Мат. VI международной конференции "Актуальные проблемы электронного приборостроения"
АПЭП-2002. Новосибирск, 2002. - Т.6. - С.36-38.

РАСПРЕДЕЛЕНИЯ СТАТИСТИК ДИСПЕРСИОННОГО АНАЛИЗА ПРИ ОТКЛОНЕНИИ ЗАКОНА РАСПРЕДЕЛЕНИЯ ОШИБОК ОТ НОРМАЛЬНОГО[1]

Б.Ю. Лемешко, В.М. Пономаренко

Новосибирский государственный технический университет
Новосибирск, Россия. E-mail: headrd@fpm.ami.nstu.ru, ponomarenkov@mail.ru

Дисперсионный анализ традиционно используется в тех областях человеческой деятельности, где возникает необходимость сравнительного анализа влияния различных факторов на некоторую интересующую величину. В основе аппарата классического дисперсионного анализа лежит предположение о принадлежности ошибки наблюдаемого отклика нормальному закону, что в случае модели с постоянными факторами приводит к нормальности самого отклика. Базируясь на этом, получены предельные распределения статистик, используемых в классическом дисперсионном анализе [1].

Основные предположения относительно рассматриваемой модели, обозначаемые через Ω, в классическом случае имеют вид [2]:

$$\Omega: \bar{y} = \bar{X}\bar{\theta} + \bar{e}, \quad rg(\bar{X}) = r, \quad \bar{e} \sim N(0, \sigma^2 I), \tag{1}$$

где $\bar{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ - вектор откликов, $\bar{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$ - вектор параметров, $\bar{X}$ - матрица планирования, $\bar{e}$ - вектор ошибок наблюдений. В рассматриваемом частном случае однофакторного анализа предположения Ω чаще всего записываются в виде [1]:

$$\Omega: \begin{cases} y_{ij} = \beta_i + e_{ij} \quad (i = 1, \ldots, I;\ j = 1, \ldots, J_i), \\ \{e_{ij}\} \ \textit{независимы и принадлежат} \ N(0, \sigma^2). \end{cases} \tag{2}$$

В этом случае $\bar{y} = (y_{11}, y_{12}, \ldots, y_{1J_1}, y_{21}, \ldots, y_{2J_2}, \ldots, y_{I1}, \ldots, y_{IJ_I})^{\mathrm{T}}$, $\bar{\theta} = (\beta_1, \ldots, \beta_I)^{\mathrm{T}}$, $n = \sum_{i=1}^{I} J_i$, $p = I$. При проведении исследований использовалась модель вида (2) размерности $I = 3$, имеющая ранг матрицы планирования $r = rg(\bar{X}) = 3$, рассматривались только сбалансированные планы наблюдений, когда $J_1 = J_2 = J_3$.

Любую гипотезу дисперсионного анализа можно задать в виде $H_0 : K^T\bar{\theta} = \bar{b}_0$ [2]. Статистика критерия проверки гипотезы

$$\Im = \frac{n-r}{g} \frac{SS_\omega - SS_\Omega}{SS_\Omega} \tag{3}$$

подчиняется $F_{g,n-r}$ -распределению Фишера, где SS_ω – минимум суммы квадратов $SS(\bar{y}, \bar{\theta}) = \|\bar{y} - \bar{X}\bar{\theta}\|$ по вектору параметров $\bar{\theta}$ при выполнении предположений ω, $\omega = H_0 \cap \Omega$, а SS_Ω - минимум суммы квадратов при выполнении предположений Ω, $g = r - m$, $m = rg(\bar{X}^{\mathrm{T}} \mid K) - rg(K)$. Если основу проверяемой гипотезы составляют ФДО (функции, допускающие оценку), что справедливо для всех рассматриваемых ниже случаев, то, как следствие из определения ФДО, $rg(X^{\mathrm{T}} \mid K) = rg(K)$, и тогда $g = rg(K)$. В итоге в нашем случае статистика (3) будет распределена как $F_{rg(K), n-rg(\bar{X})} = F_{rg(K), n-3}$. На практике вместо статистики (3) используют статистику вида

$$\Im = \frac{n-r}{g} \frac{(K^{\mathrm{T}}\bar{\theta}_0 - \bar{b}_0)^{\mathrm{T}} (K^{\mathrm{T}} G K)^{-1} (K^{\mathrm{T}}\bar{\theta}_0 - \bar{b}_0)}{(\bar{y} - \bar{X}\bar{\theta}_0)^{\mathrm{T}} (\bar{y} - \bar{X}\bar{\theta}_0)}, \tag{4}$$

[1] Работа выполнена при поддержке Российского фонда фундаментальных исследований (проект № 00-01-00913)

где G - обобщенно обратная по отношению к $\overline{X}^{\mathrm{T}}\overline{X}$ матрица, а $\overline{\theta}_0$ - оценка параметров $\overline{\theta}$ при выполнении предположений Ω. Статистика (4) совпадает со статистикой (3), если $\overline{\theta}_0$ находят по методу наименьших квадратов и при предположениях Ω, и при предположениях ω. В рассматриваемом однофакторном случае можно выделить следующие гипотезы:

1. О равенстве математических ожиданий оценок параметров известному вектору
$$H_0 : \overline{\theta} = \overline{\theta}_0. \qquad (5)$$
Такая гипотеза в рамках дисперсионного анализа допустима только для модели вида (2). Статистика критерия распределена как $F_{3,n-3}$, поскольку в данном случае формирующая гипотезу матрица K совпадает с единичной матрицей и имеет ранг, равный трем.
2. О равенстве математических ожиданий оценок ФДО, некоторым известным значениям
$$H_0 : \beta_2 - \beta_1 = \psi_1, \quad \beta_3 - \beta_1 = \psi_1, \quad \ldots, \quad \beta_l - \beta_1 = \psi_{l-1}. \qquad (6)$$

В качестве набора функций рассматривается базис ФДО размерности $l-1$. Гипотеза такого вида часто используется для проверки предположения о равенстве всех эффектов уровней фактора β_j, что можно сделать, полагая $\psi_i = 0, \quad i = 1,\ldots,l-1$. Статистика критерия распределена как $F_{2,n-3}$, так как для используемой модели $l-1=2$, следовательно, гипотезу составляют две ФДО и $rg(K)=2$.

Цель данной работы в исследовании распределений статистик, используемых при проверке гипотез (5)-(6) однофакторного анализа, при отклонении закона распределений ошибок от нормального, в частности, когда ошибки распределены по законам логистическому или Лапласа. Использование аналитических методов оказывается весьма затруднительным в силу нетривиальности самих задач. Поэтому в основу проводимого исследования положена развиваемая методика компьютерного анализа статистических закономерностей.

В работе моделировались выборки значений статистики (4) с соответствующими проверяемой гипотезе матрицей K, вектором $\overline{b}_0$ и вектором оценок параметров модели $\overline{\theta}_0$, найденным по методу наименьших квадратов. Получаемые эмпирические распределения статистик хорошо описываются бета-распределением II-го рода, частным случаем которого является $F_{g,n-r}$ -распределение Фишера. Анализ полученных эмпирических распределений статистик осуществлялся с использованием системы статистического анализа [3], разработанной на кафедре прикладной математики НГТУ.

Для каждой модели вида (2) (с соответствующим распределением ошибки) моделировалось не менее 10 выборок значений статистики (4) объемом 2000 наблюдений. Параметры моделей распределений статистик усреднялись по количеству выборок. Полученная в результате усреднения модель рассматривается как приближение предельного распределения статистики в каждом конкретном случае.

В качестве примера на рис. 1 приведены предельное распределение классической статистики при проверке гипотезы (6) (наиболее типичной гипотезы однофакторного дисперсионного анализа) и полученные в результате моделирования приближения предельных законов данной статистики в случае принадлежности ошибок законам распределения нормальному, логистическому и Лапласа. Объем выборок в этом эксперименте составлял $n=15$. Для нормального распределения проводился дополнительный эксперимент при объеме выборок $n=27$. Из рисунка видно, что все четыре кривые практически сливаются друг с другом. Это говорит о малом влиянии не только незначительных отклонений от предположения нормальности, как в случае логистического закона распределения ошибок, но и довольно существенных отклонений, как в случае распределения Лапласа. Аналогичная ситуация наблюдается и для статистики критерия проверки гипотезы (5). Следует отметить, что область корректности данных результатов пока сужена рассмотрением лишь симметричных законов распределения ошибок, статистическими “выводами о средних” [1], не касающихся гипотез относительно дисперсий, сбалансированным планом наблюдений и выполнением остальных предположений Ω: независимости и однородности наблюдений. Однако полученные ре-

зультаты позволяют надеяться, что начатое исследование распределений статистик дисперсионного анализа позволит в дальнейшем получить результаты, существенно расширяющие область корректного применения задач дисперсионного анализа.

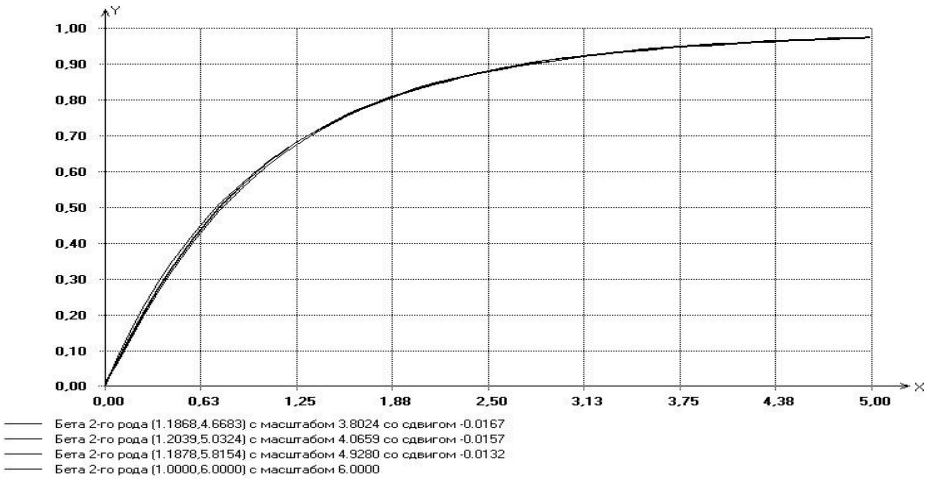

Рис. 1

Полученные нами результаты не противоречат выводам работы [1]. В [1] Шеффе показано, что при $n \to \infty$ критерии для гипотез, касающихся проверки "выводов о средних" в случае одной популяции, практически не изменятся при нарушении предположения нормальности в силу того, что лежащая в основе статистик оценка дисперсии S сходится по вероятности к σ. Предельные распределения статистик останутся теми же при замене S на σ, Это приводит к независимости распределений статистик от популяций при больших n.

В целом в критериях проверки гипотез больше всего изучено влияние ненормальности на ошибки первого рода. Это изучение осуществлялось с помощью нескольких подходов. Пирсоном использовались экспериментальные выборки для изучения влияния ненормальности ошибок на распределение статистики $\Im$ для проверки равенства средних в однофакторном анализе. Полученные им результаты позволяют говорить о том, что такое влияние не велико в случае симметричных законов ошибок и сбалансированных планах наблюдений.

Боксом и Андерсоном была предпринята попытка вычисления поправок к числам степеней свободы распределения Фишера, связанных с наблюдаемым значением эксцесса γ_2. Правда, в найденную взаимозависимость изначально заложена определенная аппроксимация, приводятся данные лишь по узкому кругу распределений, которые полностью определяются своими четырьмя моментами.

В целом, проведенные нами эксперименты показали эффективность компьютерного моделирования для исследования закономерностей в дисперсионном анализе. Очевидна необходимость расширения исследований через усложнение вида модели, расширение круга изучаемых распределений и рассматриваемых гипотез с целью установления по возможности функциональной зависимости между видом закона распределения ошибки и модели в целом и параметрами получаемых (моделей) предельных распределений статистик.

ЛИТЕРАТУРА

1. Шеффе Г. Дисперсионный анализ. – М.: Физматгиз, 1963. – 628 с.
2. Маркова Е. В., Денисов В. И., Полетаева И. А., Пономарев В. В. Дисперсионный анализ и синтез планов на ЭВМ. – М.: Наука, 1982. – 195 с.
3. Лемешко Б.Ю., Постовалов С.Н. Система статистического анализа наблюдений и исследования статистических закономерностей // Сб. "Моделирование, автоматизация и оптимизация наукоемких технологий". – Новосибирск: изд-во НГТУ, 2000. – С. 44-46.