ВЕРОЯТНОСТНЫЕ РАСПРЕДЕЛЕНИЯ ПОРЯДКОВЫХ СТАТИСТИК В АНАЛИЗЕ СВЕРХКОРОТКИХ НЕЧЕТКИХ И НЕПОЛНЫХ ВРЕМЕННЫХ РЯДОВ

ПАЩЕНКО А.Е., СУВОРОВА А.В., ТУЛУПЬЕВА Т.В., ТУЛУПЬЕВ А.Л.

УДК 311.2 + 616-036.22

Пащенко А.Е., Суворова А.В., Тулупьева Т.В., Тулупьев А.Л., **Вероятностные распределения порядковых статистик в анализе сверхкоротких нечетких и неполных временных рядов.**

Аннотация. Статья отражает результаты очередного этапа исследований, посвященных подходам к оценке интенсивности рискованного поведения. Статья содержит описание способов формирования указанной оценки интенсивности на основе максимального и минимального интервала, а также интервала-медианы между эпизодами поведения. Решение рассматриваемой задачи основывается на формировании и анализе формул для функции распределения (и совместного распределения) соответствующих порядковых статистики, плотности распределения, а также на выборе значений его параметров. Предложены подходы к анализу качества полученных оценок.

Ключевые слова: нечеткие временные ряды, дефицит информации, гранулярность данных, модели поведения, поддержка принятия решений.

Paschenko A.E., Suvorova A.V., Tulupyeva T.V., Tulupyev A.L., **Probabilistic distributions of ordinal statistics in the analysis of super-short fuzzy and incomplete time series.**

Abstract. The paper presents results of a stage of research devoted to approaches to an estimation of rate of risky behavior. The paper contains the description of ways of formation of the specified estimation of the rate on the basis of the maximum and minimum interval, and also an median interval between the behavior episodes. A solution of the problem is based on formation and the analysis of formulas for distribution (and joint distribution) functions corresponding to ordinal statistics, distribution density, and also on a choice of values of their parameters. Several approaches to the analysis of quality of the received estimations are offered.

Keywords: fuzzy time series, deficiency of the information, behavior model, decision-making support.

1. Введение. Для простоты изложения рассмотрим одну модельную задачу из области современной эпидемиологии, а затем обсудим другие возможные применения предлагаемого подхода. В современной эпидемиологии остро стоит вопрос об оценке риска передачи и приобретения опасных неизлечимых инфекций (например, ВИЧ). Наиболее точно такой риск характеризуется инсиденс-показателем (числом заразившихся за определенный период среди лиц, подвергавшихся риску заражения, отнесенным к человеко×месяцам наблюдения). Для прямого измерения инсиденс-показателя требуется организовать когортное исследование, длящееся не менее полутора лет и подразумевающее вовлечение и сопровождение 500–1000 лиц из групп риска. Однократное проведение подобного когортного исследования обхо-

дится в 1.5–2.0 млн долларов. Такой уровень расходов делает невозможным мониторинг инсиденс-показателя даже в странах с сильной экономикой Требуется предложить математические модели, позволяющие выполнить более дешевые косвенные измерения инсиденс-показателя на основе ответов респондентов, составляющих выборку из группы риска.

Инсиденс-показатель можно оценить, зная индивидуальный риск заражения за заданный период времени каждого отдельного респондента. Модель Белла—Тревино [7] увязывает оценку риска с числом эпизодов рискованного поведения. Число же эпизодов можно оценить, если, в свою очередь, известна оценка интенсивности рискованного (дизадаптивного) поведения, рассмотренного как случайный процесс определенного класса.

Следует отметить, что в случае опроса респондентов данные поступают на естественном (разговорном) языке, т. е. являются в значительной степени нечеткими и неполными. Такие высказывания необходимо систематизировать, классифицировать и формализовать для их последующей обработки. В результате интервью нам становятся известными сведения о нескольких (до трех-четырех) последних эпизодах поведения, о максимальном, минимальном и обычном интервале между эпизодами. Заметим, что ответы респондента на вопросы о последних эпизодах и о «рекордных» (максимальном и минимальном) интервалах между эпизодами рискованного поведения характеризуются стабильностью воспроизведения. Однако ограниченное число и неточность, недоопределенность, нечеткость естественно-языковых формулировок ответов (т. е. наблюдаемый сверхкороткий временной ряд) не позволяют напрямую использовать известные методы из теории массового обслуживания для оценки интенсивности поведения.

Целью настоящей работы является описание способов обработки полученной информации о минимальном и максимальном интервалах, а также интервале-медиане между эпизодами рискованного поведения, способов формирования оценки интенсивности, основанной на этих «рекордных» интервалах. Достижение данной цели основывается на формировании и анализе формул для функции распределения (совместного распределения) соответствующих порядковых статистик и плотностей распределения. Кроме того, планируется описать подходы к анализу качества полученных оценок.

Хотя в качестве модельного примера рассматривается описанная выше эпидемиологическая проблема, полученные математические модели и методы могут использоваться в других областях (например, в

маркетинге, анализе приверженности, оценки строгости соблюдения режима лечения).

2. Обработка информации о последних эпизодах. Исследуется поведение индивидов в социальной группе, при этом предполагается, что поведение состоит из серии эпизодов, а с каждым эпизодом связан риск (например, риск заражения неизлечимой инфекцией) или вероятность какого-то события (например, покупки или передачи определенного товара). Серия эпизодов рассматривается как пуассоновский случайный процесс.

В результате интервью становятся известными сведения о нескольких последних эпизодах поведения. В работе [1] приведены классификация ответов, полученных на естественном языке в результате опроса о трех последних эпизодах рискованного поведения. Представим схематично возможные варианты их взаимного расположения. Пусть (0) — это момент интервью; (1), (2), (3) — моменты на оси времени, когда произошел последний, предпоследний и предпредпоследний эпизод поведения; t_{01}, t_{12}, t_{23} — длины временных интервалов соответственно между моментом интервью и последним эпизодом, последним и вторым эпизодом, вторым и третьим эпизодом поведения в прошлом; t — весь временной промежуток, за который произошли рассматриваемые эпизоды.

Возможны три класса ответов:

1. «Вложенные интервалы», когда указывается временной интервал между моментом интервью и каждым эпизодом, например: «вчера, позавчера, неделю назад» (рис. 1).

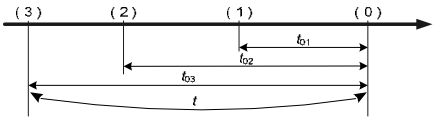

Рис. 1. Вложенные интервалы.

2. «Последовательные интервалы», когда респондент указывает эпизоды рискованного поведения, начиная с предпоследнего, отсчитывая их от момента предыдущего эпизода, например: «вчера, за неделю до этого, за неделю до этого» (рис. 2).

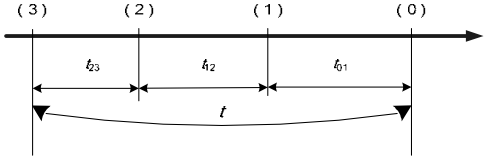

Рис. 2. Последовательные интервалы.

3. «Смешанные интервалы», являющиеся комбинацией предыдущих двух классов, например: «вчера, позавчера, еще за день до этого» (рис. 3).

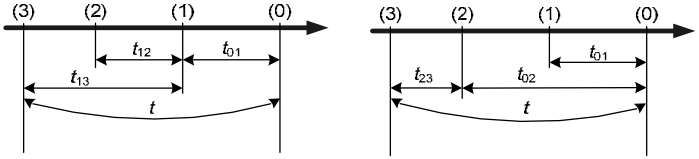

Рис. 3. Смешанные интервалы.

Для обработки последнего класса используется обобщенная схема представления ответов. В этом случае эксперт имеет дело с объединенными в последовательности эпизодами поведения, для каждой последовательности дается оценка, а итоговая оценка складывается из суммы оценок таких последовательностей (рис. 4).

Рис. 4. Обобщенная схема представления ответов.

Цель наших дальнейших исследований — определить или оценить значение параметра λ, характеризующее интенсивность участия респондента в поведении определенного вида, которое описывается пуассоновским случайным процессом. Получив оценку параметра λ, можно посчитать вероятность того, что в интервале $[t_0, t_0 + t]$ произойдут k событий:

$$\mathrm{P}\left[N\left([t_0, t_0 + t]\right) = k \right] = \frac{e^{-\lambda t} (\lambda t)^k}{k!}.$$

Один из вариантов получения оценки интенсивности рассмотрен в работе [1]. Пусть дано ν — число последовательных эпизодов от момента интервью, которые вспомнил респондент, а $\tau = t_{0\nu}$ — тот период времени, в течение которого эти эпизоды имели место. Применим метод максимального правдоподобия к основному уравнению пуассоновского процесса при вышеуказанных данных, чтобы найти соответствующую оценку интенсивности λ:

$$g(\lambda) = \frac{(\tau\lambda)^{\nu}}{\nu!} e^{-\tau\lambda},$$

$$h(\lambda) = \ln g(\lambda) = \nu \ln \lambda + \nu \ln \tau - \ln \nu! - \tau\lambda,$$

$$\frac{dh(\lambda)}{d\lambda} = \frac{\nu}{\lambda} - \tau,$$

$$\frac{dh(\lambda)}{d\lambda} = 0 \Rightarrow \lambda = \frac{\nu}{\tau}.$$

Как правило, исходя из ответов респондентов, удается дать количественную оценку числу эпизодов между моментом интервью и наиболее отдаленным эпизодом включительно. Если, в частности, респондент ответил на все вопросы о последних трех эпизодах, то $\nu = 3$.

В силу существенной неопределенности высказываний получить точную численную оценку τ (в рассматриваемом случае τ — это интервал времени между моментом интервью и самым отдаленным от него эпизодом включительно) затруднительно или даже невозможно. Однако ее можно рассмотреть как случайную величину, построенную над другими случайными величинами. Особенности построения такой случайной величины подробно рассмотрены в работе [1].

Пусть известны данные о N последних эпизодах поведения t_1, t_2, t_3,… , t_N а τ — общий временной промежуток, за который произошли эпизоды (связь между t_i и τ подробно рассмотрена в работе [1]). Интенсивность поведения λ оценивается по формуле: $\lambda = \frac{N}{\tau}$.

Для каждого эпизода со значением t_i, $1 \le i \le N$ (где N — число рассматриваемых эпизодов поведения) через характеристику разброса δ определяется интервал (*возможных* значений) в сутках:

$$[t_i - \delta x, t_i + \delta x],$$

где x — коэффициент перевода рассматриваемой единицы измерения в дни [1] (см. таблицу).

Точность и единицы измерения

Единица измерения	Значение x
Часы	1/24
Дни	1
Недели	7
Месяцы	30
Полугоды	183
Годы	365

То есть оценка в днях включает в себя минимальное значение $t_i - \delta x$, указанное в интервью значение t_i и максимальное значение $t_i + \delta x$.

Заметим, что любая точка из интервала $[t_i - \delta x, t_i + \delta x]$ возможна в качестве значения оценки t_i; что, однако, не означает, что точки из этого интервала равновероятны в качестве такого.

Сведения подобных отношениях между допустимыми значениями можно задать с помощью распределения их вероятностей. В зависимости от предположений о характере ответов респондента для задания случайной величины $\hat{t}_i$ оценки t_i используется равномерное, биномиальное или какое-либо другое вероятностное распределение.

Введенная случайная величина $\hat{t}_i$ за счет рандомизации неопределенности ответа позволяет рассмотреть интенсивность как случайную величину и вычислить характеристики последней.

Для каждого эпизода соответствующий интервал разбивается на n частей. Рассматриваются все возможные сочетания точек из интервалов, соответствующих указанным эпизодам. Расчет среднего значения для случая трех последних эпизодов производится по следующей формуле:

$$\lambda_{\text{среднее}} = \sum_{i,j,k} (\lambda_{ijk} p_i p_j p_k),$$

где p_i — вес i-й точки из первого интервала, p_j — вес j-й точки из первого интервала, p_k — вес k-й точки из первого интервала, λ_{ijk} — оценка интенсивности для соответствующего сочетания точек, т. е.

$$\lambda_{ijk} = \frac{N}{\tau_{ijk}},$$

где τ_{ijk} — соответствующая точкам i, j, k оценка величины рассматриваемого интервала.

Частные случаи этой формулы для конкретных распределений подробно рассмотрены в работе [2]. Например, для равномерного распределения

$$\lambda_{\text{среднее}} = \frac{1}{n^3} \sum_{i,j,k} \lambda_{ijk} \ .$$

Особый интерес представляет поиск распределений весов, при которых целевые показатели можно аналитически выразить через доступные исходные данные. Если связь выражается аналитически, то можно пользоваться численными методами, описанными выше.

3. Эвристическая оценка интенсивности на основе «рекордных» интервалов. Особый интерес представляют ответы респондентов, содержащие сведения о максимальном и минимальном интервале между эпизодами рискованного поведения за заданный период времени. Оценка интенсивности для рассматриваемого модельного примера на основе этих данных получена в работах [1, 3]. Отметим, что моделью поведения выступает пуассоновский процесс с уравнением

$$\Pr(\Delta t, k, \lambda) = \frac{(\lambda \Delta t)^k}{k!} e^{-\lambda \Delta t},$$

где Δt — промежуток времени наблюдения за поведением респондента; k — число эпизодов рассматриваемого поведения в течение этого промежутка; λ — интенсивность поведения; $\Pr(\Delta t, k, \lambda)$ — вероятность того, что за промежуток времени наблюдения Δt при поведении с интенсивностью[1] λ случится ровно k эпизодов указанного поведения.

Заметим, что для такого процесса известна также плотность распределения T — длины временного интервала между двумя соседними эпизодами: $p(T) = \lambda e^{-\lambda T}, T \geq 0.$

Пусть заданы $T_{\max}$ — максимальная длина временного интервала между двумя соседними эпизодами и $T_{\min}$ — минимальная длина временного интервала между двумя соседними эпизодами. Для того, чтобы получить оценку интенсивности λ по этим данным, рассмотрим подход, нацеленный на максимизацию вероятности того, что длины интервалов между эпизодами попадают в промежуток $[T_{\min}; T_{\max}]$.

[1] Под интенсивностью поведения понимается число эпизодов поведения рассматриваемого вида в определенный промежуток времени.

Вычислим эту вероятность как функцию от интенсивности λ.

$$f(\lambda) = p(T \in [T_{\min}; T_{\max}]) = \int_{T_{\min}}^{T_{\max}} \lambda e^{-\lambda\tau} \mathrm{d}\tau = e^{-\lambda T_{\min}} - e^{-\lambda T_{\max}}.$$

Определим, при каком λ функция $f(\lambda)$ принимает максимальное значение:

$$f'(\lambda) = -T_{\min} e^{-\lambda T_{\min}} + T_{\max} e^{-\lambda T_{\max}} = 0,$$

$$T_{\max} e^{-\lambda T_{\max}} = T_{\min} e^{-\lambda T_{\min}},$$

$$\ln T_{\max} - \lambda T_{\max} = \ln T_{\min} - \lambda T_{\min},$$

$$\ln T_{\max} - \ln T_{\min} = \lambda (T_{\max} - T_{\min}),$$

$$\lambda = \frac{\ln T_{\max} - \ln T_{\min}}{T_{\max} - T_{\min}}.$$

Таким образом, оценка $\hat{\lambda}$, основанная на известных экстремальных значениях, вычисляется по формуле:

$$\hat{\lambda} = \left(\ln \frac{T_{\max}}{T_{\min}} \right) \Big/ \left(T_{\max} - T_{\min} \right).$$

Как показал опыт применения выведенной формулы, получаемые оценки интенсивности правдоподобны и согласуются с мнением экспертов. Однако предлагаемый подход к оценке интенсивности требует, во-первых, более глубокого математического обоснования и, во-вторых, анализа свойств полученного результата.

4. Вероятностные распределения порядковых статистик. Поскольку связь порядковых статистик и распределений случайных величин систематически анализируется в работе [4], в изложении базовых теоретических вопросов будем следовать указанному источнику. Рассмотрим n случайных величин $X_1, X_2, ..., X_n$, заданных на одном вероятностном пространстве. Если расположить эти величины в порядке возрастания, то получится последовательность, которая называется вариационным рядом. Обозначим i-й элемент такого ряда $X_{i:n}$, т. е. $X_{1:n} \le X_{2:n} \le ... \le X_{n:n}$. Каждый элемент вариационного ряда называется порядковой статистикой.

Заметим, что

$$X_{1:n} = \min\{X_1, X_2, ..., X_n\},$$

а

$$X_{n:n} = \max\left\{X_1, X_2, ..., X_n\right\},$$

поэтому эти порядковые статистики также будем обозначать $X_{\min}$ и $X_{\max}$ соответственно.

В дальнейшем нами используется понятие медианы, обозначим ее X_{med}. Медианой вариационного ряда называется середина этого ряда:

$$X_{\text{med}} = \begin{cases} X_{0.5(n+1):n}, & \text{если } n \text{ нечетное}, \\ 0.5\left(X_{0.5n:n} + X_{(0.5n+1):n}\right), & \text{если } n \text{ четное}. \end{cases}$$

Далее в настоящей работе для простоты в рассуждениях и формулах, в которых участвует медиана, предполагается, что выборка, а значит и построенный на ее основе вариационный ряд имеют нечетное число элементов.

Мы рассмотрим только случай, когда последовательность $X_1, X_2, ..., X_n$ состоит из независимых случайных величин, имеющих общую функцию распределения F. Тогда функция распределения для последнего элемента $X_{\max}$ вариационного ряда имеет вид:

$$F_{\max}(x) = P\left\{X_{\max} < x\right\} = P\left\{X_1 < x, X_2 < x, \ldots, X_n < x\right\} =$$

$$= \prod_{i=1}^{n} P\left\{X_i < x\right\} = \left(F(x)\right)^n.$$

Аналогично для $X_{\min}$:

$$F_{\min}(x) = P\left\{X_{\min} < x\right\} = 1 - P\left\{X_{\min} \geq x\right\} = 1 - P\left\{X_1 \geq x, X_2 \geq x, \ldots, X_n \geq x\right\} =$$

$$= 1 - \prod_{i=1}^{n} P\left\{X_i \geq x\right\} = 1 - \prod_{i=1}^{n} \left(1 - P\left\{X_i < x\right\}\right) = 1 - \left(1 - F(x)\right)^n.$$

Теперь выведем формулу для функции распределения $F_{i:n}(x)$ произвольной порядковой статистики $X_{i:n}$:

$$F_{i:n}(x) = P\{X_{i:n} < x\} =$$

$$= P\left\{\begin{array}{c}\text{не менее } i \text{ элементов вариационного ряда } X_1,...,X_n \\ \text{расположены левее } x\end{array}\right\} =$$

$$= \sum_{m=i}^{n} P\{\text{ровно } m \text{ элементов из } X_1,...,X_n \text{ расположены левее } x\} =$$

$$= \sum_{m=i}^{n} C_n^m \left(F(x)\right)^m \left(1-F(x)\right)^{n-m}.$$

В частности, для медианы X_{med} при n нечетном

$$F_{\text{med}}(x) = \sum_{m=0.5(n+1)}^{n} C_n^m \left(F(x)\right)^m \left(1-F(x)\right)^{n-m}.$$

Для дальнейших рассуждений докажем тождество[4]:

$$\sum_{m=k}^{n} C_n^m y^m (1-y)^{n-m} = \int_0^y \frac{n!}{(k-1)!(n-k)!} t^{k-1}(1-t)^{n-k}\,dt, \quad 0 \le y \le 1.$$

Вычислим производную по y левой части этого тождества:

$$g'(y) = \left(\sum_{m=k}^{n} C_n^m y^m (1-y)^{n-m}\right)' =$$

$$= \sum_{m=k}^{n} m C_n^m y^{m-1}(1-y)^{n-m} + \sum_{m=k}^{n} -(n-m) C_n^m y^m (1-y)^{n-m-1} =$$

$$= \sum_{m=k}^{n} m \frac{n!}{m!(n-m)!} y^{m-1}(1-y)^{n-m} - \sum_{m=k}^{n} (n-m)\frac{n!}{m!(n-m)!} y^m (1-y)^{n-m-1} =$$

$$= \sum_{m=k}^{n} \frac{n!}{(m-1)!(n-m)!} y^{m-1}(1-y)^{n-m} - \sum_{m=k}^{n} \frac{n!}{m!(n-m-1)!} y^m (1-y)^{n-m-1} =$$

$$= \frac{n!}{(k-1)!(n-k)!} y^{k-1}(1-y)^{n-k} + \sum_{m=k+1}^{n} \frac{n!}{(m-1)!(n-m)!} y^{m-1}(1-y)^{n-m} -$$

$$- \sum_{l=k+1}^{n} \frac{n!}{(l-1)!(n-l)!} y^{l-1}(1-y)^{n-l} = \frac{n!}{(k-1)!(n-k)!} y^{k-1}(1-y)^{n-k}.$$

Производная правой части:

$$h'(y)=\left(\int_0^y \frac{n!}{(k-1)!(n-k)!}t^{k-1}(1-t)^{n-k}\,dt\right)'=$$

$$=\frac{n!}{(k-1)!(n-k)!}y^{k-1}(1-y)^{n-k}.$$

Таким образом $g'(y)=h'(y)$. Функции дифференцируемы, а значит непрерывны. Замечая, что $g(0)=h(0)=0$, получим: $g(y)=h(y),\quad 0\le y\le 1$, т. е. тождество доказано.

Используя это тождество и формулу для $F_{i:n}(x)$, получим другое выражение для функции распределения:

$$F_{i:n}(x)=\int_0^{F(x)} \frac{n!}{(i-1)!(n-i)!}t^{i-1}(1-t)^{n-i}\,dt,\quad -\infty<x<+\infty.$$

Рассмотрим совместную функцию распределения $F_{i,j:n}$ для двух порядковых статистик $X_{i:n}$ и $X_{j:n}$, где $1\le i<j\le n,\ x<y$:

$$F_{i,j:n}(x,y)=P\{X_{i:n}<x, X_{j:n}<y\}=$$

$$=P\left\{\begin{array}{c}\text{не менее } j \text{ элементов расположены левее } y,\\ \text{из них не менее } i \text{ расположены левее } x\end{array}\right\}=$$

$$=\sum_{s=j}^{n}P\left\{\begin{array}{c}\text{ровно } j \text{ элементов расположены левее } y,\\ \text{из них не менее } i \text{ расположены левее } x\end{array}\right\}=$$

$$=\sum_{s=j}^{n}\sum_{r=i}^{s}P\left\{\begin{array}{c}\text{ровно } j \text{ элементов расположены левее } y,\\ \text{из них ровно } i \text{ расположены левее } x\end{array}\right\}=$$

$$=\sum_{s=j}^{n}\sum_{r=i}^{s}C_n^r\prod_{k=1}^{r}P\{X_k<x\}\cdot C_{n-r}^{s-r}\prod_{k=1}^{s-r}P\{x\le X_k<y\}\cdot\prod_{k=1}^{n-s}P\{X_k\ge y\}=$$

$$=\sum_{s=j}^{n}\sum_{r=i}^{s}\frac{n!}{r!(s-r)!(n-s)!}\big(F(x)\big)^r\big(F(y)-F(x)\big)^{s-r}\big(1-F(y)\big)^{n-s}.$$

Например, для совместного распределения $X_{\max}$ и $X_{\min}$:

$$F_{\min,\max}(x,y)=\sum_{r=1}^{n}\frac{n!}{r!(n-r)!}\big(F(x)\big)^r\big(F(y)-F(x)\big)^{n-r}.$$

Совместное распределение $X_{\max}$ и X_{med} (n нечетное):

$$F_{\text{med,max}}(x,y)=\sum_{r=\frac{n+1}{2}}^{n}\frac{n!}{r!(n-r)!}\left(F(x)\right)^{r}\left(F(y)-F(x)\right)^{n-r}.$$

Совместное распределение $X_{\min}$ и X_{med} (n нечетное):

$$F_{\min,\text{med}}(x,y)=\sum_{s=\frac{n+1}{2}}^{n}\sum_{r=1}^{s}\frac{n!}{r!(s-r)!(n-s)!}\left(F(x)\right)^{r}\left(F(y)-F(x)\right)^{s-r}\left(1-F(y)\right)^{n-s}.$$

Предположим, что случайные величины $X_1, X_2, ..., X_n$ имеют плотность распределения f. Тогда для почти всех x $F'(x)=f(x)$ и правая часть выражения для $F_{i:n}(x)$ дифференцируема, причем $F'_{i:n}(x)=f_{i:n}(x)$, т. е. $f_{i:n}(x)$ — плотность порядковой статистики $X_{i:n}$:

$$f_{i:n}(x)=\frac{n!}{(i-1)!(n-i)!}\left(F(x)\right)^{i-1}\left(1-F(x)\right)^{n-i}f(x),\quad -\infty<x<+\infty.$$

В частности для порядковой статистики $X_{\max}$ плотность выражается формулой:

$$f_{\max}(x)=n\left(F(x)\right)^{n-1}f(x),\quad -\infty<x<+\infty.$$

Для $X_{\min}$: $f_{\min}(x)=n\left(1-F(x)\right)^{n-1}f(x),\quad -\infty<x<+\infty$.

Для X_{med} (n нечетное):

$$f_{\text{med}}(x)=\frac{n!}{\left[\left(0.5(n-1)\right)!\right]^{2}}\left[F(x)\left(1-F(x)\right)\right]^{\frac{n-1}{2}}f(x),\quad -\infty<x<+\infty.$$

Рассмотрим совместную плотность распределения r порядковых статистик $X_{k(1):n}, X_{k(2):n}, ..., X_{k(r):n}$.

Обозначим ее $f_{k(1),k(2),...k(r):n}(x_1,x_2,...,x_r)$, причем $0=k(0)<k(1)<k(2)<...<k(r)<k(r+1)=n+1$, $1\le r\le n$. Тогда эта плотность выражается следующей формулой:

$$f_{k(1),k(2),...k(r):n}(x_1,x_2,...,x_r)=$$
$$=\frac{n!}{\prod_{s=1}^{r+1}\left(k(s)-k(s-1)-1\right)!}\prod_{s=1}^{r+1}\left(F(x_s)-F(x_{s-1})\right)^{k(s)-k(s-1)-1}\prod_{s=1}^{r}f(x_s),$$

если $-\infty = x_0 < x_1 < x_2 < ... < x_r < x_{r+1} = +\infty$, и

$f_{k(1),k(2),...k(r):n}(x_1, x_2, ..., x_r) = 0$ в остальных случаях. Доказательство приводится в работе [4].

Рассмотрим частные случаи этой формулы. При $r = 2$:

$$f_{i,j:n}(x,y) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \times$$

$$\times \left(F(x)\right)^{i-1} \left(F(y) - F(x)\right)^{j-i-1} \left(1 - F(y)\right)^{n-j} f(x) f(y),$$

если $-\infty < x < y < +\infty$, и $f_{i,j:n}(x,y) = 0$ в остальных случаях.

Например, если

$$f_{\min,\max}(x,y) =$$

$$= \frac{n!}{(n-2)!} \left(F(y) - F(x)\right)^{n-2} f(x) f(y),$$

при $-\infty < x < y < +\infty$, и $f_{\min,\max}(x,y) = 0$ в остальных случаях;

а если n нечетное, то

$$f_{\min,\text{med}}(x,y) = \frac{n!}{\left(0.5(n-3)\right)!\left(0.5(n-1)\right)!} \times$$

$$\times \left(F(y) - F(x)\right)^{\frac{n-3}{2}} \left(1 - F(y)\right)^{\frac{n-1}{2}} f(x) f(y),$$

при $-\infty < x < y < +\infty$, и $f_{\min,\text{med}}(x,y) = 0$ в остальных случаях;

если же n нечетное, то

$$f_{\text{med},\max}(x,y) = \frac{n!}{\left(0.5(n-1)\right)!\left(0.5(n-3)\right)!} \times$$

$$\times \left(F(x)\right)^{0.5(n-1)} \left(F(y) - F(x)\right)^{0.5(n-3)} f(x) f(y),$$

при $-\infty < x < y < +\infty$, и $f_{\text{med},\max}(x,y) = 0$ в остальных случаях.

При $r = 3$:

$$f_{i,j,k:n}(x,y,z) = \frac{n!}{(i-1)!(j-i-1)!(k-j-1)(n-k)!} \times$$

$$\times \left(F(x)\right)^{i-1} \left(F(y) - F(x)\right)^{j-i-1} \left(F(z) - F(y)\right)^{k-j-1} \times$$

$$\times \left(1 - F(y)\right)^{n-k} f(x) f(y) f(z),$$

если $-\infty < x < y < z < +\infty$, и $f_{i,j,k:n}(x,y,z) = 0$ в остальных случаях.

Например при n нечетном

$$f_{\min,\mathrm{med},\max}(x,y,z) =$$

$$= \frac{n!}{\left[\left(0.5(n-3)\right)!\right]^2}\left[\left(F(y)-F(x)\right)\left(F(z)-F(y)\right)\right]^{0.5(n-3)} f(x)f(y)f(z),$$

если $-\infty < x < y < z < +\infty$, и $f_{\min,\mathrm{med},\max}(x,y,z) = 0$ в остальных случаях.

Выше все формулы рассматривались для общего случая — функции распределения F, удовлетворяющей определенным требованиям гладкости. Применим полученные результаты для задачи, поставленной в разделе 2, где рискованное поведение моделировалось как пуассоновский процесс. Тогда функция распределения F промежутков между эпизодами указанного поведения примет конкретный вид:

$$F(t) = 1 - e^{-\lambda t}.$$

Затем, используя рассмотренные ранее формулы, получим следующие выражения для функций распределения минимального $T_{\min}$, максимального $T_{\max}$ и обычного (считаем его медианой) T_{med} промежутков между эпизодами рискованного поведения:

$$F_{\min}(t) = 1 - e^{-n\lambda t},$$

$$F_{\max}(t) = \left(1 - e^{-\lambda t}\right)^n,$$

при n нечетном: $F_{\mathrm{med}}(t) = \sum_{m=0.5(n+1)}^{n} C_n^m \left(1 - e^{-\lambda t}\right)^m e^{-(n-m)\lambda t}$.

Теперь рассмотрим функции для различных вариантов совместного (попарного) распределения трех указанных порядковых статистик. Сначала — совместное распределение $T_{\max}$ и $T_{\min}$:

$$F_{\min,\max}(x,y) = \sum_{r=1}^{n} \frac{n!}{r!(n-r)!}\left(1 - e^{-\lambda x}\right)^r \left(e^{-\lambda x} - e^{-\lambda y}\right)^{n-r}.$$

Совместное распределение $T_{\max}$ и T_{med} (n нечетное):

$$F_{\mathrm{med},\max}(x,y) = \sum_{r=\frac{n+1}{2}}^{n} \frac{n!}{r!(n-r)!}\left(1 - e^{-\lambda x}\right)^r \left(e^{-\lambda x} - e^{-\lambda y}\right)^{n-r}.$$

Совместное распределение $T_{\min}$ и T_{med} (n нечетное):

$$F_{\min,\mathrm{med}}(x,y) =$$

$$= \sum_{s=\frac{n+1}{2}}^{n} \sum_{r=1}^{s} \frac{n!}{r!(s-r)!(n-s)!} \left(1-e^{-\lambda x}\right)^{r} \left(e^{-\lambda x}-e^{-\lambda y}\right)^{s-r} e^{-(n-s)\lambda y}.$$

Заметив, что плотность распределения f промежутков между эпизодами рискованного поведения выражается как $f(t) = \lambda e^{-\lambda t}$, выпишем формулы плотности распределения для сочетаний трех указанных порядковых статистик:

1) $f_{\min}(t) = n\lambda e^{-n\lambda t}$, $\quad -\infty < t < +\infty$;

2) $f_{\max}(t) = n\lambda e^{-\lambda t}\left(1-e^{-\lambda t}\right)^{n-1}$, $\quad -\infty < t < +\infty$;

3) $f_{\mathrm{med}}(t) = \dfrac{n!}{\left[\left(0.5(n-1)\right)!\right]^2} \lambda e^{-\lambda t\frac{n+1}{2}} \left(1-e^{-\lambda t}\right)^{\frac{n-1}{2}}$, $\quad -\infty < t < +\infty$;

4) $f_{\min,\max}(x,y) = \dfrac{n!}{(n-2)!} \lambda^2 e^{-\lambda(x+y)} \left(e^{-\lambda x}-e^{-\lambda y}\right)^{n-2}$,

если $-\infty < x < y < +\infty$, и $f_{\min,\max}(x,y) = 0$ в остальных случаях;

5) n нечетное:

$$f_{\min,\mathrm{med}}(x,y) =$$

$$= \frac{n!}{\left(0.5(n-3)\right)!\left(0.5(n-1)\right)!} \lambda^2 e^{-\lambda\left(x+\frac{n+1}{2}y\right)} \left(e^{-\lambda x}-e^{-\lambda y}\right)^{\frac{n-3}{2}},$$

если $-\infty < x < y < +\infty$, и $f_{\min,\mathrm{med}}(x,y) = 0$ в остальных случаях;

6) n нечетное:

$$f_{\mathrm{med},\max}(x,y) =$$

$$= \frac{n!}{\left(0.5(n-1)\right)!\left(0.5(n-3)\right)!} \lambda^2 e^{-\lambda(x+y)} \left(1-e^{-\lambda x}\right)^{\frac{n-1}{2}} \left(e^{-\lambda x}-e^{-\lambda y}\right)^{\frac{n-3}{2}},$$

если $-\infty < x < y < +\infty$, и $f_{\mathrm{med},\max}(x,y) = 0$ в остальных случаях;

7) n нечетное:

$$f_{\min,\mathrm{med},\max}(x,y,z) =$$

$$= \frac{n!}{\left[\left(0.5(n-3)\right)!\right]^2} \lambda^3 e^{-\lambda(x+y+z)} \left(\left(e^{-\lambda x}-e^{-\lambda y}\right)\left(e^{-\lambda y}-e^{-\lambda z}\right)\right)^{\frac{n-3}{2}},$$

если $-\infty < x < y < z < +\infty$, и $f_{\min,\mathrm{med},\max}(x,y,z) = 0$ в остальных случаях.

5. Идентификация интенсивности методом максимального правдоподобия на основе порядковых статистик. Исследуется определенный (заданный перед опросом респондента) период времени T, в течение которого имели место эпизоды рискованного поведения. Этот период отсчитывается от момента интервью назад (рис. 5). Респондент дает ответы о максимальном, минимальном и обычном интервале между эпизодами поведения в заданный период. Ответы могут содержать сведения как о всем наборе указанных величин, так и о каком-то его подмножестве.

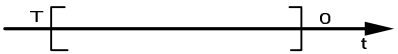

Рис. 5. Интервал ретроспективы.

Пусть $T_{\max}$ — максимальный интервал между эпизодами рискованного поведения респондента, $T_{\min}$ — минимальный и T_{med} — обычный интервал. Как отмечалось, в результате интервью нам могут быть известны различные сочетания данных: только $T_{\max}$, только $T_{\min}$, только T_{med}, $T_{\max}$ и $T_{\min}$, $T_{\max}$ и T_{med}, $T_{\min}$ и T_{med}, одновременно $T_{\max}$, $T_{\min}$, T_{med}. Для всех этих вариантов выведены в разделе 4 формулы плотности распределения соответствующих порядковых статистик и их сочетаний.

Для оценки интенсивности рискованного поведения в указанный период T применяем метод максимального правдоподобия, т. е. ищем такую оценку интенсивности λ, которая максимизирует плотность

$$f_{\min,\mathrm{med},\max}(T_{\min}, T_{\mathrm{med}}, T_{\max}; \lambda)$$

при известных ответах респондента $T_{\max}$, $T_{\min}$, T_{med}. Но эта функция, как было показано выше, зависит от числа n эпизодов рассматриваемого поведения за исследуемый промежуток времени T. В задаче, сформулированной таким образом, необходимо исследовать функцию, в которой неизвестны как интенсивность поведения λ, так и число эпизодов n. Однако можно предложить способ, который позволит перейти к использованию одного параметра.

Предполагая, что респондент не ошибся при ответе, можно оценить число эпизодов рискованного поведения. Легко заметить, что это

число должно принадлежать промежутку $\left[\frac{T}{T_{\max}}, \frac{T}{T_{\min}}\right]$. Таким образом, все возможные значения n можно перебрать. Каждому значению n сопоставляется значение интенсивности λ, вычисляемое по формуле $\lambda = \frac{n}{T}$. В этом случае функция плотности $f(t_{\min}, t_{\text{med}}, t_{\max}; \lambda)$ рассматривается как функция $\tilde{f}(t_{\min}, t_{\text{med}}, t_{\max}; n)$, зависящая не от интенсивности λ, а от числа эпизодов n. Тогда каждому значению $n_0, n_1, ..., n_k$ из интервала $\left[\frac{T}{T_{\max}}, \frac{T}{T_{\min}}\right]$ соответствует значение функции $\tilde{f}(t_{\min}, t_{\text{med}}, t_{\max}; n)$, причем $\tilde{f}_i = \tilde{f}(t_{\min}, t_{\text{med}}, t_{\max}; n_i), \quad 0 \le i \le k$.

В качестве оценки метода максимального правдоподобия берем такое n^*, что $n^* = n_j : \tilde{f}_j = \max\{\tilde{f}_i\}, 0 \le \text{i} \le \text{k}$, т. е. n^* максимизирует плотность $\tilde{f}(t_{\min}, t_{\text{med}}, t_{\max}; n)$.

Точечная оценка интенсивности в таком случае примет вид $\lambda^* = \frac{n^*}{T}$.

6. Оценка риска и оценка качества. Как описано в разделе 5, каждому значению $n_0, n_1, ..., n_k$ из интервала $\left[\frac{T}{T_{\max}}, \frac{T}{T_{\min}}\right]$ соответствует значение функции $\tilde{f}_i = \tilde{f}(t_{\min}, t_{\text{med}}, t_{\max}; n_i), \quad 0 \le i \le k$ и оценка n^* имеет вид:

$$n^* = n_j : \tilde{f}_j = \max\{\tilde{f}_i\}, 0 \le \text{i} \le \text{k}.$$

Нормируем значения $\tilde{f}_i$ и для каждого n_i получим вес (вероятность) p_i, такой, что $p_i = \tilde{f}_i \Big/ \sum_{l=0}^{k} \tilde{f}_l$.

Таким образом, чтобы учесть неполноту и неопределенность исходных данных, получаем возможность рассмотреть число эпизодов n как случайную величину, причем $P\{n = n_i\} = p_i, \quad 0 \le i \le k$. Построим отрезок, характеризующий «качество» точечной оценки числа эпизодов (указанный отрезок можно будет далее использовать для характе-

ристики качества точечной оценки риска заражения) — фактически речь пойдет о построении доверительного интервала одним из численных методов. Заметим, что по определению n^* вес $p^* = P\{n = n^*\}$ максимальный: $p^* = \max\{p_i\}$, $0 \le i \le k$. Для получения нужного отрезка будем формировать множество индексов, изначально содержащее лишь n^*. На каждом шаге указанное множество представляет собой подотрезок $n_L \dots n_R$ отрезка $n_0 \dots n_k$ натурального ряда. Если сумма вероятностей его элементов меньше заранее заданного значения α, то отрезок пополняется двумя соседними ему значениями n_{L-1} и n_{R+1}, пока их суммарный вес не достигнет α. Например, пусть $n^* = n_j$, $0 \le j \le k$. Сначала суммарный вес $p = p_j$. Если $p \ge \alpha$, то отрезок включает только n^*. Иначе — добавляем значения $n_{j-1} = n^* - 1$ и $n_{j+1} = n^* + 1$, их суммарный вес $p = p + p_{j-1} + p_{j+1}$. Если все еще $p < \alpha$, то добавляем следующие два значения n_{j-2} и n_{j+2} (или только одно из них, если с одной из сторон вышли за границы промежутка $\left[\frac{T}{T_{\max}}, \frac{T}{T_{\min}}\right]$), пересчитываем суммарный вес $p = p + p_{j-2} + p_{j+2}$. Продолжаем, пока не достигнем того, что $p \ge \alpha$. Таким образом получим отрезок, который условно можно считать доверительным интервалом для числа n эпизодов рискованного поведения.

Используя оценку интенсивности рискованного поведения, можно оценить и кумулятивный риск, связанный с указанным поведением. Зависимость выражается формулой [5]:

$$\Pr = 1 - e^{-\lambda p T},$$

где λ — интенсивность, p — риск заражения за один эпизод, T — период в днях, $\Pr$ — кумулятивный риск.

Учитывая, что $\lambda = \frac{n}{T}$, получим $\Pr = 1 - e^{-np}$.

Рассмотрим два подхода к вычислению оценки кумулятивного риска:

1) использовать значение n^*: $\Pr^* = 1 - e^{-n^* p}$;

2) рассмотреть число эпизодов n как случайную величину, т. е. использовать веса-вероятности p_i, $0 \le i \le k$:

$$\mathrm{Pr}^* = \mathsf{E}\,\mathrm{Pr} = \sum_{i=0}^{k} p_i\,\mathrm{Pr}_i = \sum_{i=0}^{k} p_i\left(1 - e^{-n_i p}\right).$$

Во втором случае можно вычислить и другие характеристики такой оценки, например дисперсию. Зная дисперсию и используя неравенство Чебышева, можно в свою очередь построить доверительный интервал, хотя последний можно далее уточнять за счет использования менее «грубых», с точки зрения оценивания, техник.

7. Дискуссия. Значительная часть положений, приведенных выше, составляют математическую основу плана дальнейших исследования. Развитие указанных положений, рассмотрение возможных вариантов исходных данных, систематический учет их особенностей и последовательное применение приемов, использующихся в мягких вычислениях, гранулярных[1] вычислениях, теории нечетких рядов, теории вероятностей и математической статистике позволит сформулировать большую часть решения поставленной выше задачи оценки риска, связанного с угрозообразующим поведением. Такая задача имеет и очевидные прикладные аспекты, и существенную фундаментальную составляющую.

Заметим, что главную роль в обработке гранулярных данных играет метод анализа и синтеза сводных показателей при информационном дефиците, предложенный проф. Н.В. Ховановым [8]. Результаты, полученные с помощью указанного метода, можно развить с помощью методов анализа нечетких временных рядов, опираясь, в частности, на результаты отечественных исследовательских коллективов, возглавляемых Н.Г. Ярушкиной [9] и С.М. Ковалевым [10].

Кроме того, на данный момент не рассмотрен вопрос о согласованности различных исходных данных, например, о согласованности данных о последних эпизодах рискованного поведения между собой, о согласованности данных о «рекордных» (минимальном и максимальном) интервалах между эпизодами рискованного поведения и, наконец, о согласованности данных о последних эпизодах и данных о «рекордных» интервалах.

Еще одно возможное направление дальнейших исследований — построение относительных оценок интенсивностей и рисков (так называемое odds ratio). Такие оценки необходимы для сравнения разных групп или же для сравнения характеристик одной и той же группы до

[1] Гранулярность — естественно-языковая неточность, нечеткость.

и после проведения поведенческой интервенции — комплекса мероприятий, нацеленного на модификацию поведения. Например, на снижение интенсивности участия в поведении, которое связано с угрозой приобретения или передачи гепатита С. Требуется оценить *априорные* (до вмешательства) и *апостериорные* (после вмешательства) оценки интенсивности и кумулятивного риска, а также отношения этих оценок, характеризующие влияние поведенческой интервенции на индивидов.

Следует заметить, что задачи, аналогичные рассматриваемому модельному примеру, но в иных терминах, встречаются, в частности, при оценке уровня террористических угроз, степени деструктивности некоторых видов социальных организаций, уровня защищенности обслуживающего персонала и пользователей информационных систем от социоинженерных атак, в маркетинговых исследованиях и других отраслях междисциплинарных исследований. Этот же подход представляется чрезвычайно интересным с точки зрения психологии, особенно отраслей, связанных с поведением человека. В частности, используя предложенный подход, можно изучать дизадаптивное поведение, в том числе связанное с функционированием психологической защиты. Исходные данные, методы их обработки и структура моделей, требующиеся для решения этих задач, с математической точки зрения, имеют одинаковую природу: решение задач в одной области легко обобщается для решения их аналогов из других областей. Однако существенными останутся усилия, посвященные адаптации «средства измерения» — опросного инструментария, который в значительной степени специфичен для конкретной предметной области, и затем развитию формальной классификации ответов, полученных на естественном языке.

Литература

1. *Пащенко А. Е., Тулупьева Т. В., Тулупьев А. Л.* Оценка интенсивности поведения респондента в условиях информационного дефицита // Тр. СПИИРАН. 2008. Вып. 7. С. 239–254.
2. *Пащенко А.Е., Суворова А.В.* Программный комплекс для экспертного оценивания интенсивности поведения респондента в условиях дефицита информации // Докл. науч.-практической конф. студентов, аспирантов, молодых ученых и специалистов «Интегрированные модели, мягкие вычисления, вероятностные системы и комплексы программ в искусственном интеллекте» (Коломна, 26–27 мая 2009 г.). Т. 2. М., 2009. С. 220–241.
3. *Пащенко А.Е., Тулупьев А.Л., Суворова А.В., Тулупьева Т.В.* Сравнение параметров угрозообразующего поведения в разных группах на основе неполных и неточных данных // Тр. СПИИРАН. 2009. Вып. 9. С. 252–261.
4. *Невзоров В. Б.* Рекорды. Математическая теория. М.: ФАЗИС, 2000. 244 с.

5. *Тулупьева Т. В., Тулупьев А. Л., Пащенко А. Е., Красносельских Т. В.* Приверженность ВААРТ и рискованное поведение среди пациентов Санкт-Петербургского Центра-СПИД: статистические модели, психологические и социо-демографические факторы // Тр. СПИИРАН. 2008. Вып. 6. С. 207–237.
6. *Тулупьева Т. В., Пащенко А. Е. , Тулупьев А. Л., Красносельских Т. В. и др*. Модели ВИЧ-рискованного поведения в контексте психологической защиты и других адаптивных стилей. СПб.: Наука, 2008. 140 с.
7. *Bell D. C., Trevino R. A.* Modeling HIV Risk [Epidemiology] // J. Acquir Immune Defic Syndr. 1999. Vol. 22, N 3. P. 280–287.
8. *Хованов Н. В*. Анализ и синтез показателей при информационном дефиците. СПб.: Изд-во СПбГУ, 1996. 196 с.
9. *Ярушкина Н. Г.* Современный интеллектуальный анализ нечетких временных рядов // Сб. науч. тр. V-й Междунар. науч.-практической конф. «Интегрированные модели и мягкие вычисления в искусственном интеллекте». Т. 1. С. 19–29.
10. *Ковалев С.М.* Гибридные коннекционистские модели извлечения темпоральных знаний // Сб. науч. тр. V-й Междунар. науч.-практической конф. «Интегрированные модели и мягкие вычисления в искусственном интеллекте». Т. 1. С. 30–40.

Пащенко Антон Евгеньевич — младший научный сотрудник научно-исследовательской группы междисциплинарных проблем информатики Учреждения Российской академии наук Санкт-Петербургский институт информатики и автоматизации РАН (СПИИРАН). Область научных интересов: математическая статистика, статистическое моделирование, применение методов биостатистики и математического моделирования в эпидемиологии. Число научных публикаций — 35. AEP@iias.spb.su, www.tulupyev.spb.ru; СПИИРАН, 14-я линия В.О., д.39, Санкт-Петербург, 199178, РФ; р.т. +7(812)328-3337, факс +7(812)328-4450.

Paschenko Anton Evgen'evich — junior researcher, Interdisciplinary Computer Science Research and Development Group, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: mathematical statistics, statistical modeling, application of biostatistics and mathematical modeling in epidemiology. The number of publications — 35. AEP@iias.spb.su, www.tulupyev.spb.ru; SPIIRAS, 39, 14th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-3337, fax +7(812)328-4450.

Суворова Алена Владимировна — студент мат.-мех. факультета Санкт-Петербургского государственного университета (СПбГУ). Область научных интересов: математическая статистика, теория вероятности. Число научных публикаций — 6. SUVALV@mail.ru; СПбГУ, математико-механический факультетт, Университетский пр., д.28, Петродворец, Санкт-Петербург, 198504, РФ. Научный руководитель — А.Л. Тулупьев.

Suvorova Alena Vladimirovna — student of the Faculty of Mathematics and Mechanics of the Saint Petersburg State University. Research interests: mathematical statistics, probability theory. The number of publications — 6. SUVALV@mail.ru; St.Petersburg State University, Faculty of Mathematics and Mechanics, 28, Universitetsky prospekt, 198504, Peterhof, St. Petersburg, Russia. Supervisor — A.L. Tulupiev.

Тулупьев Александр Львович — канд.физ.-мат. наук, доцент; ведущий научный сотрудник научно-исследовательской группы междисциплинарных проблем информатики

Учреждения Российской академии наук Санкт-Петербургский институт информатики и автоматизации РАН (СПИИРАН), доцент кафедры информатики математико-механического факультета СПбГУ. Область научных интересов: представление и обработка данных и знаний с неопределенностью, применение методов математики и информатики в социокультурных исследованиях, применение методов биостатистики и математического моделирования в эпидемиологии, технология разработки программных комплексов с СУБД. Число научных публикаций — 140. ALT@iias.spb.su, www.tulupyev.spb.ru; СПИИРАН, 14-я линия В.О., д. 39, Санкт-Петербург, 199178, РФ; р.т. +7(812)328-3337, факс +7(812)328-4450.

Tulupyev Alexander Lvovich — Ph.D. in Appl. Math. and CS, associate professor; leading researcher, Interdisciplinary Computer Science Research and Development Group, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), associate professor, Computer Science Department, Faculty of Mathematics and Mechanics, St. Petersburg State University (SPbSU). Research interests: uncertain knowledge and data representation and processing, application of mathematics and computer science in sociocultural studies, applications of biostatistics and mathematical modeling in modern epidemiology, software technologies and development of information systems with databases. The number of publications — 140. ALT@iias.spb.su, www.tulupyev.spb.ru; SPIIRAS, 39, 14th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-3337, fax +7(812)328-4450.

Тулупьева Татьяна Валентиновна — канд. психол. наук, доцент; старший научный сотрудник научно-исследовательской группы междисциплинарных проблем информатики Учреждения Российской академии наук Санкт-Петербургский институт информатики и автоматизации РАН (СПИИРАН), доцент кафедры информатики математико-механического факультета СПбГУ, доцент кафедры психологии управления и педагогики Северо-Западной академии государственной службы (СЗАГС). Область научных интересов: применение методов математики и информатики в гуманитарных исследованиях, информатизация организации и проведения психологических исследований, применение методов биостатистики в эпидемиологии, психология личности, психология управления. Число научных публикаций — 45. TVT@iias.spb.su, www.tulupyev.spb.ru; СПИИРАН, 14-я линия В.О., д.39, Санкт-Петербург, 199178, РФ; р.т. +7(812)328-3337, факс +7(812)328-4450.

Tulupyeva Tatiana Valentinovna — Ph.D. in Psychology, associate professor; senior researcher, Interdisciplinary Computer Science Research and Development Group, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), associate professor, Computer Science Department, Faculty of Mathematics and Mechanics, St. Petersburg State University (SPbSU), associate professor, Management Psychology and Pedagogic Department, North-West Academy of Public Administration (NWAPA). Research interests: application of mathematics and computer science in humanities, informatization of psychological studies, application of biostatistics in epidemiology, psychology of personality, management psychology. The number of publications — 45. TVT@iias.spb.su, www.tulupyev.spb.ru; SPIIRAS, 39, 14th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-3337, fax +7(812)328-4450.

Рекомендовано ЛПИ СПИИРАН, зав. лаб. Р.М. Юсупов , чл.-корр. РАН.
Статья поступила в редакцию 20.12.2009.

РЕФЕРАТ

Пащенко А.Е., Суворова А.В., Тулупьев А.Л., Тулупьева Т.В. **Вероятностные распределения порядковых статистик в анализе сверхкоротких нечетких и неполных временных рядов.**

В ряде отраслей научных исследований стоит задача оценки интенсивности поведения респондентов по неполным и неточным исходным данным. Источниками таких данных зачастую являются высказывания респондента на естественном (разговорном) языке, когда такие высказывания — единственный приемлемый способ получения сведений об интенсивности поведения. Полученная оценка интенсивности, как правило, используется дальше для косвенного оценивания других показателей.

Таким образом, возникает существенная потребность в развитии математических моделей (опирающихся, в том числе, и на математические методы и объекты из области мягких вычислений), позволяющих перейти от ограниченного числа неточных ответов о последних эпизодах рискованного поведения, о максимальном, минимальном и обычном интервале между эпизодами к оценке интенсивности указанного поведения, а затем и к косвенной оценке индивидуального риска, с ним связанного.

В статье особое внимание уделено обработке ответов респондентов, содержащих сведения о максимальном и минимальном интервале между эпизодами рискованного поведения за заданный период времени; предложено несколько подходов к оценке интенсивности, опирающейся на известные сведения об указанных «рекордных» интервалах. Подходы основаны на построении и анализе функций распределения и функций совместного распределения соответствующих порядковых статистик. Гранулярность (естественно-языковая неточность, нечеткость) исходных данных обрабатывается с использованием приемов, предложенных Н.В. Хованов в методе анализа и синтеза сводных показателей при информационном дефиците. Помимо скалярных оценок рассматриваются интервальные оценки интенсивности; последние играют роль оценки «качества» первых.

SUMMARY

Paschenko A.E., Syvorova A.V., Tulupyeva T.V., Tulupyev A.L,
Probabilistic distributions of ordinal statistics in the analysis of super-short fuzzy and incomplete time series

In a number of branches of scientific research, there is a problem of an estimate of rate of behavior of respondents under the incomplete and inexact initial data. Sources of such data can frequently be just statements of the respondent in a natural language, when such statements is the only mean to communicate knowledge about behavior. The calculate estimate of rate, as a rule, is used further for indirect estimate of other indicators.

Thus, an essential need arises for development of mathematical models (based, for example, on mathematical methods and objects from the area of soft calculations), allowing to pass from the limited number of inexact answers about last episodes of risky behavior, about the maximum, minimum and usual interval between episodes to an estimate of rate of the behavior, and to an indirect estimate of individual risk.

The paper pays the special attention to processing of answers of respondents with data on the maximum and minimum interval between episodes of risky behavior for the given period of time; several approaches to an estimate of the rate of risky behavior is offered. The approaches are based on construction and the analysis of functions of distributions and functions of joint distributions. Granulated (natural language discrepancy, an illegibility) initial data are processed with the method of N.V.Hovanov for the analysis and synthesis of aggregated indicators in case of information deficiency. Besides, interval estimate of rate are considered; the latter plays a special role in an estimate of "quality" of the scalar estimate.