

Пример 1

Вы экономист в государственной службе статистики. Ваша задача — оценить среднемесячный доход домохозяйства μ в крупном регионе с $N = 1.5$ млн домохозяйств.

Проблема: Провести сплошной опрос всех N домохозяйств дорого и долго. Вместо этого вы случайным образом отбираете $n = 1000$ домохозяйств и вычисляете выборочное среднее \bar{X} . Вопрос: насколько \bar{X} близко к истинному μ ?

Решение: Применяем Центральную Предельную Теорему (ЦПТ).

ЦПТ утверждает, что если X_1, X_2, \dots, X_n — независимые, одинаково распределенные случайные величины с конечным математическим ожиданием μ и дисперсией σ^2 , то при достаточно большом n распределение их выборочного среднего стремится к нормальному:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{d} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Или в стандартизированной форме:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

Распределение доходов X_i в генеральной совокупности не является нормальным. Оно асимметрично (скошено вправо). Однако, согласно ЦПТ, распределение выборочного среднего \bar{X} будет приблизительно нормальным.

Результаты выборки: Вы провели опрос и получили: Выборочное среднее: $\bar{x} = 60000$ руб. Известное из предыдущих исследований стандартное отклонение по совокупности: $\sigma = 25000$ руб. Объем выборки: $n = 1000$.

Найти интервал в котором с вероятностью больше 95 процентов лежит средний доход населения

$$\begin{aligned} P(\mu - k \leq \frac{1}{n} \sum_{i=1}^n X_i \leq \mu + k) &= P\left(\frac{-k}{\sigma} \sqrt{n} \leq \frac{\frac{1}{n} \sum_{i=1}^n X_i - \mu}{\sigma} \sqrt{n} \leq \frac{k}{\sigma} \sqrt{n}\right) = \\ &= \Phi\left(\frac{k}{\sigma} \sqrt{n}\right) - \Phi\left(\frac{-k}{\sigma} \sqrt{n}\right) = 2\Phi\left(\frac{k}{\sigma} \sqrt{n}\right) - 1 \geq 0.95 \Rightarrow \\ \Phi\left(\frac{k}{\sigma} \sqrt{n}\right) &\geq 0.975 \Rightarrow \frac{k}{\sigma} \sqrt{n} \geq 1.96 \Rightarrow k \geq 1550 \end{aligned}$$

$$P(58450 \leq \mu \leq 61550) \approx 0.95$$

Мы можем быть на 95% уверены, что истинный средний доход домохозяйств в регионе лежит в интервале от 58 450 до 61 550 рублей.

ЦПТ является фундаментальным инструментом в экономике и эконометрике, так как позволяет:

Использовать выборочные данные для оценки параметров генеральной совокупности.

Строить доверительные интервалы и проверять статистические гипотезы (t -тесты, ANOVA), даже если исходное распределение не является нормальным.

Оценки параметров. Свойства оценок

Пусть X_1, X_2, \dots, X_n - выборка из известного распределения $F(x, \theta)$, зависящего от неизвестного параметра θ .

Определение 1. $\hat{\theta}_n = f(X_1, X_2, \dots, X_n)$ - оценка параметра θ (функция от выборки, статистика)

Определение 2. Оценка $\hat{\theta}_n$ параметра θ называется несмещенной, если $M\hat{\theta}_n = \theta$.

1) Выборочное среднее \bar{X} является несмещенной оценкой математического ожидания.

2) Выборочная дисперсия S^2 является несмещенной оценкой дисперсии σ^2 .

Определение 3. Оценка $\hat{\theta}_n$ параметра θ называется состоятельной, если $\hat{\theta}_n \xrightarrow{P} \theta$

Примеры.

1) Выборочное среднее \bar{X} является состоятельной оценкой математического

ожидания в силу ЗБЧ (теорема Хинчина): $\frac{1}{n} \sum_{k=1}^n X_k \xrightarrow{P} MX_k$

2) Выборочная дисперсия S^2 является состоятельной оценкой дисперсии σ^2 , поскольку

$$S^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2 = \frac{n}{n-1} \frac{1}{n} \sum_{k=1}^n [(X_k)^2 - 2\bar{X}X_k + \bar{X}^2] = \frac{n}{n-1} \left[\frac{1}{n} \sum_{k=1}^n (X_k)^2 - \bar{X}^2 \right]$$

Первое слагаемое в квадратных скобках сходится по вероятности к MX_k^2 в силу ЗБЧ,

а второе по пункту 1) к $(MX_k)^2$. Следовательно, $\frac{1}{n} \sum_{k=1}^n (X_k)^2 - \bar{X}^2 \xrightarrow{P} MX_k^2 - (MX_k)^2 = DX_k$.

Поскольку $\frac{n}{n-1} \rightarrow 1$, то $S^2 \xrightarrow{P} DX_k$.

Определение 4. Оценка $\hat{\theta}_n$ параметра θ называется асимптотически несмещенной, если $M\hat{\theta}_n \rightarrow \theta$.

Теорема 2. Если оценка асимптотически несмещенная и ее дисперсия стремится к нулю, то она состоятельна.

Определение 5. Выборочной функцией распределения называется

$$\hat{F}_n(x) = \frac{1}{n} \sum_{k=1}^n I(x - X_k), \quad x \in \mathbb{R}, \quad I(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases},$$

То есть $\hat{F}(x) = \frac{m}{n}$, если ровно m наблюдаемых значений меньше x .

$\hat{F}_n(x)$ - статистика (при каждом x), то есть функция от выборки

Свойства эмпирической функции распределения.

1. $0 \leq \hat{F}_n(x) \leq 1$
2. $\hat{F}_n(x) \xrightarrow{n \rightarrow -\infty} 0, \hat{F}_n(x) \xrightarrow{n \rightarrow +\infty} 1$
3. $\hat{F}_n(x)$ не убывает
4. $\hat{F}_n(x)$ непрерывна слева

Теорема 3. Если взята выборка объема n из генеральной совокупности, имеющей функцию распределения $F(x)$, то $F_n(x)$ дискретная случайная величина, закон распределения которой имеет вид:

$$P\left(F_n(x) = \frac{k}{n}\right) = C_n^k (F(x))^k (1-F(x))^{n-k}, k = 0, 1, \dots, n$$

Следствия.

1. $MF_n(x) = M\left[\frac{1}{n} \sum_{k=1}^n I(x - X_k)\right] = \frac{1}{n} M\left[\sum_{k=1}^n I(x - X_k)\right] = \frac{1}{n} np = F(x)$
2. $DF_n(x) = D\left[\frac{1}{n} \sum_{k=1}^n I(x - X_k)\right] = \frac{1}{n^2} D\left[\sum_{k=1}^n I(x - X_k)\right] = \frac{1}{n^2} npq = \frac{F(x)(1-F(x))}{n}$
3. Если $X_1, X_2, \dots, X_n, \dots$ - выборка неограниченного объема, то $F_n(x) \xrightarrow{P} F(x)$.

Несмещённость: Представьте, что вы стреляете по мишени (истинное значение параметра). Несмещённость означает, что ваш прицел (метод оценки) не сбив. В среднем, после множества выстрелов (множества выборок), вы попадёте в "яблочко". На практике это значит, что у нас нет систематической ошибки: мы не завышаем и не занижаем результат постоянно. Например, если мы всегда недооцениваем инфляцию на 0.5%, наша оценка смещена.

Состоятельность: Это свойство про то, что происходит, когда мы собираем всё больше и больше данных. Состоятельная оценка — это та, которая становится тем точнее, чем больше данных у нас есть. Это как фокусировка объектива: при маленькой выборке изображение размыто, но с ростом данных оно становится чётким и сходится к истинному

значению. На практике это критически важно: мы уверены, что инвестиции в сбор больших данных окупятся повышением точности наших выводов.

Эффективность: Допустим, у нас есть две несмещённые оценки (обе в среднем попадают в цель). Эффективная оценка — это та, которая имеет меньший разброс (меньшую дисперсию). На графике с мишенями это более кучная группа попаданий. Для экономиста это значит: при прочих равных (и том же объёме данных) эффективная оценка даёт более точный и надёжный результат с меньшей ошибкой. Мы выбираем её, чтобы минимизировать риск ошибки в принятии решений.

Порядковые статистики.

Рассмотрим вариационный ряд $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$, построенный по выборке X_1, X_2, \dots, X_n из распределения $F(x)$.

Определение. Члены вариационного ряда называются порядковыми статистиками: $X_{(k)}$ - k -ая порядковая статистика

Теорема 5. Если независимая выборка взята из генеральной совокупности с функцией распределения $F(x)$, то функции распределения крайних членов вариационного ряда и их совместная функция распределения имеют вид:

$$F_{X_{(n)}}(x) = [F(x)]^n;$$

$$F_{X_{(1)}}(x) = 1 - [1 - F(x)]^n;$$

$$F_{X_{(1)}, X_{(n)}}(x, y) = [F(y)]^n - [F(y) - F(x)]^n, x < y.$$

Следствие. Если выборка взята из абсолютно непрерывного закона $F(x)$ с плотностью $p(x)$, то плотности распределения крайних членов вариационного ряда и их совместная плотность имеют вид:

$$p_{X_{(n)}}(x) = n[F(x)]^{n-1} F'(x) = n[F(x)]^{n-1} p(x)$$

$$p_{X_{(1)}}(x) = n[1 - F(x)]^{n-1} F'(x) = n[1 - F(x)]^{n-1} p(x),$$

$$p_{X_{(1)}, X_{(n)}}(x, y) = n(n-1)[F(y) - F(x)]^{n-2} p(x)p(y), x < y.$$

Теорема 6. . Если независимая выборка взята из генеральной совокупности с плотностью распределения $p(x)$, то плотность распределения k - ой порядковой статистики имеет вид:

$$p_{X_{(k)}}(x) = n C_{n-1}^{k-1} (F(x))^{k-1} [1 - F(x)]^{n-k} p(x).$$

Задачи для решения в аудитории.

Пример 1

Докажите, что оценка $\hat{\alpha} = X_{(1)}$ (первый элемент порядковой статистики) неизвестного параметра α , построенная по выборке X_1, X_2, \dots, X_n из распределения с плотностью $p(x, \alpha) = \begin{cases} 0 & x < \alpha \\ e^{\alpha-x} & x \geq \alpha \end{cases}$, является асимптотически несмещенной и состоятельной.

Решение. Функция распределения выборки $F(x, \alpha) = \begin{cases} 0 & x < \alpha \\ 1 - e^{\alpha-x} & x \geq \alpha \end{cases}$,

Математическое ожидание $M\xi = \int_{\alpha}^{+\infty} x e^{\alpha-x} dx = \int_0^{+\infty} (u + \alpha) e^{-u} du = \alpha$

По теореме плотность случайной величины $X_{(1)}$

$$p_{X_{(1)}}(x) = n [1 - F(x)]^{n-1} F'(x) = n [1 - F(x)]^{n-1} p(x),$$

Найдём мат ожидание, если оно стремится к параметру α , то оценка будет асимптотически несмещенной

$$MX_{(1)} = \int_{-\infty}^{+\infty} x \cdot n (1 - F(x))^{n-1} p(x) dx = \int_{\alpha}^{+\infty} x \cdot n (e^{\alpha-x})^{n-1} e^{\alpha-x} dx =$$

$$n \int_{\alpha}^{+\infty} x e^{n(\alpha-x)} dx = \left\{ \begin{array}{l} u = x - \alpha \\ du = dx \end{array} \right\} = n \int_0^{+\infty} (u + \alpha) e^{-nu} du$$

$$= \int_0^{+\infty} nu \cdot e^{-nu} du + \alpha \int_0^{+\infty} ne^{-nu} du = \frac{1}{n} \Gamma(2) + \alpha = \alpha + \frac{1}{n} \rightarrow \alpha$$

$$M(X_{(1)})^2 = \int_{-\infty}^{+\infty} x^2 \cdot n (1 - F(x))^{n-1} p(x) dx = \int_{\alpha}^{+\infty} x^2 \cdot n (e^{\alpha-x})^{n-1} e^{\alpha-x} dx =$$

$$n \int_{\alpha}^{+\infty} x^2 e^{n(\alpha-x)} dx = \left\{ \begin{array}{l} u = x - \alpha \\ du = dx \end{array} \right\} = n \int_0^{+\infty} (u + \alpha)^2 e^{-nu} du$$

$$= \int_0^{+\infty} nu^2 \cdot e^{-nu} du + 2\alpha \int_0^{+\infty} nue^{-nu} du + \alpha^2 \int_0^{+\infty} ne^{-nu} du =$$

$$= \frac{1}{n^2} \Gamma(3) + \frac{2\alpha}{n} \Gamma(2) + \alpha^2 = \frac{2}{n^2} + \frac{2\alpha}{n} + \alpha^2.$$

$$DX_{(1)} = \frac{2}{n^2} + \frac{2\alpha}{n} + \alpha^2 - \left(\alpha + \frac{1}{n} \right)^2 = \frac{1}{n^2} \rightarrow 0$$

По теореме 2 статистика $X_{(1)}$ является состоятельной оценкой параметра α .

Пример 2

Выборка X_1, \dots, X_n – имеет плотность распределения

$$p(x) = \begin{cases} p\lambda e^{-\lambda x} + \frac{1-p}{a}, & x \in (0; a) \\ p\lambda e^{-\lambda x}, & x \in (a; +\infty) \\ 0, & x \leq 0 \end{cases}$$

При заданных значениях параметров $\lambda = 1$ и $a = 3,2$ найдите оценку параметра p .

Таблица частот

интервалы	0-0,4	0,4-0,8	0,8-1,2	1,2-1,6	1,6-2	2-2,4	2,4-2,8	2,8-3,2	3,2-3,6	3,6-4
Частоты	217	175	151	122	96	86	78	57	11	7

Ответ. $1.27 = -0.6p + 1.6, \hat{p} = \frac{0.33}{0.6} = \frac{11}{20}.$

$$M\xi = \int_0^{+\infty} xp(x)dx = \frac{p}{\lambda} + a \frac{1-p}{2} = p + 1.6(1-p) = -0.6p + 1.6$$

$$\bar{X} = 10^{-4} (2 \cdot 217 + 6 \cdot 175 + 10 \cdot 151 + 14 \cdot 122 + 18 \cdot 96 + 22 \cdot 86 + \dots + 38 \cdot 7) = 1.27$$

Пример 3

По выборке X_1, X_2, \dots, X_5 из генеральной совокупности, распределенной по показательному закону с параметром λ , в качестве точечной оценки

математического ожидания используется $\hat{f}(\bar{X}_5) = \frac{1}{2}(X_{(1)} + X_{(5)})$. Докажите, что эта оценка смещенная.

Решение. Плотность показательного закона $p_\xi(x) = \lambda e^{-\lambda x}, x > 0$.

Функция распределения показательного закона $F(x) = 1 - e^{-\lambda x}, x > 0$.

Математическое ожидание $M\xi = \frac{1}{\lambda}$. Найдем $M\hat{f}(\overline{X}_5)$.

$$\begin{aligned} MX_{(1)} &= \int_{-\infty}^{+\infty} x \cdot n(1-F(x))^{n-1} f(x) dx = \int_0^{+\infty} x \cdot 5(e^{-\lambda x})^4 \lambda e^{-\lambda x} dx = \\ &= 5 \int_0^{+\infty} x e^{-5\lambda x} dx = \frac{5}{\lambda} \int_0^{+\infty} u e^{-5u} du = \frac{5}{\lambda} \left(-\frac{u}{5} e^{-5u} \Big|_0^{+\infty} + \frac{1}{5} \int_0^{+\infty} e^{-5u} du \right) = \\ &= \frac{1}{\lambda} \left(-\frac{1}{5} e^{-5u} \Big|_0^{+\infty} \right) = \frac{1}{5\lambda}. \end{aligned}$$

$$\begin{aligned} MX_{(5)} &= \int_{-\infty}^{+\infty} x \cdot 5(F(x))^4 f(x) dx = 5 \int_0^{+\infty} x (1 - e^{-\lambda x})^4 \lambda e^{-\lambda x} dx = \left\{ \begin{array}{l} u = \lambda x \\ du = \lambda dx \end{array} \right\} \\ &= \frac{5}{\lambda} \int_0^{+\infty} u (e^{-u} - 4e^{-2u} + 6e^{-3u} - 4e^{-4u} + e^{-5u}) du = \\ &= \frac{5}{\lambda} \left(\Gamma(2) - 4 \int_0^{+\infty} u e^{-2u} du + 6 \int_0^{+\infty} u e^{-3u} du - 4 \int_0^{+\infty} u e^{-4u} du + \int_0^{+\infty} u e^{-5u} du \right) = \\ &= \frac{5}{\lambda} \left(\Gamma(2) - \Gamma(2) + \frac{2}{3} \Gamma(2) - \frac{1}{4} \Gamma(2) + \frac{1}{25} \Gamma(2) \right) = \frac{5}{\lambda} \frac{200 - 75 + 12}{300} = \frac{137}{60\lambda} \end{aligned}$$

Найдём мат ожидание оценки

$$M\hat{f}(\overline{X}_5) = \frac{1}{2} \left(\frac{1}{5\lambda} + \frac{137}{60\lambda} \right) = \frac{149}{120\lambda} \neq \frac{1}{\lambda}.$$

Пример 4

Найдите математическое ожидание и дисперсию $X_{(k)}$, если выборка получена из равномерного распределения на $[0, a]$.

Решение.

$$MX_{(k)} = \int_0^a x \cdot n \cdot C_{n-1}^{k-1} \left(\frac{x}{a} \right)^{k-1} \left(1 - \frac{x}{a} \right)^{n-k} \frac{1}{a} dx = \left\{ \begin{array}{l} u = \frac{x}{a} \\ du = \frac{dx}{a} \end{array} \right\} = an C_{n-1}^{k-1} \int_0^1 u^k (1-u)^{n-k} du =$$

Свели вычисление интеграла в Бэта функции $B(m, n) = \int_0^1 w^{m-1} (1-w)^{n-1} dw$

$$= anC_{n-1}^{k-1}B(k+1, n-k+1) = anC_{n-1}^{k-1} \frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)} =$$

С учётом того, что k принимает целые значения, получаем

$$= an \frac{(n-1)!k!(n-k)!}{(k-1)!(n-k)!(n+1)!} = \frac{ak}{n+1}.$$

$$MX_{(k)}^2 = \int_0^a x^2 \cdot n \cdot C_{n-1}^{k-1} \left(\frac{x}{a}\right)^{k-1} \left(1 - \frac{x}{a}\right)^{n-k} \frac{1}{a} dx = a^2 C_{n-1}^{k-1} \int_0^1 nu^{k+1} (1-u)^{n-k} du =$$

$$= a^2 n C_{n-1}^{k-1} B(k+2, n-k+1) = a^2 n C_{n-1}^{k-1} \frac{\Gamma(k+2)\Gamma(n-k+1)}{\Gamma(n+3)} =$$

$$= a^2 n \frac{(n-1)!(k+1)!(n-k)!}{(k-1)!(n-k)!(n+2)!} = \frac{a^2 k(k+1)}{(n+1)(n+2)}.$$

$$DX_{(k)} = \frac{a^2 k(k+1)}{(n+1)(n+2)} - \left(\frac{ak}{n+1}\right)^2 = \frac{a^2 k((k+1)(n+1) - k(n+2))}{(n+1)^2(n+2)} = \frac{a^2 k(n-k)}{(n+1)^2(n+2)}.$$

Пример 5

Дана выборка объема n из распределения Парето с плотностью

$$p(x, \theta) = \begin{cases} \frac{2}{\theta} \left(\frac{\theta}{x}\right)^3 & x \geq \theta \\ 0 & x < \theta \end{cases}. \text{ В качестве оценки неизвестного параметра } \theta \text{ используется}$$

$$X_{(1)}.$$

Докажите, что эта оценка является асимптотически несмещенной.

$$\text{Функция распределения выборки } F(x, \alpha) = \begin{cases} 0 & x < \theta \\ 1 - \left(\frac{\theta}{x}\right)^2 & x \geq \theta \end{cases},$$

По теореме плотность случайной величины $X_{(1)}$

$$p_{X_{(1)}}(x) = n[1 - F(x)]^{n-1} F'(x) = n[1 - F(x)]^{n-1} p(x),$$

Найдём мат ожидание, если оно стремится к параметру α , то оценка будет асимптотически несмещенной

$$\begin{aligned}
 MX_{(1)} &= \int_{-\infty}^{+\infty} x \cdot n(1-F(x))^{n-1} p(x) dx = \int_{\theta}^{+\infty} x \cdot n \left(\left(\frac{\theta}{x} \right)^2 \right)^{n-1} \frac{2}{\theta} \left(\frac{\theta}{x} \right)^3 dx = \\
 &= \int_{\theta}^{+\infty} x \cdot n \left(\frac{\theta^{2n+1}}{x^{2n+1}} \right) \frac{2}{\theta} dx = 2n\theta^{2n} \int_{\theta}^{+\infty} \frac{1}{x^{2n}} dx = \\
 &= 2n\theta^{2n} \left(\frac{x^{-2n+1}}{(-2n+1)} \Big|_{\theta}^{+\infty} \right) = \frac{2n\theta^{2n}\theta^{-2n+1}}{(2n-1)} = \frac{2n}{(2n-1)}\theta \rightarrow \theta
 \end{aligned}$$

Пример 6

Постройте по методу моментов оценку неизвестного параметра θ по выборке

$$X_1, X_2, \dots, X_n \text{ из распределения с плотностью } p(x, \theta) = \begin{cases} \frac{2}{\lambda\sqrt{\pi}} e^{-\frac{(x-1)^2}{\theta^2}} & x > 1 \\ 0 & x \leq 1 \end{cases}$$

Ответ. $\hat{\lambda} = \sqrt{\pi}(\bar{X} - 1)$.

Handwritten solution for Example 6:

$$\begin{aligned}
 p(x, \lambda) &= \frac{2}{\lambda\sqrt{\pi}} e^{-\frac{(x-1)^2}{\lambda^2}}, \quad x > 1 \\
 M_1 &= \int_1^{+\infty} \frac{x \cdot 2}{\lambda\sqrt{\pi}} e^{-\frac{(x-1)^2}{\lambda^2}} dx = \int_0^{+\infty} \left(\frac{(x-1)^2}{\lambda^2} = t \right) \left(\frac{2(x-1)}{\lambda^2} dx = dt \right) = \\
 &= \int_0^{+\infty} \frac{2x-2}{\lambda\sqrt{\pi}} e^{-\frac{(x-1)^2}{\lambda^2}} dx + \int_0^{+\infty} \frac{2}{\lambda\sqrt{\pi}} e^{-\frac{(x-1)^2}{\lambda^2}} dx = \int_0^{+\infty} \frac{\lambda}{\sqrt{\pi}} e^{-t} dt + \\
 &+ \int_0^{+\infty} \frac{1}{\sqrt{\pi}} \cdot \frac{1}{\sqrt{t}} e^{-t} dt = \frac{\lambda}{\sqrt{\pi}} + \frac{1}{\sqrt{\pi}} \Gamma\left(\frac{1}{2}\right) = \frac{\lambda}{\sqrt{\pi}} + 1 \\
 \frac{\lambda}{\sqrt{\pi}} + 1 &= \bar{X} \Rightarrow \hat{\lambda} = \left(\left(\frac{1}{n} \sum x_k \right) - 1 \right) \cdot \sqrt{\pi}
 \end{aligned}$$

Контроль:

$$\begin{aligned}
 M\left(\left(\frac{1}{n} \sum x_k\right) - 1\right) \cdot \sqrt{\pi} &= \sqrt{\pi} \left(M\left(\frac{1}{n} \sum x_k\right) - 1 \right) = \\
 &= \sqrt{\pi} \cdot \left(\left(\frac{\lambda}{\sqrt{\pi}} + 1 \right) - 1 \right) = \lambda
 \end{aligned}$$

Пример 7

Постройте оценки по методу моментов для параметров α и λ по выборке X_1, X_2, \dots, X_n

из гамма-распределения с плотностью $p(x, \alpha, \lambda) = \begin{cases} \frac{\lambda^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$.

Ответ. $\hat{\lambda} = \frac{\bar{X}}{\frac{1}{n} \sum_{k=1}^n X_k^2 - \bar{X}^2}$, $\hat{\alpha} = \frac{\bar{X}^2}{\frac{1}{n} \sum_{k=1}^n X_k^2 - \bar{X}^2}$

Пример 8

Смысл свойств (кратко):

Несмещённость: Отсутствие систематической ошибки в методике измерения. В среднем — попадём в цель.

Состоятельность: Уверенность, что увеличение объёма данных (больше смен, больше дней наблюдений) приведёт к увеличению точности оценки.

Эффективность: Выбор метода, который даёт самый точный результат при фиксированном объёме данных, минимизируя "разброс" оценок.

На заводе по производству подшипников автоматическая линия должна выпускать изделия с целевым диаметром μ мм. Из-за случайных колебаний температуры и износа оборудования диаметры готовых подшипников колеблются.

Для контроля качества начальник цеха рассматривает два способа оценки среднего диаметра партии:

1. Среднее арифметическое: $\hat{\mu}_1 = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, где X_i — диаметр i -го

случайно отобранного подшипника.

2. Усечённое среднее: $\hat{\mu}_2 = \frac{1}{n-2} (X_{(2)} + X_{(3)} + \dots + X_{(n-1)})$, где $X_{(1)}$ — самый

маленький подшипник в выборке, а $X_{(n)}$ — самый большой. То есть, мы отбрасываем самое маленькое и самое большое наблюдение и усредняем оставшиеся.

Несмещённость:

Предположим, производственный процесс отлажен и распределение диаметров является симметричным относительно μ . Будут ли оценки $\hat{\mu}_1$ и $\hat{\mu}_2$ несмещёнными в этом случае?

Даёт ли метод отбрасывания экстремальных значений систематическое смещение в оценке целевого диаметра?

Эффективность:

Предположим, в процессе иногда случаются значительные выбросы из-за временного сбоя (например, один подшипник из ста имеет сильное отклонение). Какая оценка, по вашему мнению, будет более эффективной (иметь меньшую дисперсию) в такой ситуации? Почему?

Какой метод контроля качества обеспечит более стабильные и предсказуемые результаты и лучше защищён от случайного брака в выборке?

Состоятельность:

Являются ли обе оценки состоятельными? Что будет с ними при стремлении объёма выборки (количества измеряемых подшипников) к бесконечности?

Практический смысл: Если мы начнём проводить тотальный контроль всех подшипников без исключения, обе оценки приведут нас к истинному значению μ ?

X_1, \dots, X_n — случайная выборка диаметров подшипников.

$E[X_i] = \mu$, $D[X_i] = \sigma^2$ (процесс отлажен, но есть случайные колебания).

$$\hat{\mu}_1 = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$\hat{\mu}_2 = \frac{1}{n-2} (X_{(2)} + X_{(3)} + \dots + X_{(n-1)}) \text{ (усечённое среднее, без min и max).}$$

$$M[\hat{\mu}_1] = M[\bar{X}] = M\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n M[X_i] = \frac{1}{n} (n\mu) = \mu.$$

является несмещённой оценкой для μ .

Данная проверка сложнее. Упрощающее предположение: распределение диаметров симметрично относительно μ . При симметричном распределении математические ожидания крайних порядковых статистик также симметричны:

$$E[X_{(1)}] = \mu - a, \quad E[X_{(n)}] = \mu + a \quad (\text{для некоторого } a > 0).$$

Тогда математическое ожидание суммы всех наблюдений можно выразить так:

$$\begin{aligned} M\left[\sum_{i=1}^n X_i\right] &= M[X_{(1)} + X_{(2)} + \dots + X_{(n)}] = M[X_{(1)}] + M[X_{(n)}] + M\left[\sum_{i=2}^{n-1} X_i\right] = \\ &= (\mu - a) + (\mu + a) + M\left[\sum_{i=2}^{n-1} X_i\right] = 2\mu + M\left[\sum_{i=2}^{n-1} X_{(i)}\right]. \end{aligned}$$

С другой стороны, $M\left[\sum_{i=1}^n X_i\right] = n\mu$.

Приравниваем: $n\mu = 2\mu + M\left[\sum_{i=2}^{n-1} X_i\right]$

$$M\left[\sum_{i=2}^{n-1} X_i\right] = n\mu - 2\mu = (n-2)\mu$$

Следовательно,

$$M[\hat{\mu}_2] = M\left[\frac{1}{n-2} \sum_{i=2}^{n-1} X_i\right] = \frac{1}{n-2} M\left[\sum_{i=2}^{n-1} X_i\right] = \frac{1}{n-2} \cdot (n-2)\mu = \mu.$$

Вывод: При симметричном распределении $\hat{\mu}_2$ также является несмещённой оценкой для μ .

2. Проверка на состоятельность

Согласно Закону Больших Чисел (ЗБЧ), выборочное среднее сходится по вероятности к математическому ожиданию: $\bar{X} \xrightarrow[n \rightarrow \infty]{P} \mu$.

Вывод: $\hat{\mu}_1$ является состоятельной оценкой.

Для $\hat{\mu}_2$:

При увеличении объёма выборки n до бесконечности, доля отбрасываемых наблюдений (2 наблюдения) стремится к нулю: $\frac{2}{n} \rightarrow 0$. Оценка $\hat{\mu}_2$ использует

почти все данные, и её поведение также подчиняется ЗБЧ. Формально, можно показать, что разность между $\hat{\mu}_1$ и $\hat{\mu}_2$ стремится к нулю.

$$\hat{\mu}_1 = \frac{1}{n}(X_{(1)} + X_{(n)} + \sum_{i=2}^{n-1} X_i)$$

$$\hat{\mu}_2 = \frac{1}{n-2} \sum_{i=2}^{n-1} X_i = \frac{n}{n-2} \cdot \frac{1}{n} \sum_{i=2}^{n-1} X_i$$

Разность: $\hat{\mu}_1 - \hat{\mu}_2 = \frac{1}{n}(X_{(1)} + X_{(n)}) + (\frac{1}{n} - \frac{1}{n-2}) \sum_{i=2}^{n-1} X_i$. Поскольку $X_{(1)}$ и $X_{(n)}$ имеют

ограниченные дисперсии, а множитель $\frac{1}{n} \rightarrow 0$, вся разность сходится к нулю.

Следовательно, $\hat{\mu}_2$ ведёт себя так же, как и $\hat{\mu}_1$.

Вывод: $\hat{\mu}_2$ является состоятельной оценкой.

3. Сравнение эффективности

Эффективность сравнивается для несмещённых оценок. Мы установили, что обе оценки несмещённые (при симметрии).

Дисперсия $\hat{\mu}_1$: $D[\hat{\mu}_1] = D[\bar{X}] = \frac{\sigma^2}{n}$.

Дисперсия $\hat{\mu}_2$: Точное вычисление дисперсии усечённого среднего сложно и зависит от распределения. Однако, можно дать качественную оценку.

Усечённое среднее специально создано для того, чтобы уменьшить дисперсию в присутствии выбросов. Предположим, в выборке есть один подшипник с сильным отклонением (брак). В оценке $\hat{\mu}_1$ этот бракованный подшипник окажет сильное влияние на среднее. В оценке $\hat{\mu}_2$ этот подшипник с большой вероятностью окажется $X_{(1)}$ или $X_{(n)}$ и будет отброшен.

Следовательно, $\hat{\mu}_2$ будет гораздо меньше подвержна этому выбросу.

В условиях, когда возможны значительные отклонения от нормы (тяжёлые "хвосты" у распределения), дисперсия усечённого среднего меньше, чем дисперсия обычного среднего. Значит, $\hat{\mu}_2$ более эффективна, чем $\hat{\mu}_1$, в такой ситуации.

Обе оценки являются несмещёнными и состоятельными. Однако усечённое среднее ($\hat{\mu}_2$) более робастно (устойчиво) к выбросам и в условиях реального производства, где возможен брак, является более эффективной оценкой. Его использование даёт более стабильные и надёжные результаты при контроле качества.

Пример 9

Финансовый аналитик хочет оценить среднюю доходность (μ) акций некоторой компании. Доходность — это случайная величина с неизвестной дисперсией (σ^2).

У него есть два предложения:

1. Обычное среднее: $\hat{\mu}_1 = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, где X_i — доходность за i -й день.
2. Взвешенное среднее: $\hat{\mu}_2 = \frac{1}{n+1} (X_1 + 2X_2 + 3X_3 + \dots + nX_n)$. Чем позже дата, тем больший вес придаётся наблюдению (подразумевается, что последние данные более релевантны).

Несмещённость:

Является ли оценка $\hat{\mu}_2$ несмещённой для μ ? Является ли несмещённой оценка $\hat{\mu}_1$?

Склонна ли стратегия, ставящая на последние данные, систематически завышать или занижать истинную ожидаемую доходность?

Является ли оценка $\hat{\mu}_2$ состоятельной? Устремится ли она к истинному значению μ , если аналитик будет собирать данные за всё большее количество дней ($n \rightarrow \infty$)?

Если аналитик будет использовать взвешенную оценку $\hat{\mu}_2$ долгие годы, получит ли он в итоге верное представление о средней доходности актива?

В какой реальной ситуации оценка $\hat{\mu}_2$ может оказаться полезнее, чем $\hat{\mu}_1$, несмотря на её формальные недостатки? (Подсказка: если средняя

доходность μ со временем не постоянна, а медленно дрейфует — это называется "нестационарный процесс").

Практический смысл: Свойства "несмещённость" и "состоятельность" критически зависят от предположения о постоянстве оцениваемого параметра. В реальном мире это предположение часто нарушается.

X_1, \dots, X_n — доходность актива за последовательные дни.

$M[X_i] = \mu$, $D[X_i] = \sigma^2$ (предполагается, что процесс стационарен).

Оценка 1: $\hat{\mu}_1 = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

Оценка 2: $\hat{\mu}_2 = \frac{1}{n+1} (1 \cdot X_1 + 2 \cdot X_2 + \dots + n \cdot X_n)$

1. Проверка на несмещённость

$M[\hat{\mu}_1] = M[\bar{X}] = \mu$. Вывод: $\hat{\mu}_1$ — несмещённая.

$$M[\hat{\mu}_2] = M\left[\frac{1}{n+1} \sum_{i=1}^n i \cdot X_i\right] = \frac{1}{n+1} \sum_{i=1}^n i \cdot M[X_i] = \frac{1}{n+1} \sum_{i=1}^n i \cdot \mu = \frac{\mu}{n+1} \cdot \sum_{i=1}^n i.$$

Сумма первых n натуральных чисел: $\sum_{i=1}^n i = \frac{n(n+1)}{2}$.

Подставляем:

$$E[\hat{\mu}_2] = \frac{\mu}{n+1} \cdot \frac{n(n+1)}{2} = \frac{\mu n}{2}.$$

Вывод: $E[\hat{\mu}_2] = \frac{n}{2} \mu \neq \mu$ (для $n > 2$). Оценка $\hat{\mu}_2$ является смещённой. Её

смещение: $(\hat{\mu}_2) = M[\hat{\mu}_2] - \mu = \frac{n}{2} \mu - \mu = \mu \left(\frac{n}{2} - 1\right)$.

Смещение растёт с ростом объёма выборки n , что является крайне нежелательным свойством.

Мы выяснили, что $E[\hat{\mu}_2] = \frac{n}{2} \mu$. При $n \rightarrow \infty$ математическое ожидание оценки стремится к бесконечности (если $\mu > 0$) или к минус бесконечности (если $\mu < 0$). Оценка не сходится к истинному параметру μ ни по вероятности, ни в каком-либо другом смысле.

Вывод: $\hat{\mu}_2$ является несостоятельной оценкой. Это делает её совершенно непригодной для использования при большом объёме данных.

Формальный анализ показывает, что $\hat{\mu}_2$ — плохая оценка для постоянного параметра μ . Однако в реальности финансовые рынки часто нестационарны: средняя доходность μ может медленно меняться во времени (дрейфовать).

В этом случае последние данные действительно несут больше информации о текущем значении доходности, чем старые. Оценка $\hat{\mu}_1$, усредняющая все данные с одинаковым весом, будет слишком "медленной" и будет отставать от изменений рынка.

Оценка $\hat{\mu}_2$, хотя и несостоятельная для постоянного μ , может лучше отслеживать изменение. Её формальный недостаток (большие веса на последние наблюдения) становится практическим преимуществом. Более того, можно модифицировать оценку, чтобы она была состоятельной для меняющегося параметра, например, использовать экспоненциальное сглаживание с весами, затухающими в прошлое ($w_i = \lambda^{n-i}$), где сумма весов равна 1. Оценка $\hat{\mu}_2$ является грубым и неудачным аппроксимацией этой идеи.

$\hat{\mu}_1$ является несмещённой, состоятельной и эффективной (по теореме Гаусса-Маркова при нормальности ошибок).

$\hat{\mu}_2$ является смещённой и несостоятельной, её использование недопустимо.

Для отслеживания меняющейся средней доходности (μ_t меняется со временем t):

$\hat{\mu}_1$ непригодна, так как слишком инертна.

Идея, заложенная в $\hat{\mu}_2$ (учёт релевантности свежих данных), верна, но сама оценка реализована плохо. Следует использовать более сложные модели (скользящие средние, модели с изменяющейся волатильностью, GARCH), которые формально корректно определяют веса для разных моментов времени.