УДК 311.2 + 616-036.22 + 004.8

Д.В. Степанов, В.Ф. Мусина, А.В. Суворова,
А.Л. Тулупьев, А.В. Сироткин, Т.В. Тулупьева

ФУНКЦИЯ ПРАВДОПОДОБИЯ С ГЕТЕРОГЕННЫМИ АРГУМЕНТАМИ В ИДЕНТИФИКАЦИИ ПУАССОНОВСКОЙ МОДЕЛИ РИСКОВАННОГО ПОВЕДЕНИЯ В СЛУЧАЕ ИНФОРМАЦИОННОГО ДЕФИЦИТА

Степанов Д.В., Мусина В.Ф., Суворова А.В., Тулупьев А.Л., Сироткин А.В., Тулупьева Т.В. **Функция правдоподобия с гетерогенными аргументами в идентификации пуассоновской модели рискованного поведения в случае информационного дефицита.**

Аннотация. Предложен подход к решению задачи оценки интенсивности рискованного поведения индивида по данным, которые являются системами ответов респондентов на вопросы, касающиеся их поведения. Оценка строится методом максимального правдоподобия, причём функция правдоподобия характеризует правдоподобие реализации конкретной системы ответов. Построены функции правдоподобия для ситуации, когда исследователь имеет данные о нескольких последних последовательных эпизодах поведения и ситуации, когда имеются данные об одном последнем эпизоде поведения и рекордных интервалах между последовательными эпизодами процесса за заданный промежуток времени.

Ключевые слова: функция правдоподобия, рискованное поведение, гетерогенные данные, последние эпизоды, рекордные интервалы

Stepanov D.V., Musina V.F., Suvorova A.V., Tulupyev A.L., Sirotkin A.V., Tulupyeva T.V. **Risky behavior Poisson model identification: heterogeneous arguments in likelihood.**

Abstract. We consider a problem of risky behavior intensity estimation based on data received from respondent's answers about their behavior. The method of maximum likelihood estimation is used for derivation of the estimate, and likelihood function describes the likelihood of realization of the particular system of answers. The likelihood function is derived for the situation when data about several last episodes of the behavior is available and situation when data about one last episode of the behavior and record intervals between consequent episodes of the behavior process occurred within fixed time interval is available.

Keywords: likelihood function, risky behavior, heterogeneous data, last episodes, record intervals

1. Введение. Задача оценки риска передачи и приобретения опасных неизлечимых инфекций в популяции является одной из задач современной эпидемиологии [**41**]. Неизлечимость таких заболеваний обуславливает серьёзность как и социального, так и экономического ущерба, наносимого обществу при распространении

заболевания. В настоящее время значительные усилия прилагаются к мониторингу и анализу ситуации, связанной с распространением и лечением опасных неизлечимых инфекций [**18**, **31**]. Во многих странах и городах мира существуют специальные программы, направленные на сдерживание риска передачи и получения таких заболеваний [**3**]. Однако комплексный анализ ситуации, связанной с распространением опасных неизлечимых инфекций и медицинской поддержкой заразившихся, требует учёта значительного числа факторов, и, как следствие, основополагающим становится привлечение к анализу ситуации методов искусственного интеллекта, который, в свою очередь, требует математического и алгоритмического моделирования ситуации [**2**, **14**, **19**, **26**].

В настоящей работе поведение индивида, ассоциированное с риском, рассматривается как последовательность эпизодов, каждый из которых связан с возможностью нанесения того или иного типа ущерба ему самому или другим лицам, вовлечённым в те или иные эпизоды такого поведения. К примеру, ущерб в виде передачи неизлечимого вирусного заболевания, такого как ВИЧ, может быть причинён, если человек, не являющийся носителем вируса, вовлечён в определённые типы взаимодействий с носителем вируса, которые предполагают возможность передачи заболевания. Исследования [**9**, **13**, **15**, **27**] показали, что типов взаимодействий, сопряжённых с риском передачи вируса, ограниченное количество. Предполагается, что последовательность эпизодов участия в подобных взаимодействиях — это процесс рискованного поведения индивида.

Наиболее точно риск передачи или приобретения ВИЧ индивидом характеризуется [**17**, **23**] отношением числа индивидов, находившихся под риском заражения и приобретших заболевание, к человеко–месяцам наблюдения [**41**, **42**]. Такое отношение носит название инсиденс-показателя (person–time incidence rate). Прямые методы измерения инсиденс-показателя [**37**](когортные исследования масштабом в 1000–2000 человеко–лет) имеют достаточно высокую стоимость, поэтому важно разработать математические модели, позволяющие получать косвенные оценки инсиденс-показателя, например по данным на основе ответов респондентов из группы риска. Инсиденс-показатель можно косвенно оценить, зная индивидуальный риск заражения за заданный период времени каждого отдельного респондента, попавшего в выборку. В

частности, модель Белла–Тревино [**36**] увязывает оценку кумулятивного риска передачи или приобретения ВИЧ в определённый период времени с числом эпизодов рискованного поведения, произошедших в этот период времени, и вероятностью заразиться при участии в одном эпизоде рискованного поведения.

Как отмечалось, риск передачи или приобретения заболевания внутри популяции имеет также экономический смысл: лечение подобных заболеваний требует значительных затрат как со стороны индивида, так и со стороны системы здравоохранения [**39**], так что рискованное поведение может нести ущерб не только здоровью индивида, но и экономическому благосостоянию страны. В социальных и экономических целях возможно внедрение различных превентивных мер, направленных на контроль распространения заболевания [**38**]. Оценка эффективности внедрения таких мер также связана с оценкой риска передачи или приобретения заболевания индивидом до вмешательства и после вмешательства, что, в свою очередь, требует нескольких измерений инсиденс–показателя. Как уже упоминалось, стоимость прямых измерений инсиденс–показателя достаточно высока [**4**,**37**], поэтому разработка косвенных методов оценивания инсиденс-показателя важна в задаче оценки эффективности внедрения превентивных мер.

Косвенные методы получения информации о рискованном поведении часто связаны с исследованием результатов анкет, опросов, интервью, дневников и иных видов самоотчётов. С целью получения информации о числе эпизодов рискованного поведения в рассматриваемый период времени, исследователь может обратиться к тому или иному виду самоотчёта о поведении респондентов, находившихся под риском заражения, например, провести интервью или опрос. Обозначим начало рассматриваемого периода времени как 0, а момент интервью — I; а сам рассматриваемый период времени будем называть периодом наблюдения. Исследователя интересует число эпизодов, которые произошли в этом интервале времени; если число эпизодов поведения в определённом интервале времени известно, то, согласно модели Бэлла–Тревино [**36**], возможно вычислить кумулятивный риск приобретения ВИЧ в популяции. Однако если респондент может предоставить данные лишь о нескольких последних эпизодах поведения, то, чтобы *оценить* общее число эпизодов за заданный промежуток времени $[0, I]$ ис-

следователь вынужден прибегнуть к оценке параметров математической модели процесса поведения на основе имеющихся данных.

В настоящей работе мы концентрируем своё внимание на пуассоновской модели рискованного поведения. В этом случае основным параметром процесса является интенсивность, которая характеризует число эпизодов поведения, произошедших в течение заданного интервала времени. В рамках пуассоновской модели поведения, предлагается построить оценку максимального правдоподобия параметра интенсивности пуассоновского процесса. Зная такую оценку, исследователь может оценить искомое число эпизодов рискованного поведения [**35**], которые произошли в заданный период $[0, I]$, а значит и получить соответствующую оценку кумулятивного риска, ассоциированного с поведением индивида, за этот период. Пуассоновский процесс обладает рядом особенностей, поэтому авторам статьи представляется несомненной необходимость развития тематики, введение в рассмотрение иных классов случайных процессовв серии последующих работ.

Таким образом, в рамках выбранной модели задача оценки риска передачи или приобретения неизлечимой инфекции, в конце концов, сводится к задаче оценки параметра интенсивности пуассоновского процесса, характеризующего рискованное поведение индивида.

Пилотные опросы [**9**, **13**, **15**, **27**] показали, что респонденты без затруднений дают ответы на вопросы о последних эпизодах рискованного поведения и об экстремальных значениях интервалов между последовательными эпизодами за некоторый промежуток времени. Таким образом, исследователь может получить систему ответов, состоящую из длин рекордных интервалов времени между последовательными эпизодами поведения за определённый период времени и длин интервалов между несколькими последними последовательными эпизодами (включая особый интервал между последним эпизодом и моментом интервью I). Значения, составляющие такую систему, будут неточными и нечёткими (в силу того, что ответы даются на естественном языке) [**27**, **33**, **34**], так что необходимо производить соответствующую последующую обработку ответов.

Необходимо отметить, что возможно формирование различных систем ответов, описывающих поведение. Возможно, исследователю будет доступна лишь дата, когда произошёл последний эпизод

рискованного поведения, то есть интервал между последним эпизодом и моментом интервью; возможно, исследователь сможет собрать сведения о нескольких последних последовательных эпизодах поведения; возможно, ему будут доступны и сведения о значениях длин рекордных интервалов между последовательными эпизодами поведения; возможно, имеющаяся информация позволит сформировать и какую-либо иную систему ответов.

К текущему моменту коллектив имеет значительный задел в исследованиях по данной тематике. Проведено полевое исследование для сбора статистических данных и апробации разработанного опросного инструментария [**33**, **34**]. Предложены методы оценки параметров рискованного поведения по данным о последних эпизодах [**34**] и по сведениям о рекордных интервалах между эпизодами [**12**, **20**, **24**]. Кроме того, вычислены соответствующие интервальные оценки рассматриваемых параметров [**21**]. Предложены подходы, позволяющие вычислять относительные оценки интенсивности и риска [**16**], необходимые при сравнении групп между собой и при оценивании эффективности поведенческих интервенций. Предложены модификации методов, учитывающие неточность, неполноту и гранулярность данных [**23**, **24**]. Рассмотрены различные способы обработки данных об интервале между последним эпизодом поведения и моментом интервью: для пуассоновской [**19**] и гамма-пуассоновской модели [**4**,**5**]. По результатам полевого исследования выявлены психологические характеристики респондентов, ассоциированные с рискованным поведением [**4**, **22**, **33**]. Разработанные методы были реализованы в комплексах программ [**10**, **11**, **28**–**30**].

Отметим, что задача оценки интенсивности поведения индивида, ассоциированного с риском, возникает во многих областях знаний, так или иначе направленных на изучение деятельности человека и/или её последствий. К примеру, исследования социоинженерных атак, моделирующих ситуацию, в которой сторонний индивид так или иначе воздействует на информационную систему «персонал – технические средства» с целью получения критически важной информации, тесно связаны с моделированием поведения персонала, при котором возникает угроза успешной реализации атаки [**32**]. Оценка интенсивности такого поведения персонала компании в свою очередь позволяет оценить степень защищённости информационной системы [**6**, **7**].

Общая задача настоящего исследования — оценить параметры интенсивности в различных моделях рискованного поведения по описанным выше системам ответов или сведений. Цель статьи — описать в рамках пуассоновской модели поведения возможный вид функции правдоподобия для различных ситуаций, возникающих при опросе респондентов об их рискованном поведении: ситуации, когда исследователь имеет данные о нескольких последних последовательных эпизодах поведения и ситуации, когда имеются данные об одном последнем эпизоде поведения и рекордных интервалах между последовательными эпизодами процесса за заданный промежуток времени.

2. Функция правдоподобия реализации конкретной системы ответов о нескольких последних последовательных эпизодах. В качестве процесса, моделирующего процесс рискованного поведения, выбирается стандартный пуассоновский процесс с постоянной интенсивностью [**1**]. Вероятность того, что в интервале длины T произойдёт k эпизодов рискованного поведения зависит только от значения параметра интенсивности процесса и длины интервала, что выражается следующим образом:

$$P(N([t_0, t_0 + T]) = k; \lambda) = \frac{e^{-\lambda T}(\lambda T)^k}{k!}, \tag{1}$$

где $\lambda > 0$ — параметр интенсивности пуассоновского процесса [**35**].

В рамках этой модели, длина промежутка времени τ между последовательными эпизодами рискованного поведения случайна и имеет показательное распределение с плотностью

$$f(x; \lambda) = \lambda e^{-\lambda x}, x \geq 0, \tag{2}$$

где λ — параметр интенсивности пуассоновского процесса, который требуется оценить. Соответствующая функция распределения вероятности:

$$F(x; \lambda) = P(\tau \geq x; \lambda) = e^{-\lambda x}, x \geq 0. \tag{3}$$

Рассмотрим случай, когда исследователь получает информацию только об m последних последовательных эпизодах поведения респондента, которые происходят в моменты времени $t_m, t_{m-1}, \ldots t_1$, причём в момент времени t_1 произошёл последний (ближайший к моменту интервью) эпизод, в момент времени t_2 произошёл предпоследний эпизод и т.д. Исследователь получает

информацию об эпизодах поведения путём интервьюирования респондентов, момент интервью обозначим I. Таким образом, имеется m интервалов между эпизодами процесса, длины которых являются случайными: $\tau_0^* = I - t_1, \tau_1 = t_1 - t_2, \ldots, \tau_{m-1} = t_{m-1} - t_m$. Каждая из введённых длин интервалов является случайной величиной со своим распределением вероятности. Общая длина полученных интервалов $T = [t_m, I]$ также случайна. Так как в моменты времени $t_m, t_{m-1}, \ldots t_1$ происходят эпизоды поведения, которое моделируется процессом Пуассона, то случайная величина, отвечающая длине интервала $\tau_i, 1 \leq i \leq m-1$, имеет плотность распределения (2).

Интервал τ_0^* не является интервалом между эпизодами поведения, и, вообще говоря, имеет другое распределение. Однако в рамках пуассоновской модели поведения, распределение такой случайной величины представляется в явном виде [**35**], что позволит провести ревизию предлагаемых методов построения функции правдоподобия реализации конкретной системы ответов с учётом данных о конкретном распределении случайной величины τ_0^*.

Введённую систему интервалов и индексов удобно представить в виде таблицы (см. табл. 1).

Таблица 1: Введённая система обозначений

Индекс	Случайная величина	Длина интервала	Интервал
0	τ_0^*	$I - t_1$	$[t_1, I]$
1	τ_1	$t_1 - t_2$	$[t_2, t_1]$
2	τ_2	$t_2 - t_3$	$[t_3, t_2]$
...	...	...	...
i	τ_i	$t_i - t_{i+1}$	$[t_{i+1}, t_i]$
...	...	...	...
$m-2$	τ_{m-2}	$t_{m-2} - t_{m-1}$	$[t_{m-1}, t_{m-2}]$
$m-1$	τ_{m-1}	$t_{m-1} - t_m$	$[t_m, t_{m-1}]$

Обозначим $\tilde{x}$ конкретную реализацию случайной величины x.

В результате интерпретации содержания интервью или опроса респондента исследователь имеет ряд реализаций соответствующих случайных величин $\tilde{\tau}_0^*, \tilde{\tau}_1, \ldots, \tilde{\tau}_{m-1}$. Формируя систему из таких реализаций, необходимо помнить, что такая система будет гетерогенной, так как каждое значение длины интервала является реализацией отдельной случайной величины, отвечающей это-

му интервалу. В рамках выбранной пуассоновской модели поведения, каждая из реализаций $\tilde{\tau}_1, \ldots, \tilde{\tau}_{m-1}$ является реализацией случайной величины с показательным распределением с параметром λ, так как такие случайные величины отвечают длинам интервалов между последовательными эпизодами рискованного поведения. Что ещё более важно, реализация $\tilde{\tau}_0^*$ является реализацией случайной величины, отвечающей особому интервалу.

Рассмотрим случай, когда момент интервью совпадает с эпизодом процесса. Если это не так, то предположение о том, что момент интервью совпадает с эпизодом поведения, позволяет получить оценку сверху искомой оценки интенсивности рискованного поведения в силу монотонностью оценки максимального правдоподобия относительно длинны последнего интервала, что верно для рассматриваемого случая. Имеются m интервалов между последними последовательными эпизодами рискованного поведения $\tilde{\tau}_0^*, \tilde{\tau}_1, \ldots, \tilde{\tau}_{m-1}$, причём $\tilde{T} = \tilde{\tau}_0^* + \sum\limits_{i=1}^{m-1} \tilde{\tau}_i$. Функция правдоподобия есть функция плотности совместного распределения выборки, вычисленная в точке наблюдаемых значений, т.е. вероятность того, что мы имеем конкретную систему ответов $\tilde{\tau}_0^*, \tilde{\tau}_1, \ldots, \tilde{\tau}_{m-1}$ из всех возможных означиваний случайных величин $\tau_0^*, \tau_1, \ldots, \tau_{m-1}$ как функция параметра.

Пусть $p(x_0, x_1, \ldots, x_m; \lambda)$ — совместная плотность вероятности случайных величин $\tau_0^*, \tau_1, \ldots, \tau_m$

$$P(\tau_0^* < x_0, \tau_1 < x_1, \ldots, \tau_{m-1} < x_{m-1}; \lambda) = \\ = \int\limits_0^{x_0} \int\limits_0^{x_1} \ldots \int\limits_0^{x_{m-1}} p(u_0, u_1, \ldots, u_{m-1}, \lambda) du_0 \ldots du_{m-1}.$$

В силу нашего локального предположения о том, что интервал τ_0^* является интервалом между последовательными эпизодами рискованного поведения (т.е. предположения о том, что момент интервью есть эпизод поведения), в рамках пуассоновской модели поведения все случайные величины $\tau_0^*, \tau_1, \ldots, \tau_{m-1}$ являются независимыми и имеют показательное распределение с параметром интенсивности λ, откуда имеем:

$$p(\tilde{\tau}_0^*, \tilde{\tau}_1, \ldots, \tilde{\tau}_{m-1}; \lambda) = f(\tilde{\tau}_0^*, \lambda) \prod_{i=1}^{m-1} f(\tilde{\tau}_i, \lambda)$$

Функция правдоподобия выглядит следующим образом:

$$L_m^*(\lambda) = p(\tilde{\tau}_0^*, \tilde{\tau}_1, \ldots, \tilde{\tau}_{m-1}; \lambda) = f(\tilde{\tau}_0^*; \lambda) \prod_{i=1}^{m-1} f(\tilde{\tau}_i; \lambda) =$$
$$= \lambda e^{-\lambda \tilde{\tau}_0^*} \prod_{i=1}^{m-1} \lambda e^{-\lambda \tilde{\tau}_i} = \lambda^m e^{-\lambda(\tilde{\tau}_0^* + \sum\limits_{i=1}^{m} \tilde{\tau}_i)} = \tag{4}$$
$$= \lambda^m e^{-\lambda \tilde{T}}.$$

Оценка максимума правдоподобия выглядит следующим образом:

$$\hat{\lambda}^* = \arg\max_{\lambda \geq 0} L_m^*(\lambda) = \frac{m}{\tilde{T}}. \tag{5}$$

Однако полученная оценка, как уже упоминалась, является оценкой сверху оценки максимального правдоподобия интенсивности процесса, в силу предположения, что момент интервью есть эпизод процесса поведения.

Откажемся от этого предположения; теперь полагаем, что исследователь имеет данные об интервале $\tilde{\tau}_0^*$, который меньше, чем истинный интервал между эпизодом процесса, который произошёл в момент времени t_0 и некоторым «будущим» эпизодом, который произойдёт в момент времени t_{-1}. Обозначим случайную величину, моделирующую длину этого интервала, как $\tau_0 = t_{-1} - t_0$. Чтобы обработать имеющиеся у нас сведения, в рассматриваемой вероятностной модели требуется рассмотреть особый объект: не непрерывная случайная величина с конкретной реализацией, а случайное событие случайная величина длины интервала между последним эпизодом и гипотетическим последующим эпизодом больше наблюдаемой величины интервала между последним эпизодом и моментом интервью. Тогда мы построим, опираясь на изначальное понимание правдоподобия и ряд предположений о независимости, некоторые из которых могут как иметь, так и не иметь достаточного обоснования, новую, уточнённую функцию правдоподобия, являющуюся произведением вероятности указанного события и плотностей вероятности. Как и ранее, воспользуемся свойством независимости интервалов между последовательными эпизодами поведения, получим

$$L_m(\lambda) = P(\tau_0 \geq \tilde{\tau}_0^*)p(\tilde{\tau}_1, \ldots, \tilde{\tau}_{m-1}; \lambda) = \\ = P(\tau_0 \geq \tilde{\tau}_0^*) \prod_{i=1}^{m-1} f(\tilde{\tau}_i; \lambda) = F(\tilde{\tau}_0^*; \lambda) \prod_{i=1}^{m-1} \lambda e^{-\lambda \tilde{\tau}_i} = \\ = e^{-\lambda \tilde{\tau}_0^*} \lambda^{m-1} e^{-\lambda \sum_{i=1}^{m-1} \tilde{\tau}_i} = \lambda^{m-1} e^{-\lambda \tilde{T}}, \tag{6}$$

где $F(t; \lambda)$ — показательная функция распределения (3), а $p(x_1, \ldots, x_{m-1})$ как и ранее есть совместная плотность вероятности, только теперь иного набора случайных величин $\tau_1, \ldots, \tau_m$

В этом случае, оценка максимального правдоподобия будет выглядеть следующим образом:

$$\hat{\lambda} = \arg\max_{\lambda \geq 0} L_m(\lambda) = \frac{m-1}{\tilde{T}}. \tag{7}$$

Случай $m = 1$ (то есть наблюдается лишь один последний эпизод рискованного поведения), является особым: в этом случае оценка максимального правдоподобия параметра интенсивности $\hat{\lambda} = \frac{0}{\tilde{T}} = 0$. При $m = 1$ выбранный способ построения функции максимального правдоподобия не пригоден, так как не позволяет получить обоснованную оценку интенсивности поведения. Также заметим, если всё же принять предположение о том, что момент интервью является эпизодом процесса рискованного поведения, можно получить оценку сверху оценки параметра интенсивности поведения: $\hat{\lambda} = \frac{1}{\tilde{T}}$.

Удобно представить формулы для вычисления конкретных оценок в наиболее часто встречающихся случаях $m = 1, 2, 3$ в виде таблицы (см. табл. 2).

Таблица 2: Формулы для расчёта оценки интенсивности поведения при $m = 1, 2, 3$

m	L_m^*	$\hat{\lambda}^*$	L_m	$\hat{\lambda}$
1^*	$\lambda e^{-\lambda \tilde{\tau}_0^*}$	$\frac{1}{\tilde{\tau}_0^*}$	$e^{-\lambda \tilde{\tau}_0^*}$	Некорр.
2	$\lambda^2 e^{-\lambda(\tilde{\tau}_0^* + \tilde{\tau}_1)}$	$\frac{2}{\tilde{\tau}_0^* + \tilde{\tau}_1}$	$\lambda e^{-\lambda(\tilde{\tau}_0^* + \tilde{\tau}_1)}$	$\frac{1}{\tilde{\tau}_0^* + \tilde{\tau}_1}$
3	$\lambda^3 e^{-\lambda(\tilde{\tau}_0^* + \tilde{\tau}_1 + \tilde{\tau}_2)}$	$\frac{3}{\tilde{\tau}_0^* + \tilde{\tau}_1 + \tilde{\tau}_2}$	$\lambda^2 e^{-\lambda(\tilde{\tau}_0^* + \tilde{\tau}_1 + \tilde{\tau}_2)}$	$\frac{2}{\tilde{\tau}_0^* + \tilde{\tau}_1 + \tilde{\tau}_2}$

3. Учет данных о рекордных интервалах в функции правдоподобия. Часто имеющийся объём данных о поведении

респондента не ограничивается сведениями о последних последовательных эпизодах рискованного поведения. При проведении полевых исследований было обнаружено, что респонденты без затруднений сообщают сведения о минимальном, максимальном и обычном интервалах между последовательными эпизодами рискованного поведения, имевшими место в некоторый предшествующий интервью фиксированный интервал времени (в частности, месяц, три месяца, полгода). Чтобы учесть такие данные, необходимо сначала выбрать горизонт ретроспективного самоотчёта, который мы условно назовём интервал наблюдения: $[0, I]$, где за 0 обозначено «начало» наблюдения, а I — фиксированное значение момента интервью. Пусть в этом интервале произошло $N(I;\lambda)$ событий (по свойствам процесса Пуассона, число эпизодов в определённом интервале зависит лишь от длины интервала и параметра интенсивности λ), и респондент может дать информацию о m последних эпизодах рискованного поведения и о длинах рекордных интервалов между последовательными эпизодами поведения за период наблюдения $[0, I]$. Важно отметить, что $N(I;\lambda)$ является случайной величиной, и может принимать значения от 0 до ∞ согласно (8).

$$P(N([0,I]) = k;\lambda) = \frac{e^{-\lambda I}(\lambda I)^k}{k!}. \tag{8}$$

В интервале $[0, I],; I > 0$ в случайные моменты времени $t_1, \ldots, t_{N(I;\lambda)}$ происходят эпизоды рискованного поведения. Как и ранее, t_1 — момент, когда произошёл последний эпизод рискованного поведения, т.е. самый близкий к моменту интервью; в момент времени t_2 происходит предпоследний эпизод; эпизод, который произошёл в момент $t_{N(I;\lambda)}$ является первым за период наблюдения $[0, I]$, т.е. самым близким к началу наблюдения. Произошедшие эпизоды определяют $N(I;\lambda)+1$ неотрицательных случайных величин, соответствующие длинам интервалов: $\tau_0^* = I - t_1, \tau_1 = t_1 - t_2, \ldots, \tau_i = t_i - t_{i-1}, \ldots, \tau_{N(I;\lambda)-1} = t_{N(I;\lambda)-1} - t_{N(I;\lambda)}, \tau^*_{N(I;\lambda)} = t_{N(I;\lambda)} - 0$. Верхним индексом $*$ обозначены особые интервалы, вообще говоря, не являющиеся интервалами между эпизодами рискованного поведения. Схема введённых обозначений аналогична указанной в табл. 1.

Интервалы $\tau_1, \tau_2, \ldots, \tau_{N(I;\lambda)-1}$ являются интервалами между эпизодами рискованного поведения, которое моделируется пуас-

соновским процессом. Это означает, что случайные величины $\tau_1, \tau_2, \ldots, \tau_{N(I;\lambda)-1}$ имеют показательное распределение (2) с параметром λ, который является параметром интенсивности пуассоновского процесса.

Интервал $\tau^*_{N(I;\lambda)}$ есть интервал времени между началом наблюдения и первым эпизодом, произошедшим в периоде наблюдения. «Начало» наблюдения не является эпизодом рискованного поведения; вместе с тем, в силу свойства отсутствия памяти показательного распределения [**35**], хотя интервал $\tau^*_{N(I;\lambda)}$ не является «полным» интервалом между последовательными эпизодами–событиями процесса, он имеет показательное распределение с тем же параметром λ.

Перед исследователем стоит общая задача оценивания параметра интенсивности поведения λ по имеющимся данным об интервалах между последовательными последними эпизодами поведения и рекордными интервалами между последовательными эпизодами поведения за период наблюдения. Снова отметим, что при этом общее количество эпизодов в интервале времени $[0, I]$ остаётся неизвестным, и именно это число является целью исследователя.

Рассмотрим сначала частную задачу. Пусть исследователь имеет возможность получить информацию об одном последнем эпизоде, то есть интервале τ^*_0, и рекордных интервалах $\tau_{\min}, \tau_{\max}$. Чтобы построить оценку максимального правдоподобия имеющейся системы конкретных данных $\tilde{\tau}^*_0, \tilde{\tau}_{\min}, \tilde{\tau}_{max}$, требуется построить функцию правдоподобия $L_{1,\min,\max}(\lambda)$ с тем, чтобы, подставив в эту функцию конкретные значения $\tilde{\tau}^*_0, \tilde{\tau}_{\min}, \tilde{\tau}_{\max}$, решить экстремальную задачу, максимизирующую правдоподобие реализации именно этих значений,

$$L_{1,\min,\max}(\lambda) = p(\tilde{\tau}^*_0, \tilde{\tau}_{\min}, \tilde{\tau}_{\max}; \lambda) \tag{9}$$

и получить значение параметра, доставляющего максимум этой вероятности

$$\hat{\lambda} = \arg\max_{\lambda \geq 0} L_{1,\min,\max}(\lambda). \tag{10}$$

Здесь

$$p(\tilde{\tau}^*_0, \tilde{\tau}_{\min}, \tilde{\tau}_{\max}; \lambda) =$$
$$= \frac{\partial^3}{\partial x \partial y \partial z} P(\tau^*_0 < z, \tau_{\min} < x, \tau_{\max} < y; \lambda)|_{\tilde{\tau}^*_0, \tilde{\tau}_{\min}, \tilde{\tau}_{\max}}.$$

Как упоминалось в предыдущем разделе, в отличие от ряда классических случаев, аргументы функции правдоподобия, неоднородны, носят гетерогенный характер. Действительно, $\tilde{\tau}_0^*, \tilde{\tau}_{\min}, \tilde{\tau}_{\max}$ являются конкретными реализациями отдельных случайных величин $\tau_0^*, \tau_{\min}, \tau_{\max}$, каждая из которых имеет своё распределение вероятности.

Построение функции правдоподобия в этом случае будет осуществляться по следующему плану:

- переход от плотности совместного распределения случайных величин $\tau_0^*, \tau_{\min}, \tau_{\max}$ к плотности совместного распределения случайных величин $\tau_0^*, \tau_n^{(1)}, \tau_n^{(n)}$, где $\tau_n^{(1)}, \tau_n^{(n)}$ являются первым и последним членами вариационного ряда n случайных величин с показательным распределением с параметром λ; $n \geq 2$;
- вычисление функции совместного распределения случайных величин $\tau_0^*, \tau_n^{(1)}, \tau_n^{(n)}$: $Q_{n,\lambda}(z, x, y) = P_n(\tau_0^* < z, \tau_n^{(1)} < x, \tau_n^{(n)} < y)$;
- вычисление условной функции распределения при условии конкретного означивания $\tau_n^{(1)} = u, \tau_n^{(n)} = v$.

В итоге получаем вычислимое выражение.

Итак, имеются $N(I;\lambda) + 1$ случайных величин τ_0^*, τ_1, ..., $\tau_{N(I;\lambda)-1}$, $\tau_{N(I;\lambda)}^*$, из которых $\tau_1, \ldots, \tau_{N(I;\lambda)-1}$ являются интервалами между последовательными эпизодами поведения и имеют показательное распределение с параметром λ, а $\tau_{N(I;\lambda)}^*$, хоть и не является интервалом между последовательными эпизодами поведения, имеет то же распределение [**35**]. Построим вариационный ряд из случайных величин, отвечающих длинам интервалов между последовательными эпизодами поведения (в том числе $\tau_{N(I;\lambda)}^*$): $\tau_{N(I;\lambda)}^{(1)} < \tau_{N(I;\lambda)}^{(2)} < \ldots < \tau_{N(I;\lambda)}^{(N(I;\lambda))}$. Важно отметить, что при построении такого вариационного ряда в настоящей работе мы предполагаем, что интервал $\tau_{N(I;\lambda)}^*$ не может быть первым или последним членом ряда, так как респондент не воспринимает этот интервал как интервал между эпизодами. Первый и последний члены вариационного ряда отвечают, соответственно, случайным величинам $\tau_{\min}, \tau_{\max}$. Однако распределения порядковых статистик вариационного ряда зависят от его длины, в то время как

$\tilde{\tau}_{\min}, \tilde{\tau}_{\max}$ не зависят от числа наблюдений в интервале. Таким образом, распределения вероятности $\tau^{(1)}_{N(I;\lambda)}, \tau^{(N(I;\lambda))}_{N(I;\lambda)}$ зависят от числа эпизодов $N(I;\lambda)$, произошедших в интервале наблюдения $[0, I]$, а это число, в свою очередь, напрямую зависит (8) от интенсивности поведения λ. Используя формулу полной вероятности, получим

$$\begin{gathered} p(\tilde{\tau}^*_0, \tilde{\tau}_{\min}, \tilde{\tau}_{\max}; \lambda) = \\ = \sum_{n=2}^{\infty} p_n(\tilde{\tau}^*_0, \tilde{\tau}^{(1)}_n, \tilde{\tau}^{(n)}_n; \lambda) P(N(I) = n) = \\ \sum_{n=2}^{\infty} p_n(\tilde{\tau}^*_0, \tilde{\tau}^{(1)}_n, \tilde{\tau}^{(n)}_n; \lambda) \frac{e^{-\lambda I}(\lambda I)^n}{n!}. \end{gathered} \tag{11}$$

Нижний предел суммирования $n = 2$ обусловлен требованием построения вариационного ряда $\tau^{(1)}_n < \tau^{(n)}_n$ и тем, что наблюдается ещё и последний эпизод поведения. Таким образом, необходимо, чтобы были получены данные о хотя бы двух интервалах между последовательными эпизодами. В случае $n = 1$ имеется всего один интервал между последовательными эпизодами поведения, а, значит, минимальный и максимальный интервалы совпадают. В этом случае можно применить рассуждения, рассмотренные в предыдущем разделе. Вопросы суммируемости ряда оставлены за рамками работы.

Пусть $p_n(z, x, y; \lambda)$ — функция плотности, соответствующая совместной функции распределения $P_n(\tau^*_0 < z, \tau^{(1)}_n < x, \tau^{(n)}_n < y; \lambda)$:

$$\begin{gathered} p_n(\tilde{\tau}^*_0, \tilde{\tau}_{\min}, \tilde{\tau}_{\max}; \lambda) = \\ = \frac{\partial^3}{\partial x \partial y \partial z} P_n(\tau^*_0 < z, \tau_{\min} < x, \tau_{\max} < y; \lambda)|_{\tilde{\tau}^*_0, \tilde{\tau}_{\min}, \tilde{\tau}_{\max}}. \end{gathered} \tag{12}$$

Тогда в терминах плотностей вероятности целевая функция $L_{1,\min,\max}(\lambda)$ (9) переписывается следующим образом:

$$\begin{gathered} L_{1,\min,\max}(\lambda) = p(\tilde{\tau}^*_0, \tilde{\tau}_{\min}, \tilde{\tau}_{\max}; \lambda) = \\ = \sum_{n=2}^{\infty} p_n(\tilde{\tau}^*_0, \tilde{\tau}^{(1)}_n, \tilde{\tau}^{(n)}_n; \lambda) \frac{e^{-\lambda I}(\lambda I)^n}{n!} \end{gathered} \tag{13}$$

Таким образом, вместо вычисления значения функции плотности совместного распределения случайных величин $\tau^*_0, \tau_{\min}, \tau_{\max}$

в точке $\tilde{\tau}_0^*, \tilde{\tau}_{\min}, \tilde{\tau}_{\max}$, необходимо вычислить значение функции плотности совместного распределения случайных величин $\tau_0^*, \tau_n^{(1)}, \tau_n^{(n)}$ в той же точке, где $\tau_n^{(1)}, \tau_n^{(n)}$ являются первым и последним членами вариационного ряда n случайных величин с показательным распределением с параметром λ.

Перейдём к вычислению плотности вероятности $p_n(z, x, y; \lambda)$, для этого обратимся к вычислению соответствующей функции распределения вероятности (12). Имеется $n + 1, n \geq 2$ случайных величин $\tau_0^*, \tau_1, \tau_2, \ldots, \tau_{n-1}, \tau_n$, из которых каждая из величин $\tau_1, \tau_2, \ldots, \tau_{n-1}, \tau_n$ имеет показательное распределение с параметром λ, а случайная величина τ_0^* имеет иное распределение вероятности. Случайные величины $\tau_0^*, \tau_1, \tau_2, \ldots, \tau_{n-1}, \tau_n$ представляют собой интервалы между эпизодами поведения, моделируемого пуассоновским процессом, и потому независимы. Случайные величины $\tau_1, \tau_2, \ldots, \tau_{n-1}, \tau_n$ формируют вариационный ряд $\tau_n^{(1)} < \tau_n^{(2)} < \ldots < \tau_n^{(n-1)} < \tau_n^{(n)}$.

Рассмотрим интервал времени между началом наблюдения и последним эпизодом рискованного поведения, так как общая длина интервала наблюдения I фиксирована, то его длина выражается следующим образом:

$$S_n(\tau_0^*; \lambda) = I - \tau_0^* = \sum_{i=1}^{n} \tau_i = \sum_{i=1}^{n} \tau_n^{(i)}. \tag{14}$$

С учётом введённого выше обозначения, искомая вероятность (12) вычисляется следующим образом:

$$\begin{gathered} Q_{n,\lambda}(z, x, y) = P_n(\tau_0^* < z, \tau_n^{(1)} < x, \tau_n^{(n)} < y; \lambda) = \\ = P_n(S(\tau_0^*; \lambda) > I - z, \tau_n^{(1)} < x, \tau_n^{(n)} < y; \lambda) = \\ = P_n(\tau_n^{(1)} < x, \tau_n^{(n)} < y; \lambda) - \\ - P_n(S(\tau_0^*; \lambda) < I - z, \tau_n^{(1)} < x, \tau_n^{(n)} < y; \lambda). \end{gathered} \tag{15}$$

Первое слагаемое в (15) представляет собой совместную функцию распределения порядковых статистик с номерами $1, n$, плотность которой вычисляется по формуле

$$f_{1,n}^n(u, v; \lambda) = n(n-1)(F(u) - F(v))^{n-2} f(u) f(v), \tag{16}$$

где $F(x)$ — показательная функция распределения (3), $f(x)$ — соответствующая плотность распределения (2). Обозначим функцию

распределения, соответствующую $f^n_{1,n}(u,v;\lambda)$ как $F^n_{1,n}(u,v;\lambda)$:

$$F^n_{1,n}(u,v;\lambda)=\sum_{r=1}^{n}\frac{n!}{(n-r)!r!}(1-e^{-\lambda u})^r(e^{-\lambda u}-e^{-\lambda v})^{n-r}. \qquad (17)$$

Рассмотрим теперь второе слагаемое в (15). По формуле полной вероятности получим:

$$\begin{gathered}P_n(S(\tau_0^*;\lambda)<I-z,\tau_n^{(1)}<x,\tau_n^{(n)}<y;\lambda)=\\ =P_n(S(\tau_0^*;\lambda)<I-z|\tau_n^{(1)}<x,\tau_n^{(n)}<y;\lambda)\times\\ \times P_n(\tau_n^{(1)}<x,\tau_n^{(n)}<y;\lambda)=\\ =\int\limits_{\Omega(x,y)}P_n(S(\tau_0^*;\lambda)<I-z|\tau_n^{(1)}=u,\tau_n^{(n)}=v;\lambda)\times\\ \times dP_n(\tau_n^{(1)}=u,\tau_n^{(n)}=v;\lambda)=\\ =\int\limits_{\Omega(x,y)}P(S(\tau_0^*;\lambda)<I-z|\tau_n^{(1)}=u,\tau_n^{(n)}=v)f^n_{1,n}(u,v;\lambda)dudv,\end{gathered} \qquad (18)$$

где $\Omega(x,y)=\{0\le u\le x,u<v\le y,x<y\}$ — множество всех значений u,v которые могут быть выбраны в качестве значений первого и последнего членов вариационного ряда. В формуле (18) все члены возможно получить численными методами, кроме подынтегрального множителя $P_n(S(\tau_0^*;\lambda)<I-z|\tau_n^{(1)}=u,\tau_n^{(n)}=v;\lambda)$, рассмотрим его подробнее.

Обозначим

$$H_{u,v}(x;\lambda)=\frac{F(x)-F(u)}{F(u)-F(v)}=\frac{1-e^{\lambda(u-x)}}{1-e^{\lambda(u-v)}}, \qquad (19)$$

здесь $F(x)$ — показательная функция распределения (3).

Тогда, согласно [**8**, **40**],

$$P_n(S(\tau_0^*;\lambda)<I-z|\tau_n^{(1)}=u,\tau_n^{(n)}=v;\lambda)=V^{*(n-2)}_{u,v;\lambda}(I-z) \qquad (20)$$

где $V^{*(n-2)}_{u,v;\lambda}(I-z)$ — $(n-2)$–кратная свертка функций распределения (19) [**35**].

Итого:

$$\begin{aligned}\bar{Q}_{n;\lambda}(z,x,y)&=P_n(S(\tau_0^*;\lambda)<I-z,\tau_{\min}<x,\tau_{\max}<y;\lambda)=\\ &=\int\limits_{\Omega(x,y)}V^{*(n-2)}_{u,v;\lambda}(I-z)f^n_{1,n}(u,v;\lambda)dudv\end{aligned} \qquad (21)$$

и для $Q_{n,\lambda}(z,x,y)$ получаем итоговую формулу

$$Q_{n;\lambda}(z,x,y) = F^n_{1,n}(x,y;\lambda) - \bar{Q}_{n;\lambda}(z,x,y), \tag{22}$$

здесь каждое слагаемое возможно получить численными методами.

Чтобы получить функцию правдоподобия, необходимо вычислить плотность вероятности (12), затем просуммировать ряд (13). Таким образом, функция правдоподобия для реализации системы ответов, которая состоит из одного последнего эпизода поведения $\tilde{\tau}^*_0$ и рекордных интервалов $\tilde{\tau}_{\min}$ и $\tilde{\tau}_{\max}$ за период наблюдения $[0, I]$ есть

$$\begin{aligned} L_{1,\min,\max}(\lambda) = \sum_{n=2}^{\infty} \frac{e^{-\lambda I}(\lambda I)^n}{n!} \frac{\partial^3}{\partial x \partial y \partial z}[F^n_{1,n}(x,y;\lambda) - \\ - \int\limits_{\Omega(x,y)} V^{*(n-2)}_{u,v;\lambda}(I-z) f^n_{1,n}(u,v;\lambda) du dv]|_{\tilde{\tau}^*_0, \tilde{\tau}_{\min}, \tilde{\tau}_{\max}}. \end{aligned} \tag{23}$$

здесь $f^n_{1,n}(u,v;\lambda)$ — совместная плотность распределения порядковых статистик (16) с номерами $1, n$, $F^n_{1,n}(x,y;\lambda)$ — соответствующая функция распределения, $V^{*(n-2)}_{u,v;\lambda}(I-z)$ — $(n-2)$–кратная свёртка функций распределения (19), множество $\Omega(x,y) = \{0 \le u \le x, u < v \le y, x < y\}$.

Необходимо как исследовать аспекты сходимости ряда в формуле (23), так и развить процедуры и методы вычисления выражения в (23).

4. Заключение. Задачи эпидемиологии часто связаны с построением оценки кумулятивного риска индивида заразиться или приобрести заболевание за определённый период времени [**41**]. В частности, в случае распространения ВИЧ, модель Белла–Тревино увязывает кумулятивный риск с числом эпизодов, произошедших в этот период времени, и риском заразиться за один эпизод поведения. Однако во многих ситуациях прямо оценить это число эпизодов не представляется возможным [**34**]. Таким образом, возникает задача косвенной оценки числа эпизодов, произошедших в определённый период времени.

Одним из возможных решений этой задачи является идентификация параметров процесса поведения индивида на основе его самоотчётов. Такие самоотчёты формируются, в частности, при

проведении интервью или опроса. При проведении пилотных исследований [**9**,**13**,**15**,**27**] выяснилось, что индивиды часто не могут сообщить интересующую исследователя информацию о непосредственно числе эпизодов поведения за определённый период времени; исследователю становится доступна лишь система ответов определённого вида. Таким образом, необходимо оценить искомое число эпизодов поведения по имеющейся системе ответов респондента.

Пусть поведение индивида представляет собой последовательность связанных с риском эпизодов поведения, в которые вовлекается (или которые индуцирует) индивид; в настоящей работе в качестве математической модели такого поведения индивида рассматрен стандартный пуассоновский процесс [**1**]. Тогда число эпизодов, произошедших в определённый период времени, можно оценить, зная параметр интенсивности пуассоновского процесса. То есть в конце концов перед исследователем стоит частная задача оценки интенсивности поведения индивида в рамках пуассоновской модели поведения на основании имеющейся реализации системы ответов респондента.

Одним из основных методов оценивания параметров является метод максимального правдоподобия, который предполагает решение задачи максимизации функции правдоподобия, характеризующей правдоподобие реализации конкретной системы ответов. В статье предложены методы построения функции правдоподобия для двух систем ответов: система ответов из нескольких последних эпизодов поведения и система ответов из одного последнего эпизода и рекордных интервалов между последовательными эпизодами процесса поведения. Результаты для первой из рассматриваемых систем ответов представлены в формулах (4, 6) и табл.2, а функция правдоподобия для второй из рассматриваемых систем ответов — в формуле (23). Следует отметить, что требуется развить численные методы для поиска аргумента, доставляющего максимум функции правдоподобия (23).

Другим возможным способом учета данных как о последних эпизодах, так и о рекордных интервалах является построение модели в рамках теории байесовских сетей доверия [**25**, **26**], которая предполагает оценивание параметра интенсивности процесса при помощи байесовских методов оценивания.

Литература

1. *Булинский А.В., Ширяев А.Н.* Теория случайных процессов. М.: Физматлит, 2005. 408 с.

2. *Величенко В.В., Притык Д.А.* Возможности искусственного интеллекта и компьютерных технологий в построении программ лечения сложных иммунных заболеваний // Фундаментальная и прикладная математика. 2009. Т. 15. № 5. С. 21–42.

3. *Жолобов В.Е.* Итоги и перспективы реализации программ и планов мероприятий по предупреждению распространения в санкт-петербурге заболевания, вызываемого вирусом иммунодефицита человека (программы и планы "анти–ВИЧ/СПИД"1998–2012 гг.) // Вестник Российской военно-медицинской академии. 2011. Т. 1. С. 205–209.

4. *Зельтерман Д., Суворова А.В., Пащенко А.Е., Мусина В.Ф., Тулупьев А.Л., Тулупьева Т.В., Гро Л., Хаймер Р.* Диагностика регрессионных уравнений в анализе интенсивности рискованного поведения по его последним эпизодам // Труды СПИИРАН. 2011. № 17. С. 33–46.

5. *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 // Труды СПИИРАН. 2011. № 16. С. 160–185.

6. *Котенко И.В., Саенко И.Б., Юсупов Р.М.* Аналитический обзор докладов Международной конференции "Математические модели, методы и архитектуры для защиты компьютерных сетей"(MMM-ACNS-2010) // Труды СПИИРАН. 2010. № 13. С. 199–225.

7. *Котенко И.В., Степашкин М.В., Юсупов Р.М.* Математические модели, методы и архитектуры для защиты компьютерных сетей: аналитический обзор перспективных направлений исследований по результатам международного семинара MMM-ACNS-2005 // Труды СПИИРАН. 2006. № 3. Т. 2. С. 11–29.

8. *Невзоров В.Б.* Рекорды. Математическая теория.. М.: ФАЗИС, 2000. 244 с.

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. *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. Т. 2. М.: Физматлит, 2009. С. 220–241.

11. *Пащенко А.Е., Суворова А.В., Тулупьев А.Л.* Программа для расчёта нечётких оценок интенсивности угрозообразующего поведения и риска, с ним связанного, Fuzzy Risk-&-Rate Calculator (F.R.-&-R.C.). Роспатент. Свид. о гос. регистрации программы для ЭВМ № 2009614649 от 31.08.2009

12. *Пащенко А.Е., Суворова А.В., Тулупьев А.Л., Тулупьева Т.В.* Вероятностные распределения порядковых статистик в анализе сверхкоротких нечетких и неполных временных рядов // Труды СПИИРАН. 2009. № 10. СПб.: Наука С. 184–207.

13. *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. С. 33–34.

14. *Пащенко А.Е., Тулупьев А.Л., Николенко С.И.* Статистическая оценка вероятности заражения ВИЧ-инфекцией на основе данных о последних эпизодах рискованного поведения // Труды СПИИРАН. 2006. № 3. Т. 2. СПб.: Наука С. 257–268.

15. *Пащенко А.Е., Тулупьев А.Л., Тулупьева Т.В.* Апробация блока вопросов и обработка ответов о последних эпизодах рискованного поведения ВИЧ-инфицированных // X Санкт-Петербургская международная конференция «Региональная информатика–2006 (РИ-2006)»: Труды. СПб., 2007. С. 323–326.

16. *Пащенко А.Е., Тулупьев А.Л., Суворова А.В., Тулупьева Т.В.* Сравнение параметров угрозообразующего поведения в разных группах на основе неполных и неточных данных // Труды СПИИРАН. 2009. № 9. Спб.: Наука, 2009. С. 252–261.

17. *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.

18. *Рахманова А.Г., Лобзин Ю.В., Степанова Е.В., Волкова Г.В., Виноградова Е.Н.* ВИЧ-инфекция в Санкт–Петербурге // Российский медицинский журнал. 2004. № 3. С. 7–11.

19. *Суворова А.В., Лавренов А.В., Тулупьева Т.В., Тулупьев А.Л., Пащенко А.Е.* Моделирование социально-значимого поведения респондентов: аналитическая и численная оценки интенсивности в окрестности интервью при информационном дефиците // Труды СПИИРАН. 2012. № 1(20). С. 101–115.

20. *Суворова А.В., Пащенко А.Е., Тулупьева Т.В.* Оценка характеристик сверхкороткого временного ряда по гранулярным данным о рекордных интервалах между событиями // Труды СПИИРАН. 2010. № 12. С. 170–181.

21. *Суворова А.В., Пащенко А.Е., Тулупьева Т.В., Тулупьев А.Л.* Построение доверительных интервалов оценок интенсивности рискованного поведения на основе неравенства Чебышева // Труды СПИИРАН. 2009. № 10. СПб.: Наука. С. 107-120.

22. *Суворова А.В., Тулупьева Т.В., Тулупьев А.Л.* Обобщенная линейная регрессионная модель для прогноза временного интервала между последним эпизодом рискованного поведения и моментом интервью на основе социально-демографических и психологических особенностей // Труды СПИИРАН. 2012. № 2(21). С. 80-94.

23. *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. 2010. № 4. С. 30–38.

24. *Суворова А.В., Тулупьева Т.В., Тулупьев А.Л., Пащенко А.Е.* Эвристическая оценка интенсивности поведения по рекордным интервалам между эпизодами: обработка неточности ответов респондентов // XV Международная конференция по мягким вычислениям и измерениям. SCM-2012. (25-27 июня 2012 г. Санкт-Петербург). Сборник докладов. 2012. Т. 2. СПб.: Изд-во СПбГЭТУ «ЛЭТИ» С. 101–104.

25. *Суворова А.В., Тулупьева Т.В., Тулупьев А.Л., Сироткин А.В., Пащенко А.Е.* Применение байесовских сетей доверия для моделирования угрозообразующего поведения индивида по неполным и неточным данным // Тринадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2012 (16-20 октября 2012 г., г. Белгород). Труды конференции. Т. 3 Белгород: Изд-во БГТУ, 2012. С. 292–299.

26. *Суворова А.В., Тулупьева Т.В., Тулупьев А.Л., Сироткин А.В., Пащенко А.Е.* Вероятностные графические модели социально-значимого поведения индивида, учитывающие неполноту информации // Труды СПИИРАН. 2012. № 3(22). С. 101–112.

27. *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.

28. *Тулупьев А.Л., Суворова А.В., Пащенко А.Е.* Программа для учёта неточных сведений об угрозообразующем поведении Fuzzy Data Register for Risky Behavior, Version 1 (F.D.R.R.B. v. 1) // Роспатент. Свид. о гос. рег. прогр. для ЭВМ № 2010613161 от 14.05.2010.

29. *Тулупьев А.Л., Суворова А.В., Пащенко А.Е.* Программа для расчёта нечётких оценок интенсивности угрозообразующего поведения и риска, с ним связанного, Fuzzy Risk-&-Rate Calculator, Version 2(F.R.-&-R.C. v.2) // Роспатент. Свид. о гос. рег. прогр. для ЭВМ № 2010614267 от 30.06.2010.

30. *Тулупьев А.Л., Суворова А.В., Пащенко А.Е.* Программа для идентификации параметров интенсивности и риска в условиях неопределенности на основе рекордных порядковых статистик в моделях угрозообразующего поведения Record-Based Uncertain Risk-&-Rate Calculator (R.B.U.R.-&-R.C.) // Роспатент. Свид. о гос. рег. прогр. для ЭВМ № 2010614266 от 30.06.2010.

31. *Тулупьева Т.В., Тулупьев А.Л., Столярова Е.В., Пащенко А.Е.* Анализ особенностей рискованного поведения в модели адаптивных стилей ВИЧ-инфицированных (на основе результатов опроса пациентов Санкт-Петербургского СПИД-Центра) // Труды СПИИРАН. 2007. № 5. Спб.: Наука, 2008. С. 117–150.

32. *Тулупьева Т.В., Тулупьев А.Л., Пащенко А.Е., Азаров А.А., Степашкин М.В.* Социально-психологические факторы, влияющие на

степень уязвимости пользователей автоматизированных информационных систем с точки зрения социоинженерных атак // Труды СПИИРАН. 2010. № 12. С. 200–214.

33. *Тулупьева Т.В., Пащенко А.Е., Мусина В.Ф., Тулупьев А.Л., Жук С.Н., Азаров А.А., Суворова А.В., Сироткин А.В., Фильченков А.А.* Отчет о научно-исследовательской работе «Опросный инструментарий для выявления особенностей рискованного поведения в контексте адаптивных стилей и анализ результатов пилотного исследования», инвентарный № 02201259423 от 2012.06.26, по теме "Взаимосвязь адаптивных стилей ВИЧ-инфицированных и степени рискованности их поведения регистрационный № 01201262071. СПб.: СПИИРАН. 2012. 78 с. (Депонировано в ЦИТиС.)

34. *Тулупьева Т.В., Пащенко А.Е., Мусина В.Ф., Тулупьев А.Л., Азаров А.А., Жук С.Н., Казакова О.С., Красносельских Т.В., Сироткин А.В., Суворова А.В., Фильченков А.А.* Отчет о научно-исследовательской работе «Классификация ответов респондентов о последних эпизодах рискованного поведения и косвенная оценка его интенсивности» (заключительный),инвентарный № 02201259425 от 2012.06.26, по теме «Моделирование и измерение количественных характеристик ВИЧ-рискованного поведения на основе обработки ответов респондентов», регистрационный № 01201262070. СПб.: СПИИРАН. 2012. 34 с. (Депонировано в ЦИТиС.)

35. *Феллер В.* Введение в теорию вероятностей и её приложения. В 2-х томах. Т. 2. М.: Мир, 1984. 738 с.

36. *Bell D.C, Trevino R.A.* Modeling HIV Risk [Epidemiology] // JAIDS. 1999. № 22(3). P. 280–287.

37. *Hall H.I., Song R., Rhodes P., Prejean J., An Q., Lee L.M., Karon J., Brookmeyer R., Kaplan E.H., McKenna M.T., Janssen R.S.* Estimation of HIV Incidence in the United States // JAMA. 2008. № 300(5). P. 520–529.

38. *Holtgrave D.R., Qualls N.L., Graham J.D.* Economic Evaluation of HIV Prevention Programs // Annual Review of Public Health. 1996. Vol. 17. P. 467–488.

39. *Moore R.D.* Understanding the clinical and economic outcomes of HIV therapy: the Johns Hopkins HIV clinical practice cohort // Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology : Official Publication of the International Retrovirology Association. 1998. №17(1). P. 38–41.

40. *Nevzorov V.B.* Records: Mathematical Theory. Providence, Rhode Island: American Mathematical Society, 2001. 164 p.

41. *Rothman K.J.* Epidemiology: An Introduction. 2002. 223 p.

42. Федеральный научно-методический Центр по профилактике и борьбе со СПИДом. Количество ВИЧ-инфицированных в России за 2012 год. [Электронный ресурс http://www.hivrussia.ru]. По состоянию на 10.09.2012.

Степанов Денис Вячеславович — к.т.н., доцент кафедры Компьютерных технологий СПб НИУ ИТМО. Область научных интересов: случайные процессы, непараметрические методы математической статистики, биостатистика, статистическое моделирование, генетические алгоритмы, генетическое программирование. Число научных публикаций — 20. ALT@iias.spb.su, www.tulupyev.spb.ru; СПб НИУ ИТМО, Кронверкский пр. 49, Санкт-Петербург, 197101, РФ.

Stepanov Denis Vyacheslavovich — PhD in Engineering Sciences, Associate Professor, Computer Technologies Department, SPb National Research University of Information Technologies, Mechanics and Optics. Research area: stochastic processes, nonparametric statistics, biostatistics, statistic modelling, genetic algorithms, genetic programming. Number of publications — 20. ALT@iias.spb.su, www.tulupyev.spb.ru; SPb National Research University of Information Technologies, Mechanics and Optics, Kronverksky pr. 49, Saint-Petersburg, Russia; office phone +7(812)328-3337, fax +7(812)328-4450; р.т. +7(812)328-3337, факс +7(812)328-4450.

Мусина Валерия Фуатовна — младший научный сотрудник лаборатории теоретических и междисциплинарных проблем информатики СПИИРАН, студент магистратуры экономического факультета СПбГУ. Область научных интересов: случайные процессы, вероятностное и статистическое моделирование, биостатистика, вероятностные графические модели. Число научных публикаций — 13. 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.

Musina Valeriya Fuatovna — junior research fellow Theoretical and Interdisciplinary Computer Science Laboratory, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRA), graduate student of Faculty of Economics at Saint Petersburg State University. Research area: stochastic processes, probabilistic and statistic modelling, biostatistics, probabilistic graphical models. Number of publications — 13. 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.

Суворова Алена Владимировна — младший научный сотрудник лаборатории теоретических и междисциплинарных проблем информатики, Санкт-Петербургский институт информатики и автоматизации РАН, аспирант, Математико-механический факультет Санкт-Петербургского государственного университета. Область научных интересов: математическая статистика,

теория вероятности, применение методов математического моделирования в эпидемиологии. Число научных публикаций –– 37. SuvorovaAV@iias.spb.su, www.tulupyev.spb.ru, 14-я линия В.О., д.39, Санкт-Петербург, 199178, РФ; р.т.: +7(812)328-3337, факс: +7(812)328-4450.

Suvorova Alena Vladimirovna –– junior researcher, Laboratory of Theoretical and Interdisciplinary Computer Science, St. Petersburg Institute for Informatics and Automation of RAS, PhD student, The Faculty of Mathematics and Mechanics of Saint Petersburg State University. Research interests: mathematical statistics, probability theory, application of mathematical modeling in epidemiology The number of publications — 37. SuvorovaAV@iias.spb.su, www.tulupyev.spb.ru, 39, 14-th 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; заведующий лабораторией теоретических и междисциплинарных проблем информатики СПИИРАН, профессор 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, методы автоматизированной оценки защищенности персонала информационных систем от социоинженерных атак. Число научных публикаций — 230. 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.

Alexander Lvovich Tulupyev — PhD in Computer Science, Dr. of Sc., Associate Professor; Head of Theoretical and Interdisciplinary Computer Science Laboratory, SPIIRAS, 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 — 230. 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.

Сироткин Александр Владимирович — к.ф.-м.н., младший научный сотрудник лаборатории теоретических и междисциплинарных проблем информатики Учреждения Российской академии наук С.-Петербургский институт информатики и автоматизации РАН (СПИИРАН). Область научных интересов: алгебраические байесовские сети: вычислительные аспекты логико-вероятностного вывода в условиях неопределенности, математические методы анализа генома. Число научных публикаций — 64. avs@iias.spb.su; СПИИРАН, 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т. +7(812)328-3337, факс +7(812)328-4450.

Alexander Vladimirovich Sirotkin — PhD in Computer Science, junior researcher, Theoretical and Interdisciplinary Computer Science Laboratory, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: algebraic Bayesian networks, algorithms of probabilistic-logic inference under uncertainty. The number of publications —

64. avs@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.

Тулупьева Татьяна Валентиновна –– канд. психол. наук, доцент; старший научный сотрудник лаборатории теоретических и междисциплинарных проблем информатики, Санкт-Петербургский институт информатики и автоматизации РАН, доцент кафедры информатики, Математико-механический факультет Санкт-Петербургского государственного университета, доцент кафедры психологии управления и педагогики, Северо-Западная академия государственной службы. 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, 14-я линия В.О., д.39, Санкт-Петербург, 199178, РФ; рабочий телефон: +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 RAS, associate professor, Computer Science Department, The Faculty of Mathematics and Mechanics of Saint Petersburg State University, associate professor, Management Psychology and Pedagogic Department, North-Western Academy of Public Administration. Research interests: application of mathematics and computer science in humanities, informatization of psychological studies, application of biostatistics in epidemiology, psychology of personality, management psychology The number of publications –– 70. TVT@iias.spb.su, www.tulupyev.spb.ru, 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone: +7(812)328-3337, fax: +7(812)328-4450.

Поддержка исследований. Грант для молодых ученых и кандидатов наук от Правительства Санкт-Петербурга в 2009 №25.05/027/27 «Разработка математических моделей, вычислитель- ных алгоритмов и комплекса программ для оценки интенсивности рискованного поведе ния в условиях дефицита информации». Руководитель — А.Е. Пащенко.

Грант 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.

Гранты РФФИ 09-01-00861-а, 10-01-00640-а, 12-01-00945-а

Гранты Комитета по науке и высшей школе Правительства Санкт-Петербурга в 2012: «Модели и алгоритмы анализа сверхкоротких неточных временных рядов на основе гранулярных данных и знаний», руководитель — А.В. Суворова; «Разработка программного комплекса для идентификации интенсивности и производных параметров стохастических моделей рискованного поведения на основе неполных и неточных данных», руководитель — А.Е. Пащенко.

Рекомендовано ТиМПИ СПИИРАН, зав. лаб. Тулупьев А.Л., д.ф.-м.н., доцент.

Статья поступила в редакцию 16.09.2012.

РЕФЕРАТ

Степанов Д.В., Мусина В.Ф., Суворова А.В., Тулупьев А.Л., Сироткин А.В., Тулупьева Т.В. **Функция правдоподобия с гетерогенными аргументами в идентификации пуассоновской модели рискованного поведения в случае информационного дефицита.**

Задачи эпидемиологии связаны с построением оценки кумулятивного риска индивида передать или приобрести заболевание за определённый период времени. В частности, в случае распространения ВИЧ, модель Белла–Тревино увязывает кумулятивный риск с числом эпизодов, произошедших в этот период времени, и риском заразиться за один эпизод поведения. Однако во многих ситуациях прямо оценить это число эпизодов не представляется возможным. Таким образом, возникает задача косвенной оценки числа эпизодов, произошедших в определённый период времени.

Одним из возможных решений этой задачи является идентификация параметров процесса поведения индивида на основе самоотчётов о его поведении. Пилотные опросы показали, что индивиды часто не могут сообщить интересующую исследователя информацию о непосредственно числе эпизодов поведения за определённый период времени; исследователю становится доступна лишь система ответов определённого вида. Таким образом, необходимо оценить искомое число эпизодов поведения по имеющейся системе ответов респондента.

Пусть поведение индивида представляет собой последовательность связанных с риском эпизодов поведения, в которые вовлекается (или которые индуцирует) индивид; в настоящей работе в качестве математической модели такого поведения индивида рассмотрен стандартный пуассоновский процесс. Тогда число эпизодов, произошедших в определённый период времени, можно оценить, зная параметр интенсивности пуассоновского процесса.

Одним из основных методов оценивания параметров является метод максимального правдоподобия, который предполагает решение задачи максимизации функции правдоподобия, характеризующей правдоподобие реализации конкретной системы ответов. В статье предложены методы построения функции правдоподобия для двух систем ответов: система ответов из нескольких последних эпизодов поведения и система ответов из одного последнего эпизода и рекордных интервалов между последовательными эпизодами процесса поведения.

SUMMARY

Stepanov D.V., Musina V.F., Suvorova A.V., Tulupyev A.L., Sirotkin A.V., Tulupyeva T.V. **Risky behavior Poisson model identification: heterogeneous arguments in likelihood.**

Problems of the Epidemiology include ones connected with the estimation of the individual's cumulative risk of getting or transmitting the disease within specified period of time. In particular considering HIV prevalence the Bell-Trevino model links the cumulative risk with the number of episodes, which has occurred within mentioned period of time, and risk of being infected in the one episode. But in many situations one cannot get strait estimation of this number. Thus a need of indirect estimation of the number of episodes occurred in the specified period of time arises.

Identification of the parameters of the individual's behavior process based on his self-reports is one of the possible solutions of this problem. Pilot interviews showed that individuals cannot just give required information on the number of episodes of behavior occurred within the specified period of time, one can have only a system of answers of a certain type. Thus the problem of deriving an estimation of the number of episodes of the behavior on the base of available system of answers arises.

We assume that individual's behavior is a consequence of risk associated episodes of the behavior, in which individual is involved (or which he induces); in the present paper we consider the standard Poisson process as a mathematical model of the behavior. In the framework of this model one can estimate the number of episodes occurred within specified period of time knowing the intensity parameter of Poisson process.

One of the fundamental parameter estimation methods is method of maximum likelihood estimation; the main idea of this method is the maximization of the likelihood function, which describes the likelihood of particular system of answers realization. In the paper we develop methods of deriving the likelihood function for two particular systems of answers: system of answers which contains the information of several last episodes of behavior and system of answers which contains the information of one last episode and record intervals between consequent episodes of process.