## ОЦЕНИВАНИЕ В ГЕТЕРОГЕННЫХ ПОПУЛЯЦИЯХ

### 4.1 Метод максимального правдоподобия

В предыдущих главах мы описывали вероятностные закономерности, присущие процессам в гетерогенных популяциях. Это задача теории вероятностей. В настоящей главе будем рассматривать обратную задачу – по наблюдениям над процессом в популяции извлекать информацию о параметрах и структуре популяции. Это задача математической статистики. При анализе данных и оценивании параметров моделей широко применяется метод максимума правдоподобия, обязанный своему внедрению в практику Р.Фишеру [29] – статистическая процедура, направленная на построение модели, которая соответствует максимальной вероятности наблюдать конкретные экспериментальные данные. Пусть изучаемая случайная величина $X$ имеет дискретное распределение $f(x,q)$, зависящее от параметра $q$, принадлежащего некоторому множеству $\Theta$, например отрезку $[0,1]$. Вероятность наблюдать экспериментальную выборку значений $x_1,...,x_n$ является функцией параметра $q$, называется *функцией правдоподобия*, и в случае независимой выборки равна

$$L(q) = \prod_{i=1}^{n} f(x_i,q).$$

Если случайная величина $X$ принимает непрерывные значения, например как длительность безотказной работы, то в качестве функ-

ции $f(x,q)$ принимают функцию плотности вероятности случайной величины $X$. Метод максимального правдоподобия предписывает в качестве оценки параметра $q$ принимать то значение из множества $\Theta$, при котором функция $L(q)$ максимальна. Полученная оценка называется *оценкой максимального правдоподобия*

$$\hat{q} = \arg\max_{q\in\Theta} L(q).$$

В общем случае и случайная величина $X$ и параметр $q$ могут быть конечно - мерными векторами или принадлежать функциональным пространствам. При выполнении общих условий регулярности относительно оценок максимального правдоподобия доказано, что с увеличением числа независимых наблюдений эти оценки сходятся к величине параметра $q_0$, задающей то распределение, согласно которому была сгенерирована экспериментальная выборка [29]. Доказана асимптотическая нормальность оценок максимального правдоподобия. При этом, предполагается, что используемая модель верно описывает закон распределения $f(x,q)$ и ‘точное’ значение параметра $q_0$ принадлежит заданной области $\Theta$. В случае неточного описания данных моделью, построенную модель рассматривают как некоторую ‘проекцию’ реальности на заданное модельное пространство, применяя при этом специальные корректирующие приёмы [15, 30, 31].

При анализе времени безотказной работы или продолжительности жизни плотность вероятности наблюдать отказ или смерть в мо-

мент времени $t$ равна

$$f(t,q) = -\frac{d}{dt}S(t,q)$$
$$= m(t,q)S(t,q)$$
$$= m(t,q)\exp\left(-\int_0^t m(t,q)dt\right),$$

где $m(t,q)$ - интенсивность отказа при фиксированном значении параметра $q$, заданная моделью. Функция правдоподобия для независимых моментов отказов $t_1,...,t_n$ равна

$$L(q) = \prod_{i=1}^{n} m(t_i,q)S(t_i,q)$$
$$= \prod_{i=1}^{n} m(t_i,q)\exp\left(-\sum_{i=1}^{n}\int_0^{t_i} m(t,q)dt\right). \qquad (4.1)$$

Необходимым условием достижения функцией правдоподобия максимума по параметру $q = (q_1,...,q_k)$ является равенство нулю её частных производных или частных производных от её логарифма

$$\frac{\partial}{\partial q_j}\ln L(q) = 0, \quad j = 1,...,k.$$

Для функции правдоподобия (4.1) получаем систему уравнений

$$\sum_{i=1}^{n}\frac{1}{m(t_i,q)}\frac{\partial}{\partial q_j}m(t_i,q) = \sum_{i=1}^{n}\int_0^{t_i}\frac{\partial}{\partial q_j}m(t,q)dt, \quad j = 1,...,k,$$

которая принимает простой вид для постоянной интенсивности

отказа $m(t,q)=q$ , $k=1$. Из соотношения $n/q=\sum_{i=1}^{n}t_i$ следует оценка максимального правдоподобия постоянной интенсивности отказа в виде $\hat{q}=n/\sum_{i=1}^{n}t_i$ хорошо известная в литературе по теории надёжности как величина, обратная среднему времени безотказной работы. Для более сложных моделей интенсивности отказа применяют численные методы максимизации функции правдоподобия (4.1). Аналогичная схема применяется и при оценивании в гетерогенных популяциях. Функция правдоподобия для случая гамма распределённой уязвимости имеет вид

$$L=\prod_{i=1}^{n}\overline{m}(t_i)S(t_i)$$
$$=\prod_{i=1}^{n}\frac{m_0(t_i)}{\left(1+s^2\int_0^{t_i}m_0(t)dt\right)^{\left(1+s^2\right)/s^2}}.$$

Вид функции правдоподобия может изменяться в зависимости от структуры эмпирических данных. Так, часто наблюдают не длительности безотказной работы или продолжительности жизни, а число отказавших элементов или число смертей в фиксированные промежутки времени. Пусть до дня $t_i$ доработало $n_i$ независимых элементов или дожило $n_i$ особей в лабораторном эксперименте. Число элементов, отказавших на интервале $[t_i,t_{i+1})$, или число осо-

бей, погибших на этом интервале, обозначим через $d_i$. Если $S(t)$ является вероятностью безотказной работы одного элемента в течение времени $t$ или функцией дожития до возраста $t$, то вероятность отказа или смерти на интервале $[t_i, t_{i+1})$ при условии дожития до $t_i$ равна $p_i = 1 - S(t_{i+1}) / S(t_i)$. Вероятность наблюдать $d_i$ отказов или смертей среди $n_i$ независимых кандидатов задаётся биномиальным распределением

$$p(d_i, n_i) = C_{n_i}^{d_i} p_i^{d_i} (1 - p_i)^{n_i - d_i}.$$

Логарифм функции правдоподобия равен

$$\ln L = \sum_i \ln p(d_i, n_i)$$
$$= \sum_i \left( d_i \ln p_i + (n_i - d_i) \ln(1 - p_i) \right) + const$$

где слагаемое $const$ не зависит от функции дожития.

4.2 Оценивание в LCM и GOM моделях

Метод максимального правдоподобия применим и для оценки параметров модели латентных классов - LCM модели. Для построения функции правдоподобия воспользуемся обобщением биномиального закона распределения – мультиномиальным распределением. Пусть $n_{i_1 \ldots i_M}$ наблюдений из общего числа $N$ попало по координате $j$ в ячейку $i_j$ ($j$=1,…,$M$) таблицы сопряжённости. Вероятность распределения наблюдений по ячейкам таблицы сопряженности -

функция правдоподобия, описывается мультиномиальным законом распределения вероятностей и равна

$$L = N! \prod_{i_1=1}^{L_1} \dots \prod_{i_M=1}^{L_M} \frac{\left( p_{i_1 \dots i_M} \right)^{n_{i_1 \dots i_M}}}{n_{i_1 \dots i_M}!}.$$

Здесь $L_j$ - число градаций по переменной $j$. Подставив вероятность попадания наблюдения в конкретную ячейку таблицы сопряжённости, получаемое в рамках LCM модели,

$$p_{i_1 \dots i_M} = \sum_{a=1}^{K} \Theta_a \prod_{j=1}^{M} p_{i_j}^{(a)},$$

логарифм функции правдоподобия можно записать в виде

$$\ln L = \sum_{i_1=1}^{L_1} \dots \sum_{i_m=1}^{L_M} n_{i_1 \dots i_M} \ln\left( \sum_{a=1}^{K} \Theta_a \prod_{j=1}^{M} p_{i_j}^{(a)} \right) + const.$$

Во второе слагаемое включены члены, не зависящие от параметров модели. Первое слагаемое можно переписать введя двоичную переменную $y_{l,i_1 \dots i_M}$, которая принимает значение 1 если наблюдение номер $l$ попадает в ячейку с координатами $\{i_1, \dots, i_M\}$ и равную 0 в противном случае

$$\ln L \approx \sum_{l=1}^{N} \sum_{i_1=1}^{L_1} \dots \sum_{i_M=1}^{L_M} y_{l,i_1 \dots i_M} \ln\left( \sum_{a=1}^{K} \Theta_a \prod_{j=1}^{M} p_{i_j}^{(a)} \right).$$

В этом выражении внешнее суммирование ведётся по номеру эксперимента и суммируемое выражение имеет смысл логарифма правдоподобия отдельно взятого наблюдения. Оценки максималь-

ного правдоподобия стоятся путем максимизации полученной функции по параметрам $\Theta_a$, $p_{i_j}^{(a)}$ при учете ограничений

$$\sum_{a=1}^{K}\Theta_a = 1$$

$$\sum_{i_j=1}^{L_j} p_{i_j}^{(a)} = 1,\ \ j = 1,\ldots, M,\ \ \ a = 1,\ldots, K$$

$$\sum_{i_1=1}^{L_1}\ldots\sum_{i_M=1}^{L_M}\sum_{a=1}^{K}\Theta_a\prod_{j=1}^{M} p_{i_j}^{(a)} = 1.$$

Вычисление функции правдоподобия для модели степени принадлежности GOM не удаётся свести к известной схеме как для модели латентных классов. В этом случае функция правдоподобия вычисляется напрямую как произведение правдоподобий каждого независимого наблюдения. Функцию правдоподобия

$$L = \prod_{i=1}^{I} L_i$$

$$= \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{l=1}^{L_j}\left(\sum_{k=1}^{K} g_{ik}\, l_{kjl}\right)^{y_{ijl}},$$

необходимо максимизировать по значениям функций принадлежности $g_{ik}$ и профильным вероятностям $l_{kjl}$ при ограничениях $g_{ik} \geq 0$, $\sum_{k=1}^{K} g_{ik} = 1$, $i$=1,…,$I$, $l_{kjl} \geq 0$, $\sum_{l=1}^{L_j} l_{kjl} = 1$, $k$=1,…,$K$, $j$=1,…,$J$. Для решения этой задачи созданы соответствующие алгоритмы [22] и программный комплекс GOM (Decision System, Inc), доступный через Интернет по адресу www.dsisoft.com.

4.3 Выбор структуры модели

При построении моделей процессов в неоднородных популяциях, описанных в предыдущих главах, существует проблема оценки структуры модели. При анализе генетического влияния на продолжительность жизни близнецов была возможность учитывать только аддитивный генетический эффект, пренебрегая доминантным или эпистатическим эффектами, в моделях латентных классов (LCM) и степени принадлежности (GOM) число классов и чистых типов также являются параметрами структуры модели. Обычно, относительно структуры модели имеется некоторая априорная информация. Предварительные исследования указывают на факторы, играющие ведущую роль, сама рассматриваемая задача в своей постановке часто содержит заданную типизацию явления. Проблема возникает в случае, если число имеющихся наблюдений невелико. При этом часто оказывается, что “сложная” модель, учитывающая все особенности изучаемого явления, не может быть идентифицирована по имеющимся данным с достаточной точностью. В то же время более “простая” модель, пренебрегающая второстепенными деталями, может по тем же данным давать статистически более надёжный результат.

Теоретическое обоснование понятия сложности модели было заложено в работах по распознаванию математических образов и восстановлению статистических зависимостей [30, 32]. Получены соотношения между числом экспериментальных данных и числом параметров модели, достаточные для обеспечения статистической

надёжности получаемых оценок. Так, при построении регрессионных зависимостей методом наименьших квадратов по выборке из $n$ элементов, $k$ - число параметров в модели следует выбирать из условия минимизации функционала $R_k$, вычисляемого по формуле

$$R_k = R_k^n / \left( 1 - \sqrt{\frac{k}{n}\left(\ln n / k + 1\right) - \ln q} \right)$$

где $R_k^n$ - остаточная невязка для модели с $k$ параметрами, построенной по $n$ наблюдениям. Параметр $q$ задаёт величину ошибки метода и обычно равен 0.5. Такой подход получил название *метод структурной минимизации риска* [30]. Применения метода структурной минимизации риска в задачах распознавания образов, построения линейной и нелинейной регрессии, решения некорректных задач с соответствующими алгоритмами и программами описаны в [33].

Для выбора структуры моделей, построенных на основе максимизации функции правдоподобия, широко используется подход, связанный со статистической проверкой гипотез - критерий отношения правдоподобия. При этом проверяется гипотеза $H_0$, что исключение одного или нескольких параметров из модели изменит достигнутую величину функции правдоподобия несущественно. Для принятия такого решения вычисляется величина

$$Lr = 2 \times \ln\left( \max_{a \in \mathrm{A}} L(a) / \max_{b \in \mathrm{B}} L(b) \right),$$

где $\mathrm{A}$ и $\mathrm{B}$ обозначают сравниваемые классы моделей, отличающиеся числом параметров, причём $\mathrm{B} \subset \mathrm{A}$. Величина $Lr$ асимпто-

тически имеет $c^2$ распределение с числом степеней свободы $d$, равным разности между числом параметров моделей из классов $\mathrm{A}$ и $\mathrm{B}$. Переход от класса $\mathrm{A}$ к более узкому классу $\mathrm{B}$ считается целесообразным, если величина $p$-value, равная вероятности того, что случайная величина, имеющая распределение $c^2$ с $d$ степенями свободы, превосходит вычисленное значение критерий отношения правдоподобия $Lr$, достаточно мала. На практике за критическую величину для $p$-value часто принимают величину 0.01 и условие перехода к более простой модели принимает вид

$$P\left\{ c_d^2 > Lr \right\} < 0.01.$$

Использование в критерии отношения правдоподобия асимптотического распределения делает этот критерий малопригодным в условиях малого числа наблюдений или большого числа оцениваемых параметров. Среди альтернативных критериев, ориентированных на статистические выборки малого объёма, широкую популярность получил информационный критерий Акаике [34]. Для модели, построенной в классе $\mathrm{A}$ методом максимуму правдоподобия, информационный критерий имеет вид

$$L_A = \ln\left( \max_{a \subset \mathrm{A}} L(a) \right) - 2d,$$

где $d$ означает число оцениваемых параметров в модели. Модель, построенная в классе $\mathrm{B} \subset \mathrm{A}$, считается предпочтительнее модели, построенной в классе $\mathrm{A}$, если для неё значение информационного критерия больше, то есть $L_B > L_A$.