ПЕРВЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ
1886
СПбГЭТУ «ЛЭТИ»

Учебное пособие

С. В. Малов И. Ю. Малова

БАЗОВЫЕ МОДЕЛИ БИОСТАТИСТИКИ

ИСПОЛЬЗОВАНИЕ ТОЧЕЧНЫХ ПРОЦЕССОВ В АНАЛИЗЕ ДАННЫХ ТИПА ВРЕМЕНИ ЖИЗНИ

Санкт-Петербург
2024

МИНОБРНАУКИ РОССИИ

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина)

С. В. МАЛОВ И. Ю. МАЛОВА

БАЗОВЫЕ МОДЕЛИ БИОСТАТИСТИКИ

ИСПОЛЬЗОВАНИЕ ТОЧЕЧНЫХ ПРОЦЕССОВ В АНАЛИЗЕ ДАННЫХ ТИПА ВРЕМЕНИ ЖИЗНИ

Учебное пособие

Санкт-Петербург
Издательство СПбГЭТУ «ЛЭТИ»
2024

УДК 519.2(07)
ББК В 172я7
М19

Малов С. В., Малова И. Ю.

М19 Базовые модели биостатистики: использование точечных процессов в анализе данных типа времени жизни: учеб. пособие. СПб.: Изд-во СПбГЭТУ «ЛЭТИ», 2024. 80 с.

ISBN 978-5-7629-3321-6

Рассмотрено использование методов математической статистики и теории случайных процессов в анализе данных типа времени жизни.

Предназначено для поддержки дисциплины «Биостатистика» на ФКТИ СПбГЭТУ «ЛЭТИ», будет интересно студентам и аспирантам, обучающимся по программам подготовки специалистов в области информационных технологий.

УДК 519.2(07)
ББК В 172я7

Рецензенты: кафедра высшей математики ВШТЭ СПбГУПТД; д-р техн. наук, проф. Л. В. Уткин (СПбПУ им. Петра Великого).

Утверждено
редакционно-издательским советом университета
в качестве учебного пособия

ISBN 978-5-7629-3321-6

Оглавление

Введение

Настоящее издание предназначено для глубокого изучения методов анализа данных типа времени жизни с правым цензурированием. Анализ данных типа времени жизни остается актуальным направлением в современной математической и прикладной статистике, а основные его методы реализованы практически во всех статистических пакетах, таких, как R, SPSS, Statistica и SAS. Более того, модули статистического анализа данных такого типа присутствуют в программных средах MatLab, Python и др. Несмотря на имеющиеся сложности постановки экспериментов, связанных с изучением времени жизни, безотказной работы механизма или же наступления некоторого события, характеризующего переход изучаемого объекта в новое качество, приобретение или утрату им некоторого свойства, данные типа времени жизни являются важной и неотъемлемой частью научных исследований в различных областях знаний. Главная сложность обычно связана с значительной продолжительностью исследования, в связи с чем многие времена наступления событий остаются недостижимыми и в работу включается механизм цензурирования справа. Цензурирование может быть обусловлено не только истечением времени, отведенного на проведение эксперимента, но и внешними факторами, не позволяющими выявить точное время наступления события. Идея последовательного накопления статистической информации с течением времени идеально согласуется с понятиями фильтрации и стохастического базиса, играющими важную роль в теории случайных процессов. Применение точечных процессов позволяет не только получить строгие математические методы анализа для широкого класса моделей, но и создает базу для построения новых методов и расширения класса известных моделей анализа неполных данных типа времени жизни.

Основная часть учебного пособия состоит из четырех разделов. *Первый раздел* включает в себя введение в теорию случайных процессов, а также формирование основных понятий и необходимые результаты теории мартингалов с приложениями к точечным процессам, которые в дальнейшем будут использованы для построения асимптотических методов анализа данных типа времени жизни. Во *втором разделе* рассматриваются методы анализа однородных данных типа времени жизни. *Третий раздел* посвящен построению асимптотической теории линейных ранговых статистик, применяемых для проверки однородности распределений двух или нескольких групп объектов исследования. В *четвертом разделе* изучаются модели анализа неоднородных данных типа времени жизни с правым цензурированием. Основное внимание уделяется мультипликативной и аддитивно-мультипликативной регрессионным моделям с возможностью использования зависящей от времени коварианты. Будет построена асимптотическая теория, позволяющая строить доверительные множества и проверять статистические гипотезы о параметрах моделей.

Механизм *цензурирования справа* наиболее часто применяется в анализе данных типа времени жизни. Результат статистического эксперимента представляет собой пару времен отказа и цензурирования (T, U), тогда как наблюдается только пара (X, δ), где $X = T \wedge U$ и $\delta = \mathbb{I}_{\{T \leq U\}}$. Иными словами, наблюдается лишь наименьшее из времен T или U, при этом известно, какое из событий (отказ или цензурирование) произошло раньше. Отображение $\Phi : \mathbb{R}^2_+ \to \mathbb{R}_+ \times \{0, 1\}$ реализует соотношение $(T, U) \mapsto (X, \delta)$. Элементарными событиями наблюдаемой σ-алгебры $\mathfrak{D}$ будут лучи вида $((x, \infty), \{x\})$ or $(\{x\}, [x, \infty))$ в $\mathbb{R}^2$. В зависимости от модели рассматривают выборку из распределения (T, U) с правым цензурированием или независимые, но не одинаково распределенные наблюдения (несколько выборок, регрессионная модель). В последнем случае с каждой парой (T, U) и, соответственно, наблюдением (X, δ) ассоциировано значение коварианты z. Отметим, что без дополнительных предположений о распределении (T, U) по выборке невозможно восстановить распределение времени отказа, так как оцениваемый параметр не идентифицируем. Наиболее часто предполагают, что времена отказа T и цензурирования U суть независимые случайные величины. Более слабое условие независимости цензурирования, позволяющее делать статистические выводы о распределениях времен цензурирования, будет рассмотрено позднее.

1. ВВЕДЕНИЕ В ТЕОРИЮ СЛУЧАЙНЫХ ПРОЦЕССОВ

1.1. Случайные процессы и их распределения

Пусть $(\Omega, \mathfrak{F}, \mathbb{P})$ – вероятностный эксперимент, Γ – множество, $\mathfrak{D}$ – σ-алгебра подмножеств Γ. Семейство случайных величин $\{X_t\}_{t \in \Gamma}$ будем называть *случайным процессом*, заданным на множестве $(\Gamma, \mathfrak{D})$.

Случайный процесс также можно определить как отображение $X : \Gamma \times \Omega \to \mathbb{R}$, такое, что при каждом фиксированном $t_0 \in \Gamma$ отображение

$X(t_0, \omega) : \Omega \to \mathbb{R}$ является случайной величиной. Функция $X(\cdot, \omega) : \Gamma \to \mathbb{R}$ называется *траекторией* случайного процесса, соответствующей исходу ω ($\omega \in \Omega$). Свойства траекторий $X(\cdot, \omega)$ переносятся на случайный процесс, если данное свойство выполнено с вероятностью 1 (почти наверное).

Распределение случайного процесса определяют конечномерные распределения (КМР)

$$P_{t_1,\dots,t_n}(A_1 \times \ldots \times A_n) = \mathbb{P}(X_{t_1} \in A_1, \ldots, X_{t_n} \in A_n), \ A_1, \ldots, A_n \in \mathcal{B}_1,$$

при всех $t_1, \ldots, t_n \in T$ и $n \in \mathbb{N}$ удовлетворяющие условиям согласования:

(i) для любой перестановки $(\sigma_1, \ldots, \sigma_n)$ индексов $\{1, \ldots, n\}$

$$P_{t_{\sigma_1},\dots,t_{\sigma_n}}(A_{\sigma_1} \times \ldots \times A_{\sigma_n}) = P_{t_1,\dots,t_n}(A_1 \times \ldots \times A_n);$$

(ii) для любого $n \in \mathbb{N}$ и любых $t_1, \ldots, t_n \in T$

$$P_{t_1,\dots,t_{n-1}}(A_1 \times \ldots \times A_{n-1}) = P_{t_1,\dots,t_n}(A_1 \times \ldots \times A_{n-1} \times \mathbb{R}).$$

Замечания: 1. Свойства (i) и (ii) являются характеристическими свойствами КМР случайного процесса, т. е. для любого семейства распределений, удовлетворяющих условиям согласования, существует вероятностный эксперимент и заданный на нем случайный процесс, имеющий соответствующие КМР.

2. Многие свойства траекторий случайного процесса, например непрерывность, не всегда определяются распределением случайного процесса.

Случайные величины X и Y на одном вероятностном эксперименте *стохастически эквивалентны*, если $\mathbb{P}(X \neq Y) = 0$. Случайный процесс X_t называется *модификацией* процесса Y_t, если для любого $t \in \Gamma$ величины X_t и Y_t стохастически эквивалентны. Случайные процессы X_t и Y_t *неразличимы*, если $\mathbb{P}(\omega : X_t = Y_t, \ t \in \Gamma) = 1$. Ясно, что неразличимые процессы являются модификациями друг друга. Далее считаем, что $\Gamma = \mathbb{N}$ (последовательность), $\Gamma = [0, a]$ при некотором $a > 0$, или $\Gamma = \mathbb{R}_+$.

Поскольку непрерывность траекторий случайного процесса на $\mathbb{R}_+$ не определяется его распределением, важно определить достаточные условия непротиворечивого введения свойства непрерывности в определение случайного процесса. Следующая теорема Колмогорова устанавливает достаточные условия существования непрерывной модификации случайного процесса.