Bayesian Inference

Chapter 1, 2.1-2.3 BDA3

1.1 Overview of Bayesian versus Frequentist statistics

Bayesian statistics is an alternative approach to classical or frequentist statistics.

Frequentist/classical statistics

- A parameter is a fixed but unknown quantity.
- Inference is made over hypothetical random samples of data.

Bayesian statistics

- A parameter is a random variable with a probability distribution.
- Inference is made by conditioning on the observed/fixed sample of data.

Bayesian inference makes probability statements about estimands based on the observed data.

Frequentist inference makes probability statements about estimands when we repeatedly sample data from the population of interest.

Both viewpoints are useful.

Both approaches make subjective choices (though Bayesian methods tends to be a bit more transparent about these choices).

Bayes' rule is a fundamental tool in Bayesian statistics.

Consider two random variables X and Y with joint distribution $p(x, y)$.

Bayes' rule (or theorem) states that

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)},$$

where:

- $p(x|y)$ is the conditional distribution of X given Y,
- $p(y|x)$ is the conditional distribution of Y given X,
- $p(x)$ is the marginal distribution of X,
- $p(y)$ is the marginal distribution of Y.

1.2 General notation and assumptions

An **estimand** is an unobserved quantity for which statistical inference is made.

- Estimands may be observable or unobservable.

Some notation:

$y = (y_1, y_2, \dots, y_n)$	Vector of observed data
$\theta = (\theta_1, \theta_2, \dots, \theta_p)$	Vector of unobservable quantities/parameters of interest
$\tilde{y} = (\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_m)$	Vector of observable quantities of interest.
x_i or x_{ij}	Covariates or explanatory variables for observation i (or the jth explanatory variable for observation i.

The n values y_i may be regarded as **exchangeable**, i.e., the joint probability density $p(y_1, y_2, \dots, y_n)$ is invariant to permutations of the indices.

- Observations are nonexchangeable if information relevant to the outcome is conveyed in the unit indices rather than by explanatory variables.
- Exchangeable random variables as often modeled as independent and identically distributed (i.i.d.) random variables conditional on knowing the model parameters and/or enough covariate information.

1.3 Bayesian Inference

Bayesian data analysis can be idealized into three parts:

1. Set up a **full probability model**: construct a joint probability distribution for all observable and unobservable quantities in a problem.
2. Condition on the observed data: calculate and interpret the **posterior distribution**, the probability distribution of unobserved quantities of interest conditional on the observed data.
3. Evaluate model fit: Does the model fit the data? Are the modeling assumptions reasonable? How sensitive is the model to the assumptions?

Parametric statistical analysis models the random process that produced the data, y, conditional on unknown parameters θ.

$p(y\|\theta)$	The data distribution or likelihood function
$p(\theta)$	The prior distribution for θ
$p(y)$	The marginal distribution of y, a.k.a., the prior predictive distribution
$p(\theta\|y)$	The posterior distribution
$p(\tilde{y}\|y)$	The posterior predictive distribution

Bayesian conclusions about a parameter θ or unobserved data $\tilde{y}$ tend to be made in terms of **probability** statements and distributions.

Applying Bayes' rule in Bayesian Inference

By treating both y and θ as random, we can apply Bayes' rule to perform Bayesian inference.

The **prior distribution**, $p(\theta)$, describes our initial beliefs about the distribution of θ.

The **data distribution**, $p(y|\theta)$, describes our beliefs about the distribution of y if we knew the value of θ used to generate the data.

- $p(y|\theta)$ is also known as the likelihood function or sampling distribution of the data.

The **joint distribution** of y and θ, $p(y, \theta)$, is $p(y|\theta)p(\theta)$.

Bayes' rule states that the posterior distribution $p(\theta|y)$ may be found as

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)},$$

where $p(y)$ is the **marginal distribution of y** given by

$$p(y) = \sum_{\theta} p(\theta)p(y|\theta)$$

if θ is discrete or

$$p(y) = \int_{\theta} p(y|\theta)p(\theta)\, d\theta$$

if θ is continuous.

Since the denominator $p(y)$ does not depend on θ, it is simply a constant in the posterior distribution.

It is common to work with the unnormalized posterior density, which is the right side of

$$p(\theta|y) \propto p(y|\theta)p(\theta).$$

The **posterior distribution** describes our belief about the distribution of θ after seeing the data.

The marginal distribution of y is also known as the **prior predictive distribution**: prior because it is not conditional on a previously observed data and predictive because it is the distribution for a quantity that is observable.

After the data y have been observed, we can predict an unknown observable, $\tilde{y}$, of the same process.

The distribution $p(\tilde{y}|y)$ is known as the **posterior predictive distribution**.

- Posterior because it is conditional on the observed y**.**
- Predictive because it is a prediction for an observable $\tilde{y}$**.**

The posterior predictive distribution may be derived as

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

where the last line follows under the assumption that y and $\tilde{y}$ are conditionally independent given θ.

The second and third lines indicate that the posterior predictive distribution is an average of conditional predictions over the posterior distribution of θ.

Likelihood

Combining Bayes' rule with a chosen probability model means that the data y affect the posterior inference **only** through the likelihood function $p(y|\theta)$.

Bayesian inference obeys the **likelihood principle**, which states that for a given sample of data, any two probability models $p(y|\theta)$ that have the same likelihood function yield the same inference for θ.

- Inference can change in classical statistics because the data-generating mechanism for the unobserved samples impacts the results.

Bayesian inference examples

Example: Genetics (discrete)

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 gene 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 rare, since the frequency of occurrence of the gene is low in human populations.

A woman's mother carries one hemophilia gene (but not two), while her father does not. If the woman has two unaffected sons, what is the probability that the woman carries the hemophilia gene?

How does the probability change if the woman has three unaffected sons instead of two unaffected sons?

Spelling correction

Classification of words is a problem of managing uncertainty. For example, suppose someone types ‘radom.’ How should that be read? It could be a misspelling or mistyping of ‘random’ or ‘radon’ or some other alternative, or it could be the intentional typing of ‘radom’ (as in its first use in this paragraph). What is the probability that ‘radom’ actually means random?

Example: Estimating the probability of a female birth (continuous)

The proportion of female births in the European population is accepted to be less than 0.5. Let y be the number of females in n recorded births and θ be the probability of a female birth among the European population. Out of $n = 100$ randomly selected births, $y = 60$ are female. The Beta(α, β) distribution is a natural prior for θ since its support is (0,1). Without prior knowledge, any value of θ between 0 and 1 is equally likely. This is equivalent to a Beta$(1,1) = U(0,1)$ prior distribution.

Data distribution: Bin$(y|n, \theta)$

Prior distribution: Beta(α, β)

Derive the posterior distribution for θ, $p(\theta|y)$:

Derive the posterior predictive distribution of a new, independent observation, $\tilde{y}$, i.e., $p(\tilde{y}|y)$:

Summarizing (univariate) posterior distributions

The final output of a Bayesian analysis is the posterior distribution of all model parameters, and possibly, the posterior predictive distribution of observable estimands.

- The distribution may not have a simple, closed form.
- Our intuition about a distribution often relies on knowing characteristics such as the mean, median, or standard deviation instead of a formula.

Plotting

A univariate posterior is best summarized by a plot since it retains all information about the parameter.

- It still might be difficult to determine the "center" and "spread" of the posterior distribution from a plot.

Point estimation

Point estimates are often used to summarize a posterior distribution.

- This is a single value that "best" estimates the parameter based on the posterior distribution.

The **maximum a posteriori (MAP) estimator** is a common choice for summarizing a posterior distribution.

The MAP estimator is the parameter value that maximizes the posterior distribution (i.e., the posterior mode):

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \log[p(\theta|y)] = \arg\max_{\theta} (\log[p(y|\theta)] + \log[p(\theta)]).$$

Other common point estimators for a (univariate) posterior distribution include:

Posterior mean: $E(\theta|y)$

Posterior median: e.g., $\hat{\theta}_{\text{median}} = \left\{\theta^*: \int_{-\infty}^{\theta^*} p(\theta|y)\, d\theta = 0.5\right\}$

Posterior variance: $\text{var}(\theta|y)$

Posterior standard deviation: $\text{SD}(\theta|y) = \sqrt{\text{var}(\theta|y)}$

Posterior quantiles

Interval estimation

A **credible interval** (or **posterior interval)** is any interval (l, u) such that $P(l \leq \theta \leq u|y) = 1 - \alpha$.

Common methods for constructing credible intervals:

- **Central** or **equal-tailed** credible interval: l and u are the $\alpha/2$ and $1 - \alpha/2$ quantiles of the posterior distribution.
- **Highest posterior density (HPD)** credible interval: the range of values that contains $100(1 - \alpha)\%$ of the posterior probability AND the density within the region is never lower than that outside.

The central credible interval:

- Has a direct interpretation as the posterior $\alpha/2$ and $1-\alpha/2$ quantiles.
- Is invariant under one-to-one transformation of the estimand.
- Is easier to compute.
- It is more popular than the HPD interval.

The HPD interval:

- Is always at least as narrow as the central credible interval.
- May differ substantially from the central credible interval when the posterior distribution is highly skewed.

Bayesian credible interval v frequentist confidence interval

A $100(1-\alpha)\%$ credible interval for θ is an interval $[l(y), u(y)]$, based on the observed data $Y = y$, such that

$$\Pr(l(y) \leq \theta \leq u(y) | Y = y) = 1 - \alpha.$$

A $100(1-\alpha)\%$ confidence interval for θ is a random interval $[l(Y), u(Y)]$, such that before the data are gathered,

$$\Pr(l(Y) \leq \theta \leq u(Y) | \theta) = 1 - \alpha.$$

The credible interval contains the parameter with the specified probability AFTER seeing the data. The thing that is random in this setting is the PARAMETER θ.

The confidence interval will contain the parameter with the specified probability BEFORE seeing the data. Once the data is observed, the interval will either contain the parameter or it will not. The thing that is random in this setting is the DATA Y.

Monte Carlo sampling

The posterior distribution may not be describable with a closed-form expression, and a plot may not be adequate to study the posterior characteristics we care about it.

Monte Carlo sampling is a method for drawing independent samples from a distribution.

Suppose we use Monte Carlo sampling to draw $\theta^{(1)}, \ldots, \theta^{(B)}$, a sample of independent, random values of θ, from the posterior distribution $p(\theta|y)$.

Then the empirical distribution of $\{\theta^{(1)}, \ldots, \theta^{(B)}\}$ would approximate $p(\theta|y)$, and the approximation would improve as the value of B increased.

This empirical distribution is known as a **Monte Carlo approximation** to $p(\theta|y)$.

The law of large numbers says that if $\theta^{(1)}, \ldots, \theta^{(B)}$ are i.i.d. realizations from $p(\theta|y)$, then for (almost) any function $h(\theta)$[1], we have

$$\frac{1}{B}\sum_{j=1}^{B} h\left(\theta^{(j)}\right) \rightarrow E[h(\theta|y)] = \int h(\theta)p(\theta|y)d\theta \text{ as } B \rightarrow \infty.$$

[1] The expected value of $h(\theta)$ needs to be finite.

This implies that

- $\bar{\theta} = \frac{\sum \theta^{(j)}}{B} \rightarrow E(\theta|y)$
- $\frac{1}{B-1}\sum\left(\theta^{(j)} - \bar{\theta}\right)^2 \rightarrow \text{var}(\theta|y)$
- $\frac{\#\theta^{(j)} \leq c}{B} \rightarrow P(\theta \leq c|y)$
- The empirical distribution of $\left\{\theta^{(1)}, \ldots, \theta^{(B)}\right\} \rightarrow p(\theta|y)$
- The α-quantile of $\left\{\theta^{(1)}, \ldots, \theta^{(B)}\right\}$ converges to the α-quantile of $p(\theta|y)$

This is also known as **Monte Carlo integration**.

Example: Summarizing a posterior distribution

Suppose $\theta|y \sim \text{Beta}(10, 2)$.

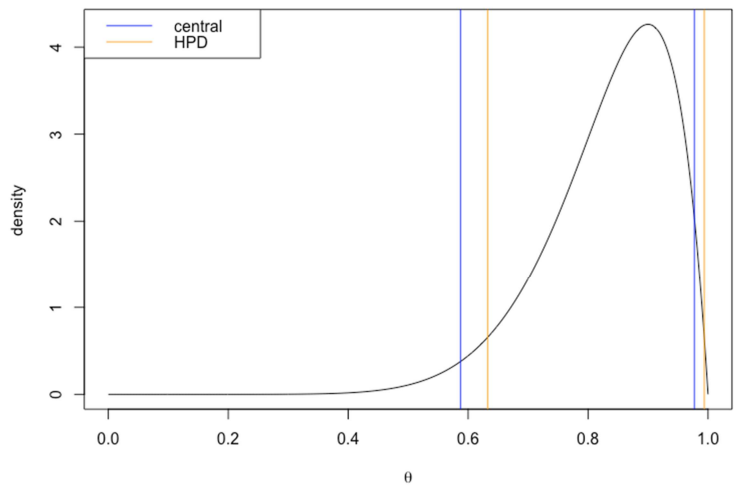
A plot of the distribution is shown below.

The 95% central credible interval is the range: [0.59, 0.98].

- This was found by determining the .025 and .975 quantiles of the Beta(10, 2) distribution.
- `qbeta(c(0.025, 0.975), 10, 2)`

The 95% HPD interval is the range: [0.63, 0.99].

- This was found numerically using the `hpd` function in the **TeachingDemos** R package.



The MAP estimate is approximately 0.9.

- `optimize(f = dbeta, interval = 0:1, shape1 = 10, shape2 = 2, maximum = TRUE)`

The posterior mean is $10/(10 + 2) = 5/6$.

- Estimate via simulation: `mean(rbeta(1000, 10, 2))`

The posterior variance is $10(2)/[(10 + 2)^2(10 + 2 + 1)] \approx 0.01$.

- Estimate via simulation: `var(rbeta(1000, 10, 2))`

Posterior distribution as compromise

Is there a general relation between the prior distribution and the posterior distribution in Bayesian inference?

Recall that

$E(\theta) = E[E(\theta|y)]$ and $\text{var}(\theta) = E[\text{var}(\theta|y)] + \text{var}[E(\theta|y)]$.

The prior mean of θ is the average of all possible posterior means over the distribution of the possible data.

The posterior variance is, on average, smaller than the prior variance, by an amount that depends on the variation in the posterior means over all the data.

The posterior distribution is a compromise between the prior information and the data, and the compromise is more heavily weighted toward the data as the sample size increases.

Example: Binomial data

In the binomial data example using the $\text{Beta}(1, 1) = U(0,1)$ prior distribution, the prior mean is $1/2$, the sample proportion is y/n, and the posterior mean is $(y + 1)/(n + 2)$.

Notice

$$\frac{y+1}{n+2} = \frac{n}{n+2} \cdot \frac{y}{n} + \frac{2}{n+2} \cdot \frac{1}{2}.$$