4
16
14
6
40

Name: Solutions

Bayesian Statistics, 22S:138
Final exam, 2010

1. Near Kiama, Australia, there is a hole in the cliffs through which ocean water erupts spectacularly with the rise and fall of waves. Jim Irish, Faculty of Engineering, University of Technology, Sydney, collected data on the waiting times (in seconds) between consecutive eruptions occurring on July 12, 1998.

The exponential distribution is often used to model waiting times between events. We plan to use Irish's data to estimate the exponential rate parameter λ in the population of all possible waiting times between eruptions. However, we are not sure whether the exponential distribution is a good fit for this population of waiting times.

As shown in your table of distributions, the variance of an exponential random variable is the square of the mean. Therefore, for an exponential random variable, the population *coefficient of variation*, defined as the 100 times the population standard deviation divided by the population mean, is 100.

$$CV = 100 \times \frac{\sigma}{\mu}$$

For sample data drawn from an exponential distribution, we would expect the sample CV to be close to 100. In the Kiama eruption data,

$$100 \times \frac{s}{\bar{x}} = 100 \times \frac{33.75051}{39.82812} = 84.7$$

Refer to the attached OpenBUGS code and output to answer the following questions. Note that the line

```
cvreal <- 100 * sd(times[1:N]) / mean(times[1:N])
```

calculates the sample CV from the observed data.

(a) Is the model given in the OpenBUGS code hierarchical? (yes / (no)) Briefly justify your answer There are only 2 stages, the likelihood and the prior on λ. A hierarchical model would have 3 or more stages.

2

(b) The `pred[i]`'s from each iteration are (circle one):

i. nuisance parameters

2

ii. missing values

iii. replicate datasets

iv. totally unnecessary for this analysis

v. none of the above

4

(c) What does the following line of code do?

```
cvpred <- 100 * sd(pred[1:N]) / mean(pred[1:N])
```

3 It computes the test quantity (also called discrepancy measure) for each replicate dataset.

(d) The trace plot of check (projected in color on the screen) looks wierd. Briefly explain why it would look like that.

3 The values of check are all 1's and 0's, so the plot has to jump back and forth between those two.

(e) What is the posterior predictive p-value? (numeric answer)

2 .1036

(f) Based on these results, what would be an appropriate next step in the Bayesian analysis (circle one):

i. Consider the analysis as presented to be adequate and report the inference on λ.

ii. Run more iterations of OpenBUGS for the exponential model because the sampler doesn't seem to have converged. Trace & BGR plots look fine.

iii. Try a more flexible model for these data instead of the exponential.

iv. None of the above.

Either is ok if you justify it.

(g) Briefly justify your answer to the previous question.

The PPP-val of .1036 is borderline - the real data is not terribly atypical of data drawn from an exponential model. However, the safest course would be to try a more flexible model and compare fit.

4

(h) Below is a list of distributions that we have studied. After each one, write "Yes" if it might be appropriate for waiting time data, and "No" if it clearly would not. Justify each answer in a few words.

i. beta no restricted to (0, 1)

ii. binomial no discrete

iii. gamma yes continuous on positive real line

iv. poisson no discrete

4

16

2

(i) If we used the Deviance Information Criterion to compare the exponential model to some other model, what value would you expect for `pD` in the exponential model? Give a numeric answer and briefly justify it.

1 There is only 1 parameter in this model.

(j) **Circle** all of the **true** statements in the following list:

5

i. The Gelman Rubin diagnostic plot for `lambda` shows failure to converge because not all of the lines are on top of each other.

ii. The numbers in the "MC error" column in WinBUGS/OpenBUGS output help us determine whether we have run enough MCMC sampler iterations.

iii. In order to use the Gelman-Rubin convergence diagnostic, one must run more than one chain.

iv. In choosing initial values for an MCMC sampler, one must not look at the current dataset being analyzed.

v. High autocorrelation in MCMC sampler output causes the Markov chain to converge slowly to its stationary distribution.

2. The breast cancer example in the first chapter of the textbook includes a table with the following information:

Model	Prior Probabilities	Likelihood for M+	Prior × Likelihood	Posterior Probabilities
H_0: No breast cancer	.9955	.0274	.0273	.893
H_a: Breast cancer	.0045	.724	.0033	.107
			.0306	1

Recall that $M+$ represented the woman having a positive mammogram. For the next two questions, write the required calculation in symbolic form using symbols such as $Pr(H_0)$ and $Pr(H_0|y)$. Then do the calculation to get a numeric answer.

2

(a) Using the data in this table, calculate the posterior odds in favor of the woman having breast cancer.

$$\frac{.107}{.893} = .1198$$

(b) Using the data in this table, calculate the Bayes factor in favor of the woman having breast cancer.

3

$$\frac{\frac{.107}{.893}}{\frac{.0045}{.9955}} = 26.48$$

(c) Explain briefly what your result in the previous problem means regarding the woman and breast cancer.

2

Strong evidence in the data alone that woman has breast cancer. For cases like this in which both H_0 and H_a are simple, the Bayes factor is not influenced by the prior at all.

3. Is the gamma family the conjugate prior for the exponential likelihood?

(a) Find the mathematical form of the posterior density $p(\lambda|y)$ for a $Gamma(\alpha, \beta)$ prior and an exponential likelihood with parameter λ. Show your work. Use your result to state whether the gamma family is the conjugate prior for the exponential likelihood.

4 $p(\lambda/y) \propto p(\lambda)\, p(y/\lambda)$

$= \lambda^{\alpha-1} e^{-\beta\lambda} \prod_{i=1}^{n} \lambda e^{-\lambda y_i}$

$= \lambda^{\alpha+n-1} e^{-\lambda(\beta+\Sigma y_i)}$

$= G(\alpha+n, \beta+\Sigma y_i)$

yes conjugate since prior and posterior are gamma. both.

(b) What is the posterior distribution of λ in problem 1 if the sum of the 64 waiting times in the sample data is 2549? If possible, identify it as a standard density.

2 $G(1+64, 60+2549)$