Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Дипломная работа

студента 522-й группы
Коробейникова Антона Ивановича

«Об оценке параметров специальной модели кривых дожития»

Научный руководитель
к.ф.-м.н., доцент А.Г. Барт
Научный руководитель
к.ф.-м.н., доцент Н.П. Алексеева

Рецензент
к.ф.-м.н., доцент В.В. Некруткин

«Допустить к защите» _____
Заведующий кафедрой
д.ф.-м.н., профессор С.М. Ермаков

Санкт-Петербург
2007г.

Содержание

1. Введение

Под *кривой дожития* понимается выборочное описание заданной характеристики объекта при его элиминации (или до *наработки на отказ*). Как правило, наблюдаемой характеристикой является само время элиминации. Широкому спектру реальных приложений кривых дожития соответствует и разнообразие математических моделей [5, 9, 19]. Основной генеральной моделью кривых дожития является функция надежности. Достаточно распространен термин *теория надежности*, отражающий, в основном, свойства именно этой модели [9], а принятая терминология ориентирована на технические приложения.

В настоящей работе будет рассмотрена специальная модель кривой дожития в виде произведения экспоненты на косинус (представлена уравнением (2.1)), впервые предложенная в [1]. Основанием для использования этой модели в биометрии служит лагранжево-гамильтоновый формализм, разработанный в [4] для биологической систематики. Согласно ему компоненты в виде произведения экспоненты на косинус являются решением системы, описывающей взаимодействие факторов органа и организма. Модель применялась для описания динамики хронического гломерулонефрита [1], гипертонической болезни [10], раневых процессов [3], хронического генерализованного пародонтита [6].

Модель представляется удобной для интерпретации экспериментаторами реальных данных, так как позволяет учесть сложный дискретно-непрерывный характер процессов патогенеза–саногенеза, выражающийся в наличии чередующихся периодов обострения и ремиссии. Кроме этого, существует возможность модификации модели, позволяющая включить управление через процедуру двойного обращения [2].

Ранее оценки параметров этой модели, как правило, получались путем минимизации расстояния между эмпирической и модельной функцией распределения. Однако статистические свойства таких оценок были неизвестны. В настоящей работе предложена новая процедура оценивания параметров модели и изучены свойства полученных оценок.

Отметим, что при изучении данных типа времени жизни, как правило, приходится сталкиваться с целым рядом проблем.

1. Проблемы, связанные с процедурой проведения эксперимента.

   Во многих экспериментальных процессах или процессах сбора данных получают не значение изучаемой случайной величины (которая ненаблюдаема по каким-либо при-

чинам), а значение менее информативной случайной величины. Среди возможных проявлений данного эффекта выделим следующие:

(a) Цензурирование (известно лишь количество наблюдений со значениями в определенной области).

(b) Усечение (наблюдения не могут принимать значения в определенной области).

(c) Группировка (для наблюдений известен лишь некоторый интервал, в пределах которого находятся значения).

(d) Определение момента начала наблюдений.

2. Проблемы, связанные с моделью кривых дожития.

   Как правило, это проблемы аналитического характера: отсутствие у модельной функции распределения (или плотности) требуемой гладкости, невыполнение различных условий регулярности и т.п., что приводит к невозможности применить большинство классических методов непосредственно.

   Так например, проблема определения момента начала наблюдений может быть решена за счет введения дополнительного параметра сдвига. Однако, поскольку наблюдаемая величина неотрицательна, то, очевидно, модельное распределение должно быть сосредоточено на неотрицательной полупрямой. Введение дополнительного параметра сдвига приводит к тому, что носитель распределения оказывается зависимым от неизвестного параметра и, как следствие этого, классическая теория оценок максимального правдоподобия не может быть применена для оценивания параметра сдвига.

Для решения задачи определения момента начала наблюдений оригинальная модель из [1] была модифицирована введением параметра сдвига. Кроме этого, считалось, что других проблем, связанных с процедурой проведения эксперимента не было, то есть отсутствовали цензурирование, усечение и группировка.

2. Математическая модель и обозначения

Модельная функция распределения задается уравнением

$$F\left(x;\,\eta,\tau,\mu\right)=1-\exp\left(-\eta\left(\frac{x-\mu}{\tau}\right)\right)\cos\left(\frac{\pi}{2}\left(\frac{x-\mu}{\tau}\right)\right),\qquad \eta>0,\ \mu<x<\mu+\tau. \tag{2.1}$$

Задача оценивания параметров данного распределения относится к так называемому «нерегулярному типу»:

1. Носитель распределения зависит от значений неизвестных параметров μ и τ.

2. Плотность распределения имеет ненулевые пределы на границе носителя.

Всюду далее η_0, τ_0, μ_0 будут обозначать истинные значения параметров η, τ, μ, соответственно. Пусть $X_1,\ldots,X_n$ — набор из n независимых одинаково распределенных случайных величин с функцией распределения, задаваемой уравнением (2.1). Обозначим логарифм функции правдоподобия через L_n:

$$L_n\left(\eta,\tau,\mu\right)=\frac{1}{n}\sum_{i=1}^{n}\log f\left(X_i;\eta,\tau,\mu\right), \tag{2.2}$$

здесь f — плотность, соответствующая функции распределения (2.1):

$$f\left(x;\,\eta,\tau,\mu\right)=\exp\left(-\frac{\eta(x-\mu)}{\tau}\right)\left(\frac{\eta}{\tau}\cos\left(\frac{\pi(x-\mu)}{2\tau}\right)+\frac{\pi}{2\tau}\sin\left(\frac{\pi(x-\mu)}{2\tau}\right)\right),\ \mu<x<\mu+\tau.$$

Порядковые статистики будут обозначаться как $X_{[1;n]}<\cdots<X_{[n;n]}$. Заметим, что функция L_n корректно определена только при $\mu<X_{[1;n]}$ и $X_{[n;n]}<\mu+\tau$.

Нижний индекс 0 будет использоваться для обозначения функции распределения и плотности с параметрами η_0,τ_0,μ_0:

$$F_0\left(x\right)=F\left(x;\,\eta_0,\tau_0,\mu_0\right),\quad f_0\left(x\right)=f\left(x;\,\eta_0,\tau_0,\mu_0\right).$$

Аналогично, символом $\mathbf{E}_0$ будем обозначать математическое ожидание с плотностью f_0.

Оценку максимального правдоподобия для η при известных значениях μ_0, τ_0 обозначим $\bar{\eta}_n$:

$$\bar{\eta}_n=\arg\max_{\eta>0}L_n\left(\eta,\tau_0,\mu_0\right).$$

Оценка $\bar{\eta}_n$ необходимо удовлетворяет уравнению

$$\frac{\partial L_n}{\partial\eta}\left(\bar{\eta}_n,\tau_0,\mu_0\right)=0. \tag{2.3}$$

Существование, состоятельность и прочие свойства $\bar{\eta}_n$ следуют из классических результатов для оценок максимального правдоподобия в регулярном случае.

Известно, что оценки максимального правдоподобия в нерегулярном случае могут быть несостоятельными (смотри, например, обзор в [11] и примеры в [21, 22]). Тем не менее, параметры μ и τ можно оценивать эффективно при помощи порядковых статистик:

$$\hat{\mu}_n = X_{[1;n]}, \quad \hat{\tau}_n = X_{[n;n]} - X_{[1;n]}. \tag{2.4}$$

Свойства этих оценок будут рассмотрены подробнее в разделе 3.1.

Для оценивания параметров воспользуемся несколько модифицированным методом максимального правдоподобия. Следуя [21], рассмотрим двухстадийную процедуру оценивания.

1. Эффективное оценивание параметров τ и μ при помощи оценок, приведенных в формуле (2.4).

2. Получение оценки $\tilde{\eta}_n$ как локального максимума логарифма модифицированной функции правдоподобия $\tilde{L}_n$:

$$\tilde{L}_n\left(\eta, \tau, \mu\right) = \frac{1}{n} \sum_{i=2}^{n-1} \log f\left(X_{[i;n]}; \eta, \tau, \mu\right), \tag{2.5}$$

$$\tilde{\eta}_n = \arg\max_{\eta > 0} \tilde{L}_n\left(\eta, \hat{\tau}_n, \hat{\mu}_n\right). \tag{2.6}$$

Следовательно, оценка $\tilde{\eta}_n$ должна удовлетворять уравнению

$$\frac{\partial \tilde{L}_n}{\partial \eta}\left(\tilde{\eta}_n, \hat{\tau}_n, \hat{\mu}_n\right) = 0. \tag{2.7}$$

Свойства этой оценки будут изучены в разделе 3.2.

3. Оценки параметров

3.1. Предварительные замечания и свойства оценок $\hat{\mu}_n$ и $\hat{\tau}_n$

В дальнейшем нам понадобятся две леммы. Лемма 1 представляет самостоятельный интерес, так как дает ответ на вопрос об асимптотических свойствах оценок $\hat{\mu}_n$ и $\hat{\tau}_n$, а лемма 2 носит сугубо технический характер.

Лемма 1. Для оценок $\hat{\tau}_n$, $\hat{\mu}_n$ выполняется:

$$\hat{\mu}_n - \mu_0 = O_p\left(\frac{1}{n}\right),\ n \to \infty, \qquad \tau_0 - \hat{\tau}_n = O_p\left(\frac{1}{n}\right),\ n \to \infty.$$

Доказательство.

$$\begin{aligned}\mathbf{P}\left(n\left(\hat{\mu}_n - \mu_0\right) > t\right) &= \mathbf{P}\left(X_{[1;n]} > \mu_0 + \frac{t}{n}\right) = \left(1 - F_0\left(\mu_0 + \frac{t}{n}\right)\right)^n = \\ &= \exp\left(-\frac{\eta_0}{\tau_0}t\right)\cos^n\left(\frac{\pi}{2\tau_0}\frac{t}{n}\right) = \exp\left(-\frac{\eta_0}{\tau_0}t\right)\left(1 - O\left(\frac{1}{n^2}\right)\right)^n = \\ &= \exp\left(-\frac{\eta_0}{\tau_0}t\right)\exp\left(n\ln\left(1 - O\left(\frac{1}{n^2}\right)\right)\right) = \exp\left(-\frac{\eta_0}{\tau_0}t\right) + O\left(\frac{1}{n}\right),\ t > 0. \quad (3.1)\end{aligned}$$

Так как $n\left(\tau_0 - \hat{\tau}_n\right) = n\left(\tau_0 + \mu_0 - X_{[n;n]}\right) + n\left(\hat{\mu}_n - \mu_0\right)$, то для доказательства второго утверждения достаточно показать, что

$$\left(\tau_0 + \mu_0\right) - X_{[n;n]} = O_p\left(\frac{1}{n}\right),\ n \to \infty.$$

Действительно, имеем

$$\begin{aligned}\mathbf{P}\left(n\left(\tau_0 + \mu_0 - X_{[n;n]}\right) > t\right) &= \mathbf{P}\left(X_{[n;n]} < \tau_0 + \mu_0 - \frac{t}{n}\right) = \\ &= \left(F_0\left(\tau_0 + \mu_0 - \frac{t}{n}\right)\right)^n = \exp\left(-\frac{\exp\left(-\eta_0\right)\pi}{2\tau_0}\,t\right) + O\left(\frac{1}{n}\right),\ t > 0. \quad (3.2)\end{aligned}$$

■

Следствие. В утверждении фактически было получено асимптотическое распределение для оценки $\hat{\mu}_n$ после соответствующего центрирования и нормирования: из формулы (3.1) непосредственно следует, что случайная величина $n\left(\hat{\mu}_n - \mu_0\right)$ имеет асимптотически экспоненциальное распределение с параметром $\dfrac{\eta_0}{\tau_0}$.

Лемма 2. Пусть случайная величина $\eta_n > 0$ сходится по вероятности к константе $\eta_0 > 0$. $\tau_n \xrightarrow[n\to\infty]{\mathbf{P}} \tau_0$, $\tau_n > 0, \tau_0 > 0$; а случайная величина μ_n ограничена: $\mu_0 < \mu_n < X_{[1;n]}$. Тогда для вторых частных производных функции правдоподобия L_n выполняется:

$$\frac{\partial^2 L_n}{\partial\eta\,\partial\mu}\left(\eta_n,\,\tau_n,\,\mu_n\right) \leqslant_{\mathbf{P}} C(\tau_0,\eta_0), \quad \frac{\partial^2 L_n}{\partial\eta\,\partial\tau}\left(\eta_n,\,\tau_n,\,\mu_n\right) \leqslant_{\mathbf{P}} C(\tau_0,\eta_0), \tag{3.3}$$

$$-\frac{\partial^2 L_n}{\partial\eta^2}\left(\eta_n,\tau_n,\mu_n\right) \xrightarrow[n\to\infty]{\mathbf{P}} \sigma^2\left(\eta_0,\tau_0\right), \tag{3.4}$$

$$\sigma^2\left(\eta_0,\tau_0\right) = -\int\limits_{\mu_0}^{\mu_0+\tau_0} \frac{\partial^2}{\partial\eta^2}\log f_0(x)\,dF_0\left(x\right). \tag{3.5}$$

Здесь символом $\leqslant_{\mathbf{P}}$ обозначена ограниченность по вероятности.

Доказательство. Ограниченность по вероятности вторых частных производных функции правдоподобия $\dfrac{\partial^2 L_n}{\partial\eta\,\partial\mu}$ и $\dfrac{\partial^2 L_n}{\partial\eta\,\partial\tau}$ следует из их явного вида и ограниченности носителя распределения (2.1): действительно, обозначая через $\theta_{i,n}$ величину $\dfrac{\pi\left(X_i-\mu_n\right)}{2\tau_n}$, имеем:

$$\frac{\partial^2 L_n}{\partial\eta\,\partial\mu}(\eta_n,\,\tau_n,\,\mu_n) = \frac{1}{\tau_n} + \frac{1}{n}\sum_{i=1}^{n}\frac{\pi^2}{\tau_n\left(2\eta_n\cos\theta_{i,n}+\pi\sin\theta_{i,n}\right)^2} \leqslant \frac{1}{\tau_n} + \frac{\pi^2}{\tau_n\left(\min\left\{\pi,2\eta_n\right\}\right)^2},$$

Аналогично,

$$\frac{\partial^2 L_n}{\partial\eta\,\partial\tau}(\eta_n,\,\tau_n,\,\mu_n) = \frac{1}{n}\sum_{i=1}^{n}\frac{\left(X_i-\mu_n\right)}{\tau_n^2}\left[1+\frac{\pi^2}{\left(2\eta_n\cos\theta_{i,n}+\pi\sin\theta_{i,n}\right)^2}\right] \leqslant \frac{1}{\tau_n} + \frac{\pi^2}{\tau_n\left(\min\left\{\pi,2\eta_n\right\}\right)^2}.$$

Но функция

$$g\left(x,y\right) = \frac{1}{x} + \frac{\pi^2}{x\left(\min\left\{\pi,2y\right\}\right)^2}$$

непрерывна в точке (τ_0,η_0), а, значит,

$$\frac{1}{\tau_n} + \frac{\pi^2}{\tau_n\left(\min\left\{\pi,2\eta_n\right\}\right)^2} \xrightarrow[n\to\infty]{\mathbf{P}} \frac{1}{\tau_0} + \frac{\pi^2}{\tau_0\left(\min\left\{\pi,2\eta_0\right\}\right)^2},$$

что доказывает ограниченность по вероятности $\dfrac{\partial^2 L_n}{\partial\eta\,\partial\mu}(\eta_n,\,\tau_n,\,\mu_n)$ и $\dfrac{\partial^2 L_n}{\partial\eta\,\partial\tau}(\eta_n,\,\tau_n,\,\mu_n)$.

Перейдем к доказательству (3.4). Заметим, что в силу закона больших чисел имеет место:

$$-\frac{\partial^2 L_n}{\partial\eta^2}\left(\eta_0,\tau_0,\mu_0\right) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\eta^2}\log f_0\left(X_i\right) \xrightarrow[n\to\infty]{\mathbf{P}} \sigma^2\left(\eta_0,\tau_0\right). \tag{3.6}$$

Функция

$$\frac{\partial^2 L_n}{\partial\eta^2}\left(\eta_n,\tau_n,\mu_n\right) = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\left(\eta_n+\frac{\pi}{2}\operatorname{tg}\theta_{i,n}\right)^2} \tag{3.7}$$

непрерывна в точке (η_0, τ_0, μ_0). Поэтому

$$\frac{\partial^2}{\partial \eta^2}\left(L_n\left(\eta_n, \tau_n, \mu_n\right) - L_n\left(\eta_0, \tau_0, \mu_0\right)\right) \xrightarrow[n\to\infty]{\mathbf{P}} 0. \tag{3.8}$$

Объединяя соотношения (3.6) и (3.8), получаем (3.4). ■

Замечание. Так как

$$\int\limits_{\mu_0}^{\mu_0+\tau_0} \frac{\partial}{\partial \eta} \log f_0\left(x\right) dF_0\left(x\right) = 0,$$

то

$$\sigma^2\left(\eta_0, \tau_0\right) = -\int\limits_{\mu_0}^{\mu_0+\tau_0} \frac{\partial^2}{\partial \eta^2} \log f_0\left(x\right) dF_0\left(x\right) = \int\limits_{\mu_0}^{\mu_0+\tau_0} \left(\frac{\partial}{\partial \eta} \log f_0\left(x\right)\right)^2 dF_0\left(x\right) > 0.$$

Замечание. Все результаты утвержения верны и для функции $\tilde{L}_n$ из формулы (2.5).

3.2. Основные свойства оценки $\tilde{\eta}_n$

В данном разделе будет доказана теорема, дающая ответ на вопрос о статистических свойствах оценки $\tilde{\eta}_n$. Для доказательства будет применена техника, использовавшаяся в [21] для получения свойств оценок в регулярном случае.

Теорема 1. Оценка $\tilde{\eta}_n$, удовлетворяющая уравнению (2.7) существует и притом единственна. Кроме того, с вероятностью, стремящейся к 1 при $n \to \infty$ имеет место

$$\tilde{\eta}_n - \bar{\eta}_n = o_p\left(\frac{1}{n^{1-\varepsilon}}\right), \ \forall \varepsilon : 0 < \varepsilon < 1. \tag{3.9}$$

Доказательство. Существование и единственность $\tilde{\eta}_n$ почти очевидны: функция $\tilde{L}_n(\eta, \hat{\tau}_n, \hat{\mu}_n)$ выпукла по η. Например, это видно из формулы (3.7).

Перейдем к доказательству соотношения (3.9). Положим

$$\begin{aligned}\xi_n = \frac{\partial}{\partial \eta}\tilde{L}_n\left(\bar{\eta}_n, \tau_0, \mu_0\right) = \frac{\partial}{\partial \eta}\left(\tilde{L}_n - L_n\right)\left(\bar{\eta}_n, \tau_0, \mu_0\right) = \\ = \frac{1}{n}\left[\frac{\partial}{\partial \eta} \log f\left(X_{[1;n]};\, \bar{\eta}_n, \tau_0, \mu_0\right) + \frac{\partial}{\partial \eta} \log f\left(X_{[n;n]};\, \bar{\eta}_n, \tau_0, \mu_0\right)\right].\end{aligned} \tag{3.10}$$

Раскладывая $\dfrac{\partial}{\partial \eta}\tilde{L}_n$ по формуле Тейлора вблизи точки $(\bar{\eta}_n, \tau_0, \mu_0)$, имеем:

$$\begin{aligned}\frac{\partial}{\partial \eta}\tilde{L}_n\left(\eta, \hat{\tau}_n, \hat{\mu}_n\right) = \xi_n + \left(\hat{\mu}_n - \mu_0\right)\frac{\partial^2 \tilde{L}_n}{\partial \eta \partial \mu}\left(\eta^*, \tau^*, \mu^*\right) + \\ + \left(\hat{\tau}_n - \tau_0\right)\frac{\partial^2 \tilde{L}_n}{\partial \eta \partial \tau}\left(\eta^*, \tau^*, \mu^*\right) + \left(\eta - \bar{\eta}_n\right)\frac{\partial^2 \tilde{L}_n}{\partial \eta^2}\left(\eta^*, \tau^*, \mu^*\right),\end{aligned} \tag{3.11}$$

Здесь:

$$(\eta^*, \tau^*, \mu^*) = \tilde{\lambda}\left(\eta, \hat{\tau}_n, \hat{\mu}_n\right) + (1-\tilde{\lambda})\left(\bar{\eta}_n, \tau_0, \mu_0\right),\ 0 < \tilde{\lambda} < 1.$$

Пусть δ_n — такая последовательность, что $\delta_n \to 0$, $n\delta_n \to \infty$. Рассмотрим функцию $g_n(y)$:

$$g_n(y) = \frac{1}{\delta_n^2}\tilde{L}_n\left(\bar{\eta}_n + \delta_n y, \hat{\tau}_n, \hat{\mu}_n\right). \tag{3.12}$$

Используя формулу (3.11), получаем:

$$\frac{\partial g_n}{\partial y}(y) = \frac{1}{\delta_n}\xi_n + (\hat{\mu}_n - \mu_0)\frac{1}{\delta_n}\frac{\partial^2 \tilde{L}_n}{\partial\eta\partial\mu}(\eta^*, \tau^*, \mu^*) + \\ + (\hat{\tau}_n - \tau_0)\frac{1}{\delta_n}\frac{\partial^2 \tilde{L}_n}{\partial\eta\partial\tau}(\eta^*, \tau^*, \mu^*) + y\frac{\partial^2 \tilde{L}_n}{\partial\eta^2}(\eta^*, \tau^*, \mu^*). \tag{3.13}$$

Заметим, кроме того, что так как $\tau_0 < \tau^* < \hat{\tau}_n$ и $\mu_0 < \mu^* < \hat{\mu}_n$, то, из леммы 1:

$$\mu^* \xrightarrow[n\to\infty]{\mathbf{P}} \mu_0, \quad \tau^* \xrightarrow[n\to\infty]{\mathbf{P}} \tau_0. \tag{3.14}$$

Далее, отметим, что оценка $\bar{\eta}_n$ состоятельна, поэтому:

$$\eta^* = \bar{\eta}_n + \tilde{\lambda} y \delta_n \xrightarrow[n\to\infty]{\mathbf{P}} \eta_0 \tag{3.15}$$

Рассмотрим слагаемые в формуле (3.13) по отдельности:

1. $$\frac{1}{\delta_n}\xi_n = \frac{1}{n\delta_n}\left[\frac{\partial}{\partial\eta}\log f\left(X_{[1;n]};\, \bar{\eta}_n, \tau_0, \mu_0\right) + \frac{\partial}{\partial\eta}\log f\left(X_{[n;n]};\, \bar{\eta}_n, \tau_0, \mu_0\right)\right].$$

   В силу леммы 1:

   $$X_{[1;n]} \xrightarrow[n\to\infty]{\mathbf{P}} \mu_0, \quad X_{[n;n]} \xrightarrow[n\to\infty]{\mathbf{P}} \mu_0 + \tau_0.$$

   Функция

   $$t(x,y) = \frac{\partial}{\partial\eta}\log f\left(x;\, y, \tau_0, \mu_0\right) = -\frac{\pi\left(x-\mu_0\right)}{\tau_0} + \frac{\cos\left(\frac{\pi(x-\mu_0)}{2\tau_0}\right)}{y\cos\left(\frac{\pi(x-\mu_0)}{2\tau_0}\right) + \frac{\pi}{2\tau_0}\sin\left(\frac{\pi(x-\mu_0)}{2\tau_0}\right)}$$

   непрерывна в точках (μ_0, η_0), $(\mu_0 + \tau_0, \eta_0)$, следовательно

   $$\frac{\partial}{\partial\eta}\log f\left(X_{[1;n]};\, \bar{\eta}_n, \tau_0, \mu_0\right) \xrightarrow[n\to\infty]{\mathbf{P}} \frac{\partial}{\partial\eta}\log f\left(\mu_0;\, \eta_0, \tau_0, \mu_0\right) = \frac{1}{\eta_0},$$

   $$\frac{\partial}{\partial\eta}\log f\left(X_{[n;n]};\, \bar{\eta}_n, \tau_0, \mu_0\right) \xrightarrow[n\to\infty]{\mathbf{P}} \frac{\partial}{\partial\eta}\log f\left(\mu_0 + \tau_0;\, \eta_0, \tau_0, \mu_0\right) = -\pi$$

   и

   $$\frac{1}{\delta_n}\xi_n \xrightarrow[n\to\infty]{\mathbf{P}} 0.$$

2. Имеет место сходимость (лемма 1):

$$\frac{\hat{\mu}_n - \mu_0}{\delta_n} \xrightarrow[n\to\infty]{\mathbf{P}} 0.$$

С учетом (3.14) и (3.15) из леммы 2 получаем:

$$\frac{\partial^2 L_n}{\partial\eta\,\partial\mu}\left(\eta^*,\,\tau^*,\,\mu^*\right) \leqslant_{\mathbf{P}} C(\tau_0,\eta_0),$$

поэтому

$$(\hat{\mu}_n - \mu_0)\,\frac{1}{\delta_n}\frac{\partial^2 L_n}{\partial\eta\,\partial\mu}\left(\eta^*,\,\tau^*,\,\mu^*\right) \xrightarrow[n\to\infty]{\mathbf{P}} 0.$$

3. Аналогично получается соотношение:

$$(\hat{\tau}_n - \tau_0)\,\frac{1}{\delta_n}\frac{\partial^2 L_n}{\partial\eta\,\partial\tau}\left(\eta^*,\,\tau^*,\,\mu^*\right) \xrightarrow[n\to\infty]{\mathbf{P}} 0.$$

4. И, наконец, применяя вторую часть леммы 2 к $\dfrac{\partial^2 \tilde{L}_n}{\partial\eta^2}\left(\eta^*,\tau^*,\mu^*\right)$, заключаем:

$$-\frac{\partial^2 \tilde{L}_n}{\partial\eta^2}\left(\eta^*,\tau^*,\mu^*\right) \xrightarrow[n\to\infty]{\mathbf{P}} \sigma^2\left(\eta_0,\tau_0\right).$$

Собирая все вместе, получаем:

$$\frac{\partial g_n}{\partial y}(y) = -y\sigma^2\left(\eta_0,\tau_0\right) + \varepsilon_n, \quad \varepsilon_n \xrightarrow[n\to\infty]{\mathbf{P}} 0.$$

Следовательно, с вероятностью, стремящейся к 1 при $n \to \infty$, имеем:

$$\frac{\partial g_n}{\partial y}(-1) > 0, \quad \frac{\partial g_n}{\partial y}(+1) < 0,$$

а значит, найдется такое число $y_0 \in (-1;1)$, что $\dfrac{\partial g_n}{\partial y}\left(y_0\right) = 0$. Обозначая $\tilde{\eta}_n = \bar{\eta}_n + \delta_n y_0$, выводим из (3.12):

$$\frac{\partial \tilde{L}_n}{\partial\eta}\left(\tilde{\eta}_n, \hat{\tau}_n, \hat{\mu}_n\right) = 0.$$

Кроме того, ясно, что $|\bar{\eta}_n - \tilde{\eta}_n| < \delta_n$. ■

Следствие. $\tilde{\eta}_n$ — состоятельная оценка для параметра η. Кроме того, $\tilde{\eta}_n$ — асимптотически эффективна, и случайная величина $\sqrt{n}\left(\tilde{\eta}_n - \eta_0\right)$ сходится к нормальному распределению с нулевым средним и дисперсией $\sigma^2\left(\eta_0,\tau_0\right)$.

Доказательство. Действительно, $\bar{\eta}_n - \eta_0 \xrightarrow[n\to\infty]{\mathbf{P}} 0$ и $\tilde{\eta}_n - \bar{\eta}_n \xrightarrow[n\to\infty]{\mathbf{P}} 0$, откуда следует состоятельность. Далее, заметим, что $\sqrt{n}\left(\tilde{\eta}_n - \eta_0\right) = \sqrt{n}\left(\bar{\eta}_n - \eta_0\right) + \sqrt{n}\left(\tilde{\eta}_n - \bar{\eta}_n\right)$, где последнее слагаемое сходится к нулю по вероятности. Следовательно, асимптотическое распределение случайной величины $\sqrt{n}\left(\tilde{\eta}_n - \eta_0\right)$ совпадает с асимптотическим распределением случайной величины $\sqrt{n}\left(\bar{\eta}_n - \eta_0\right)$, что и требовалось показать. ■

4. Проверка статистических гипотез

В данном разделе будут рассмотрены некоторые способы построения критериев: как относительно значений параметров (раздел 4.1.), так и некоторые варианты критериев согласия (раздел 4.2.).

4.1. Проверка статистических гипотез относительно значений параметров

В данном разделе будет рассмотрена задача проверки статистических гипотез относительно значений параметров μ и τ. Мы начнем со случая известного параметра η, а затем перейдем к более общему случаю оценивания этого параметра по выборке. Естественно, все результаты данного раздела пригодны и для построения доверительных областей для значений параметров.

4.1.1. Случай известного параметра η

Для начала рассмотрим гипотезу $H_0 : (\mu, \tau, \eta) = (\mu_0, \tau_0, \eta_0)$ против простой альтернативы $H_1 : (\mu, \tau, \eta) = (\mu_n, \tau_n, \eta_0)$. Критерий Неймана-Пирсона отвергает гипотезу H_0, если

$$n\left(L_n\left(\mu_n, \tau_n, \eta_0\right) - L_n\left(\mu_0, \tau_0, \eta_0\right)\right) > \alpha^* \tag{4.1}$$

для некоторого критического значения α^*. Естественно выбирать μ_n и τ_n в виде

$$\mu_n = \mu_0 + \frac{t_1}{n}, \quad \tau_n = \tau_0 + \frac{t_2}{n}$$

для фиксированых вещественных t_1, t_2, так как в этом случае критерий будет иметь мощность строго между 0 и 1.

4.1.2. Случай оценивания параметра η по выборке

Предположим теперь, что значение параметра η не известно, а оценивается по выборке посредством статистики $\tilde{\eta}_n$.

Естественным обобщением на этот случай критерия типа Неймана-Пирсона, представленного в уравнении (4.1), является критерий, отвергающий гипотезу H_0, если

$$n\left(L_n\left(\mu_n, \tau_n, \tilde{\eta}_n\right) - L_n\left(\mu_0, \tau_0, \tilde{\eta}_n\right)\right) > \alpha^{**}. \tag{4.2}$$

Для сравнения (4.2) и (4.1) заметим, что

$$\varepsilon_n = n\left[\left(L_n\left(\mu_n,\tau_n,\tilde{\eta}_n\right)-L_n\left(\mu_0,\tau_0,\tilde{\eta}_n\right)\right)-\left(L_n\left(\mu_n,\tau_n,\eta_0\right)-L_n\left(\mu_0,\tau_0,\eta_0\right)\right)\right]=$$
$$= n\left(\tilde{\eta}_n-\eta_0\right)\left[\left(\hat{\mu}_n-\mu_0\right)\frac{\partial^2\tilde{L}_n}{\partial\eta\partial\mu}\left(\eta^*,\tau^*,\mu^*\right)+\left(\hat{\tau}_n-\tau_0\right)\frac{\partial^2\tilde{L}_n}{\partial\eta\partial\tau}\left(\eta^*,\tau^*,\mu^*\right)\right] \quad (4.3)$$

для некоторых (η^*,τ^*,μ^*). В силу следствия из теоремы 1, имеем

$$\left(\tilde{\eta}_n-\eta_0\right)=O_p\left(\frac{1}{\sqrt{n}}\right), n\to\infty$$

Выражение в (4.3) в квадратных скобках есть $O_p\left(\frac{1}{n}\right)$ (рассуждения, аналогичные доказательству теоремы 1). Следовательно,

$$\varepsilon_n = O_p\left(\frac{1}{\sqrt{n}}\right), n\to\infty.$$

Таким образом, $\varepsilon_n \xrightarrow[n\to\infty]{\mathbf{P}} 0$ и критерии (4.2) и (4.1) асимптотически эквивалентны: их мощности и уровни значимости совпадают при $n\to\infty$.

Замечание. Величины α^* и α^{**} могут быть вычислены для конкретных значений параметров μ_0, τ_0, η_0, размера выборки n и уровня значимости α при помощи моделирования распределения $L_n\left(\mu_0,\tau_0,\tilde{\eta}_n\right)$.

4.2. Критерий согласия для специальной модели кривых дожитиия

4.2.1. Критерий ω^2 Крамера-Смирнова-фон Мизеса

Пусть $X_1,\ldots,X_n$ — независимые одинаково распределенные случайные величины с некоторой (неизвестной) функцией распределения $F(x)$. Рассмотрим задачу проверки гипотезы:

$$H_0: F(x)=G(x)$$

для некоторой (известной) функции распределения $G(x)$.

Обозначим через $F_n(x)$ эмпирическую функцию распределения:

$$F_n(x)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(x-X_i\right),$$

где $\mathbb{I}(x)$ - индикатор множества $(0,+\infty)$.

Статистика классического критерия ω^2 (Крамера-Смирнова-фон Мизеса) имеет вид:

$$\omega_n^2 = n\int_{-\infty}^{\infty}\left[F_n(x)-G(x)\right]^2 dG(x), \quad (4.4)$$

и гипотеза H_0 отвергается, если значение ω_n^2 достаточно велико.

Если $G(x)$ — непрерывная функция распределения, то преобразуем исходную выборку $X_1, \dots X_n$ к выборке $t_1, \dots, t_n$, где $t_i = G(X_i)$. Полученная выборка выборка в случае справедливости гипотезы H_0 будет иметь равномерное на отрезке $[0, 1]$ распределение, а статистика ω_n^2 примет вид:

$$\omega_n^2 = n \int_0^1 [F_n(t) - t)]^2 \, dt, \tag{4.5}$$

здесь $F_n(t)$ - эмпирическая функция распределения, построенная по $t_1, \dots, t_n$.

Критерий ω^2 обладает рядом достоинств, отсутствующих у другого классического критерия χ^2: он не требует субъективной группировки выборки, предельное распределение статистики ω_n^2 не зависит от функции $F(x)$, критерий является состоятельным (т.е. асимптотическая мощность критерия равна единице). Подробнее о критерии ω^2, смотри, например, [7, 8, 13, 16].

Рассмотрим более общую задачу, а именно задачу проверки сложной гипотезы вида:

$$H_0 : F(x) = G(x; \theta)$$

для некоторого (неизвестного) вещественного параметра $\theta \in [\theta_a, \theta_b]$.

Следуя [13], рассмотрим несколько измененную статистику (4.4):

$$\Omega_n^2 = n \int_{-\infty}^{\infty} \left[F_n(x) - G(x; \hat{\theta}_n) \right]^2 \, dG(x; \hat{\theta}_n), \tag{4.6}$$

где $\hat{\theta}_n$ — некоторая точечная оценка параметра θ.

Вообще говоря, распределение статистики Ω_n^2 зависит от множества факторов: вида функции распределения $F(x)$, истинного значения неизвестного параметра θ, самой оценки $\hat{\theta}_n$. Подробнее о вычислении распределения статистики Ω_n^2 смотри [12, 13, 16–18].

Замечание. Статистику Ω_n^2 можно записать в более удобном для вычислений виде [13]:

$$\Omega_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left[G(X_{[i;n]}; \hat{\theta}_n) - \frac{2i-1}{2n} \right]^2.$$

Как было показано в [13], распределение статистики Ω_n^2 существенно зависит от поведения оценки $\hat{\theta}_n$. В дальнейшем нам понадобится следующая теорема из [13] (смотри также [14], следствие 2 теоремы 1).

Теорема 2. Пусть оценка $\hat{\theta}_n$ и функция распределения $G(x;\,\theta)$ удовлетворяют следующим условиям:

1. $n\mathbf{E}(\hat{\theta}_n - \theta)^2 \xrightarrow[n\to\infty]{} 0$.
2. $G(x;\,\theta)$ удовлетворяет условию Липшица по параметру θ, то есть для $\theta_1, \theta_2 \in [\theta_a, \theta_b]$ выполняется:
$$|G(x;\,\theta_1) - G(x;\,\theta_2)| < C(x)\,|\theta_1 - \theta_2|\,, \tag{4.7}$$
для некоторой ограниченой в среднеквадратическом по распределению $G(x;\theta)$ функции $C(x)$.

Тогда $\Omega_n^2 = \omega_n^2 + \varepsilon_n$ и $\varepsilon_n \xrightarrow[n\to\infty]{\mathbf{P}} 0$.

Основной задачей данного раздела будет построение критерия типа ω^2 в случае, когда параметр сдвига μ оценивается по выборке. При этом будем предполагать, что параметры η и τ известны. Главным инструментом в построении критерия будет служить теорема 2.

4.2.2. Несмещенное оценивание параметра μ по выборке

Как и ранее, рассмотрим оценку $\hat{\mu}_n = X_{[1;n]} = \min_i X_i$. Для этой оценки несложно получить функцию распределения F_n^μ и соответствующую плотность f_n^μ:

$$F_n^\mu(x;\,\eta,\tau,\mu) = 1 - (1 - F(x))^n = 1 - \exp\left(-n\eta\left(\frac{x-\mu}{\tau}\right)\right)\cos^n\left(\frac{\pi}{2}\left(\frac{x-\mu}{\tau}\right)\right), \mu < x \leqslant \mu + \tau.$$

$$f_n^\mu(x;\,\eta,\tau,\mu) = \frac{n}{\tau}e^{-\eta y}\left[\eta\,\cos\frac{\pi y}{2} + \frac{\pi}{2}\,\sin\frac{\pi y}{2}\right]\cos^{n-1}\frac{\pi y}{2},\ \mu < x \leqslant \mu + \tau, \quad y = \frac{x-\mu}{\tau}.$$

Оценка $\hat{\mu}_n$ смещена:

$$\mathbf{E}\hat{\mu}_n = \int\limits_{\mu_0}^{\mu_0+\tau_0} x f_n^\mu(x;\,\eta_0,\tau_0,\mu_0)\,dx = \int\limits_{\mu_0}^{\mu_0+\tau_0} (x-\mu_0) f_n^\mu(x;\,\eta_0,\tau_0,\mu_0)\,dx +$$

$$+ \mu_0 \int\limits_{\mu_0}^{\mu_0+\tau_0} f_n^\mu(x;\,\eta_0,\tau_0,\mu_0)\,dx = \mu_0 + \int\limits_0^{\tau_0} x f_n^\mu(x;\,\eta_0,\tau_0,0)\,dx = \mu_0 + \varkappa_n^{(1)}(\eta_0,\tau_0),$$

здесь

$$\varkappa_n^{(1)}(\eta_0,\tau_0) = \int\limits_0^{\tau_0} x f_n^\mu(x;\,\eta_0,\tau_0,0)\,dx. \tag{4.8}$$

Константа $\varkappa_n^{(1)}(\eta_0,\tau_0)$ не зависит от значения неизвестного параметра μ и может быть вычислена отдельно (подробнее смотри раздел 4.2.4.).

После этого можно легко получить несмещенную оценку для μ:

$$\tilde{\mu} = \tilde{\mu}_n = \hat{\mu}_n - \varkappa_n^{(1)}(\eta_0,\tau_0). \tag{4.9}$$

4.2.3. Вычисление скорости сходимости оценки $\tilde{\mu}$

Теорема 2 фактически накладывает ограничение на скорость сходимости дисперсии оценки при $n \to \infty$. Ее можно получить в явном виде через плотность f_n^{μ}:

$$D_n = \mathbf{E}\left(\tilde{\mu} - \mu_0\right)^2 = \mathbf{E}\tilde{\mu}^2 - \mu_0^2 = \mathbf{E}\hat{\mu}^2 - 2\mu_0\varkappa_n^{(1)} - \left(\varkappa_n^{(1)}\right)^2 - \mu_0^2 =$$

$$= \int\limits_{\mu_0}^{\mu_0+\tau_0} x^2 f_n^{\mu}(x;\, \eta_0, \tau_0, \mu_0)\, dx - 2\mu_0\varkappa_n^{(1)} - \left(\varkappa_n^{(1)}\right)^2 - \mu_0^2 =$$

$$= \int\limits_{0}^{\tau_0} \left(x + \mu_0\right)^2 f_n^{\mu}(x;\, \eta_0, \tau_0, 0)\, dx - 2\mu_0\varkappa_n^{(1)} - \left(\varkappa_n^{(1)}\right)^2 - \mu_0^2 =$$

$$= \int\limits_{0}^{\tau_0} x^2 f_n^{\mu}(x;\, \eta_0, \tau_0, 0)\, dx + 2\mu_0 \int\limits_{0}^{\tau_0} x f_n^{\mu}(x;\, \eta_0, \tau_0, 0)\, dx - 2\mu_0\varkappa_n^{(1)} - \left(\varkappa_n^{(1)}\right)^2 =$$

$$= \varkappa_n^{(2)}(\eta_0, \tau_0) - \left(\varkappa_n^{(1)}(\eta_0, \tau_0)\right)^2,$$

здесь

$$\varkappa_n^{(2)}(\eta_0, \tau_0) = \int\limits_{0}^{\tau_0} x^2 f_n^{\mu}(x;\, \eta_0, \tau_0, 0)\, dx.$$

Для того, чтобы получить условие 1 теоремы 2, достаточно оценить величину D_n:

$$\varkappa_n^{(1)} = \frac{n}{\tau_0} \int\limits_{0}^{\tau_0} x \exp\left(-\frac{\eta_0}{\tau_0}x\right) \left(\eta_0 \cos\frac{\pi x}{2\tau_0} + \frac{\pi}{2}\sin\frac{\pi x}{2\tau_0}\right) \left(\exp\left(-\frac{\eta_0}{\tau_0}x\right)\cos\frac{\pi x}{2\tau_0}\right)^{n-1} dx =$$

$$= n\tau_0 \int\limits_{0}^{1} y e^{-n\eta_0 y} \left(\eta_0 \cos\frac{\pi}{2}y + \frac{\pi}{2}\sin\frac{\pi}{2}y\right)\cos^{n-1}\frac{\pi}{2}y\, dy \leqslant n\tau_0\left(\eta_0 + \frac{\pi}{2}\right)\int\limits_{0}^{1} y e^{-\eta_0 n y}\, dy \leqslant$$

$$\leqslant n\tau_0\left(\eta_0 + \frac{\pi}{2}\right)\int\limits_{0}^{\infty} y e^{-\eta_0 n y}\, dy = \frac{\tau_0\left(\eta_0 + \frac{\pi}{2}\right)}{\eta_0^2}\frac{1}{n} = O\left(\frac{1}{n}\right),\ n \to \infty. \quad (4.10)$$

Аналогично,

$$\varkappa_n^{(2)} = n\tau_0^2 \int\limits_{0}^{1} y^2 e^{-n\eta_0 y} \left(\eta_0 \cos\frac{\pi}{2}y + \frac{\pi}{2}\sin\frac{\pi}{2}y\right)\cos^{n-1}\frac{\pi}{2}y\, dy \leqslant n\tau_0^2\left(\eta_0 + \frac{\pi}{2}\right)\int\limits_{0}^{1} y^2 e^{-n\eta_0 y}\, dy \leqslant$$

$$\leqslant n\tau_0^2\left(\eta_0 + \frac{\pi}{2}\right)\int\limits_{0}^{\infty} y^2 e^{-n\eta_0 y}\, dy = \frac{2\tau_0^2\left(\eta_0 + \frac{\pi}{2}\right)}{\eta_0^3}\frac{1}{n^2} = O\left(\frac{1}{n^2}\right),\ n \to \infty. \quad (4.11)$$

Окончательно, имеем:

$$nD_n = n\left[\varkappa_n^{(2)} - \left(\varkappa_n^{(1)}\right)^2\right] \leqslant n\left[\varkappa_n^{(2)} + \left(\varkappa_n^{(1)}\right)^2\right] = O\left(\frac{1}{n}\right) \xrightarrow[n\to\infty]{} 0. \quad (4.12)$$

Тем самым первое условие теоремы 2 выполнено. Второе условие очевидно выполнено в силу ограниченности носителя плотности f_n^μ.

Таким образом, условия теоремы 2 полностью удовлетворены, что позволяет сформулировать следующую теорему.

Теорема 3. Пусть $X_1, \ldots, X_n$ — независимые одинаково распределенные случайные величины с функцией распределения задаваемой формулой (2.1). Кроме того, пусть параметры η и τ известны, а параметр μ оценивается по формуле (4.9). Тогда распределение статистики критерия Ω_n^2 вида (4.6) сходится при $n \to \infty$ по вероятности к (асимптотическому) распределению статистики критерия ω^2 Крамера-Смирнова-фон Мизеса.

Замечание. Повторяя доказательство теоремы 2, можно показать, что асимптотическое распределение статистики Ω_n^2 не изменится, если вместо оценки $\tilde{\mu}_n$ использовать оценку $\hat{\mu}_n$.

Замечание. На самом деле условия теоремы 2 можно ослабить: так, в частности, теорема 3 остается верной, если оценивать одновременно параметры μ и и τ по формуле (2.4), а параметр η считать известным [1].

4.2.4. Вычисление константы $\varkappa_n^{(1)}(\eta_0, \tau_0)$

Как было отмечено выше, для построения критерия согласия можно использовать смещенную оценку $\hat{\mu}_n$. Однако, при небольших объемах выборки влияние смещения на мощность критерия может оказаться значительным, поэтому в таком случае обосновано применение несмещенной оценки $\tilde{\mu}_n$. Для этого необходимо вычислить константу $\varkappa_n^{(1)}$.

Из формулы (4.8) имеем:

$$\varkappa_n^{(1)} = \int\limits_0^{\tau_0} x f_n^\mu(x;\, \eta_0, \tau_0, 0)\, dx = \frac{4\tau_0 n}{\pi^2} \int\limits_0^{\frac{\pi}{2}} y e^{-\alpha y} \left(\eta_0 \cos y + \frac{\pi}{2} \sin y\right) \left(e^{-\alpha y} \cos y\right)^{n-1}\, dy = \\ = \frac{4\tau_0 n}{\pi^2} \left(\eta_0 I_n^{(1)} + \frac{\pi}{2} I_n^{(2)}\right), \quad (4.13)$$

здесь

$$\alpha = \frac{2\eta_0}{\pi}, \quad I_n^{(1)} = \int\limits_0^{\frac{\pi}{2}} y e^{-n\alpha y} \cos^n y\, dy, \quad I_2^{(2)} = \int\limits_0^{\frac{\pi}{2}} y \sin y\, e^{-n\alpha y} \cos^{n-1} y\, dy.$$

[1] Однако, теорема 3 неверна, даже если считать известными параметры τ и μ, а в качестве оценки параметра η использовать $\bar{\eta}_n$ из формулы (2.3).

Интегрируя по частям, получаем:

$$I_n^{(2)} = -ye^{-n\alpha y}\cos^{n-1} y\cos y\Big|_0^{\frac{\pi}{2}} + \int\limits_0^{\frac{\pi}{2}} e^{-n\alpha y}\cos^n y\,dy - n\alpha\int\limits_0^{\frac{\pi}{2}} ye^{-n\alpha y}\cos^n y\,dy -$$

$$-(n-1)\int\limits_0^{\frac{\pi}{2}} y\sin y\,e^{-n\alpha y}\cos^{n-1} y\,dy = \int\limits_0^{\frac{\pi}{2}} e^{-n\alpha y}\cos^n y\,dy - n\alpha I_n^{(1)} - (n-1)I_n^{(2)}.$$

Отсюда:

$$I_n^{(2)} = \frac{1}{n}I_n^{(3)} - \alpha I_n^{(1)}, \tag{4.14}$$

где

$$I_n^{(3)} = \int\limits_0^{\frac{\pi}{2}} e^{-n\alpha y}\cos^n y\,dy.$$

Подставляя выражение для $I_n^{(2)}$ из формулы (4.14) в формулу (4.13) имеем:

$$\varkappa_n^{(1)} = \frac{4\tau_0 n}{\pi^2}\left(\eta_0 I_n^{(1)} + \frac{\pi}{2}I_n^{(2)}\right) = \frac{4\tau_0 n}{\pi^2}\left(\eta_0 I_n^{(1)} + \frac{\pi}{2n}I_n^{(3)} - \frac{\pi\alpha}{2}I_n^{(1)}\right) = \frac{2\tau_0}{\pi}I_n^{(3)}.$$

Теперь можно приступить к вычислению интеграла $I_n^{(3)}$:

$$I_n^{(3)} = \int\limits_0^{\frac{\pi}{2}} e^{-n\alpha y}\cos^n y\,dy = \frac{1}{2^{n-1}}\sum_{k=0}^{\lfloor\frac{n}{2}\rfloor} C_n^k \int\limits_0^{\frac{\pi}{2}} e^{-n\alpha y}\cos(n-2k)\,y\,dy = \frac{1}{2^{n-1}}\sum_{k=0}^{\lfloor\frac{n}{2}\rfloor} C_n^k I_{n,k}^{(3)}.$$

Дважды интегрируя по частям, получаем:

$$I_{n,k}^{(3)} = \frac{1}{n^2\alpha^2 + (n-2k)^2}\left[n\alpha + e^{-n\frac{\pi\alpha}{2}}\left((n-2k)\sin\left(\frac{\pi(n-2k)}{2}\right) - n\alpha\cos\left(\frac{\pi(n-2k)}{2}\right)\right)\right] =$$

$$= \frac{\pi^2}{4n^2\eta_0^2 + \pi^2(n-2k)^2}\left[\frac{2\eta_0 n}{\pi} + (-1)^k e^{-n\eta_0}\left((n-2k)\sin\frac{\pi n}{2} - \frac{2\eta_0 n}{\pi}\cos\frac{\pi n}{2}\right)\right]. \tag{4.15}$$

Окончательно имеем:

$$\varkappa_n^{(1)} = \frac{\tau_0}{\pi 2^{n-2}}\sum_{k=0}^{\lfloor\frac{n}{2}\rfloor} C_n^k I_{n,k}^{(3)}, \tag{4.16}$$

где $I_{n,k}^{(3)}$ вычисляются по формуле (4.15).

5. Экспериментальное изучение свойств оценок параметров

Свойства полученных оценок проверялись экспериментально: производилось моделирование распределения (2.1) и на полученных выборках изучались свойства оценок. Ввиду того, что все результаты раздела 3. являются асимптотическими, существенный интерес представлял вопрос, насколько быстро достигаются основные свойства оценок (несмещенность, нормальность $\tilde{\eta}_n$ и т.п.) в зависимости от размера выборки и истинных значений параметров. Моделированию посвящен раздел 5.2.

Кроме того, оценки использовались для исследования реальных данных (смотри раздел 5.3.).

5.1. Цензурирование

Как было сказано ранее, при изучении экспериментальных данных приходится сталкиваться с различного рода «артефактами», связанными с объективными свойствами процесса сбора данных и проведения эксперимента. Чаще всего при наблюдении реальных данных происходит *цензурирование*. В связи с этим в численных экспериментах оценки параметров были модифицированы для учета этого эффекта, и все результаты представлены в двух видах: с цензурированием и без.

Опишем подробнее процедуру цензурирования. Пусть, как и прежде, $X_1, \ldots, X_n$ — набор из n независимых одинаково распределенных случайных величин с некоторой функцией распределения $F(x)$. Пусть $\tau_1, \ldots, \tau_n$ — независимые одинаково распределенные случайные величины, имеющие распределение Бернулли с некоторым неизвестным параметром p. Величина τ_j, называемая *индикатором цензурирования*, вообще говоря, может зависеть от X_j, $1 \leqslant j \leqslant n$. Кроме того, пусть $c_1, \ldots, c_n$ — произвольные случайные величины, такие, что $c_j < X_j$ и c_j может зависеть от τ_j.

Наблюдаемой величиной является пара (Y_j, τ_j), где

$$Y_j = \tau_j c_j + (1 - \tau_j) X_j.$$

Как правило, имеет место так называемое «цензурирование справа»:

$$\tau_j = \mathbb{I}(X_j > c), \quad Y_j = \min\{c, X_j\}, \quad c_j = c,$$

где $\mathbb{I}$ — индикатор множества, а константа c известна заранее.

Ясно, что если $X_1, \ldots, X_n$ — случайные величины с функцией распределения (2.1), то процедура оценивания, представленная в разделе 3., уже не может быть применена к цензурированой выборке: статистика $\hat{\tau}_n$, вообще говоря (например, как раз в случае «цензурирования справа»), не является оценкой неизвестного параметра τ.

Как и раньше, воспользуемся методом максимального правдоподобия. Естественно считать, что $\mathbf{P}\left(Y_{[1;n]} = X_{[1;n]}\right) \xrightarrow[n\to\infty]{\mathbf{P}} 1$ (то есть, при достаточно больших n можно ожидать, что наименьшее наблюдение цензурировано не будет). Тогда, с вероятностью, стремящейся к 1 оценка $\hat{\mu}_n = Y_{[1;n]}$ по-прежнему будет являться эффективной оценкой параметра μ. Параметры η и τ будем оценивать точкой локального максимума логарифма функции правдоподобия для цензурированой выборки $\tilde{L}_n^c$:

$$\tilde{L}_n^c\left(\eta, \tau, \mu\right) = \frac{1}{n}\left[\sum_{i:\,\tau_i=0} \log f\left(Y_i; \eta, \tau, \mu\right) + \sum_{i:\,\tau_i=1} \log\left(1 - F\left(Y_i; \eta, \tau, \mu\right)\right)\right], \tag{5.1}$$

$$\left(\tilde{\eta}_n^c,\, \tilde{\tau}_n^c\right) = \underset{\eta>0,\,\tau>0}{\arg\max}\, \tilde{L}_n^c\left(\eta, \tau, \hat{\mu}_n\right). \tag{5.2}$$

Первая сумма в (5.1) соответствует нецензурированым наблюдениям, а вторая — цензурированым. Теоретические свойства этих оценок неизвестны, и поэтому все выводы будут сделаны на основании результатов моделирования.

5.2. Моделирование

Моделирование распределения с функцией распределения вида (2.1) можно проводить двумя способами:

- Методом отбора
- Методом обратных функций

Для моделирования распределения (2.1) отбор логично производить из сдвиговой модификации экспоненциального распределения с параметром сдвига μ и параметром масштаба $\frac{\eta}{\tau}$. В этом случае производная Радона-Никодима $r(x)$ имеет вид:

$$r(x) = \cos\left(\frac{\pi}{2}\left(\frac{x-\mu}{\tau}\right)\right) + \frac{\pi}{2\eta}\sin\left(\frac{\pi}{2}\left(\frac{x-\mu}{\tau}\right)\right),\ \mu < x < \mu + \tau.$$

Несложно получить константу метода отбора M:

$$r(x) \leqslant M = \sqrt{1 + \frac{\pi^2}{4\eta^2}} = \frac{\pi}{2\eta} + O(\eta),\ \eta \to 0.$$

Таким образом, видно, что метод отбора для моделирования распределения (2.1) разумно применять при не очень малых значениях параметра η.

Замкнутого аналитического выражения для функции $F^{-1}(x)$ не существует, конкретное значение может быть получено лишь численно, поэтому трудоемкость метода обратных функций для моделирования (2.1) сравнительна велика. Однако, очевидно, что она не зависит от значений параметров, и, поэтому, использование метода обратных функций обосновано для малых значений параметра η, когда трудоемкость метода отбора сильно возрастает.

5.2.1. Оценки по выборке без цензурирования

Моделировалась выборка из распределения (2.1) с некоторыми типичными параметрами (всюду далее будут показаны результаты моделирования для $\eta = 1$, $\tau = 20$, $\mu = 7$). Объём выборки n варьировался от 10 до 5000. Выборка моделировалась 100 раз, соответственно, для каждого значения n получалось выборка из $m = 100$ оценок.

Для исследования скорости сходимости строились графики зависимости стандартного отклонения оценок $\tilde{\eta}_n$, $\hat{\tau}_n$, $\hat{\mu}_n$ (точнее, для наглядности, обратного к нему для $\hat{\tau}_n$ и $\hat{\mu}_n$ и обратного к квадрату для $\tilde{\eta}_n$) от объёма выборки (рисунки 1, 2, 3, соответственно).

Из графиков видно, что теоретическая скорость сходимости оценок действительно достигается (то есть имеет порядок $\frac{1}{n}$ для $\hat{\mu}_n$ и $\hat{\tau}_n$, и $\frac{1}{\sqrt{n}}$ для $\tilde{\eta}_n$).

Перейдем к асимптотическому распределению оценок. Подробнее остановимся на распределениях $\hat{\mu}_n$ и $\tilde{\eta}_n$, так они удобны для проверки (экспоненциальное для $\hat{\mu}_n$ и нормальное для $\tilde{\eta}_n$). Для проверки на экспоненциальность кроме традиционного χ^2-критерия также использовались критерий типа Колмогорова-Смирнова [15] и ω^2 [20]. Для проверки нормальности распределения кроме χ^2-критерия использовался вариант критерия ω^2 из [7].

На модельных выборках гипотеза об экспоненциальности распределения $\hat{\mu}_n$ не отвергалась на стандартных уровнях значимости, начиная с размера выборки $n = 60$, а гипотеза о нормальности $\tilde{\eta}_n$ — с размера выборки $n = 250$. На рисунке 4 приведена эмпирическая и теоретическая функции распределения для величины $\sqrt{n}\,(\tilde{\eta}_n - \eta_0)$, а на рисунке 5 — для $n\,(\hat{\mu}_n - \mu_0)$. На графиках объём выборки n из (2.1) равнялся 500, а объём выборки непосредственно наблюдаемых величин $m = 100$.

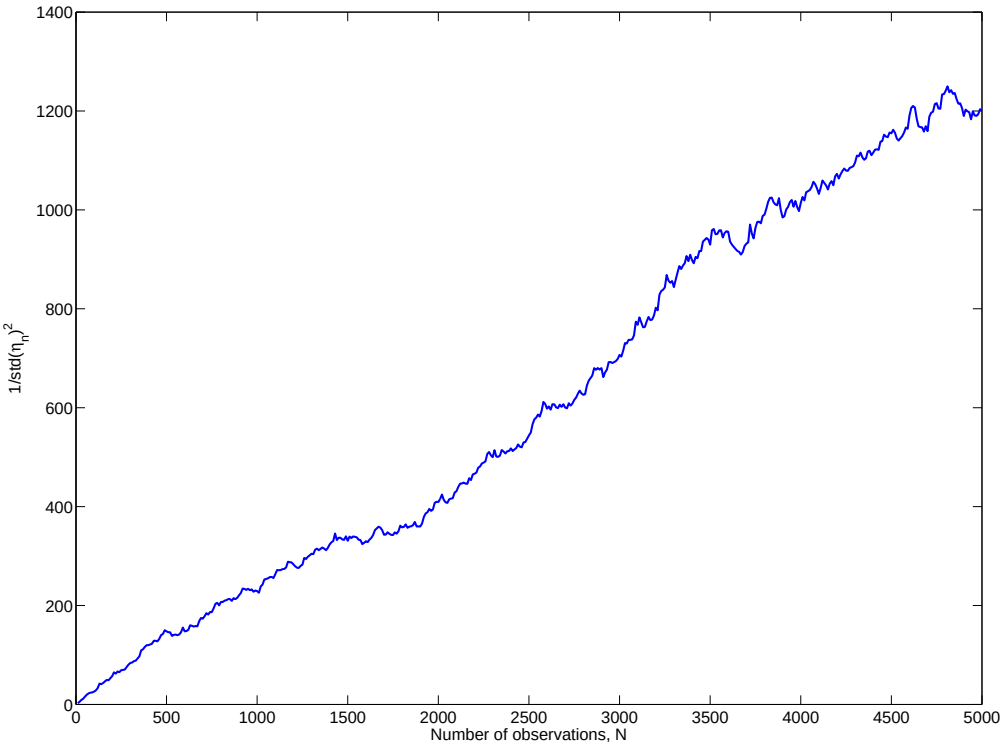


Рис. 1. Зависимость $\dfrac{1}{(\operatorname{std}\tilde{\eta}_n)^2}$ от объёма выборки n.

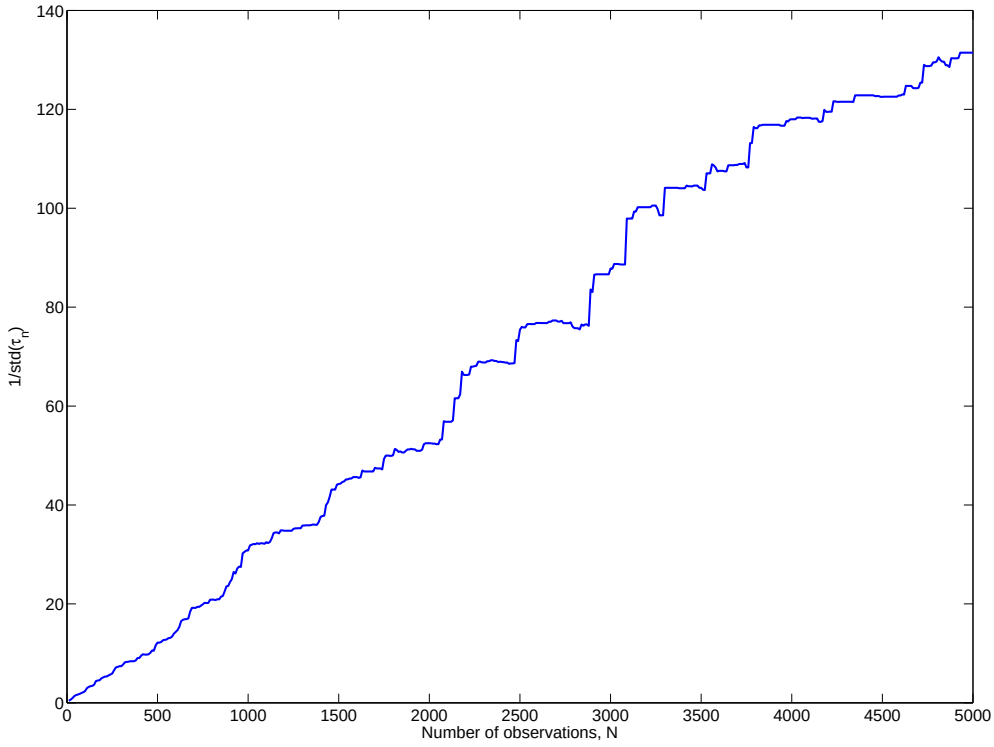


Рис. 2. Зависимость $\dfrac{1}{\operatorname{std}\hat{\tau}_n}$ от объёма выборки n.

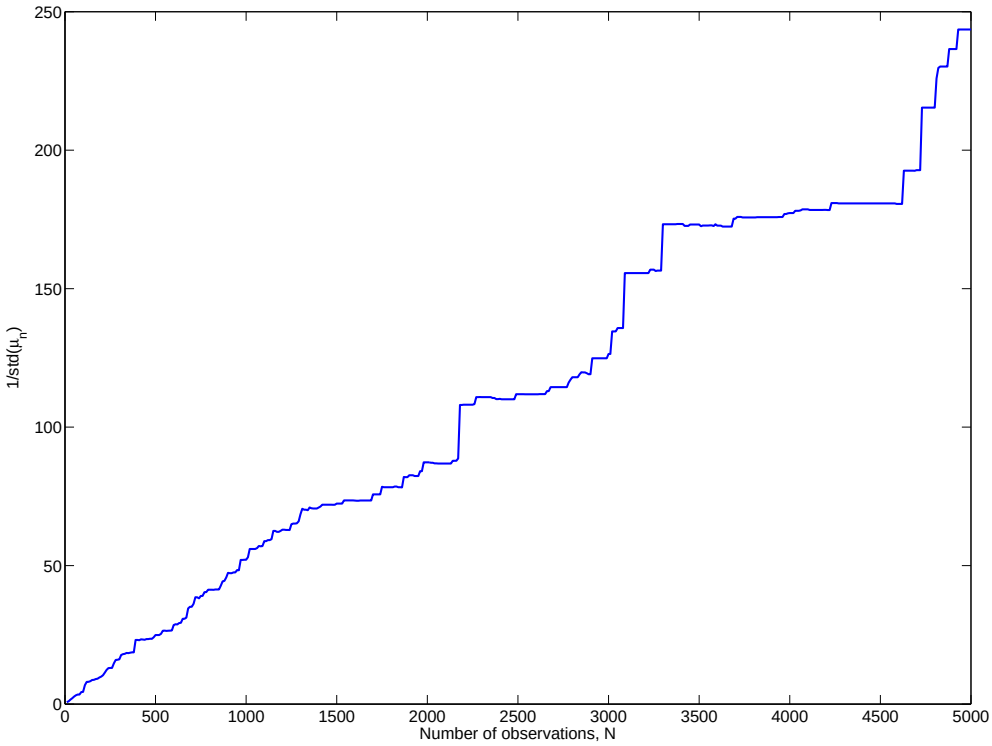


Рис. 3. Зависимость $\dfrac{1}{\operatorname{std}\hat{\mu}_n}$ от объёма выборки n.

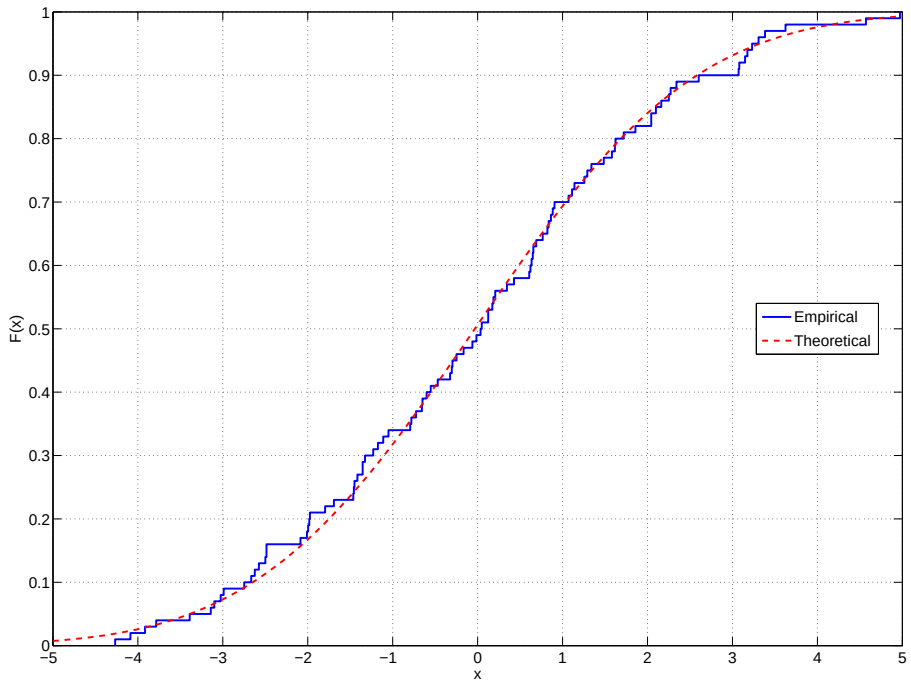


Рис. 4. Эмпирическая и теоретическая функции распределения для величины $\sqrt{n}\,(\tilde{\eta}_n - \eta_0)$, n=500, m=100.

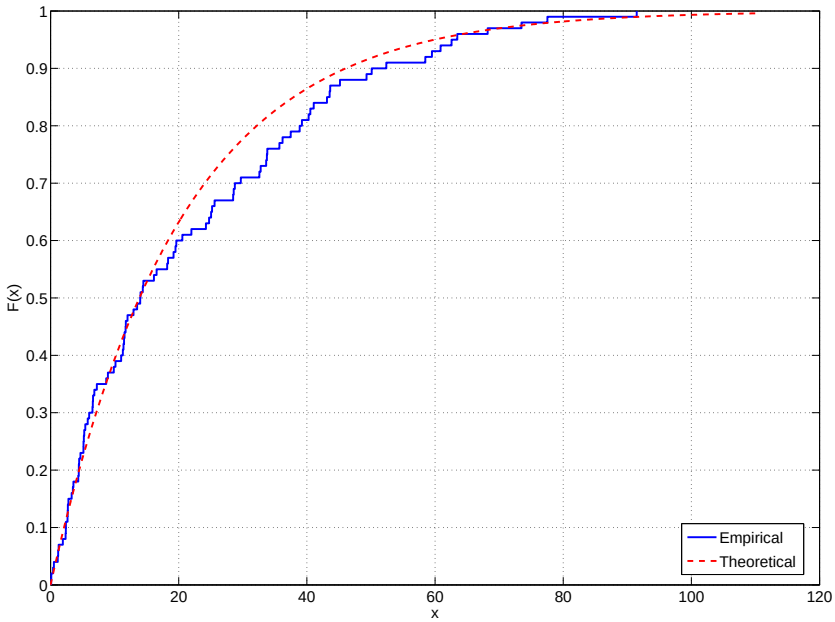


Рис. 5. Эмпирическая и теоретическая функции распределения для величины $n\,(\hat{\mu}_n - \mu_0)$, n=500, m=100.

5.2.2. Оценки по выборке с цензурированием

Все параметры моделирования были аналогичны случаю без цензурирования. В экспериментах производилось «цензурирование справа». В качестве константы цензурирования c выбиралась величина $\dfrac{\tau}{2}$.

Естественным было ожидать сохранение скорости сходимости для оценки $\hat{\mu}_n$ (на рисунке 6 приведен график зависимости величины, обратной к стандартному отклонению оценки $\hat{\mu}_n$, от объема выборки n).

Наибольший же интерес представляют скорости сходимости и асимптотические распределения оценок $\hat{\tau}_n$ и $\tilde{\eta}_n$. Видно, что скорость сходимости оценки $\tilde{\eta}_n$ осталось прежней (смотри рисунок 7), а скорость сходимости оценки $\hat{\tau}_n$ уменьшилась (отчетлива видна «нелинейность» зависимости на рисунке 8 и более похожий на линейную зависимость график на рисунке 9).

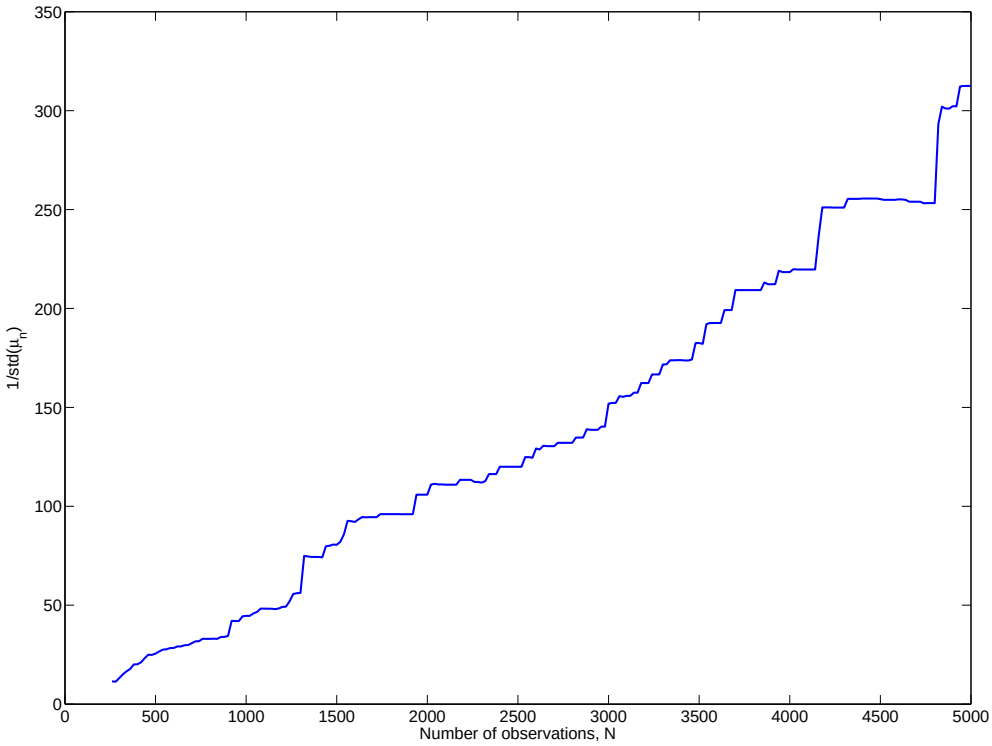


Рис. 6. Зависимость $\dfrac{1}{\operatorname{std}\hat{\mu}_n}$ от объёма выборки n (цензурированая выборка).

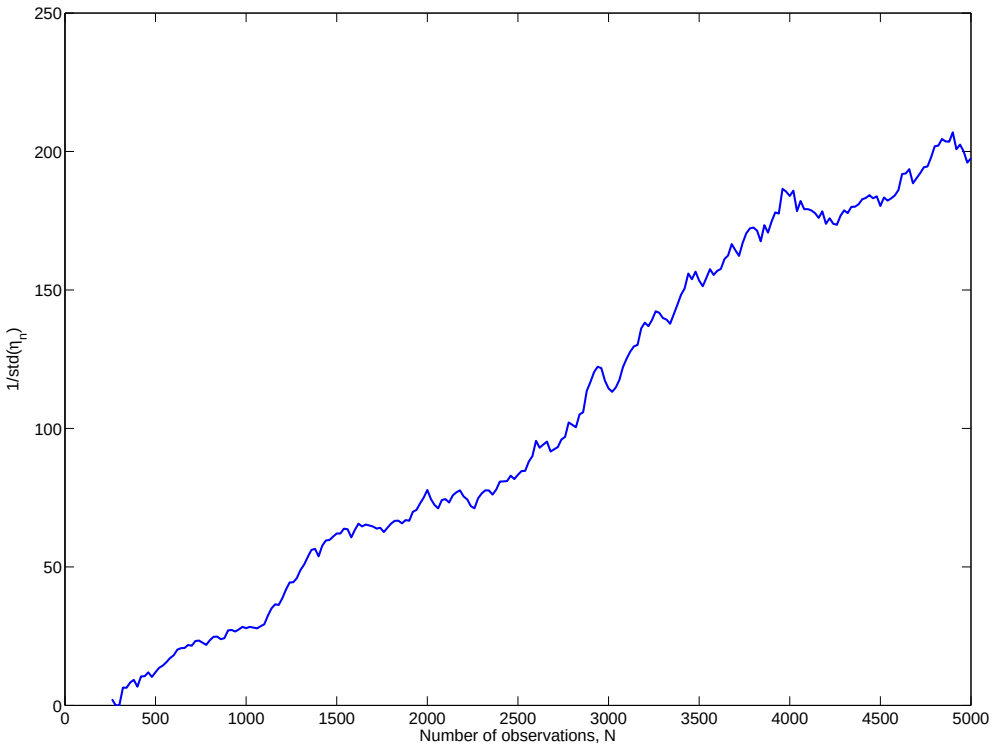


Рис. 7. Зависимость $\dfrac{1}{(\operatorname{std}\tilde{\eta}_n)^2}$ от объёма выборки n (цензурированая выборка).

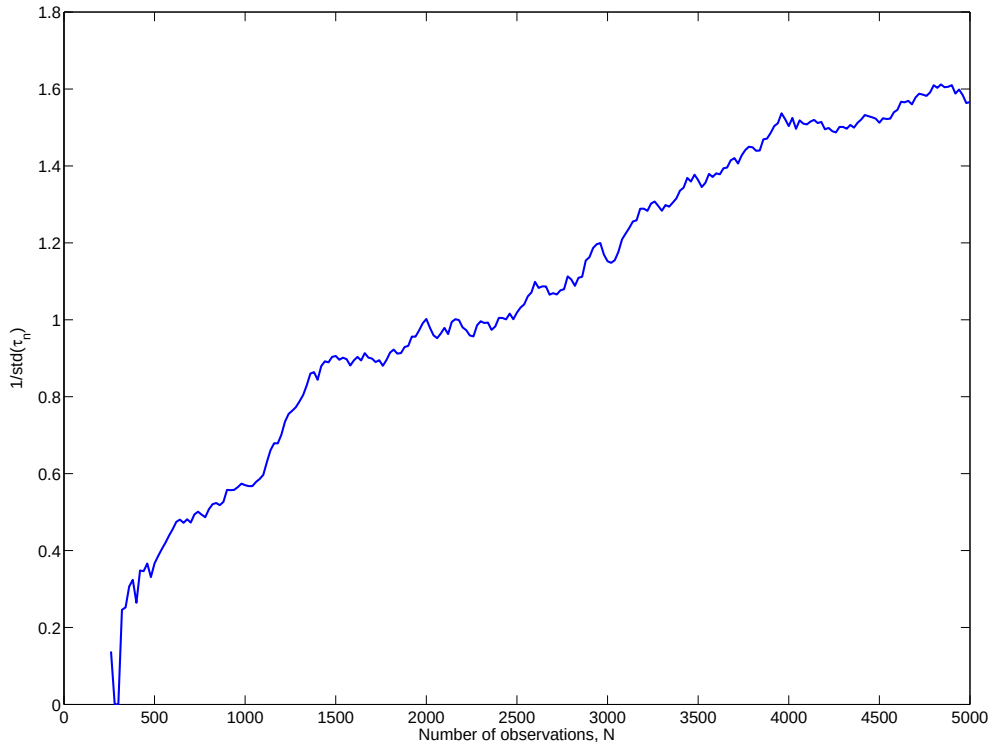


Рис. 8. Зависимость $\dfrac{1}{\operatorname{std}\hat{\tau}_n}$ от объёма выборки n (цензурированая выборка).

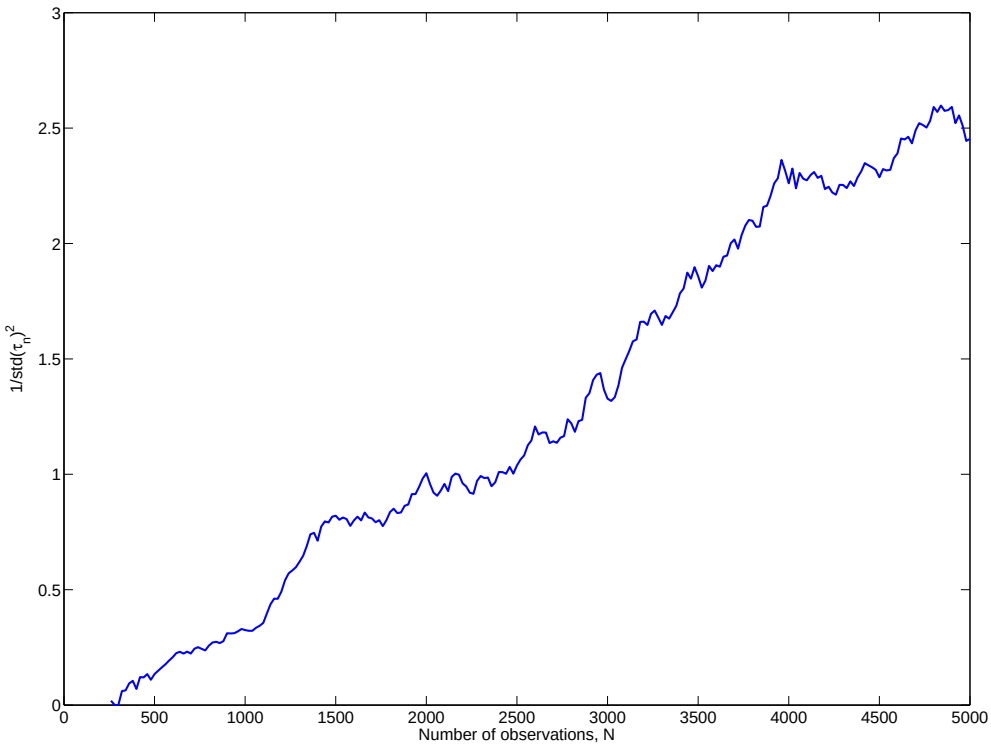


Рис. 9. Зависимость $\dfrac{1}{(\operatorname{std}\hat{\tau}_n)^2}$ от объёма выборки n (цензурированая выборка).

Гипотеза об экспоненциальности распределения $\hat{\mu}_n$ не отвергается на стандартных уровнях значимости начиная с объёма выборки $n = 350$. График эмпирической и теоретической функций распределения для $n = 500$ приведен на рисунке 10.

Кроме того, следует ожидать зависимость (возможно, даже асимптотическую) оценок $\tilde{\eta}_n$ и $\hat{\tau}_n$ при оценивании по цензурированой выборке: хотя гипотеза о нормальности распределений $\sqrt{n}\,(\tilde{\eta}_n - \eta_0)$ и $\sqrt{n}\,(\hat{\tau}_n - \tau_0)$ и не отвергается на стандартных уровнях значимости при $n = 5000$, наблюдается некоторая асимметрия распределений (на рисунках 11 и 12 приведены эмпирическая функция распределения и normal probability plot для величины $\sqrt{n}\,(\tilde{\eta}_n - \eta_0)$, соответственно). Это, вообще говоря, может привести к систематическим ошибкам в оценках параметров.

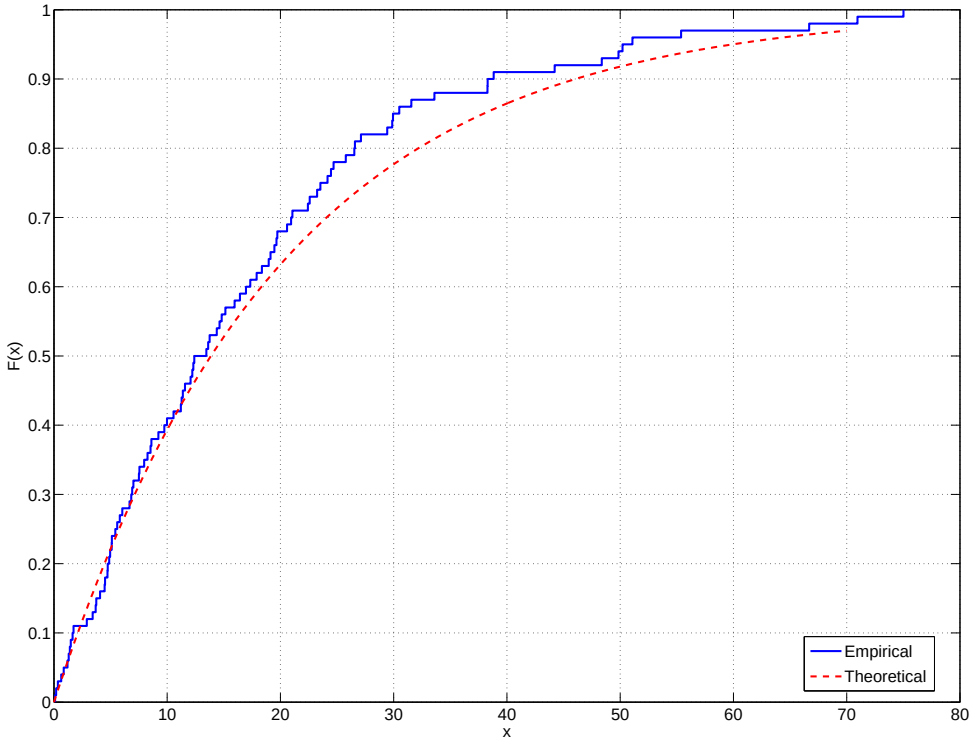


Рис. 10. Эмпирическая и теоретическая функции распределения для величины $n\,(\hat{\mu}_n - \mu_0)$, n=500, m=100 (цензурированая выборка).

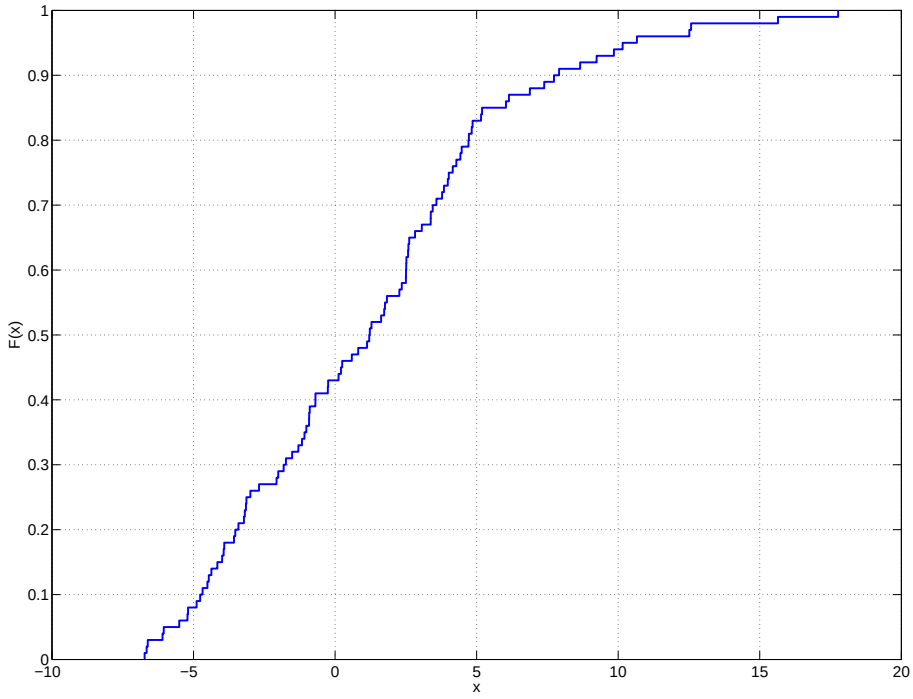


Рис. 11. Эмпирическая функция распределения для величины $\sqrt{n}\,(\tilde{\eta}_n - \eta_0)$, n=5000, m=100 (цензурированая выборка).

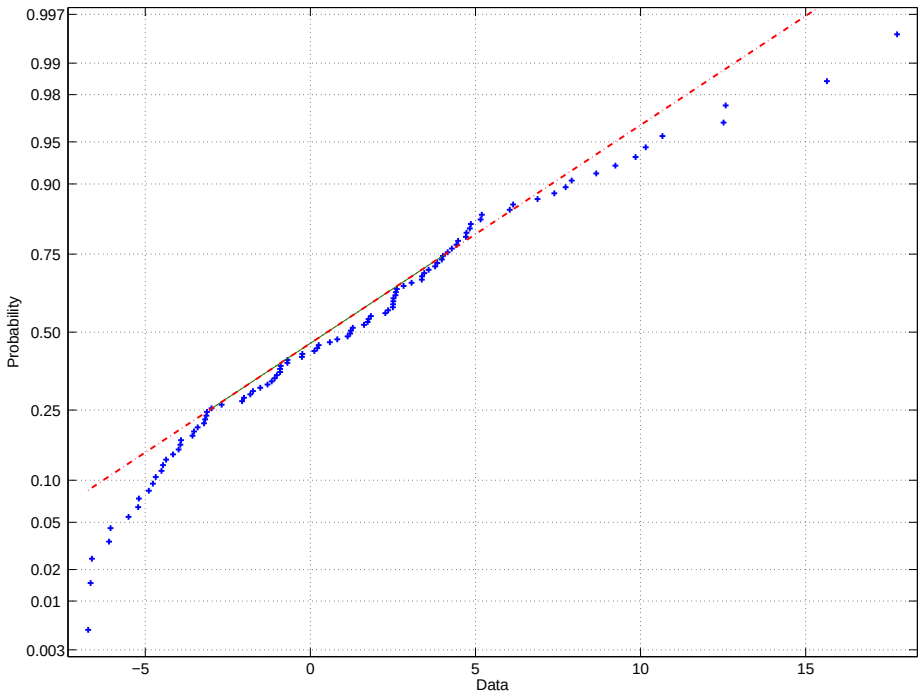


Рис. 12. Normal probability plot для величины $\sqrt{n}\,(\tilde{\eta}_n - \eta_0)$, n=5000, m=100 (цензурированая выборка).

5.3. Изучение реальных данных

В данном разделе будет рассмотрено несколько примеров обработки реальных данных из медицины. Параметры η и τ оказались удобными для интерпретации экспериментаторами: η информативен с точки зрения интенсивности наблюдаемого процесса, а τ — его максимальной продолжительности.

5.3.1. Пример из стоматологии

Сравнивались два способа лечения хронического генерализованного пародонтита у больных пожилого и старческого возраста: общепринятое лечение и лечение с применением пептидного биорегулятора «Вилон», способствующего оптимизации процессов регенерации тканей пародонта. Клинический этап исследования осуществлялся в Тарховском военном санатории МО РФ в 2002-2005 гг. [6]

Сравнение качественной динамики лечения хронического генерализованного пародонтита определялось по изменению во времени основого признака тяжести течения заболевания — средней глубины пародонтальных карманов. Для построения кривых была выбрана зависимость от времени доли больных, глубина карманов у которых не превышала 2.5 мм.

Сравнение кривых дожития производилось двумя способами: классическими непараметрическими критериями типа логарифмического рангового критерия [5] и сравнением оценок параметров модели кривых дожития. Результаты сравнения, полученные на основе классических критериев, и оценки параметров, полученные минимизацией квадрата расстояния между теоретической и эмпирической функциями распределения, были представлены в [6], поэтому подробнее остановимся на оценках параметров, полученных по формулам (2.4) и (2.5).

Объем выборки в основной группе (получавшие лечением «Вилоном») составил 38 индивидов, а контрольной (получавшими традиционное лечение) — 56 индивидов. Имело место «цензурирование справа»: наблюдения прекращались на 10 сутки. Группировка данных (наличие повторяющихся наблюдений) игнорировалась.

На рисунках 13(a) и 13(b) представлены графики эмпирической функции распределения, оцененой теоретической функции распределения и доверительные границы для теоретической функции распределения, полученные из формулы Гринвуда [5]. Оценки параметров приведены в таблице 1.

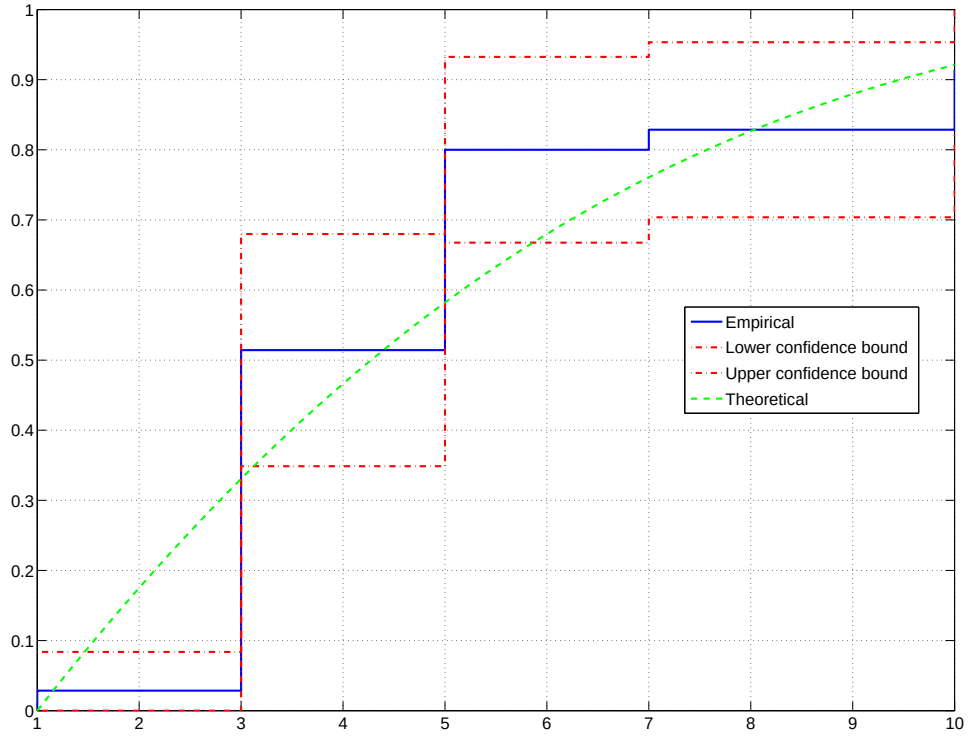
Группа	$\tilde{\eta}_n$	$\hat{\tau}_n$	$\hat{\mu}_n$
Основная	2.29	12.4	0
Контрольная	0.001	15.8	0

Таблица 1. Оценки параметров модели

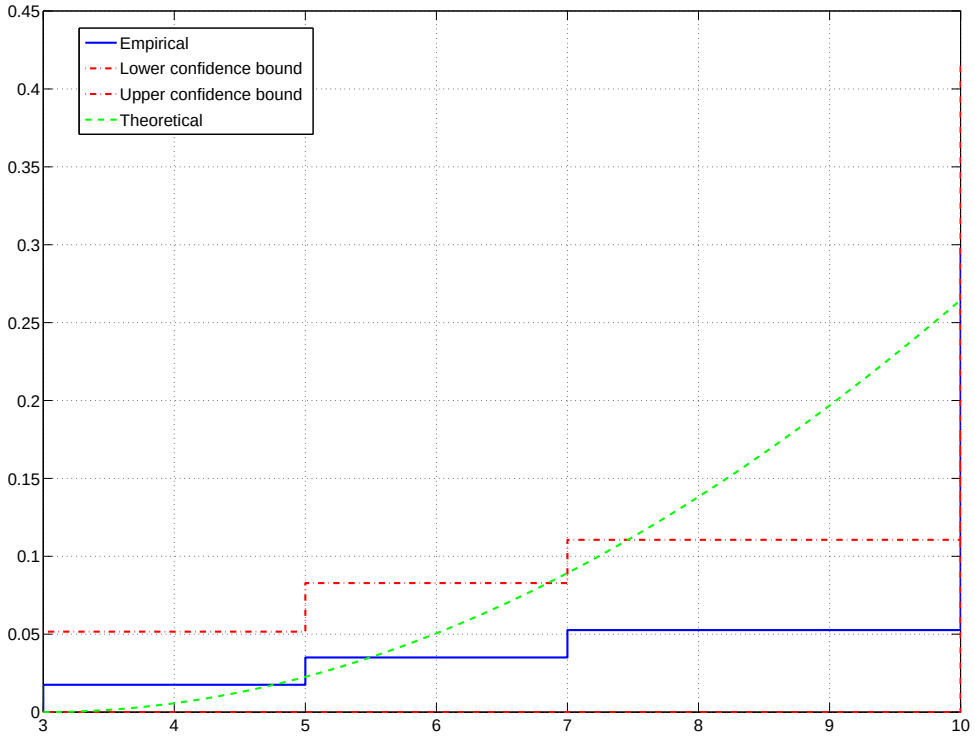
Сразу же отметим возможную недостоверность оценок для контрольной группы. Вероятнее всего, это связано со достаточно сильным цензурированием: эксперимент закончился слишком рано и существенная часть информации о выборке была потеряна.

Тем не менее, очевидно, что с точки зрения протекания процессов саногенеза-патогенеза группы существенно отличаются. Наблюдаются различия как в интенсивности процесса восстановления тканей (параметр η) так и в сроках окончания процесса (параметр τ).

Более того, следуя общему подходу к модели с точки зрения системы «орган-организм» [1,3], можно сделать вывод, что при лечении с применением «Вилона» преобладают изменения показателей состояния тканей пародонта в сторону выздоровления на уровня организма, чего нельзя сказать в случае применения только общепринятого метода лечения у больных старческого и пожилого возраста.



(a) Основная группа (с «Вилоном»).



(b) Контрольная группа (традиционное лечение).

Рис. 13. Пример из стоматологии. Эмпирическая, теоретическая функции распределения, доверительные границы для теоретической функции распределения. По оси x — сутки.

5.3.2. Пример из кардиологии

Рассматривалась группа больных гипертонической болезнью. Исследовалась зависимость частоты наступления комбинированной конечной точки (инфаркт миокарда, инсульт, смерть от сердечно-сосудистых причин) от типа гипертрофии миокарда левого желудочка.

В исследование вошло 734 больных гипертонической болезнью (преимущественно пожилого возраста) с гипертонией I-II стадии и 1-3 степени. Наблюдения проводились в течение 3-8 лет. Имело место «цензурирование справа» на 8 год наблюдений.

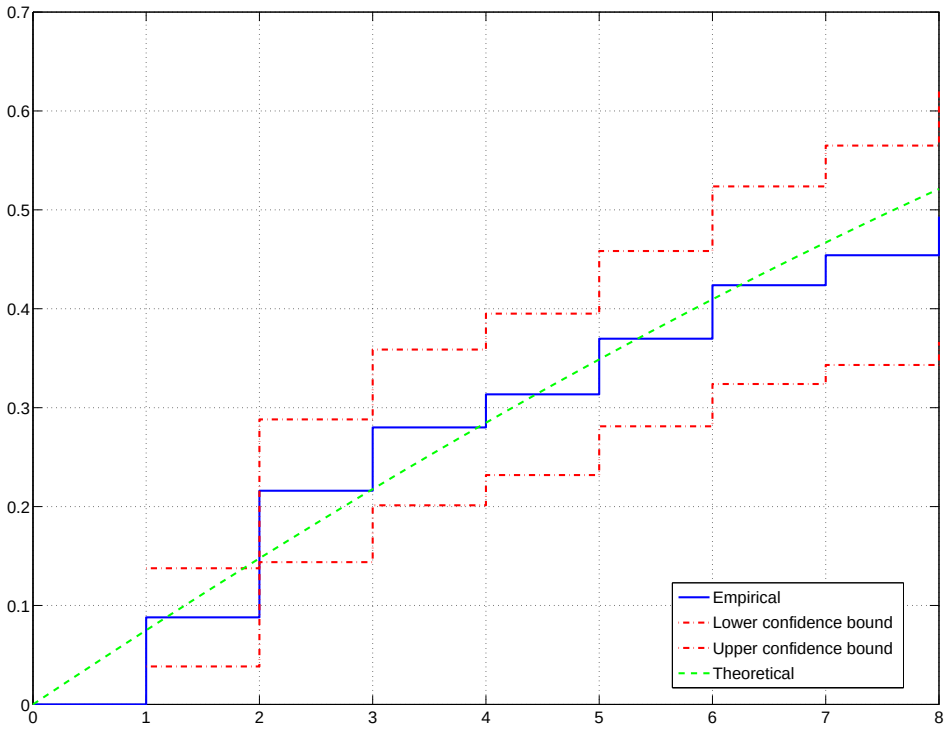
Выборка была разбита на две группы: больные с концентрической гипертрофией (тяжелое течение заболевания) и с другими типами гипертрофии (обычное течение заболевания). Объем выборки в группе с тяжелым течением заболевания составил 232 индивида, а в группе с обычным течением заболевания — 502 индивида.

На рисунках 14(a) и 14(b) представлены графики эмпирической функции распределения, оцененой теоретической функции распределения и доверительные границы для теоретической функции распределения. Оценки параметров приведены в таблице 2.

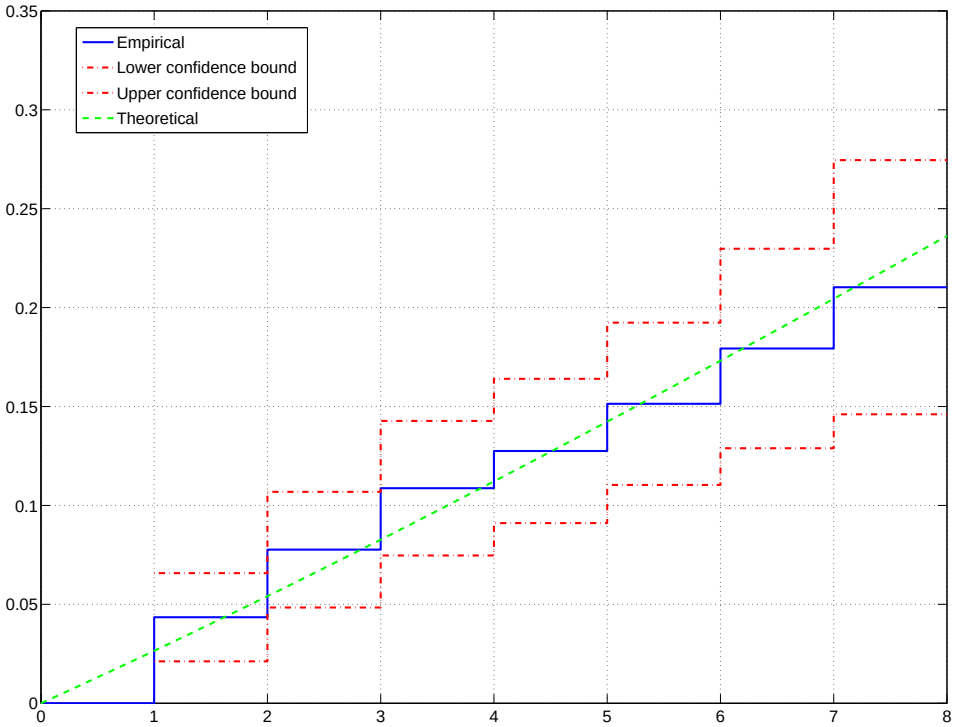
Группа	$\tilde{\eta}_n$	$\hat{\tau}_n$	$\hat{\mu}_n$
Тяжелое течение	1.9	25.4	0
Обычное течение	0.9	36.0	0

Таблица 2. Оценки параметров модели

Отличия в значениях параметра η вполне ожидаемы: параметр интерпретируется как «интенсивность» процесса. Более интересны различия в значениях параметра τ: разница в 10-11 лет хорошо согласуется с известными фактами.



(a) Группа с тяжелым течением заболевания



(b) Группа с обычным течением заболевания

Рис. 14. Пример из кардиологии. Эмпирическая, теоретическая функции распределения, доверительные границы для теоретической функции распределения. По оси x — годы.

6. Заключение

В работе предложен новый способ построения оценок для одной специальной модели кривых дожития. Исследованы некоторые важные статистические свойства полученных оценок, как правило, эти свойства носят асимптотический характер.

Из проблем, связанных с процедурой проведения эксперимента, успешно решена проблема определения начала наблюдения. Остальные проблемы, отмеченные в разделе 1. (цензурирование, группировка, усечение), требуют дополнительного рассмотрения: моделирование, проведенное в разделе 5.2.2. показывает возникновение дополнительных «эффектов», связанных с процедурой цензурирования.

Кроме того, в работе рассмотрены некоторые простейшие процедуры проверки статистических гипотез (как относительно значения параметров, так и относительно согласия эмпирической функции распределения с теоретической), которые могут быть применены при изучении реальных данных. Видится, что тема критерия согласия для рассматриваемой модели кривых дожития может быть достаточно хорошо изучена с использованием развитой в работе техники.

Также свойства оценок параметров были изучены на модельных выборках (с цензурированием и без). Полученные результаты полностью согласуются с теоретическими выводами.

Оценки были использованы для изучения реальных выборок из стоматологии и кардиологии.

7. Список литературы

1. Барт А.Г., Бондаренко Б.Б., Бойко В.И. *Математический анализ течения ХГН* // Гломерулонефрит. М.:Наука. **1980**, С. 213-225.

2. Барт А.Г., Клочкова (Алексеева) Н.П. *Критические периоды в кривых дожития* // Статистические методы в клинических испытаниях. Под редакцией А.А. Жиглявского и В.В. Некруткина — СПб.: Изд-во С.-Петерб. ун-та. **1999**.

3. Барт А.Г. *Анализ медико-биологических систем. Метод частично-обратных функций* — СПб.: Изд-во С.-Петерб. ун-та. **2003**.

4. Калинин О.М. *О единых математических трактовках в биологической систематике и динамике популяций и о связи диффузии с нелинейными уравнениями* // Проблемы кибернетики, **1972**. Вып. *25*. С. 107-117.

5. Кокс Д.Р., Оукс Д. *Анализ данных типа времени жизни.* — М.:Финансы и статистика. 1988.

6. Мадай Д.Ю., Барт А.Г., Рыжак Г.А., Коробейников А.И., Боярова С.К. *Репаративная эффективность вилона при лечении больных пожилого и старческого возраста с хроническим генерализованным пародонтитом и сопутствующими возрастными соматическими заболеваниями* — Великий Новгород: Изд-во НовГУ им. Ярослава Мудрого. **2006**.

7. Мартынов Г.В. *Критерий омега-квадрат*. М.:Наука, Гл. ред. физ.-мат. лит., **1978**.

8. Anderson T.W., Darling D.A. *Asymptotic Theory of Certain «Goodness of Fit» Criteria Based on Stochastic Processes* // The Ann. of Math. Stat., **1952**, *23*, 193-212.

9. Barlow R.E., Proschan F. *Statistical Theory of Reliability and Life Testing* — Holt, Rinehart and Winston, NewYork. **1975**.

10. Bart A.G., Bart V.A., Steland A., Zaslavskiy M.L. *Modeling Disease Dynamics and Survivor functions by Sanogenesis Curves* // Journal of Statistical Planning and Inference, **2005**, *32*, pp 33–51.

11. Cheng R.C.H., Traylor L. *Non-Regular Maximum Likelihood Problems* // Journal of the Royal Statistical Society. Series B (Methodological), **1995**, *57*, pp. 3-44.

12. Csörgő S., Faraway Julian J. *The Exact and Asymptotic Distribution of Cramer-von Mises Statistics* // J. R. Statist. Soc. Ser. B, **1996**, *58*, 221-234.

13. Darling D.A. *The Cramer-Smirnov Test in Parametric Case* // The Ann. of Math. Stat., **1955**, *26*, 1-20.

14. Durbin J. *Weak Convergence of the Sample Distribution Function when Parameters are Estimated* // The Ann. of Stat., **1973**, *1*, 279-290.

15. Durbin J. *Kolmogorov-Smirnov Tests when Parameters are Estimated with Applications to Tests of Exponentiality and Tests of Spacings* // Biometrika, **1975**, *62*, 5-22.

16. Durbin J., Knott M. *Components of Cramer-von Mises Statistics, I* // J. R. Statist. Soc. Ser. B, **1972**, *34*, 290-307.

17. Durbin J., Knott M., Taylor C.C. *Components of Cramer-von Mises Statistics, II* // J. R. Statist. Soc. Ser. B, **1975**, *37*, 216-237.

18. Durbin J., Knott M., Taylor C.C. *Corrigenda: Components of Cramer-von Mises Statistics, II* // J. R. Statist. Soc. Ser. B, **1977**, *39*, 394.

19. Klein J.P., Goel P.K. (eds.) *Survival Analysis: State of the Art*. Kluwer Academic Publication, Dordrecht-Boston-London. *1992*.

20. Pettitt A.N. *Test for the Exponential Distribution with Censored Data Using Cramer-von Mises Statistics* // Biometrika, **1977**, *64*, 629-632.

21. Smith R.L. *Maximum Likelihood Estimation in a Class of Nonregular Cases* // Biometrika, **1985**, *72*, No. 1, pp. 67-90.

22. Weiss L., Wolfowitz J. *Maximum Likelihood Estimation of a Translation Parameter of a Truncated Distribution* // The Annals of Statistics, **1973**, *1*, pp. 944-947.