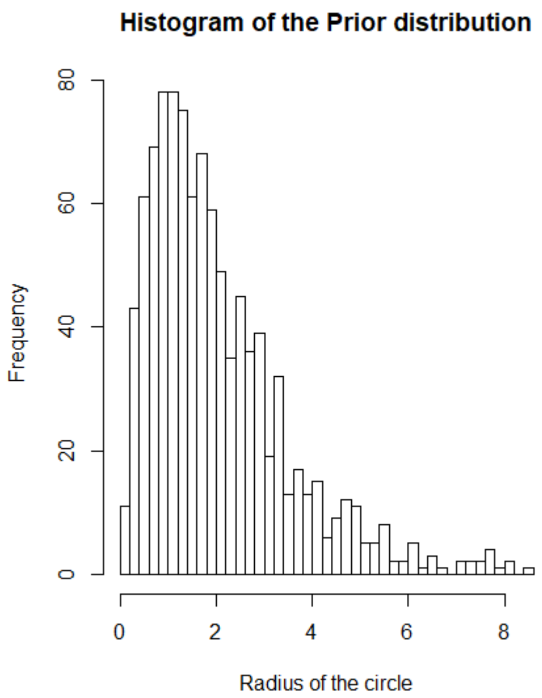
MATH5835M Practical

The aim of this report is to investigate what can be learned about the radius r of a circle in a plane where the number k of points inside the circle is observed. This will be achieved by using a Bayesian set-up where the gamma distribution is used as the prior density. Then using envelope rejection sampling, samples will be generated from the posterior density. This approach and the subsequent results will be discussed in more detail throughout the report.

Plotting the prior distribution

For the first step in the analysis, 1000 samples are generated from the prior distribution from which a histogram is plotted. Below is the histogram and the code used to generate the plot.




```
N=1000
alpha=2
Beta=1
#Generate 1000 samples from the prior distribution
prior <- rgamma(N,alpha,rate=Beta)
#Plot a histogram of these samples
hist(prior,breaks=30,xlab="Radius of the circle",
     main="Histogram of the Prior distribution")
```

Finding a formula for the likelihood p(k|r)

$$\mathrm{p(k|r)} = P(K = k \mid R = r) = P(\sum_{i=1}^{n} 1_{\{X_i^2+Y_i^2 \le r^2\}}) = P(\sum_{i=1}^{k} X_i^2 + Y_i^2 \le r^2 + \sum_{i=k+1}^{n-k} 1 - (X_i^2 + Y_i^2 \le r^2)$$

$$= \prod_{i=1}^{k} P\left(X_i^2 + Y_i^2 \le r^2\right) \cdot \prod_{i=k+1}^{n-k} 1 - P\left(X_i^2 + Y_i^2 \le r^2\right) = \left(1 - e^{-\frac{r^2}{2}}\right)^k (e^{-\frac{r^2}{2}})^{n-k}$$

(as X_i, Y_i are independent)

$$= \left(1 - e^{-\frac{r^2}{2}}\right)^k e^{-\frac{(n-k)r^2}{2}}$$

Therefore the posterior density p(r|k) is:

$$p(r|k) = \frac{p(k|r)p(r)}{p(k)} = const \cdot \left(1 - e^{-\frac{r^2}{2}}\right)^k e^{-\frac{(n-k)r^2}{2}} r^{\alpha-1} e^{-\beta r}$$

as p(k) is a constant since k is assumed to be known.

Envelope Rejection Sampling method

The proposal density is $p(r) = \frac{1}{Z} r^{\alpha-1} e^{-\beta r}$.

The target density is $p(r|k) = const \cdot \left(1 - e^{-\frac{r^2}{2}}\right)^k e^{-\frac{(n-k)r^2}{2}} r^{\alpha-1} e^{-\beta r}$

The method is as follows:

1. For n=1,2,3…
2. We generate a proposal R_n with prior density p(r)
3. Then generate U_n ~ U[0,1]
4. If $c \cdot p(R_n) \cdot U_n \le p(R_n|k)$ then accept the proposal R_n
5. Otherwise reject the proposal R_n and go back to step 2
6. Repeat until the desired number of accepted samples are obtained from the posterior density

Determining the required constant c

In order to apply the envelope rejection sampling method, a constant c >0 needs to be determined such that $f(x) \le cg(x)\ for\ all\ x \in \mathbb{R}$.

The value chosen for c should be as small as possible as this leads to a fast method and therefore

$c = sup \frac{target\ density}{proposal\ density}$ is the best choice of c to use.

$$c = sup \frac{target\ density}{proposal\ density} = \frac{p(r|k)}{p(r)} = \frac{const. \left(1 - e^{-\frac{r^2}{2}}\right)^k e^{-\frac{(n-k)r^2}{2}} r^{\alpha-1} e^{-\beta r}}{\frac{1}{Z} r^{\alpha-1} e^{-\beta r}} = \left(1 - e^{-\frac{r^2}{2}}\right)^k e^{-\frac{(n-k)r^2}{2}}$$

Substituting $p = 1 - e^{-\frac{r^2}{2}}$ into the equations gives: $c \le \left(1 - e^{-\frac{r^2}{2}}\right)^k e^{-\frac{(n-k)r^2}{2}} = p^k \cdot (1-p)^{n-k}$

To work out the maximum, first simplify using the natural logarithm then differentiate the expression and set equal to 0.

$$p^k \cdot (1-p)^{n-k} = k \ln p + (n-k) \ln 1 - p$$

$$\frac{d\ (k \ln p + (n-k) \ln 1 - p)}{dp} = \frac{k}{p} - \frac{n-k}{1-p} = 0 \quad \Rightarrow \quad k(1-p) = p(n-k) \;\Rightarrow\; p(n-k+k) = k \;\Longrightarrow\; \boldsymbol{p = \frac{k}{n}}$$

This then gives the value of p when c is at its maximum

$$\therefore p = \frac{k}{n} = 1 - e^{\frac{-r^2}{2}} \quad \Longrightarrow \quad e^{\frac{-r^2}{2}} = 1 - \frac{k}{n} = \frac{n-k}{n} \quad \Longrightarrow \quad \frac{-r^2}{2} = \ln \frac{n-k}{n} \quad \Longrightarrow \boldsymbol{r = \sqrt{2 \ln \frac{n}{n-k}}}$$

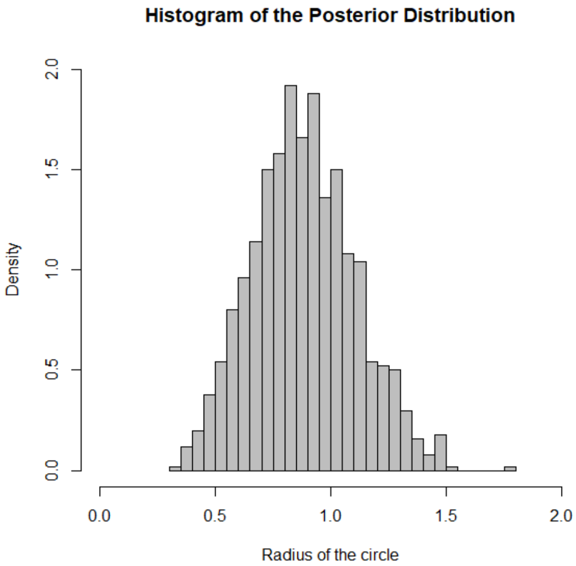
The positive square root is taken as r is always positive.

Substituting back into the equation:

$$c = \left(1 - e^{-\frac{r^2}{2}}\right)^k e^{-\frac{(n-k)r^2}{2}} = \left(1 - \frac{n-k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k} = \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k} = \boldsymbol{\frac{k^k (n-k)^{n-k}}{n^n}}$$

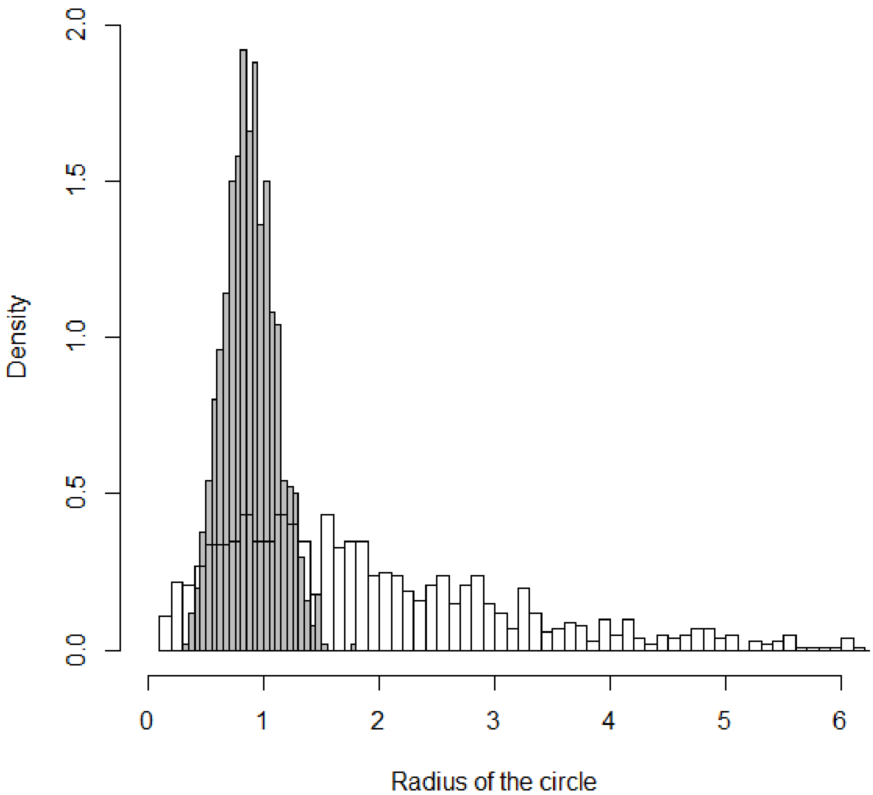
Implementing the rejection sampling method

Implementing the method detailed above, the following histogram is obtained. The code used to generate the histogram has also been included below.




```
#Envelope Rejection Sampling Method Code
N <- 1000
count <- 0
res <- NULL # empty vector
n <- 10
k <- 3
c <- ((k^k) *(n-k)^(n-k))/(n^n)
while (count < N) {
  R <- rgamma(1,alpha)
  U <- runif(1)
  if (c* (R*exp(-R) * U <= (1-exp((-R^2)/2))^k * exp(-(n-k)*(R^2)/2)*R*exp(-R)) ) {
    count <- count+1
    #Accept R value and add to res vector
    res <- c(res, R)
  } else { #Reject R value
  }
}
#Plot the histogram
hist(res,breaks=30,col="grey",probability=T,xlab="Radius of the circle",
     main="Histogram of the Posterior Distribution",ylim=c(0,2),xlim=c(0,2))

#Plotting the dual histogram
hist(res,breaks=30,col="grey",probability=T,xlab="Radius of the circle",
     ylim=c(0,2),xlim=c(0,6),main="")
hist(prior,breaks=70,xlab="Radius of the circle",probability=T,xlim=c(0,6),add=T)
```



To compare the histogram of the prior density and the histogram of the posterior density, both histograms have been drawn on the same plot. From this it can be seen that distribution of the target density is ‘enveloping’ the proposal density confirming that the method has worked. It is worth noting that the width of the target density looks to have “shrunk” in comparison to the proposal but this is fine as it lends to a simplified target density which looks to resemble a binomial distribution.

Computing the Monte Carlo Estimates

To compute the Monte Carlo estimate for the posterior expectation E(R|K=3), simply take the mean of the accepted values calculated above in the rejection sampling method. This gives the expected radius R of a circle when K=3 points have been observed inside that circle.

$$E(R|K=3) = mean(accepted\ R_n) = 0.88$$

```
#Posterior expectation E(R|K=3)
Exp = mean(res)
```

The Monte Carlo estimate for the posterior probability is computed as follows

$$P\left(R < \frac{1}{2}\middle|K=3\right) = mean\,(accepted\ R_n < 0.5) = 0.036$$

```
#Posterior prob P(R<1/2|K=3)
Prob <- mean(res<0.5)
```

This tells us the probability of the radius being less than 0.5 when K=3 points have been observed inside that circle is rather small. The code for computing both estimates can be seen on the right (where res is the vector of accepted proposals).

Discussing the choice of Monte Carlo sample size and the error of estimates

To work out the error of the estimates used, the mean square error is computed.

$$MSE = \frac{\sigma^2}{N} = \frac{Var(accepted\ X_n)}{N}$$ where N is the sample size.

It can be seen from the MSE formula above, that the error of the estimates is inversely proportional to the sample size. If the sample size increases, the error should decrease and conversely if the sample size decreases, the error of the estimates should increase. The chosen sample size should be chosen such that it has the right balance between reducing the error whilst still being computationally efficient to compute.

To illustrate this, if the number of samples in this analysis is doubled from 1000 to 2000, the MSE approximately halves from 0.000049 to 0.0000267. It doesn't decrease by a half exactly as estimates are used.

Conclusion

To conclude, the above approach will help to find radius r of a circle given the observation of k points by using an envelope rejection sampling method. The accuracy of the estimates can be improved by increasing the sample size but this should be balanced with having a computationally efficient method.