Advances and Applications in Statistics
© 2017 Pushpa Publishing House, Allahabad, India
http://www.pphmj.com
http://dx.doi.org/10.17654/AS050060469
Volume 50, Number 6, 2017, Pages 469-476 **ISSN: 0972-3617**

THE ASYMPTOTIC DISTRIBUTION OF THE PEARSON CHI-SQUARE STATISTIC FOR THE 2×2 TABLE

David Sotres-Ramos and Yoni Castillo-Tzec

Programa de Estadística
Colegio de Postgraduados
Carr. México-Texcoco, Km. 36.5
Estado de México, México

Abstract

The Pearson chi-square statistic (Q_p) is frequently used for testing hypothesis for data in a 2×2 table which arises from several different and relevant sampling frameworks. In this paper, a detailed proof is given that asymptotically Q_p has a chi-square distribution with one degree of freedom.

Introduction

Frequently, the categorical data from different type of experiments is summarized in a 2×2 contingency table like Table 1.

Received: February 16, 2017; Revised: April 12, 2017; Accepted: May 8, 2017
2010 Mathematics Subject Classification: 62-XX.
Keywords and phrases: 2×2-table, Pearson chi-square, asymptotic distribution.

Table 1. 2×2 Contingency table

	Yes	No	Total	Proportion Yes
Group 1	n_{11}	n_{12}	$n_{1.}$	$p_1 = n_{11}/n_{1.}$
Group 2	n_{21}	n_{22}	$n_{2.}$	$p_2 = n_{21}/n_{2.}$
Total	$n_{.1}$	$n_{.2}$	N	

The statistics (Q_p) is defined as follows:

$$Q_p = \frac{(p_1 - p_2)^2}{\frac{n_{.1}n_{.2}}{n_{1.}n_{2.}N}}. \tag{1}$$

Frequently the chi-square statistic Q_p is used for testing different type of hypothesis for data in a 2×2 table which arises from several different sampling frameworks like in the following investigations.

(i) Simple independent random samples of two cities, asking residents whether they desire more environmental regulations. Because interest lies in whether the proportion favoring regulations is the same for the two cities, the hypothesis of interest is: Is the distribution of the response the same in both groups?

The null and alternative hypotheses are as follows:

$$H_0 : \pi_1 = \pi_2 \quad \text{vs} \quad H_a : \pi_1 \neq \pi_2, \tag{2}$$

where π_1 and π_2 are the population proportions favoring regulations for the two cities, respectively, and Q_p is the appropriate test statistic.

(ii) In a clinical trial, patients are randomly allocated to one of two drug treatments (test and placebo), and their response to that treatment is a binary outcome. The question of interest is whether the rates of favorable response for test and placebo are the same. The null hypothesis is stated

H_0: There is no association between treatment and outcome.

There are several ways of testing this hypothesis; many of the tests are based on Q_p. However, sometimes the counts in the table cells are too small to meet the sample size requirements necessary for the chi-square distribution to apply, and exact methods based on the hypergeometric distribution are used to test the hypothesis of no association.

(iii) A simple random sample of 500 persons is questioned regarding political affiliation and attitude toward an energy-rationing program. From the data of this study, we wish to find answers to the following question: Do the data indicate that that the pattern of opinion is independent of political affiliation?

In this case, the simple random sample produces a multinomial distribution. The null hypothesis of independence is

$$H_0 : p_{ij} = p_{i0} \times p_{0j}, \text{ for all cells } (i, j),$$

where these parameters are the population probabilities of the following table:

Table 2. Population probabilities

	B_1	B_2	Row total
A_1	p_{11}	p_{12}	p_{10}
A_2	p_{21}	p_{22}	p_{20}
Column total	p_{01}	p_{02}	1

and where

$p_{ij} = P(A_i \cap B_j)$ probability of the joint occurrence of A_i and B_j.

$p_{i0} = P(A_i)$ total probability in the *i*th row.

$p_{0j} = P(B_i)$ total probability in the *j*th row.

For these and other statistical applications, Q_p is the appropriate test statistic, see for instance Seber [1], Krishnamoorthy [2], Fleiss et al. [3], Newcombe [4], Ott and Longneker [5], Plichta and Kelvin [6] and Stokes et al. [7].

To apply the chi-square statistic Q_p for testing hypothesis, in large samples, it is necessary to know the asymptotic distribution of Q_p. In this paper, a detailed proof is given that asymptotically Q_p has a chi-square distribution with one degree of freedom.

Asymptotic distribution of Q_p

Let us assume that we have two simple random samples, from two Bernoulli distributions with parameters π_1 and π_2, respectively. Therefore, we have n_1 independent and identically distributed random variables $X_1, X_2, ..., X_{n_1}$ which are distributed as

$$Bernoulli(\pi_1) \tag{3}$$

also we have n_2 independent and identically distributed random variables $Y_1, Y_2, ..., Y_{n_2}$ which are distributed as

$$Bernoulli(\pi_2). \tag{4}$$

Frequently, the data from this experiment is summarized in a 2×2 contingency table, like in Table 1, where $n_{11} = \sum_{i=1}^{n_{1.}} X_i$ and $n_{21} = \sum_{i=1}^{n_{2.}} Y_i$.

In what follows we will prove that asymptotically Q_p has a chi-square distribution with one degree of freedom. First we prove the following lemma.

Lemma 1. *Suppose we have two random samples satisfying* (3) *and* (4) *with equal sample sizes* $n_{1.} = n_{2.} = n$, *then under the null hypothesis* $H_0 : \pi_1 = \pi_2 = \pi$

$$W = \frac{(p_1 - p_2)}{\left[p(1-p)\left(\frac{1}{n} + \frac{1}{n}\right)\right]^{1/2}} \xrightarrow{L} N(0, 1), \tag{5}$$

where

$$p = \frac{n_{11} + n_{21}}{n_{1.} + n_{2.}} = \frac{\sum^{n} X_i + \sum^{n} Y_i}{2n}$$

and $N(0, 1)$ *is the standard normal random variable*.

Proof. Clearly, under $H_0 : \pi_1 = \pi_2 = \pi$

$$E(X_i) = \pi; \quad Var(X_i) = \pi(1 - \pi) \text{ and } \sigma(X_i) = [\pi(1 - \pi)]^{1/2}$$

and

$$E(Y_i) = \pi; \quad Var(Y_i) = \pi(1 - \pi) \text{ and } \sigma(Y_i) = [\pi(1 - \pi)]^{1/2}.$$

Therefore, under H_0, $X_1 - Y_1, X_2 - Y_2, \ldots, X_n - Y_n$ are independent and identically distributed random variables with

$$E(X_i - Y_i) = 0 \quad Var(X_i - Y_i) = 2\pi(1 - \pi) \text{ and}$$

$$\sigma = \sigma(X_i - Y_i) = [2\pi(1 - \pi)]^{1/2}$$

applying the central limit theorem, see for instance Hogg and Craig [8], we have that

$$U_n = \sqrt{n}[\sum (X_i - Y_i)/n\sigma] = \sqrt{n}(p_1 - p_2)/[2\pi(1 - \pi)]^{1/2} \xrightarrow{L} N(0, 1). \quad (6)$$

On the other hand, $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_n$ are $2n$ independent, identically distributed random variables with Bernoulli distribution and parameter $\pi = E(X_i) = E(Y_i)$.

Thus, by the weak law of large numbers, see Lehmann [9]

$$p = \frac{\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i}{2n} \xrightarrow{p} \pi. \quad (7)$$

Note that, $g(x) = [x(1-x)]^{1/2}$ is a continuous function. Thus

$$[p(1-p)]^{1/2} \xrightarrow{p} [\pi(1-\pi)]^{1/2}.$$

Thus,

$$W_n = \frac{[p(1-p)]^{1/2}}{[\pi(1-\pi)]^{1/2}} \xrightarrow{p} 1. \tag{8}$$

From (6) and (8) we obtain that

$$\frac{U_n}{W_n} = \frac{p_1 - p_2}{\left[\frac{2}{n} p(1-p)\right]^{1/2}} \xrightarrow{L} N(0, 1),$$

which proves Lemma 1.

Theorem 1. *Suppose we have two random samples satisfying* (3) *and* (4) *and with equal sample sizes* $n_{1.} = n_{2.} = n$, *then under the null hypothesis* $H_0 : \pi_1 = \pi_2 = \pi$

$$Q_p \xrightarrow{L} \chi_1^2, \tag{9}$$

where χ_1^2 *is a chi-square random variable with one degree of freedom.*

Proof. By Lemma 1, we have that

$$\lim_{n\to\infty} P\{W \leq y\} = \Phi(y), \quad \forall y \in R, \tag{10}$$

where Φ is the cumulative distribution function of the standard normal random variable, and

$$W = \frac{p_1 - p_2}{\left[p(1-p)\left(\frac{1}{n} + \frac{1}{n}\right)\right]^{1/2}}.$$

Since $n_{1.} = n_{2.} = n,$ we have that

$$W = \frac{p_1 - p_2}{\left[p(1-p)\left(\frac{1}{n_{1.}} + \frac{1}{n_{2.}}\right)\right]^{1/2}}. \tag{11}$$

Note that

$$p = \frac{\sum X_i + \sum Y_i}{2n} = \frac{n_{11} + n_{21}}{N} = \frac{n_{.1}}{N}$$

and

$$1 - p = 1 - \frac{n_{.1}}{N} = \frac{N - n_{.1}}{N} = \frac{n_{.2}}{N}.$$

Thus,

$$p(1-p)\left(\frac{1}{n_{1.}} + \frac{1}{n_{2.}}\right) = \left(\frac{n_{.1}}{N}\right)\left(\frac{n_{.2}}{N}\right)\left(\frac{n_{1.} + n_{2.}}{n_{1.}n_{2.}}\right) = \left(\frac{n_{.1}n_{.2}}{N^2}\right)\left(\frac{N}{n_{1.}n_{2.}}\right). \tag{12}$$

Therefore,

$$W^2 = \frac{(p_1 - p_2)^2}{p(1-p)\left(\frac{1}{n_{1.}} + \frac{1}{n_{2.}}\right)} = \frac{(p_1 - p_2)^2}{\left(\frac{n_{.1}n_{.2}}{Nn_{1.}n_{2.}}\right)} = Q_p, \tag{13}$$

where $n_{1.} = n_{2.} = n$ and $N = n + n = 2n.$

From (10) and (13) we have that

$$P\{Q_p \le z\} = P\{W^2 \le z\} = P\{W \le z^{1/2}\} - P\{W \le -z^{1/2}\},$$

$$\lim_{n\to\infty} P\{Q_p \le z\} = \lim_{n\to\infty} P\{W \le z^{1/2}\} - \lim_{n\to\infty} P\{W \le -z^{1/2}\}$$

$$= \Phi(z^{1/2}) - \Phi(-z^{1/2})$$

$$= 2\int_0^{z^{1/2}} \left(\frac{1}{\sqrt{2\pi}}\right) e^{-u^2/2} du.$$

Let

$$v = u^2 \geq 0 \;\Rightarrow\; u = v^{1/2} \;\Rightarrow\; \frac{du}{dv} = \frac{v^{-1/2}}{2}$$

thus,

$$\lim_{n\to\infty} P\{Q_p \leq z\} = \int_0^z \frac{e^{-v/2}v^{-1/2}}{\sqrt{2}\sqrt{\pi}}\, dv = \int_0^z \frac{e^{-v/2}v^{1/2-1}}{\sqrt{2}\Gamma(1/2)}\, dv.$$

Therefore, the asymptotic distribution of Q_p is the chi-square distribution with one degree of freedom, which proves (9).

Acknowledgement

The authors thank the anonymous referees for their valuable suggestions which led to the improvement of the manuscript.

References

[1] G. Seber, Statistical Models for Proportions and Probabilities, Springer, New York, 2013.

[2] K. Krishnamoorthy, Handbook of Statistical Distributions with Applications, Chapman and Hall, CRC, 2006.

[3] J. L. Fleiss, B. Levin and M. Cho Paik, Statistical Methods for Rates and Proportions, 3rd ed., Wiley and Sons, Inc., 2003.

[4] R. G. Newcombe, Confidence Intervals for Proportions and Related Measures of Effect Size, Chapman and Hall, CRC Biostatistics Series, 2013.

[5] L. Ott and M. Longneker, An Introduction to Statistical Methods and Data Analysis, 7th ed., Cengage Learning, 2016.

[6] S. B. Plichta and E. Kelvin, Statistical methods for health care research, 6th ed., Wolters Kluwer, Lippincott Williams & Wilkins, 2013.

[7] M. E. Stokes, C. S. Davis and G. G. Koch, Categorical data analysis using the SAS system, 2nd ed., SAS Institute Inc. Cary, NC, 2000.

[8] R. V. Hogg and A. T. Craig, Introduction to Mathematical Statistics, 5th ed., Upper Saddle River, New Jersey, USA, 1995.

[9] E. L. Lehmann, Elements of Large-sample Theory, 2nd ed., Springer-Verlag, New York, Inc., 2001.