

МОДЕЛИРОВАНИЕ СОЦИАЛЬНО- ЗНАЧИМОГО ПОВЕДЕНИЯ РЕСПОНДЕНТОВ ПО ДАННЫМ О ЕГО ОСОБЫХ ЭПИЗОДАХ

Алёна Владимировна СУВОРОВА, м.н.с.

Лаборатория ТуМПИ, СПИИРАН

suvalv@mail.ru

План доклада

- Постановка проблемы
- Исходные данные
- Вероятностный подход к оцениванию интенсивности
 - По последним эпизодам
 - По рекордным интервалам
 - Обработка неполных и неточных данных
 - Обобщенная модель
- Применение вероятностных графических моделей
 - Описание модели
 - Преимущества подхода
 - Расширение модели
- Разработанное программное обеспечение
- Гранты и проекты

Постановка проблемы

Рискованное поведение

- Эпидемиология

Угрожающее поведение

- Социо-инженерные атаки
- Манипулирование оператором

Исходные данные



Опросы об эпизодах поведения

Последние эпизоды поведения

- *«Когда в последний раз Вы делали ...?»*

Рекордные интервалы

- *«Каким был минимальный интервал между эпизодами... за последние 6 месяцев?»*
- *«Каким был максимальный интервал между эпизодами... за последние 6 месяцев?»*

- ✓ стабильность воспроизведения
- ✓ количественные значения
- ✓ ответы выражены на естественном языке

Цель исследований

- развитие методологии представления, агрегирования и обработки данных и знаний (полученных из самоотчетов респондентов) в условиях информационного дефицита для последующего формирования и расчета косвенных оценок интенсивности социально-значимого поведения.

Вероятностный подход

- Поведение рассматривается как случайный процесс некоторого класса



- Пуассоновский процесс

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

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

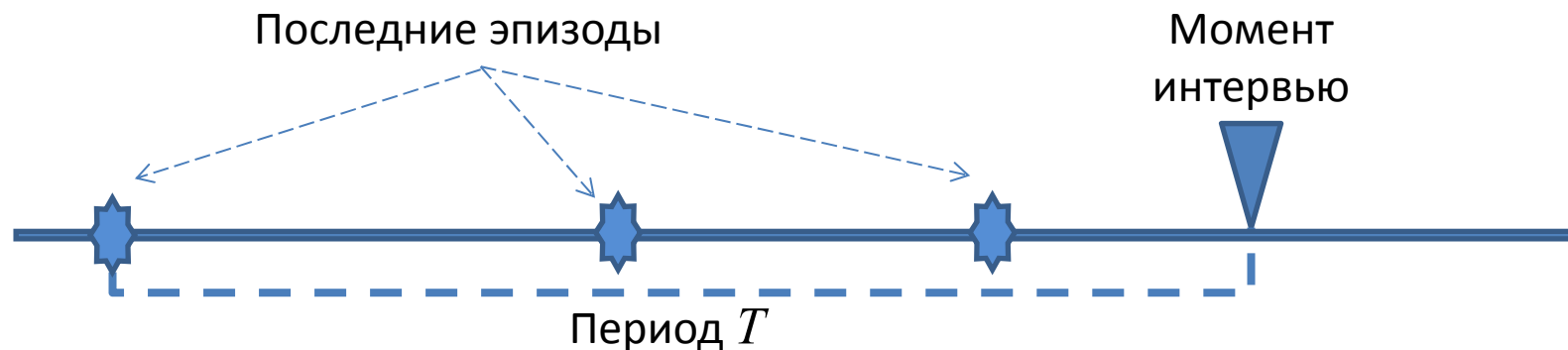
- Модификации пуассоновского процесса
- Эмпирические распределения

По последним эпизодам

Оценка методом максимального правдоподобия
(пуассоновский процесс)

$$\hat{\lambda} = \frac{k}{T}$$

k — число последовательных событий, которые вспомнил респондент, а T — тот период времени, за который эти эпизоды произошли



Последний интервал

- Неявно делается предположение, что момент интервью является эпизодом поведения
- Возникновение систематической ошибки
- Рассмотрены несколько подходов, позволяющих избежать такого предположения
 - модификация функции правдоподобия
 - коррекция вероятностного распределения для длины последнего интервала
 - рассмотрение распределения особого вида

Коррекция

$$p(t_0 = \tau_0) = \int_{\tau_0}^{\infty} \frac{p(T)}{T} dT$$

С учетом исправленной вероятности вычисление оценки интенсивности методом максимального правдоподобия сводится к решению уравнения

$$3 - \lambda(\tau_1 + \tau_2) - \frac{e^{-\lambda\tau_0}}{\left(\int_{\tau_0}^{\infty} \frac{e^{-\lambda u}}{u} du \right)} = 0$$

- Уравнение имеет единственное решение
- Знаменатель раскладывается в ряд:

$$\int_{\tau_0}^{\infty} \frac{e^{-\lambda u}}{u} du = E_1(\lambda\tau_0) = -\gamma - \ln(\lambda\tau_0) + \sum_{k=1}^{\infty} \frac{(-1)^{k+1} (\lambda\tau_0)^k}{k \cdot k!}$$

- Значение интенсивности можно найти с помощью численных методов

Численные эксперименты

Таблица. Сравнение оценок интенсивности

$\tau_1 + \tau_2$	τ_0	λ	$\lambda_{\max}$
100.0	20.0	0.0203611082700806	0.025
100.0	1.0	0.0268266111266637	0.0297029702970297
80.0	1.0	0.0332707039235333	0.0370370370370370
60.0	1.0	0.0438485906639061	0.0491803278688524
40.0	1.0	0.0644782603076189	0.0731707317073170
20.0	3.0	0.1075219188260833	0.1304347826086956
20.0	1.0	0.1230022468071118	0.1428571428571428
10.0	1.0	0.2286214963054978	0.2727272727272727
2.0	5.0	0.3228036964977513	0.4285714285714285
5.0	10.0	0.1512060697127039	0.2
2.0	15.0	0.1314253049599635	0.1764705882352941
5.0	15.0	0.1126843708081379	0.15

Бета-распределение

Наблюдение

- чем длиннее интервал между эпизодами, тем более вероятно, что момент интервью попадет в этот (более длинный) интервал

Длина интервала между последним эпизодом и моментом интервью имеет следующее распределение (принадлежащее к классу бета-простых распределений)

$$p(t) = Kt \int_0^{\infty} \lambda e^{-t\lambda} g(\lambda; \sigma, \eta) d\lambda,$$

где K — нормирующая константа, σ , η — параметры, характеризующие интенсивность.

Функция правдоподобия имеет вид:

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

По рекордным интервалам

- **Максимальный** и **минимальный** интервалы между эпизодами рискованного поведения за заданный период времени рассматриваются как члены вариационного ряда, что позволяет воспользоваться аппаратом теории рекордов
- Плотности распределения соответствующих порядковых статистик и их сочетаний имеют вид

$$f_{\min}(t) = n\lambda e^{-n\lambda t}, \quad -\infty < t < +\infty$$

$$f_{\max}(t) = n\lambda e^{-\lambda t} (1 - e^{-\lambda t})^{n-1}, \quad -\infty < t < +\infty$$

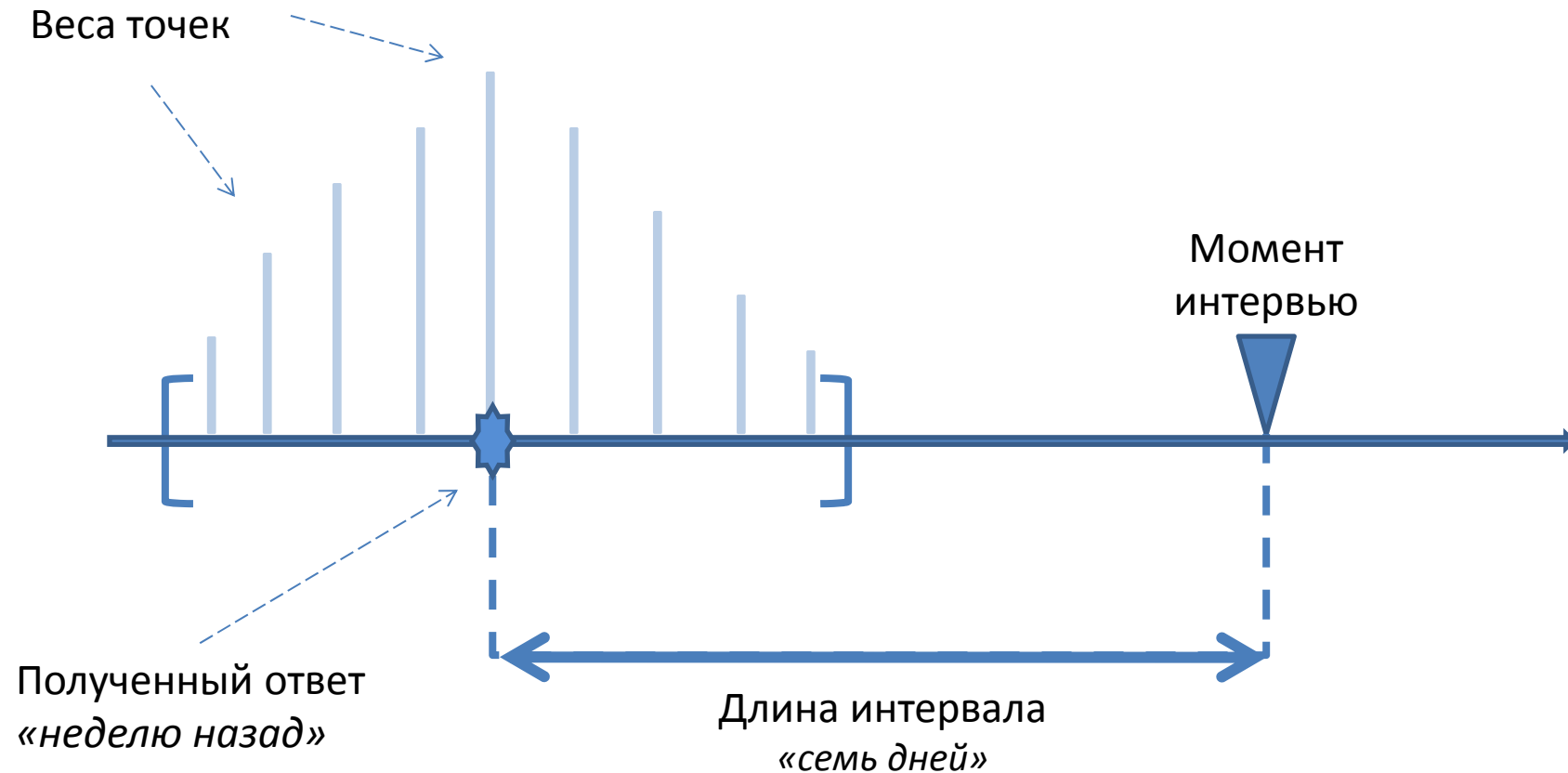
$$f_{\min, \max}(x, y) = \frac{n!}{(n-2)!} \lambda^2 e^{-\lambda(x+y)} (e^{-\lambda x} - e^{-\lambda y})^{n-2}$$

- Оценка интенсивности вычисляется методом максимального правдоподобия

Неполнота и неточность ответов

- Высказывания на естественном языке
 - Влияние единиц измерения
 - «семь дней назад» vs «неделю назад»
- ➡ Рандомизация неопределенности ответа

Обработка неопределенности



Обобщенная модель

- Предложен подход к построению модели, учитывающей одновременно данные о последних эпизодах поведения и сведения о рекордных интервалах
- Поведение рассматривается как пуассоновской процесс
- Метод максимального правдоподобия для оценки интенсивности поведения

Функция правдоподобия

$$L_{1,\min,\max}(\lambda) = \sum_{n=2}^{\infty} \frac{e^{-\lambda I} (\lambda I)^n}{n!} \frac{\partial^3}{\partial x \partial y \partial z} [F_{1,n}^n(x, y; \lambda) - \int_{\Omega(x,y)} V_{u,v;\lambda}^{*(n-2)}(I - z) f_{1,n}^n(u, v; \lambda) du dv] |_{\tilde{\tau}_0^*, \tilde{\tau}_{\min}, \tilde{\tau}_{\max}}. \quad (23)$$

здесь $f_{1,n}^n(u, v; \lambda)$ — совместная плотность распределения порядковых статистик (16) с номерами $1, n$, $F_{1,n}^n(x, y; \lambda)$ — соответствующая функция распределения, $V_{u,v;\lambda}^{*(n-2)}(I - z)$ — $(n - 2)$ -кратная свёртка функций распределения (19), множество $\Omega(x, y) = \{0 \leq u \leq x, u < v \leq y, x < y\}$.

Следующий этап

Необходимо

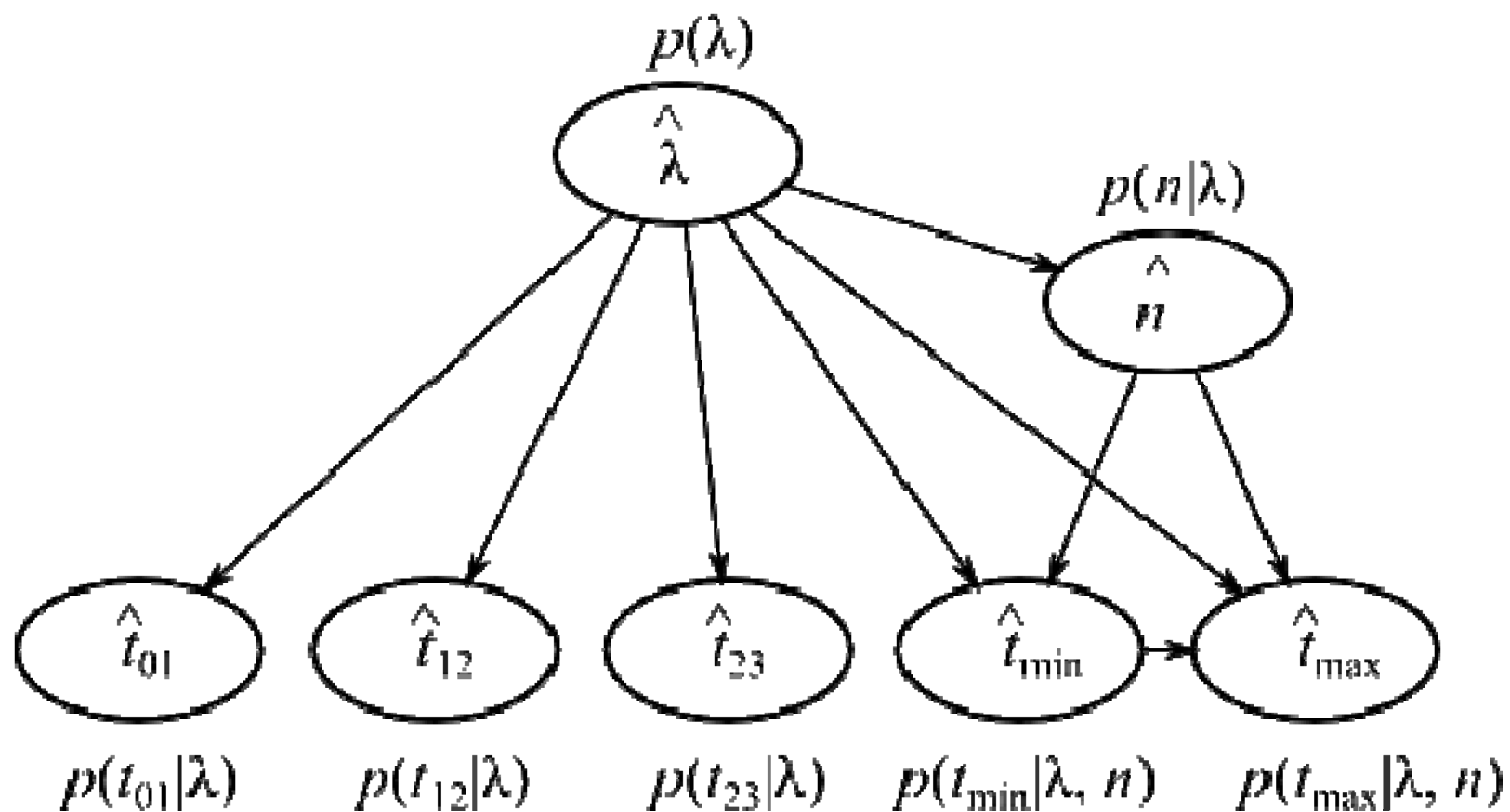
- Исследовать аспекты сходимости ряда
- Развить процедуры и методы вычисления полученного выражения
- Добавить обработку неточности исходных данных
- Развить процедуры поиска экстремума

Рассмотрение в качестве модели поведения других случайных процессов потребует повторения всех шагов

Альтернативный подход

- Задачу об оценке интенсивности социально-значимого поведения удобно свести через ряд промежуточных шагов к разработке особой **вероятностной графической модели** класса байесовских сетей доверия

Описание модели



Формальная спецификация

$$1) \quad p\left(t_{j,j+1}^{(l_j)} | \lambda\right) = e^{-a\lambda} - e^{-b\lambda}, \text{ где } t_{j,j+1}^{(l_j)} = [a; b), j = 1, 2, l_j = 1, \dots, k_j$$

$$2) \quad p\left(t_{01}^{(l_0)} | \lambda\right) = e^{-a\lambda} (a\lambda + 1) - e^{-b\lambda} (b\lambda + 1), t_{01}^{(l_0)} = [a; b)$$

$$3) \quad p(n | \lambda) = \frac{(\lambda T)^n}{n!} e^{-\lambda T}$$

$$4) \quad p\left(t_{\min}^{(l_3)} | n, \lambda\right) = \int_a^b n\lambda e^{-\lambda t} dt = n\left(e^{-a\lambda} - e^{-b\lambda}\right), t_{\min}^{(l_3)} = [a; b)$$

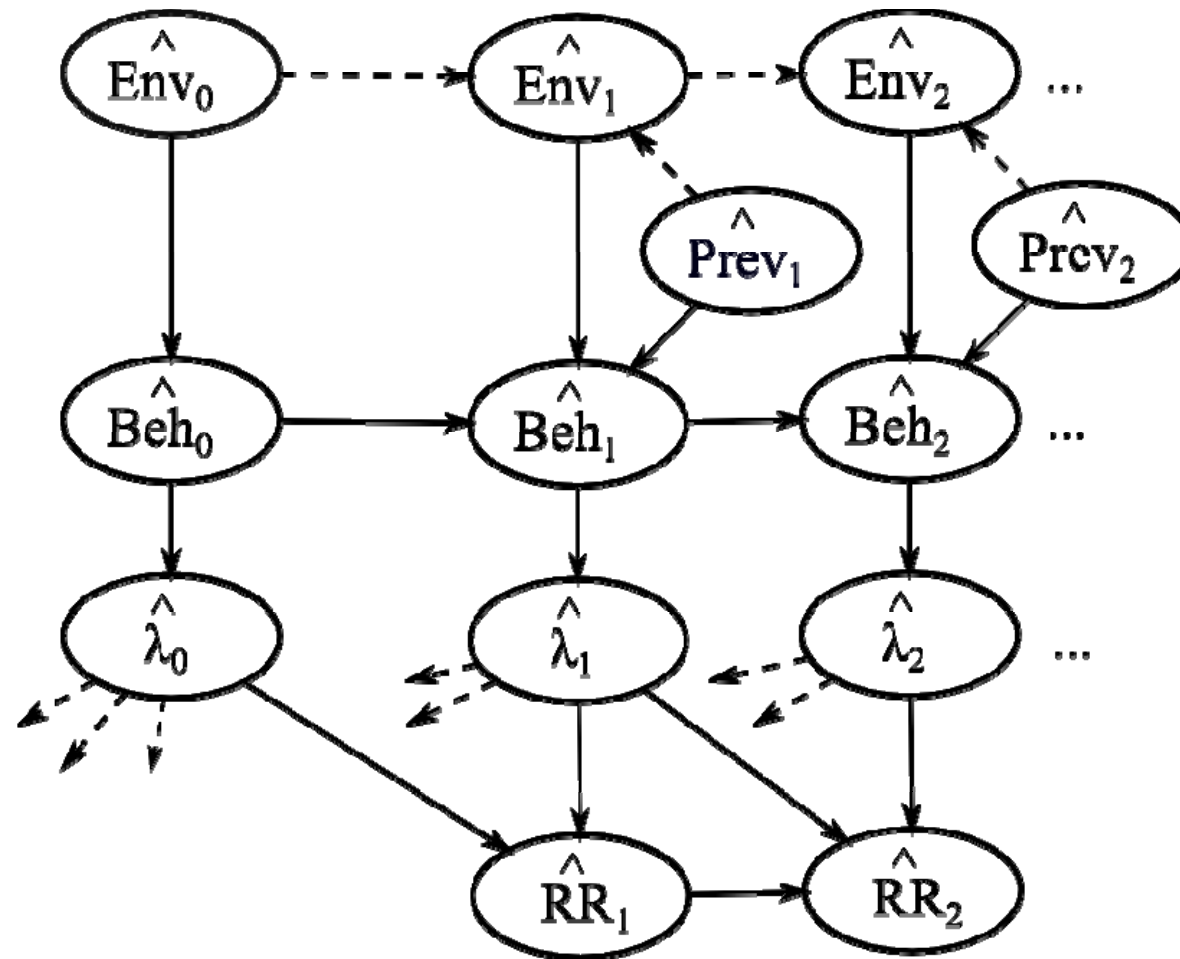
$$5) \quad p\left(t_{\max}^{(l_4)} | n, \lambda, t_{\min}\right) = \int_a^b f\left(t | n, \lambda, t_{\min}\right) dt, t_{\max}^{(l_4)} = [a; b),$$

$$\text{где } f\left(t_{\max} | n, \lambda, t_{\min}\right) = c(n-1)\lambda e^{-\lambda(t_{\max} - (n-1)t_{\min})} \left(e^{-\lambda t_{\min}} - e^{-\lambda t_{\max}}\right)^{n-2}$$

Преимущества

- Использование существующего алгоритмического аппарата
- Свободно распространяемый программный инструментарий для проведения вычислительных экспериментов и для применения в практических целях

Расширение модели



Расширение модели

- Влияние внешних факторов
- Вычисление относительных оценок
- Оценивание эффективности превенций
- Мониторинг и прогнозирование

Программный комплекс: по последним эпизодам

Fuzzy Risk-&-Rate Calculator

File Help

Абсолютные оценки по последним эпизодам Относительные оценки по последним эпизодам

Разъединить Риск за один эпизод 0.0025 Период в днях 183 Параметры

Ввести априорные данные Ввести апостериорные данные

АПОСТЕРИОРНЫЕ ДАННЫЕ

id	№ визита	тип	ответ 1	ед.измер...	кол.эпиз...	ответ 2	ед.измер...	кол.эпиз...	ответ 3	ед.измер...	кол.эпиз...
5	2	2	4	1	2	2	1	1	4	1	1

Ввести ответы самостоятельно Тип ответа смешанные Кол. эпизодов Значение Ед. измерения

Добавить новый ответ в базу данных № визита 2

Первый интервал 2 4 дни
Второй интервал 1 2 дни
Третий интервал 1 4 дни

Оценки интенсивности

априорная 0.333583875263 отклонение 0.009152883948
апостериорная 0.400243437027 отклонение 0.009880354832

0 0.1 0.2 0.3 0.4 0.5 Радиус (отклонений): 3
■ априорная оценка
■ апостериорная оценка

Оценки кумулятивного риска заражения

априорная 0.141531985859 отклонение 0.003593817240
апостериорная 0.167316081047 отклонение 0.003762965338

0 0.1 0.2 Радиус (отклонений): 3
■ априорная оценка
■ апостериорная оценка

Вычислить относительные оценки

Оценка интенсивности

относительная 1.2007303110822152
отклонение 0.044281079187896887

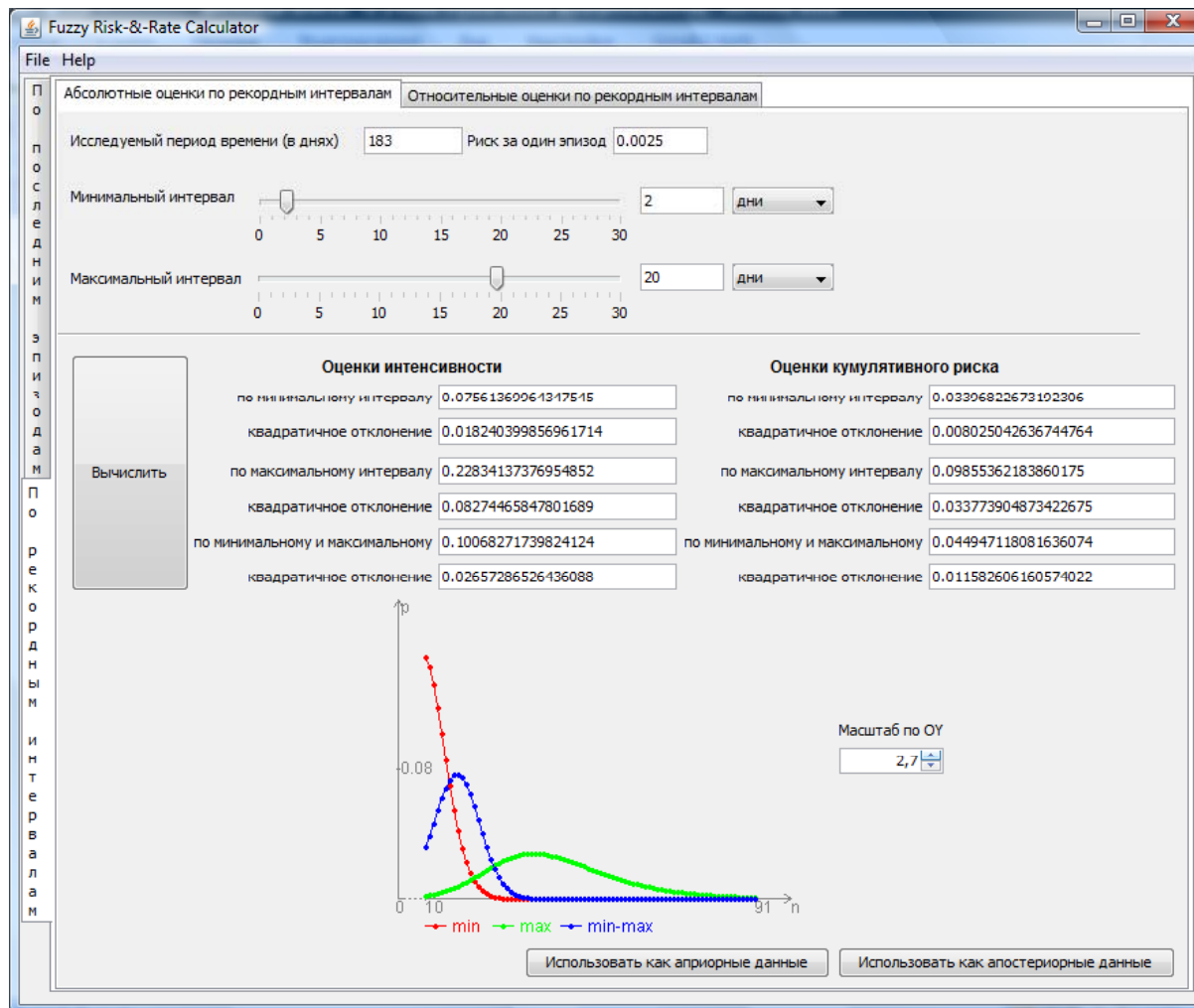
0 0.5 1 1.5 Радиус (отклонений): 3

Оценка кумулятивного риска заражения

относительная 1.1829397496286274
отклонение 0.040097042066404805

0 0.5 1 1.5 Радиус (отклонений): 3

Программный комплекс: по рекордным интервалам



Гранты и проекты

- «Alcohol and HIV Risk Reduction in St. Petersburg, R.F.» (субконтракт №М13А11589 (А06995) с Йельским университетом, США)
- Проект «Модели и алгоритмы анализа сверхкоротких неточных временных рядов на основе гранулярных данных и знаний» (Грант Правительства Санкт-Петербурга)
- «Развитие теории алгебраических байесовских сетей и родственных им логико-вероятностных графических моделей систем знаний с неопределенностью» (Грант РФФИ на 2012–2014 гг., проект № 12-01-00945-а)

МОДЕЛИРОВАНИЕ СОЦИАЛЬНО- ЗНАЧИМОГО ПОВЕДЕНИЯ РЕСПОНДЕНТОВ ПО ДАННЫМ О ЕГО ОСОБЫХ ЭПИЗОДАХ

Алёна Владимировна СУВОРОВА, м.н.с.

Лаборатория ТуМПИ, СПИИРАН

suvalv@mail.ru