# Assignment 1 (DASC 6510)

Kaleb Williams (T00567109)
Dr. Jabed Tomal

January 2024 (Winter 2024)

## Code

All codes used to create graphics and complete simulations is located here:

https://github.com/kalebwilliams774/BayesianMachineLearning

## Exercise 2.4

a) By axiom **P3** for the $j^{th}$ of $k$ entries of the partition $H = \{H_1, \dots, H_k\}$ is given by

$$P(H_j|E)P(E) = P(H_j \cap E). \tag{1}$$

b) Each $H_j$ in the partition is disjoint from every other entry meaning that there is no common elements between each entry. Therefore the partition can be represented as a union of each individual element of the partition.

$$\begin{aligned} H &= H_1 \cup \dots \cup H_k \\ &= \bigcup_{k=1}^{K} H_k \end{aligned} \tag{2}$$

By axiom **P2** the probability that the union is completely covered by the event $E$ is given by

$$\begin{aligned} P(\left(\bigcup_{k=1}^{K} H_k|E\right) &= P(H_1|E) + P(H_2|E) + \dots + P(H_k|E) \\ &= \frac{P(H_1 \cap E)}{P(E)} + \frac{P(H_2 \cap E)}{P(E)} + \dots \frac{P(H_k \cap E)}{P(E)}. \end{aligned} \tag{3}$$

If $E$ is composed of $k$ disjoint parts of the partition then $E$ is the same as the union of those $k$ elements. Then by axiom **P1** we have

$$
\begin{aligned}
P\left(\bigcup_{k=1}^{K} H_k|E\right) &= P\left(\bigcup_{k=1}^{K} H_k|E\right) \\
&= 1. \qquad (4)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
1 &= \frac{P(H_1 \cap E)}{P(E)} + \frac{P(H_2 \cap E)}{P(E)} + \ldots \frac{P(H_k \cap E)}{P(E)} \\
P(E) &= P(H_1 \cap E) + P(H_2 \cap E) + \ldots P(H_k \cap E) \\
&= P(H_1 \cap E) + P\left(\bigcup_{k=2}^{K} H_k|E\right). \qquad (5)
\end{aligned}
$$

We know that all partitions not inside of $E$ also have zero probability of occurring from axiom **P2**.

c) We have shown in part b) by axiom **P2** that the probability of the event $E$, the event covering $K$ portions of the partition is given by

$$
P(E) = P(H_1 \cap E) + P(H_2 \cap E) + \ldots P(H_k \cap E). \qquad (6)
$$

Therefore,

$$
P(E) = \sum_{k=1}^{K} P(H_k \cap E) \qquad (7)
$$

d) Combining all derived quantities we have

$$
\begin{aligned}
P(H_j|E)\left(\sum_{k=1}^{K} P(H_k \cap E)\right) &= P(E|H_j)P(H_j) \\
P(H_j|E) &= \frac{P(E|H_j)P(H_j)}{\sum_{k=1}^{K} P(H_k \cap E)} \qquad (8)
\end{aligned}
$$

# Exercise 3.7

a) From the details of the data we can conclude the sampling distribution is a binomial distribution with parameter $y_1 = 2$, two successes, with a size of $n_1 = 15$ and unknown probability parameter *theta*.

Generally, a posterior distribution for a parameter $\theta$ $p(y_1, \ldots, y_n|\theta)$ and prior distribution $\pi(\theta)$ is given by

$$p(\theta|y_1, \ldots, y_n) = \frac{1}{\int_0^\infty p(y_1, \ldots, y_n|\theta)d\theta} p(y_1, \ldots, y_n|\theta)\pi(\theta). \tag{9}$$

Our sampling distribution is then given by

$$\begin{aligned} p(y_1|\theta) &= \binom{n_1}{y_1}\theta^{y_1}(1-\theta)^{n_1-y_1} \\ &= \binom{15}{2}\theta^2(1-\theta)^{13}. \end{aligned} \tag{10}$$

The prior is described as a uniformly distributed, since $\theta$ is representative of the probability of a binomial success, the uniform prior will simply be equal to 1 as it will be able to take on possible values of 0 through 1.

The posterior distribution for our data and prior is then given by

$$\begin{aligned} p(\theta|y_1) &= \binom{15}{2}\theta^2(1-\theta)^{13}(1) \\ &= c(y)\theta^2(1-\theta)^{13} \end{aligned} \tag{11}$$

In order to find the true distribution the normalization constant must be found by taking the reciprocal of the integral over the entire range. From calculus it is know that

$$\int_0^\infty \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{12}$$

Rearranging our posterior distribution we have

$$p(y_1|\theta) = c(y)\theta^{3-1}(1-\theta)^{14-1}. \tag{13}$$

Therefore the normalization constant is given by,

$$\begin{aligned} c(y)\int_0^1 p(y_1|\theta)d\theta = c(y)\frac{\Gamma(3)\Gamma(14)}{\Gamma(14+3)} &= 1 \\ = c(y)\frac{\Gamma(3)\Gamma(14)}{\Gamma(17)} &= 1 \end{aligned}$$

$$\frac{1}{\int_0^\infty p(y_1|\theta)d\theta} = c(y) = \frac{\Gamma(17)}{\Gamma(3)\Gamma(14)}. \tag{14}$$

The posterior distribution for the binomial probability parameter $\theta$ is then given by

$$p(\theta|y_1) = \frac{\Gamma(17)}{\Gamma(3)\Gamma(14)}\theta^{3-1}(1-\theta)^{14-1} \tag{15}$$

Which tells us that $\theta$ is distributed as a $Beta(3, 14)$ distribution. Now the mean, mode and standard deviation of a Beta distribution are well known quantities.

$$E(\theta) = \frac{\alpha}{\alpha+\beta} = \frac{3}{3+14} \approx 0.1764706 \tag{16}$$

$$Mode(\theta) = \frac{\alpha-1}{\alpha+\beta-2} = \frac{3-1}{3+14-2} \approx 0.1333333 \tag{17}$$

$$\sigma(\theta) = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}} = \sqrt{\frac{(3)(14)}{(2+14)^2)(3+14+1)}} \approx 0.0571662 \tag{18}$$

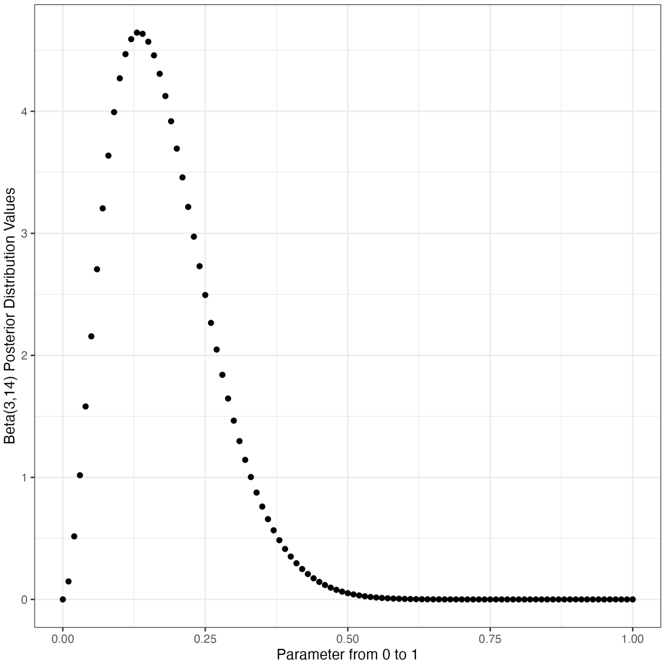

Figure 1: Posterior distribution, $Beta(3, 14)$ using a uniform prior distribution, of a binomial distribution with a single data sample $y_1 = 2$ and sample size $n_1 = 2$. Varying parameter $\theta$ from 0 to 1, possible probabilities for the success of the pilot students.

b) In order to find the posterior predictive distribution of $Y_2$ for given $Y_1$ and known sample sizes $n_1, n_2$ we must complete the following calculation

$$P(Y_2 = y_2|Y_1 = y_1) = \int_0^1 P(Y_2 = y_2|\theta)p(\theta|y_1)d\theta. \tag{19}$$

For the distribution of $Y_1$ and $Y_2$ to take on the above form, $Y_1$ and $Y_2$ must be conditionally independent of each other as well as the posterior predictive distribution does not depend on the posterior parameter $\theta$ but only the data obtained.

$P(Y_2 = y_2|\theta)$ is the sampling distribution for $Y_2$ and $p(\theta|y_1)$ is the posterior distribution for $Y_1$. The sample type of sampling was done for both $Y_1$ and $Y_2$, therefore the sampling distribution is given by a binomial probability distribution, sample size for $Y_2$ is given as $n_2 = 278$.

$$p(Y_2 = y_2|\theta) = \binom{278}{y_2}\theta^{y_2}(1-\theta)^{278-y_2} \tag{20}$$

Therefore the posterior predictive distribution for $Y_2$ is given by

$$\begin{aligned} P(Y_2 = y_2|Y_1 = y_1) &= \int_0^1 \binom{278}{y_2}\theta^{y_2}(1-\theta)^{278-y_2}\frac{\Gamma(17)}{\Gamma(3)\Gamma(14)}\theta^{3-1}(1-\theta)^{14-1}d\theta \\ &= \binom{278}{y_2}\frac{\Gamma(17)}{\Gamma(3)\Gamma(14)}\int_0^1 \theta^{(y_2+3)-1}(1-\theta)^{(292-y_2)-1}d\theta. \end{aligned} \tag{21}$$

$$\int_0^1 \theta^{(y_2+3)-1}(1-\theta)^{(292-y_2)-1}d\theta = \frac{\Gamma(y_2+3)\Gamma(292-y_2)}{\Gamma(295)}$$

$$P(Y_2 = y_2|Y_1 = y_1) = \binom{278}{y_2}\frac{\Gamma(17)}{\Gamma(3)\Gamma(14)}\frac{\Gamma(y_2+3)\Gamma(292-y_2)}{\Gamma(295)} \tag{22}$$

From properties of the gamma function we know that

$$\Gamma(x+1) = x!. \tag{23}$$

Therefore, we can simplify the gamma ratios that come from the integration and posterior distribution of $\theta$.

$$\Gamma(y_2+3) = \Gamma((y_2+2)+1) = (y_2+2)! \tag{24}$$

$$\Gamma(292-y_2) = \Gamma((291-y_2)+1) = (291-y_2)! \tag{25}$$

$$\Gamma(295) = \Gamma(294+1) = 294! \tag{26}$$

From combinatorics we know the binomial coefficient is defined as

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}. \tag{27}$$

Therefore we can write the reciprocal of the last gamma ratio in the predictive posterior distribution as

$$\begin{aligned}\frac{\Gamma(295)}{\Gamma(y_2+2)\Gamma(291-y_2)} &= \frac{294!}{(y_2+2)!(291-y_2)!} \\ &= \frac{294 \times 293!}{(y_2+2)!(291-y_2)!} \\ &= 294\binom{293}{y_2+2}\end{aligned}$$

$$\frac{\Gamma(y_2+2)\Gamma(291-y_2)}{\Gamma(295)} = \frac{1}{294\binom{293}{y_2+2}} \tag{28}$$

Apply the same method to the second gamma ratio we have

$$\begin{aligned}\frac{\Gamma(17)}{\Gamma(3)\Gamma(14)} &= \frac{16!}{2!13!} \\ &= \frac{16 \times 15!}{2!13!} \\ &= 16\binom{15}{2}.\end{aligned} \tag{29}$$

Combining all terms we arrive at the predictive posterior distribution for $Y_2$.

$$\begin{aligned}P(Y_2 = y_2|Y_1 = 2) &= \frac{\binom{278}{y_2} \times 16\binom{15}{2}}{294\binom{293}{y_2+2}} \\ &= \frac{16}{294}\frac{\binom{278}{y_2}\binom{15}{2}}{\binom{293}{y_2+2}}\end{aligned} \tag{30}$$

c)

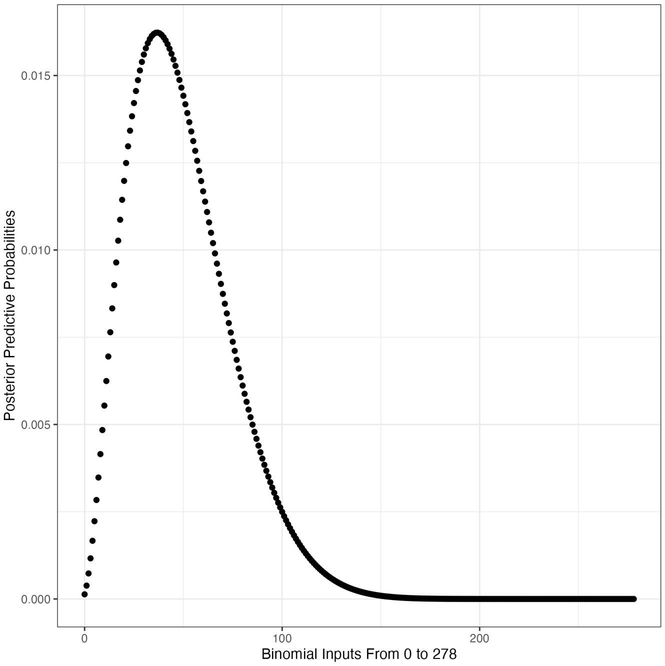

Figure 2: Posterior predictive distribution of $Y_2$ given that $Y_1 = 2$.

From R we have that the mean and standard deviation of the posterior predictive distribution of $Y_2$ are 0.003584229 and 0.005370246 respectively.

d) We want the distribution of $P(Y_2 = y_2|\theta = \hat{\theta})$ where $\theta = \hat{\ }{}^{2}/_{15}$. We know from part a) that the distribution of $Y_2$ is given by a binomial with $n_2 = 278$ and unknown probability parameter $\theta$. Using the MLE of $\hat{\theta} = {}^{2}/_{15}$ as the unknown probability parameter we have that

$$P\left(Y_2 = y_2|\theta = \frac{2}{15}\right) \sim Binomial\left(n = 278, \theta = \frac{2}{15}\right) \tag{31}$$

$$\begin{aligned} P\left(Y_2 = y_2|\theta = \frac{2}{15}\right) &= \binom{278}{y_2}\left(\frac{2}{15}\right)^{(} y_2)\left(1 - \frac{2}{15}\right)^{(} 278 - y_2) \\ &= \binom{278}{y_2}\left(\frac{2}{15}\right)^{y_2}\left(\frac{13}{15}\right)^{278-y_2} \end{aligned} \tag{32}$$

From R we have that the mean and standard deviation of the distribution of $Y_2$ given that the probability parameter is the MLE given by $\hat{\theta} = {}^{2}/_{165}$ are 0.003584229 and 0.01289371 respectively.

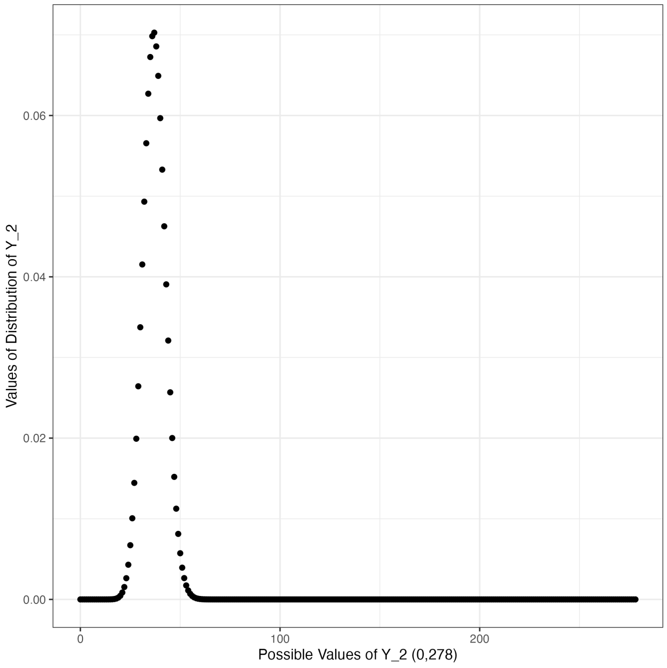

Figure 3: Plot of the distribution of $Y_2$ given that the probability parameter for the parameter is the MLE $\hat{\theta} = \frac{2}{15}$.

| | $\mu$ | $\sigma$ |
|---|---|---|
| $P(Y_2 = y_2 \vert Y_1 = 2)$ | 0.003584229 | 0.005370246 |
| $P\left(Y_2 = y_2 \vert \theta = \frac{2}{15}\right)$ | 0.003584229 | 0.01289371 |

Table 1: Mean and standard deviation for the posterior predictive distribution $Y_2$ and the distribution of $Y_2$ given that $\theta = \frac{2}{15}$.

From calculation of the means in the distributions we can see that we arrive at the same mean for both the distribution of $Y_2$ given the probability parameter is given by the MLE and the posterior predictive distribution. The variance is more then doubled when using the same MLE distribution indicating that the posterior predictive distribution will have a less spread out sample.

This is supported by the graphs as we can see for the posterior predictive distribution there is a higher concentration of non-zero sample points, then for the MLE distribution.

The obvious choice for sampling is the posterior predictive distribution. As it offers same sample mean with lower sample variance.

# Exercise 4.8

a) Let $\theta_A$ and $\theta_B$ represent the average number of men in their 30's with children with and without a bachelors degree respectively. Using a Poisson sampling distribution and a $Gamma(1, 2)$ prior distribution we will calculate the posterior distributions given both data sets.

For the posterior distribution, the sampling distribution of multiple random independent $x_i$, the joint distribution is simply given by the product of all the distributions describing each $x_i$, which is also the likelihood function.

$$\begin{aligned} p(y_1, \ldots, y_n|\theta) &= \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{y_i}}{y_i!} \\ &= \frac{e^{-\theta}\theta^{y_1}}{y_1!}\frac{e^{-\theta}\theta^{y_2}}{y_2!}\cdots\frac{e^{-\theta}\theta^{y_n}}{y_n!} \\ &= \frac{e^{-\theta-\cdots-\theta}\theta^{y_1+\cdots+y_n}}{x_1!\ldots y_n} \\ &= \frac{e^{-n\theta}\theta^{\sum\limits_{i=1}^{n} y_i}}{\prod\limits_{i=1}^{n} y_i!} \end{aligned} \tag{33}$$

A $Gamma(\alpha, \beta)$ prior is given by

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}. \tag{34}$$

Therefore, the posterior distribution for a Poisson sampling distribution of $n$ observations with a $Gamma(\alpha, \beta)$ prior is given by

$$\begin{aligned} p(\theta|y_1, \ldots, y_n) &\propto p(y_1, \ldots, y_n|\theta)\pi(\theta) \\ &= \frac{e^{-n\theta}\theta^{\sum\limits_{i=1}^{n} y_i}}{\prod\limits_{i=1}^{n} y_i!}\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \\ &= \left(\prod_{i=1}^{n} y_i!\right)^{-1}\frac{\beta^\alpha}{\Gamma(\alpha)}e^{-n\theta-\beta\theta}\theta^{\sum\limits_{i=1}^{n} y_i+\alpha-1} \\ &= c(y|\alpha, \beta)e^{-\theta(n+\beta)}\theta^{\sum\limits_{i=1}^{n} y_i+\alpha-1} \end{aligned} \tag{35}$$

In order to find the normalization constant $c(y|\alpha, \beta)$ we must integrate out all possible values of $\theta$ from the posterior. We also know the integrating a PDF over its range equates to 1.

$$c(y|\alpha|\beta) \int_0^\infty e^{-\theta(n+\beta)} \theta^{\sum\limits_{i=1}^{n} x_i + \alpha - 1} d\theta = 1 \tag{36}$$

Our posterior resembles a $Gamma(\alpha, \beta)$ distribution, applying the same technique to a regular Gamma distribution we have

$$\int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} d\theta = 1$$

$$\int_0^\infty \theta^{\alpha-1} e^{-\beta\theta} d\theta = \frac{\Gamma(\alpha)}{\beta^\alpha}. \tag{37}$$

Therefore out normalization constant is given by

$$c(y|\alpha, \beta) \frac{\Gamma\left(\sum\limits_{i=1}^{n} y_i + \alpha\right)}{(n+\beta)^{\sum\limits_{i=1}^{n} y_i + \alpha}} = 1$$

$$c(y|\alpha, \beta) = \frac{(n+\beta)^{\sum\limits_{i=1}^{n} y_i + \alpha}}{\Gamma\left(\sum\limits_{i=1}^{n} y_i + \alpha\right)}. \tag{38}$$

The full posterior distribution is then given by

$$p(\theta|y_1, \ldots, y_n) = \frac{(n+\beta)^{\sum\limits_{i=1}^{n} y_i + \alpha}}{\Gamma\left(\sum\limits_{i=1}^{n} y_i + \alpha\right)} e^{-\theta(n+\beta)} \theta^{\sum\limits_{i=1}^{n} y_i + \alpha - 1}. \tag{39}$$

Therefore our posterior distribution for a Poisson sampling distribution with a $Gamma(\alpha, \beta)$ prior distribution is

$$p(\theta|y_1, \ldots, y_n) \sim Gamma\left(\sum_{i=1}^{n} y_i + \alpha, n + \beta\right) \tag{40}$$

From the given data sets we have

| | sum($x_i$) | n |
|---|---|---|
| Children (A) | 54 | 58 |
| No Children (B) | 305 | 218 |

Table 2: Sum of all values and number of values in each data set. Men over thirty with children and a bachelors degree (Group A) and men over thirty with children without a bachelors degree (Group B).

Therefore we have that the average rate parameters for the two datasets are given by Gamma distributions.

$$\theta_A \sim Gamma(56, 59) \tag{41}$$

$$\theta_B \sim Gamma(307, 219) \tag{42}$$

Now the posterior predictive distribution can be found by taking the product of estimated Poisson distribution and the posterior distribution and integrating out the unknown parameter $\theta$.

$$\begin{aligned}
P(\tilde{y}|y_1,\ldots,y_n) &= \int_{i=0}^{n} P(\tilde{y}|\theta)p(\theta|y_1,\ldots,y_n)d\theta \\
&= \int_0^\infty \left(\frac{e^{-\theta}\theta^{\tilde{y}}}{\tilde{y}!}\right) \frac{(n+\beta)^{\sum\limits_{i=1}^{n} y_i+\alpha}}{\Gamma\left(\sum\limits_{i=1}^{n} y_i+\alpha\right)} \theta^{\sum\limits_{i=1}^{n} y_i+\alpha-1} e^{-(n+\beta)\theta}d\theta \\
&= \frac{(n+\beta)^{\sum\limits_{i=1}^{n} y_i+\alpha}}{\tilde{y}!\Gamma\left(\sum\limits_{i=1}^{n} y_i+\alpha\right)} \int_0^\infty e^{-\theta}\theta^{\tilde{y}}\theta^{\sum\limits_{i=0}^{n} y_i+\alpha} e^{-(n+\beta)\theta}d\theta \\
&= \frac{(n+\beta)^{\sum\limits_{i=1}^{n} y_i+\alpha}}{\tilde{y}!\Gamma\left(\sum\limits_{i=1}^{n} y_i+\alpha\right)} \int_0^\infty e^{-(n+\beta+1)}\theta^{\sum\limits_{i=1}^{n} y_i+\alpha+\tilde{y}-1} d\theta.
\end{aligned} \tag{43}$$

We know from earlier that

$$\int_0^\infty \theta^{\alpha-1}e^{-\beta\theta}d\theta = \frac{\Gamma(\alpha)}{\beta^\alpha}. \tag{44}$$

Therefore the posterior predictive distribution simplifies to

$$
\begin{aligned}
P(\tilde{y}|y_1,\dots,y_n) &= \frac{(n+\beta)^{\sum\limits_{i=1}^{n} y_i+\alpha}}{\tilde{y}!\Gamma\left(\sum\limits_{i=1}^{n} y_i+\alpha\right)} \frac{\Gamma\left(\sum\limits_{i=1}^{n} y_i+\alpha+\tilde{y}\right)}{(n+\beta+1)^{\sum\limits_{i=1}^{n} y_i+\alpha+\tilde{y}}} \\
&= \frac{\Gamma\left(\sum\limits_{i=1}^{n} y_i+\alpha+\tilde{y}\right)}{\tilde{y}_i\Gamma\left(\sum\limits_{i=1}^{n} y_i+\alpha\right)} \frac{(n+\beta)^{\sum\limits_{i=1}^{n} y_i+\alpha}}{(n+\beta+1)^{\sum\limits_{i=1}^{n} y_i+\alpha+\tilde{y}}} \\
&= \frac{\Gamma\left(\sum\limits_{i=1}^{n} y_i+\alpha+\tilde{y}\right)}{\Gamma(\tilde{y}+1)\Gamma\left(\sum\limits_{i=1}^{n} y_i+\alpha\right)} \frac{(n+\beta)^{\sum\limits_{i=1}^{n} y_i+\alpha}}{(n+\beta+1)^{\sum\limits_{i=1}^{n} y_i+\alpha}} \left(\frac{1}{n+\beta+1}\right)^{\tilde{y}} \\
&= \frac{\Gamma\left(\sum\limits_{i=1}^{n} y_i+\alpha+\tilde{y}\right)}{\Gamma(\tilde{y}+1)\Gamma\left(\sum\limits_{i=1}^{n} y_i+\alpha\right)} \left(\frac{n+\beta}{n+\beta+1}\right)^{\sum\limits_{i=1}^{n} y_i+\alpha} \left(\frac{1}{n+\beta+1}\right)^{\tilde{y}} \qquad (45)
\end{aligned}
$$

This tells us that our posterior predictive distribution for a Poisson sampling distribution and $\Gamma(\alpha,\beta)$ prior is a negative binomial distribution.

$$
P(\tilde{y}|y_1,\dots,y_n) \sim NB\left(\sum_{i=1}^{n} y_i+\alpha, \beta+n\right) \qquad (46)
$$

For our prior and data we have that the posterior predictive distributions for group A and group B are

$$
Y_A \sim NB(56,59) \qquad (47)
$$

$$
Y_B \sim NB(307,219). \qquad (48)
$$

In order to obtain MC samples we can first calculate a sample size of 5000 from the gamma posterior for both data sets, and use the parameter values to sample from the Poisson sampling distribution.

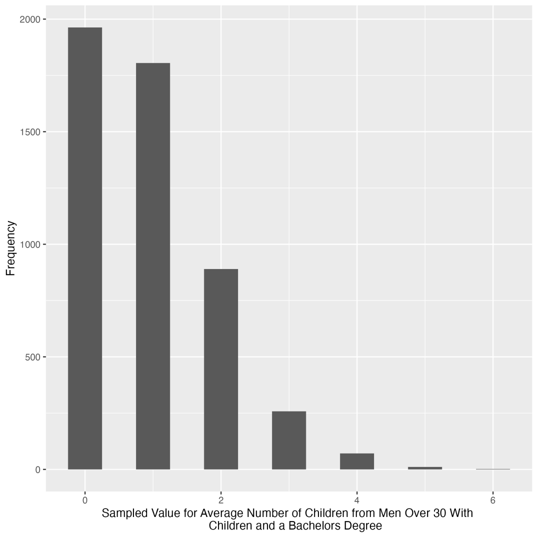

Figure 4: 5000 sized sample, $Y_A$, from the posterior predictive distribution for average number of children for men over 30 with a bachelors degree

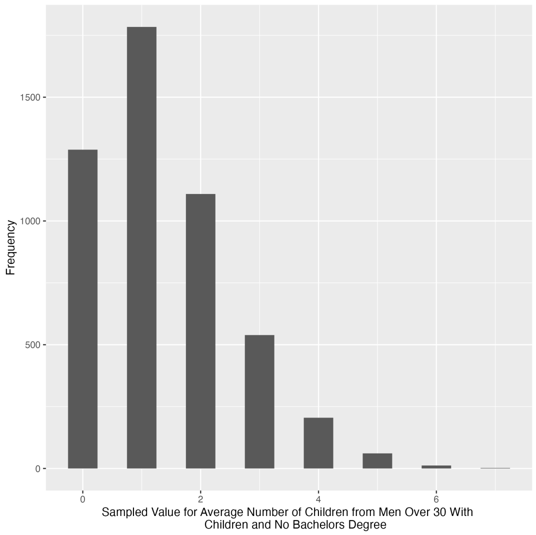

Figure 5: 5000 sized sample, $Y_B$, from the posterior predictive distribution for average number of children for men over 30 without a bachelors degree

b) The 95% confidence intervals for the difference in the parameters, $\theta_B - \theta_A$, and the sampled values, $\bar{Y}_B - \bar{Y}_A$ are given by

$$\begin{gathered} \theta_B - \theta_A : (0.1531726, 0.7352204) \\ \bar{Y}_B - \bar{Y}_A : (-2, 4) \end{gathered} \tag{49}$$

Since we are sampling from a Poisson distribution the parameters, $\theta_A$ and $\theta_B$, are representative of the average number of children the men have. The difference $\theta_B - \theta_A$ is positive for both ends of the confidence interval this indicates that men without bachelors degrees in their 30's tend to have more children then men in their 30's with bachelors degrees. From the plot we can see that the proportion of the average number of children for men without bachelors degree is higher, then that of men with bachelors degrees.

The difference $\bar{Y}_B - \bar{Y}_A$ quantifies the difference in the distributions of the number of children between the two populations. At the 95% end of the confidence interval we see that the difference between the number of children of men in their 30's without bachelors degree and men of the same age with bachelors degrees is 4, instating that for 95% of the data the difference of $\bar{Y}_B - \bar{Y}_A$ is 4 children. Meaning again, men in their 30's without bachelors degree seem to have more children then men with bachelors degrees as given 95% of the data the difference on average was 4 children in favour of men without bachelors degrees.

Investigating the means of the samples we have that $\mu_A = 0.942$ and $\mu_B = 1.366$. this demonstrates the mean number of children for both men is relatively close in contrast to the difference between the average number of children and the number of children predicted by the posterior predictive distribution.

c) The empirical distribution of the dataset is simply the distribution of the observed frequencies of the data itself. It represents the frequencies or proportions of different values or categories present in the dataset.

From figure 6 we can see that the empirical distribution consistently under performs the Poisson distribution. This meaning that the Poisson distribution will overestimate the number of children men in their 30's without bachelors degrees will have. Therefore, not the best fit for the data.

d) From figure 7 we can see the Poisson model greatly over predicts the number of children men with no bachelors degree over 30 will have, represented by the re point on the above plot. Therefore, the Poisson model is not an appropriate model as it will consistently over perform what is given by the data.

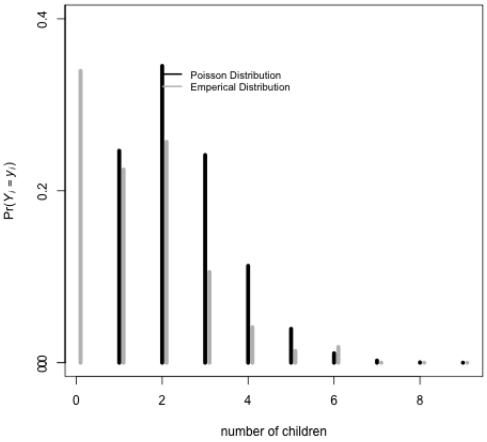

Figure 6: Poisson distribution with parameter $\hat{\theta} = 1.4$ and empirical distribution of men in their 30's without a bachelors degree.

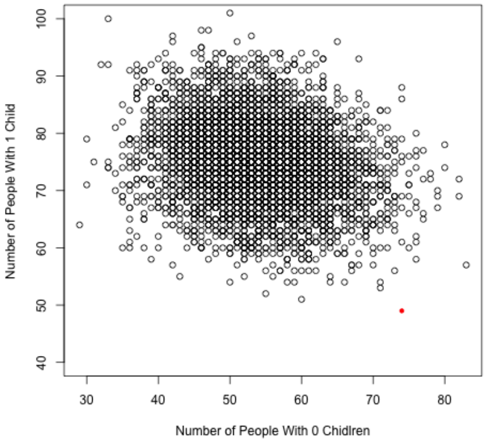

Figure 7: Plot demonstrating the amount of men over 30 without a bachelors degree have 0 and 1 child from a simulated Poisson distribution for each simulated parameter value $\theta$. The red point is demonstrating the true number of children the men have from the data given in this case: 74 with 0 children and 49 with 1 child.

## Exercise 5.2

In order to find the distribution of $P(\theta_A < \theta_B | \mathbf{y}_A, \mathbf{y}_B)$ we must first find the posterior distributions of both $\theta_A$ and $\theta_B$ to combine them into the marginal posterior distribution of both the parameters. Given that we have independent conjugate normal priors, with normal sampling distribution, with parameters $\mu_o = 75$ and $\sigma_o^2 = 100$. We wish to perform inference on the mean given the variance for the posterior have the posterior for parameters $\theta$ is given by

$$\theta \sim N\left(\mu_n, \sigma_n^2\right)$$

$$\mu_n = \frac{\kappa_o \mu_o + n\bar{y}}{\kappa_o + n} \tag{50}$$

$$\sigma_n^2 = \frac{1}{\nu_o + n}\left(\nu_o \sigma_o^2 + (n-1)s^2 + \frac{\kappa_o n}{\kappa_o + n}\left(\bar{y} - mu_o\right)^2\right)$$

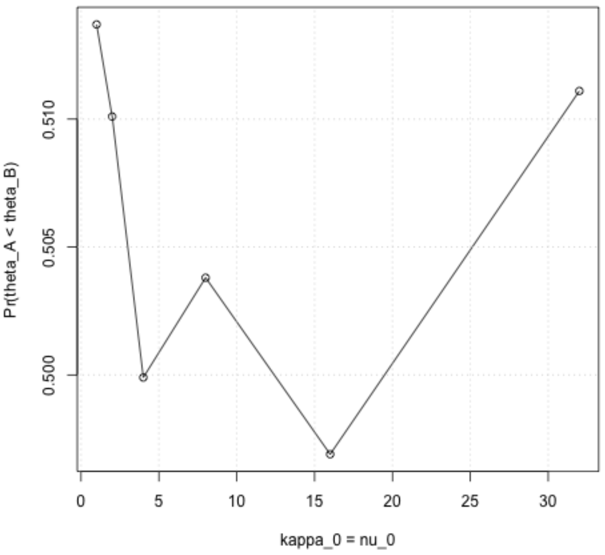

Figure 8: Plot of the probability that $\theta_A < \theta_B$ given that $\kappa_o = \nu_o \epsilon \{1, 2, 4, 16, 32\}$ from a thousand samples of each normal distribution.

From the above plot we can see that the probability that $\theta_A < \theta_B$ is greater than 50% for all values of $\kappa_o$ and $\nu_o$. For various values of the parameters we see little variation in the probability that $\theta_A < \theta_B$, meaning that the result is relatively insensitive to which values are chosen for the parameters. Furthermore, since the probability is greater than 50% we can conclude that under the normal priors $\theta_A$ will be greater than $\theta_B$ under large variation of $\kappa_o$ and $\nu_o$.

## Exercise 6.1

a) $\theta_A$ and $\theta_B$ would be dependent under this prior distribution. Since $\theta_B$ is written as a ratio of itself and $\theta_A$ the prior distribution would be a joint prior distribution as $\gamma$ simple scales the parameter $\theta$ by the ratio of the two sampling distributions parameters. This type of distribution is justified when we have belief that the parameters $\theta_A$ and $\theta_B$ can be scaled by some common parameters $\gamma$ that incorporates the two parameters.

b) The form of the joint distribution of $\theta$ given $\gamma$, $y_A$ and $y_B$ can be inferred through the use of Bayes's theorem. The full conditional distribution is given by

$$\begin{aligned} P(\theta, \gamma, y_A, y_B) &= P(y_A|\theta, \gamma)P(y_B|\theta, \gamma)\pi(\theta)\pi(\gamma) \\ &= \left( \frac{e^{-n_A\theta}\theta^{\sum_{i=1}^{n} y_{Ai}}}{\prod_{i=1}^{n} y_{Ai}} \right) \left( \frac{e^{-n_B\theta}(\theta\gamma)^{\sum_{i=1}^{n} y_{Bi}}}{\prod_{i=1}^{n} y_{Bi}} \right) \left( \frac{b_\theta^{a_\theta}}{\Gamma(a_\theta)} \theta^{a_\theta - 1} e^{-b_\theta\theta} \right) \left( \frac{b_\gamma^{a_\gamma}}{\Gamma(a_\gamma)} \theta^{a_\gamma - 1} e^{-b_\gamma\gamma} \right). \end{aligned} \tag{51}$$

From Bayes's theorem we know we can write the full conditional distribution of $\theta$ given $\gamma$, $y_A$ and $y_B$ as

$$\begin{aligned} P(\theta|\gamma, y_A, y_B) &= \frac{P(\theta, \gamma, y_A, y_B)}{P(\gamma, y_A, y_B)} \\ &= \frac{\left( \frac{e^{-n_A\theta}\theta^{\sum_{i=1}^{n_A} y_{Ai}}}{\prod_{i=1}^{n_A} y_{Ai}} \right) \left( \frac{e^{-n_B\theta}(\theta\gamma)^{\sum_{i=1}^{n_B} y_{Bi}}}{\prod_{i=1}^{n_B} y_{Bi}} \right) \left( \frac{b_\theta^{a_\theta}}{\Gamma(a_\theta)} \theta^{a_\theta - 1} e^{-b_\theta\theta} \right) \left( \frac{b_\gamma^{a_\gamma}}{\Gamma(a_\gamma)} \theta^{a_\gamma - 1} e^{-b_\gamma\gamma} \right)}{\left( \frac{e^{-n_B\gamma}\gamma^{\sum_{i=1}^{n_B} y_{Bi}}}{\prod_{i=1}^{n_B} y_{Bi}} \right) \left( \frac{b_\gamma^{a_\gamma}}{\Gamma(a_\gamma)} \theta^{a_\gamma - 1} e^{-b_\gamma\gamma} \right)} \\ &= \left( \frac{e^{-n_A\theta}\theta^{\sum_{i=1}^{n_A} y_{Ai}}}{\prod_{i=1}^{n_A} y_{Ai}} \right) \left( \frac{e^{-n_B\theta}\theta^{\sum_{i=1}^{n_B} y_{Bi}}}{\prod_{i=1}^{n_B} y_{Bi}} \right) \left( \frac{b_\theta^{a_\theta}}{\Gamma(a_\theta)} \theta^{a_\theta - 1} e^{-b_\theta\theta} \right) \\ &\propto e^{-n_A\theta}\theta^{\sum_{i=1}^{n_A} y_{Ai}} e^{-n_B\theta}\theta^{\sum_{i=1}^{n_B} y_{Bi}} \theta^{a_\theta - 1} e^{-b_\theta\theta} \\ &= \theta^{\sum_{i=1}^{n_A} y_{Ai} + \sum_{i=1}^{n_B} y_{Bi} + a_\theta - 1} e^{-n_A\theta - n_B\theta - b_\theta\theta} \end{aligned} \tag{52}$$

$$= \theta^{\sum_{i=1}^{n_A} y_{Ai} + \sum_{i=1}^{n_B} y_{Bi} + a_\theta - 1} e^{-(n_A + n_B + b_\theta)\theta} \tag{53}$$

This gives us a Gamma distribution for the full conditional of $\theta$.

$$\theta \sim Gamma\left(\sum_{i=1}^{n_A} y_{Ai} + \sum_{i=1}^{n_B} y_{Bi} + a_\theta, n_A + n_B + b_\theta\right) \tag{54}$$

c) Applying a similar technique we can find the full conditional distribution of $\gamma$ given $\theta$, $y_A$ and $y_B$.

$$\begin{aligned}
P(\gamma|\theta, y_A, y_B) &= \frac{P(\theta, \gamma, y_A, y_B)}{P(\theta, y_A, y_B)} \\
&= \frac{\left(\frac{e^{-n_A\theta}\theta^{\sum_{i=1}^{n_A} y_{Ai}}}{\prod_{i=1}^{n_A} y_{Ai}}\right)\left(\frac{e^{-n_B\theta\gamma}(\theta\gamma)^{\sum_{i=1}^{n_B} y_{Bi}}}{\prod_{i=1}^{n_B} y_{Bi}}\right)\left(\frac{b_\theta^{a_\theta}}{\Gamma(a_\theta)}\theta^{a_\theta-1}e^{-b_\theta\theta}\right)\left(\frac{b_\gamma^{a_\gamma}}{\Gamma(a_\gamma)}\theta^{a_\gamma-1}e^{-b_\gamma\gamma}\right)}{\left(\frac{e^{-n_A\theta}\theta^{\sum_{i=1}^{n_A} y_{Ai}}}{\prod_{i=1}^{n_A} y_{Ai}}\right)\left(\frac{e^{-n_B\theta}\theta^{\sum_{i=1}^{n_B} y_{Bi}}}{\prod_{i=1}^{n_B} y_{Bi}}\right)\left(\frac{b_\theta^{a_\theta}}{\Gamma(a_\theta)}\theta^{a_\theta-1}e^{-b_\theta\theta}\right)} \\
&= e^{-n_B\gamma}\gamma^{\sum_{i=1}^{n_B} y_{Bi}} \frac{b_\gamma^{a_\gamma}}{\Gamma(a_\gamma)}\gamma^{a_\gamma-1}e^{-b_\gamma\gamma} \\
&\propto \gamma^{\sum_{i=1}^{n_B} y_{Bi} + a_\gamma - 1} e^{-n_b\gamma - b_\gamma\gamma} \\
&\propto e^{-n_B\gamma}\gamma^{\sum_{i=1}^{n_B} y_{Bi}} \frac{b_\gamma^{a_\gamma}}{\Gamma(a_\gamma)}\gamma^{a_\gamma-1}e^{b_\gamma\gamma} \\
&\propto \gamma^{\sum_{i=1}^{n_B} y_{Bi} + a_\gamma - 1} e^{-(n_b + b_\gamma)\gamma}
\end{aligned} \tag{55}$$

Which again is another Gamma distribution.

$$\gamma \sim Gamma\left(\sum_{i=1}^{n_B} y_{Bi} + a_\gamma, b_\gamma + n_b\right) \tag{56}$$

d)

| $a_\gamma = b_\gamma$ | $E(\theta_B - \theta_A \vert \mathbf{y}_A, \mathbf{y}_B)$ |
|---|---|
| 8 | 0.4592726 |
| 16 | 0.4768495 |
| 32 | 0.2918187 |
| 64 | 0.1153146 |
| 128 | -0.1512068 |

Table 3: Demonstrating the expected change in the average number of children from men over 30 with no bachelors degree and men over 30 with a bachelors degree respectively as the the parameters of the scaling distribution are increased and equally valued.

As the values of $a_\gamma$ and $b_\gamma$ are increased and equally valued we see that the expected difference in the number of average children between the two groups decreases. When the prior scaling distributions, $\gamma$ parameters increase from 64 to 128 we see on average that the expected number of children is larger in the group of men with bachelors degree. This meaning with lower parameter values for the $\gamma$ prior we will be opinionated that men without bachelors degrees have more children, with higher men with bachelors degrees have more children.

## Exercise 7.5

a)

| $\hat{\theta}_A$ | $\hat{\theta}_B$ | $\hat{\sigma}_A^2$ | $\hat{\sigma}_B^2$ | $\hat{\rho}$ |
|---|---|---|---|---|
| 24.20049 | 24.80535 | 4.0928 | 4.691578 | 0.6164509 |

Table 4: Empirical estimates for the mean, variance and correlation for the two data sets.

b) After imputing the data based on the equations we arrive at the confidence interval for the paired t-test is given by

$$CI : (-3.566376, 2.697862) \tag{57}$$

c) Using Jeffrey's prior we have that the posterior distribution we have the parameters $\theta$ and $1/\sigma^2$ are given by

$$\begin{gathered} \theta \sim multivariatenormal(\mu_n, \Lambda_n) \\ \frac{1}{\sigma^2} \sim inverse - Wishart(\nu_n, \mathbf{S}_n^{-1}). \end{gathered} \tag{58}$$

The posterior mean for $\theta_A - \theta_B$ is given by 13.50383 and the 95% confidence is given by

$$CI : (5.378221, 21.629437). \tag{59}$$

From part b) we can see a major shift in the range that the 95% confidence interval covers, applying Bayesian methodology gives a wider range for the difference in the mean values of the distribution using Jeffrey's prior. A non-informative prior regarding the beliefs of the parameters.

# Q7

a) For this problem we will choose a binomial sampling distribution, with $n$ trials and $k$ successes, with a beta prior distribution. Then the posterior is also a beta distribution with parameters $\alpha + k$ and $\beta + n - k$. That is,

$$p \sim Beta(\alpha + k, \beta + n - k). \tag{60}$$

For the purposes of this portion of the question we will choose the parameter values as $n = 10$, $k = 2$, $\alpha = 0.5$ and $\beta = 0.5$. In order to calculate the credible intervals, we will need to calculate both regions of extreme values, since we are applying a confidence level of 80%, the extreme values for the intervals will be the those past 90% and 10% for the true probability parameter. The "middle of distribution" parameter will be that of 50%. This will be done using 1000 samples. After applying R code to calculate the proportion of the credible intervals in the middle and tail of the distribution we have for the middle the proportion is 0.6173333 and for the tail is 0.2916667. This demonstrates that the number of credible intervals from a true parameter in the tail of the prior is less than the nominal 80%.

In order to show the simulation variability, we can report the mean and variance of the upper and lower bounds of the credible intervals for using a true parameter value in the tail and in the middle of the prior distribution.

| | Mean | Variance |
|---|---|---|
| Middle Upper | 0.60750473 | 0.13641063 |
| Middle Lower | 0.39249527 | 0.13641063 |
| Tail Upper | 0.68362729 | 0.01133662 |
| Tail Lower | 0.39249527 | 0.01144820 |

Table 5: Means and variances for the upper and lower bounds of the credible intervals.

For the true parameter value closer to the middle of the distribution we see large variability in the upper and lower bounds of the credible intervals which can be attributed to simulation variation. For values located within the tails of the distributions we see very small fluctuation in the simulation variance.

b) Sampling the true parameter from the prior and using it to sample from our chosen binomial distribution we have that the proportion of 80% confidence intervals containing the true parameter values was 0.776. Near the true value confidence level of the credible interval.

# Q8

We are given that the sampling distribution is normal,$N(\theta, 1)$ and the prior is also normally distributed,$N(0, \tau^2)$. This meaning that the posterior distribution will also be normal. From chapter 5 we know that the posterior distribution for a normal sampling and a normal prior, with both mean and variance know is distributed as

$$\theta \sim N\left(\frac{\frac{1}{\tau_o^2}\mu_o + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_o^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\tau_o^2} + \frac{n}{\sigma^2}}\right). \tag{61}$$

Where $\tau_o^2$ is the prior variance, $\mu_o$ is the prior mean, $n$ is the sample size and $\sigma^2$ is the sampling variance. Since we are concerned with finding $E(\theta|Y)$ for the estimate of our mean parameter $\theta$ for the single observation drawn from the $N(\theta, 1)$ we must find the mean of the posterior.

We know that $\tau_o^2 = \tau^2$, $\mu_o = 0$, $n = 1$,$\sigma^2 = 1$ and since we have only one option the sample mean is equal to the sampling observation that is $\bar{Y} = Y$. Therefore the mean of the posterior is given as

$$\begin{aligned} \mu_n &= \frac{\frac{1}{\tau^2}(0) + \frac{1}{1^2}Y}{\frac{1}{\tau^2} + \frac{1}{1^2}} \\ &= \frac{Y}{\frac{1}{\tau^2} + 1} \\ &= \frac{\tau^2}{\tau^2 + 1}Y. \end{aligned} \tag{62}$$

We know have our estimate for $\theta$ and must calculated the Mean Square Error (MSE) with out estimate, also known as the Bayes's Risk, which is given by

$$\begin{aligned} MSE &= E\left((\hat{\theta} - \theta)^2\right) \\ &= E\left(\left(\frac{\tau^2}{\tau^2 + 1}Y - \theta\right)^2\right). \end{aligned} \tag{63}$$

We must calculate in terms of the "natures prior", $\theta \sim N(0, k^2)$ by fixing a value for $k$. Since we are using natures prior for our mean parameter $\theta$ we are assuming the true mean value is 0, that is our sampling distribution follows a standard normal distribution $N(0, 1)$. Bayes's Risk.

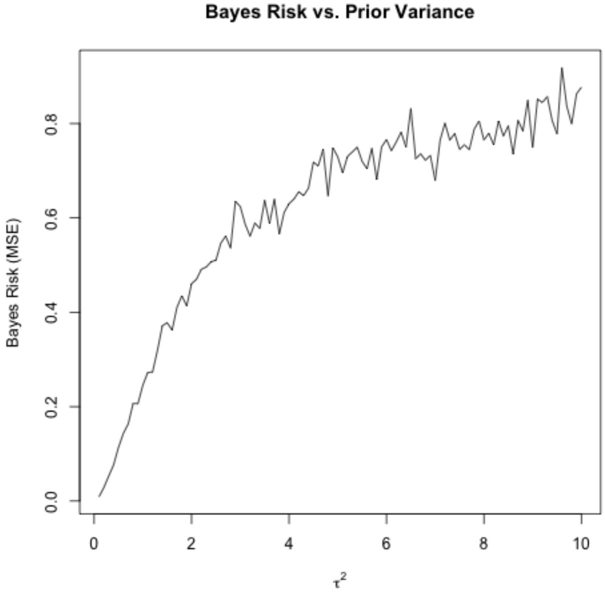

Figure 9: Bayes risk as a function of the prior distributions variance $\tau^2$.

From the above plot we can see as we increase the values of $\tau^2$ the Bayes's risk increases as well. This meaning the risk function increases as we vary the known parameter of $\tau^2$. This does not align with the facets of decision theory as the risk function increases until we hit a ceiling around a Bayes's risk of 0.80, where there is low variability and near constant value. This meaning the estimator we used does not minimize the risk as is intended within decision theory.

As Bayes risk increases as $\tau^2$ increases, it indicates that the wider prior distribution leads to poorer estimation performance, potentially due to increased bias, decreased precision, or both. This is demonstrated in the above plot meaning the observations suggest that an overly flat prior is not as effective in capturing the true underlying structure of the data compared to a concentrated more informative prior.