

University of North Carolina
Chapel Hill

Soci708-001 Statistics for Sociologists

Fall 2009

Professor François Nielsen

Stata Commands for Module 9 – Analysis of Two-Way Tables

For further information on any command in this handout, simply type `help` followed by the name of the command in Stata.

See also the Stata and SAS Guide pdf (click on Documents in side bar; guide is linked under Software Documentation).

Life is complicated. This is a lower case letter chi, squared: χ^2 ; this is an upper case chi, squared: X^2 , which looks confusingly like an upper-case x, squared: X^2 .

1 Statistical Functions in Stata

1.1 Normal Distribution Functions

The function `normal(z)` returns $P(Z \leq z)$, the area under the standard normal curve to the left of z . (Compare with Table A.)

```
. display normal(1.207)
.88628393
```

The function `invnormal(p)` returns z such that $P(Z \leq z) = p$, i.e. such that the area under the standard normal curve to the left of z is p . (Compare with Table A and Table D (bottom row).)

```
. display invnormal(0.975)
1.959964
```

1.2 Chi-Square Distribution Functions

The function `chi2tail(df, x)` returns $P(X > x)$, the area under the chi-square distribution with df degrees of freedom to the right of x . (Compare with Table F.)

```
. display chi2tail(1, 4.31)
.03788896
```

The function `invttail(df, p)` returns x such that $P(X > x) = p$, i.e. such that the area under the chi-square distribution with df degrees of freedom to the right of x is p . (Compare with Table F.)

```
. display invchi2tail(1, .03788896)
4.3100001
```

2 Chi-Square for Two-Way Table Using Summary Data

For this example, we use the `tabi` command, one of the so-called “immediate” commands in Stata. The “immediate” commands allow you to conduct analyses without having the original dataset. According to Stata Help, “`tabi` displays the $r \times c$ table using the values specified; rows are separated by `\`.” By using the option `chi2` (after a comma), you are requesting that Stata include the chi-square statistic. (You can also just type `chi`.)

The following printout uses data from IPS6e Exercise 9.11 p.549. Rows represent treatments for cocaine addiction (Desipramine, Lithium, Placebo). Columns represent Relapse, No relapse. Subjects are people addicted to cocaine. In the second printout I have specified `nofreq row` to get the percentage of Relapse/No relapse conditional on treatment.

```
. tabi 10 14\ 18 6\ 20 4, chi2 V
```

row	col 1	2	Total
1	10	14	24
2	18	6	24
3	20	4	24
Total	48	24	72

```
Pearson chi2(2) = 10.5000 Pr = 0.005  
Cramér's V = 0.3819
```

```
. tabi 10 14\ 18 6\ 20 4, chi2 V nofreq row
```

row	col 1	2	Total
1	41.67	58.33	100.00
2	75.00	25.00	100.00
3	83.33	16.67	100.00
Total	66.67	33.33	100.00

```
Pearson chi2(2) = 10.5000 Pr = 0.005  
Cramér's V = 0.3819
```

Another way to enter data already in tabular form is to create a spreadsheet in Excel, then paste the table in the Data Editor. For example using the summary data for drinking behavior by religious denomination (see next section), we could enter the data as follows.

First create the spreadsheet (with a header for each column) and copy it to the clipboard.

relig	drinks	count
1	0	39
2	0	6
3	0	2
4	0	5

1	1	94
2	1	40
3	1	21
4	1	47

Then in Stata paste the table into the Data Editor and *close* the Data Editor (click on ×). Cell frequencies are input as the variable count. Note that we need to use the expression `[fweight=count]` to tell Stata that the variable count contains cell frequencies.

```
. tab relig drinks [fweight=count], all
```

relig	drinks		Total
	0	1	
1	39	94	133
2	6	40	46
3	2	21	23
4	5	47	52
Total	52	202	254

```

      Pearson chi2(3) = 13.6827   Pr = 0.003
likelihood-ratio chi2(3) = 14.4258   Pr = 0.002
      Cramér's V = 0.2321
              gamma = 0.4915   ASE = 0.121
      Kendall's tau-b = 0.2105   ASE = 0.051

```

See next section for more details on the substantive example.

3 Chi-Square for Two-Way Table Using the Original Dataset

Another situation is when you have the original data with observations on individual cases. In this example we use the Afifi and Clark data (`survey2b.dta`) to look at the association between drinking behavior (whether a respondent drinks alcohol) and religious denomination.

The example also shows how to define value labels (to facilitate interpretation of output) for variables `drinks` and `religion`, and how to assign missing values for denomination to two doubtful cases.

The last command shows the options to correctly percentage the table to show the dependence of drinking behavior on religious denomination, in this case percentage within rows (`row nofreq`), and to request two kinds of chi-squares and Cramér's V (`V chi2 lr`).

```

. * in Stata

. use "D:\soci708\data\survey2b.dta", clear

. tab drinks

... (output not shown)

. label define noyes 0 "No" 1 "Yes"

```

```
. label values drinks noyes
```

```
. tab drinks
```

DRINKS	Freq.	Percent	Cum.
No	52	20.31	20.31
Yes	204	79.69	100.00
Total	256	100.00	

```
. tab religion
```

RELIGION	Freq.	Percent	Cum.
1	133	51.95	51.95
2	46	17.97	69.92
3	23	8.98	78.91
4	52	20.31	99.22
6	2	0.78	100.00
Total	256	100.00	

```
. * set missing to two odd answers
```

```
. replace religion=. if religion==6  
(2 real changes made, 2 to missing)
```

```
. label define denom 1 "Prot" 2 "Cath" 3 "Jewi" 4 "None"
```

```
. label values religion denom
```

```
. tab religion drinks
```

RELIGION	DRINKS		Total
	No	Yes	
Prot	39	94	133
Cath	6	40	46
Jewi	2	21	23
None	5	47	52
Total	52	202	254

```
. tab religion drinks, row nofreq V chi2 lr
```

RELIGION	DRINKS		Total
	No	Yes	
Prot	29.32	70.68	100.00
Cath	13.04	86.96	100.00
Jewi	8.70	91.30	100.00
None	9.62	90.38	100.00
Total	20.47	79.53	100.00

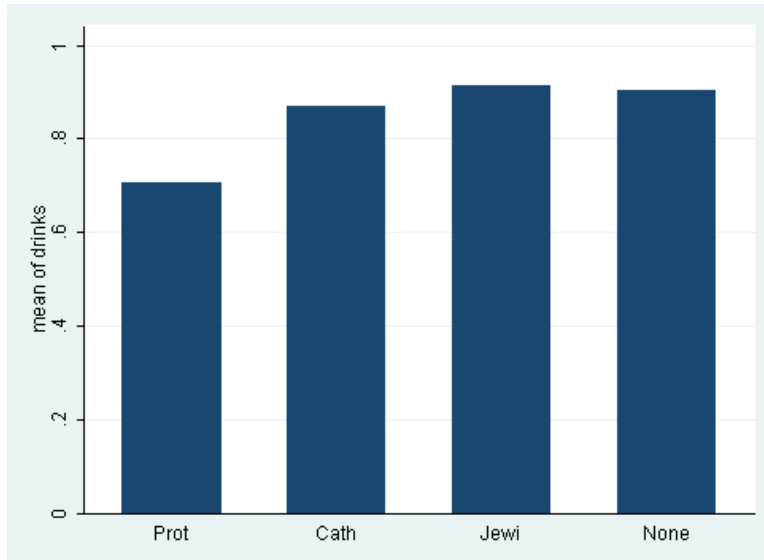


Figure 1: Proportion of respondents who drink (drinks) by religious denomination (religion).

```

Pearson chi2(3) = 13.6827 Pr = 0.003
likelihood-ratio chi2(3) = 14.4258 Pr = 0.002
Cramér's V = 0.2321

```

Finally the following command will represent the bar graph of drinks mean by denomination shown in Figure 1.

```
. graph bar drinks, over(religion)
```

4 Chi-Square Goodness of Fit Test

Some software have a dedicated procedure for the chi-square goodness of fit test. The following is an example using the program R. The test compares the age distribution of a sample from the jury pool in a district with the age distribution given by the census for the district (this is in the context of a court case, where the defense argues that the jury pool is nonrepresentative of district population). Frequencies are the numbers of individuals in age categories 18–19, 20–24, 25–29, 30–39, 40–49, 50–64, 65+. Probabilities are proportions from the census for the district.

```

> # in R
> freqs<-c(23, 96, 134, 293, 297, 380, 113)
> ps<-c(0.061, 0.150, 0.135, 0.217, 0.153, 0.182, 0.102)
> chisq.test(freqs, correct=FALSE, p=ps)

```

```
Chi-squared test for given probabilities
```

```

data: freqs
X-squared = 231.26, df = 6, p-value < 2.2e-16

```

Stata does not have a function for the chi-square goodness of fit test. I use the trick explained in IPS6e p. 547 (bottom paragraph) to input the probabilities

multiplied by a large number (I use 100,000) and use the tabi command. The chi-square of 228 is not too far off the value of 231 obtained with R. In any case the conclusion is the same that there is a significant difference between the age distribution of the jury pool and that of the district population.

. * trick for goodness of fit test in Stata (IPS6e p.547)

. tabi 23 96 134 293 297 380 113\6100 15000 13500 21700 15300 18200 10200, all

row	1	2	3	4	5	6	Total
1	23	96	134	293	297	380	1,336
2	6,100	15,000	13,500	21,700	15,300	18,200	100,000
Total	6,123	15,096	13,634	21,993	15,597	18,580	101,336

row	7	Total
1	113	1,336
2	10,200	100,000
Total	10,313	101,336

Pearson chi2(6) = 228.2214 Pr = 0.000
likelihood-ratio chi2(6) = 242.9677 Pr = 0.000
Cramér's V = 0.0475
gamma = -0.2159 ASE = 0.016
Kendall's tau-b = -0.0316 ASE = 0.002