Statistics II

Project 2: The Good estimator

Tutor: Stefan (zohren@math.leidenuniv.nl, Room: 211)

March 16, 2009

Imagine we want to explore the plant diversity in the Amazon. Suppose that the number of different plant species is infinite and we draw a random sample of size n with replacement. By X_i we denote the frequency of the i-th species in our sample. Let p_i be the probability of the i-th species to be sampled, where $\sum_{i=1}^{\infty} p_i = 1$, then the probability for observing a sequence $x_1, x_2...$ is given by a multinomial distribution

$$P(x_1, x_2...) = \binom{n}{x_1 x_2 \dots} \prod_{i=1}^{\infty} p_i^{x_i} = \frac{n!}{(n - x_1 - x_2...)! x_1! x_2! ...} \prod_{i=1}^{\infty} p_i^{x_i}. \tag{1}$$

Let us now denote the total probability mass of unobserved species by Q_n and the total number of species observed j-times in the sample by $F_j(n)$, i.e.

$$Q_n = \sum_{i=1}^{\infty} p_i I\{X_i = 0\}, \quad F_j(n) = \sum_{i=1}^{\infty} I\{X_i = j\}. \tag{2}$$

In 1953 Good proposed an estimate for the probability mass of unobserved species by

$$\hat{Q}_n = \frac{F_1}{n}. \tag{3}$$

Intuitively, this estimate makes sense: If the fraction of singletons in the sample is very high, is seems more likely that there are much more unobserved species.

Later, in 1983, Esty proved the asymptotic normality of the Good estimator. More precisely, he showed that given the condition that

$$E(F_1(n))/n \to c_1 \in (0, 1), \quad E(F_2(n))/n \to c_2 \geqslant 0 \tag{4}$$

one has

$$Z_n := \frac{n(\hat{Q}_n - Q_n)}{\sqrt{F_1(n)(1 - F_1(n)/n) + 2F_2(n)}} \xrightarrow{D} N(0, 1), \tag{5}$$

i.e. Z_n converges in distribution to the normal distribution as $n \to \infty$.

(a) In a genomic context Mao and Lindsay studied a sample of $n = 2568$ expressed sequence tags from a tomato-flower cDNA library. The different observed genes let to the following data pattern (non-zero entries): $F_1 = 1434$, $F_2 = 253$, $F_3 = 71$, $F_4 = 33$, $F_5 = 11$, $F_6 = 6$, $F_7 = 2$, $F_8 = 3$ and $F_9 = F_{10} = F_{11} = F_{12} = F_{13} = F_{14} = F_{16} = F_{23} = F_{27} = 1$.

Use this data to calculate the Good estimate $\hat{Q}_n$ for the probability mass of unobserved gene expressions. Using the asymptotic normality (5) give an 95% confidence interval for $\hat{Q}_n$.

(b) We now want to check the asymptotic normality (5) by doing some simulation studies with `R`. For the simulations we take the number of species to be $N = 10000$ (before it was infinite). Let us assume an underlying probability vector $p_i \sim i^{-2}$. For $n = 50$ sample $\vec{X} = (X_i, i = 1, ..., N)$ from a multinomial and determine Z_n. Repeat this step 100 times until you have a sample of Z_n of size 100. To check for asymptotic normality make a Z_n vs standard normal Q-Q plot (using `qqnorm(...)`). The straight line $y = x$ (`qqline(...)`) represents where asymptotic normality hold. Perform the same analysis for $n = 1000$ and comment on your results.