Cornish–Fisher expansions for distributions of generalised Hotelling–type statistics based on random size samples

M. M. Monakhov [1]

[1]Lomonosov Moscow State University, Moscow, Russian Federation, mih_monah@mail.ru

In data analysis, the problem of multiple comparisons often appears, for example, different age, professional, social strata of the population, or the influence of different doses of the drug, diagnostic methods, etc. This problem is solved by the analysis of variance, which is used to research the influence of one or more qualitative variables (factors) on one dependent quantitative variable. Variance analysis is widely used in manufacturing, healthcare, advertising, food, and service industries, and its implementations are presented in statistical packages for many programming languages. The essence of the analysis of variance is to divide the total variance of the studied trait into separate components due to the influence of specific factors, and to test hypotheses about the significance of the influence of these factors on the studied trait. Additional problems appears in the case where the observation volume is random, see Bening, Galieva, and Korolev [1].

In the problems of multivariate univariate analysis of variance, we consider q samples with a fixed size $n_1, \ldots, n_q$: $(X_{11}, \ldots, X_{1n_1}), \ldots, (X_{q1}, \ldots, X_{qn_q})$, where X_{ij} – p-dimensional observation, represented as $X_{ij} = \mu + \alpha_i + \epsilon_{ij}$, where μ and α_i – unknown vector parameters, and ϵ_{ij} are random errors which are independent equally distributed random variables (i.e.d. r.v.) with a normal distribution of $N_p(0, B)$. When we consider the main hypothesis of sample homogeneity $H_0 : \alpha_1 = \ldots = \alpha_q = 0$, the matrices S_h and S_e are defined, reflecting the inter-level and intra-level differences, respectively, for the elements of the samples.

$$S_h = \sum_{i=1}^{q} n_i(\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})', \quad S_e = \sum_{i=1}^{q}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)',$$

with $n = n_1 + \ldots + n_q$ and

$$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{q}\sum_{j=1}^{n_i} y_{ij}.$$

Assuming the validity of the main hypothesis H_0, the random matrices S_h and S_e are independent and have central Wishart distributions $W_p(q, I_p)$ and $W_p(n, I_p)$ respectively. Based on the S_h and S_e matrices, statistics to test the H_0 hypothesis are constructed, including the Lawley–Hotelling statistics.

In the papers of Bening, Galieva, and Korolev [1, 2], a general transfer theorem was proved, which allows us to estimate the convergence rate of a first-order Chebyshev–Edgeworth expansion for asymptotically normal statistics constructed from random size samples, as well as to obtain an explicit form of this expansion. In the paper of Christoph, Monakhon and Ulyanov [3] second-order Chebyshev–Edgeworth and Cornish–Fisher expansions are obtained for sample with mean-type statistics constructed from from samples with random sizes. In the present paper an analog of the transfer theorem is proved for a Hotelling-type statistic with random size, and an asymptotic Chebyshev–Edgeworth and Cornish–Fisher expansions are constructed for the distribution function (d. f.) and quantiles of this statistic.

Define the Lawley–Hotelling statistics as

$$T_n = T_0^2 = n\mathrm{tr}S_hS_e^{-1}. \tag{1}$$

Consider the case when the parameter n is not defined in advance, but is a random variable N_n. Consider now generalized normalized Hotelling statistics of random size, based on statistics (1) with i.e.d. r.v. $N_{n_1}, \ldots, N_{n_q}$

$$T_{N_n} = \widetilde{T}_0^2 = g(n)\mathrm{tr}S_hS_{N_n}^{-1}. \tag{2}$$

We write down the following theorem from the paper of Fujikoshi, Ulyanov, and Shimizu [4].

Theorem 1. Let the statistic T_n be defined in the formula 2, $G_k(x) = Pr\{\chi^2 < x\}$ – the chi-square d. f. with k degrees of freedom. There exists a real number $C_1 > 0$ such that for all integers $n \geq 1$

$$\sup_x \left|\mathbb{P}\Big(n\mathrm{tr}S_hS_e^{-1} \leq x\Big) - G_k(x) - \frac{k}{4n}\sum_{j=0}^{2} a_jG_{k+2j}(x)\right| \leq C_1n^{-2},$$

where $k = pq, a_0 = q - p - 1, a_1 = -2q$ and $a_2 = q + p + 1$.

Assume that the d. f. of the normalized random sample size N_n satisfies the following condition.

Condition 1. There exist constants $m \in \mathbb{N}, \beta > m/2, C_2 > 0$, d. f. $H(y)$ with $H(0+) = 0$, functions of bounded variation $h_i(y), i = 1, \ldots, m$, the sequence $0 < g(n) \uparrow \infty, n \to \infty$ such that for all integers $n \geq 1$

$$\sup_{y\geq 0}\left|\mathbb{P}\left(\frac{N_n}{g(n)} \leq y\right) - H(y) - \sum_{i=1}^{m}\frac{1}{n^{i/2}}h_i(y)\right| \leq C_2n^{-\beta}, n \in \mathbb{N}.$$

Analog of the transfer theorem for Hotelling-type statistics. We formulate an analog of the transfer theorem that allows us to estimate the distribution of generalized normalized Hotelling statistics of random size $g(n)\mathrm{tr}S_hS_{N_n}^{-1}$.

Theorem 2. Let the statistic T_{N_n} be defined in the formula 2 and for a random sample size N_n the Condition 1 is satisfied. Then there is a constant $C_3 > 0$ such that the inequality holds

$$\sup_x \left|\mathbb{P}\big(\, g(n)\,\mathrm{tr} S_h S_{N_n}^{-1} \le x\big) - F_n(x)\right| \le C_1 \mathbb{E} N_n^{-2} + \frac{C_3 + C_2 M_n}{n^\beta},$$

where

$$\begin{aligned} F_n(x) &= \int_{1/g(n)}^{\infty} G_k(xy) dH(y) + \sum_{i=1}^{m} \frac{1}{n^{i/2}} \int_{1/g(n)}^{\infty} G_k(xy) dh_i(y) + \\ &+ \frac{k}{4g(n)} \int_{1/g(n)}^{\infty} \sum_{j=0}^{2} \frac{a_j}{y} G_{k+2j}(xy) dH(y) + \\ &+ \frac{k}{4g(n)} \sum_{i=1}^{m} \frac{1}{n^{i/2}} \int_{1/g(n)}^{\infty} \frac{1}{y} \sum_{j=0}^{2} a_j G_{k+2j}(xy) dh_i(y), \end{aligned}$$

$$M_n = \sup_x \int_{1/g(n)}^{\infty} \left| \frac{\partial}{\partial y} \left(G_k\,(yx) + \frac{k}{4g(n)y} \sum_{j=0}^{2} a_j G_{k+2j}\,(yx) \right) \right| dy.$$

Chebyshev–Edgeworth expansions. Consider now an example of the application of the Theorem 2. Let the sample size $N_n(r)$ have a negative binomial distribution (shifted by 1) with a success probability of $1/n$ and a probability function

$$\mathbf{P}(N_n(r) = j) = \frac{\Gamma(j + r - 1)}{(j-1)!\,\Gamma(r)} \left(\frac{1}{n}\right)^r \left(1 - \frac{1}{n}\right)^{j-1}, \quad r > 0, \;\; j = 1, 2, \ldots \quad (3)$$

Using Theorem 1 from [4], we obtain the following result.

Theorem 3. Let the statistic T_m be defined by the formula 1. Let also assume that a discrete random variable $N_n = N_n(r)$ with parameter $r > 1$ has a distribution defined in 3 and is independent of $W_p(q, I_p)$ and $W_p(n, I_p)$. Consider the statistics $T_{N_n} = g(n)\,\mathrm{tr} S_h S_{N_n}^{-1}$. Asymptotic expansion for a random volume $N_n(r)$ with $r > 1$ from the paper [4, Theorem 1] is valid with $g(n) = \mathbf{E}\big(N_n(r)\big) = r(n-1) + 1$. Then there is a constant $C = C(r) > 0$ such that for all $n \in \mathbf{N}$

$$\sup_x \left|\mathbf{P}\Big(g(n)\,\mathrm{tr} S_h S_{N_n}^{-1} \le x\Big) - F_{2;n}(x)\right| \le C \begin{cases} n^{-r}, & 1 < r < 2, \\ \ln(n)\, n^{-2}, & r = 2, \\ n^{-2}, & r > 2, \end{cases}$$

where

$$F_{2;n}(x) = F\left(\frac{x}{k};k,2r\right) + \frac{1}{n}\frac{(r-2)x}{2rk}\left(f\left(\frac{x}{k};k,2r\right) - f\left(\frac{r-1}{rk}x;k,2r-2\right)\right) +$$
$$+ \frac{k}{4(r(n-1)+1)}\sum_{j=0}^{2} a_j F\left(\frac{r-1}{(k+2j)r}x;k+2j,2r-2\right) +$$
$$+ \frac{k}{4n(r(n-1)+1)}\sum_{j=0}^{2} a_j \left[\frac{2-r}{2(r-1)}F\left(\frac{r-1}{r(k+2j)}x;k+2j,2r-2\right) +\right.$$
$$+ \frac{r}{2(r-1)}F\left(\frac{r-2}{r(k+2j)}x;k+2j,2r-4\right) -$$
$$- \frac{(2-r)x}{2r(k+2j)}f\left(\frac{r-1}{r(k+2j)}x;k+2j,2r-2\right) -$$
$$\left. - \frac{(r-2)x}{2(k+2j)(r-1)}f\left(\frac{r-2}{r(k+2j)}x;k+2j,2r-4\right)\right].$$

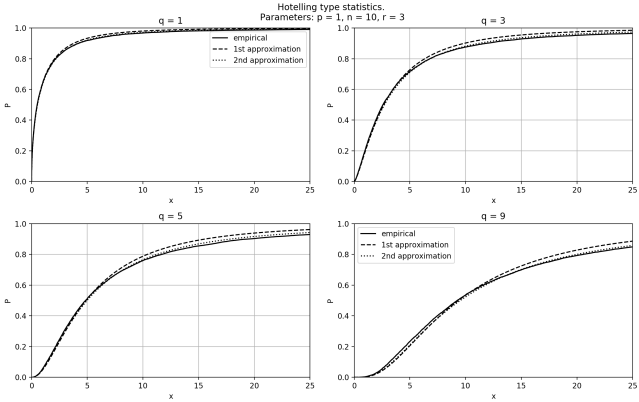

Figure 1: Empirical distribution function $\mathbb{P}\big(g(n)\mathrm{tr}S_h S_{N_n}^{-1} \le x\big)$, first-order approximation $F\left(\frac{x}{k};k,2r\right)$, and second-order approximation $F_{2;n}(x)$.

Figure 1 shows the advantage of the second-order Chebyshev–Edgeworth expansion over the first-order expansion in the approximation of the empirical distribution function.

Cornish–Fisher expansions.

Theorem 4. Under the conditions of the Theorem 3, let $x = x_\alpha$, $u = u_\alpha$ – be the α-quantiles of the normalized statistics $\mathbf{P}\Big(g(n)\,\mathrm{tr}S_h S_{N_n}^{-1} \le x\Big)$ and the limit F–distribution, respectively. Then the following asymptotic expansion

holds for $n \to \infty$:

$$\begin{aligned} x &= ku + \frac{u}{6}\left(1 - \frac{f\left(u/3; k, 1\right)}{f\left(u; k, 3\right)}\right) n^{-1} - \\ &- \frac{k}{6}\sum_{j=0}^{2} a_j \frac{F\left(\frac{k}{(k+2j)3}u; k+2j, 1\right)}{f\left(u; k, 3\right)} n^{-1} + \mathcal{O}(n^{-3/2}). \end{aligned}$$

References

1. V. E. Bening, N. K. Galieva, V. Yu. Korolev, *Asymptotic expansions for the distribution functions of statistics constructed from samples with random sizes*, Informatics and Applications, **7**:2 (2013) 75–83.

2. V. E. Bening, N. K. Galieva, V. Yu. Korolev, *On rate of convergence in distribution of asymptotically normal statistics based on samples of random size*, Annales Mathematicae et Informaticae, **39** (2012) 17–28.

3. G. Christoph, M. M. Monakhov, V. V. Ulyanov, *Second order Chebyshev–Edgeworth and Cornish–Fisher expansions for distributions of statistics constructed from samples with random sizes*, Zapiski Nauchnykh Seminarov POMI, **466** (2017) 167–207.

4. Y. Fujikoshi, V. V. Ulyanov, R. Shimizu, *L_1-norm error bounds for asymptotic expansions of multivariate scale mixtures and their applications to Hotellings generalized T_0^2*, Journal of Multivariate Analysis, **466** (2005) 1–19.