К ВОПРОСУ О ПРОВЕРКЕ «ПОКАЗАТЕЛЬНОСТИ»

Л. Н. БОЛЬШЕВ

1. Введение и постановка задачи. В книге Б. В. Гнеденко, Ю. К. Беляева и А. Д. Соловьева [1] среди многих задач теории надежности обсуждается задача «проверки типа распределения по малым выборкам» (стр. 231 и 249). Общие соображения и выводы конкретизируются на примере проверки гипотезы, согласно которой имеющиеся выборки предполагаются извлеченными из популяций с показательным распределением вероятностей.

Пусть $\xi_{i1}, \ldots, \xi_{in_i}$ $(n_i \geqslant 2;\ i = \overline{1, N})$ — взаимно независимые случайные величины. Требуется проверить гипотезу H: величины ξ_{ij} подчиняются показательным распределениям с плотностями вероятностей $a_i \exp(-a_i x)$ $(x > 0;\ j = \overline{1, n_i};\ i = \overline{1, N})$; значения параметров a_i неизвестны и, возможно, различны для разных i.

В книге [1] для проверки H рекомендуется при каждом i выбрать целое число $m_i (0 < m_i < n_i)$ и вычислить отношения

$$\eta_i = (n_i - m_i) \sum_{j=1}^{m_i} \xi_{ij} \Big/ m_i \sum_{k=m_i+1}^{n_i} \xi_{ik} \qquad (i = \overline{1, N}). \tag{1}$$

Если гипотеза H верна, то, как известно, отношение (1) подчиняется F-распределению с количествами степеней свободы $2m_i$ и $2(n_i - m_i)$. Таким образом, если функцию F-распределения со степенями свободы f_1 и f_2 обозначить $G(x; f_1, f_2)$, то можно заключить, что случайные величины

$$G[\eta_i;\ 2m_i,\ 2(n_i - m_i)] \qquad (i = \overline{1, N})$$

при гипотезе H взаимно независимы и распределены равномерно на отрезке $[0, 1]$. При отсутствии каких-либо определенных сведений о конкурирующих гипотезах этот вывод рекомендуется проверять с помощью какого-нибудь критерия значимости (например, с помощью непараметрического критерия А. Н. Колмогорова).

Основной недостаток такой процедуры заключается в том, что случайная величина η_i не характеризует гипотетическое показательное распределение. Если для проверки сложной гипотезы строится критерий значимости при отсутствии каких-либо определенных сведений о конкурирующих гипотезах, то разумно требовать, чтобы по крайней мере распределение статистики этого критерия находилось во взаимно однозначном соответствии с проверяемой гипотезой. В данном же случае такого соответствия нет: если исходные величины ξ_{ij} подчиняются показательным распределениям (гипотеза H), то η_i необходимо подчиняются F-распределениям, но обратное утверждение, как показал Лаха [2], неверно. Иными словами, *F-распределение отношений* (1) *не является характеристическим свойством показательного распределения.*

Данное сообщение посвящено уточнению изложенной процедуры проверки «показательности».

2. Основной критерий. Рассмотрим сначала лишь одну выборку $\xi_1, \ldots, \xi_n$ и, несколько расширяя задачу, предположим (гипотеза H), что ξ_i подчиняются Γ-распределениям с плотностями вероятностей

$$p(x;\ a,\ k_i) = \frac{a^{k_i}}{\Gamma(k_i)} x^{k_i - 1} e^{-ax} \qquad (x > 0,\ i = \overline{1, n}),$$

где a — неизвестный параметр и k_i — заданные числа $(a > 0,\ k_i > 0)$.

Теорема. *Сумма* $\zeta_n = \xi_1 + \ldots + \xi_n$ *и отношения*

$$\zeta_r = \sum_{j=1}^{r} \xi_j \Big/ \sum_{k=1}^{r+1} \xi_k \qquad (r = \overline{1, n-1}) \tag{2}$$

независимы в совокупности. Сумма ζ_n *имеет* Г-*распределение, а отношения* ζ_r *подчиняются* В-*распределениям с параметрами* $l_r = k_1 + \ldots + k_r$ *и* k_{r+1}.

Доказательство. Плотность совместного распределения величин ξ_i выражается формулой

$$\prod_{i=1}^{n} p(x_i;\, a,\, k_i) = a^{ln} \exp\left(-a\sum_{i=1}^{n} x_i\right)\prod_{i=1}^{n} \frac{x_i^{k_i-1}}{\Gamma(k_i)}.$$

Положим $x_1 = z_1 z_2 \ldots z_n$ и $x_i = (1 - z_{i-1}) z_i z_{i+1} \ldots z_n$ $(i = \overline{2,\ n})$ (z и x связаны такими же соотношениями, как ζ и ξ). Якобиан этого преобразования равен произведению $z_2 z_3^2 \ldots z_n^{n-1}$, поэтому плотность совместного распределения величин ζ_r выражается формулой:

$$p(z_n;\, a,\, l_n) \prod_{r=1}^{n-1} \frac{1}{B(l_r,\, k_{r+1})} z_r^{l_r-1}(1 - z_r)^{k_{r+1}-1}.$$

Теорема доказана.

Так как распределения статистик $\zeta_1, \ldots, \zeta_{n-1}$ не зависят от мешающего параметра a, то их можно заменить независимыми равномерно распределенными случайными величинами (подобно тому, как это указано во введении) и далее воспользоваться критерием А. Н. Колмогорова. Попутно заметим, что статистики ζ удобнее статистик η, так как В-распределение табулировано, а таблиц F-распределения нет.

Применяя все эти выводы к N выборкам, мы получим в результате такой процедуры не N, а $n_1 + \ldots + n_N - N$ случайных величин, которые при гипотезе H независимы и распределены одинаково равномерно на отрезке [0, 1]. Указанный способ построения критерия позволяет избежать потерь информации, заключенной в выборочных данных. Более того, если $n \geqslant 3$, то независимость статистик $\zeta_1, \ldots, \zeta_{n-1}$ и подчинение их В-распределениям с параметрами, указанными в формулировке теоремы, является свойством, *характеризующим* Г-распределение. Этот вывод — следствие одной теоремы Ю. В. Прохорова [3] о характеризации распределения с точностью до масштабного параметра. Из этой теоремы, а также из упомянутого выше результата Лахи [2] вытекает, что Г-тип характеризуется не менее чем двумя отношениями вида (2). Таким образом, например, в случае нескольких выборок критерий, основанный на статистиках вида (2), будет корректным, если все объемы выборок $n_i \geqslant 3$.

3. Критерий показательности. Вернемся к задаче, сформулированной во введении, и условимся в дальнейшем считать, что все $n_i \geqslant 3$. Рассмотрим выборку с номером i и положим

$$\zeta_{ir} = \sum_{j=1}^{r} \xi_{ij} \Big/ \sum_{k=1}^{r+1} \xi_{ik} \qquad (r = \overline{1,\ n_i - 1}).$$

Если гипотеза H верна, то согласно теореме из п. 2, статистики ζ_{ir} взаимно независимы и подчиняются В-распределениям с параметрами r и 1. Отсюда, в свою очередь, следует, что статистики ζ_{ir}^r $(r = \overline{1,\ n_i - 1};\ i = \overline{1,\ N})$ взаимно независимы и одинаково равномерно распределены на отрезке [0, 1]. К этим статистикам (в количестве $\Sigma(n_i - 1)$) и следует применять критерий А. Н. Колмогорова.

Аналогичный вывод можно сделать и в более общем случае, когда гипотетические распределения зависят от параметра сдвига и выражаются формулами $a_i \exp\{-a_i(x - b_i)\}$, $x > b_i$. Для этого нужно по каждой выборке $\xi_{i1}, \ldots, \xi_{in_i}$ предварительно вычислить разности $\xi_{ij} - \xi_{i0}$, полагая $\xi_{i0} = \min(\xi_{i1}, \ldots, \xi_{in_i})$. Эти разности следует принять за исходные значения, вычислить статистики ζ_{ir}^r и применить какой-либо критерий значимости. Разумеется, нулевые разности нужно исключить. Так как в i-й выборке количество ненулевых разностей равно $n_i - 1$, то количество статистик ζ^r равно $n_i - 2$. Следовательно, при проверке принадлежности выборок к показательному типу объемы всех выборок должны быть не менее четырех: только

при соблюдении этого условия распределение статистик ζ характеризует показательное распределение.

4. Пример (данные заимствованы из книги [1], стр. 249). В результате некоторых испытаний были получены значения $\xi_1 = 300$, $\xi_2 = 693$, $\xi_3 = 980$, $\xi_4 = 1358$, $\xi_5 = 1344$, $\xi_6 = 2090$. Согласуются ли эти наблюдения с гипотезой, что ξ_i независимы и подчиняются одному и тому же показательному распределению? Для ответа на этот вопрос в книге [1] вычисляется значение статистики (1) при $m = 3$:

$$\eta = \frac{300 + 693 + 980}{1358 + 1344 + 2090} = \frac{1973}{4792} = \frac{1}{2{,}429}$$

(в [1] ошибочно указано значение $\eta = 1/3{,}5$). Согласно двусторонним F-критериям с уровнями значимости 50, 20 и 10% критические значения для $1/\eta$ равны соответственно 1,782, 3,055 и 4,284. Иными словами, гипотеза показательности отвергается критерием с уровнем значимости 50% и не отвергается критерием с уровнем $\leqslant 20\%$. В книге [1] к указанным шести результатам наблюдений были добавлены еще четыре. По десяти наблюдениям с помощью статистики η удалось установить, что в рассматриваемом случае гипотеза показательности должна быть отвергнута с уровнем значимости 5%. Однако критерий с уровнем значимости 2% эту гипотезу не отвергает.

Применим теперь к исходным шести наблюдениям критерий показательности из п. 3. Имеем:

$$\zeta_1 = 0{,}3021, \quad \zeta_2^2 = 0{,}2533, \quad \zeta_3^3 = 0{,}2078, \quad \zeta_4^4 = 0{,}2577, \quad \zeta_5^5 = 0{,}1576.$$

Если проверяемая гипотеза верна, то ζ_r^r — значения независимых случайных величин, равномерно распределенных на отрезке [0, 1]. Статистика критерия А. Н. Колмогорова в данном случае принимает значение $D_5 = 0{,}6979$, превышающее 1%-е критическое значение 0,669 (см., например, [1], стр. 495). Поэтому проверяемую гипотезу следует отвергнуть (уровень значимости $\approx 0{,}58\%$). Для такого заключения не потребовалось никаких дополнительных наблюдений.

Поступила в редакцию
23.11.65

ЛИТЕРАТУРА

[1] Б. В. Гнеденко, Ю. К. Беляев, А. Д. Соловьев, Математические методы в теории надежности, М., Изд-во «Наука», 1965.

[2] R. G. Laha, On a problem connected with beta and gamma distributions, Trans. Amer. Math. Soc., **113**, 2 (1964), 287—298.

[3] Ю. В. Прохоров, Характеризация класса распределений распределением некоторой статистики, Теория вероят. и ее примен., X, 3 (1965), 479—487.

ON THE QUESTION OF TESTING «EXPONENTIALNESS»

L. N. BOL'ŠEV (MOSCOW)

(Summary)

The question whether the formal tests of significance for the observations to belong to the «exponential» population are correct is discussed.