Доклады Академии Наук СССР
1953. Том XCI, № 4

МАТЕМАТИКА

И. И. ГИХМАН

НЕКОТОРЫЕ ЗАМЕЧАНИЯ К КРИТЕРИЮ СОГЛАСИЯ А. Н. КОЛМОГОРОВА

(Представлено академиком А. Н. Колмогоровым 29 V 1953)

Предложенный А. Н. Колмогоровым (1) критерий согласия

$$K_N = \sup \sqrt{N} \, | F_N(x) - F(x) |, \tag{1}$$

оценивающий расхождение между предполагаемой («теоретической») функцией распределения $F(x)$ и эмпирическими данными, является, как известно, универсальным, в том смысле, что закон распределения величины K_N не зависит от непрерывной функции $F(x)$. Здесь N — число независимых измерений случайной величины, а $F_N(x)$ — эмпирическая функция распределения, построенная по результатам этих измерений. Предельное распределение величины K_N $(N \to \infty)$ табулировано ((2), см. также (3)), имеются таблицы распределений величин K_N и при конечных N(4, 5).

Во многих случаях, однако, относительно функции $F(x)$ известно только, что она принадлежит к некоторому классу распределений $\{F(x, \theta)\}$, содержащему параметр θ, значение которого определяется по эмпирическим данным. В этом случае вместо (1) приходится вводить величину

$$\bar{\tilde{K}}_N = \sup \sqrt{N} \, | F_N(x) - F(x) |, \tag{2}$$

где $\bar{\theta}$ — эмпирическое значение параметра θ, и свойства критерия согласия резко меняются. Последнее связано с тем, что величина K_N при $N \to \infty$ неустойчива относительно θ и в том случае, когда погрешность в определении параметра стремится к 0 вместе с N^{-1}. Одновременно распределение величины $\tilde{K}_N$ и ее предел при $N \to \infty$ перестают быть универсальными, даже если метод оценки параметра θ стандартизован. Из результатов настоящей заметки следует, что в этом направлении простые закономерности отсутствуют.

1. Разобьем прямую $-\infty < x < \infty$ на n интервалов, назовем их интервалами группировки, и пусть $F_i(\theta)$ — вероятность случайной величине принять значение, лежащее в первых i интервалах, а $\bar{F}_i$ — эмпирическое распределение, т. е. отношение числа наблюдений, попавших в первые i интервалов группировки, к общему числу N произведенных наблюдений. Положим

$$\xi_i = \sqrt{N} \{ \bar{F}_i - F_i(\bar{\theta}) \}, \quad \bar{\xi}_i = \sqrt{N} \left\{ \bar{F}_i - F_i(\theta_0) - (\bar{\theta} - \theta_0) \left(\frac{\partial F_i}{\partial \theta} \right)_0 \right\},$$
$$i = 1, 2, \ldots, n, \tag{3}$$

где θ_0 — истинное значение параметра θ; $\bar{\theta}$ — эмпирическая оценка величины θ_0; $(\partial F_i / \partial\theta)_0$ — значение производной $\partial F_i(\theta)/\partial\theta$ при $\theta = \theta_0$.

Легко доказать, что если функция распределения $F(x, \theta)$ в некоторой окрестности точки θ_0 имеет производную $\partial F(x, \theta)/\partial\theta$, равностепенно непрерывную по θ относительно параметра x, $-\infty < x < \infty$, и если оценка $\bar{\theta}$ параметра θ такова, что величина $\sqrt{N}(\bar{\theta} - \theta_0)$ при $N \to \infty$ слабо сходится к некоторому пределу, то последовательности $\{\max_i \xi_i, \min_i \xi_i\}$, $\{\max \bar{\xi}_i, \min \bar{\xi}_i\}$ асимптотически $(N \to \infty)$ эквивалентны. При этом мы называем две последовательности случайных векторов асимптотически эквивалентными, если они слабо сходятся одновременно, и если имеет место сходимость, то предельные распределения совпадают. Предположим теперь относительно метода определения $\bar{\theta}$, что

$$\sqrt{N'}\,(\bar{\theta} - \theta_0) = \frac{1}{\sqrt{N'}} \sum_{k=1}^{N'} \varphi(x_k') + \eta_{N'}, \tag{4}$$

где N' — число наблюдений x_k', использованных для определения $\bar{\theta}$; $\eta_{N'}$ — величина, сходящаяся по вероятности к 0 при $N' \to \infty$, и $\varphi(x)$ — функция, для которой существуют моменты первого и второго порядка, $M\varphi(x_k') = 0$, $M\varphi^2(x_k') = \sigma^2$. Это предположение оправдывается, например, если оценка $\bar{\theta}$ ищется по методу максимального правдоподобия и функция распределения $F(x, \theta)$ удовлетворяет некоторым дополнительным условиям (см., например, [6]). В последнем случае $\varphi(x) = I^{-2}\left(\frac{\partial}{\partial\theta}\log f(x, \theta)\right)_0$, $I = M\left(\frac{\partial}{\partial\theta}\log f(x, \theta)\right)_0^2$. Заметим еще, что подставляя (4) в (3) и отбрасывая $\eta_{N'}$, мы не нарушаем интересующей нас асимптотической эквивалентности. Положим $N/N' = \delta'^2$, $t = F(x, \theta_0)$, $t_i = F_i(\theta_0)$, $x = x(t)$ — функция, обратная к функции $t = F(x, \theta_0)$, $q(t) = \left(\frac{\partial}{\partial\theta} F(x(t), \theta)\right)$. Тогда

$$\bar{\xi}_i = \bar{\xi}(t_i) = \sqrt{N}\{\bar{F}_i - t_i\} - \delta'\,\Phi_{N'}\,q(t_i), \tag{5}$$

где $\Phi_{N'} = \frac{1}{\sqrt{N'}} \sum_{1}^{N'} \varphi(x_k)$.

В дальнейшем мы рассмотрим условное распределение максимального и минимального членов последовательности (5) при гипотезе $\Phi_{N'} = z$. В следующем случае соответствующее предельное распределение может быть указано из уже известных результатов без затруднений. Обычно для построения эмпирической функции распределения и оценки неизвестного значения параметра используются одни и те же наблюдения. Однако не лишен интереса и другой случай, когда для оценки параметра и проверки согласованности берутся два ряда независимых наблюдений. Остановимся на последней возможности. Последовательность $\bar{\xi}_i = \xi_i + \delta'\,\Phi_{N'}\,q(t_i)$ не зависит от гипотезы $\Phi_{N'} = z$ и ее поведение изучалось ранее (см., например, [7]).

Рассмотрим дифференциальное уравнение

$$\frac{\partial u}{\partial\tau} + \frac{1}{2}\frac{\partial^2 u}{\partial x^2} = 0$$

и пусть $u_T(t, x)$ — решение этого уравнения в области $0 \leqslant \tau < T$, $a_1(\tau) < x < a_2(\tau)$, $a_i(\tau) = \left[a_i + \delta z q\left(\frac{\tau}{1+\tau}\right)\right](1+\tau)$ $(a_1 < 0 < a_2)$, обращающееся в 0 на кривых $x = a_1(\tau)$, $x = a_2(\tau)$, и удовлетворяющее условию $u_T(T, x) = 1$, $a_1(T) < x < a_2(T)$.

Положим $u_0(\tau, x) = \lim\limits_{T\to\infty} u_T(\tau, x) = u_0(\tau, x / z)$.

Теорема 1. *Допустим*: a) *при* $N \to \infty$ *имеем* $N / N' \to \delta^2$, $\max\limits_i \Delta t_i \to 0$, *где* $\Delta t_i = t_{i+1} - t_i$; b) *функция* $F(x, \theta)$ *непрерывна по* x *и имеет равностепенно непрерывную по* θ *производную* $\partial F(x, \theta) / \partial\theta$ (x *рассматривается как параметр*), *а функция* $q(t)$ *непрерывно дифференцируема по* t; c) *оценка* $\overline{\theta}$ *параметра* θ *и эмпирическая функция распределения получены из двух независимых между собою последовательностей наблюдений*; d) *величина* $\sqrt{N'}(\overline{\theta} - \theta_0)$ *слабо сходится к некоторому пределу.*

Тогда условная вероятность совмещения неравенств

$$a_1 < \sqrt{N}\{\overline{F_i} - \overline{F_i}(\overline{\theta})\} < a_2, \quad i = 1, 2, \ldots, n, \quad a_1 < 0 < a_2,$$

при гипотезе $\sqrt{N'}(\overline{\theta} - \theta_0) = z_{N'}$ *сходится по вероятности к* $u_0(0, 0 / z)$, *когда* $N \to \infty$ *и* $z_{N'} \to z$.

Если теперь вернуться к формуле (4) и сделанным предположениям относительно моментов функции $\varphi(\xi)$, то из теоремы 1 легко получим предельную безусловную вероятность интересующего нас события. Она равна

$$\frac{1}{\sqrt{2\pi\sigma}} \int\limits_{-\infty}^{\infty} u_0(0, 0 / z)\, e^{-z^2/2\sigma^2}\, dz.$$

2. Перейдем теперь к более сложному случаю, когда эмпирическая функция распределения F_i и оценка параметра $\overline{\theta}$ получены из одних и тех же наблюдений. Создающееся при этом положение может быть охарактеризовано следующим образом.

Положим

$$\psi(t) = \varphi[x(t)], \quad \alpha(t) = \int\limits_0^t \psi(\tau)\, d\tau,$$

$$A(t, s, \sigma / z) = \frac{[(1-t)(z-\sigma) - (s + zq(t))\alpha(t)]\,[(1-t)\alpha'(t) + \alpha(t)]}{(1-t)\,[(1-t)\int\limits_t^1 \alpha'^2(t)\,dt - \alpha^2(t)]} -$$

$$- \frac{s + zq(t)}{1-t} - zq'(t)$$

и рассмотрим решение $u_\alpha(t, s, \sigma)$ параболического уравнения

$$\frac{\partial u}{\partial t} + A(t, s, \sigma / z)\left(\frac{\partial}{\partial s} + \psi(t)\frac{\partial}{\partial\sigma}\right)u + \frac{1}{2}\left(\frac{\partial}{\partial s} + \psi(t)\frac{\partial}{\partial\sigma}\right)^2 u = 0$$

в области $0 \leqslant t < \alpha < 1$, $a_1 < s < a_2$, $a_1 < 0 < a_2$, $-\infty < \sigma < \infty$, удовлетворяющее условиям $u(\alpha, s, \sigma) = 1$, $u(t, a_1, \sigma) = u(t, a_2, \sigma) = 0$, $0 \leqslant t < \alpha$. Это решение монотонно возрастает при возрастании α. Положим $u_0(t, s, \sigma) = \lim\limits_{\alpha\to 1} u_\alpha(r, s, \sigma)$.

Теорема 2. *Допустим, что выполнены предположения* a) *и* b) *теоремы* 1 *и, кроме того*: c′) *оценка неизвестного параметра и* *эмпирическая функция распределения строятся по одной и той же последовательности независимых измерений x_i, причем*

$$\sqrt{N}(\bar{\theta}-\theta_0)=\frac{1}{\sqrt{N}}\sum_{1}^{N}\varphi(x_i)+\eta_N,$$

где η_N сходится по вероятности в 0 при $N\to\infty$; d) *случайные величины $\varphi(x_i)$ имеют конечные моменты двух порядков;* e) *функция, обратная к функции $\psi(t)=\varphi[x(t)]$, абсолютно непрерывна.*

Тогда условная вероятность совмещения неравенств $a_1<\sqrt{N}\{\bar{F}_i - F_i(\bar{\theta})\}<a_2$, $i=1,\ldots,n$, при гипотезе $\sqrt{N}(\bar{\theta}-\theta_0)=z_N$ стремится, когда $N\to\infty$, $z_N\to z$, к $u_0(0, 0, 0)$.

Доказательство этой теоремы основано на том, что последовательность пар случайных величин $\{\bar{\xi}(t_i), \bar{\eta}(t_i);\ i=1,\ldots,n\}$, где

$$\bar{\eta}(t_i)=\sqrt{N}\left\{\frac{1}{N}\sum_{r=1}^{N}\chi(\tau_r, t_i)\psi(\tau_r)-\alpha(t_i)\right\},\qquad \chi(\tau_r, t)=\begin{cases}1, & F(x_r)<t,\\ 0, & F(x_r)\geqslant t,\end{cases}$$

при гипотезе $\bar{\eta}(1)=z_N$ образует цепь Маркова. При $N\to\infty$ и $\max_i \Delta t_i$ эта последовательность стремится к непрерывному процессу Маркова. Для отыскания дифференциального уравнения, управляющего предельным процессом, достаточно найти главные части моментов первых двух порядков приращений $\Delta\bar{\xi}(t_i)$, $\Delta\bar{\eta}(t_i)$ при гипотезах $\bar{\eta}(1)=z_N$, $\bar{\xi}(t_i)=s$, $\bar{\eta}(t_i)=\sigma$, больших N и малых Δt_i. Соответствующие оценки получатся, если воспользоваться известными предельными теоремами для плотностей распределения при суммировании независимых слагаемых и формулами Байесса.

Заметим еще, что известные теоремы, обосновывающие предельный переход от дискретного времени к непрерывному процессу (см. ([8], [9])), в данном случае требуют некоторых дополнений.

В заключение считаю своим долгом отметить, что рассмотренный в настоящей заметке вопрос был поставлен проф. Б. В. Гнеденко на руководимом им семинаре в Киевском государственном университете.

Поступило
29 IV 1953

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

[1] А. Н. Колмогоров, Giorn. d. Att., 4 (1933). [2] Н. В. Смирнов, Бюлл. МГУ, 2, в. 2 (1939). [3] Б. В. Гнеденко, Курс теории вероятностей, 1950. [4] F. Massey, Ann. Math. Statistics, 21, № 1 (1950). [5] Z. W. Birnbaum, J. Am. Stat. Assoc., 47, № 259 (1952). [6] Г. Крамер, Математические методы статистики, 1948. [7] И. Гихман, ДАН, 82, № 6 (1952). [8] А. Н. Колмогоров, Усп. матем. наук, в. 5 (1938). [9] А. Я. Хинчин, Асимптотические законы теории вероятностей, 1936.