

Статистические оценки параметров распределения

1. Основные понятия теории оценивания
2. Точечные оценки параметров распределения
3. Интервальные оценки параметров распределения
4. Выборочное наблюдение. Определение объема выборки

1.1 Основные понятия теории оценивания

Задача оценивания параметров распределения: по данным выборки x_1, x_2, \dots, x_n , полученным в результате n независимых наблюдений, найти *статистическую оценку* неизвестного параметра теоретического распределения – функцию от наблюдаемых случайных величин $\hat{\theta}(x_1, x_2, \dots, x_n)$.

Несмещенной называется статистическая оценка $\hat{\theta}$, математическое ожидание которой равно оцениваемому параметру θ при любом объеме выборки: $M(\hat{\theta}) = \theta$. В противном случае оценка называется *смещенной*.

Эффективной называется статистическая оценка, которая при заданном объеме выборки n имеет наименьшую возможную дисперсию.

Состоятельной называется статистическая оценка, которая при $n \rightarrow \infty$ стремится по вероятности к оцениваемому параметру:

$$\lim_{n \rightarrow \infty} P\{|\hat{\theta} - \theta| < \varepsilon\} = 1, \quad \forall \varepsilon > 0 \quad (1.1)$$

Если дисперсия несмещенной оценки при $n \rightarrow \infty$ стремится к нулю, то такая оценка оказывается и состоятельной.

1.2 Точность и доверительная вероятность оценки

Точечной называется оценка, определяемая одним числом.

Как называется метод для нахождения точечных оценок?

Интервальной называется оценка, определяемая концами интервала.

Пусть найденная по данным выборки статистическая характеристика $\hat{\theta}$ служит оценкой неизвестного параметра θ . Оценка $\hat{\theta}$ тем точнее определяет параметр θ , чем меньше разность $|\hat{\theta} - \theta|$.

Точность оценки характеризуется числом $\delta > 0$:

$$|\hat{\theta} - \theta| < \delta \quad (1.2)$$

Надежностью (доверительной вероятностью) оценки называется вероятность γ , с которой выполняется неравенство (1.2):

$$P(\hat{\theta} - \delta < \theta < \hat{\theta} + \delta) = \gamma \quad (1.3)$$

(1.3) определяет вероятность того, что интервал $(\hat{\theta} - \delta, \hat{\theta} + \delta)$ включает в себе (покрывает) неизвестный параметр θ .

Доверительным называется интервал $(\hat{\theta} - \delta, \hat{\theta} + \delta)$, который покрывает неизвестный параметр с заданной надежностью γ .

Обычно надежность оценки задается наперед, числом, близким к единице. Наиболее часто задают надежность 0,95; 0,99 и 0,999.

Доверительный интервал имеет случайные концы (доверительные границы). В разных выборках получаются различные значения $\hat{\theta}$, следовательно, от выборки к выборке будут изменяться и концы доверительного интервала, то есть доверительные границы являются случайными величинами – функциями от x_1, x_2, \dots, x_n .

Так как случайной величиной является не оцениваемый параметр θ , а доверительный интервал, то правильнее говорить не о вероятности попадания θ в доверительный интервал, а о вероятности того, что доверительный интервал покроем θ .

2. Точечные оценки параметров нормального распределения

1. Средняя \bar{X} , вычисленная по n независимым наблюдениям случайной величины X является *несмещенной* и *состоятельной* оценкой математического ожидания.

Если случайная величина X распределена нормально с параметрами $N(a, \sigma)$, то оценка \bar{X} математического ожидания $M(X)$ имеет минимальную дисперсию, равную

$$D(\bar{X}) = \frac{\sigma^2}{n} \quad (2.1)$$

следовательно, \bar{X} – *эффективная* оценка.

2. Если выборка содержит n независимых наблюдений случайной величины X с математическим ожиданием a и дисперсией σ^2 , то несмещенной оценкой дисперсии генеральной совокупности σ^2 является исправленная выборочная дисперсия:

$$s^2 = \frac{n}{n-1} D_B = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2 \quad (2.2)$$

3. Доверительные интервалы для оценки математического ожидания нормального распределения

а) среднее квадратическое отклонение σ известно

Рассмотрим выборочную среднюю \bar{x} как случайную величину \bar{X} , выборочные значения x_1, x_2, \dots, x_n – как одинаково распределенные независимые случайные величины X_1, X_2, \dots, X_n .

Если несколько случайных величин имеют одинаковые распределения, то их числовые характеристики одинаковы

Если случайная величина X распределена нормально, то и \bar{X} распределена нормально с параметрами $M(\bar{X}) = a$, $\sigma(\bar{X}) = \sigma/\sqrt{n}$:

$$M(\bar{X}) = M\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{M(X_1) + M(X_2) + \dots + M(X_n)}{n} = \frac{na}{n} = a$$

$$\sigma(\bar{X}) = \sqrt{D(\bar{X})} = \sqrt{D\left(\frac{X_1 + \dots + X_n}{n}\right)} = \sqrt{\frac{D(X_1) + \dots + D(X_n)}{n^2}} = \sqrt{\frac{n\sigma^2}{n^2}} = \frac{\sigma}{\sqrt{n}}$$

Пусть выполняется соотношение $P(|\bar{X} - a| < \delta) = \gamma$.

$$\begin{aligned} P(|\bar{X} - a| < \delta) &= P(-\delta < \bar{X} - a < \delta) = P(a - \delta < \bar{X} < a + \delta) = \\ &= \Phi\left(\frac{a + \delta - a}{\sigma(\bar{X})}\right) - \Phi\left(\frac{a - \delta - a}{\sigma(\bar{X})}\right) = 2\Phi\left(\frac{\delta}{\sigma(\bar{X})}\right) = 2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) = 2\Phi(t) \end{aligned}$$

$$t = \frac{\delta\sqrt{n}}{\sigma} \Rightarrow \delta = \frac{t\sigma}{\sqrt{n}}$$

Надежность оценки \bar{x} параметра a определяется по формуле

$$P\left(\bar{x} - \frac{t\sigma}{\sqrt{n}} < a < \bar{x} + \frac{t\sigma}{\sqrt{n}}\right) = 2\Phi(t) = \gamma \quad (3.1)$$

Доверительный интервал $(\bar{x} - t\sigma/\sqrt{n} < a < \bar{x} + t\sigma/\sqrt{n})$ с надежностью γ покрывает неизвестный параметр a . Точность оценки равна

$$\delta = \frac{t\sigma}{\sqrt{n}}$$

Число t определяется, исходя из равенства $2\Phi(t) = \gamma$, или $\Phi(t) = \gamma/2$.

Пример 3.1 Выборка из большой партии электроламп содержит 100 ламп. Средняя продолжительность горения лампы выборки оказалась равной 1000 ч. Найти с надежностью 0,95 доверительный интервал для средней продолжительности горения лампы всей партии, если известно, что среднее квадратическое отклонение продолжительности горения лампы равно 40 ч. Предполагается, что продолжительность горения ламп распределена нормально.

Решение Формализация условий задачи:

$$\bar{x} = 1000, \quad \sigma = 40, \quad n = 100, \quad \gamma = 0,95$$

Среднее квадратическое отклонение дано, поэтому воспользуемся формулой (3.1). Для надежности $\gamma = 0,95$ определим t : $t = 1,96$.

Точность оценки

$$\delta = \frac{t\sigma}{\sqrt{n}} = \frac{1,96 \cdot 40}{\sqrt{100}} \approx 7,84$$

$$\bar{x} - \delta \approx 1000 - 7,84 = 992,16, \quad \bar{x} + \delta \approx 1000 + 7,84 = 1007,84$$

Ответ: $992,16 < a < 1007,84$

б) среднее квадратическое отклонение σ неизвестно

По данным выборки можно построить случайную величину (ее возможные значения обозначены через t), которая имеет распределение Стьюдента с $k = n - 1$ степенями свободы:

$$T = \frac{\bar{x} - a}{s/\sqrt{n}}$$

\bar{x} – выборочная средняя, n – объем выборки, s – «исправленное» среднее квадратическое отклонение:

$$s = \sqrt{\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1}} \quad (3.2)$$

$$P\left(\bar{x} - \frac{t_\gamma s}{\sqrt{n}} < a < \bar{x} + \frac{t_\gamma s}{\sqrt{n}}\right) = \gamma \quad (3.3)$$

Значения $t_\gamma = t(\gamma, n)$ табулированы.

Пример 3.2 Из генеральной совокупности извлечена выборка объема $n = 10$. Найти интервальную оценку математического ожидания a нормально распределенного признака генеральной совокупности с надежностью 0,95.

x_i	-2	1	2	3	4	5
n_i	2	1	2	2	2	1

4.1 Выборочное наблюдение

Вся статистическая совокупность называется *генеральной*, а совокупность единиц, отобранных для обследования, называется *выборочной совокупностью* или *выборкой*.

Объемом совокупности (выборочной или генеральной) называется число объектов этой совокупности.

Выборка называется *повторной*, если отобранные единицы после обследования возвращаются в генеральную совокупность и снова могут попасть в выборку, или *бесповторной*, если отобранные единицы после обследования в генеральную совокупность не возвращаются.

Репрезентативность – свойство выборки объективно отображать характеристики генеральной совокупности. Выборка репрезентативна при выполнении следующих условий: 1) отбор объектов осуществляется случайным образом; 2) объем выборки достаточно велик.

Объем формируемой выборки определяется заданной точностью и зависит от ее повторности.

4.2 Определение объема выборки

Объем выборки, при которой ошибка не превысит заданную точность δ , рассчитывается по формулам (4.1) и (4.2):

1) для повторной выборки

$$n = \frac{t^2 \sigma^2}{\delta^2} \quad (4.1)$$

2) для бесповторной выборки

$$n = \frac{N t^2 \sigma^2}{N \delta^2 + t^2 \sigma^2} \quad (4.2)$$

σ – среднее квадратическое отклонение, δ – точность,

n – объем выборки, N – объем генеральной совокупности,

t – коэффициент кратности ошибки, определяемый по таблице значений функции Лапласа: $2\Phi(t) = \gamma$ ($\Phi(t) = \gamma/2$),

γ – доверительная вероятность

Величину средней ошибки простой случайной выборки при достаточно большом объеме выборки n можно определить по формуле:

для повторной выборки

$$\mu = \sqrt{\frac{\sigma^2}{n}}$$

для бесповторной выборки

$$\mu = \sqrt{\frac{\sigma^2}{n} \left(1 - \frac{n}{N}\right)}$$

$1 - n/N$ – доля единиц генеральной совокупности, не попавших в выборку.

Так как эта доля всегда меньше 1, то ошибка при бесповторном отборе при прочих равных условиях всегда меньше, чем при повторном. Если объем выборки мал по сравнению с объемом генеральной совокупности, выражением $(1-n/N)$ при вычислении средней ошибки выборки можно пренебречь

Академик А.М. Ляпунов доказал, что вероятность появления случайной ошибки выборки при достаточно большом ее объеме подчиняется нормальному закону распределения и определяется по формуле

$$F(t) = \frac{1}{\sqrt{2\pi}} \int_{-t}^t e^{-\frac{z^2}{2}} dz = \frac{2}{\sqrt{2\pi}} \int_0^t e^{-\frac{z^2}{2}} dz = 2\Phi(t)$$

t – коэффициент доверия (кратности ошибки), указывающий, во сколько раз предельная ошибка выборки δ превышает среднюю ошибку μ :

$$\delta = t\mu$$

$F(t)$ – вероятность того, что предельная ошибка выборки превосходит среднюю ошибку в t раз:

t	1	1,96	2	2,58	3
$F(t)$	0,683	0,95	0,954	0,99	0,997

Расчет предельной ошибки выборки, таким образом, осуществляется с определенной вероятностью, которой ставится в соответствие строго определенное значение коэффициента доверия.

Пример 4.1 Определить объем выборки, чтобы с вероятностью 0,95 точность при определении среднего размера деталей не превышала бы 0,4 мм. Среднее квадратическое отклонение генеральной совокупности размера деталей равно 2,5 мм.

Решение. Воспользуемся схемой повторной выборки.

$$\delta = 0,4$$

$$\sigma = 2,5$$

$$2\Phi(t) = 0,95 \Rightarrow \Phi(t) = 0,475$$

Значению 0,475 в таблице функции Лапласа соответствует значение аргумента $t = 1,96$

$$n = \frac{t^2 \sigma^2}{\delta^2} = \frac{2,5^2 \cdot 1,96^2}{0,4^2} = 150 \text{ (единиц)}$$

Ответ: объем выборки составляет 150 деталей.