Доклады Академии Наук СССР
1951. Том LXXVI, № 3

МАТЕМАТИКА

А. А. ПЕТРОВ

ПРОВЕРКА ГИПОТЕЗЫ О НОРМАЛЬНОСТИ РАСПРЕДЕЛЕНИЙ ПО МАЛЫМ ВЫБОРКАМ

(Представлено академиком А. Н. Колмогоровым 29 XI 1950)

1. Пусть дана совокупность из N выборок объема n

$$\begin{array}{c} x_{11}, \ldots, x_{1n}, \\ \cdots\cdots\cdot \\ \cdots\cdots\cdot \\ x_{N1}, \ldots, x_{Nn}. \end{array}$$

Ставится вопрос: допустимо ли предположение, что i-я выборка (при любом i) получена случайным выбором из бесконечной совокупности, имеющей интегральную функцию распределения $F(a_i x + b_i)$, где F — заданная функция, одна и та же для всех выборок, а параметры a_i и b_i могут быть различными для различных выборок и нам неизвестны. (Это предположение в дальнейшем будет называться гипотезой F.) В частности, допустимо ли предположение, что каждая из наших выборок взята из нормально распределенной совокупности с различными для разных выборок математическими ожиданиями и дисперсиями.

Для проверки высказанной гипотезы естественно рассматривать те или иные функции $\eta(x_1, x_2, \ldots, x_n)$, распределения которых, вычисленные в предположении, что $x_1, \ldots, x_n$ независимы и подчинены функции распределения $F(ax+b)$, не зависят от a и b. Можно рассматривать, например, величины

$$\eta' = \frac{x_{\max} - \bar{x}}{s}, \quad \eta'' = \frac{x_{\min} - \bar{x}}{s},$$

распределение которых в случае нормального закона F было изучено Н. В. Смирновым ([1]). Другой метод, отправляющийся от распределения величин $\eta_i = \dfrac{x_i - \bar{x}}{s}$, рассмотренного впервые Томпсоном ([2]), предложен Арлеем и Бухом ([3,4]). Предлагаемый далее метод имеет то преимущество, что не требует вычисления средних квадратических s. Кроме того, он дает $n-2$ отдельные кривые для сравнения и может поэтому являться более мощным средством различения типов распределений.

2. Пусть X есть случайная величина, имеющая распределение непрерывного типа ([5]) с функцией распределения F и плотностью вероятности $f = F'$, а $x_1' \leqslant x_2' \leqslant \ldots \leqslant x_n'$ есть последовательность n независимых наблюдений случайной величины X, упорядоченная в порядке возрастания. Рассмотрим отношения

$$\xi_k = \frac{x_k' - x_1'}{x_n' - x_1'} \quad (\text{где } 1 < k < n).$$

Эти величины инвариантны относительно выбора масштаба и начала координат, и поэтому их функции распределения будут одинаковыми для всех функций распределения $F(ax+b)$, отличающихся от $F(x)$ только значениями параметров a и b.

Можно доказать, что функция распределения случайной величины ξ_k есть

$$F_k(t)=\frac{n!}{(k-2)!\,(n-k-1)!}\iiint\limits_G [F(y)-F(x)]^{k-2}\,[F(z)-F(y)]^{n-k-1}\times$$

$$\times f(x)\,f(y)\,f(z)\,dx\,dy\,dz, \tag{1}$$

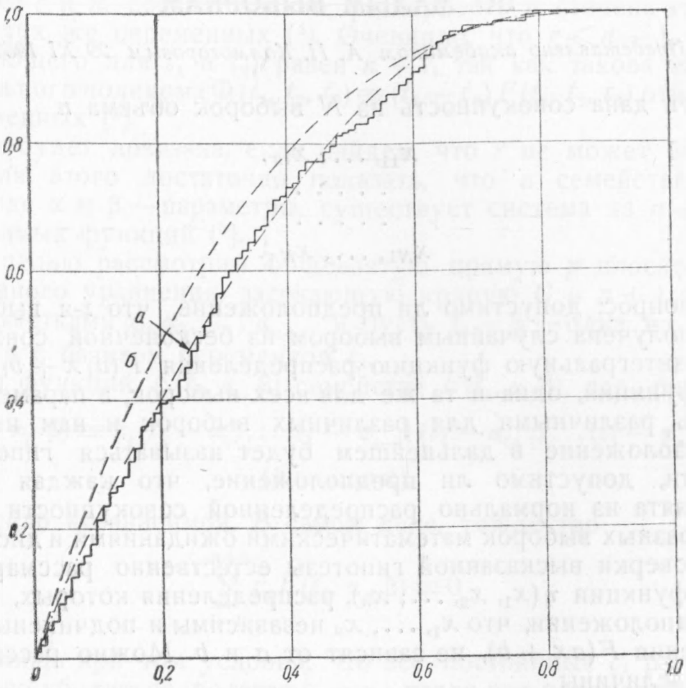

Рис. 1. $F_2(t)$

где интегрирование производится по области G, определяемой неравенствами

$$x\leqslant y\leqslant z,\qquad \frac{y-x}{z-x}<t.$$

3. Пусть величина X распределена равномерно на отрезке $(0,1)$. Тогда, в силу формулы (1),

$$F_k(t)=\frac{n!}{(k-2)!\,(n-k-1)!}\iiint\limits_{G^*}(y-x)^{k-2}\,(z-y)^{n-k-1}\,dx\,dy\,dz, \tag{2}$$

где интегрирование производится по области G^*, определяемой неравенствами

$$0\leqslant x\leqslant y\leqslant z\leqslant 1,\qquad \frac{y-x}{z-x}<t.$$

В случае $n=5$ получаем, например, отсюда

$$F_2(t)=1-(1-t)^3,\qquad F_3(t)=t^2\,(3-2t),\qquad F_4(t)=t^3.$$

4. Пусть величина X имеет непрерывную и строго монотонную функцию распределения F. С помощью подстановки $X' = F(X)$ распределение F приводится к равномерному, и функция распределения $F_k(t)$ определяется интегралом (2) по области

$$0 \leqslant x \leqslant y \leqslant z \leqslant 1, \quad \frac{F^{-1}(y) - F^{-1}(x)}{F^{-1}(z) - F^{-1}(x)} < t,$$

Рис. 2. $F_3(t)$

где F^{-1} есть функция, обратная к F. Произведя в выражении (2) интегрирование по переменной x, придем к следующей формуле, пригодной для численного интегрирования:

$$F_k(t) = \frac{n!}{(k-1)!\,(n-k-1)!} \int_0^1 dz \int_0^z (z-y)^{n-k-1} (y - x_0)^{k-1}\, dy, \tag{3}$$

где

$$x_0 = x_0(y, z, t) = F\left[\frac{F^{-1}(y) - tF^{-1}(z)}{1-t}\right].$$

5. Рекомендуемый метод проверки допустимости гипотезы F заключается в следующем. Каждая из наших N выборок дает по одному наблюденному значению для каждой из величин ξ_k. По этим N наблюденным значениям строятся эмпирические функции распределения для величин ξ_k. С помощью критерия Колмогорова [6] производится сравнение между полученными эмпирическими функциями распределения и теоретическими, вычисленными заранее по формуле (1) или (3). Если при этом обнаружатся значимые отклонения хотя бы при одном k, то испытуемая гипотеза F отвергается. В случае, когда ни при одном k таких отклонений не обнаруживается, данные, заключенные в наших выборках, хорошо согласуются с гипотезой F.

6. Для практического применения описанного метода нужно знать функции F_k, соответствующие данному значению n и испытуемой гипотезе F. Вычисления были проведены для $n=5$ и двух гипотез — нормального и равномерного распределений.

На прилагаемых графиках приведены функции F_2, F_3, F_4, вычисленные для равномерного (*б*) и нормального (*а*) распределений по формулам (2) и (3), и эмпирические функции распределения величин ξ_2, ξ_3, ξ_4, построенные по выборке в 200 пятерок, взятых из таблиц

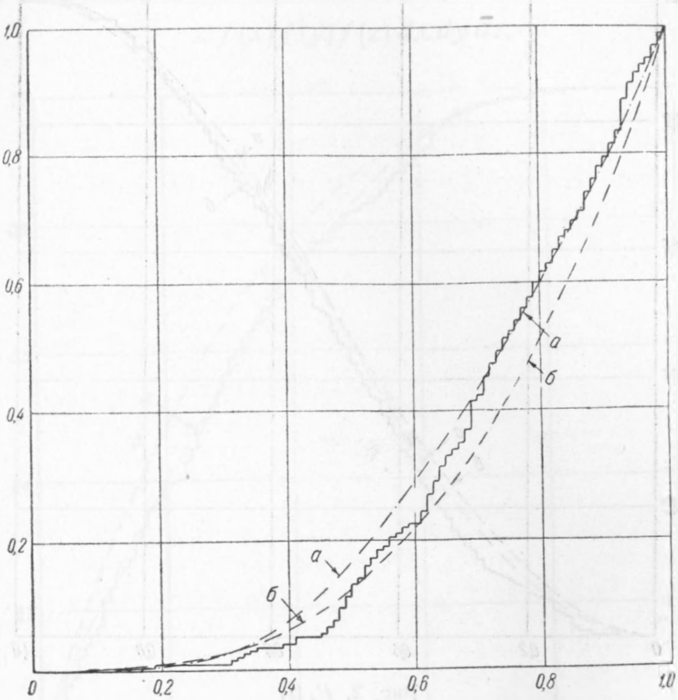

Рис. 3. $F_4(t)$

случайных чисел, подчиненных нормальному закону [7]. Во всех трех случаях эмпирические функции распределения хорошо согласуются с гипотезой нормального распределения. При сравнении с гипотезой равномерного распределения в двух случаях (ξ_2 и ξ_4) получаем значимые отклонения. Эти отклонения соответствуют в первом случае уровню значимости в 0,06% и во втором — уровню значимости в 2%.

В заключение пользуюсь случаем выразить А. Н. Колмогорову сердечную признательность за постановку задачи и руководство при выполнении этой работы.

Поступило
27 XI 1950

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

[1] Н. В. Смирнов, ДАН, **33**, 346 (1941). [2] W. R. Thompson, Ann. Math. Statistics, **6**, 214 (1935). [3] N. Arley, K. Danske, Vid. Selsk., Mat.-fys. Medd., **18**, No. 3 (1940). [4] N. Arley and K. R. Buch, Introduction to the Theory of Probability and Statistics, 1950. [5] Г. Крамер, Математические методы статистики, 1948. [6] В. И. Романовский, Применения математической статистики в опытном деле, 1947. [7] H. Wold, Random Normal Deviates, Tracts for computers, No. 25, 1948.