

Inference for Proportions

September 27, 2005

Reading HH 15

Bayesian Hypothesis Testing

Lung tissue samples from n patients are allocated into two tumor types:

- normal cells ($D_i = 0$) (n_0 non-tumors)
- tumor cells ($D_i = 1$) (n_1 tumors)

For each sample, a specific protein is recorded

- present ($p_i = 1$)
- not-present ($p_i = 0$)

It is of interest to explore whether or not the presence/absence of the protein indicates whether or not the tumor is recurrent/non-recurrent.

Data Models - Retrospective Case-Control Study

For the normal cells:

$$p_i | D_i = 0, \pi_0 \stackrel{iid}{\sim} \text{Bernoulli}(\pi_0)$$

For the tumor cells:

$$p_i | D_i = 1, \pi_1 \stackrel{iid}{\sim} \text{Bernoulli}(\pi_1)$$

where D_i is the Disease status of individual i

Odds of protein presence $\omega_1 = \pi_1 / (1 - \pi_1)$ for diseased
and $\omega_0 = \pi_0 / (1 - \pi_0)$ and non-diseased

Odds Ratio ω_1 / ω_0 is informative

Null Hypothesis

- If the frequency of the protein is the same in both groups,

$$H_0 : \pi_0 = \pi_1 = \pi$$

then the presence of the protein provides no information about tumor status.

- odds ratio

$$\frac{\pi_1 / (1 - \pi_1)}{\pi_0 / (1 - \pi_0)} = 1$$

Alternative Hypothesis

- If the protein indicator occurs with a different frequency in the two groups:

$$H_1 : \pi_0 \neq \pi_1$$

then this difference may be enough to help “predict” the status of a new biopsy.

- odds ratio not equal to 1

Classical Tests

- Test for Equality of two proportions $\pi_1 = \pi_0$
`prop.test`
- Chi-squared test for independence `chisq.test`
(equivalent)
- Fisher's Exact Test `fisher.test`

See Chapter 15 of HH

Posterior Probabilities

Goal: find the posterior probability of H_0 and H_1 given the data $P = (p_1, \dots, p_n$ (and of course $D_1, \dots, D_n$)).

Using Bayes Theorem, this is

$$P(H_1|P) = \frac{p(P|H_1)p(H_1)}{p(P|H_1)p(H_1) + p(P|H_0)p(H_0)}$$

where $p(H_i)$ is the prior probability of H_i .

Bayes Factor

The Bayes factor is defined as ratio of posterior odds to prior odds

$$BF(H_1 : H_0) = \frac{p(H_1|P)/p(H_0|P)}{p(H_1)/p(H_0)} = \frac{p(P|H_1)}{p(P|H_0)}$$

which is the ratio of the marginal likelihoods of the data under the two hypotheses. Rearranging one can express the posterior probability of H_1 as a function of the Bayes Factor and prior odds (HW).

Marginal Likelihoods

The marginal likelihood of the data is the distribution of the data under the hypothesis (or model) and does not depend on any unknown parameters.

Under H_0 , the

$$p_i | \pi \sim \text{Ber}(\pi)$$

The marginal likelihood of the data is

$$p(P|H_0) = \int_0^1 \prod_{i=1}^n \pi^{p_i} (1 - \pi)^{1-p_i} p(\pi|H_0) d\pi$$

where $p(\pi|H_0)$ is the prior for the common π under H_0 .

Marginal Likelihoods

Under H_1 , we must integrate over π_1 and π_0 to obtain the marginal likelihood of the data:

$$\iint \left(\prod_{i=1}^{n_0} \pi_0^{p_i} (1 - \pi_0)^{1-p_i} \right) \left(\prod_{i=1}^{n_1} \pi_1^{p_i} (1 - \pi_1)^{1-p_i} \right) p(\pi_0, \pi_1 | H_1) d\pi_0 d\pi_1$$

where $p(\pi_0, \pi_1 | H_1)$ is the joint prior under H_1 .

Choices?

Priors

- $\pi|H_0 \sim \text{Uniform}$
- $\pi_1|H_1 \sim \text{Uniform}$
- $\pi_2|H_1 \sim \text{Uniform}$

One set of default choices that leads to marginal consistency of beliefs.

- Marginal compatibility: Marginal distribution of π_i is the same over all hypotheses.
- Conditional compatibility: Distribution of parameters under H_0 determined by distribution of parameters in H_1 and conditioning on H_0 (change of variables)

May lead to different choices! (Borel Paradox)

Marginal Likelihood Under H_0

$$\begin{aligned} p(P|H_0) &= \int_0^1 \pi^{\sum_i^n (p_i)} (1 - \pi)^{n - \sum_i^n p_i} d\pi \\ &= \int_0^1 \pi^{\sum_i^n (p_i + 1 - 1)} (1 - \pi)^{n - \sum_i^n p_i + 1 - 1} d\pi \\ &= B\left(\sum_i p_i + 1, n - \sum_i p_i + 1\right) \end{aligned}$$

This is the kernel of a Beta integral

$$\int_0^1 t^{(a-1)} (1 - t)^{(b-1)} dt \equiv B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}$$

Calculations

Results involve beta functions (ratios of gammas)

For large values, take logs of the products/ratios and express the results in terms of sums of the log of the gamma function $\log(\Gamma(x + y)/(\Gamma(x)\Gamma(y)))$

$$\text{lgamma}(x + y) - \text{lgamma}(x) - \text{lgamma}(y)$$

Exponentiate the final result to get the solution. i.e.

$$\Gamma(x + y)/(\Gamma(x)\Gamma(y))$$

$$\exp(\text{lgamma}(x+y) - \text{lgamma}(x) - \text{lgamma}(y))$$

This is more numerically stable than using the gamma function directly. In R, you can also use the `lbeta` function directly.

Model Choice

- Select Model/Hypothesis that has the largest posterior probability
- Select Model/Hypothesis if $p(H_1|P) > \lambda$ (other costs/losses associated with making an incorrect decision)
- Don't pick a hypothesis but use the mixture distribution implied by both hypotheses.
- Find $P(D^* = 1 | p^* = 1, p_1, \dots, p_n)$

Integration

Calculating $P(D^* = 1 | p^* = 1, p_1, \dots, p_n)$ involves “averaging” or integration over unknown parameters and hypotheses.

- Find $p(\pi_i | H_j, p_1, \dots, p_n)$ and $p(H_j | p_1, \dots, p_n)$
- Find $P(D^* = 1 | p^* = 1, \pi_i, H_j)$ using Bayes Theorem
- Result will be a function of π_i and H_j , say $f(\pi_i, H_j)$
- Find the posterior distribution of the function: change of variables, and integrating over hypothesis

Stochastic Integration

For a large number of times

1. Draw H_j from $p(H_j | p_1, \dots, p_n)$ (Bernoulli draw `rbinom`)
2. Given H_j , draw π_i from its posterior (Beta draw `rbeta`)
3. Evaluate $f(\pi_i, H_j)$ for the current draw
4. Repeat

Plot histogram with kernel density estimate, calculate posterior mean, probability interval or other summaries.