BIOSTAT/STAT 570: Coursework 5

To be submitted to the course canvas site by 11:59pm Monday 8th November, 2021.

1. Consider the data given in Table 1, which are a simplified version of those reported in Breslow and Day (1980). These data arose from a case-control study that was carried out to investigate the relationship between esophageal cancer and various risk factors. Disease status is denoted Y with $Y = 0/1$ corresponding to without/with disease and alcohol consumption is represented by X with $X = 0/1$ denoting $< 80g/ \geq$ 80g on average per day. Let the probabilities of high alcohol consumption in the cases and controls be denoted

$$p_1 = \Pr(X = 1 \mid Y = 1) \quad \text{and} \quad p_2 = \Pr(X = 1 \mid Y = 0),$$

respectively. Further, let X_1 be the number exposed from n_1 cases and X_2 be the number exposed from n_2 controls. Suppose $X_i \mid p_i \sim$ Binomial(n_i, p_i) in the case ($i = 1$) and control ($i = 2$) groups.

	$X = 0$	$X = 1$	
$Y = 1$	104	96	200
$Y = 0$	666	109	775

Table 1: Case-control data: $Y = 1$ corresponds to the event of esophageal cancer, and $X = 1$ exposure to greater than 80g of alcohol per day. There are 200 cases and 775 controls.

(a) Of particular interest in studies such as this is the *odds ratio* defined by

$$\theta = \frac{\Pr(Y = 1 \mid X = 1)/\Pr(Y = 0 \mid X = 1)}{\Pr(Y = 1 \mid X = 0)/\Pr(Y = 0 \mid X = 0)}.$$

Show that the odds ratio is equal to

$$\theta = \frac{\Pr(X = 1 \mid Y = 1)/\Pr(X = 0 \mid Y = 1)}{\Pr(X = 1 \mid Y = 0)/\Pr(X = 0 \mid Y = 0)} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}.$$

Solution: Rearranging terms we have that

$$\begin{aligned}
\theta &= \frac{\Pr(Y = 1 \mid X = 1)/\Pr(Y = 0 \mid X = 1)}{\Pr(Y = 1 \mid X = 0)/\Pr(Y = 0 \mid X = 0)} \\
&= \frac{\Pr(Y = 1 \mid X = 1)\Pr(Y = 0 \mid X = 0)}{\Pr(Y = 1 \mid X = 0)\Pr(Y = 0 \mid X = 1)}
\end{aligned}$$

We can apply Bayes' Rule to each of the terms in this equation.

$$\begin{aligned}
\Pr(Y=1 \mid X=1) &= \Pr(X=1 \mid Y=1)\Pr(Y=1)/\Pr(X=1) = p_1\frac{\Pr(Y=1)}{\Pr(X=1)} \\
\Pr(Y=0 \mid X=1) &= \Pr(X=1 \mid Y=0)\Pr(Y=0)/Pr(X=1) = p_2\frac{Pr(Y=0)}{\Pr(X=1)} \\
\Pr(Y=1 \mid X=0) &= \Pr(X=0 \mid Y=1)\Pr(Y=1)/\Pr(X=1) = (1-p_1)\frac{\Pr(Y=1)}{\Pr(X=1)} \\
\Pr(Y=0 \mid X=0) &= \Pr(X=0 \mid Y=0)Pr(Y=0)/Pr(X=0 = (1-p_2)\frac{Pr(Y=0)}{Pr(X=0)}
\end{aligned}$$

Combining these results, we have that

$$\begin{aligned}
\theta &= \frac{p_1\Pr(Y=1)/\Pr(X=1)}{(1-p_1)\Pr(Y=1)/\Pr(X=1)}\frac{(1-p_2)Pr(Y=0)/Pr(X=0)}{p_2Pr(Y=0)/\Pr(X=1)} \\
&= \frac{p_1/(1-p_1)}{p_2/(1-p_2)}
\end{aligned}$$

(b) Obtain the MLE and an asymptotic 90% confidence interval for θ, for the data of Table 1.
Solution:
One can derive the MLE through solving the score equations defined by the $X_i|p_i \sim$ Bin$(n_i,p_i), i=1,2$ likelihood. Let n_1 be the number of cases and n_2 the number of controls. The likelihood and the the log-likelihood are

$$\begin{aligned}
P(\boldsymbol{x}|p_1,p_2) &\propto p_1^{x_1}(1-p_1)^{n_1-x_i}p_2^{x_2}(1-p_2)^{n_2-x_2} \\
L(p_1,p_2) &= x_1\log(p_1)+(n_1-x_1)\log(1-p_1)+x_2\log(p_2)+(n_2-x_2)\log(1-p_2)+c(\boldsymbol{x}).
\end{aligned}$$

Let $\bar{x}_{n_1}$ and $\bar{x}_{n_2}$ be the average number of exposed individuals among cases and controls, respectively. Setting the score equations to 0 yields

$$\begin{aligned}
\frac{\partial L}{\partial p_1} &= \frac{x_1}{p_1}+\frac{(x_1-n_1)}{1-p_1}=0 \iff (1-p_1)n_1\bar{x}_{n_1}+p_1(n_1\bar{x}_{n_1}-n_1)=0 \Rightarrow \widehat{p}_1=\bar{x}_{n_1}=\frac{96}{200} \\
\frac{\partial L}{\partial p_2} &= \frac{x_2}{p_2}+\frac{(x_2-n_2)}{1-p_2}=0 \iff (1-p_2)n_2\bar{x}_{n_2}+p_2(n_2\bar{x}_{n_2}-n_2)=0 \Rightarrow \widehat{p}_2=\bar{x}_{n_2}=\frac{109}{775}
\end{aligned}$$

So by invariance of the MLE, the estimated odds ratio is $\widehat{\theta}=\frac{\frac{\widehat{p}_1}{1-\widehat{p}_1}}{\frac{\widehat{p}_2}{1-\widehat{p}_2}}=5.64$

We used `glm()` to obtain the 90% CI instead of deriving the information matrix and using the delta method to obtain the asymptotic distribution for the log-odds ratio. Details of coding are included in the Appendix. The MLE for the odds ratio is 5.6 (90% CI: 4.2, 7.5).

(c) We now consider a Bayesian analysis. Assume that the prior distribution for p_i is the beta distribution Be(a, b) for $i = 1, 2$. Show that the posterior distribution $\pi(p_1, p_2 \mid x_1, x_2)$ is given by the product of the beta distributions Be$(a+x_i, b+n_i-x_i)$, $i = 1, 2$.
Solution:

$$
\begin{aligned}
\pi(p_i \mid x_i) &\propto p(x_i \mid p_i)\pi(p_i) \\
&\propto p_i^{x_i}(1-p_i)^{n_i-x_i} p_i^{a-1}(1-p_i)^{b-1} \\
&= p_i^{(a+x_i)-1}(1-p_i)^{(b+n_i-x_i)-1} \\
&\propto \text{Beta}(a+x_i, b+n_i-x_i)
\end{aligned}
$$

(d) Consider the case $a = b = 1$. Obtain expressions for the posterior mean, mode and standard deviation. Evaluate these posterior summaries for the data of Table 1. Report 90% posterior credible intervals for p_1 and p_2.
Solution: The posterior distribution is Beta$(a + x_i, b + n_i - x_i)$, which gives

$$
\begin{aligned}
\text{E}(p_i \mid x_i) &= \frac{a+x_i}{a+b+n_i} \\
\text{mode}(p_i \mid x_i) &= \frac{a+x_i-1}{a+b+n_i-2} \\
\text{sd}(p_i \mid x_i) &= \sqrt{\frac{(a+x_i)(b+n_i-x_i)}{(a+b+n_i)^2(a+b+n_i+1)}}
\end{aligned}
$$

For p_1:

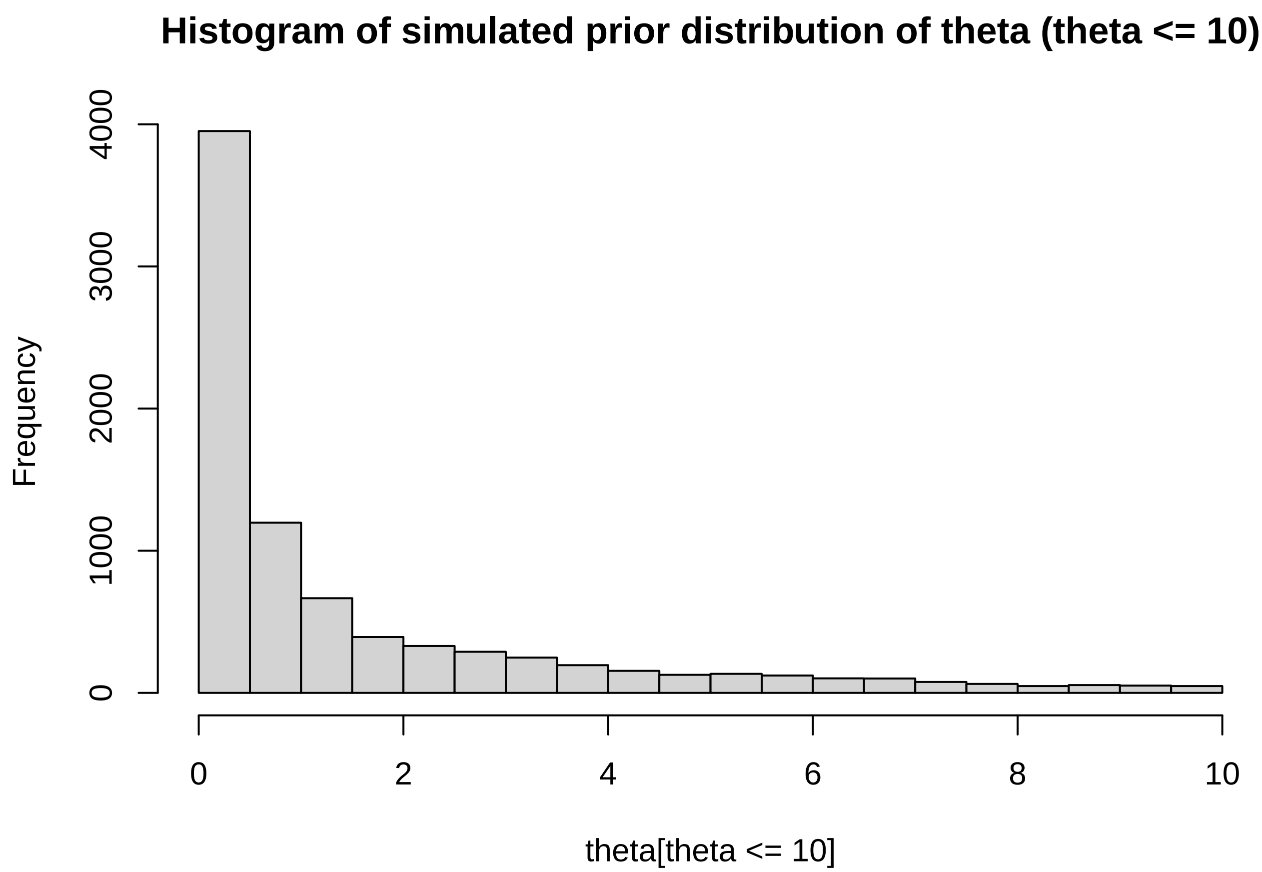
$$
\begin{aligned}
\text{E}(p_1 \mid x_1) &= 0.4802 \\
\text{mode}(p_1 \mid x_1) &= 0.4800 \\
\text{sd}(p_1 \mid x_1) &= 0.0351
\end{aligned}
$$

For p_2:

$$
\begin{aligned}
\text{E}(p_2 \mid x_2) &= 0.1416 \\
\text{mode}(p_2 \mid x_2) &= 0.1406 \\
\text{sd}(p_2 \mid x_2) &= 0.0125
\end{aligned}
$$

90% credible intervals were obtained using the `qbeta()` function. We have 90% posterior belief that p_1 is in (0.42, 0.54) and p_2 is in (0.12, 0.16).

(e) Examine the implied prior distribution for θ and give a 90% prior interval.
Solution: We simulated 10,000 samples of each of p_1 and p_2 from Beta(1,1) to obtained 10,000 samples of θ. The histogram of the samples of θ with $\theta \leq 10$ is plotted in the following. The 90% prior credible interval for θ is $(0.01, 61.11)$. We see that the prior credible interval accommodates a wide range of values. The overall distribution for θ appears to be informative based on the histogram and favors odds ratio of less than 1.



(f) Simulate samples $p_1^{(t)}, p_2^{(t)}$, $t = 1, ..., T = 1000$ from the posterior distributions $p_1 \mid x_1$ and $p_2 \mid x_2$. Form histogram representations of the posterior distributions using these samples and obtain sample-based 90% credible intervals.
Solution: We simulated 10,000 samples from the respective posteriors. These samples are shown in Figure 1.

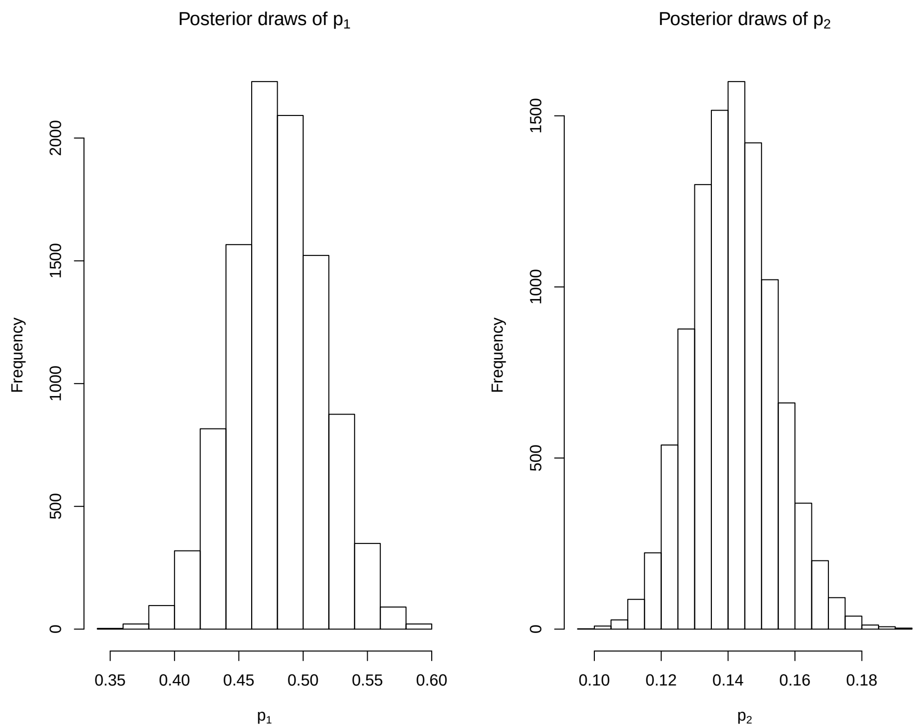


Figure 1: Histograms of samples from posteriors of p_1 and p_2.

Sample based 90% credible intervals for p_1 and p_2 were (0.44, 0.53) and (0.13, 0.16), respectively.

(g) Obtain samples from the posterior distribution of $\theta \mid x_1, x_2$ and form the histogram representation of the posterior. Obtain the posterior median and 90% credible interval for $\theta \mid x_1, x_2$ and compare with the likelihood analysis.

Solution: We took the posterior draws from p_1 and p_2 to obtain draws from the posterior distribution of θ. The histogram of the posterior distribution is shown in Figure 2.

The posterior median was 5.62, and the 90% credible interval was (4.22, 7.48), which is very similar to the MLE analysis. This is due to the non informative priors on p_1 and p_2 and the large sample sizes.

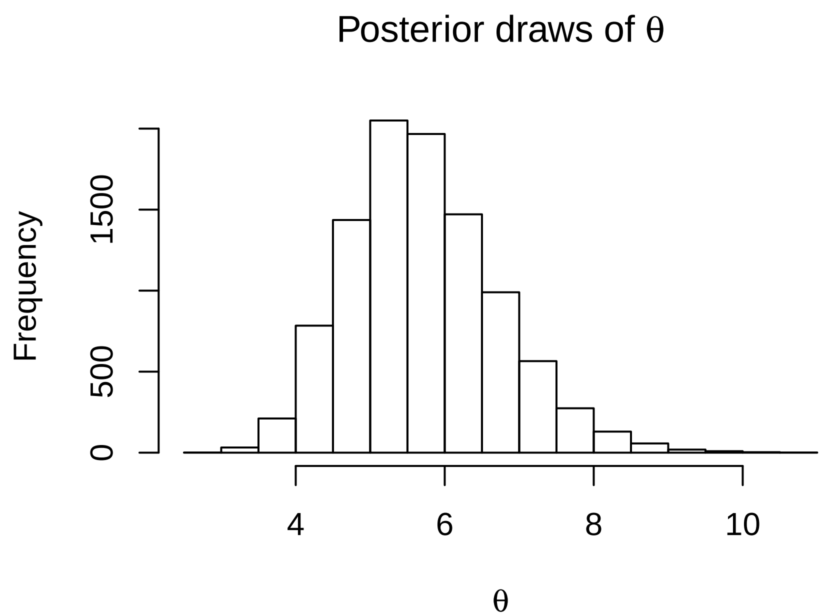


Figure 2: Histogram of samples from posterior of θ

(h) Suppose the rate of esophageal cancer is 18 in 100,000. Describe how this information may be used to evaluate

$$q_1 = \Pr(Y = 1 \mid X = 1) \quad \text{and} \quad q_0 = \Pr(Y = 1 \mid X = 0).$$

Solution: Previously we could only estimate $\Pr(X = 1 \mid Y = 1)$ and $\Pr(X = 0 \mid Y = 1)$ since this is a case control study. However, with this new information we can obtain estimates on $\Pr(Y = 1 \mid X = 1)$ and $\Pr(Y = 0 \mid X = 1)$ using Bayes Theorem:

$$\begin{aligned}
q_1 &= \Pr(Y = 1 \mid X = 1) \\
&= \frac{\Pr(X = 1 \mid Y = 1) \Pr(Y = 1)}{\Pr(X = 1)} \\
&= \frac{\Pr(X = 1 \mid Y = 1) \Pr(Y = 1)}{\Pr(X = 1 \mid Y = 1) \Pr(Y = 1) + \Pr(X = 1 \mid Y = 0) \Pr(Y = 0)} \\
&= \frac{p_1 \times 18/100000}{p_1 \times 18/100000 + p_2 \times (1 - 18/100000)}
\end{aligned}$$

Similarly for q_0

$$\begin{aligned}
q_0 &= \Pr(Y = 1 \mid X = 0) \\
&= \frac{\Pr(X = 0 \mid Y = 1)\Pr(Y = 1)}{\Pr(X = 0)} \\
&= \frac{\Pr(X = 0 \mid Y = 1)\Pr(Y = 1)}{\Pr(X = 0 \mid Y = 1)\Pr(Y = 1) + \Pr(X = 0 \mid Y = 0)\Pr(Y = 0)} \\
&= \frac{(1 - p_1) \times 18/100000}{(1 - p_1) \times 18/100000 + (1 - p_2) \times (1 - 18/100000)}
\end{aligned}$$

We take 10,000 posterior draws of p_1, p_2, and using the above formula above obtain 10,000 samples of q_1 and q_2. Using these samples of q_1 and q_2 we summarize the posterior distribution of q_1, q_0 in the histogram in Figure 3. The posterior median of q_1 and q_0 are 0.0006 and 0.0001 respectively and 90% credible interval for q_1 is $(0.0005, 0.0007)$ and q_0 is $(9.9 \times 10^{-5}, 1.2 \times 10^{-4})$.

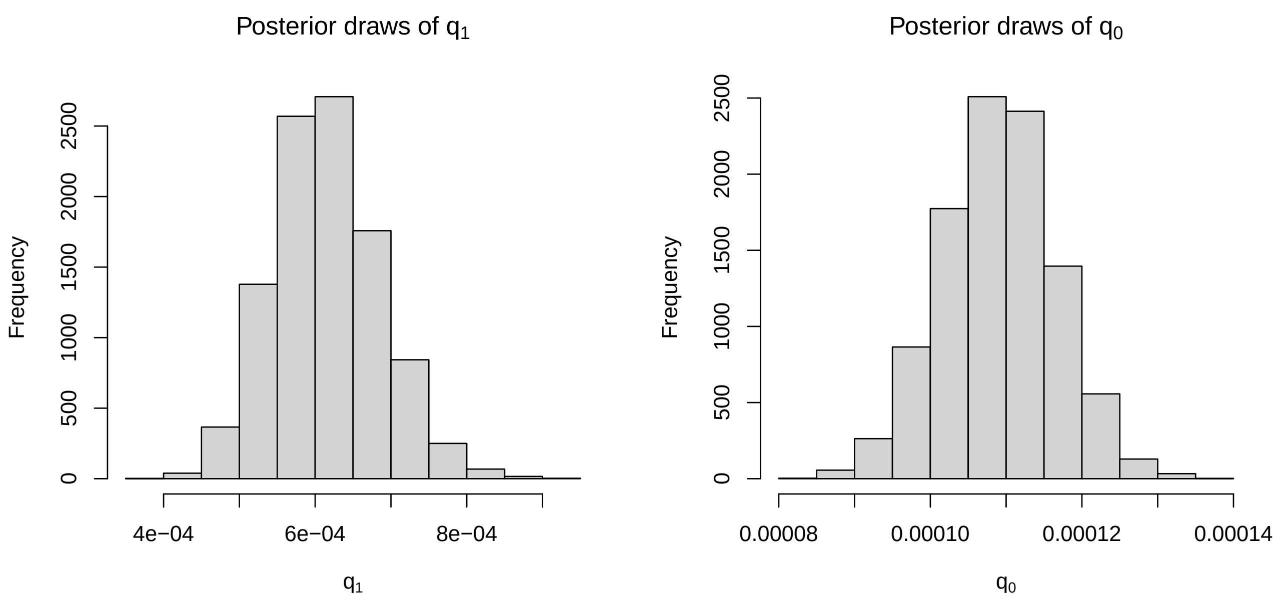


Figure 3: Histogram of posterior distribution of q_0 and q_1 for $a = b = 1$.

(i) Suppose that *a priori* you would like to select a Be(a, b) distribution on the rate of esophageal cancer with 5% of the mass less than 16 in 100,000 and 5% of the mass greater than 20 in 100,000. Find a and b to satisfy these requirements, and

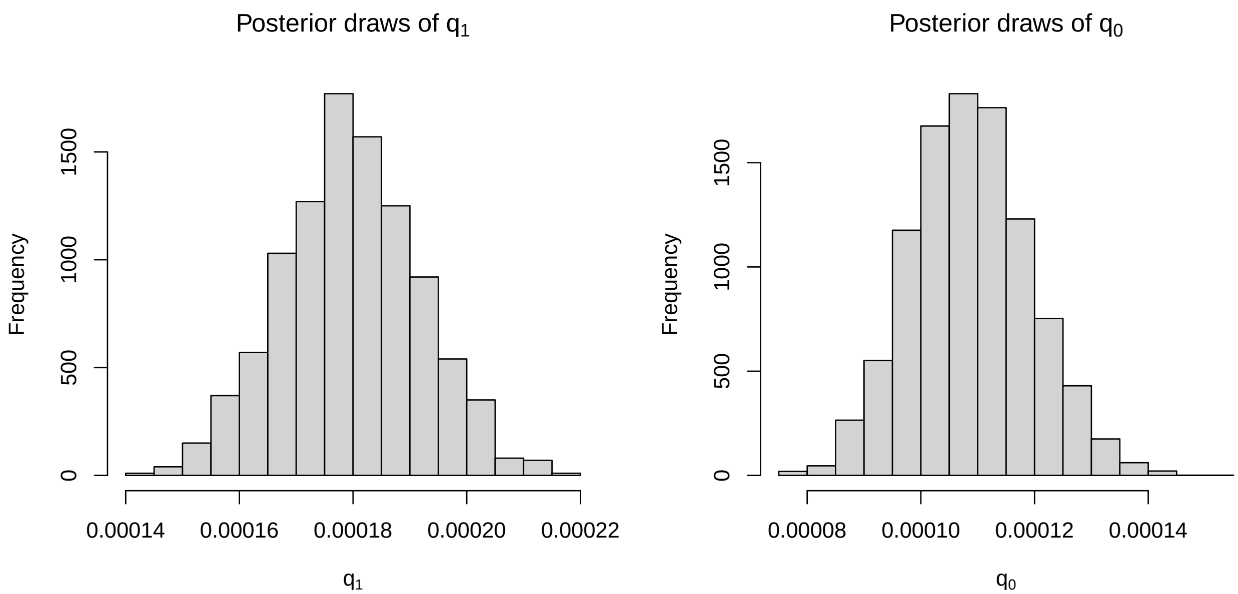


Figure 4: Histogram of posterior distribution of q_0 and q_1 for chosen a and b.

hence obtain samples from the posteriors for q_1 and q_0. **Solution:** We model $C \sim Beta(a, b)$ such that $P(C < 16/100000) = 0.05$ and $P(C > 20/100000) = 0.05$. We can find a, b by solving $[P(C < 16/100000 - 0.05]^2 + [P(C < 20/100000) - 0.95]^2 = 0$. The optimization routine in R found $a = 217.6115,\ b = 1211867.6981$. To examine the posteriors of q_1, q_0, we draw 10,000 samples of $C \sim Beta(217.6115, 1211867.6981)$ as well as posterior samples of p_1, p_2 from part f, using which we obtained 10,000 samples of $q_1 = \frac{Cp_1}{Cp_1+(1-C)p_2}$ and $q_0 = \frac{C(1-p_1)}{C(1-p_1)+(1-C)(1-p_2)}$. The histograms are plotted in Figure 4.

2. (a) Consider the “likelihood”, $\widehat{\theta} \mid \theta \sim$ N(θ, V) and the prior $\theta \sim$ N$(0, W)$ with V and W known. Show that $\theta \mid \widehat{\theta} \sim$ N$(r\widehat{\theta}, rV)$ where $r = W/(V + W)$.

Solution: We have

$$
\begin{aligned}
p(\theta \mid \widehat{\theta}) &\propto p(\widehat{\theta} \mid \theta) \times \pi(\theta) \\
&\propto \exp\left[-\frac{1}{2V}\left(\widehat{\theta}-\theta\right)^2\right] \times \exp\left[-\frac{1}{2W}\theta^2\right] \\
&\propto \exp\left[-\frac{1}{2V}\left(\theta^2 - 2\widehat{\theta}\theta\right)^2 - \frac{1}{2W}\theta^2\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\left(\frac{1}{V}+\frac{1}{W}\right)\theta^2 - 2\left(\frac{\widehat{\theta}}{V}\right)\theta\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\left(\frac{1}{rV}\right)\theta^2 - 2\left(\frac{\widehat{\theta}}{V}\right)\theta\right)\right] \\
&\propto \exp\left[-\frac{1}{2rV}\left(\theta^2 - 2\left(r\widehat{\theta}\right)\theta\right)\right],
\end{aligned}
$$

so we get $\theta \mid \widehat{\theta} \sim N(r\widehat{\theta}, rV)$, as needed.

(b) Suppose we wish to compare the models $M_0 : \theta = 0$ versus $M_1 : \theta \neq 0$. Show that the Bayes factor is given by

$$
\text{BF} = \frac{p(\widehat{\theta}|M_0)}{p(\widehat{\theta}|M_1)} = \frac{1}{\sqrt{1-r}}\exp\left(-\frac{Z^2}{2}r\right)
$$

where $Z = \widehat{\theta}/\sqrt{V}$.
Solution:

$$
\begin{aligned}
\text{BF} &= \frac{p(\widehat{\theta}|M_0)}{p(\widehat{\theta}|M_1)} = \frac{p(\widehat{\theta}|\theta_0)}{\int p(\widehat{\theta}|\theta)\pi(\theta)d\theta} \\
&= \frac{\frac{1}{\sqrt{2\pi V}}\exp\{-\frac{\widehat{\theta}^2}{2V}\}}{\frac{1}{2\pi\sqrt{VW}}\int \exp\{-\frac{(\widehat{\theta}-\theta)^2}{2V} - \frac{\theta^2}{2W}\}d\theta} = \frac{\frac{1}{\sqrt{2\pi V}}\exp\{-\frac{\widehat{\theta}^2}{2V}\}}{\frac{1}{2\pi\sqrt{VW}}\int \exp\{-\frac{(V+W)\theta^2 - 2W\theta\widehat{\theta} + W\widehat{\theta}^2}{2VW}\}d\theta} \\
&= \frac{\frac{1}{\sqrt{2\pi V}}\exp\{-\frac{\widehat{\theta}^2}{2V}\}}{\frac{1}{2\pi\sqrt{VW}}\int \exp\{-\frac{(\theta - \frac{W}{V+W}\widehat{\theta})^2 + \frac{W}{V+W}\widehat{\theta}^2 - r^2\widehat{\theta}^2}{2VW/(V+W)}\}d\theta} = \frac{\frac{1}{\sqrt{2\pi V}}\exp\{-\frac{\widehat{\theta}^2}{2V}\}}{\frac{\sqrt{Vr}}{\sqrt{2\pi VW}}\exp\{-\frac{r\widehat{\theta}^2 - r^2\widehat{\theta}^2}{2VW/(V+W)}\}d\theta} \\
&= \sqrt{\frac{W+V}{V}}\exp\{-\frac{\widehat{\theta}^2 r}{2V}\} = \frac{1}{\sqrt{1-r}}\exp\{-\frac{Z^2}{2}r\}
\end{aligned}
$$

Alternatively, we can also take use of results in part (a) to derive

$$\begin{aligned} p(\widehat{\theta}|M_1) &= \int_{M_0} p(\widehat{\theta}|\theta)\pi(\theta)d\theta = p(\widehat{\theta}) \\ &= \frac{p(\widehat{\theta}|\theta)\pi(\theta)}{p(\theta|\widehat{\theta})} \end{aligned}$$

(c) Suppose we have a prior probability $\pi_1 = \Pr(M_1)$ of model M_1 being true. Write down an expression for the posterior probability $\Pr(M_1|\widehat{\theta}_1)$, in terms of the BF.
Solution: Since

$$\frac{\Pr(M_0|\widehat{\theta})}{\Pr(M_1|\widehat{\theta})} = BF\frac{\Pr(M_0)}{\Pr(M_1)}$$

We have

$$\begin{aligned} \frac{1 - \Pr(M_1|\widehat{\theta})}{\Pr(M_1|\widehat{\theta})} &= BF\frac{1-\pi_1}{\pi_1} \\ \Rightarrow \Pr(M_1|\widehat{\theta}) &= \frac{1}{BF\frac{1-\pi_1}{\pi_1} + 1} = \frac{\pi_1}{\pi_1 + BF(1-\pi_1)} \end{aligned}$$

(d) Now suppose we have summaries from two studies, $\widehat{\theta}_j, V_j,\ j = 1, 2$. Assuming, $\widehat{\theta}_j \mid \theta \sim \mathsf{N}(\theta, V_j)$ and the prior $\theta \sim \mathsf{N}(0, W)$, derive the posterior $p(\theta|\widehat{\theta}_1, \widehat{\theta}_2)$.
Solution:

$$\begin{aligned} p(\theta|\widehat{\theta}_1, \widehat{\theta}_2) &\propto p(\widehat{\theta}_1, \widehat{\theta}_2|\theta)\pi(\theta) \\ &\propto \exp\Big\{ -\frac{(\widehat{\theta}_1 - \theta)^2}{2V_1} - \frac{(\widehat{\theta}_2 - \theta)^2}{2V_2} - \frac{\theta^2}{2W} \Big\} \\ &\propto \exp\Big\{ -\frac{1}{2V_1V_2W}\big[(V_2W + V_1W + V_1V_2)\theta - 2\theta(V_2W\widehat{\theta}_1 + V_1W\widehat{\theta}_2)\big] \Big\} \\ &\propto \exp\Bigg\{ -\frac{\Big(\theta - \frac{V_1^{-1}\widehat{\theta}_1 + V_2^{-1}\widehat{\theta}_2}{V_1^{-1} + V_2^{-2} + W^{-1}}\Big)^2}{2\Big(V_1^{-1} + V_2^{-2} + W^{-1}\Big)^{-1}} \Bigg\} \end{aligned}$$

Let $r_1 = V_1^{-1}(V_1^{-1} + V_2^{-2} + W^{-1})^{-1}$, $r_2 = V_2^{-1}(V_1^{-1} + V_2^{-2} + W^{-1})^{-1}$ and $v = (V_1^{-1} + V_2^{-2} + W^{-1})^{-1}$, then posterior distribution $p(\theta|\widehat{\theta}_1, \widehat{\theta}_2) \sim N(r_1\widehat{\theta}_1 + r_2\widehat{\theta}_2, v)$

(e) Derive the Bayes factor

$$\mathsf{BF} = \frac{p(\widehat{\theta}_1, \widehat{\theta}_2|M_0)}{p(\widehat{\theta}_1, \widehat{\theta}_2|M_1)}$$

again comparing the models $M_0 : \theta = 0$ versus $M_1 : \theta \neq 0$.
Solution: Similarly as in part (b),

$$\begin{aligned}
\text{BF} &= \frac{p(\widehat{\theta}_1, \widehat{\theta}_2|M_0)}{p(\widehat{\theta}_1, \widehat{\theta}_2|M_1)} = \frac{p(\widehat{\theta}_1, \widehat{\theta}_2|\theta_0)}{\int_{M_1} p(\widehat{\theta}_1, \widehat{\theta}_2|\theta)\pi(\theta)d\theta} \\
&= \frac{\frac{1}{2\pi\sqrt{V_1V_2}} \exp\Big\{ -\frac{\widehat{\theta}_1^{\,2}}{2V_1} - \frac{\widehat{\theta}_2^{\,2}}{2V_2} \Big\}}{\frac{1}{2\pi\sqrt{2\pi V_1V_2W}} \int \exp\Big\{ -\frac{(\widehat{\theta}_1-\theta)^2}{2V_1} - \frac{(\widehat{\theta}_2-\theta)^2}{2V_2} - \frac{\theta^2}{2W} \Big\} d\theta} \\
&= \frac{\frac{1}{2\pi\sqrt{V_1V_2}} \exp\Big\{ -\frac{\widehat{\theta}_1^{\,2}}{2V_1} - \frac{\widehat{\theta}_2^{\,2}}{2V_2} \Big\}}{\frac{1}{2\pi\sqrt{2\pi V_1V_2W}} \int \exp\Big\{ -\frac{\big(\theta-(r_1\widehat{\theta}_1+r_2\widehat{\theta}_2)\big)^2 + r_1\widehat{\theta}_1^{\,2} + r_2\widehat{\theta}_2^{\,2} - (r_1\widehat{\theta}_1+r_2\widehat{\theta}_2)^2}{2v} \Big\} d\theta} \\
&= \frac{\frac{1}{2\pi\sqrt{V_1V_2}} \exp\Big\{ -\frac{\widehat{\theta}_1^{\,2}}{2V_1} - \frac{\widehat{\theta}_2^{\,2}}{2V_2} \Big\}}{\frac{1}{2\pi\sqrt{V_1V_2W}} v^{-1/2} \exp\Big\{ -\frac{r_1\widehat{\theta}_1^{\,2} + r_2\widehat{\theta}_2^{\,2} - (r_1\widehat{\theta}_1+r_2\widehat{\theta}_2)^2}{2v} \Big\}} \\
&= \sqrt{\frac{W}{v}} \exp\Big\{ -\frac{(r_1\widehat{\theta}_1 + r_2\widehat{\theta}_2)^2}{2v} \Big\} \\
&= \sqrt{W(V_1^{-1} + V_2^{-2} + W^{-1})} \exp\Big\{ -\frac{(V_1^{-1}\widehat{\theta}_1 + V_2^{-1}\widehat{\theta}_2)^2}{2(V_1^{-1} + V_2^{-2} + W^{-1})^{-1}} \Big\}
\end{aligned}$$

We will show these results can be used in the context of a genome-wide association study on Type II diabetes, reported bu Frayling et al. (2007, Science). Two sets of data were independently collected, resulting in two log odds ratios $\widehat{\theta}_j$, $j = 1, 2$, for each SNP. For SNP rs9939609 point estimates of the odds ratio (95% confidence intervals) were 1.27 (1.16, 1.37) and 1.15 (1.09,1.23). Suppose we have a normal prior for the log odds ratio that has a 95% range [log(2/3), log(3/2)].

(f) Find W from this interval, and then calculate the posterior median and 95% intervals for θ based on (i) the first dataset only, (ii) both of the populations.
Solution: Given $\pi(\theta) \sim N(0, W)$, and let Φ to be standard normal distribution function, we have

$$\begin{aligned}
\log(3/2) &= \sqrt{W}\Phi^{-1}(0.975) \\
\Rightarrow W &= \Big(\frac{\log(3/2)}{\Phi^{-1}(0.975)}\Big)^2 \approx 0.0428
\end{aligned}$$

Similarly we calculate V_1 and V_2 by $V_j = \Big(\frac{\log CI_u - \log CI_l}{\Phi^{-1}(0.975) - \Phi^{-1}(0.025)}\Big)^2$. According to the part (a) and (d), posterior distributions given first sample and both samples are

$\theta|\widehat{\theta}_1 \sim N(r\widehat{\theta}_1, rV_1)$ and $\theta|\widehat{\theta}_1, \widehat{\theta}_2 \sim N(r_1\widehat{\theta}_1 + r_2\widehat{\theta}_2, v)$. Our result of posterior medians, credible intervals are in the following table:

	median	CL_l	CL_u
one set	0.2294	0.1479	0.3109
two sets	0.1715	0.1230	0.2201

Table 2: Posterior median and 95% credible intervals for θ

(g) Calculate the Bayes factor based on the first dataset only, and then based on both datasets.
Solution: According to part b and e, we calculate Bayes facors in following table:

	BF
one set	1.2299e-06
two sets	3.1764e-10

Table 3: Bayes factors

(h) With a prior of $\pi_1 = 1/5000$, calculate the probabilities, $\Pr(M_1|\widehat{\theta}_1)$ and $\Pr(M_1|\widehat{\theta}_1, \widehat{\theta}_2)$
Solution: According to results of part c, we calculate probabilities as following:

$$\Pr(M_1|\widehat{\theta}_1) = 0.9938892$$
$$\Pr(M_1|\widehat{\theta}_1, \widehat{\theta}_2) = 0.9999984$$

3. We will carry out a Bayesian analysis of the lung cancer and radon data, that were examined in lectures, using INLA. These data are available on the class website.

   The likelihood is

$$Y_i \mid \boldsymbol{\beta} \sim_{ind} \text{Poisson}\left[\, E_i \exp(\beta_0 + \beta_1 x_i) \,\right],$$

   where $\boldsymbol{\beta} = [\beta_0, \beta_1]^{\mathsf{T}}$, Y_i and E_i are observed and expected counts of lung cancer incidence in Minnesota in 1998–2002, and x_i is a measure of residential radon in county i, $i = 1, \ldots, n$.

   (a) Analyze these data using the default prior specifications in INLA. Produce figures of the INLA approximations to the marginal distributions of β_0 and β_1, along with the posterior means, posterior standard deviations, and 2.5%, 50%, 97.5% quantiles.
   Solution: We download the data from the textbook's website:

   - Lung cancer counts (observed and expected):
     http://faculty.washington.edu/jonno/book/MNlung.txt

- Measures of residential radon:
  http://faculty.washington.edu/jonno/book/MNradon.txt

Observed and expected counts are presented for males and females separately. In this question, we're interested in the total counts so we added the sex-specific counts ($Y_i = \texttt{obs.M} + \texttt{obs.F}$ and $E_i = \texttt{exp.M} + \texttt{exp.F}$). Multiple radon measures are available for each county, so we used the average of these as our covariate x_i. We then analysed the processed radon data using the default prior specifications in INLA. The marginal distributions of β_0 and β_1 are shown in figure 5. Summaries of the posterior distribution are shown in table 4.

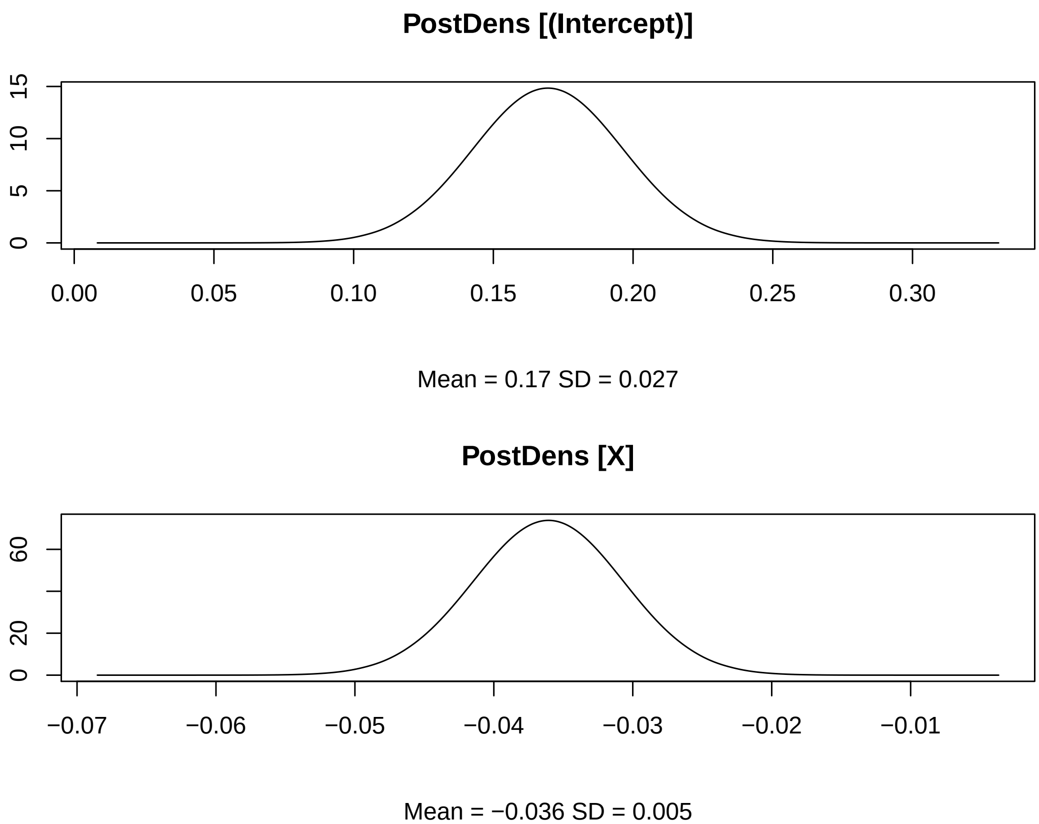


Figure 5: From top to bottom: posterior marginal distributions of β_0 and β_1 using default priors and built in plot() function from INLA package.

(b) For a more informative prior specification we may reparameterize the model as

$$Y_i \mid \boldsymbol{\theta} \sim_{ind} \text{Poisson}\left(E_i \theta_0 \theta_1^{x_i - \overline{x}}\right),$$

			Quantiles		
	Mean	Std Dev	2.5%	50%	97.5%
β_0	0.17	0.023	0.11	0.17	0.22
β_1	-0.036	0.005	-0.04	-0.36	-0.02

Table 4: Results using default priors from INLA package

where $\boldsymbol{\theta} = [\theta_0, \theta_1]^\intercal$ where

$$\theta_0 = \mathsf{E}[Y/E \mid x = \overline{x}] = \exp(\beta_0 + \beta_1 \overline{x})$$

is the expected standardized mortality ratio in an area with average radon. The parameter $\theta_1 = \exp(\beta_1)$ is the relative risk associated with a one-unit increase in radon.

For θ_0 we assume a lognormal prior with 2.5% and 97.5% quantiles of 0.67 and 1.5 to give $\mu = 0, \sigma = 0.21$. For θ_1 we again take a lognormal prior and assume the relative risk associated with a one-unit increase in radon is between 0.8 and 1.2 with probability 0.95, to give $\mu = -0.02, \sigma = 0.10$. By converting these into normal priors in INLA, rerun your analysis, and report the same summaries. **Solution:**

Our priors for θ_0, θ_1 are $\theta_0 \sim LogN(0, 0.21^2)$ and $\theta_1 \sim LogN(-0.02, 0.10^2)$. Since INLA is restricted to models with Gaussian priors, we can equivalently specify that $\log\theta_0 \sim N(0, 0.21^2)$ and $\log\theta_1 \sim N(-0.02, 0.10^2)$. Then, we have

$$\begin{aligned} Y_i|\theta &\sim \text{Poisson}(E_i \exp(\log\theta_0)\exp(\log\theta_1)^{x_i - \bar{x}}) \\ &\equiv \text{Poisson}(E_i \exp[\log\theta_0 + \log\theta_1 x_i^*]),\ \text{where } x_i^* = x_i - \bar{x} \end{aligned} \tag{1}$$

or

$$\begin{aligned} Y_i|\theta &\sim \text{Poisson}(E_i \exp(\log\theta_0)\exp(\log\theta_1)^{x_i - \bar{x}}) \\ &\equiv \text{Poisson}(E_i \exp[(\log\theta_0 - \log\theta_1\bar{x}) + \log\theta_1 x_i]), \\ &\quad \text{where} \log\theta_0 - \log\theta_1\bar{x} = \beta_0 \text{ from part (a) and } \log\theta_1 = \beta_1 \end{aligned} \tag{2}$$

If we fit model (1) (with priors $N(0, 0.21^2)$ and $N(-0.02, 0.10^2)$ on the intercept and slope, respectively), then the `summary.inla` give us inference on $\log\theta_0$ and $\log\theta_1$, noting that $\log\theta_1$ has the same interpretation as β_1 in part (a), but $\log\theta_0$ now has a different interpretation than β_0 in part (a). Using model (1), we might also be interested in inference on θ_0, θ_1 instead of $\log\theta_0, \log\theta_1$, since θ_0, θ_1 have more meaningful scientific interpretations. To get posterior quantiles we can just exponentiate quantiles for $\log\theta$. To get the posterior mean and standard deviation we cannot exponentiate the posterior mean and sd for $\log\theta$ (because $E[\exp(x)] \neq \exp(E[x])$). Instead, we can get samples from the posteriors for $\log\theta_0, \log\theta_1$ using `inla.rmarginal` and exponentiate those samples to get samples from the posteriors for θ_0, θ_1. Then we can use these samples to get the posterior mean and standard deviation.

The posterior mean, sd, and quantiles for $\log\theta_0, \log\theta_1$ using model (1) are:

			Quantiles		
	Mean	Std Dev	2.5%	50%	97.5%
$\log\theta_0$	-0.021	0.009	-0.040	-0.021	-0.003
$\log\theta_1$	-0.036	0.005	-0.047	-0.036	-0.025

Table 5: Results using informative priors and under model (1) parametrization.

The posterior mean (sd) for θ_0, θ_1, based on 10000 samples from the posterior are 0.98 (0.01) and 0.96 (0.01) respectively.

If we fit model (2) instead, then we get inference on the same β_0 and β_1 from part (a). The prior for the slope, $\log\theta_1$ is just $N(-0.02, 0.10^2)$. The prior for the intercept, $\log\theta_0 - \log\theta_1\bar{x}$ is $N([0 - \bar{x}(-0.02)], [0.21^2 + \bar{x}^2 0.10^2])$, assuming we've specified independent priors on θ_0, θ_1. Now we're fitting the same model as in part (a) except this time we're not using the default priors on β_0, β_1. We can then use the same approach as in part (a) to get inference on β_0, β_1.

The posterior mean, sd, and quantiles for β_0, β_1 using model (2) are:

			Quantiles		
	Mean	Std Dev	2.5%	50%	97.5%
β_0	0.169	0.027	0.116	0.168	0.221
β_1	-0.036	0.005	-0.047	-0.036	-0.025

Table 6: Results using informative priors and under model (2) parametrization.

Appendix

```
##################
### Question 1 ###
##################
##################
### Question 1 ###
##################
p1 <- 96/200
```

```
p2 <- 109/775
or <- (p1/(1-p1))/(p2/(1-p2))

x1 <- c(rep(1,96),rep(0,104))
x2 <- c(rep(1,109),rep(0,666))
X = c(x1,x2)
Y <- c(rep(1,200), rep(0, 775))
dat <- cbind(X,Y)

fm <- glm( X ~ factor(Y), family=binomial) #get MLE
b <- fm$coef[2]
sd <- sqrt(diag(vcov(fm))[2])
exp(b + c(qnorm((1-0.90)/2),0,-qnorm((1-0.90)/2)) %o% sd)


# posterior summaries for beta prior
a <- b <- 1

n1 <- length(x1)
n2 <- length(x2)

(a+sum(x1))/(a+b+n1) #mean p1
(a+sum(x2))/(a+b+n2) #mean p2

(a+sum(x1)-1)/(a+b+n1-2) # mode p1
(a+sum(x2)-1)/(a+b+n2-2) # mode p2

sqrt((a+sum(x1))*(b+n1-sum(x1))/((a+b+n1)^2*(a+b+n1+1))) # sd p1
sqrt((a+sum(x2))*(b+n2-sum(x2))/((a+b+n2)^2*(a+b+n2+1))) # sd p2

nsamples<- 10000
p1t <- rbeta(nsamples, a , b)
p2t <- rbeta(nsamples, a , b)
thetat <- (p1t/(1-p1t))/(p2t/(1-p2t))
hist(thetat[thetat <= 10], main = "Histogram of simulated prior distribution of
theta (theta <= 10)") # prior distribution for theta
print("90% prior credible interval for theta")
quantile(thetat, c(0.05, 0.95)) # prior credible interval of theta

# Asymptotic normality
```

```
c(qnorm(0.05, mean = p1, sd = sqrt(p1 * (1-p1) / length(x1))),
  qnorm(0.95, mean = p1, sd = sqrt(p1 * (1-p1) / length(x1))))

c(qnorm(0.05, mean = p1, sd = sqrt((p2 * (1-p2) / length(x1)))),
  qnorm(0.95, mean = p2, sd = sqrt(p2 * (1-p2) / length(x2))))

# histograms of posterior

set.seed(1)
post.p1 <- rbeta(n=10000,a + sum(x1), b + n1- sum(x1))
post.p2 <- rbeta(n=10000,a + sum(x2), b + n2 - sum(x2))

par(mfrow = c(1, 2))
hist(post.p1, main = expression(paste('Posterior draws of ', p[1])),
     xlab = expression(p[1]))
hist(post.p2, main = expression(paste('Posterior draws of ', p[2])),
     xlab = expression(p[2]))

quantile(post.p1, p = c(0.10, 0.90))
quantile(post.p2, p = c(0.10, 0.90))

par(mfrow = c(1, 1))
post.theta <- (post.p1/(1-post.p1))/(post.p2/(1-post.p2))
hist(post.theta, main = expression(paste('Posterior draws of ', theta)),
     xlab = expression(theta))

round(quantile(post.theta, prob = c(0.05, 0.5, 0.95)),2)

set.seed(1)
r <- 18/100000
post.p1 <- rbeta(n=10000,a + sum(x1), b + n1- sum(x1))
post.p2 <- rbeta(n=10000,a + sum(x2), b + n2 - sum(x2))
post.q1 <- (post.p1*r)/(post.p1*r+post.p2*(1-r))
post.q0 <- ((1-post.p1)*r)/((1-post.p1)*r+(1-post.p2)*(1-r))

par(mfrow = c(1, 2))
hist(post.q1, main = expression(paste('Posterior draws of ', q[1])),
     xlab = expression(q[1]))
hist(post.q0, main = expression(paste('Posterior draws of ', q[0])),
```

```
     xlab = expression(q[0]))

quantile(post.q1, p = c(0.10, 0.90))
quantile(post.q0, p = c(0.10, 0.90))

median(post.q1)
median(post.q0)

compute_ab <- function(params) {
  a <- params[1]
  b <- params[2]
  f <- sum((pbeta(c(16/100000, 20/100000), a, b) - c(0.05, 0.95))^2)
  return(f)
}
optimal_params <- optim(par = c(1, 1), fn = compute_ab)$par
c_samples <- rbeta(n=1000, optimal_params[1], optimal_params[2])
q1_samples <- c_samples*post.p1/(c_samples*post.p1 + (1-c_samples)*post.p1)
q0_samples <- c_samples*(1 - post.p1)/(c_samples*(1-post.p1) + (1-c_samples)*(1-post.p2))
par(mfrow = c(1, 2))
hist(q1_samples, xlab = expression(q[1]), ylab = 'Frequency', main = expression(paste('Poste
hist(q0_samples, xlab = expression(q[0]), ylab = 'Frequency', main = expression(paste('Poste

##################
### Question 2 ###
##################
tbl <- cbind(V1=c(1.27,1.16,1.37), V2=c(1.15, 1.09, 1.23
                                        ))
tbl1 <- tbl[-1,]
theta <- log(tbl[1,])
W <- (log(3/2)/qnorm(0.975))^2
V <- ((log(tbl1[2,]) - log(tbl1[1,])) / (2*qnorm(0.975)))^2
r <- W/(W+V[1])
v <- 1/(sum(1/V) + 1/W)
r12 <- 1/V*v
tbl.rst <- rbind(
qnorm(c(0.5, 0.025,0.975),mean=r*theta[1],sd=sqrt(r*V[1])),
qnorm(c(0.5,0.025,0.975),mean=sum(r12*theta),sd=sqrt(v)))
colnames(tbl.rst) <- c("median", "CL_l", "CL_u")
rownames(tbl.rst) <- c("one_set", "two_sets")
xtable(tbl.rst,digits = 4)
```

```
z <- theta[1]/sqrt(V[1])
bf1 <- 1/sqrt(1-r)*exp(-z^2/2*r)
bf2 <- sqrt(W/v)*exp(-(sum(r12*theta))^2/(2*v))
tbl.rst2 <- rbind(bf1,bf2)
colnames(tbl.rst2) <- "BF"
xtable(tbl.rst2, display = c("g","g"), digits = 5,math.style.exponents = TRUE)
bf <- c(bf1,bf2)

pi1 <- 1/5000
Prb <- pi1 / (pi1 + bf*(1-pi1))

##################
### Question 3 ###
##################
library(data.table)
lung <- as.data.frame(fread('http://faculty.washington.edu/jonno/book/MNlung.txt'))
radon <- as.data.frame(fread('http://faculty.washington.edu/jonno/book/MNradon.txt'))

# formatting: use code from Jon's website
# http://faculty.washington.edu/jonno/book/bayesian.R
Obs <- apply(cbind(lung[,3], lung[,5]), 1, sum) # add male and female observed
Exp <- apply(cbind(lung[,4], lung[,6]), 1, sum) # add male and female expected
rad.avg <- rep(0, nrow(lung))
for(i in 1:nrow(lung)) {
  rad.avg[i] <- mean(radon[radon$county==i,2]) # get average radon for each county
}
x <- rad.avg
which(!(1:87 %in% radon$county)) # check if we have radon info for all counties
# 26 63
rad.avg[26]<-0 # county with no radon info
rad.avg[63]<-0 # county with no radon info
x[26] <- NA
x[63] <- NA
newy <- Obs[is.na(x)==F] # exclude counties 26 and 63
newx <- x[is.na(x)==F]
newE <- Exp[is.na(x)==F]

# install.packages('INLA',repos='http://www.math.ntnu.no/inla/R/stable')
library(INLA)
dat <- as.data.frame(cbind(newy,newx,newE))
```

```
mod <- inla(newy ~ newx, data = dat, family = "poisson", E=newE)

mod$summary.fixed

# posterior mean, sd, quantiles
#                   mean         sd  0.025quant    0.5quant
#(Intercept)  0.16955218 0.02687972  0.11682239  0.16953610
#newx        -0.03610208 0.00540624 -0.04675376 -0.03608973
#             0.975quant        mode          kld
#(Intercept)  0.22232277  0.16950608 1.655371e-16
#newx        -0.02552853 -0.03606447 3.092889e-15

# get mean and sd by hand
inla.emarginal(function(x) x,mod$marginals.fixed$`(Intercept)`) # mean of b0
inla.emarginal(function(x) x,mod$marginals.fixed$newx) # mean of b1
sqrt(inla.emarginal(function(x) x^2,mod$marginals.fixed$`(Intercept)`) -
       (inla.emarginal(function(x) x,mod$marginals.fixed$`(Intercept)`))^2) # sd of b0
sqrt(inla.emarginal(function(x) x^2,mod$marginals.fixed$newx) -
       (inla.emarginal(function(x) x,mod$marginals.fixed$newx))^2) # sd of b0

# another approach: get samples from marginal
s <- inla.rmarginal(1000,mod$marginals.fixed$newx)
mean(s); sd(s) # estimate of posterior mean and sd for beta1

# plot marginals using plot.inla
plot(mod)

### part b ###
library(SpatialEpi)
(t0_prior <- LogNormalPriorCh(0.67,1.5,0.025,0.975)) # theta0 prior
#$mu
#[1] 0.002493771
#
#$sigma
#[1] 0.2056014
(t1_prior <- LogNormalPriorCh(0.8,1.2,0.025,0.975)) # theta1 prior
#$mu
#[1] -0.020411
#
#$sigma
```

```
#[1] 0.1034369

# plot priors
plot(seq(0, 7, 0.1), dlnorm(seq(0, 7, 0.1), meanlog = t0_prior$mu,
                            sdlog = t0_prior$sigma), type = "l", xlab = "x",
     ylab = "LogNormal Density",main=expression(paste('Prior for ',theta[0])))
plot(seq(0, 7, 0.1), dlnorm(seq(0, 7, 0.1), meanlog = t1_prior$mu,
                            sdlog = t1_prior$sigma), type = "l", xlab = "x",
     ylab = "LogNormal Density",main=expression(paste('Prior for ',theta[1])))
plot(seq(-2, 2, 0.1), dnorm(seq(-2, 2, 0.1), mean = t0_prior$mu,
                            sd = t0_prior$sigma), type = "l", xlab = "x",
     ylab = "Normal Density",main=expression(paste('Prior for log(',theta[0],')')))
plot(seq(-2, 2, 0.1), dnorm(seq(-2, 2, 0.1), mean = t1_prior$mu,
                            sd = t1_prior$sigma), type = "l", xlab = "x",
     ylab = "Normal Density",main=expression(paste('Prior for log(',theta[1],')')))

## Option 1/Model 8: Y ~ (X_i-\bar{x})
## beta0 has different interpretation now
centerX <- newx-mean(newx)

# use hyperparams from LogNormalPriorCh
mod8a <- inla(newy ~ centerX, data = dat, family = "poisson", E=newE,
              control.fixed=list(mean.intercept=t0_prior$mu,
              # prior mean for beta0
                                 prec.intercept=1/(t0_prior$sigma^2),
                                 # prior precision for beta0
                                 mean=c(t1_prior$mu),
                                  # prior mean for beta1
                                 prec=c(1/(t1_prior$sigma^2))))
                                 # prior precision for beta1
mod8a$summary.fixed
#mean          sd  0.025quant    0.5quant
#(Intercept) -0.02136778 0.009234348 -0.03954784 -0.02135136
#centerX     -0.03604975 0.005397784 -0.04668456 -0.03603750
#0.975quant        mode          kld
#(Intercept) -0.003295339 -0.02131759 1.920419e-15
#centerX     -0.025492554 -0.03601242 3.043578e-15

# plot marginals of \log\theta_0 and \log\theta_1
plot(mod8a)
```

```
# to get inference on \theta_0, \theta_1: use samples from posterior
samp_log0 <- inla.rmarginal(10000,mod8a$marginals.fixed$`(Intercept)`)
samp_log1 <- inla.rmarginal(10000,mod8a$marginals.fixed$centerX)

mean(samp_log0); sd(samp_log0) # posterior mean and sd for \log\theta_0
mean(samp_log1); sd(samp_log1) # posterior mean and sd for \log\theta_1
mean(exp(samp_log0)); sd(exp(samp_log0)) # posterior mean and sd for \theta_0
mean(exp(samp_log1)); sd(exp(samp_log1)) # posterior mean and sd for \theta_0

# plot posterior marginals
par(mfrow=c(2,1))
#hist(samp_log0,xlab=expression(paste('log(',theta[0],')')),main='Posterior Samples')
hist(exp(samp_log0),xlab=expression(theta[0]),main='10k Posterior Samples')
hist(exp(samp_log1),xlab=expression(theta[1]),main='10k Posterior Samples')
par(mfrow=c(1,1))

# use rounded hyperparams from problem statement; get slightly different answers
mod8b<- inla(newy ~ centerX, data = dat, family = "poisson", E=newE,
             control.fixed=list(mean.intercept=0,
                                prec.intercept=1/(0.21^2),
                                mean=c(-0.02),
                                prec=c(1/(0.10^2))))
mod8b$summary.fixed
#mean          sd  0.025quant    0.5quant
#(Intercept) -0.02137216 0.009234642 -0.03955280 -0.02135574
#centerX     -0.03604693 0.005397294 -0.04668076 -0.03603468
#0.975quant        mode          kld
#(Intercept) -0.003299144 -0.02132197 1.911392e-15
#centerX     -0.025490686 -0.03600961 3.048489e-15

## Option 2/Model 9: same interpretation as mod1
## speicfy independent priors for theta0 and theta1
m0 <- t0_prior$mu-t1_prior$mu*mean(newx)
sig20 <- t0_prior$sigma^2+(mean(newx)^2)*t1_prior$sigma^2

mod9a <- inla(newy ~ newx, data = dat, family = "poisson", E=newE,
              control.fixed=list(mean.intercept=m0,
                                 prec.intercept=1/(sig20),
                                 mean=c(t1_prior$mu),
```

```
                                 prec=c(1/(t1_prior$sigma^2))))
mod9a$summary.fixed
#mean           sd  0.025quant   0.5quant
#(Intercept)  0.16922669 0.026817730  0.11661785  0.1692109
#newx        -0.03603575 0.005393379 -0.04666189 -0.0360235
#0.975quant        mode          kld
#(Intercept)  0.22187499  0.16918137 1.731489e-16
#newx        -0.02548718 -0.03599845 3.065613e-15

# plot marginals
plot(mod9a)

mod9b <- inla(newy ~ newx, data = dat, family = "poisson", E=newE,
              control.fixed=list(mean.intercept=0+0.02*mean(newx),
                                 prec.intercept=1/(0.21^2+0.10^2*mean(newx)),
                                 mean=c(-0.02),
                                 prec=c(1/(0.10^2))))
mod9b$summary.fixed
#mean           sd  0.025quant   0.5quant
#(Intercept)  0.16886224 0.026743924  0.11639781  0.1688466
#newx        -0.03596607 0.005379751 -0.04656519 -0.0359539
#0.975quant        mode          kld
#(Intercept)  0.22136529  0.16881745 1.403543e-16
#newx        -0.02544401 -0.03592901 3.032177e-15
```