

СРАВНЕНИЕ ПАРАМЕТРОВ УГРООБРАЗУЮЩЕГО ПОВЕДЕНИЯ В РАЗНЫХ ГРУППАХ НА ОСНОВЕ НЕПОЛНЫХ И НЕТОЧНЫХ ДАННЫХ

ПАЩЕНКО А.Е., ТУЛУПЬЕВ А.Л., СУВОРОВА А.В., ТУЛУПЬЕВА Т.В.

УДК 311.2 + 616-036.22

Пащенко А.Е., Тулупьев А.Л., Суворова А.В., Тулупьева Т.В. Сравнение параметров угрообразующего поведения в разных группах на основе неполных и неточных данных.

Аннотация. В статье представлен способ оценки отношения показателей интенсивности поведения в двух группах. Сравнение может производиться как в одной группе, но рассмотренной до и после поведенческой интервенции, либо между двумя различными группами. Оценка интенсивности рискованного поведения может быть получена на основе либо данных о последних эпизодах рассматриваемого рискованного поведения, либо данных о минимальном и максимальном интервале между эпизодами рискованного поведения за определенный промежуток времени. Рассмотрены вопросы согласования оценок, полученных различным путем.

Ключевые слова: мягкие вычисления, дефицит информации, угрообразующее поведение, современная эпидемиология.

Paschenko A.E., Tulupyev A.L., Syvorova A.V., Tulupyeva T.V. An approach to comparison of threatening behavior parameters between social groups based upon incomplete and imprecise date.

Abstract. The paper proposes a way to estimate risky behavior rates ratio in two groups. These groups can be either the same group but considered a priori and a posteriori a behavioral intervention, or two groups/two samples from different populations to be compared simultaneously. A similar estimation is proposed to cumulative risk ratio (odds ratio). The risky behavior rate can be estimated either with data about several last episodes of the behavior or with the data on extreme intervals between the episodes during a certain period of time. For these two estimations may be different; a measure of their consistency is introduced.

Keywords: soft computations, lack of information, threatening behavior, modern epidemiology.

1. Введение. В ряде отраслей научных исследований стоит задача оценки интенсивности поведения респондентов по неполным и неточным исходным данным. Источниками таких данных зачастую являются высказывания респондента на естественном языке, когда такие высказывания — единственный приемлемый способ получения сведений об интенсивности поведения. Полученная оценка интенсивности, как правило, используется дальше для косвенного оценивания других показателей. В частности, в случае исследования рисков приобретения или передачи ВИЧ-инфекции, на основе оценки интенсивности конкретного вида ВИЧ-рискованного поведения и вероятности приобретения ВИЧ-инфекции за один эпизод можно оценить кумулятивную

вероятность приобретения ВИЧ-инфекции за заданный период времени. Полученная оценка кумулятивной вероятности приобретения ВИЧ-инфекции позволяет перейти к ключевым показателям, характеризующим эпидемиологическую ситуацию в интересующий исследователя популяции.

Под интенсивностью понимается число эпизодов поведения рассматриваемого вида в определенный промежуток времени [1]. Для получения данных о числе эпизодов и о рассматриваемом промежутке производится опрос респондента об одном или нескольких последних эпизодах его поведения [2]. Такой опрос позволяет судить о непрерывных количественных величинах — интервалах между эпизодами, а также об интервале между временем опроса и последним эпизодом.

В результате интервью становятся известными сведения о нескольких последних эпизодах поведения, о максимальном T_{\max} , минимальном T_{\min} и обычном $T_{\text{обычное}}$ интервале между эпизодами.

Требуется оценить *априорные* (до вмешательства) λ_0 , Pr_0 и *апостериорные* (после вмешательства) λ_1 , Pr_1 — оценки интенсивности и кумулятивного риска соответственно, а также отношения этих оценок $\frac{\lambda_1}{\lambda_0}$, $\frac{\text{Pr}_1}{\text{Pr}_0}$, характеризующие влияние поведенческой интервенции на индивидов.

Кроме того, требуется предложить подходы к оценке степени «сочетаемости» (непротиворечивости) двух видов исходных данных.

2. Рандомизация для представления неточности в естественно-языковых высказываниях. Пусть известны данные о N последних эпизодах поведения $t_1, t_2, t_3, \dots, t_N$ а τ — общий временной промежуток, за который произошли эпизоды (связь между t_i и τ подробно рассмотрена в [1]). Тогда интенсивность поведения λ оценивается по формуле: $\lambda = \frac{N}{\tau}$ [1].

Для каждого эпизода со значением t_i , $1 \leq i \leq N$ (N — число рассматриваемых эпизодов поведения) через характеристику разброса δ определяется интервал (*возможных значений*) в днях: $[t_i - \delta x, t_i + \delta x]$, где x — коэффициент перевода рассматриваемой единицы измерения в дни [1] (см. табл.).

То есть оценка в днях включает в себя минимальное значение $t_i - \delta x$, указанное в интервью значение t_i и максимальное значение $t_i + \delta x$.

Табл. Точность и единицы измерения

Единица измерения	Значение x
часы	1/24
дни	1
недели	7
месяцы	30
полугоды	183
годы	365

Заметим, что любая точка из интервала $[t_i - \delta x, t_i + \delta x]$ возможна в качестве значения оценки t_i ; это, однако, не означает, что точки из этого интервала равновероятны в качестве такового.

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

Введенная случайная величина \hat{t}_i за счет рандомизации неопределенности ответа позволяет рассмотреть интенсивность как случайную величину и вычислить характеристики последней.

Для каждого эпизода соответствующий интервал разбивается на n равных частей. Рассматриваются все возможные сочетания точек из интервалов, соответствующих указанным эпизодам. Расчет среднего значения для случая трех последних эпизодов производится по следующей формуле:

$$\lambda_{\text{среднее}} = \sum_{i,j,k} (\lambda_{ijk} p_i p_j p_k),$$

где p_i — вес i -й точки из первого интервала, p_j — вес j -й точки из первого интервала, p_k — вес k -й точки из первого интервала, λ_{ijk} — оценка интенсивности для соответствующего сочетания точек, т.е. $\lambda_{ijk} = \frac{N}{\tau_{ijk}}$, где τ_{ijk} — соответствующая точкам i, j, k оценка величины рассматриваемого интервала.

Среднее квадратичное отклонение σ для рассчитываемого среднего значения:

$$\sigma = \sqrt{\sum_{i,j,k} \lambda_{ijk}^2 P_i P_j P_k - \lambda_{\text{среднее}}^2}.$$

Для оценки кумулятивного риска заражения используются такие параметры, как риск заражения за один эпизод и период в днях, для которого и производятся вычисления. Зависимость выражается формулой $\text{Pr} = 1 - e^{-\lambda p \Delta}$ [2, 3], где λ — соответствующая интенсивность, p — риск заражения за один эпизод, Δ — период в днях, Pr — кумулятивный риск.

Средняя оценка кумулятивного риска заражения вычисляется так же, как и средняя оценка интенсивности поведения:

$$\text{Pr}_{\text{среднее}} = \sum_{i,j,k} (\text{Pr}_{ijk} P_i P_j P_k).$$

Для расчета квадратичного отклонения оценки кумулятивного риска заражения используется следующая формула:

$$\sigma = \sqrt{\sum_{i,j,k} \text{Pr}_{ijk}^2 P_i P_j P_k - \text{Pr}_{\text{среднее}}^2}.$$

Остановимся подробнее на дискретных распределениях, с помощью которых выбираются веса точек. При равномерном распределении получим $p_i = \frac{1}{n+1}$, $0 \leq i \leq n$, а при биномиальном с заданной вероятностью успеха π $p_i = C_n^i \pi^i (1-\pi)^{n-i}$, $\pi \in [0, 1]$, $0 \leq i \leq n$.

3. Расчет отношения величин интенсивности и кумулятивного риска. Для того, чтобы определить влияние поведенческой интервенции, рассматриваются отношения R_λ и R_{Pr} априорных и апостериорных показателей интенсивности поведения и кумулятивного риска, с ним связанного. Указанные значения R_λ и R_{Pr} вычисляются по формулам: $R_\lambda = \frac{\lambda_1}{\lambda_0}$ и $R_{\text{Pr}} = \frac{\text{Pr}_1}{\text{Pr}_0}$, где λ_0 и λ_1 — интенсивность угрожающего поведения до и после поведенческой интервенции соответственно, Pr_0 и Pr_1 — значение кумулятивного риска до и после поведенческой интервенции.

Как отмечалось ранее, данные о последних эпизодах угрожающего поведения извлекаются из естественно-языковых ответов,

поэтому являются неполными и неточными, а за счет рандомизации неопределенности ответа оценка интенсивности и кумулятивного риска рассматриваются как случайные величины. Следовательно, и отношения R_λ и R_{Pr} можно рассматривать как случайные величины.

Так же, как и при вычислении средней оценки интенсивности, интервал вариации для каждого эпизода разбивается на n равных частей. Рассматриваются все возможные сочетания точек из интервалов, соответствующих указанным эпизодам. Расчет среднего значения и квадратичного отклонения для случая трех последних эпизодов производится по следующим формулам:

$$R_\lambda = \sum_{\substack{i_0, j_0, k_0 \\ i_1, j_1, k_1}} \frac{\lambda_{1, i_1 j_1 k_1} P_{i_1} P_{j_1} P_{k_1}}{\lambda_{0, i_0 j_0 k_0} P_{i_0} P_{j_0} P_{k_0}},$$

$$R_{Pr} = \sum_{\substack{i_0, j_0, k_0 \\ i_1, j_1, k_1}} \frac{\Pr_{1, i_1 j_1 k_1} P_{i_1} P_{j_1} P_{k_1}}{\Pr_{0, i_0 j_0 k_0} P_{i_0} P_{j_0} P_{k_0}},$$

$$\sigma_{R_\lambda} = \sqrt{\sum_{\substack{i_0, j_0, k_0 \\ i_1, j_1, k_1}} \frac{\lambda_{1, i_1 j_1 k_1}^2 P_{i_1} P_{j_1} P_{k_1}}{\lambda_{0, i_0 j_0 k_0}^2 P_{i_0} P_{j_0} P_{k_0}} - R_\lambda^2},$$

$$\sigma_{R_{Pr}} = \sqrt{\sum_{\substack{i_0, j_0, k_0 \\ i_1, j_1, k_1}} \frac{\Pr_{1, i_1 j_1 k_1}^2 P_{i_1} P_{j_1} P_{k_1}}{\Pr_{0, i_0 j_0 k_0}^2 P_{i_0} P_{j_0} P_{k_0}} - R_{Pr}^2}.$$

Рассмотрим промежуток $[R_\lambda - k\sigma_{R_\lambda}, R_\lambda + k\sigma_{R_\lambda}]$, если он полностью лежит выше 1, то есть $R_\lambda - k\sigma_{R_\lambda} > 1$, то (с определенной степенью уверенности, зависящей от числа k стандартных отклонений, по которому определяется длина интервала; подход к определению степени уверенности здесь не приводится — отметим лишь, что для того, чтобы работать с доверительным интервалом, придется перейти к рассмотрению логлинейной модели) вмешательство оказало влияние на рассматриваемое поведение, причем интенсивность этого поведения возросла. Если же указанный промежуток лежит ниже 1, то интенсивность поведения снизилась. В случае, когда $1 \in [R_\lambda - k\sigma_{R_\lambda}, R_\lambda + k\sigma_{R_\lambda}]$, интервенция не оказала никакого влияния на интенсивность поведения. (Снова — два последних статистических суждения характеризуются степенью доверия, которая будет зависеть

от k .) Аналогично рассматривается изменение кумулятивного риска — промежуток $[R_{Pr} - k\sigma_{R_{Pr}}, R_{Pr} + k\sigma_{R_{Pr}}]$.

4. Максимизация вероятности интервала $[T_{\min}; T_{\max}]$. Пусть моделью поведения выступает пуассоновский процесс с уравнением

$$\Pr(\Delta t, k, \lambda) = \frac{(\lambda \Delta t)^k}{k!} e^{-\lambda \Delta t},$$

где Δt — промежуток времени наблюдения за поведением респондента; k — число эпизодов рассматриваемого поведения, случившихся в этот промежуток; λ — интенсивность поведения; $\Pr(\Delta t, k, \lambda)$ — вероятность того, что за промежуток времени наблюдения Δt при поведении с интенсивностью λ случится ровно k эпизодов указанного поведения.

Заметим, что для такого процесса известна также плотность распределения T — длины временного интервала между двумя соседними эпизодами: $p(T) = \lambda e^{-\lambda T}$, $T \geq 0$.

Пусть заданы T_{\max} — максимальная длина временного интервала между двумя соседними эпизодами и T_{\min} — минимальная длина временного интервала между двумя соседними эпизодами. Для того, чтобы получить оценку интенсивности λ по этим данным, рассмотрим подход, нацеленный на максимизацию вероятности того, что длины интервалов между эпизодами попадают в промежуток $[T_{\min}; T_{\max}]$.

Вычислим эту вероятность как функцию от интенсивности λ .

$$\begin{aligned} f(\lambda) = p(T \in [T_{\min}; T_{\max}]) &= \int_{T_{\min}}^{T_{\max}} \lambda e^{-\lambda \tau} d\tau = \\ &= e^{-\lambda T_{\min}} - e^{-\lambda T_{\max}}. \end{aligned}$$

Определим, при каком λ функция $f(\lambda)$ принимает максимальное значение:

$$\begin{aligned} f'(\lambda) &= -T_{\min} e^{-\lambda T_{\min}} + T_{\max} e^{-\lambda T_{\max}} = 0, \\ T_{\max} e^{-\lambda T_{\max}} &= T_{\min} e^{-\lambda T_{\min}}, \\ \ln T_{\max} - \lambda T_{\max} &= \ln T_{\min} - \lambda T_{\min}, \\ \ln T_{\max} - \ln T_{\min} &= \lambda(T_{\max} - T_{\min}), \end{aligned}$$

$$\lambda = \frac{\ln T_{\max} - \ln T_{\min}}{T_{\max} - T_{\min}}.$$

Таким образом, оценка $\hat{\lambda}$, основанная на известных экстремальных значениях, вычисляется по формуле:

$$\hat{\lambda} = \frac{\ln \frac{T_{\max}}{T_{\min}}}{T_{\max} - T_{\min}}.$$

5. Согласованность среднего и экстремальных значений длины временного интервала между соседними эпизодами поведения. Пусть заданы не только T_{\max} и T_{\min} — максимальная и минимальная длина временного интервала между двумя соседними эпизодами, но и $T_{\text{обычное}}$ — длина обычного интервала между двумя соседними эпизодами. Необходимо проверить непротиворечивость этих данных. Оценка интенсивности по значениям T_{\max} и T_{\min} была получена ранее:

$$\hat{\lambda} = \frac{\ln \frac{T_{\max}}{T_{\min}}}{T_{\max} - T_{\min}}.$$

С другой стороны из определения понятия интенсивности как числа эпизодов поведения рассматриваемого вида в определенный промежуток времени вытекает, что $\lambda = \frac{1}{T_{\text{среднее}}}$, то есть

$$T_{\text{среднее}} = \frac{1}{\hat{\lambda}} = \frac{T_{\max} - T_{\min}}{\ln \frac{T_{\max}}{T_{\min}}}$$

Тогда отношение $R_T = \frac{T_{\text{обычное}}}{T_{\text{среднее}}}$ является численной характери-

стикой согласованности данных. Чем более удаленной окажется величина этого значения от 1, тем менее будут представляться данные согласованными друг с другом.

С другой стороны, можно предложить несколько иной индикатор согласованности данных — при известном $T_{\text{среднее}}$ оценить вероятность попадания длины интервала между эпизодами в промежуток $[T_{\min}; T_{\max}]$, границы которого задаются также известными (сообщенными респондентом) экстремальными значениями:

$$p([T_{\min}; T_{\max}] | T_{\text{среднее}}) = \int_{T_{\min}}^{T_{\max}} \frac{1}{T_{\text{среднее}}} e^{\frac{-t}{T_{\text{среднее}}}} dt,$$

$$p([T_{\min}; T_{\max}] | T_{\text{среднее}}) = e^{\frac{-T_{\min}}{T_{\text{среднее}}}} - e^{\frac{-T_{\max}}{T_{\text{среднее}}}}.$$

Чем меньше окажется получающаяся величина, тем меньше согласованы полученные в результате интервью данные.

6. Выводы. Получены численные показатели, позволяющие сравнивать интенсивность угрожающего поведения до и после поведенческой интервенции. Кроме того, показано, как с помощью того же принципа характеризуется соотношение кумулятивных рисков до и после проведения мероприятий, направленных на модификацию поведения. Наконец, рассмотрены две численные характеристики, отражающие степень согласованности исходных данных.

Вопросы, связанные с характеристикой качества точечной оценки вышеуказанных отношений интенсивностей и рисков (то есть вопросы построения соответствующих доверительных интервалов), могут быть решены в рамках подходящей логлинейной модели. Качество характеристик согласованности требует дальнейших исследований; видимо, соответствующие доверительные интервалы будут зависеть от заданного «интервала наблюдения».

Литература

1. *Тулупьева Т. В., Тулупьев А. Л., Пащенко А. Е.* Оценка интенсивности поведения респондента в условиях информационного дефицита // Труды СПИИРАН. Вып. 7. СПб.: Наука, 2008. С. 239–254.
2. *Тулупьева Т. В., Тулупьев А. Л., Пащенко А. Е., Красносельских Т. В.* Приверженность ВААРТ и рискованное поведение среди пациентов Санкт-Петербургского Центра-СПИД: статистические модели, психологические и социо-демографические факторы // Труды СПИИРАН. 2008. Вып. 6. СПб.: Наука, 2008. С. 207–237.
3. *Тулупьева Т. В., Пащенко А. Е., Тулупьев А. Л., Красносельских Т. В., Казакова О. С.* Модели ВИЧ-рискованного поведения в контексте психологической защиты и других адаптивных стилей. СПб.: Наука, 2008. 140 с.
4. *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. СПб.: Наука, 2007. Вып. 5. С. 117–150.
5. *Bell D. C., Trevino R. A.* Modeling HIV Risk [Epidemiology] // J. Acquir Immune Defic Syndr. 1999. С. 280–287. vol.22, № 3.

6. *Bell D. C., Atkinson J. S., Mosier V., Riley M., Brown V. L.* The HIV Transmission Gradient: Relationship Patterns of Protection // *AIDS Behav.* 2007. С. 789–811. Vol. 11 № 6.
7. *Пащенко А. Е., Тулупьев А. Л., Николенко С. И.* Моделирование заражения ВИЧ-инфекцией на основе данных о последних эпизодах рискованного поведения. // *Известия высших учебных заведений: Приборостроение.* 2006. №8. 33–34 с.
8. *Пащенко А. Е.* Идентификация интенсивности пуассоновского процесса, моделирующего поведение респондента, в условиях дефицита информации. Информационно-измерительные и управляющие системы. 2009. № 4.

Тулупьев Александр Львович — к. ф.-м.н., доцент; ведущий научный сотрудник научно-исследовательской группы междисциплинарных проблем информатики СПИИРАН, доцент кафедры информатики математико-механического факультета С.-Петербургского государственного университета (СПбГУ). Область научных интересов: представление и обработка данных и знаний с неопределенностью, применение методов математики и информатики в социокультурных исследованиях, применение методов биostatистики и математического моделирования в эпидемиологии, технология разработки программных комплексов с СУБД. Число научных публикаций — 150. ALT@iias.spb.su, www.tulupyeu.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.

Тулупьева Татьяна Валентиновна — к.психол.н., доцент; старший научный сотрудник научно-исследовательской группы междисциплинарных проблем информатики Учреждения Российской академии наук С.-Петербургский институт информатики и автоматизации РАН (СПИИРАН), доцент кафедры информатики математико-механического факультета С.-Петербургского государственного университета (СПбГУ), доцент кафедры психологии управления и педагогики Северо-Западной академии государственной службы (СЗАГС). Область научных интересов: применение методов математики и информатики в гуманитарных исследованиях, информатизация организации и проведения психологических исследований, применение методов биostatистики в эпидемиологии, психология личности, психология управления. Число научных публикаций — 45. TVT@iias.spb.su, www.tulupyeu.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.

Пашенко Антон Евгеньевич — младший научный сотрудник научно-исследовательской группы междисциплинарных проблем информатики Учреждения Российской академии наук С.-Петербургский институт информатики и автоматизации РАН (СПИИРАН). Область научных интересов: математическая статистика, статистическое моделирование, применение методов биostatистики и математического моделирования в эпидемиологии. Число научных публикаций — 35. AEP@iias.spb.su, www.tulupyeu.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.

Суворова Алена Владимировна — студент мат.-мех. факультета С.-Петербургского государственного университета (СПбГУ). Область научных интересов: математическая статистика, теория вероятности SUVALV@mail.ru; СПбГУ, м/м ф-т, Университетский пр., д. 28, Старый Петергоф, Санкт-Петербург, 198504, РФ.

Поддержка исследований. Часть результатов, представленных в настоящей работе, была получена на основе результатов исследований, поддержанных грантом РГНФ

«Взаимосвязь адаптивных стилей ВИЧ-инфицированных и степени рискованности их поведения» №07-06-00738а, госконтрактом № 2.442.11.7489, шифр 2006-РИ-19.0/001/209, на НИР «Психологическая защита и копинг-стратегии ВИЧ-инфицированных с точки зрения опасности для общественного здоровья» в рамках ФЦНТП «Исследования и разработки по приоритетным направлениям развития науки и техники на 2002–2006 годы», грантом СПбНЦ РАН на 2007 год «Моделирование и измерение количественных характеристик ВИЧ-рискованного поведения на основе обработки ответов респондентов» № 2-199. Руководитель проектов — Т. В. Тулупьева.

Часть результатов получена в проекте «Оценка вероятности заражения ВИЧ-инфекцией на основе сведений о последних N эпизодах рискованного поведения, а также статистическое моделирование ограниченных указанных серий эпизодов», поддержанном грантом №02/2.1/17-03/48 (в 2007 году) Конкурса для студентов и аспирантов вузов и академических институтов, расположенных на территории Санкт-Петербурга. Руководитель проекта — А. Е. Пашенко.

Рекомендовано ЛПИ СПИИРАН, зав. лаб. Юсупов Р.М., член-корреспондент РАН.
Статья поступила в редакцию 25.06.2009.