Multinomial model for categorical data with a conjugate prior

Generalising the binomial

↳ k possible outcomes

Let $y = (y_1, \dots, y_k)$ be a vector of observed counts of each outcome

$$n = \sum_{i=1}^{k} y_i \longleftarrow \text{the number of observations}$$

· Let θ_j be the be the probability of outcome j,

$$\sum_{j=1}^{k} \theta_j = 1$$

Then the multinomial distribution for y given θ is described by

$$p(y|\theta) \propto \prod_{i=1}^{k} \theta_j^{y_j} \quad , \quad \sum_{j=1}^{k} \theta_j = 1$$

(for $k=2$, this is the binomial with $y_1 = y$ and $y_2 = n-y$)

The conjugate prior is the Dirichlet distribution (a multivariate version of the Beta)

$$p(\theta) \propto \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \quad , \quad \theta_1, \dots, \theta_k \geq 0 \quad , \quad \sum_{i=1}^{k} \theta_i = 1$$

which results in a Dirichlet posterior distribution for θ with parameters $\alpha_j + y_j$, $j = 1, \dots, k$.

Interpretation of the hyperparameters:

$\alpha_0 = \sum_{j=1}^{k} \alpha_j$ can be thought of as the prior number of observations, and then α_j as the prior counts of outcome j.

okt 6-12:17

General setup for Bayesian inference

- Formulate the model for (y, θ), i.e.
  - the sampling distribution $p(y|\theta)$
    - ↳ Write the likelihood function for θ, ignoring factors that do not depend on θ
  - the prior distribution for θ (can be quite complex)
    - ↳ Formally, it is not allowed to include any information from the data when formulating the prior
    - ↳ Sometimes, it is done when using a method called "empirical Bayes"
- Write the posterior density up to a constant of proportionality, ignoring factors that do not depend on θ

$$p(\theta|y) \propto p(\theta) \cdot p(y|\theta)$$

- Sample from the posterior distribution
  - ↳ $p(\theta|y)$ fully known: straightforward (if θ is the only quantity of interest, no need to sample!)
  - ↳ $p(\theta|y)$ not fully known:
    - Discrete approximations (as in Ex. 11, Ch. 2)
      - ↳ Not generally recommended
    - Normal approximation
      - ↳ Need a lot of data to be justified
    - Advanced computation methods

okt 6-12:41

- Suppose you are interested ~~for the~~
  (1) the posterior distribution of a function $g(\theta)$ of the parameter vector θ
  (2) the posterior predictive distribution of $\tilde{y}$

General solution:

For $i = 1, \dots, S$ do
- Sample $\theta^{(i)}$ from $p(\theta | y)$
- Compute $g(\theta^{(i)})$
- Sample $\tilde{y}^{(i)}$ from $p(\tilde{y} | \theta^{(i)})$, the sampling distr. for $\tilde{y}$ given $\theta^{(i)}$

end
- $g(\theta^{(1)}), g(\theta^{(2)}), \dots, g(\theta^{(S)})$ are samples from the posterior distr. of $g(\theta)$
- $\tilde{y}^{(1)}, \tilde{y}^{(2)}, \dots, \tilde{y}^{(S)}$ are samples from the posterior predictive distr. of $\tilde{y}$

okt 6-12:55

Asymptotic theory and the Normal approximation

- How to approximate $p(\theta | y)$ when $n \to \infty$
- Why?
  - The approximation can be easy to use, avoiding using more advanced methods
    ↳ But if we can do it more exact, we should do it!
  - The result shows that the "prior is washed away" when $n \to \infty$, and the likelihood dominates the results. Hence, different prior specifications will give more and more similar results as n grows, and the results agree more and more with maximum likelihood inference.

Example 1

$y | \theta \sim \text{Bin}(n, \theta)$, $\theta \sim \text{Beta}(\alpha, \beta)$

then we know that

$\theta | y \sim \text{Beta}(\alpha + y, \beta + n - y) \Rightarrow E[\theta | y] = \frac{\alpha + y}{\alpha + \beta + n} \underset{\substack{n \text{ large} \\ \alpha, \beta \text{ fixed}}}{\approx} \frac{y}{n} = \theta^*$ is equal to the ML estimate for θ when $y \sim \text{Bin}(n, \theta)$

$$\text{Var}[\theta | y] = \frac{E[\theta | y] \cdot (1 - E[\theta | y])}{\alpha + \beta + n + 1} \underset{\substack{n \text{ large} \\ \alpha, \beta \text{ fixed}}}{\approx} \frac{1}{n} E[\theta | y] \cdot (1 - E[\theta | y]) \approx \frac{1}{n} \theta^* (1 - \theta^*)$$

okt 6-13:16

Frequentist approach:

MLE: $\hat{\theta} = \frac{y}{n}$, $\operatorname{Var}\hat{\theta} = \frac{1}{n}\theta_0(1-\theta_0)$, where θ_0 is the true, unknown value of θ

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \theta_0(1-\theta_0))$$

95% classical CI: $\hat{\theta} \pm \frac{1.96}{\sqrt{n}} \cdot \sqrt{\hat{\theta}(1-\hat{\theta})}$

We will show that

$$[\sqrt{n}(\theta - \theta^*) \mid data] \xrightarrow{d} N(0, \theta_0(1-\theta_0))$$ where θ_0 is the true value of θ

Example 2

$y_1, \dots, y_n \mid \theta \sim N(\theta, \sigma^2)$, $\theta \sim N(0, \sigma_0^2)$, $\sigma^2 = 1$

We have previously shown that

$$\theta \mid y \sim N(\mu_n, \tau_n^2) \underset{n \text{ large}}{\approx} N(\bar{y}, \frac{\sigma^2}{n}) \underset{\theta^* = \bar{y}}{=} N(\theta^*, \frac{1}{n}) \Rightarrow \sqrt{n}(\theta - \theta^*) \mid y \underset{n \text{ large}}{\approx} N(0,1)$$

Frequentist: $\hat{\theta}_{ML} = \bar{y} \sim N(\theta, \frac{1}{n}) \Rightarrow \sqrt{n}(\theta - \hat{\theta}) \sim N(0,1)$

okt 6-13:25

General model

$y_1, \dots, y_n \mid \theta \sim p(y \mid \theta)$, prior $p(\theta)$

vector of length p

Classical frequentist theory

$\hat{\theta}$ = maximum likelihood estimator of θ such that

(1) $\hat{\theta} \xrightarrow{p} \theta_0$ = true value of θ

(2) $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N_p(0, J(\theta_0)^{-1})$

where $J(\theta)$ is the Fisher information matrix where element (i,j) is defined as

$$J_{i,j}(\theta) = E\left[\left(\frac{\partial \log p(y|\theta)}{\partial \theta_i}\right) \cdot \left(\frac{\partial \log p(y|\theta)}{\partial \theta_j}\right) \Big| \theta\right]$$
$$= E\left[-\frac{\partial^2 \log p(y|\theta)}{\partial \theta_i \partial \theta_j} \Big| \theta\right]$$

(3) (Approximate) CI for θ_i for n large

$$\hat{\theta}_i \pm \frac{1.96}{\sqrt{n}} \sqrt{(J(\hat{\theta}))^{-1}_{ii}}$$

okt 6-13:31

Bayesian asymptotic theory

$$\theta|y \sim p_n(\theta|y) = \frac{p(\theta)\cdot p_n(y|\theta)}{\int p(\theta)\cdot p_n(y|\theta)\,d\theta}$$

, Let θ^* be the maximum likelihood estimator

The result

$$\theta|y \xrightarrow{d} N_p\left(\theta^*, \tfrac{1}{n} J(\theta_0)^{-1}\right)$$

equ. $\left(\sqrt{n}\cdot(\theta-\theta^*)\,|\,y\right) \xrightarrow{d} N_p\left(0, J(\theta_0)^{-1}\right)$

θ_0 is the true value of θ

We approximate it by replacing θ_0 by θ^*

$$\theta|y \approx N\left(\theta^*, \tfrac{1}{n} J(\theta^*)^{-1}\right)$$

okt 6-13:37