BASKIN SCHOOL OF ENGINEERING
Department of Statistics
STAT 207 - Spring 2019

Name: ____________________________

Quiz 1

This quiz has two parts. Part 1 needs to be answered in class in the next 40 minutes. Once the class is over, please scan your answers and send the resulting file to my address `bruno@soe.ucsc.edu` within the following 10 minutes. Part 2 is a take home. The results need to be sent in pdf format in an e-mail to my address by 5 PM Thus 04/23/2018.
Please work individually and show your work in all the problems.

Part 1

Consider a random sample $y_1, \ldots, y_m$ of m binomial, $Bin(N, \theta)$, observations where both parameters, N, the number of trials, and θ, the probability of success are unknown.

1. Write the likelihood for N and θ given the observations. (2 pts)

2. Consider a prior for N given by a Poisson distribution with mean μ. Thus

$$p(N|\mu) = \frac{\mu^N e^{-\mu}}{N!}.$$

   Assume that $p(N, \mu, \theta) = p(N|\mu)p(\mu)p(\theta)$, where $p(\mu) \propto 1/\mu$ and $p(\theta)$ corresponds to a uniform distribution on $(0, 1)$. Obtain $p(N, \theta)$. Is this a proper distribution? (4 pts)

3. Obtain the posterior distribution $p(N, \theta|Y)$, where $Y = (y_1, \ldots, y_m)$. Obtain explicitly the two terms of the factorization $p(N, \theta|Y) = p(\theta|N, Y)p(N|Y)$. (5 pts)

4. Suppose that $m = 1$, is $p(N|Y)$ a proper distribution? (4 pts)

Hint:

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-x} dx; \quad \Gamma(x+1) = x\Gamma(x), \quad \Gamma(n) = (n-1)!$$

$$Be(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \ \ x \in (0, 1)$$

Part 2

The file `Covid19-04-13-20.csv`, available from the homework page of the class web site, contains data about the incidence and mortality due to COVID 19 in California, per county, as of 04/13/2020. The file consists of four columns: County name; Number of cases; Number of deaths; Population of the county. Let n_i be the number of cases for the i-th county. Let y_i be the corresponding number of deaths.

1. Perform an exploratory analysis of the data.

2. Consider the model
$$y_i \sim Bin(n_i, \theta), \quad \theta \sim Be(1/2, 1/2) \tag{1}$$
Obtain the posterior distribution of θ. Explore the results of fitting the California COVID 19 data using the samples obtained. Assuming that 20% of the population will become infected, what are the distributions of the number of deaths for each county?

3. Consider a second model that assumes the possibility that the data are overdispersed:
$$y_i \sim BeBi(n_i, \mu, \tau), \quad p(\mu, \tau) = \Big(\mu(1-\mu)(1+\tau)^2\Big)^{-1}. \tag{2}$$
Use a rejection sampling approach to obtain samples from $p(\mu, \tau|\boldsymbol{y})$ and apply your method to fit the California COVID 19 data. Are there any counties that are particularly influential in the analysis of the posterior for μ? Are there any important differences between the results for this model and those for model (1)?

4. Consider the hierarchical model
$$y_i \sim Bin(n_i, \theta_i), \quad \theta_i \sim Be(\mu\tau, (1-\mu)\tau), \quad p(\mu, \tau) = \Big(\mu(1-\mu)(1+\tau)^2\Big)^{-1}. \tag{3}$$
Write the posterior distribution of all model parameters as $p(\boldsymbol{\theta}, \mu, \tau|\boldsymbol{y}) = p(\boldsymbol{\theta}|\mu, \tau, \boldsymbol{y})p(\mu, \tau|\boldsymbol{y})$. Use this factorization to obtain samples from the posterior distribution of $\boldsymbol{\theta}, \mu$ and τ. Are there large differences between the counties? Compare the results from this model to those obtained from models (1) and (2).

5. Assuming that 20% of the population of California becomes infected, what are the probabilities, under the three different models, that more than 20,000 people will die of COVID 19?

Remember to use the template provided on the course web page. Start your paper with an abstract that contains a short description of the problem and the main findings. Then, the first part of the body of the paper will correspond to an introduction with a description of the problem and an exploratory data analysis. The methods and the analysis will follow. The paper will finish with concluding remarks and references. Tables and figures, if any, need to be part of the text. Do not append them to the end of the paper. You have a maximum of ten pages, including figures and tables. Do not include the code, but keep it tidy and commented, as you will have to make it available upon request.

(15 pts)

PS: For an interesting controversy regarding the incidence of Coronavirus in the Bay Area see https://www.mercurynews.com/2020/04/20/feud-over-stanford-coronavirus-study-the-authors-owe-us-all-an-apology/, or visit Andrew Gelman's blog.