OR 664 / SYST 664 / CSI 674: Final Exam

Due Monday May 18, 2020, 11:59 PM

Each question is worth ten points, for a total of 100 points for the final exam. Show all your reasoning. You will receive some credit for making an honest attempt and more credit if I can tell you were thinking along the right lines. ***Write up your solutions in a .doc, .docx, or pdf file. Your writeup for each problem should be self-contained and may include sections of code. You may attach other documents (R file, spreadsheet) separately for reference, but your solution must be understandable on its own. If you attach other documents, submit as separate files. Do not combine it all into a zip archive.*** *You are bound by the GMU honor code to work by yourself on the exam. I will be available by phone or email to answer clarification questions.*

1. The table below shows fecundity (number of eggs laid per female per day for the first 14 days of life) of three different strains of fruit flies.[1] Strain R was bred to be resistant to DDT; strain S was bred to be susceptible to DDT; and strain N was a nonselected control strain. The aim is to determine whether there is a difference in fecundity between selected and non-selected strains, and whether there is a difference between the two selected strains.

Resistant (R)	Susceptible (S)	Nonselected (N)
12.8	38.4	35.4
21.6	32.9	27.4
14.8	48.5	19.3
23.1	20.9	41.8
34.6	11.6	20.3
19.7	22.3	37.6
22.6	30.2	36.9
29.6	33.4	37.3
16.4	26.7	28.2
20.3	39.0	23.4
29.3	12.8	33.7
14.9	14.6	29.2
27.3	12.2	41.7
22.4	23.1	22.6
27.5	29.4	40.4
20.3	16.0	34.4
38.7	20.1	30.4
26.4	23.3	14.9
23.7	22.9	51.8
26.1	22.5	33.8
29.5	15.1	37.9
38.6	31.0	29.5
44.4	16.9	42.4
23.2	16.1	36.6
23.6	10.8	47.4

[1] Data set 22 from Hand et al. (1994). Original source: Sokal, R. R. and Rohlf, F.J. (1981) *Biometry*. 2nd edition. San Francisco: W.H. Freeman, 239. Data set available online at https://www2.stat.duke.edu/courses/Spring03/sta113/Data/Hand/fruitfly.dat

Assume the observations in each group are normally distributed with unknown group-specific means Θ_i and precisions P_i for i=1, 2, 3. Assume the parameters (Θ_i, P_i) are independent draws from a normal-gamma(μ, k, α, β) distribution. Find empirical Bayes estimates for the hyperparameters μ, k, α, and β as follows:

- Estimate the center μ as the grand mean of all the observations.
- Estimate the shape and scale as follows. Estimate the sample precisions $\hat{\rho}_1$, $\hat{\rho}_2$, $\hat{\rho}_3$ by calculating the sample variances and taking their inverses. Estimate the mean $\alpha\beta$ of the Gamma distribution as the average of $\hat{\rho}_1$, $\hat{\rho}_2$, $\hat{\rho}_3$. Estimate the variance $\alpha\beta^2$ as the sample variance of $\hat{\rho}_1$, $\hat{\rho}_2$, $\hat{\rho}_3$. Then solve for α and β.
- To estimate the precision multiplier k, first find the sample means $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$ for the three strains. Then calculate the sample variance of these three sample means, and invert to estimate the precision of the means. Divide this value by the average of $\hat{\rho}_1$, $\hat{\rho}_2$, $\hat{\rho}_3$ to estimate k, the ratio of the precision of the means to the precision of the observations.

Find the joint posterior distribution for Θ_1, Θ_2, and Θ_3. Find a 95% posterior credible interval for each Θ_i. Comment on your results, including whether the assumption that the observations are normally distributed is justified.

2. Using the posterior distributions from Problem 1, apply direct Monte Carlo to estimate the following probabilities: (1) the posterior probability that the mean fecundity Θ_3 for the nonselected group is the largest of the three group means; and (2) for three newly hatched individual fruitflies, one from each of the strains, the posterior probability that the non-selected fruitfly will have higher fecundity than the flies from the two selected strains. Clearly describe the process you use to find your Monte Carlo estimates. Discuss your results. Clearly explain the difference between the two estimated probabilities.

3. A blood test is designed to screen for a medical condition. The manufacturer claims the test has 90% accuracy. Based on this claim, the following prior distribution has been specified for the true positive and false positive rates of the test:
   - *True positive rate* (*sensitivity*). The probability of correctly detecting the condition if it is present has a Beta distribution with shape parameters 4.5 and 0.5.
   - *False positive rate* (1 – *specificity*). The probability that the test will erroneously report the condition if it is not present is independent of the sensitivity and has a Beta distribution with shape parameters 0.5 and 4.5.

   The test was performed on 30 blood samples for patients known to have the condition. The test correctly detected 28 out of the 30 known samples. The test was also performed on 50 samples from patients who did not have the disease. The test incorrectly reported the condition in 7 of these samples. Find the joint posterior distribution for the true positive and false positive rate of the test. Find 95% posterior credible intervals for the true and false positive rates.

4. Management at a call center is investigating the call load in order to find an efficient staffing policy. Assume that the number of calls per minute during the mid-morning period has a Poisson distribution with an unknown rate Λ. A non-informative prior distribution $g(\lambda) \propto \lambda^{-\frac{1}{2}}$ is defined on the positive real line for the unknown rate Λ. This is the Jeffreys prior, and

it is the limit of a gamma distribution with shape ½ and scale tending to infinity. Over a 1-hour period in mid-morning, 123 calls were logged. Find the posterior distribution for the unknown rate Λ.

5. Given the posterior distribution for the previous problem, find the probability that 40 or more calls will arrive in a future 15-minute mid-morning period.

6. A research group is studying violent crime in a city. The number of violent crimes in a 1-year period was recorded for 5 different wards in the city. The table below shows the data.

Ward	Number of Violent Crimes in Study Period
1	9
2	23
3	11
4	31
5	17

   - Assume that violent crime counts are independent Poisson random variables with ward-dependent means Λ_i, for i=1,…,5.
   - Assume that the means Λ_i are independent and identically distributed gamma random variables with shape α and scale β (or equivalently, shape α and mean $m = \alpha\beta$)
   - The mean $m = \alpha\beta$ of the gamma distribution is uniformly distributed on a grid of 50 equally spaced values starting at 5 and ending at 40.
   - The shape α is independent of the mean m and distributed on a grid of 50 equally spaced points starting at 1 and ending at 50, with prior probability proportional to $1/\alpha$.

   Use Gibbs sampling to draw 1000 samples from the joint posterior distribution of the mean M, the shape parameter A, and the five rate parameters Λ_i, i=1,…,5, conditional on the observed crime counts. Using your sample, calculate 95% credible intervals for the five rate parameters Λ_i. Discuss.

7. A company is considering purchasing a system for non-destructive testing of manufactured items. Assume that the system has a miss probability of 8% and a false alarm probability of 12%, where a miss is defined as failing to identify a defect when the item is defective, and a false alarm is defined as reporting a defect when there is none. The company is considering using the system to screen items for possible defects. Assume a loss of 0 for making the right choice: identifying defective items or not reporting when there is no defect. Assume that the loss for failing to identify a defect is 20 times the loss for reporting a defect when there is none. Let p be the prior probability of a defect. As a function of p, find the expected loss of three policies: (1) report all items as defective; (2) do not report any defects; and (3) report an item as defective if and only if the testing system identifies it as defective. For what range of p is each policy optimal? Comment on your results.

8. The table below contains data on ice cream consumption in pints per capita and mean daily temperature in degrees Fahrenheit over 30 four-week periods during the early 1950's.[2] The

[2] Data set 268 from Hand et al. (1994). Original source: Koteswaa Rao Kadiyala, Testing for the Independence of Regression Disturbances, *Econometrica* 38, 97-117, 970. https://www2.stat.duke.edu/courses/Spring03/sta113/Data/Hand/icecream.dat

original data set includes price per pint, weekly family income, and temperature as possible predictors. Of the three predictors, only temperature has any real effect on consumption, so we focus here on the relationship between consumption and temperature. Assume the relationship between temperature and consumption is linear with independent normally distributed errors. Assume a non-informative prior distribution $g(\eta, \beta, \rho) \propto \rho^{-1}$, for the transformed intercept η, the slope β, and the precision ρ of the regression line. Find the joint posterior distribution for (η, β, ρ). Find 95% credible intervals for the slope β and the untransformed intercept α. Comment on your results, including whether the assumptions for normal linear regression are met.

Consumption (pints)	Mean Temp (deg F)	Consumption (pints)	Mean Temp (deg F)
0.386	41	0.381	63
0.374	56	0.47	72
0.393	63	0.443	72
0.425	68	0.386	67
0.406	69	0.342	60
0.344	65	0.319	44
0.327	61	0.307	40
0.288	47	0.284	32
0.269	32	0.326	27
0.256	24	0.309	28
0.286	28	0.359	33
0.298	26	0.376	41
0.329	32	0.416	52
0.318	40	0.437	64
0.381	55	0.548	71

9. For the ice cream problem, find a 90% posterior predictive interval for ice cream consumption during a four-week period with mean daily temperature 46 degrees Fahrenheit and during a four-week period with mean daily temperature 98 degrees Fahrenheit. Which of your two predictive intervals would you trust more? Why?

10. A requirements engineering team conducted surveys of potential users of a new system. 9 out of 30 respondents stated that asynchronous interaction capability is "essential"; 15 out of 30 respondents said it was "desirable," and 6 out of 30 respondents said it was "unnecessary." Assume that the survey respondents are a representative sample from the target population of potential users. Assume a uniform prior distribution for the probability vector $(\theta_e, \theta_d, \theta_u)$ for the choice among the three responses by a user in the target population. Find the posterior distribution for the probability vector. Find 90% credible intervals for the probability that a user in the target population would choose each of the three responses. Comment on your results.