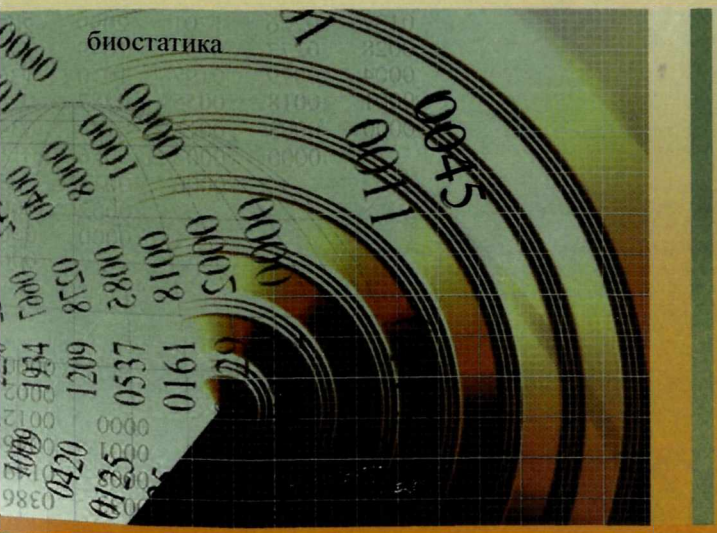
Учебное пособие

С. В. Малов И. Ю. Малова А. В. Процветкина

БАЗОВЫЕ МОДЕЛИ БИОСТАТИСТИКИ

АНАЛИЗ ДАННЫХ ТИПА ВРЕМЕНИ ЖИЗНИ



САНКТ-ПЕТЕРБУРГ
2023

В фонд библиотеки
СФУ
от автора

МИНОБРНАУКИ РОССИИ

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

С. В. МАЛОВ И. Ю. МАЛОВА А. В. ПРОЦВЕТКИНА

БАЗОВЫЕ МОДЕЛИ БИОСТАТИСТИКИ

АНАЛИЗ ДАННЫХ ТИПА ВРЕМЕНИ ЖИЗНИ

Учебное пособие

Санкт-Петербург
Издательство СПбГЭТУ «ЛЭТИ»
2023

УДК 519.2(07)
ББК В 172я7
М19

Малов С. В., Малова И. Ю., Процветкина А. В.
М19 Базовые модели биостатистики: анализ данных типа времени жизни: учеб. пособие. СПб.: Изд-во СПбГЭТУ «ЛЭТИ», 2023. 80 с.

ISBN 978-5-7629-3153-3

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 биоинформатики и математической биологии СПбАУ РАН им. Ж. И. Алфёрова; д-р техн. наук, проф. Л. В. Уткин (СПбПУ им. Петра Великого).

Утверждено
редакционно-издательским советом университета
в качестве учебного пособия

ISBN 978-5-7629-3153-3

Оглавление

Введение

Настоящее издание представляет собой описание моделей и основных методов обработки данных типа времени жизни. В современной математической статистике это направление является одним из наиболее популярных.

Существует множество прикладных задач, в которых ключевым объектом изучения является время до наступления некоторого события. В биомедицинских исследованиях таким событием может быть появление осложнений после операции, смерть, ремиссия. При анализе надежности можно говорить о выходе из строя технической системы или какой-либо ее составляющей. В общем случае, будем именовать интересующее событие отказом. Модели исследований такого вида изучаются в анализе данных типа времени жизни (анализе выживаемости).

Время отказа T – неотрицательная случайная величина, основной характеристикой которой является распределение. В определенных задачах можно заниматься исследованием распределения времени отказа в рамках одной выборки – ставить задачи оценивания этого распределения и проверки согласия с выдвигаемой гипотезой. Гораздо чаще интерес представляет сравнение распределений времени отказа в нескольких группах (выборках) – можно выдвигать и проверять гипотезы однородности. Часто исследователь вынужден учитывать влияние сопутствующих факторов, что неизбежно приводит к необходимости использования подходов, аналогичных регрессионным в классической статистике.

Исследователь должен четко представлять, что является точкой отсчета, относительно которой измеряется время отказа. Для биомедицинских исследований начало может быть как одинаковым для всех – время начала проекта (рандомизации), так и различным – момент хирургического вмешательства или усугубления болезни. Для технических систем такой точкой может быть начало эксплуатации.

Проведение биомедицинских исследований всегда ограничено по времени, вследствие чего ненаступление отказа в рамках установленного срока наблюдения вполне допустимо и, как показывает практика, весьма распространено. Довольно часто отказ не наблюдается ввиду наступления другого события, делающего невозможным продолжение наблюдения за объектом. Игнорировать такие результаты и просто исключить их из анализа нельзя, поскольку это ведет к появлению систематической ошибки оценивания. Возможны ситуации, когда точное время отказа не определено, а известен

лишь временной интервал, когда отказ произошел. Постановка эксперимента может быть такова, что отказ фиксируется только при выполнении определенных условий. Во всех случаях информация, получаемая в результате эксперимента, может быть неполной. Довольно часто часть информации теряется при статистических исследованиях. Говорят, что наблюдения потеряны совершенно случайно[1], если распределение набора пропущенных значений равномерное на множестве всех подмножеств индексов наблюдений и не зависит от самих наблюдений и их распределений, т. е. пропуск каждого значения наблюдаемого признака происходит независимо от его значения и распределения; потеряны случайно[2], если пропуск каждого значения происходит независимо от наблюдаемой величины, но может зависеть от ее распределения при различных значениях сопутствующих факторов. Потерянные случайно наблюдения можно исключить из анализа или естественным образом импутировать, но такой тип неполных данных нехарактерен для данных типа времени жизни, поэтому формальные аспекты работы с данными подобной структуры будут подробно изложены.

Обозначим

$$F(t) = \mathbb{P}(T \le t) \qquad \text{и} \qquad S(t) = 1 - F(t), \quad t \in \mathbb{R},$$

функцию распределения и функцию отказа случайной величины T соответственно. В анализе данных типа времени жизни удобнее пользоваться функцией отказа, нежели функцией распределения. Обычно предполагают, что T абсолютно непрерывна с плотностью f, но в определенных случаях имеет смысл рассматривать и распределения общего вида.

Определения: 1. Интенсивностью отказа случайной величины T по отношению к доминирующей мере μ будем называть следующую функцию:

$$\lambda(t) = \lim_{\Delta t \to 0} \mathbb{P}(t \le T < t + \Delta t | T \ge t)/\mu(\Delta t) = \\ = -\frac{dS(t)}{d\mu(t)}/S(t_-) = \frac{f(t)}{S(t_-)}, \quad t \in \mathbb{R}.$$

2. Накопленной интенсивностью будем называть функцию

$$\Lambda(t) = \int_0^t \lambda(u)\, d\mu(u) = \int_0^t \lambda(u)\, d\mu(u), \ t \in \mathbb{R}.$$

[1] Missed completely at random, *англ.*

[2] Missed at random, *англ.*

Замечания: 1. В случае абсолютно непрерывного распределения T в качестве меры μ используется мера Лебега, а следовательно, $\lambda(t) = -\frac{d \ln S(t)}{dt}$ и

$$S(t) = \exp(-\Lambda(t)), \quad t \in \mathbb{R}.$$

2. Для дискретного распределения в качестве μ используют считающую меру на множестве E, содержащем множество значений T, а следовательно, $\lambda(t) = \mathbb{P}(T = t)/S(t)$ и

$$S(t) = \prod_{u \le t}(1 - \Lambda(du)) = \prod_{u \in E \cap [0,t]}(1 - \lambda(u)).$$

3. В общем случае

$$S(t) = \int_0^t \frac{-S(du)}{S(u_-)} = \exp(-\Lambda^c(t)) \prod_{u \le t}(1 - \Lambda^d(du)),$$

где интеграл понимается в смысле Лебега–Стильтьеса; Λ^c и Λ^d – накопленные интенсивности непрерывной и дискретной компонент соответственно.

Рассмотрим примеры задач, возникающих при анализе данных типа времени жизни.

1. В период с 1962 по 1977 гг. 225 человек перенесли операцию по удалению опухоли. Из них 20 человек отказались от наблюдения, т. е. были получены данные по 205 пациентам. Цель исследования: оценить распределение времени жизни после операции. Следующие положения должны быть учтены при обработке данных:

а) пациенты поступали на операцию в различные моменты времени (время поступления для каждого больного принимается нулевым);

б) время смерти установлено лишь для погибших до 1977 г.;

в) часть пациентов погибли по причинам, не зависящим от операции;

г) возможно, имеет смысл учесть факторы риска (возраст, пол, стадия заболевания в момент обнаружения).

2. В одном из европейских городов с 1973 по 1981 гг. исследовалось влияние диабета на продолжительность жизни людей. Всего было зарегистрировано 1499 человек. Возможные задачи: оценить распределение продолжительности жизни; сравнить распределения времени жизни больного диабетом и здорового человека. При этом следует учитывать следующее:

а) начало и завершение наблюдений в различном возрасте;

б) продолжительность болезни к моменту начала наблюдений неизвестна;

в) начало наблюдений в различном возрасте.

3. Предположим, что необходимо оценить в ограниченный срок распределение времени работы некоторой (достаточно надежной) системы. Если тестировать систему в обычном режиме, то процент отказов за отведенное время в независимых испытаниях может быть крайне мал. Ясно, что не будет получено никакой информации о распределении отказа за пределами времени, отведенного на испытание. Для повышения эффективности исследования предлагается проводить испытания в более жестких условиях. Необходимо выбрать подходящую модель, позволяющую связать распределения в различных режимах. Обычно жесткую связь установить сложно, поэтому предполагается наличие параметрической связи, диктуемой свойствами системы. В этом случае оценка распределения времени отказа системы строится с учетом данной параметрической зависимости.

Учебное пособие разбито на 5 разделов. Первый раздел посвящается постановке задач, а также описанию моделей неполных данных (цензурирования и усечения) и некоторых общих подходов к их анализу. Во втором и третьем разделах обсуждаются параметрические и непараметрические методы анализа цензурированных данных типа времени жизни соответственно. В четвертом разделе рассмотрены основные методы анализа интервально-цензурированных данных. В заключительном разделе приведены компьютерные реализации и примеры использования методов анализа данных типа времени жизни.

1. МОДЕЛИ И ОСНОВНЫЕ ПОДХОДЫ К АНАЛИЗУ НЕПОЛНЫХ ДАННЫХ

1.1. Модели неполных данных

Рассмотрим некоторые модели анализа неполных данных типа времени жизни, в основе которых лежат принципы цензурирования и усечения. Цензурированием называют потерю части статистической информации, обусловленную случайным искажением наблюдения, тогда как усечение – потеря части данных ввиду нарушения специфических условий наблюдаемости, определяемых, возможно, случайными факторами[3]. Цензурированные данные обычно получают отображением исходной статистической модели эксперимента, включающей изучаемую характеристику и механизм цензурирования, в модель цензурированных данных, а усечение реализуется введением усеченных распределений.

[3] В работе Turnbull (1976) были введены 3 типа данных: полные, цензурированные и усеченные.

Сформулируем основные принципы построения моделей неполных данных. Пусть $(\mathfrak{X}^\circ \times \mathfrak{X}_1, \mathfrak{F}^\circ = \sigma(\mathfrak{F}^\circ \times \mathfrak{F}_1), \mathcal{P}^\circ)$, $\mathcal{P}^\circ = \{\mathbf{P}_\theta, \theta \in \Theta\}$ – статистический эксперимент для полных данных. Усечение подразумевает потерю части данных в эксперименте $(\mathfrak{X}^\circ, \mathfrak{F}^\circ, \mathcal{P}^\circ)$, $\mathcal{P}^\circ = \{P^\circ_\theta, \theta \in \Theta\}$, где P°_θ – соответствующие распределения компонент на $(\mathfrak{X}^\circ, \mathfrak{F}^\circ)$, а именно, в зависимости от реализации условий из $\mathfrak{X}_1$ данный статистический эксперимент преобразуется в эксперимент меньшей размерности $(\mathfrak{X}, \mathfrak{F}, \mathcal{P})$, $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, где $\mathfrak{X}$ содержит только неусеченные компоненты $\mathfrak{X}^\circ$, а их распределение $\mathbb{P}_\theta = \Psi(\mathbf{P}_\theta)$ определяется отображением $\Psi : \mathcal{P}^\circ \to \mathcal{P}$. Цензурирование предполагает, что в результате эксперимента наблюдается лишь часть событий $\mathfrak{D} \subset \mathfrak{F}$. Тогда $(\mathfrak{X}, \mathfrak{D}, \mathcal{P})$ – статистический эксперимент с цензурированием. Обычно цензурирование задают с использованием цензурирующего отображения $\Phi : (\mathfrak{X}, \mathfrak{F}) \to (\mathcal{Y}, \mathcal{C})$, такого, что $\mathfrak{D} = \{\Phi^{-1}(A) : A \in \mathcal{C}\}$. Функция Φ индуцирует семейство распределений $\mathcal{Q}_\theta = \{\mathbb{Q}_\theta : \theta \in \Theta\}$ на $(\mathcal{Y}, \mathcal{C})$: $\mathbb{Q}_\theta(\Phi(B)) = \mathbb{P}_\theta(B)$ для любого $B \in \mathfrak{F}$. Условие идентифицируемости распределения состоит в том, что существует взаимно-однозначное соответствие между θ и $\mathbb{Q}_\theta$. Наиболее часто интерес представляет лишь параметр, определяющий распределение времени отказа, поэтому идентифицируемость мешающего параметра, связанного с механизмом цензурирования, не является обязательной. Неинформативность усечения и цензурирования подразумевает использование несвязанных параметров для распределения отказа, цензурирования и условий усечения.

Модель цензурирования справа наиболее часто применяется при анализе данных типа времени жизни. Статистический эксперимент включает в себя пару времен отказа и цензурирования (T, U), но наблюдается лишь время наступления события $X = T \wedge U$ (наименьшее из времен T или U) и его тип $\delta = \mathbb{1}_{\{T \le U\}}$ (отказ при $\delta = 0$ или цензурирование при $\delta = 1$). 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 $((x, \infty), \{x\})$ и $(\{x\}, [x, \infty))$ в $\mathbb{R}^2$ при $x \ge 0$. В зависимости от плана эксперимента рассматривают выборку из распределения T с правым цензурированием U или независимые, но не одинаково распределенные наблюдения (несколько выборок, регрессионная модель). В последнем случае с каждой парой (T, U) ассоциировано значение ковариаты z. Отметим, что без дополнительных предположений относительно распределения (T, U) по выборке невозможно восстановить распределение времени отказа, так как соответ-