U-max статистики и их предельное поведение.

Симарова Е.Н., студентка 5 курса математико-механического факультата СПбГУ направления Фундаментальная математика и механика (математика), katerina.1.14@mail.ru

Аннотация

Недавно Лао и Майер (2008) рассмотрели U-max статистику, где вместо обычных сумм по подмножествам расматривается максимум ядра. Такая статистика часто появляются в стохастической геометрии. Примеры включают в себя наибольшее расстояние между случайными точками в шаре, максимальный диаметр случайного многоугольника, самый большой скалярное произведение в выборке точек и т. д. Их предельные распределения связаны с распределения экстремальных значений.

Но в статьях Лао и Майера, а также в более поздних работах на эту тематику, изучение предельного поведения производилось для конкретных значений ядер. В данной статье будет изучено предельное расределение для целого класса U-max статистик.

Введение

U-статистики

Рассмотрим $\xi_1, \ldots, \xi_n$ — независимые случайные величины, которые принимают значения в измеримом пространстве (X, A) и имеют одинаковое распределение P. Пусть $\mathscr{P} = \{P\}$ —это некоторый класс вероятностных распределений на пространстве (X, A). Рассмотрим $\Theta(P)$ — некоторый функционал на $\mathscr{P}$. Функционал $\Theta(P)$ называется регулярным, если он может быть записан в виде

$$\Theta(P) = \int_X \ldots \int_X h(x_1, \ldots, x_m) P(dx_1) ... P(dx_m),$$

где $h(x_1, \ldots, x_m)$ — некоторая вещественнозначная симметричная борелевская функция. Эту функцию называют ядром, а целое число $m \geq 1$ называют степенью ядра или (реже) функционала $\Theta(P)$.

Халмос and Хёфдинг в работах [2], [3], были первыми, кто стал расматривать класс несмещенных оценок $\Theta(P)$, получивших название U-статистики. Они определяются следующим образом. Рассмотрим функцию $h(x_1, \ldots, x_m)$ — ядро функционала $\Theta(P)$. Тогда U-статистика степени m определяется как

$$U_n = \binom{n}{m} \sum_J h(\xi_{i_1}, \ldots, \xi_{i_m}),$$

где $n \geq m$, а множество $J = \{(i_1, \ldots, i_m) : 1 \leq i_1 < \ldots < i_m \leq n\}$ — это множество упорядоченных m - элементных перестановок с множеством индексов из набора$\{1, \ldots, n\}$. Впервые такая статистика была рассмотрена в 1946 году. Оказывается, что ряд оценочных и тестовых статистик относятся к классу U-статистик. Этот факт довольно сильно повлиял на развитие дальнейшей теории.

U-max статистики, определения и история развития

U-max статистики можно рассматривать как предельный случай U-статистик. Опреляются они следующим образом:

$$H_n = \max_J h(\xi_{i_1}, \ldots, \xi_{i_m}).$$

В данном определении функция h и множество J определяются так же, как и для U-статистик. U-min статистики $H_n^{'}$ определяются аналогично. Заметим, что при замене знака ядра U-min статистика превращается в U-max статистику, поэтому эти понятия равнозначны. Приведем некоторые примеры U-max и U-min статистик.

1. Максимальное расстояние $\max\limits_{1 \leq i < j \leq n} ||\xi_i - \xi_j||$, где $\xi_1, \xi_2, \ldots, \xi_n$ — это независимо и равномерно распределенные точки на d-мерной единичной сфере $B^d, d \geq 2$.

2. Максимальное скалярное произведение $\max\limits_{1 \leq i < j \leq n} \langle \xi_i, \xi_j \rangle$, где $\xi_1, \xi_2, \ldots, \xi_n$ — это независимо и равномерно распределенные точки на d-мерной единичной сфере $B^d, d \geq 2$.

3. Максимальный периметр и площадь:

$$\max_{1 \leq i < j < l \leq n} peri(U_i, U_j, U_l) \quad \text{и} \quad \max_{1 \leq i < j < l \leq n} area(U_i, U_j, U_l)$$

среди всех вписанных треугольников, чьи вершины лежат на единичной окружности и берутся из множества $U_1, \ldots, U_n$. Точки $U_1, \ldots, U_n$. независимо и равномерно распределенны на единичной окружности S.

Лао и Майер первыми начали изучать U-max статистики, посвятив им несколько своих работ ([5], [4], [6]). Они доказали основную предельную теорему для U-max статистик. Для этого Лао и Майер использовали некоторую модификацию утверждения о сходимости Пуассона из монографии Барбура, Холста и Янсона([1]), написанной в 1992 году. Звучит основная предельная теорема следующим образом.

Теорема. *(***Теорема Лао-Майера)** *Пусть $\xi_1, \ldots, \xi_n$ — это случайные величины некоторого измеримого пространства (Θ, A), и пусть есть симмметричная борелевская функция $h : \Theta^m \to \mathbb{R}$. Обозначим через*

$$H_n = \max_J h(\xi_{i_1}, \ldots, \xi_{i_m})$$

и определим для каждого $z \in \mathbb{R}$ следующие функции:

$$p_{n,z} = \mathbb{P}\{h(\xi_1, \ldots, \xi_m) > z\},$$
$$\lambda_{n,z} = \binom{n}{m} p_{n,z},$$
$$\tau_{n,z}(r) = \frac{\mathbb{P}\{h(\xi_1, \ldots, \xi_m) > z, h(\xi_{1+m-r}, \xi_{2+m-r}, \ldots, \xi_{2m-r}) > z\}}{p_{n,z}}.$$

Тогда для всех $n \geq m$ и для всех $z \in \mathbb{R}$ верно неравенство

$$|\mathbb{P}(H_n \leq z) - e^{-\lambda_{n,z}}| \leq$$
$$\leq \left(1 - e^{-\lambda_{n,z}}\right) \cdot \left[p_{n,z} \left(\binom{n}{m} - \binom{n-m}{m} \right) + \sum_{r=1}^{m-1} \binom{m}{r} \binom{n-m}{m-r} \tau_{n,z}(r) \right].$$

При этом, при замене h на $-h$ получается соответствующее неравенство для минимума.

Сильверман и Браун в своей работе [7] предоставили условия, при которых общая теорема, использованная в [5], приводит к нетривиальному Вейбулловскому закону в пределе.

Теорема. *(***Теорема Сильвермана-Брауна)** *В условиях теоремы Лао-Майера, если для некоторой последовательности трансформаций $z_n : T \to \mathbb{R}, T \subset \mathbb{R}$ для каждого $t \in T$ выполнены условия:*

$$\lim_{n\to\infty} \lambda_{n,z_n(t)} = \lambda_t > 0, \tag{1}$$
$$\lim_{n\to\infty} n^{2m-1} p_{n,z_n(t)} \tau_{n,z_n(t)}(m-1) = 0, \tag{2}$$

тогда

$$\lim_{n\to\infty} \mathbb{P}\left(H_n \le z_n(t)\right) = e^{-\lambda_t}$$

для любого $t \in T$.

Лао и Майер придумали метод, использующий эту теорему, с помощью которого могут быть исследовано предельное поведение некоторых U-max статистик. Они использовали его для изучения предельного поведения конкретных U-max статистик, в частности, они изучили предельное поведение всех статистик, упомянутых в примерах. Например, они доказали такую предельную теорему.

Теорема (Периметр вписанного треугольника)**.** *Пусть* $U_1, U_2, \ldots, U_n$ *— независимо и равномерно распределенные точки на единичной окружности* S*. Обозначим через* $peri(U_i, U_j, U_l)$ *периметр треугольника с вершинами в точках* U_i, U_j, U_l*. Назовем*

$$H_n = \max_{1\le i<j<l\le n} peri(U_i, U_j, U_l).$$

Тогда для любого $t > 0$ *верно, что*

$$\lim_{n\to\infty} \mathbb{P}\{n^3(3\sqrt{3} - H_n) \le t\} = 1 - \exp\{-\frac{2t}{9\pi}\}.$$

Но Лао и Майер рассматривали функции h только с фиксированным числом переменных. В 2014 году Е. Королева и Ю. Никитин опубликовали статью [8], в которой были рассмотрены несколько многомерных U-max статистик с произвольной размерностью ядра и их предельное поведение. В частности, задача предельного поведения периметра вписанного треугольника обобщается до задачи предельного поведения вписанного m - угольника, где $m \ge 3$.

Теорема (Периметр вписанного m-угольника)**.** *Пусть точки* $U_1, U_2, \ldots, U_n$ *независимо и равномерно распределены на окружности* S*. Обозначим через*

$$P_{m,n} = \max_{1\le i_1<\ldots<i_m\le n} peri(U_{i_1}, \ldots, U_{i_m})$$

максимальный периметр вписанного m*-угольника с вершинами в точках* $U_1, \ldots, U_n, m \ge 3$*. Тогда для каждого* $t > 0$ *верно, что*

$$\lim_{n\to\infty} \mathbb{P}\{n^{\frac{2m}{m?1}}(2m\sin\frac{\pi}{m} - P_{m,n}) \le t\} = 1 - exp\{-\frac{t^{\frac{m-1}{2}}}{K_{1m}}\},$$

где $K_{1m} = m^{\frac{3}{2}}(\pi\sin\frac{\pi}{m})^{\frac{m-1}{2}}\Gamma(\frac{m+1}{2})$.

Новые результаты

В этой статье я обобщила метод, придуманный Лао и Майером, на более широкий класс U-max статистик. В приведенной ниже теореме показано, что для изучения предельного поведения этого класса U-max статистик необходимо только знание определителя Гессиана в точках максимума.

Теорема. *Пусть $U_1, \dots, U_m$ — независимо и равномерно распределенные точки на единичной окружности S_1 с центром в точке O. Обозначим через*

$$\beta_i = \angle U_{i+1} O U_1.$$

Рассмотрим функцию

$$f(U_1, \dots, U_m) = h(\beta_1, \dots, \beta_{m-1})$$

(иными словами, функция f не изменяется под действием поворота).

Предположим, что выполняются следующие условия:

1. *Функция f не меняется относительно перестановок U_i.*
2. *Функция h непрерывна (в топологии на $\mathbb{R} \cup -\infty$).*
3. *Функция, непрерывная на компакте даже с наличием $-\infty$, достигает своего максимума. Обозначим его через M (он же равен максимуму функции f). Потребуем того, что этот максимум достигался лишь в конечном числе точек $V_1, \dots, V_k \in [0, 2\pi]^{m-1}$. При этом предположим, что ни одна из этих точек не лежит на границе области определения. Иными словами*

   $$V_i^j > 0 \text{ и } V_i^j < 2\pi \text{ для всех } i \in \{1, \dots, k\}, j \in \{1, \dots, m-1\},$$

   где V_i^j — $j-$я координата точки V_i.
4. *Существует $\delta > 0$, что функция h трижды непрерывно дифференцируема в $\delta-$окрестности V_i для любого $i \in \{1, \dots, k\}$ (то есть в окрестностности любой точки максимума).*
5. *Рассмотрим матрицу Гессе G^i следующего вида:*

$$G^i = \begin{pmatrix} \frac{\partial^2 h(V_i)}{\partial^2 x_1} & \frac{\partial^2 h(V_i)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 h(V_i)}{\partial x_1 \partial x_{m-1}} \\ \frac{\partial^2 h(V_i)}{\partial x_1 \partial x_2} & \frac{\partial^2 h(V_i)}{\partial^2 x_2} & \cdots & \frac{\partial^2 h(V_i)}{\partial x_2 \partial x_{m-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 h(V_i)}{\partial x_{m-1} \partial x_1} & \frac{\partial^2 h(V_i)}{\partial x_{m-1} \partial x_2} & \cdots & \frac{\partial^2 h(V_i)}{\partial^2 x_{m-1}} \end{pmatrix}.$$

Предположим, что для всех точек максимума определитель таких матриц не равен 0. Будем обозначать его через $det(G^i)$.

Тогда для любого $t > 0$ верно равенство

$$\lim_{n\to\infty} \mathbb{P}\{n^{\frac{2m}{m-1}}(M - H_n) \le t)\} = 1 - e^{-t^{\frac{m-1}{2}}/K},$$

где $K = m!(2\pi)^{\frac{m-1}{2}}\Gamma(\frac{m+1}{2})\frac{1}{\left(\sum_{i=1}^{k}\frac{1}{\sqrt{det(-G^i)}}\right)}$.

Также я воспользовалась этой теоремой для изучения некоторых новых U-max статистик. Ниже приведены некоторые из полученных результатов.

Теорема. *Пусть $U_1, \dots, U_n$ — независимо и равномерно распределенные точки на единичной окружности S_1. Через $sq(V_1, V_2, V_3)$ обозначим сумму квадратов сторон треугольника с вершинами в точках V_1, V_2, V_3 (т.е. $sq(V_1, V_2, V_3) = \sum_{i=1}^{3} |V_iV_{i+1}|^2$, считаем, что $V_4 = V_1$). Пусть*

$$S_n = \max_{1\le i<j<k\le n} sq(U_i, U_j, U_k)$$

это максимальная сумма квадратов сторон треугольника среди всех треугольников с вершинами среди $U_1, \dots U_n$. Тогда для любого $t > 0$

$$\lim_{n\to\infty} \mathbb{P}\{n^3(9 - S_n) \le t\} = 1 - e^{-\frac{t}{6\sqrt{3}\pi}}.$$

Теорема. *Пусть $U_1, \dots, U_m$ — независимо и равномерно распределеныые точки на единичной окружности S_1. Рассмотрим описанный m-угольник, касающийся окружности S_1 в точках $U_1, \dots, U_m$. Вершины этого m-угольника обозначим через $A_1, \dots, A_m$ (возможно, одна из этих точек находится на бесконечности). Обозначим через*

$$h(U_1, \dots, U_m) = \sum_{i=1}^{m} OA_i$$

сумму длин отрезков, соединяющих центр единичной окружности O с вершинами описанного многоугольника. Пусть

$$H_n = \min_{1\le i_1<\dots<i_m\le n} h(U_{i_1}, \dots, U_{i_m}).$$

Тогда

$$\lim_{n\to\infty} \mathbb{P}\{n^{-\frac{2m}{m-1}}\left(-\frac{m}{\cos\frac{\pi}{m}} + H_n\right) \le t\} = 1 - e^{\frac{-t^{\frac{m-1}{2}}}{K}},$$

где $K = m\sqrt{m}\Gamma\left(\frac{m+1}{2}\right)\pi^{\frac{m-1}{2}}\left(\frac{\sin^2\frac{\pi}{m}+1}{2\cos^3\frac{\pi}{m}}\right)^{\frac{m-1}{2}}$.

Как можно видеть выше, пользоваться теоремой можно как в случае известной размерности U-max статистики, так и в случае неизвестной размерности.

Заключение

В данной работе метод изучения предельного поведения U-max статистик был обобщен на более широких класс U-max статистик. Также было обнаружено, что на предельное поведение этого класса U-max статистик оказывает влияние только определитель гессиана в точках максимума, что дает общее представление о предельном поведении U-max статистик данного класса, а также значительно упрощает изучение поведения конкретных U-max статистик.

Литература

[1] A.D. Barbour, L. Holst, S. Janson, Poisson Approximation, Oxford University Press, London, 1992.

[2] P.R. Halmos, The theory of unbiased estimation, Ann. Math. Statist. 17 (1946) 34–43.

[3] W. Hoeffding, A class of statistics with asymptotically normal distribution, Ann. Math. Statist. 19 (1948) 293–325.

[4] W. Lao, Some weak limit laws for the diameter of random point sets in bounded regions. Ph.D. Thesis, Karlsruhe, 2010.

[5] W. Lao, M. Mayer, U-max-statistics, J. Multivariate Anal. 99 (2008) 2039–2052.

[6] M. Mayer, Random Diameters and Other U-max-Statistics. Ph.D. Thesis, Bern University, 2008.

[7] F.B. Silverman, T. Brown, Short distances, flat triangles, and Poisson limits, J. Appl. Probab. 15 (1978) 815–825.

[8] E.V.Koroleva, Ya. Yu. Nikitin, U-max-statistics and limit theorems for perimeters and areas of random polygons, J. Multivariate Anal. 127 (2014) 99–111.