Компьютерные инструменты в образовании, 2018
№ 3: 22–37
УДК: 51-77
http://ipo.spb.ru/journal
doi:10.32603/2071-2340-3-22-37

КОПУЛЫ И МОДЕЛИРОВАНИЕ ЗАВИСИМОСТИ: КОСВЕННЫЕ ОЦЕНКИ ИНТЕНСИВНОСТИ РИСКОВАННОГО ПОВЕДЕНИЯ*

Столярова В. Ф.

Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук, Санкт-Петербург, Россия

Аннотация

В условиях ограниченности ресурсов наиболее доступным способом получения информации о поведении индивида является интервью. В этом случае данные о недавнем поведении индивида наименее подвержены различным типам смещения. Для построения оценки интенсивности поведения по данным о трех последних эпизодах используются математические модели поведения. В работе рассмотрена гамма–пуассоновская модель поведения. Зависимость двух интервалов между последовательными эпизодами в этой модели описана в терминах копул. Оценка параметра копулы в этом случае напрямую ведет к оценке параметров распределения интенсивности поведения в популяции. Кроме того, вид копулы позволяет определить некоторые характеристики поведения, описываемого гамма–пуассоновской моделью: кластеризованность эпизодов поведения. Возможности предложенного подхода продемонстрированы на модельных данных.

Ключевые слова: *гамма–пуассоновская модель поведения, последние эпизоды, копула.*

Цитирование: Столярова В. Ф. Копулы и моделирование зависимости: косвенные оценки интенсивности рискованного поведения // Компьютерные инструменты в образовании. 2018. № 3. С. 22–37.

1. ВВЕДЕНИЕ

Человек является неотъемлемой частью многих технических, социо–экономических, киберфизических систем. При этом поведение человека может оказаться связано с риском, то есть поступки человека могут влиять на возможность реализации некоторого негативного исхода. К примеру, употребление алкоголя [10], наркотических веществ [6] или табакокурение [25] наносит непосредственный вред здоровью человека и связано с повышенным риском смертности и нетрудоспособности. В свою очередь, высокая частота интенсивных физических упражнений связана с уменьшенным риском развития ряда заболеваний [39]. Таким образом, оценка интенсивности поведения

*Работа выполнена в рамках проекта по государственному заданию СПИИРАН № 0073-2018-0001; гранта РФФИ №18-01-00626; гранта КНВШ г. Санкт-Петербурга, 2018 г., для аспирантов.

может быть необходима при общей оценке риска смертности и нетрудоспособности при личном страховании [9].

Кроме того, подобные оценки риска, связанного с поведением, требуются в области эпидемиологии и общественного здоровья [30]. Употребление алкоголя не только повышает личный риск и приводит к дополнительным затратам в сфере государственного здравоохранения, но и является фактором риска распространения неизлечимых заболеваний, таких как ВИЧ [36]. В статье [8] была предложена формула, позволяющая вычислить кумулятивный (накопленный риск или риск приобретения инфекции за некоторый промежуток времени) риск распространения ВИЧ–инфекции в популяции по данным об участии в нескольких типах рискованного поведения. Оценка кумулятивного риска распространения инфекций важна при разработке программ превенции и мониторинга эпидемиологической ситуации [17].

1.1. Косвенные оценки интенсивности поведения.

В рамках данной работы обратимся к тем типам поведения, которые представляются и исследователем, и респондентом как последовательность четко определенных эпизодов. К примеру, к таким типам поведения относятся употребление различных продуктов питания, кофе или алкоголя. Как показали полевые исследования[58], приступы головной боли часто не воспринимаются респондентом как последовательность эпизодов.

Итак, пусть поведение человека может быть представлено как последовательность эпизодов, каждый из которых вносит вклад в риск. Базовой характеристикой такой последовательности эпизодов может выступать их *интенсивность*[50, 51, 54, 58]. Однако не всегда исследователю доступны все данные об эпизодах поведения, то есть не всегда возможна *прямая оценка* интенсивности поведения [58]. Среди косвенных оценок интенсивности поведения особую роль играют те оценки, которые доступны при проведении интервью [41]. Интервью и самоотчеты респондентов об эпизодах их поведения не только экономически выгодны, но и позволяют значительно снизить время сбора информации по сравнению с прямыми и дневниковыми методами. Однако информация, получаемая на естественном языке, является неточной и неполной, и, кроме того, в пилотных исследованиях было показано, что респонденты достаточно точно могут припомнить лишь несколько последних эпизодов поведения [58].

С учетом особенностей сбора подобной информации о поведении был предложен подход к оценке интенсивности по данным из самоотчетов респондентов о последних эпизодах этого поведения [58]. Это подход состоит из двух частей: опросного инструментария и математической модели, позволяющей строить оценки интенсивности. Например, в качестве модели точечного процесса, контролирующего появление эпизодов на временной оси, может выступать процесс Пуассона [52]. В этом случае интенсивность поведения предполагается неизменной: каждый индивид в популяции имеет одну и ту же интенсивность поведения.

Чтобы учесть индивидуальные особенности, была предложена гамма–пуассоновская модель поведения [37, 49, 53]. В рамках этой модели предполагается, что эпизоды поведения каждого индивида контролируются процессом Пуассона, но интенсивность поведения в популяции варьируется между индивидами. В качестве априорного распределения вероятности было выбрано гамма–распределение, которое обладает рядом полезных математических свойств [40] (см. раздел 3). В работе [55] был предложен байесовский подход к оценке интенсивности, в рамках которого зависимость между интенсивностью

поведения (напрямую не наблюдаемой) и интервалами между эпизодами поведения (наблюдаемыми напрямую) представлялась байесовской сетью доверия (БСД) [56, 57]. Однако байесовская сеть доверия не может учесть ряд особенностей получаемой информации [38]. Гамма–пуассоновская модель поведения индивида опирается на *непрерывные* показатели (интенсивность, длина интервала), которые связаны между собой. Использование же классической байесовской сети доверия предполагает дискретизацию переменных. Во-первых, дискретизация непрерывных по своей сути переменных связана с потерей информации. Во-вторых, чувствительность модели напрямую связана с числом уровней дискретизации, но, чем выше число уровней дискретизации, тем больше становится размерность тензоров условных вероятностей в узлах сети и тем больше параметров требуется для задания БСД. В-третьих, возникают сложности при обращении к экспертной информации: эксперту может быть трудно выстраивать взаимосвязи для переменных после дискретизации, особенно если уровни дискретизации выбираются из вычислительных соображений, а не интуитивно. Более того, изменение структуры байесовской сети доверия ведет к пересчету всех значений тензоров условных вероятностей, что требует значительных статистических мощностей, не всегда доступных в практических приложениях (особенно связанных с получением информации от респондентов).

Научной целью данной статьи является построение оценки интенсивности поведения в гамма–пуассоновской модели в случае, если доступны длины двух интервалов между последовательными эпизодами поведения. Для достижения цели используется аппарат копул. Теоретические рассуждения сопровождаются примером на модельных данных. В статье также приведен краткий обзор возможностей использования копул для моделирования зависимостей в приложениях. Дидактическая цель статьи состоит в том, чтобы продемонстрировать на практическом примере возможности и способы использования аппарата копул для моделирования данных, полученных при измерении непрерывных переменных.

2. ЗАВИСИМОСТЬ СЛУЧАЙНЫХ ВЕЛИЧИН И КОПУЛЫ

Копулы — это такие функции, которые связывают два или более распределений вероятности одномерных случайных величин в совместное распределение. Вообще говоря, копулу можно определять с двух точек зрения: вероятностной и аналитической. С аналитической точки зрения копула представляет собой функцию, обладающую свойствами:

Определение 1 *n-мерной копулой называется функция* $C : [0,1]^n \to [0,1]$, *такая что*

1. *Для каждого вектора* $\mathbf{u} \in [0,1]^n$, *такого что некоторая его координата с индексом* $k \in \{1 \dots n\}$ *обращается в* 0, *функция C принимает значение* 0;

$$\forall \mathbf{u} \in [0,1]^n : \ \exists k \in \{1 \dots n\} \ u_k = 0 : \ C(\mathbf{u}) = 0.$$

2. *Для каждого вектора* $\mathbf{u} \in [0,1]^n$, *у которого координата с некоторым индексом* k *принимает значение, отличное от 1:* $u_k \neq 1$, *а все остальные координаты принимают значение* 1, *функция C принимает значение* u_k*:*

$$\forall \mathbf{u} \in [0,1]^n : \ \exists k \in \{1 \dots n\} \ u_k \neq 1, \forall j \neq k \ u_j = 1 : \ C(\mathbf{u}) = u_k.$$

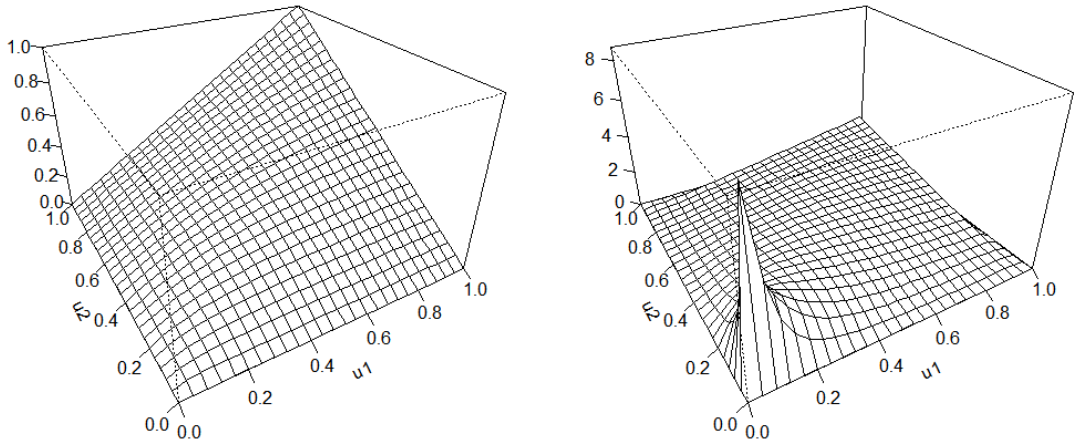
3. *(n–возрастание) Для каждых векторов* $\mathbf{a} = (a_1, \ldots, a_n)^T$ *и* $\mathbf{b} = (b_1, \ldots, b_n)^T \in [0,1]^n$*, таких что* $\mathbf{a} \le \mathbf{b}$ *(покоординатно), суперпозиция конечных разностей функции* $C(\mathbf{t})$ *по каждой координате больше или равна* 0*;*

$$\forall \mathbf{a}, \mathbf{b} \in [0,1]^n : \ \mathbf{a} \le \mathbf{b} : \ \forall \mathbf{t} \in [0,1]^n : \ \Delta_n^{b_n, a_n} \ldots \Delta_k b_k, a_k \ldots \Delta_1^{b_1, a_1} C(\mathbf{t}) \ge 0,$$

где $\Delta_k^{b_k, a_k}$ *есть оператор конечной разности в точках* a_k *и* b_k *по координате с индексом* $k \in 1 \ldots n$*:*

$$\forall C : [0,1]^n \to [0,1], \ \Delta_k^{b_k, a_k} C(\mathbf{t}) = C(t_1, \ldots, t_{k-1}, b_k, t_{k+1}, \ldots, t_n) - C(t_1, \ldots, t_{k-1}, a_k, t_{k+1}, \ldots, t_n).$$

Пример такой функции для случая $n = 2$ представлен на рисунке 1 а.



а) **б)**

Рис. 1. а) Копула–функция C (копула Клейтона с параметром 1.3), **б)** Плотность распределения копулы Клейтона с параметром 1.3

Функция $C(u, v)$ от двух переменных принимает значение 0 на осях координат: $C(0, v) = C(u, 0) = 0$. Если одна из координат обращается в 1, то функция C принимает значение, равное второй координате: $C(u, 1) = u$; $C(1, v) = v$. Свойство 3 из определения 1 копулы переписывается так ($u_1, u_2, t_1, t_2 \in [0,1]$, $u_2 > u_1$, $t_2 > t_1$):

$$C(u_2, t_2) - C(u_1, t_2) - C(u_2, t_1) + C(u_1, t_1) \ge 0.$$

Такие разности в [59] называются «смешанными разностями».

Копулы тесно связаны с триангулярными нормами, t-нормами, которые используются в нечеткой логике для описания операции нечеткой конъюнкции [3, 26, 56]. Точнее, t-норма является копулой тогда и только тогда, когда для нее выполнено свойство 2-возрастания 3 из определения копулы 1, а копула является t-нормой тогда и только тогда, когда выполнены свойства ассоциативности и коммутативности. Подобные функции применялись в исследовании поведения при описании взаимодействия атакующих действий злоумышленника и уязвимостей пользователя при моделировании социоинженерных атак [45, 46].

Заметим, что свойства 1 определяют совместную функцию распределения n случайных величин, и обратно: совместная функция распределения обладает перечисленными в 1 свойствами [59]. Таким образом, возникает вероятностное определение копулы:

Определение 2 *n-копула есть ограниченное на единичный куб* $[0,1]^n$ *совместное распределение вероятности равномерно распределенных на единичном отрезке случайных величин.*

Понятие копулы в приведенной выше постановке может казаться непригодным для практического моделирования данных, относящихся к различным типам распределений вероятности. Однако одно наблюдение позволило определить копулу для произвольных случайных величин с обратимым распределением вероятности: ведь преобразование $U = F(X)$ переводит случайную величину X с распределением вероятности F в равномерно распределенную случайную величину [59]. Основополагающая теорема теории копул, теорема Склара [26], устанавливает факт существования для случайных величин $X_1, X_2, \ldots, X_n$ с функциями распределения $F_1(x_1), F_2(x_2), \ldots, F_n(x_n)$ и совместной функцией распределения $H(x_1, x_2, \ldots, x_n)$ такой копулы C, что

$$H(x_1, x_2, \ldots, x_n) = C(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)). \tag{1}$$

В случае непрерывных маргинальных распределений такая функция C единственна. Верно и обратное утверждение: если C — копула и $F_1(x_1), F_2(x_2), \ldots, F_n(x_n)$ — функции распределения, то H — совместная функция распределения случайных величин $X_1, X_2, \ldots, X_n$.

Таким образом, копула является своего рода моделью зависимости переменных, позволяет разделить совместное распределение величин на часть, связанную с одномерными функциями распределения (маргиналами [26]), и часть, отвечающую за взаимосвязь маргинальных распределений (собственно функция копулы). Ниже перечислены основные свойства копул.

1. Копулы тесно связаны с инвариантными относительно монотонных преобразований мерами зависимости. Дело в том, что при монотонном преобразовании копула не изменяется или изменяется предсказуемым образом [26]. Это свойство важно при работе с переменными–физическими величинами, которые могут измеряться в разных единицах (например минуты, секунды…). К примеру, коэффициенты ранговой корреляции Кендалла и Спирмена являются независимыми относительно монотонных преобразований и определяются формально через копулу [26].
2. Для каждой копулы C и любых $(u, v) \in [0,1] \times [0,1]$ существуют минимальная $C^-(u, v) = \max(u + v - 1, 0)$ и максимальная копула $C^+(u, v) = \min(u, v)$, такие что:

   $$C^-(u, v) \le C(u, v) \le C^+(u, v).$$

   Такие границы для любой копулы носят названия границы Фреше–Хефдинга [26, 47].
3. Копула $C(u, v) = uv$ соответствует паре независимых случайных величин (U, V) [26].
4. В некоторых случаях существует такая непрерывная строго возрастающая функция $\phi : [0,1] \to [0, +\infty]$, $\phi(1) = 0$,что

   $$\phi(C(u, v)) = \phi(u) + \phi(v).$$

   В таком случае копула C носит название *архимедовой* [26]. Архимедовы копулы не только позволяют моделировать самые разнообразные формы зависимости (множество примеров можно найти в книге [26]), но и позволяют обращаться к любым свойствам и характеристикам этой взаимосвязи через генератор копулы — функцию ϕ. Архимедовы копулы легко обобщаются на многомерный случай.

2.1. Области применения копул

Движущей силой развития теории копул является применение их в областях экономики, связанных с управлением риском (страхование, банковское дело). К примеру, директива Solvency II, выпущенная Международной Ассоциацией Актуариев в 2004 году, рекомендует использовать копулы для моделирования взаимосвязанных рисков (портфелей). Как отмечается в [44], «копулы представляют собой исчерпывающий и гибкий инструмент для моделирования зависимостей». Интерес к копулам в области моделирования рисков и различных аггрегированных величин, таких как VaR, обусловлен значительными отклонениями от нормального закона распределения в данных [24]. Возможности применения копул к различным задачам в области управления рисками и финансов представлен в [5]. Обзор возможностей применения иерархических моделей на основе копул в финансовой сфере приведен в [1].

Другой сферой применения копул является моделирование данных времени жизни (survival data). Проблемы, связанные с анализом такого рода информации, возникают в целом ряде областей знаний: медицине, биологии, эпидемиологии, инженерных науках, анализе рисков. Данные времени жизни обладают рядом особенностей: они положительны, часто цензурированы. Модели, учитывающие индивидуальные уязвимости, (Frailty models) предполагают, что распределение времени жизни (и другие связанные характеристики, такие как функция угроз) мультипликативно зависят от случайной величины, учитывающей индивидуальные особенности [28, 40]. Два времени, наблюдаемые для одного пациента, часто бывают связаны: к примеру, в клинических исследованиях одна переменная может обозначать время до смерти пациента, а другая — время до возникновения некоторого события, как, например, рецидив или ухудшение состояния. В этом случае для оценки различных эпидемиологических показателей важно учитывать эту взаимосвязь; для этого используются копулы [18]. Другими примерами взаимосвязанных наблюдений в области медицины и эпидемиологии являются время работы каждого из парных органов тела или время выздоровления при первичном и повторном заболевании [23]. В подобных случаях также может использоваться подход на основе копул [23]. Копулы позволяют учитывать взаимосвязи в таблицах времени жизни, когда, к примеру, наблюдаются два связанных времени жизни: у родителей и у ребенка [13].

Модели на основе копул активно применяются в науках о земле, в частности, гидрологии [5]. Многие феномены в этой области описываются несколькими взаимозависимыми величинами: сила и продолжительность ливней [14, 16, 19, 42]; продолжительность, охват и интенсивность наводнения или засухи [15, 20, 29, 31, 34, 43], при этом каждая составляющая имеет свое распределение вероятности. Копулы применяются для моделирования совместного влияния нескольких источников на пиковые показатели уровня воды на дамбе [12, 15].

Другим разделом эпидемиологии и охраны общественного здоровья, в котором активно используются модели на основе копул, является мета-анализ исследований. Модели мета-анализа, собирающего воедино результаты нескольких клинических испытаний, порой опираются на анализ взаимозависимых велин: чувствительность и специфичность в мета-анализе точности диагностики [22, 27], новые конечные точки клинического исследования и исходы заболеваний [32]. Особую роль играют исследования, связанные с ВИЧ [35]. Так как распространение ВИЧ связано с поведением индивидов, то важным способом получения информации является интервьюирование индивидов. Данные, полученные при этом, могут обладать значительной неточностью и, в том числе, быть подвержены различным типам смещения: смещению припоминания, смеще-

нию социальной–желаемости ответа и др. В этом срезе возникает несколько задач, которые могут быть рассмотрены с позиции математического моделирования посредством копул: как оценить неточности при интервьюировании [4, 7], как работать с не точно измеренными данными [11].

Копулы используются в эпидемиологии для моделирования совместного поведения случайных величин, которые ассоциированы с распространением инфекционных заболеваний [2, 33]. Использование копул расширяет возможности применения существующих моделей распространения инфекций.

3. КОПУЛА ЗАВИСИМОСТИ ДЛИН ИНТЕРВАЛОВ МЕЖДУ ПОСЛЕДОВАТЕЛЬНЫМИ ЭПИЗОДАМИ

Предположим, что исследуется некоторый тип поведения, причем интенсивность поведения λ обусловлена индивидуальными особенностями каждого респондента и потому варьируется от пользователя к пользователю [37]. Априорное распределение интенсивности может быть известно в результате пилотных исследований. Однако этот шаг доступен не всегда. Поэтому в качестве предполагаемого распределения интенсивности поведения выбирается гамма–распределение вероятности $G(x)$ с параметрами формы $a > 0$ и масштаба $s > 0$ и плотностью вероятности

$$g(t) = \frac{1}{\Gamma(a)s^a} t^{a-1} \exp(-t/s). \tag{2}$$

Гамма–распределение вероятности выбирается, так как является гибким и часто используется в задачах моделирования. Ниже перечислены свойства гамма–распределения, которые играют основополагающую роль в этом выборе.

- Семейство гамма–распределений замкнуто относительно операции свертки, и, тем самым, гамма–распределение является самосопряженным. Это свойство важно при проведении байесовского вывода, позволяя не менять класс распределений при поступлении новых свидетельств, а ограничиться пересчетом параметра распределения [59].
- Гамма–распределение имеет простой вид преобразования Лапласа и потому удобно для использования в качестве смешивающего распределения [48].

Таким образом, чтобы описать интенсивность поведения в популяции, можно обратиться к оценкам параметров этого распределения: $\hat{s}$, $\hat{a}$. Имея эти оценки, можно вычислять описательные статистики, которые содержат в себе информацию о среднем значении интенсивности в популяции и могут применяться при принятии решений.

Как отмечалось, исследователю могут быть доступны различные наборы данных в силу особенностей памяти каждого индивида. Однако, как показывают пилотные исследования, последние три эпизода припоминаются легко. Три последних эпизода поведения определяют два интервала между последовательными эпизодами поведения и один особенный интервал между последним эпизодом поведения и моментом интервью. На рисунке 2 представлена диаграмма расположения рассматриваемых эпизодов и интервалов на временной оси.

В иллюстративных целях обратимся к модельным данным. Пусть параметры гамма–распределения интенсивности в популяции равны $a = 1.2$, $b = 3$. Сгенерируем выборку из $n = 10000$ наблюдений этой случайной величины. Далее для каждого значения λ_i, $i = 1 \dots n$ сгенерируем два значения экспоненциально распределенных случайных величин с параметром λ_i, соответствующие длинам интервалов τ_1 и τ_2.

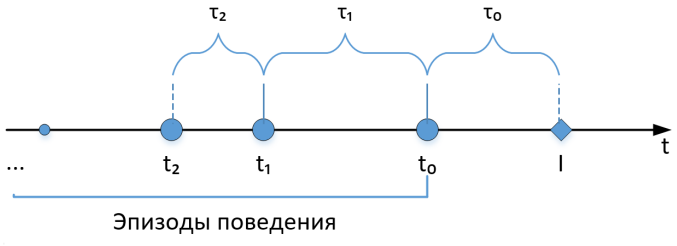


Рис. 2. Эпизоды поведения на временной оси и интервалы между ними. Здесь t_2, t_1, t_0 — моменты на временной оси, когда происходят эпизоды поведения; I — момент интервью; τ_2, τ_1 — длины интервалов между последовательными эпизодами поведения; τ_0 — длина интервала между последним эпизодом и моментом интервью [50]

Диаграмма рассеяния для полученных выборочных значений представлена на рисунке 3.

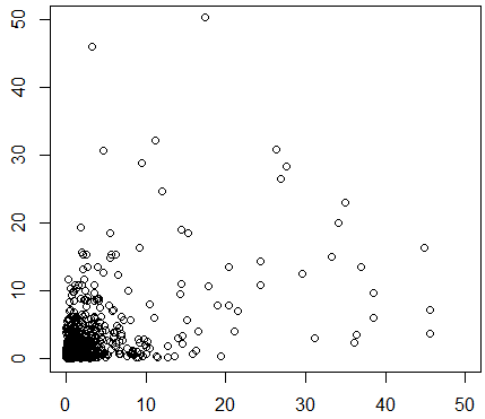


Рис. 3. Зависимость двух интервалов между последовательными эпизодами поведения в гамма–пуассоновской модели

Чтобы определить тип копулы, связывающей τ_1 и τ_2 в гамма–пуассоновской модели, выпишем их совместное распределение в форме $\bar{H}(x,y) = P[\tau_1 > x, \tau_2 > y]$. Такая форма распределения вероятности в приложениях (теория надежности, анализ рисков) называется *функцией выживаемости* или *функцией надежности*. Функция надежности обладает всеми свойствами функции распределения вероятностей в классическом виде. Итак,

$$P[\tau_1 > x, \tau_2 > y] = \int_0^\infty P[\tau_1 > x, \tau_2 > y | \lambda = t] g(t) dt =$$

$$= \int_0^\infty \exp(-tx)\exp(-ty)\frac{1}{\Gamma(a)s^a} t^{a-1}\exp(-t/s)dt =$$

$$= \frac{1}{\Gamma(a)s^a}\int_0^\infty t^{a-1}\exp(-(x+y+1/s)t)dt =$$

$$= \frac{1}{\Gamma(a)s^a}\frac{\Gamma(a)}{(x+y+1/s)^a}.$$

Таким образом,

$$P[\tau_1 > x, \tau_2 > y] = \frac{1}{(sx + sy + 1)^a}. \tag{3}$$

В случае $s = 1$ полученное распределение вероятности является *двумерным распределением Парето II типа* [5].

Рассмотрим копулу в так называемой форме survival copula [26]:

$$\hat{C}_{12}(u, v) = \bar{H}(\bar{G}_1^{(-1)}(u), \bar{G}_2^{(-1)}(v)),$$

где $\bar{G}_1(x) = \bar{H}(x, \infty)$ и $\bar{G}_2(y) = \bar{H}(\infty, y)$ — маргинальные распределения $\bar{H}$. Они имеют вид $\bar{G}_1(x) = (1 + sx)^{-a}, x > 0$. Квазиобратная функция [26] такого распределения имеет вид:

$$\bar{G}_1^{(-1)}(u) = \frac{u^{-1/a} - 1}{s}.$$

Тогда

$$\hat{C}_{12}(u, v) = (u^{-1/a} + v^{-1/a} - 1)^{-a}. \tag{4}$$

Чтобы определить параметры копулы 4 на модельных данных, сначала преобразуем выборочные переменные к равномерному распределению. На рисунке 4 представлена диаграмма рассеяния преобразованных случайных величин, дополненная контурами построенной копулы.

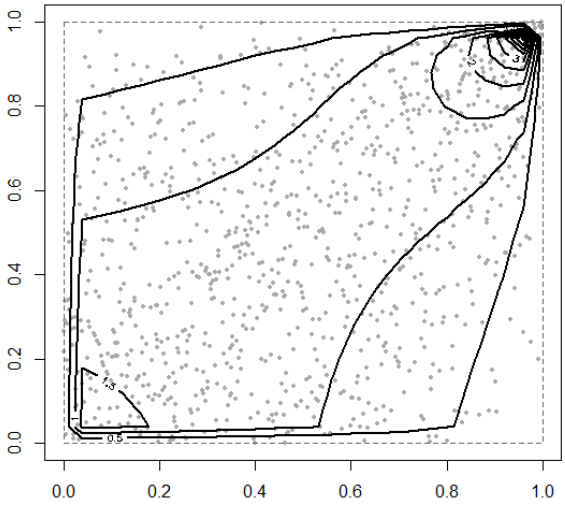


Рис. 4. Зависмость длин преобразованных к равномерному распределению интервалов и контурная диаграмма копулы их взаимосвязи

Оценка максимального правдоподобия параметра копулы Клейтона для модельных данных есть $\hat{a} = 1.23$ (95% доверительный интервал $1.18, 1.28$).

$\hat{C}_{12}(u, v)$ является *копулой Клейтона*[5, 26]. Заметим, что в 4 участвует лишь параметр a гамма–распределения 2. Это связано с тем, что собственно копула не зависит от монотонного преобразования случайных величин, является инвариантной к масштабу (scale–invariant)–подробнее понятие рассмотрено в параграфе 2), тем самым параметр масштаба s не участвует в 4.

Другой особенностью копулы Клейтона является возможность использования при построении многомерных моделей.

В нашем случае параметр копулы a может принимать значения $(0, +\infty)$. Отметим, что при $a \to 0$, копула Клейтона совпадает с копулой независимости, которая описывает случай независимых переменных: $\Pi(u, v) = uv$ [26]. Это свойство важно при построении вероятностных графичеких моделей на основе копул: непрерывных непараметрических сетей доверия (ННБСД) [21]. Графическая структура такой модели представляет собой направленный ациклический граф, вершины которого соответствуют непрерывным случайным величинам. Взаимосвязь переменных описывается типом копулы и величиной коэффициента ранговой корреляции между ними. Таким образом, чтобы вероятностная спецификация модели была корректной, необходимо, чтобы значение 0 коэффициента (условной) ранговой корреляции отвечало (условной) независимости переменных. Это свойство носит название ноль–независимости (zero independence) [21], и семейство копул Клейтона может быть расширено так, чтобы оно обладало этим свойством.

Также копула Клейтона может непосредственно быть использована для многомерного моделирования. Копула Клейтона является *архимедовой копулой* [26] с производящей функцией (generating function) $\phi(t) = t^{-1/a} - 1, t > 0$. При значениях параметра $a > 0$ можно выписать плотность копулы в явном виде [26]. Кроме того, для подкласса копул Клейтона со значениями параметра $a > 0$ возможно обобщение на многомерный случай: пусть $\mathbf{u} = (u_1, \ldots, u_n), n \geq 2$, тогда

$$C_\theta^n(\mathbf{u}) = (u_1^{-1/\theta} + u_2^{-1/\theta} + \ldots + u_n^{-1/\theta} - n + 1)^\theta, \theta > 0, n \geq 2.$$

Возможность такого обобщения важна при моделировании совместного распределения вероятности длин нескольких интервалов в гамма–пуассоновской модели.

Кроме того, даже в двумерном случае, вид копулы зависимости длин интервалов между последовательными эпизодами может прояснить некоторые моменты о природе их взаимосвязи. Во–первых, переменные, свзанные копулой Клейтона, обладают свойством PQD (Positive Quadrant Dependence):

$$P[\tau_1 \geq x, \tau_2 \geq y] > P[\tau_1 \geq x]P[\tau_2 \geq x].$$

Это свойство говорит о том, что вероятность того, что τ_1 и τ_2 одновременно принимают большие (малые) значения выше, чем если бы они были независимы. Можно сказать, что в гамма–пуассоновской модели эпизоды поведения *кластеризованы*: длинные интервалы встречаются чаще с длинными интервалами, а короткие — с короткими. Этот эффект объясняется наличием индивидуальной интенсивности поведения для каждого индивида. В пуассоновской модели поведения, в рамках которой интенсивность предполагается одинаковой для всех индивидов, длины интервалов встречаются совершенно случайно.

3.1. Компьютерное моделирование

Численное моделирование было произведено при помощи среды обработки данных R [60]. Использовались пакеты copula [61–63]. Ниже приведен код, который может использоваться для воспроизведения эксперимента.

```
#generate the sample of intensities
n<-10000
lambda <- rgamma(n, shape=1.2, rate=0.3)

#for each intensity value generate two
#exponentially distributed variables corresponding
#to inter-episode interval lengths
intervals <- data.frame(int1=double(), int2=double())

for (i in 1:n){
  intervals[i,] <- rexp(2, lambda[i])
}

library(copula)

#compute pseudo observations
CopIntervals <- pobs(as.matrix(intervals))

#fit the copula model
survClayton <- rotCopula(claytonCopula())
fittedC <- fitCopula(survClayton, CopIntervals, method="ml")

clayton <- rotCopula(claytonCopula(dim = 2, param = coef(fittedC)))

#summarize and plotting
summary(clayton)

plot(density(lambda))
tiff("plot1.tiff", height=2000, width=2000, res=300)

plot(CopIntervals[,1], CopIntervals[,2], pch=19, cex=0.2, col="dark grey",
     xlab="",
     ylab="")
contour(clayton, dCopula, col="black", lwd=2, add=T)

dev.off()
```

4. ЗАКЛЮЧЕНИЕ

Анализ современных экономических, социоэкономических процессов во многом опирается на многомерное моделирование взаимосвязанных величин. Среди множества техник, позволяющих осуществлять такое моделирование, особое место занимает аппарат копул. Являясь по своей сути одной из форм совместной функции распределения, копулы позволяют отделить задачи оценки маргинальных распределений вероятности от задачи оценки параметров структуры. Копулы успешно используются в

самых различных приложениях и, кроме того, имеют тесные связи с областью машинного обучения. Существуют вероятностные графические модели, в которых структура зависимостей описывается в терминах копул.

В статье представлена интерпретация гамма–пуассоновской модели поведения индивида с точки зрения копул. Особенность гамма–пуассоновской модели поведения заключается в том, что переменные–длины интервалов между последовательными эпизодами поведения– являются зависимыми, причем их общий параметр интенсивности поведения учитывает индивидуальные особенности и имеет гамма–распределение в популяции. В работе получен вид копулы, связывающей переменные–длины интервалов между последовательными эпизодами: копула Клейтона. Полученная копула позволяет раскрыть свойства гамма–пуассоновской модели: эпизоды в такой модели кластеризованы, длинные интервалы чаще встречаются с длинными, а короткие — с короткими. Кроме того, оценка параметра копулы, которая может быть произведена напрямую по данным о длинах интервалов между последовательными эпизодами поведения, ведет к оценке параметров гамма–распределения интенсивности поведения, что является основной целью математического моделирования поведения в этом случае.

В дальнейшем планируется рассмотреть взаимосвязи иных переменных гамма–пуассоновской модели поведения: последнего интервала между моментом интервью и первым эпизодом, рекордных интервалов.

Список литературы

1. *Aas K.* Pair-Copula Constructions for Financial Applications: A Review // Econometrics. 2016. № 4. P. 43. URL: https://doi.org/10.3390/econometrics4040043 (дата обращения 09.06.2018).
2. *El Adlouni S., Beaulieu C., Ouarda T. B., Gosselin P. L., Saint-Hilaire A.* Effects of climate on West Nile Virus transmission risk used for public health decision-making in Quebec // International journal of health geographics. 2007. № 6(1). P. 40. doi: 10.1186/1476-072X-6-40
3. *Alsina C., Schweizer B., Frank M. J.* Associative functions: triangular norms and copulas. Singapoure: World Scientific, 2006.
4. *Baek S.* A Copula-Based Method for Analyzing Bivariate Binary Longitudinal Data // Publicly Accessible Penn Dissertations 1564, 2016. URL: http://repository.upenn.edu/edissertations/1564 (дата обращения 09.06.2018).
5. *Balakrishnan N., Lai C. D.* Continuous bivariate distributions // Springer Science & Business Media, 2009. 684 p.
6. *Bargagli A. M., Hickman M., Davoli M., Perucci C. A., Schifano P., Buster M., Brugal T., Vicente J.* Drug-related mortality and its impact on adult mortality in eight European countries // The European Journal of Public Health. 2005. № 16(2). P. 198–202. doi: 10.1093/eurpub/cki168
7. *Bellamy S. L., Baek S., Troxel A. B., Ten Have T. R., Jemmott J. B.* A copula approach to estimate reliability: an application to self-reported sexual behaviors among HIV serodiscordant couples // Statistics and Its Interface. 2016. № 9(1). P. 57–67. doi: 10.4310/SII.2016.v9.n1.a6
8. *Bell D. C, Trevino R. A.* Modeling HIV Risk [Epidemiology] // JAIDS. 1999. Vol. 22(3). P. 280–287. doi: 10.1097/00126334-199911010-00010
9. *Bennett A. K.* Older age underwriting: frisky vs frail // Journal of Insurance Medicine. 2004. № 36(1). P. 74–83.
10. *Bobak M., Malyutina S., Horvat P., Pajak A., Tamosiunas A., Kubinova R., Simonova G., Topor-Madry R., Peasey A., Pikhart H., Marmot M.G.* Alcohol, drinking pattern and all-cause, cardiovascular and alcohol-related mortality in Eastern Europe // European Journal of Epidemiology. 2016. Vol. 31, № 1. P. 21–30. URL: https://doi.org/10.1007/s10654-015-0092-8 (дата обращения 09.06.2018).
11. *Cameron A.C., Li T., Trivedi P. K., Zimmer D .M.* Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts // The Econometrics Journal. 2004. № 7(2). P. 566–584. doi: 10.1111/j.1368-423X.2004.00144.x

12. *Chen L., Singh V. P., Shenglian G., Hao Z., Li T.* Flood coincidence risk analysis using multivariate copula functions // Journal of Hydrologic Engineering. 2011. № 17(6). P. 742–755. doi: 10.1061/(ASCE)HE.1943-5584.0000504
13. *Clayton D. G.* A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence // Biometrika. 1978. № 65(1). P. 141–151. doi: 10.1093/biomet/65.1.141
14. *De Michele C., Salvadori G.* A generalized Pareto № 108(D2). P. 1–11. doi: 10.1029/2002JD002534
15. *Favre A. C., El Adlouni S., Perreault L., Thiemonge A., Bobee B.* Multivariate hydrological frequency analysis using copulas // Water resources research. 2004. № 40(1). P. W01101. URL: doi:10.1029/2003WR002456 (дата обращения 09.06.2018).
16. *Grimaldi S., Serinaldi F.* Design hyetographs analysis with 3-copula function // Hydrological Scientific Journal. 2006. № 51(2). P. 223—238. doi: 10.1623/hysj.51.2.223
17. *Hargreaves J. R., Delany–Moretlwe S., Hallett T. B., Johnson S., Kapiga S., Bhattacharjee P., Dallabetta G., Garnett G. P.* The HIV prevention cascade: integrating theories of epidemiological, behavioural, and social science into programme design and monitoring // The Lancet HIV, 2016. № 3(7). P. e318–e322. URL: https://doi.org/10.1016/S2352-3018(16)30063-7 (дата обращения 09.06.2018).
18. *Jiang H., Fine J. P., Kosorok M. R., Chappell R.* Pseudo self-consistent estimation of a copula model with informative censoring // Scandinavian Journal of Statistics. 2005. № 32(1). P. 1–20. doi: 10.1111/j.1467-9469.2005.00412.x
19. *Kao S. C., Govindaraju R. S.* A bivariate frequency analysis of extreme rainfall with implications for design // Journal of Geophysical Research. 2007. № 112(D1). P. 113–119. doi: 10.1029/2007JD008522
20. *Karmakar S., Simonovic S. P.* Bivariate flood frequency analysis. Part 2: A copula-based approach with mixed marginal distributions // Journal of Flood Risk Management. 2009. № 2(1). P. 32–44. doi: 10.1111/j.1753-318X.2009.01020.x
21. *Kurowicka D., Joe H. (eds.)* Dependence modeling: vine copula handbook. World Scientific Publishing Co, 2011. 370 p. doi: 10.1142/7699
22. *Kuss O., Hoyer A., Solms A.* Meta-analysis for diagnostic accuracy studies: a new statistical model using beta-binomial distributions and bivariate copulas // Statistics in medicine, 2014. № 33(1). P. 17–30. doi: 10.1002/sim.5909
23. *Louzada F., Suzuki A. K., Cancho V. G., Prince F. L., Pereira G. A.* The long-term bivariate survival FGM copula model: an application to a brazilian HIV Data // Journal of Data Science. 2012. № 10(3). P. 511–535.
24. *Mikosch T.* Copulas: Tales and facts // Extremes. 2006. № 9. P. 3–20. URL: https://doi.org/10.1007/s10687-006-0015-x (дата обращения 09.06.2018).
25. *Mucha L., Stephenson J., Morandi N., Dirani R.* Meta-analysis of disease risk associated with smoking, by gender and intensity of smoking // Gender medicine. 2006. NN№ 3(4). P. 279–291. URL: https://doi.org/10.1016/S1550-8579(06)80216-0 (дата обращения 09.06.2018).
26. *Nelsen R. B.* An introduction to Copulas, second edition. Springer series in Statistics, 2006. 272 p. doi: 10.1007/0-387-28678-0
27. *Nikoloulopoulos A. K.* A mixed effect model for bivariate meta-analysis of diagnostic test accuracy studies using a copula representation of the random effects distribution // Statistics in medicine. 2015. № 34(29). P. 3842–3865. doi: 10.1002/sim.6595
28. *Oakes D.* Bivariate survival models induced by frailties // Journal of the American Statistical Association. 1989. № 84(406). P. 487–493. doi: 10.1080/01621459.1989.10478795
29. *Renard B., Lang M.* Use of a Gaussian copula for multivariate extreme value analysis: Some case studies in hydrology // Advances in Water Resources. 2007. № 30(4). P. 897–912. doi: 10.1016/j.advwatres.2006.08.001
30. *Sallis J. F., Owen N., Fotheringham M. J.* Behavioral epidemiology: a systematic framework to classify phases of research on health promotion and disease prevention // Annals of Behavioral Medicine. 2000. № 22(4). P. 294–298. doi: 10.1007/BF02895665
31. *Salvadori G., De Michele C.* On the use of copulas in hydrology: theory and practice // Journal of Hydrologic Engineering. 2007. № 12(4). P. 369–380. doi: 10.1061/(ASCE)1084-0699(2007)12:4(369)
32. *Shi Q., Sargent D. J.* Meta-analysis for the evaluation of surrogate endpoints in cancer clinical trials //

International journal of clinical oncology. 2009. № 14(2). P. 102–111. doi: 10.1007/s10147-009-0885-4
33. *Siettos C. I., Russo L.* Mathematical modeling of infectious disease dynamics // Virulence. 2013. № 4(4). P. 295–306. doi: 10.4161/viru.24041
34. *Shiau J. T., Wang H. Y., Chang T. T.* Bivariate frequency analysis of floods using copulas // Journal of American Water Resources Association. 2006. № 42(6). P. 1549–1564. doi: 10.1111/j.1752-1688.2006.tb06020.x
35. *Shih J. H., Louis T. A.* Inferences on the association parameter in copula models for bivariate survival data // Biometrics. 1995. № 51. P. 1384–1399. doi: 10.2307/2533269
36. *Shuper P. A., Joharchi N., Irving H., Rehm J.* Alcohol as a correlate of unprotected sexual behavior among people living with HIV/AIDS: review and meta-analysis // AIDS and Behavior. 2009. № 13(6). P. 1021–1036. doi: 10.1007/s10461-009-9589-z
37. *Tulupyev A., Suvorova A., Sousa J., Zelterman D.* Beta prime regression with application to risky behavior frequency screening // Statistics in medicine. 2013. № 32(23). P. 4044–4056. doi: 10.1002/sim.5820
38. *Uusitalo L.* Advantages and challenges of Bayesian networks in environmental modelling // Ecological modelling. 2007. № 203(3–4). P. 312–318. doi: 10.1016/j.ecolmodel.2006.11.033
39. *Warburton D. E., Nicol C. W., Bredin S. S.* Health benefits of physical activity: the evidence // Canadian medical association journal. 2006. № 174(6). P. 801–809. doi: 10.1503/cmaj.051351
40. *Wienke A.* Frailty models in survival analysis. CRC Press, 2010. 320 p.
41. *Zaba B., Slaymaker E., Urassa M., Boerma J. T.* The role of behavioral data in HIV surveillance // Aids. 2005. № 19. P. S39–S52. doi: 10.1097/01.aids.0000172876.74886.86
42. *Zhang L., Singh V. P.* Bivariate rainfall frequency distributions using Archimedean copulas // Journal of Hydrology (Amsterdam), 2007. № 332(1–2). Pp. 93–109. doi: 10.1016/j.jhydrol.2006.06.033
43. *Zhang L., Singh V. P.* Bivariate flood frequency analysis using the copula method // J. Hydrol. Eng. 2006. № 11(2). P. 150–164. doi: 10.1061/(ASCE)1084-0699(2006)11:2(150)
44. *International Actuarial Association* A global framework for insurer solvency assessment. IAA Insurer Solvency Assessment Working Party Research Report, 2004. URL: http://www.actuaries.org/LIBRARY/Papers/Global_Framework_Insurer_Solvency_Assessment-public.pdf (доступ 09.06.2018).
45. *Азаров А. А., Тулупьева Т. В., Суворова А. В., Тулупьев А. Л., Абрамов М. В., Юсупов Р. М.* Социоинженерные атаки: проблемы анализа. СПб.: Наука, 2016. 349 с.
46. *Абрамов М. В.* Методы и алгоритмы анализа защищенности пользователей информационных систем от социоинженерных атак: оценка параметров моделей: дис. на соискание степени канд. тех. наук. Санкт-Петербург, СПИИРАН, 2018. 232 с. URL: http://www.spiiras.nw.ru/dissovet/abramov/ (доступ 09.06.2018).
47. *Благовещенский Ю. Н.* Основные элементы теории копул // Прикладная эконометрика. 2012. № 2(26). С. 113–130.
48. *Гнеденко Б. В.* Курс теории вероятностей. 8-е издание. М.: Едиториал УРСС, 2005. 448 с.
49. *Зельтерман Д., Суворова А. В., Пащенко А. Е., Мусина В. Ф., Тулупьев А. Л., Тулупьева Т. В., Красносельских Т. В., Гро Л. Е., Хаймер Р.* Обработка систематической ошибки, связанной с длиной временных интервалов между интервью и последним эпизодом в гамма-пауссоновской модели поведения // Труды СПИИРАН. 2011. Вып. 16. С. 160–185.
50. *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. Вып. 3. Т. 2. С. 257–268.
51. *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. Вып. 2. С. 32–35.
52. *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 // Труды СПИИРАН. 2012. 4(23). С. 157–184.
53. *Столярова В. Ф.* Виды и свойства попарной зависимости длин интервалов между последовательными эпизодами в пуассоновской и гамма–пуассоновской моделях поведения // IV Международная летняя школа–семинар по искусственному интеллекту для студентов, аспиран-

тов, молодых ученых и специалистов интеллектуальные системы и технологии: современное состояние и перспективы–2017 (ISYT–2017) г. Санкт-Петербург, 30 июня – 3 июля, 2017 г. Сборник научных трудов. С. 160–167.
54. *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.
55. *Суворова А. В., Тулупьев А. Л., Сироткин А. В.* Байесовские сети доверия в задачах оценивания интенсивности рискованного поведения // Нечеткие системы и мягкие вычисления, 2014. Т. 9. № 2. С. 115–129.
56. *Тулупьев А. Л., Николенко С. И., Сироткин А. В.* Байесовские сети: логико-вероятностный подход. СПб.: Наука, 2006. 607 с.
57. *Тулупьев А. Л., Сироткин А. В., Николенко С. И.* Байесовские сети доверия: логико-вероятностный вывод в ациклических направленных графах. СПб.: Изд-во С.-Петерб. ун-та, 2009. 400 с.
58. *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, 2008. 140 с.
59. *Феллер В.* Введение в теорию вероятностей и её приложения. Том 1, 2. М.: Мир, 1984. 528 с., 752 с.
60. *R Core Team.* R: A language and environment for statistical computing, 2018. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (доступ 09.06.2018).
61. *Hofert M., Kojadinovic I., Maechler M. Jun Yan* copula: Multivariate Dependence with Copulas. R package version 0.999-18. 2017. URL: https://CRAN.R-project.org/package=copula (дата обращения 09.06.2018).
62. *Jun Yan* Enjoy the Joy of Copulas: With a Package copula // Journal of Statistical Software. 2017. № 21(4). P. 1-21. URL: http://www.jstatsoft.org/v21/i04/ (дата обращения 09.06.2018). doi: 10.18637/jss.v021.i04
63. *Kojadinovic I., Jun Yan* Modeling Multivariate Distributions with Continuous Margins Using the copula R Package // Journal of Statistical Software. 2010. № 34(9). P. 1–20. URL: http://www.jstatsoft.org/v34/i09/ (дата обращения 09.06.2018).

Поступила в редакцию 04.05.2018, окончательный вариант — 09.06.2018.

Computer tools in education, 2018
№ 3: 22–37
http://ipo.spb.ru/journal
doi:10.32603/2071-2340-3-22-37

COPULAE AND DEPENDENCE MODELLING: INDIRECT ESTIMATES OF A PERSON'S RISKY BEHAVIOR INTENSITY

Stoliarova V. F.

SPIIRAS, Saint Petersburg, Russia

Abstract

In conditions of limited resources, the most affordable way to obtain information about the behavior of an individual is an interview. In this case, data on the individual's recent behavior are less susceptible to various types of bias. Mathematical models of behavi-

or are used to estimate the intensity of behavior when only the data on the last three episodes are available. In the paper we consider the gamma Poisson model of behavior. The dependence of two intervals between successive episodes in this model is described in terms of copulae. Estimation of the copula parameter in this case directly leads to an estimate of the parameters of the intensity distribution in the population. In addition, knowledge of the copula type allows one to reveal some characteristics of the behavior described by the gamma Poisson model: episodes of such behavior are clustered. The possibilities of the proposed approach are demonstrated on model data.

Keywords: *gamma–Poisson model of behavior, last episodes, copula.*

Citation: V. F. Stoliarova, "Copulae and Dependence Modelling: Indirect Estimates of a Person's Risky Behavior Intensit," *Computer tools in education*, no. 3, pp. 22–37 (in Russian).

Acknowledgements: *The work is financially supported by the governmental contract SPIIRAS number № 0073-2018-0001, RFBR grant №18-01-00626, St.Petersburg Committee on Science and Higher Education grant for PhD students in 2018.*

Received 04.05.2018, the final version — 09.06.2018.

Valeriia F. Stoliarova, junior researcher SPIIRAS, valerie.stoliarova@gmail.com

Столярова Валерия Фуатовна,
младший научный сотрудник лаборатории теоретических и междисциплинарных проблем информатики СПИИРАН,
valerie.stoliarova@gmail.com