DEPARTMENT OF POLITICAL SCIENCE
AND
INTERNATIONAL RELATIONS

Posc/Uapp 816

TESTS AND MEASURES OF ASSOCIATION

I. AGENDA:
  A. Cross-classifications
    1. The chi square test of independence between X and Y
    2. Independence versus association
    3. The odds ratio
  B. Reading:
    1. Agresti and Finlay, ***Statistical Methods for the Social Sciences***, 3rd edition, Chapter 8.

II. TEST FOR STATISTICAL INDEPENDENCE:
  A. Background: we have two categorical variables, X and Y.
  B. X is normally considered the independent or explanatory variable.
    1. The example considered last time involved turnout (Y) cross-classified by level of education (X).

Education/ voted	1	2	3	4	5	6	7	Totals
Yes	**54%** 36	**58%** 80	**69%** 328	**80%** 222	**82%** 115	**89%** 249	**92%** 142	1175
No	**46** 30	**42** 57	**31** 147	**45** 55	**18** 26	**11** 32	**8** 12	359
Totals	100% 66	100% 137	100% 475	100% 277	100% 141	100% 260	100% 154	1534

  C. We want to test the hypothesis that the X is statistically independent of Y.
    1. Statistical independence holds when

$$\boxed{\pi_{ij} = \pi_{i+} \, \pi_{+j}}$$

      a. where π_{ij} be the probability that a randomly chosen member of the population being classified in category i of Y and j of X and π_{i+} is the (marginal) probabilities of a randomly selected observation being in the ith category of Y and π_{+j} is the probability that a randomly selected observation is in the jth category of X.

2. We use the chi square statistic to test the hypothesis that the variables are independent.

D. Sampling distribution

1. Reprinted in part from the last (Class 5) set of notes.
2. When dealing with a mean or difference of means we are of course looking at sample statistics and their distributions.
3. We need to do the same in this case, namely find an appropriate sample statistic and determine its sampling distribution.
4. We proceed by asking what we would **expect** to observe **if** the null hypothesis of independence were true. What, in other words, would a sample drawn from a population of two independent variables be expected to "look" like?
5. We then compare this expected value with the actually observed frequency.
6. We do this for each frequency in the I X J table using the formula for finding estimated expected frequencies:

$$\boxed{\hat{F}_{ij} = \left(\frac{F_{i+}}{F_{++}}\frac{F_{+j}}{F_{++}}\right)F_{++} = \frac{F_{i+}F_{+j}}{F_{++}}}$$

   a. Here we use F's (instead of N's) to denote frequencies.
   b. F_{++} = N, the table total.

7. If H_0 holds, then the expected and observed frequencies (F_{ij}) should be the same except for sampling error.
   a. We have to estimate the expected frequencies since we do not normally know the population probabilities.

E. Test statistic:

1. That is, we compare F_{ij} ***with*** $\hat{F}_{ij}$, where F_{ij} is the **observed** frequency in the ijth combination of X and Y and $\hat{F}_{ij}$ is the **estimated expected** frequency under the null hypothesis.
   a. The observed and estimated expected frequencies (in parentheses) for the voting data are:

Observed/ expected	1	2	3	4	5	6	7	Totals
Yes	36 (50.55)	80 (104.93)	328 (363.84)	222 (212.17)	115 (108)	249 (199.15)	142 (117.96)	1175
No	30 (15.45)	57 (32.06)	147 (111.64)	55 (64.83)	26 (33)	32 (60.85)	12 (36.04)	359
Totals	66	137	475	277	141	260	154	1534

i. Note that the sum of expected frequencies equals the sum of the observed frequencies by both column and row.

ii. The expected frequencies are said to "fit" the row and column totals.

2. It's instructive and helpful to look through the contingency table cell by cell to see where departures from expected and observed frequencies are greatest.
   a. If the null hypothesis of independence does not hold, one wants to know where it breaks down. There may be, in other words, an association in one or more parts or sub-tables.
   b. For example, the expected and observed frequencies for levels 4 and 5 of education are relatively close together. The major discrepancies are in the ends of the table.
   c. In the 7th level, for instance, we have more voters than we would expect to find if the variables were independent.
   d. And we have fewer voters than expected in the 1st level of education.
3. We aggregate the differences between each observed and expected frequency into a chi square statistic, denoted χ^2
4. The formula for the observed chi square with (I - 1)(J - 1) degrees of freedom is

$$\chi^2_{obs} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(F_{ij} - \hat{F}_{ij})^2}{\hat{F}_{ij}}$$

5. The probability of observed chi square, χ^2, is given by the chi square distribution with degrees of freedom:

$$df = (I - 1)(J - 1)$$

   a. Once again, I is the number of categories of Y and J is the number of categories of X.

F. Critical value:

1. As with the means test, we need to find a critical value such that if the null hypothesis is true and if the sample result equals or exceeds it we have reason to reject the null hypothesis.
2. As noted above the sampling statistic, χ^2 , will have a chi square distribution with degrees of freedom (I - 1)(J - 1).
3. This distribution has been extensively tabulated.
    a. See Agresti and Finlay, ***Statistical Methods in the Social Sciences***, 3rd edition, Table C (page 670).
4. The voting data example table has (2 - 1)(7 - 1) = 6 degrees of freedom.
    a. We are interested in right tailed probabilities since we are only concerned with chi square values quite far 0.
    b. With 6 degrees of freedom the table entry is 12.59 for the .05 level; 16.81 for the .01 level; and 22.46 for the .001 level.

G. Decision:

1. In the voting example the components of the chi square--the individual squared deviations--are given in the following table.

Observed/ expected/ component	1	2	3	4	5	6	7	Totals
Yes	36 (50.55) **[4.19]**	80 (104.93) **[5.92]**	328 (363.84) **[3.53]**	222 (212.17) **[.00]**	115 (108) **[.454]**	249 (199.15) **[12.48]**	142 (117.96) **[4.90]**	1175
No	30 (15.45) **[13.70]**	57 (32.06) **[19.40]**	147 (111.64) **[11.20]**	55 (64.83) **[1.49]**	26 (33) **[1.48]**	32 (60.85) **[13.68]**	12 (36.04) **[14.27]**	359
Totals	66	137	475	277	141	260	154	1534

        i. We can see by looking at these components where independence breaks down.

2. The total chi square is 106.70.
3. Compared with the critical chi square the observed value is much larger so we reject the null hypothesis of statistical independence and conclude that there is some type of relationship between voting and education.

H. Interpretation

1. Note that although the chi square test is commonly used--it's a standard part of every statistical package--it is not as useful as it might seem, as the next section indicates.
2. Look in the body of the table to see the nature of the association.

III. REMARKS AND INTERPRETATION:

A. Reprinted from Class 5 notes.

B. Many of these comments apply to statistical testing in general.

C. The chi square test only allows us to infer independence or non independence.

1. If we reject H_0, we still do not know the "strength" or magnitude or nature of the relationship between X and Y.
2. Hence is usually important to calculate or obtain a measure of the strength of the relationship.

D. Chi square's numerical magnitude is affected by the sample size: as N (= F_{++}) increases, the chi square will always (except possibly under unusual circumstances) increase as well. This holds even if the nature or strength of the relationship stays the same.

1. Here's an example. Consider first the following hypothetical table in which there really isn't much of a relationship between the variables.

Y/X				Totals
	30	30	30	90
	30	30	36	96
	40	40	34	114
Totals	100	100	100	**300**

   a. In this case the cell entries can be interpreted as either frequencies or percentages. In either case, there is no real strong relationship between the variables. For instance, 30 percent of the each column is in the first row.

   b. The weak relationship is indicated by the observed chi square of 1.38 with 4 degrees of freedom. If these were random sample data, we could not reject the null hypothesis of statistical independence between X and Y.

2. Now simply multiply every frequency in the body of the table by 10. Doing so does not change the nature of the relationship. (30 percent of each column is still in the first category of Y.) But the magnitude of chi square increases by 10 times.

Y/X				Totals
	300	300	300	900
	300	300	360	960
	400	400	340	1140
Totals	1000	1000	1000	**3000**

   a. So the chi square is now 13.82, again with 4 degrees of freedom.

This is significant at the .01 level, meaning we would on the basis of this sample reject the null hypothesis of independence.

b. And perhaps we should since there is a very small departure from independence.

c. Nevertheless the fact that we obtained a highly significant chi square should not obscure the fact the relationship is trivial.

3. For further information see H. T. Reynolds ***The Analysis of Nominal Data***, 2nd edition.

E. Note also that the chi square test is not directed at any particular alternative hypothesis.

1. That is, it tests only statistical independence. But there might be a “linear’ trend in the population that we want to estimate or test for.

2. We’ll discuss these sorts of issues further.

F. In general, one can always find statistical significance if the sample size is large enough.

1. Hence we need to discuss the magnitude and nature of relationships.

IV. ASSOCIATION VERSUS INDEPENDENCE:

A. It is one thing to find that two variables are statistically independent. Usually, however, an investigator knows ahead of time that this hypothesis is untenable. After all, people do not usually compare randomly selected pairs of variables but examines those that theory or commonsense indicates will be related.

B. Consequently, we often want to know more than a statistically dependent relationship exists; we want to explore the **strength** and **nature** of the relationship.

C. We’ll start with a simple case.

V. THE ODDS RATIO:

A. The odds ratio $\boldsymbol{\Omega}$ is an index that measures the strength and “direction” of association in a 2 X 2 table.

B. Since we will normally deal with I X J tables, where I and J are the number of rows and columns respectively, we need a set of basic odds ratio to show the nature and direction of the relationships.

1. Having a set of $\boldsymbol{\Omega}$”s has both advantages and disadvantages. It’s a drawback because most analysts want to summarize the information in a table with a single number. On the other hand, a set of indexes encourages us to look for different relationships in different parts of the table.

a. We can frequently investigate complex research questions with a set of odds ratios.

C. The odds ratio in a 2 X 2 table.

1. Consider first a2 X 2 table (i.e., a table with 2 rows and 2 columns) formed by taking the first and seventh columns of the previous table. We’ll leave out the percentages because we don’t need them now.

a. This sub-table effectively compares the two extreme educational groups' turnout rates.

Education/ voted	1	7
Yes	36	142
No	30	12

2. First ask what are the odds that someone in the first educational category (i.e., column 1) voted? They are easy enough to calculate, namely:

$$\hat{\ }_{1} = \frac{36}{30} = 1.2$$

   a. This number suggests that the odds are 1.2 to 1 that someone in the first educational category votes. That's almost even, since 1 to one would mean an equal chance of voting and not voting.
      i. Notice the hat over **o**. It means we are estimating the corresponding population value.
      ii. Note also that this is not a percentage; it's the odds of voting as opposed to not voting.

3. Now, what are the odds that someone with the highest level of education (i.e., educational level = 7) voted?

$$\hat{\ }_{7} = \frac{142}{12} = 11.83333$$

   a. In this category the odds are almost 12 to 1 that a person votes. A person with the highest education, in other words, is 12 times more likely to vote than not to vote.

4. We can compare the two educational levels by taking the **ratio** of their respective odds.

$$\hat{\ } = \frac{1.2}{11.833333} = 0.101$$

   a. That is the odds of voting are among those with the lowest education are only one tenth as large as those for the most highly educated.

D. Now let's look at a compare a different part of the table, say, the first level of education with level 2. That is, let's compare the odds of voting among the two lowest categories..

1. The odds of voting the lowest category are, we just saw, 1.2 to one; and the odds for those in category 2 (refer back to the table of frequencies) are:

$$\hat{\ }_2 = \frac{80}{57} = 1.40$$

2. The odds are thus about the same, 1.4 to one for voting.
3. The **ratio** of the odds is

$$\hat{\ } = \frac{1.2}{1.40} = 0.857$$

   a. This means the ratio of the odds of first category people compared to those in the next category is about 9 tenths to one. Stated differently but equivalently, the odds of the lowest group voting are about .9 of those of the second category, which in turn suggests that the odds are not too different.

E. General formula for the odds ratio in a 2 X 2 table.

1. Let F_{11} be the frequency in the 1st cell, F_{12} be the frequency in the second cell of row 1, F_{21} be the frequency of the cell in the 2nd row 1st column; and F_{22} be the frequency in the last cell. For instance:

Column/ row	1	2
1	F_{11}	F_{12}
2	F_{21}	F_{22}

   a. The row and column numbers in this table are used merely for convenience.

2. The formula for the odds ratio is:

$$_{11} = \frac{\frac{F_{11}}{F_{21}}}{\frac{F_{21}}{F_{22}}} = \frac{F_{11}F_{22}}{F_{12}F_{21}}$$

a. This formula makes sense: it just shows the ratio of two odds.

F. Properties of the odds ratio:

1. See Agresti and Finlay, ***Statistical Methods for Social Sciences***, 3rd edition, pages 265 to 272 for a brief discussion.
2. The numerical value of the odds ratio varies between 0 and (plus) infinity.
   a. That is to say **Ω** can be zero or any positive number. (If your calculated value is negative, you did something wrong.)
3. When the odds of being in the second category of one variable are the **same** for both categories of the other variable, **Ω** = 1.0.
   a. In effect, when the odds ratio equals 1.0, there is no relationship between the variables.
   b. Example. Consider the frequencies in this hypothetical table:

25	90
75	270

   c. The odds ratio is 1.0 since the odds are of being in the second row are the same (1 to 3) in **both** categories (columns) of the other variable.
   d. Consequently departures from 1.0 in either direction represent an association between the variables.
4. The farther or greater the departure from 1.0, the greater the strength or magnitude of association between X and Y.
5. More exactly, if **Ω** < 1.0, the odds of being in the second category of Y, the row variable, are **lower** in the first category or column of X than in the second category.
6. On the other hand, if **Ω** > 1.0, then the odds of being in the first category of Y given that one is in the first category of X are greater than if one is in the second category of X.
7. Look at a couple of examples

2	1
1	1

   a. The odds ratio is 2/1 to 1/1 or 2.0. It's twice as likely that a randomly selected case from column 1 will be in row 1 than if a case is chosen from column two.

1	2
1	1

b. Now the odds ratio is 1/1 to 2/1 = .5.

8. "Two values for [**Ω**] represent the same strength of association, but in opposite directions, when one value is the inverse of the other." Agresti and Finlay, ***Statistical Methods for the Social Sciences***, 3rd edition, page 271.

a. Example: see the table above.

G. Estimation

1. As with many other sample statistics we can test hypotheses about the odds ratio, determine its expected value, find its standard error, and place confidence intervals around it.

H. Log odds

1. Because the range of the odds ratio runs from 0 to (plus) infinity with 1.0 indicating no relationship, one sometimes wants measure that is more symmetric or "balanced" so that say 0 represents statistical independence.

2. If we take the natural logarithm of **Ω** we get such a measure:

a. The **log odds ratio** is

$$_{11} = \ln\ _{11} = \ln\left(\frac{F_{11}F_{22}}{F_{12}F_{21}}\right)$$

b. Here **ln** stands for the natural logarithm.

3. The log odds ratio of voters in levels of education 1 and 7 is

$$_{11} = \ln(.101) = \quad 2.29$$

VI. NEXT TIME:

A. More on measures of association for I by J tables.

B. Linear regression.

Go to Notes page Go to Statistics page