УДК 519.2

М.С.Тихов
(Горьковский университет)

СТАТИСТИЧЕСКИЙ АНАЛИЗ ПО ГРУППИРОВАННЫМ И ЦЕНЗУРИРОВАННЫМ НАБЛЮДЕНИЯМ

I. Во многих экспериментальных процессах или процессах сбора данных получают не значение изучаемой случайной величины, которая ненаблюдаема , а значение менее информативной случайной величины. Потеря информации может происходить, например, вследствие группировки или цензурирования выборочных данных [1] , [2] . Для рассматривавшейся нами ситуации группированные данные возникают вследствие принятой методики наблюдений, которая позволяет зафиксировать попадание случайной величины в заранее данные интервалы. Цензурированные наблюдения возникают из-за того, что принятая методика не позволяет получить данные за некоторой границей. Поэтому встает вопрос о возможности наилучшего использования имеющейся информации. В настоящей заметке в основном рассматриваются некоторые вопросы, связанные с задачей оценки параметра неизвестного распределения по группированным и цензурированным данным.

Статистический контроль качества основывается часто на статистиках, являющихся состоятельными оценками параметров. Известно [1], что критерии, основанные на эффективных оценках, имеют максимальную асимптотическую относительную эффективность. Поэтому изучение качеств критериев сводится к изучению качеств оценок, что еще раз подтверждает необходимость изучения качества оценок параметра, получаемых тем или иным путем.

2. Пусть $\mathcal{F}(x,\theta)$ - функция распределения (Ф.Р.) изучаемой случайной величины (с.в.) X, $\theta\in\Theta$ - интервал в R_1, $-\infty=u_0<u_1<\ldots<u_{s-1}<\infty=u_s$ - разбиение области значений с.в. X ,

$$\pi_j(\theta)=\mathcal{F}(u_j,\theta)-\mathcal{F}(u_{j-1},\theta)\,,\ j=1,\ldots,s\ -$$

вероятности того, что наблюденное значение величины X попадает в интервалы $(-\infty\, u_1]$, $(u_1,u_2]$, ..., $(u_{s-1}\,,\,\infty)$, $\nu_j=n_j/n$, $j=1,\ldots,s$ - наблюденные частоты, соответствующие выбранным интервалам группировки (выборка повторная).

Задавшись неотрицательной измеримой функцией $\varphi(y)$, $y\in R_1$, рассмотрим оценку параметра θ, определенную равенством

$$\hat{\theta}_n = \int_{\Theta} y p_n(y) dy \,, \tag{I}$$

где

$$p_n(y) = \frac{\varphi(y) \prod_{j=1}^{s} \pi_j^{n_j}(y)}{\int_{\Theta} \prod_{j=1}^{s} \pi_j^{n_j}(y)\varphi(y)dy} \,, \quad y \in \Theta \,. \tag{2}$$

Если $\varphi(y)$ – плотность распределения вероятностей на Θ, то функцию $p_n(y)$ и оценку $\hat{\Theta}_n$ естественно трактовать как апостериорную плотность и соответственно апостериорное среднее с.в. θ, имеющей априорное распределение с плотностью $\varphi(y)$. Мы сохраним название апостериорной плотности и байесовской оценки за правыми частями равенств (I) и (2), хотя в наших рассмотрениях параметр θ не обязательно есть с.в., а функция $\varphi(y)$ не обязана быть плотностью распределения вероятностей.

Сделаем в интервале (I) замену $y = \theta_0 + \frac{z}{\sqrt{n}}$, где θ_0 – истинное значение параметра.
Тогда

$$\sqrt{n}\,(\hat{\theta}_n - \theta_0) = \frac{\int_{R_1} z \prod_{j=1}^{s} \pi_j^{n_j}(\theta_0 + \frac{z}{\sqrt{n}})\varphi(\theta_0 + \frac{z}{\sqrt{n}})dz}{\int_{R_1} \prod_{j=1}^{s} \pi_j^{n_j}(\theta_0 + \frac{z}{\sqrt{n}})\varphi(\theta_0 + \frac{z}{\sqrt{n}})dz} \tag{3}$$

(полагая всегда $\varphi(y) = 0$ для $y \in R_1 \setminus \Theta$, можно считать $\Theta = R_1$).
Определим функцию

$$Y_n(z) = \sum_{j=1}^{s} n_j \ln \frac{\pi_j(\theta_0 + \frac{z}{\sqrt{n}})}{\pi_j(\theta_0)} \tag{4}$$

и будем трактовать ее как случайную функцию z. Оказывается, что при определенных условиях, наложенных на распределение с.в. X, конечномерные распределения $Y_n(z)$ сходятся к конечномерным распределениям некоторой предельной функции $Y(z)$.

Из соотношений (3) и (4) имеем

$$\sqrt{n}\,(\hat{\theta}_n - \theta_0) = \frac{\int_{R_1} z \exp\{Y_n(z)\}\varphi(\theta_0 + \frac{z}{\sqrt{n}})dz}{\int_{R_1} \exp\{Y_n(z)\}\varphi(\theta_0 + \frac{z}{\sqrt{n}})dz} \,. \tag{5}$$

Ниже будет доказано, что при $n \to \infty$ возможны предельные переходы под знаками интегралов. Если $\varphi(\theta)$ - непрерывна в θ_0 и положительна, мы найдем из формулы (5), что предельное распределение $\sqrt{n}(\hat{\theta}_n - \theta_0)$ не зависит от φ и совпадает с распределением величины

$$\frac{\int\limits_{R_1} z\, exp\{Y(z)\}\, dz}{\int\limits_{R_1} exp\{Y(z)\}\, dz} .$$

Оказывается, что при сделанных ниже предположениях процесс $Y(z)$ имеет вид

$$Y(z) = \sqrt{\mathcal{J}_S(\theta_0)}\, z\xi - \frac{z^2}{2}\mathcal{J}_S(\theta_0) ,$$

где

$$\mathcal{J}_S(\theta) = \sum_{j=1}^{S} \frac{1}{\mathfrak{X}_j(\theta)} \left(\frac{\partial \mathfrak{X}_j(\theta)}{\partial \theta}\right)^2 \quad \text{- количество}$$

информации в группированных наблюдениях, а ξ - нормальная с.в. с параметрами (0,I).

Введем следующие условия:

(I.I) Параметрическое множество Θ есть интервал R_1 .

(I.2) Если $\theta \neq \beta$, то $\mathfrak{X}_j(\theta) \neq \mathfrak{X}_j(\beta)$ хотя бы при одном j .

(I.3) Функции $\mathfrak{X}_j(\theta)$ имеют производные первого порядка по θ .

(I.4) Информационное количество $\mathcal{J}_S(\theta_0) \neq 0$ в точке θ_0 .

(П.I) Функция $\varphi(\theta)$ - измеримая функция $\theta \in R_1$ равна нулю вне множества Θ^c - замыкания множества Θ .

(П.2) Функция $\varphi(\theta)$ непрерывна в окрестности точки θ_0 и $\varphi(\theta_0) \neq 0$.

3. Займемся исследованием процесса $Y_n(z)$. Основной результат содержится в следующей теореме.

Т е о р е м а I. Конечномерные распределения процесса $Y_n(z)$ сходятся к конечномерным распределениям процесса $Y(z) = z\sqrt{\mathcal{J}_S(\theta_0)}\,\xi - \frac{z^2}{2}\mathcal{J}_S(\theta_0)$ где ξ - нормальная с.в. с параметрами (0,I) и

$$\mathcal{J}_S(\theta_0) = \sum_{j=1}^{S} \frac{1}{\mathfrak{X}_j(\theta_0)} \left(\frac{\partial \mathfrak{X}_j(\theta_0)}{\partial \theta_0}\right)^2 .$$

Д о к а з а т е л ь с т в о. Достаточно доказать два следующих утверждения:

А) Распределения с.в. $Y_n(z)$ сходятся к распределению с.в.

$$z\sqrt{\mathcal{J}_s(\theta_0)}\,\xi - \frac{1}{2} z^2 \mathcal{J}_s(\theta_0)\ ;$$

в) При $n \to \infty$ разность

$$(Y_n(z_2)/z_2 + \frac{1}{2} z_2 \mathcal{J}_s(\theta_0)) - (Y_n(z_1)/z_1 + \frac{1}{2} z_1 \mathcal{J}_s(\theta_0))$$

сходится по вероятности к нулю для любых $z_1, z_2 \in \Theta$. Доказательству теоремы предпошлем ряд лемм.

Л е м м а I. При условиях I-П для любого z

$$\lim_{n\to\infty} \frac{n}{z^2} \sum_{j=1}^{s} \left(\sqrt{\mathfrak{X}_j(\theta_0 + \frac{z}{\sqrt{n}})} - \sqrt{\mathfrak{X}_j(\theta_0)}\right)^2 = \frac{1}{4} \mathcal{J}_s(\theta_0). \tag{6}$$

Д о к а з а т е л ь с т в о. Заметим, что

$$\lim_{n\to\infty} \frac{n}{z^2} \sum_{j=1}^{s} \left(\sqrt{\mathfrak{X}_j(\theta_0 + \frac{z}{\sqrt{n}})} - \sqrt{\mathfrak{X}_j(\theta_0)}\right)^2 =$$

$$= \sum_{j=1}^{s} \lim_{n\to\infty} \left(\frac{\sqrt{\mathfrak{X}_j(\theta_0 + z/\sqrt{n})} - \sqrt{\mathfrak{X}_j(\theta_0)}}{z/\sqrt{n}}\right)^2 =$$

$$= \sum_{j=1}^{s} \frac{1}{4} \left(\frac{\partial \mathfrak{X}_j(\theta_0)}{\partial \theta_0}\right)^2 \cdot \frac{1}{\mathfrak{X}_j(\theta_0)} = \frac{1}{4} \mathcal{J}_s(\theta_0)\ .$$

Л е м м а 2. Для любой точки $z \in \Theta$ при $n \to \infty$

$$\sum_{j=1}^{s} \sqrt{\mathfrak{X}_j(\theta_0 + \frac{z}{\sqrt{n}}) \mathfrak{X}_j(\theta_0)} = 1 - \frac{1}{8} \mathcal{J}_s(\theta_0) \frac{z^2}{n} + o\left(\frac{1}{n}\right)\ . \tag{7}$$

Д о к а з а т е л ь с т в о. По лемме I

$$\sum_{j=1}^{s} \left(\sqrt{\mathfrak{X}_j(\theta_0 + z/\sqrt{n})} - \sqrt{\mathfrak{X}_j(\theta_0)}\right)^2 = 2 - 2 \sum_{j=1}^{s} \sqrt{\mathfrak{X}_j(\theta_0 + z/\sqrt{n}) \mathfrak{X}_j(\theta_0)} =$$

$$= \frac{1}{4} \mathcal{J}_s(\theta_0) \frac{z^2}{n} + o\left(\frac{1}{n}\right)$$, что эквивалентно (7).

Л е м м а 3. Для $\forall z \in \Theta$ при $n \to \infty$

$$\sum_{j=1}^{s} \pi_j(\theta_0) \ln \frac{\pi_j(\theta_0 + \frac{z}{\sqrt{n}})}{\pi_j(\theta_0)} = -\frac{z^2}{2n} \mathcal{J}_s(\theta_0) + o\left(\frac{1}{n}\right) . \tag{8}$$

Д о к а з а т е л ь с т в о. Разлагая

$$\ln \pi_j(\theta_0 + z/\sqrt{n})/\pi_j(\theta_0) = 2 \ln \sqrt{\pi_j(\theta_0 + z/\sqrt{n})/\pi_j(\theta_0)}$$

по степеням $\left(\sqrt{\pi_j(\theta_0 + z/\sqrt{n})} - \sqrt{\pi_j(\theta_0)}\right)/\sqrt{\pi_j(\theta_0)}$, найдем

$$\sum_{j=1}^{s} \pi_j(\theta_0) \ln \pi_j(\theta_0 + z\sqrt{n})/\pi_j(\theta_0) = 2 \sum_{j=1}^{s} \pi_j(\theta_0) \left[\left(\sqrt{\pi_j(\theta_0 + z/\sqrt{n})} - \sqrt{\pi_j(\theta_0)}\right) :\right.$$

$$\left. : \pi_j(\theta_0)\right] - \sum_{j=1}^{s} \pi_j(\theta_0) \left[\left(\sqrt{\pi_j(\theta_0 + z\sqrt{n})} - \sqrt{\pi_j(\theta_0)}\right)/\sqrt{\pi_j(\theta_0)}\right]^2 + o\left(\frac{1}{n}\right) =$$

$$= 2 \sum_{j=1}^{s} \sqrt{\pi_j(\theta_0 + \frac{z}{\sqrt{n}}) \pi_j(\theta_0)} - 2 - \frac{1}{4} \mathcal{J}_s(\theta_0) \frac{z^2}{n} + o\left(\frac{1}{n}\right) = -\frac{1}{2} \mathcal{J}_s(\theta_0) \frac{z^2}{n} + o\left(\frac{1}{n}\right) .$$

Л е м м а 4. Для любых $z_1, z_2 \in \Theta$ при $n \to \infty$

$$\sum_{j=1}^{s} \pi_j(\theta_0) \ln \frac{\pi_j(\theta_0 + z_1/\sqrt{n})}{\pi_j(\theta_0)} \ln \frac{\pi_j(\theta_0 + z_2/\sqrt{n})}{\pi_j(\theta_0)} = \mathcal{J}_s(\theta_0) \frac{z_1 z_2}{n} + o\left(\frac{1}{n}\right) , \tag{9}$$

в частности,

$$\sum_{j=1}^{s} \pi_j(\theta_0) \ln^2 \frac{\pi_j(\theta_0 + z/\sqrt{n})}{\pi_j(\theta_0)} = \mathcal{J}_s(\theta_0) \frac{z^2}{n} + o\left(\frac{1}{n}\right) .$$

Д о к а з а т е л ь с т в о. Равенство (9) доказывается так же как и (8):

$$\sum_{j=1}^{s} \pi_j(\theta_0) \ln \frac{\pi_j(\theta_0 + \frac{z_1}{\sqrt{n}})}{\pi_j(\theta_0)} \ln \frac{\pi_j(\theta_0 + \frac{z_2}{\sqrt{n}})}{\pi_j(\theta_0)} = 4 \sum_{j=1}^{s} \left(\sqrt{\pi_j(\theta_0 + \frac{z_1}{\sqrt{n}})} - \sqrt{\pi_j(\theta_0)}\right) .$$

$$\left(\sqrt{\pi_j(\theta_0 + \frac{z_2}{\sqrt{n}})} - \sqrt{\pi_j(\theta_0)}\right) + o\left(\frac{1}{n}\right) = \mathcal{J}_s(\theta_0) \frac{z_1 z_2}{n} + o\left(\frac{1}{n}\right) .$$

Теперь мы в состоянии доказать утверждение А). Справедливы равенства

$$Y_n(z)+\frac{1}{2}z^2\mathcal{J}_s(\theta_0)=\sum_{j=1}^{s}n_j\ln\pi_j(\theta_0+z/\sqrt{n})/\pi_j(\theta_0)+\frac{1}{2}z^2\mathcal{J}_s(\theta_0)=$$

$$=n\sum_{j=1}^{s}(\nu_j-\pi_j(\theta_0))\ln\pi_j(\theta_0+\frac{z}{\sqrt{n}})/\pi_j(\theta_0)+n\sum_{j=1}^{s}\pi_j(\theta_0)\ln\pi_j(\theta_0+z/\sqrt{n})/\pi_j(\theta_0)+$$

$$+\frac{1}{2}z^2\mathcal{J}_s(\theta_0)=n\sum_{j=1}^{s}(\nu_j-\pi_j(\theta_0))\ln\pi_j(\theta_0+z/\sqrt{n})/\pi_j(\theta_0)+o(1)=$$

$$=\sum_{j=1}^{s}z\sqrt{n}(\nu_j-\pi_j(\theta_0))\sqrt{n}/z\;\ln\pi_j(\theta_0+z/\sqrt{n})/\pi_j(\theta_0)+o(1),$$

так как $\sqrt{n}(\nu_j-\pi_j(\theta_0))$ имеет нормальное предельное распределение, а $\frac{\sqrt{n}}{z}\ln\frac{\pi_j(\theta_0+z/\sqrt{n})}{\pi_j(\theta_0)}\to\frac{1}{\pi_j(\theta_0)}\frac{\partial}{\partial\theta_0}\pi_j(\theta)$. Действительно, пусть

$$\nu'=\left(\frac{n_1-n\pi_1}{\sqrt{n\pi_1}}\right),\dots,\left(\frac{n_s-n\pi_s}{\sqrt{n\pi_s}}\right),\ \varphi'=(\sqrt{\pi_1},\dots,\sqrt{\pi_s}),\ b'=(b_1,\dots,b_s).$$

Тогда ([4] с. 332-333) линейная форма

$$b'\nu=\sum_{j=1}^{s}b_j\ \frac{n_j-n\pi_j}{\sqrt{n\pi_j}}$$

асимптотически распределена нормально с нулевым математическим ожиданием и дисперсией $bb'-(b'\varphi)^2$. В нашем случае

$$b_j=z\frac{\partial}{\partial\theta_0}\pi_j(\theta_0)/\sqrt{\pi_j(\theta_0)},\quad b'b=\sum_{j=1}^{s}b_j^2=\mathcal{J}_s(\theta_0)z^2,$$

$$b'\varphi=\sum_{j=1}^{s}b_j\sqrt{\pi_j(\theta_0)}=\sum_{j=1}^{s}z\,\frac{\frac{\partial}{\partial\theta_0}\pi_j(\theta_0)}{\sqrt{\pi_j(\theta_0)}}\cdot\sqrt{\pi_j(\theta_0)}=z\frac{\partial}{\partial\theta_0}\sum_{j=1}^{s}\pi_j(\theta_0)=0.$$

Поэтому $\quad b'b-(b'\varphi)^2=\mathcal{J}_s(\theta_0)z^2.$

Из предыдущих рассуждений выводим, что $Y_n(z)+\frac{z^2}{2}\mathcal{J}_s(\theta_0)$ асимптотически распределена как $z\sqrt{\mathcal{J}_s(\theta_0)}\,\xi$, где ξ - нормальная с.в. с параметрами (0,I), т.е.

$$Y_n(z)\overset{d}{=}-\frac{1}{2}z^2\mathcal{J}_s(\theta_0)+z\sqrt{\mathcal{J}_s(\theta_0)}\,\xi\,.$$

Докажем теперь утверждение В).

По лемме 3

$$E_{\theta_0}\left\{\left(\frac{Y_n(z_2)}{z_2}+\frac{z_2\mathcal{J}_s(\theta_0)}{2}\right)-\left(\frac{Y_n(z_1)}{z_1}+\frac{z_1\mathcal{J}_s(\theta_0)}{2}\right)\right\}=o(1)$$

при $n\to\infty$, поэтому достаточно проверить, что

$$E_{\theta_0}\left[(Y_n(z_2)/z_2+\frac{1}{2}z_2\mathcal{J}_s(\theta_0))-(Y_n(z_1)/z_1+\frac{1}{2}z_1\mathcal{J}_s(\theta_0)\right]^2\to 0\,.$$

Последнее соотношение следует из леммы 4.

4. Докажем предельную теорему для разностей $\sqrt{n}(\hat{\theta}_n-\theta_0)$.

Теорема 2. Пусть выполнены условия I-II. Тогда предельное распределение разностей $\sqrt{n}(\hat{\theta}_n-\theta_0)$ нормально с параметрами $\left(0,\ \frac{1}{\mathcal{J}_s(\theta_0)}\right)$.

Доказательство. В п.2 было показано, что

$$\sqrt{n}(\hat{\theta}_n-\theta_0)=\frac{\int\limits_{R_1}\theta\exp\{Y_n(\theta)\}\varphi(\theta_0+\theta/\sqrt{n})d\theta}{\int\limits_{R_1}\exp\{Y_n(\theta)\}\varphi(\theta_0+\theta/\sqrt{n})d\theta}\,. \qquad (10)$$

По теореме I конечномерные распределения процесса $Y_n(\theta)$ сходятся к конечномерным распределениям процесса

$$Y(\theta)=\xi\sqrt{\mathcal{J}_s(\theta_0)}\,\theta-\frac{1}{2}\mathcal{J}_s(\theta_0)\theta^2\,.$$

Переходя в правой части (10) формально к пределу под знаком интеграла, мы придем к выражению

$$\frac{\int\limits_{R_1}\theta\exp\{Y(\theta)\}d\theta}{\int\limits_{R_1}\exp\{Y(\theta)\}d\theta}=\frac{\xi}{\sqrt{\mathcal{J}_s(\theta_0)}}$$

и остается только обосновать эти формальные выкладки.

Фиксируем некоторое положительное число A . При любых t_1,t_2 функционал

$$f_{t_1t_2}(x)=t_1\int\limits_{-A}^{A}\theta x(\theta)d\theta+t_2\int\limits_{-A}^{A}x(\theta)d\theta$$

непрерывен в $C_0(-\infty,\infty)$.

Рассмотрим произвольные z_1,z_2 такие, что $\theta_0+z_1,\theta_0+z_2\in[-A,A]$, и проверим, что для $n\geqslant n_0$

$$E_{\theta_0}[Y_n(z_1)-Y_n(z_2)]^2\leq H(z_1-z_2)^2,$$

где постоянная $H>0$. Действительно,

$$E_{\theta_0}[Y_n(z_1)-Y_n(z_2)]^2 = E_{\theta_0}\Big[\sum_{j=1}^{s} n_j\Big(\ln\frac{\pi_j(\theta_0+z_1/\sqrt{n})}{\pi_j(\theta_0)} -$$

$$-\ln\frac{\pi_j(\theta_0+z_2/\sqrt{n})}{\pi_j(\theta_0)}\Big)\Big]^2 \le n\sum_{j=1}^{s}\pi_j(\theta_0)\Big[\ln\pi_j\Big(\theta_0+\frac{z_1}{\sqrt{n}}\Big) -$$

$$-\ln\pi_j\Big(\theta_0+\frac{z_2}{\sqrt{n}}\Big)\Big]^2 .$$

Поскольку существуют $\frac{\partial}{\partial\theta}\ln\pi_j(\theta)$ для $\theta_0-\frac{A}{\sqrt{n}}\le\theta\le\theta_0+\frac{A}{\sqrt{n}}$, то существует n_0 и H такие, что для $n\geqslant n_0$

$$\Big[\ln\pi_j\Big(\theta_0+\frac{z_1}{\sqrt{n}}\Big)-\ln\pi_j\Big(\theta_0+\frac{z_2}{\sqrt{n}}\Big)\Big]^2\le H\cdot\frac{(z_1-z_2)^2}{n} ,$$

следовательно,

$$E_{\theta_0}[Y_n(z_1)-Y_n(z_2)]^2\le H(z_1-z_2)^2.$$

В силу теоремы 2([9], с. 583), непрерывности процесса $Y_n(z)$ и непрерывности функции $\varphi(\theta)$ в точке θ_0 следует, что распределения случайных векторов $\Big(\int_{-A}^{A}\theta\exp\{Y_n(\theta)\}\varphi(\theta_0+\theta/\sqrt{n})d\theta,$ $\int_{-A}^{A}\exp\{Y_n(\theta)\}\varphi(\theta_0+\theta/\sqrt{n})d\theta$ сходятся к распределению вектора

$$\Big(\int_{-A}^{A}\theta\exp\{Y(\theta)\}\varphi(\theta_0)d\theta,\ \int_{-A}^{A}\exp\{Y(\theta)\}\varphi(\theta_0)d\theta\Big).$$

Далее,

$$P\Big\{\int_{|\theta|>A}(|\theta|+1)e^{Y_n(\theta)}\varphi\Big(\theta_0+\frac{\theta}{\sqrt{n}}\Big)d\theta>\frac{1}{A^N}\Big\}\le\sum_{|k|>A}P\Big\{\int_{k}^{k+1}(|\theta|+1)\cdot$$

$$\cdot e^{Y_n(\theta)}\varphi\Big(\theta_0+\frac{\theta}{\sqrt{n}}\Big)d\theta>\frac{1}{k^N}\Big\}\le\sum_{|k|>A}P\Big\{\max_{\theta\in[k,k+1]}e^{Y_n(\theta)}>k^{-(N+\delta+2)}\Big\}\le\frac{C_N}{A^N},$$

если только числа A и n выбраны достаточно большими (здесь снова, как и выше, использована лемма Прохорова [10]). Для процесса $e^{Y(\theta)}$ по любому $\varepsilon>0$ существует $\delta>0$ так, чтобы

$$P\Big\{\int_{|\theta|>A}\theta e^{Y(\theta)}\varphi(\theta_0)d\theta>\delta\Big\}<\varepsilon,\ P\Big\{\int_{|\theta|>A}e^{Y(\theta)}\varphi(\theta_0)d\theta>\delta\Big\}<\varepsilon.$$

Но тогда будут выполнены аналогичные неравенства для процесса $e^{y_n(\theta)}$ при достаточно больших n.

Таким образом, каково бы ни было $\varepsilon > 0$ при больших n,

$$\sqrt{n}\,(\hat{\theta}_n - \theta_0) = \frac{\int_{-A}^{A} \theta e^{y_n(\theta)}\, d\theta}{\int_{-A}^{A} e^{y_n(\theta)} d\theta} (1 + r_n(A)),$$

где $P\{|r_n(A)| > \delta\} < \varepsilon$, что доказывает наши формальные выкладки.

5. Заметим, что оценки максимального правдоподобия укладываются в нашу схему. Действительно, рассмотрим функцию потерь $W(d-\theta)$. Оценку t_n параметра θ назовем байесовской относительно функции потерь $W(d-\theta)$ и апостериорной плотности $\varphi(\theta)$, если

$$\int_{R_1} W(t_n-\theta) p_n(\theta) d\theta = \inf_d \int_{R_1} W(d-\theta) p_n(\theta) d\theta .$$

Тогда оценки максимального правдоподобия можно считать байесовскими по отношению к функции потерь $W(d-\theta) = -\delta(d-\theta)$ и $\varphi(\theta) = 1$, а затем нетрудно провести обоснование предельного перехода под знаком интеграла, откуда получаются известные результаты [4]. Если $W(d-\theta) = -\delta(d-\theta)$ при произвольной $\varphi(\theta)$, то предыдущее соотношение определяет оценки максимальной апостериорной плотности.

6. Если $\tilde{\theta}_n$ - оценка параметра θ по группированным наблюдениям такая, что $E_\theta \tilde{\theta}_n = \theta + B(\theta)$ существует $B_\theta'(\theta)$, то при условии, что можно дифференцировать под знаком интеграла, справедливо неравенство Крамера-Рао, которое в нашем случае имеет вид

$$E_\theta (\tilde{\theta}_n - \theta)^2 \geqslant B^2(\theta) + \frac{(1 + B_\theta'(\theta))}{n\, \mathcal{I}_s(\theta)} . \tag{II}$$

Если τ - (марковский) момент остановки, $E_\theta \tau < \infty$, то справедливо также неравенство Волфовитца

$$E_\theta (\tilde{\theta}_\tau - \theta)^2 \geqslant B^2(\theta) + \frac{(1 + B_\theta'(\theta))^2}{(E_\theta \tau)\, \mathcal{I}_s(\theta)}, \tag{I2}$$

где $\tilde{\theta}_\tau = \tilde{\theta}_n$ на множестве $(\tau = n)$. Если рассмотреть для примера показательное распределение с плотностью $p(x,\theta) = \frac{1}{\theta} e^{-x/\theta}$,

$x>0$, то минимизацией $1/\mathcal{J}_s(\theta_0)$ можно получить оптимальное разбиение как в работе [5], полученное там из других соображений.

Обозначим через $I(\theta)=E_\theta\left(\frac{\partial \ln p(x,\theta)}{\partial\theta}\right)^2$ информационное количество Фишера, $p(x,\theta)$ - плотность ф.р. $\mathcal{F}(x,\theta)$; предположим, что $I(\theta)<\infty$. Покажем, что $\mathcal{J}_s(\theta)\leq I(\theta)$, последнее говорит о частичной потере информации при группировке. Для этого достаточно показать, что

$$\frac{1}{\pi_j}\left(\frac{\partial\pi_j}{\partial\theta}\right)^2\leq\int_{u_{j-1}}^{u_j}\left(\frac{\partial p(u,\theta)}{\partial\theta}\right)^2\frac{1}{p(u,\theta)}\,du\,. \tag{13}$$

предполагая возможность дифференцирования под знаком интеграла, будем иметь следующее неравенство, эквивалентное (13):

$$\left[\int_{u_{j-1}}^{u_j}\frac{\partial p(u,\theta)}{\partial\theta}du\right]^2\leq\left(\int_{u_{j-1}}^{u_j}\left[\frac{\partial p(u,\theta)}{\partial\theta}\right]^2\frac{1}{p(u,\theta)}\,du\right)\left(\int_{u_{j-1}}^{u_j}p(u,\theta)du\right)\cdot$$

Заметим, что последнее неравенство эквивалентно, в свою очередь,

$$\left(\int_{u_{j-1}}^{u_j}\frac{p'_\theta}{\sqrt{p}}\sqrt{p}\,du\right)^2\leq\left(\int_{u_{j-1}}^{u_j}\left(\frac{p'_\theta}{\sqrt{p}}\right)^2du\right)\left(\int_{u_{j-1}}^{u_j}(\sqrt{p})^2du\right),$$

а последнее неравенство есть хорошо известное неравенство Коши-Шварца.

Пусть теперь $u_1=a$, $u_{s-1}=b$, .

$\Delta_j=(u_j-u_{j-1})$, $\max\limits_{1\leq j\leq s-1}\Delta_j\to 0$ при $s\to\infty$.

Условие (Ш.I).

Предположим также, что функция $\frac{\partial^2}{\partial u\partial\theta}\mathcal{F}(u,\theta)$ непрерывна по u .

Рассмотрим снова отдельное слагаемое в $\mathcal{J}_s(\theta)$. По определению производной имеем

$$\frac{1}{\pi_j}\left(\frac{\partial\pi_j}{\partial\theta}\right)^2=\frac{\left[\frac{\partial}{\partial\theta}\mathcal{F}(u_j,\theta)-\frac{\partial}{\partial\theta}\mathcal{F}(u_{j-1},\theta)\right]^2}{\mathcal{F}(u_j,\theta)-\mathcal{F}(u_{j-1},\theta)}=$$

$$=\frac{\Delta_j^2\left[\frac{\partial}{\partial\theta}p(u_j,\theta)+o(1)\right]^2}{\Delta_j\left[p(u_j,\theta)+o(1)\right]}=\left[\left(\frac{\partial}{\partial\theta}\ln p(u_j,\theta)\right)^2p(u_j,\theta)+o(1)\right]\Delta_j$$

при $\max\limits_{1\leq j\leq s-1}\Delta_j\to 0$, $s\to\infty$.

Отсюда следует, что

$$\mathcal{J}_s(\theta) \to \frac{\left[\frac{\partial}{\partial\theta}\mathcal{F}(a,\theta)\right]^2}{\mathcal{F}(a,\theta)} + \int_a^b \left[\frac{\partial \ln p(u,\theta)}{\partial\theta}\right]^2 p(u,\theta)du + $$

$$+ \frac{\left[\frac{\partial}{\partial\theta}\mathcal{F}(b,\theta)\right]^2}{1-\mathcal{F}(b,\theta)} = I_{a,b}(\theta),$$

т.е.

$$D\{\hat{\theta}_n\} \sim [n\, I_{a,b}(\theta)]^{-1},$$

а распределение $\sqrt{n}\,(\hat{\theta}_n - \theta_0)$ – асимптотически нормальное с параметрами $(0,\ 1/I_{a,b}(\theta_0))$. Нетрудно показать, что неравенство (II) преобразуется в неравенство вида

$$E_\theta(\tilde{\theta}_n - \theta)^2 \geqslant B^2(\theta) + \frac{(1+B'_\theta(\theta))^2}{n\, I_{a,b}(\theta)}.$$

Соответствующее неравенство Волфовитца принимает вид

$$E_\theta(\tilde{\theta}_\tau - \theta)^2 \geqslant B^2(\theta) + \frac{(1+B'_\theta(\theta))^2}{(E_\theta\tau)\, I_{a,b}(\theta)}.$$

Здесь уже $\tilde{\theta}_n$ – оценка по наблюдениям, дважды цензурированным по типу I в точках a и b. Отсюда, в частности, следует, что если не наблюдается r_1 наименьших и r_2 наибольших элементов выборки, то для этого цензурирования функция правдоподобия

$$L[\mathcal{X}(\theta)] = \frac{n!}{n_1!\dots n_s!}\,\pi_1^{n_1}(\theta)\,\pi_2^{n_2}(\theta)\dots\pi_s^{n_s}(\theta)$$

полиномиального распределения преобразуется в функцию правдоподобия

$$L_x(\theta) = \left[\int_{-\infty}^{a} p(x,\theta)dx\right]^{r_1} \prod_{i=1}^{n-r_1-r_2} p(x_i,\theta) \left[\int_B^{\infty} p(x,\theta)dx\right]^{r_2},$$

а оценки, полученные по методу МП, состоятельны и асимптотически нормальны (и асимптотически эффективны). Напомним, что основное ограничение состоит в требовании существования первой производной от плотности по параметру (сравни условия [7]). Как и выше, $\hat{\theta}_n$ – обобщенная байесовская оценка, для последовательности $\sqrt{n}\,(\hat{\theta}_n - \theta_0)$

предельное распределение - нормальное. АОЭ цензурированной оценки (соответственно и критерия, построенного на основе этой оценки) по сравнению с обычной оценкой метода МП (соответственно критерия, основанного на оценке метода МП по полной выборке)

$$АОЭ = \frac{[n\,I(\theta)]^{-1}}{D\{\hat{\theta}_n\}} = \frac{I_{a,b}(\theta)}{I(\theta)} \quad .$$

Пусть теперь $a \to -\infty$, $b \to \infty$. Покажем, что

$$I_{a,b}(\theta) \to I(\theta).$$

Действительно,

$$\frac{\left[\frac{\partial}{\partial\theta}\mathcal{F}(a,\theta)\right]^2}{\mathcal{F}(a,\theta)} = \frac{\left[\int_{-\infty}^{a} p'_\theta(u,\theta)du\right]^2}{\int_{-\infty}^{a} p(u,\theta)du} =$$

$$= \frac{\left[\int_{-\infty}^{a} \frac{p'_\theta}{\sqrt{p}}\sqrt{p}\,du\right]^2}{\int_{-\infty}^{a} p(u,\theta)du} \le \frac{\left(\int_{-\infty}^{a} \frac{p'^2_\theta}{p}du\right)\int_{-\infty}^{a} p\,du}{\int_{-\infty}^{a} p\,du} =$$

$$= \int_{-\infty}^{a} \frac{p'^2_\theta}{p}du \to 0 \qquad , \text{ если } \quad I(\theta) < \infty .$$

Заметим, что отношение полагаем равным нулю, если числитель равен нулю. Аналогичные рассуждения можно провести и для точки b . Последнее и доказывает требуемое.

Если $a \to -\infty$, а b - фиксировано, что соответствует цензурированию типа I на уровне b [1], то при соответствующих условиях

$$\mathcal{I}_b(\theta) \to \int_{-\infty}^{b}\left(\frac{\partial \ln p}{\partial\theta}\right)^2 p\,du + \frac{\left[\frac{\partial}{\partial\theta}\mathcal{F}(b,\theta)\right]}{1-\mathcal{F}(b,\theta)} = I_b(\theta),$$

$$D\{\hat{\theta}_n\} \sim \left\{n\left[\int_{-\infty}^{b}\left(\frac{\partial \ln p}{\partial\theta}\right)^2 p\,du + \frac{\left[\frac{\partial}{\partial\theta}\mathcal{F}(b,\theta)\right]^2}{1-\mathcal{F}(b,\theta)}\right]\right\}^{-1},$$

$$АОЭ = \frac{I_b(\theta)}{I(\theta)} \quad .$$

В частности, для показательного распределения

$$АОЭ = 1 - e^{-b/\theta} .$$

7. Усложним задачу оценивания по цензурированной выборке, которую представим следующей ситуацией.

Рассмотрим случай, когда n изделий, поставленных на испытания, разбиты на q групп. В первой группе k_1 изделий, во второй – k_2 и в последней – k_q ($k_1+\ldots+k_q$, $k_1,\ldots,k_q \geqslant 0$). Для каждой из этих групп производится цензурирование по типу I на уровне b_i для i – й группы (уровень b_i фиксируется до наблюдений). В этом случае функция правдоподобия

$$L_x(\theta)=\prod_{i=1}^{q}\Big[\Big(\prod_{j=1}^{k_i-r_i}p(x_j,\theta)\Big)\Big(\int_{b_i}^{\infty}p(x,\theta)dx\Big)^{r_i}\Big].$$

Введем обозначение

$$I_i(\theta)=I_{b_i}(\theta)=\Big\{\int_{-\infty}^{b_i}\Big(\frac{\partial \ln p}{\partial\theta}\Big)^2 p\,du+(1-\mathcal{F}(b,\theta))\Big[\frac{\partial \ln(1-\mathcal{F}(b_i,\theta))}{\partial\theta}\Big]^2.$$

Предположим, что $k_i/n \to \lambda$ при $n\to\infty$ (k_i – неслучайны). Обобщая предыдущие рассуждения на этот случаи, можно показать, что при условиях регулярности I–III асимптотическое распределение $\sqrt{n}(\hat\theta_n-\theta_0)$ – нормальное $N(0,\sum_{i=1}^{q}\lambda_i I_i^{-1}(\theta_0))$. Отсюда оценки $\hat\theta_n$, например, метода МП состоятельны и

$$AOЭ=\Big\{I(\theta)\Big[\sum_{i=1}^{q}\lambda_i I_i^{-1}(\theta)\Big]\Big\}^{-1}, \text{ где}$$

$$\lambda_i\geqslant 0,\ \sum_{i=1}^{q}\lambda_i=1,$$

по сравнению с оценками метода МП по полной выборке.

Можно рассмотреть, как и выше, и обобщенные байесовские оценки.

Предыдущую схему можно рассматривать как предельный случай оценивания параметра θ по группированным наблюдениям, когда для каждой из q групп интервалы группирования произведены по-разному. Пусть для i – й группы разбиение имеет вид $(-\infty,u_1^i],\ldots,(u_{s_i-1}^i,\infty)$, $\mathcal{P}_j^i(\theta)$, $j=1,\ldots,s_i$ – соответствующие вероятности $\mathcal{I}_s^i(\theta)=\sum_{1}^{s_i}\frac{1}{\mathcal{P}_j^i}\Big(\frac{\partial\mathcal{P}_j^i}{\partial\theta}\Big)^2$. Тогда асимптотическое распределение $\sqrt{n}(\hat\theta_n-\theta_0)$ – нормальное $N\big(0,\sum_{i=1}^{q}\lambda_i[\mathcal{I}_s^i(\theta_0)]^{-1}\big)$.

8. Рассмотрим теперь следующую схему наблюдений. Пусть $n+m$ изделий поставлено на контроль. Для каждого изделия проведен

случайный эксперимент с двумя исходами A и $\bar{A}$, вероятности этих исходов равны p и $q = 1 - p$.

Если в результате испытания происходит A, то наблюдается попадание в один из интервалов $(-\infty, u_1], \ldots, (u_{s-2}, u_{s-1}], (u_{s-1}, \infty)$. с вероятностями $\pi_1(\theta), \ldots, \pi_s(\theta)$ соответственно; $\nu_j = n_j / n$, $j = 1, \ldots, s$ – частоты.

Если произошло $\bar{A}$, то наблюдается попадание в один из интервалов $(-\infty, u_1], \ldots, (u_{s-2}, u_{s-1}], (u_{s-1}, u_s], (u_s, \infty)$ с вероятностями $\mu_1, \ldots, \mu_s, \mu_{s+1}$, где $\mu_j = \pi_j(\theta)$, $j = 1, \ldots, s-1$, $\mu_s + \mu_{s+1} = \pi_s(\theta)$, $\varkappa_j = m_j / m$ – частоты.

Пусть имеется повторная выборка объема $n + m = N$, N – фиксировано.

По аналогии с предыдущим составим обобщенную байесовскую оценку

$$\hat{\theta}_N = \frac{\int_{R_1} \theta \varphi(\theta) \prod_{j=1}^{s} \pi_j^{n_j}(\theta) \prod_{j=1}^{s+1} \mu_j^{m_j}(\theta) d\theta}{\int_{R_1} \varphi(\theta) \prod_{j=1}^{s} \pi_j^{n_j}(\theta) \prod_{j=1}^{s+1} \mu_j^{m_j}(\theta) d\theta},$$

$$\sqrt{N}(\hat{\theta}_n - \theta_0) = \frac{\int_{R_1} z \varphi(\theta_0 + z/\sqrt{N}) e^{Y_N(z)} dz}{\int_{R_1} \varphi(\theta_0 + z/\sqrt{N}) e^{Y_N(z)} dz}, \text{ где}$$

$$Y_N(z) = \sum_{j=1}^{s} n_j \ln \frac{\pi_j(\theta_0 + z/\sqrt{N})}{\pi_j(\theta_0)} + \sum_{j}^{s+1} m_j \ln \frac{\mu_j(\theta_0 + z/\sqrt{n})}{\mu_j(\theta_0)} =$$

$$= N \sum_{j=1}^{s} \left(\frac{n_j}{N} - p\pi_j(\theta_0)\right) \ln \frac{\pi_j(\theta_0 + z/\sqrt{N})}{\pi_j(\theta_0)} + N \sum_{j=1}^{s} p\pi_j(\theta_0) \ln \frac{\pi_j(\theta_0 + z/\sqrt{N})}{\pi_j(\theta_0)} +$$

$$+ N \sum_{j=1}^{s+1} \left(\frac{m_j}{N} - q\mu_j(\theta_0)\right) \ln \frac{\mu_j(\theta_0 + z/\sqrt{N})}{\mu_j(\theta_0)} + N \sum_{j=1}^{s+1} q\mu_j(\theta_0) \ln \frac{\mu_j(\theta_0 + z/\sqrt{N})}{\mu_j(\theta_0)}.$$

При фиксированном z $Y_N(z)$ асимптотически распределена как величина

$$z\left(p\sqrt{\mathcal{I}_s(\theta_0)} + q\sqrt{M_{s+1}(\theta_0)}\right)\zeta - \frac{z^2}{2}\left(p\mathcal{I}_s(\theta_0) + qM_{s+1}(\theta_0)\right),$$

где ζ распределена нормально $N(0,1)$, а

$$M_{s+1}(\theta_0) = \sum_{j=1}^{s+1} \frac{1}{\mu_j(\theta_0)} \left(\frac{\partial \mu_j(\theta_0)}{\partial \theta_0}\right)^2.$$

В таком случае выводим, что $\sqrt{N}\,(\hat{\theta}_N - \theta_0)$ распределена при $N\to\infty$ асимптотически нормально с параметрами $(0,[p\mathcal{J}_s(\theta_0)+qM_{s+1}(\theta_0)]^{-1})$. Если теперь снова уменьшать интервалы, т.е. рассмотреть случай, когда $\max\limits_{1\le j\le s-1}\Delta_j \to 0$ при $s\to\infty$, то получим, что соответствующие $\sqrt{N}(\hat{\theta}_N-\theta_0)$ асимптотически распределены нормально с параметрами $(0,[\int_{-\infty}^{\infty} I_y(\theta_0)g(y)dy]^{-1})$, где $g(y)$ – плотность распределения величины, полученной как предел простой случайной величины, когда рассматривается обобщение предыдущей схемы с двумя исходами.

Предельная схема описывается следующим образом.

Наблюдается величина $Z=\min(X,Y)$, где $p(x,\theta)$ – плотность распределения величины X, $g(y)$ – плотность распределения величины Y. Величины X и Y независимы. Иными словами

$$Z=\begin{cases} X & \text{, если } \{X<Y\}, \\ Y & \text{, если } \{X\geqslant Y\}. \end{cases}$$

По аналогии с классическим случаем можно рассмотреть функцию правдоподобия (ФП)

$$L_{x,y}(\theta)=\prod_{k=1}^{n} p(x_k,\theta)\prod_{i=1}^{m}(1-\mathcal{F}(y_i,\theta)) \tag{14}$$

и получить оценки метода максимального правдоподобия, развивая идеи п. 5. Функция правдоподобия $L_{x,y}(\theta)$ в дискретной схеме при фиксированном θ как функция (x,y) есть совместная вероятность появления событий $B_k=\{X=x_k\}\cap\{Y>x_k\}$ и $C_i=\{Y=y_i\}\cap\{X\geqslant y_i\}$. Переходя к непрерывному случаю, будем иметь

$$L_{x,y}(\theta)=\prod_{k=1}^{n} p(x_k,\theta)(1-G(x_k))\prod_{i=1}^{m} g(y_i)(1-\mathcal{F}(y_i,\theta)):$$

Поскольку $(1-G(x_k))$, где $G(x)=\int_{-\infty}^{x} g(u)du$, $g(y_i)$ не содержат параметра θ, то ФП можно рассматривать согласно равенству (14).

Предыдущие рассуждения говорят о том, что оценки по методу МП $L_{x,y}(\theta)$ состоятельны.

Асимптотическая дисперсия зависит от плотности $g(y)$, которая, как правило, неизвестна. Можно рассмотреть и обобщенные байесовские оценки.

Литература

I. Кендалл М.Дж., Стьюарт А. Статистические выводы и связи. М., Наука, 1973.

2. Закс Ш. Теория статистических выводов. М., Мир, 1975.

3. Ибрагимов И.А., Хасьминский Р.З. Асимптотическое поведение некоторых статистических оценок в гладком случае. - Теория вероятн. и ее примен., 1972, ч. I, ХУП. 3, 469-486; 1973, ч. П, ХУШ. I, 78-93.

4. Рао С.Р. Линейные статистические методы и их применения. М., Наука, 1968.

5. Введение в теорию порядковых статистик. Пер. с англ./Под ред. А.Я.Боярского. М., Статистика, 1970.

6. Тихов М.С., Сокова И.Н. Об одном подходе к оптимизации при управлении качеством производственного процесса. - Математическое обеспечение АСУП (тезисы П Всесоюзн. семинара). ЦЭМИ АН СССР, М.-Горький, 1975, 184-186.

7. Halperin M. Maximum Likelihood Estimation in Truncated Samples.- Ann. Math. Stat., 23, 226, 1952.

8. Кулльдорф Г. Введение в теорию оценивания. М., Наука, 1966.

9. Гихман И.И., Скороход А.В. Введение в теорию случайных процессов. М., Наука, 1965.

IO. Прохоров Ю.В. Сходимость случайных процессов и предельные теоремы теории вероятностей. - Теория вероятн. и ее примен., 1965, I, 177-238.