Nonparametric test for conditional independence in two-way contingency tables

Gery Geenens
Institut de Statistique, Université catholique de Louvain, voie du roman pays, 20, 1348 Louvain-la-Neuve, Belgium

Keywords: two-way contingency tables; chi-square test; likelihood ratio test; nonparametric regression; conditional independence; adjusted Nadaraya-Watson estimator; long-range dependent process.
AMS: 62H17, 62E20, 62G08, 62G20, 62M07.

Abstract

Consider a two-way contingency table, built on two categorical variables R and S. A fundamental question in this context is whether R and S are independent or not. This over-studied testing problem mostly relies on the chi-square or the likelihood ratio test statistics. The principal drawback of the classical way of doing is that probabilities for an individual to fall in a cell of the table are assumed to be equal from one individual to another, so that we are not treating each individual as such but rather a group of supposed homogeneous units. This is often highly unrealistic, since in most of the practical situations some possibly known characteristics of each individual ought to influence or be associated with R, S, or both, and therefore influence the whole dependence structure of the table. In this case, a more judicious idea seems to analyze the conditional joint distribution of R and S given the vector of covariates, say X, and then to test for the conditional independence between R and S given X. Such a test requires the estimation of the conditional probabilities of each cell, given the values of X. In the literature, a striking fact is that the estimation of conditional probabilities associated with categorical responses, given a vector of covariates, is almost always treated via logistic regression methods, most of the time with very few validation of this parametric assumption. In this work, we first present a nonparametric estimation procedure for the conditional probabilities to fall in each cell of the table. These estimates can be used as such, or be employed to validate a parametric assumption, like the logistic one. Secondly, we propose a generalization of the chi-square and the likelihood ratio tests to the case of testing for conditional independence, based on the above-mentioned nonparametric estimates of the conditional probabilities. The asymptotic law of the proposed test statistics is derived.

1. Introduction

Let R and S be two categorical variables, with r and s levels respectively, and consider a sample of n individuals for which R and S are known. A contingency table is built by cross-classifying the sample with respect to the

levels of R and S. Quantities of interest facing such a table are typically the joint probability distribution $\pi = \{\pi_{ij} : 1 \leq i \leq r, 1 \leq j \leq s\}$ of R and S, and ensuing marginal distributions $\{\pi_{i\cdot} : 1 \leq i \leq r\}$ and $\{\pi_{\cdot j} : 1 \leq j \leq s\}$. A fundamental question in this context is whether R and S are independent or not. Formally, we wish to test the null hypothesis that the joint probability distribution of R and S is equal to the product of their marginal distributions, i.e.

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j} \qquad \forall (i,j).$$

Most of the procedures for this testing problem rely on a divergence criterion between $\{\hat{p}_{ij}\}$ and $\{\hat{p}_{i\cdot}\hat{p}_{\cdot j}\}$, where $\hat{p} = \{\hat{p}_{ij} : 1 \leq i \leq r, 1 \leq j \leq s\}$ is the MLE of π. In particular are used the Pearson's chi-square statistic and the likelihood ratio test statistic. It is well known that these two statistics are asymptotically equivalent and asymptotically follow a χ^2 distribution with $(r-1)(s-1)$ degrees of freedom under the null hypothesis, what permits to define an asymptotic rejection criterion for H_0.

But this classical way of doing suffers from an important drawback: the distribution π is assumed to be equal from one individual to another, so that we are not treating each individual as such but rather a group of supposed homogeneous units. This is often unrealistic, since in most of the practical situations some possibly known characteristics of each individual ought to influence or be associated with R, S, or both. Then, a more judicious idea seems to analyze the conditional joint distribution of R and S given the vector of covariates, say X, i.e. $\pi(x) = \{\pi_{ij}(x) : 1 \leq i \leq r, 1 \leq j \leq s\}$, with

$$\pi_{ij}(x) = P\left(R = i, S = j | X = x\right).$$

The fact is that the characteristics X could also influence the eventual relationship between R and S, so that the above mentionned tests of independence have to be transformed to take into account, and possibly remove, this eventual effect of X. The general aim of this paper is therefore to propose test procedures for conditional independence of two categorical variables, given a set of extra covariates. If $S_X \subset \mathbb{R}^p$ is the support of vector X, this conditional independence hypothesis writes

$$H_0 : \pi_{ij}(x) = \pi_{i\cdot}(x)\,\pi_{\cdot j}(x) \qquad \forall x \in S_X, \forall (i,j). \tag{1}$$

Now, testing procedures for (1) have the need of estimates of $\pi(x)$. In this work, we concentrate on nonparametric estimation of this vector of functions. The motivation to favour this kind of methods rather than parametric ones, such as logistic regression, is threefold and is clearly exposed in section 2, as well as some theoretical results about the nonparametric estimation of $\pi(x)$. In section 3, the proposed test procedures are described, as well as the asymptotic distribution of the test statistics is derived. Section 4 presents some concluding remarks and some leads for future work.

2. Nonparametric estimation of $\pi(x)$

Defining a new random vector $Z = \{Z^{(ij)} : 1 \leq i \leq r, 1 \leq j \leq s\}$, with $Z^{(ij)}$ taking the value 1 if the individual belongs to cell (i, j) and 0 otherwise, it is seen that $\pi_{ij}(x) = E(Z^{(ij)}|X = x)$, so that the estimation of $\pi(x)$ is clearly nothing else but a regression problem. Several parametric models have been proposed in this purpose, belonging for most of them to the class of the Generalized Linear Models. The main one is the multivariate logistic regression model, introduced by McCullagh and Nelder (1989, section 6.5.4). These models take advantage of the usual properties of the parametric models, but we can however point out three important drawbacks. First, the binary character of the responses $Z^{(ij)}$ leads to difficulties in analyzing scatter-plots. Since those plots are often the primary tool for defining a reliable parametric pattern for a regression function, the risk of misspecification, already pointed out by Horowitz and Savin (2001), is very high. Besides, you should have here to analyze not one, but simultaneously rs scatter-plots, what obviously makes the exercice still more delicate. Second, again because of the 0-1 response, standard residuals-based model checking techniques are not adapted: residuals are here differences between discrete and continuous quantities, so that their basic representation is not informative, and more elaborate validation procedures (see e.g. Landwehr et al (1984) or Fowlkes (1987)) are themselves based on nonparametric estimation of the conditional probabilities. Finally, a third disadvantage is that the number of parameters to be estimated in these models are $(rs - 1)(p + 1)$, what can already be very large for a moderate size of the table and a moderate number of covariates. The model is therefore rarely simple and concise. Moreover, the Maximum Likelihood estimation of these parameters is based on a large dimensional optimization problem, from which practical difficulties frequently arise.

For these reasons, we consider here a nonparametric estimation of the conditional probabilities $\pi_{ij}(x)$. In this work we will use the Adjusted Nadaraya-Watson kernel regression estimator (ANW), introduced in Hall et al (1999), and discussed in Hall and Presnell (1999) and Cai (2001). For simplicity, we will consider here only the case where X is univariate[1]. Given a kernel function K and a bandwidth h, this estimator of $\pi_{ij}(x)$ writes

$$\hat{p}_{ij}(x) = \sum_{k=1}^{n} W_h(x, X_k) Z_k^{(ij)}, \tag{2}$$

[1] See Geenens and Simar (2007) for results in a multivariate setting.

with[2]

$$W_h(x, X_k) = \frac{w_k(x)K_h(x - X_k)}{\sum_{k'=1}^{n} w_{k'}(x)K_h(x - X_{k'})},$$

where $\{w_k(x), 1 \leq k \leq n\}$ are extra weights satisfying for all x the following properties : $w_k(x) \geq 0$, $\sum_k w_k(x) = 1$,

$$\sum_{k=1}^{n} w_k(x)(x - X_k)K_h(x - X_k) = 0 \tag{3}$$

and $\prod_k w_k(x)$ as large as possible. The motivation to use this estimator is the following: first, and contrary to the Local Linear kernel estimator (LL), the estimation always belongs to the range of the responses, which is here $[0, 1]$ and hence of prime interest as the estimation of probabilities is concerned. Secondly, esentially due to (3), it can be shown (see the above mentionned references) that estimator (2) is first-order equivalent to LL, and therefore preserves the bias, the variance and the automatic good boundary behaviour properties of LL estimator. Note that the same bandwidth h is used for the estimation of the conditional probability to fall in each cell (i, j), in order to keep, for $\{\hat{p}_{ij}(x)\}$, essential properties of the underlying $\{\pi_{ij}(x)\}$, mainly the fact that they have to sum to one for any x. It should not be the case if different bandwidths h_{ij} were used.

Suppose now that the following mild assumptions hold:

(A1) $\{(X_k, Z_k), k = 1, \ldots, n\}$ are i.i.d. random vectors of compact support $D = S_X \times \{z \in \{0, 1\}^{rs} : \sum_q z^{(q)} = 1\}$, such that $Z|X$ follows a multinomial distribution of parameters $(1; \pi(X))$;

(A2) the random variable X admits a density, denoted f, which is differentiable, bounded, and bounded away from zero for any $x \in S_X$;

(A3) the conditional probabilities $\pi_{ij}(x)$ have two continuous derivatives, and are bounded away form 0 and 1 for any $x \in S_X$;

(A4) the kernel K is a symmetric Lipshitz continuous probability density on $[-1, 1]$;

(A5) the bandwidth sequence is such that $h \sim O(n^{-\beta})$ with $\beta \in]1/5, 1/2[$.

The multinomial sampling supposed in (A1) is usual when working in contingency tables, as well as (A2), (A3) and (A4) are classical in nonparametric

[2] K_h is the usual normalized version of K : $K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$

regression. Assumption (A5) is designed such that, first, the bias of estimator $\hat{p}$ is implicitly treated via undersmoothing ($h = o(n^{-1/5})$, see o.a. Hall (1992)), and second, the variance of the test statistic remains "tractable" ($nh^2 \to \infty$, see (5) hereafter). Then, we have, for any fixed x and $\forall(i,j)$:

$$(nh)^{1/2}\left(\hat{p}_{ij}(x) - \pi_{ij}(x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_{ij}^2(x)), \tag{4}$$

where $\sigma_{ij}^2(x) = \nu_0 \frac{\pi_{ij}(x)(1-\pi_{ij}(x))}{f(x)}$, with $\nu_0 = \int K^2(x)dx$. Defining vectors

$$\pi(x) = \big(\pi_{11}(x), \pi_{12}(x), \ldots, \pi_{r(s-1)}(x), \pi_{rs}(x)\big)^t \quad \text{and}$$
$$\hat{p}(x) = \big(\hat{p}_{11}(x), \hat{p}_{12}(x), \ldots, \hat{p}_{r(s-1)}(x), \hat{p}_{rs}(x)\big)^t,$$

the vector analogue of (4) can be written

$$(nh)^{1/2}\left(\hat{p}(x) - \pi(x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\nu_0}{f(x)}\left(\mathrm{diag}(\pi(x)) - \pi(x)\pi(x)^t\right)\right).$$

Finally, define $\nu_0(u) = (K * K)(u)$, i.e. the convolution of the kernel K with itself[3], it can also be shown (see Geenens and Simar (2007) for details) that

$$(nh)\,\mathrm{cov}\left(\hat{p}_{ij}(x_1), \hat{p}_{ij}(x_2)\right)) = \frac{\nu_0(\delta)}{f(x_1)}\pi_{ij}(x_1)(1-\pi_{ij}(x_1)) + o(1) \qquad \text{and}$$
$$(nh)\,\mathrm{cov}\left(\hat{p}_{i_1j_1}(x_1), \hat{p}_{i_2j_2}(x_2)\right)) = -\frac{\nu_0(\delta)}{f(x_1)}\pi_{i_1j_1}(x_1)\pi_{i_2j_2}(x_1) + o(1),$$

as $n \to \infty$, with $\delta = \frac{x_1 - x_2}{h}$.

3. Testing for conditional independence

In order to test the null hypothesis (1), the basic idea is to use, for x fixed, a divergence criterion based on $\hat{p}(x)$, e.g. similar to Pearson's chi-square or likelihood ratio, and then integrate it with respect to x. For example, in view of (4), we could use an estimated version of

$$\int_{S_X} \frac{nhf(x)}{\nu_0} \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(\hat{p}_{ij}(x) - \pi_{i\cdot}(x)\pi_{\cdot j}(x)\right)^2}{\pi_{i\cdot}(x)\pi_{\cdot j}(x)} f(x)dx.$$

The unknown quantities, namely $\pi(x)$ and $f(x)$, could respectively be estimated by $\hat{p}(x)$ and by the usual nonparametric kernel density estimator

$$\hat{f}(x) = \frac{1}{n}\sum_{k=1}^{n} K_h(x - X_k),$$

[3]Note that $\nu_0 = \nu_0(0)$.

while a natural estimation of the integral, seen as an expectation, could be the empirical mean on the observed X_k. The proposed test statistic is therefore given by

$$V^2 = \frac{1}{n}\sum_{k=1}^{n} \frac{nh\hat{f}(X_k)}{\nu_0} \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(\hat{p}_{ij}(X_k) - \hat{p}_{i\cdot}(X_k)\hat{p}_{\cdot j}(X_k)\right)^2}{\hat{p}_{i\cdot}(X_k)\hat{p}_{\cdot j}(X_k)}$$
$$\doteq \frac{1}{n}\sum_{k=1}^{n} \|\hat{v}(X_k)\|^2.$$

In the same way, a generalization of the likelihood ratio test statistic should be given by

$$G = \frac{1}{n}\sum_{k=1}^{n} \frac{2nh\hat{f}(X_k)}{\nu_0} \sum_{i=1}^{r}\sum_{j=1}^{s} \hat{p}_{ij}(X_k) \log \frac{\hat{p}_{ij}(X_k)}{\hat{p}_{i\cdot}(X_k)\hat{p}_{\cdot j}(X_k)}.$$

Geenens and Simar (2007) show that, under assumptions (A1)-(A5) and under H_0,

$$E(V^2) = (r-1)(s-1) + o(1)$$

and

$$\mathrm{var}(V^2) = 2h(r-1)(s-1)\frac{\phi_0 N_0}{\nu_0^2} + o(h), \tag{5}$$

with $\nu_0(u) = (K * K)(u)$, $N_0 = \int \nu_0^2(u)du$ and $\phi_0 = \int f^2(x)dx$.

Note that the variance of V^2 is of order $O(h)$, what is not usual. This comes from the obvious fact that $\{\|\hat{v}(X_k)\|^2\}$ does not form a sequence of independent random variables, as $\|\hat{v}(X_k)\|^2$ and $\|\hat{v}(X_{k'})\|^2$ for which $|X_k - X_{k'}| \leq 2h$ are partially built, through $\hat{p}$, on a common set of observations, so that there exists strong dependence between such two random variables. Moreover, for a fixed k, the cardinality of the set $\{k' : \|X_k - X_{k'}\| < 2h\}$ is of order $O(nh)$, and therefore tends to ∞, so that this dependence is growing with the sample size. Hence, the sequence $\{\|\hat{v}(X_k)\|^2\}$ rather forms a long-range dependent process, which has to be treated with attention. Nevertheless, a Central Limit Theorem can be proven for this kind of dependent variables (see Romano and Wolf (2000)), and it can be stated:

Theorem 3.1 *Under the assumptions (A1-A5), if H_0 hold, we have:*

$$\frac{1}{\sqrt{h}} \frac{\nu_0}{\sqrt{2\phi_0 N_0 (r-1)(s-1)}} (V^2 - (r-1)(s-1)) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1). \tag{6}$$

Similarly to the classical unconditional case, statistics G and V^2 can be proven to be asymptotically equivalent, so that it directly follows

Theorem 3.2 *Under the assumptions (A1-A5), if H_0 hold, we have:*

$$\frac{1}{\sqrt{h}} \frac{\nu_0}{\sqrt{2\phi_0 N_0 (r-1)(s-1)}} (G - (r-1)(s-1)) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1). \tag{7}$$

Now, in order to use distribution (6) or (7) in a practical rejection criterion for hypothesis H_0, it is necessary to estimate the unknown coefficient ϕ_0. This could be done by

$$\hat{\phi}_0 = \frac{1}{n} \sum_{k=1}^{n} \hat{f}(X_k),$$

without altering the distribution of the normalized test statistic (see Geenens and Simar (2007)). Finally, two rejection criteria for the null hypothesis (1), of asymptotic level α, are given by

$$V^2 > (r-1)(s-1) + z_{1-\alpha} \frac{\sqrt{h}}{\nu_0} \sqrt{2(r-1)(s-1)\hat{\phi}_0 N_0} \tag{8}$$

and

$$G > (r-1)(s-1) + z_{1-\alpha} \frac{\sqrt{h}}{\nu_0} \sqrt{2(r-1)(s-1)\hat{\phi}_0 N_0},$$

with $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of a standard normal distribution. The consistency of these procedures are proven in Geenens ans Simar (2007).

4. Conclusion

This paper summarized the main results obtained in Geenens and Simar (2007). It deals with testing for independence between two categorical variables, given a vector X of continuous covariates. This kind of procedure is particularly useful when it is known that some characteristics of the individuals influence the two categorical variables, as well as their eventual relationship, as it permits to take into account, and to remove in some way, the effect of those covariates. The test procedure is based on a nonparametric estimation of the conditional probabilities to fall in any cell of the contingency table. The Adjusted Nadaraya-Watson estimator of Hall et al (1999) is used in that purpose, as it enjoys the good theoretical properties of the Local Linear kernel estimator, while keeping the estimates in the range $[0, 1]$ of the observed responses. Next, for any fixed x in the support of X, we compute some divergence criterion - basically some generalization of the chi-square criterion or the likelihood ratio criterion in classical independence tests in contingency tables - between estimated joint conditional distribution and estimated product of the two conditional marginal distributions. Finally, the information for all x is gathered by taking the empirical mean of the criterion on the observations, what will

be the test statistic. This one is shown to asymptotically follow a normal distribution, from which a practical rejection criterion can be derived. Further extensions of the procedure could consist on the development of a bootstrap algorithm, in order to better approximate the distribution of the test statistic when the number of observations is not large enough to make the asymptotic results reliable, of well on the development of semiparametric estimation of the conditional probabilities, e.g. single-index regression models (SIM), to avoid the well known "curse of dimensionality" affecting any nonparametric method.

Acknowledgements: Financial support from the "Interuniversity Attraction Pole", Phase VI from the Belgian Science Policy is gratefully acknowledged

5. Bibliography

[1] Cai, Z. (2001). *Weighted Nadaraya-Watson regression estimation*. Statist. Probab. Lett., 51, 307-318.

[2] Fowlkes, E.B. (1987). *Some diagnostics for binary logistic regression via smoothing*. Biometrika, 74, 503-515.

[3] Geenens, G. and Simar, L. (2007). *Nonparametric test for conditional independence in two-way contingency tables*. Discussion Paper, Institut de Statistique, Université catholique de Louvain.

[4] Hall, P. (1992). *On bootstrap confidence intervals in nonparametric regression*. Ann. Stat., 20 (2), 695-711.

[5] Hall, P., Wolff, R.C.L. and Yao, Q. (1999). *Methods for estimating a conditional distribution function*. J. Amer. Statist. Assoc., 94, 154-163.

[6] Hall, P. and Presnell, B. (1999). *Intentionally biased bootstrap methods*. J. Roy. Statist. Soc. Ser. B, 61, 143-158.

[7] Horowitz, J.L. and Savin, N.E. (2001). *0Binary Response Models: Logits, Probits and Semiparametrics*. J. Econ. Persp., 15 (4), 43-56.

[8] Landwehr, J.M., Pregibon, D. and Shoemaker, A.C. (1984). *Graphical methods for assessing logistic regression models*. J. Amer. Statist. Assoc., 79, 61-83.

[9] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

[10] Romano, J.P. and Wolf, M. (2000). *A more general central limit theorem for m-dependent random variables with unbounded m*. Statist. Probab. Lett., 47, 115-124.