Chapter 4: Monte Carlo Approximation

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

- This chapter corresponds to Chapter 4 of Peter Hoff's book.
- The use of conjugate priors lead to posterior distributions for which there were simple formula for posterior means and variances.
- However, we often want to summarize other aspects of a posterior, e.g.
  - Calculate $\mathrm{P}(\theta \in A|\boldsymbol{y})$ for arbitrary sets A;
  - Compute means and standard deviations of some function of θ;
  - Find predictive distribution of missing or unobserved data;
  - Find posterior distribution of $|\theta_1 - \theta_2|$, θ_1/θ_2 or $\max\{\theta_1, \ldots, \theta_m\}$ when comparing two or more populations.
- Obtaining exact values for these posterior quantities may be difficult or analytically impossible. However, if we can generate random samples of the parameters from their posterior distributions, then these quantities can be approximated using the **Monte Carlo method**.

Outline

4.1 The Monte Carlo Method

- Let θ be a parameter of interest and $\boldsymbol{y} = (y_1, \ldots, y_n)$ be observations from a distribution $p(y|\theta)$. Suppose we could sample S independent, random θ-values from the posterior distribution $p(\theta|\boldsymbol{y})$:

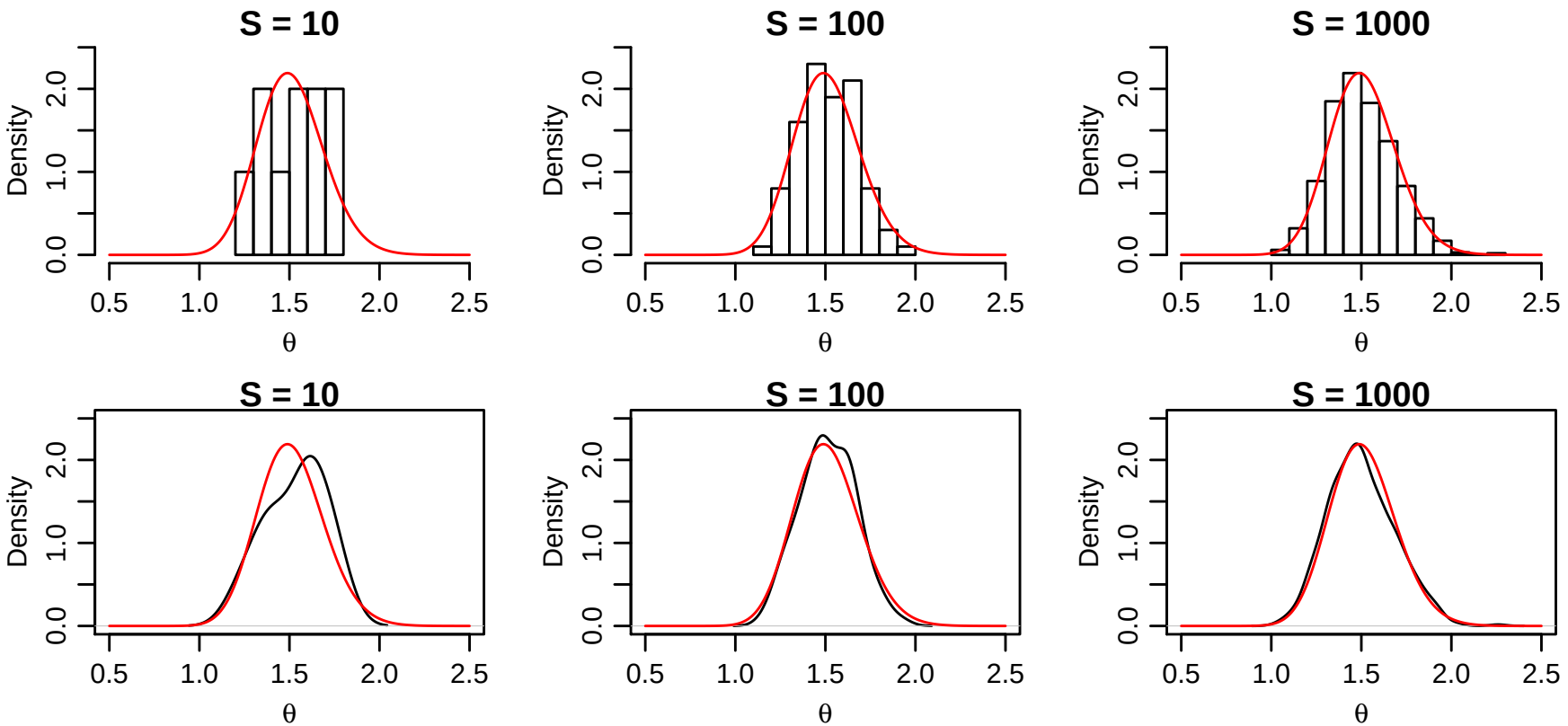
$$\theta^{(1)}, \ldots, \theta^{(S)} \overset{\text{i.i.d.}}{\sim} p(\theta|\boldsymbol{y}).$$

- The empirical distribution of the samples $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ provides a **Monte Carlo approximation** to $p(\theta|\boldsymbol{y})$, and the quality of approximation improves as S increases.

- Many computing environments have procedures for simulating this sampling process. `R` has built in functions (e.g. `rnorm, rgamma` etc) to generate i.i.d. samples from standard probability distributions.

4.1 The Monte Carlo Method

- The figure below shows Monte Carlo approximations to the density of Gamma(68,45). The first row shows histograms and the second row shows kernel estimates based on S samples generated from Gamma(68,45). The true density is superimposed in red.



4.1 The Monte Carlo Method

- The empirical distribution of the Monte Carlo samples provides an increasingly close approximation to the true density as S increases.
- The previous figure can be produced using the R-code below.
- To enable reproducible results, it may be desirable to set a random seed when doing simulations.
- We store the values of S that we are investigating in the vector `Svalues` and create a sequence of θ-values stored in the vector `theta`. Then we set up a 2×3 plot.

```
set.seed(1)
Svalues <- c(10,100,1000) # sample size
theta <- seq(from=0.5, to=2.5, by=0.01)
par(mfcol=c(2,3))
```

4.1 The Monte Carlo Method

- For each value of S, generate S samples from Gamma(68,45) and store these in the vector `samples`.
- Plot the histogram of the random samples, setting `freq=FALSE` to plot densities instead of frequencies (total area of histogram is 1).
- Use `dgamma` to compute the true density and superimpose this curve on the histogram.
- Use `density` to obtain kernel density estimates from the random samples and plot this estimated kernel density.

```
for (i in 1:3){
  S <- Svalues[i]
  samples <- rgamma(S,68,45)
  hist(samples, freq=FALSE, xlim=c(0.5,2.5), ylim=c(0,2.5))
  points(theta, dgamma(theta,68,45), type="l", col="red")
  plot(density(samples), xlim=c(0.5,2.5), ylim=c(0,2.5))
  points(theta, dgamma(theta,68,45), type="l", col="red")
}
```

4.1.1 Law of Large Numbers

- Let $g(\theta)$ be (just about) any function of the parameter θ. The **law of large numbers** says that if $\theta^{(1)}, \ldots, \theta^{(S)}$ are i.i.d. samples from $p(\theta|\boldsymbol{y})$, then as $S \to \infty$,

$$\frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)}) \to \mathrm{E}[g(\theta)|\boldsymbol{y}] = \int g(\theta)p(\theta|\boldsymbol{y})\,\mathrm{d}\theta.$$

  The convergence happens in certain sense (in probability, or almost surely).

4.1.1 Law of Large Numbers

- The law of large numbers implies that as $S \to \infty$,
  - $\bar{\theta} = \frac{1}{S}\sum_{s=1}^{S} \theta^{(s)} \to \mathrm{E}(\theta|\boldsymbol{y})$;
  - $\frac{1}{S-1}\sum_{s=1}^{S} \left(\theta^{(s)} - \bar{\theta}\right)^2 \to \mathrm{Var}(\theta|\boldsymbol{y})$;
  - $\frac{1}{S}\sum_{s=1}^{S} \mathbb{1}\{\theta^{(s)} \leq c\} \to \mathrm{P}(\theta \leq c|\boldsymbol{y})$, for any real number c;
  - The empirical distribution of $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ converges to $p(\theta|\boldsymbol{y})$;
  - The $\alpha \times 100$-percentile of $\{\theta^{(1)}, \ldots, \theta^{(S)}\} \to \theta_\alpha$, where $\mathrm{P}(\theta \leq \theta_\alpha|\boldsymbol{y}) = \alpha$.
- Replace $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ above with $\{g(\theta^{(1)}), \ldots, g(\theta^{(S)})\}$ for any function g, usually these convergence results still hold true.

4.1.2 Numerical Evaluation

- Just about any aspect of a posterior distribution can be approximated arbitrarily exactly with a large enough Monte Carlo sample.
- Let us gain some familiarity with the Monte Carlo procedure by first comparing its approximations to a few posterior quantities that can be computed exactly (or nearly so) by other methods.
- Suppose $Y_1, \ldots, Y_n | \theta \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\theta)$ and a $\text{Gamma}(a_0, b_0)$ prior is considered for θ. Having observed $\boldsymbol{y} = (y_1, \ldots, y_n)$, the posterior is

$$\theta | \boldsymbol{y} \sim \text{Gamma}\left(a_0 + \sum_{i=1}^{n} y_i, b_0 + n\right).$$

- Let $a_0 = 2$, $b_0 = 1$, $n = 44$, $\sum_{i=1}^{n} y_i = 66$ so that

$$\theta | \boldsymbol{y} \sim \text{Gamma}(68, 45).$$

4.1.2 Numerical Evaluation

Expectation

- The posterior mean is $68/45 = 1.51$.
- Monte Carlo approximations of the posterior mean for $S \in \{10, 100, 1000\}$ can be obtained in R as follows.

```
> set.seed(1)
> Svalues <- c(10,100,1000)
> m <- rep(0,3)
> for (i in 1:3){
+  S <- Svalues[i]
+  samples <- rgamma(S,68,45)
+  m[i] <- mean(samples)
+ }
> m
[1] 1.532794 1.513947 1.501015
```

4.1.2 Numerical Evaluation

Probabilities

- The posterior probability $\mathrm{P}(\theta < 1.75|\boldsymbol{y})$ can be computed in R using pgamma which yields 0.8998.
- The Monte Carlo approximations of this posterior probability for $S \in \{10, 100, 1000\}$ are as follows.

```
> pgamma(1.75,68,45)
[1] 0.8998286
```

```
> set.seed(1)
> Svalues <- c(10,100,1000)
> prob <- rep(0,3)
> for (i in 1:3){
+  S <- Svalues[i]
+  samples <- rgamma(S,68,45)
+  prob[i] <- sum(samples<1.75)/S
+ }
> prob
[1] 0.900 0.940 0.899
```

4.1.2 Numerical Evaluation

Quantiles

- A 95% quantile-based CI can be obtained using `qgamma` which gives (1.173, 1.891).
- Approximate 95% CI can also be obtained from Monte Carlo samples.

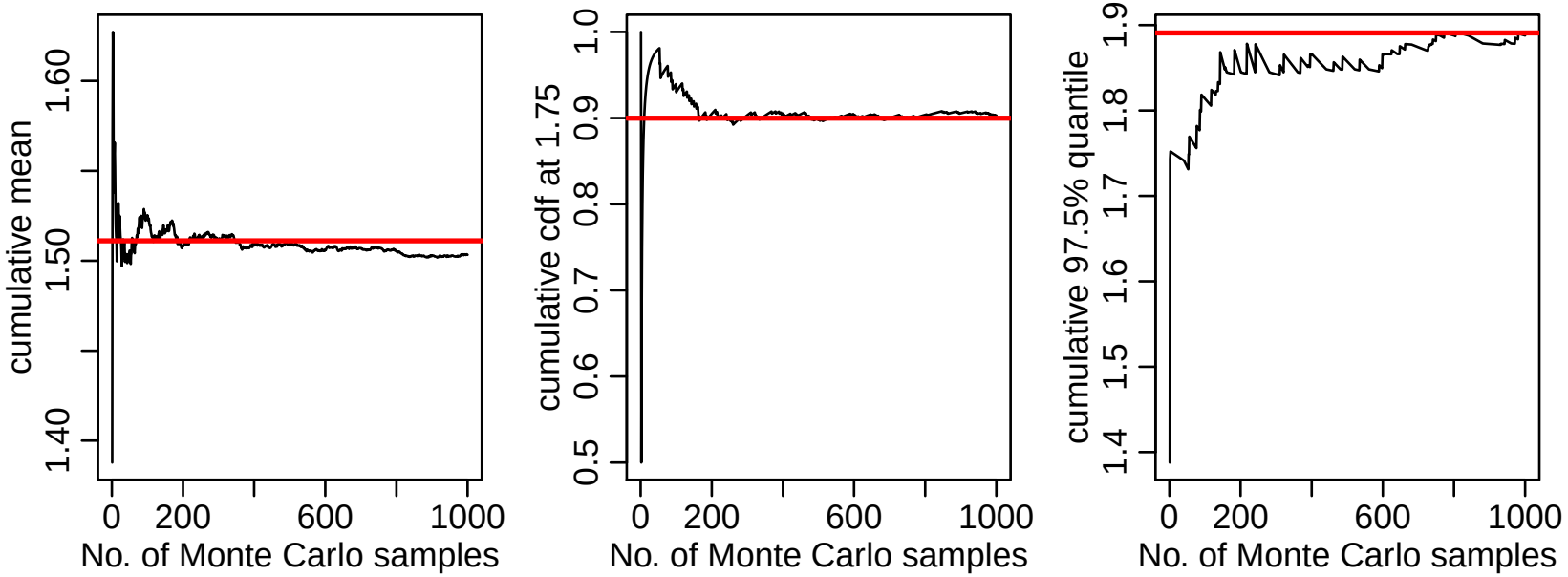
```
> qgamma(c(0.025,0.975),68,45)
[1] 1.173437 1.890836

> set.seed(1)
> Svalues <- c(10,100,1000)
> CI <- matrix(0,3,2)
> for (i in 1:3){
+  S <- Svalues[i]
+  samples <- rgamma(S,68,45)
+  CI[i,] <- quantile(samples,
+                     c(0.025,0.975))
+ }
```

```
> CI
          [,1]     [,2]
[1,] 1.260291 1.750068
[2,] 1.231646 1.813752
[3,] 1.180194 1.892473
```

4.1.2 Numerical Evaluation

- The figure below shows the convergence of the Monte Carlo estimates to the true values (in red) graphically, based on cumulative estimates from a sequence of $S = 1000$ samples from Gamma(68,45).



- Such plots can help indicate when enough Monte Carlo samples have been made.

4.1.3 Accuracy of Monte Carlo estimates

- Monte Carlo standard errors can be used to assess the accuracy of approximations to posterior means.
- Let $\bar{\theta} = \sum_{s=1}^{S} \theta^{(s)}/S$ be the sample mean of the Monte Carlo samples.
- The **central limit theorem** says that

$$\bar{\theta}|\boldsymbol{y} \sim \mathrm{N}\left(\mathrm{E}(\theta|\boldsymbol{y}), \frac{\mathrm{Var}(\theta|\boldsymbol{y})}{S}\right) \quad \text{approximately.}$$

- Let $\hat{\sigma}^2 = \frac{1}{S-1}\sum_{s=1}^{S}(\theta^{(s)} - \bar{\theta})^2$ be the Monte Carlo estimate of $\mathrm{Var}(\theta|\boldsymbol{y})$. Then the Monte Carlo standard error is $\sqrt{\hat{\sigma}^2/S}$.
- An approximate 95% Monte Carlo confidence interval for $\mathrm{E}(\theta|\boldsymbol{y})$ is

$$\bar{\theta} \pm 2\sqrt{\hat{\sigma}^2/S}$$

.

4.1.3 Accuracy of Monte Carlo estimates

- S can be chosen large enough so that the Monte Carlo standard error is less than the precision to which we want to report $\mathrm{E}(\theta|\boldsymbol{y})$.
- For example, if we had generated a Monte Carlo sample of size $S = 100$ for which the estimate of $\mathrm{Var}(\theta|\boldsymbol{y})$ was 0.024. Then the Monte Carlo standard error is $\sqrt{0.024/100} = 0.015$.
- If we want the difference between $\mathrm{E}(\theta|\boldsymbol{y})$ and its Monte Carlo estimate to be less than 0.01 with high probability (a probability $> 95\%$), we would need to increase the Monte Carlo sample size so that

$$2\sqrt{0.024/S} < 0.01 \Rightarrow S > 960.$$

Outline

4.2 Posterior inference for arbitrary functions

- Suppose we are interested in the posterior distribution of some computable function $g(\theta)$ of θ. For example, in the binomial model, we are sometimes interested in the log odds:

$$\gamma = \log \frac{\theta}{1-\theta}.$$

- From the law of large numbers, if we generate $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ from $p(\theta|\boldsymbol{y})$, then $\frac{1}{S}\sum_{s=1}^{S} \log \frac{\theta^{(s)}}{1-\theta^{(s)}}$ converges to $\mathrm{E}\left[\log \frac{\theta}{1-\theta} \,\middle|\, \boldsymbol{y}\right]$ as $S \to \infty$.

- However, we may also be interested in other aspects of the posterior distribution of $\gamma = \log \frac{\theta}{1-\theta}$.

4.2 Posterior inference for arbitrary functions

- These too can be computed using a Monte Carlo approach:

$$\left.\begin{array}{l} \text{sample } \theta^{(1)} \sim p(\theta|\boldsymbol{y}) \text{ and compute } \gamma^{(1)} = g(\theta^{(1)}) \\ \vdots \\ \text{sample } \theta^{(S)} \sim p(\theta|\boldsymbol{y}) \text{ and compute } \gamma^{(S)} = g(\theta^{(S)}) \end{array}\right\} \text{independently.}$$

- The sequence $\{\gamma^{(1)}, \ldots, \gamma^{(S)}\}$ constitutes S independent samples from $p(\gamma|\boldsymbol{y})$, and so as $S \to \infty$
  - $\bar{\gamma} = \sum_{s=1}^{S} \gamma^{(s)}/S \to \mathrm{E}(\gamma|\boldsymbol{y})$
  - $\frac{1}{S-1}\sum_{s=1}^{S}(\gamma^{(s)} - \bar{\gamma})^2 \to \mathrm{Var}(\gamma|\boldsymbol{y})$
  - the empirical distribution of $\{\gamma^{(1)}, \ldots, \gamma^{(S)}\} \to p(\gamma|\boldsymbol{y})$

  as before.

4.2.1 Example: Log-odds

- In the 1988 General Social Survey, respondents were asked if they agreed with a Supreme Court ruling that prohibited governments from requiring the reading of religious texts in public schools.
- Of the respondents, $y = 441$ out of $n = 860$ non-Protestants agreed with the ruling.
- Let θ be the population proportion agreeing with the ruling among non-Protestants.
- Using a Binomial sampling model and a uniform prior distribution, the posterior $p(\theta|y)$ is
  $\text{Beta}(1 + 441, 1 + 860 - 441) = \text{Beta}(442, 420)$.
- Using the Monte Carlo approach described earlier, we can obtain samples of the log-odds $\gamma = \log \frac{\theta}{1-\theta}$ from both the prior and posterior distributions of γ.

4.2.1 Example: Log-odds

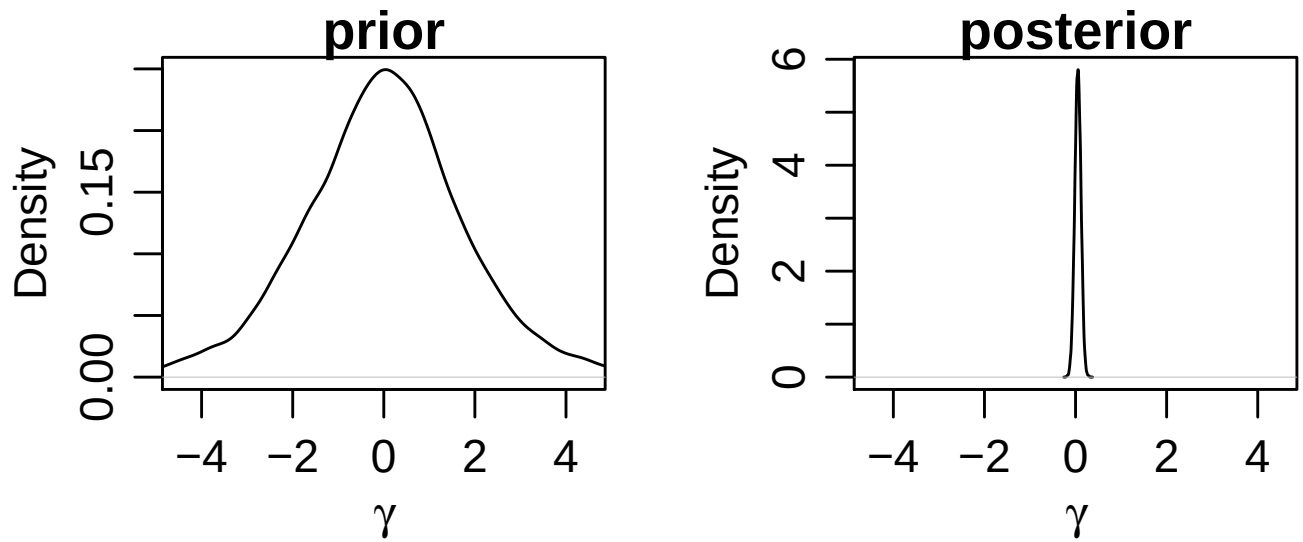
- The R-code below sets the sample size S as 10000 and the Beta prior parameters, `a0` and `b0`, both as 1 (corresponding to a uniform prior).
- Then we use `rbeta` to generate samples from the prior and posterior of θ. The posterior $p(\theta|y)$ is a Beta$(a_0 + y, b_0 + n - y)$ distribution. The θ-samples are then transformed to obtain samples of γ.

```
S <- 10000
a0 <- 1; b0 <- 1
y <- 441; n <- 860
theta_samples_prior <- rbeta(S,a0,b0)
gamma_samples_prior <- log(theta_samples_prior/
                           (1-theta_samples_prior))
theta_samples_post <- rbeta(S,a0+y,b0+n-y)
gamma_samples_post <- log(theta_samples_post/
                          (1-theta_samples_post))
```

4.2.1 Example: Log-odds

- Using the `density` function, we can plot smooth kernel density approximations of the prior and posterior distributions of γ.

```
plot(density(gamma_samples_prior),xlim=c(-4.5,4.5))
plot(density(gamma_samples_post),xlim=c(-4.5,4.5))
```



4.2.2 Example: Birth rates

- The General Social Survey gathered data on the educational attainment and number of children of 155 women who were 40 years old at the time of survey. We will compare the women with college degrees to those without in terms of their numbers of children.

- Let $Y_{1,1}, \ldots, Y_{n_1,1}$ denote the numbers of children for the $n_1 = 111$ women without college degrees and $Y_{1,2}, \ldots Y_{n_2,2}$ be the corresponding data for the $n_2 = 44$ women with degrees.

- We assume $Y_{i,1}$ is independent of $Y_{j,2}$ for any i, j and use the sampling models

$$Y_{1,1}, \ldots, Y_{n_1,1} \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\theta_1),$$
$$Y_{1,2}, \ldots, Y_{n_2,2} \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\theta_2).$$

and the conjugate priors $\{\theta_1, \theta_2\} \overset{\text{i.i.d.}}{\sim} \text{Gamma}(2, 1)$.

4.2.2 Example: Birth rates

- Let $\boldsymbol{y}_1 = (y_{1,1}, \ldots, y_{n_1,1})$ and $\boldsymbol{y}_2 = (y_{1,2}, \ldots, y_{n_2,2})$. We have $\sum_{i=1}^{n_1} y_{i,1} = 217$ and $\sum_{i=1}^{n_2} y_{i,2} = 66$. The joint posterior distribution is

$$\begin{aligned} p(\theta_1, \theta_2 | \boldsymbol{y}_1, \boldsymbol{y}_2) &\propto p(\boldsymbol{y}_1, \boldsymbol{y}_2 | \theta_1, \theta_2) p(\theta_1, \theta_2) \\ &\propto \{p(\boldsymbol{y}_1 | \theta_1) p(\theta_1)\} \{p(\boldsymbol{y}_2 | \theta_2) p(\theta_2)\} \\ &\propto p(\theta_1 | \boldsymbol{y}_1) p(\theta_2 | \boldsymbol{y}_2). \end{aligned}$$

Hence, θ_1 and θ_2 are independent *a posteriori* and

$$\begin{aligned} \theta_1 | \boldsymbol{y}_1 &\sim \text{Gamma}(2 + 217, 1 + 111) = \text{Gamma}(219, 112), \\ \theta_2 | \boldsymbol{y}_2 &\sim \text{Gamma}(2 + 66, 1 + 44) = \text{Gamma}(68, 45). \end{aligned}$$

4.2.2 Example: Birth rates

- To describe our knowledge about the difference between θ_1 and θ_2, we may be interested in the value of $\mathrm{P}(\theta_1 > \theta_2 | \boldsymbol{y}_1, \boldsymbol{y}_2)$ or the posterior distribution of θ_1/θ_2.

- Both of these quantities can be obtained using Monte Carlo sampling:

$$\text{Sample} \quad \theta_1^{(1)}, \ldots, \theta_1^{(S)} \sim p(\theta_1 | \boldsymbol{y}) \quad \text{and} \quad \theta_2^{(1)}, \ldots, \theta_2^{(S)} \sim p(\theta_2 | \boldsymbol{y}).$$

- The sequence $\{(\theta_1^{(1)}, \theta_2^{(1)}), \ldots, (\theta_1^{(S)}, \theta_2^{(S)})\}$ consists of S independent samples from the joint posterior distribution of θ_1 and θ_2 and can be used to make Monte Carlo approximations to posterior quantities of interest.

4.2.2 Example: Birth rates

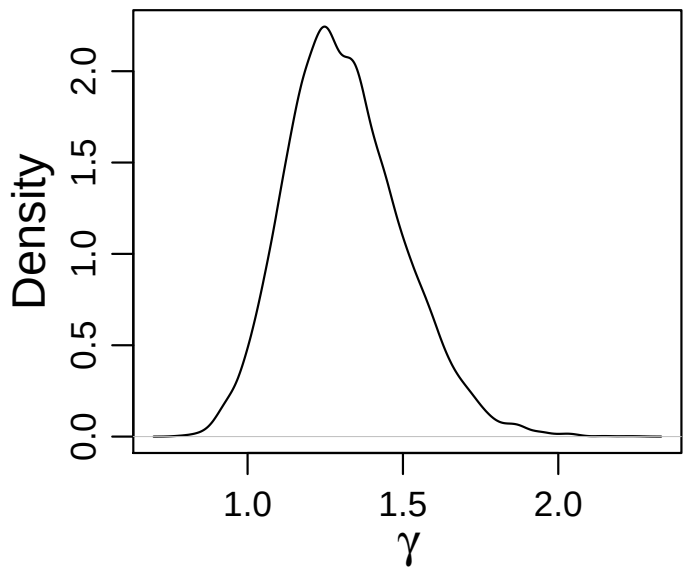
- $\mathrm{P}(\theta_1 > \theta_2 | \boldsymbol{y}_1, \boldsymbol{y}_2)$ can be approximated by $\frac{1}{S}\sum_{s=1}^{S} \mathbb{1}(\theta_1^{(s)} > \theta_2^{(s)})$.

```
> set.seed(1)
> S <- 10^4
> theta1.samples <- rgamma(S,219,112)
> theta2.samples <- rgamma(S,68,45)
> sum(theta1.samples>theta2.samples)/S
[1] 0.9708
```

Hence, $\mathrm{P}(\theta_1 > \theta_2 | \boldsymbol{y}_1, \boldsymbol{y}_2) \approx 0.9708$. The posterior indicates substantial evidence that $\theta_1 > \theta_2$.

4.2.2 Example: Birth rates

- If we are interested in the means of the two groups, we could consider the empirical distribution of $\{\theta_1^{(1)}/\theta_2^{(1)}, \ldots, \theta_1^{(S)}/\theta_2^{(S)}\}$. A Monte Carlo estimate of this posterior density of $\gamma = \theta_1/\theta_2$ is given in the figure below.




```
gamma.samples <- theta1.samples/theta2.samples
plot(density(gamma.samples),xlab=expression(gamma),main="")
```

Outline

4.3 Sampling from predictive distributions

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

$$p(\tilde{y}|\mathbf{y}) = \int p(\tilde{y}|\theta)\, p(\theta|\mathbf{y})\, \mathrm{d}\theta. \tag{1}$$

- In some situations, we may be able to sample from $p(\theta|\boldsymbol{y})$ and $p(y|\theta)$, but $p(\tilde{y}|\mathbf{y})$ is too complicated to sample from directly.

- If so, we can sample from the posterior predictive distribution indirectly using a Monte Carlo procedure, based on the following observation: From Equation (1), $p(\tilde{y}|\mathbf{y})$ can be written as the posterior expectation of $p(\tilde{y}|\theta)$, i.e.

$$p(\tilde{y}|\mathbf{y}) = \mathrm{E}[p(\tilde{y}|\theta)|\boldsymbol{y}].$$

4.3 Sampling from predictive distributions

- Hence, if $p(y|\theta)$ is discrete and $\{\theta^{(1)}, \ldots, \theta^{(S)}\} \overset{\text{i.i.d.}}{\sim} p(\theta|\boldsymbol{y})$,

$$\begin{aligned} \mathrm{P}(\tilde{Y} = \tilde{y}|\boldsymbol{y}) &= \int \mathrm{P}(\tilde{Y} = \tilde{y}|\theta)\, p(\theta|\mathbf{y})\, \mathrm{d}\theta \\ &= \mathrm{E}[\mathrm{P}(\tilde{Y} = \tilde{y}|\theta)|\boldsymbol{y}] \approx \frac{1}{S}\sum_{s=1}^{S} P(\tilde{Y} = \tilde{y}|\theta^{(s)}) \end{aligned}$$

- More generally, it is useful to have a set of samples of $\tilde{y}$ from its posterior predictive distribution. These samples can be obtained as follows:

$$\begin{aligned} &\text{sample} \quad \theta^{(1)} \sim p(\theta|\boldsymbol{y}) \quad \text{and sample} \quad \tilde{y}^{(1)} \sim p(\tilde{y}|\theta^{(1)}) \\ &\qquad \vdots \\ &\text{sample} \quad \theta^{(S)} \sim p(\theta|\boldsymbol{y}) \quad \text{and sample} \quad \tilde{y}^{(S)} \sim p(\tilde{y}|\theta^{(S)}) \end{aligned}$$

  The sequence $\{\tilde{y}^{(1)}, \ldots, \tilde{y}^{(S)}\}$ constitutes S independent samples from the posterior predictive distribution of $\tilde{y}$.

4.3.1 Example: Birth rates

- In the birth rates example, suppose we sample two individuals randomly, one from each of the two populations (with and without college degree).
- Let $\tilde{Y}_1$ and $\tilde{Y}_2$ denote the numbers of children of the women without and with college degree respectively.
- To what extent do we expect the woman without college degree to have more children than the other?

4.3.1 Example: Birth rates

- To answer this question, we need to find $\mathrm{P}(\tilde{Y}_1 > \tilde{Y}_2 | \boldsymbol{y}_1, \boldsymbol{y}_2)$. First note that

$$\begin{aligned} p(\tilde{Y}_1, \tilde{Y}_2 | \boldsymbol{y}_1, \boldsymbol{y}_2) &= \int \int p(\tilde{Y}_1, \tilde{Y}_2 | \theta_1, \theta_2) p(\theta_1, \theta_2 | \boldsymbol{y}_1, \boldsymbol{y}_2) \, \mathrm{d}\theta_1 \, \mathrm{d}\theta_2 \\ &= \int \int p(\tilde{Y}_1 | \theta_1) p(\tilde{Y}_2 | \theta_2) p(\theta_1 | \boldsymbol{y}_1) p(\theta_2 | \boldsymbol{y}_2) \, \mathrm{d}\theta_1 \, \mathrm{d}\theta_2 \\ &= \int p(\tilde{Y}_1 | \theta_1) p(\theta_1 | \boldsymbol{y}_1) \, \mathrm{d}\theta_1 \int p(\tilde{Y}_2 | \theta_2) p(\theta_2 | \boldsymbol{y}_2) \, \mathrm{d}\theta_2 \\ &= p(\tilde{Y}_1 | \boldsymbol{y}_1) p(\tilde{Y}_2 | \boldsymbol{y}_2). \end{aligned}$$

  That is, $\tilde{Y}_1$ and $\tilde{Y}_2$ are *a posteriori* independent.

- We can approximate $\mathrm{P}(\tilde{Y}_1 > \tilde{Y}_2 | \boldsymbol{y}_1, \boldsymbol{y}_2)$ with Monte Carlo sampling. Since $\tilde{Y}_1$ and $\tilde{Y}_2$ are a posteriori independent, we can sample each variable separately from their individual posterior distribution.

4.3.1 Example: Birth rates

- For $s = 1, \ldots, S$,

$$\text{Sample } \theta_1^{(s)} \sim \underbrace{p(\theta_1|\boldsymbol{y}_1)}_{\text{Gamma(219,112)}} \quad \text{and } \tilde{y}_1^{(s)} \sim \underbrace{p(\tilde{y}_1|\theta_1^{(s)})}_{\text{Poisson}(\theta_1^{(s)})},$$

$$\text{Sample } \theta_2^{(s)} \sim \underbrace{p(\theta_2|\boldsymbol{y}_2)}_{\text{Gamma(68,45)}} \quad \text{and } \tilde{y}_2^{(s)} \sim \underbrace{p(\tilde{y}_2|\theta_2^{(s)})}_{\text{Poisson}(\theta_2^{(s)})}.$$

- Then

$$\text{P}(\tilde{Y}_1 > \tilde{Y}_2|\boldsymbol{y}_1, \boldsymbol{y}_2) \approx \frac{1}{S}\sum_{s=1}^{S} \mathbb{1}\left\{\tilde{y}_1^{(s)} > \tilde{y}_2^{(s)}\right\},$$

where $\mathbb{1}(A) = 1$ if the event A happens, and $\mathbb{1}(A) = 0$ otherwise.

4.3.1 Example: Birth rates

- Let $S = 10^6$. First, we generate samples of θ_1 and θ_2 from their respective posterior distributions. Then we generate samples of $\tilde{Y}_1$ and $\tilde{Y}_2$ from their sampling models conditional on θ_1 and θ_2 respectively.
- We can now compute the proportion of samples for which $\tilde{Y}_1 > \tilde{Y}_2$.

```
> S <- 10^6
> # Method 2
> set.seed(1)
> theta1.samples <- rgamma(S,219,112)
> theta2.samples <- rgamma(S,68,45)
> predy1.samples <- rpois(S,theta1.samples)
> predy2.samples <- rpois(S,theta2.samples)
> sum(predy1.samples > predy2.samples)/S
[1] 0.48228
```

- Hence $\mathrm{P}(\tilde{Y}_1 > \tilde{Y}_2 | \boldsymbol{y}_1, \boldsymbol{y}_2) \approx 0.482$.

4.3.1 Example: Birth rates

- You may recall from Chapter 2 that

$$\tilde{Y}_1|\boldsymbol{y}_1 \sim \text{Negative Binomial}\left(219, \frac{112}{112+1}\right),$$

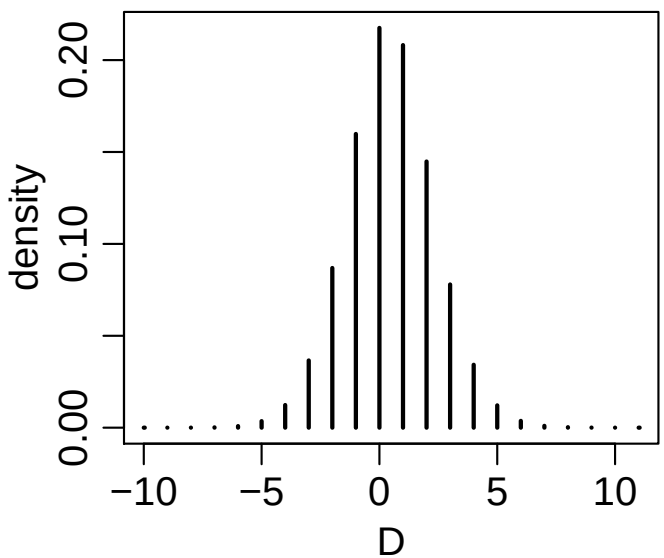
$$\tilde{Y}_2|\boldsymbol{y}_2 \sim \text{Negative Binomial}\left(68, \frac{45}{45+1}\right).$$

Thus an alternative method is to generate samples of $\tilde{Y}_1$ and $\tilde{Y}_2$ from their posterior predictive distributions directly. We see below that the estimate of $\text{P}(\tilde{Y}_1 > \tilde{Y}_2|\boldsymbol{y}_1, \boldsymbol{y}_2)$ agrees with that obtained previously.

```
> S <- 10^6
> set.seed(1)
> r1 <- 219;  p1 <- 112/113
> r2 <- 68;   p2 <- 45/46
> predy1.samples <- rnbinom(S, r1, p1)
> predy2.samples <- rnbinom(S, r2, p2)
> sum(predy1.samples > predy2.samples)/S
[1] 0.482089
```

4.3.1 Example: Birth rates

- Once we have generated these Monte Carlo samples from the posterior predictive distribution, we can use them to calculate other posterior quantities of interest.
- The figure below shows the Monte Carlo approximation to the posterior distribution of $D = \tilde{Y}_1 - \tilde{Y}_2$, the difference in number of children between two individuals, one sampled from each of the two groups.



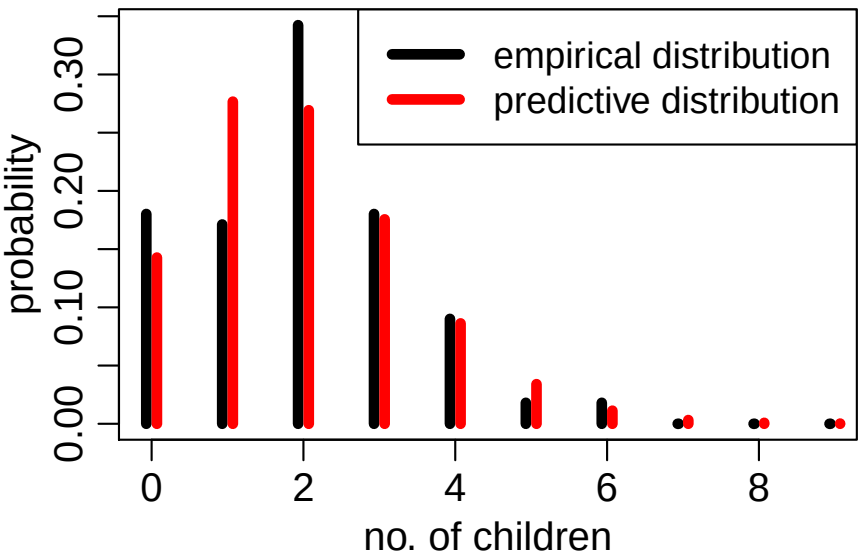

```
> D <- predy1.samples - predy2.samples
> D <- as.factor(D)
> prob <- summary(D)/S
> sum(prob)
[1] 1
> plot(as.numeric(levels(D)),prob,
+   type="h",xlab="D",ylab="density")
```

Outline

4.4 Posterior predictive model checking

Example: Birth rates

- Consider the sample of $n_1 = 111$ women without college degrees. The figure below shows the empirical distribution of the number of children of these women, along with the posterior predictive distribution.



- The number of women with two children is 38, which is twice the number of women with one child. In contrast, the predictive distribution suggest that the probability of sampling a woman with two children is slightly less probable than sampling a woman with one.

- These two distributions seem to be in conflict.
- One explanation for the large number of women in the sample with two children is sampling variability: the empirical distribution of sampled data does not generally exactly match the distribution of the population from which the data were sampled, and may look quite different if the sample size is small.
- A smooth population distribution can produce sample empirical distributions that are quite bumpy. In such cases, having a predictive distribution that smooths over the bumps of the empirical distribution may be desirable.

4.4 Posterior predictive model checking

Example: Birth rates

- An alternative explanation for the large number of women in the sample with two children is that this is indeed a feature of the population and the data are correctly reflecting this feature.
- The Poisson model is unable to represent this feature because there is no Poisson distribution that has such a sharp peak at $y = 2$.

4.4 Posterior predictive model checking

Example: Birth rates

- The discrepancy between the empirical and predictive distributions can be assessed numerically with Monte Carlo simulation.
- For every vector $\boldsymbol{y}$ of length 111, let $t(\boldsymbol{y})$ be the ratio of the number of 2's to the number of 1's in $\boldsymbol{y}$. For our observed data $\boldsymbol{y}_{\text{obs}}$, $t(\boldsymbol{y}_{\text{obs}}) = 38/19 = 2$.
- Now suppose we were to sample a different set of 111 women, obtaining a data vector $\tilde{\mathbf{Y}}$ of length 111 recording their number of children. What sort of values of $t(\tilde{\mathbf{Y}})$ would we expect?

4.4 Posterior predictive model checking

Example: Birth rates

- Monte Carlo samples from the posterior predictive distribution of $t(\tilde{\mathbf{Y}})$ can be obtained with the following procedure:

  For each $s = 1, \ldots, S$,

  1. Sample $\theta^{(s)} \sim \underbrace{p(\theta|\boldsymbol{y}_{\text{obs}})}_{\text{Gamma(219,112)}}$,
  2. Sample $\tilde{\mathbf{Y}}^{(s)} = (\tilde{y}_1^{(s)}, \ldots, \tilde{y}_{111}^{(s)}) \overset{\text{i.i.d.}}{\sim} \underbrace{p(y|\theta^{(s)})}_{\text{Poisson}(\theta^{(s)})}$,
  3. Compute $t^{(s)} = t(\tilde{\mathbf{Y}}^{(s)})$.

- $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ are samples from the posterior distribution of θ, $\{\tilde{\mathbf{Y}}^{(1)}, \ldots, \tilde{\mathbf{Y}}^{(S)}\}$ are posterior predictive datasets, each of size 111; $\{t^{(1)}, \ldots, t^{(S)}\}$ are samples from the posterior predictive distribution of $t(\tilde{Y})$.

4.4 Posterior predictive model checking

Example: Birth rates

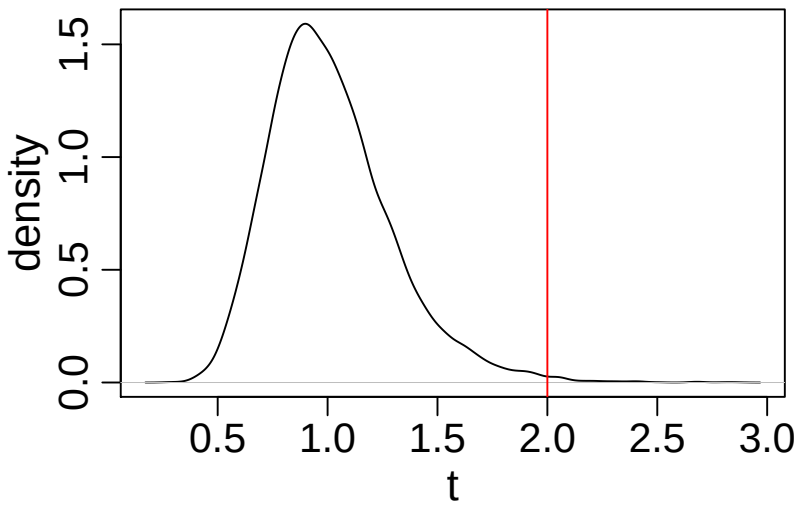
- The Monte carlo sampling can be performed using the R-code below.

```
set.seed(1)
S <- 10^4
n1 <- 111
theta.samples <- rgamma(S,219,112)
t.samples <- rep(0,S)
for (s in 1:S){
  Y.samples <- rpois(n1,theta.samples[s])
  t.samples[s] <- sum(Y.samples==2)/sum(Y.samples==1)
}
plot(density(t.samples),xlab="t",ylab="density")
abline(v=2,col="red")
> sum(t.samples>=2)/S
[1] 0.0056
```

4.4 Posterior predictive model checking

Example: Birth rates

- The figure below plots the Monte Carlo approximation to the distribution of $t(\tilde{Y})$, with the observed value $t(\boldsymbol{y}_{\text{obs}})$ indicated with a red vertical line.



- Of 10000 Monte Carlo datasets, only about 0.5% had values of $t(\boldsymbol{y})$ that equaled or exceeded $t(\boldsymbol{y}_{\text{obs}})$. This indicates that our Poisson model is flawed: it predicts that we would hardly ever see a dataset that resembled our observed one in terms of $t(\boldsymbol{y})$.

4.4 Posterior predictive model checking

- In terms of data description, we should make sure that our model generates predictive datasets $\tilde{\mathbf{Y}}$ that resemble the observed dataset in terms of features that are of interest. If this condition is not met, we may want to consider using a more complex model.
- However, an incorrect model can still provide correct inference for some aspects of the true population. For example, the Poisson model provides consistent estimation of the population mean as well as accurate confidence intervals if the population mean is approximately equal to the variance.