Stat 574B: HW 4 Problem 1

Nate Hattersley

November 15, 2017

part a.

Assume Y and θ are distributed as given. Then the integrand $p(Y|\theta)p(\theta)$ looks like:

$$p(Y|\theta)p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{n+\alpha-1}e^{-\theta(\beta+\sum\limits_i Y_i)}$$

Integrating out theta, we recognize the terms dependent on theta as the kernel of a Gamma($n+\alpha$, $\beta+\sum\limits_i Y_i$). This means that

$$\begin{aligned} p(Y) &= \frac{\beta^\alpha}{\Gamma(\alpha)}\int\limits_0^\infty \theta^{n+\alpha-1}e^{-\theta(\beta+\sum\limits_i Y_i)}d\theta \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(n+\alpha)}{(\beta+\sum\limits_i Y_i)^{n+\alpha}} \end{aligned}$$

I'm not sure how to interpret the results of this, since it seems that α, β, n are fixed. But, it's simple enough to see that you can factor the PDF according to Neyman-Fisher to show that $\sum\limits_i Y_i$ is a sufficient statistic. The average is a one-to-one transformation of the sum, hence $\bar{Y}$ is also sufficient.

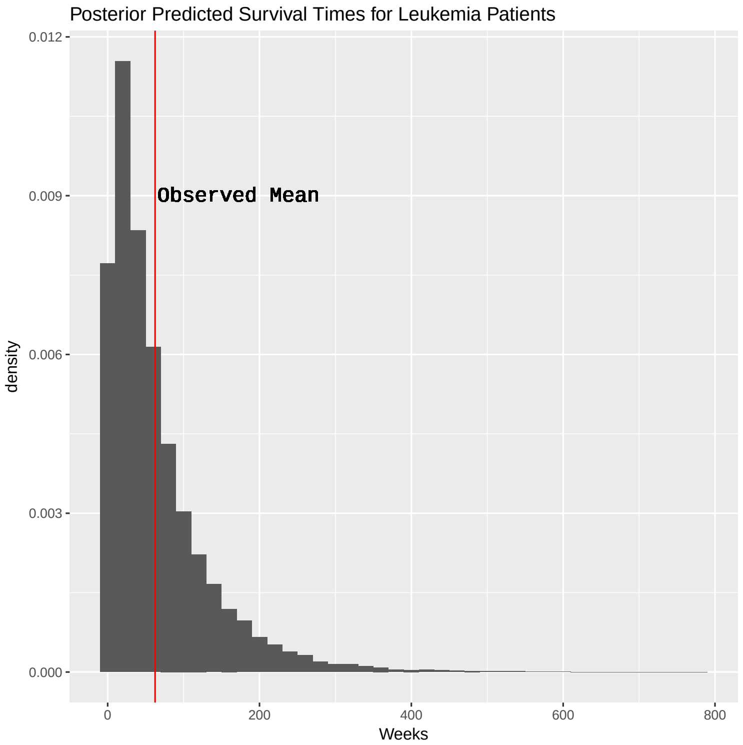
part b.

This predictive PDF is obviously a decreasing function of $\sum\limits_i Y_i$ and by extension $\bar{Y}$. Hence if $x > y$, $p(x) < p(y)$; and if $p(x) < p(y)$, $x > y$. ■

part c.

The sampling scheme is relatively straightforward. First, I will sample from the posterior distribution of $\theta|Y_{obs}$, and use those samples as parameters to sample from the exponential distribution. These are the posterior predictive samples. I took 1.7×10^4 samples. I create a table with summary statistics (mean, standard deviation, 95% credible interval) of the PPD sample as well as a histogram.

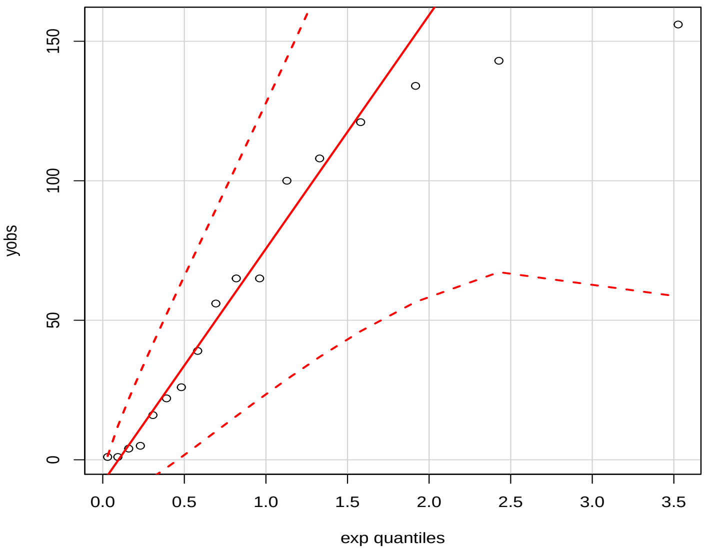
Mean	SD	Lower	Upper
64.64	69.69	1.48	253.79

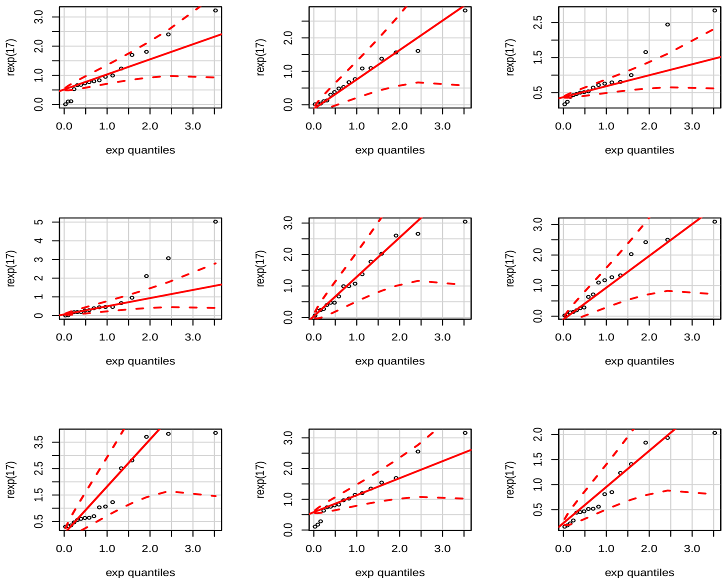


Now, to model the p-value, note that we can break the output vector into 1000 samples of 17 numbers. Hence, I will iterate over the PPD sample in chunks of 17 and compare the chunk's sample mean to our data mean. The proportion of chunks with sample averages greater than observed is the p-value, which I found to be 0.51. Given such a large p value, we conclude that we fail to reject the null assumption of the data being exponentially distributed.

part d.

For this part, I will just randomly pick 9 chunks of 17 like I mentioned above, and make QQ plots for each one of them. First, I include the QQ plot of the observed data for reference:



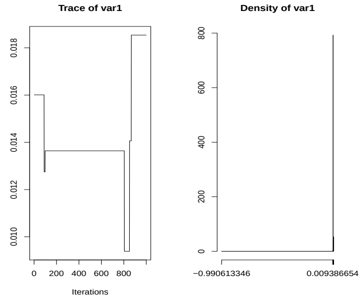


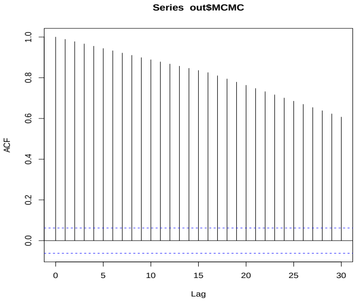
Qualitatively, it looks like there is room for huge outliers from generating a small sample of the exponential distribution. My conclusion is that our finding from part c is justified, and that more anamolous results than observed are possible from a random sample of 17 exponential random variables.

part e.

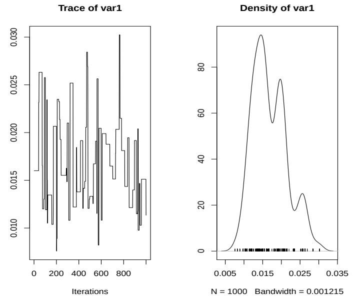
My starting value for θ was $1/\bar{Y}$, where $\bar{Y}$ is the data mean. This was my naive best guess for $E(\theta)$ and should lie in a region with good probability mass. I chose a selection of δ values of different orders of magnitude (10^{-i} for $i \in 1 : 4$) in order to visualize the effects on the acceptance ratio and the trace plot. If δ is too large, then the acceptance ratio becomes exceedingly small and you won't get a good approximation of the posterior. I figured that the best δ value would be somewhere close to the standard deviation of the distribution, perhaps some small multiple of it. I know the posterior and so I know that its standard deviation is .0038, hence I posit that .01 and .001 will provide the best mixing.

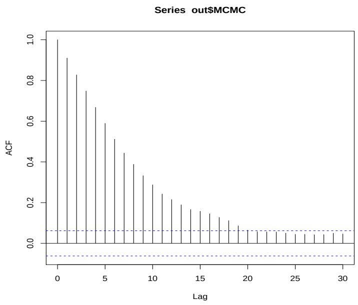
Delta: 1.000
Acceptance Ratio: 0.50%
Effective Size: 9.96



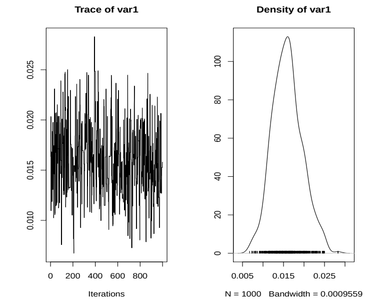


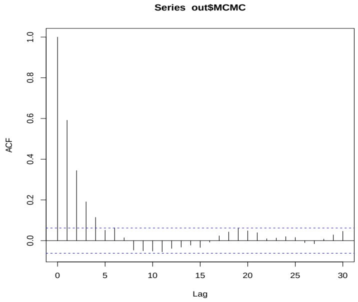
Delta: 0.100
Acceptance Ratio: 7.20%
Effective Size: 47.00



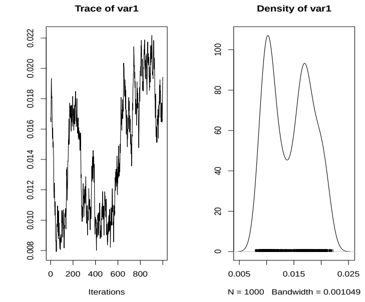


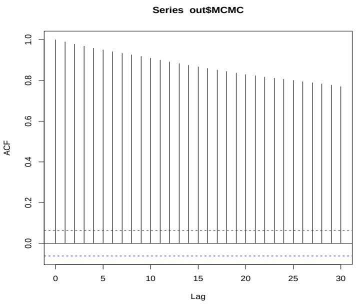
Delta: 0.010
Acceptance Ratio: 49.60%
Effective Size: 256.48





Delta: 0.001
Acceptance Ratio: 93.70%
Effective Size: 5.20





Interestingly, $\delta = .001$ provided poor mixing because the delta value was so small that almost every point was accepted. The best mixing came when $\delta = .01$. To illustrate what that means for patient outcomes, I took the mean, median, and a 95% credible interval on the MH posterior sample. The mean of the resultant exponential distribution given θ is $1/\theta$. Hence, I inverted the summaries of θ to give an idea of the possible spread of population means.

Mean	Median	Lower	Upper
61.86	63.54	42.45	102.57

Code

```
## PART C
set.seed(42)
yobs <- c(65, 156, 100, 134, 16, 108, 121, 4, 39, 143, 56, 26,
    22, 1, 1, 5, 65)
n <- 17
a <- 1
b <- 52

size <- 1000

## Generate PPD Sample
theta_new <- rgamma(17 * size, n + a, b + sum(yobs))
ynew <- rexp(17 * size, theta_new)
## Summaries of PPD
ynew %>% (function(x) {
    data.frame(Mean = mean(x), SD = sd(x), Lower = quantile(x,
        0.025), Upper = quantile(x, 0.975))
}) %>% xtable %>% print(include.rownames = F)

ggplot(data.frame(Weeks = ynew)) + geom_histogram(aes(x = Weeks,
    y = ..density..), binwidth = 20) + ggtitle("Posterior Predicted Survival Times for Leukemia Patients
    geom_vline(aes(xintercept = mean(yobs)), col = "red") + geom_text(aes(x = mean(yobs) +
    110, y = 0.009, label = "Observed Mean"), alpha = 0.75, family = "mono",
    fontface = "plain", size = 5)

## Computing P value

p.val <- 0

chunk_i <- function(i) {
    ynew[(17 * i + 1):(17 * (i + 1))]
}

for (i in 0:(size - 1)) {
    subst <- chunk_i(i)
    p.val %<>% c(mean(subst) > mean(yobs)) %>% weighted.mean(c(i,
        1))
}

## PART D
qqPlot(yobs, d = "exp")
par(mfrow = c(3, 3))
for (i in 1:9) {
    qqPlot(rexp(17), dist = "exp")
}
## PART E
size <- 1000

test.delta <- function(d) {
```

```r
    theta <- 1/mean(yobs)
    THETA <- c()
    ACCEPTRATIO <- 0
    for (i in 1:size) {

        theta_prop <- runif(1, theta - d, theta + d)
        dtheta <- dgamma(theta, n + a, b + sum(yobs))
        dtheta_prop <- dgamma(theta_prop, n + a, b + sum(yobs))

        if (!dtheta || is.na(dtheta))
            accept <- 1 else if (is.na(dtheta_prop))
            accept <- 0 else accept <- dtheta_prop/dtheta


        if (accept > runif(1)) {
            ACCEPTRATIO %<>% c(1) %>% weighted.mean(c(i - 1,
                1))
            theta <- theta_prop
        } else {
            ACCEPTRATIO %<>% c(0) %>% weighted.mean(c(i - 1,
                1))
        }
        THETA[i] <- theta
    }
    list(MCMC = as.mcmc(THETA), RATIO = ACCEPTRATIO)
}

test.deltas <- function(...) {
    for (i in c(...)) {
        out <- test.delta(i)
        cat(sprintf("Delta: %.3f\\\\", i))
        cat(sprintf("Acceptance Ratio: %.2f\\%%\\\\", 100 * out$RATIO))
        cat(sprintf("Effective Size: %.2f", effectiveSize(out$MCMC)))
        plot(out$MCMC)
        acf(out$MCMC)
        cat("\\noindent\\rule{\\textwidth}{1pt}\\\\\\\\")
    }
}

test.deltas(1, 0.1, 0.01, 0.001)

## one-liner to make a variable and print it
(best <- test.delta(0.01) %>% with(data.frame(Mean = 1/mean(MCMC),
    Median = 1/median(MCMC), Lower = 1/quantile(MCMC, 0.975),
    Upper = 1/quantile(MCMC, 0.025)))) %>% xtable %>% print(include.rownames = F)
```