Homework 3 - Stat 488 Bayesian Analysis

Nicholas Marey

September 18, 2017

Question 1

Here you will write an R function to perform a Bayesian analysis of count data. The response is the number of events $Y \in 0, 1, 2, ...$ that occurs in a time interval of $M > 0$ hours, and the unknown parameter is the mean count per hour, θ. Assume the Bayesian model with likelihood $Y|\theta \sim Poisson(M\theta)$, prior $\theta \sim Gamma(a, b)$, and thus posterior $\theta|Y \sim Gamma(Y + a, M + b)$.

Part a

Write an R function that takes Y, M, a, and b as inputs. The function should produce a plot (clearly labeled!) that overlays the prior and posterior density functions (both using the dgamma function), and it should return a list with the posterior mean, posterior standard deviation, and a 95% credible interval.

```
pg.posterior <- function(Y, M, a, b) {


    grid <- seq(0.01, 3, 0.01)

    like <- dpois(Y, M * grid)
    like <- like/sum(like)  #standardize

    prior <- dgamma(grid, a, b)
    prior <- prior/sum(prior)  #standardize

    post <- like * prior
    post <- post/sum(post)  #standardize

    plot(grid, prior, type = "l", xlab = "lambda", ylab = "Density", col = "blue")
    lines(grid, post, col = "red", lty = 2, lwd = 2)
    legend("topright", c("Prior", "Posterior"), lwd = c(1, 1), col = c("blue",
        "red"), lty = c(1, 2))

    post.ci <- qgamma(c(0.025, 0.975), Y + a, M + b)  #Posterior 95% Credible Interval
    post.mean <- (Y + a)/(M + b)  #Posterior mean
    post.sd <- sqrt((Y + a)/(M + b))  #Posterior sd

    posterior <- data.frame(post.mean, post.sd, post.ci[1], post.ci[2])
    colnames(posterior) = c("mean", "Standard Deviation", "Lower CI", "Upper CI")
    return(posterior)
}
```

Part b

What values of a and b would make good default values to represent a prior that carries little information about θ? Make these the default values in your function.

0.01 for a and b will not affect the posterior distribution greatly so they were chosen as defaults values that will carry little information about θ.

```
pg.posterior <- function(Y, M, a = 0.01, b = 0.01) {


    grid <- seq(0.01, 3, 0.01)

    like <- dpois(Y, M * grid)
    like <- like/sum(like)  #standardize

    prior <- dgamma(grid, a, b)
    prior <- prior/sum(prior)  #standardize

    post <- like * prior
    post <- post/sum(post)  #standardize

    plot(grid, prior, type = "l", xlab = "lambda", ylab = "Density", col = "blue")
    lines(grid, post, col = "red", lty = 2, lwd = 2)
    legend("topright", c("Prior", "Posterior"), lwd = c(1, 1), col = c("blue",
        "red"), lty = c(1, 2))

    post.ci <- qgamma(c(0.025, 0.975), Y + a, M + b)  #Posterior 95% Credible Interval
    post.mean <- (Y + a)/(M + b)  #Posterior mean
    post.sd <- sqrt((Y + a)/(M + b))  #Posterior sd

    posterior <- data.frame(post.mean, post.sd, post.ci[1], post.ci[2])
    colnames(posterior) = c("mean", "Standard Deviation", "Lower CI", "Upper CI")
    return(posterior)
}
```

Part c

What values of a and b give prior mean 1 and prior standard deviation 2?

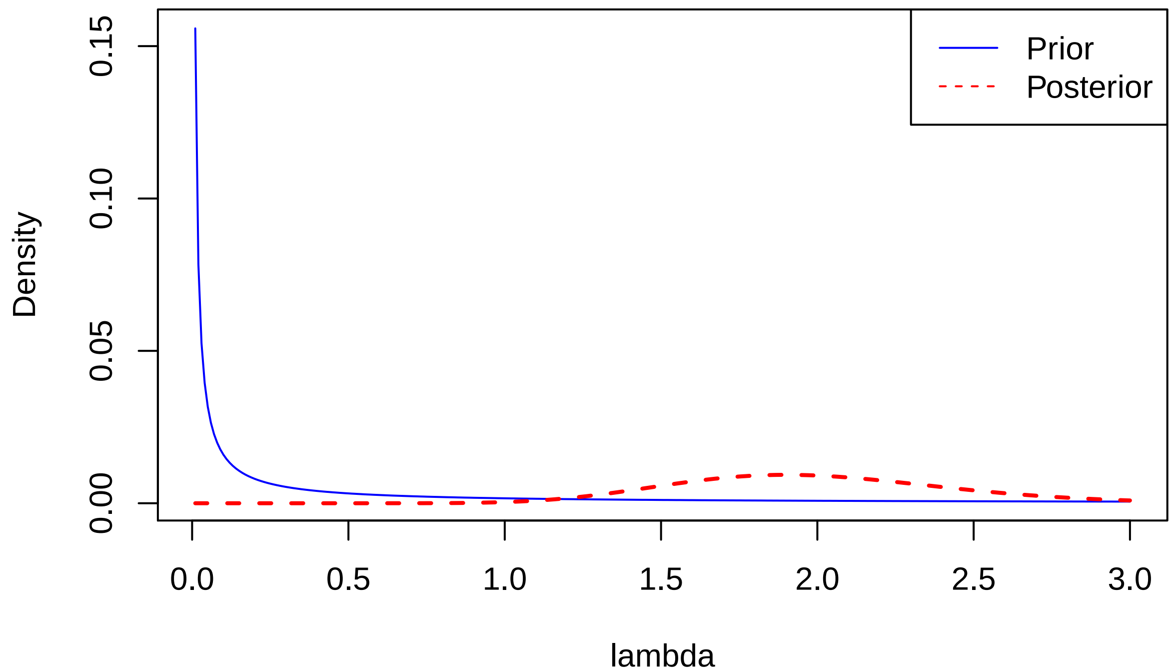
$$\frac{\alpha}{\beta} = mean = 1$$

$$\alpha = \beta$$
$$\frac{\alpha}{\beta^2} = \sigma^2 = 4$$
$$\frac{\alpha}{\alpha^2} = 4$$
$$\frac{1}{\alpha} = 4$$
$$\alpha = \beta = \frac{1}{4}$$

Part d

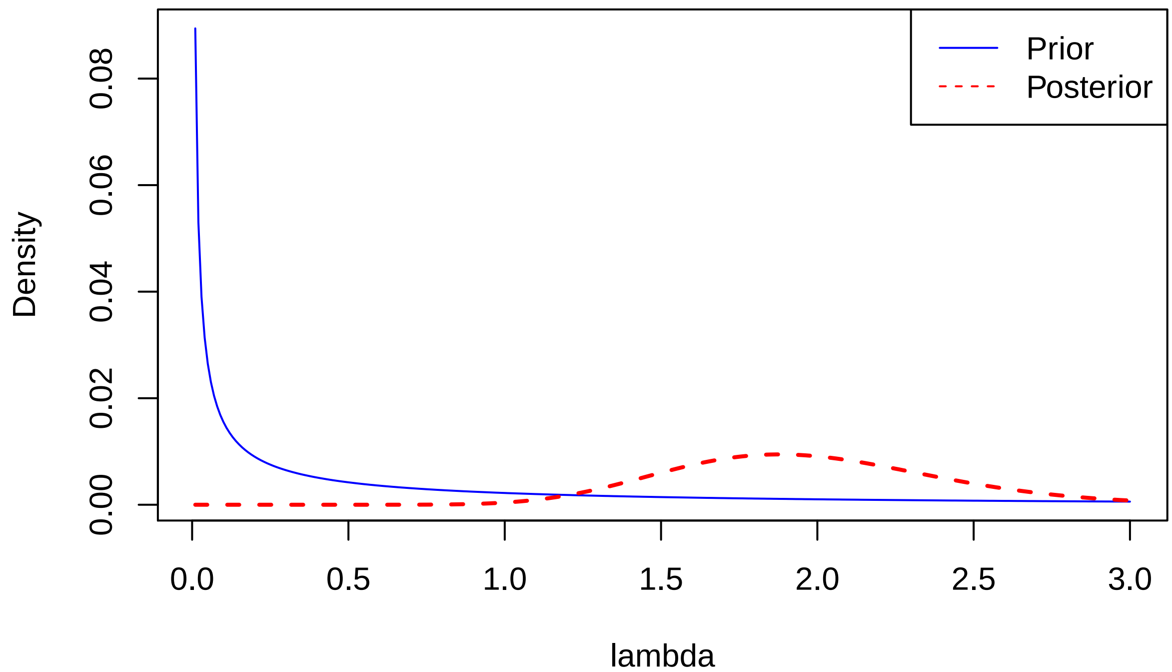
Now we observe $Y = 20$ events in $M = 10$ hours. Use your code from (a) to conduct a Bayesian analysis of these data. Perform the analysis twice, once with the uninformative prior from (b) and once with the informative prior in (c).

```r
pg.posterior(20, 10)
```




```
##       mean Standard Deviation Lower CI Upper CI
## 1 1.999001            1.41386 1.221211  2.96534
```

```r
pg.posterior(20, 10, 0.25, 0.25)
```




```
##      mean Standard Deviation Lower CI Upper CI
## 1 1.97561           1.405564   1.2109 2.924462
```

Part e

Summarize the results. In particular, how does this analysis compare to a frequentist analysis and how much are the results affected by the prior?

Model	Mean	Standard Deviation	Lower 95% CI	Upper 95% CI
Uninformative	1.999	1.414	1.221	2.965
Informed	1.976	1.406	1.211	2.924
Frequentist	2	1.414	0	5

The results obtain from the bayesian method are affected by the prior since they do not match what a frequentist would say. Assuming a poisson distribution and based on the observation with 2 events happening per hour, lambda would be equal to 2. This in turn would cause the mean to be equal to 2 and the standard deviation equal to 1.414. With the introduction of the prior, the mean and the standard deviation are both shifted towards the prior. With the uniformed prior the mean barely shifts toward the prior, by 0.001, and the standard deviation stays the same at 1.414. With the informed prior, there is a greater shift toward the prior, of 0.024, and the standard deviation shifts toward the prior as well by 0.008.

Overall, the results are not affected very greatly by the prior in this scenario.

Question 2

A steel plant relies on a machine that produces devices guaranteed to be defecting with probability less than 0.10. You are in change of quality control. To ensure that the machine is working properly you take 100 samples each day (independent across sample and day) and record the number of samples that are defective. The data for a recent 10-day stretch are given below:

Days	1	2	3	4	5	6	7	8	9	10
Samples	12	11	9	15	18	15	11	8	12	16

Part a

Assuming the true defect probability is constant, plot the posterior probability that the defect probability is greater than 0.1 as a function of the day. For day t, use the samples from day 1 through day t in this calculation.

Calculating α and β for prior probability based on belief that mean of θ is 0.10 with variance of 0.05

$$E[\theta] = \frac{\alpha}{\alpha + \beta} = 0.1$$

$$10\alpha = \alpha + \beta$$

$$9\alpha = \beta$$

$$V[\theta] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} = 0.05$$

$$\frac{\alpha * 9\alpha}{(\alpha + 9\alpha)^2(\alpha + 9\alpha + 1)} = 0.05$$

$$\frac{9}{100 * 0.05} = 10\alpha + 1$$

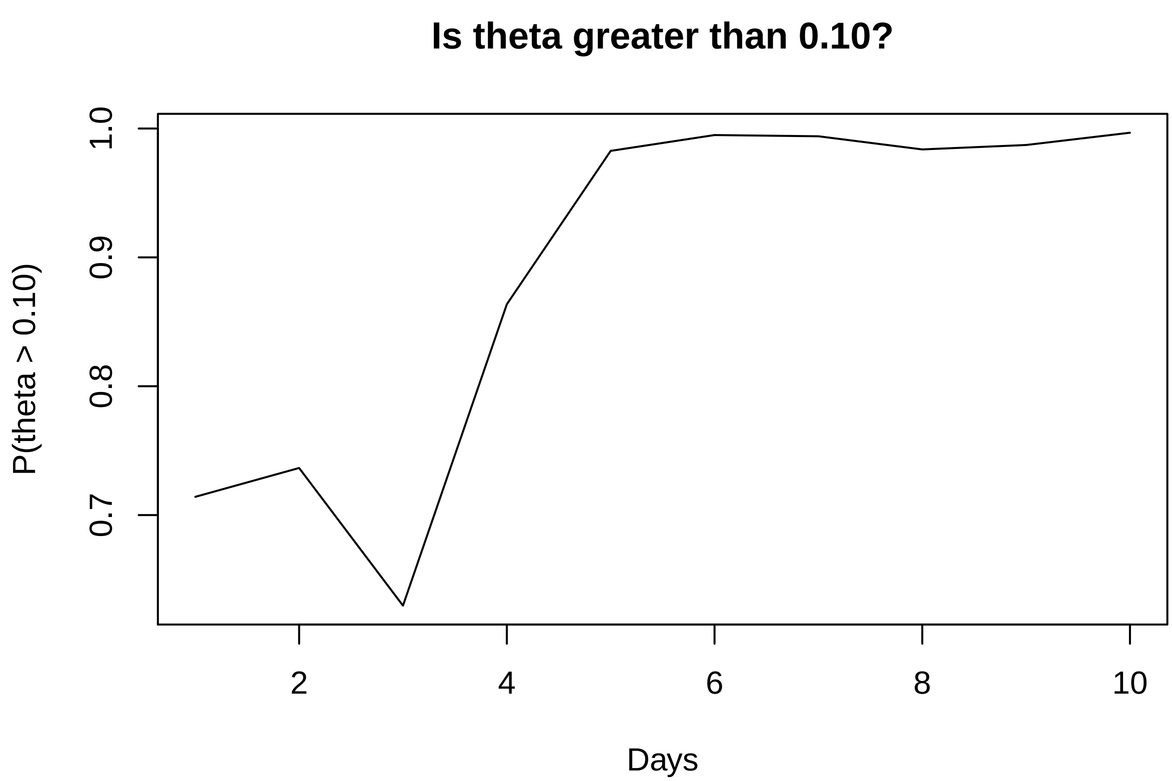
$$\frac{9}{5} - 1 = 10\alpha$$

$$\alpha = 0.08$$

$$9\alpha = \beta = 0.08 * 9 = 0.72$$

```r
# Inputing data
samples <- c(12, 11, 9, 15, 18, 15, 11, 8, 12, 16)
days <- 1:10

# priors
a <- 0.08
b <- 0.72

# calculating number of samples through day t calculating number of defects
# through day t
cumulative.samples <- seq(100, 1000, 100)
cumulative.defects <- rep(NA, 10)

for (i in 1:10) {
    if (i == 1) {
        cumulative.defects[i] = samples[i]
```

```r
    } else {
        cumulative.defects[i] <- cumulative.defects[i - 1] + samples[i]
    }

}

# Combining into a data frame
dat <- data.frame(days, cumulative.defects, cumulative.samples)

# Calculating probability theta > 0.10
prob <- rep(NA, 10)

for (i in 1:10) {
    prob[i] <- 1 - pbeta(0.1, cumulative.defects[i] + a, cumulative.samples[i] -
        cumulative.defects[i] + b)
}

# Combining into data frame
dat2 <- data.frame(prob, days)

# Plotting results
plot(dat2$days, dat2$prob, type = "l", xlab = "Days", ylab = "P(theta > 0.10)",
    main = "Is theta greater than 0.10?")
```



Part b

When, if ever, would you sound the alarm and claim that the machine is out of order?

I would probably sound the alarm on the first day, with a probability of 0.714 that $\theta > 0.1$ on a piece of equipment where this is never suppose to occur, I would find it odd. By day 4, I would certainly sound the alarm. The probability that theta is greater than 0.10 through day 4 is 86.36%. This is large enough to conclude that the machine is out of order. Furthermore, the probability that theta is greater than 0.10 never goes down after day 4. Therefore by day 5 my suspicions have been confirmed when I notice the probability that theta is greater than 0.10 is 98.26% and I would have to call in the mechanic to take a look at the machine.