6. Робастные оценки. Медиана, усеченное среднее,

Как мы успели убедиться за прошлые занятия, зачастую очень удобной оценкой математического ожидания оказывается $\overline{X}$. Это несмещенная, состоятельная, а при наличии у распределения дисперсии, еще и асимптотически нормальная оценка. Однако, $\overline{X}$ обладает некоторыми недостатками, которые заметно осложняют ее использование.

Пример 1. Предположим, что в точке $(\theta, 1)$ находится источник излучения, который испускает в сторону оси OX γ-частицы под случайными углами, равномерно распределенным на $(-\pi/2, \pi/2)$. Мы регистрируем координаты X_i точек касания оси частицами и хотим оценить с их помощью параметр θ. Давайте найдем распределение X_i.

$$P(X_i \le x) = P(X_i - \theta \le x - \theta) = P(\tan(\phi) \le x - \theta) = \frac{\pi/2 + \arctan(x-\theta)}{\pi}.$$

Его плотность

$$f_X(x) = (P(X \le x))' = \frac{1}{\pi(1 + (x-\theta)^2)}.$$

Таким образом, величины $X_i - \theta$ имеют распределение Коши. Его характеристическая функция имеет вид $e^{-|t|}$, откуда характеристическая функция $\overline{X}$ примет вид

$$\psi_{\overline{X}}(t) = Ee^{i(X_1+\ldots+X_n)t/n} = (Ee^{iX_1 t/n})^n = \psi^n_{X_1}(t/n) = e^{-n|t|/n + i\theta t} = e^{i\theta t - |t|} = \psi_{X_1}(t).$$

Следовательно, $\overline{X}$ ничуть не лучше оценивает θ, чем каждый из X_i. У распределения Коши "тяжелые хвосты", среди X_i появляются "большие" значения, которые искажают $\overline{X}$.

Для того, чтобы справиться с этой проблемой, нам нужны оценки, которые меньше подвержены искажению в результате появления больших значений в выборке. Одной из таких оценок является выборочная медиана. Напомню,

$$MED = \begin{cases} X_{(k)}, & n = 2k-1, \\ \frac{X_{(k)} + X_{(k+1)}}{2}, & n = 2k, \end{cases}$$

где $X_{(1)}, \ldots, X_{(n)}$ - вариационный ряд, то есть упорядоченные по возрастанию X_i. Медиана является оценкой теоретической медианы, то решения уравнения $med = x_{1/2}$, такой что $F(x_{1/2}) = 1/2$. Если таких решений несколько (т.е. отрезок, то будем полагать med серединой этого отрезка). Если их нет совсем, то выберем $med = F^{-1}(1/2) = \inf\{x : F(x) \ge 1/2\}$.

В последних двух случаях теоретическая медиана зачастую ведет себя не очень удачно:

Пример 2. Рассмотрим бернуллиевскую случайную величину с параметром 1/2. Тогда медиана окажется крайне неудачной оценкой чего бы то ни было, поскольку всегда равна 0, 0.5 или 1, причем по ЦПТ мы будем периодически менять значение с 0 на 1, проходя по пути через 1/2. Никакой стабилизации с ростом размера выборки мы не увидим.

Для величин с равномерным распределением на объединении отрезков $[-2, -1]$ и $[1, 2]$ при нечетном n медиана будет лежать или в отрезке $[-2, -1]$ или в $[1, 2]$, причем опять же постоянно перескакивать из одного отрезка в другой.

Поэтому ограничимся распределениями, у которых существует единственное решение этого уравения. Предположим, что наше распределение абсолютно непрерывно, а плотность непрерывна в окрестности $x_{1/2}$ и положительна в самой $x_{1/2}$. При таких условиях выборочная медиана оказывается состоятельной асимптотически нормальной оценкой теоретической:

Теорема 1.

$$\sqrt{n}(MED - x_{1/2}) \xrightarrow{d} Z \sim \mathcal{N}\left(0, \frac{1}{4f_X^2(x_{1/2})}\right).$$

Заметим, что для симметричных около θ распределений, $x_{1/2} = \theta$ в силу симметрии. Поэтому, например, для нормального $\mathcal{N}(\theta, \sigma^2)$, $R[\theta - a, \theta + a]$ или распределения Коши из примера 1 медиана будет асимптотически нормальной оценкой θ.

Распределение MED можно отыскать, используя формулу

$$f_{X_{(k)}}(x) = nC_{n-1}^{k-1}F(x)^{k-1}(1-F(x))^{n-k}p(x)$$

для плотности k-ой порядковой статистики $X_{(k)}$. Эта формула достаточно естественна - мы должны выбрать один номер для k-ой по порядку величины и $k-1$ для тех, которые меньше.

Для симметричных относительно θ $\overline{X}$ и MED оценивают одну и ту же величину. В этом случае, осмысленно сравнивать их с помощью асимптотической эффективности.

Пример 3. Для $\mathcal{N}(\theta, 1)$ асимптотическая дисперсия $\overline{X}$ есть 1, дисперсия медианы — $(4/(\sqrt{2\pi})^2)^{-1} = \pi/2$. Следовательно, $e_{MED,\overline{X}} = 2/\pi \approx 0.64$, т.е. выборочное среднее примерно в полтора раза эффективнее медианы.

Для равномерного $R[\theta-a, \theta+a]$ аналогичным образом можно убедиться, что среднее втрое эффективнее медианы.

Но для распределения Лапласа $f_\theta(x) = e^{-|x-\theta|}$ преимущество на стороне медианы, которая вдвое эффективнее.

Для распределения Коши, как мы выяснили, оценка $\overline{X}$ не асимптотически нормальна, а вот медиана работает вполне прилично.

Снижение эффективности на моделях "с легкими хвостами" по сравнению с $\overline{X}$ связано с тем, что медиана отбрасывает большую долю часть выборки. Давайте попробуем найти некоторый баланс между $\overline{X}$ и MED. В связи с этим рассматривают так называемое усеченное среднее порядка $0 \le \alpha \le 1/2$

$$\overline{X}_\alpha = \frac{X_{([n\alpha]+1)} + ... + X_{(n-[n\alpha])}}{n - 2[\alpha n])}$$

При $\alpha = 0$ это $\overline{X}$, при $\alpha = 1/2$ оценка не всегда определена, но ее доопределяют равной MED.

Рассмотрим класс $Symm_\theta$ распределений, имеющих непрерывную плотность, симметричную около некоторой точки θ, причем носитель этих распределений представляет собой конечный или бесконечный интервал $(\theta - c, \theta + c)$. Тогда если F_θ лежит в $Symm_\theta$, то оценка $\overline{X}_\alpha$ будет асимптотически нормальной оценкой θ:

Теорема 2.

$$\sqrt{n}(\overline{X}_\alpha - \theta) \xrightarrow{d} Z \sim \mathcal{N}(0, \sigma_\alpha^2),$$

$\sigma_\alpha^2 = \frac{2}{(1-2\alpha)^2}\int_\theta^{y_{1-\alpha}}(t-\theta)^2 f_\theta(t)dt + \alpha(y_{1-\alpha} - \theta)^2$, y_α - так называемая α-квантиль распределения F_θ, т.е. решение уравнения $F_\theta(y_\alpha) = \alpha$. В нашем случае функция распределения непрерывна и строго монотонна на $(-c, c)$, а значит такое y_α единственно.

Замечание. Если мы рассматриваем модель $F_\theta() = F(x - \theta)$, где F — заданное распределение из $Symm_0$ (т.е. модель сдвига фиксированного распределения), то σ_α^2 и $y_\alpha - \theta$ не будут зависеть от θ и наша дисперсия примет более удобную форму

$$\sigma_\alpha^2 = \frac{2}{(1-2\alpha)^2}\int_0^{y_{1-\alpha}} t^2 f(t)dt + \alpha y_{1-\alpha}^2,$$

где y — квантили распределения F, а f — его плотность.

Пример 4 Для нормального распределения с параметрами $\mathcal{N}(\theta, 1)$ $e_{\overline{X}_\alpha,\overline{X}}$ приближенно выражается следующим образом:

α	0	1/20	1/8	1/4	3/8	1/2
$e_{\overline{X}_\alpha,\overline{X}}$	1,00	0,99	0,94	0,84	0,74	0,64

Опять же на нормальном распределении $\overline{X}$ более выгодно, однако, скажем при $\alpha = 1/20$ разница

между ними незначительна. При этом $\overline{X}_{1/20}$ допускает 1/20 выбросов, которые никак не повлияют на нашу оценку, т.е. значительно более устойчива к выбросам.

Зачастую на практике наша выборка зашумлена, порядка 1-10 процентов выборки может быть ошибками. Это связано со сбоями оборудования, ошибками эксперимента и человеческим фактором — опечатками и прочим. Удалить эти "выбросы" не всегда представляется возможным. Одна из проблем выборочного среднего в том, что если у нас, скажем, 99 данных от 0 до 1, а одно данное 300, то выборочное среднее окажется не меньше 3. В связи с этим важным качеством оценок является так называемая *робастность* — устойчивость к искажению данных. Мы ограничимся рассмотрением вопросов, связанных с тем, насколько можно исказить выборку, чтобы оценка ушла в бесконечность.

Для этого нам понадобится понятие *обобщенной функции распределения*. Рассмотрим функцию F — монотонную и непрерывную справа, принимающую значения из $[0,1]$, но не обязательно стремящуююся к 0 на $-\infty$ и к 1 на ∞. В таком случае естественно считать, что это ф.р. случайной величины, принимающей значения $-\infty$, ∞ с вероятностями $F(-\infty)$, $1-F(\infty)$. В частности, $\int_{-\infty}^{\infty} x dF(x)$ будем считать с учетом бесконечных значений.

Множество всех обобщенных ф.р. назовем $\mathcal{M}$. Введем на них метрику Леви — расстоянием Леви $d(F,G)$ между обобщенными функциями распределения F и G назовем $\inf\{\varepsilon : F(x-\varepsilon)-\varepsilon < G(x) < F(x+\varepsilon)+\varepsilon\}$. ε-окрестностью $F_0 \in \mathcal{M}$ назовем $U_\varepsilon = \{F \in \mathcal{M} : d(F,F_0) < \varepsilon\}$.

Пример 5. Для функций распределения двух бернуллиевских величин с параметрами p_1, p_2 расстояние Леви совпадает с равномерным и равно $|p_1-p_2|$. Подсчитаем его для ф.р. F, F_c двух равномерных распределений $R[a,b]$, $R[a+c,b+c]$, $c>0$.
Докажем, что расстояние Леви есть $\rho = \min(1, c/(b-a+1))$.
Пусть оно $\varepsilon < \rho$. Рассмотрим точку $x = a+c-\varepsilon$, $x > a$. В силу определения расстояния

$$F(x) \le F_c(x+\varepsilon)+\varepsilon = F(a+c)+\varepsilon = \varepsilon.$$

Т.к. $F(x) < \varepsilon < 1$, то $a < x < a+b$, откуда

$$F(x) = \frac{x-a}{b-a} = \frac{c-\varepsilon}{b-a} \le \varepsilon.$$

Решая последнее неравенство, получаем противоречие. Значит расстояние не меньше указанной величины
Остается доказать, что для всех x выполнено

$$F_c(x-\rho)-\rho \le F(x) \le F_c(x+\rho)+\rho.$$

Левое неравенство очевидно из того, что $F_c(x) \le F(x)$. Давайте докажем правое при $x=b$.
Если $c \ge b-a+1$, то $\rho = 1$ и неравенство очевидно. Пусть $c < b-a+1$. Тогда $b+c > b+c/(b-a+1) > a+c$, откуда

$$F_c(b+\rho)+\rho = \frac{b+\frac{c}{b-a+1}-a-c}{b-a} + \frac{c}{b-a+1} = 1 = F(b).$$

Но тогда то же неравенство выполнено при всех x. Действительно, при $x < a$ $F(x) = 0$ и неравенство тривиально, при $x > b$

$$F(x) = F(b) \le F_c(b+\rho)+\rho \le F_c(x+\rho)+\rho.$$

Если же $x \in [a,b]$, то

$$F(x) = F(b) - \frac{b-x}{b-a} \le F_c(b+\rho)+\rho-\frac{b-x}{b-a} \le F_c(x+\rho)+\rho,$$

поскольку $F_c(x+y) \le F_c(x)+y/(b-a)$.

Представим оценку $\hat{\theta}$ в виде отображения $f(\hat{F})$ от ЭФР. Так, например, в задаче 1.1.3 показывалось,

что

$$\overline{X} = \int_{\mathbb{R}} x d\hat{F}(x),\ S^2 = \int_{\mathbb{R}} x^2 d\hat{F}(x) - \overline{X}^2.$$

В целом, эмпирическая функция распределения при фиксированном размере выборки определяет вариаицонный ряд выборки $X_{(1)}, ..., X_{(n)}$, а значит и все функции от него.

Будем смотреть на то, насколько $\hat{\theta}$ изменяется при изменении F. Рассмотрим ее максимальное изменении в окрестности ф.р. F_0

$$b(\varepsilon) = \sup\{|f(F) - f(F_0)|, F \in U_\varepsilon\}.$$

Тогда $b(1)$ — максимальное возможное значение $f(F) - f(F_0)$ на всем $\mathcal{M}$. Назовем асимптотическим пороговым значением f

$$\tau(\hat{\theta}) = \sup \varepsilon : b(\varepsilon) < b(1).$$

Давайте разбираться с тем, что это же такое:
Пример 6. Рассмотрим оценку $\overline{X} = \int_{-\infty}^{\infty} x d\hat{F}(x)$. Для нее $b(\varepsilon) = \infty$ при любой F_0, поскольку функция

$$F = \varepsilon I_{x<F_0^{-1}(\varepsilon)} + F_0(x) I_{x \geq \varepsilon}$$

лежит в $U_\varepsilon(F_0)$, а функционал $\overline{X}$ на ней бесконечен. Таким образом, асимптотическое пороговое значение у этой статистики равно 0.

Пример 7 Оценку MED можно представить в виде $(\hat{F}^{-1}(1/2)+\hat{F}^{-1}(1/2+0))/2$ в силу 6.2.1. Покажем, что $\tau(\hat{\theta}) = 1/2$ при любой F_0.
Действительно, $b(1) = \infty$, поскольку для $F_1 = 1/2 - \delta$ $f(F_1) = \infty$ при любом $\delta > 0$. При этом F_1 удалена от любой функции из $\mathcal{M}$ не более чем на $1/2 + \delta$, т.е. лежит в $U_{1/2+\delta}$.
Значит $\tau(MED) \leq 1/2$.
Когда же $f(F) = \pm\infty$? Когда $F(-\infty) \geq 1/2$ или $F(\infty) \geq 1/2$. Однако, в U_ε $F(-\infty) \leq \varepsilon$, $F(\infty) \geq 1 - \varepsilon$.
Значит $\tau(MED) \geq 1/2$. Следовательно, $\tau(MED) = 1/2$.

В какой-то мере асимпотическое пороговое значение отражает, какую часть выборки мне нужно испортить, чтобы изменить значение моей статистики сколь угодно далеко от исходного. Для $\overline{X}$ достаточно испортить даже одно наблюдение, т.е. 0%. Для MED потребуется не меньше половины выборки.

Для нее нетрудно найти представления через ЭФР (6.1.1):

$$\overline{X}_\alpha = (1 - 2\alpha)^{-1} \int_\alpha^{1-\alpha} \hat{F}^{-1}(x) dx,$$

Строго говоря, это выражение верно только при целых $n\alpha$, при нецелых оно немного отличается. Тем не менее при больших n разница между этими статистиками близка к нулю. Для $\overline{X}_\alpha$ можно показать (6.3.1), что пороговым значением будет α, как и следовало ожидать.

А как у нее с эффективностью в сравнении с $\overline{X}$? Об этом говорит теорема 3:
Теорема 3. Для распределений, удовлетворяющих теореме 2, справедливо неравенство

$$e_{\overline{X}_\alpha, \overline{X}} \geq (1 - 2\alpha)^2,$$

а для распределений, которые помимо всего прочего имеют единственный максимум (из симметричности он расположен в точке θ), справедливо неравенство

$$e_{\overline{X}_\alpha, \overline{X}} \geq 1/(1 + 4\alpha).$$

Нижняя граница при этом, к сожалению, достижима.

Таким образом, при небольших α усеченное среднее практически ничего не теряет, но приобретает заметную устойчивость, например, при работе с распределениями типа Коши. Но если я хочу пороговое

значение, скажем, 1/4, то на распределениях с единственным максимумом я потеряют 16 процентов, а в худшем случае из теоремы 3 и вовсе ухудшу дисперсию в 4 раза. Нет ли более эффективного решения при необходимости большого порогового значения? Оказывается, такая оценка существует.

Рассмотрим оценку $W = MED((X_i + X_j)/2)$. Такую оценку называют оценкой Ходжеса-Лемана, но поскольку мы уже использовали этот термин, то будем звать ее медианой средних Уошла. Ее пороговое значение примерно 0.3 (см. задачу 5.2.2), при этом справедлива следующая теорема:

Теорема 4. Для распределения класса Symm справедливо соотношение

$$\sqrt{n}(W - \theta) \xrightarrow{d} Z \sim \mathcal{N}(0, \sigma^2),$$

где $\sigma^2 = (\int f_X^2(t)dt)^{-2}/12$. При этом $e_{W,\overline{X}} \geq 108/125$.

Зачастую рассматривают оценку W, выбирая медиану не по всем i, j, а только по $i \leq j$ или $i < j$, чтобы не рассматривать повторяющиеся пары. Асимптотически все три оценки одинаково себя ведут.

Таким образом, оценка медиана средних Уолша достаточно толерантна, имеет очень хорошую относительную эффективность (около 0.86) относительно $\overline{X}$ на любом распределении (а для нормального распределения 0.955). Стоит отметить, что и она не идеальна — при несимметричном распределении у усеченного среднего не сильно изменяется поведение дисперсии, а вот W потеряет все свои положительные свойства.

Под $F * F$ подразумевается $\int_{\mathbb{R}} F(t - x)dF(x)$ — ф.р. суммы независимых $X, Y \sim F$.

6.1.1 Доказать, что $\overline{X}_\alpha = (1 - 2\alpha)^{-1} \int_\alpha^{1-\alpha} \hat{F}^{-1}(x)dx$ при целых $n\alpha$.

6.2.1 Доказать, что $MED = (\hat{F}^{-1}(1/2) + \hat{F}^{-1}(1/2 + 0))/2$.

6.3.1 Доказать, что $\tau(\overline{X}_\alpha) = \alpha$.

6.1.2 Доказать, что $\tau(W) \leq 1 - 1/\sqrt{2}$.

6.2.2 Доказать, что $W = med(\hat{F} * \hat{F})/2$, где $\hat{F}$ — э.ф.р.

6.3.2 Доказать, что $\tau(W) \geq 1 - 1/\sqrt{2}$.

6.1.3 Найти асимптотическую дисперсию для $\overline{X}_{1/4}$, $\overline{X}$ и W для выборки, имеющей распределение Лапласа ($f_X(x) = e^{-|x-\theta|}/2$). Моделировать выборку размера 20 для некоторого θ и сравнить точность наших оценок.

6.2.3 Моделировать выборку размера 20 с равномерным распределением на $[\theta - 1/2, \theta + 1/2]$, где θ — выбранный вами параметр. Сравнить, какая из оценок $\hat{\theta} = (\max(X_1, ..., X_n) + \min(X_1, ..., X_n))/2$ и W оказалась точнее. Доказать, что $D\hat{\theta}$ быстрее DW стремится к 0 с ростом n.

6.3.3 Для X_i, имеющих ф.р. $F(x, \theta, 1) = (1 + e^{-(x-\theta)})^{-1}$, найти асимптотическую дисперсию W. Выбрать произвольное θ и посмотреть, насколько хорошо по выборке размера 20 оценка W оценит θ.