Coursework 5: STAT 570

Philip Pham

November 1, 2018

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 $Y = 0$ and $Y = 1$ corresponding to without/with disease and alcohol consumption is represented by X with $X = 0$ and $X = 1$ denoting less than 80g and greater than or equal to 80g on average per day. Let the probabilities of high alcohol consumption in the cases and controls be denoted

$$p_1 = \mathbb{P}(X = 1 \mid Y = 1) \text{ and } p_2 = \mathbb{P}(X = 1 \mid Y = 0), \tag{1}$$

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 $X_i \mid p_i \sim \text{Binomial}(n_i, p_i)$ in the case $(i = 1)$ and control $(i = 2)$ groups.

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

$$\theta = \frac{\mathbb{P}(Y = 1 \mid X = 1)/\mathbb{P}(Y = 0 \mid X = 1)}{\mathbb{P}(Y = 1 \mid X = 0)/\mathbb{P}(Y = 0 \mid X = 0)}. \tag{2}$$

   Show that the odds ratio is equal to

$$\theta = \frac{\mathbb{P}(X = 1 \mid Y = 1)/\mathbb{P}(X = 0 \mid Y = 1)}{\mathbb{P}(X = 1 \mid Y = 0)/\mathbb{P}(X = 0 \mid Y = 0)} = \frac{p_1/(1 - p_1)}{p_2/(1 - p_2)}. \tag{3}$$

   Solution: We have that

$$\mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(X = x \mid Y = y)\,\mathbb{P}(Y = y)}{\mathbb{P}(X = x)} \tag{4}$$

   by Bayes' rule. Applying Equation 4 to Equation 2, we get

$$\theta = \frac{[\mathbb{P}(X = 1 \mid Y = 1)\,\mathbb{P}(Y = 1)]\,/\,[\mathbb{P}(X = 0 \mid Y = 1)\,\mathbb{P}(Y = 0)]}{[\mathbb{P}(X = 0 \mid Y = 1)\,\mathbb{P}(Y = 1)]\,/\,[\mathbb{P}(X = 0 \mid Y = 0)\,\mathbb{P}(Y = 0)]}. \tag{5}$$

   The $\mathbb{P}(Y = y)$ factors cancel and we obtain the first part of Equation 3. Using Equation 1, we substitute to obtain the second part of Equation 3.

	$X = 0$	$X = 1$	
$Y = 1$	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.

(b) Obtain the MLE and a 90% confidence interval for θ, for the data of Table 1.

Solution: The likelihood and log-likelihood functions are

$$L(p_1, p_2) = \binom{n_1}{x_1} p_1^{x_1} (1-p_1)^{n_1-x_1} + \binom{n_2}{x_2} p_2^{x_2} (1-p_2)^{n_2-x_2} \tag{6}$$

$$\begin{aligned} l(p_1, p_2) &= \log L(p_1, p_2) \\ &= \sum_{i=1}^{2} \left[\log \binom{n_i}{x_i} + x_i \log p_i + (n_i - x_i) \log (1-p_i) \right], \end{aligned}$$

so the score function is

$$S(p_1, p_2) = \nabla \log L(p_1, p_2) = \begin{pmatrix} \frac{x_1 - n_1 p_1}{p_1(1-p_1)} \\ \frac{x_2 - n_2 p_2}{p_2(1-p_2)}. \end{pmatrix} \tag{7}$$

Thus, the Fisher information is

$$I(p_1, p_2) = mathbbE\left[S(p_1, p_2)\, S(p_1, p_2)^\intercal\right] = \begin{pmatrix} \frac{n_1}{p_1(1-p_1)} & 0 \\ 0 & \frac{n_2}{p_2(1-p_2)} \end{pmatrix}. \tag{8}$$

From Equation 7, we can solve $S(\hat{p}_1, \hat{p}_2) = \mathbf{0}$ to get the MLEs $\hat{p}_1 = x_1/n_1$ and $\hat{p}_2 = x_2/n_2$. Since the MLE is invariant to reparameterization, we have the MLE for θ:

$$\boxed{\hat{\theta} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = \frac{1992}{1417} \approx 5.640.} \tag{9}$$

We estimate the confidence interval for $\log \hat{\theta}$ which works since log is a monotonic transform. Using the delta method and Equation 8, we have that

$$\begin{aligned} \mathrm{Var}\left(\log \hat{\theta}\right) &\approx \left(\nabla \log \hat{\theta}\right)^\intercal \left(I(\hat{p}_1, \hat{p}_2)\right)^{-1} \left(\nabla \log \hat{\theta}\right) \\ &= \begin{pmatrix} \frac{1}{\hat{p}_1(1-\hat{p}_1)} & \frac{1}{\hat{p}_2(1-\hat{p}_2)} \end{pmatrix} \begin{pmatrix} \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} & 0 \\ 0 & \frac{\hat{p}_2(1-\hat{p}_2)}{n_2} \end{pmatrix} \begin{pmatrix} \frac{1}{\hat{p}_1(1-\hat{p}_1)} \\ \frac{1}{\hat{p}_2(1-\hat{p}_2)} \end{pmatrix} \\ &= \frac{1}{n_1 \hat{p}_1 (1-\hat{p}_1)} + \frac{1}{n_2 \hat{p}_2 (1-\hat{p}_2)} \\ &= \frac{1}{n_1 \hat{p}_1} + \frac{1}{n_1 (1-\hat{p}_1)} + \frac{1}{n_2 \hat{p}_2} + \frac{1}{n_2 (1-\hat{p}_2)}. \end{aligned} \tag{10}$$

Numerically, this is $\mathrm{Var}\left(\log \hat{\theta}\right) \approx 0.0307$.

The 90% confidence interval for $\log \hat{\theta}$ is approximately

$$\left(\log \hat{\theta} - \Phi^{-1}(0.95) \sqrt{\mathrm{Var}\left(\log \hat{\theta}\right)}, \log \hat{\theta} + \Phi^{-1}(0.95) \sqrt{\mathrm{Var}\left(\log \hat{\theta}\right)} \right), \tag{11}$$

which is about $(1.441, 2.018)$. Taking the exponent of both sides, we have a 90% confidence interval for $\hat{\theta}$ of $\boxed{(4.228, 7.524)\,.}$

(c) We now consider a Bayesian analysis. Assume that the prior distribution for p_i is the beta distribution Beta (a, b) for $i = 1, 2$. Show that the posterior distribution $p_i \mid x_i$ is given by the beta distribution Beta $(a + x_i, b + n_i - x_i)$, $i = 1, 2$.

Solution: From Equation 6, we have that the posterior:

$$\begin{aligned} p(p_i \mid X_i = x_i) &\propto \mathbb{P}(X_i = x_i \mid p_i)\, p(p_i) \\ &\propto p_i^{x_i+a-1} (1-p_i)^{n_i-x_i+b-1}. \end{aligned}$$

Integration from 0 to 1, we have the beta fnction, so

$$p(p_i \mid X_i = x_i) = \frac{\Gamma(a+x_i+b+n_i-x_i)}{\Gamma(a+x_i)\,\Gamma(b+n_i-x_i)} p_i^{a+x_i-1} (1-p_i)^{b+n_i-x_i-1}, \quad (12)$$

which is the Beta $(a+x_i, b+n_i-x_i)$ distribution.

(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: For $a = b = 1$, we have that $p_1 \mid x_1 \sim \text{Beta}(97, 105)$ and $p_2 \mid x_2 \sim \text{Beta}(110, 667)$.
For the posterior means, we have that $\mathbb{E}[p_1 \mid x_1] = 97/202$ and $\mathbb{E}[p_2 \mid x_2] = 110/777$.
The mode of a Beta (α, β) distributed random variable is $\frac{\alpha-1}{\alpha+\beta-2}$. So, for the posterior modes, we have that $\text{mode}(p_1 \mid x_1) = 12/25$ and $\text{mode}(p_2 \mid x_2) = 109/775$.
The variance of a Beta (α, β) distributed random variable is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
For $p_1 \mid x_1$ and $p_2 \mid x_2$, we have standard errors:

$$\begin{aligned} \sigma_{p_1|x_1} &= \frac{1}{202}\sqrt{\frac{10185}{203}} \approx 0.0351 \\ \sigma_{p_2|x_2} &= \frac{1}{777}\sqrt{\frac{36685}{389}} \approx 0.0125. \end{aligned}$$

For the 90% credible interval, I choose l and u such that $\mathbb{P}([l,u]) = 0.9$, $\mathbb{P}((-\infty, l)) = 0.05$ and $\mathbb{P}((u,\infty)) = 0.05$. This is called the *equal-tailed interval*.
For $p_1 \mid x_1$, the interval is $[0.4226, 0.5380]$. For $p_2 \mid x_2$, the inverval is $[0.1215, 0.1626]$ This is computed numerically with `scipy.stats.beta.interval` in `case_control.ipynb`.

(e) Obtain the asymptotic form of the posterior distribution and obtain 90% credible intervals for p_1 and p_2. Compare this interval with the exact calculation of the previous part.

Solution:

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

Solution: This result follows from the conjugacy of the normal distribution with

itself:

$$\begin{aligned} p\left(\theta \mid \hat{\theta}\right) &\propto p\left(\hat{\theta} \mid \theta\right) p\left(\theta\right) \\ &\propto \exp\left(-\frac{1}{2V}\left(\hat{\theta}-\theta\right)^2 - \frac{1}{2W}\theta^2\right) \\ &\propto \exp\left(-\frac{V+W}{2\left(VW\right)}\left(\frac{W}{V+W}\hat{\theta}^2 - 2\frac{W}{V+W}\hat{\theta}\theta + \theta^2\right)\right) \\ &\propto \exp\left(-\frac{V+W}{2\left(VW\right)}\left(\theta - \frac{W}{V+W}\hat{\theta}\right)^2\right) = \exp\left(-\frac{1}{2\left(rV\right)}\left(\theta - r\hat{\theta}\right)^2\right) \end{aligned}$$

after completing the square. We recognize this distribution as being part of the normal family, which gives us the result.

(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

$$\frac{p\left(\hat{\theta} \mid M_0\right)}{p\left(\hat{\theta} \mid M_1\right)} = \frac{1}{\sqrt{1-r}} \exp\left(-\frac{Z^2}{2}r\right), \tag{13}$$

where $Z = \hat{\theta}/\sqrt{V}$.

Solution: We have that

$$\begin{aligned} p\left(\hat{\theta} \mid M_0\right) &= p\left(\hat{\theta} \mid \theta = 0\right) = \frac{1}{\sqrt{2\pi V}} \exp\left(-\frac{1}{2V}\hat{\theta}^2\right) \\ p\left(\hat{\theta} \mid M_1\right) &= \int_{-\infty}^{\infty} p\left(\hat{\theta} \mid \theta\right) p\left(\theta\right) \, \mathrm{d}\theta \\ &= \frac{1}{\sqrt{2\pi\left(V+W\right)}} \exp\left(-\frac{1}{2\left(V+W\right)}\hat{\theta}^2\right) \end{aligned}$$

after completing the square. Substituting into the left-hand side of Equation 13, we obtain

$$\frac{p\left(\hat{\theta} \mid M_0\right)}{p\left(\hat{\theta} \mid M_1\right)} = \sqrt{\frac{V+W}{V}} \exp\left(-\frac{1}{2} \cdot \frac{W}{V+W} \cdot \frac{\hat{\theta}^2}{V}\right) = \frac{1}{\sqrt{1-r}} \exp\left(-\frac{Z^2}{2}r\right)$$

as desired.

(c) Suppose we have a prior probability $\pi_1 = \mathbb{P}\left(M_1\right)$ of model M_1 being true. Write down an expression for the posterior probability $\mathbb{P}\left(M_1 \mid \hat{\theta}\right)$ in terms of the Bayes factor.

Solution: Let K be the Bayes factor. By applying Bayes' rule, we have that

$$\begin{aligned} \mathbb{P}\left(M_1 \mid \hat{\theta}\right) &= \frac{\mathbb{P}\left(\hat{\theta} \mid M_1\right)\mathbb{P}\left(M_1\right)}{\mathbb{P}\left(\hat{\theta} \mid M_0\right)\mathbb{P}\left(M_0\right) + \mathbb{P}\left(\hat{\theta} \mid M_1\right)\mathbb{P}\left(M_1\right)} \\ &= \frac{K^{-1}\mathbb{P}\left(\hat{\theta} \mid M_0\right)\pi_1}{\mathbb{P}\left(\hat{\theta} \mid M_0\right)\left(1-\pi_1\right) + K^{-1}\mathbb{P}\left(\hat{\theta} \mid M_0\right)\pi_1} \\ &= \frac{K^{-1}\pi_1}{\left(1-\pi_1\right) + K^{-1}\pi_1} = \frac{\pi_1}{K\left(1-\pi_1\right) + \pi_1}. \end{aligned}$$

(d) Now suppose we have summaries from two studies, θ_j, V_j, $j = 1, 2$. Assuming, $\theta_j \mid \theta \sim \mathcal{N}(\theta, V_j)$ and the prior $\theta \sim \mathcal{N}(0, W)$, derive the posterior $p(\theta \mid \theta_1, \theta_2)$.

Solution: We have

$$\begin{aligned} p(\theta \mid \theta_1, \theta_2) &\propto p(\theta_2 \mid \theta_1, \theta)\, p(\theta_1 \mid \theta)\, p(\theta) = p(\theta_2 \mid \theta)\, p(\theta_1 \mid \theta)\, p(\theta) \\ &\propto \exp\left(-\frac{1}{2V_2}(\theta_2 - \theta)^2\right) \exp\left(-\frac{V_1 + W}{2(V_1 W)}\left(\theta - \frac{W}{V_1 + W}\theta_1\right)^2\right) \\ &\propto \exp\left(-\frac{V_1V_2 + V_1W + V_2W}{2(V_1V_2W)}\left(\theta - \frac{V_2W\theta_1 + V_1W\theta_2}{V_1V_2 + V_1W + V_2W}\right)^2\right), \end{aligned}$$

after repeatedly completing the square and dropping factors that don't depend on θ.
Thus, we have that

$$\theta \mid \theta_1, \theta_2 \sim \mathcal{N}\left(\frac{V_2W\theta_1 + V_1W\theta_2}{V_1V_2 + V_1W + V_2W}, \frac{V_1V_2W}{V_1V_2 + V_1W + V_2W}\right). \tag{14}$$

(e) Derive the Bayes factor

$$\frac{p(\theta_1, \theta_2 \mid M_0)}{p(\theta_1, \theta_2 \mid M_1)}, \tag{15}$$

again comparing the models M_0: $\theta = 0$ versus M_1: $\theta \neq 0$.

Solution: (θ_1, θ_2) have a bivariate normal distribution. Under M_0, we have that

$$\begin{aligned} p(\theta_1, \theta_2 \mid M_0) &= p(\theta_1, \theta_2 \mid \theta = 0) \\ &= \frac{1}{2\pi\sqrt{V_1V_2}} \exp\left(-\frac{1}{2}\begin{pmatrix}\theta_1 & \theta_2\end{pmatrix}\begin{pmatrix}\frac{1}{V_1} & 0 \\ 0 & \frac{1}{V_2}\end{pmatrix}\begin{pmatrix}\theta_1 \\ \theta_2\end{pmatrix}\right). \end{aligned} \tag{16}$$

Under M_1, we have that

$$p(\theta_1, \theta_2 \mid M_1) = \int_{-\infty}^{\infty} p(\theta_1, \theta_2 \mid \theta)\, p(\theta)\, \mathrm{d}\theta. \tag{17}$$

We can consider θ as having the improper prior $\mathcal{N}\left(\mathbf{0}, \begin{pmatrix}W & W \\ W & W\end{pmatrix}\right)$, which results in

$$\theta_1, \theta_2 \mid M_1 \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix}V_1 + W & W \\ W & V_2 + W\end{pmatrix}\right) \tag{18}$$

by conjugacy of the multivariate normal distribution.
The Bayes factor can then be computed:

$$\sqrt{\frac{V_1V_2 + V_1W + V_2W}{V_1V_2}} \exp\left(-\frac{1}{2}\begin{pmatrix}\theta_1 & \theta_2\end{pmatrix}\Lambda\begin{pmatrix}\theta_1 \\ \theta_2\end{pmatrix}\right), \tag{19}$$

where

$$\Lambda = \begin{pmatrix}\frac{1}{V_1} & 0 \\ 0 & \frac{1}{V_2}\end{pmatrix} + \frac{1}{V_1V_2 + V_1W + V_2W}\begin{pmatrix}V_2 + W & -W \\ -W & V_1 + W\end{pmatrix}. \tag{20}$$

(f) We will show these results can be used in the context of a genome-wide association study on Type II diabetes, reported by Frayling et al. (2007, Science). Two sets of data were independently collected, resulting in two log odds ratios $\hat{\theta}_j$, $j = 1, 2$, for each SNP.

For SNP rs9939609 point estimates (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 odds ratio that has a 95% range (0.67, 1.50).

i. 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:

ii.

iii.