

AnSc 5403
Biometry

Lecture Notes 27

I. Chi-square tests for frequency data – Reference – Remington and Schork (1970) and Conover, W. J. (1971), *Practical Nonparametric Statistics*, John Wiley & Sons, Inc.

A. Observational data are often in the form of counts or frequencies

1. The basic question that is typically asked in such frequency of count data is whether the observed frequencies or counts conform to (agree with) the expected frequencies or counts
2. The basic test statistic used to deal with such data is the chi-square (χ^2) statistic, which is given by:

a.
$$\chi^2 = \sum_{i=1}^k (O_i - E_i)^2 / E_i$$

- b. Karl Pearson showed that this statistic is approximately distributed by the chi-square distribution
- c. The df for the statistic is given by –
 - (a) Where k is the total number of observed frequencies and m is the number of parameters in the model that must be estimated from the data
3. Our basic hypothesis in the χ^2 test is that the observed frequencies or counts agree with the expected frequencies or counts
 - a. If the hypothesis is false, we would expect the values of “O – E” to be large, which would yield a large value for the χ^2 test statistic
 - (a) Because the “O – E” values are squared, the value of χ^2 will always be positive, so chi-square frequency tests are inherently –
 - (i) The critical value of the test statistic (χ^2) is the (100- α) percentile of the chi-square distribution, and the critical region is values of χ^2 that are greater than or equal to the value of the tabular value for $\chi^2_{1-\alpha}$

B. Contingency tables

1. A contingency table is an array of numbers in a matrix form, where the numbers represent –
2. Example – an entomologist observes 37 insects and counts the number of live and dead insects within three categories

	Moths	Grasshoppers	Others	Total
Alive	3	21	3	27
Dead	9	1	0	10
Total	12	22	3	37

- a. This would be a 2 x 3 (rows x columns or r x c) contingency table
3. We will consider two examples of using r x c contingency tables:
 - a. The chi-square test for –
 - b. The chi-square test for –
- C. The chi-square test for differences in probabilities
 1. In this test we have “r” populations, and –
 - a. Each observation in each random sample may be –

	Class 1	Class 2	...	Class c	Totals
Population 1	O_{11}	O_{12}	...	O_{1c}	n_1
Population 2	O_{21}	O_{22}	...	O_{2c}	n_2
...
Population r	O_{r1}	O_{r2}		O_{rc}	n_r
Totals	C_1	C_2	...	C_c	N

2. Thus, our assumptions for this test are as follows:
 - a. Each sample is a random sample
 - b. The outcomes of among samples are mutually independent
 - c. Each observation may be classified into one of the “c” classes
3. Hypotheses – the most common hypotheses tested with this type of r x c contingency table are as follows:

- a. H_0 : All probabilities in the same column are equal to each other
 - b. H_A : At least two of the probabilities in the same column are not equal to each other
4. The test statistic is given by:
- a. $\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$, where $E_{ij} = \frac{n_i C_j}{N}$, where n_i is the row total for a given population, C_j is the column total for a given class, and $N =$ the grand total
5. Example (from Conover, 1971):
- a. Randomly selected samples of students from private and public schools were given standardized achievement tests, with the following results:

	0 - 275	276 - 350	351 - 425	426 - 500	Totals
Private school	6	14	17	9	46
Public school	30	32	17	3	82
Totals	36	46	34	12	128

- b. We want to test the H_0 that the –
 - (a) NOTE: Before we calculate the test statistic, we need to note that in this type of contingency table, an important rule to follow is that if any of the E_{ij} is less than 1, and 20% of the E_{ij} are less than 5, the chi-square approximation is likely poor
- c. First, let's compute the values for E_{ij}
 - (a) For the C_{11} , the E_{ij} would equal the Row 1 total (46) multiplied by the Column 1 total (36), with the product divided by the grand total (128)
 - (i) $(46 \times 36)/128 = 12.93$ (we will round to 12.9 for the table)
 - (b) For C_{24} , the comparable value would be given by $(82 \times 12)/128 = 7.7$

(c) Each E_{ij} can be calculated in the same way, yielding the following table of E_{ij}

0 - 275	276 - 350	351 - 425	426 - 500
12.9	16.5	12.2	4.3
23.1	29.5	21.8	7.7

6. Now we can calculate the test statistic as follows:

$$\begin{aligned} \text{a. } & ([6 - 12.9]^2/12.9) + ([14 - 16.5]^2/16.5) + ([17 - 12.2]^2/12.2) + ([9 - 4.3]^2/4.3) + \\ & ([30 - 23.1]^2/23.1) + ([32 - 29.5]^2/29.5) + ([17 - 21.8]^2/21.8) + ([3 - 7.7]^2/7.7) = \\ & 17.3 \end{aligned}$$

7. This test statistic has $(r - 1)(c - 1)$ df, or $(2 - 1)(4 - 1) = 3$ df

a. Note that if we use our rule of $k - 1 - m$, that means k would equal $r \times c$, or 8 in this case

(a) So, $8 - 1 - 4 = 3$, which means that we had to estimate four parameters in the model (in this case, the value of four parameters estimated is derived from three probabilities from the columns and one for the rows)

8. Our rule about any of the E_{ij} being less than 1 and 20% being less than 5 is met, so we can evaluate our test statistic

a. If we set $\alpha = 0.05$, the critical value for the $\chi^2_{1-\alpha}$ with 3 df is –

b. Thus, we would conclude that the distribution of test scores differs between public and private school students

D. The chi-square test for independence

1. In this type of chi-square test, a single random sample of size N is taken

a. Each observation in the random sample can be classified according to two criteria

(a) One of these criteria is associated with one of the “r” rows

(b) The other is associated with the “c” columns

- b. The “layout” of the data would be as follows:

	Column 1	Column 2	...	Column c	Totals
Row 1	O_{11}	O_{12}	...	O_{1c}	R_1
Row 2	O_{21}	O_{22}	...	O_{2c}	R_2
...
Row r	O_{r1}	O_{r2}		O_{rc}	R_r
Totals	C_1	C_2	...	C_c	N

2. Our assumptions for this test will be:
 - a. The sample of N observations is a random sample (thus, each observation has the same probability of being classified in a row or column independent of other observations)
 - b. Each observation may be classified into exactly one of the “ r ” categories and exactly one of the “ c ” categories
3. Hypotheses – the most common hypotheses tested with this type of $r \times c$ contingency table are as follows:
 - a. H_0 : The event that an observation is in a particular row is independent of the event that an observation is in a particular column
 - (a) Essentially, this means that the probability that a given observation is in a particular row and column is equal to the probability that it is in a given row multiplied by the probability that it is in a given column
 - b. H_A : The event that an observation is in a particular row is **not** independent of the event that an observation is in a particular column
4. Our test statistic is calculated exactly as before:
 - a. $\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$, where $E_{ij} = \frac{R_i C_j}{N}$, with R_i and C_j being the row and column totals and N being the grand total

5. Example (from Conover, 1971):

- a. A random sample of students ($N = 94$) at a certain university was classified according to the college in which they enrolled and by whether they graduated in state or out of state

	Engineering	Arts & Sciences	Home Economics	Other	Totals
In state	16	14	13	13	56
Out of state	14	6	10	8	38
Totals	30	20	23	21	94

- b. H_0 : The college in which the student enrolled is independent of whether they went to high school in state or out of state
- c. Using our formula for the test statistic, we would calculate that:
- (a) $\chi^2 = 1.55$
- (i) The df for this χ^2 is the same as before $(r - 1)(c - 1) = 3$
- d. With $\alpha = 0.05$, (our critical value of $\chi^2_{1-\alpha} = 7.81$) we would conclude we would –
6. You might be wondering how this test differs from the previous one (the chi-square test for differences in probabilities)
- a. The primary difference in the independence test is that we have only one random sample that is categorized in two ways (rows and columns).
- b. In the chi-square test for difference in probabilities, each row represented a random sample of known size (n_i)
- (a) With the chi-square test for independence, the row totals are –
- c. This difference actually affects the calculation of exact probabilities for the test statistics with each method
- (a) Nonetheless, the distributions of both test statistics can be approximated with the chi-square distribution with $(r - 1)(c - 1)$ df