Факультет ВМК
Кафедра ПТВ

Курсовая работа

Тема: "Оценивание распределения по случайно цензурированным выборкам"

Исполнитель:
студент гр. 8302 Кочеганов В. М.

Научный руководитель:
д. ф.-м. н. профессор Тихов М. С.

Н. Новгород 2010 г.

Вступительные замечания

Предположим, что целью математического исследования является некоторый статистически устойчивый эксперимент E, который задается комплексом условий его проведения $\Sigma = \{u_1; u_2; \ldots\}$ и множеством $\mathfrak{F} = \{A; B; A_1; B_1; \ldots\}$ допустимых исходов. Если комплекс Σ таков, что возможно построение достаточно точной вероятностной модели $(\Omega; \mathcal{F}; P)$, то дальнейшее изучение эксперимента возлагается на аппарат теории вероятностей (хотя, необходимо подчеркнуть, что целью теории вероятностей является и само построение тройки $(\Omega; \mathcal{F}; P)$).

Совершенно другой предстает та же задача исследования эксперимента, если условия его проведения не позволяют создать точную математическую модель в виде $(\Omega; \mathcal{F}; P)$ (существует, хотя бы и небольшая, но неопределенность в задании вероятностной меры P на измеримом пространстве $(\Omega; \mathcal{F})$). Причем, иногда удается выбрать некоторый класс $\mathcal{P}$ вероятностных мер, в котором предположительно находится "истинная" мера P (хотя это и не обязательно!).

Таким образом, мы пришли к одной из важнейших задач математической статистики: нахождение меры $P \in \mathcal{P}$, исходя из математической модели эксперимента

$$E \leftrightarrow (\Omega; \mathcal{F}; \mathcal{P}),$$

называемой статистической структурой.

Практически во всех прикладных вопросах сама природа пространства $(\Omega; \mathcal{F})$ не играет никакой роли, но важными являются некоторые числовые характеристики, присущие каждому элементарному исходу. Необходи-

мость искать вероятностную меру, определенную для всех мыслимых наблюдаемых исходов эксперимента E, в этом случае просто отпадает, и вместо неё ищется только распределение P_ξ (интегральная функция распределения) интересующей нас случайной величины ξ, принимающей значения в R^k. Таким образом, не ограничивая общности, можем рассматривать в роли статистической структуры тройку $(R^k; \mathfrak{B}_X; \mathcal{P}_\xi)$, где $\mathfrak{B}_X$— борелевская σ-алгебра на R^k.

Далее, в математической статистике предполагается, что исследуемый эксперимент E состоит в проведении n эквивалентных между собой экспериментов E_i, для каждого из которых определена случайная величина ξ_i, отвечающая за один и тот же количественный признак (возможно векторный) во всех n экспериментах. Причем за случайную переменную ξ, о которой говорилось выше, принимают вектор $(\xi_1; \xi_2; \ldots; \xi_n)$, где ξ_i принимает значения из R^m. Её называют выборкой. Конкретную

реализа-
цию вектора $(\xi_1;\xi_2;\ldots;\xi_n)$ будем обозначать $(x_1;x_2;\ldots;x_n)$ и называть также — выборкой. Таким образом, за сложным экспериментом E, на основе которого и проводятся математические выкладки, на самом деле скрывается один единственный интересующий нас эксперимент E_1 (E_2, $\ldots, E_n$), который, в силу некоторых обстоятельств, не представляется возможным рассматривать отдельно.

Замечание. Важно подчеркнуть, что список вопросов, решаемых математической статистикой, далеко не исчерпывается нахождением
неизвестного распределения. К нему могут быть также отнесены, например, проверка статистических гипотез, оценка зависимости, управление процессами и др.

Итак, ниже будет рассматриваться статистическая структура вида $(R^{m\times n};\mathfrak{B}_X;\mathcal{P}_\xi)$, где множество вероятностных мер $\mathcal{P}_\xi$ будем считать таким, что относительно любой меры $P \in \mathcal{P}$ вектор $\xi = (\xi_1;\xi_2;\ldots;\xi_n)$ состоит из независимых и одинаково распределенных компонент. А поскольку любое распределение находится во взаимно однозначном соответствии с функцией распределения, то эквивалентной задачей является поиск (оценивание) функции распределения случайной величины ξ, представимой в следующем виде:

$$F_\xi(\mathrm{x}) = \prod_{i=1}^{n} F(x_i),$$

где $F(x_1)$ — функция распределения случайной величины ξ_1.

В случае $m = 1$ задача хорошо известна и имеет следующий подход к решению, который порождает глубокие результаты. Вводится эмпирическая функция распределения $F_n(x)$ (которую и выбирают в качестве оценки для неизвестной функции $F(x)$), где

$$F_n(x) = \frac{\sum_{i=1}^{n} I(X_i < x)}{n}$$

и $I(A)$-индикатор события A. При увеличении объема выборки, т. е. при $n \to \infty$, данная функция обладает следующими замечательными свойствами:

1) $F_n(x) \xrightarrow{p} F(x)$, x фиксирован (следует из теоремы Бернулли);

2) $P(F_n(x) \to F(x)) = 1$, x фиксирован (следует из теоремы Бореля);

3) положим $\Delta_n = \sup_x |F_n(x) - F(x)|$, тогда

$$P(\Delta_n \to 0) = 1, \text{ равномерно по всем } x \epsilon R$$

(Теорема Гливенко);

4) $\sqrt{n}(F_n(x) - F(x)) \Rightarrow_D \eta \epsilon N\Big(0, F(x)(1 - F(x))\Big)$, x фиксирован

(следует из теоремы Муавра-Лапласа);

5) $P(\sqrt{n}\Delta_n < x) \to K(x) = \sum_{j=-\infty}^{+\infty}(-1)^j \exp(-2j^2x^2), x > 0.$

Однако вопрос становится совсем другим, если не во всех опытах значение числовой характеристики мы наблюдаем явно.

1. Постановка задачи

Частным случаем рассмотренной выше задачи является оценивание функции распределения, когда измерение числовых характеристик некоторых объектов подвергается воздействию внешних факторов и, следовательно, не может рассматриваться как полноценный элемент выборки.

Цель данной работы - рассмотреть ещё более узкий круг задач, в которых роль воздействий играют различные виды цензурирования. Оно возникает, например, практически во всех экспериментах по наблюдению за объектами, связанными с временем жизни (медицина, испытания изделий на надежность и т. п.). Числовой характеристикой, подвергающейся измерению, является время наступления некоторого события, связанного с объектом (время его отказа, время начала работы и т. п.). Ей соответствует случайная величина X с неизвестной функцией распределения $F(x)$, которую необходимо оценить по имеющейся выборке.

Именно, мы рассмотрим следующие конкретные случаи.

1) Для одной части экспериментов (их количество обозначим n_1) время отказа наблюдаются непосредственно, а для другой — с точностью до некоторого определенного интервала (partly interval-censored data, $n - n_1$). Здесь задачу также разбобьем на два подслучая.

   (1.1) *"Case 1" partly interval-censored data.*

   Об i-ом ($i = n_1 + 1, \ldots, n$) цензурированном объекте имеем информацию только о его статусе (уже отказал/еще не отказал). Причем эту информацию получаем относительно случайного числа U_i. С практической точки зрения, U_i — это время единственной проверки состояния объекта (наступило интересующее нас событие или еще нет). Информация в таком случае может быть представлена выборкой следующего вида:

$$\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_{n_1}, \ (\delta_i, U_i)_{n_1+1 \ \le \ i \ \le \ n},$$

где $\delta_i = 1$, если неизвестное время отказа $\mathrm{X}_i \le U_i$, и $\delta_i = 0$ иначе.

(1.2) *("Case 2") General partly interval-censored data.*

В данном случае для каждого объекта имеем уже не один, а несколько моментов проверки $U_1 < U_2 < \ldots < U_m$ ($U_0 = 0$, $U_{m+1} = \infty$), также являющихся случайными. Положим

$$\Delta = (\delta_1, \delta_2, \ldots, \delta_{m+1}), \ \delta_k = I(U_{k-1} < X \le U_k),$$

где $I(A)$ — индикатор события A (т. е. $\delta_k = 1$, если отказ произошел между моментами U_{k-1} и U_k). Эффективную информацию тогда можно представить в выборке

$$\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_{n_1}, \ (\Delta_i, U_{m_i-1,i}, U_{m_i,i}), \quad i = n_1 + 1, \ldots, n.$$

2) Предположим, что каждый(!) объект может быть подвержен цензурированию. Более конкретно, имеем следующую ситуацию.

При проведении эксперимента E_i одновременно с интересующим нас временем отказа объекта X_i имеем ещё одну числовую характеристику Y_i (независимую от X_i) — время срабатывания внешнего фактора на наблюдение, при котором наблюдение сразу прекращается. Если последняя оказалась меньше, то в выборке имеем значение для Y_i, иначе имеем истинное время отказа объекта X_i. Положим

$$\zeta_i = min(X_i, Y_i), \quad \delta_i = I(X_i < Y_i), \quad 1 \ \le \ i \ \le \ n.$$

Тогда выборка будет иметь вид

$$\xi = (\zeta_i, \delta_i)_{1 \ \le \ i \ \le \ n}.$$

Она носит название случайно цензурировнной выборки (СЦВ).

2. Первый случай цензурирования

2.1. Задачи из практики

В этом пункте речь пойдет о двух видах группового цензурирования ("Case 1" partly interval-censored data, General partly interval-censored data), которых объединяет тот факт, что значение некоторых моментов отказа неизвестно и заключено в случайном интервале.

Ситуации, в которых данные предстают именно в таком виде, в практике встречаются достаточно часто. Например, при оценке распределения такой числовой характеристики, как возраст человека в самом начале его хронического заболевания(время начала хронической болезни). Для одних этот возраст известен точно. Однако, таких людей немного, и основную часть данных составляют случаи, когда известно лишь текущее состояние человека. Для уже заболевших время начала меньше текущего возраста, для здоровых — больше. Это первый вид цензурирования ("Case 1").

Второй вид выборок возникает при, так называемых, исследованиях по текущим данным (follow-up studies). Например, речь может идти о времени начала стенокардии у пациента, который находится на учете. Для некоторых людей это время записано точно, но для остальных время заболевания находится между двумя приемами у врача.

2.2. Подход к задаче

Описываемый подход представлен в [3] и заключается в следующем. Пусть $F(x)$ — неизвестная функция распределения случайной величины X, моментов прекращения. Введем функцию

$$L_n(F) = \prod_{i=1}^{n_1} dF(X_i) \prod_{i=n_1+1}^{n} [F(U_{m_i,i}) - F(U_{m_i-1,i})], \tag{1}$$

где через $dF(x)$ обозначим значение плотности в точке x в непрерывном и вероятность попадания в точку x в дискретном случае.

Определение 1. Непараметрической оценкой максимального правдоподобия F_n (nonparametric maximum likelihood estimator, NPMLE) называется функция, на которой достигается максимальное значение функционала (1) среди всех функций распределения, т. е.

$$F_n = arg \max_{F \, \epsilon \, \mathcal{D}} (L_n(F)),$$

где $\mathcal{D}$ - класс всех функций распределения.

Для удобства вместо функции L_n рассматривают функцию $\log L_n$, поскольку произведение в (1) переходит в сумму

$$\log L_n(F) = \sum_{i=1}^{n_1} \log dF(X_i) + \sum_{i=n_1+1}^{n} \log F(U_{m_i,i}) - F(U_{m_i-1,i}). \tag{2}$$

Ясно, что поставленная выше задача эквивалентна следующей

$$F_n = arg \max_{F \in \mathcal{D}} (\log L_n(F)),$$

Очевидно, что если такая функция существует, то она будет иметь скачки только в точках времени отказа $(X_i)_{1 \le i \le n_1}$ и времени проверок $(U_{m_i-1,i}, U_{m_i,i})_{n_1+1 \le i \le n}$. В таком случае нет необходимости различать $F(x)$ и x. Тогда задача сводится к известной задаче выпуклого математического программирования

$$s^* = arg \max_{s \in S} (\log L_n(s)),$$

где $S = \{(s_1, \ldots, s_k : 0 < s_1 \le \ldots \le s_k < 1)\}$ — выпуклое множество некоторой размерности k, а функция $\log L_n(s)$, вогнутая на этом множестве, имеет единственный максимум.

Таким образом, данное выше определение корректно и функция NPMLE существует и единственна.

2.3. Основные теоремы о сходимости NPMLE

В [3] показано, что введенная выше функция $F_n(x)$ обладает рядом полезных свойств, которые позволяют использовать её в качестве оценки для неизвестной функции $F(x)$. В частности, результаты экспериментов показывают, что NPMLE, основанная на всех наблюдениях (в том числе и на цензурированных), существенно превосходит эмпирическую функцию распределения, построенную только на явно наблюдаемых значениях времен отказов.

Сделаем несколько предположений об эксперименте E.

Первое предположение состоит в том, что количество наблюдаемых отказов не пренебрежимо мало по сравнению со всем объемом n.

Предположение $(A1): \ n_1/n \xrightarrow[n\to\infty]{} \alpha_1 > 0.$

Это предположение является ключевым, поскольку оно порождает высокий порядок сходимости для F_n: $\sqrt{n}$. Если $n_1/n \xrightarrow[n\to\infty]{} 0$, то сходимость будет медленнее чем $\sqrt{n}$. В частности, для $n_1 = 0$ сходимость будет порядка $\sqrt[3]{n}$.

Теорема 1. ("Case 1"partly interval-censored data) Предположим, что выполнено $(A1)$ и функция $F(x)$ абсолютно непрерывна. Тогда

$$\sqrt{n}(F_n(x) - F(x)) \Rightarrow_D Z_1,$$

где $\Rightarrow_D$ означает сходимость по распределению, Z_1 — Гауссовский процесс с нулевым математическим ожиданием и ковариационной функцией,
достигающей нижней границы оценки для $F(x)$. Вид ковариационной функции можно найти в [3].

Предположение $(A2)$: функция распределения $F(x)$ является абсолютно непрерывной и строго возрастающей.

Предположение $(A3)$: совместная функция распределения $G(u_1, \ldots, u_m)$ случайной величины $(U_1, \ldots, U_m)$ является абсолютно непрерывной. Более того, существует положительное число $\eta_0 > 0$ такое, что

$$P(U_{k+1} - U_k \geq \eta_0) = 1, \; k = 1, \ldots, m-1.$$

Предположение $(A3)^*$: $(U_1, \ldots, U_m)$ - дискретная случайная величина. Причем вероятность того, что она примет значение из некоторого конечного множества, равна единице.

Предположение $(A3)$ предполагает, что между любыми двумя смежными проверками $(U_i, \; U_{i+1})$ должно пройти как минимум некоторое время η_0.

Теорема 2. (General partly interval-censored data) Предположим, что выполнены $(A1)$ и $(A2)$. Пусть также выполнено одно из условий $(A3)$ или $(A3)^*$. Тогда

$$\sqrt{n}(F_n(x) - F(x)) \Rightarrow_D Z_2,$$

где Z_2 - Гауссовский процесс с нулевым математическим ожиданием и ковариационной функцией, достигающей нижней границы оценки для $F(x)$. Для случая $m = 2$ вид ковариационной функции можно найти в [3].

Доказательство сформулированных теорем также можно найти в [3].

2.4. Эффективность при использовании цензурированных данных

Для определения практической эффективности, даваемой NPMLE на основе всех наблюдений относительно эмпирической функции распределения, использующей только явно наблюдаемые отказы, а также для определения зависимости её от коэффициента α_1, доли явно наблюдаемых отказов, были проведены несколько экспериментов. Роль этих экспериментов намного больше, чем может показаться на первый взгляд, поскольку аналитически разрешить вопрос об эффективности и её зависимости от параметра α_1 достаточно сложно (слишком сложную и слишком скрытую структуру имеет предельная ковариационная функция NPMLE).

Моделирование экспериментов было разбито на две части.

В первой части рассматривалось цензурированые выборки первого вида ("Case 1"). Было взято четыре различных вида распределения: равномерное на $[0;1]$, экспоненциальное со средним 0.5, два распределения Вейбулла с параметрами $(k_1,\lambda_1)=(1.4,0.55)$ и $(k_2,\lambda_2)=(0.7,0.4)$. Распределение проверочного времени U — равномерное на $[0;1]$.

Во второй части эксперимента рассматривались цензурированные выборки второго вида ("Case 2"). Здесь также было взято четыре распределения: равномерное на $[0;1]$, экспоненциальное со средним значением 2 и два распределения Вейбулла с параметрами $(k_1,\lambda_1)=(1.5,2.25)$ и $(k_2,\lambda_2)=(0.7,1.65)$. Распределение случайной величины U_1 является равномерным на $[0;1]$. После реализации U_1, получают реализацию U_2, добавляя к уже полученному значению U_1 случайное число, полученное из равномерного на $[0;1]$ распределения.

Размер выборки $n=n_1+n_2=100$ и количество повторов для каждой модели — 500. Всего было использовано пять различных комбинаций $(n_1;n_2)$: $\alpha_1=0.1,\ 0.3,\ 0.5,\ 0.7,\ 0.9$. Среднеквадратическая ошибка(MSE) вычислена как среднее арифметическое квадратов разностей значений функций $F_n(x)$ и $F(x)$ в моменты набюдения отказов. Результаты моделирования представлены в таблицах 1 и 2.

Через F_n^{emp} обозначена эмпирическая функция распределения; через Rate — отношение ошибки для F_n^{emp} к F_n.

Из таблиц 1 и 2 видно, что среднеквадратическая ошибка NPMLE F_n меньше (относительно эмпирической функции) для значений параметра $\alpha_1=0.1$ до 0.7. Для этих значений относительная эффективность F_n находится примерно в интервале от 1.2 до 8.2. Даже в случае $\alpha_1 = 0.9$ есть небольшое превосходство в оценке функцией F_n. В следствие сказанного, результаты, представленные в таблицах 1 и 2, показывают преимущество оценок, проводимых с использованием ценурированных данных в дополнение к явно наблюдаемым отказам, над оценками, использующими только явные отказы. Из таблиц также можно заключить, что увеличение ошибки при уменьшении n_1, при постоянном n, не существенно. Последний факт подсказывает, что производительность оценки NPMLE стабильно для широких рамок значений α_1 и постоянном n.

3. Случайно цензурированные выборки (СЦВ)

3.1. Задачи из практики

В данной части работы речь пойдет о случайно цензурированных выборках, являющихся, в некотором роде, частным случаем групповых цензурированных выборок, рассмотренных выше.

Как уже отмечалось, практические задачи, порождающие данную ситуацию, связаны с временем жизни некоторых объектов. Ситуация говорит сама за себя, однако было бы неплохо ее уточнить.

Существует большой класс технических объектов, статистическая информация о надежности которых по определению не может быть полной и достаточно определенной. Это, прежде всего, высоконадежные и мало-
серийные объекты, а также уникальные объекты, для которых нет прототипов (аналогов). Подразумевается, безусловно, что объекты, для которых решается задача оценки надежности, созданы по единой технологии. За длительный период наблюдения такие объекты могут иметь ограниченное количество отказов. Наработки неотказавших за этот же период объектов являются дополнительной (и, как мы видели в предыдущем случае цензурирования, весьма ценной) информацией о надежности всей контролируемой совокупности объектов.

3.2. Оценка Нельсона

Одной из статистик, используемых для оценки функции распределения моментов отказа $F(x)$, является оценка Нельсона.

Итак, пусть $(X_i; Y_i)_{1 \le i \le n}$ — последовательность независимых одинаково распределенных пар случайных величин, причем X_i и Y_i для $1 \le i \le n$ также независимы между собой. Обозначим $F(x)$ и $G(x)$ функции распределения случайных величин X_1 и Y_1 соответственно; $\zeta_i = \min(X_i; Y_i)$ и $\delta_i = I(X_i < Y_i)$, $1 \le i \le n$.

Заметим, что $1 - P(\zeta_i < x) = (1 - F(x))(1 - G(x))$.

Рассмотрим статистику

$$W_n = \frac{1}{n} \sum_{j=1}^{n-1} \frac{a(\zeta_n^{(j)})\delta_n^{[j]}}{1 - j/n},$$

где $a(y)$ — измеримая функция, $\zeta_n^{(j)}$ - j-ая порядковая статистика, а $\delta_n^{[j]}$- индуцированная порядковая статистика (т. е. если пары $(\zeta_j; \delta_j)_{1 \le j \le n}$

упорядочены по первой компоненте и $\zeta_n^{(j)} = \zeta_i$, то $\delta_n^{[j]} = \delta_i$).

Известно, что последовательность $\sqrt{n}(W_n - \mu)$ сходится по распределению к нормальной случайной величине с нулевым математическим ожиданием и дисперсией

$$\sigma^2 = \int\limits_{-\infty}^{+\infty} \frac{a^2(y)dF(y)}{(1-F(y))^2(1-G(y))}.$$

Здесь

$$\mu = \int\limits_{-\infty}^{+\infty} \frac{a(y)dF(y)}{1-F(y)}. \tag{3}$$

Статистика

$$S_n^2 = \frac{1}{n}\sum_{j=1}^{n-1} \frac{a(\zeta_n^{(j)})\delta_n^{[j]}}{(1-j/n)^2} \tag{4}$$

асимптотически нормальна с математическим ожиданием, равным

$$\mu_S = \int\limits_{-\infty}^{+\infty} \frac{a^2(y)dF(y)}{(1-F(y))^2(1-G(y))},$$

поэтому ее можно использовать для оценки дисперсии σ^2.

Оценка Нельсона получается из предыдущего при $a(y) = I(y < x)$ и определяется следующим образом:

$$V_n(x) = \frac{1}{n}\sum_{j=1}^{n-1} \frac{I(\zeta_n^{(j)} < x)\delta_n^{[j]}}{1-j/n}.$$

Из (3) следует:

$$\mu = \int\limits_{-\infty}^{+\infty} \frac{I(y<x)dF(y)}{1-F(y)} = \int\limits_{-\infty}^{x} \frac{dF(y)}{1-F(y)} = \int\limits_{F(-\infty)}^{F(x)} \frac{dz}{1-z} = -\int\limits_{0}^{F(x)} d\ln(1-z) =$$

$$= -\ln(1-z)\big|_0^{F(x)} = -\ln(1-F(x)).$$

Откуда, в частности, следует состоятельность оценки Нельсона. Для доказательства докажем более общее утверждение.

Утверждение 1. Пусть α — случайная величина, а $(\alpha_n)_{n\geq 1}$ — последовательность случайных величин на том же вероятностном пространстве, и

$$\sqrt{n}(\alpha_n - \alpha) \Rightarrow_D \beta.$$

Тогда

$$\alpha_n \xrightarrow{p} \alpha.$$

Доказательство:

Действительно, пусть $\sqrt{n}(\alpha_n - \alpha) \Rightarrow_D \beta$. Тогда

$$P(\sqrt{n}|\alpha_n - \alpha| < x) \xrightarrow[n\to\infty]{} F_\beta(x) - F_\beta(-x+0)$$

и

$$P(|\alpha_n - \alpha| < \epsilon) = P(\sqrt{n}|\alpha_n - \alpha| < \sqrt{n}\epsilon) \geq P(\sqrt{n}|\alpha_n - \alpha| < k) \xrightarrow[n\to\infty]{}$$

$$\xrightarrow[n\to\infty]{} F_\beta(k) - F_\beta(-k+0),$$

где $k \in N$.

И следовательно, выбирая k так, чтобы $F_\beta(k) - F_\beta(-k+0) \geq 1 - \frac{\delta}{2}$ и n, чтобы $\sqrt{n}\epsilon \geq k$ и $P(\sqrt{n}|\alpha_n - \alpha| < k) \geq F_\beta(k) - F_\beta(-k+0)$, получим:

$$\forall \epsilon > 0,\ \forall \delta > 0 \exists N = n^* :\ \forall n > n^* P(|\alpha_n - \alpha| < \epsilon) \geq 1 - \delta,$$

что и требовалось доказать.■

Взяв в качестве α_n оценку Нельсона $V_n(x)$, а в качестве α — $(-\ln(1 - F(x)))$, получаем, что $V_n(x) \xrightarrow[n\to\infty]{p} -\ln(1 - F(x))$.

В качестве оценки для дисперсии можно взять величину, определяемую из формулы (4):

$$S_n^2 = \frac{1}{n}\sum_{j=1}^{n-1} \frac{I(\zeta_n^{(j)} < x)\delta_n^{[j]}}{(1 - j/n)^2}.$$

3.3. Оценка Каплана-Мейера

Попробуем выразить из оценки Нельсона оценку непосредственно для функции $1 - F(x)$ (а не для $-\ln(1 - F(x))$).

Обозначим $I(\zeta_n^{(j)} < x) = I_j$ и $\delta_n^{[j]} = \delta_j$. Учитывая разложение в ряд Тейлора

$$\ln(1 - x) = -x - \frac{x^2}{2(1-\theta)^2},\quad \theta \in (0, x),$$

имеем последовательно:

$$V_n(x) = \frac{1}{n}\sum_{j=1}^{n-1} \frac{I_j\delta_j}{1 - j/n} = -\sum_{j=1}^{n-1}\Big(-\frac{I_j\delta_j}{n-j}\Big) = -\sum_{j=1}^{n-1} I_j\delta_j\Big(\ln\big(1 - \frac{1}{n-j}\big) +$$

$$+\frac{1}{2(1-\theta_j)^2(n-j)^2}\Big) = -\sum_{j=1}^{n-1} I_j\delta_j \ln\big(1-\frac{1}{n-j}\big) - \sum_{j=1}^{n-1} I_j\delta_j \frac{1}{2(1-\theta_j)^2(n-j)^2}.$$

Обозначим последнюю сумму через γ_n, а первую(с минусом) через $v_n(x)$. Далее, если мы покажем, что

$$\sqrt{n}(V_n(x)-\mu) - \sqrt{n}(v_n(x)-\mu) = \sqrt{n}(V_n(x)-v_n(x)) = -\sqrt{n}\gamma_n \xrightarrow{p} 0, \quad (5)$$

то отсюда будет следовать, что

$$\sqrt{n}(v_n(x)-\mu) \Rightarrow_D N(0,\sigma^2), \quad (6)$$

поскольку $\sqrt{n}(V_n(x)-\mu) \Rightarrow_D N(0,\sigma^2)$, $\mu = -\ln(1-F(x))$. Действительно, поскольку если случайные величины $(X_n)_{n\geq 1}$, $(Y_n)_{n\geq 1}$ и Y таковы, что $(X_n - Y_n) \xrightarrow{p} 0$ и $Y_n \Rightarrow_D Y$, то $X_n \Rightarrow_D Y$.

Докажем (5). Пусть $j^* = \max\{j : 1 \leq j \leq n-1,\ I_j = 1\}$. Имеем последовательно:

$$\sum_{j=1}^{n-1} I_j\delta_j \frac{\sqrt{n}}{2(1-\theta_j)^2(n-j)^2} \leq \sum_{j=1}^{n-1} I_j\delta_j \frac{\sqrt{n}}{2(1-\frac{1}{n-j})^2(n-j)^2} =$$

$$= \sum_{j=1}^{n-1} I_j\delta_j \frac{\sqrt{n}}{2(n-j-1)^2} \leq \frac{\sqrt{n}}{2(n-j^*-1)} \sum_{j=1}^{n-1} I_j\delta_j \frac{1}{n-j-1} = \frac{\sqrt{n}}{2(n-j^*-1)} \times$$

$$\times \sum_{j=1}^{n-1} I_j\delta_j \Big(\frac{1}{n-j} + \frac{1}{(n-j-1)(n-j)}\Big) \leq \frac{\sqrt{n}}{2(n-j^*-1)} \sum_{j=1}^{n-1} I_j\delta_j \frac{1}{n-j} +$$

$$+\frac{\sqrt{n}}{2(n-j^*-1)^2} \sum_{j=1}^{n-1} I_j\delta_j \frac{1}{n-j}.$$

И, в итоге,

$$\sum_{j=1}^{n-1} I_j\delta_j \frac{\sqrt{n}}{2(1-\theta_j)^2(n-j)^2} \leq V_n(x)\Big(\frac{\sqrt{n}(n-j^*)}{2(n-j^*-1)^2}\Big) \quad (7)$$

Лемма 1. $\frac{\sqrt{n}}{n-j^*} \xrightarrow{p} 0$

Доказательство:

По своему смыслу, величина $(n-j^*)$ означает количество отказов (X_i), время которых превышает x. Тогда, если ввести бернуллиевские случайные величины $(\eta_i)_{1\ \leq\ i\ \leq\ n}$, причем

$$(\eta_i=1)\ \Leftrightarrow\ (X_i>x),\ P(\eta_i=1)=p=1-F(x+0)>0,$$

то $\sum_{i=1}^n \eta_i = S_n = n-j^*$ и

$$P(\frac{\sqrt{n}}{n-j^*}<\epsilon)=P(\frac{n-j^*}{\sqrt{n}}>\frac{1}{\epsilon})=P(\frac{S_n}{\sqrt{n}}>\frac{1}{\epsilon})=P(\frac{S_n}{n}>\frac{1}{\sqrt{n}\epsilon})=$$

$$P(\frac{S_n}{n}-p>\frac{1}{\sqrt{n}\epsilon}-p)\geq_{(\forall n>n^*)} P(\frac{S_n}{n}-p>\frac{p}{2}-p)=P(\frac{S_n}{n}-p>-\frac{p}{2})=$$

$$=1-P(\frac{S_n}{n}-p\leq-\frac{p}{2})\xrightarrow[n\to\infty]{}1.$$

Последнее следует из теоремы Бернулли.■

Проделанные выкладки позволяют утверждать, как уже отмечалось, что

$$\sqrt{n}\bigg(v_n(x)-\Big(-\ln\big(1-F(x)\big)\Big)\bigg)\Rightarrow_D N(0,\sigma^2). \qquad (8)$$

Лемма 2. Если

$$\sqrt{n}(T_n-\theta)\Rightarrow_D N(0,\tau^2),$$

то

$$\sqrt{n}(f(T_n)-f(\theta))\Rightarrow_D N(0,\big(\frac{d}{d\theta}f(\theta)\big)^2\tau^2),$$

при условии, что $\frac{d}{d\theta}f(\theta)$ существует, непрерывна и не равна нулю.

Доказательство:

Разложим $f(T_n)$ по формуле Тейлора в точке θ:

$$f(T_n)=f(\theta)+(T_n-\theta)(\frac{d}{d\theta}f(\theta)+R_n), \qquad (9)$$

где $R_n \xrightarrow[T_n\to\theta]{} 0$. Из условия леммы следует, что $T_n \xrightarrow{p} \theta$ и, следовательно, $R_n \xrightarrow{p} 0$. Легко показать, что если $Y_n \Rightarrow_D Y$ и $A_n,\ B_n$ сходятся по вероятности к a и b, то $A_n+B_nY_n \Rightarrow_D a+bY$.

Из разложения (9) находим, что

$$\sqrt{n}(f(T_n)-f(\theta))=\sqrt{n}(T_n-\theta)(\frac{d}{d\theta}f(\theta)+R_n)$$

и, положив $A_n = 0,\ B_n = \sqrt{n}(T_n - \theta),\ Y_n = \frac{d}{d\theta}f(\theta) + R_n$, имеем результат леммы 2.■

Преобразуем выражение для $-v_n$

$$\sum_{j=1}^{n-1} I_j\delta_j \ln\left(1 - \frac{1}{n-j}\right) = \ln\prod_{j=1}^{n-1}\left(1 - \frac{1}{n-j}\right)^{I_j\delta_j} = \ln\prod_{j=1}^{n-1}\left(\frac{n-j-1}{n-j}\right)^{I_j\delta_j}$$

Применим для выражения

$$\sqrt{n}\Big(\big(-v_n(x)\big) - \ln\big(1 - F(x)\big)\Big) \Rightarrow_D N(0, \sigma^2)$$

и функции $f(x) = e^x$ лемму 2. Получим:

$$\sqrt{n}\left(\prod_{j=1}^{n-1}\left(\frac{n-j-1}{n-j}\right)^{I_j\delta_j} - \Big(1 - F(x)\Big)\right) \Rightarrow_D N(0, \big(1 - F(x)\big)^2\sigma^2).$$

Оценка $S_n^{KM} = \prod_{j=1}^{n-1}\left(\frac{n-j-1}{n-j}\right)^{I_j\delta_j}$ называется оценкой Каплана-Мейера. Очевидно, $S_n^{KM} \xrightarrow{p} \big(1 - F(x)\big)$ и ассимптотическая дисперсия имеет вид

$$\rho^2 = \big(1 - F(x)\big)^2 \int\limits_{-\infty}^{x} \frac{dF(y)}{(1 - F(y))^2(1 - G(y))}. \tag{10}$$

3.4. Модель пропорциональных рисков

Пусть

$$S_X(x) = 1 - F(x),$$
$$S_Y(x) = 1 - G(x),$$
$$1 - Z(x) = (1 - F(x))(1 - G(x)) = P(\zeta_1 \geq x),$$
$$F(x) = G(x) = 0,\ \forall x \leq 0.$$

Некоторые практические ситуации позволяют сделать сильные предположения относительно функций $S_X(x)$ и $S_Y(x)$. Например, при диагностировании осложнений некоторых болезней применяют модель пропорциональных рисков, которая предполагает следующий вид функций выживаемости:

$$S_Y(x) = S_X^{\beta}(x).$$

Рассмотрим эмпирическую функцию распределения величины ζ:

$$Z_n(x) = \frac{\sum_{i=1}^{n} I(\zeta_i < x)}{n}$$

и пусть $\delta^* = \frac{\sum_{i=1}^{n} I(\delta_i=1)}{n}$. Рассмотрим в качестве оценки для функции выживаемости $S_X(x)$ оценку Абдушукурова-Ченга-Лина (ACL-оценка):

$$S_X^{ACL} = (1 - Z_n(x))^{\delta^*}.$$

Докажем сначала состоятельность этой оценки. Из теоремы Бернулли следует, что $\delta^* \xrightarrow{p} P(\delta = 1)$ и $1 - Z_n(x) \xrightarrow{p} 1 - Z(x)$.

Для $P(\delta = 1)$ последовательно имеем

$$P(X < Y) \leq \sum_{i=1}^{\infty} P(X \epsilon [x_i, x_{i+1}), X \leq Y) \leq \sum_{i=1}^{\infty} P(X \epsilon [x_i, x_{i+1}), x_i \leq Y) \leq$$

$$\leq \sum_{i=1}^{\infty} P(X \epsilon [x_i, x_{i+1}))P(x_i \leq Y) = \sum_{i=1}^{\infty} (F(x_{i+1}) - F(x_i))S_Y(x_i) \xrightarrow[\Delta_i \to 0]{} .$$

$$\xrightarrow[\Delta_i \to 0]{} \int\limits_0^{\infty} S_y(x) dF(x).$$

Аналогично показывается, что $P(X < Y) \geq \int\limits_0^{\infty} S_y(x) dF(x)$, и, значит, $P(X < Y) = \int\limits_0^{\infty} S_y(x) dF(x)$. Применительно к нашей модели последнее выражение упрощается в

$$P(X < Y) = -\int\limits_0^{\infty} S_X^{\beta}(x) dS_X(x) = -\int\limits_{S(0)}^{S(+\infty)} z^{\beta} dz = \left.\frac{z^{\beta+1}}{\beta+1}\right|_0^1 = \frac{1}{\beta+1}.$$

Одновременно с этим

$$1 - Z(x) = S_X(x) S_Y(x) = S_X^{\beta+1}(x).$$

Имеет место утверждение, доказательство которого аналогично доказательству леммы 2: из того факта, что $\alpha_n \xrightarrow{p} \alpha$, $\beta_n \xrightarrow{p} \beta$ и $f(x, y)$ дифференцируема, следует, что $f(\alpha_n, \beta_n) \xrightarrow{p} f(\alpha, \beta)$. В данной задаче $f(x, y) = (1 - x)^y$, $\alpha_n = H_n$, $\beta_n = \delta^*$. Теперь легко видеть, что $S_X^{ACL} \xrightarrow{p} S_X(x)$.

Решим задачу о нахождении асимптотической дисперсии для ACL-оценки. Сначала упростим выражение для $D(S_X^{ACL}(x))$. Именно, запишем неоднократно использовавшийся нами ряд Тейлора для логарифма:

$$\ln S_X^{ACL}(x) = \ln S_X(x) + \frac{S_X^{ACL}(x) - S_X(x)}{S_X(x)} + o(S_X^{ACL}(x) - S_X(x)),$$

откуда видно, что

$$\ln S_X^{ACL}(x) - \left(\ln S_X(x) + \frac{S_X^{ACL}(x) - S_X(x)}{S_X(x)} \right) \xrightarrow{p} 0,$$

и, значит, в качестве асимптотической дисперсии для $S_X^{ACL}(x)$ можно взять $\lim_{n\to\infty} \Big(S_X^2(x) D(\ln(S_X^{ACL}(x))) \Big)$, а в качестве математического ожидания для $\ln S_X^{ACL}(x)$ величину $\ln(S_X(x))$.

Для еще более наглядного вывода воспользуемся следующим утверждением.

Утверждение 2. Пусть $X_1,\ X_2$ независимые случайные величины с функцией времени жизни $S_1,\ S_2$ соответственно. Пусть $Z = \min(X_1, X_2)$ и $I = i$ тогда и только тогда, когда $Z = X_i,\ (i = 1, 2)$ и пусть $0 < P(I = i) < 1$. Тогда Z и I независимы тогда и только тогда, когда существует действительное положительное число β такое, что $S_2(x) = S_1^\beta(x)$.

Доказательство:

Предположим, что $S_2(x) = S_1^\beta(x)$. Заметим, что, во-первых,

$$P(I = 1) = P(X_1 < X_2) = -\int\limits_0^\infty S_2(x) dS_1(x) = -\int\limits_0^\infty S_1^\beta(x) dS_1(x) = \frac{1}{1+\beta}$$

и аналогично $P(I = 2) = \frac{\beta}{1+\beta}$.
Во-вторых,

$$P(Z > x) = P(X_1 > x, X_2 > x) = S_1(x) S_2(x) = S_1^{1+\beta}(x).$$

Тогда

$$P(I = 2, Z > x) = P(X_2 < X_1, X_2 > x) = -\int\limits_x^\infty S_1(x) dS_2(x) =$$

$$= -\int\limits_x^\infty S_1(x) dS_1^\beta(x) = \frac{\beta}{\beta+1} S_1^{1+\beta}(x) = P(I = 2) P(Z > x)$$

и аналогично $P(I = 1, Z > x) = P(I = 1)P(Z > x)$, $\forall x \,\epsilon\, R$, откуда следует независимость величин I и Z.

Обратно, пусть $\forall x \,\epsilon\, R$

$$P(I=1,Z>x) = P(I=1)P(Z>x) = -\int\limits_x^\infty S_2(x)dS_1(x) = S_1(x)S_2(x)\frac{1}{\beta+1}.$$

Предположив для функций $S_1(x)$ и $S_2(x)$ существование непрерывных производных, продифференцируем последнее равенство. Получим

$$S_2(x)d(S_1(x))(1+\beta) = dS_1(x)S_2(x) + dS_2(x)S_1(x),$$

что равносильно дифференциальному следующему уравнению с разделяющимися переменными

$$\beta\frac{dS_1(x)}{S_1(x)} = \frac{dS_2(x)}{S_2(x)},$$

решая которое, находим

$$S_2(x) = S_1(x)^\beta C.$$

Константа C находится предельным переходом при $x \to 0$, т. к. $S_1(0) = S_2(0) = 1$, и равна 1. Что и требовалось доказать.

Далее, учитывая полученные результаты, займемся непосредственным вычислением асимптотической дисперсии для ACL-оценки. Последовательно имеем

$$DS_X^{ACL}(x) \sim S_X^2(x)D\ln S_X^{ACL}(x) = S_X^2(x)D\big[\delta^*\ln(1-Z_n(x))\big] =$$

$$= S_X^2(x)\Big(M\Big[\delta^*\ln(1-Z_n(x))\Big]^2 - M^2\Big[\delta^*\ln(1-Z_n(x))\Big]\Big);$$

в силу независимости δ^* и $Z_n(x)$

$$M\big[\delta^*\ln(1-Z_n(x))\big]^2 = M\delta^{*2}M\big[\ln(1-Z_n(x))\big]^2 = \left(\frac{p(1-p)}{n}+p^2\right)\times$$

$$\times\Big(D\Big[\ln(1-Z_n(x))\Big]+M^2\Big[\ln(1-Z_n(x))\Big]\Big) \sim \left(\frac{p(1-p)}{n}+p^2\right)\left(\frac{D\big[1-Z_n(x)\big]}{(1-Z(x))^2}+\right.$$

$$\left.+\ln^2(1-Z(x))\right) = \left(\frac{p(1-p)}{n}+p^2\right)\left(\frac{Z(x)}{n(1-Z(x))}+\ln^2(1-Z(x))\right)$$

и

$$M^2\big[\delta^* \ln(1 - Z_n(x))\big] = M^2\delta^* M^2\big[\ln(1 - Z_n(x))\big] \sim p^2 \ln^2(1 - Z(x)).$$

Подставляя все в выражение для $DS_X^{ACL}(x)$, получаем:

$$nDS_X^{ACL}(x) = nS_X^2(x)\Big(\frac{p(1-p)Z(x)}{n^2(1-Z(x))} + \frac{p(1-p)\ln^2(1-Z(x))}{n} + \frac{p^2Z(x)}{n(1-Z(x))}\Big)$$

и, переходя к пределу при $n \to \infty$, заключаем, что

$$nDS_X^{ACL}(x) \xrightarrow[n\to\infty]{} S_X^2(x)\Big(p(1-p)\ln^2(1-Z(x)) + \frac{p^2Z(x)}{1-Z(x)}\Big).$$

3.5. Сравнение асимптотики оценок KM и ACL в случае модели пропорциональных рисков

В модели пропорциональных рисков выражение для асимптотической дисперсии ACL-оценки упрощается в

$$nDS_X^{ACL}(x) \sim S_X^2(x)\beta \ln^2 S_X(x) + (\beta+1)^{-2}(1 - S_X^{\beta+1})S_X^{1-\beta}. \qquad (11)$$

Аналогично, из выражения (10) находим выражение для асимптотической дисперсии оценки Каплана-Мейера:

$$\rho^2 = -S_X^2(x)\int\limits_0^x \frac{dS_X(y)}{S_X^2(y)S_X^\beta(y)} = (1+\beta)^{-1}S_X^2(x)(S_X^{-1-\beta}(x) - 1). \qquad (12)$$

Взяв отношение (11) к (12) получим эффективность оценки Каплана-Мейера относительно ACL-оценки, а именно

$$\varepsilon_1 = \varepsilon_1(x) = (\beta+1)\beta\frac{\ln^2 S_X(x)S_X^{\beta+1}(x)}{1 - S_X^{\beta+1}(x)} + \frac{1}{\beta+1}.$$

Для большей наглядности предположим, что $S_X(x) = 1 - x = t$ и последнее выражение примет вид:

$$\varepsilon_1 = \varepsilon_1(t,\beta) = (\beta+1)\beta\frac{t^{\beta+1}\ln^2 t}{1 - t^{\beta+1}} + \frac{1}{\beta+1},\ 0 < t < 1.$$

На Рис.1 приведен график функции $\varepsilon_1(t,\beta)$ для значений параметра $\beta = 0.1, 0.2, 0.5, 1.0, 5.0, 10.0, 20.0, 100.0$(кривые располагаются сверху вниз). Графики функций показывают, что эффективность KM-оценок падает, если значение параметра β увеличивается. Кроме того, для больших значений параметра β вместо оценки $S_n^{KM}(x)$ лучше использовать оценку $S_X^{ACL}(x)$.

Рис. 1. График функции $\varepsilon_1(t, \beta)$

3.6. Другой способ обоснования оценки КМ

Большое количество математических открытий делается на основе некоторых субъективных рассуждений, которые в последствии строго обосновываются (конечно, если они привели к продуктивным результатам на гипотетическом уровне). Сейчас же рассмотрим вопрос о том, как можно было придти к оценке Каплана-Мейера

$$S_n^{KM}(t) = \prod_{j=1}^{n-1} \left(\frac{n-j-1}{n-j} \right)^{I(\zeta_n^{(j)} < t)\delta^{[j]}},$$

используя простейший математический аппарат и немного допущений.

Предположим, что моменты прекращения, соответствующие нецензурированным наблюдениям, расположены в порядке возрастания: $t_1 < t_2 < \ldots < t_k$.

Пусть h_j—количество нереализованных моментов цензурирования, т. е. количество наблюдений, моменты прекращения которых действительно наблюдались, с длительностями t_j при $j = 1, 2, \ldots, k$. Очевидно, что при отсутствии в выборке длительностей с совпадающими значениями, все h_j равны единице. Введем также величину m_j, показывающую количество наблюдений, цензурированных в момент времени между t_j и t_{j+1}; в данном случае m_k будет показывать количество наблюдений, продолжающихся дольше, чем длительность t_k, соответствующая наиболее продолжительном нецензурированному наблюдению. Таким образом, величина n_j, которая показывает количество либо незаконченных, либо цензурированных к моменту t_j моментов остановок, равна

$$n_j = \sum_{i \geq j}^{k} (m_i + h_i).$$

Все это позволяет определить оценку функции риска $\lambda(t_j)$ как $\lambda_j^* = \lambda^*(t_j) = \frac{h_j}{n_j}$, т. е. как отношение количества моментов прекращения, которые действительно имели место к моменту t_j, к количеству наблюдений, которые потенциально могли окончиться к этому моменту. Соответственно, оценка функции выживания $S(t_j)$ в данном случае будет собой представлять совместную вероятность того, что к моменту t_j событие не завершится:

$$S_n^{KM}(t_j) = \prod_{i=1}^{j} \frac{n_i - h_i}{n_i} = \prod_{i=1}^{j} (1 - \lambda_i^*).$$

Последнее выражение в точности совпадает с оценкой КМ, если положить все $h_j = 1$.

Данные рассуждения также дают толчок полагать, что оценку КМ можно использовать и для наблюдений с повторяющимися временами отказа.

Список литературы

[1] Крамер Г. Математические методы статистики. — М.: Мир, 1975.

[2] Гнеденко Б. В.Курс теории вероятностей. — М.: Наука, 1988.

[3] Huang J. Asymptotic properties of nonparametric estimation based on partly interval-censored data. — Statistica Sinica, 9, 501-519, 1999.

[4] Godambe, V. P. An optimum property of regular maximum likelihood estimation. — Ann. Math. Statist. 31, 1208-1211, 1960.

[5] Рыбалко В. В. Определение закона надежности высоконадежных и малосерийных объектов по случайно цензурированным выборкам. — EXPonenta Pro, 1, 44-48, 2003.

Содержание

	(n_1, n_2)	(10,90)	(30,70)	(50,50)	(70,30)	(90,10)
Равномерное	F_n^{emp}	116.27	30.95	19.48	12.58	9.12
$[0; 1]$	NPMLE	14.16	14.07	12.08	10.44	8.60
	Rate	8.21	2.20	1.61	1.20	1.06
Экспон.	F_n^{emp}	98.16	31.73	19.33	11.73	8.19
среднее= 0.5	NPMLE	16.24	14.12	13.07	10.8	7.88
	Rate	6.04	2.25	1.48	1.09	1.04
Вейбулл	F_n^{emp}	120.19	31.28	17.14	11.81	10.10
$k = 1.4$	NPMLE	15.89	12.47	11.65	9.51	9.80
$\lambda = 0.55$	Rate	7.56	2.51	1.47	1.24	1.03
Вейбулл	F_n^{emp}	108.7	30.67	18.01	12.73	8.97
$k = 0.7$	NPMLE	18.5	15.24	13.32	10.83	8.77
$\lambda = 0.4$	Rate	5.88	2.01	1.35	1.18	1.02

Таблица 1. MSE×10^4: "Case 1" partly interval censoring

	(n_1, n_2)	(10,90)	(30,70)	(50,50)	(70,30)	(90,10)
Равномерное	F_n^{emp}	109.51	31.20	18.87	11.77	10.21
$[0; 4]$	NPMLE	14.10	11.53	10.97	9.47	9.41
	Rate	7.77	2.71	1.72	1.24	1.09
Экспон.	F_n^{emp}	105.47	29.63	17.55	14.35	9.9
среднее= 2	NPMLE	15.94	14.49	12.09	11.46	9.59
	Rate	6.61	2.04	1.45	1.25	1.03
Вейбулл	F_n^{emp}	94.09	31.71	19.25	11.62	9.98
$k = 1.5$	NPMLE	12.56	11.82	11.07	9.35	9.16
$\lambda = 2.25$	Rate	7.49	2.68	1.74	1.24	1.09
Вейбулл	F_n^{emp}	109.98	30.34	19.51	14.86	10.27
$k = 0.7$	NPMLE	18.95	15.62	14.08	12.08	9.85
$\lambda = 1.65$	Rate	5.7	1.94	1.38	1.23	1.04

Таблица 2. MSE×10^4: "Case 2" partly interval censoring