Repetitive experiments: probability and frequency

← Back to Chapters

A version of Exercise 9.1

For more on the "missing species" problem see Wikipedia, the paper of Fisher and Section 6.2 of the Computer Age Statistical Inferenceby Efron and Hastie.

We follow the above references answer a different, albeit related, question: namely, how many new species we may hope to find if we sample more i.e. collect more specimens.

According to Fisher,

$S = -\alpha \ln(1-x)$, $N = \alpha \frac{x}{1-x}$

where α is some kind of "average Poisson rate" parameter and is independent of size of sample, while $x = \frac{\sigma}{\sigma+1}$ where $\sigma = \frac{x}{1-x}$ the scale parameter in the prior Gamma distribution over rates of Poissons, and grows linearly with sample size/sample time.

Then

$S = \alpha \ln(1+\sigma)$, $N = \alpha\sigma$, $\frac{S}{N} = \frac{\ln(1+\sigma)}{\sigma}$.

For $N = 122$, $S = 19$ this gives $\frac{\ln(1+t)}{t} = \frac{19}{122}$

and

$$\sigma = 19.345837627074542762469631647085763256768649243492746396 66$$

This can be expected to be reasonably accurate up to about doubling the sample size/σ.

After double the sample size i.e. doubling σ, the expected number of species S becomes

$$S_{new} = S_{init}\frac{\ln(1+2\sigma)}{\ln(1+\sigma)} = 23.214257688437440491078932702669742639041624950243016919 01$$

So after collecting 122 more samples we expect to discover about 4 new species.

Exercise 9.2

Presumably we are to rederive formula 3.89 (and no 3.77).

Given k linear functions $\mathcal{G}_1, \ldots \mathcal{G}_k$ of n_j with $\mathcal{G}_l(n_1, \ldots, n_m) = \sum g_j^l n_j$ we introduce $M(n, G_1, \ldots, G_k)$ as the number of tuples $\vec{n} = (n_1, \ldots, n_m)$ such that $\mathcal{G}_l(\vec{n}) = G_l$.

Recursion 9.19 becomes

$$M(n, G_1, \ldots, G_k) = \sum_{j=1}^{m} M(n-1, G_1 - g_j^1, \ldots, G_k - g_j^k)$$

Ansatz 9.20 becomes

$$M(n, G_1, \ldots, G_k) = \exp\{\alpha n + \sum_l \lambda_l G_l\}$$

and condition 9.21

$$\exp\{\alpha\} = Z(\lambda_1, \ldots, \lambda_k) = \sum_{j=1}^{m} \exp\{\sum_{l=1}^{k} -\lambda_l g_j^l\}$$

To shorten notation we write $\vec{G} = (G_1, \ldots, G_k)$ and $\vec{\lambda} = (\lambda_1, \ldots, \lambda_k)$. Then general solution 9.22 becomes

$$H(n, \vec{G}) = \int d\vec{\lambda} \;\; Z^n(\vec{\lambda}) \exp\{\vec{\lambda} \cdot \vec{G}\} \, h(\vec{\lambda})$$

Then with initial condition $M(0, \vec{G}) = \delta(\vec{G}, \vec{0})$ we have as before

$$Z^n(\vec{\lambda}) = \sum_{\vec{G}} M(n, \vec{G}) \exp\{-\vec{\lambda} \cdot \vec{g}\}$$

Now we apply this to derive multinomial distribution. Partition outcomes n_j into k sets $S_1, \ldots, S_k$ of size $s_1, \ldots, s_k$ and let $\mathcal{G}_l(n_j) = \begin{cases} 1 \text{ if } n_j \in S_l \\ 0 \text{ else} \end{cases}$

Then $M(n, c_1, \ldots, c_k)$ is the number of ways to get a tuple $\vec{n}$ with $\sum_{j \in S_l} n_j = c_l$ and we can compute

$$Z(\vec{\lambda}) = \sum_l s_l \exp\{-\lambda_l\} = \sum_l s_l x_l$$

$$Z^n(\vec{\lambda}) = \sum_{(c_1, \ldots c_k)} \binom{n}{c_1, \ldots c_k} \prod_l (s_l x_l)^{c_l}$$

and so

$$M(n, c_1, \ldots, c_k) = \binom{n}{c_1, \ldots c_k} \prod_l s_l^{c_l}$$

Probability of getting sequence of this type is $\frac{M(n,c_1,\ldots,c_k)}{n^m}$ and denoting $\frac{s_l}{n} = f_l$ we have

$$P(c_1, \ldots, c_k) = \binom{n}{c_1, \ldots c_k} \prod_l f_l^{c_l}$$

in agreement with 3.89.

Comments on Section 9.7

"If A is linear in the n_j, then it is the same as our G in (9.17)" – I think Jaynes means that "for a given linear function $\mathcal{G}$ of n_js, let A be the proposition '$\mathcal{G}$ takes value G' ". In that case, indeed, $M(n, G) = M(n, A)$ by definition.

"the notion of entropy inherent in probability theory independently of the work of Shannon" – in the work of Shannon, the notion of entropy is derived from combinatorics of sequences, much like here.

Exercise 9.3 (the first one)

$$f_k'' = \frac{n_k - \delta_{jk} - \delta_{tk}}{n - 2}$$

$$\delta f_k = f_k'' - f_k = f_k'' - \frac{n_k}{n} = \frac{2f_k - \delta_{jk} - \delta_{tk}}{n - 2}$$

so

$$\delta H = \sum_k (-1 - \ln(f_k)) \frac{2f_k - \delta_{jk} - \delta_{tk}}{n - 2}$$

$$\frac{2H + \ln(f_j) + \ln(f_t)}{n - 2}$$

and

$$H'' = \frac{nH + \ln(f_j) + \ln(f_t)}{n - 2}$$

$$M(n - 2, G - g_j - g_t) = \exp\{(n - 2)H''\} = f_j f_t \exp\{nH\}$$

so

$$p(r_i = j, r_s = t|GnI_0) = f_j f_t$$

The trials are still pairwise independent (and by similar analysis l-wise independent as long as $k \ll n$).

Formula 9.78

Assuming not all g_j are the same, strict convexity of $\ln Z$ implies that the derivative $-\frac{\partial \ln Z}{\partial \lambda}$ attains each value in its range once. However, it only attains values $\min g_j < \lambda < \max g_j$. Averages $\bar{G}$ outside this range - with exception of the endpoints $\min g_j$ and $\max g_j$ - are impossible (the endpoints correspond to $\lambda = \pm\infty$ and would require those u_i with non-extremal g_i to be set to zero, while not restricting the others; this is a rather special case, not directly covered by Jaynes' analysis in this section). We better condition on an achievable average.