ISyE 8843 Bayes Statistics
Fall 2004

# Midterm

Version 2.0

James D. Delaney

October 27, 2004

## Problem 1: Nematodes

The Nematode data is an example of an unbalanced one-way layout. Assuming the one-way ANOVA model:

$$y_{ij} = \mu_i + \varepsilon_{ij} \qquad \varepsilon_{ij}|\sigma_i^2 \sim \mathcal{N}(0, \sigma_i^2)$$

The practical objective of the experiment would be to determine which nematocide is "best". That is, the question that needs to be answered is: *which nematocide results in the highest yield?* The equivalent statistical questions are: *a) Is there evidence to conclude that at least one treatment mean is different from the other two?* and *b) If so, is there a treatment mean that is significantly larger than the other two?* The typical frequentist approach in the ANOVA framework to answering these questions are an *F-test* and then possibly *simultaneous t-tests* for comparing treatment means.

For this exercise, a hierarchical Bayesian framework is suggested. The explicit forms of the full conditionals are provided. This enables an estimate of the joint posterior distribution of the treatment means to be constructed via Gibbs sampling MCMC. The code to perform this simulation was written in *R*. The output was $40,000$ 9-dimensional simulated points, the first $10,000$ of which are discarded as "burn-in". Below are some time series plots and histograms of some of the simulated data. The marginal densities of the treatment means seem to be roughly symmetric, bell-shaped. Perhaps that for $\mu_1$ is a little skewed to the downside. The histograms of the within-treatment variation and the three non-negative hyper-parameters seem to be skewed to varying degrees to the upside. That for $\tau^2$ being very highly skewed. Some approximate summary statistics for the simulated parameters are provided below:

| Statistic | $\mu_1$ | $\mu_2$ | $\mu_3$ | $\sigma_1^2$ | $\sigma_2^2$ | $\sigma_3^2$ | $\sigma^2$ | $\psi$ | $\tau^2$ |
|---|---|---|---|---|---|---|---|---|---|
| mean | 26.64 | 20.46 | 21.64 | 30.36 | 47.36 | 39.65 | 20.46 | 22.50 | 22.10 |
| var | 3.75 | 3.58 | 2.78 | 424.38 | 618.19 | 382.70 | 222.11 | 8.14 | 1753.07 |

To address the question of most practical interest, the simulated joint density of the three contrasts: $\mu_2 - \mu_1$, $\mu_3 - \mu_1$, and $\mu_3 - \mu_2$ is estimated based on the simulated data. From this a (joint) 95% credible set on the contrasts can be examined to see if it contains the point $[0, 0, 0]$. This is equivalent to testing the hypothesis: $H_0 : \mu_1 = \mu_2 = \mu_3$ vs. $H_A$ : *at least one* $\mu_i$ *is different.* That is, reject $H_0$ if the credible set does not contain the origin. Since the sample variances of the $\hat{\mu}_i$'s are nearly the same for each $i = 1, 2, 3$, a 95% *credible sphere* is constructed, centered at the contrast means: $[-6.19, -5.00, 1.18]$. This is approximated to be the set such that the $L^2$-*distance* from the mean is less than or equal to 7.65. Since $\sqrt{6.19^2 + 5.00^2 + 1.18^2} \approx 8.04$ the origin is *not* in the 95% credible sphere, so it could be concluded that at least one of the contrasts is significantly different from 0. From the mean contrast vector, it is clear that at the very least, treatment 1 is significantly better than treatment 3. In addition, the contrasts can be tested individually, which would reveal that the first two contrasts are significantly different from zero, whereas there is not enough evidence to conclude that the third contrast is different from zero. Incidentally, referring to the *R* output, scaling the contrasts by their sample standard deviations to form a 95% *credible* **ellipsoid** leads to the same conclusion, that:

**Use of nematocide 1 results in higher tomato yields than the other two nematocides studied**.

## *R* Code

```
GibbsANOVA<-function(data0,par0=c(1,1,1,1,0.1,10,0.1),M=40000) {
# data0 is a data.frame of summary data named: ybar, s2, ni
# The length of each column is the number of treatments, k.
# par0=c(a_0, c_0, d_0, f_0, g_0, psi_0, zeta_0)
# theta=c(mu_1,...mu_k,sigma_1^2,...sigma_k^2,sigma^2,psi,tau^2)
# The starting values for theta are set somewhat arbitrarily.
k<-length(data0)
# set starting values
theta<-matrix(0,M,2*k+3)
theta[1,]<-c(rep(par0[6],k),rep(par0[4]/par0[5],(k+1)),par0[6],par0[3]/par0[2])
# cycle through generating random variates from full conditionals
for(i in 2:M) {
# mu_i's
num<-(data0$ybar*data0$ni)/theta[(i-1),(k+1):(k+3)]+
theta[(i-1),(2*k+2)]/theta[(i-1),(2*k+3)]
den<-data0$ni/theta[(i-1),(k+1):(k+3)]+1/theta[(i-1),(2*k+3)]
theta[i,1:k]<-rnorm(n=k,mean=num/den,sd=sqrt(1/den))
# psi
num<-sum(theta[i,1:k])+par0[7]*par0[6]
den<-k+par0[7]
theta[i,(2*k+2)]<-rnorm(n=1,mean=num/den,
sd=sqrt(theta[(i-1),(2*k+3)]/den))
# tau^2
theta[i,(2*k+3)]<-1/rgamma(n=1,shape=(par0[2]+k+1)/2, rate=(par0[3]+sum(
(theta[i,1:k]-theta[i,(2*k+2)])^2)+par0[7]*(theta[i,(2*k+2)]- par0[6])^2)/2)
# sigma_i^2
theta[i,(k+1):(k+3)]<-1/rgamma(n=k,shape=(par0[1]+data0$ni)/2,
rate=(par0[1]*theta[(i-1),(2*k+1)]+(data0$ni-1)*data0$s2+
data0$ni*(data0$ybar-theta[i,(1:k)])^2)/2)
# sigma^2
theta[i,(2*k+1)]<-rgamma(n=1,shape=(par0[4]+k*par0[1])/2,
rate=(par0[5]+par0[1]*sum(1/theta[i,(k+1):(k+3)]))/2)
}
return(theta)
}


# example of testing contrasts: 3 treatments (95% credible ellipsoid)
#> ANOVAsims<-GibbsANOVA(data0)
#> c1<-ANOVAsims[10001:40000,2]-ANOVAsims[10001:40000,1]
#> c2<-ANOVAsims[10001:40000,3]-ANOVAsims[10001:40000,1]
#> c3<-ANOVAsims[10001:40000,3]-ANOVAsims[10001:40000,2]
#> contrasts0<-cbind(c1,c2,c3)
#> cov(contrasts0)
#          c1       c2        c3
#c1  7.171884 3.753002 -3.418882
#c2  3.753002 6.068929  2.315927
#c3 -3.418882 2.315927  5.734809
#> contrastss<-contrasts0/sqrt(diag(cov(contrasts0)))
#> apply(contrastss,1,mean)
#        c1         c2         c3
#-2.3098178 -2.0306339  0.4941118
#> sum(sqrt((contrastss[1,]+2.31)^2+(contrastss[2,]+2.03)^2+(contrastss[3,]-0.494)^2)<3.04)/30000
#[1] 0.9496
#> sqrt((2.31)^2+(2.03)^2+(-0.494)^2)
#[1] 3.114649
#> sum(sqrt((contrastss[1,]+2.31)^2+(contrastss[2,]+2.03)^2+(-0.494)^2)<3.04)/30000
#[1] 0.9769
#so treatment 2 and 3 are not significantly different.

#whereas treatment 1 differs from both treatment 2 and treatment 3:
#> sum(sqrt((contrastss[1,]+2.31)^2+(+2.03)^2+(contrastss[3,]-0.494)^2)<3.04)/30000
#[1] 0.9108
#> sum(sqrt((+2.31)^2+(contrastss[2,]+2.03)^2+(contrastss[3,]-0.494)^2)<3.04)/30000
#[1] 0.8581667
```

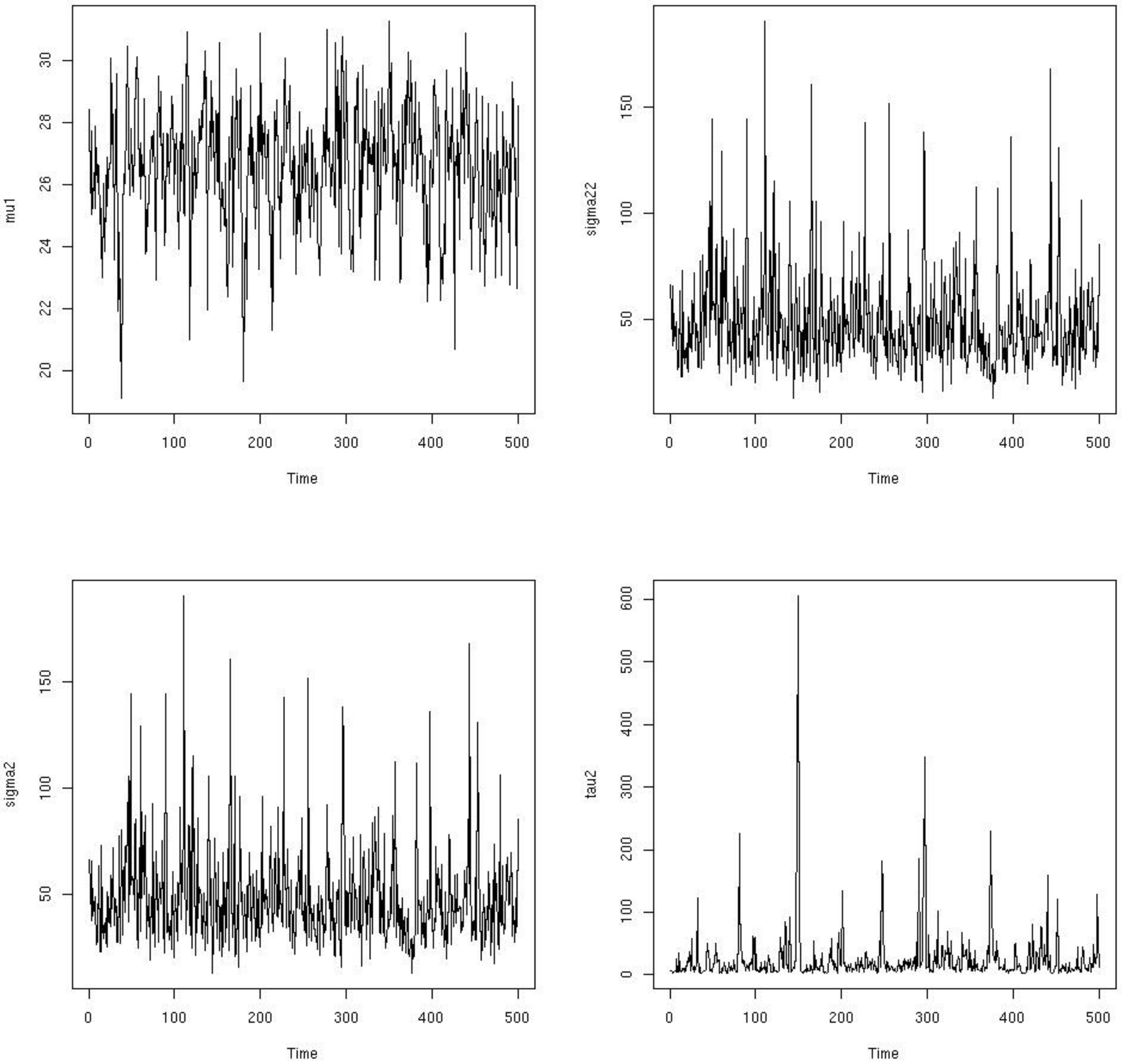

Figure 1: Last 500 Gibbs Samples of some selected Parameters

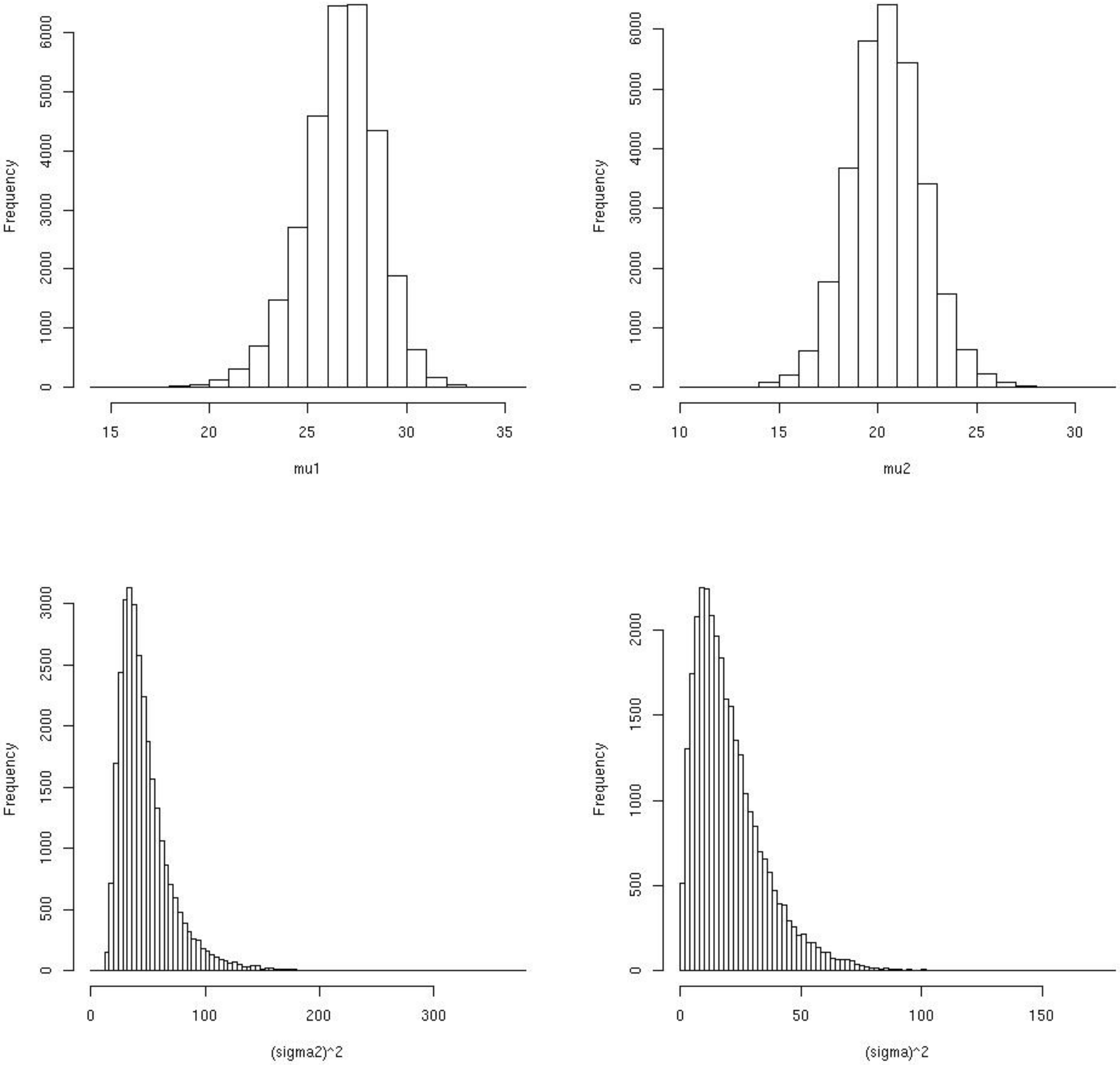

Figure 2: Marginal Histograms of Last 30,000 Parameter Gibbs Samples

**Problem 2: Rainfall Data in Marquitia**

**a) Posterior Density:**

In general,

$$f(\mu, \sigma|y_1, \ldots y_n) \propto \pi(\mu, \sigma) \prod_{i=1}^{n} f(y_i|\mu, \sigma)$$

So using the priors for $\mu$ and $\sigma$ and the likelihood specified for this problem:

$$f(\mu, \sigma|y_1, \ldots y_n) \propto \frac{1}{10\sqrt{2\pi}} \exp{(-\frac{\mu^2}{200})} \frac{1}{10\sigma\sqrt{2\pi}} \exp{(-\frac{(\ln\sigma)^2}{200})} \sigma^{-n} \exp\left\{-\sum_{i=1}^{n}\left[\frac{y_i-\mu}{\sigma} + \exp\left(-\frac{y_i-\mu}{\sigma}\right)\right]\right\}$$

Which can be simplified a little bit:

$$f(\mu, \sigma|y_1, \ldots y_n) \propto \sigma^{-(n+1)} \exp\left\{-\sum_{i=1}^{n}\left[\frac{y_i}{\sigma} + \exp\left(-\frac{y_i-\mu}{\sigma}\right)\right] - \frac{n\mu}{\sigma} - \frac{\mu^2+(\ln\sigma)^2}{200}\right\}$$

Utilizing the 48 observations $y_1 = 154, y_2 = 49.6, \ldots, y_{48} = 44.3$:

$$f(\mu, \sigma|154, \ldots, 44.3) \propto \sigma^{-49} \exp\left\{-\left[\frac{154+\ldots+44.3}{\sigma} + e^{-\frac{154-\mu}{\sigma}} + \ldots + e^{-\frac{44.3-\mu}{\sigma}}\right] - \frac{48\mu}{\sigma} - \frac{\mu^2+(\ln\sigma)^2}{200}\right\}$$

**b) Metropolis-Hastings Algorithm:**

Earlier in the semester, I adapted the Metropolis for Weibull practice example to *R*. For this problem, only a little modification is required to make it *Metropolis for Gumbel.* For simplicity, the proposal joint density is assumed to be the product of densities for $\mu'$ and $\sigma'$. To enable the flexibility of two tuning parameters, these are chosen to be $\mathcal{N}(\mu, s_1^2)$ and $\mathcal{LN}(\log(\sigma), s_2^2)$, respectively, giving $\mu'$ support on $\mathscr{R}$ and $\sigma'$ support on $\mathscr{R}^+$. That is:

$$q(\mu', \sigma'|\mu, \sigma) = \frac{1}{2\pi\sigma s_1 s_2} \exp\left\{-\frac{(\log(\sigma') - \log(\sigma))^2}{2s_2^2} - \frac{(\mu'-\mu)^2}{2s_1^2}\right\}$$

As can be seen in the time series plots of the simulated parameter values, with this proposal distribution, the simulated bivariate posterior distribution could possibly be improved to mix better than with the "tuning" parameter pair: $s_1 = 2, s_2 = 2$. A little bit of tuning, say $s_1 = 0.5, s_2 = 0.1$ could be used to arrive at better performance. The density of $\mu$ seems to be skewed slightly to the left and the marginal density of $\sigma$ is skewed to the right. The histograms for $\mu$ and $\sigma$ are attached. Summary statistics are as follows (after a burn-in of 10000, 30000 realizations from the bivariate posterior distribution are used):

| | | |
|---|---|---|
| $\mu$ | mean | 45.77 |
| | var | 11.22 |
| $\sigma$ | mean | 22.44 |
| | var | 8.56 |

**c) Prediction:**

There are several ways that the probability: $P(y^* \geq 410|y)$ can be estimated based on the above output. One possibility is to simply "plug-in" the means of the simulated parameters into the Gumbel cdf. This yields a probability of, $8.953704e-08$, very close to zero. A better way to estimate this probability is based on actually estimating the *posterior predictive distribution.* This is like treating $F(y^*|y)$ as a mixture of $30,001$ equally weighted Gumbel distributions with parameters from the sequence of simulated parameter pairs. So for each of the $30,001$ pairs, the expression $1 - F(410|\mu, \sigma)$ can be evaluated. Then, the estimate would be the mean of *this* posterior distribution. In this case, $\widehat{P(y^* \geq 410|y)} = 4.6021e - 07 \approx \mathbf{0}$. That is, the probability of observing a maximum daily rainfall as large or greater than that tragic day in 1999 in any given year is essentially zero.

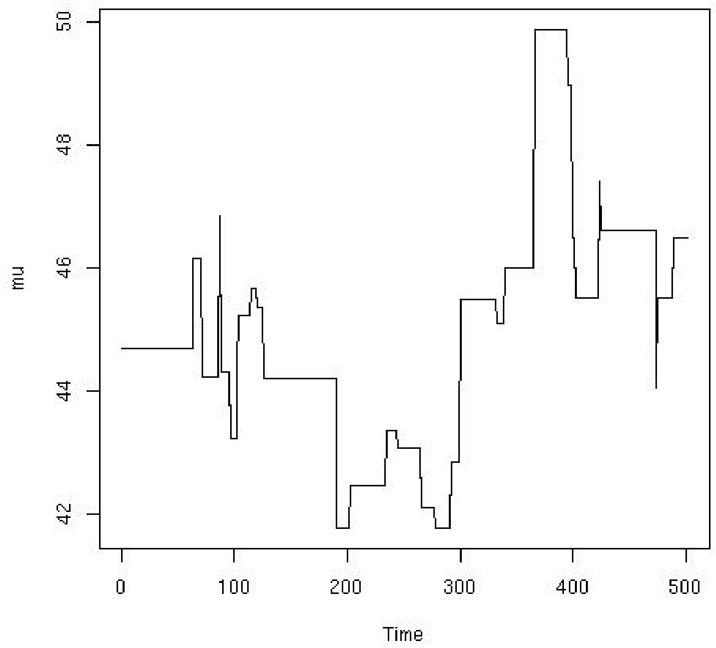

Figure 3: Last 500 Simulations of the Parameter $\mu$ $s_1 = 2, s_2 = 2 \ldots$Poor Mixing

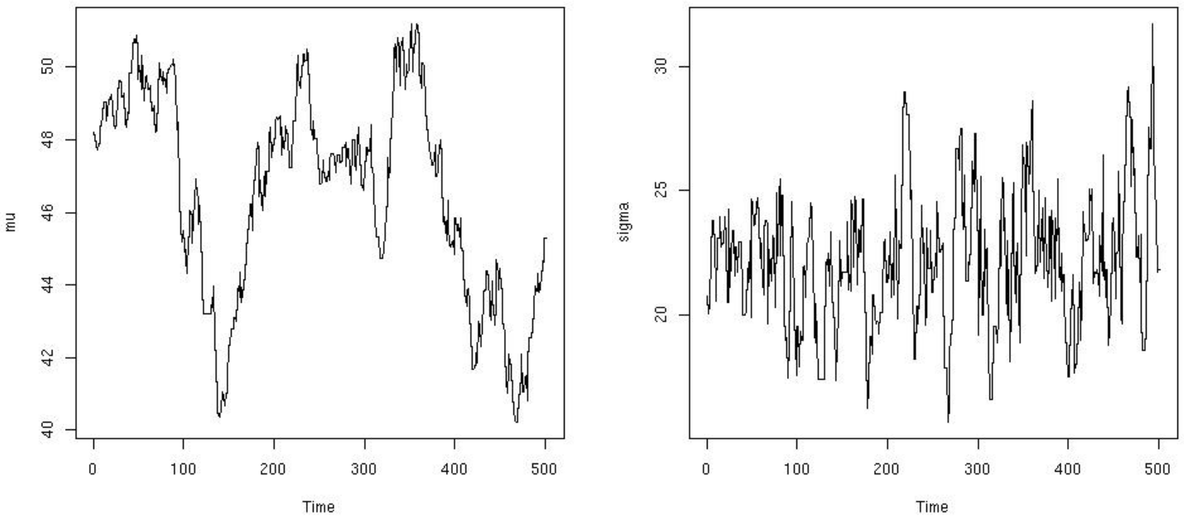

Figure 4: Last 500 Simulations of the Parameters $s_1 = 0.5, s_2 = 0.1$

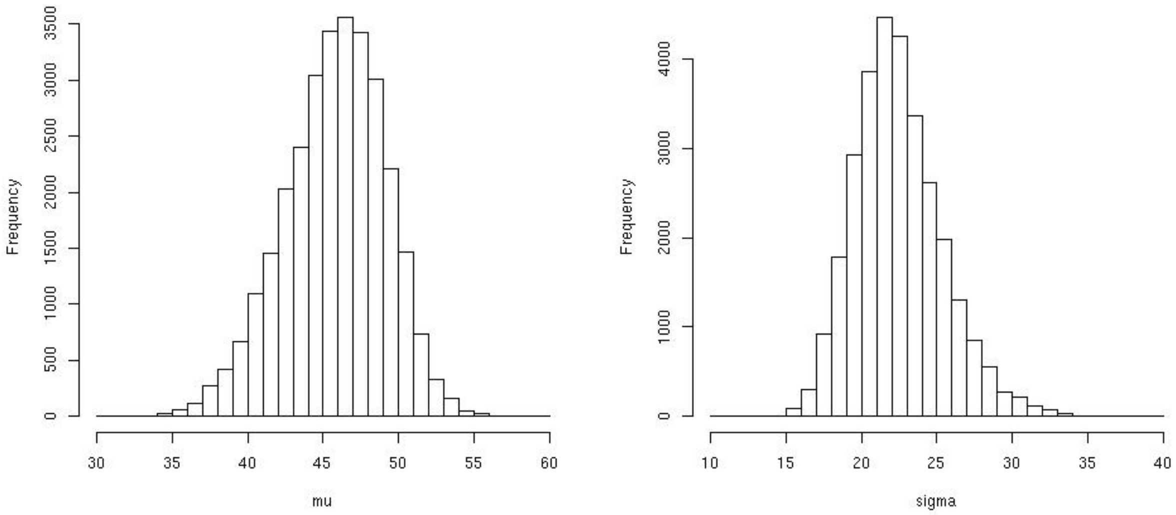

Figure 5: Marginal Histograms of Last 30,000 Parameter Pairs $s_1 = 0.5, s_2 = 0.1$

### *R* Code

```
MH4gum<-function(data0,s1=1,s2=1,mu0=0,sigma0=1,M=40000) {
n<-length(data0) # two tuning parameters s1 and s2.
ru<-runif(M)   # we will be testing rho M times.
rn<-rnorm(M) # these random variates can be
re<-rnorm(M) # updated to reflect their means.
xi<-matrix(0,M,2)
xi[1,]<-c(mu0,sigma0)

for (i in 2:M) {
xi[i,1]<-xi[i-1,1]+rn[i]*s1
xi[i,2]<-xi[i-1,2]*exp(re[i]*s2)
rho<-(xi[i-1,2]/xi[i,2])^n*exp(sum(data0/xi[i-1,2]-data0/xi[i,2]+
exp(-(data0-xi[i-1,1])/xi[i-1,2])-exp(-(data0-xi[i,1])/xi[i,2]))+
n*(xi[i,1]/xi[i,2]-xi[i-1,1]/xi[i-1,2])+(xi[i-1,1]^2-xi[i,1]^2+
log(xi[i-1,2])^2-log(xi[i,2])^2)/200+xi[i,2]/xi[i-1,2]-xi[i-1,2]/xi[i,2])
if(ru[i]>rho)
xi[i,]<-xi[i-1,]
}
return(xi)
}

pgumbel<-function(x,mu,sigma) {
return(exp(-exp(-(x-mu)/sigma)))
}
```

**Problem 3: MARTA Passenger Arrivals (BUGS Example)**

For this hierarchical Bayesian nonparametric problem it was not difficult to have *BUGS* simulate from the posterior distribution of the model parameters because of the excellent, clear instructions provided in *Handout 14*. Some traces of the simulated values were examined and appeared to exhibit no indication that the simulation required any tuning. This page is attached. The summary posterior statistics were found directly from within *BUGS*:

| **node** | **mean** | **sd** | **MC error** | **2.5%** | **median** | **97.5%** | **start** | **sample** |
|---:|---|---|---:|---:|---:|---:|---|---:|
| lambda | 8.63700 | 0.66660 | 0.004298 | 8 | 9 | 10 | 10001 | 30000 |
| P[1] | 0.02030 | 0.01998 | 1.121E-4 | 5.436E-4 | 0.01431 | 0.07419 | 10001 | 30000 |
| P[2] | 0.02037 | 0.01997 | 1.158E-4 | 5.062E-4 | 0.01423 | 0.07416 | 10001 | 30000 |
| P[3] | 0.02063 | 0.02028 | 1.175E-4 | 5.767E-4 | 0.01460 | 0.07552 | 10001 | 30000 |
| P[4] | 0.04080 | 0.02792 | 1.638E-4 | 0.005010 | 0.03477 | 0.11040 | 10001 | 30000 |
| P[5] | 0.04071 | 0.02801 | 1.555E-4 | 0.005076 | 0.03449 | 0.11090 | 10001 | 30000 |
| P[6] | 0.06124 | 0.03401 | 2.030E-4 | 0.013010 | 0.05519 | 0.14200 | 10001 | 30000 |
| P[7] | 0.06162 | 0.03430 | 1.990E-4 | 0.013050 | 0.05546 | 0.14350 | 10001 | 30000 |
| P[8] | 0.09009 | 0.04188 | 2.804E-4 | 0.026620 | 0.08411 | 0.18750 | 10001 | 30000 |
| P[9] | 0.09164 | 0.04199 | 2.523E-4 | 0.026740 | 0.08639 | 0.18850 | 10001 | 30000 |
| P[10] | 0.10390 | 0.04330 | 2.607E-4 | 0.035670 | 0.09861 | 0.20240 | 10001 | 30000 |
| P[11] | 0.12260 | 0.04625 | 2.609E-4 | 0.046680 | 0.11760 | 0.22650 | 10001 | 30000 |
| P[12] | 0.10220 | 0.04284 | 2.438E-4 | 0.034300 | 0.09691 | 0.19940 | 10001 | 30000 |
| P[13] | 0.08148 | 0.03862 | 2.142E-4 | 0.023130 | 0.07580 | 0.17240 | 10001 | 30000 |
| P[14] | 0.06095 | 0.03376 | 1.894E-4 | 0.013140 | 0.05521 | 0.14160 | 10001 | 30000 |
| P[15] | 0.04076 | 0.02794 | 1.531E-4 | 0.005245 | 0.03455 | 0.11040 | 10001 | 30000 |
| P[16] | 0.02048 | 0.01992 | 1.157E-4 | 5.324E-4 | 0.01446 | 0.07309 | 10001 | 30000 |
| P[17] | 0.02018 | 0.01981 | 1.114E-4 | 5.219E-4 | 0.01407 | 0.07420 | 10001 | 30000 |

The main parameter of interest is the arrival rate, $\lambda$. The posterior mean of $\lambda$ is 8.64. The median is 9 passengers every four minutes. Either number could be justified as an estimate of the passenger arrival rate per four minute interval. *BUGS* provides an easy way to save the simulated parameter values, in order, to a text file. This then enables the data to be easily imported into another environment, such as *R* or *MATLAB*, for data analysis and graphing. In this example, *MATLAB* was used to provide the histograms for $\lambda$, $p_2$, $p_7$, $p_9$, $p_{10}$, $p_{13}$, and $p_{17}$. The histograms illustrate that $\lambda$ is pretty much confined to the five integers 7, 8, 9, 10, and 11. The mode being 9. What can also be seen is that the probabilities that correspond to integers far away from 9 seem to have a mode at zero. Whereas, the other four are very similar in shape with modes in the general vicinity of 0.10.

# References

[1] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Boca Raton: Chapman and Hall/CRC, 2nd edition, 2004.

[2] Brani Vidakovic. Lecture notes for isye 8843. *Working Manuscript*, 2004.

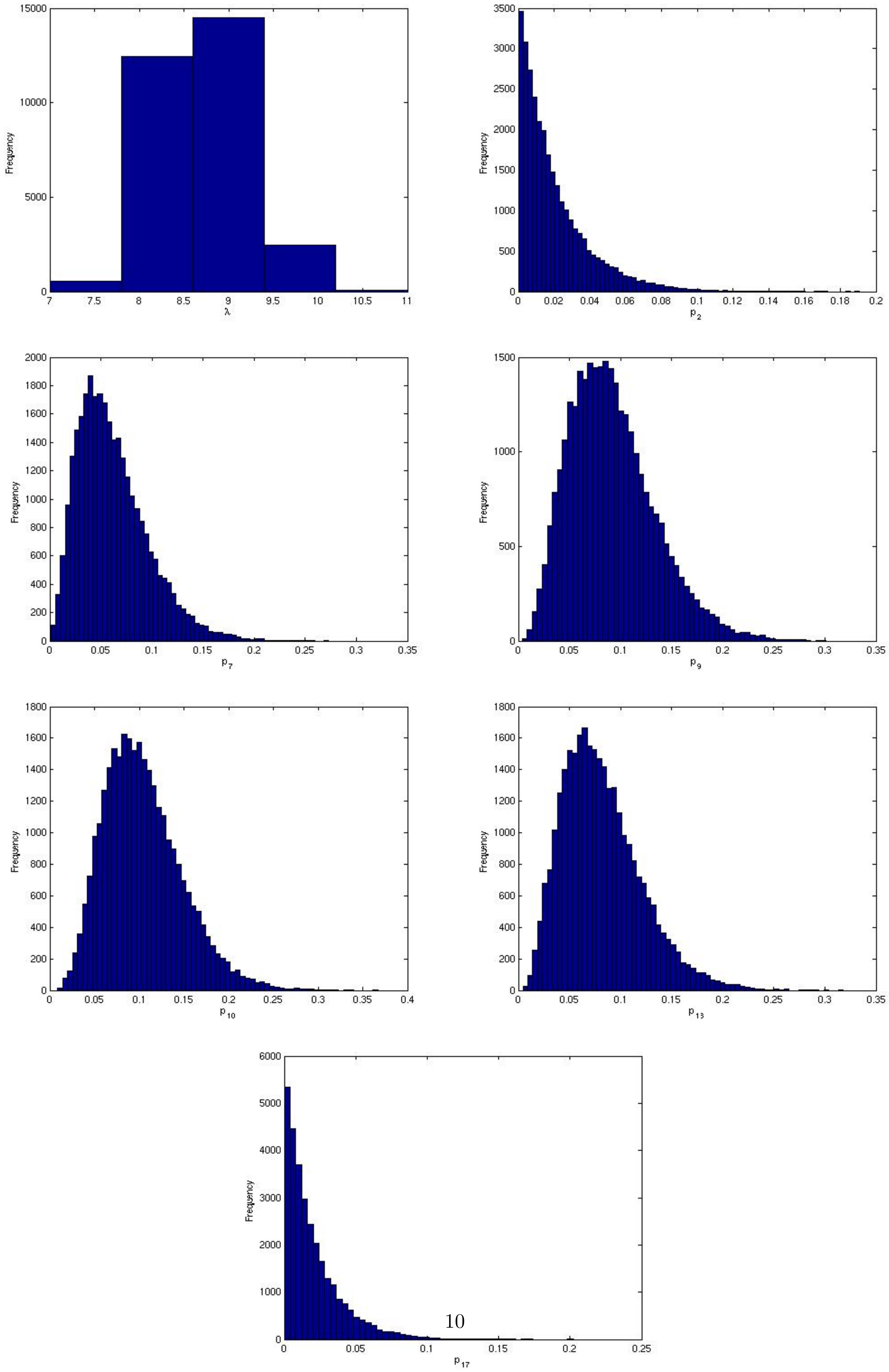

Figure 6: Marginal Histograms of 30,000 Posterior Samples via *BUGS/MATLAB*