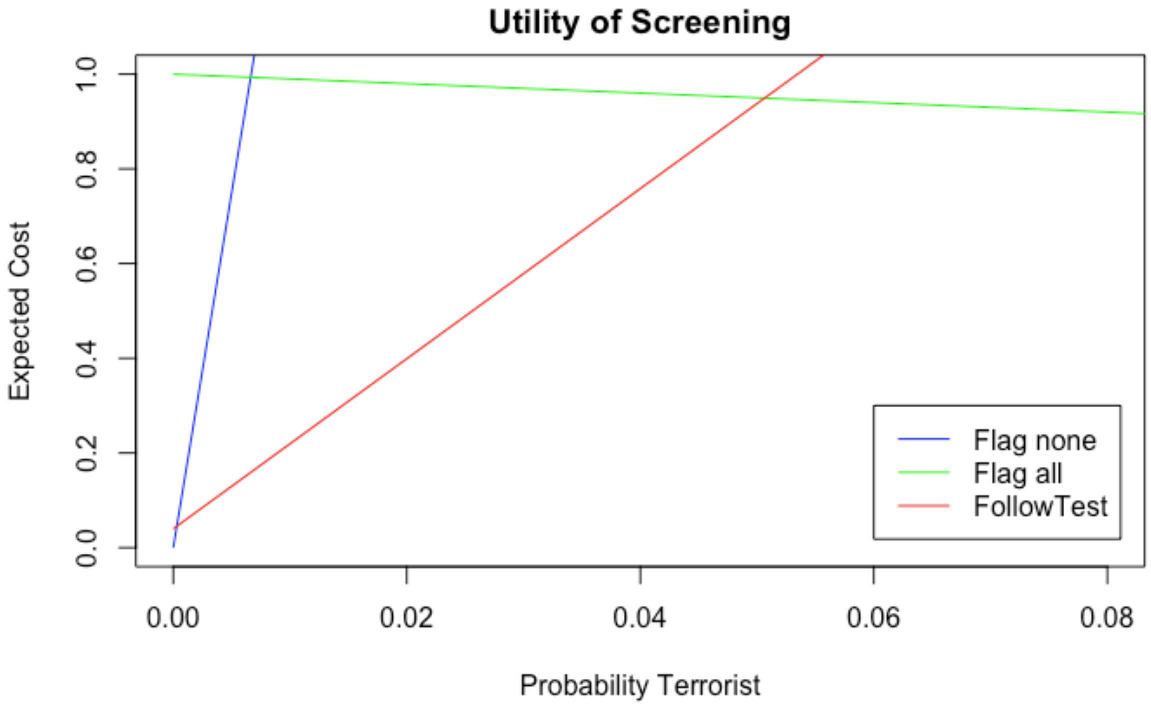
Problem 1

A security screening system photographs people entering a facility, matches them to a database of terrorist suspects, and alerts security to stop an individual for further scrutiny if a match is found. The system has a miss probability of 12% and a false alarm probability of 4%, where a miss is defined as failing to issue an alert for a person in the database, and a false alarm is defined as issuing an alert for a person who is not in the database. Assume that true positives and true negatives cost nothing, and the cost of a miss is 150 times the cost of a false alarm. Let p be the prior probability that a person is in the database. Plot the expected loss of three policies: (1) stop everyone for questioning; (2) stop no one for questioning; and (3) stop someone for questioning if an alert is issued. For what range of p is each policy optimal? Comment on your answer.



'Flag none' is only an optimum strategy when $P=0$.
For $0 < P < \sim 0.05$, following the test the is the best strategy.
For $P > \sim 0.05$, talking to everyone is the best strategy.

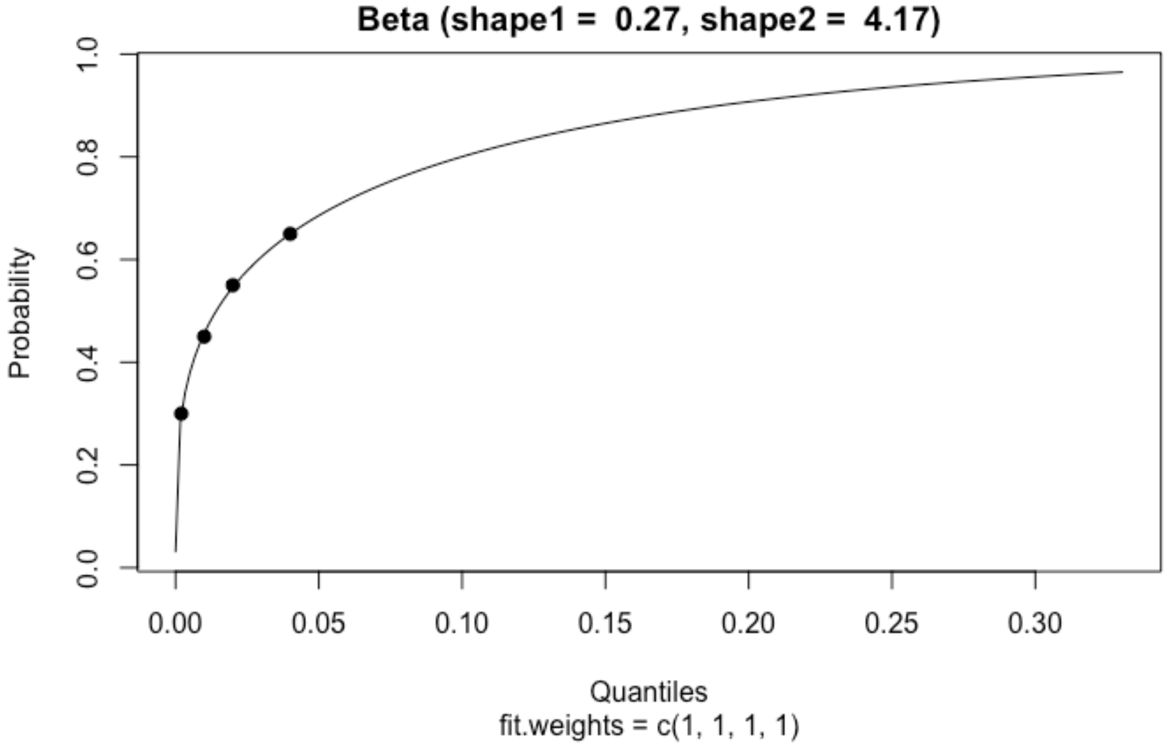
These are notably small probabilities, and this relates directly to the fact that the cost of a miss is 150 times that of incorrectly flagging someone who is not in the DB of potential terrorists. Additionally, there are no costs for correct classifications which further simplifies the utility model here.

Problem 1 Code

```
p_vals <- seq(0,100,1)/100      # represents terrorists
flag_none  <- 150*p_vals        # no cost to citizens
flag_all <- 1*(1-p_vals)        # no costs attributed to terrorists
follow_test <- ((1-p_vals)*1*0.04)+(p_vals*150*0.12)

plot(seq(0,1,0.01),flag_none,
     type="l",
     col="blue",
     main="Utility of Screening",
     xlab="Probability Terrorist",
     ylab="Expected Cost",
     xlim=c(0,.08),
     ylim=c(0,1))
lines(seq(0,1,0.01),flag_all,col='green')
lines(seq(0,1,0.01),follow_test,col="red")
legend(0.06,0.30,c("Flag none","Flag
all","FollowTest"),col=c("blue","green","red"),lty=c(1,1,1))
```

Problem 2

Historically, one in every 50 fourth-graders at an elementary school fails to reach minimum threshold on a state-mandated reading exam. The school is planning a study to evaluate a new reading curriculum. A school administrator assesses a 55% chance that the new curriculum would reduce the failure rate for the exam. That is, there is a 55% chance that if the new curriculum were implemented, fewer than 1 in 50 students would fail the exam. Suppose the administrator also assesses a 45% chance that the failure rate would be cut in half and a 30% chance that the fail rate is improved by a factor of 10. That is, there is a 45% chance that fewer than 1 in 100 students would fail the exam, and a 30% chance that fewer than 2 in 1000 students would fail the exam under the new curriculum. The administrator also assesses a 35% chance that the failure rate would more than double, or at least 4% of the students would fail the exam. If you were to use a Beta distribution to fit these judgments, what parameters would you use? Do you think it provides a good fit? Justify your answer.



The parameters I would use are shape = 0.27 and scale = 4.17. This is a good fit in terms of variance from the data. However, it is necessary to change the 35% chance of the failure rate being > 4% to match the rest of the inputs. It makes more sense to say there is a 65% chance that < 4% will fail. This then gives us a good beta quantile distribution in terms of fitting the data points, and it passes a basic logic test of sensibility when reviewing the plot.

Problem 2 Code

```
require(rriskDistributions)

#Inputting Data
fail_count <- c( 0.002, 0.01, 0.02, 0.04)
chance <- c( 0.30, 0.45, 0.55, 0.65)

#Using a fitting function to determnie the shape and scale
params <- get.beta.par(p=chance,q=fail_count)

#The following is not strictly  necessary; I just wanted
#to be sure I understood the auto-generated chart.
shape <- params[1]
scale <- params[2]
betas <- qbeta(probas,shape1=shape,shape2=scale)

#Plot the expert opinion with the fitted beta distribution
plot(fail_count,chance,xlim=c(0,1),ylab='Probability',xlab='Quantiles'
,ylim=c(0,1))
lines(betas,probas,type='l')
```

Problem 3

A shape recognition system classifies objects as round, rectangular, or irregular. It correctly classifies round objects 80% of the time, rectangular objects 85% of the time, and irregular objects 70% of the time. Incorrectly classified objects are equally likely to be classified as either of the two incorrect object types. Assume in a given environment that 15% of the objects are round, 25% of the objects are rectangular, and 60% of the objects are irregularly shaped. Find the joint distribution of object shapes and classification results. If the system reports that an object is rectangular, what is the posterior probability that the object has each of the three shapes given the system report?

Confusion Matrix:

Confusion Matrix	**Class: Round**	**Class: Rectangle**	**Class: Irregular**
Actual: Round	0.80	0.10	0.10
Actual: Rectangle	0.075	0.85	0.75
Actual: Irregular	0.15	0.15	0.70

For an object reported as a rectangle, the posterior probability that it is a rectangle is 77%, the probability that it is round is 9%, and the probability that it is irregular is 14% (P = 0.7727273, 0.09090909, and 0.1363636, respectively, for exact figures).

Problem 3 Code

```
#Posterior Probabilities  for Class: Rectangle:
prob_round <- 0.1/(0.1+0.85+0.15)
prob_rect <- 0.85/(0.1+0.85+0.15)
prob_irreg <- 0.15/(0.1+0.85+0.15)
```

Jericho McLeod
G00986513

Problem 4

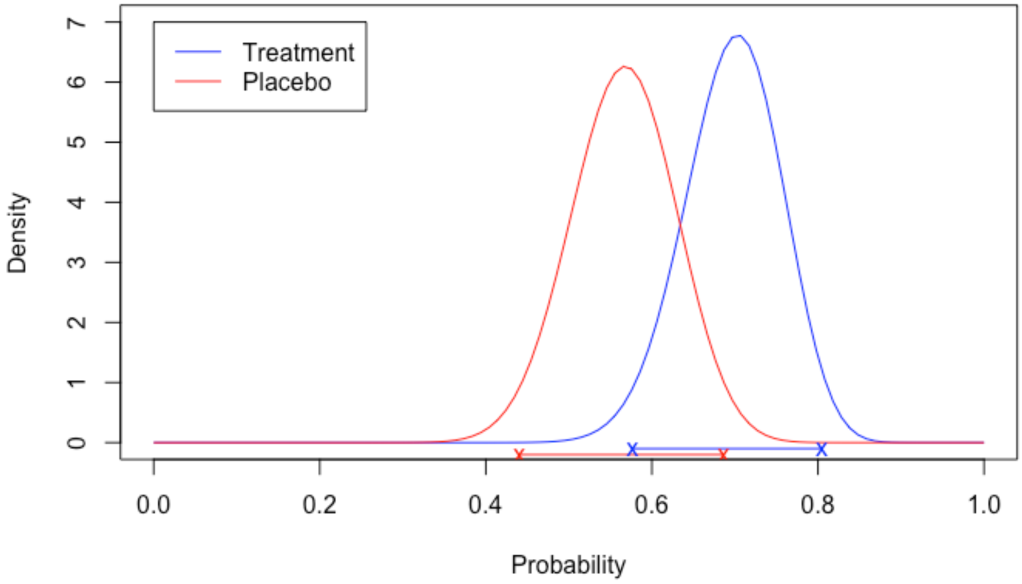
A drug for treating depression is undergoing clinical trials. A total of 120 patients were enrolled in a randomized, double-blind, placebo-controlled study. Half of the patients were randomly selected to receive the drug; the remaining patients were given a placebo. Questionnaires were administered at the start of the study and 30 days after treatment began. One of the questions on the post-study questionnaire asked whether patients experienced improvement in their mood since the start of the study. The table below shows responses to this question.

	No Improvement	Improvement	Total
Treatment	18	42	60
Placebo	26	34	60
Total	44	76	120

Let Y_1 and Y_2 be the number of patients in the treatment and placebo groups who experienced improvement. Assume Y_1 and Y_2 are independent random variables with Binomial(60, Q_i) distributions, for i=1,2. Assume Q_1 and Q_2 are independent with a Beta distribution with shape parameters 1/2 and 1/2 (this is the Jeffreys prior distribution). Find the joint posterior distribution for Q_1 and Q_2. Name the distribution type and its hyperparameters. Plot the posterior density functions for Q_1 and Q_2 on the same axes. Comment on your results.

The joint distribution between Theta_1 and Theta_2 is a joint beta distribution. The shape for treatments is 42.5 and the scale is 18.5. The shape for placebos is 34.5 and the scale is 26.5.

The credible interval for treatment showing improved mood is 0.58 to 0.80, while that of placebo is 0.44 to 0.69. The joint beta posterior distribution and credible intervals are shown in the following chart.



The distributions overlap to some degree, but the means for each distribution lies outside of the credible interval of the opposite distribution. This chart retains some ambiguity, but it appears to show that treatment has a higher likelihood of leading to an improved outcome.

Problem 4 Code

```
#Entering Jeffreys  prior
treatment_prior_shape  <- 1/2
treatment_prior_scale  <- 1/2
placebo_prior_shape <- 1/2
placebo_prior_scale <- 1/2

#Find posteriors  given data
treatment_posterior_shape <- treatment_prior_shape + 42
treatment_posterior_scale <- treatment_prior_scale + 18
placebo_posterior_shape <- placebo_prior_shape + 34
placebo_posterior_scale <- placebo_prior_scale + 26


#Calculate the beta distributions and credible intervals to plot
x <- seq(0,1,length=100)

lambda1 <-
dbeta(x,shape1=treatment_posterior_shape,shape2=treatment_posterior_sc
ale)

lambda2 <-
dbeta(x,shape1=placebo_posterior_shape,shape2=placebo_posterior_scale)

cred_int1 <-
qbeta(c(.025,.975),shape1=treatment_posterior_shape,shape2=treatment_p
osterior_scale)

cred_int2 <-
qbeta(c(.025,.975),shape1=placebo_posterior_shape,shape2=placebo_poste
rior_scale)

#Draw the actual plot, adding lines for the second distribution,
# segments for the credible intervals, and points (x’s) for the
# ends of the credible interval line segments
plot(x,lambda1,type='l',col='blue',ylim=c(0,7),ylab='Density',xlab='Pr
obability')


lines(x,lambda2,col='red')
segments(x0 = c_int1[1], y0 = -.10, x1 = c_int1[2], y1 = -
.10,col='blue')
```

```
segments(x0 = c_int2[1], y0 = -.20, x1 = c_int2[2], y1 = -
.20,col='red')

points(cred_int1[1],-.1,col='blue',pch='x')
points(cred_int1[2],-.1,col='blue',pch='x')
points(cred_int2[1],-.2,col='red',pch='x')
points(cred_int2[2],-.2,col='red',pch='x')

legend(0,7,c("Treatment","Placebo"),col=c("blue","red"),lty=c(1,1))
```

Problem 5

Generate 5000 random pairs (q_{1i}, q_{2i}), i=1, ..., 5000 from the joint posterior distribution for (Q_1, Q_2). Use this random sample to estimate the posterior probability that the rate of improvement is higher for treatment than for placebo. Does your analysis support the hypothesis that the drug alleviates symptoms of depression? Explain clearly your process for generating the sample. Discuss your analysis and results.

- Initially, I use a sample size of 5000 and created a random sampling from the given distributions
    - Sample 1 was 5000 examples using the rbeta function and the shape and scale parameters for "Treatment"
    - Sample 2 was the same using "Placebo" parameters
- Then I took the count of times the difference was greater than zero divided by the total observations
- This gave me a probability that treatment improved outcomes, the result from problem 4, of 94%.
- This answered the question asked, but I explored a bit further by repeating the steps above 10,000 times
- This game me a normal distribution of the probability treatment improved outcomes compared to placebos with a mean approximately the same as above; 94%
- The standard deviation of this distribution was 0.35%
- This dMC shows a 94% probability that treatment has a higher outcome than placebos. That is not to say it works 94% better.

Problem 5 Code

```
library(MASS)

#Monte Carlo sample  size
sample_size <- 5000

#calculate the lambda values in monte carlo  and take the difference
lambda1 <-
rbeta(sample_size,shape1=treatment_posterior_shape,shape2=treatment_po
sterior_scale)
lambda2 <-
rbeta(sample_size,shape1=placebo_posterior_shape,shape2=placebo_poster
ior_scale)

Difference  <-  lambda1-lambda2
probTreatBetter <- sum(Difference>0)/length(Difference)

print(probTreatBetter)
# output: [1] 0.9376
```

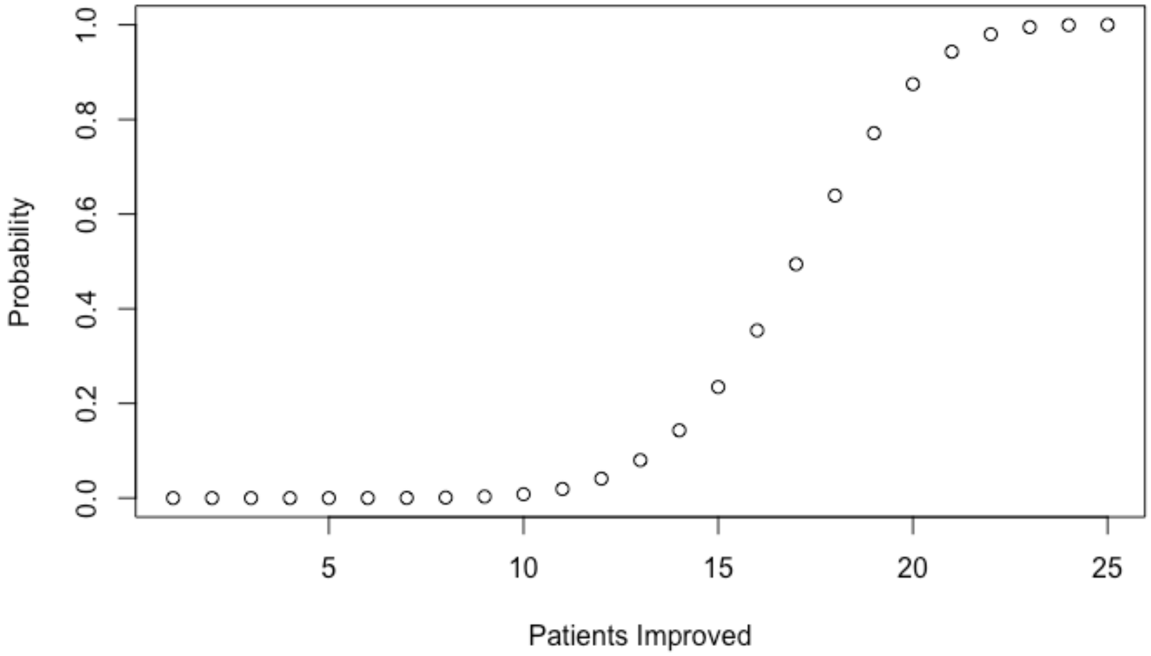
```
#Iterate over Monte Carlo simulation to get sample distribution of
differences
iters  <- seq(0,1,length=10000)
mc_result <- c()
for (i in iters){
  lambda1 <-
rbeta(sample_size,shape1=treatment_posterior_shape,shape2=treatment_po
sterior_scale)
  lambda2 <-
rbeta(sample_size,shape1=placebo_posterior_shape,shape2=placebo_poster
ior_scale)
  Difference  <-  lambda1-lambda2
  probTreatBetter <- sum(Difference>0)/length(Difference)
  mc_result  <- c(mc_result,probTreatBetter)
}

#plotting the histogram  of results
hist_data  <- hist(mc_result,prob=TRUE)
lines(density(mc_result))

#Fitting normal distribution
#This can also be done analytically  by just taking
#the mean and standard distribution
fit_params  <-fitdistr(mc_result,'normal')
print(fit_params)

#  Values I got running this:
#       mean            sd
#  9.354622e-01   3.492441e-03
```

Problem 6

Suppose 25 new patients are given the drug. Using the posterior distribution from problem 4 as the prior distribution for the probability of reporting improvement, find the predictive distribution for the number of patients who will report improvement 30 days after receiving the drug. State the distribution type and parameters. Find the posterior probability that 20 or more patients will report improvement in 30 days.

To calculate this I used a betabinomial as I was looking for the number of successes. The parameters used for the dbetabinom function are p, m, and n, for which I used, respectively, 25, alpha / (alpha+beta), and (alpha+beta), where alpha and beta are the posterior shape and scale from problems 4 and 5.

The chance that 20 or more patients will report improvement in 30 days (the initial success condition) is 87%. Further, the distribution or reporting improvement out of the existing distribution is shown below.



Problem 6 Code

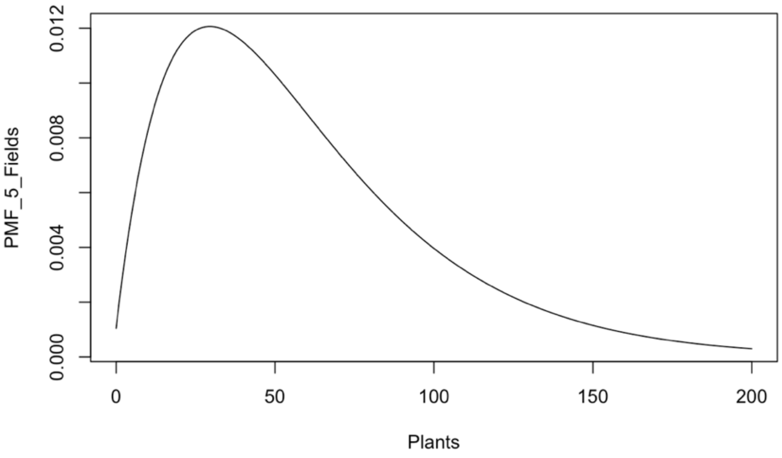
(the treatment_posterior_shape and treatment_posterior_scale variables are from Problem 5)

```
library("rmutil")

p = treatment_posterior_shape  /
(treatment_posterior_shape+treatment_posterior_scale)
m = treatment_posterior_shape+treatment_posterior_scale
n = 25

# Predictive Dist
pred_dist <-  dbetabinom(seq(0,25,1),n,p,m)
plot(seq(0,25,1),pred_dist,type="l",ylim=c(0,.15),col='black',
xlab="Trials",
     ylab="Probability  ",)


#Visually checking to make sure this makes sense
probability_20 <- pbetabinom(seq(1,25,length=25),n,p,m)
plot(probability_20)

#getting the actual answer required
probability_20 <- pbetabinom(20,n,p,m)
print(probability_20)
```

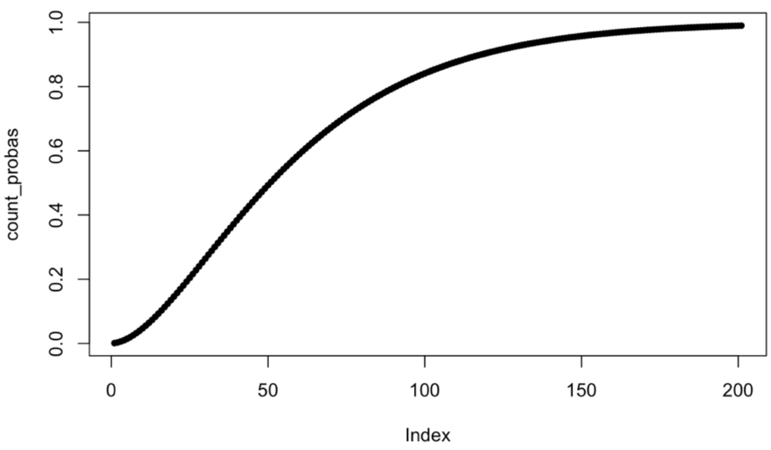
Problem 7

A researcher is performing a study of plant growth. To model the number of plants of a given species that will be found on a 10 m^2 plot of land, she uses a Poisson distribution with unknown rate parameter Λ that depends on the species and the soil conditions. Based on her knowledge of plants and soil conditions, she assesses a conjugate Gamma prior distribution with shape 2 and scale 6 for the mean number of plants growing on a 10m^2 plot of land of a given type. She plans to count the number of plants growing on five plots of this given type. What is her predictive distribution for the total number of plants she will find growing on the five plots? Find the probability that she will count fewer than a total of 75 plants growing on the five plots.

The predictive distribution for a poisson with an unknown rate parameter is a negative binomial distribution. In this case, the size is 2 and the probability is 0.032, or 1/1+5*scale. The density plot of this distribution is shown below.



The probability that she will count fewer than 75 plants is 71.4%, which can be obtained by using the same shape and scale from the density function shown about in a probability function in R; pnbinom. Plotting this for 0 to 200 plants yields the graph below.



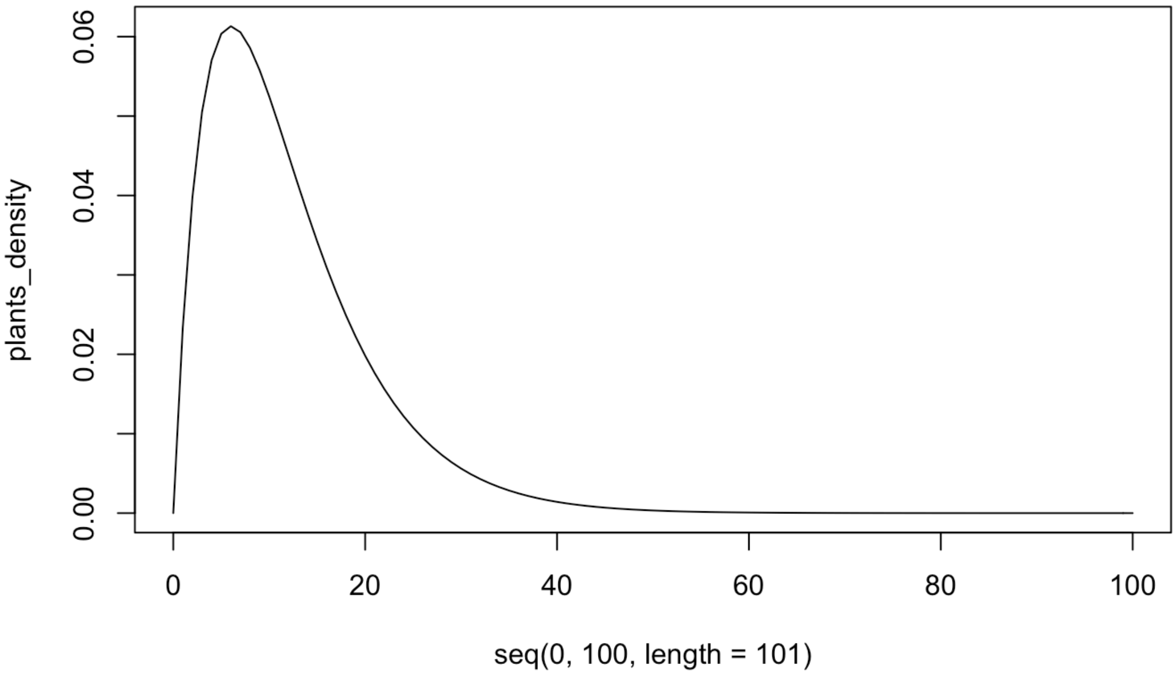
Problem 7 Code

```
g_shape_prior <- 2
g_scale_prior <- 6

# Plotting the gamma distribution just to make sure it makes sense
plants_density  <-
dgamma(seq(0,100,length=101),shape=g_shape_prior,scale=g_scale_prior)
plot(seq(0,100,length=101),plants_density,type='l')

# Plotting the negative binomial distribution
PMF_5_Fields <- dnbinom(seq(0,200,length=201), size=g_shape_prior,
prob=1/(1+5*g_scale_prior))
Plants <- seq(0,200,length=201)
plot(Plants,PMF_5_Fields,type='l')

# Getting the likelihood of <75 plants counted
prob75 <- pnbinom(75, size=g_shape_prior, prob=1/(1+5*g_scale_prior))
prob75

# Plotting the probabilities of all counts from 0 to 200
count_probas <- pnbinom(seq(0,200,length=201), size=g_shape_prior,
prob=1/(1+5*g_scale_prior))
plot(count_probas,pch=20)
```

Problem 8

The researcher goes out into the field and counts the plants she finds on the five plots. She counts a total of 72 plants. Find her posterior distribution for the Poisson parameter Λ. Name the type of distribution and hyperparameters. Find a 95% credible interval for Λ.

The posterior distribution for Lambda is a Gamma Distribution with shape 74 and scale 0.1935484. The credible interval for this 11.7 to 17.2 plants per plot. It is visualized in the chart below.



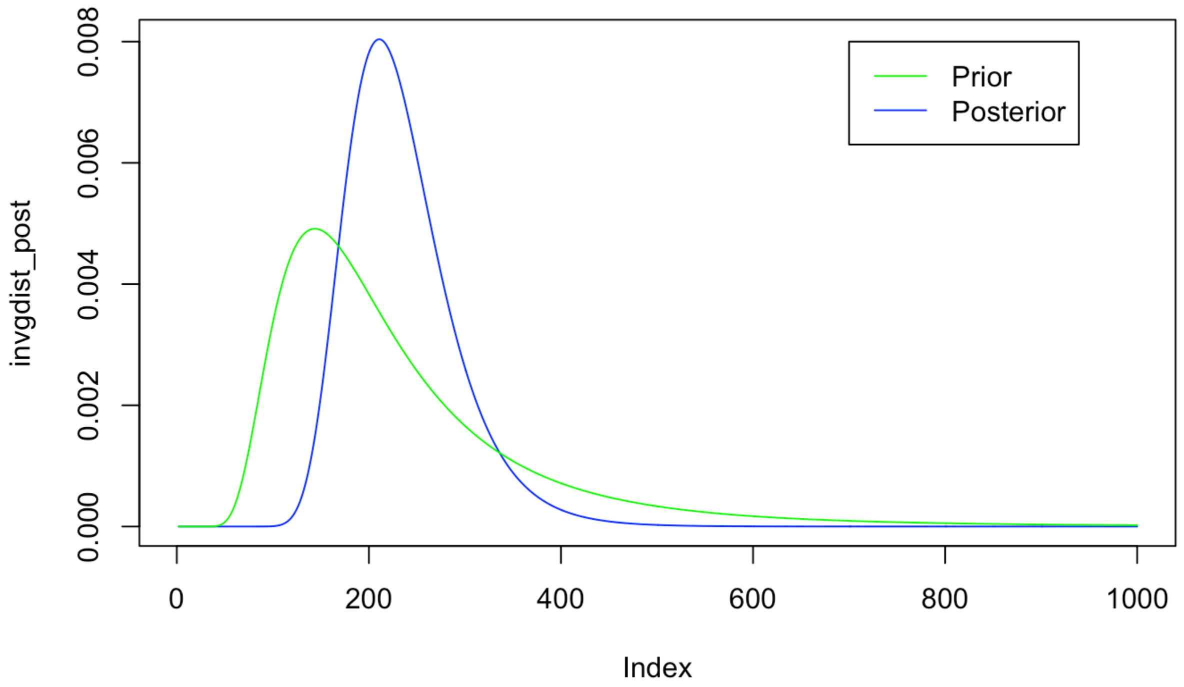
Problem 8 Code

```
g_shape_posterior <- 2 +  72
g_scale_posterior <- 1/  (1/6 + 5)
print(g_shape_posterior)
print(g_scale_posterior)
plant_post <-
dgamma(seq(0,100,length=101),shape=g_shape_posterior,scale=g_scale_pos
terior)
plot(seq(0,100,length=101),plants_density,type='l')

cred_int <- qgamma(c(0.05,0.95), shape=g_shape_posterior, scale
=g_scale_posterior)
print(cred_int)
```

Problem 9

Management at a call center is investigating the call load in order to find an efficient staffing policy. Assume that time intervals between calls are exponentially distributed. Assume the mean time between calls Q is constant during the mid-morning period. Assume an inverse Gamma prior distribution with shape a =4 and scale b = 0.0015 for Q. The following sequence of call times was collected during mid-Amorning, measured in seconds after the start of data collection: 168, 314, 560, 754, 1215, 1493, 1757, 1820, 1871, 1982, 2134, 2430, 3187, 3388, 3485. Find the posterior distribution for Q. Find the prior and posterior mean and standard deviation for Q. Discuss. (Note: Because of the memoryless property of the exponential distribution, you can treat the time until the first call as having an exponential distribution.)

The posterior shape is 19 and the posterior scale is 0.0002381357. The prior mean and standard deviations are 238.1and 168.4, while the posterior mean and standard deviation are 233.3 56.6. This shows that the initial estimate was actually a pretty good one; the mean moved by only about 2%. The dispersion of the distribution decreased by two thirds, but this is expected when observing new data that matches the prior distribution well.



Problem 9 Code

```
#visualizing the prior
require(invgamma)
alpha = 4
beta =  0.0014
v <- seq(0,1000,length=1000)
invgdist <- dinvgamma(v,shape=alpha,scale=beta)
plot(invgdist,type='l')

#updating data to be in the form I prefer (time since last)
obs <- c(168, 314, 560, 754, 1215, 1493, 1757, 1820, 1871, 1982, 2134,
2430, 3187, 3388, 3485)
obs_diff <- diff(obs)
obs <- c(168,obs_diff) # adding back the first time period
obs

theta  <- mean(obs)
theta

# Formulas to use:
# alpha* = alpha X n
# beta* =  (beta^-1 + sum(X_i))^-1
alpha_posterior <- alpha + length(obs)
beta_posterior <- (1/beta + sum(obs))^-1
alpha_posterior
beta_posterior

invgdist_post <-
dinvgamma(v,shape=alpha_posterior,scale=beta_posterior)
plot(invgdist_post,type='l',col='blue')
lines(invgdist,col='green')
legend(700,.008,c("Prior","Posterior"),col=c("green","blue"),lty=c(1,1
))

prior_mean <- 1 / (beta * (alpha-1))
prior_sd <- (1 / (beta^2 * (alpha-1)^2  * (alpha-2) ))^0.5
posterior_mean <- 1 / (beta_posterior * (alpha_posterior-1))
posterior_sd <- (1 / (beta_posterior^2 * (alpha_posterior-1)^2  *
(alpha_posterior-2) ))^0.5

print(prior_mean)
print(prior_sd)
print(posterior_mean)
print(posterior_sd)


(prior_mean-posterior_mean) / prior_mean
```

Problem 10

Do you think the model of independent and identically distributed exponential observations is a good model for the data of Problem 9? Explain your reasoning.

Partially. I think the events measured are independent- the time until the next call is not dependent on the time since the last call. In the example we have about 1 hour of observation time- 3,485 seconds. It may have been an hour with no calls in the last two minutes, but that is unknown. To me, this means the data may in fact be identically distributed in this case. However, I do not think all hours are identically distributed in reality. For example, if we created a distribution for each hour, we might find that most hours are from the same distribution, but the lunch hour is different. Given the purpose of being useful for a staffing policy, I think the distribution may be correct, but not necessarily the data collection itself.