Some of Bayesian Statistics using `R`

Rebecca C. Steorts

Last LaTeX'd Wednesday 13th August, 2014

Contents

Chapter 1

Motivations and Introduction to Bayesian Inference

1.1 Motivational Examples

Example 1.1: The New York Times article listed below talks about the financial crisis of 2008. Specifically it points to risk analysis and model failure. We will discuss risk in our last chapter of this class. What can we learn from this article?

`http://www.nytimes.com/2009/09/13/business/13unboxed.html?_r=1`

Example 1.2: We want to generate a random integer between 1 and 10. Let's try doing this ourselves and see what happens. Everyone in the class choose a random number quickly and write it down. What has happened?

The goal of generating a random integer between 1 and 10 is that each integer should have an equally likely chance of being chosen.

From our experiment, did this occur? What went wrong?

Since we cannot generate a random number ourselves because we are likely to choose some numbers more frequently than others, we can generate this number using software. We will come back to this example and go through how to generate such an integer in R. When you go home today, download R and begin reading Chapter 1 of *Using R for Introductory Statistics* if you bought the book. Also, read the `R` Help document I have posted on the course webpage when you go home.

Example 1.3: The article listed below in the New York Times from April 25, 2010, talks about the confusion that students as well as professionals such as physicians have regarding Bayes' Theorem and conditional probabilities.

http://opinionator.blogs.nytimes.com/2010/04/25/chances-are/

The article deals with calculating the probability that a woman has breast cancer given she has a positive mammogram, which is a calculation that needs to be done correctly in a real-world situation.

Let's take a few minutes to read through the article. Focus specifically on the section where a doctor estimates a womans probability of having breast cancer, given that she has a positive mammogram, to be 90 percent. What is wrong with his answer intuitively? Notice that as Gigerenzer talks to many other doctors their estimates are very different, some giving estimates at 10 percent and others giving estimates between 50 and 80 percent. So, what's the correct answer and how do we reason this out? First, let's look at all the information we have.

Let's look at an example from the article where we're interested in the probability that a woman has breast cancer given she has a positive mammogram even though she is in a low-risk group (i.e., 40–50 years old with no family history of breast cancer). The probability that one of these women has breast cancer is 0.008. If a woman has breast cancer, then 90 percent of the time she will have a positive mammogram. If a woman does not have breast cancer, the probability is 0.07 that she will still have a positive mammogram. Suppose a woman has a positive mammogram. What is the probability that she actually has breast cancer?

Solution: Natural Frequencies Approach (Gigerenzer's Approach). We start with a simple, intuitive approach, bypassing Bayes' method since often times people confuse the conditional probability that A occurs given B, $P(A|B)$, with the conditional probability that B occurs given A, $P(B|A)$. Another common mistake is that people mistake $P(A|B)$ with $P(A,B)$. We will discuss these more in the next chapter.

Let BC denote the event that a woman has breast cancer and let M^+ denote the event that a woman receives a positive mammogram. We are interested in finding $P(BC|M^+)$.

From our assumptions, we know P(BC) = 0.008. That is, 8 out of 1000 women have breast cancer. Also, 90 percent of 8 women have a positive mammogram which implies that approximately 7 women will have a positive mammogram. Of the remaining 992 women who don't have breast cancer, $(992)(0.07) \approx 70$ have a positive mammogram. Thus, given a sample of women who have a positive mammogram screening, what percentage of them have breast cancer? Since we found that $7 + 70 = 70$ have a positive mammogram, this means

$$P(BC|M^+) = 7/77 \approx 9\%.$$

> *Comment*: Although using this approach is easier and more intuitive than using Bayes' Rule, note that we have arrived at our answer by a few assumptions. First, the assumption that 8 out of 1000 women have breast cancer isn't necessarily true. In reality, events don't have to follow their corresponding probabilities. For example, if we flipped a coin 1000 times, we should not expect to get exactly 500 heads. We also rounded to whole numbers, so our final answer was

an approximation. Keep these ideas in mind when solving problems this way.

Alternative Solution: Bayes' Rule
Recall Bayes' Rule from STA 2023:

$$P(A|B) = P(A, B)/P(B).$$

We want to calculate

$$\begin{aligned}P(BC|M^+) &= P(BC, M^+)/P(M^+)\\ &= P(M^+|BC)P(BC)/P(M^+)\\ &= (0.9)(0.008)/P(M^+)\end{aligned}$$

Also,

$$\begin{aligned}P(M^+) &= P(M^+|BC)P(BC) + P(M^+|\text{no } BC)P(\text{no } BC) \qquad (1.1)\\ &= (0.9)(0.008) + (.07)(1 - .008)\\ &= 0.07664\end{aligned}$$

Plugging back in to $P(BC|M^+)$, we find

$$P(BC|M^+) = 0.0072/.07664 = 0.0939.$$

In equation (1.1), we are using the law of total probability, which recall says that

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

We will review this more completely in Chapter 2.

1.2 Introduction to Bayesian Inference

1.2.1 Simple Bayesian Example

Suppose you are a biology student submitting your first paper to a journal and you have assessed your chances of your paper being published. Suppose this particular journal has a particularly low acceptance rate of 20 percent, but your paper ended up being accepted! What is your updated assessment that your *next* submission will be accepted (assume the topic is similar)?

Using intuition, we know that the direct estimate or frequentist estimate is 100 percent since we only have one sample, however, after thinking more, this answer seems very naive given that we know this journal's acceptance rate is 20 percent. Thus, it makes sense to pick a number smaller than 100 percent. If you do this, you are behaving like a Bayesian because you are adjusting the direct estimate due to knowing some *prior* information.

That is, it makes sense to incorporate the prior information, i.e., in this case the acceptance rate of 20 percent into our calculation so that our answer will be more accurate. We will explore this idea in more detail soon. The important point is that you understand the basic idea behind the example.

1.2.2 Frequentist versus Bayesian

In statistics, there are two approaches that are mainly used by researches—frequentist and Bayesian. The *frequentist* approach evaluates procedures based on sampling from a particular model (the likelihood) repeatedly. The likelihood defines the distribution of the observed data conditional on the unknown parameter(s). For example, we denote the likelihood as $L(X_1, \ldots, X_n|\theta)$, where the X_i's are the data and θ is the parameter we are trying to estimate.

Example 1.4: Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ where σ^2 is known. From STA 2023, we know that $\bar{X}$ (the sample average) estimates μ well. We can show (using calculus) that $\bar{X}$ is the frequentist estimate for μ in this situation.

Moving along to the *Bayesian* approach, this requires sampling a model (likelihood) and also knowing a prior distribution on *all* unknown parameters in the model. In the easiest case, where we only have one unknown parameter, the likelihood and prior are combined in such a way that we compute the distribution of the unknown parameter given the data (*posterior distribution*). We'll write out some notation for what this means later. For now, it's important to understand the basic idea.

We have briefly discussed the two approaches. We will outline how both conceptually differ from each other so it is very clear.

Bayesian Methods

- The parameter is considered to be a random variable and has a distribution.
- The prior distribution placed on the unknown parameter quantifies our beliefs regarding the unknown parameter.
- We use the laws of probability to make inferences about the unknown parameter of interest.
- We update our beliefs about the unknown parameter after getting data (likelihood). This yields the posterior distribution which reweights things according to the prior distribution and the data (likelihood).
- The Bayesian approach makes sense even when we treat the experiment as if it is only occurring one time.

Frequentist Methods

- For a frequentist, the parameter is fixed but an unknown constant.
- The probabilities are always interpreted as long-run relative frequencies.
- Statistical procedures are judged by how well they perform in the long-run over some infinite number of repetitions of the experiment.

1.2.3 Frequentist or Bayesian

Let's consider why both models might have criticism.

Bayesian Criticisms

- Bayesian methods require choosing some prior with *known parameters*. One question that is often asked is how to choose the prior as well as the prior parameters.

- Bayesians are often criticized for choosing priors out of *convenience*. Such priors are called conjugate priors and and allow us to compute the posterior distribution very easily as we will soon see.

- Bayesian methods are sometimes not used because inference may depend on the choice of a specific prior. This causes a lack of model *robustness*.

- Recent computational advances (WinBUGS and using MCMC) have basically eliminated any problems in the past regarding choice of priors, however difficulties of prior selection and possible non-robustness remain.

Frequentist Criticisms

- Bayesians have criticized frequentist methods for failure to incorporate prior information, which is often available in problems today.

- Frequentist methods are often also avoided due to inefficiency, inflexibility, and incoherence. By incoherence, we mean that these methods fail to process information systematically.

- Another criticism is that even though frequentists avoid dependence on prior beliefs, they still require some set of simple model assumptions must be made that are free of confounding, selection bias, measurement error, etc.

Since frequentist methods are the ones stressed in undergraduate studies whether you realize it or not, for the remainder of this course we will concentrate on learning Bayesian methods. Bayesian methods are becoming more and more popular in industrial work and in areas outside of statistics, so it's important that you have a firm understanding of Bayesian methods, how to apply them, and how they differ from the frequentist approach.

However, before we can introduce any Bayesian notation or methods, we will need to review many probability concepts to make sure we are familiar with some statistical tools.

Chapter 2

Probability and Random Variables

We'll review some concepts that are purely frequentist in nature as well as cover some ideas that you may have not seen before that will be useful in future chapters.

When we seek to calculate a probability, we are wanting to measure our belief in the occurrence of a future event. For example, if we flip a coin, we do not know for certain if it will land on heads, but we have learned from introductory courses that on average it will land on heads half the time and on tails the other half. When we speak of random occurrences that cannot be predicted with certainty, we say that these events are *random* or *stochastic.*

Why do we need to learn probability theory? Consider a gambler who is interested in whether or not a (six-sided) die is balanced. If the die is perfectly balanced and we did millions of experiments, we would expect that each side would come up with probability 1/6. However, the gambler sets out to prove that the die is loaded. He takes a sample of ten tosses, all resulting in 1s. The gambler decides from this that the die must be loaded. But can we be sure? The gambler decided that the die was loaded not from using calculations, but from intuition. He decided that it wasn't *impossible* to get ten 1s but *improbable.*

Probability theory will provide us with a rigorous method for finding a number that will agree with the relative frequency of occurrence of some event in a long series of trials. We can see how this is needed if, for example, the gambler had instead rolled five 1s, two 2s, one 3, one 4, and one 6 instead of ten 1s. Using intuition alone, how can we make a decision regarding whether or not the die is balanced? Thus, in this chapter we will create a foundation necessary for computing these quantities so we aren't sitting around guessing like our friend the gambler.

2.1 First Principles

Definition 2.1: A *sample* is a group of subjects for which we have data and about which we want to study information. A sample might be 18 UF psychology majors, randomly chosen by UFID or 100 possible flips of the quarter in my pocket.

Definition 2.2: The group of subjects that is the target of interest is known as the *population*. Examples of populations include the following:

- All UF biology majors
- All adults that smoke in the United States
- All possible flips of the quarter in my pocket

Definition 2.3: A *parameter* is a measurement describing some characteristic of a population.

Definition 2.4: A *statistic* is a numerical measurement describing some characteristic of a sample. Let $x_1, x_2, \ldots, x_n$ be a sample of measured values.

Definition 2.5: The *sample mean* is defined as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The *population mean* is denoted by μ.

Definition 2.6: The *sample median* is the middle value when the original data values are arranged in order of increasing (or decreasing) magnitude. We'll denote the median by $\tilde{x}$.

Example 2.1: Suppose we have observations that are already ordered: $2, 5, 7, 10, 49$. Then 7 is the median. However, if we have an even number of observations, such as 2, 4, 6, 10. Then the median is the average of the middle two observations. The median here is $10/2 = 5$. □

Definition 2.7: The *sample variance* is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

The *population variance* is denoted by σ^2.

Example 2.2: (Aqua Running) Aqua running has been suggested as a method of cardiovascular conditioning for injured athletes and others who desire a low-impact aerobics program. In a study to investigate the relationship between exercise cadence and heart rate, the heart rates of 20 healthy volunteers were

87	109	79	80	96	95	90	92	96	98
101	91	78	112	94	98	94	107	81	96

measured at a cadence of 48 cycles per minute (a cycle consists of two steps). The data are as follows:

Calculate the sample mean and sample variance.

Solution: The sample mean is simply

$$\bar{X} = \frac{87 + 109 + 79 + \ldots + 96}{20} = 1874/20 = 93.7.$$

The sample variance is

$$\frac{1}{20}[(87 - 93.7)^2 + \ldots + (96 - 93.7)^2] = 91.27.$$

At the end of the chapter, we will look at how to do computations such as these in `R`. □

2.2 Properties of Probability

In order to calculate probabilities of events, we will need to establish some notation.

Definition 2.8: An *experiment* is a process where we record an observation. Examples of experiments include flipping a coin, tossing a die, measuring the IQ score of a student, or administering a drug to patient and asking whether the pain is relieved.

> Remark: You may notice this definition is somewhat broader than the one you learned in STA 2023. For convenience, any data that is collected in a random way will be referred to as an experiment.

We want to calculate the probability of *something* happening. We call this *something* an *event*. If we're speaking of one event, we'll denote the event by A.

Definition 2.9: The *sample space*, S, associated with an experiment is the set of all possible outcomes.

Definition 2.10: The *outcome* is the result achieved from running one trial of an experiment.

Definition 2.11: Some particular set of outcomes that we are interested in is called an *event*.

Recall from your introductory statistics course that if each outcome is equally likely, then we can easily find the probability of event A, denoted by $P(A)$. We find it by

$$P(A) = \frac{\#\text{ of outcomes in } A}{\text{total } \#\text{ of outcomes in sample space } S}.$$

So far, our definitions above may seem a bit abstract. Let's look at an example to make things more clear.

Example 2.3 (Rolling a Die)**:** Consider rolling a six-sided die on a flat surface. The die's sides are labeled 1,2,3,4,5,6. Suppose we are interested in betting on the roll being a 6. Then there are 6 possible outcomes (rolling a 1,2,3,4,5, or 6). Each outcome listed above is also an event. For example, rolling a 3 is an event. In fact, there are more than 6 events.

To see this, suppose we are instead interested in betting that the roll is odd. We denote this event by $\{1, 3, 5\}$. As we can see, there are many events that can be considered. Or we could be interested in not rolling a 6. We denote this by 6^c. Then $6^c = \{1, 2, 3, 4, 5\}$. This notation is called the *complement* of 6. □

Definition 2.12: More formally, the *complement* of event A, denoted by A^c, consists of all the outcomes in the sample space S that are not in A.

Definition 2.13: Let S be the sample space associated with an experiment. To every event A in S, we assign a number, $P(A)$, called the *probability of* A. The following results hold:

1. $P(A) \geq 0$
2. $P(S) = 1$
3. If events A and B are mutually exclusive (i.e., have no outcomes in common), then $P(A \cup B) = P(A) + P(B)$.

   Remark: Since $P(A) \geq 0$ and $P(S) = 1$, we know

   $$0 \leq P(A) \leq 1.$$

   Why? S is the entire sample space, so A must live somewhere in S. This means that $A \subset S \implies P(A) \leq P(S) = 1$.

Theorem 2.1: $P(A) = 1 - P(A^c)$.

Proof: Note that $S = A \cup A^c$. Also, A and A^c are mutually exclusive, so by above,

$$1 = P(S) = P(A \cup A^c) = P(A) + P(A^c).$$

Definition 2.14: We use $\emptyset$ to denote an impossible event. Naturally, P($\emptyset$) = 0.

Example 2.4: (Coaching Jobs) Four equally qualified people apply for two coaching jobs on the Florida football team. One and only one applicant is a member of a minority group. The positions are filled by choosing two of the applicants at random.

1. What are the possible outcomes for this experiment?
   Denote the applicants by person A, B, C, and D. Then the possible outcomes are all the combinations of choices of the candidates, i.e.,

   $$\{A, B\}, \{A, C\}, \{A, D\}, \{B, C\}, \{B, D\}, \{C, D\}.$$

2. Assign reasonable probabilities to each of the sample points. Since two applicants are chosen at random, each outcome has an equally likely probability of being chosen, which is 1/6.

3. Find the probability that the applicant from the minority group is selected for a position. Suppose person A is from a minority group. Person A is in 3 of the 6 possible outcomes. So, $P(\text{Person } A \text{ selected}) = 3/6 = 1/2$.

□

Example 2.5: (Coin Tossing) A balanced coin is tossed three times. Calculate the probability that exactly two of the three tosses results in heads. Denote this event by A.

Solution: Denote heads by H and tails by T. Let's list out the set of all possible outcomes. The first letter we write will represent the outcome of the first flip, the second letter will represent the outcome of the second flip, etc. Assume that the coin is fair.

There are eight possible events which are

$$HHH, HHT, HTT, HTH, TTT, TTH, THH, THT.$$

Since we assume the coin is fair, let's now list out the outcomes that have two of the three tosses with heads. They are

$$HHT, HTH, THH.$$

Thus, $P(A) = 3/8$. □

2.3 Conditional Probability and Independence

2.3.1 Probability Motivation

In the last section, we looked at an example of flipping a fair coin three times. Suppose we are instead interested in the probability of the following:

- getting a H on toss three given that first two tosses were HT
- getting a T on toss three given that the first two tosses were HH

This illustrates the Bayesian idea that we have prior information to condition on that we should use to calculate our probability. Let's see the simplest form of how this prior information is used.

Remember that Bayesian probability conditions on the probability of knowing information, while classical statistics views probability as the frequency after a hypothetical set of trials.

Definition 2.15 (Conditional Probability):

$$P(A|B) = P(A, B)/P(B) \quad \text{provided} \quad P(B) > 0.$$

This easily extends to what is called the *Multiplicative Law of Probability*, i.e.,

$$P(A,B) = P(A|B)P(B).$$

Then we can extend this to find the probability of the intersection of any number of events.

Theorem 2.2 (Additive Law of Probability)**:**

$$P(A \cup B) = P(A) + P(B) - P(A,B).$$

Definition 2.16 (Independence): Events A and B are *independent* if *any* of the following holds:

$$P(A|B) = P(A).$$
$$P(B|A) = P(B).$$
$$P(A,B) = P(A)P(B).$$

To see why for example, $P(A,B) = P(A)P(B)$ is equivalent to $P(B|A) = P(B)$ when A and B are independent, simply recall the conditional probability formula:

$$\begin{aligned} P(B|A) =& P(B,A)/P(A) \\ &= P(A)P(B)/P(A) \\ &= P(B). \end{aligned}$$

Examples of independence include the following:

- Tosses of the same fair coin. Once we toss the coin once, on the next toss, the coin forgets about the last toss.
- Drawing items out of a bag with replacement.

Theorem 2.3 (Law of Total Probability)**:** Assume that we have sets $A_1, A_2, \ldots, A_k$ such that

1. $S = A_1 \cup A_2 \cup \cdots \cup A_k$.
2. A_i and A_j are mutually exclusive for all $i \neq j$.

Then $\{A_1, A_2 \ldots, A_k\}$ are said to partition S. If $P(A_i) > 0$ for all i then for any event B,

$$P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i).$$

As a special case, $S = A \cup A^c$ and suppose $P(A) > 0, P(A^c) > 0$. We know A and A^c are mutually exclusive. Then

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c).$$

Example 2.6: (Aces) Suppose we are playing cards and we're interested in the probability of getting certain cards. Let A represent the event that the first card we draw is an ace. Let B represent the event that the second card we draw is an ace. What is the probability that we select two aces?

Solution: That is, find $P(A, B) = P(B|A)P(A)$ by the conditional probability. $P(B|A) = 3/51$ since we have only have 3 aces left and 51 cards total remaining. Also, $P(A) = 4/52$. Then $P(A, B) = P(B|A)P(A) = (3/51)(4/52) = 1/221$. Think about calculating the probability of getting three aces at home.

What is the probability that the second card is an ace? That is, we are looking for $P(B)$. Naturally, we will apply the law of total probability in order to calculate $P(B)$. $P(B) = P(B|A)P(A) + P(B|A^c)P(A^c)$. We are apply the law of total probability by conditioning on whether or not we had an ace on the first draw.

We have already calculated the first part of the expression above which is 1/221. The second part is very similar. We find that $P(B|A^c)P(A^c) = (4/51)(48/52) = 16/221$.

Putting this all together, we know $P(B) = 1/221 + 16/221 = 17/221 = 1/13$. □

The next theorem is the fundamental theorem that led to Bayesian statistics. We'll go through the result and then work through some examples.

Theorem 2.4 (Bayes' Rule)**:**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}.$$

Proof: $P(A|B) = P(A, B)/P(B) = P(B|A)P(A)/P(B)$. Now apply the law of total probability to $P(B)$ to get the result.

> Remark: Notice that Bayes' Rule is just an application of the law of conditional probability and the law of total probability. If you need to use this for a problem, you could simply just use these two laws instead of memorizing Bayes' Rule. However, do what's easiest for you.

Example 2.7: A diagnostic test for a disease is advertised such that it correctly detects a disease in 90 percent of the individuals who actually have the disease. Also, if a person does not have the disease, the test will report that he or she does not have it with probability 0.9. Only 1 percent of the population has the disease in question. If Jack is chosen randomly from the population and the diagnostic test indicates that he has the disease, what is the conditional probability that he really has the disease? Are you surprised by the answer? Would you call this test reliable?

We want to calculate $P(D|PT)$. We'll let D denote disease and ND denote no disease. Similarly, PT will stand for a positive test and NT will denote a negative test. From the information, we know the following:

$$P(PT|D) = 0.9, P(NT|ND) = 0.9, \text{ and } P(D) = 0.01.$$

Also, since $P(NT|ND) = 0.9$, this implies $P(PT|ND) = 0.1$.

Then

$$\begin{aligned}
P(D|PT) &= \frac{P(D, PT)}{P(PT)} \\
&= \frac{P(PT|D)P(D)}{P(PT|D)P(D) + P(PT|ND)P(ND)} \\
&= \frac{(0.9)(0.01)}{(0.9)(0.01) + (0.1)(1 - 0.01)} \\
&= 1/12 \approx 0.083.
\end{aligned}$$

Alternative Solution: Natural Frequencies Approach

Again we want to find $P(D|PT)$.

1. Since $P(D) = 0.01$, we assume that 10 out of 1000 people have the disease. (We could assume 1 out of 100, but we'd like to take a large sample than one person).

2. 90% of 10 people have a positive test, which implies that 9 people have PT.

3. There are 991 people remaining who don't have the disease and $P(PT|ND) = 0.1$. So, $(991)(.1) \approx 99$ people who have PT.

4. To summarize, we have found 108 people who have a PT. Thus, $P(D|PT) = 9/108 \approx 0.083$.

This test clearly isn't really worth that much since it doesn't give us that much information. However, it could be used as a screening test for a disease if the test is cheap and quick. For example, suppose that there is a $\$15,000$ test for the disease and $P(D|PT)$ is very high. The downside here is that the test is very expensive. The solution is to have patients do the cheap test, where a patient has about an 8 percent chance of having the disease given the test is positive. If the screening test comes back positive, then it makes more sense to have the patient do the more expensive test. So, the test in the above example isn't completely worthless to us.

2.4 Discrete Random Variables and Their Probability Distributions

When we are performing experiments, we often find that we are interested in functions of an outcome instead of the outcome itself. For example, suppose we are tossing two dice. We might be interested in the sum of two dice being seven instead of the outcome being (1,6), (2,5), etc. *Random variables* are real-valued functions defined on the sample space. The value of a random variable is determined by the outcome of some trial, so naturally probabilities will correspond to possible values of each random variable.

2.4.1 Discrete Random Variables

Definition 2.17: A *random variable* Y is said to be discrete if it can assume only a finite or countably infinite number of distinct values.

> Remark: Suppose Y is countably infinite. This means that the value of Y can assume values that you could count if you had an infinite amount of time.

We now have a random variable Y, so we can think about the probability that Y takes some real value y. We denote this by $P(Y = y)$. Keep in your mind that Y is a random variable, but y is a fixed quantity.

Definition 2.18: The probability that Y takes on value y, $P(Y = y)$, is defined as the sum of the probabilities of all the sample points in S that are assigned the value y. We will sometimes denote $P(Y = y)$ by $p(y)$.

Definition 2.19: The *probability distribution* for a discrete random variable Y can be represented by a formula, table, or graph that provides $p(y) = P(Y = y)$ for all y.

Theorem 2.5: For any discrete probability distribution, the following must be true:

1. $0 \leq p(y) \leq 1$ for all y.
2. $\sum_y p(y) = 1$, where we sum over all values of y with positive probability.

To represent the probability distribution of Y, we will usually form a table where we find $p(y)$ for each value of y in the sample space. The usual way to find $p(y)$ is to identify the possible values of y. Then calculate $p(y)$ for each value of y. This is made more clear by the following example:

Example 2.8: (Urn Example) Suppose we have an urn containing five balls. Each ball is labeled 1 to 5. Two balls are randomly selected from the urn. Find the probability distribution for the sum of the two sampled numbers.

Solution: Let Y = the sum of the two sampled numbers. Let's first list out all the possible outcomes (20 of them).

(1,2), (1,3), (1,4), (1,5)
(2,1), (2,3), (2,4), (2,5)
(3,1), (3,2), (3,4), (3,5)
(4,1), (4,2), (4,3), (4,5)
(5,1), (5,2), (5,3), (5,4)

We're not interested in the outcomes above, but in the distribution of Y. From the outcomes above, it's easy to see what values Y can take. Y can be 3,4,5,6,7,8, or 9. Then using the information above, we find $p(y)$ in the table below. Notice in the table that all the probabilities add up to 1. This isn't a coincidence. This will always occur and is something you should check to make sure that you have a valid probability distribution. □

y	$p(y)$
3	2/20
4	2/20
5	4/20
6	4/20
7	4/20
8	2/20
9	2/20

Table 2.1: Urn Example

Definition 2.20: Let Y be a discrete random variable with probability distribution $p(y)$. Then the *expected value* of Y is defined to be

$$E(Y) = \sum_y y p(y).$$

Definition 2.21: Suppose Y is a discrete random variable with $E(Y) = \mu$. Then the *variance* of Y is defined as

$$V(Y) = E[(Y - \mu)^2] = \sum_y (y - \mu)^2 p(y).$$

Theorem 2.6: Suppose Y is a discrete random variable with $E(Y) = \mu$. The variance of Y can also be expressed as

$$V(Y) = E(Y^2) - E(Y)^2.$$

> Remark: The above result is often very useful in calculating the variance instead of using the definition.

Example 2.9: (Urn Example Revisited) Let's look back at the urn example where we found the following probability distribution earlier:

y	$p(y)$
3	2/20
4	2/20
5	4/20
6	4/20
7	4/20
8	2/20
9	2/20

Table 2.2: Urn Example

Let's calculate $E(Y)$ and $V(Y)$.

Solution: Recall that Y represented the sum of the two sampled balls from the five urns.

$$\begin{aligned} E(Y) &= \sum_y yp(y) \\ &= 3(2/20) + 4(2/20) + 5(4/20) + 6(4/20) + 7(4/20) + 8(2/20) + 9(2/20) \\ &= 6. \end{aligned}$$

In order to calculate, $V(Y)$ we must find $E(Y^2)$.

$$\begin{aligned} E(Y^2) &= \sum_y y^2p(y) \\ &= 3^2(2/20) + 4^2(2/20) + 5^2(4/20) + 6^2(4/20) + 7^2(4/20) + 8^2(2/20) + 9^2(2/20) \\ &= 39. \end{aligned}$$

Therefore,

$$V(Y) = E(Y^2) - E(Y)^2 = 39 - 6^2 = 39 - 36 = 3.$$

□

We have given the definitions for calculating the mean and variance of discrete random variable. We can also use `R` to calculate the mean and variance for discrete random variables. We will eventually see that we can do this for continuous random variables as well. In the meantime, we will next discuss some discrete distributions that will be relevant in upcoming chapters.

2.4.2 Discrete Distributions

Now that we have covered what random variables are and what probability distributions are in the discrete case, it will be helpful to look at some specific ones that will come up in the upcoming chapters. They include the Binomial, Poisson, and the Geometric distributions.

Binomial Distribution

Here is the setup:

- X represents the result of a trial (say flipping a coin) that results in a success or failure.
- Let p denote a success (and so $1-p$ represents a failure).
- We decide to run an experiment for 1 trial to start.

Then $P(X=x) = p^x(1-p)^{1-x}$. This is called a *Bernoulli trial.*

For a Binomial experiment, we have n trials instead, and we assume that the trials are independent. To calculate $P(X=x)$, think about flipping a coin, and suppose a head is a success. We are interested in finding the probability that we get x heads, where the $P(H)=p$. Since the trials are independent, we know that the probability we get x heads is just p^x. Similarly, if we have x heads, we must have $n-x$ tails, and the probability of this is $(1-p)^{n-x}$. So we might think that

$$P(X=x) = p^x(1-p)^{n-x},$$

but this isn't quite right. We need to take into account the different orderings that could occur. Recall that to take into account ordering we simply use $\binom{n}{x}$, which counts the numbers of ways to place x heads when we have n spots available. Once these are placed, the tails are fixed. Thus, the *Binomial distribution* is defined as:

$$p(x) = \binom{n}{x} p^x(1-p)^{n-x}, \; x = 0,1,2,\ldots,n \;\text{ and }\; 0 \le p \le 1.$$

> Remark: $E(X) = np$ and $V(X) = np(1-p)$. You don't need to know how to derive the expectation and variance.

Poisson Distribution

Suppose that X is the number of rare events that occur over time and λ is the average value of X. For example, we might be interested in the number of accidents in a given time interval or the number of radioactive particles that decay over a particular time period.

A random variable Y has a *Poisson* distribution if and only if

$$p(y) = \frac{\lambda^x e^{-\lambda}}{x!}, \;\; x = 0,1,2,\ldots, \;\; \lambda > 0.$$

> Remark: $E(X) = \lambda$ and $V(X) = \lambda$. You don't need to know how to derive the expectation and variance.

The Poisson distribution can be derived as an approximation to the binomial distribution when n is large, p is small, and np is not too large and not too small. In this case, define $\lambda = np$. Consider each piece of the binomial distribution.

n is much larger than x, so

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} = \frac{\overbrace{n(n-1)(n-2)\cdots(n-x+1)}^{x \text{ terms}}(n-x)!}{x!(n-x)!} \approx \frac{n^x}{x!}.$$

Also, again because n is much larger than x,

$$(1-p)^{n-x} \approx (1-p)^n = \left(1 - \frac{np}{n}\right)^n = \left(1 - \frac{\lambda}{n}\right)^n \approx e^{-\lambda}.$$

Then the binomial distribution becomes

$$\binom{n}{x} p^x (1-p)^{n-x} \approx \frac{n^x}{x!} p^x e^{-\lambda} = \frac{(np)^x e^{-\lambda}}{x!} = \frac{\lambda^x e^{-\lambda}}{x!},$$

which is the Poisson distribution.

Geometric Distribution

Suppose the probability of a success is equal to p and this is true for each trial. We run an experiment that involves independent and identical trials, each of which can result in a success or a failure. Suppose we have n trials and the geometric random variable Y is number of the trial on which the first success occurs.

Suppose we want to calculate $P(Y = 1)$? Clearly, P(Y=1) = p since we must get a success. What about calculating $P(Y = 2)$? Here, we are calculating the probability our first success is on trial 2, which means the first trial was a failure. This means we ended up with FS. Thus, $P(Y = 2) = p(1-p)$. What about the other events that could occur with $3, 4, 5$, etc. trials. Denote these events by E_3, E_4, E_5, etc. Then

$$\begin{aligned} &E_3 : FFS \\ &E_4 : FFFS \\ &E_5 : FFFFS \\ &\quad\vdots \\ &E_k : \underbrace{FFF\ldots F}_{k-1}S \\ &\quad\vdots \end{aligned}$$

So, A random variable Y is said to have a *geometric probability distribution* if and only if

$$p(y) = P(Y = y) = (1 - p)^{y-1}p, \;\; y = 1, 2, 3 \ldots, \;\; 0 \leq p \leq 1.$$

> Remark: $E(X) = \frac{1}{p}$ and $V(X) = \frac{1-p}{p^2}$. You don't need to know how to derive the expectation and variance.

There are many other discrete probability distributions, but we have just covered a few.

2.5 Continuous Random Variables and Their Probability Distributions

We just finished discussing random variables that could take on a finite number of values or a countable number of values. We now discuss what we mean by a continuous random variable.

2.5.1 Continuous Random Variables

For example, if you think about random variables we encounter in the real world, ones of interest are not just discrete. The number of days that it rains is a discrete random variable because it must take a value $0, 1, 2, \ldots, n$. However, if we think about daily rainfall in Alachua County, it can be anywhere from 0 inches to 5 inches perhaps in a given month. A random variable that can take on any value in any interval is considered continuous. Now we will look at these types of distributions.

Definition 2.22: A *continuous random variable* Y can take on a continuum of possible values. An example of this would be Y representing the lifetime of a car, where the lifetime is assumed to take any value in some interval $[a, b]$.

In this section, we will gloss over some concepts known as limits and derivatives since calculus is not assumed in this course. What we would like to talk about is the *probability density function* $f(x)$ that corresponds to some random variable X.

> Connection to Calculus: If you have had calculus before, in this course a random variable X will always have a distribution function $F(X) = P(X \leq x)$ as well as a density function $f(x)$. Recall the relationship between these two from your calculus course that $f(x) = F'(x)$. If you haven't had any calculus before, just ignore these comments.

Theorem 2.7: If $f(x)$ is a *probability density function* for a continuous random variable X, then

1. $f(x) \geq 0$ for all $x, -\infty < x < \infty$.
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$.

> Remark: The first equation simply says that the density must be nonnegative for all values of x. The second involves an integral. In non-calculus language, this says that if we look at the area underneath the density, it must be equal to 1. For example, think about a standard normal density from your STA 2023 course (a bell shaped curve). The area underneath that density curve equals 1.

Definition 2.23: Let Y be a continuous random variable with probability density $f(y)$. Then the *expected value* of Y is defined to be

$$E(Y) = \int_y y f(y)\, dy.$$

Definition 2.24: Suppose Y is a continuous random variable with $E(Y) = \mu$. Then the *variance* of Y is defined as

$$V(Y) = E[(Y - \mu)^2] = \int_y (y - \mu)^2 f(y)\, dy.$$

Theorem 2.8: Suppose Y is a continuous random variable with $E(Y) = \mu$. The variance of Y can also be expressed as

$$V(Y) = E(Y^2) - E(Y)^2.$$

Notice that this theorem holds for continuous and discrete random variables.

> Remark: The above result is often very useful in calculating the variance instead of using the definition.

2.5.2 Continuous Distributions

We have covered many discrete distributions already. We will now look at some continuous distributions that will be of use to us in future chapters. They include the Uniform, Normal, and Beta distributions.

Uniform Distribution

Suppose a bus arrives between 9:00 and 9:10 every morning and the probability that the bus will arrive in any subinterval of time is proportional to the length of the subinterval. For example, the bus is as likely to arrive between 9:02 and 9:04 as 9:08 and 9:10. Let X denote the length of time a person must wait for the bus if the person arrived at the bus stop at 9:00 a.m. sharp.

If we carefully noted how long (in minutes) the person had to wait after 9:00 a.m. for several mornings, we could construct a relative frequency histogram for the data. We would expect to see from this histogram that the relative frequency with which we observed a value of Y between 2 and 4 would be approximately the same as the relative frequency for that of 8 and 10.

The random variable X is an example of a *uniform distribution.* If $a < b$, X is said to have a *uniform probability distribution* on the interval $[a, b]$ if and only if the density function of X is

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise.} \end{cases}$$

> Remark: $E(X) = \frac{b+a}{2}$ and $V(X) = \frac{(b-a)^2}{12}$. You don't need to know how to derive the expectation and variance.

Normal Distribution

The *normal distribution* is the most widely used continuous probability distribution. Recall from STA 2023 that it has a bell-shaped curve centered about its mean μ.

A random variable X has a *normal probability distribution* if and only if for $\sigma > 0$ and $-\infty < \mu < \infty$, the density function of X is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-1}{2\sigma^2}(x-\mu)^2 \right\}, \quad -\infty < x < \infty.$$

> Remark: $E(X) = \mu$ and $V(X) = \sigma^2$. You don't need to know how to derive the expectation and variance.

Beta Distribution

A random variable has a *beta distribution* if and only if the density function of X is

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \;\; 0 \leq x \leq 1, \; \alpha > 0, \; \beta > 0.$$

There are many other discrete and continuous distributions that are useful that we have not covered. As they are needed, we will introduce them as needed.

2.6 A Brief Introduction to R

These instructions will walk you through using R. At first glance, R may seem intimidating, but everything should be easy if you follow the instructions.

What is R?

R is a free and flexible collection of software for statistical computing and graphics. It's designed to work on any operating system, and one of its strengths is that users around the world can write “packages” to implement new statistical techniques as they are developed.

Downloading and Installing R

Obviously what you'll need to download depends on what operating system you're using.

- If you use Windows, download this:
  `http://cran.opensourceresources.org/bin/windows/base/R-2.12.1-win.exe`
- If you use Mac OS X, download this:
  `http://cran.opensourceresources.org/bin/macosx/R-2.12.1.pkg`
- If you use Linux, talk to me, because you're going to need a lot more help to get things set up (but you can still do it).

Double-click the file you downloaded to start the Installer. Just follow along with the default installation options until the Installer finishes. You should now either see `R.exe` in your `Program Files` area (Windows) or see `R.app` in your `Applications` area (Mac OS X). This is what you double-click to start R.

The R Console

Once you start R, the main thing you see is a window called the R Console. The window contains a command prompt `>` with a blinking cursor. One way to use R is to type commands here. Try typing `2+2` at the prompt and pressing Enter. You should see

```
> 2+2
[1] 4
>
```

You can see that R has calculated your answer, `4` (ignore the `[1]` before it), and is now ready for another command.

R Source Files

Entering long lists of commands into the Console is not the most efficient way to use R. Instead, it's easier to type up a list of commands that we can execute all at once and that we can also save to run again later (or for someone else to run). We do this by making something called an R source file.

You can create a new R source file by pressing Ctrl+N (Windows) or Command-N (Mac). You will notice during the semester that I will post R source files on the course website that we go through in class so you can look at them later on.

You can simply save these source files to your machine and run them in R and you will notice that the output is the same from class.

Executing R Source Files

To run all the commands in an R source file, open it up, select all the commands in the file by pressing Ctrl+A (Windows) or Command+A (Mac), and then press Ctrl+R (Windows) or Command+Enter (Mac) to run everything. The results will be displayed in the R Console.

Example Calculations in R

The R source file `rhelp.R` from the course website will show you various calculations that you can do in R. I have outlined the commands from the source file below so that you can follow along.

Functions

Many mathematical and statistical functions are available in R. We can add or subtract numbers, multiply and divide, or use functions like square root or the exponential function. Running the R source file, we find

```
> 2+2
[1] 4
> (2-4)*5
[1] -10
> sqrt(9)
[1] 3
> exp(1)
[1] 2.718282
```

Assignment

It is often very convenient to give a value a name that we intend to use very frequently. For example, if we have 100 trees that we have sampled, we would make this assignment. We can always adjust our assignment at anytime, which is most convenient.

```
> trees = 100
> trees + 5
[1] 105
```

Setting the Working Directory

The working directory is where R looks for files and where it saves any files that it creates. You can select any directory on your computer to be the working directory. There are two ways to set the working directory:

- Select Miscellaneous > Change Working Directory... from the menu at the top of R. Then select your desired directory.

- Execute the command setwd("MyDirectory") in R or at the beginning of your source file. Note that exactly what MyDirectory should look like will depend on whether you have a PC or a Mac.

Reading in a File

Often we will want to read a table of data from our working directory into R. The most common way to do this is with the read.table function. This function can be used when the entries in the data table are separated by white space (and do not contain any white space themselves). There may or may not be a header at the top specifying variable names; this should be indicated by using the header=T or header=F option.

When we read in a table, R stores it in an object called a data frame. To use the variables stored in the data frame, we must attach the data frame.

```
olympics = read.table("olym-400.txt",header=T)
attach(olympics)
```

R has a variety of functions similar to read.table for other types of file formats, such as read.csv for comma-separated files and read.fwf for fixed-width table files (where the table is built using ”spaces” that make the columns line up). We won't be using these in this class.

Once, we have read the data in, we can look at variable names directly by calling them now by their name since we have attached the data. We can also apply functions to our data and manipulate it so as to examine our data. For example, we can type olympics in R, which in return will output the data. Alternatively, we can type year, which will list the data for years out in R. If we want a generic summary output from R we can type summary(olympics), which prints information for each variable such as the minimum, maximum, mean, *etc.*. Finally, we calculate the mean, variance, or median of years using the functions mean, var, median. After we're done with the data frame, we should detach it in case we want to re-use the same variable names for new data later.

```
> setwd("~/Desktop/sta4930/rhelp/")
> olympics = read.table("olym-400.txt",header=T)
> attach(olympics)
> olympics
   years times
1   1896 54.20
2   1900 49.40
3   1904 49.20
4   1908 50.00
5   1912 48.20
```

```
6   1920 49.60
7   1924 47.60
8   1928 47.80
9   1932 46.20
10  1936 46.50
11  1948 46.20
12  1952 45.90
13  1956 46.70
14  1960 44.90
15  1964 45.10
16  1968 43.80
17  1972 44.66
18  1976 44.26
19  1980 44.60
20  1984 44.27
21  1988 43.87
22  1992 43.50
23  1996 43.49
24  2000 43.84
25  2004 44.00
26  2008 43.75
> years
 [1] 1896 1900 1904 1908 1912 1920 1924 1928 1932 1936 1948 1952 1956 1960 1964 1968
[17] 1972 1976 1980 1984 1988 1992 1996 2000 2004 2008
> summary(olympics)
     years          times      
 Min.   :1896   Min.   :43.49  
 1st Qu.:1925   1st Qu.:44.06  
 Median :1958   Median :45.50  
 Mean   :1954   Mean   :46.21  
 3rd Qu.:1983   3rd Qu.:47.75  
 Max.   :2008   Max.   :54.20  
> mean(years)
[1] 1954.154
> var(years)
[1] 1234.215
> median(years)
[1] 1958
> detach(olympics)
```

We have presented just a few basic tools that will help you get started in R. For more information, be sure to read the first chapter of *Using R for Introductory Statistics* by John Verzani.

2.7 Exploring an Example in R

We will explore a dataset for 657 (Albert, 2009) students recorded in a spreadsheet and saved as the file "studentdata.txt" with tabs between each field. The first line of the data file contains a header with the variable names. Students at Bowling Green University in Ohio were asked many questions including:

- What is your gender?
- What is your height in inches?
- Choose a whole number between 1 and 10.
- Give the time you went to bed last night.
- Give the time you woke up this morning.
- What was the cost in dollars of your last haircut including the tip?
- Do you prefer water, pop, or milk with dinner?

2.7.1 Setting Things Up

First, download and install R from the web. Then open up a new R file. You'll want to save your R file in some folder/directory on your computer. You'll need to tell R where you've saved your new file. To do so, go to the R menu and select **Misc**. Under **Misc**, select option **Change Working Directory**. Then find the folder that your new file is saved in and select it. You can change the directory within the source file using the `setwd("")` command. Your directory of choice should go within the quotations.

If you are unsure of what the paths are your computer look like, you can use the `getwd("")` command to guide you. If you have additional questions on this, see me after class or in office hours.

Now install the LearnBayes package. We usually read the file into R using the `read.table` command. However, since the file is already in R, we can access the data simply using the `data` command. If we then use the command `studentdata[1,]`, we find that this prints the first row of the data frame. Finally, to make the variable names such as Height visible in R, we much `attach` the data.

So far we have done the following:

```
>install.packages("LearnBayes")
>library(LearnBayes)
>data(studentdata)
> studentdata[1,]
  Student Height Gender Shoes Number Dvds ToSleep WakeUp Haircut Job Drink
1       1     67 female    10      5   10    -2.5    5.5      60  30 water
> attach(studentdata)
```

2.7.2 Analyzing the Data

One categorical variable in this dataset is Drink, which indicates whether students prefer water, pop, or milk with dinner. We can visually see the more popular choice using the `table` command. We can then visualize this using the `barplot` command. To save the plot to a file we use the `pdf` command and close it with `dev.off()`

Now we have the following commands and plot:

```
table(Drink)
t = table(Drink)
pdf(file = "ch2/drink.pdf", width  = 5, height = 4.5)
barplot(t, xlab="Drink", ylab="Count")
dev.off()
```

In the `pdf` command, we give the file a name as well as set the width and height of the file that will be saved. The file will appear in the same directory as your R file. Once you have plotted your function, you must use the command `dev.off` to close the file. Otherwise, you will have problems.

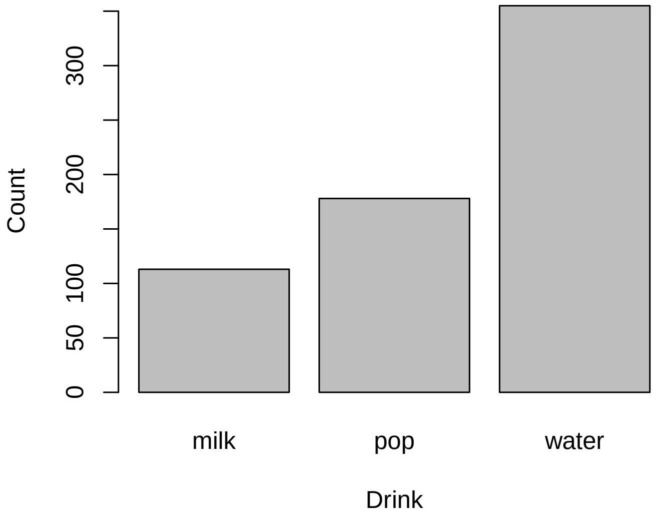


Figure 2.1: Bar plot of drinking preferences of statistics students

From the table, we can see that more than half the students preferred water. Also, pop was more popular than milk.

Suppose next we're interested in how long the students slept the previous night. We didn't ask the students how long they slept, but we can easily calculate this value by subtracting each student's go-to-bed time from his wakeup time. We'll call this new variable `hours.of.sleep`. From the summary statistics computed, we see that students slept 7.5 hours on average (not taking into account the students that didn't answer) and 50 percent of the students slept between 6.5 and 8.5 hours (recall half our data is between the 1st and 3rd quantiles). We can also calculate the sample mean and sample variance of number of hours slept using commands `mean` and `var`. In order to calculate both of these quantities, we need to let R know that some of the data values are NA. We list the commands for our calculations below.

```
> hours.of.sleep = WakeUp - ToSleep
> summary(hours.of.sleep)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  2.500   6.500   7.500   7.385   8.500  12.500   4.000
> mean(hours.of.sleep, na.rm=T)
[1] 7.385191
> var(hours.of.sleep, na.rm=T)
[1] 2.28557
```

Finally, suppose we are interested if the length of sleep is related to the time a student goes to sleep. We can easily plot the data to try and answer this question. We can also put a line of best fit through the data to see if it is linear. If you've seen linear regression, we're fitting a simple linear model to the data in `R` and plotting the regression line. We fit the linear model using the `lm` function, where the dependent variable (y) is the first argument and the independent variable (x) is the second argument. To put the line through the points on the plot we use the `abline` command.

```
pdf(file = "sleep.pdf", width  = 5, height = 4.5)
plot(ToSleep,hours.of.sleep,xlab="time to bed",ylab="length of sleep")
fit = lm(hours.of.sleep˜ToSleep)
abline(fit)
dev.off()
detach(studentdata)
```

From the regression plot, the relationship between the two variables seems to be linear, although we can see the data has high variance.

We could easily explore this dataset more in R, but this shows you an idea of the types of calculations and plots that R can provide us.

2.8 Exercises

1. (4) Give two reasons we might not want to use frequentist methods as well as two reasons we might not want to use Bayesian methods.

2. (8) An elevator with two passengers stops at the second, third, and fourth floors. Suppose it is equally likely that a passengers gets off at any of the three floors.

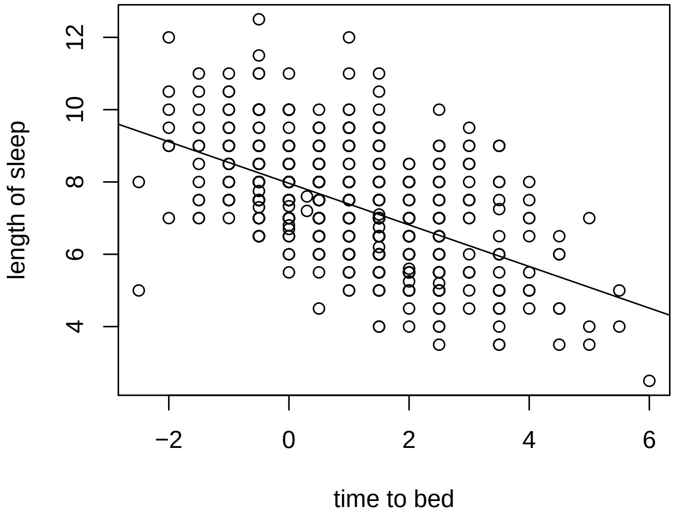


Figure 2.2: Plot of length of sleep versus time to bed

Let a and b denote the two passengers and a_2b_4 denote that a gets off at the second floor and b gets off at the fourth floor, with similar representations for the other cases.

(a) (4) List the sample space.

(b) (4) Let A be the event that the passengers get off at different floors. Find $P(A)$.

3. (12) Suppose we are playing cards using a standard deck containing 52 cards.

   (a) (4) What is the sample space?

   (b) (4) Find the probability of drawing an ace from the deck.

   (c) (4) Let A represent drawing a spade from the deck. Let B represent drawing a face card (K, Q, or J). Find P(A,B).

4. (28) Complete the following example using R. This should be similar to the example we went through in class on Thursday, January 13.

   To receive full credit on this assignment, you must turn in two things: your source code from R and the answers to each part of the questions below (which you will obtain from the R console). Do NOT include your output from the R console.

In writing your source code in R, make sure you do all parts in order. Note that you will no credit if you do not attach your source code from R.

(a) Calculate $120 \times 6 - 150$ in R.

(b) Calvin is 182.88 cm tall. Convert his height to units of meters in R. In doing this calculation, make a variable assignment, using x, so that his height in meters is saved in R.
Hint: This should only be one line of code in R.

(c) Using your variable assignment from part (b), add 10 meters to x. What is the new value of x?

(d) I have uploaded a data file to the Datasets part of the website. It contains data that shows the improvement of 8 students in a speed reading program and the number of weeks they have been in the program. Read the file `reading.txt` into R and assign it the name read. Remember to attach it.

(e) Find the mean and variance of speed gain.

(f) Find the minimum and maximum of speed gain.

(g) What's the last thing you need to do in your R source code? Do this in your code and write a short explanation of what command you need and why it's needed.

5. Suppose that one of the genes associated with the control of carbohydrate metabolism exhibits two alleles—a dominant W and a recessive w. The probabilities of the WW, Ww,and ww genotypes in the present generation are p, q, and r, respectively, for both males and females. Assume that the genotype of the father and the genotype of the mother are independent of each other, e.g., the mother is WW with probability p regardless of the genotype of the father. What are the chances that an individual in the *next* generation will be a ww? Let A denote the event that an offspring receives a w allele from its father. Let B denote the event that it receives the recessive allele from its mother.

6. A box contains one two-headed coin and eight fair coins. One is drawn at random and tossed seven times. Suppose all seven tosses come up heads. What is the probability the coin is fair?

7. Suppose we roll two fair four sided dice. We define the following events:

$$A : \text{the maximum roll is 4}$$
$$B : \text{the sum of the two rolls is 3}$$
$$C : \text{the minimum roll is 1.}$$

(a) Are A and B independent?

(b) Are A and B mutually exclusive?

(c) Give a short proof to show that if two events A and B are mutually exclusive, they must be dependent provided $P(A) \neq 0$ and $P(B) \neq 0$.

8. A lab blood test is 95 percent effective in detecting a certain disease when it is, in fact, present. However, the test also yields a “false positive” result for 1 percent of the healthy persons tested. This means that if a healthy person is tested, then w.p. 0.01 the test result will imply he has the disease. If 0.5 percent of the population actually has the disease, what is the probability a person has the disease given that his test result is positive?

   Let D denote disease, ND denote no disease, PT denote a positive test, and NT a negative test.

   (a) Answer the question above using probability formulas to reach your final answer.

   (b) Answer the question using the natural frequencies approach.

9. A dice game called craps is played as follows. Two dice are rolled by the shooter. The shooter can win by rolling a 7 or an 11 on his first roll (called a natural). The shooter can also win by throwing a 4,5,6,8,9, or 10 on this first roll and throwing the same number again before rolling a 7.

   Let A_1 denote the event that the shooter rolls a natural. Let $A_4, A_5, A_6, A_8, A_9, A_{10}$ denote the respective events where the shooter wins when his point is a 4,5,6,8,9, or 10 respectively. Denote the rolls by R_i in setting up the problem. What is the probability that the shooter wins the game? Comment on your answer and based on your knowledge of casinos and gambling, does this seem like a good game to play if you went to Vegas?

   Hints:

   - Notice that the A_i are mutually exclusive.
   - Fact: $\sum_{k=0}^{\infty} p^r = \frac{1}{1-p}$, $|p| < 1$.
   - In calculating $P(\text{shooter wins})$, take advantage of some symmetry properties of the A_i to save yourself some work.

Chapter 3

Bayesian Statistics

3.0.1 Discrete Case

The foundation for Bayesian inference begins with Bayes' Rule, which we already presented in Chapter 2. Recall that as long as $P(B) > 0$,

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\ &\propto P(B|A)P(A), \end{aligned}$$

where $P(B) = \sum_i P(B|A_i)P(A_i)$, where the A_i are all mutually exclusive. We refer to $P(B)$ as a *normalizing constant.*

> Writing $P(A|B)$ using proportionality isn't really useful in the discrete setting because you have to put the normalizing constant back in eventually. We'll see how proportionality is useful later on.

In the formula above, we can think of $P(A)$ as our *prior distribution*, meaning our prior information about the data. We can think of $P(B|A)$ as the *likelihood.* The likelihood function is a function of the states of nature A for the fixed data B that we observe. $P(B)$ is called the *marginal distribution* or marginal probability of the data. Recall that we can calculate $P(B)$ using the law of total probability. That is,

$$P(B) = \sum_i P(B|A_i)P(A_i) = \sum_i P(B, A_i),$$

where the events A_i are all mutually exclusive.

Here, when we talk about the discrete case, *we mean the state of nature A is discrete.* For all practical purposes, both the data and the state of nature A will be discrete. You could certainly come up with an odd problem (not very realistic) where the data was continuous and the parameter was discrete. For example, we could have continuous data, but we would have to be very careful to make sure we did not divide by zero in calculating our posterior distribution.

3.0.2 Continuous Case

If we want to view Bayes' Rule in the continuous case where we have probability densities, we find very similar results. Let θ represent the parameter of the density and x denote the observed data. For the continuous case, we mean that θ is *always continuous.* The data could be *either discrete or continuous.* We can then write Bayes' Rule as

$$\begin{aligned} p(\theta|x) &= \frac{p(x|\theta)p(\theta)}{p(x)} \\ &\propto p(x|\theta)p(\theta), \end{aligned}$$

where $p(x) = \int_\theta p(x|\theta)p(\theta)\ d\theta$. So, $p(x)$ is a normalizing constant, and its distribution does not depend on the value of θ.

> Remark: In the examples we'll look at, we won't have to worry about the marginal distribution of x, $p(x)$. However, in more complicated problems, you would need to calculate $p(x)$ and could not take advantage of proportionality. In these more complicated problems, you would need knowledge of calculus.

Theorem 3.1: In order to perform valid inference, the posterior must be *proper*. In the discrete case, this means that the posterior distribution must sum to 1. In the continuous case, this means $\int_\theta p(\theta|x)\ d\theta = 1$.

Comment: Due to Theorem 3.1, in the discrete/continuous settings that we will soon cover, if we can recognize the posterior distribution as a common distribution, then we know it must be a proper distribution (since all continuous distributions integrate to 1 and discrete distributions sum to 1).

We will take advantage of proportionality by writing down $p(\theta|x)$ and simplifying it to some common distribution we recognize. If we encounter a posterior distribution that we do not recognize, then calculus would be needed to ensure the posterior is proper.

3.1 Discrete Examples

Example 3.1 (Simple Posterior Distribution)**:** Suppose we have three states of nature, denoted by A_1, A_2, A_3. We represent the likelihood and prior information in the following table:

	$P(A)$	$P(D\|A)$	
	Prior	D_1	D_2
A_1	0.3	0.0	1.0
A_2	0.5	0.7	0.3
A_3	0.2	0.2	0.8

Table 3.1: Simple Likelihood and Prior

Let's clarify what's going on in Table 3.1. Suppose we assume state of nature A_1. Then we observe D_2 100% of the time and we never observe D_1. On the other hand, suppose we instead assume state of nature A_2. Then we observe D_2 30% of the time and we observe D_1 70% of the time.

Suppose we observe D_1. How might we calculate the posterior of A given D_1?

	$P(A)$	$P(D\|A)$		$P(D_1, A)$	$P(A\|D_1)$
	Prior	D_1	D_2	Joint	Posterior
A_1	0.3	0.0	1.0	0.00	0.00
A_2	0.5	0.7	0.3	0.35	0.90
A_3	0.2	0.2	0.8	0.04	0.10
				0.39	1.00

Table 3.2: Posterior distribution of A given D_1

Let's look at one of the calculations in Table 3.1 to make sure we understand what's going on. Consider calculating

$$\begin{aligned} P(A_2|D_1) &= \frac{P(A_2, D_1)}{P(D_1)} \\ &= 0.35/0.39 \\ &= 0.90 \end{aligned}$$

Go home and calculate the posterior of A given D_2 at home as an exercise.

Remark: To compute each posterior calculation, we first write out each prior probability for A as well as the likelihood, i.e., $P(D_1|A)$. For ease of computation, we then write out the joint probabilities for each value of A. When we sum the joint probabilities up (over all values of A), we end up with the marginal distribution of D_1. With this information, we can now calculate the posterior probabilities as shown in the table.

Example 3.2 (Colon Cancer)**:** Doctors at Shands Hospital are interested in a hemoccult test for colon cancer. This screening procedure is not as accurate as a colonoscopy, but it is a much less expensive procedure for annual screening tests of colon cancer. Let A be the event that a patient has the cancer, let $+$ denote a patient tests positive for the condition, and $-$ denote a patient tests negative.

In the general population, only 0.3% have undiagnosed colon cancer. The hemoccult test will be positive 50% of the time if a patient has the disease and will be positive 3% of the time if the patient doesn't have the disease. Doctors are most interested in the proportion of *false positives* and *false negatives* that would occur if we used the test to screen the general population.

Let's find the posterior distribution of test result given whether or not the patient has colon cancer.

		Likelihood		Joint		Posterior	
	Prior	$+$	$-$	$+$	$-$	$+$	$-$
A	0.003	0.5	0.5	0.0015	0.0015	0.048	0.002
A^c	0.997	0.03	0.97	0.0299	0.9671	0.952	0.998
				0.0314	0.9686	1.00	1.00

Table 3.3: Posterior distribution

We'll calculate $P(A|-)$ to illustrate how one calculation in the table above is done. Consider calculating the probability that a patient does have colon cancer given they have a positive test result.

$$\begin{aligned} P(A|+) &= \frac{P(A,+)}{P(+)} \\ &= 0.0015/0.0314 \\ &= 0.048. \end{aligned}$$

Conclusions:

1. From this we can conclude that if a patient from the population tests positive for colon cancer, there is less than a 5 percent chance that the person has colon cancer.

2. There are many false positives, meaning that if a patient tests positive for colon cancer, there is about a 95 percent chance that the person doesn't have cancer. (This is why this is simply a screening test and must be followed up by a more accurate test for those who test positive).

3. In terms of false negatives, there aren't very many. That is, if a patient tests negative for the test, the probability the patient has the cancer is only 0.002.

Example 3.3 (Colon Cancer Example Using Natural Frequencies)**:** Recall the last example where we looked at posterior probabilities involving patients and colon cancer. Many doctors and patients have trouble understanding such examples as the one we just did. They have an easier time understanding the concept through an idea called *natural frequencies*, which involves considering a particular size for the population and then computing the expected number in each category for the population. This is one way for doctors and patients to better understand statistics and the real meaning of test results or a way for a statistician to relay information to a client.

Calculate the posterior probabilities using relative frequencies for the colon cancer example.

Consider the following information:

- We originally assumed a population of infinite size. When using the natural frequencies method, we need to assume the population has a finite size. Without loss of generality, we will assume that we are screening $10,000$ patients.

- From our original assumptions, we know 0.3% of the population have colon cancer. This means (0.003)(10,000) = 30 patients have colon cancer.

- If a patient has colon cancer, we know 50% test positive for colon cancer. So, (0.5)(30) = 15 patients have colon cancer and test positive for it.

- This also means that 15 patients have colon cancer and test negative for it.

- On the other hand, we know that 9970 patients don't have colon cancer. Of these patients, 3% have a positive test result, meaning (0.03)(9970) = 299 patients don't have cancer but test positive for the disease.

- Also, 9671 patients don't have cancer and test negative for it.

Now we have all the information we need to calculate the posterior probabilities that were in the table from the colon cancer example. Recall A is the event that a patient has colon cancer. Here are the four posterior probabilities:

1. $P(A|+) = \dfrac{15}{299+15} = 0.048.$

2. $P(A^c|+) = \dfrac{299}{299+15} = 0.952.$

3. $P(A|-) = \dfrac{15}{9671+15} = 0.002.$

4. $P(A^c|-) = \dfrac{9671}{9671+15} = 0.998.$

Remark: The posterior probabilities match those in Table 3.3. The same interpretations apply as well. While this method may be easier or more intuitive than the first approach, we have arrived at our matching answers by making some assumptions that aren't necessarily true. Making an assumption that 30 patients out of $10,000$ have colon cancer isn't an assumption that will always hold. We have invented a population size for this example. Also, events don't have to follow their corresponding probabilities as we talked about in Chapter 1. We have also rounded many of the answers above, which is another glossed over assumption. So, use this method for intuition, but also make sure you understand that it is not an exact method or theorem like Bayes' Rule.

3.2 Continuous and Discrete Examples

Definition 3.1: A distribution or density is *conjugate* if the prior and posterior distributions have the same type of distribution.

Example 3.4: Suppose that the likelihood for a particular problem is Binomial(n, θ) and the prior,θ, has a density that is Beta(a, b). We will soon show the posterior will also be Beta with different parameters. Thus, the prior distribution is conjugate for the binomial likelihood. Remember that conjugacy means that when we update using the prior information, we obtain a posterior that has the same distribution as the prior with possibly different parameter values.

> When dealing with continuous/discrete distributions in the rest of our material, we will almost always have conjugacy. Note you still need to check that you have conjugacy to proceed on in your calculations. Otherwise, remember that we have to compute an integral, which we don't know how to compute in this course.

Theorem 3.2:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}.$$

> Remark: This theorem only provides valid inference results if the posterior is proper, meaning the posterior integrates or sums to 1. For the purposes of this course, we will always deal with proper posteriors. However in real-world problems, posteriors can be improper and can lead to invalid inference.

Example 3.5: Suppose

$$\begin{aligned} X \mid \theta &\sim \text{Binomial}(n, \theta) \\ \theta &\sim \text{Beta}(a, b). \end{aligned}$$

Then

$$\begin{aligned} p(\theta|x) &\propto p(x|\theta)p(\theta) \\ &\propto \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\ &\propto \theta^x(1-\theta)^{n-x}\theta^{a-1}(1-\theta)^{b-1} \\ &\propto \theta^{x+a-1}(1-\theta)^{n-x+b-1}. \end{aligned}$$

The above expression is just an updated Beta with parameters $x+a$ and $n-x+b$, so

$$\theta \mid x \sim \text{Beta}(x+a, n-x+b).$$

□

How to Choose a and b

In Example 3.5, you may wonder how we choose the values of a and b. In practice, we often have ideas regarding the value of the prior mean and variance. Past studies or expert opinions often help us choose these values. Suppose we know the prior mean of Beta(a, b) to be μ. We can then say

$$\mu = \frac{a}{(a+b)},$$

however this doesn't determine the values of a and b uniquely. We still need more information.

If we know the prior variance to be σ^2 then

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

and we have a system of two equations involving two unknowns (a and b). Hence, we can uniquely solve for both parameters.

Example 3.6: (Normal-Uniform Prior)

$$X_1, \ldots, X_n|\theta \overset{iid}{\sim} \text{Normal}(\theta, \sigma^2),\ \sigma^2 \text{ known}$$
$$\theta \sim \text{Uniform}(-\infty, \infty),$$

where $\theta \sim$ Uniform$(-\infty, \infty)$ means that $p(\theta) \propto 1$.

Calculate the posterior distribution of θ given the data.

$$\begin{aligned} p(\theta|x) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-1}{2\sigma^2}(x_i - \theta)^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\right\} \\ &\propto \exp\left\{\frac{-1}{2\sigma^2}\sum_i (x_i - \theta)^2\right\}. \end{aligned}$$

Note that $\sum_i (x_i - \theta)^2 = \sum_i (x_i - \bar{x})^2 + n(\bar{x} - \theta)^2$. Then

$$\begin{aligned} p(\theta|x) &\propto \exp\left\{\frac{-1}{2\sigma^2}\sum_i (x_i - \bar{x})^2\right\} \exp\left\{\frac{-n}{2\sigma^2}(\bar{x} - \theta)^2\right\} \\ &\propto \exp\left\{\frac{-n}{2\sigma^2}(\bar{x} - \theta)^2\right\} \\ &= \exp\left\{\frac{-n}{2\sigma^2}(\theta - \bar{x})^2\right\}. \end{aligned}$$

Thus,

$$\theta | x_1, \ldots, x_n \sim \text{Normal}(\bar{x}, \sigma^2/n).$$

Example 3.7: (Beta-Binomial Example in R) We are interested in a population of American college students and the proportion of the population that sleep at least eight hours a night, which we denote by θ.

The student newspaper, *The Gamecock*, at the University of South Carolina printed an internet article "College Students Don't Get Enough Sleep" on April 20, 2004. The article reported that most students spend six hours sleeping each night. A year earlier, a similar article appeared in the University of Notre Dame's paper, *Fresh Writing*. The article entitled *Sleep on It: Implementing a Relaxation Program into the College Curriculum* reported that based on a sample of 100 students "approximately 70% reported to receiving only five to six hours of sleep on the weekdays, 28% receiving seven to eight, and only 2% receiving the healthy nine hours for teenagers."

As for our data, a random sample of 27 students is taken from the University of Florida. 11 students record that they sleep at least eight hours each night. Based on this information, we are interested in estimating θ.

Since we are Bayesians, we need to figure out our prior distribution for θ. We read both the articles from USC and UND and believe it's probably true that most college students get less than eight hours of sleep. Therefore, we want our prior to assign most of the probability to values of θ less than 0.5. From the information given, we decide that our best guess for θ is 0.3, although we think it is very possible that θ could be any value in $[0, 0.5]$.

Given this information, we believe that the median of θ is 0.3 and the 90th percentile is 0.5. Knowing this allows us to estimate the unknown values of a and b. After some calculations we find that $a = 3.3$ and $b = 7.2$. How did we actually calculate a and b?

We would need to solve the following equations:

$$\int_0^{0.3} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}\, d\theta = 0.5$$

$$\int_0^{0.5} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}\, d\theta = 0.9$$

In non-calculus language, this means the 0.5 quantile (50th percentile) = 0.3. The 0.9 quantile (90th percentile) = 0.5. We can easily solve this numerically in `R` using a numerical solver `BBsolve`.

Since you won't have used this command before, you'll need to install the package `BB` and load the library.

Here is the code in `R` to find a and b.

```
## install the BBsolve package
install.packages("BB", repos="http://cran.r-project.org")
library(BB)
fn = function(x){qbeta(c(0.5,0.9),x[1],x[2])-c(0.3,0.5)}
BBsolve(c(1,1),fn)
```

BBsolve is a numerical solver. You give it two arguments. First, you give it a set of initial values as your best guess of a and b. Then you define a function which must take arguments as vectors for the unknown values of a and b. The solver sets the function to zero and then finds the best solutions for a and b (we won't go into the details of how it solves for the values in this class).

So, in place of a and b, we use $x[1]$ and $x[2]$ in our function (since it requires vectors). We also use the qbeta command because we're interested in the 0.5 and 0.9 quantiles. We must subtract 0.3 and 0.5 from each respective quantile since BBsolve sets our function to zero and solves. To do all of this in one step, we represent everything as a vector using c(). For example, c(0.3,0.5) represents the vector with elements 0.3 and 0.5.

If you have more questions on function or BBsolve remember you can get help by ?function or ?BBsolve.

Using our calculations from Example 3.5 our model is

$$
\begin{aligned}
X \mid \theta &\sim \text{Binomial}(27, \theta)\\
\theta &\sim \text{Beta}(3.3, 7.2)\\
\theta \mid x &\sim \text{Beta}(x + 3.3, 27 - x + 7.2)\\
\theta \mid 11 &\sim \text{Beta}(14.3, 23.2)
\end{aligned}
$$

We can easily plot the likelihood, prior, and posterior in R. The command seq below creates a sequence of 500 numbers between 0 and 1. Using the plot command, we plot the posterior distribution using lty (line type) 2 and lwd (line width) 3. For the xlab (x-label), we use expression(theta) so that a θ will print on the x-axis. The command lines allows us to add more plots onto the original plot. The legend creates a legend which is mostly straightforward. From Figure 4.2, we can see that the posterior is a mixture of the likelihood and the prior. By visual inspection, the posterior mean appears to be around 0.4. We will come back to this example later to analyze other questions of interest.

```
th = seq(0,1,length=500)
a = 3.3
b = 7.2
n = 27
x = 11
prior = dbeta(th,a,b)
like = dbeta(th,x+1,n-x+1)
post = dbeta(th,x+a,n-x+b)
pdf("sleep.pdf",width=7,height=5)
plot(th,post,type="l",ylab="Density",lty=2,lwd=3,xlab = expression(theta))
lines(th,like,lty=1,lwd=3)
lines(th,prior,lty=3,lwd=3)
legend(0.7,4,c("Prior","Likelihood","Posterior"),lty=c(3,1,2),lwd=c(3,3,3))
dev.off()
```

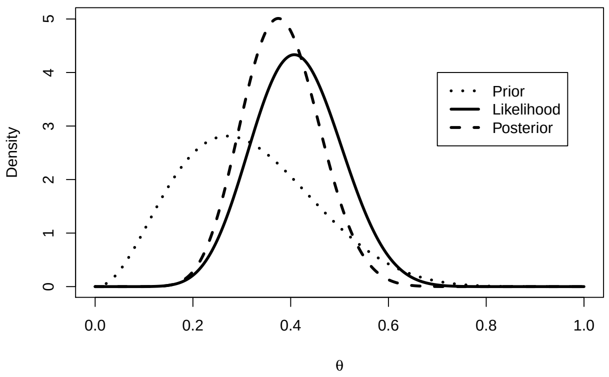


Figure 3.1: Likelihood $p(X|\theta)$, Prior $p(\theta)$, and Posterior Distribution $p(\theta|X)$

□

Example 3.8:

$$\begin{aligned} X_1, \dots, X_n|\theta &\overset{iid}{\sim} \mathrm{N}(\theta, \sigma^2) \\ \theta &\sim \mathrm{N}(\mu, \tau^2), \end{aligned}$$

where σ^2 is known. Calculate the distribution of $\theta|x_1, \dots, x_n$.

$$\begin{aligned} p(\theta|x_1, \dots, x_n) &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-1}{2\sigma^2}(x_i - \theta)^2 \right\} \times \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{ \frac{-1}{2\tau^2}(\theta - \mu)^2 \right\} \\ &\propto \exp\left\{ \frac{-1}{2\sigma^2} \sum_i (x_i - \theta)^2 \right\} \exp\left\{ \frac{-1}{2\tau^2}(\theta - \mu)^2 \right\}. \end{aligned}$$

Consider

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

Then

$$
\begin{aligned}
p(\theta|x_1,\ldots,x_n) &= \exp\left\{\frac{-1}{2\sigma^2}\sum_i (x_i-\bar{x})^2\right\} \times \exp\left\{\frac{-1}{2\sigma^2}n(\bar{x}-\theta)^2\right\} \times \exp\left\{\frac{-1}{2\tau^2}(\theta-\mu)^2\right\} \\
&\propto \exp\left\{\frac{-1}{2\sigma^2}n(\bar{x}-\theta)^2\right\}\exp\left\{\frac{-1}{2\tau^2}(\theta-\mu)^2\right\} \\
&= \exp\left\{\frac{-1}{2}\left[\frac{n}{\sigma^2}\left(\bar{x}^2-2\bar{x}\theta+\theta^2\right)+\frac{1}{\tau^2}\left(\theta^2-2\theta\mu+\mu^2\right)\right]\right\} \\
&= \exp\left\{\frac{-1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)\theta^2-2\theta\left(\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{\tau^2}\right)+\frac{n\bar{x}^2}{\sigma^2}+\frac{\mu^2}{\tau^2}\right]\right\} \\
&\propto \exp\left\{\frac{-1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)\left(\theta^2-2\theta\frac{\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\right)\right]\right\} \\
&\propto \exp\left\{\frac{-1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)\left(\theta-\frac{\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\right)^2\right]\right\}.
\end{aligned}
$$

Recall what it means to complete the square as we did above.[1] Thus,

$$
\begin{aligned}
\theta|x_1,\ldots,x_n &\sim N\left(\frac{\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}, \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\right) \\
&= N\left(\frac{n\bar{x}\tau^2+\mu\sigma^2}{n\tau^2+\sigma^2}, \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}\right).
\end{aligned}
$$

[1] Recall from algebra that $(x-b)^2 = x^2-2bx+b^2$. We want to complete something that resembles $x^2-2bx = x^2+2bx+(2b/2)^2-(2b/2)^2 = (x-b)^2-b^2$.

Definition 3.2: The reciprocal of the variance is referred to as the *precision.* That is,

$$\text{Precision} = \frac{1}{\text{Variance}}.$$

Example 3.9: (Normal-Normal Revisited) Recall Example 3.8. We write the posterior mean as $E(\theta|x)$. Let's write the posterior mean in this example as

$$\begin{aligned} E(\theta|x) &= \frac{\frac{n\bar{x}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}. \\ &= \frac{\frac{n\bar{x}}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} + \frac{\frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}. \end{aligned}$$

We also write the posterior variance as

$$V(\theta|x) = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}.$$

We can see that the posterior mean is a weighted average of the sample mean and the prior mean. The weights are proportional to the reciprocal of the respective variances (precision). In statistics, we often say that the reciprocal of the variance is called the precision. In this case,

$$\begin{aligned} \text{Posterior Precision} &= \frac{1}{\text{Posterior Variance}} \\ &= (n/\sigma^2) + (1/\tau^2) \\ &= \text{Sample Precision} + \text{Prior Precision}. \end{aligned}$$

The posterior precision is larger than either the sample precision or the prior precision. Equivalently, the posterior variance, denoted by $V(\theta|x)$, is smaller than either the sample variance or the prior variance.

What happens as n gets large?

Divide the posterior mean (numerator and denominator) by n. Then

$$E(\theta|x) = \frac{\dfrac{1}{n}\dfrac{n\bar{x}}{\sigma^2} + \dfrac{1}{n}\dfrac{\mu}{\tau^2}}{\dfrac{1}{n}\dfrac{n}{\sigma^2} + \dfrac{1}{n}\dfrac{1}{\tau^2}} \approx \frac{\dfrac{\bar{x}}{\sigma^2}}{\dfrac{1}{\sigma^2}} \approx \bar{x} \quad \text{for large } n.$$

In the case of the posterior variance, divide the denominator and numerator by n. Then

$$V(\theta|x) = \frac{\dfrac{1}{n}}{\dfrac{1}{n}\dfrac{n}{\sigma^2} + \dfrac{1}{n}\dfrac{1}{\tau^2}} \approx \frac{\sigma^2}{n},$$

which goes to zero as n gets large.

Gamma and Inverse Gamma Distributions

Before we introduce the next example, we will first introduce two continuous probability densities—the Gamma and the Inverse Gamma.

Some random variables (such as the Beta distribution) are always non-negative and for various reasons yield distributions of data that are skewed to the right. That is, most of the area under the density function is located near the origin, and the density function drops gradually as x increases.

The lengths of time between malfunctions for aircraft engines possess a skewed frequency distribution, as do the lengths of time between arrivals at a supermarket checkout queue. Furthermore, the lengths of time to complete a routine checkup for a car or aircraft carrier possess a skewed frequency distribution. The populations associated with these random variables often possess distributions that are adequately modeled by the gamma density function.

Suppose $X \sim \text{Gamma}(\alpha, \beta)$. Then

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}, \;\; x > 0, \; \alpha > 0, \; \beta > 0.$$

Suppose $Y \sim \text{IG}(a, b)$, where IG stands for Inverse Gamma. Then

$$f(y) = \frac{b^{a}}{\Gamma(a)} y^{-a-1} e^{-b/y}, \;\; y > 0, \; a > 0, \; b > 0.$$

There is a special relationship between the Gamma and Inverse Gamma random variables. As defined above, using calculus, it is easy to show that $Y = 1/X$. Even though you don't need to know how to prove this fact in this course, it's one worth knowing.

Example 3.10:

$$\begin{aligned} X|\alpha,\beta &\sim \text{Gamma}(\alpha,\beta),\ \alpha \text{ known},\ \beta \text{ unknown} \\ \beta &\sim \text{IG}(a,b). \end{aligned}$$

Calculate the posterior distribution of $\beta|x$.

$$\begin{aligned} p(\beta|x) &\propto \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} \times \frac{b^a}{\Gamma(a)} \beta^{-a-1} e^{-b/\beta} \\ &\propto \frac{1}{\beta^\alpha} e^{-x/\beta} \beta^{-a-1} e^{-b/\beta} \\ &= \beta^{-\alpha-a-1} e^{-(x+b)/\beta}. \end{aligned}$$

Notice that this looks like an Inverse Gamma distribution with parameters $\alpha + a$ and $x + b$. Thus,

$$\beta|x \sim IG(\alpha + a, x + b).$$

Example 3.11: (Bayesian versus Frequentist)
Suppose a child is given an IQ test and his score is X. We assume that

$$\begin{aligned} X|\theta &\sim \text{Normal}(\theta, 100) \\ \theta &\sim \text{Normal}(100, 225) \end{aligned}$$

From previous calculations, we know that the posterior is

$$\theta|x \sim \text{Normal}\left(\frac{400 + 9x}{13}, \frac{900}{13}\right).$$

In the next chapter we will learn about the posterior mean and variance. Here the posterior mean is $(400 + 9x)/13$. Suppose $x = 115$. Then the posterior mean becomes 110.4. Contrasting this, we know that the frequentist estimate is the mle, which is $x = 115$ in this example.

The posterior variance is $900/13 = 69.23$, whereas the variance of the data is $\sigma^2 = 100$.

Now suppose we take the Uniform$(-\infty, \infty)$ prior on θ. From an earlier example, we found that the posterior is

$$\theta|x \sim \text{Normal}(115, 100).$$

Notice that the posterior mean and mle are both 115 and the posterior variance and variance of the data are both 100.

When we put little/no prior information on θ, the data washes away most/all of the prior information (and the results of frequentist and Bayesian estimation are similar or equivalent in this case).

Example 3.12: (Normal Example with Unknown Variance)
Consider

$$\begin{aligned} X_1, \ldots, X_n|\theta &\overset{iid}{\sim} \text{Normal}(\theta, \sigma^2),\ \theta \text{ known},\ \sigma^2 \text{ unknown} \\ p(\sigma^2) &\propto (\sigma^2)^{-1}. \end{aligned}$$

Calculate $p(\sigma^2|x_1,\ldots,x_n)$.

$$
\begin{aligned}
p(\sigma^2|x_1,\ldots,x_n) &\propto (2\pi\sigma^2)^{-n/2}\exp\left\{\frac{-1}{2\sigma^2}\sum_i(x_i-\theta)^2\right\}(\sigma^2)^{-1} \\
&\propto (\sigma^2)^{-n/2-1}\exp\left\{\frac{-1}{2\sigma^2}\sum_i(x_i-\theta)^2\right\}.
\end{aligned}
$$

Recall, if $Y \sim \text{IG}(a,b)$, then $f(y) = \dfrac{b^a}{\Gamma(a)}y^{-a-1}e^{-b/y}$. Thus,

$$
\sigma^2|x_1,\ldots,x_n \sim \text{IG}\left(n/2, \frac{\sum_{i=1}^n(x_i-\theta)^2}{2}\right).
$$

Example 3.13: (Football Data) Gelman *et. al* (2003) consider the problem of estimating an unknown variance using American football scores. The focus is on the difference d between a game outcome (winning score minus losing score) and a published point spread.

We observe $d_1,\ldots,d_n$, the observed differences between game outcomes and point spreads for $n = 2240$ football games. We assume these differences are a random sample from a Normal distribution with mean 0 and unknown variance σ^2. Our goal is to find inference regarding the unknown parameter σ^2, which represents the variability in the game outcomes and point spreads.

We can refer to Example 3.12, since the setup here is the same. Hence the posterior becomes

$$
\sigma^2|d_1,\ldots,d_n \sim \text{IG}(n/2, \sum_i d_i^2/2).
$$

The next logical step would be plotting the posterior distribution in R. As far as I can tell, there is not a built in function predefined in R for the Inverse Gamma density. However, someone saw the need for it and built one in using the pscl package.

Proceeding below, we try and calculate the posterior using the function densigamma, which corresponds to the Inverse Gamma density. However, running this line in the code gives the following error:

```
Warning message:
In densigamma(sigmas, n/2, sum(d^2)/2) : value out of range in 'gammafn'
```

What's the problem? Think about the what the posterior looks like. Recall that

$$
p(\sigma^2|\boldsymbol{d}) = \frac{(\sum_i d_i^2/2)^{n/2}}{\Gamma(n/2)}(\sigma^2)^{-n/2-1}e^{-(\sum_i d_i^2)/2\sigma^2}.
$$

In the calculation R is doing, it's dividing by $\Gamma(1120)$, which is a very large factorial. This is too large for even R to compute, so we're out of luck here. So, what can we do to analyze the data?

```
setwd("~/Desktop/sta4930/football")
data = read.table("football.txt",header=T)
names(data)
attach(data)
score = favorite-underdog
d = score-spread
n = length(d)
hist(d)
install.packages("pscl",repos="http://cran.opensourceresources.org")
library(pscl)
?densigamma
sigmas = seq(10,20,by=0.1)
post = densigamma(sigmas,n/2,sum(d^2)/2)
v = sum(d^2)
```

We know we can't use the Inverse Gamma density (because of the function in R), but we do know a relationship regarding the Inverse Gamma and Gamma distributions. So, let's apply this fact.

You may be thinking, we're going to run into the same problem because we'll still be dividing by $\Gamma(1120)$. This is true, except the Gamma density function dgamma was built into R by the original writers. The dgamma function is able to do some internal tricks that let it calculate the gamma density even though the individual piece $\Gamma(n/2)$ by itself is too large for R to handle. So, moving forward, we will apply the following fact that we already learned:

$$\text{If } X \sim \text{IG}(a,b), \text{ then } 1/X \sim \text{Gamma}(a, 1/b).$$

Since

$$\sigma^2|d_1,\ldots,d_n \sim \text{IG}(n/2, \sum_i d_i^2/2),$$

we know that

$$\frac{1}{\sigma^2}|d_1,\ldots,d_n \sim \text{Gamma}(n/2, 2/v), \quad \text{where } v = \sum_i d_i^2.$$

Now we can plot this posterior distribution in R, however in terms of learning inference about σ^2, this isn't going to be very useful. This is where we begin to need calculus methods and reach the outer limits that will be taught in this course.

In the code below, we plot the posterior of $\frac{1}{\sigma^2}|\boldsymbol{d}$. In order to do so, we must create a new sequence of x-values since the mean of our gamma will be at $n/v \approx 0.0053$.

```
xnew = seq(0.004,0.007,.000001)
pdf("football_sigmainv.pdf", width = 5, height = 4.5)
post.d = dgamma(xnew,n/2,scale = 2/v)
plot(xnew,post.d, type= "l", xlab = expression(1/sigma^2), ylab= "density")
dev.off()
```

As we can see from the plot below, viewing the posterior of $\frac{1}{\sigma^2}|\boldsymbol{d}$ isn't very useful. We would like to get the parameter in terms of σ^2, so that we could plot the posterior distribution of interest as well as calculate the posterior mean and variance. From this point on we need to use a bit of calculus (for which you are not responsible). The calculations are provided just for interest and so you can see when calculus is needed and when situations arise in application that don't proceed as we might have expected.

Calculus details start here.
To recap, we know

$$\frac{1}{\sigma^2}|d_1,\ldots,d_n \sim \text{Gamma}(n/2, 2/v), \quad \text{where } v = \sum_i d_i^2.$$

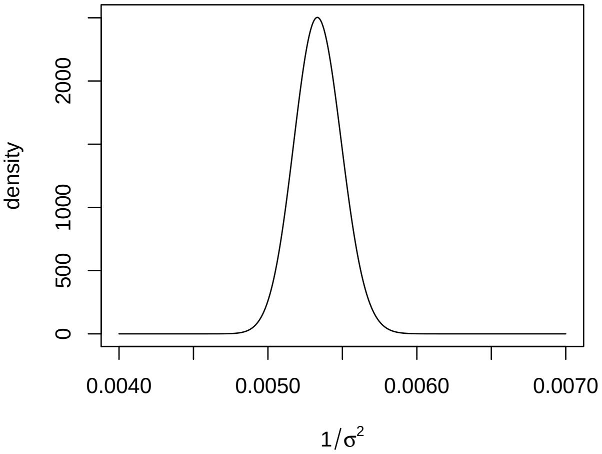


Figure 3.2: Posterior Distribution $p(\frac{1}{\sigma^2}|d_1,\ldots,d_n)$

Let $u = \frac{1}{\sigma^2}$. We are going to make a transformation of variables now to write the density in terms of σ^2. (Here comes the calculus).

Since $u = \frac{1}{\sigma^2}$, this implies $\sigma^2 = \frac{1}{u}$. Then $\left|\dfrac{\partial u}{\partial \sigma^2}\right| = \dfrac{1}{\sigma^4}$.

Now applying the transformation of variables we find that

$$f(\sigma^2|d_1,\ldots,d_n) = \frac{1}{\Gamma(n/2)(2/v)^{n/2}}\left(\frac{1}{\sigma^2}\right)^{n/2-1} e^{-\frac{v}{2\sigma^2}}\left(\frac{1}{\sigma^4}\right).$$

Thus,

$$\sigma^2|\boldsymbol{d} \sim \text{Gamma}(n/2, 2/v)\left(\frac{1}{\sigma^4}\right).$$

Calculus details end here. You are not responsible for knowing any of the results involving calculus. They are simply here in case you are interested in them.

Now, we know the density of $\sigma^2|\boldsymbol{d}$ in a form we can calculate in R.

```
x.s = seq(150,250,1)
pdf("football_sigma.pdf", height = 5, width = 4.5)
post.s = dgamma(1/x.s,n/2, scale = 2/v)*(1/x.s^2)
plot(x.s,post.s, type="l", xlab = expression(sigma^2), ylab="density")
dev.off()
detach(data)
```

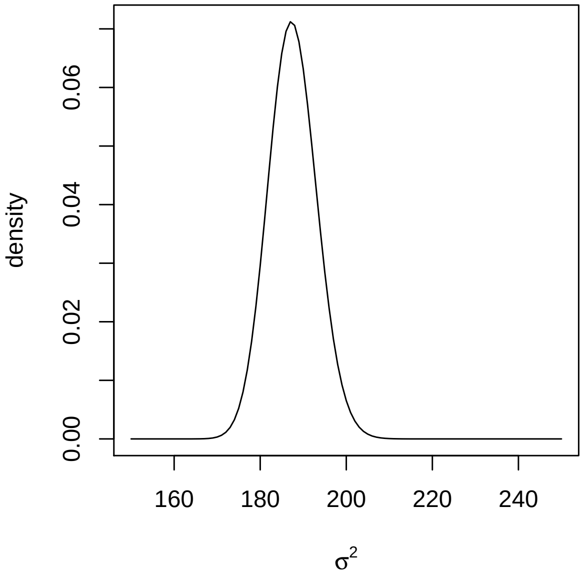


Figure 3.3: Posterior Distribution $p(\sigma^2|d_1,\ldots,d_n)$

From the posterior plot in Figure 3.3 we can see that the posterior mean is around 185. This means that the variability of the actual game result around the point spread has a standard deviation around 14 points. If you wanted to actually calculate the posterior mean and variance you would need to use calculus or some sort of numerical method, which I will not go into since this is beyond the scope of this course.

What's interesting about this example is that there is a lot more variability in football games than the average person would most likely think. Assume that (1) the standard deviation actually is 14 points, and (2) game result is normally distributed (which it's not, exactly, but this is a reasonable approximation). Things with a normal distribution fall two or more standard deviations from their mean about 5% of the time, so this means that, roughly speaking, about 5% of football games end up 28 or more points away from their spread.

Example 3.14:

$$
\begin{aligned}
Y_1, \ldots, Y_n | \mu, \sigma^2 &\overset{iid}{\sim} \text{Normal}(\mu, \sigma^2), \\
\mu | \sigma^2 &\sim \text{Normal}(\mu_0, \frac{\sigma^2}{\kappa_0}), \\
\sigma^2 &\sim \text{IG}(\frac{\nu_0}{2}, \frac{\sigma_0^2}{2}),
\end{aligned}
$$

where $\mu_0, \kappa_0, \nu_0, \sigma_0^2$ are constant.

Find $p(\mu, \sigma^2 | y_1 \ldots, y_n)$. Notice that

$$
\begin{aligned}
p(\mu, \sigma^2 | y_1 \ldots, y_n) &= \frac{p(\mu, \sigma^2, y_1 \ldots, y_n)}{p(y_1 \ldots, y_n)} \\
&\propto p(y_1 \ldots, y_n | \mu, \sigma^2) p(\mu, \sigma^2) \\
&= p(y_1 \ldots, y_n | \mu, \sigma^2) p(\mu | \sigma^2) p(\sigma^2).
\end{aligned}
$$

Then

$$
\begin{aligned}
p(\mu, \sigma^2 | y_1 \ldots, y_n) &\propto p(y_1 \ldots, y_n | \mu, \sigma^2) p(\mu | \sigma^2) p(\sigma^2) \\
&\propto (\sigma^2)^{-n/2} \exp\left\{ \frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\} (\sigma^2)^{-1/2} \exp\left\{ \frac{-\kappa_0}{2\sigma^2} (\mu - \mu_0)^2 \right\} \\
&\times (\sigma^2)^{-\nu_0/2-1} \exp\left\{ \frac{-\sigma_0^2}{2\sigma^2} \right\}.
\end{aligned}
$$

Consider $\sum_i (y_i - \mu)^2 = \sum_i (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2$.

Then

$$
\begin{aligned}
n(\bar{y} - \mu)^2 + \kappa_0(\mu - \mu_0)^2 &= n\bar{y}^2 - 2n\bar{y}\mu + n\mu^2 + \kappa_0\mu^2 - 2\kappa_0\mu\mu_0 + \kappa_0{\mu_0}^2 \\
&= (n + \kappa_0)\mu^2 - 2(n\bar{y} + \kappa_0\mu_0)\mu + n\bar{y}^2 + \kappa_0{\mu_0}^2 \\
&= (n + \kappa_0) \left(\mu - \frac{n\bar{y} + \kappa_0\mu_0}{n + \kappa_0} \right)^2 - \frac{(n\bar{y} + \kappa_0\mu_0)^2}{n + \kappa_0} + n\bar{y}^2 + \kappa_0{\mu_0}^2.
\end{aligned}
$$

Now consider

$$
\begin{aligned}
n\bar{y}^2 + \kappa_0{\mu_0}^2 - \frac{(n\bar{y} + \kappa_0\mu_0)^2}{n + \kappa_0} &= n\bar{y}^2 + \kappa_0{\mu_0}^2 + \frac{-n^2\bar{y}^2 - 2n\kappa_0\mu_0\bar{y} - {\kappa_0}^2{\mu_0}^2}{n + \kappa_0} \\
&= \frac{n^2\bar{y}^2 + n\kappa_0{\mu_0}^2 + n\kappa_0\bar{y}^2 + {\kappa_0}^2{\mu_0}^2 - n^2\bar{y}^2 - 2n\kappa_0\mu_0\bar{y} - {\kappa_0}^2{\mu_0}^2}{n + \kappa_0} \\
&= \frac{n\kappa_0{\mu_0}^2 + n\kappa_0\bar{y}^2 - 2n\kappa_0\mu_0\bar{y}}{n + \kappa_0} \\
&= \frac{n\kappa_0({\mu_0}^2 - 2\mu_0\bar{y} + \bar{y}^2)}{n + \kappa_0} \\
&= \frac{n\kappa_0(\mu_0 - \bar{y})^2}{n + \kappa_0}.
\end{aligned}
$$

Putting this all together, we find

$$
\begin{aligned}
p(\mu, \sigma^2 | y_1 \ldots, y_n) &\propto \exp\left\{ \frac{-n}{2\sigma^2} (\bar{y} - \mu)^2 \right\} \exp\left\{ \frac{-1}{2\sigma^2} \sum_i (y_i - \bar{y})^2 \right\} \\
&\quad \times \exp\left\{ \frac{-\kappa_0}{2\sigma^2} \sum_i (\mu - \mu_0)^2 \right\} (\sigma^2)^{-n/2-1/2} (\sigma^2)^{-\nu_0/2-1} \exp\left\{ \frac{-\sigma_0^2}{2\sigma^2} \right\} \\
&= \exp\left\{ \frac{-n\kappa_0}{2\sigma^2(n+\kappa_0)} (\mu_0 - \bar{y})^2 \right\} \exp\left\{ \frac{-1}{2\sigma^2} \sum_i (y_i - \bar{y})^2 \right\} \\
&\quad \times \exp\left\{ -\frac{(n+\kappa_0)}{2\sigma^2} \left(\mu - \frac{n\bar{y} + \kappa_0\mu_0}{n+\kappa_0} \right)^2 \right\} (\sigma^2)^{-\nu_0/2-1} (\sigma^2)^{-n/2-1} \exp\left\{ \frac{-\sigma_0^2}{2\sigma^2} \right\} \\
&= \exp\left\{ \frac{-1}{2\sigma^2} \sum_i (y_i - \bar{y})^2 - \frac{n\kappa_0}{2\sigma^2(n+\kappa_0)} (\mu_0 - \bar{y})^2 - \frac{\sigma_0^2}{2\sigma^2} \right\} (\sigma^2)^{-(n+\nu_0)/2-1} \\
&\quad \times \exp\left\{ -\frac{(n+\kappa_0)}{2\sigma^2} \left(\mu - \frac{n\bar{y} + \kappa_0\mu_0}{n+\kappa_0} \right)^2 \right\} (\sigma^2)^{-1/2}.
\end{aligned}
$$

Since the posterior above factors, we find

$$
\mu | \sigma^2, \boldsymbol{y} \sim \text{Normal}\left(\frac{n\bar{y} + \kappa_0\mu_0}{n+\kappa_0}, \frac{\sigma^2}{n+\kappa_0} \right),
$$

$$
\sigma^2 | \boldsymbol{y} \sim IG\left(\frac{n+\nu_0}{2}, \frac{1}{2} \left(\sum_i (y_i - \bar{y})^2 + \frac{n\kappa_0}{(n+\kappa_0)} (\mu_0 - \bar{y})^2 + \sigma_0^2 \right) \right).
$$

3.3 Posterior Predictive Distributions

We have just gone through many examples illustrating how to calculate many simple posterior distributions. This is the main goal of a Bayesian analysis. Another goal might be prediction. That is given some data, y, and a new observation, $\tilde{y}$, we may wish to find the conditional distribution of $\tilde{y}$ given y. This distribution is referred to as the *posterior predictive distribution*. That is, our goal is to find $p(\tilde{y}|y)$.

We'll derive the posterior predictive distribution for the discrete case (θ is discrete). It's the same for the continuous case, with the sums replaced with integrals.

Consider

$$\begin{aligned}
p(\tilde{y}|y) &= \frac{p(\tilde{y},y)}{p(y)} \\
&= \frac{\sum_\theta p(\tilde{y},y,\theta)}{p(y)} \\
&= \frac{\sum_\theta p(\tilde{y}|y,\theta)p(y,\theta)}{p(y)} \\
&= \sum_\theta p(\tilde{y}|y,\theta)p(\theta|y).
\end{aligned}$$

In most contexts, if θ is given, then $\tilde{y}|\theta$ is independent of y, i.e., the value of θ determines the distribution of $\tilde{y}$, without needing to also know y. When this is the case, we say that $\tilde{y}$ and y are *conditionally independent* given θ. Then the above becomes

$$p(\tilde{y}|y) = \sum_\theta p(\tilde{y}|\theta)p(\theta|y).$$

Theorem 3.3: If θ is discrete and $\tilde{y}$ and y are conditionally independent given θ, then the posterior predictive distribution is

$$p(\tilde{y}|y) = \sum_\theta p(\tilde{y}|\theta)p(\theta|y).$$

If θ is continuous and $\tilde{y}$ and y are conditionally independent given θ, then the posterior predictive distribution is

$$p(\tilde{y}|y) = \int_\theta p(\tilde{y}|\theta)p(\theta|y)\,d\theta.$$

Theorem 3.4: Suppose $p(x)$ is a pdf that looks like $p(x) = cf(x)$, where c is a constant and f is a continuous function of x. Since

$$\int_x p(x)\ dx = \int_x cf(x)\ dx = 1,$$

then

$$\int_x f(x) dx = 1/c.$$

Note: No calculus is needed to compute $\int_x f(x)\ dx$ if $f(x)$ looks like a known pdf.

Example 3.15: Human males have one X-chromosome and one Y-chromosome, whereas females have two X-chromosomes, each chromosome being inherited from one parent. Hemophilia is a disease that exhibits X-chromosome-linked recessive inheritance, meaning that a male who inherits the gene that causes the disease on the X-chromosome is affected, whereas a female carrying the gene on only one of her X-chromosomes is not affected. The disease is generally fatal for women who inherit two such genes, and this is very rare, since the frequency of occurrence of the gene is very low in human populations.

Consider a woman who has an affected brother (xY), which implies that her mother must be a carrier of the hemophilia gene (xX). We are also told that her father is not affected (XY), thus the woman herself has a fifty-fifty chance of having the gene.

Let θ denote the state of the woman. It can take two values: the woman is a carrier ($\theta = 1$) or not ($\theta = 0$). Based on this, the prior can be written as

$$P(\theta = 1) = P(\theta = 0) = 1/2.$$

Suppose the woman has a son who does not have hemophilia ($S1 = 0$). Now suppose the woman has another son. Calculate the probability that this second son also will not have hemophilia ($S2 = 0$), given that the first son does not have hemophilia. Assume son one and son two are conditionally independent given θ.

Solution:

$$p(S2 = 0|S1 = 0) = \sum_\theta p(S2 = 0|\theta)p(\theta|S1 = 0).$$

First compute

$$\begin{aligned} p(\theta|S1 = 0) &= \frac{p(S1 = 0|\theta)p(\theta)}{p(S1 = 0|\theta = 0)p(\theta = 0) + p(S1 = 0|\theta = 1)p(\theta = 1)} \\ &= \begin{cases} \frac{(1)(1/2)}{(1)(1/2)+(1/2)(1/2)} = \frac{2}{3} & \text{if } \theta = 0 \\ \frac{1}{3} & \text{if } \theta = 1. \end{cases} \end{aligned}$$

Then

$$\begin{aligned} p(S2 = 0|S1 = 0) &= p(S2 = 0|\theta = 0)p(\theta = 0|S1 = 0) + p(S2 = 0|\theta = 1)p(\theta = 1|S1 = 0) \\ &= (1)(2/3) + (1/2)(1/3) = 5/6. \end{aligned}$$

Negative Binomial Distribution

Before doing the next example, we will introduce the Negative Binomial distribution. The binomial distribution counts the numbers of successes in a fixed number of Bernoulli trials. Recall, a Bernoulli trial has a fixed success probability p, where the trials are iid (independent and identically distributed).

Suppose instead, we count the number of Bernoulli trials required to get a fixed number of successes. This formulation leads to the *Negative Binomial distribution.*

In a sequence of independent Bernoulli(p) trials, let X denote the trial at which the rth success occurs, where r is a fixed integer.

Then

$$f(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r},\ x = r, r+1, \ldots$$

and we say $X \sim$ Negative Binom(r, p).

There is another useful formulation of the Negative Binomial distribution. In many cases, it is defined as $Y =$ number of failures before the rth success. This formulation is statistically equivalent to the one given above in term of $X =$ trial at which the rth success occurs, since $Y = X - r$. Then

$$f(y) = \binom{r+y-1}{y} p^r (1-p)^y,\ y = 0, 1, 2, \ldots$$

and we say $Y \sim$ Negative Binom(r, p).

When we refer to the Negative Binomial distribution in this class, we will refer to the second one defined unless we indicate otherwise.

Example 3.16: (Poisson-Gamma)

$$\begin{aligned} X|\lambda &\sim \text{Poisson}(\lambda) \\ \lambda &\sim \text{Gamma}(a, b) \end{aligned}$$

Assume that $\tilde{X}|\lambda \sim$ Poisson(λ) is independent of X. Assume we have a new observation $\tilde{x}$. Find the posterior predictive distribution, $p(\tilde{x}|x)$. Assume that a is an integer.

Solution:

First, we must find $p(\lambda|x)$.

Recall

$$\begin{aligned} p(\lambda|x) &\propto p(x|\lambda)(p(\lambda) \\ &\propto e^{-\lambda}\lambda^x \lambda^{a-1} e^{-\lambda/b} \\ &= \lambda^{x+a-1} e^{-\lambda(1+1/b)}. \end{aligned}$$

Thus, $\lambda|x \sim \text{Gamma}(x + a, \frac{1}{1+1/b})$, i.e., $\lambda|x \sim \text{Gamma}(x + a, \frac{b}{b+1})$.

It then follows that

$$\begin{aligned}
p(\tilde{x}|x) &= \int_\lambda p(\tilde{x}|\lambda)p(\lambda|x)\;d\lambda \\
&= \int_\lambda \frac{e^{-\lambda}\lambda^{\tilde{x}}}{\tilde{x}!} \frac{1}{\Gamma(x+a)(\frac{b}{b+1})^{x+a}} \lambda^{x+a-1} e^{-\lambda(b+1)/b}\;d\lambda \\
&= \frac{1}{\tilde{x}!\;\Gamma(x+a)(\frac{b}{b+1})^{x+a}} \int_\lambda \lambda^{\tilde{x}+x+a-1} e^{-\lambda(2b+1/b)}\;d\lambda \\
&= \frac{1}{\tilde{x}!\;\Gamma(x+a)(\frac{b}{b+1})^{x+a}} \Gamma(\tilde{x}+x+a)(b/(2b+1)^{\tilde{x}+x+a} \\
&= \frac{\Gamma(\tilde{x}+x+a)(b/(2b+1)^{\tilde{x}+x+a}}{\tilde{x}!\;\Gamma(x+a)(\frac{b}{b+1})^{x+a}} \\
&= \frac{\Gamma(\tilde{x}+x+a)}{\tilde{x}!\;\Gamma(x+a)}\;\frac{b^{\,\tilde{x}+x+a}}{b^{x+a}}\,\frac{(b+1)^{x+a}}{(2b+1)^{\,\tilde{x}+x+a}} \\
&= \frac{(\;\tilde{x}+x+a-1)!}{(x+a-1)!\;\tilde{x}!}\,\frac{b^{\tilde{x}}\;(b+1)^{x+a}}{(2b+1)^{\tilde{x}+x+a}} \\
&= \binom{\tilde{x}+x+a-1}{\tilde{x}} \left(\frac{b}{2b+1}\right)^{\tilde{x}} \left(\frac{b+1}{2b+1}\right)^{x+a}.
\end{aligned}$$

Let $p = b/(2b+1)$, which implies $1 - p = (b+1)/(2b+1)$.

Then

$$p(\tilde{x}|x) = \binom{\tilde{x}+x+a-1}{\tilde{x}} p^{\tilde{x}}(1-p)^{x+a}.$$

Thus,

$$\tilde{x}|x \sim \text{Negative Binom}\left(x+a, \frac{b}{2b+1}\right).$$

Example 3.17: Suppose that X is the number of pregnant women arriving at a particular hospital to deliver their babies during a given month. The discrete count nature of the data plus its natural interpretation as an arrival rate suggest modeling it with a Poisson likelihood.

To use a Bayesian analysis, we require a prior distribution for θ having support on the positive real line. A convenient choice is given by the Gamma distribution, since it's conjugate for the Poisson likelihood.

The model is given by:

$$\begin{aligned} X|\lambda &\sim \text{Poisson}(\lambda) \\ \lambda &\sim \text{Gamma}(a, b). \end{aligned}$$

We are also told 42 moms are observed arriving at the particular hospital during December 2007. Using prior study information given, we are told $a = 5$ and $b = 6$. (We found a, b by working backwards from a prior mean of 30 and prior variance of 180).

We would like to find several things in this example:

1. Plot the likelihood, prior, and posterior distributions as functions of λ in `R`.
2. Plot the posterior predictive distribution where the number of pregnant women arriving falls between [0,100], integer valued.
3. Find the posterior predictive probability that the number of pregnant women arrive is between 40 and 45 (inclusive).

Solution: The first thing we need to know to do this problem are $p(\lambda|x)$ and $p(\tilde{x}|x)$. We found these in Example 3.16. So,

$$\lambda|x \sim \text{Gamma}\left(x + a, \frac{b}{b+1}\right),$$

and

$$\tilde{x}|x \sim \text{Negative Binom}\left(x + a, \frac{b}{2b+1}\right).$$

Next, we can move right into R for our analysis.

```
setwd("~/Desktop/sta4930/ch3")
lam = seq(0,100, length=500)
x = 42
a = 5
b = 6
like = dgamma(lam,x+1,scale=1)
prior = dgamma(lam,5,scale=6)
post = dgamma(lam,x+a,scale=b/(b+1))
pdf("preg.pdf", width = 5, height = 4.5)
plot(lam, post, xlab = expression(lambda), ylab= "Density", lty=2, lwd=3, type="l")
lines(lam,like, lty=1,lwd=3)
lines(lam,prior, lty=3,lwd=3)
legend(70,.06,c("Prior", "Likelihood","Posterior"), lty = c(2,1,3),
lwd=c(3,3,3))
dev.off()

##posterior predictive distribution

xnew = seq(0,100) ## will all be ints
post_pred_values = dnbinom(xnew,x+a,b/(2*b+1))
plot(xnew, post_pred_values, type="h", xlab = "x", ylab="Posterior Predictive Distribution")

## what is posterior predictive prob that number
of pregnant women arrive is between 40 and 45 (inclusive)

(ans = sum(post_pred_values[41:46])) ##recall we included 0
```

In the first part of the code, we plot the posterior, likelihood, and posterior. This should be self explanatory since we have already done an example of this.

When we find our posterior predictive distribution, we must create a sequence of integers from 0 to 100 (inclusive) using the `seq` command. Then we find the posterior predictive values using the function `dnbinom`. Then we simply plot the sequence of x_{new} on the x-axis and the corresponding posterior predictive values on the y-axis. We set `type="h"` so that our plot will appear somewhat like a smooth histogram.

Finally, in order to calculate the posterior predictive probability that the number of pregnant women arrive is between 40 and 45, we simply add up the posterior predictive probabilities that correspond to these values. We find that the posterior predictive probability that the number of pregnant women arrive is between 40 and 45 is 0.1284.

3.4 Exercises

1. Consider the following

$$
\begin{aligned}
X|\theta &\sim \text{Normal}(\theta, \sigma^2) \\
\theta &\sim \text{Normal}(\mu, \tau^2)
\end{aligned}
$$

   (a) Find the posterior distribution of θ given the data. Work out all the details (don't just simply give the result).

(b) Write down the posterior mean, $E(\theta|x)$. Write the posterior mean as a weighted average of x and the prior mean.

2. Let

$$p(y|\theta) = \theta^{-1}e^{-y/\theta},\ y > 0,\ \theta > 0.$$
$$p(\theta) = \theta^{-a}e^{-b/\theta},\ \theta > 0,\ a > 2,\ b > 0.$$

(a) Find the posterior distribution of $\theta|y$.

(b) Calculate the posterior mean and posterior variance.

(c) Notice the prior is still proper when $1 < a \leq 2$. How would such a change affect the posterior mean and posterior variance?

3. Let

$$X|\theta \sim \text{Pareto}(\theta, \beta),\ \theta \text{ unknown},\ \beta \text{ known}$$
$$\theta \sim \text{Gamma}(a, b).$$

Find the posterior distribution of $\theta|x$.

4. Consider

$$X_1, \ldots, X_n|\theta \overset{iid}{\sim} \text{Geometric}(\theta).$$

(a) Find the posterior for θ that is conjugate to the Geometric likelihood and give any relevant parameters.

(b) Using your answer in (a), derive the posterior distribution of $\theta|x_1, \ldots, x_n$.

5. Consider

$$X|\theta \sim \text{NegBin}(r, \theta), r \text{ known},\ \theta \text{ unknown},$$

where $f(x|\theta) = \binom{x-1}{r-1}\theta^r(1-\theta)^{x-r},\ x = r, r+1, \ldots;\ r > 0,\ 0 < \theta \leq 1.$

(a) Find the prior on θ that is conjugate to the likelihood above.

(b) Using your answer in (a), derive the posterior distribution of $\theta|x$.

(c) Consider $\tilde{X}|\theta \sim \text{NegBin}(r, \theta), r$ known, θ unknown. Assume that $\tilde{X}$ and X are conditionally independent given θ. Show that

$$p(\tilde{x}|x) = \binom{\tilde{x}-1}{r-1}\frac{\Gamma(2r+a)}{\Gamma(a+r)\Gamma(x+b-r)}\frac{\Gamma(x+a+b)\Gamma(\tilde{x}+x+b-2r)}{\Gamma(\tilde{x}+x+a+b)}.$$

6. Let X be a random variable with Geometric probability density function (pdf). That is,

$$P(X = x|\theta) = \theta(1-\theta)^{x-1},\ x = 1, 2, \ldots$$

Consider the prior under which $P(\theta = 1/4) = 2/3, P(\theta = 1) = 1/3$. Find the posterior distribution of θ given the data. *Hint*: You'll need to consider two cases. Look at $x = 1$ and $x > 1$ separately.

7. Suppose $X_1, \ldots, X_n|\theta \overset{iid}{\sim} \text{Exp}(\theta)$. Then $f(x) = \theta e^{-\theta x},\ x > 0,\ \theta > 0$.

   (a) State the prior distribution on θ that is conjugate to the exponential likelihood above (give its distribution and parameters). Your prior should be a continuous distribution function and take two parameter values. (Other priors you give will not be accepted for credit.)

   (b) Then derive the posterior distribution of $\theta|x_1, \ldots, x_n$.

   (c) Also give the posterior mean and variance, $E(\theta|x_1, \ldots, x_n)$ and $V(\theta|x_1, \ldots, x_n)$.

8. Let

$$\begin{aligned} X_1, \ldots, X_n|\alpha, \beta &\overset{iid}{\sim} \text{IG}(\alpha, \beta),\ \alpha \text{ known},\ \beta \text{ unknown} \\ \beta &\sim \text{Gamma}(a, b), \end{aligned}$$

Calculate the posterior distribution of $\beta|x_1, \ldots, x_n$.

9. Suppose $a < x < b$. Consider the notation $I_{(a,b)}(x)$, where I denotes the indicator function. We define $I_{(a,b)}(x)$ to be the following:

$$I_{(a,b)}(x) = \begin{cases} 1 & \text{if } a < x < b, \\ 0 & \text{otherwise.} \end{cases}$$

Let

$$\begin{aligned} X|\theta &\sim \text{Uniform}(0, \theta) \\ \theta &\sim \text{Pareto}(\alpha, \beta), \end{aligned}$$

where $p(\theta) = \dfrac{\alpha\beta^\alpha}{\theta^{\alpha+1}} I_{(\beta,\infty)}(\theta)$. Calculate the posterior distribution of $\theta|x$.

Chapter 4

Bayesian Inference

4.1 Posterior Mean and Variance

Now that we have learned how to compute the posterior distribution, we will want to compute simpler statistics that summarize the idea of the posterior. For example, in the sleep example, it would be of interest to know the average number of hours that students sleep as well as an estimate of the variance. The posterior mean and variance have different interpretations than the mean and variance did when we did frequentist calculations.

Definition 4.1: The posterior mean is $E[\theta|x]$. Recall that for discrete random variables this is

$$E[\theta|x] = \sum_{\text{all } \theta} \theta p(\theta|x).$$

For continuous random variables, we get the following:

$$E[\theta|x] = \int_\theta \theta p(\theta|x)\; d\theta.$$

Definition 4.2: The posterior variance is defined as $V(\theta|x)$.

Theorem 4.1: The posterior variance, $V(\theta|x)$, can be calculated via

$$V(\theta|x) = E(\theta^2|x) - E(\theta|x)^2.$$

Example 4.1: (Beta-Binomial Posterior Mean and Variance)
Recall in Chapter 3, we looked at the following model

$$\begin{aligned} X \mid \theta &\sim \text{Binomial}(n, \theta)\\ \theta &\sim \text{Beta}(a, b). \end{aligned}$$

We found that

$$\theta|x \sim \text{Beta}(x + a, n - x + b).$$

Recall if $Y \sim$ Beta(a,b), then $E(Y) = \dfrac{a}{a+b}$ and $V(Y) = \dfrac{ab}{(a+b)^2(a+b+1)}$. Thus,

$$E(\theta|x) = \frac{x+a}{x+a+n-x+b} = \frac{x+a}{a+b+n}.$$

Also,

$$V(\theta|x) = \frac{x+a}{(a+b+n)^2(a+b+n+1)}.$$

4.1.1 Interpretation of the Posterior Mean and Posterior Variance

In order to understand how we interpret the posterior mean, $E(\theta|x)$, and the posterior variance, $V(\theta|x)$, we must first explain a bit about a subject called decision theory.

Suppose the following:

- θ is the true parameter value (unknown) that we want to estimate, i.e., this is our target.
- It would be optimal if $\hat{\theta}$ was equal to θ every time.
- In reality, our estimate will be sometimes too low or too high compared to the target.

Consider the following:

- We suffer a penalty depending on how far off the estimate, $\hat{\theta}$, is from the target, θ.
- If the estimate is close to the target, we suffer little penalty.
- If the estimate if far away from the target, we suffer a larger penalty.

We can think of the penalty term as taking the difference between the estimate and the target and then squaring it. We call this the *squared error loss function* between $\hat{\theta}$ and θ. There are other loss functions that could be taken, for example, we could take the absolute difference of the estimate and true value. This loss function is referred to as absolute error loss.

Definition 4.3: The squared error loss function is defined to be

$$L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$$

and $\hat{\theta}$ is our estimate of θ.

If you think about the loss function defined, we would like the loss function to be as small as possible as we average over the probability distribution described by the posterior. Since we want the loss function to be as small as possible as we average over the posterior distribution, this translates to finding the value of $\hat{\theta}$ that minimizes $E[L(\theta, \hat{\theta})|X = x]$. The value that minimizes the above expression is called the *Bayes estimate,* $\hat{\theta}^B$.

Definition 4.4: The *Bayes estimate*, $\hat{\theta}^B$, is the value of θ that minimizes $E[L(\theta, \hat{\theta})|X = x]$.

Theorem 4.2: If the loss function is squared error, i.e.,

$$L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$$

then the Bayes estimate is $\hat{\theta}^B = E(\theta|x)$. In other words, the Bayes estimate is simply the posterior mean.

Theorem 4.3: If the loss function is absolute error, i.e.,

$$L(\theta, \hat{\theta}) = |\hat{\theta} - \theta|$$

then the Bayes estimate is the posterior median. If your posterior is symmetric, then the posterior mean and median are the same. For example, the Normal is symmetric. Also, if $a = b$, then the Beta(a, b) distribution is symmetric about 1/2.

Example 4.2: Suppose the University of Florida is building a new basketball arena. Define the following:

- Let θ be the number of people who want to attend games in the new arena.

- Let $\hat{\theta}$ be the number of seats in the new arena.

If the new arena contains too few seats, then the school will lose out on ticket sales. However, if the arena has too many seats, some number of seats will remain empty and the school will lose money. The school could quantify this using some loss function and then find the number of seats, $\hat{\theta}$, that minimizes the loss function, based on some data they collect.

For example, we could define

$$L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|.$$

Then the school collects some data that depends on θ. Suppose the school believes the following information:

$$\begin{aligned} X_1, \ldots, X_n | \theta &\overset{iid}{\sim} \text{Normal}(\theta, \sigma^2) \\ \theta &\sim \text{Normal}(\mu, \tau^2). \end{aligned}$$

Finally, the school would seek to find the value of $\hat{\theta}$ that minimizes

$$E[\, |\hat{\theta} - \theta| \, \big| X = x].$$

By Theorem 4.3, we know that the Bayes estimate, i.e., the posterior median, minimizes $E[\, |\hat{\theta} - \theta| \, \big| X = x]$.

Recall the posterior distribution is

$$\theta | x_1, \ldots, x_n \sim N\left(\frac{n\bar{x}\tau^2 + \mu\sigma^2}{n\tau^2 + \sigma^2}, \frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2} \right).$$

Since the Normal is symmetric about its mean, the posterior mean and median are the same. Thus, the Bayes estimate is

$$\hat{\theta}^B = \frac{n\bar{x}\tau^2 + \mu\sigma^2}{n\tau^2 + \sigma^2}.$$

Posterior Variance

Another idea we might want to consider is the precision about our posterior beliefs about θ. We may be very sure that θ is close to some value, or we may be unsure about the value of θ. One way to quantify this idea is by talking about the spread of the posterior distribution. The most common form here is calculating the *posterior variance*.

Example 4.3: Recall Example 3.7 where we had the following model:

$$\begin{aligned} X_1, \ldots, X_n | \theta &\sim \text{N}(\theta, \sigma^2) \\ \theta &\sim \text{N}(\mu, \tau^2), \end{aligned}$$

where σ^2 is known. Recall that

$$\theta|x_1,\ldots,x_n \sim N\left(\frac{n\bar{x}\tau^2+\mu\sigma^2}{n\tau^2+\sigma^2}, \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}\right).$$

Calculate the posterior mean and posterior variance under squared error loss. Also, find the Bayes estimate.

$$E[\theta|X_1,\ldots,X_n] = \frac{n\bar{x}\tau^2+\mu\sigma^2}{n\tau^2+\sigma^2}.$$

$$V[\theta|X_1,\ldots,X_n] = \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}.$$

The Bayes estimate is simply the posterior mean since we are using squared error loss.

Example 4.4: Sleep Example: Posterior Mean and Variance
Recall the Beta-Binomial example from Section 3.2 where we were interested in the proportion of the population of American college students that sleep at least eight hours each night (θ).

Recall that we had a random sample of 27 students from UF, where 11 students recorded they slept at least eight hours each night. So, we assume the data is distributed as Binomial$(27,\theta)$.

If you look back to the example, we assumed that the prior on θ was Beta(3.3,7.2). Thus, the posterior distribution is

$$\begin{aligned}&\theta|11 \sim \text{Beta}(11+3.3, 27-11+7.2), \text{ i.e.,}\\&\theta|11 \sim \text{Beta}(14.3, 23.2).\end{aligned}$$

We could easily calculate the posterior mean and variance out by hand or in R. Using the code below, we find that $E(\theta|x) = 0.381$ and $V(\theta|x) = 0.0061$. You could just as easily do this using your calculator.

Interpreting the above results, this means that on average that approximately 38 percent of college students sleep at least eight hours each night. We are averaging over the posterior beliefs about θ. We are not averaging over repeated samples as is done in frequentist analysis.

Interpreting the posterior standard deviation is a bit more natural since it's on the same scale as the posterior mean. Here, the posterior standard deviation is $\sqrt{0.0061} = 0.078$. This gives us an idea of the amount of spread in the posterior distribution around its mean.

Suppose our loss function was squared error, what would the Bayes estimate be?

```
a = 3.3
b = 7.2
n = 27
s = seq(0,1,0.01)
```

```
x = 11
post = dbeta(s,x+a,n-x+b)
a.new = x+a
b.new = n-x+b
post.mean = a.new/(a.new + b.new)
post.var = a.new*b.new/((a.new + b.new)^2*(a.new + b.new+1))}
```

Squared Error versus Absolute Error

We have discussed two loss function so far: squared error and absolute error. Why would we choose one over the other?

- Squared error is usually chosen out of convenience. It's differentiable and very easy to work with, even for very complicated expresssions.
- Absolute error is not differentiable but is robust to outliers.

4.2 Confidence Intervals versus Credible Intervals

One major difference between Bayesians and frequentists is how they interpret intervals. Let's quickly review what a frequentist confidence interval is and how to interpret one.

Frequentist Confidence Intervals

A confidence interval for an unknown (fixed) parameter, θ, is an interval of numbers that we believe is likely to contain the true value of θ. Intervals are important because they provide us with an idea of how well we can estimate θ.

DEFINITION 4.5: A *confidence interval* is constructed to contain θ a percentage of the time, say 95%. Suppose our confidence level is 95% and our interval is (L, U). Then we are 95% confident that the true value of θ is contained in (L, U) in *the long run.* In the long run means that this would occur nearly 95% of the time if we repeated our study millions and millions of times.

Misconceptions

- A confidence interval is a statement about θ (a population parameter). It is not a statement about the sample.

- Remember that a confidence interval is *not* a statement about individual subjects in the population. As an example, suppose that I tell you that a 95% confidence interval for the average amount of television watched by Americans is (2.69, 6.04) hours. This *doesn't* mean we can say that 95% of all Americans watch between 2.69 and 6.04 hours of television. We also *cannot* say that 95% of Americans in the sample watch between 2.69 and 6.04 hours of television. Beware that statements such as these are false. However, we can say that we are 95 percent confident that the *average* amount of televison watched by Americans is between 2.69 and 6.04 hours.

Bayesian Credible Intervals

Recall that frequentists treat θ as fixed, but Bayesians treat θ as a random variable. The main difference between frequentist confidence intervals and Bayesian credible intervals is the following:

- Frequentists invoke the concept of probability before observing the data. For any fixed value of θ, a frequentist confidence interval will contain the true parameter θ with some probability, e.g., 0.95.

- Bayesians invoke the concept of probability after observing the data. For some particular set of data $X = x$, the random variable θ lies in a Bayesian credible interval with some probability, e.g., 0.95.

Assumptions

In lower level classes, you wrote down assumptions whenever you did confidence intervals. This is redundant for any problem we construct in this course since we always know the data is randomly distributed and we assume it comes from some underlying distribution, say Normal, Gamma, etc. We also always assume our observations are i.i.d. (independent and identically distributed), meaning that the observations are all independent and they all have the same variance. Thus, when working a particular problem, we will assume these assumptions are satisfied given the proposed model holds.

Definition 4.6: A Bayesian credible interval of size $1 - \alpha$ is an interval (a, b) such that

$$P(a \leq \theta \leq b|x) = 1 - \alpha.$$

$$\int_a^b p(\theta|x)\, d\theta = 1 - \alpha.$$

> Remark: When you're calculating credible intervals, you'll find the values of a and b by several means. You could be asked do the following:
>
> - Use a Z-table when appropriate.
> - Use `R` to approximate the values of a and b.
> - You could be given `R` code/output and asked to find the values of a and b.

Important Point

Our definition for the credible interval could lead to many choices of (a, b) for particular problems. For this course, we will require that our credible interval have equal probability $\alpha/2$ in each tail so that our interval will be unique. That is, we will assume

$$P(\theta < a|x) = \alpha/2$$

and

$$P(\theta > b|x) = \alpha/2.$$

To see this more clearly, see Figure 5.1.

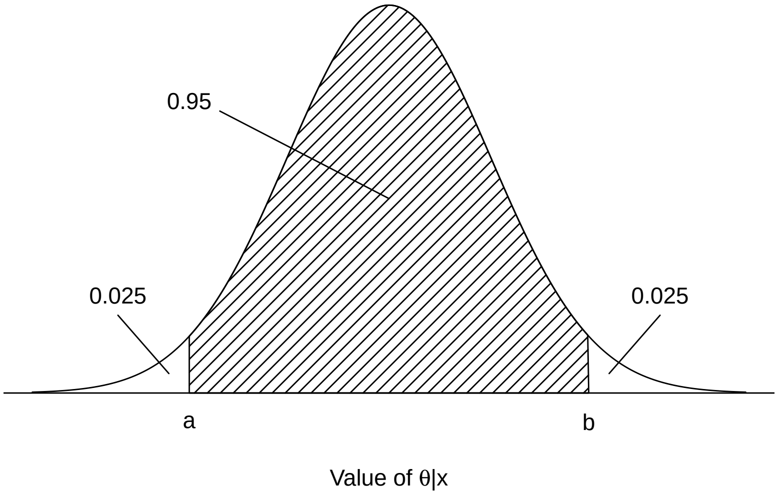


Figure 4.1: Illustration of 95% credible interval

Interpretation

We interpret Bayesian credible intervals as follows: There is a 95% probability that the true value of θ is in the interval (a, b), given the data.

Comparisons

- Conceptually, probability comes into play in a frequentist confidence interval *before* collecting the data, i.e., there is a 95% probability that we will collect data that produces an interval that contains the true parameter value. However, this is awkward, because we would like to make statements about the probability that the interval contains the true parameter value given the data that we actually observed.
- Meanwhile, probability comes into play in a Bayesian credible interval *after* collecting the data, i.e., based on the data, we now think there is a 95% probability that the true parameter value is in the interval. This is more natural because we want to make a probability statement regarding that data after we have observed it.

Example 4.5: Suppose

$$\begin{aligned} X_1 \ldots, X_n|\theta &\sim N(\theta, \sigma^2) \\ \theta &\sim N(\mu, \tau^2), \end{aligned}$$

where μ, σ^2, and τ^2 are known. Calculate a 95% credible interval for θ.

Recall

$$\theta|x_1, \ldots x_n \sim N\left(\frac{n\bar{x}\tau^2 + \mu\sigma^2}{n\tau^2 + \sigma^2}, \frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}\right).$$

Let

$$\begin{aligned} \mu^* &= \frac{n\bar{x}\tau^2 + \mu\sigma^2}{n\tau^2 + \sigma^2}, \\ \sigma^{*2} &= \frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}. \end{aligned}$$

We want to calculate a and b such that $P(\theta < a|x_1, \ldots, x_n) = 0.05/2 = 0.025$ and $P(\theta > b|x_1, \ldots, x_n) = 0.05/2 = 0.025$. So,

$$\begin{aligned} 0.025 &= P(\theta < a|x_1, \ldots, x_n) \\ &= P\left(\frac{\theta - \mu^*}{\sigma^*} < \frac{a - \mu^*}{\sigma^*}\middle| x_1, \ldots, x_n\right) \\ &= P\left(Z < \frac{a - \mu^*}{\sigma^*}\middle| x_1, \ldots, x_n\right), \text{ where } Z \sim N(0, 1). \end{aligned}$$

Thus, we now must find an a such that $P\left(Z < \frac{a-\mu^*}{\sigma^*}\middle| x_1, \ldots, x_n\right) = 0.025$. From a Z-table, we know that

$$\frac{a - \mu^*}{\sigma^*} = -1.96.$$

This tells us that $a = \mu^* - 1.96\sigma^*$. Similarly, $b = \mu^* + 1.96\sigma^*$. (Work this part out on your own at home). Therefore, a 95% credible interval is

$$\mu^* \pm 1.96\sigma^*.$$

Example 4.6: We're interested in knowing the true average number of ornaments on a Christmas tree. Call this θ. We take a random sample of n Christmas trees, count the ornaments on each one, and call the results $X_1, \ldots, X_n$. Let the prior on θ be Normal$(75, 225)$.

Using data (`trees.txt`) we have, we will calculate the 95% credible interval and confidence interval for θ. In `R` we first read in the data file `trees.txt`. We then set the initial values for our known parameters, n, σ, μ, and τ.

Next, we refer to Example 4.5, and calculate the values of μ^* and σ^* using this example. Finally, again referring to Example 4.5, we recall that the formula for a 95% credible interval here is

$$\mu^* \pm 1.96\sigma^*.$$

On the other hand, recalling back to STA 2023, a 95% confidence interval in this situation is

$$\bar{x} \pm 1.96\sigma/\sqrt{n}.$$

From the `R` code, we find that there is a 95% probability that the average number of ornaments per tree is in (45.00, 57.13) given the data. We also find that we are 95% confident that the average number of ornaments per tree is contained in (43.80, 56.20). If we compare the width of each interval, we see that the credible interval is slightly narrower. It is also shifted towards slightly higher values than the confidence interval for this data, which makes sense because the prior mean was higher than the sample mean. What would happen to the width of the intervals if we increased n? Does this make sense?

```
x = read.table("trees.txt",header=T)
attach(x)
n = 10
sigma = 10
mu = 75
tau = 15

mu.star = (n*mean(orn)*tau^2+mu*sigma^2)/(n*tau^2+sigma^2)
sigma.star = sqrt((sigma^2*tau^2)/(n*tau^2+sigma^2))

(cred.i = mu.star+c(-1,1)*qnorm(0.975)*sigma.star)
(conf.i = mean(orn)+c(-1,1)*qnorm(0.975)*sigma/sqrt(n))

diff(cred.i)
diff(conf.i)
detach(x)
```

Example 4.7: (Sleep Example Revisited)
Recall the Beta-Binomial example from Section 3.2 where we were interested in the proportion of the population of American college students that sleep at least eight hours each night (θ).

Recall that we had a random sample of 27 students from UF, where 11 students recorded they slept at least eight hours each night. So, we assume the data is distributed as Binomial$(27, \theta)$.

If you look back to the example, we assumed that the prior on θ was Beta(3.3,7.2). Thus, the posterior distribution is

$$\begin{aligned} &\theta|11 \sim \text{Beta}(11+3.3, 27-11+7.2), \text{ i.e.,} \\ &\theta|11 \sim \text{Beta}(14.3, 23.2). \end{aligned}$$

Suppose now we would like to find a 90 percent credible interval for θ. We cannot compute this in closed form since computing probabilities for Beta distributions involves messy integrals that we do not know how to compute. However, we can use `R` to find the interval.

We need to solve

$$P(\theta < c|x) = 0.05$$

and

$$P(\theta > d|x) = 0.05 \text{ for } c \text{ and } d.$$

The reason we cannot compute this in closed form is because we need to compute

$$\int_0^c \text{Beta}(14.3, 23.2)\ d\theta = 0.05$$

and

$$\int_d^1 \text{Beta}(14.3, 23.2)\ d\theta = 0.05.$$

Note that Beta(14.3,23.2) represents

$$f(\theta) = \frac{\Gamma(37.5)}{\Gamma(14.3)\Gamma(23.2)} \theta^{14.3-1}(1-\theta)^{23.2-1}.$$

The `R` code for this is very straightforward:

```
a = 3.3
b = 7.2
n = 27
x = 11
a.star = x+a
b.star = n-x+b

c = qbeta(0.05,a.star,b.star)
d = qbeta(1-0.05,a.star,b.star)
```

Running the code in `R`, we find that a 90 percent credible interval for θ is (0.256,0.514), meaning that there is a 90 percent probability that the proportion of UF students who sleep eight or more hours per night is between 0.256 and 0.514 given the data.

4.3 Bayesian Robustness

4.3.1 Motivation: Heart Transplant Example

We consider the problem of learning about the rate of success of heart transplant surgery of a particular hospital in the United States. For this hospital, we record the number of transplant surgeries, n, and the number of deaths, y, within 30 days of surgery. We assume the following model:

$$Y|\lambda \sim \text{Poisson}(e\lambda),$$

where e is the exposure and λ is the mortality rate.

A frequentist would estimate λ using $\hat{\lambda} = y/e$, however this estimate is poor when the number of deaths, y, is close to zero. In the situation when small death counts are possible, it is desirable to use Bayesian inference and take advantage of prior knowledge of the size of the mortality rate. A convenient choice for the prior distribution is a Gamma(α, β) due to conjugacy.

There are two formulations for the Gamma distribution. There is the one we have dealt with all semester in class (where β is the scale parameter), however there is another formulation that is very convenient at times (here β is taken to be the rate parameter). The second formulation is as follows:

Definition 4.7: Suppose $X \sim \text{Gamma}(\alpha, \beta)$, where β is the rate parameter. Then

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},\ x > 0,\ \alpha > 0,\ \beta > 0.$$

Suppose the prior density here is expressed as

$$p(\lambda) \propto \lambda^{\alpha-1} \exp\{-\beta\lambda\}.$$

We first consider a small hospital in Florida which experienced only one death and has exposure, $e = 66$. We will look at two different Gamma posteriors, i.e., they will have different values of α and β. We will want to see how the two priors, denoted by Prior A and Prior B, influence their corresponding posterior distributions.

- Prior A: $\lambda \sim \text{Gamma}(16, 15174)$.
- Prior B: $\lambda \sim \text{Gamma}(20, 10000)$.

We easily find that the posterior distribution

$$\lambda|y \sim \text{Gamma}(\alpha + y, \beta + e).$$

You can check that this is the correct posterior on your own.

Let's now compare the two posterior distributions of $\lambda|y$ in `R`.

To see the impact the prior has on its corresponding posterior distribution, it is helpful to display the densities on the same plot. From Figure 4.2, we see that Prior A results in a posterior with a smaller mortality rate. The estimated mortality rate using Prior B will be about twice that using Prior A.

- Clearly the choice of prior matters in this example.
- With a bigger sample size, we would see less of an effect of the prior.
- A noninformative prior might be a good choice here. We will talk more about noninformative priors in the next section.

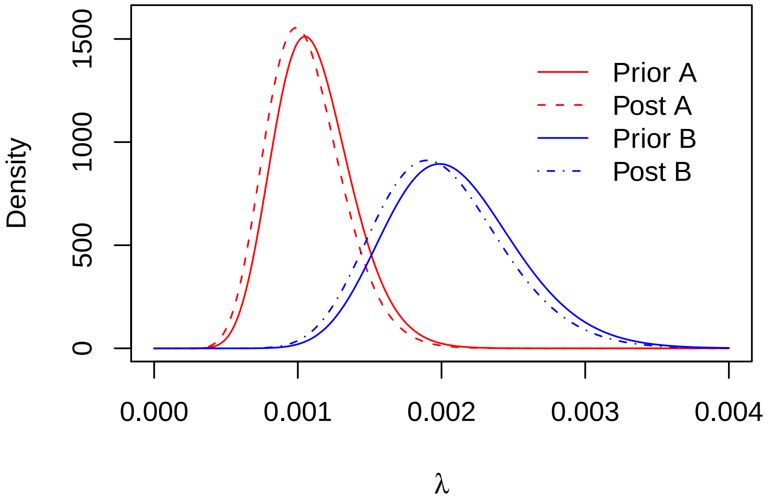


Figure 4.2: Prior $p(\lambda)$ and Posterior Distribution $p(\lambda|y)$ using Priors A and B

Below we find the R code for analyzing this example.

```
setwd("/Users/resteorts/Desktop/sta4930/ch4")
a=16; b=15174; y=1; e=66
lam = seq(0,0.004,length=500)
postA = dgamma(lam,shape = a+y, rate = b + e)
priorA = dgamma(lam,shape=a,rate=b)
a = 20; b=10000;
postB = dgamma(lam,shape = a+y, rate = b + e)
priorB = dgamma(lam,shape=a,rate=b)
pdf("heart.pdf",width=5,height=4)
plot(lam,postA, type = "l", col = "red", lty = 1,
xlab=expression(lambda),ylab="Density",ylim=c(0,1600))
lines(lam, priorA, lty = 2, col ="red")
lines(lam, postB, lty = 1, col ="blue")
lines(lam, priorB, lty = 4, col ="blue")
legend(0.0025,1500,c("Prior A","Post A","Prior B","Post B"),
lty=c(1,2,1,4),col=c("red","red","blue","blue"),bty="n")
dev.off()
```

□

Definition 4.8: A Bayesian analysis is said to be *robust* to the choice of prior if the inference is insensitive to the different priors that match the user's beliefs.

Recall that in Example 3.9, we showed that for the Normal-Normal setup, for large n, the posterior mean was approximately $\bar{x}$ and the posterior variance was approximately σ^2/n. Notice that these results do not depend on the values of μ and τ^2 in the prior. This illustrates a general property of many Bayesian models: for large n, a Bayesian analysis is usually robust. This holds not only for the Normal-Normal setup, but for most other simple setups as well.

4.4 Informative and Noninformative Priors

Thus far in this course, we have mostly considered *informative* or *subjective* priors. Ideally, we want to choose a prior reflecting our beliefs about the unknown parameter of interest. This is a *subjective* choice. All Bayesians agree that wherever prior information is available, one should try to incorporate a prior reflecting this information as much as possible. We have mentioned how incorporation of a prior expert opinion would strengthen purely data-based analysis in real-life decision problems. Using prior information can also be useful in problems of statistical inference when your sample size is small or you have a high or infinite dimensional parameter space.

However, in dealing with real-life problems you may run into problems such as

- not having past historical data

- not having an expert opinion to base your prior knowledge on (perhaps your research is cutting edge and new)
- as your model becomes more complicated, it becomes hard to know what priors to put on each unknown parameter

The problems we have dealt with all semester have been very simple in nature. We have only had one parameter to estimate (except for one example). Think about a more complex problem such as the following (we looked at this problem in Ch 3):

$$\begin{aligned} X|\theta &\sim N(\theta, \sigma^2) \\ \theta|\sigma^2 &\sim N(\mu, \tau) \\ \sigma^2 &\sim \text{IG}(a, b) \end{aligned}$$

where now θ and σ^2 are both unknown and we must find the posterior distributions of $\theta|X, \sigma^2$ and $\sigma^2|X$. For this slightly more complex problem, it is much harder to think about what values μ, τ, a, b should take for a particular problem. What should we do in these type of situations?

Often no reliable prior information concerning θ exists or inference based completely on the data is desired. It might appear that inference in such settings would be inappropriate, but reaching this conclusion is too hasty.

Suppose we could find a distribution $p(\theta)$ that contained no or little information about θ in the sense that it didn't favor one value of θ over another (provided this is possible). Then it would be natural to refer to such a distribution as a *noninformative prior.* We could also argue that all or most of the information contained in the posterior distribution, $p(\theta|x)$, came from the data. Thus, all resulting inferences were *objective* and not subjective.

Definition 4.9: *Informative/subjective priors* represent our prior beliefs about parameter values before collecting any data. For example, in reality, if statisticians are unsure about specifying the prior, they will turn to the experts in the field or experimenters to look at past data to help fix the prior.

Example 4.8: (Pregnant Mothers) Suppose that X is the number of pregnant mothers arriving at a hospital to deliver their babies during a given month. The discrete count nature of the data as well as its natural interpretation leads to adopting a Poisson likelihood,

$$p(x|\theta) = \frac{e^{-\theta}\theta^x}{x!},\ x \in \{0, 1, 2, \ldots\},\ \theta > 0.$$

A convenient choice for the prior distribution here is a Gamma(a, b) since it is conjugate for the Poisson likelihood. To illustrate the example further, suppose that 42 moms deliver babies during the month of December. Suppose from past data at this hospital, we assume a prior of Gamma$(5, 6)$. From this, we can easily calculate the posterior distribution, posterior mean and variance, and do various calculations of interest in `R`.

Definition 4.10: *Noninformative/objective priors* contain little or no information about θ in the sense that they do not favor one value of θ over another. Therefore, when we calculate the posterior distribution, most if not all of the inference will arise from the likelihood. Inferences in this case are *objective and not subjective.* Note that objective priors are typically improper, yet have proper posteriors. Let's look at the following example to see why we might consider such priors.

Example 4.9: (Pregnant Mothers Continued) Recall Example 4.8. As we noted earlier, it would be natural to take the prior on θ as Gamma(a, b) since it is the conjugate prior for the Poisson likelihood, however suppose that for this data set we do not have any information on the number of pregnant mothers arriving at the hospital so there is no basis for using a Gamma prior or any other *informative* prior. In this situation, we could take some noninformative prior.

Comment: It is worth noting that many of the objective priors are improper, so we must check that the posterior is proper.

Theorem 4.4: Propriety of the Posterior

- If the prior is proper, then the posterior will *always* be proper.
- If the prior is improper, you must check that the posterior is proper.

4.4.1 Meaning Of Flat

What does a "flat prior" really mean? People really abuse the word flat and interchange it for noninformative. Let's talk about what people really mean when they use the term "flat," since it can have different meanings.

Example 4.10: Often statisticians will refer to a prior as being flat, when it actually looks flats. An example of this would be taking such a prior to be

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

We can plot the density of this prior to see that the density is flat.

Example 4.11: Now suppose we consider Jeffreys' prior, $p_J(\theta)$, where $X \sim \text{Bin}(n, \theta)$.

We calculate Jeffreys' prior by finding the Fisher information. The Fisher information tells us how much information the data gives us for certain parameter values.

In this example, it can be shown that $p_J(\theta) \propto \text{Beta}(1/2, 1/2)$. Let's consider the plot of this prior. Flat here is a purely abstract idea. In order to achieve objective inference, we need to compensate more for values on the boundary than values in the middle.

Example 4.12: Finally, we consider the following prior on θ :

$$\theta \sim N(0, 1000).$$

What happens in this situation? We look at two plots to consider the behavior of this prior.

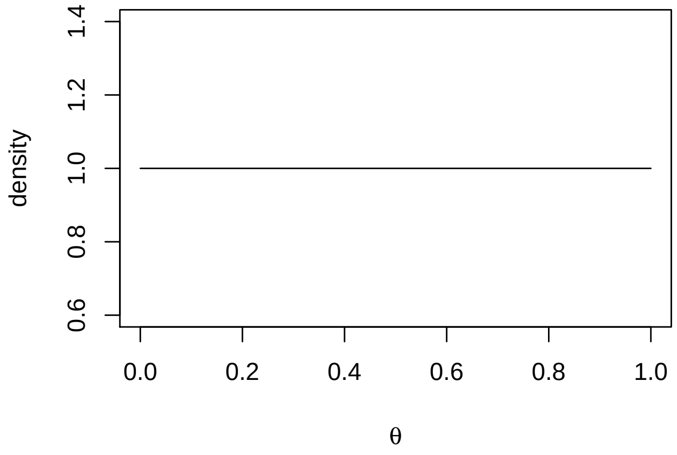


Figure 4.3: Unif(0,1) prior

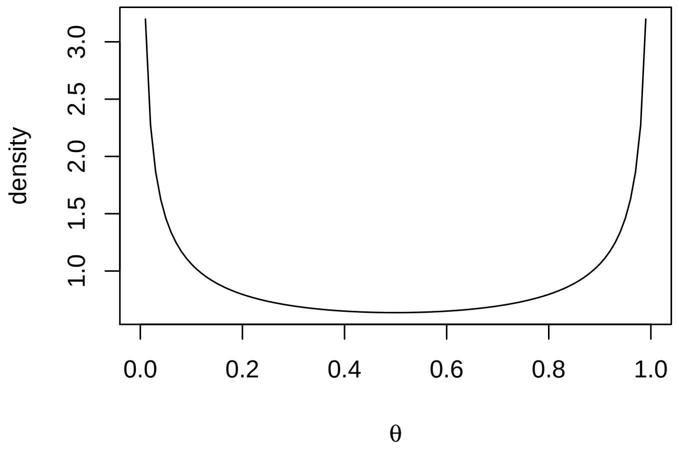


Figure 4.4: Jeffreys' prior for Binom likelihood

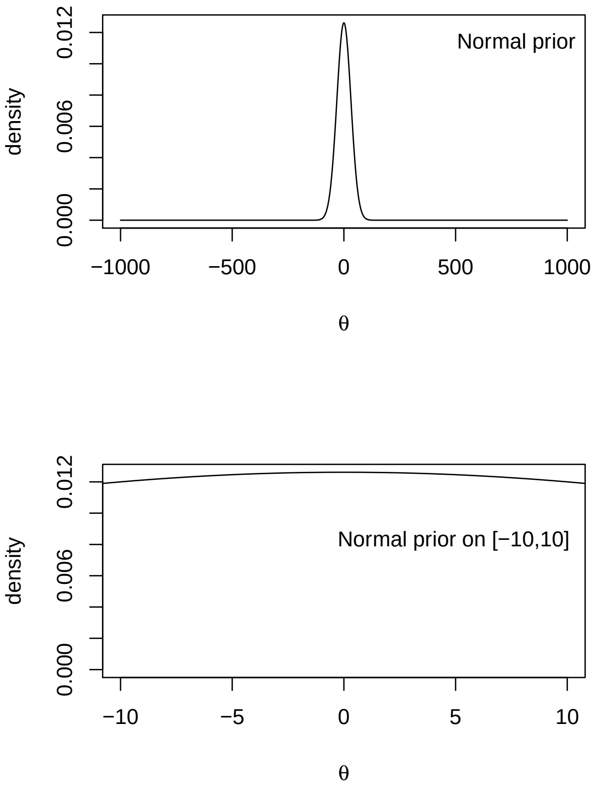


Figure 4.5: Normal priors

4.4.2 Objective Priors in More Detail

Uniform Prior of Bayes and Laplace

Example 4.13: (Thomas Bayes) In 1763, Thomas Bayes considered the question of what prior to use when estimating a binomial success probability p. He described the problem quite differently back then by considering throwing balls onto a billiard table. He separated the billiard table into many different intervals and considered different events. By doing so (and not going into the details of this), he argued that a Uniform(0,1) prior was appropriate for p.

Example 4.14: (Laplace) In 1814, Pierre-Simon Laplace wanted to know the probability that the sun will rise tomorrow. He answered this question using the following Bayesian analysis:

- Let X represent the number of days the sun rises. Let p be the probability the sun will rise tomorrow.
- Let $X|p \sim \text{Bin}(n, p)$.
- Suppose $p \sim \text{Uniform}(0, 1)$.
- Based on reading the Bible, we compute the total number of days, n, and the total number of days, x, recorded in history in which the sun rose. Clearly, $x = n$.

Then

$$\begin{aligned}\pi(p|x) &\propto \binom{n}{x} p^x (1-p)^{n-x} \cdot 1 \\ &\propto p^{x+1-1}(1-p)^{n-x+1-1}\end{aligned}$$

This implies

$$p|x \sim \text{Beta}(x+1, n-x+1)$$

Then

$$\hat{p} = E[p|x] = \frac{x+1}{x+1+n-x+1} = \frac{x+1}{n+2} = \frac{n+1}{n+2}.$$

Thus, Laplace's estimate for the probability that the sun rises tomorrow is $(n+1)/(n+2)$, where n is the total number of days recorded in history. For instance, if so far we have encountered 100 days in the history of our universe, this would say that the probability the sun will rise tomorrow is $101/102 \approx 0.9902$. However, we know that this calculation is ridiculous because the sun always rises. Thus, objective Bayesian methods shouldn't be recklessly applied to every problem we study.

Criticism of the Uniform Prior

The Uniform prior of Bayes and Laplace and has been criticized for many different reasons. We will discuss one important reason for criticism and not go into the other reasons since they go beyond the scope of this course.

In statistics, it is often a good property when a rule for choosing a prior is *invariant* under what are called one-to-one transformations. Invariant basically means unchanging in some sense. The invariance principle means that a rule for choosing a prior should provide equivalent beliefs even if we consider a transformed version of our parameter, like p^2 or $\log p$ instead of p.

Jeffreys' Prior

One prior that is invariant under one-to-one transformations is Jeffreys' prior (*Bayesian Data Analysis*, Gelman, *et al.*, p. 63). You are *not* responsible for this invariance proof (calculus is involved).

What does the invariance principle mean? Suppose our prior parameter is θ, however we would like to transform to ϕ.

Define $\phi = f(\theta)$.

Jeffreys' prior says that if θ has the distribution specified by Jeffreys' prior for θ, then $f(\theta)$ will have the distribution specified by Jeffreys' prior for ϕ. If this isn't clear, we will make this more clear by going over two examples to illustrate this idea.

However, suppose θ has a Uniform prior. Then one can show $\phi = f(\theta)$ will not have a Uniform prior.

Aside from the invariance property of Jeffreys' prior, in the univariate case, Jeffreys' prior satisfies many optimality criteria that statisticians are interested in. One such that you might have heard of is called the Kullback-Leibler divergence.

Definition 4.11: Define

$$I(\theta) = -E\left[\frac{\partial^2 \log p(y|\theta)}{\partial\theta^2}\right],$$

where $I(\theta)$ is called the Fisher information. Then *Jeffreys' prior* is defined to be

$$p_J(\theta) = \sqrt{I(\theta)}.$$

> Since calculating Jeffreys' prior involves calculus, for a problem you will be either given the Fisher information or Jeffreys' prior. You will *not* be expected to compute the Fisher information. You should understand why it would be desirable to use Jeffreys' prior though and that its form changes whenever the likelihood changes.

Example 4.15: (Uniform Prior is Not Invariant to Transformation)
Any calculus that is given, you are not responsible for knowing. This example is given to simply illustrate the ideas just presented.

Let $\theta \sim \text{Uniform}(0, 1)$. Suppose now we would like to transform from θ to θ^2.

Let $\phi = \theta^2$. Then $\theta = \sqrt{\phi}$. It follows (by calculus) that

$$\frac{\partial\theta}{\partial\phi} = \frac{1}{2\sqrt{\phi}}.$$

Thus, $p(\phi) = \dfrac{1}{2\sqrt{\phi}}$, $0 < \phi < 1$ which shows that ϕ is not Uniform on $(0, 1)$. Hence, the transformation is not invariant. Criticism such as this led to consideration of Jeffreys' prior.

Example 4.16: (Jeffreys' Prior Invariance Example)
Any calculus that is given, you are not responsible for knowing. This example is given to simply illustrate the ideas just presented.

Suppose

$$X|\theta \sim \text{Exp}(\theta)$$

One can show using calculus that $I(\theta) = 1/\theta^2$. Then $p_J(\theta) = 1/\theta$. Suppose that $\phi = \theta^2$. It follows (by calculus) that

$$\frac{\partial\theta}{\partial\phi} = \frac{1}{2\sqrt{\phi}}.$$

By calculus, we know

$$\begin{aligned} p_J(\phi) &= p_J(\sqrt{\phi}) \left|\frac{\partial\theta}{\partial\phi}\right| \\ &= \frac{1}{\sqrt{\phi}} \frac{1}{\sqrt{2\phi}} \propto \frac{1}{\phi}. \end{aligned}$$

Hence, we have shown for this example, that Jeffreys' prior is invariant under the transformation $\phi = \theta^2$.

Example 4.17: (Jeffreys' prior) Suppose

$$X|\theta \sim \text{Binomial}(n, \theta)$$

Let's calculate the posterior using Jeffrey's prior. We are given that

$$\begin{aligned} p_J(\theta) &= \sqrt{\frac{n}{\theta(1-\theta)}} \\ &\propto \text{Beta}(1/2, 1/2). \end{aligned}$$

We then find that

$$\begin{aligned} p(x|\theta) &\propto \theta^x(1-\theta)^{n-x}\theta^{1/2-1}(1-\theta)^{1/2-1} \\ &= \theta^{x-1/2}(1-\theta)^{n-x-1/2} \\ &= \theta^{x-1/2+1-1}(1-\theta)^{n-x-1/2+1-1}. \end{aligned}$$

Thus, $\theta|x \sim \text{Beta}(x + 1/2, n - x + 1/2)$, which is a proper posterior since the prior is proper.

Note: It is very important to check that your posterior is proper.

Haldane's Prior

In 1963, Haldane introduced the following improper prior:

$$p(\theta) \propto \theta^{-1}(1-\theta)^{-1}.$$

It can be shown to be improper using simple calculus, which we will not go into. However, the posterior is proper under certain conditions.
Let

$$Y|\theta \sim \text{Bin}(n, \theta).$$

Calculate $p(\theta|y)$ and show that it is improper when $y = 0$ or $y = n$.

> Remark: Recall that for a Binomial distribution, Y can take values $y = 0, 1, 2, \ldots, n$.

We will first calculate $p(\theta|y)$.

$$\begin{aligned} p(\theta|y) &\propto \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}}{\theta(1-\theta)} \\ &\propto \theta^{y-1}(1-\theta)^{n-y-1} \\ &= \theta^{y-1}(1-\theta)^{(n-y)-1}. \end{aligned}$$

The density of a Beta(a, b) is the following:

$$f(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad \theta > 0$$

This implies that $\theta|Y \sim \text{Beta}(y, n-y)$.

We need to check that our posterior is proper finally. Recall that the parameters of the Beta need to be positive. Thus, $y > 0$ and $n - y > 0$. This means that $y \neq 0$ and $n \neq y$ in order for the posterior to be proper.

> Remark: Recall that the Beta density must integrate to 1 whenever the parameter values are positive. Hence, when they are not positive, the density does not integrate to 1 and integrates to ∞. Thus, for the problem above, when $y = 0$ and $n = y$ the density is improper.

There are many other objective priors that are used in Bayesian inference, however, this is the level of exposure that we will cover in this course. If you're interested in learning more about objective priors (g-prior, probability matching priors), see me and I can give you some references.

4.5 Exercises

1. Suppose that
$$X_1, \ldots, X_n|\theta \overset{iid}{\sim} \text{Poisson}(\theta).$$
   (a) What prior is conjugate for the Poisson likelihood? Give the distribution for θ along with any associated parameters.

   (b) Calculate the posterior distribution of $\theta|x_1, \ldots, x_n$ using your prior in (a).

   (c) Find the posterior mean.

   (d) Write the posterior mean as a weighted average of the prior mean and the sample average. What are the weights of the prior mean and the sample average?

   (e) What happens to the posterior mean if $n >> b$? What happens to the posterior mean if $n << b$? The symbol $>>$ stands for much greater than, and $<<$ represents much less than.

2. Let
$$\begin{aligned} Y|\theta &\sim \text{Exp}(\theta) \\ \theta &\sim \text{Gamma}(a, b). \end{aligned}$$
   Suppose we have a new observation $\tilde{Y}|\theta \sim \text{Exp}(\theta)$, where conditional on θ, Y and $\tilde{Y}$ are independent. Show that
$$p(\tilde{y}|y) = \frac{b(a+1)(by+1)^{a+1}}{(b\tilde{y} + by + 1)^{a+2}},$$
   where a is an integer. (Note that this is a valid density function that integrates to 1).

3. Let
$$X|\theta \sim \text{Geometric}(\theta).$$
Consider the prior under which $P(\theta = 1/4) = 2/3, P(\theta = 1) = 1/3.$
$$p(\theta|x > 1) = \begin{cases} 0 & \text{if } \theta = 1 \\ 1 & \text{if } \theta = 1/4. \end{cases}$$
Suppose $\tilde{X}|\theta \sim \text{Geometric}(\theta)$. Assume that given θ, $\tilde{X}$ and X are independent.

   Derive $p(\tilde{x}|x)$ for the cases $x = 1$ and $x > 1$. That is, show the following:

   - If $x = 1$,
   $$p(\tilde{x}|x = 1) = \begin{cases} 3/4 & \text{if } \tilde{x} = 1 \\ \dfrac{1}{12}\left(\dfrac{3}{4}\right)^{\tilde{x}-1} & \text{if } \tilde{x} > 1. \end{cases}$$
   - If $x > 1$, $\tilde{X}|x \sim \text{Geometric}(1/4)$.

4. Refer to Example 3.10. Recall
$$X_1, \ldots, X_n|\alpha, \beta \overset{iid}{\sim} \text{Gamma}(\alpha, \beta), \alpha \text{ known}, \beta \text{ unknown}$$
$$\beta \sim \text{IG}(a, b).$$
   (a) Find the posterior of $\beta|x_1, \ldots, x_n$.

   (b) Suppose we take our loss function to be squared error loss. What is the Bayes estimate in terms of the problem (be specific)? Explain what the Bayes estimate means in words.

   (c) If we used absolute error loss instead, what would the Bayes estimate be?

5. Find a 90 percent credible interval for λ using the methods of Chapter 4 for the following model:
$$X|\lambda \sim \text{Poisson}(\lambda)$$
$$\lambda \sim \text{Gamma}(a, b),$$
where X is the number of babies delivered at a hospital during a particular month and λ is the average number of babies that are delivered per month. Be sure to interpret your interval in terms of the problem. *Hint:* You'll want to use R to solve this problem, so include any code with your write up. For full credit make it very clear what you're doing in your write up and in R. Assume the following: $x = 42, a = 5, b = 6$.

6. Suppose

$$X_1, \ldots, X_n | \theta \overset{iid}{\sim} \text{Poisson}(\theta).$$

Assume $I(\theta) = \dfrac{n}{\theta}$.

(a) Find Jeffreys' prior. Is it proper or improper? Show that it's proper or improper without using any calculus.

(b) Find $p(\theta | x_1, \ldots, x_n)$ under Jeffreys' prior.

7. Suppose

$$\begin{aligned} X_1, \ldots, X_n | \sigma^2 &\overset{iid}{\sim} \text{Normal}(0, \sigma^2) \\ p_J(\sigma^2) &\propto \sigma^{-2}. \end{aligned}$$

Find $p(\sigma^2 | x_1, \ldots, x_n)$.

Chapter 5

Bayesian Risk and Minimaxity

5.1 Motivations

What is risk and why do we study it? *Risk* is the average amount we expect to lose in the long run. For example, if we own 10 shares of Coca-Cola stock, then the risk is the average amount we expect to lose over the long run (or we could interpret this to the be average amount we gain).

We think about risk inadvertently whether we realize it or not when we make important decisions. As statisticians, risk is something we would like to minimize.

Example 5.1: (Nancy Reagan) Faced with early-stage breast cancer, former First Lady Nancy Reagan weighed the pros and cons that went along with having a mastectomy, instead of a lumpectomy. At the time, Nancy Reagan said,

"At the time of my operation, there were some people, including doctors, who thought I have taken too drastic a step in choosing the mastectomy instead of lumpectomy, which involved removing only the tumor itself and a small amount of tissue—but also weeks of radiation. I resented these statements, and I still do. This is a very personal decision, one that each woman must make for herself. This was my choice, and I don't believe I should have been criticized for it. For some women, it would have been wrong, but for me it was right. Perhaps, if I had been 20 years old and unmarried, I would have made a different decision. But I've already had my children and I have a wonderful understanding husband."

This example shows the risk analysis that Nancy Reagan went through in deciding between two surgery operations. People make these same types of decisions every day. We will show how to make them from a statistical point of view so that we can make the best choice possible.

5.2 Introduction to Loss and Risk

After we observe the data $\boldsymbol{X} = \boldsymbol{x}$, we can make a decision regarding θ. We define the set of allowable decisions to be the action space, $\mathcal{A}$. Often $\mathcal{A}$ is the same as the parameter space, in which case the decision d is equivalent to estimating the value of θ.

Definition 5.1: (Loss Function) The *loss function*, $L(\theta, d)$, is the amount of loss incurred from estimating the parameter θ by some statistic d (a function of the data).

The loss function in an estimation problem reflects the fact that if some decision d is close to an unknown parameter θ, then the decision d is reasonable and little loss is incurred. However, if d is far away from θ, a large loss is incurred. Our decision is based on the data X, so we will usually write the decision as $d(X)$.

Example 5.2: (Examples of Loss Functions)

- Squared Error: $L(\theta, d) = (d - \theta)^2$.
- Absolute Value: $L(\theta, d) = |d - \theta|$.

The squared error loss function penalizes more heavily than the absolute value loss function when d is farther away from θ.

Definition 5.2: (Risk) The (frequentist) risk is defined as

$$R(\theta, d) = E_\theta[L(\theta, d(X))],$$

which is the average loss over time that results from using the estimator d.

> Remark 1: We write the subscript θ on the expectation because the risk is calculated as if the value of θ is fixed, i.e., each value of θ gives a corresponding value of the risk.
>
> Remark 2: We can write the risk as
>
> $$R(\theta, d) = \int_x L(\theta, d(x))p(x|\theta)\ dx$$
>
> or
>
> $$R(\theta, d) = \sum_x L(\theta, d(x))p(x|\theta).$$

Definition 5.3: (Bayes Risk) The Bayes risk is defined as

$$r(d) = E[R(\theta, d)],$$

where the expectation is taken over some prior $p(\theta)$ that specifies the probabilities of different values of θ.

Remark 1: We can write the Bayes risk as

$$r(d) = \int_\theta R(\theta, d)p(\theta)\, d\theta$$

or

$$r(d) = \sum_\theta R(\theta, d)p(\theta).$$

Definition 5.4: Basic Decision Problem
The general non-sequential decision theory consists of three basic principles.

1. The possible states of nature, denoted by Θ, sometimes referred to as the parameter space.
2. A set of actions, $\mathcal{A}$, available to the statistician.
3. A loss function, $L(\theta, a)$, where θ is any state of nature in Θ and a is any action in $\mathcal{A}$.

The triplet expression $(\Theta, \mathcal{A}, L)$ defines what is called a game. The game is played as follows. Nature selects a point in Θ, and the statistician without being informed of the choice of nature, chooses an action (or decision) in $\mathcal{A}$. As a consequence of these two choices, the statistician loses an amount $L(\theta, a)$.

Example 5.3: Two Contestants
Two contestants simultaneously put up either one or two fingers. One of the players, say player I, wins if the sum of the digits showing is **odd**. The other player, say player II, wins if the sum of the digits showing is **even**. The winner in all cases receives in dollars the sum of the digits showing, which is paid to him by the loser.

We label player I as nature and player II as the statistician. If we denote 1 and 2 by the respective decisions to put up one or two fingers, then $\Theta = \mathcal{A} = \{1, 2\}$.

We can easily write down the loss functions for this example:

$$L(1,1) = -2,\ L(1,2) = L(2,1) = 3,\ L(2,2) = -4.$$

We can also view the loss function in the form of a table.

$\mathcal{A}, \Theta$	1	2
1	-2	3
2	3	-4

Table 5.1: Loss function of two players from view of statistician

Example 5.4: Recall the set up of Example 5.3. Suppose before the game is played, player II (statistician) is allowed to ask player I (nature) how many fingers he intends to put up. Suppose nature must answer truthfully with probability 3/4 (and thus untruthfully with probability 1/4).

Therefore, the statistician therefore observes a random variable X (the answer nature gives) taking either the values 1 or 2. Note that

$$f_1(1) = 3/4,\ f_1(2) = 1/4,\ f_2(1) = 1/4,\ f_2(2) = 3/4.$$

There are exactly four possible decision rules d_1, d_2, d_3, d_4 with

$$d_1(1) = d_1(2) = 1$$

$$d_2(1) = 1, d_2(2) = 2$$
$$d_3(1) = 2, d_3(2) = 1$$
$$d_4(1) = d_4(2) = 2$$

Let's calculate $R(\theta, d_1)$. Suppose

$$\begin{aligned} R(\theta, d_1) &= \sum_x L(\theta, d_1) p(x|\theta) \\ &= L(\theta, d_1(1)) f_\theta(1) + L(\theta, d_1(2)) f_\theta(2). \end{aligned}$$

Recall $\theta = 1, 2$.

$$\begin{aligned} R(1, d_1) &= L(1, d_1(1)) f_1(1) + L(1, d_1(2)) f_1(2) \\ &= L(1, 1) \\ &= -2. \end{aligned}$$

$$\begin{aligned} R(2, d_1) &= L(2, d_1(2)) f_2(1) + L(2, d_1(2)) f_2(2) \\ &= L(2, 1) \\ &= 3. \end{aligned}$$

It is easy to calculate the (frequentist) risk for d_2, d_3, d_4.

5.2.1 How Often Should the Statistician Carry His Umbrella?

Suppose a statistician at the University of Florida doesn't particularly like getting wet, so he's interested in knowing how often he should carry his umbrella given the probability that it will rain on a given day.

		Statistician	
		Takes Umbrella (T)	Doesn't Take (D)
Nature	Rain (R)	0	10
	No Rain (N)	1	0

Table 5.2: Loss Function

First, let's consider what our loss function is in terms of the table above. Note that

$$L(\theta, d) = \begin{cases} 0, & \text{for } L(R,T), \\ 10, & \text{for } L(R,D), \\ 1, & \text{for } L(N,T), \\ 0, & \text{for } L(N,D). \end{cases}$$

Notice that in this example, there is no data X. That means that for any particular state of nature (R or N) and any particular decision (T or D), there is nothing random going on. Thus, $R(\theta, d) = E_\theta[L(\theta, d(X))]$ is really just $E_\theta[L(\theta, d)] = L(\theta, d)$, so the risk and the loss are the same. This is *always* the case in no data problems.

Suppose we can predict "Rain" with 100% accuracy. Let's now find the value of d that minimizes $R(\theta, d)$.

Solution:

$$R(R, d) = \begin{cases} 0, & \text{if } d = T, \\ 10, & \text{if } d = D. \end{cases}$$

The d that minimizes the risk above is $d = T$, meaning the statistician would take his umbrella. Similarly, if we know $\theta = N$ with 100% accuracy, then we decide D, and the statistician decides to leave his umbrella at home.

The cases above are unreasonable, so let's consider the situation where we know

$$\theta = \begin{cases} R, & \text{with probability } p, \\ N, & \text{with probability } 1 - p. \end{cases}$$

This is a prior $p(\theta)$ on the values of θ.

Now we would like to minimize the Bayes risk, $r(d) = E[R(\theta, d)]$.

Solution: If the statistician takes the umbrella, then

$$r(d) = E[R(\theta, T)] = p \cdot 0 + (1-p) \cdot 1 = 1 - p.$$

If the statistician decides to leave his umbrella at home then

$$r(d) = E[R(\theta, D)] = p \cdot 10 + (1-p) \cdot 0 = 10p.$$

If $1 - p < 10p$, then the statistician should take his umbrella. On the other hand, if $1 - p > 10p$, the statistician should leave his umbrella at home. Note that we have minimized the Bayes risk.

As an example, if $p = 0.2$, then the risk from taking is $1 - p = 0.8$ and from not taking is $10p = 2$. So, the statistician assumes less Bayes risk of getting wet by carrying his umbrella around.

The statistician might now wonder what the changepoint value is for this problem. That is, what is the value of p when the statistician takes and doesn't take his umbrella that results in the two situations having equal Bayes risk. To find this value of p, we simply solve

$$1 - p = 10p$$

and find that $p = 1/11$.

As a homework exercise, redo with a general $L(R, D) = q$ and see how sensitive the risk and Bayes risk are for different values of q. You may use `R` for this. □

5.3 Minimaxity

We now introduce a concept that appears as an "insurance against the worst case" in the sense that it aims to minimize the expected loss in the least favorable case. This idea appears in economics and not only in statistics (a subject called game theory), where two adversaries compete. In our case, they will usually be the Statistician and Nature. Once the statistician has determined the procedure at hand, nature selects the state of nature, i.e., the parameter, which maximizes the loss of the statistician.

Definition 5.5: An estimator is minimax if it minimizes the maximum risk. For solving problems, we will first maximize the risk over all possible parameter values. Then we find the estimator that minimizes this maximum risk.

Theorem 5.1: If the Bayes estimate, $\hat{\theta}_B$, has constant (frequentist) risk under the given prior, then $\hat{\theta}_B$ is considered to be minimax.

Example 5.5: The first oil drilling platforms in the North Sea were designed according to a minimax principle. They were supposed to resist the conjugate action of the worst gale (wind) and worst storm ever observed, at the minimal record temperature. This strategy obviously gives a very comfortable margin of safety, but at a great margin of cost. For more recent platforms, engineers have taken into account the distribution of these weather phenomena in order to reduce the production cost.

Example 5.6: (Umbrella Example Continued) Recall our example where we considered a statistician carrying an umbrella and we found the following loss function:

$$L(\theta, d) = \begin{cases} 0, & \text{for } L(R,T), \\ 10, & \text{for } L(R,D), \\ 1, & \text{for } L(N,T), \\ 0, & \text{for } L(N,D). \end{cases}$$

Since this is a no data problem, $R(\theta, d) = L(\theta, d)$. Then to find the minimax estimator d, we first maximize over all possible values of θ for each estimator d, i.e., we maximize over rain and not rain. The maximum risks for D and T are $R(R, D) = 10$ and $R(N, T) = 1$. Then minimizing the risk functions over the estimators, we find $R(N, T) = 1$. So, the minimax estimator is $d = T$, or rather for the statistician to *always* carry his umbrella. □

Example 5.7: Two people, A and B, are suspected of committing a robbery and have been apprehended. They are placed in separate cells. Both suspects are questioned and enticed to confess to the burglary. If neither person talks, then they will both serve very little jail time. However, the incentive offered by the police is that the person who cooperates first will get immunity from prosecution. Table 5.3 gives the loss as perceived by person A (in years of prison time), where a_1 denotes that person A talks and a_2 denotes that A doesn't talk. Similarly, θ_1 denotes the state of nature that B talks, while θ_2 represents the state of nature that B doesn't talk.

	a_1	a_2
θ_1	10	20
θ_2	0	1

Table 5.3: Loss Function for Robbery Example

Note that this is a no data problem, so the frequentist risk equals the loss.

We want to find the minimax estimator a_1 or a_2 in terms of person A.

Solution: Step 1: We want to calculate the maximum risk over θ_1 and θ_2 for each choice of a_1 and a_2. Doing so, we get $R(\theta_1, a_1) = 10$ and $R(\theta_1, a_2) = 20$.

Step 2: Now we minimize over all values of a. This leads to $R(\theta_1, a_1) = 10$. This implies the minimax estimator is a_1 or rather that person A should confess.

Similarly, if we look at the risk for person B, we find the minimax estimator should be that person B should also confess. Thus, minimaxity leads to both people going to jail for the full 20 years.

Intuitively above, we're picking the worst case scenario for each situation. The we pick the one with the "best" worst case scenario.

□

Example 5.8: Minimax Example
Suppose $X \sim \text{Geo}(\theta)$. Then

$$P(X = x|\theta) = \theta(1-\theta)^{x-1},\ x = 1, 2, \ldots$$

Consider the prior under which $P(\theta = 1/4) = 2/3,\ P(\theta = 1) = 1/3$. Find the Bayes estimate of θ under the above prior using squared error loss. Recall from a previous homework assignment

$$P(\theta = 1/4|x = 1) = 1/3$$

and

$$P(\theta = 1|x = 1) = 2/3.$$

Also for

$$k > 1, P(\theta = 1|X = k) = 0,$$

which implies

$$P(\theta = 1/4|X = k) = 1 \text{ for all } k > 1.$$

Hence, the Bayes estimate under squared error loss under the given prior is simply the posterior mean. If $k = 1$,

$$d(1) = \hat{\theta}^B = (1/4)(1/3) + (1)(2/3) = 3/4.$$

If $k > 1$,

$$d(k) = \hat{\theta}^B = 1/4.$$

Now consider

$$\begin{aligned}
R(\theta, d) &= \sum_x L(\theta, \hat{\theta}) p(x|\theta) \\
&= (3/4-\theta)^2 P(X = 1|\theta) + (1/4-\theta)^2 P(X \neq 1|\theta) \\
&= (3/4-\theta)^2\theta + (1/4-\theta)^2(1-\theta) \\
&= 1/16.
\end{aligned}$$

Since the Bayes estimate has constant risk under the prior given, we know that the Bayes estimate is minimax.

5.4 Choosing the Best Estimator

We've looked at different estimators throughout the course such as the mle and the Bayes estimate. But which one is best? The is often a very hard question to answer and is in a sense beyond the scope of this course. The main conclusion we can reach, however, is that usually one estimator is not uniformly better than another one. The next example illustrates why.

Example 5.9: Let $X_1, \ldots, X_{16}|\theta$ constitute a random sample of size 16 from $\mathrm{N}(\theta, 1)$. We will assume the squared error loss function out of convenience. That is,

$$L(\theta, d) = (\theta - d)^2.$$

The following two estimators are suggested:

$$d_1(x) = \frac{1}{16} \sum_i x_i$$

and

$$d_2(x) = 0.$$

We can easily compute the frequentist risk.

$$R(\theta, d_1(x)) = E_\theta[(\bar{X} - \theta)^2].$$

Recall

$$\bar{X} \sim \mathrm{N}(\theta, 1/16).$$

This implies

$$\bar{X} - \theta \sim \mathrm{N}(0, 1/16).$$

Thus,

$$R(\theta, d_1(x)) = E_\theta[(\bar{X} - \theta)^2] = 1/16.$$

Clearly,

$$R(\theta, d_2(x)) = \theta^2.$$

Plotting both of these, we find that neither d_1 nor d_2 is uniformly better than the other.

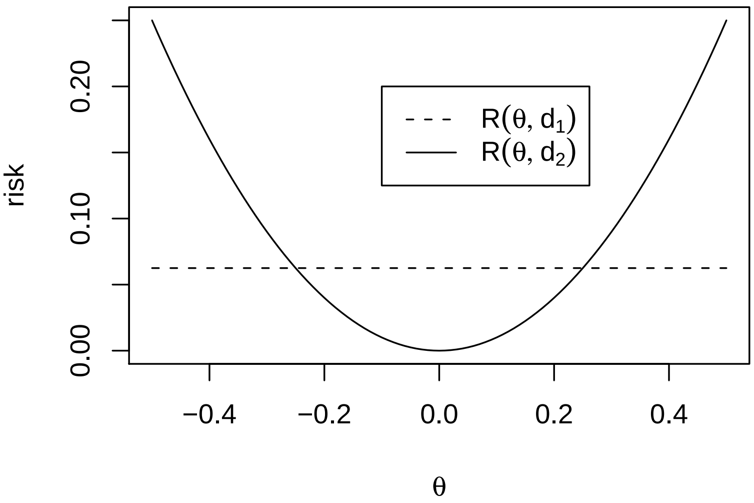


Figure 5.1: Frequentist risk plot for d_1 and d_2

Chapter 6

Monte Carlo Methods

6.1 A Quick Review of Monte Carlo Methods

Monte Carlo methods are needed in situations when we have some integral form of a density that we cannot evaluate (even if we could use calculus tools). Luckily, we can numerically do things to help. What happens is that we will often in real problems encounter such functions as $\int h(x)f(x)dx$, where $h(x)f(x)$ is a complicated density function where we cannot evaluate $\int h(x)f(x)dx$ at all. What should we do? We will go through many different types of solutions, where we will estimate the integral function. A Monte Carlo method is one way of doing this. Basically, we simulate or approximate what we can not evaluate in software and then evaluate how well we are doing using the standard error of our estimate.

One motivation for Monte Carlo methods is to approximate an integral of the form $\int_X h(x)f(x)\ dx$ that is intractable, where f is a probability density. You might wonder why we wouldn't just use numerical integration techniques.

There are a few reasons:

- The most serious problem is the so-called "curse of dimensionality." Suppose we have a p-dimensional integral. Numerical integration typically entails evaluating the integrand over some grid of points. However, if p is even moderately large, then any reasonably fine grid will contain an impractically large number of points. For example if $p = 6$, then a grid with just ten points in each dimension—already too coarse for any sensible amount of precision—will consist of 10^6 points. If $p = 50$, then even an absurdly coarse grid with just *two* points in each dimension will consist of 2^{50} points (note that $2^{50} > 10^{15}$).
- There can still be problems even when the dimensionality is small. There are packages in R called **area** and **integrate**, however, area cannot deal with infinite bounds in the integral, and even though integrate can handle infinite bounds, it is fragile and often produces output that's not trustworthy (Robert and Casella, 2010).

6.1.1 Classical Monte Carlo Integration

The generic problem here is to evaluate $E_f[h(x)] = \int_X h(x)f(x)\,dx$. The classical way to solve this is to generate a sample $(X_1, \ldots, X_n)$ from f and propose as an approximation the empirical average

$$\bar{h}_n = \frac{1}{n}\sum_{j=n}^{n} h(x_j).$$

Why? It can be shown that $\bar{h}_n$ converges a.s. (i.e. for almost every generated sequence) to $E_f[h(X)]$ by the Strong Law of Large Numbers.

Also, under certain assumptions (which we won't get into, see Casella and Robert, page 65, for details), the asymptotic variance can be approximated and then can be estimated from the sample $(X_1, \ldots, X_n)$ by

$$v_n = 1/n^2 \sum_{j=1}^{n} [h(x_j) - \bar{h}_n]^2.$$

Finally, by the CLT (for large n),

$$\frac{\bar{h}_n - E_f[h(X)]}{\sqrt{v}_n} \overset{\text{approx.}}{\sim} N(0, 1).$$

There are examples in Casella and Robert (2010) along with `R` code for those that haven't seen these methods before or want to review them.

6.1.2 Importance Sampling

Importance sampling involves generating random variables from a different distribution and then reweighing the output. It's name is given since the new distribution is chosen to give greater mass to regions where h is large (the important part of the space).

Let g be an arbitrary density function and then we can write

$$I = E_f[h(x)] = \int_X h(x)\frac{f(x)}{g(x)} g(x)\,dx = E_g\left[\frac{h(x)f(x)}{g(x)}\right]. \tag{6.1}$$

This is estimated by

$$\hat{I} = \frac{1}{n}\sum_{j=1}^{n} \frac{f(X_j)}{g(X_j)} h(X_j) \longrightarrow E_f[h(X)] \tag{6.2}$$

based on a sample generated from g (not f). Since (6.1) can be written as an expectation under g, (6.2) converges to (6.1) for the same reason the Monte carlo estimator $\bar{h}_n$ converges.

Remark: Calculating the variance of $\hat{I}$, we find

$$Var(\hat{I}) = \frac{1}{n^2}\sum_i Var\left(\frac{h(X_i)f(X_i)}{g(X_i)}\right) = \frac{1}{n}Var\left(\frac{h(X_i)f(X_i)}{g(X_i)}\right) \implies$$
$$\widehat{Var}(\hat{I}) = \frac{1}{n}\widehat{Var}\left(\frac{h(X_i)f(X_i)}{g(X_i)}\right).$$

Example 6.1: Suppose we want to estimate $P(X > 5)$, where $X \sim N(0,1)$.

Naive method: Generate n iid standard normals and use the proportion $\hat{p}$ that are larger than 5.

Importance sampling: We will sample from a distribution that gives high probability to the "important region" (the set $(5,\infty)$) and then reweight.

Solution: Let ϕ_o and ϕ_θ be the densities of the $N(0,1)$ and $N(\theta,1)$ distributions (θ taken around 5 will work). We have

$$p = \int I(u > 5)\phi_o(u)\,du = \int \left[I(u > 5)\frac{\phi_o(u)}{\phi_\theta(u)}\right]\phi_\theta(u)\,du.$$

In other words, if

$$h(u) = I(u > 5)\frac{\phi_o(u)}{\phi_\theta(u)}$$

then $p = E_{\phi_\theta}[h(X)]$. If $X_1,\ldots,X_n \sim N(\theta,1)$, then an unbiased estimate is $\hat{p} = \frac{1}{n}\sum_i h(X_i)$.

We implement this in R as follows:

```
1 - pnorm(5)                    # gives 2.866516e-07

# Naive method
set.seed(1)
ss <- 100000
x <- rnorm(n=ss)
phat <- sum(x>5)/length(x)
sdphat <- sqrt(phat*(1-phat)/length(x)) # gives 0

# IS method

set.seed(1)
y <- rnorm(n=ss, mean=5)
h <- dnorm(y, mean=0)/dnorm(y, mean=5) * I(y>5)
mean(h)                         # gives 2.865596e-07
sd(h)/sqrt(length(h))           # gives 2.157211e-09
```

Example 6.2: Let $f(x)$ be the pdf of a $N(0,1)$. Assume we want to compute

$$a = \int_{-1}^{1} f(x)dx = \int_{-1}^{1} N(0,1)dx$$

We can use importance sampling to do this calculation. Let $g(X)$ be an arbitrary pdf,

$$a(x) = \int_{-1}^{1} \frac{f(x)}{g(x)} g(x)\, dx.$$

We want to be able to draw $g(x) \sim Y$ easily. But how should we go about choosing $g(x)$?

- Note that if $g \sim Y$, then $a = E[I_{[-1,1]}(Y)\frac{f(Y)}{g(Y)}]$.
- The variance of $I_{[-1,1]}(Y)\dfrac{f(Y)}{g(Y)}$ is minimized picking $g \propto I_{[-1,1]}(x)f(x)$. Nevertheless simulating from this g is usually expensive.
- Some g's which are easy to simulate from are the pdf's of the Uniform$(-1,1)$, the Normal$(0,1)$ and a Cauchy with location parameter 0.
- Below, there is code of how to get a sample from $I_{[-1,1]}(Y)\dfrac{f(Y)}{g(Y)}$ for these distributions,

```
uniformIS <- function(nn) {
  sapply(runif(nn,-1,1),
    function(xx) dnorm(xx,0,1)/dunif(xx,-1,1)) }

cauchyIS <- function(nn) {
  sapply(rt(nn,1),
    function(xx) (xx <= 1)*(xx >= -1)*dnorm(xx,0,1)/dt(xx,2)) }

gaussianIS <- function(nn) {
  sapply(rnorm(nn,0,1),
    function(xx) (xx <= 1)*(xx >= -1)) }
```

Figure 6.1 presents histograms for a sample size 1000 from each of these distributions. The sample variance of $I_{[-1,1]}(Y)\frac{f(Y)}{g(Y)}$ was, respectively, $red0.009$, $red0.349$ and $red0.227$ (for the uniform, cauchy, and the normal).

- Even though the shape of the uniform distribution is very different from $f(x)$, a standard normal, in $(-1,1)$, $f(x)$ has a lot of mass outside of $(-1,1)$.
- This is why the histograms for the Cauchy and the Normal have big bars on 0 and the variance obtained from the uniform distribution is the lowest.
- How would these results change if we wanted to compute the integral over the range $(-3,3)$ instead of $(-1,1)$? This is left as a homework exercise.

Figure 6.1: Histograms for samples from $I_{[-1,1]}(Y)\frac{f(Y)}{g(Y)}$ when g is, respectively, a uniform, a Cauchy and a Normal pdf.

6.1.3 Importance Sampling with unknown normalizing constant

Often we have sample from μ, but know $\pi(x)$ except for a multiplicative $\mu(x)$ constant. Typical example is Bayesian situation:

- $\pi = \nu_Y =$ posterior density of θ given Y when prior density is ν.
- $\mu = \lambda_Y =$ posterior density of θ given Y when prior density is λ.

  We want to estimate $\dfrac{\pi(x)}{\mu(x)} = \dfrac{c_\nu L(\theta)\nu(\theta)}{c_\lambda L(\theta)\lambda(\theta)} = c\dfrac{\nu(\theta)}{\lambda(\theta)} = c\,\ell(x)$,

  where $\ell(x)$ is known and c is unknown.

  Remark: get a ratio of priors.

  Then if we're estimating $h(x)$, we find

$$\begin{aligned}\int h(x)\pi(x)\,dx &= \int h(x)\;c\;\ell(x)\mu(x)\;d(x)\\ &= \frac{\int h(x)\;c\;\ell(x)\mu(x)\;d(x)}{\int \mu(x)\;d(x)}\\ &= \frac{\int h(x)\;c\;\ell(x)\mu(x)\;d(x)}{\int c\;\ell(x)\mu(x)\;d(x)}\\ &= \frac{\int h(x)\;\ell(x)\mu(x)\;d(x)}{\int \ell(x)\mu(x)\;d(x)}.\end{aligned}$$

  Generate $X_1, \ldots, X_n \sim \mu$ and estimate via

$$\sum_i \frac{h(X_i)\;\ell(X_i)}{\ell(X_i)} = \sum_i h(X_i)\left(\frac{\ell(X_i)}{\sum_j \ell(X_j)}\right) = \sum_i w_i h(X_i)$$

  where $w_i = \dfrac{\ell(X_i)}{\sum_j \ell(X_j)} = \dfrac{\nu(\theta_i)/\lambda(\theta_i)}{\sum_j \nu(\theta_j)/\lambda(\theta_j)}$.

  Motivation
  Why the choice above for $\ell(X)$? Just taking a ratio of priors. The motivation is the following for example:

  – Suppose our application is to Bayesian statistics where $\theta_1, \ldots, \theta_n \sim \lambda_Y$.

- Think about the posterior corresponding here is an essay to deal with conjugate prior λ.
- Think of $\pi = \nu$ as a complicated prior and $\mu = \lambda$ as a conjugate prior.
- Then the weights are $w_i = \dfrac{\nu(\theta_i)/\lambda(\theta_i)}{\sum_j \nu(\theta_j)/\lambda(\theta_j)}$.

1. If μ and π i.e. ν and λ differ greatly most of the weight will be taken up by a few observations resulting in an unstable estimate.
2. We can get an estimate of the variance of $\sum_i \dfrac{h(X_i)\ \ell(X_i)}{\ell(X_i)}$ but we need to use theorems from advance probability theory (The Cramer-Wold device and the Multivariate Delta Method). We'll skip these details.
3. In the application of Bayesian statistics, the cancellation of a potentially very complicated likelihood can lead to a great simplification.
4. The original purpose of importance sampling was to sample more heavily from regions that are important. So, we may do importance sampling using a density μ because it's more convenient than using a density π. (These could also be measures if the densities don't exist for those taking measure theory).

6.1.4 Rejection Sampling

Suppose π is a density on the reals and suppose $\pi(x) = c\, l(x)$ where l is known, c is not known. We are interested in case where π is complicated. Want to generate $X \sim \pi$.

Motivating idea: look at a very simple case of rejection sampling.

Suppose first that l is bounded and is zero outside of $[0, 1]$. Suppose also l is constant on the intervals $((j-1)/k, j/k), j = 1, \ldots, k$. Let M be such that $M \geq l(x)$ for all x.

For very simple case, consider the following procedure.

1. Generate a point (U_1, U_2) uniformly at random from the rectangle of height M sitting on top of the interval $[0, 1]$.
2. If the point is below the graph of the function l, retain U_1. Else, reject the point and go back to (1).

Remark: Using the Probability Integral Transformation in reverse. If $X \sim F^{-1}(U)$, then $X \sim F$ where $U \sim \text{Uniform}(0, 1)$.

Remark: Think about what this is doing, we're generating many draws that are wasting time. Think about the restriction on $[0, 1]$ and if this makes sense.

General Case:

Suppose the density g is such that for some known constant M, $Mg(x) \geq l(x)$ for all x. Procedure:

1. Generate $X \sim g$, and calculate $r(X) = \dfrac{l(X)}{M\ g(X)}$.
2. Flip a coin with probability of success $r(X)$. If we have a success, retain X. Else return to (1).

To show that an accepted point has distribution π, let $I =$ indicator that the point is accepted. Then

$$P(I = 1) = \int P(I = 1 \mid X = x) g(x)\ dx = \int \frac{\pi(x)/c}{M\ g(x)} g(x)\ dx = \frac{1}{c\ M}.$$

Thus, if g_l is the conditional distribution of X given I, we have

$$g_I(x|I = 1) = g(x) \frac{\pi(x)/c}{M\ g(x)} / P(I = 1) = \pi(x).$$

Example 6.3: Suppose we want to generate random variables from the Beta(5.5,5.5) distribution. Note: There are no direct methods for generating from Beta(a,b) if a,b are not integers.

One possibility is to use a Uniform(0,1) as the trial distribution. A better idea is to use an approximating normal distribution.

```
##simple rejection sampler for Beta(5.5,5.5), 3.26.13

a <- 5.5; b <- 5.5
m <- a/(a+b); s <- sqrt((a/(a+b))*(b/(a+b))/(a+b+1))
funct1 <- function(x) {dnorm(x, mean=m, sd=s)}
funct2 <- function(x) {dbeta(x, shape1=a, shape2=b)}

##plotting normal and beta densities
pdf(file = "beta1.pdf", height = 4.5, width = 5)
plot(funct1, from=0, to=1, col="blue", ylab="")
plot(funct2, from=0, to=1, col="red", add=T)
dev.off()

##M=1.3 (this is trial and error to get a good M)
funct1 <- function(x) {1.3*dnorm(x, mean=m, sd=s)}
funct2 <- function(x) {dbeta(x, shape1=a, shape2=b)}
pdf(file = "beta2.pdf", height = 4.5, width = 5)
plot(funct1, from=0, to=1, col="blue", ylab="")
plot(funct2, from=0, to=1, col="red", add=T)
dev.off()

##Doing accept-reject
##substance of code
set.seed(1); nsim <- 1e5
x <- rnorm(n=nsim, mean=m, sd=s)
```

```
u <- runif(n=nsim)
ratio <- dbeta(x, shape1=a, shape2=b) /
           (1.3*dnorm(x, mean=m, sd=s))
ind <- I(u < ratio)
betas <- x[ind==1]
# as a check to make sure we have enough
length(betas) # gives 76836

funct2 <- function(x) {dbeta(x, shape1=a, shape2=b)}
pdf(file = "beta3.pdf", height = 4.5, width = 5)
plot(density(betas))
plot(funct2, from=0, to=1, col="red", lty=2, add=T)
dev.off()
```

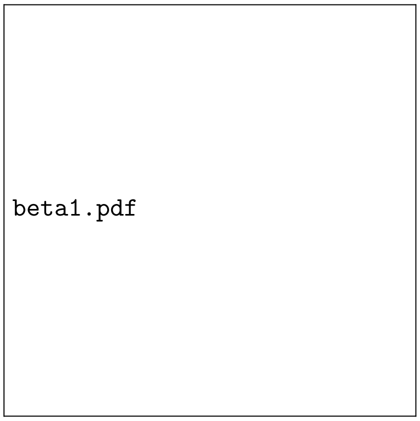


Figure 6.2: Normal enveloping Beta

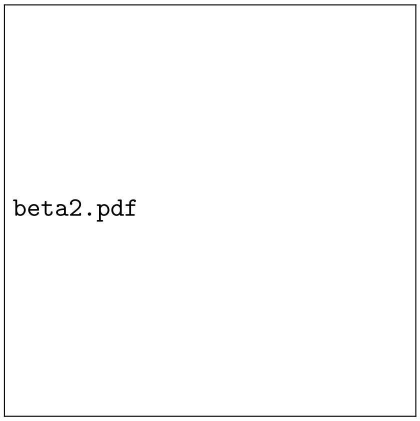


Figure 6.3: Naive rejection sampling, M=1.3

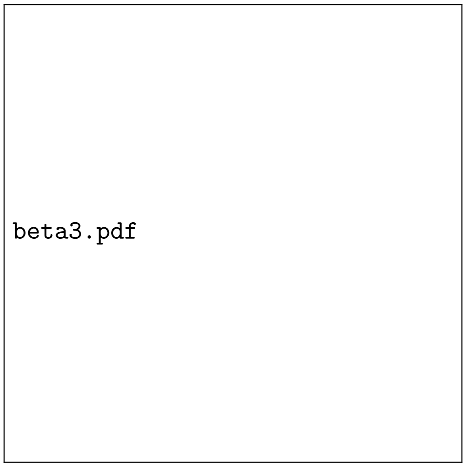


Figure 6.4: Rejection sampler

6.2 Introduction to Gibbs and MCMC

The main idea here involves iterative simulation. We sample values on a random variable from a sequence of distributions that converge as iterations continue to a target distribution. The simulated values are generated by a Markov chain whose stationary distribution is the target distribution, i.e., the posterior distribution.

Geman and Geman (1994) introduced Gibbs sampling for simulating a multivariate probability distribution $p(x)$ using as random walk on a vector x, where $p(x)$ is not necessarily a posterior density.

6.2.1 Markov Chains and Gibbs Samplers

We have a probability distribution π on some space X and we are interested in estimating π or $\int h(x)\pi(x)dx$, where h is some function. We are considering situation where π is analytically intractable.

The Basic idea of MCMC

- Construct a sequence of random variables $X_1, X_2, \ldots$ with the property that the distribution of X_n converges to π as $n \to \infty$.
- If n_o is large, then $X_{n_o}, X_{n0+1} \ldots$ all have the distribution π and these can be used to estimate π and $\int h(x)\pi(x)dx$.

Two problems:

1. The distribution of $X_{n_o}, X_{n0+1} \ldots$ is only approximately π.
2. The random variables $X_{n_o}, X_{n0+1} \ldots$ are NOT independent; they may be correlated.

The MCMC Method Setup: We have a probability distribution π which is analytically intractable. Want to estimate π or $\int h(x)\pi(x)dx$, where h is some function.

The MCMC method consider of coming up with a transition probability function $P(x, A)$ with the property that it has station distribution π.

A Markov chain with Markov transition function $P(\cdot, \cdot)$ is a sequence of random variables $X_1, X_2, \ldots$ on a measurable space such that:

1. $P(X_{n+1} \in A | X_n = x) = P(x, A)$.
2. $P(X_{n+1} \in A | X_1, X_2, \ldots, X_n) = P(X_{n+1} \in A | X_n = x)$.

1.) is called a Markov transition function and 2.) is the Markov property, which says "where I'm going next only depends on where I am right now."

Coming back to the MCMC method, we fix a starting point x_o and generate an observation from X_1 from $P(x_o, \cdot)$, generate an observation from X_2 from $P(X_1, \cdot)$, etc. This generates the Markov chain $x_o = X_o, X_1, X_2, \ldots,$

If we can show that

$$\sup_{C \in B} |P^n(x, C) - \pi(C)| \to 0 \text{ for all } x \in X$$

then by running the chain sufficiently long enough, we succeed in generating an observation X_n with distribution approximately π.

What is a Markov chain?

Start with a sequence of dependent random variables, $\{X^{(t)}\}$. That is we have the sequence

$$X^{(0)}, X^{(1)}, \dots, X^{(t)}, \dots$$

such that the probability distribution of $X^{(t)}$ given all the past variables only depends on the very last one $X^{(t-1)}$. This conditional probability is called the transition kernel or Markov kernel K, i.e.,

$$X^{(t+1)} | X^{(0)}, X^{(1)}, \dots, X^{(t)} \sim K(X^{(t)}, X^{(t+1)}).$$

- For a given Markov kernel K, there may exist a distribution f such that

$$\int_X K(x,y) f(x)\, dx = f(y).$$

- If f satisfies this equation, we call f a stationary distribution of K. What this means is that if $X^{(t)} \sim f$, then $X^{(t+1)} \sim f$ as well.

The theory of Markov chains provides various results about the existence and uniqueness of stationary distributions, but such results are beyond the scope of this course. However, one specific result is that under fairly general conditions that are typically satisfied in practice, if a stationary distribution f exists, then f is the limiting distribution of $\{X^{(t)}\}$ is f for almost any initial value or distribution of $X^{(0)}$. This property is called redergodicity. From a simulation point of view, it means that if a given kernel K produces an ergodic Markov chain with stationary distribution f, generating a chain from this kernel will eventually produce simulations that are redapproximately from f.

In particular, a very important result can be derived. For integrable functions h, the standard average

$$\frac{1}{M} \sum_{i=1}^{M} h(X^{(t)}) \longrightarrow E_f[h(X)].$$

This means that the LLN lies at the basis of Monte Carlo methods which can be applied in MCMC settings. The result shown above is called the Ergodic Theorem.

Of course, even in applied settings, it should always be confirmed that the Markov chain in question behaves as desired before blindly using MCMC to perform Bayesian calculations. Again, such theoretical verifications are beyond the scope of this course. Practically speaking, however, the MCMC methods we will discuss do indeed behave nicely in an extremely wide variety on problems.

Now we turn to Gibbs. The name Gibbs sampling comes from a paper by Geman and Geman (1984), which first applied a Gibbs sampler on a Gibbs random field. The name stuck from there. It's actually a special case of something from Markov chain Monte Carlo (MCMC), and more specifically a method called Metropolis-Hastings, which we will hopefully get to. We'll start by studying the simple case of the two-stage sampler and then look at the multi-stage sampler.

6.2.2 The Two-Stage Gibbs Sampler

The two-stage Gibbs sampler creates a Markov chain from a joint distribution. Suppose we have two random variables X and Y with joint density $f(x, y)$. They also have respective conditional densities $f_{Y|X}$ and $f_{X|Y}$. The two-stage sampler generates a Markov chain $\{(X_t, Y_t)\}$ according to the following steps:

Two-stage Gibbs Sampler
Take $X_0 = x_0$. Then for $t = 1, 2, \ldots,$ generate

1. $X_t \sim f_{X|Y}(\cdot|y_{t-1})$
2. $Y_t \sim f_{Y|X}(\cdot|x_t)$.

As long as we can write down both conditionals (and simulate from them), it is easy to implement the algorithm above.

Example 6.4: Bivariate Normal
Consider the bivariate normal model

$$(X, Y) \sim N_2\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

Recall the following fact from Casella and Berger (2009): If

$$(X, Y) \sim N_2\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right),$$

then

$$Y|X = x \sim N\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X), \sigma_Y^2(1 - \rho^2)\right).$$

Suppose we calculate the Gibbs sampler just given the starting point (x_0, y_0). Since this is a toy example, let's suppose we only care about X. Note that we don't really need both components of the starting point, since if we pick x_0, we can generate Y_0 from $f_{Y|X}(\cdot|x_0)$.

We know that $Y_0 \sim N(\rho x_0, 1 - \rho^2)$ and $X_1|Y_0 = y_0 \sim N(\rho y_0, 1 - \rho^2)$. Then

$$E[X_1] = E[E[X_1|Y_0]] = \rho x_0$$

and

$$\mathbf{v}[X_1] = E\mathbf{v}[X_1|Y_0]] + \mathbf{v}E[X_1|Y_0]] = 1 - \rho^4.$$

Then

$$X_1 \sim N(\rho^2 x_0, 1 - \rho^4).$$

We want the unconditional distribution of X_2 eventually. So, we need to update (X_2, Y_2). So we need Y_1 so we can generate $Y_1|X_1 = x_1$. Since we only care about X, we can use the conditional distribution formula to find that $Y_1|X_1 = x_1 \sim N(\rho x_1, 1 - \rho)$. Then using iterated expectation and iterated variance, we can show that

$$X_2 \sim N(\rho^4 x_o, 1 - \rho^8).$$

If we keep iterating, we find that

$$X_n \sim N(\rho^{2n} x_o, 1 - \rho^{4n}).$$

(To see this, iterate a few times and find the pattern.) What happens as $n \to \infty$?

$$X_n \overset{\text{approx.}}{\sim} N(0, 1).$$

Example 6.5: Binomial-Beta
Suppose $X|\theta \sim \text{Bin}(n, \theta)$ and $\theta \sim \text{Beta}(a, b)$. Then the joint distribution is

$$f(x, \theta) = \binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{x+a-1}(1-\theta)^{n-x+b-1}.$$

The distribution of $X|\theta$ is given above, and $\theta|X \sim \text{Beta}(x + a, n - x + b)$.

We can implement the Gibbs sampler in R as

```
gibbs_beta_bin <- function(nsim, nn, aa, bb)
{
  xx <- rep(NA,nsim)
  tt <- rep(NA,nsim)

  tt[1] <- rbeta(1,aa,bb)
  xx[1] <- rbinom(1,nn,tt[1])

  for(ii in 2:nsim)
  {
    tt[ii] <- rbeta(1,aa+xx[ii-1],bb+nn-xx[ii-1])
    xx[ii] <- rbinom(1,nn,tt[ii])
  }

  return(list(beta_bin=xx,beta=tt))
}
```

Since X has a discrete distribution, we can use a rootogram to check if the Gibbs sampler performed a good approximation. The rootogram plot is implemented in the library `vcd` in `R`. The following are the commands to generate this rootogram:

```
gibbs_sample <- gibbs_beta_bin(5000,15,3,7)

# Density of a beta-binomial distribution with parameters
# nn: sample size of the binomial
# aa: first parameter of the beta
# bb: second parameter of the beta
dbetabi <- function(xx, nn, aa, bb)
{
  return(choose(nn,xx)*exp(lgamma(aa+xx)-lgamma(aa)+lgamma(nn-xx+bb)-
```

Figure 6.5: Rootogram from a Beta-Binomial(15,3,7)

Figure 6.6: Histogram for a Beta(3,7)

```
                             lgamma(bb)-lgamma(nn+aa+bb)+lgamma(aa+bb)))
}

#Rootogram for the marginal distribution of X.
library(vcd)
beta_bin_sample <- gibbs_sample$beta_bin
max_observed <- max(beta_bin_sample)
rootogram(table(beta_bin_sample),5000*dbetabi(0:max_observed,15,3,7),
  scale="raw",xlab="X",main="Rootogram for Beta Binomial sample")
```

Figure 6.5 presents the rootogram for the Gibbs sample for the Beta-Binomial distribution. Similarly, Figure 6.6 shows the same for the marginal distribution of θ obtained through the following commands:

```
#Histogram for the marginal distribution of Theta.
beta_sample <- gibbs_sample$beta
hist(beta_sample,probability=TRUE,xlab=expression(theta),
  ylab="Marginal Density", main="Histogram for Beta sample")
curve(dbeta(x,3,7),from=0,to=1,add=TRUE)
```

Example 6.6: Consider the posterior on (θ, σ^2) associated with the following model:

$$\begin{aligned} X_i|\theta &\sim N(\theta, \sigma^2),\ i = 1, \ldots, n, \\ \theta &\sim N(\theta_o, \tau^2) \\ \sigma^2 &\sim \text{InverseGamma}(a, b), \end{aligned}$$

where θ_o, τ^2, a, b known. Recall that $p(\sigma^2) = \dfrac{b^a}{\Gamma(a)} \dfrac{e^{-b/x}}{x^{a+1}}$.

The Gibbs sampler for these conditional distributions can be coded in R as follows:

```
# gibbs_gaussian: Gibbs sampler for marginal of theta|X=xx and sigma2|X=xx
# when Theta ~ Normal(theta0,tau2) and Sigma2 ~ Inv-Gamma(aa,bb) and
# X|Theta=tt,Sigma2=ss ~ Normal(tt,ss)
#
# returns a list gibbs_sample
# gibbs_sample$theta : sample from the marginal distribution of Theta|X=xx
# gibbs_sample$sigma2: sample from the marginal distribution of Sigma2|X=xx
```

Figure 6.7: Histograms for posterior mean and standard deviation.

```
gibbs_gaussian <- function(nsim,xx,theta0,tau2,aa,bb)
{
  nn <- length(xx)
  xbar <- mean(xx)
  RSS <- sum((xx-xbar)^2)
  post_sigma_shape <- aa + nn/2

  theta <- rep(NA,nsim)
  sigma2 <- rep(NA,nsim)

  sigma2[1] <- 1/rgamma(1,shape=aa,rate=bb)
  ww <- sigma2[1]/(sigma2[1]+nn*tau2)
  theta[1] <- rnorm(1,mean=ww*theta0+(1-ww)*xbar, sd=sqrt(tau2*ww))

  for(ii in 2:nsim)
  {
    new_post_sigma_rate <- (1/2)*(RSS+ nn*(xbar-theta[ii-1])^2) + bb
    sigma2[ii] <- 1/rgamma(1,shape=post_sigma_shape,
      rate=new_post_sigma_rate)

    new_ww <- sigma2[ii]/(sigma2[ii]+nn*tau2)
    theta[ii] <- rnorm(1,mean=new_ww*theta0+(1-new_ww)*xbar,
      sd=sqrt(tau2*new_ww))
  }

  return(list(theta=theta,sigma2=sigma2))
}
```

The histograms in Figure 6.7 for the posterior for θ and σ^2 are obtained as follows:

```
library(mcsm)
data(Energy)
gibbs_sample <- gibbs_gaussian(5000,log(Energy[,1]),5,10,3,3)

par(mfrow=c(1,2))
hist(gibbs_sample$theta,xlab=expression(theta~"|X=x"),main="")
hist(sqrt(gibbs_sample$sigma2),xlab=expression(sigma~"|X=x"),main="")
```

6.2.3 The Multistage Gibbs Sampler

There is a natural extension from the two-stage Gibbs sampler to the general multistage Gibbs sampler. Suppose that for $p > 1$, we can write the random variable $\mathbf{X} = (X_1, \dots, X_p)$, where the X_i's are either unidimensional or multidimensional components. Suppose that we can simulate from corresponding conditional densities $f_1, \dots, f_p$. That is, we can simulate

$$X_i | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_p \sim f(x_i | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_p)$$

for $i = 1, \dots, p$. The associated Gibbs sampling algorithm is given by the following transition from $X^{(t)}$ to $X^{(t+1)}$:

The Multistage Gibbs sampler
At iteration $t = 1, 2, \dots$ given $x^{(t-1)} = (x_1^{(t-1)}, \dots, x_p^{(t-1)})$, generate

1. $X_1^{(t)} \sim f(x_1 | x_2^{(t-1)}, \dots, x_p^{(t-1)})$,

2. $X_2^{(t)} \sim f(x_2 | x_1^{(t)}, x_3^{(t-1)} \dots, x_p^{(t-1)})$,

$\vdots$

p−1. $X_{p-1}^{(t)} \sim f(x_{p-1} | x_1^{(t)}, \dots, x_{p-2}^{(t)}, x_p^{(t-1)})$,

p. $X_p^{(t)} \sim f(x_p | x_1^{(t)}, \dots, x_{p-1}^{(t)})$.

The densities $f_1, \dots, f_p$ are called the full conditionals, and a particular feature of the Gibbs sampler is that these are the only densities used for simulation. Hence, even for high-dimensional problems, all of the simulations may be univariate, which is a major advantage.

Example 6.7: (Casella and Robert, p. 207) Consider the following model:

$$\begin{aligned} X_{ij} | \theta_i, \sigma^2 &\overset{ind}{\sim} N(\theta_i, \sigma^2) \qquad 1 \le i \le k, \;\; 1 \le j \le n_i \\ \theta_i | \mu, \tau^2 &\overset{iid}{\sim} N(\mu, \tau^2) \\ \mu | \sigma_\mu^2 &\sim N(\mu_0, \sigma_\mu^2) \\ \sigma^2 &\sim IG(a_1, b_1) \\ \tau^2 &\sim IG(a_2, b_2) \\ \sigma_\mu^2 &\sim IG(a_3, b_3) \end{aligned}$$

The conditional independencies in this example can be visualized by the Bayesian Network in Figure 6.8. Using these conditional independencies, we can compute

[-¿,¿=stealth',shorten ¿=1pt,auto,node distance=3cm, thick,main node/.style=circle,fill=blue!20,draw,font=]
[main node] (1) σ_μ^2; [main node] (2) [left of=1] μ; [main node] (3) [left of=2] τ^2; [main node] (4) [below of=2] θ_i; [main node] (5) [below right of=2] σ^2; [main node] (6) [below of=4] X_{ij};
`everynode/.style={font=\sffamily\small}](1)edgenode`left] (2) (2) edge node [right] (4) (3) edge node [right] (4) (4) edge node [left] (6) (5) edge node [left] (6);

Figure 6.8: Bayesian Network for Example 6.7.

Figure 6.9: Histograms for posterior quantities.

the complete conditional distributions for each of the variables as

$$\begin{aligned}
\theta_i &\sim N\left(\frac{\sigma^2}{\sigma^2+n_i\tau^2}\mu+\frac{n_i\tau^2}{\sigma^2+n_i\tau^2}\bar{X}_i,\frac{\sigma^2\tau^2}{\sigma^2+n_i\tau^2}\right),\\
\mu &\sim N\left(\frac{\tau^2}{\tau^2+k\sigma_\mu^2}\mu_0+\frac{k\sigma_\mu^2}{\tau^2+k\sigma_\mu^2}\bar{\theta},\frac{\sigma_\mu^2\tau^2}{\tau^2+k\sigma_\mu^2}\right),\\
\sigma^2 &\sim IG\left(\sum_i n_i/2+a_1,(1/2)\sum_{i,j}(X_{i,j}-\theta_i)^2+b_1\right),\\
\tau^2 &\sim IG\left(k/2+a_2,(1/2)\sum_i(\theta_i-\mu)^2+b_2\right),\\
\sigma_\mu^2 &\sim IG\left(1/2+a_3,1/2(\mu-\mu_0)^2+b_3\right),
\end{aligned}$$

where $\bar{\theta}=\sum_i n_i\theta_i/\sum_i n_i$.

Running the chain with $\mu_0=5$ and $a_1=a_2=a_3=b_1=b_2=b_3=3$ and chain size 5000, we get the histograms in Figure 6.9.

6.2.4 Application of the GS to latent variable models

We give an example of Gibbs sampling to an data augmentation example. We look at the example from a genetic linkage analysis. This example is given in Rao (1973, pp. 3689) where it is analyzed in a frequentist setting; it was re-analyzed in Dempster, Laird and Rubin (1977), and re-analyzed in a Bayesian framework in Tanner and Wong (1987).

Example 6.8: A genetic model specifies that 197 animals are distributed multinomially into four categories, with cell probabilities given by

$$\pi=(1/2+\theta/4,(1-\theta)/4,(1-\theta)/4,\theta/4).(*)$$

The actual observations are $y=(125,18,20,34)$. We want to estimate θ.

Biological basis for model:

Suppose we have two factors, call them α and β (say eye color and leg length).

- Each comes at two levels: α comes in levels A and a, and β comes in levels B and b.
- Suppose A is dominant, a is recessive; also B dominant, b recessive.
- Suppose further that $P(A) = 1/2 = P(a)$ [and similarly for the other factor].
- Now suppose that the two factors are related: $P(B|A) = 1 - \eta$ and $P(b|A) = \eta$.
- Similarly, $P(B|a) = \eta$ and $P(b|a) = 1 - \eta$.

To calculate probability of the phenotypes AB, Ab, aB and ab in an offspring (phenotype is what we actually see in the offspring), we suppose that mother and father are chosen independently from the population, and make following table, involving the genotypes (genotype is what is actually in the genes, and this is not seen).

Then

$$P(\text{Father is AB}) = P(B|A)P(B) = \frac{1}{2}(1-\eta).$$

$$P(\text{Mother is AB}) = P(B|A)P(B) = \frac{1}{2}(1-\eta).$$

$$P(\text{O.S. is AB}) = P(B|A)P(B) = \frac{1}{4}(1-\eta)^2.$$

redNote: $\eta = 1/2$ means no linkage and people like to estimate η.

Table 6.1: default

	AB	Ab	aB	ab
AB	$\frac{1}{4}(1-\eta)^2$	$\frac{1}{4}(1-\eta)\eta$	$\frac{1}{4}(1-\eta)\eta$	$\frac{1}{4}(1-\eta)^2$
Ab	$\frac{1}{4}(1-\eta)\eta$			
aB				
ab				

There are 9 cases where we would see the phenotype AB and adding up their probabilities, we get $\dfrac{3-2\eta+\eta^2}{4}$. You can find similar probabilities for the other phenotypes. Writing

$$\frac{(3-2\eta+\eta^2)}{4} = \frac{1}{2} + \frac{1-2\eta+\eta^2}{4}$$

and letting $\theta = (1-\eta)^2$, we find the model specified in (*).

What now?

Suppose we put the prior Beta(a, b) on θ. How do we get the posterior?

Here is one method, using the Gibbs sampler.

Split first cell into two cells, one with probability 1/2, the other with probability $\theta/4$.

Augment the data into a 5-category multinomial, call it X, where X_1 is Bernoulli with parameter 1/2. Now consider $p_{data}(X_1|\theta)$. Will run a Gibbs sampler of length 2:

- The conditional distribution of $X_1 \mid \theta$ (and given the data) is $Bin(125, \frac{1/2}{1/2+\theta/4})$.
- Given the data, conditional on X1 the model is simply a binomial with $n = 197 - X_1$, and probability of success θ, and data consisting of $(125 - X_1 + X_5)$successes, $(X_3 + X_4)$ failures
- Thus, conditional distribution of $\theta \mid X_1$ and the data is

$$\text{Beta}(a + 125 - X_1 - X_5, b + X_3 + X_4).$$

R Code to implement G.S.:

```
set.seed(1)
a <- 1; b <- 1
z <- c(125,18,20,34)
x <- c(z[1]/2, z[1]/2, z[2:4])
nsim <- 50000 # runs in about 2 seconds on 3.8GHz P4
theta <- rep(a/(a+b), nsim)
for (j in 1:nsim)
{
  theta[j] <- rbeta(n=1, shape1=a+125-x[1]+x[5],
                    shape2=b+x[3]+x[4])
  x[1] <- rbinom(n=1, z[1], (2/(2+theta[j])))
}
mean(theta) # gives 0.623
pdf(file="post-dist-theta.pdf",
               horiz=F, height=5.0, width=5.0)
plot(density(theta), xlab=expression(theta), ylab="",
     main=expression(paste("Post Dist of ", theta)))
dev.off()
eta <- 1 - sqrt(theta) # Variable of actual interest
plot(density(eta))
sum(eta > .4)/nsim # gives 0
```

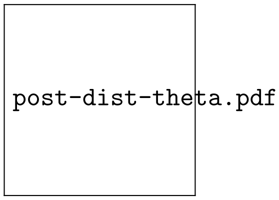


Figure 6.10: **Posterior Distribution of θ for Genetic Linkage**

6.3 MCMC Diagnostics

We will want to check any chain that we run to assess any lack of convergence.

The adequate length of a run will depend on

- a burn-in period (debatable topic).
- mixing rate.
- variance of quantity we are monitoring.

Quick checks:

- trace plots: a times series plot of the parameters of interest; indicates how quickly the chain is mixing of failure to mix.
- Autocorrelations plots.
- Plots of log posterior densities – used mostly in high dimensional problems.
- Multiple starting points – diagnostic to attempt to handle problems when we obtain different estimates when we start with multiple (different) starting values.

Definition: An autocorrelation plot graphically measures the correlation between X_i and each X_{k+i} variable in the chain.

- The Lag-k correlation is the $\text{Corr}(X_i, X_{k+i})$.
- By looking at autocorrelation plots of parameters that we are interested in, we can decide how much to thin or subsample our chain by.
- Then rerun Gibbs sampler using new thin value.

For a real data example that I'm working on:

diag/trace_ucla_bayes2.pdf

Figure 6.11: Trace Plot for RL Example

diag/acf_ucla_bayes2.pdf

Figure 6.12: Max Autocorrelation Plot for RL Example

Multiple Starting Points: Can help determine if burn-in is long enough.

- Basic idea: want to estimate the mean of a parameter θ.
- Run chain 1 starting at x_o. Estimate the mean to be 10 ± 0.1.
- Run chain 2 starting at x_1. Estimate the mean to be 11 ± 0.1.
- Then we know that the effort of the starting point hasn't been forgotten.
- Maybe the chain hasn't reached the area of high probability yet and need to be run for longer?
- Try running multiple chains.

Gelman-Rubin

- Idea is that if we run several chains, the behavior of the chains should be basically the same.
- Check informally using trace plots.
- Check using the Gelman-Rubin diagnostic – but can fail like any test.
- Suggestions – Geweke – more robust when normality fails.

6.4 Theory and Application Based Example

6.4.1 PlA2 Example

Twelve studies run to investigate potential link between presence of a certain genetic trait and risk of heart attack. Each was case-control, and considered a group of individuals with coronary heart disease and another group with no history of heart disease. For each study i (i = 1, . . . , 12) the proportion having the genetic trait in each group was noted and a log odds ratio $\hat{\psi}_i$ was calculated, together with a standard error σ_i. Results are summarized in table below (data from Burr et al. 2003).

i	1	2	3	4	5	6
$\hat{\psi}_i$	1.06	-0.10	0.62	0.02	1.07	-0.02
σ_i	0.37	0.11	0.22	0.11	0.12	0.12

i	7	8	9	10	11	12
$\hat{\psi}_i$	-0.12	-0.38	0.51	0.00	0.38	0.40
σ_i	0.22	0.23	0.18	0.32	0.20	0.25

Setup:

- Twelve studies were run to investigate the potential link between presence of a certain genetic trait and risk of heart attack.
- Each study was case-control and considered a group of individuals with coronary heart disease and another group with no history of coronary heart disease.
- For each study i ($i = 1, \cdots, 12$) the proportion having the genetic trait in each group was recorded.
- For each study, a log odds ratio, $\hat{\psi}_i$, and standard error, σ_i, were calculated.

Let ψ_i represent the true log odds ratio for study i. Then a typical hierarchical model would look like:

$$\begin{aligned} \hat{\psi}_i \mid \psi_i &\overset{ind}{\sim} N(\psi_i, \sigma_i^2) \; i = 1, \ldots, 12 \\ \psi_i \mid \mu, \tau &\overset{iid}{\sim} N(\mu, \tau^2) \; i = 1, \ldots, 12 \\ (\mu, \tau) &\sim \nu. \end{aligned}$$

From this, the likelihood is

$$L(\mu, \tau) = \int \ldots \int \prod_{i=1}^{12} N_{\psi_i, \sigma_i}(\hat{\psi}_i) \prod_{i=1}^{12} N_{\mu, \tau}(\psi_i) \; d\psi_1 \ldots d\psi_{12}.$$

The posterior can be written as (as long as μ and τ have densities), as

$$\pi(\mu,\tau \mid \hat{\psi}_i) = c^{-1}\left[\int \cdots \int \prod_{i=1}^{12} N_{\psi_i,\sigma_i}(\hat{\psi}_i) \prod_{i=1}^{12} N_{\mu,\tau}(\psi_i)\, d\psi_1 \ldots d\psi_{12}\right] p(\mu,\tau).$$

Supppose we take $\nu =$ "Normal/Inverse Gamme prior." Then conditional on $\tau, \mu \sim N(c, d\tau^2)$, $\gamma = 1/\tau^2 \sim Gamma(a,b)$.

Remark: The reason for taking this prior is that it is conjugate for the normal distribution with both mean and variance unknown (that is, it is conjugate for the model in which the ψ_i's are observed).

We will use the notation NIG(a, b, c, d) to denote this prior. Taking a = .1, b = .1, c = 0, and d = 1000 gives a flat prior.

- If we are frequentists, then we need to calculated the likelihood

$$L(\mu,\tau) = \int \cdots \int \prod_{i=1}^{12} N_{\psi_i,\sigma_i}(\hat{\psi}_i) \prod_{i=1}^{12} N_{\mu,\tau}(\psi_i)\, d\psi_1 \ldots d\psi_{12}.$$

- If we are Bayesians, we need to calculate the likelihood and in addition we need to calculate the normalizing constant in order to find the posterior

$$pi(\mu,\tau \mid \hat{\psi}_i) = \frac{L(\mu,\tau)p(\mu,\tau)}{\int L(\mu,\tau)p(\mu,\tau) d\mu d\tau}$$

- Neither above is easy to do.

We have a choice:

- Select a model that doesn't fit the data well but gives answers that are easy to obtain, i.e. in closed form.
- Select a model that is appropriate for the data but i computationally difficult to deal with.

MCMC methods often allow us (in many cases) to make the second choice.

Going back to the example and fitting a model:

Recall the general model:

$$\begin{aligned}\hat{\psi}_i \mid \psi_i &\overset{ind}{\sim} N(\psi_i, \sigma_i^2)\ i = 1, \ldots, 12\\ \psi_i \mid \mu, \tau &\overset{iid}{\sim} N(\mu, \tau^2)\ i = 1, \ldots, 12\\ (\mu,\tau) &\sim NIG(a,b,c,d).\end{aligned}$$

Then the posterior of (μ, τ) is NIG(a',b',c',d'), with

$$a' = a + n/2 \qquad b' = b + \frac{1}{2}\sum_i (X_i - \bar{X})^2 + \frac{n(\bar{X}-c)^2}{2(1+nd)}$$

and

$$c' = \frac{c + nd\bar{X}}{nd+1} \qquad d' = \frac{1}{n + d^{-1}}.$$

This means that

$$\mu \mid \tau^2, y \sim N(c', d')$$

and

$$\tau^2 \mid y \sim \text{InverseGamma}(a', b')$$

Implementing the Gibbs sampler:
Want the posterior distribution of (μ, τ).

- In order to clarify what we are doing, we use the notation that subscripting a distribution by a random variable denotes conditioning.
- Thus, if U and V are two random variables, $L(U|V)$ and $L_V(U)$ will both denote the conditional distribution of U given V.

We want to find $L_{\hat{\psi}}\left(\mu, \tau, \psi\right)$. We'll run a Gibbs sampler of length 2:

- Given (μ, τ), the ψ's are independent. The conditional distribution of ψ given $\hat{\psi}$ is the conditional distribution of ψ given only $\hat{\psi}$. This conditional distribution is given by a standard result for the conjugate normal/normal situation: it is $N(\mu^{'}, \tau^{'2})$, where redStart typing page 87

$$\mu^{'} = \frac{\sigma_i^2 \mu + \tau^2 \hat{\psi}_i}{\sigma_i^2 \ + \tau^2} \qquad \tau^{'2} = \frac{\sigma_i^2 \tau^2}{\sigma_i^2 + \tau^2}$$

- Given the (ψ's, the data) are superfluous, i.e. $L_{\hat{\psi}}\left(\mu, \tau \mid \psi\right)$= $L\left(\mu, \tau \mid \psi\right)$. This conditional distribution is given by the conjugacy of the Normal / Inverse gamma prior: $L\left(\mu, \tau \mid \psi\right) = \text{NIG}(a', b', c', d')$, where

$$a' = a + n/2 \qquad b' = b + \frac{1}{2}\sum_i (\psi_i \bar{\psi})^2 + \frac{n(\bar{\psi} - c)^2}{2(1 + nd)}$$

and

$$c' = \frac{c + nd\bar{\psi}}{nd+1} \qquad d' = \frac{1}{n + d^{-1}}.$$

This gives us a sequence $\mu, \tau, \psi_1, \ldots, \psi_n; g = 1, ..., G$, from $L_{\hat{\psi}}\left(\mu, \tau, \psi\right)$. If were interested in, e.g., the posterior distribution of μ, we just retain the first coordinate in the sequence.

Specific Example for PlA2 data
Our proposed hierarchical model is

$$\hat{\psi}_i \mid \psi_i \overset{ind}{\sim} N(\psi_i, \sigma_i^2) \quad i = 1, \cdots, 12$$

$$\psi_i \mid \mu, \tau^2 \overset{iid}{\sim} N(\mu, \tau^2) \quad i = 1, \cdots, 12$$

$$\mu \mid \tau^2 \sim N(0, 1000\tau^2)$$

$$\gamma = 1/\tau^2 \sim \text{Gamma}(0.1, 0.1)$$

Why is a normal prior taken? It's conjugate for the normal distribution with the mean and variance known. The two priors above with the chosen hyperparameters result in noninformative hyperpriors.

```
%\frame[containsverbatim]{
%\frametitle{PlA2 Example}

The file model.txt contains

\begin{verbatim}
model {
for (i in 1:N) {
psihat[i] ~ dnorm(psi[i],1/(sigma[i])^2)
psi[i] ~ dnorm(mu,1/tau^2)
}
mu ~ dnorm(0,1/(1000*tau^2))
tau <- 1/sqrt(gam)
gam ~ dgamma(0.1,0.1)
}
```

Note: In BUGS, use dnorm(mean,precision), where precision = 1/variance.

```
%\frame[containsverbatim]{
%\frametitle{PlA2 Example}

The file data.txt contains

\begin{verbatim}
"N" <- 12
"psihat" <-  c(1.055, -0.097, 0.626, 0.017, 1.068,
-0.025, -0.117, -0.381, 0.507, 0, 0.385, 0.405)
"sigma" <- c(0.373, 0.116, 0.229, 0.117, 0.471,
0.120, 0.220, 0.239, 0.186, 0.328, 0.206, 0.254)
```

The file inits_1.txt contains

```
".RNG.name" <- "base::Super-Duper"
".RNG.seed" <- 12
"psi" <- c(0,0,0,0,0,0,0,0,0,0,0,0)
"mu" <- 0
"gam" <- 1
```

```
%\frame[containsverbatim]{
%\frametitle{PlA2 Example}

The file script.txt contains
\small
\begin{verbatim}
model clear
data clear
model in "model"
data in "data"
compile, nchains(2)
inits in "inits1", chain(1)
inits in "inits2", chain(2)
initialize
update 10000
monitor mu
monitor psi
monitor gam
update 100000
coda *, stem(CODA1)
coda *, stem(CODA2)
```

Now, we read in the coda files into R from the current directory and continue our analysis. The first part of our analysis will consist of some diagnostic procedures.

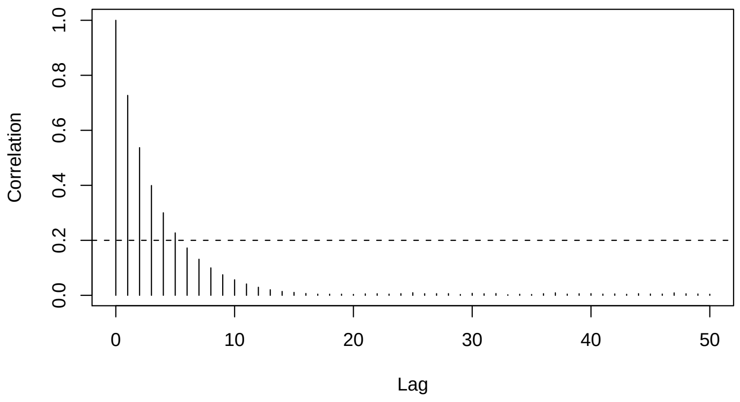
We will consider

- Autocorrelation Plots
- Trace Plots
- Gelman-Rubin Diagnostic
- Geweke Diagnostic

Definition: An autocorrelation plot graphically measures the correlation between X_i and each X_{k+i} variable in the chain.

- The Lag-k correlation is the $\text{Corr}(X_i, X_{k+i})$.
- By looking at autocorrelation plots of parameters that we are interested in, we can decide how much to thin or subsample our chain by.
- We can rerun our JAGS script using our thin value.

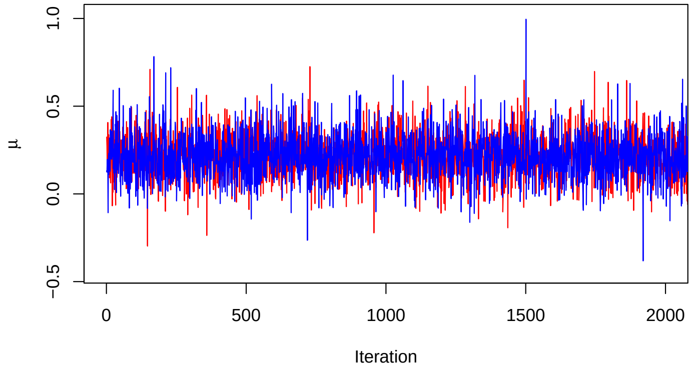
We take the thin value to be the first lag whose correlation ≤ 0.2. For this plot, we take a thin of 2. We will go back and rerun our JAGS script and skip every other value in each chain. After thinning, we will proceed with other diagnostic procedures of interest.



```
%\frame[containsverbatim]{
%\frametitle{PlA2 Example}

The file script\_thin.txt contains
\small
\begin{verbatim}
model clear
data clear
model in "model"
data in "data"
compile, nchains(2)
inits in "inits1", chain(1)
inits in "inits2", chain(2)
initialize
update 10000
monitor mu, thin(6)
monitor psi, thin(6)
monitor gam, thin(6)
update 100000
coda *, stem(CODA1_thin)
coda *, stem(CODA2_thin)
```

Definition: A trace plot is a time series plot of the parameter, say μ, that we monitor as the Markov chain(s) proceed(s).



Definition: The Gelman-Rubin diagnostic tests that burn-in is adequate and requires that multiple starting points be used.

To compute the G-R statistic, we must

- Run two chains in JAGS using two different sets of initial values (and two different seeds).
- Load coda package in R and run gelman.diag(mcmc.list(chain1,chain2)).

How do we interpret the Gelman-Rubin Diagnostic?

- If the chain has reached convergence, the G-R test statistic R ≈ 1. We conclude that burn-in is adequate.
- Values above 1.05 indicate lack of convergence.

Warning: The distribution of R under the null hypothesis is essentially an F distribution. Recall that the F-test for comparing two variances is not robust to violations of normality. Thus, we want to be cautious in using the G-R diagnostic.

```
%\frame[containsverbatim]{
%\frametitle{Gelman-Rubin Diagnostic}

Doing this in R, for the PlA-2 example, we find

\begin{verbatim}
        Point est. 97.5% quantile
mu               1              1
psi[1]           1              1
psi[2]           1              1
```

```
...
psi[11]          1              1
psi[12]          1              1
gam              1              1
```

Since 1 is in all the 95% CI, we can conclude that we have not failed to converge.

Suppose μ is the parameter of interest.

Main Idea: If burn-in is adequate, the mean of the posterior distribution of μ from the first half of the chain should equal the mean from the second half of the chain.

To compute the Geweke statistic, we must

- Run a chain in JAGS along with a set of initial values.
- Load the coda package in R and run geweke.diag(mcmc.list(chain)).
- The Geweke statistic asymptotically has a standard normal distribution, so if the values from R are outside -2.5 or 2.5, this indicates nonstationarity of chain and that burn-in is not sufficient.
- Using the Geweke diagnostic on the PlA2 data indicates that burn-in of 10,000 is sufficient (the largest absolute Z-score is 1.75).
- Observe that the Geweke diagnostic does not require multiple starting points as Gelman-Rubin does.
- The Geweke statistic (based on a T-test) is robust against violations of normality so the Geweke test is preferred to Gelman-Rubin.

Using Gelman-Rubin and Geweke we have shown that burn-in is "sufficient."

- We can look at summary statistics such as means, standard errors, and credible intervals using the summary function.
- We can use kernel density functions in R to estimate posterior distributions that we are interested in using the density function.

The posterior of $\mu \mid$ data is

Alternatively, we can estimate the conditional distributions of $\exp(\psi_i)$'s given the data. A few are shown below.

- So, here we're looking at the odds ratio's of the prob of getting heart disease given you have the genetic trait over the prob of not getting heart disease given you have the trait. Note that all estimates are pulled toward the mean showing a Bayesian Stein effect.

	Post Mean	Post SD	Post Naive SE
μ	0.217272	0.127	0.0009834
ψ_1	0.594141	0.2883	0.0022334
ψ_2	-0.062498	0.1108	0.0008583
ψ_3	0.490872	0.2012	0.0015588
ψ_4	0.040284	0.1118	0.0008658
ψ_5	0.51521	0.3157	0.0024453
ψ_6	0.003678	0.114	0.0008831
ψ_7	-0.015558	0.1883	0.0014586
ψ_8	-0.175852	0.2064	0.0015988
ψ_9	0.433525	0.1689	0.0013084
ψ_{10}	0.101912	0.2423	0.0018769
ψ_{11}	0.332775	0.1803	0.0013965
ψ_{12}	0.331466	0.2107	0.0016318
γ	10.465411	6.6611	0.051596

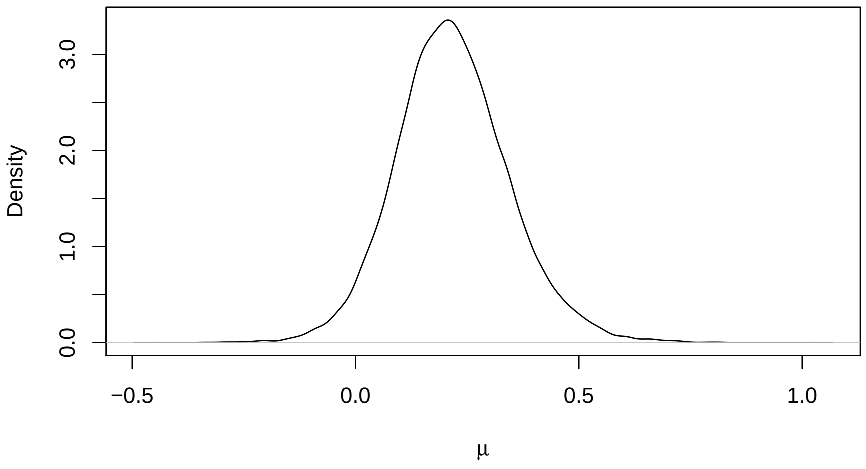
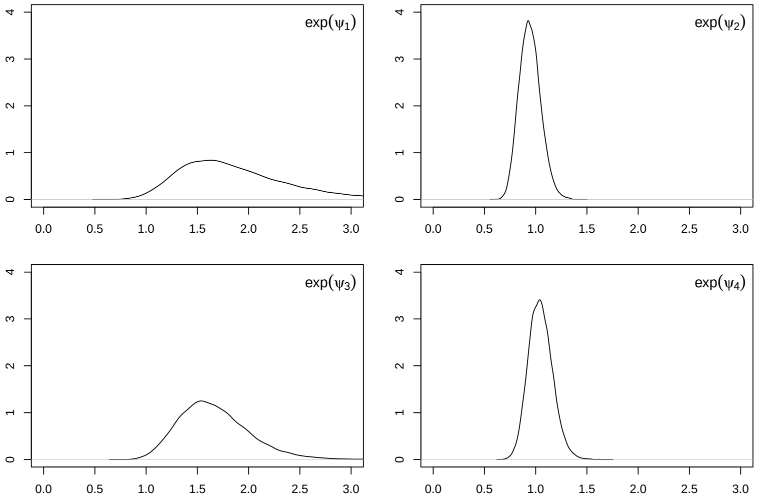


Figure 6.13: Posterior of μ | data

- This is the odds ratio of having a heart attack for those who have the genetic trait versus those who don't (looking at study i).



```
%\frame[containsverbatim]{
Moreover, we could have just have easily done this analysis in WinBUGS. Below is the correspondin
\begin{verbatim}
model{
for (i in 1:N) {
psihat[i] ~ dnorm(psi[i],rho[i])
psi[i] ~ dnorm(mu,gam)
rho[i] <- 1/pow(sigma[i],2)
}

mu ~ dnorm(0,gamt)
gam ~ dgamma(0.1,0.1)
gamt <- gam/1000
}
```

Finally, we can either run the analysis using WinBUGS or JAGS and R. I will demonstrate how to do this using JAGS for this example. I have included the basic code to run this on a Windows machine via WinBUGS. Both methods yield essentially the same results.

To run WinBUGS within `R`, you need the following:

- Load the `R2WinBUGS` library.
- Read in data and format it as a `list()`.
- Format intital values as a `list()`.

- Format the unknown parameters using `c()`.
- Run the `bugs()` command to open/run WinBUGS.
- Read in the G.S. values using `read.coda()`.

```
\scriptsize
\begin{verbatim}
setwd("C:/Documents and Settings/Tina Greenly
/Desktop/beka_winbugs/novartis/pla2")
library(R2WinBUGS)
pla2 <- read.table("pla2_data.txt",header=T)
attach(pla2)
names(pla2)
N<-length(psihat)
data <- list("psihat", "sigma", "N")

inits1 <- list(psi = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0), mu = 0, gam = 1)
inits2 <- list(psi = c(2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2), mu = 1, gam = 2)
inits = list(inits1, inits2)
parameters <- c("mu", "psi", "gam")
pla2.sim <- bugs(data, inits, parameters,
  "pla2.bug", n.chains=2, n.iter = 110000,
  codaPkg=T,debug=T,n.burnin = 10000,n.thin=1,bugs.seed=c(12,13),
working.directory="C:/Documents and Settings/Tina Greenly/
Desktop/beka_winbugs/novartis/pla2")
detach(pla2)
coda1 = read.coda("coda1.txt","codaIndex.txt")
coda2 = read.coda("coda2.txt","codaIndex.txt")
```

6.5 Metropolis and Metropolis-Hastings

The Metropolis-Hastings algorithm is a general term for a family of Markov chain simulation methods that are useful for drawing samples from Bayesian posterior distributions. The Gibbs sampler can be viewed as a special case of Metropolis-Hastings (as well will soon see). Here, we present the basic Metropolis algorithm and its generalization to the Metropolis-Hastings algorithm, which is often useful in applications (and has many extensions).

Suppose we can sample from $p(\theta|y)$. Then we could generate

$$\theta^{(1)}, \ldots, \theta^{(S)} \overset{iid}{\sim} p(\theta|y)$$

and obtain Monte Carlo approximations of posterior quantities

$$E[g(\theta)|y] \to 1/S \sum_{i=1}^{S} g(\theta^{(i)}).$$

But what if we cannot sample directly from $p(\theta|y)$? The important concept here is that we are able to construct a large collection of θ values (rather than them being iid, since this most certain for most realistic situations will not hold). Thus, for any two different θ values θ_a and θ_b, we need

$$\frac{\#\theta's \text{ in the collection } = \theta_a}{\#\theta's \text{ in the collection } = \theta_b} \approx \frac{p(\theta_a|y)}{p(\theta_b|y)}.$$

How might we intuitively construct such a collection?

- Suppose we have a working collection $\{\theta^{(1)}, \ldots, \theta^{(s)}\}$ and we want to add a new value $\theta^{(s+1)}$.
- Consider adding a value θ^* which is nearby $\theta^{(s)}$.
- Should we include θ^* or not?
- If $p(\theta^*|y) > p(\theta^{(s)}|y)$, then we want more θ^*'s in the set than $\theta^{(s)}$'s.
- But if $p(\theta^*|y) < p(\theta^{(s)}|y)$, we shouldn't necessarily include θ^*.

Based on the above, perhaps our decision to include θ^* or not should be based upon a comparison of $p(\theta^*|y)$ and $p(\theta^{(s)}|y)$. We can do this by computing r:

$$r = \frac{p(\theta^*|y)}{p(\theta^{(s)}|y)} = \frac{p(y \mid \theta^*)p(\theta^*)}{p(y \mid \theta^{(s)})p(\theta^{(s)})}.$$

Having computed r, what should we do next?

- If $r > 1$ blue(intuition): Since $\theta^{(s)}$ is already in our set, we should include θ^* as it has a higher probability than $\theta^{(s)}$.

  red(procedure): Accept θ^* into our set and let $\theta^{(s+1)} = \theta^*$.

- If $r < 1$ blue(intuition): The relative frequency of θ-values in our set equal to θ^* compared to those equal to $\theta^{(s)}$ should be

  $$\frac{p(\theta^*|y)}{p(\theta^{(s)}|y)} = r.$$

  This means that for every instance of $\theta^{(s)}$, we should only have a fraction of an instance of a θ^* value.

  red(procedure): Set $\theta^{(s+1)}$ equal to either θ^* or $\theta^{(s)}$ with probability r and $1 - r$ respectively.

This is basic intuition behind the Metropolis (1953) algorithm. More formally, it

- It proceeds by sampling a proposal value θ^* nearby the current value $\theta^{(s)}$ using a *symmetric proposal distribution* $J(\theta^* \mid \theta^{(s)})$.

- What does symmetry mean here? It means that $J(\theta_a \mid \theta_b) = J(\theta_b \mid \theta_a)$. That is, the probability of proposing $\theta^* = \theta_a$ given that $\theta^{(s)} = \theta_b$ is equal to the probability of proposing $\theta^* = \theta_b$ given that $\theta^{(s)} = \theta_a$.

- Symmetric proposals include:

  $$J(\theta^* \mid \theta^{(s)}) = \text{Uniform}(\theta^{(s)} - \delta, \theta^{(s)} + \delta)$$

  and

  $$J(\theta^* \mid \theta^{(s)}) = \text{Normal}(\theta^{(s)}, \delta^2).$$

The Metropolis algorithm proceeds as follows:

1. Sample $\theta^* \sim J(\theta \mid \theta^{(s)})$.

2. Compute the acceptance ratio (r):

   $$r = \frac{p(\theta^*|y)}{p(\theta^{(s)}|y)} = \frac{p(y \mid \theta^*)p(\theta^*)}{p(y \mid \theta^{(s)})p(\theta^{(s)})}.$$

3. Let

   $$\theta^{(s+1)} = \begin{cases} \theta^* & \text{with prob min(r,1)} \\ \theta^{(s)} & \text{otherwise.} \end{cases}$$

Remark: Step 3 can be accomplished by sampling $u \sim \text{Uniform}(0,1)$ and setting $\theta^{(s+1)} = \theta^*$ if $u < r$ and setting $\theta^{(s+1)} = \theta^{(s)}$ otherwise.

Example 6.9: Metropolis for Normal-Normal
Let's test out the Metropolis algorithm for the conjugate Normal-Normal model with a known variance situation.

That is let

$$\begin{aligned} X_1, \ldots, X_n \mid \theta &\overset{iid}{\sim} \text{Normal}(\theta, \sigma^2) \\ \theta &\sim \text{Normal}(\mu, \tau^2). \end{aligned}$$

Recall that the posterior of θ is Normal(μ_n, τ_n^2), where

$$\mu_n = \bar{x}\frac{n/\sigma^2}{n/\sigma^2 + 1/\tau^2} + \mu\frac{1/\tau^2}{n/\sigma^2 + 1/\tau^2}$$

and

$$\tau_n^2 = \frac{1}{n/\sigma^2 + 1/\tau^2}.$$

Suppose (taken from Hoff, 2009), $\sigma^2 = 1, \tau^2 = 10$, $\mu = 5$, $n = 5$, and $y = (9.37, 10.18, 9.16, 11.60, 10.33)$. For these data, $\mu_n = 10.03$ and $\tau_n^2 = 0.20$.

Suppose that for some ridiculous reason we cannot come up with the posterior distribution and instead we need the Metropolis algorithm to approximate it (please note how incredible silly this example is and it's just to illustrate the method).

Based on this model and prior, we need to compute the acceptance ratio r

$$r = \frac{p(\theta^*|x)}{p(\theta^{(s)}|x)} = \frac{p(x|t^*)p(\theta^*)}{p(x|\theta^{(s)})p(\theta^{(s)})} = \left(\frac{\prod_i \text{dnorm}(x_i, \theta^*, \sigma)}{\prod_i \text{dnorm}(x_i, \theta^{(s)}, \sigma)}\right)\left(\frac{\prod_i \text{dnorm}(\theta^*, \mu, \sigma)}{\prod_i \text{dnorm}(\theta^{(s)}, \mu, \sigma)}\right)$$

In many cases, computing the ratio r directly can be numerically unstable, however, this can be modified by taking $\log r$.

This results in

$$\begin{aligned} \log r = &\sum_i \left[\log \text{dnorm}(x_i, \theta^*, \sigma) - \log \text{dnorm}(x_i, \theta^{(s)}, \sigma)\right] \\ &+ \sum_i \left[\log \text{dnorm}(\theta^*, \mu, \sigma) - \log \text{dnorm}(\theta^{(s)}, \mu, \sigma)\right]. \end{aligned}$$

Then a proposal is accepted if $\log u < \log r$, where u is sample from the Uniform(0,1).

The R-code below generates 10,000 iterations of the Metropolis algorithm stating at $\theta^{(0)} = 0.$ and using a normal proposal distribution, where

$$\theta^{(s+1)} \sim \text{Normal}(\theta^{(s)}, 2).$$

Below is R-code for running the above model. Figure 6.14 shows a trace plot for this run as well as a histogram for the Metropolis algorithm compared with a draw from the true normal density. From the trace plot, although the value of θ does not start near the posterior mean of 10.03, it quickly arrives there after just a few iterations. The second plot shows that the empirical distribution of the simulated values is very close to the true posterior distribution.

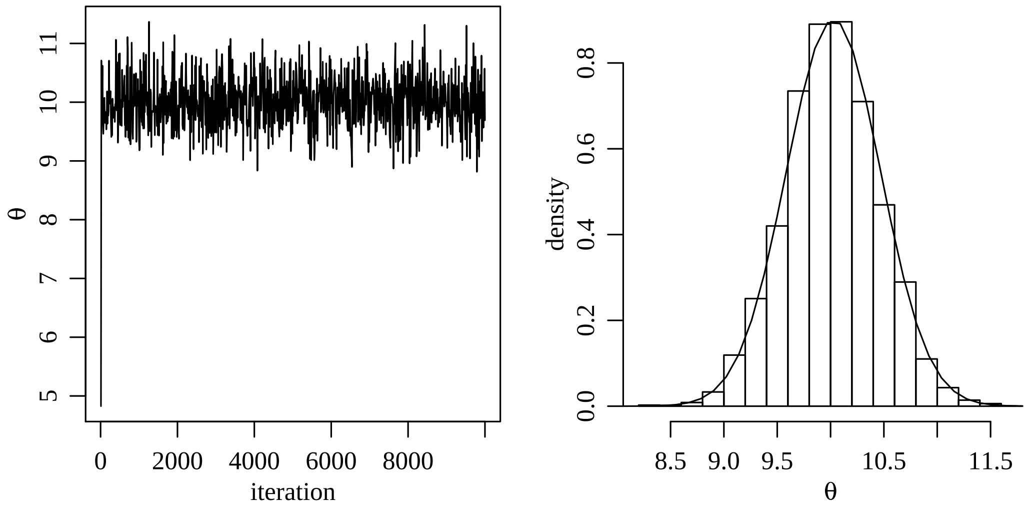


Figure 6.14: Results from the Metropolis sampler for the normal model.

```
## initialing values for normal-normal example and setting seed
# MH algorithm for one-sample normal problem with known variance

s2<-1
t2<-10 ; mu<-5; set.seed(1); n<-5; y<-round(rnorm(n,10,1),2)
mu.n<-( mean(y)*n/s2 + mu/t2 )/( n/s2+1/t2)
t2.n<-1/(n/s2+1/t2)

####metropolis part####
y<-c(9.37, 10.18, 9.16, 11.60, 10.33)
##S = total num of simulations
theta<-0 ; delta<-2 ; S<-10000 ; THETA<-NULL ; set.seed(1)

for(s in 1:S)
{

## simulating our proposal
  theta.star<-rnorm(1,theta,sqrt(delta))

##taking the log of the ratio r
  log.r<-( sum(dnorm(y,theta.star,sqrt(s2),log=TRUE)) +
             dnorm(theta.star,mu,sqrt(t2),log=TRUE) )  -
         ( sum(dnorm(y,theta,sqrt(s2),log=TRUE)) +
             dnorm(theta,mu,sqrt(t2),log=TRUE) )
  if(log(runif(1))<log.r) { theta<-theta.star }

##updating THETA

  THETA<-c(THETA,theta)

}

##two plots: trace of theta and comparing the empirical distribution
##of simulated values to the true posterior

pdf("metropolis_normal.pdf",family="Times",height=3.5,width=7)
par(mar=c(3,3,1,1),mgp=c(1.75,.75,0))
par(mfrow=c(1,2))

skeep<-seq(10,S,by=10)
plot(skeep,THETA[skeep],type="l",xlab="iteration",ylab=expression(theta))

hist(THETA[-(1:50)],prob=TRUE,main="",xlab=expression(theta),ylab="density")
th<-seq(min(THETA),max(THETA),length=100)
lines(th,dnorm(th,mu.n,sqrt(t2.n)) )
dev.off()
```

6.5.1 Metropolis-Hastings Algorithm

Recall that a Markov chain is a sequentially generated sequence $\{x^{(1)}, , x^{(2)}, \ldots\}$ such that the mechanism that generates $x^{(s+1)}$ can depend on the value of $x^{(s)}$ but not on anything that was in the sequence before it. A better way of putting this: for a Markov chain, the future depends on the present and not on the past.

The Gibbs sampler and the Metropolis algorithm are both ways of generating Markov chains that approximate a target probability distribution.

We first consider a simple example where our target probability distribution is $p_o(u, v)$, a bivariate distribution for two random variables U and V. In the one-sample normal problem, we would have $U = \theta$, $V = \sigma^2$ and $p_o(u, v) = p(\theta, \sigma^2|y)$.

What does the Gibbs sampler have us do? It has us iteratively sample values of U and V from their conditional distributions. That is,

1. update U : sample $u^{(s+1)} \sim p_o(u \mid v^{(s)})$
2. update V : sample $v^{(s+1)} \sim p_o(v \mid u^{(s+1)})$.

In contrast, Metropolis proposes changes to $X = (U, V)$ and then accepts or rejects those changes based on p_o. An alternative way to implement the Metropolis algorithm if to propose and then accept or reject change to one element at a time:

1. update U :

   (a) sample $u^* \sim J_u(u \mid u^{(s)})$
   (b) compute $r = \dfrac{p_o(u^*, v^{(s)})}{p_o(u^{(s)}, v^{(s)})}$
   (c) set $u^{(s+1)}$ equal to u^* or $u^{(s+1)}$ with prob min(1,r) and max(0,1-r).

2. update V : sample $v^{(s+1)} \sim p_o(v \mid u^{(s+1)})$.
   (a) sample $v^* \sim J_u(v \mid v^{(s)})$
   (b) compute $r = \dfrac{p_o(u^{(s+1)}, v^*)}{p_o(u^{(s+1)}, v^{(s)})}$
   (c) set $v^{(s+1)}$ equal to v^* or $v^{(s)}$ with prob min(1,r) and max(0,1-r).

Here, J_u and J_v are separate symmetric proposal distributions for U and V.

- The Metropolis algorithm generates proposals from J_u and J_v
- It accepts them with some probability min(1,r).
- Similarly, each step of Gibbs can be seen as generating a proposal from a full conditional and then accepting it with probability 1.
- The Metropolis-Hastings (MH) algorithm generalizes both of these approaches by allowing arbitrary proposal distributions.
- The proposal distributions can be symmetric around the current values, full conditionals, or something else entirely. A MH algorithm for approximating $p_o(u, v)$ runs as follows:

1. update U :
   (a) sample $u^* \sim J_u(u \mid u^{(s)}, v^{(s)})$
   (b) compute
   $$r = \frac{p_o(u^*, v^{(s)})}{p_o(u^{(s)}, v^{(s)})} \times \frac{J_u(u^{(s)} \mid u^*, v^{(s)})}{J_u(u^* \mid u^{(s)}, v^{(s)})}$$
   (c) set $u^{(s+1)}$ equal to u^* or $u^{(s+1)}$ with prob min(1,r) and max(0,1-r).

2. update V :
   (a) sample $v^* \sim J_v(u \mid u^{(s+1)}, v^{(s)})$
   (b) compute
   $$r = \frac{p_o(u^{(s+1)}, v^*)}{p_o(u^{(s+1)}, v^{(s)})} \times \frac{J_u(v^{(s+1)} \mid u^{(s+1)}, v^*)}{J_u(v^* \mid u^{(s+1)}, v^{(s)})}$$
   (c) set $v^{(s+1)}$ equal to v^* or $v^{(s+1)}$ with prob min(1,r) and max(0,1-r).

In the above algorithm, the proposal distributions J_u and J_v are not required to be symmetric. The only requirement is that they not depend on U or V values in our sequence previous to the most current values. This requirement ensures that the sequence is a Markov chain.

Doesn't the algorithm above look familiar? Yes, it looks a lot like Metropolis, except the acceptance ratio r contains an extra factor:

- It contains the ratio of the prob of generating the redcurrent value from the proposed to the prob of generating the redproposed from the current.
- This can be viewed as a correction factor.
- If a value u^* is much more likely to be proposed than the current value $u^{(s)}$ then we must reddown-weight the probability of accepting u.
- Otherwise, such a value u^* will be overrepresented in the chain.

Exercise 1: Show that Metropolis is a special case of MH. Hint: Think about the jumps J.

Exercise 2: Show that Gibbs is a special case of MH. Hint: Show that r = 1.

Note: The MH algorithm can easily be generalized.

Example 6.10: Poisson Regression We implement the Metropolis algorithm for a Poisson regression model.

- We have a sample from a population of 52 song sparrows that was studied over the course of a summer and their reproductive activities were recorded.
- In particular, their age and number of new offspring were recorded for each sparrow (Arcese et al., 1992).
- A simple probability model to fit the data would be a Poisson regression where, Y = number of offspring conditional on x = age.

Thus, we assume that

$$Y|\theta_x \sim \text{Poisson}(\theta_x).$$

For stability of the model, we assume that the mean number of offspring θ_x is a smoother function of age. Thus, we express $\theta_x = \beta_1 + \beta_2 x_+ \beta_3 x^2$.

Remark: This parameterization allows some values of θ_x to be negative, so as an alternative we reparameterize and model the log-mean of Y, so that

$$\log E(Y|x) = \log \theta_x = \log(\beta_1 + \beta_2 x_+ \beta_3 x^2)$$

which implies that $\theta_x = \exp(\beta_1 + \beta_2 x_+ \beta_3 x^2) = \exp(\boldsymbol{\beta}^T \boldsymbol{x})$.

Now back to the problem of implementing Metropolis. For this problem, we will write

$$\log E(Y_i|x_i) = \log(\beta_1 + \beta_2 x_i + \beta_3 x_i^2) = \boldsymbol{\beta}^T \boldsymbol{x_i},$$

where x_i is the age of sparrow i. We will abuse notation slightly and write $\boldsymbol{x_i} = (1, x_i, x_i^2)$.

- We will assume the prior on the regression coefficients is iid Normal(0,100).
- Given a current value $\boldsymbol{\beta^{(s)}}$ and a value $\boldsymbol{\beta^*}$ generated from $J(\boldsymbol{\beta^*}, \boldsymbol{\beta^{(s)}})$ the acceptance ration for the Metropolis algorithm is:

$$r = \frac{p(\boldsymbol{\beta^*}|\boldsymbol{X}, \boldsymbol{y})}{p(\boldsymbol{\beta^{(s)}}|\boldsymbol{X}, \boldsymbol{y})} = \frac{\prod_{i=1}^n \text{dpois}(y_i, x_i^T \beta^*)}{\prod_{i=1}^n \text{dpois}(y_i, x_i^T \beta^{(s)})} \times \frac{\prod_{j=1}^3 \text{dnorm}(\beta_j^*, 0, 10)}{\prod_{j=1}^3 \text{dnorm}(\beta_j^{(s)}, 0, 10)}.$$

- We just need to specify the proposal distribution for θ^*
- A convenient choice is a multivariate normal distribution with mean $\boldsymbol{\beta^{(s)}}$.
- In many problems, the posterior variance can be an efficient choice of a proposal variance. But we don't know it here.
- However, it's often sufficient to use a rough approximation. In a normal regression problem, the posterior variance will be close to $\sigma^2(X^T X)^{-1}$ where σ^2 is the variance of Y.

In our problem: $E \log Y = \beta^T x$ so we can try a proposal variance of $\hat{\sigma}^2(X^T X)^{-1}$ where $\hat{\sigma}^2$ is the sample variance of $\log(y + 1/2)$.

Remark: Note we add $1/2$ because otherwise $\log 0$ is undefined. The code of implementing the algorithm is included below.

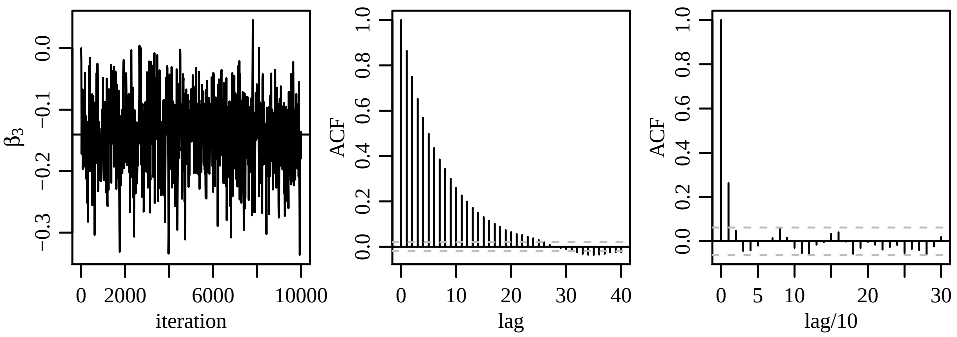


Figure 6.15: Plot of the Markov chain in β_3 along with autocorrelations functions

```
###example 5.10 -- sparrow Poisson regression
yX.sparrow<-dget("http://www.stat.washington.edu/~hoff/Book/Data/data/yX.sparrow")

### sample from the multivariate normal distribution
rmvnorm<-function(n,mu,Sigma)
{
  p<-length(mu)
  res<-matrix(0,nrow=n,ncol=p)
  if( n>0 & p>0 )
  {
    E<-matrix(rnorm(n*p),n,p)
    res<-t(  t(E%*%chol(Sigma)) +c(mu))
  }
  res
}


y<- yX.sparrow[,1]; X<- yX.sparrow[,-1]
n<-length(y) ; p<-dim(X)[2]

pmn.beta<-rep(0,p)
psd.beta<-rep(10,p)

var.prop<- var(log(y+1/2))*solve( t(X)%*%X )
beta<-rep(0,p)
S<-10000
BETA<-matrix(0,nrow=S,ncol=p)
ac<-0
set.seed(1)

for(s in 1:S) {

#propose a new beta

beta.p<- t(rmvnorm(1, beta, var.prop ))

lhr<- sum(dpois(y,exp(X%*%beta.p),log=T)) -
      sum(dpois(y,exp(X%*%beta),log=T)) +
      sum(dnorm(beta.p,pmn.beta,psd.beta,log=T)) -
      sum(dnorm(beta,pmn.beta,psd.beta,log=T))

if( log(runif(1))< lhr ) { beta<-beta.p ; ac<-ac+1 }

BETA[s,]<-beta
                    }
cat(ac/S,"\n")

#######

library(coda)
apply(BETA,2,effectiveSize)
```

```
####
pdf("sparrow_plot1.pdf",family="Times",height=1.75,width=5)
par(mar=c(2.75,2.75,.5,.5),mgp=c(1.7,.7,0))
par(mfrow=c(1,3))
blabs<-c(expression(beta[1]),expression(beta[2]),expression(beta[3]))
thin<-c(1,(1:1000)*(S/1000))
j<-3
plot(thin,BETA[thin,j],type="l",xlab="iteration",ylab=blabs[j])
abline(h=mean(BETA[,j]) )

acf(BETA[,j],ci.col="gray",xlab="lag")
acf(BETA[thin,j],xlab="lag/10",ci.col="gray")
dev.off()
####
```

6.5.2 Metropolis and Gibbs Combined

In complex models, it is often the case that the conditional distributions are available for some parameters but not for others. What can we do then? In these situations we can combine Gibbs and Metropolis-type proposal distributions to generate a Markov chain to approximate the joint posterior distribution of all the parameters.

- Here, we look at an example of estimating the parameters in a regression model for time-series data, where the errors are temporally correlated.
- The full conditionals are available for the regression parameters here, but not the parameter describing the dependence among the observations.

Example 6.11: Historical CO_2 and temperature data

Analyses of ice cores from East Antarctica have allowed scientists to deduce historical atmospheric conditions of law few hundred years (Petit et al, 1999). Figure 5.18 plots time-series of redtemperature and redcarbon dioxide concentration on a standardized scale (centered and called to have mean of zero and variance of 1).

- The data include 200 values of temperature measured at roughly equal time intervals, with time between consecutive measurements being around 2,000 years.
- For each value of temperature there is a CO_2 concentration value that corresponds to data that is about 1,000 years previous to the temperature value (on average).
- Temperature is recorded in terms of its difference from current present temperature in degrees Celsius and CO_2 concentration is recorded in parts per million by volume.

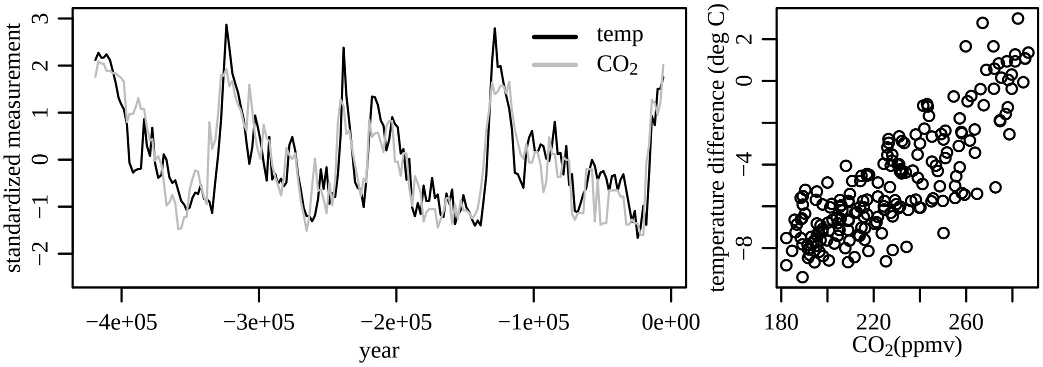


Figure 6.16: Temperature and carbon dioxide data.

- The plot indicates the temporal history of temperature and CO_2 follow very similar patterns.
- The second plot in Figure 5.18 indicates that CO_2 concentration at a given time is predictive of temperature following that time point.
- We can quantify this using a linear regression model for temperature (Y) as a function of $(CO_2)(x)$.
- The validity of the standard error relies on the error terms in the regression model being iid and standard confidence intervals further rely on the errors being normally distributed.
- These two assumptions are examined in the two residual diagnostic plots in Figure 5.19.

- The first plot shows a histogram of the residuals and indicates redno serious deviation from non-normality.
- The second plot gives the autocorrelation function of the residuals, indicating a rednontrivial correlation of 0.52 between residuals at consecutive time points.
- Such a positive correlation generally implies there is less information in the data and less evidence for a relations between the two variables than is assumed by the OLS regression analysis.

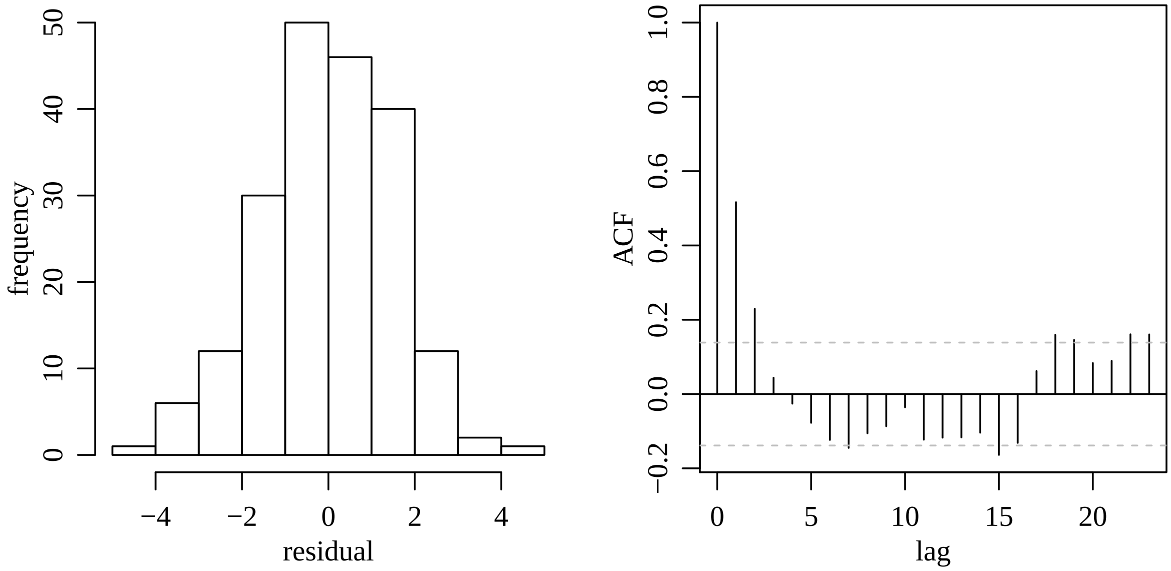


Figure 6.17: Temperature and carbon dioxide data.

The ordinary regression model is

$$\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}).$$

The diagnostic plots suggest that a more appropriate model for the ice core data is one in which the error terms are not independent, but temporally correlated.

We will replace $\sigma^2\boldsymbol{I}$ with a covariance matrix Σ that can represent the positive correlation between sequential observations. One simple, popular class of covariance matrices for temporally correlated data are those having *first order autoregressive structure*:

$$\Sigma = \sigma^2 C_p = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \\ & & & & \\ \vdots & \vdots & & \ddots & \\ & & & & \\ \rho^{n-1} & \rho^{n-2} & & & 1 \end{pmatrix}$$

Under this covariance matrix the variance of $Y_i|\beta, \boldsymbol{x}_i$ is σ^2 but the correlation between Y_i and Y_{i+t} is ρ^t. Using the multivariate normal and inverse gamma prior (it is left as an exercise to show that)

$$\begin{aligned}\boldsymbol{\beta} \mid \boldsymbol{X}, \boldsymbol{y}, \sigma^2, \rho &\sim N(\beta_n, \Sigma_n),\\ \sigma^2 \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\beta} &\sim IG((\nu_o + n)/2, [\nu_o \sigma_o^2 + SSR_\rho]/2)\end{aligned}$$

where $\beta_n = \Sigma_n \left(\boldsymbol{X}^T C_p^{-1} \boldsymbol{X}/\sigma^2 + \Sigma_o^{-1}\beta_o\right)^{-1}$ and $\Sigma_n = \left(\boldsymbol{X}^T C_p^{-1} \boldsymbol{X}/\sigma^2 + \Sigma_o^{-1}\right)^{-1}$ and $SSR_\rho = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T C_p^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$

- If β_o and Σ_o has large diagonal entries, then β_n is very close to

$$(\boldsymbol{X}^T C_p^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T C_p^{-1} \boldsymbol{y}$$

- If ρ were known this would be the generalized least squares (GLS) estimate of β.
- This is a type of weighted LS estimate that is used when the error terms are not iid. In such situations, both OLS and GLS provide unbiased estimates of β but the GLS has lower variance.
- Bayesian analysis using a model that accounts for correlation errors provides parameter estimates that are similar to those of GLS, so for convenience we will refer to our analysis as "Bayesian GLS."

If we knew the value of ρ we could just implement Gibbs to approximate $p(\beta, \sigma^2 | X, y, \rho)$. However, ρ is unknown and typically the distribution of ρ is nonstandard for most prior distributions, suggesting that the Gibbs sampler isn't applicable. What can we do instead?

We can use the redgenerality of the MH algorithm. Recall we are allowed to use different proposals at each step. We can iteratively update β, σ^2, and ρ at different steps (using Gibbs proposals). That is:

- We will make proposals for β and σ^2 using the full conditionals and
- make a symmetric proposal for ρ.
- Following the rules of MH, we accept with prob 1 any proposal coming from a full conditional distribution, whereas we have to calcite an acceptance probability for proposals of ρ.

We run the following algorithm:

1. Update β: Sample $\beta^{(s+1)} \sim N(\beta_n, \Sigma_n)$, where β_n and Σ_n depend on $\sigma^{2(s)}$ and $\rho^{(s)}$.
2. Update σ^2: Sample $\sigma^{2(s+1)}$ ~IG($(\nu_o+n)/2, [\nu_o\sigma_o^2+SSR_\rho]/2)$ where SSR_ρ depends on $\beta^{(s+1)}$ and $\rho^{(s)}$.
3. Update ρ : (a): Propose $\rho^* \sim$ Uniform$(\rho^{(s)} - \delta, \rho^{(s)} + \delta)$. If $\rho^* < 0$ then reassign it to be $|p^*|$. If $\rho^* > 1$ then reassign it to be $2 - \rho^*$.

   (b) Compute the acceptance ratio

$$r = \frac{p(y \mid X, \beta^{(s+1)}, \sigma^{2(s+1)}, \rho^*)\; p(\rho^*)}{p(y \mid X, \beta^{(s+1)}, \sigma^{2(s+1)}, \rho^{(s)})\; p(\rho^{(s)})} and sample$$

   u $\sim$ Uniform$(0, 1)$. If $u < r$, set $\rho^{(s+1)} = \rho^*$, otherwise $\rho^{(s+1)} = \rho^{(s)}$.

The proposal used in Step 3(a) is called *reflecting random walk*, which insures that $0 < \rho < 1$. Note that a sequence of MH steps in which each parameter is updated is often referred to as a *scan* of the algorithm.

Exercise: Show that the proposal is symmetric.

For convenience and ease, we're going to use diffuse priors for the parameters with $\beta_o = 0, \Sigma_o = diag(1000), \nu_o = 1$, and $\sigma^2 = 1$. Our prior on ρ will be Uniform(0,1). We first run 1000 iterations of the MH algorithm and show a trace plot of ρ as well as an autocorrelation plot (Figure 5.20).

Suppose now we want to generate 25,000 scans for a total of 100,000 parameter values. The MC is highly correlated, so we will thin every 25th value in the chain. This reduces the autocorrelation.

The Monte Carlo approximation of the posterior density of β_2 (the slope) appears in the Figure 5.20. The posterior mean is 0.028 with 95 percent posterior credible interval of (0.01,0.05), indicating that the relationship between temperature and CO_2 is positive. As indicated in the second plot this relationship seems much weaker than suggested by the OLS estimate of 0.08. For the OLS estimation, the small number of data points with high y-values have a large influence on the estimate of β. On the other hand, the GLS model recognizes many of these extreme points are highly correlated with one another and down weights their influence.

Remark: this weaker regression coefficient is a result of the temporally correlated data and not of the particular prior distribution we used or the Bayesian approach in general.

Exercise: Repeat the analysis with different prior distributions and perform non-Bayesian GLS for comparison.

```
#####
##example 5.10 in notes
# MH and Gibbs problem
```

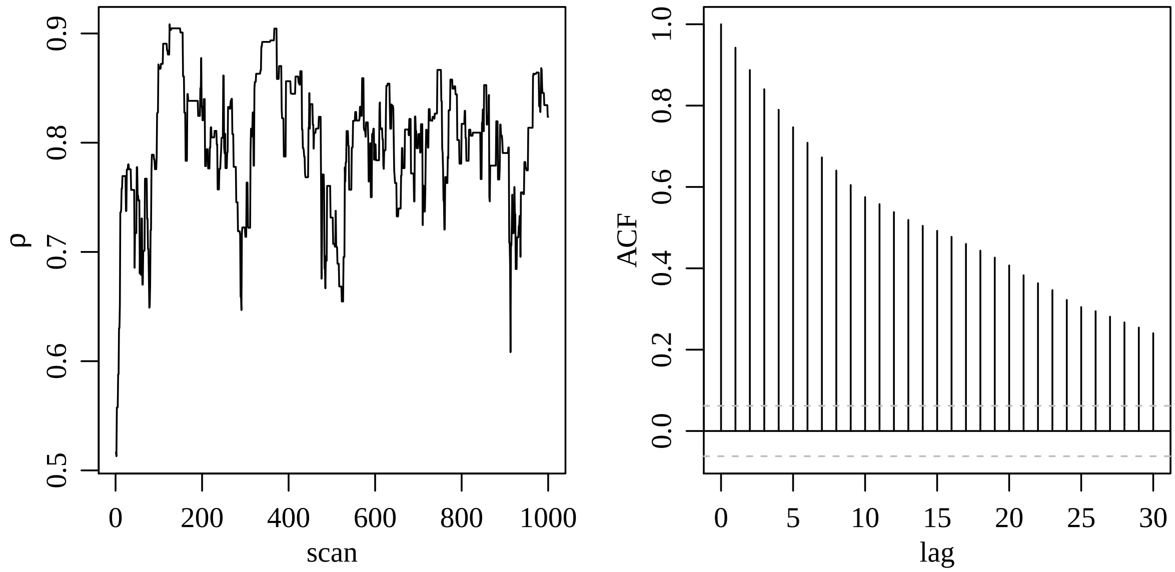


Figure 6.18: The first 1,000 values of ρ generated from the Markov chain.

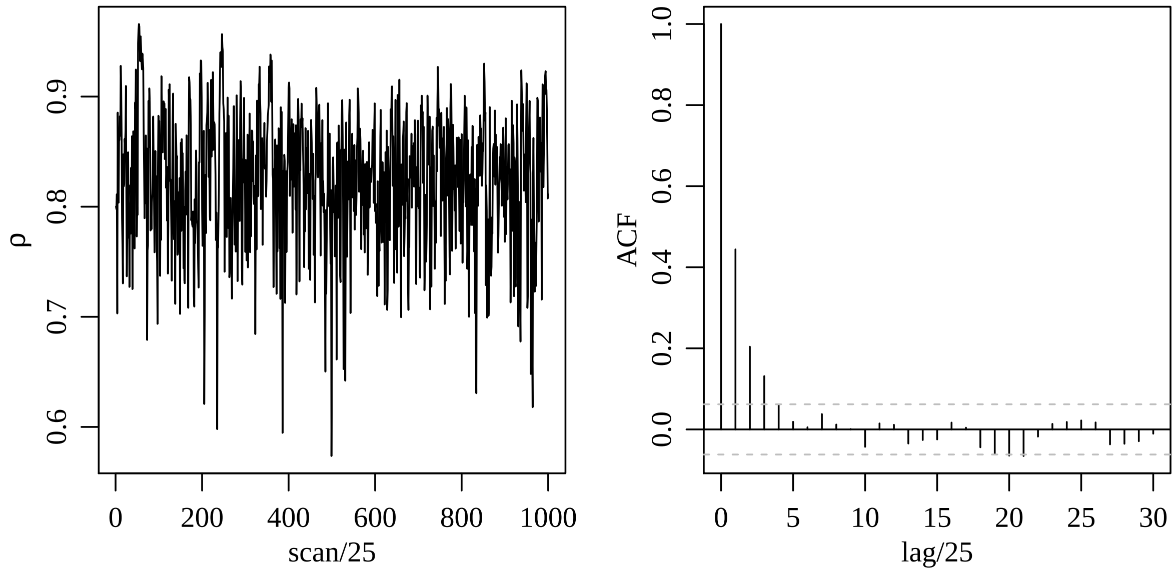


Figure 6.19: Every 25th value of ρ generated from the Markov chain of length 25,000.

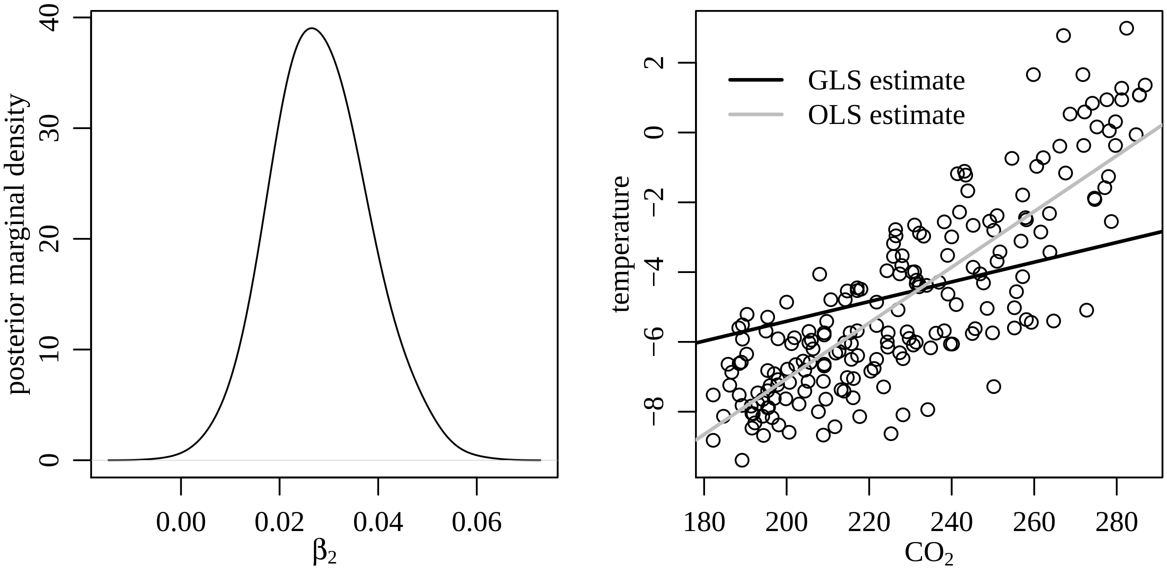


Figure 6.20: Posterior distribution of the slope parameter β_2 and posterior mean regression line (after generating the Markov chain with length 25,000 with thin 25).

```
##temperature and co2 problem

source("http://www.stat.washington.edu/~hoff/Book/Data/data/chapter10.r")


### sample from the multivariate normal distribution
rmvnorm<-function(n,mu,Sigma)
{
  p<-length(mu)
  res<-matrix(0,nrow=n,ncol=p)
  if( n>0 & p>0 )
  {
    E<-matrix(rnorm(n*p),n,p)
    res<-t(  t(E%*%chol(Sigma)) +c(mu))
  }
  res
}
###

##reading in the data and storing it
dtmp<-as.matrix(read.table("volstok.txt",header=F), sep = "-")
dco2<-as.matrix(read.table("co2.txt",header=F, sep = "\t"))
dtmp[,2]<- -dtmp[,2]
dco2[,2]<- -dco2[,2]
library(nlme)

#### get evenly spaced temperature points
ymin<-max( c(min(dtmp[,2]),min(dco2[,2])))
ymax<-min( c(max(dtmp[,2]),max(dco2[,2])))
n<-200
syear<-seq(ymin,ymax,length=n)
dat<-NULL
for(i in 1:n) {
 tmp<-dtmp[ dtmp[,2]>=syear[i] ,]
 dat<-rbind(dat,  tmp[dim(tmp)[1],c(2,4)] )
               }
dat<-as.matrix(dat)
####

####
dct<-NULL
for(i in 1:n) {
  xc<-dco2[ dco2[,2] < dat[i,1] ,,drop=FALSE]
  xc<-xc[ 1, ]
  dct<-rbind(dct, c( xc[c(2,4)], dat[i,] ) )
               }
```

```
mean( dct[,3]-dct[,1])


dct<-dct[,c(3,2,4)]
colnames(dct)<-c("year","co2","tmp")
rownames(dct)<-NULL
dct<-as.data.frame(dct)

##looking at temporal history of co2 and temperature
########
pdf("temp_co2.pdf",family="Times",height=1.75,width=5)
par(mar=c(2.75,2.75,.5,.5),mgp=c(1.7,.7,0))
layout(matrix( c(1,1,2),nrow=1,ncol=3) )

#plot(dct[,1],qnorm( rank(dct[,3])/(length(dct[,3])+1 )) ,
plot(dct[,1],  (dct[,3]-mean(dct[,3]))/sd(dct[,3]) ,
   type="l",col="black",
   xlab="year",ylab="standardized measurement",ylim=c(-2.5,3))
legend(-115000,3.2,legend=c("temp",expression(CO[2])),bty="n",
       lwd=c(2,2),col=c("black","gray"))
lines(dct[,1],  (dct[,2]-mean(dct[,2]))/sd(dct[,2]),
#lines(dct[,1],qnorm( rank(dct[,2])/(length(dct[,2])+1 )),
  type="l",col="gray")

plot(dct[,2], dct[,3],xlab=expression(paste(CO[2],"(ppmv)")),
ylab="temperature difference (deg C)")
dev.off()
########


##residual analysis for the least squares estimation
########
pdf("residual_analysis.pdf",family="Times",height=3.5,width=7)
par(mar=c(3,3,1,1),mgp=c(1.75,.75,0))
par(mfrow=c(1,2))

lmfit<-lm(dct$tmp~dct$co2)
hist(lmfit$res,main="",xlab="residual",ylab="frequency")
#plot(dct$year, lmfit$res,xlab="year",ylab="residual",type="l" ); abline(h=0)
acf(lmfit$res,ci.col="gray",xlab="lag")
dev.off()
########

##BEGINNING THE GIBBS WITHIN METROPOLIS
```

```
######## starting values (DIFFUSE)
n<-dim(dct)[1]
y<-dct[,3]
X<-cbind(rep(1,n),dct[,2])
DY<-abs(outer( (1:n),(1:n) ,"-"))

lmfit<-lm(y~-1+X)
fit.gls <- gls(y~X[,2], correlation=corARMA(p=1), method="ML")
beta<-lmfit$coef
s2<-summary(lmfit)$sigma^2
phi<-acf(lmfit$res,plot=FALSE)$acf[2]
nu0<-1 ; s20<-1 ; T0<-diag(1/1000,nrow=2)
###
set.seed(1)

###number of MH steps
S<-25000 ; odens<-S/1000
OUT<-NULL ; ac<-0 ; par(mfrow=c(1,2))
library(psych)
for(s in 1:S)
{

  Cor<-phi^DY  ; iCor<-solve(Cor)
  V.beta<- solve( t(X)%*%iCor%*%X/s2 + T0)
  E.beta<- V.beta%*%( t(X)%*%iCor%*%y/s2  )
  beta<-t(rmvnorm(1,E.beta,V.beta)  )

  s2<-1/rgamma(1,(nu0+n)/2,(nu0*s20+t(y-X%*%beta)%*%iCor%*%(y-X%*%beta)) /2 )

  phi.p<-abs(runif(1,phi-.1,phi+.1))
  phi.p<- min( phi.p, 2-phi.p)
  lr<- -.5*( determinant(phi.p^DY,log=TRUE)$mod -
             determinant(phi^DY,log=TRUE)$mod  +
   tr( (y-X%*%beta)%*%t(y-X%*%beta)%*%(solve(phi.p^DY) -solve(phi^DY)) )/s2 )

  if( log(runif(1)) < lr ) { phi<-phi.p ; ac<-ac+1 }

  if(s%%odens==0)
    {
      cat(s,ac/s,beta,s2,phi,"\n") ; OUT<-rbind(OUT,c(beta,s2,phi))
#      par(mfrow=c(2,2))
#      plot(OUT[,1]) ; abline(h=fit.gls$coef[1])
#      plot(OUT[,2]) ; abline(h=fit.gls$coef[2])
#      plot(OUT[,3]) ; abline(h=fit.gls$sigma^2)
#      plot(OUT[,4]) ; abline(h=.8284)
```

```
    }
}
#####

OUT.25000<-OUT
library(coda)
apply(OUT,2,effectiveSize )


OUT.25000<-dget("data.f10_10.f10_11")
apply(OUT.25000,2,effectiveSize )


pdf("trace_auto_1000.pdf",family="Times",height=3.5,width=7)
par(mar=c(3,3,1,1),mgp=c(1.75,.75,0))
par(mfrow=c(1,2))
plot(OUT.1000[,4],xlab="scan",ylab=expression(rho),type="l")
acf(OUT.1000[,4],ci.col="gray",xlab="lag")
dev.off()


pdf("trace_thin_25.pdf",family="Times",height=3.5,width=7)
par(mar=c(3,3,1,1),mgp=c(1.75,.75,0))
par(mfrow=c(1,2))
plot(OUT.25000[,4],xlab="scan/25",ylab=expression(rho),type="l")
acf(OUT.25000[,4],ci.col="gray",xlab="lag/25")
dev.off()

pdf("fig10_11.pdf",family="Times",height=3.5,width=7)
par(mar=c(3,3,1,1),mgp=c(1.75,.75,0))
par(mfrow=c(1,2))

plot(density(OUT.25000[,2],adj=2),xlab=expression(beta[2]),
   ylab="posterior marginal density",main="")

plot(y~X[,2],xlab=expression(CO[2]),ylab="temperature")
abline(mean(OUT.25000[,1]),mean(OUT.25000[,2]),lwd=2)
abline(lmfit$coef,col="gray",lwd=2)
legend(180,2.5,legend=c("GLS estimate","OLS estimate"),bty="n",
      lwd=c(2,2),col=c("black","gray"))
dev.off()

quantile(OUT.25000[,2],probs=c(.025,.975) )

plot(X[,2],y,type="l")
points(X[,2],y,cex=2,pch=19)
points(X[,2],y,cex=1.9,pch=19,col="white")
text(X[,2],y,1:n)

iC<-solve( mean(OUT[,4])^DY )
Lev.gls<-solve(t(X)%*%iC%*%X)%*%t(X)%*%iC
Lev.ols<-solve(t(X)%*%X)%*%t(X)

plot(y,Lev.ols[2,] )
plot(y,Lev.gls[2,] )
```

6.6 Introduction to Nonparametric Bayes

As we have seen, Bayesian parametric methods takes classical methodology for prior and posterior distributions in models with a finite number of parameters. It is often the case that the number of parameters taken in such model is low for computational complexity, however, in current research problem we deal with high dimensional data and high dimensional parameters. The origins of Bayesian methods have been around since the mid-1700's and are still thriving today. The applicability of Bayesian parametric models still remains and has widened with the increased advancements made in modern computing and the growth of methods available in Markov chain Monte Carlo.

Frequentist nonparametrics covers a wide array of areas in statistics. The area is well known for being associated with testing procedures that are or become asymptotically distribution free, which lead to nonparametric confidence intervals, bands, etc. (Hjors et al., 2010). Further information can be found on these methods in Wasserman (2006).

Nonparametric Bayesian methods are models and methods characterized generally by large parameter spaces, such as unknown density and regression functions and construction of probability measures over these spaces. Typical examples seen in practice include density estimation, nonparametric regression with fixed error distributions, hazard rate and survival function estimation. For a thorough introduction into this subject see (Hjors et al., 2010).

6.6.1 Motivations

The motivation is the following:

- We have $X_1 \dots X_n \overset{iid}{\sim} F, F \in \mathcal{F}$. We usually assume that $\mathcal{F}$ is a parametric family.
- Then, putting a prior on $\mathcal{F}$ amounts to putting a prior on $\mathbb{R}^d$ for some d.
- We would like to be able to put a prior on all the set of cdf's. And we would like the prior to have some basic features:

1. The prior should have large support.

2. The prior should give rise to priors which are analytically tractable or computationally manageable.

3. We should be able to center the prior around a given parametric family.

6.6.2 The Dirichlet Process

Review of Finite Dimensional Dirichlet Distribution

- This is a distribution on the k-dimensional simplex.
- Let $(\alpha_1, \dots, \alpha_k)$ be such that $\alpha_j > 0$ for all j.
- The Dirichlet distribution with parameter vector $(\alpha_1, \dots, \alpha_k)$ has density

$$p(\theta) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\prod_{j=1}^k \Gamma(\alpha_j)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}.$$

- It is conjugate to the Multinomial distribution. That is if $\boldsymbol{Y} \sim Multinomial(n, \boldsymbol{\theta})$ and $\boldsymbol{\theta} \sim Dir(\alpha_1, \dots, \alpha_k)$, then it can be shown that

$$\boldsymbol{\theta}|\boldsymbol{y} \sim Dir(\alpha_1 + N_1, \dots, \alpha_k + N_k).$$

- It can be shown that

$$E(\theta_j) = \alpha_j/\alpha$$

  where $\alpha = \sum_j \alpha_j$. It can also be shown that

$$Var(\theta_j) = \frac{\alpha_j(\alpha - \alpha_j)}{\alpha^2(\alpha + 1)}.$$

Infinite Dimensional Dirichlet Distribution

Let α be a finite (non-null) measure (or think probability distribution) on $\mathbb{R}$. Sometimes α will be called the concentration parameter (in scenarios when we might hand wave the measure theory for example or it's not needed).

You should think about the Infinite Dimension Dirichlet Distribution as a reddistribution of distributions as we will soon see.

Definition 6.1: F has the Dirichlet distribution with parameter (measure) α if for every finite measurable partition $A_1, \ldots, A_k$ of $\mathbb{R}$ the k-dimensional random vector $(F(\{A_1\}), \ldots, F(\{A_k\}))$ has the finite k-dimensional Dirichlet distribution

$$\text{Dir}(\alpha(A_1), \ldots, \alpha(A_k)).$$

For more on this see: Freedman (1963), Ferguson (1973, 1974).

Intuition: Each $F(\{A_k\}) \in [0,1]$ since F is some cumulative distribution function. Also,

$$F(\{A_1\}) + \cdots + F(\{A_k\}),$$

thus, $(F(\{A_1\}), \ldots, F(\{A_k\}))$ lives on the k-dimensional simplex.

Remark: For those with measure theory: You can't have a measure that is 0. Note that Lebesgue measure isn't finite on the reals.

We will construct the Dirichlet process to intuitively understand it based on the "Polya Urn Scheme" of Blackwell and MacQueen (1973). This is one of the most intuitive approaches. Others in the literature include Ferguson (1973, 1974), which include two constructions. There is an incorrect constructions involve the Kolmovgorov extension theorem (the problem is that the sets aren't measurable, so an important technical detail). The other is a correct construction based on something called the gamma process (this involves much overhead and existence of the gamma process).

6.6.3 Polya Urn Scheme on Urn With Finitely Many Colors

- Consider an urn containing a finite number of balls of k different colors.
- There are $\alpha_1, \ldots, \alpha_k$ balls of colors $1 \ldots, k$, respectively.
- We pick a ball at random, look at its color, return it to the urn together with another ball of the same color.
- We repeat this indefinitely.
- Let $p_1(n), \ldots p_k(n)$ be the proportions of balls of colors $1 \ldots, k$ at time n.

Example 6.12: Polya Urn for Three Balls
Suppose we have three balls in our urn. Let redred correspond the ball 1. Let blueblue correspond the ball 2. Let greengreen correspond the ball 3. Furthermore, suppose that $P(redred) = 2/9), P(blueblue) = 3/9$ and $P(greengreen) = 4/9$.

Let α be a the following probability measure (or rather discrete probability distribution):

- $\alpha_o(1) = 2/9$.
- $\alpha_o(2) = 3/9$.
- $\alpha_o(3) = 4/9$.

Another way of writing this is define $\alpha_o = 2/9\ \delta_1 + 3/9\ \delta_2 + 4/9\ \delta_3$ where

$$\delta_1(A) = \begin{cases} 1 & \text{if } 1 \in A \\ 0 & \text{otherwise.} \end{cases}$$

6.6.4 Polya Urn Scheme in General

Let α be a finite measure on a space $\mathcal{X} = \mathbb{R}$.

1. (Step 1) Suppose $X_1 \sim \alpha_o$.
2. (Step 2) Now create a new measure $\alpha + \delta_{X_1}$ where

$$\delta_{X_1}(A) = \begin{cases} 1 & \text{if } X_1 \in A \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$X_2 \sim \frac{\alpha + \delta_{X_1}}{\alpha(\mathtt{R}) + \delta_{X_1}(\mathtt{R})} = \frac{\alpha + \delta_{X_1}}{\alpha(\mathtt{R}) + 1}.$$

Fact: $\delta_{X_1}(\mathtt{R}) = 1$. Think about why this is intuitively true.

What does the above equation really mean?

– α represents the original distribution of balls.

– δ_{X_1} represents the ball we just added.

Deeper understanding

– Suppose the urn contained N total balls when we started.

– Then the probability that the second ball drawn X_2 will be of the original N balls is $N/(N+1)$.

– This represents the α part of the distribution of X_2.

– We want the probability of drawing a new ball to be $1/(N+1)$. This goes with δ_{X_1}.

– When we write $X_2 \sim \dfrac{\alpha + \delta_{X_1}}{\text{norm. constant}}$ we want $N/(N+1)$ of the probability to go to $\dfrac{\alpha}{\text{norm. constant}}$ and $1/(N+1)$ to go to $\dfrac{\delta_{X_1}}{\text{norm. constant}}$.

How does this continue? Since we want

$$= \frac{\delta_{X_1}(\mathtt{R})}{\alpha(\mathtt{R}) + 1} = 1/(N+1)$$

this implies that

$$\frac{1}{\alpha(\mathtt{R}) + 1} = 1/(N+1) \implies \alpha(\mathtt{R}) = N.$$

Hence, we take $\alpha(\mathtt{R}) = N$ and then we plug back in and find that $\alpha_o = \alpha/N \implies \alpha = \alpha_o\, N$.

This implies that

$$X_2 \sim \frac{\alpha_o\, N + \delta_{X_1}}{N+1},$$

which is now in terms of α_o and N (which we know).

(Step 3) Continue forming new measures: $\alpha + \delta_{X_1} + \delta_{X_2}$. Then

$$X_3 \sim \frac{\alpha + \delta_{X_1} + \delta_{X_2}}{\alpha(\mathtt{R}) + \delta_{X_1}(\mathtt{R}) + \delta_{X_2}(\mathtt{R})} = \frac{\alpha + \delta_{X_1} + \delta_{X_2}}{\alpha(\mathtt{R}) + 2}.$$

In general, it can be shown that

$$P(X_{n+1} \mid X_1 \dots X_n) = \frac{\alpha(A) + \sum_{i=1}^n \delta_{X_i}(A)}{\alpha(\mathtt{R}) + n} = \frac{\alpha_o\, N + \sum_{i=1}^n \delta_{X_i}(A)}{N + n}.$$

Polya Urn Scheme in General Case: Theorem

- Let α be a finite measure on a space X (this space can be very general, but we will assume it's the reals).
- Define a sequence $\{X_1, X_2, \dots\}$ of random variables to be a *Polya urn sequence with parameter measure* α if
  - $P(X_1 \in B) = \alpha(B)/\alpha(\mathtt{R})$.
  - For every n,

$$P(X_{n+1} \in B \mid X_1 \dots, X_n) = \frac{\alpha(B) + \sum_i \delta_{X_i}(B)}{\alpha(\mathtt{R}) + n}.$$

Specifically, $X_1, X_2, \dots,$ is PUS(α) if

$$P(X_1 \in A) = \frac{\alpha(A)}{\alpha(\mathtt{R})} = \alpha_o$$

for every $A \in \mathtt{R}$ and for every n

$$P(X_{n+1} \in B \mid X_1 \dots, X_n) = \frac{\alpha(B) + \sum_i \delta_{X_i}(B)}{\alpha(\mathtt{R}) + n}$$

for every $A \in \mathtt{R}$.

6.6.5 De Finetti and Exchaneability

Recall what exchangeability means. Suppose that $Y_1, Y_2, \dots, Y_n$ is a sequence of random variables. This sequence is said to be exchangeable if the distribution of

$$(Y_1, Y_2, \dots, Y_n) \overset{d}{=} (Y_{\pi(1)}, Y_{\pi(2)}, \dots, Y_{\pi(n)})$$

for every permutation π of $1, \dots, n$.

Note: This means that we can permute the random variables and the distribution doesn't change.

An infinite sequence is said to be exchangeable if for every n, $Y_1, Y_2, \dots, Y_n$ is exchangeable. That is, we don't require exchangeability for infinite permutations, but it must be true for every "chunk" that we take that is of length or size n.

Definition 6.2: De Finetti's General Theorem
Let $X_1, X_2 \dots$ be an infinite exchangeable sequence of random variables. Then there exists a probability measure π such that

$$X_1, X_2 \dots, \mid F \overset{iid}{\sim} F$$

$$F \sim \pi$$

for any $x_1, \dots x_n \in \{0, 1\}$.

Remark: Suppose that $X_1, X_2 \dots$ is an infinite exchangeable sequence of binary random variables. Then there exists a probability measure (distribution) on $[0, 1]$ such that for every n

$$P(X_1 = x_1, \dots, X_n = x_n) = \int_0^1 p^{\sum_i x_i} (1-p)^{n - \sum_i x_i} \mu(p) dp$$

where $\mu(p)$ is the measure or probability distribution or prior that we take on p.

Theorem 6.1: A General Result (without proof)
Let $X_1, X_2 \ldots$ be PUS(α). Then this can be thought of as a two-stage process where

- $F \sim \mathrm{Dir}(\alpha)$
- $X_1, X_2 \ldots, \mid F \overset{iid}{\sim} F$

If we consider the process of the PUS consisting of $X_2, X_3 \ldots$, then it's a PUS($\alpha + \delta_{X1}$). That is, $F \mid X_1 \sim \mathrm{Dir}(\alpha + \delta_{X1})$.

More generally, it can be shown that
$F \mid X_1 \ldots X_n \sim \mathrm{Dir}(\alpha + \sum_{i=1}^{n} \delta_{Xi})$.

6.6.6 Chinese Restaurant Process

- There are Bayesian NP approaches to many of the main issues in statistics including:
  - regression.
  - classification.
  - clustering.
  - survival analysis.
  - time series analysis.
  - spatial data analysis.
- These generally involve assumptions of redexchangeability or partial redexchangeability.
  - and corresponding distributions on random objects of various kinds (functions, partitions, measures, etc.)
- We look at the problem of clustering for concreteness.

6.6.7 Clustering: How to choose K?

- Adhoc approaches (hierarchical clustering)
  - these methods do yield a data-drive choice of K
  - there is little understanding how good these choices are (meaning the checks are adhoc based on some criterion)
- Methods based on objective functions (M-estimators)
  - K-means, spectral clustering
  - they come with some frequentist guarantees
  - it's often hard to turn these into data-driven choices of K
- Parametric likelihood-based methods
  - finite mixture models, Bayesian variants
  - various model choice methods: hypothesis testing, cross-validation, bootstrap, AIC, BIC, Laplace, reversible jump MCMC
  - do the assumptions underlying the method apply to the setting (not very often)
- Something different: The Chinese restaurant process.

Basic idea: In many data analysis settings, we don't know the number of latent clusters and would like to learn it from the data. BNP clustering addresses this by assuming there is an infinite number of latent clusters, but that only a finite number of them is used to generate the observed data. Under these assumptions, the posterior yields a distribution over the number of clusters, the assign of data to clusters, and the parameters associated with each cluster. In addition, the predictive distribution, the assignment of the next data point, allows for new data to be assign to a previously unseen cluster.

How does it work: The BNP problem addresses and finesses the clustering problem by choosing the number of clusters by assuming it is infinite, however it specifies a prior over the infinite groupings $P(c)$ in such a way that favors assigning data to a small number of groups, where c refers to the cluster assignments. The prior over groupings is a well known problem called the Chinese restaurant process (CRP), which is a distribution over infinite partition of the integers (Aldous, 1985; Pitman, 2002).

Where does the name come from?

- Imagine that Sam and Mike own a restaurant with an infinite number of tables.
- Imagine a sequence of customers entering their restaurant and sitting down.
- The first customer (Liz) enters and sits at the first table.
- The second customer enters and sits at the first table with probability $\frac{1}{1+\alpha}$ and a new table with probability $\frac{\alpha}{1+\alpha}$, where α is positive and real.
- Liz is friendly and people would want to sit and talk with her. So, we would assume that $\frac{1}{1+\alpha}$ is a high probability, meaning that α is a small number.
- What happens with the nth customer?
  * He sits at each of the previously occupied tables with probability proportional to the number previous customers sitting there.
  * He sits at the next unoccupied table with probability proportional to α.

More formally, let c_n be the table assigned me of customer n. A draw from this distribution can be generated by sequentially assigning observations with probability

$$P(c_n = k \mid c) = \begin{cases} \frac{m_k}{n-1+\alpha} & \text{if } k \leq K_+ \text{ (i.e. k is a previously occupied table),} \\ \frac{\alpha}{n-1+\alpha} & \text{otherwise (i.e. k is the next unoccupied table),} \end{cases}$$

where m_k is the number of customers sitting at table k and K_+ is the number of table for which $m_k > 0$. The parameter α is called the concentration parameter.

The rich just get richer

- CRP rule: next customer sits at a table with prob. proportional to number of customers already sitting at it (and sits at new table with prob. proportional to α).
- Customers tend to sit at most popular tables.
- Most popular tables attract the most new customers, and become even more popular.
- CRPs exhibit power law behavior, where a few tables attract the bulk of the customers.
- The concentration parameter α determines how likely a customer is to sit at a fresh table.

More formally stated:

- A larger value of α will produce more occupied tables (and fewer customers per table).
- Thus, a small value of α produces more customers at each table.
- The CRP exhibits an important invariance property: the cluster assignments under this distribution are exchangeable.
- This means that $p(c)$ is unchanged if the order of customers is shuffled (up to label changes). This may be counter-intuitive since the process was just described sequentially.

The CRP and Clustering

- The data points refer to the customers and the tables are the clusters.
  - Then the CRP defines a prior distribution on the partition of the data and on the number of tables.
- The prior can be completed with:
  - A likelihood, meaning there needs to be an parameterized probability distribution that corresponds to each table
  - A prior for the parameters –the first customer to sit at table k chooses the parameter vector for that table (ϕ_k) from the prior
- Now that we have a distribution for any quantity we care about in some clustering setting.

Now, let's think about how we would write down this process out formally. We're writing out a mixture model with a component that's nonparametric.

Let's define the following:

- y_n are the observations at time n.
- c_n are the latent clusters that generate c_n.
- F is a parametric family of distributions for y_n.
- θ_k represent the clustering parameters.
- G_o represents a general prior for the clustering parameters (this is the nonparametric part).

We also assume that each observation is conditionally independent given its latent cluster assignment and its cluster parameters.

Using the CRP, we can view the model as

$$\begin{aligned} y_n \mid c_n, \theta &\sim F(\theta_{c_n}) \\ c_n &\propto p(c_n) \\ \theta_k &\propto G_o. \end{aligned}$$

We want to know $p(y \mid c)$.

Then by Bayes' rule, $p(c|y) = \dfrac{p(y|c)p(c)}{\sum_c p(y|c)p(c)}$, where

$$p(y \mid c) = \int_\theta \left[\prod_{n=1}^{N} F(y|\theta_{c_n}) \prod_{k=1}^{K} G_o(\theta_k) \right] \, d\theta.$$

A G_o that is conjugate allow this integral to be calculated analytically. For example, the Gaussian is the conjugate prior to a Gaussian with fixed variance (and thus a mixture of Gaussians model is computationally convenient). We illustrate this specific example below.

Example 6.13: Suppose

$$\begin{aligned} y_n \mid c_n, \theta &\sim N(\theta_{c_n}, 1) \\ c_n &\sim \text{Multinomial}(1, p) \\ \theta_k &\sim N(\mu, \tau^2), \end{aligned}$$

where p, μ, and τ^2 are known.

Then

$$p(y|c) = \int_\theta \left[\prod_{n=1}^{N} \text{Normal}(\theta_{c_n}, 1)(y_n) \times \prod_{k=1}^{K} \text{Normal}(\mu, \tau^2)(\theta_k) \right] d\theta.$$

The term above (inside the integral) is just another normal as a function of θ. Then we can integrate θ out as we have in problems before.

Once we calculate $p(y|c)$, we can simply plug this and $p(c)$ into

$$p(c|y) = \frac{p(y|c)p(c)}{\sum_c p(y|c)p(c)}.$$

Example 6.14: Gaussian Mixture using `R`

Information on the `R` package `profdpm`:

This package facilitates inference at the posterior mode in a class of conjugate product partition models (PPM) by approximating the maximum a posteriori data (MAP) partition. The class of PPMs is motivated by an augmented formulation of the Dirichlet process mixture, which is currently the ONLY available member of this class. The profdpm package consists of two model fittting functions, `profBinary` and `profLinear`, their associated summary methods summary.profBinary and summary.profLinear, and a function (pci) that computes several metrics of agreement between two data partitions. However, the profdpm package was designed to be extensible to other types of product partition models. For more on this package, see help(profdpm) after installation.

- The following example simulates a dataset consisting of 99 longitudinal measurements on 33 units of observation, or subjects.
- Each subject is measured at three times, drawn uniformly and independently from the unit interval.
- Each of the three measurements per subject are drawn independently from the normal distribution with one of three linear mean functions of time, and with unit variance.
- The linear mean functions vary by intercept and slope. The longitudinal structure imposes a grouping among measurements on a single subject.

- Observations grouped in this way should always cluster together. A grouping structure is specified using the group parameter; a factor that behaves similarly to the groups parameter of lattice graphics functions.
- For the PPM of conjugate binary models, the grouping structure is imposed by the model formula.
- Grouped observations correspond to rows of the model matrix, resulting from a call to model.matrix on the formula passed to profBinary. Hence, the profBinary function does not have a group parameter in its prototype.
- The goal of the following example is to recover the simulated partition and to create simultaneous 95% credible bands for the mean within each cluster. The following R code block creates and the simulated dataset.

```
set.seed(42)
sim <- function(multiplier = 1) {
x <- as.matrix(runif(99))
a <- multiplier * c(5,0,-5)
s <- multiplier * c(-10,0,10)
y <- c(a[1]+s[1]*x[1:33],
a[2]+s[2]*x[34:66],
a[3]+s[3]*x[67:99]) + rnorm(99)
group <- rep(1:33, rep(3,33))
return(data.frame(x=x,y=y,gr=group))
}
dat <- sim()
library("profdpm")
fitL <- profLinear(y ~ x, group=gr, data=dat)
sfitL <- summary(fitL)
%pdf(np_plot.pdf)
plot(fitL$x[,2], fitL$y, col=grey(0.9), xlab="x", ylab="y")
for(grp in unique(fitL$group)) {
ind <- which(fitL$group==grp)
ord <- order(fitL$x[ind,2])
lines(fitL$x[ind,2][ord],
fitL$y[ind][ord],
col=grey(0.9))
}
for(cls in 1:length(sfitL)) {
# The following implements the (3rd) method of
# Hanson & McMillan (2012) for simultaneous credible bands
# Generate coefficients from profile posterior
n <- 1e4
```

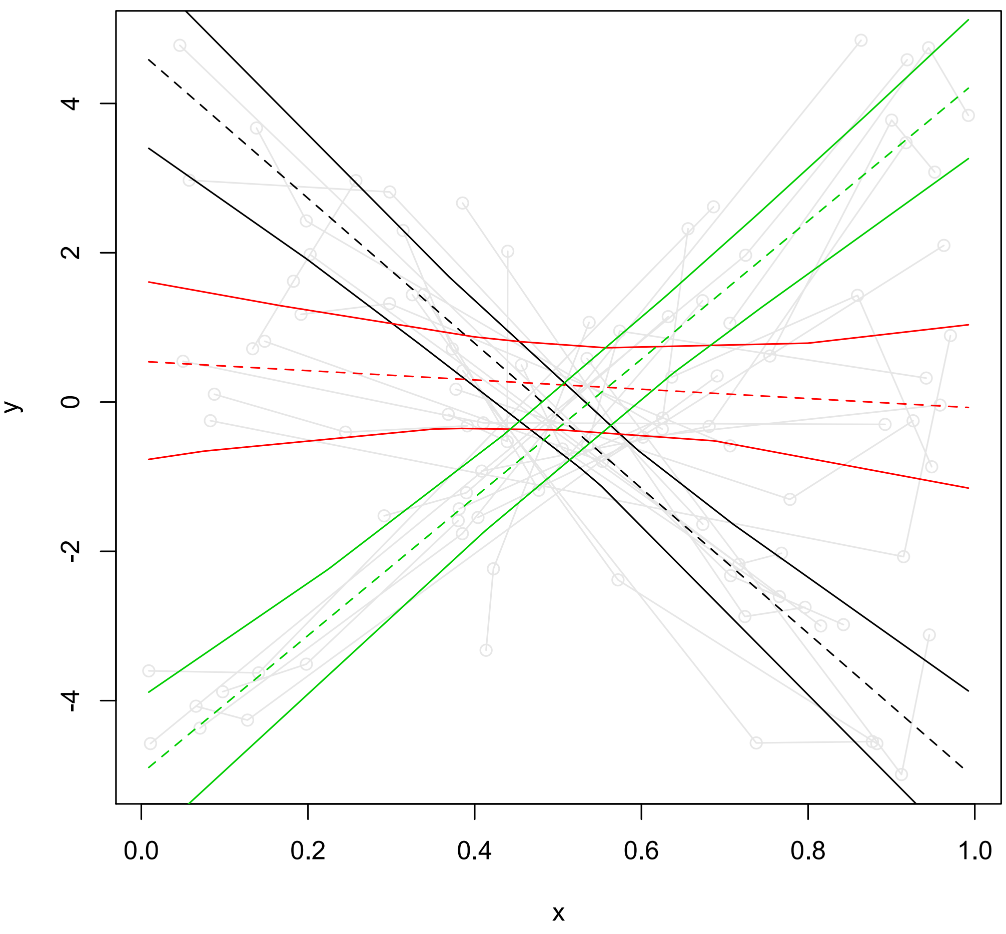


Figure 6.21: Simulated data; 99 longitudinal measurements on 33 subjects. Simultaneous confidence bands for the mean within each of the three clusters.

```
tau <- rgamma(n, shape=fitL$a[[cls]]/2, scale=2/fitL$b[[cls]])
muz <- matrix(rnorm(n*2, 0, 1),n,2)
mus <- (muz / sqrt(tau)) %*% chol(solve(fitL$s[[cls]]))
mu <- outer(rep(1,n), fitL$m[[cls]]) + mus

# Compute Mahalanobis distances
mhd <- rowSums(muz^2)

# Find the smallest 95% in terms of Mahalanobis distance
# I.e., a 95% credible region for mu
ord <- order(mhd, decreasing=TRUE)[-(1:floor(n*0.05))]
mu <- mu[ord,]
#Compute the 95% credible band
plotx <- seq(min(dat$x), max(dat$x), length.out=200)
ral <- apply(mu, 1, function(m) m[1] + m[2] * plotx)
rlo <- apply(ral, 1, min)
rhi <- apply(ral, 1, max)
rmd <- fitL$m[[cls]][1] + fitL$m[[cls]][2] * plotx

lines(plotx, rmd, col=cls, lty=2)
lines(plotx, rhi, col=cls)
lines(plotx, rlo, col=cls)
}
%dev.off()
```