Homework 04

Spencer Pease

4/27/2020

Questions

Q1

Q1.a

We can generally define the posterior distribution as the product of a likelihood and prior function:

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$
$$P(\theta \mid X = x) \propto P(X = x \mid \theta) \times P(\theta)$$

where, given the parameters of the problem, the likelihood takes the form of a binomial distribution, and the prior probability follows a uniform distribution (or, a beta distribution with $\alpha = \beta = 1$):

$$X \mid \theta \sim \text{Binomial}(X, \theta)$$
$$\theta \sim \text{Beta}(1, 1)$$

We can define our binomial as the function $f_{bin}(k, n, p)$, where this describes getting exactly k successes in n trials, each with a probability p of occurring.

$$f_{bin}(k, n, p) = f_{bin}(k, n, \theta) = P(X = k \mid \theta; n) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$$

Since the prior, θ, is a beta distribution, it acts as a conjugate prior for the binomial likelihood function. From this, we know that the posterior distribution will also be a beta distribution in the form:

$$f_{post}(\theta, k, \alpha, \beta) = P(\theta \mid X = k; \alpha, \beta) = \text{Beta}(\alpha + k, n - k + \beta)$$

From the problem, we know $k = 43$ divorces occurred over the period 2005-2015 from a sample of $n = 112$ married people in 2005. This leads us to our final analytic posterior distribution:

$$\begin{aligned} P(\theta \mid X = k; \alpha, \beta, n) &= \text{Beta}(\alpha + k, n - k + \beta) \\ P(\theta \mid X = 43; 1, 1, 112) &= \text{Beta}(1 + 43, 112 - 43 + 1) \\ &= \text{Beta}(44, 70) \end{aligned}$$

Q1.b

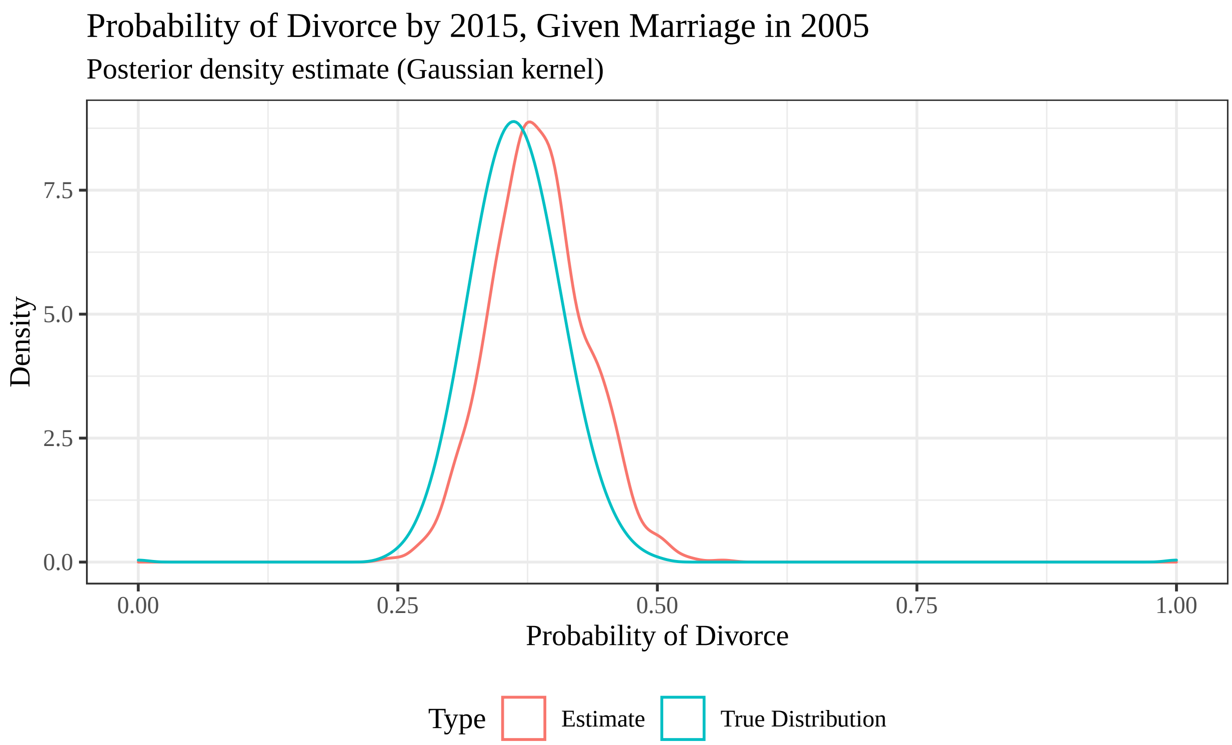
From our analytic posterior distribution of θ, we simulate a new sample of $1,000$ people. From this sample, we can get a 95% Bayesian confidence interval for θ by pulling the 2.5^{th} and 97.5^{th} percentile of the data.

Table 1: Posterior distribution summary

Mean	Median	95% Low	95% High
0.385	0.383	0.301	0.476

Q1.c

With the new sample from our posterior, we can plot a nonparametric density estimate for our posterior. The true distribution is also included for reference. Both density curves use a Gaussian kernel.



Q2

In this scenario, we take both our likelihood and prior distributions to be normal:

$$X \mid \theta \propto N(\theta, \sigma^2)$$
$$\theta \propto N(\mu, \tau^2)$$

where we are given $\mu = 10$, $\tau = 3$, and $\sigma = 4$.

Since the normal distribution is a conjugate prior with itself, we know that the form of the the posterior distribution must be:

$$
\begin{aligned}
P(X = x; \mu, \tau, \sigma) &= N\left(\frac{\tau^2}{\sigma^2+\tau^2}x + \frac{\sigma^2}{\sigma^2+\tau^2}\mu, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right) \\
P(X = 9.46; 10, 3, 4) &= N\left(\frac{3^2}{4^2+3^2}9.46 + \frac{4^2}{4^2+3^2}(10), \frac{4^2 3^2}{4^2+3^2}\right) \\
&= N\left(\frac{9}{25}9.46 + \frac{160}{25}, \frac{144}{25}\right) \\
&= N(9.81, 5.76)
\end{aligned}
$$

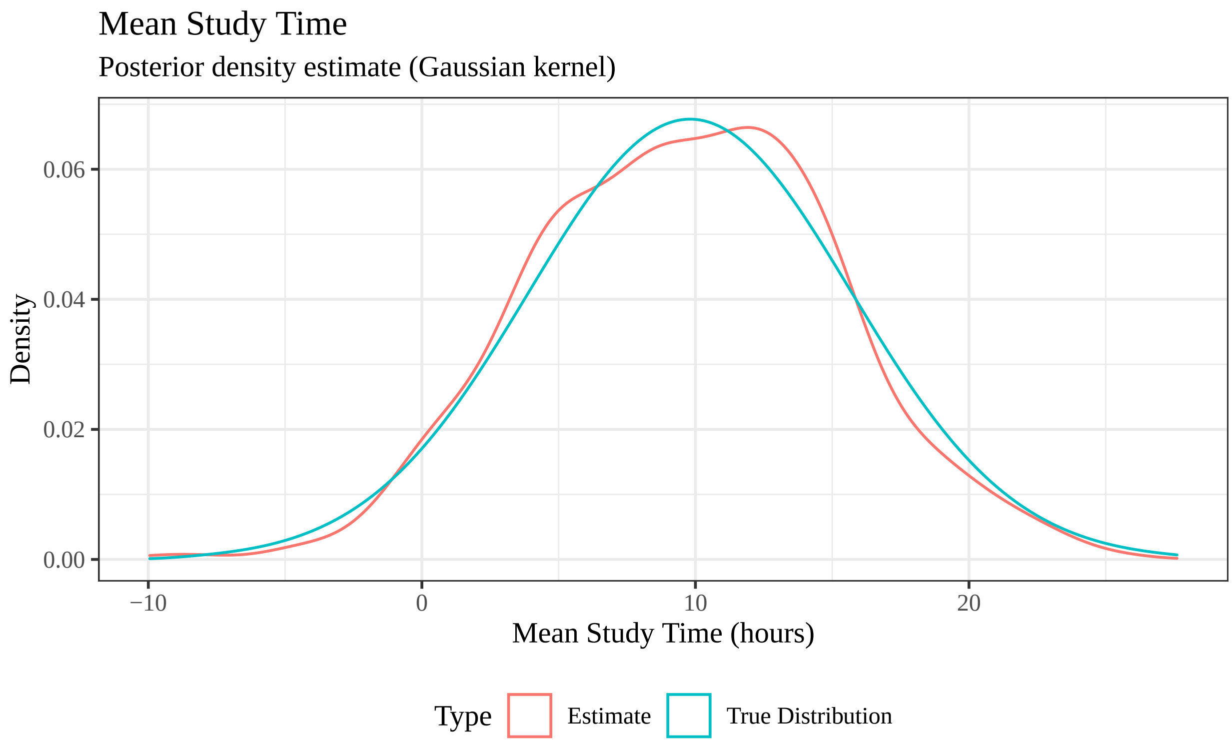
with $x = 9.46$, our observed mean.

We can draw $1,000$ new values from this posterior distribution to get a new sample. From this sample, we can get a 95% Bayesian confidence interval for θ by pulling the 2.5^{th} and 97.5^{th} percentile of the data.

Table 2: Posterior distribution summary

Mean	Median	95% Low	95% High
9.617	9.72	-0.803	20.439

Again, we can compare the posterior kernel density estimate of the new draws to the true distribution (with a Gaussian kernel):



Appendix

```r
# Prep work ---------------------------------------------------------------

# Load libraries
library(ggplot2)
library(tibble)
library(tidyr)

# Helper functions
post_draws_density <- function(data, ...) {


  ggplot(data, aes(...)) +
    geom_density() +
    theme_bw() +
    theme(
      text = element_text(family = "serif"),
      legend.position = "bottom"
    ) +
    labs(
      subtitle = paste0("Posterior density estimate (Gaussian kernel)"),
      y = "Density",
      color = "Type"
    )

}

# Control randomness
set.seed(9876)


# Question 1 --------------------------------------------------------------

prior_a <- 1
prior_b <- 1

n_married <- 112
divorced_n_obs <- 43


# Question 1a -------------------------------------------------------------

divorced_posterior <- function(n) {
  rbeta(n, prior_a + divorced_n_obs, n_married - divorced_n_obs + prior_b)
}


# Question 1b -------------------------------------------------------------

divorced_n_sims <- 1000

## NOTE: Not the way to do this, but still useful reference
#
# divorced_prob_obs <- divorced_n_obs / n_married
```

```r
# divorced_dist <- function(n, p) rbinom(n, n_married, p)
# divorced_prior_dist <- function(n) rbeta(n, 1, 1)
#
# divorced_prob_draws <- divorced_prior_dist(divorced_n_sims)
# divorced_dist_draws <- divorced_dist(divorced_n_sims, divorced_prob_draws)
#
# divorced_post_dist <- divorced_prob_draws[divorced_dist_draws == divorced_n_obs]

divorced_posterior_draws <- divorced_posterior(divorced_n_sims)

divorced_post_tbl <- tibble(
  Mean = mean(divorced_posterior_draws),
  Median = quantile(divorced_posterior_draws, .5),
  `95% Low` = quantile(divorced_posterior_draws, .025),
  `95% High` = quantile(divorced_posterior_draws, .975)
)

knitr::kable(
  divorced_post_tbl,
  booktabs = TRUE,
  digits = 3,
  caption = "Posterior distribution summary"
)


# Question 1c ------------------------------------------------------------

divorced_true_post <- tibble(
  value = qbeta(seq(0, 1, .001), 44, 77),
  type = "True Distribution"
)

divorced_est_post <- tibble(
  value = divorced_posterior_draws,
  type = "Estimate"
)

divorced_post_compare_tbl <- dplyr::bind_rows(
  divorced_true_post, divorced_est_post
)

divorced_post_plot <-
  post_draws_density(divorced_post_compare_tbl, x = value, color = type) +
  labs(
    title = "Probability of Divorce by 2015, Given Marriage in 2005",
    x = "Probability of Divorce"
  )

divorced_post_plot

# Question 2 -------------------------------------------------------------

study_obs <- c(
```

```r
  2.1, 9.8, 13.9, 11.3, 8.9, 15.7, 16.4, 4.5, 8.9, 11.9, 12.5, 11.1, 11.6,
  14.5, 9.6,  7.4,  3.3, 9.1,  9.4,  6.0, 7.4, 8.5,  1.6, 11.4,  9.7
)

study_obs_mean <- mean(study_obs)

study_true_sd <- 4
study_prior_mean <- 10
study_prior_sd <- 3

study_post_mean <- `+`(
  study_prior_sd^2 / (study_true_sd^2 + study_prior_sd^2) * study_obs_mean,
  study_true_sd^2 / (study_true_sd^2 + study_prior_sd^2) * study_prior_mean
)

study_post_sd <- study_true_sd^2 * study_prior_sd^2 / (study_true_sd^2 + study_prior_sd^2)


study_posterior <- function(n) rnorm(n, study_post_mean, study_post_sd)

study_n_sims <- 1000

study_posterior_draws <- study_posterior(study_n_sims)

study_post_tbl <- tibble(
  Mean = mean(study_posterior_draws),
  Median = quantile(study_posterior_draws, .5),
  `95% Low` = quantile(study_posterior_draws, .025),
  `95% High` = quantile(study_posterior_draws, .975)
)


knitr::kable(
  study_post_tbl,
  booktabs = TRUE,
  digits = 3,
  caption = "Posterior distribution summary"
)

study_true_post <- tibble(
  value = qnorm(seq(0, 1, .001), study_post_mean, study_post_sd),
  type = "True Distribution"
)

study_est_post <- tibble(
  value = study_posterior_draws,
  type = "Estimate"
)

study_post_compare_tbl <- dplyr::bind_rows(
  study_true_post, study_est_post
)
```

```r
divorced_post_plot <-
  post_draws_density(study_post_compare_tbl, x = value, color = type) +
  labs(
    title = "Mean Study Time",
    x = "Mean Study Time (hours)"
  )

divorced_post_plot
```