Goodness-of-Fit Tests and Categorical Data Analysis (Devore Chapter Fourteen)

MATH-252-01: Probability and Statistics II[*]

Spring 2018

Contents

Thursday 5 April 2018

1 Chi-Squared Tests with Known Probabilities

1.1 Chi-Squared Testing

The next sort of inference we consider is known as *categorical data analysis*. Consider an experiment with a set of k possible discrete outcomes. If we do n independent repetitions of that experiment, we find n_1 that end in outcome #1, n_2 that end in outcome #2, up to n_k ending in outcome #k, where each n_i is a non-negative integer, and $n_1 + n_2 + \cdots + n_k = n$. There are k different categories into which each observation can fall, and we count up the numbers of each. We have a model which tells us the expected probabilities p_1, p_2, …, p_k of the different outcomes, where $0 \leq p_i \leq 1$ and $p_1 + p_2 + \cdots p_k = 1$. The expected numbers of observations in each category according to the model are np_1, np_2, …, np_k. (Note that these will in general not be integers.) If we consider this model to be the null hypothesis, we're interested in a test which rejects the model if the observed number counts in the categories are very different from their expected values.

To see the test statistic, we take a slight diversion and recall that if $X_1, \ldots, X_k$ are independent random normal variables such that $X_i \sim N(\mu_i, \sigma_i^2)$, then

$$U = \sum_{i=1}^{k} \frac{(X_i - \mu_i)^2}{\sigma_i^2} = \sum_{i=1}^{k} {Z_i}^2 \tag{1.1}$$

is a chi-square random variable with k degrees of freedom, also known as $\chi^2(k)$. Similarly, one of the consequences of Student's theorem is that if $\{X_i\}$ is a sample from $N(\mu, \sigma^2)$

$$V = \sum_{i=1}^{k} \frac{(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2(k-1) \tag{1.2}$$

as we saw when we considered confidence intervals for the population variance, if we have a statistic $W \sim \chi^2(\nu)$,

$$P(W \geq \chi^2_{\alpha,\nu}) = \alpha \tag{1.3}$$

so comparing the statistic value to $\chi^2_{\alpha,\nu}$ gives a test at significance level α of the original model.

To connect this to the categorical data analysis problem, consider the $k = 2$ case, where there are two possible outcomes. Then $N_1 \sim \mathrm{Bin}(n, p_1)$ and $N_2 = n - N_1$. If np_1 and $np_2 = n(1-p_1)$ are both more than about 5, we can treat the binomial random variable N_1 as approximately $N(np_1, np_1p_2)$ so

$$W = \frac{(N_1 - np_1)^2}{np_1p_2} \tag{1.4}$$

Is approximately $\chi^2(1)$-distributed. If we use a little algebra, specifically $\frac{1}{p_1p_2} = \frac{1}{p_1} + \frac{1}{p_2}$ and $N_1 - np_1 = -(N_2 - np_2)$, we can show

$$W = \frac{(N_1 - np_1)^2}{np_1} + \frac{(N_2 - np_2)^2}{np_2} \tag{1.5}$$

which is a form of the chi-squared statistic (with $k-1 = 1$ degree of freedom) which treats outcomes 1 and 2 on equal footing.

Now consider the case where $k > 2$. The random variables $N_1, N_2, \ldots, N_k$ are not independent, but obey what's called a *multinomial distribution*, whose pmf (for every possible combination of non-negative integers with $n_1 + n_2 + \cdots + n_k = n$) is

$$p(n_1, \ldots, n_k) = \frac{n!}{n_1!n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k} \tag{1.6}$$

In fact, the last random variable $N_k = n - N_1 - N_2 - \cdots - N_{k-1}$ is redundant; all the information is carried in the first $k-1$ variables. It turns out that, as long as all the expected number counts $\{np_i\}$ (including np_k) are at least 5, the statistic

$$W = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} \tag{1.7}$$

written in analogy with (1.5) is a chi-squared random variable with $k-1$ degrees of freedom. So the statistic value (to be compared to $\chi^2_{\alpha,k-1}$) for a specific set of number counts is

$$\sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \tag{1.8}$$

1.1.1 Example: Twins

Consider 50 sets of twins. If they are all fraternal, and boys and girls are equally likely, we'd expect on average 12.5 boy-boy pairs, 25 mixed pairs, and 12.5 girl-girl pairs.

	boy-boy	mixed	girl-girl
obs	8	23	19
exp	12.5	25	12.5

We can calculate

$$\chi^2 = \frac{(8-12.5)^2}{12.5} + \frac{(23-25)^2}{25} + \frac{(19-12.5)^2}{12.5} = 5.16 \tag{1.9}$$

and note that $\chi^2_{.10,2} \approx 4.605$ while $\chi^2_{.05,2} \approx 5.992$ so this hypothesis would be rejected at the 10% level but not the 5% level.

1.2 Chi-Squared Test for a Specified Distribution

We can also apply the chi-squared test to a histogram of data hypothesized to be a sample drawn from a specified distribution. Then the number counts are the number of observations which fall into each bin. For example, suppose we want to test the hypothesis that a set of 80 observations is a sample from an exponential distribution with rate parameter $\lambda = \ln 2$. The cdf is

$$F(x) = 1 - e^{-x \ln 2} = 1 - \frac{1}{2^x} \tag{1.10}$$

so the expected number of observations lying between x_1 and x_2 is $\frac{n}{2^{x_1}} - \frac{n}{2^{x_2}}$. Suppose we divide into four bins, and obtain the following observed and expected histogram.

	$x < 1$	$1 \le x < 2$	$2 \le x < 3$	$3 \le x$
obs	40	21	14	5
exp	40	20	10	10

The chi-squared statistic will be

$$\begin{aligned}\chi^2 &= \frac{(40-40)^2}{40} + \frac{(21-20)^2}{20} + \frac{(14-10)^2}{10} + \frac{(5-10)^2}{10} \\ &= \frac{0}{40} + \frac{1}{20} + \frac{16}{10} + \frac{25}{10} = \frac{1+32+50}{20} = \frac{83}{20} = 4.15\end{aligned} \tag{1.11}$$

We can compare this to the percentiles of $\chi^2(3)$ distribution. Since $\chi^2_{.10,3} \approx 6.251$, we see that the P-value is greater than 10%, and we would not reject this model at any reasonable confidence level. Note that we could choose different bins than the ones we did; in practice it's good to choose the bins to have similar numbers of expected observations.

Practice Problems

14.1, 14.3

Tuesday 10 April 2018

2 Chi-Squared Tests with Estimated Parameters

In both of the examples from last time, we assumed that the expected numbers for each category were known, but that required an assumption in each case. Considering the problem of twin distribution, we assumed that half of all children were male, leading to expected numbers of

boy-boy	mixed	girl-girl
$n/4$	$n/2$	$n/4$

If the fraction of male children is θ, this becomes

boy-boy	mixed	girl-girl
$n\theta^2$	$2n\theta(1-\theta)$	$n(1-\theta)^2$

we've replaced the fraction expected in category k, which was previously a fixed number p_i, with a function $\pi_i(\theta)$. And in principle, we can imagine that there would be not just one parameter θ, but a set of parameters $\boldsymbol{\theta} \equiv \{\theta_1, \theta_2, \ldots, \theta_m\}$, where

$m < k - 1$. The chi-square statistic would then be

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{k} \frac{[n_i - n\pi_i(\boldsymbol{\theta})]^2}{n\pi_i(\boldsymbol{\theta})} \tag{2.1}$$

To actually evaluate this for a test, we need to put in an estimate $\hat{\boldsymbol{\theta}}$ for the parameters. One choice would be to find the one which minimizes $\chi^2(\boldsymbol{\theta})$ for the observed $\{n_i\}$, but instead, we will use the maximum likelihood estimate, i.e., the one which maximizes

$$f(n_1, \ldots, n_k; \boldsymbol{\theta}) \propto [\pi_1(\boldsymbol{\theta})]^{n_1} [\pi_2(\boldsymbol{\theta})]^{n_2} \cdots [\pi_k(\boldsymbol{\theta})]^{n_k} \tag{2.2}$$

or equivalently

$$\ln f(n_1, \ldots, n_k; \boldsymbol{\theta}) = \sum_{i=1}^{k} n_i \ln \pi_i(\boldsymbol{\theta}) + \ln \frac{n!}{n_1! \cdots n_k!} \tag{2.3}$$

where the $n_1, \ldots, n_k$ we use are the actual observed values. For example, in the twin example, the log-likelihood is

$$\begin{aligned} &\ln f(n_1, \ldots, n_k; \theta) \\ &\quad = 2n_1 \ln\theta + n_2[\ln\theta + \ln(1-\theta))] + 2n_3 \ln(1-\theta) + \text{const} \\ &\qquad = (2n_1 + n_2)\ln\theta + (n_2 + 2n_3)\ln(1-\theta) + \text{const} \end{aligned} \tag{2.4}$$

where the constant depends on n_1, n_2 and n_3, but not θ. Differentiating with respect to θ gives the maximum likelihood equation

$$\frac{2n_1 + n_2}{\hat{\theta}} - \frac{n_2 + 2n_3}{1 - \hat{\theta}} = 0 \tag{2.5}$$

which has the solution

$$\hat{\theta} = \frac{2n_1 + n_2}{2n} = \frac{16 + 23}{100} = .39 \tag{2.6}$$

We can plug that into the formulas for the expected numbers of twins in each category (for $n = 50$) and get

	boy-boy	mixed	girl-girl
obs	8	23	19
exp	$n\theta^2$	$2n\theta(1-\theta)$	$n(1-\theta)^2$
exp for $\theta = .39$	7.605	23.790	18.605

The chi-squared is then

$$\chi^2 = \frac{(8 - 7.605)^2}{7.605} + \frac{(23 - 23.790)^2}{23.790} + \frac{(19 - 18.605)^2}{18.605} \approx .0551 \tag{2.7}$$

However, since we've used the data to fit one parameter, the percentiles we should compare it to are not $\chi^2(2)$ but $\chi^2(2-1) = \chi^2(1)$. In general, if we have k categories and m functionally independent parameters (meaning no one parameter can be completely determined from the other $m - 1$), the resulting statistic will be approximately $\chi^2(k - 1 - m)$.

2.1 Chi-Squared for a Parametrized Distribution

We can also return to the application where we apply the chi-squared test to a binned histogram to check a hypothesized distribution. Last time we gave an example where we observed the following number counts for a data sample of size $n = 80$:

$x < 1$	$1 \le x < 2$	$2 \le x < 3$	$3 \le x$
40	21	14	5

Last time we hypothesized an exponential distribution with $\lambda = \ln 2$, but suppose we allowed λ to be an unknown parameter. Then the cdf of the distribution would be

$$F(x) = 1 - e^{-\lambda x} \tag{2.8}$$

and the expected number counts in the bins would be

$x < 1$	$1 \le x < 2$	$2 \le x < 3$	$3 \le x$
$n(1 - e^{-\lambda})$	$n(e^{-\lambda} - e^{-2\lambda})$	$n(e^{-2\lambda} - e^{-3\lambda})$	$ne^{-3\lambda}$

If we wanted to follow the maximum likelihood prescription from the previous section, we would need to find the λ value which maximizes

$$(1 - e^{-\lambda})^{n_1}(e^{-\lambda} - e^{-2\lambda})^{n_2}(e^{-2\lambda} - e^{-3\lambda})^{n_3}(e^{-3\lambda})^{n_4} \tag{2.9}$$

This seems like a pretty complicated function, although it turns out that it is maximized by setting λ to

$$\hat{\lambda} = \ln \frac{n_1 + 2n_2 + 3n_3 + 3n_4}{n_2 + 2n_3 + 3n_4} \tag{2.10}$$

and the resulting statistic is approximately $\chi^2(3)$-distributed. But in general it may be difficult and/or require a numerical method to find the parameters $\hat{\boldsymbol{\theta}}$ which maximize

$$\sum_{i=1}^{k} n_i \ln[F(b_i; \boldsymbol{\theta}) - F(a_i; \boldsymbol{\theta})] \tag{2.11}$$

where the ith histogram bin covers $a_i < x \le b_i$, which is what we'd need to get a statistic which was χ^2 distributed with $k - 1 - m$ degrees of freedom. Instead, it is often convenient to use some other point estimate constructed not just from the binned histogram, but from the full sample of n data values. For instance, for the exponential model we could use $\lambda = 1/\overline{x}$. If we do this, though the χ^2 value will generally not be as low as if we had really minimized over $\boldsymbol{\theta}$. This means that the threshold c_α for a test with false alarm probability α will be between the value $\chi^2_{\alpha,k-1}$ we'd use if we had fixed the parameters arbitrarily and the (lower) value $\chi^2_{\alpha,k-1-m}$ we'd use if we had minimized the χ^2 over the m parameters, i.e.,

$$\chi^2_{\alpha,k-1-m} \le c_\alpha \le \chi^2_{\alpha,k-1} \tag{2.12}$$

Practice Problems

14.15, 14.17

Thursday 12 April 2018

3 Two-Way Contingency Tables

We now consider another categorical data analysis problem, in which we have two sets of categories, and each observation falls into one category from the first set and one from the second set. We'd like to know if the data indicate any statistical relationship between which of the first set of categories an observation falls into and which of the second. For instance, suppose we survey students majoring in four disciplines about their food choices:

	Vegan	Vegetarian	Non-Veg	Total
Math & Stat	9	23	49	81
Physics	6	15	30	51
Chemistry	12	28	62	102
Biology	25	50	98	173
Total	52	116	239	407

We'd like to know if the data indicate any significant tendencies for students in one major to have one diet or another. There are two ways to pose the question:

1. Is there any difference in the tendencies of students in one major or another to have a vegan, vegetarian, or non-vegetarian diet? (homogeneity)
2. Is there any correlation between the major chosen by a student and their dietary choices? (independence)

The two questions have different interpretations in the framework of classical statistics, but they turn out to lead to identical data analysis procedures.

As a matter of notation, we'll consider the table to have I rows and J columns, labelled by $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$, respectively. (In the table above, I is 4 and J is 3.) We'll call the observed number in each cell $n(i, j)$. (Devore calls this N_{ij}.) We'll write the total of the numbers in row i as $n(i, \cdot) = \sum_{j=1}^{J} n(i, j)$ (Devore: $n_{i\cdot}$) and in column j as $n(\cdot, j) = \sum_{i=1}^{I} n(i, j)$ (Devore: $n_{\cdot j}$). The total number of observations is

$$n = \sum_{i=1}^{I} n(i, \cdot) = \sum_{i=1}^{I} \sum_{j=1}^{J} n(i, j) = \sum_{j=1}^{J} n(\cdot, j) \tag{3.1}$$

3.1 Perspective 1: Test for homogeneity

In the first way of looking at things, the number $n(i, \cdot)$ in each row is a given, and for each i we consider $\{N(i, j)|j = 1, \ldots J\}$ to be a multinomial random vector with probabilities $p(j|i)$. (Devore calls this p_{ij}.) Note that this is not a conditional probability according to the set theory definition of Devore Chapter Two, but it's the probability for an observation which is in row i to fall into column j. These probabilities obey $\sum_{j=1}^{J} p(j|i) = 1$ for each i. The most important property of the multinomial for us is that

$$E\left(N(i, j)\right) = n(i, \cdot) p(j|i) \tag{3.2}$$

The null hypothesis of homogeneity is that each of these I sets of J probabilities is the same:

$$H_0 : \qquad p(j|i) = p(\cdot, j) \quad \text{for all } I \tag{3.3}$$

We can estimate this common probability as $\hat{p}(\cdot, j) = \frac{n(\cdot, j)}{n}$ so that the estimated expected number of items in each cell is

$$\hat{e}(i, j) = n(i, \cdot) \hat{p}(\cdot, j) = \frac{n(i, \cdot)\, n(\cdot, j)}{n} \tag{3.4}$$

We then use these estimated expected numbers to make a chi-squared statistic for the whole table's divergence from homogeneity:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{[n(i, j) - \hat{e}(i, j)]^2}{\hat{e}(i, j)} \tag{3.5}$$

We have I multinomial random variables with J categories each, which means we've observed $I(J - 1)$ independent numbers. We've estimated J probabilities, but only $J - 1$ of them were independent because they had to add to 1. Thus the number of degrees of freedom for the chi-squared should be

$$I(J - 1) - (J - 1) = (I - 1)(J - 1) \tag{3.6}$$

Note that although the formalism treats the rows and columns rather differently, the final data analysis prescription treats them symmetrically.

3.2 Perspective 2: Test for independence

In the second way of looking at things, the only thing we treat as given is the total number of observations n, and the observation $\{N(i, j)|i = 1, \ldots I; j = 1, \ldots J\}$ is a (IJ-dimensional) multinomial random vector with probabilities $p(i, j)$. (Devore *also* calls this p_{ij}, which is one reason I wanted to make the notation more descriptive.) These probabilities obey $\sum_{i=1}^{I} \sum_{j=1}^{J} p(i, j) = 1$. Now the multinomial says

$$E\left(N(i, j)\right) = np(i, j) \tag{3.7}$$

The null hypothesis of independence is that the probabilities can be decomposed assuming independence of the two sets of categories:

$$H_0 : \qquad p(i, j) = p(i, \cdot)\, p(\cdot, j) \tag{3.8}$$

We can estimate the probabilties as $\hat{p}(i,\cdot) = \frac{n(i,\cdot)}{n}$ and $\hat{p}(\cdot,j) = \frac{n(\cdot,j)}{n}$ so that the estimated expected number of items in each cell is

$$\hat{e}(i,j) = n\hat{p}(i,\cdot)\hat{p}(\cdot,j) = n\,\frac{n(i,\cdot)\,n(\cdot,j)}{n^2} = \frac{n(i,\cdot)\,n(\cdot,j)}{n} \tag{3.9}$$

which is exactly what we saw above[1] We then make the chi-squared as before:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{[n(i,j)-\hat{e}(i,j)]^2}{\hat{e}(i,j)} \tag{3.10}$$

We have a single multinomial with IJ categories now, so there are $IJ-1$ independent observations. We've estimated I probabilities $\{\hat{p}(i,\cdot)$, $I-1$ of which are independent, and J probabilities $\{\hat{p}(\cdot,j)$, $J-1$ of which are independent, so we have

$$IJ-1-(I-1)-(J-1) = IJ-I-J+1 = (I-1)(J-1) \tag{3.11}$$

which, again, is the same number of degrees of freedom as the homogeneity test told us.

3.2.1 Example

We return to the example above. The estimates according to the model are $\frac{(81)(52)}{407} \approx 10.35$, $\frac{(81)(116)}{407} \approx 23.09$, etc.:

	Vegan	Vegetarian	Non-Veg	Total
Math & Stat	10.35	23.09	47.57	81
Physics	6.52	14.54	29.95	51
Chemistry	13.03	29.07	59.90	102
Biology	22.10	49.31	101.59	173
Total	52	116	239	407

Note that it's really easy to do this in a simple spreadsheet program. If I construct the chi-squared statistic I get 0.99. I need to compare this to a chi-squared distribution with $(4-1)(3-1) = (3)(2) = 6$ degrees of freedom, so 0.99 is rather low indeed and the data show no significant correlation.

Practice Problems

14.27, 14.34, 14.35, 14.41

[1] Note that viewed as a statistic, the estimate from independence is $\hat{e}(i,j) = \frac{N(i,\cdot)\,N(\cdot,j)}{n}$ while that from homogeneity is $\hat{e}(i,j) = \frac{n(i,\cdot)\,N(\cdot,j)}{n}$, but this distinction has no effect on the data analysis prescription.