Stat 136 (Bayesian Statistics)

Lesson 1.1: Introduction to Bayesian Statistics

Main Approaches to Statistics

There are two main philosophical approaches to statistics. The first is often referred to as the **frequentist** approach. Sometimes it is called the **classical** approach. Procedures are developed by looking at how they perform over all possible random samples. The probabilities do not relate to the particular random sample that was obtained. In many ways this indirect method places the **cart before the horse**.

The alternative approach that we take in this course is the **Bayesian** approach. It applies the laws of probability directly to the problem. This offers many fundamental advantages over the more commonly used frequentist approach.

Frequentist Approach to Statistics

The frequentist approach to statistics is based on the following ideas:

- Parameters, the numerical characteristics of the population, are fixed but unknown constants.
- Probabilities are always interpreted as long-run relative frequency.
- Statistical procedures are judged by how well they perform in the long run over an infinite number of hypothetical repetitions of the experiment.

Probability statements are only allowed for random quantities. The unknown parameters are fixed, not random, so probability statements cannot be made about their value. Instead, a sample is drawn from the population, and a sample statistic is calculated. The probability distribution of the statistic over all possible random samples from the population is determined and is known as the sampling distribution of the statistic. A parameter of the population will also be a parameter of the sampling distribution. The probability statement that can be made about the statistic based on its sampling distribution is converted to a confidence statement about the parameter. The confidence is based on the average behavior of the procedure over all possible samples.

Bayesian Approach to Statistics

Bayesian approach to statistics put forward the ideas that:

- Since we are uncertain about the true value of the parameters, we will consider them to be random variables.
- The rules of probability are used directly to make inferences about the parameters.
- Probability statements about parameters must be interpreted as *degree of belief.* The *prior* distribution must be subjective. Each person can have his/her own prior, which contains the relative weights that person gives to every possible parameter value. It measures how *plausible* the person considers each parameter value to be before observing the data.
- We revise our beliefs about parameters after getting the data by using Bayes' theorem. This gives our posterior distribution which gives the relative weights we give to each parameter value after analyzing the data. The posterior distribution comes from two sources: the prior distribution and the observed data.

This has a number of advantages over the conventional frequentist approach. Bayes' theorem is the only consistent way to modify our beliefs about the parameters given the data that actually occurred. This means that the inference is based on the actual occurring data, not all possible data sets that might have occurred but did not! Allowing the parameter to be a random variable allows us make probability statements about it, posterior to the data.

This contrasts with the conventional approach where inference probabilities are based on all possible data sets that could have occurred for the fixed parameter value. Given the actual data, there is nothing random left with a fixed parameter value, so one can only make confidence statements, based on what could have occurred.

Bayesian statistics also has a general way of dealing with a nuisance parameter. A nuisance parameter is one which we do not want to make inference about, but we do not want them to interfere with the inferences we are making about the main parameters. Frequentist statistics does not have a general procedure for dealing with them. Bayesian statistics is predictive, unlike conventional frequentist statistics. This means that we can easily find the conditional probability distribution of the next observation given the sample data.

Why Bayesian statistics?

- Long history: named after the 18^{th} century Presbyterian minister and mathematician Thomas Bayes (1701 - 1761).

- Modeling: incorporate prior belief or domain experts knowledge.
- Theoretical: doesn't need large sample assumption.
- Computational: Markov chain Monte Carlo (MCMC).

Bayesian approaches are largely popularized by revolutionary advance in computational technology during the last twenty five years (the invention of Gibbs sampler around 1990 - we will read a research paper about it later).

The Bayes Theorem: The Key to Bayesian Statistics

Review: Conditional Probability and Law of Total Probability

Recall that by definition, the conditional probability of A given B, is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{provided } P(B) > 0. \tag{1}$$

Similarly, we have

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \text{provided } P(A) > 0. \tag{2}$$

By multiplication, we obtain from (2),

$$P(A \cap B) = P(B|A) \times P(A) \tag{3}$$

Recall further that the total probability, $P(B)$, can be computed as,

$$P(B) = P(B|A) \times P(A) + P(B|\overline{A}) \times P(\overline{A}) \tag{4}$$

where: $P(\overline{A}) = 1 - P(A)$.

The Bayes Theorem

Applying (3) and (4) to (1) gives rise to the Bayes formula

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B|A) \times P(A) + P(B|\overline{A}) \times P(\overline{A})} \tag{5}$$

As regards terminology, we call $P(A)$ the **prior** probability of A (meaning the probability of A before B is known to have occurred), and we call $P(A|B)$ the **posterior** probability of A given B (meaning the probability of A after B is known to have occurred).

We may also say that $P(A)$ represents our a *priori* beliefs regarding A, and $P(A|B)$ represents our a *posteriori* beliefs regarding A.

The above result can be easily extended to more than 2 partitions of set B. Let $A_1, A_2, \cdots, A_k$ forms a partition of set B. Then for any $i = 1, 2, \cdots, k$

$$\begin{aligned} P(A_i|B) &= \frac{P(A_i \cap B)}{P(B)} \\ &= \frac{P(B|A_i) \times P(A_i)}{\sum_{j=1}^{k} P(B|A_j) \times P(A_j)} \end{aligned} \tag{6}$$

Example 1.1.1

The incidence of a disease in the population is 1%. A medical test for the disease is 90% accurate in the sense that it produces a false reading 10% of the time, both: (a) when the test is applied to a person with the disease; and (b) when the test is applied to a person without the disease. A person is randomly selected from population and given the test. The test result is positive (i.e. it indicates that the person has the disease). What is the probability that the person actually has the disease?

SOLUTION:

Let A be the event that the person has the disease, and let B be the event that they test positive for the disease. Then:

- $P(A) = 0.01$ (the prior probability of the person having the disease)
- $P(B|A) = 0.9$ (the **true positive** rate, also called the **sensitivity** of the test)
- $P(\overline{B}|\overline{A}) = 0.9$ (the **true negative** rate, also called the **specificity** of the test)

So,

$P(A \cap B) = P(A) \times P(B|A) = 0.01 \times 0.9 = 0.009$ and

$P(\overline{A} \cap B) = P(\overline{A}) \times P(B|\overline{A}) = 0.99 \times 0.1 = 0.099.$

Thus, the unconditional (or *prior*) probability of the person testing positive is

$$P(B) = P(A) \times P(B|A) + P(\overline{A}) \times P(B|\overline{A}) = 0.009 + 0.099 = 0.108$$

So the required posterior probability of the person having the disease is

$$\begin{aligned} P(A|B) &= \frac{P(A) \times P(B|A)}{P(A) \times P(B|A) + P(\overline{A}) \times P(B|\overline{A})} \\ &= \frac{0.009}{0.108} \\ &\approx 0.083 \end{aligned}$$

Example 1.1.2

In a particular population:

- 10% of persons have Type 1 blood, and of these, 2% have a particular disease;
- 30% of persons have Type 2 blood, and of these, 4% have the disease;
- 60% of persons have Type 3 blood, and of these, 3% have the disease.

A person is randomly selected from the population and found to have the disease. What is the probability that this person has Type 3 blood?

SOLUTION:

Let

- A = ‘The person has Type 1 blood’
- B = ‘The person has Type 2 blood’
- C = ‘The person has Type 3 blood’
- D = ‘The person has the disease’

Then,

- $P(A) = 0.1, P(B) = 0.3, P(C) = 0.6$
- $P(D|A) = 0.02, P(D|B) = 0.04, P(D|C) = 0.03$

By the law of total probability we have

$$\begin{aligned} P(D) &= P(D|A) \times P(A) + P(D|B) \times P(B) + P(D|C) \times P(C) \\ &= 0.02 \times 0.1 + 0.04 \times 0.3 + 0.03 \times 0.6 \\ &= 0.032 \end{aligned}$$

Therefore,

$$\begin{aligned} P(C|D) &= \frac{P(D|C) \times P(C)}{P(D)} \\ &= \frac{0.03 \times 0.6}{0.032} \\ &= 0.5625 \end{aligned}$$

Elements of a Bayesian model

Bayes' formula extends naturally to statistical models. A Bayesian model is a parametric model in the classical (or frequentist) sense, but with the addition of a prior probability distribution for the model parameter, which is treated as a random variable rather than an unknown constant. The basic elements of a Bayesian model may be listed as:

1. The **parameter** of interest, say θ. Note that this is completely general, since θ may be vector valued. So θ might be a binomial parameter, or the mean and variance of a Normal distribution, or an odds ratio, or a set of regression coefficients, etc. The parameter of interest is sometimes usefully thought of as the "true state of nature".

2. The **prior distribution** of θ, $f(\theta)$. This prior distribution summarizes what is known about θ before the experiment is carried out. It is "subjective", so may vary from investigator to investigator.

3. The **likelihood function**, $f(y|\theta)$. The likelihood function provides the distribution of the data, y, given the parameter value θ). So it may be the binomial likelihood, a normal likelihood, a likelihood from a regression equation with associated normal residual variance, logistic regression model, etc.

4. The **posterior distribution**, $f(\theta|y)$. The posterior distribution summarizes the information in the data, y, together with the information in the prior distribution, $f(\theta)$. Thus, it summarizes what is known about the parameter of interest θ after the data are collected.

5. **Bayes Theorem**. This theorem relates the above quantities:

$$\text{posterior distribution} = \frac{\text{likelihood of the data} \times \text{prior distribution}}{\text{normalizing constant}}$$

or

$$f(\theta|y) = \frac{f(y|\theta) \times f(\theta)}{f(y)}$$

where:

$$f(y) = \begin{cases} \sum_\theta f(\theta) \times f(y|\theta), & \text{if } \theta \text{ is discrete} \\ \int f(\theta) \times f(y|\theta) d\theta, & \text{if } \theta \text{ is continuous} \end{cases}$$

Ignoring the normalizing constant, we get

$$f(\theta|y) \propto f(y|\theta) \times f(\theta)$$

Thus we "update" the prior distribution to a posterior distribution after seeing the data via Bayes Theorem.

6. The **action**, a. The action is the decision or action that is taken after the analysis is completed. For example, one may decide to treat a patient with *Drug 1* or *Drug 2*, depending on the data collected in a clinical trial. Thus our action will either be to use *Drug 1* (so that $a = 1$) or *Drug 2* (so that $a = 2$).

7. The **loss function**, $L(\theta, a)$. Each time we choose an action, there is some loss we incur, which depends on what the true state of nature is, and what action we decide to take. For example, if the true state of nature is that *Drug 1* is in fact superior to Drug 2, then choosing action $a = 1$ will incur a smaller loss than choosing $a = 2$. Now, the usual problem is that we do not know the true state of nature, we only have data that lets us make probabilistic statements about it (ie, we have a posterior distribution for θ, but do not usually know the exact value of θ). Also, we rarely make decisions before seeing the data, so that in general, $a = a(y)$ is a function of the data. Note that while we will refer to these as "losses", we could equally well use "gains".

8. **Expected Bayes Loss (Bayes Risk)**: We do not know the true value of θ, but we do have a posterior distribution once the data are known, $f(\theta|y)$. Hence, to make a "coherent" Bayesian decision, we minimize the Expected Bayesian Loss, defined by:

$$EBL = \int L(\theta, a(y)) f(\theta|y) d\theta$$

In other words, we choose the action $a(y)$ such that the EBL is minimized.

The first five elements in the above list comprise a non-decision theoretic Bayesian approach to statistical inference. This type of analysis (ie, nondecision theoretic) is what most of us are used to seeing in the medical literature. However, many Bayesians argue that the main reason we carry out any statistical analyses is to help in making decisions, so that elements 6, 7, and 8 are crucial. There is little doubt that we will see more such analyses in the near future, but it remains to be seen how popular the decision theoretic framework will become in medicine. The main problem is to specify the loss functions, since there are so many possible consequences (main outcomes, side effects, costs, etc.) to medical decisions, and it is difficult to combine these into a single loss function. My guess is that much work will have to be done on developing loss functions before the decision theoretic approach becomes mainstream. This course, therefore, will focus on elements 1 through 5.

Steps of Bayesian Data Analysis

1. **Identify/Collect the data** required to answer the research questions. As a general recommendation, it is helpful to visualize the data to get a sense of how the data look, as well as to inspect for any potential anomalies in the data collection.

2. **Choose a statistical model** for the data in relation to the research questions. The model should have good theoretical justification and have parameters that are meaningful for the research questions.

3. **Specify prior distributions** for the model parameters. Although this is a subjective endeavor, the priors chosen should be at least sensible to audience who are skeptical.

4. **Obtain the posterior distributions** for the model parameters. As described below, this can be obtained by analytical or various mathematical approximations. For mathematical approximations, one should check the algorithms for convergence to make sure the results closely mimic the target posterior distributions

5. **Conduct a posterior predictive check** to examine the fit between the model and the data, i.e., whether the chosen model with the estimated parameters generate predictions that deviate form the data being analyzed on important features. If the model does not fit the data, one should go back to step 2 to specify a different model

6. If the fit between the model and the data is deemed satisfactory, one can proceed to **interpret the results** in the context of the research questions. It is also important to **visualize the results** in ways that are meaningful for the analysis.