

Л.Н. Белых, А.П. Бирюков, Е.В. Васильев, В.П. Невзоров

О ТЕОРЕТИЧЕСКИХ ОЦЕНКАХ СРЕДНЕГО РИСКА ОБЩЕЙ СМЕРТНОСТИ И ПРАВОМЕРНОСТИ ПРИМЕНЕНИЯ РАЗЛИЧНЫХ ЗАКОНОВ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ В ЭПИДЕМИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

L.N. Belyh, A.P. Biryukov, E.V. Vasiliev, V.P. Nevzorov

About the Theoretical Estimates the Average Risk of Death and Legality of the Application of Various Laws of Probability Distributions in Epidemiological Studies

РЕФЕРАТ

Цель: Теоретически подтвердить целесообразность использования эмпирических формул для расчета средних рисков смертности и применения различных законов распределения вероятностей для статистического анализа данных наблюдений.

Результаты: Получены оценки среднего значения риска смерти с различными степенями точности через функцию риска и функцию распределения времени наступления нежелательного события. Получена статистическая оценка риска в частном случае (экспоненциальное распределение) методом максимального правдоподобия. Продемонстрированы взаимозаменяемости законов распределения вероятностей в эпидемиологии.

Выводы: Рассмотрение времени наступления нежелательного события в виде непрерывной конечной положительной случайной величины позволило при помощи функции риска построить ее функцию распределения и получить с различными степенями точности оценки среднего значения риска.

Проведенная работа может служить теоретической основой для использования эмпирических формул при расчетах средних рисков смертности с применением различных законов распределения вероятностей статистического анализа данных наблюдений.

Ключевые слова: функция риска, среднее значение риска, законы распределения, статистическая оценка риска

ABSTRACT

Purpose: To confirm theoretically the usefulness of empirical formulas for the calculation of the average risk of mortality and the use of various laws of probability distribution for the statistical analysis of observational data.

Results: Estimations of the average value of the risk of death with varying degrees of accuracy by the risk function, the distribution function of the risk, time of occurrence of an undesirable event. Obtain statistical assessment of risk in a particular case (the exponential distribution) by maximum likelihood. Show cased interchangeability laws of probability distribution in epidemiology.

Conclusion: The time of onset of adverse events issued in the form of a continuous finite positive random variable. It enabled to construct a distribution function and to obtain estimates of the average value of the risk with varying degrees of accuracy.

Existing work can serve as a theoretical basis for the use of empirical formulas for calculation of average value of the risk function with various laws of probability distribution of statistical data analysis.

Key words: risk function, average value of the risk function, distribution laws, statistical evaluation of the risk

Введение

В настоящее время эмпирические формулы для расчета средних рисков смертности и применения различных законов распределения вероятностей для статистического анализа результатов наблюдений находят все большее распространение. Так, например, специалисты IARC [1] представляют риск смерти как отношение числа случаев смерти к количеству человеко-лет наблюдения и рассматривают три закона распределения результатов (нормальный, пуассоновский и биномиальный) при обработке данных. Понятие функции риска упоминают специалисты в документах НКДАР ООН [2], но не рассматривают его как строгое математическое определение функции рисков.

В работе предложен вариант представления времени наступления смерти в виде непрерывной случайной величины в характеристике показателя

функции распределения по уровню среднего значения риска смерти, где предусмотрена возможность использования интервала человеко-лет наблюдения с выходом на разную степень точности относительно используемых эмпирических формул.

Переход к характеристике среднего значения риска смерти на соответствующей территории является источником для получения статистических характеристик при эпидемиологических исследованиях по выявлению риска наступления нежелательного события (заболевания, смерти) для человека, находящегося на рассматриваемой территории.

Такой подход предназначен для специалистов, занимающихся проблемой состояния здоровья населения и оценивающих значимость влияния регистрируемого фактора воздействия на исследуемой территории, в частности, для специалистов по радиационной эпидемиологии.

Федеральный медицинский биофизический центр им. А.И.Бурназяна ФМБА России, Москва. E-mail: valerij.nevzorov@mail.ru

A.I. Burnasyan Federal Medical Biophysical Center of FMBA, Moscow, Russia. E-mail: valerij.nevzorov@mail.ru

Теоретическое подтверждение такой возможности открывает дорогу практическому использованию описанного подхода практическому здравоохранению вне зависимости от конкретного фактора, учитываемого в каждом отдельном случае, в том числе и радиационной эпидемиологии.

Определение функции риска, ее свойства и оценки среднего значения риска

Рассмотрим положительную конечную непрерывную случайную величину $t_{\text{наст}}$ — момент наступления некоего нежелательного бесповторного события в жизни человека. Таким событием является смерть человека по любой причине. По определению $0 < t_{\text{наст}} < \infty$. Непрерывность случайной величины $t_{\text{наст}}$ предполагает существование непрерывной функции ее распределения [3]:

$$R(t) = P(t_{\text{наст}} < t), t \in [0, \infty), \quad (1)$$

$$R(0) = 0, \quad (2)$$

$$R(\infty) = 1 \quad (3)$$

и кусочно-непрерывной производной этой функции — плотности распределения вероятностей [3]:

$$r(t) = \frac{dR}{dt} = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq t_{\text{наст}} < t + \Delta t)}{\Delta t}. \quad (4)$$

Функция $R(t)$ является неотрицательной неубывающей ограниченной функцией: $0 \leq R(t) \leq 1$. Справедливо равенство $R(t) = \int_0^t r(t)dt$ и цепочка равенств

$$P(a \leq t_{\text{наст}} \leq b) = P(a \leq t_{\text{наст}} < b) = P(a < t_{\text{наст}} \leq b) = P(a < t_{\text{наст}} < b) = \int_a^b r(t)dt = R(b) - R(a).$$

Условия (2) и (3) являются следствием определения случайной величины $t_{\text{наст}}$. В самом деле, $R(0) = P(t_{\text{наст}} < 0) = 0$ как вероятность невозможного события и $R(\infty) = P(t_{\text{наст}} < \infty) = 1$ как вероятность достоверного события. Независимую переменную можно интерпретировать как текущее время или как текущий возраст человека.

Таким образом, функция $R(t)$ есть вероятность того, что рассматриваемое нежелательное событие наступит в полуоткрытом интервале $[0, t)$. Тогда вероятность того, что в указанном интервале событие не произойдет, есть

$$Q(t) = P(t_{\text{наст}} \geq t) = 1 - P(t_{\text{наст}} < t) = 1 - R(t). \quad (5)$$

Введем в рассмотрение вероятность того, что нежелательное событие произойдет в малом интервале $[t, t + \Delta t)$ при условии, что оно не произошло ранее, т.е.

$$P(t \leq t_{\text{наст}} < t + \Delta t / t_{\text{наст}} \geq t) = P(A / B),$$

где $A = \{t_{\text{наст}} : t \leq t_{\text{наст}} < t + \Delta t\}$

$$\text{и } B = \{t_{\text{наст}} : t_{\text{наст}} \geq t\}. \quad (6)$$

Очевидно, что пересечение $A \cdot B = A$ и

$$P(A / B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)}, \quad (7)$$

где

$$P(A) = P(t \leq t_{\text{наст}} < t + \Delta t) = R(t + \Delta t) - R(t) = \Delta R(t),$$

$$P(B) = P(t_{\text{наст}} \geq t) = 1 - P(t_{\text{наст}} < t) = 1 - R(t). \quad (8)$$

Определим величину $\lambda(t)$ как

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq t_{\text{наст}} < t + \Delta t / t_{\text{наст}} \geq t)}{\Delta t} \quad (9)$$

и назовем ее функцией риска. Согласно (9), функция риска — это вероятность того, что нежелательное событие наступит в течение короткого интервала времени $[t, t + \Delta t)$, $\Delta t \rightarrow 0$, при условии, что оно не произошло ранее.

С учетом (7) и (8) получим

$$\begin{aligned} \lambda(t) &= \lim_{\Delta t \rightarrow 0} \frac{P(A / B)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{P(A)}{\Delta t \cdot P(B)} = \\ &= \lim_{\Delta t \rightarrow 0} \frac{\Delta R(t)}{\Delta t (1 - R(t))} = \frac{dR/dt}{1 - R(t)}. \end{aligned}$$

Из последнего равенства, с учетом (2) и (4) следует

$$\lambda(t) = \frac{r(t)}{1 - R(t)} \quad (10)$$

и

$$R(t) = 1 - e^{-\int_0^t \lambda(t)dt}. \quad (11)$$

Введем индикатор рассматриваемого события как дискретную случайную величину вида:

$$S(t) = \begin{cases} 1 & \text{с вероятностью } R(t), \\ 0 & \text{с вероятностью } Q(t) = 1 - R(t). \end{cases} \quad (12)$$

Единица указывает на то, что событие произошло, а ноль — событие не произошло.

Вычислим математическое ожидание и дисперсию случайной величины S :

$$\begin{aligned} M(S(t)) &= 1 \cdot R(t) + 0 \cdot Q(t) = R(t) \\ D(S(t)) &= M(S^2(t)) - M^2(S(t)) = \\ &= 1^2 \cdot R(t) + 0^2 \cdot Q(t) - R^2(t) = R(t) \cdot Q(t). \end{aligned}$$

Пусть имеется группа из N человек, каждый со своим индикатором $S_i(t)$ в виде (12), $i = \overline{1, N}$. Тогда математическое ожидание и дисперсия суммы показателей всех индикаторов равны

$$\begin{aligned} M(\sum_{i=1}^N S_i(t)) &= \sum_{i=1}^N M(S_i(t)) = \\ &= \sum_{i=1}^N R(t) = N \cdot R(t) \end{aligned} \quad (13)$$

$$D(\sum_{i=1}^N S_i(t)) = \sum_{i=1}^N D(S_i(t)) = \\ = \sum_{i=1}^N Q(t) \cdot R(t) = N \cdot R(t) \cdot Q(t). \quad (14)$$

При вычислении предполагалась независимость показаний индикаторов.

Тогда функция риска $\lambda(t)$ есть:

$$\lambda(t) = \frac{\frac{dR}{dt}}{1-R(t)} = \frac{N \frac{dR}{dt}}{N(1-R(t))} = \frac{\frac{dm(t)}{dt}}{N-m(t)}, \quad (15)$$

где $m(t) = N \cdot R(t)$ – математическое ожидание числа событий, произошедших в интервале $[0, t)$.

Введем $N(t) = N - m(t)$ – математическое ожидание числа не наступивших из N возможных в интервале $[0, t)$ событий. Функция $N(t)$ – неотрицательная невозрастающая функция, потому что $N = \text{const} > 0$, $0 \leq m(t) = N \cdot R(t) \leq N$ и $m(t)$ – неубывающая функция в силу $\frac{dm}{dt} = N \frac{dR}{dt} = N \cdot r(t) \geq 0$.

Тогда, согласно (15):

$$\lambda(t) = \frac{\frac{dm(t)}{dt}}{N-m(t)} = -\frac{\frac{dN(t)}{dt}}{N(t)} = -\frac{d \ln N(t)}{dt} \geq 0, \quad (16)$$

то есть функция риска $\lambda(t)$ есть абсолютная величина скорости убывания логарифма математического ожидания числа не наступивших в интервале $[0, t)$ нежелательных событий из N возможных $N(t)$. Отсюда при очевидном начальном условии $N(0) = N$ следует

$$\int_0^t \lambda(t) dt = -\ln N(t) \Big|_0^t = \ln N(0) - \ln N(t) = \\ = \ln \frac{N(0)}{N(t)} = \ln \frac{N}{N-m(t)} = \ln \left(1 + \frac{m(t)}{N-m(t)}\right). \quad (17)$$

Интеграл слева по теореме о среднем значении [3] представим в виде:

$$\int_0^t \lambda(t) dt = \lambda(\xi) \cdot (t - 0) = \lambda(\xi) \cdot t, \quad (18)$$

где $\xi \in [0, t)$.

Зафиксируем некий момент времени $t = t_{fin}$, который можно интерпретировать как время окончания эксперимента или наблюдения. Пусть T – длина интервала $[0, t_{fin})$ (интервала наблюдения), т.е. $T = t_{fin} - 0 = t_{fin}$. Тогда $t = t_{fin} = T$ и $\xi \in [0, T)$. Обозначив $\lambda_{cp} = \lambda(\xi)$, из (18) и (17) получим:

$$\int_0^T \lambda(t) dt = \lambda_{cp} \cdot T = \ln \left(1 + \frac{m}{N-m}\right), \quad (19)$$

где $m = m(T)$ – математическое ожидание числа наступивших в интервале $[0, T]$ нежелательных событий.

Из (19) следует, что среднее в смысле (18) значение функции риска $\lambda(t)$ на интервале $[0, T]$ есть:

$$\lambda_{cp} = \frac{1}{T} \ln \left(1 + \frac{m}{N-m}\right). \quad (20)$$

Укажем приближения правой части равенства (20) (приближения логарифмической функции), основанные на свойствах логарифмической функции $\ln(1+x)$ (выпуклость, разложение в ряд Тейлора, тео-

рема Лейбница об остатке знакопередающегося ряда [3]).

Прежде всего, при $N > m$ выполняется строгое двойное неравенство

$$\frac{m}{NT} < \lambda_{cp} < \frac{m}{(N-m)T}. \quad (21)$$

В самом деле, для любого $x \geq 0$ справедливо $\frac{1}{(1+x)^2} \leq \frac{1}{(1+x)} \leq 1$, причем равенство возможно только при $x = 0$. Тогда по свойствам определенного интеграла $\forall x^* \geq 0$ справедливо:

$$\int_0^{x^*} \frac{dx}{(1+x)^2} \leq \int_0^{x^*} \frac{dx}{(1+x)} \leq \int_0^{x^*} 1 \cdot dx,$$

или

$$\frac{x^*}{1+x^*} \leq \ln(1+x^*) \leq x^*,$$

причем равенство возможно только при $x^* = 0$. Подставив сюда значение $x^* = \frac{m}{N-m} > 0$, получим:

$$\frac{m}{N} < \ln \left(1 + \frac{m}{N-m}\right) < \frac{m}{N-m}.$$

Разделив все три части этого неравенства на продолжительность интервала наблюдения T , придем к искомому результату – строгому неравенству (21).

Следует иметь в виду, что среднее значение риска $\lambda_{cp} = \lambda(\xi) = \frac{1}{T} \int_0^T \lambda(t) dt$, где $\xi \in [0, T]$, не является постоянной величиной, а зависит от интервала наблюдения T . Даже при исследованиях одной и той же выборки на разных интервалах наблюдения будут получены разные средние значения риска. Однако, как будет показано ниже, при $m \ll N$, левая и правая границы неравенства (21) достаточно близки (являются эквивалентными бесконечно малым величинам, т.е. предел их отношения при $m \rightarrow 0$ равен единице) и могут быть использованы в качестве оценок среднего значения риска.

Приближенные значения для λ_{cp} находятся из следующих равенств, справедливых с различными степенями точности:

а) с точностью до величин, эквивалентных $O\left(\frac{1}{2} \left(\frac{m}{N}\right)^2\right)$,

$$\lambda_{cp} T = \frac{m}{(N-m)} \quad (\text{то есть } \lambda_{cp} = \frac{m}{(N-m)T}). \quad (22)$$

и

$$\lambda_{cp} T = \frac{m}{N} \quad (\text{то есть } \lambda_{cp} = \frac{m}{NT}). \quad (23)$$

б) с точностью до величин, эквивалентных $O\left(\frac{1}{3} \left(\frac{m}{N}\right)^3\right)$,

$$\lambda_{cp} T = \frac{m}{(N-m)} - \frac{1}{2} \left(\frac{m}{N-m}\right)^2 \quad (24)$$

и

$$\lambda_{cp} T = \frac{m}{N} + \frac{1}{2} \left(\frac{m}{N}\right)^2. \quad (25)$$

Пример.

Пусть $m = 1 \cdot 10^3, N = 25 \cdot 10^3, T = 20$ (лет). Тогда согласно формуле (20) точное значение $\lambda_{\text{ср}}$ есть:

$$\lambda_{\text{ср}} = \frac{1}{T} \ln\left(1 + \frac{m}{N-m}\right) = \frac{1}{20} \ln\left(1 + \frac{1}{25-1}\right) = \\ = \frac{1}{20} \ln \frac{25}{24} = \frac{1}{20} (\ln 25 - \ln 24) = 204,11 \cdot 10^{-5}.$$

Получим оценки $\lambda_{\text{ср}}$

а) по формулам (22) и (23):

$$\lambda_{\text{ср}} = \frac{m}{(N-m)T} = \frac{1}{24 \cdot 20} = 208,33 \cdot 10^{-5}$$

и

$$\lambda_{\text{ср}} = \frac{m}{NT} = \frac{1}{25 \cdot 20} = 200,00 \cdot 10^{-5};$$

б) по формулам (24) и (25)

$$\lambda_{\text{ср}} = \frac{1}{T} \left(\frac{m}{(N-m)} - \frac{1}{2} \left(\frac{m}{(N-m)} \right)^2 \right) = 203,99 \cdot 10^{-5}$$

и

$$\lambda_{\text{ср}} = \frac{1}{T} \left(\frac{m}{N} + \frac{1}{2} \left(\frac{m}{N} \right)^2 \right) = 204,00 \cdot 10^{-5}.$$

Статистическая оценка риска методом максимального правдоподобия в экспоненциальном случае

Предположим, что функция риска (или просто риск) является постоянной положительной величиной λ , т.е. $\lambda(t) \equiv \lambda = \text{const} > 0$ для всех $t \geq 0$. Тогда из общего вида функции распределения (11) следует

$$R(t) = 1 - e^{-\int_0^t \lambda dt} = 1 - e^{-\lambda t}. \quad (26)$$

Закон распределения (26) называется экспоненциальным, или показательным.

Пусть $t \in [0, T]$, где T – одновременно и продолжительность интервала наблюдения, и время окончания наблюдения $t_{\text{fin}} = T$. Это возможно, если время начала наблюдения считать нулем: $t_0 = 0$. Рассмотрим группу из N индивидуумов на интервале $[0, T]$, в котором произошло m случаев наступления нежелательных событий, причем каждый i -ый случай произошел в свой i -ый момент времени $t_i, i = \overline{1, m}$.

Индикатором состояния i -го индивидуума ($i = \overline{1, m}$) является дискретная случайная величина:

$$S_i = \begin{cases} 1 \text{ с вероятностью } R(t_i) = 1 - e^{-\lambda t_i}, \\ 0 \text{ с вероятностью } Q(T) = 1 - R(T) = e^{-\lambda T}. \end{cases} \quad (27)$$

Единица указывает на то, что событие с вероятностью $R(t_i)$ произошло в интервале $[0, t_i] \subset [0, T]$, а ноль – на то, что с вероятностью $Q(T)$ в интервале $[0, T]$ ничего не случилось (события не произошло).

Запишем $R(t_i)$ в виде, удобном для анализа, опуская индекс i :

$$R(t) = 1 - e^{-\lambda t} = e^{-\lambda t} (e^{\lambda t} - 1) = \\ = e^{-\lambda t} \left(\lambda t + e^{\lambda \delta} \frac{(\lambda t)^2}{2} \right) = \lambda t e^{-\lambda t} + e^{-\lambda(t-\delta)} \frac{(\lambda t)^2}{2} = \\ = \lambda t e^{-\lambda t} + O\left(\frac{(\lambda t)^2}{2}\right), \delta \in (0, t).$$

Здесь использовались разложение функции e^x , где $x = \lambda t$, в ряд Маклорена и формула его остаточного члена [3]. Тогда индикатор события, с точностью до величин, эквивалентных $O\left(\frac{x^2}{2}\right)$, примет вид

$$S_i = \begin{cases} 1 \text{ с вероятностью } R(t_i) = \lambda t_i e^{-\lambda t_i}, \\ 0 \text{ с вероятностью } Q(T) = e^{-\lambda T}. \end{cases} \quad (28)$$

Функция правдоподобия для описанной ситуации (m событий из N возможных на интервале $[0, T]$) записывается в виде:

$$L(\lambda) = \prod_{i=1}^m R(t_i) \cdot \prod_{i=m+1}^N Q(T) = \\ = \prod_{i=1}^m \lambda t_i e^{-\lambda t_i} \cdot \prod_{i=m+1}^N e^{-\lambda T}. \quad (29)$$

Первое произведение есть вероятность наступления независимых нежелательных событий на интервале $[0, T]$, каждое i -ое из которых наступает в свой момент времени t_i . Второе произведение есть вероятность того, что на всем интервале $[0, T]$ с $(N-m)$ индивидуумами ничего нежелательного не случилось.

Натуральный логарифм функции $L(\lambda)$ равен

$$\ln L(\lambda) = \sum_{i=1}^m \ln(\lambda t_i e^{-\lambda t_i}) + \sum_{i=m+1}^N \ln(e^{-\lambda T}) = \\ = m \ln \lambda + \sum_{i=1}^m \ln t_i - \lambda \sum_{i=1}^m t_i - \lambda \sum_{i=m+1}^N T = \\ = m \ln \lambda - \lambda (\sum_{i=1}^m t_i + (N-m)T) + \sum_{i=1}^m \ln t_i = \\ = m \ln \lambda - \lambda \cdot PY + \sum_{i=1}^m \ln t_i.$$

Здесь $PY = \sum_{i=1}^m t_i + (N-m)T$ – количество человеко-лет наблюдения.

Согласно методу максимального правдоподобия [3], для нахождения оценки риска λ следует найти максимум функции $L(\lambda)$ или, что то же самое и удобнее, максимум ее логарифма $\ln L(\lambda)$. Для этого необходимо и достаточно приравнять производную функции $\ln L(\lambda)$ нулю и убедиться, что ее вторая производная при найденном значении λ отрицательна.

$$\frac{d \ln L(\lambda)}{d \lambda} = m \cdot \frac{1}{\lambda} - PY = 0, \frac{d^2 \ln L(\lambda)}{d \lambda^2} = -\frac{m}{\lambda^2} < 0$$

для всех $\lambda > 0$.

Из первого равенства следует оценка:

$$\lambda = \frac{m}{PY} = \frac{m}{\sum_{i=1}^m t_i + (N-m)T}. \quad (30)$$

Полученная оценка риска λ является эффективной, состоятельной, асимптотически несмещенной и асимптотически нормальной [3, 5].

Если предположить, что у каждого i -го индивидуума на интервале $[0, t)$ вероятность наступления события определяется законом Пуассона, то:

$$P(k_i, N, t) = \frac{(\lambda t)^{k_i}}{k_i!} e^{-\lambda t}, k_i = 0; 1. \quad (31)$$

Оценим, насколько сильно ограничение $k_i = 0; 1$ в законе (31), т.к. должно быть $k_i = 0, 1, 2, \dots$

Прежде всего, это ограничение возникает в связи с тем, что у i -го индивидуума нежелательное событие на интервале может наступить или один раз ($k_i = 1$), или ни разу ($k_i = 0$). Вычислим соответствующие вероятности:

$$p = P(1, N, t) = \lambda t \cdot e^{-\lambda t},$$

$$q = P(0, N, t) = e^{-\lambda t}$$

и оценим, с какой точностью их сумма приближается к единице, так как сумма вероятностей противоположных событий должна равняться единице. Найдем разность

$$\begin{aligned} \Delta &= 1 - (p + q) = e^{\lambda t} \cdot e^{-\lambda t} - \lambda t \cdot e^{-\lambda t} - e^{-\lambda t} = \\ &= e^{-\lambda t} (e^{\lambda t} - 1 - \lambda t) = e^{-\lambda t} \left(1 + \lambda t + e^{\delta t} \frac{(\lambda t)^2}{2} - 1 - \lambda t \right) \\ &= e^{-\lambda t} \cdot e^{\delta t} \frac{(\lambda t)^2}{2} = O \left(\frac{(\lambda t)^2}{2} \right), \delta \in (0, \lambda). \end{aligned}$$

Таким образом, индикатор события у i -го индивидуума S_i имеет тот же самый вид (27) и с той же самой точностью $O \left(\frac{(\lambda t)^2}{2} \right)$:

$$S_i = \begin{cases} 1 \text{ с вероятностью } p_i = \lambda t_i \cdot e^{-\lambda t_i}, \\ 0 \text{ с вероятностью } q_i = e^{-\lambda t_i}. \end{cases} \quad (32)$$

Разумеется, метод максимального правдоподобия для пуассоновских индикаторов дает ту же самую оценку:

$$\lambda = \frac{m}{PY} = \frac{m}{\sum_{i=1}^m t_i + (N-m)T}. \quad (33)$$

Тем самым мы продемонстрировали эквивалентность экспоненциального и пуассоновского подходов к статистической оценке рисков λ .

Таким образом, в данном разделе рассмотрение времени наступления нежелательного события (смерти) в виде непрерывной случайной величины $t_{\text{наст}}$ позволило параметризовать функцию ее распределения $R(t)$ при помощи функции риска $\lambda(t)$ и получить оценки среднего значения риска смерти с разными степенями точности, которые отражают структуру эмпирических формул — отношение числа умерших за интервал наблюдения к числу человеко-лет наблюдения.

Взаимозаменяемость законов распределения вероятностей в эпидемиологии

Рассмотрим группу из N индивидуумов, у каждого из которых с вероятностью $R(t)$ в интервале $[0, t)$ наступает нежелательное событие. Тогда вероятность $P_j(t, N)$ того, что в группе из N лиц наступит j событий в интервале $[0, t)$ подчиняется биномиальному закону [4]:

$$P_j(t, N) = \frac{N!}{j!(N-j)!} R^j(t) \cdot (1 - R(t))^{N-j} \quad (34)$$

с математическим ожиданием для j наступивших в интервале $[0, t)$ событий $M(j, t) = N \cdot R(t) = m(t)$ и дисперсией $D(j, t) = N \cdot R(t) \cdot (1 - R(t))$, $j = 0, 1, 2, \dots, N$.

При достаточно больших $N(N \rightarrow \infty)$ и $R(t) = \frac{m(t)}{N} \ll 1$ (достаточно $\frac{m(t)}{N} < 0,1$) биномиальное распределение (34) аппроксимируется законом Пуассона [3, 4]:

$$P_j(t, N) = e^{-N \cdot R(t)} \cdot \frac{(N \cdot R(t))^j}{j!} \quad (35)$$

с одинаковыми математическим ожиданием и дисперсией

$$M(j, t) = D(j, t) = N \cdot R(t) = m(t). \quad (36)$$

По локальной теореме Муавра–Лапласа [3] при $N \rightarrow \infty$ биномиальное распределение (34) становится нормальным:

$$P_j(t, N) = \frac{1}{\sqrt{2\pi N \cdot R(t) \cdot Q(t)}} \cdot e^{-\frac{(j - N \cdot R(t))^2}{2N \cdot R(t) \cdot Q(t)}}, \quad (37)$$

где $Q(t) = 1 - R(t)$,

и с учетом пуассоновости принимает вид

$$P_j(t, N) = \frac{1}{\sqrt{2\pi N \cdot R(t)}} \cdot e^{-\frac{(j - N \cdot R(t))^2}{2N \cdot R(t)}} \quad (38)$$

с равными математическим ожиданием и дисперсией

$$M(j, t) = D(j, t) = N \cdot R(t) = m(t). \quad (39)$$

Частным подтверждением эквивалентности нормального (38) и пуассоновского (35) законов распределения является случай, когда число наступивших в интервале $[0, t)$ событий равно своему математическому ожиданию, т.е. случай $j = N \cdot R(t) = m(t) = m$. При этом условии нормальная вероятность есть:

$$P_m(t, N) = \frac{1}{\sqrt{2\pi m}},$$

а пуассоновская вероятность равна:

$$P_m(t, N) = e^{-m} \cdot \frac{m^m}{m!} = \frac{1}{\sqrt{2\pi m}}.$$

Здесь использована формула Стирлинга [3] при достаточно больших m , например, при $m > 10$:

$$m! = \sqrt{2\pi m} \cdot m^m \cdot e^{-m}.$$

Формула Стирлинга означает, что отношение величин, стоящих в ней слева и справа при $m \rightarrow \infty$ стремится к единице, причем чем больше m , тем точнее равенство.

Пример. Пусть $m = 10$. Тогда:

$$m! = 10! = 1 \cdot 2 \cdot 3 \cdot \dots \cdot 8 \cdot 9 \cdot 10 = 3628800 = 3,6288 \cdot 10^6,$$

$$e^{-m} = e^{-10} = \frac{1}{e^{10}} = \frac{1}{2,2026462 \cdot 10^4} = 0,4539993 \cdot 10^{-4},$$

$$\pi = 3,1415927,$$

$$M = \sqrt{2\pi m} \cdot m^m \cdot e^{-m} = \sqrt{20\pi} \cdot 10^{10} \cdot e^{-10} =$$

$$= 7,9266546 \cdot 10^{10} \cdot 0,4539993 \cdot 10^{-4} = 3,5986956 \cdot 10^6.$$

Абсолютная ошибка

$$\Delta = |m! - M| = 0,0301044 \cdot 10^6.$$

Относительная ошибка в процентах

$$\delta = \frac{\Delta}{m!} \times 100 \% = 0,8267195 \% \approx 0,83 \%.$$

Отношение

$$\frac{m!}{M} = \frac{3,6288000 \cdot 10^6}{3,5986956 \cdot 10^6} = 1,0083653.$$

Биномиальное (дискретное) распределение является адекватным описанием (приближением) реального процесса формирования выборки. Переход к нормальному (непрерывному) закону позволяет использовать для обработки данных наблюдений хорошо изученные и детально разработанные статистические методы, формулируемые, в основном, на языке математического ожидания m и дисперсии σ^2 . Закон Пуассона не только упрощает статистический анализ ($m = \sigma^2$), но и предназначен для описания редких дискретных случайных событий. Все три закона являются устойчивыми в том смысле, что сумма независимых случайных величин, распределенных по любому из них, распределена по тому же закону, что и ее слагаемые [3].

Необходимость рассмотрения изложенного здесь материала связана с тем, что он является теоретическим (абстрактным) описанием и в каком-то смысле обоснованием практических статистических результатов обработки данных наблюдений, представленных, например, специалистами IARC [1]. В [1] под интенсивностью относительной частоты смерти от рака (по сути — функция риска) понимается отношение $\lambda = \frac{m}{PY}$, где m — число умерших за период наблюдения T ; $PY = \sum_{i=1}^N \tau_i = \sum_{i=1}^m t_i + (N - m)T$ — число человеко-лет наблюдения; $\tau_i = t_i = t_{\text{наст}_i}$ — время наступления смерти у i -го умершего, отсчитываемого от начала наблюдения ($i = \overline{1, m}$); $\tau_i = T$ для i от

$(m + 1)$ до N , если за период наблюдения T никакого нежелательного события не произошло. Справедливо неравенство $(N - m)T < PY < NT$ и, следовательно, $\frac{m}{NT} < \lambda = \frac{m}{PY} < \frac{m}{(N - m)T}$, т.е. величина λ находится в тех же достаточно узких ($m \ll N$) пределах, что и представленное здесь среднее значение функции риска $\lambda_{\text{ср}}$ (см. неравенство (21)).

Кроме того, в [1] используются, в основном, «нормальные» методы обработки данных. Однако, в Таблице 6 и Таблице 7 из [1] представлены результаты обработки данных по соответственно биномиальному и пуассоновскому законам, которые практически совпадают. Тем самым демонстрируется практическая эквивалентность трёх описанных выше законов распределения. Наше рассмотрение объясняет, почему это происходит.

Заключение

Таким образом, рассмотрение времени наступления нежелательного события в виде непрерывной конечной положительной случайной величины позволило построить при помощи функции риска ее функцию распределения и получить с различными степенями точности оценки среднего значения риска. В частном случае (экспоненциальное распределение) методом максимального правдоподобия получена статистическая оценка риска.

СПИСОК ЛИТЕРАТУРЫ

1. Boyle P., Parkin D. Statistical methods for registries. // In: Cancer Registration (Principles and Methods). — IARC Publication, Lyon, 1991, No. 95, P. 126–158.
2. Report of the United Nations Scientific Committee on the Effects of Atomic Radiation to the General Assembly. Annex A. Epidemiological studies of radiation and cancer, 2006, 310 pp.
3. Корн Г., Корн Т. Справочник по математике (для научных работников и инженеров). — М.: Наука, 1974, 832 с.
4. Asachenkov A., Marchuk G., Mohler R., Zuev S. Disease Models. International Institute for Applied System Analysis. — Laxenburg, Austria, 1994, 316 pp.
5. Волков И.К., Зув С.М., Цветкова Г.М. Случайные процессы. — М.: МГТУ им. Н.Э. Баумана, 1999, 448 с.

Поступила: 26.06.2015

Принята к публикации: 15.07.2015