Самарская Лука. 2007. – Т. 16, № 3(21). – С. 585-587.

Э.Х. Симпсон**

ИЗМЕРЕНИЕ РАЗНООБРАЗИЯ*

«Характеристика», определенная Юлом[1] и «индекс разнообразия», определенный Фишером[2], – две меры степени концентрации или достигнутого разнообразия, когда особи популяции разбиты на группы. Оба [эти показателя] определены как статистики, которые вычисляются на основе выборочных данных, а не в терминах популяционных констант. Прежний индекс разнообразия основан на логарифмическом распределении. Это не общее допущение, поскольку не всегда дает значения, которые независимы от объема выборки; такой индекс нельзя использовать, например, применительно к бесконечному числу особей популяции, разбитых на конечное число групп. Уильямс[3] указал на зависимость между «характеристикой» и «индексом разнообразия», когда оба используют логарифмическое распределение. Цель настоящего сообщения состоит в том, чтобы определить и исследовать меру концентрации в терминах популяционных констант.

Пусть бесконечная популяция такова, что каждая особь принадлежит одной из $\boldsymbol{Z}$ групп, и пусть $\boldsymbol{\pi_1 ... \pi_Z}$ $(\boldsymbol{\Sigma\pi = 1})$ – доли особей в различных группах. Тогда $\boldsymbol{\lambda}$, определяемая как $\boldsymbol{\Sigma\pi^2}$, – мера концентрации классификации. Эта величина [$\boldsymbol{\lambda}$] может принимать любое значение между $\boldsymbol{1/Z}$ и $\boldsymbol{1}$, старое представление наименьшей концентрации или наибольшего разнообразия, возможного с $\boldsymbol{Z}$ группами, и последней полной концентрацией, когда все особи, находятся в одной группе. Показатель $\boldsymbol{\lambda}$ может быть просто интерпретирован как вероятность того, что две особи, выбранные случайно и независимо из популяции, будут принадлежать одной и той же группе.

Теперь предположим, что имеем выборку из $\boldsymbol{N}$ особей, случайно отобранных из популяции и $\boldsymbol{n}_1, \boldsymbol{n}_2, \ldots, \boldsymbol{n_Z}$ $(\boldsymbol{\Sigma n = N})$ – это число особей в разных группах. Легко показать, что $\boldsymbol{L} = (\boldsymbol{\Sigma n(n\text{-}1)}) / (\boldsymbol{N(N\text{-}1)})$ – несмещенная оценка $\boldsymbol{\lambda}$; это почти очевидно, если принять во внимание, что $0{,}5\boldsymbol{N(N\text{-}1)}$ – число пар в выборке и $0{,}5\boldsymbol{n(n\text{-}1)}$ – число пар, с учетом разбиения на группы.

$\boldsymbol{L}$ – также является несмещенной оценкой $\boldsymbol{\lambda}$ для переменного объема выборки, если отсутствуют выборки объема 0 или 1 и вероятность получения выборки $(\boldsymbol{n}_1, \boldsymbol{n}_2, \ldots, \boldsymbol{n_Z})$ раскладывается в эти два фактора:

** *Симпсон Эдуард (Edward Hugh Simpson; г.р. 1927) – математик, статистик; член Британского королевского статистического общества.*

* Simpson E.H. Measurement of diversity // Nature. – 1949. – V. 163, № 688. – P. 688. (перевод Г.С. Розенберга).

$$P(n_1, n_2, ..., n_Z) = P(N)\frac{N!}{n_1!n_2!...}(\pi_1)^{n_1}(\pi_2)^{n_2}... \quad ,$$

где $\boldsymbol{P}(\boldsymbol{N})$ задает распределение вероятности объема выборки, $\mathbf{2} \le \boldsymbol{N} \le \infty$. Это тем более верно, когда выборки получены методом «постоянных (стационарных) воздействий» (fixed-exposure), традиционным в биологических исследованиях; $\boldsymbol{N}$ тогда имеет распределение Пуассона, пригодное для описания выборок с отсутствием первых двух [выборок объема 0 и 1].

Если повторная выборка объема $\boldsymbol{N}$ получена из той же самой популяции, вычисленное значение $\boldsymbol{L}$ будет распределено как $\boldsymbol{\lambda}$ с дисперсией

$$\frac{4N(N-1)(N-2)\sum\pi^3 + 2N(N-1)\sum\pi^2 - 2N(N-1)(2N-3)(\sum\pi^2)^2}{\left[N(N-1)\right]^2} \;;$$

или, если $\boldsymbol{N}$ очень велико, приблизительно

$$\frac{4}{N}\left[\sum\pi^3 - (\sum\pi^2)^2\right] .$$

Третьи и четвертые кумулянты [*накопленные частоты. – Г.Р.*] распределения $\boldsymbol{L}$ также были точно определены. Они показывают, что при росте $\boldsymbol{N}$ распределение стремится к нормальному, кроме того случая, когда $\boldsymbol{\lambda} = 1/\boldsymbol{Z}$; в этом случае распределение $\boldsymbol{LNZ}$ стремится к χ^2-распределению с $(\boldsymbol{Z}\text{-}1)$ степенью свободы, но со средней смещенной от $\boldsymbol{Z}$-1 к $\boldsymbol{N}$.

«Характеристика», определенная Юлом[1], – 1000 $(\Sigma \boldsymbol{n}(\boldsymbol{n}\mathbf{-1})) / \boldsymbol{N}^2$, что отличается от $\boldsymbol{L}$, оценивающей $\boldsymbol{\lambda}$, только наличием $\boldsymbol{N}$ вместо $\boldsymbol{N}$**-*1*** в знаменателе и коэффициентом масштаба 1000.

Теперь, позвольте показать, что значение $\boldsymbol{\lambda}$, взятое для популяции, состоящей из Z групп, частоты которых $\boldsymbol{\pi}_\mathbf{i} = \mathbf{w_i}/\Sigma\mathbf{w}$, где $\mathbf{w_i}$ выбраны случайно и независимо, соответствует распределению III-го типа:

$$dF = \frac{1}{(k-1)!}e^{-w}w^{k-1}dw \quad , \qquad \mathbf{0} \le \boldsymbol{w} \le \infty \, .$$

Это можно назвать «отрицательной биномиальной популяцией», так как выборки, полученные методом «постоянных (стационарных) воздействий», будут подчиняться отрицательному биномиальному распределению. Можно вычислить соответствующее этому значение $\boldsymbol{\lambda}$, со средним значением $\Sigma\mathbf{w_i}^2/(\Sigma\mathbf{w_i})^2$ по всем наборам $(\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_Z)$, которое может быть получено по популяционным значениям $\boldsymbol{w}$. Таким образом,

$$\lambda = \int_0^{\infty}...\int_0^{\infty}\left[\frac{1}{(k-1)!}\right]e^{-\sum w}\left[w_1...w_Z\right]^{k-1}dw_1...dw_Z = \frac{k+1}{Zk+1} \quad .$$

Распределение Пуассона – специальный случай отрицательного биномиального распределения, в котором $\boldsymbol{k}$ стремится к бесконечности. При этом условии, $\boldsymbol{\lambda} = 1/\boldsymbol{Z}$. Это именно то, что мы и ожидали, так как распределение Пуассона соответствует популяции, в которой все группы представлены одинаково, и, таким образом, вероятность, что две случайно выбранные особи будут принадлежать одной группе, должна быть $1/\boldsymbol{Z}$.

Другой особый случай отрицательного биномиального распределения – «логарифмическая популяция», которая получается, если одновременно $\boldsymbol{Z}$ стремится к бесконечности, а $\boldsymbol{k}$ – к нулю так, что произведение $\boldsymbol{Zk}$ остается конечным и стремится к величине α. (Это – не совсем то же самое предположение, которое использует Фишер[2], но количественно α и есть его «индекс разнообразия»). Полученное значение для λ, снизу ограничено $1/(\alpha+1)$.

Заметим, что это последнее значение не совместимо с уравнением, приводимым Уильямсом[3], а именно, «характеристика» Юла[1] имеет вид $1000/\alpha$, для логарифмического распределения. Этот результат был получен с использованием формулы Юла к рядам вероятных значений, тогда как предлагаемая процедура эквивалентна применению формулы сначала и затем усреднения результата. Некоторая поддержка новому уравнению найдена при рассмотрении рангов связанных переменных. Так как «характеристика» не может превысить 1000, более раннее уравнение [$1000/\alpha$. – *Г.Р.*] отрицало бы все значения α меньше чем 1; но предлагаемое уравнение задает диапазон $\mathbf{0 \le \alpha \le \infty}$, в то время как $\mathbf{1 \ge \lambda \ge 0}$.

3 West End Avenue,
Pinner (*пригород в северо-западном Лондоне. – Г.Р.*).
Jan. 29.

[1] Yule, "Statistical Study of Literary Vocabulary" (Cambridge, 1944).
[2] Fisher, Corbet and Williams, *J. Animal Ecol.*, **12**, 42 (1948).
[3] Williams, *Nature*, **157**, 482 (1946).