УДК 311.2 + 616-036.22

А.В. Суворова, А.Е. Пащенко, Т.В. Тулупьева

ОЦЕНКА ХАРАКТЕРИСТИК СВЕРХКОРОТКОГО ВРЕМЕННОГО РЯДА ПО ГРАНУЛЯРНЫМ ДАННЫМ О РЕКОРДНЫХ ИНТЕРВАЛАХ МЕЖДУ СОБЫТИЯМИ

Суворова А.В., Пащенко А.Е., Тулупьева Т.В. **Оценка характеристик сверхкороткого временного ряда по гранулярным данным о рекордных интервалах между событиями.**

Аннотация. Рассматривается развитие процедур оценивания интенсивности поведения по данным о минимальном, максимальном и обычном интервале между эпизодами указанного поведения. Математической моделью поведения выступает пуассоновский процесс; наблюдения представляют собой ответы респондентов на естественном языке об указанных интервалах. Описанный в статье подход учитывает гранулярность исходных данных и основывается на использовании аппарата порядковых статистик и применении метода рандомизации. Рассмотрены примеры применения полученных оценок для анализа не только рискованного, но и других видов социально-значимого поведения.

Ключевые слова: гранулярность данных, сверхкороткие временные ряды, модели поведения, дефицит информации.

Suvorova A.V., Paschenko A.E., Tulupyeva T.V. **Super-short time series' parameters estimate on the base of granular data about record intervals between episodes.**

Abstract. An improved approach to behavior rate estimates on the base of data about minimum, maximum and usual interval between behavior episodes is considered. The mathematical model of this behavior is a Poisson stochastic process; the observations are respondents' natural language answers about mentioned intervals. The considered method takes account of data granularity; it is based on the analysis of ordinal statistics and method of randomization. In the paper, there are several examples of the estimates applications to some other types of socially significant behavior, including risky behavior.

Keywords: data granularity, super-short time series, behavior models, information deficiency.

1. Введение. В ряде отраслей социологических, психологических, маркетинговых исследований стоят задачи оценивания интенсивности поведения респондентов, причем эти оценки вычисляются по данным, полученным по самоотчетам респондентов об их поведении (ответы на вопросы интервью или анкеты). Например, в эпидемиологии важен вопрос оценки риска передачи или приобретения неизлечимых инфекций (например, ВИЧ), а для этого необходимо знать интенсивность рискованного поведения респондентов.

Следует отметить, что в случае опроса респондентов данные поступают на естественном языке, т.е. являются в значительной степени нечеткими и неполными. Такие высказывания необходимо систематизировать, классифицировать и формализовать для их последующей

обработки. Методы такой формализации и классификации подробно рассмотрены в [1–3]. 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 естественно-языковых формулировок ответов (т.е. наблюдаемый сверхкороткий временной ряд [4]) не позволяют напрямую использовать известные методы из теории массового обслуживания для оценки интенсивности поведения.

Методы построения точечных оценок интенсивности рискованного поведения респондентов и риска, с ним связанного по данным о максимальном, минимальном и обычном интервале между эпизодами рассматриваемого поведения подробно описаны в работе [5]. В ней сформированы на основе материалов книги В.Б. Невзорова [6] и проанализированы формулы для функций распределения (как совместных, так и по отдельности) соответствующих порядковых статистик и плотностей распределения.

Как уже отмечалось, сведения о рекордных интервалах представлены на естественном языке, что и определяет такую их особенность как гранулярность формулировок. Другими словами бытовой язык диктует использование привычных оценок длительности, не отличающихся особой точностью. Например, высказывание «неделя» для характеристики максимального интервала между эпизодами не означает то же самое, что «семь дней» или «168 часов». Приходится учитывать гранулярность данных.

Целью предлагаемого исследования является развитие математических моделей обработки полученной информации о минимальном и максимальном интервалах, а также интервале-медиане между эпизодами рискованного поведения, и развитие способов формирования оценки интенсивности, основанной на этих «рекордных» интервалах.

2. Постановка задачи. Рассмотрим формальную постановку задачи. Исследуется определенный (заданный перед опросом респондента) период времени T, за который произошли эпизоды рискованного поведения. Этот период отсчитывается от момента интервью назад (рис. 1). Респондент дает ответы о максимальном, минимальном и обычном интервале между эпизодами поведения, случившимися в за-

данный период. Ответы могут содержать сведения как обо всем наборе указанных величин, так и о каком-то его подмножестве.

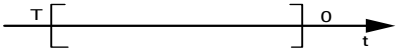

Рис. 1. Интервал ретроспективы.

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 формулы плотности распределения соответствующих порядковых статистик и их сочетаний [5]:

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

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:

$$f_{\min,\text{med}}(x,y) = \frac{n!}{\left(\dfrac{n-3}{2}\right)!\left(\dfrac{n-1}{2}\right)!}\lambda^2 e^{-\lambda\left(x+\frac{n+1}{2}y\right)}\left(e^{-\lambda x}-e^{-\lambda y}\right)^{\frac{n-3}{2}},$$

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

6) n нечетное:

$$f_{\text{med},\max}(x,y) = \frac{n!}{\left(\dfrac{n-1}{2}\right)!\left(\dfrac{n-3}{2}\right)!}\lambda^2 e^{-\lambda(x+y)}\left(1-e^{-\lambda x}\right)^{\frac{n-1}{2}}\left(e^{-\lambda x}-e^{-\lambda y}\right)^{\frac{n-3}{2}},$$

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

7) n нечетное:

$$f_{\text{min,med,max}}(x, y, z) = \frac{n!}{\left(\left(\frac{n-3}{2}\right)!\right)^2} \lambda^3 e^{-\lambda(x+y+z)} \left(\left(e^{-\lambda x} - e^{-\lambda y}\right)\left(e^{-\lambda y} - e^{-\lambda z}\right)\right)^{\frac{n-3}{2}},$$

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

Процедура построения оценки интенсивности рискованного поведения в указанный период T подробно описана в [5]. Она основана на применении метода максимального правдоподобия. То есть мы ищем такую оценку интенсивности λ, которая максимизирует плотность $f_{\text{min,med,max}}(T_{\min}, T_{\text{med}}, T_{\max}; \lambda)$ при известных ответах респондента $T_{\max}$, $T_{\min}$, T_{med}. Но эта функция, как показано выше, зависит от числа n эпизодов рассматриваемого поведения за исследуемый промежуток времени T. Поэтому в работе [5] предложен следующий метод. 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 $n^* = n_j : \tilde{f}_j = \max\left\{\tilde{f}_i\right\}, 0 \le i \le k$, т.е. n^* максимизирует

плотность $\tilde{f}\left(t_{\min}, t_{\text{med}}, t_{\max}; n\right)$. Точечная оценка интенсивности в таком случае имеет вид: $\lambda^* = \dfrac{n^*}{T}$.

Отметим, что респонденты используют в своих высказываниях преимущественно следующие единицы измерения: часы, дни, недели, месяцы, полугода, года. Причем использованная единица измерения несет в себе информацию о точности измерения. Поясним это на примере двух высказываний о длине интервала между эпизодами: «семь дней» и «неделя». Когда респондент использует формулировку «семь дней» (рис. 2а), это свидетельствует о высокой «надежности» припоминания и его уверенности в том, что событие произошло ровно 7 дней назад. Когда респондент использует формулировку «неделя», он априорно снижает точность высказывания (рис. 2б). Неделя — это и шесть дней и восемь. Таким образом, можно говорить о гранулярности получаемых ответов и рассматривать не точечное значение промежутка между эпизодами, а интервальное.

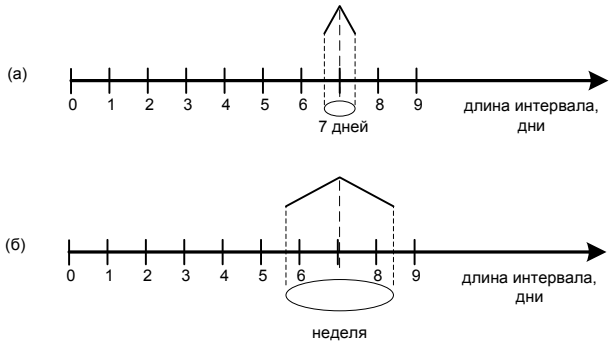

Рис. 2. Точность единиц измерения.

3. Рандомизация. В силу существенной неопределенности высказываний на естественном языке получить точную численную оценку t (в рассматриваемом случае t — это указываемая респондентом длина одного из рекордных интервалов между эпизодами) затруднительно или даже невозможно. Однако ее можно рассмотреть как случайную величину, построенную над другими случайными величинами. Особенности процесса построения такой случайной величины подробно рассмотрены в [2, 3, 7].

Для указанной респондентом длины t одного из рекордных интервалов между эпизодами определяется промежуток (возможных значений — или гранула на рис. 2) в днях $[t-\delta x, t+\delta x]$, где x — коэффициент перевода рассматриваемой единицы измерения в дни, δ — характеристика разброса (радиус гранулы на рис. 2).

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

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

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

Рассмотрим сначала случай, когда нам известно только одно значение t — либо длина минимального промежутка между эпизодами, либо максимального, либо обычного. Интервал $[t-\delta x, t+\delta x]$ разбивается на L частей. Каждая из получившихся таким образом точек t_i, $0 \le i \le L$, выступает как самостоятельное значение, для которого выполняется описанная выше процедура получения оценки интенсивности поведения методом максимального правдоподобия. Т.е. каждому t_i соответствует свое значение $\lambda_i^* = \frac{n_i^*}{T}$. Используя одно из подходящих вероятностных распределений, зададим веса p_i, $0 \le i \le L$. Например, для равномерного распределения $p_i = \frac{1}{L}$. Расчет весов для некоторых других распределений приведен в [8]. Таким образом мы получили дискретную случайную величину $\tilde{\lambda}^*$, принимающую значения λ_i^* с вероятностями p_i, $0 \le i \le L$. Расчет среднего, являющегося искомой оценкой интенсивности, производится по следующей формуле:

$$\lambda_{\text{avg}} = \sum_{i=0}^{L} \lambda_i^* p_i \ .$$

Продолжим этот подход на случай, когда нам известно более одного данного о рекордных интервалах. В этом случае каждый из соответствующих промежутков $[t_{\min} - \delta x, t_{\min} + \delta x]$, $[t_{\text{med}} - \delta x, t_{\text{med}} + \delta x]$, $[t_{\max} - \delta x, t_{\max} + \delta x]$ (или два из них — в зависимости от исследуемых данных) разбивается на L частей и рассматриваются все возможные сочетания точек из получившихся промежутков. При этом возможными сочетаниями считаются те, для которых выполняется соотношение $t_{i,\min} \le t_{\text{i,med}} \le t_{i,\max}$, $0 \le i \le L$. В качестве функции $f(t;n)$, используемой в процедуре получения точечной оценки интенсивности, применяется соответствующая совместная функция плотности распределения. Если промежутки

$$[t_{\min} - \delta x, t_{\min} + \delta x],\ [t_{\text{med}} - \delta x, t_{\text{med}} + \delta x],\ [t_{\max} - \delta x, t_{\max} + \delta x]$$

не пересекаются (рис. 3а), то возможны все сочетания точек, т.е. всего L^3 вариантов (когда даны сведения обо всех трех рекордных интервалах) или L^2 вариантов в случае, когда известны только два каких-то значения длин интервалов. Среднее значение рассчитывается по формуле:

$$\lambda_{\text{avg}} = \sum_{i=0}^{L} \left(\lambda_i^{\,*} p_i p_j p_r \right),$$

где p_i — вес i-й точки из $[t_{\min} - \delta x, t_{\min} + \delta x]$, p_j — вес j-й точки из $[t_{\text{med}} - \delta x, t_{\text{med}} + \delta x]$, p_r — вес r-й точки из $[t_{\max} - \delta x, t_{\max} + \delta x]$.

Если какое-либо значение неизвестно, то веса точек из соответствующего промежутка не учитываются.

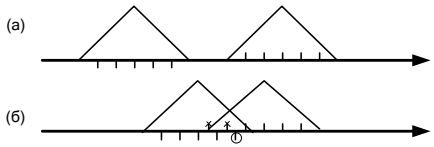

Рис. 3. Возможные сочетания точек.

В случае пересечения рассматриваемых промежутков среднее вычисляется по той же формуле, только не учитываются невозможные сочетания, при этом веса пересчитываются (нормируются) таким образом, чтобы для каждой фиксированной точки из одного промежутка сумма весов возможных точек из другого промежутка была равна единице.

Рассмотрим пример: пусть на рис. 3б левый промежуток (обозначенный треугольником) — это «гранула» для минимального интервала между эпизодами рискованного поведения, а правый — для максимального; а распределение вероятностей на каждом промежутке равномерное. Тогда для предпоследней, выделенной на рис. 3б, точки левого интервала первые две точки из правого невозможны — иначе получится, что максимальный интервал меньше минимального. В этом случае веса оставшихся пяти точек правого интервала при равномерном распределении будут равны не $\frac{1}{7}$, как раньше, а $\frac{1}{5}$.

В общем случае алгоритм пересчета весов при пересекающихся промежутках описывается следующим образом. В промежутке, строящемся по минимальному значению, веса определяются только заданным вероятностным распределением. В следующем веса нормируются для каждой точки из первого промежутка, и в третьем (если он есть) — нормируются по отношению к каждой точке из второго промежутка.

4. Заключение. На основе метода рандомизации, рассмотренного Н.В. Хованов[7] с использованием аппарата порядковых статистик были предложены методы оценки интенсивности поведения на основе рекордных интервалов между эпизодами рассматриваемого поведения. При этом данные оценки построены с учетом неточности, неполноты, а также гранулярности исходных данных, вызванной естественно-языковым характером изложения этих данных.

Литература

1. *Суворова А.В., Тулупьев А.Л., Пащенко А.Е., Тулупьева Т.В. и др.* Анализ гранулярных данных и знаний в задачах исследования социально значимых видов поведения // Компьютерные инструменты в образовании. 2010. № 4. С. 30–38.
2. *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.
3. *Пащенко А.Е., Тулупьева Т.В.* Применение процедуры рандомизации для оценки интенсивности поведения респондента в условиях информационного дефицита // Сб. науч. тр. V Междунар. науч.-практ. конф. «Интегрированные модели и мягкие вычисления в искусственном интеллекте». Т. 1. С. 743–751.
4. *Ярушкина Н.Г.* Современный интеллектуальный анализ нечетких временных рядов // Сб. науч. тр. V Междунар. науч.-практ. конф. «Интегрированные модели и мягкие вычисления в искусственном интеллекте». Т. 1. С. 19–29.
5. *Пащенко А.Е., Суворова А.В., Тулупьева Т.В. Тулупьев А.Л.* Вероятностные распределения порядковых статистик в анализе сверхкоротких нечетких и неполных временных рядов // Тр. СПИИРАН. 2009. Вып. 10. С. 184–207.
6. *Невзоров В. Б.* Рекорды. Математическая теория. М.: Изд. ФАЗИС, 2000. 244 с.
7. *Хованов Н. В.* Анализ и синтез показателей при информационном дефиците // СПб.: Изд-во СПбГУ, 1996. 196 с.

8. *Пащенко А.Е., Суворова А.В.* Программный комплекс для экспертного оценивания интенсивности поведения респондента в условиях дефицита информации // Докл. науч.-практ. 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.: Физматлит, 2009. С. 220–241.

Суворова Алена Владимировна — м. н. с. лаборатории теоретических и междисциплинарных проблем информатики Учреждения Российской академии наук Санкт-Петербургский институт информатики и автоматизации РАН (СПИИРАН), аспирант математико-механического факультета Санкт-Петербургского государственного университета (СПбГУ). Область научных интересов: математическая статистика, теория вероятности, применение методов математического моделирования в эпидемиологии. Число научных публикаций — 21. suvalv@mail.ru, 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. Научный руководитель — д-р физ.-мат. наук, доцент А.Л. Тулупьев.

Suvorova Alena Vladimirovna — junior researcher, Laboratory of Theoretical and Interdisciplinary Computer Science, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), PhD student, Faculty of Mathematics and Mechanics of St. Petersburg State University (SPbSU). Research interests: mathematical statistics, probability theory, application of mathematical modeling in epidemiology. The number of publications — 21. SUVALV@mail.ru, 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. Scientific advisor — PhD in Appl. Math. and CS, Dr. Sci. in CS A.L. Tulupiev.

Пащенко Антон Евгеньевич — м. н. с.лаборатории теоретических и 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 — 45. 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, Laboratory of Theoretical and Interdisciplinary Computer Science, 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 — 45. 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.

Тулупьева Татьяна Валентиновна — канд. психол. наук, доцент; старший научный сотрудник лаборатории теоретических и 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 исследова-

ниях, информатизация организации и проведения психологических исследований, применение методов биостатистики в эпидемиологии, психология личности, психология управления. Число научных публикаций — 70. 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 — PhD in Psychology, associate professor; senior researcher, Laboratory of Theoretical and Interdisciplinary Computer Science, 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 — 70. 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 ТиМПИ СПИИРАН, зав. лаб. А.Л. Тулупьев, д-р физ.-мат. наук, доцент.
Статья поступила в редакцию 06.12.2010.

Поддержка исследований. В публикации представлены результаты исследований, поддержанные грантом для молодых ученых и кандидатов наук от Правительства Санкт-Петербурга в 2009 №25.05/027/27 «Разработка математических моделей, вычислительных алгоритмов и комплекса программ для оценки интенсивности рискованного поведения в условиях дефицита информации». Руководитель — А.Е. Пащенко. Также исследования поддержаны грантом для молодых ученых и кандидатов наук от Правительства Санкт-Петербурга в 2010 «Разработка математических моделей, алгоритмов и распределенного комплекса программ для косвенной оценки рисков, связанных с угрозообразующим поведением». Руководитель — А.Е. Пащенко.

РЕФЕРАТ

Суворова А.В., Пащенко А.Е., Тулупьева Т.В. **Оценка характеристик сверхкороткого временного ряда по гранулярным данным о рекордных интервалах между событиями.**

В ряде отраслей социологических, психологических, маркетинговых исследований стоят задачи оценивания интенсивности поведения респондентов, причем эти оценки вычисляются по данным на естественном языке, полученным из ответов респондентов на вопросы об их поведении. Таким образом, данные опираются на память респондентов, поэтому приходится работать с очень небольшим числом параметров. Сведения о таких величинах, как максимальный, минимальный, обычный интервалы между эпизодами поведения, как правило, легко вспоминаются и отличаются стабильностью воспроизведения. Методы построения оценок интенсивности поведения именно на основе таких данных и рассмотрены в статье.

Использование бытового языка диктует применение привычных оценок длительности, не отличающихся особой точностью. Отметим, что, чем меньше единица измерения, тем большую точность ответа можно ожидать. Такую зависимость необходимо учитывать. Таким образом, можно говорить о гранулярности получаемых ответов и рассматривать не точечное значение величины промежутка между эпизодами, а интервальное.

Методы построения точечных оценок интенсивности рискованного поведения респондентов и риска, с ним связанного по данным о максимальном, минимальном и обычном интервале между эпизодами рассматриваемого поведения подробно рассмотрены в предыдущих работах. Указанные методы опирались, прежде всего, на использование формул для функций распределения (как совместных, так и по отдельности) соответствующих порядковых статистик. В данной статье рассмотрены подходы к развитию этих методов на случай, когда исходные данные представлены не точечным значением, а некоторым интервалом.

В силу существенной неопределенности высказываний на естественном языке, гранулярности исходных данных получить точную численную оценку указываемой респондентом длины одного из рекордных интервалов между эпизодами затруднительно или даже невозможно. В статье обработка таких данных основана на методе рандомизации, согласно которому исходная величина рассмотрена как случайная, построенная над другими случайными величинами.

Нами описана процедура построения такой случайной величины, проанализированы проблемы, возникающие при этом, и предложены методы их решения. На основе анализа этой случайной величины разработан подход к оценке интенсивности поведения, учитывающий гранулярность данных о минимальном, максимальном и обычном интервале между эпизодами рассматриваемого поведения.

SUMMARY

Suvorova A.V., Paschenko A.E., Tulupyeva T.V. **Super-short time series' parameters estimate on the base of granular data about record intervals between episodes.**

In some branches of sociological, psychological, epidemiological, marketing scientific research there is a problem of respondents' behavior rate estimate; these estimates are calculated on the base of natural language data, derived from respondents' answers to questions about their behavior. For the data based on respondents' memory we have to work with a very small number of parameters. Respondents often easily remember data about maximum, minimum and usual intervals between their behavior episodes. This paper proposes methods for behavior rate estimate based on such data.

Everyday language dictates the usage of usual duration estimates which are not very accurate. Note if we have the smaller measurement unit, we expect to get the more precise answer. We should consider this dependence. Thus, we can talk about the data granularity and we can consider the interval value of the time interval length between episodes.

Processing of respondents' risky behavior rate and risk estimate according to the data about the maximum, minimum and usual interval between behavior episodes considered in detail in previous papers. These methods are based on the applying of formulas for distribution functions (both together and separately) of the ordinal statistics. In this paper we consider the approaches to the improvement of these methods to the case of interval initial data.

Natural language is characterized by the uncertainty of statements used in, at the same time the source data is granular — that's why it is difficult or even impossible to obtain precise numerical estimate of record intervals between respondent's behavior episodes. In this paper the processing of such data is based on the method of randomization, that means that the value of interval length is considered as a random variable, based on other random variables.

We have described a procedure for constructing such random variable. The issues arise in this process are analyzed; several approaches to the issues elimination are proposed. In the paper we propose an approach to behavior rate estimate based on the analysis of described random variable; our approach takes the granularity of the data about minimum, maximum, and the usual interval between considered behavior episodes into account.