

Multic Lean Statistical Tutorial

John D. Cook

Department of Biostatistics and Applied Mathematics

University of Texas M. D. Anderson Cancer Center

April 5, 2006

Contents

1 Preliminaries	3
1.1 Beta distribution	3
1.1.1 Basic properties	3
1.1.2 Asymptotic properties	4
1.1.3 Exercises	5
1.2 Posterior probabilities	5
1.2.1 Exercises	7
1.3 Random inequalities	7
1.3.1 Exercises	8
1.4 Multinomial and Dirichlet distributions	9
1.4.1 Exercises	11
2 Clinical trials	12
2.1 Response-only trial monitoring	12
2.1.1 Method description	12
2.1.2 Example	14
2.1.3 Criticism	15
2.1.4 Exercises	17
2.2 Response and toxicity trial monitoring	17
2.2.1 Example	21

2.2.2	Continuous monitoring	24
3	Scenarios and simulations	26
3.1	Philosophical considerations	26
3.1.1	Exercises	27
3.2	Simulations	27
3.2.1	Exercises	28
4	Solutions and references	29
4.1	Solutions	29
4.1.1	Beta distribution	29
4.1.2	Posterior probabilities	30
4.1.3	Random inequalities	30
4.1.4	Multinomial and Dirichlet distributions	32
4.1.5	Response-only trial monitoring	32
4.1.6	Philosophical considerations	33
4.1.7	Simulations	35
4.2	References	37

Chapter 1

Preliminaries

1.1 Beta distribution

1.1.1 Basic properties

The beta distribution is important for two main reasons: it is the natural conjugate prior for the binomial distribution, and is a flexible distribution, capable of assuming a wide variety of shapes. The beta distribution has two positive parameters, commonly denoted simply a and b .

Let X be a beta(a, b) random variable. Then the PDF of X is

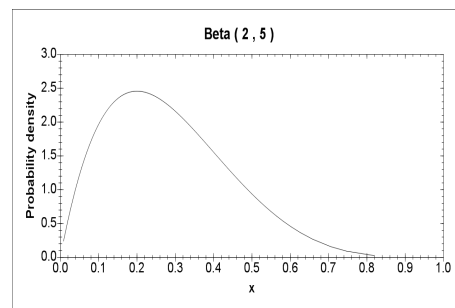
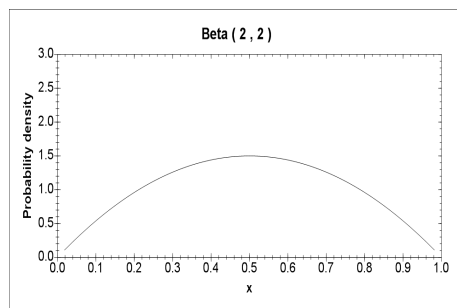
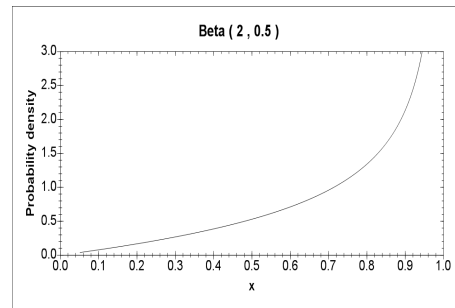
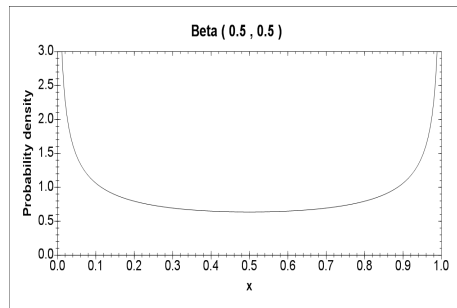
$$f_X(x) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1}$$

for $0 \leq x \leq 1$. The normalizing factor $B(a, b)$ is defined as

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Note that f_X has a singularity at 0 for $0 < a < 1$ and a singularity at 1 for $0 < b < 1$. If $a > 1$ and $b > 1$, the distribution is unimodal with mode

$$m = \frac{a-1}{a+b-2}.$$



The mean of X is

$$\mu = \frac{a}{a+b}$$

and the variance is

$$\sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}.$$

Note that

$$\sigma^2 = \frac{\mu(1-\mu)}{(a+b+1)} < \mu(1-\mu).$$

This says the mean of a beta random variable determines an upper bound on the variance. This upper bound takes on its maximum value of $1/4$ when $\mu = 1/2$.

1.1.2 Asymptotic properties

For large values of a and b , X can be approximated by a normal random variable with the same mean and variance. For example, a beta(40,60) random variable is well approximated by a normal random variable with mean 0.4 and variance

0.00238. Also, for large parameters, the mean and mode of X are approximately equal. Finally, note that if $a = \mathcal{O}(n)$ and $b = \mathcal{O}(n)$ then $\sigma^2 = \mathcal{O}(1/n)$.

1.1.3 Exercises

For the following exercises, assume that X is a random variable with a beta(a, b) distribution and PDF f_X . These exercises should be done with pencil and paper, without a computer.

1. For what values of a and b does X have a uniform distribution?
2. Suppose the graph of f_X is a straight line with a positive slope. What are the values of a and b ?
3. If $a = b = 13$, what is $P(X > 0.5)$?
4. If $a = b = 100$, is $P(X > 0.60)$ greater than 0.01? Hint: use a normal approximation.
5. Describe the graph of f_X if $a = 100$ and $b = 20$.
6. Describe the graph of f_X if $a = 0.5$ and $b = 3$.
7. If the graph of f_X has a bathtub shape, what can you say about a and b ?

1.2 Posterior probabilities

Let θ be a Bernoulli random variable distributed a priori as beta(a, b) and let y be an observation from θ , $y = 0$ or $y = 1$. Then the distribution on $\theta|y$ is beta($a + y, b + 1 - y$). If we observe n independent samples from θ and find s 1's (successes) and f 0's (failures) then the posterior distribution of θ is beta($a+s, b+f$). We update our prior on θ by adding the number of successes to the a parameter and the number of failures to the b parameter. Note that as more data accumulates, *i.e.*, as s and f increase, the effect of a and b fades and the

posterior distribution converges to a spike centered at the sample mean of the data.

Because of the rule for updating beta priors given data, the parameters a and b are sometimes referred to as the number of prior successes and prior failures, respectively. Also, $a + b$ is sometimes referred to as the number of prior observations and serves as a measure of how informative a prior is. Using this terminology, we would say that a $\text{beta}(0.4, 1.6)$ prior is a slightly informative prior, containing as much information as two observations. A $\text{beta}(10,13)$ prior is highly informative, containing as much information as 23 observations.

As an example, consider an experiment to determine the probability θ of a certain bent coin coming up heads. Suppose we start with a uniform prior on θ and then observe 6 heads and 4 tails. The posterior distribution on θ given this data is $\text{beta}(7, 5)$. This distribution has a mean of 0.583 and a variance of 0.131. Now suppose after further observation we have a total of 600 heads and 400 tails. The posterior distribution is now $\text{beta}(601, 401)$. This distribution has a mean of 0.5998 and a variance of 0.000239.

When first encountering Bayesian statistics, people commonly believe that a uniform distribution is the ultimate uninformative prior. It is an *egalitarian* prior in the sense that all probabilities have the same weight, but other distributions are less informative in the sense that they lead to a posterior distribution that converges more quickly to the sample mean. For instance, in the example above, if the prior had been $\text{beta}(0.01, 0.01)$, the posterior after the first ten observations would have been $\text{beta}(6.01, 4.01)$. However, it is odd to call a $\text{beta}(0.01, 0.01)$ distribution uninformative: it implies a strong belief that θ is either 0 or 1 and very unlikely to have any value in between.

1.2.1 Exercises

1. Suppose the prior probability of response on a particular drug is distributed as $\text{beta}(0.3, 0.7)$. After 6 responses and 4 failures have been observed, what is the posterior distribution on the probability of response?
2. Suppose you are quite certain that a coin is fair. Your prior probability of heads is distributed as $\text{beta}(1000, 1000)$. You then do an experiment and obtain 60 heads and 40 tails. What is your posterior distribution on the probability of heads? What is the posterior mean and variance? What can you say about the posterior probability that the probability of heads is 0.60 or greater?

1.3 Random inequalities

For any independent random variables X and Y , the probability that a sample from X is greater than a sample from Y is given by

$$P(X > Y) = \int_{-\infty}^{\infty} f_X(x)F_Y(x) dx$$

where f_X is the PDF of X and F_Y is the CDF of Y . One could approximate the above integral by sampling from the two distributions. While sampling is often the simplest approach, numerical integration can be far more accurate and efficient. See [Cook 2003].

We are interested in the special case of X and Y both being beta random variables. In this case the infinite integral becomes an integral over $[0, 1]$ since the PDF of a beta random variable is zero outside of this interval. See [Cook 2003] for numerical techniques for computing beta inequalities. When applied to clinical trials, X may be the probability of some outcome using an experimental treatment while Y is the corresponding probability for a standard treatment.

Reducing the random variables to their means can lead to the following erroneous reasoning: samples from X center around its mean μ_X and samples

from Y around its mean μ_Y and so if $\mu_X > \mu_Y$ then $P(X > Y)$ is large. One must keep the variances of X and Y in mind. If the probability masses of X and Y are highly concentrated relative to $\mu_X - \mu_Y$ then the preceding reasoning is correct. But if the variance of one or both variables is large, the reasoning breaks down. For example, if X has a beta(0.4, 0.6) distribution and Y a beta(35,65) distribution, $\mu_X = 0.4$ and $\mu_Y = 0.35$. In this case $P(X > Y) = 0.48$. In other words, even though an average sample from X is larger than an average sample from Y , when independent samples are drawn from each, the sample from Y is larger than its counterpart from X more than half of the time.

Often we are interested in random inequalities with a shift term δ . In this case, we are interested in $P(X > Y + \delta)$. In the previous discussion δ was zero. But we may be interested in positive or negative values of δ . For example, if X represents patient response on an experimental treatment and Y the patient response on the standard treatment, setting $\delta = 0.1$ says we are interested in the probability that X is at least a 0.1 improvement over Y . On the other hand, if $\delta = -0.05$ then we are interested in the probability that X is no more than 0.05 worse than Y .

1.3.1 Exercises

For the following exercises, assume that X and Y are independent random variables with beta(a_X, b_X) and beta(a_Y, b_Y) distributions, respectively. Define

$$\psi(a_X, b_X, a_Y, b_Y) = P(X > Y).$$

Also define

$$\varphi(a_X, b_X, a_Y, b_Y, \delta) = P(X > Y + \delta).$$

These exercises should be done with pencil and paper, without a computer.

1. $\psi(3.1, 2, 3.1, 2) =$
2. $\psi(23, 8, 7, 2) + \psi(7, 2, 23, 8) =$

3. What can you say about the magnitude of $\psi(30, 70, 90, 10)$?
4. $\lim_{c \rightarrow \infty} \psi(a, b, c, d) =$
5. $\varphi(5, 4, 3, 2, 2) =$
6. $\varphi(8, 2, 1, 3, -3) =$
7. What can you say about φ as a function of δ ?
8. $\varphi(4, 3, 2, 1, 0.5) + \varphi(2, 1, 4, 3, -0.5) =$

1.4 Multinomial and Dirichlet distributions

The multinomial distribution is a generalization of the binomial distribution. Rather than simply counting the number of successes (for example, heads in coin tosses), the multinomial counts the number of events in a vector of categories (for example, how many times each side of a die comes up in a set of rolls). A multinomial distribution has parameters n and

$$\mathbf{p} = (p_1, \dots, p_k)$$

where each p_i is non-negative and the p_i 's sum to 1. In the case of multiple rolls of a six-sided die, $k = 6$ and each p_i is the probability of the i th face appearing on a single roll. A multinomial distribution with parameters n and \mathbf{p} assigns probabilities to k -tuples of integers. The probability assigned to a k -tuple is zero if one of the components is negative or if the sum of the components exceeds n .

For non-negative integers a_1, \dots, a_k with $a_1 + \dots + a_k = n$ the PDF of a multinomial distribution with parameters n and (p_1, p_2, \dots, p_k) is

$$f(a_1, \dots, a_k) = \frac{n!}{a_1! \dots a_k!} p_1^{a_1} \dots p_k^{a_k}.$$

The marginal distribution of the i th component is binomial with parameters n and p_k .

One could consider a binomial random variable as a special case of a multinomial with $k = 2$. Instead of a single value for the number of successes, we consider it to return two values, the number of successes and the number of failures. Here $p_1 = p$ and $p_2 = 1 - p$.

The Dirichlet distribution generalizes the beta distribution in much the same way that the multinomial generalizes the binomial. The Dirichlet random variable

$$\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$$

with positive parameters a_1, a_2, \dots, a_k has PDF

$$f(\mathbf{p}) = \frac{\Gamma(a_1 + \dots + a_k)}{\Gamma(a_1) \dots \Gamma(a_k)} p_1^{a_1-1} \dots p_k^{a_k-1}$$

where $\mathbf{p} = (p_1, \dots, p_k)$. If $p_1 + \dots + p_k \neq 1$ then the probability density is zero.

Let $a = a_1 + \dots + a_k$. Then the component θ_i has the expected value $\mu_i = a_i/a$ and variance $\mu_i(1-\mu_i)/(a+1)$. The covariance of a pair of components (θ_i, θ_j) with $i \neq j$ is $-\mu_i\mu_j/(a+1)$. Note that as with the beta distribution, changing all the parameters proportionately changes the component variances by the same amount but does not change the component means.

When viewed as a special case of the Dirichlet, a beta distribution assigns probability densities to the pair $(p, 1 - p)$.

The Dirichlet distribution is the natural (conjugate) prior for the multinomial, just as the beta is the natural (conjugate) prior for the binomial. Suppose a vector $\boldsymbol{\theta}$ has prior distribution $\text{Dir}(a_1, \dots, a_k)$. If we tally a series of observations from $\boldsymbol{\theta}$ thus forming a multinomial sample $Y = (y_1, \dots, y_k)$, then the posterior distribution on $\boldsymbol{\theta}$ given the data Y is $\text{Dir}(a_1 + y_1, \dots, a_k + y_k)$. The successes in each category are added to that category's parameter in the Dirichlet distribution.

1.4.1 Exercises

1. Suppose there are four possible outcomes for a patient in a clinical trial: A_1 , A_2 , A_3 , and A_4 . Each event A_i has probability p_i . Three patients are treated. What is the probability of outcome A_3 occurring twice and A_2 once?
2. Suppose the events A_i are as described in the table below.

A_1	response: yes,	toxicity: yes
A_2	response: yes,	toxicity: no
A_3	response: no,	toxicity: yes
A_4	response: no,	toxicity: no

What is the marginal probability of response, *i.e.*, the probability of response either with or without toxicity? What is the marginal probability of toxicity?

3. Suppose a trial stops if there are either three non-responses or three toxicities from among the first three patients. What is the probability of the trial stopping? Show that the marginal probabilities of toxicity and response are not sufficient to determine this probability.

Chapter 2

Clinical trials

2.1 Response-only trial monitoring

2.1.1 Method description

The simplest case of the multiple comparison method is not to compare multiple events, but to only consider a single event: patient response. We present this method as an intermediate step along the way toward a realistic trial design. *We do not endorse the method in this section as it stands*; see the criticisms in section 2.1.3. However, it does illustrate important features of the full method in a simplified setting.

Suppose we are designing a single-arm phase II clinical trial for an experimental treatment E . The standard treatment is denoted by S . The purpose of the following design is to provide a rule to stop the trial if, based on interim data, it becomes apparent that treatment E is less effective than treatment S .

Since this is a single-arm trial, the method will compare the posterior probabilities of response for patients in the trial on treatment E with historical data on treatment S ; thus, we do not learn more about S during the trial. We assume *a priori* that θ_E , the probability of response on treatment E , has a beta(a_E, b_E)

distribution and that θ_S , the probability of response on the standard treatment, has a $\text{beta}(a_S, b_S)$ distribution.

We will stop the trial if

$$P(\theta_S > \theta_E \mid \text{data}) > \pi^*.$$

Typically π^* is in the range $[0.90, 0.99]$ with 0.95 being a common choice.

The distribution on θ_S should be based on the number of successes and failures for previous patients treated under similar circumstances, if possible. (See section 2.1.3 below.)

The choice of prior on θ_E is somewhat arbitrary, and it isn't critical: for any weakly informative prior, the actual data will quickly wash out the effect of the prior. We recommend the following procedure for selecting the prior on θ_E .

Rule of thumb: The prior on θ_E should have the same mean as θ_S and its parameters should sum to 2.

The first half of this rule says that we believe *a priori* that both treatments have the same average effectiveness. If we believed that θ_S were more effective, it would be unethical to conduct the trial. On the other hand, it would be presumptuous to build into the method an assumed superiority for the experimental treatment. Therefore we assume that both distributions have the same mean. The additional assumption that $a + b = 2$ uniquely determines a and b , and insures that the prior distribution is not too informative.

Though it may not be obvious at first, one also does not want a prior which is too *uninformative*. For example, a $\text{beta}(0.01, 0.01)$ prior is generally regarded as very uninformative, though it is actually *informative* in an odd way: the prior says we are quite certain that the probability of response is either near 0 or near 1, that the treatment being investigated is either virtually ineffective or highly effective. If the first patient fails to respond, the posterior distribution becomes $\text{beta}(1.01, 0.01)$, indicating a strong belief that the treatment is ineffective. This

may cause the trial to stop after only one patient! This behavior would actually be reasonable if we truly did believe that the treatment were either completely effective or completely ineffective. It would be like someone saying they have either a double-headed or double-tailed coin in their pocket: after one coin toss, you know what is going to happen on all future tosses.

2.1.2 Example

For the upcoming trial of treatment E , suppose we are limited to a maximum of 30 patients. Suppose further that we have searched for historical data on patients similar to those in the upcoming trial of treatment E and found 100 patients. Of these patients, 30 responded to treatment S and 70 did not. We therefore take θ_S , the historical response distribution on treatment S , to be $\text{beta}(30, 70)$.

The rule of thumb from the previous section suggests a $\text{beta}(0.6, 1.4)$ prior distribution on θ_E . We choose $\pi^* = 0.95$. Roughly speaking, we will stop the trial if we're 95% sure the standard treatment is superior to the experimental treatment.

We could enroll and observe one patient at a time, updating the posterior distribution on θ_E and checking the stopping rule. However, it is possible to look ahead and completely determine under what circumstances the trial will stop. We call these circumstances *stopping boundaries*.

For example, we know that under the worst circumstances, the stopping rule cannot kick in until we have treated at least six patients. If the first five patients do not respond, the posterior probability of a sample from θ_E being less than a sample from θ_S is 0.949, but if the first six do not respond, this probability goes up to 0.965. Therefore if no patients out of the first six respond, we stop the trial, but we cannot stop any earlier.

The full set of stopping boundaries are

Response count	Boundary
0	6
1	12
2	17
3	22
4	27
5	30

This says we stop if we see no response from among the first six patients, or only one response among the first 12, or only two among the first 17, etc. We must stop after 30 no matter how well the experimental treatment is performing.

2.1.3 Criticism

Single-arm trials are not ideal: the patients for whom we have historical data were not treated under the same conditions as those in the current trial, and so it is quite possible that factors over which we have no control have a larger effect than the differences between drugs treatments E and S . We therefore assume that for some reason we are not able to conduct a two-arm trial. The methodology presented in this tutorial falls under the heading “How to Conduct a Single-Arm Trial If You Must” with our strong recommendation to randomize patients between treatments E and S if logistically and ethically possible.

One way to mitigate the problem of using historical data is to *discount* the data, retaining the mean but increasing the variance. For a beta distribution, this means reducing the a and b parameters by the same proportion. How much of the historical data should be retained? A general rule is to retain no more than half the data. In the preceding example, this would result in a beta(15,35) rather than a beta(30,70) distribution on θ_S . Depending on the relevance of the historical data to the trial at hand, the historical data may need to be

discounted more.

Sometimes the difficulties with the historical data are so great that one must construct a standard distribution θ_S by more creative methods. (In which case there is all the more reason to question the validity of the single-arm approach, but again we assume that the designer has no choice.) One may determine, for example, the desired mean and variance for θ_S and solve for the parameters a and b . This is possible as long as the variance is below the upper bound determined by the mean. Alternatively, one may wish to specify two quantile values and solve for the beta parameters. This capability is built into the accompanying software.

One may wish to stop a trial if the experimental outcomes appear no better than the standard. If both treatments are equally effective, the stopping criterion

$$P(\theta_S > \theta_E | \text{data}) > \pi^*$$

would rarely be satisfied for any reasonable value of π^* . (The left side of the inequality would hover around 0.5 while π^* is most often around 0.95.) In this case we may compare the shifted probabilities, as introduced in section 1.3. The stopping criterion is generalized to

$$P(\theta_S + \delta > \theta_E | \text{data}) > \pi^*$$

for some value of δ . Setting δ to a value greater than zero requires the experimental treatment to be an improvement over the standard.

A remaining flaw of the design in this section is that it allows the trial to continue regardless of the number of toxicities. The most common justification for not monitoring toxicity in the design of phase II trials is that the toxicity *will* be monitored, albeit implicitly: the investigator will stop the trial if things “go bad.” No doubt this is true: if every single patient experiences adverse effects from drug toxicity, most investigators will eventually stop the trial. In this case the trial effectively has two stopping rules: an explicit statistical rule for

monitoring response and an implicit subjective rule for monitoring toxicity. The former is subjected to formal review; the latter depends on the investigator's psychological discomfort with the results. In section 2.2 we extend our method to include monitoring both toxicity and response.

2.1.4 Exercises

1. Assume a $\text{beta}(15,30)$ distribution on the probability of response on the standard treatment and a $\text{beta}(0.6, 1.3)$ prior on the probability of response on the experimental treatment. If five patients have been treated on the experimental treatment and three of these have responded favorably, what is the posterior probability that the standard treatment is more effective?
2. Re-do the example of this section using a $\text{beta}(15,35)$ distribution on θ_S .
3. Consider the generalized stopping criterion, *i.e.*, the random inequality with a δ term. One might argue that the δ term is unnecessary: rather than inserting a δ term, one could replace θ_S with another distribution with the same variance but with its mean shifted by δ . Explain why this proposed method is not equivalent, though it may be approximately equivalent under some circumstances. (Hint: it *would* be equivalent if we were comparing normal random variables rather than beta random variables.)

2.2 Response and toxicity trial monitoring

As before, we are designing a single-arm phase II clinical trial for an experimental treatment E . We will compare the posterior probabilities of response *and toxicity* of patients in the trial on treatment E with historical data on treatment S . The objections to conducting single-arm trials given in section 2.1.3

still apply.

We could model the probabilities of response and toxicity separately as beta random variables. This would be the simplest generalization of the response-only method, and in fact the development presented here ultimately reduces to a pair of beta random variables for each treatment. However, we develop a Dirichlet model because it gives more detailed insight to the method and how it can be generalized to more complex settings. See [Thall and Sung].

In the response-only model, we simply monitored the probability of response. There were two elementary events — response and non-response — and monitoring was based on these directly. In the response and toxicity model there are four elementary events and our monitoring is based not on these elementary events directly but on their unions into compound events. The four elementary events are as follows.

A_1 response: yes, toxicity: yes

A_2 response: yes, toxicity: no

A_3 response: no, toxicity: yes

A_4 response: no, toxicity: no

Thus, response is the union of the two elementary events A_1 and A_2 . Similarly, toxicity is the union of the two elementary outcomes A_1 and A_3 . Denote the four elementary event probabilities by

$$\theta_j = \Pr(A_j)$$

for $j = 1, 2, 3, 4$ so that

$$\theta_1 + \theta_2 + \theta_3 + \theta_4 = 1.$$

The probability of response is given by

$$\pi_R = \theta_1 + \theta_2$$

and the probability of toxicity is given by

$$\pi_T = \theta_1 + \theta_3.$$

Let

$$\mathbf{X} = (X_1, X_2, X_3, X_4)$$

denote the observed vector. That is, at any given point in the trial when n patients have been evaluated, X_j is the number of patients falling in category A_j . For example, X_1 is the number of patients out of the n evaluated who experienced both response and toxicity. The method will require two vectors of probabilities, one for each treatment:

$$\begin{aligned}\boldsymbol{\theta}_E &= (\theta_{E,1}, \theta_{E,2}, \theta_{E,3}, \theta_{E,4}) \\ \boldsymbol{\theta}_S &= (\theta_{S,1}, \theta_{S,2}, \theta_{S,3}, \theta_{S,4}).\end{aligned}$$

Note that the probabilities of response and toxicity on each treatment are given by

$$\begin{aligned}\pi_{E,R} &= \theta_{E,1} + \theta_{E,2} \\ \pi_{E,T} &= \theta_{E,1} + \theta_{E,3} \\ \pi_{S,R} &= \theta_{S,1} + \theta_{S,2} \\ \pi_{S,T} &= \theta_{S,1} + \theta_{S,3}\end{aligned}$$

Monitoring is based on the compound probabilities denoted by π 's and not the elementary probabilities denoted by θ 's.

The vector $\boldsymbol{\theta}_S$ has a $\text{Dir}(\mathbf{a}_S)$ distribution. This distribution is informative since it represents the standard treatment to which the experimental treatment is being compared. As before, the components of \mathbf{a}_S could be based on counts for their corresponding events in historical data, appropriately discounted. The vector $\boldsymbol{\theta}_E$ has a $\text{Dir}(\mathbf{a}_E)$ distribution and is noninformative. As before, we recommend forming the prior for the experimental treatment by scaling the parameters of the distribution for the standard treatment. We recommend that

the Dirichlet parameters for the experimental treatment sum to no more than 4.

The number of observed responses, $X_1 + X_2$, is binomial with parameters n and $\pi_{E,R}$ and the number of observed toxicities, $X_1 + X_3$, is binomial with parameters n and $\pi_{E,T}$. The posterior distributions on $\pi_{E,R}$ and $\pi_{E,T}$ have beta distributions with parameters

$$(a_{E,1} + X_1 + a_{E,2} + X_2, a_{E,3} + X_3 + a_{E,4} + X_4)$$

and

$$(a_{E,1} + X_1 + a_{E,3} + X_3, a_{E,2} + X_2 + a_{E,4} + X_4).$$

Note that $X_1 + X_2$ and $X_1 + X_3$ are not independent, nor are $\pi_{E,R}$ and $\pi_{E,T}$ independent *a posteriori*.

The first beta parameter for the distribution of $\pi_{E,R}$ is the number of prior responses, $a_{E,1} + a_{E,2}$, plus the number of observed responses, $X_1 + X_2$. The second beta parameter is the number of prior non-responses, $a_{E,3} + a_{E,4}$, plus the number of observed non-responses, $X_1 + X_3$. An analogous interpretation applies to $\pi_{E,T}$.

We say that treatment E is worse than treatment S with respect to patient response if

$$P(\pi_{S,R} + \delta_R > \pi_{E,R} \mid \text{data}) > \pi^*$$

for a given large π^* . Typically, π^* is in the range $[0.90, 0.99]$ with 0.95 being a common choice. Similarly, we say treatment E is worse than treatment S with respect to toxicity if

$$P(\pi_{S,T} + \delta_T < \pi_{E,T} \mid \text{data}) > \pi_*.$$

As with π^* , π_* is typically in the range $[0.90, 0.99]$ with 0.95 being a common choice.

Setting $\delta_R > 0$ says that we require the probability of response on the experimental treatment to be better than that on the standard by a factor

δ_R . Setting $\delta_T > 0$ says that we allow the experimental treatment to be more toxic than the standard by an amount δ_T . Together these two terms say that we are willing to accept a certain increase in toxicity for a given increase in response. This is the most common situation. However, one may want the opposite trade-off, allowing a decrease in response in exchange for a decrease in toxicity, corresponding to negative values of both δ_R and δ_T . The most demanding option would be $\delta_R > 0$ and $\delta_T < 0$, requiring an increase in response and a decrease in toxicity. The only sign combination that makes no sense is $\delta_R < 0$ and $\delta_T > 0$. This would say we would allow the experimental treatment to be less responsive and more toxic.

Increasing π^* (π_*) makes the method less likely to stop due to poor responses (increased toxicity). Setting π^* (π_*) to 1 effectively eliminates the response (toxicity) stopping rule.

2.2.1 Example

Suppose for the upcoming trial of treatment E we are limited to a maximum of 30 patients.

Suppose we have searched for historical data on patients similar to those in the upcoming trial of treatment E and found 200 patients. Of these patients, 60 responded to treatment S and 140 did not. We therefore take $\pi_{S,R}$, the historical response distribution on treatment S , to be $\text{beta}(30, 70)$, applying our rule of discounting the historical data by half. Suppose further that of the 200 patients we found, we were only able to determine the toxicity status on 160, but out of those 160, 40 had experienced toxic effects from the drug. We therefore take the historical toxicity distribution to be $\text{beta}(20, 60)$.

Scaling the standard treatment beta parameters to sum to 2 to obtain non-informative priors, we have a $\text{beta}(0.6, 1.4)$ prior distribution on $\pi_{E,R}$ and a $\text{beta}(0.5, 1.5)$ prior on $\pi_{E,T}$.

We choose $\pi^* = \pi_* = 0.95$, the default values provided by the software. Also, we set $\delta_R = \delta_T = 0$.

As before, we are able to determine the stopping boundaries in advance.

For example, we know that under the worse circumstances, the response stopping rule cannot kick in until we have treated at least six patients. If the first five patients do not respond, the posterior probability of $\pi_{E,R}$ being less than $\pi_{E,T}$ is 0.949, but if the first six do not respond, this probability goes up to 0.965. Therefore if no patients out of the first six respond, we stop the trial, but we cannot stop due to response any earlier.

The full set of stopping boundaries for patient response are

Response count	Boundary
0	6
1	12
2	17
3	22
4	27
5	30

This says we stop if we see no response among the first six patients, or only one response among the first 12, or only two among the first 17, etc.

The toxicity stopping boundaries are

Toxicity count	Boundary
3	3
3	4
4	6
5	8
6	10
6	11
7	13
7	14
8	16
8	17
9	19
10	21
10	22
11	24
11	25
12	27
12	28
13	30

This says we stop if all of the first three patients experience toxic effects from the drug, or three among the first four, or four among the first six, etc.

Note that only certain sample sizes are possible. For example, the trial cannot possibly stop after 9 patients. If we treated the 9th patient, we must have had at least one response out of the first 6 patients (since we didn't stop at the first response stopping boundary) and will continue until at least 12 until the response stopping boundary has another chance to stop the trial. Similarly, we must have had fewer than 5 drug toxicities among the first 8 patients, and

the toxicity stopping rule cannot stop the trial until at least 10 patients have been treated.

2.2.2 Continuous monitoring

It has been suggested that continuous monitoring is impractical in phase II trials. However, this is not necessarily true if one looks ahead, not suspending accrual to wait for irrelevant outcomes.

In the example, one could potentially stop the trial at 18 points during the trial, combining the response and toxicity stopping boundaries. However, this does not mean that a trial actually would suspend accrual, waiting for all patient observations to come in, to test the stopping criteria. For example, the trial could stop after three patients if the first three patients all experienced toxicity. Suppose the first patient has passed the observation window and is known to have not experienced toxicity, though the second and third patients are still being observed. There is no need to wait on these two missing outcomes before treating the fourth patient: the missing data could not possibly effect the decision whether to terminate the trial. While there are 18 points where one could conceivably suspend accrual, it is most likely that one would suspend much less often, perhaps never suspending accrual at all.

Continuous monitoring with look-ahead only slows down the trial accrual when it is most desirable to do so, when the performance of the experimental treatment is so marginal that missing data has the potential to activate a stopping rule.

The expected time required to conduct a trial depends on the arrival rate, the observation window, and the probabilities of toxicity and response on the experimental treatment. However, simulation studies have shown that under many common scenarios, continuous monitoring with look-ahead does not slow down a trial significantly. See [Cook 2004]. The accompanying software gives

the user the ability to carry out trial duration simulations.

Chapter 3

Scenarios and simulations

3.1 Philosophical considerations

The subject of scenarios is controversial because it touches on issues of philosophy and tradition.

From a purely Bayesian perspective, one determines the stopping rules for a trial by asking how sure one needs to be that the experimental treatment is inferior before stopping the trial. The required degrees of certainty are expressed in the values of π^* and π_* . Once these values are settled, no more work is necessary. There is no need to examine how the rule works out under specific scenarios. Set the Bayesian parameters and let the scenario chips fall where they may.

The opposite approach is to view the Bayesian machinery as a method for designing frequentist trials. The goal is to get a design in which the probability of early stopping is sufficiently high under a particular bad scenario and sufficiently low during a particular good scenario. The parameters π^* and π_* are not interpreted as probabilities but simply parameters to be adjusted until the desired scenario outcomes are achieved. Set the behavior on a couple scenarios

and let the Bayesian chips fall where they may.

There is a continuum of positions between these extremes. What one might call a “pragmatic” Bayesian position would be to view the stopping rules from a Bayesian perspective, but run a few scenarios in order to understand better how the design performs and to present the design to others who are more accustomed to frequentist designs.

3.1.1 Exercises

1. Elaborate on why a Bayesian may object to letting the operating characteristics on a couple scenarios dictate the values of π^* and π_* .
2. Present a frequentist argument for setting the design parameters to achieve power and specificity.
3. How might a Bayesian and a frequentist disagree about the role of maximum sample size?

3.2 Simulations

This chapter has consistently used the word “scenario” rather than simulation. The model is simple enough that simulation is not necessary in order to calculate the stopping probabilities: the computed probabilities are exact, apart from the limitations of the finite-precision arithmetic used in their calculation.

Here is where the dependent nature of response and toxicity becomes important. The *locations* of the stopping boundaries do not depend on the association between response and toxicity, but the probabilities of *hitting* those stopping boundaries do. For the purpose of calculating stopping boundaries, the Dirichlet structure is irrelevant; one could just as easily assume that the probabilities of response and toxicity are independent beta random variables. But when calculating the probabilities of a trial stopping at various boundaries under various

scenarios, one must take the full Dirichlet model into effect. See Exercise 3 in section 1.4 for an example illustrating this point.

3.2.1 Exercises

1. Suppose a trial of 10 patients is monitoring only response and the early stopping boundaries are $0/3$, $1/6$, and $2/9$. That is, the trial will stop if there is no response among the first three patients, or only one among the first six, or only two among the first nine. If the probability of response is p , what is the probability of the trial treating 10 patients?
2. Suppose a trial of 10 patients is monitoring only toxicity and the early stopping boundaries are $3/3$ and $6/7$. That is, the trial will stop if all of the first three patients or if six out of the first seven experience drug toxicity. If the probability of toxicity is s , what is the probability of the trial treating 10 patients?
3. Combine the preceding exercises into a single trial monitoring both response and toxicity. Assume that toxicity and response are independent random variables.
4. Refer to Exercise 3 in section 1.4. Compute the stopping probabilities in the cases of

$$\mathbf{p} = (0.3, 0.1, 0.0, 0.6)$$

and

$$\mathbf{p} = (0.0, 0.4, 0.3, 0.3).$$

Note that the marginal probabilities of response and toxicity are the same for both scenarios.

Chapter 4

Solutions and references

4.1 Solutions

4.1.1 Beta distribution

The following exercises assume that X is a beta(a, b) random variable.

1. If $a = b = 1$ then X has a uniform distribution.
2. If the graph of f_X is a straight line with a positive slope, then $a = 2$ and $b = 1$.
3. If $a = b$ then X is symmetrically distributed around $1/2$ and so $P(X > 1/2) = 1/2$.
4. If $a = b = 100$, X has approximately the same distribution as a normal random variable with mean $1/2$ and variance $1/801$. This corresponds to a standard deviation of approximately 0.035 . If a sample from X is greater than 0.60 then it is nearly three standard deviations from its mean, an event that occurs with a probability smaller than 0.01 . A direct calculation of the probability without using the normal approximation finds the probability to be 0.00216 .

5. The graph of f_X where $a = 100$ and $b = 20$ has an approximately normal shape centered near $100/120$.
6. The graph of f_X where $a = 0.5$ and $b = 3$ has a vertical asymptote at $x = 0$ and monotonically decreases to a value of zero at $x = 1$.
7. If the graph of f_X has a bathtub shape, a and b must both be less than 1.

4.1.2 Posterior probabilities

1. If X is distributed *a priori* as $\text{beta}(0.3, 0.7)$ then given 6 successes and 4 failures X has a posterior distribution that is $\text{beta}(6.3, 4.7)$.
2. If X is distributed *a priori* as $\text{beta}(1000, 1000)$ then given 60 responses and 40 failures X has a posterior distribution that is $\text{beta}(1060, 1040)$. This posterior distribution has a mean of 0.5048, a variance of 0.000119, and a standard deviation of 0.011. Using a normal approximation, the probability of the variable being more than eight standard deviations from its mean is extremely small. Direct calculation not using a normal approximation shows the probability to be less than 10^{-18} . The moral of this story is that a highly informative prior can stubbornly resist the influence of data.

4.1.3 Random inequalities

Recall that

$$\psi(a_X, b_X, a_Y, b_Y) = P(X > Y)$$

where X and Y are independent random variables with $\text{beta}(a_X, b_X)$ and $\text{beta}(a_Y, b_Y)$ distributions, respectively. Also,

$$\varphi(a_X, b_X, a_Y, b_Y, \delta) = P(X > Y + \delta).$$

1. If any two random variables X and Y are identically distributed then $P(X > Y) = P(Y < X) = 1/2$. [This assumes that $P(X = Y)$ is zero, which is true for beta random variables since they contain no point masses.] Therefore $\psi(a, b, a, b) = 1/2$ for all $a, b > 0$.
2. $\psi(a, b, c, d) + \psi(c, d, a, b) = 1$ because $P(X > Y) + P(Y > X) = 1$.
3. Normal approximation shows that $\psi(30, 70, 90, 10)$ is vanishingly small. X is centered around 0.3 with a standard deviation of 0.045 and Y is centered around 0.9 with a standard deviation of 0.03. In order for a sample from X to be larger than a sample from Y , one or the other sample would have to be many standard deviations away from its mean.
4. As $c \rightarrow \infty$, a $\text{beta}(c, d)$ random variable approaches a point mass at $x = 1$ and so the probability of a sample from a fixed distribution X being larger than a sample from Y goes to zero. $\lim_{c \rightarrow \infty} \psi(a, b, c, d) = 0$.
5. If $\delta > 1$, $P(X > Y + \delta) = 0$ because $X \leq 1$ and $Y \geq 0$. Thus $\varphi(5, 4, 3, 2, 2) = 0$.
6. If $\delta < -1$, $P(X > Y + \delta) = 1$ because $X \geq 0$ and $Y \leq 1$. Thus $\varphi(8, 2, 1, 3, -3) = 1$.
7. $\varphi(a, b, c, d, \delta) = 1$ for $\delta \leq -1$ and $\varphi(a, b, c, d, \delta) = 0$ for $\delta \geq 1$. For $-1 < \delta < 1$, φ is a decreasing function of δ .
8. For any random variables X and Y without point masses and any constant δ ,

$$P(X > Y + \delta) + P(Y > X - \delta) = P(X > Y + \delta) + P(Y + \delta > X) = 1.$$

$$\text{Therefore } \varphi(4, 3, 2, 1, 0.5) + \varphi(2, 1, 4, 3, -0.5) = 1.$$

4.1.4 Multinomial and Dirichlet distributions

1. Suppose there are four possible outcomes for a patient in a clinical trial: A_1 , A_2 , A_3 , and A_4 . The probability of outcome A_3 occurring twice and A_2 once is

$$\frac{3!}{0! 1! 2! 0!} p_1^0 p_2^1 p_3^2 p_4^0 = 3p_2 p_3^2.$$

2. The marginal probability of response is $p_1 + p_2$. The marginal probability of toxicity is $p_1 + p_3$.
3. Suppose a trial stops if either there are three non-responses or three toxicities out of the first three patients. Let N be the event of three non-responses and T be the event of three toxicities.

$$\begin{aligned} P(N \cup T) &= P(N) + P(T) - P(N \cap T) \\ &= (p_3 + p_4)^3 + (p_1 + p_3)^3 - p_3^3 \end{aligned}$$

Note that $P(N \cap T)$ is not determined by marginal probabilities.

4.1.5 Response-only trial monitoring

1. Assume a beta(15,30) distribution on the probability of response on the standard treatment and a beta(0.6, 1.3) prior on the probability of response on the experimental treatment. If five patients have been treated on the experimental treatment and three of these have responded, what is the posterior probability that the standard treatment is more effective? We find the probability that a beta(3.6, 3.3) random variable is larger than a beta(15,30) random variable, which is 0.174.
2. If we use a beta(15,35) distribution on θ_S , the stopping boundaries for the example in this section become

Response count	Boundary
0	6
1	13
2	18
3	24
4	29
5	30

The boundaries do not change much, but some move up a little. This is to be expected: if we are less certain about the historical distribution, we will also be less certain that the standard treatment is better and will require more evidence that the experimental treatment is inferior before we will stop the trial.

3. One cannot change a beta random variable's mean without also changing its distribution's shape. However, for a normal random variable one can shift the mean without changing the shape. If the standard distribution is approximately normal, the proposed method is approximately equivalent. Since the standard distributions are typically highly informative, the parameters are large and the normal approximation is good.

4.1.6 Philosophical considerations

1. Any clinical trial, viewed from any statistical perspective, must balance the interests of the patients and the interests of the investigators. If one is sufficiently convinced that a drug is inferior, one is ethically obligated to stop the trial. However, the standard of "sufficiently convinced" cannot be too low. Otherwise one could run a significant chance of dismissing a superior drug due to bad luck early on in a trial. The Bayesian approach

seeks to address this trade-off directly by specifying probabilities that quantify the degree of certainty required to stop the trial.

From a purely Bayesian perspective, this direct approach is most rational. Any departure from this position has a degree of arbitrariness. Although there is a natural scenario corresponding to a frequentist null hypothesis (i.e., that both experimental and standard treatments have equal characteristics), the choice of an alternative hypothesis is arbitrary. Put another way, out of the infinite space of possible scenarios, how does one choose two scenarios that are so special that the design of the trial should be subject to the operating characteristics for these two scenarios? Why should the amount of certainty of inferiority needed to stop the trial be a function of one's arbitrary choice of scenarios?

2. Power and specificity are well-established criteria for evaluating clinical trial designs. The results of a clinical trial need to be convincing to a wider audience than just the statistician who designed the trial. Editors of medical journals, physicians, and regulatory agents may need to be persuaded by the results of a trial, and these constituencies are likely to be more familiar and more comfortable with traditional frequentist reasoning.
3. How might a Bayesian and a frequentist disagree about the role of maximum sample size? A frequentist design has certain quantitative goals that may not be possible to meet with a given sample size. It will take a certain number of patients, for example, to achieve 5% power and 80% specificity. Without that minimum number, nothing can be done. (There is no logical reason why one couldn't adjust the numbers 0.05 and 0.80, but these numbers are well established in tradition and one does not change them. Instead one might try changing the alternative hypothesis.)

The Bayesian approach is more continuous. One can have a trial of any

size. The more patients, the smaller the variance in the posterior probabilities of toxicity and response on the experimental treatment. But there is no break where n patients are too few but $n+1$ patients are enough. Trials with n patients simply lead to slightly more diffuse posterior distributions than trials with $n+1$ patients.

4.1.7 Simulations

1. Suppose a trial of 10 patients is monitoring only response and the early stopping boundaries are $0/3$, $1/6$, and $2/9$. If the probability of response is p , what is the probability of the trial treating 10 patients?

Let the true probability of response be p , and the probability of failure $q = 1 - p$. The probability of stopping at 3 is q^3 .

Stopping at 6 requires having one response out of the first three and one out of the next three. This has probability $(3pq^2)(q^3) = 3pq^5$.

There are two ways to stop at 9. First, one response out of the first three, one out of the next three, and none out of the final three. Second, get two responses out of the first three, and none out of the next six. Thus the probability of stopping after the 9th patient is $(3pq^2)(3pq^2)(q^3) + (3p^2q)q^6 = 12p^2q^7$.

The probability of treating the 10th patient is the probability of not stopping after the 3rd, 6th, or 9th patients. This equals $1 - q^3 - 3pq^5 - 12p^2q^7$.

2. Suppose a trial of 10 patients is monitoring only toxicity and the early stopping boundaries are $3/3$ and $6/7$. If the probability of toxicity is s , what is the probability of the trial treating 12 patients?

Denote the probability of non-toxicity by $r = 1 - s$. The probability of stopping at 3 is s^3 . The probability of stopping at 7 is the probability of one non-toxicity out of the first three followed by all toxicities for the next

four. This has the probability $3rs^2s^4 = 3rs^6$.

The probability of stopping at 10 is the probability of not stopping at 3 or 7. This equals $1 - s^3 - 3rs^6$.

3. Combine the preceding exercises into a single trial monitoring both response and toxicity. Assume that toxicity and response are independent random variables.

The possible early stopping boundaries are 3, 6, 7, and 9.

The probability of continuing past patient 3 is the product of the probabilities for not stopping for response and not stopping for toxicity. This equals $1 - q^3(1 - s^3)$.

The probability of continuing past patient 6 is the product of the probabilities for not stopping at patient 6 for response and not stopping at patient 3 for toxicity. This equals $(1 - 3pq^5)(1 - s^3)$.

The probability of continuing past patient 7 is the product of the probabilities for not stopping at patient 6 for response and not stopping at patient 7 for toxicity. This equals $(1 - 3pq^5)(1 - 3rs^6)$.

The probability of continuing past patient 9 is the product of the probabilities for not stopping at patient 9 for response and not stopping at patient 7 for toxicity. This equals $(1 - 12p^2q^7)(1 - 3rs^6)$.

4. Combine the preceding exercises into a single trial monitoring both response and toxicity. Assume that toxicity and response are independent random variables.

5. If $\mathbf{p} = (0.3, 0.1, 0.0, 0.6)$ then

$$(p_3 + p_4)^3 + (p_1 + p_3)^3 - p_3^3 = 0.243.$$

- If $\mathbf{p} = (0.0, 0.4, 0.3, 0.3)$ then

$$(p_3 + p_4)^3 + (p_1 + p_3)^3 - p_3^3 = 0.216.$$

4.2 References

The M. D. Anderson technical reports listed below may be found by following the “Research” link on <http://www.mdanderson.org/departments/biostats/>.

John D. Cook Numerical computation of stochastic inequality probabilities. M. D. Anderson Biostatistics and Applied Mathematics technical report UTMDABTR-008-03, 2003.

John D. Cook Simulation results for phase II clinical trial durations. M. D. Anderson Biostatistics and Applied Mathematics technical report UTMDABTR-014-04, 2004.

Peter Thall, Richard Simon, and Eli Estey. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Statistics in Medicine* 14:357-379, 1995.

Peter Thall, and Hsi-Guang Sung. Some extensions and applications of Bayesian strategy for monitoring multiple outcomes in clinical trials. *Statistics in Medicine*, 17:1563-1580, 1998.