Ma 322: Biostatistics
Brillouin and Shannon Diversity

Prof. Wickerhauser

Wednesday, January 15th, 2020

Suppose that N objects fall into K categories with n_k elements in category k for $k = 1, \ldots, K$. Thus $N = n_1 + \cdots + n_K$, with $0 \le n_k \le N$ for each k.

There are two commonly used methods to compute the *diversity* of such a set, namely to measure how evenly filled are the categories.

1 Brillouin's diversity.

This is defined by

$$H_B \stackrel{\text{def}}{=} \frac{1}{N} \ln \left(\frac{N!}{n_1! \cdots n_K!} \right) = \frac{1}{N} \left(\ln N! - \sum_{k=1}^{K} \ln n_k! \right).$$

It is used when the whole population is categorised into a known set of categories.

Because the argument of ln() is a multinomial coefficient, which must be a positive integer, it follws that $H_B \ge 0$.

The minimum value $H_B = 0$ is achieved by $n_1 = N$ and $n_2 = \cdots = n_K = 0$. Zero diversity means all elements are in the same category because if there are two or more nonempty categories then $H_B > 0$.

If N is a multiple of K, say $N = cK$, then the maximum value is achieved by $n_k = c = N/K$ for each $k = 1, \ldots, K$, giving

$$H_B^{\max} = \frac{1}{N} \left(\ln N! - K \ln c! \right).$$

However, if $N = cK + d$ with a remainder $0 < d < K$, then it is impossible to fill all categories with the same number of elements. The greatest diversity is then achieved by

$$n_k = \begin{cases} c+1, & k = 1, \ldots, d, \\ \lfloor K/N \rfloor = c, & k = d+1, \ldots, K, \end{cases}$$

which is as close to equal filling as possible. This gives the more general formula

$$H_B^{\max} = \frac{1}{N}\left(\ln N! - (K-d)\ln c! - d\ln(c+1)!\right).$$

The ratio $H_B/H_B^{\max}$ is called the *Brillouin evenness* or *relative Brillouin diversity* of the population.

2 Shannon's diversity

This is defined by

$$H_S \stackrel{\text{def}}{=} -\sum_{k=1}^{K} \frac{n_k}{N} \ln \frac{n_k}{N} = \ln N - \frac{1}{N}\sum_{k=1}^{K} n_k \ln n_k,$$

with the convention (justified by L'Hôpital's rule) that $0\ln(0) = 0$.

Shannon's diversity is most often computed for random samples of a multicategory population, and then the number of categories K is also a random variable determined by the measurement.

The minimum Shannon diversity $H_S = 0$ is achived by $n_1 = N$ with $n_2 = \cdots = n_K = 0$. If more than one category has nonzero elements, then $H_S > 0$, so that just as for Brillouin, zero Shannon diversity means all samples are in the same category.

The maximum Shannon diversity is achieved by equal numbers $n_k = N/K$ in all categories $k = 1, \ldots, K$, giving

$$H_S^{\max} = \ln N - \frac{1}{N}\sum_{k=1}^{K} \frac{K}{N} \ln \frac{K}{N} = \ln N - (\frac{1}{N})(\frac{N}{K})(K)\ln \frac{N}{K} = \ln N - \ln \frac{N}{K} = \ln K.$$

The ratio $H_S/H_S^{\max} = H_S/\ln K$ is called the *Shannon evenness* or *relative Shannon diversity* of the sample. Note that both numerator and denominator of these ratios may depend on the sample.