Ömer Yiğit – r0767950
Rana Cansu Kebabcı – r0772826
Hesna Aksoy – r0772816

CONCEPTS OF BAYESIAN DATA ANALYSIS: PROJECT

1) Lyme disease

100 ticks are collected from field and grasslands. It is found that 12 of them carry Borrelia bacteria. In a previous experiment, 10 ticks were collected and it was found that 2 of them carry Borrelia bacteria.

1.) The random variable is the occurrence of Borrelia bacteria. Likelihood for the historical data follows a binomial distribution.

$$\mathrm{L}(\theta|\mathrm{y}_0) = \binom{\mathrm{n}_0}{\mathrm{y}_0}\theta^{\mathrm{y}_0}(1-\theta)^{(\mathrm{n}_0-\mathrm{y}_0)}$$
$$y_0 = 2, n_0 = 10$$

Kernel of the binomial likelihood (part of the function which contains information about the parameter θ) shows similar properties as the beta density function. Matching the parameters gives us the prior distribution, which is a beta(3,9) distribution.

$$\mathrm{p}(\theta) = \frac{1}{\mathrm{B}(\alpha_0,\beta_0)}\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1}$$
$$\mathrm{y}_0 = \alpha_0 - 1\,, \mathrm{n}_0 - \mathrm{y}_0 = \beta_0 - 1$$
$$\alpha_0 = 3\,, \beta_0 = 9$$

Likelihood of the experiment also follows a binomial distribution. Similarly, posterior distribution can also be found easily with kernel of the beta distribution. The posterior is beta(15,97) distribution. The calculations can easily be done, thanks to conjugacy property.

$$\mathrm{L}(\theta|\mathrm{y}) = \binom{\mathrm{n}}{\mathrm{y}}\theta^{\mathrm{y}}(1-\theta)^{(\mathrm{n}-\mathrm{y})}$$
$$y = 12, n = 100$$
$$\mathrm{p}(\theta|\mathrm{y}) \propto \mathrm{L}(\theta|\mathrm{y})\mathrm{p}(\theta)$$
$$p(\theta|y) = \frac{1}{B(\bar{\alpha},\bar{\beta})}\theta^{\bar{\alpha}-1}(1-\theta)^{\bar{\beta}-1}$$
$$\bar{\alpha} = \alpha_0 + \mathrm{y}\,, \bar{\beta} = \beta_0 + \mathrm{n} - \mathrm{y}$$
$$\bar{\alpha}=15\,, \bar{\beta}=97$$

Prior, scaled likelihood and posterior is plotted and the plot is given below in *Figure 1*.

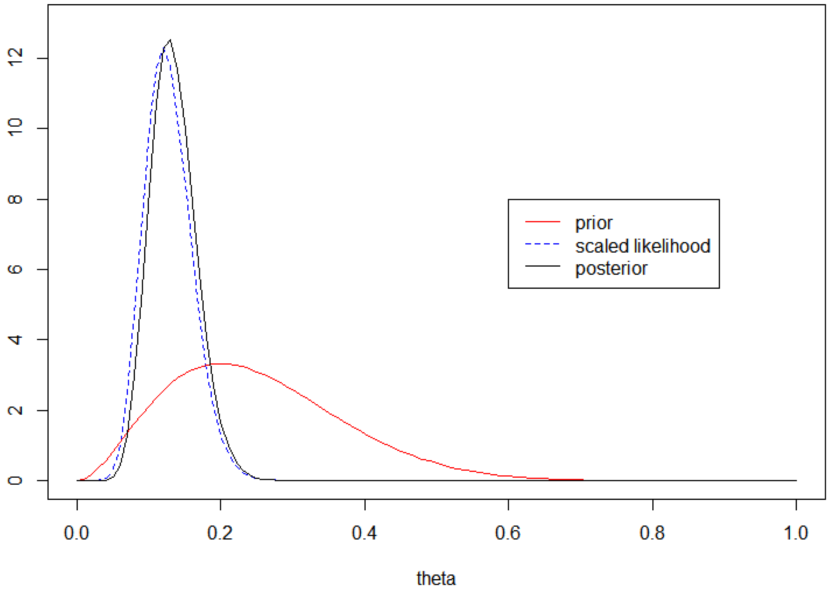

Figure 1: Distributions of the parameter θ

Posterior summary measures are calculated according to formulas and they are given below.

- Posterior mode: $\hat{\theta}_M$=0.127
- Posterior mean: $\bar{\theta} = 0.134$
- Posterior median: $\bar{\theta}_M = 0.132$
- Posterior variance: $\bar{\sigma}^2$=0.032^2
- Credible interval – HPD: (0.074 , 0.198) with α=0.05
- Credible interval – equal tail: (0.078 , 0.203) with α = 0.05

2.) Predictions can be done with posterior predictive distributions (PPD). The uncertainty in the parameter θ is also taken into account. In a future experiment with 50 ticks ($m = 50$), for the binomial case, PPD is a beta-binomial distribution with parameters $m, \bar{\alpha}$ *and* $\bar{\beta}$ – BB(50,15,97). Analytically, the PPD can be found by the following way.

$$p(\tilde{y}|y) = \int p(\tilde{y},\theta|y)d\theta = \int p(\tilde{y}\,|\theta,y)p(\theta|y)d\theta$$

$$p(\tilde{y}|y) = \int_0^1 \binom{m}{\tilde{y}} \theta^{\tilde{y}}(1-\theta)^{(m-\tilde{y})} \frac{\theta^{\alpha-1}(1-\theta)^{(\beta-1)}}{B(\alpha,\beta)} d\theta$$

$$p(\tilde{y}|y) = \binom{m}{\tilde{y}} \frac{B(\tilde{y}+\alpha, m-\tilde{y}+\beta)}{B(\alpha,\beta)}$$

The corresponding density plot of the PPD is given below in *Figure 2*. According to the expected value of the distribution above, the occurrence of 6.696 Borrelia is expected. The probability of finding 4 out of 50 ticks carrying the disease is 0.107 – the length of the red bar corresponding to $\tilde{y} = 4$.

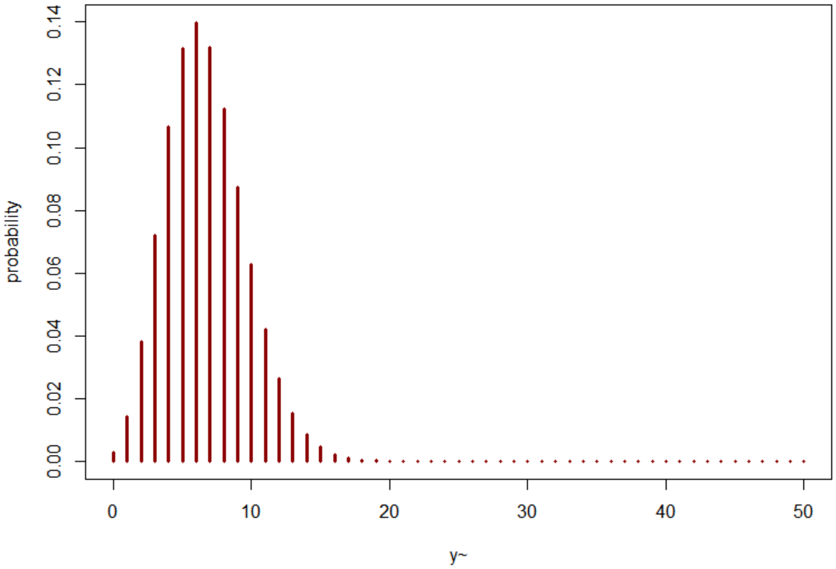

Figure 2: Posterior predictive density (PPD) plot

2. Sleep-deprived reaction

The data describes the effect of sleep deprivation on cognitive performance. 18 subjects were restricted to 3 hours of sleep. Their reaction time to a visual stimulus (in ms) was measured. A simple linear regression equation of the data is given as $\mathrm{R_{it}}=\alpha+\beta \mathrm{t}+\epsilon_{\mathrm{it}}$ with R_{it} the reaction time of subject i at day t and $\epsilon_{it} \sim N(0, \sigma^2)$. The data is used in mean-centered format to avoid of non-convergence caused by autocorrelation.

1.) As priors for the parameters of the Bayesian model, non-informative priors are selected. 2 MCMC chains are taken. Completely random initial values may cause finding sub-optimal results; therefore, a systematic approach was preferred. The initial values selected are based on maximum likelihood parameter estimates and standard errors. α and β parameters are assumed to be normally distributed and initial values are randomly sampled from their corresponding distributions. The details of the model can be found in Appendix. At the first step, the model is updated with 1000 runs and diagnostic measures of convergence are examined. The diagnostic plots are given below in *Figure 3.1* and *3.2*.

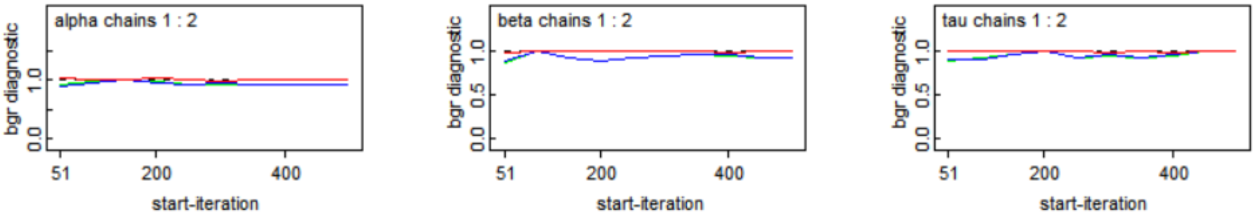

Figure 3.1: bgr diagnostic measures of the parameters of interest

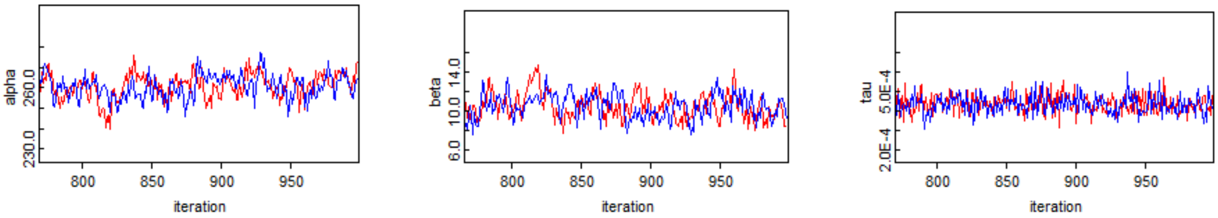

Figure 3.2: Trace plots for the first 1000 iterations

In *Figure 3.1*, The lines on the value of 1 indicate that the convergence seems to be satisfied quite early on the sampling. However, as a rule of thumb, MC error of the parameters should be lower than 5% of the standard deviation. Also, burn-in period should be discarded and hence, the sampling chain should be extended. In order to do so, 2000 more runs are iterated on the existing chains. Based on above figures, it is decided to discard 500 iterations as burn-in part. It is seen that the convergence properties are still satisfied. Trace plots, which are given below in *Figure 4* are also consistent with the achieved result.

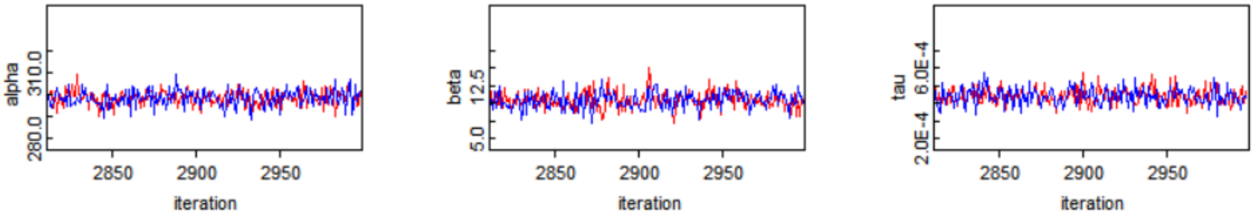

Figure 4: Trace plots of the parameters of interest

2.) Probability density functions and summary statistics of the parameters are given below in, *Figure 5* and *Table 1*, respectively.

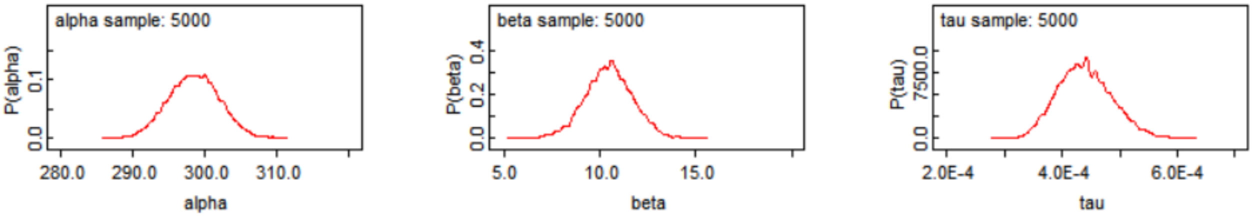

Figure 5: Posterior distributions of the parameters

parameter	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
alpha	298.5	3.554	0.05443	291.6	298.5	305.4	501	5000
beta	10.46	1.24	0.01775	7.971	10.48	12.87	501	5000
sigma	47.89	2.572	0.03525	43.14	47.8	53.15	501	5000
tau	4.397E-4	4.711E-5	6.473E-7	3.541E-4	4.378E-4	5.377E-4	501	5000

Table 1: Summary statistics of the parameters

Since non-informative priors are used, all contribution to posterior distributions of the parameters came from the likelihood functions, which are normally distributed. Therefore, posteriors are also normally distributed.

3.) The posterior probability that reaction time is increased by 10ms per day depends on the posterior distribution of β. According to the distribution parameters given above, it is assumed that β is normally distributed with mean of 10.46 and standard deviation of 1.24.
`dnorm(10, mean=10.46, sd=1.24)` code in R provided the result as 0.30 (30%).

4.) Posterior predictive distributions at days 0-9 are obtained from the same sampling model. Visualization as OpenBUGS output is provided in *Figure 6* and the statistics are summarized in *Table 2*. ytilde[t+1] is the posterior predictive distribution of day t.

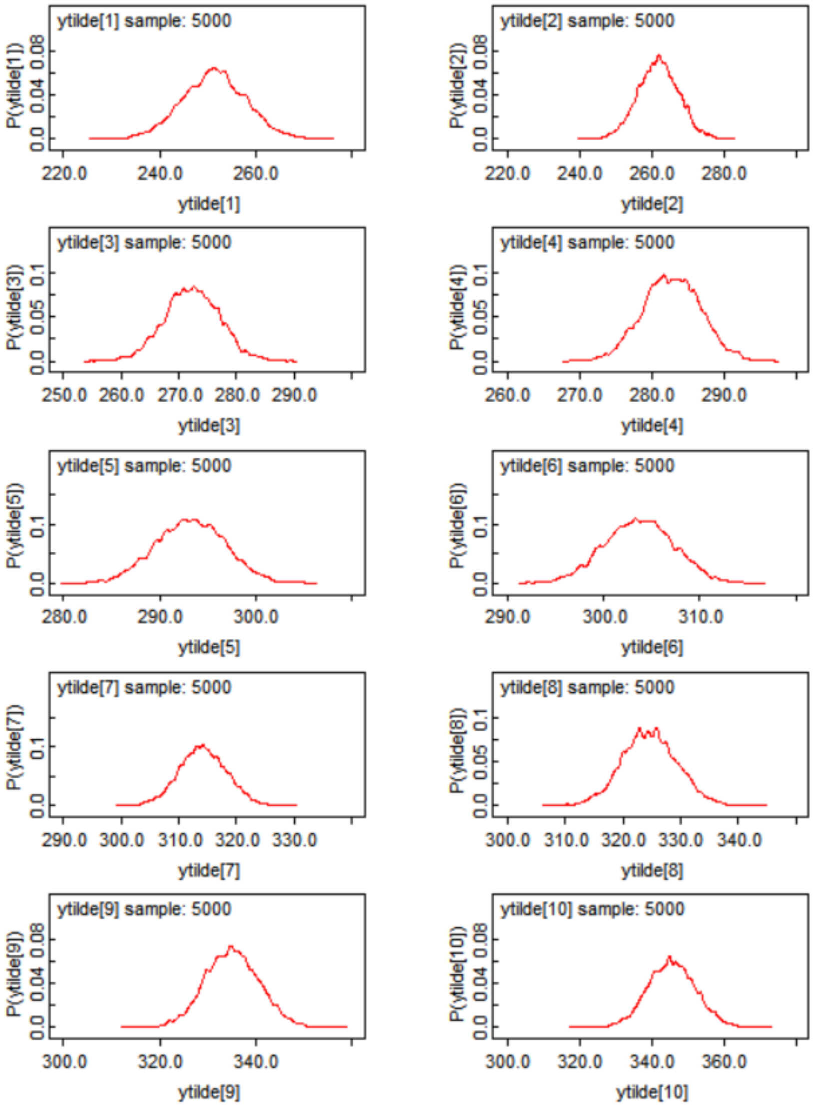

Figure 6: Posterior predictive distributions (PPD) of days 0-9

parameter	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
ytilde[1]	251.4	6.624	0.1013	238.5	251.4	264.6	501	5000
ytilde[2]	261.9	5.617	0.08682	250.9	261.9	272.9	501	5000
ytilde[3]	272.3	4.723	0.07378	263.0	272.3	281.7	501	5000
ytilde[4]	282.8	4.016	0.06308	275.0	282.8	290.8	501	5000
ytilde[5]	293.3	3.61	0.05607	286.2	293.3	300.3	501	5000
ytilde[6]	303.7	3.606	0.05422	296.7	303.7	310.6	501	5000
ytilde[7]	314.2	4.006	0.05801	306.3	314.1	322.0	501	5000
ytilde[8]	324.7	4.708	0.06649	315.4	324.6	333.9	501	5000
ytilde[9]	335.1	5.599	0.07815	324.2	335.1	346.2	501	5000
ytilde[10]	345.6	6.604	0.09178	332.6	345.5	358.5	501	5000

Table 2: Summary statistics according to PPDs of days 0-9

Means of ytilde[t+1] parameters are the equivalent of $\tilde{\mu}_t$. Also, the mean of the sigma estimate, which is 47.89, is the equivalent of $\tilde{\sigma}$. Hence, according to the formula given below, ppd of each day can be found.

$$\tilde{\sigma}_t^2 = \tilde{\sigma}^2 \cdot (1 + [1,t](X^T X)^{-1}[1,t]^T)$$

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \end{bmatrix}$$

PPD of each day and similar statistics for the actual data are provided in *Table 3*. Assuming normal distribution, a comparison is visualized in *Figure 7*. Dashed lines represent the means of the corresponding distributions.

	Actual data		**PPD**	
Days	**mean**	**sd**	**mean**	**sd**
0	256.6518	32.12945	251.4	55.54940
1	264.4958	33.43033	261.9	53.51019
2	265.3619	29.47342	272.3	51.92826
3	282.9920	38.85774	282.8	50.84630
4	288.6494	42.53789	293.3	50.29659
5	308.5185	51.76962	303.7	50.29659
6	312.1783	63.17372	314.2	50.84630
7	318.7506	50.10396	324.7	51.92826
8	336.6295	60.19972	335.1	53.51019
9	350.8512	66.98616	345.6	55.54940

Table 3: Summary statistics for the actual data

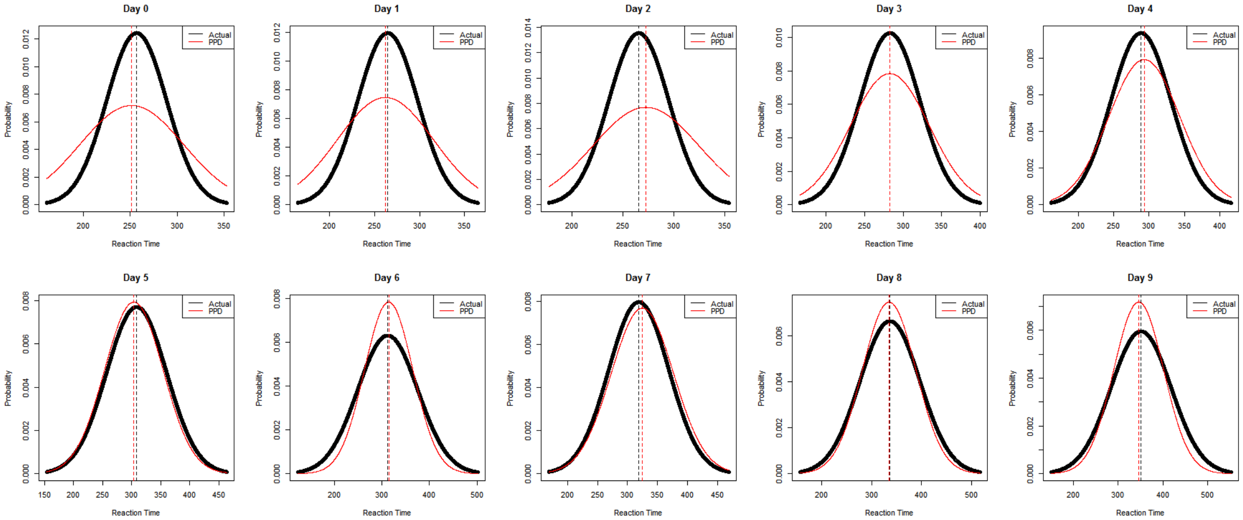

Figure 7: Comparison of PPD and actual data distribution

Means of PPDs are very close to the means of the distributions of the actual data per day. The first notable observation is that the variances of PPDs are closer the variances of the distributions of the actual data around the middle days. The reason is that the mean-centered data is used for prediction in the regression analysis. Another observation is that the variance of the actual data tends to increase by time. The reason is that there may be different unobservable (or uninteresting) effects that have an impact on the reaction time.

5.) Now, two models are compared. In the first model, as explained above, likelihood is assumed to have normal distribution. In the second model, likelihood is assumed to have t-distribution. The main metric to compare two models is deviance information criterion (DIC). First model has the DIC value of 1906.0, whereas the second model has the DIC value of 1916.0. It means that the model with normally distributed likelihood is better. Changing the likelihood from a normal distribution to t-distribution showed no improvement, although the DIC values are really close to each other. As the last remark, it should be noted that the results may slightly change due to random sampling. However, these changes are not significant enough to reach different conclusions.

APPENDIX

1. R codes for lyme disease study

```
## 1: ANALYTICAL DERIVATIONS

## Data specification
s <- 12
n <- 100

## Prior specification
alpha <- 3
beta <- 9

## Calculations for posterior distribution
theta <- seq(0, 1, 0.01)
#Scaled likelihood, easy to compare with prior and posterior
likelihood <- dbeta(theta, s+1, n-s+1)
prior <- dbeta(theta, alpha, beta)
posterior <- dbeta(theta, alpha+s, beta+n-s)

## Plotting likelihood, prior and posterior
plot(theta,likelihood,type="l",ylim=c(0,13),col="blue",ylab="", lty=2)
lines(theta,prior,col="red")
lines(theta,posterior)
legend(0.6,8, c("prior","scaled likelihood","posterior"),
       lty=c(1,2,1), col=c("red","blue","black"))
#The prior is not so informative (small sample size).
#since the likelihood sample size is way larger than prior,
#it dominates the posterior distribution.

## Posterior summary measures
alpha.post<-alpha+s
beta.post<-beta+n-s
posterior_mode <- (alpha.post-1)/(alpha.post+beta.post-2)
posterior_mean <- alpha.post/(alpha.post+beta.post)
posterior_median <- qbeta(0.5,alpha.post,beta.post)
posterior_variance <- alpha.post*beta.post/((alpha.post+beta.post)^2*(alpha.post+beta.post+1))

## Credibility intervals - HPD
hpdbeta <- function(alpha,beta)
{
        p2 <- alpha
        q2 <- beta
        f <- function(x,p=p2,q=q2){
                b<-qbeta(pbeta(x,p,q)+0.95,p,q);(dbeta(x,p,q)-dbeta(b,p,q))^2}
        hpdmin <- optimize(f,lower=0,upper=qbeta(0.05,p2,q2),p=p2,q=q2)$minimum
        hpdmax <- qbeta(pbeta(hpdmin,p2,q2)+0.95,p2,q2)
        return(c(hpdmin,hpdmax))
}
hpd.CI<-hpdbeta(alpha.post,beta.post)

## Credibility intervals - equal tail
plot(theta,posterior,type="l",ylim=c(0,13),col="black",ylab="")
Eq.CI<-c(qbeta(0.025,alpha.post,beta.post),qbeta(0.975,alpha.post,beta.post))
abline(v=Eq.CI[1])
abline(v=Eq.CI[2])

## 2. PREDICTONS

## Predictive posterior distribution (page 155)
par(mfrow=c(1,1))
m <- 50
ytilde <- 0:50
#log of combination (m choose ytilde)
term1 <- log(choose(m, ytilde))
#nominator of betabinomial dist
term2 <- lgamma(alpha.post+ytilde)+lgamma(m-ytilde+beta.post)-lgamma(alpha.post+beta.post+m)
#constant part B(alfa,beta) in beta dist = inverse gamma dist. page 78
#calculations here are in log scale
#denominator of betabinomial dist
term3 <- lgamma(alpha.post)+lgamma(beta.post)-lgamma(alpha.post+beta.post)
#everything together
lpart <- term1+term2-term3
#back to original scale
dytilde <- exp(lpart)

## Plotting PPD
plot(ytilde, dytilde, type="n", xlab="y~", ylab="probability")
lines(ytilde, dytilde, type="h", col="dark red", lwd=3)
## Expectation
expected <- m*alpha.post/(alpha.post+beta.post)
```

```
## Prediction for ytilde=4
pred_func<-function(alpha, beta, ytilde, m)
{
  term1 <-  log(choose(m, ytilde))
  term2 <- lgamma(alpha+ytilde)+lgamma(m-ytilde+beta)-lgamma(alpha+beta+m)
  term3 <- lgamma(alpha)+lgamma(beta)-lgamma(alpha+beta)
  lpart <- term1+term2-term3
  dytilde <- exp(lpart)
  return(dytilde)
}
pred_func(alpha.post, beta.post, 4, 50)
```

2. OpenBUGS codes for sleep deprivation study

```
### LIKELIHOOD WITH NORMAL DISTRIBUTION
model
{
    for (j in 1:10)
    {
           for(i in 1:18)
           {
                   R[j, i] ~ dnorm(mu[j], tau)
        }
           mu[j] <- alpha+ beta * (t[j] - tbar)
           ytilde[j] <- alpha + beta * (tpred[j] - tbar)
    }
    alpha ~ dnorm(0.0, 1.0E-6)
    beta ~ dnorm(0.0, 1.0E-6)
    tau ~ dgamma(0.001, 0.001)
    sigma <- 1 / sqrt(tau)
}

### LIKELIHOOD WITH T DISTRIBUTION
model
{
    for (j in 1:10)
    {
           for(i in 1:18)
           {
                   R[j, i] ~ dt(mu[j], tau, k)
           }
           mu[j] <- alpha + beta *(t[j] - tbar)
           ytilde[j] <- alpha + beta * (tpred[j] - tbar)
    }
    alpha ~ dnorm(0.0, 1.0E-6)
    beta ~ dnorm(0.0, 1.0E-6)
    k ~ dgamma(2, 0.01)
    tau <- (k - 2) / k
}

### MAIN DATA
list(t=c(0,1,2,3,4,5,6,7,8,9), tpred=c(0,1,2,3,4,5,6,7,8,9), tbar=4.5,
                   R=structure(

.Data=c(249.56,222.7339,199.0539,321.5426,287.6079,234.8606,283.8424,265.4731,241.6083,312.36
66,236.1032,256.2968,250.5265,221.6771,271.9235,225.264,269.8804,269.4117,

258.7047,205.2658,194.3322,300.4002,285,242.8118,289.555,276.2012,273.9472,313.8058,230.3167,
243.4543,300.0576,298.1939,268.4369,234.5235,272.4428,273.474,

250.8006,202.9778,234.32,283.8565,301.8206,272.9613,276.7693,243.3647,254.4907,291.6112,238.9
256,256.2046,269.8939,326.8785,257.2424,238.9008,277.8989,297.5968,

321.4398,204.707,232.8416,285.133,320.1153,309.7688,299.8097,254.6723,270.8021,346.1222,254.9
22,255.5271,280.5891,346.8555,277.6566,240.473,281.7895,310.6316,

356.8519,207.7161,229.3074,285.7973,316.2773,317.4629,297.171,279.0244,251.4519,365.7324,250.
7103,268.9165,271.8274,348.7402,314.8222,267.5373,279.1705,287.1726,

414.6901,215.9618,220.4579,297.5855,293.3187,309.9976,338.1665,284.1912,254.6362,391.8385,269
.7744,329.7247,304.6336,352.8287,317.2135,344.1937,284.512,329.6076,

382.2038,213.6303,235.4208,280.2396,290.075,454.1619,332.0265,305.5248,245.4523,404.2601,281.
5648,379.4445,287.7466,354.4266,298.1353,281.1481,259.2658,334.4818,

290.1486,217.7272,255.7511,318.2613,334.8177,346.8311,348.8399,331.5229,235.311,416.6923,308.
102,362.9184,266.5955,360.4326,348.1229,347.5855,304.6306,343.2199,

430.5853,224.2957,261.0125,305.3495,293.7469,330.3003,333.36,335.7469,235.7541,455.8643,336.2
806,394.4872,321.5418,375.6406,340.28,365.163,350.7807,369.1417,

466.3535,237.3142,247.5153,354.0487,371.5811,253.8644,362.0428,377.299,237.2466,458.9167,351.
6451,389.0527,347.5655,388.5417,366.5131,372.2288,369.4692,364.1236),
                           .Dim=c(10,18)))

### INITIALS FOR NORMAL LIKELIHOOD MODEL
list(alpha=258.7, beta=10.80, tau= 2)
list(alpha=254.05, beta= 2.19, tau= 3)

### INITIALS FOR T LIKELIHOOD MODEL
list(alpha=258.7, beta=10.80, k= 2)
list(alpha=254.05, beta=12.19, k=3)
```