УДК 311.2 + 616-036.22

А.В. Лавренов, А.В. Суворова, А.Е. Пащенко

ОСОБЕННОСТИ ОБРАБОТКИ ДАННЫХ И ЗНАНИЙ ОБ ЭПИЗОДАХ СОЦИАЛЬНО-ЗНАЧИМОГО ПОВЕДЕНИЯ В ОКРЕСТНОСТИ ИНТЕРВЬЮ

Лавренов А.В., Суворова А.В., Пащенко А.Е. **Особенности обработки данных и знаний об эпизодах социально-значимого поведения в окрестности интервью.**
Аннотация. Рассматривается подход к улучшению процедур построения оценок различных параметров поведения респондентов по сведениям об их последних эпизодах. Предпринята попытка избежать неявного предположения о том, что следующий эпизод происходит в момент интервью (или в ближайшее время после него), поскольку такое утверждение зачастую не соответствует действительности. В работе подробно описаны недостатки такого подхода, а также предложены способы моделирования и обработки неопределенности, связанной с корректным учетом момента интервью и прогнозом следующего эпизода. Разработаны программные приложения, обеспечивающие возможность проведения численных экспериментов, реализующих предложенные модели обработки.
Ключевые слова: оценка интенсивности, модели поведения, последние эпизоды, неопределенность.

Lavrenov A.V., Suvorova A.V., Paschenko A.E. **Processing issues of data and knowledge about socially significant behavior' episodes near the interview moment.**
Abstract. The paper offers an improved method for respondents' behavior parameters estimates based on data about several last episodes. We try to avoid an assumption (which often is not correct) that the next behavior episode happens at the same time as the moment of interview does. The paper outlines some disadvantages of previous method and proposes modeling and processing of uncertainty related to a correct representation of the interview time moment and the next episode. The developed software calculates estimates according to the considered models and supports making numerical experiment.
Keywords: rate estimate, behavior models, last episodes, uncertainty.

1. Введение. Задачи оценивания интенсивности социально-значимого поведения респондентов возникают во многих отраслях социологических, психологических, маркетинговых исследований. Например, в эпидемиологии важен вопрос оценки риска передачи или приобретения неизлечимых инфекций (например, ВИЧ), а для этого необходимо знать интенсивность рискованного поведения респондентов. причем эти оценки вычисляются по данным, полученным по самоотчетам респондентов об их поведении (ответы на вопросы интервью или анкеты). Выделение групп потребителей, существенно различающихся интенсивностью потребления продуктов, товаров или услуг, в маркетинговых исследованиях позволяет сосредоточить усилия на тех группах, которые многочисленны, но товар потребляют неинтен-

сивно. Выявление их особенностей этих групп может привести к разработке стратегии, ведущей к существенному увеличению объема продаж.

Отметим, что наиболее доступными исходными данными для анализа поведения выступают самоотчеты респондентов о его поведении, то есть ответы в анкете на блок вопросов или результаты проведения интервью. На данный момент разработаны и применяются в опросах респондентов два подхода к оцениванию интенсивности поведения: прямые вопросы и Лайкерт-шкалы — каждый из которых имеет недостатки [1, 2]. Одной из возможных альтернатив представляется опрос респондента об одном или нескольких последних эпизодах его поведения. Заметим, что ответы респондента на такие вопросы о последних эпизодах характеризуются стабильностью воспроизведения. Однако ограниченное число и неточность, недоопределенность, нечеткость естественно-языковых формулировок ответов (то есть наблюдаемый сверхкороткий временной ряд) не позволяют напрямую использовать известные методы из теории массового обслуживания для оценки интенсивности поведения, поэтому возникает необходимость в предложении новых математических моделей.

Поведение рассматривается как случайный процесс некоторого класса. При этом встают вопросы о том, какой процесс лучше описывает поведение, как меняются параметры этого процесса, как осуществляется обработка неполных исходных данных. Цель данной статьи — описать проблемы, возникающие при анализе данных о последних эпизодах социально-значимого поведения, и предложить некоторые пути их решения.

2. Описание используемых подходов и некоторых их недостатков. Рассмотрим используемый подход к анализу данных в условиях дефицита информации [2]. Более подробно он описан в [1, 3–10].

В результате интервью становятся известными сведения о нескольких (в рассматриваемом случае — трех) последних эпизодах поведения. Серия эпизодов рассматривается как пуассоновский случайный процесс с основным уравнением [11]

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

где t_0 — начальный момент времени, k — число последовательных событий, которые вспомнил респондент, а T — тот период времени, за который эти эпизоды произошли, λ — интенсивность.

Цель исследований — определить или оценить величину параметра λ, характеризующего интенсивность участия респондента в поведении определенного вида, а также его производные характеристики.

Применим метод максимального правдоподобия [12] к основному уравнению пуассоновского процесса при вышеуказанных данных, чтобы найти соответствующую оценку интенсивности λ:

$$g(\lambda) = \frac{T\lambda^{\nu}}{k!} e^{-T\lambda};$$

$$h(\lambda) = \ln g(\lambda) = k\ln\lambda + k\ln T - \ln k! - T\lambda;$$

$$\frac{dh(\lambda)}{d\lambda} = \frac{k}{\lambda} - T;$$

$$\frac{dh(\lambda)}{d\lambda} = 0 \Rightarrow \lambda = \frac{k}{T}.$$

Рассмотренный подход, несмотря на логичные и правомерные мотивирующие соображения, имеет ряд недостатков из-за некоторых оставленных без внимания деталей. Одна из таких деталей и будет обсуждаться в данной работе.

Предположим, что мы опрашиваем людей на улице, как часто они ездят в большие гипермаркеты. Если человек посещает магазин строго один раз в n дней, то интенсивность его поведения $\lambda = \frac{1}{n}$. Предположим, что первый респондент посещает магазин по воскресениям, а в субботу мы спросили его о трёх последних эпизодах. Он ответил, что они происходили 6, 13 и 20 дней назад. Тем самым, по нашим оценкам, интенсивность его поведения $\lambda_1 = 3/20$, что очень близко к реальной интенсивности 1/7. Теперь предположим, что второй респондент посещает магазин раз в 10 дней, и на наши вопросы он ответил 1, 11 и 21 день назад. Его реальная интенсивность — 1/10, однако наши оценки говорят, что $\lambda_2 = 1/7$. Такое сильное отличие показывает, что использовав имеющиеся данные только частично, (а именно, мы пока пользовались только одним из трёх сообщённых респондентами интервалов), мы получили слишком приблизительную оценку интенсивности поведения. Для того, чтобы уточнить полученные выше оценки, рассмотрим другой подход к анализу параметров пуассоновского процесса.

Отметим, что для пуассоновского процесса нам известна функция распределения длин интервалов между соседними эпизодами поведе-

ния $F(\tau) = 1 - \lambda e^{-\lambda\tau}$, где λ — интенсивность пуассоновского процесса, τ — длина интервала. Другими словами, по определению функции распределения вероятность того, что случайный интервал окажется меньше интервала τ равна $F(\tau)$. Функция плотности имеет вид:

$$p(\tau) = \frac{dF(\tau)}{d\tau} = \lambda e^{-\lambda\tau}.$$

В рассматриваемой нами задаче обозначим через τ_0 длину интервала между моментом интервью и последним эпизодом, τ_1 — между последним и предпоследним эпизодами, τ_2 — между предпоследним и третьим с конца, T — по-прежнему, между моментом интервью и третьим с конца эпизодом. Чтобы учесть все сообщённые респондентом интервалы, воспользуемся принципом максимального правдоподобия в следующем смысле: найдём такое λ, для которого вероятность того, что все три эпизода τ_0, τ_1, τ_2 произошли, максимальна. Поскольку мы рассматриваем поведение как пуассоновский процесс, и, следовательно, длины интервалов независимы [11], то получим:

$$g(\lambda) = p(\tau_0)p(\tau_1)p(\tau_2) = \lambda^3 e^{-\lambda T},$$

$$\frac{dg(\lambda)}{d\lambda} = \lambda^2 e^{-\lambda T}(3 - \lambda T) = 0,$$

$$\lambda = \frac{3}{T}. \qquad (1)$$

С одной стороны, такой подход приводит нас к уже полученному ранее результату, а значит, не решает обнаруженных проблем. С другой же стороны, такой подход показывает, где мы делаем слишком сильное предположение, и таким образом подсказывает методы устранения выявленных несоответствий.

В приведенных выше рассуждениях мы неявно предполагаем, что в день интервью также произошел эпизод поведения (поскольку мы рассматриваем интервал τ_0 как интервал между эпизодами). Отметим, что сходное предположение мы делали и при вычислении интенсивности первым описанным способом, а именно, мы считали, что за интервал длины T происходило ровно три события, однако, если ближайший эпизод поведения после интервью произойдёт ещё не скоро, то интервалы вида $(t_0 + \varepsilon, t_0 + T + \varepsilon)$ (являющиеся интервалами длины T) вносят значительный вклад в интенсивность, не учтённый при наших

оценках, и этот вклад является нулевым как раз при условии, что в день интервью произошёл ещё один эпизод.

Таким образом, нам известны достоверно только два интервала между эпизодами поведения, и в этом смысле более достоверной оценкой интенсивности будет $2/(\tau_1+\tau_2)$. Полученная же выше оценка (как видно из второго же подхода) является оценкой сверху для интенсивности. Обозначим в связи с этим

$$\lambda_{\text{rel}} = \frac{2}{\tau_1+\tau_2}, \tag{2}$$

$$\lambda_{\max} = \frac{3}{T}. \tag{3}$$

Как легко убедиться, λ_{rel} также не является удовлетворительной оценкой интенсивности. Скажем, первый респондент дал ответы 1 день назад, 2 и 3, а второй — 20, 21 и 22. В обоих случаях λ_{rel} окажется одинаковой, однако же для второго респондента она превысит $\lambda_{\max}$, а значит, как интенсивность не будет выдерживать ни какой критики. Тем самым, при оценке интенсивности необходимо учитывать величину τ_0, и тривиальный способ это сделать — предположить, что в момент интервью также происходит эпизод поведения. Однако возможны и другие подходы, некоторые из которых описаны ниже.

3. Простейшие дополнительные условия. В этой части мы введём некоторые дополнительные условия, позволяющие скорректировать оценку интенсивности для приведённых в п. 2 двух модельных примеров её неточности.

В первом приближении поставим целью добиться следующего поведения интенсивности λ: если $\lambda_{\text{rel}} < \lambda_{\max}$, то в качестве интенсивности положим λ_{rel}, иначе — $\lambda_{\max}$ (см. (2), (3)).

$$\lambda = \lambda_1 = \begin{cases} \lambda_{\text{rel}}, \text{ если } \lambda_{\text{rel}} < \lambda_{\max}; \\ \lambda_{\max}, \text{ иначе.} \end{cases} \tag{4}$$

Для этого наложим на исходные данные такие условия: будем считать, что величина $\Theta = \tau_0 + \theta$ не является независимой, и, более того, определяется следующим соотношением:

$$\Theta = \begin{cases} \dfrac{\tau_1 + \tau_2}{2}, \text{ если } \dfrac{\tau_1 + \tau_2}{2} > \tau_0; \\ \tau_0, \text{ иначе.} \end{cases}$$

Рассмотренное поведение интенсивности отвечает следующему интуитивному пониманию ситуации. Предположим, что со времени последнего эпизода до момента интервью прошло меньше времени, чем можно в среднем ожидать, исходя из двух достоверных величин интервалов τ_1 и τ_2. Тогда разумно считать, что τ_0 не привносит никакой дополнительной информации к оценке интенсивности, так как мы могли с равной вероятностью провести интервью как в начале длинного, так и в конце короткого интервалов. Но если между последним эпизодом и моментом интервью прошло больше времени, чем можно ожидать, исходя из τ_1 и τ_2, то разумно считать, что следующий эпизод поведения произойдет вскоре после момента интервью.

При наложенных ограничениях мы имеем только три случайные величины: τ_1, τ_2 и τ_0. Будем полагать, что интервью может выпасть на любой момент интервала Θ. Таким образом, для нахождения интенсивности применим метод максимального правдоподобия к следующей функции:

$$g(\lambda) = p(\tau_0 + \theta)\, p(\tau_1)\, p(\tau_2) = \lambda^3 e^{-\lambda(T+\theta)};$$

$$\frac{dg(\lambda)}{d\lambda} = \lambda^2 e^{-\lambda T}(3 - \lambda(T+\theta)) = 0;$$

$$\lambda = \frac{3}{T+\theta}.$$

$$\lambda = \begin{cases} \dfrac{3}{T + \dfrac{\tau_1 + \tau_2}{2} - \tau_0}, \text{ если } \dfrac{\tau_1 + \tau_2}{2} > \tau_0; \\ \dfrac{3}{T}, \text{ иначе.} \end{cases}$$

$$\lambda = \begin{cases} \lambda_{\text{rel}}, \text{ если } \dfrac{\tau_1 + \tau_2}{2} > \tau_0; \\ \lambda_{\max}, \text{ иначе.} \end{cases}$$

Осталось заметить, что $\lambda_{\text{rel}} < \lambda_{\max}$ как раз при $\dfrac{\tau_1 + \tau_2}{2} > \tau_0$.

К минусам рассмотренного подхода можно отнести следующее. Мы оцениваем неизвестную нам величину $\tau_0 + \theta$ исходя из двух известных τ_1 и τ_2, что, хотя и согласуется с нашей интуицией, противоречит рассмотрению поведения респондента как пуассоновского процесса, где величины интервалов независимы.

4. Разные подходы к учёту τ_0. Для оценки интенсивности мы применяли принцип максимального правдоподобия для длин интервалов между эпизодами. Проще говоря, мы искали такую величину λ, чтобы вероятность того, что произошли три эпизода поведения с определенными длинами интервалов между соседними из них, была наибольшей. Отметим, что на самом деле случайных событий происходит не три, а четыре (одно из которых связано с будущим), а именно, интервалы между эпизодами, интервал между последним эпизодом и моментом интервью и интервал между моментом интервью и следующим эпизодом. Последнюю величину, когда это необходимо, можно рассматривать несколько другим образом: как момент, в который интервал между последним эпизодом и следующим за ним был прерван интервью. Таким образом, из необходимых для оценки интенсивности четырёх случайных величин нам известно только три: τ_0, τ_1 и τ_2. В связи с этим, применить предложенный выше подход напрямую не удастся, но можно предложить некоторые его вариации. Обозначим (рис. 1) через θ неизвестную нам четвёртую из случайных величин — интервал между моментом интервью и воображаемым следующим эпизодом поведения. Сразу отметим, что в некоторых анкетах возможно спрашивать, когда планируется следующий эпизод поведения и получать, возможно, более точную оценку, чем все предложенные ниже. Однако же, такой подход не всегда возможен.

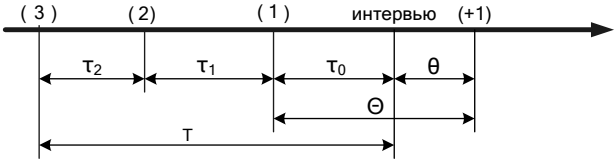

Рис. 1. Обозначения интервалов.

Первый способ справиться с возникающими трудностями — это сделать дополнительные предположения, связывающие известные нам длины эпизодов и неизвестные λ и θ. Тогда можно будет выразить, скажем, вторую неизвестную через первую и снова работать с одной

переменной. Такой подход — в некоторой степени творческий, и дополнительные условия следует выбирать, имея в виду детали конкретного исследования. Пример таких условий, позволяющих справиться с выявленными ранее недостатками оценки (1), был описан в п. 3, но более подробное его рассмотрение выходит за рамки данной работы.

Другой путь — попытаться, применив метод максимального правдоподобия для произошедших случайных событий, найти и λ, и θ. Все события можно считать независимыми, тогда наша задача сводится к нахождению максимума произведения функций плотности этих случайных величин как функции двух переменных. Функции плотности вероятности для длин первых двух эпизодов нам известны, но осталось найти функцию плотности для третьего. Она равна $\lambda e^{-\lambda(\tau_0+\theta)}$, но величина θ нам неизвестна.

Пусть нам известна величина эпизода $\Theta = \tau_0 + \theta$ и мы знаем вероятность $P(\tau_0)$ того, что интервью будет проведено в момент τ_0 со дня последнего эпизода. $P(\tau_0)$ зависит от величины интервала Θ, то есть, мы можем сказать, что $P(\tau_0) = P_\Theta(\tau_0) = Pr[\tau_0|\Theta]$. Мы же ищем вероятность Θ при условии, что случилось τ_0. Тогда по теореме Байеса получаем:

$$\Pr[\Theta|\tau_0] = \frac{\Pr[\tau_0|\Theta]\cdot Pr[\Theta]}{\Pr[\tau_0]}$$

Вероятность $\Pr[\Theta]$ равна $\lambda e^{-\lambda\Theta}$, так как это интервал между событиями пуассоновского процесса. Вероятность $\Pr[\tau_0]$ не зависит от Θ, но зависит от λ. Обозначим $\rho(\lambda, \tau_0) = \frac{1}{\Pr[\tau_0]}$. Тогда

$$\Pr[\Theta|\tau_0] = p_\Theta(\tau_0)\cdot\lambda e^{-\lambda\Theta}\cdot\rho(\lambda, \tau_0) \qquad (5)$$

Остаются открытыми два вопроса — какую выбирать функцию для $P(\tau_0)$ и как вычислять функцию $\rho(\lambda, \tau_0)$. Ответ на первый вопрос зависит от конкретного исследования, но всё же в большинстве случаев естественно считать, что интервью может с равной вероятностью выпасть на любой момент интервала между эпизодами и в качестве $P(\tau_0)$ брать постоянную функцию (отвечающую равномерному распределению), т.е., равную $\frac{1}{\Theta}$. На второй вопрос можно ответить при-

ближённо: функцию $\rho(\lambda, \tau_0)$ в небольшой окрестности можно интерполировать многочленом высокой степени с очень большой точностью.

Этот подход к уточнению оценки интенсивности поведения подробнее рассматривается в п. 5.

Соответствующий алгоритм приближенного вычисления $\rho(\lambda, \tau_0)$ реализован на языке Java в виде программного приложения (рис. 2.). Если считать, что фиксирован некоторый отрезок и точки на нём равномерно распределены, не составляет труда вычислить вероятность заданной точки. К этой вероятности можно относиться как к условной, поскольку она зависит от величины фиксированного отрезка. Приложение приближённо вычисляет вероятность заданной точки в предположении, что отрезки распределены экспоненциально (см. формулу (5) и её описание в статье). Пользователю предоставляется возможность регулировать параметры, по которым строится приближение. Таким образом, приложение можно рассматривать как иллюстрацию и подтверждение многих ссылок, данных в статье на использование приближенных вычислений.

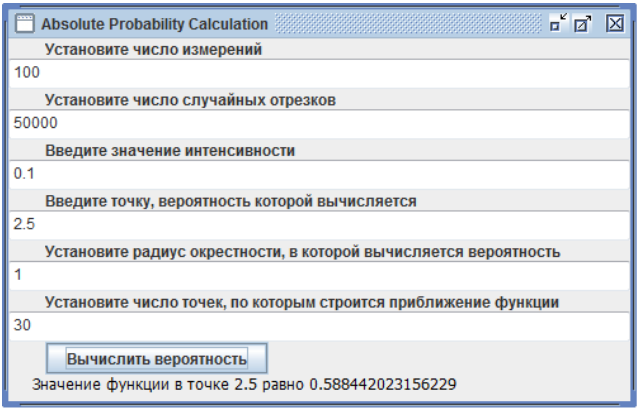

Рис. 2. Вычисление значения функции $\rho(\lambda, \tau_0)$.

Описание интерфейса. В первой графе пользователю предлагается ввести число точек, которые будут равномерно распределятся на интервалах случайной длины. Во второй графе можно задать количество случайных интервалов, по которым будет строится приближение. Ниже можно установить значение интенсивности экспоненциального распределения, которому подчинены моделируемые отрезки, и значе-

ние точки, абсолютная вероятность которой будет вычисляться. Далее предлагается ввести радиус окрестности, в которой будет вычисляться функция вероятности и количество точек, по которым она будет строиться.

Описание алгоритма. Программа моделирует заданное количество отрезков, длины которых подчиняются экспоненциальному закону распределения с заданной интенсивностью. На каждом из них она равномерно распределяет точки (количество которых также может регулироваться пользователем). Затем программа приближённо считает функцию распределения для заданного количества точек из окрестности заданного радиуса (то есть, вероятность того, что выпавшая случайная точка будет меньше данной), и решает для этих точек интерполяционную задачу. После чего, она вычисляет производную полученного многочлена и считает её значение в точке, вероятность которой необходимо найти.

5. Уточнение оценки интенсивности поведения. Будем искать такие λ и θ, чтобы вероятность того, что произойдут все три случайных события τ_2, τ_1 и мнимый эпизод Θ, была максимальной. Для этого найдём максимум функции двух переменных

$$g(\lambda,\theta) = \lambda^3 e^{-\lambda(T+\theta)} \cdot \frac{1}{\tau_0 + \theta} \cdot \rho(\lambda, \tau_0)$$

(см. (5) и п. 4). Эта функция монотонно убывает при θ, стремящемся к бесконечности, поэтому её максимум при ограничении $\theta = 0$ (ежели таковой существует) будет максимумом функции и на всей области задания. Тем самым, достаточно искать максимум функции

$$h(\lambda) = g(\lambda,0) = \lambda^3 e^{-\lambda T} \cdot \frac{\rho(\lambda,\tau_0)}{\tau_0}.$$

Эта функция принимает положительные значения для некоторых точек и, как показывают компьютерные вычисления, стремится к нулю при λ стремящемся к бесконечности. Поэтому можно выбрать достаточно большой отрезок, вне которого функция меньше некоторого своего положительного значения, а тогда максимум h на этом отрезке будет максимумом на всей области задания. А так как h неотрицательна и $h(0) = 0$, достаточно проверить на максимум все локальные экстремумы функции h.

Для этого решим уравнение:

$$h'(\lambda) = 3\lambda^2 e^{-\lambda T} \cdot \frac{\rho(\lambda, \tau_0)}{\tau_0} + \lambda^3 \cdot (-T) e^{-\lambda T} \cdot \frac{\rho(\lambda, \tau_0)}{\tau_0} + \lambda^3 e^{-\lambda T} \cdot \frac{\rho'(\lambda)}{\tau_0} =$$

$$= \frac{\lambda^2 e^{-\lambda T}}{\tau_0} \cdot \left(3\rho(\lambda, \tau_0) - \lambda T \cdot \rho(\lambda, \tau_0) + \lambda \cdot \rho'_\lambda(\lambda, \tau_0)\right) = 0,$$

$$\left[\begin{array}{l} \lambda = 0, \\ \lambda = \dfrac{3\rho(\lambda, \tau_0)}{T \cdot \rho(\lambda, \tau_0) - \rho'(\lambda, \tau_0)}. \end{array}\right.$$

Поскольку $\lambda = 0$ не является точкой максимума функции h, искомая интенсивность является решением второго уравнения. Его можно переписать следующим образом:

$$\lambda = \lambda_2 = \frac{3}{T - \dfrac{\rho'(\lambda, \tau_0)}{\rho(\lambda, \tau_0)}} \qquad (6)$$

Уравнение (6) приближённо решается с погрешностью не больше 0.001 за приемлемое время. Например, для интервалов 1, 10, 10, $\lambda_2 \approx 0.131$ (а $\lambda_1 = 0.1$). Таким образом, если в нашем исследовании нельзя считать, что эпизоды происходят в среднем, через равные промежутки времени (как в случае с походами в магазин), то нельзя выкинуть из рассмотрения даже очень маленький τ_0. Для другого нашего примера с интервалами 20, 1, 1, $\lambda_2 \approx 0.126$ (а $\lambda_1 = \frac{3}{22} \approx 0.136$). Таким образом, при отсутствии дополнительных условий, $\lambda_{\max}$ не является удовлетворительной оценкой интенсивности даже в тех случаях, когда это согласуется с нашей интуицией. Это происходит из-за того, что интервалы распределены не равномерно, как мы подсознательно считаем, а экспоненциально.

Алгоритм, вычисляющий значение интенсивности исходя из тождества (6) в статье, реализован в программном приложении на языке Java (рис. 3).

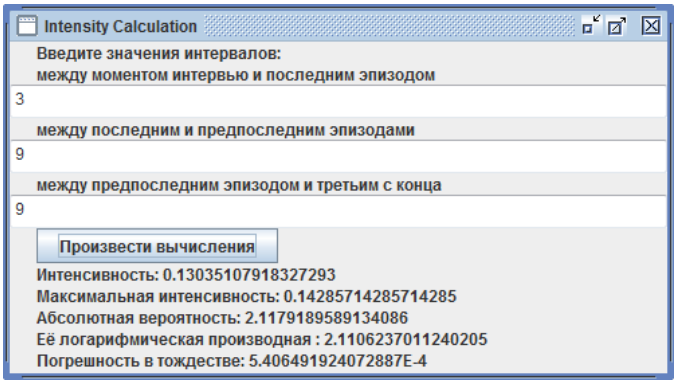

Рис. 3. Вычисление уточненной оценки интенсивности.

Описание интерфейса. В три графы пользователю предлагается ввести величины интервалов между эпизодами поведения, то есть, данные, взятые из анкеты респондента. Помимо интенсивности, программа вычислит максимальную интенсивность (см. формулу (3)), значение которой обычно брали в качестве интенсивности, также выведет данные промежуточных вычислений: вероятность интервала между последним эпизодом и моментом интервью (см. формулу (5) и описание приближенного вычисления $\rho(\lambda, \tau_0)$) и логарифмическую производную этой вероятности, как функции, зависящей от интенсивности. Также программа считает величину погрешности в формуле (6) для полученного значения интенсивности.

Описание алгоритма. Приложение итеративно вычисляет значение интенсивности: сначала в качестве λ в формуле (6) берётся значение максимальной интенсивности и вычисляется правая часть формулы (6). Затем в качестве λ используется полученное значение и шаг повторяется до тех пор, пока погрешность в тождестве не станет достаточно мала.

Алгоритм вычисления значения функции $\rho(\lambda, \tau_0)$ описан выше. Для вычисления её производной как функции от λ вновь приближённо вычисляется её значения в точках из окрестности текущего значения λ, а затем по ним строится многочлен и считается его производная. Имея эти два значения, арифметическими операциями приложение вычисляет логарифмическую производную от вероятности и правую часть формулы (6).

6. Заключение. В разных исследованиях правильный учёт интервала между последним эпизодом и моментом интервью может разли-

чаться, и универсального метода не существует, как мы убедились уже на примере λ_1 и λ_2 выше (см. (4),(6)). В каждом конкретном исследовании нужно делать разумные дополнительные предположения, которые часто совершенно не применимы в любом другом, потому что когда события не являются совершенно случайными, а скованны определёнными условиями, то эти условия должны находить отражение и в математической модели. В качестве дальнейших шагов в этом направлении можно рассмотреть другие распределения момента интервью внутри эпизода или же заняться оптимизацией приближённого решения уравнения.

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.

Литература

1. *Суворова А.В., Тулупьев А.Л., Пащенко А.Е., Тулупьева Т.В., Красносельских Т.В.* Анализ гранулярных данных и знаний в задачах исследования социально значимых видов поведения // Компьютерные инструменты в образовании. №4. 2010. С. 30–38.
2. *Хованов Н.В.* Анализ и синтез показателей при информационном дефиците. СПб.: Изд-во СПбГУ, 1996. 196 с.
3. *Тулупьева Т.В., Пащенко А.Е., Тулупьев А.Л., Красносельских Т.В., Казакова О.С.* Модели ВИЧ-рискованного поведения в контексте психологической защиты и других адаптивных стилей. СПб.: Наука, 2008. 140 с.
4. *Пащенко А. Е.* Идентификация интенсивности пуассоновского процесса, моделирующего поведение респондента, в условиях дефицита информации // Информационно-измерительные и управляющие системы. 2009. № 4. т. 7, С. 45–48.
5. *Пащенко А.Е., Суворова А.В.* Программный комплекс для экспертного оценивания интенсивности поведения респондента в условиях дефицита информации // Интегрированные модели, мягкие вычисления, вероятностные системы и комплексы программ в искусственном интеллекте. Научно-практическая конференция студентов, аспирантов, молодых ученых и специалистов (Коломна, 26–27 мая 2009 г.). Научные доклады. В 2-х т. Т. 2. М.: Физматлит, 2009. С. 220–241.
6. *Пащенко А. Е., Тулупьев А. Л., Николенко С. И.* Моделирование заражения ВИЧ-инфекцией на основе данных о последних эпизодах рискованного поведения // Известия высших учебных заведений: Приборостроение. 2006. №8. 33–34 с.
7. *Пащенко А.Е., Тулупьев А.Л., Тулупьева Т.В., Красносельских Т.В., Соколовский Е.В.* Косвенная оценка вероятности заражения ВИЧ-инфекцией на основе данных о последних эпизодах рискованного поведения // Здравоохранение Российской Федерации. 2010. № 2. С. 32–35.

8. *Пащенко А.Е., Тулупьева Т.В., Суворова А.В., Тулупьев А.Л.* Интеллектуальная система для экспертного оценивания интенсивности рискованного поведения в условиях информационного дефицита // Региональная информатика-2008 (РИ-2008). XI Санкт-Петербургская международная конференция. Санкт-Петербург, 22–24 октября, 2008 г.: Материалы конференции / СПОИСУ. СПб., 2009. С. 285–291.
9. *Тулупьева Т.В., Тулупьев А.Л., Пащенко А.Е.* Оценка интенсивности поведения респондента в условиях информационного дефицита // Труды СПИИРАН. Вып. 7. СПб.: Наука, 2008. С. 239–254.
10. *Тулупьева Т.В., Пащенко А.Е., Тулупьев А.Л., Голянич В.М.* Модели ВИЧ-рискованного поведения в контексте психологической защиты и адаптации // Вестник СПбГУ. Серия 12. Серия 12. Вып. 1. С. 95–104.
11. *Розанов Ю.А.* Случайные процессы (краткий курс). М.: Главная редакция физико-математической литературы издательства «Наука», 1971. 288 с.
12. *Крамер Г.* Математические методы статистики. М.: Мир, 1975. 648 с.

Лавренов Андрей Валентинович — студент математико-механического факультета Санкт-Петербургского государственного университета (СПбГУ). Область научных интересов: математическая статистика, теория вероятности. Число научных публикаций — 1. vedrfiolnir@gmail.com; СПИИРАН, 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т. +7(812)328-3337, факс +7(812)328-4450. Научный руководитель — А.Л. Тулупьев.

Lavrenov Andrey Valentinovich —student, Faculty of Mathematics and Mechanics of St. Petersburg State University (SPbSU). Research interests: mathematical statistics, probability theory. The number of publications — 1. vedrfiolnir@gmail.com; SPIIRAS, 39, 14th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-3337, fax +7(812)328-4450. Scientific advisor — A.L. Tulupiev.

Суворова Алена Владимировна — младший научный сотрудник 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, г. Санкт-Петербург, 199178, РФ; р.т. +7(812)328-3337, факс +7(812)328-4450. Научный руководитель — 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 — A.L. Tulupiev.

Пащенко Антон Евгеньевич — младший научный сотрудник 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.

Тулупьев Александр Львович — д.ф.-м.н., доцент; заведующего лабораторией теоретических и междисциплинарных проблем информатики СПИИРАН, доцент кафедры информатики математико-механического факультета С.-Петербургского государственного университета (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 — 220. ALT@iias.spb.su, www.tulupyev.spb.ru; СПИИРАН, 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т. +7(812)328-3337, факс +7(812)328-4450.

Tulupyev Alexander Lvovich — PhD in Computer Science, Dr. of Sc.. Associate Professor; Head of Theoretical and Interdisciplinary Computer Science Laboratory, SPIIRAS, Associate Professor of Computer Science Department, SPbSU. Research area: uncertain data and knowledge representation and processing, mathematics and computer science applications in socio-cultural studies, biostatistics, simulation, and mathematical modeling applications in epidemiology, data intensive software systems development technology. Number of publications — 220. ALT@iias.spb.su, www.tulupyev.spb.ru; SPIIRAS, 14-th line V.O., 39, St. Petersburg, 199178, Russia; office phone +7(812)328-3337, fax +7(812)328-4450.

Рекомендовано ТиМПИ СПИИРАН, зав. лаб. А.Л. Тулупьев, д.ф.-м.н., доцент.
Статья поступила в редакцию 06.12.2010.

РЕФЕРАТ

Лавренов А.В., Суворова А.В., Пащенко А.Е. **Особенности обработки данных и знаний об эпизодах социально-значимого поведения в окрестности интервью.**

Задачи оценивания интенсивности социально-значимого поведения респондентов по их самоотчетам об эпизодах поведения возникают во многих отраслях социологических, психологических, маркетинговых исследований.

Заметим, что ответы респондента на вопросы о последних эпизодах характеризуются стабильностью воспроизведения. Однако ограниченное число и неточность, недоопределенность, нечеткость естественно-языковых формулировок ответов (то есть наблюдаемый сверхкороткий временной ряд) не позволяют напрямую использовать известные методы из теории массового обслуживания для оценки интенсивности поведения, поэтому возникает необходимость в предложении новых математических моделей.

Поведение рассматривается как случайный процесс некоторого класса. При этом встают вопросы о том, какой процесс лучше описывает поведение, как меняются параметры этого процесса, как осуществляется обработка неполных исходных данных. Цель данной статьи — описать проблемы, возникающие при анализе данных о последних эпизодах социально-значимого поведения, и предложить некоторые пути их решения.

Используемые в настоящее время методы нахождения интенсивности, несмотря на логичные и правомерные мотивирующие соображения, имеют ряд недостатков из-за некоторых оставленных без внимания деталей. Одна из таких деталей связана с тем, что в момент интервью принимается случившимся ещё один эпизод поведения. В статье обсуждается уместность дополнительных предположений для решения данной задачи, приводится её решения как с некоторыми предположениями, так и без них. Для решения задачи без дополнительных предположений используются компьютерные вычисления, точность которых заведомо больше точности используемых в настоящее время оценок интенсивности. В заключение, демонстрируется, что в каждом конкретном исследовании следует делать разумные дополнительные предположения, потому что когда события не являются совершенно случайными, а скованны определёнными условиями, то эти условия должны находить отражение и в математической модели.

SUMMARY

Lavrenov A.V., Suvorova A.V., Paschenko A.E. **Processing issues of data and knowledge about socially significant behavior' episodes near the interview moment.**

We are faced with the problem of socially significant behavior rate estimate on the base of respondents' self-reports about their behavior episodes in many fields of sociological, psychological and marketing research.

Note that respondents stably reproduce their answers about last behavior episodes. Limited number of natural language responses' forms and their fuzziness, uncertainty, imprecision (or super-short time series) do not allow direct use of known methods of queuing theory for behavior rate estimate, that's why we need to propose new mathematical behavior models.

The behavior is described by a random process. And we have to find such answers as what random process type is the best for describing behavior, how the parameters of this process change, how incomplete data is handled. This paper represents issues arise in data about the socially significant behavior last episodes analysis. Several approaches to the issues elimination are proposed.

Despite logical and valid motivating arguments ways of finding the rate used nowadays have several drawbacks because of some unnoticed details. One of such details is associated with considering one more behavior episode takes place at the time of interview. Appropriateness of extra arguments for solution of the problem is developed in this paper, solutions of the problem are provided, both with and without any arguments. Computer calculations which are a priory more exact than used nowadays intensity estimation are needed for solution a problem without extra arguments. In conclusion, it's shown that reasonable extra arguments should be taken a place during each certain research, because in case of events being not completely occasional and happening within the confines of certain conditions, such conditions must have an effect on a mathematical model.