Introduction to Bayesian Data Analysis
Tutorial 3 Solutions

(1) (a)

$$p(y|\theta) = \frac{e^{-\theta}\theta^y}{y!}$$

$$\log p(y|\theta) = -\theta + y\log\theta - \log y!$$

$$\frac{\partial \log p(y|\theta)}{\partial\theta} = -1 + \frac{y}{\theta}$$

$$\frac{\partial^2 \log p(y|\theta)}{\partial\theta^2} = -\frac{y}{\theta^2}$$

$$\mathrm{I}(\theta) = -E\left[\frac{\partial^2 \log p(y|\theta)}{\partial\theta^2}\right] = \frac{1}{\theta}$$

$$p_J(\theta) \propto \theta^{-1/2}$$

Following the distributional form of the family of Gamma distributions, Jeffreys' prior implies a Gamma(1/2,0) distribution which is not a proper distribution.

(b)

$$f(\theta, y) = \theta^{1/2-1}\frac{e^{-\theta}\theta^y}{y!} \propto \theta^{1/2-1}e^{-\theta}\theta^y = \text{Gamma}(y+1/2, 1)$$

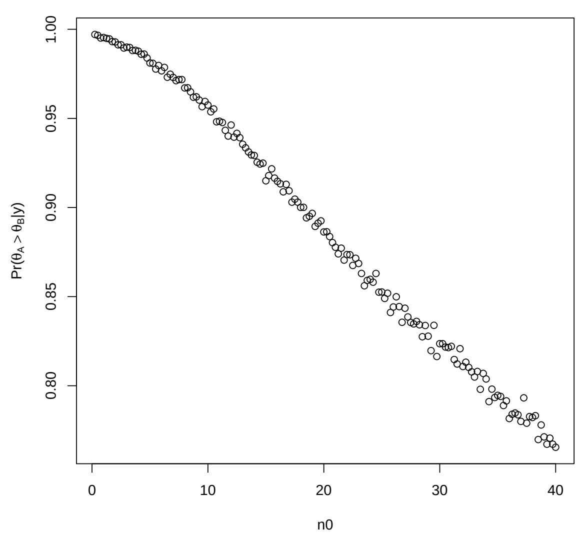
which is a proper posterior density.

(2) $p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta = E[p(\tilde{y}|\theta)|y]$. That is, the integral is finding the weighted average value of $p(\tilde{y}|\theta)$ over all possible values of θ where the weights are given by the posterior density $p(\theta|y)$.

(3) (a) $Pr(\theta_B < \theta_A|\mathbf{y_A}, \mathbf{y_B}) = 0.99$

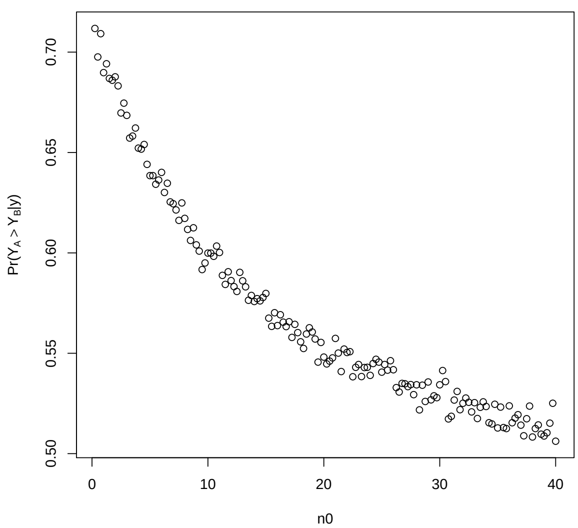
```
> y_A<-c(12,9,12,14,13,13,15,8,15,6)
> syA<-sum(y_A)
> nA<-length(y_A)
> y_B<-c(11,11,10,9,9,8,7,10,6,8,8,9,7)
> syB<-sum(y_B)
> nB<-length(y_B)
>
> a1<-120
> b1<-10
> a2<-12
> b2<-1
>
> theta1.mc<-rgamma(10000,a1+syA,b1+nA)
> theta2.mc<-rgamma(10000,a2+syB,b2+nB)
> mean(theta1.mc>theta2.mc)
[1] 0.9939
```

(b) The $Pr(\theta_B < \theta_A|\mathbf{y_A}, \mathbf{y_B})$ decreases as n_0 increases, but still remains well above 0.5. The results are not sensitive to n_0. Why??



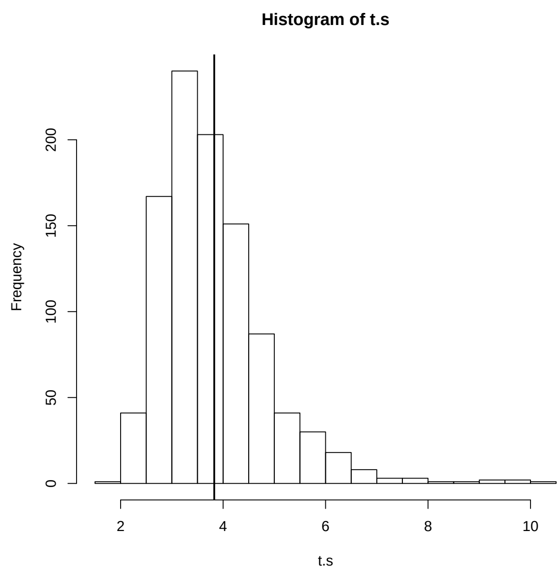
(c) The estimated posterior probability probability based on the posterior predictive probability $Pr(\tilde{Y}_B < \tilde{Y}_A|\mathbf{y_A}, \mathbf{y_B})$ is less than the posterior probability $Pr(\theta_B < \theta_A|\mathbf{y_A}, \mathbf{y_B})$. After allowing for sampling variability in predicting the counts for a new patient, the probability that the counts for a new patient B is less than the counts for a new patient A is only 0.6945. Also note that for large n_0, $Pr(\tilde{Y}_B < \tilde{Y}_A|\mathbf{y_A}, \mathbf{y_B}) \to 0.5$. That is, $Pr(\tilde{Y}_B < \tilde{Y}_A|\mathbf{y_A}, \mathbf{y_B})$ is sensitive to the value of n_0.

```
> count<-0
> for(i in 1:10000){
+ yA.mc<-rpois(1,theta1.mc[i])
+ yB.mc<-rpois(1,theta2.mc[i])
+ count<-count+(yA.mc>yB.mc)*1}
> count/10000
[1] 0.6945
```



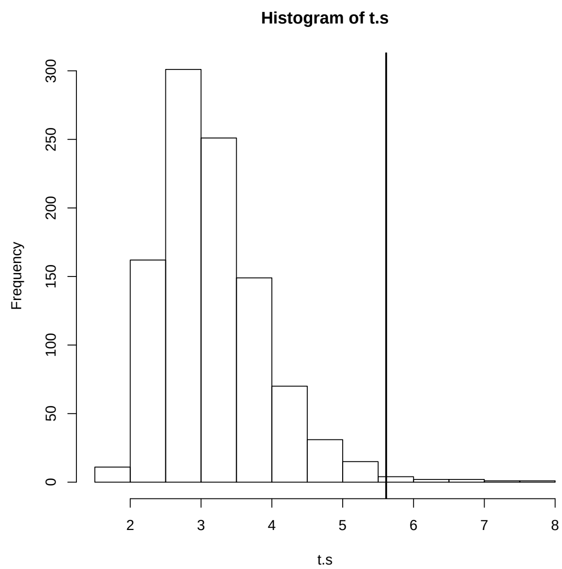
(4) (a) The posterior predictive p-value is 0.418. Based on this statistic, the Poisson model is a good fit to capture the ratio of the mean to variance for population A (but it may not be a good model to capture other aspects of the true probability distribution).

```
yA.mc<-NULL
for (i in 1:1000){
theta1.mc<-rgamma(1,a1+syA,b1+nA)
yA.mc<-cbind(yA.mc,rpois(nA,theta1.mc))
}
t.s<-apply(yA.mc,2,mean)/apply(yA.mc,2,sd)
mean(t.s>mean(y_A)/sd(y_A))
[1] 0.418
```

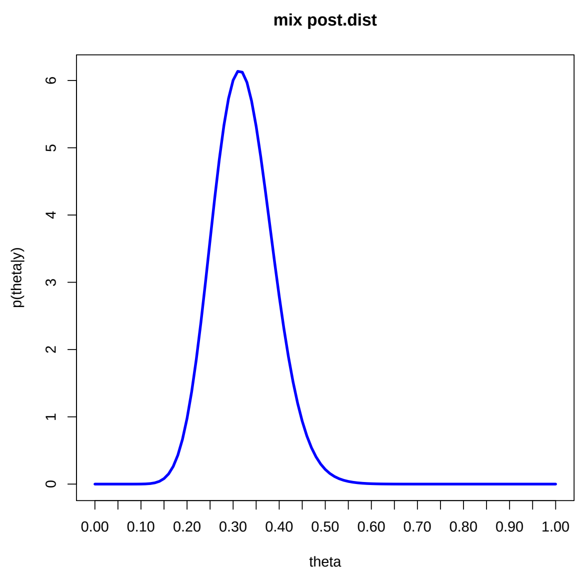


(b) The posterior predictive p-value is 0.008. Based on this statistic, the Poisson model is not a good fit to capture the ratio of the mean to variance for population B.

```
for (i in 1:1000){
theta2.mc<-rgamma(1,a2+syB,b2+nB)
yB.mc<-cbind(yB.mc,rpois(nB,theta2.mc))
}
mean(t.s>mean(y_B)/sd(y_B))
[1] 0.008
```



(5) (a) Using a discrete approximation, a 95% quantile-based posterior interval is (0.20, 0.46)




```
> post.theta<-function(theta) p1*dbeta(theta,a1+y,n-y+b1)+
      p2*dbeta(theta,a2+y,n-y+b2)
> theta_0.025<-seq(0,0.25,0.01)
> n_0.025<-length(theta_0.025)
> theta_0.975<-seq(1,0.4,-0.01)
> n_0.975<-length(theta_0.975)
> d1<-post.theta(theta_0.025)*0.01
#*0.01 because increment size of theta_seq is 0.01
```

```
> d2<-post.theta(theta_0.975)*0.01
>
> cdf_0.025<-d1[1]
> k<-0
> for (i in 2:n_0.025){
+ if(k==0){
+ cdf_0.025<-c(cdf_0.025,cdf_0.025[i-1]+d1[i])
+ k<-(cdf_0.025[i]>=0.025)*1
+ } else {
+ k<-length(cdf_0.025)
+ }
+ }
> theta_0.025[k]
[1] 0.2
> cdf_0.975<-d2[1]
> k<-0
> for (i in 2:n_0.975){
+ if(k==0){
+ cdf_0.975<-c(cdf_0.975,cdf_0.975[i-1]+d2[i])
+ k<-(cdf_0.975[i]>=0.025)*1
+ } else {
+ k<-length(cdf_0.975)
+ }
+ }
> theta_0.975[k]
[1] 0.46
```

(b) Using Monte-Carlo simulation, a 95% quantile based posterior interval for θ is (0.20, 0.46). The discrete approximation gives similar values.

```
> theta.mc<-NULL
> for (i in 1:1000){
+ x<-rbinom(1,1,p1)
+ if (x==1){
+  theta <-rbeta(1,a1+y,n-y+b1)
+ } else {
+  theta <-rbeta(1,a2+y,n-y+b2)
+ }
+ theta.mc<-c(theta.mc,theta)
+ }
> quantile(theta.mc,c(0.025,0.975))
     2.5%     97.5%
0.2049685 0.4604861
```