Comparing groups – Solution

Math 315, Adam Loy

Example 1. Comparing proportions across 3+ groups

A Washington Post-ABC News Tracking poll was conducted from November 3 to 6, 2016. A total of 2,220 likely voters were polled and asked to indicate their preference in the presidential election. Let $y_1 = 1043$ supported Clinton, $y_2 = 955$ supported Trump, and $y_3 = 222$ supported other candidates or expressed no opinion.

The counts an be modeled as arising from a multinomial distribution with sample size $n = 2220$ and respective probabilities θ_1, θ_2, and θ_3.

For this example, use a uniform (uninformative) prior on (θ_1, θ_2, θ_3); that is,

$$\pi(\theta_1, \theta_2, \theta_3) \propto 1.$$

(a) Derive the joint posterior distribution of $(\theta_1, \theta_2, \theta_3)$. Look in appendix A.1 of your textbook and determine which multivariate distribution matches your posterior.

$$p(\theta_1, \theta_2, \theta_3 \mid y_1, y_2, y_3) \propto 1 \cdot \frac{n!}{y_1! y_2! y_3!} \theta_1^{y_1} \theta_2^{y_2} \theta_3^{y_3}$$

$$\propto \theta_1^{y_1} \cdot \theta_2^{y_2} \theta_3^{y_3}$$

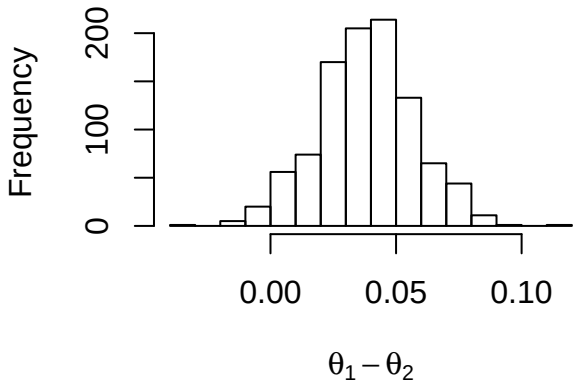
Thus $(\theta_1, \theta_2, \theta_3) \sim \text{Dirichlet}(y_1+1, y_2+1, y_3+1)$

(b) You can use the `rdirichlet()` function in the **MCMCpack** R package to simulate from your posterior from part (a). To do this, store the parameters from your posterior in the vector `alpha`, then draw 1000 simulations from the posterior.

```
# Fill in the blanks with the posterior parameter values
alpha <- c(1043 + 1, 955 + 1, 222 + 1)

# Draw samples from the posterior
library(MCMCpack)
theta <- rdirichlet(1000, alpha)
```

(c) Using your posterior draws, find the posterior density of $\theta_1 - \theta_2$, the difference in proportions between Clinton and Trump. Create a histogram of this posterior.

```
diff_prop <- theta[,1] - theta[,2]
hist(diff_prop, xlab = expression(theta[1] - theta[2]), main = "Posterior: Diff. in proportions")
```



(d) What is the posterior probability that a larger proportion of likely voters support Clinton?

Note: This estimate will change due to Monte Carlo error, but your solution should be close to mine.

```
mean(diff_prop > 0)
```

```
## [1] 0.974
```

Example 2. A Bayesian two-sample t-test

Suppose that we observe two independent normal samples, the first distributed according to an $\mathcal{N}(\mu_1, \sigma_1^2)$ distribution, the second according to an $\mathcal{N}(\mu_2, \sigma_2^2)$ distribution. Denote the first sample by $X_1, \ldots, X_m$ and the second sample by $Y_1, \ldots, Y_n$. Suppose also that the parameters $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ are assigned the vague prior

$$\pi(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) \propto \frac{1}{\sigma_1^2 \sigma_2^2}.$$

(a) Write down an expression for the posterior density.

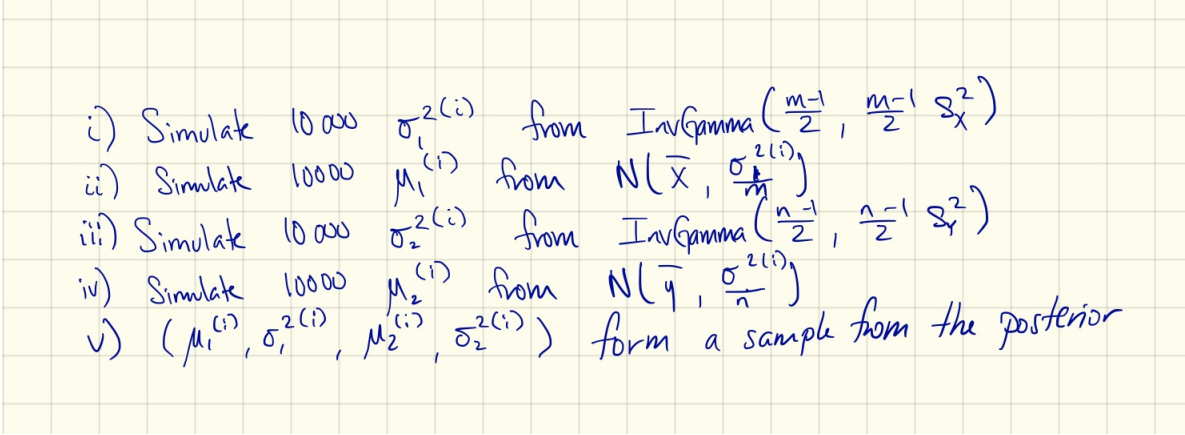
$$p(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2 \mid \text{data}) \propto \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(x_i - \mu_1)^2} \cdot \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2\sigma_2^2}(y_j - \mu_2)^2}$$

(b) Show that the vectors (μ_1, σ_1^2) and (μ_2, σ_2^2) are independent. To do this, show that the posterior distribution derived in (a) can be factored into the product of two posteriors, where these posteriors are the same posterior we derived during class on Monday for the one group problem.

$$p(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2 \mid \text{data}) \propto p(\mu_1, \sigma_1^2 \mid x_1, \ldots, x_m) \cdot p(\mu_2, \sigma_2^2 \mid y_1, \ldots, y_n)$$

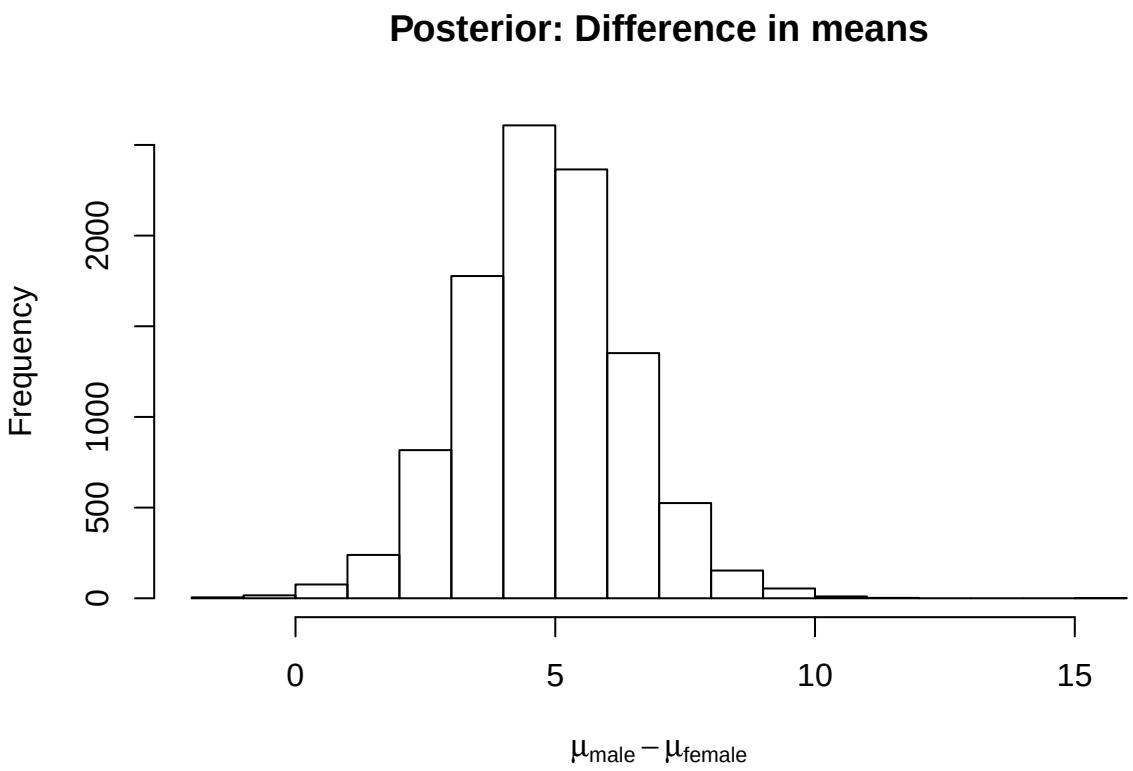
Thus, (μ_1, σ_1^2) and (μ_2, σ_2^2) are independent.

(c) Keeping in mind that (μ_1, σ_1^2) and (μ_2, σ_2^2) are independent, describe how to simulate from the joint posterior density of $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$. (Do this in words or using psuedo code.)



(d) The following data give the mandible lengths in millimeters for 10 male and ten female golden jackals in the collection of the British Museum. Using simulation, find the posterior density of the difference in mean mandible length between the sexes. Is there sufficient evidence to conclude that the males have a larger average?

```r
# Load the data
males   <- c(120, 107, 110, 116, 114, 111, 113, 117, 114, 112)
females <- c(110, 111, 107, 108, 110, 105, 107, 106, 111, 111)

# Calculating summary statistics
avg_male <- mean(males)
avg_female <- mean(females)

var_male <- var(males)
var_female <- var(females)

m <- length(males)
n <- length(females)

#Simulating from the posterior using conditional distributions
S <- 1e4

sigma2_male <- 1 / rgamma(S, (m - 1) / 2, ((m - 1) * var_male) / 2)
mu_male <- rnorm(S, avg_male, sqrt(sigma2_male / m))

sigma2_female <- 1 / rgamma(S, (n - 1) / 2, ((n - 1) * var_female) / 2)
mu_female <- rnorm(S, avg_female, sqrt(sigma2_female / n))

# Now, find posterior distribution of mean difference
post_diff <- mu_male - mu_female
hist(post_diff, xlab = expression(mu[male] - mu[female]), main = "Posterior: Difference in means")
```




```
# Calculate the posterior probability that mu[male] > mu[female]
mean(post_diff > 0)
```

```
## [1] 0.9979
```