Statistics and Probability Letters 119 (2016) 317–325

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Double asymptotics for the chi-square statistic

Grzegorz A. Rempała [a,*], Jacek Wesołowski [b]

[a] Division of Biostatistics and Mathematical Biosciences Institute, The Ohio State University, 43210 Columbus, OH, USA
[b] Wydział Matematyki i Nauk Informacyjnych, Politechnika Warszawska, Warsaw, Poland

ARTICLE INFO

Article history:
Received 26 April 2016
Received in revised form 3 September 2016
Accepted 4 September 2016
Available online 10 September 2016

Keywords:
Pearson chi-square statistic
Central limit theorem
Poisson limit theorem
Weak convergence

ABSTRACT

We consider distributional limit of the Pearson chi-square statistic when the number of classes m_n increases with the sample size n and $n/\sqrt{m_n} \to \lambda$. Under mild moment conditions, the limit is Gaussian for $\lambda = \infty$, Poisson for finite $\lambda > 0$, and degenerate for $\lambda = 0$.

1. Preliminaries

The Pearson chi-square statistic is probably one of the best-known and most important objects of statistical science and has played a major role in statistical applications ever since its first appearance in Karl Pearson's work on "randomness testing" (Pearson, 1900). The standard test for goodness-of-fit with the Pearson chi-square statistic tacitly assumes that the support of the discrete distribution of interest is fixed (whether finite or not) and unaffected by the sampling process. However, this assumption may be unrealistic for modern 'big-data' problems which involve complex, adaptive data acquisition processes (see, e.g., Grotzinger et al., 2014 for an example in astro-biology). In many such cases the associated statistical testing problems may be more accurately described in terms of triangular arrays of discrete distributions whose finite supports are dependent upon the collected samples and increase with the samples' size (Pietrzak et al., 2016). Motivated by 'big-data' applications, in this note we establish some asymptotic results for the Pearson chi-square statistic for triangular arrays of discrete random variables for which their number of classes m_n grows with the sample size n. Specifically, let $X_{n,k}$, $k = 1, \ldots, n$, be i.i.d. random variables having the same distribution as X_n, where

$$\mathbb{P}(X_n = i) = p_n(i) > 0, \quad i = 1, 2, \ldots, m_n < \infty, \ n = 1, 2, \ldots.$$

Recall that the standard Pearson chi-square statistic is defined as

$$\chi_n^2 = n \sum_{i=1}^{m_n} \frac{\left(\hat{p}_n(i) - p_n(i)\right)^2}{p_n(i)}, \tag{1}$$

where the empirical frequencies $\hat{p}_n(i)$ are

$$\hat{p}_n(i) = n^{-1} \sum_{k=1}^{n} I(X_{n,k} = i), \quad i = 1, \ldots, m_n.$$

* Corresponding author.
E-mail address: rempala.3@osu.edu (G.A. Rempała).

http://dx.doi.org/10.1016/j.spl.2016.09.004
0167-7152/

As stated above, in what follows we will be interested in the *double asymptotic* analysis of the weak limit of χ_n^2, that is, the case when $m_n \to \infty$ as $n \to \infty$.

Observe that χ_n^2 given in (1) can be decomposed into a sum of two uncorrelated components as follows

$$\chi_n^2 = n^{-1}(U_n + S_n) - n, \tag{2}$$

where

$$U_n = \sum_{1\le k\ne l\le n} \frac{I(X_{n,k} = X_{n,l})}{p_n(X_{n,k})} \tag{3}$$

and

$$S_n = \sum_{k=1}^{n} \frac{1}{p_n(X_{n,k})} = \sum_{k=1}^{n} p_n^{-1}(X_{n,k}). \tag{4}$$

The second equality above introduces notational convention we use throughout. Note that for fixed n the statistic S_n is simply a sum of i.i.d. random variables and U_n is an unnormalized U-statistic (see, e.g., Korolyuk and Borovskich, 2013). It is routine to check that

$$\mathbb{E}\, U_n = n(n-1) \quad \text{and} \quad \mathbb{E}\, S_n = nm_n$$

and consequently

$$\mathbb{E}\, \chi_n^2 = m_n - 1.$$

Moreover, since we also have $\mathbb{C}\mathrm{ov}(U_n, S_n) = 0$, it follows that

$$\mathbb{V}\mathrm{ar}\, \chi_n^2 = n^{-2}(\mathbb{V}\mathrm{ar}\, S_n + \mathbb{V}\mathrm{ar}\, U_n) = n^{-1}[\mathbb{V}\mathrm{ar}\, p_n^{-1}(X_n) + 2(n-1)(m_n - 1)].$$

When $m_n = m$ is a constant then the classical result (see, e.g., Shao, 2003, chapter 6) implies that the statistic χ_n^2 asymptotically follows the χ^2-distribution with $(m-1)$ degrees of freedom. Consequently, when m is large the standardized statistic $(\chi_n^2 - (m-1))/\sqrt{2(m-1)}$ may be approximated by the standard normal distribution. However, in the case when $m_n \to \infty$ as $n \to \infty$ the matters appear to be more subtle and the above normal approximation may or may not be valid depending upon the asymptotic relation of m_n and n, as described below. Since S_n is a sum of i.i.d. random variables, the case when S_n contributes to the limit of normalized χ_n^2 may be largely handled with the standard theory for arrays of i.i.d. variables. Consequently, we focus here on a seemingly more interesting case when the asymptotic influence of U_n dominates over that of S_n. Specifically, throughout the paper we assume that as n, $m_n \to \infty$

(C) $(m_n n)^{-1}\mathbb{V}\mathrm{ar}\, p_n^{-1}(X_n) \to 0.$

Note that (C) implies $n^{-1}(S_n - nm_n)/\sqrt{2m_n} \to 0$ in probability and, in particular, is trivially satisfied when X_n is a uniform random variable on the integer lattice $1, \ldots, m_n$, that is, when $p_n(i) = m_n^{-1}$ for $i = 1 \ldots, m_n$. Under condition (C) we get a rather complete picture of the limiting behavior of χ_n^2. Our main results are presented in Section 2 where we discuss the Poissonian and Gaussian asymptotics. Some examples, relations to asymptotics known in the literature and further discussions are provided in Section 3. The basic tools used in our derivations are listed in the appendix. In what follows limits are taken as $n \to \infty$ with $m_n \to \infty$ and $\xrightarrow{d}$ stands for convergence in distribution.

2. Poissonian and Gaussian asymptotics

We start with the case when a naive normal approximation for the standardized χ_n^2 statistic fails. Indeed, as it turns out, when m_n is asymptotically of order n^2, we have the following Poisson limit theorem for χ_n^2.

Theorem 2.1. *Assume that the condition* (C) *holds, as well as*

$$\frac{n}{\sqrt{m_n}} \to \lambda \in (0, \infty). \tag{5}$$

Then

$$\frac{\chi_n^2 - m_n}{\sqrt{2m_n}} \xrightarrow{d} \frac{\sqrt{2}}{\lambda} Z - \frac{\lambda}{\sqrt{2}}, \quad Z \sim \mathrm{Pois}\left(\frac{\lambda^2}{2}\right). \tag{6}$$