Математическая статистика

Антон Ермилов
По лекциям Руслана Пусева

18 декабря 2018 г.

Содержание

1. Введение **1**
1.1 Задачи математической статистики . . . 1
1.2 Математическая постановка задач статистики . . . 2

2. Квантили **4**
2.1 Определение квантили . . . 4
2.2 Оценка генеральных квантилей . . . 4

3. Случайные векторы **6**
3.1 Гауссовский вектор . . . 6

4. Статистические оценки **8**
4.1 Требования, предъявляемые к оценкам . . . 8
4.2 Лемма Фишера . . . 8
4.3 Построение оценок. Метод моментов . . . 10
4.4 Построение оценок. Метод максимального правдоподобия . . . 13
4.5 Достаточные статистики . . . 19
4.6 Доверительные интервалы . . . 24
4.7 Байесовские и минимаксные оценки . . . 29

5. Регрессионный анализ **33**
5.1 Линейная регрессия . . . 33

6. Проверки гипотез **40**
6.1 Постановка задачи . . . 40
6.2 Наиболее мощные критерии . . . 42
6.3 Критерии для непрерывных числовых данных . . . 45
6.4 Критерии для категориальных и дискретных числовых данных . . . 50
6.5 Ранги. Наиболее мощные ранговые критерии . . . 54

6.6 Ранговые критерии. Гипотеза случайности 57
6.7 Ранговые критерии. Гипотеза симметрии 60
6.8 Ранговые критерии. Гипотеза независимости 62

7. Моделирование распределений **64**
7.1 Метод обратной функции 64
7.2 Метод отбора 66
7.3 Метод декомпозиции 67

1. Введение

1.1. Задачи математической статистики

Предисловие.

Обычно задача теории вероятностей формулируется как “По заданной случайной величине X и ее распределению вычислить распределение случайной величины $f(X)$”.

С точки зрения математической статистики происходит нечто похожее: нам дана случайная величина X, распределение которой нам неизвестно, но мы хотим сделать какие-то выводы об этом самом неизвестном распределении и о каких-то функциях от X.

Иными словами, мы хотим делать выводы об окружающем мире, основываясь на наблюдениях.

Замечание.

Математическая статистика занимается исследованием данных вероятностной природы.

Исследованием же данных невероятностной природы занимается машинное обучение.

Тем не менее, такая классификация отнюдь не является строгой, нередко то и другое относят к задачам статистики.

Примеры статистических задач.

1. Оценка вероятности успеха:

   $X_1, \ldots, X_n$ — независимые испытания

   $\mathrm{P}(X_i = 0) = p,\ \mathrm{P}(X_i = 1) = 1 - p$

   Как оценить p?

2. Проверка биологических, медицинских и прочих гипотез:

   Хорошим примером являются опыты Менделя по скрещиванию чистых линий гороха.

   Так, в процессе скрещивания во втором поколении у него получилось 8506 желтых горошин и 23174 зеленых горошин.

   Выдвинутая же им гипотеза утверждала, что соотношение фенотипов должно быть $1 : 3$.

   В данном случае задача математической статистики заключается в том, чтобы понять, насколько полученные отличия существенны.

3. GWAS — genome-wide association study.

   Известно, что существует взаимосвязь между генотипом и фенотипом. Более того, фенотип определяется по большей части (хотя и не только) генотипом.

   Задача (неформально) — научиться по генотипу предсказывать фенотип.

4. Проверка авторства текстов.

   К примеру, в качестве одного из критериев можно рассматривать распределение слов по длине в различных текстах автора.

1.2. Математическая постановка задач статистики

***Определение* 1.1.**

$(\Omega, \mathcal{F}, \mathrm{P})$ — вероятностное пространство.

$X_1, \ldots, X_n$ — набор одинаково распределенных случайных величин.

$\{\mathrm{P}_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ — семейство вероятностных распределений.

P_θ заданы на том же множестве, что и X_i.

Распределение X_i — одно из P_θ (но мы не знаем, с каким θ).

Обычно X_i называют наблюдением, а сам набор $X_1, \ldots, X_n$ — выборкой.

***Определение* 1.2** (Задача оценивания параметров)**.**

$X_1, \ldots, X_n$ — наблюдения, $X_i \sim \mathrm{P}_\theta$, $\theta \in \Theta$.

Требуется построить:

a) функцию $T(X_1, \ldots, X_n)$, которая хорошо бы приближала θ

b) интервал $(T_1(X_1, \ldots, X_n), T_2(X_1, \ldots, X_n))$, т.ч. $\mathrm{P}(T_1 < \theta < T_2) \geqslant \gamma$ (считаем, что $\theta \in \mathbb{R}$)

***Определение* 1.3** (Задача проверки статистических гипотез)**.**

$\Theta = \Theta_0 \sqcup \Theta_1$

H_0 и H_1 — гипотезы

$H_0 : \ \theta \in \Theta_0$

$H_1 : \ \theta \in \Theta_1$

Требуется различить H_0 и H_1 на $X_1, \ldots, X_n$, т.е. понять, какая из гипотез верна.

В общем случае $\Theta = \bigsqcup\limits_{i=1}^{n} \Theta_i$ и $H_i : \ \theta \in \Theta_i$.

Пример.

$X_1, \ldots, X_n$ — наблюдения, $X_i \sim \mathrm{F}$

$\mathrm{F}_n(t) = \frac{1}{n} \cdot \#\{i : X_i < t\}$ — эмпирическая функция распределения; доля наблюдений, принимающих значение $< t$.

Запишем иначе:

$\mathrm{F}_n(t) = \frac{1}{n} \cdot \sum\limits_{i=1}^{n} \mathbb{1}\{X_i < t\}$

Отсюда получаем матожидание:

$\mathrm{E}\,\mathrm{F}_n(t) = \frac{1}{n} \sum\limits_{i=1}^{n} \mathrm{P}(X_i < t) = \mathrm{F}(t)$

Получили, что в среднем мы не промахиваемся.

Посчитаем также дисперсию:

$\mathrm{D}\,\mathrm{F}_n(t) = \frac{1}{n^2} \sum\limits_{i=1}^{n} \mathrm{D}\,\mathbb{1}\{X_i < t\} = \frac{1}{n^2} \sum\limits_{i=1}^{n} \mathrm{F}_{X_i}(t)(1 - \mathrm{F}_{X_i}(t)) = \frac{1}{n}\,\mathrm{F}(t)(1 - \mathrm{F}(t))$

Таким образом, из закона больших чисел получаем, что $\mathrm{F}_n(t) \xrightarrow{\mathrm{P}} \mathrm{F}(t)$.

Теорема 1.1 (Гливенко-Кантелли)**.**

$$\mathrm{P}\left(\sup_{t \in \mathbb{R}} |\mathrm{F}_n(t) - \mathrm{F}(t)| \to 0\right) = 1$$

Доказательство. Без доказательства. □

Замечание.

Получили, что эмпирическая функция распределения сходится по вероятности к функции распределения рассматриваемой случайной величины.

Таким образом, если вспомнить главную задачу математической статистики, можно подумать, что мы научились ее решать.

Однако в действительности скорость сходимости эмпирической функции очень плохая, потому так просто статистические задачи не решаются.

Пример.

Научимся оценивать матожидание.

$X_1, \ldots, X_n$ — наблюдения, $m = \mathrm{E}\, X_i$

Как оценить m?

Естественная мысль — оценивать выборочным средним:

$\overline{X} = \frac{X_1 + \ldots + X_n}{n}$

Посчитаем матожидание для $\overline{X}$:

$\mathrm{E}\, \overline{X} = \frac{m + \ldots + m}{n} = m$

Получили, что в среднем не промахиваемся.

Если при всем этом $\mathrm{E}\, X_i^2 < +\infty$, то $\overline{X} \xrightarrow{\mathrm{P}} m$ (можно воспользоваться законом больших чисел).

Кроме того, если $\mathrm{E}\, X_i^2 < +\infty$, то есть асимптотическая нормальность, т.е. $\sqrt{n}\frac{\overline{X} - m}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1)$.

Аналогично, если хотим оценить $m_k = \mathrm{E}\, X_i^k$, то рассматриваем $\frac{X_1^k + \ldots + X_n^k}{n}$.

Напоминание.

Сходимость по распределению $X_n \xrightarrow{d} X$ — просто поточечная сходимость F_{X_n} к функции распределения F_X (в случае, если последняя непрерывна).

2. Квантили

2.1. Определение квантили

***Определение* 2.1.**

X — некоторая случайная величина, $X \sim \mathrm{F}$

$p \in (0,1)$ и $\exists t:\ \mathrm{F}(t) = p$ и в окрестности t функция F непрерывна и строго возрастает

Квантиль порядка p — $\zeta_p = \mathrm{F}^{-1}(p)$

Замечание.

Возникает вопрос, что делать, если функция распределения не непрерывна (а именно с такими нам приходится иметь дело в математической статистике).

Отсюда цель: научиться оценивать квантили.

***Определение* 2.2.**

$X_1, \ldots, X_n$ — наблюдения.

Рассмотрим случайную величину ξ, принимающую значения $X_1, \ldots, X_n$ с вероятностью $\frac{1}{n}$.

Тогда наши выборочные моменты — это просто матожидание для ξ.

Действительно, ранее мы уже поняли, что $\mathrm{E}\, X$ можно оценивать выборочным средним $\overline{X}$. А оно в точности совпадает с $\mathrm{E}\,\xi$.

Аналогично, $\mathrm{F}(t)$ мы научились оценивать через эмпирическую функцию распределения $\mathrm{F}_n(t)$, которая также совпадает с функцией распределения F_ξ.

Тогда выборочная квантиль порядка p — это $Z_p = X_{([np]+1)}$, т.е. порядковая статистика с индексом $[np]+1$.

2.2. Оценка генеральных квантилей

Теорема 2.1.

$X_1, \ldots, X_n$ — наблюдения, $X_i \sim \mathrm{F}$, $f = \mathrm{F}'$

$p \in (0,1)$, ζ_p — генеральная квантиль порядка p и $f(\zeta_p) \neq 0$

Тогда выборочная квантиль будет асимптотически нормальной, т.е. $\sqrt{n}(Z_p - \zeta_p) \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{f^2(\zeta_p)}\right)$.

Доказательство.

$k := [np] + 1$

$\mathrm{P}\left(\sqrt{n}(Z_p - \zeta_p) < t\right) = \mathrm{P}\left(Z_p < \zeta_p + \frac{t}{\sqrt{n}}\right) = \mathrm{P}\left(X_{(k)} < \zeta_p + \frac{t}{\sqrt{n}}\right) =$

$= \mathrm{P}\left(\text{хотя бы } k \text{ наблюдений } < \zeta_p + \frac{t}{\sqrt{n}}\right) = \mathrm{P}\left(\sum_{i=1}^n \mathbb{1}\left\{X_i < \zeta_p + \frac{t}{\sqrt{n}}\right\} \geqslant k\right)$

В левой части мы получили сумму независимых одинаково распределенных случайных величин, можем использовать центральную предельную теорему. Для этого посчитаем матожидание и дисперсию:

$\mathrm{E}\,\mathbb{1}\left\{X_i < \zeta_p + \frac{t}{\sqrt{n}}\right\} = \mathrm{P}\left(X_i < \zeta_p + \frac{t}{\sqrt{n}}\right) = \mathrm{F}\left(\zeta_p + \frac{t}{\sqrt{n}}\right) \overset{\text{Тейлор}}{=} \mathrm{F}(\zeta_p) + \frac{t}{\sqrt{n}}\mathrm{F}'(\zeta_p) + o\left(\frac{t}{\sqrt{n}}\right)$

Вспомнив, что $\mathrm{F}(\zeta_p) = p$ и $\mathrm{F}'(\zeta_p) = f(\zeta_p)$, получаем:

$\mathrm{E}\,\mathbb{1}\left\{X_i < \zeta_p + \frac{t}{\sqrt{n}}\right\} = p + \frac{t}{\sqrt{n}} f(\zeta_p) + o\left(\frac{t}{\sqrt{n}}\right)$

Посчитаем дисперсию:

$\mathrm{D}\,\mathbb{1}\left\{X_i < \zeta_p + \frac{t}{\sqrt{n}}\right\} = \mathrm{F}\left(\zeta_p + \frac{t}{\sqrt{n}}\right)\left(1 - \mathrm{F}\left(\zeta_p + \frac{t}{\sqrt{n}}\right)\right) \overset{\text{Тейлор}}{=} \left(\mathrm{F}(\zeta_p) + o(1)\right)\left(1 - \mathrm{F}(\zeta_p) + o(1)\right)$

После подстановки аналогично получаем:

$\mathrm{D}\,\mathbb{1}\left\{X_i < \zeta_p + \frac{t}{\sqrt{n}}\right\} = p(1-p)(1+o(1))$

Положим $S_n = \sum_{i=1}^{n} \mathbb{1}\left\{X_i < \zeta_p + \frac{t}{\sqrt{n}}\right\}$. Кроме того, введем $\Phi(t)$ — функцию распределения для $\mathcal{N}(0,1)$. Тогда по ЦПТ имеем:

$$\mathrm{P}\left(S_n \geqslant k\right) = 1 - \mathrm{P}\left(S_n < k\right) = 1 - \mathrm{P}\left(\frac{S_n - \mathrm{E}\,S_n}{\sqrt{\mathrm{D}\,S_n}} < \frac{k - \mathrm{E}\,S_n}{\sqrt{\mathrm{D}\,S_n}}\right) \overset{\text{ЦПТ}}{=} 1 - \Phi\left(\frac{k - n\left(p + \frac{t}{\sqrt{n}} f(\zeta_p) + o\left(\frac{t}{\sqrt{n}}\right)\right)}{\sqrt{p(1-p)(1+o(1))\cdot n}} + o(1)\right) =$$

$$= 1 - \Phi\left(\frac{[np] + 1 - np - \sqrt{n}\cdot t f(\zeta_p) + o\left(\sqrt{n}\right)}{\sqrt{p(1-p)}\cdot\sqrt{n}\cdot(1+o(1))}\right) + o(1) = 1 - \Phi\left(\frac{-\sqrt{n}\cdot t f(\zeta_p) + o(\sqrt{n})}{\sqrt{p(1-p)}\cdot\sqrt{n}\cdot(1+o(1))}\right) + o(1) =$$

$$= 1 - \Phi\left(\frac{-t f(\zeta_p) + o(1)}{\sqrt{p(1-p)}} + o(1)\right) + o(1) = 1 - \Phi\left(\frac{-t f(\zeta_p)}{\sqrt{p(1-p)}}\right) + o(1) = \Phi\left(\frac{t f(\zeta_p)}{\sqrt{p(1-p)}}\right) + o(1)$$

Получили, что $\mathrm{P}\left(\sqrt{n}(Z_p - \zeta_p) < t\right) \to \Phi\left(\frac{t f(\zeta_p)}{\sqrt{p(1-p)}}\right) = \Phi\left(\frac{t}{\sigma}\right)$, где $\sigma = \frac{\sqrt{p(1-p)}}{f(\zeta_p)}$.

А это в точности то, что мы и хотели. □

Утверждение 2.2.

$\sqrt{n}\left(T_n - \theta\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \implies T_n - \theta \xrightarrow{\mathrm{P}} 0$

Доказательство.

Выпишем условие сходимости по вероятности:

$\forall \varepsilon > 0: \ \mathrm{P}\left(|T_n - \theta| > \varepsilon\right) \to 0$

$\mathrm{P}\left(|T_n - \theta| > \varepsilon\right) = 1 - \mathrm{P}\left(\sqrt{n}\cdot|T_n - \theta| \leqslant \sqrt{n}\cdot\varepsilon\right) = 1 - \mathrm{P}\left(-\sqrt{n}\cdot\varepsilon \leqslant \sqrt{n}\cdot(T_n - \theta) \leqslant \sqrt{n}\cdot\varepsilon\right)$

Тогда, воспользовавшись сходимостью по распределению, имеем:

$\mathrm{P}\left(|T_n - \theta| > \varepsilon\right) = 1 - \left(\Phi\left(\frac{\sqrt{n}\cdot\varepsilon}{\sigma}\right) - \Phi\left(-\frac{\sqrt{n}\cdot\varepsilon}{\sigma}\right) + o(1)\right) \underset{n\to\infty}{\to} 0$ □

3. Случайные векторы

3.1. Гауссовский вектор

***Определение* 3.1.**
Стандартный гауссовский вектор — $(X_1, \ldots, X_n)$, где X_i — независимые случайные величины из $\mathcal{N}(0,1)$.

Вектор Y имеет гауссовское распределение, если представим в виде $Y = a + LX$, где $a \in \mathbb{R}^n$, а $L : \mathbb{R}^n \to \mathbb{R}^n$ — линейное отображение.

***Определение* 3.2.**
Характеристическая функция случайного вектора: $\varphi_\xi(t) = \mathrm{E}\, e^{i\langle t,\xi\rangle}$

Утверждение 3.1.
Поймем, как выглядит характеристическая функция стандартного гауссовского вектора:

$\varphi_X(t) = \mathrm{E}\, e^{i\langle t,X\rangle} = \mathrm{E}\, e^{i\sum_{j=1}^n t_j X_j} = \prod_{j=1}^n \mathrm{E}\, e^{it_jX_j} = \prod_{j=1}^n e^{-\frac{1}{2}\cdot t_j^2} = e^{-\frac{1}{2}\langle t,t\rangle}$

Для произвольного же вектора с гауссовским распределением имеем:

$\varphi_Y(t) = \mathrm{E}\, e^{i\langle t,Y\rangle} = \mathrm{E}\, e^{i\langle t,a+LX\rangle} = \mathrm{E}\, e^{i\langle t,a\rangle}\cdot e^{i\langle t,LX\rangle} = e^{i\langle t,a\rangle}\cdot \mathrm{E}\, e^{i\langle t,LX\rangle} = e^{i\langle t,a\rangle}\cdot \mathrm{E}\, e^{i\langle L^*t,X\rangle}$

Тогда, воспользовавшись видом характеристической функции для стандартного гауссовского вектора, получаем характеристическую функцию для произвольного вектора Y, имеющего гауссовское распределение:

$\varphi_Y(t) = e^{i\langle t,a\rangle}\cdot e^{-\frac{1}{2}\langle L^*t,L^*t\rangle} = e^{i\langle t,a\rangle}\cdot e^{-\frac{1}{2}\langle LL^*t,t\rangle}$, где $Y = a + LX$

Утверждение 3.2.
Поймем, как определяется плотность для гауссовского вектора

Положим $Y = a + LX$.

Кроме того, потребуем обратимость L, т.к. иначе Y живет на некотором подпространстве и тогда плотности никакой нет.

Тогда имеем:

$$\mathrm{P}(X \in A) = \int\limits_A f_X(t)\, dt = \int\limits_A f_{X_1}(t_1)\cdot\ldots\cdot f_{X_n}(t_n)\, dt = \int\limits_A \tfrac{1}{(2\pi)^{n/2}}\cdot e^{-\frac{1}{2}\langle t,t\rangle}\, dt$$

Поймем, что происходит с Y.

Понятно, что сдвиг на плотность не влияет, потому будем считать, что $a = 0$.

$$\mathrm{P}(LX \in A) = \int\limits_{L^{-1}A} f_X(t)\, dt = \int\limits_A f_X(L^{-1}t)\cdot|\det L^{-1}|\, dt \implies f_Y(t) = f_X(L^{-1}t)\cdot|\det L^{-1}|$$

В общем же случае вид плотности будет следующий:

$f_Y(t) = f_{a+LX}(t) = f_{LX}(t-a) = f_X(L^{-1}(t-a))\cdot|\det L^{-1}|$

Утверждение 3.3.
Посмотрим на матожидание стандартного гауссовского вектора X:

$\mathrm{E}\, X = 0$, т.к. $X_i \sim \mathcal{N}(0,1)$

$\mathrm{E}[XX^T]$ — ковариационная матрица (ее элементы как раз имеют вид $\mathrm{E}[X_iX_j]$), причем эта матрица будет единичной, т.к. дисперсия равна единице.

Поймем, что происходит с произвольным гауссовским вектором $Y = a + LX$:

$\mathrm{E}\, Y = \mathrm{E}[a + LX] = a + \mathrm{E}[LX] = a$

$\mathrm{E}[(Y - a)(Y - a)^T] = \mathrm{E}[LXX^TL^T] = L(\mathrm{E}[XX^T])L^T = LL^T$

Если же нам загадали некоторый гауссовский случайный вектор с ковариационной матрицей R, то построить сам вектор мы можем путем извлечения корня из R (это возможно сделать в силу того, что R симметричная и положительно определенная).

Таким образом, получив матрицу $R^{\frac{1}{2}}$, мы в то же время получаем линейное отображение L, а значит — и сам вектор Y (с точностью до сдвига).

4. Статистические оценки

4.1. Требования, предъявляемые к оценкам

***Определение* 4.1.**
$X_1, \ldots, X_n$ — выборка, $X_i \sim \mathrm{P}_\theta,\ \theta \in \Theta$

$T = T(X_1, \ldots, X_n)$ — произвольная оценка.

Определим критерии "хорошести" нашей оценки:

1. Несмещенность: $\forall\theta\ \ \mathrm{E}_\theta\, T = \theta$
2. Асимптотическая несмещенность: $\forall\theta\ \ \mathrm{E}_\theta\, T_n \to \theta$
3. Состоятельность: $\forall\theta\ \ T_n \overset{\mathrm{P}_\theta}{\to} \theta$
4. Сильная состоятельность: $\forall\theta\ \ T_n \overset{\text{п.н.}}{\to} \theta$
5. Асимптотическая нормальность: $\sqrt{n}(T_n - \theta) \overset{d}{\to} \mathcal{N}(0, \sigma^2(\theta))$
6. Эффективность: T_1 "эффективнее" T_2, если $\forall\theta\ \ \mathrm{E}_\theta(T_1 - \theta)^2 \leqslant \mathrm{E}_\theta(T_2 - \theta)^2$

Замечание.
Тем не менее, совсем необязательно, чтобы наши оценки удовлетворяли всем вышеописанным критериям. Даже отсутствие несмещенности как правило не делает оценку "плохой".

4.2. Лемма Фишера

Напоминание.
Матрица C является ортогональной, если выполняется $C^TC = CC^T = E$, т.е. если $C^T = C^{-1}$.

Утверждение 4.1.
$X = (X_1, \ldots, X_n)^T,\ X_i$ — независимые компоненты, $X_i \sim \mathcal{N}(0, \sigma^2)$.

C — ортогональная матрица.

Тогда $CX \overset{d}{=} X$.

Доказательство.
$\mathrm{E}\, e^{i\langle t, X\rangle} = e^{-\frac{\sigma^2}{2}\langle t,t\rangle}$

$\mathrm{E}\, e^{i\langle t, CX\rangle} = e^{-\frac{\sigma^2}{2}\langle C^Tt, C^Tt\rangle} = e^{-\frac{\sigma^2}{2}\langle t, CC^Tt\rangle} = e^{-\frac{\sigma^2}{2}\langle t,t\rangle}$

Получили, что характеристические функции X и CX совпадают. А значит, по теореме единственности и сами распределения совпадают. □

***Определение* 4.2** (Распределение хи-квадрат)**.**
$X_1, \ldots, X_n$ — независимые случайные величины, $X_i \sim \mathcal{N}(0, 1)$.

Тогда распределение хи-квадрат определяется как $\chi^2_n = X_1^2 + \ldots + X_n^2$.

***Определение* 4.3** (Распределение Стьюдента)**.**
$X, X_1, \ldots, X_n$ — независимые случайные величины, $X_i \sim \mathcal{N}(0, 1)$.

Тогда распределение Стьюдента определяется как $T_n = \frac{X}{\sqrt{\frac{1}{n}(X_1^2 + \ldots + X_n^2)}}$.

***Определение* 4.4.**

Пусть $X_1, \ldots, X_n$ — независимые наблюдения, $\overline{X} = \frac{X_1+\ldots+X_n}{n}$ — выборочное среднее.

Тогда $S^2 = \frac{1}{n}\sum\limits_{k=1}^{n}(X_k - \overline{X})^2$ называется выборочной дисперсией.

В частности, если раскрыть все скобки в вышеприведенной сумме, то можно получить следующее представление для выборочной дисперсии, аналогичное представлению для обычной дисперсии:

$$S^2 = \left(\frac{1}{n}\sum_{k=1}^{n} X_k^2\right) - \overline{X}^2 \iff \mathrm{D}\,\xi = \mathrm{E}\,\xi^2 - (\mathrm{E}\,\xi)^2$$

Лемма Фишера.

$X_1, \ldots, X_n$ — независимые наблюдения, $X_i \sim \mathcal{N}(\theta, \sigma^2)$.

$\overline{X} = \frac{X_1+\ldots+X_n}{n}$

$S^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^2$

Тогда:

1) $\sqrt{n}\frac{\overline{X}-\theta}{\sigma} \sim \mathcal{N}(0,1)$

2) $\frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$

3) $\overline{X}$ и S^2 — независимы

4) $\sqrt{n-1}\frac{\overline{X}-\theta}{S} \sim T_{n-1}$

Доказательство.

1. $\overline{X}$ — линейная комбинация независимых нормальных случайных величин, а значит имеет нормальное распределение.

   В этом легко убедиться, если взглянуть на характеристическую функцию $\overline{X}$.

   В частности, $\overline{X} \sim \mathcal{N}\left(\mathrm{E}\,\overline{X}, \mathrm{D}\,\overline{X}\right) = \mathcal{N}\left(\theta, \frac{\sigma^2}{n}\right)$.

   А значит, $\frac{\overline{X}-\theta}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1)$.

2, 3. Можно считать, что $\theta = 0$.

   Действительно, если мы все наблюдения сдвинем на θ, то S^2 никак не изменится.

   Тем не менее, $\overline{X}$ изменится. Но т.к. изменение происходит на константу, то на независимость S^2 и $\overline{X}$ это никак не повлияет.

   Рассмотрим тогда вектор $X = (X_1, \ldots, X_n)^T$, где X_k — независимы, $X_k \sim \mathcal{N}(0, \sigma^2)$.

   И рассмотрим ортогональную матрицу C следующего вида:

$$C = \begin{pmatrix} \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{pmatrix}$$

   Здесь важно, чтобы первая строка этой матрицы имела ровно такой вид. В частности, можно заметить, что такая матрица всегда существует.

   Тогда имеем $Y = CX \stackrel{d}{=} X$.

   Попытаемся выразить выборочное среднее и выборочную дисперсию через Y_k:

   $Y_1 = \frac{1}{\sqrt{n}}(X_1 + \ldots + X_n) = \sqrt{n}\cdot\overline{X} \implies \overline{X} = \frac{Y_1}{\sqrt{n}}$

Распишем теперь выборочную дисперсию. Воспользуемся тем фактом, что C не меняет длину вектора:

$S^2 = \frac{1}{n}\sum\limits_{k=1}^{n} X_k^2 - \overline{X}^2 = \frac{1}{n}\sum\limits_{k=1}^{n} Y_k^2 - \frac{Y_1^2}{n} = \frac{1}{n}\sum\limits_{k=2}^{n} Y_k^2$

Таким образом, мы доказали оба пункта.

А именно, второй пункт получается элементарной подстановкой вместо S^2 полученного результата.

Третий же пункт следует из того, что $\overline{X}$ и S^2 выражаются через разные Y_k, а потому независимы.

4. $T_n = \frac{X}{\sqrt{\frac{1}{n}(X_1^2+\ldots+X_n^2)}}$, где $X, X_i \sim \mathcal{N}(0,1)$ и независимы.

   Положим $X = \frac{\overline{X}-\theta}{\sqrt{\sigma^2/n}} = \sqrt{n}\frac{\overline{X}-\theta}{\sigma}$.

   Тогда имеем:

   $T_{n-1} \sim \frac{X}{\sqrt{\frac{1}{n-1}\chi^2_{n-1}}} = \frac{\sqrt{n}\frac{\overline{X}-\theta}{\sigma}}{\sqrt{\frac{1}{n-1}\cdot\frac{nS^2}{\sigma^2}}} = \sqrt{n-1}\frac{\overline{X}-\theta}{S}$

□

4.3. Построение оценок. Метод моментов

Мотивация.

До сих пор мы старались по имеющимся у нас наблюдениям восстановить информацию о распределении.

К примеру, мы уже выяснили, что функция распределения неплохо приближается эмпирической функцией распределения, а матожидание — выборочным средним.

Более того, таким же образом можно пробовать оценивать и различные функционалы (к примеру, интегральные):

$$\int g(x)\, d\,\mathrm{F}(x) = \int g(x)\, d\,\mathrm{F}_n(x) = \frac{1}{n}\sum_{k=1}^{n} g(X_k)$$

Хотелось бы применить текущие знания для построения некоторого метода, который бы позволил хорошо приближать интересующие нас параметры.

Задача.

Вернемся к общей задаче. Будем рассматривать простой случай, когда все наблюдения и параметры существуют в $\mathbb{R}$.

Пусть $X_1, \ldots, X_n \sim P_\theta$ — наблюдения, $\theta \in \Theta$.

Как и всегда, сам параметр θ нам неизвестен, но мы хотим его оценить.

Решение.

Рассмотрим некоторую функцию $g : \mathbb{R} \to \mathbb{R}$, такую что генеральный момент $\mathrm{E}_\theta\, g(X) =: h(\theta)$ — достаточно хорошая функция. На текущий момент будем требовать, чтобы h была обратима.

Предположим при этом, что наш генеральный момент $\mathrm{E}_\theta\, g(X)$ достаточно хорошо приближается выборочным. На самом деле, в идеальном случае мы бы имели следующее равенство:

$$\frac{1}{n}\sum_{k=1}^{n} g(X_k) = \int g(x)\, d\,\mathrm{F}_n(x) = \int g(x)\, d\,\mathrm{F}_\theta(x) = \mathrm{E}_\theta\, g(X) = h(\theta)$$

Отсюда появляется идея оценивать параметр как $\overline{\theta} = h^{-1}\left(\frac{1}{n}\sum\limits_{k=1}^{n} g(X_k)\right)$.

Такая идея действительно кажется вполне разумной. Тем не менее, требуется исследовать свойства, которыми обладает такая оценка параметров, а также проверить, как хорошо она работает на каких-нибудь известных распределениях.

Замечание.

В качестве $g(x)$ обычно выбирают функции какого-нибудь простого и удобного вида. Самый типичный вариант — $g(x) = x^k$, поскольку в таком случае нам приходится работать с моментами $\mathrm{E}\, X^k$, а о них мы уже многое знаем.

Замечание.

Если параметр лежит в пространстве большей размерности (т.е. $\Theta \subset \mathbb{R}^d$, где $d > 1$), то такой подход по-прежнему имеет право на существование. Однако теперь, чтобы получить $\overline{\theta}$, нам потребуется по крайней мере d различных уравнений.

Пример.

1. Рассмотрим нормальное распределение $\mathcal{N}(\theta, \sigma^2)$.

   В данном случае мы имеем два неизвестных параметра. Для их оценки попробуем выписать первые два момента, т.е. выберем $g_1(x) = x$ и $g_2(x) = x^2$:

   $$\begin{cases} \mathrm{E}_{(\overline{\theta},\overline{\sigma})}\, X = \frac{1}{n}\sum\limits_{k=1}^{n} X_k \\ \mathrm{E}_{(\overline{\theta},\overline{\sigma})}\, X^2 = \frac{1}{n}\sum\limits_{k=1}^{n} X_k^2 \end{cases} \overset{\mathrm{D}\, X = \mathrm{E}\, X^2 - (\mathrm{E}\, X)^2}{\Longleftrightarrow} \begin{cases} \overline{\theta} = \overline{X} \\ \overline{\sigma}^2 + \overline{\theta}^2 = \frac{1}{n}\sum\limits_{k=1}^{n} X_k^2 \end{cases} \iff \begin{cases} \overline{\theta} = \overline{X} \\ \overline{\sigma}^2 = \left(\frac{1}{n}\sum\limits_{k=1}^{n} X_k^2\right) - \overline{X}^2 \end{cases}$$

   Получили, что параметры нормального распределения можно оценить как $\overline{X}$ и S^2.

   Но мы уже знаем, что эти оценки хорошо приближают матожидание и дисперсию произвольного случайного распределения.

   А так как параметрами нормального распределения как раз являются его матожидание и дисперсия, то мы действительно получили неплохую оценку его параметров.

   Значит, есть надежда, что этот метод действительно работает неплохо.

2. Другой пример — распределение Коши.

   Напомним, что распределение Коши — это класс абсолютно непрерывных распределений.

   Плотность таких распределений имеет вид $f(x) = \frac{1}{\pi}\left[\frac{1}{(x-\theta)^2+1}\right]$, где θ — параметр сдвига.

   Это распределение примечательно тем, что оно не имеет математического ожидания. А потому оценивать это распределение с помощью первого момента просто не получится.

Теперь нам остается лишь проверить свойства, которыми обладают выведенные нами оценки.

Но прежде чем приступать к их рассмотрению, докажем следующее утверждение:

Утверждение 4.2.

$\sqrt{n}\frac{T_n - a}{\sigma} \to \mathcal{N}(0, 1)$

f — непрерывная функция, $f'(a) \neq 0$

Тогда $f(T_n)$ асимптотически нормальна.

Доказательство.

$f(T_n) = f(a) + (T_n - a)f'(a) + o(T_n - a)$

$\sqrt{n}\,(f(T_n) - f(a)) = \sqrt{n}(T_n - a)f'(a) + o\,(\sqrt{n}(T_n - a))$

$\sqrt{n}(T_n - a) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$

$o\,(\sqrt{n}(T_n - a)) = o(1) \cdot \sqrt{n}(T_n - a) \xrightarrow{d} 0$

Т.е. $\sqrt{n}\,(f(T_n) - f(a)) \xrightarrow{d} \mathcal{N}(0, \sigma^2 \cdot (f'(a))^2)$ □

Замечание.

Важно, что $f'(a)$ не обращается в ноль, поскольку иначе пришлось бы расписывать последующие члены в ряде Тейлора.

Свойства.

Поймем, что происходит со свойствами таких оценок:

1. Несмещенность.

   Хотим проверить равенство $\mathrm{E}\,\overline{\theta} \stackrel{?}{=} \theta$.

   Положим $Z := \frac{1}{n}\sum\limits_{k=1}^{n} g(X_k)$. Тогда имеем:

   $\mathrm{E}\,\overline{\theta} = \mathrm{E}\,h^{-1}\left(\frac{1}{n}\sum\limits_{k=1}^{n} g(X_k)\right) = \mathrm{E}\,h^{-1}(Z)$

   $\theta = h^{-1}(h(\theta)) \stackrel{\text{def.}}{=} h^{-1}(\mathrm{E}\,g(X)) = h^{-1}\left(\frac{1}{n}\sum\limits_{k=1}^{n}\mathrm{E}\,g(X_k)\right) = h^{-1}\left(\mathrm{E}\left[\frac{1}{n}\sum\limits_{k=1}^{n} g(X_k)\right]\right) = h^{-1}(\mathrm{E}\,Z)$

   Таким образом, требуется проверить равенство $h^{-1}(\mathrm{E}\,Z)$ и $\mathrm{E}\,h^{-1}(Z)$.

   Но если h^{-1} строго выпукла или строго вогнута (а такая ситуация вполне обычная), то по неравенству Йенсена получается строгое неравенство.

   Значит, несмещенность в общем случае может и не существовать.

2. Состоятельность.

   Если $\mathrm{D}\,g(X) < \infty$ и h^{-1} непрерывна, то получаем:

   $\frac{1}{n}\sum\limits_{k=1}^{n} g(X_k) \xrightarrow{\mathrm{P}} \mathrm{E}_\theta\,g(X) = h(\theta)$ (закон больших чисел)

   $\overline{\theta} = h^{-1}\left(\frac{1}{n}\sum\limits_{k=1}^{n} g(X_k)\right) \xrightarrow{\mathrm{P}} h^{-1}(h(\theta)) = \theta$

3. Асимптотическая нормальность.

   Если $\mathrm{D}\,g(X) < \infty$, а h^{-1} непрерывно дифференцируема и не обращается в ноль, то получаем:

   $\dfrac{\frac{1}{n}\sum\limits_{k=1}^{n} g(X_k) - \mathrm{E}_\theta\,g(X)}{\sqrt{\frac{\mathrm{D}\,g(X)}{n}}} \xrightarrow{d} \mathcal{N}(0,1)$ (центральная предельная теорема)

   Тогда, применяя к нашей оценке h^{-1}, по доказанному выше утверждению получаем:

   $\sqrt{n}(\overline{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \mathrm{D}[g(X)] \cdot ((h^{-1})' \circ h(\theta))^2\right)$

4.4. Построение оценок. Метод максимального правдоподобия

Мотивация.

Мы уже выяснили, что метод моментов умеет неплохо приближать реальные оценки. Тем не менее, он требует от нас вычисления обратной функции, что в случае произвольного распределения очень сложно и нетривиально. В некоторых же ситуациях он и вовсе не способен выдать никакого решения (к примеру, если матожидание отсутствует).

Потому хотелось бы рассмотреть метод получения оценок, который избавил бы нас по крайней мере от подобных проблем.

Кроме того, мы также покажем, что рассматриваемый далее метод дает асимптотически эффективные оценки, в то время как про эффективность метода моментов в общем случае сложно что-то сказать.

Напоминание.

Распределение P_θ абсолютно непрерывно относительно меры μ (обозначается как $\mathrm{P}_\theta \ll \mu$), если $\mu(A) = 0 \implies \mathrm{P}_\theta(A) = 0$.

***Определение* 4.5.**

$X_1, \ldots, X_n \sim \mathrm{P}_\theta,\ \theta \in \Theta$

$\mathrm{P}_\theta \ll \mu$, где μ — σ-конечная мера (к примеру, мера Лебега или считающая мера)

По теореме Радона-Никодима P_θ представима как $\mathrm{P}_\theta(A) = \int\limits_A f(x;\, \theta)\ \mu(dx)$ для некоторой f.

В данном случае, $f(x;\, \theta)$ — или плотность распределения (в случае меры Лебега), или вероятность $\mathrm{P}_\theta(X = x)$ события x (в случае считающей меры).

Тогда для произвольной заданной выборки $\mathbf{X} = (X_1, \ldots, X_n)$ можно определить функцию правдоподобия L (от likelihood):

$$L(\mathbf{X};\, \theta) := L(X_1, \ldots, X_n;\, \theta) = \prod_{i=1}^{n} f(X_i;\, \theta)$$

$\widehat{\theta}$ — оценка максимального правдоподобия (ОМП), если $L(\mathbf{X};\, \widehat{\theta}) = \max\limits_{\theta} L(\mathbf{X};\, \theta)$.

Соответственно, основная идея метода максимального правдоподобия — подобрать такой набор параметров $\widehat{\theta}$, при котором вероятность получения данной выборки максимальна, т.е. достигается максимум $L(\mathbf{X};\, \theta)$.

Кроме того, нередко вместо $L(\mathbf{X};\, \theta)$ максимизируют $\ln L(\mathbf{X};\, \theta)$, поскольку можно перейти от произведения к сумме.

Замечание.

На самом деле, когда речь идет о функции правдоподобия, подразумевается именно функция одного аргумента $L_{\mathbf{X}}(\theta)$, т.е. сама выборка фиксируется. Тем не менее, в литературе часто выборка $\mathbf{X}$ также указывается в качестве аргумента L, а потому дальше будет использоваться именно такая нотация.

Пример.

Пусть $X_1, \ldots, X_n \sim \mathrm{Poisson}(\theta),\ \ \theta \in \Theta$.

Выпишем функцию правдоподобия для нашей выборки:

$$L(\mathbf{X};\, \theta) = \prod_{i=1}^{n} \frac{\theta^{X_i}}{X_i!} \cdot e^{-\theta} = \frac{\theta^{X_1 + \ldots + X_n}}{X_1! \cdot \ldots \cdot X_n!} \cdot e^{-n\theta}$$

Хотим найти параметр θ, который бы максимизировал значение функции правдоподобия. Для этого достаточно продифференцировать $L(\mathbf{X};\,\theta)$ по параметру и найти ноль производной, после чего проверить, что он действительно является максимумом.

Тем не менее, искать производную суммы легче, чем производную произведения, а потому прологарифмируем нашу функцию:

$$\ln L(\mathbf{X};\,\theta) = -\ln(X_1! \cdot \ldots \cdot X_n!) + \sum_{i=1}^{n} X_i \cdot \ln\theta - n\theta$$

$$\frac{\partial}{\partial\,\theta} \ln L(\mathbf{X};\,\theta) = \frac{1}{\theta} \cdot \sum_{i=1}^{n} X_i - n = 0 \implies \widehat{\theta} = \overline{X}$$

$$\frac{\partial^2}{\partial\,\theta^2} \ln L(\mathbf{X};\,\theta) = -\frac{1}{\theta^2} \cdot \sum_{i=1}^{n} X_i < 0 \implies \widehat{\theta} = \arg\max_{\theta} L(\mathbf{X};\,\theta)$$

***Определение* 4.6.**

$X_1,\,\ldots,\,X_n \sim \mathrm{P}_\theta,\ \ \theta \in \Theta$

Информацией Фишера называется функция $I_n(\theta) = \mathrm{E}_\theta \left(\dfrac{\partial \ln f(X_1,\,\ldots,\,X_n;\,\theta)}{\partial\,\theta} \right)^2$

Замечание.

Так как наблюдения у нас независимы, $f(X_1,\,\ldots,\,X_n;\,\theta) = \prod\limits_{i=1}^{n} f(X_i;\,\theta) = L(\mathbf{X};\,\theta)$.

А потому информацию Фишера можно записать как $I_n(\theta) = \mathrm{E} \left(\dfrac{\partial \ln L(\mathbf{X};\,\theta)}{\partial\,\theta} \right)^2$.

Свойства информации Фишера.

1. Информация Фишера для выборки размера n представима как $I_n(\theta) = \displaystyle\int_{\mathcal{X}} \frac{(f'_\theta)^2}{f}\,d\mu$

   Доказательство.

   $$I_n(\theta) = \mathrm{E}_\theta \left(\frac{\partial \ln f(X_1,\,\ldots,\,X_n;\,\theta)}{\partial\,\theta} \right)^2 = \mathrm{E} \left(\frac{f'_\theta}{f} \right)^2 = \int_{\mathcal{X}} \left(\frac{f'_\theta}{f} \right)^2 \cdot f\,d\mu = \int_{\mathcal{X}} \frac{(f'_\theta)^2}{f}\,d\mu$$

   □

2. Пусть для плотности распределения $f(X_1,\,\ldots,\,X_n;\,\theta)$ выполняется равенство:

   $$\int_{\mathcal{X}} \frac{\partial}{\partial\,\theta} f(X_1,\,\ldots,\,X_n;\,\theta)\,d\mu = \frac{\partial}{\partial\,\theta} \int_{\mathcal{X}} f(X_1,\,\ldots,\,X_n;\,\theta)\,d\mu$$

   Тогда $I_n(\theta) = nI_1(\theta)$.

   Доказательство.
   Поскольку f — плотность случайного распределения, интеграл по ней равен 1 и, соответственно, никаким образом не зависит от параметра θ. Отсюда имеем:

   $$\mathrm{E} \left(\frac{\partial \ln f}{\partial\,\theta} \right) = \int_{\mathcal{X}} \frac{f'_\theta}{f} \cdot f\,d\mu = \int_{\mathcal{X}} \frac{\partial}{\partial\,\theta} f\,d\mu = \frac{\partial}{\partial\,\theta} \int_{\mathcal{X}} f\,d\mu = 0$$

Этим свойством мы воспользуемся дальше в доказательстве. На текущий же момент хотим по-другому записать определение информации Фишера.

Распишем чуть подробнее $\dfrac{\partial \ln f}{\partial \theta}$:

$$\frac{\partial \ln f(X_1, \ldots, X_n; \theta)}{\partial \theta} = \frac{\partial}{\partial \theta} \ln \prod_{i=1}^{n} f(X_i; \theta) = \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \ln f(X_i; \theta) = \sum_{i=1}^{n} \left(\frac{\partial}{\partial \theta} \ln f(X_i; \theta) \right)$$

Отсюда уже становится видно, каким образом свести $I_n(\theta)$ к $I_1(\theta)$. Просто выпишем определение информации Фишера, раскрыв квадрат суммы и воспользовавшись свойствами матожидания. Тогда получим:

$$I_n(\theta) = \mathrm{E} \left(\frac{\partial \ln f(X_1, \ldots, X_n)}{\partial \theta} \right)^2 = \mathrm{E} \left[\sum_{i=1}^{n} \left(\frac{\partial}{\partial \theta} \ln f(X_i; \theta) \right) \right]^2 =$$

$$= \mathrm{E} \left[\sum_{i=1}^{n} \left(\frac{\partial}{\partial \theta} \ln f(X_i; \theta) \right)^2 \right] + \mathrm{E} \left[\sum_{i \neq j} \left(\frac{\partial}{\partial \theta} \ln f(X_i; \theta) \cdot \frac{\partial}{\partial \theta} \ln f(X_j; \theta) \right) \right] =$$

$$= nI_1(\theta) + \sum_{i \neq j} \left(\mathrm{E} \left[\frac{\partial}{\partial \theta} \ln f(X_i; \theta) \right] \cdot \mathrm{E} \left[\frac{\partial}{\partial \theta} \ln f(X_j; \theta) \right] \right) = nI_1(\theta)$$

□

Напоминание.

Из курса математического анализа мы знаем, что если $\frac{\partial^k f}{\partial \theta^k}(x, \theta)$ существует и, к примеру, непрерывно дифференцируема по параметру θ, то верно равенство:

$$\frac{\partial^{k+1}}{\partial \theta^{k+1}} \int_{\mathcal{X}} f(x; \theta)\, d\mu = \int_{\mathcal{X}} \frac{\partial^{k+1}}{\partial \theta^{k+1}} f(x; \theta)\, d\mu \tag{1}$$

Тем не менее, вместо того, чтобы проверять непрерывную дифференцируемость функции, иногда может быть проще проверить такое равенство вручную.

Теорема 4.3.

$X_1, \ldots, X_n \sim \mathrm{P}_\theta$ — произвольная выборка, $\theta \in \Theta \subset \mathbb{R}$

$d\,\mathrm{P}_\theta = f(x, \theta)\, d\mu$

$0 \neq I_1(\theta) < +\infty$

$\dfrac{1}{n} \cdot \dfrac{\partial^3 \ln L}{\partial \theta^3}(X; \theta)$ — ограничено по θ

И пусть выполняется условие (1) для первых двух производных.

Тогда верны следующие утверждения:

1. ОМП состоятельна, т.е. $\widehat{\theta}_n \xrightarrow{\mathrm{P}_\theta} \theta \ \ \forall \theta \in \Theta$
2. ОМП асимптотически нормальна, т.е. $\sqrt{n} \left(\widehat{\theta}_n - \theta \right) \xrightarrow{d} \mathcal{N} \left(0, \dfrac{1}{I_1(\theta)} \right) \ \ \forall \theta \in \Theta$

Доказательство.

1. Без доказательства

2. Разложим $\dfrac{\partial \ln L}{\partial \theta}(\mathbf{X};\, \theta)$ в ряд Тейлора:

$$\frac{\partial \ln L}{\partial \theta}(\mathbf{X};\, \theta) = \frac{\partial \ln L}{\partial \theta}(\mathbf{X};\, \theta_0) + (\theta - \theta_0) \cdot \frac{\partial^2 \ln L}{\partial \theta^2}(\mathbf{X};\, \theta_0) + \frac{1}{2}(\theta - \theta_0)^2 \cdot \frac{\partial^3 \ln L}{\partial \theta^3}(\mathbf{X};\, \tilde{\theta}), \quad \tilde{\theta} \in (\theta_0,\, \theta)$$

Подставим теперь вместо θ нашу оценку максимального правдоподобия $\widehat{\theta}$. Поскольку в точке $\widehat{\theta}$ достигается максимум функции правдоподобия, значение производной в ней равно 0. Т.е. получаем:

$$0 = \frac{\partial \ln L}{\partial \theta}(\mathbf{X};\, \theta_0) + (\widehat{\theta} - \theta_0) \left(\frac{\partial^2 \ln L}{\partial \theta^2}(\mathbf{X};\, \theta_0) + \frac{1}{2}(\widehat{\theta} - \theta_0) \cdot \frac{\partial^3 \ln L}{\partial \theta^3}(\mathbf{X};\, \tilde{\theta}) \right), \quad \tilde{\theta} \in (\theta_0,\, \theta)$$

Попробуем вычленить из полученного выражения значение $\widehat{\theta} - \theta_0$, домноженное на $\sqrt{n}$.

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) = -\frac{\dfrac{1}{\sqrt{n}} \cdot \dfrac{\partial \ln L}{\partial \theta}(\mathbf{X};\, \theta_0)}{\dfrac{1}{n} \cdot \dfrac{\partial^2 \ln L}{\partial \theta^2}(\mathbf{X};\, \theta_0) + \dfrac{1}{2n}(\widehat{\theta} - \theta_0) \cdot \dfrac{\partial^3 \ln L}{\partial \theta^3}(\mathbf{X};\, \tilde{\theta})}, \quad \tilde{\theta} \in (\theta_0,\, \theta)$$

Теперь можем пробовать отдельно рассматривать каждую из частей полученного выражения. Начнем с первой производной:

$$\frac{\partial \ln L}{\partial \theta}(\mathbf{X};\, \theta_0) = \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \ln f(X_i;\, \theta_0) = \sum_{i=1}^{n} \frac{f'_\theta}{f}(X_i;\, \theta_0)$$

Получили сумму независимых случайных величин. Кроме того, нормировка $\frac{1}{\sqrt{n}}$ намекает нам на возможность использования центральной предельной теоремы. Остается посчитать для этого матожидание и дисперсию:

$$\mathrm{E}\left[\frac{f'_\theta}{f}(X_i;\, \theta_0)\right] = \int\limits_{\mathcal{X}} \frac{f'_\theta}{f} \cdot f(x;\, \theta_0)\, d\mu = \int\limits_{\mathcal{X}} f'_\theta(x;\, \theta_0)\, d\mu = 0$$

$$\mathrm{D}\left[\frac{f'_\theta}{f}(X_i;\, \theta_0)\right] = \mathrm{E}\left[\frac{f'_\theta}{f}(X_i;\, \theta_0)\right]^2 = \mathrm{E}\left[\frac{\partial \ln f(X_i;\, \theta_0)}{\partial \theta}\right]^2 = I_1(\theta_0)$$

Воспользовавшись ЦПТ, получаем:

$$\frac{1}{\sqrt{n}} \cdot \frac{\partial \ln L}{\partial \theta}(\mathbf{X};\, \theta_0) \xrightarrow{d} \mathcal{N}(0,\, I_1(\theta_0))$$

Мы теперь знаем, как распределен числитель. Остается лишь понять, что делать со знаменателем. Посчитаем для начала вторую производную:

$$\frac{\partial^2 \ln L}{\partial \theta^2}(\mathbf{X};\, \theta_0) = \frac{\partial}{\partial \theta}\left(\frac{\partial \ln L}{\partial \theta}(\mathbf{X};\, \theta_0)\right) = \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \frac{f'_\theta}{f}(X_i;\, \theta_0) = \sum_{i=1}^{n} \frac{f''_\theta \cdot f - (f'_\theta)^2}{f^2}(X_i;\, \theta_0)$$

Мы снова получили сумму независимых случайных величин. Нормировка $\frac{1}{n}$ намекает при этом на возможность использования закона больших чисел. Для этого посчитаем матожидание:

$$\mathrm{E}\left[\frac{f''_\theta \cdot f - (f'_\theta)^2}{f^2}(X_i;\, \theta_0)\right] = \mathrm{E}\left[\frac{f''_\theta}{f}(X_i;\, \theta_0)\right] - \mathrm{E}\left[\frac{f'_\theta}{f}(X_i;\, \theta_0)\right]^2 =$$

$$= \int\limits_{\mathcal{X}} \frac{\partial^2 f}{\partial \theta^2}(x;\, \theta_0)\, d\mu - I_1(\theta_0) = \frac{\partial^2}{\partial \theta^2} \int\limits_{\mathcal{X}} f(x;\, \theta_0)\, d\mu - I_1(\theta_0) = -I_1(\theta_0)$$

Получили, что первое слагаемое в знаменателе сходится по вероятности к $-I_1(\theta_0)$, т.е.

$$\frac{1}{n} \cdot \frac{\partial^2 \ln L}{\partial \theta^2}(\mathbf{X};\, \theta_0) \xrightarrow{\mathrm{P}_{\theta_0}} -I_1(\theta_0)$$

Осталось рассмотреть второе слагаемое знаменателя.

С одной стороны, из первого пункта теоремы мы знаем, что оценка $\widehat{\theta}$ состоятельна, т.е. $\widehat{\theta} - \theta \xrightarrow{\mathrm{P}_\theta} 0$. С другой стороны, оставшийся сомножитель у нас ограничен по условию. А значит, всё слагаемое по вероятности стремится к 0.

В итоге мы выяснили, что числитель по распределению сходится к $\mathcal{N}\,(0,\, I_1(\theta_0))$, а знаменатель сходится по вероятности к $-I_1(\theta_0)$. А значит получаем:

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} \frac{1}{I_1(\theta_0)}\mathcal{N}(0,\, I_1(\theta_0)) = \mathcal{N}\left(0,\, \frac{1}{I_1(\theta_0)}\right)$$

□

Замечание.

Приведенная выше теорема верна и для $\Theta \subset \mathbb{R}^k$. Доказательство ее отличается лишь использованием формулой Тейлора для многомерного случая.

Таким образом, мы показали, что оценки максимального правдоподобия состоятельны и асимптотически нормальны. Осталось показать, что, ко всему прочему, они нередко являются еще и асимптотически эффективными. Для этого докажем следующую теорему.

Теорема 4.4 (Неравенство Рао-Крамéра)**.**

$X_1, \ldots, X_n \sim \mathrm{P}_\theta$ — произвольная выборка, $\theta \in \Theta$

$0 \neq I_1(\theta) < \infty$

И пусть выполняется условие (1) для первой производной.

Рассмотрим произвольную оценку T параметра θ.

$\mathrm{E}_\theta\, T = \theta + b(\theta)$, где $b(\theta)$ — смещение (от bias)

Тогда $\mathrm{E}_\theta (T - \theta)^2 \geqslant \dfrac{(1 + b'(\theta))^2}{I_n(\theta)} + b^2(\theta)$

Доказательство.

Заметим, что чтобы получить значение $1 + b'(\theta)$, нам всего лишь необходимо продифференцировать наше матожидание по параметру.

Запишем тогда это значение в немного другом виде, воспользовавшись возможностью переставлять интеграл и производную местами:

$$1 + b'(\theta) = (\theta + b(\theta))' = (\mathrm{E}_\theta\, T)' = \frac{\partial}{\partial \theta} \int\limits_{\mathcal{X}} T \cdot f(x;\, \theta)\, d\mu = \int\limits_{\mathcal{X}} T \cdot f'_\theta(x;\, \theta)\, d\mu =$$

$$= \int\limits_{\mathcal{X}} T \cdot f'_\theta(x;\, \theta)\, d\mu - \mathrm{E}_\theta\, T \cdot 0 = \int\limits_{\mathcal{X}} (T - \mathrm{E}_\theta\, T) \cdot f'_\theta(x;\, \theta)\, d\mu$$

Попробуем теперь получить некоторую оценку сверху непосредственно на значение $(1 + b'(\theta))^2$. Для этого чуть иначе запишем полученный интеграл, а затем воспользуемся неравенством Коши-Буняковского:

$$(1 + b'(\theta))^2 = \left[\int\limits_{\mathcal{X}} (T - \mathrm{E}_\theta T) \cdot f'_\theta \, d\mu\right]^2 = \left[\int\limits_{\mathcal{X}} (T - \mathrm{E}_\theta T) \cdot \sqrt{f} \cdot \frac{f'_\theta}{\sqrt{f}} \, d\mu\right]^2 \overset{\text{КБ}}{\leqslant}$$

$$\overset{\text{КБ}}{\leqslant} \int\limits_{\mathcal{X}} \left((T - \mathrm{E}_\theta T) \cdot \sqrt{f}\right)^2 d\mu \cdot \int\limits_{\mathcal{X}} \left(\frac{f'_\theta}{\sqrt{f}}\right)^2 d\mu$$

Заметим, что левый интеграл — это в точности дисперсия $\mathrm{D}_\theta(T - \mathrm{E}_\theta T) = \mathrm{E}_\theta(T - \mathrm{E}_\theta T)^2$ отклонения нашей оценки. Правый интеграл — это информация Фишера. А потому имеем:

$$(1 + b'(\theta))^2 \leqslant \mathrm{E}_\theta(T - \mathrm{E}_\theta T)^2 \cdot I_n(\theta)$$

$$\frac{(1 + b'(\theta))^2}{I_n(\theta)} \leqslant \mathrm{E}_\theta(T - \mathrm{E}_\theta T)^2 = \mathrm{E}_\theta\left((T - \theta) - b(\theta)\right)^2 = \mathrm{E}_\theta(T - \theta)^2 - 2b(\theta)\,\mathrm{E}_\theta(T - \theta) + b^2(\theta)$$

Вспомнив теперь, что $\mathrm{E}_\theta(T - \theta) = b(\theta)$, получаем оценку снизу на $\mathrm{E}_\theta(T - \theta)^2$:

$$\frac{(1 + b'(\theta))^2}{I_n(\theta)} \leqslant \mathrm{E}_\theta(T - \theta)^2 - b^2(\theta) \implies \mathrm{E}_\theta(T - \theta)^2 \geqslant \frac{(1 + b'(\theta))^2}{I_n(\theta)} + b^2(\theta)$$

□

Следствие.

Если в условии теоремы $b(\theta) \equiv 0$, то $\mathrm{E}_\theta(T - \theta)^2 \geqslant \dfrac{1}{I_n(\theta)}$

Доказательство.

$$\mathrm{E}_\theta(T - \theta)^2 \geqslant \frac{(1 + b'(\theta))^2}{I_n(\theta)} + b^2(\theta) \overset{b,\, b' \equiv 0}{=} \frac{1}{I_n(\theta)}$$

□

Следствие.

Если в условии теоремы T — несмещенная оценка максимального правдоподобия, причем условие (1) выполняется вплоть до второй производной, то оценка T асимптотически эффективна.

Доказательство.

Мы знаем, что наша оценка несмещенная. А потому $\mathrm{D}(T - \theta)$ оценивается снизу как $I_n^{-1}(\theta)$.

Для удобства же будем сейчас рассматривать дисперсию величины $\sqrt{n}\,(T - \theta)$:

$$\mathrm{D}\left(\sqrt{n}\,(T - \theta)\right) = \mathrm{E}\left(\sqrt{n}\,(T - \theta)\right)^2 = n\,\mathrm{E}\,(T - \theta)^2 \geqslant n \cdot \frac{1}{I_n(\theta)} = \frac{1}{I_1(\theta)}$$

Вспомним, что оценка максимального правдоподобия является асимптотически нормальной:

$$\sqrt{n}\,(T - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I_1(\theta)}\right)$$

Таким образом получили, что дисперсия отклонения нашей оценки сходится к своей нижней границе. А значит, несмещенная оценка максимального правдоподобия является асимптотически эффективной.

□

4.5. Достаточные статистики

Предисловие.

Далее мы рассмотрим метод, который в некоторых ситуациях позволит нам получать эффективные оценки параметров.

Однако для начала нам потребуется вспомнить понятие условного математического ожидания, а также его свойства. Доказательства этих свойств по большей части будут опускаться, их всегда можно найти в соответствующей главе в конспекте по теории вероятностей.

***Определение* 4.7.**

Рассмотрим вероятностное пространство $(\Omega, \mathcal{F}, \mathrm{P})$.

Пусть X — произвольная случайная величина, $\mathcal{A} \subset \mathcal{F}$ — некоторая σ-алгебра.

Случайная величина $Y := \mathrm{E}(X \mid \mathcal{A})$ называется условным матожиданием X относительно σ-алгебры $\mathcal{A}$, если:

1. Y измерима относительно σ-алгебры $\mathcal{A}$
2. Y "в среднем" совпадает с X на элементах из $\mathcal{A}$, т.е. $\int\limits_A Y \, d\,\mathrm{P} = \int\limits_A X \, d\,\mathrm{P} \quad \forall A \in \mathcal{A}$

Кроме того, если $\mathrm{E}\,|X| < \infty$, то условное математическое ожидание существует (теорема Радона-Никодима) и единственно с точностью до значений на множестве вероятности 0.

Свойства.

1. Линейность: $\mathrm{E}(c_1 X_1 + c_2 X_2 \mid \mathcal{A}) = c_1\,\mathrm{E}(X_1 \mid \mathcal{A}) + c_2\,\mathrm{E}(X_2 \mid \mathcal{A})$
2. Пусть имеются вложенные σ-алгебры $\mathcal{A}_1 \subset \mathcal{A}_2 \subset \mathcal{F}$.

   Тогда $\mathrm{E}\Big(\mathrm{E}(X \mid \mathcal{A}_1) \mid \mathcal{A}_2\Big) = \mathrm{E}(X \mid \mathcal{A}_1) = \mathrm{E}\Big(\mathrm{E}(X \mid \mathcal{A}_2) \mid \mathcal{A}_1\Big)$.
3. $\mathrm{E}(X \mid \{\varnothing, \Omega\}) = \mathrm{E}\,X$
4. $\mathrm{E}\Big(\mathrm{E}(X \mid \mathcal{A})\Big) = \mathrm{E}\,X$

   Доказательство.

   $\mathrm{E}\Big(\mathrm{E}(X \mid \mathcal{A})\Big) = \mathrm{E}\Big(\mathrm{E}(X \mid \mathcal{A}) \mid \{\varnothing, \Omega\}\Big) = \mathrm{E}(X \mid \{\varnothing, \Omega\}) = \mathrm{E}\,X$ □
5. Если Y измерима относительно $\mathcal{A}$, то $\mathrm{E}_\theta(XY \mid \mathcal{A}) = Y\,\mathrm{E}_\theta(X \mid \mathcal{A})$

Мы познакомились с базовыми свойствами условного математического ожидания. Хотелось бы теперь понять, каким образом его вычислять. Посмотрим для начала, как это делается в самом простом случае — дискретном.

Утверждение 4.5.

Пусть Ω — не более чем счетное пространство событий, $\mathcal{F} = 2^\Omega$.

Рассмотрим произвольную σ-алгебру $\mathcal{A} \subset \mathcal{F}$.

Тогда существует разбиение $\Omega = C_1 \sqcup C_2 \sqcup \ldots$, т.ч. $C_i \in \mathcal{A}$ и $\forall A \in \mathcal{A}: \; A = C_{i_1} \sqcup C_{i_2} \sqcup \ldots$

Соответственно, каждое из таких разбиений — конечное либо счетное.

Доказательство.

Заведем на Ω следующее отношение эквивалентности:

$$x \sim y \iff \forall A \in \mathcal{A} : \; x, y \in A \text{ или } x, y \notin A$$

Положим $\Omega = C_1 \sqcup C_2 \sqcup \ldots$ — разбиение на классы эквивалентности по введенному нами отношению. Покажем, что $C_i \in \mathcal{A}$.

Рассмотрим произвольный элемент $x \in \Omega$.

Определим $C_x := \bigcap\limits_{z \not\sim x} A_{x,z}$, где $A_{x,z} \in \mathcal{A}$, т.ч. $x \in A_{x,z}$, $z \notin A_{x,z}$.

Заметим, что для любого $z \not\sim x$ всегда найдется такое множество $A_{x,z}$, это следует из определения эквивалентности. Кроме того, из того же определения эквивалентности имеем, что любой $y \sim x$ будет лежать в $A_{x,z}$. Таким образом, полученный C_x будет содержать лишь эквивалентные x элементы, т.е. будет являться в точности классом эквивалентности элемента x.

А так как C_x является пересечением не более чем счетного числа множеств из $\mathcal{A}$, то C_x также лежит в $\mathcal{A}$. □

Поймем, что нам дает возможность такого разбиения Ω.

Утверждение 4.6.

Пусть $Y = \mathrm{E}(X \mid \mathcal{A})$. Тогда верны следующие утверждения:

1. Y постоянен на C_i (напрямую следует из измеримости Y относительно $\mathcal{A}$).
2. $\int\limits_{C_i} X \, d\,\mathrm{P} = \int\limits_{C_i} Y \, d\,\mathrm{P} = c_i \cdot \mathrm{P}(C_i)$, где c_i — значение, которое принимает Y на множестве C_i.

Отсюда получаем следующее выражение для условного матожидания:

$$Y := \mathrm{E}(X \mid \mathcal{A}) = \sum_i \left[\frac{1}{\mathrm{P}(C_i)} \int\limits_{C_i} X \, d\,\mathrm{P} \cdot \mathbb{1}_{C_i} \right]$$

Мы поняли, как вычислять условное математическое ожидание в дискретном случае. На самом деле, в непрерывном случае получается довольно похожий результат, хотя строгого объяснения далее приводиться не будет.

***Определение* 4.8.**

Пусть X имеет непрерывное распределение, T — некоторая случайная величина, $\mathcal{A}_T$ — минимальная σ-алгебра, относительно которой T измерима.

Введем условное матожидание X относительно σ-алгебры $\mathcal{A}_T$, порожденной T:

$$E(X \mid T) := \mathrm{E}(X \mid \mathcal{A}_T)$$

Утверждение 4.7.

Условное матожидание такого вида представимо как функция от T, т.е.

$$\exists h : \; \mathrm{E}(X \mid \mathcal{A}_T)(\omega) = h(T(\omega)) \;\; \forall \omega \in \Omega$$

***Определение* 4.9.**

В соответствии с последним утверждением можно определить условное матожидание на множествах $\{\omega \in \Omega : T(\omega) = t\}$:

$$\mathrm{E}(X \mid T = t) := h(t)$$

Утверждение 4.8.

В непрерывном случае имеем следующее представление для условного матожидания:

$$\mathrm{E}\Big(g(X)\,|\,T=t\Big) := \frac{\int\limits_{-\infty}^{\infty} g(x)f(x,t)\,dx}{\int\limits_{-\infty}^{\infty} f(x,t)\,dx}$$

Плотность нашей случайной величины, соответственно, принимает такой вид:

$$f_{X\,|\,T}(x\,|\,t) = \frac{f(x,t)}{\int\limits_{-\infty}^{\infty} f(x,t)\,dx}$$

Перейдем теперь непосредственно к обсуждению достаточных статистик. Начнем с определения.

***Определение* 4.10.**

Пусть имеется выборка $X_1, \ldots, X_n \sim \mathrm{P}_\theta,\ \ \theta \in \Theta$.

T — достаточная статистика для $\{\mathrm{P}_\theta \mid \theta \in \Theta\}$, если $\mathrm{P}_\theta(\mathbf{X} \in A\,|\,T=t)$ не зависит от θ.

Замечание.

Такая статистика T, по сути, содержит всю существенную информацию о параметре. Соответственно, если мы знаем значение этой статистики, то мы знаем и распределение нашей выборки.

Пример.

Пусть элементы выборки $X_1, \ldots, X_n$ распределены по Бернулли с параметром θ, т.е.

$$\mathrm{P}(X_i = 1) = \theta,\ \ \mathrm{P}(X_i = 0) = 1-\theta$$

Покажем, что статистика $T = \sum\limits_{k=1}^{n} X_k$ является достаточной для нашего семейства.

Проверим выполнимость условия достаточности на элементарный событиях.

$$\mathrm{P}_\theta(X_1 = x_1, \ldots, X_n = x_n\,|\,T=t) = \frac{\mathrm{P}_\theta\left(X_1 = x_1, \ldots, X_n = x_n; \sum\limits_{i=1}^{n} X_i = t\right)}{\mathrm{P}\left(\sum\limits_{i=1}^{n} X_i = t\right)}$$

Заметим, что если $\sum\limits_{i=1}^{n} x_i \neq t$, то вероятность такого элементарного события равна 0.

Если же $\sum\limits_{i=1}^{n} x_i = t$, то наша условная вероятность представима в следующем виде:

$$\mathrm{P}_\theta(X_1 = x_1, \ldots, X_n = x_n\,|\,T=t) = \frac{\mathrm{P}_\theta\left(X_1 = x_1, \ldots, X_n = x_n\right)}{\mathrm{P}_\theta\left(\sum\limits_{i=1}^{n} X_i = t\right)} =$$

$$= \frac{\prod\limits_{i=1}^{n} \mathrm{P}_\theta(X_i = x_i)}{\mathrm{P}_\theta\left(\sum\limits_{i=1}^{n} X_i = t\right)} = \frac{\prod\limits_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}}$$

Получили, что при известном значении статистики T наша условная вероятность никак не зависит от параметра θ, что и требовалось.

Заметим, что проверять по определению условие достаточности порой может быть очень сложно. Потому хотелось бы получить какой-нибудь простой метод проверки такого условия. Для этого докажем следующую теорему, также известную как теорему факторизации.

Теорема 4.9 (Fisher, Neyman)**.**
Пусть $X_1, \ldots, X_n$ имеют плотность $f(x;\, \theta)$ относительно некоторой σ-конечной меры μ.
Тогда статистика T достаточна $\iff$ $f(x_1, \ldots, x_n;\, \theta) = g_\theta(T(x_1, \ldots, x_n)) \cdot h(x_1, \ldots, x_n)$.

Доказательство.
Без доказательства. □

Замечание.
На самом деле, эта теорема позволяет не только проверять достаточность статистик, но также дает возможность находить их: достаточно привести требуемое разложение f на g_θ и h.

Пример.
Рассмотрим всё тот же пример с бернуллиевским распределением.

$$f(x_1, \ldots, x_n;\, \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i} \cdot (1-\theta)^{n-\sum_{i=1}^n x_i} = \theta^T \cdot (1-\theta)^{n-T} =: g_\theta(T)$$

Таким образом, если положим $h \equiv 1$, то получим в точности разложение совместной плотности в произведение двух функций $g_\theta(T)$ и h. А значит, наша статистика T является достаточной.

Кроме того, на этом же примере можно видеть, каким образом можно было бы получить эту самую достаточную статистику T, если бы мы о ней не знали.

Пример.
Разберем теперь чуть менее приятный пример, а именно — рассмотрим распределение Коши с параметром θ.

Его совместная плотность имеет следующий вид:

$$f(X_1, \ldots, X_n;\, \theta) = \frac{1}{\pi^n} \cdot \prod_{i=1}^{n} \frac{1}{(1 + (x_i - \theta))^2}$$

В этом случае сходу неочевидно, каким образом можно было бы факторизовать нашу плотность так, чтобы она удовлетворяла условиям теоремы.

Более того, утверждается, что для распределения Коши не существует достаточной статистики размерности меньше, чем размер выборки.

Таким образом, в целом теорема работает хорошо, но все-таки не всегда.

Перейдем теперь непосредственно к построению эффективных оценок. Именно здесь нам как раз и потребуется знание достаточных статистик.

Теорема 4.10 (Rao, Blackwell, Kolmogorov)**.**

Пусть T — достаточная статистика, S — произвольная несмещенная оценка параметра θ.

Тогда существует S_1 — несмещенная оценка параметра θ, которая зависит от T и является более эффективной, нежели S.

Иными словами, S_1 удовлетворяет условию $\mathrm{E}_\theta(S_1-\theta)^2 \leqslant \mathrm{E}_\theta(S-\theta)^2 \ \ \forall\theta$.

Доказательство.

Положим $S_1 := \mathrm{E}_\theta(S \,|\, T)$.

Заметим, что т.к. T — достаточная статистика, то S_1 никак не зависит от θ. А значит, S_1 действительно является некоторой оценкой.

Кроме того, $\mathrm{E}_\theta\, S_1 = \mathrm{E}_\theta(\mathrm{E}_\theta(S \,|\, T)) = \mathrm{E}_\theta\, S = \theta$, т.е. S_1 является несмещенной оценкой.

Осталось показать, что S_1 эффективнее S. Действительно:

$$\mathrm{E}_\theta(S-\theta)^2 = \mathrm{E}_\theta(S-S_1+S_1-\theta)^2 = \mathrm{E}_\theta(S_1-\theta)^2 + \mathrm{E}_\theta(S-S_1)^2 + 2\,\mathrm{E}_\theta\Big[(S-S_1)(S_1-\theta)\Big]$$

Разберем чуть подробнее третье слагаемое:

$$\mathrm{E}_\theta\Big[(S-S_1)(S_1-\theta)\Big] = \mathrm{E}_\theta\Big(\mathrm{E}_\theta\Big[(S-S_1)(S_1-\theta)\,|\,T\Big]\Big) = \mathrm{E}_\theta\Big((S_1-\theta)\,\mathrm{E}_\theta(S-S_1\,|\,T)\Big) =$$

$$= \mathrm{E}_\theta\Big((S_1-\theta)\Big[\,\mathrm{E}_\theta(S\,|\,T) - S_1\Big]\Big) = \mathrm{E}_\theta\Big((S_1-\theta)(S_1-S_1)\Big) = 0$$

Здесь во втором равенстве мы воспользовались свойством условного матожидания, которое позволяет выносить измеримые сомножители за пределы матожидания.

Соответсвенно, поскольку третье слагаемое равно 0, а второе — неотрицательно и оценивается снизу 0, то получаем в точности желаемое неравенство. □

Мы, в каком-то смысле, таким образом показали, что эффективная оценка существует. Не хватает лишь единственности.

***Определение* 4.11.**

Достаточная статистика T называется полной, если для любой измеримой функции φ выполняется условие:

$$\Big(\forall\theta:\ \mathrm{E}_\theta\,\varphi(T)=0\Big) \implies \varphi(T)\equiv 0$$

Теорема 4.11 (Lehmann, Scheffé)**.**

Пусть T — полная достаточная статистика.

Тогда существует не более одной (с точностью до значения на T) несмещенной оценки, зависящей от T.

Более того, если такая оценка существует, то она эффективна.

Доказательство.

Пусть S — несмещенная оценка, зависящая от T. Будем считать, что хотя бы одна такая оценка все-таки существует.

Покажем, что в таком случае она единственна.

От противного. Пусть есть две несмещенные оценки $S_1(T)$ и $S_2(T)$. Тогда:

$$\mathrm{E}_\theta\, S_1(T) = \mathrm{E}_\theta\, S_2(T) = \theta \implies \mathrm{E}_\theta\Big[(S_1-S_2)(T)\Big] = 0$$

Если же мы теперь вспомним, что наша статистика T — полная, то как раз получим:

$$\mathrm{E}_\theta\Big[(S_1 - S_2)(T)\Big] = 0 \implies (S_1 - S_2)(T) \equiv 0 \implies S_1(T) \equiv S_2(T)$$

Поймем, почему S будет эффективна. Снова пойдем от противного.

Пусть есть несмещенная оценка R, которая эффективнее S. Но тогда мы знаем, что существует несмещенная оценка $R_1 := \mathrm{E}_\theta(R \,|\, T)$, которая зависит от T и эффективнее S. Пришли к противоречию с единственностью несмещенной оценки, зависящей от T.

Значит, S является эффективной оценкой, что и требовалось. □

4.6. Доверительные интервалы

Предисловие.

Иногда, если размеры выборок невелики, мы не можем достаточно точно определить значение оцениваемого параметра. Тем не менее, нам может быть интересно, в каких пределах он "скорее всего" лежит.

***Определение* 4.12.**

$X_1, \ldots, X_n \sim \mathrm{P}_\theta$ — наблюдения, $\theta \in \Theta \subset \mathbb{R}$

Доверительным интервалом порядка γ называется такой интервал $(T_1(\mathbf{X}),\, T_2(\mathbf{X}))$, что:

$$\mathrm{P}_\theta\Big(T_1(\mathbf{X}) < \theta < T_2(\mathbf{X})\Big) \geqslant \gamma$$

Замечание.

Заметим, что данное выражение следует трактовать не как "вероятность того, что θ попадает в интервал $(T_1,\, T_2)$", а как "вероятность того, что интервал $(T_1,\, T_2)$ содержит параметр θ", поскольку θ в нашем случае является вполне конкретной величиной.

Замечание.

Поскольку нас устраивает любой интервал $(T_1,\, T_2)$, который содержит θ с вероятностью $\geqslant \gamma$, то ограничимся лишь случаем, когда достигается равенство γ.

Замечание.

В общем случае мы хотим выбрать T_1 и T_2 так, чтобы доверительный интервал в среднем был как можно меньше, т.е. хотим минимизировать $\mathrm{E}_\theta(T_2 - T_1)$.

Тем не менее, задача построения "идеального" доверительного интервала как правило не имеет конкретного решения. А потому далее мы лишь поймем, как в принципе можно строить доверительные интервалы.

Для этого рассмотрим следующую пару примеров.

Примеры.

1. Пусть $X_1, \ldots, X_n \sim \mathcal{N}(\theta,\, \sigma^2)$, причем σ^2 нам известна.

   Научимся строить для нашей выборки доверительный интервал параметра θ порядка γ.

   Из леммы Фишера мы знаем, что $\sqrt{n}\dfrac{\overline{X} - \theta}{\sigma} \sim \mathcal{N}(0, 1)$.

   Тогда, с одной стороны, мы имеем

   $$\mathrm{P}_\theta\left(\sqrt{n}\frac{\big|\overline{X} - \theta\big|}{\sigma} < z\right) = \Phi(z) - \Phi(-z) = 2\Phi(z) - 1$$

С другой стороны,

$$\mathrm{P}_\theta\left(\sqrt{n}\frac{|\overline{X}-\theta|}{\sigma} < z\right) = \mathrm{P}_\theta\left(\overline{X} - \frac{\sigma z}{\sqrt{n}} < \theta < \overline{X} + \frac{\sigma z}{\sqrt{n}}\right)$$

Таким образом, мы уже получили некоторый общий вид доверительного интервала. Осталось лишь подобрать такое значение z, при котором наша вероятность будет хотя бы γ.

Для этого заметим, что равенство достигается при $z_\gamma = \Phi^{-1}\left(\frac{1+\gamma}{2}\right)$.

А значит, наш доверительный интервал будет иметь следующий вид:

$$\left(\overline{X} - \frac{\sigma z_\gamma}{\sqrt{n}},\ \overline{X} + \frac{\sigma z_\gamma}{\sqrt{n}}\right),\quad z_\gamma = \Phi^{-1}\left(\frac{1+\gamma}{2}\right)$$

2. Пусть $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2)$, причем θ и σ^2 нам неизвестны.

   Научимся строить доверительный интервал порядка γ для параметра θ.

   Для этого вспомним, что мы умеем оценивать как матожидание, так и дисперсию распределения с помощью соответствующих выборочных величин.

   Из леммы Фишера знаем, что $\sqrt{n-1}\dfrac{\overline{X}-\theta}{S} \sim T_{n-1}$, где T_{n-1} — распределение Стьюдента.

   Поскольку T_{n-1} симметрично, то можем провернуть тот же "трюк", что и в прошлый раз:

$$\mathrm{P}_\theta\left(\sqrt{n-1}\frac{|\overline{X}-\theta|}{S} < z\right) = \mathrm{F}_{T_{n-1}}(z) - \mathrm{F}_{T_{n-1}}(-z) = 2\,\mathrm{F}_{T_{n-1}}(z) - 1$$

$$\mathrm{P}_\theta\left(\sqrt{n-1}\frac{|\overline{X}-\theta|}{S} < z\right) = \mathrm{P}_\theta\left(\overline{X} - \frac{Sz}{\sqrt{n-1}} < \theta < \overline{X} + \frac{Sz}{\sqrt{n-1}}\right)$$

   Тогда если выберем $z_\gamma := \mathrm{F}^{-1}_{T_{n-1}}\left(\frac{1+\gamma}{2}\right)$, то получим доверительный интервал порядка γ:

$$\left(\overline{X} - \frac{Sz_\gamma}{\sqrt{n-1}},\ \overline{X} + \frac{Sz_\gamma}{\sqrt{n-1}}\right),\quad z_\gamma = \mathrm{F}^{-1}_{T_{n-1}}\left(\frac{1+\gamma}{2}\right)$$

Замечание.

Поскольку мы верим в то, что ЦПТ часто применима к реальным данным, то этот пример является одним из самых важных, поскольку распределение ошибки в пределе как раз получается нормальным.

Тем не менее, в жизни встречается не только нормальное распределение, а потому хотелось бы понять, как строить доверительные в общем случае.

Утверждение 4.12.

$X_1, \ldots, X_n \sim \mathrm{P}_\theta,\ \ \theta \in \Theta$

Пусть нам дана такая функция $G(\mathbf{X}; \theta)$, которая удовлетворяет следующим условиям:

1) распределение G не зависит от θ

2) G строго монотонна по θ

Поскольку распределение G не зависит от параметра θ, то можно выбрать c_1 и c_2, такие что:

$$\mathrm{P}_\theta\Big(c_1 < G(\mathbf{X}; \theta) < c_2\Big) = \gamma$$

А так как G монотонна по θ, то неравенство выше можно решить относительно θ:

$$\mathrm{P}_\theta\Big(c_1 < G(\mathbf{X};\,\theta) < c_2\Big) = \mathrm{P}_\theta\Big(d_1(c_1) < \theta < d_2(c_2)\Big)$$

Замечание.

Соответственно, мы в некотором смысле научились находить доверительные интервалы в случае, если нам известная эта функция G.

Такую функцию G иногда еще называют центральной статистикой.

Замечание.

Можно заметить, что приведенное утверждение обобщает методы построения доверительных интервалов, которые мы использовали в примерах выше.

В качестве G в них мы выбирали функции $\sqrt{n}\dfrac{\overline{X}-\theta}{\sigma}$ и $\sqrt{n-1}\dfrac{\overline{X}-\theta}{S}$ соответственно.

С одной стороны, мы по крайней мере поняли, функцию какого вида следует искать, чтобы строить с помощью нее доверительный интервал. С другой стороны, непонятно, как эту функцию в общем случае находить.

Таким образом, хотелось бы все-таки получить способ построения хотя бы какого-нибудь доверительного интервала для первоначального приближения.

Утверждение 4.13.

Пусть $X \sim \mathrm{F}(x)$, где F непрерывна и строго возрастает.

Тогда $\mathrm{F}(X) \sim U[0,1]$.

В частности, $X \sim \mathrm{F}^{-1}(U[0,1])$.

Доказательство.

Просто возьмем и явно выпишем функцию распределения для $\mathrm{F}(x)$:

$$\mathrm{P}\Big(\mathrm{F}(x) < t\Big) = \mathrm{P}\Big(x < \mathrm{F}^{-1}(t)\Big) = \mathrm{F}\Big(\mathrm{F}^{-1}(t)\Big) = t$$

Аналогично:

$$\mathrm{P}\Big(\mathrm{F}^{-1}(t) < x\Big) = \mathrm{P}\Big(t < \mathrm{F}(x)\Big) = \mathrm{F}(x)$$

□

Утверждение 4.14.

Пусть $X_1, \ldots, X_n \sim \mathrm{F}(x;\,\theta)$, где F непрерывна и строго возрастает по x и θ.

$H : \mathbb{R} \to [0,1]$ — произвольная непрерывная строго возрастающая функция.

Тогда в качестве G из утверждения выше можно выбрать функцию следующего вида:

$$G(\mathbf{X};\,\theta) := \sum_{k=1}^{n} H^{-1}\Big(\mathrm{F}(X_k;\,\theta)\Big)$$

Доказательство.

Как мы уже выяснили, распределение $\mathrm{F}(X_k;\,\theta)$ имеет вполне конкретный вид и никак не зависит от параметра θ.

А потому распределение G от θ также никак зависеть не будет.

Кроме того, H непрерывна и строго возрастает. Значит, H обратима, и ее обратная функция также непрерывна и строго возрастает.

Следовательно, функция G также является строго возрастающей как композиция строго возрастающих функций. □

Замечание.

Данное утверждение верно и в случае, когда F строго убывает по θ при любом x, поскольку нам достаточно, чтобы G была строго монотонна по θ. Сам тип монотонности при этом не важен.

Замечание.

Обычно функцию H выбирают таким образом, чтобы случайная величина $H^{-1}(U[0,1])$ имела какое-нибудь известное хорошее распределение. В частности, выбор нередко ложится на различные функции распределения.

Чтобы лучше разобраться в этом, рассмотрим следующий пример.

Пример.

Рассмотрим распределение Коши. В стандартном случае его функция распределения имеет вполне хороший вид:

$$\mathrm{F}(x) = \frac{1}{2} + \frac{1}{\pi}\operatorname{arctg} x, \;\; \mathrm{F}^{-1}(y) = \operatorname{tg}\left(\pi\left(y - \frac{1}{2}\right)\right)$$

Предположим теперь, что наша выборка $\mathbf{X}$ получена из некоторого распределения Коши с параметром θ. Тогда имеем:

$$\mathrm{F}^{-1}\Big(\mathrm{F}(X_k;\,\theta)\Big) = \operatorname{tg}\Big(\pi\,\mathrm{F}(X_k;\,\theta) - \frac{\pi}{2}\Big) \sim \mathrm{Cauchy}(0)$$

Если же мы просуммируем стандартные распределения Коши и отнормируем результат, то снова получим стандартное распределение Коши:

$$G(\mathbf{X};\,\theta) := \frac{1}{n}\sum_{i=1}^{n}\mathrm{F}^{-1}\Big(\mathrm{F}(X_k;\,\theta)\Big) = \frac{1}{n}\sum_{i=1}^{n}\operatorname{tg}\Big(\pi\,\mathrm{F}(X_k;\,\theta) - \frac{\pi}{2}\Big) \sim \mathrm{Cauchy}(0)$$

В корректности этого выражения несложно убедиться, если вспомнить, что характеристическая функция стандартного распределения Коши имеет вид $\varphi(x) = e^{-|x|}$.

Ну а тогда мы, по сути, получили центральную статистику, с помощью которой можно построить и сам доверительный интервал для нашего распределения.

Таким образом, хоть мы так и не научились получать хорошие оценки для распределения Коши, но по крайней мере поняли, каким образом можно построить для него доверительный интервал (хотя в реальности выражение θ через полученную функцию G — довольно противная задача).

Пример.

Попробуем для того же распределения Коши получить какую-нибудь другую центральную статистику. А именно, выберем ее следующим образом:

$$G(\mathbf{X};\,\theta) := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Phi^{-1}\Big(\mathrm{F}(X_k;\,\theta)\Big)$$

Поскольку G — сумма стандартных нормальных распределений, то ее распределение также нормально, и с учетом нормировки получаем $G \sim \mathcal{N}(0,1)$.

Получили еще один вариант центральной статистики, которую можно использовать для построения доверительного интервала.

Замечание.

Пусть мы для прошлой задачи хотим построить доверительный интервал порядка γ.

Поскольку в том и в другом случае функция G распределена симметрично относительно 0, то минимальный доверительный интервал будет также симметричен относительно 0.

А значит, надо просто найти такую константу $c > 0$, которая бы удовлетворяла условию $\mathrm{P}_\theta(-c < G(\mathbf{X};\, \theta) < c) = \gamma$, а затем переписать наше неравенство относительно θ для получения самого интервала.

Таким образом, мы показали, как для одного и того же распределения построить различные центральные статистики. Каждая из этих статистик при этом будет давать свой доверительный интервал. И как раз здесь проявляется проблема построения хорошего (не говоря уже об "идеальном") доверительного интервала, поскольку нельзя с ходу сказать, какой из выбранных нами способов окажется лучше и даст более точную оценку.

Тем не менее, иногда мы все-таки готовы пожертвовать точностью, чтобы получить более приятные и простые выражения, а потому введем определение асимптотического доверительного интервала.

***Определение* 4.13.**

Интервал $(T_1,\, T_2)$ называется асимптотическим доверительным интервалом порядка γ, если

$$\varliminf_{n\to\infty} \mathrm{P}_\theta\Big(T_1(\mathbf{X}) < \theta < T_2(\mathbf{X})\Big) \geqslant \gamma$$

Поймем, что очень часто строить асимптотические доверительные интервалы гораздо проще, нежели обычные.

Утверждение 4.15.

Пусть у нас есть асимптотически нормальная оценка T параметра θ:

$$\sqrt{n}(T(\mathbf{X}) - \theta) \xrightarrow{d} \mathcal{N}(0,\, \sigma^2(\theta))$$

И предположим, что $\sigma(\theta)$ непрерывна.

Научимся тогда строить асимптотический доверительный интервал для параметра θ.

Для начала перепишем определение асимптотической нормальности и получим распределение, не зависящее от параметра:

$$G(\mathbf{X};\, \theta) := \sqrt{n}\frac{T_n(\mathbf{X}) - \theta}{\sigma(\theta)} \xrightarrow{d} \mathcal{N}(0, 1)$$

Это практически то, что нужно. Проблема лишь в том, что знаменатель у нас зависит от θ. Но если бы мы могли избавиться от θ в знаменателе, то без проблем смогли бы получить доверительный интервал.

Заметим тогда, что если есть асимптотическая нормальность, то есть и состоятельность. А если есть состоятельность, то есть и непрерывность:

$$T_n \xrightarrow{\mathrm{P}} \theta \implies \sigma(T_n) \xrightarrow{\mathrm{P}} \sigma(\theta)$$

А тогда можем получить следующий вид нашей центральной статистики:

$$G(\mathbf{X};\, \theta) := \sqrt{n}\frac{T_n - \theta}{\sigma(T_n)} \xrightarrow{d} \mathcal{N}(0, 1)$$

По сути, мы свели задачу построения асимптотического доверительного интервала к разбиравшемуся ранее примеру.

Соответственно, можем теперь получить и сам доверительный интервал порядка γ:

$$\mathrm{P}\left(-z_\gamma < \sqrt{n}\frac{T_n(\mathbf{X}) - \theta}{\sigma(T_n)} < z_\gamma\right) = \mathrm{P}\left(T_n - \frac{\sigma(T_n)z_\gamma}{\sqrt{n}} < \theta < T_n + \frac{\sigma(T_n)z_\gamma}{\sqrt{n}}\right) \to \gamma$$

$$, \text{ где } z_\gamma = \Phi^{-1}\left(\frac{1+\gamma}{2}\right)$$

TODO: important example of asymptotic confidence intervals

4.7. Байесовские и минимаксные оценки

Предисловие.

До сих пор мы разговаривали о фишеровской статистике. Такая модель предполагает, что ошибка в оценке параметров возникает исключительно из-за того, что сбор данных происходит на выборке, а не на всей популяции. По сути, единственная ошибка, существование которой допускается — это ошибка выборки, поскольку выборка может оказаться недостаточно сбалансированной.

В противовес такой модели существует байесовская статистика, которая пытается заранее предсказать, какие значения параметров распределения более вероятны, а затем корректирует свое представление о них в соответствии с последующими наблюдениями.

Обе эти модели в зависимости от конкретной задачи могут вести себя как хорошо, так и плохо. Соответственно, дальше речь как раз пойдет о том, каким образом оценивать параметры распределения в терминах байесовской статистики.

Для начала введем несколько важных для байесовской статистики определений.

***Определение* 4.14.**

Априорным распределением параметра θ называется распределение $q(\theta)$, которое отражает наше представление о распределении параметра до учета экспериментальных данных.

Апостериорным распределением параметра θ называется условное распределение $q(\theta \mid \mathbf{X})$, которое учитывает данные, полученные в результате эксперимента. В частности, оно определяется как

$$q(\theta \mid \mathbf{X}) = \frac{f(\mathbf{X} \mid \theta) \cdot q(\theta)}{f(\mathbf{X})} = \frac{f(\mathbf{X} \mid \theta) \cdot q(\theta)}{\int\limits_\Theta f(\mathbf{X} \mid \tilde{\theta}) \cdot q(\tilde{\theta})\, d\tilde{\theta}}$$

В дискретном случае f и q можно воспринимать как вероятности соответствующих событий.

***Определение* 4.15.**

Функцией потерь называется такая функция $W(T, \theta)$, которая определяет, насколько хорошо наша статистика T оценивает параметр θ при заданной выборке. По сути, эта функция должна удовлетворять следующим условиям:

1. $W(T, \theta) \geqslant 0$, и $W(T, \theta) = 0$, если $T(\mathbf{X}) = \theta$
2. Если $W(T_1, \theta) < W(T_2, \theta)$, то при текущей выборке статистика T_1 лучше оценивает параметр, нежели T_2

Таким образом, задавая W, мы определяем некоторое правило, согласно которому решаем, насколько наша оценка нас устраивает.

Замечание.

Как правило, в качестве функции потерь выбирают $|T-\theta|$ и $(T-\theta)^2$.

Именно с функциями такого вида мы и будем работать дальше.

***Определение* 4.16.**

$R(T,\,\theta) := \mathrm{E}_\theta\, W(T,\,\theta)$ — риск оценки T в точке θ.

По сути, мы таким образом оцениваем средние потери для выбранной нами статистики.

Замечание.

Вполне логичное желание — найти такую оценку T, которая бы минимизировала риск при любом значении параметра θ.

Тем не менее, можно показать, что так не бывает.

А именно, пусть бы такая оценка T существовала.

Тогда выполнялось бы следующее условие $R(T,\,\theta) \leqslant R(S,\,\theta)\ \ \forall S\ \forall\theta$

В частности, мы могли бы выбрать $S \equiv \theta$. А тогда $R(T,\,\theta) \leqslant R(\theta,\,\theta) = 0\ \ \forall\theta$

Но это возможно лишь при условии $W(T,\theta) \equiv 0\ \ \forall\theta$.

А значит, $T \equiv \theta\ \ \forall\theta$, чего не бывает.

Мы показали, что функция риска хоть и позволяет оценивать ошибку нашей оценки, но не дает нам никакой информации о том, какая из оценок действительно лучше в той или иной ситуации.

Тем не менее, с этой проблемой можно побороться, если немного иначе определить риск, который мы хотим минимизировать. Для этого введем понятие минимаксной и байесовской оценки.

***Определение* 4.17.**

Оценка T называется минимаксной, если она минимизирует максимальный риск, т.е

$$\sup_\theta R(T,\,\theta) = \inf_S \sup_\theta R(S,\,\theta)$$

***Определение* 4.18.**

Пусть на нашем параметрическом множестве Θ задано вероятностное распределение Q.

Определим байесовский риск оценки T по отношению к априорному распределению Q как

$$R(T) := \int_\Theta R(T,\,\theta)\,dQ(\theta)$$

Оценка T называется байесовской оценкой по отношению к априорному распределению Q, если она минимизирует байесовский риск, т.е.

$$R(T) = \inf_S R(S)$$

Как и всегда, нам хотелось бы теперь понять, каким образом строить эти оценки.

Утверждение 4.16.

Пусть $X_1,\,\ldots,\,X_n$ имеют плотность $f(x;\,\theta)$ относительно σ-конечной меры μ.

И пусть над Θ задано априорное распределение Q с соответствующей плотностью q.

Тогда байесовская оценка в случае квадратичной функции потерь будет иметь вид

$$T(\mathbf{X}) = \int\limits_{\Theta} \theta \cdot q(\theta \,|\, \mathbf{X})\, d\theta$$

, где $q(\theta \,|\, \mathbf{X})$ — плотность соответствующего апостериорного распределения Q.

Доказательство.

Мы знаем, что байесовская оценка должна минимизировать байесовский риск. Т.е. хотим

$$\int\limits_{\Theta} R(T,\, \theta) \cdot q(\theta)\, d\theta = \int\limits_{\Theta} \int\limits_{\mathcal{X}^n} \Big(T(\mathbf{x}) - \theta\Big)^2 \cdot f(\mathbf{x},\, \theta) \cdot q(\theta)\, d\mu\, d\theta \to \min$$

Поскольку мы интегрируем неотрицательную функцию, то можем переставлять интегралы местами. А тогда можно пробовать для каждого x минимизировать подынтегральное выражение:

$$\int\limits_{\mathcal{X}^n} \int\limits_{\Theta} \Big(T(\mathbf{x}) - \theta\Big)^2 \cdot f(\mathbf{x} \,|\, \theta) \cdot q(\theta)\, d\theta\, d\mu \to \min \iff \int\limits_{\Theta} \Big(T(\mathbf{x}) - \theta\Big)^2 \cdot f(\mathbf{x} \,|\, \theta) \cdot q(\theta)\, d\theta \to \min$$

Раскроем квадрат в полученном выражении:

$$T(\mathbf{x})^2 \int\limits_{\Theta} f(\mathbf{x} \,|\, \theta) \cdot q(\theta)\, d\theta - 2T(\mathbf{x}) \int\limits_{\Theta} \theta \cdot f(\mathbf{x} \,|\, \theta) \cdot q(\theta)\, d\theta + \int\limits_{\Theta} \theta^2 \cdot f(\mathbf{x} \,|\, \theta) \cdot q(\theta)\, d\theta \to \min$$

Мы получили некоторую квадратичную функцию. А искать минимум квадратичной функции мы умеем:

$$T(\mathbf{x}) = \frac{\displaystyle\int\limits_{\Theta} \theta \cdot f(\mathbf{x} \,|\, \theta) \cdot q(\theta)\, d\theta}{\displaystyle\int\limits_{\Theta} f(\mathbf{x} \,|\, \theta) \cdot q(\theta)\, d\theta} = \int\limits_{\Theta} \theta \cdot q(\theta \,|\, \mathbf{x})\, d\theta$$

□

Замечание.

Мы выяснили, что в случае квадратичной функции потерь наша оценка — это апостериорное среднее.

Несложно убедиться, что в случае, когда функция потерь определяется как модуль разности, наша оценка будет определяться как апостериорная медиана.

Перейдем теперь к минимаксным оценкам.

Теорема 4.17.

Пусть T_k — байесовская оценка по отношению к априорному распределению Q_k.

T — произвольная оценка параметра θ.

Тогда если $\sup\limits_{\theta} R(T,\, \theta) \leqslant \overline{\lim\limits_{k}}\, R(T_k)$, то T — минимаксная оценка.

Доказательство.

Рассмотрим произвольную оценку S. Тогда верно:

$$\sup_\theta R(S,\,\theta) \geqslant \int\limits_\Theta R(S,\,\theta)\,dQ_k(\theta) \geqslant \int\limits_\Theta R(T_k,\,\theta)\,dQ_k(\theta) = R(T_k)$$

Заметим, что это верно для любого k. А тогда можем перейти к пределу по k и воспользоваться условием теоремы:

$$\sup_\theta R(S,\,\theta) \geqslant \overline{\lim_k}\, R(T_k) \geqslant \sup_\theta R(T,\,\theta)$$

Получили определение минимаксной оценки. □

Следствие.

Пусть T — байесовская оценка с постоянным риском, т.е. $R(T,\,\theta)$ не зависит от θ.

Тогда T — минимаксная оценка.

Доказательство.

В условии теоремы возьмем последовательность $T_k = T$.

А так как $R(T,\,\theta)$ никак не зависит от θ, то:

$$\sup_\theta R(T,\,\theta) = \int\limits_\Theta R(T,\,\theta)\,dQ = R(T)$$

Таким образом, условие теоремы выполняется и T — минимаксная оценка. □

TODO: examples

5. Регрессионный анализ

5.1. Линейная регрессия

Предисловие.

Очень часто нам хочется определить, как зависит наблюдаемая случайная величина от каких-то других случайных величин. При этом в реальности так получается, что исследуемые зависимости почти всегда являются лишь статистическими, но не функциональными.

К примеру, довольно типичной задачей является определение зависимости стоимости квартиры от каких-либо ее параметров: площади, этажа, района. При равных значениях независимых переменных стоимость квартиры может отличаться. По сути, она является некоторой случайной величиной, зависящей от входных параметров. А потому речь здесь может идти лишь о статистической зависимости.

Собственно, регрессионный анализ как раз и занимается подобным исследованием поведения зависимой случайной величины от набора различных независимых переменных.

Попробуем теперь формально определить понятие регрессии.

***Определение* 5.1.**

Пусть $\vec{X} = (X_1, \ldots, X_k)^T$ — набор независимых случайных величин, а Y — произвольная случайная величина, зависящая от $\vec{X}$.

Тогда функция $f(\vec{x}) = \mathrm{E}(Y \mid \vec{X} = \vec{x})$ называется регрессией Y на $\vec{X}$.

Разумеется, в общем случае Y может вести себя как угодно в зависимости от значений независимых переменных $\vec{X}$. А потому хочется для начала рассмотреть какую-нибудь простую модель регрессии.

***Определение* 5.2.**

Будем считать, что поведение Y описывается следующим образом:

$$Y \mid \vec{X} = \vec{X}^T\vec{\theta} + \varepsilon$$

Здесь $\vec{\theta}$ — набор из k параметров, который нам неизвестен, но заранее фиксирован. Именно его мы и будем пытаться найти.

Величина ε при этом является некоторой случайной величиной (шумом), которая обладает следующими свойствами:

$$\mathrm{E}\,\varepsilon = 0, \;\; \mathrm{D}\,\varepsilon = \sigma^2, \text{ причем } \sigma^2 \text{ нам неизвестна}$$

Сама регрессия в таком случае будет иметь вид

$$f(\vec{x}) := \mathrm{E}(Y \mid \vec{X} = \vec{x}) = \vec{x}^T\vec{\theta}$$

Такая модель регрессии, в которой значение функции f линейно зависит от значений $\vec{X}$, называется линейной регрессией.

Замечание.

Рассмотрим все тот же пример с квартирой.

В этом случае значения X_i — это некоторые количественные параметры квартиры: площадь, этаж, год постройки здания и т.п.

Соответственно, в модели линейной регрессии мы предполагаем, что стоимость квартиры имеет некоторую линейную зависимость от значений наших переменных. Но при этом допускаем отклонения стоимости в ту и другую сторону.

Замечание.

Далее мы везде будем работать с параметрами и независимыми переменными как с векторами. Явно при этом обозначение их "векторной природы" использоваться не будет.

Мы ввели понятие линейной регрессии и даже определили задачу: найти набор параметров θ, который бы описывал поведение зависимой переменной Y. Вопрос лишь в том, как этот набор параметров находить.

Как и всегда, будем пробовать восстанавливать параметры на основе наблюдений.

Утверждение 5.1.

Пусть $(X_1, Y_1), \ldots, (X_n, Y_n)$ — наши наблюдения.

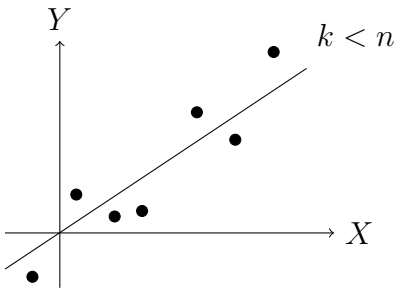
Тогда в нашей линейной модели вектор значений $\mathbf{Y}$ представим следующим образом:

$$\mathbf{Y} = \mathbf{X} \cdot \theta + \varepsilon$$

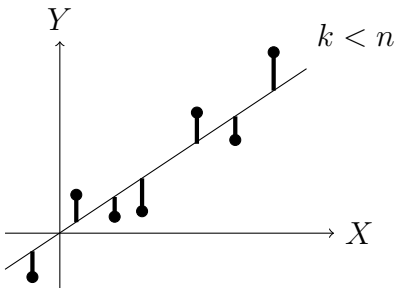
$$, \text{ где } \mathbf{X} = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \mathrm{E}\,\varepsilon_i\varepsilon_j = 0 \text{ при } i \neq j$$

Заметим, что если бы мы могли "забыть" про существующий шум ε и не учитывать его, то мы бы получили СЛАУ, которую можно было бы попробовать решить с помощью того же метода Гаусса.

Тем не менее, обычно количество наших наблюдений (n) оказывается сильно больше, чем количество параметров (k). А потому такая система в общем случае вообще не будет иметь решения.



Тогда возникает другая идея: искать набор параметров θ таким образом, чтобы минимизировать сумму квадратов отклонений значений зависимой переменной Y от нашей прямой (или гиперплоскости в случае пространства большей размерности).



Формально, хочется найти такой набор параметров θ, который бы минимизировал функцию

$$S(\theta) = \sum_{i=1}^{n} \left(Y_i - X_i^T\theta\right)^2 = \|\mathbf{Y} - \mathbf{X}\,\theta\|^2$$

***Определение* 5.3.**
Оценка θ^*, такая что $S(\theta^*) = \min\limits_{\theta} S(\theta)$, называется оценкой наименьших квадратов (ОНК).

Осталось понять, каким образом эту оценку получить. Для начала введем следующее обозначение и докажем пару связанных с ним утверждений.

Обозначение.
Введем следующее обозначение для вектора из частных производных:

$$\frac{df}{d\theta} := \begin{pmatrix} \dfrac{\partial f}{\partial \theta_1} \\ \vdots \\ \dfrac{\partial f}{\partial \theta_k} \end{pmatrix}$$

Утверждение 5.2.
Пусть α — произвольный вектор из k компонент. Тогда верно равенство

$$\frac{d}{d\theta}\left(\alpha^T\theta\right) = \frac{d}{d\theta}\left(\theta^T\alpha\right) = \alpha$$

Доказательство.
Поскольку $\alpha^T\theta = \sum\limits_{i=1}^{k} \alpha_i\theta_i = \theta^T\alpha$, то частная производная по θ_i будет в точности равна α_i. □

Утверждение 5.3.
Пусть A — симметричная матрица, т.е. $A^T = A$. Тогда верно равенство

$$\frac{d}{d\theta}\left(\theta^T A\theta\right) = 2A\theta$$

Доказательство.
Распишем нашу квадратичную форму:

$$\theta^T A\theta = \sum_{i,j} A_{i,j}\theta_i\theta_j$$

Остается лишь посмотреть на вид ее частной производной по параметру θ_i и убедиться, что она совпадает с i-ой строкой матрицы $2A\theta$:

$$\frac{\partial}{\partial\,\theta_i}\left(\sum_{i,j} A_{i,j}\theta_i\theta_j\right) = \sum_{j=1}^{k}\left(A_{i,j} + A_{j,i}\right)\theta_j = 2\sum_{j=1}^{k} A_{i,j}\theta_j = \Big(2A\theta\Big)_i$$

□

Перейдем теперь к нахождению оценки наименьших квадратов.

Утверждение 5.4.

Пусть (X_1, Y_1), …, (X_n, Y_n) — наши наблюдения.

Тогда ОНК θ^* достигается на любом решении системы $\mathbf{X}^T\mathbf{X}\,\theta = \mathbf{X}^T\mathbf{Y}$.

В частности, если матрица $\mathbf{X}^T\mathbf{X}$ обратима, ОМП имеет вид $\theta^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.

Доказательство.

Распишем чуть иначе нашу функцию $S(\theta)$:

$$S(\theta) = \|\mathbf{Y} - \mathbf{X}\,\theta\|^2 = \Big\langle \mathbf{Y} - \mathbf{X}\,\theta,\, \mathbf{Y} - \mathbf{X}\,\theta \Big\rangle = \theta^T\,\mathbf{X}^T\mathbf{X}\,\theta - 2\mathbf{Y}^T\,\mathbf{X}\,\theta + \mathbf{Y}^T\,\mathbf{Y}$$

Такая функция всегда достигает своего минимума, поскольку полученная квадратичная форма неотрицательно определена. В частности, в точке минимума значение производной нашей функции равно 0.

Посмотрим для начала на то, как наша производная вообще выглядит:

$$\frac{dS(\theta)}{d\theta} = 2\mathbf{X}^T\mathbf{X}\,\theta - 2\mathbf{X}^T\mathbf{Y}$$

Отсюда уже видно, что нас интересуют решения системы $\mathbf{X}^T\mathbf{X}\,\theta = \mathbf{X}^T\mathbf{Y}$.

А тогда если $\mathbf{X}^T\mathbf{X}$ обратима, то решение единственно, и наша ОНК принимает вид

$$\theta^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Остается показать, что если матрица $\mathbf{X}^T\mathbf{X}$ необратима, то нам подойдет любое решение системы.

Для этого рассмотрим какое-нибудь решение θ^*. И покажем, что при любом другом значении параметра θ мы получим решение не лучше:

$$S(\theta) = \|\mathbf{Y} - \mathbf{X}\,\theta\|^2 = \|\mathbf{Y} - \mathbf{X}\,\theta^* + \mathbf{X}\,(\theta^* - \theta)\|^2 =$$

$$= \|\mathbf{Y} - \mathbf{X}\,\theta^*\|^2 + \|\mathbf{X}\,(\theta^* - \theta)\|^2 + 2\Big\langle \mathbf{X}\,(\theta^* - \theta), \mathbf{Y} - \mathbf{X}\,\theta^* \Big\rangle =$$

$$= \|\mathbf{Y} - \mathbf{X}\,\theta^*\|^2 + \|\mathbf{X}\,(\theta^* - \theta)\|^2 + 2(\theta^* - \theta)^T\Big(\mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{X}\,\theta^*\Big) =$$

$$= \|\mathbf{Y} - \mathbf{X}\,\theta^*\|^2 + \|\mathbf{X}\,(\theta^* - \theta)\|^2 + 0 \geqslant \|\mathbf{Y} - \mathbf{X}\,\theta^*\|^2 = S(\theta^*)$$

□

Замечание.

Поскольку выбор матрицы $\mathbf{X}$ в нашей власти (она формируется из наших наблюдений), то всегда можно добиться того, чтобы матрица $\mathbf{X}^T\mathbf{X}$ была обратима. Хотя с точки зрения нахождения решения разницы никакой не будет (придется или обращать матрицу, или находить явно какое-нибудь решение с помощью того же метода Гаусса).

Таким образом, мы получили некоторую оценку параметров для нашей модели линейной регрессии. Вопрос лишь в том, насколько наша оценка в действительности хорошая.

***Определение* 5.4.**

Пусть θ_1 и θ_2 — оценки набора параметров θ.

Тогда θ_1 эффективнее θ_2, если $\mathrm{cov}(\theta_2) - \mathrm{cov}(\theta_1)$ — неотрицательно определенная матрица.

Здесь $\mathrm{cov}(\theta)$ — ковариационная матрица случайного вектора θ.

Замечание.

Данное определение эффективности является обобщением соответствующего определения для одномерного случая.

Только если раньше мы пытались оценивать дисперсию нашей оценки, то теперь можем пробовать оценивать ее "многомерный вариант", т.е. ковариационную матрицу.

Замечание.

Введем отношение порядка на матрицах как "$A \leqslant B$, если $B - A$ неотрицательно определена".

Тогда оценка θ_1 эффективнее оценки θ_2, если $\operatorname{cov}(\theta_1) \leqslant \operatorname{cov}(\theta_2)$.

Теорема 5.5 (Gauss, Markov)**.**

Пусть $(X_1, Y_1), \ldots, (X_n, Y_n)$ — наши наблюдения, и существует $(\mathbf{X}^T\mathbf{X})^{-1}$.

Причем будем рассматривать модель, в которой Y имеет распределение $Y \mid X = X^T\theta + \varepsilon$.

Тогда $\theta^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ — наилучшая линейная несмещенная оценка параметра θ.

В английской литературе она известна как best linear unbiased estimator, или попросту BLUE.

Доказательство.

Линейность нашей оценки очевидна (делаем линейные преобразования с элементами из $\mathbf{Y}$).

Проверим тогда несмещенность нашей оценки θ^*:

$$\mathrm{E}_\theta\, \theta^* = \mathrm{E}_\theta \left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\right] = \mathrm{E}_\theta \left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\,\theta + \varepsilon)\right] =$$

$$= \mathrm{E}_\theta\, \theta + \mathrm{E}_\theta \left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon\right] = \theta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\, \mathrm{E}_\theta\, \varepsilon = \theta$$

Последний переход здесь верен в силу линейности матожидания и того, что распределение $\mathbf{X}$ никак не зависит от θ.

Осталось проверить эффективность. Для этого возьмем какую-нибудь другую линейную несмещенную оценку $\tilde{\theta}$ и покажем, что она не лучше.

Во-первых, эта оценка будет иметь вид $\tilde{\theta} = L\mathbf{Y}$. А так как мы хотим, чтобы эта оценка была несмещенной, то матрица L должна удовлетворять условию $\mathrm{E}_\theta(L\mathbf{Y}) = \theta \;\; \forall\theta$, то есть:

$$\forall\theta: \;\; \theta = \mathrm{E}_\theta(L\mathbf{Y}) = \mathrm{E}_\theta(L\mathbf{X}\,\theta + L\varepsilon) = L\mathbf{X}\,\theta$$

$$\Longrightarrow \;\; L\mathbf{X} = E_k, \text{ где } E_k \text{ — единичная матрица размера } k$$

Посмотрим теперь на ковариационную матрицу θ^*:

$$\operatorname{cov}(\theta^*) = \mathrm{E}_\theta \left[(\theta^* - \theta)(\theta^* - \theta)^T\right] =$$

$$= \mathrm{E}_\theta \left[\Big((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\,\theta + \varepsilon) - \theta\Big)\Big((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\,\theta + \varepsilon) - \theta\Big)^T\right] =$$

$$= \mathrm{E}_\theta \left[\Big((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon\Big)\Big((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon\Big)^T\right] = \mathrm{E}_\theta \left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon\varepsilon^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\right] =$$

Заметим, что в последнем выражении случайной является только величина $\varepsilon\varepsilon^T$. А потому всю остальную часть по линейности можно вынести из-под матожидания.

$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \cdot \mathrm{E}_\theta(\varepsilon\varepsilon^T) \cdot \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} =$$

Заметим, что $\mathrm{E}(\varepsilon\varepsilon^T)$ — это, по сути, ковариационная матрица "шумов". Более того, на диагонали этой матрицы будут стоять дисперсии шумов, а в остальных позициях — нули, поскольку различные шумы не коррелируют между собой.

$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \cdot \sigma^2 E_n \cdot \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2 \cdot (\mathbf{X}^T\mathbf{X})^{-1}$$

Таким образом, мы смогли посчитать ковариационную матрицу для оценки θ^*.

Осталось посчитать ковариационную матрицу для $\tilde{\theta} = L\mathbf{Y}$, при условии что $L\mathbf{X} = E_k$.

Заведем для начала следующее обозначение:

$$C := L - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \implies L = C + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

Теперь попробуем расписать саму ковариационную матрицу $\mathrm{cov}(\tilde{\theta})$:

$$\mathrm{cov}(\tilde{\theta}) = \mathrm{E}_\theta\left[(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^T\right] = \mathrm{E}_\theta\left[(L\mathbf{Y} - \theta)(L\mathbf{Y} - \theta)^T\right] =$$

$$= \mathrm{E}_\theta\left[\Big((C + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)(\mathbf{X}\,\theta + \varepsilon) - \theta\Big)\Big((C + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)(\mathbf{X}\,\theta + \varepsilon) - \theta\Big)^T\right] =$$

$$= \mathrm{E}_\theta\left[\Big(C\mathbf{X}\,\theta + C\varepsilon + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon\Big)\Big(C\mathbf{X}\,\theta + C\varepsilon + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon\Big)^T\right] =$$

Из определения C и условий на L заметим, что $C\mathbf{X} = 0$. Тогда:

$$= \mathrm{E}_\theta\left[\Big(C\varepsilon + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon\Big)\Big(C\varepsilon + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon\Big)^T\right] =$$

$$= \mathrm{E}_\theta\left[C\varepsilon\varepsilon^T C^T + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon\varepsilon^T C^T + C\varepsilon\varepsilon^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon\varepsilon^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\right] =$$

По линейности матожидания попробуем оценить каждое слагаемое по отдельности.

$$= \sigma^2 \cdot CC^T + \left[\sigma^2 \cdot (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T C^T + \sigma^2 \cdot C\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\right] + \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} =$$

$$= \sigma^2 \cdot CC^T + 0 + \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \mathrm{cov}(\theta^*) + \sigma^2 \cdot CC^T \geqslant \mathrm{cov}(\theta^*)$$

Таким образом мы получили, что наша оценка θ^* является эффективной, что и требовалось. □

Напоследок получим еще один маленький результат. А именно, попробуем оценить дисперсию шумов, которая изначально была нам неизвестна.

Утверждение 5.6.

$$\frac{1}{n-k}S(\theta^*) \text{ — несмещенная оценка } \sigma^2$$

Доказательство.

Оценим матожидание величины $S(\theta^*)$:

$$\mathrm{E}_{\theta,\sigma}\, S(\theta^*) = \mathrm{E}_{\theta,\sigma}\left[(\mathbf{Y} - \mathbf{X}\,\theta^*)^T(\mathbf{Y} - \mathbf{X}\,\theta^*)\right] =$$

$$= \mathrm{E}_{\theta,\sigma}\left[\Big(\mathbf{X}\,\theta + \varepsilon - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\theta + \varepsilon)\Big)^T\Big(\mathbf{X}\,\theta + \varepsilon - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\theta + \varepsilon)\Big)\right] =$$

$$= \mathrm{E}_\sigma \left[\Big(\varepsilon - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon \Big)^T \Big(\varepsilon - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon \Big) \right] =$$

$$= \mathrm{E}_\sigma \left[\varepsilon^T \Big(E_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \Big)^T \Big(E_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \Big) \varepsilon \right] =: \mathrm{E}_\sigma \left(\varepsilon^T B^T B \varepsilon \right)$$

Поскольку наши шумы независимы, а их дисперсия совпадает, то можем переписать полученное матожидание в следующем виде:

$$\mathrm{E}_\sigma \left(\varepsilon^T B^T B \varepsilon \right) = \sum_{i,j} (B^T B)_{i,j} \, \mathrm{E}_\sigma \, \varepsilon_i \varepsilon_j = \sum_{i=1}^{n} (B^T B)_{i,i} \, \mathrm{E}_\sigma \, \varepsilon_i^2 = \sigma^2 \cdot \mathrm{tr}\big(B^T B \big)$$

Осталось понять, что след нашей матрицы будет равен $n - k$. Действительно:

$$B^T B = E_n - 2\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \cdot \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = E_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

$$\mathrm{tr}\big(B^T B \big) = \mathrm{tr}(E_n) - \mathrm{tr}\big(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \big) \overset{\mathrm{tr}(AB) = \mathrm{tr}(BA)}{=} n - \mathrm{tr}\big((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X} \big) = n - k$$

□

6. Проверки гипотез

6.1. Постановка задачи

Напомним формулировку задачи проверки гипотез в самом простом ее случае.

***Определение* 6.1.**
Пусть $X_1, \ldots, X_n \sim \mathrm{P}_\theta$ — произвольная выборка, распределение которой зависит от параметра $\theta \in \Theta$.

Положим, что параметрическое множество Θ представимо как $\Theta = \Theta_0 \sqcup \Theta_1$.

Тогда гипотезами H_0 и H_1 формально называются утверждения о том, какому из множеств Θ_0 и Θ_1 принадлежит наш параметр θ.

При этом обычно гипотезу H_0 называют основной (или нулевой), а гипотезу H_1 — альтернативной.

Соответственно наша задача заключается в том, чтобы на основе выборки понять, какую из двух гипотез следует принять, а какую — отвергнуть.

Поскольку мы изначально не знаем, в каком из множеств Θ_0 и Θ_1 лежит параметр θ, принимая какую-то из гипотез мы можем как угадать, так и ошибиться. Таким образом, можем определить четыре следующих варианта исхода событий:

	принять H_0	принять H_1
верна H_0	✓	ошибка I рода
верна H_1	ошибка II рода	✓

Все, что нам теперь хочется — научиться принимать одну из гипотез некоторым “оптимальным” образом.

***Определение* 6.2.**
Пусть $X_1, \ldots, X_n \sim \mathrm{P}_\theta$ — произвольная выборка, $\theta \in \Theta$

И пусть выдвинуты гипотезы $H_0 : \theta \in \Theta_0$ и $H_1 : \theta \in \Theta_1$, где $\Theta = \Theta_0 \sqcup \Theta_1$.

Тогда критерием $\varphi : \mathcal{X}^n \to [0, 1]$ называется такая функция, которая по выборке $\mathbf{X}$ выдает нам вероятность отвергнуть гипотезу H_0.

На основе этого критерия мы и будем в дальнейшем выбирать, какую из гипотез следует принять.

Замечание.
Подразумевается, что если критерий φ для нашей выборки выдает 0 или 1, то мы однозначно выбираем одну из двух гипотез. В противном же случае нам потребуется провести некоторый независимый эксперимент и принять решение на основе него.

К примеру, в таком случае можно подкинуть “кривую монетку”, одно из значений которой выпадает с вероятностью $\varphi(\mathbf{X})$.

Замечание.
Также можно заметить, что вероятности ошибок выражаются следующим образом:

$$\mathrm{P}(\text{ошибка I рода}) = \mathrm{E}_0\,\varphi(\mathbf{X}), \quad \mathrm{P}(\text{ошибка II рода}) = \mathrm{E}_1(1 - \varphi(\mathbf{X}))$$

Опишем теперь подробно возможную постановку нашей задачи.

Задача.

Пусть нам задано некоторое число $\alpha \in (0, 1)$.

Тогда для заданного значения (уровня) α требуется построить критерий φ, т.ч.

1) P(ошибка I рода) $\leqslant \alpha$

2) P(ошибка II рода) $\to \min$

Мощность такого критерия при этом определяется как $1 - \mathrm{P}$(ошибка II рода).

Замечание.

Как следствие из предыдущего замечания, можем получить альтернативное выражение для мощности критерия:

$$\text{Мощность критерия} = \mathrm{E}_1\, \varphi(\mathbf{X})$$

Рассмотрим пару примеров, чтобы лучше разобраться в том, как вообще могут быть устроены критерии. Ограничимся при этом нахождением каких-нибудь критериев уровня α, не задумываясь об их мощности.

Примеры.

1. Пусть $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2)$ — выборка из нормального распределения, σ^2 известна.

   Определим следующие гипотезы:

   $$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0$$

   Заметим, что если верна гипотеза H_0, то по лемме Фишера выполняется условие:

   $$Y := \sqrt{n}\frac{\overline{X} - \theta_0}{\sigma} \sim \mathcal{N}(0, 1)$$

   Научимся на основе этого знания строить критерий уровня α. Для этого воспользуемся тем же "трюком", что и при построении доверительных интервалов:

   $$\alpha = \mathrm{P}(|Y| > z_\alpha) = 2(1 - \Phi(z_\alpha)) \implies z_\alpha := \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

   Отсюда построим и сам критерий:

   $$\varphi(\mathbf{X}) := \begin{cases} 1, & \sqrt{n}\dfrac{\left|\overline{X} - \theta_0\right|}{\sigma} > z_\alpha \\ 0, & \sqrt{n}\dfrac{\left|\overline{X} - \theta_0\right|}{\sigma} \leqslant z_\alpha \end{cases}$$

   Заметим, что мы выбирали z_α таким образом, что $\mathrm{P}\left(\sqrt{n}\dfrac{\left|\overline{X} - \theta_0\right|}{\sigma} > z_\alpha\right) \leqslant \alpha$.

   Следовательно, вероятность ошибки первого рода (т.е. ошибки в случае, когда распределение выборки зависит от θ_0) точно будет не больше, чем α.

   С мощностью же этого критерия мы разберемся чуть позднее.

2. Пусть в условии предыдущего примера нам неизвестна дисперсия σ^2.

   Тогда мы снова можем воспользоваться леммой Фишера, оценивая σ^2 с помощью выборочной дисперсии, и получить следующий вариант критерия:

$$\varphi(\mathbf{X}) := \begin{cases} 1, & \sqrt{n-1}\dfrac{|\overline{X} - \theta_0|}{S} > z_\alpha \\ 0, & \sqrt{n-1}\dfrac{|\overline{X} - \theta_0|}{S} \leqslant z_\alpha \end{cases}, \text{ где } z_\alpha := \mathrm{F}^{-1}_{T_{n-1}}\left(1 - \frac{\alpha}{2}\right)$$

6.2. Наиболее мощные критерии

Обычно нам хочется не только ограничить вероятность ошибки первого рода, но при прочих равных еще и минимизировать вероятность ошибки второго рода.

Чтобы справиться с этой задачей, попробуем рассмотреть еще более простой вариант ее постановки и понять, как в таком случае будет устроен критерий.

Замечание.

Если гипотеза предполагает лишь одно распределение, то она называется простой, иначе — сложной.

Лемма (Neyman, Pearson)**.**

Пусть $X_1, \ldots, X_n \sim \mathrm{P}$ — произвольная выборка из некоторого распределения.

И пусть мы рассматриваем простые гипотезы, которые имеют вид $H_0 : \mathrm{P} = \mathrm{P}_0$ и $H_1 : \mathrm{P} = \mathrm{P}_1$.

Положим f_0 и f_1 — плотности P_0 и P_1 относительно некоторой меры μ, и рассмотрим семейство критериев следующего вида:

$$\varphi(\mathbf{X}) := \begin{cases} 1, & f_1(\mathbf{X}) > k \cdot f_0(\mathbf{X}) \\ p(k), & f_1(\mathbf{X}) = k \cdot f_0(\mathbf{X}) \\ 0, & f_1(\mathbf{X}) < k \cdot f_0(\mathbf{X}) \end{cases}, \qquad (*)$$

$$\text{где } p(k) = \begin{cases} 0, & \mathrm{P}\left(f_1(\mathbf{X}) = k \cdot f_0(\mathbf{X})\right) = 0 \\ \dfrac{\alpha_0 - \alpha(k)}{\alpha(k-0) - \alpha(k)}, & \mathrm{P}\left(f_1(\mathbf{X}) = k \cdot f_0(\mathbf{X})\right) > 0 \end{cases},$$

$$\alpha_0 \in (0,1), \quad \alpha(k) = \mathrm{P}_0\left(f_1(\mathbf{X}) > k \cdot f_0(\mathbf{X})\right)$$

Тогда верны следующие утверждения:

1) $\forall\, \alpha_0 \in (0,1)$ существует критерий уровня ровно α_0 вида $(*)$

2) Этот критерий — наиболее мощный среди всех критериев уровня $\leqslant \alpha_0$

3) Любой наиболее мощный критерий уровня α_0 имеет вид $(*)$ с точностью до значений критерия при $f_1(\mathbf{X}) = k \cdot f_2(\mathbf{X})$

Доказательство.

1. Посмотрим внимательнее на функцию $\alpha(k)$, которая как раз соответствует вероятности ошибки первого рода и определяет уровень критерия.

   Эта функция не возрастает и непрерывна справа, причем $\alpha(-\infty) = 1$ и $\alpha(+\infty) = 0$.

   А тогда возможны две ситуации.

В первом случае существует такое k, при котором $\alpha(k) = \alpha_0$.

В данной ситуации по непрерывности справа получаем, что $\mathrm{P}(f_1(\mathbf{X}) = k \cdot f_0(\mathbf{X})) = 0$. А тогда сам критерий имеет вид:

$$\varphi(\mathbf{X}) := \begin{cases} 1, & f_1(\mathbf{X}) > k \cdot f_0(\mathbf{X}) \\ 0, & f_1(\mathbf{X}) < k \cdot f_0(\mathbf{X}) \end{cases}$$

Иначе в силу непрерывности справа найдется такое k, что $\alpha(k-0) \geqslant \alpha_0 > \alpha(k)$.

Поймем, что в данном случае критерий будет иметь в точности вид $(*)$.

Действительно, посчитаем вероятность ошибки первого рода.

$$\mathrm{E}_0\,\varphi = \mathrm{P}_0\left(f_1(\mathbf{X}) > k \cdot f_0(\mathbf{X})\right) + \mathrm{P}_0\left(f_1(\mathbf{X}) = k \cdot f_0(\mathbf{X})\right) \cdot \frac{\alpha_0 - \alpha(k)}{\alpha(k-0) - \alpha(k)} =$$

$$= \alpha(k) + \big(\alpha(k-0) - \alpha(k)\big) \cdot \frac{\alpha_0 - \alpha_k}{\alpha(k-0) - \alpha(k)} = \alpha(k) + \big(\alpha_0 - \alpha(k)\big) = \alpha_0$$

2. Рассмотрим какой-нибудь другой критерий $\tilde{\varphi}$, т.ч. $\tilde{\alpha}_0 = \mathrm{E}_0\,\tilde{\varphi} \leqslant \alpha_0$. Покажем, что этот критерий не мощнее, т.е.
$$\mathrm{E}_1\,\varphi \geqslant \mathrm{E}_1\,\tilde{\varphi}$$

Действительно:

$$\mathrm{E}_1(\varphi - \tilde{\varphi}) \geqslant \mathrm{E}_1(\varphi - \tilde{\varphi}) - k \cdot \mathrm{E}_0(\varphi - \tilde{\varphi}) = \int\limits_{\mathcal{X}^n} (\varphi - \tilde{\varphi})\big(f_1(x) - k f_0(x)\big)\,d\mu \geqslant 0$$

Поймем последнее неравенство. Для этого рассмотрим три случая:

1) $f_1(x) > k \cdot f_0(x) \implies \varphi = 1 \implies \varphi - \tilde{\varphi} \geqslant 0 \implies (\varphi - \tilde{\varphi})\big(f_1(x) - k \cdot f_0(x)\big) \geqslant 0$

2) $f_1(x) = k \cdot f_0(x) \implies f_1(x) - k \cdot f_0(x) = 0 \implies (\varphi - \tilde{\varphi})\big(f_1(x) - k \cdot f_0(x)\big) = 0$

3) $f_1(x) < k \cdot f_0(x) \implies \varphi = 0 \implies \varphi - \tilde{\varphi} \leqslant 0 \implies (\varphi - \tilde{\varphi})\big(f_1(x) - k \cdot f_0(x)\big) \geqslant 0$

Получили, что во всех трех ситуациях подынтегральное выражение оказывается неотрицательным. А значит, и сам интеграл неотрицателен, что и требовалось.

3. Пусть $\tilde{\varphi}$ — наиболее мощный критерий уровня $\leqslant \alpha_0$.

По доказанному выше мы тогда имеем:

$$\mathrm{E}_1(\varphi - \tilde{\varphi}) \geqslant 0, \quad \mathrm{E}_1(\varphi - \tilde{\varphi}) \geqslant 0 \implies \alpha_0 = \mathrm{E}_1\,\varphi = \mathrm{E}_1\,\tilde{\varphi}$$

В частности, из того же второго пункта получаем, что:

$$\int\limits_{\mathcal{X}^n} (\varphi - \tilde{\varphi})\big(f_1(x) - k f_0(x)\big)\,d\mu = 0$$

Поскольку подынтегральная функция неотрицательна, то 0 мы могли получить только в том случае, когда эта функция сама равна 0.

А тогда получаем, что $\varphi - \tilde{\varphi} \equiv 0$ при $f_1(x) \neq k \cdot f_0(x)$, что и требовалось.

□

Тем не менее, в реальной жизни нам довольно редко требуется делать выбор между простой гипотезой и простой альтернативой.

Поймем, что можно делать в случае, если альтернатива сложная.

Утверждение 6.1.
Пусть $X_1, \ldots, X_n$ — выборка из распределения P_θ.

И пусть для любой пары параметров θ_1 и θ_2 найдется такая функция ψ_{θ_1,θ_2}, что:

$$\frac{f_{\theta_2}(\mathbf{X})}{f_{\theta_1}(\mathbf{X})} = \psi_{\theta_1,\theta_2}\Big(T(\mathbf{X})\Big),$$

где $T(\mathbf{X})$ — некоторая статистика, причем $\forall\, \theta_1 < \theta_2$ функция ψ_{θ_1,θ_2} возрастает и непрерывна.

Научимся в таком случае различать следующие гипотезы:

$$H_0 : \theta = \theta_0, \quad H_1 : \theta > \theta_0$$

А именно, рассмотрим произвольный параметр $\theta_1 > \theta_0$.

Для проверки нулевой гипотезы хотелось бы получить критерий вида $\dfrac{f_{\theta_1}(\mathbf{X})}{f_{\theta_0}(\mathbf{X})} > k$. То есть:

$$\frac{f_{\theta_1}(\mathbf{X})}{f_{\theta_0}(\mathbf{X})} > k \iff \psi_{\theta_0,\theta_1}\Big(T(\mathbf{X})\Big) > k \iff T(\mathbf{X}) > c$$

Как и ранее, если данное условие выполняется, то первая гипотеза отвергается, иначе же принимается.

А тогда если мы хотим получить критерий для уровня значимости α, то c следует выбирать таким образом, чтобы $\mathrm{P}_{\theta_0}\big(T(\mathbf{X}) > c\big) = \alpha$ (поскольку в случае нулевой гипотезы мы хотим ошибаться с вероятностью не больше α).

Отсюда видно, что c зависит лишь от θ_0 и α и никак не зависит от альтернативного значения параметра θ_1. То есть получили критерий для тестирования простой гипотезы против сложной альтернативы (для описанной в условии ситуации).

Замечание.
Если же ψ_{θ_1,θ_2} убывает при $\theta_1 < \theta_2$, то становится возможным проверять такие гипотезы:

$$H_0 : \theta = \theta_0, \quad H_1 : \theta < \theta_0$$

Пример.
Пусть есть выборка $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2)$, причем σ^2 известна.

Протестируем следующие гипотезы:

$$H_0 : \theta = \theta_0, \quad H_1 : \theta = \theta_1 > \theta_0$$

По лемме Неймана-Пирсона для построения критерия нам требуется рассмотреть отношение функций правдоподобия:

$$\frac{f_1(\mathbf{X})}{f_0(\mathbf{X})} = \frac{\dfrac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\dfrac{1}{2\sigma^2}\sum\limits_{k=1}^{n}(X_k-\theta_1)^2\right)}{\dfrac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\dfrac{1}{2\sigma^2}\sum\limits_{k=1}^{n}(X_k-\theta_0)^2\right)} = \exp\left(\frac{1}{2\sigma^2}\sum_{k=1}^{n}\Big((X_k-\theta_0)^2-(X_k-\theta_1)^2\Big)\right) =$$

$$= \exp\left(\frac{\theta_1 - \theta_0}{2\sigma^2}\sum_{k=1}^{n}(2X_k - \theta_0 - \theta_1)\right) > k$$

Таким образом, мы уже получили некоторую критическую область. Прологарифмируем теперь полученное выражение и получим иной вид для нашей критической области:

$$\sum_{k=1}^{n}(2X_k - \theta_0 - \theta_1) > \frac{2\sigma^2 \ln k}{\theta_1 - \theta_0} = c \iff \overline{X} > \tilde{c}$$

Здесь c и $\tilde{c}$ — некоторые константы, т.к. зависят только от заранее известных константных параметров.

А тогда все, что нам теперь остается — по уровню зачимости α найти такое $\tilde{c}$, что полученное условие выполняется с вероятностью $\leqslant \alpha$ (т.к. именно в таком случае мы отвергаем нулевую гипотезу и допускаем ошибку первого рода):

$$\alpha = \mathrm{P}_{\theta_0}(\overline{X} > \tilde{c}) = \mathrm{P}_{\theta_0}\left(\sqrt{n}\frac{\overline{X} - \theta_0}{\sigma} > \sqrt{n}\frac{\tilde{c} - \theta_0}{\sigma}\right) = 1 - \Phi\left(\sqrt{n}\frac{\tilde{c} - \theta_0}{\sigma}\right) \implies \tilde{c} = \theta_0 + \frac{\sigma\Phi^{-1}(1-\alpha)}{\sqrt{n}}$$

В частности, можем попробовать посчитать мощность нашего критерия:

$$\mathrm{P}_{\theta_1}(\overline{X} > \tilde{c}) = \mathrm{P}_{\theta_1}\left(\sqrt{n}\frac{\overline{X} - \theta_1}{\sigma} > \sqrt{n}\frac{\tilde{c} - \theta_1}{\sigma}\right) = 1 - \Phi\left(\sqrt{n}\frac{\tilde{c} - \theta_1}{\sigma}\right) =$$

$$= 1 - \Phi\left(\sqrt{n}\frac{\tilde{c} - \theta_0}{\sigma} + \sqrt{n}\frac{\theta_0 - \theta_1}{\sigma}\right) = 1 - \Phi\left(\Phi^{-1}(1-\alpha) + \sqrt{n}\frac{\theta_0 - \theta_1}{\sigma}\right) \overset{\theta_0 - \theta_1 < 0}{\underset{n\to\infty}{\to}} 1 - 0 = 1$$

То есть мы на самом деле получили критерий уровня α, который при увеличении объема нашей выборки еще и сводит ошибки второго рода к нулю.

6.3. Критерии для непрерывных числовых данных

Перейдем теперь к рассмотрению способов проверки различных статистических гипотез. Как правило, применимость этих методов сильно зависит от природы данных, а потому начнем с рассмотрения методов, хорошо работающих для непрерывных данных.

И для начала разберем еще пару примеров того, как можно проверять некоторые статистические гипотезы для данных, имеющих нормальное распределение.

Примеры.

В случае наличия всего одной выборки $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2)$ мы уже умеем проверять гипотезу $H_0 : \theta = \theta_0$ для некоторого θ_0 (как в случае, когда σ^2 нам известна, так и в случае, когда неизвестна).

Пусть у нас теперь имеются выборки $X_1, \ldots, X_n \sim \mathcal{N}(\theta_1, \sigma_1^2)$ и $Y_1, \ldots, Y_m \sim \mathcal{N}(\theta_2, \sigma_2^2)$.

Научимся проверять гипотезу $H_0 : \theta_1 = \theta_2$.

1. Пусть σ_1^2 и σ_2^2 нам известны.

   Заметим, что если нулевая гипотеза верна, то

   $$\overline{X} - \overline{Y} \sim \mathcal{N}\left(0, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right) \implies \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim \mathcal{N}(0,1)$$

   Дальше можем поступить так же, как и в случае одной выборки. А именно, использовать модуль данной статистики для получения критерия.

2. Пусть $\sigma_1^2 = \sigma_2^2 = \sigma^2$, но σ^2 нам неизвестна.

   Рассмотрим тогда такую статистику:

$$\frac{\dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}}}{\sqrt{\dfrac{1}{n+m-2}\left(\dfrac{nS_x^2}{\sigma^2} + \dfrac{mS_y^2}{\sigma^2}\right)}} \sim \frac{\mathcal{N}(0,1)}{\sqrt{\dfrac{1}{n+m-2}\left(\chi_{n-1}^2 + \chi_{m-1}^2\right)}} \sim T_{n+m-2}$$

   Заметим, что полученная слева статистика никак не зависит от σ^2, а потому корректно определена.

   Таким образом, мы снова получили некоторую статистику, на основе которой уже несложно построить и сам критерий (аналогично тому, как делали ранее).

3. Пусть σ_1^2 и σ_2^2 неизвестна.

   Утверждается, что в таком случае хорошего критерия не существует, а существуют лишь приближенные некоторые приближенные (проблема Беренса-Фишера).

Тем не менее, данные у нас далеко не всегда распределены нормально. Но прежде чем приступать к рассмотрению способов проверки гипотез в общем случае, определим виды гипотез, которые вообще чаще всего встречаются в этой науке.

***Определение* 6.3.**
Пусть дана выборка $X_1, \ldots, X_n \sim \mathrm{P}$.

Гипотеза $H_0 : \mathrm{P} \sim \mathrm{P}_0$ называется простой гипотезой согласия (goodness of fit).

Тем не менее, иногда нам хочется проверить не равенство конкретному распределению, а принадлежность целому семейству распределений P_θ. В таком случае гипотеза согласия называется сложной.

***Определение* 6.4.**
Пусть даны выборки $X_1, \ldots, X_n \sim \mathrm{P}_1$ и $Y_1, \ldots, Y_m \sim \mathrm{P}_2$.

Гипотеза $H_0 : \mathrm{P}_1 \sim \mathrm{P}_2$ называется гипотезой однородности (homogeneity).

***Определение* 6.5.**
Пусть дана совместная выборка $(X_1, Y_1), \ldots, (X_n, Y_n) \sim \mathrm{P}$.

Гипотеза $H_0 : \Big(\exists\, \mathrm{P}_1, \mathrm{P}_2 : \mathrm{P}(x,y) = \mathrm{P}_1(x) \cdot \mathrm{P}_2(y)\Big)$ называется гипотезой независимости.

Научимся для начала проверять гипотезу согласия.

В непрерывном случае для этого надо проверить, что функция распределения элементов из выборки совпадает с заданной. Но т.к. саму функцию распределения наблюдений F мы не знаем, то возникает идея использовать вместо нее эмпирическую функцию распределения F_n. И, соответственно, смотреть на ее близость к заданной функции распределения по некоторой метрике.

***Определение* 6.6.**
Наиболее известные и часто используемые метрики:

1. $\sup\limits_x |\mathrm{F}_n(x) - \mathrm{F}_0(x)|$

2. $\sup\limits_x \Big(\mathrm{F}_n(x) - \mathrm{F}_0(x) \Big)$
3. $\sup\limits_x \Big(\mathrm{F}_0(x) - \mathrm{F}_n(x) \Big)$
4. $\int\limits_{-\infty}^{\infty} \Big(\mathrm{F}_n(x) - \mathrm{F}_0(x) \Big)^2 \, d\,\mathrm{F}_0(x)$

Первая из этих метрик называется статистикой Колмогорова, вторая и третья — статистиками Смирнова, последняя — статистика Крамера-фон Мизеса (также известная как омега-квадрат, ω^2).

Кроме того, первые три метрики нередко объединяют под названием статистик Колмогорова-Смирнова.

Поймем, чем эти метрики так хороши.

Теорема 6.2.
Если F_0 непрерывна и гипотеза H_0 верна, то распределение статистик 1-4 не зависит от F_0.

Доказательство.
Разберем доказательство теоремы для статистики Колмогорова.

Предположим, что F_0 строго возрастает. И положим $t = \mathrm{F}_0(x)$.

Тогда:

$$T = \sup_x |\mathrm{F}_n(x) - \mathrm{F}_0(x)| = \sup_{0 \leqslant t \leqslant 1} \big|\mathrm{F}_n(\mathrm{F}_0^{-1}(t)) - t\big|$$

Перепишем чуть иначе F_n:

$$\mathrm{F}_n(\mathrm{F}_0^{-1}(t)) = \frac{1}{n}\{k : X_k < \mathrm{F}_0^{-1}(t)\} = \frac{1}{n}\{k : \mathrm{F}_0(X_k) < t\}$$

Заметим, что $\mathrm{F}_0(X) \sim \mathrm{U}[0,1]$.

То есть $\mathrm{F}_0(X_1), \ldots, \mathrm{F}_0(X_n)$ — выборка из стандартного равномерного распределения. А сама статистика может быть переписана в следующем виде:

$$T = \sup_{0 \leqslant t \leqslant 1} \left| \frac{1}{n} \#\{k : U_k < t\} - t \right|, \text{ где } U_k = \mathrm{F}_0(X_k) \sim \mathrm{U}[0,1]$$

Значит, распределение исходной статистики никак не зависит от распределения F_0, что и требовалось. □

Мы выяснили, что распределения данных метрик не зависят от самой F_0. Более того, верны следующие утверждения:

Утверждение 6.3.

1. $\sqrt{n} \sup\limits_x |\mathrm{F}_n(x) - \mathrm{F}_0(x)| \xrightarrow{d} K$, где K — распределение Колмогорова
2. $\sqrt{n} \sup\limits_x \Big(\mathrm{F}_n(x) - \mathrm{F}_0(x) \Big) \xrightarrow{d} \mathrm{Rayleigh}\left(\frac{1}{2}\right)$, где Rayleigh — распределение Рэлея

3. $n \int\limits_{-\infty}^{\infty} \Big(\mathrm{F}_n(x) - \mathrm{F}_0(x) \Big)^2 \, d\,\mathrm{F}_0(x) \xrightarrow{d} \omega^2$, где ω^2 — распределение омега-квадрат

Все, что нам теперь остается для проверки гипотезы согласия — посчитать расстояние между эмпирической функцией распределения и заданной по одной из метрик, а затем найти соответствующую квантиль предельной функции распределения (с помощью существующих таблиц или библиотек) чтобы понять, следует нам принять гипотезу или отвергнуть.

Рассмотрим теперь гипотезу однородности.

Пусть $X_1, \ldots, X_n \sim \mathrm{F}$ и $Y_1, \ldots, Y_m \sim \mathrm{G}$. Будем проверять гипотезу $H_0 : \mathrm{F} = \mathrm{G}$.

И снова будем пробовать оценивать расстояние между эмпирическими функциями распределения F_n и G_m с помощью одной из уже описанных метрик.

Для этих метрик все так же верна теорема о том, что при справедливости нулевой гипотезы и непрерывности F и G распределение статистик не зависит от прочих распределений. Более того, при должной нормировке мы в пределе получим все те же распределения Колмогорова, Рэлея и омега-квадрат.

Утверждение 6.4.

1. $\sqrt{\dfrac{nm}{n+m}} \sup\limits_x |\mathrm{F}_n(x) - \mathrm{G}_m(x)| \xrightarrow{d} K$
2. $\sqrt{\dfrac{nm}{n+m}} \sup\limits_x \Big(\mathrm{F}_n(x) - \mathrm{G}_m(x) \Big) \xrightarrow{d} \mathrm{Rayleigh}\left(\frac{1}{2}\right)$
3. $\dfrac{nm}{n+m} \int\limits_{-\infty}^{\infty} \Big(\mathrm{F}_n(x) - \mathrm{G}_m(x) \Big)^2 \, d\,\mathrm{F}_0(x) \xrightarrow{d} \omega^2$

Докажем второе утверждение в случае, когда размеры выборок равны:

Утверждение 6.5.

Пусть даны выборки $X_1, \ldots, X_n \sim \mathrm{F}$ и $Y_1, \ldots, Y_n \sim \mathrm{G}$.

Тогда если $\mathrm{F} = \mathrm{G}$ и F, G — непрерывны, то:

$$\mathrm{P}\left(\sqrt{\frac{n}{2}} \sup_t \Big(\mathrm{F}_n(t) - \mathrm{G}_n(t) \Big) < z \right) \to \mathrm{F}_{\mathrm{Rayleigh}(1/2)}(z) = 1 - e^{-2z^2}$$

Доказательство.

Перепишем нашу вероятность:

$$\mathrm{P}\left(\sqrt{\frac{n}{2}} \sup_t \Big(\mathrm{F}_n(t) - \mathrm{G}_n(t) \Big) > z \right) = \mathrm{P}\left(n \sup_t \Big(\mathrm{F}_n(t) - \mathrm{G}_n(t) \Big) > z\sqrt{2n} \right) =$$

Поймем, как иначе можно выразить левую часть полученного неравенства. Для этого рассмотрим общий вариационный ряд наших выборок $Z_1 < Z_2 < \ldots < Z_{2n}$ и введем следующую случайную величину:

$$\eta_k := \begin{cases} 1, & \text{если } Z_k \in \mathbf{X} \\ -1, & \text{если } Z_k \in \mathbf{Y} \end{cases}, \quad S_k := \sum_{i=1}^{k} \eta_i$$

Тогда:

$$n \sup_t \Big(\mathrm{F}_n(t) - \mathrm{G}_n(t) \Big) = \sup_t \left(\sum_{k=1}^{n} \mathbb{1}\{X_k < t\} - \sum_{k=1}^{n} \mathbb{1}\{Y_k < t\} \right) = \sup_t \left(\sum_{Z_k < t} \eta_k \right) = \max_k S_k$$

А значит, исходную вероятность можно переписать так:

$$= \mathrm{P}\left(\max_k S_k > z\sqrt{2n} \right) = \mathrm{P}\left(\max_k S_k \geqslant r \right), \text{ где } r = \left[z\sqrt{2n} \right] + 1$$

По сути, мы свели теперь нашу задачу к задаче о случайных блужданиях.

В частности, любой последовательности η_k можно сопоставить некоторый диагональный путь на координатной сетке из точки $(0, 0)$ в точку $(2n, 0)$. И в терминах путей нас будет интересовать вероятность того, что этот путь пересечет прямую $y = r$.

TODO: picture

Более того, поскольку при $\mathrm{F} = \mathrm{G}$ каждое из значений η_k выпадает с вероятностью $\frac{1}{2}$, эта задача аналогична подсчету количества таких путей. А ее мы можем решить с помощью метода отражения.

А именно, рассмотрим произвольный путь, касающийся прямой $y = r$.

Тогда если мы отразим всю правую часть пути относительно прямой $y = r$, то получим некоторый новый путь из точки $(0, 0)$ в точку $(2n, 2r)$. Причем очевидно, что между путями первого и второго вида существует явная биекция.

А тогда задача свелась к тому, чтобы посчитать количество диагональных путей из точки $(0, 0)$ в точку $(2n, 2r)$. Искомая же вероятность может быть тогда переписана в таком виде:

$$\mathrm{P}\left(\max_k S_k \geqslant r \right) = \mathrm{P}\left(S_{2n} = 2r \right) = \frac{\#\{\text{пути из (0, 0) в (2n, 2r)}\}}{\binom{2n}{n}} = \frac{\binom{2n}{n+r}}{\binom{2n}{n}}$$

Все, что остается — посчитать асимптотику и убедиться, что такое выражение в пределе стремится к e^{-2z^2}.

$$\ln\left(\frac{\binom{2n}{n+r}}{\binom{2n}{n}} \right) = \ln\left(\frac{(2n)!}{(n+r)!(n-r)!} \cdot \frac{n! \cdot n!}{(2n)!} \right) = \ln\left(\frac{(n!)^2}{(n+r)!(n-r)!} \right)$$

$$= \ln\left(\frac{2\pi n \cdot \left(\frac{n}{e}\right)^{2n}}{\sqrt{2\pi \cdot (n+r)} \left(\frac{n+r}{e}\right)^{n+r} \cdot \sqrt{2\pi \cdot (n-r)} \left(\frac{n-r}{e}\right)^{n-r}} \cdot \right) + o(1) =$$

$$= \ln\left(\frac{n^{2n+1}}{(n+r)^{n+r+\frac{1}{2}} \cdot (n-r)^{n-r+\frac{1}{2}}} \right) + o(1) = \ln\left(\frac{n^{n+r+\frac{1}{2}}}{(n+r)^{n+r+\frac{1}{2}}} \cdot \frac{n^{n-r+\frac{1}{2}}}{(n-r)^{n-r+\frac{1}{2}}} \right) + o(1) =$$

$$= -\ln\left(1 + \frac{r}{n}\right)^{n+r+\frac{1}{2}} - \ln\left(1 - \frac{r}{n}\right)^{n-r+\frac{1}{2}} =$$

Распишем логарифм по Тейлору, вспомним дополнительно о том, что $r \sim \sqrt{n}$:

$$= -\left(n + r + \frac{1}{2}\right) \cdot \left(\frac{r}{n} - \frac{r^2}{2n^2} + o\left(\frac{1}{n}\right) \right) - \left(n - r + \frac{1}{2}\right) \cdot \left(-\frac{r}{n} - \frac{r^2}{2n^2} + o\left(\frac{1}{n}\right) \right) =$$

$$= -\left(r - \frac{r^2}{2n} + \frac{r^2}{n} + o(1) \right) - \left(-r + \frac{r^2}{n} - \frac{r^2}{2n} + o(1) \right) = -\frac{r^2}{n} + o(1) = -\frac{z^2 \cdot 2n}{n} + o(1) = -2z^2$$

□

6.4. Критерии для категориальных и дискретных числовых данных

Мы обсудили некоторые критерии, неплохо работающие в случае, когда наши данные имеют непрерывную природу.

Но такие критерии плохо работают для данных, принимающих лишь значения из некоторого дискретного множества. К примеру, это могут быть числовые данные, принимающие лишь целочисленные значения, либо нечисловые данные, в которых каждому элементу соответствует какая-то категория (label).

Тем не менее, для данных именно такой природы тоже существует большое количество критериев. И одним из наиболее сильных таких критериев является критерий Пирсона, также известный как критерий хи-квадрат.

Теорема 6.6 (Pearson's chi-squared test)**.**
Пусть дана выборка $X_1, \ldots, X_n \sim \mathrm{P}$, где X_i — независимы.

И пусть верна гипотеза $H_0 : \mathrm{P} \sim \mathrm{P}_0$, где

$$\mathrm{P}_0 : \begin{array}{c|c|c|c} x_1 & x_2 & \ldots & x_s \\ \hline p_1 & p_2 & \ldots & p_s \end{array}$$

Тогда верно следующее утверждение:

$$\chi^2 := \sum_{i=1}^{s} \frac{(\nu_i - np_i)^2}{np_i} \underset{n\to\infty}{\overset{H_0}{\to}} \chi^2_{s-1}, \text{ где } \nu_i = \#\{k : X_k = x_i\}$$

Доказательство.
Введем для начала несколько обозначений:

$$a_i(X_j) := \mathbb{1}\{X_j = x_i\}$$

$$\nu := \left(\frac{\nu_1 - np_1}{\sqrt{np_1}}, \ldots, \frac{\nu_s - np_s}{\sqrt{np_s}} \right)^T, \quad t := (t_1, \ldots, t_s)^T$$

Здесь ν — случайный вектор, а t — некоторый произвольный набор коэффициентов.

Рассмотрим теперь скалярное произведение $\langle \nu, t \rangle$. Попробуем переписать его в некотором более удобном виде:

$$\langle \nu, t \rangle = \sum_{i=1}^{s} \frac{\nu_i - np_i}{\sqrt{np_i}} \cdot t_i = \left[\nu_i = \sum_{j=1}^{n} a_i(X_j) \right] = \sum_{i=1}^{s} \left(\frac{t_i}{\sqrt{np_i}} \sum_{j=1}^{n} a_i(X_j) - t_i \sqrt{np_i} \right) =$$

$$= \left[t_i \sqrt{np_i} = \frac{t_i}{\sqrt{np_i}} \sum_{j=1}^{n} p_i \right] = \sum_{i=1}^{s} \left(\sum_{j=1}^{n} \frac{t_i}{\sqrt{np_i}} \big(a_i(X_j) - p_i \big) \right) =$$

$$= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left(\sum_{i=1}^{s} \frac{t_i}{\sqrt{p_i}} \big(a_i(X_j) - p_i \big) \right) = \left[Y_j := \sum_{i=1}^{s} \frac{t_i}{\sqrt{p_i}} \big(a_i(X_j) - p_i \big) \right] = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} Y_j$$

Заметим, что т.к. X_j независимы, то Y_j также независимы. А значит, можно попробовать воспользоваться центральной предельной теоремой. Для этого требуется посчитать матожидание и дисперсию Y_j.

Посчитаем матожидание:

$$\mathrm{E}\, a_i(X_j) = p_i \implies \mathrm{E}\, Y_j = \sum_{i=1}^{s} \frac{t_i}{\sqrt{p_i}} \big(\mathrm{E}\, a_i(X_j) - p_i \big) = 0$$

И дисперсию:

$$\mathrm{D}\,Y_j = \mathrm{E}\,Y_j^2 = \sum_{i=1}^{s} \frac{t_i^2}{p_i}\,\mathrm{E}\left(a_i(X_j) - p_i\right)^2 + \sum_{i\neq k} \frac{t_i t_k}{\sqrt{p_i p_k}}\,\mathrm{E}\left(\left(a_i(X_j) - p_i\right)\left(a_k(X_j) - p_k\right)\right) =$$

$$\mathrm{E}\left(a_i(X_j) - p_i\right)^2 = \mathrm{E}\,a_i(X_j)^2 - 2p_i\,\mathrm{E}\,a_i(X_j) + p_i^2 = p_i - p_i^2 = p_i(1-p_i)$$

$$\mathrm{E}\left(\left(a_i(X_j) - p_i\right)\left(a_k(X_j) - p_k\right)\right) = \mathrm{E}\,a_i(X_j)a_k(X_j) - p_i\,\mathrm{E}\,a_k(X_j) - p_k\,\mathrm{E}\,a_i(X_j) + p_i p_k = -p_i p_k$$

$$= \sum_{i=1}^{s} \frac{t_i^2}{p_i} p_i(1-p_i) + \sum_{i\neq k} \frac{t_i t_k}{\sqrt{p_i p_k}}(-p_i p_k) = \sum_{i=1}^{s} t_i^2 - \sum_{i=1}^{s} (t_i\sqrt{p_i})^2 - \sum_{i\neq k} t_i\sqrt{p_i}\cdot t_k\sqrt{p_k} =$$

$$= \sum_{i=1}^{s} t_i^2 - \left(\sum_{i=1}^{s} t_i\sqrt{p_i}\right)^2$$

Соберем вместе все полученные результаты:

$$q(t_1, \ldots, t_s) := \sum_{i=1}^{s} t_i^2 - \left(\sum_{i=1}^{s} t_i\sqrt{p_i}\right)^2$$

$$\langle \nu, t\rangle = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} Y_j, \quad \mathrm{E}\,Y_j = 0, \quad \mathrm{D}\,Y_j = q(t_1, \ldots, t_s)$$

По центральной предельной теореме мы имеем:

$$\langle \nu, t\rangle \xrightarrow[n\to\infty]{d} \mathcal{N}(0, q(t_1, \ldots, t_s))$$

Поскольку мы знаем вид характеристической функции для нормального распределения, то можем посчитать предельный вид следующей величины:

$$\mathrm{E}\exp\left(i\,\lambda\,\langle \nu, t\rangle\right) \to \exp\left(-\frac{\lambda^2}{2} q(t_1, \ldots, t_s)\right)$$

А т.к. это верно для любого λ, то при $\lambda = 1$ получаем в точности предельный вид характеристической функции вектора ν:

$$\varphi_\nu(t) = \mathrm{E}\exp\left(i\,\langle \nu, t\rangle\right) \to \exp\left(-\frac{1}{2} q(t_1, \ldots, t_s)\right)$$

Заметим, что $q(t_1, \ldots, t_s)$ — некоторая квадратичная функция. Положим Q — соответствующая этой квадратичной функции матрица.

Если же мы вспомним, как выглядит характеристическая функция гауссовского вектора, то запросто получим и предельное распределение вектора ν:

$$\varphi_\nu(t) \to \exp\left(-\frac{1}{2} q(t_1, \ldots, t_s)\right) = \exp\left(-\frac{1}{2}\,\langle Qt, t\rangle\right) \implies \nu \xrightarrow{d} \eta \sim \mathcal{N}(0, Q)$$

А тогда:

$$\chi^2 = \|\nu\|^2 \to \|\eta\|^2 = \sum_{i=1}^{s} \eta_i^2$$

Остается лишь понять, как распределена сумма квадратов компонент вектора η.

Для этого воспользуемся похожим трюком, что и при доказательстве леммы Фишера. А именно, рассмотрим произвольную ортогональную матрицу C следующего вида:

$$C = \begin{pmatrix} \sqrt{p_1} & \cdots & \sqrt{p_s} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{pmatrix}$$

Поскольку C — ортогональная матрица, то $\eta \stackrel{d}{=} C\eta$.

Поймем, как устроена ковариационная матрица вектора $C\eta$. Для этого выпишем его характеристическую функцию:

$$\varphi_{C\eta}(\tau) = \varphi_\eta(C^{-1}\tau) \stackrel{C^{-1}=C^T}{=} \varphi_\eta(C^T\tau) = \exp\left(-\frac{1}{2}\left\langle QC^T\tau,\, C^T\tau\right\rangle\right)$$

$$\left\langle QC^T\tau,\, C^T\tau\right\rangle = \sum_{i=1}^{s}(C^T\tau)_i^2 - \left(\sum_{i=1}^{s}(C^T\tau)_i\cdot\sqrt{p_i}\right)^2 = \left\|C^T\tau\right\|^2 - \left(CC^T\tau\right)_1^2 = \|\tau\|^2 - \tau_1^2 = \sum_{i=2}^{s}\tau_i^2$$

То есть мы получили, что ковариационная матрица вектора $C\eta$ содержит только $s-1$ единицу на главной диагонали. А значит, распределение $\|C\eta\|^2$ в точности соответствует распределению χ^2_{s-1}, что и требовалось. □

Замечание.

В доказанной нами теореме мы пользовались тем, что наше пространство конечно.

Но на самом деле, ее можно использовать и в случае, когда наше пространство бесконечно (в т.ч. если оно непрерывно). Для этого достаточно выбрать произвольное разбиение пространства $\mathcal{X} = \mathcal{X}_1 \sqcup \ldots \sqcup \mathcal{X}_s$ и определить $p_i := \mathrm{P}(X \in \mathcal{X}_i)$.

Замечание.

Кроме того, мы в данной теореме нигде не используем структуру нашего пространства. Все, что нас интересует — это вероятность p_i попадания события в каждый из блоков. А потому полученный критерий работает и в случае категориальных данных.

Мы теперь знаем, как с помощью критерия Пирсона проверять простые гипотезы согласия. Для сложных же гипотез верна следующая теорема:

Теорема 6.7.

Пусть дана выборка $X_1, \ldots, X_n \sim \mathrm{P}$.

И пусть дано семейство распределений $\{\mathrm{P}_\theta \mid \theta \in \Theta \subset \mathbb{R}^d\}$, такое что:

$$\mathrm{P}_\theta : \begin{array}{c|c|c|c} x_1 & x_2 & \ldots & x_s \\ \hline p_1(\theta) & p_2(\theta) & \ldots & p_s(\theta) \end{array}$$

Определим $\chi^2(\theta)$ следующим образом:

$$\chi^2(\theta) := \sum_{i=1}^{s}\frac{\left(\nu_i - np_i(\theta)\right)^2}{np_i(\theta)},\ \text{где } \nu_i = \#\{k : X_k = x_i\}$$

И выберем такое $\widehat{\theta}$, что $\chi^2(\widehat{\theta}) = \inf\limits_\theta \chi^2(\theta)$.

Тогда если верна гипотеза $H_0 : \mathrm{P} \in \{\mathrm{P}_\theta\}$, то верно следующее утверждение:

$$\chi^2(\widehat{\theta}) = \sum_{i=1}^{s}\frac{\left(\nu_i - np_i(\widehat{\theta})\right)^2}{np_i} \underset{n\to\infty}{\overset{H_0}{\to}} \chi^2_{s-d-1}$$

Доказательство.

Без доказательства. □

Замечание.

На самом деле достаточно использовать не точное значение $\widehat{\theta}$, а лишь какую-нибудь близкую оценку, т.к. при небольшом изменении параметра значение функции $\chi^2(\theta)$ также меняется несильно.

Замечание.

Количество степеней свободы распределения хи-квадрат следует воспринимать следующим образом.

Будем считать, что наше распределение однозначно задается с помощью $s-1$ параметра, т.к. все вероятности, кроме одной выбираются случайным образом, последняя же получается однозначно. А поскольку мы делаем оценку для d параметров, то количество случайных параметров уменьшается до $s-d-1$.

Эта теорема хороша не только тем, что позволяет строить критерии для сложных гипотез согласия, но и тем, что с ее помощью довольно просто строятся критерии для гипотез однородности и независимости.

Утверждение 6.8.

Пусть даны выборки $X_1, \ldots, X_{n_1} \sim \mathrm{F}$ и $Y_1, \ldots, Y_{n_2} \sim \mathrm{G}$, где F и G — дискретные распределения со значениями $x_1, \ldots, x_s$.

И пусть мы проверяем гипотезу однородности $H_0 : \mathrm{F} = \mathrm{G}$.

Посчитаем для каждого исхода количество соответствующих событий:

	x_1	x_2	$\ldots$	x_s
X	$\nu_{1,1}$	$\nu_{1,2}$	$\ldots$	$\nu_{1,s}$
Y	$\nu_{2,1}$	$\nu_{2,2}$	$\ldots$	$\nu_{2,s}$
	$\nu_{\cdot,1}$	$\nu_{\cdot,2}$	$\ldots$	$\nu_{\cdot,s}$

, где $\nu_{\cdot,i} = \nu_{1,i} + \nu_{2,i}$

И построим оценки для вероятностей как $\widehat{p}_i = \dfrac{\nu_{\cdot,i}}{n_1+n_2}$.

Тогда если гипотеза H_0 верна, то верно следующее утверждение:

$$\sum_{\substack{1\leqslant i\leqslant 2\\ 1\leqslant j\leqslant s}} \frac{(\nu_{i,j} - n_i\widehat{p}_j)^2}{n_i\widehat{p}_j} \xrightarrow{d} \chi^2_{s-1}$$

Доказательство.

Без доказательства. □

Замечание.

Неформальное доказательство можно описать следующим образом: внутри полученной таблички каждая из строк содержит по $s-1$ независимому параметру, а т.к. мы строим по ней оценку для $s-1$ параметра, то в результате случайность остается лишь в $2(s-1)-(s-1) = s-1$ параметре.

Утверждение 6.9.

Пусть дана совместная выборка $(X_1, Y_1), \ldots, (X_n, Y_n) \sim \mathrm{P}$.

И пусть мы проверяем гипотезу независимости $H_0 : \Big(\exists\, \mathrm{P}_1,\, \mathrm{P}_2 : \mathrm{P}(x,y) = \mathrm{P}_1(x)\cdot \mathrm{P}_2(y)\Big)$.

Посчитаем для каждого исхода количество соответствующих событий:

$X \setminus Y$	y_1	y_2	$\cdots$	y_r	
x_1	$\nu_{1,1}$	$\nu_{1,2}$	$\cdots$	$\nu_{1,r}$	$\nu_{1,\cdot}$
x_2	$\nu_{2,1}$	$\nu_{2,2}$	$\cdots$	$\nu_{2,r}$	$\nu_{2,\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_s	$\nu_{s,1}$	$\nu_{s,2}$	$\cdots$	$\nu_{s,r}$	$\nu_{s,\cdot}$
	$\nu_{\cdot,1}$	$\nu_{\cdot,2}$	$\cdots$	$\nu_{\cdot,r}$	

, где $\nu_{\cdot,i} = \sum\limits_{j=1}^{s} \nu_{j,i},\ \nu_{i,\cdot} = \sum\limits_{j=1}^{r} \nu_{i,j}$

И построим на основе полученной таблицы оценки для вероятностей как $\widehat{p}_{i,j} = \dfrac{\nu_{i,\cdot} \cdot \nu_{\cdot,j}}{n^2}$.

Тогда если гипотеза H_0 верна, то верно следующее утверждение:

$$\sum_{\substack{1 \leqslant i \leqslant s \\ 1 \leqslant j \leqslant r}} \frac{(\nu_{i,j} - n\widehat{p}_{i,j})^2}{n\widehat{p}_{i,j}} \xrightarrow{d} \chi^2_{(s-1)(r-1)}$$

Доказательство.

Без доказательства. □

Замечание.

Снова приведем неформальное доказательство сего утверждения: сама табличка изначально содержит $rs - 1$ независимый параметр, оценку же мы приводим для $(r-1) + (s-1)$ параметров (соответственно, это $\nu_{i,\cdot}$ и $\nu_{\cdot,j}$, на основе которых вычисляются $\widehat{p}_{i,j}$). Отсюда получаем, что количество степеней свободы хи-квадрат — в точности $rs - 1 - (r-1) - (s-1) = (r-1)(s-1)$.

6.5. Ранги. Наиболее мощные ранговые критерии

Предположим, что наши данные непрерывны, но устроены таким образом, что мы умеем лишь сравнивать наблюдения между собой и говорить, какое из них “меньше”. Причем будем считать такое отношение транзитивно.

Такие данные называются ранговыми, и именно их мы и будем обсуждать сейчас. Но для этого нам сначала придется чуть ближе познакомиться с порядковыми статистиками.

Напоминание.

Пусть $X_1, \ldots, X_n$ — некоторые наблюдения, обладающие описанным выше свойством.

$X_{(i)}$ — i-ая порядковая статистика, она же значение i-ого по величине наблюдения из $X_1, \ldots, X_n$.

Обозначение.

Пусть $X_1, \ldots, X_n$ — некоторые наблюдения, обладающие описанным выше свойством.

Введем следующие обозначения:

$\mathbf{X} := (X_1, \ldots, X_n)$ — вектор наблюдений

$\mathbf{X}_{(\cdot)} := (X_{(1)}, \ldots, X_{(n)})$ — вектор порядковых статистик (вариационный ряд)

***Определение* 6.7.**

Пусть $X_1, \ldots, X_n \sim \mathrm{F}$ — непрерывно распределенные наблюдения. Непрерывность будем требовать для того, чтобы вероятность совпадения двух событий была равна 0.

Определим ранг R_i как позицию элемента X_i в вариационном ряду $\mathbf{X}_{(\cdot)}$, т.е. $X_i = X_{(R_i)}$.

$\mathbf{R} := (R_1, \ldots, R_n)$ — вектор рангов.

Замечание.

Из определения видно, что вектор $\mathbf{R}$ является перестановкой чисел $\{1, \ldots, n\}$, т.е. $\mathbf{R} \in S_n$.

Теорема 6.10.

Пусть f — совместная плотность случайного вектора $\mathbf{X}$.

Тогда вариационный ряд $\mathbf{X}_{(\cdot)}$ имеет распределение с плотностью

$$\overline{f}(\mathbf{x}_{(\cdot)}) := \sum_{\mathbf{r} \in S_n} f(x_{(r_1)}, \ldots, x_{(r_n)})$$

Кроме того,

$$\mathrm{P}(\mathbf{R} = \mathbf{r} \,|\, \mathbf{X}_{(\cdot)} = \mathbf{x}_{(\cdot)}) = \frac{f(x_{(r_1)}, \ldots, x_{(r_n)})}{\overline{f}(\mathbf{x}_{(\cdot)})}$$

Доказательство.

Пусть $\mathcal{A}_{(\cdot)}$ — σ-алгебра борелевских подмножеств $\mathcal{X}_{(\cdot)}$, где $\mathcal{X}_{(\cdot)}$ — подпространство $\mathcal{X}$, на котором определены все порядковые статистики.

Рассмотрим произвольное множество $A \in \mathcal{A}_{(\cdot)}$. Тогда для него будет верно:

$$\mathrm{P}(\mathbf{X}_{(\cdot)} \in A) = \int\limits_{\mathbf{x}_{(\cdot)} \in A} f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n = \sum_{\mathbf{r} \in S_n} \int\limits_{\substack{\mathbf{x}_{(\cdot)} \in A \\ x_i = x_{(r_i)}}} f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n =$$

Заметим, что по вектору порядковых статистик $\mathbf{x}_{(\cdot)}$ и вектору рангов $\mathbf{r}$ однозначно восстанавливается исходный вектор $\mathbf{x}$. А потому можем сделать замену переменной и перейти к интегрированию по множеству $\mathbf{x} \in A$.

В частности, поскольку матрица замены соответствует некоторой перестановке элементов, модуль якобиана замены будет равен единице.

$$= \sum_{\mathbf{r} \in S_n} \int\limits_{\mathbf{x} \in A} f(x_{(r_1)}, \ldots, x_{(r_n)})\, dx_{(1)} \ldots dx_{(n)} \overset{f \geqslant 0}{=} \int\limits_{\mathbf{x} \in A} \sum_{\mathbf{r} \in S_n} f(x_{(r_1)}, \ldots, x_{(r_n)})\, dx_{(1)} \ldots dx_{(n)}$$

С другой стороны, если $\overline{f}$ — плотность распределения вариационного ряда, то для любого $A \in \mathcal{A}_{(\cdot)}$ выполняется:

$$\mathrm{P}(\mathbf{X}_{(\cdot)} \in A) = \int\limits_{\mathbf{x} \in A} \overline{f}(\mathbf{x}_{(\cdot)})\, dx_{(1)} \ldots dx_{(n)}$$

А значит,

$$\overline{f}(\mathbf{x}_{(\cdot)}) = \sum_{\mathbf{r} \in S_n} f(x_{(r_1)}, \ldots, x_{(r_n)})$$

В частности, на основе приведённых выше выкладок мы можем запросто посчитать следующую совместную вероятность:

$$\mathrm{P}(\mathbf{R} = \mathbf{r},\, \mathbf{X}_{(\cdot)} \in A) = \int\limits_{\substack{\mathbf{x}_{(\cdot)} \in A \\ x_i = x_{(r_i)}}} f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n = \int\limits_{\mathbf{x} \in A} f(x_{(r_1)}, \ldots, x_{(r_n)})\, dx_{(1)} \ldots dx_{(n)}$$

А тогда вероятность $\mathrm{P}(\mathbf{R} = \mathbf{r} \,|\, \mathbf{X}_{(\cdot)} = \mathbf{x}_{(\cdot)})$ вычисляется по формуле условной вероятности просто как отношение соответствующих плотностей, т.е.

$$\mathrm{P}(\mathbf{R} = \mathbf{r} \,|\, \mathbf{X}_{(\cdot)} = \mathbf{x}_{(\cdot)}) = \frac{f(x_{(r_1)}, \ldots, x_{(r_n)})}{\overline{f}(\mathbf{x}_{(\cdot)})}$$

□

Мы сейчас доказали теорему о распределении рангов произвольного случайного вектора. Когда же мы работаем с реальными данными, мы все-таки привыкли предполагать, что наши наблюдения в идеальном случае независимы и одинаково распределены. Как мы сейчас увидим, в таком случае распределение вектора рангов и вариационного ряда принимает довольно простой и приятный вид.

Теорема 6.11.

Пусть $X_1, \ldots, X_n$ — независимые одинаково распределенные случайные величины с совместной плотностью q.

Тогда случайные векторы $\mathbf{R}$ и $\mathbf{X}_{(\cdot)}$ независимы, причем

$$\mathrm{P}(\mathbf{R} = \mathbf{r}) = \frac{1}{n!}, \qquad \overline{q}(\mathbf{x}_{(\cdot)}) = n!\, q(\mathbf{x}_{(\cdot)})$$

Доказательство.

Заметим, что если случайные величины независимы и одинаково распределены, то

$$q(x_{(1)}, \ldots, x_{(n)}) = q(x_{(r_1)}, \ldots, x_{(r_n)}) \;\; \forall r \in S_n$$

А тогда

$$\overline{q}(\mathbf{x}_{(\cdot)}) = \sum_{r \in S_n} q(x_{(r_1)}, \ldots, x_{(r_n)}) = n!\, q(\mathbf{x}_{(\cdot)})$$

В частности, $\forall \mathbf{x}_{(\cdot)} \in \mathcal{X}_{(\cdot)}$ выполняется следующее равенство:

$$\mathrm{P}(\mathbf{R} = \mathbf{r} \,|\, \mathbf{X}_{(\cdot)} = \mathbf{x}_{(\cdot)}) = \frac{q(x_{(r_1)}, \ldots, x_{(r_n)})}{\overline{q}(\mathbf{x}_{(\cdot)})} = \frac{q(\mathbf{x}_{(\cdot)})}{n!\, q(\mathbf{x}_{(\cdot)})} = \frac{1}{n!}$$

То есть получили, что распределение вектора рангов никак не зависит от распределения вариационного ряда. А значит, векторы $\mathbf{R}$ и $\mathbf{X}_{(\cdot)}$ независимы. □

Осталось теперь немного поговорить о том, что такое ранговые критерии и как они устроены.

***Определение* 6.8.**

Ранговые критерии — такие критерии, что их критическая область зависит только от вектора рангов $\mathbf{R}$.

Теорема 6.12.

Пусть дана выборка $X_1, \ldots, X_n$, такая что $\mathbf{X} \sim \mathrm{P}$.

И пусть проверяются следующие простые гипотезы:

$$H_0 : \mathrm{P} = \mathrm{P}_0, \quad H_1 : \mathrm{P} = \mathrm{P}_1$$

Тогда если распределение P_0 соответствует совместному распределению независимых одинаково распределенных случайных величин, то наиболее мощный ранговый критерий имеет следующую структуру:

$$\varphi(\mathbf{X}) = \begin{cases} 1, & \mathrm{P}_1(\mathbf{R} = \mathbf{r}) \geqslant k \\ 0, & \mathrm{P}_1(\mathbf{R} = \mathbf{r}) < k \end{cases}, \text{ где } \mathbf{r} \text{ — ранги элементов } \mathbf{X}$$

Доказательство.

Воспользуемся леммой Неймана-Пирсона.

По определению ранговых критериев наша критическая область зависит только от вектора рангов, а потому и интересовать нас будет только распределение вектора рангов.

Но если наблюдения независимы, то $\mathrm{P}_0(\mathbf{R} = \mathbf{r}) = \frac{1}{n!}$, т.е. эта величина является константой. □

Но, как и прежде, нас чаще все-таки будет интересовать проверка простой гипотезы против сложной альтернативы. А потому введем понятие локально наиболее мощного критерия.

***Определение* 6.9.**

Пусть нам дано семейство вероятностных распределений $\{\mathrm{P}_\delta\}$, запараметризованных некоторым параметром δ.

И пусть совместное распределение элементов выборки $X_1, \ldots, X_n$ лежит в этом семействе.

Будем проверять следующие гипотезы:

$$H_0 : \delta = 0, \quad H_1 : \delta > 0$$

Критерий называется локально наиболее мощным, если $\exists \varepsilon > 0 : \forall \delta_0 \in (0, \varepsilon)$ φ — наиболее мощный критерий против альтернативы $H_1 : \delta = \delta_0$.

Замечание.

В дальнейшем мы будем считать, что плотности распределений элементов выборки принадлежат семейству плотностей $\{f(X, \theta)\}$, а совместная плотность распределения имеет следующий вид:

$$q_\delta(\mathbf{X}) = \prod_{i=1}^{n} f(X_i, \delta c_i), \quad \delta > 0$$

Кроме того, для определения последующих критериев мы будем требовать от плотностей некоторых условий регулярности. А именно:

Условия регулярности.

1. $f(x, \theta)$ абсолютно непрерывна по θ (или $f(x, \cdot) \in C^1$)
2. $\exists\ f'(x, 0) = \lim\limits_{\theta \to 0} \dfrac{f(x, \theta) - f(x, 0)}{\theta}$
3. $\lim\limits_{\theta \to 0} \int\limits_{-\infty}^{\infty} |f'(x, \theta)|\, dx = \int\limits_{-\infty}^{\infty} |f'(x, 0)|\, dx < \infty$

Перейдем теперь непосредственно к основным гипотезам и критериям для их проверки.

6.6. Ранговые критерии. Гипотеза случайности

Начнем с определения гипотезы случайности.

***Определение* 6.10.**

$X_1, \ldots, X_n$ — произвольная выборка, имеющая совместную плотность q.

Гипотеза $H_0 : q(X_1, \ldots, X_n) = \prod\limits_{i=1}^{n} f(X_i)$ называется гипотезой случайности.

Иными словами, эта гипотеза попросту предполагает, что все наблюдения независимы и одинаково распределены.

Мы уже знаем как распределены $\mathbf{R}$ и $\mathbf{X}_{(\cdot)}$ в случае, если гипотеза случайности верна.

Получим теперь вид локально наиболее мощного критерия для проверки этой гипотезы, чтобы на основе него затем построить остальные критерии.

Обозначение.
Будем называть метками (score) функции вида:

$$a_n(i, f) = \mathrm{E}_0 \frac{f'(X_{(i)}, 0)}{f(X_{(i)}, 0)}$$

Здесь n — размер выборки, E_0 — матожидание, соответствующее плотности $f(x, 0)$.

Теорема 6.13.
Пусть выполняются условия регулярности.

Тогда локально наиболее мощный ранговый критерий для проверки гипотезы случайности — критерий с критической областью следующего вида:

$$\sum_{i=1}^{n} c_i \, a_n(R_i, f) \geqslant k$$

Доказательство.
Распишем вероятность встретить конкретный вектор рангов при распределении P_δ:

$$\mathrm{P}_\delta(\mathbf{R} = \mathbf{r}) = \int\limits_{\mathbf{R}=\mathbf{r}} q_\delta(x_1, \ldots, x_n)\, dx_1 \ldots dx_n = \int\limits_{\mathbf{R}=\mathbf{r}} \prod_{i=1}^{n} f(x_i, \delta c_i)\, dx_1 \ldots dx_n =$$

$$= \int\limits_{\mathbf{R}=\mathbf{r}} \prod_{i=1}^{n} f(x_i, 0)\, dx_1 \ldots dx_n + \delta \cdot \int\limits_{\mathbf{R}=\mathbf{r}} \frac{1}{\delta} \left(\prod_{i=1}^{n} f(x_i, \delta c_i) - \prod_{i=1}^{n} f(x_i, 0) \right) dx_1 \ldots dx_n =$$

Заметим, что первый интеграл — вероятность появления конкретного ранга при справедливости нулевой гипотезы. Это значение нам уже известно и равно $\frac{1}{n!}$.

Второй же интеграл можно несколько переписать, воспользовавшись свойством следующего вида:

$$\prod_{i=1}^{n} a_i - \prod_{i=1}^{n} b_i = \sum_{i=1}^{n} \left((a_i - b_i) \cdot \prod_{j=1}^{i-1} a_j \cdot \prod_{j=i+1}^{n} b_j \right)$$

А тогда получаем:

$$= \frac{1}{n!} + \delta \cdot \int\limits_{\mathbf{R}=\mathbf{r}} \sum_{i=1}^{n} \left(c_i \cdot \frac{f(x_i, \delta c_i) - f(x_i, 0)}{\delta c_i} \cdot \prod_{j=1}^{i-1} f(x_j, \delta c_j) \cdot \prod_{j=i+1}^{n} f(x_j, 0) \right) dx_1 \ldots dx_n = s$$

Воспользуемся описанными выше условиями регулярностями и перейдем к пределу по δ:

$$= \frac{1}{n!} + \delta \cdot \int\limits_{\mathbf{R}=\mathbf{r}} \sum_{i=1}^{n} \left(c_i f'(x_i, 0) \cdot \prod_{j \neq i} f(x_j, 0) \right) dx_1 \ldots dx_n + o(\delta) =$$

$$= \frac{1}{n!} + \delta \cdot \sum_{i=1}^{n} \int\limits_{\mathbf{R}=\mathbf{r}} \left(c_i \cdot \frac{f'(x_i, 0)}{f(x_i, 0)} \cdot \prod_{j=1}^{n} f(x_j, 0) \right) dx_1 \ldots dx_n + o(\delta) =$$

$$= \frac{1}{n!} + \delta \cdot \sum_{i=1}^{n} c_i \, \mathrm{E}_0 \left(\frac{f'(x_i, 0)}{f(x_i, 0)} \,\middle|\, \mathbf{R} = \mathbf{r} \right) + o(\delta) = \frac{1}{n!} + \delta \cdot \sum_{i=1}^{n} c_i \, \mathrm{E}_0 \, \frac{f'(x_{(r_i)}, 0)}{f(x_{(r_i)}, 0)} + o(\delta) =$$

$$= \frac{1}{n!} + \delta \cdot \sum_{i=1}^{n} c_i a_n(r_i, f) + o(\delta)$$

Вспомним теперь, что при фиксированной δ критическая область имеет следующий вид:

$$\mathrm{P}_\delta(\mathbf{R} = \mathbf{r}) \geqslant \tilde{k}$$

А тогда можно найти такое k, что в некоторой окрестности нуля будет выполняться:

$$\frac{1}{n!} + \delta \cdot \sum_{i=1}^{n} c_i a_n(r_i, f) + o(\delta) \geqslant \tilde{k}, \;\; \delta > 0 \implies \sum_{i=1}^{n} c_i a_n(r_i, f) \geqslant k$$

□

Мы теперь знаем общий вид локально наиболее мощного критерия для гипотезы случайности. Построим на основе него несколько критериев для тестирования этой гипотезы против гипотезы сдвига.

А именно, пусть дана пара выборок $X_1, \ldots, X_m$ и $X_{m+1}, \ldots, X_{m+n}$, и мы проверяем гипотезы следующего вида:

$$H_0 : q_0(\mathbf{X}) = \prod_{i=1}^{n+m} f(X_i), \quad H_1 : q_\delta(\mathbf{X}) = \prod_{i=1}^{m} f(X_i - \delta) \cdot \prod_{i=1}^{n} f(X_{m+i})$$

1. ***Критерий нормальных меток***

   Пусть наши выборки были получены из нормального распределения, т.е.

   $$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad \text{и} \quad f(x, \theta) := f(x + \theta)$$

   Поймем, как в этом случае устроены метки:

   $$\frac{f'(x, 0)}{f(x, 0)} = (\ln f(x, 0))' = \left(-\frac{x^2}{2} \right)' = -x$$

   $$a_N(i, f) = \mathrm{E}_0 \, \frac{f'(X_{(i)}, 0)}{f(X_{(i)}, 0)} = \mathrm{E}_0(-X_{(i)}) = -\,\mathrm{E}\, \Phi^{-1}(\mathrm{U}_{(i)})$$

   А тогда, поскольку $c_1, \ldots, c_m$ равны -1 и $c_{m+1}, \ldots, c_{m+n}$ равны 0, сама критическая область может быть записана как

   $$\sum_{i=1}^{m} -a_{n+m}(R_i, f) = \sum_{i=1}^{m} \mathrm{E}\, \Phi^{-1}(\mathrm{U}_{(R_i)}) \geqslant k$$

2. ***Критерий ван дер Вардена***

   Главная проблема критерия нормальных меток заключается в сложности вычисления $a_n(i, f)$.

   Но заметим, что в силу $\mathrm{E}\,\Phi^{-1}(\mathrm{U}_{(i)}) \approx \Phi^{-1}(\mathrm{E}\,\mathrm{U}_{(i)})$ его можно переписать в более удобном виде:

$$\sum_{i=1}^{m} a_{n+m}(R_i, f) = \sum_{i=1}^{m} \Phi^{-1}\left(\frac{R_i}{n+m+1}\right) \geqslant k$$

3. ***Критерий Вилкóксона***

   Рассмотрим теперь случай логистического распределения с плотностью $f(x) = \dfrac{e^{-x}}{(1+e^{-x})^2}$.

   Снова поймем, как в этом случае будут устроены метки:

$$\frac{f'(x,0)}{f(x,0)} = (\ln f(x,0))' = \left(-x - 2\ln\left(1+e^{-x}\right)\right)' = -1 + 2\frac{e^{-x}}{1+e^{-x}} = \frac{e^{-x}-1}{e^{-x}+1}$$

$$\mathrm{F}(x) = \frac{1}{1+e^{-x}}, \;\; \mathrm{F}^{-1}(y) = -\ln\left(\frac{1}{y} - 1\right) = \ln\frac{y}{1-y}$$

$$\frac{f'}{f}\left(\mathrm{F}^{-1}(y)\right) = \left(\frac{e^{-x}-1}{e^{-x}+1}\right) \circ \ln\frac{y}{1-y} = \frac{\dfrac{1-y}{y} - 1}{\dfrac{1-y}{y} + 1} = 1 - 2y$$

$$a_N(i, f) = \mathrm{E}_0\left(\frac{f'}{f}(X_{(i)}, 0)\right) = \mathrm{E}\left(\frac{f'}{f}\left(\mathrm{F}^{-1}(\mathrm{U}_{(i)}), 0\right)\right) = \mathrm{E}\left(1 - 2\mathrm{U}_{(i)}\right) = 1 - \frac{2i}{N+1}$$

   Как и раньше, какие-то c_i у нас равны -1, а какие-то равны 0. Сам же критерий может быть переписан теперь в таком виде:

$$\sum_{i=1}^{m} R_i \geqslant k$$

6.7. Ранговые критерии. Гипотеза симметрии

Перейдем теперь к гипотезе симметрии.

***Определение* 6.11.**
$X_1, \ldots, X_n$ — произвольная выборка, имеющая совместную плотность q.

Гипотеза $H_0 : q(X_1, \ldots, X_n) = \prod_{i=1}^{n} f(X_i),\ f(x) = f(-x)$ называется гипотезой симметрии.

По сути, эта та же гипотеза случайности, но дополнительно проверяется, что распределение симметрично относительно нуля.

Заведем еще несколько обозначений, которые потребуются при обсуждении критериев для тестирования этой гипотезы.

Обозначение.
Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — исходная выборка. Тогда:

$|\mathbf{X}| := (|X_1|, \ldots, |X_n|)$ — вектор абсолютных величин.

$\mathbf{R}^+ := R(|\mathbf{X}|)$ — вектор рангов абсолютных величин.

$\mathrm{sign}(\mathbf{X}) := (\mathrm{sign}(X_1), \ldots, \mathrm{sign}(X_2))$ — вектор знаков исходных величин.

В соответствии с введенными обозначениями можем попытаться понять, как устроены распределения соответствующих векторов. Для этого докажем следующую теорему:

Теорема 6.14.
$X_1, \ldots, X_n$ — произвольная выборка, имеющая совместную плотность q.

Если гипотеза симметрии верна, то векторы $\operatorname{sign}(\mathbf{X})$, $\mathbf{R}^+$, $|\mathbf{X}|_{(\cdot)}$ независимы, причем:

$$\mathrm{P}(\operatorname{sign}(\mathbf{X}) = \mathbf{s}) = \frac{1}{2^n}, \quad \mathrm{P}(\mathbf{R}^+ = \mathbf{r}) = \frac{1}{n!}, \quad \overline{q}(|\mathbf{X}|_{(\cdot)}) = n! \cdot 2^n \cdot \prod_{i=1}^{n} f(X_i)$$

Доказательство.
Независимость следует из того, что по этой тройке векторов мы всегда можем однозначно восстановить исходный вектор $\mathbf{X}$, ровно как и наоборот.

Перейдем теперь к вероятности встретить конкретный вектор знаков.

Заметим, что если наше распределение симметрично относительно нуля, то каждое наблюдение с равной вероятностью оказывается положительным либо отрицательным. А т.к. наблюдения независимы, то вероятность вектора — произведение вероятностей событий по каждой координате.

Что же касается распределений $\mathbf{R}^+$ и $|\mathbf{X}|_{(\cdot)}$, то для таких векторов доказательство уже было представлено выше. □

Предположим теперь, что мы хотим тестировать гипотезу симметрии против гипотезы сдвига, т.е. для некоторой выборки $X_1, \ldots, X_n$ проверяем гипотезы:

$$H_0 : q_0(\mathbf{X}) = \prod_{i=1}^{n} f(X_i), \quad H_1 : q_\delta(\mathbf{X}) = \prod_{i=1}^{n} f(X_i - \delta)$$

Аналогично гипотезе случайности, получим для этого случая вид локально наиболее мощного критерия для тестирования против гипотезы сдвига.

Обозначение.
Введем альтернативное определение меток для получения вида локально наиболее мощного критерия

$$a_n^+(i, f) := \mathrm{E}_0 \frac{f'(|X|_{(i)})}{f(|X|_{(i)})}$$

Теорема 6.15.
Локально наиболее мощный ранговый критерий для тестирования гипотезы симметрии против гипотезы сдвига — критерий с критической областью следующего вида:

$$\sum_{i=1}^{n} a_n^+(R_i^+, f) \operatorname{sign}(X_i) \geqslant k$$

Доказательство.
Оставляется в качестве упражнения, в точности повторяет доказательство теоремы для гипотезы случайности. □

Замечание.
Перепишем вид критической области в более удобном виде:

$$\sum_{X_i > 0} a_n^+(R_i^+, f) = \frac{1}{2}\left(\sum_{i=1}^{n} a_n^+(R_i^+, f) \operatorname{sign}(X_i) + \sum_{i=1}^{n} a_n^+(R_i^+, f)\right) =$$

$$= \frac{1}{2}\left(\sum_{i=1}^{n} a_n^+(R_i^+, f)\,\mathrm{sign}(X_i) + \sum_{i=1}^{n} a_n^+(i, f)\right) \geqslant \frac{1}{2}(k + C) = \tilde{k}$$

Соответственно, критическая область локально наиболее мощного критерия для тестирования гипотезы симметрии против гипотезы сдвига представима как

$$\sum_{X_i>0} a_n^+(R_i^+, f) \geqslant k$$

1. ***Одновыборочный критерий Вилкóксона***

   Пусть дана выборка $X_1, \ldots, X_n \sim f$, где $f(x)$ — плотность логистического распределения.

   Тогда критическая область переписывается в виде

   $$\sum_{X_i>0} R_i \geqslant k$$

   Доказывается аналогично критерию Вилкоксона для тестирования гипотезы случайности.

2. ***Критерий знаков***

   Пусть дана выборка $X_1, \ldots, X_n \sim f$, где $f(x) = \frac{1}{2}e^{-|x|}$.

   Поймем, как устроены метки:

   $$\frac{f'}{f}(x) = \mathrm{sign}(x) \implies a_n^+(R_i, f) = 1$$

   А тогда критическая область записывается как

   $$\sum_{X_i>0} 1 = \#\{X_i > 0\} \geqslant k$$

6.8. Ранговые критерии. Гипотеза независимости

Опишем постановку задачи в случае гипотезы независимости.

Пусть нам дана совместная выборка $(X_1, Y_1), \ldots, (X_n, Y_n)$, где $X_i \sim f$, $Y_i \sim g$.

Положим $\mathbf{R}$ и $\mathbf{Q}$ — ранги $\mathbf{X}$ и $\mathbf{Y}$ соответственно.

И будем проверять следующие гипотезы:

$$H_0 :\ q_0(\mathbf{X}, \mathbf{Y}) = \prod_{i=1}^{n} f(X_i)g(Y_i)$$

$$H_1 :\ q_\delta(\mathbf{X}, \mathbf{Y}) = \prod_{i=1}^{n}\left(\int_{-\infty}^{\infty} f(x - \delta z)g(y - \delta z)\, d\,H(z)\right)$$

Здесь $H(z)$ — функция распределения некоторой случайной величины Z. И по сути, альтернативная гипотеза утверждает, что элементы выборки получены с помощью независимых случайных величин X^* и Y^* как

$$\begin{cases} X_i = X_i^* + \delta Z_i \\ Y_i = Y_i^* + \delta Z_i \end{cases}$$

И для такого случая также известен вид локально наиболее мощного критерия.

Теорема 6.16.

Локально наиболее мощный ранговый критерий для тестирования гипотезы независимости — критерий с критической областью следующего вида:

$$\sum_{i=1}^{n} a_n(R_i, f) \cdot a_n(Q_i, g) \geqslant k$$

Доказательство.

Без доказательства. □

Рассмотрим теперь конкретные критерии для тестирования гипотезы независимости:

1. ***Коэффициент корреляции Спирмена***

   Пусть f и g — плотности логистического распределения.

   Тогда критическая область представима как

   $$\sum_{i=1}^{n} R_i Q_i \geqslant k$$

   Доказывается это утверждение так же, как и раньше. Возникает лишь вопрос, при чем здесь коэффициент корреляции.

   На самом деле, коэффициент корреляции Спирмена задается следующим образом:

   $$\rho = \frac{12}{n^3} \sum_{i=1}^{n} (R_i - \mathrm{E}\, R_i)(Q_i - \mathrm{E}\, Q_i) = \frac{12}{n^3} \sum_{i=1}^{n} \left(R_i - \frac{n+1}{2}\right)\left(Q_i - \frac{n+1}{2}\right)$$

   Этот коэффициент действительно отражает то, насколько ранги коррелируют друг с другом. В частности, несложно убедиться, что он принимает значения из $[-1, 1]$.

   А если раскрыть скобки, то можно увидеть, что приведенная критическая область эквивалентна критической области $\rho \geqslant \tilde{k}$ для некоторого $\tilde{k}$.

7. Моделирование распределений

Напоследок обсудим следующую тему, связанную с моделированием распределений.

Разумеется, сейчас существует большое количество различных библиотек, умеющих моделировать многие стандартные распределения. Тем не менее, какие-то более хитрые распределения приходится реализовывать самостоятельно, и именно об этом будет рассказываться в этой главе.

Далее все предлагаемые методы будут основываться на предположении, что мы некоторым образом умеем генерировать выборки из стандартного равномерного распределения.

7.1. Метод обратной функции

Обозначение.

Пусть F — функция распределения желаемой случайной величины X.

Определим функцию $\mathrm{G} : (0,1) \to \mathbb{R}$ следующим образом:

$$\mathrm{G}(y) := \inf_x \{x : \mathrm{F}(x) > y\}$$

Лемма.

$\mathrm{G}(y) < x \iff y < \mathrm{F}(x)$

Доказательство.

Обозначим $A_y = \{x \in \mathbb{R} : \mathrm{F}(x) > y\}$ и положим $z = \mathrm{G}(y)$.

Заметим, что т.к. F монотонно неубывает, то любое значение, большее чем z, лежит в множестве. Само же z может как лежать в нем, так и не лежать. То есть, $A_y = \langle z, +\infty)$.

Кроме того, $\mathrm{F}(z) \leqslant y$, т.к. иначе по непрерывности слева нашлась бы такая точка $z^* < z$, что $\mathrm{F}(z^*) > y$. А это противоречит определению $z = \mathrm{G}(y)$.

Таким образом мы получили, что на самом деле $A_y = (z, +\infty)$. А поскольку при заданном y каждое из условий $z = G(y) < x$ и $y < F(x)$ задает одно и то же множество, то эти условия равносильны. □

Лемма.

$\mathrm{G}(\mathrm{U}[0,1]) \sim X$, где X — моделируемая случайная величина

Доказательство.

Пусть $\alpha \sim \mathrm{U}[0,1]$.

Тогда $\mathrm{P}(\mathrm{G}(\alpha) < t) = \mathrm{P}(\alpha < \mathrm{F}(t)) = \mathrm{F}(t)$. □

Две приведенные выше леммы показывают, каким образом моделировать произвольные случайные величины в случае, когда мы можем посчитать функцию G. По сути, мы таким образом просто свели генерацию случайной выборки из заданного распределения к генерации выборки из стандартного равномерного распределения.

Кроме того, в некоторых случаях G вычисляется довольно просто. А именно, если F обратима и мы можем легко вычислить обратную функцию.

Рассмотрим несколько примеров, чтобы лучше разобраться в этом.

Примеры.

1. Пусть $X \sim \mathrm{Exp}(\lambda)$, $\alpha \sim \mathrm{U}[0,1]$.

   Мы знаем, что функция распределения нашей случайной величины имеет вид

   $$\mathrm{F}(x) = 1 - e^{-\lambda x},\ x > 0$$

   Легко проверить, что обратная функция будет, соответственно, иметь вид

   $$\mathrm{G}(y) = -\frac{\ln(1-y)}{\lambda}$$

   А тогда можем моделировать нашу случайную величину по правилу

   $$-\frac{\ln(1-\alpha)}{\lambda} \sim \mathrm{Exp}(\lambda)$$

2. Пусть $X \sim \mathrm{Cauchy}$.

   Мы знаем, что функция распределения нашей случайной величины имеет вид

   $$\mathrm{F}(x) = \frac{1}{2} + \frac{1}{\pi}\arctan x$$

   Для этого распределения обратная функция тоже считается довольно просто:

   $$\mathrm{G}(y) = \tan\left(\pi\left(y - \frac{1}{2}\right)\right)$$

   Само же распределение моделируется все тем же способом на основе значений $\alpha \sim \mathrm{U}[0,1]$.

3. Пусть $X \sim \mathcal{N}(0,1)$.

   Поскольку функция распределения Φ не имеет хорошего представления (вообще говоря, является некоторым интегралом), то с ходу и непонятно, каким образом искать обратную функцию $\mathrm{G} = \Phi^{-1}$.

   Потому воспользуемся трюком и научимся генерировать сразу пару значений из нормального распределения. А именно, научимся получать стандартный гауссовский вектор (ξ, η), где $\xi, \eta \sim \mathcal{N}(0,1)$ и независимы.

   Для этого перейдем к полярным координатам: $(\xi, \eta) \to (\rho, \varphi)$, где $\xi = \rho\cos\varphi$, $\eta = \rho\sin\varphi$.

   Тогда если в старых координатах плотность гауссовского вектора имела вид

   $$f_{(\xi,\eta)}(x,y) = \frac{1}{2\pi}e^{-\frac{1}{2}(x^2+y^2)},$$

   то в новых координатах плотность будет иметь вид

   $$f_{(\rho,\varphi)}(r,\alpha) = \frac{1}{2\pi}e^{-\frac{1}{2}r^2}\cdot r.$$

   Заметим, что если распределения φ и ρ у нас независимы, то плотность их совместного распределения попросту равна произведению плотностей этих распределений.

   Тогда скажем, что $f_\varphi(\alpha) = \dfrac{1}{2\pi}$, $f_\rho(r) = re^{-\frac{1}{2}r^2}$.

   Очевидно, что тогда в качестве φ мы можем просто выбрать $\mathrm{U}[0, 2\pi]$. Осталось разобраться, как получить распределение ρ по его плотности. Для этого взглянем на его функцию распределения:

   $$\mathrm{F}_\rho(r) = 1 - e^{-\frac{1}{2}r^2} \implies G(y) = \sqrt{-2\ln(1-y)} \overset{\alpha\sim\mathrm{U}[0,1]}{\Longrightarrow} G(\alpha) = \sqrt{-2\ln(1-\alpha)} \overset{\zeta\sim\mathrm{Exp}(1)}{\sim} \sqrt{2\zeta}$$

   Таким образом, мы научились генерировать сразу пару значений из $\mathcal{N}(0,1)$ как

   $$\xi = \sqrt{2\zeta}\cdot\cos\varphi,\ \eta = \sqrt{2\zeta}\cdot\sin\varphi,\ \text{где } \zeta \sim \mathrm{Exp}(1),\ \varphi \sim \mathrm{U}[0, 2\pi]$$

7.2. Метод отбора

Предположим, что плотность $p(x)$ моделируемой случайной величины η существует на некотором отрезке $[a, b]$. И пусть при этом плотность ограничена. То есть:

$$p(x) = 0 \text{ при } x \notin [a, b], \quad p(x) \leqslant M$$

Приведем алгоритм моделирования распределении для такой ситуации.

Алгоритм.

Представим себе график плотности. В силу описанных условий, его ненулевая часть полностью помещается в прямоугольник $[a, b] \times [0, M]$.

TODO: picture

Выберем внутри этого прямоугольника случайную точку $(\xi_1, \xi_2) \sim \mathrm{U}\big([a, b] \times [0, M]\big)$.

Если $p(\xi_1) > \xi_2$, т.е. точка попала в подграфик p, то положим $\eta = \xi_1$.

Иначе повторим предыдущий шаг.

TODO: one more picture

Поймем, почему описанный алгоритм действительно моделирует наше случайное распределение, а также насколько хорошо он работает. Начнем с последнего.

Утверждение 7.1.

Матожидание количества действий, необходимых для генерации одного элемента из распределения $p(x)$ с помощью нашего алгоритма, равно $M(b - a)$.

Доказательство.

Действительно. Так как мы выбираем случайную точку в прямоугольнике размера $M(b-a)$, в то время как площадь подграфика p равна 1, нам в среднем потребуется сгенерировать как раз $M(b - a)$ точек, прежде чем мы попадем в подграфик.

Соответственно, имеет смысл применять алгоритм в таком виде только тогда, когда $M(b-a)$ не слишком велико. □

Утверждение 7.2.

Случайная величина η, моделируемая с помощью описанного алгоритма, распределена с плотностью p.

Доказательство.

Заметим, что генерируя точки равномерно на $[a, b] \times [0, M]$, мы в том числе получаем равномерное распределение на подграфике $\mathcal{P}_p$ нашей плотности.

А тогда имеем:

$$\mathrm{P}(\eta < t) = \iint\limits_{\substack{(x,y)\in\mathcal{P}_p \\ x<t}} 1\,dx\,dy = \int\limits_a^t \left(\int\limits_0^{p(x)} 1\,dy \right) dx = \int\limits_a^t p(x)\,dx$$

□

С одной стороны, мы привели метод, который позволяет довольно просто генерировать многие распределения. Тем не менее, его главными минусами являются допущения, описанные в

начале. В частности, такой способ уже не подходит для распределений, плотность которых не ограничена.

Тем не менее, этой проблемы можно попробовать избежать, если считать нашу плотность не относительно меры Лебега, а относительно какой-нибудь другой меры.

А именно, предположим, что нам даны распределения $\mathcal{P}$ и $\mathcal{Q}$ с соответствующими плотностями $p(x)$ и $q(x)$, причем $\mathcal{P}$ абсолютно непрерывно относительно $\mathcal{Q}$.

Плотность распределения $\mathcal{P}$ относительно $\mathcal{Q}$ определяется как $\dfrac{d\mathcal{P}}{d\mathcal{Q}} = \dfrac{p}{q} = r$.

Тогда если $r(x) \leqslant M$ на $[a, b]$, то можем воспользоваться похожим алгоритмом для моделирования распределения p.

Алгоритм.

Выберем случайную точку (ξ_1, ξ_2), т.ч. $\xi_1 \sim q(x)$, $\xi_2 \sim \mathrm{U}[0, M]$.

Если $r(\xi_1) > \xi_2$, т.е. точка попала в подграфик r, то положим $\eta = \xi_1$.

Иначе повторим предыдущий шаг.

7.3. Метод декомпозиции

Рассмотрим еще один метод генерации данных, в основе которого лежит представление нашей случайной величины в виде линейной комбинации других случайных величин, каждую из которых мы умеем моделировать.

Лемма.

Пусть функция распределения исследуемой случайной величины X имеет вид

$$\mathrm{F}(x) = \sum_{i=1}^{n} p_i \, \mathrm{F}_i(x), \text{ где } p_i \geqslant 0,\ p_1 + \ldots + p_n = 1$$

Здесь F_i — функции распределения, соответствующие некоторым независимым случайным величинам ξ_i.

Кроме того, пусть нам дана независимая случайная величина τ, принимающая значение k с вероятностью p_k.

Тогда $X \sim \xi_\tau$.

Доказательство.

$$\mathrm{P}(X < t) = \sum_{i=1}^{n} \mathrm{P}(\tau = i) \cdot \mathrm{P}(\xi_i < t \,|\, \tau = i) = \sum_{i=1}^{n} \mathrm{P}(\tau = i) \cdot \mathrm{P}(\xi_i < t) = \sum_{i=1}^{n} p_i \, \mathrm{P}(\xi_i < t)$$

□