Тр. международной конференции "Актуальные проблемы электронного приборостроения". Новосибирск, 1998. Т.3. - С. 17-23.

К ОЦЕНИВАНИЮ ПАРАМЕТРОВ ЗАКОНОВ РАСПРЕДЕЛЕНИЙ ПО НЕПОЛНЫМ ВЫБОРКАМ

Б.Ю.Лемешко, С.Я.Гильдебрант, С.Н.Постовалов

Новосибирский государственный технический университет
Новосибирск, Россия. E-mail: headrd@ nstu.nsk.su

Аннотация. Рассматриваются вопросы оценивания параметров законов распределений по цензурированным выборкам, в которых доступными для наблюдения оказываются не более 50% объема выборки. Исследуются потери в информации Фишера и влияние объема выборки и её наблюдаемой части на точность оценивания параметров распределений.

1. Введение.

При исследовании надежности и контроле качества типична ситуация оценивания параметров распределений по цензурированным слева и/или справа наблюдениям [1]. При этом к моменту прекращения испытаниий большой партии изделий наблюдается выход из строя лишь части из них, обычно достаточно малой по сравнению с объемом всей партии.

При двустороннем цензурировании доступными для наблюдения оказываются $N - n_1 - n_2$ измерений в области левее некоторого значения $x_{(1)}$ и правее $x_{(2)}$

$$\cdots\ x_{(1)} < X_{n_1+1} < X_{n_1+2} < \ldots < X_{N-n_1-n_2} < \ \cdots.$$

Очевидно, что в такой неполной (цензурированной) выборке содержится меньше информации, чем в полной и это, естественно, отражается на точности оценивания параметров.

Наиболее универсальным методом по отношению к форме представления выборочных данных является метод максимального правдоподобия. Оценки максимального правдоподобия (ОМП) неизвестного параметра по цензурированным наблюдениям называется такое значение параметра, при котором функция правдоподобия

$$L(\theta) = \gamma P_1^{n_1}(\theta) P_3^{n_2}(\theta) \prod_{j=n_1+1}^{N-n_1-n_2} f(X_j, \theta), \qquad (1)$$

где γ - некоторая константа; $f(x,\theta)$ - функция плотности случайной величины; $P_1(\theta) = \int_{-\infty}^{x_{(1)}} f(x,\theta)dx$, $P_3(\theta) = \int_{x_{(2)}}^{\infty} f(x,\theta)dx$ - вероятности попадания наблюдений в интервалы левее $x_{(1)}$ и правее $x_{(2)}$, достигает максимума на множестве возможных значений параметра.

При выборке, цензурированной с двух сторон, за редким исключением ОМП вычисляются только численными методами.

2. Количество информации Фишера в наблюдениях цензурированной выборки

Если оценивается скалярный параметр, то асимптотическая дисперсия его ОМП определяется соотношением

$$D(\hat{\theta}) = n^{-1} J_c^{-1}(\hat{\theta}), \qquad (2)$$

где информационное количество Фишера определяется соотношением

$$J_c(\theta)=\frac{1}{P_1(\theta)}\left[\frac{\partial P_1(\theta)}{\partial\theta}\right]^2+\int_{x_{(1)}}^{x_{(2)}}\left[\frac{\partial \ln f(x,\theta)}{\partial\theta}\right]^2 f(x,\theta)dx+\frac{1}{P_3(\theta)}\left[\frac{\partial P_3(\theta)}{\partial\theta}\right]^2. \quad (3)$$

Если выборка цензурирована только справа, то в выражении исчезает левое слагаемое, только слева - правое слагаемое. Это соотношение позволяет судить о потерях информации о параметре распределения в зависимости от степени цензурирования слева или справа.

В случае экспоненциального распределения с функцией плотности $f(x)=\theta\, e^{-\theta x}$ при двустороннем цензурировании количество информации Фишера имеет вид

$$J_c(\theta)=\frac{1}{\theta^2}\left[\frac{1}{P_1(\theta)}\big(1-P_1(\theta)\big)^2 t_{(1)}^2+\big(1-P_1(\theta)\big)t_{(1)}^2+1-P_1(\theta)-P_3(\theta)\right], \quad (4)$$

где $t_{(i)}=\theta x_{(i)}$, $P_1(\theta)=1-e^{-\theta x_{(1)}}$, $P_3(\theta)=e^{-\theta x_{(2)}}$. При правом цензурировании получаем

$$J_c^{n}(\theta)=\frac{1}{\theta^2}\big(1-P_3(\theta)\big), \quad (5)$$

при левом -

$$J_c^{л}(\theta)=\frac{1}{\theta^2}\left[\frac{1}{P_1(\theta)}\big(1-P_1(\theta)\big)^2 t_{(1)}^2+\big(1-P_1(\theta)\big)t_{(1)}^2+1-P_1(\theta)\right]. \quad (6)$$

В случае векторного параметра элементы информационной матрицы Фишера $\mathbf{J}_c$ определяются соотношениями, аналогичными (3).

В табл. 1 значения относительной информации, представляющие собой отношения количества информации Фишера о параметре в цензурированной выборке к количеству информации в нецензурированной выборке $J_c(\theta)/J(\theta)$, в зависимости от степени цензурирования приведены для распределений экспоненциального, Вейбулла, нормального (логарифмически нормального), Лапласа, Рэлея, гамма-распределения. В случае вектора параметров в таблице представлены значения отношения определителей соответствующих информационных матриц $\det\mathbf{J}_c(\theta)/\det\mathbf{J}(\theta)$. В зависимости от закона цензурирование справа и слева по разному влияет на потери информаци о параметрах. Так о параметре экспоненциального распределения при той же степени цензурировании слева в выборке сохраняется существенно больше информации, чем при цензурировании справа. Это же характерно для распределения Вейбулла. В случае гамма-распределения величина относительной информации зависит от параметра формы θ этого распределения и “перераспределяется” между параметрами закона с его ростом. В табл.1 ее значения для гамма-распределения приведены для значений параметра формы 0.5 и 2.

Если задаться максимально допустимой асимптотической дисперсией, величина которой определяется соотношением (3), то в зависимости от степени цензурирования можно оценить минимально необходимый объем выборки, при котором она должна быть не хуже заданной. Или наоборот по объему выборки оценить максимально возможную степень цензурирования, еще обеспечивающую требуемую точность оценивания.

Асимптотическая дисперсия является теоретической характеристикой точности оценивания. Реально же точность оценивания (дисперсия оценки) не в последнюю очередь зависит от особенностей случайной выборки (от возможного наличия в ней аномальных наблюдений, от того, действительно ли наблюдаемая выборка принадлежит предполагаемому закону). На моделируемых выборках различного объема и принадлежащих различным зако-

нам исследована возможность получения достаточно хороших оценок параметров распределений по неполным выборкам при наблюдении только части области определения случайной величины.

Таблица 1.

Относительное количество информации Фишера в наблюдении цензурированной выборки $J_c(\theta)/J(\theta) \quad \left(\det \mathbf{J}_c(\theta)/\det \mathbf{J}(\theta)\right)$

Наблюдаемая часть %	О масштабном параметре распределений экспоненциального и Вейбулла		О параметре формы распределения Вейбулла		О двух параметрах распределения Вейбулла	
	Цензурирование слева	Цензурирование справа	Цензурирование слева	Цензурирование справа	Цензурирование слева	Цензурирование справа
100	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
60	0.9914	0.6000	0.7091	0.4343	0.6389	0.2658
50	0.9805	0.5000	0.6343	0.4011	0.5256	0.1771
40	0.9597	0.4000	0.5680	0.3878	0.4076	0.1093
30	0.9212	0.3000	0.5168	0.3859	0.2878	0.0595
20	0.8476	0.2000	0.4883	0.3814	0.1707	0.0257
10	0.6891	0.1000	0.4830	0.3405	0.0654	0.0063
5	0.5223	0.0500	0.4654	0.2718	0.0234	0.0015
Наблюдаемая часть %	О параметре сдвига нормального распределения	О параметре масштаба нормального распределения	О двух параметрах нормального распределения	О параметре масштаба распределения Лапласа	О параметре распределения Рэлея	
	Цензурирование слева*)	Цензурирование слева*)	Цензурирование слева*)	Цензурирование слева*)	Цензурирование слева	Цензурирование справа
100	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
60	0.8753	0.5599	0.4399	0.6131	0.9914	0.6000
50	0.8183	0.5000	0.3296	0.6103	0.9805	0.5000
40	0.7467	0.4601	0.2311	0.5918	0.9597	0.4000
30	0.6550	0.4399	0.1457	0.5538	0.9212	0.3000
20	0.5336	0.4309	0.0754	0.4885	0.8476	0.2000
10	0.3591	0.4252	0.0239	0.3740	0.6891	0.1000
5	0.2318	0.3795	0.0073	0.2730	0.5223	0.0500
Наблюдаемая часть %	О параметре формы гамма-распределения ($\theta = 0.5$)		О параметре масштаба гамма-распределения ($\theta = 0.5$)		О двух параметрах гамма-распределения ($\theta = 0.5$)	
	Цензурирование слева	Цензурирование справа	Цензурирование слева	Цензурирование справа	Цензурирование слева	Цензурирование справа
100	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
60	0.6693	0.9698	0.9984	0.4750	0.4756	0.3201
50	0.5819	0.9484	0.9955	0.3715	0.3589	0.2163
40	0.4902	0.9157	0.9876	0.2778	0.2522	0.1343
30	0.3927	0.8646	0.9681	0.1939	0.1586	0.0730
20	0.2865	0.7796	0.9208	0.1196	0.0812	0.0311
10	0.1651	0.6192	0.7925	0.0554	0.0251	0.0075
5	0.0935	0.4615	0.6321	0.0265	0.0076	0.0018
Наблюдаемая часть %	О параметре формы гамма-распределения ($\theta = 2$)		О параметре масштаба гамма-распределения ($\theta = 2$)		О двух параметрах гамма-распределения ($\theta = 2$)	
	Цензурирование слева	Цензурирование справа	Цензурирование слева	Цензурирование справа	Цензурирование слева	Цензурирование справа
100	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
60	0.7770	0.9353	0.9759	0.6895	0.4395	0.4171

50	0.7022	0.8985	0.9548	0.5981	0.3284	0.3071
40	0.6168	0.8475	0.9208	0.5005	0.2296	0.2108
30	0.5177	0.7758	0.8659	0.3957	0.1444	0.1292
20	0.3993	0.6697	0.7737	0.2815	0.0749	0.0642
10	0.2483	0.4955	0.6002	0.1543	0.0248	0.0190
5	0.1498	0.3470	0.4369	0.0825	0.0089	0.0054

3. Исследование точности оценивания по неполным выборкам в зависимости от ее объема

В табл. 2 приведены результаты исследования точности вычисления оценки максимального правдоподобия (ОМП) параметра экспоненциального закона по цензурированной справа выборке, когда для наблюдения доступен определенный процент от полного объёма выборки n. В строке "100%" приведена ОМП по полной выборке. Для закона, найденного по цензурированной выборке, по шести используемым в системе [2] критериям проверялось согласие с полной выборкой. Если различие между законами, найденными по полной и по цензурированной выборке, становилось ощутимым (не было оснований для принятия гипотезы о согласии при уровне значимости близком к $\alpha = 0.01$), то данные в таблице помечены одной звездочкой, при полном "несогласии" - двумя звездочками. Из таблицы видно, что при объемах выборки 500-2000 достаточно хорошие оценки получаются по 30 % выборки, то есть при 70 % цензурирования справа. А при объеме выборки в 100 наблюдений достаточно хорошие оценки получаются уже только по 50 % выборки.

Таблица 2.

Исследование точности оценивания параметра экспоненциального закона по цензурированной справа выборке

Наблюдаемая часть %	n=2000	n=1000	n=500	n=300	n=200	n=100
	θ	θ	θ	θ	θ	θ
100%	1.0300	1.0220	1.0592	1.0056	1.0715	1.0864
60%	1.0242	1.0109	1.0018	0.9976	0.9410	0.9348
50%	1.0227	0.9803	1.0009	0.9597	0.9359	0.9757
40%	1.0266	0.9905	0.9769	0.9853	0.9529	*0.8928
30%	1.0146	1.0111	0.9799	*0.9161	*0.9138	*0.9282
20%	*0.9801	**0.9392	0.9825	**0.8619	**0.8190	**0.8288
10%	*0.9793	*0.9468	**0.8205	0.9849	1.0329	**1.4165
5%	**0.9105	**0.8609	**0.8544	**0.6949	1.0015	**1.5764

Таблица 3.

Исследование точности оценивания параметра экспоненциального закона по цензурированной слева выборке

Наблюдаемая часть %	n=2000	n=1000	n=500	n=300	n=200	n=100
	θ	θ	θ	θ	θ	θ
100%	1.0300	1.0220	1.0592	1.0056	1.0715	1.0864
60%	1.0317	1.0229	1.0599	1.0123	1.0747	1.1080
50%	1.0307	1.0219	1.0610	1.0045	1.0797	1.0980
40%	1.0311	1.0280	1.0675	1.0160	1.0787	1.1576

30%	1.0290	1.0280	1.0810	1.0089	1.0949	1.1061
20%	1.0347	1.0145	1.0696	1.0086	1.1118	1.1057
10%	1.0271	1.0512	1.0840	1.0277	1.1616	1.1945
5%	1.0059	1.0664	1.0925	0.9698	1.1403	**1.4304

В табл. 3 представлены аналогичные результаты при вычислении ОМП параметра экспоненциального закона по выборке, цензурированной слева. Как видим, в данном случае оценки, получаемые по сильно цензурированной выборке, в целом ближе к оценкам по полной выборке, чем при цензурировании справа. Это различие в точности оценивания напрямую коррелируется с количеством информации Фишера о параметре, сохранившемся в цензурированной выборке (см. табл. 1): для одной и той же степени цензурирования при цензурировании слева сохраняется больше информации о параметре, чем при цензурировании справа. Больше информации, следовательно, выше точность оценивания.

Наличие в выборке аномальных измерений оказывает существенное влияние на получаемые оценки параметров. Причем наличие аномальных измерений в наблюдаемой области цензурированной выборки отражается на оценках еще более заметно.

4. Заключение

Опираясь на весь цикл проведенных исследований, полностью не вошедших в краткий текст данного сообщения, зафиксируем основные факты.

В аналитически простом виде выражения для оценок параметров по цензурированным выборкам получаются лишь в некоторых частных случаях. Более перспективно определение оценок численными методами, для чего, вообще говоря, нет принципиальных трудностей.

При условии, что соответствующая параметрическая модель хорошо описывает закон распределения наблюдаемой случайной величины, можно достаточно точно оценивать параметры закона даже при очень сильно цензурированных выборках.

Идентифицировать параметрическую модель по малой выборке чрезвычайно сложно, так как можно указать (подобрать) множество моделей, одинаково хорошо описывающих выборочные данные с позиций различных критериев согласия. Особенности генеральной совокупности более четко проявляются с ростом объема выборки. Поэтому очевидно, что при больших объемах выборки можно достаточно точно находить оценки при большей степени цензурирования.

Цензурирование справа и слева для несимметричных законов может быть связано с различными потерями в количестве информации Фишера. Чем больше потери, тем больше асимптотическая дисперсия вычисляемых оценок. Величина потерь адекватно отражается на точности оценивания параметров.

При сильном цензурировании мы оцениваем параметры закона по левому или правому “хвосту” эмпирической функции распределения. И в этом случае на оценках в большей степени могут сказываться имеющиеся в выборке случайные отклонения от предполагаемого закона. Особенно резко это будет проявляться при малых объемах выборок. Поэтому естественно, что более предпочтительным является использование робастных оценок, в том числе соответствующих *L*- и *MD*-оценок.

Литература

1. Тихов М.С. Оценивание показателей качества по неполным выборка // Надежность и контроль качества. 1996. № 11. С. 16-24.
2. Лемешко Б.Ю. Статистический анализ одномерных наблюдений случайных величин: Программная система. - Новосибирск: Изд-во НГТУ, 1995. - 125 с.