Tests in contingency tables as regression tests

by

Stanislav Anatolyev* and Grigory Kosenok
New Economic School

Abstract

Applied researchers often use tests based on contingency tables, especially in preliminary data analysis and diagnostic testing. We show that many such tests may be alternatively implemented by testing for coefficient restrictions in linear regression systems.

JEL classification codes: C12, C22, C32, C53

Key words: Contingency table, linear regression, chi-squared test, Wald test, ranks

*Corresponding author. Address: Stanislav Anatolyev, New Economic School, Nakhimovsky Pr., 47, Moscow, 117418 Russia. E-mail: sanatoly@nes.ru.

1 Introduction

Often applied economists and financiers use tests associated with contingency tables. Such tests are designed for verifying independence or homogeneity properties of original data or regression residuals, and are heavily used in preliminary data analysis and diagnostic testing. For example, the phrase "contingency table" leads to 185 and 61 hits using advance search (performed in July 2008) in JSTOR economics (31 journals) and finance (7 journals) collections, respectively. Some of associated tests are even more frequently mentioned.

The tests related to contingency tables are performed by utilizing particular formulas, often quite complex ones. In this paper we give a number of results showing that typically such tests may be alternatively implemented via a system of linear regressions. This concerns tests for independence, tests for accordance with distribution, tests for symmetry, and tests based on ranks. Such reformulation is useful for a number of reasons. First, it unifies the regression analysis with the theory of contingency tables. Second, the bridge between the two theories sheds more light on intuitive contents of contingency table tests and test statistics. Third, running regressions may be more convenient and familiar for practitioners.

A complete working paper version of the paper containing some references to applications, more related discussions, and proofs of all results can be found on the website
`http://www.nes.ru/~sanatoly/Papers/CT.pdf`.

2 Two-way contingency tables

We begin by introducing some basic notation. The sample size is denoted by n. Bars denote taking sample averages, i.e., for example, $\overline{a}_{ij} = n^{-1}\sum_{t=1}^{n} a_{ij}$. By $\|a_i\|_{i=1}^{\ell}$ we mean a column $\ell \times 1$ vector with i^{th} element a_i. By $\|a_{i,j}\|_{i=1}^{\ell_1}{}_{j=1}^{\ell_2}$ we mean an $\ell_1 \times \ell_2$ matrix whose i^{th}, j^{th} element equals $a_{i,j}$.

We consider a two-way $(\ell_x + 1) \times (\ell_y + 1)$ contingency table. The state space Ω_x of one variable, x, is partitioned into the cover $\{K_i\}_{i=1}^{\ell_x+1}$; similarly, the state space Ω_y of the other variable, y, is partitioned into the cover $\{\Lambda_j\}_{j=1}^{\ell_y+1}$. Let us denote

$$\mathbb{I}_{i,\cdot} = \mathbb{I}(x \in K_i), \quad \mathbb{I}_{\cdot,j} = \mathbb{I}(y \in \Lambda_j), \quad \mathbb{I}_{i,j} = \mathbb{I}(x \in K_i)\,\mathbb{I}(y \in \Lambda_j)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Let

$$\pi_{i,\cdot} = \Pr\{x \in K_i\}, \quad \pi_{\cdot,j} = \Pr\{y \in \Lambda_j\}, \quad \pi_{i,j} = \Pr\{x \in K_i, y \in \Lambda_j\},$$

and define

$$\pi_x = \|\pi_{i,\cdot}\|_{i=1}^{\ell_x}, \quad \pi_y = \|\pi_{\cdot,j}\|_{j=1}^{\ell_y} \quad \pi = \|\pi_{i,j}\|_{i=1}^{\ell_x}{}_{j=1}^{\ell_y}.$$

We assume that $\pi_{i,j} > 0$ for all i and j.

The contingency table looks as follows:

		y					
		Λ_1	Λ_2	$\cdots$	Λ_{ℓ_y}	Λ_{ℓ_y+1}	Ω_y
x	K_1	$p_{1,1}$	$p_{1,2}$	$\cdots$	p_{1,ℓ_y}	p_{1,ℓ_y+1}	$p_{1,\cdot}$
	K_2	$p_{2,1}$	$p_{2,2}$	$\cdots$	p_{2,ℓ_y}	p_{2,ℓ_y+1}	$p_{2,\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
	K_{ℓ_x}	$p_{\ell_x,1}$	$p_{\ell_x,2}$	$\cdots$	p_{ℓ_x,ℓ_y}	p_{ℓ_x,ℓ_y+1}	$p_{\ell_x,\cdot}$
	K_{ℓ_x+1}	$p_{\ell_x+1,1}$	$p_{\ell_x+1,2}$	$\cdots$	p_{ℓ_x+1,ℓ_y}	p_{ℓ_x+1,ℓ_y+1}	$p_{\ell_x+1,\cdot}$
	Ω_x	$p_{\cdot,1}$	$p_{\cdot,2}$	$\cdots$	$p_{\cdot,\ell_y}$	$p_{\cdot,\ell_y+1}$	1

The figures in the tables are empirical probabilities of falling into corresponding cells

$$p_{i,j} = \bar{\mathbb{I}}_{i,j}$$

and marginal empirical probabilities

$$p_{i,\cdot} = \bar{\mathbb{I}}_{i,\cdot}, \quad p_{\cdot,j} = \bar{\mathbb{I}}_{\cdot,j}.$$

In many applications, each K_i is an interval $[\kappa_{i-1}, \kappa_i)$ and each Λ_j is an interval $[\lambda_{j-1}, \lambda_j)$, where $-\infty = \kappa_0 < \kappa_1 < \cdots < \kappa_{\ell_x} < \kappa_{\ell_x+1} = +\infty$ and $-\infty = \lambda_0 < \lambda_1 < \cdots < \lambda_{\ell_y} < \lambda_{\ell_y+1} = +\infty$. When $\ell_x = \ell_y$ and $K_i = \Lambda_i$ for all i, the contingency table is referred to as one with identical categorizations. However, rows and columns of a contingency table need not correspond to partitionings of a real axis, and categorizations need not be identical.

3 Tests and their equivalences

3.1 Tests for independence

The classical χ^2-test statistic for independence between the variables x and y (more precisely, for no association between x and y) is equal to

$$X^2 = n \sum_{i=1}^{\ell_x+1} \sum_{j=1}^{\ell_y+1} \frac{(p_{i,j} - p_{i,\cdot}p_{\cdot,j})^2}{p_{i,\cdot}p_{\cdot,j}}, \tag{1}$$

and is asymptotically distributed as $\chi^2(\ell_x \ell_y)$.

Theorem 1 *The χ^2 test (1) is asymptotically equivalent to an OLS-based Wald test for the nullity of all slope coefficients in a linear multiple regression of $\mathbb{I}_{\cdot,j}$ on $\mathbb{I}_{i,\cdot}$ with a constant in each equation, i.e. for the null*

$$H_0 : \beta_{ji} = 0,\ i = 1, \cdots, \ell_x,\ j = 1, \cdots, \ell_y$$

in the regression system

$$\mathbb{I}_{\cdot,j} = \alpha_j + \sum_{i=1}^{\ell_x} \beta_{ji} \mathbb{I}_{i,\cdot} + \eta_j, \quad j = 1, \cdots, \ell_y. \tag{2}$$

Alternatively, $\mathbb{I}_{i,\cdot}$ may be regressed on $\mathbb{I}_{\cdot,j}$ rather than $\mathbb{I}_{\cdot,j}$ are regressed on $\mathbb{I}_{i,\cdot}$.

Note that when the contingency table is 2×2, the test for independence can be run using only one bivariate regression.

3.2 Tests for accordance with distribution

The previous χ^2 test may be also interpreted as a test for homogeneity of $\ell_x + 1$ subsamples $y|(x \in K_i)$, i.e. that the conditional distribution of y does not depend on x. Suppose now that the marginals $\{\pi_{\cdot,j}\}_{j=1}^{\ell_y+1}$ are known a priori. Then the χ^2 test, whose statistic can be modified to take account of this knowledge, may be interpreted as a test for accordance of $\ell_x + 1$ independent subsamples with a known multinomial distribution. The modified test statistic is equal to

$$X^2 = n \sum_{i=1}^{\ell_x+1} \sum_{j=1}^{\ell_y+1} \frac{\left(p_{i,j} - p_{i,\cdot}\pi_{\cdot,j}\right)^2}{p_{i,\cdot}\pi_{\cdot,j}}, \tag{3}$$

and is asymptotically distributed as $\chi^2\left(\left(\ell_x + 1\right)\ell_y\right)$.

Theorem 2 *The χ^2 test (3) is asymptotically equivalent to an OLS-based Wald test for the nullity of all coefficients in a linear multiple regression of $\mathbb{I}_{\cdot,j} - \pi_{\cdot,j}$ on $\mathbb{I}_{i,\cdot}$ with a constant in each equation, i.e. for the null*

$$H_0 : \alpha_j = \beta_{ji} = 0,\ i = 1, \cdots, \ell_x,\ j = 1, \cdots, \ell_y$$

in the regression system

$$\mathbb{I}_{\cdot,j} - \pi_{\cdot,j} = \alpha_j + \sum_{i=1}^{\ell_x} \beta_{ji}\mathbb{I}_{i,\cdot} + \eta_j, \quad j = 1, \cdots, \ell_y. \tag{4}$$

Remark 1 *The regression from theorem 2 is the same as that from theorem 1. However, the set of restrictions is expanded by additional ℓ_y restrictions of equality of the intercepts in (2) to the known a priori y-marginals. This explains the additional ℓ_y degrees of freedom in the asymptotic χ^2 distribution.*

When $\ell_x = 0$, the contingency table essentially becomes one-way, and there is only one subsample $y|(x \in \Omega_x)$. Then the test is called the Pearson (1900) test for goodness of fit, and it simply verifies if the sample is drawn from a given multinomial distribution. The test statistic becomes Pearson's

$$X^2 = n \sum_{j=1}^{\ell_y+1} \frac{\left(p_{\cdot,j} - \pi_{\cdot,j}\right)^2}{\pi_{\cdot,j}}, \tag{5}$$

and is asymptotically distributed as $\chi^2\left(\ell_y\right)$.

Corollary 1 *The Pearson test (5) is asymptotically equivalent to an OLS-based Wald test for the nullity of all coefficients in a linear multiple regression of $\mathbb{I}_{\cdot,j} - \pi_{\cdot,j}$ on a constant in each equation, i.e. for the null*

$$H_0 : \alpha_j = 0,\ j = 1, \cdots, \ell_y$$

in the regression system

$$\mathbb{I}_{\cdot,j} - \pi_{\cdot,j} = \alpha_j + \eta_j, \quad j = 1, \cdots, \ell_y. \tag{6}$$

Applications of the Pearson test in economics can be divided into two categories. In the first category, frequencies of simulated values from an estimated model falling into pre-specified bins are compared to a multinomial distribution implied by an assumed continuous distribution, the latter possibly having shape parameters estimated. Most often, however, the reference distribution is uniform so that $\pi_{\cdot,j} = 1/(\ell_y + 1)$, and the leading application is evaluation of conditional density forecasts. In the second category of applications, one compares frequencies of model-generated predictions falling into pre-specified bins to a multinomial distribution implied by an empirical density.

3.3 Tests for symmetry

Next we analyze two tests for symmetry in contingency tables with identical categorizations. Stuart (1955) suggested a test for homogeneity of the marginal distributions of x and y. Formally, the null is

$$H_0 : \pi_{i,\cdot} = \pi_{\cdot,i},\ i = 1, \cdots, \ell,$$

which automatically implies also $\pi_{\ell+1,\cdot} = \pi_{\cdot,\ell+1}$. The test is based on the $\ell \times 1$ vector of differences $p_{i,\cdot} - p_{\cdot,i},\ i = 1, \cdots, \ell$. Let $d_n = \|p_{i,\cdot} - p_{\cdot,i}\|_{i=1}^{\ell}$. The test statistic is

$$Q_n = n d_n' V^{-1} d_n,$$

where $V = \|V_{i,j}\|_{i=1\ j=1}^{\ell\quad\ell}$, and $V_{i,i} = p_{i,\cdot} + p_{\cdot,i} - 2p_{i,i} - (p_{i,\cdot} - p_{\cdot,i})^2$, $V_{i,j} = -p_{i,j} - p_{j,i} - (p_{i,\cdot} - p_{\cdot,i})(p_{j,\cdot} - p_{\cdot,j})$, $i \neq j$. Under H_0, Q_n is asymptotically distributed as $\chi^2(\ell)$.

Theorem 3 *The Stuart Q_n test is asymptotically equivalent to an OLS-based Wald test for the nullity of all intercepts in a linear multiple regression of $\mathbb{I}_{i,\cdot} - \mathbb{I}_{\cdot,i}$ on a constant, i.e. for the null*

$$H_0 : \alpha_i = 0,\ i = 1, \cdots, \ell$$

in the regression system

$$\mathbb{I}_{i,\cdot} - \mathbb{I}_{\cdot,i} = \alpha_i + \eta_i, \quad i = 1, \cdots, \ell. \tag{7}$$

Bowker (1948) suggested a test for complete symmetry of the contingency table. Such symmetry implies a stronger equivalence between the two classifications than equality of marginal distributions. In fact, it is the two conditional distributions that are compared. Formally, the null is

$$H_0 : \pi_{i,j} = \pi_{j,i},\ i = 2, \cdots, \ell + 1,\ j = 1, \cdots, i - 1.$$

The test is based on $\ell(\ell+1)/2$ differences $p_{i,j} - p_{j,i},\ i = 2, \cdots, \ell+1,\ j = 1, \cdots, i-1$. The test statistic is

$$U_n = n \sum_{i=2}^{\ell+1} \sum_{j=1}^{i-1} \frac{(p_{i,j} - p_{j,i})^2}{p_{i,j} + p_{j,i}}.$$

Under H_0, U_n is asymptotically distributed as $\chi^2(\ell(\ell+1)/2)$.

Theorem 4 *The Bowker U_n test is asymptotically equivalent to an OLS-based Wald test for the nullity of all intercepts in a linear multiple regression of $\mathbb{I}_{i,j} - \mathbb{I}_{j,i}$ on a constant, i.e. for the null*

$$H_0 : \alpha_{ij} = 0,\ i = 2, \cdots, \ell + 1,\ j = 1, \cdots, i - 1$$

in the regression system

$$\mathbb{I}_{i,j} - \mathbb{I}_{j,i} = \alpha_{ij} + \eta_{ij}, \quad i = 2, \cdots, \ell + 1,\ j = 1, \cdots, i - 1. \tag{8}$$

3.4 Tests based on ranks

Often researchers carry out testing for independence or homogeneity using rank transformed data rather than the original data, the idea being to compare more objective "ordinal" data characteristics instead of "cardinal" ones. In the rest of the paper we review two class of tests based on ranks – the Kruskal–Wallis test and the Spearman rank test.

Suppose that k random samples of size $n_1, ..., n_k$ are tested for identity of distributions they come from. Let the vector of ranks $(r_{1,1}, ..., r_{1,n_1}, ..., r_{k,1}, ..., r_{k,n_k})'$ correspond to the pooled sample (of length $n = \sum_{j=1}^{k} n_j$). Let j index samples, while i index observations within a sample. The sums of ranks for the separate samples is denoted by $R_j = \sum_{i=1}^{n_j} r_{j,i}$. The Kruskal–Wallis test statistic is

$$KW = \frac{12}{(n-1)\, n} \sum_{j=1}^{k} \frac{1}{n_j} \left(R_j - \frac{n+1}{2} n_j \right)^2.$$

Let asymptotically $n \to \infty$ and $\min_j n_j \to \infty$ so that $\lambda_j \equiv \lim_{\min_j n_j \to \infty} n_j / n \neq 0$ for $j = 1, ..., k$. Under these circumstances, KW is asymptotically distributed as $\chi^2 (k-1)$.

Theorem 5 *The Kruskal–Wallis KW test is asymptotically equivalent to an OLS-based Wald test for the nullity of all intercepts in a linear multiple regression of $r_{j,i} - \frac{n+1}{2}$, $j = 1, \cdots, k-1$, on a constant, i.e. for the null*

$$H_0 : \alpha_j = 0,\ j = 1, \cdots, k - 1$$

in the regression system

$$r_{j,i} - \frac{n+1}{2} = \alpha_j + \eta_{ij}, \quad j = 1, \cdots, k - 1, \tag{9}$$

with observations running from $i = 1$ to $i = n_j$ for equation j.

Remark 2 *A similar, but different, situation is considered in statistical literature. A fixed number s of products are ranked by a fixed number k of experts. Denote by $K_{i,j}$ the ranking that the expert j gave the product i ($K_{i,j}$ varies from 1 to s), and by $N_{i,j}$ the number of times the product i received ranking j.*
The Friedman test statistic

$$F = \frac{12}{ks\,(s+1)} \sum_{j=1}^{s} \left(\sum_{j=1}^{k} K_{i,j} - \frac{s+1}{2} k \right)^2$$

is used to test for homogeneity of products. It is asymptotically distributed as $\chi^2(s-1)$ *as the number of experts increases. It is possible to show that* F *is asymptotically equivalent to an OLS-based Wald test for*

$$H_0 : \alpha_i = 0,\ i = 1, \cdots, s-1$$

in the regression system

$$K_{i,j} - \frac{s+1}{2} = \alpha_i + \eta_{ij}, \quad i = 1, \cdots, s-1, \tag{10}$$

with observations running from $j = 1$ *to* $j = k$.
The Anderson test statistic

$$A = \frac{s}{k} \sum_{i=1}^{s} \sum_{j=1}^{s} \left(N_{i,j} - \frac{k}{s} \right)^2,$$

asymptotically distributed as $\chi^2\left((s-1)^2\right)$ *as the number of experts increases, is used to test for homogeneity of products. It is possible to show that* A *is asymptotically equivalent to an OLS-based Wald test for*

$$H_0 : \alpha_{i,j} = 0,\ i, j = 1, \cdots, s-1$$

in the regression system

$$N_{i,j} - \frac{k}{s} = \alpha_{i,j} + \eta_{ij}, \quad i, j = 1, \cdots, s-1, \tag{11}$$

with one observation per equation.

Suppose the vector of ranks $(R_1, ..., R_n)'$ corresponds to a random sample of length n. The Spearman rank statistic ρ is defined as

$$\rho = \frac{12}{(n-1)\, n} \sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right) \left(R_i - \frac{n+1}{2} \right).$$

Theorem 6 *The Spearman rank statistic* ρ *is equal to* $n+1$ *times the OLS slope coefficient in a linear regression of* R_i *on a constant and observation number* i

$$R_i = \alpha + \beta i + \eta, \tag{12}$$

with observations running from $i = 1$ *to* $i = n$.

As a consequence, an OLS-based t test for the null $H_0 : \beta = 0$ may be used to test for independence of elements in a given sample, which is often realized in practice by informally comparing ρ to zero.

The pairwise Spearman rank correlation coefficient ρ between two vectors of ranks $(R_1, ..., R_n)'$ and $(S_1, ..., S_n)'$ corresponding to two random samples of length n is defined as

$$\rho = \frac{\sum_{i,j=1}^{n} (S_j - S_i)(R_j - R_i)}{\sqrt{\sum_{i,j=1}^{n} (S_j - S_i)^2 \sum_{i,j=1}^{n} (R_j - R_i)^2}}.$$

Theorem 7 *The Spearman rank correlation coefficient ρ is equal the OLS slope coefficient in a linear regression of $R_j - R_i$ on $S_j - S_i$*

$$R_j - R_i = \beta (S_j - S_i) + \eta, \tag{13}$$

with n^2 observations for all possible pairs (i, j) where each index runs from 1 to n. Alternatively, one may switch $S_j - S_i$ and $R_j - R_i$ in the regression (13). A constant term may be innocuously introduced into the regression (13).

As a consequence, an OLS-based t test for the null $H_0 : \beta = 0$ may be used to test for independence of two given random samples.

Remark 3 *In the context of Remark 2, the Umbrella test statistic*

$$U = \frac{12}{\sqrt{k(s-1)}s(s+1)} \left(\sum_{i=1}^{s} i \sum_{j=1}^{k} K_{i,j} - \frac{1}{2} ks(s+1)^2 \right),$$

asymptotically distributed as $N(0, 1)$ as the number of experts increases, is used to test for homogeneity of products. It is possible to show that U is asymptotically equivalent to an OLS-based t test for

$$H_0 : \beta = 0$$

in the regression

$$K_{i,j} = \alpha + \beta i + \eta_{ij}, \tag{14}$$

with observations running from $i, j = 1$ to $i = s, j = k$.

References

Bowker, A.H. (1948) A test for symmetry in contingency tables. *Journal of the American Statistical Association* 43, 572–574.

Pearson, K. (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series* 50, 157–175.

Stuart, A. (1955) A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika* 42, 412–416.