УДК 311.2 + 616-036.22

Д. Зельтерман, А.Л. Тулупьев, А.В. Суворова, А.Е. Пащенко,
В.Ф. Мусина, Т.В. Тулупьева, Т.В. Красносельских,
Л. Гро, Р. Хаймер

ОБРАБОТКА СИСТЕМАТИЧЕСКОЙ ОШИБКИ, СВЯЗАННОЙ С ДЛИНОЙ ВРЕМЕННЫХ ИНТЕРВАЛОВ МЕЖДУ ИНТЕРВЬЮ И ПОСЛЕДНИМ ЭПИЗОДОМ В ГАММА-ПУАССОНОВСКОЙ МОДЕЛИ ПОВЕДЕНИЯ

Зельтерман Д., Тулупьев А.Л., Суворова А.В., Пащенко А.Е., Мусина В.Ф., Тулупьева Т.В., Красносельских Т.В., Гро Л., Хаймер Р. **Обработка систематической ошибки, связанной с длиной временных интервалов между интервью и последним эпизодом в гамма-пауссоновской модели поведения.**

Аннотация. Рассматривается подход к оцениванию интенсивности и производных параметров поведения респондентов по сведениям о последнем эпизоде их поведения. В качестве модели поведения предложен гамма-пуассоновский процесс, описаны его характеристики, а также различные варианты его параметризации. Разработан метод, позволяющий обработать систематическую ошибку, возникающую из-за неявного предположения, что момент интервью является эпизодом поведения. В работе также предложены способы обработки исходных данных, характеризующихся гранулярностью.

Ключевые слова: оценка интенсивности, модели поведения, последние эпизоды, неопределенность, гранулярность.

Zelterman D., Tulupyev A.L., Suvorova A.V., Paschenko A.E., Musina V.F., Tulupyeva T.V., Krasnoselskikh T.V., Grau L., Heimer R. **Processing length bias of time intervals between the last episode and the interview in Gamma-Poisson models of behavior.**

Abstract. We develop a technique for quantitative estimates of respondents' behavior that uses respondents' answers about the time interval since the last episode. The paper provides the block of questions and formalized set of answers to be used in a questionnaire as well as the mathematical approach for data processing and making the estimates. The respondents' behavior mathematical model under discussion belongs to the class of generalized Gamma-Poisson stochastic process and takes into account the length bias inherent to the data collected from the respondents' answers about the last episodes of their behavior.

Keywords: rate estimate, behavior models, last episodes, uncertainty, granularity.

1. Введение. Во многих отраслях социологических, психологических, маркетинговых исследований возникают задачи оценивания интенсивности социально-значимого поведения респондентов [1]. Например, в настоящее время наиболее острой эпидемиологической проблемой является оценка риска передачи и приобретения такой опасной и неизлечимой инфекции как инфекция вирусом иммунодефицита человека (ВИЧ) в зависимости от особенностей инъекционного

и сексуального поведения индивида. Наиболее точно такой риск характеризуется инсиденс-показателем [2] — числом заразившихся за определенный период среди лиц, подвергавшихся риску заражения, отнесенным к человеко×месяцам наблюдения.

Для прямого измерения инсиденс-показателя требуется организовать когортное исследование, подразумевающее вовлечение, как правило, не менее 500 представителей группы риска и их медицинское и социальное сопровождение в течение значительного периода времени. Однократное проведение подобного когортного исследования обходится в полтора-два миллиона долларов. Такой уровень расходов делает затруднительным мониторинг инсиденс-показателя даже в странах с сильной экономикой. [2, 3] Требуется предложить математические модели, позволяющие выполнить более дешевые косвенные измерения инсиденс-показателя на основе ответов респондентов, составляющих выборку из группы риска. Один из таких способов опирается на модель Белла–Тревино [4]. Инсиденс-показатель можно оценить, зная индивидуальный риск заражения каждого отдельного респондента за заданный период времени. Модель Белла–Тревино увязывает оценку риска с числом эпизодов рискованного поведения. Число же эпизодов можно оценить, если, в свою очередь, известна оценка интенсивности рискованного поведения, рассмотренного как случайный процесс определенного класса [5].

В маркетинговых исследованиях выделение групп потребителей, существенно различающихся интенсивностью потребления продуктов, товаров или услуг, позволяет сосредоточить усилия на тех группах, которые многочисленны, но товар потребляют неинтенсивно. Выявление особенностей этих групп позволяет разработать стратегию, ведущую к существенному увеличению объема продаж. Следует отметить, что необходимые данные об интенсивности потребления невозможно получить из анализа продаж, то есть недостаточно изучить «чеки» — данные о состоявшихся продажах. Это дает возможность принять во внимание лишь те группы, которые и так уже покупают данный товар. Не исключено, что в таком случае из анализа выпадут многочисленные потенциальные потребители, которые ни разу еще не употребляли интересующие маркетологов товары или услуги и которые как раз и составляют «неосвоенную» нишу для продаж.

Приведем другой пример. При одном из самых распространенных заболеваний эндокринной системы — сахарном диабете, — в основе которого лежит относительный или абсолютный дефицит инсулина, важнейшее значение имеет строгое соблюдение диеты. В некоторых

случаях при диабете 2-го типа для компенсации нарушения углеводного обмена и прекращения прогрессирования заболевания достаточно лишь ограничить употребление легкоусвояемых углеводов и жиров. При 1-м типе сахарного диабета соблюдение диеты жизненно важно для больного, ее нарушение может привести к гипо- или гипергликемической коме, а иногда — к смерти. Лечащему врачу необходимо оценивать частоту и обстоятельства отклонения пациента от диеты, чтобы иметь возможность корректировать дозировку назначаемых пероральных сахароснижающих препаратов или инсулинов, а также понимать, какие вмешательства необходимо предпринять для повышения приверженности больного рекомендациям по соблюдению диеты. Таким образом, выбор тактики ведения больного сахарным диабетом во многом основывается на степени интенсивности отклонения от диеты.

Рассмотрим в качестве иллюстрации упрощенный пример. Предположим, что мы хотим оценить интенсивность потребления бананов. Как правило, бананы в розницу покупаются не поштучно, а гроздями, чаще всего в одну–две–три грозди. Таким образом, число бананов варьируется, варьируется и другие обстоятельства, влияющие на интервалы времени между покупками (другими словами, между эпизодами поведения). Например, если сегодня куплено больше бананов, то следующий раз за ними в магазин мы придем позже (рис. 1). Однако интенсивность потребления бананов в относительно короткие промежутки времени (1–3 месяца) не сильно варьируется, и ее можно оценить. Более того, интересно оценить интенсивность потребления бананов (и характеристики ее изменчивости) в течение долгого времени, когда удается учесть сезонные и ценовые факторы, а также особенности личности респондента, его социально-демографических и социально-экономических характеристик.

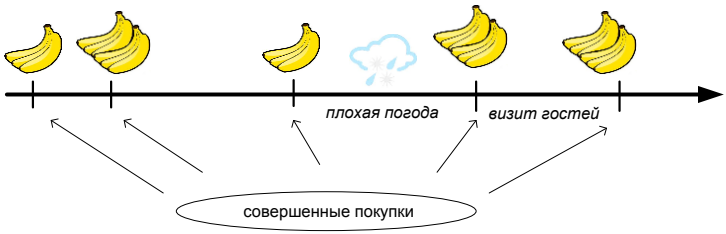


Рис. 1. К примеру о покупке бананов.

2. Подходы к оценке интенсивности. Отметим, что наиболее доступными исходными данными для анализа поведения выступают са-

моотчеты 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 [8, Appendix A], каждый из которых имеет недостатки [5–7]. Первый метод — прямые вопросы: «Сколько раз Вы *делали так* в течение последнего месяца (трех, шести, года)?». На такие вопросы респонденты обычно дают практически не соотносящиеся с реальностью ответы. Действительно, можно задать себе вопрос: «Сколько раз за последние полгода я покупал бананы?». Попытка ответа даже самому себе даст четкую картину незначительной достоверности такого ответа.

Второй метод — Лайкерт-шкалы — опросники, в которых используются качественные, а не количественные варианты: «Никогда», «Редко», «Иногда», «Часто», «Всегда» и подобные им возможности для ответа. Вопрос ставится легко, ответ тоже получить несложно, однако эти ответы не несут никаких полезных сведений относительно числа эпизодов: то, что «Часто» для одного человека, может быть «Редко» для другого, а то, что «Часто» в одном виде поведения, может быть «Редко» для другого вида поведения. Кроме того, «расстояние» между «Всегда» и «Очень часто» совершенно не обязательно совпадает с расстоянием между «Редко» и «Никогда». На практике шкалы арифметизируют, но за этой арифметизацией не стоит никакой достоверной гипотезы; получающиеся расчеты ситуацию с интенсивностью поведения не характеризуют вообще никак. Таким образом, возникает потребность в более адекватных источниках сведений о социально значимом поведении и методиках их обработки, которые сделают возможной более обоснованную оценку числа эпизодов.

Отметим, что использование «прямых» вопросов о числе эпизодов поведения респондента в заданный длительный промежуток времени, применение Лайкерт-шкал или категоризованных ответов является классическим приемом и упоминается в весьма авторитетных источниках (см. в частности, Appendix A в [8]). Вместе с тем стоит учитывать, что указанный опросный инструментарий разрабатывался без учета потребностей, которые возникли позже, например — получение количественных оценок интенсивности рискованного поведения и риска, с ним связанного в эпидемиологических исследованиях ВИЧ/СПИД.

Одной из возможных альтернатив представляется опрос респондента об одном или нескольких последних эпизодах его поведения [5–7, 9–12]. Заметим, что ответы респондента на такие вопросы о послед-

них 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.

Интенсивность поведения предлагается оценивать по данным о последних эпизодах рассматриваемого поведения или, другими словами, по известным длинам интервалов между последовательными эпизодами поведения. Так, в случае трех последних эпизодов известны значения длин интервалов между моментом интервью и последним эпизодом, между последним и предпоследним эпизодами и между предпоследним и третьим с конца. Отметим, что момент интервью не является эпизодом поведения, таким образом, рассмотрение первого из перечисленных интервалов как интервала между последовательными эпизодами приводит к возникновению систематической ошибки, способам учета которой посвящены последующие разделы.

3. Основные предположения о распределениях. В качестве модели поведения респондента рассмотрим обобщенный пуассоновский процесс с параметром λ, а именно: поведение представляет собой череду эпизодов, которые характеризуются интенсивностью λ, которая сама по себе является случайной величиной. Предположим, что интенсивность поведения имеет гамма-распределение:

$$\lambda \sim g(\lambda;\mu,\eta),$$

$$g(\lambda;\mu,\eta) = \Gamma(\eta)^{-1}(\eta/\mu)^{\eta}\lambda^{\eta-1}e^{-\lambda\frac{\eta}{\mu}},$$

где параметры $\lambda > 0$, $\mu > 0$, и $\eta > 0$. Чтобы каким-то образом оценить поведение параметра λ в предложенной модели, достаточно определить, как ведут себя величины $\mu > 0$, и $\eta > 0$.

В [13] было показано, что $\mathrm{E}\lambda = \mu$, $\mathrm{D}\lambda = \dfrac{\mu^2}{\eta}$.

Предположим, нас интересует интенсивность определенного поведения, и чтобы оценивать параметры, связанные с этим поведением, мы сформировали выборку. Перед нами стоит задача: по ней оценить параметры μ и η. Одними из наиболее доступных данных, связанных с поведением респондента, являются данные о последнем эпизоде поведения, а именно: нас интересует интервал между последним эпизодом и эпизодом интервью, обозначим соответствующую ему случайную величину за T. Заметим, что чем длиннее интервал между эпизодами, тем более вероятно, что момент интервью попадет в этот (более длинный) интервал. Чтобы отразить это наблюдение в дальнейших вычислениях, введем в распределение множитель t, соответствующий длине интервала между соседними эпизодами поведения.

$$T \sim Kt\int_0^\infty \lambda e^{-t\lambda} g\left(\lambda;\mu,\eta\right) d\lambda. \tag{1}$$

Подставим в эту формулу распределение параметра λ, вид которого мы предложили выше. Получим, что

$$Kt\int_0^\infty \lambda e^{-t\lambda} g\left(\lambda;\mu,\eta\right) d\lambda = Kt\int_0^\infty \lambda e^{-t\lambda} \Gamma(\eta)^{-1} (\eta/\mu)^{\eta} \lambda^{\eta-1} e^{-\lambda\frac{\eta}{\mu}} d\lambda =$$

$$= K_1 t\int_0^\infty \lambda e^{-t\lambda} \lambda^{\eta-1} e^{-\lambda\frac{\eta}{\mu}} d\lambda = K_1 t\int_0^\infty \lambda^{(\eta+1)-1} e^{-\lambda\left(\frac{\eta}{\mu}+t\right)} d\lambda = \ldots$$

Сделаем замену переменных:

$$\dot{\eta} = \eta + 1, \qquad \frac{\dot{\eta}}{\dot{\mu}} = \frac{\eta}{\mu} + t;$$

$$\eta = \dot{\eta} - 1, \qquad \dot{\mu} = \frac{(\eta+1)\mu}{\eta + \mu t}.$$

$$\ldots = K_1 t\int_0^\infty \underbrace{\lambda^{\dot{\eta}-1} e^{-\lambda\frac{\dot{\eta}}{\dot{\mu}}}}_{\text{new } \dot{C}^{-1}\cdot\text{p.d.f.}} d\lambda = K_1 t \frac{\Gamma(\dot{\eta})}{(\dot{\eta}/\dot{\mu})^{\dot{\eta}}} = K_2 t \frac{1}{\left(\frac{\eta}{\mu}+t\right)^{\eta+1}} = \underbrace{K_2 \mu^{\eta+1}}_{C} \frac{t}{(\eta+\mu t)^{\eta+1}}.$$

Таким образом, исходная формула (1) преобразовывается к виду:

$$T \sim C \cdot \frac{t}{(\eta + \mu t)^{\eta+1}}, \qquad (2)$$

$$C = \mu^2 \eta^{\eta} (\eta - 1).$$

Моменты случайной величины T выражаются следующим образом (отметим, что для существования моментов необходимо потребовать $\eta > 3$):

$$\mathrm{E}T = C \int_0^{\infty} \frac{t^2}{(\eta + \mu t)^{\eta+1}}\, dt = 2 \cdot \frac{\eta^2}{\mu} \cdot \frac{\eta^2 - 3\eta + 1}{\eta^2 - 2\eta},$$

$$\mathrm{E}T^2 = C \int_0^{\infty} \frac{t^3}{(\eta + \mu t)^{\eta+1}} dt = \frac{\eta^4}{\mu^2} \cdot \frac{\eta - 1}{\eta} \cdot \left(\frac{1}{3-\eta} - \frac{3}{2-\eta} + \frac{3}{1-\eta} - \frac{1}{\eta} \right),$$

$$\mathrm{D}T = \mathrm{E}T^2 - \left(\mathrm{E}T\right)^2.$$

4. Связь с простым бета-распределением. Заметим, что полученное распределение случайной величины T (2) может быть классифицировано как бета-простое (beta-prime) распределение [14–17], плотность которого выражается формулой

$$\mathrm{BP}(x; \alpha, \beta) = \frac{x^{\alpha-1}(1+x)^{-\alpha-\beta}}{\mathrm{B}(\alpha, \beta)} = \frac{1}{\mathrm{B}(\alpha, \beta)} \cdot \frac{x^{\alpha-1}}{(1+x)^{\alpha+\beta}}, \qquad (3)$$

где $\mathrm{B}(\alpha, \beta)$ — бета-функция:

$$\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx = \int_0^{\infty} \frac{x^{\alpha-1}}{(1-x)^{\alpha+\beta}} dx;$$

В работах [14–17] отражены различные примеры применения такого распределения. В частности, в статье [16] подробно описаны взаимосвязи бета-распределений различных типов с другими распределениями и введено обобщение: бета-распределение с пятью параметрами. Последнее в свою очередь позволяет ввести в рассмотрение экспоненциальное обобщенное бета-распределение, включающее в себя логистическое, экспоненциальное и нормальное распределения в качестве частных случаев. Это семейство используется для построения частично адаптивных оценок в эконометрических моделях со скошенными и островершинными распределениями ошибки (т.е. отличных от нормального распределения). В работе уделяется внимание вопросам гибкости распределений с различным количеством параметров в усло-

виях различных моделей и возможным методам оценок их параметров. В качестве примера применения приводится задача поиска распределений, наилучшим образом описывающих номинальный семейный доход в 1985 году. Кроме того, рассматривается модель рынка и для неё предлагается оптимальное распределение ошибки.

Для преобразования выражения (2) к виду (3), производится замена переменной:

$$T \overset{d}{=} \frac{\eta}{\mu} \cdot S, \tag{4}$$

$$S \sim \mathrm{BP}(2, \eta - 1). \tag{5}$$

Отметим, что введенная случайная величина S не зависит от параметра μ, который участвует только при переходе от T к S в уравнении (4). Вычислим моменты случайной величины S. Для случайно величины, распределенной по закону бета-простого распределения $\mathrm{B}(\alpha, \beta)$ [18], они вычисляются следующим образом:

- если $\beta > 1$, то математическое ожидание случайной величины S равно $\frac{\alpha}{\beta - 1}$,
- если $\beta > 2$, то дисперсия случайной величины S равна $\frac{\alpha(\alpha + \beta - 1)}{(\beta - 2)(\beta - 1)^2}$.

Применительно к рассматриваемому случаю для $\eta > 2$ и $\eta > 3$ соответственно получим:

$$\mathrm{E}S = \frac{2}{\eta - 2}; \ \mathrm{D}S = \frac{2\eta}{(\eta - 3)(\eta - 2)^2};$$

и, следовательно, в терминах исходной случайной величины T:

$$\mathrm{E}T = \frac{2}{\mu} \cdot \frac{\eta}{\eta - 2}; \ \mathrm{D}T = \frac{2}{\mu^2} \cdot \frac{\eta^3}{(\eta - 3)(\eta - 2)^2}.$$

Заметим также, что если $\eta \to \infty$, то моменты стремятся к вполне определенному пределу, а именно: $\mathrm{E}T = \frac{2}{\mu}$ и $\mathrm{D}T = \frac{2}{\mu^2}$. Формально этот факт соответствует характеристикам параметра λ в случае, когда

интенсивность поведения равна постоянна (тогда модель поведения представляется пуассоновским , а параметр λ — константой).

Уравнения (4–5) допускают переход к более удобной для дальнейших исследований системе параметров, такой, что $\mathrm{E}T = \varphi$, т. е. $\varphi = \frac{2}{\mu}\cdot\frac{\eta}{\eta-2}$.

Перепишем полученную формулу в терминах параметра μ и обратного к нему:

$$\mu = \frac{1}{\varphi}\cdot\frac{2\eta}{\eta-2},\ \frac{1}{\mu} = \varphi\cdot\frac{\eta-2}{2\eta}.$$

Таким образом, исходное распределение случайной величины T будет выглядеть так:

$$T \overset{d}{=} \varphi\cdot\frac{\eta-2}{2}\cdot S.$$

Соответственно моменты выражаются следующим образом:

$$\mathrm{E}T = \varphi,$$

$$\mathrm{D}T = \varphi^2\cdot\frac{\eta}{2(\eta-3)},$$

$$\lim_{\eta\to+\infty}\mathrm{D}T = \frac{\varphi^2}{2}.$$

5. Оценка методом максимального правдоподобия. Выпишем функцию правдоподобия для распределения T:

$$\Lambda(\sigma,\eta;t_1,...,t_n) = \sigma^{-n}\prod_{i=1}^{n}\frac{1}{\mathrm{B}(2,\eta-1)}\frac{t_i/\sigma}{(1+t_i/\sigma)^{\eta+1}},$$

где $\sigma = \frac{\eta}{\mu}$. Учитывая, что $\mathrm{B}(2,\eta-1) = \frac{1}{\eta(\eta-1)}$, получим:

$$\Lambda(\sigma,\eta;t_1,...,t_n) = \sigma^{-n}\eta^n(\eta-1)^n\prod_{i=1}^{n}\frac{t_i/\sigma}{(1+t_i/\sigma)^{\eta+1}},$$

$$L(\sigma,\eta;t_1,...,t_n) = \ln\Lambda(\sigma,\eta;t_1,...,t_n) =$$

$$= n\ln\eta + n\ln(\eta-1) - n\ln\sigma + \sum_{i=1}^{n}\ln t_i/\sigma - (\eta+1)\sum_{i=1}^{n}\ln(1+t_i/\sigma);$$

$$\begin{cases} \dfrac{\partial L(\sigma,\eta;t_1,...,t_n)}{\partial\sigma}=0 \\ \dfrac{\partial L(\sigma,\eta;t_1,...,t_n)}{\partial\eta}=0 \end{cases} \Rightarrow \begin{cases} \dfrac{2n}{\sigma}-\dfrac{\alpha+1}{\sigma}\sum\dfrac{t_i}{t_i+\sigma}=0 \\ \dfrac{n}{\eta-1}+\dfrac{n}{\eta}-\sum\limits_{i=1}^{n}\ln(1+t_i/\sigma)=0 \end{cases}.$$

Современные статистические пакеты позволят найти сочетания параметров σ и η, на которых достигается максимум. Однако из-за неточности исходных данных такие оценки придется искать несколько раз, а потом сворачивать их.

6. Возможные модификации модели. Для некоторых моделей поведения параметр t в уравнении (1) может оказаться растущим слишком быстро, так что в некоторых случаях может возникнуть потребность заменить его на некоторую неубывающую (или даже ограниченную неубывающую) функцию $\phi(t)$, то есть ввести в рассмотрение модели типа:

$$T \sim K\phi(t)\int_0^{\infty}\lambda e^{-t\lambda}g\left(\lambda;\mu,\eta\right)d\lambda.$$

Статистический критерий для того, чтобы принять гипотезу о том, что параметр является (или несколько параметров являются) детерминированными величинами или случайными величинами предложен в статье [19].

Также для построения различных моделей поведения, возможно рассмотрение иных законов распределения длины интервала между соседними эпизодами, а также иных законов для моделирования вариаций параметров указанных законов распределения длины интервала.

Отметим, что эпизоду может быть приписан количественный атрибут (например, риск, связанным именно с этим эпизодом; объем алкоголя, выпитый именно в этот эпизод; число/масса бананов, купленных в этот раз). Таким образом, представляется обоснованным рассмотрение моделей, учитывающих и такое дополнительное данное.

7. Опросный инструментарий и его программная реализация. Для получения необходимых для анализа статистических данных разработан опросный инструментарий, включающий закрытые вопросы о последних эпизодах употребления алкоголя и участия в незащищенных сексуальных контактах с постоянным(и) партнером(ами) и с непостоянным(и) партнером(ами). Блоки вопросов о каждом из указанных

видов поведения имеют одинаковую структуру, поэтому приведем часть анкеты, связанную с информацией об употреблении алкоголя:

---

1. Употребляли ли Вы когда-либо алкоголь?

Да = *1*

Нет = *0* →**переход к вопросу 8.**

2. Когда Вы <u>в последний раз</u> употребляли алкоголь *(возможен только один вариант ответа)*?

A. В течение последних 24 часов (суток) = *1*

B. Более 24 часов (суток) назад = *2*

Если =1, то → **Сколько часов назад Вы употребляли алкоголь?**
(Вводится число от 0 до 24 в соответствии с ответом респондента)

Если = 2, то учитываем, каким образом выражен ответ:

a. Если ответ выражен в часах, днях, неделях, месяцах или годах

...часов назад = *1*
... дней назад = *2*
... недель назад = *3*
... месяцев назад = *4*
... лет назад = *5*

} *Далее введите число часов, дней, недель, месяцев, лет (вводится число от 1 до 999)*

b. Если ответ выражен точной датой = *6* →*Укажите дату: дд/мм/гг (необходим календарь для выбора даты)*

c. Если ответ выражен разнообразными временными характеристиками:

Вчера = *7*
Позавчера = *8*
Поза-позавчера (три дня назад) = *9*
В прошлый понедельник = *10*
В прошлый вторник = *11*
В прошлую среду = *12*
В прошлый четверг = *13*
В прошлую пятницу = *14*
В прошлую субботу = *15*
В прошлое воскресенье =*16*
Не помню, когда это было = *17*
Другой вариант ответа = *18* *(указать словами)*

3. Когда Вы <u>в предпоследний раз</u> употребляли алкоголь *(возможен только один вариант ответа)*?

A. Не употреблял(а) алкоголь до описанного выше случая = *0* → **переход к вопросу 8.**

B. В течение последних 24 часов (суток)

Интервьюер уточняет, от какого момента ведется отсчет:

от момента интервью = *1*

от момента последнего эпизода употребления алкоголя = *2*

C. Более 24 часов (суток) назад

Интервьюер уточняет, от какого момента ведется отсчет:

от момента интервью = *3*

от момента последнего употребления алкоголя = *4*

Если =1 или =2, то → **Сколько часов назад Вы употребляли алкоголь?**
(Вводится число от 0 до 24)

Если = 3 или = 4, то учитываем, каким образом выражен ответ:

a. Если ответ выражен в часах, днях, неделях, месяцах или годах

...часов назад = *1*
... дней назад = *2*
... недель назад = *3*
... месяцев назад = *4*
... лет назад = *5*

Далее введите число часов, дней, недель, месяцев, лет (вводится число от 1 до 999)

b. Если ответ выражен точной датой = *6* →*Укажите дату: дд/мм/гг (необходим календарь для выбора даты)*

c. Если ответ выражен разнообразными временными характеристиками:

Вчера = *7*
Позавчера = *8*
Поза-позавчера (три дня назад) = *9*
В прошлый понедельник = *10*
В прошлый вторник = *11*
В прошлую среду = *12*
В прошлый четверг = *13*
В прошлую пятницу = *14*
В прошлую субботу = *15*
В прошлое воскресенье = *16*
Не помню, когда это было = *17*
Другой вариант ответа = *18* *(указать словами)*

4. Когда Вы <u>в пред-предпоследний раз</u> употребляли алкоголь?

A. Не употреблял(а) алкоголь до описанных выше двух случаев = *0* → **переход к вопросу 8.**

B. В течение последних 24 часов (суток)

Интервьюер уточняет, от какого момента ведется отсчет:

от момента интервью = *1*

от момента последнего эпизода употребления алкоголя = *2*

C. Более 24 часов (суток) назад

Интервьюер уточняет, от какого момента ведется отсчет:

от момента интервью = *3*

от момента последнего употребления алкоголя = *4*

Если =1 или =2, то → **Сколько часов назад Вы употребляли алкоголь?** *(Вводится число от 0 до 24)*

Если = 3 или =4, то учитываем, каким образом выражен ответ:

a. Если ответ выражен в часах, днях, неделях, месяцах или годах

...часов назад = *1*
... дней назад = *2*
... недель назад = *3*
... месяцев назад = *4*
... лет назад = *5*

Далее введите число часов, дней, недель, месяцев, лет (вводится число от 1 до 999)

b. Если ответ выражен точной датой = *6* →*Укажите дату: дд/мм/гг*

c. Если ответ выражен разнообразными временными характеристиками:

Вчера = *7*
Позавчера = *8*
Поза-позавчера (три дня назад) = *9*
В прошлый понедельник = *10*
В прошлый вторник = *11*
В прошлую среду = *12*
В прошлый четверг = *13*
В прошлую пятницу = *14*
В прошлую субботу = *15*
В прошлое воскресенье = *16*
Не помню, когда это было = *17*
Другой вариант ответа = *18* *(указать словами)*

5. Каким был САМЫЙ КОРОТКИЙ промежуток времени между <u>последовательными</u> случаями употребления Вами алкоголя за <u>последние 6 месяцев</u>?

A. Я ни разу не употреблял алкоголь в течение последних 6 месяцев. = 0→ переход к вопросу 8.

B. Я только один раз употреблял алкоголь в течение последних 6 месяцев. = 0→ переход к вопросу 8.

C. Если употреблял более одного раза за последние 6 месяцев, то *учитываем, каким образом выражен ответ:*

Часы = *1*	Месяцы = *4*	*Далее введите число часов, дней, недель, месяцев, лет (вводится число от 1 до 999)*
Дни = *2*	Годы = *5*	
Недели = *3*	Не помню = *6*	

6. Каким был САМЫЙ ДЛИННЫЙ промежуток времени между <u>последовательными</u> случаями употребления Вами алкоголя за <u>последние 6 месяцев</u>?

Учитываем, каким образом выражен ответ:

Часы = *1*	Месяцы = *4*	*Далее введите число часов, дней, недель, месяцев, лет (вводится число от 1 до 999)*
Дни = *2*	Годы = *5*	
Недели = *3*	Не помню = *6*	

7. Каким был ОБЫЧНЫЙ промежуток времени между <u>последовательными</u> случаями употребления Вами алкоголя за <u>последние 6 месяцев</u>?

Учитываем, каким образом выражен ответ:

Часы = *1*	Месяцы = *4*	*Далее введите число часов, дней, недель, месяцев, лет (вводится число от 1 до 999)*
Дни = *2*	Годы = *5*	
Недели = *3*	Не помню = *6*	

X. Комментарии интервьюера о частоте употребления респондентом алкоголя (ОТКРЫТЫЙ ВОПРОС…)

Респонденты могут указать особую регулярность употребления (например, по пятницам, или какую-то иную закономерность), а также нарушения обычной закономерности употребления в какой-то момент (например, в связи с праздниками)

В англоязычном варианте такой блок вопросов имеет вид:

---

1. Have you ever consumed any alcoholic beverage in your life?

Yes = *1*

No = *0* ***If NO, go to question 8.***

2. When was the last time you consumed an alcoholic beverage? *Only one response permitted*

Within last 24 hours *= 1*

a. How many hours ago did you consume an alcoholic beverage?

Enter number 0 – 24

More than 24 hours ago = *2*

a. Would you like to answer in hours, days, weeks, months, or years?

Hours = *1*
Days = *2*
Weeks = *3*
Months = *4*
Years = *5*

b. Enter number *Enter 1-999*

Exact date = *3*

a. Select the date

Yesterday = *4*
The day before yesterday = *5*
The day before
the day before yesterday = *6*
Last Sunday = *7*
Last Monday = *8*
Last Tuesday = *9*
Last Wednesday = *10*
Last Thursday = *11*
Last Friday = *12*
Last Saturday = *13*
The episode happened
but I cannot remember
when it happened = *14*

3. **When did you consume an alcoholic beverage the <u>second</u> to last time?** *Only one response permitted*

I did not consume an alcoholic beverage any other time before the instance I described above. = *0*

Within last 24 hours = *1*

a. How many hours ago did you consume an alcoholic beverage the second to last time? <u>OR</u> How many hours ago was this?

Enter number 0 - 24

More than 24 hours ago = *2*

a. Would you like to count from the time of this interview or from the last time you consumed an alcoholic beverage?

From the time of this interview = *1*
From the last time I drank an alcoholic beverage = *2*

b. Would you like to answer in hours, days, weeks, months, or years?

Hours = *1*
Days = *2*
Weeks = *3*
Months = *4*
Years = *5*

c. Enter number *Enter 1-999*

Exact date = *3*

a. Select the date

Yesterday = *4*
The day before yesterday = *5*
The day before
the day before yesterday = *6*
Last Sunday = *7*
Last Monday = *8*
Last Tuesday = *9*
Last Wednesday = *10*
Last Thursday = *11*
Last Friday = *12*
Last Saturday = *13*
The episode happened
but I cannot remember
when it happened = *14*

4. When did you consume an alcoholic beverage the <u>third</u> to last time?

I did not consume an alcoholic beverage any other time before the second to last time. = *0*

Within last 24 hours = *1*

a. How many hours ago did you consume an alcoholic beverage the third to last time? <u>OR</u> How many hours ago was this?

Enter number 0 - 24

More than 24 hours ago = *2*

a. Would you like to count from the time of this interview or from the <u>second</u> to last time you consumed an alcoholic beverage?

From the time of this interview = *1*
From the second to last time I drank an alcoholic beverage = *2*

b. Would you like to answer in hours, days, weeks, months, or years?

Hours = *1*
Days = *2*
Weeks = *3*
Months = *4*
Years = *5*

c. Enter number *Enter 1-999*

Exact date = *3*

a. Select the date

Yesterday = *4*
The day before yesterday = *5*
The day before
the day before yesterday = *6*
Last Sunday = *7*
Last Monday = *8*
Last Tuesday = *9*
Last Wednesday = *10*
Last Thursday = *11*
Last Friday = *12*
Last Saturday = *13*
The episode happened
but I cannot remember
when it happened = *14*

5. In the past 6 months, what was the SHORTEST time to elapse between any given episode of consuming an alcoholic beverage and the very next time you drank an alcoholic beverage again?

I did not drink any alcoholic beverage in the past 6 months. = *0 Go to #8*
I only drank an alcoholic beverage once in the past 6 months. = *5 Go to #8*
Would you like to answer in hours, days, weeks, or months?

Hours = *1* Weeks = *3*
Days = *2* Months = *4*

a. Enter number *Enter 1-999*

Do not remember = *6*

6. In the past 6 months, what was the LONGEST time to elapse between any given episode of consuming an alcoholic beverage and the very next time you drank an alcoholic beverage again?

Would you like to answer in hours, days, weeks, or months?

Hours = *1* Weeks = *3*
Days = *2* Months = *4*

a. Enter number *Enter 1-999*

Do not remember = *6*

7. In the past 6 months, about how much time USUALLY or typically elapsed between any given episode of consuming an alcoholic beverage and the very next time you drank an alcoholic beverage again?

Would you like to answer in hours, days, weeks, or months?

Hours = *1* Weeks = *3*
Days = *2* Months = *4*

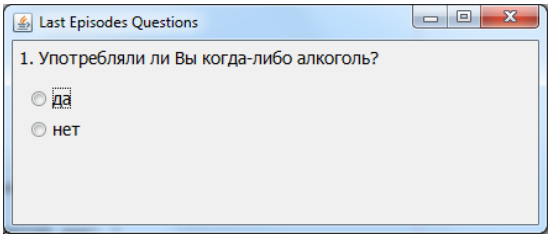
a. Enter number *Enter 1-999*

Do not remember = *6*

XX. You can add more information on the frequency of your consuming an alcoholic beverage (OPEN QUESTION…)

Разработанный инструментарий удобен для дальнейшей программной реализации, упрощающей условные переходы между вопросами. Элементы простейшего интерфейса представлены на рис. 2а–б.

(а)



(б)

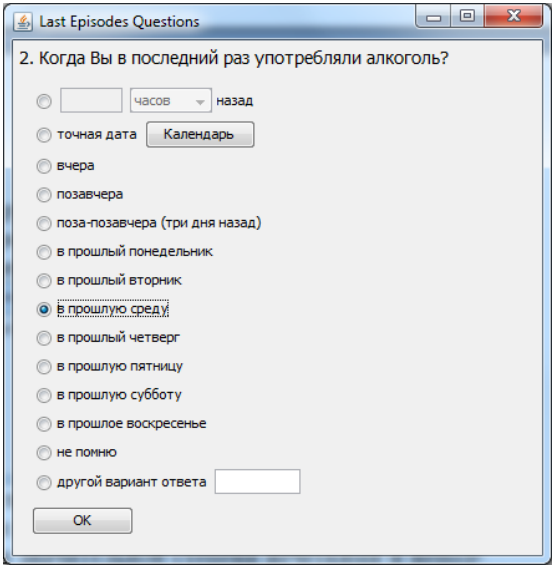


Рис. 2. Элементы пользовательского интерфейса.

8. Гранулярность исходных данных. В случае опроса респондентов об особенностях их поведения 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. Ограниченное число и неточность, неопределенность, нечеткость естественно-языковых формулировок ответов не позволяют напрямую использовать известные методы из теории массового обслуживания для оценки интенсивности поведения.

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: «четыре недели назад» и «месяц назад». В обыденном языке эти формулировки часто обозначают одно и то же, пренебрегая при этом тем фактом, что месяц — это четыре недели плюс два–три дня. Когда респондент использует формулировку «четыре недели назад», это свидетельствует о более высокой «надежности» припоминания и его уверенности в том, что событие произошло именно четыре недели назад, с погрешностью в несколько дней. Когда респондент использует формулировку «месяц назад», он априорно снижает точность высказывания. Месяц назад — это может быть около трех недель назад, а в каких-то случаях — шесть недель. Отметим, что ответ в виде «28 дней назад» свидетельствует об уверенности респондента в относительно точной дате события. Таким образом, можно говорить о гранулярности получаемых ответов (рис. 3). На рис. 3 схематично представлены сведения о нескольких последних эпизодах поведения. Пусть (интервью) — это момент интервью, (1) — момент на оси времени, когда произошел последний эпизод поведения. $T_1, T_2, ..., T_n$ — длины временных интервалов между моментом интервью и последним эпизодом, полученные по результатам опроса.

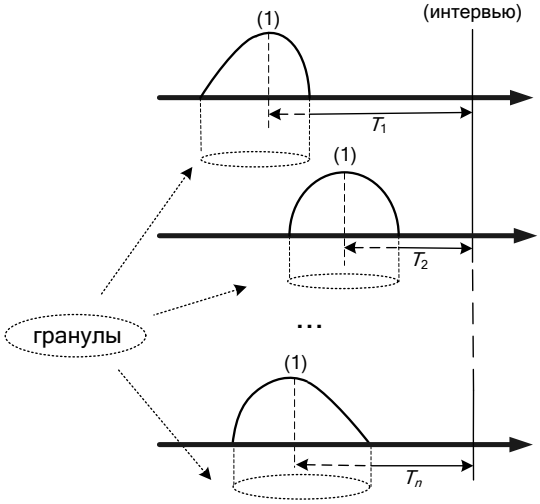


Рис. 3. Гранулярность ответов.

В силу существенной неопределенности высказываний на естественном языке получить точную численную оценку $\hat{t}_i$ затруднительно. Однако ее можно рассмотреть как случайную величину, построенную над другими случайными величинами. Особенности процесса построения такой случайной величины подробно рассмотрены в [11].

Для каждого значения T_i, $1 \le i \le n$ (n — размер выборки) через характеристику разброса δ определяется интервал (*возможных* значений) в днях: $[T_i - \delta x, T_i + \delta x]$, где x — коэффициент перевода рассматриваемой единицы измерения в дни [11].

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

Введенная случайная величина $\hat{T}_i$ за счет рандомизации [20–22] неопределенности ответа позволяет рассмотреть длину интервала, а следовательно, и параметры, определяющие распределение интенсивности, как случайную величину и вычислить характеристики последней.

Для каждого значения соответствующий интервал разбивается на m частей. Рассматриваются все возможные сочетания точек из интервалов, вычисляются их веса и рассчитывается среднее значение параметров. Однако при увеличении размера выборки такой метод может оказаться вычислительно сложным, поэтому необходимы другие методы обработки неопределенности.

9. Заключение. Необходимость оценки интенсивности поведения и нахождения ее связей с другими параметрами исследуемых объектов возникает в современных науках о человеке и обществе при решении многих задач. Одним из способов такого оценивания является рассмотрение данных об интервалах между последним эпизодом поведения и моментом интервью, однако при таком подходе на результат оказывает сильное влияние систематическая ошибка выборки. Предложенные изменения, внесенные в уравнение плотности распределения длин интервалов, позволяют обрабатывать такую ошибку. Для

гамма-пуассоновской модели поведения полученное распределение в результате перехода к новой системе параметров преобразуется к виду, указывающему на принадлежность рассматриваемого распределения к классу бета-распределений.

Результатов, полученных в рамках классических операций с распределениями вероятности, моментами случайных величин и функциями правдоподобия, оказывается недостаточно при переходе к неточным данным, доступным в результате проведения интервьюирования или опроса респондентов. То есть, для дальнейшей обработки неопределенности, связанной с гранулярностью исходных данных, требуется использовать метод сводных показателей [20-22]; кроме того, полученные результаты предполагается развить с помощью методов анализа нечетких временных рядов, опираясь, в частности, на работы [23–25].

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. Также работа поддержана грантом AIDS International Training and Research Program “Training and Research in HIV Prevention in Russia” (2 D43 TW001028) от Fogarty International Center, National Institutes of Health, USA.

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. *Rothman K.J.* Epidemiology: An Introduction. Oxford etc.: Oxford University Press, 2002. 223 p.
3. *Bonita R., Beaglehole R., Kjellstrom T.* Basic epidemiology. Geneva: WHO, 2006. 226 p.
4. *Bell D.C, Trevino R.A.* Modeling HIV Risk [Epidemiology] // JAIDS. 1999. Vol. 22(3). P. 280–287.
5. *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.
6. *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.
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. *Frowler F.J.* Improving survey questions: design and evaluation. Thousand Oaks, CA: SAGE Publications, 1995. 200 p. (Applied social research methods series, v. 38.)
9. *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.
10. *Тулупьев А.Л., Суворова А.В., Тулупьева Т.В., Пащенко А.Е.* Косвенные оценки и сравнение параметров угрообразующего поведения в разных группах по неполным и неточным данным // Международная конференция по мягким вычислениям и измерениям. Сборник докладов. 2009. Т. 2. СПб.: Изд-во СПбГЭТУ «ЛЭТИ», 2009. С. 110–114.
11. *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.
12. *Zelterman D., Tulupyev A., Heimer R., Abdala N.* Statistical design for a small serial dilution series // Statistics in Medicine. 2010. Vol. 29. P. 411–420.
13. *Zelterman D.* Models for Discrete Data: Revised Edition. New York: Oxford University Press, 2006. 285 p.
14. *Bekker A., Roux J., Pham-Gia T.* Sankhyā:The type I distribution of the ratio of independent "Weibullized" generalized beta-prime variables // Stat Papers. 2009. Vol. 50. P.323–338.
15. *Coelho C., Mexia J.* On the Distribution of the Product and Ratio Independent Generalized Gamma-Ratio Random Variables // The Indian Journal of Statistics. 2007. Vol. 69. Part. 2. P. 221–255.
16. *McDonald J., Xu Y.* A generalization of the beta distribution with applications // Journal of Econometrics. 1995. Vol. 66. P. 133–152.
17. *Pham-Gia T.* Exact distribution of the generalized Wilks's statistic and applications // Journal of Multivariate Analysis. 2008. Vol. 99. P. 1698–1716.
18. Beta Prime Distribution // Wikipedia. (Access: March 17, 2011.) <http://en.wikipedia.org/wiki/Beta_prime_distribution>.
19. *Zelterman D., Chen Ch.* Homogeneity test against central-mixture alternatives // Journal of the American Statistical Association. 1988. Vol. 83. No. 401. P. 179–182.
20. *Hovanov N., Yudaeva M., Hovanov K.* Multicriteria estimation of probabilities on basis of expert non-numeric, non-exact and non-complete knowledge // European Journal of Operational Research. 2009. Vol. 195. Issue 3. P. 857–863.
21. *Хованов Н.В.* Анализ и синтез показателей при информационном дефиците. СПб.: Изд-во СПбГУ, 1996. 196 с.
22. *Хованов Н.В.* Метод рандомизированных траекторий в задачах оценки функциональной зависимости // Труды СПИИРАН. 2009. Вып. 9. С. 262–279.
23. *Ярушкина Н. Г.* Современный интеллектуальный анализ нечетких временных рядов // Интегрированные модели и мягкие вычисления в искусственном интеллекте. V-я Международная научно-практическая конференция. Сборник научных трудов. В 2-х т. Т. 1. С. 19–29.
24. *Ковалев С.М.* Гибридные коннекционистские модели извлечения темпоральных знаний // Интегрированные модели и мягкие вычисления в искусственном интеллекте. V-я Международная научно-практическая конференция. Сборник научных

трудов. В 2-х т. Т. 1. С. 30–40.
25. Нечеткие гибридные системы. Теория и практика / под ред. Ярушкиной Н.Г. М.: Физматлит, 2007. 208 с.

Зельтерман Даниэл — Ph.D., Full Professor; профессор отделения биостатистики, факультет эпидемиологии и общественного здоровья, медицинский факультет, Йельский университет. Область научных интересов: разработка статистических методов обработки категориальных и данных о выживаемости, прикладная биостатистика. Число научных публикаций — 150. daniel.zelterman@yale.edu; 60 College St, LEPH 204, EPH, Yale University, New Haven, CT, 06510-3210, USA; ph.: +1 203 785-5574, fax: +1 203 785-6912.

Zelterman Daniel — Ph.D., Full Professor; Professor, Division of Biostatistics, Yale School of Epidemiology and Public Health, Yale University School of Medicine, Yale University. Research area: statistical methodology developments for categorical and survival data, applied biostatistics. Number of publications — 150. daniel.zelterman@yale.edu; 60 College St, LEPH 204, EPH, Yale University, New Haven, CT, 06510-3210, USA; ph.: +1 203 785-5574, fax: +1 203 785-6912.

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. SuvorovaAV@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. 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. SuvorovaAV@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. 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 num-

ber 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.

Мусина Валерия Фуатовна — программист лаборатории теоретических и междисциплинарных проблем информатики Учреждения Российской академии наук Санкт-Петербургский институт информатики и автоматизации РАН (СПИИРАН), студент математико-механического факультета Санкт-Петербургского государственного университета (СПбГУ). Область научных интересов: математическая статистика, теория вероятности, биостатистика, обработка данных. Число научных публикаций — 2. valery.musina@gmail.com, 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.

Musina Valery Fuatovna — programmer, Laboratory of Theoretical and Interdisciplinary Computer Science, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), student, Faculty of Mathematics and Mechanics of St. Petersburg State University (SPbSU). Research interests: mathematical statistics, probability theory, biostatistics, data processing. The number of publications — 2. valery.musina@gmail.com, 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.

Тулупьев Александр Львович — д.ф.-м.н., доцент; заведующий 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.

Тулупьева Татьяна Валентиновна — канд. психол. наук, доцент; старший научный сотрудник лаборатории теоретических и междисциплинарных проблем информатики Учреждения Российской академии наук С.-Петербургский институт информатики и автоматизации РАН (СПИИРАН), доцент кафедры информатики математико-механического факультета С.-Петербургского государственного университета (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 — 70. TVT@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.

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. Number of publications — 70. TVT@iias.spb.su, www.tulupyev.spb.ru; SPIIRAS, 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-3337, fax +7(812)328-4450.

Красносельских Татьяна Валерьевна — канд. мед. наук, доцент кафедры дерматовенерологии с клиникой Санкт-Петербургского Государственного медицинского университета им. акад. И.П.Павлова (СПбГМУ), начальник информационно-аналитического отдела Управления научных исследований СПбГМУ. Область научных интересов: разработка превентивных технологий, основанных на мультидисциплинарном подходе, для групп повышенного поведенческого риска заражения инфекциями, передаваемыми половым путем (ИППП), в том числе ВИЧ-инфекцией. Число научных публикаций — 85. tatiana.krasnoselskikh@gmail.com; кафедра дерматовенерологии с клиникой, ул. Льва Толстого, д. 6/8, г. Санкт-Петербург, 197022, РФ; тел. +7 921 764-1612.

Krasnoselskikh Tatiana Valerievna — MD, PhD, Associate Professor, Department of Dermatology and Venereology, Pavlov State Medical University, St. Petersburg (PSMU); Head of Division of Information and Analysis, Department of Scientific Research, PSMU. Research interests: development of preventive technologies based on the multidisciplinary approach for the populations under high risk of sexually transmitted infections (STIs) including HIV infection. Number of publications — 85. tatiana.krasnoselskikh@gmail.com; Department of Skin and Venereal Diseases, 6/8 Leo Tolstoy Str., St. Petersburg, 197022, Russia; phone +7 921 764-1612.

Гро Лоретта — PhD; научный сотрудник, медицинский факультет, Йельский университет. Область научных интересов: исследования среди потребителей инъекционных наркотиков, обнаружение и предупреждение передозировки наркотиками, количественные и качественные методы анализа данных, разработка и проверка методик тестирования, качественные методы интервьюирования, когнитивные и эмоциональные соотношения между рискованным поведением и профилактическими мероприятиями. Число научных публикаций — 26. lauretta.grau@yale.edu; 60 College St, LEPH 504, EPH, Yale University, New Haven, CT, 06510-3210, USA; ph.: +1 203 785-2904, fax: +1 203 785-3260.

Lauretta E. Grau — PhD; Associate Research Scientist, Yale School of Public Health, Yale University. Research area: health promotion among injection drug users and opioid overdose recognition and prevention, quantitative and qualitative data analysis, the development and validation of quantitative instruments, qualitative interviewing skills, and the cognitive and emotional correlates of risk and preventive health behaviors. Number of publications — 26.

lauretta.grau@yale.edu; 60 College St, LEPH 504, EPH, Yale University, New Haven, CT, 06510-3210, USA; ph.: +1 203 785-2904, fax: +1 203 785-3260.

Хаймер Роберт — Ph.D., Full Professor; профессор, медицинский факультет, Йельский университет. Область научных интересов: эпидемиология инфекционных заболеваний (в особенности ВИЧ, гепатит, ИППП). Число научных публикаций — 114. robert.heimer@yale.edu; 60 College St, LEPH 504, EPH, Yale University, New Haven, CT, 06510-3210, USA; ph.: +1 203 785-6732, fax: +1 203 785-3260.

Heimer Robert — Ph.D., Full Professor; Professor, Yale School of Public Health, Yale University. Research area: epidemiology of infectious diseases (with the focus on HIV, hepatitis, STD). Number of publications — 114. robert.heimer@yale.edu; 60 College St, LEPH 504, EPH, Yale University, New Haven, CT, 06510-3210, USA; ph.: +1 203 785-6732, fax: +1 203 785-3260.

Рекомендовано ТиМПИ СПИИРАН, зав. лаб. А.Л. Тулупьев, д.ф.-м.н., доцент.
Статья поступила в редакцию 25.03.2011.

РЕФЕРАТ

Зельтерман Д., Тулупьев А.Л., Суворова А.В., Пащенко А.Е., Мусина В.Ф., Тулупьева Т.В., Красносельских Т.В., Гро Л., Хаймер Р. **Обработка систематической ошибки, связанной с длиной временных интервалов между интервью и последним эпизодом в гамма-пауссоновской модели поведения.**

Задачи оценивания интенсивности и производных характеристик 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 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 гамма-пуассоновский случайный процесс. Интенсивность поведения предлагается оценивать по данным о последних эпизодах рассматриваемого поведения или, другими словами, по известным длинам интервалов между последовательными эпизодами поведения. Так, в случае трех последних эпизодов известны значения длин интервалов между моментом интервью и последним эпизодом, между последним и предпоследним эпизодами и между предпоследним и третьим с конца. Отметим, что момент интервью не является эпизодом поведения, таким образом, рассмотрение первого из перечисленных интервалов как интервала между последовательными эпизодами приводит к возникновению систематической ошибки, способам учета которой посвящена данная работа.

Для анализа особенностей длины интервала между интервью и последним эпизодом рассмотрено вероятностное распределение, отражающее наблюдение, что чем длиннее интервал между эпизодами, тем более вероятно, что момент интервью попадет в этот (более длинный) интервал. Описаны различные характеристики такого распределения, предложены варианты его перепараметризации. Показано, что полученное распределение принадлежит более широкому классу — является частным случаем бета-простого распределения.

В работе предложены методы, позволяющие обработать гранулярные исходные данные.

SUMMARY

Zelterman D., Tulupyev A.L., Suvorova A.V., Paschenko A.E., Musina V.F., Tulupyeva T.V., Krasnoselskikh T.V., Grau L., Heimer R. **Processing length bias of time intervals between the last episode and the interview.**

In many fields of sociological, psychological and marketing research, we face the problem of socially significant behavior rate or frequency estimate on the base of respondents' self-reports about their behavior. The traditional approaches to ask respondents about their behavior frequency fall into two categories. The first category relies upon question about the number of episodes of the behavior that have happened during month, 3 months, 6 months or another period of time; and it is highly implausible that respondents are able to recall all the episodes. The second category allows for collecting answers in Likert scale (e.g. "always", "very often", "often", "sometimes", "rarely", "never"); this type of answers does not provide information enough for making quantitative estimates of the behavior rate or frequency.

Our earlier studies have shown that respondents can stably provide their answers about last episodes of their behavior. In this paper, we describe a model that allows for making quantitative estimates of behavior rate or frequency based on the respondents' responses about the last episodes of their behavior.

The mathematical model of this behavior is a generalized Poisson stochastic process (Gamma-Poisson stochastic process); the observations are respondents' natural language answers about time intervals between the interview and the last episode of the respondent's behavior. Our approach takes into account the length bias that is represented in the model with the assumption that interviews happen more likely in longer intervals.

We have inferred that the random variable representing the time intervals between the interview and last episode is distributed according a Beta prime distribution law. We offer two different parameterization of the random variable distribution and derive the expected value and variance of the random variable as well as the likelihood function. Based on maximum likelihood approach and a sample of intervals, we can numerically estimate the distribution parameters with R software. We notice that the distribution parameters should satisfy certain conditions so that the mathematical expectation and variance exist.

Finally, we discuss ways to construct alternative stochastic models for the behavior and issues related to the granularity of the initial data.