Математическая статистика

лектор: Дарина Михайловна Двинских

19 июня 2025 г.

Содержание

Обозначения 3

Лекции 4

1 Основные понятия математической статистики. Вероятностно-статистическая модель, выборка, статистика, оценка. Свойства оценок. . . . 4
- 1.1 Мотивационный пример: выборы . . . 4
- 1.2 Замечание . . . 5
- 1.3 Основные определения . . . 5
- 1.4 Свойства оценок . . . 6

2 Введение в теорию точечного оценивания параметров. Сравнение оценок, функция потерь, функция риска. Оптимальная оценка. . . . 8
- 2.1 Сравнение оценок . . . 8
- 2.2 Оптимальные оценки . . . 10

3 Эмпирическая функция распределения, ее свойства. Порядковые статистики, вариационный ряд выборки, выборочные квантили. . . . 12
- 3.1 Эмпирическое функция распределения . . . 12
- 3.2 Порядковые статистики. . . . 14
- 3.3 Выборочные моменты. . . . 15

4 Неравенство Рао-Крамера и эффективные оценки. Сверхэффективные оценки. Экспоненциальное семейство распределений. . . . 18
- 4.1 Обобщенная плотность. Функция правдоподобия . . . 18
- 4.2 Информация Фишера. Неравенство Рао-Крамера . . . 19

5 Экспоненциальное семейство распределений. Эффективные оценки в экспоненциальном семействе. . . . 25
- 5.1 Экспоненциальное семейство распределений. . . . 25
- 5.2 Эффективные оценки в экспоненциальном семействе . . . 26
- 5.3 Многомерное неравенство Рао-Крамера. . . . 28

6 Оптимальная оценка в модели линейной регрессии . . . 29
- 6.1 Мотивационный пример . . . 29
- 6.2 Линейная регрессия . . . 29
- 6.3 Метод наименьших квадратов . . . 30
- 6.4 Статистические свойства и геометрическая интерпретация . . . 31
- 6.5 Теорема Гаусса-Маркова . . . 32

7 Метод максимального правдоподобия. Состоятельность и асимптотическая нормальность оценок максимального правдоподобия. . . . 34
- 7.1 Метод максимального правдоподобия. . . . 34

7.2 Дивергенция Кульбака-Лейблера. 35
7.3 Асимптотические свойства оценок максимального правдоподобия. 36
8 Доверительные интервалы: точные и асимптотические. 38
8.1 Метод центральной статистики 38
8.2 Асимптотические доверительные интервалы 40
9 Проверка статистических гипотез. 43
9.1 Ошибки первого и второго рода, уровень значимости. 44
9.2 Параметрические гипотезы. 44
10 Критерий Неймана-Пирсона для проверки параметрических гипотез. 47
10.1 Простые гипотезы 47
10.2 Сложные гипотезы 49
11 Критерий хи-квадрат Пирсона. p-value. 51
11.1 Критерий согласия хи-квадрат Пирсона. Простая гипотеза 51
11.2 Критерий согласия хи-квадрат Пирсона. Сложная гипотеза 53
11.3 p-value 54
12 Гипотеза и критерий однородности. 55
12.1 Критерий хи-квадрат Пирсона. 55
12.2 Параметрический случай 57
13 Байесовская статистика. Априорное и апостериорное распределение. Байесовская оценка. Сопряженные семейства распределений. 59
13.1 Сопряженные распределения. 61
13.2 Байесовским доверительные интервалы 62
14 Байесовские и минимаксные оценки 63
14.1 Три подхода в сравнении оценок 64

Вспомогательные материалы **66**

Обозначения

Распределения

$\mathrm{Bin}(n, p)$	Биномиальное распределение с n экспериментами и вероятностью успеха p
$\chi^2(n)$	Распределение хи-квадрат с n степенями свободы
$\mathcal{N}(\mu, \Sigma)$	Многомерное нормальное распределение со средним μ и ковариационной матрицей Σ
$\mathcal{N}(\mu, \sigma^2)$	Одномерное нормальное распределение со средним μ и дисперсией σ^2
$\mathrm{Exp}(\lambda)$	Экспоненциальное (показательное) распределение с параметром λ
$\Gamma(k, \theta)$	Гамма-распределение с параметрами (k, θ)
$\mathrm{Pois}(\theta)$	Распределение Пуассона с параметром $\theta > 0$
$\mathcal{U}(a, b)$	Равномерное распределение на интервале $(a; b)$
$F(m, n)$	Распределение Фишера (F-распределение)
$t(n)$	Распределение Стьюдента (t-распределение) с n степенями свободы

Вероятность

$\mathbb{P}(A)$	Вероятность события A
$\mathbb{E}[X]$	Математическое ожидание с.в. X
$\mathrm{Var}[X]$	Дисперсия с.в. X
$\mathrm{cov}(X, Y)$	Ковариация с.в. X и Y
$\overset{\text{i.i.d}}{\sim}$	независимые и одинаково распредленные (independent and identically distributed)
$\xrightarrow{a.s.}$	Сходимость почти наверное (almost surely)
$\xrightarrow{\mathbb{P}}$	Сходимость по вероятности (in probability)
$\xrightarrow{d}$	Сходимость по распределению (in distribution)

Лекции

1 Основные понятия математической статистики. Вероятностно-статистическая модель, выборка, статистика, оценка. Свойства оценок.

1.1 Мотивационный пример: выборы

Перед тем как перейти к основным определениям рассмотрим задачу предсказания итогов выборов (есть два кандидата: A и B). Опросим n человек ($n << N$, где N - численность населения). Пусть

- θ – доля людей, которые будут голосовать за кандидата A на выборах
- $1-\theta$ – доля людей, который будут голосовать за кандидата B

Вопрос: как построить оценку на θ предсказать, кто победит ($\theta > 0.5$?)?

Модель опроса: построим упрощенную модель опроса, где люди выбираются с возвращением, всего n человек: $X_1, X_2, \ldots, X_n$. Будем считать, что

$$X_i = \begin{cases} 1, & i\text{-ый человек голосует за } A \\ 0, & \text{иначе.} \end{cases}$$

$X_1, X_2, \ldots, X_n$ – независимые случайные величины, так как опрос с возвращением, более того

$$X_i \overset{\text{i.i.d}}{\sim} \mathcal{B}\text{er}(\theta).$$

Интуиция подсказывает, что можно воспользоваться следующей оценкой для неизвестного параметра θ

$$\widehat{\theta} = \frac{\text{доля людей за A из опроса}}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Такую оценку в дальнейшем будем называть выборочным средним. Поскольку $X_1, X_2, \ldots, X_n$ – случайные величины, то и $\widehat{\theta}$ – случайная величина. Посчитаем ее математическое ожидание и дисперсию

- $$\mathbb{E}\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}X_i = \theta$$
- для подсчета дисперсии воспользуемся независимостью $X_1, X_2, \ldots, X_n$:

$$\text{Var}(\widehat{\theta}) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{\theta(1-\theta)}{n}.$$

Итого делаем вывод:

- в среднем оценка $\widehat{\theta}$ совпадает с истинным значением параметра θ. Такие оценки в дальнейшем будем называть несмещенными;
- чем больше n, тем меньше разброс оценки $\widehat{\theta}$ относительно истинного значения параметра θ.

Далее мы будем называть такие оценки (для которых выполняются оба эти свойства) состоятельными.

1.2 Замечание

При практических применениях статистики никогда не стоит забывать о том, что к статистическим выводам нужно подходить с осторожностью. Обоснованность выводов зависит не только от корректности примененных математических методов, но и от адекватности статистической модели (предположений, лежащих в ее основе) и от качества входных данных. Если входные данные модели недостаточно репрезентативны, то все выводы, сделанные на ее основе, могут оказаться ошибочными. Приведем исторический пример. В 1936 году влиятельный американский еженедельный журнал 'The Literary Digest' провел беспрецедентный по охвату аудитории предвыборный опрос, чтобы предсказать результат выборов президента США. Нужно отметить, что до этого момента журнал проводил такие опросы регулярно перед каждыми выборами, начиная с 1916 года, и неизменно правильно предсказывал победителя. Итак, журнал разослал почти 10 миллионов анкет с вопросом: 'За кого Вы будете голосовать на предстоящих выборах: Франклина Рузвельта (демократа) или Альфреда Лэндона (республиканца)?'. Журнал получил примерно 2.27 миллиона ответов – астрономическое число для любого социологического опроса. По итогам обработки ответов в редакции пришли к выводу, что Альфред Лэндон должен победить Франклина Рузвельта с большим отрывом. В реальности же произошла оглушительная победа Франклина Рузвельта. Республиканец смог победить всего в 2 штатах – небывалая история для американских выборов. Удар по репутации журнала оказался настолько сильным, что меньше, чем через полтора года, в 1938-м, журнал закрылся.

Итак, где же ошиблись представители 'The Literary Digest'? Ведь учтенные ими 2.27 миллиона голосов – это действительно очень много. Здесь сказались сразу 2 критических ошибки. Во-первых, сказался принцип формирования базы данных адресов для рассылки анкет. Журнал разослал анкеты своим подписчикам, а также владельцам телефонов и автомобилей (эти базы данных были доступны). Но в 1936 году в Америке на фоне Великой Депрессии позволить себе телефон, подписку на литературный журнал или автомобиль могли только состоятельные люди, которые составляли меньшинство избирателей, и были более склонны голосовать за республиканцев. Таким образом, не было соблюдено требование репрезентативности данных для опроса. Данный эффект в англоязычной статистической литературе называется **'sampling bias'**. Наряду с этим, возможно, сказался и еще один эффект. Люди, которым не нравился (действующий на тот момент) президент Рузвельт просто проявили больше желания потратить свое время, чтобы отправить анкету с ответом. Подобное смещение также часто отражается в социологических опросах и также имеет специальное название **'participation bias'** (или **'non-response bias'**).

Будем считать, что с вводной частью мы закончили и теперь мы наконец переходим к основным определениям, на которых строится весь курс.

1.3 Основные определения

Определение 1 (Выборка)**.** Набор $\mathbf{X} = (X_1, \ldots, X_n)$ независимых и одинаково распределенных случайных величин с распределением $\mathbb{P}_X$ называется *выборкой* размера n.

Определение 2 (Выборочное пространство)**.** Множество всех возможных значений $\mathbf{X}$ называется *выборочным пространством* и обозначается $\mathcal{X}$.

Определение 3 (Вероятностно-статистическая модель)**.** *Вероятностно-статистическая модель* - это тройка $(\mathcal{X}, \mathcal{B}_\mathcal{X}, \mathcal{P})$, где $\mathcal{X}$ – выборочное пространство, $\mathcal{B}_\mathcal{X}$ – σ-алгебра на нем, а $\mathcal{P}$ – семейство распределений (вероятностных мер) на $(\mathcal{X}, \mathcal{B}_\mathcal{X})$.

Замечание 1. Хотя нам изначально неизвестно $\mathbb{P}_X$, зачастую нам известен класс распределений $\mathcal{P}$, которому принадлежит $\mathbb{P}_X$.

Задача математической статистики заключается в том, чтобы восстановить те или иные свойства неизвестного распределения $\mathbb{P}_X \in \mathcal{P}$ на основе выборки $\mathbf{X}$ из данного распределения. Если у нас нет никакой априорной информации о распределении $\mathbb{P}_X$ и семействе допустимых мер $\mathcal{P}$, данная задача может быть очень сложной. Поэтому на практике часто рассматриваются параметрические вероятностно-статистические модели.

Определение 4 (Параметрическая модель)**.** Вероятностно-статистическая модель $(\mathcal{X}, \mathcal{B}_\mathcal{X}, \mathcal{P})$ называется параметрической, если семейство вероятностных мер $\mathcal{P}$ может быть параметризовано конечным числом параметров, то есть

$$\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta \subset \mathbb{R}^r\}.$$

При этом $\mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2}$, если $\theta_1 \neq \theta_2$.

Преимущество параметрического класса моделей в том, что вместо задачи восстановления распределения $\mathbb{P}_X$ достаточно решить более простую задачу оценивания параметра $\theta \in \Theta$. В дальнейшем, если не указано иного, мы будем рассматривать именно параметрическую модель $(\mathcal{X}, \mathcal{B}_\mathcal{X}, \mathcal{P})$, $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ и выборку $\mathbf{X}$ из неизвестного распределения $\mathbb{P}_X \in \mathcal{P}$.

Пример 1. Примеры параметрических моделей:

1. Биномиальное распределение $\mathrm{Bin}(n, \theta)$, $\theta \in (0; 1)$;
2. Нормальное распределение $\mathcal{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$;
3. Равномерное распределение $\mathcal{U}(\theta_1, \theta_2)$, $\theta_1 < \theta_2$, $\theta = (\theta_1, \theta_2)$.

Определение 5 (Статистика)**.** *Статистикой* $T(\mathbf{X})$ называется измеримое отображение $T : \mathcal{X} \to \mathbb{R}^r$.

Подчеркнем, что T в определении выше – это функция от выборки, она не должна зависеть от неизвестного параметра θ. Так как $\mathbf{X}$ – это случайный вектор, то и $T(\mathbf{X})$ будет случайна и имеет смысл говорить о ее *распределении*. Измеримость нужна для того, чтобы были определены следующие вероятности при всех t:

$$F_T(t) = \mathbb{P}(T(\mathbf{X}) \leqslant t).$$

Определение 6 (Оценка)**.** Если множество возможных значений $T(\mathbf{X}) \in \Theta$, то статистика $T(\mathbf{X})$ называется оценкой.

Ниже мы рассмотрим свойства оценок.

1.4 Свойства оценок

Прежде чем перейти к свойствам оценок, сформулируем вспомогательное определение.

Определение 7 (Смещение)**.** *Смещением* оценки $\widehat{\theta}_n(\boldsymbol{X})$ называется величина

$$b_n(\theta) = \mathbb{E}_\theta \widehat{\theta}_n(\boldsymbol{X}) - \theta.$$

Запись $\mathbb{E}_\theta$ означает, что математическое ожидание берется по выборке $\mathbf{X}$ из истинного распределения $\mathbb{P}_\theta$. Например, если вероятностная мера $\mathbb{P}_\theta$ абсолютно непрерывна и имеет плотность $p_\theta(x)$, то плотность вектора $\boldsymbol{X}$

$$p_{\boldsymbol{X}}(\boldsymbol{x}, \theta) = p_\theta(x_1) \cdot \cdots \cdot p_\theta(x_n), \qquad \boldsymbol{x} = (x_1, \ldots, x_n),$$

а математическое ожидание

$$\mathbb{E}_\theta[\widehat{\theta}(\mathbf{X})] = \int_{\mathcal{X}} \widehat{\theta}(\boldsymbol{x}) p_{\boldsymbol{X}}(\boldsymbol{x}, \theta) d\boldsymbol{x}.$$

Теперь перейдем к основным свойствам.

Определение 8 (Несмещенность)**.** Оценка $\widehat{\theta}_n(\boldsymbol{X})$ (или последовательность оценок) параметра θ называется

- *несмещенной*, если для всех $\theta \in \Theta : b_n(\theta) = 0$
- *асимптотически несмещенной*, если для всех $\theta \in \Theta : b_n(\theta) \xrightarrow{n\to\infty} 0$.

Определение несмещенной оценки для функции $\tau(\theta)$ формулируется аналогично.

Определение 9 (Состоятельность)**.** Оценка $\widehat{\theta}_n(\boldsymbol{X})$ (или последовательность оценок) называется *состоятельной* оценкой параметра θ, если для любого $\theta \in \Theta$

$$\widehat{\theta}_n(\boldsymbol{X}) \xrightarrow{\mathbb{P}_\theta} \theta, \text{ при } \ n \to \infty\,,$$

то есть для любого $\theta \in \Theta$ и любого $\varepsilon > 0$

$$\mathbb{P}_\theta\Big(\|\widehat{\theta}_n(\boldsymbol{X}) - \theta\| \geqslant \varepsilon\Big) \to 0, \ n \to \infty\,.$$

Определение 10 (Сильная состоятельность)**.** Оценка $\widehat{\theta}_n(\boldsymbol{X})$ (или последовательность оценок) называется *сильно состоятельной* оценкой параметра θ, если для любого $\theta \in \Theta$

$$\widehat{\theta}_n(\boldsymbol{X}) \xrightarrow{a.s.} \theta, \ \text{ при } n \to \infty\,,$$

то есть для любого $\theta \in \Theta$

$$\mathbb{P}_\theta\Big(\lim_{n\to\infty} \widehat{\theta}_n(\boldsymbol{X}) = \theta\Big) = 1\,.$$

Определение 11 (Асимптотическая нормальность)**.** Оценка $\widehat{\theta}_n(\boldsymbol{X})$ (или последовательность оценок) называется *асимптотически нормальной* оценкой параметра θ если для любого $\theta \in \Theta$

$$\sqrt{n}\Big(\widehat{\theta}_n(\mathbf{X}) - \theta\Big) \xrightarrow{d_\theta} \mathcal{N}(0, \sigma^2(\theta)), \ n \to \infty,$$

где $\sigma^2(\theta)$ называется *асимптотической дисперсией* оценки $\hat{\theta}_n$.

2 Введение в теорию точечного оценивания параметров. Сравнение оценок, функция потерь, функция риска. Оптимальная оценка.

Задача оценивания является одной из центральных в математической статистике. Предположим, что имеется выборка $\mathbf{X} = (X_1, \dots, X_n)$ на вероятностном пространстве $(\mathcal{X}, \mathcal{B}_\mathcal{X}, \mathbb{P})$. Допустим, что распределение элементов выборки принадлежит некоторому параметрическому семейству $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, где Θ – множество значений параметра. Поскольку истинное значение параметра θ неизвестно, задача состоит в том, чтобы попытаться сделать вывод о распределении наблюдений на основе конечной выборки. В данном случае достаточно оценить параметр θ, то есть, получив наблюдения $X_1, \dots, X_n$, получить некоторую оценку $\widehat{\theta}(\mathbf{X}) = \widehat{\theta}(X_1, \dots, X_n)$, которая будет отражать информацию о θ. В дальнейшем мы будем иногда опускать зависимость от выборки $\mathbf{X}$, и писать просто $\widehat{\theta}$ вместо $\widehat{\theta}(\mathbf{X})$. Строго говоря, оценкой $\widehat{\theta}$ параметра θ называется любая $\mathcal{B}_\mathcal{X}$-измеримая функция от $X_1, \dots, X_n$. Возможна также постановка задачи, при которой интересен не сам параметр θ, а некоторая функция от него $\tau(\theta)$. В этом случае оценка $\widehat{\tau}$ должна быть близка к $\tau(\theta)$. Стоит отметить, что точно восстановить значение параметра θ невозможно, хотя, если объем выборки достаточно велик, $\widehat{\theta}$ может быть очень близка к θ.

2.1 Сравнение оценок

Задача с предсказанием результатов выборов является частным случаем более общей задачи: по выборке $\mathbf{X} = (X_1, \dots, X_n)$ из неизвестного распределения $\mathbb{P}$ оценить это распределение. Пока что мы ограничиваемся параметрическим случаем ($\mathbb{P} = \mathbb{P}_\theta, \theta \in \Theta$) и вместо оценки неизвестного распределения $\mathbb{P}$ нам достаточно лишь оценить неизвестный параметр. Предположим, мы смогли привести две оценки, тогда как же понять какая из них лучше? Чтобы сравнивать различные оценки, необходимо ввести метрику качества. Обычно для ее определения вводится борелевская функцию (далее именуемая как *функция потерь*)

$$\rho : \Theta \times \Theta \to \mathbb{R}.$$

Если $\widehat{\theta}$ – оценка параметра θ, то $\rho(\widehat{\theta}, \theta)$ показывает величину потерь. Как правило, функции потерь оказываются неотрицательными, но это не является обязательным условием.

Пример 2. Примеры функций потерь.

1. Абсолютная функция потерь: $\rho(x, y) = |x - y|$.
2. Квадратичная функция потерь: $\rho(x, y) = (x - y)^2$.
3. Логарифмические потери: $\rho(x, y) = \log y - \log x$ (встречаются в задаче оценки плотности).
4. Если $\theta \in \mathbb{R}^d$, $d > 1$, то для симметричной неотрицательно определенной матрицы $A \in \mathbb{R}^{d \times d}$ можно определить
$$\rho(x, y) = \langle A(x - y), x - y \rangle.$$

Далее в рамках курса будем считать функцию потерь квадратичной, если не сказано обратное. Поскольку $\widehat{\theta} = \widehat{\theta}(\mathbf{X})$ - это случайная величина, зависящая от выборки, то и функция потерь $\rho(\widehat{\theta}, \theta)$ будет тоже случайной величиной. Для того чтобы мера качества оценки не зависела от конкретной реализации выборки, вводят функцию риска.

Определение 12 (Функция риска)**.** При заданной функции потерь $\rho(x, y)$ функцией риска оценки $\widehat{\theta}$ называется

$$R(\widehat{\theta}, \theta) = \mathbb{E}_\theta \rho(\widehat{\theta}, \theta).$$

При квадратичной функции потерь функция риска именуется среднеквадратичной ошибкой. В качестве метрики качества будем рассматривать функцию риска. Теперь перейдем к сравнению оценок, базируясь на введенной функции риска.

Определение 13. Говорят, что оценка $\widehat{\theta}$ лучше оценивает параметр θ, чем оценка $\widetilde{\theta}$, если для всех $\theta \in \Theta$ выполнено

$$R(\widehat{\theta}, \theta) \leqslant R(\widetilde{\theta}, \theta),$$

причем для некоторого $\theta \in \Theta$ неравенство строгое.

Определение 14. Оценка $\widehat{\theta}$ называется наилучшей в классе $\mathcal{K}$, если она лучше любой другой оценки $\widetilde{\theta} \in \mathcal{K}$.

Вообще говоря, в определении 14 сравниваются различные функции $R_{\widetilde{\theta}}(\theta) := R(\widehat{\theta}, \theta)$ от параметра θ и неочевидно, что существует такая функция $R_{\widehat{\theta}}(\theta) := R(\widehat{\theta}, \theta)$, которая будет подпирать снизу остальные функции. Действительно, покажем, что наилучшая оценка не всегда существует.

Замечание 2. Рассмотрим $\Theta \subseteq \mathbb{R}$, класс всевозможных оценок $\mathcal{K}$, и квадратичную функцию потерь $\rho(x, y) = (x - y)^2$, Тогда наилучшей оценки не существует.

Доказательство. Для любого $\theta_0 \in \Theta$ можно рассмотреть оценку $\widehat{\theta}_0(\mathbf{X}) \equiv \theta_0$. Тогда в точке $\theta = \theta_0$ ее функция риска обращается в нуль:

$$R(\widehat{\theta}_0(\mathbf{X}), \theta_0) = \mathbb{E}_{\theta_0}\Big(\widehat{\theta}_0(\mathbf{X}) - \theta_0\Big)^2 = 0$$

Это означает, что если наилучшая оценка $\widehat{\theta}(\mathbf{X})$ существует, то в каждом $\theta \in \Theta$ должно достигаться значение $R(\widehat{\theta}(\mathbf{X}), \theta) = 0$, а значит должно быть выполнено $R(\widehat{\theta}(\mathbf{X}), \theta) \equiv 0$. Поскольку $R(\widehat{\theta}(\mathbf{X}), \theta) = \mathbb{E}_\theta[(\widehat{\theta}(\mathbf{X}) - \theta)^2]$, должно выполняться $\widehat{\theta}(x) \equiv \theta$ для всех $\theta \in \Theta$. Противоречие, так как оценка не может зависеть от неизвестного параметра. □

В данном примере отсутствие наилучшей оценки объясняется тем, что рассматриваемый класс оценок чрезмерно широкий. Оказывается если его сузить, то оптимальная оценка может существовать, например если рассматривать только несмещенные оценки. Покажем, как к этому можно прийти, но сперва докажем вспомогательную теорему о том, что в случае квадратичной функции потерь функция риска раскладывается на сумму квадрата смещения оценки и ее дисперсии.

Теорема 1. *Пусть дана выборка $\mathbf{X} = (X_1, \ldots, X_n)$ из распределения $\mathbb{P}_\theta, \theta \in \Theta \in \mathbb{R}^d$. Пусть задана квадратичная функция потерь $\rho(x, y) = \|x - y\|^2$, и $\widehat{\theta}$ – произвольная оценка параметра θ. Тогда*

$$R(\widehat{\theta}, \theta) = \mathbb{E}_\theta\Big\|\widehat{\theta} - \theta\Big\|^2 = \|b(\theta)\|^2 + \mathrm{Tr}\Big(\mathrm{Var}(\widehat{\theta})\Big),$$

где $\mathrm{Var}(\widehat{\theta})$ – ковариационная матрица $\widehat{\theta}$

$$\mathrm{Var}(\widehat{\theta}) = \mathbb{E}_\theta\Big(\widehat{\theta} - \mathbb{E}_\theta\widehat{\theta}\Big)^\mathsf{T}\Big(\widehat{\theta} - \mathbb{E}_\theta\widehat{\theta}\Big)^\mathsf{T},$$

а $\mathrm{Tr}(\cdot)$ – след матрицы.

Доказательство. Добавим и вычтем $\mathbb{E}_\theta\widehat{\theta}$ из среднеквадратичной ошибки, а далее раскроем квадрат нормы

$$\begin{aligned} R(\widehat{\theta},\theta) &= \mathbb{E}_\theta\left\|\widehat{\theta}-\theta\right\|^2 = \mathbb{E}_\theta\left\|\widehat{\theta}-\mathbb{E}_\theta\widehat{\theta}+\mathbb{E}_\theta\widehat{\theta}-\theta\right\|^2 \\ &= \mathbb{E}_\theta\left\|\widehat{\theta}-\mathbb{E}_\theta\widehat{\theta}\right\|^2 + 2\mathbb{E}_\theta\left(\widehat{\theta}-\mathbb{E}_\theta\widehat{\theta}\right)^\top\left(\mathbb{E}_\theta\widehat{\theta}-\theta\right) + \left\|\mathbb{E}_\theta\widehat{\theta}-\theta\right\|^2 \\ &= \mathbb{E}_\theta\left\|\widehat{\theta}-\mathbb{E}_\theta\widehat{\theta}\right\|^2 + \left\|\mathbb{E}_\theta\widehat{\theta}-\theta\right\|^2 \\ &= \mathrm{Tr}\Big(\mathrm{Var}(\widehat{\theta})\Big) + \|b(\theta)\|^2. \end{aligned}$$

□

Опираясь на данную теорему, сформулируем лемму, которая бывает очень полезной для доказательства состоятельности. Ограничимся для простоты случаем скалярного параметра.

Лемма 1 (Достаточное условие состоятельности)**.** *Если смещение* $b(\theta) \xrightarrow{n\to\infty} 0$ *и* $\mathrm{Var}(\widehat{\theta}) \xrightarrow{n\to\infty} 0$*, то оценка* $\widehat{\theta}_n$ *состоятельна.*

Доказательство. Воспользуемся неравенством Маркова для фиксированного $\varepsilon > 0$

$$\mathbb{P}(|\widehat{\theta}-\theta| > \varepsilon) \leqslant \frac{\mathbb{E}_\theta(\widehat{\theta}-\theta)^2}{\varepsilon^2} = \frac{b(\theta)^2+\mathrm{Var}(\widehat{\theta})}{\varepsilon^2} \xrightarrow{n\to\infty} 0.$$

□

Теорема 1 говорит о том, что для сравнения оценок достаточно сравнить квадраты их смещения и дисперсии. А что если мы будем рассматривать только несмещенные оценки (это не значит, что смещенные оценки чем-то плохи, но для простоты ограничимся несмещенными). Тогда будет достаточно сравнивать лишь дисперсии оценок. Вспомним, что дисперсия оценки показывает ее разброс относительно истинного значения параметра и чем этот разброс меньше, тем точнее оценка. Тогда уточним определение 14 на случай, когда $\mathcal{K}$ – это класс несмещенных оценок и сформулируем понятие оптимальной оценки.

2.2 Оптимальные оценки

Пусть

$$\mathcal{K} \text{ - класс несмещенных оценок с ограниченной дисперсией.}$$

Определение 15 (Оптимальная оценка)**.** Оптимальной оценкой параметра θ называется такая оценка $\widehat{\theta}$, что

- $\widehat{\theta} \in \mathcal{K}$;
- Оценка $\widehat{\theta}$ имеет равномерно минимальную дисперсию в классе $\mathcal{K}$, то есть для любой *несмещенной* оценки $\widetilde{\theta} \in \mathcal{K}$ при всех $\theta \in \Theta$ выполнено

$$\mathrm{Var}_\theta(\widehat{\theta}) \leqslant \mathrm{Var}_\theta(\widetilde{\theta}).$$

А теперь наконец сформулируем основное преимущество нашего выбора рассматривать лишь несмещенные оценки.

Лемма 2. *Если оптимальная оценка существует, то она единственна.*

Доказательство. Докажем от противного. Пусть существуют две оптимальные оценки $\widehat{\theta}_1$ и $\widehat{\theta}_2$ для параметра θ. Тогда $\mathbb{E}_\theta(\widehat{\theta}_1) = \mathbb{E}_\theta(\widehat{\theta}_2) = \theta$, и в силу их оптимальности

$$\mathrm{Var}_\theta(\widehat{\theta}_1) = \mathrm{Var}_\theta(\widehat{\theta}_2)\,. \tag{1}$$

Построим теперь новую оценку $\widehat{\theta}_3 = \frac{\widehat{\theta}_1+\widehat{\theta}_2}{2}$ и вычислим ее математическое ожидание и дисперсию:

$$\mathbb{E}_\theta\widehat{\theta}_3 = \frac{1}{2}\mathbb{E}_\theta(\widehat{\theta}_1+\widehat{\theta}_2) = \theta\,,$$

то есть $\widehat{\theta}_3$ также является несмещенной. Вычислим теперь дисперсию оценки $\widehat{\theta}_3$:

$$\mathrm{Var}_\theta(\widehat{\theta}_3) = \frac{1}{4}\Big(\mathrm{Var}_\theta(\widehat{\theta}_1) + \mathrm{Var}_\theta(\widehat{\theta}_2) + 2\mathrm{cov}_\theta(\widehat{\theta}_1,\widehat{\theta}_2)\Big) = \frac{1}{2}\Big(\mathrm{Var}_\theta(\widehat{\theta}_1) + \mathrm{cov}_\theta(\widehat{\theta}_1,\widehat{\theta}_2)\Big).$$

По неравенству Коши-Буняковского-Шварца (КБШ)

$$\mathrm{cov}_\theta(\widehat{\theta}_1,\widehat{\theta}_2) \leqslant \sqrt{\mathrm{Var}_\theta(\widehat{\theta}_1)}\sqrt{\mathrm{Var}_\theta(\widehat{\theta}_2)} \leqslant \mathrm{Var}_\theta(\widehat{\theta}_1),$$

значит, с учетом (1), получим

$$\mathrm{Var}_\theta(\widehat{\theta}_3) \leqslant \mathrm{Var}_\theta(\widehat{\theta}_1)\,.$$

Так как оценки $\widehat{\theta}_1$ и $\widehat{\theta}_2$ – оптимальные, предыдущее неравенство должно обратиться в равенство. Это возможно, только если $\widehat{\theta}_1$ и $\widehat{\theta}_2$ линейно связаны

$$\widehat{\theta}_2 = a\widehat{\theta}_1 + b,$$

для некоторых постоянных a и b. Найдем эти константы. Из КБШ имеем

$$\mathrm{cov}_\theta(\widehat{\theta}_1,\widehat{\theta}_2) = \mathrm{cov}_\theta(\widehat{\theta}_1, a\widehat{\theta}_1+b) = \mathrm{cov}_\theta(\widehat{\theta}_1, a\widehat{\theta}_1) = \mathbb{E}_\theta(\widehat{\theta}_1-\theta)(a\widehat{\theta}_1 - a\theta) = a\mathbb{E}_\theta(\widehat{\theta}_1-\theta)^2 = a\mathrm{Var}_\theta(\widehat{\theta}_1).$$

Но $\mathrm{cov}_\theta(\widehat{\theta}_1,\widehat{\theta}_2) = \mathrm{Var}_\theta(\widehat{\theta}_1)$, значит $a = 1$. Из несмещенности получим $b = 0$. Следовательно, оценки $\widehat{\theta}_1$ и $\widehat{\theta}_2$ совпадают и значит если существует оптимальная оценка, то она единственна. □

Прокомментируем результаты этой леммы. Если бы мы рассматривали квадратичную функцию потерь и всевозможные оценки, то наилучшей оценки бы не существовало. Но мы ограничились лишь несмещенными оценками, и в этом классе наилучшая оценка уже однозначно определяется.

3 Эмпирическая функция распределения, ее свойства. Порядковые статистики, вариационный ряд выборки, выборочные квантили.

3.1 Эмпирическое функция распределения

В некоторых задачах распределение элементов выборки не получается охарактеризовать каким-либо конечномерным параметром. Например, о функции распределения $X_1, \dots, X_n$ может быть известно только то, что она непрерывна. Иногда, чтобы подчеркнуть широту класса допустимых распределений $\mathcal{P} = \{\mathbb{P}_X\}$, говорят о задаче *непараметрического* оценивания. В такой задаче объектом, представляющим интерес, обычно является функция распределения или плотность распределения элементов выборки. Оценим функцию распределения F. Для начала вспомним, что

$$F(x) := \mathbb{P}_X(X \leqslant x) = \mathbb{E}\mathbb{1}(X \leqslant x), \quad x \in \mathbb{R}.$$

ЗБЧ говорит, что в качестве оценки неизвестного математического ожидания можно взять среднее. Это приводит нас к определению *эмпирической функции распределения* (ЭФР), которая будет выступать в качестве оценки неизвестной $F(x)$.

Определение 16 (Эмпирическая функция распределения)**.** Пусть $\mathbf{X} = (X_1, \dots, X_n)$ есть выборка размера n из распределения $\mathbb{P}_X$. Тогда случайная величина

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \leqslant x\}, \quad x \in \mathbb{R},$$

называется *эмпирической функцией распределения* (ЭФР).

Отметим, что ЭФР является неубывающей функцией и ее значения лежат на отрезке $[0, 1]$, как и у истинной функции рапспределения $F(x)$, поэтому ЭФР может быть использована в качестве оценки для $F(x)$. Отметим ряд свойств ЭФР.

Теорема 2. *Пусть $\mathbf{X} = (X_1, \dots, X_n)$ - выборка из распределения $\mathbb{P}_X$. Пусть $F(x) = \mathbb{P}_X(X_i \leqslant x)$ – это истинная функция распределения, а $\widehat{F}_n(x)$ – эмпирическая функция распределения, построенная по этой выборке. Тогда для любого $x \in \mathbb{R}$ выполнено:*

1. *несмещенность: $\mathbb{E}[\widehat{F}_n(x)] = F(x)$;*
2. *$\mathrm{Var}(\widehat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}$;*
3. *$n\hat{F}_n(x) \sim \mathrm{Bin}(n, F(x))$;*
4. *сильная состоятельность: $\widehat{F}_n(x) \xrightarrow{a.s.} F(x)$ при $n \to \infty$;*
5. *асимптотическая нормальность:*

$$\sqrt{n}(\widehat{F}_n(x) - F(x)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(x)),$$

где $\sigma^2(x) = F(x)(1 - F(x))$ – асимптотическая дисперсия;

Доказательство. 1. несмещенность:

$$\mathbb{E}\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\mathbb{1}\{X_i \leqslant x\} = \mathbb{E}\mathbb{1}\{X_1 \leqslant x\} = \mathbb{P}(X_1 \leqslant x) = F_\theta(x).$$

2. Из независимости $X_1, \dots, X_n$ получаем

$$\mathrm{Var}(\widehat{F}_n(x)) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(\mathbb{1}\{X_i \leqslant x\}) = \frac{1}{n}\mathrm{Var}(\mathbb{1}\{X_1 \leqslant x\}).$$

Определим случайную величину $\eta = \mathbb{1}\{X_1 \leqslant x\}$:

$$\eta = \begin{cases} 1, & \text{с вероятностью } \mathbb{P}(X_1 \leqslant x) = F(x); \\ 0, & \text{иначе,} \end{cases}$$

то есть $\eta \sim \mathcal{B}\mathrm{er}(F(x))$. Тогда

$$\mathrm{Var}(\widehat{F}_n(x)) = \frac{1}{n}\mathrm{Var}(\mathbb{1}\{X_1 \leqslant x\}) = \frac{1}{n}\mathrm{Var}(\eta) = \frac{1}{n}F(x)(1-F(x)).$$

3. Из того, что $\eta_i = \mathbb{1}\{X_i \leqslant x\} \sim \mathcal{B}\mathrm{er}(F(x))$ следует

$$n\widehat{F}_n(x) = \sum_{i=1}^{n}\mathbb{1}\{X_i \leqslant x\} = \sum_{i=1}^{n}\eta_i \sim \mathrm{Bin}(n, F(x)).$$

4. достаточно применить УЗБЧ Колмогорова к независимым, одинаково распределенным случайным величинам $\mathbb{1}\{X_k \leqslant x\}$, $k \in \{1, \dots, n\}$.

5. применяя ЦПТ для $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\eta_i$:

$$\sqrt{n}(\widehat{F}_n(x) - F(x)) = \sqrt{n}(\overline{\eta} - \mathbb{E}\eta_1) \xrightarrow{d} \mathcal{N}(0, \mathrm{Var}(\eta_1)),$$

то есть асимптотическая дисперсия $\sigma^2(x) = \mathrm{Var}(\eta_1) = F(x)(1-F(x))$.

□

На самом деле верен более сильный результат: $\widehat{F}_n(x)$ равномерно приближает $F(x)$ в каждой точке. Этот результат известен как теорема Гливенко-Кантелли (ее доказывать мы не будем).

Теорема 3 (Гливенко-Кантелли)**.** *Пусть $\widehat{F}_n(x)$ – эмпирическая функция распределения. Тогда*

$$\sup_{x\in\mathbb{R}}|\widehat{F}_n(x) - F(x)| \xrightarrow{a.s.} 0, \quad n \to \infty.$$

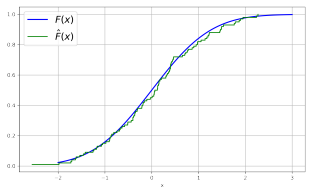

Рис. 1: Эмпирическая функция распределения приближает истинную функцию распределения

3.2 Порядковые статистики.

Упорядочим выборку $\mathbf{X} = (X_1, \dots, X_n)$ по возрастанию:

$$\begin{aligned} X_{(1)} &= \min\{X_1, \dots, X_n\}, \\ X_{(2)} &= \min\{(X_1, \dots, X_n) \setminus X_{(1)}\}, \\ &\dots \\ X_{(n)} &= \max\{X_1, \dots, X_n\}. \end{aligned}$$

Определение 17 (вариационный ряд)**.** Упорядоченный набор $X_{(1)} \leqslant X_{(2)} \leqslant \dots \leqslant X_{(n)}$ называется *вариационным рядом*, построенным по выборке $\mathbf{X}$.

Определение 18 ($k-$я порядковая статистика)**.** Элемент $X_{(k)}$ будем называть k-й порядковой статистикой. Элементы $X_{(1)}$ и $X_{(n)}$ называются минимальной и максимальной порядковыми статистиками соответственно.

Используя понятия $k-$х порядковых статистик, определение эмпирической функции распределения 16 упрощается:

$$\widehat{F}_n(x) = \begin{cases} 0, & x < X_{(1)}; \\ \frac{k}{n}, & X_{(k)} \leqslant x < X_{(k+1)}, \quad k \in \{1, \dots, n-1\}; \\ 1, & x \geqslant X_{(n)}. \end{cases}$$

То есть именно в точках $X_{(k)}$ происходят скачки ЭФР.

Вопрос: является ли вариационный ряд статистикой, а выборкой? Для ответа на этот вопрос найдем распределение $k-$й порядковой статистики $X_{(k)}$.

Лемма 3. *Пусть* $\mathbf{X} = (X_1, \dots, X_n)$ *- выборка из распределения* $\mathbb{P}_X$ *с функцией распределения* $F(x) = \mathbb{P}_X(X_i \leqslant x)$*. Тогда функция распределения* $k-$*ой порядковой статистики* $X_{(k)}$ *задается формулой*

$$F_{X_{(k)}}(x) = \mathbb{P}(X_{(k)} \leqslant x) = \sum_{j=k}^{n} \binom{n}{j} \big[F(x)\big]^j \big[1 - F(x)\big]^{n-j}.$$

Доказательство. Воспользуемся тем, что ЭФР монотонная

$$\mathbb{P}(X_{(k)} \leqslant x) = \mathbb{P}\Big(\widehat{F}_n(x) \geqslant \frac{k}{n}\Big) = \mathbb{P}\Big(n\widehat{F}_n(x) \geqslant k\Big).$$

Из теоремы 2 имеем $n\hat{F}_n(x) \sim \mathrm{Bin}(n, F(x))$. Значит,

$$\mathbb{P}\Big(n\widehat{F}_n(x) \geqslant k\Big) = \sum_{j=k}^{n} \mathbb{P}(\widehat{F}_n(x) = j) = \sum_{j=k}^{n} \binom{n}{j} \big[F(x)\big]^j \big[1 - F(x)\big]^{n-j}.$$

□

Видим, что распределение $k-$ой порядковой статистики $X_{(k)}$ различно для каждого $k \in \mathbb{N}$, поэтому вариационный ряд не является выборкой, однако он будет являться статистикой, поскольку он зависит лишь от элементов выборки.

На ЭФР $\widehat{F}_n(x)$ можно смотреть как на функцию распределения некоторой дискретной случайной величины ξ, принимающей n значений: $X_1, \dots, X_n$ – с вероятностями, равными $1/n$. Для ЭФР как и

для обычного распределения можно вводить различные числовые характеристики (математическое ожидание, дисперсию, моменты):

$$\widehat{g}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i), \quad \text{для различных функций } g.$$

3.3 Выборочные моменты.

Определение 19 (Выборочный и центральный моменты)**.** Пусть дана выборка $X_1, ..., X_n$. Для любого $k \in \mathbb{N}$ величины

$$\overline{X^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k \quad \text{и} \quad \alpha^k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$$

называются *k-ым выборочным моментом* и *k-ым центральным выборочным моментом* соответственно.

Замечание 3. Первый выборочный момент называют выборочным средним и обычно обозначают $\overline{X}$. Второй центральный выборочный момент носит название выборочной дисперсии, ее будем обозначать через S^2.

Лемма 4. *Пусть* $\mathbf{X} = (X_1, \ldots, X_n)$ *- выборка из распределения* $\mathbb{P}$*, имеющего конечный момент порядка* k*. Тогда* $\overline{X^k}$ *является несмещённой и (сильно) состоятельной оценкой для теоретического* k*-го момента* $\mathbb{E}_\theta X_1^k$*:*

1. $\mathbb{E}[\overline{X^k}] = \mathbb{E}X_1^k$*;*
2. $\overline{X^k} \xrightarrow{a.s.} \mathbb{E}X_1^k$ *при* $n \to \infty$*.*

Доказательство данной леммы следует из определения выборочных моментов и УЗБЧ в форме Колмогорова (см. семинары).

Ввиду особой практической важности оценивания теоретических среднего и дисперсии, сформулируем следующую теорему.

Теорема 4. *Пусть дана выборка* $X_1, ..., X_n$ *из распределения с конечным вторым моментом* $\mathbb{E}X_1^2$*. Тогда. выборочное среднее*

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

и статистика

$$S_0^2 := \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

являются несмещенными и состоятельными оценками для $\mathbb{E}X_1$ *и* $\mathrm{Var}(X_1)$*.*

В отличие от выборочного среднего $\overline{X}$, которое является несмещенной оценкой своего математического ожидания, выборочная дисперсия S^2 будет смещенной оценкой для истинной дисперсии. Поэтому в теореме 4 ее пришлось подправить и ввести статистику S_0^2. С точки зрения состоятельности, обе оценки S^2 и S_0^2 будут состоятельными.

Лемма 5. *Пусть* $\mathbf{X} = (X_1, \dots, X_n)$ *- выборка из распределения* $\mathbb{P}$*, имеющего конечный момент порядка* $2k$*. Тогда* k*-ый выборочный момент* $\overline{X^k}$ *является асимптотически нормальной оценкой для теоретического* k*-го момента* $\mathbb{E}_\theta X_1^k$*:*

$$\sqrt{n}(\overline{X^k} - \mathbb{E}X_1^k) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}_\theta X_1^{2k} - (\mathbb{E}X_1^k)^2).$$

В случае выборки $\mathbf{X} = (X_1, \dots, X_n)$ из нормального распределения, можно гарантировать, что .выборочное среднее $\overline{X}$ и выборочная дисперсия S^2 будут независимы. Более того, можно предъявить конкретные распределения для некоторых функций от этих статистик. Этот результат известен как теорема Фишера. Но перед тем как перейти к ней, вспомним одно свойства нормального (гауссовского) вектора.

Лемма 6 (Свойства многомерного нормального распределения)**.** *1. Если* $\xi \in \mathbb{R}^n$ *– многомерный нормальный вектор со средним* μ *и ковариационной матрицей* Σ*, а* $A \in \mathbb{R}^{m\times n}$ *– детерминированная матрица, то*

$$A\xi \sim \mathcal{N}(A\mu, A\Sigma A^\intercal).$$

(Доказать через характеристическую функцию).

2. *Если* $\xi \in \mathbb{R}^n$ *– многомерный нормальный вектор со средним* μ *и невырожденной ковариационной матрицей* Σ*, то плотность его распределения*

$$p_\xi(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} e^{-(x-\mu)^\intercal \Sigma^{-1}(x-\mu)/2}, \quad x \in \mathbb{R}^n.$$

3. *Компоненты гауссовского случайного вектора независимы тогда и только тогда, когда они некоррелированы.*

Лемма доказывается через характеристические функции.

Теорема 5 (Фишер)**.** *Пусть* $(X_1, \dots, X_n)$ *– выборка объема* n *из распределения* $\mathcal{N}(\mu, \sigma^2)$*. Пусть*

$$S_0^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$$

Тогда статистики $\overline{X}$ *и* S_0^2 *независимы и имеют нормальное распределение и хи-квадрат распределение соответственно*

$$\sqrt{n}\cdot\frac{\overline{X}-\mu}{\sigma} \sim \mathcal{N}(0,1), \qquad \frac{(n-1)S_0^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \overline{X})^2 \sim \chi^2(n-1).$$

Доказательство. 1. Докажем для начала независимость $\overline{X}$ и S_0^2. Доказательство данной теоремы разбиралось на семинарах, однако приведем его здесь тоже для справки. По условию $(X_1, \dots, X_n)$ - нормальный (гауссовский) вектор. Тогда, пользуясь свойством гауссовского вектора (см. лемму 6), можно получить, что вектор $(\overline{X}, X_1 - \overline{X}, \dots, X_n - \overline{X})$ – тоже гауссовский вектор (то есть можно предъявить матрицу A такую, что итоговый вектор – это результат умножения этой матрицы A на исходный нормальный вектор).

Проверим, что компонента нормального вектора $\overline{X}$ независима с остальными $(X_1 - \overline{X}, \dots, X_n - \overline{X})$. В случае гауссовского вектора для независимости достаточно проверить некоррелированность (см. лемму 6).

$$\operatorname{cov}(\overline{X}, X_j - \overline{X}) = \operatorname{cov}(\overline{X}, X_j) - \operatorname{cov}(\overline{X}, \overline{X}) = \frac{\sigma^2}{n} - \frac{n\sigma^2}{n^2} = 0$$

Теперь остается заметить, что S^2 – это функция от с.в. $X_j - \overline{X}$, а значит она тоже независима с $\overline{X}$. Что и требовалось доказать.

2. Теперь перейдем к распределениям. Первое утверждение $\sqrt{n} \cdot \frac{\overline{X}-\mu}{\sigma} \sim \mathcal{N}(0,1)$ следует из независимости элементов выборки и свойств гауссовского вектора (см. лемма 6). Перейдем к доказательству второго утверждения. Заметим, что

$$\begin{aligned}\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 &= \sum_{i=1}^{n}\left(\frac{(X_i-\overline{X})+(\overline{X}-\mu)}{\sigma}\right)^2 \\ &= \sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 + \sum_{i=1}^{n}\left(\frac{\overline{X}-\mu}{\sigma}\right)^2 + 2\left(\frac{\overline{X}-\mu}{\sigma}\right)\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right) \\ &= \sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 + n\left(\frac{\overline{X}-\mu}{\sigma}\right)^2.\end{aligned}$$

Значит

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 + \left(\sqrt{n}\cdot\frac{\overline{X}-\mu}{\sigma}\right)^2$$

Так как $X_i \sim \mathcal{N}(\mu,\sigma^2)$ для всех $i = 1,\dots,n$,то

$$\frac{X_j-\mu}{\sigma} \sim \mathcal{N}(0,1) \quad \text{и} \quad \sqrt{n}\cdot\frac{\overline{X}-\mu}{\sigma} \sim \mathcal{N}(0,1)$$

Тогда по определению распределения хи-квадрат понятно, что

$$\underbrace{\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2}_{\sim\chi^2_n} = \underbrace{\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2}_{Y} + \underbrace{\left(\sqrt{n}\cdot\frac{\overline{X}-\mu}{\sigma}\right)^2}_{\sim\chi^2_1}$$

Докажем через характеристические функции, что $Y \sim \chi^2(n-1)$. Для χ^2_n характеристическая функция имеет вид $\varphi_X(t) = (1-2it)^{-n/2}$. Тогда из независимости слагаемых справа от равенства получаем

$$(1-2it)^{-n/2} = \varphi_Y(t)\cdot(1-2it)^{-1/2}, \quad \varphi_Y(t) = (1-2it)^{-(n-1)/2}$$

Замечаем, что характеристическая функция Y соответствует распределению χ^2_{n-1}. Остается добавить, что характеристическая функция однозначно задает распределение. Поэтому

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 \sim \chi^2_{n-1}$$

Что и требовалось доказать.

□

4 Неравенство Рао-Крамера и эффективные оценки. Сверхэффективные оценки. Экспоненциальное семейство распределений.

На прошлых лекциях мы обсудили сравнение оценок посредством функции риска и установили факт, что в классе $\mathcal{K}$ несмещенных оценок *оптимальная оценка* (см. Определение 15) единственна, если она существует (лемма 2). Возникают следующие резонные вопросы:

1. Как искать оптимальную оценку в классе $\mathcal{K}$?
2. как проверить, является ли данная оценка $\widehat{\theta}(\mathbf{X}) \in \mathcal{K}$ оптимальной?
3. можно ли по заданной выборке $\mathbf{X}$ построить оценку $\widehat{\theta}(\mathbf{X})$ со сколь угодно малой дисперсией?

Далее мы покажем, что если рассматриваемая вероятностно-статистическая модель удовлетворяет *условиям регулярности*, то имеет место нижняя оценка для дисперсии любой несмещенной оценки. Но прежде чем переходить к формулировке основных результатов, нам потребуется ввести несколько определений.

4.1 Обобщенная плотность. Функция правдоподобия

Для того чтобы не дублировать некоторые утверждения для дискретных и абсолютно непрерывных распределений введем понятие *обобщенной плотности* $p_\theta(x)$. Нам понадобятся два случая:

1. Пусть X – абсолютно непрерывная случайная величина, то есть существует такая неотрицательная функция $p(x) \geqslant 0$, что для любого $B \in \mathcal{B}(\mathbb{R})$

$$\mathbb{P}_X(B) = \int_B p(x)dx\,.$$

   Тогда ее *обобщенная плотность* и есть обычная плотность $p_\theta(x)$.

2. Пусть X – дискретная случайная величина, принимающая не более чем счетный набор значений $(x_k)_{k\in\mathbb{N}}$. Тогда ее *обобщенной плотностью* будем называть величину

$$p(x) = \begin{cases} \mathbb{P}(X = x_k), & \text{если } x = x_k; \\ 0, \text{ иначе.} \end{cases}$$

Определение 20 (Функция правдоподобия)**.** Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ – выборка с обобщенной плотностью $p_\theta(x)$. *Функцией правдоподобия* называется функция

$$\mathcal{L}(\theta) = p_\theta(\mathbf{X}) = \prod_{i=1}^{n} p_\theta(X_i).$$

Замечание 4. Функция правдоподобия – это совместная плотность данных (плотность вектора $\mathbf{X}$), за исключением того, что это функция от параметра θ, то есть $\mathcal{L} : \Theta \to \mathbb{R}_+$. Функция правдоподобия как функция от θ показывает насколько правдоподобно значение параметра θ для данной выборки $\mathbf{X} = (X_1, \ldots, X_n)$. Она не является функцией плотности: в общем случае $\int_\Theta \mathcal{L} d\theta \neq 1$.

Приведем несколько примеров вычисления функции правдоподобия и обобщенной плотности. Начнем с абсолютно непрерывного распределения.

Exercise 1. Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка из нормального распределения $\mathcal{N}(\theta, \sigma^2)$ с известной дисперсией σ^2. Тогда обобщенная плотность $p_\theta(x)$ величин X_i совпадает с обычной плотностью

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\}.$$

Функция правдоподобия параметра θ для выборки $\mathbf{X} = (X_1, \ldots, X_n)$:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} p_\theta(X_i) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{-\frac{\sum_{i=1}^n (X_i-\theta)^2}{2\sigma^2}\right\}.$$

Теперь рассмотрим дискретное распределение.

Exercise 2. Пусть $X_1, \ldots, X_n$ – выборка из распределения $\mathrm{Bin}(1, \theta)$, то есть

$$X_i = \begin{cases} 1, \text{ с вероятностью } \theta; \\ 0, \text{ с вероятностью } 1-\theta\,. \end{cases}$$

Легко видеть, что (проверьте!) величины X_i имеют обобщенную плотность

$$p_\theta(x) = \theta^x (1-\theta)^{1-x}\,, \quad x \in \{0, 1\}\,.$$

Соответствующая функция правдоподобия принимает вид

$$\mathcal{L}(\theta) = \theta^{\sum_{i=1}^n X_i} (1-\theta)^{n-\sum_{i=1}^n X_i}\,.$$

4.2 Информация Фишера. Неравенство Рао-Крамера

Далее определим понятия *вклада выборки* и *информацию Фишера* через правдоподобие данных.

Определение 21 (Вклад выборки)**.** Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ – выборка с обобщенной плотностью $p_\theta(x)$. *Вкладом* выборки $\mathbf{X}$ называется

$$U_\theta(\boldsymbol{X}) = \frac{\partial}{\partial\theta} \ln p_\theta(\boldsymbol{X}).$$

Определение 22 (Информация Фишера)**.** Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ – выборка с обобщенной плотностью $p_\theta(x)$. *Информацией Фишера* выборки $\mathbf{X}$ называется величина

$$I_n(\theta) = \mathbb{E}_\theta\big[U_\theta(\boldsymbol{X})^2\big].$$

Информация Фишера показывает количество информации, содержащееся в выборке $\mathbf{X}$, о параметре θ. Информацией, содержащейся в одном наблюдении, называют величину

$$i(\theta) = I_1(\theta) = \mathbb{E}_\theta\big[U_\theta(X_1)^2\big].$$

Exercise 3. Пусть X_1 - выборка размера 1 из нормального распределения $\mathcal{N}(\theta, \sigma^2)$ с известной дисперсией σ^2. Посчитаем информацию о неизвестном параметре θ, содержащуюся в одном элементе. Начнем с лог-правдоподобия.

$$\ln p_\theta(X_1) = -\frac{1}{2}\ln\big(2\pi\sigma^2\big) - \frac{(X_1-\theta)^2}{\sigma^2}.$$

Теперь посчитаем вклад одного элемента

$$U_\theta(X_1) = \frac{(X_1 - \theta)}{\sigma^2}.$$

Значит информация Фишера равна

$$i(\theta) = \frac{1}{\sigma^4}\mathbb{E}_\theta(X_1 - \theta) = \frac{\mathrm{Var}_\theta(X_1)}{\sigma^4} = \frac{1}{\sigma^2}.$$

Получили, что информация о неизвестном параметре θ обратно пропорциональна дисперсии, то есть чем меньше дисперсия, тем больше информации содержится в выборке, тем точнее мы сможем оценить θ. И действительно, чем меньше разброс и тем лучше случайная величина будет концентрироваться около своего среднего, в данном случае около θ.

Оказывается при выполнении некоторых условий (*условий регулярности*, перечисляемых ниже) можно указать нижнюю оценку на дисперсию несмещенных оценок. Данное утверждение будет называется неравенством Рао-Крамера.[1]

Предположение 1 (Условия регулярности).

1. Параметрическое множество Θ – открытое подмножество $\mathbb{R}$.
2. Множество $A = \{\boldsymbol{x} : p_\theta(\boldsymbol{x}) > 0\}$ не зависит от параметра θ.[2]
3. Для любой статистики $T(\boldsymbol{X})$ такой, что $\mathbb{E}_\theta[T^2(\boldsymbol{X})] < +\infty$, законно дифференцирование по параметру под знаком интеграла:

$$\frac{\partial}{\partial\theta}\mathbb{E}_\theta T(\boldsymbol{X}) = \mathbb{E}_\theta T(\boldsymbol{X})U_\theta(\boldsymbol{X}), \text{ где } U_\theta(\boldsymbol{X}) = \frac{\partial}{\partial\theta}\ln p_\theta(\boldsymbol{X}).$$

Покажем, почему соотношение выше и есть дифференцирование под знаком интеграла:[3]

$$\begin{aligned}\frac{\partial}{\partial\theta}\mathbb{E}_\theta T(\boldsymbol{X}) &= \frac{\partial}{\partial\theta}\int_A T(\boldsymbol{x})p_\theta(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_A T(\boldsymbol{x})\frac{\partial p_\theta(\boldsymbol{x})}{\partial\theta}\mathrm{d}\boldsymbol{x} \\ &= \int_A T(\boldsymbol{x})\left(\frac{1}{p_\theta(\boldsymbol{x})}\frac{\partial p_\theta(\boldsymbol{x})}{\partial\theta}\right)p_\theta(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= \int_A T(\boldsymbol{x})\frac{\partial \ln p_\theta(\boldsymbol{x})}{\partial\theta}p_\theta(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \mathbb{E}_\theta T(\boldsymbol{X})U_\theta(\boldsymbol{X}).\end{aligned}$$

4. Информация Фишера выборки $\boldsymbol{X}$ положительна и конечна при всех $\theta \in \Theta$: $0 < I_n(\theta) < \infty$.

Перед тем как перейти к неравенству Рао-Крамера, докажем ряд полезных утверждений.

Лемма 7. *Пусть* $\mathbf{X} = (X_1, \ldots, X_n)$ *– выборка из регулярного семейства с обобщенной плотностью* $p_\theta(x)$. *Тогда* $\mathbb{E}_\theta U_\theta(\mathbf{X}) = 0$.

[1] Неравенство Рао-Крамера может быть также обобщено на случай смещенных оценок Ивченко and Медведев [1984]. Мы же рассматриваем более простой вариант этого неравенства, потому что именно в случае несмещенных оценок среднеквадратичный риск совпадает с дисперсией.

[2] Данное требование обусловлено тем, что нам придется неоднократно дифференцировать по θ под знаком интеграла функции, определенные на выборочном пространстве.

[3] В машинном обучении и особенно в обучении с подкреплением такой переход к логарифму функции правдоподобия называют log-derivative trick. Он часто возникает при практическом применении оценок, полученных с помощью методов Монте-Карло. Например, см. ссылку и блог.

Доказательство. Воспользуемся условием регулярности 3 для $T(\mathbf{X}) \equiv 1$:

$$\frac{\partial}{\partial\theta}\mathbb{E}_\theta T(\mathbf{X}) = \frac{\partial}{\partial\theta}\mathbb{E}_\theta 1 = 0 = \mathbb{E}_\theta U_\theta(\mathbf{X}).$$

Следовательно, $\mathbb{E}_\theta U_\theta(\mathbf{X}) = 0$. □

Следующая лемма говорит о том, что в регулярных моделях, информация Фишера линейно возрастает с ростом объема выборки n.

Лемма 8. *Пусть* $\mathbf{X} = (X_1, \ldots, X_n)$ *– выборка из регулярного семейства с обобщенной плотностью* $p_\theta(x)$*. Тогда*

$$I_n(\theta) = ni(\theta)\,.$$

Доказательство. Воспользуемся леммой 7 и определением вклада выборки

$$I_n(\theta) = \mathbb{E}_\theta\Big[(U_\theta(\boldsymbol{X}))^2\Big] = \mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\ln p_\theta(\boldsymbol{X})\right] = \mathrm{Var}_\theta\left[\sum_{i=1}^n \frac{\partial}{\partial\theta}\ln p_\theta(X_i)\right] = n\mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\ln p_\theta(X_1)\right] = ni(\theta)\,.$$

□

Следующая лемма дает альтернативный способ вычисления информации Фишера в регулярных моделях.

Лемма 9. *Пусть* $\mathbf{X} = (X_1, \ldots, X_n)$ *– выборка из регулярного семейства с обобщенной плотностью* $p_\theta(x)$*, дважды дифференцируемая по* θ*. Тогда информация Фишера*

$$I_n(\theta) = -\mathbb{E}\left[\frac{\partial}{\partial\theta}U_\theta(\boldsymbol{X})\right].$$

Доказательство. Воспользуемся леммой 7, условием регулярности 3 и продифференцируем еще раз по θ

$$\begin{aligned}
0 = \frac{\partial}{\partial\theta}\mathbb{E}_\theta U_\theta(\mathbf{X}) &= \frac{\partial}{\partial\theta}\int_A U_\theta(\mathbf{X})p_\theta(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \frac{\partial}{\partial\theta}\int_A\left(\frac{\partial}{\partial\theta}\ln p_\theta(\boldsymbol{x})\right)p_\theta(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \int_A\left(\frac{\partial^2}{\partial\theta^2}\ln p_\theta(\boldsymbol{x})\right)p_\theta(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_A\left(\frac{\partial}{\partial\theta}\ln p_\theta(\boldsymbol{x})\right)\frac{\partial}{\partial\theta}p_\theta(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \int_A\left(\frac{\partial^2}{\partial\theta^2}\ln p_\theta(\boldsymbol{x})\right)p_\theta(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_A\left(\frac{\partial}{\partial\theta}\ln p_\theta(\boldsymbol{x})\right)^2 p_\theta(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \mathbb{E}\left[\frac{\partial}{\partial\theta}U_\theta(\mathbf{X})\right] + \mathbb{E}\big[U_\theta(\mathbf{X})^2\big] = \mathbb{E}\left[\frac{\partial}{\partial\theta}U_\theta(\mathbf{X})\right] + I_n(\theta).
\end{aligned}$$

Откуда получаем требуемое равенство. □

Теперь перейдем к основному результату, а имменно к нижней границе на дисперсию несмещенных оценок в регулярных моделях.

Теорема 6 (Неравенство Рао-Крамéра)**.** *Пусть* $\mathbf{X} = (X_1, \dots, X_n)$ *– выборка из регулярного семейства с обобщенной плотностью* $p_\theta(x)$*. Пусть* $\tau(\theta)$ *— дифференцируемая функция, и* $\widehat{\theta}(\mathbf{X})$ *— несмещённая оценка* $\tau(\theta)$ *такая, что* $\mathbb{E}_\theta[\widehat{\theta}^2(\mathbf{X})] < \infty$*. Тогда для всех* $\theta \in \Theta$

$$\mathrm{Var}_\theta(\widehat{\theta}(\boldsymbol{X})) \geqslant \frac{(\tau'(\theta))^2}{ni(\theta)}.$$

Доказательство. Из несмещенности имеем

$$\mathbb{E}_\theta[\widehat{\theta}(\boldsymbol{X})] = \tau(\theta).$$

воспользуемся свойством регулярности 3, продифференцировав это уравнение по θ

$$\tau'(\theta) = \frac{\partial}{\partial\theta}\mathbb{E}_\theta\widehat{\theta}(\boldsymbol{X}) = \mathbb{E}_\theta[\widehat{\theta}(\boldsymbol{X})U_\theta(\boldsymbol{X})]$$

Сведем математическое ожидание в правой части равенство к ковариации, а затем воспользуемся неравенством Коши-Буняковского-Шварца. Для этого применим лемму 7 к уравнению выше

$$\begin{aligned}\tau'(\theta) = \mathbb{E}_\theta[\widehat{\theta}(\boldsymbol{X})U_\theta(\boldsymbol{X})] &= \mathbb{E}_\theta[(\widehat{\theta}(\boldsymbol{X}) - \tau(\theta))U_\theta(\boldsymbol{X})] + \mathbb{E}_\theta[\tau(\theta)U_\theta(\boldsymbol{X})] \\ &= \mathbb{E}_\theta[(\widehat{\theta}(\boldsymbol{X}) - \tau(\theta))U_\theta(\boldsymbol{X})] \\ &= \mathbb{E}_\theta[(\widehat{\theta}(\boldsymbol{X}) - \tau(\theta))(U_\theta(\boldsymbol{X}) - \mathbb{E}_\theta U_\theta(\boldsymbol{X}))] \\ &= \mathrm{cov}(\widehat{\theta}(\boldsymbol{X}), U_\theta(\boldsymbol{X})).\end{aligned}$$

Теперь, применяя неравенство Коши-Буняковского-Шварца, получим

$$(\tau'(\theta))^2 = \mathrm{cov}(\widehat{\theta}(\boldsymbol{X}), U_\theta(\boldsymbol{X}))^2 \leqslant \mathrm{Var}_\theta(\widehat{\theta}(\boldsymbol{X}))\mathrm{Var}_\theta(U_\theta(\boldsymbol{X})) = \mathrm{Var}_\theta(\widehat{\theta}(\boldsymbol{X}))I_n(\theta). \tag{2}$$

Следовательно,

$$\mathrm{Var}_\theta(\widehat{\theta}(\boldsymbol{X})) \geqslant \frac{(\tau'(\theta))^2}{I_n(\theta)}.$$

Далее остается только применить лемму 8 и получить требуемое неравенство. □

Таким образом, в регулярной модели мы не можем рассчитывать на оценку искомого параметра θ с дисперсией, убывающей быстрее чем $1/n$ с ростом размера выборки n.

Определение 23 (Эффективная оценка)**.** Пусть выполнены условия регулярности. Несмещенная оценка $\widehat{\theta}(\boldsymbol{X})$ для $\tau(\theta)$ называется *эффективной*, если при любом $\theta \in \Theta$ для неё достигается равенство в неравенстве Рао-Крамера, то есть для всех $\theta \in \Theta$

$$\mathrm{Var}_\theta(\widehat{\theta}(\boldsymbol{X})) = \frac{(\tau'(\theta))^2}{ni(\theta)}.$$

Замечание 5. Если $\tau(\theta) = \theta$, то определение (23) упрощается

$$\mathrm{Var}_\theta(\widehat{\theta}(\boldsymbol{X})) = \frac{1}{ni(\theta)}.$$

Для доказательства эффективности часто бывает полезен следующий критерий эффективности

Следствие 1 (Критерий эффективности)**.** *Пусть выполнены условия регулярности. Оценка эффективна тогда и только тогда, когда оценка $\widehat{\theta}(\mathbf{X})$ является линейной функцией вклада выборки $U_\theta(\boldsymbol{X})$, то есть*

$$\widehat{\theta}(\boldsymbol{X}) - \tau(\theta) = c(\theta)U_\theta(\boldsymbol{X})\,, \text{ где } c(\theta) = \frac{\tau'(\theta)}{ni(\theta)}\,.$$

Доказательство. Неравенство Коши-Шварца (2) переходит в равенство тогда и только тогда, когда между случайными величинами есть линейная зависимость, то есть существуют функции $c(\theta)$ и $a(\theta)$ такие, что

$$\widehat{\theta}(\boldsymbol{X}) = c(\theta)U_\theta(\boldsymbol{X}) + a(\theta)\,. \tag{3}$$

Найдем эти функции. Для начала возьмём матожидание от этого равенства и воспользуемся леммой 7

$$\mathbb{E}_\theta\widehat{\theta}(\boldsymbol{X}) = \mathbb{E}_\theta[c(\theta)U_\theta(\boldsymbol{X}) + a(\theta)].$$

Тогда получаем

$$\tau(\theta) = c(\theta)\cdot 0 + a(\theta) \implies a(\theta) = \tau(\theta).$$

Теперь умножим обе части равенства (3) на $U_\theta(\boldsymbol{X})$ и возьмём матожидание:

$$\mathbb{E}_\theta[(\widehat{\theta}(\boldsymbol{X}) - \tau(\theta))U_\theta(\boldsymbol{X})] = \mathbb{E}_\theta[c(\theta)U_\theta^2(\boldsymbol{X})].$$

Тогда по свойству регулярности 3 получаем

$$\tau'(\theta) = c(\theta)I_n(\theta) \quad \implies \quad c(\theta) = \frac{\tau'(\theta)}{I_n(\theta)}.$$

□

Лемма 10. *Если оценка $\widehat{\theta}(\mathbf{X})$ — эффективная оценка для $\tau(\theta)$, то она также является оптимальной. Обратное неверно.*

Доказательство. Утверждение леммы следует из теоремы 6 и определения оптимальной оценки 15.

Обратное утверждение неверно, потому что в нерегулярных моделях дисперсия может быть меньше нижней границы неравенства Рао-Крамера. □

Следующий пример показывает, что в нерегулярных моделях можно привести несмещенные оценки с дисперсией меньшей, чем нижняя граница неравенства Рао-Крамера.

Пример 3. Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ – выборка из равномерного распределения $U[0, \theta]$, где $\theta > 0$. Покажем, что несмещенная оценка параметра θ

$$\widehat{\theta}(\mathbf{X}) = \frac{n+1}{n}X_{(n)}.$$

обладает дисперсией $\mathrm{Var}(\widehat{\theta}(\mathbf{X})) = \frac{\theta^2}{n(n+2)}$

Доказательство. Данная модель не будет регулярна, поскольку нарушено условие 2 (пределы интегрирования во всех последующих вычислениях зависят от θ, и поэтому дифференцировать под

знаком интеграла уже не получится). Заметим, что функция распределения $n-$й порядковой статистики имеет вид

$$\mathbb{P}_\theta(X_{(n)} \leqslant t) = \mathbb{P}_\theta(X_1 \leqslant t, \dots, X_n \leqslant t) = \prod_{i=1}^{n} \mathbb{P}_\theta(X_i \leqslant t) = \begin{cases} 0, t \leqslant 0 \\ \left(\frac{t}{\theta}\right)^n, 0 < t < \theta \\ 1, t \geqslant \theta\,. \end{cases}$$

Следовательно, ее плотность распределения имеет вид

$$p_\theta(t) = \begin{cases} \frac{nt^{n-1}}{\theta^n}, 0 < t < \theta \\ 0, \text{ иначе }. \end{cases}$$

Посчитаем моменты, а затем вычислим дисперсию

$$\mathbb{E}_\theta\left[\frac{n+1}{n} X_{(n)}\right] = \frac{n+1}{n} \int_0^\theta \frac{nt^n}{\theta^n} \mathrm{d}t = \theta$$
$$\mathbb{E}_\theta\left[\left(\frac{n+1}{n} X_{(n)}\right)^2\right] = \frac{(n+1)^2}{n^2} \int_0^\theta \frac{nt^{n+1}}{\theta^n} \mathrm{d}t = \frac{(n+1)^2}{n(n+2)}\theta^2$$

Тогда дисперсия оценки будет

$$\mathrm{Var}_\theta\left[\frac{n+1}{n} X_{(n)}\right] = \mathbb{E}_\theta\left[\left(\frac{n+1}{n} X_{(n)}\right)^2\right] - \left(\mathbb{E}_\theta\left[\frac{n+1}{n} X_{(n)}\right]\right)^2 = \frac{(n+1)^2}{n(n+2)}\theta^2 - \theta^2 = \frac{\theta^2}{n(n+2)}\,.$$

□

Таким образом мы получили, что дисперсия оценки из примера 3 убывает с ростом n как $1/n^2$ в то время как нижняяя граница в неравенстве Рао-Крамера соответствует $1/n$. Подобные оценки называются *сверх-эффективными*, они возможны только в нерегулярных моделях.

5 Экспоненциальное семейство распределений. Эффективные оценки в экспоненциальном семействе.

5.1 Экспоненциальное семейство распределений.

Рассмотрим класс параметрических моделей, для которых всегда существует эффективная (а, значит, и оптимальная) оценка некоторой параметрической функции. Модели из данного класса называются *экспоненциальными*.

Определение 24 (Экспоненциальное семейство распределений)**.** Пусть $\mathcal{P} = \{\mathbb{P}_\theta \colon \theta \in \Theta \in \mathbb{R}\}$ – параметрическое семейство с обобщенной плотностью $p_\theta(x)$. Будем говорить, что распределение $\mathbb{P}_\theta$ принадлежит к экспоненциальному классу распределений, если его обобщенная плотность представляется в виде

$$p_\theta(x) = \exp\{\eta(\theta)T(x) - A(\theta) + H(x)\}, \tag{4}$$

где η, T, H — некоторые функции, а A – нормировочная константа.

Замечание 6. В случае векторного параметра $\theta \in \Theta \in \mathbb{R}^r$ равенство (4) заменяется на

$$p_\theta(x) = \exp\left\{\sum_{k=1}^{r} \eta_k(\theta)T_k(x) - A(\theta) + H(x)\right\}.$$

Класс экспоненциальных распределений достаточно широк: почти все известные распределения принадлежат экспоненциальному семейству, за исключением равномерное распределения. Покажем это.

Пример 4. Рассмотрим распределение Пуассона $\mathrm{Pois}(\theta)$ и $\theta > 0$. Обобщенная плотность

$$p_\theta(x) = \frac{\theta^x e^{-\theta}}{x!} = \exp\left\{\ln\left(\frac{\theta^x e^{-\theta}}{x!}\right)\right\} = \exp\{x\ln\theta - \theta - \ln x!\}.$$

Тогда $\eta(\theta) = \ln\theta,\ T(x) = x,\ A(\theta) = -\theta,\ H(x) = -\ln x!$.

Пример 5. Рассмотрим распределение Бернулли $\mathrm{Bin}(1,\theta)$ и $\theta \in \{0,1\}$. Обобщенная плотность

$$p_\theta(x) = \theta^x(1-\theta)^{1-x} = \exp\{x\ln\theta + (1-x)\ln(1-\theta)\} = \exp\{x(\ln\theta - \ln(1-\theta)) + \ln(1-\theta)\}.$$

Пример 6. Рассмотрим нормальное распределение $\mathcal{N}(\theta,\sigma^2)$ с неизвестным средним θ и известной дисперсией σ^2. Рассмотрим плотность распределения

$$p_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\} = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{\frac{\theta x}{\sigma^2} - \frac{\theta^2}{2\sigma^2} - \frac{x^2}{2\sigma^2}\right\}.$$

Свойства экспоненциального семейства

1. Пусть дана выборка $\mathbf{X} = (X_1,\ldots,X_n)$ из экспоненциального семейства, тогда и распределение вектора $\mathbf{X}$ также принадлежит экспоненциальному семейству.

   Действительно

$$\begin{aligned} p_\theta(\boldsymbol{x}) = \prod_{i=1}^{n} p_\theta(x_i) &= \prod_{i=1}^{n} \exp\{\eta(\theta)T(x_i) - A(\theta) + H(x_i)\} \\ &= \exp\left\{\eta(\theta)\sum_{i=1}^{n} T(x_i) - nA(\theta) + \sum_{i=1}^{n} H(x_i)\right\} \end{aligned}$$

Статистику $\sum_{i=1}^n T(X_i)$ обычно называют *достаточной статистикой*, так как для того, чтобы посчитать вклад выборки или информацию Фишера достаточно лишь знать $\sum_{i=1}^n T(X_i)$. То есть вместо того, чтобы хранить всю выборку, достаточно лишь хранить $\sum_{i=1}^n T(X_i)$. Покажем это, распишем вклад выборки для экспоненциального семейства

$$U_\theta(\mathbf{X}) = \frac{\partial}{\partial\theta}\ln p_\theta(\mathbf{X}) = \eta'(\theta)\sum_{i=1}^n T(x_i) - nA'(\theta).$$

То есть в статистике $\sum_{i=1}^n T(X_i)$ содержится столько же информации о неизвестном параметре θ сколько и во всей выборке $X_1, \dots, X_n$.

Пример 7. Рассмотрим распределение Пуассона $\mathrm{Pois}(\theta)$, $\theta > 0$, для которого $T(x) = x$ (см. пример (4)). Тогда $\sum_{i=1}^n T(X_i) = \sum_{i=1}^n X_i$.

Замечание 7. Можно показать, что в равномерном распределении $U[0;\theta]$ достаточной будет статистика $T(\mathbf{X}) = X_{(n)} \neq \sum_{i=1}^n T(X_i)$.

2. Вогнутость логарифма правдоподобия относительно параметра $\eta = \eta(\theta)$.

5.2 Эффективные оценки в экспоненциальном семействе

Следующая теорема позволяет наиболее просто находить эффективные оценки и функции, для которых существуют эффективные оценки для распределений из экспоненциального семейства.

Теорема 7. *Пусть $\mathbf{X} = (X_1, X_2, \dots, X_n)$ — выборка из регулярной модели $\{\mathbb{P}_\theta\colon \theta \in \Theta \subseteq \mathbb{R}\}$. Эффективная оценка $\widehat{\theta}(\mathbf{X})$ для некоторой функции $\tau(\theta)$ существует тогда и только тогда, когда $\mathbb{P}_\theta$ принадлежит экспоненциальному семейству, причём*

$$\widehat{\theta}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n T(X_i), \quad \text{и} \quad \tau(\theta) = \frac{A'(\theta)}{\eta'(\theta)}. \tag{5}$$

(последние два равенства следует понимать с точностью до линейного преобразования).

Доказательство. 1. Докажем сначала прямое утверждение, то есть пусть существует эффективная оценка $\widehat{\theta}(\mathbf{X})$ для некоторой $\tau(\theta)$, тогда покажем, что распределение принадлежит экспоненциальному семейству.

Если $\widehat{\theta}(\mathbf{X})$ – эффективная оценка, то для нее должен выполняться критерий эффективности (Следствие 1)

$$\widehat{\theta}(\mathbf{X}) = \tau(\theta) + \frac{\tau'(\theta)}{ni(\theta)}U_\theta(\mathbf{X}) = \tau(\theta) + \frac{\tau'(\theta)}{ni(\theta)}\sum_{i=1}^n U_\theta(X_i), \tag{6}$$

так как

$$U_\theta(\mathbf{X}) = \frac{\partial}{\partial\theta}\ln p_\theta(\mathbf{X}) = \frac{\partial}{\partial\theta}\prod_{i=1}^n \ln p_\theta(X_i) = \sum_{i=1}^n \frac{\partial}{\partial\theta}\ln p_\theta(X_i) = \sum_{i=1}^n U_\theta(X_i).$$

Заметим теперь, что левая часть (6) не зависит от неизвестного параметра θ (по определению оценки), значит и правая часть (6) не должна зависеть от θ. Тогда обозначим правую часть (6) за $\frac{1}{n}\sum_{i=1}^n C(X_i)$, то есть

$$\widehat{\theta}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n C(X_i) = \frac{1}{n}\sum_{i=1}^n\left(\tau(\theta) + \frac{\tau'(\theta)}{i(\theta)}U_\theta(X_i)\right). \tag{7}$$

Теперь из этого уравнения выразим вклад одного наблюдения

$$U_\theta(x) = \frac{\partial}{\partial\theta}\ln p_\theta(x) = \frac{i(\theta)}{\tau'(\theta)}C(x) - \frac{i(\theta)\tau(\theta)}{\tau'(\theta)}.$$

Возьмем неопределенный интеграл от этого выражения

$$\ln p_\theta(x) = C(x)\int\frac{i(\theta)}{\tau'(\theta)}\mathrm{d}\theta - \int\frac{i(\theta)\tau(\theta)}{\tau'(\theta)}\mathrm{d}\theta + B(x).$$

Тогда получаем, что распределение принадлежит экспоненциальному семейству с

$$\eta(\theta) = \int\frac{i(\theta)}{\tau'(\theta)}\,\mathrm{d}\theta\,, \quad T(x) = C(x), \quad A(\theta) = \int\frac{i(\theta)\tau(\theta)}{\tau'(\theta)}\,\mathrm{d}\theta\,.$$

Используем это в (7) и получим вид эффективных оценок

$$\widehat{\theta}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}T(X_i). \tag{8}$$

Теперь найдем вклад выборки, информацию Фишера и подставим их совместно с видом эффективной оценки в критерий эффективности (Следствие 1), чтобы найти вид $\tau(\theta)$. Вклад выборки одного элемента

$$U_\theta(X_1) = \frac{\partial}{\partial\theta}(\eta(\theta)T(X_1) - A(\theta) + H(X_1)) = \eta'(\theta)T(X_1) - A'(\theta) + H(X_1). \tag{9}$$

Тогда информация Фишера одного элемента

$$i(\theta) = -\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}U_\theta(X_1)\right] = -\eta''(\theta)\mathbb{E}_\theta T(X_1) + A''(\theta) = -\eta''(\theta)\tau(\theta) + A''(\theta), \tag{10}$$

где в последнем равенстве мы воспользовались несмещенностью оценки $\mathbb{E}_\theta T(X_1) = \tau(\theta)$. Теперь подставим (9), (10) и (8) в критерий эффективности (Следствие 1)

$$\frac{1}{n}\sum_{i=1}^{n}T(X_i) = \frac{1}{n}\sum_{i=1}^{n}\left(\tau(\theta) - \frac{\tau'(\theta)(\eta'(\theta)T(X_i) - A'(\theta))}{A''(\theta) - \eta''(\theta)\tau(\theta)}\right).$$

Заметим снова, что в данном тождестве слева перед функцией $T(X_i)$ стоит множитель 1, значит и справа коэффициент перед $T(X_i)$ должен равняться 1. Получаем

$$\frac{\tau'(\theta)\eta'(\theta)}{A''(\theta) - \eta''(\theta)\tau(\theta)} = 1.$$

Приравняем числитель и знаменатель в тождестве выше

$$\tau'(\theta)\eta'(\theta) = A''(\theta) - \eta''(\theta)\tau(\theta)$$

Данное тождество можно переписать следующим образом

$$\big(\tau(\theta)\eta'(\theta)\big)' = A''(\theta).$$

Возьмем неопределенный интеграл по θ и получим требуемое утверждение (5) с точностью до линейных преобразований.

2. Чтобы доказать утверждение в обратную сторону, достаточно лишь проверить, что критерий эффективности выполняется для $\tau(\theta)$ и $\widehat{\theta}(\mathbf{X})$ из (5).

□

5.3 Многомерное неравенство Рао-Крамера.

Определенные выше понятия и критерии оптимальности непосредственно переносятся и на случай векторного параметра $\theta = (\theta_1, \dots, \theta_r) \in \mathbb{R}^r$. В этом случае под вкладом выборки $\mathbf{X}$ мы понимаем вектор $\mathbf{U}_\theta(\mathbf{X}) = (U_{\theta,1}(\mathbf{X}), \dots, U_{\theta,r}(\mathbf{X}))$, где

$$U_{\theta,i}(\mathbf{X}) = \frac{\partial}{\partial \theta_i} \ln p_\theta(\boldsymbol{X}) \,.$$

Аналогом информации Фишера (см. определение 22) является информационная матрица Фишера:

$$\mathbf{I}_n(\theta) \in \mathbb{R}^{r\times r} \,, (\mathbf{I}_n(\theta))_{i,j} = \mathbb{E}_\theta[U_{\theta,i}(\mathbf{X}) U_{\theta,j}(\mathbf{X})] \,.$$

При этом информационная матрица одного наблюдения имеет вид

$$\mathbf{I}_1(\theta) \in \mathbb{R}^{m\times m} \,, (\mathbf{I}_1(\theta))_{i,j} = \mathbb{E}_\theta \left[\frac{\partial}{\partial \theta_i} \ln p_\theta(X_1) \frac{\partial}{\partial \theta_j} \ln p_\theta(X_1) \right] = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln p_\theta(X_1) \right] .$$

Последнее равенство верно для дважды дифференцируемых плотностей и проверяется интегрированием по частям. Теперь мы можем сформулировать многомерный аналог неравенства Рао-Крамера:

Теорема 8 (Многомерное неравенство Рао-Крамéра)**.** *Пусть* $\mathbf{X}$ *- выборка из* $\mathbb{P} \in \{\mathbb{P}_\theta : \theta \in \Theta \in \mathbb{R}^r\}$*, удовлетворяющего условиям регулярности. Пусть* $\tau(\theta) : \mathbb{R}^r \mapsto \mathbb{R}$ *— дифференцируемая функция, и* $\widehat{\theta}(\mathbf{X})$ *— несмещённая оценка* $\tau(\theta)$ *такая, что* $\mathbb{E}_\theta[\widehat{\theta}^2(\mathbf{X})] < \infty$*. Тогда для всех* $\theta \in \Theta$*:*

$$\mathrm{Var}_\theta(\widehat{\theta}(\boldsymbol{X})) \geqslant \frac{1}{n} (\nabla \tau(\theta))^\intercal \mathbf{I}_1^{-1}(\theta) \nabla \tau(\theta) \,. \tag{11}$$

Данное неравенство обращается в равенство тогда и только тогда, когда оценка $\widehat{\theta}(\mathbf{X})$ *является линейной функцией вклада выборки, то есть*

$$\widehat{\theta}(\boldsymbol{X}) - \tau(\theta) = c(\theta)^\intercal \mathbf{U}_\theta(\boldsymbol{X}) \,, \text{ где } c(\theta) = \frac{1}{n} \mathbf{I}_1(\theta)^{-1} \nabla \tau(\theta) \,.$$

Доказательство теоремы аналогично доказательству в одномерном случае.

6 Оптимальная оценка в модели линейной регрессии

6.1 Мотивационный пример

До сих пор в рамках курса рассматривались задачи оценивания и статистического вывода по выборке из независимых одинаково распределенных случайных элементов. Однако имеется широкий класс прикладных задач, в которых наблюдения зависят от одного или нескольких факторов, значения которых могут отличаться для различных элементов выборки. Рассмотрим следующий пример. Допустим, что в среднем ежемесячный объем продаж Y магазина мобильных устройств зависит от устанавливаемых магазином цен P и от затрат на рекламу A по следующей формуле:

$$\mathbb{E}Y = \theta_1 + \theta_2 P + \theta_3 A + \theta_4 A^2.$$

Отметим, что фактический объем продаж Y может меняться от месяца к месяцу даже если цены устройств и затраты на рекламу оставались неизменными. Это связано с тем, что значение Y определяется также и некоторыми случайными факторами. Поэтому можно составить следующую вероятностную модель:

$$Y = \theta_1 + \theta_2 P + \theta_3 A + \theta_4 A^2 + \varepsilon, \quad \text{где } \mathbb{E}(\varepsilon \,|\, P, A) = 0.$$

Форма зависимости $\mathbb{E}Y$ от факторов P и A может быть продиктована общими соображениями, экономическими законами или результатами исследований. Например, добавление слагаемого $\theta_4 A^2$ при отрицательных значениях коэффициента θ_4 может быть обусловлено тем, чрезмерные траты на рекламу не принесут желаемого эффекта. Определение этой формы выходит за рамки предмета изучения математической статистики. Однако математическая статистика позволяет оценить значения коэффициентов $\theta_1, \theta_2, \theta_3$ и θ_4 по эмпирическим наблюдениям. Допустим, что имеется набор наблюдений $\{(P_i, A_i, Y_i) : 1 \leqslant i \leqslant n\}$, причем

$$Y_i = \theta_1 + \theta_2 P_i + \theta_3 A_i + \theta_4 A_i^2 + \varepsilon_i, \quad \text{где } \mathbb{E}(\varepsilon \,|\, P, A) = 0 \text{ для всех } i \in \{1, \ldots, n\}. \tag{12}$$

Хорошая оценка $\widehat{\boldsymbol{\theta}}_{LSE} = (\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3, \widehat{\theta}_4)^\intercal$ вектора коэффициентов $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)^\intercal$ позволит построить оценку систематической составляющей объема продаж:

$$\widehat{Y} = \widehat{\theta}_1 + \widehat{\theta}_2 P + \widehat{\theta}_3 A + \widehat{\theta}_4 A^2,$$

а следовательно, поможет определить оптимальный уровень цен и затрат на рекламу для максимизации выручки. Задача определения зависимости интересующей нас величины (“отклика”) от различных факторов называется задачей регрессии. Заметим, что в модели (12) значение $\mathbb{E}Y$ зависит линейно от неизвестного вектора параметров θ, поэтому (12) относится к модели *линейной* регрессии, несмотря на то, что фактор A входит в нее нелинейно.

6.2 Линейная регрессия

В общем виде регрессионную задачу можно сформулировать следующим образом. По имеющимся данным $\{(X_i, Y_i) : 1 \leqslant i \leqslant n\}$ нужно для нового входного x предсказать y, используя регрессионную функцию $f(x)$ такую, что $y \approx f(x)$. То есть предполагаем, что случайные наблюдения $Y_1, \ldots, Y_n$ получены из модели

$$Y_i = f(X_i) + \varepsilon_i, \quad i \in \{1, \ldots, n\}, \tag{13}$$

где $\mathbb{E}(\varepsilon_i \,|\, X_1, \ldots, X_n) = 0$ для всех i от 1 до n. Далее для простоты считаем, что $X_1, \ldots, X_n$ – детерминированные величины. Тогда условие $\mathbb{E}(\varepsilon_i \,|\, X_1, \ldots, X_n) = 0$, $1 \leqslant i \leqslant n$ можно заменить на $\mathbb{E}\varepsilon_i = 0$ для всех $i \in \{1, \ldots, n\}$.

Будем предполагать, что регрессионная функция f принадлежит некоторому параметрическому семейству $\{f_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta \in \mathbb{R}^p\}$. Более того, будем рассматривать лишь модели, которые линейны по параметру $\boldsymbol{\theta}$. Важно заметить, что мы предполагаем линейность по параметру $\boldsymbol{\theta}$, в то время как по входной переменной x $f(x)$ может быть нелинейна

$$f(x) = \sum_{j=1}^{p} \psi_j(x)\theta_j = \boldsymbol{\psi}(x)^\intercal \boldsymbol{\theta}, \tag{14}$$

где $\psi_i : \mathcal{X} \to \mathbb{R}$ известны $(i = 1, \ldots, p)$, а $\boldsymbol{\psi}(x) = (\psi_1(x), \ldots, \psi_p(x))$ – вектор-функция.

Пример 8. $f(x)$ – полиномиальная функция степени $p-1$

$$f(x) = \sum_{j=1}^{p} \theta_j x^{j-1}.$$

Тогда задачу регрессии (13) можно переписать следующим образом

$$Y_i = \boldsymbol{\psi}(X_i)^\intercal \boldsymbol{\theta} + \varepsilon_i, \quad i \in \{1, \ldots, n\},$$

В матрично-векторной форме:

$$\mathbf{Y} = \Psi^\intercal \boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad \mathbb{E}\boldsymbol{\varepsilon} = \mathbf{0}. \tag{15}$$

где $\Psi = (\boldsymbol{\psi}(X_1), \ldots, \boldsymbol{\psi}(X_n)) \in \mathbb{R}^{p \times n}$ называется *матрицей плана*. Мы предполагаем, что матрица $\Psi \in \mathbb{R}^{p \times n}$ имеет полный ранг, то есть ранг Ψ равен p. Эквивалентно, мы предполагаем, что $\Psi\Psi^\intercal \in \mathbb{R}^{p \times p}$ обратима.

Наша задача состоит в оценке параметра $\boldsymbol{\theta}$ и регрессионной функции $f(x) = \boldsymbol{\psi}(x)^\intercal \boldsymbol{\theta}$. В случае, когда истинная функция f не принадлежит параметрическому классу (14), модель называют мисспецифицированной.

6.3 Метод наименьших квадратов

Широко распространенным методом оценивания вектора коэффициентов $\boldsymbol{\theta}$ в модели линейной регрессии (15) является метод наименьших квадратов. Будем рассматривать квадратичные потери и задачу минимизации эмпирического риска

$$\widehat{\boldsymbol{\theta}}_{LSE} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^{n} (Y_i - \boldsymbol{\psi}(X_i)^\intercal \boldsymbol{\theta})^2 = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \|\mathbf{Y} - \Psi^\intercal \boldsymbol{\theta}\|^2. \tag{16}$$

Находя градиент функции $\|\mathbf{Y} - \Psi^\intercal \boldsymbol{\theta}\|^2$ и приравнивая его к нулю, получаем, что $\widehat{\boldsymbol{\theta}}_{LSE}$ удовлетворяет уравнению

$$2\Psi(\Psi^\intercal \widehat{\boldsymbol{\theta}}_{LSE} - \mathbf{Y}) = \mathbf{0}.$$

Если матрица Ψ является матрицей ранга p, то матрица $\Psi\Psi^\intercal$ невырождена и значение оценки метода наименьших квадратов $\widehat{\boldsymbol{\theta}}$ можно найти по формуле[4]

$$\widehat{\boldsymbol{\theta}}_{LSE} = (\Psi\Psi^\intercal)^{-1}\Psi\mathbf{Y}.$$

[4] В общем случае в качестве оценки $\widehat{\boldsymbol{\theta}}_{LSE}$ можно взять $(\Psi\Psi^\intercal)^\dagger \Psi Y$, где $(\Psi\Psi^\intercal)^\dagger$ – псевдообратная к $\Psi\Psi^\intercal$ матрица. Заметим, что если ранг матрицы Ψ меньше размерности вектора параметров $\boldsymbol{\theta}$, то решение задачи оптимизации (16) неединственно.

6.4 Статистические свойства и геометрическая интерпретация

Перейдем к рассмотрению статистических свойств оценки $\widehat{\boldsymbol{\theta}}_{LSE}$.

1. Линейность (по Y):
$$\widehat{\boldsymbol{\theta}}_{LSE} = (\Psi\Psi^\intercal)^{-1}\Psi\mathbf{Y} =: AY.$$

2. Несмещенность
$$\mathbb{E}_{\boldsymbol{\theta}}\boldsymbol{\theta}_{LSE} = \mathbb{E}_{\boldsymbol{\theta}}(\Psi\Psi^\intercal)^{-1}\Psi\mathbf{Y} = \mathbb{E}_{\boldsymbol{\theta}}(\Psi\Psi^\intercal)^{-1}\Psi(\Psi^\intercal\boldsymbol{\theta} + \varepsilon) = \boldsymbol{\theta}.$$

3. Дисперсия:
$$\begin{aligned}\mathrm{Var}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_{LSE}) &= \mathrm{Var}_{\boldsymbol{\theta}}((\Psi\Psi^\intercal)^{-1}\Psi\mathbf{Y}) = \mathrm{Var}_{\boldsymbol{\theta}}((\Psi\Psi^\intercal)^{-1}\Psi(\Psi^\intercal\boldsymbol{\theta} + \varepsilon)) = \mathrm{Var}_{\boldsymbol{\theta}}((\Psi\Psi^\intercal)^{-1}\Psi\varepsilon)\\ &= \mathbb{E}_{\boldsymbol{\theta}}((\Psi\Psi^\intercal)^{-1}\Psi\varepsilon)(\Psi\Psi^\intercal)^{-1}\Psi\varepsilon)^\intercal = (\Psi\Psi^\intercal)^{-1}\Psi\mathbb{E}_{\boldsymbol{\theta}}[\varepsilon\varepsilon^\intercal]\Psi^\intercal(\Psi\Psi^\intercal)^{-1}\\ &= \sigma^2(\Psi\Psi^\intercal)^{-1}.\end{aligned}$$

Теперь перейдем к геометрической интерпретации оценки регрессионной функции $f(x) = \boldsymbol{\psi}(x)^\intercal\boldsymbol{\theta}$:
$$\widehat{f} = \Psi^\intercal\widehat{\boldsymbol{\theta}}_{LSE} = \Psi^\intercal(\Psi\Psi^\intercal)^{-1}\Psi\mathbf{Y},$$

Обозначим
$$\Pi = \Psi^\intercal(\Psi\Psi^\intercal)^{-1}\Psi \in \mathbb{R}^{n\times n}.$$

Утверждение 1. $\widehat{f} = \Psi^\intercal\widehat{\boldsymbol{\theta}}_{LSE} \equiv \Pi\mathbf{Y}$ *является проекцией вектора наблюдений на линейную оболочку строк матрицы плана* Ψ*, то есть на* $\mathrm{Im}(\Psi^\intercal)$.

Доказательство. Непосредственно проверяется, что Π – ортопроектор, то есть $\Pi^\intercal = \Pi$ и $\Pi^2 = \Pi$:

1.
$$\Pi^\intercal = (\Psi^\intercal(\Psi\Psi^\intercal)^{-1}\Psi)^\intercal = \Psi^\intercal(\Psi^\intercal(\Psi\Psi^\intercal)^{-1})^\intercal = \Psi^\intercal(\Psi\Psi^\intercal)^{-1}\Psi = \Pi$$

2.
$$\Pi^2 = (\Psi^\intercal(\Psi\Psi^\intercal)^{-1}\Psi)^\intercal\cdot(\Psi^\intercal(\Psi\Psi^\intercal)^{-1}\Psi)^\intercal = (\Psi^\intercal(\Psi\Psi^\intercal)^{-1}\Psi)^\intercal = \Pi.$$

Из определения ортопроектора следует, что все собственные значения матрицы Π удовлетворяют уравнению
$$\lambda^2 = \lambda,$$
то есть принимают значение либо 0, либо 1. Следовательно, все пространство $\mathbb{R}^n$ разбивается на прямую сумму собственного подпространства оператора Π, отвечающего значению 1, и его ядра. То есть Π задает ортогональную проекцию на некоторое линейное подпространство в $\mathbb{R}^n$. Покажем теперь, что Π – ортопроектор y на $\mathrm{Im}(\Psi^\intercal)$ (линейную оболочку строк матрицы плана).

1. Для начала покажем, что $\Pi u = u$ для любого $u \in \mathrm{Im}(\Psi^\intercal)$. Для любого $a \in \mathbb{R}^p$, $u := \Psi^\intercal a \in \mathrm{Im}(\Psi^\intercal)$, тогда
$$\Pi u = \Pi\Psi^\intercal a = \Psi^\intercal(\Psi\Psi^\intercal)^{-1}\Psi\Psi^\intercal a = \Psi^\intercal a = u.$$
То есть $\Pi u = u$ для любого $u \in \mathrm{Im}(\Psi^\intercal)$.

2. осталось показать, что $\Pi v = 0$ для любого $v \in \mathrm{Im}(\Psi^\intercal)^\perp$. Рассмотрим $z \in \mathrm{Ker}(\Psi)$ и $x \in \mathbb{R}^n$
$$0 = \langle x, \Psi z\rangle = \langle\Psi^\intercal x, z\rangle = \langle\mathrm{Im}(\Psi^\intercal), z\rangle.$$
Так как $z \in \mathrm{Ker}(\Psi)$, получаем, что $\mathrm{Im}(\Psi^\intercal) = \mathrm{Ker}(\Psi)^\perp$.

□

6.5 Теорема Гаусса-Маркова

Перейдем к рассмотрению статистических свойств оценки метода наименьших квадратов. Оказывается, что она является оптимальной в широком классе оценок при довольно слабых предположениях.

Теорема 9 (Гаусс, Марков)**.** *Пусть компоненты вектора $\boldsymbol{\varepsilon}$ независимы и $\mathbb{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\intercal = \sigma^2 I_n$. Тогда оценка $\widehat{\boldsymbol{\theta}}_{LSE} = (\Psi\Psi^\intercal)^{-1}\Psi\mathbf{Y}$ является оптимальной среди всех линейных несмещенных оценок параметра $\boldsymbol{\theta}$. Более того, если $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$, то оценка $\widehat{\boldsymbol{\theta}}_{LSE}$ является оптимальной среди всех несмещенных оценок.*

Доказательство. 1. Рассмотрим произвольную несмещенную линейную оценку $\widetilde{\boldsymbol{\theta}} = A\mathbf{Y}$. Из условия несмещенности следует, что

$$\boldsymbol{\theta} = \mathbb{E}\widetilde{\boldsymbol{\theta}} = A\Psi^\intercal\boldsymbol{\theta} \quad \text{для всех } \boldsymbol{\theta} \in \mathbb{R}^p$$

Значит, $A\Psi^\intercal = I_p$. Рассмотрим ковариационную матрицу оценки $\widetilde{\boldsymbol{\theta}}$

$$\mathrm{Var}(\widetilde{\boldsymbol{\theta}}) = \mathrm{Var}(\widetilde{\boldsymbol{\theta}} + \widehat{\boldsymbol{\theta}}_{LSE} - \widehat{\boldsymbol{\theta}}_{LSE}) = \mathrm{Var}(\widehat{\boldsymbol{\theta}}) + \mathbb{E}\widehat{\boldsymbol{\theta}}_{LSE}(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{LSE})^\intercal + \mathbb{E}(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{LSE})\widehat{\boldsymbol{\theta}}^\intercal + \mathrm{Var}(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}).$$

Покажем, что $\mathbb{E}\widehat{\boldsymbol{\theta}}_{LSE}(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{LSE})^\intercal = O$, где $O \in \mathbb{R}^{p\times p}$ – матрица с нулевыми компонентами. В самом деле,

$$\begin{aligned}
\mathbb{E}\widehat{\boldsymbol{\theta}}_{LSE}(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{LSE})^\intercal &= \mathbb{E}(\Psi\Psi^\intercal)^{-1}\Psi\mathbf{Y}\mathbf{Y}^\intercal(A - (\Psi\Psi^\intercal)^{-1}\Psi)^\intercal \\
&= (\Psi\Psi^\intercal)^{-1}\Psi\mathbb{E}(\Psi^\intercal\boldsymbol{\theta} + \boldsymbol{\varepsilon})(\Psi^\intercal\boldsymbol{\theta} + \boldsymbol{\varepsilon})^\intercal(A - (\Psi\Psi^\intercal)^{-1}\Psi)^\intercal \\
&= \boldsymbol{\theta}(\boldsymbol{\theta})^\intercal\Psi(A - (\Psi\Psi^\intercal)^{-1}\Psi)^\intercal + (\Psi\Psi^\intercal)^{-1}\Psi\mathbb{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\intercal(A - (\Psi\Psi^\intercal)^{-1}\Psi)^\intercal \\
&= O + \sigma^2(\Psi\Psi^\intercal)^{-1}\Psi(A - (\Psi\Psi^\intercal)^{-1}\Psi)^\intercal = O.
\end{aligned}$$

Тогда

$$\mathrm{Var}(\widetilde{\boldsymbol{\theta}}) = \mathrm{Var}(\widehat{\boldsymbol{\theta}}_{LSE}) + \mathrm{Var}(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}).$$

Так как $\mathrm{Var}(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{LSE}) \succcurlyeq 0$, то $\mathrm{Var}(\widetilde{\boldsymbol{\theta}}) \succeq \mathrm{Var}(\widehat{\boldsymbol{\theta}}_{LSE})$, причем равенство достигается тогда и только тогда, когда $\widetilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_{LSE}$ с вероятностью 1. Из положительной полуопределенности матрицы $\mathrm{Var}(\widetilde{\boldsymbol{\theta}}) - \mathrm{Var}(\widehat{\boldsymbol{\theta}})$ следует, что $\mathrm{Tr}\big(\mathrm{Var}(\widetilde{\boldsymbol{\theta}})\big) \geqslant \mathrm{Tr}\big(\mathrm{Var}(\widehat{\boldsymbol{\theta}})\big)$. Используя несмещенность $\widetilde{\boldsymbol{\theta}}$ и теорему 1, получаем, что среднеквадратичная ошибка

$$\mathbb{E}\Big\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\Big\|^2 = \mathrm{Tr}\Big(\mathrm{Var}(\widetilde{\boldsymbol{\theta}})\Big).$$

Аналогично получаем, что среднеквадратичная ошибка

$$\mathbb{E}\Big\|\widehat{\boldsymbol{\theta}}_{LSE} - \boldsymbol{\theta}\Big\|^2 = \mathrm{Tr}\Big(\mathrm{Var}(\widehat{\boldsymbol{\theta}}_{LSE})\Big),$$

а значит,

$$\mathbb{E}\Big\|\widehat{\boldsymbol{\theta}}_{LSE} - \boldsymbol{\theta}\Big\|^2 = \mathrm{Tr}\Big[\mathrm{Var}(\widehat{\boldsymbol{\theta}}_{LSE})\Big] \geqslant \mathrm{Tr}\Big[\mathrm{Var}(\widetilde{\boldsymbol{\theta}}_{LSE})\Big] = \mathbb{E}\Big\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\Big\|^2,$$

что означает оптимальность оценки метода наименьших квадратов в классе всех линейных несмещенных оценок параметра $\boldsymbol{\theta}$. Заметим, что при доказательстве первого утверждения теоремы не использовалась гауссовость вектора ошибок $\boldsymbol{\varepsilon}$.

2. Теперь докажем вторую часть теоремы. Пусть $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, тогда $Y_i - \boldsymbol{\psi}(X_i)^\intercal\boldsymbol{\theta} \sim \mathcal{N}(0, \sigma^2)$, $i \in \{1, \ldots, n\}$. Напомним, что логарифм правдоподобия наблюдений $L(\mathbf{Y}, \boldsymbol{\theta})$ можно записать в виде (см. пример 1)

$$\ln p_{\boldsymbol{\theta}}(\mathbf{Y}) = -\frac{1}{2\sigma^2}\|\mathbf{Y} - \Psi^\intercal\boldsymbol{\theta}\|^2 - \frac{n}{2}\log\big(2\pi\sigma^2\big).$$

Найдем информационную матрицу Фишера:

$$I(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}}\left(\frac{\partial^2 \ln p_{\boldsymbol{\theta}}(\mathbf{Y})}{\partial \boldsymbol{\theta}^2}\right) = \frac{1}{\sigma^2}\,\Psi\Psi^{\mathsf{T}}.$$

С другой стороны,

$$\mathrm{Var}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_{LSE}) = \sigma^2(\Psi\Psi^{\mathsf{T}})^{-1}.$$

Таким образом, $\mathrm{Var}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_{LSE}) = I(\boldsymbol{\theta})^{-1}$.

То есть для оценки $\widehat{\boldsymbol{\theta}}_{LSE}$ достигается равенство в неравенстве Рао-Крамера, из чего следует ее эффективность, а значит и оптимальность.

□

7 Метод максимального правдоподобия. Состоятельность и асимптотическая нормальность оценок максимального правдоподобия.

7.1 Метод максимального правдоподобия.

Метод максимального правдоподобия (см. определение 20) – один из ключевых результатов в статистике.

Определение 25. *Оценкой максимального правдоподобия* (ОМП) параметра θ называется статистика

$$\widehat{\theta}_{MLE}(\boldsymbol{X}) = \arg\max_{\theta\in\Theta} \mathcal{L}(\theta).$$

Данное определение уже накладывает несколько ограничений: как минимум, то, что максимум существует и он единственен. Но в дальнейшем будем считать, что они выполнены.

Интуиция метода проста: вспомним, что правдоподобие показывает насколько вероятно данное значение параметра для конкретной выборки, то есть мы берем ту оценку, при которой данная выборка $\mathbf{X}$ наиболее вероятна.

Замечание 8. Обычно удобнее работать с логарифмом правдоподобия, а поскольку логарифм - монотонная функция, то часто оценкой максимального правдоподобия называют

$$\widehat{\theta}_{MLE}(\boldsymbol{X}) = \arg\max_{\theta\in\Theta} \ln\mathcal{L}(\theta). \tag{17}$$

Более того, логарифм естественно возникает из минимизации дивергенции Кульбака-Лейблера (см. ниже).

Сперва рассмотрим пример нахождения оценки максимального правдоподобия.

Пример 9. Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка из нормального распределения $\mathcal{N}(\mu, \sigma^2)$ с неизвестными математическим ожиданием и дисперсией. Найдите оценку максимального правдоподобия параметра $\theta = (\mu, \sigma^2)$.

Решение. По определению функции правдоподобия

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} p_\theta(X_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(X_i-\mu)^2}{2\sigma^2}\right\} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2\right\}.$$

Нужно максимизировать её одновременно по μ и по σ^2, но работать с экспонентой не очень удобно, поэтому прологарифмируем

$$\ln\mathcal{L}(\theta) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2.$$

Эта функция дифференцируема, поэтому можно достаточно легко найти всех претендентов на точки экстремума, приравняв градиент к нулю:

$$0 = \frac{\partial}{\partial\mu}\ln\mathcal{L}(\theta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu),$$

$$0 = \frac{\partial}{\partial\sigma^2}\ln\mathcal{L}(\theta) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i-\mu)^2.$$

Отсюда получаем, что претендентами на точки экстремума будут

$$\widehat{\mu}(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{\boldsymbol{X}},$$
$$\widehat{\sigma}^2(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = S^2.$$

Получили уже знакомые выборочное среднее и выборочную дисперсию. Далее необходимо проверить, что $\widehat{\theta}(\boldsymbol{X}) = (\overline{\boldsymbol{X}}, S^2)$ действительно максимум. Для этого нужно проверить, что гессиан в точке $(\overline{\boldsymbol{X}}, S^2)$ отрицательно определён. □

7.2 Дивергенция Кульбака-Лейблера.

Поймем откуда возникает метод максимизации лог-правдоподобия (17). Пусть есть выборка $\mathbf{X}$ из неизвестного распределения $\mathbb{P}_\theta$. Наша цель, найти такую оценку $\widehat{\theta}$, чтобы распределение $\mathbb{P}_{\widehat{\theta}}$ было наиболее близко к истинному распределению $\mathbb{P}_\theta$. Но как сравнивать вероятностные меры? Очевидно обычное евклидово расстояние здесь не подходит. Вместо него рассмотрим дивергенцию Кульбака-Лейблера, обобщенное расстояние (не удовлетворяет неравенству треугольника и симметрии).

Определение 26 (Дивергенция Кульбака-Лейблера)**.** Пусть $\mathbb{P}_\theta$ и $\mathbb{P}_{\widehat{\theta}}$ – две абсолютно непрерывные вероятностные меры. Тогда дивергенции Кульбака-Лейблера (KL-дивергенция) определяется как

$$\mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_{\widehat{\theta}}) = \int_{-\infty}^{\infty} p_\theta(x)\log\left(\frac{p_\theta(x)}{p_{\widehat{\theta}}(x)}\right)\mathrm{d}x = \mathbb{E}_\theta \log\left(\frac{p_\theta(x)}{p_{\widehat{\theta}}(x)}\right),$$

где $p_\theta(x)$ и $p_{\widehat{\theta}}(x)$ – это плотности распределений $\mathbb{P}_\theta$ и $\mathbb{P}_{\widehat{\theta}}$ соответственно.

Дивергенция Кульбака-Лейблера показывает величину потерь информации при замене истинной вероятностной меры $\mathbb{P}_\theta$ на ее апрексимацию $\mathbb{P}_{\widehat{\theta}}$. Для дискретных распределений дивергенции Кульбака-Лейблера представляется через сумму.

Лемма 11 (Свойства KL-дивергенции)**.** *1.* $\mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_{\widehat{\theta}}) \geqslant 0$

2. $\mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_{\widehat{\theta}}) \iff \mathbb{P}_\theta = \mathbb{P}_{\widehat{\theta}}$ *почти наверное.*

Доказательство. 1. Воспользуемся интегральным неравенством Йенсена для вогнутой функции $\varphi(y) = \log y$:

$$\varphi(\mathbb{E}Y) \geqslant \mathbb{E}\varphi(Y).$$

Данное неравенство обращается в равенство тогда и только тогда, когда $\varphi(y)$ и/или Y – это константа почти наверное. В нашем случае равенство будет достигаться только тогда, когда Y – это константа почти наверное, так как $\varphi(y) = \log y$ – вогнутая. Воспользуемся неравенством Йенсена:

$$-\mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_{\widehat{\theta}}) = \int_{-\infty}^{\infty} p_\theta(x)\log\left(\frac{p_{\widehat{\theta}}(x)}{p_\theta(x)}\right)\mathrm{d}x = \mathbb{E}_\theta\log\left(\frac{p_{\widehat{\theta}}(x)}{p_\theta(x)}\right)$$
$$\leqslant \log\left(\mathbb{E}_\theta\frac{p_{\widehat{\theta}}(x)}{p_\theta(x)}\right) = \log\left(\int_{-\infty}^{\infty}\frac{p_{\widehat{\theta}}(x)}{p_\theta(x)}p_\theta(x)\mathrm{d}x\right) = \log\left(\int_{-\infty}^{\infty} p_{\widehat{\theta}}(x)\mathrm{d}x\right) = \log 1 = 0.$$

Значит, $\mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_{\widehat{\theta}}) \geqslant 0$.

2. В обратную сторону очевидно.

□

Изучим более подробно дивергенцию Кульбака-Лейблера:

$$\mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_{\widehat{\theta}}) = \int_{-\infty}^{\infty} p_\theta(x) \log\left(\frac{p_\theta(x)}{p_{\widehat{\theta}}(x)}\right) \mathrm{d}x = \mathbb{E}_\theta \log\left(\frac{p_\theta(x)}{p_{\widehat{\theta}}(x)}\right) = \mathbb{E}_\theta \log p_\theta(x) - \mathbb{E}_\theta \log p_{\widehat{\theta}}(x)$$
$$= \text{‘const’} - \mathbb{E}_\theta \log p_{\widehat{\theta}}(x).$$

Поскольку истинный параметр θ неизвестен, то посчитать дивергенцию Кульбака Лейблера $\mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_{\widehat{\theta}})$ не представляется возможным. Оценим ее с помощью закона больших чисел

$$\widehat{KL}(\mathbb{P}_\theta, \mathbb{P}_{\widehat{\theta}}) = \text{‘const’} - \frac{1}{n}\sum_{i=1}^{n} \log p_{\widehat{\theta}}(X_i). \tag{18}$$

Тогда наилучшей оценкой, доставляющей минимум (18) будет

$$\widehat{\theta}(\mathbf{X}) = \arg\min_{\widehat{\theta}} \widehat{KL}(\mathbb{P}_\theta, \mathbb{P}_{\widehat{\theta}}) = \arg\min_{\widehat{\theta}} -\frac{1}{n}\sum_{i=1}^{n} \log p_{\widehat{\theta}}(X_i) = \arg\max_{\widehat{\theta}} \frac{1}{n}\sum_{i=1}^{n} \log p_{\widehat{\theta}}(X_i).$$

Заметим, что полученная оценка совпадает с оценкой максимального правдоподобия (17).

7.3 Асимптотические свойства оценок максимального правдоподобия.

Теорема 10. *Пусть дана выборка $\mathbf{X} = (X_1, \ldots, X_n)$ из регулярного семейства (см. условия регулярности 1) с обобщенной плотностью $p_\theta(x)$. Пусть дополнительно выполнены следующие условия:*

1. *функция правдоподобия $\mathcal{L}(\theta)$ имеет лишь один максимум, и притом достижимый;*
2. *$p_\theta(x)$ трижды непрерывно дифференцируема по θ и при этом существует независящая от θ функция $M(x)$ такая, что для всех $\theta \in \Theta$*

$$\left|\frac{\partial^3}{\partial\theta^3} \ln p_\theta(x)\right| \leqslant M(x), \qquad \text{причём} \qquad \mathbb{E}_\theta M(X_1) < +\infty.$$

Тогда оценка $\widehat{\theta}_{MLE}$, найденная как решение уравнения правдоподобия

$$\mathcal{L}'(\theta)\Big|_{\theta=\widehat{\theta}_{MLE}} = 0$$

является

1. *состоятельной;*
2. *асимптотически нормальной с асимптотической дисперсией $i^{-1}(\theta)$: для всех $\theta \in \Theta$*

$$\sqrt{n}(\widehat{\theta}_{MLE} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{i(\theta)}\right).$$

3. *асимптотически эффективной;*
4. *асимптотически оптимальной;*

5. *инвариантной: если $\widehat{\theta}_{MLE}$ – ОМП для θ и отображение $\theta \mapsto g(\theta)$ – биективно, то $g(\widehat{\theta}_{MLE})$ – ОМП для $g(\theta)$;*

Асимптотическая эффективность оценки максимального правдопободия следует из того факта, что $\mathrm{Var}(\widehat{\theta}_{MLE}) \to i(\theta)$.

Следствие 2. *Пусть выполняются условия теоремы 10. Тогда если существует эффективная оценка для θ, то она совпадает с оценкой максимального правдоподобия $\widehat{\theta}_{MLE}$.*

Доказательство. Пусть $\widehat{\theta}$ – эффективная оценка, тогда она удовлетворяет критерию эффективности (следствие 1):

$$\widehat{\theta} - \theta = c(\theta)U_\theta(\mathbf{X}) = c(\theta)\frac{\partial}{\partial\theta}\ln\mathcal{L}(\theta).$$

В точке $\theta = \widehat{\theta}_{MLE}$:

$$\widehat{\theta} - \widehat{\theta}_{MLE} = c(\theta)\frac{\partial}{\partial\theta}\ln\mathcal{L}(\theta)\bigg|_{\theta=\widehat{\theta}_{MLE}} = 0$$

Значит, $\widehat{\theta} = \widehat{\theta}_{MLE}$. □

8 Доверительные интервалы: точные и асимптотические.

До сих пор мы имели дело только с точечными оценками. Теперь мы переходим к интервальным. Для простоты будем считать, что мы рассматриваем параметрические семейства с одномерным параметром. Под термином интервал в дальнейшем мы будем подразумевать некоторое множество, не обязательно интервал в геометрическом смысле.

Определение 27 (Доверительный интервал)**.** Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ – выборка из распределения $\mathbb{P} \in \{\mathbb{P}_\theta \colon \theta \in \Theta\}$, где $\Theta \subseteq \mathbb{R}$. Будем называть пару статистик $(T_1(\boldsymbol{X}), T_2(\boldsymbol{X}))$ *доверительным интервалом* для параметра θ с уровнем доверия $1 - \alpha \in (0, 1)$, если

- для всех $\theta \in \Theta$ выполнено $T_1(\boldsymbol{X}) \leqslant T_2(\boldsymbol{X})$ с вероятностью 1;
- для всех $\theta \in \Theta$ выполнено
$$\mathbb{P}_\theta(T_1(\boldsymbol{X}) < \theta < T_2(\boldsymbol{X})) \geqslant 1 - \alpha. \tag{19}$$

Если в неравенстве (19) всегда достигается равенство (т.е. $\mathbb{P}_\theta(T_1(\boldsymbol{X}) < \theta < T_2(\boldsymbol{X})) = 1 - \alpha$ для всех $\theta \in \Theta$), то доверительный интервал называют *точным*.

Замечание 9. Подчеркнем, что неизвестный параметр θ не является случайной величиной, а $T_1(\boldsymbol{X}), T_2(\boldsymbol{X})$ – это случайные величины, не зависящие от θ.

При таком определении точность оценивания есть длина интервала $(T_1(\boldsymbol{X}), T_2(\boldsymbol{X}))$, а наша уверенность в оценивании — вероятность попадания параметра в этот интервал.

Определение 28. Подмножество $T(\boldsymbol{X}) \subseteq \Theta$ называется доверительной областью для θ с уровнем доверия $1 - \alpha$, если для всех $\theta \in \Theta$ выполнено

$$\mathbb{P}_\theta(\theta \in T(\boldsymbol{X})) \geqslant 1 - \alpha$$

8.1 Метод центральной статистики

Одним из основных методов поиска доверительного интервала является метод *центральной статистики*. Данный метод применим в случае выборки $\boldsymbol{X}$ из абсолютно непрерывного распределения $\mathbb{P}_\theta$.

Определение 29 (Центральная статистика)**.** Функция от выборки $G(\boldsymbol{X}, \theta)$, которая может зависеть от параметра θ, называется *центральной статистикой*, если

- $G(\boldsymbol{X}, \theta)$ непрерывна и монотонна по θ при любом фиксированном $\boldsymbol{X}$;
- $\mathbb{P}_\theta(G(\boldsymbol{X}, \theta) \leqslant t) = F_G(t)$ – непрерывная функция по t, не зависящая от θ (т.е. распределение $G(\boldsymbol{X}, \theta)$ не зависит от θ).

Таким образом, если $G(\boldsymbol{X}, \theta)$ – центральная статистика, то множество $\{\theta : t_1 < G(\boldsymbol{X}, \theta) < t_2\}$ является интервалом. Заметьте, что формально центральная статистика не является статистикой с точки зрения определения 5, так как $G(\boldsymbol{X}, \theta)$ зависит от неизвестного параметра θ.

Если для конкретной задачи нам удалось построить центральную статистику $G(\boldsymbol{X}, \theta)$, то доверительные интервалы для параметра θ строятся следующим образом:

1. Фиксируются $t_1 < t_2 \in \mathbb{R}$ такие, что выполнено

$$\mathbb{P}_\theta(t_1 \leqslant G(\boldsymbol{X}, \theta) \leqslant t_2) = F_G(t_2) - F_G(t_1) = 1 - \alpha\,.$$

Точки t_1 и t_2 есть квантили в распределении F_G. Дадим определение $\gamma-$квантиля.

Определение 30 ($\gamma-$квантиль)**.** x_γ называется $\gamma-$квантилем некоторого распределения с функцией распределения F, если

$$F(x_\gamma) = \gamma$$

Заметим, что выбор точек t_1 и t_2 определен неоднозначно. Возьмем некоторую $0 < \gamma < \alpha$ и $t_1 = x_\gamma$, тогда $t_2 = x_{1-\alpha+\gamma}$, где x_γ и $x_{1-\alpha+\gamma}$ – это $\gamma-$квантиль и $1-\alpha+\gamma-$квантиль соответственно. Квантили могут быть найдены из соответствующих таблиц с квантилями или численно, например в python'е.

2. Из условий

$$\begin{cases} G(\mathbf{X}, \theta) = t_2 \\ G(\mathbf{X}, \theta) = t_1 \end{cases}$$

можно найти статистики $T_1(\mathbf{X}, t_1)$ и $T_2(\mathbf{X}, t_2)$, такие что

$$\begin{cases} G(\mathbf{X}, T_2(\mathbf{X}, t_2)) = t_2 \\ G(\mathbf{X}, T_1(\mathbf{X}, t_1)) = t_1\,. \end{cases}$$

Из неравенств выше и определения центральной статистики следует, что

$$\mathbb{P}_\theta(t_1 \leqslant G(\boldsymbol{X}, \theta) \leqslant t_2) = \mathbb{P}_\theta(T_1(\boldsymbol{X}, t_1) \leqslant \theta \leqslant T_2(\boldsymbol{X}, t_2)) = 1 - \alpha\,.$$

Таким образом, мы получили $1-\alpha$–доверительный интервал $(T_1(\boldsymbol{X}, t_1), T_2(\boldsymbol{X}, t_2))$ для параметра $\theta \in \Theta$.

Для разрешения проблемы неоднозначности выбора границ t_1 и t_2, можно строить *центральные* доверительные интервалы. Но наилучшим же интервалом будет тот, у которого наименьшая длина, то есть наикратчайший интервал.

Определение 31 (Центральный интервал)**.** Доверительный интервал называют *центральным*, если

$$\mathbb{P}_\theta(G(\boldsymbol{X}, \theta) \leqslant t_1) = \mathbb{P}_\theta(G(\boldsymbol{X}, \theta) \geqslant t_2) = \frac{\alpha}{2}\,.$$

Для построения центрального доверительного интервала квантили выбираются следующим образом

$$t_1 = x_{\alpha/2}, \quad t_2 = x_{1-\alpha/2}.$$

Пример 10. Пусть у нас есть выборка $\boldsymbol{X} = (X_1, \ldots, X_n)$ из распределения $\mathcal{N}(\theta, \sigma^2)$ с известным параметром σ^2. Наша задача состоит в построении доверительного интервала для неизвестного математического ожидания θ. Часто центральная статистика строится на основе точечной оценки неизвестного параметра θ. В нормальной модели $\mathcal{N}(\theta, \sigma^2)$ математическое ожидание можно оценить методом максимального правдоподобия и получить $\widehat{\theta}_{MLE} = \overline{X} \sim \mathcal{N}(\theta, \sigma^2/n)$. Для того чтобы получить распределение, не зависящее от неизвестного параметра, проведем стандартизацию, в результате чего получим центральную статистику:

$$G(\boldsymbol{X}, \theta) = \frac{\sqrt{n}(\overline{X} - \theta)}{\sigma},$$

которая имеет нормальное распределение $\mathcal{N}(0,1)$. $G(\boldsymbol{X},\theta)$ строго монотонна по θ и ее распределение не зависит от θ. Тогда необходимо выбрать t_1 и t_2, такие что

$$\mathbb{P}_\theta\left(t_1 < \frac{\sqrt{n}(\overline{X}-\theta)}{\sigma} < t_2\right) = \mathbb{P}_\theta\left(\overline{X} + \frac{t_1\sigma}{\sqrt{n}} < \theta < \overline{X} + \frac{t_2\sigma}{\sqrt{n}}\right) = 1-\alpha\,.$$

Заметим, что в данном случае наикратчайший интервал будет совпадать с центральным. Для построения центрального доверительного интервала заметим, что распределение $G(\boldsymbol{X},\theta)$ симметрично относительно 0. Тогда достаточно ввести один квантиль

$$t_1 = x_{\alpha/2}, \quad t_2 = x_{1-\alpha/2} = -x_{\alpha/2},$$

где $x_{\alpha/2}$ – $\alpha/2$-квантиль распределения $\mathcal{N}(0,1)$. Искомый доверительный интервал будет иметь вид

$$\left(\overline{X} + \frac{\sigma}{\sqrt{n}}x_{\alpha/2},\ \overline{X} - \frac{\sigma}{\sqrt{n}}x_{\alpha/2}\right).$$

Пример 11. Пусть теперь математическое ожидание в нормальной модели будет известно, а дисперсия σ^2 нет. То есть дана выборка $X_1,\dots,X_n$ из распределения $\mathcal{N}(\mu,\theta^2)$. Наша задача построить центральный доверительный интервал для θ^2. Неизвестная дисперсия может быть оценена с помощью выборочной дисперсии. Воспользуемся этой оценкой, а также информацией о том, что математическое ожидание нам известно и построим центральную статистику

$$G(\boldsymbol{X},\theta) = \frac{1}{\theta^2}\sum_{i=1}^n (X_i-\mu)^2,$$

имеющую распределение хи-квадрат $\chi^2(n)$. Данная центральная статистика является строго монотонной функцией от θ^2, а также ее распределение не зависит от θ^2. Распределение $\chi^2(n)$ не симметрично, поэтому понадобятся два квантиля

$$t_1 = x_{\alpha/2}, \quad t_2 = x_{1-\alpha/2}$$

где $x_{\alpha/2}$ и $x_{1-\alpha/2}$ есть $\alpha/2$-квантиль и $1-\alpha/2$-квантиль распределения χ^2 соответственно. Таким образом, получаем

$$\mathbb{P}_\theta\left(x_{\alpha/2} < \frac{1}{\theta^2}\sum_{i=1}^n (X_i-\mu)^2 < x_{1-\alpha/2}\right) = \mathbb{P}_\theta\left(\frac{\sum_{i=1}^n (X_i-\mu)^2}{x_{1-\alpha/2}} < \theta^2 < \frac{\sum_{i=1}^n (X_i-\mu)^2}{x_{\alpha/2}}\right).$$

То есть искомый доверительный интервал

$$\left(\frac{\sum_{i=1}^n (X_i-\mu)^2}{x_{1-\alpha/2}}, \frac{\sum_{i=1}^n (X_i-\mu)^2}{x_{\alpha/2}}\right)$$

8.2 Асимптотические доверительные интервалы

Не во всех задачах удается легко найти центральную статистику, распределение которой известно точно. Намного проще предъявить центральную статистику, у которой было бы известно лишь предельное распределение. Доверительные интервалы, построенные по предельному распределению, мы будем называть *асимптотическими*. Таким образом, в определении (27) неравенство (19) заменится на

$$\mathbb{P}_\theta(T_1(\boldsymbol{X}) < \theta < T_2(\boldsymbol{X})) \xrightarrow{n\to\infty} 1-\alpha.$$

Пусть размер выборки n достаточно большой, покажем, как на основе асимптотически нормальной оценки $\widehat{\theta}$, то есть такой что

$$\frac{\sqrt{n}(\widehat{\theta}-\theta)}{\sigma(\theta)} \xrightarrow{d} \mathcal{N}(0,1), \quad \text{где } \sigma^2(\theta) \text{ – это асимптотическая дисперсия,}$$

можно строить асимптотические доверительные интервалы. Пусть $\sigma^2(\theta)$ непрерывна при любом $\theta \in \Theta$.

1. Покажем, что следующая величина будет иметь распределение $\mathcal{N}(0,1)$

$$\frac{\sqrt{n}(\widehat{\theta}-\theta)}{\sigma(\widehat{\theta})} = \frac{\sqrt{n}(\widehat{\theta}-\theta)}{\sigma(\theta)} \cdot \frac{\sigma(\theta)}{\sigma(\widehat{\theta})}.$$

2. Из определения асимптотической нормальности следует состоятельность $\widehat{\theta}$. Далее, пользуясь теоремой о наследовании сходимости 22, получаем

$$\sigma(\widehat{\theta}) \xrightarrow{\mathbb{P}} \sigma(\theta)$$

3. По лемме Слуцкого 14 и из определения асимптотической нормальности следует

$$\frac{\sqrt{n}(\widehat{\theta}-\theta)}{\sigma(\widehat{\theta})} \xrightarrow{d} \mathcal{N}(0,1).$$

4. Далее по аналогии с методом центральной статистики строим асимптотический интервал для $\frac{\sqrt{n}(\widehat{\theta}-\theta)}{\sigma(\widehat{\theta})}$

$$\mathbb{P}_\theta\left(x_{\alpha/2} < \frac{\sqrt{n}(\widehat{\theta}-\theta)}{\sigma(\widehat{\theta})} < -x_{\alpha/2}\right) \xrightarrow{n\to\infty} 1-\alpha,$$

где $x_{\alpha/2}$ есть $\alpha/2$-квантиль распределения $\mathcal{N}(0,1)$. Значит асимптотически доверительный интервал для θ есть

$$\left(\widehat{\theta} + \frac{\sigma(\widehat{\theta})}{\sqrt{n}} x_{\alpha/2},\ \widehat{\theta} - \frac{\sigma(\widehat{\theta})}{\sqrt{n}} x_{\alpha/2}\right)$$

Замечание 10. Если модель удовлетворяет условиям регулярности, то в качестве асимптотически нормальной оценки можно взять оценку максимального правдоподобия $\widehat{\theta}_{MLE}$. Тогда асимптотическая дисперсия может быть вычислена как $\sigma^2(\theta) = \frac{1}{i_1(\theta)}$ (см. теорему 7.3). Более того, так как $\widehat{\theta}_{MLE}$ обладает наименьшей асимптотической дисперсией, то и доверительный интервал, построенный с ее использованием, будет наикратчайший.

Пример 12. Пусть дана выборка $(X_1, \dots, X_n)$ из распределения с Бернулли $\mathcal{B}\text{er}(\theta)$, $\theta \in (0,1)$. Оценкой максимального правдоподобия $\widehat{\theta}_{MLE} = \overline{X}$. В данном примере можно не искать асимптотическую дисперсию через информацию Фишера, так как она будет совпадать с дисперсией в распределении $\mathcal{B}\text{er}(\theta)$ по ЦПТ 21

$$\frac{\sqrt{n}(\overline{X}-\theta)}{\sqrt{\theta(1-\theta)}} \xrightarrow{d} \mathcal{N}(0,1).$$

Теперь рассмотрим

$$\frac{\sqrt{n}(\overline{X}-\theta)}{\sqrt{\overline{X}(1-\overline{X})}} = \frac{\sqrt{n}(\overline{X}-\theta)}{\sqrt{\theta(1-\theta)}} \cdot \frac{\sqrt{\theta(1-\theta)}}{\sqrt{\overline{X}(1-\overline{X})}}$$

Пользуясь теоремой о наследовании сходимости 22 и состоятельностью оценки $\widehat{\theta}_{MLE} = \overline{X}$, получаем

$$\sqrt{\overline{X}(1-\overline{X})} \xrightarrow{\mathbb{P}} \sqrt{\theta(1-\theta)}.$$

Тогда по лемме Слуцкого 14 произведение сойдется к $\mathcal{N}(0,1)$. Значит

$$\mathbb{P}_\theta\left(x_{\alpha/2} < \frac{\sqrt{n}(\overline{X}-\theta)}{\sqrt{\overline{X}(1-\overline{X})}} < -x_{\alpha/2}\right) \xrightarrow{n\to\infty} 1-\alpha,$$

где $x_{\alpha/2}$ есть $\alpha/2$-квантиль распределения $\mathcal{N}(0,1)$. Далее осталось выразить θ и получить асимптотически доверительный интервал для θ

$$\left(\overline{X} + \sqrt{\frac{\overline{X}(1-\overline{X})}{n}}x_{\alpha/2},\ \overline{X} - \sqrt{\frac{\overline{X}(1-\overline{X})}{n}}x_{\alpha/2}\right)$$

9 Проверка статистических гипотез.

До сих пор нашей основной задачей было оценивание неизвестных параметров распределений. Разумеется, статистика не исчерпывается подобными задачами. Часто на практике возникает необходимость проверки тех или иных утверждений о виде и свойствах распределения наблюдаемых случайных величин, такие утверждения называются *статистическими гипотезами*. В общем случае задача проверки статистических гипотез формулируется следующим образом. Пусть дана выборка $\boldsymbol{X} = (X_1, \ldots, X_n)$ из некоторого распределения $\mathbb{P}$. Известно, что данное распределение принадлежит некоторому классу распределений $\mathcal{P}$, $\mathbb{P}_{\boldsymbol{X}} \in \mathcal{P}$. Например, в задаче параметрического оценивания $\mathcal{P} = \{\mathbb{P}_\theta\}_{\theta \in \Theta}$, однако в общем случае $\mathcal{P}$ может быть устроено и значительно сложнее. В классе распределений $\mathcal{P}$ выделяется некоторое подмножество $\mathcal{P}_0 \subset \mathcal{P}$, и требуется по выборке $\boldsymbol{X}$ проверить, справедливо ли утверждение

$$H_0 : \mathbb{P} \in \mathcal{P}_0 \,.$$

В этом случае утверждение H_0 называется *основной (нулевой) гипотезой*. При этом противоположное утверждение

$$H_1 : \mathbb{P} \in \mathcal{P} \setminus \mathcal{P}_0$$

называется *альтернативной* гипотезой. В таком случае также говорят, что проверяется гипотеза H_0 против альтернативы H_1. Если подмножество $\mathcal{P}_0$ состоит из одного элемента, то H_0 называется *простой* гипотезой, в противном случае – *сложной* гипотезой. При проверки гипотезы на основании наблюдения выборки $\boldsymbol{X}$ мы либо отклоняем гипотезу H_0 как ложную, т.е. принимаем альтернативную гипотезу H_1, либо не отклоняем H_0. Правило, в соответствии с которым мы делаем выбор в пользу одной из гипотез называется *статистическим критерием*

$$\delta : \mathcal{X} \to \{H_0, H_1\}.$$

Представим выборочное пространство $\mathcal{X}$ как объединение двух непересекающихся множеств $\mathcal{X}_0$ и $\mathcal{X}_1$, где $\mathcal{X}_1$ – область отклонения, а $\mathcal{X}_0$ – область не отклонения. То есть

- $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1$
- $\mathcal{X}_0 \cap \mathcal{X}_1 = \emptyset$

Множество $\mathcal{X}_1$ мы будем называть *критической областью*. Таким образом,

$$H_0 \text{ отвергается} \quad \Longleftrightarrow \quad \boldsymbol{x} \in \mathcal{X}_1,$$

то есть мы отвергаем основную гипотезу тогда и только тогда, когда наблюдаемые случайные величины попали в критическую область. При построении статистических критериев используется общий подход: в качестве критической области $\mathcal{X}_1$ выбирается такое подмножество выборочного пространства, попадание в которое маловероятно при $\mathbb{P} \in \mathcal{P}_0$, то есть если H_0 верна.

В конкретных задачах критическая функция обычно задается с помощью некоторой *тестовой статистики* $T(\mathbf{X})$, которая измеряет отклонение эмпирических данных от гипотетических (соответствующее гипотезе H_0). Распределение $T(\mathbf{X})$ должно быть известно точно или хотя бы приближенно (для больших выборок). В частности, если гипотеза H_0 сложная, то желательно, чтобы распределение $T(\mathbf{X})$ было одним и тем же для всех $\mathbb{P} \in \mathcal{P}_0$. Тогда в терминах тестовой статистики критерий будет иметь вид

$$H_0 \text{ отвергается} \quad \Longleftrightarrow \quad T(\boldsymbol{x}) \in \mathcal{T}_1,$$

где $\mathcal{T}_1 = \{t = T(\mathbf{X}), \boldsymbol{x} \in \mathcal{X}_1\}$ – критическая область для тестовой статистики.

9.1 Ошибки первого и второго рода, уровень значимости.

Вывод, сделанный на основе статистического критерия, не обязательно оказывается правильным. Например, может оказаться, что статистический критерий отверг гипотезу H_0, хотя она была верна. Либо наоборот принял решение в пользу основной гипотезы, в то время как на самом деле имеет место альтернативная. Таким образом, приходим к понятию ошибок первого и второго рода.

Определение 32 (Ошибки 1-го и 2-го рода)**.** Пусть выдвинута основная гипотеза H_0 против альтернативны гипотеза H_1. Рассмотрим статистический критерий δ. Вероятность того, что мы отвергнем верную гипотезу H_0 называется вероятностью *ошибки первого рода*

$$\alpha(\delta) = \mathbb{P}_{H_0}(\delta(\mathbf{X}) = H_1) = \mathbb{P}_{H_0}(\mathbf{X} \in \mathcal{X}_1).$$

Вероятность того, что мы не отвергнем неверную гипотезу H_0 называется вероятностью *ошибки первого рода*

$$\beta(\delta) = \mathbb{P}_{H_1}(\delta(\mathbf{X}) = H_0) = \mathbb{P}_{H_1}(\mathbf{X} \notin \mathcal{X}_1)$$

называется вероятностью *ошибки второго рода.*

Как было сказано ранее, статистический критерий составляется таким образом, чтобы вероятность отвергнуть основную гипотезу при ее справедливости была небольшой. Обычно фиксируется некоторая верхняя граница ошибки первого рода.

Определение 33 (Уровень значимости)**.** Пусть выдвинута гипотеза $H_0 : \mathbb{P} \in \mathcal{P}_0$, и пусть дан статистический критерий δ. Если $\alpha(\delta) \leqslant \alpha_0$ при всех $\mathbb{P} \in \mathcal{P}_0$, то говорят, что критерий δ имеет уровень значимости α_0.

Если гипотеза H_0 простая, то понятия уровня значимости критерия α_0 и ошибки первого рода $\alpha(\delta)$ совпадают. В случае сложной основной гипотезы это не так.

9.2 Параметрические гипотезы.

В ряде задач, например, в задаче параметрического оценивания, класс допустимых распределений $\mathcal{P}$ наблюдаемых случайных величин является параметрическим $\mathcal{P} = \{\mathbb{P}_\theta\}_{\theta\in\Theta}$. Пусть Θ_0 и Θ_1 – непустые непересекающиеся подмножества Θ. Далее будем рассматривать задачу проверки гипотезы вида

$$H_0 : \theta \in \Theta_0$$

против альтернативы

$$H_1 : \theta \in \Theta_1$$

Параметрическая модель позволяет учитывать дополнительные аспекты при построении статистических критериев, например, вычислять функцию мощности.

Определение 34 (Функция мощности)**.** Функция

$$W(\delta, \theta) = \mathbb{P}_\theta(\delta(\mathbf{X}) = H_1) = \mathbb{P}_\theta(\mathbf{X} \in \mathcal{X}_1), \quad \theta \in \Theta$$

называется функцией мощности.

Определение 35 (Мощность)**.** Функция

$$W(\delta,\theta)=\mathbb{P}_\theta(\delta(\mathbf{X})=H_1)=\mathbb{P}_\theta(\mathbf{X}\in\mathcal{X}_1),\quad \theta\in\Theta_1$$

называется мощностью.

Заметим, что если $\theta\in\Theta_0$, то функция мощности совпадает с ошибкой первого рода; если $\theta\in\Theta_1$ – с мощностью критерия, равной $1-\beta(\delta)$, где $\beta(\delta)$ – ошибка второго рода. Функция мощности позволяет сравнивать критерии одинакового уровня значимости: из двух критериев с одинаковым уровнем значимости лучше тот, мощность которого выше. Если альтернатива сложная (Θ_1 состоит больше, чем из одной точки), то как и при сравнении оценок возникает проблема сравнения двух функций $W(\theta,\delta_1)$ и $W(\theta,\delta_2)$. В случае если и сама гипотеза, и альтернатива простые, то такая проблема отсутствует можно и построить наиболее мощный критерий. В общем случае задача построения наилучшего критерия формулируется как задача оптимизации

$$\max_\delta W(\delta,\theta),\quad \theta\in\Theta_1,\quad \text{при условии } W(\delta,\theta)\leqslant\alpha_0,\quad \theta\in\Theta_0.$$

Также для нас будет полезно следующее свойство критериев.

Определение 36 (Несмещенность критерия)**.** Статистический критерий δ уровня значимости α_0, то есть $W(\delta,\theta)\leqslant\alpha_0,\theta\in\Theta_0$, называется несмещенным, если

$$W(\delta,\theta)\geqslant\alpha_0,\quad \theta\in\Theta_1.$$

Несмещенность означает, что попадание в критическое множество $\mathcal{X}_1$ (а следовательно и отвержение гипотезы H_0) при справедливости любой из альтернатив не менее вероятно, чем попадание в него при справедливости H_0, то есть правильное отвержение имеет не меньшую вероятность, чем неправильное.

Определение 37 (Состоятельность критерия)**.** Пусть $\mathbf{X}=(X_1,\dots,X_n)$ – выборка объема n. Статистический критерий δ уровня значимости α_0 называется состоятельным, если

$$W(\delta,\theta)\xrightarrow{n\to\infty}1,\quad \theta\in\Theta_1.$$

Состоятельность означает, что если истинной является некоторая альтернатива, то при большом числе наблюдений мы будем попадать в критическую область с вероятностью, близкой к 1, то есть отклоняя нулевую гипотезу, мы будем принимать правильное решение.

Пример 13. Пусть дана выборка $\mathbf{X}=(X_1,\dots,X_n)$ из нормального распределения $\mathcal{N}(\theta,\sigma^2)$. Построить критерий для проверки гипотезы $H_0:\theta=\theta_0$ против альтернативы $H_1:\theta=\theta_1$ ($\theta_1>\theta_0$) на уровне значимости $\alpha\in(0,1)$. Найти мощность этого критерия.

Доказательство. В данной задаче логика подсказывает выбрать выборочное среднее в качестве тестовой статистики:

$$T(\mathbf{X})=\overline{X}\sim\mathcal{N}(\theta,\sigma^2/n).$$

Тогда критическое множество может быть задано

$$\mathcal{X}_1=\{\boldsymbol{x}\in X:\overline{X}\geqslant t_\alpha\},$$

где порог t_α выбирается из уровня значимости критерия:

$$\mathbb{P}_{\theta_0}(\overline{X}>t_\alpha)=\alpha.$$

При справедливости H_0 $\overline{X} \sim \mathcal{N}(\theta_0, \sigma^2/n)$. Тогда

$$\alpha = \mathbb{P}_{\theta_0}\left(\frac{\sqrt{n}(\overline{X} - \theta_0)}{\sigma} > \frac{\sqrt{n}(t_\alpha - \theta_0)}{\sigma}\right) = 1 - \Phi\left(\frac{\sqrt{n}(t_\alpha - \theta_0)}{\sigma}\right),$$

Φ – функция распределения стандартного нормального распределения. Значит

$$\frac{\sqrt{n}(t_\alpha - \theta_0)}{\sigma} = z_{1-\alpha},$$

где $z_{1-\alpha}$ — квантиль стандартного нормального распределения. Отсюда:

$$t_\alpha = \theta_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}.$$

Заметим, что критическая граница не зависит от альтернативы, а зависит лишь от знака $\theta_1 > \theta_0$. Значит, для задачи с односторонней альтернативой $\theta > \theta_0$ критерий будет точно такой же.

Рассчитаем мощность критерия, используя распределение при альтернативе $\overline{X} \sim \mathcal{N}(\theta_1, \sigma^2/n)$

$$W(\theta_1) = \mathbb{P}_{\theta_1}(\overline{X} > t_\alpha) = \mathbb{P}_{\theta_1}\left(\frac{\sqrt{n}(\overline{X} - \theta_1)}{\sigma} > \frac{\sqrt{n}(t_\alpha - \theta_1}{\sigma}\right) = 1 - \Phi\left(\frac{\sqrt{n}(t_\alpha - \theta_1)}{\sigma}\right).$$

Подставляя $t_\alpha = \theta_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$, получаем:

$$W(\theta_1) = 1 - \Phi\left(\frac{\sqrt{n}(\theta_0 - \theta_1)}{\sigma} + z_{1-\alpha}\right).$$

□

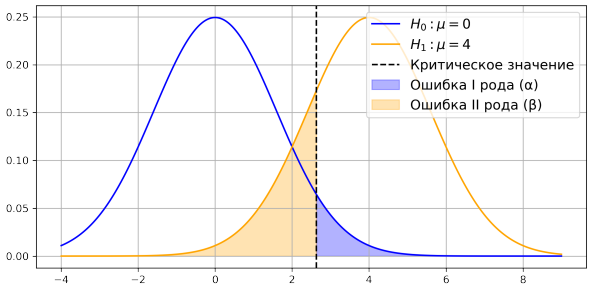

Рис. 2: Ошибки 1-го и 2-го рода

10 Критерий Неймана-Пирсона для проверки параметрических гипотез.

10.1 Простые гипотезы

Рассмотрим параметрическую модель

$$\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}.$$

Пусть мы имеем дело с простой гипотезой и простой альтернативой, то есть когда $\Theta_0 = \{\theta_0\}$, а $\Theta_1 = \{\theta_1\}$

$$\begin{aligned} &H_0 : \theta = \theta_0; \\ &H_1 : \theta = \theta_1 \ (\theta_1 > \theta_0). \end{aligned} \tag{20}$$

Оказывается, что для данной задачи при любом $\alpha \in (0,1)$ можно построить наиболее мощный критерий уровня значимости α.

Пусть дана простая выборка $\mathbf{X} = (X_1, \ldots, X_n)$ из распределения $\mathbb{P}$, и рассмотрим следующую задачу проверки простой гипотезы против простой альтернативы

$$H_0 : \mathbb{P} = \mathbb{P}_0; \tag{21}$$

$$H_1 : \mathbb{P} = \mathbb{P}_1. \tag{22}$$

Данная задача также относится к параметрическим, так как семейство $\{\mathbb{P}_0, \mathbb{P}_1\}$, состоящее всего из двух элементов, можно представить в виде

$$\{\mathbb{P}_0, \mathbb{P}_1\} = \{\mathbb{P}_\theta = \theta\mathbb{P}_1 + (1-\theta)\mathbb{P}_0, \theta \in \{0,1\}\}.$$

Тогда задача (21) сведется к параметрической

$$\begin{aligned} &H_0 : \theta = 0; \\ &H_1 : \theta = 1. \end{aligned}$$

Непрерывный случай. Рассмотрим непрерывный случай, когда оба распределения $\mathbb{P}_{\theta_0}, \mathbb{P}_{\theta_1}$ – абсолютно непрерывны с соответствующими плотностями p_{θ_0} и p_{θ_1}.

Определение 38 (Статистика отношения правдоподобий)**.** Статистикой отношения правдоподобий называется статистика

$$\Lambda(\boldsymbol{x}) = \frac{p_{\theta_1}(\boldsymbol{x})}{p_{\theta_0}(\boldsymbol{x})} = \frac{\prod_{i=1}^n p_{\theta_1}(x_i)}{\prod_{i=1}^n p_{\theta_0}(x_i)}.$$

Большие значения этой статистики будут свидетельствовать о том, что значение θ_1 более вероятно, чем значение θ_0. Критерий, построенный на основе статистики отношения правдоподобия, будем называть критерием Неймана-Пирсона. Рассмотрим ошибку 1-го рода и введем функцию

$$\varphi(t) = \mathbb{P}_{\theta_0}(\Lambda(\mathbf{X}) \geqslant t) \tag{23}$$

и предположим, что она непрерывна.

Теорема 11 (Нейман, Пирсон)**.** *Пусть дана выборка $\mathbf{X} = (X_1, \ldots, X_n)$ из распределения $\mathbb{P}_\theta, \theta = \{\theta_0, \theta_1\}$. Пусть $p_{\theta_0}, p_{\theta_1} > 0$. Допустим, что функция φ, определенная в (23) непрерывна. Зафиксируем произвольное $\alpha \in (0,1)$ и найдем $t_\alpha \geqslant 0$, удовлетворяющее равенству $\varphi(t_\alpha) = \alpha$ для φ, определенной в (23). Тогда в задаче (20) статистический критерий δ^* с критической областью*

$$\mathcal{X}_1^* = \{\boldsymbol{x} \in \mathbb{R}^n \mid \Lambda(\boldsymbol{x}) \geqslant t_\alpha\} \tag{24}$$

является наиболее мощным среди всех критериев уровня значимости α.

Доказательство. Пусть δ^* – это критерий Неймана-Пирсона с критической областью $\mathcal{X}_1^*$, а $\delta(\mathbf{X})$ – любой другой критерий уровня значимости α с критической областью $\mathcal{X}_1$. Тогда нужно показать

$$W(\delta^*, \theta_1) \geqslant W(\delta, \theta_1).$$

Во-первых, в силу определения t_α, критерий δ^* удовлетворяет равенству

$$\alpha(\delta^*) = \mathbb{P}_{\theta_0}(\delta^*(\mathbf{X}) = H_1) = \mathbb{P}_{\theta_0}(\mathbf{X} \in \mathcal{X}_1^*) = \mathbb{P}_{\theta_0}(\Lambda(\mathbf{X}) \geqslant t_\alpha) = \alpha.$$

Отсюда и из того, что уровень значимости критерия δ также равняется α, получаем

$$\alpha(\delta) = \mathbb{P}_{\theta_0}(\delta(\mathbf{X}) = H_1) = \mathbb{P}_{\theta_0}(\mathbf{X} \in \mathcal{X}_1) \leqslant \alpha = \mathbb{P}_{\theta_0}(\mathbf{X} \in \mathcal{X}_1^*). \tag{25}$$

Обозначим через $I(\boldsymbol{x}) = \mathbb{1}(\boldsymbol{x} \in \mathcal{X}_1)$ и $I^*(\boldsymbol{x}) = \mathbb{1}(\boldsymbol{x} \in \mathcal{X}_1^*)$ индикаторы множества $\mathcal{X}_1$ и $\mathcal{X}_1^*$ соответственно и рассмотрим функцию

$$f(\boldsymbol{x}) = (I^*(\boldsymbol{x}) - I(\boldsymbol{x}))(p_{\theta_1}(\boldsymbol{x}) - t_\alpha p_{\theta_0}(\boldsymbol{x})), \tag{26}$$

где $p_{\theta_1}(\boldsymbol{x})$ и $p_{\theta_0}(\boldsymbol{x})$ – это плотности при гипотезах H_0 и H_1 соответственно. Покажем, что $f(\boldsymbol{x}) \geqslant 0$ при всех $\boldsymbol{x} \in \mathcal{X}$: рассмотрим

1. $\boldsymbol{x} \in \mathcal{X}_1^*$: $I^*(\boldsymbol{x}) = 1$ и $p_{\theta_1}(\boldsymbol{x}) \geqslant t_\alpha p_{\theta_0}(\boldsymbol{x})$, значит оба сомножителя в формуле (26) неотрицательны, поэтому $f(\boldsymbol{x}) \geqslant 0$;
2. $\boldsymbol{x} \notin \mathcal{X}_1^*$: $I^*(\boldsymbol{x}) = 0$ и $p_{\theta_1}(\boldsymbol{x}) < t_\alpha p_{\theta_0}(\boldsymbol{x})$, значит оба сомножителя неположительны, поэтому $f(\boldsymbol{x}) \geqslant 0$.

Значит, $f(\boldsymbol{x})$ действительно неотрицательна, тогда

$$\begin{aligned}
0 &\leqslant \int_{\mathcal{X}} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \int_{\mathcal{X}} (I^*(\boldsymbol{x})p_{\theta_1}(\boldsymbol{x}) - I(\boldsymbol{x})p_{\theta_1}(\boldsymbol{x}) - t_\alpha(I^*(\boldsymbol{x})p_{\theta_0}(\boldsymbol{x}) - I(\boldsymbol{x})p_{\theta_0}(\boldsymbol{x})))\mathrm{d}\boldsymbol{x} \\
&= \int_{\mathcal{X}_1^*} p_{\theta_1}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_{\mathcal{X}_1} p_{\theta_1}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - t_\alpha\left(\int_{\mathcal{X}_1^*} p_{\theta_0}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_{\mathcal{X}_1} p_{\theta_0}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right) \\
&= \mathbb{P}_{\theta_1}(\mathbf{X} \in \mathcal{X}_1^*) - \mathbb{P}_{\theta_1}(\mathbf{X} \in \mathcal{X}_1) - t_\alpha(\mathbb{P}_{\theta_0}(\mathbf{X} \in \mathcal{X}_1^*) - \mathbb{P}_{\theta_0}(\mathbf{X} \in \mathcal{X}_1)).
\end{aligned}$$

Отсюда и из неравенства (25) вытекает, что

$$\mathbb{P}_{\theta_1}(\boldsymbol{X} \in \mathcal{X}_1^*) \geqslant \mathbb{P}_{\theta_1}(\boldsymbol{X} \in \mathcal{X}_1),$$

то есть мощность δ^* не меньше мощности любого другого критерия с точно таким же уровнем значимости

$$W(\delta^*, \theta_1) \geqslant W(\delta, \theta_1).$$

□

Далее покажем, что критерий Неймана-Пирсона будет строго несмещенным.

Лемма 12. *Критерий Неймана-Пирсона δ^*, задаваемый критической областью (24), является строго несмещенным*

$$W(\delta^*, \theta) > \alpha.$$

Доказательство. Рассмотрим два случая

1. $t_\alpha > 1$: из определения $\mathcal{X}_1^*$ следует, что для любого $\boldsymbol{x} \in \mathcal{X}_1^*$: $p_{\theta_1}(\boldsymbol{x}) \geqslant t_\alpha p_{\theta_0}(\boldsymbol{x}) > p_{\theta_0}(\boldsymbol{x})$. Тогда

$$W(\delta^*, \theta_1) = \mathbb{P}_{\theta_1}(\boldsymbol{X} \in \mathcal{X}_1^*) = \int_{\mathcal{X}_1^*} p_{\theta_1}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} > \int_{\mathcal{X}_1^*} p_{\theta_0}(\boldsymbol{x})\mathrm{d}x = \mathbb{P}_{\theta_0}(\boldsymbol{X} \in \mathcal{X}_1^*) = \alpha.$$

2. $t_\alpha \leqslant 1$: из определения $\mathcal{X}_1^*$ следует, что при $\boldsymbol{x} \in \mathcal{X}_0 := \mathcal{X} \backslash \mathcal{X}_1^*$, $p_{\theta_1}(\boldsymbol{x}) < t_\alpha p_{\theta_0}(\boldsymbol{x}) \leqslant p_{\theta_0}(\boldsymbol{x})$. Тогда

$$W(\delta^*, \theta_1) = \mathbb{P}_{\theta_1}(\boldsymbol{X} \in \mathcal{X}_1^*) = 1 - \int_{\mathcal{X}_0} p_{\theta_1}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} > 1 - \int_{\mathcal{X}_0} p_{\theta_0}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_{\mathcal{X}_1^*} p_{\theta_0}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \alpha.$$

То есть в любом случае $W(\delta^*, \theta_1) > \alpha$, что и требовалось доказать. □

10.2 Сложные гипотезы

Пусть теперь альтернатива сложная односторонняя

$$\begin{aligned} &H_0 : \theta = \theta_0; \\ &H_1 : \theta > \theta_0. \end{aligned} \tag{27}$$

Теорема 12 (О равномерно наиболее мощном критерии)**.** *Пусть статистика отношения правдоподобия $\Lambda(\mathbf{X})$ является монотонной функцией от достаточной статистики $T(\boldsymbol{X})$, то есть такой, что правдоподобие представляется в виде*

$$L_{\mathbf{X}}(\theta) = g(T(\mathbf{X}), \theta)h(\mathbf{X}),$$

где g и h – некоторые функции. Тогда в задаче (27) существует равномерно наиболее мощный критерий, совпадающий с критерием Неймана-Пирсона для задачи (20).

Доказательство. Рассмотрим статистику отношения правдоподобия

$$\Lambda(\mathbf{X}) = \frac{g(T(\mathbf{X}), \theta_1)}{g(T(\mathbf{X}), \theta_0)},$$

и пусть $\Lambda(\mathbf{X})$ строго возрастающая функция от $T(\mathbf{X})$, тогда критическая область в критерии Неймана-Пирсона для задачи (20) может быть записана эквивалентным образом

$$\Lambda(\mathbf{X}) \geqslant c_\alpha \quad \Longleftrightarrow \quad T(\mathbf{X}) \geqslant t_\alpha,$$

причем граница t_α определяется из уровня значимости критерия

$$\mathbb{P}_{\theta_0}(T(\mathbf{X}) \geqslant t_\alpha) = \alpha.$$

То есть граница t_α одна и та же при любой альтернативе $\theta_1 > \theta_0$, значит и критерий будет точно такой же, как и для задачи (20). Случай строго убывающей фукнции рассматривается аналогично, но знак будет в другую сторону

$$\Lambda(\mathbf{X}) \geqslant c_\alpha \quad \Longleftrightarrow \quad T(\mathbf{X}) \leqslant t_\alpha.$$

□

Замечание 11. В условиях Теоремы (12) гипотезы (H_0, H_1) со знаками $(=, >)$ могут быть заменены на гипотезы $(=, <)$, $(\geqslant, <)$ или $(\leqslant, >)$; критерий останется таким же. В случае монотонного убывания отношения правдоподобия знак в критерии меняется на противоположный.

Замечание 12 (Экспоненциальное семейство)**.** Для распределений из экспоненциального семейства, то есть таких, для которых правдоподобие представляется в виде

$$L_{\mathbf{X}}(\theta) = \exp\{\eta(\theta)T(\mathbf{X}) - A(\theta)\}h(\mathbf{X}),$$

где $T(\mathbf{X})$ – достаточная статистика, всегда существует равномерно наиболее мощный критерий в задаче (27), если $\eta(\theta)$ – строго монотонная функция, так как тогда и $\Lambda(\mathbf{X})$ будет строго монотонной функцией от достаточной статистики $T(\mathbf{X})$.

Замечание 13 (Дискретные распределения)**.** В случае дискретных распределений не для любого $\alpha \in (0, 1)$ существует такое $t_\alpha \geqslant 0$, удовлетворяющее равенству $\varphi(t_\alpha) = \alpha$ для φ, определенной в (23). Если требуется построить критерий с уровнем значимости в точности равным α, то прибегают к рандомизации. Однако на практике зачастую строят критерий Неймана-Пирсона с уровнем значимости $\alpha_0 \leqslant \alpha$.

11 Критерий хи-квадрат Пирсона. p-value.

Критерий хи-квадрат Пирсона является наиболее универсальным методом проверки простых гипотез о виде распределения. Критерий применим как для дискретных распределений, так и для абсолютно непрерывных. Для абсолютно непрерывных распределений понадобится предварительная группировка соответствующие наблюдений, сведя задачу к дискретному случаю.

11.1 Критерий согласия хи-квадрат Пирсона. Простая гипотеза

Пусть дана выборка $\boldsymbol{X} = (X_1, \ldots, X_n)$ из неизвестного распределения $F(x)$. И пусть выдвинута простая гипотеза

$$H_0 : F(x) = F_0(x)\,, \tag{28}$$

где функция распределения $F_0(x)$ фиксирована. В качестве альтернативы берутся все прочие распределения. Критерии для проверки такой гипотезы мы будем называть критериями согласия, так как альтернатива никак не конретизирована и речь идет лишь о согласии данных с H_0.

Дискретные распределения Пусть мы знаем, что элементы выборки $X_1, \ldots, X_n$ принимают целочисленные значения от 1 до N, каждое с вероятностью

$$p_j := \mathbb{P}(X_i = j), \quad i = 1, \ldots, n,\ j = 1, \ldots, N.$$

Определим частоты

$$\nu_j = \sum_{i=1}^{n} \mathbb{1}\{X_i \in j\}\,, \quad j \in \{1, \ldots, N\}\,.$$

Теперь покажем, как абсолютно непрерывный случай может быть сведен к дискретному.

Абсолютно непрерывные распределения Проведем группировку наблюдений: разобьем числовую прямую на N непересекающихся промежутков $\Delta_1, \ldots, \Delta_N$, таких что $\cup\Delta_i = \mathbb{R}$. Тогда вероятность того, что элемент выборки X_i попал в Δ_j Тогда частоты могут быть определены как

$$p_j := \int_{\Delta_j} \mathrm{d}F(x), \quad i = \{1, \ldots, n\},\ j = \{1, \ldots, N\}.$$

При справедливости гипотезы H_0 эти вероятности должны равняться

$$p_j^0 := \mathbb{P}(X_i = j) = \int_{\Delta_j} \mathrm{d}F_0(x), \quad i = 1, \ldots, n,\ j = \{1, \ldots, N\}.$$

Соответствующие частоты

$$\nu_j = \sum_{j=1}^{n} \mathbb{1}\{X_i \in \Delta_j\}\,, \quad j \in \{1, \ldots, N\}\,, i \in \{1, \ldots, n\} \tag{29}$$

Замечание 14. Разумеется группировка слагаемых сопряжена потерей информации. Также возникает вопрос выбора числа интервалов N и их вида. Здесь можно предложить несколько эвристических правил, в частности, $N \approx \sqrt{n}$.

Зададим векторы частот $\nu = (\nu_1, \dots, \nu_N)$, $\boldsymbol{p} = (p_1, \dots, p_N)$, $\boldsymbol{p}^{(0)} = (p_1^{(0)}, \dots, p_N^{(0)})$. Тогда гипотеза (28) может быть эквивалентно переформулирована как

$$H_0 : \boldsymbol{p} = \boldsymbol{p^{(0)}} \,. \tag{30}$$

При этом, в силу УЗБЧ, выполнено

$$\frac{\nu_i}{n} = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{X_i \in \Delta_j\} \xrightarrow{a.s.} \mathbb{E}\mathbb{1}\{X_1 \in \Delta_j\} = \mathbb{P}(\xi_1 = i) = p_i \,.$$

Поэтому при справедливости гипотезы H_0 значения ν_j должны быть близки к np_j. Введем статистику Пирсона, значения которой должны быть малы при справедливости гипотезы H_0.

Определение 39 (Статистика Пирсона)**.**

$$T_{\chi^2} = \sum_{j=1}^{N} \frac{(\nu_j - np_j^{(0)})^2}{np_j^{(0)}} \,.$$

Очевидно, если гипотеза H_0 справедлива, значения статистики χ^2 не должны быть слишком большими. Поэтому естественно выбрать критическую область для гипотезы H_0 в виде $\{\chi^2 > t_\alpha\}$, где критическое значение t_α при заданном уровне значимости α должно определяться из условия

$$\mathbb{P}(T_{\chi^2} > t_\alpha | H_0) = \alpha.$$

Точное распределение статистики $\chi^2_{0,n}$ неудобно для расчетов, однако, у этой статистики есть простое предельное распределение, не зависящее от $\boldsymbol{p^{(0)}}$. Справедлива следующая теорема:

Теорема 13 (Предельное распределение статистики Пирсона)**.** *Если* $0 < p_j^{(0)} < 1$ *при* $j \in \{1, \dots, N\}$*, то при условии справедливости* H_0 *выполнено*

$$T_{\chi^2} \xrightarrow{d} \chi^2(N-1) \quad n \to \infty,$$

где $\chi^2(N-1)$ *– распределение* χ^2 *с* $N-1$ *степенями свободы.*

Тогда критическая область критерия хи-квадрат Пирсона может быть задается следующим способом

$$\mathcal{X}_1 = \{\boldsymbol{x} : T_{\chi^2} \geqslant t_\alpha\},$$

где t_α – $1-\alpha$-квантиль в распределении $\chi^2(N-1)$, так как критическая граница t_α определяется из уровня значимости критерия

$$\alpha = \mathbb{P}_{H_0}(T_{\chi^2} \geqslant t_\alpha) = 1 - F_{T_{\chi^2}}(t_\alpha).$$

Замечание 15 (Условия применимости)**.** На практике критерий хи-квадрат хорошо работает уже для выборок объема $n \geqslant 50$ и при частотах $\nu_j \geqslant 5$ для любого $j = \{1, \dots N\}$.

11.2 Критерий согласия хи-квадрат Пирсона. Сложная гипотеза

Пусть теперь функция распределения $F_0(x,(\boldsymbol{\theta}))$ задана с точностью до неизвестного параметра $(\boldsymbol{\theta}) = (\theta_1,\dots,\theta_r)$. Например проверяется гипотеза о том, что выборка $\boldsymbol{X} = (X_1,\dots,X_n)$ распределена по закону Бернулли.

$$H_0 : F(x) \in \mathcal{F}_0(x) = \{F_0(x,(\boldsymbol{\theta})) : \boldsymbol{\theta} = (\theta_1,\dots,\theta_r) \in \Theta \subset \mathbb{R}^r\},$$

где функция распределения $F_0(x)$ фиксирована. В качестве альтернативны берутся все прочие распределения. Критерии для проверки такой гипотезы мы будем называть критериями согласия, так как альтернатива никак не конретизирована и речь идет лишь о согласии данных с H_0. Тогда задача (30) заменится на

$$H_0 : \boldsymbol{p}(\boldsymbol{\theta}) = \boldsymbol{p^{(0)}}(\boldsymbol{\theta}).$$

Поскольку параметр $\boldsymbol{\theta}$ оценим его методом максимального правдоподобия. Для этого поймем, как распределен вектор частот $\boldsymbol{\nu}$: вектор частот $\boldsymbol{\nu} \sim M(n,\boldsymbol{p}(\boldsymbol{\theta}))$ имеет мультиномиальное (полиномиальное) распределение, которое задается следующей вероятностью

$$\mathbb{P}(\boldsymbol{\nu} = \boldsymbol{y}) = \frac{n!}{y_1!\dots,y_N!}(p_1^{(0)}(\boldsymbol{\theta}))^{y_1}\dots(p_N^{(0)}(\boldsymbol{\theta}))^{y_N},$$

где $\boldsymbol{y} = (y_1,\dots,y_N), y_j \geqslant 0$ для всех $j = \{1,\dots N\}$, $\sum_{j=1}^n y_j = n$.

Пример 14. При $N = 2$ мультиномиальное (полиномиальное) распределение $M(n,p_1,p_2)$ есть биномиальное распределение $\mathrm{Bin}(n,p_1)$.

Тогда правдоподобие параметра $\boldsymbol{\theta}$ вектора $\boldsymbol{\nu}$

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{n!}{\nu_1!\dots,\nu_N!}\prod_{j=1}^N (p_j^{(0)}(\boldsymbol{\theta}))^{\nu_j}.$$

Оценка максимального правдоподобия будет равна

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{MLE} &= \arg\max_{\boldsymbol{\theta}\in\Theta} \ln \mathcal{L}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}\in\Theta}\left\{\ln\left(\frac{n!}{\prod_{j=1}^N \nu_j!}\right) + \ln\left(\prod_{j=1}^N \ln\left(p_j^{(0)}(\boldsymbol{\theta})\right)^{\nu_j}\right)\right\}\\ &= \arg\max_{\boldsymbol{\theta}\in\Theta} \ln \mathcal{L}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}\in\Theta}\sum_{j=1}^N \nu_j \ln p_j^{(0)}(\boldsymbol{\theta}). \qquad (31)\end{aligned}$$

Подставим эту оценку в статистику T_{χ^2}

$$T_{\chi^2} = \sum_{j=1}^N \frac{(\nu_j - np_j^{(0)}(\widehat{\boldsymbol{\theta}}_{MLE}))^2}{np_j^{(0)}(\widehat{\boldsymbol{\theta}}_{MLE})}.$$

Теорема 14 (Предельное распределение статистики T_{χ^2})**.** *Пусть для любого* $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^r$

$$\sum_{i=1}^N p_i^{(0)}(\boldsymbol{\theta}) = 1, \quad 0 < p_i^{(0)}(\boldsymbol{\theta}) < 1,$$

пусть существуют непрерывные производные

$$\frac{\partial p_j^{(0)}(\boldsymbol{\theta})}{\partial \theta_k}, \quad \frac{\partial^2 p_j^{(0)}(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_\ell}, \quad k, \ell = 1, \ldots, r$$

и пусть матрица $\left(\frac{\partial p_j^{(0)}(\boldsymbol{\theta})}{\partial \theta_k} \right)$ *размера* $N \times r$ *имеет ранг* r.

Тогда при справедливости H_0

$$T_{\chi^2} \xrightarrow{d} \chi^2(N - r - 1), \quad n \to \infty.$$

где $\chi^2(N - r - 1)$ *- распределение хи-квадрат с* $N - r - 1$ *степенями свободы.*

11.3 p-value

Пусть гипотеза отвергается при уровне значимости α, значит она будет отвергаться и при уровне $\alpha' > \alpha$. То есть существует наименьшее α_0, при котором гипотеза отвергается, назовем это α_0 p-value.

Определение 40 (p-value)**.** Пусть для любого $\alpha \in (0, 1)$ построен критерий уровня значимости α с критической областью $\mathcal{X}_\alpha$ и тестовой статистикой $T(\mathbf{X})$, тогда

$$\text{p-value} = \min_{\alpha \in (0,1)} \{T(\mathbf{X}) \in \mathcal{X}_\alpha\}$$

p-value является мерой свидетельства против гипотезы H_0: чем меньше p-value, тем более сильное свидетельство против H_0.

Замечание 16. Большое p-value не говорит о том, что H_0 верна. Причины большого p-value: H_0 действительно верна либо у критерия слабая мощность.

Теорема 15. *Пусть критерий уровня значимости* α *с тестовой статистикой* $T(\mathbf{X})$ *имеет критическую форму вида*

$$\mathcal{X}_\alpha = \{\boldsymbol{x} \in \mathcal{X} : T(\boldsymbol{x}) \geqslant c_\alpha\}.$$

Тогда

$$\text{p-value} = \sup_{F \in \mathcal{F}_0} \mathbb{P}_{H_0}(T(\mathbf{X}) \geqslant T(\boldsymbol{x})).$$

12 Гипотеза и критерий однородности.

Пусть имеется K серий независимых наблюдений

$$X_1^{(1)}, \dots, X_{n_1}^{(1)} \sim F_1(x)$$
$$\vdots$$
$$X_1^{(K)}, \dots, X_{n_K}^{(K)} \sim F_K(x).$$

Гипотезой однородности является утверждение, что все эти выборки были сгенерированы из одного и того же распределения, то есть

$$H_0 : F_1(x) = \dots = F_m(x) =: F(x).$$

В отличие от гипотезы о виде распределения функция $F(x)$ неизвестна и не задается нулевой гипотезой. Это задача непараметрическогй статистики. В случае справедливости нулевой гипотезы все выборки можно объединить в одну большую выборку. Также проверка гипотез однородности представляет собой A/B тестирование, при котором зачастую требуется определить, достигнуты ли какие-то улучшения.

12.1 Критерий хи-квадрат Пирсона.

Критерий хи-квадрат Пирсона – это универсальный критерий, применимый к анализу данных любой природы (как мы уже видели в предыдущем разделе метод группировки слагаемых позволяет использование этого критерия не только для дискретных случайных величин, но и абсолютно непрерывных). Поэтому для простоты рассмотрим только дискретные распределения.

Дискретные распределения. Пусть неизвестное гипотетическое распределение таково, что случайная величина X_i^k ($i = \{1, \dots, n_k\}, k = \{1, \dots, K\}$) принимают целочисленные значения от 1 до N, каждое с вероятностью

$$p_j^{(k)} := \mathbb{P}(X_i = j).$$

Определим частоты

$$\nu_j^{(k)} = \sum_{i=1}^{n} \mathbb{1}\{X_i \in j\}.$$

Зададим векторы частот $\nu^{(k)} = (\nu_1, \dots, \nu_N)$, $\boldsymbol{p}^{(k)} = (p_1^{(k)}, \dots, p_N^{(k)})$ для любого $k = \{\mathbb{1}, \dots, K\}$. Тогда нулевая гипотеза может быть эквивалентно переформулирована

$$H_0 : \boldsymbol{p}^{(1)} = \boldsymbol{p}^{(2)} = \dots = \boldsymbol{p}^{(K)} =: \boldsymbol{p}. \tag{32}$$

Вектор $\boldsymbol{p}$ не задается нулевой гипотезой, поэтому его необходимо оценить. Интуитивно понятно, что если нулевая гипотеза верна, то выборки можно объединить в одну большую выборку и по УЗБЧ оценить $p_j \approx \sum_{k=1}^K \nu_j^{(k)} / \sum_{k=1}^K n_k$ для $j = \{1, \dots, N\}$. Докажем это утверждение методом максимального правдоподобия.

Лемма 13. *Оценка максимальноого правдоподобия для неизвестного вектора вероятностей $\boldsymbol{p}$ в нулевой гипотезе (32) задается следующей формулой*

$$\widehat{p}_j = \frac{\sum_{k=1}^K \nu_j^{(k)}}{n} \quad \text{для каждого } j = \{1, \dots, N\},$$

где $n = \sum_{k=1}^K n_k$.

Доказательство. При справедливости нулевой гипотезы векторы частот имеет мультиномиальное (полиномиальное) распределение

$$\boldsymbol{\nu}^{(k)} \sim M(n_k, \boldsymbol{p})$$

Напомним что, мультиномиальное распределение задается следующей вероятностью

$$\mathbb{P}(\boldsymbol{\nu}^k = \boldsymbol{y}) = \frac{n!}{y_1! \dots, y_N!} p_1^{y_1} \dots p_N^{y_N},$$

где $\boldsymbol{y} = (y_1, \dots, y_N), y_j \geqslant 0$ для всех $j = \{1, \dots N\}$, $\sum_{j=1}^n y_j = n$. Тогда правдоподобие вектора $\boldsymbol{p}$ для $\boldsymbol{\nu}^{(1)}, \dots, \boldsymbol{\nu}^{(K)}$

$$\mathcal{L}(\boldsymbol{p}) = \prod_{k=1}^K \frac{n_k!}{\prod_{j=1}^N \nu_j^{(k)}!} \prod_{j=1}^N p_j^{\nu_j^{(k)}}$$

Оценка максимального правдоподобия вектора $\boldsymbol{p}$ будет находиться из решения оптимизационной задачи

$$\begin{aligned}
\widehat{\boldsymbol{p}} &= \arg \max_{\substack{\boldsymbol{p}:\ p_j \geqslant 0,\ j=\{1,\dots N\},\\ \sum_{j=1}^N p_j = 1}} \ln \mathcal{L}(\boldsymbol{p}) = \arg \max_{\substack{\boldsymbol{p}:\ p_j \geqslant 0,\ j=\{1,\dots N\},\\ \sum_{j=1}^N p_j = 1}} \left\{ \ln \left(\prod_{k=1}^K \prod_{j=1}^N p_j^{\nu_j^{(k)}} \right) + \ln \left(\prod_{k=1}^K \frac{n_k!}{\prod_{j=1}^N \nu_j^{(k)}!} \right) \right\} \\
&= \arg \max_{\substack{\boldsymbol{p}:\ p_j \geqslant 0,\ j=\{1,\dots N\},\\ \sum_{j=1}^N p_j = 1}} \ln \left(\prod_{k=1}^K \prod_{j=1}^N p_j^{\nu_j^{(k)}} \right) \\
&= \arg \max_{\substack{\boldsymbol{p}:\ p_j \geqslant 0,\ j=\{1,\dots N\},\\ \sum_{j=1}^N p_j = 1}} \sum_{k=1}^K \sum_{j=1}^N \nu_j^{(k)} \ln p_j
\end{aligned}$$

Решим данную задачу минимизации методом множителей Лагранжа: внесем ограничение $\sum_{j=1}^N p_j = 1$ в оптимизируемую фукнцию со штрафом $\lambda \in \mathbb{R}$

$$\min_{\lambda \in \mathbb{R}} \max_{\boldsymbol{p}:\ p_j \geqslant 0,\ j=\{1,\dots N\}} \left\{ \sum_{k=1}^K \sum_{j=1}^N \nu_j^{(k)} \ln p_j + \lambda (1 - \sum_{j=1}^N p_j) \right\}.$$

Продифференцируем по p_j и приравняем к нулю, так как теперь полученная задача без ограничений (почти)

$$\sum_{k=1}^K \nu_j^{(k)} \ln p_j + \lambda = 0$$

Тогда

$$\widehat{p}_j = \frac{\sum_{k=1}^K \nu_j^{(k)}}{\lambda}. \tag{33}$$

Поскольку λ неизвестна, воспользуемся граничным условием $\sum_{j=1}^N p_j = 1$: просуммируем (33) и найдем λ

$$\sum_{j=1}^N \widehat{p}_j = \sum_{j=1}^N \frac{\sum_{k=1}^K \nu_j^{(k)}}{\lambda} = \frac{1}{\lambda} \sum_{j=1}^N \sum_{k=1}^K \nu_j^{(k)}.$$

Значит

$$\lambda = \sum_{j=1}^N \sum_{k=1}^K \nu_j^{(k)} = \sum_{k=1}^K n_k = n.$$

Подставляем это выражение в (33) и завершаем доказательство. □

Преобразуем статистику Пирсона (39) для данной задачи

$$T_{\chi^2} = \sum_{k,j=1}^{K,N} \frac{(\nu_j^{(k)} - n_k \widehat{p}_j)^2}{n_k \widehat{p}_j},$$

где

$$\widehat{p}_j = \frac{\sum_{k=1}^m \nu_{kj}}{n}, \quad n = \sum_{k=1}^m n_k.$$

При справедливости гипотезы H_0 аналогично теореме (13) будет иметь место следующий предельный результат

$$T_{\chi^2} = \sum_{k,j=1}^{K,N} \frac{(\nu_{kj} - n_k \widehat{p}_j)^2}{n_k \widehat{p}_j} \xrightarrow{d} \chi^2((K-1)(N-1))$$

при всех $n_k \to \infty$, $k = 1, \ldots, K$. При справедливости гипотезы H_0 значение данной статистики не должно быть слишком большим, поэтому асимптотический критерий однородности будет выглядеть следующим образом.

Асимптотический критерий однородности.

$$H_0 \text{ отвергается} \iff T_{\chi^2} \geqslant t_\alpha,$$

где t_α – $1-\alpha$-квантиль в распределении $\chi^2((K-1)(N-1))$, так как критическая граница t_α определяется из уровня значимости критерия

$$\alpha = \mathbb{P}_{H_0}(T_{\chi^2} \geqslant t_\alpha) = 1 - F_{T_{\chi^2}}(t_\alpha).$$

Поскольку критерий асимптотический, применять его можно лишь для больших выборок.

12.2 Параметрический случай

Эту же методику можно использовать и для проверки гипотезы о том, что гипотетические вероятности $\boldsymbol{p}^{(k)}$ имеют заданную параметрическую форму $\boldsymbol{p}^{(k)} = \boldsymbol{p}^{(k)}(\boldsymbol{\theta})$. Такая ситуация возникает, например, когда требуется проверить гипотезу о том, что K серий наблюдений произведены над одной и той же случайной величиной с рапспределением из заданного параметрического семейства $\mathcal{P} = \{\mathbb{P}_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$.

Статистика Пирсона (39) для данной задачи

$$T_{\chi^2} = \sum_{k,j=1}^{K,N} \frac{(\nu_j^{(k)} - n_k p_j(\widehat{\boldsymbol{\theta}})^2}{n_k p_j(\widehat{\boldsymbol{\theta}})},$$

где

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \ln \mathcal{L}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}\in\Theta} \sum_{j=1}^N \nu_j \ln p_j^{(0)}$$

Аналогично теореме (13) будет иметь место следующий предельный результат

$$T_{\chi^2} = \sum_{k,j=1}^{K,N} \frac{(\nu_{kj} - n_k \widehat{p}_j)^2}{n_k \widehat{p}_j} \xrightarrow{d} \chi^2(K(N-1) - r).$$

при справедливости гипотезы H_0 при всех $n_k \to \infty$, $k = 1, \ldots, K$. Тогда асимптотический критерий однородности будет выглядеть следующим образом.

Асимптотический критерий однородности.

$$H_0 \text{ отвергается} \quad \Longleftrightarrow \quad T_{\chi^2} \geqslant t_\alpha,$$

где t_α – $1-\alpha$-квантиль в распределении $\chi^2(K(N-1)-r)$.

13 Байесовская статистика. Априорное и апостериорное распределение. Байесовская оценка. Сопряженные семейства распределений.

Ранее в курсе при рассмотрении задачи точечного оценивания параметров мы так или иначе подходили к вопросу со стороны равномерного подхода: одна оценка считается не хуже другой, если при всех θ ее риск не больше. Такой строгий подход к сравнению оценок накладывал существенные ограничения на класс оценок, среди которых мы ищем наилучшую.

На практике же зачастую у нас есть некоторые *априорные* предположения о том, как устроены возможные значения параметра. В классическом примере с оцениванием вероятности θ выпадения монетки орлом чисто теоретически параметр может принимать все значения в отрезке $[0, 1]$. С точки зрения же здравого смысла, θ скорее всего будет принимать близкие к $1/2$ значения, если мы знаем, что работаем с обычной монеткой, «нечестность» которой достигается разве что за счет неидеальной формы и других физических особенностей. В такой ситуации очень важно, чтобы оценка имела небольшой риск при значениях θ, близких к $1/2$. Риск же оценки при экстремальных θ несущественен. Проводя такие логические рассуждения, мы фактически присваиваем каждому значению θ некоторый «вес», соответствующий тому, насколько нам важно получить низкие значения риска при данном значении параметра. Разумной же задачей после введения весов становится нахождение оценки, минимизирующей взвешенную сумму рисков при разных значениях θ. Формализовать же понятие веса удобнее всего в виде вероятностного распределения, такое распределение называют *априорным* и обозначают $\pi(\theta)$. После того, как была сгенерирована выборка, можно обновить распределение $\pi(\theta)$ через условное $\pi(\theta|\boldsymbol{x})$, такое распределение называют *апостериорным*. Если априорное распределение $\pi(\theta)$ имеет интерпретацию наших изначальных представлений о том, какие значения параметра более и менее вероятные, то $\pi(\theta|\boldsymbol{x})$ представляет собой обновленные знания о θ, учитывающие информацию о параметре, содержащуюся в выборке.

Теорема Байеса позволяет связать априорное и апостериорное распределение. Алгоритм можно описать следующими действиями.

1. Введем априорное распределение на Θ с обобщенной плотностью $\pi(\theta)$. Распределение задает наши изначальные предположения о том, какое должно быть истинное значение неизвестного параметра θ

2. Сгенерируем выборку
$$X_1, \dots, X_n \sim p(x|\theta), \quad \theta \sim \pi(\theta).$$
Заметим, что сейчас мы обозначаем обобщенную плотность распределения, из которого была сгенерирована выборка, как условную плотность, потому что в отличие от классического случая, θ теперь случайная величина, а выборка генерируется по-прежнему из какого-то распределения с фиксированным значением θ.

3. Обновляем наши предположения о неизвестном параметре θ, вычисляя апостериорное распределение с обобщенной плотностью $\pi(\theta|\boldsymbol{x})$. По теореме Байеса условную плотность можно расписать следующим образом
$$\pi(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)\pi(\theta)}{\int\limits_{\Theta} p(\boldsymbol{x}|\theta)\pi(\theta)\mathrm{d}\theta}. \tag{34}$$

Заметим, что условная плотность $p(\boldsymbol{x}|\theta)$ – это функция правдоподобия, если вместо $\boldsymbol{x}$ подставить выборку X. Тогда из (34) получаем, что апостериорная плотность должна быть прямо пропорциональна произведению функции правдоподобия на априорную обобщенную плотность

$$\pi(\theta|\boldsymbol{x}) \propto \mathcal{L}(\theta)\pi(\theta).$$

Зная апостериорное распределение, можно вычислить *апостериорные потери*.

Определение 41 (Апостериориорные потери)**.** Пусть $\rho(\cdot,\cdot)$ – это функция потерь, тогда апостериорными потерями называется величина

$$r(\widehat{\theta}|\boldsymbol{x}) = \mathbb{E}[\rho(\theta,\widehat{\theta})|\mathbf{X}=\boldsymbol{x}] = \int_\Theta \rho(\theta,\widehat{\theta})\pi(\theta|\boldsymbol{x})\mathrm{d}\theta.$$

В байесовском подходе в качестве оценки неизвестного параметра θ берется та оценка, которая минимизирует байесовский риск.

Определение 42 (Байесовский риск)**.** Пусть $\rho(\cdot,\cdot)$ — функция потерь и $R(\widehat{\theta}(\boldsymbol{X}),\theta) = \mathbb{E}_\theta\rho(\widehat{\theta}(\boldsymbol{X}),\theta)$ — соответствующий (классический) риск оценки $\widehat{\theta}(\boldsymbol{X})$. Байесовским риском для *априорного* распределения с плотностью $\pi(\theta)$ называется

$$B_\pi(\widehat{\theta}(\boldsymbol{X})) = \mathbb{E}R(\theta,\widehat{\theta}(\boldsymbol{X})) = \int\limits_\Theta R(\theta,\widehat{\theta}(\boldsymbol{X}))\pi(\theta)\mathrm{d}\theta. \tag{35}$$

Байесовский риск является естественной формализацией взвешенного подсчета риска, про который мы говорили выше. Априорное распределение $\pi(\theta)$ соответствует тем самым изначальным предположениям о том, какие значения параметра более и менее вероятны.

Определение 43 (Байесовская оценка)**.** Байесовской оценкой для априорного распределения $\pi(\theta)$ называется оценка $\widehat{\theta}_\pi(\boldsymbol{X})$, минимизирующая байесовский риск

$$\widehat{\theta}_\pi(\boldsymbol{X}) = \arg\min_{\widetilde{\theta}\in\Theta} B_\pi(\widetilde{\theta}).$$

Теорема 16. *Определим оценку*

$$\widehat{\theta}(\boldsymbol{x}) = \arg\min_{\widetilde{\theta}} r(\widetilde{\theta}|\boldsymbol{x}),$$

тогда $\widehat{\theta}(\boldsymbol{x})$ – байесовская оценка,

Доказательство. Пусть $f(\boldsymbol{x},\theta)$ – это совместная плотность, тогда определим маргинальную плотность

$$f(\boldsymbol{x}) = \int_\Theta f(\boldsymbol{x},\theta)\pi(\theta)\mathrm{d}\theta.$$

Рассмотрим байесовский риск

$$\begin{aligned}
B_\pi(\widehat{\theta}(\boldsymbol{x})) = \int\limits_\Theta R(\theta,\widehat{\theta}(\boldsymbol{x}))\pi(\theta)\mathrm{d}\theta &= \int\limits_\Theta\int_{\mathcal{X}} \rho(\theta,\widehat{\theta}(\boldsymbol{x}))p(\boldsymbol{x}|\theta)\mathrm{d}\boldsymbol{x}\pi(\theta)\mathrm{d}\theta \\
&\int_\Theta\int_{\mathcal{X}} \rho(\theta,\widehat{\theta}(\boldsymbol{x}))f(\boldsymbol{x},\theta)\mathrm{d}\boldsymbol{x}\mathrm{d}\theta \\
&= \int_\Theta\int_{\mathcal{X}} \rho(\theta,\widehat{\theta}(\boldsymbol{x}))\pi(\theta|\boldsymbol{x})f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}\theta \\
&= \int_{\mathcal{X}} r(\widehat{\theta}(\boldsymbol{x})|\boldsymbol{x})f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.
\end{aligned}$$

Для любого $\boldsymbol{x}\in\mathcal{X}$ определим

$$\widehat{\theta}(\boldsymbol{x}) = \arg\min_{\widetilde{\theta}} r(\widetilde{\theta}|\boldsymbol{x}).$$

Значит она будет минимизировать и $\int_{\mathcal{X}} r(\widehat{\theta}(\boldsymbol{x})|\boldsymbol{x})f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, то есть $\widehat{\theta}(\boldsymbol{x})$ – байесовская. □

Для заданного априорного распределения $\pi(\theta)$ нужно найти соответствующую байесовскую оценку $\widehat{\theta}_\pi(\boldsymbol{X})$. Но как ее искать? В общем случае это сложный вопрос, но если $\Theta \subset \mathbb{R}^n$, а функция потерь ρ квадратичная, то байесовская оценка будет равняться математическому ожиданию апостериорного распределения (апостериорному среднему). Для простоты, далее мы будем считать, что $\Theta \subset \mathbb{R}$.

Теорема 17. *Пусть выбрана квадратичная функция потерь $\rho(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$. Тогда байесовской оценка – это апостериорное среднее*

$$\widehat{\theta}_\pi(\boldsymbol{x}) = \int_\Theta \theta \pi(\theta|\boldsymbol{x}) \mathrm{d}\theta. \tag{36}$$

Доказательство. Воспользуемся теоремой (16) и тем фактом, что потери квадратичные

$$\widehat{\theta}_\pi(\boldsymbol{x}) = \arg\min_{\widetilde{\theta}} r(\widetilde{\theta}|\boldsymbol{x}) = \arg\min_{\widetilde{\theta}} \int_\Theta (\theta - \widetilde{\theta})^2 \pi(\theta|\boldsymbol{x}) \mathrm{d}\theta.$$

Продифференцируем по $\widetilde{\theta}$

$$\int_\Theta 2(\theta - \widetilde{\theta}) \pi(\theta|\boldsymbol{x}) \mathrm{d}\theta = 0.$$

Теперь выразим $\widetilde{\theta}$

$$\widetilde{\theta} = \frac{\int_\Theta \theta \pi(\theta|\boldsymbol{x}) \mathrm{d}\theta}{\int_\Theta \pi(\theta|\boldsymbol{x}) \mathrm{d}\theta} = \int_\Theta \theta \pi(\theta|\boldsymbol{x}) \mathrm{d}\theta.$$

□

Замечание 17. Если π — дискретное распределение, то апостериорная плотность и байесовская оценка переписываются через суммы:

$$\pi(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)\pi(\theta)}{\sum\limits_\theta p(\boldsymbol{x}|\theta)\pi(\theta)};$$

$$\widehat{\theta}_\pi(\boldsymbol{x}) = \sum_\theta \theta \pi(\theta|\boldsymbol{x}).$$

13.1 Сопряженные распределения.

Существенным ограничением в байесовском подходе являются вычислительные аспекты – нужно взять интеграл в знаменателе, чтобы посчитать апостериорную плотность. Если параметр θ имеет большую размерность, задача подсчета интеграла становится очень трудоемкой, поэтому хотелось бы иметь какой-то способ этот подсчет избегать. Один из таких способов — рассматривать задачу байесовского оценивания для определенного класса априорных распределений, называющихся *сопряженными.*

Определение 44 (Сопряжённые априорные распределения)**.** Если апостериорное распределение $\pi(\theta|\boldsymbol{x})$ принадлежит тому же семейству вероятностных распределений, что и априорное распределение $\pi(\theta)$ (т.е. имеет тот же вид, но с другими параметрами), то это семейство распределений называется сопряжённым семейству функций правдоподобия $p(\boldsymbol{x}|\theta)$. При этом распределение $\pi(\theta)$ называется сопряжённым априорным распределением к семейству функций правдоподобия $\pi(\theta|\boldsymbol{x})$.

Пример 15. Пусть $\boldsymbol{X}$ — выборка из распределения Бернулли $\mathrm{Ber}(\theta)$. Покажем, что семейство бета-распределений является сопряженным к семейству распределений Бернулли и найдем байесовскую оценку параметра θ.

Решение. В качестве априорного распределения возьмем бета-распределение $\mathrm{Beta}(\alpha, \beta)$ с известными произвольными параметрами α, β и посчитаем апостериорную плотность. Начинаем, как обычно, с совместной

$$p(\boldsymbol{X}|\theta)\pi(\theta) = \mathcal{L}(\theta)q(\theta) = \theta^{\sum_i X_i}(1-\theta)^{n-\sum_i X_i} \cdot \frac{1}{\mathrm{B}(\alpha,\beta)} t^{\alpha-1}(1-\theta)^{\beta-1}$$
$$= \frac{1}{\mathrm{B}(\alpha,\beta)} t^{\sum_i X_i+\alpha-1}(1-t)^{n-\sum_i X_i+\beta-1}.$$

Получилось что-то очень похожее на плотность бета-распределения: на самом деле это в точности плотность распределения $\mathrm{Beta}(\sum_i X_i + \alpha, n - \sum_i X_i + \beta)$, только умноженная на нормировочную константу, не зависящую от θ.

Замечательно здесь то, что совместная плотность в свою очередь отличается от апостериорной плотности делением на $\int_\Theta p(\boldsymbol{X}|\theta)\pi(\theta)\mathrm{d}\theta$, величину, не зависящую от t. Тогда апостериорная плотность тоже отличается от плотности $\mathrm{Beta}(\sum_i X_i + \alpha, n - \sum_i X_i + \beta)$ домножением на константу. Но если две плотности $p_1(\theta)$ и $p_2(\theta)$ отличаются домножением на константу C, то $C = 1$, так как

$$\int\limits_\Theta p_1(\theta)\mathrm{d}\theta = \int\limits_\Theta p_2(\theta)\mathrm{d}\theta = 1.$$

Значит, $\pi(\cdot|\boldsymbol{X}) = \mathrm{Beta}(\sum_i X_i + \alpha, n - \sum_i X_i + \beta)$, то есть апостериорное распределение лежит в том же семействе, что и априорное, и семейства действительно сопряжены.

Наконец, зная, что байесовская оценка — математического ожидания апостериорного распределения, достаточно вспомнить математического ожидания произвольного бета-распределения. Для $\xi \sim \mathrm{Beta}(\alpha, \beta)$ верно $\mathbb{E}\xi = \frac{\alpha}{\alpha+\beta}$. Тогда байесовская оценка равна

$$\widehat{\theta}_\pi(\boldsymbol{X}) = \frac{\sum_i X_i + \alpha}{\sum_i X_i + \alpha + n - \sum_i X_i + \beta} = \frac{\sum_i X_i + \alpha}{n + \alpha + \beta}.$$

Примечательно здесь то, что байесовская оценка является выпуклой комбинацией оценки максимального правдоподобия $\overline{X}$ и математического ожидания априорного распределения $\alpha/(\alpha+\beta)$

$$\widehat{\theta}_\pi(\boldsymbol{X}) = \frac{n}{n+\alpha+\beta} \cdot \overline{X} + \frac{\alpha+\beta}{n+\alpha+\beta} \cdot \frac{\alpha}{\alpha+\beta}.$$

□

13.2 Байесовским доверительные интервалы

Байесовская статистика позволяет довольно интуитивно строить доверительные интервалы. Поскольку неизвестный параметр θ рассматривается как случайная величина, то доверительный интервал для него может быть построен напрямую через вероятность попадания в этот интервал

$$\mathbb{P}_\theta(T_1(\boldsymbol{X}) < \theta < T_2(\boldsymbol{X})) \geqslant 1 - \alpha,$$

где $\mathbb{P}_\theta$ – апостериорное распределение с обобщенной плотностью $\pi(\theta|\boldsymbol{x})$.

14 Байесовские и минимаксные оценки

Если априорная информация об оцениваемом параметре θ отсутствует, для построения оценки можно использовать так называемый *минимаксный подход*, при котором за меру точности оценки принимается максимальное значение ее риска, или максимальный риск.

Определение 45 (Максимальный риск)**.**

$$R_{\max}(\widehat{\theta}) = \sup_{\theta\in\Theta} R(\theta, \widehat{\theta}).$$

Определение 46 (Максимальная оценка)**.**

$$\widehat{\theta} = \arg\inf_{\widetilde{\theta}} R_{\max}(\widetilde{\theta}) = \arg\inf_{\widetilde{\theta}} \sup_{\theta\in\Theta} R(\theta, \widetilde{\theta})$$

В общем случае вопрос об отыскании такой оценки весьма трудный, но в ряде случаев она может быть найдена через байесовскую оценку: предположим, что существует априорное распределение $\pi(\theta) > 0$, для которого соответствующая байесовская оценка $\widehat{\theta}_\pi$ имеет постоянный риск ($R(\theta, \widehat{\theta}_\pi) = \text{const}$), тогда минимаксная оценка будет байесовской оценкой $\widehat{\theta}_\pi$. Вспомним также, что, например, в случае квадратичных потерь байесовская оценка может быть достаотчно просто найдена, а именно как математическое ожидание апостериорного распределения.

Теорема 18. *Пусть $\widehat{\theta}_\pi$ – байесовская оценка для некоторого априорного распределения $\pi(\theta) > 0$ и для всех $\theta \in \Theta$*

$$R(\theta, \widehat{\theta}_\pi) \leqslant B_\pi(\widehat{\theta}_\pi),$$

то $\widehat{\theta}_\pi$ будет минимаксной оценкой, а соответствующее распределение π будет называться наименее благоприятным распределением.

Доказательство. От противного: пусть $\widehat{\theta}_\pi$ не минимаксная оценка, значит существует $\widehat{\theta}_0$ такая, что

$$R_{\max}(\widehat{\theta}_0) < R_{\max}(\widehat{\theta}_\pi). \tag{37}$$

Поскольку максимальное значение всегда не меньше (взвешенного) среднего, то для любой оценки, а значит и для $\widehat{\theta}_0$

$$B_\pi(\widehat{\theta}_0) \leqslant R_{\max}(\widehat{\theta}_0). \tag{38}$$

По условию для всех $\theta \in \Theta$, $R(\theta, \widehat{\theta}_\pi) \leqslant B_\pi(\widehat{\theta}_\pi)$, значит

$$R_{\max}(\widehat{\theta}_\pi) \leqslant B_\pi(\widehat{\theta}_\pi). \tag{39}$$

Используем (38) и (39) для (37)

$$B_\pi(\widehat{\theta}_0) \leqslant R_{\max}(\widehat{\theta}_0) < R_{\max}(\widehat{\theta}_\pi) \leqslant B_\pi(\widehat{\theta}_\pi).$$

Значит, $\widehat{\theta}_\pi$ не может быть байесовской оценкой, поскольку байесовский риск для оценки $\widehat{\theta}_0$ меньше. Получили противоречие, значит $\widehat{\theta}_\pi$ – минимаксная оценка. □

Замечание 18. Термин *наименее благоприятное рапспределение* отражает следующую особенность такого априорного распределения. Рассмотрим байесовский риск при любом другом априорном распределении π', тогда

$$B_{\pi'}(\widehat{\theta}_\pi) = \int_\Theta R(\theta, \widehat{\theta}_\pi)\pi'(\theta)\mathrm{d}\theta \leqslant \int_\Theta B_\pi(\widehat{\theta}_\pi)\pi'(\theta)\mathrm{d}\theta = B_\pi(\widehat{\theta}_\pi)\int_\Theta \pi'(\theta)\mathrm{d}\theta = B_\pi(\widehat{\theta}_\pi).$$

То есть байесовский риск оценки $\widehat{\theta}_\pi$ при произвольном априорном распределении π' не хуже, чем при априорном распределении π. Поэтому говорят, что распределение π доставляет наибольшие (взвешенные) средние потери.

Следствие 3. *Пусть $\widehat{\theta}_\pi$ – байесовская оценка для некоторого априорного распределения $\pi(\theta) > 0$ и для всех $\theta \in \Theta$*

$$R(\theta, \widehat{\theta}_\pi) = \text{const},$$

то $\widehat{\theta}_\pi$ будет минимаксной оценкой, а соответствующее распределение π будет называться наименее благоприятным распределением.

Доказательство. Рассмотрим

$$B_\pi(\widehat{\theta}_\pi) = \int_\Theta R(\theta, \widehat{\theta}_\pi)\pi(\theta)\mathrm{d}\theta = \text{const} \in_\Theta \pi(\theta)\mathrm{d}\theta = \text{const}.$$

Далее применяем теорему (18), так как $R(\theta, \widehat{\theta}_\pi) \leqslant B_\pi(\widehat{\theta}_\pi) = \text{const}$. □

Однако не всегда минимаксная оценка будет байесовская. В таком случае весьма полезной может оказаться следующая теорема.

Теорема 19. *Пусть $\pi_m(\theta)$ – последовательность из m априорныз распределений таких, что*

$$r_m := \inf_{\widetilde{\theta}} B_\pi(\widetilde{\theta}) \xrightarrow{m\to\infty} r < \infty.$$

Пусть существует оценка $\widehat{\theta}$ такая, что

$$R_{\max}(\widehat{\theta}) = r.$$

Тогда $\widehat{\theta}$ – минимаксная оценка.

Доказательство. От противного: пусть $\widehat{\theta}$ не минимаксная оценка, значит существует $\widehat{\theta}_0$ такая, что

$$R_{\max}(\widehat{\theta}_0) < R_{\max}(\widehat{\theta}). \tag{40}$$

Поскольку максимальное значение всегда не меньше (взвешенного) среднего, то для любой оценки, а значит и для $\widehat{\theta}_0$

$$R_{\max}(\widehat{\theta}_0) \geqslant B_\pi(\widehat{\theta}_0) \geqslant \inf_{\widetilde{\theta}} B_\pi(\widetilde{\theta}) := r_m \xrightarrow{m\to\infty} r = R_{\max}(\widehat{\theta}). \tag{41}$$

Получили противоречие с (40), значит значит $\widehat{\theta}$ – минимаксная оценка. □

14.1 Три подхода в сравнении оценок

Рассмотрим две оценки $\widehat{\theta}_1$ и $\widehat{\theta}_2$, для которых функция риска изображена на графике 3.

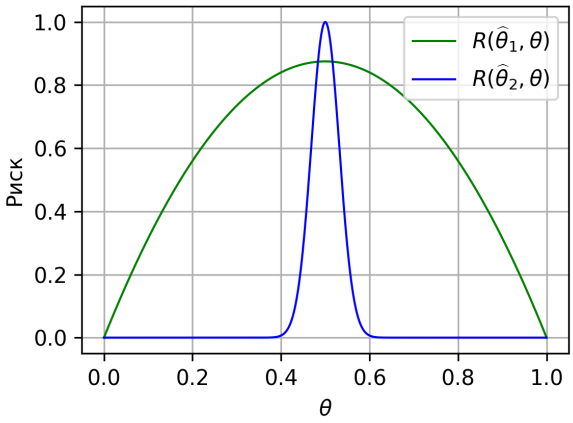

Рис. 3: График функции риска

1. **Равномерный подход:** это наш классический подход к сравнению оценок, который мы изучали в самом начале курса. При данном подходе все значения неизвестного параметра θ считаются равновероятными и оценка $\widehat{\theta}_1$ лучше, чем $\widehat{\theta}_2$ если для всех значений $\theta \in \Theta$

$$R(\widehat{\theta}_1, \theta) \leqslant R(\widehat{\theta}_2, \theta)$$

и существует хотя бы одно значение θ, при котором знак неравенства строгий. В данном подходе для рисунка 3 нельзя сказать какая оценка лучше: $\widehat{\theta}_1$ или $\widehat{\theta}_2$. Более того, в классе всех возможных оценок наилучшей оценки вообще не существует при данном подходе, поэтому мы ограничивались лишь рассмотрением класса несмещенных оценок.

2. **Байесовский подход:** мы предполагаем некоторые значения параметра θ более вероятными, чем другие и выражаем это через априорное распределение $\pi(\theta)$. В этом подходе главное, чтобы риск был меньше на наиболее вероятных значениях θ. Здесь сравниваем $B_\pi(\widehat{\theta}_1) = \mathbb{E}R(\widehat{\theta}_1, \theta)$ и $B_\pi(\widehat{\theta}_2) = \mathbb{E}R(\widehat{\theta}_2, \theta)$. Если, например, более вероятны значения в окрестности 0.5, то может получиться, что оценка $\widehat{\theta}_1$ лучше, чем $\widehat{\theta}_2$. Если же в качестве априорного распределения взять равномерное, то оценка $\widehat{\theta}_2$ будет лучше, чем $\widehat{\theta}_1$, поскольку в этом случае байесовский риск будет равняться площади под графиком

$$B_\pi(\widehat{\theta}) = \mathbb{E}R(\widehat{\theta}, \theta) = \int_\Theta R(\widehat{\theta}, \theta)\pi(\theta)\mathrm{d}\theta = \int_\Theta R(\widehat{\theta}, \theta)\mathrm{d}\theta$$

3. **Минимаксный подход:** мы стремимся избегать больших рисков, поэтому мы сравниваем оценки через максимальное значение риска. В данном подходе оценка $\widehat{\theta}_1$ будет лучше, чем $\widehat{\theta}_2$. Данный пример показывает несовершенство максимального подхода, потому что в данном примере он противоречит интуиции: априорная информация о параметре θ неизвестна, и за исключением малой окрестности вблизи 0.5 риск оценки $\widehat{\theta}_2$ почти равен нулю.

Вспомогательные материалы

Теорема 20. *(Закон больших чисел) Пусть $X_1, X_2, \ldots, X_n$ – выборка из распределения с конечными математическим ожиданием μ. Обозначим $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ – выборочное среднее. Тогда при $n \to \infty$*

$$\overline{X} \xrightarrow{\mathbb{P}} \mu.$$

Теорема 21. *(Центральная предельная теорема) Пусть $X_1, X_2, \ldots, X_n$ – выборка из распределения с конечными математическим ожиданием μ и дисперсией σ^2. Тогда при $n \to \infty$*

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

где $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ – выборочное среднее.

Теорема 22. *(О наследовании сходимости) Если $\xi_n \xrightarrow{a.s./\mathbb{P}/d} \xi$, а h — непрерывная относительно распределения ξ функция (т.е. существует множество B такое, что $\mathbb{P}(\xi \in B) = 1$ и h непрерывна на B), тогда*

$$h(\xi_n) \xrightarrow{a.s./\mathbb{P}/d} h(\xi).$$

Лемма 14. *(Слуцкого) Пусть $\xi_n \xrightarrow{d} \xi$ и $\eta_n \xrightarrow{d} C$, где C — константа. Тогда:*

$$\xi_n + \eta_n \xrightarrow{d} \xi + C$$

$$\xi_n \cdot \eta_n \xrightarrow{d} \xi \cdot C$$

Лемма 15. *Если C — константа, то $\xi_n \xrightarrow{\mathsf{P}} C$ тогда и только тогда, когда $\xi_n \xrightarrow{d} C$.*

Список литературы

Г.И. Ивченко and Ю.И Медведев. *Математическая статистика.* М. Высшая школа, 1984.