Chapter 10

Categorical Data Analysis

In Section 8.6, we learned to compare two population proportions. We can extend this approach to more than two populations (groups) by the means of a chi-square test.

Consider the experiment of randomly selecting n items, each of which belongs to one of k categories (for example, we collect a sample of 100 people and look at their blood types, and there are $k = 4$ types). We will count the number of items in our sample of the type i and denote that X_i. We will refer to X_i as *observed count* for category i. Note that $X_1 + X_2 + ... + X_k = n$.

We will be concerned with estimating or testing the probabilities (or proportions) of ith category, p_i, $i = 1, ..., k$. Also, keep in mind the restriction $\sum_i p_i = 1$.

There are two types of tests considered in this Chapter:

- A test for **goodness-of-fit**, that is, how well do the observed counts X_i fit a given distribution.

- A test for **independence**, for which there are two classification categories (variables), and we are testing the independence of these variables.

10.1 Chi-square goodness-of-fit test

This is a test for the fit of the sample proportions to given numbers. Suppose that we have observations that can be classified into each of k groups (categorical data). We would like to test

$$H_0 : p_1 = p_1^0,\ p_2 = p_2^0,\ \dots\ , p_k = p_k^0$$

$$H_A :\ \text{some of the } p_i\text{'s are unequal to } p_i^0\text{'s}$$

where p_i is the probability that a subject will belong to group i and $p_i^0, i = 1, ..., k$ are given numbers. (Note that $\sum p_i = \sum p_i^0 = 1$, so that p_k can actually be obtained from the rest of p_i's.)

Our data (*Observed counts*) are the counts of each category in the sample, $X_1, X_2, ..., X_k$ such that $\sum_{i=1}^{k} X_i = n$. The total sample size is n. For $k = 2$ we would get X_1 = number of successes, and $X_2 = n - X_1$ = number of failures, that is, Binomial distribution. For $k > 2$ we deal with *Multinomial distribution*.

For testing H_0, we compare the observed counts X_i to the ones we would expect under null hypothesis, that is,

$$\text{Expected counts} \qquad E_1 = np_1^0, \; .... \; , E_k = np_k^0$$

To adjust for the size of each group, we would take the squared difference divided by E_i, that is $(E_i - X_i)^2/E_i$. Adding up, we obtain the

$$\text{Chi-square statistic} \qquad \chi^2 = \sum_{i=1}^{k} \frac{(E_i - X_i)^2}{E_i} \tag{10.1}$$

with $k-1$ degrees of freedom

We would reject H_0 when χ^2 statistic is large (that is, the Observed counts are far from Expected counts). Thus, our test is always *one-sided.* To find the p-value, use χ^2 upper-tail probability table very much like the t-table. See Table C.

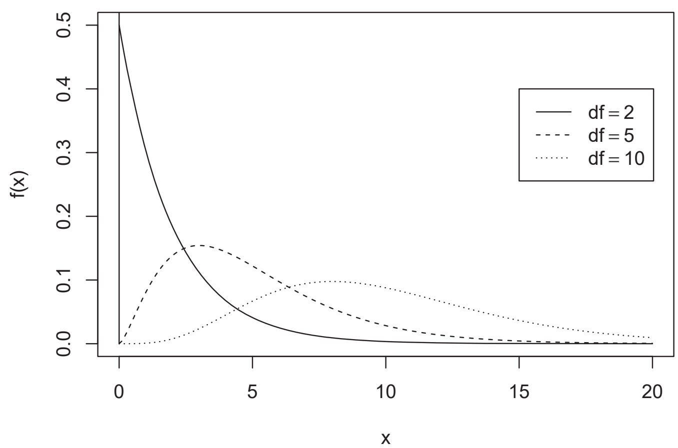

Figure 10.1: Chi-square densities

Assumption for chi-square test: all Expected counts should be ≥ 5 (this is necessary so that the normal approximation for counts X_i holds.) Some details: see below[1]

[1]Chi-square distribution with degrees of freedom $= k$ is related to Normal distribution as follows:

$$\chi^2 = Z_1^2 + Z_2^2 + ... + Z_k^2,$$

where $Z_1, ..., Z_k$ are independent, standard Normal r.v.'s.

Also, it can be shown that chi-square (`df` $= k$) distribution is simply *Gamma*$(\alpha = k/2, \beta = 2)$– sorry, this α and the significance level for testing are not the same!

For example, Chi-square(`df` $= 2$) is the same as Exponential ($\beta = 2$). (Why?)
Note that this distribution has positive values and is not symmetric!

Table C: Critical points of the chi-square distribution

	Upper tail probability						
Degrees of freedom	0.100	0.050	0.025	0.010	0.005	0.001	0.0005
1	2.706	3.841	5.024	6.635	7.879	10.828	12.116
2	4.605	5.991	7.378	9.210	10.597	13.816	15.202
3	6.251	7.815	9.348	11.345	12.838	16.266	17.730
4	7.779	9.488	11.143	13.277	14.860	18.467	19.997
5	9.236	11.070	12.833	15.086	16.750	20.515	22.105
6	10.645	12.592	14.449	16.812	18.548	22.458	24.103
7	12.017	14.067	16.013	18.475	20.278	24.322	26.018
8	13.362	15.507	17.535	20.090	21.955	26.124	27.868
9	14.684	16.919	19.023	21.666	23.589	27.877	29.666
10	15.987	18.307	20.483	23.209	25.188	29.588	31.420
11	17.275	19.675	21.920	24.725	26.757	31.264	33.137
12	18.549	21.026	23.337	26.217	28.300	32.909	34.821
13	19.812	22.362	24.736	27.688	29.819	34.528	36.478
14	21.064	23.685	26.119	29.141	31.319	36.123	38.109
15	22.307	24.996	27.488	30.578	32.801	37.697	39.719
16	23.542	26.296	28.845	32.000	34.267	39.252	41.308
17	24.769	27.587	30.191	33.409	35.718	40.790	42.879
18	25.989	28.869	31.526	34.805	37.156	42.312	44.434
19	27.204	30.144	32.852	36.191	38.582	43.820	45.973
20	28.412	31.410	34.170	37.566	39.997	45.315	47.498
21	29.615	32.671	35.479	38.932	41.401	46.797	49.011
22	30.813	33.924	36.781	40.289	42.796	48.268	50.511
23	32.007	35.172	38.076	41.638	44.181	49.728	52.000
24	33.196	36.415	39.364	42.980	45.559	51.179	53.479
25	34.382	37.652	40.646	44.314	46.928	52.620	54.947
30	40.256	43.773	46.979	50.892	53.672	59.703	62.162
40	51.805	55.758	59.342	63.691	66.766	73.402	76.095
60	74.397	79.082	83.298	88.379	91.952	99.607	102.695
80	96.578	101.879	106.629	112.329	116.321	124.839	128.261
100	118.498	124.342	129.561	135.807	140.169	149.449	153.167

Example 10.1.

When studying earthquakes, we recorded the following numbers of earthquakes (1 and above on Richter scale) for 7 consecutive days in January 2008.

Day	1	2	3	4	5	6	7	Total
Count	85	98	79	118	112	135	137	764
Expected	109.1	109.1	109.1	109.1	109.1	109.1	109.1	764

Here, $n = 764$. Is there evidence that the rate of earthquake activity changes during this week?

Solution. If the null hypothesis $H_0 : p_1 = p_2 = ... = p_7$ were true, then each $p_i = 1/7$, $i = 1, ..., 7$. Thus, we can find the expected counts $E_i = 764/7 = 109.1$.

Results: $\chi^2 = 28.8$, $df = 6$, p-value < 0.0005 from Table C.(The highest number there, 24.103, corresponds to upper tail area 0.0005.) Since the p-value is small, we reject H_0 and claim that the earthquake frequency **does** change during the week.[2] □

Example 10.2.

In this example, we will test whether a paricular distribution matches our experimental results. These are the data from the probability board (quincunx), we test if the distribution is really Binomial (as is often claimed). The slots are labeled 0-19. Some slots were merged together (why?)

Slots	0-6	7	8	9	10	11	12	13-19	Total
Observed	16	2	11	18	14	14	7	18	100
Expected	8.4	9.6	14.4	17.6	17.6	14.4	9.6	8.4	100

Solution. The expected counts are computed using Binomial($n = 19, p = 0.5$) distribution, and then multiplying by the $Total = 100$. For example,

$$E_9 = \binom{19}{9} 0.5^9 (1 - 0.5)^{19-9} \times 100 = 17.6$$

Next, $\chi^2 = 26.45$, $df = 7$, and p-value < 0.0005.
Conclusion: Reject H_0, the distribution is not exactly Binomial. □

10.2 Chi-square test for independence

This test is applied to the category probabilities for two variables. Each case is classified according to variable 1 (for example, Gender) and variable 2 (for example, College Major). The data are usually given in a *cross-classification* table (a 2-way table). Let X_{ij} be the observed table counts for row i and column j.

We are interested in testing whether Variable 1 (in r rows) is independent of Variable 2 (in c columns).[3]

[2]We did not specify α for this example. As mentioned earlier, $\alpha = 0.05$ is a good "default" choice. Even if we pick a conservative $\alpha = 0.01$, we would still reject H_0 here.

[3]These are not random variables in the sense of Chapter 3, because they are *categorical*, not numerical.

In this situation, we set up a chi-square statistic following equation (10.1). However, now the table is bigger. The Expected counts will be found using independence assumption, as

$$\text{Expected counts } E_{ij} = \frac{R_i C_j}{n}, \quad i = 1, ..., r \quad j = 1, ..., c$$

where R_i and C_j are the row and column totals.

Theorem 10.1. Chi-square test for independence

To test

$$H_0: \text{ Variable 1 is independent of Variable 2 } vs$$

$$H_A: \text{ Variable 1 is } \textbf{not} \text{ independent of Variable 2}$$

we can use the χ^2 random variable with $df = (r-1)(c-1)$, where

$$\text{test statistic} \quad \chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(E_{ij} - X_{ij})^2}{E_{ij}} \tag{10.2}$$

Example 10.3.
Suppose that we ordered 50 components from each of the vendors A, B and C, and the results are as follows

	Succeeded	Failed	Total
Vendor A	49	1	50
Vendor B	45	5	50
Vendor C	41	9	50

We would like to investigate whether all the vendors are equally reliable. That is,

$$H_0: \text{ Failure rate is independent of Vendor}$$

$$H_A: \text{ Not all Vendors have the same failure rate}$$

Solution. We'll put all the expected counts into the table

```
Expected counts:
            Succeeded   Failed   Total
Vendor A       45          5       50
Vendor B       45          5       50
Vendor C       45          5       50
 -----------------------------------------------------
Total         135         15      150
```

The χ^2 statistic will have $df = (3-1)(2-1) = 2$.

Here, $\chi^2 = (45-49)^2/45 + (1-5)^2/5 + ... = 7.11$. Since χ^2 statistic is between table values 5.991 and 7.378, the p-value is between 0.025 and 0.05. At the standard $\alpha = 0.05$ we are rejecting H_0. Thus, there **is** evidence that vendors have different failure rates.[4] □

[4]For this particular example, since `df` $= 2$, there is a more exact p-value calculation based on Exponential distribution: $P(Y > 7.11) = \exp(-7.11/2) = 0.0286$. For `df` $\neq 2$, we can use `R` function `pchisq`, Excel function `chidist` or other software to compute the exact p-values.

Exercises

10.1.
In testing how well people can generate random patterns, the researchers asked everyone in a group of 20 people to write a list of 5 random digits. The results are tabulated below

Digits	0	1	2	3	4	5	6	7	8	9	Total
Observed	6	11	10	13	8	13	7	17	8	7	100

Are the digits completely random or do humans have preference for some particular digits over the others?

10.2.
Forensic statistics. To uncover rigged elections, a variety of statistical tests might be applied. For example, made-up precinct totals are sometimes likely to have an excess of 0 or 5 as their last digits. For a city election, the observers counted that 21 precinct totals had the last digit 0, 18 had the last digit 5, while 102 had some other last digit. Is there evidence that the elections were rigged?

10.3.
In an earlier example of Poisson distribution, we discussed the number of Nazi bombs hitting $0.5 \times 0.5km$ squares in London. The following were counts of squares that have $0, 1, 2, ...$ hits:

number of hits	0	1	2	3	4 and up
count	229	211	93	35	8

Test whether the data fit the Poisson distribution (for $p_1^0, ...p_k^0$ use the Poisson probabilities, with the parameter μ estimated as average number of hits per square, $\mu = 0.9288$).

10.4.
To test the attitudes to a tax reform, the state officials collected data of the opinions of likely voters, along with their income level

	Income Level:		
	Low	Medium	High
For	182	213	203
Against	154	138	110

Do the people with different incomes have significantly different opinions on tax reform? (That is, test whether the Opinion variable is independent of Income variable.)

10.5.
Using exponential distribution, confirm the calculation of chi-square ($\mathtt{df} = 2$) critical points from Table C for upper tail area $\alpha = 0.1$ and $\alpha = 0.005$. Find the point for $\chi^2(\mathtt{df} = 2)$ distribution with $\alpha = 0.2$

Notes

[t] Kotswara Rao Kadilyala (1970). "Testing for the independence of regression disturbances" Econometrica, 38, 97-117. Appears in: *A Handbook of Small Data Sets*, D. J. Hand, et al, editors (1994). Chapman and Hall, London.

[u]from *The R book* by Michael Crawley

[v]Mlodinow again. The director, Sherry Lansing, was subsequently fired only to see several films developed during her tenure, including *Men In Black*, hit it big.

[w]see `http://www.akdart.com/postrate.html`