Chapter 1

An introduction to Bayesian Methods

1.1 The Scientific Method

The scientific method is a process of devising studies and updating knowledge using evidence from these studies. It has several steps:

(1) Ask a question

(2) Assemble and evaluate the relevant information

(3) Based on current information, design an investigation or experiment (or perhaps no experiment) to address the question posed in step 1. Consider costs and benefits of any investigation, including the value of any information they may contain. Recognize 6 is coming.

(4) Carry out the investigation or experiment

(5) Use the evidence from step 4 to update the previously available information: draw conclusions, if only tentative ones.

(6) Repeat steps 3 through 5 as necessary.

The scientist formulates the question. This step is not really statistical, but an effective use of statistics can help show that a question being addressed is inappropriate or impossible to answer. *The two principal parts of the learning process are designing studies or experiments* (what data should be collected and how should they be collected?) and *analyzing data from experiments* (how do we learn from the data collected?). The second part is the main focus of this course. We need a formalism for learning, for incorporating the results of studies or experiments into what we already know. This is step 5 of the scientific method. Such a formalism should allow for describing knowledge available to us at any time. Knowledge can then be converted into inferences, decisions and designs for additional studies. This course uses the *Bayesian formalism*. There are no other approaches which can provide a unified treatment for combining all available information.

1.2 Components of Statistical Inference

The goal of statistical inference is to use**information** available to make **inferences** about **unknown quantities** in a **population**.

1.2.1 Information

One important source of information is **data**, but there is an undeniable role for **non-data-based information**.

Data

Data is comprised of observations. Observations made in a specified way (for example, carrying out a particular experiment repeatedly) form a sample.

A sample is a collection of observations

The simplest observation is made on an experimental or observational unit (depending on the study design). Examples of units are plots of land, patients, petri dishes, households, ants, engines, and so on. Identifying the experimental unit or unit of observation is important in drawing conclusions from studies and extrapolating to some larger population of units.

Sometimes there is more than one type of experimental unit. For example, to judge customer satisfaction with a product, a company may select 10 grocery stores and survey 20 customers at each store. Responses may vary from store to store and they may vary from customer to customer within the stores. At one level, the experimental unit is the store, but at another level, it is the customer. [1]

The sample should be tied in some way to the question that is being addressed (step 1 of the scientific method).

The question must deal with the process that produces the observations. The set of all possible observations (real or imagined) is called a population. A population may be finite or infinite.

A population is the collection of potential observations of which the sample is a part

Other Information

We have concluded that the information available can be be partitioned into information obtained from the data as well as other information obtained independently or prior to the data.

Information can also come from theories of behavior, "subjective views" that there is a structure underlying unknowns, expectations that quantities take particular range of values, prior analyses of other data, including data that is only loosely related to the data set under investigation.

1.2.2 Unknown quantities

Unknown quantity is a generic term referring to any value not known to the investigator. Certainly, parameters of probability distributions can be considered unknown since these are purely abstractions that index a class of models. Future values and missing observations are also unknown.

1.2.3 Study design

Most scientific questions are relative, or comparative. "How big is it?" Compared with what? "How effective is it?" Compared with what? Controls are essential for interpreting results. Designing experiments plays a fundamental role in the scientific method.

There are two principal types of studies, randomized and observational. Randomization minimizes bias in treatment assignment. While observational studies are problematic, they are essential in science.

In a randomized study the researcher assigns treatment using a randomization device. In an observational study, a researcher observes differences between two groups-treatment and control- but does not assign treatment. When subjects choose their own treatments, comparisons are inherently flawed. Self-selected samples are practically worthless for scientific inference.

Showing relationships is not the same a showing cause and effect.

A low rate of response in surveys can weaken and even invalidate conclusions.

Although it is not our goal to focus on study design, the interpretation of our results will be affected by how the data is collected, so, indirectly, the study design will matter to us and we will discuss it.

[1] Wide application of hierarchical linear models is one of the major success stories of modern Bayesian statistics since the late eighties. This was substantially driven by the invigorating breakthroughs in Bayesian computation via Markov chain simulation. Hierarchical models will allow us to address, later in the course, these situations with more than one experimental unit.

1.3 Bayesian Statistical inference

Bayesian inference utilizes probability statements as the basis for inference. What this means is that our goal is to make probability statements about unknown quantities based on the sample and prior information.

A particular type of inference is making conclusions or predictions about the next observation from the population. Predictions are not usually specific, such as "the next observation is a 7." Rather, they are subject to error.

A prediction is an inference about the next sample observation or set of observations.

Bayesian methods have become common in advanced training and research in statistics, but elementary training appears to have lagged, despite arguments for reform by many statisticians. Observational researchers (not just statisticians) need training in subjective Bayesianism. Bayesian approaches make explicit those subjective and arbitrary elements that are shared by all statistical inferences.

A critical aspect of Bayesian modeling is that all relevant uncertainties are represented by probability distributions

Denote the set of unknowns as θ. Our prior beliefs based on available information are expressed as probability distributions, $p(\theta)$. This is called the *prior*. In most cases, this represents a density function, but it can also represent at probability mass function for discrete parameter spaces, or a mixed continuous-discrete distribution. The information provided by the data is introduced via the probability distribution for the data, $p(D \mid \theta)$, where D denotes the observable data. This distribution is usually called the "likelihood function." Given D, any function which is proportional to $p(D \mid \theta)$ is called the *likelihood*, $l(\theta)$. Modeling is the art of choosing appropriate models for the prior and for the data. To deliver on the goal of inference, we must combine the likelihood and prior to produce the distribution of the unknowns conditional on the data and the prior. Bayes' theorem gives

$$p(\theta \mid D) = \frac{p(D, \theta)}{p(D)} = \frac{p(D \mid \theta)p(\theta)}{p(D)}$$

$p(\theta \mid D)$ is called the *posterior* distribution and reflects the combined data and prior information. Bayes theorem is often expressed using the likelihood function. The shape of the posterior distribution is determined entirely by the likelihood and prior in the numerator, and this is often emphasized by rewriting the equation:

$$p(\theta \mid D) \propto l(\theta)p(\theta). \tag{1.1}$$ {post1}

If $l(\theta) = p(D \mid \theta)$, then the constant of proportionality is the marginal distribution of the data $p(D) = \int p(D, \theta)d\theta$. Of course, we are assuming here that this normalizing constant exists. If $p(\theta)$ represents a proper distribution (that is, it integrates to 1)), then it exists.

Bayesian analyses need not be limited to using a single prior or likelihood function. Acceptability of an analysis is often enhanced by presenting results from different priors and different likelihood functions.

The randomness in the unknowns, as reflected in the prior distribution, represents personal uncertainty. It is not a property of the unknowns. Statistical results in a Bayesian analysis depend as much on the model chosen for the data as on the prior. These models represent our uncertainty, not any characteristic of the phenomenon under study. They are not "laws." Probabilities are nothing more than expressions of opinions, as in common phrasings such as "It will probably rain tomorrow."

There are many varieties of Bayesian analysis. The fullest version of the Bayesian paradigm casts statistical problems in the framework of decision making. It entails formulating subjective prior probabilities to express pre-existing information, careful modelling of the data structure, checking and allowing for uncertainty in model assumptions, formulating a set of possible decisions and a utility function to express how the value of each alternative decision is affected by the unknown model parameters. But each of these components can be omitted. Many users of Bayesian methods do not employ genuine prior information, either because it is insubstantial or because they are uncomfortable with subjectivity. The decision-theoretic framework is also widely omitted, with many feeling that statistical inference should not really be formulated as a decision. So there are varieties of Bayesian analysis and varieties of Bayesian analysts. But the common strand that underlies this variation is the basic principle of using Bayes' theorem and expressing uncertainty about unknown unknowns probabilistically.

1.3.1 Prediction and Bayes

One of the appeals of the Bayesian approach is that all unknowns are treated the same. Prediction is defined as making probability statements about the distribution of as yet unobserved data, denoted by D_f. The only real distinction between "other unknowns" and unobserved data D_f is that D_f is potentially observable:

$$P(D_f \mid D) = \int p(D_f, \theta \mid D) d\theta = \int p(D_f \mid \theta, D) p(\theta \mid D) d\theta$$

This last equation defines the predictive distribution of D_f given the observed data. In many cases, we assume that D and D_f are independent, conditional on θ. In this case, the predictive distribution simplifies to

$$p(D_f \mid D) = \int p(D_f \mid \theta) p(\theta \mid D) d\theta$$

In this last equation, we average the likelihood for the unobserved data over the posterior for θ. This averaging properly accounts for uncertainty in θ when forming predictive statements about D_f.

1.3.2 Summarizing the Posterior

For any problem of practical interest, the posterior distribution is a high-dimensional object. Therefore, summaries of the posterior play an important role in Bayesian statistics. Reporting moments of the marginal distributions of unknowns such as the posterior mean and standard deviation of a distribution's parameters is common practice. It is far more useful and informative to produce the marginal distribution of parameters or relevant functions of parameters as the output of the analysis. Simulation methods are ideally suited for this. If we can simulate from the posterior distribution of the parameters and other unknowns, then we can simply construct the marginal of any function of interest. Typically, we describe these marginals graphically. As these distributions are often nonnormal, the mean and standard deviations are not particularly useful. One of the purposes of this course is to introduce a set of tools to achieve this goal of simulating from posterior distributions.

Prior to the advent of powerful simulation methods, attention focused on the evaluation of specific integrals of the posterior distributions as a way of summarizing this high-dimensional object. For many years, only problems for which the integrals could be performed analytically were analyzed by Bayesians. Obviously, this restricts the set of priors and likelihoods to a very small set that produces posteriors of known distributional form and for which these integrals can be evaluated analytically.

Until the mid-1980s, Bayesian methods appeared to be impractical since the class of models for which the posterior inference could be computed was no larger than the class of models for which exact sampling results were available. Moreover, the Bayes approach does require assessment of a prior, which some feel to be an extra cost. Simulation methods, in particular Markov chain Monte Carlo (MCMC) methods, have freed us from computational constraints for a very wide class of models. MCMC methods are ideally suited for models built from a sequence of conditional distributions, often called hierarchical models. Bayesian hierarchical models offer tremendous flexibility and modularity and are particularly useful for marketing, bioinformatics, engineering, education, economics, etc...

1.3.3 Example: inference for a proportion with a discrete prior

Think of a population of college students and let p represent the (unknown) proportion of students that sleep at least 8 hours. We are interested in learning about the location of p.

In the Bayesian viewpoint, a person's beliefs about the uncertainty in this proportion are represented by a probability distribution placed on this parameter and called the prior probability. This distribution reflects the person's subjective prior opinion about plausible values of p. Based on some surveys read, let's say we believe that college students generally get less than eight hours of sleep and so p, the proportion that sleep at least 8 hours is likely smaller than 0.5. After some reflection, her best guess at the value of p is 0.3. But it is very plausible that this proportion could be any value in the interval from 0 to 0.5.

To see if this prior opinion about p is supported by information, a study is designed to obtain data (D). A random sample of 27 students is taken –in this group, 11 record that they had at least eight hours of sleep the previous night.

Based on the prior information and this observed data, we are interested in estimating the proportion p. In addition, we are interested in predicting the number of students that get at least eight hours of sleep if a new sample of 20 students is taken.

Suppose that our prior density for p is denoted by $g(p$. If we regard a "success" as sleeping at least eight hours and we take a random sample with s successes and f failures, then the likelihoof function is given by

$$l(D \mid p) = p^s(a-p)^f \qquad 0 < p < 1$$

The posterior density for p, by Bayes' rule, is obtained, up to a proportionality constant, by multiplying the prior density by the likelihood.

$$g(p|D) \propto g(p)l(D \mid p)$$

1.4 Using a discrete prior

If we believe

$$0.05,\ \ 0.15\ \ 0.25\ \ 0.35\ \ 0.45\ \ 0.55\ \ 0.65\ \ 0.75\ \ 0.85\ \ 0.95$$

are possible values for p, and we assign the following weights to these values

$$2, 4, 8, 8, 4, 2, 1, 1, 1, 1$$

we can convert both to prior probabilities by dividing each weight by the sum. In *R*, we define p to be the vector of proportion values and *prior* the corresponding weights that we normalize to probabilities. The *plot* command is used with the "histogram" type option to graph the prior distribution, and Figure 1.1 left-hand side displays this graph.

```
 p = seq(0.05, 0.95, by=0.1)
prior.weights= c(2, 4, 8, 8, 4, 2, 1, 1, 1, 1)
 prior.prob=prior.weights/sum(prior.weights)
plot(p, prior.prob, type="h", ylab="Prior probability")
```

In our example, 11 of 27 students sleep at least eight hours, so $s = 11$ and $f = 16$, and the likelihood function is

$$l(p) = p^{11}(1-p)^{16}, \qquad 0 < p < 1$$

Note that the likelihood is a beta density with parameters $s + 1 = 12$ and $f + 1 = 17$. We will take advantage of this later.

We will now compute the posterior probabilities. One inputs the vector of proportion values p, the vector of prior probabilities `prior` and s and f. The output is a vector of posterior probabilities. The `cbind` command is used to display a table of the prior and posterior probabilities, and Figure 1.2 displays a line graph of the posterior probabilities.

```
s=11; f=16
 likelihood=(p^s)*((1-p)^f)      # this is l
 prop.post= likelihood*prior.prob         # this is proportional to g(p | D)
 post=(prop.post/sum(prop.post))  # this is g(p|D)
 cbind(p, prior.prob, post)     # show p, prior and post in one single table
 par(mfrow=c(1,3))
  plot(p, post, type="h", ylab="Posterior Probability -black line",
  main="Prior (dash-red) and posterior (black-solid) probabilities")
   lines(p,prior.prob,type="h",lty=2,col="red")
```

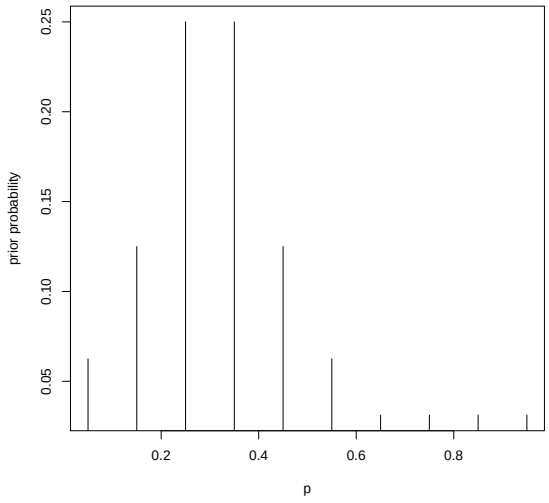

Figure 1.1: Discrete prior distribution for a proportion p {fig:sleeppric

Here, in the R output given below and in Figure we note that most of the posterior probability is concentrated on the values $p = 0.35$ and $p = 0.45$. If we combine the probabilities for the three most likely values, we can say the posterior probability that p falls in the set $\{0.25, 0.35, 0.45\}$ is equal to 0.942.

```
        p         prior.prob              post
 [1,] 0.05    0.06250       2.882642e-08
 [2,] 0.15    0.12500       1.722978e-03
 [3,] 0.25    0.25000       1.282104e-01
 [4,] 0.35    0.25000       5.259751e-01
 [5,] 0.45    0.12500       2.882131e-01
 [6,] 0.55    0.06250       5.283635e-02
 [7,] 0.65    0.03125       2.976107e-03
 [8,] 0.75    0.03125       6.595185e-05
 [9,] 0.85    0.03125       7.371932e-08
[10,] 0.95   0.03125       5.820934e-15
```

1.5 Additional examples

For this section, read Hoff's textbook [**?**, Section 1.2.1]. In this course, we do not talk about non-Bayesian methods. Hence, skip the subsections "Comparison to non-Bayesian methods," "General estimation of a population mean."

Read also section 1.2.2. Skip the paragraph on page 10 starting with "How does this estimate of beta compare . . . " As indicated above, we do not compare Bayesian methods to other methods.

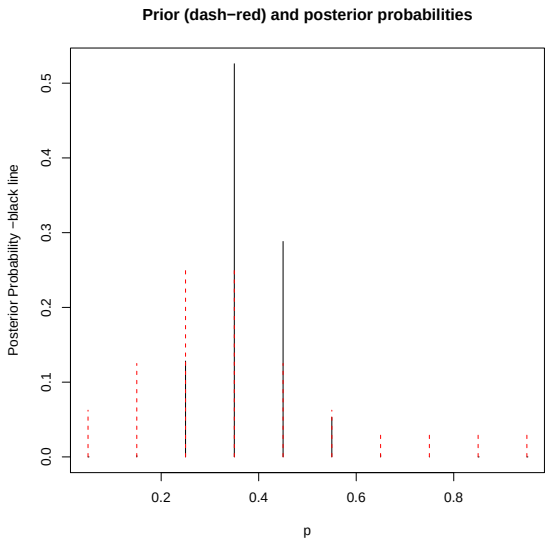

Figure 1.2: Discrete prior distribution for a proportion p

leepposterior}