1 Basics of likelihood based statistics

reading Chap. 1 and 4 of Gelman et al. (1995)

1.1 Conditional probability and Bayes theorem

Traditionally, statisticians have modeled phenomena by supposing there are population parameters which are fixed at certain unknown values, and then tried to estimate (i.e. give a best guess of) these parameters. In this framework, you observe data given that the parameter value takes some unknown fixed value. In contrast, Bayesian statisticians treat all quantities, data and parameters, as random variables. In the Bayesian framework, we envision the following process as responsible for the data we observe: the parameter values are drawn from a probability distribution, then we observe the data conditional on the values of the parameters. The problem of statistical inference is how to convert the observations we have conditional on the parameter values into information about the parameters.

Bayes theorem provides a mathematically rigorous way to convert from the distribution of the data given the parameter values into statements about the parameter values conditional on the data. Let θ represent a parameter in a statistical model (for example the mean of a distribution) and let y represent the data we have observed (in general y may be a vector, i.e. a collection of measurements). We denote the density of θ by $p(\theta)$ and the conditional distribution of θ given y by $p(\theta|y)$. This notation deviates from the usual practice of denoting the density of a random variable by expressions like $p_\theta(x)$ or $p_{\theta|y}(x)$ where x is the argument of the density, but it simplifies the exposition in many respects. Then Bayes theorem is derived by 2 applications of the definition of conditional probability:

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

We call $p(\theta|y)$ the *posterior distribution* of θ given the data y, $p(y|\theta)$ is called the *likelihood* and $p(\theta)$ is called the *prior distribution*. Since $p(\theta|y)$ is a probability density in the argument θ, we know $\int p(\theta|y)\, d\theta = 1$, hence if we have the likelihood and the prior we can always compute $p(y)$ by integrating the product of the likelihood and the prior, setting the result to one and using this to obtain $p(y)$, but in fact we often don't need $p(y)$ at all.

The prior distribution is chosen to reflect any knowledge you may have about the parameters prior to the current data analysis. Often a great deal of research has already been conducted on a topic, so the prior lets the researcher incorporate this knowledge and sharpen the results already obtained, in contrast to simply replicating results already available in the scientific literature. In other contexts, although there is no previous research on a topic, one can supply a prior distribution. As an example, suppose you are trying to locate the chromosome on which a gene which is responsible for some Mendelian trait resides. In this context, if you have no idea where the gene is, your prior distribution could be summarized by saying you think the gene is equally likely to be on any of the chromosomes. More elaborate priors could be constructed by taking into account that not all chromosomes are the same length or have the same number of genes. The first prior is an example of a *noninformative prior* because the prior distribution merely expresses our total ignorance about the chromosome which has the gene. For continuous parameters, like the mean of a normal distribution, we may suppose that prior to seeing the data, the mean is equally likely to be any real number. We can approximate such a prior with the density $p(\theta) = 1/(2k)$ for $-k \leq \theta \leq k$ then let k become arbitrarily large.

1.2 Likelihood based inference

When we use a noninformative prior, our posterior is often proportional to the likelihood, hence we can simply use the likelihood to conduct inference when we do not have useful prior information or are reluctant to attempt to incorporate the prior information in our analysis. One constructs the likelihood for a model with reference to a probability model. The likelihood is the joint distribution of the data given the parameter values, so a probability model supplies this joint distribution. Often we have data which we are willing to model as independent measurements drawn from a common distribution, hence we can find the joint distribution of the data given the parameters by multiplying all of the marginal distributions

$$p(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta).$$

Often Bayesian methods are questioned because of the need to specify the prior distribution, but the real challenge to successful statistical modeling is the intelligent specification of the likelihood. In particular, the assumption of independence is often the most questionable and important assumption made.

While we can always write down a mathematical expression for the posterior distribution given the 2 key ingredients, how to use this expression to draw useful conclusions is not necessarily so straightforward. One idea is to report the mean of the posterior distribution, or find an interval so that there is a specified probability of the parameter falling in this interval. These 2 summaries require integration of the posterior, and generally this integration can not be done analytically. An alternative is to use the posterior *mode* (i.e. the value of θ so that $p(\theta|y)$ is as large as possible) to summarize the posterior distribution. If we think in terms of our data generation model, if we use θ_0 to denote the value of the parameter which is drawn in the initial draw from the parameter distribution, then there is an important result which states that, under a variety of assumptions, the probability that the posterior mode differs from θ_0 by any (small) amount will become zero as the sample size increases. Some of the more important assumptions on which this claim rests are

1. observations are independent, or "almost" independent
2. *identifiability*, i.e. distinct parameter values give rise to different values for the posterior-otherwise the mode is not a well defined object

When the sample size is not large, the posterior mode may not be a useful summary of the posterior distribution. In these cases, one must resort to numerical integration or simulation. When there are only a few parameters, numerical integration is often feasible, but the contemporary approach to these problems usually involves simulation. Simulation provides a very powerful approach to inference since given a set of random draws from the posterior distribution we can construct probability intervals simply by sorting the simulated values. For example, suppose we want a confidence interval for the inverse of a regression coefficient. If we have simulations of the regression coefficient we can invert all these simulated values, sort these and construct probability intervals. This saves us from having to approximate the distribution of a random variable that is defined by inverting a random variable whose distribution is known (using the delta method for example).

1.2.1 Maximum likelihood estimates

Under the 2 conditions listed above (and a number of technical conditions), for large sample sizes, the posterior distribution becomes centered around the posterior mode, denoted $\hat{\theta}$, hence we can Taylor expand

the posterior around the posterior mode (provided this mode isn't on the boundary of the allowed set of θ)

$$\begin{aligned} p(\theta|y) &= \exp\{\log p(\theta|y)\} \\ &= \exp\left\{\log p(\hat{\theta}|y) + (\theta-\hat{\theta})\frac{d}{d\theta}\log p(\theta|y)|_{\theta=\hat{\theta}} + \frac{1}{2}(\theta-\hat{\theta})^2\frac{d^2}{d\theta^2}\log p(\theta|y)|_{\theta=\hat{\theta}}\right\} \\ &= p(\hat{\theta}|y)\exp\left\{-\frac{1}{2}(\theta-\hat{\theta})^2[-\frac{d^2}{d\theta^2}\log p(\theta|y)|_{\theta=\hat{\theta}}]\right\}. \end{aligned}$$

This expansion indicates that the posterior distribution is approximately normal with mean $\hat{\theta}$ and variance $(-\frac{d^2}{d\theta^2}\log p(\theta|y)|_{\theta=\hat{\theta}})^{-1}$. This mean is known as the maximum likelihood estimate (MLE) and the variance is the inverse of a quantity called the *observed Fisher information*, and is often denoted $I(\hat{\theta})$. When we refer to a distribution which holds when the sample size is large, we speak of an *asymptotic distribution*. As the sample size becomes increasingly large, the prior becomes increasingly irrelevant, hence if we have lots of data we don't need to be so concerned about the specification of the prior distribution. Non-Bayesian statistics relies quite heavily on MLEs, and they use the same distribution we have shown is the approximate posterior, but since non-Bayesians typically think of the parameter fixed at some value and make statements about what will happen in repeated sampling from the data distribution, the interpretation of the results differ. In addition, we have argued that the posterior distribution converges pointwise to a normal density, whereas in the usual treatment of MLEs one only argues that the MLE converges in distribution to a normal density. (The latter convergence can be shown to hold under weaker technical conditions on the likelihood.)

1.2.2 Likelihood ratio tests

Hypothesis testing can be thought of as trying to use the data to decide which of 2 competing models, here denoted M_1 and M_2, is more likely. We will here discuss the special situation in which M_2 is a special case of M_1. As an example, suppose M_1 specifies that $y_i \sim \mathrm{N}(\mu, 1)$ and M_2 specifies $y_i \sim \mathrm{N}(0, 1)$, equivalently, we specify the null hypothesis that $\mu = 0$ and want to test this against the 2 sided alternative. Since models often depend on parameters, as in the previous example, it makes sense to treat these models as random objects and enquire about the posterior probability of a model $p(M_1|y)$. If we want to compare 2 models then we should look at the ratio of the posterior probabilities. Using Bayes theorem it is easy to show

$$\frac{p(M_1|y)}{p(M_2|y)} = \frac{p(y|M_1)}{p(y|M_2)}\frac{p(M_1)}{p(M_2)}.$$

This states that the posterior odds $\frac{p(M_1|y)}{p(M_2|y)}$ is equal to the product of the prior odds, $\frac{p(M_1)}{p(M_2)}$, and the likelihood ratio, $\frac{p(y|M_1)}{p(y|M_2)}$. If we suppose that prior to seeing the data the 2 models are equally likely, then the prior odds is one so the posterior odds is determined solely by the likelihood ratio. Likelihood ratios provide a way of testing a hypothesis in a general probability model since one only needs to be able to write down the likelihood in order to compute the likelihood ratio.

In order to calculate the p-value for a hypothesis based on the likelihood ratio, we need to determine the distribution of the likelihood ratio assuming the null hypothesis is true. It transpires that under assumptions similar to those used to prove that the posterior is asymptotically normally distributed, $2\log\frac{p(y|M_1)}{p(y|M_2)}$ is distributed according to a χ^2 distribution with degrees of freedom given by the difference in the number of parameters in the 2 models. Note that the p-value we obtain in this fashion is not a posterior probability but rather is the usual sort of p-value. On the other hand, the Bayesian framework provides a justification for

examing the likelihood ratio (because the likelihood ratio is equal to the posterior odds when the 2 models are *a priori* equally likely).

Exercises

1. Suppose X is Binomially distributed, i.e. $P(X = k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$. Compute the MLE of θ and find its asymptotic variance.
2. Show that the posterior odds is equal to the product of the likelihood ratio and the prior odds.

References

Gelman, A., Carlin, J., Stern, H., Rubin, D. (1995), *Bayesian Data Analysis*, Chapman and Hall, New York.