Теория вероятностей и статистика

Тема 6. Статистические оценки параметров распределения

Белов А.И.

Уральский федеральный университет

Екатеринбург, 2020

Точечные оценки

Необходимо определить значение неизвестного параметра θ распределения случайной величины X по выборке $x_1, x_2, \ldots, x_n$.

Определение

Функцию $\theta^* = \theta^*(x_1, x_2, \ldots, x_n)$ называют точечной оценкой (статистикой) параметра θ, если мы принимаем $\theta \approx \theta^*$.

Значения выборки x_k реализуют наблюдения случайной величины X в k-м эксперименте, так что можно считать, что

$$\theta^* = \theta^*(X_1, X_2, \ldots, X_n)$$

является случайной величиной, а именно функцией от случайных величин $X_1, X_2, \ldots, X_n$, где X_k распределены одинаково с X и независимы в совокупности.

Несмещенные оценки

Определение

Оценка θ^* называется несмещенной, если

$$M(\theta^*) = \theta.$$

В противном случае оценка называется смещенной.

Разность

$$d(\theta^*) = M(\theta^*) - \theta$$

называется смещением оценки θ^*.

Определение

Оценка θ^* называется асимптотически несмещенной, если

$$\lim_{n\to\infty} M(\theta^*) = \theta.$$

Состоятельные оценки

Определение

Оценка θ^* называется состоятельной, если она сходится по вероятности к оцениваемому параметру θ, т. е. для любого $\varepsilon > 0$

$$\lim_{n\to\infty} P(|\theta^* - \theta| < \varepsilon) = 1.$$

Всякая состоятельная оценка является асимптотически несмещенной (без док-ва).

Эффективные оценки

Пусть θ_1^* и θ_2^* — две несмещенные оценки для θ и объем выборки фиксирован.

Определение

Оценка θ_1^* называется более эффективной, чем оценка θ_2^*, если

$$D(\theta_1^*) < D(\theta_2^*).$$

Несмещенная оценка θ^* называется эффективной, если она имеет наименьшую дисперсию среди всех несмещенных оценок.

Выборочная средняя

Определение

Выборочной средней называют

$$\overline{x}^* = \frac{x_1 + \ldots + x_n}{n} = \frac{1}{n}\sum_{k=1}^{n} x_k.$$

Если $\widetilde{x}_1, \ldots, \widetilde{x}_m$ — варианты выборки, $n_1, \ldots, n_m$ — соответствующие частоты, то

$$\overline{x}^* = \frac{1}{n}\sum_{i=1}^{m} n_i \widetilde{x}_i.$$

Выборочная средняя как оценка математического ожидания

Теорема

Выборочная средняя является состоятельной несмещённой оценкой математического ожидания.

Без доказательства.

$$D(\overline{x}^*) = \frac{1}{n} D(X),\ \sigma(\overline{x}^*) = \frac{\sigma(X)}{\sqrt{n}}.$$

Согласно теореме Ляпунова оценка $\overline{x}^*$ асимптотически нормальна, т. е. при больших объемах выборки имеет распределение, близкое к нормальному с $M(\overline{x}^*) = M(X)$ и $\sigma(\overline{x}^*) \to 0$ при $n \to \infty$.

Выборочная медиана

Определение

Пусть выборка упорядочена по возрастанию, то есть

$$x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n.$$

Если объём выборки — нечетное число ($n = 2m + 1$), то выборочной медианой называют число $M_e^* = x_{m+1}$.

Если объём выборки — четное число ($n = 2m$), то выборочной медианой называют $M_e^* = \dfrac{x_m + x_{m+1}}{2}$.

Выборочная мода дискретной случайной величины

Определение

Если X — дискретная случайная величина с небольшим числом возможных значений, а $\widetilde{x}_1, \ldots, \widetilde{x}_m$ — варианты выборки, а $n_1, \ldots, n_m$ — соответствующие частоты, то выборочной модой называют варианту с наибольшей частотой, т. е. $M_0^* = \widetilde{x}_i$, где $n_i = \max\{n_1, \ldots, n_m\}$.

Если в статистическом распределении несколько подряд идущих вариант имеют наибольшую частоту, то в качестве выборочной моды берут их среднее арифметическое.

Выборочная мода непрерывной случайной величины

Определение

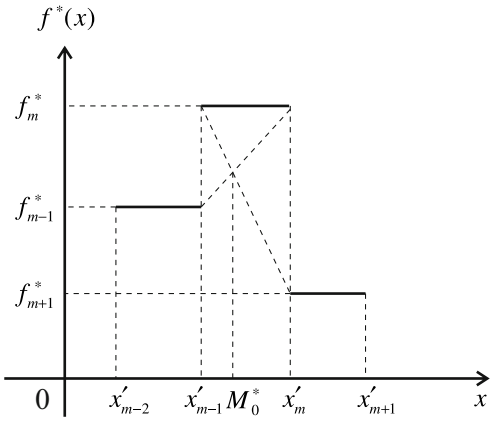
Пусть X — непрерывная случайная величина или дискретная случайная величина с большим числом возможных значений и $x'_0 < x'_1 < \ldots < x'_n$ — границы интервалов группировки.

Пусть также m такое, что $f^*_{m-1} < f^*_m > f^*_{m+1}$, где $f^*_k = f^*(\overline{x}_k)$, а $f^*(x)$ — эмпирическая функция распределения.

Тогда выборочной модой называют

$$M^*_0 = x'_{m-1} + \frac{f^*_m - f^*_{m-1}}{2f^*_m - f^*_{m-1} - f^*_{m+1}} \left(x'_m - x'_{m-1}\right).$$

Геометрический смысл выборочной моды непрерывной случайной величины



Выборочные дисперсия и среднее квадратическое отклонение

Определение

Выборочной дисперсией называют среднее арифметическое квадратов отклонений значений выборки от выборочного среднего:

$$D^* = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x}^*)^2.$$

Выборочным средним квадратическим отклонением называют

$$\sigma^* = \sqrt{D^*}.$$

Очевидно, что если $\widetilde{x}_1, \ldots, \widetilde{x}_m$ — варианты выборки, а $n_1, \ldots, n_m$ — соответствующие частоты, то

$$D^* = \frac{1}{n}\sum_{i=1}^{m} n_i(\widetilde{x}_i - \overline{x}^*)^2.$$

Формула вычисления выборочной дисперсии. Состоятельность выборочной дисперсии

Легко показать, что

$$D^* = \overline{(x^2)}^* - \overline{x}^{*2} = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - \overline{x}^{*2} = \frac{1}{n}\sum_{i=1}^{m} n_i \widetilde{x}_i^2 - \overline{x}^{*2}.$$

Теорема

Выборочная дисперсия является состоятельной оценкой дисперсии.

Без доказательства. □

Смещенность выборочной дисперсии

Теорема

Выборочная дисперсия является смещенной оценкой дисперсии со смещением

$$d(D^*) = -\frac{D(X)}{n}.$$

Выборочная дисперсия является асимптотически несмещенной оценкой дисперсии.

Доказательство теоремы о смещении выборочной дисперсии

Непосредственными выкладками получаем, что

$$M(D^*) = \frac{n-1}{n} D(X).$$

$$d(D^*) = M(D^*) - D(X) = \frac{n-1}{n} D(X) - D(X) = -\frac{D(X)}{n}.$$

Поскольку $\lim\limits_{n\to\infty} d(D^*) = \lim\limits_{n\to\infty} \left(-\frac{D(X)}{n}\right) = 0,$
то D^* — асимптотически несмещенная. □

Исправленные выборочные дисперсия и среднее квадратическое отклонение

Для получения несмещенной оценки дисперсии, что особенно важно для выборок небольшого объема, используют исправленную выборочную дисперсию

$$S^2 = \frac{n}{n-1} D^* = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x}^*)^2.$$

Для оценки среднего квадратического отклонения используют исправленное выборочное среднее квадратическое отклонение

$$S = \sqrt{\frac{n}{n-1} D^*} = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x}^*)^2}.$$

Эмпирический начальный момент k-го порядка

Определение

Эмпирическим (выборочным) начальным моментом k-го порядка называется

$$\nu_k^* = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

При $k = 1$ имеем $\nu_1^* = \overline{x}^*$.

Теорема

Эмпирический начальный момент k-порядка является несмещенной оценкой начального момента k-го порядка.

Эмпирический центральный момент k-го порядка

Определение

Эмпирическим (выборочным) центральным моментом k-го порядка называется

$$\mu_k^* = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}^*)^k.$$

При $k = 2$ имеем $\mu_2^* = D^*$.

Согласно теореме о смещении выборочной дисперсии это смещенная оценка.

Состоятельность эмпирических центральных моментов. Исправленные эмпирические центральные моменты

Теорема

Эмпирический центральный момент k-порядка является состоятельной и, следовательно, асимптотически несмещенной оценкой центрального момента k-го порядка.

Без доказательства. □

Исправленные эмпирические центральные моменты малых порядков:

$$m_2^* = S^2 = \frac{n}{n-1}\mu_2^*,$$

$$m_3^* = \frac{n^2}{(n-1)(n-2)}\mu_3^*,$$

$$m_4^* = \frac{n(n^2-2n+3)\mu_4^* - 3n(2n-3)(\mu_2^*)^2}{(n-1)(n-2)(n-3)}.$$

Оценка коэффициентов асимметрии и эксцесса

Смещенные и несмещенные оценки коэффициента симметрии

$$\gamma_1^* = \frac{\mu_3^*}{\sigma^{*3}}, \qquad g_1^* = \frac{m_3^*}{S^3}$$

Смещенные и несмещенные оценки коэффициента эксцесса

$$\gamma_2^* = \frac{\mu_4^*}{\sigma^{*4}} - 3, \qquad g_2^* = \frac{m_4^*}{S^4} - 3.$$

Метод моментов

Когда известен тип распределения исследуемой случайной величины, то встает вопрос об оценке параметров этого распределения.

Метод моментов состоит в том, что по выборке вычисляются эмпирические (или исправленные эмпирические) моменты, которые затем приравниваются к теоретическим моментам.

Как правило для известных типов распределений известны зависимости между параметрами распределения и центральными или начальными моментами.

Таким образом можно получить систему уравнений, из которых находятся значения параметров, которые принимаются в качестве точечных оценок параметров распределения.

Пример

Рассмотрим пример выборки 3, 2, 2, 3, 2, 3, 3, 3, 2, 2, 2, 3, 3, 2, 1, 1, 1, 2, 3, 3. Статистическое распределение

$\widetilde{x}_k$	1	2	3
n_k	3	8	9

.

Пусть известно, что исследуемая случайная величина распределена по биномиальному закону. Этот закон имеет два параметра — n и p.

Получим точечные оценки этих параметров методом моментов.

Для биномиально распределенной случайной величины X
$\nu_1 = np$ и $\mu_2 = npq = np(1-p)$.

$\nu_1^* = \frac{1}{20}(3 \cdot 1 + 8 \cdot 2 + 9 \cdot 3) = \frac{46}{20} = 2{,}3.$

$\mu_2^* = D^* = \frac{1}{20}(3 \cdot 1 + 8 \cdot 4 + 9 \cdot 9) - 2{,}3^2 = \frac{116}{20} - 5{,}29 = 0{,}51.$

Пример (окончание)

Объем выборки небольшой: $n = 20$.
Лучше использовать исправленный эмпирический центральный момент второго порядка $m_2^* = \frac{n}{n-1}\mu_2^* = \frac{20}{19} \cdot 0{,}51 \approx 0{,}537$.

Теперь решим систему

$$\begin{cases} n^*p^* &= 2{,}3 \\ n^*p^*(1-p^*) &= 0{,}537 \end{cases}$$

$1 - p^* = \frac{0{,}537}{2{,}3} \approx 0{,}233$;
$p^* = 0{,}767$;
$n^* = \text{Round}\left(\frac{2{,}3}{0{,}767}\right) = \text{Round}(3{,}129) = 3$,
где $\text{Round}(x)$ — функция округления x до ближайшего целого.

При использовании $\mu_2^* = 0{,}51$ получится $p^* = 0,778$, $n^* = 3$.

Теоретические значения $n = 3$, $p = 0{,}75$.

Надежность оценки

Определение

Пусть θ^* — точечная оценка параметра θ. Число $\varepsilon > 0$ такое, что $|\theta^* - \theta| < \varepsilon$ называется точностью оценки θ^*.

Пусть точность оценки ε зафиксирована. Тогда $|\theta^* - \theta| < \varepsilon$ — случайное событие и, следовательно, имеет некоторую вероятность.

Определение

Для заданной точности оценки $\varepsilon > 0$ надежностью или доверительной вероятностью называется

$$\gamma = P(|\theta^* - \theta| < \varepsilon).$$

Обычно γ выбирают близко к 1: 0,95; 0,99; 0,999.

Доверительный интервал

Если мы зафиксируем γ, то пусть $\varepsilon_\gamma > 0$ такое число, что

$$P(|\theta^* - \theta| < \varepsilon_\gamma) = \gamma.$$

$\theta \in (\theta^* - \varepsilon_\gamma, \theta^* + \varepsilon_\gamma)$ с вероятностью γ.

Определение

Интервал $(\theta^* - \varepsilon_\gamma, \theta^* + \varepsilon_\gamma)$ называется доверительным интервалом, который покрывает неизвестный параметр θ с заданной надежностью γ.

Границы доверительного интервала — случайные величины.

Интервальная оценка вероятности события по частоте

Рассмотрим схему Бернулли с параметрами n и p.
Пусть m — число успехов в серии из n испытаний.
Из закона больших чисел следует, что частота появления успеха $\frac{m}{n}$ является состоятельной и несмещённой оценкой вероятности успеха p.
Зафиксируем надежность $\gamma > 0$.
Возьмём $t > 0$ такое, что $2\Phi(t) = \gamma$, т. е. $t = \Phi^{-1}\left(\frac{\gamma}{2}\right)$.

Если $F(x)$ — функции нормального распределения с параметрами $a = 0$, $\sigma = 1$, то $t = F^{-1}\left(\frac{1+\gamma}{2}\right)$. В Microsoft Excel можно использовать НОРМ.СТ.ОБР().

Обозначим через w^* эмпирическое значение частоты $\frac{m}{n}$.

Интервальная оценка вероятности события по частоте (окончание)

Доверительный интервал вероятности события (p_1, p_2):

$$p_{1,2} = \frac{n}{t^2 + n}\left(w^* + \frac{t^2}{2n} \mp t\sqrt{\frac{w^*(1-w^*)}{n} + \left(\frac{t}{2n}\right)^2}\right).$$

Без доказательства.

Для больших выборок ($n > 100$) границы доверительного интервала можно вычислять по упрощенной формуле

$$p_{1,2} = w^* \mp t\sqrt{\frac{w^*(1-w^*)}{n}}.$$

Интервальная оценка a нормального распределения

Пусть $x_1, \ldots, x_2$ — выборка, $\overline{x}^*$ — выборочная средняя,
а S — исправленное выборочное среднее квадратическое отклонение.

Доверительный интервал для a

$$\left(\overline{x}^* - \frac{t_\gamma S}{\sqrt{n}}, \overline{x}^* + \frac{t_\gamma S}{\sqrt{n}}\right)$$

Значения t_γ можно найти по таблицам.

В Microsoft Excel можно найти $t_\gamma =$ СТЬЮДЕНТ.ОБР.2Х$(1-\gamma, n-1)$.

Интервальная оценка σ нормального распределения

Пусть $x_1, \ldots, x_n$ — выборка нормально распределенной случайной величины. В качестве точечной оценки σ возьмем исправленное выборочное среднее квадратическое отклонение S.

Доверительный интервал для σ:

$$(S(1-q), S(1+q)).$$

Значение q берут из таблиц.