



> Конспект > 0 урок > Введение: основы статистики

> Оглавление

> Оглавление

> Точечное оценивание

> Определения

> Свойства оценок

> Несмещённая оценка

> Состоятельная оценка

> Критерий состоятельности

> Свойства оценок

> Оценка максимального правдоподобия

> Пример ОМП. Нормальное распределение

> Основные свойства ОМП

> Экспоненциальное семейство распределений

> Центральная предельная теорема и Закон больших чисел

> Закон больших чисел

> Центральная предельная теорема

> Доверительный интервал

> Резюме

> Дополнения

> Plug-in estimators

> Эмпирическая функция распределения

> Оценка для часто используемых функций

> Использование ECDF в качестве оценки функции распределения

> Выборочные оценки

> Характеристические функции

> Дискретная с.в.

- > [Абсолютно непрерывная с.в.](#)
- > [Применение Plug-in для характеристических функций](#)
 - > [Характеристические функции для разных распределений](#)
 - > [Нормальное распределение](#)
 - > [Распределение Коши](#)
 - > [Регрессия](#)

> Точечное оценивание

> Определения

Определение

Будем называть **выборкой** набор случайных величин.

В теории вероятностей и статистике отдельно выделяют случай **независимых одинаково распределённых случайных величин** (н.о.р.с.в.). Каждая из них распределена так же, как и остальные, и все они независимы в совокупности.

Это принято обозначать как $X_1, \dots, X_n \sim F$.

Определение

Статистика — любая измеримая функция от выборки.

Определение

Точечная оценка параметра распределения — статистика, которую мы могли бы рассматривать как предполагаемое значение оцениваемого параметра.

> Свойства оценок

> Несмещённая оценка

$\hat{\theta}$ или $\hat{\theta}_n$ — **оценка параметра** θ .

$\hat{\theta}_n = g(X_1, \dots, X_n)$ — случайная величина, т.к. зависит от данных.

Определение

Оценка $\hat{\theta}_n$ **несмещённая**, если $E(\hat{\theta}_n) = \theta$

$bias(\hat{\theta}_n) = E(\hat{\theta}_n) - \theta$ — смещение оценки.

Пример

$X_1, \dots, X_n \sim F$, покажем, что $\bar{X}$ является несмещённой оценкой $EX = m$:

$$E\hat{m} = E\left(n^{-1} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n EX_i = \frac{1}{n} \sum_{i=1}^n m = \frac{nm}{n} = m$$

> Состоятельная оценка

Сходимость по вероятности: $\forall \epsilon > 0$ выполняется $P(|\hat{\theta}_n - \theta| > \epsilon) \xrightarrow{n \rightarrow \infty} 0$.

Определение

Оценка $\hat{\Theta}_n$ **состоятельная**, если $\hat{\theta}_n \xrightarrow{P} \theta$.

Пример

Дана выборка независимых одинаково распределённых случайных величин $X_1, \dots, X_n \sim F$.

Покажем, что оценка математического ожидания $\hat{\theta}_n = \frac{1}{n} \sum_i X_i$ является состоятельной:

$$V(\hat{\theta}) = V\left(\frac{1}{n} \sum_i X_i\right) = \frac{1}{n^2} V\left(\sum_i X_i\right) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Неравенство Чебышёва $P(|\hat{\theta} - \theta| > \epsilon) \leq \frac{V(\hat{\theta})}{\epsilon^2}$. Получаем:

$$P(|\hat{\theta} - \theta| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \rightarrow 0, n \rightarrow \infty$$

> Критерий состоятельности

Теорема

Если $bias \rightarrow 0$ и $se \rightarrow 0$ при $n \rightarrow \infty$, то оценка $\hat{\theta}$ состоятельная.

Пример

Дана выборка из распределения Бернулли $X_1, \dots, X_n \sim Bernoulli(p)$.

Оценка параметра распределения $\hat{p}_n = \frac{1}{n} \sum_i X_i$.

$E(\hat{p}_n) = p$, тогда при $n \rightarrow \infty$:

$$bias = p - \hat{p} = 0, se = \sqrt{\frac{p(1-p)}{n}} \rightarrow 0$$

Оценка $\hat{p}_n$ состоятельная.

> Свойства оценок

Определение

Оценка является **асимптотически нормальной**, если $\frac{\hat{\theta}_n - \theta}{se} \rightsquigarrow N(0, 1)$

Определение

τ_θ — семейство несмещённых оценок для Θ .

Оценка $\theta_{opt}^* \in \tau_\theta$ называется **оптимальной** оценкой для Θ , если

$$\forall \theta^* \in \tau_\Theta \quad V \theta_{opt}^* \leq V \theta^*$$

> Оценка максимального правдоподобия

Дана выборка из параметрического распределения $X_1, \dots, X_n \sim F_\theta, \theta \in \Theta$.

Функция правдоподобия $L(\theta) = \prod_{i=1}^n p(x_i; \theta)$.

Оценка максимального правдоподобия параметра $\theta : \hat{\theta} = \arg \max_\theta L(\theta)$.

> Пример ОМП. Нормальное распределение

Пусть $X_1, \dots, X_n \sim N(\mu, \sigma^2)$. Оценить параметры распределения μ, σ .

1. Функции правдоподобия (без домножения на константу):

$$L(\mu, \sigma) = \prod_i \frac{1}{\sigma} \exp\left\{-\frac{(X_i - \mu)^2}{2\sigma^2}\right\} = \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2} \sum_i (X_i - \mu)^2\right\} =$$

$$\sigma^{-n} \exp\left\{-\frac{nS^2}{2\sigma^2}\right\} \exp\left\{-\frac{n(\bar{X} - \mu)^2}{2\sigma^2}\right\}, \text{ где}$$

$$\bar{X} = \frac{1}{n} \sum_i X_i, S^2 = \frac{1}{n} \sum_i (X_i - \bar{X})^2$$

Последнее равенство верно, так как $\sum_i (X_i - \mu)^2 = nS^2 + n(\bar{X} - \mu)^2$,

что легко доказать из равенства $\sum_i (X_i - \mu)^2 = \sum_i (X_i - \bar{X} + \bar{X} - \mu)^2$

2. Логарифм функции правдоподобия:

$$\ln L(\mu, \sigma) = -n \ln \sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\bar{X} - \mu)^2}{2\sigma^2}$$

3. Возьмём производные по параметрам и приравняем их к нулю:

$$\frac{\delta \ln L(\mu, \sigma)}{\delta \mu} = \frac{n(\bar{X} - \mu)}{\sigma^2} = 0 \Rightarrow \hat{\mu} = \bar{X}$$

$$\frac{\delta \ln L(\mu, \sigma)}{\delta \sigma} = -\frac{n}{\sigma} + \frac{nS^2}{\sigma^3} + \frac{n(\bar{X} - \mu)^2}{\sigma^3} = 0 \Rightarrow \hat{\sigma} = S$$

> Основные свойства ОМП

- ОМП состоятельная: $\hat{\theta} \xrightarrow{P} \theta^*$, где θ^* — истинное значение параметра θ .
- **Инвариантность** ОМП: если $\hat{\theta}$ — ОМП параметра θ , то $g(\hat{\theta})$ — ОМП для $g(\theta)$.
- ОМП **асимптотически нормальная**: $(\hat{\theta} - \theta^*)/\hat{se} \rightsquigarrow N(0, 1)$.
- ОМП является **асимптотически оптимальной** или **эффективной**.

Среди всех хороших оценок ОМП имеет наименьшую дисперсию, по крайней мере, для больших выборок.

> Экспоненциальное семейство распределений

Многие известные и популярные распределения могут быть представлены в обобщённом виде:

$$f(x) = \frac{1}{h(\theta)} g(x) e^{\theta^T u(x)}$$

$$h(\theta) \in R^1, x \in R^m, \theta \in R^\alpha, \alpha \ll m, u(x) = (u_1(x), \dots, u_\alpha(x))^T.$$

Распределение	Плотность	$u(x)$	θ
Бернулли	$q^x(1-q)^{1-x}$	x	$\log \frac{q}{1-q}$
Мультиномиальное	$\prod_k \mu_k^{x_k}$	$[x_1, \dots, x_{K-1}]$	$\theta_i = \log \frac{\mu_i}{1-\sum_j \mu_j}$
Нормальное	$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$[x, x^2]$	$[-\frac{1}{2\sigma}, \frac{\mu}{\sigma^2}]$
Гамма	$\frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$	$[\log x, x]$	$[a-1, -b]$
Бета	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$	$[\log(x), \log(1-x)]$	$[a-1, b-1]$
Пуассон	$\exp(-\lambda) \frac{\lambda^x}{x!}$	$[x, \log \Gamma(x+1)]$	$[k, -1]$

Для распределений из экспоненциального семейства **ОМП существует и единственна.**

$$\begin{aligned}
\ln L(X^n, \theta) &= -n \ln h(\theta) + \sum_{i=1}^n \ln g(X_i) + \sum_{i=1}^n \theta^T u(X_i) \\
\frac{\partial \ln L(X^n, \theta)}{\partial \theta_i} &= \frac{\partial}{\partial \theta_i} \left(-\ln h(\theta) + \theta^T \left(\frac{1}{n} \sum_{i=1}^n u(X_i) \right) \right) \\
&= -\frac{1}{h(\theta)} \frac{\partial(h\theta)}{\partial \theta_i} + \frac{1}{n} \sum_{j=1}^n u_i(X_j) \\
\frac{\partial^2 \ln L(X^n, \theta)}{\partial \theta_i \partial \theta_j} &= \frac{1}{h^2(\theta)} \frac{\partial h(\theta)}{\partial \theta_i} \frac{\partial h(\theta)}{\partial \theta_j} - \frac{1}{h(\theta)} \frac{\partial^2 h(\theta)}{\partial \theta_i \partial \theta_j}
\end{aligned}$$

Распишем нормировочный множитель

$$h(\theta) = \int g(x) e^{\theta^T u(x)} dx$$

$$\frac{\partial h(\theta)}{\partial \theta_i} = \int u_i(x) g(x) e^{\theta^T u(x)} dx$$

$$\frac{1}{h(\theta)} \frac{\partial h(\theta)}{\partial \theta_i} = \int \frac{1}{h(\theta)} u_i(x) g(x) e^{\theta^T u(x)} dx = \mathbb{E} u_i(x)$$

Тогда

$$\frac{\partial^2 \ln L(X^n, \theta)}{\partial \theta_i \partial \theta_j} = \mathbb{E} u_i \mathbb{E} u_j - \mathbb{E}(u_i u_j) = -\text{cov}(u)$$

> Центральная предельная теорема и Закон больших чисел

> Закон больших чисел

Пусть $X_1, \dots, X_n$ н.о.р.с.в. с конечным вторым моментом $\mathbb{E} X_1^2 < \infty$.

Тогда:

$$\frac{X_1 + \dots + X_n}{n} \xrightarrow{P} \mathbb{E} X_1$$

> Центральная предельная теорема

Пусть $X_1, \dots, X_n$ н.о.р.с.в.

$$m = \mathbb{E} X_1, \sigma^2 = V X_1, \sigma \in (0, \infty)$$

Тогда $\forall y \in \mathbb{R}$:

$$P\left(\frac{X_1 + \dots + X_n - nm}{\sigma\sqrt{n}} < y\right) \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y e^{-\frac{x^2}{2}} dx = \Phi(y)$$

> Доверительный интервал

Определение

Доверительным интервалом с доверительной вероятностью $1 - \alpha$ для параметра θ называется интервал $C_n = (a, b)$, где $a = a(X_1, \dots, X_n)$ и $b = b(X_1, \dots, X_n)$ — такие функции выборки, что $P(\theta \in C_n) \geq 1 - \alpha$.

Пример

Дана выборка $X_1, \dots, X_n \sim N(\mu, \sigma^2)$. Построить доверительный интервал для μ .

Точечная оценка $\hat{\mu} = \bar{X}$.

Распределение точечной оценки $\hat{\mu} \sim N(\mu, \sigma^2/n)$.

Тогда:

$$\frac{\sqrt{n}(\hat{\mu} - \mu)}{\sigma} \sim N(0, 1)$$

Получаем доверительный интервал:

$$P\left\{-z_{\frac{\alpha}{2}} \leq \frac{\sqrt{n}(\hat{\mu} - \mu)}{\sigma} \leq z_{\frac{\alpha}{2}}\right\} = 1 - \alpha$$

$$C_n = \left(\hat{\mu} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \hat{\mu} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$$

$$L = p_n(X, \theta) = \prod_{i=1}^n p(x_i, \theta)$$

$$\frac{\partial \ln L}{\partial \theta} = \sum_{i=1}^n \frac{\partial \ln p(x_i, \theta)}{\partial \theta}$$

Покажем, что $\mathbb{E} \frac{\partial \ln p(x, \theta)}{\partial \theta} = 0$

$$1 = \int p(x, \theta) dx$$

$$\begin{aligned} 0 &= \int \frac{\partial p(x, \theta)}{\partial \theta} dx = \int \frac{\frac{\partial p(x, \theta)}{\partial \theta} p(x, \theta)}{p(x, \theta)} dx \\ &= \int \frac{\partial \ln p(x, \theta)}{\partial \theta} p(x, \theta) dx = \mathbb{E} \frac{\partial \ln p(x, \theta)}{\partial \theta} \end{aligned}$$

Дисперсия

$$\mathbb{V} \left(\frac{\partial \ln p(x, \theta)}{\partial \theta} \right) = \mathbb{E} \left(\frac{\partial \ln L}{\partial \theta} \right)^2 = -\mathbb{E} \frac{\partial^2 \ln p(x, \theta)}{\partial \theta^2}$$

> Резюме

- Рассмотрели основные определения, познакомились с точечным оцениванием, узнали, какие естественные требования нужно предъявлять к оценкам параметров.
- Обсудили метод максимального правдоподобия, научились его применять и поговорили о его сильных и слабых сторонах.
- Для экспоненциального семейства распределений доказали корректность применения метода максимального правдоподобия.
- Обсудили центральную предельную теорему и закон больших чисел, что позволяет нам доказывать сходимость наших оценок по вероятности.
- Научились строить доверительные интервалы для точечных оценок.

> Дополнения

> Plug-in estimators

Оцениваемый параметр можно представить как функцию от распределения.

Мы можем представить параметры как функцию от распределения:

Среднее значение: $\mu(F) = \int x dF(x)$

Дисперсия: $\sigma^2(F) = \int x^2 dF - \mu^2(F)$

Медиана: $m(F) = \inf\{x | F(x) \geq 1/2\}$

Plug-in оценивание превращает проблему получения оценки θ в проблему оценивания распределения F . Но как это сделать?

> Эмпирическая функция распределения

Определение

Эмпирическая функция распределения $\hat{F}_n$ выборки $X_1, \dots, X_n$ имеет вид:

$$\hat{F}_n = \frac{\sum_{i=1}^n I(X_i \leq x)}{n}, \quad \text{где} \quad I(X_i \leq x) = \begin{cases} 1, & X_i \leq x \\ 0, & X_i > x \end{cases}$$

Теорема

Пусть $\hat{F}_n$ - эмпирическая функция распределения, построенная по выборке $X_1, \dots, X_n \sim F$. Тогда:

$$E(\hat{F}_n(x)) = F(x)$$

$$V(\hat{F}_n(x)) = \frac{F(x)(1 - F(x))}{n}$$

$$\hat{F}_n(x) \xrightarrow{P} F(x)$$

> Оценка для часто используемых функций

> Использование ECDF в качестве оценки функции распределения

Мы можем использовать ECDF в качестве оценки функции распределения. При этом дифференциал превращается в сумму δ -функций

$$d\hat{F}_n = \frac{1}{n} \sum_{i=1}^n \delta(x - X_i),$$

которая при интегрировании трансформируется в обычную сумму.

> Выборочные оценки

Выборочное среднее: $\mu(\hat{F}_n) = \int x d\hat{F}_n(x) = \int x \frac{1}{n} \sum_{i=1}^n \delta(x - X_i) =$

$$\frac{1}{n} \sum_{i=1}^n X_i$$

Выборочная дисперсия: $\sigma^2(\hat{F}_n) = \int x^2 d\hat{F}_n(x) - \mu^2(\hat{F}_n) = \frac{1}{n} \sum_{i=1}^n X_i^2 -$

$$\left(\frac{1}{n} \sum_{i=1}^n X_i\right)^2$$

> Характеристические функции

Определение

Пусть задана случайная величина X с распределением P , тогда характеристическая функция задаётся формулой:

$$\phi_X(t) = E[\exp(itX)] = \int_{-\infty}^{+\infty} \exp(itX) P(x) dx$$

> Дискретная с.в.

Для дискретной с.в. со значениями x_k и вероятностями p_k х.ф. принимает вид:

$$\phi_X(t) = \sum_{k=1}^N p_k \exp(itx_k)$$

Пример: распределение Бернулли $\phi_X(t) = 1 + p \cdot (\exp(it) - 1)$

> Абсолютно непрерывная с.в.

Пусть с.в. X имеет плотность распределения f_X :

$$\phi_X(t) = \int_{-\infty}^{+\infty} \exp(itx) f_X(x) dx$$

Пример: $X \sim U[0, 1]$

$$\phi_X(t) = \frac{\exp(it) - 1}{it}$$

> Применение Plug-in для характеристических функций

> Характеристические функции для разных распределений

Различные параметрические семейства имеют разный вид характеристических функций. Мы можем построить характеристическую функцию на основе ECDF оценки и сравнить с настоящим видом распределения:

$$\phi_{ECDF} = \frac{1}{N} \sum_{k=1}^N \exp(itX_k)$$

> Нормальное распределение

$$\phi_X(t) = \exp(i \cdot \mu t - \sigma^2 t^2 / 2)$$

> Распределение Коши

$$\phi_X(t) = \exp(i \cdot x_0 t - \gamma |t|)$$

> Регрессия

Регрессионный анализ позволяет понять, какому распределению в большей степени соответствует наша выборка.
