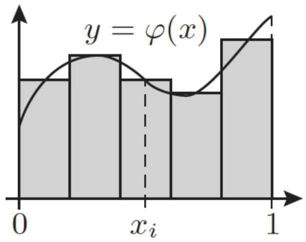
Метод Монте-Карло

Метод Монте-Карло — метод решения математических задач при помощи случайных чисел. Знакомство с методом начнем с рассмотрения задачи численного интегрирования функции $\phi(x)$, заданной на отрезке $[0,1]$. Как вычислить приближенно интеграл $I = \int_0^1 \phi(x)dx$? Пожалуй, простейший способ — *метод прямоугольников.* Он состоит в замене I на интегральнуюvсумму $I_n = \frac{1}{n}\sum_{i=1}^n \phi(x_i)$, где $x_i = \frac{i-1/2}{n}$ — это *«узлы» равномерной сетки*, т. е. середины интервалов разбиения отрезка $[0,1]$ на n равных частей.



При условии, что $\phi(x)$ дважды непрерывно дифференцируема, можно показать, что *погрешность метода прямоугольников* $\delta_n = |I - I_n|$ оцениваться сверху так:

$$\delta_n \le \frac{M}{24n^2}, \quad \text{где} \quad M = \max_{x\in[0,1]} |\phi''(x)|.$$

Таким образом, для гладких функций погрешность метода прямоугольников имеет порядок малости $1/n^2$.

Метод Монте-Карло для вычисления интеграла I отличается от метода прямоугольников тем, что в качестве «узлов» используются случайные числа $y_1, \dots, y_n$. Пусть случайные величины $X_1, \dots, X_n$ — независимы и равномерно распределены на $[0,1]$. Тогда

$$\mathbb{E}\phi(X_i) = \int_{-\infty}^{+\infty} \phi(u) f_{X_i}(u)du = \int_0^1 \phi(u)du = I.$$

Согласно закону больших чисел

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i) \xrightarrow{P} I \qquad \text{при } n \to \infty.$$

Здесь $\xrightarrow{P}$ обозначает *сходимость по вероятности*:

$$\text{для любого } \varepsilon > 0 \qquad \mathbb{P}(|\hat{I}_n - I| > \varepsilon) \to 0 \text{ при } n \to \infty.$$

Таким образом, с ростом n погрешность приближения должна стремиться к нулю. Кроме того, если $\sigma^2 = \mathrm{Var}\phi(X_i) < \infty$, то по центральной предельной теореме (ЦПТ)

$$\mathbb{P}\left(x \le \frac{\sqrt{n}|\hat{I}_n - I|}{\sigma} \le y\right) \to \mathbb{P}(x \le Z \le y),$$

где $Z \sim \mathcal{N}(0,1)$. Если положить $x = -3$ и $y = 3$, то мы получим $P(-3 \le Z \le 3) \approx 0.997$. Мы сейчас получили, что при достаточно больших n (когда ЦПТ дает хорошую аппроксимацию) выполняется неравенство

$$|\hat{I}_n - I| \le \frac{3\sigma}{\sqrt{n}} \quad \text{с вероятностью близкой к 1.}$$

Заметим, что ошибка оценки интеграла $|\hat{I}_n - I|$ убывает как $1/\sqrt{n}$ с ростом n. Говорят, что $\hat{I}_n$ является «оценкой» I, а интервал

$$\hat{I}_n - \frac{3\sigma}{\sqrt{n}} \leq I \leq \hat{I}_n + \frac{3\sigma}{\sqrt{n}}$$

называют «доверительным интервалом» оценки $\hat{I}_n$.

Отметим, что в выводе мы воспользовались так называемым *«правилом трех сигм»*: случайная величина $X \sim \mathcal{N}(\mu, \sigma^2)$ принимает значения из отрезка $[\mu - 3\sigma, \mu + 3\sigma]$ с вероятностью 0.997, которую зачастую не отличают от 1.

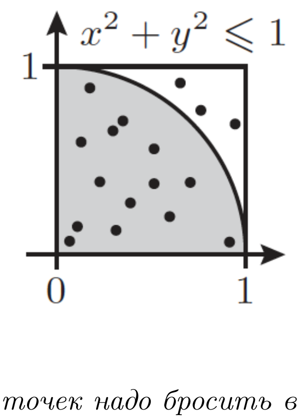
Неразумно использовать метод Монте-Карло для вычисления одномерных интегралов — для этого существуют квадратурные формулы, простейшая из которых — рассмотренная выше формула метода прямоугольников. Дело в том, что погрешность метода Монте-Карло больше, чем погрешность метода прямоугольников (да и верна эта оценка погрешности лишь с некоторой вероятностью).

Тем не менее, метод Монте-Карло (или его модификации) часто оказывается единственным численным методом, позволяющим решить задачу вычисления интеграла большой кратности. Дело в том, что число «узлов» сетки возрастает как n^k, где k — кратность интеграла (так называемое «проклятие размерности»). Так, чтобы найти интеграл по десятимерному кубу, используя в качестве «узлов» только его вершины, надо $2^{10} = 1024$ раза вычислить значение интегрируемой функции. В практических задачах эти вычисления могут оказаться довольно долгими, например, когда для расчета значений требуется численное решение систем нелинейных или дифференциальных уравнений.

Напротив, метод Монте-Карло не зависит от размерности: чтобы найти приближенное значение интеграла

$$\int_0^1 \ldots \int_0^1 \phi(u_1, \ldots, u_n) du_1 \ldots du_n$$

с точностью порядка $1/\sqrt{n}$ достаточно случайно набросать n точек в k-мерный единичный куб (разбив случайные числа на группы из k элементов) и вычислить среднее арифметическое значений ϕ в этих точках. В частности, если функция — индикатор некоторой области D, то с помощью метода Монте-Карло можно приближенно определить объем этой области. Например, частота случайных точек, попавших под дугу окружности будет служить приближением к $\pi/4$.



Упражнение 1. *Сколько случайных точек надо бросить в единичный квадрат, чтобы получить площадь под дугой окружности (см. рисунок выше) с точностью* 0.001 *и с вероятностью* 0.997*?*

Сравнение оценок

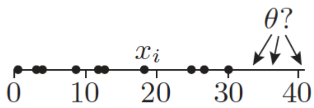
Анализируемые методами математической статистики данные обычно рассматриваются как реализация выборки из некоторого распределения, известного с точностью до параметра (или нескольких параметров). При таком подходе для определения распределения, наиболее подходящего для описания данных, достаточно уметь оценивать значение параметра по реализации. В этой главе будет рассказано, как сравнивать различные оценки по точности.

Статистическая модель

Эксперимент. Пусть θ — некоторое *неизвестное* положительное число. Ниже приведены (с точностью до $0,1$) координаты x_i десяти точек, взятых наудачу из отрезка $[0,\theta]$.

$$3.5 \quad 3.2 \quad 25.6 \quad 8.8 \quad 11.6 \quad 26.6 \quad 18.2 \quad 0.4 \quad 12.3 \quad 30.1.$$

Попробуйте угадать значение параметра θ, на котором изображены эти точки.



С формальной точки зрения мы имеем дело со следующей моделью: набор x_i — это реализация независимых и равномерно распределенных на отрезке $[0,\theta]$ случайных величин X_i с функцией распределения

$$F_\theta(x) = \begin{cases} 0, & \text{если } x \le 0, \\ x/\theta, & \text{если } 0 < x < \theta, \\ 1, & \text{если } x \ge \theta. \end{cases}$$

Здесь $\theta \in \Theta = (0,+\infty)$ — неизвестный параметр масштаба.

Статистическая модель. В общем случае задается семейство функций распределения $\{F_\theta(x), \theta \in \Theta\}$, где Θ — множество возможных значений параметра; данные $x_1,\dots,x_n$ рассматриваются как реализация выборки $X_1,\dots,X_n$, элементы которой имеют функцию распределения $F_{\theta_0}(x)$ при некотором неизвестном значении $\theta_0 \in \Theta$. Задача состоит в том, чтобы оценить (восстановить) θ_0 по выборке $x_1,\dots,x_n$, по возможности, *наиболее точно.*

Как «угадать» задуманное значение, основываясь на наблюдениях $x_1,\dots,x_n$? Будем оценивать θ_0 при помощи некоторых функций $\hat\theta$ от n переменных $x_1,\dots,x_n$, которые называются *оценками* или *статистиками.*

Для приведенных выше данных эксперимента в качестве оценок неизвестного параметра масштаба можно использовать, скажем, $\hat\theta_1 = x_{(n)} = \max\{x_1,\dots,x_n\}$ и $\hat\theta_2 = 2(x_1+\ldots+x_n)/n$. Интуитивно понятно, что при увеличении n каждая из оценок будет приближаться именно к тому значению θ, с которым моделировалась выборка. Но какая из них точнее? Каким образом вообще можно сравнивать оценки? Прежде чем дать ответы на эти вопросы, познакомимся с важнейшими свойствами оценок — несмещенностью и состоятельностью.

Несмещенность и состоятельность

Определение. Оценка $\hat\theta(x_1,\dots,x_n)$ параметра θ называется *несмещенной*, если $\mathbb{E}_\theta\,\hat\theta(X_1,\dots,X_n) = \theta$ для всех $\theta \in \Theta$.

Здесь индекс θ у $\mathbb{E}_\theta$ означает, что имеется в виду математическое ожидание случайной величины

$\hat{\theta}(X_1, \ldots, X_n)$, где X_i распределены с функцией распределения $F_\theta(x)$. В дальнейшем этот индекс будет опускаться, чтобы формулы не выглядели слишком громоздко.

Замечание. Важно, чтобы условие несмещенности выполнялось для всех $\theta \in \Theta$. *Тривиальный контрпример*: оценка $\hat{\theta}(x_1, \ldots, x_n) = 1$, идеальная при $\theta = 1$, при других значениях θ имеет смещение

$$b(\theta) = \mathbb{E}\hat{\theta} - \theta = 1 - \theta.$$

Иногда представляет интерес получение оценки не для самого параметра θ, а для некоторой заданной функции $\phi(\theta)$.

Пример 1. Для выборочного контроля из партии готовой продукции отобраны n приборов. Пусть величины $X_1, \ldots, X_n$ — их времена работы до поломки. Допустим, что X_i одинаково показательно распределены с неизвестным параметром $\theta : F_\theta(x) = 1 - e^{-\theta x}, x > 0$. Требуется оценить *среднее время до поломки прибора*

$$\phi(\theta) = \mathbb{E}X_1 = \theta \int_0^{+\infty} xe^{-\theta x} dx = \frac{1}{\theta} \int_0^{+\infty} ye^{-y} dy = \frac{1}{\theta}.$$

По свойствам математического ожидания *выборочное среднее* $\overline{X} = \frac{X_1 + \ldots + X_n}{n}$ будет несмещенной оценкой для функции $\phi(\theta)$: $\mathbb{E}\overline{X} = \phi(\theta)$.

Пример 2. Рассмотрим выборку из какого-либо распределения с двумя параметрами μ и σ, где $\mu = \mathbb{E}X_1$ и $\sigma^2 = \mathrm{Var}X_1$ (скажем, нормального закона $\mathcal{N}(\mu, \sigma^2)$). По свойствам математического ожидания выборочное среднее $\overline{X}$ несмещенно оценивает параметр μ. В качестве оценки для неизвестной дисперсии $\phi(\sigma) = \sigma^2$ можно взять *выборочную дисперсию*

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X}^2.$$

Однако, оценка S^2 *имеет смещение*. Действительно, так как случайные величины X_i независимы и одинаково распределены, то, применяя свойства математического ожидания, получаем:

$$\begin{aligned}
\mathbb{E}S^2 &= \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} X_i^2 - \frac{1}{n^2} \sum_{i,j=1}^{n} X_i X_j\right) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}X_i^2 - \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}X_i^2 - \frac{1}{n^2} \sum_{i \neq j}^{n} \mathbb{E}X_i \mathbb{E}X_j \\
&= \mathbb{E}X_1^2 - \frac{1}{n}\mathbb{E}X_1^2 - \frac{n-1}{n}(\mathbb{E}X_1)^2 = \frac{n-1}{n}\mathrm{Var}X_1.
\end{aligned}$$

Чтобы устранить смещение, достаточно домножить S^2 на $n/(n-1)$.

Само по себе свойство несмещенности *не достаточно* для того, чтобы оценка хорошо приближала неизвестный параметр. Например, первый элемент X_1 выборки из закона Бернулли служит несмещенной оценкой для θ: $\mathbb{E}X_1 = 0 \cdot (1 - \theta) + 1 \cdot \theta = \theta$. Однако, его возможные значения 0 и 1 даже не принадлежат $\Theta = (0, 1)$. Необходимо, чтобы погрешность приближения стремилась к нулю с увеличением размера выборки. Это свойство в математической статистике называется *состоятельностью.*

Определение. Оценка $\hat{\theta}(x_1, \ldots, x_n)$ параметра θ называется *состоятельной*, если для всех $\theta \in \Theta$ последовательность

$$\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n) \xrightarrow{P} \theta \quad \text{при } n \to \infty.$$

Состоятельность оценки (а точнее — последовательности оценок $\{\hat{\theta}_n\}$) означает концентрацию вероятностной массы около истинного значения параметра с ростом размера выборки n.

Как установить, будет ли данная оценка состоятельной? Обычно оказывается полезным один из следующих трех способов:

1. Иногда удается доказать состоятельность, непосредственно вычисляя функцию распределения оценки.

2. Другой способ проверки состоит в использовании закона больших чисел и свойства сходимости: если случайные величины ξ_n сходятся по вероятности к случайной величине ξ, то есть $\xi_n \xrightarrow{P} \xi$, и функция $\phi(x)$ непрерывна, то $\phi(\xi_n) \xrightarrow{P} \phi(\xi)$.

3. Часто установить состоятельность помогает следующая лемма.

Лемма. *Если оценка $\hat{\theta}_n$ не смещена, $\mathbb{E}\hat{\theta}_n = \theta$, и дисперсия $\mathrm{Var}\hat{\theta}_n$ стремятся к нулю при $n \to \infty$, то оценка $\hat{\theta}_n$ состоятельна.*

Доказательство. Согласно неравенству Чебышева:

$$\mathbb{P}(|\hat{\theta}_n - \theta| > \varepsilon) \leq \frac{\mathrm{Var}\hat{\theta}_n}{\varepsilon^2} \to 0 \text{ при } n \to \infty.$$

Лемма доказана. □

Упражнение 2. *Для случайных величин $X_i, i = 1, \ldots, n$, взятых наудачу из отрезка $[0, \theta]$, докажите состоятельность оценки $\hat{\theta} = X_{(n)} = \max\{X_1, \ldots, X_n\}$ первым способом.*

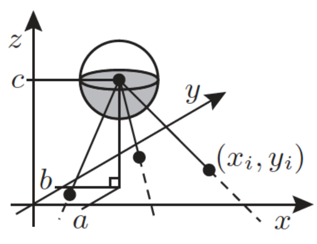
Упражнение 3. *Для случайных величин, распределенных экспоненциально с неизвестным параметром $\theta > 0$ (см. Пример 1), докажите состоятельность оценки $\hat{\theta} = 1/\overline{X}$ параметра θ вторым способом.*

Упражнение 4. *Для случайных величин $X_i, i = 1, \ldots, n$, взятых из распределения $\mathcal{N}(\theta, 1)$, докажите состоятельность оценки $\hat{\theta} = \overline{X} = (X_1 + \ldots + X_n)/n$ третьим способом.*

Тяжелые хвосты

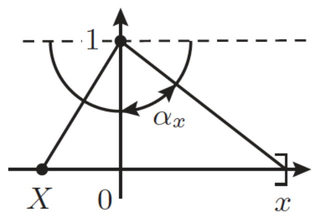
Распределение Коши

Рассмотрим следующий эксперимент по локализации источника излучения. В некоторой точке трехмерного пространства с неизвестными координатами (a, b, c) находится источник γ-излучения. Регистрируются координаты (x_i, y_i), $i = 1, \ldots, n$, точек пересечения траекторий γ-квантов с поверхностью детекторной плоскости $z = 0$. Требуется оценить параметры a и b по этим данным, предполагая, что направления траекторий γ-квантов случайны, т. е. равномерно распределены на сфере с центром в точке (a, b, c).



Какую оценку можно было бы предложить для (a, b)? Первое, что приходит в голову, — это $(\overline{X}, \overline{Y})$. Ясно, что точки пересечения траекторий с плоскостью $z = 0$ располагаются гуще непосредственно под источником излучения. В подобных случаях прибегают к усреднению данных, чтобы, по возможности, устранить разброс измерений (предполагается, что при этом происходит взаимная компенсация отклонений в разные стороны).

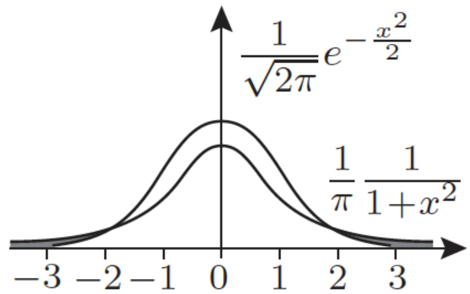
Однако, в данном случае усреднение *совершенно бесполезно*. Для объяснения, почему это так, рассмотрим одномерный аналог эксперимента: из точки (0, 1) выходит случайный луч, направление которого равномерно распределено на нижней полуокружности с центром $(0, 1)$. Случайная величина X — координата пересечения этого луча с осью абсцисс. Какая плотность $p(x)$ у этой величины?



Решение. Понятно, что плотность — четная функция. Вычислим ее для $x \geq 0$. Найдем сначала функцию распределения $F(x) = \mathbb{P}(X \geq x)$:

$$F(x) = \mathbb{P}(X \geq 0) + \mathbb{P}(0 < X \leq x) = \frac{1}{2} + \frac{\alpha_x}{\pi} = \frac{1}{2} + \frac{1}{\pi}\arctan(x).$$

Отсюда $p(x) = F'(x) = 1/[\pi(1+x^2)]$. Это — *плотность Коши.* На первый взгляд она похожа на плотность стандартного нормального закона $\mathcal{N}(0, 1)$. Однако, они различаются по скорости убывания к нулю при $x \to \infty$ вероятностей $\mathbb{P}(X \leq -x)$ и $\mathbb{P}(X \geq x)$ (так называемых «*хвостов распределения*»). У закона Коши «хвосты» намного «тяжелее».



Чем опасны «тяжелые хвосты»? Тем, что случайная величина с таким распределением с довольно существенной вероятностью может принимать большие по абсолютной величине значения. Поэтому в реализации выборки большого размера из такого закона обязательно появятся одно или несколько наблюдений, которые сильно отличаются от остальных (их называют «*выбросами*»). В этом случае при оценивании «центра» распределения при помощи выборочного среднего X произойдет резкое смещение оценки в сторону наибольшего «выброса».

Из-за слишком «тяжелых хвостов» у закона Коши не существует даже математического ожидания (проверьте!). Если бы оно существовало, то по усиленному закону больших чисел среднее арифметическое сходилось бы к $\mathbb{E}X_1$ по вероятности при $n \to \infty$. А что происходит с $\overline{X}$ для выборки из распределения Коши?

Ответ такой: $\overline{X}$ также имеет распределение Коши при любом n. Поэтому наблюдаемое значение $\overline{X}$ будет отклоняться от 0 ничуть не меньше значений самих X_i.

Как же, все-таки, *состоятельно* оценить θ в *модели сдвига* $F(x - \theta)$, когда F — функция распределения закона Коши? Подходящей оказывается, например, оценка, определяемая в следующем параграфе.

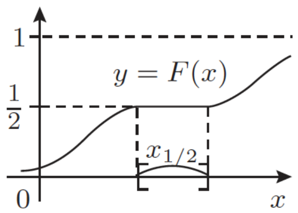
Выборочная медиана

Рассмотрим так называемый *вариационный ряд* $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$, состоящий из упорядоченных по возрастанию элементов выборки $(X_1, \ldots, X_n)$.

Определение. *Выборочной медианой* называется оценка

$$\mathrm{MED} = \begin{cases} X_{(k+1)}, & \text{если } n = 2k+1, \\ (X_{(k)} + X_{(k+1)})/2, & \text{если } n = 2k. \end{cases}$$

Выборочная медиана MED служит оценкой для *теоретической медианы* $x_{1/2}$, которая определяется как решение уравнения $F(x) = 1/2$, где $F(x)$ — функция распределения элементов выборки. Для непрерывной функции $F(x)$ решение всегда существует, но может быть не единственным.



Подобно математическому ожиданию, медиана $x_{1/2}$ является характеристикой, показывающей, где располагается «центр» распределения:

$$\mathbb{P}(X \le x_{1/2}) = \mathbb{P}(X \ge x_{1/2}) = \frac{1}{2}.$$

Пример 3 (Модель радиоактивного распада)**.** Как известно, радий (Ra) с течением времени превращается в радон (Rn). В момент распада атом радия излучает α-частицу — ядро атома гелия (He), и происходит переход $Ra \to Rn$. Допустим, что время τ до распада отдельного атома Ra не зависит от состояния других атомов и имеет показательное распределение: $p_t = \mathbb{P}(\tau > t) = e^{-\lambda t}$. Если имеется всего n атомов радия (в одном грамме насчитывается приблизительно 1022 атомов), то среднее число остающихся через время t атомов есть $n(t) = npt = ne^{-\lambda t}$. Определяемая из равенства $n(T) = n/2$ величина T (период полураспада) не зависит от исходного количества Ra: $T = \ln 2/\lambda$ (для радия $T \approx 1600$ лет). На языке теории вероятностей T — медиана показательного распределения.

Какими свойствами обладает MED как оценка для $x_{1/2}$?

Теорема 1. *Пусть элементы выборки имеют плотность $p(x)$, причем $p(x_{1/2}) > 0$. Тогда при $n \to \infty$ для любых $x, y \in \mathbb{R}$*

$$\mathbb{P}\big(x < \sqrt{n}(\mathrm{MED} - x_{1/2}) < y\big) \to \mathbb{P}(x < Z < y),$$

где $Z \sim \mathcal{N}\left(0, \frac{1}{4p^2(x_{1/2})}\right)$.

При выполнении условий Теоремы 1 выборочная медиана будет состоятельной оценкой для $x_{1/2}$.

Упражнение 5. *Почему* MED *состоятельна?*

Более того, из теоремы 1 следует, что точность оценки MED при больших n имеет порядок малости $1/\sqrt{n}$. Действительно, умножая $(\mathrm{MED} - x_{1/2})$ на $\sqrt{n}$ («коэффициент увеличения микроскопа»), мы получаем нечто «практически ограниченное» (с вероятностью 0,997 по «правилу трех сигм»).

Возвращаясь к оценке параметра сдвига распределения Коши, видим, что $x_{1/2} = \theta$ (это вытекает из симметрии плотности случайной величины X_i относительно θ), причем $p(x_{1/2}) = 1/\pi > 0$.

Следовательно, MED — состоятельная оценка для параметра сдвига.

Рассмотренный пример поучителен тем, что такая «естественная» оценка центра симметрии распределения и «сгущения» наблюдаемых значений элементов выборки, как $\overline{X}$, оказывается несостоятельной, и поэтому требуются более сложные оценки.

Методы получения оценок

В этой главе рассматриваются несколько методов получения оценок параметров статистических моделей, в том числе — метод моментов и метод максимального правдоподобия.

Plug-in оценки

Если нам необходимо оценить параметр θ и $\theta = \mathbb{E}\phi(X_1)$ для некоторой функции $\phi(x)$, то мы можем рассмотреть оценку вида

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\phi(X_i).$$

Нам уже встречались оценки такого типа. Например, в модели Бернулли, когда X_i имею вероятность успеха с неизвестным параметром $\theta \in (0,1)$, мы оценивали $\theta = \mathbb{E}X_1$ с помощью среднего $\overline{X}$. По закону больших чисел такие оценки оказываются несмещенными и состоятельными. Если $\mathrm{Var}\phi(X_1) < +\infty$, то можно построить доверительный интервал для plug-in оценки (как мы делали в главе про Монте-Карло).

Метод моментов

Моментом k-го порядка случайной величины X называется величина $\alpha_k = \mathbb{E}X^k$. Моменты существуют не всегда. Например, у закона Коши математическое ожидание α_1 не определено.

Положим $A_k = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^k$. Если момент α_k существует, то в силу закона больших чисел $A_k \xrightarrow{P} \alpha_k$ (это plug-in оценки моментов). Поэтому для реализации $x_1, \ldots, x_n$ выборки достаточно большого размера можно утверждать, что $a_k = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i^k \approx \alpha_k$, т. е. эмпирические моменты k-го порядка a_k близки к теоретическим моментам α_k. На этом соображении основывается так называемый метод моментов.

Допустим, что распределение элементов выборки зависит от m неизвестных параметров $\theta_1, \ldots, \theta_m$, где вектор $\theta = (\theta_1, \ldots, \theta_m)$ принадлежит некоторой области Θ в $\mathbb{R}^m$. Пусть $\mathbb{E}|X|^m < \infty$ для всех $\theta \in \Theta$ (отсюда следует конечность всех моментов до m из неравенства Ляпунова). Тогда существуют все $\alpha_k = \alpha_k(\theta)$, $k = 1, \ldots, m$, и можно записать систему из m (вообще говоря, нелинейных) уравнений

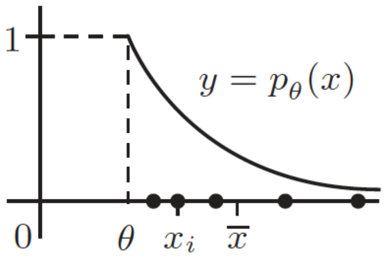
$$\alpha_k(\theta) = a_k, \quad k = 1, \ldots, m.$$

Предположим, что левая часть системы задает взаимно однозначное отображение $g : \Theta \to B$, где B — некоторая область в $\mathbb{R}^m$, и что обратное отображение $g^{-1} : B \to \Theta$ непрерывно. Другими словами, для всех $(y_1, ..., y_m)$ из B система имеет единственное решение, которое непрерывно зависит от правой части. Компоненты решения $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_m)$ при $y_k = A_k$ называются *оценками метода моментов*.

Пример 4. Рассмотрим модель сдвига показательного закона, в которой плотностью распределения величин X_i служит функция $p_\theta(x) = e^{-(x-\theta)} \cdot \mathbf{1}_{\{x\geq\theta\}}$. Здесь

$$\alpha_1(\theta) = \mathbb{E}X_1 = \int_\theta^{+\infty} xe^{-(x-\theta)}dx = \int_0^{+\infty} (y+\theta)e^{-y}dy = 1+\theta.$$

Из уравнения $1+\theta = A_1 = \overline{X}$, находим по методу моментов оценку $\hat{\theta} = \overline{X} - 1$.



Какими статистическими свойствами обладают оценки, полученные методом моментов? Их состоятельность вытекает из непрерывности определенного выше отображения g^{-1}.

Метод максимального правдоподобия

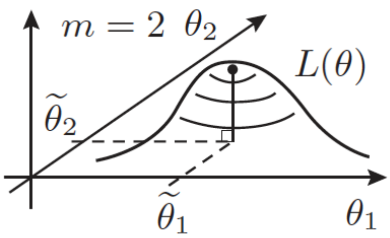
Метод получил распространение после появления в 1912 г. статьи Р. Фишера, где было доказано, что получаемые этим методом оценки являются асимптотически наиболее точными при выполнении некоторых условий регулярности модели.

Для знакомства с методом предположим для простоты, что элементы выборки X_i имеют дискретное распределение: $f(x,\theta) = P(X_1 = x)$ (здесь $\theta = (\theta_1, \ldots, \theta_m)$ — вектор неизвестных параметров модели). Тогда совместная вероятность выборки

$$f(x,\theta) = f(x_1,\theta) \cdot \ldots \cdot f(x_n,\theta)$$

зависит от $n+m$ аргументов (здесь $x = (x_1, ..., x_n)$). Рассматриваемая как функция от $\theta_1, \ldots, \theta_m$ при фиксированных значениях элементов выборки $x_1, \ldots, x_n$, она называется *функцией правдоподобия* и обычно обозначается через $L(\theta)$. Величину $L(\theta)$ можно считать мерой правдоподобия значения $L(\theta)$ при заданной реализации x.

Представляется разумным в качестве оценок параметров $\theta_1, \ldots, \theta_m$ взять наиболее правдоподобные значения $\tilde{\theta}_1, \ldots, \tilde{\theta}_m$ которые получаются при максимизации функции $L(\theta)$. Такие оценки называются *оценками максимального правдоподобия* (ОМП).



Часто проще искать точку максимума функции $\ln L(\theta)$, которая совпадает с $\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_m)$ в силу монотонности логарифма.

Пример 5. Для схемы Бернулли $X_1, \ldots, X_n$ с вероятностью «успеха» θ имеем:

$$f(x, \theta) = P(X_1 = x) = \theta^x (1-\theta)^{1-x},$$

где x принимает значения 0 или 1. Поэтому функция правдоподобия $L(\theta) = \theta^{s_n}(1-\theta)^{n-s_n}$, где $s_n = x_1 + \ldots + x_n$, представляет собой многочлен n-й степени. Найдем точку максимума

$$\ln L(\theta) = s_n \ln \theta + (n - s_n) \ln(1-\theta).$$

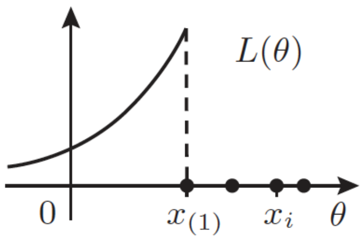
Дифференцируя по θ, получаем уравнение $s_n/\theta - (n - s_n)/(1-\theta) = 0$, откуда $\tilde{\theta} = s_n/n = \bar{x}$. Таким образом, ОМП в схеме Бернулли — это частота «успехов» в реализации $x_1, \ldots, x_n$.

В случае непрерывных моделей будем использовать обозначение $f(x, \theta)$ для плотности распределения случайной величины X_1.

Пример 6. Рассмотрим модель сдвига показательного закона с плотностью $f(x, \theta) = e^{-(x-\theta)}\, \mathbf{1}_{\{x \geq \theta\}}$. В этом случае функция правдоподобия равна

$$L(\theta) = \prod_{i=1}^{n} f(x_i, \theta) = e^{-(x_1 + \ldots + x_n)} e^{n\theta}\, \mathbf{1}_{\{x_{(1)} \geq \theta\}}.$$

Отсюда (см. рис. ниже) получаем в качестве ОМП $\tilde{\theta} = x_{(1)}$, которая отлична от оценки метода моментов $\hat{\theta} = \bar{x} - 1$, найденной ранее для этой модели в Примере 4. Заметим также, что здесь $L(\theta)$ не является гладкой функцией, и поэтому ОМП нельзя вычислять, приравнивая нулю производную функции правдоподобия.



В случае, когда $L(\theta)$ гладко зависит от $\theta_1, \ldots, \theta_m$, оценки максимального правдоподобия являются компонентами решения (вообще говоря, нелинейной) системы уравнений:

$$\frac{\partial}{\partial \theta_j} \ln L(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta_j} \ln f(x_i, \theta) = 0, \quad j = 1, \ldots, m.$$

Доверительные интервалы

Вместо того, чтобы приближать неизвестный скалярный параметр θ с помощью «точечной» оценки $\hat{\theta}$, можно локализовать его иначе — указать случайный интервал $(\hat{\theta}_1, \hat{\theta}_2)$, который накрывает θ с вероятностью близкой к единице.

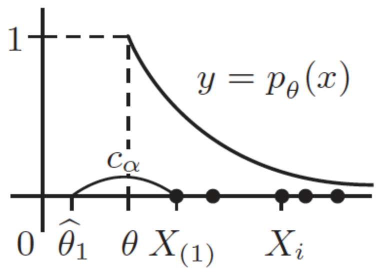
Определение. Пусть $\alpha \in (0, 1)$. Две статистики $\hat{\theta}_1$ и $\hat{\theta}_1$ определяют границы *доверительного интервала для параметра θ с коэффициентом доверия* $1 - \alpha$, если при всех $\theta \in \Theta$ для выборки $\mathbf{X} = (X_1, \ldots, X_n)$ из закона распределения $F_\theta(x)$ справедливо неравенство

$$\mathbb{P}\big(\hat{\theta}_1(\mathbf{X}) < \theta < \hat{\theta}_2(\mathbf{X})\big) \geq 1 - \alpha.$$

Часто на практике полагают $\alpha = 0.05$. Если вероятность в левой части неравенства стремится к $1 - \alpha$ при $n \to \infty$, то интервал называется *асимптотическим* (так у нас было, например, в главе про метод Монте-Карло). Как правило, длина доверительного интервала возрастает при увеличении коэффициента доверия $1-\alpha$ и стремится к нулю с ростом размера выборки n.

Пример 7. Для модели сдвига показательного закона с плотностью $p_\theta(x) = e^{-(x-\theta)}\,\mathbf{1}_{\{x\geq\theta\}}$ оценкой максимального правдоподобия является $X_{(1)} = \min\{X_1,\ldots,X_n\}$. Поскольку $\theta < X_{(1)}$, можно взять $X_{(1)}$ в качестве $\hat\theta_2$. Попробуем подобрать константу c_α так, чтобы для $\hat\theta_1 = X_{(1)} - c_\alpha$ (см. рис.) при всех θ выполнялось тождество

$$\mathbb{P}(X_{(1)} - c_\alpha < \theta < X_{(1)}) = \mathbb{P}(X_{(1)} - c_\alpha < \theta) = 1-\alpha.$$



Используя независимость и показательность величин $X_i - \theta$, перепишем условие:

$$\alpha = \mathbb{P}(X_{(1)} - \theta \geq c_\alpha) = \mathbb{P}(X_i - \theta \geq c_\alpha,\ i=1,\ldots,n) = e^{-nc_\alpha}.$$

Откуда находим, что длина интервала $c_\alpha = (-\ln\alpha)/n$. Отметим, что $c_\alpha \to \infty$ при $\alpha \to 0$ и $c_\alpha \to 0$ при $n \to \infty$.

Пример 8. Допустим, что элементы выборки X_i распределены по закону $\mathcal{N}(\theta,\sigma^2)$, причем параметр масштаба σ известен, а параметр сдвига θ — нет. Эту модель часто применяют к данным, полученным при независимых измерениях некоторой величины θ с помощью прибора (или метода), имеющего известную среднюю погрешность (стандартную ошибку) σ.

Пусть $\Phi(x)$ — функция распределения закона $\mathcal{N}(0,1)$. Для $0<\alpha<1$ обозначим через x_α так называемую α-квантиль этого закона, т. е. решение уравнения $\Phi(x_\alpha) = \alpha$. Приведем некоторые значения $x_{1-\alpha/2}$:

α	0,05	10^{-2}	10^{-3}	10^{-5}
$x_{1-\alpha/2}$	1,96	2,58	3,29	4,26

ОМП оценкой для θ служит $\overline{X}$. Известно, что $\overline{X} \sim \mathcal{N}(\theta,\sigma^2/n)$. Тогда $\sqrt{n}(\overline{X}-\theta)/\sigma \sim \mathcal{N}(0,1)$. Поэтому в качестве границ интервала с коэффициентом доверия $1-\alpha$ можно взять

$$\hat\theta_1 = \overline{X} - \frac{\sigma x_{1-\alpha/2}}{\sqrt{n}} \quad \text{и} \quad \hat\theta_2 = \overline{X} - \frac{\sigma x_{\alpha/2}}{\sqrt{n}}.$$

Мы получили

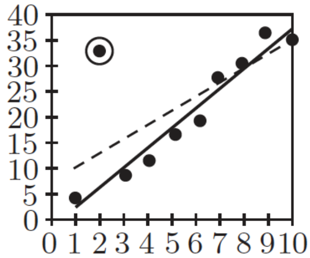
$$\mathbb{P}(\hat\theta_1 < \theta < \hat\theta_2) = \mathbb{P}(x_{\alpha/2} < \sqrt{n}(\overline{X}-\theta)/\sigma < x_{1-\alpha/2}) = 1-\alpha.$$

В силу четности плотности закона $\mathcal{N}(0,1)$ верно равенство $x_{\alpha/2} = -x_{1-\alpha/2}$. Таким образом, из приведенной выше таблицы видим, что с вероятностью $0,95$ истинное значение параметра сдвига θ находится в интервале $\overline{X} \pm 1.96\sigma/\sqrt{n} \approx 2\sigma/\sqrt{n}$ (правило двух сигм).

Робастность

В реальных данных доля «выбросов» (выделяющихся значений) обычно составляет от 1% до 10%. Это происходит из-за большого числа неучтенных факторов (в медицине, психологии), сбоев оборудования, скажем, скачков напряжения в электросети (в экспериментальной физике), ошибках при вводе с клавиатуры чисел в компьютер и т. д. Даже в астрономических таблицах встречается до 0,1% ошибок.

Казалось бы, можно придерживаться такой стратегии борьбы с «выбросами»: найти их и исключить, а затем применить эффективные методы для анализа оставшихся данных. Конечно, среди точек на прямой «выброс» хорошо заметен. Но реальные данные, как правило, многомерные. На рисунке ниже приведена двумерная выборка, где обведенная кружком точка (очевидный «выброс») не выделяется среди остальных ни по координате X, ни по координате Y. Однако, если попытаться формально подогнать прямую под это «облако» точек, то ее угловой коэффициент будет существенно искажен под влиянием «выброса».



Возможна ситуация, когда даже проецирование многомерных данных на всевозможные двумерные плоскости не позволит выявить выделяющиеся наблюдения.

Так что исключение «многомерного выброса» (или группы «выбросов») — весьма непростая задача. Методы, которые устойчивы к выбросам, называются *робастными*. Как правило, они менее точны на «чистых данных» (бесплатным бывает только сыр в мышеловке), но зато меньше подвержены влиянию выделяющихся наблюдений и не допускают существенного смещения оценок параметров модели.

Список литературы

[1] М. Б. Лагутин. Наглядная математическая статистика. Бином, 2009.