Using Stata for Categorical Data Analysis

NOTE: These problems make extensive use of Nick Cox's `tab_chi`, which is actually a collection of routines, and Adrian Mander's `ipf` command. From within Stata, use the commands `ssc install tab_chi` and `ssc install ipf` to get the most current versions of these programs. Thanks to Nick Cox, Richard Campbell and Philip Ender for helping me to identify the Stata routines needed for this handout.

This handout shows how to work the problems in Stata; see the related handouts for the underlying statistical theory and for SPSS solutions. Most of the commands have additional optional parameters that may be useful; type `help commandname` for more information.

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 <u>mutually exclusive and exhaustive</u>.

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.

Stata Solution. Surprisingly, Stata does not seem to have any built-in routines for Case I, but luckily Nick Cox's `chitesti` routine (part of his `tab_chi` package) is available. Like other Stata "immediate" commands, `chitesti` obtains data not from the data stored in memory but from numbers typed as arguments. The format (without optional parameters) is

```
chitesti #obs1 #obs2 [...] [ \ #exp1 #exp2 [...] ]
```

In this case,

```
. chitesti 35 40 83 16 26 0 \ 36 34 64 26 34 6, sep(6)

observed frequencies from keyboard; expected frequencies from keyboard

         Pearson chi2(5) =  18.4557   Pr =  0.002
likelihood-ratio chi2(5) =  24.6965   Pr =  0.000

  +-------------------------------------------+
  | observed   expected   obs - exp   Pearson |
  |-------------------------------------------|
  |       35     36.000      -1.000    -0.167 |
  |       40     34.000       6.000     1.029 |
  |       83     64.000      19.000     2.375 |
  |       16     26.000     -10.000    -1.961 |
  |       26     34.000      -8.000    -1.372 |
  |        0      6.000      -6.000    -2.449 |
  +-------------------------------------------+
```

The significant chi-square statistics imply that the null should be rejected, i.e. the distribution today is not the same as 10 years ago.

Alternatively, we could have the data in a file and then use the `chitest` command, e.g. the data would be

```
. list  observed expected, sep(6)

     +---------------------+
     | observed   expected |
     |---------------------|
  1. |       35         36 |
  2. |       40         34 |
  3. |       83         64 |
  4. |       16         26 |
  5. |       26         34 |
  6. |        0          6 |
     +---------------------+
```

We then give the command

```
. chitest  observed expected, sep(6)

observed frequencies from observed; expected frequencies from expected

         Pearson chi2(5) =  18.4557   Pr =  0.002
likelihood-ratio chi2(5) =  24.6965   Pr =  0.000

  +--------------------------------------------+
  | observed   expected   obs - exp   Pearson |
  |--------------------------------------------|
  |       35     36.000      -1.000    -0.167 |
  |       40     34.000       6.000     1.029 |
  |       83     64.000      19.000     2.375 |
  |       16     26.000     -10.000    -1.961 |
  |       26     34.000      -8.000    -1.372 |
  |        0      6.000      -6.000    -2.449 |
  +--------------------------------------------+
```

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 for some bizarre reason 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.

With the `chitesti` and `chitest` commands, if you DON'T specify expected frequencies, the equi-probability model is assumed. Hence,

```
. chitesti 35 40 83 16 26 0, sep(6)

observed frequencies from keyboard; expected frequencies equal

         Pearson chi2(5) = 119.3800   Pr =  0.000
likelihood-ratio chi2(5) = 133.0330   Pr =  0.000

  +--------------------------------------------+
  | observed   expected   obs - exp   Pearson |
  |--------------------------------------------|
  |       35     33.333       1.667     0.289 |
  |       40     33.333       6.667     1.155 |
  |       83     33.333      49.667     8.603 |
  |       16     33.333     -17.333    -3.002 |
  |       26     33.333      -7.333    -1.270 |
  |        0     33.333     -33.333    -5.774 |
  +--------------------------------------------+
```

Or, using a data file,

```
. chitest observed, sep(6)

observed frequencies from observed; expected frequencies equal

         Pearson chi2(5) = 119.3800   Pr =  0.000
likelihood-ratio chi2(5) = 133.0330   Pr =  0.000

  +-------------------------------------------+
  | observed   expected   obs - exp   Pearson |
  |-------------------------------------------|
  |       35     33.333       1.667     0.289 |
  |       40     33.333       6.667     1.155 |
  |       83     33.333      49.667     8.603 |
  |       16     33.333     -17.333    -3.002 |
  |       26     33.333      -7.333    -1.270 |
  |        0     33.333     -33.333    -5.774 |
  +-------------------------------------------+
```

Obviously, the equi-probability model does not work very well in this case, but there is no reason we would have expected it to.

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 α = .05.

Stata Solution. There are various ways to do this in Stata. Nick Cox's `tabchii` and `tabchi` commands, which are part of his `tab_chi` package, can be used. See their help files. But, Stata's `tabi` and `tabulate` commands are already available for Case II. `tabi` has the following format:

```
tabi #11 #12 [...] \ #21 #22 [...] [\ ...], tabulate_options
```

i.e. you enter the data for row 1, then row 2, etc. The command also includes several options for displaying various statistics and other types of information, e.g. `chi2` gives you the Pearson chi-square, `lrchi2` gives you the Likelihood Ratio Chi-Square, and `exact` gives you Fisher's Exact Test. For this problem,

```
. tabi 55 65 \50 30, chi2 lrchi2 exact

           |          col
       row |         1          2 |     Total
-----------+----------------------+----------
         1 |        55         65 |       120
         2 |        50         30 |        80
-----------+----------------------+----------
     Total |       105         95 |       200

          Pearson chi2(1) =   5.3467   Pr = 0.021
 likelihood-ratio chi2(1) =   5.3875   Pr = 0.020
           Fisher's exact =                 0.022
   1-sided Fisher's exact =                 0.015
```

You could also enter the data like this: let gender = 1 if male, 2 if female; party = 1 if Democrat, 2 = Republican; wgt = frequency. Then,

```
. list  gender party wgt

     +----------------------+
     | gender   party   wgt |
     |----------------------|
  1. |      1       1    55 |
  2. |      1       2    65 |
  3. |      2       1    50 |
  4. |      2       2    30 |
     +----------------------+
```

We can now use Stata's `tabulate` command (which can be abbreviated `tab`). The `[freq=wgt]` parameter tells it to weight each of the four combinations by its frequency.

```
. tab gender party [freq = wgt], chi2 lrchi2 exact

-> tabulation of gender by party

           |         party
    gender |         1          2 |     Total
-----------+----------------------+----------
         1 |        55         65 |       120
         2 |        50         30 |        80
-----------+----------------------+----------
     Total |       105         95 |       200

          Pearson chi2(1) =   5.3467   Pr = 0.021
 likelihood-ratio chi2(1) =   5.3875   Pr = 0.020
           Fisher's exact =                 0.022
   1-sided Fisher's exact =                 0.015
```

If you have individual-level data, e.g. in this case the data set would have 200 individual-level records, the `tab` command is

```
. tab gender party, chi2 lrchi2 exact

-> tabulation of gender by party

           |         party
    gender |         1          2 |     Total
-----------+----------------------+----------
         1 |        55         65 |       120
         2 |        50         30 |        80
-----------+----------------------+----------
     Total |       105         95 |       200

          Pearson chi2(1) =   5.3467   Pr = 0.021
 likelihood-ratio chi2(1) =   5.3875   Pr = 0.020
           Fisher's exact =                 0.022
   1-sided Fisher's exact =                 0.015
```

Sidelights. (1) I used the command `expand wgt` to create an individual-level dataset. This duplicated records based on their frequencies, i.e. it took the tabled data and expanded it into 200 individual-level records. (2) Yates correction for continuity is sometimes used for 1 X 2 and 2 X 2 tables. I personally don't know of any straightforward way to do this in Stata. [UPDATE OCT 2014: The user-written `exactcc` command (`findit exactcc`) can calculate the Yates correction if you really need it.] Fisher's Exact Test is generally better anyway. (3) Fisher's Exact Test is most useful when the sample is small, e.g. one or more expected values is less than 5. With larger N, it might take a while to calculate.

Alternative Approach for 2 X 2 tables. 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 Republican party. Hence, we could also treat this as a two sample problem, case V, test of $p_1 = p_2$. We can use the `prtesti` and `prtest` commands. We'll let p = the probability of being Republican. Using `prtesti`,

```
. prtesti 120 65 80 30, count

Two-sample test of proportion                      x: Number of obs =      120
                                                   y: Number of obs =       80

------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .5416667   .0454848                      .4525181    .6308152
           y |       .375   .0541266                      .2689138    .4810862
-------------+----------------------------------------------------------------
        diff |   .1666667   .0707004                      .0280963     .305237
             |  under Ho:   .0720785     2.31   0.021
------------------------------------------------------------------------------

               Ho: proportion(x) - proportion(y) = diff = 0

       Ha: diff < 0            Ha: diff != 0             Ha: diff > 0
        z =  2.312               z =  2.312                z =  2.312
    P < z =  0.9896        P > |z| =  0.0208          P > z =  0.0104
```

Using a data file, we first create a new version of party that is coded 0 = Democrat, 1 = Republican, and then use the `prtest` command.

```
. gen party2 = party - 1
. prtest  party2, by( gender)

Two-sample test of proportion                     Male: Number of obs =      120
                                                Female: Number of obs =       80

------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Male |   .5416667   .0454848                      .4525181    .6308152
      Female |       .375   .0541266                      .2689138    .4810862
-------------+----------------------------------------------------------------
        diff |   .1666667   .0707004                      .0280963     .305237
             |  under Ho:   .0720785     2.31   0.021
------------------------------------------------------------------------------

           Ho: proportion(Male) - proportion(Female) = diff = 0

       Ha: diff < 0            Ha: diff != 0            Ha: diff > 0
        z =  2.312               z =  2.312              z =  2.312
    P < z =  0.9896        P > |z| =  0.0208         P > z =  0.0104

. * z squared = the chi-square value we got earlier
. display r(z) ^ 2
5.3467001
```

A small advantage of this approach in this case is that the sign of the test statistic is meaningful. The positive and significant z value tells us men are more likely than women to be Republicans.

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 α = .10.

Stata Solution. Problems like this can be addressed using advanced Stata routines like `poisson` and `glm`. For our current purposes, however, Adrian Mander’s `ipf` command (iterative proportional fitting) provides a simple, straightforward solution. (`ipf` also could have been used for some of the previous problems.)

The format of the `ipf` command depends on how the data have been entered. One approach is to enter the data as 8 cases, with the variables gender, race, party and freq:

```
. list , sep(4)

     +---------------------------------------+
     | gender       race        party   freq |
     |---------------------------------------|
  1. |   Male      White   Republican     20 |
  2. |   Male   NonWhite   Republican      5 |
  3. |   Male      White     Democrat     20 |
  4. |   Male   NonWhite     Democrat     15 |
     |---------------------------------------|
  5. | Female      White   Republican     18 |
  6. | Female   NonWhite   Republican      2 |
  7. | Female      White     Democrat     15 |
  8. | Female   NonWhite     Democrat      5 |
     +---------------------------------------+
```

You then use `ipf` specifying `[fw = freq]`, i.e. you weight by the frequency count. (If instead your data set consists of the 100 individual-level cases, then just leave this parameter off.)

The `fit` parameter tells `ipf` what model to fit; by specifying `fit(gender+race+party)` we tell `ipf` to fit the model of independence, i.e. we fit the main effects only but do not allow for any interactions (dependence) among the variables.

```
. ipf [fw = freq], fit(gender + race + party)
Deleting all matrices......

Expansion of the various marginal models
----------------------------------------
marginal model 1 varlist :  gender
marginal model 2 varlist :  race
marginal model 3 varlist :  party
unique varlist  gender race party

N.B.  structural/sampling zeroes may lead to an incorrect df
Residual degrees of freedom = 4
Number of parameters        = 4
Number of cells             = 8

Loglikelihood = 166.0760865136649
Loglikelihood = 166.076086513665

Goodness of Fit Tests
---------------------
df = 4
Likelihood Ratio Statistic G^2 =   9.0042 p-value = 0.061
Pearson Statistic          X^2 =   9.2798 p-value = 0.054
```

These are the same chi-square statistics we got before. If we are using the (rather generous) .10 level of significance, we should reject the model of independence. However, we do not know where the dependence is at this point.

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 $\alpha = .05$.

Stata 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.

To estimate this model with `ipf`, we use the * parameter to allow for an interaction (dependence) between gender and race, but we do not allow for gender or race to interact with party:

```
. ipf [fw = freq], fit(gender + race + party + gender*race)
Deleting all matrices......

Expansion of the various marginal models
----------------------------------------
marginal model 1 varlist :  gender
marginal model 2 varlist :  race
marginal model 3 varlist :  party
marginal model 4 varlist :  gender race
unique varlist  gender race party

N.B.  structural/sampling zeroes may lead to an incorrect df
Residual degrees of freedom = 3
Number of parameters        = 5
Number of cells             = 8

Loglikelihood = 167.6620628360595
Loglikelihood = 167.6620628360595

Goodness of Fit Tests
---------------------
df = 3
Likelihood Ratio Statistic G^2 =   5.8322 p-value = 0.120
Pearson Statistic          X^2 =   5.6146 p-value = 0.132
```

Again, the chi-square statistics are the same as before. Because they are not significant at the .05 level (or .10 for that matter) we do NOT reject the model of conditional independence. Having said that, however, it can be noted that the model probably should include an effect of race on party affiliation, as the fit improves significantly when this interaction is added to the model:

```
. ipf [fw = freq], fit(gender + race + party + gender*race + race*party)

N.B.  structural/sampling zeroes may lead to an incorrect df
Residual degrees of freedom = 2
Number of parameters        = 6
Number of cells             = 8

Loglikelihood = 170.486282357668
Loglikelihood = 170.4862823576681

Goodness of Fit Tests
---------------------
df = 2
Likelihood Ratio Statistic G^2 =   0.1838 p-value = 0.912
Pearson Statistic          X^2 =   0.1841 p-value = 0.912

. display 5.8322-.1838
5.6484

. display chi2tail(1, 5.6484)
.01747131
```

Note that, when the race*party interaction is added to the model, the Likelihood Ratio Chi-Square drops from 5.8322 to .1838, i.e. by 5.6484. This change (which has 1 degree of freedom) is significant at the .0175 level, implying that we should allow for a race*party interaction. We'll talk more about chi-square contrasts between models during 2nd semester.