УДК 001.8

В. В. Нешитой

Метод наибольшего правдоподобия, устойчивый метод и энтропия

Исследуется близость двух методов оценивания параметров непрерывных, в том числе ранговых, распределений – метода наибольшего правдоподобия Р. Фишера и устойчивого метода автора. Показывается, что оба метода базируются на одном и том же равенстве p(x) = βzp(z), устанавливающем взаимосвязь между обобщенной плотностью p(x) и двухпараметрической плотностью p(z). В методе наибольшего правдоподобия используется логарифм этого равенства. Показывается, что логарифмическая функция правдоподобия, взятая с обратным знаком, представляет собой энтропию распределения. При преобразовании распределений второй системы к форме первой системы энтропия плотности уменьшается, т.е. появляется новая информация.

Ключевые слова: *системы непрерывных распределений, методы оценивания параметров, метод наибольшего правдоподобия, устойчивый метод, логарифмическая функция правдоподобия, критерий согласия, энтропия распределения, извлечение информации из распределения*

МЕТОД НАИБОЛЬШЕГО ПРАВДОПОДОБИЯ

В 1912 г. Р. Фишер предложил для нахождения оценок параметров аппроксимирующих распределений метод наибольшего (максимального) правдоподобия. Суть метода сводится к тому, что наступившие события имели наибольшую вероятность наступить при заданном комплексе условий. Вероятность совместного наступления событий при условии их независимости равна произведению вероятностей наступивших событий. Это произведение называется функцией правдоподобия:

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

В качестве оценок максимального правдоподобия параметров Θ_j принимаются те их значения, при которых функция правдоподобия имеет максимум. Дифференцируя функцию правдоподобия по параметрам Θ_j и приравнивая частные производные к нулю, получают систему уравнений правдоподобия, решая которую, находят оценки параметров. Однако здесь следует отметить, что в случае распределений с тремя и тем более с четырьмя параметрами получаются весьма сложные уравнения, решение которых сопряжено с большими трудностями.

Ниже будут рассматриваться четырехпараметрические непрерывные распределения автора [1], задающие три системы непрерывных распределений:

$$p(x) = Ne^{k\beta x}(1 - \alpha u e^{\beta x})^{\frac{1}{u}-1} \ ;$$

$$p(t) = Nt^{k\beta-1}(1 - \alpha u t^{\beta})^{\frac{1}{u}-1} \ ;$$

$$p(y) = \frac{N}{y}(\ln y)^{k\beta-1}\left[1 - \alpha u(\ln y)^{\beta}\right]^{\frac{1}{u}-1} .$$

Приведенные плотности распределения включают как частные случаи широкое разнообразие непрерывных распределений.

ДРУГИЕ ФОРМЫ ЗАДАНИЯ ФУНКЦИИ ПРАВДОПОДОБИЯ

Приведенная форма задания функции правдоподобия (1) не является единственно возможной. Суть метода наибольшего правдоподобия не изменится, если из функции правдоподобия L извлечь корень n-ой степени:

$$L^* = \sqrt[n]{\prod_{i=1}^{n} f(x_i, \theta_j)} = \left[\prod_{i=1}^{n} f(x_i, \theta_j)\right]^{\frac{1}{n}} . \qquad (2)$$

Здесь функция правдоподобия задана средним геометрическим вероятностей n независимых событий. Это вторая возможная форма задания функции правдоподобия.

Третья и четвертая формы получаются путем логарифмирования первых двух форм:

$$\ln L = \sum_{i=1}^{n} \ln f(x_i, \theta_j) \ ; \qquad (3)$$

$$\ln L^* = \frac{1}{n}\sum_{i=1}^{n} \ln f(x_i, \theta_j) = \overline{\ln f(x_i, \theta_j)} \ . \qquad (4)$$

Здесь форма (3) задана суммой логарифмов вероятностей n независимых событий, а форма (4) – средним значением логарифмов вероятностей.

Все четыре формы представления функции правдоподобия являются равноправными в том смысле, что дают одни и те же оценки параметров аппроксимирующего распределения. При этих оценках каждая функция правдоподобия принимает свое максимальное значение.

Рассмотрим пример.

Пусть случайная величина T имеет показательный закон распределения

$$p(t) = \frac{\alpha}{e^{\alpha t}} \ . \qquad (5)$$

Необходимо оценить параметр α по результатам наблюдений $t_1, t_2, ..., t_n$, где n – объем выборки.

РЕПОЗИТОРИЙ БГУКИ

Запишем для показательного закона все рассмотренные нами формы задания функции правдоподобия:

1. $L=\prod_{i=1}^{n}p(t_i)=\frac{\alpha}{e^{\alpha t_1}}\cdot\frac{\alpha}{e^{\alpha t_2}}\cdots\frac{\alpha}{e^{\alpha t_n}}=\frac{\alpha^n}{e^{\alpha\sum t_i}}$;
2. $L^*=\left[\prod_{i=1}^{n}p(t_i)\right]^{\frac{1}{n}}=\frac{\alpha}{e^{\alpha(\sum t_i)/n}}=\frac{\alpha}{e^{\alpha\bar{t}}}$;
3. $\ln L=n\ln\alpha-\alpha\sum_{i=1}^{n}t_i$;
4. $\ln L^*=\ln\alpha-\alpha\bar{t}$.

Дифференцируя любую из них по параметру α и приравнивая первую производную к нулю, найдем:

$$\alpha=\frac{1}{\bar{t}}. \qquad (6)$$

МОДИФИЦИРОВАННЫЙ МЕТОД НАИБОЛЬШЕГО ПРАВДОПОДОБИЯ

В качестве логарифмической функции правдоподобия для непрерывных распределений в случае генеральной совокупности может быть принято математическое ожидание логарифма плотности распределения [2]:

$$M\left[\ln p(x)\right]=\ln L. \qquad (7)$$

Использование этой функции правдоподобия позволяет значительно проще решать такие задачи, как вычисление оценок параметров, вычисление значений функции правдоподобия при заданных значениях параметров, проведение научных исследований. Эта форма задания функции правдоподобия следует из формулы (4), в которой среднее значение логарифмической функции правдоподобия заменено математическим ожиданием.

Используем функцию правдоподобия (7) для нахождения оценки параметра α показательного закона распределения (5). Здесь порядок расчета следующий:

- логарифмируем плотность распределения

$$\ln p(t)=\ln\alpha-\alpha t;$$

- находим математическое ожидание логарифма плотности

$$M\left[\ln p(t)\right]=\ln\alpha-\alpha M(t);$$

- находим частную производную по параметру α и приравниваем ее к нулю

$$\frac{\partial M\left[\ln p(t)\right]}{\partial\alpha}=\frac{1}{\alpha}-M(t)=0;$$

- из полученного уравнения правдоподобия находим зависимость между параметром α и математическим ожиданием случайной величины T в случае показательного закона распределения

$$\alpha=\frac{1}{M(t)}.$$

Последняя формула справедлива для генеральной совокупности. Для перехода к выборочной совокупности заменяем математическое ожидание случайной величины T его оценкой, вычисленной по выборке объемом n:

$$\bar{t}=\frac{1}{n}\sum_{i=1}^{n}t_i.$$

Оценка параметра α будет равна $\alpha=1/\bar{t}$, т.е. вычисляется по прежней формуле (6).

Рассмотрим пример на вычисление M[lnp(x)]. Логарифмическую функцию правдоподобия можно вычислить двумя способами. Первый, традиционный, способ – это ее вычисление путем интегрирования:

$$M\left[\ln p(x)\right]=\int\left[\ln p(x)\right]p(x)dx.$$

Пусть плотность p(x) задается четырехпараметрической формулой

$$p(x)=\frac{\beta(\alpha u)^k\Gamma(\mathrm{k}+1/\mathrm{u})}{\Gamma(\mathrm{k})\Gamma(1/\mathrm{u})}e^{k\beta x}\left(1-\alpha u e^{\beta x}\right)^{\frac{1}{u}-1}, \qquad (8)$$

которая относится к первому типу первой системы непрерывных распределений [1].

Тогда логарифмическая функция правдоподобия будет выражаться интегралом

$$M\left[\ln p(x)\right]=\int\left[\begin{array}{l}\ln\frac{\beta(\alpha u)^k\Gamma(\mathrm{k}+1/\mathrm{u})}{\Gamma(\mathrm{k})\Gamma(1/\mathrm{u})}+k\beta x+\\+\left(\frac{1}{u}-1\right)\ln\left(1-\alpha u e^{\beta x}\right)\end{array}\right]p(x)dx,$$

где плотность p(x) задается формулой (8).

При втором способе используется метод дифференцирования [2].

Прологарифмируем плотность распределения (8):

$$\ln p(x)=\ln\beta+k\ln\alpha u+\ln\Gamma(\mathrm{k}+1/\mathrm{u})-\ln\Gamma(\mathrm{k})-\\-\ln\Gamma(1/\mathrm{u})+\mathrm{k}\beta\mathrm{x}+(1/\mathrm{u}-1)\ln\left(1-\alpha\mathrm{u}e^{\beta\mathrm{x}}\right).$$

На основании последнего равенства запишем логарифмическую функцию правдоподобия

$$\ln L=M\left[\ln p(x)\right]=\ln\beta+k\ln\alpha u+\ln\Gamma(\mathrm{k}+1/\mathrm{u})-\ln\Gamma(\mathrm{k})-\\-\ln\Gamma(1/\mathrm{u})+k\beta M(x)+(1/\mathrm{u}-1)\mathrm{M}\left[\ln\left(1-\alpha\mathrm{u}e^{\beta\mathrm{x}}\right)\right]. \qquad (9)$$

Возьмем от функции (9) частные производные по параметрам α, β, k, u и приравняем их к нулю. В результате получим систему четырех уравнений правдоподобия с четырьмя неизвестными параметрами:

$$\begin{cases}\frac{\partial\ln L}{\partial\alpha}=\frac{k}{\alpha}+\left(\frac{1}{u}-1\right)M\left(\frac{-ue^{\beta x}}{1-\alpha ue^{\beta x}}\right)=0\\ \frac{\partial\ln L}{\partial\beta}=\frac{1}{\beta}+kM(x)+\left(\frac{1}{u}-1\right)M\left(\frac{-\alpha ue^{\beta x}x}{1-\alpha ue^{\beta x}}\right)=0\\ \frac{\partial\ln L}{\partial k}=\ln\alpha u+\psi\left(k+\frac{1}{u}\right)-\psi(k)+\beta M(x)=0\\ \frac{\partial\ln L}{\partial u}=\frac{k}{u}+\psi\left(k+\frac{1}{u}\right)\left(-\frac{1}{u^2}\right)-\psi\left(\frac{1}{u}\right)\left(-\frac{1}{u^2}\right)+\\ +\left(-\frac{1}{u^2}\right)M\left[\ln\left(1-\alpha ue^{\beta x}\right)\right]+\left(\frac{1}{u}-1\right)M\left(\frac{-\alpha ue^{\beta x}}{1-\alpha ue^{\beta x}}\right)=0.\end{cases} \qquad (10)$$

Оценки четырех параметров находятся путем решения системы уравнений правдоподобия (10).

Полученные уравнения можно несколько упростить. Так, из первого уравнения правдоподобия имеем

$$k=\alpha(1-u)M\left(\frac{e^{\beta x}}{1-\alpha ue^{\beta x}}\right). \qquad (11)$$

РЕПОЗИТОРИЙ БГУКИ

Умножив четвертое уравнение правдоподобия на u, получим:

$$k+\psi\left(k+\frac{1}{u}\right)\left(-\frac{1}{u}\right)-\psi\left(\frac{1}{u}\right)\left(-\frac{1}{u}\right)+\left(-\frac{1}{u}\right)M\left[\ln\left(1-\alpha u e^{\beta x}\right)\right]-$$

$$-\alpha(1-u)M\left(\frac{e^{\beta x}}{1-\alpha u e^{\beta x}}\right)=0.$$

С учетом (11) последнее равенство примет вид:

$$\psi\left(k+\frac{1}{u}\right)-\psi\left(\frac{1}{u}\right)+M\left[\ln\left(1-\alpha u e^{\beta x}\right)\right]=0\ . \quad (12)$$

Перепишем систему уравнений правдоподобия для первого типа распределений, заданных плотностью (8):

$$\begin{cases} k-\alpha(1-u)M\left(\dfrac{e^{\beta x}}{1-\alpha u e^{\beta x}}\right)=0 \\ \dfrac{1}{\beta}+kM(x)-\alpha(1-u)M\left(\dfrac{xe^{\beta x}}{1-\alpha u e^{\beta x}}\right)=0 \\ \ln\alpha u+\psi\left(k+\dfrac{1}{u}\right)-\psi(k)+\beta M(x)=0 \\ \psi\left(k+\dfrac{1}{u}\right)-\psi\left(\dfrac{1}{u}\right)+M\left[\ln\left(1-\alpha u e^{\beta x}\right)\right]=0\ . \end{cases} \quad (13)$$

С учетом двух последних уравнений системы (13) логарифмическая функция правдоподобия запишется в окончательном виде [2]:

$$\ln L=M[\ln p(x)]=\ln\beta+\ln\frac{\Gamma(\mathrm{k}+1/u)}{\Gamma(\mathrm{k})\Gamma(1/\mathrm{u})}+$$

$$+k\left[\psi(k)-\psi\left(k+\frac{1}{u}\right)\right]+\left(\frac{1}{u}-1\right)\left[\psi\left(\frac{1}{u}\right)-\psi\left(k+\frac{1}{u}\right)\right]. \quad (14)$$

Таким образом, логарифмическая функция правдоподобия получена достаточно простым способом без интегрирования.

Из (14) следует, что логарифмическая функция правдоподобия зависит от трех параметров: β, k, u. Она может быть вычислена по этой формуле для распределений первого типа не только первой системы непрерывных распределений, но и второй и третьей систем, если их привести к форме плотности p(x), т.е. представить в виде tp(t)=f(lnt), yp(y)lny=f(lnlny).

ОЦЕНИВАНИЕ ПАРАМЕТРОВ ПО МЕТОДУ НАИБОЛЬШЕГО ПРАВДОПОДОБИЯ

Традиционно считается, что оценки параметров по методу наибольшего правдоподобия находятся путем решения системы уравнений правдоподобия, при этом число уравнений равно числу параметров.

В нашем примере обобщенное распределение (8) содержит четыре параметра: два параметра формы k, u, масштабный параметр β и параметр сдвига α. Главными здесь являются параметры формы.

Уравнения правдоподобия (13) оказались весьма сложными, причем в каждом из них имеются все четыре параметра. Решить такую систему практически невозможно. Однако, как показали исследования, в этом нет необходимости. Чтобы упростить решение поставленной задачи, ее надо разделить на два этапа.

На первом этапе разрабатываются два критерия (показателя), которые зависят только от двух параметров формы k, u. По этим критериям устанавливается тип аппроксимирующего распределения, а оценки параметров k, u вычисляются либо графическим методом, либо путем решения системы **двух уравнений с двумя неизвестными**. Такая система уравнений решается значительно проще, чем система четырех уравнений с четырьмя неизвестными, заданная, например, формулами (13).

На втором этапе при известных оценках параметров формы по простым формулам вычисляются оценки параметров α, β.

Такой подход позволил автору разработать устойчивый метод вычисления наилучшего аппроксимирующего распределения и нахождения оценок его параметров.

Начнем решать задачу оценивания параметров по методу наибольшего правдоподобия со второго этапа.

Рассмотрим случайную величину

$$Z=\alpha u e^{\beta X}\ . \quad (15)$$

и найдем плотность распределения случайной величины Z по известной формуле *p(z)=p(x)(dx/dz)=p(x)/(dz/dx).*

Первая производная от z по x равна $dz/dx=\alpha u\beta e^{\beta x}$. Тогда из плотности (8) получим

$$p(z)=\frac{\Gamma(\mathrm{k}+1/\mathrm{u})}{\Gamma(\mathrm{k})\Gamma(1/\mathrm{u})}z^{k-1}(1-z)^{\frac{1}{u}-1}, \quad (16)$$

т.е. имеем бета-распределение.

Приведем плотность (16) к форме плотности (8), т.е. представим ее в виде зависимости *zp(z)=f(lnz)*:

$$zp(z)=\frac{\Gamma(\mathrm{k}+1/\mathrm{u})}{\Gamma(\mathrm{k})\Gamma(1/\mathrm{u})}e^{k\ln z}\left(1-e^{\ln z}\right)^{\frac{1}{u}-1}.$$

Запишем для последней плотности логарифмическую функцию правдоподобия:

$$\ln L_{(z)}=M[\ln zp(z)]=\ln\Gamma\left(\mathrm{k}+\frac{1}{\mathrm{u}}\right)-$$

$$-\ln\Gamma(\mathrm{k})-\ln\Gamma\left(\frac{1}{\mathrm{u}}\right)+kM(\ln z)+\left(\frac{1}{u}-1\right)M[\ln(1-z)]\ . \quad (17)$$

Найдем уравнения правдоподобия:

$$\begin{cases} \dfrac{\partial\ln L_{(z)}}{\partial k}=\psi\left(k+\dfrac{1}{u}\right)-\psi(k)+M(\ln z)=0 \\ \dfrac{\partial\ln L_{(z)}}{\partial u}=\psi\left(k+\dfrac{1}{u}\right)\left(-\dfrac{1}{u^2}\right)-\psi\left(\dfrac{1}{u}\right)\left(-\dfrac{1}{u^2}\right)+ \\ +\left(-\dfrac{1}{u^2}\right)M[\ln(1-z)]=0. \end{cases}$$

Из первого уравнения правдоподобия имеем:

$$M(\ln z)=\psi(k)-\psi\left(k+\frac{1}{u}\right). \quad (18)$$

Из второго уравнения правдоподобия находим:

$$M[\ln(1-z)]=\psi\left(\frac{1}{u}\right)-\psi\left(k+\frac{1}{u}\right). \quad (19)$$

РЕПОЗИТОРИЙ БГУКИ

Логарифмическая функция правдоподобия (17) с учетом формул (18) и (19) перепишется в виде:

$$\ln L_{(z)} = M[\ln zp(z)] = \ln\frac{\Gamma(k+1/u)}{\Gamma(k)\Gamma(1/u)} + \\ +k\left[\psi(k)-\psi\left(k+\frac{1}{u}\right)\right]+\left(\frac{1}{u}-1\right)\left[\psi\left(\frac{1}{u}\right)-\psi\left(k+\frac{1}{u}\right)\right]. \quad (20)$$

Из сопоставления функций правдоподобия (14) и (20) следует, что первая из них зависит от трех параметров β, k, u, а вторая – только от двух параметров k, u.

На основании формул (14) и (20) можем записать равенство

$$M[\ln p(x)] = \ln\beta + M[\ln zp(z)], \quad (21)$$

которое справедливо также для других типов распределений.

Равенство (21) позволяет вычислять оценку наибольшего правдоподобия параметра β при известных оценках двух параметров формы k, u:

$$\beta = e^{M[\ln p(x)]-M[\ln zp(z)]} = \frac{e^{M[\ln p(x)]}}{e^{M[\ln zp(z)]}} = \frac{\overline{p(x)}_{геом.}}{\overline{\mathrm{zp(z)}}_{геом.}}. \quad (22)$$

Перепишем далее третье уравнение правдоподобия системы (13) с учетом формулы (18):

$$\beta M(x) = M(\ln z) - \ln\alpha u .$$

Отсюда найдем оценку наибольшего правдоподобия произведения αu:

$$\alpha u = e^{M(\ln z)-\beta M(x)}. \quad (23)$$

Входящие в формулы (22), (23) величины *M(lnzp(z))* и *M(lnz)* зависят от двух параметров формы k, u. Они вычисляются по формулам (20) и (18). Другие величины – *M(lnp(x))* и *M(x)* – в общем случае зависят от четырех параметров, но эти величины вычисляются по статистическому распределению.

Переходим к первому этапу решения.

Итак, при известных оценках параметров формы оценки наибольшего правдоподобия параметра β и произведения параметров αu вычисляются по формулам (22) и (23). Остается невыясненным вопрос о вычислении типа аппроксимирующего распределения и оценок параметров k, u по двум показателям, зависящим от этих параметров. Нахождение таких показателей является наиважнейшей задачей любого метода оценивания. Успешное решение этой задачи позволяет отказаться от выдвижения гипотез о виде аппроксимирующего распределения и проверки каждой из них по критериям согласия. **Наличие таких показателей позволяет вычислять наилучшее аппроксимирующее распределение из трех систем непрерывных распределений автора без выдвижения гипотез,** легко решать и другие задачи.

В случае метода наибольшего правдоподобия такие показатели можно получить из равенства (21) в виде центральных моментов второго – четвертого порядков

$$\mu_r = M\left[\ln p(x) - M\left(\ln p(x)\right)\right]^r = \\ = M\left[\ln zp(z) - M\left(\ln zp(z)\right)\right]^r , \quad (24)$$

которые зависят от двух параметров формы k, u [1].

В качестве первого показателя целесообразно принять центральный момент второго порядка. В качестве второго показателя могут быть приняты либо центральный момент третьего порядка, либо разность $\Delta = \ln M[p(x)] - M[\ln p(x)]$. Но решение этой задачи еще требует серьезных исследований.

Если искомые показатели найдены, необходимо при заданных значениях параметров формы построить бинарную сетку (номограмму) зависимости одного показателя от другого. С помощью построенной номограммы можно решать задачу установления типа аппроксимирующего распределения и нахождения в первом приближении оценок двух параметров формы в ручном режиме. Оценки параметров находятся по номограмме при известных значениях двух показателей, вычисленных по статистическому распределению.

В заключение следует отметить, что хорошие показатели должны обеспечить построение номограммы с высокой разрешающей способностью.

УСТОЙЧИВЫЙ МЕТОД

Устойчивым называется метод оценивания параметров, который не чувствителен к выбросам на концах статистического распределения.

Рассмотрим плотности (8) и (16). Случайные величины X и Z связаны соотношением (15)

$$Z = \alpha u e^{\beta X} .$$

Базой устойчивого метода является равенство, устанавливающее взаимосвязь между плотностями *p(x)* и *p(z)*. Найдем его.

Поскольку $p(z) = p(x)(dx/dz)$, $dx/dz = 1/\beta z$, то $p(z) = p(x)/\beta z$, откуда и следует искомое равенство

$$p(x) = \beta z p(z) . \quad (25)$$

Запишем на основе (25) новые равенства [1]

$$M[p(x)]^r = \beta^r M[zp(z)]^r \quad (26)$$

или

$$S_r^{(x)} = \beta^r S_r^{(z)} . \quad (27)$$

При известных оценках параметров k, u оценка параметра β устойчивого метода задается равенством (при r=1 в формулах (26), (27))

$$\beta = \frac{M[p(x)]}{M[zp(z)]} = \frac{S_1^{(x)}}{S_1^{(z)}} . \quad (28)$$

Здесь математическое ожидание плотности p(x) заменяется средним значением, которое вычисляется по статистическому распределению.

Логарифмируя равенство (25) и переходя к математическим ожиданиям, получим формулу (21), которая является базой метода наибольшего правдоподобия.

Формула (22) дает оценку наибольшего правдоподобия параметра β как **отношение средних геометрических** значений величин *p(x)* и *zp(z)*, а формула (28) – как **отношение средних арифметических** тех же величин в случае устойчивого метода. Отсюда следует, что оценки параметра β, вычисленные по обоим методам, должны быть одинаковыми. Оценка

РЕПОЗИТОРИЙ БГУКИ

параметра α (или произведения αu) вычисляется в обоих методах по одним и тем же формулам.

Итак, устойчивый метод близок к методу наибольшего правдоподобия, но в то же время он значительно проще последнего. Устойчивый метод обладает тем несомненным преимуществом перед методом наибольшего правдоподобия, что для него разработаны два показателя (асимметрии $B = M[p(x)(x - M(x)]$ и островершинности $H = S_3 / S_1^3$), с помощью которых по заранее построенной бинарной сетке (номограмме) или соответствующей компьютерной программе автора легко вычисляются аппроксимирующие распределения и оценки двух параметров формы k, u [1, 2 – 5]. Для метода наибольшего правдоподобия такие показатели еще предстоит разработать.

С другой стороны, метод наибольшего правдоподобия позволяет легко выражать через параметры распределения математическое ожидание логарифма плотности распределения. Например, для плотности p(x) (см. формулу (8)) величина M[lnp(x)] задается формулой (14). Эту величину можно использовать как естественный критерий близости статистического распределения и вычисленного закона распределения. По найденным оценкам параметров β, k, u следует вычислить теоретическое значение величины M[lnp(x)] и сравнить его с эмпирическим значением $\overline{\ln p(x)}$, рассчитанным непосредственно по статистическому распределению. Оба значения должны практически совпадать.

ФУНКЦИЯ ПРАВДОПОДОБИЯ КАК КРИТЕРИЙ СОГЛАСИЯ

Итак, важнейшим свойством функции правдоподобия является то, что **только при точно установленном аппроксимирующем распределении функция правдоподобия, рассчитанная по оценкам его параметров, будет равна статистической функции правдоподобия, т.е. вычисленной по статистическому распределению.**

При неправильно выбранном теоретическом законе распределения функция правдоподобия, вычисленная по оценкам его параметров, которые в свою очередь вычислены по статистическому распределению, будет **меньше статистической функции правдоподобия, т.е. меньше максимального ее значения.**

К сожалению, метод наибольшего правдоподобия не дает рекомендаций по ***вычислению*** наилучшего аппроксимирующего распределения. Поэтому при использовании этого метода приходится выдвигать различные гипотезы и проверять их с помощью критериев согласия. Поскольку все неподходящие аппроксимирующие распределения будут иметь меньшие значения функции правдоподобия, чем вычисленное непосредственно по статистическому распределению, **в качестве критерия согласия целесообразно использовать степень близости теоретической и статистической функций правдоподобия.**

Рассмотрим пример.

Пусть показательный закон распределения (5) имеет параметр α=1. В этом случае $M(t) = 1$. Тогда логарифмическая функция правдоподобия будет равна

$$\ln L = M[\ln p(t)] = \ln \alpha - \alpha M(t) = -1.$$

Вычислим далее для показательного закона распределения некоторые числовые характеристики случайной величины T, необходимые для дальнейших расчетов: начальный момент второго порядка, дисперсию, среднее квадратическое отклонение, математическое ожидание логарифма случайной величины и математическое ожидание случайной величины в степени ½. При α=1 они равны:

$$M(t^2) = 2;\ D(t) = 1;\ S(t) = 1;$$

$$M(\ln t) = \psi(1) = -0{,}5772164;\ M(\sqrt{t}) = \sqrt{\pi}/2.$$

Теперь используем **метод выдвижения гипотез** для нахождения наилучшего аппроксимирующего распределения по приведенным числовым характеристикам случайной величины.

Рассмотрим три гипотезы. Запишем закон распределения и его логарифмическую функцию правдоподобия, вычисленную по числовым характеристикам показательного закона.

1. Закон Вейбулла $p(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^{\beta}}$ при β=1/2:

$$p(t) = \frac{\alpha}{2\sqrt{t}\, e^{\alpha\sqrt{t}}};$$

$$M[\ln p(t)] = -\ln\sqrt{\pi} - 0{,}5M(\ln t) - 1 = -1{,}283757.$$

2. Закон Вейбулла при β=2 (распределение Релея):

$$p(t) = \frac{2\alpha t}{e^{\alpha t^2}};$$

$$M[\ln p(t)] = \ln 2 - \ln M(t^2) + M(\ln t) - 1 = -1{,}577216.$$

3. Нормальный закон:

$$p(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-a)^2}{2\sigma^2}};$$

$$M[\ln p(t)] = -\ln\sigma - 0{,}5(\ln 2\pi + 1) = -1{,}418939.$$

Из полученных результатов следует, что наиболее близким к показательному распределению, для которого $M[\ln p(t)] = -1$, оказался закон Вейбулла с параметром β=1/2. Однако ни одна из выдвинутых гипотез не может быть принята, так как во всех трех случаях логарифмическая функция правдоподобия оказалась значительно меньше –1.

Приведенные расчеты свидетельствуют о неэффективности метода выдвижения гипотез. **Теоретический закон распределения необходимо вычислять по статистическому распределению.** Для решения этой задачи автором разработаны три системы непрерывных четырехпараметрических распределений, методы вычисления типа теоретического распределения и оценок его параметров (универсальный метод моментов и общий устойчивый метод), а также серия компьютерных программ под общим названием SNR (системы непрерывных распределений).

РЕПОЗИТОРИЙ БГУКИ

ЭНТРОПИЯ

В качестве меры неопределенности системы принята энтропия. В случае непрерывных распределений она представляет собой математическое ожидание логарифма плотности, взятое с обратным знаком. Другими словами, энтропия – это взятая с обратным знаком логарифмическая функция правдоподобия (7):

$$H_x = -M\left[\ln p(x)\right]. \qquad (29)$$

Рассмотрим единицы измерения энтропии. В приведенной формуле логарифм плотности взят по основанию e=2.71828…. В данном случае в качестве единицы измерения энтропии принят «нат». При основании 10 единица измерения энтропии называется «дит», а при основании 2 – «бит».

С энтропией связано понятие информации. Количество полученной информации о системе уменьшает энтропию системы на это количество информации. Если о системе известно все, то количество информации равно энтропии этой системы, т.е. $I_x = H_x$ [5].

Рассмотрим один способ извлечения информации из ранговых распределений. Пусть ранговое распределение задано плотностью

$$p(t) = \frac{\beta\alpha^k}{\Gamma(\mathrm{k})} t^{k\beta-1} e^{-\alpha t^\beta}, \qquad (30)$$

которая относится ко второму типу второй системы непрерывных распределений автора. Запишем логарифмическую функцию правдоподобия:

$$\ln L = M\left[\ln p(t)\right] = \ln\beta + k\ln\alpha - \ln\Gamma(\mathrm{k}) + (\mathrm{k}\beta\text{-}1)\mathrm{M}(\ln t) - \alpha\mathrm{M}(t^\beta). \qquad (31)$$

Для того чтобы выразить ее через параметры распределения, найдем уравнения правдоподобия:

$$\frac{\partial \ln L}{\partial \alpha} = \frac{k}{\alpha} - M\left(t^\beta\right) = 0;$$
$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\beta} + kM(\ln t) - \alpha M\left(t^\beta \ln t\right) = 0;$$
$$\frac{\partial \ln L}{\partial k} = \ln\alpha - \psi(k) + \beta M(\ln t) = 0.$$

Из первого и последнего уравнений правдоподобия имеем:

$$k = \alpha M\left(t^\beta\right), \qquad (32)$$

$$M(\ln t) = \frac{1}{\beta}\left[\psi(k) - \ln\alpha\right]. \qquad (33)$$

Подставляя значения величин k, M(lnt) в формулу (31), найдем:

$$\ln L = M\left[\ln p(t)\right] = \ln\beta - \ln\Gamma(\mathrm{k}) + \mathrm{k}\psi(\mathrm{k}) - \mathrm{k} - \mathrm{M}(\ln t). \qquad (34)$$

Следовательно, энтропия плотности (30) равна

$$H_{p(t)} = -M\left[\ln p(t)\right] = M(\ln t) + \ln\Gamma(\mathrm{k}) + \mathrm{k} - \mathrm{k}\psi(\mathrm{k}) - \ln\beta. \qquad (35)$$

Пусть параметры рангового распределения (30) равны: α=0,5; β=0,4; k=2. Вычислим его энтропию. Для этого найдем вначале M(lnt) по формуле (33):

$$M(\ln t) = \frac{1}{\beta}\left[\psi(k) - \ln\alpha\right] = \frac{1}{0,4}\left[\psi(2) - \ln 0,5\right] = 2,7898287.$$

Здесь значение пси-функции Ψ(2)=0,4227843 взято из таблицы в [5, с. 53]. Тогда

$$H_{p(t)} = -M\left[\ln p(t)\right] = 2,7898287 + \ln\Gamma(2) + 2 - 2\psi(2) - \ln 0,4 = 4,860551 \text{ (нат)}.$$

Преобразуем далее плотность (30) к форме соответствующей плотности p(x) первой системы непрерывных распределений, т.е. представим плотность p(t) в виде tp(t)=f(lnt) (где tp(t)=p(x), lnt=x):

$$tp(t) = \frac{\beta\alpha^k}{\Gamma(\mathrm{k})} e^{k\beta\ln t} e^{-\alpha e^{\beta\ln t}}. \qquad (36)$$

Запишем для нее логарифмическую функцию правдоподобия:

$$\ln L = M\left[\ln tp(t)\right] = \ln\beta + k\ln\alpha - \ln\Gamma(\mathrm{k}) + \mathrm{k}\beta\mathrm{M}(\ln t) - \alpha\mathrm{M}\left(e^{\beta\ln t}\right). \qquad (37)$$

Выраженная через параметры распределения, она равна:

$$\ln L = M\left[\ln tp(t)\right] = \ln\beta - \ln\Gamma(\mathrm{k}) + \mathrm{k}\psi(\mathrm{k}) - \mathrm{k}. \qquad (38)$$

Следовательно, энтропия плотности (36) равна:

$$H_{tp(t)} = -M\left[\ln tp(t)\right] = \ln\Gamma(\mathrm{k}) + \mathrm{k} - \mathrm{k}\psi(\mathrm{k}) - \ln\beta. \qquad (39)$$

Сравнивая это равенство с (35), имеем:

$$H_{tp(t)} = H_{p(t)} - M(\ln t), \qquad (40)$$

т.е. энтропия (степень неопределенности) плотности tp(t)=p(lnt)=p(x) оказалась меньше энтропии плотности p(t) на величину M(lnt)=2,7898287 и составила $H_{tp(t)} = 2,0707223$ (нат), т.е. меньше энтропии $H_{p(t)}$ в 2,347 раза.

Таким образом, приведение второй системы непрерывных распределений к форме первой системы уменьшает энтропию второй системы.

Аналогично, приведение третьей системы непрерывных распределений к форме второй (или первой) системы также уменьшает ее энтропию.

Что касается ранговых (убывающих) распределений, то здесь наиболее ярко проявляется уменьшение энтропии, или появление новой информации в виде трех характерных точек: моды и двух точек перегиба, – которые позволяют объективно разделить ранговое распределение на ядро и три зоны рассеяния [3, 4]. Действительно, убывающее ранговое распределение, будучи представленным в виде графика зависимости tp(t)=f(lnt), превращается в одновершинную кривую распределения с тремя характерными точками, которые нельзя обнаружить непосредственно на убывающей кривой.

Два метода оценивания параметров (универсальный метод моментов и общий устойчивый метод) разработаны автором для первой системы непрерывных распределений, заданной плотностью p(x), а другие системы непрерывных распределений при нахождении оценок параметров по этим методам приводятся к первой системе. Это преобразование уменьшает энтропию распределений второй и третьей систем и позволяет находить оценки их параметров методом, пригодным для первой системы непрерывных распределений.

РЕПОЗИТОРИЙ БГУКИ

Действительно, метод моментов не может быть применен непосредственно к плотности (30), где значения случайной величины T возводятся в степень β. Но после преобразования той же плотности к виду (36) параметр β уже не является степенью случайной величины T, при этом вычисляются моменты не самой случайной величины T, а ее логарифма. Фактически в этом случае находятся оценки параметров плотности:

$$p(x) = \frac{\beta\alpha^k}{\Gamma(\mathrm{k})} e^{k\beta x} e^{-\alpha e^{\beta x}},$$

где x=lnt, p(x)=tp(t)=p(lnt) [3]. Найденные оценки являются также оценками параметров исходной плотности p(t), которая задана формулой (30).

ЗАКЛЮЧЕНИЕ

В статье рассмотрены четыре формы задания функции правдоподобия. Для проведения научных исследований введена пятая форма как математическое ожидание логарифма плотности распределения.

Дана сравнительная характеристика двух методов оценивания параметров четырехпараметрических непрерывных распределений: метода наибольшего правдоподобия Р. Фишера и устойчивого метода автора. Показано, что оба метода базируются на одном и том же равенстве

$$p(x) = \beta z p(z),$$

где плотность p(x) зависит от четырех параметров, а плотность p(z) – от двух.

В методе наибольшего правдоподобия используется логарифмическая форма приведенного равенства, где в качестве логарифмической функции правдоподобия используется математическое ожидание логарифма плотности:

$$M[\ln p(x)] = \ln\beta + M[\ln zp(z)].$$

Устойчивый метод позволяет вычислять закон распределения на базе четырехпараметрических систем непрерывных распределений с помощью показателей асимметрии $B = M[p(x)(x - M(x)]$ и островершинности $H = S_3 / S_1^3$. Оценки последних находятся по статистическому распределению и зависят от двух параметров формы. Тип распределения и оценки параметров формы находятся по заранее построенной бинарной сетке (номограмме) [3, 4] либо рассчитываются по соответствующей компьютерной программе автора. При этом подходящая система непрерывных распределений выбирается в зависимости от свойств случайной величины: это либо первая система, либо вторая и реже – третья система.

В методе наибольшего правдоподобия такие показатели отсутствуют. Поэтому **вид аппроксимирующего распределения подбирается традиционным методом – путем выдвижения гипотез**.

Логарифмическая функция правдоподобия, заданная формулой $\ln L = M[\ln p(x)]$ и взятая с обратным знаком, представляет собой энтропию. При преобразовании распределений второй системы к форме первой системы энтропия плотности уменьшается, т.е. появляется новая информация.

СПИСОК ЛИТЕРАТУРЫ

1. Нешитой В. В. Элементы теории обобщенных распределений. – Минск : РИВШ, 2009. – 204 с.
2. Нешитой В. В. Исследование статистических закономерностей текста и информационных потоков : дис. … д-ра техн. наук. – Минск, 1987. – 505 с.
3. Нешитой В. В. Законы Ципфа, Бредфорда и универсальные модели // НТИ. Сер. 2. – 2010. – №1. – С. 26 – 33; Neshitoi V. V. Zipf's and Bradford's laws and universal models // Automatic Documentation and Mathematical Linguistics. – 2010. – Vol. 44, №1. – P. 30 – 37.
4. Нешитой В. В. Методы статанализа в библиотечной деятельности: вычисление непрерывных распределений : учеб.-метод. пособие. – Минск : БГУ культуры и искусств, 2010. – 61 с.
5. Вентцель Е. С. Теория вероятностей. – М. : Физматгиз, 1969. – 576 с.

Материал поступил в редакцию 20.10.11.

Сведения об авторе

НЕШИТОЙ Василий Васильевич – доктор технических наук, профессор, заведующий кафедрой информационных ресурсов УО «Белорусский государственный университет культуры и искусств»

E-mail: neshitoi_vv@tut.by

РЕПОЗИТОРИЙ БГУКИ