Bayesian Modeling

One-parameter models

Yixin Wang

3 One-parameter models

A one-parameter model is a class of sampling distribution that is indexed by a single unknown parameter. We will study Bayesian inference with several such models. Although simple, they will help to illustrate several key concepts in Bayesian data analysis, including conjugate priors, predictive distributions and confidence regions.

3.1 The binomial model

Example 3.1. (Happiness data) In a General Social Survey conducted in 1998, each female of age 65 or over was asked whether or not they were generally happy or not. Let $Y_i = 1$ if respondent i reported being generally happy, and 0 otherwise. The label i is given arbitrarily before the data are collected; we do not assume to have any further information distinguishing these individuals. As before, we use $p(y_1, \ldots, y_n)$ as the shorthand notation for $\Pr(Y_1 = y_1, \ldots, Y_n = y_n)$ and so on.

We shall assume a binomial distribution to describe our sampling model. Associated with this model is a parameter $\theta \in [0, 1]$ and that

$$Y_1, \ldots, Y_n | \theta \overset{i.i.d.}{\sim} \text{Bernoulli}(\theta).$$

Accordingly,

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

It is reported that out of $n = 129$ respondents, 118 individuals report being generally happy (91%), and 11 individuals do not report being generally happy (9%).

Uniform prior To continue with Bayesian analysis, we need to give θ a prior distribution. Let us take the uniform prior, so that

$$p(\theta) = 1 \text{ for all } \theta \in [0,1].$$

Uniform prior is considered a ”vague” or ”non-informative” prior, and referred as such in the literature. [whether it is truly non-informative is a different matter!] Now, we are ready to apply the Bayes’ rule to obtain

$$p(\theta|y_1,\ldots,y_{129}) \propto p(\theta)p(y_1,\ldots,y_{129}|\theta) = \theta^{118}(1-\theta)^{11}.$$

In the above expression, we drop the normalizing constant, which is $p(y_1,\ldots,y_{129})$.

To find the mode of the posterior distribution, we need to solve the optimization problem

$$\max_{\theta\in[0,1]} \log\{\theta^{118}(1-\theta)^{11}\}.$$

Taking derivative with respect to θ and setting to zero, we obtain the maximizer to be $\hat{\theta} = 118/129 = .91$, the fraction of respondents who report being generally happy. The reader might think: so much for all the math, only to get such an obvious answer?

But what about other quantities relevant to the posterior distribution of θ? The normalizing constant of the posterior density is $p(y_1,\ldots,y_{129})$. Why in general this quantity is difficult to calculate, for this specific example it has a closed form: the expression defining the posterior distribution should remind us of the beta distribution. A beta pdf is defined on $[0,1]$ and takes the form

$$p(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}. \qquad (4)$$

Here $a,b>0$ are the parameters. Since the density function integrates to one, this implies that

$$\int_0^1 \theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Exercise 3.1. Based on the above identity, prove the following: under the beta distribution beta(a,b)

$$\begin{aligned}
\text{mode}[\theta] &= (a-1)/[(a-1)+(b-1)]\text{if}a>1,b>1,\\
\mathbb{E}[\theta] &= a/(a+b),\\
\text{Var}[\theta] &= ab/[(a+b+1)(a+b)^2].
\end{aligned}$$

□

Back to our example, then we have

$$p(y) = \int_0^1 \theta^{118}(1-\theta)^{11}\mathrm{d}\theta = \frac{\Gamma(119)\Gamma(12)}{\Gamma(131)}.$$

In fact, the posterior distribution of θ is indeed beta$(119,12)$.

Beta prior The uniform distribution of $[0, 1]$ is an instance of the beta distribution for $a = b = 1$. Employing the beta prior instead, and apply Bayes' rule

$$\begin{aligned} p(\theta|y_1, \ldots, y_n) &\propto p(\theta)p(y_1, \ldots, y_n|\theta) \\ &\propto \theta^{a-1}(1-\theta)^{b-1} \times \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} \\ &= \text{beta}(\theta|a + \sum_{i=1}^n y_i, b + n - \sum_{i=1}^n y_i). \end{aligned}$$

This is an instance of *conjugacy*: a beta prior, when combined with a binomial likelihood, yields a beta posterior distribution. Conjugacy is the property of a prior relative to a given likelihood: a prior is conjugate with respect to a likelihood if the resulting posterior distribution takes the same form.

Conjugacy a treasured property in Bayesian statistics because it simplifies posterior computation, a considerable bottleneck. Once we know the form of the posterior density, we only need to concern with the posterior distribution's parameters, which reflects the posterior updates that combines both prior information and the information gleaned from the data.

Example 3.2. In Example 3.1, the posterior distribution of θ receives the update from the data via the statistic $\sum_{i=1}^{n} y_i$. This reflects the fact that $\sum_{i=1}^{n} Y_i$ is the *sufficient statistic for* θ under the Bernoulli sampling model. In our Bayesian framework, we may express this as

$$p(\theta|Y_1, \ldots, Y_n) = p(\theta|\sum_{i=1}^{n} Y_i).$$

In other words, the information contained in the observed in the data $\{Y_1 = y_1, \ldots, Y_n = y_n\}$ is the same as the information contained in $Y = y$, where $Y = \sum_{i=1}^{n} Y_i$ and $y = \sum_{i=1}^{n} y_i$.

Alternatively, we may consider a sampling model in which the data are the count of people who report to be ”generally happy”, as opposed to ”not generally happy”. Thus, instead of the Bernoulli sampling model, the suitable sampling model is a binomial distribution. Applying the same computation as above, the reader should be able to derive that if we posit

$$\begin{aligned} \text{prior: } \theta &\sim \text{beta}(a, b) \\ \text{sampling: } Y = y &\sim \text{binomial}(n, \theta), \end{aligned}$$

then by the Bayes’ rule we obtain

$$\text{posterior: } \theta | Y = y \sim \text{beta}(a + y, b + n - y).$$

This is also the calculation that we relied on in Example 1.1.

Prediction After having obtained data sample $\{y_1, \ldots, y_n\}$ we are also interested in the distribution of new observations. This is called the *predictive distribution*. Suppose that $\tilde{Y}$ is an additional outcome of the same population as the observed sample via the sampling model

$$Y_1, \ldots, Y_n, \tilde{Y}|\theta \overset{i.i.d.}{\sim} p(.|\theta).$$

Under the prior distribution

$$\theta \sim p(\theta)$$

the predictive distribution of $\tilde{Y}$ given $\{Y_1 = y_1, \ldots, Y_n = y_n\}$ takes the form

$$\begin{aligned} p(\tilde{Y} = \tilde{y}|y_1, \ldots, y_n) &= \int p(\tilde{y}, \theta|y_1, \ldots, y_n)\mathrm{d}\theta \\ &= \int p(\tilde{y}|\theta, y_1, \ldots, y_n)p(\theta|y_1, \ldots, y_n)\mathrm{d}\theta \\ &= \int p(\tilde{y}|\theta)p(\theta|y_1, \ldots, y_n)\mathrm{d}\theta. \end{aligned}$$

The last identity is due to the i.i.d. assumption in the sampling model.

Some remarks

- The predictive distribution depends on the observed data. It does not depend on the unknown θ.
- The unknown θ is integrated out in the formula via the posterior distribution. Thus the predictive distribution takes into account both the observed data and the prior distribution.
- Contrast this with a frequentist approach: one can obtain a point-estimate $\hat{\theta}$ based on the observed data, and then plug-in the sampling model to produce a predictive distribution of new observation:

  $$p_{\text{plug-in}}(\tilde{Y} = \tilde{y}) := p(\tilde{y}|\hat{\theta}).$$

  Because the Bayesian approach relies on a distribution over the unknown θ rather than a single numerical value of θ, it allows for a broader range of predictive distributions than a plug-in approach.

Example 3.3. Continue from Example 3.1 (Binomial sampling and uniform prior). We use the uniform distribution as the prior for happiness level θ. The uniform distribution is beta(a, b), where $a = b = 1$. The predictive distribution of the answer "I'm generally happy" for the next respondent is

$$\begin{aligned}\Pr(\tilde{Y} = 1|y_1, \ldots, y_n) &= \int p(1|\theta)p(\theta|y_1, \ldots, y_n)\mathrm{d}\theta \\ &= \int \theta p(\theta|y_1, \ldots, y_n) \\ &= \frac{a + \sum_{i=1}^n y_i}{a + b + n}.\end{aligned}$$

Suppose that out of 20 people, none is reportedly happy, then the probability that the next person is reportedly happy will be $a/(a + b + 20) = 1/22$.

Contrast this with the non-Bayesian plug-in approach: the mode of $p(\theta|y_1, \ldots, y_n)$ is the same as the mode of the likelihood function $p(y_1, \ldots, y_n|\theta)$, which is equal

$$\frac{a + \sum_{i=1}^{n} y_i - 1}{a + b + n - 2} = 0.$$

If we plug in $\hat{\theta} = 0$, then the predictive probability that the next person is reportedly happy will be 0.

3.2 Confidence regions

It is of interest to identify regions of the parameter space that are likely to contain the true value of the unknown parameter. The following definition for scalar parameter can be extended to multidimensional domains.

Definition 3.1 (Bayesian coverage)**.** An interval $[l(y), u(y)]$, based on the data observed data $Y = y$, has 95% Bayesian coverage for θ if

$$\Pr(l(y) < \theta < u(y)|Y = y) = .95.$$

Note: in the above probability expression, it is θ that is random, $Y = y$ fixed. Interpretation: having observed the data and calculated the conditional probability, the unknown θ is in the given interval with probability 95%.

Frequentist approach provides point estimates for unknown θ, not a distribution. To quantify for the uncertainty of the estimate, there is a notion of confidence interval defined as follows.

Definition 3.2 (Frequentist coverage)**.** A *random* interval $[l(Y), u(Y)]$ has 95% frequentist coverage for θ if, *before* the data are gathered,

$$\Pr(l(Y) < \theta < u(Y)|\theta) = .95.$$

Note: in the above probability expression, it is Y that is random, θ is unknown but fixed. Once you observe $Y = y$, you cannot provide any guarantee for $[l(y), u(y)]$ regarding the unknown θ. What frequentist coverage means is: if we are to run a large number of unrelated (independent) experiments and create the interval $[l(y), u(y)]$ for each one of them, then we can expect that 95% of the intervals contain the correct parameter value.

Some remarks

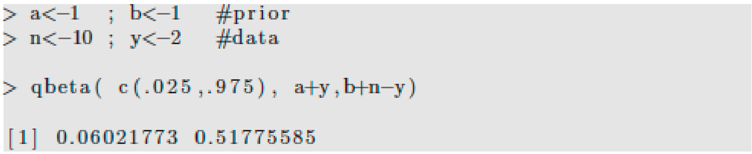
- Both notions are useful.
- The frequentist coverage describes the pre-experiment coverage, i.e., it promises a guarantee if the experiments are to be repeated many times in the future.
- The Bayesian coverage describes the post-experiment coverage, i.e., it is applicable to the data at hand, under a prior specification.
- When sample size gets large, usually the two coverages tend toward the same interval.

Example 3.4. Suppose that $Y = y$ represents a text document, while θ represents the probability that y is generated by a generative AI tool such as ChatGPT. A bot detector is designed to estimate not only θ, but also provide a coverage in the form of interval $[(l(y), u(y))]$ for θ. Discuss how the two notions of uncertainty quantification are distinct and can be useful in different ways.

Quantile-based interval This is the easiest way to obtain a Bayesian coverage: take $l(y) := \theta_{\alpha/2}$ and $u(y) := \theta_{1-\alpha/2}$, the left and right threshold for the $\alpha/2$ probability tail of the posterior distribution:

$$\Pr(\theta < \theta_{\alpha/2}|Y = y) = \Pr(\theta > \theta_{1-\alpha/2}|Y = y) = \alpha/2.$$

In R programming language:



```
> a<-1  ; b<-1   #prior
> n<-10 ; y<-2   #data

> qbeta( c(.025,.975), a+y,b+n-y)

[1] 0.06021773 0.51775585
```

A potential problem with this interval is that some θ-values outside the quantile-based interval may have higher probability (density) than some points inside the interval. In addition, for multi-modal posterior distribution (having multiple peaks), this choice of interval may be not very useful.

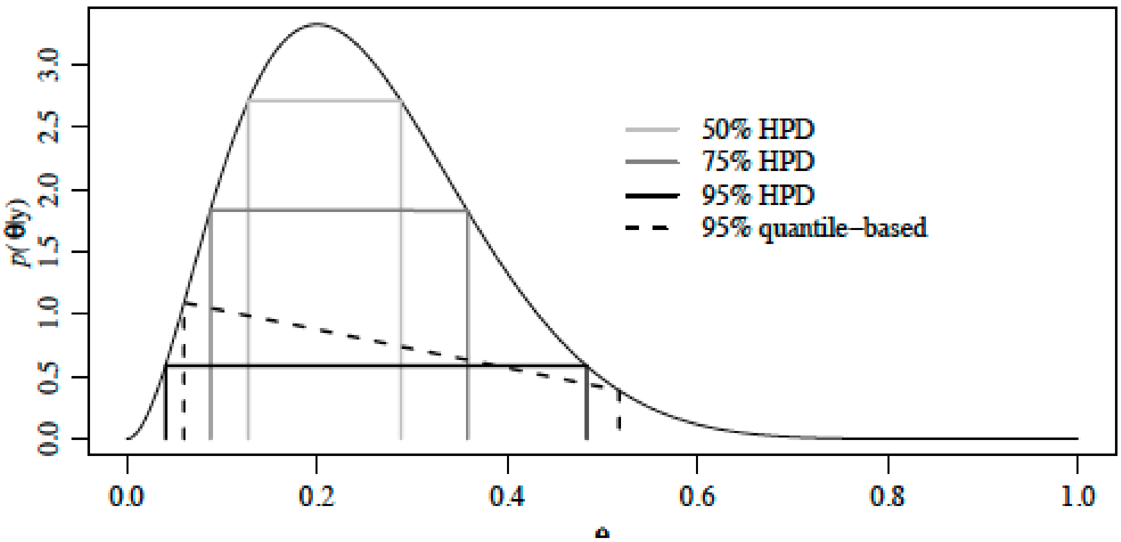


Figure 3.1: Quantile-based interval and highest posterior density regions.

An alternative is the so-called "highest posterior density (HPD)" region: it is the subset $s(y) \subset \Theta$ such that

(i) $\Pr(\theta \in s(y)|Y = y) = 1 - \alpha$.

(ii) If $\theta_a \in s(y)$ and $\theta_b \notin s(y)$, then $p(\theta_a|Y = y) > p(\theta_b|Y = y)$.

See Fig. 3.1 for an illustration. The HPD is characterized by threshold $c > 0$ of the posterior density. By sliding the threshold up and down the real axis we obtain different α. When the posterior density is a multi-modal function, the HPD may be composed of multiple disconnected subsets.

3.3 The Poisson model

Poisson is a probability distribution whose domain is the unbounded set of natural numbers. It is a useful modeling tool for count data.

Consider the Poisson sampling model: $Y|\theta \sim$ Poisson(θ). That is, for $y = 0, 1, \ldots$,

$$\Pr(Y = y|\theta) = \theta^y e^{-\theta}/y!.$$

Poisson random variables have an interesting feature in that both the mean and the variance are determined by the same parameter θ and in fact, $\mathbb{E}[Y|\theta] = \mathrm{Var}[Y|\theta] = \theta$.

Given n-i.i.d. sample: $Y_1, \ldots, Y_n|\theta \overset{iid}{\sim}$ Poisson(θ). We have

$$\begin{aligned}
\Pr(Y_1 = y_1, \ldots, Y_n = y_n|\theta) &= \prod_{i=1}^{n} p(y_i|\theta) \\
&= \prod_{i=1}^{n} \theta^{y_i} e^{-\theta}/y_i! \\
&=: c(y_1, \ldots, y_n)\theta^{\sum_i y_i} e^{-n\theta}.
\end{aligned}$$

From the above expression we find that $\sum_{i=1}^n Y_i$ is the sufficient statistic of the Poisson sampling model. Moreover, it can be verified that $\sum_{i=1}^n Y_i|\theta \sim$ Poisson$(n\theta)$.

We proceed to give a prior distribution for $\theta \in \mathbb{R}_+$. By Bayes' rule, we know that a prior pdf $p(\theta)$ yields the posterior pdf of the form

$$\begin{array}{rcl} p(\theta|y_1, \ldots, y_n) & \propto & p(\theta)p(y_1, \ldots, y_n|\theta) \\ & \propto & p(\theta)\theta^{\sum_i y_i} e^{-n\theta}. \end{array}$$

If we want a conjugate prior, then $p(\theta)$ must be of the form $\theta^{c_1} e^{-c_2\theta}$, up to a multiplying constant. The pdf that has this form is given by the Gamma distribution.

Gamma distribution Endow θ with the Gamma prior: $\theta|a,b \sim \text{Gamma}(a,b)$, for some (hyper) parameters $a, b > 0$:

$$p(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta}.$$

a is called the shape parameter, and b the rate parameter of Gamma distributions.

With this prior in place, the posterior pdf takes the form

$$p(\theta|y_1, \ldots, y_n) \propto \theta^{a+\sum_i y_i - 1} e^{-(b+n)\theta}.$$

The proportional operator simplifies the expression by allowing us to keep only terms that vary with θ. This shows that the posterior pdf of another Gamma distribution. In other words, we have shown that the Gamma is a conjugate prior with respect to the Poisson sampling/ likelihood model:

$$\theta|Y_1, \ldots, Y_n \sim \text{Gamma}(a + \sum_{i=1}^{n} Y_i, b+n).$$

Based on basic properties of the Gamma distribution, we find

$$\begin{aligned}
\mathbb{E}[\theta|y_1, \ldots, y_n] &= \frac{a + \sum y_i}{b+n} \\
&= \frac{b}{b+n}(a/b) + \frac{n}{b+n} \sum y_i/n \\
\text{Var}[\theta|y_1, \ldots, y_n] &= \frac{a + \sum y_i}{(b+n)^2}.
\end{aligned}$$

We find that the posterior mean is, again, a convex combination of the prior expectation and the sample average. Note the impact of increasing the sample size n.

We proceed to the posterior predictive distribution. For $\tilde{y} = 0, 1, 2, \ldots$,

$$\begin{aligned}
p(\tilde{y}|y_1, \ldots, y_n) &= \int_0^\infty p(\tilde{y}|\theta, y_1, \ldots, y_n) p(\theta|y_1, \ldots, y_n) \mathrm{d}\theta \\
&= \int p(\tilde{y}|\theta) p(\theta|y_1, \ldots, y_n) \mathrm{d}\theta \\
&= \int \text{Poisson}(\tilde{y}|\theta) \text{Gamma}(\theta|a + \sum y_i, b + n) \mathrm{d}\theta \\
&= \int \left\{ \frac{1}{\tilde{y}!} \theta^{\tilde{y}} e^{-\theta} \right\} \left\{ \frac{b + n)^{a + \sum y_i}}{\Gamma(a + \sum y_i)} \theta^{a + \sum y_i - 1} e^{-(b+n)\theta} \right\} \mathrm{d}\theta \\
&= \frac{(b + n)^{a + \sum y_i}}{\Gamma(\tilde{y} + 1)\Gamma(a + \sum y_i)} \int \theta^{a + \sum y_i + \tilde{y} - 1} e^{-(b+n+1)\theta} \mathrm{d}\theta.
\end{aligned}$$

Exploiting the identity that follows from the definition of Gamma density

$$\int \theta^{a-1} e^{-b\theta} = \Gamma(a)/b^a$$

to obtain

$$p(\tilde{y}|y_1, \ldots, y_n) = \frac{\Gamma(a + \sum y_i + \tilde{y})}{\Gamma(\tilde{y} + 1)\Gamma(a + \sum y_i)} \left(\frac{b + n}{b + n + 1} \right)^{a + \sum y_i} \left(\frac{1}{b + n + 1} \right)^{\tilde{y}}.$$

This is a *negative binomial distribution* with parameters $(a + \sum y_i, b + n)$ (i.e., the number of $\tilde{y}$ failures until $a + \sum y_i$ successes), for which

$$\begin{aligned}
\mathbb{E}[\tilde{Y}|y_1, \ldots, y_n] &= \frac{1/(b+n+1)}{(b+n)/(b+n+1)}(a + \sum y_i) \\
&= \frac{a + \sum y_i}{b+n} = \mathbb{E}[\theta|y_1, \ldots, y_n] \\
\mathrm{Var}[\tilde{Y}|y_1, \ldots, y_n] &= \mathbb{E}[\tilde{Y}|y_1, \ldots, y_n]\frac{b+n+1}{b+n} \\
&= \frac{a + \sum y_i}{b+n}\frac{b+n+1}{b+n} = \mathrm{Var}[\theta|y_1, \ldots, y_n](b+n+1).
\end{aligned}$$

Note how the predictive posterior mean of $\tilde{Y}$ is the same as that of θ. This is due to the fact of Poisson sampling model: $\mathbb{E}[\tilde{Y}|\theta] = \theta$. Note also under Poisson, $\mathrm{Var}[\tilde{Y}|\theta] = \theta$. The predictive posterior variance of $\tilde{Y}$ is quite a bit larger than that of θ: the sources of its variability are that of the Poisson sampling model and the parameter θ itself. As n gets large, the posterior of θ contracts considerably, so the variability of $\tilde{Y}$ stems primarily from that of the Poisson sampling model rather than the parameter's. [3]

[3] Instead of exploiting properties of the negative binomial distribution, we may appeal to the iterated expectation and iterated variance formula to arrive at the above formula for the predictive posterior distribution.

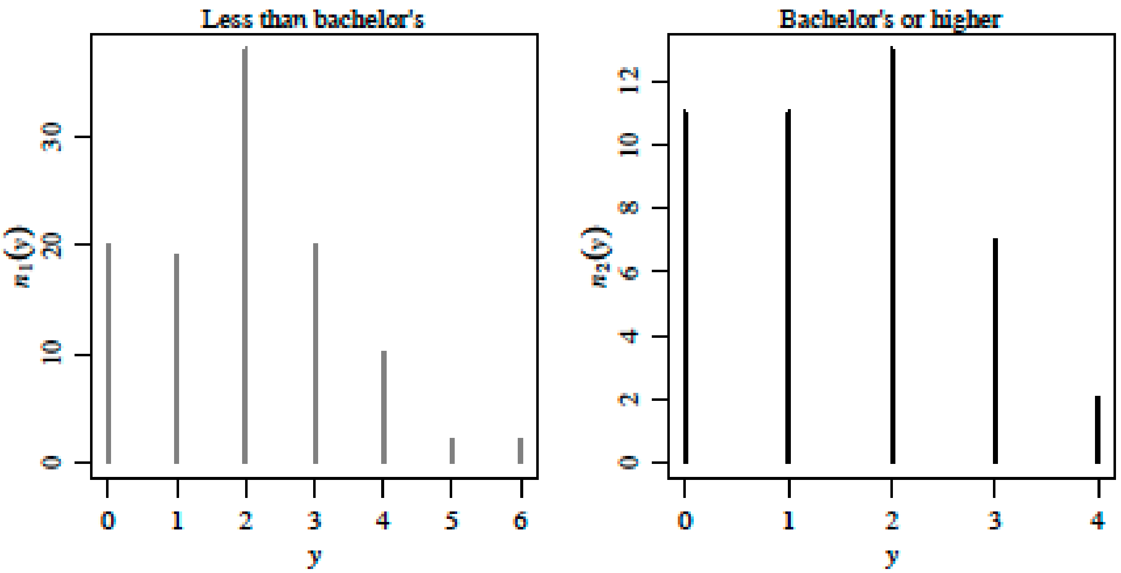


Figure 3.2: Birthrate data from the 1990s General Social Survey: number of children for the two groups of women.

3.4 Example: birth rates

We follow the example in PH (2009), Chapter 3. Fig. 3.2 illustrates the data collected on the number of children of 155 women who were 40 year of age at the time of the survey. The women are divided into two groups, those with college degrees and those without.

Let $\{Y_{i,1}\}_{i=1}^{n_1}$ denote the data from the first group, and $\{Y_{i,2}\}_{i=1}^{n_2}$ from the second group. To compare between these two groups, we shall make use of the Poisson sampling model:

$$Y_{1,1}, \ldots, Y_{n_1,1}|\theta_1 \overset{iid}{\sim} \text{poisson}(\theta_1)$$
$$Y_{1,2}, \ldots, Y_{n_2,2}|\theta_1 \overset{iid}{\sim} \text{poisson}(\theta_2).$$

Some basic statistics:

- Less than bachelor's: $n_1 = 111$, $\sum Y_{i,1} = 217$, $\bar{Y}_1 = 1.95$
- Bachelor's or higher: $n_2 = 44$, $\sum Y_{i,2} = 66$, $\bar{Y}_2 = 1.50$.

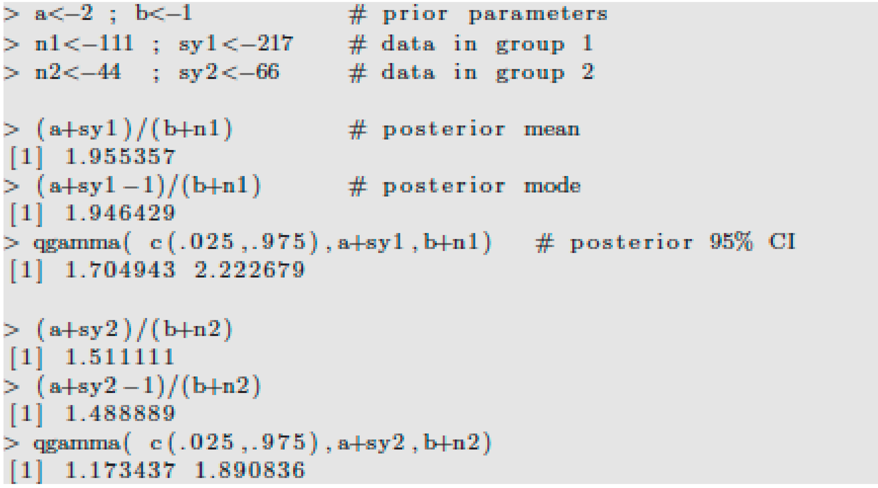
Let us endow θ_1 and θ_2 with the same prior:

$$\theta_1, \theta_2 \overset{iid}{\sim} \text{gamma}(a = 2, b = 1).$$

Then we obtain the following posterior ditsributions

$$\theta_1 | \{n_1 = 111, \sum Y_{i,1} = 217\} \sim \text{gamma}(2 + 217, 1 + 111)$$
$$\theta_2 | \{n_2 = 44, \sum Y_{i,2} = 66\} \sim \text{gamma}(68, 45)$$

In R codes:



```
> a<-2 ; b<-1              # prior parameters
> n1<-111 ; sy1<-217       # data in group 1
> n2<-44  ; sy2<-66        # data in group 2

> (a+sy1)/(b+n1)           # posterior mean
[1] 1.955357
> (a+sy1-1)/(b+n1)         # posterior mode
[1] 1.946429
> qgamma( c(.025,.975),a+sy1,b+n1)   # posterior 95% CI
[1] 1.704943 2.222679

> (a+sy2)/(b+n2)
[1] 1.511111
> (a+sy2-1)/(b+n2)
[1] 1.488889
> qgamma( c(.025,.975),a+sy2,b+n2)
[1] 1.173437 1.890836
```

The posterior distributions give substantial evidence that $\theta_1 > \theta_2$. For example, it can be computed that $\Pr(\theta_1 > \theta_2 | \sum Y_i, 1 = 217, \sum Y_i, 2 = 66) = .97$.

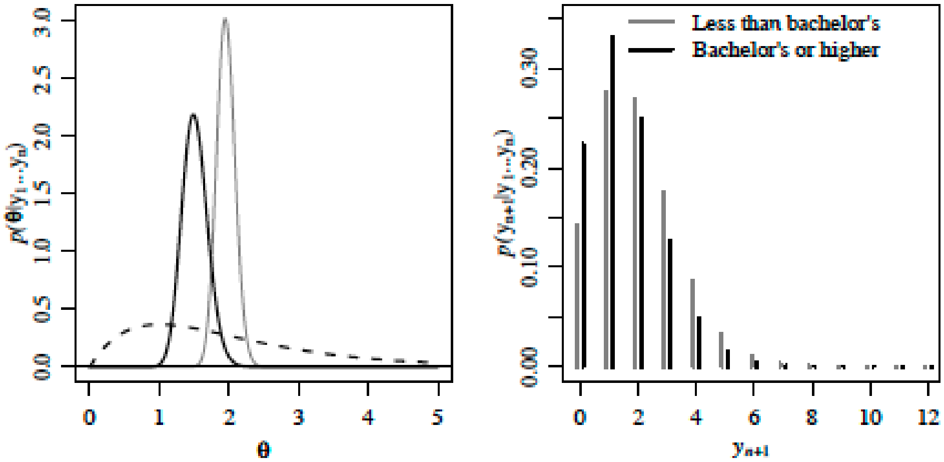
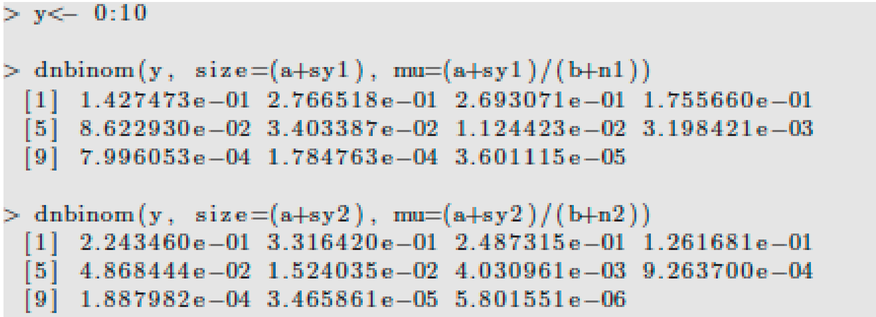


Figure 3.3: Posterior distributions of mean birth rates with the common prior given by the dashed line, and the posterior predictive distributions for number of children.

To what extent do we expect that a woman without the bachelor's degree to have more children than the other? See the right panel in Fig. 3.3.

In R codes:



```
> y<- 0:10

> dnbinom(y, size=(a+sy1), mu=(a+sy1)/(b+n1))
 [1] 1.427473e-01 2.766518e-01 2.693071e-01 1.755660e-01
 [5] 8.622930e-02 3.403387e-02 1.124423e-02 3.198421e-03
 [9] 7.996053e-04 1.784763e-04 3.601115e-05

> dnbinom(y, size=(a+sy2), mu=(a+sy2)/(b+n2))
 [1] 2.243460e-01 3.316420e-01 2.487315e-01 1.261681e-01
 [5] 4.868444e-02 1.524035e-02 4.030961e-03 9.263700e-04
 [9] 1.887982e-04 3.465861e-05 5.801551e-06
```

There is considerable overlap between the two predictive posterior distributions of $\tilde{Y}_1$ and $\tilde{Y}_2$. We can compute that

$$\Pr(\tilde{Y}_1 > \tilde{Y}_2 | \sum Y_{i,1} = 217, \sum Y_{i,2} = 66) = .48$$
$$\Pr(\tilde{Y}_1 = \tilde{Y}_2 | \sum Y_{i,1} = 217, \sum Y_{i,2} = 66) = .22.$$

It is a reminder that the Poisson sampling model has very high variance, so that the strong evidence in the difference of two population's does not mean the individual observations are overtly different.

The materials in this course are adapted from materials created by David Blei, Yang Chen, Andrew Gelman, Scott Linderman, Long Nguyen, and the 3blue1brown channel.