Лекция 15

СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

ЦЕЛЬ ЛЕКЦИИ: ввести понятие оценки неизвестного параметра распределения и дать классификацию таких оценок; получить точечные оценки математического ожидания и дисперсии.

На практике в большинстве случаев закон распределения случайной величины X неизвестен, и по результатам наблюдений $x_1, x_2, \ldots, x_n$ необходимо оценить числовые характеристики (например, математическое ожидание, дисперсию или другие моменты) или неизвестный параметр a, который определяет закон распределения (плотность распределения) $f(x,a)$ изучаемой случайной величины. Так, для показательного распределения или распределения Пуассона достаточно оценить один параметр, а для нормального распределения подлежат оценке уже два параметра – математическое ожидание и дисперсия.

Виды оценок

Случайная величина X имеет плотность вероятности $f(x,a)$, где a – неизвестный параметр распределения. В результате эксперимента получены значения этой случайной величины: $x_1, x_2, \ldots, x_n$. Произвести оценку по существу означает, что выборочным значениям случайной величины необходимо поставить в соответствие некоторое значение параметра a, т. е. создать некоторую функцию результатов наблюдений $u(x_1, x_2, \ldots, x_n)$, значение которой принимается за оценку $\hat{a}_n$ параметра a. Индекс n указывает на количество проведенных опытов.

Любая функция, зависящая от результатов наблюдений, называется статистикой. Так как результаты наблюдений являются случайными величинами, то и статистика тоже будет случайной величиной. Следовательно, оценку $\hat{a}_n = u(x_1, x_2, \ldots, x_n)$ неизвестного параметра a следует рассматривать как случайную величину, а ее значение, вычисленное по экспериментальным данным объемом n, – как одно из возможных значений этой случайной величины.

Оценки параметров распределений (числовых характеристик случайной величины) подразделяются на точечные и интервальные. Точечная оценка параметра a определяется одним числом $\hat{a}_n$, и ее точность характеризуется дисперсией оценки. Интервальной оценкой называют оценку, которая определяется двумя числами, $\hat{a}_1$ и $\hat{a}_2$ – концами интервала, накрывающего оцениваемый параметр a с заданной доверительной вероятностью.

Классификация точечных оценок

Чтобы точечная оценка неизвестного параметра $\hat{a}_n = u(x_1, x_2, \ldots, x_n)$ была наилучшей с точки зрения точности, необходимо, чтобы она была состоятельной, несмещенной и эффективной.

Состоятельной называется оценка $\hat{a}_n = u(x_1, x_2, \ldots, x_n)$ параметра a, если она сходится по вероятности к оцениваемому параметру, т. е.

$$\hat{a}_n \underset{n\to\infty}{\overset{p}{\longrightarrow}} a. \tag{8.8}$$

На основании неравенства Чебышева можно показать, что достаточным условием выполнения соотношения (8.8) является равенство

$$\lim_{n\to\infty} M[\hat{a}_n] = a.$$

Состоятельность является асимптотической характеристикой оценки при $n \to \infty$.

Несмещенной называется оценка $\hat{a}_n = u(x_1, x_2, \ldots, x_n)$ (оценка без систематической ошибки), математическое ожидание которой равно оцениваемому параметру, т. е.

$$M[\hat{a}_n] = a. \tag{8.9}$$

Если равенство (8.9) не выполняется, то оценка называется смещенной. Разность $\Delta a = M[\hat{a}_n] - a$ называется смещением или систематической ошибкой оценки. Если же равенство (8.9) выполняется лишь при $n \to \infty$, то соответствующая оценка называется асимптотически несмещенной.

Необходимо отметить, что если состоятельность – практически обязательное условие всех используемых на практике оценок (несостоятельные оценки используются крайне редко), то свойство несмещенности является лишь желательным. Многие часто применяемые оценки свойством несмещенности не обладают.

В общем случае точность оценки некоторого параметра a, полученная на основании опытных данных $x_1, x_2, \ldots, x_n$, характеризуется средним квадратом ошибки

$$\varepsilon^2 = M[(\hat{a}_n - a)^2],$$

который можно привести к виду

$$\varepsilon^2 = D[\hat{a}_n] + (\Delta a)^2,$$

где $D[\hat{a}_n] = M[(\hat{a}_n - M[\hat{a}_n])^2]$ – дисперсия, $(\Delta a)^2 = (M[\hat{a}_n] - a)^2$ – квадрат смещения оценки.

Если оценка несмещенная, то

$$\varepsilon^2 = D[\hat{a}_n] = M[(\hat{a}_n - a)^2].$$

При конечных n оценки могут различаться средним квадратом ошибки ε^2. Естественно, что, чем меньше эта ошибка, тем теснее группируются значения оценки около оцениваемого параметра. Поэтому всегда желательно, чтобы ошибка оценки была по возможности наименьшей, т. е. выполнялось условие

$$\varepsilon^2 = M[(\hat{a}_n - a)^2] = \min_{\hat{a}_n}. \tag{8.10}$$

Оценку $\hat{a}_n$, удовлетворяющую условию (8.10), называют оценкой с минимальным квадратом ошибки.

<u>Эффективной</u> называется оценка $\hat{a}_n = u(x_1, x_2, \ldots, x_n)$, для которой средний квадрат ошибки не больше среднего квадрата ошибки любой другой оценки, т. е.

$$M[(\hat{a}_n - a)^2] \le M[(\hat{a}_n^* - a)^2],$$

где $\hat{a}_n^*$ – любая другая оценка параметра a.

Известно, что дисперсия любой несмещенной оценки одного параметра a удовлетворяет неравенству Крамера – Рао

$$D[\hat{a}_n] \geq \frac{1}{M\left[\frac{\partial}{\partial a}\ln f(x_1,\ldots,x_n / a)\right]} = \frac{-1}{M\left[\frac{\partial^2}{\partial a^2}\ln f(x_1,\ldots,x_n / a)\right]},$$

где $f(x_1,\ldots,x_n / a)$ – условная плотность распределения вероятностей полученных значений случайной величины при истинном значении параметра a.

Таким образом, несмещенная оценка $\hat{a}_n = u(x_1, x_2, \ldots, x_n)$, для которой неравенство Крамера – Рао обращается в равенство, будет эффективной, т. е. такая оценка имеет минимальную дисперсию.

Точечные оценки математического ожидания и дисперсии

Если рассматривается случайная величина X, имеющая математическое ожидание m и дисперсию D, то оба эти параметра считаются неизвестными. Поэтому над случайной величиной X производится n независимых опытов, которые дают результаты: $x_1, x_2, \ldots, x_n$. Необходимо найти состоятельные и несмещенные оценки неизвестных параметров m и D.

В качестве оценок $\hat{m}$ и $\hat{D}$ обычно выбираются соответственно статистическое (выборочное) среднее значение и статистическая (выборочная) дисперсия:

$$\hat{m} = m^* = \frac{1}{n}\sum_{i=1}^{n} x_i; \tag{8.11}$$

$$\hat{D} = D^* = \frac{1}{n}\sum_{i=1}^{n}(x_i - m^*)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (m^*)^2. \tag{8.12}$$

Оценка математического ожидания (8.11) является состоятельной согласно закону больших чисел (теорема Чебышева):

$$m^* = \frac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow[n\to\infty]{p} m.$$

Математическое ожидание случайной величины m^*

$$M[m^*] = M\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n} M\left[\sum_{i=1}^{n} x_i\right] = \frac{1}{n} nm = m\,.$$

Следовательно, оценка m^* является несмещенной.

Дисперсия оценки математического ожидания:

$$D[m^*] = M[(m^* - m)^2] = M\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i - m\right)^2\right] =$$

$$= \frac{1}{n^2} M\left[\left(\sum_{i=1}^{n} (x_i - m)\right)^2\right] = \frac{1}{n^2} M\left[\sum_{i=1}^{n} (x_i - m)^2\right] = \frac{1}{n^2}(nD) = \frac{D}{n}$$

.

Если случайная величина X распределена по нормальному закону, то оценка m^* является также и эффективной.

Математическое ожидание оценки дисперсии D^*

$$M[D^*] = M\left[\frac{1}{n}\sum_{i=1}^{n} (x_i - m^*)^2\right] = \frac{1}{n} M\left[\sum_{i=1}^{n} (x_i - m^*)^2\right].$$

В то же время

$$\sum_{i=1}^{n} (x_i - m^*)^2 = \sum_{i=1}^{n} (x_i - m + m - m^*)^2 =$$

$$= \sum_{i=1}^{n} (x_i - m)^2 - 2(m^* - m)\sum_{i=1}^{n} (x_i - m) + \sum_{i=1}^{n} (m^* - m)^2 =$$

$$= \sum_{i=1}^{n} (x_i - m)^2 - 2(m^* - m)\, n(m^* - m) + n(m^* - m)^2 =$$

$$= \sum_{i=1}^{n} (x_i - m)^2 - n(m^* - m)^2\,.$$

Так как $M[(x_i - m)^2] = D$, а $M[(m^* - m)^2] = D/n$, то получаем

$$M[D^*] = \frac{1}{n}(nD - D) = \frac{n-1}{n} D\,. \qquad (8.13)$$

Таким образом, D^* – смещенная оценка, хотя является состоятельной и эффективной.

Из формулы (8.13) следует, что для получения несмещенной оценки D^* следует видоизменить выборочную дисперсию (8.12) следующим образом:

$$\overline{D}^* = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m^*)^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2 - (m^*)^2\right), \qquad (8.14)$$

которая считается "лучшей" по сравнению с оценкой (8.12), хотя при больших n эти оценки практически равны друг другу.

Методы получения оценок параметров распределения

Часто на практике на основании анализа физического механизма, порождающего случайную величину X, можно сделать вывод о законе распределения этой случайной величины. Однако параметры этого распределения неизвестны, и их необходимо оценить по результатам эксперимента, обычно представленных в виде конечной выборки $x_1, x_2, \ldots, x_n$. Для решения такой задачи чаще всего применяются два метода: метод моментов и метод максимального правдоподобия.

Метод моментов. Метод состоит в приравнивании теоретических моментов соответствующим эмпирическим моментам того же порядка.

Эмпирические начальные моменты k-го порядка определяются формулами:

$$\alpha_k^* = \frac{1}{n}\sum_{i=1}^{n}(x_i)^k ,$$

а соответствующие им теоретические начальные моменты k-го порядка – формулами:

$$\alpha_k = \sum_{i=1}^{n}(x_i)^k p(x,a) \text{ для дискретных случайных величин,}$$

$$\alpha_k = \int_{-\infty}^{\infty} x^k f(x,a)dx \text{ для непрерывных случайных величин,}$$

где a – оцениваемый параметр распределения.

Для получения оценок параметров распределения, содержащего два неизвестных параметра a_1 и a_2, составляется система из двух уравнений

$$\begin{cases} \alpha_1(\hat{a}_1, \hat{a}_2) = \alpha_1^*(\hat{a}_1, \hat{a}_2) \\ \mu_2(\hat{a}_1, \hat{a}_2) = \mu_2^*(\hat{a}_1, \hat{a}_2), \end{cases}$$

где μ_2 и μ_2^* – теоретический и эмпирический центральные моменты второго порядка.

Решением системы уравнений являются оценки $\hat{a}_1$ и $\hat{a}_2$ неизвестных параметров распределения a_1 и a_2.

Приравняв теоретический эмпирический начальные моменты первого порядка, получаем, что оценкой математического ожидания случайной величины X, имеющей произвольное распределение, будет выборочное среднее, т. е. $M[X] = m^* = 1/n \sum_{i=1}^{n} x_i$. Затем, приравняв теоретический и эмпирический центральные моменты второго порядка, получим, что оценка дисперсии случайной величины X, имеющей произвольное распределение, определяется формулой

$$D[X] = D^* = \frac{1}{n} \sum_{i=1}^{n} (x_i - m^*).$$

Подобным образом можно найти оценки теоретических моментов любого порядка.

Метод моментов отличается простотой и не требует сложных вычислений, но полученные этим методом оценки часто являются неэффективными.

Метод максимального правдоподобия. Метод максимального правдоподобия точечной оценки неизвестных параметров распределения сводится к отысканию максимума функции одного или нескольких оцениваемых параметров.

Пусть X – непрерывная случайная величина, которая в результате n испытаний приняла значения $x_1, x_2, \ldots, x_n$. Для получения оценки неизвестного параметра a необходимо найти такое значение $\hat{a}$, при котором вероятность реализации полученной выборки была бы максимальной. Так как $x_1, x_2, \ldots, x_n$ представляют собой взаимно независимые величины с одинаковой плотностью вероятности $f(x)$, то функцией правдоподобия называют функцию аргумента a:

$$L(x_1, x_2, \ldots, x_n; a) = f(x_1; a) \cdot f(x_2; a) \ldots f(x_n; a).$$

Оценкой максимального правдоподобия параметра a называется такое значение $\hat{a}$, при котором функция правдоподобия достигает максимума, т. е. является решением уравнения

$$\left.\frac{dL(a)}{da}\right|_{a=\hat{a}} = 0,$$

которое явно зависит от результатов испытаний $x_1, x_2, \ldots, x_n$.

Поскольку функции $L(a)$ и $\ln L(a)$ достигают максимума при одних и тех же значениях $\hat{a} = u(x_1, x_2, \ldots, x_n)$, то часто для упрощения расчетов используют логарифмическую функцию правдоподобия и ищут корень соответствующего уравнения

$$\left.\frac{d \ln L(a)}{da}\right|_{a=\hat{a}} = 0,$$

которое называется уравнением правдоподобия.

Если необходимо оценить несколько параметров $a_1, a_2, \ldots, a_k$ распределения $f(x; a_1, a_2, \ldots, a_k)$, то функция правдоподобия будет зависеть от этих параметров. Для нахождения оценок $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_k$ параметров распределения необходимо решить систему k уравнений правдоподобия

$$\left.\frac{d}{da} \ln L(a_1, a_2, \ldots, a_k,)\right|_{a_i=\hat{a}_i} = 0, i = 1, 2, \ldots, k.$$

Метод максимального правдоподобия дает состоятельные и асимптотически эффективные оценки. Однако получаемые методом максимального правдоподобия оценки бывают смещенными, и, кроме того, для нахождения оценок часто приходится решать достаточно сложные системы уравнений.