Bayesian Statistics

A Companion to the Statistics with R Coursera Course

Christine Chai

Last built on 2017-08-16

Contents

Welcome

$0n_0$

This book is a written companion for the Coursera Course 'Bayesian Statistics' from the Statistics with R specialization. Materials and examples from the course are discussed more extensively and extra examples and exercises are provided.

Chapter 1

The Basics of Bayesian Statistics

Bayesian statistics mostly involves **conditional probability**, which is the the probability of an event A given event B, and it can be calculated using the Bayes' rule. The concept of conditional probability is widely used in medical testing, in which false positives and false negatives may occur. A false positive can be defined as a positive outcome on a medical test when the patient does not actually have the disease they are being tested for. In other words, it's the probability of testing positive given no disease. Similarly, a false negative can be defined as a negative outcome on a medical test when the patient does have the disease. In other words, testing negative given disease. Both indicators are critical for any medical decisions.

For how the Bayes' rule is applied, we can set up a prior, then calculate posterior probabilities based on a prior and likelihood. That is to say, the prior probabilities are updated through an iterative process of data collection.

1.1 Bayes' Rule

NEED INTRO TEXT OTHERWISE LABEL MULTIPLIED DEFINED IN LATEX/PDF

1.1.1 Conditional Probabilities & Bayes' Rule

Consider Table 1.1. It shows the results of a poll among 1738 adult Americans. This table allows us to calculate probabilities.

For instance, the probability of an adult American using an online dating site can be calculated as

$$P(\text{using an online dating site}) = \frac{\text{Number that indicated they used an online dating site}}{\text{Total number of people in the poll}} = \frac{225}{1738} \approx 13\%.$$

Table 1.1: Results from a 2015 Gallup poll on the use of online dating sites by age group

	18-29	30-49	50-64	65+	Total
Used online dating site	60	86	58	21	225
Did not use online dating site	255	426	450	382	1513
Total	316	512	508	403	1738

This is the overall probability of using an online dating site. Say, we are now interested in the probability of using an online dating site if one falls in the age group 30-49. Similar to the above, we have

$$P(\text{using an online dating site} \mid \text{in age group 30-49}) = \frac{\text{Number in age group 30-49 that indicated they used an online dating site}}{\text{Total number in age group 30-49}} = \frac{86}{512} \approx 17\%.$$

Here, the pipe symbol '|' means *conditional on*. This is a *conditional probability* as one can consider it the probability of using an online dating site conditional on being in age group 30-49.

We can rewrite this conditional probability in terms of 'regular' probabilities by dividing both numerator and the denominator by the total number of people in the poll. That is,

$$\begin{aligned} &P(\text{using an online dating site} \mid \text{in age group 30-49}) \\ &\quad = \frac{\text{Number in age group 30-49 that indicated they used an online dating site}}{\text{Total number in age group 30-49}} \\ &\quad = \frac{\frac{\text{Number in age group 30-49 that indicated they used an online dating site}}{\text{Total number of people in the poll}}}{\frac{\text{Total number in age group 30-49}}{\text{Total number of people in the poll}}} \\ &\quad = \frac{P(\text{using an online dating site \& falling in age group 30-49})}{P(\text{Falling in age group 30-49})}. \end{aligned}$$

It turns out this relationship holds true for any conditional probability and is known as Bayes' rule:

Definition 1.1 (Bayes' Rule)**.** The conditional probability of the event A conditional on the event B is given by

$$P(A \mid B) = \frac{P(A \,\&\, B)}{P(B)}.$$

Example 1.1. What is the probability that an 18-29 year old from Table 1.1 uses online dating sites?

Note that the question asks a question about 18-29 year olds. Therefore, it conditions on being 18-29 years old. Bayes' rule provides a way to compute this conditional probability:

$$\begin{aligned} &P(\text{using an online dating site} \mid \text{in age group 18-29}) \\ &\quad = \frac{P(\text{using an online dating site \& falling in age group 18-29})}{P(\text{Falling in age group 18-29})} \\ &\quad = \frac{\frac{\text{Number in age group 18-29 that indicated they used an online dating site}}{\text{Total number of people in the poll}}}{\frac{\text{Total number in age group 18-29}}{\text{Total number of people in the poll}}} \\ &\quad = \frac{\text{Number in age group 18-29 that indicated they used an online dating site}}{\text{Total number in age group 18-29}} = \frac{60}{315} \approx 19\%. \end{aligned}$$

1.1.2 Bayes' Rule and Diagnostic Testing

To better understand conditional probabilities and their importance, let us consider an example involving the human immunodeficiency virus (HIV). In the early 1980s, HIV had just been discovered and was rapidly

expanding. There was major concern with the safety of the blood supply. Also, virtually no cure existed making an HIV diagnosis basically a death sentence, in addition to the stigma that was attached to the disease.

These made false positives and false negatives in HIV testing highly undesirable. A *false positive* is when a test returns postive while the truth is negative. That would for instance be that someone without HIV is wrongly diagnosed with HIV, wrongly telling that person they are going to die and casting the stigma on them. A *false negative* is when a test returns negative while the truth is positive. That is when someone with HIV undergoes an HIV test which wrongly comes back negative. The latter poses a threat to the blood supply if that person is about to donate blood.

The probability of a false positive if the truth is negative is called the false positive rate. Similarly, the false negative rate is the probability of a false negative if the truth is positive. Note that both these rates are conditional probabilities: The false positive rate of an HIV test is the probability of a positive result *conditional on* the person tested having no HIV.

The HIV test we consider is an enzyme-linked immunosorbent assay, commonly known as an ELISA. We would like to know the probability that someone (in the early 1980s) has HIV if ELISA tests positive. For this, we need the following information. ELISA's true positive rate (one minus the false negative rate), also referred to as sensitivity, recall, or probability of detection, is estimated as

$$P(\text{ELISA is positive} \mid \text{Person tested has HIV}) = 93\% = 0.93.$$

Its true negative rate (one minus the false positive rate), also referred to as specificity, is estimated as

$$P(\text{ELISA is negative} \mid \text{Person tested has no HIV}) = 99\% = 0.99.$$

Also relevant to our question is the prevalence of HIV in the overall population, which is estimated to be 1.48 out of every 1000 American adults. We therefore assume

$$P(\text{Person tested has HIV}) = \frac{1.48}{1000} = 0.00148. \tag{1.1}$$

Note that the above numbers are estimates. For our purposes, however, we will treat them as if they were exact.

Our goal is to compute the probability of HIV if ELISA is positive, that is $P(\text{Person tested has HIV} \mid \text{ELISA is positive})$. In none of the above numbers did we condition on the outcome of ELISA. Fortunately, Bayes' rule allows is to use the above numbers to compute the probability we seek. Bayes' rule states that

$$\begin{aligned} &P(\text{Person tested has HIV} \mid \text{ELISA is positive}) \\ &\quad = \frac{P(\text{Person tested has HIV} \,\&\, \text{ELISA is positive})}{P(\text{ELISA is positive})}. \end{aligned}$$

The can be derived as follows. For someone to test positive and be HIV positive, that person first needs to be HIV positive and then seconldy test positive. The probability of the first thing happening is $P(\text{HIV positive}) = 0.00148$. The probability of then testing positive is $P(\text{ELISA is positive} \mid \text{Person tested has HIV}) = 0.93$, the true positive rate. This yields for the numerator

$$\begin{aligned} &P(\text{Person tested has HIV} \,\&\, \text{ELISA is positive}) \\ &\quad = P(\text{Person tested has HIV}) P(\text{ELISA is positive} \mid \text{Person tested has HIV}) \\ &\quad = 0.00148 \cdot 0.93 = 0.0013764. \end{aligned}$$

The first step in the above equation is implied by Bayes' rule: By multiplying the left- and right-hand side of Bayes' rule as presented in Section 1.1.1 by $P(B)$, we obtain

$$P(A \mid B)P(B) = P(A \,\&\, B).$$

The denominator in (1.1.2) can be expanded as

$$\begin{aligned} &P(\text{ELISA is positive}) \\ &\quad = P(\text{Person tested has HIV} \,\&\, \text{ELISA is positive}) + P(\text{Person tested has no HIV} \,\&\, \text{ELISA is positive}) \\ &\quad = 0.0013764 + 0.0099852 = 0.0113616 \end{aligned}$$

where we used (**??**) and

$$\begin{aligned} &P(\text{Person tested has no HIV} \,\&\, \text{ELISA is positive}) \\ &= P(\text{Person tested has no HIV})P(\text{ELISA is positive} \mid \text{Person tested has no HIV}) = (1 - P(\text{Person tested has HIV}))\cdot(1 - P(\end{aligned}$$

Putting this all together and inserting into (1.1.2) reveals

$$P(\text{Person tested has HIV} \mid \text{ELISA is positive}) = \frac{0.0013764}{0.0113616} \approx 0.12. \tag{1.2}$$

So even when the ELISA returns positive, the probability of having HIV is only 12%. An important reason why this number is so low is due to the prevalence of HIV. Before testing, one's probability of HIV was 0.148%, so the positive test changes that probability dramatically, but it is still below 50%. That is, it is more likely that one is HIV negative rather than positive after one positive ELISA test.

Questions like the one we just answered (What is the probability of a disease if a test returns positive?) are crucial to make medical diagnoses. As we saw, just the true positive and true negative rates of a test do not tell the full story, but also a disease's prevalence plays a role. Bayes' rule is a tool to synthesize such numbers into a more useful probability of having a disease after a test result.

If the an individual is at a higher risk for having HIV than a randomly sampled person from the population considered, how, if at all, would you expect $P(\text{Person tested has HIV} \mid \text{ELISA is positive})$ to change?

Example 1.2. What is the probability that someone who tests positive does not actually have HIV?

We found in (1.2) that someone who tests positive has a 0.12 probability of having HIV. That implies that the same person has a $1 - 0.12 = 0.88$ probability of not having HIV, despite testing positive.

Example 1.3. If the an individual is at a higher risk for having HIV than a randomly sampled person from the population considered, how, if at all, would you expect $P(\text{Person tested has HIV} \mid \text{ELISA is positive})$ to change?

If the person has a priori a higher risk for HIV and tests positive, then the probability of having HIV must be higher than for someone not at increased risk who also tests positive. Therefore, $P(\text{Person tested has HIV} \mid \text{ELISA is positive}) > 0.12$ where 0.12 comes from (1.2).

One can derive this mathematically by plugging in a larger number in (1.1) than 0.00148, as that number represents the prior risk of HIV. Changing the calculations accordingly shows $P(\text{Person tested has HIV} \mid \text{ELISA is positive}) > 0.12$.

Example 1.4. If the false positive rate of the test is higher than 1%, how, if at all, would you expect $P(\text{Person tested has HIV} \mid \text{ELISA is positive})$ to change?

If the false positive rate increases, the probability of a wrong positive result increases. That means that a positive test result is more likely to be wrong and thus less indicative of HIV. Therefore, the probability of HIV after a positive ELISA goes down such that $P(\text{Person tested has HIV} \mid \text{ELISA is positive}) < 0.12$.

1.1.3 Bayes Updating

In the previous section, we saw that one positive ELISA test yields a probability of having HIV of 12%. To obtain a more convincing probability, one might want to do a second ELISA test after a first one comes up positive. What is the probability of being HIV positive of also the second ELISA test comes back positive?

To solve this problem, we will assume that the correctness of this second test is not influenced by the first ELISA, that is, the tests are independent from each other. This assumption probably does not hold true as it is plausible that if the first test was a false positive, it is more likely that the second one will be one as well. Nonetheless, we stick with the independence assumption for simplicity.

In the last section, we used $P(\text{Person tested has HIV}) = 0.00148$, see (1.1), to compute the probability of HIV after one positive test. If we repeat those steps but now with $P(\text{Person tested has HIV}) = 0.12$, the probability that a person with one positive test has HIV, we exactly obtain the probability of HIV after two positive tests. Repeating the maths from the previous section, involving Bayes' rule, gives

$$\begin{aligned} &P(\text{Person tested has HIV} \mid \text{Second ELISA is also positive}) \\ &\quad = \frac{P(\text{Person tested has HIV})P(\text{Second ELISA is positive} \mid \text{Person tested has HIV})}{P(\text{Second ELISA is also positive})} \\ &\quad = \frac{0.12 \cdot 0.93}{P(\text{Person tested has HIV})P(\text{Second ELISA is positive} \mid \text{Has HIV}) \\ &\qquad + P(\text{Person tested has no HIV})P(\text{Second ELISA is positive} \mid \text{Has no HIV})} \\ &\quad = \frac{0.1116}{0.12 \cdot 0.93 + (1 - 0.12) \cdot (1 - 0.99)} \approx 0.93. \end{aligned}$$

Since we are considering the same ELISA test, we used the same true positive and true negative rates as in Section 1.1.2. We see that two positive tests makes it much more probable for someone to have HIV than when only one test comes up positive.

This process, of using Bayes' rule to update a probability based on an event affecting it, is called Bayes' updating. More generally, the what one tries to update can be considered 'prior' information, sometimes simply called the *prior*. The event providing information about this can also be data. Then, updating this prior using Bayes' rule gives the information conditional on the data, also known as the *posterior*, as in the information *after* having seen the data. Going from the prior to the posterior is Bayes updating.

The probability of HIV after one positive ELISA, 0.12, was the posterior in the previous section as it was an update of the overall prevalence of HIV, (1.1). However, in this section we answered a question where we used this posterior information as the prior. This process of using a posterior as prior in a new problem is natural in the Bayesian framework of updating knowledge based on the data.

Example 1.5. What is the probability that one actually has HIV after testing positive 3 times on the ELISA? Again, assume that all three ELISAs are independent.

Analogous to what we did in this section, we can use Bayes' updating for this. However, now the prior is the probability of HIV after two positive ELISAs, that is $P(\text{Person tested has HIV}) = 0.93$. Analogous to (**??**), the answer follows as

$$
\begin{aligned}
&P(\text{Person tested has HIV} \mid \text{Third ELISA is also positive}) \\
&\quad = \frac{P(\text{Person tested has HIV})P(\text{Third ELISA is positive} \mid \text{Person tested has HIV})}{P(\text{Third ELISA is also positive})} \\
&\quad = \frac{0.93 \cdot 0.93}{P(\text{Person tested has HIV})P(\text{Third ELISA is positive} \mid \text{Has HIV})} \\
&\qquad + P(\text{Person tested has no HIV})P(\text{Third ELISA is positive} \mid \text{Has no HIV}) \\
&\quad = \frac{0.8649}{0.93 \cdot 0.93 + (1 - 0.93) \cdot (1 - 0.99)} \approx 0.999.
\end{aligned}
$$

1.1.4 Bayesian vs. Frequentist Definitions of Probability

The frequentist definition of probability is based on observation of a large number of trials. The probability for an event E to occur is $P(E)$, and assume we get n_E successes out of n trials. Then we have

$$P(E) = \lim_{n \to \infty} \frac{n_E}{n}. \tag{1.3}$$

On the other hand, the Bayesian definition of probability $P(E)$ reflects our prior beliefs, so $P(E)$ can be any probability distribution, provided that it is consistent with all of our beliefs. (For example, we cannot believe that the probability of a coin landing heads is 0.7 and that the probability of getting tails is 0.8, because they are inconsistent.)

The two definitions result in different methods of inference. Using the frequentist approach, we describe the confidence level as the proportion of random samples from the same population that produced confidence intervals which contain the true population parameter. For example, if we generated 100 random samples from the population, and 95 of the samples contain the true parameter, then the confidence level is 95%. Note that each sample either contains the true parameter or does not, so the confidence level is NOT the probability that a given interval includes the true population parameter.

Example 1.6. Based on a 2015 Pew Research poll on 1,500 adults: "We are 95% confident that 60% to 64% of Americans think the federal government does not do enough for middle class people.

The correct interpretation is: 95% of random samples of 1,500 adults will produce confidence intervals that contain the true proportion of Americans who think the federal government does not do enough for middle class people.

Here are two common misconceptions:

- There is a 95% chance that this confidence interval includes the true population proportion.
- The true population proportion is in this interval 95% of the time.

The probability that a given confidence interval captures the true parameter is either zero or one. To a frequentist, the problem is that one never knows whether a specific interval contains the true value with probability zero or one. So a frequentist says that "95% of similarly constructed intervals contain the true value".

The second (incorrect) statement sounds like the true proportion is a value that moves around that is sometimes in the given interval and sometimes not in it. Actually the true proportion is constant, it's the various intervals constructed based on new samples that are different.

The Bayesian alternative is the credible interval, which has a definition that is easier to interpret. Since a Bayesian is allowed to express uncertainty in terms of probability, a Bayesian credible interval is a range for which the Bayesian thinks that the probability of including the true value is, say, 0.95. Thus a Bayesian can say that there is a 95% chance that the credible interval contains the true parameter value.

Example 1.7. The posterior distribution yields a 95% credible interval of 60% to 64% for the proportion of Americans who think the federal government does not do enough for middle class people.

We can say that there is a 95% probability that the proportion is between 60% and 64% because this is a **credible** interval, and more details will be introduced later in the course.

1.2 Inference for a Proportion

1.2.1 Inference for a Proportion: Frequentist Approach

Example 1.8. RU-486 is claimed to be an effective "morning after" contraceptive pill, but is it really effective?

Data: A total of 40 women came to a health clinic asking for emergency contraception (usually to prevent pregnancy after unprotected sex). They were randomly assigned to RU-486 (treatment) or standard therapy (control), 20 in each group. In the treatment group, 4 out of 20 became pregnant. In the control group, the pregnancy rate is 16 out of 20.

Question: How strongly do these data indicate that the treatment is more effective than the control?

To simplify the framework, let's make it a one proportion problem and just consider the 20 total pregnancies because the two groups have the same sample size. If the treatment and control are equally effective, then the probability that a pregnancy comes from the treatment group (p) should be 0.5. If RU-486 is more effective, then the probability that a pregnancy comes from the treatment group (p) should be less than 0.5.

Therefore, we can form the hypotheses as below:

- $p =$ probability that a given pregnancy comes from the treatment group
- $H_0 : p = 0.5$ (no difference, a pregnancy is equally likely to come from the treatment or control group)
- $H_A : p < 0.5$ (treatment is more effective, a pregnancy is less likely to come from the treatment group)

A p-value is needed to make an inference decision with the frequentist approach. The definition of p-value is the probability of observing something *at least* as extreme as the data, given that the null hypothesis (H_0) is true. "More extreme" means in the direction of the alternative hypothesis (H_A).

Since H_0 states that the probability of success (pregnancy) is 0.5, we can calculate the p-value from 20 independent Bernoulli trials where the probability of success is 0.5. The outcome of this experiment is 4 successes in 20 trials, so the goal is to obtain 4 or fewer successes in the 20 Bernoulli trials.

This probability can be calculated exactly from a binomial distribution with $n = 20$ trials and success probability $p = 0.5$. Assume k is the actual number of successes observed, the p-value is

$$P(k \leq 4) = P(k = 0) + P(k = 1) + P(k = 2) + P(k = 3) + P(k = 4)$$

.

```
sum(dbinom(0:4, size = 20, p = 0.5))
```

```
## [1] 0.005908966
```

According to R, the probability of getting 4 or fewer successes in 20 trials is 0.0059. Therefore, given that pregnancy is equally likely in the two groups, we get the chance of observing 4 or fewer preganancy in the treatment group is 0.0059. With such a small probability, we reject the null hypothesis and conclude that the data provide convincing evidence for the treatment being more effective than the control.

Table 1.2: Prior, likelihood, and posterior probabilities for each of the 9 models

Model (p)	0.1000	0.2000	0.3000	0.4000	0.5000	6e-01	0.70	0.80	0.90
Prior $P(model)$	0.0600	0.0600	0.0600	0.0600	0.5200	6e-02	0.06	0.06	0.06
Likelihood $P(data\|model)$	0.0898	0.2182	0.1304	0.0350	0.0046	3e-04	0.00	0.00	0.00
$P(data\|model)$ x $P(model)$	0.0054	0.0131	0.0078	0.0021	0.0024	0e+00	0.00	0.00	0.00
Posterior $P(model\|data)$	0.1748	0.4248	0.2539	0.0681	0.0780	5e-04	0.00	0.00	0.00

1.2.2 Inference for a Proportion: Bayesian Approach

This section uses the same example, but this time we make the inference for the proportion from a Bayesian approach. Recall that we still consider only the 20 total pregnancies, 4 of which come from the treatment group. The question we would like to answer is that how likely is for 4 pregnancies to occur in the treatment group. Also remember that if the treatment and control are equally effective, and the sample sizes for the two groups are the same, then the probability (p) that the pregnancy comes from the treatment group is 0.5.

Within the Bayesian framework, we need to make some assumptions on the models which generated the data. First, p is a probability, so it can take on any value between 0 and 1. However, let's simplify by using discrete cases – assume p, the chance of a pregnancy comes from the treatment group, can take on nine values, from 10%, 20%, 30%, up to 90%. For example, $p = 20\%$ means that among 10 pregnancies, it is expected that 2 of them will occur in the treatment group. Note that we consider all nine models, compared with the frequentist paradigm that whe consider only one model.

Table 1.2 specifies the prior probabilities that we want to assign to our assumption. There is no unique correct prior, but any prior probability should reflect our beliefs prior to the experiement. The prior probabilities should incorporate the information from all relevant research before we perform the current experiement.

This prior incorporates two beliefs: the probability of $p = 0.5$ is highest, and the benefit of the treatment is symmetric. The second belief means that the treatment is equally likely to be better or worse than the standard treatment. Now it is natural to ask how I came up with this prior, and the specification will be discussed in detail later in the course.

Next, let's calculate the likelihood – the probability of observed data for each model considered. In mathematical terms, we have

$$P(\text{data}|\text{model}) = P(k = 4|n = 20, p)$$

The likelihood can be computed as a binomial with 4 successes and 20 trials with p is equal to the assumed value in each model. The values are listed in Table 1.2.

After setting up the prior and computing the likelihood, we are ready to calculate the posterior using the Bayes' rule, that is,

$$P(\text{model}|\text{data}) = \frac{P(\text{model})P(\text{data}|\text{model})}{P(\text{data})}$$

The posterior probability values are also listed in Table 1.2, and the highest probability occurs at $p = 0.2$, which is 42.48%. Note that the priors and posteriors across all models both sum to 1.

In decision making, we choose the model with the highest posterior probability, which is $p = 0.2$. In comparison, the highest prior probability is at $p = 0.5$ with 52%, and the posterior probability of $p = 0.5$ drops to 7.8%. This demonstrates how we update our beliefs based on observed data. Note that the calculation of posterior, likelihood, and prior is unrelated to the frequentist concept (data "at least as extreme as observed").

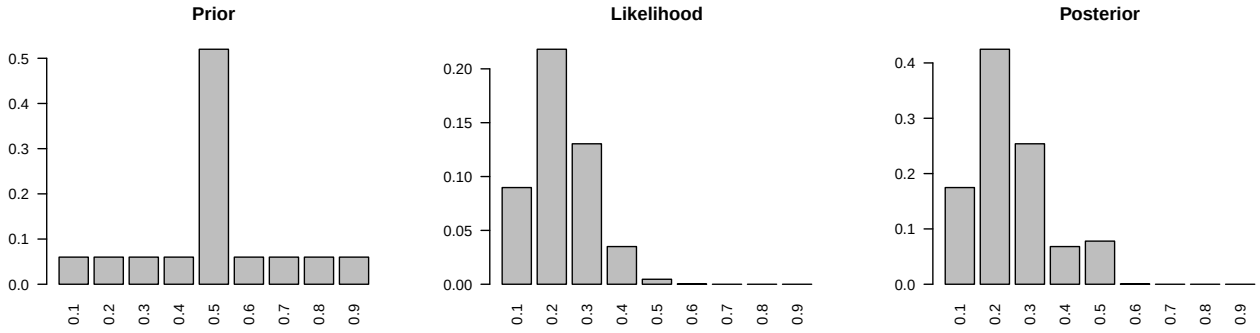


Figure 1.1: Original: sample size $n = 20$ and number of successes $k = 4$

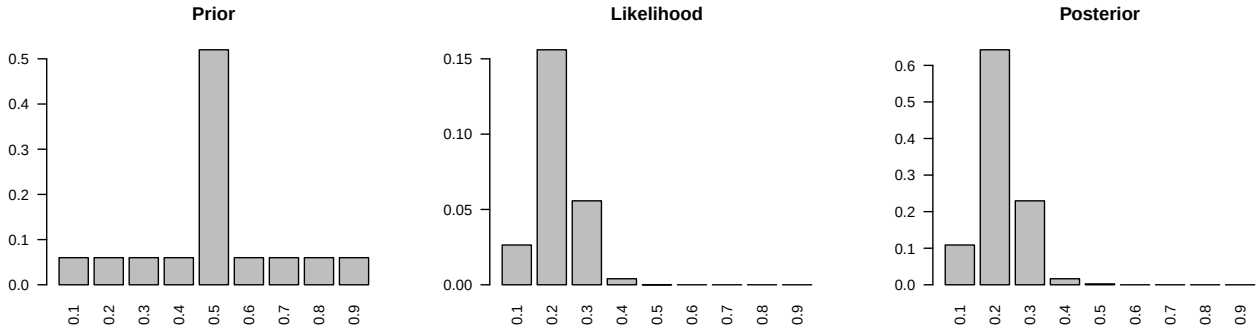


Figure 1.2: More data: sample size $n = 40$ and number of successes $k = 8$

Here are the histograms of the prior, the likelihood, and the posterior probabilities:

We started with the high prior at $p = 0.5$, but the data likelihood peaks at $p = 0.2$. And we updated our prior based on observed data to find the posterior. The Bayesian paradigm, unlike the frequentist approach, allows us to make direct probability statements about our models. For example, we can calculate the probability that RU-486, the treatment, is more effective than the control as the sum of the posteriors of the models where $p < 0.5$. Adding up the relevant posterior probabilities in Table 1.2, we get the chance that the treatment is more effective than the control is 92.16%.

1.2.3 Effect of Sample Size on the Posterior

The RU-486 example is summarized in Figure 1.1, and let's look at what the posterior distribution would look like if we had more data.

Suppose our sample size was 40 instead of 20, and the number of successes was 8 instead of 4. Note that the ratio between the sample size and the number of successes is still 20%. We will start with the same prior distribution. Then calculate the likelihood of the data which is also centered at 0.20, but is less variable than the original likelihood we had with the smaller sample size. And finally put these two together to obtain the posterior distribution. The posterior also has a peak at p is equal to 0.20, but the peak is taller, as shown in Figure 1.2. In other words, there is more mass on that model, and less on the others.

To illustrate the effect of the sample size even further, we're going to keep increasing our sample size, but still maintain the the 20% ratio between the sample size and the number of successes. So let's consider a sample with 200 observations and 40 successes. Once again, we're going to use the same prior and the likelihood is again centered at 20% and almost all of the probability mass in the posterior is at p is equal to 0.20. The other models do not have zero probability mass, but they're posterior probabilities are very close to zero.

Figure 1.3 demonstrates that **as more data are collected, the likelihood ends up dominating the**

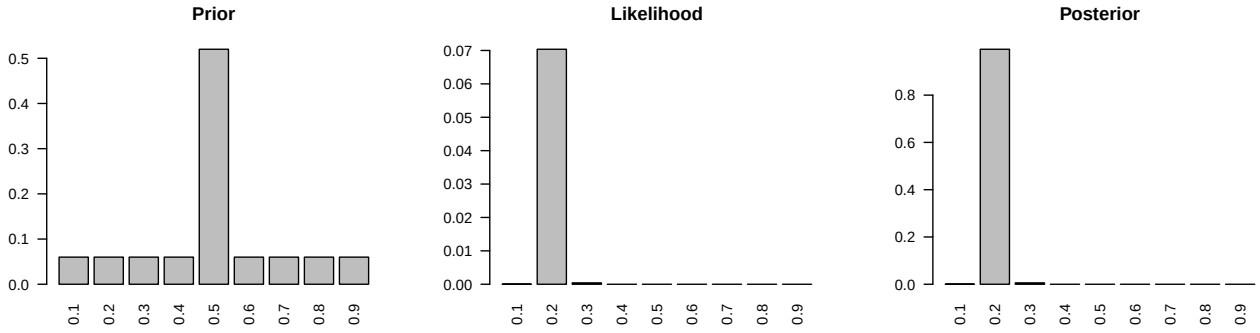


Figure 1.3: More data: sample size $n = 200$ and number of successes $k = 40$

prior. This is why, while a good prior helps, a bad prior can be overcome with a large sample. However, it's important to note that this will only work as long as we don't place a zero probability mass on any of the models in the prior.

1.3 Frequentist vs. Bayesian Inference

1.3.1 Frequentist vs. Bayesian Inference

In this section, we will solve a simple inference problem using both frequentist and Bayesian approaches. Then we will compare our results based on decisions based on the two methods, to see whether we get the same answer or not. If we do not, we will discuss why that happens.

Example 1.9. We have a population of M&M's, and in this population the percentage of yellow M&M's is either 10% or 20%. You've been hired as a statistical consultant to decide whether the true percentage of yellow M&M's is 10% or 20%.

Payoffs/losses: You are being asked to make a decision, and there are associated payoff/losses that you should consider. If you make the correct decision, your boss gives you a bonus. On the other hand, if you make the wrong decision, you lose your job.

Data: You can "buy" a random sample from the population – You pay \$200 for each M&M, and you must buy in \$1,000 increments (5 M&Ms at a time). You have a total of \$4,000 to spend, i.e., you may buy 5, 10, 15, or 20 M&Ms.

Remark: Remember that the cost of making a wrong decision is high, so you want to be fairly confident of your decision. At the same time, though, data collection is also costly, so you don't want to pay for a sample larger than you need. If you believe that you could actually make a correct decision using a smaller sample size, you might choose to do so and save money and resources.

Let's start with the frequentist inference.

- Hypothesis: H_0 is 10% yellow M&Ms, and H_A is >10% yellow M&Ms.
- Significance level: $\alpha = 0.05$.
- Sample: red, green, **yellow**, blue, orange
- Observed data: $k = 1, n = 5$
- P-value: $P(k \geq 1 | n = 5, p = 0.10) = 1 - P(k = 0 | n = 5, p = 0.10) = 1 - 0.90^5 \approx 0.41$

Note that the p-value is the probability of observed or more extreme outcome given that the null hypothesis is true.

Table 1.3: Frequentist and Bayesian probabilities for larger sample sizes

	Frequentist	Bayesian H_1	Bayesian H_2
Observed Data	P(k or more | 10% yellow)	P(10% yellow | n, k)	P(20% yellow | n, k)
n = 5, k = 1	0.41	0.45	0.55
n = 10, k = 2	0.26	0.39	0.61
n = 15, k = 3	0.18	0.34	0.66
n = 20, k = 4	0.13	0.29	0.71

Therefore, we fail to reject H_0 and conclude that the data do not provide convincing evidence that the proportion of yellow M&M's is greater than 10%. This means that if we had to pick between 10% and 20% for the proportion of M&M's, even though this hypothesis testing procedure does not actually confirm the null hypothesis, we would likely stick with 10% since we couldn't find evidence that the proportion of yellow M&M's is greater than 10%.

The Bayesian inference works differently as below.

- Hypotheses: H_1 is 10% yellow M&Ms, and H_2 is 20% yellow M&Ms.
- Prior: $P(H_1) = P(H_2) = 0.5$
- Sample: red, green, **yellow**, blue, orange
- Observed data: $k = 1, n = 5$
- Likelihood:

$$P(k = 1|H_1) = \begin{pmatrix} 5 \\ 1 \end{pmatrix} \times 0.10 \times 0.90^4 \approx 0.33$$
$$P(k = 1|H_2) = \begin{pmatrix} 5 \\ 1 \end{pmatrix} \times 0.20 \times 0.80^4 \approx 0.41$$

- Posterior

$$P(H_1|k = 1) = \frac{P(H_1)P(k = 1|H_1)}{P(k = 1)} = \frac{0.5 \times 0.33}{0.5 \times 0.33 + 0.5 \times 0.41} \approx 0.45$$
$$P(H_2|k = 1) = 1 - 0.45 = 0.55$$

The posterior probabilities of whether H_1 or H_2 is correct are close to each other. As a result, with equal priors and a low sample size, it is difficult to make a decision with a strong confidence, given the observed data. However, H_2 has a higher posterior probability than H_1, so if we had to make a decision at this point, we should pick H_2, i.e., the proportion of yellow M&Ms is 20%. Note that this decision contradicts with the decision based on the frequentist approach.

Table 1.3 summarizes what the results would look like if we had chosen larger sample sizes. Under each of these scenarios, the frequentist method yields a higher p-value than our significance level, so we would fail to reject the null hypothesis with any of these samples. On the other hand, the Bayesian method always yields a higher posterior for the second model where p is equal to 0.20. So the decisions that we would make are contradictory to each other.

However, if we had set up our framework differently in the frequentist method and set our null hypothesis to be $p = 0.20$ and our alternative to be $p < 0.20$, we would obtain different results. This shows that **the frequentist method is highly sensitive to the null hypothesis**, while in the Bayesian method, our results would be the same regardless of which order we evaluate our models.

1.4 Exercises

1. **Conditioning on dating site usage.** Recall Table 1.1. What is the probability that an online dating site user from this sample is 18-29 years old?
2. **Probability of no HIV.** Consider the ELISA test from Section 1.1.2. What is the probability that someone has no HIV if that person has a negative ELISA result? How does this compare to the probability of having no HIV before any test was done?
3. **Probability of no HIV after contradictive tests.** Consider the ELISA test from Section 1.1.2. What is the probability that someone has no HIV if that person first tests positive on the ELISA and secondly test negative? Assume that the tests are independent from each other.

Chapter 2

Bayesian Inference

This chapter is focused on the continuous version of Bayes' rule and how to use it in a conjugate family. The RU-486 example will allow us to discuss Bayesian modeling in a concrete way. It also leads naturally to a Bayesian analysis without conjugacy. For the non-conjugate case, there's usually no simple mathematical expression, and one must resort to computation. Finally, we discuss credible intervals, i.e., the Bayesian analog of frequentist confidence intervals, and Bayesian estimation and prediction.

It is assumed that the readers have mastered the concept of conditional probability and the Bayes' rule for discrete random variables. Calculus is not required for this chapter; however, for those who do, we shall briefly look at an integral.

2.1 Continuous Variables and Eliciting Probability Distributions

Missing: Continuous random variable's pdf plot in 2.1.1

2.1.1 From the Discrete to the Continuous

This section leads the reader from the discrete random variable to continuous random variables. Let's start with the binomial random variable such as the number of heads in ten coin tosses, can only take a discrete number of values – 0, 1, 2, up to 10.

When the probability of a coin landing heads is p, the chance of getting k heads in n tosses is

$$P(X = k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}$$

.

This formula is called the **probability mass function** (pmf) for the binomial.

The probability mass function can be visualized as a histogram in Figure 2.1. The area under the histogram is one, and the area of each bar is the probability of seeing a binomial random variable, whose value is equal to the x-value at the center of the bars base.

In contrast, the normal distribution, a.k.a. Gaussian distribution or the bell-shaped curve, can take any numerical value in $(-\infty, +\infty)$. A random variable generated from a normal distribution because it can take a continuum of values.

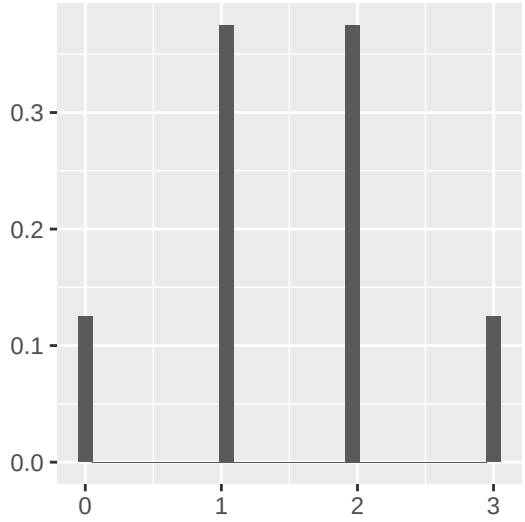


Figure 2.1: Histogram of binomial random variable

In general, if the set of possible values a random variable can take are separated points, it is a discrete random variable. But if it can take any value in some (possibly infinite) interval, then it is a continuous random variable.

When the random variable is **discrete**, it has a **probability mass function** or pmf. That pmf tells us the probability that the random variable takes each of the possible values. But when the random variable is continuous, it has probability zero of taking any single value. (Hence probability zero does not equal to impossible, an event of probabilty zero can still happen.)

We can only talk about the probability of a continuous random variable lined within some interval. For example, suppose that heights are approximately normally distributed. The probability of finding someone who is exactly 6 feet tall at 0.0000 inches tall for an infinite number of 0s after the decimal point is 0. But we can easily calculate the probability of finding someone who is between 5'11" inches tall and 6'1" inches tall.

A **continuous** random variable has a **probability density function** or pdf, instead of probability mass functions. The probability of finding someone whose height lies between 5'11" and 6'1" is the area under the pdf curve for height between those two values.

NEED TO GET THE PLOTS HERE

For example, a normal distribution with mean μ and standard deviation σ (i.e., variance σ^2) is defined as

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2\sigma^2}(x-\mu)^2],$$

where x is any value the random variable X can take. This is denoted as $X \sim N(\mu, \sigma^2)$, where μ and σ^2 are the parameters of the normal distribution.

Recall that a probability mass function assigns the probability that a random variable takes a specific value for the discrete set of possible values. The sum of those probabilities over all possible values must equal one.

Similarly, a probability density function is any $f(x)$ that is non-negative and has area one underneath its curve. The pdf can be regarded as the limit of histograms made from its sample data. As the sample size becomes infinitely large, the bin width of the histogram shrinks to zero.

There are infinite number of pmf's and an infinite number of pdf's. Some distributions are so important that they have been given names:

- Continuous: normal, uniform, beta, gamma
- Discrete: binomial, Poisson

Here is a summary of the key ideas in this section:

1. Continuous random variables exist and they can take any value within some possibly infinite range.
2. The probability that a continuous random variable takes a specific value is zero.
3. Probabilities from a continuous random variable are determined by the density function with this non-negative and the area beneath it is one.
4. We can find the probability that a random variable lies between two values (c and d) as the area under the density function that lies between them.

2.1.2 Elicitation

Next, we introduce the concept of prior elicitation in base and statistics. Often, one has a belief about the distribution of one's data. You may think that your data come from a binomial distribution and in that case you typically know the n, the number of trials but you usually do not know p, the probability of success. Or you may think that your data come from a normal distribution. But you do not know the mean μ or the standard deviation σ of the normal. Beside to knowing the distribution of one's data, you may also have beliefs about the unknown p in the binomial or the unknown mean μ in the normal.

Bayesians express their belief in terms of personal probabilities. These personal probabilities encapsulate everything a Bayesian knows or believes about the problem. But these beliefs must obey the laws of probability, and be consistent with everything else the Bayesian knows.

Example 2.1. You cannot say that your probability of passing this course is 200%, no matter how confident you are. A probability value must be between zero and one. (If you still think you have a probability of 200% to pass the course, you are definitely not going to pass it.)

Example 2.2. You may know nothing at all about the value of p that generated some binomial data. In which case any value between zero and one is equally likely, you may want to make an inference on the proportion of people who would buy a new band of toothpaste. If you have industry experience, you may have a strong belief about the value of p, but if you are new to the industry you would do nothing about p. In any value between zero and one seems equally like a deal. This major personal probability is the uniform distribution whose probably density function is flat, denoted as Unif$(0, 1)$.

Example 2.3. If you were tossing a coin, most people believed that the probability of heads is pretty close to half. They know that some coin are loaded and they know that some coins may have two heads or two tails. And they probably also know that coins are not perfectly balanced. Nonetheless, before they start to collect data by tossing the coin and counting the number of heads their belief is that values of p near 0.5 are very likely, where's values of p near 0 or 1 are very unlikely.

Example 2.4. In real life, here are two ways to elicit a probability that you cousin will get married. A frequentist might go to the U.S. Census records and determine what proportion of people get married (or, better, what proportion of people of your cousin's ethnicity, education level, religion, and age cohort are married). In contrast, a Bayesian might think "My cousin is brilliant, attractive, and fun. The probability that my cousin gets married is really high – probably around 0.97."

So a base angle sits to express their belief about the value of p through a probability distribution, and a very flexible family of distributions for this purpose is the **beta family**. A member of the beta family is specified by two parameters, α and β; we denote this as $p \sim \text{beta}(\alpha, \beta)$. The probability density function is

$$f(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \tag{2.1}$$

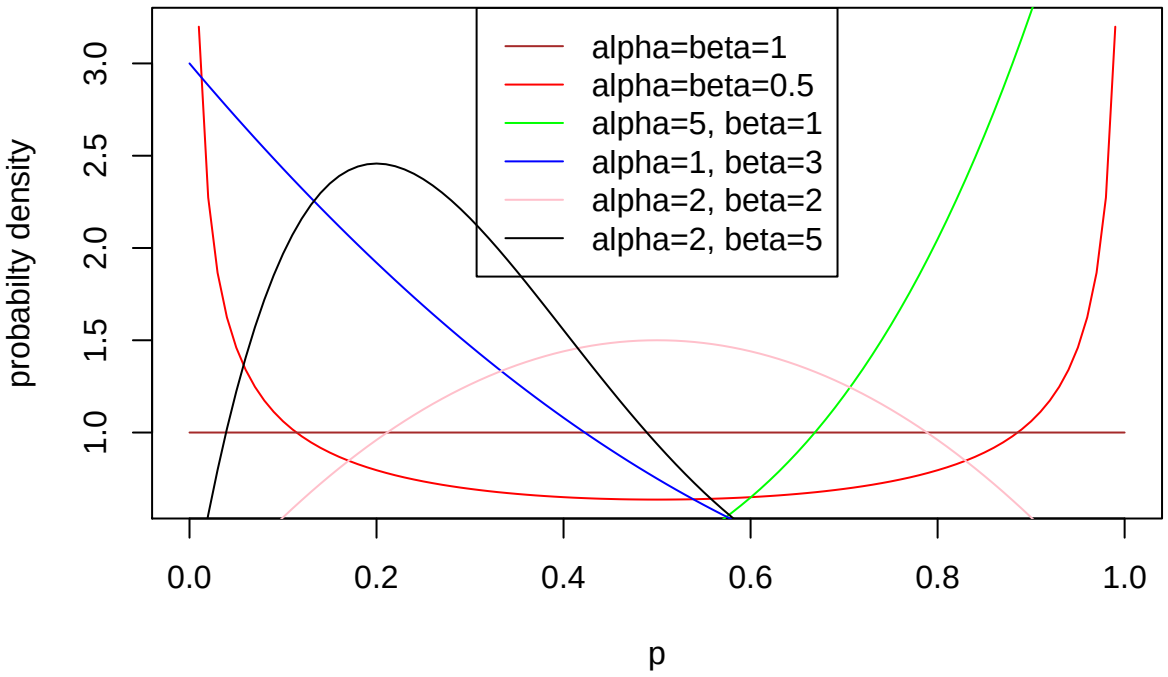


Figure 2.2: Beta family

where $0 \leq p \leq 1, \alpha > 0, \beta > 0$, and Γ is a factorial:

$$\Gamma(n) = (n-1)! = (n-1) \times (n-2) \times \cdots \times 1$$

When $\alpha = \beta = 1$, the beta distribution becomes a uniform distribution, i.e. the probabilty density function is a flat line. In other words, the uniform distribution is a special case of the beta family.

The expected value of p is $\frac{\alpha}{\alpha+\beta}$, so α can be regarded as the prior number of successes, and β the prior number of failures. When $\alpha = \beta$, then one gets a symmetrical pdf around 0.5. For large but equal values of α and β, the area under the beta probability density near 0.5 is very large. Figure 2.2 compares the beta distribution with different parameter values.

These kinds of priors are probably appropriate if you want to infer the probability of getting heads in a coin toss. The beta family also includes skewed densities, which is appropriate if you think that p the probability of success in ths binomial trial is close to zero or one.

Bayes' rule is a machine to turn one's prior beliefs into posterior beliefs. With binomial data you start with whatever beliefs you may have about p, then you observe data in the form of the number of head, say 20 tosses of a coin with 15 heads.

Next, Bayes' rule tells you how the data changes your opinion about p. The same principle applies to all other inferences. You start with your prior probability distribution over some parameter, then you use data to update that distribution to become the posterior distribution that expresses your new belief.

These rules ensure that the change in distributions from prior to posterior is the uniquely rational solution. So, as long as you begin with the prior distribution that reflects your true opinion, you can hardly go wrong.

However, expressing that prior can be difficult. There are proofs and methods whereby a rational and coherent thinker can self-illicit their true prior distribution, but these are impractical and people are rarely rational and coherent.

The good news is that with the few simple conditions no matter what part distribution you choose. If enough data are observed, you will converge to an accurate posterior distribution. So, two bayesians, say the

reference Thomas Bayes and the agnostic Ajay Good can start with different priors but, observe the same data. As the amount of data increases, they will converge to the same posterior distribution.

Here is a summary of the key ideas in this section:

1. Bayesians express their uncertainty through probability distributions.
2. One can think about the situation and self-elicit a probability distribution that approximately reflects his/her personal probability.
3. One's personal probability should change according Bayes' rule, as new data are observed.
4. The beta family of distribution can describe a wide range of prior beliefs.

2.1.3 Conjugacy

Next, let's introduce the concept of conjugacy in Bayesian statistics.

Suppose we have the prior beliefs about the data as below:

- Binomial distribution $\text{Bin}(n, p)$ with n known and p unknown
- Prior belief about p is $\text{beta}(\alpha, \beta)$

Then we observe x success in n trials, and it turns out the Bayes' rule implies that our new belief about the probability density of p is also the beta distribution, but with different parameters. In mathematical terms,

$$p|x \sim \text{beta}(\alpha + x, \beta + n - x). \tag{2.2}$$

This is an example of conjugacy. Conjugacy occurs when the **posterior distribution** is in the **same family** of probability density functions as the prior belief, but with **new parameter values**, which have been updated to reflect what we have learned from the data.

Why are the beta binomial families conjugate? Here is a mathematical explanation.

Recall the discrete form of the Bayes' rule:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)}$$

However, this formula does not apply to continuous random variables, such as the p which follows a beta distribution, because the denominator sums over all possible values (must be finitely many) of the random variable.

But the good news is that the p has a finite range – it can tak any value **only** between 0 and 1. Hence we can perform integration, which is a generalization of the summation. The Bayes' rule can also be written in continuous form as:

$$\pi^*(p|x) = \frac{P(x|p)\pi(p)}{\int_0^1 P(x|p)\pi(p)dp}.$$

This is analogus to the discrete form, since the integral in the denominator will also be equal to some constant, just like a summation. This constant ensures that the total area under the curve, i.e. the posterior density function, equals 1.

Note that in the numerator, the first term, $P(x|p)$, is the data likelihood – the probability of observing the data given a specific value of p. The second term, $\pi(p)$, is the probability density function that reflects the prior belief about p.

In the beta-binomial case, we have $P(x|p) = \text{Bin}(n, p)$ and $\pi(p) = \text{beta}(\alpha, \beta)$.

Plugging in these distributions, we get

$$\begin{aligned}\pi^*(p|x) &= \frac{1}{\text{some number}} \times P(x|p)\pi(p) \\ &= \frac{1}{\text{some number}}[\binom{n}{x} p^x(1-p)^{n-x}][\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}] \\ &= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)} \times p^{\alpha+x-1}(1-p)^{\beta+n-x-1}\end{aligned}$$

Let $\alpha^* = \alpha + x$ and $\beta^* = \beta + n - x$, and we get

$$\pi^*(p|x) = \text{beta}(\alpha^*, \beta^*) = \text{beta}(\alpha + x, \beta + n - x),$$

same as the posterior formula in Equation (2.2).

We can recognize the posterior distribution from the numerator $p^{\alpha+x-1}$ and $(1-p)^{\beta+n-x-1}$. Everything else are just constants, and they must take the unique value, which is needed to ensure that the area under the curve between 0 and 1 equals 1. So they have to take the values of the beta, which has parameters $\alpha + x$ and $\beta + n - x$.

This is a cute trick. We can find the answer without doing the integral simply by looking at form of the numerator.

Without conjugacy, one has to do the integral. Often, the integral is impossible to evaluate. That obstacle is the primary reason that most statistical theory in the 20th century was not Bayesian. The situation didn't change until modern computing allowed researchers to compute integrals numerically.

In summary, some pairs of distributions are conjugate. If your prior is in one and your data comes from the other, then your posterior is in the same family as the prior, but with new parameters. We explored this in the context of the beta-binomial conjugate families. And we saw that conjugacy meant that we could apply the continuous version of Bayes' rule without having to do any integration.

2.2 Three Conjugate Families

Missing: Gamma plot in 2.2.2

2.2.1 Inference on a Binomial Proportion

Example 2.5. Recall Example 1.8, a simplified version of a real clinical trial taken in Scotland. It concerned RU-486, a morning after pill that was being studied to determine whether it was effective at preventing unwanted pregnancies. It had 800 women, each of whom had intercourse no more than 72 hours before reporting to a family planning clinic to seek contraception.

Half of these women were randomly assigned to the standard contraceptive, a large dose of estrogen and progesterone. And half of the women were assigned RU-486. Among the RU-486 group, there were no pregnancies. Among those receiving the standard therapy, four became pregnant.

Statistically, one can model these data as coming from a binomial distribution. Imagine a coin with two sides. One side is labeled standard therapy and the other is labeled RU-486. The coin was tossed four times, and each time it landed with the standard therapy side face up.

A frequentist would analyze the problem as below:

- The parameter p is the probability of a preganancy comes from the standard treatment.
- $H_0 : p \geq 0.5$ and $H_A : p < 0.5$
- The p-value is $0.5^4 = 0.0625 > 0.05$

Therefore, the frequentist fails to reject the null hypothesis, and will not conclude that RU-486 is superior to standard therapy.

Remark: The significance probability, or p-value, is the chance of observing data that are as or more supportive of the alternative hypothesis than the data that were collected, when the null hypothesis is true.

Now suppose a Bayesian performed the analysis. She may set her beliefs about the drug and decide that she has no prior knowledge about the efficacy of RU-486 at all. This would be reasonable if, for example, it were the first clinical trial of the drug. In that case, she would be using the uniform distribution on the interval from 0 to 1, which corresponds to the beta$(1, 1)$ density. In mathematical terms,

$$p \sim \text{Unif}(0, 1) = \text{beta}(1, 1).$$

From conjugacy, we know that since there were four failures for RU-486 and no successes, that her posterior probability of an RU-486 child is

$$p|x \sim \text{beta}(1 + 0, 1 + 4) = \text{beta}(1, 5).$$

This is a beta that has much more area near p equal to 0. The mean of beta(α, β) is $\frac{\alpha}{\alpha+\beta}$. So this Bayesian now believes that the unknown p, the probability of an RU-468 child, is about 1 over 6.

The standard deviation of a beta distribution with parameters in alpha and beta also has a closed form:

$$p \sim \text{beta}(\alpha, \beta) \Rightarrow \text{Standard deviation} = \sqrt{\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}}$$

Before she saw the data, the Bayesian's uncertainty expressed by her standard deviation was 0.71. After seeing the data, it was much reduced – her posterior standard deviation is just 0.13.

We promised not to do much calculus, so I hope you will trust me to tell you that this Bayesian now believes that her posterior probability that $p < 0.5$ is 0.96875. She thought there was a 50-50 chance that RU-486 is better. But now she thinks there's about a 97% chance that RU-486 is better.

Suppose a fifth child were born, also to a mother who received standard chip therapy. Now the Bayesian's prior is beta(1, 5) and the additional data point further updates her to a new posterior beta of 1 and 6. **As data comes in, the Bayesian's previous posterior becomes her new prior, so learning is self-consistent.**

This example has taught us several things:

1. We saw how to build a statistical model for an applied problem.
2. We could compare the frequentist and Bayesian approaches to inference and see large differences in the conclusions.
3. We saw how the data changed the Bayesian's opinion with a new mean for p and less uncertainty.
4. We learned that Bayesian's continually update as new data arrive. **Yesterday's posterior is today's prior.**

2.2.2 The Gamma-Poisson Conjugate Families

A second important case is the gamma-Poisson conjugate families. In this case the data come from a Poisson distribution, and the prior and posterior are both gamma distributions.

The Poisson random variable can take any **non-negative integer value** all the way up to infinity. It is used in describing **count data**, where one counts the number of independent events that occur in a fixed amount of time, a fixed area, or a fixed volume.

Moreover, the Poisson distribution has been used to describe the number of phone calls one receives in an hour. Or, the number of pediatric cancer cases in the city, for example, to see if pollution has elevated the cancer rate above that of in previous years or for similar cities. It is also used in medical screening for diseases, such as HIV, where one can count the number of T-cells in the tissue sample.

The Poisson distribution has a single parameter λ, and it is denoted as $X \sim \text{Pois}(\lambda)$ with $\lambda > 0$. The probability mass function is

$$P(X = k) = \frac{\lambda^k}{k!} \exp^{-\lambda} \text{ for } k = 0, 1, \cdots,$$

where $k! = k \times (k-1) \times \cdots \times 1$. This gives the probability of observing a random variable equal to k.

Note that λ is both the mean and the variance of the Poisson random variable. It is obvious that λ must be greater than zero, because it represents the mean number of counts, and the variance should be greater than zero (except for constants, which have zero variance).

Example 2.6. Famously, von Bortkiewicz used the Poisson distribution to study the number of Prussian cavalrymen who were kicked to death by a horse each year. This is count data over the course of a year, and the events are probably independent, so the Poisson model makes sense.

He had data on 15 cavalry units for the 20 years between 1875 and 1894, inclusive. The total number of cavalrymen who died by horse kick was 200.

One can imagine that a Prussian general might want to estimate λ. The average number per year, per unit. Perhaps in order to see whether some educational campaign about best practices for equine safety would make a difference.

Suppose the Prussian general is a Bayesian. Introspective elicitation leads him to think that $\lambda = 0.75$ and standard deviation 1.

Modern computing was unavailable at that time yet, so the general will need to express his prior as a member of a family conjugate to the Poisson. It turns out that this family consists of the gamma distributions. Gamma distributions describe continuous non-negative random variables. As we know, the value of lambda in the Poisson can take any non-negative value so this fits.

And, the gamma family is pretty flexible, one can see a wide range of gamma shapes.

NEED TO GET THE GAMMA PLOT HERE

The probability density function for the gamma is indexed by shape k and scale θ, denoted as $\text{Gamma}(k, \theta)$ with $k, \theta > 0$. The mathematical form of the distribution is

$$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta},$$

where

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx.$$

Table 2.1: Before and after seeing the data

	lambda	Standard Deviation
Before	0.75	1.000
After	0.67	0.047

$\Gamma(z)$, the gamma function, is simply a constant that ensures the area under curve between 0 and 1 sums to 1, just like in the beta probability distribution case of Equation (2.1). A special case is that $\Gamma(n) = (n-1)!$ when n is a positive integer.

However, some books parameterize the gamma distribution in a slightly different way with shape $\alpha = k$ and rate (inverse scale) $\beta = 1/\theta$:

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

For this example, we use the k-θ parameterization, but you should always check which parameterization is being used. For example, R uses the α-β parameterization by default.
In the the later material we find that using the rate parameterization is more convenient.

** ANY WAY TO MAKE THE MATERIAL MORE IN SYNC AS LABS LATER SECTIONS ALL USE THE RATE PARAMETERIZATION **

For our parameterization, the mean of Gamma(k, θ) is $k\theta$, and the variance is $k\theta^2$. We can get the general's prior as below:

$$\begin{aligned} \text{Mean} &= k\theta = 0.75 \\ \text{Standard deviation} &= \theta\sqrt{k} = 1 \end{aligned}$$

Hence

$$k = \frac{9}{16} \text{ and } \theta = \frac{4}{3}$$

For the gamma Poisson conjugate family, suppose we observed data $x_1, x_2, \cdots, x_n$ that follow a Poisson distribution.Then similar to the previous section, we would recognize the kernel of the gamma when using the gamma-Poisson family. The posterior Gamma(k^*, θ^*) has parameters

$$k^* = k + \sum_{i=1}^{n} x_i \text{ and } \theta^* = \frac{\theta}{(n\theta + 1)}.$$

For this dataset, $N = 15 \times 20 = 300$ observations, and the number of casualities is 200. Therefore, the general now thinks that the average number of Prussian cavalry officers who die at the hoofs of their horses follows a gamma distribution with the parameters below:

$$\begin{aligned} k^* &= k + \sum_{i=1}^{n} x_i = \frac{9}{16} + 200 = 200.5625 \\ \theta^* &= \frac{\theta}{(n\theta + 1)} = \frac{4/3}{300 \times (4/3)} = 0.0033 \end{aligned}$$

How the general has changed his mind is described in Table 2.1. After seeing the data, his uncertainty about lambda, expressed as a standard deviation, shrunk from 1 to 0.047.

In summary, we learned about the Poisson and gamma distributions; we also knew that the gamma-Poisson families are conjugate. Moreover, we learned the updating fomula, and applied it to a classical dataset.

2.2.3 The Normal-Normal Conjugate Families

There are other conjugate families, and one is the normal-normal pair. If your data come from a normal distribution with known standard deviation σ but unknown mean μ, and if your prior on the mean μ, has a normal distribution with self-elicited mean ν and self-elicited standard deviation τ, then your posterior density for the mean, after seeing a sample of size n with sample mean $\bar{x}$, is also normal. In mathematical notation, we have

$$\begin{aligned} x|\mu &\sim N(\mu, \sigma) \\ \mu &\sim N(\nu, \tau) \end{aligned}$$

As a practical matter, one often does not know sigma, the standard deviation of the normal from which the data come. In that case, you could use a more advanced conjugate family that we will describe in 3.2.1. But there are cases in which it is reasonable to treat the σ as known.

Example 2.7. An analytical chemist whose balance produces measurements that are normally distributed with mean equal to the true mass of the sample and standard deviation that has been estimated by the manufacturer balance and confirmed against calibration standards provided by the National Institute of Standards and Technology.

Note that this normal-normal assumption made by the anayltical chemist is technically wrong, but still reasonable.

1. The normal family puts some probability on all possible values between $(-\infty, +\infty)$. But the mass on the balance can **never** be negative. However, the normal prior on the unknown mass is usually so concentrated on positive values that the normal distribution is still a good approximation.
2. Even if the chemist has repeatedly calibrated her balance with standards from the National Institute of Standards and Technology, she still will not know its standard deviation precisely. However, if she has done it often and well, it is probably a sufficiently good approximation to assume that the standard deviation is known.

For the normal-normal conjugate families, assume the prior on the unknown mean follows a normal distribution, i.e. $\mu \sim N(\nu, \tau)$. We also assume that the data $x_1, x_2, \cdots, x_n$ are independent and come from a normal with standard deviation σ.

Then the posterior distribution of μ is also normal, with mean as a weighted average of the prior mean and the sample mean. We have

$$\mu|x_1, x_2, \cdots, x_n \sim N(\nu^*, \tau^*),$$

where

$$\nu^* = \frac{\nu\sigma^2 + n\bar{x}\tau^2}{\sigma^2 + n\tau^2} \text{ and } \tau^* = \sqrt{\frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}}.$$

Let's continue from Example 2.7, and suppose she wants to measure the mass of a sample of ammonium nitrate.

Her balance has a known standard deviation of 0.2 milligrams. By looking at the sample, she thinks this mass is about 10 milligrams and based on her previous experience in estimating masses, her guess has the

standard deviation of 2. So she decides that her prior for the mass of the sample is a normal distribution with mean, 10 milligrams, and standard deviation, 2 milligrams.

Now she collects five measurements on the sample and finds that the average of those is 10.5. By conjugacy of the normal-normal family, our posterior belief about the mass of the sample has the normal distribution.

The new mean of that posterior normal is found by plugging into the formula:

$$\begin{aligned}
\mu &\sim N(\nu = 10, \tau = 2) \\
\nu^* &= \frac{\nu\sigma^2 + n\bar{x}\tau^2}{\sigma^2 + n\tau^2} = \frac{10 \times (0.2)^2 + 5 \times 10.5 \times 2^2}{(0.2)^2 + 5 \times 2^2} = 10.499 \\
\tau^* &= \sqrt{\frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}} = \sqrt{(0.2)^2 \times 2^2}(0.2)^2 + 5 \times 2^2 = 0.089.
\end{aligned}$$

Before seeing the data, the Bayesian analytical chemist thinks the ammonium nitrate has mass 10 mg and uncertainty (standard deviation) 2 mg. After seeing the data, she thinks the mass is 10.499 mg and standard deviation 0.089 mg. Her posterior mean has shifted quite a bit and her uncertainty has dropped by a lot. That's exactly what an analytical chemist wants.

This is the last of the three examples of conjugate families. There are many more, but they do not suffice for every situation one might have.

We learned several things in this lecture. First, we learned the new pair of conjugate families and the relevant updating formula. Also, we worked a realistic example problem that can arise in practical situations.

2.3 Credible Intervals and Predictive Inference

Missing: JAGS plot for 2.3.1

2.3.1 Non-Conjugate Priors

In many applications, a Bayesian may not be able to use a conjugate prior. Sometimes she may want to use a reference prior, which injects the minimum amount of personal belief into the analysis. But most often, a Bayesian will have a personal belief about the problem that cannot be expressed in terms of a convenient conjugate prior.

For example, we shall reconsider the RU-486 case from earlier in which four children were born to standard therapy mothers. But no children were born to RU-486 mothers. This time, the Bayesian believes that the probability p of an RU-486 baby is uniformly distributed between 0 and one-half, but has a point mass of 0.5 at one-half. That is, she believes there's a 50% chance that there is no difference between standard therapy and RU-486. But if there is a difference, she thinks that RU-486 is better, but she is completely unsure about how much better it would be.

In mathematical notation, the probability density function of p is

$$f_p(x) = \begin{cases} 1 & \text{for} \quad 0 \le x < 0.5 \\ 1 & \text{for} \quad x = 0.5 \\ 0 & \text{for} \quad x < 0 \text{ or } x > 0.5 \end{cases}$$

We can check that the area under the density curve, plus the amount of the point mass, equals 1.

The cumulative distribution function of p is

$$F_p(p \leq x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ x & \text{for} \quad 0 \leq x < 0.5 \\ 1 & \text{for} \quad x \geq 0.5 \end{cases}$$

Why would this be a reasonable prior for an analyst to self-elicit? One reason is that in clinical trials, there's actually quite a lot of preceding research on the efficacy of the drug. This research might be based on animal studies or knowledge of the chemical activity of the molecule. So the Bayesian might feel sure that there is no possibility that RU-486 is worse than the standard treatment. And her interest is on whether the therapies are equivalent and if not, how much better RU-486 is than the standard therapy.

As previously mentioned, there is no way to compute the posterior distribution for p in a simple or even a complex mathematical form. And that is why Bayesian inference languished for so many decades until computational power enabled numerical solutions. But now we have such tools, and one of them is called **JAGS (Just Another Gibbs Sampler)**.

If we apply JAGS to the RU-486 data with this non-conjugate prior, we can find the posterior distribution. At a high level, this program is defining the binomial probability, that is the likelihood of seeing 0 RU-486 children, which is binomial. And then it defines the prior by using a few tricks to draw from either a uniform on the interval from 0 to one-half, or else draw from the point mass at one-half. Then it calls the JAGS model function, and draws 5,000 times from the posterior and creates a histogram of the results.

NEED TO GET THE JAGS PLOT HERE

That histogram is lightly smooth to generate the posterior density you see. There is still a point mass of probability at 0.5, but now it has less weight than before. Also, note how the data have changed the posterior away from the prior. The analyst sees a lot of probability under the curve near 0.2, but responds to the fact that no children were born to RU-486 mothers.

This section is mostly a look-ahead to future material. We have seen that a Bayesian might reasonably employ a non-conjugate prior in a practical application. But then she will need to employ some kind of numerical computation to approximate the posterior distribution. Additionally, we have used a computational tool, JAGS, to approximate the posterior for p, and identified its three important elements, the probability of the data given p, that is the likelihood, and the prior, and the call to the Gibbs sampler.

2.3.2 Credible Intervals

In this section, we introduce credible intervals, the Bayesian alternative to confidence intervals. Let's start with the confidence intervals, which are the frequentist way to express uncertainty about an estimate of a population mean, a population proportion or some other parameter.

A confidence interval has the form of an upper and lower bound.

$$L, U = \text{pe} \pm \text{se} \times \text{cv}$$

- L = lower, U = upper
- pe = point estimate, se = standard error, cv = critical value

Most importantly, the interpretation of a 95% confidence interval on the mean is that **"95% of similarly constructed intervals will contain the true mean"**, not "the probability that true mean lies between L and U is 0.95".

The reason for this frequentist wording is that a frequentist may not express his uncertainty as a probability. The true mean is either within the interval or not, so the probability is zero or one. The problem is that the frequentist does not know which is the case.

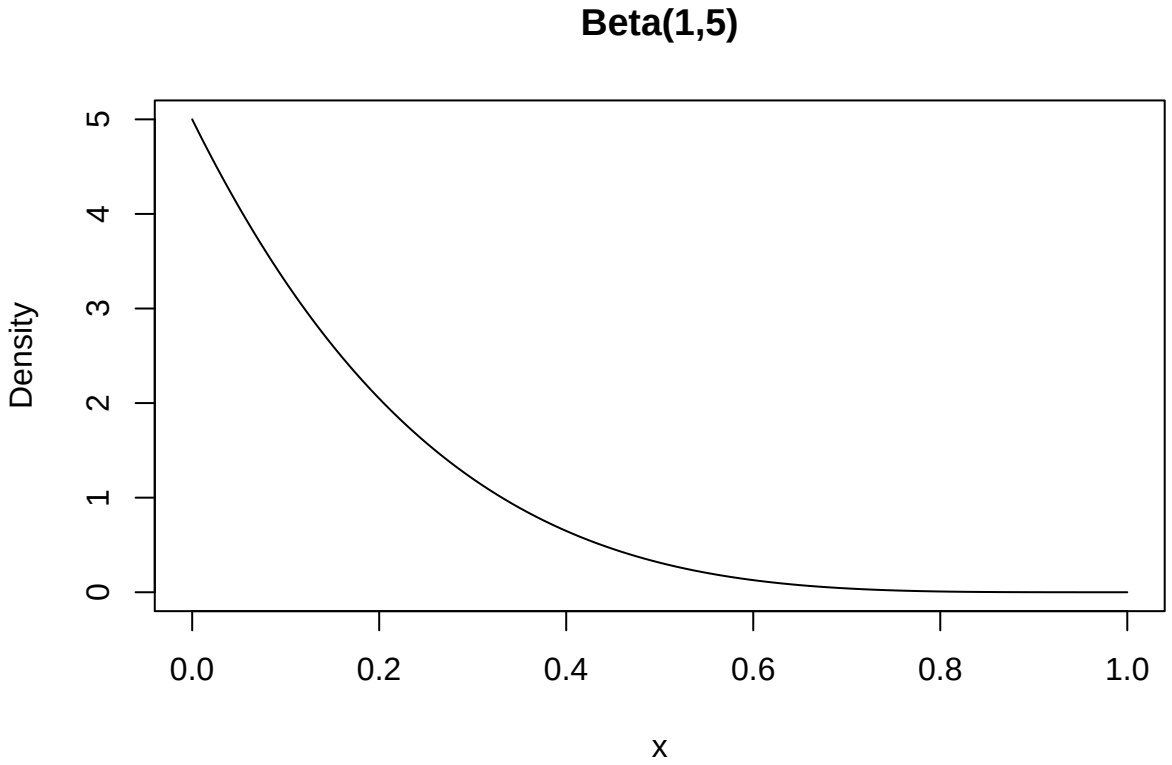


Figure 2.3: RU-486 Posterior

On the other hand, Bayesians have no such qualms. It is fine for us to say that **"the probability that the true mean is contained within a given interval is 0.95"**. To distinguish our intervals from confidence intervals, we call them **credible intervals**.

Recall the RU-486 example. When the analyst used the beta-binomial family, she took the prior as $p \sim \text{beta}(1,1)$, the uniform distribution, where p is the probability of a child having a mother who received RU-486.

After we observed four children born to mothers who received conventional therapy, her posterior is $p|x \sim \text{beta}(1,5)$. In Figure 2.3, the posterior probability density for $\text{beta}(1,5)$ puts a lot of probability near zero and very little probability near one.

For the Bayesian, her 95% credible interval is just any L and U such that the posterior probability that $L < p < U$ is 0.95. The shortest such interval is obviously preferable.

To find this interval, the Bayesian looks at the area under the $\text{beta}(1,5)$ distribution, that lies to the left of a value x.

The density of the $\text{beta}(1,5)$ is

$$f(p) = 5(1-p)^4 \text{ for } 0 \leq p \leq 1,$$

and the area under the density between 0 and x is

$$F(x) = 1 - (1-x)^5 \text{ for } 0 \leq p \leq 1.$$

The Bayesian can use this to find L, U with area 0.95 under the density curve between them, i.e. $F(U)F(L) = 0.95$. Note that the Bayesian credible interval is asymmetric, unlike the symmetric confidence intervals that frequentists often obtain. It turns out that $L = 0$ and $U = 0.45$ is the shortest interval with probability 0.95 of containing p.

What have we done? We have seen the difference in interpretations between the frequentist confidence interval and the Bayesian credible interval. Also, we have seen the general form of a credible interval. Finally, we have done a practical example constructing a 95% credible interval for the RU-486 data set.

2.3.3 Predictive Inference

Predictive inference arises when the goal is not to find a posterior distribution over some parameter, but rather to find a posterior distribution over some random variable depends on the parameter.

Specifically, we want to make an inference on a random variable X with probability densifity function $f(x|\theta)$, where you have some personal probability distribution $\pi(\theta)$ for the θ.

To solve this, one needs to integrate:

$$P(X \leq x) = \int_{-\infty}^{\infty} P(X \leq x|\theta)\pi(\theta)d\theta$$

The equation gives us the weighted average of the probabilities for X, where the weights correspond to the personal probability on θ. But we won't do an integral; instead, we will illustrate the thinking with a trivial example.

Example 2.8. Suppose you have two coins. One coin has probability 0.7 of coming up heads, and the other has probability 0.4 of coming up heads. You are playing a gambling game with a friend, and you draw one of those two coins at random from a bag.

Before you start the game, your prior belief is that the probability of choosing the 0.7 coin is 0.5. This is reasonable, because both coins were equally likely to be drawn. In this game, you win if the coin comes up heads.

Suppose the game starts, you have tossed twice, and have obtained two heads. Then what is your new belief about p, the probability that you are using the 0.7 coin?

This is just a simple application of the discrete form of Bayes' rule.

- Prior: $p = 0.5$
- Posterior:

$$p^* = \frac{P(2 \text{ heads}|0.7) \times 0.5}{P(2 \text{ heads}|0.7) \times 0.5 + P(2 \text{ heads}|0.4) \times 0.5} = 0.754.$$

However, this does not answer the important question – What is the predictive probability that the next toss will come up heads? This is of interest because you are gambling on getting heads.

Fortunately, the predictive probablity of getting heads is not difficult to calculate:

- p^* of 0.7 coin $= 0.754$
- p^* of 0.4 coin $= 1 0.754 = 0.246$
- $P(\text{heads}) = P(\text{heads}|0.7) \times 0.754 + P(\text{heads}|0.4) \times 0.246 = 0.626$

Therefore, the predictive probability that the next toss will come up heads is 0.626.

Note that most realistic predictive inference problems are more complicated and require one to use integrals. For example, one might want to know the chance that a fifth child born in the RU-486 clinical trial will have a mother who received RU-486. Or you might want to know the probability that your stock broker's next recommendation will be profitable.

We have learned three things in this section. First, often the real goal is **a prediction about the value of a future random variable**, rather than making an estimate of a parameter. Second, these are deep waters, and often one needs to integrate. Finally, in certain simple cases where the parameter can only take discrete values, one can find a solution without integration. In our example, the parameter could only take two values to indicate which of the two coins was being used.

Chapter 3

Introduction to Losses and Decision-making

3.1 Losses and Decision Making

3.1.1 Loss Functions

3.1.2 Working with Loss Functions

3.1.3 Minimizing Expectated Loss for Hypothesis Testing

3.1.4 Posterior Probabilities of Hypotheses and Bayes Factors

3.2 Inference and Decision-Making with Multiple Parameters

This section is focused on the extending the Normal-Normal conjugate family introduced in 2.2.3 to the problem of inference in a Normal population with an unknown mean and variance. We will introduce the Normal-Gamma conjugate family for inference about the unknown mean and variance and will present Monte Carlo simulation for inference about functions of the parameters as well as sampling from predictive distributions, which can assist with prior elucidation. For situations when limited prior information is available, we discuss a limiting case of the Normal-Gamma conjugate family, leading to priors that can be used for a reference analysis. Finally, we will show how to create a more flexible and robust prior distribution by using mixtures of the Normal-Gamma conjugate prior. For inference in this case we will introduce Markov Chain Monte Carlo, a powerful simulation method for Bayesian inference.

It is assumed that the readers have mastered the concepts of one-parameter Normal-Normal conjugate priors. Calculus is not required for this section; however, for those who are comfortable with calculus and would like to go deeper, we shall present starred sections with more details on the derivations.

3.2.1 Inference for a Normal Mean with Unknown Variance

In 2.2.3 we described the normal-normal conjugate family for inference about an unknown mean μ with a known standard deviation σ when the data were assumed to be a random sample from a normal population. In this section we will introduce the normal-gamma conjugate family for the common situation when σ is

unknown. As both μ and σ^2 unknown, we will need to specify a **joint** prior distribution to describe our prior uncertainty about them.

Sampling Model

Recall that a conjugate pair is a sampling model for the data and prior distribution for the unknown parameters such that the posterior distribution is in the same family of distributions as the prior distribution. We will assume that the data are a random sample of size n from a normal population with mean μ and variance σ^2; the following is a mathematical shorthand to represent this distribution assumption

$$Y_1, \ldots Y_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$$

where the 'iid' above the distributed as symbol '$\sim$' indicates that each of the observations are **i**ndependent of the others (given μ and σ^2) and are **i**dentically **d**istributed.

Conjugate prior Back in 2.2.3, we found that with normal data, the conjugate prior for μ when the standard deviation σ was known was a normal distribution. We will build on this to specify a conditional prior distribution for μ as

$$\mu \mid \sigma^2 \sim \mathsf{N}(m_0, \sigma^2/n_0) \tag{3.1}$$

with hyper-parameters m_0, the prior mean for μ, and σ^2/n_0 the prior variance. While previously the variance was a known constant τ^2, replacing τ^2 with a multiple of σ^2 is needed for representing the joint conjugate prior for the mean and variance. Because σ has the same units as the data, the hyper-parameter n_0 is unitless, but is used to express our prior precision about μ with larger values of n_0 indicating more precision and smaller values less precision. We will see later how the hyper-parameter n_0 may be interpreted as a prior sample size.

As σ^2 is unknown, a Bayesian would use a prior distribution to describe the uncertainty about the variance before seeing data. Since the variance is non-negative, continuous, and with no upper limit, a gamma distribution is a candidate prior for the variance, based on the distributions that we have seen so far. However, that choice does not lead to a posterior distribution in the same family or that is recognizable as any common distribution. It turns out that the the inverse of the variance, which is known as the precision, has a conjugate gamma prior distribution. Letting $\phi = 1/\sigma^2$ denote the precision or inverse variance, the conjugate prior for ϕ,

$$\phi \sim \mathsf{Gamma}\left(\frac{v_0}{2}, \frac{v_0 s_0^2}{2}\right) \tag{3.2}$$

is a gamma distribution with hyper-parameters v_0, prior degrees of freedom, and s_0^2 a prior variance or guess for σ^2. Equivalently we may say that the inverse of the variance has a

$$1/\sigma^2 \sim \mathsf{Gamma}(v_0/2, s_0^2 v_0/2)$$

gamma distribution to avoid using a new symbol. Together the Normal conditional distribution for μ given σ^2 in (3.1) and the marginal Gamma distribution for ϕ in (3.2) lead to a joint distribution for the pair (μ, ϕ) that we will call the Normal-Gamma family of distributions:

$$(\mu, \phi) \sim \mathsf{NormalGamma}(m_0, n_0, s_0^2, v_0) \tag{3.3}$$

with the four hyper-parameters m_0, n_0, s_0^2, and v_0.

Posterior Distribution

As a conjugate family, the posterior distribution of the pair of parameters (μ, ϕ) is in the same family as the prior distribution when the sample data arise from a normal distribution, that is the posterior is also Normal-Gamma

$$(\mu, \phi) \mid \text{data} \sim \mathsf{NormalGamma}(m_n, n_n, s_n^2, v_n) \tag{3.4}$$

where the subscript n on the hyper-parameters indicates the updated values after seeing the n observations. One attraction to conjugate families is there are relatively simple updating rules for obtaining the new hyper-parameters:

$$\begin{aligned} m_n &= \frac{n\bar{Y} + n_0 m_0}{n + n_0} \\ n_n &= n_0 + n \\ v_n &= v_0 + n \\ s_n^2 &= \frac{1}{v_n}\left[s_0^2 v_0 + s^2(n-1) + \frac{n_0 n}{n_n}(\bar{Y} - m_0)^2\right]. \end{aligned}$$

The updated hyper-parameter m_n in the posterior distribution of μ is the posterior mean, which is a weighted average of the sample mean $\bar{Y}$ and prior mean m_0 with weights $n/(n + n_0$ and $n_0/(n + n_0)$ respectively and does not depend on σ^2. The posterior sample size n_n is the sum of the prior sample size n_n and the sample size n, representing the combined precision of the estimate for μ. The posterior degrees of freedom v_n are also increased by adding the sample size n to the prior degrees of freedom v_0. Finally, the posterior variance hyper-parameter s_n^2 combines three sources of information about σ in terms of sums of squared deviations. **FILL IN MORE DETAILS** The first term in the square brackets is the sample variance times the sample degrees of freedom which is the sample sum of squares. The second term represents the prior sum of squares, while the third term is based on the squared difference of the sample mean and prior mean. We then divide by the posterior degrees of freedom to get the new hyper-parameter.

The joint Normal-Gamma distribution for the pair μ and ϕ,

$$(\mu, \phi) \mid \text{data} \sim \mathsf{NormalGamma}(m_n, n_n, s_n^2, v_n)$$

is equivalent to a **hierarchical model** specified in two stages with μ given σ having a conditional normal distribution

$$\mu \mid \text{data}, \sigma^2 \sim \mathsf{N}(m_n, \sigma^2/n_n)$$

and the inverse variance marginally

$$1/\sigma^2 \mid \text{data} \sim \mathsf{Gamma}(v_n/2, s_n^2 v_n/2)$$

having a gamma distribution. We will see in the next section how this representation is convenient for generating samples from the posterior distribution.

Marginal Distribution for μ

We are generally interested in inference about μ unconditionally as σ^2 is unknown. This marginal inference requires the unconditional or marginal distribution of μ that 'averages' over the uncertainty in σ. For continuous variables like σ, this averaging is performed by integration leading to the following result:

μ given the data is a

$$\mu \mid \text{data} \sim \mathsf{t}(v_n, m_n, s_n^2/n_n)$$

with density

$$p(\mu) = \frac{\Gamma\left(\frac{v_n+1}{2}\right)}{\sqrt{\pi v_n}\frac{s_n}{\sqrt{n_n}}\Gamma\left(\frac{v_n}{2}\right)}\left(1 + \frac{1}{v_n}\frac{(\mu - m_n)^2}{s_n^2/n_n}\right)^{-\frac{v_n+1}{2}}$$

with the degrees of freedom v_n, a location parameter m_n and squared scale parameter that is the posterior variance parameter divided by the posterior sample size. A standard Student t random variable can be obtained by taking μ and subtracting the location m_n and dividing by the scale $s_n/\sqrt{n}$:

$$\frac{\mu - m_n}{s_n/\sqrt{n_n}} \equiv t \sim \mathsf{t}(v_n, 0, 1)$$

with degrees of freedom v_n, location 0 and scale 1 in the expression for the density in (3.2.1). This latter representation allows us to use standard statistical functions for posterior inference such as finding credible intervals.

The Student t distribution is similar to the normal distribution as it is symmetric and bell shaped, however, the **tails** of the distribution are fatter or heavier than the normal distribution. The parameters m_n and s_n^2 play similar roles in determining the center and spread of the distribution, as in the Normal distribution, however, as Student t distributions with degrees of freedom less than 3 do not have a mean or variance, the parameter m_n is called the location or center of the distribution and the $s_n/\sqrt{n}$ is the scale.

Example

Let's look at an example based on a sample of total trihalomethanes or TTHM in tap water from a city in NC. The data can be loaded from the `statsr` package

```
library(statsr)
data(tapwater)
```

Using prior information about TTHM from the city, we will use a Normal-Gamma prior distribution, $\mathsf{NormalGamma}(35, 25, 156.25, 24)$ with a prior mean of 35 parts per billion, a prior sample size of 25, an estimate of the variance of 156.25 with degrees of freedom 24. In section 3.2.3, we will describe how we arrived at these values.

Using the summaries of the data, $\bar{Y} = 55.5$, variance $s^2 = 540.7$ and sample size of $n = 28$ with the prior hyper-parameters from above, the posterior hyper-parameters are updated as follows:

$$\begin{aligned}
n_n &= 25 + 28 = 53 \\
m_n &= \frac{28 \times 55.5 + 25 \times 35}{53} = 45.8 \\
v_n &= 24 + 28 = 52 \\
s_n^2 &= \frac{(n-1)s^2 + v_0 s_0^2 + n_0 n (m_0 - \bar{Y})^2 / n_n}{v_n} \\
&= \frac{1}{52}\left[27 \times 540.7 + 24 \times 156.25 + \frac{25 \times 28}{53} \times (35 - 55.5)^2\right] = 459.9
\end{aligned}$$

in the conjugate $\mathsf{NormalGamma}(45.8, 53, 459.9, 52)$ posterior distribution that now summarizes our uncertainty about μ and ϕ (σ^2) after seeing the data.

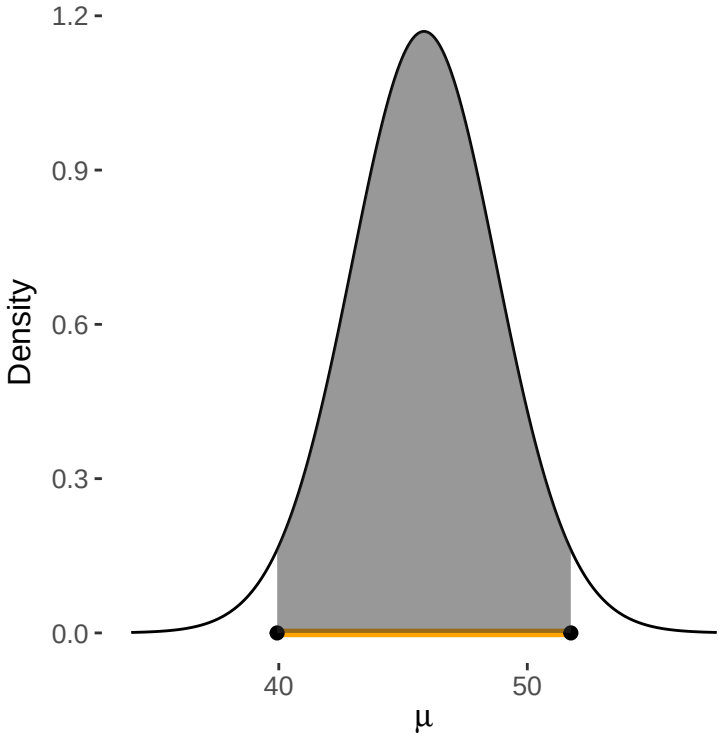
We can obtain the updated hyper-parameters in `R` using the following code in `R`

```
# prior hyperparameters
m_0 = 35; n_0 = 25;  s2_0 = 156.25; v_0 = n_0 - 1
# sample summaries
Y = tapwater$tthm
ybar = mean(Y)
s2 = var(Y)
n = length(Y)
# posterior hyperparamters
n_n = n_0 + n
```

```
m_n = (n*ybar + n_0*m_0)/n_n
v_n = v_0 + n
s2_n = ((n-1)*s2 + v_0*s2_0 + n_0*n*(m_0 - ybar)^2/n_n)/v_n
```

Credible intervals for μ

To find a credible interval for the mean μ, we use the Student t distribution. Since the distribution of μ is unimodal and symmetric, the shortest 95 percent credible interval or the **Highest Posterior Density** interval, HPD for short,



is the orange interval given by the Lower endpoint L and upper endpoint U where the probability that mu is in the interval (L, U) is the shaded area which is equal to zero point nine five.

using the standardized t distribution and some algebra, these values are

$$L = m_n + t_{0.025}\sqrt{s_n^2/n_n}$$
$$U = m_n + t_{0.975}\sqrt{s_n^2/n_n}$$

or the posterior mean (our point estimate) plus quantiles of the standard t distribution times the scale. Because of the symmetry in the Student t distribution, the credible interval is $m_n \pm t_{0.975}\sqrt{s_n^2/n_n}$, which should look familiar to expressions for confidence intervals.

Using the following code in R the 95% credible interval for the tap water data is

```
m_n + qt(c(0.025, 0.975), v_n)*sqrt(s2_n/n_n)
```

```
## [1] 39.93192 51.75374
```

Based on the updated posterior, we find that there is a 95 chance that the mean TTHM concentration is between 39.9 parts per billion and 51.7 parts per billion.

Summary The Normal-Gamma conjugate prior for inference about an unknown mean and variance for samples from a normal distribution allows simple expressions for updating prior beliefs given the data. The joint Normal-Gamma distribution leads to the Student t distribution for inference about μ when σ is

unknown. The Student t distribution can be used to provide credible intervals for μ using R or other software that provides quantiles of a standard t distribution.

For the energetic learner who is comfortable with calculus, the following optional material provides more details on how the posterior distributions were obtained and other results in this section.

For those that are ready to move on, we will introduce Monte Carlo sampling in the next section; Monte Carlo Sampling is a simulation method that will allow us to approximate distributions of transformations of the parameters without using calculus or change of variables, as well as aid exploratory data analysis of the prior or posterior distribution.

Details of Results (optional reading)

TBA

3.2.2 Monte Carlo Inference

3.2.3 Predictive Distributions

3.2.4 Reference Priors

3.2.5 Mixtures of Conjugate Priors

3.2.6 MCMC

3.3 Hypothesis Testing with Normal Populations

3.3.1 Bayes Factors for Testing a Normal Mean: variance known

3.3.2 Bayes Factors for Testing a Normal Mean: unknown variance

3.3.3 Testing Normal Means: paired data

3.3.4 Testing Normal Means: independent groups

3.3.5 Inference after Testing

3.4 Excercises

Bibliography