DEPARTMENT OF POLITICAL SCIENCE
AND
INTERNATIONAL RELATIONS

Posc/Uapp 816

CONTINGENCY TABLES

I. AGENDA:
- A. Cross-classifications
  1. Two-by-two and R by C tables
  2. Statistical independence
  3. The interpretation of chi square
- B. Reading:
  1. Agresti and Finlay, ***Statistical Methods for the Social Sciences***, 3rd edition, Chapter 8.

II. WHAT EXPLAINS VOTING TURNOUT:
- A. An extremely important topic in political science and sociology is the explanation of voting turnout or more precisely non-turnout,
  1. Despite efforts to guarantee voting rights, the trend in voting participation since 1960 and indeed since 1920 has been mostly down.
- B. Example:
  1. Every two years the Center for Political Studies at the University of Michigan conducts a national survey of Americans in order to measure their political behavior and attitudes.
  2. Each time the respondents are asked if they voted in the latest national election.
     a. Many respondents claim to have voted when in fact they probably did not. Still, it's common to use these "self-reports."
  3. Here are the "marginal" totals for the 1996 election.
     a. **Frequency** refers to the number of people who gave a particular response.
     b. **Code** refers to the numeric code (i.e., number) given to each type of response. (Hence, people who said they voted are coded "1.")
     c. Note that some people's responses were not recorded. These "missing values" are coded 0.
     d. The numeric codes are arbitrary. They are usually not used in any statistical analysis and could be any other convenient characters.

In talking to people about elections, we often find that a lot of people were not able to vote because they weren't registered, they were sick, or they just didn't have time. How about you--did you vote in the elections this November?

Frequency	Code	Response
1175	1.	Yes, voted
359	5.	No, didn't vote
180	0.	No post interview (missing data)

C. The goal is to explain **why** people do and do not vote.
   1. We'll think of turnout or voting as the **dependent** variable and denote it Y.

III. ONE-WAY FREQUENCY DISTRIBUTION:
   A. It's convenient to find the percentage of each response category after the missing data have been removed.
   B. The next table, which incidentally was produced by SPSS, provides both the percentages and frequency distribution.

	Code	Frequency	Percent	Valid Percent	Cumulative Percent
	1	1175	68.6	76.6	76.6
	5	359	20.9	23.4	100.0
Sub-total		1534	89.5	100.0	
Missing	0	180	10.5		

   1. For example, 1175 is 68.6 percent of **all** responses but 76.6 of the "valid" or non-missing responses.
   C. Again, our objective is to explain or account for the variation in the valid responses.

IV. A CROSS-CLASSIFICATION OR TWO-WAY DISTRIBUTION:
   A. We might hypothesize that education accounts partly for voting: those with higher levels of education are more likely to vote than those without much formal education.
   B. A way to investigate this possibility is to form a table that shows percentages of the values of a dependent variable within categories of an **independent** variable, which we'll denote X.

C. A cross-classification or contingency table arrays the categories of categorical dependent and independent variables.

1. Example:

Independent/ dependent variable	Category 1 of X	Category 2 of X	...	Next to last category of X	Last category of X
Category 1 of Y			...		
Category 2 of Y			...		
...			...		
Last category of Y			...		

2. The "cell" entries are usually the number of cases with the joint response (e.g., the number of respondents who gave response 1 on the dependent variable and response 1 on the independent variable) or various percentages.

D. Voting example.

1. Let's see if voting is "associated" with turnout. The election study classifies respondents into 7 levels of education. This seven-category variable can be cross-classified with vote to form a 2 by 7 table, each entry of which is the number of people who have a particular combination of education and vote response.

Education/ voted	1	2	3	4	5	6	7
Yes	36	80	328	222	115	249	142
No	30	57	147	55	26	32	12

2. So we see that 36 people said they voted **and** were classified in the first education level. The **joint** frequency is, in other words, 36.
3. Similarly, 12 people in level 7 of education said they did not vote.

E. It hard to interpret these **raw** or joint frequencies so most analysts convert them to percentages.

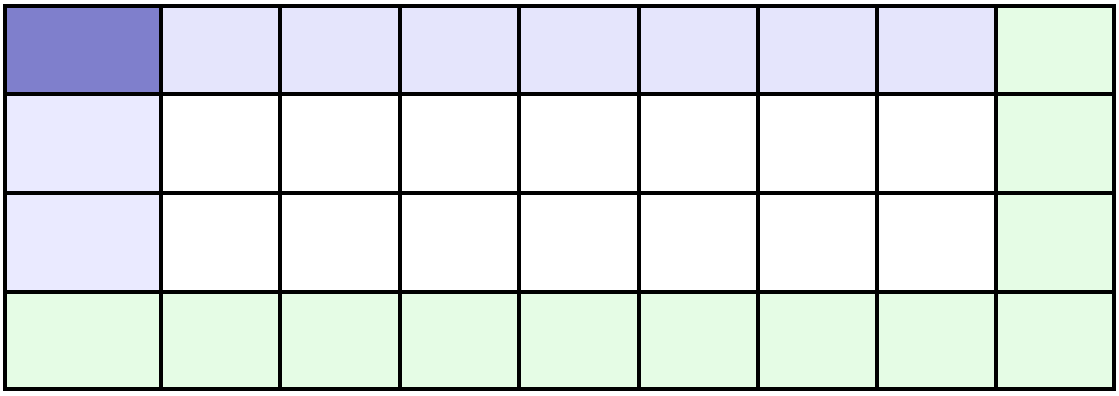
1. We usually want to know the percent of responses in each category of the independent variable so we want the so-called column percentages.
2. That is, we add each column to find a column total and then obtain the percent in row, as in the example below.
   a. See Agresti and Finlay, ***Statistical Methods***, 3rd edition, pages 249 to 253.

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

b. Numbers in bold are the column percents
   i. Example: There are a total of 66 people classified in the first level of education (look in the second column). 36 of them voted. This 54 percent of the 66.
   ii. Similarly, there are 154 people classified in level 7, of whom only 12 or 8 percent reported not voting.

F. Interpretation.
1. By examining the column percentages one can see a pattern: As the level of education increases, the percent of voters also increases.
   a. Note; we can make this statement because we are using column percentages which show relative frequencies within categories of the independent variable.
   b. Always calculate percentages within categories of the independent variable.
2. Here is a particularly clear and common method of explanation:
   a. There is variation on the dependent variable: some people reported voting; others did not. Why?
   b. We use education as a possible explanation: people with less education tend to be less likely to vote than those with higher degrees of training.
   c. In other words some categories of X are associated with categories of Y.

V. TABLE WITH MARGINAL TOTALS:
A. It is frequently useful to include in the cross-classification table **marginal totals**, the number of cases in each category of the independent and dependent variables as well as the grand total.
1. The above table contains column marginal totals but not those for the rows or the grand total.
B. Here's a "complete" table:



$$N = \sum_{i=1}^{I}$$

VI. STATISTICAL INDEPENDENCE:

A. It is apparent from the pattern of the column percentages that voting is related to or associated with education.
   1. Another way of saying this is that the two variables are not independent.

B. But of course the table describes a sample, not the population of American citizens. So one wonders whether this pattern of association holds for the population as a whole.

C. We can use sample data to test the hypothesis that two variables, X and Y, are statistically independent.
   1. Here, X is the independent variable, education, and Y is the dependent variable "Did R vote?"

D. The formal definition of statistical independence turns on "joint" and "marginal" probabilities
   1. Let π_{ij} be the probability that a randomly chosen member of the population being classified in category i of Y and j of X.
   2. π_{ij} , in other words, is the probability that Y = i **and** X = j.
   3. For example, using the previous data we see that the (estimated) probability of being a non-voter (category 2 of Y) and a member of education category 1 is $\hat{\pi}_{21} = 30/66 = .46$.

E. Statistical independence for a population is defined as:

$$\pi_{ij} = \pi_i \quad \pi_j$$

   1. In words, statistical independence means that the probability of being in category i **and** category j is just the overall probability of being in i times the overall probability of being in j.
      a. π_{i+} is the marginal probability of being in category i regardless of what category X takes.
      b. Similarly π_{+j} is the "marginal" probability of being in the jth category of X.
      c. If voting were statistically independent of education, then the probability of a person in level 1 of education not voting would just be the probability of being in level 1 times the probability of non-voting.
      d. If there is a statistical association between the variables, then the joint probability will not be the product of the two marginal probabilities.
   2. Two variables are not statistical independent if the equality does not hold; that is, if

$$\pi_{ij} \neq \pi_i \quad \pi_j$$

a. If the variables are not independent we cannot find the probability of Y = i and X = j just by multiplying marginal probabilities.

F. See Agresti and Finlay, ***Statistical Methods for Social Sciences***, 3rd edition, pages 252 to 253 for further discussion.

VII. TESTING FOR INDEPENDENCE:

A. Given sample data how can be determine if two variables in a population are statistically independent?

1. The "chi square" test is one method.

B. The chi square method tests the hypothesis that one variable is statistically independent of another.

1. If we reject the hypothesis of independence, then we can go further to determine the nature and magnitude of the association.

C. Hypotheses:

1. H_0: X and Y are statistically indpendent.
2. H_A: X and Y are not statistically independent.

D. Sampling distribution

1. When dealing with a mean or difference of means we are of course looking at sample statistics and their distributions.
2. We need to do the same in this case, namely find an appropriate sample statistic and determine its sampling distribution.
3. 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?
4. If the probability of being a voter is $\pi_{1+} = .7$ and the probability of being in the first education category is $\pi_{+j} = .4$, and if the null hypothesis of independence is true, then the probability of being a non-voting education level 1 person would be (according to the definition above)

$$\textbf{\textit{Probability}}_{\textit{voter and education} = 1} = \textbf{(.7)(.4)} = \textbf{.28}$$

5. And if we had a sample of, say, 2,000 people we would expect .28 X 2000 = 280 of them would be voters with level 1 of education.
   a. This number is called the **expected frequency** under the null hypothesis of independence.
6. Moreover, this expected value could be compared to the actually observed frequency of voting and level 1 education.
7. If H_0 holds, then the expected and observed frequencies (F) 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.

8. So it is reasonable to 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.
9. Of course, there are I times J joint frequencies to estimate so we can't look at just one.
   a. The procedure to compute the estimated expected values is to estimate probabilities:

$$\frac{N_i}{N} \textit{ is an estimator of } \pi_i$$

   b. We in effect estimate the marginal probabilities by dividing marginal frequencies by the total N.
   c. Since we want estimated **frequencies** we have to multiply these probabilities by the sample size. That is, for instance:

$$\hat{F}_{ij} = \left(\frac{N_i}{N} \frac{N_j}{N} \right) N = \frac{N_i N_j}{N}$$

   d. All of the I times J expected frequencies are estimated in this manner.
   e. Example: here are the estimated expected frequencies for some of the combinations of X and Y in the voting table:

$$\hat{F}_{11} = \frac{(1175)(66)}{1534} = 50.55$$

$$\hat{F}_{12} = \frac{(1175)(137)}{1544} = 104.93$$

$$\hat{F}_{27} = \frac{(359)(154)}{1544} = 336.04$$

   f. Notice that the expected frequencies do not (and normally will not) be whole numbers.
   g. Here are the observed and estimated expected frequencies (in

parentheses) for the voting data.

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.

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

10. We aggregate the differences between each observed and expected frequency into a chi square statistic, denoted χ^2

11. The chi square, χ^2, has a sampling distribution called (logically enough) the chi square, which has degrees of freedom:

$$df = (I \quad 1)\ (J \quad 1)$$

a. Where I is the number of categories of Y and J is the number of categories of X.

E. Sample statistic:

1. Procedure in short:

a. We have to determine **IJ** expected values--expected under the hypotheses of independence.

b. We then compare each of them with their respective observed frequencies.

c. We finally sum the discrepancies to create the observed **chi square** statistic.

2. The sample or observed chi square is given by:

$$\chi^2_{obs} = \sum \frac{(F_{ij} \quad \hat{F}_{ij})^2}{\hat{F}_{ij}}$$

a. Where the sum is over all of the IJ categories.

3. These numbers are calculated from a sample that has been drawn

randomly from a population.

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. Example: voting data 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. Compare the observed with the critical chi square. Since it is much larger, with 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.

VIII. REMARKS AND INTERPRETATION:

A. 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.

B. Chi square's numerical magnitude is affected by the sample size: as N 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 trival.

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

C. 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.

IX. NEXT TIME:
A. Further analysis of cross-classification tables.

Go to Notes page
Go to Statistics page