STATISTICS 225 – Fall 2018

Homework 4 - solutions

1. **Posterior predictive model checking.**

   (a) The marginal distribution of y is given by $p(y) = \int p(y|\theta)p(\theta)d\theta$. It is a bit easier to get by noting that y can be written as the sum of θ (a $N(0,9)$ variable according to our prior distribution) and a $N(0,1)$ random variable. Thus the marginal distribution of y is $N(0,10)$. Based on this prior probability we find $\Pr(Y \geq 10) = .0008$. This suggests that the observed data ($y = 10$) causes us to reject our model (the combined prior/sampling distribution specification).

   (b) The posterior distribution of θ, derived from results for the conjugate normal family, is a normal distribution with mean equal to $(1 \cdot 10 + (1/9) \cdot 0)/(1 + (1/9)) = 9$ and variance $1/(1 + (1/9)) = 0.9$. Then the posterior predictive distribution of y^{rep} is $N(9, 1.9)$. Then $\Pr(y^{rep} \geq 10) = .23$. Thus the posterior predictive distribution does not find the observed value of y to be very unusual. Though the prior is "misleading" in that it suggests $y = 10$ is highly unlikely; the prior-to-posterior analysis (in this simple case) seems to suggest the observed data are consistent with the model (as judged by the realistic values of θ under the model).

2. **Model checking.**

   (a) If $y_i, i = 1, \ldots, 10$ are independent identically distributed Poisson(θ) random variables (conditional on θ) and the prior distribution is $p(\theta) \propto 1/\theta$, then

   $$p(\theta|y) \propto \frac{1}{\theta}\prod_{i=1}^{10}(\theta^{y_i}e^{-\theta}) = \theta^{\sum_i y_i - 1}e^{-10\theta}.$$

   We recognize this as the kernel of a gamma distribution and thus $\theta|y \sim Gamma(\sum_i y_i, 10) = Gamma(1180, 10)$.

   (b) Naive model

      i. R code is listed below. (I have also posted Stan code for fitting this model though it is trivial to do in R as seen below.) The naive model does not generate data that is consistent with the observed data. $T(y,\theta)$ is always greater than $T(y^{rep},\theta)$. Indeed $T(y,\theta)$ is always bigger than 400 while the largest observed value of $T(y^{rep},\theta)$ is about 30. The posterior probability of replicate data being further from θ than the observed data (as measured by T) is zero.

      ii. The idea here is that for large θ a Poisson(θ) random variable has an approximate Gaussian distribution. If true, then $(y-\theta)/\sqrt{\theta}$ is approximately $N(0,1)$, $(y-\theta)^2/\theta$ is approximately χ^2_1 and $T(y^{rep},\theta)$ is approximately χ^2_{10}. The mean of 1000 draws from $T(y^{rep},\theta)$ was 9.8 in my simulations and the variance was 19.2. These would be 10 and 20 under the chi-square distribution. The differences here are consistent with simulation error (i.e., if I used more than 1000 draws the values would likely be closer to 10 and 20).

```
y <- c(74, 99, 58, 70, 122, 77, 104, 129, 328, 119)
n <- length(y)
nsims <- 1000
thetasim <- rgamma(nsims,sum(y),n)
Tobs <- rep(0,nsims)
Trep <- rep(0,nsims)
for (i in (1:nsims)) {
  Tobs[i] <- sum((y-thetasim[i])^2/thetasim[i])
  yrep <- rpois(n,thetasim[i])
  Trep[i] <- sum((yrep - thetasim[i])^2/thetasim[i])
}
plot(Tobs,Trep)
sum(Trep > Tobs)/length(Trep)
describe(Trep)
```

   (c) Poisson-Gamma hierarchical model - Stan program is available on the website. The R code to run Stan is listed here along with the posterior summaries.

```
# Stan analysis of Poisson-Gamma hierarchical model
traffic.data <- list(M=n, y=y)
stan_fit <- stan(file="H://Stat225//poissongamma.stan",data=traffic.data, iter=1000, chains=4)
print(stan_fit)
```

```
# cut and paste of results
              mean se_mean    sd   2.5%    25%    50%    75%  97.5% n_eff Rhat
alpha         3.43    0.04  1.67   1.06   2.25   3.13   4.27   7.49  1404    1
beta          0.03    0.00  0.02   0.01   0.02   0.02   0.04   0.07  1462    1
theta[1]     75.13    0.19  8.61  59.30  69.19  74.36  80.89  92.94  2000    1
theta[2]     99.73    0.22  9.97  81.25  92.52  99.36 106.44 120.00  2000    1
theta[3]     59.99    0.17  7.75  45.51  54.41  59.77  64.81  76.32  2000    1
theta[4]     71.75    0.19  8.42  56.28  65.82  71.34  77.50  88.28  2000    1
theta[5]    122.24    0.25 11.01 101.96 114.69 122.02 129.21 144.71  2000    1
theta[6]     78.21    0.20  9.08  61.42  71.83  77.77  83.98  97.22  2000    1
theta[7]    104.50    0.24 10.54  85.93  96.97 103.89 111.06 127.54  1876    1
theta[8]    128.95    0.26 11.49 107.77 121.15 128.55 136.42 152.28  2000    1
theta[9]    322.42    0.39 17.59 289.18 310.93 321.76 333.67 358.07  2000    1
theta[10]   119.26    0.24 10.75  98.33 112.03 119.18 126.34 141.52  2000    1
```

(d) The posted Stan code includes generation of Tobs and Trep. R code to carry out this part of the problem is included below along with summaries of Tobs and Trep. In this case the empirical probability (based on 2000 draws) that the replicate data would be more extreme than the observed data (as measured by T) is .48. Thus the hierarchical model appears to generate replicate data sets that resemble the observed data. Also note that the distribution of $Trep$ for this model is the same (approximately chi-squared 10 d.f.) as it was in part (b). This is to be expected since in each case we are generating Poisson data.

```
d <- as.data.frame(stan_fit)
sum(d$trep > d$tobs)
describe(d$tobs)
describe(d$trep)
              mean se_mean    sd   2.5%    25%    50%    75%  97.5% n_eff Rhat
tobs         10.20    0.16  4.55   3.15   6.98   9.37  12.70  20.53   852    1
trep          9.99    0.10  4.41   3.20   6.73   9.46  12.51  20.56  1829    1
```