Bayesian Statistics

A Companion to the Statistics with R Coursera Course

Christine Chai
Merlise Clyde
Lizzy Huang
Colin Rundel

Last built on 2017-12-19

Contents

Welcome

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.

Authors are listed in alphabetical order.

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

This section introduces how the Bayes' rule is applied to calculating conditional probability, and several real-life examples are demonstrated. Finally, we compare the Bayesian and frequentist definition of probability.

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

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

$P(\text{using an online dating site}) =$

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

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,

$P(\text{using an online dating site} \mid \text{in age group 30-49})$

$$= \frac{\text{Number in age group 30-49 that indicated they used an online dating site}}{\text{Total number in age group 30-49}}$$

$$= \frac{\frac{\text{Number in age group 30-49 that indicated they used an online dating site}}{\text{Total number of people in the poll}}}{\frac{\text{Total number in age group 30-49}}{\text{Total number of people in the poll}}}$$

$$= \frac{P(\text{using an online dating site \& falling in age group 30-49})}{P(\text{Falling in age group 30-49})}.$$

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:

$P(\text{using an online dating site} \mid \text{in age group 18-29})$

$$= \frac{P(\text{using an online dating site \& falling in age group 18-29})}{P(\text{Falling in age group 18-29})}$$

$$= \frac{\frac{\text{Number in age group 18-29 that indicated they used an online dating site}}{\text{Total number of people in the poll}}}{\frac{\text{Total number in age group 18-29}}{\text{Total number of people in the poll}}}$$

$$= \frac{\text{Number in age group 18-29 that indicated they used an online dating site}}{\text{Total number in age group 18-29}} = \frac{60}{315} \approx 19\%.$$

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

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

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

$$P(\text{Person tested has HIV} \,\&\, \text{ELISA is positive})$$
$$= P(\text{Person tested has HIV})P(\text{ELISA is positive} \mid \text{Person tested has HIV})$$
$$= 0.00148 \cdot 0.93 = 0.0013764. \tag{1.2}$$

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

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

where we used (1.2) and

$P(\text{Person tested has no HIV} \& \text{ELISA is positive})$
$$= P(\text{Person tested has no HIV})P(\text{ELISA is positive} \mid \text{Person tested has no HIV})$$
$$= (1 - P(\text{Person tested has HIV})) \cdot (1 - P(\text{ELISA is negative} \mid \text{Person tested has no HIV}))$$
$$= (1 - 0.00148) \cdot (1 - 0.99) = 0.0099852.$$

Putting this all together and inserting into (1.1.2) reveals

$$P(\text{Person tested has HIV} \mid \text{ELISA is positive}) = \frac{0.0013764}{0.0113616} \approx 0.12. \tag{1.3}$$

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 (1.3) 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 (1.3).

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(Person tested has HIV | 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(Person tested has HIV | 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(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(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

$$
\begin{aligned}
&P(\text{Person tested has HIV} \mid \text{Second ELISA is also positive}) \\
&= \frac{P(\text{Person tested has HIV})P(\text{Second ELISA is positive} \mid \text{Person tested has HIV})}{P(\text{Second ELISA is also positive})} \\
&= \frac{0.12 \cdot 0.93}{\begin{aligned}&P(\text{Person tested has HIV})P(\text{Second ELISA is positive} \mid \text{Has HIV}) \\ &+ P(\text{Person tested has no HIV})P(\text{Second ELISA is positive} \mid \text{Has no HIV})\end{aligned}} \\
&= \frac{0.1116}{0.12 \cdot 0.93 + (1 - 0.12) \cdot (1 - 0.99)} \approx 0.93.
\end{aligned}
\tag{1.4}
$$

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(Person tested has HIV) = 0.93. Analogous to (1.4), the answer follows as

$P(\text{Person tested has HIV} \mid \text{Third ELISA is also positive})$

$$
\begin{aligned}
&= \frac{P(\text{Person tested has HIV})P(\text{Third ELISA is positive} \mid \text{Person tested has HIV})}{P(\text{Third ELISA is also positive})} \\
&= \frac{0.93 \cdot 0.93}{\begin{array}{l} P(\text{Person tested has HIV})P(\text{Third ELISA is positive} \mid \text{Has HIV}) \\ +P(\text{Person tested has no HIV})P(\text{Third ELISA is positive} \mid \text{Has no HIV}) \end{array}} \\
&= \frac{0.8649}{0.93 \cdot 0.93 + (1 - 0.93) \cdot (1 - 0.99)} \approx 0.999.
\end{aligned}
\tag{1.5}
$$

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

$$
P(E) = \lim_{n \to \infty} \frac{n_E}{n}.
\tag{1.6}
$$

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

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

.

```
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").

Here are the histograms of the prior, the likelihood, and the posterior probabilities:
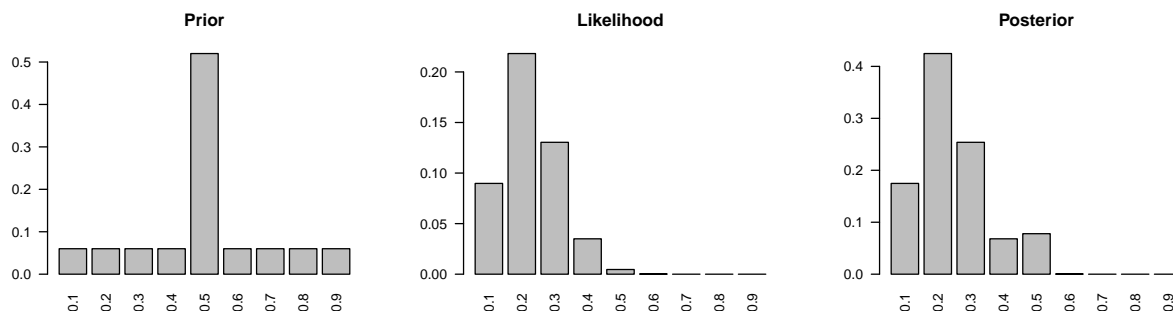
Figure 1.1: Original: sample size $n = 20$ and number of successes $k = 4$

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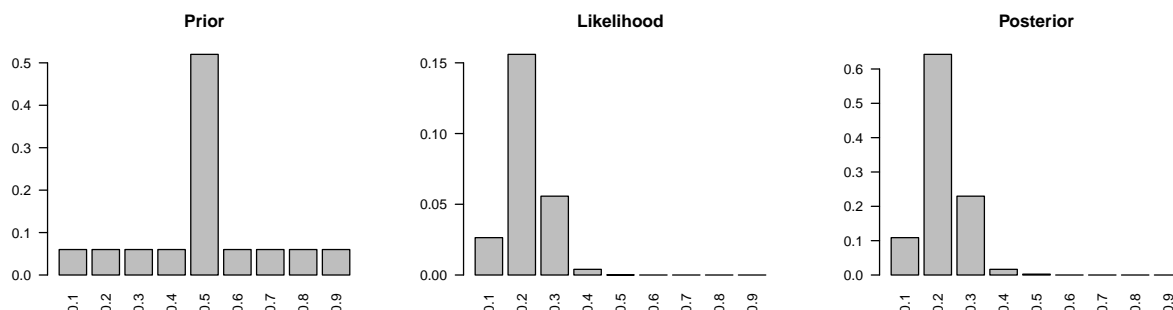
Figure 1.2: More data: sample size $n = 40$ and number of successes $k = 8$

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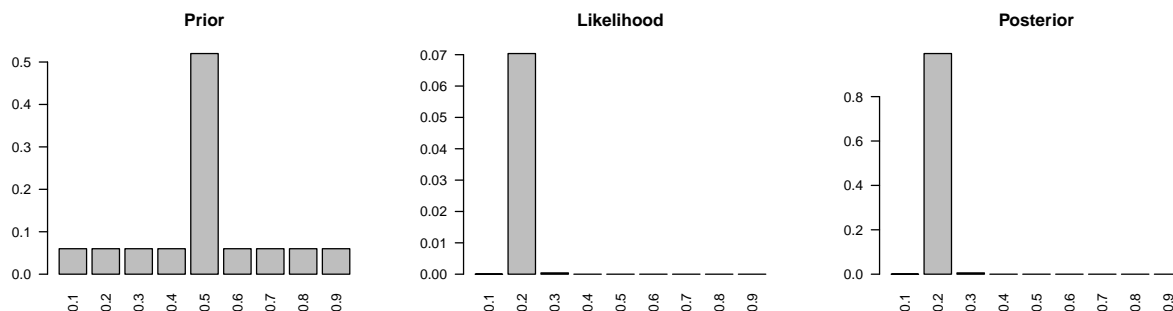
Figure 1.3: More data: sample size $n = 200$ and number of successes $k = 40$

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

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.

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

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

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.

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
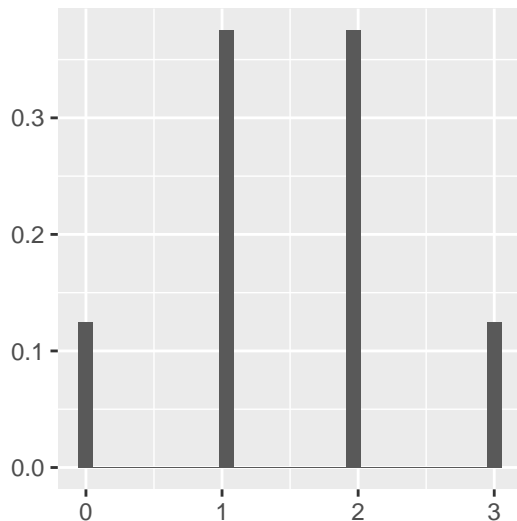
Figure 2.1: Histogram of binomial random variable

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" (71 inches) and 6'1" (73 inches) is the area under the pdf curve for height between those two values, as shown in the blue area of Figure 2.2.[1]

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

---

[1]Code reference: http://www.statmethods.net/advgraphs/probability.html
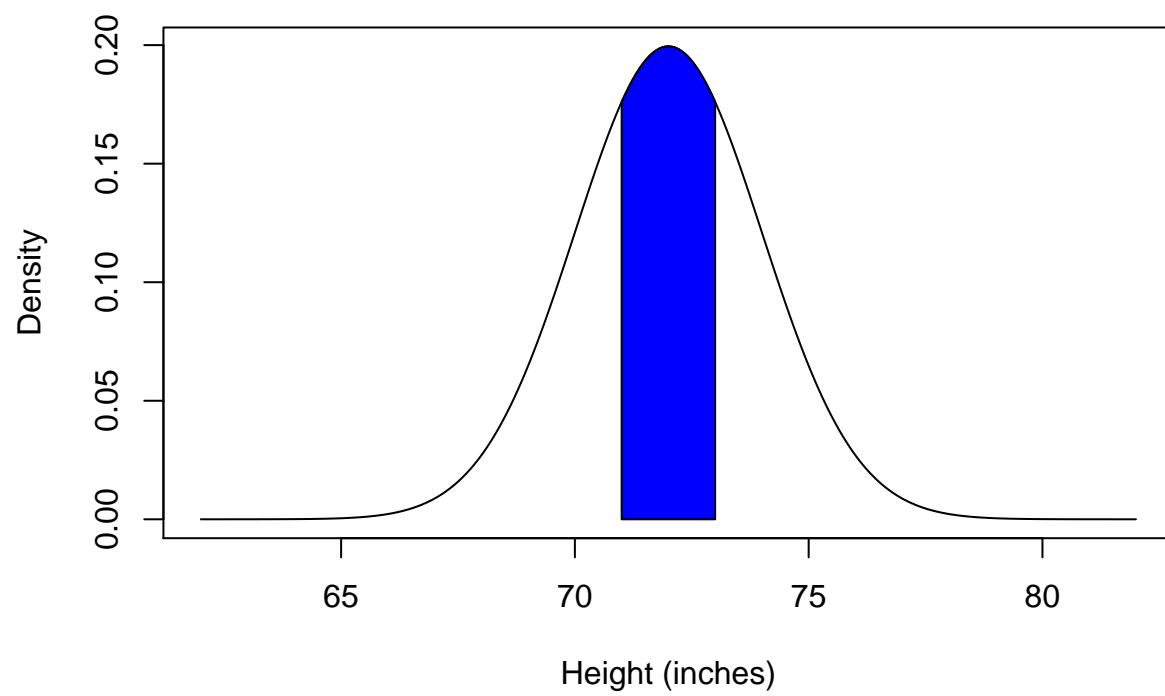
Figure 2.2: Area under curve for the probability density function

- 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 $\text{Unif}(0, 1)$.

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

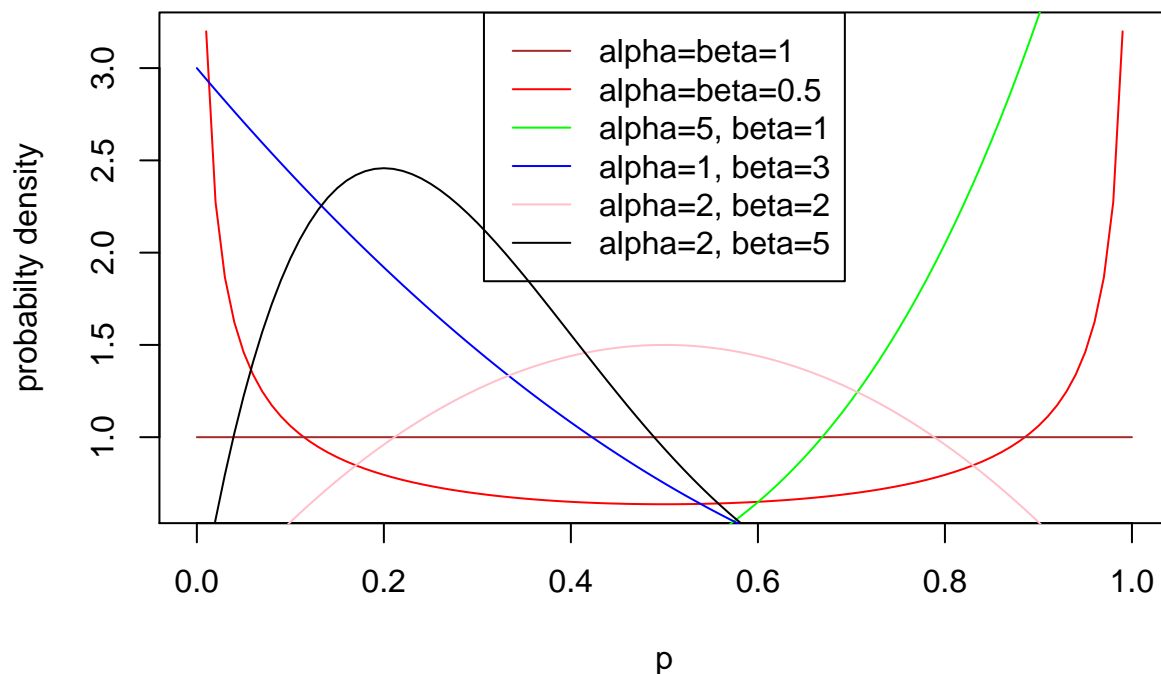$$f(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1 - p)^{\beta-1}, \tag{2.1}$$

Figure 2.3: 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 2.3 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

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

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

In this section, the three conjugate families are beta-binomial, normal-gamma, and normal-normal pairs. Each of them has its own applications in everyday life.

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 λ in the Poisson can take any non-negative value so this fits.

The gamma family is flexible, and Figure 2.4 illustrates a wide range of gamma shapes.

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
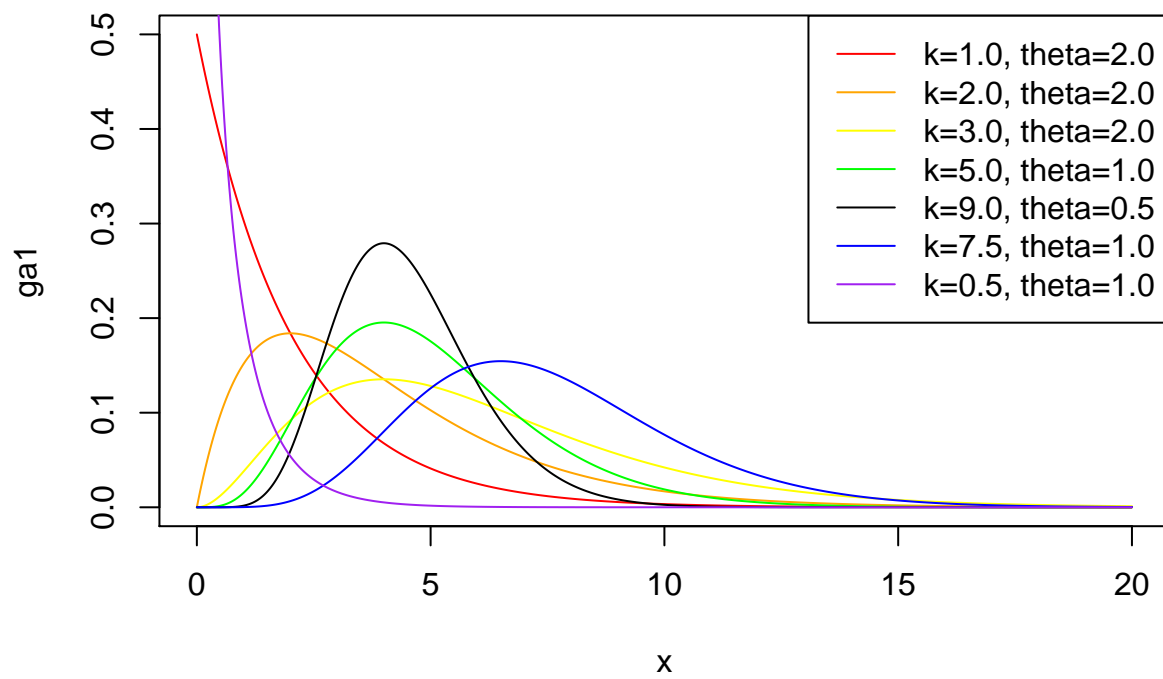
Figure 2.4: Gamma family

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

where

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

$\Gamma(z)$, 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 $\text{Gamma}(k, \theta)$ is $k\theta$, and the variance is $k\theta^2$. We can get the general's prior as below:

$$\text{Mean} = k\theta = 0.75$$
$$\text{Standard deviation} = \theta\sqrt{k} = 1$$

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 $\text{Gamma}(k^*, \theta^*)$ 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:

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

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.

Table 2.1: Before and after seeing the data

	lambda	Standard Deviation
Before	0.75	1.000
After	0.67	0.047

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

$$x|\mu \sim N(\mu, \sigma)$$
$$\mu \sim N(\nu, \tau)$$

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 4.1. 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:

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

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

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

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

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

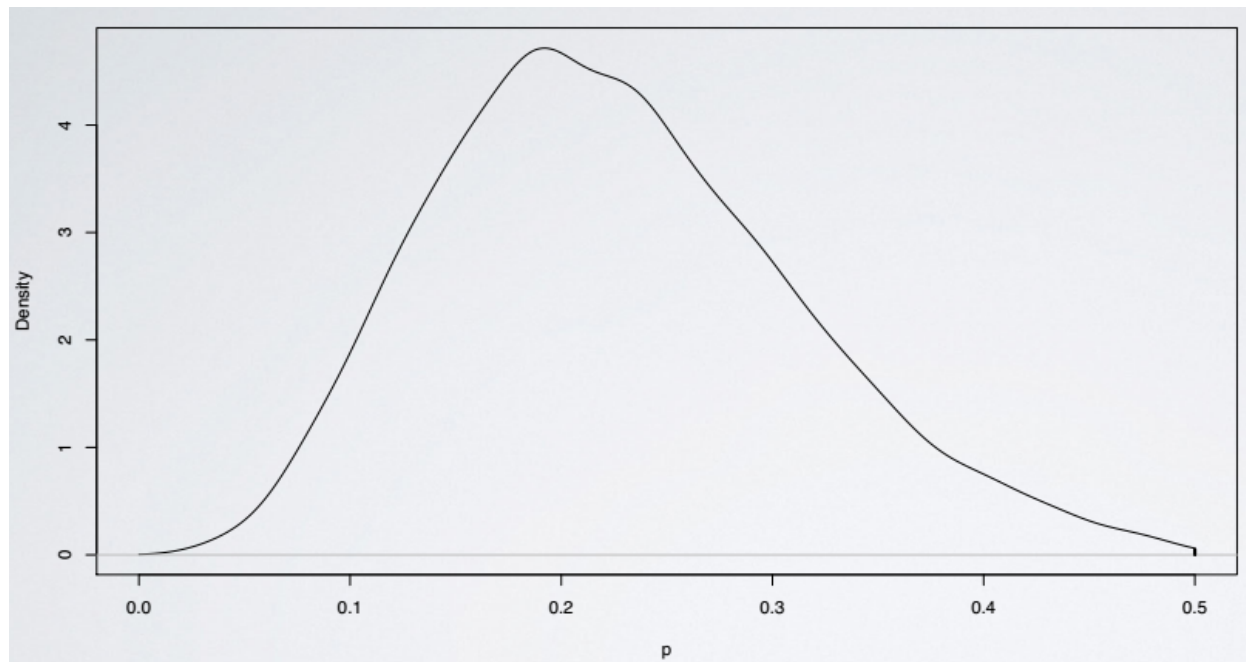The cumulative distribution function of p is

Figure 2.5: Posterior with JAGS

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

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, as in Figure 2.5. 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.

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.

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 $p \sim$ 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 $p|x \sim$ beta$(1, 5)$. In Figure 2.6, the posterior probability density for 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 beta$(1, 5)$ distribution, that lies to the left of a value x.

The density of the 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.
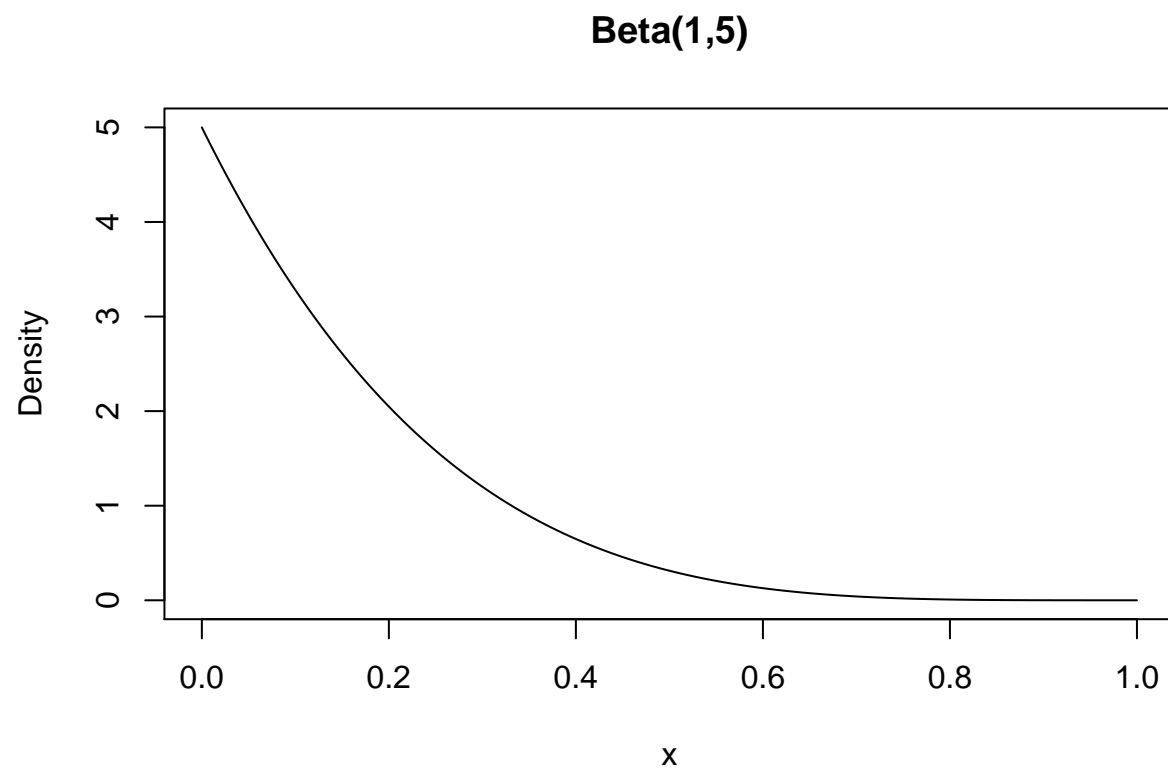
Beta(1,5)

Figure 2.6: RU-486 Posterior

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:

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

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 $= 10.754 = 0.246$
- $P(\text{heads}) = P(\text{heads}|0.7) \times 0.754 + P(\text{heads}|0.4) \times 0.246 = 0.626$

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

In the previous chapter, we learned about continuous random variables. That enabled us to study conjugate families, such as the beta binomial, the poisson gamma, and the normal normal. We also considered the difficulties of eliciting a personal prior, and of handling inference in nonconjugate cases. Finally, we introduced the credible interval and studied predictive inference.

In this new chapter, we will introduce loss functions and Bayesian decision making, minimizing expected loss for hypothesis testing, and define posterior probabilities of hypothesis and base factors. We will then outline Bayesian testing for two proportions and two means, discuss how findings from credible intervals compare to those from our hypothesis test, and finally discuss when to reject, accept, or wait.

3.1 Losses and Decision Making

To a Bayesian, the posterior distribution is the basis of any inference, since it integrates both his/her prior opinions and knowledge and the new information provided by the data. It also contains everything she believes about the distribution of the unknown parameter of interest.

However, the posterior distribution on its own is not always sufficient. Sometimes the inference we want to express is a **credible interval**, because it indicates a range of likely values for the parameter. That would be helpful if you wanted to say that you are **95% certain** the probability of an RU-486 pregnancy lies between some number L and some number U. And on other occasions, one needs to make a single number guess about the value of the parameter. For example, you might want to declare the average payoff for an insurance claim or tell a patient how much longer he/she has to live.

Therefore, the Bayesian perspective leads directly to **decision theory**. And in decision theory, one seeks to minimize one's expected loss.

3.1.1 Loss Functions

Quantifying the loss can be tricky, and Table 3.1 summarizes three different examples with three different loss functions.

If you're declaring the average payoff for an insurance claim, and if you are **linear** in how you value money, that is, twice as much money is exactly twice as good, then one can prove that the optimal one-number estimate is the **median** of the posterior distribution. But in different situations, other measures of loss may apply.

Table 3.1: Loss Functions

Loss	Best Estimate
Linear	Median
Quadratic	Mean
0/1	Mode

If you are advising a patient on his/her life expectancy, it is easy to imagine that large errors are far more problematic than small ones. And perhaps the loss increases as the **square** of how far off your single number estimate is from the truth. For example, if she's told that her average life expectancy is two years, and it is actually ten, then her estate planning will be catastrophically bad, and she will die in poverty. In the case when the loss is proportional to the **quadratic** error, one can show that the optimal one-number estimate is the **mean** of the posterior distribution.

Finally, in some cases, the penalty is 0 if you are exactly correct, but constant if you're at all wrong. This is the case with the old saying that close only counts with horseshoes and hand grenades; i.e., coming close but not succeeding is not good enough. And it would apply if you want a prize for correctly guessing the number of jelly beans in a jar. Here, of course, instead of minimizing expected losses, we want to **maximize the expected gain**. If a Bayesian is in such a situation, then his/her best one-number estimate is the **mode** of his/her posterior distribution, which is the most likely value.

There is a large literature on decision theory, and it is directly linked to risk analysis, which arises in many fields. Although it is possible for frequentists to employ a certain kind of decision theory, it is much more natural for Bayesians.

When making point estimates of unknown parameters, we should make the choices that minimize the loss. Nevertheless, the best estimate depends on the kind of loss function we are using, and we will discuss in more depth how these best estimates are determined in the next section.

3.1.2   Working with Loss Functions

Now we illustrate why certain estimates minimize certain loss functions.

Example 3.1. You work at a car dealership. Your boss wants to know how many cars the dealership will sell per month. An analyst who has worked with past data from your company provided you a distribution that shows the probability of number of cars the dealership will sell per month. In Bayesian lingo, this is called the posterior distribution. A dot plot of that posterior is shown in Figure 3.1. The mean, median and the mode of the distribution are also marked on the plot. Your boss doesn't know any Bayesian statistics though, so he/she wants you to report **a single number** for the number of cars the dealership will sell per month.

Suppose your single guess is 30, and we call this g in the following calculations. If your loss function is L_0 (i.e., a 0/1 loss), then you lose a point for each value in your posterior that differs from your guess and do not lose any points for values that exactly equal your guess. The total loss is the sum of the losses from each value in the posterior.

In mathematical terms, we define L_0 (0/1 loss) as

$$L_{0,i}(0,g) = \begin{cases} 0 & \text{if } g = x_i \\ 1 & \text{otherwise} \end{cases}$$
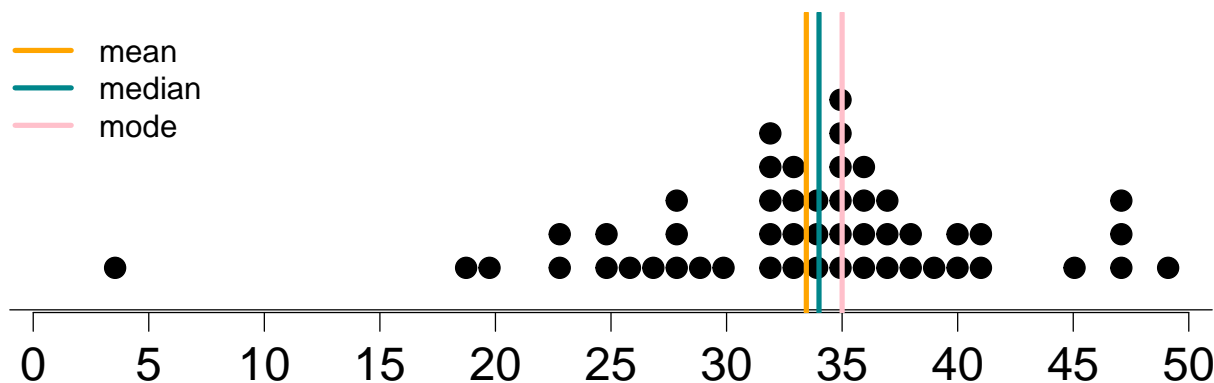
The total loss is $L_0 = \sum_i L_{0,i}(0,g)$.

Figure 3.1: Posterior

Table 3.2: L0: 0/1 loss for g = 30

i	x_i	L0: 0/1
1	4	1
2	19	1
3	20	1
	...	...
14	30	0
	...	...
50	47	1
51	49	1
	Total	50

Let's calculate what the total loss would be if your guess is 30. Table 3.2 summarizes the values in the posterior distribution sorted in descending order.

The first value is 4, which is not equal to your guess of 30, so the loss for that value is 1. The second value is 19, also not equal to your guess of 30, and the loss for that value is also 1. The third value is 20, also not equal to your guess of 30, and the loss for this value is also 1.

There is only one 30 in your posterior, and the loss for this value is 0 – since it's equal to your guess (good news!). The remaining values in the posterior are all different than 30 hence, the loss for them are all ones as well.

To find the total loss, we simply sum over these individual losses in the posterior distribution with 51 observations where only one of them equals our guess and the remainder are different. Hence, the total loss is 50.

Figure 3.2 is a visualization of the posterior distribution, along with the 0-1 loss calculated for a series of possible guesses within the range of the posterior distribution. To create this visualization of the loss function, we went through the process we described earlier for a guess of 30 for all guesses considered, and we recorded the total loss. We can see that the loss function has the lowest value when g, our guess, is equal to **the most frequent observation** in the posterior. Hence, L_0 is minimized at the **mode** of the posterior, which means that if we use the 0/1 loss, the best point estimate is the mode of the posterior.

Let's consider another loss function. If your loss function is L_1 (i.e., linear loss), then the total loss for a guess is the sum of the **absolute values** of the difference between that guess and each value in the posterior. Note that the absolute value function is required, because overestimates and underestimates do not cancel out.
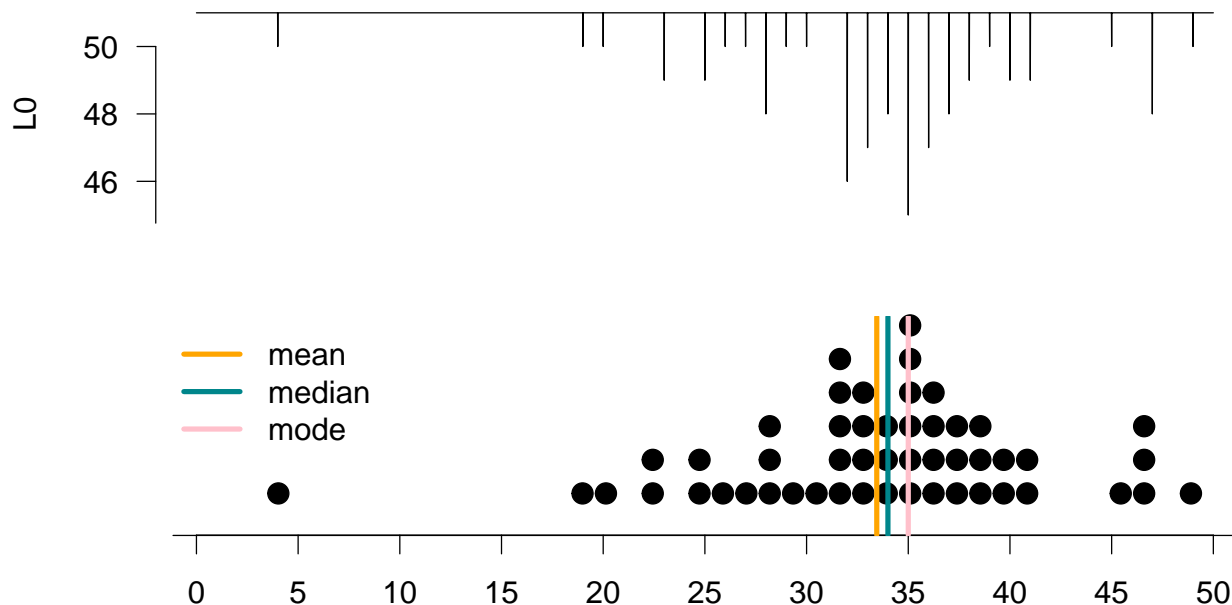
Figure 3.2: L0 is minimized at the mode of the posterior

In mathematical terms, L_1 (linear loss) is calculated as $L_1(g) = \sum_i |x_i - g|$.

We can once again calculate the total loss under L_1 if your guess is 30. Table 3.3 summarizes the values in the posterior distribution sorted in descending order.

The first value is 4, and the absolute value of the difference between 4 and 30 is 26. The second value is 19, and the absolute value of the difference between 19 and 30 is 11. The third value is 20 and the absolute value of the difference between 20 and 30 is 10.

There is only one 30 in your posterior, and the loss for this value is 0 since it is equal to your guess. The remaining value in the posterior are all different than 30 hence their losses are different than 0.

To find the total loss, we again simply sum over these individual losses, and the total is to 346.

Again, Figure 3.3 is a visualization of the posterior distribution, along with a linear loss function calculated for a series of possible guesses within the range of the posterior distribution. To create this visualization of the loss function, we went through the same process we described earlier for all of the guesses considered. This time, the function has the lowest value when g is equal to the **median** of the posterior. Hence, L_1 is minimized at the **median** of the posterior one other loss function.

If your loss function is L_2 (i.e. a squared loss), then the total loss for a guess is the sum of the squared differences between that guess and each value in the posterior.

We can once again calculate the total loss under L_2 if your guess is 30. Table 3.4 summarizes the posterior distribution again, sorted in ascending order.

The first value is 4, and the squared difference between 4 and 30 is 676. The second value is 19, and the square of the difference between 19 and 30 is 121. The third value is 20, and the square difference between 20 and 30 is 100.

There is only one 30 in your posterior, and the loss for this value is 0 since it is equal to your guess. The remaining values in the posterior are again all different than 30, hence their losses are all different than 0.

To find the total loss, we simply sum over these individual losses again and the total loss comes out to 3,732. We have the visualization of the posterior distribution. Again, this time along with the squared loss function calculated for a possible serious of possible guesses within the range of the posterior distribution.

Table 3.3: L1: linear loss for g = 30

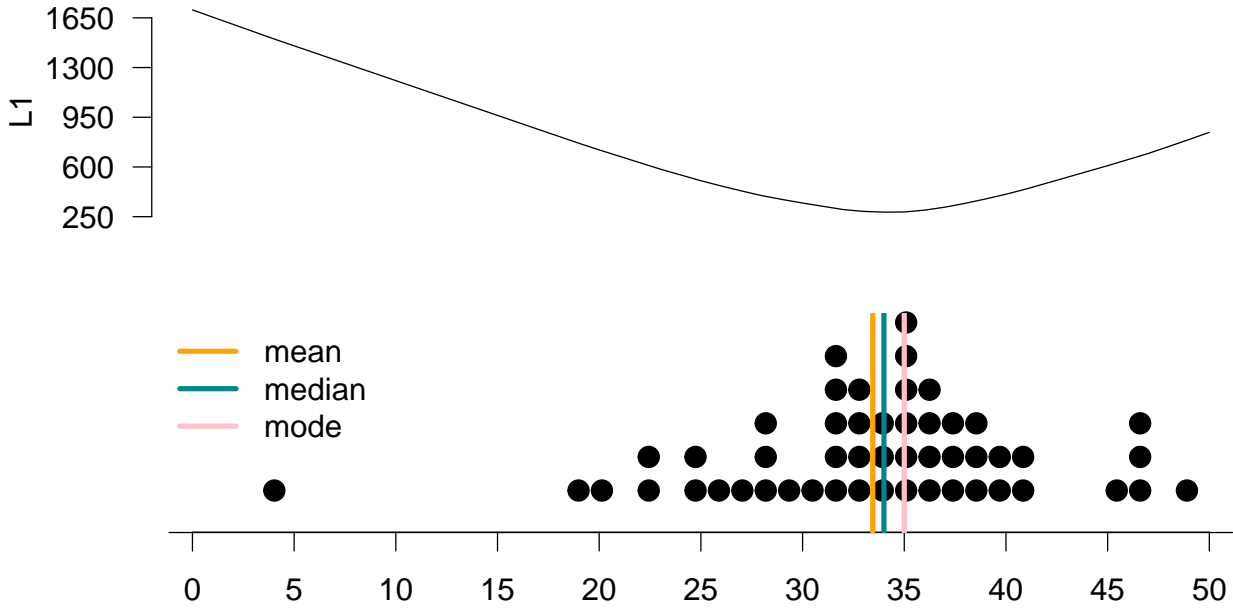| i	x_i	L1: |x_i-30|
1	4	1
2	19	1
3	20	1
	...	...
14	30	0
	...	...
50	47	1
51	49	1
	Total	346

Figure 3.3: L1 is minimized at the median of the posterior

Table 3.4: L2: squared loss for g = 30

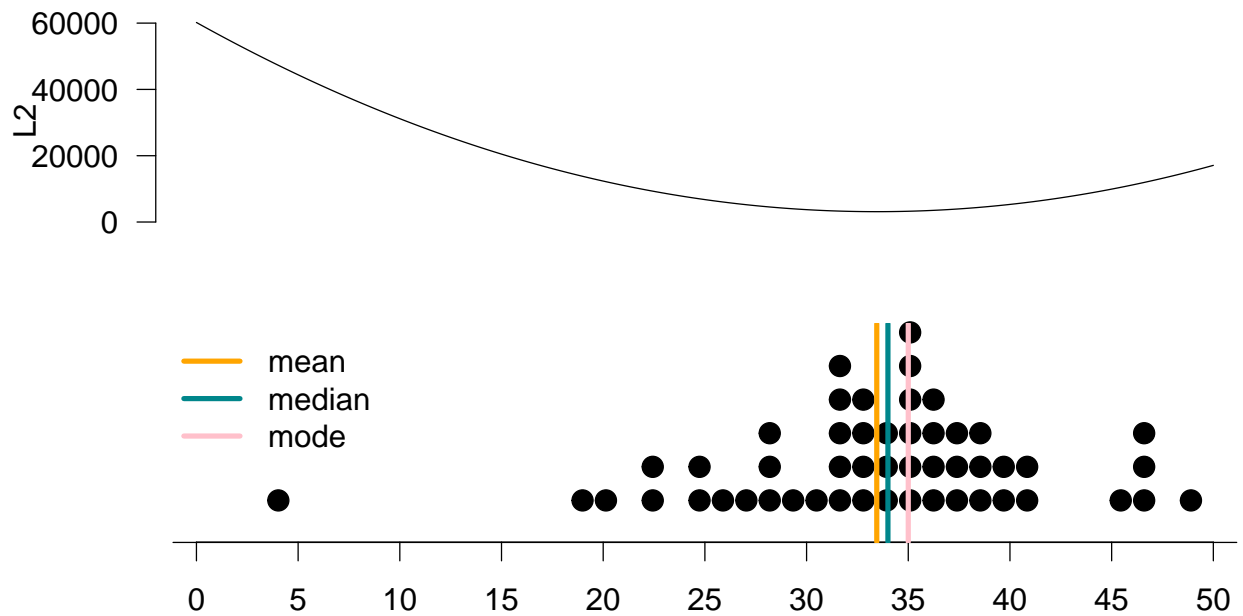| i	x_i	L2: (x_i-30)^2
1	4	1
2	19	1
3	20	1
	...	...
14	30	0
	...	...
50	47	1
51	49	1
	Total	3732

Figure 3.4: L2 is minimized at the mean of the posterior

Creating the visualization in Figure 3.4 had the same steps. Go through the same process described earlier for a guess of 30, for all guesses considered, and record the total loss. This time, the function has the lowest value when g is equal to the **mean** of the posterior. Hence, L_2 is minimized at the **mean** of the posterior distribution.

To sum up, the point estimate to report to your boss about the number of cars the dealership will sell per month **depends on your loss function**. In any case, you will choose to report the estimate that minimizes the loss.

- L_0 is minimized at the **mode** of the posterior distribution.
- L_1 is minimized at the **median** of the posterior distribution.
- L_2 is minimized at the **mean** of the posterior distribution.

3.1.3 Minimizing Expected Loss for Hypothesis Testing

In Bayesian statistics, the inference about a parameter is made based on the posterior distribution, and let's include this in the hypothesis test setting.

Suppose we have two competing hypothesis, H_1 and H_2. Then we get

- $P(H_1 \text{ is true} \mid \text{data}) = $ posterior probability of H_1
- $P(H_2 \text{ is true} \mid \text{data}) = $ posterior probability of H_2

One straightforward way of choosing between H_1 and H_2 would be to **choose the one with the higher posterior probability**. In other words, the potential decision criterion is to

- Reject H_1 if $P(H_1 \text{ is true} \mid \text{data}) < P(H_1 \text{ is true} \mid \text{data})$.

However, since hypothesis testing is a decision problem, we should also consider a loss function. Let's revisit the HIV testing example in Section 1.1.2, and suppose we want to test the two competing hypotheses below:

- H_1: Patient does not have HIV
- H_2: Patient has HIV

These are the only two possibilities, so they are mutually exclusive hypotheses that cover the entire decision space.

We can define the loss function as $L(d)$ – the loss that occurs when decision d is made. Then the Bayesian testing procedure minimizes the posterior expected loss.

The possible decisions (actions) are:

- d_1: Choose H_1 - decide that the patient does not have HIV
- d_2: Choose H_2 - decide that the patient has HIV

For each decision d, we might be right, or we might be wrong. If the decision is right, the loss $L(d)$ associated with the decision d is zero, i.e. no loss. If the decision is wrong, the loss $L(d)$ associated with the decision d is some positive value w.

For $d = d_1$, we have

- **Right**: Decide patient does not have HIV, and indeed they do not. $\Rightarrow L(d_1) = 0$
- **Wrong**: Decide patient does not have HIV, but they do. $\Rightarrow L(d_1) = w_1$

For $d = d_2$, we also have

- **Right**: Decide patient has HIV, and indeed they do. $\Rightarrow L(d_2) = 0$
- **Wrong**: Decide patient has HIV, but they don't $\Rightarrow L(d2) = w_2$

The consequences of making a wrong decision d_1 or d_2 are different.

Wrong d_1 is a **false negative**:

- We decide that patient does not have HIV when in reality they do.
- Potential consequences: no treatment and premature death! (severe)

Wrong d_2 is a **false positive**:

- We decide that the patient has HIV when in reality they do not.
- Potential consequences: distress and unnecessary further investigation. (not ideal but less severe than the consequences of a false negative decision)

Let's put these definitions in the context of the HIV testing example with ELISA.

Hypotheses

- H_1: Patient does not have HIV
- H_2: Patient has HIV

Decision

- d_1: Choose H_1 - decide that the patient does not have HIV
- d_2: Choose H_2 - decide that the patient has HIV

Losses

- $L(d_1) = \begin{cases} 0 & \text{if } d_1 \text{ is right} \\ w_1 = 1000 & \text{if } d_1 \text{ is wrong} \end{cases}$

- $L(d_2) = \begin{cases} 0 & \text{if } d_2 \text{ is right} \\ w_2 = 10 & \text{if } d_2 \text{ is wrong} \end{cases}$

The values of w_1 and w_2 are arbitrarily chosen. But the important thing is that w_1, the loss associated with a false negative determination, is much higher than w_2, the loss associated with a false positive determination.

Posteriors

The plus sign means that our patient had tested positive on the ELISA.

- $P(H_1|+) \approx 0.88$ - the posterior probability of the patient **not** having HIV given positive ELISA result
- $P(H_2|+) \approx 0.12$ - the posterior probability of the patient having HIV given positive ELISA result, as the complement value of $P(H_1|+)$

Expected losses

- $E[L(d_1)] = 0.88 \times 0 + 0.12 \times 1000 = 120$
- $E[L(d_2)] = 0.88 \times 10 + 0.12 \times 0 = 8.8$

Since the expected loss for d_2 is lower, we should make this decision – the patient has HIV.

Note that our decision is highly influenced by the losses we assigned to d_1 and d_2.

If the losses were symmetric, say $w_1 = w_2 = 10$, then the expected loss for d_1 becomes

$$E[L(d_1)] = 0.88 \times 0 + 0.12 \times 10 = 1.2,$$

while the expected loss for d_2 would not change. Therefore, we would choose d_1 instead; that is, we would decide that the patient does not have HIV.

To recap, Bayesian methodologies allow for the integration of losses into the decision making framework easily. And in Bayesian testing, we minimize the posterior expected loss.

3.1.4 Posterior Probabilities of Hypotheses and Bayes Factors

In this section, we will continue with the HIV testing example to introduce the concept of Bayes factors. Earlier, we introduced the concept of priors and posteriors. The **prior odds** is defined as **the ratio of the prior probabilities of hypotheses**.

Therefore, if there are two competing hypotheses being considered, then the prior odds of H_1 to H_2 can be defined as $O[H_1 : H_2]$, which is equal to $P(H_1)$ over probability of $P(H_2)$. In mathematical terms,

$$O[H_1 : H_2] = \frac{P(H_1)}{P(H_2)}$$

Similarly, the **posterior odds** is **the ratio of the two posterior probabilities of hypotheses**, written as

$$PO[H_1 : H_2] = \frac{P(H_1|\text{data})}{P(H_2|\text{data})}$$

Using Bayes' rule, we can rewrite the posterior probabilities as below:

$$
\begin{aligned}
PO[H_1 : H_2] &= \frac{P(H_1|\text{data})}{P(H_2|\text{data})} \\
&= \frac{(P(\text{data}|H_1) \times P(H_1))/P(\text{data})}{(P(\text{data}|H_2) \times P(H_2))/P(\text{data})} \\
&= \frac{(P(\text{data}|H_1) \times P(H_1))}{(P(\text{data}|H_2) \times P(H_2))} \\
&= \boxed{\frac{P(\text{data}|H_1)}{P(\text{data}|H_2)}} \times \boxed{\frac{P(H_1)}{P(H_2)}} \\
&= \textbf{Bayes factor} \times \textbf{prior odds}
\end{aligned}
$$

In mathematical notation, we have

$$PO[H_1 : H_2] = BF[H_1 : H_2] \times O[H_1 : H_2]$$

In other words, the posterior odds is the product of the Bayes factor and the prior odds for these two hypotheses.

The Bayes factor quantifies the evidence of data arising from H_1 versus H_2.

In a discrete case, the Bayes factor is simply the ratio of the likelihoods of the observed data under the two hypotheses, written as

$$BF[H_1 : H_2] = \frac{P(\text{data}|H_1)}{P(\text{data}|H_2)}.$$

On the other hand, in a continuous case, the Bayes factor is the ratio of the marginal likelihoods, written as

$$BF[H_1 : H_2] = \frac{\int P(\text{data}|\theta, H_1)d\theta}{\int P(\text{data}|\theta, H_2)d\theta}.$$

Note that θ is the set formed by all possible values of the model parameters.

In this section, we will stick with the simpler discrete case. And in upcoming sections, we will revisit calculating Bayes factors for more complicated models.

Let's return to the HIV testing example from earlier, where our patient had tested positive in the ELISA.

Hypotheses

- H_1: Patient does not have HIV
- H_2: Patient has HIV

Priors

The prior probabilities we place on these hypothesis came from the prevalence of HIV at the time in the general population. We were told that the prevalence of HIV in the population was 1.48 out of 1000, hence the prior probability assigned to H_2 is 0.00148. And the prior assigned to H_1 is simply the complement of this.

- $P(H_1) = 0.99852$ and $P(H_2) = 0.00148$

The prior odds are

- $O[H_1 : H_2] = \dfrac{P(H_1)}{P(H_2)} = \dfrac{0.99852}{0.00148} = 674.6757$

Posteriors

Given a positive ELISA result, the posterior probabilities of these hypotheses can also be calculated, and these are approximately 0.88 and 0.12. We will hold on to more decimal places in our calculations to avoid rounding errors later.

- $P(H_1|+) = 0.8788551$ and $P(H_2|+) = 0.1211449$

The posterior odds are

- $PO[H_1 : H_2] = \dfrac{P(H_1|+)}{P(H_2|+)} = \dfrac{0.8788551}{0.1211449} = 7.254578$

Bayes Factor

Finally, we can calculate the Bayes factor as the ratio of the posterior odds to prior odds, which comes out to approximately 0.0108. Note that in this simple discrete case the Bayes factor, it simplifies to the ratio of the likelihoods of the observed data under the two hypotheses.

$$BF[H_1 : H_2] = \frac{PO[H_1 : H_2]}{O[H_1 : H_2]} = \frac{7.25457}{674.6757} \approx 0.0108$$
$$= \frac{P(+|H_1)}{P(+|H_2)} = \frac{0.01}{0.93} \approx 0.0108$$

Alternatively, remember that the true positive rate of the test was 0.93 and the false positive rate was 0.01. Using these two values, the Bayes factor also comes out to approximately 0.0108.

So now that we calculated the Bayes factor, the next natural question is, what does this number mean? A commonly used scale for interpreting Bayes factors is proposed by Jeffreys (1961), as in Table 3.5. If the Bayes factor is between 1 and 3, the evidence against H_2 is not worth a bare mention. If it is 3 to 20, the

Table 3.5: Interpreting the Bayes factor

BF[H_1:H_2]	Evidence against H_2
1 to 3	Not worth a bare mention
3 to 20	Positive
20 to 150	Strong
> 150	Very strong

Table 3.6: Interpreting the Bayes factor

2*log(BF[H_2:H_1])	Evidence against H_1
0 to 2	Not worth a bare mention
2 to 6	Positive
6 to 10	Strong
> 10	Very strong

evidence is positive. If it is 20 to 150, the evidence is strong. If it is greater than 150, the evidence is very strong.

It might have caught your attention that the Bayes factor we calculated does not even appear on the scale. To obtain a Bayes factor value on the scale, we will need to change the order of our hypotheses and calculate $BF[H_2 : H_1]$, i.e. the Bayes factor for H_2 to H_1. Then we look for evidence against H_1 instead.

We can calculate $BF[H_2 : H_1]$ as a reciprocal of $BF[H_1 : H_2]$ as below:

$$BF[H_2 : H_1] = \frac{1}{BF[H_1 : H_2]} = \frac{1}{0.0108} = 92.59259$$

For our data, this comes out to approximately 93. Hence the evidence against H_1 (the patient does not have HIV) is strong. Therefore, even though the posterior for having HIV given a positive result, i.e. $P(H_2|+)$, was low, we would still decide that the patient has HIV, according to the scale based on a positive ELISA result.

An intuitive way of thinking about this is to consider not only the posteriors, but also the priors assigned to these hypotheses. Bayes factor is the ratio of the posterior odds to prior odds. While 12% is a low posterior probability for having HIV given a positive ELISA result, this value is still much higher than the overall prevalence of HIV in the population (in other words, the prior probability for that hypothesis).

Another commonly used scale for interpreting Bayes factors is proposed by Kass and Raftery (1995), and it deals with the natural logarithm of the calculated Bayes factor. The values can be interpreted in Table 3.6.

Reporting of the log scale can be helpful for numerical accuracy reasons when the likelihoods are very small. Taking two times the natural logarithm of the Bayes factor we calculated earlier, we would end up with the same decision that the evidence against H_1 is strong.

$$2 \times \log(92.59259) = 9.056418$$

To recap, we defined prior odds, posterior odds, and the Bayes factor. We learned about scales by which we can interpret these values for model selection. We also re-emphasize that in Bayesian testing, the order in which we evaluate the models of hypotheses does **not** matter. The Bayes factor of H_2 versus H_1, $BF[H_2 : H_1]$, is simply the reciprocal of the Bayes factor for H_1 versus H_2, that is, $BF[H_1 : H_2]$.

3.2   Exercises

Test something here

Chapter 4

Inference and Decision-Making with Multiple Parameters

This chapter 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 optional sections with more details on the derivations.

4.1 The Normal-Gamma Conjugate Family

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

$$\mu \mid \sigma^2 \sim \mathsf{N}(m_0, \sigma^2/n_0) \tag{4.1}$$

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,

$$\phi \sim \mathsf{Gamma}\left(\frac{v_0}{2}, \frac{v_0 s_0^2}{2}\right) \tag{4.2}$$

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 (4.1) and the marginal Gamma distribution for ϕ in (4.2) lead to a joint distribution for the pair (μ, ϕ) that we will call the Normal-Gamma family of distributions:

$$(\mu, \phi) \sim \mathsf{NormalGamma}(m_0, n_0, s_0^2, v_0) \tag{4.3}$$

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

$$(\mu, \phi) \mid \text{data} \sim \mathsf{NormalGamma}(m_n, n_n, s_n^2, v_n) \tag{4.4}$$

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:

$$m_n = \frac{n\bar{Y} + n_0 m_0}{n + n_0}$$

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

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

$$p(\mu) = \frac{\Gamma\left(\frac{v_n+1}{2}\right)}{\sqrt{\pi v_n} \frac{s_n}{\sqrt{n_n}} \Gamma\left(\frac{v_n}{2}\right)} \left(1 + \frac{1}{v_n} \frac{(\mu - m_n)^2}{s_n^2/n_n}\right)^{-\frac{v_n+1}{2}} \tag{4.5}$$

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 (4.5). 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

```r
library(statsr)
library(ggplot2)
```

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 4.2.1, 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

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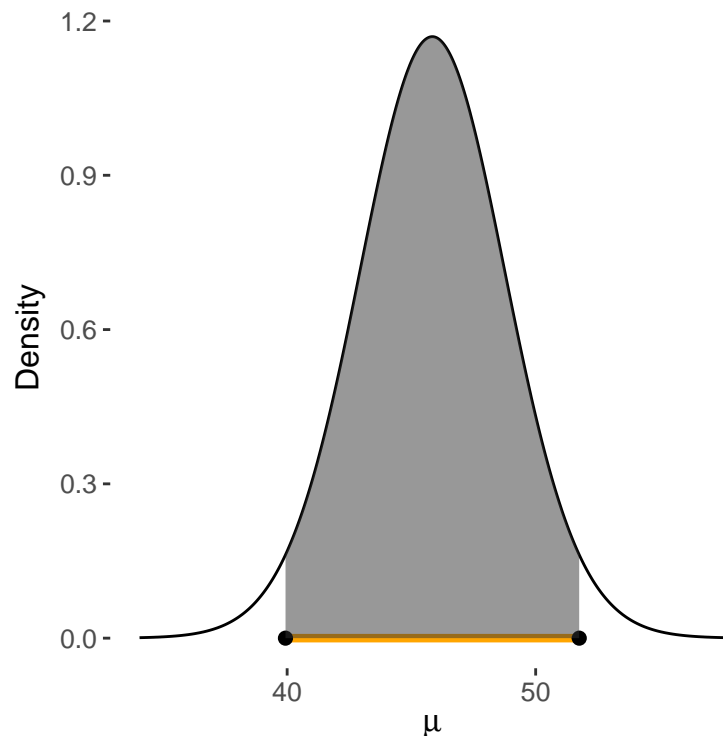
is the orange interval given by the Lower endpoint L and upper endpoint U where the probability that mu is in the interval (L, U) is the shaded area which is equal to zero point nine five.

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.

4.1.1   Details of Results (optional reading)**

TBA

4.2   Monte Carlo Inference

Figure out how to reuse code chunks with purl(), read_chunk() ref.label as in post http://jeromyanglim. tumblr.com/post/113132013271/how-to-run-r-code-in-subset-of-named-code-chunks

```
library(statsr)
library(ggplot2)
```

In this section, we will illustrate Monte Carlo sampling from posterior distributions and how it can be used for inference. In Section 4.1, we can get the posterior distributions for the precision (inverse variance), and the mean given the precision. Then the marginal distribution of the mean can be obtained via integration.

Here is a recap of the joint posterior distribution for the mean μ and the precision $\phi = 1/\sigma^2$:

- Conditional posterior $\mu \mid \text{data}, \sigma^2 \sim \mathsf{N}(m_n, \sigma^2/n_n)$
- Marginal posterior $1/\sigma^2 = \phi \mid \text{data} \sim \mathsf{Gamma}(v_n/2, s_n^2 v_n/2)$
- Marginal posterior $\mu \mid \text{data} \sim \mathsf{t}(v_n, m_n, s_n^2/n_n)$

What if we are interested in the distribution of the standard deviation σ itself, or other transformations of the parameters? There may not be a closed-form expression for the distributions.

However, it turns out that **Monte Carlo sampling** is an easy way to make inference, when we cannot analytically calculate distributions of parameters, expectations, or probabilities. Monte Carlo methods are computational algorithms that rely on repeated random sampling to calculate numerical results. The name refers to the famous Monte Carlo Casino in Monaco, home to games of chance such as Roulette.

Let's start with a case where we know the posterior distribution.

For posterior inference about the precision ϕ using Monte Carlo simulation, we generate S random samples from the posterior distribution:

$$\phi^{(1)}, \phi^{(2)}, \cdots, \phi^{(S)} \overset{\text{iid}}{\sim} \mathsf{Gamma}(v_n/2, s_n^2 v_n/2)$$

The term **iid** stands for **i**ndependent and **i**dentically **d**istributed. In other words, the S draws of ϕ are independent and identically distributed from the gamma distribution.

Then the empirical distribution of the S samples is used to approximate the actual posterior distribution. From the samples, the sample mean of the draws of ϕ can be used to approximate the posterior mean of ϕ.

Likewise, we can calculate probabilities, quantiles and other functions using the samples from the posterior distribution. For example, if we want to calculate the posterior expectation of some function of ϕ, written as $g(\phi)$, we can approximate that by taking the average of the function, and evaluate it at the S draws of ϕ, written as $\frac{1}{S} \sum_{i=1}^{S} g(\phi^{(i)})$.

The approximation improves as the size of the Monte Carlo simulation S increases.

$$\frac{1}{S}\sum_{i=1}^{S} g(\phi^{(i)}) \to E(g(\phi \mid \text{data}))$$

Example

We will apply this with the tap water example. To start, we will set a random seed, which allows the results to be replicated. To generate 1,000 draws from the gamma posterior distribution, we use the `rgamma` function R with the posterior hyperparameters from last time.

```
data(tapwater)
m_0 = 35;
n_0 = 25;
s2_0 = 156.25;
v_0 = n_0 - 1
Y = tapwater$tthm
ybar = mean(Y)
s2 = round(var(Y),1)
n = length(Y)
n_n = n_0 + n
m_n = round((n*ybar + n_0*m_0)/n_n, 1)
v_n = v_0 + n
s2_n = round( ((n-1)*s2 + v_0*s2_0 + n_0*n*(m_0 - ybar)^2/n_n)/v_n, 1)
L = qt(.025, v_n)*sqrt(s2_n/n_n) + m_n
U = qt(.975, v_n)*sqrt(s2_n/n_n) + m_n
```

```
set.seed(8675309)
phi = rgamma(1000, shape = v_n/2, rate=s2_n*v_n/2)
```

Figure 4.1 shows the histogram of the 1,000 draws of ϕ generated from the Monte Carlo simulation, representing the empirical distribution. The orange line represents the actual gamma posterior density.

Try changing the random seed or increasing the number of simulations, and see how the approximation changes.

We will now use Monte Carlo simulations to approximate the distribution of σ. Since $\sigma = 1/\sqrt{\phi}$, we can apply the transformation to the 1,000 draws of ϕ to obtain a random sample of σ. We can then estimate the posterior mean of σ by calculating the sample mean of the 1,000 draws.

```
sigma = 1/sqrt(phi)
mean(sigma) # posterior mean of sigma
```

```
## [1] 21.814
```

Similarly, we can obtain a 95% credible interval for σ by finding the sample quantiles of the distribution.

```
quantile(sigma, c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 18.20700 26.53736
```
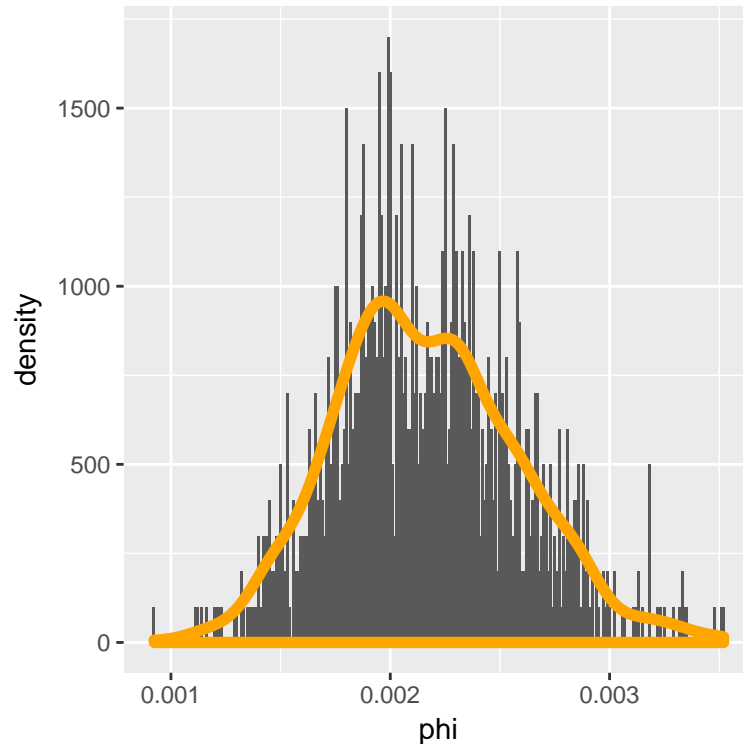
Figure 4.1: Empirical distribution of the tap water example

ADD HPD interval ???

Summary

To recap, we have introduced the powerful method of Monte Carlo simulation for posterior inference. Monte Carlo methods provide estimates of expectations, probabilities, and quantiles of distributions from the simulated values. Monte Carlo simulation also allows us to approximate distributions of functions of the parameters, or the transformations of the parameters.

Next we will discuss predictive distributions and show how Monte Carlo simulation may be used to help choose prior hyperparameters, using the prior predictive distribution of data.

4.2.1   Predictive Distributions

```r
library(statsr)
library(ggplot2)
```

In this section, we will discuss prior and posterior **predictive** distributions of the data and show how Monte Carlo sampling from the prior predictive distribution can help select hyper parameters.

We can obtain the prior predictive distribution of the data, by taking the joint distribution of the data and the parameters in averaging over the possible values of the parameters from the prior.

- Prior:

$$\frac{1}{\sigma^2} = \phi \sim \mathsf{Gamma}\left(\frac{v_0}{2}, \frac{v_0 s_0^2}{2}\right)$$

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

- Sampling model:

$$Y_i \mid \mu, \sigma^2 \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$$

- Prior predictive distribution for Y:

$$p(Y) = \iint p(Y \mid \mu, \sigma^2) p(\mu \mid \sigma^2) p(\sigma^2) d\mu \, d\sigma^2$$

$$Y \sim t(v_0, m_0, s_0^2 + s_0^2/n_0)$$

This distribution of the observables can be used to help elicit prior hyper parameters as in the tap water example.

A report from the city water department suggests that levels of TTHM are expected to be between 10-60 parts per billion (ppb).

- Set the prior mean μ to be at the midpoint of the interval: $m_0 = (60 + 10)/2 = 35$

- Standard deviation: Based on the empirical rule, 95% observations are within $\pm 2\sigma$ of μ, we expect that the range of the data should be 4σ.

- Prior estimate of sigma: $s_0 = (60 - 10)/4 = 12.5$ or $s_0^2 = [(60 - 10)/4]^2 = 156.25$

To complete the specification, we also need to choose the prior sample size n_0 and degrees of freedom v_0. As the degrees of freedom of the variance are $n - 1$, we set $v_0 = n_0 - 1$. We will draw samples from the prior predictive distribution and modify n_0 so that the simulated data agree with our prior assumptions.

The following `R` code shows a simulation from the predictive distribution with the prior sample size of 2. Please note that the number of Monte Carlo simulations should not be confused with the prior sample size n_0.

We begin by simulating ϕ, transfering ϕ to calculate σ, and then simulating values of μ. Finally, the simulated values of μ, σ are used to generate possible values of TTHM denoted by Y.

```
m_0 = (60+10)/2; s2_0 = ((60-10)/4)^2;
n_0 = 2; v_0 = n_0 - 1
set.seed(1234)
phi = rgamma(10000, v_0/2, s2_0*v_0/2)
sigma = 1/sqrt(phi)
mu = rnorm(10000, mean=m_0, sd=sigma/(sqrt(n_0)))
y = rnorm(10000, mu, sigma)
quantile(y, c(0.025,0.975))
```

```
##      2.5%     97.5%
## -140.1391  217.7050
```
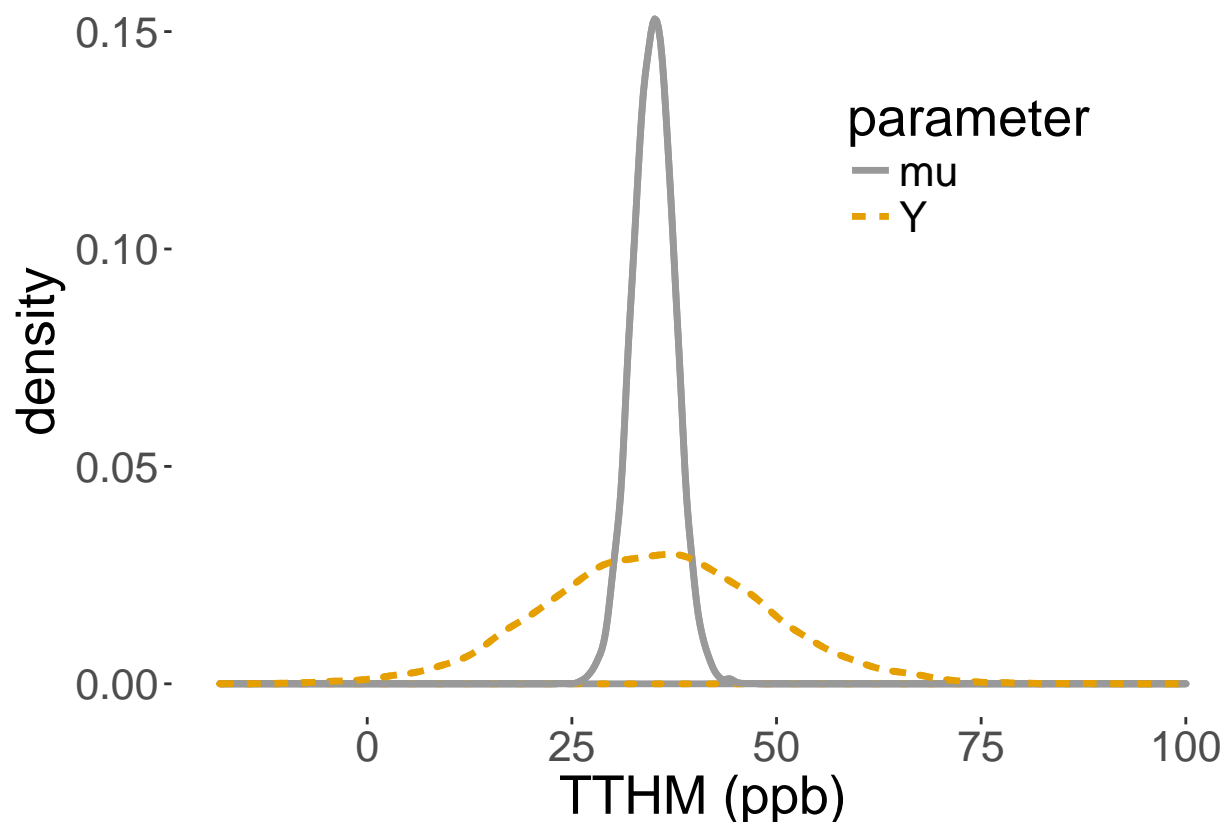
Figure 4.2: Prior density

This forward simulation propagates uncertainty in μ, σ to the prior predictive distribution of the data. Calculating the sample quantiles from the samples of the prior predictive for Y, we see that the 95% predictive interval includes negative values. Since TTHM is non-negative, we need to adjust n_0 and repeat.

After some trial and error, we find that the prior sample size of 25 (in fact the Central Limit Theorem suggests at least 25 or 30 to be "sufficiently large"), the empirical quantiles from the prior predictive distribution are close to the range of 10 to 16 that we were given as prior information.

```
m_0 = (60+10)/2; s2_0 = ((60-10)/4)^2;
n_0 = 25; v_0 = n_0 - 1
set.seed(1234)
phi = rgamma(10000, v_0/2, s2_0*v_0/2)
sigma = 1/sqrt(phi)
mu = rnorm(10000, mean=m_0, sd=sigma/(sqrt(n_0)))
y = rnorm(10000, mu, sigma)
quantile(y, c(0.025,0.975))
```

```
##      2.5%     97.5%
## 8.802515 61.857350
```

Figure 4.2 shows an estimate of the prior distribution of μ in gray and the more dispersed prior predictive distribution in TTHM in orange, obtained from the Monte Carlo samples.

Using the Monte Carlo samples, we can also estimate the prior probability of negative values of TTHM by counting the number of times the simulated values are less than zero out of the total number of simulations.

```
sum(y < 0)/length(y)  # P(Y < 0) a priori
```

```
## [1] 0.0049
```

With the normal prior distribution, this probability will never be zero, but may be acceptably small, so we can still use the conjugate normal gamma model for analysis.

We can use the same strategy to generate samples from the predictive distribution of a new measurement Y_{n+1} given the observed data. In mathematical terms, the posterior predictive distribution is written as

$$Y_{n+1} \mid Y_1, \ldots, Y_n \sim \mathsf{t}(v_n, m_n, s_n^2(1 + 1/n_n))$$

In the code, we replace the prior hyper parameters with the posterior hyper parameters from last time.

```
set.seed(1234)
phi = rgamma(10000, v_n/2, s2_n*v_n/2)
sigma = 1/sqrt(phi)
post_mu = rnorm(10000, mean=m_n, sd=sigma/(sqrt(n_n)))
pred_y =  rnorm(10000,post_mu, sigma)
quantile(pred_y, c(.025, .975))
```

```
##     2.5%     97.5%
##  3.280216 89.830212
```

Figure 4.3 shows the Monte Carlo approximation to the prior distribution of μ, and the posterior distribution of μ which is shifted to the right. The prior and posterior predictive distributions are also depicted, showing how the data have updated the prior information.

Using the Monte-Carlo samples from the posterior predictive distribution, we can estimate the probability that a new TTHM sample will exceed the legal limit of 80 parts per billion, which is approximately 0.06.

```
sum(pred_y > 80)/length(pred_y)  # P(Y > 80 | data)
```

```
## [1] 0.0619
```

By using Monte-Carlo methods, we can obtain prior and posterior predictive distributions of the data.

- Sampling from the prior predictive distribution can help with the selection of prior hyper parameters and verify that these choices reflect the prior information that is available.

- Visualizing prior predictive distributions based on Monte Carlo simulations can help explore implications of our prior assumptions such as the choice of the hyper parameters or even assume distributions.

- If samples are incompatible with known information, such as support on positive values, we may need to modify assumptions and look at other families of prior distributions.
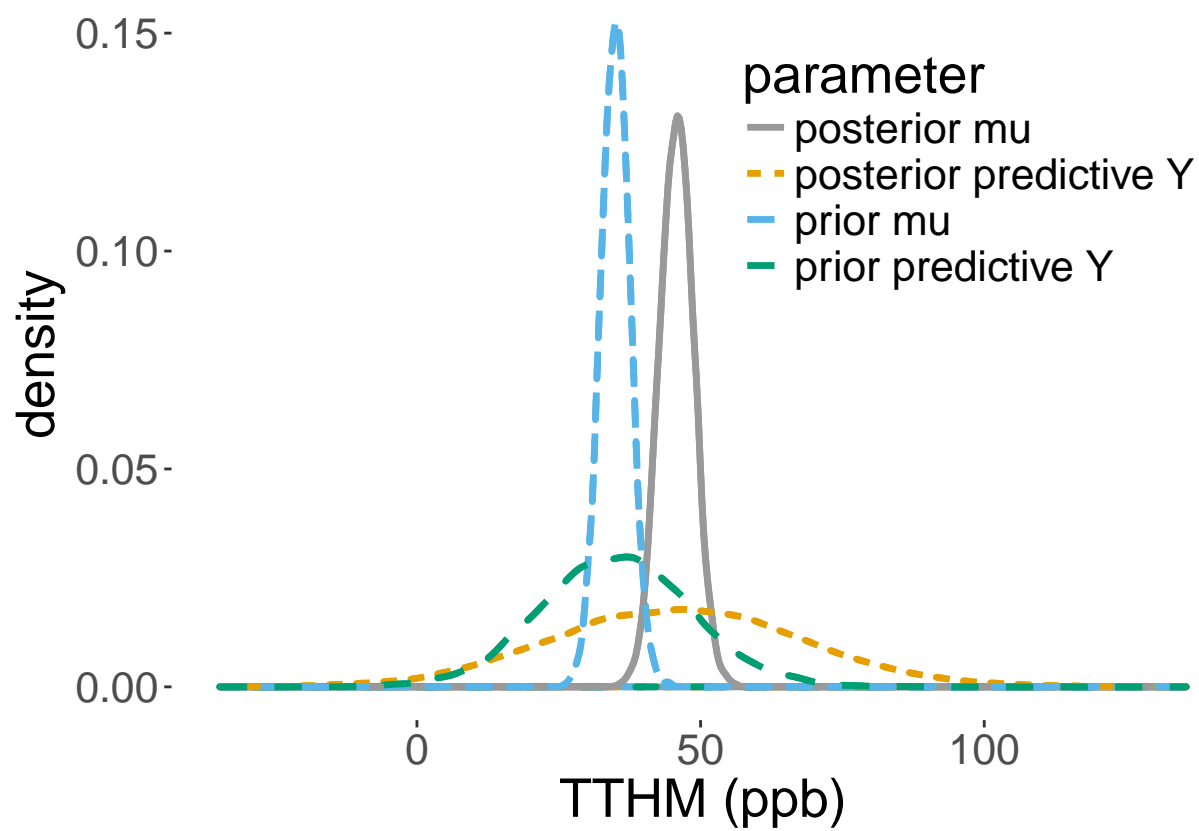
4.3   Reference Priors

Figure 4.3: Posterior densities

```
library(statsr)
library(ggplot2)
```

In Section 4.2.1, we described a way of specifying an informative prior distribution for inference about TTHM in tapwater based on additional prior information. We had to use a prior sample size that was comparable to the observed sample size for the prior predictive under the Normal-Gamma distribution to agree with the reported prior interval.

However, there may be cases where prior information is not available, or you may wish to present an objective analysis where minimal prior information is used. Or perhaps, you want to use the Bayesian paradigm to make probability statements about parameters, but not use any prior information.

In this section, we will present reference priors for normal data, which can be viewed as a limiting form of the Normal-Gamma conjugate prior distribution. **Can you actually perform a Bayesian analysis without using prior information?**

Conjugate priors can be interpreted to be based on a prior sample. What happens in the conjugate Normal-Gamma prior if we take our prior sample size n_0 to go to zero? If we have no data, then we will define the prior sample variance s_0^2 to go to 0, and based on the relationship between prior sample sized and prior degrees of freedom, we will let the prior degrees of freedom go to the prior sample size minus one, or negative 1, i.e. $v_0 = n_0 - 1 \to -1$.

With this limit, we have the following properties:

- The posterior mean goes to the sample mean.

- The posterior sample size is the observed sample size.

- The posterior degrees of freedom go to the sample degrees of freedom.

- The posterior variance parameter goes to the sample variance.

In this limit, the posterior hyperparameters do not depend on the prior hyperparameters.

Since $n_0 \to 0, s_0^2 \to 0, v_0 = n_0 - 1 \to -1$, we have in mathematical terms:

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

This limiting Normal-gamma distribution, $\mathsf{N-Gamma}(0,0,0,-1)$, is not really a normal gamma distribution, as the density does not integrate to 1. The form of the limit can be viewed as a prior for μ that is proportional to a constant, or uniform/flat on the whole real line. And a prior for the variance is proportional to 1 over the variance. The joint prior is taken as the product of the two.

$$p(\mu \mid \sigma^2) \propto 1$$
$$p(\sigma^2) \propto 1/\sigma^2$$
$$p(\mu, \sigma^2) \propto 1/\sigma^2$$

This is refered to as a **reference prior** because the posterior hyperparameters do not depend on the prior hyperparameters.

In addition, $\mathsf{NormalGamma}(0, 0, 0, -1)$ is a special case of a reference prior, known as the independent Jeffreys prior. While Jeffreys used other arguments to arrive at the form of the prior, the goal was to have an **objective prior** invariant to shifting the data by a constant or multiplying by a constant.

Now, a naive approach to constructing a non-informative distribution might be to use a uniform distribution to represent lack of knowledge. However, would you use a uniform distribution for σ^2, or a uniform distribution for the precision $1/\sigma^2$? Or perhaps a uniform distribution for σ? These would all lead to different posteriors with little justification for any of them. This ambiguity led Sir Harold Jeffreys to propose reference distributions for the mean and variance for situations where prior information was limited. These priors are **invariant** to the units of the data.

The unnormalized priors that do not integrate to a constant are called **improper distributions**. An important consideration in using them is that one cannot generate samples from the prior or the prior predictive distribution to data and are referred to as **non-generative distributions**.

While the reference prior is not a proper prior distribution, and cannot reflect anyone's actual prior beliefs, the formal application phase rule can still be used to show that **the posterior distribution is a valid normal gamma distribution**, leading to a formal phase posterior distribution. That depends only on summary statistics of the data.

The posterior distribution $\mathsf{NormalGamma}(\bar{Y}, n, s^2, n-1)$ breaks down to

$$\mu \mid \sigma^2, \text{data} \sim \mathsf{N}(\bar{Y}, \sigma^2/n)$$
$$1/\sigma^2 \mid \text{data} \sim \mathsf{Gamma}((n-1)/2, s^2(n-1)/2).$$

- Under the reference prior $p(\mu, \sigma^2) \propto 1/\sigma^2$, the posterior distribution after standardizing μ has a Student t distribution with n minus one degrees of freedom.

$$\frac{\mu - \bar{Y}}{\sqrt{s^2/n}} \mid \text{data} \sim \mathsf{t}(n-1, 0, 1)$$

* Prior to seeing the data, the distribution of the standardized sample mean given μ and σ also has a Student t distribution.

$$\frac{\mu - \bar{Y}}{\sqrt{s^2/n}} \mid \mu, \sigma^2 \sim \mathsf{t}(n-1, 0, 1)$$

- Both frequentist sampling distributions and Bayesian reference posterior distributions lead to intervals of this form:

$$(\bar{Y} - t_{1-\alpha/2}s/\sqrt{n},\ \bar{Y} + t_{1-\alpha/2}s/\sqrt{n})$$

- However, only the Bayesian approach justifies the probability statements about μ being in the interval after seeing the data.

$$P(\bar{Y} - t_{1-\alpha/2}s/\sqrt{n} < \mu < \bar{Y} + t_{1-\alpha/2}s/\sqrt{n}) = 1 - \alpha$$

We can use either analytic expressions based on the t-distribution, or Monte Carlo samples from the posterior predictive distribution, to make predictions about a new sample.

Here is some code to generate the Monte Carlo samples from the tap water example:
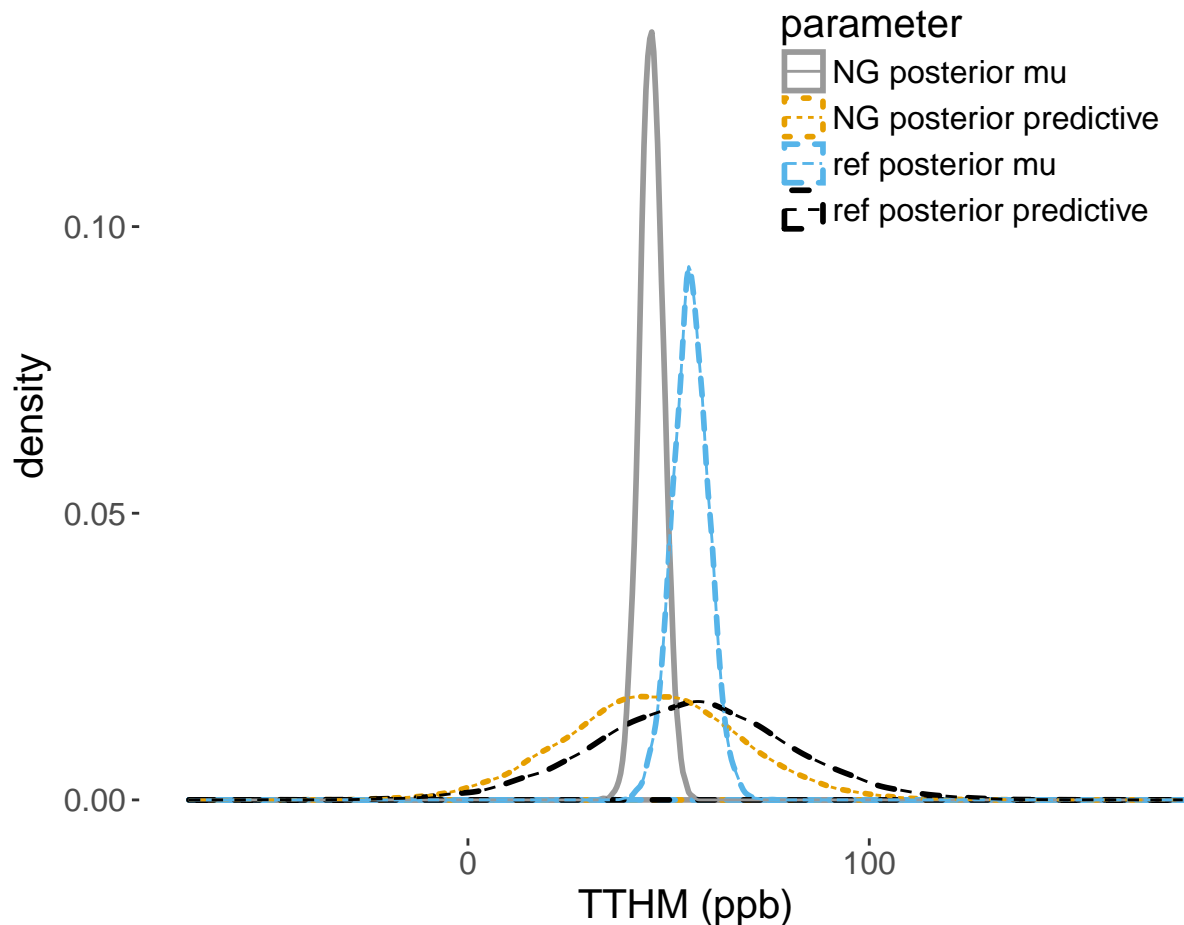
Figure 4.4: Comparison of posterior densities

```
phi = rgamma(10000, (n-1)/2, s2*(n-1)/2)
sigma = 1/sqrt(phi)
post_mu = rnorm(10000, mean=ybar, sd=sigma/(sqrt(n)))
pred_y =  rnorm(10000,post_mu, sigma)
quantile(pred_y, c(.025, .975))
```

```
##      2.5%     97.5%
##   6.692877 104.225954
```

Using the Monte Carlo samples, Figure 4.4 shows the posterior distribution based on the informative Normal-Gamma prior and the reference prior. Both the posterior distribution for μ and the posterior predictive distribution for a new sample are shifted to the right, and are centered at the sample mean. The posterior for μ under the reference prior is less concentrated around its mean than the posterior under the informative prior, which leads to an increased posterior sample size and hence increased precision.

The posterior probability that a new sample will exceed the legal limit of 80 ppb under the reference prior is roughly 0.15, which is more than double the probability of 0.06 from the posterior under the informative prior.

```
sum(pred_y > 80)/length(pred_y)   # P(Y > 80 | data)
```

```
## [1] 0.1534
```

In constructing the informative prior from the reported interval, there are two critical assumptions. First, the prior data are exchangeable with the observed data. Second, the conjugate normal gamma distribution is suitable for representing the prior information. These assumptions may or may not be verifiable, but they should be considered carefully when using informative conjugate priors.

In the case of the tap water example, there are several concerns: One, it is unclear that the prior data are exchangeable with the observed data. For example, water treatment conditions may have changed. Two, the prior sample size was not based on a real prior sample, but instead selected so that the prior predictive intervals under the normal gamma model agreed with the prior data. As we do not have access to the prior data, we cannot check assumptions about normality that would help justify the prior. Other skewed distributions may be consistent with the prior interval, but lead to different conclusions.

To recap, we have introduced a reference prior for inference for normal data with an unknown mean and variance. Reference priors are often part of a prior sensitivity study and are used when objectivity is of utmost importance.

If conclusions are fundamentally different with an informative prior and a reference prior, one may wish to carefully examine assumputions that led to the informative prior.

- Is the prior information based on a prior sample that is exchangable with the observed data?

- Is the normal-gamma assumption appropriate?

Informative priors can provide more accurate inference when data are limited, and the transparency of explicitly laying out prior assumptions is an important aspect of reproducible research. However, one needs to be careful that certain prior assumptions may lead to un-intended consequences.

Next, we will investigate a prior distribution that is a mixture of conjugate priors, so the new prior distribution provides robustness to prior mis-specification in the prior sample size.

While we will no longer have nice analytical expressions for the posterior, we can simulate from the posterior distribution using a Monte Carlo algorithm called Markov chain Monte Carlo (MCMC).

4.4   Mixtures of Conjugate Priors

```
library(statsr)
library(ggplot2)
```

In this section, we will describe priors that are constructed as a mixture of conjugate priors – in particular, the Cauchy distribution. As these are no longer conjugate priors, nice analytic expressions for the posterior distribution are not available. However, we can use a Monte Carlo algorithm called Markov chain Monte Carlo (MCMC) for posterior inference.

In many situations, we may have reasonable prior information about the mean μ, but we are less confident in how many observations our prior beliefs are equivalent to. We can address this uncertainty in the prior sample size, through an additional prior distribution on a n_0 via a hierarchical prior.

The hierarchical prior for the normal gamma distribution is written as

$$\mu \mid \sigma^2, n_0 \sim \mathsf{N}(m_0, \sigma^2/n_0)$$
$$n_0 \mid \sigma^2 \sim \mathsf{Gamma}(1/2, r^2/2)$$

If $r = 1$, then this corresponds to a prior expected sample size of one because the expectation of $\mathsf{Gamma}(1/2, 1/2)$ is one.

The marginal prior distribution from μ can be attained via integration, and we get

$$\mu \mid \sigma^2 \sim \mathsf{C}(m_0, \sigma^2 r^2)$$

This is a **Cauchy distribution** centered at the prior mean m_0, with the scale parameter $\sigma^2 r^2$. The probability density function (pdf) is:

$$p(\mu \mid \sigma) = \frac{1}{\pi \sigma r} \left(1 + \frac{(\mu - m_0)^2}{\sigma^2 r^2} \right)^{-1}$$

The Cauchy distribution does not have a mean or standard deviation, but the center (location) and the scale play a similar role to the mean and standard deviation of the normal distribution. The Cauchy distribution is a special case of a student t distribution with one degree of freedom.

As Figure 4.5 shows, the standard Cauchy distribution with $r = 1$ and the standard normal distribution $\mathsf{N}(0, 1)$ are centered at the same location. But the Cauchy distribution has heavier tails – more probability on extreme values than the normal distribution with the same scale parameter σ. Cauchy priors were recommended by Sir Harold Jeffreys as a default objective prior for both estimation and testing.

ADD MORE DETAILS BEYOND THE VIDEO?

4.5   Markov Chain Monte Carlo (MCMC)

```r
library(statsr)
library(ggplot2)
```

The Cauchy prior described in Section 4.4 is not a contrary prior, and therefore, the posterior distribution from $(\mu \mid \sigma)$, is not Cauchy or any well-known distribution. Fortunately, the conditional distribution of $(\mu, \sigma \mid n_0, \text{data})$, is normal gamma and easy to simulate from, as we learned in the previous sections. The conditional distribution of $(n_0 \mid \mu, \sigma, \text{data})$ is a gamma distribution, also easy to simulate from the given μ, σ.

It turns out that if we alternate generating Monte Carlo samples from these conditional distributions, the sequence of samples converges to samples from the joint distribution of (μ, σ, n_0), as the number of simulated values increases. The Monte Carlo algorithm we have just described is a special case of Markov chain Monte Carlo (MCMC), known as the Gibbs sampler.

Let's look at the pseudo code for the algorithm.

```r
# initialize MCMC
sigma2[1] = 1; n_0[1]=1; mu[1]=m_0

#draw from full conditional distributions
for (i in 2:S) {
  mu[i]     = p_mu(sigma2[i-1], n_0[i-1],  m_0, r, data)
  sigma2[i] = p_sigma2(mu[i], n_0[i-1],    m_0, r, data)
  n_0[i]    = p_n_0(mu[i], sigma2[i],      m_0, r, data)
}
```
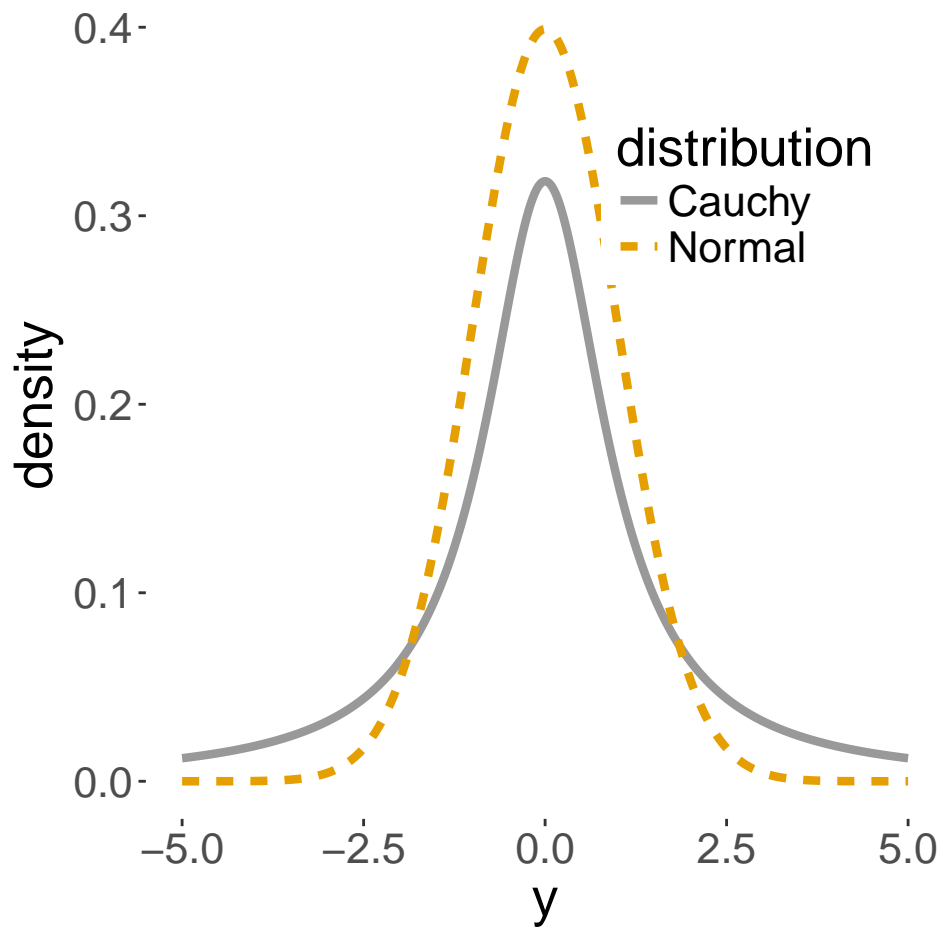
Figure 4.5: Cauchy distribution

We start with the initial values of each of the parameters for $i = 1$. In theory, these can be completely arbitrary, as long as they are allowed values for the parameters.

For each iteration i, the algorithm will cycle through generating each parameter, given the **current** value of the other parameters. The functions `p_mu`, `p_sigma2`, and `p_n_0` return a simulated value from the respective distribution conditional on the inputs.

Whenever we update a parameter, we use the **new value** in the subsequent steps as the n draws for σ, n_0. We will repeat this until we reach iteration S, leading to a dependent sequence of s draws from the joint posterior distribution.

Incorporating the tap water example in Section 4.1, we will use MCMC to generate samples under the Cauchy prior. We set 35 as the location parameter and $r = 1$. To complete our prior specification, we use the Jeffrey's reference prior on σ^2. This combination is referred to as the Jeffrey's Zellner-Siow Cauchy prior or "JZS" in the R `statsr` package.

```r
bayes_inference(y=tthm, data=tapwater, statistic="mean",
                mu_0 = 35, rscale=1, prior="JZS",
                type="ci", method="sim")
```

```
## Single numerical variable
## n = 28, y-bar = 55.5239, s = 23.254
## (Assuming Zellner-Siow Cauchy prior:  mu | sigma^2 ~ C(35, 1*sigma)
## (Assuming improper Jeffreys prior: p(sigma^2) = 1/sigma^2
##
## Posterior Summaries
##             2.5%       25%      50%      75%     97.5%
## mu    45.5713714 51.820910 54.87345 57.87171 64.20477
## sigma 18.4996738 21.810376 23.84572 26.30359 32.11330
## n_0    0.2512834  2.512059  6.13636 12.66747 36.37425
##
## 95% CI for mu: (45.5714, 64.2048)
```
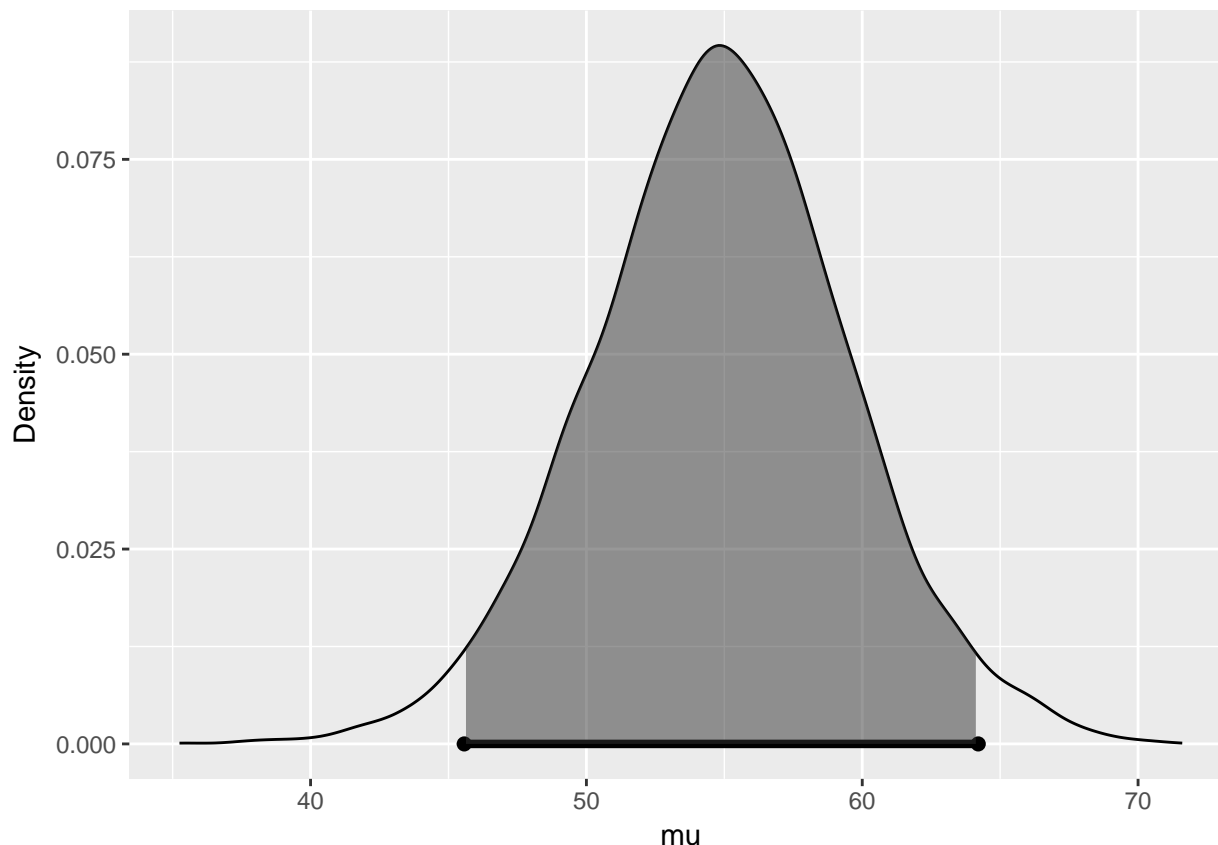
Using the `bayes_inference` function from the `statsr` package, we can obtain summary statistics and a plot from the MCMC output – not only μ, but also inference about σ^2 and the prior sample size.

The posterior mean under the JZS model is much closer to the sample mean than what the normal gamma prior used previously. Under the informative normal gamma prior, the sample made a 55.5, about eight standard deviations above the mean – a surprising value under the normal prior. Under the Cauchy prior, the informative prior location has much less influence.

This is **the robustness property of the Cauchy prior**, leading the posterior to put more weight on the sample mean than the prior mean, especially when the prior location is not close to the sample mean. We can see that the central 50% interval for n_0 is well below the value 25 used in the normal prior, which placed almost equal weight on the prior in sample mean.

Using the MCMC draws of μ, σ, we can obtain Monte Carlo samples from the predictive distribution of y, by plugging μ and σ into the corresponding functions. Figure 4.6 compares the posterior densities estimated from the simulative values of μ and the predicted draws of TTHM under the Jeffrey Zellner-Siow prior, and the informative normal prior from μ with $n_0 = 25$ and the reference prior on σ^2.

To recap, we have shown how to create more flexible prior distributions, such as the Cauchy distribution using mixtures of conjugate priors. As the posterior distributions are not available in closed form, we demonstrated how MCMC can be used for inference using the hierarchical prior distribution. Starting in the late 1980's, MCMC algorithms have led to an exponential rise in the use of Bayes in methods, because complex models built through hierarchical distributions suddenly were tractable. The Cauchy prior is well-known for being robust prior mis-specifications. For example, having a prior mean that is far from the observed mean. This provides an alternative to the reference prior as a default or objective distribution that is proper.

In the next sections, we will return to Bayes factors and hypothesis testing where the Cauchy prior plays an important role.
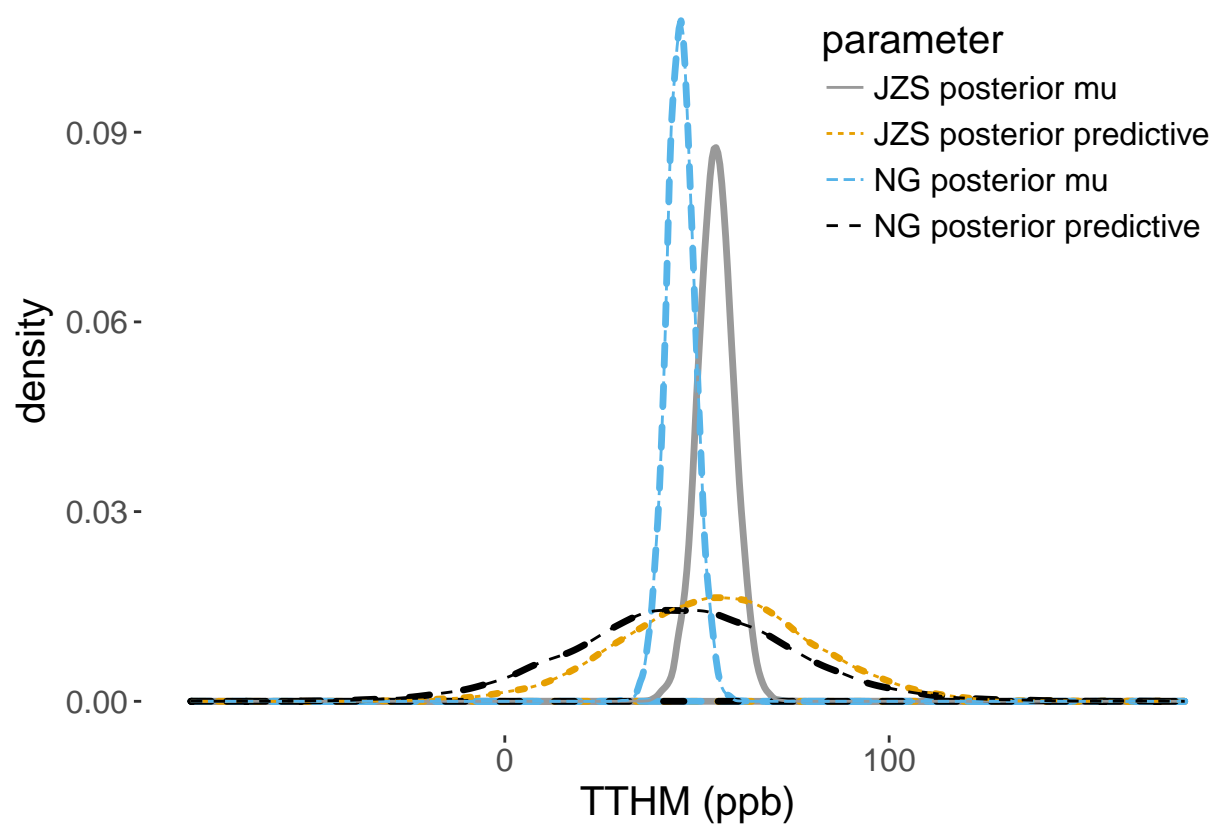
Figure 4.6: Comparison of posterior densities

Chapter 5

Hypothesis Testing with Normal Populations

In Section 3.1.4, we described how the Bayes factors can be used for hypothesis testing. Now we will use the Bayes factors to compare normal means, i.e. test whether the mean of a population is zero or compare two groups of normally-distributed populations. We divide this mission into three cases: known variance for a single population, unknown variance for a single population using paired data, and unknown variance using two independent groups.

5.1   Bayes Factors for Testing a Normal Mean: variance known

Now we show how to obtain base factors for testing hypothesis about a normal mean, where **the variance is known**. To start, let's consider a random sample of observations from a normal population with mean μ and pre-specified variance σ^2. We consider testing whether the population mean μ is equal to m_0 or not.

Therefore, we can formulate the data and hypotheses as below:

Data

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

Hypotheses

- $H_1 : \mu = m_0$
- $H_2 : \mu \neq m_0$

Priors

We also need to specify priors for μ under both hypotheses. Under H_1, we assume that μ is exactly m_0, so this occurs with probability 1 under H_1. Now under H_2, μ is unspecified, so we describe our prior uncertainty with the conjugate normal distribution centered at m_0 and with a variance $\sigma^2/\mathbf{n_0}$. This is centered at the hypothesized value m_0, and it seems that the mean is equally likely to be larger or smaller than m_0, so a dividing factor n_0 is given to the variance. The hyper parameter n_0 controls the precision of the prior as before.

In mathematical terms, the priors are:

- $H_1 : \mu = m_0$ with probability 1
- $H_2 : \mu \sim \mathsf{N}(m_0, \sigma^2/\mathbf{n_0})$

Bayes Factor

Now the Bayes factor for comparing H_1 to H_2 is the ratio of the distribution, the data under the assumption that $\mu = m_0$ to the distribution of the data under H_2.

$$\mathsf{BF}[H_1 : H_2] = \frac{p(\text{data} \mid \mu = m_0, \sigma^2)}{\int p(\text{data} \mid \mu, \sigma^2) p(\mu \mid m_0, \mathbf{n_0}, \sigma^2) \, d\mu}$$

$$\mathsf{BF}[H_1 : H_2] = \left(\frac{n + \mathbf{n_0}}{\mathbf{n_0}}\right)^{1/2} \exp\left\{-\frac{1}{2}\frac{n}{n + \mathbf{n_0}}Z^2\right\}$$

$$Z = \frac{(\bar{Y} - m_0)}{\sigma/\sqrt{n}}$$

The term in the denominator requires integration to account for the uncertainty in μ under H_2. And it can be shown that the Bayes factor is a function of the observed sampled size, the prior sample size n_0 and a Z score.

Let's explore how the hyperparameters in n_0 influences the Bayes factor in Equation (5.1). For illustration we will use the sample size of 100. Recall that for estimation, we interpreted n_0 as a prior sample size and considered the limiting case where n_0 goes to zero as a non-informative or reference prior.

$$\mathsf{BF}[H_1 : H_2] = \left(\frac{n + \mathbf{n_0}}{\mathbf{n_0}}\right)^{1/2} \exp\left\{-\frac{1}{2}\frac{n}{n + \mathbf{n_0}}Z^2\right\} \tag{5.1}$$

Figure 5.1 shows the Bayes factor for comparing H_1 to H_2 on the y-axis as n_0 changes on the x-axis. The different lines correspond to different values of the Z score or how many standard errors $\bar{y}$ is from the hypothesized mean. As expected, larger values of the Z score favor H_2.

But as n_0 becomes smaller and approaches 0, the first term in the Bayes factor goes to infinity, while the exponential term involving the data goes to a constant and is ignored. In the limit as $n_0 \to 0$ under this noninformative prior, the Bayes factor paradoxically ends up favoring H_1 regardless of the value of $\bar{y}$.

The takeaway from this is that we cannot use improper priors with $n_0 = 0$, if we are going to test our hypothesis that $\mu = n_0$. Similarly, vague priors that use a small value of n_0 are not recommended due to the sensitivity of the results to the choice of an arbitrarily small value of n_0.

This problem arises with vague priors – the Bayes factor favors the null model H_1 even when the data are far away from the value under the null – are known as the Bartlett's paradox or the Jeffrey's-Lindleys paradox.

Now one way to try to understand the effect of prior is through the standard effect size

$$\delta = \frac{\mu - m_0}{\sigma}.$$

The prior of the standard effect size is

$$\delta \mid H_2 \sim \mathsf{N}(0, \frac{1}{\mathbf{n_0}})$$

This allows us to think about a standardized effect independent of the units of the problem. One default choice is using the unit information prior, where the prior sample size n_0 is 1, leading to a standard normal for the standardized effect size. This is depicted with the blue normal density in Figure 5.2. This suggested that we expect that the mean will be within ± 1.96 standard deviations of the hypothesized mean **with probability 0.95**. (Note that we can say this only under a Bayesian setting.)

In many fields we expect that the effect will be small relative to σ. If we do not expect to see large effects, then we may want to use a more informative prior on the effect size as the density in orange with $n_0 = 4$. So they expected the mean to be within $\pm 1/\sqrt{n_0}$ or five standard deviations of the prior mean.
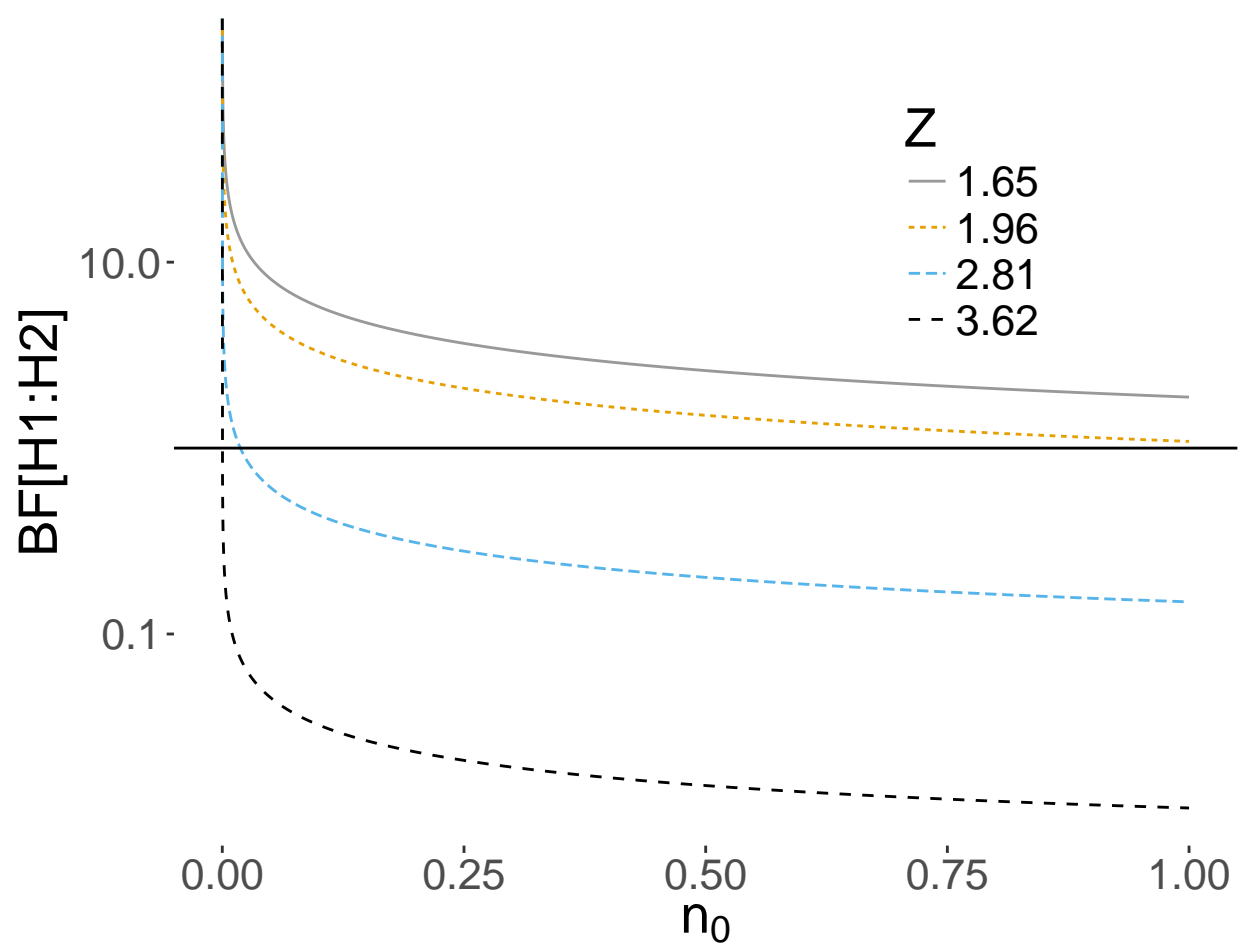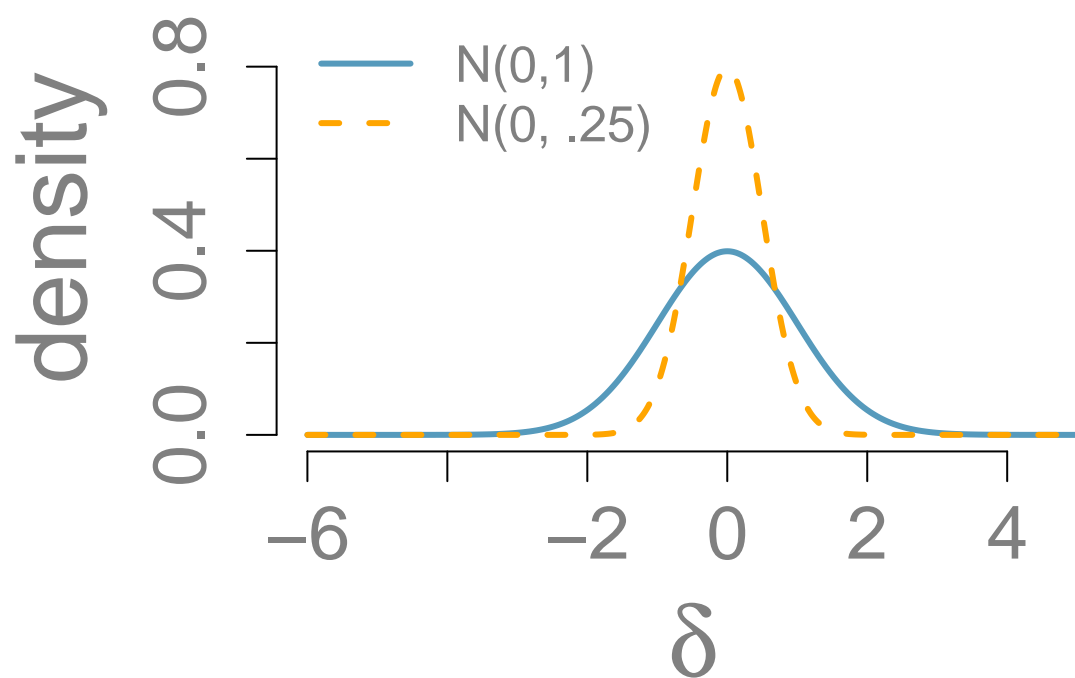
Figure 5.1: Vague prior for mu: n=100

Figure 5.2: Prior on standard effect size

Example 5.1. To illustrate, we give an example from parapsychological research. The case involved the test of the subject's claim to affect a series of randomly generated 0's and 1's by means of extra sensory perception (ESP). The random sequence of 0's and 1's are generated by a machine with probability of generating 1 being 0.5. The subject claims that his ESP would make the sample mean differ significantly from 0.5.

Therefore, we are testing $H_1 : \mu = 0.5$ versus $H_2 : \mu \neq 0.5$. Let's use a prior that suggests we do not expect a large effect which leads the following solution for n_0. Assume we want a standard effect of 0.03, there is a 95% chance that it is between $(-0.03/\sigma, 0.03/\sigma)$, with $n_0 = (1.96\sigma/0.03)^2 = 32.7^2$.

Figure 5.3 shows our informative prior in blue, while the unit information prior is in orange. On this scale, the unit information prior needs to be almost uniform for the range that we are interested.
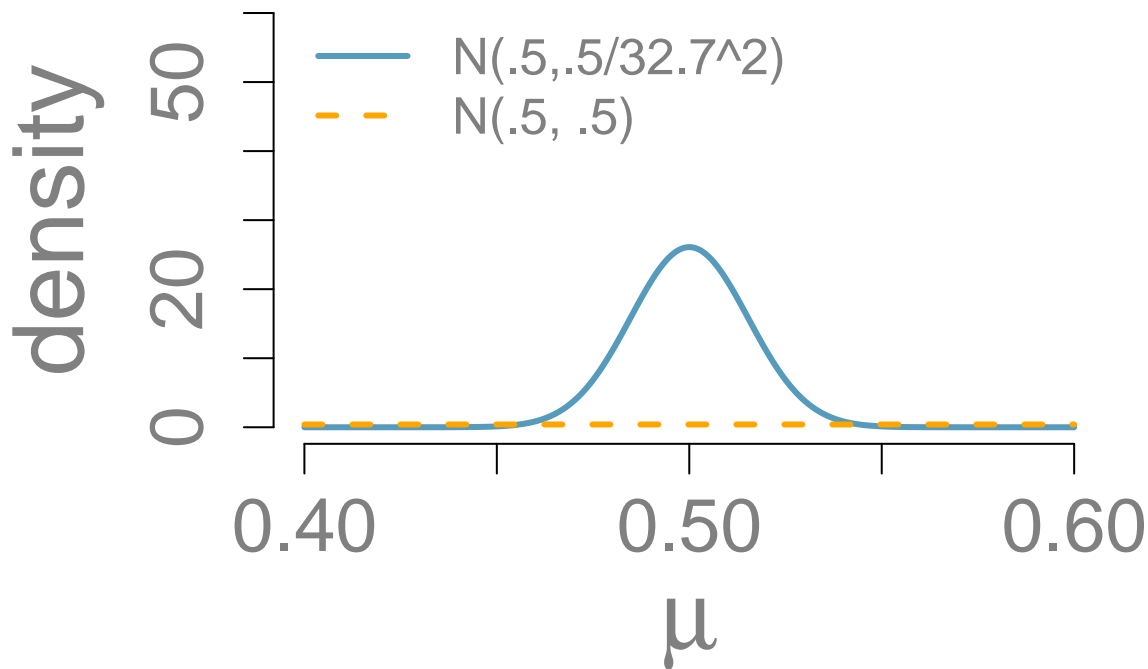
Figure 5.3: Prior effect in the extra sensory perception test

A very large data set with over 104 million trials was collected to test this hypothesis, so we use a normal distribution to approximate the distribution the sample mean.

- Sample size: $n = 1.0449 \times 10^8$
- Sample mean: $\bar{y} = 0.500177$, standard deviation $\sigma = 0.5$
- Z-score: 3.61

Now using our prior in the data, the Bayes factor for H_1 to H_2 was 0.46, implying evidence against the hypothesis H_1 that $\mu = 0.5$.

- Informative $\mathsf{BF}[H_1 : H_2] = 0.46$
- $\mathsf{BF}[H_2 : H_1] = 1/\mathsf{BF}[H_1 : H_2] = 2.19$

Now, this can be inverted to provide the evidence in favor of H_2. The evidence suggests that the hypothesis that the machine operates with a probability that is not 0.5, is 2.19 times more likely than the hypothesis the probability is 0.5. Based on the interpretation of Bayes factors from Table 3.5, this is in the range of "not worth the bare mention".

To recap, we present expressions for calculating Bayes factors for a normal model with a specified variance. We show that the improper reference priors for μ when $n_0 = 0$, or vague priors where n_0 is arbitrarily small, lead to Bayes factors that favor the null hypothesis regardless of the data, and thus should not be used for hypothesis testing.

Bayes factors with normal priors can be sensitive to the choice of the n_0. While the default value of $n_0 = 1$ is reasonable in many cases, this may be too non-informative if one expects more effects. Wherever possible, think about how large an effect you expect and use that information to help select the n_0.

All the ESP examples suggest weak evidence and favored the machine generating random 0's and 1's with a probability that is different from 0.5. Note that ESP is not the only explanation – a deviation from 0.5 can also occur if the random number generator is biased. Bias in the stream of random numbers in our pseudorandom numbers has huge implications for numerous fields that depend on simulation. If the context had been about detecting a small bias in random numbers what prior would you use and how would it change the outcome? You can try to experiment in R or other software packages that generate random Bernoullis.

Next, we will look at Bayes factors in normal models with unknown variances using the Cauchy prior so that results are less sensitive to the choice of n_0.

5.2   Comparing Two Paired Means using Bayes Factors

We previously learned that we can use a paired t-test to compare means from two paired samples. In this section, we will show how Bayes factors can be expressed as a function of the t-statistic for comparing the means and provide posterior probabilities of the hypothesis that whether the means are equal or different.

Example 5.2. Trace metals in drinking water affect the flavor, and unusually high concentrations can pose a health hazard. Ten pairs of data were taken measuring the zinc concentration in bottom and surface water at ten randomly sampled locations, as listed in Table 5.1.

Water samples collected at the the same location, on the surface and the bottom, cannot be assumed to be independent of each other. However, it may be reasonable to assume that the differences in the concentration at the bottom and the surface in randomly sampled locations are independent of each other.

```
## Loading required package: BayesFactor

## Loading required package: coda

## Loading required package: Matrix


## ************
## Welcome to BayesFactor 0.9.12-2. If you have questions, please contact Richard Morey (richarddmorey@g
##
## Type BFManual() to open the manual.
## ************
```

To start modeling, we will treat the ten differences as a random sample from a normal population where the parameter of interest is the difference between the average zinc concentration at the bottom and the average zinc concentration at the surface, or the main difference, μ.

In mathematical terms, we have

Table 5.1: Zinc in drinking water

location	bottom	surface	difference
1	0.430	0.415	0.015
2	0.266	0.238	0.028
3	0.567	0.390	0.177
4	0.531	0.410	0.121
5	0.707	0.605	0.102
6	0.716	0.609	0.107
7	0.651	0.632	0.019
8	0.589	0.523	0.066
9	0.469	0.411	0.058
10	0.723	0.612	0.111

- Random sample of $n = 10$ differences $Y_1, \ldots, Y_n$
- Normal population with mean $\mu \equiv \mu_B - \mu_S$

In this case, we have no information about the variability in the data, and we will treat the variance, σ^2, as unknown.

The hypothesis of the main concentration at the surface and bottom are the same is equivalent to saying $\mu = 0$. The second hypothesis is that the difference between the mean bottom and surface concentrations, or equivalently that the mean difference $\mu \neq 0$.

In other words, we are going to compare the following hypotheses:

- $H_1 : \mu_B = \mu_S \Leftrightarrow \mu = 0$
- $H_2 : \mu_B \neq \mu_S \Leftrightarrow \mu \neq 0$

The Bayes factor is the ratio between the distributions of the data under each hypothesis, which does not depend on any unknown parameters.

$$\mathsf{BF}[H_1 : H_2] = \frac{p(\text{data} \mid H_1)}{p(\text{data} \mid H_2)}$$

To obtain the Bayes factor, we need to use integration over the prior distributions under each hypothesis to obtain those distributions of the data.

$$\mathsf{BF}[H_1 : H_2] = \iint p(\text{data} \mid \mu, \sigma^2) p(\mu \mid \sigma^2) p(\sigma^2 \mid H_2) \, d\mu \, d\sigma^2$$

This requires specifying the following priors:

- $\mu \mid \sigma^2, H_2 \sim \mathsf{N}(0, \sigma^2/n_0)$
- $p(\sigma^2) \propto 1/\sigma^2$ for both H_1 and H_2

μ is exactly zero under the hypothesis H_1. For μ in H_2, we start with the same conjugate normal prior as we used in Section 5.1 – testing the normal mean with known variance. Since we assume that σ^2 is known, we model $\mu \mid \sigma^2$ instead of μ itself.

The σ^2 appears in both the numerator and denominator of the Bayes factor. For default or reference case, we use the Jeffreys prior (a.k.a. reference prior) on σ^2. As long as we have more than two observations, this (improper) prior will lead to a proper posterior.

After integration and rearranging, one can derive a simple expression for the Bayes factor:

$$\mathsf{BF}[H_1 : H_2] = \left(\frac{n + n_0}{n_0}\right)^{1/2} \left(\frac{t^2 \frac{n_0}{n+n_0} + \nu}{t^2 + \nu}\right)^{\frac{\nu+1}{2}}$$

This is a function of the t-statistic

$$t = \frac{|\bar{Y}|}{s/\sqrt{n}}$$

,

where s is the sample standard deviation and the degrees of freedom $\nu = n - 1$ (sample size minus one).

As we saw in the case of Bayes factors with known variance, we cannot use the improper prior on μ because when $n_0 \to 0$, then $\mathsf{BF}[H1 : H_2] \to \infty$ favoring H_1 regardless of the magnitude of the t-statistic. Arbitrary, vague small choices for n_0 also lead to arbitrary large Bayes factors in favor of H_1. Another example of the Barlett's or Jeffreys-Lindley paradox.

Sir Herald Jeffrey discovered another paradox testing using the conjugant normal prior, known as the **information paradox**. His thought experiment assumed that our sample size n and the prior sample size n_0. He then considered what would happen to the Bayes factor as the sample mean moved further and further away from the hypothesized mean, measured in terms standard errors with the t-statistic, i.e., $|t| \to \infty$. As the t-statistic or information about the mean moved further and further from zero, the Bayes factor goes to a constant depending on n, n_0 rather than providing overwhelming support for H_2.

The bounded Bayes factor is

$$\mathsf{BF}[H_1 : H_2] \to \left(\frac{n_0}{n_0 + n}\right)^{\frac{n-1}{2}}$$

Jeffrey wanted a prior with $\mathsf{BF}[H_1 : H_2] \to 0$ (or equivalently, $\mathsf{BF}[H_2 : H_1] \to \infty$), as the information from the t-statistic grows, indicating the sample mean is as far as from the hypothesized mean and should favor H_2.

To resolve the paradox when the information the t-statistic favors H_2 but the Bayes factor does not, Jeffreys showed that **no normal prior could resolve the paradox**.

But a **Cauchy prior** on μ, would resolve it. In this way, $\mathsf{BF}[H_2 : H_1]$ goes to infinity as the sample mean becomes further away from the hypothesized mean. Recall that the Cauchy prior is written as $\mathsf{C}(0, r^2\sigma^2)$. While Jeffreys used a default of $r = 1$, smaller values of r can be used if smaller effects are expected.

The combination of the Jeffrey's prior on σ^2 and this Cauchy prior on μ under H_2 is sometimes referred to as the **Jeffrey-Zellener-Siow prior**.

However, there is no closed form expressions for the Bayes factor under the Cauchy distribution. To obtain the Bayes factor, we must use the numerical integration or simulation methods.

We will use the `bayes_inference` function from the **statsr** package to test whether the mean difference is zero in Example 5.2 (zinc), using the JZS (Jeffreys-Zellener-Siow) prior.

```
library(statsr)
bayes_inference(difference, data=zinc, statistic="mean", type="ht",
               prior="JZS", mu_0=0, method="theo", alt="twosided")
```

```
## Single numerical variable
## n = 10, y-bar = 0.0804, s = 0.0523
```

```
## (Using Zellner-Siow Cauchy prior:  mu ~ C(0, 1*sigma)
## (Using Jeffreys prior: p(sigma^2) = 1/sigma^2
##
## Hypotheses:
## H1: mu = 0 versus H2: mu != 0
## Priors:
## P(H1) = 0.5 , P(H2) = 0.5
## Results:
## BF[H2:H1] = 50.7757
## P(H1|data) = 0.0193  P(H2|data) = 0.9807
##
## Posterior summaries for mu under H2:
## Single numerical variable
## n = 10, y-bar = 0.0804, s = 0.0523
## (Assuming Zellner-Siow Cauchy prior:  mu | sigma^2 ~ C(0, 1*sigma)
## (Assuming improper Jeffreys prior: p(sigma^2) = 1/sigma^2
##
## Posterior Summaries
##               2.5%         25%         50%          75%         97.5%
## mu     0.03651159 0.06335608 0.07541097  0.08720882  0.11231749
## sigma 0.03667844 0.04740997 0.05532057  0.06554686  0.09552602
## n_0    0.16159352 1.88750112 4.73144753 10.08484315 32.39528910
##
## 95% CI for mu: (0.0365, 0.1123)
```
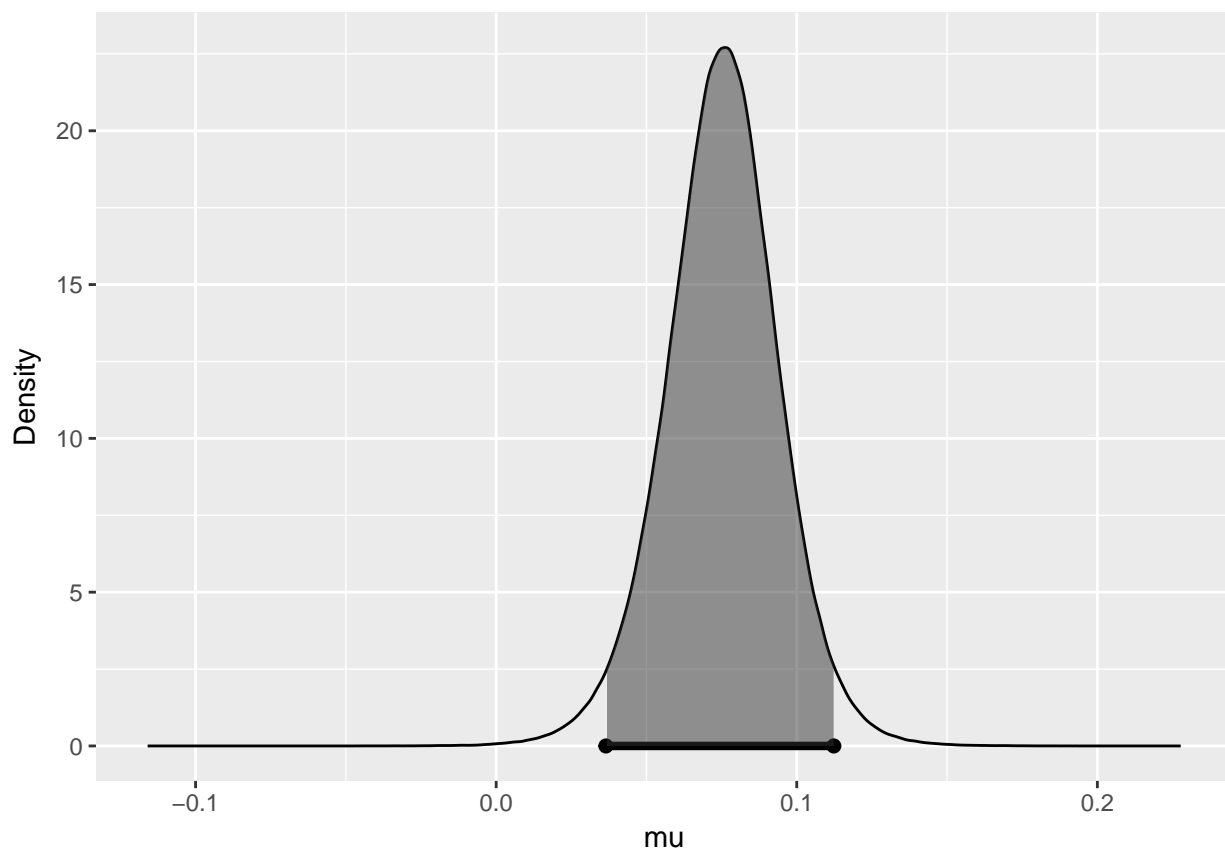

With equal prior probabilities on the two hypothesis, the Bayes factor is the posterior odds. From the output,

we see this indicates that the hypothesis H_2, the mean difference is different from 0, is almost 51 times more likely than the hypothesis H_1 that the average concentration is the same at the surface and the bottom.

To sum up, we have used the **Cauchy prior** as a default prior testing hypothesis about a normal mean when variances are unknown. This does require numerical integration, but it is available in the `bayes_inference` function from the `statsr` package. If you expect that the effect sizes will be small, smaller values of r are recommended.

It is often important to quantify the magnitude of the difference in addition to testing. The Cauchy Prior provides a default prior for both testing and inference; it avoids problems that arise with choosing a value of n_0 (prior sample size) in both cases.

Next, we will illustrate using the Cauchy prior for comparing two means from independent normal samples.

5.3   Comparing Independent Means: hypothesis testing

In the previous section, we described Bayes factors for testing whether the mean difference of **paired** samples was zero. In this section, we will consider a slightly different problem – we have two **independent** samples, and we would like to test the hypothesis that the means are different or equal.

Example 5.3. We illustrate the testing of independent groups with data from a 2004 survey of birth records from North Carolina, which are available in the `statsr` package.

The variable of interest is `gained` – the weight gain of mothers during pregnancy. We have two groups defined by the categorical variable, `mature`, with levels, younger mom and older mom.

Question of interest: Do the data provide convincing evidence of a difference between the average weight gain of older moms and the average weight gain of younger moms?

We will view the data as a random sample from two populations, older and younger moms. The two groups are modeled as:

$$
\begin{aligned}
Y_{O,i} &\overset{\text{iid}}{\sim} \mathsf{N}(\mu + \alpha/2, \sigma^2) \\
Y_{Y,i} &\overset{\text{iid}}{\sim} \mathsf{N}(\mu - \alpha/2, \sigma^2)
\end{aligned}
\tag{5.2}
$$

The model for weight gain for older moms using the subscript O, and it assumes that the observations Y are independent and identically distributed, with a mean $\mu + \alpha/2$ and variance σ^2.

For the younger women, the observations with the subscript Y are independent and identically distributed with a mean $\mu - \alpha/2$ and variance σ^2.

Using this representation of the means in the two groups, the difference in means simplifies to α – the parameter of interest.

$$
(\mu + \alpha/2) - (\mu - \alpha/2) = \alpha
$$

You may ask, "Why don't we set the average weight gain of older women to $\mu + \alpha$, and the average weight gain of younger women to μ?" We need the parameter α to be present in both $Y_{O,i}$ (the group of older women) and $Y_{Y,i}$ (the group of younger women).

We have the following competing hypotheses:

- $H_1 : \alpha = 0 \Leftrightarrow$ The means are not different.
- $H_2 : \alpha \neq 0 \Leftrightarrow$ The means are different.

In this representation, μ represents the overall weight gain for all women. (Does the model in Equation (5.2) make more sense now?) To test the hypothesis, we need to specify prior distributions for α under H_2 (c.f. $\alpha = 0$ under H_1) and priors for μ, σ^2 under both hypotheses.

Recall that the Bayes factor is the ratio of the distribution of the data under the two hypotheses.

$$\mathsf{BF}[H_1 : H_2] = \frac{p(\text{data} \mid H_1)}{p(\text{data} \mid H_2)}$$

$$= \frac{\iint p(\text{data} \mid \alpha = 0, \mu, \sigma^2) p(\mu, \sigma^2 \mid H_1) \, d\mu \, d\sigma^2}{\int \iint p(\text{data} \mid \alpha, \mu, \sigma^2) p(\alpha \mid \sigma^2) p(\mu, \sigma^2 \mid H_2) \, d\mu \, d\sigma^2 \, d\alpha}$$

As before, we need to average over uncertainty and the parameters to obtain the unconditional distribution of the data. Now, as in the test about a single mean, we cannot use improper or non-informative priors for α for testing.

Under H_2, we use the Cauchy prior for α, or equivalently, the Cauchy prior on the standardized effect δ with the scale of r:

$$\delta = \alpha/\sigma^2 \sim \mathsf{C}(0, r^2)$$

Now, under both H_1 and H_2, we use the Jeffrey's reference prior on μ and σ^2:

$$p(\mu, \sigma^2) \propto 1/\sigma^2$$

While this is an improper prior on μ, this does not suffer from the Bartlett's-Lindley's-Jeffreys' paradox as μ is a common parameter in the model in H_1 and H_2. This is another example of the Jeffreys-Zellner-Siow prior.

As in the single mean case, we will need numerical algorithms to obtain the Bayes factor. Now the following output illustrates testing of Bayes factors, using the Bayes inference function from the `statsr` package.

```r
library(statsr)
data(nc)
bayes_inference(y=gained, x=mature, data=nc,type='ht',
                statistic='mean',  alternative='twosided', null=0,
                prior='JZS', r=1, method='theo', show_summ=FALSE)
```

```
## Hypotheses:
## H1: mu_mature mom  = mu_younger mom
## H2: mu_mature mom != mu_younger mom
##
## Priors: P(H1) = 0.5  P(H2) = 0.5
##
## Results:
## BF[H1:H2] = 5.7162
## P(H1|data) = 0.8511
## P(H2|data) = 0.1489
##
## Posterior summaries for under H2:
## 95% Cred. Int.: (-4.3828 , 0.8451)
```
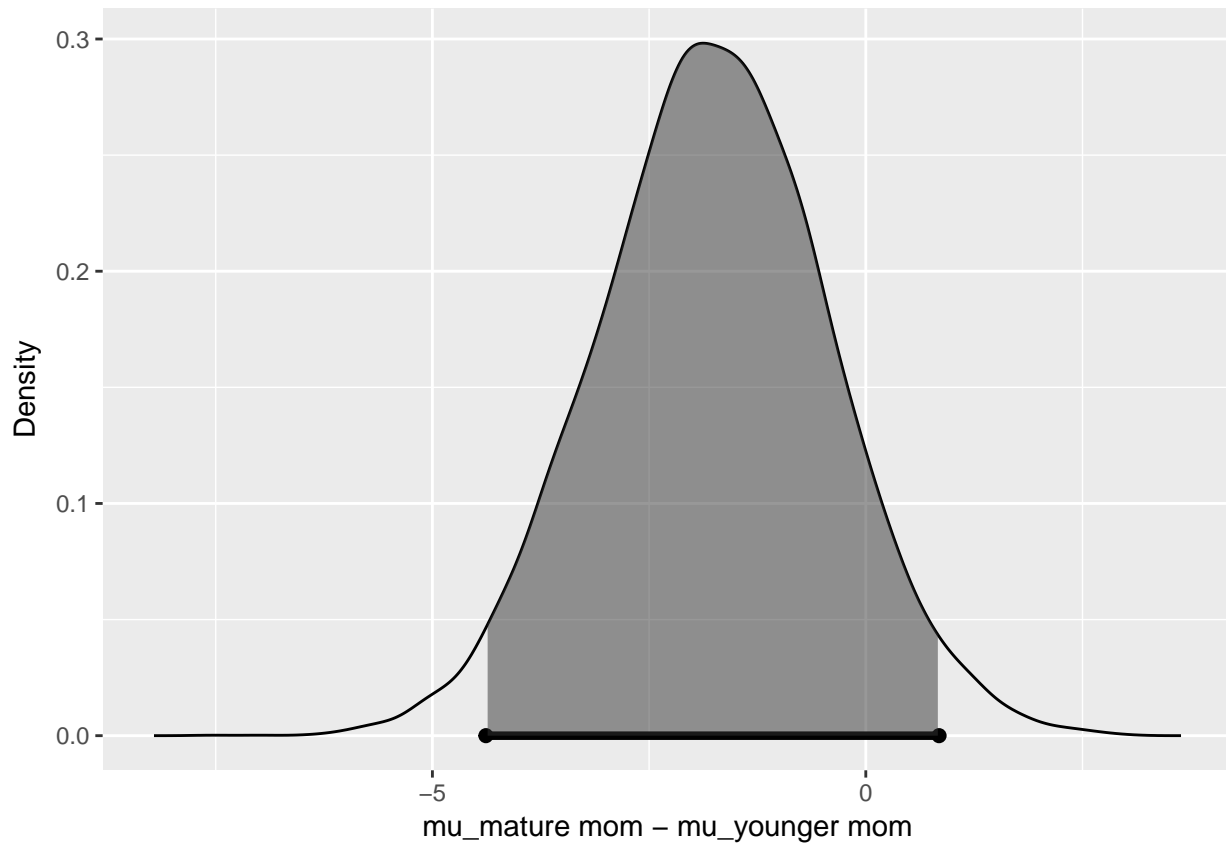
We see that the Bayes factor for H_1 to H_2 is about 5.7, with positive support for H_1 that there is no difference in average weight gain between younger and older women. Using equal prior probabilities, the probability that there is a difference in average weight gain between the two groups is about 0.15 given the data. Based on the interpretation of Bayes factors from Table 3.5, this is in the range of "positive" (between 3 and 20).

To recap, we have illustrated testing hypotheses about population means with two independent samples, using a Cauchy prior on the difference in the means. One assumption that we have made is that **the variances are equal in both groups**. The case where the variances are unequal is referred to as the Behren-Fisher problem, and this is beyond the scope for this course. Next, we will look at another example to put everything together with testing and discuss summarizing results.

5.4   Inference after Testing

TBA

Chapter 6

Introduction to Bayesian Regression

In the previous chapter, we introduced Bayesian decision making using posterior probabilities and a variety of loss functions. We discussed how to minimize the expected loss for hypothesis testing. Moreover, we instroduce the concept of Bayes factors and gave some examples of how they can be used in Bayesian hypothesis testing for comparison of two means. We also discussed how to choose appropriate and robust priors. When there is no conjugacy, we applied Markov Chain Monte Carlo simulation to approximate the posterior distributions of parameters of interest.

In this chapter, we will apply Bayesian inference to linear regression. We will first apply Bayesian statistics to simple linear regression models, then generalize the results to multiple linear regression models. We will see when using the reference prior, the posterior means, posterior standard deviations, and credible intervals of the coefficients will coincide with the counterparts in the frequentist ordinary least square (OLS) linear regression models. However, using the Bayes framework, we can now interpret the credible intervals as the probability of the coefficients lie in such intervals.

6.1 Bayesian Simple Linear Regression

In this section, we turn to Bayesian inference in simple linear regression. We will use reference prior distribution which will provide a connection between the frequentist solution and Bayesian answers. This provides a baseline analysis for comparions with more informative prior distributions. To illustrate the ideas, we will use an example of predicting body fat.

6.1.1 Frequentist Ordinary Least Square Simple Linear Regression

Obtaining accurate measurements of body fat is expensive and not easy to be done. Instaed, predictive models which predict the percentage of body fat using readily available measurements such as abdominal circumference are easy to use and inexpensive. We will illustrate this using the `bodyfat` data from the library `BAS`.

To start, we load the `BAS` library (you may download the package from CRAN) to access the dataframe. We print out a summary of the variables in this dataframe.

```
# Suppress warning messages
library(BAS)
data(bodyfat)
summary(bodyfat)
```

```
##      Density          Bodyfat            Age             Weight
##   Min.   :0.995   Min.   : 0.00   Min.   :22.00   Min.   :118.5
##   1st Qu.:1.041   1st Qu.:12.47   1st Qu.:35.75   1st Qu.:159.0
##   Median :1.055   Median :19.20   Median :43.00   Median :176.5
##   Mean   :1.056   Mean   :19.15   Mean   :44.88   Mean   :178.9
##   3rd Qu.:1.070   3rd Qu.:25.30   3rd Qu.:54.00   3rd Qu.:197.0
##   Max.   :1.109   Max.   :47.50   Max.   :81.00   Max.   :363.1
##       Height           Neck            Chest           Abdomen
##   Min.   :29.50   Min.   :31.10   Min.   : 79.30   Min.   : 69.40
##   1st Qu.:68.25   1st Qu.:36.40   1st Qu.: 94.35   1st Qu.: 84.58
##   Median :70.00   Median :38.00   Median : 99.65   Median : 90.95
##   Mean   :70.15   Mean   :37.99   Mean   :100.82   Mean   : 92.56
##   3rd Qu.:72.25   3rd Qu.:39.42   3rd Qu.:105.38   3rd Qu.: 99.33
##   Max.   :77.75   Max.   :51.20   Max.   :136.20   Max.   :148.10
##        Hip            Thigh            Knee            Ankle
##   Min.   : 85.0   Min.   :47.20   Min.   :33.00   Min.   :19.1
##   1st Qu.: 95.5   1st Qu.:56.00   1st Qu.:36.98   1st Qu.:22.0
##   Median : 99.3   Median :59.00   Median :38.50   Median :22.8
##   Mean   : 99.9   Mean   :59.41   Mean   :38.59   Mean   :23.1
##   3rd Qu.:103.5   3rd Qu.:62.35   3rd Qu.:39.92   3rd Qu.:24.0
##   Max.   :147.7   Max.   :87.30   Max.   :49.10   Max.   :33.9
##       Biceps          Forearm          Wrist
##   Min.   :24.80   Min.   :21.00   Min.   :15.80
##   1st Qu.:30.20   1st Qu.:27.30   1st Qu.:17.60
##   Median :32.05   Median :28.70   Median :18.30
##   Mean   :32.27   Mean   :28.66   Mean   :18.23
##   3rd Qu.:34.33   3rd Qu.:30.00   3rd Qu.:18.80
##   Max.   :45.00   Max.   :34.90   Max.   :21.40
```

This dataframe includes 252 measurements on men of body fat and other measurements, such as waist circumference (`Abdomen`). We will use `Abdomen` to illustrate Bayesian simple linear regression. We regress the response variable `Bodyfat` on the predictor `Abdomen`, which gives us the model

$$y_i = \alpha + \beta x_i + \epsilon_i,$$

which the assumption that the errors ϵ_i are independent and identically distributed as normal random variables with mean zero and constant variance σ^2.

The figure below shows the percentage body fat obtained from under water weighing and the abdominal circumference for 252 men. To predict body fat, the line overlayed on the scatter plot illustrates the best fitting ordinary least squares line obtained with the `lm` function in R.

```r
plot(Bodyfat ~ Abdomen, data = bodyfat,
     xlab = "abdomen circumference (cm)",
     col = "blue", pch = 16, main = "")

# Ordinary least square linear regression
bodyfat.lm = lm(Bodyfat ~ Abdomen, data = bodyfat)
summary(bodyfat.lm)
```

```
##
## Call:
## lm(formula = Bodyfat ~ Abdomen, data = bodyfat)
##
```

```
## Residuals:
##      Min       1Q    Median       3Q       Max
## -19.0160  -3.7557    0.0554   3.4215   12.9007
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -39.28018    2.66034  -14.77   <2e-16 ***
## Abdomen       0.63130    0.02855   22.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.877 on 250 degrees of freedom
## Multiple R-squared:  0.6617, Adjusted R-squared:  0.6603
## F-statistic: 488.9 on 1 and 250 DF,  p-value: < 2.2e-16
```

```
beta = coef(bodyfat.lm)
abline(beta, lwd = 4, col = 1)
```
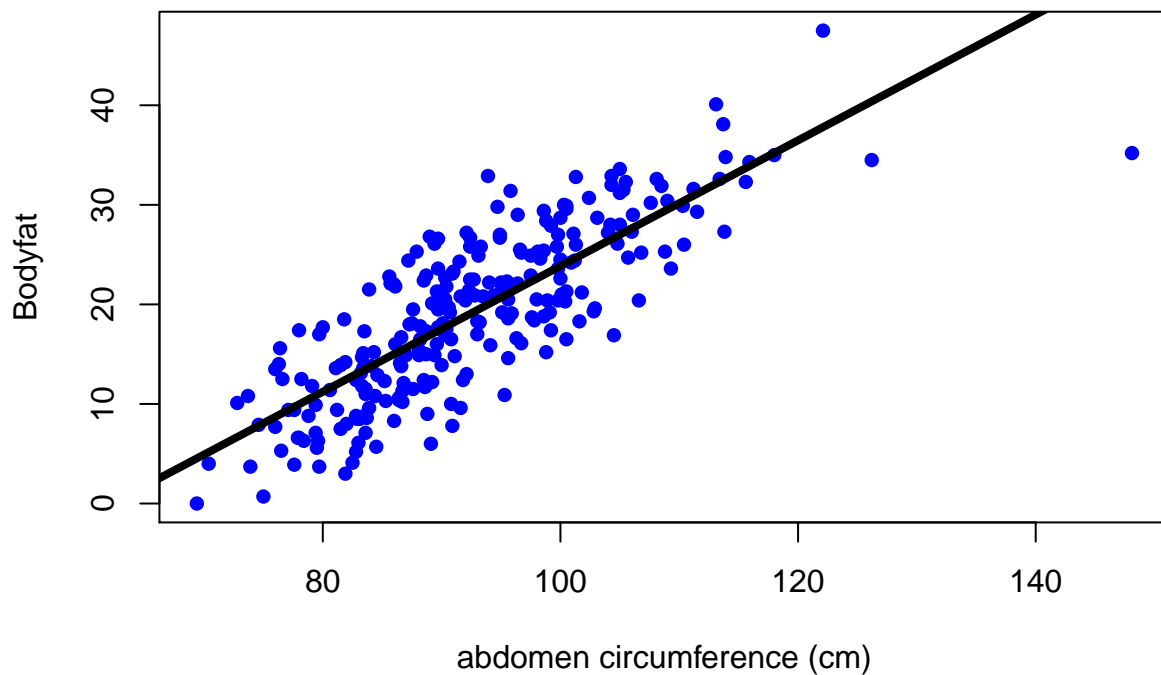

From the summary, we see that this model has an estimated slope, $\hat{\beta}$, of 0.63 and an estimated intercept, $\hat{\alpha}$, of about -39.28%. For every additional centimeter, we expect body fat to increase by 0.63%. The negative interceptive course does not make sense as a physical model, but neither does predicting a male with a waist of zero centimeters. Nevertheless, this linear regression may be an accurate approximation for prediction purposes for measurements that are in the observed range for this population.

The residuals, which provide an estimate of the fitting error, are equal to $\hat{\epsilon}_i = Y_i - \hat{Y}_i$, the difference between the observed values Y_i and the fited values $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$, where X_i is the abdominal circumference for the

ith male. $\hat{\epsilon}_i$ are used for diagnostics as well as estimating the constant variance in the assumption of the model σ^2 via the mean squared error (MSE):

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_i^n (y_i - \hat{y}_i)^2 = \frac{1}{n-2} \sum_i^n \hat{\epsilon}_i^2.$$

Here the degrees of freedom $n - 2$ are the number of observations adjusted for the number of parameters that we estimated in the regression. The MSE, $\hat{\sigma}^2$, may be obtained from the output as the square of the entry labeled "residual standard error".

Since residuals and fitted values are uncorrelated with the expected value of the residuals equal to zero if the model is correct, the scatterplot of residuals versus fitted values provides an additional visual check of the model adequacy.

```
plot(residuals(bodyfat.lm) ~ fitted(bodyfat.lm))
abline(h = 0)
```
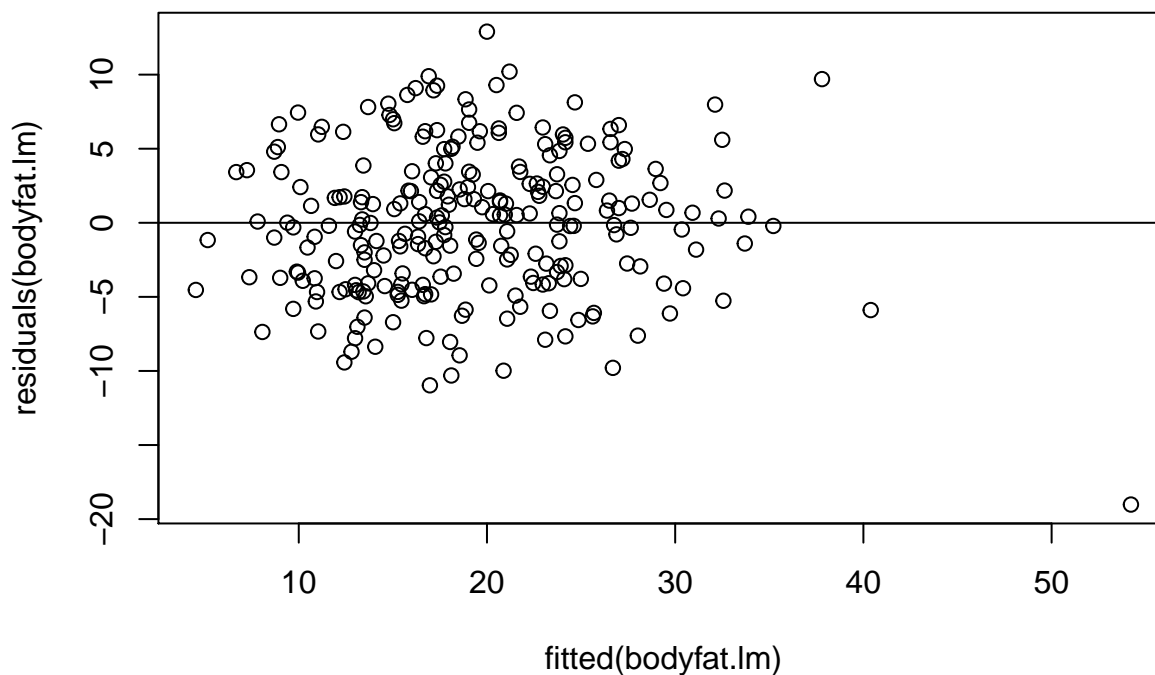

With the exception of the one observation for the individual with the largest waist measurement, the residual plot suggests that the linear regression is a reasonable approximation.

Furthermore, we can check the normal probability plot of the residuals for the assumption of normally distributed errors. We see that only observation 39 is exceptionally away from the normal quantile.

```
plot(bodyfat.lm, which = 2)
```
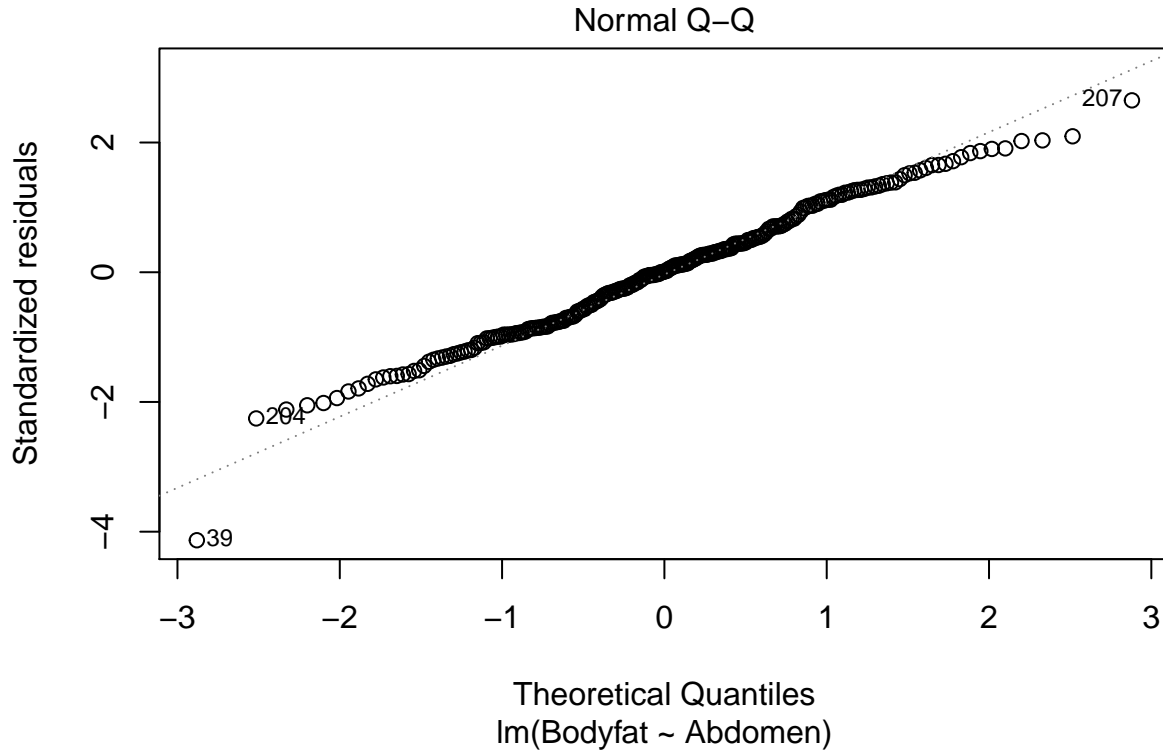
The confidence interval of α and β can be constructed using the standard errors $\mathrm{sd}_{\hat{\alpha}}$ and $\mathrm{sd}_{\hat{\beta}}$ respectively. To proceed, we introduce notations of some "sums of squares"

$$S_{xx} = \sum_{i}^{n}(x_i - \bar{x})^2$$

$$S_{yy} = \sum_{i}^{n}(y_i - \bar{x})^2$$

$$S_{xy} = \sum_{i}^{n}(x_i - barx)(y_i - \bar{y})$$

$$SSE = \sum_{i}^{n}(y_i - \hat{y}_i)^2 = \sum_{i}^{n}\hat{\epsilon}_i^2.$$

$\hat{\alpha}$ and $\hat{\beta}$ can be calculated using these "sums of squares"

$$\hat{\beta} = \frac{\sum_i(x_i - \bar{x})(y_i - \bar{y})}{\sum_i(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x}$$

The last "sum of square" is the *sum of squars of errors*. Its sample mean is exactly the mean squared error (MSE) we introduced previously

$$\hat{\sigma}^2 = \frac{SSE}{n - 2} = MSE$$

$\text{sd}_{\hat{\alpha}}$ and $\text{sd}_{\hat{\beta}}$ are given as

$$\text{sd}_{\hat{\alpha}} = \sqrt{\frac{\text{SSE}}{n-2}\left(\frac{1}{n} + \frac{\bar{x}^2}{\text{S}_{xx}}\right)} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\text{S}_{xx}}}$$

$$\text{sd}_{\hat{\beta}} = \sqrt{\frac{\text{SSE}}{n-2}\frac{1}{\text{S}_x x}} = \frac{\hat{\sigma}}{\sqrt{\text{S}_{xx}}}$$

We may construct the confidence intervals of α and β using the t-statistics

$$t_{\alpha}^* = \frac{\alpha - \hat{\alpha}}{\text{sd}_{\hat{\alpha}}}, \qquad t_{\beta}^* = \frac{\beta - \hat{\beta}}{\text{sd}_{\hat{\beta}}}$$

They both have degrees of freedom $n-2$.

6.1.2 Bayesian Simple Linear Regression Using Reference Prior

Let us now turn to the Bayesian version and show how to obtain the posterior distributions of α and β under the reference prior.

The Bayesian model starts with the same model as the classical frequentist approach:

$$y_i = \alpha + \beta x_i + \epsilon_i,$$

with the assumption that the errors, ϵ_i, are independent and identically distributed as normal random variables with mean zero and constant variance σ^2. This assumption is exactly the same as the classical inference for testing and constructing confidence intervals for α and β.

Our goal is to update the distributions of the unknown parameters α, β, and σ^2, based on the data $x_1, y_1, \cdots, x_n, y_n$, where n is the number of observations.

Under the assumption that the errors ϵ_i are normally distributed with constant variance σ^2, we have for each response y_i, conditioning on the observed data x_i and the parameters α, β, σ^2, is normally distributed:

$$y_i \mid x_i, \alpha, \beta, \sigma^2 \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2), \qquad i = 1, \cdots, n.$$

That is, the likelihood of y_i given x_i, α, β, and σ^2 is

$$\pi(y_i \mid x_i, \alpha, \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - (\alpha + \beta x_i))^2}{2\sigma^2}\right).$$

We will first consider the standard noninformative prior (reference prior). Using the reference prior, we will obtain familiar distributions as the posterior distributions of α, β, and σ^2, which gives the analogue to the frequentist results. Here we assume

$$\pi(\alpha, \beta, \sigma^2) \propto \frac{1}{\sigma^2}.$$

Using the hierachical model framework, we may equivalently assume that the joint prior distribution of α and β under σ^2 is the uniform prior, while the prior distribution of σ^2 is proportional to $\frac{1}{\sigma^2}$. That is

$$\pi(\alpha, \beta \mid \sigma^2) \propto 1, \qquad \pi(\sigma^2) \propto \frac{1}{\sigma^2},$$

Combining the two using conditional probability, we will get the same joint prior distribution.

Then we apply the Bayes' rule to derive the posterior joint distribution after observing $y_1, \cdots, y_n$:

$$\pi^*(\alpha, \beta, \sigma^2 \mid y_1, \cdots, y_n) \propto \left[\prod_i^n \pi(y_i \mid x_i, \alpha, \beta, \sigma^2)\right] \pi(\alpha, \beta, \sigma^2)$$

$$\propto \left[\left(\frac{1}{(\sigma^2)^{1/2}} \exp\left(-\frac{(y_1 - (\alpha + \beta x_1))^2}{2\sigma^2}\right)\right) \times \cdots\right.$$

$$\left. \times \left(\frac{1}{(\sigma^2)^{1/2}} \exp\left(-\frac{(y_n - (\alpha + \beta x_n))^2}{2\sigma^2}\right)\right)\right] \times \left(\frac{1}{\sigma^2}\right)$$

$$\propto \frac{1}{(\sigma^2)^{(n+2)/2}} \exp\left(-\frac{\sum_i (y_i - \alpha - \beta x_i)^2}{2\sigma^2}\right)$$

For example, to obtain the marginal distribution of β, the slope of the linear regression model, we integrate α and σ^2 out:

$$\pi^*(\beta \mid y_1, \cdots, y_n) = \int_0^\infty \left(\int_{-\infty}^\infty \pi^*(\alpha, \beta, \sigma^2 \mid y_1, \cdots, y_n)\, d\alpha\right) d\sigma^2.$$

It can be shown that the marginal posterior distribution of β is the Student's t-distribution

$$\beta \mid y_1, \cdots, y_n \sim t_{n-2}\left(\hat{\beta}, \frac{\hat{\sigma}^2}{S_{xx}}\right) = t_{n-2}\left(\hat{\beta}, (sd_{\hat{\beta}})^2\right)$$

with degrees of freedom $n - 2$, center at $\hat{\beta}$, the coefficient we obtain from the frequentist OLS model, and scale parameter $\dfrac{\hat{\sigma}^2}{S_{xx}} = \left(sd_{\hat{\beta}}\right)^2$, which is the square of the standard error of $\hat{\beta}$ under the frequentist OLS model.

Similarly, we can integrate out β and σ^2 from the joint posterior distribution to get the marginal posterior distribution of α, $\pi^*(\alpha \mid y_1, \cdots, y_n)$. It turns out that $\pi^*(\alpha \mid y_1, \cdots, y_n)$ is again the Student's t-distribution with degrees of freedom $n - 2$, center at $\hat{\alpha}$, the y-intercept from the frequentist OLS model, and scale parameter $\hat{\sigma}^2\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right) = (sd_{\hat{\alpha}})^2$, which is the square of the standard error of $\hat{\alpha}$ under the frequentist OLS model.

Moreover, we can show that the marginal posterior distribution of σ^2 is the inverse Gamma distribution

$$\sigma^2 \mid y_1, \cdots, y_n \sim IG\left(\frac{n-2}{2}, \frac{SSE}{2}\right).$$

We provide the detailed deviations in section ??? (how to do reference in R Markdown?)

Credible Intervals for the Slope β and the Intercept α

The Bayesian posterior distribution results of α and β show that under the reference prior, the posterior credible intervals are in fact numerically equivalent to the confidence intervals from the classical frequentist analysis, providing a baseline for other Bayesian analyses with more informative prior distributions or perhaps other "objective" prior distributions, such as a Student's t prior with 1 degree of freedome which is known as the Cauchy distribution.

We can use the `lm` function to obtain the OLS estimates and construct the credible intervals of α and β

```
output = summary(bodyfat.lm)$coef[, 1:2]
output
```

```
##               Estimate  Std. Error
## (Intercept) -39.2801847 2.66033696
## Abdomen       0.6313044 0.02855067
```

The columns labeled `Estimate` and `Std. Error` are equivalent to the centers (or posterior means) and scale parameters (or standard deviations) in the two t-distributions respetively. The credible intervals of α and β are the same as the frequentist ocnfidence intervals, but we now can interpret them from a Bayesian perspective.

The `confint` function provides 95% confidence intervals, which under the reference prior are equivalent to the 95% credible intervals. The code below extracts them and relabels the output as the Bayesian results.

```
out = cbind(output, confint(bodyfat.lm))
colnames(out) = c("posterior mean", "posterior std", "2.5", "97.5")
round(out, 2)
```

```
##            posterior mean posterior std   2.5   97.5
## (Intercept)        -39.28          2.66 -44.52 -34.04
## Abdomen              0.63          0.03   0.58   0.69
```

These intervals coincide with the confidence intervals from the frequentist approach. The primary difference is the interpretation. For example, based on the data, we believe that there is 95% chance that body fat will increase by 5.8% up to 6.9% for every additional 10 centimeter increase in the waist circumference.

Credible Intervals for the Mean μ_Y and the Prediction y_{n+1}

Under the reference prior, the mean of Y, μ_Y, at the point x_i is given by $\alpha + \beta x_i$, with a posterior distributuion

$$\alpha + \beta x_i \mid \text{data} \sim t_{n-2}(\hat{\alpha} + \hat{\beta} x_i, S^2_{Y|X_i}),$$

where

$$S^2_{Y|X_i} = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right)$$

A new prediction y_{n+1} at a point x_{n+1} follows the Student's t-distribution

$$y_{n+1} \mid \text{data}, x_{n+1} \sim t_{n-2}\left(\hat{\alpha} + \hat{\beta} x_{n+1}, S^2_{Y|X_{n+1}} \right),$$

where

$$S^2_{Y|X_{n+1}} = \hat{\sigma}^2 + \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}} \right) = \hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}} \right).$$

The variance for predicting a new observation y_{n+1} has an extra $\hat{\sigma}^2$ which comes from the uncertainty of a new observation about the mean μ_Y estimated by the regressio line.

We can extract these intervals using the `predict` function

```
x = bodyfat$Abdomen
y = bodyfat$Bodyfat
xnew <- seq(min(x), max(x), length.out = 100)
ynew <- data.frame(predict(bodyfat.lm, newdata = data.frame(Abdomen = xnew),
                    interval = "confidence", level = 0.95))

# plot the data with the credible intervals of the mean
plot(x, y, xlab = "abdomen", ylab = "bodyfat", col = "blue", pch = 16)
lines(ynew$lwr ~ xnew, lty = 2, lwd = 3, col = "darkgrey")
lines(ynew$upr ~ xnew, lty = 2, lwd = 3, col = "darkgrey")
abline(bodyfat.lm, col = "orange")
```

```r
# plot the credible intervals of the prediction
ypred <- data.frame(predict(bodyfat.lm, newdata = data.frame(Abdomen = xnew),
                            interval = "prediction", level = 0.95))
lines(ypred$lwr ~ xnew, lty = 3, lwd = 3, col = "darkgrey")
lines(ypred$upr ~ xnew, lty = 3, lwd = 3, col = "darkgrey")

# identify the potential outlier: observation 39
points(bodyfat[39, "Abdomen"], bodyfat[39, "Bodyfat"], col = "orange", cex = 5)
legend(110, 15, legend = c("Posterior mean", "95% CI for mean", "95% CI for predictions"),
       col = c("orange", rep("darkgrey", 2)), lwd = 3, lty = c(1, 2, 3))
```
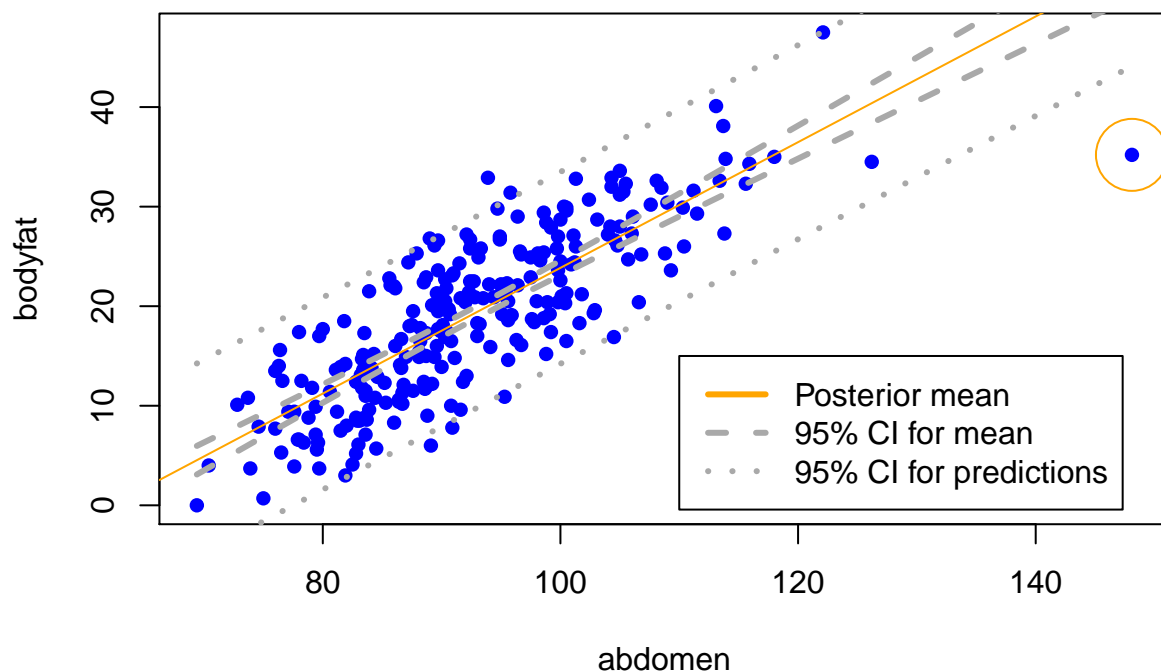

Note in the ablve the legend "CI" can mean either confidence interval or credible interval. The difference comes down to the interpretation. For example, the prediction at the same abdominal circumference as in Case 39 is

```r
pred.39 = predict(bodyfat.lm, newdata = bodyfat[39, ], interval = "prediction", level = 0.95)
out <- cbind(bodyfat[39,]$Abdomen, pred.39)
colnames(out) <- c("abdomen", "prediction", "lower", "upper")
out
```

```
##    abdomen prediction   lower    upper
## 39   148.1   54.21599 44.0967 64.33528
```

Based on the data, a Bayesian would expect that a man with waist circumference of 148.1 centermeters should have bodyfat of 54.216 with 95% chance thta it is between 44.097 and 64.335 percent.

While we expect the majority of the data will be within the prediction intervals (the short dashed grey lines), Case 39 seems to be well below the interval. We next use Bayesian methods to calculate the probability that this case is abnormal or an outlier by falling more than k standard deviations from either side of the mean.

6.1.3   Informative Priors

Except from the noninformative reference prior, we may also consider using a more general semi-conjugate prior distribution of α, β, and σ^2 when there is information available about the parameters.

Since the data $y_1, \cdots, y_n$ are normally distributed, from Chapter 3 we see that a Normal-Gamma distribution will form a conjugacy in this situation. We then set up the prior distributions through a hierarchical model. We first assume that, given σ^2, α and β together follow the bivariate normal prior distribution, from which their marginal distributions are both normal,

$$\alpha \mid \sigma^2 \sim \mathcal{N}(a_0, \sigma^2 \mathrm{S}_\alpha)$$
$$\beta \mid \sigma^2 \sim \mathcal{N}(b_0, \sigma^2 \mathrm{S}_\beta),$$

with covariance

$$\mathrm{Cov}(\alpha, \beta \mid \sigma^2) = \sigma^2 \mathrm{S}_{\alpha\beta}.$$

This is equivalent to setting the coefficient vector (α, β) to have a biraviate normal distribution with convariance matrix Σ_0 given by S_α, S_β, and $\mathrm{S}_{\alpha\beta}$

$$(\alpha, \beta) \mid \sigma^2 \sim \mathrm{BiraviateNormal}((a_0, b_0), \sigma^2 \Sigma_0).$$

Then for σ^2, we will use a Inverse Gamma distribution

$$\sigma^2 \sim \mathrm{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0}{2}\right).$$

Now the join prior distribution of α, β, and σ^2 form a distribution that is analogous to the Normal-Gamma distribution. Prior information about α, β, and σ^2 are encoded in the hyperparameters $a_0, b_0, \mathrm{S}_\alpha, \mathrm{S}_\beta, \mathrm{S}_{\alpha\beta}$, ν_0, and σ_0.

The marginal posterior distribution of the coefficient vector (α, β) will be biraviate normal, and the marginal posterior distribution of σ^2 is again a Inverse Gamma distribution

$$\sigma^2 \mid y_1, \cdots, y_n \sim \mathrm{IG}\left(\frac{\nu_0 + n}{2}, \frac{\nu_0 \sigma_0^2 + \mathrm{SSE}}{2}\right).$$

One can see that the reference prior is the limiting case of this conjugate prior we impose. We usually use Gibbs sampling to approximate the joint posterior distribution instead of using the result directly, especially later in multiple linear regression when we have more coefficients. We omit the deviations of the posterior distributions due to the heavy use of advanced linear algebra. One can refer to (Hopf's book, how to reference????) for more details.

Based on any prior information we have for the model, we can also impose other priors and assumptions on α, β, and σ^2 to get different Bayesian results. Most of these priors will not form any conjugacy and will require us to use simulation methods such as MCMC for approximations.

6.1.4   Deviations of Marginal Posterior Distributions of α, β, and σ^2

In this section, we will use the notations we introduced earlier such as SSE, the sum of squares of errors, $\hat{\sigma}^2$, the mean squared error, S_{xx}, $\mathrm{sd}_{\hat{\alpha}}$, $\mathrm{sd}_{\hat{\beta}}$ and so on to simplify our calculations.

We will also use the following quantities derived from the formula of $\bar{x}$, $\bar{y}$, $\hat{\alpha}$, and $\hat{\beta}$

$$\sum_i^n (x_i - \bar{x}) = 0$$

$$\sum_i^n (y_i - \bar{y}) = 0$$

$$\sum_i^n (y_i - \hat{y}_i) = \sum_i^n (y_i - (\hat{\alpha} + \hat{\beta}x_i)) = 0$$

$$\sum_i^n (x_i - \bar{x})(y_i - \hat{y}_i) = \sum_i^n (x_i - \bar{x})(y_i - \bar{y} - \hat{\beta}(x_i - \bar{x})) = \sum_i^n (x_i - \bar{x})(y_i - \bar{y}) - \hat{\beta}\sum_i^n (x_i - \bar{x})^2 = 0$$

$$\sum_i^n x_i^2 = \sum_i^n (x_i - \bar{x})^2 + n\bar{x}^2 = S_{xx} + n\bar{x}^2$$

We first further simplify the numerator inside the exponential function in the formula of $\pi^*(\alpha, \beta, \sigma^2 \mid y_1, \cdots, y_n)$:

$$\sum_i^n (y_i - \alpha - \beta x_i)^2 = \sum_i^n \left(y_i - \hat{\alpha} - \hat{\beta}x_i - (\alpha - \hat{\alpha}) - (\beta - \hat{\beta})x_i \right)^2$$

$$= \sum_i^n \left(y_i - \hat{\alpha} - \hat{\beta}x_i \right)^2 + \sum_i^n (\alpha - \hat{\alpha})^2 + \sum_i^n (\beta - \hat{\beta})^2 (x_i)^2$$

$$- 2\sum_i^n (\alpha - \hat{\alpha})(y_i - \hat{\alpha} - \hat{\beta}x_i)$$

$$- 2\sum_i^n (\beta - \hat{\beta})(x_i)(y_i - \hat{\alpha} - \hat{\beta}x_i)$$

$$+ 2\sum_i^n (\alpha - \hat{\alpha})(\beta - \hat{\beta})(x_i)$$

$$= SSE + n(\alpha - \hat{\alpha})^2 + (\beta - \hat{\beta})^2 \sum_i^n x_i^2 - 2(\alpha - \hat{\alpha}) \sum_i^n (y_i - \hat{y}_i)$$

$$- 2(\beta - \hat{\beta}) \sum_i^n x_i(y_i - \hat{y}_i) + 2(\alpha - \hat{\alpha})(\beta - \hat{\beta})(n\bar{x})$$

It is clear that

$$-2(\alpha - \hat{\alpha}) \sum_i^n (y_i - \hat{y}_i) = 0$$

And

$$-2(\beta - \hat{\beta}) \sum_i^n x_i(y_i - \hat{y}_i) = -2(\beta - \hat{\beta}) \sum_i^n (x_i - \bar{x})(y_i - \hat{y}_i) - 2(\beta - \hat{\beta}) \sum_i^n \bar{x}(y_i - \hat{y}_i)$$

$$= -2(\beta - \hat{\beta}) \times 0 - 2(\beta - \hat{\beta})\bar{x} \sum_i^n (y_i - \hat{y}_i) = 0$$

Finally, we use the quantity that $\sum_i^n x_i^2 = \sum_i^n (x_i - \bar{x})^2 + n\bar{x}^2$ to combine the terms $n(\alpha - \hat{\alpha})^2$, $2(\alpha - \hat{\alpha})(\beta - \hat{\beta}) \sum_i^n x_i$, and $(\beta - \hat{\beta})^2 \sum_i^n x_i^2$ together.

$$\sum_i^n (y_i - \alpha - \beta x_i)^2 = \text{SSE} + n(\alpha - \hat{\alpha})^2 + (\beta - \hat{\beta})^2 \sum_i^n (x_i - \bar{x})^2 + (\beta - \hat{\beta})^2 (n\bar{x}^2)$$

$$+ 2(\alpha - \hat{\alpha})(\beta - \hat{\beta})(n\bar{x})$$

$$= \text{SSE} + (\beta - \hat{\beta})^2 \text{S}_{xx} + n \left[(\alpha - \hat{\alpha}) + (\beta - \hat{\beta})\bar{x} \right]^2$$

Therefore, the posterior joint distribution of α, β, σ^2 is

$$\pi^*(\alpha^*, \beta, \sigma^2 \mid y_1, \cdots, y_n) \propto \frac{1}{(\sigma^2)^{(n+2)/2}} \exp\left(-\frac{\sum_i (y_i - \alpha - \beta x_i)^2}{2\sigma^2} \right)$$

$$= \frac{1}{(\sigma^2)^{(n+2)/2}} \exp\left(-\frac{\text{SSE} + n(\alpha - \hat{\alpha} + (\beta - \hat{\beta})\bar{x})^2 + (\beta - \hat{\beta})^2 \sum_i (x_i - \bar{x})^2}{2\sigma^2} \right)$$

Marginal Posterior Distribution of β

To get the marginal posterior distribution of β, we need to integrate out α and σ^2:

$$\pi^*(\beta \mid y_1, \cdots, y_n) = \int_0^\infty \int_{-\infty}^\infty \pi^*(\alpha, \beta, \sigma^2 \mid y_1, \cdots, y_n) \, d\alpha \, d\sigma^2$$

$$= \int_0^\infty \left(\int_{-\infty}^\infty \frac{1}{(\sigma^2)^{(n+2)/2}} \exp\left(-\frac{\text{SSE} + n(\alpha - \hat{\alpha} + (\beta - \hat{\beta})\bar{x})^2 + (\beta - \hat{\beta}) \sum_i (x_i - \bar{x})^2}{2\sigma^2} \right) d\alpha \right) d\sigma^2$$

The integral inside is the joint posterior distribution of β and σ^2

$$\pi^*(\beta, \sigma^2 \mid y_1, \cdots, y_n)$$

$$= \int_{-\infty}^\infty \frac{1}{(\sigma^2)^{(n+2)/2}} \exp\left(-\frac{\text{SSE} + n(\alpha - \hat{\alpha} + (\beta - \hat{\beta})\bar{x})^2 + (\beta - \hat{\beta})^2 \sum_i (x_i - \bar{x})^2}{2\sigma^2} \right) d\alpha$$

$$= \int_{-\infty}^\infty \frac{1}{(\sigma^2)^{(n+2)/2}} \exp\left(-\frac{\text{SSE} + (\beta - \hat{\beta})^2 \sum_i (x_i - \bar{x})^2}{2\sigma^2} \right) \exp\left(-\frac{n(\alpha - \hat{\alpha} + (\beta - \hat{\beta})\bar{x})^2}{2\sigma^2} \right) d\alpha$$

$$= \frac{1}{(\sigma^2)^{(n+2)/2}} \exp\left(-\frac{\text{SSE} + (\beta - \hat{\beta})^2 \sum_i (x_i - \bar{x})^2}{2\sigma^2} \right) \int_{-\infty}^\infty \exp\left(-\frac{n(\alpha - \hat{\alpha} + (\beta - \hat{\beta})\bar{x})^2}{2\sigma^2} \right) d\alpha$$

Here,

$$\exp\left(-\frac{n(\alpha - \hat{\alpha} + (\beta - \hat{\beta})\bar{x})^2}{2\sigma^2} \right)$$

can be viewed as part of a normal distribution of α, with mean $\hat{\alpha} - (\beta - \hat{\beta})\bar{x}$, and variance σ^2/n. Therefore, the integral from the last line above is proportional to $\sqrt{\sigma^2/n}$. We get

$$\pi^*(\beta, \sigma^2 \mid y_1, \cdots, y_n) \propto \frac{1}{(\sigma^2)^{(n+2)/2}} \exp\left(-\frac{\text{SSE} + (\beta - \hat{\beta})^2 \sum_i (x_i - \bar{x})^2}{2\sigma^2} \right) \times \sqrt{\sigma^2}$$

$$= \frac{1}{(\sigma^2)^{(n+1)/2}} \exp\left(-\frac{\text{SSE} + (\beta - \hat{\beta})^2 \sum_i (x_i - \bar{x})^2}{2\sigma^2} \right)$$

We then integrate σ^2 out to get the marginal distribution of β. Here we first perform change of variable and set $\sigma^2 = \frac{1}{\phi}$. Then the integral becomes

$$\pi^*(\beta \mid y_1, \cdots, y_n) \propto \int_0^\infty \frac{1}{(\sigma^2)^{(n+1)/2}} \exp\left(-\frac{\text{SSE} + (\beta - \hat{\beta})^2 \sum_i (x_i - \bar{x})^2}{2\sigma^2}\right) d\sigma^2$$

$$\propto \int_0^\infty \phi^{\frac{n-3}{2}} \exp\left(-\frac{\text{SSE} + (\beta - \hat{\beta})^2 \sum_i (x_i - \bar{x})^2}{2}\phi\right) d\phi$$

$$\propto \left(\frac{\text{SSE} + (\beta - \hat{\beta})^2 \sum_i (x_i - \bar{x})^2}{2}\right)^{-\frac{(n-2)+1}{2}} \int_0^\infty s^{\frac{n-3}{2}} e^{-s} \, ds$$

Here we use another change of variable by setting $s = \dfrac{\text{SSE} + (\beta - \hat{\beta})^2 \sum_i (x_i - \bar{x})^2}{2}\phi$, and the fact that $\displaystyle\int_0^\infty s^{(n-3)/2} e^{-s} \, ds$ gives us the Gamma function, which is a constant.

We can rewrite the last line from above to obtain the marginal posterior distribution of β to be the Student's t-distribution with degrees of freedom $n - 2$, center $\hat{\beta}$, and scale parameter $\dfrac{\hat{\sigma}^2}{\sum_i (x_i - \bar{x})^2}$

$$\pi^*(\beta \mid y_1, \cdots, y_n) \propto \left[1 + \frac{1}{n-2}\frac{(\beta - \hat{\beta})^2}{\frac{\text{SSE}}{n-2}/(\sum_i (x_i - \bar{x})^2)}\right]^{-\frac{(n-2)+1}{2}} = \left[1 + \frac{1}{n-2}\frac{(\beta - \hat{\beta})^2}{\hat{\sigma}^2/(\sum_i (x_i - \bar{x})^2)}\right]^{-\frac{(n-2)+1}{2}},$$

where $\dfrac{\hat{\sigma}^2}{\sum_i (x_i - \bar{x})^2}$ is exactly the square of the standard error of $\hat{\beta}$ from the frequentist OLS model.

To summarize, under the reference prior, the marginal posterior distribution of the slope of the Bayesian simple linear regression follows the Student's t-distribution

$$\beta \mid y_1, \cdots, y_n \sim t_{n-2}\left(\hat{\beta}, \left(\text{sd}_{\hat{\beta}}\right)^2\right)$$

Marginal Posterior Distribution of α

A similar approach will lead us to the marginal distribution of α. We again start from the joint posterior distribution

$$\pi^*(\alpha, \beta, \sigma^2 \mid y_1, \cdots, y_n) \propto \frac{1}{(\sigma^2)^{(n+2)/2}} \exp\left(-\frac{\text{SSE} + n(\alpha - \hat{\alpha} - (\beta - \hat{\beta})\bar{x})^2 + (\beta - \hat{\beta})^2 \sum_i (x_i - \bar{x})^2}{2\sigma^2}\right)$$

This time we integrate β and σ^2 out to get the marginal posterior distribution of α. We first compute the integral

$$\pi^*(\alpha, \sigma^2 \mid y_1, \cdots, y_n) = \int_{-\infty}^\infty \pi^*(\alpha, \beta, \sigma^2 \mid y_1, \cdots, y_n) \, d\beta$$

$$= \int_{-\infty}^\infty \frac{1}{(\sigma^2)^{(n+2)/2}} \exp\left(-\frac{\text{SSE} + n(\alpha - \hat{\alpha} + (\beta - \hat{\beta})\bar{x})^2 + (\beta - \hat{\beta})^2 \sum_i (x_i - \bar{x})^2}{2\sigma^2}\right) d\beta$$

Here we group the terms with $\beta - \hat{\beta}$ together, then complete the square so that we can treat is as part of a

normal distribution function to simplify the integral

$$n(\alpha - \hat{\alpha} + (\beta - \hat{\beta})\bar{x})^2 + (\beta - \hat{\beta})^2 \sum_i (x_i - \bar{x})^2$$

$$= (\beta - \hat{\beta})^2 \left(\sum_i (x_i - \bar{x})^2 + n\bar{x}^2 \right) + 2n\bar{x}(\alpha - \hat{\alpha})(\beta - \hat{\beta}) + n(\alpha - \hat{\alpha})^2$$

$$= \left(\sum_i (x_i - \bar{x})^2 + n\bar{x}^2 \right) \left[(\beta - \hat{\beta}) + \frac{n\bar{x}(\alpha - \hat{\alpha})}{\sum_i (x_i - \bar{x})^2 + n\bar{x}^2} \right]^2 + n(\alpha - \hat{\alpha})^2 \left[\frac{\sum_i (x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2 + n\bar{x}^2} \right]$$

$$= \left(\sum_i (x_i - \bar{x})^2 + n\bar{x}^2 \right) \left[(\beta - \hat{\beta}) + \frac{n\bar{x}(\alpha - \hat{\alpha})}{\sum_i (x_i - \bar{x})^2 + n\bar{x}^2} \right]^2 + \frac{(\alpha - \hat{\alpha})^2}{\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}}$$

When integrating, we can then view

$$\exp\left(-\frac{\sum_i (x_i - \bar{x})^2 + n\bar{x}^2}{2\sigma^2} \left(\beta - \hat{\beta} + \frac{n\bar{x}(\alpha - \hat{\alpha})}{\sum_i (x_i - \bar{x})^2 + n\bar{x}^2} \right)^2 \right)$$

as part of a normal distribution function, and get

$$\pi^*(\alpha, \sigma^2 \mid y_1, \cdots, y_n)$$

$$\propto \frac{1}{(\sigma^2)^{(n+2)/2}} \exp\left(-\frac{\text{SSE} + (\alpha - \hat{\alpha})^2 / (\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2})}{2\sigma^2} \right)$$

$$\times \int_{-\infty}^{\infty} \exp\left(-\frac{\sum_i (x_i - \bar{x})^2 + n\bar{x}^2}{2\sigma^2} \left(\beta - \hat{\beta} + \frac{n\bar{x}(\alpha - \hat{\alpha})}{\sum_i (x_i - \bar{x})^2 + n\bar{x}^2} \right)^2 \right) d\beta$$

$$\propto \frac{1}{(\sigma^2)^{(n+1)/2}} \exp\left(-\frac{\text{SSE} + (\alpha - \hat{\alpha})^2 / (\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2})}{2\sigma^2} \right)$$

To get the marginal posterior distribution of α, we again integrate σ^2 out. using the same change of variable $\sigma^2 = \frac{1}{\phi}$, and $s = \dfrac{\text{SSE} + (\alpha - \hat{\alpha})^2 / (\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2})}{2} \phi$.

$$\pi^*(\alpha \mid y_1, \cdots, y_n)$$

$$= \int_0^{\infty} \pi^*(\alpha, \sigma^2 \mid y_1, \cdots, y_n) \, d\sigma^2$$

$$\propto \int_0^{\infty} \phi^{(n-3)/2} \exp\left(-\frac{\text{SSE} + (\alpha - \hat{\alpha})^2 / (\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2})}{2} \phi \right) d\phi$$

$$\propto \left(\text{SSE} + (\alpha - \hat{\alpha})^2 / (\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}) \right)^{-\frac{(n-2)+1}{2}} \int_0^{\infty} s^{(n-3)/2} e^{-s} \, ds$$

$$\propto \left[1 + \frac{1}{n-2} \frac{(\alpha - \hat{\alpha})^2}{\frac{\text{SSE}}{n-2} \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right)} \right]^{-\frac{(n-2)+1}{2}} \qquad = \left[1 + \frac{1}{n-2} \left(\frac{\alpha - \hat{\alpha}}{\text{sd}_{\hat{\alpha}}} \right)^2 \right]^{-\frac{(n-2)+1}{2}}$$

This shows that the marginal posterior distribution of α also follows a Student's t-distribution, with $n-2$ degrees of freedom. Its center is $\hat{\alpha}$, the usual coefficient in the frequentist OLS estimate, and its scale parameter is $\hat{\sigma}^2 \left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right)$, which is the square of the standard error of $\hat{\alpha}$.

Marginal Posterior Distribution of σ^2

To show that the marginal posterior distribution of σ^2 follows the inverse Gamma distribution, we only need to show the precision $\phi = \dfrac{1}{\sigma^2}$ follows a Gamma distribution.

We have shown in Week 3 that taking the prior distribution of σ^2 proportional to $\dfrac{1}{\sigma^2}$ is equivalent to taking the prior distribution of ϕ proportional to $\dfrac{1}{\phi}$

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2} \qquad \Longrightarrow \qquad \pi(\phi) \propto \frac{1}{\phi}$$

Therefore, under the parameters α, β, and the precision ϕ, we have the joint prior distribution as

$$\pi(\alpha, \beta, \phi) \propto \frac{1}{\phi}$$

and the joint posterior distribution as

$$\pi^*(\alpha, \beta, \phi \mid y_1, \cdots, y_n) \propto \phi^{\frac{n}{2}-1} \exp\left(-\frac{\sum_i (y_i - \alpha - \beta x_i)}{2} \phi \right)$$

Using the partial results we have calculated previously, we get

$$\pi^*(\beta, \phi \mid y_1, \cdots, y_n) = \int_{-\infty}^{\infty} \pi^*(\alpha, \beta, \phi \mid y_1, \cdots, y_n)\, d\alpha \propto \phi^{\frac{n-3}{2}} \exp\left(-\frac{\mathrm{SSE} + (\beta - \hat{\beta})^2 \sum_i (x_i - \bar{x})^2}{2} \phi \right)$$

Intergrating over β, we finally have

$$\pi^*(\phi \mid y_1, \cdots, y_n)$$
$$\propto \int_{-\infty}^{\infty} \phi^{\frac{n-3}{2}} \exp\left(-\frac{\mathrm{SSE} + (\beta - \hat{\beta})^2 \sum_i (x_i - \bar{x})^2}{2} \phi \right) d\beta$$
$$= \phi^{\frac{n-3}{2}} \exp\left(-\frac{\mathrm{SSE}}{2} \phi \right) \int_{-\infty}^{\infty} \exp\left(-\frac{(\beta - \hat{\beta})^2 \sum_i (x_i - \bar{x})^2}{2} \phi \right) d\beta$$
$$\propto \phi^{\frac{n-4}{2}} \exp\left(-\frac{\mathrm{SSE}}{2} \phi \right) = \phi^{\frac{n-2}{2}-1} \exp\left(-\frac{\mathrm{SSE}}{2} \phi \right).$$

This is a Gamma distribution with shape parameter $\dfrac{n-2}{2}$ and rate parameter $\dfrac{\mathrm{SSE}}{2}$. Therefore, the updated σ^2 follows the inverse Gamma distribution

$$\sigma^2 \mid y_1, \cdots, y_n \sim \mathrm{IG}\left(\frac{n-2}{2}, \frac{\mathrm{SSE}}{2} \right).$$

6.2 Checking Outliers

The plot and predictive intervals suggest that predictions for Case 39 are not well captured by the model. There is always the possibility that this case does not meet the assumptions of the simple linear regression

model (wrong mean or variance) or could be in error. Model diagnostics such as plots of residuals versus fitted values are useful in identifying potential outliers. Now with the interpretation of Bayesian paradigm, we can go further to calculate the probability to demonstrate whether a case falls too far from the mean.

The article by Chaloner and Brant (1988) suggested an approach for defining outliers and then calculating the probability that a case or multiple cases were outliers. The assumed model for our simple linear regression is $y_i = \alpha + \beta x_i + \epsilon_i$, with ϵ_i having independent, identical distributions that are normal with mean zero and constant variance σ^2, i.e., $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Chaloner & Brant consider outliers to be points where the error or the model discrepancy ϵ_i is greater than k standard deviation for some large k, and then proceed to calculate the posterior probability that a case is an outlier to be

$$\mathbb{P}(|\epsilon_i| > k\sigma \mid \text{data})$$

Since $\epsilon_i = y_i - \alpha - \beta x_i$, this is equivalent to calculating

$$\mathbb{P}(|y_i - \alpha - \beta x_i| > k\sigma \mid \text{data}).$$

6.2.1 Posterior Distribution of ϵ_i Conditioning On σ

6.2.2 Implementation Using `BAS` Package

The code for calculating the probability of outliers involves integration. We have implemented this in the function `Bayes.outlier.prob` that can be sourced from the file `bayes-outliers.R`. Applying this to the `bodyfat` data for Case 39, we get

```
library(BAS)
data(bodyfat)
#source("bayes-outliers.R")
#library(mvtnorm)
#outliers = Bayes.outlier.prob(bodyfat.lm)

# The default is to consider k=3
#prob.39 = outliers$prob.outlier[39]
#prob.39
```

We see that this case has an extremely high probability (0.9917) of being more an outlier, that is, the error is greater than $k = 3$ standard deviations, based on the fitted model and data.

With $k = 3$, however, there may be a high probability a priori of at least one outlier in a large sample. We can compute this using

```
n = nrow(bodyfat)
# probability of no outliers if outliers are error greater than 3 standard deviation
prob = (1 - (2 * pnorm(-3))) ^ n
prob
```

```
## [1] 0.5059747
```

```
# probability of at least one outlier
prob.least1 = 1 - (1 - (2 * pnorm(-3))) ^ n
prob.least1
```

```
## [1] 0.4940253
```

With $n = 252$, the probability of at least one outlier is much larger than say the marginal probability that one point is an outlier of 0.05. So we would expect that there will be at least one point where the error is more than 3 standard deviations from zero almost 50% of the time. Rather than fixing k, we can fix the prior probability of no outliers to be say 0.95, and solve for the value of k.

```
k = qnorm(0.5 + 0.5 * 0.95 ^ (1 / n))
k
```

```
## [1] 3.714602
```

This leads to a larger value of k after adjusting k so that there is at least one outlier. We examine Case 39 again under this k

```
#outliers.no = Bayes.outlier.prob(bodyfat.lm, k = k)
#prob.no.39 = outliers.no$prob.outlier[39]
#prob.no.39
```

The posterior probability of Case 39 being an outlier is 0.68475. While this is not strikingly large, it is much larger than the marginal prior probability of

```
2 * pnorm(-k)
```

```
## [1] 0.0002035241
```

6.2.3 Summary

There is a substantial probability that Case 39 is an outlier. If you do view it as an outlier, what are your options? One option is to investigate the case and determine if the data are input incorrectly, and fix it. Another option is when you cannot confirm there is a data entry error, you may delete the observation from the analysis and refit the model without the case. If you do take this option, be sure to describe what you did so that your research is reproducible. You may want to apply diagnostics and calculate the probability of a case being an outlier using this reduced data. As a word of caution, if you discover that there are a large number of points that appear to be outliers, take a second look at your model assumptions, since the problem may be with the model rather than the data! A third option we will talk about later, is to combine inference under the model that retains this case as part of the population, and the model that treats it as coming from another population. This approach incorporates our uncertainty about whether the case is an outlier given the data.

The source code is based on using a reference prior for the linear model and extends to multiple regression.

6.3 Bayesian Multiple Linear Regression

In this section, we will discuss Bayesian inference in multiple linear regression. We will use the reference prior to provide the default, or base line analysis of the model, which provides the correspondence between Bayesian and frequentist approaches.

6.3.1   The Model

To illustrate the idea, we use the data set that we examined earlier on kids' cognitive scores, where we predicted the value of the kid's cognitive score from the mother's high school status, mother's IQ score, whether or not the mother worked during the first three years of the kid's life, and the mother's age. We set up the model as follows

$$\text{score}_i = \beta_0 + \beta_1 \text{hs}_i + \beta_2 \text{IQ}_i + \beta_3 \text{work}_i + \beta_4 \text{age}_i + \epsilon_i.$$

Here, score_i is the ith kid's cognitive score. hs_i, IQ_i, work_i, and age_i represent the high school status, the IQ score, work status during the first three years and the age of the ith kid's mother. ϵ_i is the error term.

6.3.2   Data Pre-processing

We can download the data set from Gelman's website and read the summary information of the data set.

```
cognitive = foreign::read.dta("http://www.stat.columbia.edu/~gelman/arm/examples/child.iq/kidiq.dta")
summary(cognitive)
```

```
##    kid_score          mom_hs          mom_iq          mom_work
##  Min.   : 20.0   Min.   :0.0000   Min.   : 71.04   Min.   :1.000
##  1st Qu.: 74.0   1st Qu.:1.0000   1st Qu.: 88.66   1st Qu.:2.000
##  Median : 90.0   Median :1.0000   Median : 97.92   Median :3.000
##  Mean   : 86.8   Mean   :0.7857   Mean   :100.00   Mean   :2.896
##  3rd Qu.:102.0   3rd Qu.:1.0000   3rd Qu.:110.27   3rd Qu.:4.000
##  Max.   :144.0   Max.   :1.0000   Max.   :138.89   Max.   :4.000
##    mom_age
##  Min.   :17.00
##  1st Qu.:21.00
##  Median :23.00
##  Mean   :22.79
##  3rd Qu.:25.00
##  Max.   :29.00
```

As we see that variables `mom_hs` and `mom_work` should be considered as categorical variables. We transform them into indicator variables where `mom_work` shows whether the mother worked for 1 or more years, and `mom_hs` indicates whether the mother had more than a high school education.

The code is as below:[1]

```
cognitive$mom_work <- as.numeric(cognitive$mom_work > 1)
cognitive$mom_hs <- as.numeric(cognitive$mom_hs > 0)

# Modify column names of the data set
colnames(cognitive) <- c("kid_score", "hs", "IQ", "work", "age")
```

6.3.3   Specify Bayesian Prior Distributions

For Bayesian inference, we need to specify a prior distribution of the error terms. Since the kid's cognitive scores score_i are continuous, we assume that ϵ_i are independent, and identically distributed with the normal distribution

$$\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2),$$

---

[1]Note: `as.numeric` is not necessary here. We use `as.numeric` to keep the names of the levels of the two variables short.

where σ^2 is the commonly shared variance of all observations.

We will also need to specify the prior distribution for all the coefficients β_0, β_1, β_2, β_3, and β_4. An informative prior, which assumes that the β's follow the multivariate normal distribution with covariance matrix $\sigma^2 \Sigma_0$ can be used. We may further impose the inverse Gamma distribution to σ^2, to complete the hierachical model

$$\beta_0, \beta_1, \beta_2, \beta_3, \beta_4 \mid \sigma^2 \ \sim \mathcal{N}((b_0, b_1, b_2, b_3, b_4), \sigma^2 \Sigma_0)$$
$$\sigma^2 \ \sim \text{IG}(\nu_0/2, \nu_0 \sigma_0^2/2)$$

This gives us the multivariate normal-gamma conjugate family, with hyperparameters $b_0, b_1, b_2, b_3, b_4, \Sigma_0, \nu_0$, and σ_0^2. For this prior, we will need to specify the values of all the hyperparameters. This elicitation can be quite involved, especially when we do not have enough prior information about the variances, covariances of the coefficients and other prior hyperparameters. Therefore, we are going to adopt the noninformative prior (i.e., the reference prior), which is a limiting case of this multivariate normal-gamma prior.

The reference prior in the multiple linear regression model is similar to the reference prior we used in the simple linear regression model, where the prior distribution of all the coefficients β's is the uniform prior, and the prior of σ^2 is proportional to its reciprocal

$$\pi(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4 \mid \sigma^2) \propto 1, \qquad \pi(\sigma^2) \propto \frac{1}{\sigma^2}.$$

Under this reference prior, the posterior distributions of the coefficients, β's, are parallel to the ones in simple linear regression. The marginal posterior distributions of β's are the Student's t-distributions with centers given by the frequentist OLS estimates $\hat{\beta}$'s, scale parameters given by the standard errors $\text{SE}_{\hat{\beta}}^2$ obtained from the OLS estimates

$$\beta_j \mid y_1, \cdots, y_n \ \sim t_{n-p-1}(\hat{\beta}_j, \text{SE}_{\hat{\beta}_j}^2), \qquad j = 0, 1, \cdots, p.$$

The degrees of freedom of these t-distributions are $n - p - 1$, where p is the number of predictors. In the kid cognitive score example, $p = 4$. The posterior mean, $\hat{\beta}_j$, is the center of the t-distributions of β_j, which is the same as the OLS estimates of β_j. The posterior standard deviation of β_j, which is the square root of the scale parameter of the t-distribution, is SE_{β_j}, the standard error of β_j under the OLS estimates. That means, under the reference prior, we can easily obtain the posterior mean and posterior standard deviation from using the `lm` function, since they are numerically equivalent to the counterpart of the frequentist approach.

6.3.4 Fitting the Bayesian Model

To gain more flexibility in choosing priors, we will instead use the `bas.lm` function in the `BAS` library, which allows us to specify different model priors.

```
library(BAS)
cog.bas = bas.lm(kid_score ~ ., data = cognitive, prior = "BIC",
                 modelprior = Bernoulli(1), bestmodel = rep(1, 5), n.models = 1)
```

The above `bas.lm` function uses the model formula the same as in the `lm`. It first specifies the response and predictor variables, a data argument to provide the data frame. The addition arguments further include the prior on the coefficients. We use `"BIC"` here to indicate that the model is based on the non-informative reference prior. (We will explain in the later section why we use the name `"BIC"`.) The `modelprior` argument tells the function to include all variables, which can be viewed as every variable is included with probability 1 under a Bernoulli trial. Because we want to fit just the full model, we use `bestmodel = rep(1,5)` to indicate that the intercept and all 4 predictors are included. The argument `n.models = 1` fit just this one model.

6.3.5   Posterior Means and Posterior Standard Deviations

Similar to the OLS regression process, we can extract the posterior means and standard deviations of the coefficients using the `coef` function

```
cog.coef <- coef(cog.bas)
cog.coef
```

```
##
##  Marginal Posterior Summaries of Coefficients:
##
##  Using  BMA
##
##  Based on the top  1 models
##            post mean  post SD   post p(B != 0)
## Intercept  86.79724   0.87092   1.00000
## hs          5.09482   2.31450   1.00000
## IQ          0.56147   0.06064   1.00000
## work        2.53718   2.35067   1.00000
## age         0.21802   0.33074   1.00000
```

From the last column in this summary, we see that the probability of the coefficients to be non-zero is always 1. This is because we specify the argument `Bernoulli(1)` to force the model to include all variables.

We can visualize the coefficients in front of the predictors $\beta_1, \beta_2, \beta_3, \beta_4$ using the `plot` function. We use the `subset` argument to plot only the coefficients of the predictors.

```
par(mfrow = c(2, 2), col.lab = "darkgrey", col.axis = "darkgrey", col = "darkgrey")
plot(cog.coef, subset = 2:5, ask = F)
```
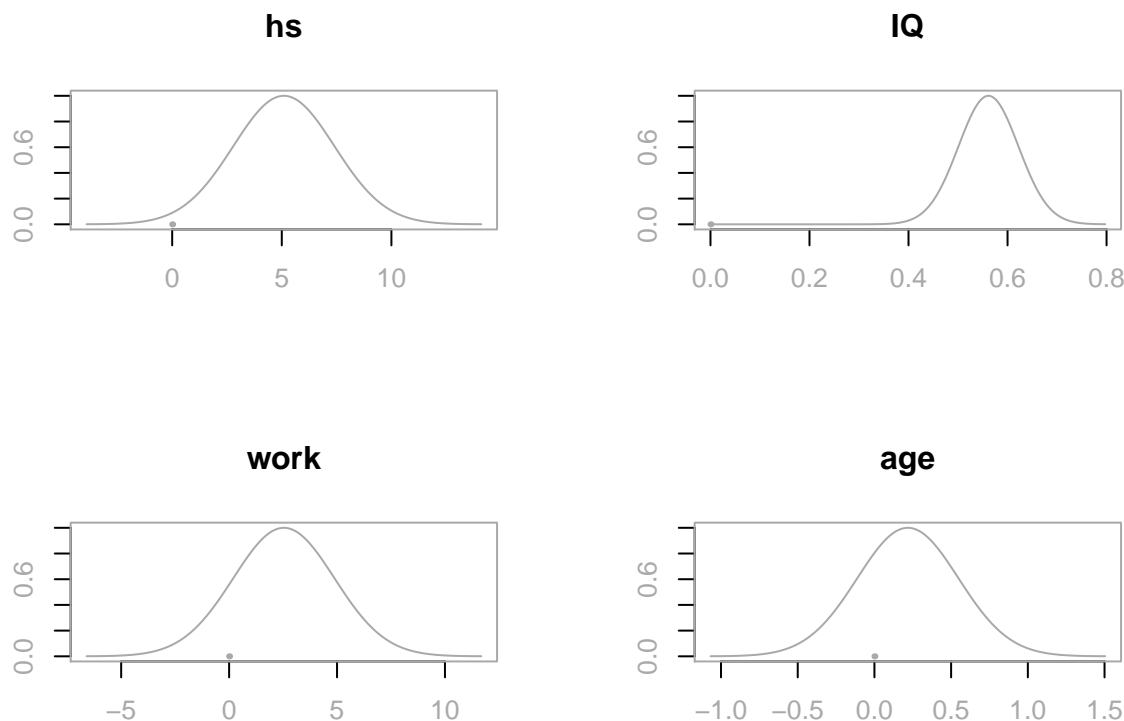
These distributions all center at their respetive OLS estimates $\hat{\beta}_j$, with the spread of the distribution related to the standard errors.

6.3.6   Credible Intervals Summary

We can also report the posterior means, posterior standard deviations, and the 95% credible intervals of the coefficients of all 4 predictors, which may give a clearer and more useful summary. The **BAS** library provides the method `confint` to extract the credible intervals from the output `cog.coef`. If we are only interested in the distributions of the coefficients of the 4 predictors, we may use the `parm` argument to restrict the variables shown in the summary

```
confint(cog.coef, parm = 2:5)
```

```
##             2.5%      97.5%       beta
## hs     0.5456507 9.6439990 5.0948248
## IQ     0.4422784 0.6806616 0.5614700
## work -2.0830879 7.1574454 2.5371788
## age  -0.4320547 0.8680925 0.2180189
## attr(,"Probability")
## [1] 0.95
## attr(,"class")
## [1] "confint.bas"
```

All together, we can generate a summary table showing the posterior means, posterior standard deviations, the upper and lower bounds of the 95% credible intervals of all coefficients $\beta_0, \beta_1, \beta_2, \beta_3$, and β_4.

```r
out <- confint(cog.coef)[, 1:2]   # only extract the upper and lower bounds of the credible intervals
names <- c("posterior mean", "posterior std", colnames(out))
out <- cbind(cog.coef$postmean, cog.coef$postsd, out)
colnames(out) <- names

round(out, 2)
```

```
##           posterior mean posterior std  2.5% 97.5%
## Intercept          86.80          0.87 85.09 88.51
## hs                  5.09          2.31  0.55  9.64
## IQ                  0.56          0.06  0.44  0.68
## work                2.54          2.35 -2.08  7.16
## age                 0.22          0.33 -0.43  0.87
```

As in the simple linear aggression, the posterior estimates from the reference prior, that are in the table, are equivalent to the numbers reported from the lm function in R, or using the confident function in the OLS estimates. These intervals are centered at the posterior mean with width given by the appropriate t quantile with $n - p - 1$ degrees of freedom times the posterior standard deviation. The primary difference is the interpretation of the intervals. For example, given this data we believe there is a 95% chance that the kid's cognitive score increases by 0.44 to 0.68 with an additional increase of the mother's IQ score. The mother's high school status has a larger effect where we believe that there is a 95% chance the kid would score of 0.55 up to 9.64 points higher if the mother had three or more years of high school. The credible intervals of the predictors work and age include 0, which implies that we may improve this model so that the model will accomplish a desired level of explanation or prediction with fewer predictors.

Chapter 7

Bayesian Model Selection

In the last section, we provided a Bayesian inference analysis for kid's cognitive scores using multiple linear regression. We found that several credible intervals of the coefficients contain zero, suggesting that we could potentially simplify the model. In this section, we will discuss model selection, which is picking variables for multiple linear regression based on Bayesian information criterion, or BIC. In the next section, we will discuss other model selection methods, such as using Bayes factors.

7.1 Bayesian Information Criterion (BIC)

Bayesian information criterion (BIC) is one of the Bayesian criteria used for Bayesian model selection, and tends to be one of the most popular criteria.

Bibliography

Chaloner, K. and Brant, R. (1988). A bayesian approach to outlier detection and residual analysis. *Biometrika*, 75(4):651–659.

Jeffreys, S. H. (1961). *Theory of Probability: 3rd Edition*. Clarendon Press.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, 90(430):773–795.