Глава 9

Правдоподобие и оценки максимального правдоподобия

Содержание

9.1. Физическая задача: эффективность регистрации 120
9.2. Оценка и её значение 120
9.3. Функция правдоподобия 121
9.4. Метод максимального правдоподобия 122
9.5. Биномиальная модель 123
9.6. Пуассоновский счёт 124
9.7. Среднее гауссова распределения 125
9.8. Неизвестны среднее и дисперсия 126
9.9. Что уже можно прочитать по функции правдоподобия 126
9.10. Задачи 127
Литература 128

Аннотация

В этой главе мы перейдём от вероятностной модели данных к оценке её параметров. Начнём с калибровки эффективности детектора, введём функцию правдоподобия и найдём её максимум. Затем тот же метод применим к пуассоновскому счёту и гауссову распределению. Последний пример покажет, что оценка максимального правдоподобия может быть смещённой.

9.1. Физическая задача: эффективность регистрации

Перед исследуемым детектором установлен другой прибор, который отмечает прохождение заряженной частицы. За время калибровки он зарегистрировал $N = 20$ частиц. Исследуемый детектор сработал для $k = 7$ из них. Требуется оценить его эффективность ε. Ошибками первого прибора и случайными совпадениями пока пренебрежём.

Каждая частица даёт один из двух результатов: детектор сработал или не сработал. Если результаты независимы, а эффективность во время калибровки постоянна, число зарегистрированных частиц имеет биномиальное распределение:

$$K \sim \mathrm{Bin}(N, \varepsilon). \tag{9.1}$$

Первая оценка напрашивается сама:

$$\hat{\varepsilon} = \frac{k}{N} = \frac{7}{20} = 0.35. \tag{9.2}$$

Почему именно эта оценка? Что означает её шляпка? Как изменится вывод, если детектор сработает семь раз из двадцати или семьдесят раз из двухсот? И что делать, если число зарегистрированных событий складывается из сигнала и известного фона? Для ответа запишем функцию правдоподобия.

9.2. Оценка и её значение

Пусть данные описываются распределением

$$X \sim f(x \mid \theta), \tag{9.3}$$

где θ — параметр или набор параметров модели. Это может быть эффективность детектора, среднее число событий, масса частицы, сечение процесса или коэффициент корреляции.

В частотном подходе θ фиксирован, но неизвестен. До эксперимента случайны данные. **Оценкой** параметра называется правило, которое ставит данным в соответствие число или вектор чисел:

$$\hat{\theta} = T(X_1, \dots, X_N). \tag{9.4}$$

После подстановки конкретной выборки получается значение оценки:

$$\hat{\theta}_{\text{data}} = T(x_1, \dots, x_N). \tag{9.5}$$

Это различие понадобится дальше. В формуле 9.4 оценка является случайной величиной и меняется от опыта к опыту. Число в формуле 9.5 уже получено из данных.

9.2.1. Один параметр — разные оценки

Пусть $X_1, \dots, X_N$ имеют гауссово распределение со средним μ. Оценивать μ можно выборочным средним, медианой или, например, полусуммой первого и последнего элементов выборки:

$$\hat{\mu}_1 = \frac{1}{N}\sum_{i=1}^{N} X_i, \qquad \hat{\mu}_2 = \operatorname{median}(X_1, \dots, X_N), \qquad \hat{\mu}_3 = \frac{X_1 + X_N}{2}. \tag{9.6}$$

Все три правила дают оценку одного параметра. При гауссовом распределении они несмещённые, но разброс у них различен. Третья оценка, кроме того, выбрасывает почти всю собранную информацию. Условие задачи само по себе ещё не выбирает оценку. Нужен принцип, по которому этот выбор будет сделан.

Точечная оценка даёт одно значение $\hat{\theta}$. Интервальная оценка задаёт область допустимых значений параметра. В этой главе мы строим точечные оценки. Интервалы потребуют отдельной статистической процедуры, к которой мы перейдём позже.

9.3. Функция правдоподобия

При заданном θ функция $f(x \mid \theta)$ описывает распределение возможных данных. После эксперимента данные $\mathbf{x}$ фиксированы, а параметр остаётся неизвестным. Рассмотрим ту же функцию как функцию параметра:

$$\boxed{L(\theta; \mathbf{x}) = f(\mathbf{x} \mid \theta)}. \tag{9.7}$$

Она называется **функцией правдоподобия**. Вертикальная черта в $f(\mathbf{x} \mid \theta)$ напоминает, что распределение данных задаётся при фиксированных параметрах. Точка с запятой в $L(\theta; \mathbf{x})$ отделяет аргумент функции от уже полученных данных.

Таблица 9.1. Вероятностная модель и функция правдоподобия используют одну математическую функцию, но отвечают на разные вопросы.

Функция	Фиксировано	Что меняется	Вопрос
$f(\mathbf{x} \mid \theta)$	θ	$\mathbf{x}$	Как распределены результаты повторных опытов?
$L(\theta; \mathbf{x})$	$\mathbf{x}$	θ	Какие параметры лучше согласуются с этими данными?

Правдоподобие не является распределением вероятности параметра. Для него нет условия нормировки по θ. Более того, множитель, который не зависит от θ, можно отбросить: положение максимума и отношения значений L от этого не изменятся.

9.3.1. Независимая выборка

Для независимых измерений совместное правдоподобие равно произведению:

$$L(\theta) = \prod_{i=1}^{N} f_i(x_i \mid \theta). \tag{9.8}$$

Индекс у f_i допускает разные распределения или разные ошибки отдельных измерений. Для одинаково распределённых величин все f_i совпадают.

Произведение многих малых чисел быстро достигает машинного нуля. Поэтому в вычислениях используют логарифм правдоподобия:

$$\ell(\theta) = \ln L(\theta) = \sum_{i=1}^{N} \ln f_i(x_i \mid \theta). \tag{9.9}$$

Логарифм монотонен, поэтому L и ℓ достигают максимума при одних и тех же параметрах. Независимые наборы данных добавляют к общей функции по одному слагаемому:

$$\ell_{\text{total}}(\theta) = \ell_1(\theta) + \ell_2(\theta) + \cdots. \tag{9.10}$$

В больших экспериментах результаты разных каналов и периодов набора данных объединяются сложением соответствующих логарифмов.

9.4. Метод максимального правдоподобия

Оценка максимального правдоподобия выбирает параметры, при которых полученные данные имеют наибольшее правдоподобие [1, 2]:

$$\boxed{\hat{\theta}_{\text{ММП}} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \ell(\theta)}. \tag{9.11}$$

Если максимум находится внутри допустимой области, а функция гладкая, то для одного параметра выполняются условия

$$\left.\frac{d\ell}{d\theta}\right|_{\theta=\hat{\theta}} = 0, \qquad \left.\frac{d^2\ell}{d\theta^2}\right|_{\theta=\hat{\theta}} < 0. \tag{9.12}$$

Слово «внутри» здесь существенно. Если параметр ограничен, максимум может оказаться на границе, где первая производная не равна нулю. Такой случай мы увидим уже в биномиальной задаче.

БИОГРАФИЯ

Рональд Эйлмер Фишер

1890–1962

Рональд Фишер — британский статистик и генетик. В работах 1920-х годов он придал современную форму теории статистического оценивания, ввёл понятия достаточной статистики и информации Фишера и систематически развил метод максимального правдоподобия.

Работая на Ротамстедской опытной станции, Фишер создал методы планирования и анализа эксперимента, включая дисперсионный анализ. В генетике он связал менделевское наследование с естественным отбором и стал одним из основателей популяционной генетики. Его имя встретится ещё раз при обсуждении информации, которую данные содержат о параметре.

Источник: MacTutor; фото неизвестный фотограф, 1913; общественное достояние

9.5. Биномиальная модель

Вернёмся к калибровке детектора. При N независимых частицах и эффективности ε вероятность зарегистрировать ровно k частиц равна

$$P(K = k \mid \varepsilon) = \binom{N}{k} \varepsilon^k (1-\varepsilon)^{N-k}. \tag{9.13}$$

После того как N и k получены, эта формула рассматривается как функция ε. Комбинаторный множитель от эффективности не зависит, поэтому

$$L(\varepsilon) \propto \varepsilon^k (1-\varepsilon)^{N-k}. \tag{9.14}$$

9.5.1. Оценка эффективности

Сначала рассмотрим $0 < k < N$, когда максимум лежит внутри интервала $0 < \varepsilon < 1$. Логарифм правдоподобия с точностью до постоянного слагаемого имеет вид

$$\ell(\varepsilon) = k \ln \varepsilon + (N-k) \ln(1-\varepsilon) + \text{const}. \tag{9.15}$$

Его производная равна

$$\frac{d\ell}{d\varepsilon} = \frac{k}{\varepsilon} - \frac{N-k}{1-\varepsilon}. \tag{9.16}$$

Условие максимума даёт

$$k(1-\varepsilon) = \varepsilon(N-k),$$

откуда

$$\boxed{\hat{\varepsilon}_{\text{ММП}} = \frac{k}{N}}. \tag{9.17}$$

Так получено значение оценки для конкретных данных. До эксперимента вместо k следует писать случайную величину K:

$$\hat{\varepsilon} = \frac{K}{N}. \tag{9.18}$$

Из свойств биномиального распределения следуют среднее и дисперсия оценки:

$$\mathbb{E}[\hat{\varepsilon}] = \varepsilon, \qquad \operatorname{Var}(\hat{\varepsilon}) = \frac{\varepsilon(1-\varepsilon)}{N}. \tag{9.19}$$

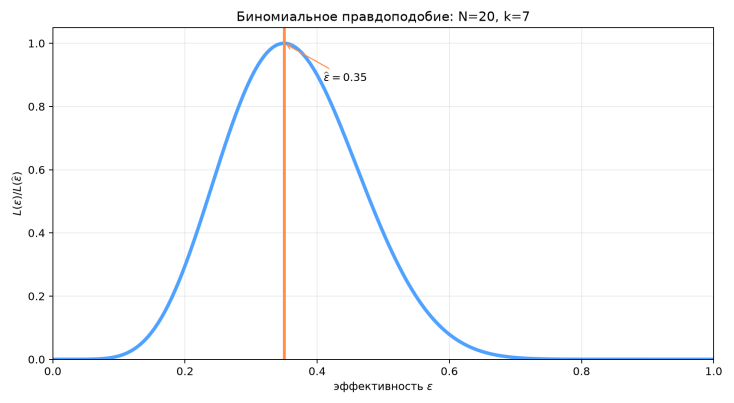
Значит, в этой задаче оценка максимального правдоподобия несмещённая, а её разброс убывает как $1/\sqrt{N}$. Неизвестную ε в формуле для стандартного отклонения часто заменяют её оценкой:

$$\hat{\sigma}_{\hat{\varepsilon}} = \sqrt{\frac{\hat{\varepsilon}(1-\hat{\varepsilon})}{N}}. \tag{9.20}$$

Для нашей калибровки $\hat{\varepsilon} = 0.35$ и $\hat{\sigma}_{\hat{\varepsilon}} \simeq 0.107$. Последнее число пока не следует читать как готовый доверительный интервал. Особенно плохо такая интерпретация работает возле $\varepsilon = 0$ и $\varepsilon = 1$.

Если $k = 0$, функция правдоподобия монотонно убывает и максимум находится при $\hat{\varepsilon} = 0$. При $k = N$ она монотонно возрастает и максимум равен $\hat{\varepsilon} = 1$. В обоих случаях максимум лежит на физической границе, а уравнение 9.16 для внутренней точки применять нельзя.

9.5.2. Интерактив: форма биномиального правдоподобия



Интерактив. Биномиальное правдоподобие
Меняйте число частиц N и число срабатываний k. Вертикальная линия показывает оценку $\hat{\varepsilon} = k/N$. При увеличении выборки максимум становится уже.

На рисунке 9.1 отношение k/N сохранено, но число испытаний увеличено в пять раз. Положение максимума не изменилось, а значения эффективности вдали от 0.35 теперь подавлены гораздо сильнее.

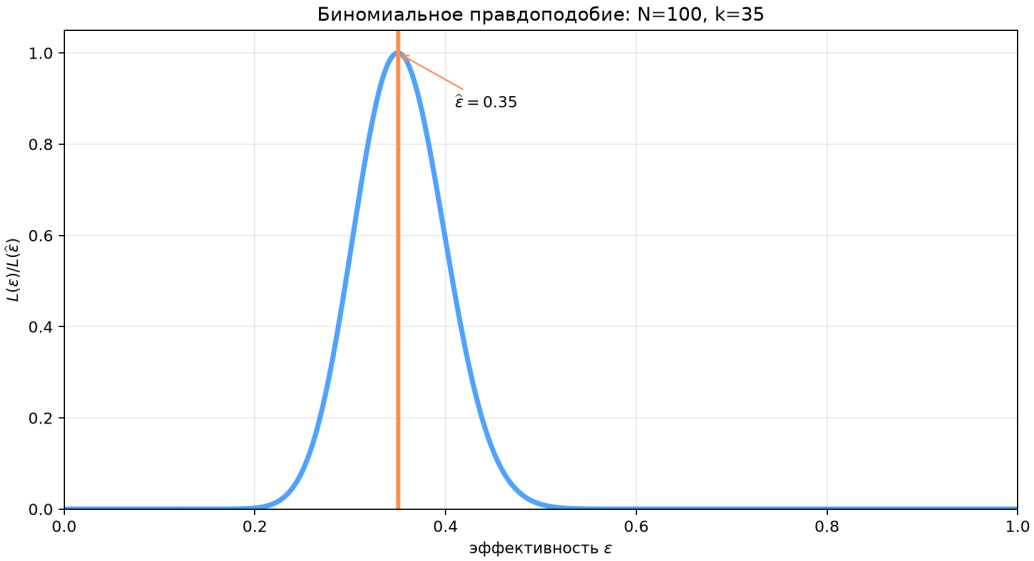


Рис. 9.1. *Относительное правдоподобие эффективности при $N = 100$ и $k = 35$. Вертикальная линия соответствует $\hat{\varepsilon} = 0.35$.*

Биномиальная модель предполагает, что каждая отмеченная первым прибором частица действительно прошла через исследуемый детектор, а вероятность его срабатывания оставалась постоянной. Фоновые совпадения, зависимость эффективности от энергии и неоднородность детектора добавят в правдоподобие новые параметры. Одна дробь k/N такую калибровку уже не описывает.

9.6. Пуассоновский счёт

Пусть N — число событий, зарегистрированных за фиксированное время, а μ — ожидаемое число событий:

$$N \sim \mathrm{Pois}(\mu). \tag{9.21}$$

Если получено n событий, правдоподобие равно

$$L(\mu) = \frac{\mu^n e^{-\mu}}{n!}, \qquad \mu \geq 0. \tag{9.22}$$

Для $n > 0$

$$\ell(\mu) = n \ln \mu - \mu - \ln n!, \qquad \frac{d\ell}{d\mu} = \frac{n}{\mu} - 1. \tag{9.23}$$

Следовательно,

$$\boxed{\hat{\mu}_{\text{ММП}} = n}. \tag{9.24}$$

При $n = 0$ максимум находится на границе $\mu = 0$, и ответ остаётся тем же. Поскольку до эксперимента $\hat{\mu} = N$, а $\mathbb{E}[N] = \mu$, эта оценка несмещённая.

9.6.1. Известный фон

Теперь пусть среднее число событий складывается из неизвестного сигнала s и известного фона b:

$$N \sim \mathrm{Pois}(s + b), \qquad s \geq 0. \tag{9.25}$$

Если временно разрешить s принимать любые значения при условии $s + b \geq 0$, максимум находится при

$$\hat{s}_{\text{unconstrained}} = n - b. \tag{9.26}$$

Однако отрицательное среднее число сигнальных событий не имеет физического смысла. С учётом ограничения $s \geq 0$ получаем

$$\boxed{\hat{s}_{\text{ММП}} = \max(0, n - b)}. \tag{9.27}$$

При $n < b$ производная логарифма правдоподобия в разрешённой области не обращается в нуль. Максимум находится в точке $s = 0$. Физическая граница меняет статистическую процедуру, и обычное гауссово приближение возле максимума в этой ситуации требует отдельной проверки.

9.7. Среднее гауссова распределения

Пусть независимые измерения имеют одинаковую известную ошибку σ:

$$X_i \sim \mathrm{N}(\mu, \sigma^2), \qquad i = 1, \ldots, N. \tag{9.28}$$

Требуется оценить μ. Логарифм правдоподобия с точностью до постоянного слагаемого равен

$$\ell(\mu) = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2 + \text{const}. \tag{9.29}$$

Максимизация ℓ совпадает с минимизацией суммы квадратов отклонений. Из условия

$$\frac{d\ell}{d\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i - \mu) = 0 \tag{9.30}$$

получаем

$$\boxed{\hat{\mu}_{\text{ММП}} = \frac{1}{N}\sum_{i=1}^{N} x_i = \overline{x}}. \tag{9.31}$$

Метод максимального правдоподобия выбрал из трёх оценок 9.6 выборочное среднее. Связь между гауссовым правдоподобием и методом наименьших квадратов будет подробно использована в следующих главах.

Если ошибки отдельных измерений σ_i различаются, в сумме появляется вес $1/\sigma_i^2$. Максимум правдоподобия тогда даёт взвешенное среднее, полученное в формуле 7.36.

9.8. Неизвестны среднее и дисперсия

Пусть теперь требуется одновременно оценить μ и σ^2. Полный логарифм правдоподобия имеет вид

$$\ell(\mu, \sigma^2) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2. \tag{9.32}$$

Нормировочный множитель гауссова распределения теперь зависит от σ^2, поэтому отбрасывать первое слагаемое нельзя. Максимизация по двум параметрам даёт

$$\boxed{\hat{\mu}_{\text{ММП}} = \overline{x}, \qquad \widehat{\sigma^2}_{\text{ММП}} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}. \tag{9.33}$$

Оценка среднего несмещённая. Для дисперсии получается другой результат:

$$\mathbb{E}\left[\widehat{\sigma^2}_{\text{ММП}}\right] = \frac{N-1}{N}\sigma^2. \tag{9.34}$$

Делитель N появился из условия максимума правдоподобия. Несмещённая оценка дисперсии использует делитель $N - 1$:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2, \qquad \mathbb{E}[s^2] = \sigma^2. \tag{9.35}$$

Здесь оценка максимального правдоподобия и несмещённая оценка расходятся. Метод максимального правдоподобия задаёт способ построения оценки. Несмещённость — отдельное свойство этой оценки. Для $\widehat{\sigma^2}_{\text{ММП}}$ смещение исчезает при $N \to \infty$, но при конечном N оно существует.

9.9. Что уже можно прочитать по функции правдоподобия

Положение максимума даёт точечную оценку. Ширина максимума показывает, как быстро ухудшается согласие модели с данными при смещении параметра. С ростом выборки максимум обычно сужается, как в биномиальном примере. Если максимум прижат к границе, форма функции становится несимметричной.

Эти наблюдения пока остаются качественными. Чтобы превратить ширину правдоподобия в ошибку оценки или интервал, потребуется определить статистическую процедуру и проверить её свойства. Этому посвящена следующая глава.

Итоги главы

- Оценка является функцией случайных данных; после подстановки выборки она принимает конкретное значение.
- Правдоподобие сравнивает значения параметров при фиксированных данных и не является распределением вероятности параметра.
- Для независимых измерений правдоподобия перемножаются, а их логарифмы складываются.
- Оценка максимального правдоподобия соответствует максимуму L в допустимой области параметров.
- Метод максимального правдоподобия не гарантирует несмещённость: для гауссовой дисперсии оценка с делителем N смещена при конечной выборке.

9.10. Задачи

Задача 1. Вероятность и правдоподобие

Одно наблюдение x получено из экспоненциального распределения

$$f(x \mid \lambda) = \lambda e^{-\lambda x}, \qquad x \geq 0.$$

1. Нарисуйте $f(x \mid \lambda)$ как функцию x при нескольких фиксированных λ.
2. После наблюдения $x = x_{\text{obs}}$ запишите $L(\lambda; x_{\text{obs}})$.
3. Вычислите $\int_0^\infty L(\lambda; x_{\text{obs}})\, d\lambda$. Почему этот интеграл не является условием нормировки вероятности параметра?
4. Найдите $\hat{\lambda}_{\text{ММП}}$.
5. Объясните словами различие между плотностью вероятности и функцией правдоподобия.

Задача 2. Биномиальная оценка

В N независимых испытаниях наблюдалось k успехов.

1. Запишите биномиальное правдоподобие $L(p)$.
2. Найдите $\hat{p}_{\text{ММП}}$.
3. Вычислите $\mathrm{E}[\hat{p}]$ и $\mathrm{Var}(\hat{p})$.
4. Покажите, что оценка несмещённая и состоятельная.
5. Исследуйте форму $L(p)$ для случаев $k = 0$, $k = N/2$ и $k = N$.

Задача 3. Пуассоновский счёт с известным фоном

Пусть

$$N \sim \mathrm{Pois}(s + b),$$

где $s \geq 0$ — неизвестный сигнал, а фон b известен точно.

1. Запишите $L(s)$ и $\ell(s)$.
2. Найдите неограниченную оценку максимального правдоподобия сигнала.
3. Найдите оценку максимального правдоподобия при физическом ограничении $s \geq 0$.
4. Объясните, что происходит при $n < b$.

Задача 4. Гауссово среднее

Пусть

$$X_i \sim \mathrm{N}(\mu, \sigma^2),$$

где σ известна.

1. Запишите функцию правдоподобия независимой выборки.
2. Покажите, что $\hat{\mu}_{\mathrm{ММП}} = \langle X \rangle$.
3. Найдите $\mathrm{Var}(\hat{\mu}_{\mathrm{ММП}})$.

Задача 5. Взвешенное среднее

Имеются независимые измерения

$$X_i \sim \mathrm{N}(\mu, \sigma_i^2),$$

где σ_i известны и различаются.

1. Запишите совместный логарифм функции правдоподобия.
2. Найдите оценку максимального правдоподобия параметра μ.
3. Покажите, что результат является взвешенным средним с весами $w_i = 1/\sigma_i^2$.
4. Найдите дисперсию оценки.
5. Объясните, почему измерения с меньшей ошибкой получают больший вес.

Литература

[1] Fisher, R. A. On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London. Series A*, **222**, 309–368. 1922. doi:10.1098/rsta.1922.0009.
[2] Cowan, G. *Statistical Data Analysis.* Oxford University Press. 1998.