Categorical Data Analysis

Related topics/headings: Categorical data analysis; or, Nonparametric statistics; or, chi-square tests for the analysis of categorical data.

OVERVIEW

For our hypothesis testing so far, we have been using parametric statistical methods. Parametric methods (1) assume some knowledge about the characteristics of the parent population (e.g. normality) (2) require measurement equivalent to at least an interval scale (calculating a mean or a variance makes no sense otherwise).

Frequently, however, there are research problems in which one wants to make direct inferences about two or more distributions, either by asking if a population distribution has some particular specifiable form, or by asking if two or more population distributions are identical. These questions occur most often when variables are qualitative in nature, making it impossible to carry out the usual inferences in terms of means or variances. For such problems, we use nonparametric methods. Nonparametric methods (1) do not depend on any assumptions about the parameters of the parent population (2) generally assume data are only measured at the nominal or ordinal level.

There are two common types of hypothesis-testing problems that are addressed with nonparametric methods:

(1) How well does a sample distribution correspond with a hypothetical population distribution? As you might guess, the best evidence one has about a population distribution is the sample distribution. The greater the discrepancy between the sample and theoretical distributions, the more we question the "goodness" of the theory.

EX: Suppose we wanted to see whether the distribution of educational achievement had changed over the last 25 years. We might take as our null hypothesis that the distribution of educational achievement had not changed, and see how well our modern-day sample supported that theory.

(2) We often wish to find evidence for association between two qualitative variables - hence we analyze cross-classifications of two discrete distributions.

EX: What is the relationship between sex and party vote - are women more likely than men to support Democratic party candidates?

CASE I. COMPARING SAMPLE AND POPULATION DISTRIBUTIONS.

Suppose that a study of educational achievement of American men were being carried on. The population studied is the set of all American males who are 25 years old at the time of the

study. Each subject observed can be put into 1 and only 1 of the following categories, based on his maximum formal educational achievement:

1 = college grad
2 = some college
3 = high school grad
4 = some high school
5 = finished 8th grade
6 = did not finish 8th grade

Note that these categories are mutually exclusive and exhaustive.

The researcher happens to know that 10 years ago the distribution of educational achievement on this scale for 25 year old men was:

1 - 18%
2 - 17%
3 - 32%
4 - 13%
5 - 17%
6 - 3%

A random sample of 200 subjects is drawn from the current population of 25 year old males, and the following frequency distribution obtained:

1 - 35
2 - 40
3 - 83
4 - 16
5 - 26
6 - 0

The researcher would like to ask if the present population distribution on this scale is exactly like that of 10 years ago. That is, he would like to test

H_0: There has been no change across time. The distribution of education in the present population is the same as the distribution of education in the population 10 years ago

H_A: There has been change across time. The present population distribution differs from the population distribution of 10 years ago.

PROCEDURE: Assume that there has been "no change" over the last 10 years. In a sample of 200, how many men would be expected to fall into each category?

For each category, the expected frequency is

$$N * p_j = E_j = \textit{expected frequency for jth category,}$$

where N = sample size = 200 (for this sample), and p_j = the relative frequency for category j dictated by the null hypothesis. For example, since 18% of all 25 year old males 10 years ago were college graduates, we would expect 18% of the current sample, or 36 males, to be college graduates today if there has been no change. We can therefore construct the following table:

Category	Observed freq (O_j)	Expected freq (E_j)
1	35	36 = 200 * .18
2	40	34 = 200 * .17
3	83	64 = 200 * .32
4	16	26 = 200 * .13
5	26	34 = 200 * .17
6	0	6 = 200 * .03

Question: The observed and expected frequencies obviously differ from each other - but we expect some discrepancies, just because of sampling variability. How do we decide whether the discrepancies are too large to attribute simply to chance?

Answer: We need a test statistic that measures the "goodness of fit" between the observed frequencies and the frequencies expected under the null hypothesis. The Pearson chi-square statistic is one appropriate choice. (The Likelihood Ratio Chi-Square, sometimes referred to as L^2, is another commonly used alternative, but we won't discuss it this semester.) The formula for this statistic is

$$\chi^2_{c-1} = \Sigma\,(O_j - E_j)^2/E_j$$

Calculating χ^2_{c-1} for the above, we get

Category	O_j	E_j	$(O_j - E_j)$	$(O_j - E_j)^2/E_j$
1	35	36	-1	1/36 = 0.0278
2	40	34	6	36/34 = 1.0588
3	83	64	19	361/64 = 5.6406
4	16	26	-10	100/26 = 3.8462
5	26	34	-8	64/34 = 1.8824
6	0	6	-6	36/ 6 = 6.0000

Summing the last column, we get $\chi^2_{c-1} = 18.46$.

Q: Now that we have computed χ^2_{c-1}, what do we do with it???

A: Note that, the closer O_j is to E_j, the smaller χ^2_{c-1} is. Hence, small values of χ^2_{c-1} imply good fits (i.e. the distribution specified in the null hypothesis is similar to the distribution found in the sample), big values imply poor fits (implying that the hypothesized distribution and the sample distribution are probably not one and the same).

To determine what "small" and "big" are, note that, when each expected frequency is as little as 5 or more (and possibly as little as 1 or more),

χ^2_{c-1} ~ Chi-square(c-1), where
c = the number of categories,
c - 1 = v = degrees of freedom

Q: Why does d.f. = c - 1?
A: When working with tabled data (a frequency distribution can be thought of as a 1 dimensional table) the general formula for degrees of freedom is

d.f. = number of cells - # of pieces of sample information required for computing expected cell frequencies.

In the present example, there are c = 6 cells (or categories). In order to come up with expected frequencies for those 6 cells, we only had to have 1 piece of sample information, namely, N, the sample size. (The values for p_j were all contained in the null hypothesis.)

Q: What is a chi-square distribution, and how does one work with it?

Appendix E, Table IV (Hayes pp. 933-934) gives critical values for the Chi-square distribution. The second page of the table has the values you will be most interested in, e.g. Q = .05, Q =.01.

A: The chi-square distribution is easy to work with, but there are some important differences between it and the Normal distribution or the T distribution. Note that

- ✓ The chi-square distribution is NOT symmetric
- ✓ All chi-square values are positive
- ✓ As with the T distribution, the shape of the chi-square distribution depends on the degrees of freedom.
- ✓ Hypothesis tests involving chi-square are usually one-tailed. We are only interested in whether the observed sample distribution significantly differs from the hypothesized distribution. We therefore look at values that occur in the upper tail of the chi-square distribution. That is, low values of chi-square indicate that the sample distribution and the hypothetical distribution are similar to each other, high values indicate that the distributions are dissimilar.
- ✓ A random variable has a chi-square distribution with N degrees of freedom if it has the same distribution as the sum of the squares of N independent variables, each

normally distributed, and each having expectation 0 and variance 1. For example, if $Z \sim N(0,1)$, then $Z^2 \sim$ Chi-square(1). If Z_1 and Z_2 are both $\sim N(0,1)$, then $Z_1^2 + Z_2^2 \sim$ Chi-square(2).

EXAMPLES:

Q. If v = d.f. = 1, what is $P(\chi^2_1 \geq 3.84)$?

A. Note that, for v = 1 and $\chi^2_1 = 3.84$, Q = .05. i.e. $F(3.84) = P(\chi^2_1 \leq 3.84) = .95$, hence $P(\chi^2_1 \geq 3.84) = 1 - .95 = .05$. (Incidentally, note that $1.96^2 = 3.84$. If $Z \sim N(0,1)$, then $P(-1.96 \leq Z \leq 1.96) = .95 = P(Z^2 \leq 3.84)$. Recall that $Z^2 \sim$ Chi-square(1).)

Q. If v = 5, what is the critical value for χ^2_5 such that $P(\chi^2_5 \geq \chi^2_5) = .01$?

A. Note that, for v = 5 and Q = .01, the critical value is 15.0863. Ergo, $P(\chi^2_5 \geq 15.1) = 1 - .99 = .01$

Returning to our present problem - we had six categories of education. Hence, we want to know $P(\chi^2_5 \geq 18.46)$. That is, how likely is it, if the null hypothesis is true, that we could get a Pearson chi-square value of this big or bigger in a sample? Looking at Table IV, v = 5, we see that this value is around .003 (look at Q = .005 and Q = .001). That is, if the null hypothesis is true, we would expect to observe a sample distribution that differed this much from the hypothesized distribution fewer than 3 times out of a thousand. Hence, we should probably reject the null hypothesis.

To put this problem in our usual hypothesis testing format,

Step 1: H_0: Distribution now is the same as 10 years ago
H_A: Distribution now and 10 years ago differ

Step 2: An appropriate test statistic is

$$\chi^2_{c-1} = \Sigma (O_j - E_j)^2/E_j, \qquad \text{where } E_j = Np_j$$

Step 3: Acceptance region: Accept H_0 if

$$P(\chi^2_{c-1} \leq \chi^2_{c-1}) = 1 - \alpha.$$

In the present example, let us use $\alpha = .01$. Since v = 5, accept H_0 if

$\chi^2_{c-1} \leq 15.1$ (see v =5, Q = .01)

Step 4. The computed test statistic = 18.46.

Step 5. Reject H_0. The value of the computed test statistic lies outside of the acceptance region.

SPSS Solution. The NPAR TESTS Command can be used to estimate this model in SPSS. If using the pull-down menus in SPSS, choose ANALYZE/ NONPARAMETRIC TESTS/ CHI-SQUARE.

```
* Case I:  Comparing sample and population distributions
* Educ distribution same as 10 years ago.
data list free / educ wgt.
begin data.
1 35
2 40
3 83
4 16
5 26
end data.
weight by wgt.

NPAR TEST
  /CHISQUARE=educ  (1,6)
  /EXPECTED=36 34 64 26 34 6
  /STATISTICS  DESCRIPTIVES
  /MISSING ANALYSIS.
```

NPar Tests

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
EDUC	200	2.7900	1.20963	1.00	5.00

Chi-Square Test

Frequencies

	EDUC			
	Category	Observed N	Expected N	Residual
1	1.00	35	36.0	-1.0
2	2.00	40	34.0	6.0
3	3.00	83	64.0	19.0
4	4.00	16	26.0	-10.0
5	5.00	26	34.0	-8.0
6		0	6.0	-6.0
Total		200		

Test Statistics

	EDUC
Chi-Square[a]	18.456
df	5
Asymp. Sig.	.002

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 6.0.

OTHER HYPOTHETICAL DISTRIBUTIONS: In the above example, the hypothetical distribution we used was the known population distribution of 10 years ago. Another possible hypothetical distribution that is sometimes used is specified by the equi-probability model. The equi-probability model claims that the expected number of cases is the same for each category; that is, we test

H_0: $E_1 = E_2 = ... = E_c$
H_A: The frequencies are not all equal.

The expected frequency for each cell is (Sample size/Number of categories). Such a model might be plausible if we were interested in, say, whether birth rates differed across months. If we believed the equi-probability model might apply to educational achievement, we would hypothesize that 33.33 people would fall into each of our 6 categories.

Calculating χ^2_{c-1} for the equi-probability model, we get

Category	O_j	E_j	$(O_j - E_j)^2/E_j$
1	35	33.33	0.0837
2	40	33.33	1.3348
3	83	33.33	74.0207
4	16	33.33	9.0108
5	26	33.33	1.6120
6	0	33.33	33.3333

Summing the last column, we get $\chi^2_{c-1} = 119.39$. Obviously, the equi-probability model does not provide a very good description of educational achievement in the United States.

SPSS Solution. Again use the NPAR TESTS Command.

```
* Equi-probability model.  Same observed data as before.
NPAR TEST
  /CHISQUARE=educ  (1,6)
  /EXPECTED=EQUAL
  /STATISTICS  DESCRIPTIVES
  /MISSING ANALYSIS.
```

NPar Tests

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
EDUC	200	2.7900	1.20963	1.00	5.00

Chi-Square Test

Frequencies

	EDUC			
	Category	Observed N	Expected N	Residual
1	1.00	35	33.3	1.7
2	2.00	40	33.3	6.7
3	3.00	83	33.3	49.7
4	4.00	16	33.3	-17.3
5	5.00	26	33.3	-7.3
6		0	33.3	-33.3
Total		200		

Test Statistics

	EDUC
Chi-Square[a]	119.380
df	5
Asymp. Sig.	.000

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 33.3.

CASE II. TESTS OF ASSOCIATION

A researcher wants to know whether men and women in a particular community differ in their political party preferences. She collects data from a random sample of 200 registered voters, and observes the following:

	Dem	**Rep**
Male	55	65
Female	50	30

Do men and women significantly differ in their political preferences? Use $\alpha = .05$.

PROCEDURE. The researcher wants to test what we call the model of independence (the reason for that name will become apparent in a moment). That is, she wants to test

H_0: Men and women do not differ in their political preferences
H_A: Men and women do differ in their political preferences.

Suppose H_0 (the model of independence) were true. What joint distribution of sex and party preference would we expect to observe?

Let A = Sex, A_1 = male, A_2 = female, B = political party preference, B_1 = Democrat, B_2 = Republican. Note that $P(A_1) = .6$ (since there are 120 males in a sample of 200), $P(A_2) = .4$, $P(B_1) = .525$ (105 Democrats out of a sample of 200) and $P(B_2) = .475$.

If men and women do not differ, then the variables A (sex) and B (party vote) should be independent of each other. That is, $P(A_i \cap B_j) = P(A_i)P(B_j)$. Hence, for a sample of size N,

$$E_{ij} = P(A_i) * P(B_j) * N$$

For example, if the null hypothesis were true, we would expect 31.5% of the sample (i.e. 63 of the 200 sample members) to consist of male democrats, since 60% of the sample is male and 52.5% of the sample is Democratic. The complete set of observed and expected frequencies is

Sex/Party	Observed	Expected
Male Dem	55	P(Male)*P(Dem) =.6 * .525 * 200 = 63
Male Rep	65	P(Male)*P(Rep) = .6 * .475 * 200 = 57
Female Dem	50	P(Fem)*P(Dem) = .4 * .525 * 200 = 42
Female Rep	30	(P(Fem)*P(Rep) = .4 * .475 * 200 = 38

Q: The observed and expected frequencies obviously differ - but we expect some differences, just because of sampling variability. How do we decide if the differences are too large to attribute simply to chance?

A: Once again, the Pearson chi-square is an appropriate test statistic. The appropriate formula is

$$\chi^2_v = \Sigma\Sigma\, (O_{ij} - E_{ij})^2/E_{ij},$$

where r is the number of rows (i.e. the number of different possible values for sex), c is the number of columns (i.e. the number of different possible values for party preference), and v = degrees of freedom = rc - 1 - (r-1) - (c-1) = (r-1)(c-1).

Q: Why does d.f. = rc - 1 - (r-1) - (c-1) = (r-1)(c-1)?
A: Recall our general formula from above:

d.f. = number of cells - # of pieces of sample information required for computing expected cell frequencies.

In this example, the number of cells is rc = 2*2 = 4. The pieces of sample information required for computing the expected cell frequencies are N, the sample size; $P(A_1)$ = P(Male) = .6; and $P(B_1)$ = P(Democrat) = .525. Note that, once we knew $P(A_1)$ and $P(B_1)$, we immediately knew $P(A_2)$ and $P(B_2)$, since probabilities sum to 1; we don't have to use additional degrees of freedom to estimate them. Hence, there are 4 cells, we had to know 3 pieces of sample information to get expected frequencies for those 4 cells, hence there is 1 d.f. NOTE: In a 2-dimensional table, it happens to work out that, for the model of independence, d.f. = (r-1)(c-1). It is NOT the case that in a 3-dimensional table d.f. = (r-1)(c-1)(l-1), where l is the number of categories for the 3rd variable; rather, d.f. = rcl - 1 - (r-1) - (c-1) - (l-1).

Returning to the problem - we can compute

Sex/Party	Observed	Expected	$(O_{ij} - E_{ij})^2/E_{ij}$
Male Dem	55	63	64/63 = 1.0159
Male Rep	65	57	64/57 = 0.9552
Female Dem	50	42	64/42 = 1.5238
Female Rep	30	38	64/38 = 1.6842

Note that v = (r - 1)(c - 1) = 1. Adding up the numbers on the right-hand column, we get χ^2_1 = 5.347. Looking at table IV, we see that we would get a test statistic this large only about 2% of the time if H_0 were true, hence we reject H_0.

To put things more formally then,

Step 1.

H_0: Men and women do not differ in their political preferences
H_A: Men and women do differ in their political preferences.

or, equivalently,

H_0: $P(A_i \cap B_j) = P(A_i)P(B_j)$ (Model of independence)
H_A: $P(A_i \cap B_j) <> P(A_i)P(B_j)$ for some i, j

Step 2. An appropriate test statistic is

$\chi^2_v = \Sigma\Sigma\ (O_{ij} - E_{ij})^2/E_{ij}$, v = rc-1-(r-1)-(c-1) = (r-1)(c-1)

Step 3. For α = .05 and v = 1, accept H_0 if $\chi^2_v \leq 3.84$

Step 4. The computed value of the test statistic is 5.347

Step 5. Reject H_0, the computed test statistic is too high.

Yates Correction for Continuity. Sometimes in a 1 X 2 or 2 X 2 table (but not for other size tables), Yates Correction for Continuity is applied. This involves subtracting 0.5 from positive differences between observed and expected frequencies, and adding .5 to negative differences before squaring. This will reduce the magnitude of the test statistic. To apply the correction in the above example,

Sex/Party	Observed	Expected (with correction)	$(O_{ij} - E_{ij})^2/E_{ij}$
Male Dem	55	62.5	$-7.5^2/62.5 = .9$
Male Rep	65	57.5	$7.5^2/57.5 = .9783$
Female Dem	50	42.5	$7.5^2/42.5 = 1.3235$
Female Rep	30	37.5	$-7.5^2/37.5 = 1.5$

After applying the correction, the computed value of the test statistic is 4.70.

Fisher's Exact Test. The Pearson Chi-Square test and the Yates Correction for Continuity are actually just approximations of the exact probability; and particularly when some expected frequencies are small (5 or less) they may be somewhat inaccurate. As Stata 8's Reference Manual S-Z, p. 219 notes, "Fisher's exact test yields the probability of observing a table that gives at least as much evidence of association as the one actually observed under the assumption of no association." In other words, if the model of independence holds, how likely would you be to see a table that deviated this much or more from the expected frequencies?

You are most likely to see Fisher's exact test used with 2 X 2 tables where one or more expected frequencies is less than 5, but it can be computed in other situations. It can be hard to do by hand though and even computers can have problems when the sample size or number of cells is large. SPSS can optionally report Fisher's exact test for 2 X 2 tables but apparently won't do it for larger tables (unless perhaps you buy some of its additional modules). Stata can, by request, compute Fisher's exact test for any size two dimensional table, but it may take a while to do so.

You don't get a test statistic with Fisher's exact test; instead, you just get the probabilities. For the current example, the 2-sided probability of getting a table where the observed frequencies differed this much or more from the expected frequencies if the model of indepdence is true is .022; the one-sided probability is .015.

SPSS Solution. SPSS Has a couple of ways of doing this. The easiest is probably the crosstabs command. On the SPSS pulldown menus, look for ANALYZE/ DESCRIPTIVE STATISTICS/ CROSSTABS.

```
* Case II:  Tests of association.
Data list free / Sex Party Wgt.
Begin data.
1 1 55
2 1 50
1 2 65
2 2 30
End data.
Weight by Wgt.

CROSSTABS
  /TABLES=sex  BY party
  /FORMAT= AVALUE NOINDEX BOX LABELS TABLES
  /STATISTIC=CHISQ
  /CELLS= COUNT EXPECTED .
```

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
SEX * PARTY	200	100.0%	0	.0%	200	100.0%

SEX * PARTY Crosstabulation

			PARTY		
			1.00	2.00	Total
SEX	1.00	Count	55	65	120
		Expected Count	63.0	57.0	120.0
	2.00	Count	50	30	80
		Expected Count	42.0	38.0	80.0
Total		Count	105	95	200
		Expected Count	105.0	95.0	200.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	5.347[b]	1	.021		
Continuity Correction[a]	4.699	1	.030		
Likelihood Ratio	5.388	1	.020		
Fisher's Exact Test				.022	.015
Linear-by-Linear Association	5.320	1	.021		
N of Valid Cases	200				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 38.00.

CHI-SQUARE TESTS OF ASSOCIATION FOR 2 X 2 TABLES (NONPARAMETRIC TESTS, CASE II) VS. TWO SAMPLE TESTS, CASE V, TEST OF P1-P2 = 0.

Consider again the following sample data.

	Dem	**Rep**
Male	55	65
Female	50	30

Note that, instead of viewing this as one sample of 200 men and women, we could view it as two samples, a sample of 120 men and another sample of 80 women. Further, since there are only two categories for political party, testing whether men and women have the same distribution of party preferences is equivalent to testing whether the same proportion of men and women support the Democratic party. Hence, we could also treat this as a two sample problem, case V, test of $p_1 = p_2$. The computed test statistic is

$$z = \frac{\frac{X_1}{N_1} - \frac{X_2}{N_2}}{\sqrt{\left(\frac{N_1 + N_2}{N_1 N_2}\right)\left(\frac{X_1 + X_2}{N_1 + N_2}\right)\left(1 - \frac{X_1 + X_2}{N_1 + N_2}\right)}} = \frac{\frac{55}{120} - \frac{50}{80}}{\sqrt{\left(\frac{120 + 80}{120 * 80}\right)\left(\frac{55 + 50}{120 + 80}\right)\left(1 - \frac{55 + 50}{120 + 80}\right)}} = 2.31$$

Hence, using $\alpha = .05$, we again reject H_0.

NOTE: Recall that if $Z \sim N(0,1)$, $Z^2 \sim$ Chi-square(1). If we square 2.31229, we get 5.347 - which was the value we got for χ^2 with 1 d.f. when we did the chi-square test for association. For a 2 X 2 table, a chi-square test for the model of independence and a 2 sample test of $p_1 - p_2 = 0$ (with a 2-tailed alternative) will yield the same results. Once you get bigger tables, of course, the tests are no longer equivalent (since you either have more than 2 samples, or you have more than just p_1 and p_2).

CASE III: CHI-SQUARE TESTS OF ASSOCIATION FOR N-DIMENSIONAL TABLES

A researcher collects the following data:

Gender/Party	*Republican*		*Democrat*	
	W	*NW*	*W*	*NW*
Male	20	5	20	15
Female	18	2	15	5

Test the hypothesis that sex, race, and party affiliation are independent of each other. Use $\alpha = .10$.

Solution. Let A = Sex, A_1 = male, A_2 = female, B = Race, B_1 = white, B_2 = nonwhite, C = Party affiliation, C_1 = Republican, C_2 = Democrat. Note that N = 100, $P(A_1) = .60$, $P(A_2) = 1 - .60 = .40$, $P(B_1) = .73$, $P(B_2) = 1 - .73 = .27$, $P(C_1) = .45$, $P(C_2) = 1 - .45 = .55$.

Step 1.
H_0: $P(A_i \cap B_j \cap C_k) = P(A_i) * P(B_j) * P(C_k)$ (Independence model)
H_A: $P(A_i \cap B_j \cap C_k) <> P(A_i) * P(B_j) * P(C_k)$ for some i, j, k

Step 2. The appropriate test statistic is

$$\chi^2_v = \Sigma\Sigma\Sigma (O_{ijk} - E_{ijk})^2/E_{ijk}$$

Note that $E_{ijk} = P(A_i) * P(B_j) * P(C_k) * N$.

Since A, B, and C each have two categories, the sample information required for computing the expected frequencies is $P(A_1)$, $P(B_1)$, $P(C_1)$, and N. (Note that once we know $P(A_1)$, $P(B_1)$, and $P(C_1)$, we automatically know $P(A_2)$, $P(B_2)$, and $P(C_2)$). Hence, there are 8 cells in the table, we need 4 pieces of sample information to compute the expected frequencies for those 8 cells, hence d.f. = 8 - 4. More generally, for a three-dimensional table, the model of independence has d.f. = v = rcl - 1 - (r - 1) - (c - 1) - (l - 1).

Step 3. Accept H_0 if $\chi^2_4 \leq 7.78$ (see v = 4 and Q = .10)

Step 4. To compute the Pearson Chi-square:

Sex/Race/Party	O_{ijk}	E_{ijk}	$(O-E)^2/E$
M / W / R	20	19.71 = .6*.73*.45*100	0.0043
F / W / R	18	13.14 = .4*.73*.45*100	1.7975
M / NW/ R	5	7.29 = .6*.27*.45*100	0.7194
F / NW/ R	2	4.86 = .4*.27*.45*100	1.6830
M / W / D	20	24.09 = .6*.73*.55*100	0.6944
F / W / D	15	16.06 = .4*.73*.55*100	0.0700
M / NW/ D	15	8.91 = .6*.27*.55*100	4.1625
F / NW/ D	5	5.94 = .4*.27*.55*100	0.1488

Summing the last column, we get a computed test statistic value of 9.28.

Step 5. Reject H_0, the computed test statistic value lies outside the acceptance region. (Note that we would not reject if we used $\alpha = .05$.)

SPSS Solution. You can still do Crosstabs but SPSS doesn't report the test statistics in a particularly useful fashion. The SPSS GENLOG command provides one way of dealing with more complicated tables, and lets you also estimate more sophisticated models. On the SPSS menus, use ANALYZE/ LOGLINEAR/ GENERAL. I'm only showing the most important parts of the printout below.

```
* N-Dimensional tables.

Data list free / sex party race wgt.
begin data.
1 1 1 20
1 1 2  5
1 2 1 20
1 2 2 15
2 1 1 18
2 1 2  2
2 2 1 15
2 2 2  5
end data.
weight by wgt.

* Model of independence.
GENLOG
  party race sex
  /MODEL=POISSON
  /PRINT FREQ
  /PLOT NONE
  /CRITERIA =CIN(95) ITERATE(20) CONVERGE(.001) DELTA(.5)
  /DESIGN party race sex  .
```

General Loglinear

```
Table Information

                 Observed            Expected
Factor    Value     Count       %       Count       %

PARTY   1.00
 RACE    1.00
  SEX     1.00      20.00 ( 20.00)      19.71 ( 19.71)
  SEX     2.00      18.00 ( 18.00)      13.14 ( 13.14)
 RACE    2.00
  SEX     1.00       5.00 (  5.00)       7.29 (  7.29)
  SEX     2.00       2.00 (  2.00)       4.86 (  4.86)

PARTY   2.00
 RACE    1.00
  SEX     1.00      20.00 ( 20.00)      24.09 ( 24.09)
  SEX     2.00      15.00 ( 15.00)      16.06 ( 16.06)
 RACE    2.00
  SEX     1.00      15.00 ( 15.00)       8.91 (  8.91)
  SEX     2.00       5.00 (  5.00)       5.94 (  5.94)

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Goodness-of-fit Statistics

                    Chi-Square       DF       Sig.

Likelihood Ratio        9.0042        4      .0610
         Pearson        9.2798        4      .0545
```

CONDITIONAL INDEPENDENCE IN N-DIMENSIONAL TABLES

Using the same data as in the last problem, test whether party vote is independent of sex and race, WITHOUT assuming that sex and race are independent of each other. Use $\alpha = .05$.

Solution. We are being asked to test the model of conditional independence. This model says that party vote is not affected by either race or sex, although race and sex may be associated with each other. Such a model makes sense if we are primarily interested in the determinants of party vote, and do not care whether other variables happen to be associated with each other.

Note that $P(A_1 \cap B_1) = .40$, $P(A_2 \cap B_1) = .33$, $P(A_1 \cap B_2) = .20$, $P(A_2 \cap B_2) = 1 - .40 - .33 - .20 = .07$, $P(C_1) = .45$, $P(C_2) = 1 - .45 = .55$, and $N = 100$.

Step 1.
H_0: $P(A_i \cap B_j \cap C_k) = P(A_i \cap B_j) * P(C_k)$
H_A: $P(A_i \cap B_j \cap C_k) <> P(A_i \cap B_j) * P(C_k)$ for some i, j, k

Step 2. The Pearson chi-square is again an appropriate test statistic. However, the expected values for the model of conditional independence are

$$E_{ijk} = P(A_i \cap B_j) * P(C_k) * N.$$

To compute the expected values, we need 5 pieces of sample information (N, $P(C_1)$, $P(A_1 \cap B_1)$, $P(A_2 \cap B_1)$, and $P(A_1 \cap B_2)$), hence d.f. =

$v = rcl - 1 - (rc - 1) - (l - 1) = 8 - 1 - (4 - 1) - (2 - 1) = 3.$

Step 3. For $\alpha = .05$ and $v = 3$, accept H_0 if $\chi^2_3 \leq 7.81$.
Step 4. To compute the Pearson Chi-square:

Sex-Race/Party	O_{ijk}	E_{ijk}	$(O-E)^2/E$
M-W / R	20	18.00 = .40*.45*100	0.2222
F-W / R	18	14.85 = .33*.45*100	0.6682
M-NW/ R	5	9.00 = .20*.45*100	1.7778
F-NW/ R	2	3.15 = .07*.45*100	0.4198
M-W / D	20	22.00 = .40*.55*100	0.1818
F-W / D	15	18.15 = .33*.55*100	0.5467
M-NW/ D	15	11.00 = .20*.55*100	1.4545
F-NW/ D	5	3.85 = .07*.55*100	0.3435

Summing the last column, the computed test statistic = 5.61.

Step 5. Accept H_0; the computed test statistic falls within the acceptance region.

SPSS Solution. You can again use GENLOG.

```
* Model of conditional independence.  Same data as above.

GENLOG
  party race sex
  /MODEL=POISSON
  /PRINT FREQ
  /PLOT NONE
  /CRITERIA =CIN(95) ITERATE(20) CONVERGE(.001) DELTA(.5)
  /DESIGN party race sex race*sex  .
```

General Loglinear

```
Table Information

                 Observed            Expected
Factor    Value     Count       %       Count       %

PARTY   1.00
 RACE    1.00
  SEX     1.00      20.00 ( 20.00)      18.00 ( 18.00)
  SEX     2.00      18.00 ( 18.00)      14.85 ( 14.85)
 RACE    2.00
  SEX     1.00       5.00 (  5.00)       9.00 (  9.00)
  SEX     2.00       2.00 (  2.00)       3.15 (  3.15)

PARTY   2.00
 RACE    1.00
  SEX     1.00      20.00 ( 20.00)      22.00 ( 22.00)
  SEX     2.00      15.00 ( 15.00)      18.15 ( 18.15)
 RACE    2.00
  SEX     1.00      15.00 ( 15.00)      11.00 ( 11.00)
  SEX     2.00       5.00 (  5.00)       3.85 (  3.85)

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Goodness-of-fit Statistics

                    Chi-Square       DF       Sig.

Likelihood Ratio        5.8322        3      .1201
         Pearson        5.6146        3      .1319
```