APPROXIMATIONS FOR DISTRIBUTIONS OF THE POWER DIVERGENCE GOODNESS-OF-FIT STATISTICS

Vladimir ULYANOV
Faculty of Computational Mathematics and Cybernetics
Moscow State University

Let $\boldsymbol{X} = (X_1, \ldots, X_{k+1})$ be a random vector in $\mathbf{R}^{k+1}$ with multinomial distribution with parameters n, $\pi_1, \pi_2, \ldots, \pi_{k+1}$, where $\pi_1 + \ldots + \pi_{k+1} = 1$, i.e. for the integers $n_j : 0 \le n_j \le n$ for $j = 1, 2, \ldots, k+1$, and $n_1 + \ldots + n_{k+1} = n$, we have

$$\mathrm{P}(X_1 = n_1, \ldots, X_{k+1} = n_{k+1}) = n! \prod_{j=1}^{k+1} \pi_j^{n_j}/n_j!.$$

For testing the simple hypothesis $H_0 : \pi = \mathbf{p}$ ($\mathbf{p}$ is a fixed vector) against $H_1 : \pi \ne \mathbf{p}$ three statistics are often used:
Karl Pearson's chi-square test: $T_1 = \sum_{j=1}^{k+1}(X_j - np_j)^2/(np_j)$,
Log-likelihood ratio statistic: $T_2 = 2\sum_{j=1}^{k+1} X_j \log\{X_j/(np_j)\}$,
Freeman-Tukey statistic: $T_3 = 4\sum_{j=1}^{k+1}(\sqrt{X_j} - \sqrt{np_j})^2$.

By multivariate CLT one can show

$$\mathrm{P}(T_i < c) = G_k(c) + R \quad \textit{with} \quad R = O(n^{-1/2}),$$

where $G_k(x)$ is the distribution function of a chi-square random variable with k degrees of freedom. We would like to consider problem: can we get more precise result for order of R ?

Let $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n$ be i.i.d. random vectors in $\mathbf{R}^{k+1}$ with multinomial distribution with parameters 1, $p_1, p_2, \ldots, p_{k+1}$. Using $\boldsymbol{Y}_1 = (Y_{11}, \ldots, Y_{1,k+1})^T$ we define a random vector $\boldsymbol{Z}_1 = (Z_{11}, \ldots, Z_{1k})^T$ by the formula $Z_{1i} = Y_{1i} - p_i$ for $i = 1, 2, \ldots, k$. Then $\boldsymbol{Z}_1$ lies with probability 1 on the lattice

$$U = \left\{ m - p : m \text{ is an integer vector in } \mathbf{R}^k \right\},$$

$\boldsymbol{Z}_1$ has mean zero and its covariance matrix V is known.
Put $S_n = n^{-1/2}(\boldsymbol{Z}_1 + \ldots + \boldsymbol{Z}_n)$. Then

$$\mathrm{P}(T_1 < c) = \mathrm{P}(S_n^T V^{-1} S_n < c) = \mathrm{P}(S_n \in A),$$

where $A = \left\{ x = (x_1, \ldots, x_k)^T : x^T V^{-1} x < c \right\}$ is an ellipsoid.

Additional notation: $N(nc)$ – number of integer vectors m in the ellipsoid $(m - np)^T V^{-1}(m - np) < nc$; $V(nc)$ – the volume of this ellipsoid; $V(nc) = (\pi nc)^{k/2}|V|^{1/2}/\Gamma(k/2 + 1)$.

Theorem 1. Yarnold (1972) *If* $\mathrm{E}|\boldsymbol{Z}_1|^4 < \infty$ *then* $\mathrm{P}(S_n^T V^{-1} S_n < c) - G_k(c) = J_1 + O\left(n^{-1}\right)$ *uniformly in* c, *and*

$$J_1 = (N(nc) - V(nc))\frac{\exp(-c/2)}{(2\pi n)^{k/2}|V|^{1/2}} = O(n^{-k/(k+1)}).$$

Theorem 2. Götze and Ulyanov (2003) *If* $k : k \ge 5$ *then there exists a positive constant* $c_1(k)$ *depending on* k *only such that*

$$|J_1| \le \frac{c_1(k)}{n}\left(\frac{\sigma_1}{\sigma_k}\right)^{k+1}\left(1 + \log\frac{\sigma_1}{\sigma_k}\right),$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k$ *denote the eigenvalues of* V^{-1}.

Siotani and Fujikoshi (1984) applied the local asymptotic expansions to construct asymptotic expansions for the distribution functions of the log-likelihood ratio statistic and the Freeman-Tukey statistic. Later the results by Siotani and Fujikoshi were extended to the so-called power divergence family of statistics introduced in Cressie and Read (1984):

$$T_\lambda = 2(\lambda(\lambda+1))^{-1} \sum_{j=1}^{k+1} X_j \left[(X_j/(np_j))^\lambda - 1 \right],$$

where λ is a real number. However a new problem arises here.

It is connected with the fact that here the set A is defined as $A = \left\{ x = (x_1, \ldots, x_k)^T : T_2(x) < c \right\}$ or as $A = \{x : T_3(x) < c\}$, where e.g. for log-likelihood statistic

$$T_2(x) = 2 \sum_{j=1}^{k+1} (np_j + \sqrt{n}x_j) \log\{1 + x_j/(\sqrt{n}p_j)\}.$$

This means that comparing with situation for Pearson's test the set A is not exactly an ellipsoid but only "approximated" by ellipsoid. Therefore, one needs limit theorems for "almost" ellipsoids and results for lattice point problems in these cases.

Theorem (Huxley (1993)). *Let* Ω *be a convex Euclidean plane domain of area* V*, bounded by a simple closed curve* C*, composed of finitely many pieces* C_i*, which are three times continuously differentiable in the following sense: on each piece* C_i *the radius of curvature* ρ *is non-zero and continuously differentiable with respect to the tangent angle* ψ*. Let* M *be sufficiently large and let* $M\Omega$ *denote the set formed by expanding* Ω *linearly by a factor* M*. Then for any isometric embedding of* $M\Omega$ *in the Euclidean plane, the number of integer points* (m, n) *in* $M\Omega$ *is*

$$VM^2 + O\left(IM^{131/208} (\log M)^{18627/8320} \right).$$

Theorem. Assylbekov, Zubov and Ulyanov (2007).
For $k = 2$ *and any* λ *we have*

$$\mathrm{P}(T_\lambda < c) = G_2(c) + J_1(A^\lambda) + O(n^{-1})$$

with

$$J_1(A^\lambda) = \left(n^{-3/4+7/104} (\log n)^{18627/8320} \right).$$

Similar problems arise for approximations for the distributions of multinomial goodness-of-fit statistics under local alternatives (see Taneichi, Sekiya and Suzukawa (2002)).

In the above arguments we constructed asymptotic expansions for the distribution functions of the statistics applying local expansions. One of the problems in using local Edgeworth expansions lies in locating the lattice points near the boundary. However, with the availability of sufficient computing power this problem of tracking lattice points near the boundary can be tackled effectively especially for small and moderate sample sizes (see Bhattacharya and Chan (1996)).