

Categorical Data Analysis

In this ‘chapter’ we consider: (1) a test for fit of a distribution, focusing on completely specified distributions, with some broader discussion; and (2) contingency table tests, including testing independence of two categorical variables, plus testing equality of multiple categorical distributions based on independent random samples. The standard tests are Chi-squared tests—Pearson’s chi-square tests—and one useful reference is

http://en.wikipedia.org/wiki/Pearson's_chi-square_test

Problem (1) is more complicated if the distribution being tested depends on one or more unspecified parameters that need to be estimated, and we will only address this case superficially.

Goodness-of-Fit Tests when Category Probabilities are Completely Specified

Consider a random sample of size n for a categorical variable with k categories. Consider testing the null hypothesis that the corresponding categories have probabilities p_{1o}, \dots, p_{ko} , respectively—namely, $H_o : p_i = p_{io}$ for $i = 1, \dots, k$.

To test this null hypothesis, one would want to reject it if the observed sample was surprisingly far from what would be expected if the null hypothesis were true, where the i th cell count N_i has expected value $E[N_i] = np_{io}$ if H_o is true.

Example: Consider rolling a die 60 times to test whether the die is fair. If fair, each possible value should have equal probability, so $p_i = 1/6$ for $i = 1, \dots, 6$. Suppose 60 rolls of a die yield the counts 5, 9, 4, 12, 18, 12 for the numbers 1–6, respectively. Is this evidence that the die is not fair?

Exact Test: The distribution of the observations into the categories, or cells, follows a multinomial distribution. In particular, if N_i denotes the number of observations in the i th cell, then $(N_1, \dots, N_k) \sim \text{multinomial}(n, p_1, \dots, p_k)$, where $\sum_i p_i = 1$, with $n = \sum_i N_i$.

If the die is fair, then each expected count is 10 (i.e. $np_i = 60(\frac{1}{6})$), and one should reject the null hypothesis that the die is fair if the observed sample deviates surprisingly from what is expected to happen when the null hypothesis is true—namely, if the observed significance level is small. In theory, one could compute the probability of every possible outcome under the null hypothesis and compute the observed significance level associated with this outcome *if* one could order (or partially order) the possible outcomes w.r.t. how far they are from what is expected. How might one do this? Perhaps less likely outcomes are further. Alternatively (or is this the same idea?!), one could determine a test statistic and use it to order the possible outcomes. Consider for example the following test statistic:

$$\chi^2 = \sum_i \frac{(N_i - E[N_i])^2}{E[N_i]}$$

We won’t attempt to carry out the exact test in the example above.

Approximate Test: If the sample size n is sufficiently large, then under the null hypothesis the above test statistic has approximately a $\chi^2(k-1)$ distribution, (i.e. $\nu = k-1$). So, an approximate size- α test rejects the null hypothesis if $\chi^2 > \chi^2(1-\alpha; k-1)$.

This approximation works well if all expected cell counts are at least five, and this restriction can be relaxed some.

Example, continued: For the above die data:

$$\begin{aligned} \chi^2 &= [(5-10)^2 + (9-10)^2 + (4-10)^2 + (12-10)^2 + (18-10)^2 + (12-10)^2]/10 \\ &= (25 + 1 + 36 + 4 + 64 + 4)/10 = 13.4 \end{aligned}$$

What is the corresponding observed significance level? What is the conclusion?

These multinomial experiments reduce to a binomial experiment if $k = 2$

In a binomial experiment, a binomial observation of $X \sim B(n, p)$ is used to test $H_o : p = p_o$. You are probably familiar with a test of this null hypothesis based on the test statistic

$$z = \frac{\hat{p} - p_o}{\sqrt{p_o q_o / n}},$$

where $q_o = 1 - p_o$. For a two-tailed test, one would reject H_o if $|z|$ is large, say $|z| > z(1 - \alpha/2)$, or equivalently if z^2 is large, say $z^2 > \chi^2(1 - \alpha; 1)$. Note that

$$z^2 = (\hat{p} - p_o)^2 / (p_o q_o / n) = (x - np_o)^2 / (np_o q_o),$$

and compare this with the standard Chi-squared test statistic

$$\begin{aligned} \chi^2 &= (x - np_o)^2 / (np_o) + (n - x - np_o)^2 / (np_o) \\ &= [q_o(x - np_o)^2 + p_o(n - x - np_o)^2] / (np_o q_o) \\ &= [q_o(x - np_o)^2 + p_o(np_o - x)^2] / (np_o q_o) \\ &= (x - np_o)^2 / (np_o q_o). \end{aligned}$$

These are the same! So, rejecting for $|z|$ or z^2 large is equivalent to rejecting for large values of the standard Chi-squared statistic.

Note: With $k = 2$, there is a natural way to order the outcomes, and so it is clear how to compute observed significance levels for the exact test, at least for one-sided tests.

Extensions of the Goodness-of-Fit test

Here we will simply note that there are interesting and useful variations on the test we have considered for a *simple null hypothesis*—namely, for a completely specified distribution—with only k cells.

The number of cells could be infinite. For example, we might test that our random sample comes from a Poisson distribution with mean $\lambda = 5$. One would typically need to collapse cells so one has only k categories, doing so in such a way that expected cell counts are not too small.

We could test a *composite null hypothesis* that may require the estimation of one or more parameters. For example, we might test that our random sample comes from a Poisson distribution with mean λ unspecified. One would need to estimate λ from the data. One could do this by maximum likelihood, and one would also need to collapse to k categories, say. The MLE depends on whether it is obtained from the original data or from the data after collapsing the data into categories. If one computes the MLE from the collapsed data then uses the MLE to compute expected cell counts to determine the χ^2 statistic, in this case under the null hypothesis the test statistic is asymptotically $\chi^2(k - 1 - 1)$, where the one additional d.f. is lost because of the estimation of one parameter. Unfortunately, it is easier to compute the MLE from the original data, but then the Chi-squared approximation is less appropriate.

One could use the same general approach to test goodness-of-fit of a continuous distribution. For example, given a random sample of size 50 from some distribution, we could test the null hypothesis that the distribution is normal with $\mu = 100$ and $\sigma = 20$. To do so, we could concoct categories by dividing or partitioning this normal distribution into say 10 equally likely intervals by finding the deciles of the distribution. (Why 10?) We could likewise test normality without specification of the parameters, but then we would have a composite hypothesis and the parameters would need

to be estimated. It should be noted that this is not the best way to test normality, but it is one way. Some better tests involve comparing the *order statistics* to the *empirical distribution function*, analogous to looking at a normal probability plot, basing the test for example on the greatest deviation between the empirical distribution function and the fitted normal distribution function.

Two-Way Contingency Tables

These are used for: (1) testing for homogeneity of populations w.r.t. a categorical variable given independent random samples, and (2) testing for independence of two categorical variables given a single random sample.

(1) Testing homogeneity of populations arises given independent random samples from r populations, where each population can be divided into the same c categories. Let p_{ij} denote the probability of a random observation from the i th population falling into the j th category. The populations are homogeneous if, for each category j , p_{ij} does not depend on the population i .

(2) Testing independence arises given a random sample from a single population where each observation can be categorized w.r.t. two categorical variables, with r and c levels respectively, say. Suppose we lay this out in a two-way table with r rows and c columns.

Approximate Test: Both cases introduced above are generally tested in the same way, using Pearson's χ^2 test statistic already introduced—namely,

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(N_{ij} - E[N_{ij}])^2}{E[N_{ij}]},$$

though here the cells simply depend on two indices rather than one. The test statistic is computed conditioned on the row and column totals, $N_{i.}$ and $N_{.j}$, say, and so there are only $(r - 1)(c - 1)$ degrees of freedom. Also, one needs to know how to compute the expected cell counts in each of the two settings.

For (1), if the populations are homogeneous, then given the marginal totals, one would estimate the probability of the j th category by $N_{.j}/n$, for $n = N_{..}$ the total number of observations. So, the expected value of the count N_{ij} in the i th sample falling into this j th category would be $E[N_{ij}] = N_{i.}(N_{.j}/n)$.

For (2), if the two categorical variables are independent, then given the marginal totals, one would estimate the probability of a random observation falling into the i th row (i.e. the i th level of the row variable) by $N_{i.}/n$, one would estimate the probability of a random observation falling into the j th column (i.e. the j th level of the column variable) by $N_{.j}/n$, under independence one would estimate the probability of a random observation falling into the ij th cell (i.e. row i and column j) by the product of these probability estimates, so $(N_{i.}/n)(N_{.j}/n)$, and so the expected value of the count N_{ij} in the i th sample falling into this ij th category would be $E[N_{ij}] = n(N_{i.}/n)(N_{.j}/n) = N_{i.}(N_{.j}/n)$ —the same as for case (1).

The approximate tests reject the appropriate null hypothesis if $\chi^2 > \chi^2(1 - \alpha; (r - 1)(c - 1))$. As before, these approximate tests work well if all expected cell counts are at least five, and this restriction can be relaxed some.

Fisher's exact test: Sir R.A. Fisher developed an exact test for the above cases (1) and (2) tests, given the marginal totals. For a good presentation on *Fisher's exact test*, see the Wikipedia presentation at the following link:

http://en.wikipedia.org/wiki/Fisher's_exact_test

In particular, note the following. When the sample (or samples) is not large enough to justify the

Chi-squared approximation, one should use an exact test, and exact probabilities can be computed under the null distribution by using the hypergeometric distribution. An example is given for a 2×2 table, which could fit either case (1) or case (2), which is amenable to illustrating exact calculations, and which illustrates the fact that there is a natural order of the possible outcomes in a 2×2 table given the marginal totals.

The three external links at the bottom of the Wikipedia page are all interesting and useful.

The first external link is to an “On-line exact test calculator with examples”. At the outset it provides links to: an exact (one-sided) analysis of 2×2 contingency tables, and an exact two-sided analysis of up to 6×6 contingency tables. Examples are also provided. Examples 1 and 2 are of types (1) and (2), respectively. Example 3 is a simple example of a 2×2 table for which Fisher’s exact test is easily illustrated.

The second external link is to an “On-line exact test calculator that accepts larger cell counts”. This gets forwarded to a new link—

<http://www.langsrud.com/fisher.htm>

—where p -values can be computed for 1- or 2-tailed tests for 2×2 tables.

The third external link is to the “mathworld.wolfram.com Page detailing the extension of Fisher’s exact test to $m \times n$ contingency tables”. This page provides a nice presentation on Fisher’s exact test, including the role of the hypergeometric distribution, the need to determine an order for the possible outcomes, computational issues arising if the table is more than 2×2 , and a simple example of the exact test for a 2×2 table.

See also *Yate’s correction for continuity*, intended to improve the accuracy of the Chi-squared test for 2×2 tables when the sample size is small—namely, with an expected cell frequency less than five. See

http://en.wikipedia.org/wiki/Yates'_correction_for_continuity

Yate’s correction was presumably more useful before we had computers to facilitate use of exact methods, especially for cases where the sample size is not large enough to justify the Chi-squared approximation.

Homework: see next page

Homework:

1. The following is the distribution of the hourly number of trucks arriving at a company's warehouse:

Number of trucks	0	1	2	3	4	5	6	7	8
Frequency	52	151	130	102	45	12	5	1	2

Find the sample mean of this distribution, use it (rounded to one decimal place) to estimate the parameter λ , and fit a Poisson distribution—namely, use the corresponding Poisson distribution to estimate cell probabilities. Test for goodness of fit at the 5% level of significance, reducing the number of degrees of freedom by one due to estimation of the single parameter, and collapsing the cells for 7 and above into one category. Also compute the observed significance level. Interpret the results.

2. Criminologists have long debated whether there is a relationship between weather conditions and the incidence of violent crime. A survey classified 1359 homicides according to season, resulting in the accompanying data. Test the null hypothesis of equal proportions for all seasons.

	Winter	Spring	Summer	Fall	Total
observed	328	334	372		1359
expected					

- a) State H_0 and H_A for this problem.
 - b) Complete the table.
 - c) Conduct the test using $\alpha = 0.05$. What is your conclusion?
3. Seldane-D is an over-the-counter drug designed to relieve sneezing, nasal congestion, and other symptoms of allergic rhinitis. General adverse effects of Seldane-D were investigated in a double-blind, controlled study of 58 patients suffering from allergic rhinitis. A sample of 37 patients were given Seldane-D, whereas a second sample of 19 patients were given a placebo (no drug). The number of patients reporting insomnia in each of the two groups are given in the table. Test to determine whether the proportion of patients taking Seldane-D who experience insomnia differs from the corresponding proportion for patients receiving the placebo. Use $\alpha = 0.05$.

	Seldane-D	Placebo
Insomnia	10	2
No Insomnia	27	17
Totals	37	19

- a) Use the χ^2 test.
- b) Use Fisher's exact test.