Homework 7 Solution

Math 315, Fall 2019

1.

Suppose that $y_1, \ldots, y_n$ are a random sample from the Poisson/gamma density

$$f(y|\alpha, \beta) = \frac{\Gamma(y+\alpha)}{\Gamma(\alpha)y!} \cdot \frac{\beta^\alpha}{(\beta+1)^{y+\alpha}},$$

where $\alpha > 0$ and $\beta > 0$. This density is an appropriate model for observed counts that show more dispersion (i.e. variability) than predicted under a Poisson model. Suppose that (α, β) are assigned the noninformative prior proportional to $1/(\alpha\beta)^2$. If we transform to the real-valued parameters $\theta_1 = log(\alpha)$ and $\theta_2 = log(\beta)$, the posterior density is proportional to

$$p(\theta_1, \theta_2|y_1, \ldots, y_n) \propto \frac{1}{(\alpha\beta)^2} \prod_{i=1}^{n} \frac{\Gamma(y_i+\alpha)}{\Gamma(\alpha)y_i!} \cdot \frac{\beta^\alpha}{(\beta+1)^{y_i+\alpha}},$$

where $\alpha = e^{\theta_1}$ and $\beta = e^{\theta_2}$. Use this framework to model data collected by Gilchrist (1984), in which a series of 33 insect traps were set across sand dunes and the numbers of different insects caught over a fixed time were recorded. The number of insects of the taxa Staphylinoidea caught in the traps are given in the `insects.csv` data file and can be loaded using the below command:

```
insects <- read.csv("http://aloy.rbind.io/data/insects.csv")
```

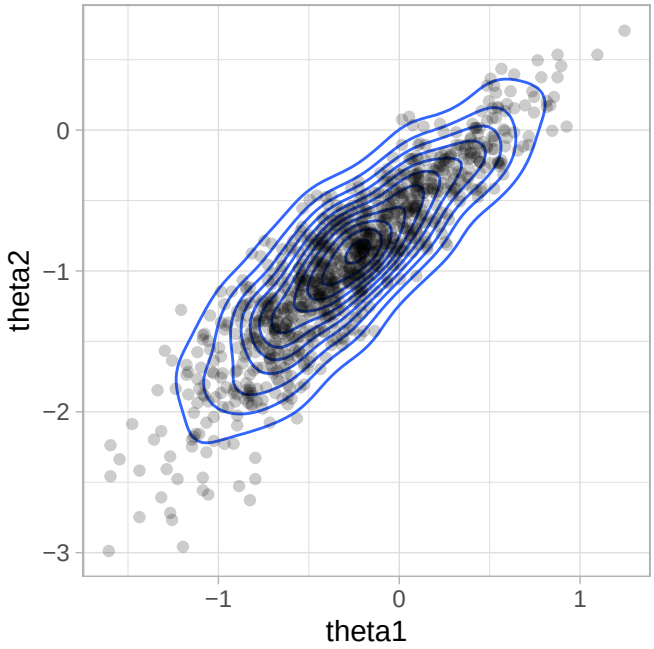
(a)

Use grid approximation to approximate the posterior density, and use Monte Carlo sampling to simulate 1000 draws from the joint posterior density of (θ_1, θ_2).

First, let's derive the log posterior for computational stability. (This wasn't required, but is a good idea. If you used a grid approximation on the original posterior, that's OK.) Notice that I formed the grid over (θ_1, θ_2) for computational stability. (Notice $\theta_1, \theta_2 \in \mathbb{R}$.)

```
param_grid <- expand.grid(
  theta1 = seq(-5, 5, length.out = 1000),
  theta2 = seq(-5, 5, length.out = 1000)
)

theta1 <- param_grid$theta1
theta2 <- param_grid$theta2

n <- nrow(insects)
y <- insects$n

logpost <- numeric(nrow(param_grid))
for(i in 1:n) {
  logpost <- logpost + lgamma(y[i] + exp(theta1)) - lfactorial(y[i])  -
  (y[i] + exp(theta1)) * log(exp(theta2) + 1)
}
logpost <- logpost - 2 * theta1 - 2 * theta2 - n * lgamma(exp(theta1)) + n * theta2 * exp(theta1)

unstd_posterior <- exp(logpost - max(logpost)) # numeric stability
```

```r
posterior <- unstd_posterior / sum(unstd_posterior)

mc_sample <- dplyr::sample_n(param_grid, size = 1000, replace = TRUE, weight = posterior)

library(ggplot2)
ggplot(mc_sample) +
  geom_density2d(aes(x = theta1, y = theta2)) +
  geom_point(aes(x = theta1, y = theta2), alpha = 0.2) +
  theme_light()
```



Note: the plot wasn't required, but it's a good sanity check.

Mathematical details: When I coded my log posterior, I pulled out the constants with respects to y_i from the sum. Below is my log posterior in terms of α and β:

$$\begin{aligned}\ell(\theta_1, \theta_2 | y_1, \ldots, y_n) &\propto -2\log(\alpha) - 2\log(\beta) + \sum_{i=1}^{n} [\log(\Gamma(y_i + \alpha)) - \log(\Gamma(\alpha)) - \log(y_i!) + \alpha\log(\beta) - (y_i + \alpha)\log(\beta + 1)] \\ &= -2\log(\alpha) - 2\log(\beta) - n\log(\Gamma(\alpha)) + n\alpha\log(\beta) + \sum_{i=1}^{n} [\log(\Gamma(y_i + \alpha)) - \log(y_i!) - (y_i + \alpha)\log(\beta + 1)]\end{aligned}$$

Then, plug in e^{θ_1} for α and e^{θ_2} for β.

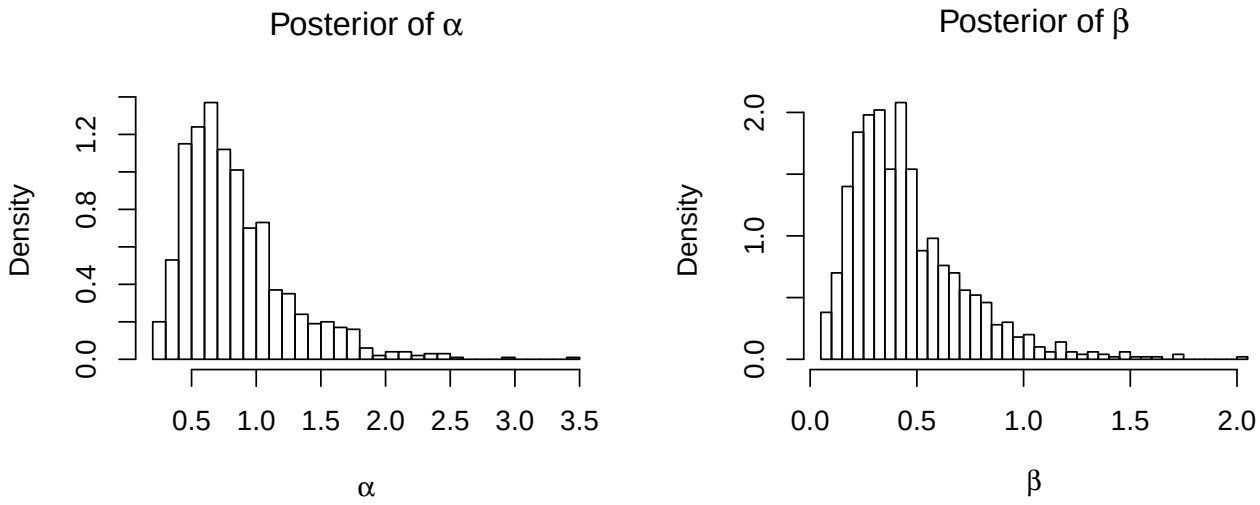
(b)

From your Monte Carlo sample, calculate 90% interval estimates for the parameters α and β.

Remember that you construct the marginal posterior distributions by simply grabbing the appropriate column of the `mc_sample` data frame. Here, we first need to extract the appropriate parameter and transform it back to α and β:

```
alpha_draws <- exp(mc_sample$theta1)
beta_draws <- exp(mc_sample$theta2)
```

Below is a plot of the marginal posteriors, but this wasn't required.



Given the data and the prior, there is a 95% chance that α is between 0.351 and 1.675.

Given the data and the prior, there is a 95% chance that β is between 0.147 and 0.976.

```
(alpha_ci <- quantile(alpha_draws, probs = c(0.05, 0.95)))
```

```
##        5%       95%
## 0.3513241 1.6745015
```

```
(beta_ci  <- quantile(beta_draws, probs = c(0.05, 0.95)))
```

```
##        5%       95%
## 0.1469864 0.9757761
```

2.

The following table gives the records of accidents in 1998 compiled by the Department of Highway Safety and Motor Vehicles in Florida.

	Fatal	Nonfatal
None	1601	162527
Seat belt	510	412368

Denote the number of accidents and fatalities when no safety equipment was in use by n_N and y_N , respectively. Similarly, let n_S and y_S denote the number of accidents and fatalities when a seat belt was in use. Assume that y_N and y_S are independent with $y_N \sim \text{Binomial}(n_N, p_N)$ and $y_S \sim \text{Binomial}(n_S, p_S)$. Assume a uniform prior is placed on the vector of probabilities (p_N, p_S). In this problem, treat n_S and n_N as fixed and known.

(a)

Derive the joint posterior distribution of (p_N, p_S).

$$\begin{aligned} p(p_N, p_S | y_N, y_S, n_N, n_S) &\propto \binom{n_N}{y_N} p_N^{y_N}(1-p_n)^{n_N - y_N} \cdot \binom{n_S}{y_S} p_S^{y_S}(1-p_S)^{n_S - y_S} \\ &\propto p_N^{(y_N+1)-1}(1-p_n)^{(n_N - y_N+1)-1} \cdot p_S^{(y_S+1)-1}(1-p_S)^{(n_S - y_S+1)-1} \end{aligned}$$

(b)

Show that p_N and p_S have independent beta posterior distributions.

The joint posterior distribution is the product of $p_N \sim \text{Beta}(y_N + 1, n_N - y_N + 1)$ and $p_S \sim \text{Beta}(y_S + 1, n_S - y_S + 1)$; thus, p_N and p_S are independent.

(c)

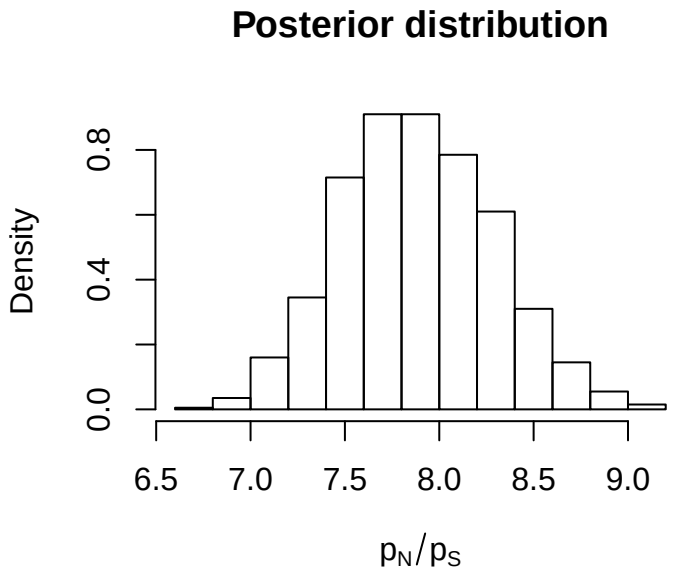
Use the function `rbeta()` to simulate 1000 values from the joint posterior distribution of (p_N, p_S).

```
p_n <- rbeta(1000, 1601 + 1, 162527 + 1)
p_s <- rbeta(1000,  510 + 1, 412368 + 1)
```

(d)

Using your sample, construct a histogram of the relative risk p_N/p_S. Calculate a 95% interval estimate of this relative risk.

```
hist(p_n / p_s, xlab = expression(p[N]/p[S]), main = "Posterior distribution", freq = FALSE)
```



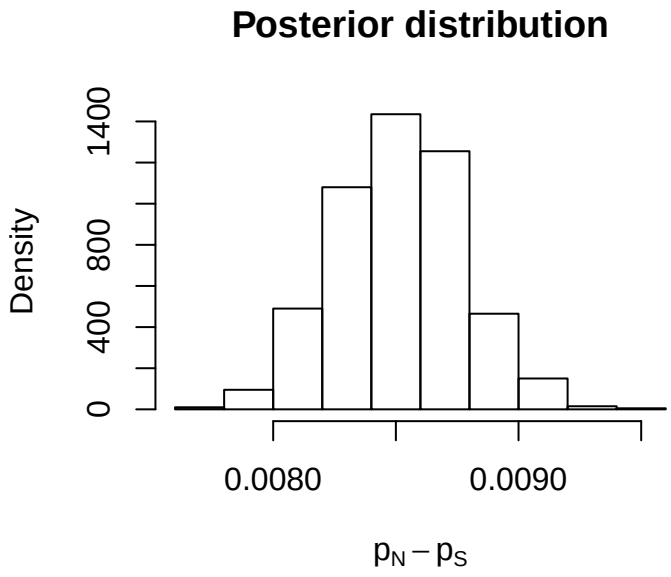

```
quantile(p_n / p_s, probs = c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 7.128767 8.706900
```

(e)

Construct a histogram of the difference in risks $p_N - p_S$.

```
hist(p_n - p_s, xlab = expression(p[N]-p[S]), main = "Posterior distribution", freq = FALSE)
```



(f)

Compute the posterior probability that the difference in risks exceeds 0.

Given the data and the prior, the probability that the difference in risks exceeds 0 is 1.

```
mean(p_n - p_s > 0)
```

```
## [1] 1
```