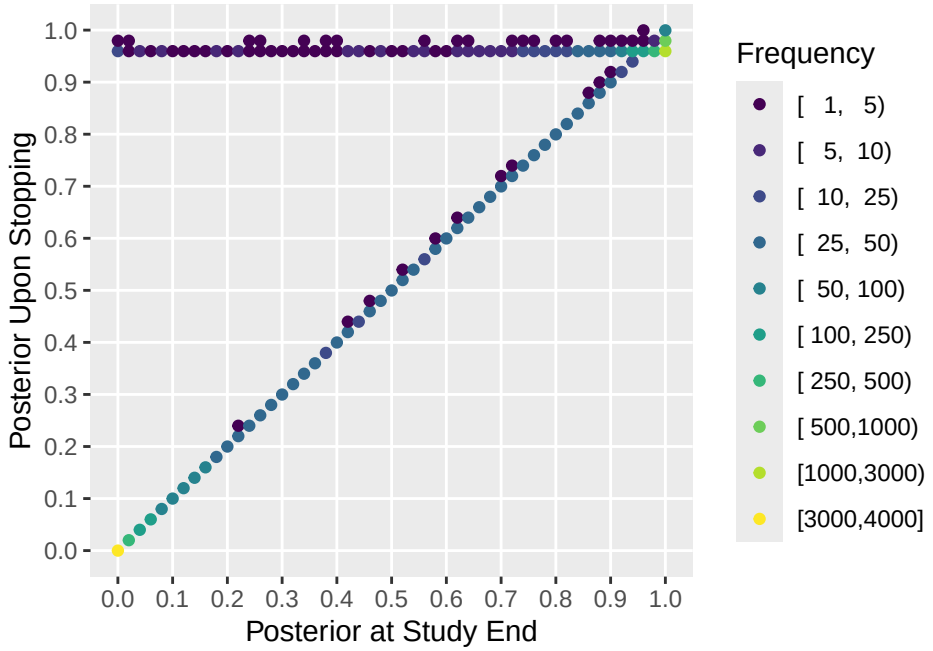


Figure 9: Scatterplot of posterior $P(\mu > 0|y)$ at final assessment after all 500 subjects vs. posterior at stopping for efficacy. Points were binned using a 50 × 50 grid, with frequency of simulated trials indicated by colors. The prior in Figure 8 was used.

The proportion of trials stopped early with PP ≥ 0.95 for which the final PP after all 500 analyses was < 0.7 is 0.039. This could be taken as estimating the probability of being misled by early looks.

To see how the PPs are calibrated against the true probability of efficacy, the *loess* smoother is used to relate final PPs to the binary variable indicating that the true $\mu > 0$ (Figure 10).

```
v ← val.prob(postn, mu > 0, m=400, logistic.cal=FALSE,   # Fig. 10
        xlab=expression(paste('Posterior Probability ', mu > 0, ' at Study End'
             )),
        ylab=expression(paste('Proportion of Trials with ',  mu > 0)))
```

```
hist(mu[post ≥ 0.95], nclass=50, xlim=c(-1,4),
     xlab=expression(mu), main='')   # Fig. 11
abline(v=0, col='red', lwd=0.5)
k ← post ≥ 0.95
regret ← mean(mu[k] ≤ 0)
text(-0.5, 500, paste0('Proportion regret=', round(regret, 3)), srt=90)
```

4321 of 10,000 trials were stopped early for efficacy with a PP ≥ 0.95. Of these, 169 actually had $\mu \leq 0$ (proportion of 0.039 as shown in Figure 11). 1732 of the trials stopped before the 21^{st}

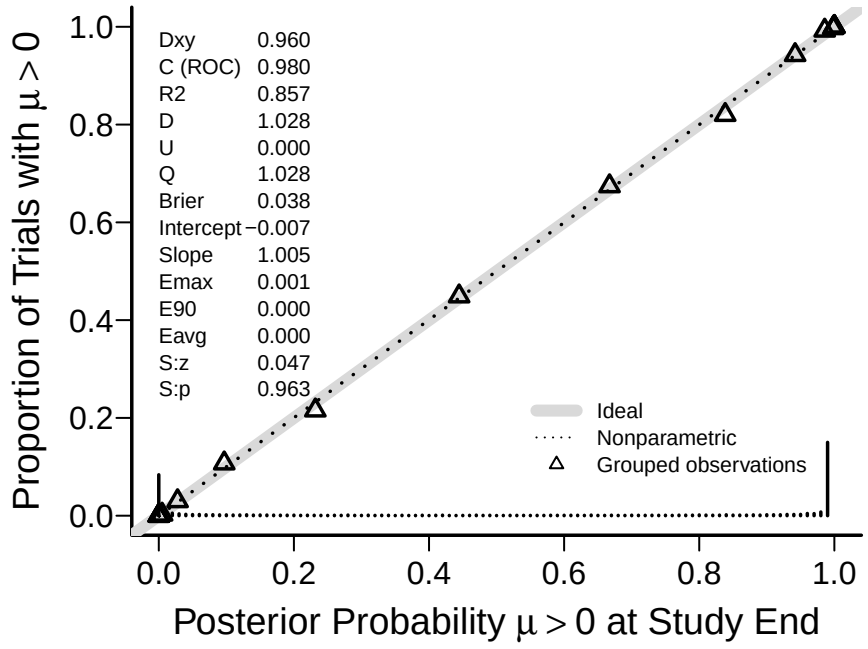


Figure 10: Calibration curve for the posterior probability of efficacy at study end, estimated using the *loess* nonparametric smoother (dotted line). Line of identity is the thick grayscale line. Simple grouped proportions based on intervals of posterior probability containing 400 trials per group are shown as triangles. Unlike with smooth nonparametric estimates, grouping must be done to get an adequate denominator for proportions. The frequency distribution of posterior probabilities is depicted with vertical line segments. The prior in Figure 8 was used.

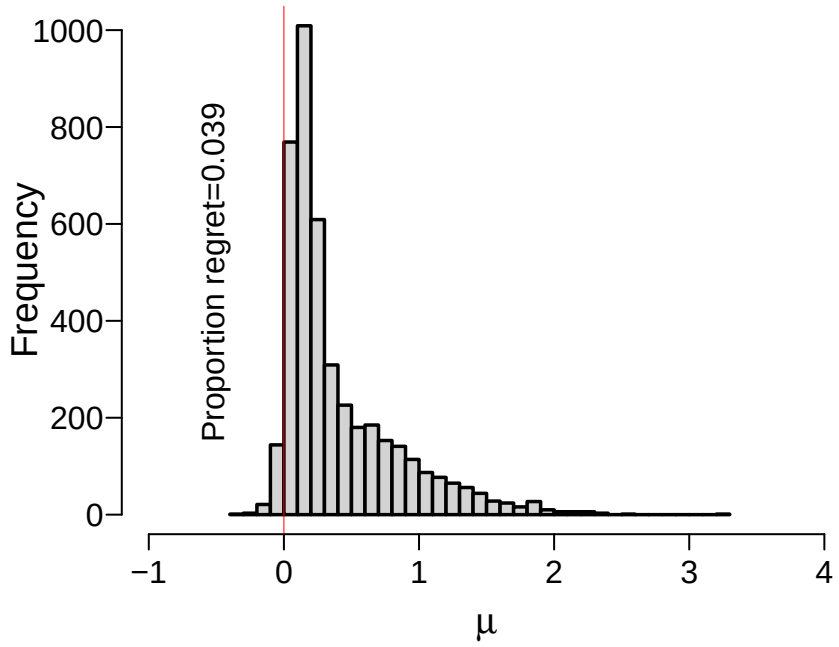


Figure 11: Frequency distribution of true values of μ when stopping early for efficacy (concluding $\mu > 0$), using prior in Figure 8

look. If we did not look before the 21^{st} subject, the number of stopped trials actually having $\mu \leq 0$ was 137.

The proportion of trials stopped with a PP ≥ 0.95 was 0.432. This is the Bayesian power of the study design. Note that in this calculation $\frac{1}{2}$ of the trials had negative efficacy, i.e., $\frac{1}{2}$ of the simulated values of μ used to simulate the trials were negative. It is important to note that to achieve a high probability of detecting a clinically important effect one would have to be optimistic, i.e., to suspend the belief that the prior distribution is symmetric about zero. This can perhaps be better dealt with by leaving the skeptical prior intact and computing the proportion of trials that were stopped early for efficacy when $\mu > 0.25$, which was 1.000. The proportion stopped early when $\mu \in [0.15, 0.20]$ was 0.987. The relationship between true μ and the stopping time can be estimated. 14 of the 10,000 studies were stopped after one observation, with a mean PP of 0.978 and a mean true value of μ of 1.257. Stopping after one observation would not be possible had the variance of the outcome variable not been known. The average sample size (and stopping time) was 318 if we treat non-stopped studies as having a sample size 500. A better approach treats non-stopped studies as having a right-censored time-to-stopping of 500; the results are in Figure 12.

```
dd ← datadist(mu); options(datadist='dd')
f ← psm(Surv(st, st < 500) ~ rcs(mu, 6), dist='lognormal')
plot(Predict(f, mu, fun=exp, conf.int=FALSE), xlim=c(-.25, 1), ylim=c(0, 500),
     ylab='Median Stopping Time',    # Fig. 12
     abline=list(list(v=0, col=gray(.9))))
```

Cost savings of the Bayesian approach is even more obvious when one tests continually for futility. Let's define futility as a posterior $P(\mu < 0.05) \geq 0.9$. The number of trials stopped early for either efficacy or futility was 9744, with a mean sample size at stopping of 65. The number of trials that were stopped early for futility, ignoring efficacy, is 5884. The mean sample size at stopping only for futility was 235.

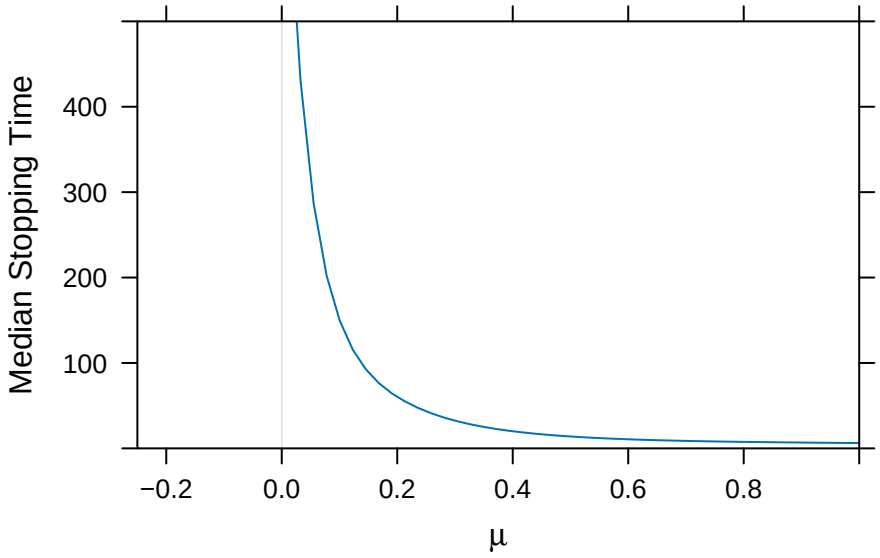


Figure 12: Estimated median stopping time for efficacy as a function of true value of μ, using a log-normal survival time distribution and a restricted cubic spline in μ with 6 default knots. Studies never stopping are right censored at 500 trials. The prior in Figure 8 was used.

Next examine the relationship between (1) the sample mean and true μ at the time of stopping and (2) the posterior mean and true μ at the time of stopping, using the nonparametric regression "super smoother" [26]. These are shown in Figure 13.

```
plot(0, 0, xlab=expression(paste('Estimated ', mu)),    # Fig. 13
     ylab=expression(mu), type='n', xlim=c(-2, 4), ylim=c(-2, 4))
abline(a=0, b=1, col=gray(.9), lwd=4)
lines(supsmu(cmean, mu))
lines(supsmu(postmean, mu), col='blue')
```

It can readily be seen that the ordinary sample mean is biased high when studies are stopped early because $\overline{Y}$ is large when large true values of μ are not favored by the prior. But the posterior mean is perfectly calibrated. The same would be found for the posterior median. Had the prior distribution had heavy tails, i.e., had we believed that very large treatment effects were likely, the sample mean would not have been so biased (see Figure 19). Note that in the frequentist setting how one adjusts for bias in point effect estimates when early stopping has occurred is unclear, and proposed solutions are complex.

Returning to calibration of PPs, the proportion of trials that stopped with a PP ≥ 0.95 that actually had a true value of $\mu > 0$ was 0.961. The mean PP when stopping early was 0.960.

The proportion of trials that did not stop early that had a true value of $\mu \leq 0$ was 0.870. The mean PP $P(\mu \leq 0)$ at the end of such studies was 0.869.

From these estimates, calibration of PPs is perfect, as expected. Figure 14 shows calibration in more detail by estimating the relationship between the PP at stopping to the true $P(\mu > 0)$.

```
v ← val.prob(post, mu > 0, m=400, logistic.cal=FALSE,    # Fig. 14
         xlab=expression(paste('Posterior Probability ', mu > 0, ' Upon Stopping
```

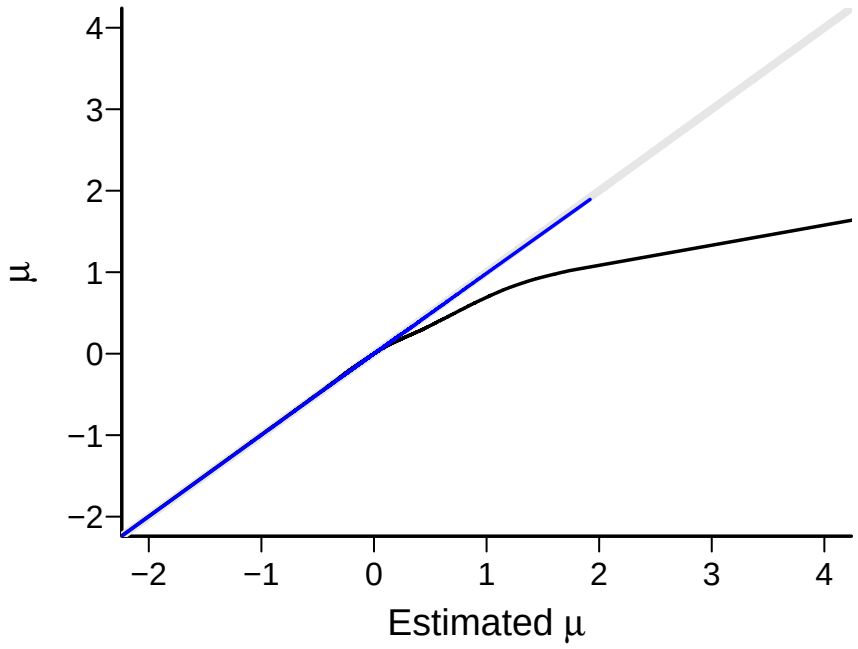


Figure 13: True μ vs. sample mean at stopping (black line) and vs. posterior mean at stopping (blue line) using the prior in Figure 8. Thick grayscale line is the line of identity.

```
        ')),
      ylab=expression(paste('Proportion of Trials with ',  mu > 0)))
```

The interpretation of PPs is independent of the stopping rule, which allows for painless unblinded sample size re-estimation as well as allowing studies to begin without preliminary data needed for sample size estimation.

If efficacy is ignored and we considered stopping early only for futility, the PP of futility at the time of stopping is well calibrated as shown in Figure 15.

```
v ← val.prob(postf, mu < 0.05, m=400, logistic.cal=FALSE,   # Fig. 15
      xlab=expression(paste('Posterior Probability ', mu < 0.05, ' Upon
           Stopping')),
      ylab=expression(paste('Proportion of Trials with ',  mu < 0.05)))
```

5.2 Flatter Prior

To understand the effect of using a flatter prior for the unknown mean μ, we now let the prior distribution be Gaussian still with mean 0 but with standard deviation of 3, with no mixing with another normal distribution. The prior is shown in Figure 16.

```
mu ← seq(-6, 6, length=200)   # Fig. 16
plot(mu, dnorm(mu, 0, 3), type='l', ylim=c(0, 1.5),
     xlab=expression(mu), ylab='Prior Density')
```

Figure 17 shows the frequency distribution of true values of
mu for the subset of studies that reached a PP of efficacy of 0.95 at any of the sequential tests.

```
set.seed(4)
```

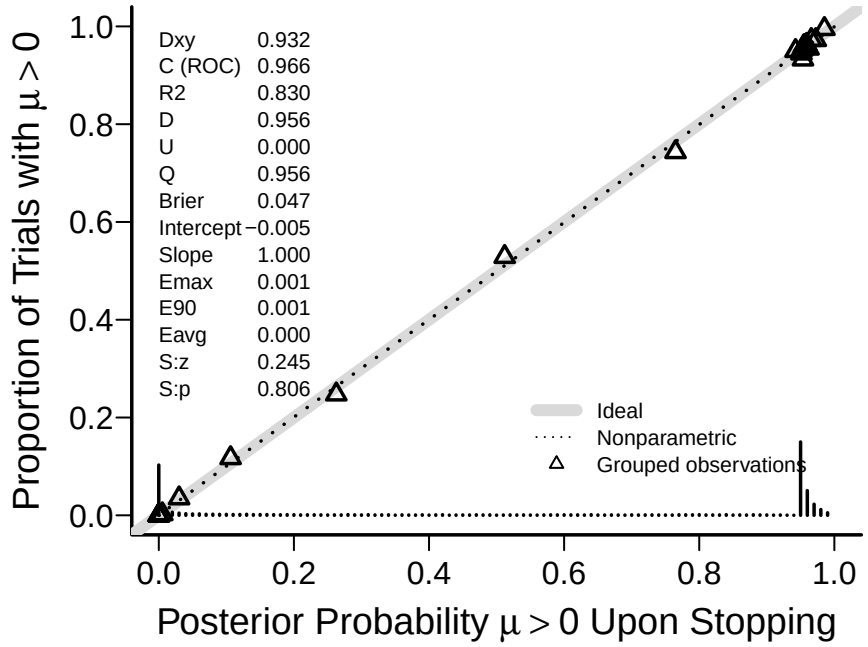


Figure 14: Calibration curve for the posterior probability of efficacy upon stopping or the posterior at the final sample size if no stopping, estimated using the *loess* nonparametric smoother (dotted line). Line of identity is the thick grayscale line. Simple grouped proportions based on intervals of posterior probability containing 400 trials per group are shown as triangles. The frequency distribution of posterior probabilities is depicted with vertical line segments. The prior in Figure 8 was used.

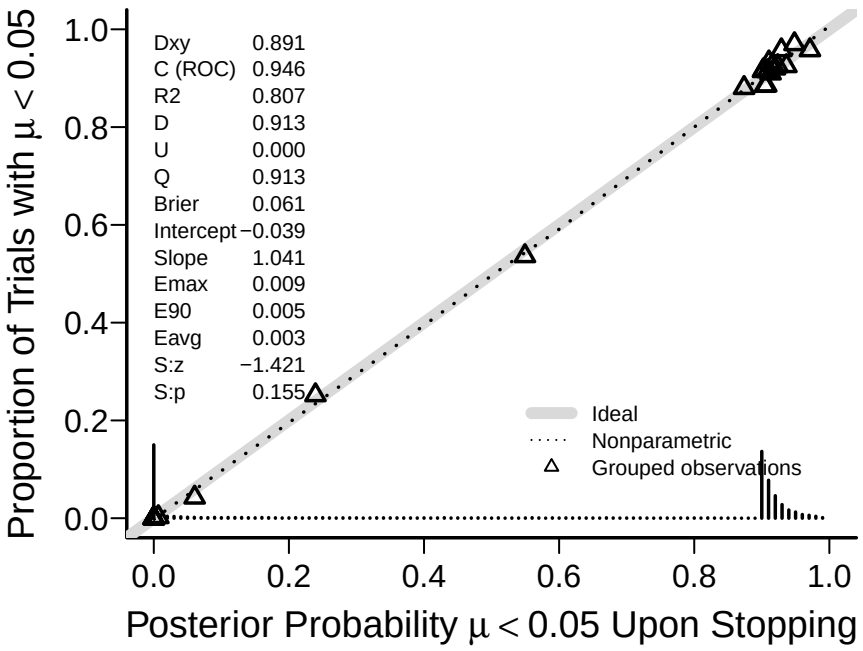


Figure 15: Calibration curve for the posterior probability of futility upon stopping for futility, or the posterior at the final sample size if no stopping for futility, estimated using the *loess* nonparametric smoother (dotted line). Line of identity is the thick grayscale line. Simple grouped proportions based on intervals of posterior probability containing 400 trials per group are shown as triangles. The frequency distribution of posterior probabilities is depicted with vertical line segments. The prior in Figure 8 was used.

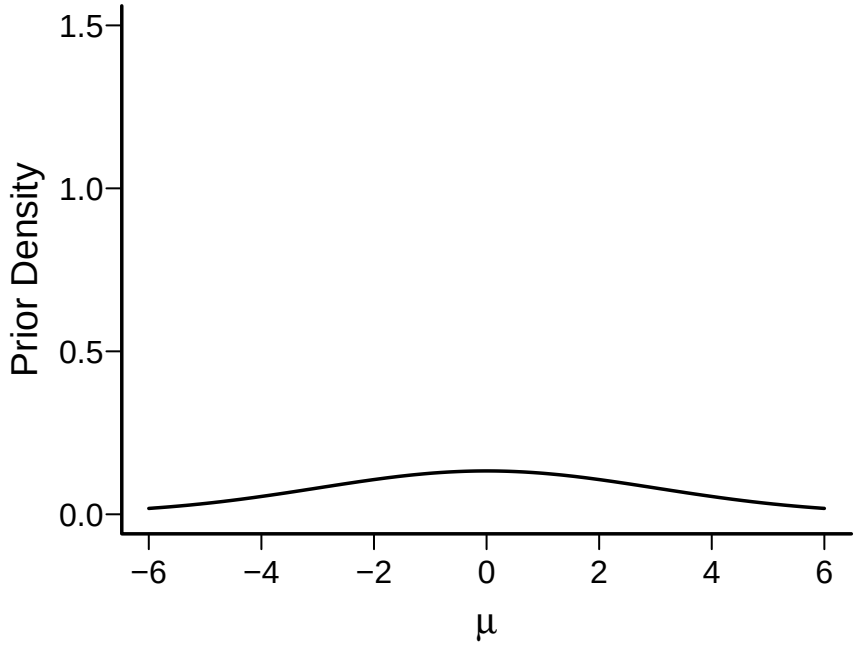


Figure 16: Flatter prior; normal with mean 0 and $\sigma = 3$, using same y-axis scale as skeptical prior.

```
z ← simseq(500, prior.mu=0, prior.sd=3, wt=1, postcut=0.95,
           postcutf=0.9, nsim=10000)
mu       ← z$mu
post     ← z$post
st       ← z$stopped
sti      ← z$stoppedi
stf      ← z$stoppedf
stfu     ← z$stoppedfu
cmean    ← z$cmean
postmean← z$postmean
postf    ← z$postf

hist(mu[post ≥ 0.95], nclass=50, # xlim=c(-1,4),
     xlab=expression(mu), main='')    # Fig. 17
k ← post ≥ 0.95
abline(v=0, col='red', lwd=0.5)
regret ← mean(mu[k] ≤ 0)
text(-.6, 150, paste0('Proportion regret=', round(regret, 3)), srt=90)
```

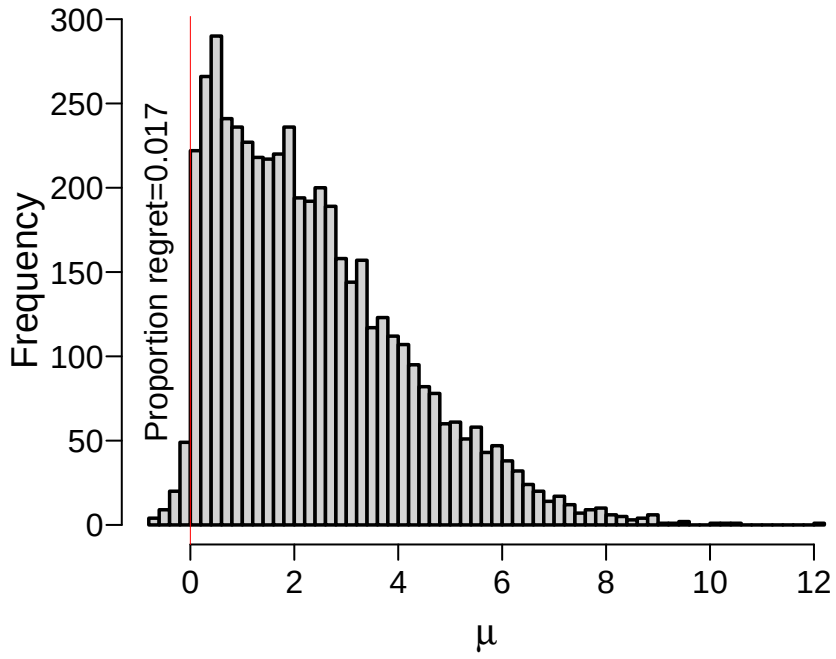


Figure 17: Frequency distribution of actual μ when using a flatter prior (Figure 16) and stopping early for efficacy.

4938 of 10,000 trials were stopped early for efficacy with a PP ≥ 0.95. Of these, 82 actually had $\mu \leq 0$. 4618 of the trials stopped before the 21^{st} look. If we did not look before the 21^{st} subject, the number of stopped trials actually having $\mu \leq 0$ was 25.

The Bayesian power, ie., proportion of trials stopped with a PP ≥ 0.95, was 0.494. The proportion of trials that were stopped early for efficacy when $\mu > 0.25$ was 1.000. The proportion stopped early when $\mu \in [0.15, 0.20]$ was 1.000. 2844 of the 10,000 studies were stopped after one observation, with a mean PP of 0.993 and a mean true value of μ of 3.342. The average sample size (and stopping time) was 257 if we treat non-stopped studies as having a sample size 500. Treating non-stopped studies as having a right-censored time-to-stopping leads to the results in Figure 18.

```
dd ← datadist(mu); options(datadist='dd')
```

```
# lognormal time to event model, log median a restricted cubic spline in true mu
f ← psm(Surv(st, st < 500) ~ rcs(mu, 6), dist='lognormal')
plot(Predict(f, mu, fun=exp, conf.int=FALSE), xlim=c(-.5, 1), ylim=c(0, 500),
     ylab='Median Stopping Time',    # Fig. 18
     abline=list(list(v=0, col=gray(.9))))
```

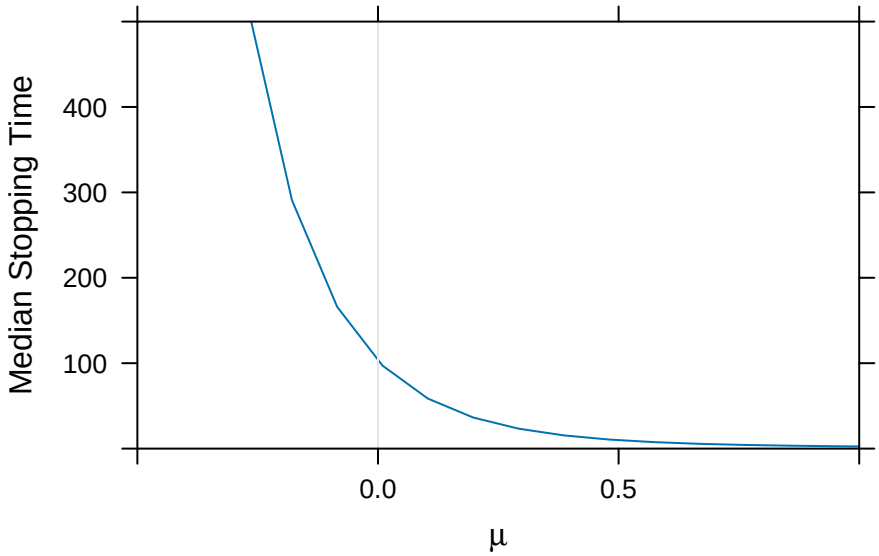


Figure 18: Estimated median stopping time for efficacy as a function of true value of μ, using a log-normal survival time distribution and a restricted cubic spline in μ with 6 default knots. The prior has mean 0, $\sigma = 3$ (Figure 16). Studies never stopping are right censored at 500 trials.

The number of trials stopped early for either efficacy or futility was 9994, with a mean sample size at stopping of 5. The number of trials that were stopped early for futility, ignoring efficacy, is 5289. The mean sample size at stopping only for futility was 239.

Figure 19 shows the relationships between estimated and actual μ. Since the prior now allows for a wider distribution for μ, the sample mean upon stopping early is less biased than with the more skeptical prior.

```
plot(0, 0, xlab=expression(paste('Estimated ', mu)),    # Fig. 19
     ylab=expression(mu), type='n', xlim=c(-2, 4), ylim=c(-2, 4))
abline(a=0, b=1, col=gray(.9), lwd=6)
lines(supsmu(cmean, mu))
lines(supsmu(postmean, mu), col='blue')
```

Returning to calibration of PPs, the proportion of trials that stopped with a PP ≥ 0.95 that actually had a true value of $\mu > 0$ was 0.983. The mean PP when stopping early was 0.985.

The proportion of trials that did not stop early that had a true value of $\mu \leq 0$ was 0.992. The mean PP $P(\mu \leq 0)$ at the end of such studies was 0.991.

From these estimates, calibration of PPs is perfect, as expected. Figure 20 shows calibration in more detail by estimating the relationship between the PP at stopping to the true $P(\mu > 0)$.

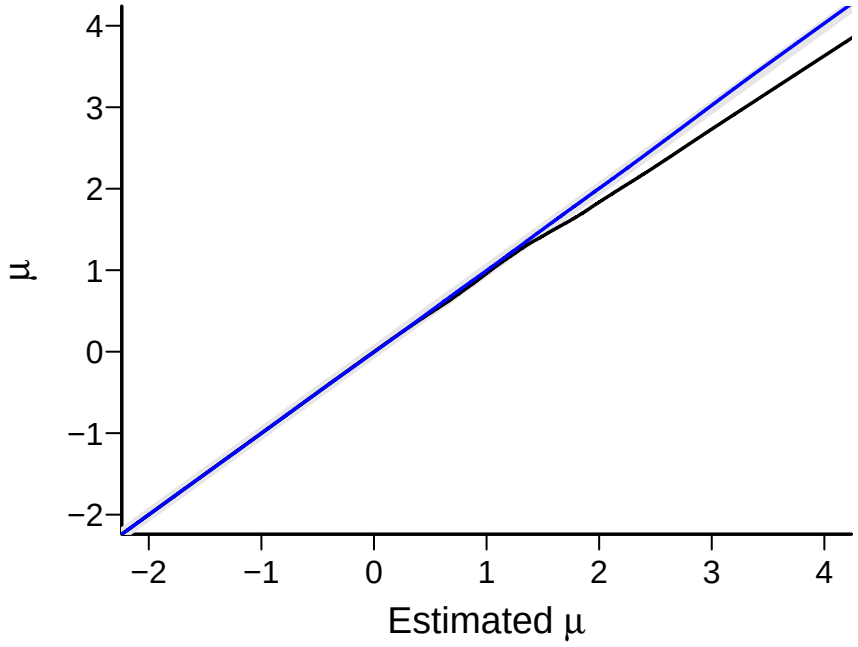


Figure 19: True μ vs. sample mean at stopping (black line) and vs. posterior mean at stopping (blue line) when the prior (Figure 16) is flatter. Thick grayscale line is the line of identity.

```
v ← val.prob(post, mu > 0, m=400, logistic.cal=FALSE,    # Fig. 20
         xlab=expression(paste('Posterior Probability ', mu > 0, ' Upon Stopping
              ')),
         ylab=expression(paste('Proportion of Trials with ',  mu > 0)))
```

If efficacy is ignored and we considered stopping early only for futility, the posterior probability of futility at the time of stopping is well calibrated as shown in Figure 21.

```
v ← val.prob(postf, mu < 0.05, m=400, logistic.cal=FALSE,    # Fig. 21
         xlab=expression(paste('Posterior Probability ', mu < 0.05, ' Upon
              Stopping')),
         ylab=expression(paste('Proportion of Trials with ',  mu < 0.05)))
```

6 Posterior Probabilities in Decision Making

The optimum Bayes decision is the one that optimizes expected utility/loss/cost. This decision is a function of the posterior distribution and the utility function. The utility function is very difficult to specify, and there is likely as much disagreement about utilities among reviewers as there is disagreement about likely efficacy. In practice, reviewers will make decisions on the basis of PPs, taking into account their own utilities in an informal way. For example, a new treatment for an incurable disease may be judged more liberally than when there are already five effective drugs on the market for the disease. "Hard" endpoints such as death may be judged more liberally than "soft" endpoints such as quality of life. Evidence for *any* reduction in mortality may be judged more strictly than evidence for a major reduction.

Though some would prefer to have hard cutoffs on PPs for demonstration of efficacy, we prefer not to entertain prescribed cutoffs because that would prevent utilities related to the points just discussed from being used. That being said, the following table provides examples of how PPs

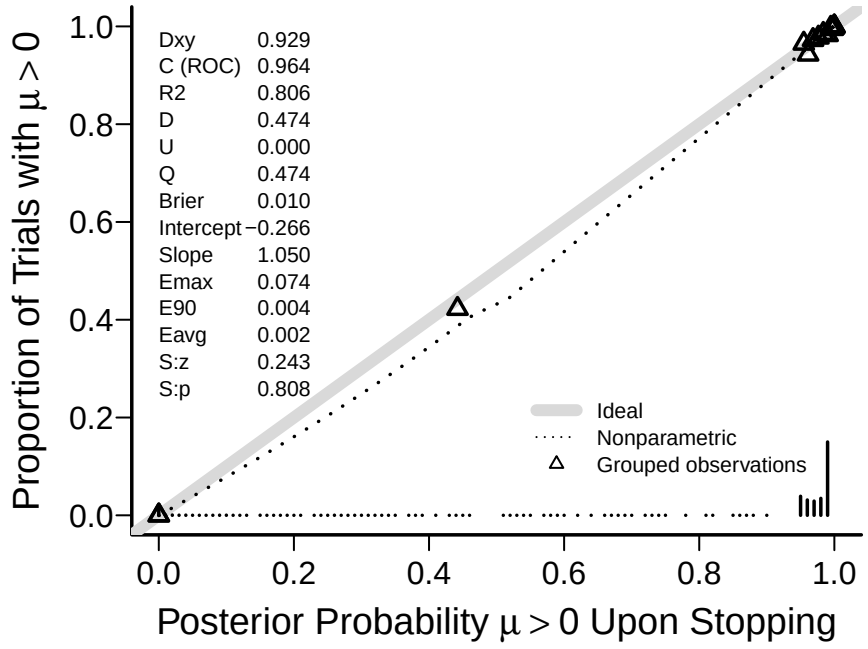


Figure 20: Calibration curve for the posterior probability of efficacy upon stopping or the posterior at the final sample size if no stopping, estimated using the *loess* nonparametric smoother (dotted line). Line of identity is the thick grayscale line. Simple grouped proportions based on intervals of posterior probability containing 400 trials per group are shown as triangles. The frequency distribution of posterior probabilities is depicted with vertical line segments. The prior is normal with mean 0 $\sigma = 3$ as shown in Figure 16.

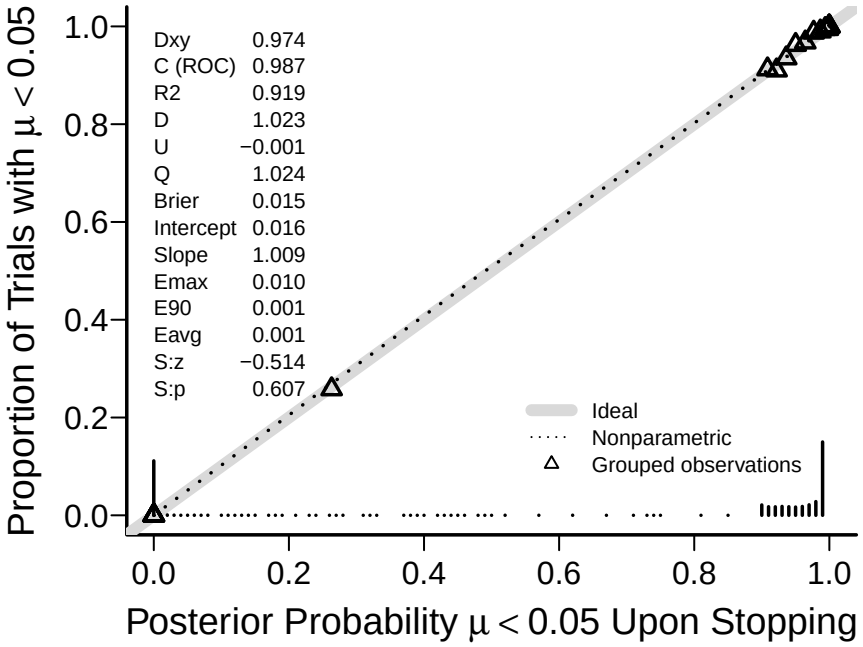


Figure 21: Calibration curve for the posterior probability of futility upon stopping for futility, or the posterior at the final sample size if no stopping for futility, estimated using the *loess* nonparametric smoother (dotted line). A flatter prior (Figure 16) is used. Line of identity is the thick grayscale line. Simple grouped proportions based on intervals of posterior probability containing 400 trials per group are shown as triangles. The frequency distribution of posterior probabilities is depicted with vertical line segments.

might be considered in decision making, taking into account only some of the types of utilities relevant in practice. In the table, λ represents the true treatment B : A hazard ratio in a time to death analysis, LVEF is left ventricular ejection fraction, Δ is the true reduction in mean LVEF. For multiple outcomes see also the next section.

Indication or Harm	Posterior Probability
Mortality reduction	$P(\lambda < 1) \geq 0.95$ or $P(\lambda < 0.8) \geq 0.8$
Any mortality reduction or large improvement in LVEF	$P(\lambda < 1) \geq 0.9$ or $P(\Delta > 0.15) \geq 0.95$
Mortality Increase	$P(\lambda > 1) \geq 0.9$
Mortality reduction in in a Phase 2 trial	$P(\lambda < 1) \geq 0.8$
Major improvement in one target or improvement in any 3 of 5 targets	P(specific target $\downarrow$ 20%) ≥ 0.95 **or** $P(\geq 3$ targets improved any) ≥ 0.95

7 Multiple Outcomes and Totality of Evidence

Because Bayes provides a direct probability for an assertion of interest, the assertion can be a compound one involving multiple patient outcomes. In this way one can make an overall evidentiary statement about multiple efficacy parameters as well as compute PPs simultaneously involving efficacy and safety. The following list contains examples of conditions for which a single PP may easily be calculated. When using a simulation technique to make draws of efficacy and safety parameters from all of the posterior distributions, computation of the overall PP is as simple as computing the fraction of posterior draws satisfying one of the conditions below as done in Section 8.

Type of Assessment	Assertion/Condition
Efficacy	Mean blood pressure $\downarrow$ 5 mmHg **or** exercise time $\uparrow$ 4m
	Mean blood pressure $\downarrow$ 5 mmHg **and** exercise time $\uparrow$ 4m
	(Any mortality $\downarrow$ and exercise time $\uparrow$ 4m) **or** mortality $\downarrow >$ 0.02
	Improvement in any two of blood pressure, exercise time, LV function, or need for diuretics
Efficacy or non-inferiority	Mortality $\downarrow$ **or** (mortality $\uparrow$ by < 0.02 and blood pressure $\uparrow$ by < 3 mmHg)
Risk/benefit	Incidence of stroke $\downarrow$ **and** significant bleeding $\uparrow$ by factor < 1.1

One could imagine a clinical trial with 5 endpoints where success is declared if the probability of hitting any two of them is greater than 0.95. This could be a more honest way to deal with the fact that one seldom has an unarguable unique clinically or patient-guided list of endpoints, especially in view of the compromises made in the choice of endpoints due to statistical power considerations. The bar can be set higher by making the five targets be non-null targets, i.e., clinically non-trivial improvements in patient outcomes.

7.1 Example: Acute Treatment of Migraine

In the FDA Center for Drug Evaluation and Research's draft guidance for industry *Migraine: Developing Drugs for Acute Treatment*, October 2014, it is stated that

> "... approval of drugs for the acute treatment of migraine involved the demonstration of an effect on 4 co-primary endpoints: pain, nausea, photophobia, and phonophobia. More recently, approval based on an effect on headache pain and nausea as co-primary endpoints has been considered. An alternative approach would consist of having patients prospectively identify their most bothersome migraine-associated symptom in addition to pain. Using this approach, the two co-primary endpoints would be (1) having no headache pain at 2 hours after dosing and (2) a demonstrated effect on the most bothersome migraine-associated symptom at 2 hours after dose. Regardless of the associated symptom identified as the most bothersome, al three important migrained-associated symptoms (i.e., nausea, photophobia, and phonophobia) should be assessed as secondary endpoints."

Bayesian PPs can be easily computed for any combination of the four symptoms, and can even handle the situation where the patient is unsure of which symptom is the most bothersome, by attaching patient-provided probabilities of bothersomeness. Here are some example PPs that can be calculated for migraine trials. Let A, B, C, D denote the events of hitting the 4 targets, respectively. Let Y_i^B denote a binary outcome for patient i achieving relief from her most bothersome symptom within 2 hours and target E denote achieving an increase in the odds of achieving this target for drug compared to placebo.

Hit all 4 targets: PP(A and B and C and D)

Hit 2 and at least one other: PP(A and B and (C or D))

Hit any 3: PP(number of A, B, C, D ≥ 3)

Pain-free and improve most bothersome: PP(A and E)

8 Bayesian Analysis of Simulated RCT with Two Endpoints

Simulated randomized clinical trials (RCT) are useful because we know the true treatment effects being estimated. Consider a two-treatment (A, B) RCT for hypertension where covariate adjustment is used for baseline systolic blood pressure (SBP) and where there are two outcomes: (1) incidence of death or stroke (DS) within one year and (2) systolic blood pressure at one year after randomization. Even though time to DS would be a preferred outcome (and would handle censoring) we ignore the timing of events for this example and use a binary logistic model for DS. SBP is assumed to follow a normal distribution with constant SD $\sigma = 7$ given baseline SBP. The true B:A treatment effect is assumed to be a mean 3mmHg difference in SBP with B. We assume a correlation between SBP reduction and incidence of DS by forming a true logistic model for DS in which baseline and 1y SBP each have a regression coefficient of 0.05 for predicting DS and treatment has a coefficient of log(0.8) corresponding to B:A odds ratio of 0.8. To estimate the true effect of treatment on DS not adjusted for follow-up SBP we first simulate a trial with $n = 40000$.

```
# Simulate a single trial with sample size n
sim ← function(n) {
    trt  ← c(rep('A', n / 2), rep('B', n / 2))
    sbp0 ← rnorm(n, 140, 7)
    sbp  ← sbp0 - 5 - 3 * (trt == 'B') + rnorm(n, sd=7)
    logit ← -2.6 + log(0.8) * (trt == 'B') + 0.05 * (sbp0 - 140) +
             0.05 * (sbp - 130)
    ds   ← ifelse(runif(n) ≤ plogis(logit), 1, 0)
    data.frame(trt, sbp0, sbp, ds)
}
```

```
set.seed(1)
d ← sim(n=40000)
require(rms)
ols(sbp ~ sbp0 + trt, data=d)
```

Linear Regression Model

```
ols(formula = sbp ~ sbp0 + trt, data = d)
```

	Model Likelihood Ratio Test		Discrimination Indexes	
Obs 40000	LR χ^2	28287.46	R^2	0.507
σ 7.0270	d.f.	2	R^2_{adj}	0.507
d.f. 39997	Pr($>\chi^2$)	0.0000	g	8.038

Residuals

Min	1Q	Median	3Q	Max
−30.12	−4.736	−0.0007268	4.737	27.48

	$\hat{\beta}$	S.E.	t	Pr($>\|t\|$)
Intercept	-4.2576	0.7026	-6.06	<0.0001
sbp0	0.9944	0.0050	198.58	<0.0001
trt=B	-2.9380	0.0703	-41.81	<0.0001

```
lrm(ds ~ sbp0 + sbp + trt, data=d)
```

Logistic Regression Model

```
lrm(formula = ds ~ sbp0 + sbp + trt, data = d)
```

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	40000	LR χ^2	2142.26	R^2	0.113	C	0.717
0	36293	d.f.	3	$R^2_{3,40000}$	0.052	D_{xy}	0.434
1	3707	Pr($>\chi^2$)	<0.0001	$R^2_{3,10090.4}$	0.191	γ	0.435
$\max\|\frac{\partial \log L}{\partial \beta}\|$	2×10^{-9}			Brier	0.079	τ_a	0.073

	$\hat{\beta}$	S.E.	Wald Z	Pr($>\|Z\|$)
Intercept	-16.6218	0.3847	-43.21	<0.0001
sbp0	0.0517	0.0036	14.40	<0.0001
sbp	0.0522	0.0026	20.32	<0.0001
trt=B	-0.2594	0.0367	-7.07	<0.0001

```
f ← lrm(ds ~ sbp0 + trt, data=d)
print(f)
```

Logistic Regression Model

```
lrm(formula = ds ~ sbp0 + trt, data = d)
```

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	40000	LR χ^2	1719.66	R^2	0.091	C	0.697
0	36293	d.f.	2	$R^2_{2,40000}$	0.042	D_{xy}	0.394
1	3707	Pr($>\chi^2$)	<0.0001	$R^2_{2,10090.4}$	0.157	γ	0.394
$\max \lvert \frac{\partial \log L}{\partial \beta} \rvert$	3×10^{-11}			Brier	0.080	τ_a	0.066

	$\hat{\beta}$	S.E.	Wald Z	Pr($>\lvert Z\rvert$)
Intercept	-16.5559	0.3807	-43.49	<0.0001
sbp0	0.1019	0.0026	38.49	<0.0001
trt=B	-0.4045	0.0358	-11.31	<0.0001

```
btrt ← coef(f)['trt=B']
```

The regression coefficient for treatment in the proper outcome model that did not adjust for 1y SBP was -0.4045 which corresponds to a B:A odds ratio of 0.6673, taken as the true treatment effect on the binary outcome.

The trial could easily be run sequentially but we treat the sample size as fixed at $n = 1500$ and simulate the trial data as such. The traditional frequentist analysis follows.

```
set.seed(7)
d ← sim(n=1500)
re ← round(cor(d$ds, d$sbp), 3)
dd ← datadist(d); options(datadist='dd')
f ← ols(sbp ~ sbp0 + trt, data=d)
f
```

Linear Regression Model

```
ols(formula = sbp ~ sbp0 + trt, data = d)
```

	Model Likelihood Ratio Test		Discrimination Indexes	
Obs 1500	LR χ^2	1087.47	R^2	0.516
σ 7.0059	d.f.	2	R^2_{adj}	0.515
d.f. 1497	Pr($>\chi^2$)	0.0000	g	8.177

Residuals

Min	1Q	Median	3Q	Max
−24.58	−4.619	0.154	4.241	24.29

	$\hat{\beta}$	S.E.	t	Pr($>\lvert t\rvert$)
Intercept	-5.4630	3.6567	-1.49	0.1354
sbp0	1.0048	0.0260	38.62	<0.0001
trt=B	-3.1831	0.3620	-8.79	<0.0001

```
summary(f)
```

	Low	High	Δ	Effect	S.E.	Lower 0.95	Upper 0.95
sbp0	135.22	144.52	9.2964	9.3411	0.24186	8.8667	9.8156
trt — B:A	1.00	2.00		-3.1831	0.36199	-3.8932	-2.4730

```
f ← lrm(ds  ~ sbp0 + trt, data=d)
f
```

Logistic Regression Model

`lrm(formula = ds ~ sbp0 + trt, data = d)`

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	1500	LR χ^2	53.82	R^2	0.075	C	0.670
0	1357	d.f.	2	$R^2_{2,1500}$	0.034	D_{xy}	0.339
1	143	Pr($>\chi^2$)	<0.0001	$R^2_{2,388.1}$	0.125	γ	0.340
$\max \lvert \frac{\partial \log L}{\partial \beta} \rvert$	2×10^{-5}			Brier	0.083	τ_a	0.059

	$\hat{\beta}$	S.E.	Wald Z	Pr($>\lvert Z\rvert$)
Intercept	-15.1566	1.8989	-7.98	<0.0001
sbp0	0.0920	0.0132	6.96	<0.0001
trt=B	-0.2715	0.1807	-1.50	0.1330

```
summary(f)
```

	Low	High	Δ	Effect	S.E.	Lower 0.95	Upper 0.95
sbp0	135.22	144.52	9.2964	0.85487	0.12286	0.61408	1.095700
Odds Ratio	135.22	144.52	9.2964	2.35110		1.84800	2.991200
trt — B:A	1.00	2.00		-0.27146	0.18069	-0.62559	0.082684
Odds Ratio	1.00	2.00		0.76227		0.53494	1.086200

Pearson's r correlation between the SBP outcome and the death/stroke outcome is 0.22. If the frequentist analysis with $n = 1500$ is repeated 2500 times, the correlation across the 2500 of the estimated treatment effects on DS and the estimated treatment effects on SBP is 0.142.

For the Bayesian analysis, we use use `Stan` and the `rstan R` package [14][14]. Two models are analyzed simultaneously. The regression coefficients in both models have a prior distribution which is multivariate normal with means equal to zero. The standard deviation of the prior for the treatment effect is specified so that the prior probability that the blood pressure reduction is more than 10mmHg is 0.1. The standard deviation for the prior for the B:A log odds ratio is taken so that the prior probability that the odds ratio is less than 0.5 (regression coefficient $< \log(0.5)$) is 0.05. A flat distribution on $[0, \infty]$ is used as the prior for the residual standard deviation. 5,000 post-warmup iterations are run using Stan's No-U-turn sampler in 4 chains, resulting in 20,000 draws from the posterior distributions and taking 10 minutes of run time on a 4-core machine.

```
require(rstan)
rstan_options(auto_write = TRUE)
options(mc.cores = parallel::detectCores())  # 4 CPUs used

model ← "
data {
    int       n;
    vector[n] x;
    real      y1[n];
    int       y2[n];
    vector[n] treat;
```

[14]Thanks to Prof. Chris Fonnesbeck, Vanderbilt Department of Biostatistics, for writing the Stan script for this model.

```
    vector[2] Zero;
    vector<lower=0>[2] sigma_b;
}

parameters {
    vector[2] alpha;
    vector[2] beta;
    vector[2] mu;
    real<lower=0> sigma_y;
    cholesky_factor_corr[2] L_b;
}

transformed parameters {
    vector[n]     theta1;
    vector[n]     theta2;

    theta1 = mu[1] + alpha[1]*x + beta[1]*treat;
    theta2 = mu[2] + alpha[2]*x + beta[2]*treat;
}

model {
    beta ~ multi_normal_cholesky(Zero, diag_pre_multiply(sigma_b, L_b));
    L_b ~ lkj_corr_cholesky(1);  // correlation matrix for reg. parameters, LKJ
         prior

    y1 ~ normal(theta1, sigma_y);
    y2 ~ bernoulli_logit(theta2);
}

generated quantities {
  matrix[2,2] Omega;
  matrix[2,2] Sigma;
  Omega = multiply_lower_tri_self_transpose(L_b);
  Sigma = quad_form_diag(Omega, sigma_b);
}
"

s <- stan(model_code = model, iter=10000, seed=7,
          data=with(d, list(x=sbp0, treat=1*(trt == 'B'),
                            y1=sbp, y2=ds,
                            sigma_b=c(-10 / qnorm(0.1),
                                      log(0.5) / qnorm(0.05)),
                            Zero=c(0,0), n=nrow(d))))
s
betas <- extract(s, pars='beta')$beta
```

```
Inference for Stan model: cd388c9aa01c1c78a612ddca57e2c5c6.
4 chains, each with iter=10000; warmup=5000; thin=1;
post-warmup draws per chain=5000, total post-warmup draws=20000.

             mean se_mean     sd      2.5%    97.5% n_eff   Rhat
alpha[1]   1.0047  0.0002 0.0259    0.9539   1.0557 11616 1.0006
alpha[2]   0.0923  0.0001 0.0133    0.0666   0.1182 11596 1.0006
beta[1]   -3.1780  0.0027 0.3607   -3.8797  -2.4695 18285 1.0001
beta[2]   -0.2129  0.0012 0.1596   -0.5325   0.1026 16387 1.0001
mu[1]     -5.4464  0.0339 3.6436  -12.6012   1.6968 11581 1.0006
mu[2]    -15.2366  0.0177 1.9063  -18.9511 -11.5603 11583 1.0006

Samples were drawn using NUTS(diag_e) at Mon Dec 19 13:51:42 2016.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

Posterior densities are obtained using kernel density estimators, shown in Figure 22.

```
b1 ← betas[, 1]
b2 ← betas[, 2]
plot(density(b1), type='l', xlab='B-A SBP Difference', main='')    # Fig. 22
plot(density(exp(b2)), type='l', xlab='B:A OR for Death or Stroke', main='')
```

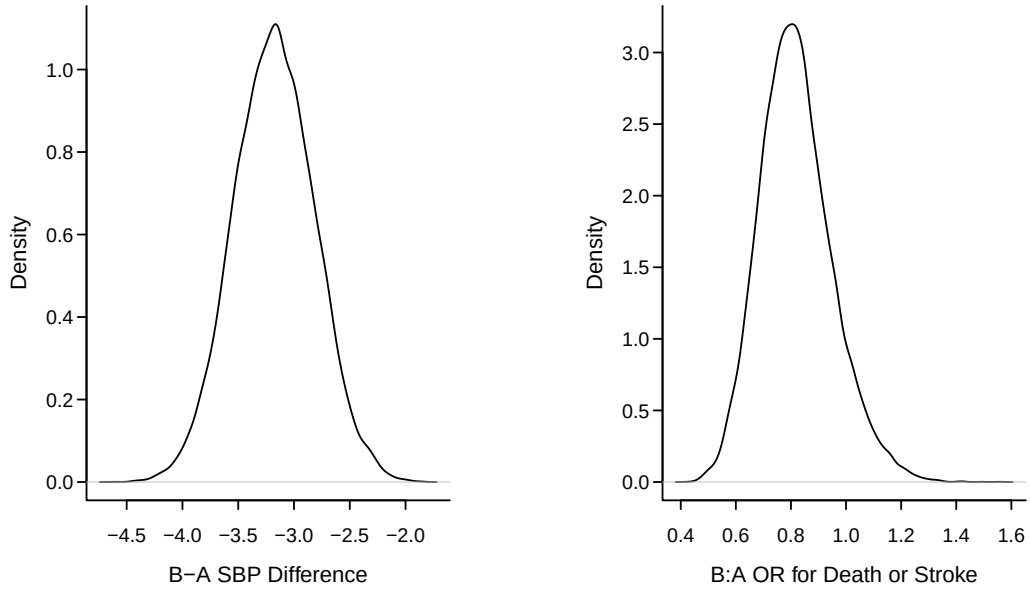


Figure 22: Posterior densities for treatment effects on two outcomes

Posterior means, medians, and 0.95 credible intervals are computed simply by using ordinary samples estimates on the posterior draws.

```
ci1 ← quantile(b1, c(0.025, 0.975))
ci2 ← quantile(b2, c(0.025, 0.975))
data.frame(Mean=c(mean(b1), mean(b2)), Median=c(median(b1), median(b2)),
           Lower=c(ci1[1], ci2[1]), Upper=c(ci2[1], ci2[2]), row.names=c('b1', '
                b2'))
```

```
         Mean      Median       Lower       Upper
b1 -3.178025 -3.1782677 -3.8796797 -0.5324795
b2 -0.212865 -0.2117236 -0.5324795  0.1025766
```

The posterior mean and median log odds ratio is less impressive than the frequentist maximum likelihood estimate of -0.271 because of the skeptical prior.

A variety of posterior probabilities are easily calculated. Here non-inferiority is defined by SBP increase less than 1mmHg and DS increased by an odds ratio less than 1.05. Similarity with respect to the effect on DS is taken to be an odds ratio between 0.85 and $\frac{1}{0.85}$. The final calculation is the mean number of targets achieved when the targets are any SBP reduction and any reduction of odds of DS.

```
cat('Prob(SBP reduced at least 2 mmHg) = ', rmean(b1 < -2), '\n',
    'Prob(B:A OR for DS < 1)           = ', rmean(b2 < 0),  '\n',
    'Prob(SBP reduced by 2 and OR < 1) = ', rmean(b1 < -2 & b2 < 0), '\n',
    'Prob(SBP reduced by 2 or  OR < 1) = ', rmean(b1 < -2 | b2 < 0), '\n',
    'Prob(Non-inferiority)             = ', rmean(b1 < 1 & b2 < log(1.05)),'\n',
```

```
    'Prob(DS similar)                  = ', rmean(exp(b2) > 0.85 &
                                                  exp(b2) < 1 / 0.85), '\n',
    'E(# targets achieved)             = ', rmean((b1 < 0) + (b2 < 0)), '\n',
    sep='')
```

```
Prob(SBP reduced at least 2 mmHg) = 0.999
Prob(B:A OR for DS < 1)           = 0.908
Prob(SBP reduced by 2 and OR < 1) = 0.908
Prob(SBP reduced by 2 or  OR < 1) = 1.000
Prob(Non-inferiority)             = 0.948
Prob(DS similar)                  = 0.363
E(# targets achieved)             = 1.908
```

The six probabilities above are forward probabilities directly addressing the clinical questions. Frequentist solutions for these questions are either highly indirect or unavailable.

9 Bayesian Clinical Trial Design

Here we consider standard non-adaptive trial designs with possible sequential monitoring.

9.1 Sample Size Estimation

For fixed sample size designs, there are several Bayesian approaches that have been developed for sample size estimation. These approaches are a bit more honest than frequentist ones because they admit uncertainty in key parameters such as subject-to-subject standard deviations and effect sizes of interest. An excellent recent paper showing how to estimate the sample size needed to compare control with a set of active treatments (e.g., different doses of one drug) is by Whitehead et al. [67]. Their goal is to have high PP that one or more treatments is better than control or high PP that none of the treatments is better than control. See also [43, 58, 32, 60, 1, 37, 51, 55, 64, 12]. Spiegelhalter et al. [58] provide a clear rationale for computing the *unconditional power*, which integrates over the prior distribution for the treatment effect, providing a "realistic assessment of the predictive probability of obtaining a 'significant' result." They also discuss the damage done by using a fixed, optimistic, treatment effect in a standard power calculation.

9.2 Bayesian Power Example

Bayesian power may be defined as the probability that the posterior probability will exceed a certain value such as 0.95. Instead of the more honest calculations discussed above that incorporate uncertainty in the true treatment effect, one may compute Bayesian power as a function of the unknown effect. Consider the situation used in Section 3.2.1 where the data are normal with variance 1.0 and the prior is normal with mean zero and variance $\frac{1}{\tau}$. Take $\mu > 0$ to indicate efficacy. The PP for $P(\mu > 0|\text{data})$ is $\Phi(\frac{n\overline{Y}}{\sqrt{n+\tau}})$.

$\overline{Y}$ is normally distributed with mean μ and variance $\frac{1}{n}$. Since the long-run average of $\overline{Y}$ is μ, the long-run median of $\overline{Y}$ is also μ so the median PP is $\Phi(\frac{n\mu}{\sqrt{n+\tau}})$. Letting $z = \Phi^{-1}(0.95)$, the chance that the PP exceeds 0.95 is $P(\Phi(\frac{n\overline{Y}}{\sqrt{n+\tau}}) > 0.95) = P(\frac{n\overline{Y}}{\sqrt{n+\tau}} > z) = \Phi(\frac{n\mu - z\sqrt{n+\tau}}{\sqrt{n}})$. For varying n and τ the median PP and the Bayesian power are shown in Figure 23.

```
require(ggplot2)
d ← expand.grid(mu=seq(-2, 2, length=200),
                n=c(1, 2, 5, 10),
                tau=c(0, 2, 10),
                what=1:2)
d ← transform(d,
```

```
               p = ifelse(what == 1,
                    pnorm(n * mu / sqrt(n + tau)),
                    pnorm((n * mu - qnorm(0.95) * sqrt(n + tau)) / sqrt(n))),
               n = factor(n),
               tau = factor(tau)
)
d$Probability ← factor(d$what)
levels(d$n)   ← paste0('n==', levels(d$n))
levels(d$tau) ← paste0('tau==', levels(d$tau))

ggplot(d, aes(x=mu, y=p, col=Probability)) + geom_line() +
  scale_color_discrete(labels=expression(Median~P(mu > 0),
                                         P(P(mu > 0) > 0.95))) +
  facet_grid(tau ~ n, labeller=label_parsed) +
  xlab(expression(mu)) +
  ylab('Probability')   # Fig. 23
```

9.3 Sequential Monitoring and Futility Analysis

As discussed above, standard posterior probabilities may be used as often as desired to monitor a study, either at prespecified times, at will, or continuously. Often a clinical trial is designed to yield more evidence than necessary for efficacy because safety endpoints are rare. The Bayesian approach can easily provide a formal approach to satisfying this need. A criterion for study termination for success could be for example a PP > 0.95 for efficacy and a PP > 0.95 for safety where "safety" is interpreted to mean that the experimental:control hazard ratio for a safety event ≤ 1.2.

Spiegelhalter et al. [58] argue that a 'range of equivalence' be used in monitoring. For example, one might terminate a study if the probability that a new agent is more than δ better than standard therapy is high, or if the probability that the standard treatment is better than the new agent by any amount is high. Here δ could be chosen as a clinical equivalence threshold to take into account toxicity tradeoffs. The value of δ is related to the 'uncertainty principle' that allows one to ethically randomize subjects [58].

The Bayesian sequential designs discussed above are written as if there is no cap on the ultimate sample size. That may be the case in some studies, but one often reaches a point where clinically important efficacy is unlikely to be achieved within a reasonable sample size. Bayesian methods can in many cases cut costs by declaring futility earlier than frequentist methods, and futility can be more formally defined with Bayes. There are three overall Bayesian attacks.

1. Stop when the current PP that the efficacy is very small or is negative exceeds some probability level such as 0.9
2. Use the *predictive distribution* at a planned ultimate sample size to decide on futility [7].
3. Pre-specify the maximum sample size and continue the study until that is reached if the trial is not stopped early for efficacy or harm (a more expensive approach).

Spiegelhalter et al. [58] have a very insightful equation that for a simple statistical setup and a flat prior estimates the chance of ultimate success given only the Z-statistic at an interim look that was based on a fraction f of subjects randomized to date. This is shown in Figure 24. Spiegelhalter et al. take issue with the practice of stochastic curtailment or conditional power analysis that assumes a single value of the true unknown efficacy parameter. This Bayesian predictive approach requires no such choice.

```
pf ← function(z, f) pnorm(z/sqrt(1 - f) - 1.96 * sqrt(f) / sqrt(1 - f))
zs ← seq(-1, 3, length=200)
fs ← c(.1, .25, .4, .75, .9)
d ← expand.grid(Z=zs, f=fs)
f ← d$f
```

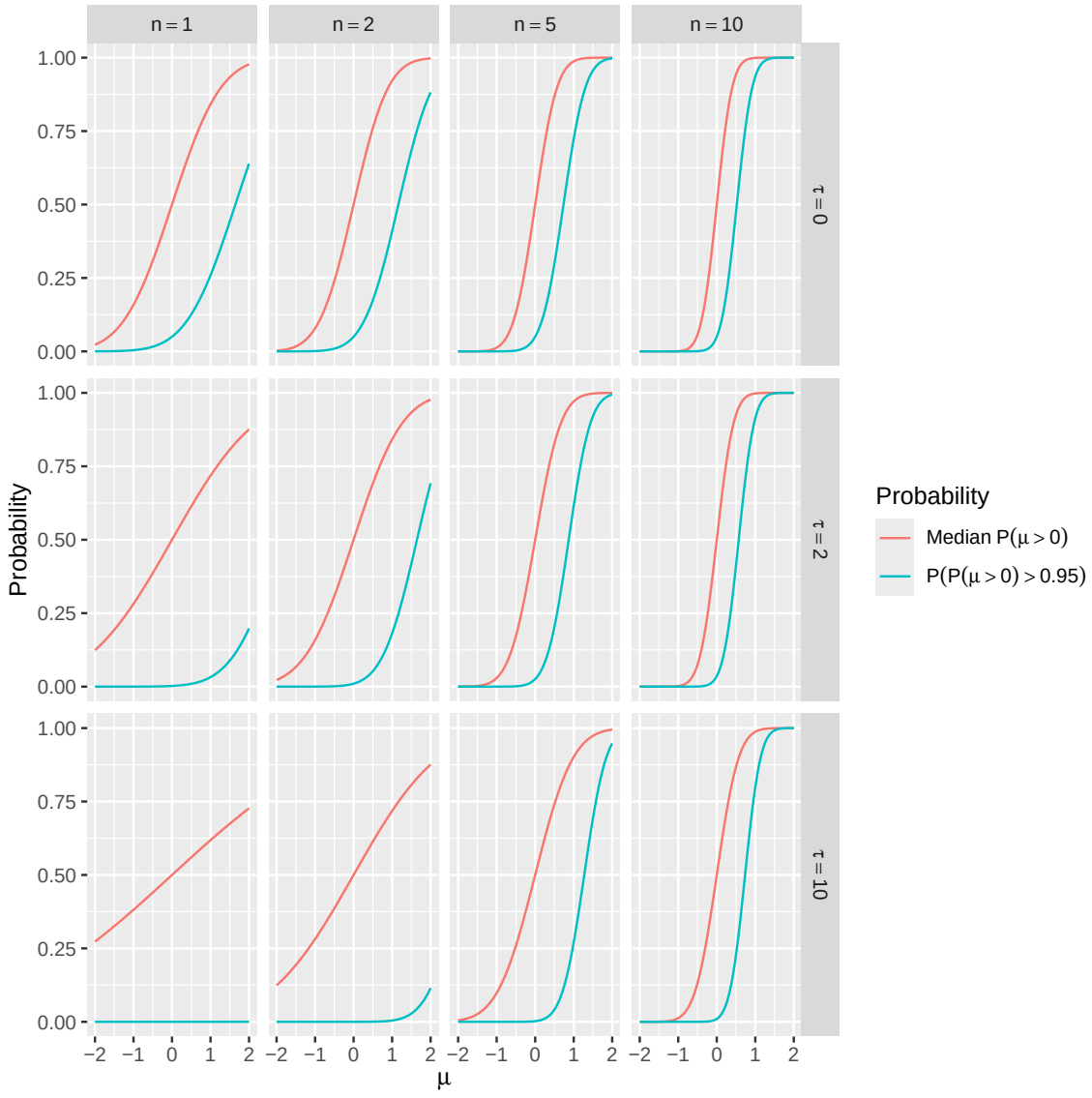


Figure 23: Bayesian power and median posterior probability for varying true effect values, sample sizes, and degrees of skepticism

```
d$p ← with(d, pf(Z, f))
d$f ← NULL
p ← split(d, f)
labcurve(p, pl=TRUE, ylab='Predictive Probability')   # Fig. 24
```

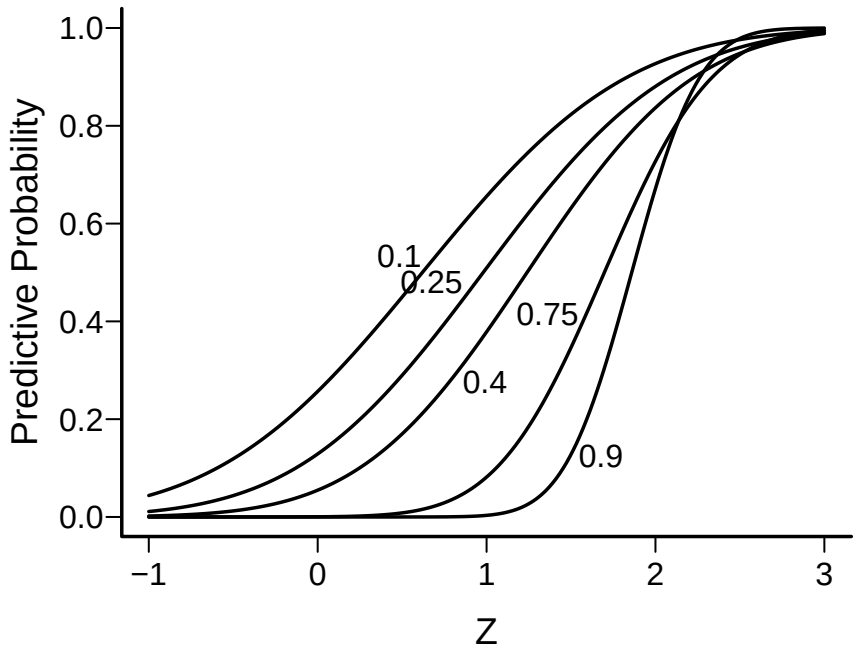


Figure 24: Predictive probability of the final 0.95 credible interval excluding zero, and the treatment effect being in the right direction, given the fraction f of study completed and the current test statistic Z when the prior is flat. f values are written beside curves.

For example, to have any reasonable hope of demonstrating efficacy if an interim Z value is 1.0, one must be less than $\frac{1}{10}^{\text{th}}$ of the way through the trial.

> The published report of any study attempts to answer the crucial question: *What does this trial add to our knowledge?* The strength of the Bayesian approach is that it allows one to express the answer *formally*. It therefore provides a rational framework for dealing with many crucial issues in trial design, monitoring and reporting. In particular, by making explicitly different prior opinions about efficacy, and differing demands on new therapies, it may shed light on the varying attitudes to balancing the clinical duty to provide the best treatment for individual patients, against the desire to provide sufficient evidence to convince the general body of clinical practice.
>
> Spiegelhalter, Freedman, and Parmar [58]

10 General Recommendations

It is clear that there is a major role for Bayesian methods in all aspects of drug development. Though the following ideas are only the authors' personal recommendations, they are a reasonable starting place.

When a sponsor launches a study to be used for their own drug development purposes, the sponsor may reasonably decide that the choice of the prior is entirely up to them. But it

should be specified in a statistical analysis plan before the new study begins. When a trial is to play an important role in a regulatory submission, the prior distribution to be used in the primary analyses should be developed in an iterative process between the sponsor and medical and statistical reviewers in the regulatory agency, before the study begins.

When trustworthy, relevant information exists before a study, and the information comes from a similar disease, treatment, and dosing setting, it may be entirely reasonable to use such information when developing the prior. It will seldom be appropriate for expert opinion alone to be used in prior formulation, but strong pharmacology and existing drug class information could be used to form a weakly optimistic prior. When a Phase 2 study has a strong design and has strong similarities to a proposed Phase 3 study, it may be appropriate to use the Phase 2 posterior distribution as a prior for Phase 3. In some situations, the previous results my be discounted using, for example, a mixture of a skeptical prior and the previous study posterior as in our pediatrics example above.

When there is no applicable prior information, it is usually appropriate to use a somewhat skeptical prior for the new study. In exchange, the sponsor would have the ability to take unlimited, unscheduled looks at the data, and at early looks the skeptical prior would properly pull back the efficacy estimates. As shown in Figure 7, the effect of skepticism is equivalent to ignoring data on a small number of subjects, and this effect wears off as n gets large.

When final results are reported, it is important to provide details about the prior, how it was developed, and when. Entire posterior distributions should be emphasized, and posterior probabilities of any efficacy as well as of more-than-trivially-important efficacy should be reported. We do not recommend a hard cutoff for "winning" on a posterior probability, but instead recommend provision of evidence of efficacy and clinically significant efficacy, with the posterior probability of the latter necessarily being less. When applicable, totality of evidence should be summarized by a posterior probability of a compound condition.

For treatments deemed not to be superior, it would be useful to future drug developers to report the probability of similarily between acive and control treatments. For non-inferiority studies, the PP of similarity should be reported as well as the PP of non-inferiority.

For reporting of main efficacy results, we recommend reporting the posterior probability of any efficacy, the PP of more than the minimally clinically significant efficacy, and a 0.95 credible interval for the unknown efficacy parameter. It should be noted that the exlusion of zero from this two-sided credible interval is not what one should emphasize in the final judgment of efficacy. Instead, the directional PPs should be used.

11 Summary

Characteristics of frequentist and Bayesian approaches are summarized in the table below.

Attribute	Frequentist	Bayesian
Nature of probabilities	long-run relative frequencies	degree of belief[a]
Probabilities calculated	P(data|no effect)	P(effect $> c$|data)
Timing of arguments	After the study, influenced by data	Before the study
Type of arguments	Multiplicity re:multiple endpoints, treatments, times; clinical significance; α-spending function; complex designs; how to accurately compute p-value; how to use outside information	Prior distribution
Everyday challenges	Conceptual	Computational
Type I assertion probability	Can be controlled but arbitrary if multiple tests. Never zero regardless of n; does not prevent detection of clinically trivial effects; **NOT** the probability of regulator's regret	Not relevant; can prevent declaring evidence for trivial effects by directly computing probability of non-trivial effect
Efficacy probability	Not available	Posterior probability; If approve drug with PP=0.96, probability of error=0.04 (regulator's regret)
Clinical relevance	Tests must be augmented by confidence limits	Built-in because of direct estimation of P(effect)
Sample size	Guessed; hard to adjust once study starts	Savings due to unlimited looks with no penalty; can stop early for harm, futility, or efficacy; can extend any study; sample size estimate can incorporate uncertainty
Effect estimates if stop early	Overstated	Perfectly calibrated by prior
Skepticism	Effect of multiplicity adjustment is constant	Wears off as $n \uparrow$
Design	Does not extend to complex designs such as response-adaptive randomization and incorporating prior information	Extends to complex designs and has formal mechanism for incorporating relevant prior information

[a]Bayesian probabilities can also be actual probabilities in the frequentist sense. See the biased coin example in Section 2.4.

The Bayesian approach provides direct measures of evidence that are on the clinical scale and not the randomness scale. PPs have meaning no matter what the context, including aggressive sequential testing. Examining how the Bayesian approach works in an extreme multiple comparisons situation sheds light on the much cleaner interpretation of PPs than p-values. Bayesian inference works well in standard fixed sample size clinical trials but also allows one to use highly flexible designs that allow earlier learning, while achieving reliability of results without any notion of type I assertion probability. PPs are perfectly calibrated even when used as a stopping rule. Bayesian effect estimates, e.g., posterior means, modes, or medians, are also perfectly calibrated even with early stopping. Fully sequential designs with no need to plan the look frequency in advance, but rather allowing it to be dictated by how outside knowledge or within-study data evolve, are easily allowed with Bayes. There is a potential to stop studies

earlier for futility or harm, and sometimes for efficacy. Simultaneous probability statements about multiple endpoints are easily made. When historical data are justified in formulating the prior distribution, Bayes is the only formal approach available.

For More Information
See Harrell's blog at fharrell.com along with comments others have posted there, especially fharrell.com/post/journey. Post your own comments and questions.

Some useful interactive demonstrations of Bayesian calculations for a two-sample t test may be found at

- rpsychologist.com/d3/bayes
- sumsar.net/best_online

An excellent resource for Bayesian methods in clinical trials may be found at trialdesign.org, especially for Phase 1 and 2 studies.

A large number of R scripts illustrating Bayesian analysis are available from github.com/avehtari/BDA_R_demos.

References

[1] C. J. Adcock. "Sample Size Determination: A Review". In: *The Statistician* 46 (1997), pp. 261–283.

[2] Alan Agresti and Brent A. Coull. "Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions". In: *Am Statistician* 52 (1998), pp. 119–126.

[3] Jim Albert. *Bayesian Computation with R*. New York: Springer, 2007.

[4] D. G. Altman and J. M. Bland. "Absence of Evidence Is Not Evidence of Absence". In: *BMJ* 311 (1995), p. 485.

[5] Peter C. Austin and Ewout W. Steyerberg. "Graphical Assessment of Internal and External Calibration of Logistic Regression Models by Using Loess Smoothers." In: *Stat Med* 33.3 (Feb. 2014), pp. 517–535. ISSN: 1097-0258. DOI: 10.1002/sim.5941. pmid: 24002997. URL: http://dx.doi.org/10.1002/sim.5941.

[6] James Berger and L. Mark Berliner. "Robust Bayes and Empirical Bayes Analysis with -Contaminated Priors". In: *The Annals of Statistics* 14.2 (1986), pp. 461–486. ISSN: 0090-5364. JSTOR: 2241230. URL: https://www.jstor.org/stable/2241230 (visited on 08/23/2019).

[7] Donald A. Berry. "Bayesian Clinical Trials". In: *Nat Rev* 5 (2006). excellent review of Bayesian approaches in clinical trials; "The greatest virtue of the traditional approach may be its extreme rigour and narrowness of focus to the experiment at hand, but a side effect of this virtue is inflexibility, which in turn limits innovation in the design and analysis of clinical trials. ... The set of 'other possible results' depends on the experimental design. ... Everything that is known is taken as given and all probabilities are calculated conditionally on known values. ... in contrast to the frequentist approach, only the probabilities of the observed results matter. ... The continuous learning that is possible in the Bayesian approach enables investigators to modify trials in midcourse. ... it is possible to learn from small samples, depending on the results, ... it is possible to adapt to what is learned to enable better treatment of patients. ... subjectivity in prior distributions is explicit and

open to examination (and critique) by all. ... The Bayesian approach has several advantages in drug development. One is the process of updating knowledge gradually rather than restricting revisions in study design to large, discrete steps measured in trials or phases.", pp. 27–36.

Editorial, p. 3
.

[8] Donald A. Berry. "Interim Analysis in Clinical Trials: The Role of the Likelihood Principle". In: *Am Statistician* 41 (1987), pp. 117–122. DOI: 10.1080/00031305.1987.10475458. URL: http://dx.doi.org/10.1080/00031305.1987.10475458.

[9] J. D. Blume. "How Often Likelihood Ratios Are Misleading in Sequential Trials". In: *Comm Stat Th Meth* 37.8 (2008), pp. 1193–1206.

[10] J. D. Blume. "Likelihood Methods for Measuring Statistical Evidence". In: *Stat Med* 21.17 (2002), pp. 2563–2599.

[11] Jeffrey D. Blume. "Likelihood and Its Evidential Framework". In: *Handbook of the Philosophy of Science: Philosophy of Statistics*. Ed. by Dov M. Gabbay and John Woods. San Diego: North Holland, 2011, pp. 493–511.

[12] Thomas M. Braun. "Motivating Sample Sizes in Adaptive Phase I Trials via Bayesian Posterior Credible Intervals". In: *Biom* (), n/a. DOI: 10.1111/biom.12872. URL: http://dx.doi.org/10.1111/biom.12872.

[13] William M. Briggs. "The Substitute for P-Values". In: *JASA* 112.519 (July 2017), pp. 897–898. DOI: 10.1080/01621459.2017.1311264. URL: http://dx.doi.org/10.1080/01621459.2017.1311264.

[14] Bob Carpenter et al. "Stan: A Probabilistic Programming Language". In: *J Stat Software* 76.1 (2017), pp. 1–32. DOI: 10.18637/jss.v076.i01. URL: https://www.jstatsoft.org/v076/i01.

[15] Leena Choi, Jeffrey D. Blume, and William D. Dupont. "Elucidating the Foundations of Statistical Inference with 2 x 2 Tables". In: *PLoS ONE* 10.4 (Apr. 2015), e0121263+. DOI: 10.1371/journal.pone.0121263. URL: http://dx.doi.org/10.1371/journal.pone.0121263.

[16] Jacob Cohen. "The Earth Is Round ($p < .05$)". In: *Am Psychologist* 49.12 (1994), pp. 997–1003. ISSN: 1935-990X. DOI: 10.1037/0003-066x.49.12.997. URL: http://dx.doi.org/10.1037/0003-066x.49.12.997.

[17] Maria J. Costa et al. "The Case for a Bayesian Approach to Benefit-Risk Assessment:" in: *Therapeutic Innovation & Regulatory Science* 51.5 (Apr. 2017), pp. 568–574. ISSN: 2168-4790. DOI: 10.1177/2168479017698190. URL: http://dx.doi.org/10.1177/2168479017698190.

[18] Gerald G. Crans and Jonathan J. Shuster. "How Conservative Is Fisher's Exact Test? A Quantitative Evaluation of the Two-Sample Comparative Binomial Trial". In: *Stat Med* 27 (2008). comparative binomial experiment;first paper to truly quantify the conservativeness of Fisher's test;"the test size of FET was less than 0.035 for nearly all sample sizes before 50 and did not approach 0.05 even for sample sizes over 100.";conservativeness of "exact" methods;see Stat in Med 28:173-179, 2009 for a criticism which was unanswered, pp. 3598–3611. DOI: 10.1002/sim.3221. URL: http://dx.doi.org/10.1002/sim.3221.

[19] Angel M. Cronin and Andrew J. Vickers. "Statistical Methods to Correct for Verification Bias in Diagnostic Studies Are Inadequate When There Are Few False Negatives: A Simulation Study". In: *BMC Med Res Methodol* 8.1 (Nov. 2008), pp. 75+. ISSN: 1471-2288. DOI: 10.1186/1471-2288-8-75. pmid: 19014457. URL: http://dx.doi.org/10.1186/1471-2288-8-75.

[20] Nigel Dallow, Nicky Best, and Timothy H. Montague. "Better Decision Making in Drug Development through Adoption of Formal Prior Elicitation". In: *Pharm Stat* 0.0 (2018). DOI: 10.1002/pst.1854. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pst.1854. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pst.1854.

[21] A. P. Dawid. "Comment on "the Philosophy of Statistics" by D. V. Lindley". In: *The Statistician* 49 (2000), pp. 325–326.

[22] W. Edwards Deming. "On Probability as a Basis for Action". In: *Am Statistician* 29.4 (Nov. 1975), pp. 146–152. DOI: 10.1080/00031305.1975.10477402. URL: http://dx.doi.org/10.1080/00031305.1975.10477402.

[23] Ward Edwards, Harold Lindman, and Leonard J. Savage. "Bayesian Statistical Inference for Psychological Research". In: *Psych Rev* 70.3 (May 1963), pp. 193–242. URL: http://psycnet.apa.org/doi/10.1037/h0044139.

[24] Scott S. Emerson. "Stopping a Clinical Trial Very Early Based on Unplanned Interim Analysis: A Group Sequential Approach". In: *Biometrics* 51 (1995), pp. 1152–1162.

[25] Alvan R. Feinstein. *Clinical Biostatistics*. St. Louis: C. V. Mosby, 1977.

[26] J. H. Friedman. *A Variable Span Smoother*. Technical Report 5. Laboratory for Computational Statistics, Department of Statistics, Stanford University, 1984.

[27] Andrew Gelman. "Bayesian and Frequentist Regression Methods". In: *Stat Med* 34.7 (Mar. 2015), pp. 1259–1260. ISSN: 02776715. DOI: 10.1002/sim.6427. URL: http://dx.doi.org/10.1002/sim.6427.

[28] Andrew Gelman. "P Values and Statistical Practice". In: *Epi* 24.1 (Jan. 2013), pp. 69–72. ISSN: 1044-3983. DOI: 10.1097/ede.0b013e31827886f7. pmid: 23232612. URL: http://dx.doi.org/10.1097/ede.0b013e31827886f7.

[29] Steven N. Goodman. "Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy". In: *Ann Int Med* 130.12 (June 1999). Nice language for what happens when scientists use NHST to justify strong statements in their conclusions and interpretation; p-value fallacy, pp. 995+. ISSN: 0003-4819. DOI: 10.7326/0003-4819-130-12-199906150-00008. URL: http://dx.doi.org/10.7326/0003-4819-130-12-199906150-00008.

[30] Sander Greenland et al. "Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations". In: *Eur J Epi* 31.4 (2016). Best article on misinterpretation of p-values. Pithy summaries., pp. 337–350. DOI: 10.1007/s10654-016-0149-3. URL: http://dx.doi.org/10.1007/s10654-016-0149-3.

[31] Anthony G. Greenwald et al. "Effect Sizes and p Values: What Should Be Reported and What Should Be Replicated?" In: *Psychophysiology* 33.2 (1996), pp. 175–183. DOI: 10.1111/j.1469-8986.1996.tb02121.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8986.1996.tb02121.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8986.1996.tb02121.x.

[32] Jean-Marie Grouin et al. "Bayesian Sample Size Determination in Non-Sequential Clinical Trials: Statistical Aspects and Some Regulatory Considerations". In: *Stat Med* 26 (2007), pp. 4914–4924.

[33] Frank E. Harrell and Tina Shih. "Using Full Probability Models to Compute Probabilities of Actual Interest to Decision-Makers". In: *Int J Tech Assess Hlth Care* 17 (2001), pp. 17–26.

[34] Frank E. Harrell and James C. Slaughter. "Biostatistics for Biomedical Research". In: (2020). URL: https://hbiostat.org/bbr.

[35] M. A. Hlatky et al. "Rethinking Sensitivity and Specificity". In: *Am J Card* 59 (1987), pp. 1195–1198.

[36] Alexei C. Ionan et al. "Bayesian Methods in Human Drug and Biological Products Development in CDER and CBER". In: *Ther Innov Regul Sci* (Dec. 2, 2022). Examples of use of Bayes at FDA CDER and CBER. ISSN: 2168-4804. DOI: 10.1007/s43441-022-00483-0. URL: https://doi.org/10.1007/s43441-022-00483-0 (visited on 12/02/2022).

[37] Lawrence Joseph and Patrick Bélisle. "Bayesian Sample Size Determination for Normal Means and Differences between Normal Means". In: *The Statistician* 46 (1997), pp. 209–226.

[38] Lee Kennedy-Shaffer. "When the Alpha Is the Omega: P-Values, "Substantial Evidence," and the 0.05 Standard at FDA". In: *Food Drug Law J* 72.4 (2017), pp. 595–635. ISSN: 1064-590X. pmid: 30294197. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6169785/ (visited on 08/06/2020).

[39] Annette Kopp-Schneider, Silvia Calderazzo, and Manuel Wiesenfarth. "Power Gains by Using External Information in Clinical Trials Are Typically Not Possible When Requiring Strict Type I Error Control". In: *Biometrical Journal* 0.0 (2019). ISSN: 1521-4036. DOI: 10.1002/bimj.201800395. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201800395 (visited on 07/07/2019).

[40] John K. Kruschke. "Bayesian Estimation Supersedes the t Test." In: *J Exp Psych* 142.2 (May 2013), pp. 573–603. ISSN: 1939-2222. DOI: 10.1037/a0029146. pmid: 22774788. URL: http://dx.doi.org/10.1037/a0029146.

[41] John K. Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Second Edition. Waltham MA: Academic Press, 2015. ISBN: 978-0-12-405888-0. URL: http://www.sciencedirect.com/science/book/9780124058880.

[42] John K. Kruschke and Torrin M. Liddell. "Bayesian Data Analysis for Newcomers". In: (2017). Excellent for teaching Bayesian methods and explaining the advantages, pp. 1–23. DOI: 10.3758/s13423-017-1272-1. URL: http://dx.doi.org/10.3758/s13423-017-1272-1.

[43] Kevin Kunzmann et al. "A Review of Bayesian Perspectives on Sample Size Derivation for Confirmatory Trials". In: *The American Statistician* 0.0 (Mar. 22, 2021), pp. 1–9. ISSN: 0003-1305. DOI: 10.1080/00031305.2021.1901782. URL: https://doi.org/10.1080/00031305.2021.1901782 (visited on 04/23/2021).

[44] Dennis V. Lindley. "The Analysis of Experimental Data: The Appreciation of Tea and Wine". In: *Teaching Statistics* 15.1 (Mar. 1993), pp. 22–25. DOI: 10.1111/j.1467-9639.1993.tb00252.x. URL: http://dx.doi.org/10.1111/j.1467-9639.1993.tb00252.x.

[45] Daniel B. Mark, Kerry L. Lee, and Frank E. Harrell. "Understanding the Role of P Values and Hypothesis Tests in Clinical Research". In: *JAMA Card* 1.9 (Dec. 2016), pp. 1048–1054. ISSN: 2380-6583. DOI: 10.1001/jamacardio.2016.3312. URL: http://dx.doi.org/10.1001/jamacardio.2016.3312.

[46] N. Maxwell. *Data Matters: Conceptual Statistics for a Random World*. Key College Pub., 2004. URL: https://books.google.com/books?id=KH5GAAAAYAAJ.

[47] Richard McElreath. *Statistical Rethinking : A Bayesian Course with Examples in R and Stan*. 2016. ISBN: 978-1-4822-5344-3. URL: http://www.worldcat.org/isbn/9781482253443.

[48] Blakeley B. McShane and David Gal. "Statistical Significance and the Dichotomization of Evidence". In: *JASA* 112.519 (Oct. 2017), pp. 885–895. ISSN: 0162-1459. DOI: 10.1080/01621459.2017.1289846. URL: http://dx.doi.org/10.1080/01621459.2017.1289846.

[49] Fanni Natanegara et al. "The Current State of Bayesian Methods in Medical Product Development: Survey Results and Recommendations from the DIA Bayesian Scientific Working Group". In: *Pharm Stat* 13.1 (Jan. 2014), pp. 3–12. ISSN: 15391604. DOI: 10.1002/pst.1595. URL: http://dx.doi.org/10.1002/pst.1595.

[50] M. Oakes. *Statistical Inference: A Commentary for the Social and Behavioral Sciences*. "It is incomparably more useful to have a plausible range for the value of a parameter than to know, with whatever degree of certitude, what single value is untenable.". New York: Wiley, 1986.

[51] Hamid Pezeshk and John Gittins. "A Fully Bayesian Approach to Calculating Sample Sizes for Clinical Trials with Binary Reponses". In: *Drug Info J* 36 (2002), pp. 143–150.

[52] Stephen J. Ruberg et al. "Application of Bayesian Approaches in Drug Development: Starting a Virtuous Cycle". In: *Nat Rev Drug Discov* (Feb. 15, 2023), pp. 1–16. ISSN: 1474-1784. DOI: 10.1038/s41573-023-00638-0. URL: https://www.nature.com/articles/s41573-023-00638-0 (visited on 02/16/2023).

[53] Heinz Schmidli et al. "Robust Meta-Analytic-Predictive Priors in Clinical Trials with Historical Control Information". In: *Biometrics* 70.4 (Dec. 2014), pp. 1023–1032. ISSN: 0006341X. DOI: 10.1111/biom.12242. URL: http://dx.doi.org/10.1111/biom.12242.

[54] Nate Silver. *The Signal and the Noise: Why So Many Predictions Fail–but Some Don't*. 1st ed. Penguin Books, 2012. ISBN: 0-14-312508-7. URL: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0143125087.

[55] Richard Simon and Laurence S. Freedman. "Bayesian Design and Analysis of Two Two Factorial Clinical Trials". In: *Biometrics* 53 (1997), pp. 456–464.

[56] D. J. Spiegelhalter. "Probabilistic Prediction in Patient Management and Clinical Trials". In: *Stat Med* 5 (1986). z-test for calibration inaccuracy (implemented in Stata, and R Hmisc package's val.prob function), pp. 421–433. DOI: 10.1002/sim.4780050506. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780050506.

[57] David J. Spiegelhalter, Keith R. Abrams, and Jonathan P. Myles. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. New York: Wiley, 2004.

[58] David J. Spiegelhalter, L. S. Freedman, and M. K. B. Parmar. "Applying Bayesian Ideas in Drug Development and Clinical Trials". In: *Stat Med* 12 (1993), pp. 1501–1511. DOI: 10.1002/sim.4780121516. URL: http://dx.doi.org/10.1002/sim.4780121516.

[59] David J. Spiegelhalter, Laurence S. Freedman, and Mahesh K. B. Parmar. "Bayesian Approaches to Randomized Trials". In: *J Roy Stat Soc A* 157 (1994), pp. 357–416. DOI: 10.2307/2983527. URL: https://doi.org/10.2307/2983527.

[60] David J. Spiegelhalter and Lawrence S. Freedman. "A Predictive Approach to Selecting the Size of a Clinical Trial, Based on Subjective Clinical Opinion". In: *Stat Med* 5 (1986), pp. 1–13. DOI: 10.1002/sim.4780050103. URL: http://dx.doi.org/10.1002/sim.4780050103.

[61] R Development Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2020. ISBN: 3-900051-07-0. URL: http://www.R-project.org.

[62] Andrew J. Vickers. "Decision Analysis for the Evaluation of Diagnostic Tests, Prediction Models, and Molecular Markers". In: *Am Statistician* 62.4 (2008). limitations of accuracy metrics;incorporating clinical consequences;nice example of calculation of expected outcome;drawbacks of conventional decision analysis, especially because of the difficulty of eliciting the expected harm of a missed diagnosis;use of a threshold on the probability of disease for taking some action;decision curve;has other good references to decision analysis, pp. 314–320.

[63] Eric-Jan Wagenmakers et al. "Bayesian Inference for Psychology. Part I: Theoretical Advantages and Practical Ramifications". In: (2017), pp. 1–23. DOI: 10.3758/s13423-017-1343-3. URL: http://dx.doi.org/10.3758/s13423-017-1343-3.

[64] Hansheng Wang, Shein-Chung Chow, and Murphy Chen. "A Bayesian Approach on Sample Size Calculation for Comparing Means". In: *J Biopharm Stat* 15.5 (Sept. 2005). analytic form for posterior for normal t-test case, pp. 799–807. ISSN: 1054-3406. DOI: 10.1081/bip-200067789. URL: http://dx.doi.org/10.1081/bip-200067789.

[65] Ronald L. Wasserstein and Nicole A. Lazar. "The ASA's Statement on p-Values: Context, Process, and Purpose". In: *Am Statistician* 70.2 (Apr. 2016), pp. 129–133. ISSN: 0003-1305. DOI: 10.1080/00031305.2016.1154108. URL: http://dx.doi.org/10.1080/00031305.2016.1154108.

[66] Kristina Weber, Rob Hemmings, and Armin Koch. "How to Use Prior Knowledge and Still Give New Data a Chance?" In: *Pharmaceutical Statistics* 17.4 (2018), pp. 329–341. ISSN: 1539-1612. DOI: 10.1002/pst.1862. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pst.1862 (visited on 07/13/2018).

[67] John Whitehead, Faye Cleary, and Amanda Turner. "Bayesian Sample Sizes for Exploratory Clinical Trials Comparing Multiple Experimental Treatments with a Control". In: *Stat Med* 34.12 (May 2015), pp. 2048–2061. ISSN: 02776715. DOI: 10.1002/sim.6469. URL: http://dx.doi.org/10.1002/sim.6469.

[68] Manuel Wiesenfarth and Silvia Calderazzo. "Quantification of Prior Impact in Terms of Effective Current Sample Size". In: *Biometrics* 0 (ja 2019). ISSN: 1541-0420. DOI: 10.1111/biom.13124. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13124 (visited on 07/31/2019).