Numerical Studies on Convergence of Multinomial Goodness-of-fit Statistics to Chisquare Distribution

Zh. Assylbekov
Graduate School of Science, Hiroshima University

October 1, 2008

Abstract

Let $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_k)'$ be a random vector with multinomial distribution. In this report we investigate numerically the convergence rate of so-called power divergence family of statistics $\{I^\lambda(\boldsymbol{Y}), \lambda \in \mathbb{R}\}$ introduced by Cressie and Read (1984) to chi-square distribution for $k = 4, 5, 6$.

1 Introduction

Let $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_k)'$ be a random vector with the multinomial distribution $M_k(n, \boldsymbol{\pi})$, i.e.,

$$\Pr(Y_1 = n_1, Y_2 = n_2, \ldots, Y_k = n_k) = \begin{cases} n! \prod_{j=1}^k \frac{\pi_j^{n_j}}{n_j!} & \sum_{j=1}^k n_j = n \\ 0 & \text{otherwise,} \end{cases}$$

where $n_j = 0, 1, \ldots, n$, $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_k)'$, $\pi_j > 0$, $\sum_{j=1}^k \pi_j = 1$. For testing the simple hypothesis H : $\boldsymbol{\pi} = \boldsymbol{p}$ ($\boldsymbol{p}$ is a fixed vector) against K : $\boldsymbol{\pi} \neq \boldsymbol{p}$ the power divergence statistics (introduced by Cressie and Read in [2]) can be used:

$$2nI^\lambda = \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^k Y_j \left(\left(\frac{Y_j}{np_j} \right)^\lambda - 1 \right), \ \lambda \in \mathbb{R},$$

where $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)'$, $p_j > 0$ $(j = 1, 2, \ldots, k)$ and $\sum_{j=1}^k p_j = 1$.

Throughout this paper we will use the following notation:

$$\boldsymbol{x} = (x_1, \dots, x_r)',$$
$$\boldsymbol{x}^* = (x_1, \dots, x_{l-1}, x_{l+1}, \dots, x_r)',$$

For any $B \subset \mathbb{R}^r$ and for any $l \in \{1, \dots, r\}$ denote

$$B_l = \{\boldsymbol{x}^* : \boldsymbol{x} \in B\}.$$

Definition 1. A set $B \subset \mathbb{R}^r$ is called an *extended convex set*, if B has the following representation for every $l \in \{1, 2, \dots, r\}$:

$$B = \{\boldsymbol{x} : \lambda_l(\boldsymbol{x}^*) < x_l < \theta_l(\boldsymbol{x}^*),\ \boldsymbol{x}^* \in B_l\},$$

where λ_l,θ_l are continuous functions on B_l.

It is known (see Cressie, Read [2]), that under the null hypothesis $2nI^\lambda$ has the chisquare distribution with $r = k - 1$ degrees of freedom in the limit. Moreover the distribution function of $2nI^\lambda$ has the following expansion:

$$\Pr\left(2nI^\lambda < c\right) = \Pr\left(\chi_r^2 < c\right) + J_2 + O\left(n^{-1}\right), \tag{1}$$

where

$$J_2 = -\frac{1}{\sqrt{n}} \sum_{l=1}^{r} n^{-\frac{r-l}{2}} \sum_{x_{l+1} \in L_{l+1}} \cdots \sum_{x_r \in L_r}$$
$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \chi_{B_l^\lambda}(\boldsymbol{x}^*) \left[S_1\left(\sqrt{n}x_l + p_l n\right) \phi(\boldsymbol{x})\right]_{\lambda_l(\boldsymbol{x}^*)}^{\theta_l(\boldsymbol{x}^*)} dx_1 \dots dx_{l-1}, \tag{2}$$

$$L_j = \{x_j : x_j = \frac{1}{\sqrt{n}}(n_j - np_j), n_j \in \mathbb{Z}\}, \tag{3}$$
$$S_1(x) = x - [x] - \frac{1}{2},$$
$$[h(\boldsymbol{x})]_{\lambda(\boldsymbol{x}^*)}^{\theta(\boldsymbol{x}^*)} = h(x_1, \dots, x_{l-1}, \theta_l(\boldsymbol{x}^*), x_{l+1}, \dots, x_r)$$
$$- h(x_1, \dots, x_{l-1}, \lambda_l(\boldsymbol{x}^*), x_{l+1}, \dots, x_r),$$
$$\phi(\boldsymbol{x}) = (2\pi)^{-\frac{r}{2}} |\Omega|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{x}'\Omega^{-1}\boldsymbol{x}\right),$$
$$\Omega = \mathrm{diag}(p_1, \dots, p_r) - (p_1, \dots, p_r)'(p_1, \dots, p_r).$$

Here $\chi_A(x)$ is an indicator function, $\theta_l(\boldsymbol{x}^*)$ and $\lambda_l(\boldsymbol{x}^*)$ are continuous functions from Definition 1 for the set

$$B^\lambda = \{\boldsymbol{x} : 2nI^\lambda(\boldsymbol{x}) < c\} \tag{4}$$

with

$$2nI^{\lambda}(\boldsymbol{x}) = \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^{k} (np_j + \sqrt{n}x_j) \left(\left(1 + \frac{x_j}{\sqrt{np_j}} \right)^{\lambda} - 1 \right), \tag{5}$$

$$x_k = -(x_1 + \cdots + x_r).$$

It follows from Yarnold's result [5] that

$$J_2 = O(n^{-1/2}).$$

Zubov and Ulyanov in [6] showed that

$$J_2 = O(n^{-1+\frac{1}{r+1}}).$$

This was improved by the author in [1], where it was shown that

$$J_2 = O\left(n^{-1+\mu(r)}\right),$$

with

$$\mu(r) = \begin{cases} 6/(7r+4) \text{ for } 3 \le r \le 7, \\ 5/(6r+2) \text{ for } r \ge 8. \end{cases}$$

In the present report we investigate numerically whether the upper bound for J_2 can be improved.

2 Preliminaries

By definition put

$$L = \left\{ \boldsymbol{x} : x_j = \frac{1}{\sqrt{n}}(m_j - np_j),\ m_j \in \mathbb{Z},\ j = \overline{1,r} \right\},$$

This means that L is an r-dimensional lattice in $\mathbb{R}^r$ and lattice spacing of L is $\frac{1}{\sqrt{n}}$. Let N^{λ} be the number of lattice points in the ellipsoid B^{λ}, i.e, $N^{\lambda} = \#(L \cap B^{\lambda})$. Let V^{λ} be the volume of B^{λ}.

Lemma 1. *Let J_2 be the term defined by (2); then*

$$J_2 = dn^{-\frac{r}{2}}(N^{\lambda} - n^{\frac{r}{2}}V^{\lambda}) + O(n^{-1}), \tag{6}$$

where $d = const > 0$.

Proof. The proof is given in [1], Proposition 1. □

3 Numerical studies

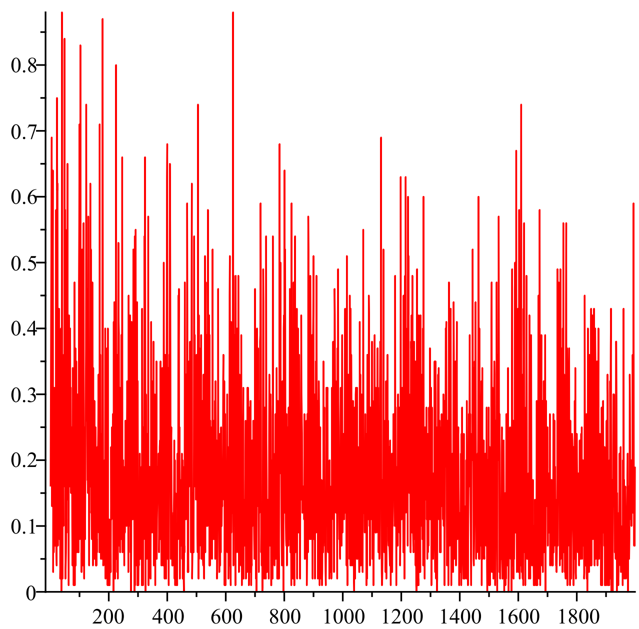
In [1] it was shown that B^λ is a convex boundy which has smooth boundary with nonvanishing and bounded Gaussian curvature throughout. Hence Lemma 1 reduces the original problem of estimating J_2 (see (2)) to the lattice point problem inside convex body B^λ (see (6)). In [4], W. Müller made a conjecture, which in terms of our problem reads

$$n^{-\frac{r}{2}}(N^\lambda - n^{\frac{r}{2}}V^\lambda) = \begin{cases} O(n^{-1+\varepsilon}), & r = 3, 4, \\ O(n^{-1}), & r \geq 5. \end{cases}$$

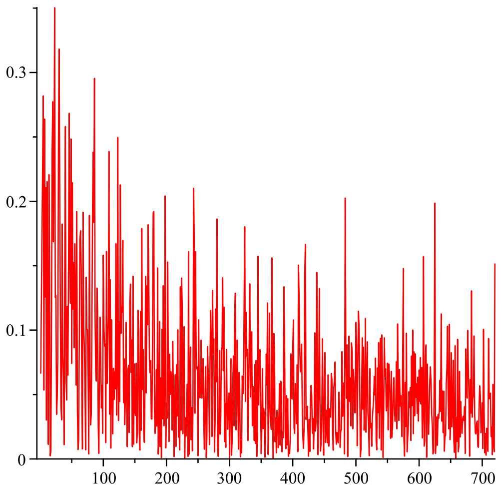
In order to test this conjecture we calculate the expressions

$$\begin{cases} n^{-\frac{r}{2}} \left| N^\lambda - n^{\frac{r}{2}}V^\lambda \right| n^{0.9}, & r = 3, 4, \\ n^{-\frac{r}{2}} \left| N^\lambda - n^{\frac{r}{2}}V^\lambda \right| n, & r = 5. \end{cases} \tag{7}$$

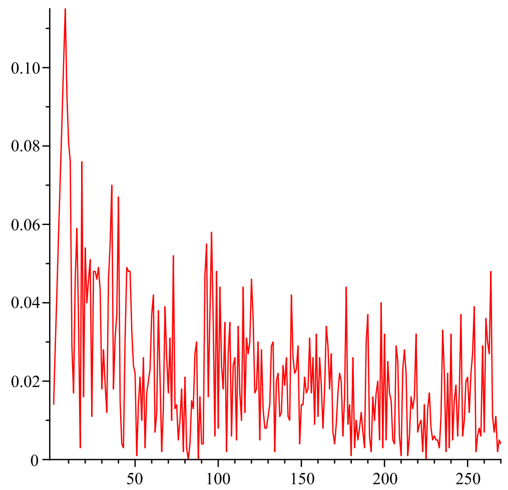
If Müller's conjecture is true then the expressions in (7) should be bounded by above. The results of our computations for $r = 3$ and $n = \overline{1, 2000}$ are given below



For $r = 4$ and $n = \overline{1,720}$ we have



And, finally, for $r = 5$ and $n = \overline{1,270}$ we have



As it is seen on these pictures, our results fit Müller's conjecture. The algorithm of computations is rather straightforward and is given for $r = 3$ in Appendix A.

A Algorithm of computations in C

```
#include <stdio.h>
#include <stdlib.h>
#include <math.h>
#define C 1.0
#define PI 3.14159265358979323846
#define K 4

double T(int n, double *x, double *p)
{
  double result = 0;
  int j;
  x[K-1] = 0;
  for (j = 0; j < K - 1; j++)
  {
    result += 2 * (n*p[j] + sqrt(n)*x[j])*log(1 + x[j]/sqrt(n)/p[j]);
    x[K-1] -= x[j];
  }
  result += 2 * (n*p[K-1] + sqrt(n)*x[K-1])*log(1 + x[j]/sqrt(n)/p[K-1]);
  return result;
}

int N(int n, double *p)
{
  double x[K];
  int i, j, k, count = 0;
  for (i = 1; i < n; i++)
    for (j = 1; j < n - i; j++)
      for (k = 1; k < n - i - j; k++)
      {
        x[0] = 1/sqrt(n)*(i - n*p[0]);
        x[1] = 1/sqrt(n)*(j - n*p[1]);
        x[2] = 1/sqrt(n)*(k - n*p[2]);
        if (T(n,x,p) < C)
        {
          count++;
        }
      }
  return count;
}
```

```
double V(int n, double *p)
{
  return 4*pow(PI*C, 3.0/2)*sqrt(p[0]*p[1]*p[2]*p[3])/3/sqrt(PI);
}

int main(int argc, char *argv[])
{
  double p[K], volume;
  int n;

  p[0] = 0.1;
  p[1] = 0.1;
  p[2] = 0.3;
  p[K-1] = 0.5;

  volume = V(n,p);

  for(n = 1; n < 1000; n++)
  {
    printf("%f\n", (pow(n, -3.0/2)*N(n, p)-volume));
    fprintf(stderr, "%d ", n);
  }

  return 0;
}
```

References

[1] Zh. Assylbekov, Convergence of multinomial goodness-of-fit statistics to chisquare distribution, submitted.

[2] N. C. Cressie and T. R. C. Read, Multinomial goodness-of-fit tests, J. R. Statist. Soc. B (1984) 46, No. 3, 440-464.

[3] E. Hlawka, Über Integrale auf konvexen Körpern I, II, Monatsh. Math. **54**, 1-36, 81-99 (1950).

[4] W. Müller, Lattice points in large convex bodies, Mh. Math. 128, 315–330 (1999).

[5] J. K. Yarnold, Asymptotic approximations for the probability that a sum of lattice random vectors lies in a convex set, The Annals of Mathematical Statistics 1972, Vol. 43, No. 5, 1566–1580.

[6] V. N. Zubov, V. V. Ulyanov, Refinement on the convergence of one family of goodness-of-fit statistics to chi-squared distribution, submitted.