Bayesian Theory and Examples

Theory

A note about notation: In general, we use capital letters to note a random variable (RV) that has not yet taken on a value, and a lower-case letter to indicate an “instatiation” or “manifestation” of the RV.

Probability Distribution

These aren’t all of the distributions, but these are brought up in the examples that follow. (Note that $f(x)$ represents the probability mass function (https://en.wikipedia.org/wiki/Probability_mass_function) (PMF) for discrete distribution and the probability density function (https://en.wikipedia.org/wiki/Probability_density_function) (PDF) for continuous distribution

Binomial

$$X \sim \mathcal{B}in(n,p)$$
$$f(x) = \binom{n}{k} p^X q^{n-X}, k = 0, 1, \ldots, n$$
$$\mathrm{E}[X] = np,$$
$$\mathrm{Var}(X) = npq.$$

This is used to model the number of successes x in n trials, where the probability of success is p. (It’s a generalization of the Bernoulli distribution, which is limited to just 1 trial.)

Uniform

$$X \sim \mathcal{U}(a,b)$$
$$f(x) = \tfrac{1}{b-a}, a \leq x \leq b,$$
$$\mathrm{E}[X] = \tfrac{a+b}{2},$$
$$\mathrm{Var}(X) = \tfrac{(b-a)^2}{12}.$$

This is commonly used if we only really know the upper and lower bounds of a distribution, and not much else.

Normal

$$X \sim \mathcal{N}(\mu, \sigma^2)$$
$$f(x) = \tfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\tfrac{-(x-\mu)^2}{2\sigma^2}\right]$$
$$\mathrm{E}[X] = \mu,$$
$$\mathrm{Var}(X) = \sigma^2.$$

This is the “mother” of all distributions. The sum of the values coming from any distribution will eventually (approximately) follow the normal distribution. (See the Central Limit Theorem.)

Gamma

$$X \sim \mathcal{G}a(\alpha, \beta)$$
$$f(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}, x \geq 0$$
$$\mathrm{E}[X] = \tfrac{\alpha}{\beta},$$
$$\mathrm{Var}(X) = \tfrac{\alpha}{\beta^2}.$$

Note that $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt.$

The Gamma distribution is a generalization of two other well-known distributions—the exponential and chi-squared distributions (which themselves draw upon attributes of other distributions). It shows up in a couple of conjugate pairs (to be discussed).

Beta

$$\begin{gathered} X \sim \mathcal{B}e(\alpha, \beta) \\ f(x) = \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0 \leq x \leq 1, \alpha, \beta > 0. \\ \mathrm{E}[X] = \frac{\alpha}{\alpha+\beta}, \\ \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \end{gathered}$$

Note that $\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

The Beta distribution is a generalization of the Gamma distribution. This distribution is probably the best distribution for modeling other distributions.

Bayes' Theorem

"Bayesian probability is an interpretation of the concept of probability, in which … probability is interpreted … as quantification of a personal belief." [1]

Although there is a "formula" for Bayes' theorem (see below), Bayes statistics is more about the "mentality". (See the Bayesian vs. Frequentist section below.)

$$\begin{aligned} \text{Posterior Probability} &= \frac{(\text{ Likelihood })(\text{ Prior Probability })}{\text{Marginal Likelihood}} \\ P(H|E) &= \frac{P(E|H)P(H)}{P(E)} \end{aligned}$$

Note the notation: We use H to indicate "hypothesis" and E to indicate evidence. Sometimes evidence is also referred to as "data" (D) in the literature.

Also, note that the Bayes' formula is really just a different way of stating the conditional probability of events. (i.e. The probability of A given B is $P(A|B) = \frac{P(B|A)P(B)}{P(A)}$.)

Regarding the components of Bayes' formula:

- The prior $P(H)$ is the probability of H before observing the data. Referring back to the definition of Bayes probability, it is the "personal" component. It is "elicited" by the person observing the data/performing the experiment, and its reliability is dependent on the knowledge of this person.

- The likelihood $P(E|H)$ is the evidence about H provided by the data.

- The marginal $P(E)$ (i.e. the "normalizing constant") is the probability of the observing the data, accounting for all possibile hypotheses. It is usually what causes difficulty in calculations. We can avoid difficulty in calculations if our problem can be modeled with conjugate pairs (to be discussed). Also, note that the marginal may also be referred to as the predictive prior. A more in-depth discussion of this discussion would clarify why this is.

- The posterior $P(H|E)$ is the probability that H is true after the data is considered.

Bayesian vs. Frequentist [2]

"Disagreements are in the nature of (1) model parameters and (2) use of conditioning:

- *Frequentists: (1) Parameters are **fixed numbers**. (2) Inference involves **optimization.***

- *Bayesians: Parameters are **random variables**. (2) Inference involves **integration**."*

(See the Beta-Binomial example below.)

Conjugacy

"Conjugacy occurs when the posterior distribution is in the same family of probability density functions as the prior belief, but with new parameter values, which have been updated to reflect what we have learned from the data." [3]

Beta-Binomial

*“Suppose we perform an experiment and estimate that the data comes from a binomial distribution $\mathcal{B}in(n,p)$ with **known** n (number of trials) and **unknown** p (probability). (This is the likelihood.) Let’s say that we have a prior belief that p can be modeled with the Beta distribution $\mathcal{B}e(\alpha,\beta)$ (with α,β that we choose, presumably with our”expertise“). (This is the prior.) In the experiment we observed x successes in n trials. Then Bayes’ rule implies that our new belief about the probability density of p—the posterior distribution, of $p|x$—is also the Beta distribution, with”updated" parameters."* [4]

$$p|x \sim \mathcal{B}e(\alpha+x,\beta+n-x).$$

We can formalize this problem set-up as follows.

$$\begin{array}{rcccl} \text{Likelihood} & : & x|p & \sim & \mathcal{B}in(n,p) \\ \text{Prior} & : & p & \sim & \mathcal{B}e(\alpha,\beta) \\ \text{Posterior} & : & p|x & \sim & \mathcal{B}e(\alpha+x,\beta+n-x) \end{array}$$

Note that it is common to “simplify” notation, e.g. $P(p|x)$ is express simply as $p|x$.

The posterior mean reflects an “update” to the prior mean given x and n.

$$\begin{array}{ccc} \mathrm{E}[X] & \to & \mathrm{E}[p|X] \\ \frac{\alpha}{\alpha+\beta} & \to & \frac{\alpha+x}{\alpha+\beta+n}. \end{array}$$

Example

We want to estimate the probability p that a coin falls heads up. After $n=10$ flips, we observe $X=0$ heads. we can model the likelihood as a $\mathcal{B}in(n=10,p)$ (where p is unknown). A “flat” prior for this kind of experiment is a $\mathcal{U}(0,1)$ distribution, which happens to be equivalent to $\mathcal{B}e(1,1)$. What does the Beta-Binomial conjugate pair lead us to conclude about p?

Using $\alpha=1,\beta=1$ for our prior distribution, the posterior probability is $\mathcal{B}e(1+(0),1+(10)-(0))=\mathcal{B}e(1,11)$. Then the posterior mean (i.e. average or expectation) is $\hat{p}=\frac{(1)}{(1)+(11)}=\frac{1}{12}$ (because the expectation of the $\mathcal{B}e$ distribution is $\mathrm{E}[X]=\frac{\alpha}{\alpha+\beta}$.) Note that

If we had taken the Frequentist approach (which does not incorporate priors), then we would have concluded that $\hat{p}=\frac{X}{n}=\frac{0}{10}=0$.

Note that the Bayes estimation of the posterior mean would be much closer to $p = 0.5$ if we had used a stronger prior (which is more realistic for something like coin flips, where it is physically very difficult to create a “rigged” coin). For example, if we had used the prior $\mathcal{B}e(1000, 1000)$, then our posterior estimate of the mean would have been $\hat{p} = \frac{(10000)}{(10000)+(11)} \approx 0.5.$

Normal-Normal

For the Normal-Normal conjugate pair, we assume that the data comes from a normal distribution with **known** variance σ^2 and **unknown** mean μ, which we want to estimate. Also, we must “elicit” value for mean μ_0 and variance σ_0^2 for the prior distribution (which is itself normal). (This is like how we choose α and β for the prior for the Beta-Binomial conjugate pair.)

$$\begin{array}{rcccl} \text{Likelihood} & : & x|\mu & \sim & \mathcal{N}(\mu, \sigma^2) \\ \text{Prior} & : & \mu & \sim & \mathcal{N}(\mu_0, \sigma_0^2) \\ \text{Posterior} & : & \mu|x & \sim & \mathcal{N}(\frac{\mu_0\sigma^2 + nx\sigma_0^2}{\sigma^2 + n\sigma_0^2}, \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2}) \end{array}$$

If you look closely, you’ll see that the posterior mean μ is actually a weighted average of the prior mean μ_0 and the observed mean $\frac{x}{n}$.

Note that this conjugate pair can also be used if assuming a known prior mean and an unknown prior variance.

Example

Joe models his IQ as $X \sim \mathcal{N}(\mu, 80)$. The distribution of IQs of students at Joe’s university is $\mathcal{N}(110, 120)$. Joe takes a single IQ test and scores 98.

$$\begin{array}{rcccl} \text{Likelihood} & : & x|\mu & \sim & \mathcal{N}(\mu, 80) \\ \text{Prior} & : & \mu & \sim & \mathcal{N}(110, 120) \\ \text{Posterior} & : & \mu|x & \sim & \mathcal{N}(\frac{(110)(80)+(1)(98)(120)}{(80)+(1)(120)}, \frac{(80)(120)}{(80+(1)(120)}) \approx \mathcal{N}(102.8, 48) \end{array}$$

Normal-Gamma

The Normal-Gamma conjugate is used to estimate two unknown parameters—mean μ and variance σ^2. In may be thought of as a more “complex” extension of the Normal-Normal conjugate pair. It is more applicable to more real-world contexts, where both the mean and variance are **unknown** (but the posterior is assumed to be normal). As you can imagine, the math gets more complicated…

General Bayes Computation

So we have seen how posterior distributions for parameters can be generated via conjugacy. Unfortunately, very few settings can be modeled “well” with conjugate pairs–which are nice because they have analytical, closed form solutions. Probably the first “computational” approach to try is a grid approximation. (In fact, this approach is useful for verifying closed-form problems.) However, this approach tends to be too simplistic and is usually not very “useful”.

Going a step beyond grid approximation would be Laplace approximation, which is also known as quadratic approximation. It represents an improvement over simple grid approximation, but it’s not the most popular approach.

The most popular computation approach–and one that is used a lot in practice–is Markov chain Monte Carlo (https://en.wikipedia.org/wiki/Markov_chain_Monte_Carlo) (MCMC). One of the terms you’ll see commonly used with MCMC is “Gibbs sampling” (https://en.wikipedia.org/wiki/Gibbs_sampling)—this is just one of many approaches to implement MCMC. Also, MCMC is where “specialized” Bayesian software that you may have heard about (e.g. OpenBUGS, JAGS, Stan) comes into play. However, these are not “native” to `R`. (There are `R` packages that “wrap” the functionality of these software, but I wouldn’t consider this “native”.)

In the end, no matter which of these approaches you use, each is just the first step towards implementing “Bayesian linear regression”, which can be thought of as “typical” linear regression with approximate parameters. Grid approximation, Laplace approximation, and MCMC are just methods for performing Bayesian estimation. They do not do inference

Bayesian Linear Regression

The best package I’ve found for implementing Bayesian linear regression with pure `R` is the `{BAS}` package (https://cran.r-project.org/web/packages/BAS/vignettes/BAS-vignette.html). It provides an interface similar to the canonical `lm()`, but with additional functionality to implement the Bayesian approach. (Most notably, you can specify `prior`s for the coefficients.)

[1] https://en.wikipedia.org/wiki/Bayesian_probability (https://en.wikipedia.org/wiki/Bayesian_probability).↩

[2] Tony’s Bayes Statistics class.↩

[3] https://statswithr.github.io/book/bayesian-inference.html#conjugacy (https://statswithr.github.io/book/bayesian-inference.html#conjugacy)↩

[4] https://statswithr.github.io/book/bayesian-inference.html#conjugacy (https://statswithr.github.io/book/bayesian-inference.html#conjugacy)↩