Significance testing in 454 pyrosequencing

David Knowles

December 10, 2008

The control population is denoted as population 1, and the treated population is denoted population 2. In population i at a particular position, we have k_i errors out of a coverage of N_i, with an underlying "true" error rate of p_i.

There is actually no need to use the Poisson approximation: we can do all the calculations with the full binomial distribution, given by

$$P(k_i|N_i, p_i) = \binom{N_i}{k_i} p_i^{k_i}(1-p_i)^{N_i-k_i}$$

(the | symbol means "given")

Under the null hypothesis, $p_1 = p_2 = p$. To calculate the distribution over p implied by the control population, we need to use Bayes' rule:

$$P(p|k_1, N_1) \propto P(k_1|p, N_1)P(p)$$

Here $P(p|k_1, N_1)$ is known as the posterior, $P(k_1|p, N_1)$ as a function of p is the likelihood function, and $P(p)$ is the prior on p. The conjugate prior for a Binomial likelihood function is the Beta distribution:

$$P(p) = \text{Beta}(p; a, b) = \frac{1}{B(a, b)} p^{a-1}(1-p)^{b-1}$$

If you are unhappy with the idea of using prior knowledge (which many biologists seems to be!) you can use the uninformative settings $a = b = 0$. My recommendation however would be to fit a, b using Type II maximum likelihood over your whole dataset. This isn't entirely straightforward - I used a Newton-Raphson gradient descent algorithm which I could provide the code for. A proper prior does help you: for example, it deals with the situation when $k_1 = 0$ which would otherwise force $p = 0$. As a quick solution, try setting $a = 1, b = \frac{1}{\bar{p}}$ where $\bar{p}$ is the overall subsitution rate you see. To give you a bit more intuition about this, you can view a, b as pseudocounts of how many errorenous vs. non-errorenous base calls you have seen.

Because of the conjugacy between the binomial and beta distributions, the posterior on p also becomes a beta distribution (I won't go through the working here, but it's not difficult):

$$P(p|k_1, N_1) = \text{Beta}(p; k_1 + a, N_1 - k_1 + b)$$

The p-value you are trying to calculate is

$$P(k \geq k_2 | k_1, N_1, N_2) = 1 - \sum_{k=0}^{k_2-1} P(k|k_1, N_1, N_2)$$

To calculate the sum we need the indivdiuals terms:

$$\begin{aligned} P(k_2|k_1, N_1, N_2) &= \int P(k_2|N_2, p) P(p|k_1, N_1) dp && (1) \\ &= \int \text{Binomial}(k_2; N_2, p) \, \text{Beta}(p; k_1 + a, N_1 - k_1 + b) dp && (2) \\ &= \tbinom{N_2}{k_2} \tfrac{\text{Beta}(k_1+k_2+a, N_1-k_1+N_2-k_2+b)}{\text{Beta}(k_1+a, N_1-k_1+b)} && (3) \end{aligned}$$

Note that the integral here is straightforward again due to the conjugacy of the binomial and beta distributions. The integrand has the functional form of a Beta distribution, which means that the whole integral is simply equal to the normalising constant of the corresponding distribution.