Lietuvos matematikos rinkinys. LMD darbai ISSN 0132-2818
52 tomas, 2011, 327–331 www.mii.lt/LMR/

Inconsistency of χ^2 test for sparse categorical data under multinomial sampling

Pavel Samusenko

Faculty of Fundamental Sciences, Vilniaus Gedimino Technical University
Saulėtekio 11, LT-10223 Vilnius
E-mail: pavels.vgtu@gmail.com

Abstract. Simple conditions for the inconsistency of Pearson's χ^2 test in case of very sparse categorical data are given. The conditions illustrate the phenomenon of "reversed consistency": the greater deviation from the null hypothesis the less power of the test.

Keywords: χ^2 test, categorical data, inconsistency, sparse contingency table.

1 Introduction

Statistical inference problems caused by sparsity of contingency tables are widely discussed in the literature. According to a rule of thumb, expected (under the null hypothesis) frequencies in a contingency table is required to exceed 5 in the majority of its cells [6]. If this condition is violated, the χ^2 approximation of Pearson's χ^2 test statistic may be inaccurate and the contingency table is said to be *sparse*.

There is a vast literature dealing with approximation problems resulting from the sparsity, see, e.g., [1, 7, 8, 2, 3] and references therein). In this paper, it is shown that, for very sparse categorical data, the χ^2 test can become completely uninformative (inconsistent) and hence there is no sense to approximate or adjust its distribution. For the likelihood ratio test, analogous results are presented in [4] and [5].

In the next section we introduce notation, present some background and specify a sparsity condition. The inconsistency of Pearson's χ^2 test is proved in Section 3. A simple example and simulation results provided in the last section illustrate the inconsistency and "reversed consistency" phenomena for a finite sample.

2 Notation and background

Let y_j denote an observed frequency of the category $j \in J = J_n := \{1, \ldots, n\}$ in a sample of N iid observations. Hence $Y := (y_1, \ldots, y_n) \sim \mathit{Multinomial}_n(N, P)$ where $P := (p_1, \ldots, p_n) \in \mathcal{P}$,

$$\mathcal{P} := \left\{ q \in \mathbf{R}^n \colon q_j \geqslant 0,\ j = 1, \ldots, n,\ \sum_{i=1}^{n} q_i = 1 \right\}.$$

Let us assume that a simple hypothesis

$$H_0\colon P = P_0 \quad \text{versus} \quad H_1\colon P \neq P_0$$

is to be tested on the basis of the observed frequencies Y with a given $P_0 = (p_1^0, \ldots, p_n^0) \in \mathcal{P}_+ := \{q \in \mathcal{P},\ q_i > 0,\ \forall i\}$.

We consider very sparse categorical data (contingency tables). Here it means that

$$n = n(N), \qquad N = o(n), \qquad P = P(N), \qquad P_0 = P_0(N) \quad (N \to \infty).$$

We shall also use additional (technical) conditions related to the sparseness, see Proposition 1.

Perason's χ^2 statistic

$$e\chi^2 := \sum_{j\in J} \frac{(y_j - Np_j^0)^2}{Np_j^0} = \sum_{j\in J} \frac{y_j^2}{Np_j^0} - N. \tag{1}$$

Using moment generation function one can find the means

$$\mathbf{E}\chi^2 = (N-1)\sum_{j\in J} \frac{p_j^2}{p_j^0} + \sum_{j\in J} \frac{p_j}{p_j^0} - N, \qquad \mathbf{E}_0\chi^2 = n-1, \tag{2}$$

and the variances

$$\begin{aligned} \mathbf{D}\chi^2 \quad = \quad & \frac{1}{N}\sum_{j\in J} \frac{p_j}{(p_{0j})^2} + 6\Big(1-\frac{1}{N}\Big)\sum_{j\in J}\Big(\frac{p_j}{p_{0j}}\Big)^2 \\ & + 4N\Big(1-\frac{1}{N}\Big)\Big(1-\frac{2}{N}\Big)\sum_{j\in J}\frac{(p_j)^3}{(p_{0j})^2} \\ & - \frac{1}{N}\Big(\sum_{j\in J}\frac{p_j}{p_{0j}}\Big)^2 - \Big(1-\frac{1}{N}\Big)\Big(\sum_{j\in J}\frac{p_j}{p_{0j}}\Big)\Big(\sum_{i\in J}\frac{(p_i)^2}{p_{0i}}\Big) \\ & - \Big(4N-6\Big)\Big(1-\frac{1}{N}\Big)\Big(\sum_{j\in J}\frac{(p_j)^2}{p_{0j}}\Big)^2, \end{aligned} \tag{3}$$

$$\mathbf{D}_0\chi^2 = \frac{1}{N}\sum_{j\in J}\frac{1}{p_{0j}} - \frac{n^2}{N} + 2(n-1)\Big(1-\frac{1}{N}\Big) \tag{4}$$

of the χ^2 statistic. Here and in the sequel $\mathbf{E}, \mathbf{D}$, and $\mathbf{P}$ ($\mathbf{E}_0, \mathbf{D}_0$, and $\mathbf{P}_0$) denote, respectively, the expectation, the variance, and the probability for $Y \sim \mathit{Multinomial}_n(N, P)$ (respectively, $Y \sim \mathit{Multinomial}_n(N, P_0)$).

3 Inconsistency

In this section the inconsistency of the χ^2 statistic is derived under additional conditions related to and quite natural for (very) sparse categorical data.

Definition 1. Let $T_N := T(S_N)$ be a statistic of a sample S_N with N being the sample size. A test (criterion) based on the statistic T_N is said to be *consistent* for testing H_0 versus H_1 if there exists a sequence c_N such that

$$\mathbf{P}_0\{T_N > c_N\} + \mathbf{P}\{T_N < c_N\} \to 0, \quad N \to \infty.$$

Otherwise, the test is called *inconsistent*.

Proposition 1. *Suppose that*

$$\Delta_N := \mathbf{E}\chi^2 - \mathbf{E}_0\chi^2 = \sum_{j\in J}\frac{p_j}{p_j^0} + (N-1)\sum_{j\in J}\frac{p_j^2}{p_j^0} - N - (n-1) < 0, \tag{5}$$

and the asymptotic relation

$$\rho_N^2 := \frac{\Delta_N^2}{D_N^2} \to \infty \quad (N\to\infty) \tag{6}$$

is valid with $D_N := \sqrt{\mathbf{D}_0\chi^2} + \sqrt{\mathbf{D}\chi^2}$. *Then the* χ^2 *test is inconsistent.*

On the other hand, the test based on the statistic $T_N := |\chi^2 - (n-1)|$ *is consistent with* $c_N = |\Delta_N|/2$ *provided* (6) holds.

Proof. The Tchebyshev's inequality implies

$$\mathbf{P}_0\bigl\{\chi^2 \le \mathbf{E}_0\chi^2 - 2\sqrt{\mathbf{D}_0\chi^2}\,\bigr\} \leqslant 1/4, \tag{7}$$

$$\mathbf{P}\bigl\{\chi^2 \geqslant \mathbf{E}\chi^2 + 2\sqrt{\mathbf{D}\chi^2}\,\bigr\} \leqslant 1/4. \tag{8}$$

Consequently,

$$\begin{aligned}&\mathbf{P}_0\bigl\{\chi^2 > c_N\bigr\} + \mathbf{P}\bigl\{\chi^2 < c_N\bigr\}\\ &\quad\geqslant \mathbf{P}_0\bigl\{\chi^2 > \max(c_N, c_{0N})\bigr\} + \mathbf{P}\bigl\{\chi^2 < \min(c_N, c_{1N})\bigr\}\end{aligned} \tag{9}$$

where $c_{0N} := \mathbf{E}_0\chi^2 - 2\sqrt{\mathbf{D}_0\chi^2}$ and $c_{1N} := \mathbf{E}\chi^2 + 2\sqrt{\mathbf{D}\chi^2}$. Since, in view of (5) and (6),

$$c_{1N} - c_{0N} = \Delta_N + 2\bigl(\sqrt{\mathbf{D}_0\chi^2} + \sqrt{\mathbf{D}\chi^2}\,\bigr) < 0$$

for all sufficiently large N, we then get $c_{0N} \geqslant c_{1N}$ and hence either $\max(c_N, c_{0N}) = c_{0N}$ or $\min(c_N, c_{1N}) = c_{1N}$. From (7), (8) and (9) we derive inconsistency of χ^2 test:

$$\mathbf{P}_0\bigl\{\chi^2 > c_N\bigr\} + \mathbf{P}\bigl\{\chi^2 < c_N\bigr\} \geqslant \max\bigl(\mathbf{P}_0\bigl\{\chi^2 > c_{0N}\bigr\}, \mathbf{P}\bigl\{\chi^2 < c_{1N})\bigr\}\bigr) \geqslant 3/4.$$

The consistency of T_N follows from (2) and the Tchebyshev inequality:

$$\begin{aligned}&\mathbf{P}_0\bigl\{T_N > |\Delta_N|/2\bigr\} + \mathbf{P}\bigl\{T_N < |\Delta_N|/2\bigr\}\\ &\quad\leqslant \mathbf{P}_0\bigl\{T_N^2 > \Delta_N^2/4\bigr\} + \mathbf{P}\bigl\{\bigl|\chi^2 - \mathbf{E}_P\chi^2\bigr| > |\Delta_N|/2\bigr\}\\ &\quad\leqslant 4\frac{\mathbf{E}_0 T_N^2 + \mathbf{D}\chi^2}{\Delta_N^2} = \frac{4}{\rho_N^2} \to 0 \quad (N\to\infty)\end{aligned}$$

due to (6).

Proposition 1 shows that (5) is the key condition which determines the inconsistency of χ^2 test. When P_0 is the uniform distribution, $\Delta \geqslant 0$ for any P and hence, for any P, condition (5) is not satisfied. In the next section we present a simple example when conditions (5) and (6) are fulfilled.

Remark 1. By definition (5)

$$\Delta_N = \sum_{j\in J} \frac{p_j - p_j^0}{p_j^0} + (N-1)\sum_{j\in J} \frac{(p_j - p_j^0)^2}{p_j^0}. \tag{10}$$

Since the second term in this expression is nonnegative the requirement $\Delta < 0$ implies that the absolute value of the first term in (10) should dominate second one.

4 Example

For a given $\beta > 1$ and $q_0, q \in (0, 1/2)$, set $m = [N^\beta]$, $n = 2m$, $j_0 = m$,

$$p_j^0 = q_0/m, \quad \forall j \leqslant m, \qquad p_j^0 = (1-q_0)/m, \quad \forall j > m,$$
$$p_j = q/m, \quad \forall j \leqslant m, \qquad p_j = (1-q)/m, \quad \forall j > m.$$

Then the conditions of Proposition 1 are fulfilled.

If $q = 0$, means (2) and variances (4), (3) are given by

$$\mathbf{E}\chi^2 = \frac{N-1}{1-q_0} + \frac{m}{1-q_0} - N, \qquad \mathbf{E}_0\chi^2 = n-1,$$
$$\mathbf{D}_0\chi^2 = \frac{m^2}{Nq_0(1-q_0)} - \frac{n^2}{N} + 2(n-1)\left(1 - \frac{1}{N}\right),$$
$$\mathbf{D}\chi^2 = \frac{2(m-1)}{(1-q_0)^2}\left(1 - \frac{1}{N}\right).$$

Consequently,

$$\Delta_N = -\frac{1-2q_0}{2-2q_0} n + \mathcal{O}(N),$$

$\mathbf{D}\chi^2 = \mathcal{O}(n)$, and

$$\mathbf{D}_{p^0}\chi^2 = \frac{n^2}{N}\left[\frac{1}{4q_0(1-q_0)} - 1\right] + \mathcal{O}(n).$$

Thus

$$\rho_N = -\sqrt{\frac{Nq_0}{1-q_0}}\left[1 + \mathcal{O}\left(\frac{N}{n}\right)\right].$$

A computer experiment illustrates the asymptotic findings in case of finite samples. In the simulations, the number of observations $N = 200$, the number of cells $n = 2m = 600$. Two cases are considered: (a) $q_0 = 0.2$, $q = 0$ and (b) $q_0 = 0.2$, $q = 0.1$. The number of repetitions is set to 100. The histograms of the χ^2 statistic for the null hypothesis H_0 and the alternative H_1 are represented in Fig. 1.

The figure clearly demonstrates the inconsistency of the χ^2 statistic. Actually, in the first case (case (a)), the phenomenon of the "*reversed consistency*" is observed: although the values of the χ^2 statistic under the null hypothesis H_0 are significantly greater than its values under the alternative H_1 (the data under the alternative "fits" the null hypothesis better than the data under the null hypothesis itself) the latter is evidently separable from the former. Thus Pearson's χ^2 test is completely uninformative in this case.

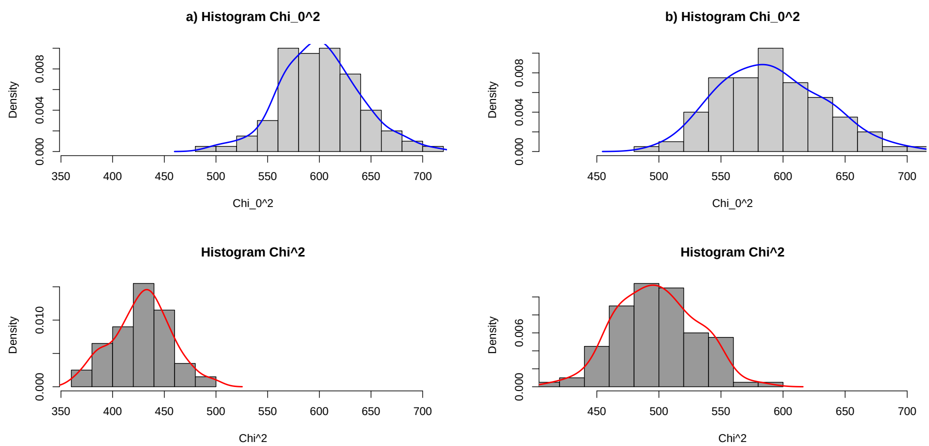

Fig. 1. Histograms of the χ^2 statistic under the null hypothesis and under the alternative in case (a) $q_0 = 0.2$, $q = 0$ and in case (b) $q_0 = 0.2$, $q = 0.1$.

References

[1] A. Agresti. *Categorical Data Analysis*. Wiley & Sons, New York, 1990.

[2] A. Agresti and B.D. Hitchcock. Bayes inference for categorical data analysis. *Stat. Meth. Appl.*, **14**:297–330, 2005.

[3] O. Kuss. Global goodness-of-fit tests in logistic regression with sparse data. *Stat. Med.*, **21**:3789–3801, 2002.

[4] M. Radavičius and P. Samusenko. Profile statistics for sparse contingency tables. *CDAM*, **40**:115–123, 2010.

[5] M. Radavičius and P. Samusenko. Profile statistics for sparse contingency tables under poisson sampling. *Austrian J. Stat.*, **40**:115–123, 2010.

[6] C.R. Rao. *Linear Statistical Inference and its Applications*. John Wiley, New York, 1965.

[7] J.S. Simonoff. Smoothing categorical data. *J. Stat. Plann. Inf.*, **47**:41–69, 1995.

[8] M. von Davier. Bootstraping goodness-of-fit statistics for sparse categorical data. Results of a Monte Carlo study. *Meth. Psych. Res.*, **2**(2), 1997. Available from Internet: www.pabst-publishers.de/mpr.

REZIUMĖ

χ^2 testo nepagrįstumas išsklaidytiems kategoriniams duomenims polinominio ėmimo atveju

P. Samusenko

Pateiktos paprastos Pearsono χ^2 testo nepagrįstumo sąlygos labai išsklaidytų kategorinių duomenų atveju. Tos sąlygos iliustruoja „atvirkščio pagrįstumo" reiškinį: kuo didesnis alternatyvos ir nulines hipotezes skirtumas tuo mažesnė χ^2 testo galia.

Raktiniai žodžiai: χ^2 testas, kategoriniai duomenys, nepagrįstumas, išsklaidyta dažnių lentelė.