

Мы приступаем к изучению более точного метода оценки неизвестного значения параметра. Он превосходит метод моментов и при наличии достаточных статистик дает оптимальные оценки с точки зрения квадратичного риска. Более того, при выполнении определенных условий регулярности этот метод приводит к асимптотически ($n \rightarrow \infty$) оптимальным оценкам для широкого класса вероятностных моделей и практически при любых функциях потерь.

Идея метода состоит в математической формализации “разумного” поведения человека в условиях неопределенности. Представим себе ситуацию, что мы ожидаем появления одного из нескольких событий, вероятности которых нам неизвестны и нас интересуют не столько значения этих вероятностей, сколько то событие, которое происходит наиболее часто. Ситуация осложняется тем, что мы располагаем всего одним испытанием, в результате которого произошло некоторое событие A . Конечно, мы примем решение, что A обладает наибольшей вероятностью, и вряд ли можно предложить нечто более разумное, чем такое правило принятия решения.

В этом и состоит *принцип максимального правдоподобия*, который буквально пронизывает всю теорию оптимального статистического вывода. Применение этого принципа к проблеме оценки параметров приводит к следующему статистическому правилу: *если $x^{(n)}$ – результат наблюдения случайной выборки $X^{(n)}$, то за оценку параметра следует брать то его значение, при котором результат $x^{(n)}$ обладает наибольшим правдоподобием.*

Вы спросите, что такое “правдоподобие” результата $x^{(n)}$? Давайте формализуем это понятие.

Если наблюдается дискретная случайная величина, то естественно назвать правдоподобием результата $x^{(n)}$ при фиксированном значении параметра θ вероятность его наблюдения в статистическом эксперименте. Но в дискретном случае эта вероятность совпадает со значением функции плотности в точке $x^{(n)}$: $P_\theta(X^{(n)} = x^{(n)}) = f_n(x^{(n)} | \theta)$. Следовательно, оценка по методу максимального правдоподобия определяется точкой достижения максимума у функции плотности слу-

чайной выборки, то есть

$$\hat{\theta}_n(X^{(n)}) = \arg \max_{\theta \in \Theta} f_n(X^{(n)} | \theta). \quad (1)$$

Рассмотрим сразу же простой пример. Пусть $X^{(n)}$ – выборка в схеме Бернулли, и мы оцениваем вероятность θ успешного исхода. В этой модели

$$f(X^{(n)} | \theta) = \theta^{\sum_1^n X_k} (1 - \theta)^{n - \sum_1^n X_k}.$$

Дифференцируя эту функцию по θ и приравнявая производную нулю, находим оценку максимального правдоподобия $\theta = (1/n) \sum_1^n X_k$. Это – давно знакомая нам оценка вероятности успеха в испытаниях Бернулли, которую мы получили с помощью моментов и постоянно использовали при иллюстрации закона больших чисел.

Теперь определим правдоподобие в случае выбора из непрерывного распределения с функцией плотности (по мере Лебега) $f_n(x^{(n)} | \theta)$, $x^{(n)} \in \mathbb{R}^n$, $\theta \in \Theta$. Пусть $x^{(n)}$ – совокупность выборочных данных, то есть точка в n -мерном выборочном пространстве \mathbb{R}^n . Окружим эту точку прямоугольным параллелепипедом малого размера, скажем, $V_\varepsilon = \prod_1^n [x_k - \varepsilon/2; x_k + \varepsilon/2]$. В силу теоремы о среднем для кратного интеграла вероятность того, что выборочный вектор попадет в этот параллелепипед $P(X^{(n)} \in V_\varepsilon) \sim f_n(x^{(n)} | \theta) \cdot \varepsilon^n$, когда $\varepsilon \rightarrow 0$. Если трактовать эту вероятность, как правдоподобие результата $x^{(n)}$, которое, конечно, зависит от выбора малого ε , мы видим, что проблема максимизации правдоподобия сводится к проблеме отыскания точки достижения максимума по всем $\theta \in \Theta$ у функции плотности f_n . Таким образом, и в случае непрерывного распределения разумно назвать правдоподобием результата $x^{(n)}$ при фиксированном значении параметра θ опять-таки величину функции плотности выборки, то есть $f_n(x^{(n)} | \theta)$, и определить оценку максимального правдоподобия той же формулой (1).

Рассмотрим пример на построение такой оценки в случае выбора из непрерывного распределения. Пусть наблюдается случайная величина $X \sim \mathcal{N}(\mu, \sigma^2)$, так что функция плотности выборки

$$f_n(x^{(n)} | \theta) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_1^n (x_k - \mu)^2 \right\},$$

где $\theta = (\mu, \sigma)$ – двумерный параметр, значение которого нам неизвестно. В соответствии с формулой (1) необходимо отыскать точку достижения максимума функции $f_n(X^{(n)} | \mu, \theta)$ по переменным $\mu \in \mathbb{R}$ и $\sigma \in \mathbb{R}_+$. Естественно, логарифм этой функции имеет те же точки экстремума, что и сама функция, но логарифмирование упрощает выкладки, поэтому ищем максимум функции

$$\mathcal{L}(\theta | X^{(n)}) = \ln f_n(X^{(n)} | \theta) = -\frac{n}{2} \ln 2\pi - n \ln \sigma - \frac{1}{2\sigma^2} \sum_1^n (X_k - \mu)^2.$$

Составляем уравнения, определяющие точки экстремума:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mu} &= \frac{1}{2\sigma^2} \sum_1^n (X_k - \mu) = 0, \\ \frac{\partial \mathcal{L}}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_1^n (X_k - \mu)^2 = 0. \end{aligned}$$

Из первого уравнения сразу находим оценку параметра μ : $\hat{\mu}_n = \bar{X}$. Подставляя \bar{X} вместо μ во второе уравнение, находим оценку σ : $\hat{\sigma}_n = S$ (выборочное стандартное отклонение). Очевидно, (\bar{X}, S) – точка максимума. Таким образом, метод максимального правдоподобия приводит к тем же оценкам \bar{X} и S^2 параметров μ и σ^2 , что и метод моментов.

Теперь дадим строгое определение правдоподобия и рассмотрим еще несколько примеров, в которых метод максимального правдоподобия дает оценки, отличные от метода моментов.

Определение 4.1. Случайная функция

$$L(\theta | X^{(n)}) = \prod_{i=1}^n f(X_i | \theta)$$

на параметрическом пространстве Θ называется *функцией правдоподобия*, а значение ее реализации $L(\theta_0 | x^{(n)})$ при результате наблюдения $X^{(n)} = x^{(n)}$ и фиксированном $\theta = \theta_0$ – *правдоподобием значения θ_0 при результате $x^{(n)}$* . Любая точка $\hat{\theta}_n = \hat{\theta}_n(X^{(n)})$ (статистика) параметрического пространства Θ , доставляющая абсолютный максимум функции правдоподобия, называется *оценкой максимального правдоподобия* параметра θ .

Поскольку функция правдоподобия представляет собой произведение функций от θ , то при отыскании ее максимума методами дифференциального исчисления удобнее иметь дело с логарифмом этой

функции. Естественно, точки экстремума у *функции логарифмического правдоподобия*

$$\mathcal{L}(\theta | X^{(n)}) = \sum_{i=1}^n \ln f(X_i | \theta)$$

те же, что и у функции L , но если функция $L(\cdot | x^{(n)})$ имеет непрерывные частные производные по компонентам $\theta_1, \dots, \theta_k$ параметрического вектора θ , то проще дифференцировать \mathcal{L} чем L . В этом случае система уравнений

$$\frac{\partial \mathcal{L}(\theta | X^{(n)})}{\partial \theta_i} = 0, \quad i = 1, \dots, k \quad (2)$$

называется *уравнениями правдоподобия*. Это еще одна разновидность так называемых *оценочных уравнений*, – в предыдущем параграфе мы имели дело с уравнениями метода моментов.

Любое решение системы уравнений (2), доставляющее максимум функции $\mathcal{L}(\cdot | X^{(n)})$, может рассматриваться как оценка θ по методу максимального правдоподобия. Мы не будем изучать случаи, когда система (2) имеет несколько решений с возможно одинаковыми значениями функции правдоподобия в этих точках, так что требуются дополнительные априорные знания относительно вероятностной модели, позволяющие выбрать одно из этих решений. Во всех рассмотренных ниже примерах оценка максимального правдоподобия единственна.

Пример 4.1. *Оценка параметра положения равномерного распределения $U(0, \theta)$.* Равномерное на отрезке $[0; \theta]$ распределение имеет функцию плотности $f(x | \theta) = \theta^{-1}$, если $0 \leq x \leq \theta$, и $f(x | \theta) = 0$ вне этого отрезка. Следовательно, функция $L(\theta | X^{(n)})$ отлична от нуля и равна θ^{-n} только в области $\theta \geq X_{(n)} = \max_{1 \leq k \leq n} X_k$. Ее максимум по θ достигается в граничной точке $\theta = X_{(n)}$, так что наибольшее значение $X_{(n)}$ выборки $X^{(n)}$ есть оценка максимального правдоподобия параметра θ .

Легко видеть, что оценка θ по методу моментов равна $2\bar{X}$. Эта оценка на порядок хуже оценки максимального правдоподобия с точки зрения квадратичного риска $R(\theta; \hat{\theta}_n) = \mathbf{E}_\theta \left(\hat{\theta}_n(X^{(n)}) - \theta \right)^2$. Простые вычисления соответствующих математических ожиданий показывают, что $R(\theta; 2\bar{X}) = O(n^{-1})$, в то время как $R(\theta; X_{(n)}) = O(n^{-2})$.

Данный пример интересен тем, что здесь функция правдоподобия не имеет гладкого максимума, и именно это обстоятельство, как будет видно в дальнейшем, обеспечивает такое различное поведение риска рассматриваемых оценок.

Пример 4.2. Оценка параметров гамма-распределения $G(a, \lambda)$. У этого распределения функция плотности

$$f(x | \theta) = \frac{1}{a^\lambda \Gamma(\lambda)} x^{\lambda-1} \exp\left\{-\frac{x}{a}\right\}, \quad x > 0, \quad \theta = (a, \lambda),$$

отлична от нуля только на положительной полуоси, и логарифмическое правдоподобие

$$\mathcal{L}(a, \lambda | X^{(n)}) = -n\lambda \ln a - n \ln \Gamma(\lambda) + (\lambda - 1) \sum_1^n \ln X_k - \frac{1}{a} \sum_1^n X_k.$$

Составляем уравнения правдоподобия:

$$\frac{\partial \mathcal{L}}{\partial a} = -\frac{n\lambda}{a} + \frac{1}{a^2} \sum_1^n X_k = 0,$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = -n \ln a - n\psi(\lambda) + \sum_1^n \ln X_k = 0,$$

где $\psi(\lambda) = d \ln \Gamma(\lambda) / d\lambda$ – так называемая пси-функция Эйлера. Исключая из первого уравнения параметр a и подставляя полученный результат во второе, получаем трансцендентное уравнение

$$\ln \lambda - \psi(\lambda) = \ln \bar{X} - \frac{1}{n} \sum_1^n \ln X_k,$$

которое в силу свойства монотонности функции $\ln \lambda - \psi(\lambda)$ имеет единственное решение. При численном решении этого уравнения может оказаться полезной асимптотическая ($\lambda \rightarrow \infty$) формула

$$\ln \lambda - \psi(\lambda) = \frac{1}{2\lambda} + \frac{1}{12\lambda^2} + O\left(\frac{1}{\lambda^4}\right).$$

Пример 4.3. Оценка параметров структурированного среднего при нормальном распределении отклика. Данная задача весьма часто возникает при калибровке шкалы прибора. Две переменные x и y связаны линейным соотношением $y = a + bx$, и для градуировки

значений y на шкале прибора необходимо знать значения параметров a и b этой зависимости. Однако, для каждого стандартного фиксированного значения x прибор замеряет значение y с ошибкой, так что замеры происходят в рамках вероятностной модели $Y = a + bx + \xi$, где ошибка измерения (случайная величина) ξ имеет нормальное распределение с нулевым средним и некоторой дисперсией σ^2 , значение которой, как правило, также не известно. Случайная величина Y обычно называется *откликом* на значение *регрессора* x ; ее распределение при фиксированном x очевидно нормально $(a + bx, \sigma^2)$.

Для оценки a и b производится n измерений отклика y_1, \dots, y_n при некоторых фиксированных значениях x_1, \dots, x_n регрессора x , оптимальный выбор которых, обеспечивающий наибольшую точность и надежность калибровки, составляет самостоятельную задачу особой области математической статистики – *планирование регрессионных экспериментов*. Мы будем предполагать, что значения x_1, \dots, x_n априори фиксированы. В таком случае значения y_1, \dots, y_n представляют реализации n независимых случайных величин Y_1, \dots, Y_n , и $Y_k \sim \mathcal{N}(a + bx_k, \sigma^2)$, $k = 1, \dots, n$. Совместная функция плотности Y_1, \dots, Y_n равна

$$f_n(y^{(n)} | a, b, \sigma) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_1^n (y_k - a - bx_k)^2 \right\},$$

так что логарифмическая функция правдоподобия, необходимая для оценки параметров a , b и σ методом максимального правдоподобия имеет вид

$$\mathcal{L}(a, b, \sigma | Y^{(n)}) = -\frac{n}{2} \ln 2\pi - n \ln \sigma - \frac{1}{2\sigma^2} \sum_1^n (Y_k - a - bx_k)^2.$$

Вычисляя производные этой функции по переменным a , b и σ , получаем уравнения правдоподобия

$$\begin{aligned} \sum_1^n (Y_k - a - bx_k) &= 0, \\ \sum_1^n x_k (Y_k - a - bx_k) &= 0, \\ n\sigma^2 - \sum_1^n (Y_k - a - bx_k)^2 &= 0. \end{aligned}$$

Конечно, это очень простая система уравнений, решение которой не может вызывать каких-либо затруднений, и мы сразу пишем оценки максимального правдоподобия

$$\hat{a}_n = \bar{Y} - \frac{m_{xY}}{s_x^2} \bar{x}, \quad \hat{b}_n = \frac{m_{xY}}{s_x^2}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_1^n \left(Y_k - \hat{a}_n - \hat{b}_n x_k \right)^2,$$

где

$$\bar{x} = \frac{1}{n} \sum_1^n x_k, \quad \bar{Y} = \frac{1}{n} \sum_1^n Y_k, \quad s_x = \frac{1}{n} \sum_1^n (x_k - \bar{x})^2, \quad S_Y = \frac{1}{n} \sum_1^n (Y_k - \bar{Y})^2,$$

$$m_{xY} = \frac{1}{n} \sum_1^n (x_k - \bar{x})(Y_k - \bar{Y}).$$

Легко видеть, что оценки по методу максимального правдоподобия параметров a и b совпадают с их оценками по *методу наименьших квадратов*. В этом методе “выравнивания” экспериментальных данных оценки ищутся из условия минимизации суммы квадратов *невязок*: $\sum_1^n (Y_k - a - bx_k)^2$, причем под невязкой понимается разность между откликом Y и его “теоретическим” средним значением $a + bx$.

Пример 4.4. *Оценка параметров двумерного нормального распределения: задачи регрессии и прогноза.* Оценка по методу максимального правдоподобия пяти параметров $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho$ двумерного нормального распределения с функцией плотности

$$f(x, y | \theta) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left(\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right) \right\}$$

не представляет особой технической сложности. Эти оценки совпадают с оценками по методу моментов и, таким образом, равны выборочным аналогам тех характеристик двумерного нормального распределения, которые соответствуют указанным пяти параметрам:

$$\hat{\mu}_{1,n} = \bar{X}, \quad \hat{\mu}_{2,n} = \bar{Y}, \quad \hat{\sigma}_{1,n}^2 = S_X^2, \quad \hat{\sigma}_{2,n}^2 = S_Y^2, \quad \hat{\rho}_n = r.$$

Формулы для вычисления выборочных средних \bar{X} и \bar{Y} , выборочных дисперсий S_1^2 и S_2^2 , а также выборочного коэффициента корреляции r приведены в конце §2.

Полученные оценки часто используются для оценки параметров линейного прогноза $Y = a + bX$ значений случайной величины Y по результатам наблюдений X . В случае нормального распределения линейный прогноз обладает свойством оптимальности с точки зрения малости средней квадратичной ошибки и совпадает с кривой средней квадратичной регрессии (см. предложение 10.3 курса ТВ)

$$y = \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1).$$

Однако формальная подгонка прогностической кривой с помощью прямой линии используется и вне рамок нормальной модели, и в этом случае оценки

$$\hat{a}_n = \bar{Y} - r \frac{S_2}{S_1} \bar{X}, \quad \hat{b}_n = r \frac{S_2}{S_1}$$

совпадают с оценками по *методу наименьших квадратов*: минимизируется, как и в примере 4.3, сумма квадратов невязок

$$\sum_1^n (Y_k - a - bX_k)^2.$$

Хотя оценки в обоих примерах имеют одинаковый вид, но решаемые в них статистические проблемы весьма различны: в примере 4.3 оценивались параметры некоторой функциональной зависимости с ошибками в наблюдениях отклика, в то время как в примере 4.4 решается задача выявления корреляционной связи и использования этой связи для прогноза.

Лекция 7

Исследуем теперь асимптотические свойства оценок по методу максимального правдоподобия.

Начнем с выяснения достаточных условий состоятельности этих оценок. Такие ограничения на вероятностную модель обычно называются *условиями регулярности*, и в данном случае они имеют следующий вид.

(R1) Параметрическое пространство Θ есть открытый интервал на прямой \mathbb{R} .

- (R2) Носитель \mathcal{X} распределения P_θ наблюдаемой случайной величины X не зависит от θ , то есть все множества $\mathcal{X} = \{x : f(x|\theta) > 0\}$ можно считать одинаковыми, каково бы ни было $\theta \in \Theta$.
- (R3) Распределения P_θ различны при разных θ , то есть при любых $\theta_1 \neq \theta_2$, $\theta_1, \theta_2 \in \Theta$, имеет место тождество $\mu\{x : x \in \mathcal{X}, f(x|\theta_1) = f(x|\theta_2)\} = 0$, где μ – мера, по которой вычисляется плотность $f(x|\theta)$ распределения P_θ .

Доказательство состоятельности оценок максимального правдоподобия, как и оценок по методу моментов, опирается на закон больших чисел, но при этом используется следующее достаточно простое, но играющее большую роль в теории вероятностей, неравенство.

Лемма 4.1. (неравенство Йенсена) Пусть X – случайная величина с конечным математическим ожиданием. Если функция $g(\cdot)$ дважды дифференцируема и выпукла ($g'' > 0$) на некотором интервале, содержащем носитель распределения X , и математическое ожидание $\mathbf{E} g(X)$ существует, то справедливо неравенство $\mathbf{E} g(X) \geq g(\mathbf{E} X)$, причем знак равенства достигается тогда и только тогда, когда распределение X сосредоточено в одной точке ($X = \text{const.}$).

Доказательство. Так как функция g дважды дифференцируема, то справедливо следующее представление Тейлора в окрестности точки $\mu = \mathbf{E} X$:

$$g(X) = g(\mu) + (X - \mu)g'(\mu) + (X - \mu)^2 g''(\mu + \lambda(X - \mu))/2, \quad 0 < \lambda < 1.$$

Вычисляя математическое ожидание от обеих частей этого равенства, получаем

$$\mathbf{E} g(X) = g(\mathbf{E} X) + \mathbf{E}(X - \mu)^2 g''(\mu + \lambda(X - \mu))/2 \geq g(\mathbf{E} X).$$

Знак равенства возможен только в случае $\mathbf{E}(X - \mu)^2 g''(\mu + \lambda(X - \mu)) = 0$. Но поскольку $g'' > 0$, то последнее равенство с необходимостью влечет $(X - \mu)^2 = 0$, то есть $X = \text{const.}$

Покажем теперь, что справедлива

Теорема 4.1 (состоятельность). Если функция логарифмического правдоподобия

$$\mathcal{L}(\theta | X^{(n)}) = \sum_{k=1}^n \ln f(X_k | \theta) \tag{3}$$

имеет единственный максимум, то при выполнении условий регулярности (R1)–(R3) точка $\hat{\theta}_n$ достижения максимума этой функции (оценка максимального правдоподобия) является состоятельной оценкой параметра θ .

Доказательство. Покажем, что для любого фиксированного $\theta_0 \in \Theta$ и любого $\varepsilon > 0$ вероятность $P_{\theta_0}(|\hat{\theta}_n - \theta_0| < \varepsilon) \rightarrow 1$.

Если θ_0 – истинное значение параметра θ , то в силу условия (R1) θ_0 – внутренняя точка Θ . Тогда сформулированная выше задача состоит в доказательстве следующего утверждения: в некоторой ε -окрестности $(\theta_0 - \varepsilon; \theta_0 + \varepsilon)$ функция $\mathcal{L}(\cdot | X^{(n)})$ обладает локальным максимумом с вероятностью, стремящейся к единице при $n \rightarrow \infty$.

Но если происходит событие

$$A_n = \{\mathcal{L}(\theta_0 | X^{(n)}) > \mathcal{L}(\theta_0 \pm \varepsilon | X^{(n)})\},$$

то внутри этой окрестности имеется точка максимума, и нам остается только показать, что $P_{\theta_0}(A_n) \rightarrow 1$, ибо $P_{\theta_0}(|\hat{\theta}_n - \theta_0| < \varepsilon) \geq P_{\theta_0}(A_n)$.

Используя условие (R2) и вид функции \mathcal{L} (см. (3)), представим неравенство, определяющее событие A_n , в виде

$$\frac{1}{n} \sum_{k=1}^n \ln \frac{f(X_k | \theta_0 \pm \varepsilon)}{f(X_k | \theta_0)} < 0.$$

В силу закона больших чисел Хинчина левая часть этого неравенства сходится по вероятности к

$$\mathbf{E}_{\theta_0} \ln \frac{f(X_k | \theta_0 \pm \varepsilon)}{f(X_k | \theta_0)}, \quad (4)$$

и для доказательства утверждения достаточно показать, что это математическое ожидание строго меньше нуля (кстати, докажите сами, что при справедливости условий теоремы математическое ожидание (4) всегда существует, в противном случае закон больших чисел Хинчина не применим).

Так как $g(x) = -\ln x$ – выпуклая функция, то в силу неравенства Йенсена

$$\begin{aligned} \mathbf{E}_{\theta_0} \ln \frac{f(X | \theta_0 \pm \varepsilon)}{f(X | \theta_0)} &\leq \ln \mathbf{E}_{\theta_0} \frac{f(X | \theta_0 \pm \varepsilon)}{f(X | \theta_0)} = \\ &\ln \int_{\mathcal{X}} \frac{f(x | \theta_0 \pm \varepsilon)}{f(x | \theta_0)} \cdot f(x | \theta_0) d\mu(x) = \ln 1 = 0, \end{aligned}$$

причем равенство нулю первого члена в этой цепочке неравенств возможно лишь в случае

$$\frac{f(X | \theta_0 \pm \varepsilon)}{f(X | \theta_0)} = \text{const.},$$

то есть, поскольку интеграл от плотности равен 1, лишь в случае $f(X | \theta_0 \pm \varepsilon) = f(X | \theta_0)$, что невозможно в силу условия (R3). Таким образом, математическое ожидание (4) строго меньше нуля, и состоятельность оценки максимального правдоподобия доказана.

Изучим теперь асимптотическое распределение оценки максимального правдоподобия. Для этого нам потребуется ввести дополнительные условия регулярности.

(R4) Для каждой точки θ_0 параметрического пространства Θ существует некоторая ее окрестность, в которой функция плотности $f(x | \theta)$ трижды дифференцируема по параметру θ и

$$\left| \frac{\partial f(x | \theta)}{\partial \theta} \right| \leq H_1(x), \quad (5)$$

$$\left| \frac{\partial^2 f(x | \theta)}{\partial \theta^2} \right| \leq H_2(x), \quad (6)$$

$$\left| \frac{\partial^3 \ln f(x | \theta)}{\partial \theta^3} \right| \leq H_3(x),$$

причем функции H_1, H_2 и H_3 интегрируемы по мере μ на носителе \mathcal{X} распределения X .

(R5) Функция

$$I(\theta) = \mathbf{E}_\theta \left(\frac{\partial \ln f(X | \theta)}{\partial \theta} \right)^2 = \int_{\mathcal{X}} \left(\frac{\partial \ln f(x | \theta)}{\partial \theta} \right)^2 f(x | \theta) d\mu(x) > 0,$$

каково бы ни было $\theta \in \Theta$.

Естественно, столь громоздкие и, на первый взгляд, странные условия требуют некоторого комментария.

Условие (R4) означает, что соответствующие производные функции плотности равномерно интегрируемы на \mathcal{X} , и поэтому можно выносить производную по θ за знак интеграла.

Условие (R5) требует положительности очень важной, с точки зрения состоятельности статистического вывода, характеристики вероятностной модели: $I(\theta)$ называется *информацией по Фишеру* в точке θ , содержащейся в наблюдении случайной величины X . Если $I(\theta) = 0$, то возникают непреодолимые трудности с принятием корректного решения, соответствующего этой параметрической точке θ . Понятно, что аналогичным образом можно определить и информацию по Фишеру, содержащуюся в случайной выборке $X^{(n)}$:

$$I_n(\theta) = \mathbf{E}_\theta \left(\frac{\partial \ln f_n(X^{(n)} | \theta)}{\partial \theta} \right)^2.$$

Приведем несколько утверждений, касающихся свойств информации по Фишеру.

Лемма 4.2. 1^0 . При выполнении условия (R4) в части (6) для вычисления информации по Фишеру можно использовать формулу

$$I(\theta) = - \mathbf{E}_\theta \frac{\partial^2 \ln f(X | \theta)}{\partial \theta^2}.$$

2^0 . При выполнении условия (R4) в части (5) информация по Фишеру обладает свойством аддитивности $I_n(\theta) = nI(\theta)$ – информация, содержащаяся в выборке, равна сумме информации, содержащихся в наблюдении каждой ее компоненты.

Доказательство. 1^0 . Условие (R4) в части (6) обеспечивает возможность смены порядка дифференцирования и интегрирования функции плотности, поэтому

$$\begin{aligned} \mathbf{E}_\theta \frac{\partial^2 \ln f(X | \theta)}{\partial \theta^2} &= \mathbf{E}_\theta \left(\frac{f''_{\theta\theta}(X | \theta)}{f(X | \theta)} - \left(\frac{f'_\theta(X | \theta)}{f(X | \theta)} \right)^2 \right) = \\ \int_{\mathfrak{X}} \frac{f''_{\theta\theta}(x | \theta)}{f(x | \theta)} \cdot f(x | \theta) d\mu(x) - I(\theta) &= \frac{d^2}{d\theta^2} \int_{\mathfrak{X}} f(x | \theta) d\mu(x) - I(\theta) = -I(\theta). \end{aligned}$$

2^0 . Используя независимость и одинаковую распределенность компонент случайной выборки, получаем, что

$$I_n(\theta) = \mathbf{E}_\theta \left(\frac{\partial \sum_1^n \ln f(X_k | \theta)}{\partial \theta} \right)^2 =$$

$$\begin{aligned} & \mathbf{E}_\theta \left(\sum_{k=1}^n \left(\frac{\partial \ln f(X_k | \theta)}{\partial \theta} \right)^2 - \sum_{i \neq j} \frac{\partial \ln f(X_i | \theta)}{\partial \theta} \cdot \frac{\partial \ln f(X_j | \theta)}{\partial \theta} \right) = \\ & \sum_{k=1}^n \mathbf{E}_\theta \left(\frac{\partial \ln f(X_k | \theta)}{\partial \theta} \right)^2 - \sum_{i \neq j} \mathbf{E}_\theta \frac{\partial \ln f(X_i | \theta)}{\partial \theta} \cdot \mathbf{E}_\theta \frac{\partial \ln f(X_j | \theta)}{\partial \theta} = nI(\theta), \end{aligned}$$

поскольку, в силу неравенства (5) в условии (R4), математическое ожидание

$$\mathbf{E}_\theta \frac{\partial \ln f(X | \theta)}{\partial \theta} = \int_{\mathcal{X}} \frac{f'_\theta(x | \theta)}{f(x | \theta)} \cdot f(x | \theta) d\mu(x) = \frac{d}{d\theta} \int_{\mathcal{X}} f(x | \theta) d\mu(x) = 0.$$

Теперь приступим к выводу асимптотического распределения оценки максимального правдоподобия скалярного параметра θ .

Теорема 4.2 (асимптотическая нормальность). *При выполнении условий (R1)–(R5) любой корень $\hat{\theta}_n = \hat{\theta}_n(X^{(n)})$ уравнения правдоподобия $\partial \mathcal{L}(\theta | X^{(n)})/\partial \theta = 0$ асимптотически ($n \rightarrow \infty$) нормален со средним θ и дисперсией $(nI(\theta))^{-1}$, то есть*

$$\lim_{n \rightarrow \infty} P_\theta \left((\hat{\theta}_n - \theta) \sqrt{nI(\theta)} < x \right) = \Phi(x).$$

Доказательство. Если $\hat{\theta}_n$ – оценка по методу максимального правдоподобия (корень уравнения правдоподобия), то имеет место тождество $\partial \mathcal{L}(\hat{\theta}_n | X^{(n)})/\partial \theta = 0$. Используя условие (R4), разложим его левую часть по формуле Тейлора в окрестности истинного значения θ_0 параметра θ :

$$\begin{aligned} \partial \mathcal{L}(\hat{\theta}_n | X^{(n)})/\partial \theta &= \mathcal{L}'(\theta_0 | X^{(n)}) + \\ & (\hat{\theta}_n - \theta_0) \mathcal{L}''(\theta_0 | X^{(n)}) + (\hat{\theta}_n - \theta_0)^2 \mathcal{L}'''(\theta_1 | X^{(n)})/2 = 0, \end{aligned}$$

где производные от функции правдоподобия \mathcal{L} вычисляются по параметру θ , а $\theta_1 = \theta_0 + \lambda(\hat{\theta}_n - \theta_0)$, $0 < \lambda < 1$.

Разрешим полученное уравнение относительно величины $\sqrt{n}(\hat{\theta}_n - \theta_0)$, которая, согласно утверждению теоремы, должна иметь в пределе при $n \rightarrow \infty$ нормальное распределение со средним 0 и дисперсией $[I(\theta_0)]^{-1}$:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{\mathcal{L}'(\theta_0 | X^{(n)})/\sqrt{n}}{-\mathcal{L}''(\theta_0 | X^{(n)})/n - (\hat{\theta}_n - \theta_0) \mathcal{L}'''(\theta_1 | X^{(n)})/2n}. \quad (7)$$

Числитель правой части этого представления

$$\frac{1}{\sqrt{n}} \mathcal{L}'(\theta_0 | X^{(n)}) = \frac{1}{\sqrt{n}} \sum_1^n \frac{\partial \ln f(X_k | \theta)}{\partial \theta}$$

есть нормированная на \sqrt{n} сумма независимых, одинаково распределенных случайных величин с нулевыми средними и дисперсиями $I(\theta_0) > 0$ (см. доказательство пункта 2⁰ леммы 4.2). Таким образом, в силу центральной предельной теоремы числитель правой части (7) асимптотически нормален с этими параметрами, и для завершения доказательства теоремы достаточно показать, что знаменатель (7) сходится по вероятности к постоянной $I(\theta_0)$, и сослаться на пункт (2) предложения 11.1 (теорема типа Слуцкого) курса ТВ.

В силу закона больших чисел и утверждения 1⁰ леммы 4.2 первое слагаемое в знаменателе (7)

$$-\frac{1}{n} \mathcal{L}''(\theta_0 | X^{(n)}) = -\frac{1}{n} \sum_1^n \frac{\partial^2 \ln f(X_k | \theta_0)}{\partial \theta^2} \xrightarrow{P} -\mathbf{E}_{\theta_0} \frac{\partial^2 \ln f(X | \theta_0)}{\partial \theta^2} = I(\theta_0),$$

так что остается показать, что и второе слагаемое сходится по вероятности к нулю.

Так как при выполнении условий (R1)–(R3) оценка максимального правдоподобия состоятельна, то $\hat{\theta}_n - \theta_0 \xrightarrow{P} 0$. Множитель при этой разности

$$\frac{1}{n} \mathcal{L}'''(\theta_1 | X^{(n)}) = \frac{1}{n} \sum_1^n \frac{\partial^3 \ln f(X_k | \theta_1)}{\partial \theta^3}$$

в силу условия (R4), начиная с некоторого n по абсолютной величине не превосходит $(1/n) \sum_1^n H_3(X_k)$ (это то n , при котором θ_1 попадает в окрестность точки θ_0). Применяя к этой сумме закон больших чисел, получаем, что она сходится по вероятности к

$$\mathbf{E}_{\theta_0} H_3(X) \leq \int_{\mathcal{X}} H_3(x) d\mu(x) < \infty,$$

и поэтому указанный выше сомножитель ограничен с вероятностью единица, а все второе слагаемое в знаменателе правой части (7) сходится по вероятности к нулю.

Доказанная теорема, как будет видно из основного результата следующего параграфа, устанавливает асимптотическую оптимальность оценок максимального правдоподобия с точки зрения квадратичного риска.