Chapter 2: One-parameter Model

ST4234: Bayesian Statistics

Semester 2, AY 2019/2020

Department of Statistics and Applied Probability

National University of Singapore

LI Cheng

stalic@nus.edu.sg

Introduction

- In this chapter, we consider Bayesian inference for two one-parameter models: the Binomial model and the Poisson model. This corresponds to Chapter 3 in Peter Hoff's book.
- These models also provide an environment within which we will learn the basics of Bayesian data analysis, including conjugate prior distributions, predictive distributions and summaries of posterior distributions.
- We first quickly revisit the fundamentals of Bayesian inference in the previous chapter.

Outline

- Bayesian Inference
- The Binomial Model
    - Model, Prior, and Posterior
    - Combining Information
    - Conjugacy
    - Sufficient Statistics
    - Prediction
    - Summaries of posterior distribution
- The Poisson model
    - Combining information
    - Prediction
- Mixtures of conjugate priors
- Exponential families and conjugate priors
- Summary

2.1 Bayesian Inference

Bayesian Framework

- Suppose we have some **unknown quantity** θ (possibly a vector) that we wish to learn about and we observe some**data** y relevant to their values.

- In Bayesian statistics, we need to specify

  - a **sampling model** or a probability distribution which describes how y depends on θ. This is expressed as a probability density function (pdf) $p(y|\theta)$, which we call the **likelihood function**.

  - a **prior distribution** $p(\theta)$ which expresses any prior knowledge or beliefs that we have about their values before observing the data.

- Inference concerning θ is based on the **posterior distribution** $p(\theta|y)$, which is computed from the likelihood and prior using Bayes' Theorem.

$$p(\theta|y) = \frac{p(y|\theta)\,p(\theta)}{\int p(y|\theta')\,p(\theta')\,\mathrm{d}\theta'}.$$

2.1 Bayesian Inference

Marginal Distribution

- The posterior distribution describes our beliefs about θ after our prior beliefs have been updated using the observed data.
- The integral in the denominator is called the **marginal distribution** of the data y, and we denote it as

$$m(y) = \int p(y|\theta')\, p(\theta')\, \mathrm{d}\theta'.$$

- As the denominator $m(y)$ does not depend on θ,

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

- The likelihood may be multiplied by any constant (or function of y alone) without altering the posterior. Hence, when finding the posterior, we can leave out any multiplicative constants and any factor of $p(y|\theta)$ that is not a function of θ.

2.1 Bayesian Inference

Sequential use of Bayes' Theorem

- Suppose we have two sets of independently collected observations y_1 and y_2.
- The posterior for the full dataset (y_1, y_2) is

$$\begin{aligned} p(\theta|y_1, y_2) &\propto p(y_1, y_2|\theta)\ p(\theta) \\ &= p(y_2|\theta)\ p(y_1|\theta)\ p(\theta) \quad [y_1 \text{ and } y_2 \text{ are independent}] \\ &\propto p(y_2|\theta)\ p(\theta|y_1) \end{aligned}$$

- Hence we can find the posterior of θ given (y_1, y_2), by treating the posterior of $p(\theta|y_1)$ as the prior for y_2.
- This formula for updating the posterior can be applied naturally when the data arrive sequentially and independently over time.

2.1 Bayesian Inference

Predictive Posterior

- Suppose that we have observed data $\boldsymbol{y} = (y_1, \ldots, y_n)$ and are interested in predicting the value of a future observation $\tilde{y}$.
- The posterior predictive distribution of $\tilde{y}$ is

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

[assume $\tilde{y}$ is independent of $\boldsymbol{y}$ given θ]

- The last step follows because if $\tilde{y}$ and $\boldsymbol{y}$ are independent given θ, then $p(\tilde{y}|\theta, \boldsymbol{y}) = p(\tilde{y}|\theta)$.
- This distribution $p(\tilde{y}|\boldsymbol{y})$ summarizes the information concerning the likely value of a new observation, given the likelihood, the prior and the data we have observed so far.

- The marginal distribution of $\tilde{y}$ is

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

  This is also referred to as the prior predictive distribution of $\tilde{y}$ as it summarizes our information about $\tilde{y}$ **before** observing the data.

- We will see how to compute the predictive distribution $p(\tilde{y}|\boldsymbol{y})$ in the next examples of one-parameter models.

Outline

Bayesian Inference

The Binomial Model

- Model, Prior, and Posterior
- Combining Information
- Conjugacy
- Sufficient Statistics
- Prediction
- Summaries of posterior distribution

The Poisson model

- Combining information
- Prediction

Mixtures of conjugate priors

Exponential families and conjugate priors

Summary

2.2.1 Model, Prior, and Posterior

Happiness data

- In the 1998 General Social Survey, 129 females age 65 or over were asked if they were generally happy. The results were:

Happy	Not happy	Total
118	11	129

- What can we infer about the proportion θ of females age 65 or over who are generally happy?
- Let us frame this problem using the binomial model.

2.2.1 Model, Prior, and Posterior

The Binomial Model

- Suppose the parameter of interest is the probability of success θ in n independent trials which can result in either success or failure.
- Assume n is fixed and the probability of success is the same in each trial.
- The number of successes is a random variable $Y \in \{0, 1, \dots, n\}$ which has a binomial distribution with index n and parameter θ, denoted

$$Y \sim \text{Binomial}(n, \theta).$$

- The probability of an observation y is

$$\text{P}(Y = y|\theta) = p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}.$$

- $\text{E}(Y|\theta) = n\theta$ and $\text{Var}(Y|\theta) = n\theta(1-\theta)$.

2.2.1 Model, Prior, and Posterior

R commands for binomial distribution

- `dbinom`: density (probability in the case of discrete random variable)
- `pbinom`: cumulative distribution function
- `qbinom`: quantile function
- `rbinom`: random generation

Example: Suppose $Y \sim \text{Binomial}(10, 0.2)$.

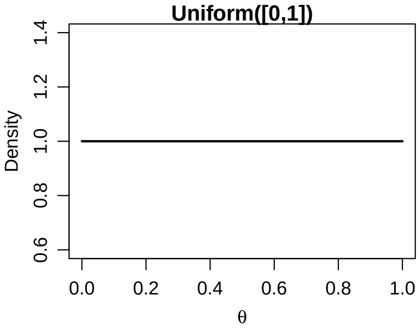
- Find $\text{P}(Y = 2)$: `dbinom(x=2, size=10, prob=0.2)` (Ans: 0.302)
- Find $\text{P}(Y \leq 2)$: `pbinom(q=2, size=10, prob=0.2)` (Ans: 0.678)
- Find the smallest integer α such that $P(Y \leq \alpha) \geq 0.7$: `qbinom(p=0.7, size=10, prob=0.2)` (Ans: 3)
- Generate 50 random numbers from $\text{Binomial}(10, 0.2)$: `rbinom(n=50, size=10, prob=0.2)`

2.2.1 Model, Prior, and Posterior

Uniform Prior

- The parameter θ is some unknown value between 0 and 1.
- First let us consider a uniform density on $[0, 1]$ as the prior for θ:

$$p(\theta) = 1 \text{ for } 0 \leq \theta \leq 1.$$



2.2.1 Model, Prior, and Posterior

Posterior Distribution

- Applying Bayes' Theorem, the posterior distribution of θ is

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &= \binom{n}{y}\theta^{y}(1-\theta)^{n-y}\cdot 1 \\ &\propto \theta^{y}(1-\theta)^{n-y} \quad [\text{omit } \binom{n}{y} \text{ as it is not a function of } \theta]. \end{aligned}$$

Thus $p(\theta|y) = \dfrac{\theta^{y}(1-\theta)^{n-y}}{C}$ for some proportionality constant C.

2.2.1 Model, Prior, and Posterior

Normalization Constant

- For $p(\theta|y)$ to be a proper density,

$$\int_0^1 p(\theta|y)\,\mathrm{d}\theta = \frac{1}{C}\int_0^1 \theta^y(1-\theta)^{n-y}\,\mathrm{d}\theta = 1$$
$$\Rightarrow C = \int_0^1 \theta^y(1-\theta)^{n-y}\,\mathrm{d}\theta.$$

  We say that C is a normalization constant as it ensures $p(\theta|y)$ integrates to 1.

- From calculus, the Beta function $\mathrm{B}(a,b)$ is defined as

$$\mathrm{B}(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}\,\mathrm{d}\theta$$

  Thus $C = \mathrm{B}(y+1, n-y+1)$.

- To compute $\mathrm{B}(a,b)$ in R, just use `beta(a,b)`.

2.2.1 Model, Prior, and Posterior

Posterior Distribution

- The posterior distribution of θ is

$$p(\theta|y) = \frac{\theta^{y}(1-\theta)^{n-y}}{\mathrm{B}(y+1, n-y+1)}, \quad 0 \le \theta \le 1.$$

  This distribution belongs to the family of beta distributions.

- We say that θ follows the beta distribution with parameters $a > 0$ and $b > 0$, denoted $\theta \sim \mathrm{Beta}(a, b)$ if

$$p(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{\mathrm{B}(a, b)}, \quad 0 \le \theta \le 1.$$

- Thus, we have shown that

$$\begin{array}{r} \theta \sim \mathrm{Uniform}[0,1] \\ Y \sim \mathrm{Binomial}(n, \theta) \end{array} \;\Rightarrow\; \theta|y \sim \mathrm{Beta}(y+1, n-y+1).$$

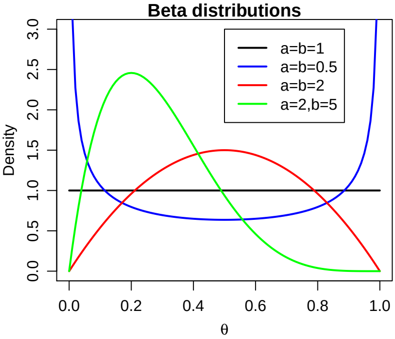
2.2.1 Model, Prior, and Posterior

Properties of beta distribution

If $\theta \sim \text{Beta}(a, b)$,

- $\text{E}(\theta) = \dfrac{a}{a+b}$, $\text{Var}(\theta) = \dfrac{ab}{(a+b)^2(a+b+1)}$.
- If $a > 1$, $b > 1$, then the mode of θ is $\dfrac{a-1}{a+b-2}$.

- This plot shows some members in the family of beta distributions.
- Note that the uniform distribution is a member with $a = b = 1$.



2.2.1 Model, Prior, and Posterior

R commands for beta distribution

- The functions `dbeta, pbeta, qbeta, rbeta` provide the density, cumulative distribution function, quantile function and random generation respectively for the beta distribution in `R`.

Example: Suppose $\theta \sim \text{Beta}(3, 9)$.

- Find $p(0.5)$: `dbeta(x = 0.5, shape1 = 3, shape2 = 9)`
  (Ans: 0.483)
- Find $\text{P}(\theta \leq 0.1)$: `pbeta(q = 0.1, shape1 = 3, shape2 = 9)`
  (Ans: 0.0896)
- Find α such that $P(\theta \leq \alpha) = 0.025$:
  `qbeta(p = 0.025, shape1 = 3, shape2 = 9)` (Ans: 0.06)
- Generate 50 random numbers from $\text{Beta}(3, 9)$:
  `rbeta(n = 50, shape1 = 3, shape2 = 9)`

2.2.1 Model, Prior, and Posterior

Happiness data

- Let θ denote the proportion of females age 65 or over who are generally happy. Let us assume the prior

$$p(\theta) = 1 \;\text{ for }\; 0 \leq \theta \leq 1.$$

- Let Y denote the number of females who reported being generally happy out of 129 females age 65 or over surveyed. Then

$$Y \sim \text{Binomial}(129, \theta).$$

$$\text{P}(Y = y|\theta) = p(y|\theta) = \binom{129}{\theta} \theta^y (1-\theta)^{129-y}.$$

2.2.1 Model, Prior, and Posterior

Happiness data

- The observed value of Y is $y = 118$. The likelihood is thus

$$p(y|\theta) = \binom{129}{118}\theta^{118}(1-\theta)^{11}.$$

- From our earlier results, the posterior is

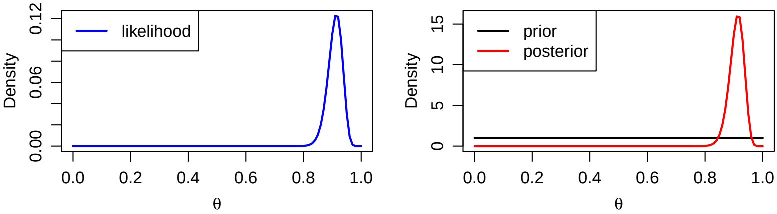
$$\theta|y \sim \text{Beta}(118+1, 129-118+1).$$

$$p(\theta|y) = \frac{\theta^{118}(1-\theta)^{11}}{B(119, 12)} \text{ for } 0 \leq \theta \leq 1.$$

- The likelihood and posterior are proportional to each other as functions of θ. They have the same shape but not the same scale.

2.2.1 Model, Prior, and Posterior

Happiness data



- As $\theta|y \sim \text{Beta}(119, 12)$,
  - Posterior mean: $\text{E}(\theta|y) = 119/(119+12) = 0.908$.
  - Posterior standard deviation:

$$\sqrt{\text{Var}(\theta|y)} = \sqrt{\frac{119 \cdot 12}{(119+12)^2(119+12+1)}} = 0.025.$$

  - Posterior mode: $(119-1)/(119+12-2) = 0.915$.

2.2.1 Model, Prior, and Posterior

General Beta Prior

- The uniform distribution is a special case of Beta(a_0, b_0) distribution, with $a_0 = b_0 = 1$.
- Now we can consider the more general beta prior.
- Suppose $\theta \sim \text{Beta}(a_0, b_0)$, then

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &\propto \binom{n}{y}\theta^y(1-\theta)^{n-y} \cdot \frac{1}{B(a_0, b_0)}\theta^{a_0-1}(1-\theta)^{b_0-1} \\ &\propto \theta^{a_0+y-1}(1-\theta)^{b_0+n-y-1}. \end{aligned}$$

- $p(\theta|y)$ has the same shape as the pdf of Beta$(a_0 + y, b_0 + n - y)$. They must also have the same scale since both integrate to 1 over $\theta \in [0, 1]$.
- Thus $p(\theta|y)$ and the pdf of Beta$(a_0 + y, b_0 + n - y)$ must be the same and

$$\theta|y \sim \text{Beta}(a_0 + y, b_0 + n - y).$$

2.2.1 Model, Prior, and Posterior

Posterior corresponding to a beta prior

- The derivations on the previous slide have shown that

$$\begin{aligned} \theta &\sim \text{Beta}(a_0, b_0) \\ Y &\sim \text{Binomial}(n, \theta) \end{aligned} \quad \Rightarrow \quad \theta | y \sim \text{Beta}(a_0 + y, b_0 + n - y).$$

- Throughout this course, we will use this trick to identify posterior distributions: we will recognize that the posterior distribution is proportional to a known probability density, and therefore must equal that density.

2.2.2 Combining Information

- If $\theta|y \sim \text{Beta}(a_0 + y, b_0 + n - y)$, then

$$\begin{aligned} \text{E}(\theta|y) &= \frac{a_0 + y}{a_0 + b_0 + n} \\ &= \frac{\cancel{a_0 + b_0}}{a_0 + b_0 + n} \underbrace{\frac{a_0}{\cancel{a_0 + b_0}}}_{\text{prior mean}} + \frac{\cancel{n}}{a_0 + b_0 + n} \underbrace{\frac{y}{\cancel{n}}}_{\text{sample mean}} . \end{aligned}$$

- Note that $\frac{y}{n} = \frac{\sum_{i=1}^{n} y_i}{n}$ is the sample mean.
- The posterior mean is a weighted average of the prior mean and the sample mean with weights proportional to $a_0 + b_0$ and n respectively.

2.2.2 Combining Information

- From

$$\mathrm{E}(\theta|y) = \frac{a_0 + b_0}{a_0 + b_0 + n} \underbrace{\frac{a_0}{a_0 + b_0}}_{\text{prior mean}} + \frac{n}{a_0 + b_0 + n} \underbrace{\frac{y}{n}}_{\text{sample mean}},$$

  we can relate a_0 with y and $a_0 + b_0$ with n.

- This leads to an interpretation of a_0 and b_0 as "prior data":

$$\begin{aligned} a_0 &: \text{prior number of 1s} \\ b_0 &: \text{prior number of 0s} \\ a_0 + b_0 &: \text{prior sample size} \end{aligned}$$

- If $n > a_0 + b_0$, the majority of our information about θ would come from the data as opposed to the prior distribution.

- For example, if $n \gg a_0 + b_0$, then $\mathrm{E}(\theta|y) \approx \frac{y}{n}$.

2.2.3 Conjugacy

- We say that the class of beta priors is conjugate for the binomial sampling model as the prior and posterior distributions come from the same family of distributions (the beta distribution).

- A class $\mathcal{P}$ of prior distributions for θ is called **conjugate** for a sampling model $p(y|\theta)$, if

$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|y) \in \mathcal{P}.$$

- If $p(y|\theta)$ is the binomial distribution, then $\mathcal{P}$ can be taken as the family of all beta distributions.

- The family of conjugate priors $\mathcal{P}$ is not unique. However, there are commonly used conjugate priors for most parametric distributions. We will introduce them as we cover model parametric distributions later.

2.2.4 Sufficient Statistics

Happiness data

- In the General Social Survey, 129 females age 65 and over were asked if they were generally happy.
- Let $Y_i = 1$ if respondent i reported being generally happy, and $Y_i = 0$ otherwise. Then

$$
\begin{aligned}
&Y_i \sim \text{Bernoulli}(\theta)\\
&\text{P}(Y_i = y_i) = \theta^{y_i}(1-\theta)^{1-y_i}.
\end{aligned}
$$

- In finding the posterior distribution of θ, we made use of the summary statistic that a total of $Y = 118$ women reported being happy out of 129.
- Would the posterior distribution change if we consider the full data $Y_1 = y_1 \ldots, Y_n = y_n$ instead?

2.2.4 Sufficient Statistics

Happiness data

- Let $\boldsymbol{y} = (y_1, \dots, y_n)$. The likelihood

$$p(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} [\theta^{y_i}(1-\theta)^{1-y_i}] = \theta^{y}(1-\theta)^{n-y},$$

  where $y = \sum_{i=1}^{n} y_i$.

- It is easy to verify that $p(\theta|\boldsymbol{y}) = p(\theta|y)$ (Try it!).
- This implies that the statistic $Y = \sum_{i=1}^{n} Y_i$ contains all the information about θ available from the data.
- As long as we know the value of this statistic, we can ignore what the original data $\boldsymbol{y} = (y_1, \dots, y_n)$.
- We say that Y is a sufficient statistic for θ given $\boldsymbol{y} = (y_1, \dots, y_n)$. It is "sufficient" to know Y in order to make inference about θ.

2.2.4 Sufficient Statistics

Definition

- Let t be a function of the observations $\boldsymbol{y} = (y_1, \ldots, y_n)$.
- The statistic t is called **a sufficient statistic** for θ given $\boldsymbol{y}$ if

$$p(\boldsymbol{y}|t, \theta) = p(\boldsymbol{y}|t).$$

- The above definition says that the conditional distribution of $\boldsymbol{y}$, given the statistic t, does not depend on the parameter θ.
- If t is a sufficient statistic, then

$$\begin{aligned} p(\boldsymbol{y}|\theta) &= p(\boldsymbol{y}, t|\theta) \\ &= p(\boldsymbol{y}|t, \theta)\, p(t|\theta) \\ &= p(\boldsymbol{y}|t)\, p(t|\theta). \end{aligned} \tag{1}$$

2.2.4 Sufficient Statistics

Let $Y_i \sim \text{Bernoulli}(\theta)$ for $i = 1, \ldots, n$ and $T = \sum_{i=1}^{n} Y_i$. Then we can show that $t = \sum_{i=1}^{n} y_i$ is a sufficient statistic for θ given $\boldsymbol{y} = (y_1, \ldots, y_n)$.

- We need to show that $p(\boldsymbol{y}|t, \theta)$ is independent of θ. Note that $T \sim \text{Binomial}(n, \theta)$.
- Then,

$$p(\boldsymbol{y}|t, \theta) = \frac{p(\boldsymbol{y}, t|\theta)}{p(t|\theta)} = \begin{cases} \dfrac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \dfrac{1}{\binom{n}{t}} & \text{if } t = \sum_{i=1}^{n} y_i, \\ 0 & \text{if } t \neq \sum_{i=1}^{n} y_i. \end{cases}$$

which is independent of θ, Thus $p(\boldsymbol{y}|t, \theta) = p(\boldsymbol{y}|t)$.

2.2.4 Sufficient Statistics

- For any prior distribution, the posterior distribution of θ given $\boldsymbol{y}$ is the same as the posterior distribution of θ given a sufficient statistic t.
- Proof (not examinable):

$$\begin{aligned} p(\theta|\boldsymbol{y}) &= \frac{p(\boldsymbol{y}|\theta)\ p(\theta)}{\int p(\boldsymbol{y}|\theta')\ p(\theta')\ \mathrm{d}\theta'} \quad \text{[Bayes' Theorem]} \\ &= \frac{p(\boldsymbol{y}|t)\ p(t|\theta)\ p(\theta)}{\int p(\boldsymbol{y}|t)\ p(t|\theta')\ p(\theta')\ \mathrm{d}\theta'} \quad \text{[from (1)]} \\ &= \frac{p(t|\theta)\ p(\theta)}{\int p(t|\theta')\ p(\theta')\ \mathrm{d}\theta'} \\ &= p(\theta|t). \end{aligned}$$

2.2.4 Sufficient Statistics

The following theorem helps us find sufficient statistics more readily.

Theorem (Factorization Theorem (Fisher-Neyman))

A statistic t is sufficient for θ given $\boldsymbol{y}$ if and only if there are functions f and g such that

$$p(\boldsymbol{y}|\theta) = f(t,\theta)g(\boldsymbol{y}),$$

where $t = t(\boldsymbol{y})$.

In the binomial model, we can take

$$f(t,\theta) = \theta^t(1-\theta)^{n-t}, \quad g(\boldsymbol{y}) = 1.$$

2.2.5 Prediction

- Suppose we wish to predict the response of a female (age 65 and over) who was not included in the survey.
- Let $\widetilde{Y} \in \{0, 1\}$ denote the response of the female who was not surveyed previously. Then $\widetilde{Y}|\theta \sim \text{Bernoulli}(\theta)$.
- The posterior predictive distribution of $\widetilde{Y}$ is the conditional distribution of $\tilde{Y}$ given the observations $\boldsymbol{y} = (y_1, \ldots, y_n)$.

2.2.5 Prediction

Happiness data

$$
\begin{aligned}
\mathrm{P}(\widetilde{Y} = 1|\boldsymbol{y}) &= \int \mathrm{P}(\widetilde{Y} = 1, \theta|\boldsymbol{y})\ \mathrm{d}\theta \\
&= \int \mathrm{P}(\widetilde{Y} = 1|\theta)\ p(\theta|\boldsymbol{y})\ \mathrm{d}\theta \\
&= \int \theta\ p(\theta|y)\ \mathrm{d}\theta = \mathrm{E}(\theta|y) = \frac{a_0 + y}{a_0 + b_0 + n}.
\end{aligned}
$$

$$
\mathrm{P}(\widetilde{Y} = 0|\boldsymbol{y}) = 1 - \mathrm{P}(\widetilde{Y} = 1|\boldsymbol{y}) = \frac{b_0 + n - y}{a_0 + b_0 + n}.
$$

- $\widetilde{Y}$ is (unconditionally) not independent of $Y_1, \ldots, Y_n$, because observing $Y_1, \ldots, Y_n$, gives information about θ, which in turn gives information about $\widetilde{Y}$.
- Conditional on the value of θ, $\widetilde{Y}$ is independent of $Y_1, \ldots, Y_n$ and is distributed as Bernoulli(θ). Our current beliefs about θ are contained in the posterior $p(\theta|\boldsymbol{y})$.

2.2.6 Summaries of posterior distribution

Point Estimation

- To obtain a point estimate $\hat{\theta}$ of θ, we may select a summary feature of $p(\theta|y)$ such as its mean, median, or mode.
- As $p(\theta|y) \propto p(\theta)p(y|\theta)$, the posterior mode is the value of θ that maximizes the R.H.S. Moreover, when the prior is flat (e.g. Uniform[0,1]), the posterior mode is equal to the maximum likelihood estimate of θ (value of θ that maximizes the likelihood $p(y|\theta)$).
- For symmetric posterior densities, the mean and median are equal (why?). If the posterior is unimodal as well, then all three measures coincide (e.g. normal distribution).
- For asymmetric posteriors, the median is often preferred as the mode considers only the value corresponding to the maximum value of the density while the mean gives too much weight to extreme outliers.

2.2.6 Summaries of posterior distribution

Interval Estimation

- Let Θ denote the parameter space of θ. It is often desirable to identify regions that are likely to contain the true value of θ.
- In Bayesian inference, such a region is referred to as a **credible set**, although people also use the term "Bayesian confidence interval" or simply "confidence interval (CI)".

2.2.6 Summaries of posterior distribution

Credible Set

- A $100(1-\alpha)\%$ credible set for θ is a subset $\mathcal{C}$ of Θ such that

$$P(\theta \in \mathcal{C} \mid y) = \int_{\mathcal{C}} p(\theta|y)\mathrm{d}\theta \geq 1-\alpha,$$

  where integration is replaced by summation if θ is discrete.

- We used $\geq$ instead of $=$ in the above definition to accommodate discrete settings, i.e. when the posterior $p(\theta|y)$ is a discrete distribution and it may not be possible to obtain an interval with coverage probability exactly $(1-\alpha)$.

- However, in continuous settings, i.e. when the posterior $p(\theta|y)$ is a density function, we would like credible sets with exact coverage to minimize their size and for more precision.

2.2.6 Summaries of posterior distribution

Interpretation of Credible Sets

The interpretation of credible set in Bayesian statistics is different from that of confidence interval in frequentist statistics (what you have learned before in mathematical statistics class).

- In Bayesian statistics, the unknown parameter θ is regarded as a random variable, and the interval is fixed once data is observed. Thus we can make direct probabilistic statements like

  "The probability that θ lies in $\mathcal{C}$ given observed data y is $(1-\alpha)$."

2.2.6 Summaries of posterior distribution

Interpretation of Credible Sets

- In the frequentist statistics, it is Y that is regarded as random and giving rise to a random interval which has probability $(1 - \alpha)$ of containing the fixed but unknown θ. The corresponding statement would be,

  "If we could recompute $\mathcal{C}$ for a large number of datasets collected in the same way as ours, about $100(1 - \alpha)\%$ of them will contain the true value of θ."

- This statement is not very comforting as we may not be able to even imagine repeating our experiment a large number of times.
- In comparison, the Bayesian interpretation of credible sets is purely probabilistic and more straightforward.

2.2.6 Summaries of posterior distribution

Interpretation of Credible Sets

- From another perspective, the frequentist and Bayesian notions of coverage describe pre- and post-experimental coverage, respectively.

- In the frequentist approach, once we observe the data $Y = y$ and use it to compute the confidence interval. The resulting confidence interval will either contain the true value of θ or not, so the actual coverage probability will be either 1 or 0.

- In the Bayesian approach, the credible set is a truly random set. The randomness comes from the randomness in the data $Y = y$.

- Researchers have shown that in general, the Bayesian credible sets constructed using the methods we are going to introduce next, will also have "almost" the correct frequentist coverage. Roughly speaking, a 95% credible set from a Bayesian approach can also be approximately used as a 95% confidence interval in the frequentist setup.

2.2.6 Summaries of posterior distribution

Quantile-based/equal-tails intervals

- An easy way to obtain a $100(1-\alpha)\%$ confidence interval (CI, or credible interval, credible set) for θ is to take the $\alpha/2$ and $(1-\alpha/2)$ quantiles of $p(\theta|y)$.
- Quantile-based/equal-tails CI: We find two numbers $\theta_{\alpha/2} < \theta_{1-\alpha/2}$, such that

$$\mathrm{P}(\theta < \theta_{\alpha/2} \mid y) = \alpha/2 \quad \text{and} \quad \mathrm{P}(\theta > \theta_{1-\alpha/2} \mid y) = \alpha/2.$$

- This implies $\mathrm{P}(\theta_{\alpha/2} < \theta < \theta_{1-\alpha/2} \mid y) = 1-\alpha$.
- Hence the $100(1-\alpha)\%$ quantile-based CI is $[\theta_{\alpha/2}, \theta_{1-\alpha/2}]$.

2.2.6 Summaries of posterior distribution

Quantile-based/equal-tails intervals

Example: Binomial sampling and uniform prior

Suppose we observe $Y = 2$ ones out of $n = 10$ independent draws of a binary random variable. Let θ denote the probability of observing a one in each draw. Using a uniform prior distribution for θ, find a quantile-based 95% posterior confidence interval for θ.

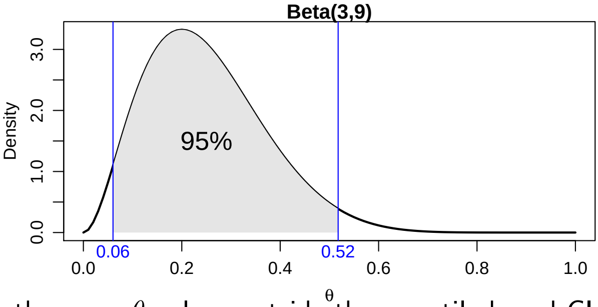
- Let $y = 2$. The posterior distribution is $\theta|y \sim \text{Beta}(1+2, 1+8)$.
- A 95% posterior CI can be obtained from the .025 and .975 quantiles of this beta distribution. Hence the 95% quantile-based CI is [0.06, 0.52].

```
> qbeta(c(0.025, 0.975), 3, 9)
[1] 0.06021773 0.51775585
```

2.2.6 Summaries of posterior distribution

Quantile-based/equal-tails Intervals

- The figure below show a plot of the posterior distribution Beta$(3, 9)$ and the quantile-based CI [0.06, 0.52].



- Note that there are θ-values outside the quantile-based CI that have higher probability than some points inside the interval. This suggests a more restrictive type of interval.

2.2.6 Summaries of posterior distribution

Highest posterior density (HPD) region

- The highest posterior density (HPD) credible set is defined as the set

  $$\mathcal{C} = \{\theta \in \Theta : p(\theta|y) \geq k(\alpha)\}$$

  where $k(\alpha)$ is the largest constant satisfying

  $$P(\theta \in \mathcal{C} \mid y) \geq 1 - \alpha.$$

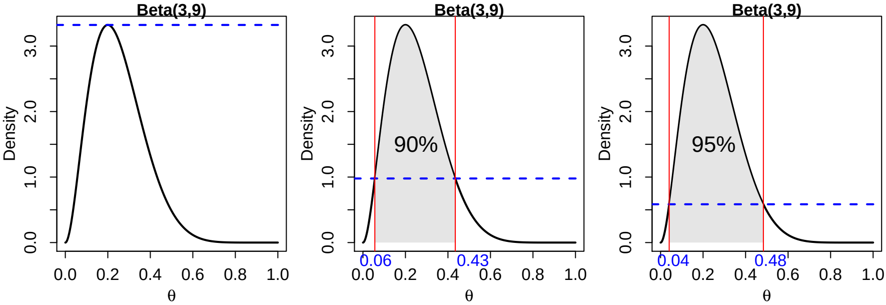
- All points in a HPD region have higher posterior density than points outside the region.
- Such a credible set is very appealing as it groups together the "most likely" θ values.
- It is easier to understand this definition by visualization.

2.2.6 Summaries of posterior distribution

Highest posterior density (HPD) region

Visualization

- To visualize the HPD region, imagine drawing a horizontal line across the graph at the mode of the posterior distribution, and then "pushing down" the horizontal line until the corresponding values on the θ-axis trapped the appropriate probability.



2.2.6 Summaries of posterior distribution

Highest posterior density (HPD) region

- Finding HPD regions usually require the use of numerical methods. There are some packages in `R` that provide functions for finding HPD regions. We introduce two packages: `TeachingDemos` and `coda`.
- If the posterior comes from a known family of densities (e.g. beta) where the quantile function is available (e.g. `qbeta`), we can use the `hpd` function in `TeachingDemos`.

2.2.6 Summaries of posterior distribution

Highest posterior density (HPD) region

- The arguments of `hpd` are the name of the quantile function and its parameters, followed by the confidence level.

```
> a <- 3
> b <- 9

> require(TeachingDemos)
Loading required package: TeachingDemos

> hpd(qbeta, shape1 = a, shape2 = b, conf = 0.90)
[1] 0.05597708 0.43439374

> hpd(qbeta, shape1 = a, shape2 = b, conf = 0.95)
[1] 0.04055517 0.48372366
```

2.2.6 Summaries of posterior distribution

Highest posterior density (HPD) region

- Later in the semester, we will come across posterior distributions which do not belong to any known family of densities.
- In that case, we can use the `HPDinterval` function in the `coda` package to find the HPD region based on random samples from the posterior distribution.

2.2.6 Summaries of posterior distribution

Highest posterior density (HPD) region

Example using coda:

- Generate 1 million random samples from the beta distribution using rbeta.
- Create an mcmc object using mcmc.
- HPDinterval computes the HPD region using the mcmc object. Confidence level is specified under “prob”.

```
> a <- 3
> b <- 9
> require(coda)
Loading required package: coda

> rand_sp <- rbeta(10^6, a, b)
> mcmc_obj <- mcmc(rand_sp)
> HPDinterval(mcmc_obj, prob = 0.95)
          lower     upper
var1 0.03962348 0.4821268
attr(,"Probability")
[1] 0.95
```

2.2.6 Summaries of posterior distribution

Highest posterior density (HPD) region

- A HPD region might not be an interval if the posterior density is multimodal (having multiple peaks). The `HPDinterval` function in the `coda` package assumes that the distribution is not severely multimodal.

- In our example, the quantile-based CI is slightly wider $(0.52 - 0.06 = 0.46)$ than the HPD region $(0.48 - 0.04 = 0.44)$ even though both contain 95% of the posterior probability.

- Generally, the quantile-based CI will be equal to the HPD region if the posterior is symmetric and unimodal, but will be wider otherwise.

- In fact, it can be shown that for unimodal posterior densities, the HPD interval has the shortest length among all intervals with the same level of coverage (the same $1 - \alpha$).

Outline

Bayesian Inference

The Binomial Model

- Model, Prior, and Posterior
- Combining Information
- Conjugacy
- Sufficient Statistics
- Prediction
- Summaries of posterior distribution

The Poisson model

- Combining information
- Prediction

Mixtures of conjugate priors

Exponential families and conjugate priors

Summary

2.3 The Poisson model

- Some measurements, such as a person's number of friends, or the occurrence of certain event in a certain amount of time, have values that are non-negative integers. In these cases, the sample space is $\mathcal{Y} = \{0, 1, 2, \dots\}$. A simple probability model on $\mathcal{Y}$ is the Poisson model.

- A random variable Y has a Poisson distribution with mean θ, denoted as $Y \sim \text{Poisson}(\theta)$ if

$$\text{P}(Y = y|\theta) = \frac{\theta^y}{y!}\exp(-\theta) \text{ for } y \in \{0, 1, 2, \dots\}.$$

- $\text{E}(Y|\theta) = \text{Var}(Y|\theta) = \theta.$

2.3 The Poisson model

R commands for Poisson distribution

- The functions `dpois, ppois, qpois, rpois` provide the density, cumulative distribution function, quantile function and random generation respectively for the Poisson distribution in `R`.

Example: Suppose $Y \sim \text{Poisson}(1)$.

- Find $\text{P}(Y = 2)$: `dpois(x = 2, lambda = 1)` (Ans: 0.184)
- Find $\text{P}(Y \leq 2)$: `ppois(q = 2, lambda = 1)` (Ans: 0.920)
- Find the smallest integer α such that $P(Y \leq \alpha) \geq 0.7$: `qpois(p = 0.9, lambda = 1)` (Ans: 1)
- Generate 50 random numbers from Poisson(1): `rpois(n = 50, lambda = 1)`

2.3 The Poisson model

- Suppose we have n observations $y_1, \ldots, y_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\theta)$. The joint pdf of $\boldsymbol{y} = (y_1, \ldots, y_n)$ is

$$p(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} p(y_i|\theta) = \prod_{i=1}^{n} \left\{ \exp(-\theta) \frac{\theta^{y_i}}{y_i!} \right\}$$

$$= \exp(-n\theta) \theta^{\sum_{i=1}^{n} y_i} \frac{1}{\prod_{i=1}^{n} y_i!}$$

- From the Factorization Theorem, $t = \sum_{i=1}^{n} y_i$ is a sufficient statistic for θ. This is because $p(\boldsymbol{y}|\theta)$ is of the form $f(t, \theta) g(\boldsymbol{y})$, where $f(t, \theta) = \theta^t \exp(-n\theta)$ and $g(\boldsymbol{y}) = 1 / \prod_{i=1}^{n} y_i!$.

- Moreover, $T = \sum_{i=1}^{n} Y_i \sim \text{Poisson}(n\theta)$.

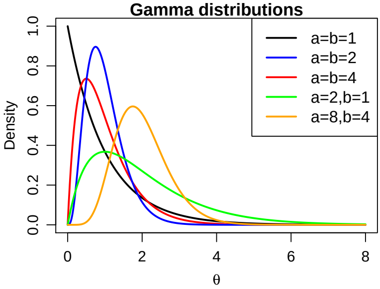
2.3 The Poisson model

Gamma distribution

- A random variable θ is said to follow the gamma distribution with shape parameter $a > 0$ and rate parameter $b > 0$, denoted $\theta \sim \text{Gamma}(a, b)$ if

$$p(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} \exp(-b\theta) \text{ for } \theta > 0.$$

- $\text{E}(\theta) = a/b$, $\text{Var}(\theta) = a/b^2$.
- Mode of θ is $(a-1)/b$ if $a > 1$ and 0 if $a \leq 1$.



R commands for gamma distribution

- The functions `dgamma, pgamma, qgamma, rgamma` provide the density, cumulative distribution function, quantile function and random generation respectively for the gamma distribution in R.

Example: Suppose $\theta \sim \text{Gamma}(68, 45)$.

- Find $p(1)$: `dgamma(x = 1, shape = 68, rate = 45)` (Ans: 0.021)
- Find $\text{P}(\theta \leq 2)$: `pgamma(q = 2, shape = 68, rate = 45)` (Ans: 0.993)
- Find α such that $P(\theta \leq \alpha) = 0.025$: `qgamma(p = 0.025, shape = 68, rate = 45)` (Ans: 1.17)
- Generate 50 random numbers from $\text{Gamma}(68, 45)$: `rgamma(n = 50, shape = 68, rate = 45)`

Posterior distribution

- Suppose $Y_1, \ldots, Y_n | \theta \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\theta)$ and $\theta \sim \text{Gamma}(a_0, b_0)$. Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ and $t = \sum_{i=1}^{n} y_i$. Then

$$\begin{aligned} p(\theta|\boldsymbol{y}) = p(\theta|t) &\propto p(t|\theta)p(\theta) \\ &\propto \theta^t \exp(-n\theta) \cdot \theta^{a_0 - 1} \exp(-b_0\theta) \\ &\propto \theta^{a_0 + t - 1} \exp(-(b_0 + n)\theta). \end{aligned}$$

- Therefore, the posterior is again a gamma distribution and

$$\begin{matrix} \theta \sim \text{Gamma}(a_0, b_0) \\ Y_1, \ldots, Y_n | \theta \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\theta) \end{matrix} \quad \Rightarrow \quad \theta|\boldsymbol{y} \sim \text{Gamma}(a, b),$$

where $a = a_0 + t$ and $b = b_0 + n$.

2.3.1 Combining information

$$\mathrm{E}(\theta|\boldsymbol{y}) = \frac{a}{b} = \frac{a_0 + t}{b_0 + n} = \frac{\cancel{b_0}}{b_0 + n} \underbrace{\frac{a_0}{\cancel{b_0}}}_{\text{prior mean}} + \frac{\cancel{n}}{b_0 + n} \underbrace{\frac{t}{\cancel{n}}}_{\text{sample mean}} .$$

$$\mathrm{Var}(\theta|\boldsymbol{y}) = \frac{a}{b^2} = \frac{a_0 + t}{(b_0 + n)^2}.$$

- The posterior mean of θ is a weighted average of the prior mean and sample mean.
- We can interpret
  - b_0 as the number of prior observations
  - a_0 as the sum of counts from b_0 prior observations.
- When n is large, $\mathrm{E}(\theta|\boldsymbol{y}) \approx t/n = \bar{y}$, where $\bar{y}$ is the sample mean.
- Similarly, we can show that $\mathrm{Var}(\theta|\boldsymbol{y}) \approx \bar{y}/n$.

2.3.2 Prediction

- The posterior predictive distribution for a future observation $\tilde{Y}$ is

$$\mathrm{P}(\tilde{Y}=\tilde{y}|\boldsymbol{y}) = \int_0^\infty \mathrm{P}(\tilde{Y}=\tilde{y}|\theta)\, p(\theta|\boldsymbol{y})\, \mathrm{d}\theta$$

$$= \int_0^\infty \mathrm{e}^{-\theta}\frac{\theta^{\tilde{y}}}{\tilde{y}!}\cdot\frac{b^a}{\Gamma(a)}\theta^{a-1}\mathrm{e}^{-b\theta}\, \mathrm{d}\theta$$

$$= \frac{b^a\Gamma(a+\tilde{y})}{\Gamma(a)\tilde{y}!(b+1)^{a+\tilde{y}}}\underbrace{\int_0^\infty \frac{(b+1)^{a+\tilde{y}}}{\Gamma(a+\tilde{y})}\theta^{a+\tilde{y}-1}\mathrm{e}^{-(b+1)\theta}\, \mathrm{d}\theta}_{=1}$$

$$= \frac{\Gamma(a+\tilde{y})}{\Gamma(a)\tilde{y}!}\left(\frac{b}{b+1}\right)^a\left(\frac{1}{b+1}\right)^{\tilde{y}}.$$

where $\tilde{y}$ can take any value in $\{0, 1, 2, \ldots\}$.

2.3.2 Prediction

Predictive distribution

- Note that $\Gamma(n) = (n-1)!$ if n is a positive integer.
- Thus if a is a positive integer, then

$$\begin{aligned}
\mathrm{P}(\tilde{Y} = \tilde{y} | \boldsymbol{y}) &= \frac{(a+\tilde{y}-1)!}{(a-1)!\tilde{y}!} \left(\frac{b}{b+1}\right)^{a} \left(\frac{1}{b+1}\right)^{\tilde{y}} \\
&= \binom{a+\tilde{y}-1}{\tilde{y}} \left(\frac{b}{b+1}\right)^{a} \left(\frac{1}{b+1}\right)^{\tilde{y}} .
\end{aligned}$$

This function is the pdf of a negative binomial distribution.

2.3.2 Prediction

Negative binomial distribution

- In a sequence of independent Bernoulli trials with probability of success p, let the random variable X be the number of failures until the rth success. Then X follows a negative binomial distribution with parameters r and p, denoted as

$$X \sim \mathrm{NB}(r, p),$$

where
$$\mathrm{P}(X = x) = \binom{r + x - 1}{x} p^r (1 - p)^x, \quad x = 0, 1, , 2, \ldots,$$

- $\mathrm{E}(X) = \dfrac{r(1-p)}{p}$ and $\mathrm{Var}(X) = \dfrac{r(1-p)}{p^2}$.
- The functions `dnbinom, pnbinom, qnbinom, rnbinom` provide the density, cumulative distribution function, quantile function and random generation respectively for the $\mathrm{NB}(r, p)$ distribution in `R`.
- They work in a similar manner as corresponding functions for the binomial distribution. Set `size` to be r and `prob` to be p.

2.3.2 Prediction

Predictive distribution

- Therefore

$$\tilde{Y}|\boldsymbol{y} \sim \text{NB}\left(a, \frac{b}{b+1}\right).$$

$$\text{E}(\tilde{Y}|\boldsymbol{y}) = a\frac{1/(b+1)}{b/(b+1)} = \frac{a}{b} = \text{E}(\theta|\boldsymbol{y}).$$

$$\begin{aligned}\text{Var}(\tilde{Y}|\boldsymbol{y}) &= a\frac{1/(b+1)}{b^2/(b+1)^2} = \frac{a(b+1)}{b^2}.\\ &= \text{Var}(\theta|\boldsymbol{y})(b+1) = \text{E}(\theta|\boldsymbol{y})\frac{b+1}{b}.\end{aligned}$$

- Note that $\text{E}(\theta|\boldsymbol{y}) = a/b = (a_0 + \sum_{i=1}^{n} y_i)/(b_0 + n)$ and $\text{Var}(\theta|\boldsymbol{y}) = a/b^2 = (a_0 + \sum_{i=1}^{n} y_i)/(b_0 + n)^2$.

2.3.2 Prediction

Predictive distribution

- The predictive variance is a measure of the uncertainty about a new sample $\tilde{Y}$ from the population.

- Uncertainty about $\tilde{Y}$ comes from two sources:
  1. uncertainty about the population and
  2. the variability in sampling from the population.

- For large n, uncertainty about θ is small and uncertainty about $\tilde{Y}$ stems primarily from sampling variability, which for the Poisson model is equal to θ (since $\frac{b+1}{b} = \frac{b_0+n+1}{b_0+n} \approx 1$).

- For small n, uncertainty about $\tilde{Y}$ also includes the uncertainty about θ, and so the total uncertainty is larger than just the sampling variability (since $\frac{b_0+n+1}{b_0+n} > 1$).

Outline

Bayesian Inference

The Binomial Model

- Model, Prior, and Posterior
- Combining Information
- Conjugacy
- Sufficient Statistics
- Prediction
- Summaries of posterior distribution

The Poisson model

- Combining information
- Prediction

Mixtures of conjugate priors

Exponential families and conjugate priors

Summary

2.4 Mixtures of conjugate priors

- While a single conjugate prior may be inadequate to accurately reflect prior knowledge, a finite mixture of conjugate priors may be sufficiently flexible (allowing multimodality, heavier tails, etc) while still enabling simplified posterior calculations.
- Suppose we have a likelihood $p(\theta|\boldsymbol{y})$ and prior densities, $p_1(\theta)$ and $p_2(\theta)$, from a conjugate family $\mathcal{P}$, which give rise to posteriors $p_1(\theta|\boldsymbol{y})$ and $p_2(\theta|\boldsymbol{y})$ respectively.
- Consider the mixture prior

$$p(\theta) = \alpha p_1(\theta) + (1-\alpha)p_2(\theta).$$

where $0 \leq \alpha \leq 1$. Can you show that $p(\theta)$ is a well-defined probability density function?

2.4 Mixtures of conjugate priors

- The posterior corresponding to the prior $p(\theta)$ is

$$\begin{aligned}
p(\theta|\boldsymbol{y}) &= \frac{p(\boldsymbol{y}|\theta)p(\theta)}{\int p(\boldsymbol{y}|\theta')p(\theta')\ \mathrm{d}\theta'} \\
&= \frac{\alpha p(\boldsymbol{y}|\theta)p_1(\theta) + (1-\alpha)p(\boldsymbol{y}|\theta)p_2(\theta)}{\alpha \int p(\boldsymbol{y}|\theta')p_1(\theta')\ \mathrm{d}\theta' + (1-\alpha)\int p(\boldsymbol{y}|\theta')p_2(\theta')\ \mathrm{d}\theta'} \\
&= \frac{\alpha m_1(\boldsymbol{y})p_1(\theta|\boldsymbol{y}) + (1-\alpha)m_2(\boldsymbol{y})p_2(\theta|\boldsymbol{y})}{\alpha m_1(\boldsymbol{y}) + (1-\alpha)m_2(\boldsymbol{y})} \\
&= w_1 p_1(\theta|\boldsymbol{y}) + w_2 p_2(\theta|\boldsymbol{y}).
\end{aligned}$$

where $m_i(\boldsymbol{y}) = \displaystyle\int p_i(\theta)p(\boldsymbol{y}|\theta)\ \mathrm{d}\theta$ for $i = 1, 2$,

$w_1 = \dfrac{\alpha m_1(\boldsymbol{y})}{\alpha m_1(\boldsymbol{y}) + (1-\alpha)m_2(\boldsymbol{y})}$ and $w_2 = 1 - w_1$.

- Note that $p_i(\theta|\boldsymbol{y}) = p(\boldsymbol{y}|\theta)p_i(\theta)/m_i(\boldsymbol{y})$, for $i = 1, 2$.

2.4 Mixtures of conjugate priors

- Thus a mixture of conjugate priors leads to a mixture of respective posteriors.

- If we have a likelihood $p(\theta|\boldsymbol{y})$ and $p_1(\theta|\boldsymbol{y})$ and $p_2(\theta|\boldsymbol{y})$ are the posteriors corresponding to the priors, $p_1(\theta)$ and $p_2(\theta)$, from a conjugate family $\mathcal{P}$, then

$$p(\theta) = \alpha p_1(\theta) + (1-\alpha)p_2(\theta)$$
$$\Rightarrow \begin{cases} p(\theta|\boldsymbol{y}) = w_1 p_1(\theta|\boldsymbol{y}) + w_2 p_2(\theta|\boldsymbol{y}), \\ w_1 = \dfrac{\alpha m_1(\boldsymbol{y})}{\alpha m_1(\boldsymbol{y}) + (1-\alpha)m_2(\boldsymbol{y})} \\ w_2 = 1 - w_1. \end{cases}$$

2.4 Mixtures of conjugate priors

- More generally, it is possible to take any convex combination of more than two priors in $\mathcal{P}$ and get a corresponding convex combination of the respective posteriors.
- A convex combination is a linear combination of points where all coefficients are non-negative and sum to 1.

2.4 Mixtures of conjugate priors

Example: Spinning a coin

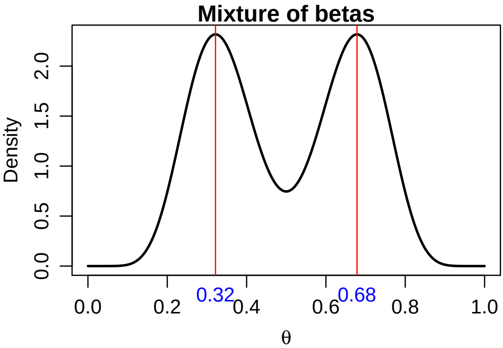
- Diaconis and Ylvisaker (1985) observe that there is a big difference between spinning a coin on a table and tossing it in the air.
- While tossing often leads to an even proportion of 'heads' and 'tails', spinning often leads to proportions like 1/3 or 2/3 (the shape of the edge of a coin is a strong determining factor for this bias).
- Let θ denote the proportion of 'heads' for a particular coin. Suppose we consider a prior for θ that is a fifty-fifty mixture of two beta densities $\text{Beta}(10, 20)$ and $\text{Beta}(20, 10)$.
- The function "beta_mixture" computes the pdf of the mixture of betas prior.

```
> beta_mixture <- function(theta){
+   0.5*dbeta(theta,10,20) + 0.5*dbeta(theta,20,10)  }
```

2.4 Mixtures of conjugate priors

Example: Spinning a coin

- This mixture of beta densities prior is a bimodal distribution, with modes close to the modes, 0.32 and 0.68, of its components.



2.4 Mixtures of conjugate priors

Example: Spinning a coin

- The modes of the mixture prior can be found using the `optimize` function. To capture the first and second mode, we specify the search interval to be (0.2, 0.5). Then we change the search interval to (0.5, 0.7) to capture the second mode.

```
> optimize(f=beta_mixture, interval=c(0.2,0.5), maximum=TRUE)
$maximum
[1] 0.3216435
$objective
[1] 2.317263

> optimize(f=beta_mixture, interval=c(0.5,0.7), maximum=TRUE)
$maximum
[1] 0.6783853
$objective
[1] 2.317263
```

- `optimize` is for 1-dimensional functions. `optim` is for multi-dimensional functions.

2.4 Mixtures of conjugate priors

Example: Spinning a coin

- Suppose the coin is then spun ten times, of which three were "heads". Let Y denote the number of heads out of ten spins. Therefore $y = 3$ and

$$Y \sim \text{Binomial}(10, \theta).$$

- The posterior is

$$p(\theta|y) = w_1 p_1(\theta|y) + w_2 p_2(\theta|y),$$

  where

  - $p_1(\theta|y)$ is the posterior corresponding to the Beta(10, 20) prior, Beta(13, 27).
  - $p_2(\theta|y)$ is the posterior corresponding to the Beta(20, 10) prior, Beta(23, 17).

2.4 Mixtures of conjugate priors

Example: Spinning a coin

- The marginal distribution corresponding to a $\mathrm{Beta}(a_0, b_0)$ prior is

$$\begin{aligned} m(y) &= \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} \frac{\theta^{a_0-1}(1-\theta)^{b_0-1}}{\mathrm{B}(a_0, b_0)} \mathrm{d}\theta \\ &= \binom{n}{y} \frac{\mathrm{B}(a_0+y, b_0+n-y)}{\mathrm{B}(a_0, b_0)}. \end{aligned}$$

$$\therefore m_1(y) = \binom{10}{3} \frac{\mathrm{B}(13, 27)}{\mathrm{B}(10, 20)}, \quad m_2(y) = \binom{10}{3} \frac{\mathrm{B}(23, 17)}{\mathrm{B}(20, 10)}.$$

- The coefficients are

$$w_1 = \frac{0.5\mathrm{B}(13, 27)}{0.5\mathrm{B}(13, 27) + 0.5\mathrm{B}(23, 17)} = \frac{115}{129}, \quad w_2 = 1 - \frac{115}{129} = \frac{14}{129}.$$

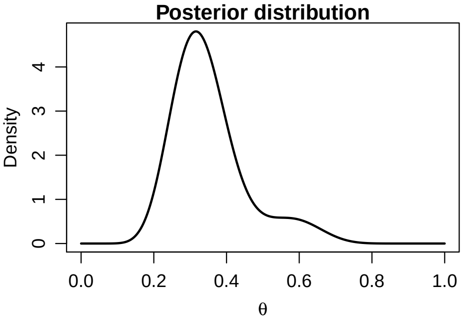
- Note that $\binom{10}{3}$ and $\mathrm{B}(10, 20) = \mathrm{B}(20, 10)$ cancel out.

- $\mathrm{B}(a, b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ and $\Gamma(n) = (n-1)!$ for integer-valued n.

2.4 Mixtures of conjugate priors

Example: Spinning a coin

- The posterior $p(\theta|y) = \dfrac{115}{129} p_1(\theta|y) + \dfrac{14}{129} p_2(\theta|y)$ is plotted below.



- Properties of this posterior can be deduced conveniently from those of the component betas. For example,

$$\begin{aligned}\mathsf{E}(\theta|y) &= \int_0^1 \theta p(\theta|y)\mathrm{d}\theta \\ &= \int_0^1 \theta \left\{ \frac{115}{129} p_1(\theta|y) + \frac{14}{129} p_2(\theta|y) \right\} \mathrm{d}\theta \\ &= \frac{115}{129}\mathsf{E}_1(\theta|y) + \frac{14}{129}\mathsf{E}_2(\theta|y),\end{aligned}$$

where $\mathsf{E}_1(\theta|y)$ and $\mathsf{E}_2(\theta|y)$ denote the respective means of the posterior distributions $p_1(\theta|y)$ and $p_2(\theta|y)$.

Outline

Bayesian Inference

The Binomial Model

- Model, Prior, and Posterior
- Combining Information
- Conjugacy
- Sufficient Statistics
- Prediction
- Summaries of posterior distribution

The Poisson model

- Combining information
- Prediction

Mixtures of conjugate priors

Exponential families and conjugate priors

Summary

2.5 Exponential families and conjugate priors

- The binomial and Poisson models discussed in this chapter are examples of one-parameter exponential family models.

- A one-parameter exponential family model is any model whose density can be expressed as

$$p(y|\theta) = h(y)g(\theta)\exp\{\eta(\theta)t(y)\}, \tag{2}$$

  where θ is the parameter of the family and $t(y)$ is the sufficient statistic for θ.

2.5 Exponential families and conjugate priors

Example: Binomial model

- If $Y \overset{\text{i.i.d.}}{\sim} \text{Binomial}(\theta)$,

$$
\begin{aligned}
p(y|\theta) &= \binom{n}{y} \theta^y (1-\theta)^{n-y} \\
&= \binom{n}{y} (1-\theta)^n \exp\left\{ y \log\left(\frac{\theta}{1-\theta}\right)\right\},
\end{aligned}
$$

  where $h(y) = \binom{n}{y}$, $g(\theta) = (1-\theta)^n$, $\eta(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$ and $t(y) = y$.

- Therefore the binomial model belongs to the one-parameter exponential family.

2.5 Exponential families and conjugate priors

Example: Poisson model

- If $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim}$ Poisson(θ),

$$p(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} \left\{ \frac{\theta^{y_i}}{y_i!} \exp(-\theta) \right\}$$

$$= \frac{1}{\prod_{i=1}^{n} y_i!} \exp(-n\theta) \exp\left(\log(\theta) \sum_{i=1}^{n} y_i \right),$$

where $h(y) = 1/\prod_{i=1}^{n} y_i!$, $g(\theta) = \exp(-n\theta)$, $\eta(\theta) = \log(\theta)$ and $t(\boldsymbol{y}) = \sum_{i=1}^{n} y_i$.

- Thus $p(\boldsymbol{y}|\theta)$ belongs to the one-parameter exponential family.

2.5 Exponential families and conjugate priors

- When a model belongs to the one-parameter exponential family in (2), a family of conjugate prior distributions is given by

  $$p(\theta) \propto g(\theta)^{\nu} \exp\{\eta(\theta)\tau\},$$

  where ν and τ are parameters of the prior.
- ν and τ are allowed to take values for which $p(\theta)$ is a well-defined pdf (e.g. $p(\theta) \geq 0$ for $\theta \in \Theta$ and $\int_{\Theta} p(\theta) \, \mathrm{d}\theta = 1$).

2.5 Exponential families and conjugate priors

- Combining such prior with the sampling model $Y \sim p(y|\theta)$ yields the posterior:

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &\propto g(\theta)\exp\{\eta(\theta)t(y)\} \cdot g(\theta)^{\nu}\exp\{\eta(\theta)\tau\} \\ &\propto g(\theta)^{\nu+1}\exp\{\eta(\theta)[\tau + t(y)]\} \end{aligned}$$

  which belongs to the same family as the prior distribution with parameters $\nu + 1$ and $\tau + t(y)$.

2.5 Exponential families and conjugate priors

Example: binomial model

- If $Y \sim \text{Binomial}(n, \theta)$, $p(y|\theta) \propto (1-\theta)^n \exp\left\{ y \log\left(\frac{\theta}{1-\theta}\right)\right\}$.
- The conjugate prior is given by

$$\begin{aligned} p(\theta) &\propto (1-\theta)^{n\nu} \exp\left\{ \tau \log\left(\frac{\theta}{1-\theta}\right)\right\} \\ &\propto \theta^{\tau}(1-\theta)^{n\nu-\tau} \end{aligned}$$

  which is a Beta(a, b) distribution where $a = \tau + 1$, $b = n\nu - \tau + 1$. Note that the prior is well-defined only for $a > 0$, $b > 0$. This implies $\tau > -1$ and $\nu > \frac{\tau - 1}{n}$.
- Hence we see again that the beta family provides a class of conjugate priors for the binomial model.

Outline

Bayesian Inference

The Binomial Model

- Model, Prior, and Posterior
- Combining Information
- Conjugacy
- Sufficient Statistics
- Prediction
- Summaries of posterior distribution

The Poisson model

- Combining information
- Prediction

Mixtures of conjugate priors

Exponential families and conjugate priors

Summary

2.6 Summary

- In this chapter, we have studied
  - Bayesian inference for the binomial model and Poisson model
  - Posterior predictive distributions
  - Summaries of posterior distributions: point and interval estimates
  - Conjugate priors and mixtures of conjugate priors
  - Sufficient statistics and exponential families