Chapter 2

Conditional Probability and Bayes' Rule

2.1 Introduction

As we said earlier, the fundamental problem of statistics is drawing conclusions about unknowns (parameters, missing data, future observations) based on a sample of data from the population and prior information. The Bayesian formalism, we also said, follows closely the scientific method by giving us a rule to update prior information via Bayes theorem. Our uncertainties about the unknowns are given as probability statements.

Bayes theorem is not only a method to do formal statistical analysis. It also resembles the way we think. To illustrate what we mean by this, consider the following two examples.

Example 1. In observing cards selected from a deck, we may have prior information that only one of these two models is possible: $p_1 = 0.1$, the deck contains 10% spades. $p_2 = 0.2$, the deck of cards contains 20% spades. As we select cards (either with or without replacement) from the deck, i.e., as we observe data, we decide how strongly the observations support p_1 as opposed to p_2. We will never be sure which model is correct, but we ought to be able to discriminate reasonably well if the number of cards selected is large. And we ought to be willing to rethink our set of models should we get something inconsistent with both, such as 40 spades among the the first 50 cards observed. That is, we will update our prior opinion about the model and perhaps be more inclined to think that the model is perhaps closer to $p = 0.6$ or $p = 0.7$. This would be our posterior belief.

Every set of models is limited by the process of specifying it, so any particular set may fail to contain the true model. When a set of models does not contain the true model, a large sample will be most consistent with those models in the set that are close to the true one. A realistic set to specify is one in which the models are sufficiently spread out that at least some will be close to those models in the realm of possibility, and therefore also close to the true model.

Example 2. Viscosity of dimethylaniline. Suppose we don't have any prior information about the viscosity of dimethylaniline at $20°C$, that is, our prior information model is that the average viscosity is uniformly distributed. To obtain information about the average viscosity of dimethylaniline at $20°C$, a scientist makes the following 12 measurements (in centipoises or cP): 146, 154, 141, 140, 136, 132, 147, 140, 147, 139, 140, 140.

The measurements vary because the scientist cannot control all aspects of the procedure. A stem-and-leaf plot will convey the accuracy and variability of the measuring process and show how the 12 observations relate to each other. A stem-and-leaf plot like the one below has the first two digits of each number on the stem and the last digit on the leaf. When there are replicates, as in this example, the last digit will repeat, thus showing a longer leaf. The extreme viscosities are easy to see in the stem-and-leaf plot. Something else we can see is that the measurements seem to be centered near 140 cP. The following R code and output contain all we need for now to see the data.

```
viscosity=c(146,154,141,140,136,132,147,140,147,139,140,140)  # enter data manually
sorted.viscosity=sort(viscosity)     # sort the data
```

```
 sorted.viscosity      # view the sorted data
 [1] 132 136 139 140 140 140 140 141 146 147 147 154  # output is the sorted data
 pdf('viscosity.pdf')   # I type this to save the plot in a file. You do not need it if copy pasting
 stem(viscosity)     # do a stem-and-leaf plot

#########  This is the stem-and-leaf plot given by R 33333333
  The decimal point is 1 digit(s) to the right of the |

  13 | 2
  13 | 69
  14 | 00001
  14 | 677
  15 | 4

> dev.off()    # you need this only if you saved the file
null device     #  R response to closing the file
          1       # more R response to closing the file
```

Thus since the biggest observation is 154cP, it would be surprising if the next observation were 180 cP. So based on the data and the prior information, we would not predict our next observation to be larger than 180 cP. On the other hand, four of the 12 observations were 140cP. The sample size is small. But it is large enough to suspect that the process producing these observations - the population – produces fewer at 180 cP than at 140 cP.

If we knew the population, we would have complete information when predicting the next observation. We do not know the population, but we do have partial information, since the sample we have is a subset of the population. The basic problem of statistics is to make inferences about the population and, in turn, to predict future observations.

Using –or conditioning on– information is the subject of our course.

We use probability to measure uncertainties such as those just discussed. Using information means updating probabilities. How do we modify probabilities in the light of accumulating evidence? For example, after the San Francisco Bay Area earthquake in October 1989, the *New York Times* reported that examination of surface fissures had some geologists wondering "whether [earlier] probability estimates of future quakes are twoo low." The fundamental problem of statistics is using observations to update probabilities concerning the models that have produced these observations.

> Assigning probabilities to potential models a priori, observing data, and modifying (or not) the a priori probabilities of the models based on the data is the method of learning that is most consistent with the scientific method. Bayes theorem encompasses that method in one single formula.

We will do in this chapter a basic review of probability theory and models relevant to the course. We start with generic events, and then move quickly to specifying the events in terms of random variables and their probability distributions.

2.2 Joint probabilities

Describing how probabilities change as we condition on events requires joint probabilities. These are probabilities of several events occurring simultaneously. The simultaneous occurrence of two events A and B is their product or intersection. The standard notation for the intersection of events A and B is $A \cap B$. The event $A \cap B$ contains those outcomes that are in both A and B. For example, $\{1, 2, 3, 4\} \cap \{3, 4, 5\} = \{3, 4\}$. An outcome is in the intersection of two events if it is in the first and it is also in the second.

When two events contain no outcomes in common, then their intersection is empty. For example, $\{1, 2, 3\} \cap \{5, 6\} =$. Such events are mutually exclusive or disjoint.

The use of the word *and* is ambiguous and makes the distinction between union $\cup$ and intersection $\cap$ confusing. The union of two events includes those outcomes in the first *and* those in the second. The intersection of two events

includes those outcomes in the first *and* in the second. The first *and* refers to adding outcomes; the second *and* refers to the simultaneous occurrence of two conditions.

Independent events and independent experiments

$A, B, C, ...$ are independent events whenever

$$P(A \cap B \cap C \cap .....) = P(A)P(B)P(C).....$$

and any subset of events is also independent.

Now consider two (or more) experiments. In addition, consider any combination of outcomes, one from each of the experiments. If these outcomes are independent, so are the experiments: Experiments are independent if the outcomes from one are independent of the outcomes from the other, and this holds for any pair of outcomes, one from each experiment.

Example 3. Important control systems on aircraft have built-in redundancies. Some systems have as many as three alternatives in case they fail. Suppose four systems function independently and the probability of a failure on any of them during a flight is 1%. What is the probability that all four fails?

The answer is $0.01^4 = 0.00000008$.

2.3 Conditional Probability

Conditioning on events or circumstances –assuming that they have occurred or that they apply–is standard in our thinking and in our discourse. But it is difficult to keep the logic straight. For example, during the health care debate of 1993-1994, and ad placed by the League of Women Voters and the Kaiser Family Fund said: " 84 percent of Americans who lack health insurance are in families that work hard and pay taxes, but don't get health insurance on the job." The less than meticulous reader may read "84 percent of Americans... don't get health insurance on the job." Clearly wrong. Only about 15% of workers did not have health insurance. To further confuse the issue, the ad continues" "That's eight out ten of us, and that's a fact." The conditional in the latter statement is "of us.", apparently referring to all "all Americans" rather than "Americans who lack health insurance" The problems with language in this example are easy for the careful reader to sort out. Other examples are more difficult to decipher and you will encounter some of them in this chapter.

Suppose you assess your probability of B. Then I tell you that event A happened. I ask you how this knowledge changes your probability of B. You may tell me that it does not. An obvious example is when A is the universe U, and so I have given you no new information. More generally, A and B are independent if the information contained in A is irrelevant for B.

But your probabilities will sometimes change. For example, suppose $A = B^c$. Now when I tell you A happened, you know that B did not happen. The new probability of B is 0. Or, suppose $A = B$. Then you know B did happen. Its new probability is 1. So your probability of B can change dramatically -regardless of how big or small it was initially– depending on the new information, A.

Example 4. Probability of Survival with breast cancer . A 60-year-old woman has just had a lumpectomy. My probability that she will survive at least 5 more years with optimal care is 90%. Now we are told that she had 25 axillary lymph nodes removed and 15 of them tested positive, meaning that her breast cancer has spread to the lymph nodes. My probability of her survival decreases to 50%. Now suppose tests carried out before her lumpectomy are reviewed and it is discovered her breast cancer has spread to her liver. Now my probability that she survives the next 5 years drops to 10%.

It is easier to define conditional probability if we have a notation for the probability of B given A. We will use $P(B|A)$. The definition is

$$P(B|A) = \frac{P(B \cap A)}{P(A)}.$$

When A contains no information about B, then $P(B) = P(B \mid |A)$ and we say that A and B are independent.

Example 5. How many girls? A family has two children. You are told that at least one is a girl. Call this the event C. What is your probability that the other is a girl? Call D the event "both children are girls."

$$P(D \mid C) = \frac{P(C \cap D)}{P(C)} = \frac{P(both \quad girls)}{P(C)} = \frac{1/4}{3/4} = 1/3$$

There is a neat generalization of the multiplication rule for independent events. For any two events A and B, multiply both sides of the definition of conditional probaility by $P(A)$:

$$P(A)P(B \mid A) = P(A)\frac{P(B \cap A)}{P(A)} = P(A \cap B)$$

The multiplication rule is then

$$P(A \cap B) = P(A)P(B \mid A)$$

when the events are independent, it reduces to the formula given above

$$P(A \cap B) = P(A)P(B)$$

2.4 Exchangeable Experiments

The definition of independence is quite restrictive. For example, repeated observations to generate a sample from a population are not independent. Conditioning on the first observation changes the probabilities for the second. An assumption that is not so restrictive is exchangeability. Two experiments are *exchangeable for you* if all the following hold:

(1) The possible outcomes are the same in both

(2) The probability of each outcome in one is the same as it is in the other.

(3) The conditional probabilities for the second experiment, given the results of the first, are the same as the conditional probability for the first, given the results of the second.

2.5 Law of total probability

There are many circumstances in which you would like to know the probability of an event, but you can not calculate it directly. You may be able to find it if you know its probability under some conditions. The desired probability is a weighted average of the various conditional probabilities.

For example, suppose you are in a chess tournament and will play your next game against one of several opponents: A, B, C,..... The event W represents winning (W) can be split up as $W = (W \cap A) \cup (W \cap B) \cup (W \cap C)$..... Because the events on the right hand side are disjoint, the $P(W) = P(W \cap A) + (W \cap B) + (W \cap C) +$ ..... and so on. But then, we can apply the general product rule to obtain

$$P(W) = P(W \mid A)P(A) + (W \mid B)P(B) + (W \mid C)P(C) + ...$$

Example 6. You are in the chess tournament and will play your next game against either Joe or Mary, depending on results of some other games. Suppose your probability of beating Joe is $7/10$, but of beating Mary is only $2/10$. You assess your probability of playing joe as $1/4$. How likely is it that you win your next game?

$$\begin{aligned} P(W) &= P(W \mid J)P(J) + P(W \mid J^c)P(J^c) && (2.1) \\ &= (7/10)(1/4) + (2/10)(3/4) = 0.325 && (2.2) \end{aligned}$$

2.6 Bayes' Rule

Bayes' rule indicates how probabilities change in the light of evidence. So it is the most important tool in statistics, wherein the evidence is usually data. We will use it in every statistics problem we address.

Example 7. Who was your opponent? In an earlier example, we calculated the probability that you would win your next chess game by averaging over your possible opponents. Now suppose you tell me you won your next chess game -the "evidence" mentioned above. Who was your opponent?

Without conditioning on the evidence, you were three times as likely to have played Mary as Joe. Now I learn that you won and I want to find $P(J \mid W)$. Since you are more likely to win playing Joe than playing Mary, it seems reasonable to expect $P(J \mid W)$ to be bigger than $P(J)$. Bayes' rule gives $P(J \mid W)$-and verifies that it is bigger.

The definition of conditional probability says

$$P(J \mid W) = \frac{P(J \cap W)}{P(W)}$$

Using the multiplication rule in the numerator gives

$$P(J \mid W) = \frac{P(W \mid J)P(J)}{P(W)}$$

Using the law of total probability to expand the denominator gives the following important result known as Bayes' rule:

$$P(J \mid W) = \frac{P(W \mid J)P(J)}{P(W \mid J)P(J) + P(W \mid J^c)P(J^c)}$$

This applies to any events W and J in any setting. In our example,

$$P(J \mid W) = \frac{(7/10)(1/4)}{(7/10)(1/4) + (2/10)(3/4)}$$

Bayes' rule relates inverse probabilities, giving $P(J \mid W)$ in terms of $P(W \mid J)$.

The conventional terminology for $P(J \mid W)$ is **the posterior probability of J given W** and for $P(J)$ is the **prior probability of J** since it applies *before* or not conditionally on the information that W occurred.

An equivalent expression for Bayes' theorem is

$$\frac{P(J \mid W)}{P(J^c \mid W)} = \frac{P(W \mid J)}{P(W \mid J^c)} \frac{P(J)}{P(J^c)}$$

The ratio $P(J \mid W)/P(J^c \mid W)$ is the posterior odds in favor of J. The second factor on the right, $P(J)/P(J^c)$, is the prior odds in favor of J. $P(W \mid J)/P(W \mid J^c)$ is the Bayes factor in favor of J.

Bayes factors are ratios of probabilities of the information at hand (W in our example....). These probabilities are called **likelihoods**. For example, $P(W \mid J)$ is the likelihood of J (for observation W).

Def: The likelihood of a model is the probability of the observations assuming the model.

The unconditional probability P(W) is the denominator of Bayes' theorem but to a Bayesian this is an unimportant probability statement since W has already been observed and therefore has probability one of occurrence. So there is no point in treating W probabilistically if the actual facts are sitting on our desk right now. So the only purpose for P(W) in this context is to make sure that $P(J \mid W)$ sums or integrates to one.

This last discussion suggests simply treating $P(W)$ as a normalizing constant since it does not change the relative probabilities for J. Maybe this is a big conceptual leap, but if we could recover unconditional $P(W)$ later, it is convenient to just use it then to make the conditional statement, $P(J \mid W)$ a properly scaled probability statement. So if $P(J \mid W)$ summed or integrated to five instead of one, we would simply divide everywhere by five and lose nothing but the agony of carrying $P(W)$ through the calculations. If we temporarily ignore $P(W)$, then:

$$P(J \mid W) \propto P(J)P(W \mid J)$$

So the final estimated probability of interest in the left hand side is a balance between things we have already seen or believe, $P(J)$, and the contribution from the new observation, $P(W \mid J)$.

As described earlier, this is an ideal paradigm for inference in the social and behavioral sciences, as well as the physical sciences.

Example 8. Inference about a genetic probability. 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 disease on only one of her two 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 low in human populations.

The prior distribution

Consider a woman that has an affected brother, which implies that her mother must be a carrier of the hemophilia gene with one "good" and one "bad" hemophilia gene. We are also told that her father is not affected; thus the woman herself has a fifty-fifty chance of having the gene. The unknown quantity of interest, the state of the woman, has just two values: the woman is either a carrier of the gene ($\theta = 1$) or not ($\theta = 0$). Based on the information provided thus far, the prior distribution for the unknown θ can be expressed simply as $Pr(\theta = 1) = Pr(\theta = 0) = \frac{1}{2}$.

The model and likelihood

The data used to update this prior information consists of the affection status of the woman's sons. Suppose she has two sons, neither of whom is affected. Let $y_i = 1$ or 0 denote an affected or unaffected son, respectively. The outcomes of the two sons are exchangeable and, conditional on the unknown θ are independent; we assume the sons are not identical twins. The two items of independent data generate the following likelihood function:

$$P(y_1 = 0, y_2 = 0|\theta = 1) = (0.5)(0.5) = 0.25$$

$$P(y_1 = 0, y_2 = 0) = (1)(1) = 1$$

These expressions follow from the fact that if the woman is a carrier, then each of her sons will have a 50% chance or inheriting the gene and so being affected, whereas if she is not a carrier then there is a probability very close to 1 that a son of hers will be unaffected. (In fact, there is a nonzero probability of being affected even if the mother is not a carrier, but this risk –the mutation rate– is very small and can be ignored for this example.

The posterior distribution

Bayes rule can now be used to combine the information in the data with the prior probability, in particular, interest is likely to focus on the posterior probability that the woman is a carrier. Using y to denote the joint data (y_1, y_2), this is simply

$$Pr(\theta = 1|y) = \frac{p(y|\theta = 1)Pr(\theta = 1)}{p(y|\theta = 1)Pr(\theta = 1) + p(y|\theta = 0)Pr(\theta = 0)} = \frac{(0.25)(0.5)}{(0.25)(0.5) + 1(0.5)} = \frac{0.125}{0.625} = 0.2$$

Intuitively, it is clear that if a woman has unaffected children, it is less probable that she is a carrier, and Bayes' rule provides a formal mechanism for determining the extent of the correction. The results can also be described in terms of prior and posterior odds. The prior odds of the woman being a carrier are $\frac{0.5}{0.5} = 1$. The likelihood ratio based on the information about her two unaffected sons is $\frac{0.25}{1} = 0.25$, so the posterior odds are obtained very simply as 0.25. Converting back to a probability, we obtain $\frac{0.25}{1+0.25} = 0.2$, just as before.

Adding more data

A key aspect of Bayesian analysis is the ease with which sequential analyses can be performed. For example, suppose that the woman has a third son, who is also unaffected. The entire calculation does not need to be redone; rather we use the previous posterior distribution as the new prior distribution, to obtain

$$Pr(\theta = 1|y_1, y_2, y_3) = \frac{(0.5)(0.2)}{(0.5)(0.2) + (1)(0.8)}$$

Alternatively, if we suppose that the third son is affected, it is easy to check that the posterior probability of the woman being a carrier becomes 1 (again ignoring the possibility of a mutation).

2.7 Statistical Models with Bayes's theorem

The statistical role of the quantities seen earlier has not yet been identified, since we have been talking abstractly about "events" rather than conventional data. The goal of inference is to make claims about unknown quantities using data currently in hand and prior information. We designate a generic Greek character to denote an unknown that is the objective of our analysis. As is typical in these endeavors, we will use θ for this purpose. What we usually have available to us is generally (and perhaps a little vaguely) labeled D for data and some prior information about θ. Therefore, the objective is to obtain a probabilistic statement about θ given D and the prior information: $p(\theta \mid D)$.

Inferences are made by first specifying a parameter model for the data generating process. This defines what the data should be expected to look like given a specific probabilistic function conditional on unknown variable values. These are the common probability density functions (continuous data) and probability mass functions (discrete data) such as normal, binomial, chi-square, etc. denoted by $p(D \mid \theta)$.

Now we can relate these two conditional probabilities $p(\theta \mid D) \propto p(\theta)p(D \mid \theta)$ where $p(\theta)$ is a formalized statement of the prior knowledge about θ before observing the data. If we know little, then this prior distribution should be a vague probabilistic statement and if we know a lot then this should be a very narrow and specific claim. The right-hand side of the equation above implies that the post-data,post-prior information inference for θ is a compromise between prior information and the information provided by the new data, and the left-hand side is the posterior distribution of θ since it provides the updated distribution of θ after conditioning on the data.

Bayesians describe $p(\theta \mid D)$ to readers via distributional summaries such as means, modes, quantiles, probabilities over regions, traditional-level probability intervals, and graphical displays. Once the posterior distribution has been calculated, everything about it is known and it is entirely up to the researcher to highlight features of interest. Often it is convenient to report the posterior mean and variance in papers and reports by default. We can calculate the posterior mean using an expected value calculation, confining ourselves here to the continuous case:

$$E[\theta \mid D] = \int \theta p(\theta \mid D) d\theta$$

and the posterior variance via a similar process

$$Var[\theta \mid D] = E\left[(\theta - E[\theta \mid D])^2 \mid D\right] = \int (\theta - E[\theta \mid D])^2 \, p(\theta \mid D) d\theta$$

It is more interesting to obtain other summaries of the posterior, such as $p(\theta > value)$ or the expectation of some function of θ. Seldom, except in a few special cases, we will be able to find those integrals exactly.

2.8 Concepts from Distribution Theory

Bayesian inference relies heavily on probability theory and, in particular, distributional theory. This section provides a review of basic distributional theory with examples designed to be relevant to Bayesian applications.

2.8.1 Discrete random variables and their distributions

A basic starting point for probability theory is a *discrete random variable*, X. X can take on a countable number of values, each with some probability. The classic example would be a Bernoulli random variable, where X takes the value 1 with probability p and 0 with probability $1 - p$. X denotes some event such as whether a company will sell a product tomorrow. p represents the probability of a sale. We can easily extend this example to the number of units sold tomorrow. Then X is still discrete but take on the values $0, 1, 2 \cdots, m$ with probabilities, $p_1, p_2, \cdots, p_m$. X now has a nontrivial probability distribution. With knowledge of this distribution, we can answer any question such as the probability that there will be at least one sale tomorrow, the probability that there will be between 1 and 10 sales, etc. In general, we can compute the probability that sales will be in any set simply by summing over the probabilities of the elements in the set:

$$Pr(X \in A) = \sum_{x \in A} p_x$$

. We can also compute the *expectation* of the number of units sold tomorrow as the average over the probability distribution.

$$E[X] = \sum_{i=0}^{m} ip_i$$

2.8.2 Continuous random variables

If we are looking at aggregate sales of a popular consumer product, we might approximate sales as a continuous random variable which can take on any nonnegative real number. For this situation, we must summarize the probability distribution of X by a probability density. A density function is a *rate* function which tells us the probability per volume of unit of X. X has a density function $f(X)$; $f(X)$ is a positive-valued function which integrates to one. To find the probability that X takes on any set of values we must integrate $f(X)$ over this set:

$$Pr(X \in A) = \int_A f(x \mid \theta)dx$$

This is very much the analog of the discrete sum. We can easily find the expectation of any function of X by computing the appropriate integral:

$$E[h(X)] = \int (h(x)f(x)dx$$

2.8.3 Joint distributions of two random variables

In many situations, we will want to consider the joint distribution of two or more random variables, both of which are continuous. For example, we might consider the joint distribution of sales tomorrow in two different markets. Let X denote the sales in market A and Y denote the sales in market B. For this situation, there is a bivariate density function $f(X, Y)$ This density gives the probability rate per unit of area in the plane. With the joint density, we compute the probability of any set of (X, Y) values. For example, we can compute the probability that both X and Y are positive. This is the area under the density for the positive quadrant:

$$Pr(X > 0, Y > 0) = \int_0^\infty \int_0^\infty f(x, y)dxdy.$$

Marginal distributions obtained from the joint distribution

Given the joint density, we can also compute the marginal densities of each of the variables X and Y. That is to say, if we know everything about the joint distribution, we certainly know everything about the marginal distribution. The way to think of this is via simulation. Suppose we were able to simulate from the joint distribution. If we look at the simulated distribution of either X or Y alone, we have simulated the marginal distribution.

To find the marginal density of X, we must average the *joint* density over all possible values of Y :

$$f(X) = \int f(x, y)dy$$

A simple example will make this idea clear.

Example 9. Suppose X, Y are uniformly distributed over the triangle $\{X, Y : 0 < Y < 1 \text{ and } Y < X < 1\}$. A uniform distribution means that the density is constant over the shaded triangle. The area of this triangle is 1/2, so this means that the density must be 2 in order to endure that the joint density integrates to 1:

$$\int_0^1 \int_0^1 f(x, y)dxdy = \int_0^1 \int_0^1 2dxdy = 1$$

The marginal density of X is

$$f(X) = \int_0^1 f(x, y)dy = 2x$$

Thus the marginal distribution of X is not uniform! The marginal density of Y can be found to be $f(y) = 2 - 2y$.

Conditional distributions obtained from the joint distribution

We can also define the concept of a *conditional* distribution or density. For the continuous case,

$$f(Y \mid X) = \frac{f(X,Y)}{f(X)}$$

There is a different density for every value of Y. We note that the conditional density is proportional to the joint. The marginal only serves to get the right normalization.

Example 10. Returning to our last example,

$$f(Y \mid X) = \frac{2}{2x}, \qquad y \in (0,x).$$

Thus, if $x = 1$, then the density is uniform over (0,1) with height 1. The dependence betwen X and Y is only evidenced by the fact that the range of Y is restricted by the value of x.

2.9 List of distributions

We will be using many probability distributions throughout the course. They will be introduced one by one in their corresponding lecture. You will be given a handout in lecture, separately, with a list of the most common distributions used in Bayesian analysis.

This handout will be helpful in the homework, where you will have to simulate random values from these distributions to start getting familiarized with the notion of sampling from probability distributions, which is a key component of the computational methods we will use.

2.10 Drawing random numbers from some of these distributions

Most of the simulation methods that we will use rely on random sampling from the posterior distribution. Hence, it is important that you become familiar with some of these distributions, how we summarize them and how we plot them. Please, be in lecture when we do that.

2.11 Additional Required Reading

Chapter 2 of the required textbook (Hoff's)