Contingency Tables (Lecture 2)

Heather Weaver & Mark Fulginiti

September 14, 2017

Bayesian Multinomial Example

Adapted from BDA3 by Gelman eta.

Election Polling Results (n =1,447)

Obama: $y_1 = 727$

Romney: $y_2 = 583$

Other / NA: $y_3 = 137$

Note: Assuming simple random sampling (i.e. ignoring any biases)

$$(y_1, y_2, y_3 \mid n) \sim Multnomial(\theta_1, \theta_2, \theta_3)$$

Estimand:

$$\theta_1 - \theta_2$$

$$H_o : \theta_1 = \theta_2$$

Prior: one choice

$$(y_1, y_2, y_3) \sim Dirichlet(\alpha_1, \alpha_2, \alpha_3)$$

$$\alpha_1 = \alpha_2 = \alpha_3 = 1$$

Posterior:

$$P(\theta_1, \theta_2, \theta_3 \mid \vec{y}\ ) \sim Dirichlet(y_1 + 1,\ y_2 + 1,\ y_3 + 1)$$

$$P(\theta_1, \theta_2, \theta_3 \mid \vec{y}\ ) \sim Dirichlet(728,\ 584,\ 138)$$

- If a closed form solution exists, we could solve for quantiles, CI bounds, etc.

Computationally:

1. Set $i = 1$
2. Draw $\vec{\theta}^{\,[\ i\ ]} \sim\ Dirichlet()$
3. Calculate $\vec{\theta}_1^{\,[\ i\ ]} -\ \vec{\theta}_2^{\,[\ i\ ]}$
4. set $i = i + 1$
5. If $i = 1,000$ skip
6. Compute summary / quantiles of $(\theta_1 - \theta_2)$

- 80% Credible Interval for $(\theta_1 - \theta_2)$ [10^{th} percentile, 90^{th} percentile]

Bayesian emample for small-sample cells

- Exact tests are a frequentest solution for low cell counts

Following example from a Wikipedia article on contingency tables, lingpipe blog (Bob Carpenter)

Gender	Left Handed	Right Handed	Total
Male	43	$y_1 = 9$	$m_1 = 52$
Female	44	$y_2 = 4$	$m_2 = 48$

- Question: Are men and women equally likely to be left-handed?

$$H_o : P(LH \mid M) = P(LH \mid F)$$
$$H_o : P(\pi_1) = P(\pi_2)$$
$$H_o : \pi_1 - \pi_2 = 0$$

We could also look at the risk ratio $\frac{\pi_1}{\pi_2}$ or the odds ratio $\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$.

$Y_1 \sim Binomial(n_1,\ \pi_1)$ the number of men who are left-handed

$Y_2 \sim Binomial(n_2,\ \pi_2)$ the number of women who are left-handed

We are assuming Independence of the above variables

Prior: Assume uniform priors

$$\pi_1 \sim Uniform(0,1) = Beta(1,1)$$
$$\pi_2 \sim Uniform(0,1) = Beta(1,1)$$

Posterior for π:

$$\pi_i \mid y_i, n_i \sim Beta(y_i + 1,\ n - y_i + 1)$$

$$\begin{pmatrix} {\pi_1}^{(1)} \\ {\pi_1}^{(2)} \\ \vdots \\ {\pi_1}^{(k)} \end{pmatrix} - \begin{pmatrix} {\pi_2}^{(1)} \\ {\pi_2}^{(2)} \\ \vdots \\ {\pi_2}^{(k)} \end{pmatrix} \implies \begin{pmatrix} (\pi_1 - \pi_2)^{(1)} \\ (\pi_1 - \pi_2)^{(2)} \\ \vdots \\ (\pi_1 - \pi_2)^{(k)} \end{pmatrix}$$

- Advantage: No assumptions about large sample size needed.
- Disadvantage: Results are sensitive to choice to prior.

Confidence intervals from likelihood ratios

Define a $(1-\alpha)\%$ credible interval/region. The set of Θ for which the Likelihood Ratio Test Statistic (LRST)

$$LRTS(\Theta) < \chi^2_{df}(1-\alpha)$$

$$LRTS(\Theta) = -2[L(\Theta) - L(\hat{\Theta})]$$

where $L(\Theta)$ is the log-likelihood from Θ, across a subset of components of Θ, and $L(\hat{\Theta})$ is the fixed log-likelihood for Θ calculated at the MLE

- $\chi^2_{df}(1-\alpha)$ is the $(1-\alpha)^{th}$ quantile of χ^2_{df}
- The df is the number of parameters varying in Θ

e.g. 1-Parameter CI

$$\Theta = \mu$$

$$\text{LRTS} = -2(L(\mu) - L(\hat{\mu}))$$

- 95% CI for $\mu = \{\ \mu : LRTS < \chi_1^2(0.95)\ \}$

Example

$$y_i \sim Poisson(\mu), \quad i = 1, 2, 3$$

$$P(y \mid \mu) = \frac{(e^{-\mu}\mu^y)}{y!}$$

$$L(\mu \mid y) = log\left(\prod_{i=1}^{3} \frac{e^{-\mu}\mu^y}{y!}\right) = \sum_{i=1}^{3} log\left(\frac{e^{-\mu}\mu^y}{y!}\right) =$$

$$= \sum(-\mu + y_i log(\mu) - log(y!)) \propto -n\mu + log(\mu) * \sum y_i + C$$

$$L(\mu \mid \vec{y}) \propto -3\mu + 7log(\mu)$$

$$LRTS = -2(L(\mu) - L(\hat{\mu})) < \chi_1^2(0.95)$$

$$\rightarrow (L(\mu) > L(\hat{\mu})) - \frac{\chi_1^2(0.95)}{2}$$

$$\frac{\chi_1^2(0.95)}{2} = \frac{3.84}{2}$$

Example with Multiple parameters

$$\Theta = (\theta_1,\ \theta_2,\ \dots,\ \theta_r,\ \theta_{r+1},\ \dots,\ \theta_p)$$

$$\text{of particular interests are}(\theta_{r+1},\ \dots,\ \theta_p)$$

$$LRST = -2\left\{L(\theta_{r+1},\ \dots,\ \theta_p \mid \hat{\theta}_1,\ \dots,\ \hat{\theta}_r) - L(\hat{\theta}_{r+1},\ \dots,\ \hat{\theta}_p \mid \hat{\theta}_1,\ \dots,\ \hat{\theta}_r)\right\}$$

Calculate the likelihood on a grid of $(\theta_{r+1},\ \dots,\ \theta_p) = \theta^{\ *}$

There are p - r parameters, so the df is p - r

95 % CI for $\theta^{\ *} = \left\{\theta^{\ *}\ : LRTS < \chi^2_{p-r}\ (0.95)\right\}$

- Note: This whole assumption is based on large sample approximation

Diagnostic Tests

Disease Presence	Test Positive	Test Negative	Total
Yes	π_{11}	π_{12}	1
No	π_{21}	π_{22}	1

- Sensitivity: Pr(+ | D) = $\pi_{1|1}$ = probability the test is positive given that you have the disease
- Specificity: Pr(- | D^C) = $\pi_{2|2}$ = probability that you test negative given that you do not have the disease

Breast Cancer Example

Cancer Presence	Test Positive	Test Negative	Total
Yes	86	14	$100 = n_1$
No	12	88	$100 = n_2$

Gold standard would be to have $\alpha = 0,\ \beta = 0$.

- Sensitivity: Pr(+ | D) = $\pi_{1|1}$ = 86%
- Specificity: Pr(- | D^C) = $\pi_{2|2}$ = 88%

From a clinical perspective we are much more interested in positive/negative predicted value.

P(D | +) sometimes = $\frac{\text{True Positive}}{\text{True Positive + False Positive}}$

e.g. P(Breast Cancer | +) = $\frac{86}{86+12}$

Note: This is only true if $(\frac{n_1}{n_1+n_2}) \approx$ P(D) = prevalence of the disease

$$P(D \mid\ +\ ) = \frac{P(D \cap +)}{P(+)} = \frac{P(\ +\ \mid D) * P(D)}{P(+ \cap D)*} * P(+) \cap D^C$$

$$= \frac{Sensetivity * Population\ Prevalance}{P(\ +\ \mid D)P(D) + P(\ +\ \mid D^C)P(D^C)}$$

$$= \frac{Sensetivity * Population\ Prevalance}{(Sensetivity * Population\ Prevalance) + ((1 - Specicifity) * (1 - Population\ Prevalance))}$$