Chapter 6 Stata v10.1 Analysis Examples Syntax and Output

General Notes on Stata 10.1

Given that this tool is used throughout the ASDA textbook this chapter includes only the syntax and output for the analysis examples provided in Chapter 6. Stata 10.1 is an excellent tool for survey data analysis as well as graphing and related data management tasks. It offers a very comprehensive set of svy commands as well as weighted graphics and convenient syntax and data management abilities. For these reasons, we use Stata as the primary software for the ASDA text.

The examples and syntax presented here assume that all data management including variable construction, labels for variable values and other preparation steps are complete. See the Stata documentation for assistance with these issues.

One analysis included in this chapter cannot be run in Stata 10.1. For example, the CMH trends test (example 6.1) is run in Sudaan rather than Stata.

Please check the Stata documentation and also the ASDA web site for updates to Stata as new versions are released. For example, we have already included an example of how to use Stata 11.0 with the new "factor" variable features/syntax and compared this to the older “xi” type of syntax for including categorical variables in data analysis.
. * 6.1 NHANES DATA
. svy, subpop(age18p): tab irregular, se ci col deff
(running tabulate on estimation sample)

Number of strata   =        15                 Number of obs      =       9906
Number of PSUs     =        30                 Population size    =  284231791
                           Subpop. no. of obs =        5121      Subpop. size     =  210315370
                           Design df          =         15

1=yes     |     column          se          lb          ub        deff
----------
0 |      .9705       .0067       .9525       .9818       11.33
1 |      .0295       .0067       .0182       .0475       11.33

Total |          1

Key: column    =  column proportions
      se        =  linearized standard errors of column proportions
      lb        =  lower 95% confidence bounds for column proportions
      ub        =  upper 95% confidence bounds for column proportions
      deff      =  deff for variances of column proportions

. svy, subpop(age18p): prop irregular
(running proportion on estimation sample)

Survey: Proportion estimation
Number of strata =      15       Number of obs    =       9906
Number of PSUs   =      30       Population size  =  284231791
Subpop. no. obs  =       5121  Subpop. size     =  210315370
Design df        =         15

|             Linearized
| Proportion   Std. Err.     [95% Conf. Interval]
-------------
irregular    |
0 |    .970469   .0066546       .956285    .9846529
1 |    .029531   .0066546      .0153471     .043715

. estat effects

|             Linearized
| Proportion   Std. Err.       DEFF      DEFT
-------------
-------------
irregular    |
0 |    .970469   .0066546      11.325   3.36526
1 |    .029531   .0066546      11.325   3.36526

. svy: mean irregular, subpop(age18p)
(running mean on estimation sample)

Survey: Mean estimation
Number of strata =      15       Number of obs    =       9906
Number of PSUs   =      30       Population size  =  284231791
Subpop. no. obs  =       5121  Subpop. size     =  210315370
Design df        =         15

|             Linearized
|       Mean   Std. Err.     [95% Conf. Interval]
-------------
irregular    |
0 |    .029531   .0066546      .0153471     .043715

. estat effects

|             Linearized
|       Mean   Std. Err.       DEFF      DEFT
-------------
-------------
irregular    |
0 |    .029531   .0066546      11.325   3.36526
* example 6.2
* add a table of race/eth from NHANES file
.svy: proportion ridreth1, subpop(age18p)
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata = 15       Number of obs = 10348
Number of PSUs = 30       Population size = 291616892
Subpop. no. obs = 5334       Subpop. size = 217700471
Design df = 15

--------------------------------------------------------------
|             Linearized
| Proportion   Std. Err.     [95% Conf. Interval]
-------------
ridreth1     |
1 |   .0807834   .0100534   .059355   .1022117
2 |   .0337852   .0074214   .0179669   .0496035
3 |   .7141428   .0276981   .6551057   .77318
4 |   .1172628   .0198491   .0749555   1.159701
5 |   .0540257   .005825    .04161   .0664414

. estat effects

----------------------------------------------------------
|             Linearized
| Proportion   Std. Err.       DEFF      DEFT
-------------
ridreth1     |
1 |   .0807834   .0100534     10.2808   3.20638
2 |   .0337852   .0074214     12.7441   3.56989
3 |   .7141428   .0276981     28.3863   5.32788
4 |   .1172628   .0198491     28.7495   5.36185
5 |   .0540257   .005825     5.01483   2.23938

. tab ridreth1 if age18p==1

1=mex 2=oth
3=white
4=black
5=other

<table>
<thead>
<tr>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
</table>
1 | 1,185 | 21.30 | 21.30 |
2 | 171  | 3.07  | 24.38 |
3 | 2,633 | 47.33 | 71.71 |
4 | 1,341 | 24.11 | 95.81 |
5 | 233  | 4.19  | 100.00 |

Total | 5,563 | 100.00 |
* example 6.3

```
. svy: tab bp_cat , subpop(age18p) obs se ci
(running tabulate on estimation sample)
```

Number of strata = 15                 Number of obs      = 9842
Number of PSUs   = 30                 Population size    = 281904771
Subpop. no. of obs = 5057
Subpop. size     = 207988351
Design df         = 15

```
<table>
<thead>
<tr>
<th>bp_cat</th>
<th>proportions</th>
<th>se</th>
<th>lb</th>
<th>ub</th>
<th>obs</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>.4711</td>
<td>.0111</td>
<td>.4475</td>
<td>.4948</td>
<td>2441</td>
</tr>
<tr>
<td>pre-hype</td>
<td>.4185</td>
<td>.0118</td>
<td>.3937</td>
<td>.4439</td>
<td>1988</td>
</tr>
<tr>
<td>stage 1</td>
<td>.0864</td>
<td>.0062</td>
<td>.0741</td>
<td>.1006</td>
<td>470</td>
</tr>
<tr>
<td>stage 2</td>
<td>.024</td>
<td>.0024</td>
<td>.0193</td>
<td>.0297</td>
<td>158</td>
</tr>
<tr>
<td>Total</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>5057</td>
</tr>
</tbody>
</table>

Key:  proportions  =  cell proportions
      se      =  linearized standard errors of cell proportions
      lb      =  lower 95% confidence bounds for cell proportions
      ub      =  upper 95% confidence bounds for cell proportions
      obs     =  number of observations

. svy: tab bp_cat , subpop (age18p) deff
(running tabulate on estimation sample)

Number of strata = 15                 Number of obs      = 9842
Number of PSUs   = 30                 Population size    = 281904771
Subpop. no. of obs = 5057
Subpop. size     = 207988351
Design df         = 15

```
<table>
<thead>
<tr>
<th>bp_cat</th>
<th>proportions</th>
<th>deff</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>.4711</td>
<td>3.581</td>
</tr>
<tr>
<td>pre-hype</td>
<td>.4185</td>
<td>4.145</td>
</tr>
<tr>
<td>stage 1</td>
<td>.0864</td>
<td>3.544</td>
</tr>
<tr>
<td>stage 2</td>
<td>.024</td>
<td>1.794</td>
</tr>
<tr>
<td>Total</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Key:  proportions  =  cell proportions
      deff      =  deff for variances of cell proportions
* example 6.4
* use 4 category blood pressure variable to examine multinomial gof tool
* note bp_cat 1=normal 2=pre-hypertensive 3=stage 1 high blood pressure 4=stage 2 high blood pressure

.svym: tab bp_cat, subpop(age18p) se ci col deff
(running tabulate on estimation sample)

Number of strata = 15 Number of obs = 9842
Number of PSUs = 30 Population size = 281904771
Subpop. no. of obs = 5057
Subpop. size = 207988351
Design df = 15

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

bp_cat | column se lb ub deff
----------
1 | .4711 .0111 .4475 .4948 3.581
2 | .4185 .0118 .3937 .4439 4.145
3 | .0864 .0062 .0741 .1006 3.544
4 | .024 .0024 .0193 .0297 1.794
Total | 1

Key: column = column proportions
se = linearized standard errors of column proportions
lb = lower 95% confidence bounds for column proportions
ub = upper 95% confidence bounds for column proportions
deff = deff for variances of column proportions

.recode bp_cat (1=.5) (2=.3) (3=.15) (4=.05), generate (p)
(7359 differences between bp_cat and p)

.mgof bp_cat=p if age18p==1, svym

Number of strata = 15 Number of obs = 5057
Number of PSUs = 30 Pop size = 2.1e+08
Design df = 15
N of outcomes = 4
F df1 = 2.62707
F df2 = 39.406

Goodness-of-fit Coef. F-value P-value
-----------------------------------------
Pearson's X2 | 450.1971 69.5663 0.0000
Log likelihood ratio | 465.482 71.9281 0.0000
* figure 6.3
  . graph pie bp_cat1 bp_cat2 bp_cat3 bp_cat4 [pweight=wtmec2yr] if age18p==1, plabel(_all percent, format(%9.1f)) ///
  > scheme(s2mono) ///
  > legend (label(1 "Normal")label(2 "Pre-Hypertensive") label(3 "Stage 1 Hypertensive") label (4 "Stage 2 Hypertensive"))
* figure 6.4
graph bar (mean) bp_cat1 bp_cat2 bp_cat3 bp_cat4 [pweight=wtmec2yr] if age18p==1, percentages ///
> bar(1,color(gs12)) bar(2,color(gs4)) bar(3,color(gs8)) bar(4,color(black)) ///
> blabel(bar, format(%5.1f)) bargap(7) scheme(s2mono) ///
> legend (label(1 "Normal") label(2 "Pre-Hypertensive") label(3 "Stage1 Hypertensive") label(4 "Stage 2 Hypertensive")) ///
> ytitle ("Percentage")
Example 6.6 NCS-R DATA

```
svset seclustr [pweight=ncsrwtsh], strata(sestrat)
pweight: ncsrwtsh
VCE: linearized
Single unit: missing
Strata 1: sestrat
SU 1: seclustr
FPC 1: <zero>
```

```
.svy: tab sex mde, se ci deff
```

```
(running tabulate on estimation sample)
```

<table>
<thead>
<tr>
<th>sex</th>
<th>MajorDepEpisode</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
<td>Total</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.4066</td>
<td>0.0722</td>
<td>0.4789</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.007)</td>
<td>(0.0034)</td>
<td>(0.0053)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[.3926,</td>
<td>[.0656,</td>
<td>[.4681,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.4208]</td>
<td>.0795]</td>
<td>.4681]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.874</td>
<td>1.637</td>
<td>1.051</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.4016</td>
<td>0.1195</td>
<td>0.5211</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0054)</td>
<td>(0.003)</td>
<td>(0.0053)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[.3909,</td>
<td>[.1135,</td>
<td>[.5104,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.4125]</td>
<td>.1258]</td>
<td>.5104]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.11</td>
<td>0.8086</td>
<td>1.051</td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>.8083</td>
<td>.1917</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0049)</td>
<td>(0.0049)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[.7983,</td>
<td>[.1821,</td>
<td>.8179]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.424</td>
<td>1.424</td>
<td></td>
</tr>
</tbody>
</table>

Key: cell proportions
(Linearized standard errors of cell proportions)
[95% confidence intervals for cell proportions]
deff for variances of cell proportions

Pearson:
Uncorrected  ch2(1) = 92.1499
Design-based  F(1, 42) = 57.9784  P = 0.0000

Mean generalized deff = 1.3727
CV of generalized deffs = 0.0000

```
.svy: tab sex mde, row se ci deff
```

```
(running tabulate on estimation sample)
```

<table>
<thead>
<tr>
<th>sex</th>
<th>MajorDepEpisode</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
<td>Total</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.8492</td>
<td>0.1508</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0077)</td>
<td>(0.0077)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[.8329,</td>
<td>[.1358,</td>
<td>[.8642]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.8642]</td>
<td>.1671]</td>
<td>.8329]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.083</td>
<td>2.083</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.7707</td>
<td>0.2293</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0056)</td>
<td>(0.0056)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[.7591,</td>
<td>[.2181,</td>
<td>[.7819]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.7819]</td>
<td>.2409]</td>
<td>.7591]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.8728</td>
<td>.8728</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>.8083</td>
<td>.1917</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0049)</td>
<td>(0.0049)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[.7983,</td>
<td>[.1821,</td>
<td>.8179]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.424</td>
<td>1.424</td>
<td></td>
</tr>
</tbody>
</table>

Key: row proportions
(Linearized standard errors of row proportions)
[95% confidence intervals for row proportions]
deff for variances of row proportions

Pearson:
Uncorrected  ch2(1) = 92.1499
Design-based  F(1, 42) = 57.9784  P = 0.0000

Mean generalized deff = 1.3727
CV of generalized deffs = 0.0000
* example 6.7 mde with over statement for sex NCS-R DATA

.svy: prop mde
(running proportion on estimation sample)

Survey: Proportion estimation

<table>
<thead>
<tr>
<th>Number of strata = 42</th>
<th>Number of obs = 9282</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of PSUs = 84</td>
<td>Population size = 9282</td>
</tr>
<tr>
<td>Design df = 42</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Linearized</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>mde</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>.8082888 .0048768</td>
</tr>
<tr>
<td></td>
<td>.798447 .8181306</td>
</tr>
<tr>
<td>1</td>
<td>.1917112 .0048768</td>
</tr>
<tr>
<td></td>
<td>.1818694 .201553</td>
</tr>
</tbody>
</table>

.svy: proportion mde, over(sex)
(running proportion on estimation sample)

Survey: Proportion estimation

<table>
<thead>
<tr>
<th>Number of strata = 42</th>
<th>Number of obs = 9282</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of PSUs = 84</td>
<td>Population size = 9282</td>
</tr>
<tr>
<td>Design df = 42</td>
<td></td>
</tr>
</tbody>
</table>

._prop_1: mde = 0
._prop_2: mde = 1

Male: sex = Male
Female: sex = Female

<table>
<thead>
<tr>
<th>Linearized</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>_prop_1</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>.8492067 .0077478</td>
</tr>
<tr>
<td></td>
<td>.8335709 .8648424</td>
</tr>
<tr>
<td>Female</td>
<td>.7706917 .0056473</td>
</tr>
<tr>
<td></td>
<td>.7592951 .7820883</td>
</tr>
<tr>
<td>_prop_2</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>.1507933 .0077478</td>
</tr>
<tr>
<td></td>
<td>.1351576 .1664291</td>
</tr>
<tr>
<td>Female</td>
<td>.2293083 .0056473</td>
</tr>
<tr>
<td></td>
<td>.2179117 .2407049</td>
</tr>
</tbody>
</table>

.lincom [_prop_2]Male - [_prop_2]Female
( 1) [._prop_2]Male - [._prop_2]Female = 0

| Coef. | Std. Err. | t   | P>|t| | [95% Conf. Interval] |
|-------|-----------|-----|------|----------------------|
| (1)   | -.078515  | .0095518 | -8.22 | 0.000 | -.0977914 -.0592386 |
. * example 6.8 use part 2 weight for alcohol dependence NCS-R DATA
. svyset seclustr [pweight=ncsrwtlg], strata(seestrat) vce(linearized) singleunit(missing)

    pweight: ncsrwtlg
    VCE: linearized
    Single unit: missing
    Strata 1: seestrat
    SU 1: seclustr
    FPC 1: <zero>

. svy: tab ed4cat ald, subpop (if age < 29) se ci row deff
(running tabulate on estimation sample)

<table>
<thead>
<tr>
<th>Number of strata</th>
<th>Number of obs</th>
<th>Population size</th>
<th>Subpop. no. of obs</th>
<th>Subpop. size</th>
<th>Design df</th>
</tr>
</thead>
<tbody>
<tr>
<td>42</td>
<td>5692</td>
<td>5692.0005</td>
<td>1275</td>
<td>1266.5565</td>
<td>42</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>years of education</th>
<th>0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-11</td>
<td>.9087</td>
<td>.0913</td>
<td>1</td>
</tr>
<tr>
<td>(0.0294)</td>
<td>(0.0294)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[.8297,.9531]</td>
<td>[.0469,.1703]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.3</td>
<td>2.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>.9514</td>
<td>.0486</td>
<td>1</td>
</tr>
<tr>
<td>(0.0135)</td>
<td>(0.0135)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[.9159,.9724]</td>
<td>[.0276,.0841]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.585</td>
<td>1.585</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13-15</td>
<td>.951</td>
<td>.049</td>
<td>1</td>
</tr>
<tr>
<td>(0.01)</td>
<td>(0.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[.9263,.9678]</td>
<td>[.0322,.0737]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.9407</td>
<td>.9407</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16+</td>
<td>.931</td>
<td>.069</td>
<td>1</td>
</tr>
<tr>
<td>(0.0136)</td>
<td>(0.0136)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[.8978,.9539]</td>
<td>[.0461,.1022]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.5983</td>
<td>.5983</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>.9405</td>
<td>.0595</td>
<td>1</td>
</tr>
<tr>
<td>(0.0088)</td>
<td>(0.0088)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[.92,.956]</td>
<td>[.044,.08]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.0855</td>
<td>.0855</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Key: row proportions
(Linearized standard errors of row proportions)
[95% confidence intervals for row proportions]
deff for variances of row proportions

Pearson:
Uncorrected chi2(3) = 27.2130
Design-based F(2.75, 115.53) = 1.6498    P = 0.1857

Mean generalized deff = 6.6250
CV of generalized deffs = 0.5594
. * example 6.9 simple logistic regression of mde on male part 1 weight NCS-R DATA

. svyset seclustr [pw=ncsrwtsh], strata(sestrat) vce(linearized) singleunit(missing)

    pweight: ncsrwtsh
    VCE: linearized
    Single unit: missing
    Strata 1: sestrat
    SU 1: seclustr
    FPC 1: <zero>

. svy: logistic mde sexm
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        42                  Number of obs      =      9282
Number of PSUs     =        84                  Population size    = 9282.0002
Design df          =        42                  F(1, 42) = 57.28
Prob > F           =    0.0000

------------------------------------------------------------------------------
       |             Linearized         |                |
       | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]     |
-------------|----------------------|----------------------|----------------------|----------------------|
sexm | .5968012   .0407038 -7.57   0.000     .5200601   .6848663 |
------------------------------------------------------------------------------

. * figure 6.8

. graph bar (mean) bp_cat1 bp_cat2 bp_cat3 bp_cat4 [pw=wtme2yr] if age18p==1, blabel(bar, format(%9.1f) color(none)) ///
> bar(1,color(gs12)) bar(2,color(gs4)) bar(3,color(gs8)) bar(4,color(black)) ///
> bargap(7) scheme(s2mono) over(riagendr) percentages ///
> legend (label(1 "Normal")label(2 "Pre-Hypertensive") label(3 "Stage 1 Hypertensive") label (4 "Stage 2 Hypertensive") ///
> ytitle ("Percentage")
*example 6.10 done in Sudaan as Stata does not offer a complex design corrected CMH trends test