Семинар №14.
Робастные статистики.

В реальных данных доля "выбросов" (выделяющихся значений) составляет от 1% до 10% процентов. Это происходит от большого количества неучтенных факторов (в медицине, психологии и пр.), сбоев оборудования (типа скачков напряжения в электросети – в экспериментальной физике), даже в астрономических таблицах встречается до 0,1% ошибок. Поэтому надо как-то реагировать на эти ошибки, когда мы работаем со статистическими данными. Например, отсекать далеко отстоящие наблюдения, переходить от значений членов выборки к их рангам и прочим методам, которые помогают устранить ошибки. Однако, надо следить и за тем, чтобы эффективность получаемых робастных (т.е. устойчивых к выбросам) оценок по сравнению с обычными оценками (т.е. отношение их асимптотических дисперсий) не была слишком низкой.

Одной из простых мер робастности является асимптотическая толерантность – это наибольшая доля выбросов в выборке, которую "выдерживает" статистика, не смещаясь вслед за выбросами на $+\infty$ или $-\infty$. Дадим формальное определение толерантности:
Определение. Пусть для оценки $\widehat{\theta}(x_{(1)}, \ldots, x_{(n)})$ (где $x_{(1)} \le \ldots \le x_{(n)}$ – вариационный ряд числовой последовательности $\{x_i\}_{i=1}^n$) найдётся целое число k, $0 \le k < n$ такое, что
1) если $x_{(k+2)}, \ldots, x_{(n)}$ – фиксированы, а $x_{(k+1)} \to -\infty$, то $\widehat{\theta}(x_{(1)}, \ldots, x_{(n)}) \to -\infty$;
2) если $x_{(1)}, \ldots, x_{(n-k-1)}$ – фиксированы, а $x_{(n-k)} \to +\infty$, то $\widehat{\theta}(x_{(1)}, \ldots, x_{(n)}) \to +\infty$.
Обозначим через k_n^* наименьшее такое k. Асимптотической толерантностью оценки $\widehat{\theta}$ называется предел $\tau_{\widehat{\theta}} = \lim n \to \infty \frac{k_n^*}{n}$, если этот предел существует.

Очевидно, что толерантность выборочного среднего $\tau_{\overline{X}} = 0$, толерантность выборочной медианы $\tau_{\widehat{\mu}} = \frac{1}{2}$, а толерантность введенного ниже усеченного среднего $\tau_{\overline{X_\alpha}} = \alpha$.

L-оценки.

Пусть $X_{(1)} \le \ldots \le X_{(n)}$ – вариационный ряд выборки $\{X_i\}_{i=1}^n$.
Определение. L-оценка – линейная комбинация вида $\sum\limits_{i=1}^{n} \omega_i X_{(i)}$.
Пример. Усеченное среднее $\overline{X}_\alpha = \frac{1}{n-2k}(X_{(k+1)} + \ldots + X_{(n-k)})$, где $k = [\alpha n]$ и $0 < \alpha < 0,5$. Это несмещенная и состоятельная оценка параметра θ для симметричного относительно θ распределения.

Чтобы изучить асимптотическое поведение L-оценок, обозначим $\omega_{in} = \frac{1}{n}\lambda\left(\frac{i}{n+1}\right)$, где λ – некая функция, определенная на отрезке $[0,1]$. Определим также $\mu(F,\lambda) = \int\limits_0^1 \lambda(t)F^{-1}(t)dt$ и $\sigma^2(F,\lambda) = \int\limits_0^1 G^2(t)dt - \left(\int\limits_0^1 G(t)dt\right)^2$, где $F(t)$ – некая функция распределения с плотностью $p(t)$, а $G(t)$ – любая функция с $G'(t) = \frac{\lambda(t)}{p(F^{-1}(t))}$.

Теорема 1 *$\{X_i\}_{i=1}^n$ – н.о.р.сл.в. на интервале (a,b), где $-\infty \le a < b \le +\infty$ с функцией распределения F, для которой выполнены следующие свойства:*
1) $p(x) > 0 \ \ \forall x \in (a,b)$, где $p(x)$ – плотность функции распределения F.
2) $EX_1^2 < +\infty$.
Потребуем еще, что $\lambda(t)$ – непрерывна почти всюду и ограничена, $\int_0^1 \lambda(t)dt = 1$.
Тогда для L-оценки $L_n = \frac{1}{n}\sum\limits_{i=1}^{n} \lambda\left(\frac{i}{n+1}\right) X_{(i)}$ справедлива сходимость

$$\sqrt{n}(L_n - \mu(F,\lambda)) \xrightarrow{d} \xi \sim N(0, \sigma^2(F,\lambda)) \text{ при } n \to \infty$$

Следствие. Пусть элементы выборки распределены согласно функции распределения $F(x-\theta)$, где F – симметрична относительно 0, а λ – симметрична относительно $\frac{1}{2}$. Тогда $\mu(F,\lambda) = \theta$.

Зададимся вопросом: при каком λ при фиксированной функции распределения F достигается минимум дисперсии?

Определение. Оценка $\widehat{\theta}_n$ называется асимптотически эффективной оценкой параметра θ, если $\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, \frac{1}{i(\theta)})$, где $i(\theta)$ – информация Фишера распределения P_θ.

Теорема 2 *Пусть плотность $p(x)$ функции распределения $F(x)$ дважды дифференцируема почти всюду и $p(x) \to 0$ при $x \to \pm\infty$. Положим $\gamma(x) = -\frac{p'(x)}{p(x)}$. Тогда при $\lambda^*(t) = C\gamma'(F^{-1}(t))$ $\sigma^2(F, \lambda^*) = \frac{1}{i(\theta)}$, т.е. оценка L_n является асимптотически эффективной оценкой параметра $\mu(F, \lambda)$.*

M-оценки.

Усеченное среднее $\overline{X}_\alpha$ – своеобразный компромисс между выборочным средним $\overline{X}$ $(\alpha = 0)$ и выборочной медианой $\widehat{\mu}$ $(\alpha \to 0.5)$ (т.е. и робастна, и эффективность не сильно уменьшается). Но можно действовать иначе: заметим, что $\overline{X}$ и $\widehat{\mu}$ минимизируют соответственно $\sum(X_i - \theta)^2$ и $\sum|X_i - \theta|$. Хьюбер предложил строить оценки параметра θ с помощью минимизации "меры разброса" $\sum \rho_b(X_i - \theta)$, где $\rho_b(x) = \frac{x^2}{2} I\{|x| \le b\} + (b|x| - \frac{b^2}{2}) I\{|x| > b\}$ (такие оценки называются оценками Хьюбера). Обобщим этот метод в следующем определении.

Определение. M-оценка M_n параметра сдвига θ определяется для некоторой функции $\rho(x)$ как точка минимума по θ функции $\sum \rho(X_i - \theta)$.

Очевидно, что если $\rho(x)$ строго выпукла, то минимизирующее значение единственно. Примером M-оценок могут служить только что введенные нами оценки Хьюбера.

Кроме того, понятно, что если у функции $\rho(x)$ есть производная, то M_n есть одно из решений уравнения $\sum \rho'(X_i - \theta) = 0$.

Теорема 3 *Пусть $\{X_i\}_{i \ge 1}$ – н.о.р.сл.в. с ф.р. $F(x - \theta)$, где $p(x) = F'(x)$ – четная. Тогда при слабых предположениях относительно F и ρ' выполнено*

$$\sqrt{n}(M_n - \theta) \xrightarrow{d} \xi \sim N(0, \sigma^2(F, \rho')), \text{ при } n \to \infty,$$

где асимптотическая дисперсия равна $\sigma^2(F, \rho') = \frac{\int(\rho'(x))^2 p(x)dx}{(\int \rho''(x)p(x)dx)^2}$.

Замечание. Если $\rho(x) = -\ln p(x)$, то $\rho'(x) = \gamma(x) = -\frac{p'(x)}{p(x)}$. В этом случае минимизация суммы $\sum \rho(X_i - \theta)$ эквивалентно максимизации правдоподобия $\prod p(X_i - \theta)$, т.е. M-оценка параметра сдвига совпадает с оценкой максимального правдоподобия (ОМП) и поэтому является асимптотически эффективной.

R-оценки.

R-оценки ведут своё происхождение от ранговых критериев для проверки гипотез о значении параметра сдвига θ (см. статистику Манна-Уитни, критерий Уилкоксона, медиану Ходжеса-Лемана и прочие методы непараметрической статистики).

Пусть $X_{(1)} < \ldots < X_{(n)}$ – вариационный ряд выборки $\{X_i\}_{i=1}^n$, $d_1, \ldots, d_n$ – набор неотрицательных чисел. Рассмотрим $\frac{n(n+1)}{2}$ полусумм вида $\frac{X_{(j)} + X_{(k)}}{2}$ при $1 \le j \le k \le n$. Каждой такой полусумме припишем вес $\omega_{jk} = \frac{d_{n-(k-j)}}{\sum_{i=1}^n i d_i}$. Легко убедиться, что $\sum\limits_{j,k} \omega_{jk} = 1$.

Определение. Рассмотрим дискретное распределение, присваивающее вероятности ω_{jk} значениям $\frac{X_{(j)} + X_{(k)}}{2}$. Тогда R-оценка R_n есть медиана этого распределения.

Пример. 1) Если $d_1 = \ldots = d_{n-1} = 0$, $d_n = 1$, то $\sum i d_i = n$ и $\omega_{jk} = \frac{1}{n}$ при $j = k$, тогда дискретное распределение из определения R-оценки – равномерное на множестве значений $\{X_{(i)}\}$, $i \in \{1, \ldots, n\}$. Т.е. $R_n = med\{X_{(1)}, \ldots, X_{(n)}\}$ – выборочная медиана.

2) Если $d_1 = \ldots = d_n = 1$, то все величины $\frac{X_{(j)}+X_{(k)}}{2}$ имеют одинаковый вес $\frac{2}{n(n+1)}$. Тогда $R_n = med\left\{\frac{X_{(j)}+X_{(k)}}{2}, j,k \in \{1,\ldots,n\}\right\}$ – так называемая медиана средних Уолша (называемая также медианой Ходжеса-Лемана) W.

Изучим теперь асимптотические свойства R-оценок. Пусть функция $K(t)$ определена на отрезке $[0,1]$, не убывает на нём и $K(1-t) = -K(t)$. Положим $d_i := d_{in} = K\left(\frac{i+1}{2n+1}\right) - K\left(\frac{i}{2n+1}\right)$.

Теорема 4 *В модели* $X_i \sim F(x-\theta)$, *где* $F'(x) = p(x)$ – *четная, при некоторых условиях на* F *и* K *выполнено*

$$\sqrt{n}(R_n - \theta) \xrightarrow{d} N(0, \sigma^2(F,K)),$$

где $\sigma^2(F,K) = \int_0^1 K^2(t)dt / \left(\int_{\mathbb{R}} K'[F(x)]p^2(x)dx\right)^2$.

Асимптотическая эффективность R-оценки для заданной функции распределения F достигается при $K(t) = \gamma(F^{-1}(x))$, где $\gamma(x) = -\frac{p'(x)}{p(x)}$, как и ранее.