
Basics of Bayesian Statistics

Suppose a woman believes she may be pregnant after a single sexual encounter, but she is unsure. So, she takes a pregnancy test that is known to be 90% accurate—meaning it gives positive results to positive cases 90% of the time—and the test produces a positive result.¹ Ultimately, she would like to know the probability she is pregnant, given a positive test ($p(\text{preg} | \text{test } +)$); however, what she knows is the probability of obtaining a positive test result if she is pregnant ($p(\text{test } + | \text{preg})$), and she knows the result of the test.

In a similar type of problem, suppose a 30-year-old man has a positive blood test for a prostate cancer marker (PSA). Assume this test is also approximately 90% accurate. Once again, in this situation, the individual would like to know the probability that he has prostate cancer, given the positive test, but the information at hand is simply the probability of testing positive if he has prostate cancer, coupled with the knowledge that he tested positive.

Bayes' Theorem offers a way to reverse conditional probabilities and, hence, provides a way to answer these questions. In this chapter, I first show how Bayes' Theorem can be applied to answer these questions, but then I expand the discussion to show how the theorem can be applied to probability distributions to answer the type of questions that social scientists commonly ask. For that, I return to the polling data described in the previous chapter.

3.1 Bayes' Theorem for point probabilities

Bayes' original theorem applied to point probabilities. The basic theorem states simply:

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}. \quad (3.1)$$

¹ In fact, most pregnancy tests today have a higher accuracy rate, but the accuracy rate depends on the proper use of the test as well as other factors.

In English, the theorem says that a conditional probability for event B given event A is equal to the conditional probability of event A given event B , multiplied by the marginal probability for event B and divided by the marginal probability for event A .

Proof: From the probability rules introduced in Chapter 2, we know that $p(A, B) = p(A|B)p(B)$. Similarly, we can state that $p(B, A) = p(B|A)p(A)$. Obviously, $p(A, B) = p(B, A)$, so we can set the right sides of each of these equations equal to each other to obtain:

$$p(B|A)p(A) = p(A|B)p(B).$$

Dividing both sides by $p(A)$ leaves us with Equation 3.1.

The left side of Equation 3.1 is the conditional probability in which we are interested, whereas the right side consists of three components. $p(A|B)$ is the conditional probability we are interested in reversing. $p(B)$ is the unconditional (marginal) probability of the event of interest. Finally, $p(A)$ is the marginal probability of event A . This quantity is computed as the sum of the conditional probability of A under all possible events B_i in the sample space: Either the woman is pregnant or she is not. Stated mathematically for a discrete sample space:

$$p(A) = \sum_{B_i \in S_B} p(A | B_i)p(B_i).$$

Returning to the pregnancy example to make the theorem more concrete, suppose that, in addition to the 90% accuracy rate, we also know that the test gives false-positive results 50% of the time. In other words, in cases in which a woman is *not* pregnant, she will test positive 50% of the time. Thus, there are two possible events B_i : $B_1 = \text{preg}$ and $B_2 = \text{not preg}$. Additionally, given the accuracy and false-positive rates, we know the conditional probabilities of obtaining a positive test under these events: $p(\text{test } + | \text{preg}) = .9$ and $p(\text{test } + | \text{not preg}) = .5$. With this information, combined with some “prior” information concerning the probability of becoming pregnant from a single sexual encounter, Bayes’ theorem provides a prescription for determining the probability of interest.

The “prior” information we need, $p(B) \equiv p(\text{preg})$, is the marginal probability of being pregnant, not knowing anything beyond the fact that the woman has had a single sexual encounter. This information is considered prior information, because it is relevant information that exists prior to the test. We may know from previous research that, without any additional information (e.g., concerning date of last menstrual cycle), the probability of conception for any single sexual encounter is approximately 15%. (In a similar fashion, concerning the prostate cancer scenario, we may know that the prostate cancer incidence rate for 30-year-olds is .00001—see Exercises). With this information, we can determine $p(B | A) \equiv p(\text{preg} | \text{test } +)$ as:

$$p(\text{preg} \mid \text{test } +) = \frac{p(\text{test } + \mid \text{preg})p(\text{preg})}{p(\text{test } + \mid \text{preg})p(\text{preg}) + p(\text{test } + \mid \text{not preg})p(\text{not preg})}.$$

Filling in the known information yields:

$$p(\text{preg} \mid \text{test } +) = \frac{(.90)(.15)}{(.90)(.15) + (.50)(.85)} = \frac{.135}{.135 + .425} = .241.$$

Thus, the probability the woman is pregnant, given the positive test, is only .241. Using Bayesian terminology, this probability is called a “posterior probability,” because it is the estimated probability of being pregnant obtained *after* observing the data (the positive test). The posterior probability is quite small, which is surprising, given a test with so-called 90% “accuracy.” However, a few things affect this probability. First is the relatively low probability of becoming pregnant from a single sexual encounter (.15). Second is the extremely high probability of a false-positive test (.50), especially given the high probability of not becoming pregnant from a single sexual encounter ($p = .85$) (see Exercises).

If the woman is aware of the test’s limitations, she may choose to repeat the test. Now, she can use the “updated” probability of being pregnant ($p = .241$) as the new $p(B)$; that is, the prior probability for being pregnant has now been updated to reflect the results of the first test. If she repeats the test and again observes a positive result, her new “posterior probability” of being pregnant is:

$$p(\text{preg} \mid \text{test } +) = \frac{(.90)(.241)}{(.90)(.241) + (.50)(.759)} = \frac{.135}{.135 + .425} = .364.$$

This result is still not very convincing evidence that she is pregnant, but if she repeats the test again and finds a positive result, her probability increases to .507 (for general interest, subsequent positive tests yield the following probabilities: test 4 = .649, test 5 = .769, test 6 = .857, test 7 = .915, test 8 = .951, test 9 = .972, test 10 = .984).

This process of repeating the test and recomputing the probability of interest is the basic process of concern in Bayesian statistics. From a Bayesian perspective, we begin with some prior probability for some event, and we update this prior probability with new information to obtain a posterior probability. The posterior probability can then be used as a prior probability in a subsequent analysis. From a Bayesian point of view, this is an appropriate strategy for conducting scientific research: We continue to gather data to evaluate a particular scientific hypothesis; we do not begin anew (ignorant) each time we attempt to answer a hypothesis, because previous research provides us with *a priori* information concerning the merit of the hypothesis.

3.2 Bayes' Theorem applied to probability distributions

Bayes' theorem, and indeed, its repeated application in cases such as the example above, is beyond mathematical dispute. However, Bayesian statistics typically involves using probability *distributions* rather than point probabilities for the quantities in the theorem. In the pregnancy example, we assumed the prior probability for pregnancy was a known quantity of exactly .15. However, it is unreasonable to believe that this probability of .15 is in fact this precise. A cursory glance at various websites, for example, reveals a wide range for this probability, depending on a woman's age, the date of her last menstrual cycle, her use of contraception, etc. Perhaps even more importantly, even if these factors were not relevant in determining the prior probability for being pregnant, our knowledge of this prior probability is not likely to be perfect because it is simply derived from previous samples and is not a known and fixed population quantity (which is precisely why different sources may give different estimates of this prior probability!). From a Bayesian perspective, then, we may replace this value of .15 with a distribution for the prior pregnancy probability that captures our prior uncertainty about its true value. The inclusion of a prior probability distribution ultimately produces a posterior probability that is also no longer a single quantity; instead, the posterior becomes a probability distribution as well. This distribution combines the information from the positive test with the prior probability distribution to provide an updated distribution concerning our knowledge of the probability the woman is pregnant.

Put generally, the goal of Bayesian statistics is to represent prior uncertainty about model parameters with a probability distribution and to update this prior uncertainty with current data to produce a posterior probability distribution for the parameter that contains less uncertainty. This perspective implies a subjective view of probability—probability represents uncertainty—and it contrasts with the classical perspective. From the Bayesian perspective, any quantity for which the true value is uncertain, including model parameters, can be represented with probability distributions. From the classical perspective, however, it is unacceptable to place probability distributions on parameters, because parameters are assumed to be fixed quantities: Only the data are random, and thus, probability distributions can only be used to represent the data.

Bayes' Theorem, expressed in terms of probability distributions, appears as:

$$f(\theta|\text{data}) = \frac{f(\text{data}|\theta)f(\theta)}{f(\text{data})}, \quad (3.2)$$

where $f(\theta|\text{data})$ is the posterior distribution for the parameter θ , $f(\text{data}|\theta)$ is the *sampling density* for the data—which is proportional to the Likelihood function, only differing by a constant that makes it a proper density function— $f(\theta)$ is the prior distribution for the parameter, and $f(\text{data})$ is the

marginal probability of the data. For a continuous sample space, this marginal probability is computed as:

$$f(\text{data}) = \int f(\text{data}|\theta)f(\theta)d\theta,$$

the integral of the sampling density multiplied by the prior over the sample space for θ . This quantity is sometimes called the “marginal likelihood” for the data and acts as a normalizing constant to make the posterior density proper (but see Raftery 1995 for an important use of this marginal likelihood). Because this denominator simply scales the posterior density to make it a proper density, and because the sampling density is proportional to the likelihood function, Bayes' Theorem for probability distributions is often stated as:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}, \quad (3.3)$$

where the symbol “ \propto ” means “is proportional to.”

3.2.1 Proportionality

As Equation 3.3 shows, the posterior density is proportional to the likelihood function for the data (given the model parameters) multiplied by the prior for the parameters. The prior distribution is often—but not always—normalized so that it is a true density function for the parameter. The likelihood function, however, as we saw in the previous chapter, is not itself a density; instead, it is a product of densities and thus lacks a normalizing constant to make it a true density function. Consider, for example, the Bernoulli versus binomial specifications of the likelihood function for the dichotomous voting data. First, the Bernoulli specification lacked the combinatorial expression to make the likelihood function a true density function for either the data or the parameter. Second, although the binomial representation for the likelihood function constituted a true density function, it only constituted a true density *for the data* and not for the parameter p . Thus, when the prior distribution for a parameter is multiplied by the likelihood function, the result is also not a proper density function. Indeed, Equation 3.3 will be “off” by the denominator on the right side of Equation 3.2, in addition to whatever normalizing constant is needed to equalize the likelihood function and the sampling density $p(\text{data} | \theta)$.

Fortunately, the fact that the posterior density is only proportional to the product of the likelihood function and prior is not generally a problem in Bayesian analysis, as the remainder of the book will demonstrate. However, a note is in order regarding what proportionality actually means. In brief, if a is proportional to b , then a and b only differ by a multiplicative constant. How does this translate to probability distributions? First, we need to keep in mind that, in a Bayesian analysis, model parameters are considered random quantities, whereas the data, having been already observed, are considered fixed quantities. This view is completely opposite that assumed under the

classical approach. Second, we need to recall from Chapter 2 that potential density functions often need to have a normalizing constant included to make them proper density functions, but we also need to recall that this normalizing constant only has the effect of scaling the density—it does not fundamentally change the relative frequencies of different values of the random variable. As we saw in Chapter 2, the normalizing constant is sometimes simply a true constant—a number—but sometimes the constant involves the random variable(s) themselves.

As a general rule, when considering a univariate density, any term, say Q (no matter how complicated), that can be factored away from the random variable in the density—so that all the term(s) involving the random variable are simply multiples of Q —can be considered an irrelevant proportionality constant and can be eliminated from the density without affecting the results.

In theory, this rule is fairly straightforward, but it is often difficult to apply for two key reasons. First, it is sometimes difficult to see whether a term can be factored out. For example, consider the following function for θ :

$$f(\theta) = e^{-\theta+Q}.$$

It may not be immediately clear that Q here is an arbitrary constant with respect to θ , but it is. This function can be rewritten as:

$$f(\theta) = e^{-\theta} \times e^Q,$$

using the algebraic rule that $e^{a+b} = e^a e^b$. Thus, if we are considering $f(\theta)$ as a density function for θ , e^Q would be an arbitrary constant and could be removed without affecting inference about θ . Thus, we could state without loss of information that:

$$f(\theta) \propto e^{-\theta}.$$

In fact, this particular function, without Q , is an exponential density for θ with parameter $\beta = 1$ (see the end of this chapter). With Q , it is proportional to an exponential density; it simply needs a normalizing constant of e^{-Q} so that the function integrates to 1 over the sample space $S = \{\theta : \theta > 0\}$:

$$\int_0^{\infty} e^{-\theta+Q} d\theta = -\frac{1}{e^{\infty-Q}} + e^Q = e^Q.$$

Thus, given that this function integrates to e^Q , e^{-Q} renormalizes the integral to 1.

A second difficulty with this rule is that multivariate densities sometimes make it difficult to determine what is an irrelevant constant and what is not. With Gibbs sampling, as we will discuss in the next chapter and throughout the remainder of the book, we generally break down multivariate densities into univariate conditional densities. When we do this, we can consider all terms not involving the random variable to which the conditional density applies to

be proportionality constants. I will show this shortly in the last example in this chapter.

3.3 Bayes' Theorem with distributions: A voting example

To make the notion of Bayes' Theorem applied to probability distributions concrete, consider the polling data from the previous chapter. In the previous chapter, we attempted to determine whether John F. Kerry would win the popular vote in Ohio, using the most recent CNN/USAToday/Gallup polling data. When we have a sample of data, such as potential votes for and against a candidate, and we assume they arise from a particular probability distribution, the construction of a likelihood function gives us the joint probability of the events, conditional on the parameter of interest: $p(\text{data}|\text{parameter})$. In the election polling example, we maximized this likelihood function to obtain a value for the parameter of interest—the proportion of Kerry voters in Ohio—that maximized the probability of obtaining the polling data we did. That estimated proportion (let's call it K to minimize confusion) was .521. We then determined how uncertain we were about our finding that $K = .521$. To be more precise, we determined under some assumptions how far K may reasonably be from .521 and still produce the polling data we observed.

This process of maximizing the likelihood function ultimately simply tells us how probable the data are under different values for K —indeed, that is precisely what a likelihood function *is*—but our ultimate question is really whether Kerry will win, given the polling data. Thus, our question of interest is “what is $p(K > .5)$,” but the likelihood function gives us $p(\text{poll data} | K)$ —that is, the probability of the data given different values of K .

In order to answer the question of interest, we need to apply Bayes' Theorem in order to obtain a posterior distribution for K and then evaluate $p(K > .5)$ using this distribution. Bayes' Theorem says:

$$f(K|\text{poll data}) \propto f(\text{poll data}|K)f(K),$$

or verbally: The posterior distribution for K , given the sample data, is proportional to the probability of the sample data, given K , multiplied by the prior probability for K . $f(\text{poll data}|K)$ is the likelihood function (or sampling density for the data). As we discussed in the previous chapter, it can be viewed as a binomial distribution with $x = 556$ “successes” (votes for Kerry) and $n - x = 511$ “failures” (votes for Bush), with $n = 1,067$ total votes between the two candidates. Thus,

$$f(\text{poll data}|K) \propto K^{556}(1 - K)^{511}.$$

What remains to be specified to complete the Bayesian development of the model is a prior probability distribution for K . The important question is: How do we do construct a prior?

3.3.1 Specification of a prior: The beta distribution

Specification of an appropriate prior distribution for a parameter is the most substantial aspect of a Bayesian analysis that differentiates it from a classical analysis. In the pregnancy example, the prior probability for pregnancy was said to be a point estimate of .15. However, as we discussed earlier, that specification did not consider that that prior probability is not known with complete certainty. Thus, if we wanted to be more realistic in our estimate of the posterior probability of pregnancy, we could compute the posterior probability under different values for the prior probability to obtain a collection of possible posterior probabilities that we could then consider and compare to determine which estimated posterior probability we thought was more reasonable. More efficiently, we could replace the point estimate of .15 with a probability distribution that represented (1) the plausible values of the prior probability of pregnancy and (2) their relative merit. For example, we may give considerable prior weight to the value .15 with diminishing weight to values of the prior probability that are far from .15.

Similarly, in the polling data example, we can use a distribution to represent our prior knowledge and uncertainty regarding K . An appropriate prior distribution for an unknown proportion such as K is a beta distribution. The pdf of the beta distribution is:

$$f(K | \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} K^{\alpha-1} (1 - K)^{\beta-1},$$

where $\Gamma(a)$ is the gamma function applied to a and $0 < K < 1$.² The parameters α and β can be thought of as prior “successes” and “failures,” respectively. The mean and variance of a beta distribution are determined by these parameters:

$$E(K | \alpha, \beta) = \frac{\alpha}{\alpha + \beta}$$

and

$$\text{Var}(K | \alpha, \beta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

This distribution looks similar to the binomial distribution we have already discussed. The key difference is that, whereas the random variable is x and the key parameter is K in the binomial distribution, the random variable is K and the parameters are α and β in the beta distribution. Keep in mind, however, from a Bayesian perspective, all unknown quantities can be considered random variables.

² The gamma function is the generalization of the factorial to nonintegers. For integers, $\Gamma(a) = (a - 1)!$. For nonintegers, $\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$. Most software packages will compute this function, but it is often unnecessary in practice, because it tends to be part of the normalizing constant in most problems.

How do we choose α and β for our prior distribution? The answer to this question depends on at least two factors. First, how much information prior to this poll do we have about the parameter K ? Second, how much stock do we want to put into this prior information? These are questions that all Bayesian analyses must face, but contrary to the view that this is a limitation of Bayesian statistics, the incorporation of prior information can actually be an advantage and provides us considerable flexibility. If we have little or no prior information, or we want to put very little stock in the information we have, we can choose values for α and β that reduce the distribution to a uniform distribution. For example, if we let $\alpha = 1$ and $\beta = 1$, we get

$$f(p|\alpha = 1, \beta = 1) \propto K^{1-1=0}(1 - K)^{1-1=0} = 1,$$

which is proportional to a uniform distribution on the allowable interval for K ($[0,1]$). That is, the prior distribution is flat, not producing greater *a priori* weight for any value of K over another. Thus, the prior distribution will have little effect on the posterior distribution. For this reason, this type of prior is called “noninformative.”³

At the opposite extreme, if we have considerable prior information and we want it to weigh heavily relative to the current data, we can use large values of α and β . A little algebraic manipulation of the formula for the variance reveals that, as α and β increase, the variance decreases, which makes sense, because adding additional prior information *ought* to reduce our uncertainty about the parameter. Thus, adding more prior successes and failures (increasing both parameters) reduces prior uncertainty about the parameter of interest (K). Finally, if we have considerable prior information but we do not wish for it to weigh heavily in the posterior distribution, we can choose moderate values of the parameters that yield a mean that is consistent with the previous research but that also produce a variance around that mean that is broad.

Figure 3.1 displays some beta distributions with different values of α and β in order to clarify these ideas. All three displayed beta distributions have a mean of .5, but they each have different variances as a result of having α and β parameters of different magnitude. The most-peaked beta distribution has parameters $\alpha = \beta = 50$. The least-peaked distribution is actually flat—uniform—with parameters $\alpha = \beta = 1$. As with the binomial distribution, the beta distribution becomes skewed if α and β are unequal, but the basic idea is the same: the larger the parameters, the more prior information and the narrower the density.

Returning to the voting example, CNN/USAToday/Gallup had conducted three previous polls, the results of which could be treated as prior information.

³ Virtually all priors, despite sometimes being called “noninformative,” impart some information to the posterior distribution. Another way to say this is that claiming ignorance is, in fact, providing some information! However, flat priors generally have little weight in affecting posterior inference, and so they are called noninformative. See Box and Tiao 1973; Gelman et al. 1995; and Lee 1989.

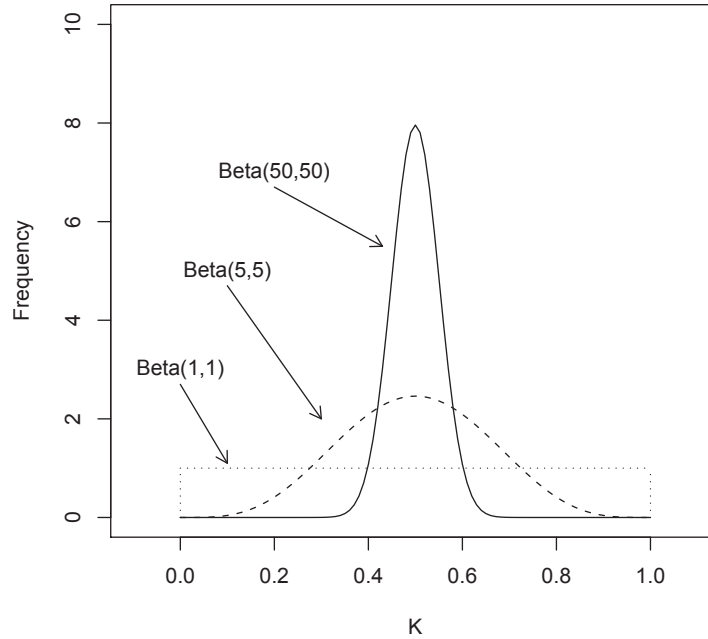


Fig. 3.1. Three beta distributions with mean $\alpha/(\alpha + \beta) = .5$.

These additional polling data are shown in Table 3.1.⁴ If we consider these previous polls to provide us prior knowledge about the election, then our prior information consists of 1,008 (339 + 325 + 344) votes for Bush and 942 votes for Kerry (346 + 312 + 284) out of a total of 1,950 votes.

This prior information can be included by using a beta distribution with parameters $\alpha = 942$ and $\beta = 1008$:

$$f(K | \alpha, \beta) \propto K^{942-1}(1 - K)^{1008-1}.$$

⁴ The data appear to show some trending, in the sense that the proportion stating that they would vote for Bush declined across time, whereas the proportion stating that they would vote for Kerry increased. This fact may suggest consideration of a more complex model than discussed here. Nonetheless, given a margin of error of $\pm 4\%$ for each of these additional polls, it is unclear whether the trend is meaningful. In other words, we could simply consider these polls as repeated samples from the same, unchanging population. Indeed, the website shows the results of 22 polls taken by various organizations, and no trending is apparent in the proportions from late September on.

Table 3.1. CNN/USAToday/Gallup 2004 presidential election polls.

Date	n	% for Bush	$\approx n$	% for Kerry	$\approx n$
Oct 17-20	706	48%	339	49%	346
Sep 25-28	664	49%	325	47%	312
Sep 4-7	661	52%	344	43%	284
TOTAL	2,031		1,008		942

Note: Proportions and candidate-specific sample sizes may not add to 100% of total sample n , because proportions opting for third-party candidates have been excluded.

After combining this prior with the binomial likelihood for the current sample, we obtain the following posterior density for K :

$$p(K \mid \alpha, \beta, x) \propto K^{556}(1 - K)^{511}K^{941}(1 - K)^{1007} = K^{1497}(1 - K)^{1518}.$$

This posterior density is also a beta density, with parameters $\alpha = 1498$ and $\beta = 1519$, and highlights the important concept of “conjugacy” in Bayesian statistics. When the prior and likelihood are of such a form that the posterior distribution follows the same form as the prior, the prior and likelihood are said to be conjugate. Historically, conjugacy has been very important to Bayesians, because, prior to the development of the methods discussed in this book, using conjugate priors/likelihoods with known forms ensured that the posterior would be a known distribution that could be easily evaluated to answer the scientific question of interest.

Figure 3.2 shows the prior, likelihood, and posterior densities. The likelihood function has been normalized as a proper density for K , rather than x . The figure shows that the posterior density is a compromise between the prior distribution and the likelihood (current data). The prior is on the left side of the figure; the likelihood is on the right side; and the posterior is between, but closer to the prior. The reason the posterior is closer to the prior is that the prior contained more information than the likelihood: There were 1,950 previously sampled persons and only 1,067 in the current sample.⁵

With the posterior density determined, we now can summarize our updated knowledge about K , the proportion of voters in Ohio who will vote for Kerry, and answer our question of interest: What is the probability that Kerry would win Ohio? A number of summaries are possible, given that we have a posterior distribution with a known form (a beta density). First, the mean of K is $1498/(1498 + 1519) = .497$, and the median is also .497 (found using the `qbeta` function in R). The variance of this beta distribution is .00008283 (standard deviation=.0091). If we are willing to assume that this beta distribution is approximately normal, then we could construct a 95% interval based on a normal approximation and conclude that the proportion of Ohio voters

⁵ This movement of the posterior distribution away from the prior and toward the likelihood is sometimes called “Bayesian shrinkage” (see Gelman et al. 1995).

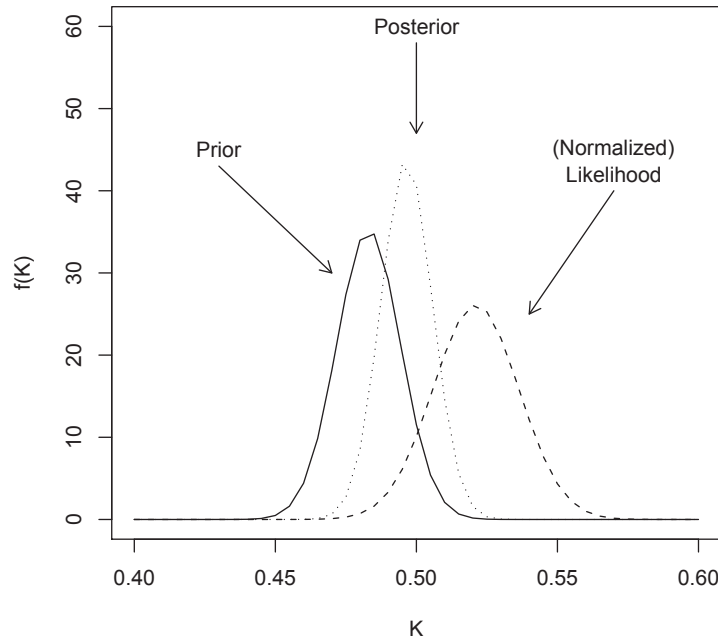


Fig. 3.2. Prior, likelihood, and posterior for polling data example: The likelihood function has been normalized as a density for the parameter K .

who would vote for Kerry falls between .479 and .515 ($.497 \pm 1.96 \times .0091$). This interval is called a “credible interval,” a “posterior probability interval,” or a “probability interval,” and it has a simpler interpretation than the classical confidence interval. Using this interval, we can say simply that the proportion K falls in this interval with probability .95.

If, on the other hand, we are not willing to assume that this posterior density is approximately normal, we can directly compute a 95% probability interval by selecting the lower and upper values of this beta density that produce the desired interval. That is, we can determine the values of this beta density below which 2.5% of the distribution falls and above which 2.5% of the distribution falls. These values are .479 and .514, which are quite close to those under the normal approximation.

These results suggest that, even with the prior information, the election may have been too close to call, given that the interval estimate for K captures .5. However, the substantive question—what is the probability that Kerry would win—can also be answered within the Bayesian framework. This probability is the probability that Kerry will get more than half of the votes, which

is simply the probability that $K > .5$. This probability can be directly computed from the beta distribution as the integral of this density from .5 to 1 (the mass of the curve to the right of .5; see Figure 3.3). The result is .351, which means that Kerry did not have a favorable chance to win Ohio, given the complete polling data.

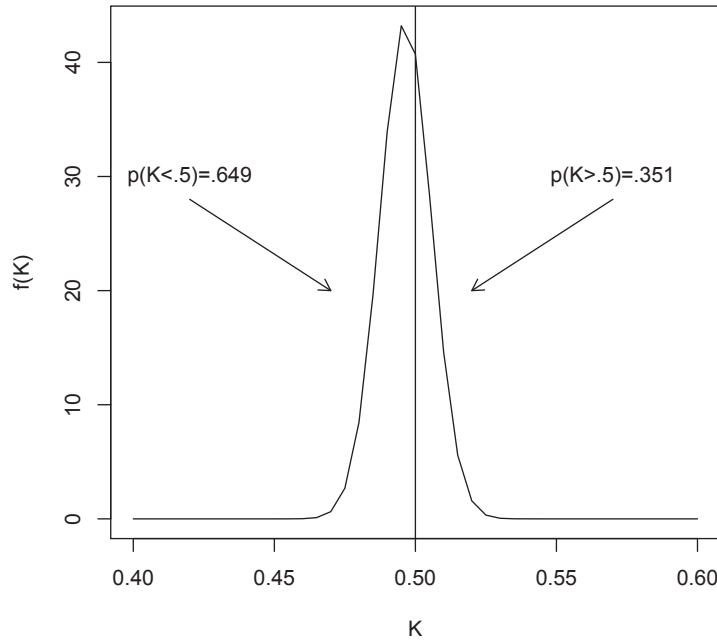


Fig. 3.3. Posterior for polling data example: A vertical line at $K = .5$ is included to show the area needed to be computed to estimate the probability that Kerry would win Ohio.

In fact, Kerry did not win Ohio; he obtained 48.9% of the votes cast for either Kerry or Bush. The classical analysis did not yield this conclusion: It simply suggested that the results were too close to call. The Bayesian analysis, on the other hand, while recognizing that the election would be close, suggested that there was not a very high probability that Kerry would win. The price that had to be paid for reaching this conclusion, however, was (1) we had to be willing to specify a prior probability for K , and (2) we had to be willing to treat the parameter of interest as a random, and not a fixed, quantity.

3.3.2 An alternative model for the polling data: A gamma prior/Poisson likelihood approach

In this section, I repeat the analysis from the previous section. However, instead of considering the problem as a binomial problem with the proportion parameter p , I consider the problem as a Poisson distribution problem with rate parameter λ . As we discussed in the previous chapter, the Poisson distribution is a distribution for count variables; we can consider an individual's potential vote for Kerry as a discrete count that takes values of either 0 or 1. From that perspective, the likelihood function for the 1,067 sample members in the most recent survey prior to the election is:

$$L(\lambda|Y) = \prod_{i=1}^{1067} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} = \frac{\lambda^{\sum_{i=1}^{1067} y_i} e^{-1067\lambda}}{\prod_{i=1}^{1067} y_i!},$$

where y_i is the 0 (Bush) or 1 (Kerry) vote of the i^{th} individual.

As in the binomial example, we would probably like to include the previous survey data in our prior distribution. A conjugate prior for the Poisson distribution is a gamma distribution. The pdf of the gamma distribution is as follows. If $x \sim \text{gamma}(\alpha, \beta)$, then:

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$

The parameters α and β in the gamma distribution are shape and inverse-scale parameters, respectively. The mean of a gamma distribution is α/β , and the variance is α/β^2 . Figure 3.4 shows four different gamma distributions. As the plot shows, the distribution is very flexible: Slight changes in the α and β parameters—which can take any non-negative value—yield highly variable shapes and scales for the density.

For the moment, we will leave α and β unspecified in our voting model so that we can see how they enter into the posterior distribution. If we combine this gamma prior with the likelihood function, we obtain:

$$p(\lambda | Y) \propto \left(\frac{\beta^\alpha}{\Gamma(\alpha)} \right) \lambda^{\alpha-1} e^{-\beta\lambda} \left(\frac{1}{\prod_{i=1}^{1067} y_i!} \right) \lambda^{\sum_{i=1}^{1067} y_i} e^{-1067\lambda}.$$

This expression can be simplified by combining like terms and excluding the arbitrary proportionality constants (the terms in parentheses, which do not include λ) to obtain:

$$p(\lambda | y) \propto \lambda^{\sum_{i=1}^{1067} y_i + \alpha - 1} e^{-(1067 + \beta)\lambda}.$$

Given that each y_i is either a 0 (vote for Bush) or 1 (vote for Kerry), $\sum_{i=1}^{1067} y_i$ is simply the count of votes for Kerry in the current sample (=556). Thus, just as in the binomial example, the parameters α and β —at least in this

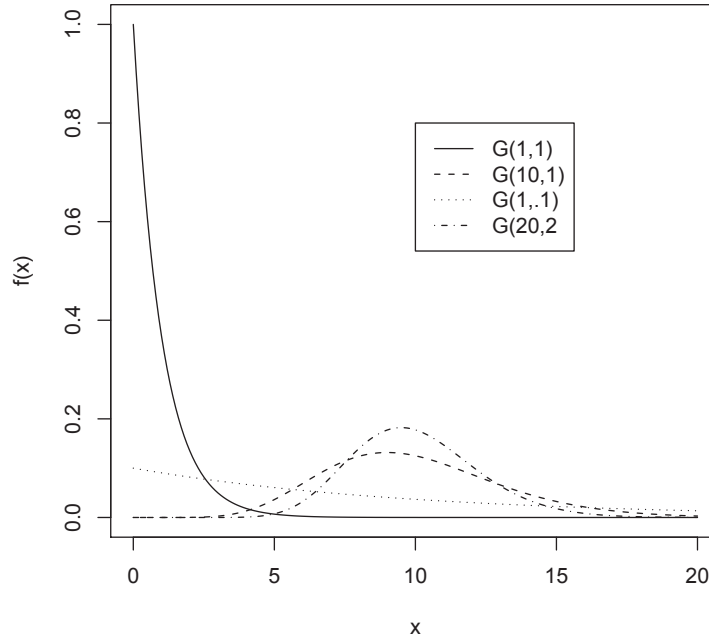


Fig. 3.4. Some examples of the gamma distribution.

particular model—appear to capture prior “successes” and “failures.” Specifically, α is the count of prior “successes,” and β is the total number of prior observations. The mean of the gamma distribution (α/β) also supports this conclusion. Thus, as in the beta prior/binomial likelihood example, if we want to incorporate the data from previous survey into the prior distribution, we can set $\alpha = 942$ and $\beta = 942 + 1008 = 1950$ to obtain the following posterior:

$$p(\lambda | Y) \propto \lambda^{556+942-1} e^{-(1067+1950)\lambda} = \lambda^{1497} e^{-3017\lambda}.$$

Thus, the posterior density is also a gamma density with parameters $\alpha = 1498$ and $\beta = 3017$. Because the gamma density is a known density, we can immediately compute the posterior mean and standard deviation for λ : $\bar{\lambda} = .497$; $\hat{\sigma}_{\lambda} = .0128$. If we wish to construct a 95% probability/credible interval for λ , and we are willing to make a normal approximation given the large sample size, we can construct the interval as $.497 \pm 1.96 \times .0128$. This result gives us an interval estimate of $[.472, .522]$ for λ . On the other hand, if we wish to compute the interval directly using integration of the gamma density (i.e., the cdf for the gamma distribution), we obtain an interval of $[.472, .522]$.

In this case, the normal-theory interval and the analytically derived interval are the same when rounded to three decimal places.

How does this posterior inference compare with that obtained using the beta prior/binomial likelihood approach? The means for K in the beta/binomial approach and for λ in the gamma/Poisson approach are identical. The intervals are also quite comparable, but the interval in this latter approach is wider—about 42% wider. If we wish to determine the probability that Kerry would win Ohio, we simply need to compute $p(\lambda > .5)$, which equals .390. Thus, under this model, Kerry had a probability of winning of .390, which is still an unfavorable result, although it is a slightly greater probability than the beta/binomial result of .351.

Which model is to be preferred? In this case, the substantive conclusion we reached was comparable for the two models: Kerry was unlikely to win Ohio. So, it does not matter which model we choose. The fact that the two models produced comparable results is reassuring, because the conclusion does not appear to be very sensitive to choice of model. Ultimately, however, we should probably place greater emphasis on the beta/binomial model, because the Poisson distribution is a distribution for counts, and our data, which consisted of dichotomous outcomes, really does not fit the bill. Consider the parameter λ : There is no guarantee with the gamma/Poisson setup that λ will be less than 1. This lack of limit could certainly be problematic if we had less data, or if the underlying proportion favoring Kerry were closer to 1. In such a case, the upper bound on the interval for λ may have exceeded 1, and our results would therefore be suspect. In this particular case, however, we had enough data and prior information that ultimately made the interval width very narrow, and so the bounding problem was not an issue. Nonetheless, the beta/binomial setup is a more natural model for the voting data.

3.4 A normal prior–normal likelihood example with σ^2 known

The normal distribution is one of the most common distributions used in statistics by social scientists, in part because many social phenomena in fact follow a normal distribution. Thus, it is not uncommon for a social scientist to use a normal distribution as the basis for a likelihood function for a set of data. Here, I develop a normal distribution problem, but for the sake of keeping this example general for use in later chapters, I used a contrived scenario and keep the mathematics fairly general. The purpose at this point is simply to illustrate a Bayesian approach with a multivariate posterior distribution.⁶

⁶ The normal distribution involves two parameters: the mean (μ) and variance (σ^2). When considered as a density for x , it is univariate, but when a normal likelihood and some prior for the parameters are combined, the result is a joint posterior distribution for μ and σ^2 , which makes the posterior a multivariate density.

Suppose that we have a class of 30 students who have recently taken a midterm exam, and the mean grade was $\bar{x} = 75$ with a standard deviation of $\sigma = 10$. Note that for now we have assumed that the variance is known, hence, the use of σ rather than s . We have taught the course repeatedly, semester after semester, and past test means have given us an overall mean μ of 70, but the class means have varied from class to class, giving us a standard deviation for the class means of $\tau = 5$. That is, τ reflects how much our class means have varied and does not directly reflect the variability of individual test scores. We will discuss this more in depth momentarily.

Our goal is ultimately to update our knowledge of μ , the unobservable population mean test score with the new test grade data. In other words, we wish to find $f(\mu|x)$. Bayes' Theorem tells us that:

$$f(\mu|X) \propto f(X|\mu)f(\mu),$$

where $f(X|\mu)$ is the likelihood function for the current data, and $f(\mu)$ is the prior for the test mean. (At the moment, I am omitting σ^2 from the notation). If we assume the current test scores are normally distributed with a mean equal to μ and variance σ^2 , then our likelihood function for X is:

$$f(X|\mu) \propto L(\mu|X) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\}.$$

Furthermore, our previous test results have provided us with an overall mean of 70, but we are uncertain about μ 's actual value, given that class means vary semester by semester (giving us $\tau = 5$). So our prior distribution for μ is:

$$f(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{-\frac{(\mu - M)^2}{2\tau^2}\right\},$$

where in this expression, μ is the random variable, with M as the prior mean (=70), and τ^2 (=25) reflects the variation of μ around M .

Our posterior is the product of the likelihood and prior, which gives us:

$$f(\mu|X) \propto \frac{1}{\sqrt{\tau^2\sigma^2}} \exp\left\{\frac{-(\mu - M)^2}{2\tau^2} + \frac{-\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}\right\}.$$

This posterior can be reexpressed as a normal distribution for μ , but it takes some algebra in order to see this. First, since the terms outside the exponential are simply normalizing constants with respect to μ , we can drop them and work with the terms inside the exponential function. Second, let's expand the quadratic components and the summations. For the sake of simplicity, I temporarily drop the exponential function in this expression:

$$(-1/2) \left[\frac{\mu^2 - 2\mu M + M^2}{\tau^2} + \frac{\sum x^2 - 2n\bar{x}\mu + n\mu^2}{\sigma^2} \right].$$

Using this expression, any term that does not include μ can be viewed as a proportionality constant, can be factored out of the exponent, and can be dropped (recall that $e^{a+b} = e^a e^b$). After obtaining common denominators for the remaining terms by cross-multiplying by each of the individual denominators and dropping proportionality constants, we are left with:

$$(-1/2) \left[\frac{\sigma^2 \mu^2 - 2\sigma^2 \mu M - 2\tau^2 n \bar{x} \mu + \tau^2 n \mu^2}{\sigma^2 \tau^2} \right].$$

From here, we need to combine terms involving μ^2 and those involving μ :

$$(-1/2) \left[\frac{(n\tau^2 + \sigma^2)\mu^2 - 2(\sigma^2 M + \tau^2 n \bar{x})\mu}{\sigma^2 \tau^2} \right].$$

Dividing the numerator and denominator of this fraction by the $(n\tau^2 + \sigma^2)$ in front of μ^2 yields:

$$(-1/2) \left[\frac{\mu^2 - 2\mu \frac{(\sigma^2 M + n\tau^2 \bar{x})}{(n\tau^2 + \sigma^2)}}{\frac{\sigma^2 \tau^2}{(n\tau^2 + \sigma^2)}} \right].$$

Finally, all we need to do is to complete the square in μ and discard any remaining constants to obtain:

$$(-1/2) \left[\frac{\left(\mu - \frac{(\sigma^2 M + n\tau^2 \bar{x})}{(n\tau^2 + \sigma^2)} \right)^2}{\frac{\sigma^2 \tau^2}{(n\tau^2 + \sigma^2)}} \right].$$

This result shows that our updated μ is normally distributed with mean $(\sigma^2 M + \tau^2 n \bar{x}) / (n\tau^2 + \sigma^2)$ and variance $(\sigma^2 \tau^2) / (n\tau^2 + \sigma^2)$. Notice how the posterior mean is a weighted combination of the prior mean and the sample mean. The prior mean is multiplied by the known variance of test scores in the sample, σ^2 , whereas the sample mean \bar{x} is multiplied by n and by the prior variance τ^2 . This shows first that the sample mean will tend to have more weight than the prior mean (because of the n multiple), but also that the prior and sample variances affect the weighting of the means. If the sample variance is large, then the prior mean has considerable weight in the posterior; if the prior variance is large, the sample mean has considerable weight in the posterior. If the two quantities are equal ($\sigma^2 = \tau^2$), then the calculation reduces to $(M + n\bar{x}) / (n + 1)$, which means that the prior mean will only have a weight of $1 / (n + 1)$ in the posterior.

In this particular example, our posterior mean would be:

$$(100 \times 70) + (25 \times 30 \times 75) / (30 \times 25 + 100) = 74.4.$$

Thus, our result is clearly more heavily influenced by the sample data than by the prior. One thing that must be kept in mind but is easily forgotten is that our updated variance parameter (which is 20—the standard deviation is

therefore 4.47) reflects our uncertainty about μ . This estimate is smaller than both the prior variance and the sample variance, and it is much closer to τ^2 than to σ^2 . Why? Again, this quantity reflects how much μ varies (or, put another way, how much uncertainty we have in knowing M , the true value of μ) and not how much we know about any particular sample. Thus, the fact that our sample standard deviation was 10 does not play a large role in changing our minds about uncertainty in μ , especially given that the sample mean was not that different from the prior mean. In other words, our sample mean is sufficiently close to our prior mean μ so that we are unconvinced that the variance of μ around M should be larger than it was. Indeed, the data convince us that our prior variance should actually be smaller, because the current sample mean is well within the range around M implied by our prior value for τ .

3.4.1 Extending the normal distribution example

The natural extension of the previous example in which the variance σ^2 was considered known is to consider the more realistic case in which the variance is not known. Recall that, ultimately in the previous example, we were interested in the quantity μ —the overall mean test score. Previous data had given us an estimate of μ , but we were still uncertain about its value, and thus, we used τ to represent our uncertainty in μ . We considered σ^2 to be a known quantity (10). In reality, we typically do not know σ^2 any more than we know μ , and thus we have two quantities of interest that we should be updating with new information. A full probability model for μ and σ^2 would look like:

$$f(\mu, \sigma^2 | x) \propto f(x | \mu, \sigma^2) f(\mu, \sigma^2).$$

This model is similar to the one in the example above, but we have now explicitly noted that σ^2 is also an unknown quantity, by including it in the prior distribution. Therefore, we now need to specify a joint prior for both μ and σ^2 , and not just a prior for μ . If we assume μ and σ^2 are independent—and this is a reasonable assumption as we mentioned in the previous chapter; there’s no reason the two parameters need be related—then we can consider $p(\mu, \sigma^2) = p(\mu)p(\sigma^2)$ and establish separate priors for each.

In the example above, we established the prior for μ to be $\mu \sim N(M, \tau^2)$, where M was the prior mean (70) and τ^2 was the measure of uncertainty we had in μ . We did not, however, specify a prior for σ^2 , but we used σ^2 to update our knowledge of τ .⁷

How do we specify a prior distribution for μ and σ^2 in a more general case? Unlike in the previous example, we often do not have prior information about these parameters, and so we often wish to develop noninformative priors for

⁷ Recall from the CLT that $\bar{x} \sim N(\mu, \sigma^2/n)$; thus σ^2 and τ^2 are related: σ^2/n should be an estimate for τ^2 , and so treating σ^2 as fixed yields an updated τ^2 that depends heavily on the new sample data.

them. There are several ways to do this in the normal distribution problem, but two of the most common approaches lead to the same prior. One approach is to assign a uniform prior over the real line for μ and the same uniform prior for $\log(\sigma^2)$. We assign a uniform prior on $\log(\sigma^2)$ because σ^2 is a nonnegative quantity, and the transformation to $\log(\sigma^2)$ stretches this new parameter across the real line. If we transform the uniform prior on $\log(\sigma^2)$ into a density for σ^2 , we obtain $p(\sigma^2) \propto 1/\sigma^2$.⁸ Thus, our joint prior is: $p(\mu, \sigma^2) \propto 1/\sigma^2$.

A second way to obtain this prior is to give μ and σ^2 proper prior distributions (not uniform over the real line, which is improper). If we continue with the assumption that $\mu \sim N(M, \tau^2)$, we can choose values of M and τ^2 that yield a flat distribution. For example, if we let $\mu \sim N(0, 10000)$, we have a very flat prior for μ . We can also choose a relatively noninformative prior for σ^2 by first noting that variance parameters follow an inverse gamma distribution (see the next section) and then choosing values for the inverse gamma distribution that produce a noninformative prior. If $\sigma^2 \sim IG(a, b)$, the pdf appears as:

$$f(\sigma^2|a, b) \propto (\sigma^2)^{-(a+1)} e^{-b/(\sigma^2)}.$$

In the limit, if we let the parameters a and b approach 0, a noninformative prior is obtained as $1/\sigma^2$. Strictly speaking, however, if a and b are 0, the distribution is improper, but we can let both parameters *approach* 0. We can then use this as our prior for σ^2 (that is, $\sigma^2 \sim IG(0, 0)$; $p(\sigma^2) \propto 1/\sigma^2$). There are other ways to arrive at this choice for the prior distribution for μ and σ , but I will not address them here (see Gelman et al. 1995).

The resulting posterior for μ and σ^2 , if we assume a joint prior of $1/\sigma^2$ for these parameters, is:

$$f(\mu, \sigma^2|X) \propto \frac{1}{\sigma^2} \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\}. \quad (3.4)$$

Unlike in the previous example, however, this is a joint posterior density for two parameters rather than one. Yet we can determine the *conditional* posterior distributions for both parameters, using the rule discussed in the previous chapter that, generally, $f(x|y) \propto f(x, y)$.

Determining the form for the posterior density for μ follows the same logic as in the previous section. First, we carry out the product over all observations. Next, we expand the quadratic, eliminate terms that are constant with respect to μ and rearrange the terms with the μ^2 term first. Doing so yields:

⁸ This transformation of variables involves a Jacobian, as discussed in the previous chapter. Let $m = \log(\sigma^2)$, and let $p(m) \propto \text{constant}$. Then $p(\sigma^2) \propto \text{constant} \times J$, where J is the Jacobian of the transformation from m to σ^2 . The Jacobian is then $dm/d\sigma^2 = 1/\sigma^2$. See DeGroot (1986) for a fuller exposition of this process, and see any introductory calculus book for a general discussion of transformations of variables. See Gelman et al. 1995 for further discussion of this prior.

$$f(\mu|X, \sigma^2) \propto \exp \left\{ -\frac{n\mu^2 - 2n\bar{x}\mu}{2\sigma^2} \right\}.$$

Next, to isolate μ^2 , we can divide the numerator and denominator by n . Finally, we can complete the square in μ to find:

$$f(\mu|X, \sigma^2) \propto \exp \left\{ -\frac{(\mu - \bar{x})^2}{2\sigma^2/n} \right\}.$$

This result shows us that the conditional distribution for $\mu|X, \sigma^2 \sim N(\bar{x}, \frac{\sigma^2}{n})$, which should look familiar. That is, this is a similar result to what the Central Limit Theorem in classical statistics claims regarding the sampling distribution for \bar{x} .

What about the posterior distribution for σ^2 ? There are at least two ways to approach this derivation. First, we could consider the conditional distribution for $\sigma^2|\mu, X$. If we take this approach, then we again begin with the full posterior density, but we now must consider all terms that involve σ^2 . If we carry out the multiplication in the posterior density and combine like terms, we obtain:

$$f(\mu, \sigma^2) \propto \frac{1}{(\sigma^2)^{n/2+1}} \exp \left\{ -\frac{\sum (x_i - \mu)^2}{2\sigma^2} \right\}.$$

Referring back to the above description of the inverse gamma distribution, it is clear that, if μ is considered fixed, the conditional posterior density for σ^2 is inverse gamma with parameters $a = n/2$ and $b = \sum (x_i - \mu)^2/2$.

A second way to approach this problem is to consider that the joint posterior density for μ and σ^2 can be factored using the conditional probability rule as:

$$f(\mu, \sigma^2|X) = f(\mu|\sigma^2, X)f(\sigma^2|X).$$

The first term on the right-hand side we have already considered in the previous example with σ^2 considered to be a known, fixed quantity. The latter term, however, is the *marginal* posterior density for σ^2 . Technically, an exact expression for it can be found by integrating the joint posterior density over μ (i.e., $\int f(\mu, \sigma^2)d\mu$.) (see Gelman et al. 1995). Alternatively, we can find an expression proportional to it by factoring Equation 3.4. We know that the distribution for $\mu|\sigma^2, X$ is proportional to a normal density with mean \bar{x} and variance σ^2/n . Thus, if we factor this term out of the posterior, what is left is proportional to the marginal density for σ^2 .

In order to factor the posterior, first, expand the quadratic again to obtain:

$$\frac{1}{(\sigma^2)^{n/2+1}} \exp \left\{ -\frac{\sum x_i^2 - 2n\bar{x}\mu + n\mu^2}{2\sigma^2} \right\}.$$

Next, rearrange terms to put μ^2 first, and divide the numerator and denominator by n . Once again, complete the square to obtain:

$$\frac{1}{(\sigma^2)^{n/2+1}} \exp \left\{ -\frac{(\mu - \bar{x})^2 + \sum x_i^2/n - \bar{x}^2}{2\sigma^2/n} \right\}.$$

We can now separate the two parts of the exponential to obtain:

$$\frac{1}{\sigma} \exp \left\{ -\frac{(\mu - \bar{x})^2}{2\sigma^2/n} \right\} \times \frac{1}{(\sigma^2)^{n/2}} \exp \left\{ \frac{\sum x_i^2 - n\bar{x}^2}{2\sigma^2} \right\}.$$

The first term is the conditional posterior for μ . The latter term is proportional to the marginal posterior density for σ^2 . The numerator in the exponential is the numerator for the computational version of the sample variance, $\sum(x_i - \bar{x})^2$, and so, the result is recognizable as an inverse gamma distribution with parameters $a = (n - 1)/2$ and $b = (n - 1)\text{var}(x)/2$.

3.5 Some useful prior distributions

Thus far, we have discussed the use of a beta prior for proportion parameter p combined with a binomial likelihood function, a gamma prior for a Poisson rate parameter λ , a normal prior for a mean parameter combined with a normal likelihood function for the case in which the variance parameter σ^2 was assumed to be known, and a reference prior of $1/\sigma^2$ —a special case of an inverse gamma distribution—for a normal likelihood function for the case in which neither μ nor σ^2 were assumed to be known. In this section, I discuss a few additional distributions that are commonly used as priors for parameters in social science models. These distributions are commonly used as priors, because they are conjugate for certain sampling densities/likelihood functions. Specifically, I discuss the Dirichlet, the inverse gamma (in some more depth), and the Wishart and inverse Wishart distributions.

One thing that must be kept in mind when considering distributions as priors and/or sampling densities is *what symbols in the density are parameters versus what symbols are the random variables*. For example, take the binomial distribution discussed in Chapter 2. In the binomial mass function, the random variable is represented by x , whereas the parameter is represented by p . However, in the beta distribution, the random variable is represented by p and the parameters are α and β . From a Bayesian perspective, parameters *are* random variables or at least can be treated as such. Thus, what is important to realize is that we may need to change notation in the pdf so that we maintain the appropriate notation for representing the prior distribution for the parameter(s). For example, if we used θ to represent the parameter p in the binomial likelihood function, while p is used as the random variable in the beta distribution, the two distributions, when multiplied together, would contain p , θ , and x , and it would be unclear how θ and p were related. In fact, in the beta-binomial setup, $\theta = p$, but we need to make sure our notation is clear so that that can be immediately seen.

3.5.1 The Dirichlet distribution

Just as the multinomial distribution is a multivariate extension of the binomial distribution, the Dirichlet distribution is a multivariate generalization of the beta distribution. If X is a k -dimensional vector and $X \sim \text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_k)$, then:

$$f(X) = \frac{\Gamma(\alpha_1 + \dots + \alpha_k)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_k)} x_1^{\alpha_1-1} \dots x_k^{\alpha_k-1}.$$

Just as the beta distribution is a conjugate prior for the binomial distribution, the Dirichlet is a conjugate prior for the multinomial distribution. We can see this result clearly, if we combine a Dirichlet distribution as a prior with a multinomial distribution likelihood:

$$\begin{aligned} f(p_1 \dots p_k | X) &\propto f(X | p_1 \dots p_k) f(p_1 \dots p_k) \\ &\propto \text{Multinomial}(X | p_1 \dots p_k) \text{Dirichlet}(p_1 \dots p_k | \alpha_1 \dots \alpha_k) \\ &\propto \text{Dirichlet}(p_1 \dots p_k | \alpha_1 + x_1, \alpha_2 + x_2, \dots, \alpha_k + x_k) \\ &\propto p_1^{\alpha_1+x_1-1} p_2^{\alpha_2+x_2-1} \dots p_k^{\alpha_k+x_k-1}. \end{aligned}$$

Notice how here, as we discussed at the beginning of the section, the vector X in the original specification of the Dirichlet pdf has been changed to a vector p . In this specification, p is the random variable in the Dirichlet distribution, whereas $\alpha_1 \dots \alpha_k$ are the parameters representing prior counts of outcomes in each of the k possible outcome categories.

Also observe how the resulting Dirichlet posterior distribution looks just like the resulting beta posterior distribution, only with more possible outcomes.

3.5.2 The inverse gamma distribution

We have already discussed the gamma distribution in the Poisson/gamma example, and we have briefly discussed the inverse gamma distribution. If $1/x \sim \text{gamma}(\alpha, \beta)$, then $x \sim \text{IG}(\alpha, \beta)$. The density function for the inverse gamma distribution is:

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\beta/x},$$

with $x > 0$. Just as in the gamma distribution, the parameters α and β affect the shape and scale of the curve (respectively), and both must be greater than 0 to make the density proper.

As discussed earlier, the inverse gamma distribution is used as a conjugate prior for the variance in a normal model. If the normal distribution is parameterized with a precision parameter rather than with a variance parameter, where the precision parameter is simply the inverse of the variance, the

gamma distribution is appropriate as a conjugate prior distribution for the precision parameter. In a normal model, if an inverse gamma distribution is used as the prior for the variance, the marginal distribution for the mean is a t distribution.

The gamma and inverse gamma distributions are general distributions; other distributions arise by fixing the parameters to specific values. For example, if α is set to 1, the exponential distribution results:

$$f(x) = (1/\beta)e^{-x/\beta},$$

or, more commonly $f(x) = \beta e^{-\beta x}$, where β is an inverse scale parameter. Under this parameterization, $\beta_{\text{inverse scale}} = 1/\beta_{\text{scale}}$.

If α is set to $v/2$, where v is the degrees of freedom, and β is set to $1/2$, the chi-square distribution results. Setting the parameters equal to the same value in the inverse-gamma distribution yields an inverse-chi-square distribution.

3.5.3 Wishart and inverse Wishart distributions

The Wishart and inverse Wishart distributions are complex in appearance; they are multivariate generalizations of the gamma and inverse gamma distributions, respectively. Thus, just as the inverse gamma is a conjugate prior density for the variance in a univariate normal model, the inverse Wishart is a conjugate prior density for the variance-covariance matrix in a multivariate normal model. With an inverse Wishart distribution for the variance-covariance matrix in a multivariate normal model, the marginal distribution for the mean vector is multivariate t .

If $X \sim \text{Wishart}(S)$, where S is a scale matrix of dimension d , then

$$f(X) \propto |X|^{(v-d-1)/2} \exp\left\{-\frac{1}{2}\text{tr}(S^{-1}X)\right\},$$

where v is the degrees of freedom.

If $X \sim \text{inverse Wishart}(S^{-1})$, then:

$$f(X) \propto |X|^{-(v+d+1)/2} \exp\left\{-\frac{1}{2}\text{tr}(SX^{-1})\right\}.$$

The assumption for both the Wishart and inverse Wishart distributions is that X and S are both positive definite; that is, $z^T X z > 0$ and $z^T S z > 0$ for any non-zero vector z of length d .

3.6 Criticism against Bayesian statistics

As we have seen in the examples, the development of a Bayesian model requires the inclusion of a prior distribution for the parameters in the model. The notion of using prior research or other information to inform a current

analysis and to produce an updated prior for subsequent use seems quite reasonable, if not very appropriate, for the advancement of research toward a more refined knowledge of the parameters that govern social processes. However, the Bayesian approach to updating knowledge of parameters has been criticized on philosophical grounds for more than a century, providing one reason its adoption has been relatively limited in mainstream social science research.

What is in philosophical dispute between Bayesians and classical statisticians includes: (1) whether data and hypotheses (which are simply statements about parameters of distributions⁹) can hold the same status as random variables, and (2) whether the use of a prior probability injects too much subjectivity into the modeling process.

The first standard argument presented against the Bayesian approach is that, because parameters are fixed, it is unreasonable to place a probability distribution on them (they simply are what they are). More formally, parameters and data cannot share the same sample space. However, recall that the Bayesian perspective on probability is that probability is a subjective approach to uncertainty. Whether a parameter is indeed fixed, to a Bayesian, is irrelevant, because we are still uncertain about its true value. Thus, imposing a probability distribution over a parameter space is reasonable, because it provides a method to reflect our uncertainty about the parameter's true value.

Bayesians argue that doing so has some significant advantages. First, as we have seen, Bayesian interval estimates have a clearer and more direct interpretation than classical confidence intervals. That is, we can directly conclude that a parameter falls in some interval with some probability. This is a common but incorrect interpretation of classical confidence intervals, which simply reflect the probability of obtaining an interval estimate that contains the parameter of interest under repeated sampling. Second, the Bayesian approach can naturally incorporate the findings of previous research with the prior, whereas the classical approach to statistics really has no coherent means of using previous results in current analyses beyond assisting with the specification of a hypothesis. That is, the Bayesian approach formalizes the process of hypothesis construction by incorporating it as part of the model. Third, the Bayesian approach more easily allows more detailed summaries concerning parameters. Instead of simply obtaining a maximum likelihood estimate and standard error, we have an entire distribution that can be summarized using various measures (e.g., mean, median, mode, and interquartile range).

⁹ An alternative representation of Bayes' Theorem is $p(\text{Hypothesis} \mid \text{data}) \propto p(\text{data} \mid \text{Hypothesis}) \times p(\text{Hypothesis})$, which shows that, from a Bayesian perspective, we can place a probability (distribution) on a scientific hypothesis. See Jeffreys 1961 for a detailed discussion of the theory of "inverse probability," which describes the Bayesian approach in these terms.

The second general argument that has been advanced against Bayesian analysis is that incorporating a prior injects too much subjectivity into statistical modeling. The Bayesian response to this argument is multifaceted. First, all statistics is subjective. The choice of sampling density (likelihood) to use in a specific project is a subjective determination. For example, when faced with an ordinal outcome, some choose to use a normal likelihood function, leading to the ordinary least squares (OLS) regression model. Others choose a binomial likelihood with a link function, leading to an ordinal logit or probit regression model. These are subjective choices.

Second, the choice of cut-point (α) at which to declare a result “statistically significant” in a classical sense is a purely subjective determination. Also, similarly, the decision to declare a statistically significant result substantively meaningful is a subjective decision.

A third response to the subjectivity criticism is that priors tend to be overwhelmed by data, especially in social science research. The prior distribution generally contributes to the posterior once, whereas data enter into the likelihood function multiple times. As $n \rightarrow \infty$, the prior’s influence on the posterior often becomes negligible.

Fourth, priors can be quite noninformative, obviating the need for large quantities of data to “outweigh” them. In other words, a prior can be made to contribute little information to the posterior. That is, given that the posterior density is simply a weighted likelihood function, where the weighting is imposed by the prior, we can simply choose a prior distribution for the parameters that assigns approximately equal weight to all possible values of the parameters. The simplest noninformative prior that is often used is thus a uniform prior. Use of this prior yields a posterior density that is proportional to the likelihood function. In that case, the mode of the likelihood function (the maximum likelihood estimate) is the same as the Bayesian *maximum a posteriori* (MAP) estimate, and the substantive conclusions reached by both approaches may be similar, only differing in interpretation.

In defense of the classical criticism, although uniform densities for parameters are often used as priors, transformation from one parameterization of a parameter to another may yield an informative prior. However, alternative approaches have been developed for generating noninformative priors, including the development of Jeffreys priors and other priors. These noninformative priors tend to be based on the information matrix and are invariant under parameter transformation. An in-depth discussion of such priors is beyond the scope of this book, given the goal of a general introduction to estimation. For more details, see Gelman et al. (1995) or see Gill (2002) for a more in-depth discussion of the history of the use and construction of noninformative priors.

A fourth response is that the influence of priors can be evaluated after modeling the data to determine whether posterior inference is reasonable. Ultimately, the results of any statistical analysis, whether Bayesian or classical, must be subjectively evaluated to determine whether they are reasonable, and so, the use of informative priors cannot introduce any more subjectivity than

could be included via other means in any analysis. Another response along these lines is that we can use priors to our advantage to examine how powerful the data are at invalidating the prior. For example, we may establish a conservative prior for a regression coefficient that claims that the *a priori* probability for a regression coefficient is heavily concentrated around 0 (i.e., the covariate has no effect on the outcome). We can then examine the strength of the data in rejecting this prior, providing a conservative test of a covariate's effect.

In general, the historical criticisms of Bayesian statistics are philosophical in nature and cannot be conclusively adjudicated. Instead, the rise in the use of Bayesian statistics over the last few decades has largely occurred for pragmatic reasons, including (1) that many contemporary research questions readily lend themselves to a Bayesian approach, and (2) that the development of sampling methods used to estimate model parameters has increased their ease of use. The remaining chapters attempt to demonstrate these points.

3.7 Conclusions

In this chapter, we have developed the basics of the Bayesian approach to statistical inference. First, we derived Bayes' Theorem from the probability rules developed in the previous chapter, and we applied Bayes' Theorem to problems requiring point estimates for probabilities. We then extended the Bayesian approach to handle prior *distributions* for parameters rather than simply point estimates for prior probabilities. The result was that our posterior probability became a distribution, rather than a point estimate. Next, we discussed how to summarize posterior probability distributions, and we demonstrated how to do so using several common examples. Finally, we discussed some common criticisms of the Bayesian approach that have been advanced over the last century, and we reviewed some common Bayesian responses to them. Although the material presented in this chapter is sufficient for gaining a basic understanding of the Bayesian approach to statistics, I recommend several additional sources for more in-depth coverage. I recommend Lee 1989 for an extremely thorough but accessible exposition of the Bayesian paradigm, and I recommend Box and Tiao (1973) for a more advanced exposition.

In the next chapter, we will continue exploring the Bayesian approach to posterior summarization and inference, but we will ultimately focus on multivariate posterior distributions—the most common type of posterior distribution found in social science research—where the multivariate posterior distribution may not be as easy to summarize directly as the univariate posterior densities shown in this chapter.

3.8 Exercises

1. In your own words, state what Bayes' Theorem for point probabilities actually does. For example, refer to Chapter 2 where I defined conditional probability, and use the same sort of discussion to describe how the theorem works.
2. The pregnancy example was completely contrived. In fact, most pregnancy tests today do not have such high rates of false positives. The "accuracy rate" is usually determined by computing the percent of correct answers the test gives; that is, the combined percent of positive results for positive cases and negative results for negative cases (versus false positives and false negatives). Recompute the posterior probability for being pregnant based on an accuracy rate of 90% defined in this manner. Assume that false positives and false negatives occur equally frequently under this 90% rate. What changes in the calculation?
3. Determine the posterior probability that a 30-year-old male has prostate cancer, given (1) a positive PSA test result; (2) a 90% accuracy rate (as defined in the pregnancy example), coupled with a 90% false positive rate; and (3) a prior probability of .00001 for a 30-year-old male having prostate cancer. Based on the result, why might a physician consider *not* testing a 30-year-old male using the PSA test?
4. Find and plot the posterior distribution for a binomial likelihood with $x = 5$ successes out of $n = 10$ trials using at least three different beta prior distributions. Does the prior make a large difference in the outcome—when?
5. Find and plot the posterior distribution for a normal distribution likelihood with a sample mean $\bar{x} = 100$ and variance $\text{var}(x) = 144$ (assume $n = 169$) using at least three different normal priors for the mean. When does the prior make the largest difference in the outcome—when the prior mean varies substantially from the sample mean, or when the prior variance is small or large?
6. Reconsider the pregnancy example from the beginning of the chapter. I showed the posterior probabilities for the second through the tenth subsequent tests. Reproduce these results, using the posterior obtained from the k^{th} test as the prior for the $(k + 1)^{\text{st}}$ test. Next, assume the original prior ($p = .15$) and assume the 10 tests were taken simultaneously and all yielded a positive result. What is the posterior probability for pregnancy? Finally, reconduct the pregnancy example with the 10 positive tests treated simultaneously as the current data, and use a beta prior distribution. Interpret the results.
7. In the 2004 U.S. presidential election, surveys throughout the fall constantly reversed the projected victor. As each survey was conducted, would it have been appropriate to incorporate the results of previous surveys as priors and treat the current survey as new data to update the prior in a Bayesian fashion? If so, do you think a more consistent picture of the

winner would have emerged before the election? If a Bayesian approach would not have been appropriate, why not?

8. Give two simple examples showing a case in which a prior distribution would *not* be overwhelmed by data, regardless of the sample size.
9. Show how the multinomial likelihood and Dirichlet prior are simply a multivariate generalization of the binomial likelihood and beta prior.
10. Show how the Wishart distribution reduces to the gamma distribution when the number of dimensions of the random variable is 1.
11. I said throughout the chapter that the inverse gamma distribution was the appropriate distribution for a variance parameter. It could be said that variance parameter could be considered to be distributed as an inverse chi-square random variable. Both of these statements are true. How?
12. Why can a prior distribution that equals a constant be considered proportional to a uniform distribution?