STA402L, Lab 2: The Beta-Binomial model

Due: XXX

Turning in solutions

This lab is part of Homework 2. Solutions to the exercises, as well as the non-lab homework exercises are to be written up and uploaded to Gradescope as a PDF.

Getting started

You will need the following R packages. If you do not already have them installed, please do so first using the `install.packages` function.

```
require(tidyverse)
require(rstanarm)
require(magrittr)
require(rstan)
```

While you will be expected to be able to write your own samplers for homework assignments (for the most part), we will rely on stan a lot for the labs. We will look at writing sampling methods in stan later on, but for now, you will be provided with the stan scripts needed for inference and the code needed to grab posterior samples from stan objects. Your focus then should be on using those posterior samples to answer several kinds of questions. For this lab, you will need two stan files `lab-02-pool.stan` and `lab-02-nopool.stan`, which you can download here:

- https://sta-602l-s21.github.io/Course-Website/labs/lab-02-nopool.stan; and
- https://sta-602l-s21.github.io/Course-Website/labs/lab-02-pool.stan.

Download both and make sure to save them in the same folder as the R script or R markdown file you are working from.

Also, start browsing some of the references on programming with Stan to gain some familiarity with stan.

Repeated Binomial Trials

By this point, you are familiar with binomial data: If $Y \sim Binom(n, \theta)$, we assume the data are such that over n trials with success probability θ, we observe y successes. Let us consider multiple binomial realizations. We have data on rat tumor development from Tarone (1982). Specifically, we have the number of incidences of endometrial stromal polyps in 71 different groups of female lab rats of type F344. We begin by loading in the data:

```
tumors <- read.csv(file = url("http://www.stat.columbia.edu/~gelman/book/data/rats.asc"),
                   skip = 2, header = T, sep = " ")[,c(1,2)]
y <- tumors$y
N <- tumors$N
n <- length(y)
```

Each row represents a group, or a draw from a binomial distribution. The y variable denotes the number of succcesses and the N variable denotes the total number of rats in that control group. For example the first group consists of 20 rats, with 0 of these 20 having developed a tumor. n is the number of groups.

If we assume that the probability of developing a tumor is the same across groups, then for each of the $i = 1, 2, \ldots, n$ groups, we have $y_i \sim Binom(N_i, \theta)$. We have learned that the Beta distribution is conjugate for Binomial data. For now, we place a $Beta(1, 1)$ prior on θ, which corresponds to a uniform density on the interval $[0, 1]$.

```r
plot(seq(0, 1, length.out = 1000),
     dbeta(seq(0, 1, length.out = 1000), 1, 1),
     type = 'l',
     xlab = expression(theta), ylab = "Density",
     main = "The Beta(1, 1) density")
```

```r
stan_dat <- list(n = n, N = N, y =y, a = 1, b = 1)
fit_pool <- stan('lab-02-pool.stan', data = stan_dat, chains = 2, refresh = 0)
pool_output <- rstan::extract(fit_pool)
mean(pool_output$theta)
```

---

1. Plot a histogram of θ from the `rstan` object called `pool_output`. Describe the distribution.

Alternatively, we may not have reason to believe that the probability of a rat developing a tumor should be the same across groups. Then we have the model $y_i \sim Binom(N_i, \theta_i)$ for $i = 1, 2, \ldots, n$. If we had expert knowledge about the different groups of rats, we might place different priors on each of the n θ_i's. However, for simplicity we choose to model the θ_i as i.i.d. $Beta(1, 1)$.

```r
stan_dat <- list(n = n, N = N, y =y, a = 1, b = 1)
fit_nopool <- stan('lab-02-nopool.stan', data = stan_dat, chains = 2, refresh = 0)
nopool_output <- rstan::extract(fit_nopool)
apply(nopool_output$theta,2,mean)
```

---

2. Visualize the posterior distributions of the θ_i with boxplots. In the plot, there should be one box and whiskers object for each θ_i.

What is actually being plotted here? What does each point represent?

---

3. Take a few minutes to look at the contents of the two files `lab-02-pool.stan` and `lab-02-nopool.stan`. How are they different?

Sensitivity analysis

With the Beta-Binomial model, we know that the posterior is $\theta|Y \sim Beta(a + \sum y_i, b + \sum N_i - \sum y_i)$. Therefore, the posterior mean is

$$E[\theta|Y] = \frac{a + \sum y_i}{a + b + \sum N_i}$$

We fit the above models with $a = 1, b = 1$, but it is good practice to perform an analysis to determine how sensitive the posterior is to the choice of prior. Considering the first model where we assumed the same success probability θ across groups, let us sample from the posterior distribution of θ over a range of a and b values. These parameter settings produce very different pictures of our prior beliefs about θ:

```
par(mfrow = c(4, 4))
par(mar=c(2,2,2,2))
for(a_val in c(1, 10, 25, 100)){
  for(b_val in rev(c(1, 10, 25, 100))){
    plot(seq(0, 1, length.out = 1000),
     dbeta(seq(0, 1, length.out = 1000), a_val, b_val),
     type = 'l',
     xlab = expression(theta), ylab = "Density",
     main = paste0("Beta(", a_val, ", ", b_val, ")"))
  }
}
```

To get samples from the posterior distribution of θ for each one of the prior distributions above, we run:

```
output_list <- list()
for(a_val in c(1, 10, 25, 100)){
  for(b_val in c(1, 10, 25, 100)){
    stan_dat <- list(n = n, N = N, y = y, a = a_val, b = b_val)
    fit_pool <- stan('lab-02-pool.stan', data = stan_dat, chains = 2, refresh = 0)
    output_list[[paste0("a_", a_val, ":b_", b_val)]] <- rstan::extract(fit_pool)[["theta"]]
  }
}
```

We then compile the samples from the different prior specifications into a data.frame, which will help us visualize the results.

```
output_list %>%
  plyr::ldply(function(theta){
    reshape2::melt(theta) %>%
      dplyr::mutate(post_mean = mean(theta))
  }, .id = "prior") %>%
  tidyr::separate("prior", into = c("a", "b"), sep = ":") %>%
  dplyr::mutate(a = as.numeric(gsub("._", "", a)),
                b = as.numeric(gsub("._", "", b))) %>%
  ggplot2::ggplot() +
  geom_density(aes(x = value)) +
  geom_vline(aes(xintercept = post_mean)) +
  facet_grid(a~factor(b, levels = rev(c(1, 10, 25, 100)))) +
  scale_colour_brewer(palette = "Set1") +
  labs(x = expression(theta), y = "Density")
```

In the plot above, increasing values of the parameter a are displayed moving from top to bottom along the vertical direction. Decreasing values of the parameter b are displayed moving from left to right along the horizontal direction. We can see that all of the posterior distributions look roughly normal with roughly equal variance. They are all fairly concentrated on values of θ within the range $[0.1, 0.2]$.

Here is a further observation that is specific to the concept of sensitivity analysis: for fixed a, as b increases the posterior mean shifts slightly towards lower values of θ. But for fixed b, as a increases the posterior mean shifts more dramatically towards higher values of θ. We might say that the posterior mean of θ is more sensitive to our prior beliefs about a than it is to our prior beliefs about b. Why might this be the case?

We can look at the formula for the posterior mean above to find an explicit answer. For that, recall that if $Y \sim Binom(N, \theta)$ and $\theta \sim Beta(a, b)$, then:

$$E[\theta|Y] = \frac{a+Y}{a+b+N} = \frac{a+b}{a+b+N}\left(\frac{a}{a+b}\right) + \frac{N}{a+b+N}\left(\frac{Y}{N}\right)$$

So the posterior mean is a weighted average of the prior mean and the MLE.

What might be a more intuitive explanation for the posterior's high sensitivity to the parameter a?

---

4. What observable quantity does the parameter a represent about our prior beliefs with respect to these data? What does b represent?
5. What do we actually observe in the rat tumor data with respect to these quantities?
6. How well do our different prior beliefs – the ones represented by the different parameter settings above – match up with the data?

Returning to the initial exploration where we considered a single θ versus allowing θ_i to vary across groups: You should have noticed that if we allow the groups to have different success probabilities, then our estimates $\hat{\theta}_i$ vary from 0.05 to 0.30. However when we assumed a single probability of success, we obtained $\hat{\theta} \approx 0.15$. In this first approach and assuming the 71 groups are independent, we essentially have one large binomial trial: $Y^* = \sum y_i \sim Binom(\sum N_i, \theta)$. Applying ideas from the sensitivity analysis:

---

7. Why might we have observed such a difference between the two approaches when using the prior $Beta(1, 1)$? Consider calculating the MLEs for θ and θ_i and comparing these values to the values obtained with the Bayesian approach:

```r
# approach 1
mle.1 <- sum(y)/sum(N)

# approach 2
mle.2 <- y/N
```

Acknowledgement

This lab was created by Jordan Bryan and Becky Tang.