УДК 519.2

ОЦЕНИВАНИЕ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ ВЕЙБУЛЛА ПО СЛУЧАЙНО ЦЕНЗУРИРОВАННЫМ ВЫБОРКАМ

М.С. Тихов, В.В. Агеев, Т.С. Бородина

Нижегородский госуниверситет им. Н.И. Лобачевского

tikhovm@mail.ru

Поступила в редакцию 23.09.2009

Распределение Вейбулла является важным распределением, особенно для анализа надежности и ремонтопригодности. В статье представлены некоторые методы оценивания параметров распределения Вейбулла, именно параметра формы α и масштабного параметра σ. Доказана состоятельность и асимптотическая нормальность полученных оценок.

Ключевые слова: распределение Вейбулла, случайно цензурированная выборка, параметрическое оценивание, метод наименьших квадратов.

Введение

Распределение Вейбулла является одним из важнейших в теории надежности. Оно имеет функцию распределения

$$F(x) = 1 - \exp\left(-\left(\frac{x-\mu}{\sigma}\right)^{\alpha}\right), \quad (1)$$

$$x > \mu, \sigma > 0, \alpha > 0.$$

Это распределение является также одним из трех предельных распределений нормированных разностей минимума (максимума) независимых и одинаково распределенных случайных величин. Статистическому анализу распределения Вейбулла и его параметров посвящены две большие недавно вышедшие монографии (см. [1, 2]), список цитируемых статей которых насчитывает свыше тысячи источников. Основной метод оценивания параметров этого распределения – метод максимального правдоподобия. Однако ввиду аналитической сложности получаемых при этом уравнений для нахождения значений параметров распределения Вейбулла он используется в совокупности с численными методами, что вносит дополнительную погрешность в определение значений оценок. Другим способом оценивания является довольно простой и аналитически удобный метод спрямления, который сводится к следующему нелинейному функциональному преобразованию уравнения (1) (полагая $\mu = 0$):

$$z = \alpha y - \lambda, \quad (2)$$

где $z = \ln(-\ln(1 - F(x)))$, $y = \ln x$, $\lambda = \alpha \ln \sigma$.

Таким образом, в системе координат $y0z$ уравнение (2) есть уравнение прямой и оценку параметров α и λ, а значит, и σ можно производить по методу наименьших квадратов. Для полных, цензурированных по типу I и типу II, прогрессивно цензурированных выборок такие оценки предложены в работах [1–5] и были исследованы в основном численными методами. Однако в литературе совсем не представлены подходы к построению и не исследовано поведение оценок по случайно цензурированным выборкам, что важно для практических приложений. В данной работе мы рассматриваем задачу построения оценок параметров распределения Вейбулла по случайно цензурированным выборкам. В преобразовании (2) мы используем, в основном, множительные оценки Каплана–Мейера [6] и оценки Нельсона–Аалена [7–9], а также метод наименьших квадратов. Мы показываем, что полученные оценки являются состоятельными и асимптотически нормальными. Рассмотрены и другие, важные в теории надежности и эконометрике, распределения, для которых применим предложенный подход.

1. Построение оценок

Мы рассматриваем случайно цензурированную выборку, когда вместо величин $X_i, 1 \le i \le n$, наблюдаются только пары (Z_1, δ_1), (Z_2, δ_2), …, (Z_n, δ_n), $Z_i = \min(X_i, T_i)$, $\delta_i = I(X_i \le T_i)$, $i = 1, 2, \ldots$, где $\{T_i,\ 1 \le i \le n\}$ есть цензурирующие величины, $I(A)$ обозначает индикатор события A. Оценки параметров распределения Вейбулла будем строить по случайно цензурированным данным $\{(Z_i, \delta_i), 1 \le i \le N\}$.

Имея в виду практическое применение модели случайного цензурирования, будем считать, что X_i, $1 \le i \le n$, и T_j, $1 \le j \le n$, являются неотрицательными и независимыми случайными величинами, кроме того, каждая из случайных величин T_j не зависит от каждой из величин X_i, где X_i имеет функцию распределения $F(x) = 1 - \exp(-(x/\sigma)^\alpha)$. Цензурирующие величины T_1, T_2, … предполагаются независимыми, одинаково распределенными случайными величинами с известной или неизвестной функцией распределения $G(x)$. В таком случае функция распределения величины Z_i равна

$$H(x) = \mathbf{P}(Z_i \le x) = 1 - (1 - F(x))G(x) = \\ = 1 - S(x)\overline{G}(x),$$

где $S(x) = 1 - F(x)$, $\overline{G}(x) = \mathbf{P}(T_i > x)$.

Хорошо известно, что наилучшей в определенном смысле оценкой функции распределения $F(x)$ по выборке $(Z_i, \delta_i), 1 \le i \le n$, является оценка Каплана–Мейера (КМ) [6], определенная следующим образом:

$$\hat{S}_n(x) = 1 - \hat{F}_n(x) = \\ = \prod_{i=1}^{n}\left(1 - \frac{\delta_{[i:n]}}{n-i+1}\right)^{I(Z_{i:n} \le x)} = \tag{3} \\ = \prod_{u \le x}\left(1 - \frac{d\overline{N}(u)}{\overline{Y}(u)}\right),$$

где $Z_{i:n}$ есть *i*-я порядковая статистика, построенная по выборке $Z_1, Z_2, \ldots, Z_n$, а $\delta_{[i:n]}$ – *i*-я сопутствующая порядковая статистика: если $Z_{i:n} = Z_j$, то $\delta_{[i:n]} = \delta_j$; $\overline{N}(t) = \sum_{j=1}^{N} N_j(t)$, $N_j(t) = I(X_j \le t, \delta_j = 1)$, $\overline{Y}(t) = \sum_{j=1}^{n} Y_j(t)$, $Y_j(t) = I(X_j > t)$.

Известно (см. [10]), что для оценки (3) имеет место соотношение:

$$\mathbf{P}\left(\sup_{-\infty < x < +\infty} |\hat{F}_n(x) - F(x)| = O\left(\sqrt{\frac{\ln n}{n}}\right)\right) = 1,$$

т.е. почти наверное $\hat{F}_n(x)$ равномерно сходится к истинной функции распределения $F(x)$.

Известно также, что эти оценки асимптотически нормальны с ожиданием $F(x)$ и асимптотической дисперсией $\dfrac{S^2(x)}{n}\displaystyle\int_0^x \frac{dF(y)}{S^2(y)\overline{G}(y)}$.

Оценки параметров распределения Вейбулла будем находить из условия минимума по α и λ выражения

$$Q \equiv Q(\alpha, \lambda) = \int_0^\infty (\ln(-\ln(1 - \hat{F}_n(x))) - \\ - \alpha \ln x + \lambda)^2 \, d\hat{F}_n(x).$$

Исходя из необходимого условия экстремума (для квадратичной функции оно является и достаточным), получаем систему уравнений:

$$\begin{cases} \displaystyle\int_0^\infty \ln x \cdot \ln(-\ln(1 - \hat{F}_n(x)))\, d\hat{F}_n(x) = \\ \displaystyle\quad = \alpha \int_0^\infty \ln^2 x \, d\hat{F}_n(x) - \lambda \int_0^\infty \ln x \, d\hat{F}_n(x), \\ \displaystyle\int_0^\infty \ln(-\ln(1 - \hat{F}_N(x)))\, d\hat{F}_n(x) = \\ \displaystyle\quad = \alpha \int_0^\infty \ln x \, d\hat{F}_n(x) - \lambda. \end{cases} \tag{4}$$

Положим:

$$\hat{\nu}_1 = \int_0^\infty \ln(-\ln(1 - \hat{F}_n(x)))\, d\hat{F}_n(x),$$

$$\hat{\nu}_2 = \int_0^\infty \ln x \cdot \ln(-\ln(1 - \hat{F}_n(x)))\, d\hat{F}_n(x),$$

$$\hat{\tau}_1 = \int_0^\infty \ln x \, d\hat{F}_n(x), \quad \hat{\tau}_2 = \int_0^\infty \ln^2 x \, d\hat{F}_n(x).$$

Тогда система (4) примет вид:

$$\begin{cases} \hat{\nu}_2 = \hat{\tau}_2\alpha - \hat{\tau}_1\lambda, \\ \hat{\nu}_1 = \hat{\tau}_1\alpha - \lambda, \end{cases}$$

решением ее будет

$$\hat{\alpha} = \frac{\hat{\nu}_1\hat{\tau}_1 - \hat{\nu}_2}{\hat{\tau}_1^2 - \hat{\tau}_2}, \quad \hat{\lambda} = \frac{\hat{\nu}_1\hat{\tau}_2 - \hat{\nu}_2\hat{\tau}_1}{\hat{\tau}_1^2 - \hat{\tau}_2}.$$

2. Состоятельность оценок

Имеет место следующий результат.

Теорема 1. *Предположим, что* $S(x) = \exp(-(x/\sigma)^\alpha)$, *а* $G(x)$ – *непрерывная функция. Тогда* $\hat{\alpha} \xrightarrow[n\to\infty]{p} \alpha$, $\hat{\lambda} \xrightarrow[n\to\infty]{p} \lambda$.

Доказательство. Учитывая представление

$$\hat{\nu}_1 = \int_0^\infty \ln(-\ln(1 - \hat{F}_n(x)))\, d\hat{F}_n(x) = \\ = \int_0^\infty \ln(-\ln(1 - F(x)))\, d\hat{F}_n(x) + \\ + \int_0^\infty \ln(-\ln(1 - \hat{F}_n(x)) - \ln(-\ln(1 - F(x)))\, d\hat{F}_n(x)$$

и что при $n \to \infty$

$$\sup_{x\le x_0}|\ln(-\ln(1-\hat F_n(x)))-\ln(-\ln(1-F(x)))|\le$$

$$\le C_1\sup_{x\le x_0}|\ln(1-\hat F_n(x))-\ln(1-F(x))|\le$$

$$\le C_2\sup_{x\le x_0}|\hat F_n(x)-F(x)|\,, \qquad (6)$$

используя результаты работ [11, 12], получаем: $\hat\nu_1-\nu_1\xrightarrow[n\to\infty]{p}0$, где

$$\nu_1=\int_0^\infty \ln(-\ln(1-F(x)))\,dF(x)=$$

$$=\int_0^\infty(\alpha\ln x+\lambda)\,dF(x).$$

Аналогично,

$$\hat\nu_2=\int_0^\infty \ln x\cdot\ln(-\ln(1-\hat F_n(x)))\,d\hat F_n(x)\xrightarrow[n\to\infty]{p}\nu_2\,,$$

где

$$\nu_2=\int_0^\infty \ln x\ln(-\ln(1-F(x)))\,dF(x)=$$

$$=\int_0^\infty \ln x(\alpha\ln x+\lambda)\,dF(x),$$

$$\hat\tau_1=\int_0^\infty \ln x\,d\hat F_n(x)\xrightarrow[n\to\infty]{p}\tau_1=\int_0^\infty \ln x\,dF(x),$$

$$\hat\tau_2=\int_0^\infty \ln^2 x\,d\hat F_n(x)\xrightarrow[n\to\infty]{p}\tau_2=\int_0^\infty \ln^2 x\,dF(x).$$

Из этих соотношений следует, что

$$\hat\alpha\xrightarrow[n\to\infty]{p}\frac{\nu_1\tau_1-\nu_2}{\tau_1^2-\tau_2}=\alpha \;\text{ и }\; \hat\lambda\xrightarrow[n\to\infty]{p}\lambda.$$

3. Асимптотическая нормальность

Поскольку $H(x)$ есть функция распределения величины $Z_i=\min(X_i,T_i)$, то положим $\overline{H}=1-H$ и определим $\tau_H=\inf\{x\ge 0:H(x)=1\}$.

Теорема 2. *При условиях теоремы 1* $\sqrt{n}(\hat\alpha-\alpha)$ *и* $\sqrt{n}(\hat\lambda-\lambda)$ *асимптотически* (*при* $n\to\infty$) *нормальны.*

Доказательство. Для установления сходимости нормированных разностей оценок $\sqrt{n}(\hat\alpha-\alpha)$ и $\sqrt{n}(\hat\lambda-\lambda)$ к нормальным случайным величинам воспользуемся результатами работ [11–13], в которых показано, что

$$\sqrt{n}\left(\int\varphi\,d\hat F_n-\int_0^{\tau_H}\varphi\,dF\right)\xrightarrow[n\to\infty]{d}\mathcal{N}(0,\sigma_1^2),$$

$$\sigma_1^2=\int_0^{\tau_H}\frac{\varphi^2(x)}{\overline{G}(x)}dF(x)-\left(\int_0^{\tau_H}\varphi(x)\,dF(x)\right)^2-$$

$$-\int_0^{\tau_H}\frac{S(x)}{1-H(x)}\left\{\int_x^{\tau_H}\varphi(y)\,dF(y)\right\}dG(x),$$

где измеримая функция $\varphi:R\to R$ такова, что $\int\varphi^2\,dF<\infty$.

Из этого результата следует, например, что $\sqrt{n}(\hat\nu_1-\nu_1)$ и $\sqrt{n}(\hat\tau_1-\tau_1)$ асимптотически нормальны $\mathcal{N}(0,\sigma_2^2)$ и $\mathcal{N}(0,\sigma_3^2)$, если в качестве функции $\varphi(x)$ взять $\varphi_2(x)=\ln(-\ln(1-F(x)))$ и $\varphi_3(x)=\ln x$ соответственно.

Из разложения

$$\hat\nu_1\hat\tau_1-\nu_1\tau_1=\hat\nu_1(\hat\tau_1-\tau_1)+\tau_1(\hat\nu_1-\nu_1)$$

выводим, что последовательность $\sqrt{n}(\hat\nu_1\hat\tau_1-\nu_1\tau_1)$ при $n\to\infty$ будет асимптотически нормальна $\mathcal{N}(0,\sigma_4^2)$, где $\sigma_4^2=\sigma_2^2+\sigma_3^2+2\sigma_{23}$, $\sigma_{23}=\lim\limits_{N\to\infty}\sqrt{n}\,\mathbf{E}((\hat\nu_1-\nu_1)(\hat\tau_1-\tau_1))$.

Аналогично, $\sqrt{n}(\hat\nu_1\hat\tau_2-\nu_1\tau_2)$, $\sqrt{n}(\hat\nu_2\hat\tau_1-\nu_2\tau_1)$ и $\sqrt{n}(\hat\tau_1^2-\tau_1^2)$ будут асимптотически нормальны $\mathcal{N}(0,\sigma_5^2)$, $\mathcal{N}(0,\sigma_6^2)$ и $\mathcal{N}(0,\sigma_7^2)$.

Далее, пусть $\sqrt{n}(\hat T_{1n}-T_1)$ и $\sqrt{n}(\hat T_{2n}-T_2)$ асимптотически нормальны $\mathcal{N}(0,\sigma_8^2)$, $\mathcal{N}(0,\sigma_9^2)$. Имеем представление (см. [14]):

$$\frac{\hat T_{2n}}{\hat T_{1n}}-\frac{T_2}{T_1}=\frac{\hat T_{2n}-T_2}{T_1}-\frac{T_2(\hat T_{1n}-T_1)}{T_1^2}+$$

$$+O\left(\frac{(\hat T_{2n}-T_2)(\hat T_{1n}-T_1)}{T_1^2}\right)+$$

$$+O\left(\frac{T_2(\hat T_{1n}-T_1)^2}{T_1^3}\right),$$

из которого следует асимптотическая нормальность последовательностей $\sqrt{n}(\hat\alpha-\alpha)$ и $\sqrt{n}(\hat\lambda-\lambda)$. Теорема доказана.

В представленных рассуждениях вместо $\ln(1-\hat F_n(x))$ можно взять другие оценки функции распределения (например, оценки Нельсона–Аалена $A_n(x)$ [7–9]) и мы получим те же предельные распределения оценок параметров

α и λ, когда используются статистики $A_n(x)$, поскольку $\hat{F}_n$ и $\exp(-A_n)$ имеют одни и те же предельные распределения. Однако оценка $S_n^* = \exp(-A_n)$, являясь асимптотически несмещенной, при конечных n имеет смещение, и в некоторых случаях ее значения могут сильно отличаться от оценки $\hat{S}_n(x)$, поэтому для уменьшения этого смещения предлагается воспользоваться аппроксимациями Паде (см. [15]) для функции $\ln(1-x)$. Эти аппроксимации можно найти, воспользовавшись пакетом символьной математики Maple 12. Именно, находим, что

$$Pade\,[1,1] = -2x/(2-x),$$

$$Pade\,[2,2] = (3x^2 - x)/(6 - 6x + x^2).$$

Для демонстрации точности аппроксимации рассмотрим следующий пример.

Дана выборка объема $N = 80$ на интервале $[0, 4]$, в которой

z_j	d_j	c_j
1	38	1
2	16	2
3	10	2
4	8	3

где d_j – число отказов в момент z_j (наблюдалась величина X_j и $\delta_j = 1$), c_j – число приостановок в момент z_j (наблюдалась величина T_j и $\delta_j = 0$).

Результаты расчетов приведены в таблице.

Таблица

(1)	(2)	(3)	(4)	(5)
1	0.5250	0.6219	0.5364	0.5253
2	0.3200	0.4210	0.3303	0.3204
3	0.1809	0.2725	0.1895	0.1811
4	0.0494	0.1317	0.0604	0.0502

Здесь в первом столбце приведены значения переменной x, во втором – значения оценки Каплана–Мейера $\hat{S}(x)$, в 3-м – значения функции $\exp(-A_n(x))$, в 4-м – значения функции $\exp(-Pade\,[1,1])$, в 5-м – значения функции $\exp(-Pade[2,2])$. Таким образом, представленный пример говорит о том, что уже аппроксимация [2/2] достаточно хорошо приближает оценку КМ.

4. Другие распределения и оценивание их параметров

Подход, аналогичный представленному, можно использовать и для других распределений: распределения, которые являются предельными для нормированных разностей экстремальных порядковых статистик (см. пп. 4.1 и 4.2), а также для распределений 4.3 и 4.4.

4.1. Распределение Фреше (обратное распределение Вейбулла (см. [2], 3.3.3, с. 129):

$$F_1(x) = \exp\left(-\left(\frac{x}{\sigma}\right)^{-\alpha}\right), x > 0, \sigma > 0, \alpha > 0.$$ Здесь необходимое преобразование есть $\ln(-\ln F_1(x)) = -\alpha\ (\ln x - \ln\sigma)$.

4.2. Распределение Гнеденко–Гумбеля (лог-распределение Вейбулла) (см. [2], 3.3.4, с.131)

$$F_2(x) = 1 - \exp(-\exp(\alpha(x - \ln\sigma))).$$

Преобразование:

$$\ln(-\ln(1 - F_2(x))) = -\alpha\ (\ln x - \ln\sigma).$$

4.3. Распределение Парето – при больших значениях переменной x является аппроксимацией распределения Фреше:

$$F_3(x) = 1 - \left(\frac{x}{\sigma}\right)^{-\alpha}, x > \sigma, \sigma > 0, \alpha > 0.$$

Преобразование:

$$\ln(1 - F_3(x)) = -\alpha(\ln x - \ln\sigma).$$

4.4. Логистическое распределение. Является сверткой двух распределений 4.2:

$$F_4(x) = \frac{1}{1 + \exp\left(\left(\frac{|x|}{\sigma}\right)^{\alpha} \cdot \operatorname{sign} x\right)},$$

$$-\infty < x < \infty, \sigma > 0, \alpha > 0.$$

Преобразование:

$$\ln\left((\operatorname{sign}x \cdot \ln)\frac{1}{F(x)} - 1\right) = \alpha\left(\ln|x| - \sigma\right).$$

Список литературы

1. Dobson B. The Weibull analysis handbook. ASQ Quality Press, 2006. 167 p.

2. Rinne H. The Weibull distribution. A Handbook. CRC Press, 2009. 762 p.

3. Cohen A.C. Maximum likelihood estimation in the Weibull distribution based on complete and on censored samples // Technometrics. 1965. V. 7, No. 4. P. 579–588.

4. Al-Fawzan M.A. Methods for estimating the parameters of the Weibull distribution. URL: http:/interstat.statjournals.net/YEAR/2000/articles/0010001.pdf.

5. Romeu J.L. Empirical assessment of Weibull distribution // RAC START. V. 10, No. 3. URL: http://src.alionscience.com/pdf/WEIBULL.

6. Kaplan E.L., Meier P. Nonparametric estimation from incomplete observations // J. Amer. Statist. Assoc. 1958. V. 53. P. 457–481.

7. Nelson W. Hasard plotting for incomplete failure data // J. Qual. Tech. 1969. V. 1. P. 27–52.

8. Aalen O. Nonparametric inference in connection with multiple decrement models // Scandinavian J. Statist. 1976. P. 15–27.

9. Aalen O. Nonparametric inference for a family of counting process // Ann. Statist. 1978. V. 6, No. 4. P. 701–726.

10. Foldes A., Rejto L. Strong uniform consistency for nonparametric survival curve estimators from randomly censored data // Ann. Statist., 1981. V. 9, No. 1. P. 122–129.

11. Stute W. The statistical Analysis of Kaplan–Meier integrals // Lecture Notes. 1995. V. 27. P. 231–254.

12. Stute W. The central limit theorem under random censorship // Ann. Statist. 1995. V. 23, No. 2. P. 422–439.

13. Suzukawa A. Asymptotic properties of Aalen–Johansen integrals for competing risk data // J. Japan Statist. Soc. 2002. V. 32, No. 1. P. 77–93.

14. Tikhov M.S. Statistical estimation on the basis of interval-censored data // Journal of Math. Sciences. 2004. V.119, No. 3. P. 321–335.

15. Бейкер Дж., Грейс-Моррис П. Аппроксимации Паде. М.: Мир, 1986. 502 с.

16. Breslow N.E. Discussion of professor Cox's paper // Journal of the Royal Statistical Society, Series B. 1972. V. 34. V. 216–217.

17. Тихов М.С. Линейные функции индуцированных порядковых статистик и непараметрическое оценивание распределений в зависимости доза–эффект // Обозрение прикл. и промышл. мат. 1999. Т. 6, № 1. С. 244.

RANDOMLY CENSORED SAMPLE ESTIMATION OF WEIBULL DISTRIBUTION PARAMETERS

M.S. Tikhov, V.V. Ageev, T.S. Borodina

The Weibull distribution is an important distribution for reliability and maintainability analysis. Some methods for estimating Weibull parameters, namely, shape parameter α and scale parameter σ are presented. Consistency and asymptotic normality of the estimations obtained have been proved.

Keywords: Weibull distribution, randomly censored sample, parametric estimation, least-square method.