Statistics/Data Analysis

```
      log:  /scratch/OhioFamilyHealthSurvey/analysislog.smcl
  log type:  smcl
 opened on:  14 Jul 2008, 15:09:22

1 . do /tmp/SD12619.000000

2 . /*************************************************************/
3 . /*  INTRODUCTION TO STATA - toys using brain-IQ data         */
4 . /*************************************************************/
5 .
6 . /*input data using 'point and click'*/
7 .
8 . /*clear the memory*/
9 . clear

10.
11. /*same thing using code*/
12. insheet using "/scratch/OhioFamilyHealthSurvey/brainIQ.dat"
  (7 vars, 38 obs)

13.
14.
15. /*look at the data using the data editor*/
16. /*look at the data using 'list'*/
17. list
```

	gender	fsiq	viq	piq	weight	height	mri_co~t
1.	1	133	132	124	118	64.5	816932
2.	0	139	123	150	143	73.3	1038437
3.	0	133	129	128	172	68.8	965353
4.	1	137	132	134	147	65	951545
5.	1	99	90	110	146	69	928799
6.	1	138	136	131	138	64.5	991305
7.	1	92	90	98	175	66	854258
8.	0	89	93	84	134	66.3	904858
9.	0	133	114	147	172	68.8	955466
10.	1	132	129	124	118	64.5	833868
11.	0	141	150	128	151	70	1079549
12.	0	135	129	124	155	69	924059
13.	1	140	120	147	155	70.5	856472
14.	1	96	100	90	146	66	878897
15.	1	83	71	96	135	68	865363
16.	1	132	132	120	127	68.5	852244
17.	0	100	96	102	178	73.5	945088
18.	1	101	112	84	136	66.3	808020
19.	0	80	77	86	180	70	889083
20.	0	97	107	84	186	76.5	905940
21.	1	135	129	134	122	62	790619
22.	0	139	145	128	132	68	955003
23.	1	91	86	102	114	63	831772
24.	0	141	145	131	171	72	935494
25.	1	85	90	84	140	68	798612
26.	0	103	96	110	187	77	1062462
27.	1	77	83	72	106	63	793549
28.	1	130	126	124	159	66.5	866662
29.	1	133	126	132	127	62.5	857782
30.	0	144	145	137	191	67	949589
31.	0	103	96	110	192	75.5	997925
32.	0	90	96	86	181	69	879987
33.	1	83	90	81	143	66.5	834344
34.	1	133	129	128	153	66.5	948066
35.	0	140	150	124	144	70.5	949395

36.	1	88	86	94	139	64.5	893983
37.	0	81	90	74	148	74	930016
38.	0	89	91	89	179	75.5	935863

```
18.
19. set more off /*tells STATA to not wait for scrolling*/

20.
21. /*variable summaries*/
22. summarize                          /*summarize all the variables*/
```

Variable	Obs	Mean	Std. Dev.	Min	Max
gender	38	.5263158	.5060094	0	1
fsiq	38	113.5526	23.81539	77	144
viq	38	112.1316	22.9396	71	150
piq	38	111.3421	22.59787	72	150
weight	38	151.0526	23.47851	106	192
height	38	68.42105	3.99379	62	77
mri_count	38	906754.2	72561.75	790619	1079549

```
23. summarize gender fsiq mri_count   /*summarize just some of the variables*/
```

Variable	Obs	Mean	Std. Dev.	Min	Max
gender	38	.5263158	.5060094	0	1
fsiq	38	113.5526	23.81539	77	144
mri_count	38	906754.2	72561.75	790619	1079549

```
24.
25. help summarize

26. summarize fsiq, detail            /*use options to show more detail*/
```

FSIQ

Percentiles		Smallest		
1%	77	77		
5%	80	80		
10%	83	81	Obs	38
25%	90	83	Sum of Wgt.	38
50%	116.5		Mean	113.5526
		Largest	Std. Dev.	23.81539
75%	135	140		
90%	140	141	Variance	567.1728
95%	141	141	Skewness	-.1248064
99%	144	144	Kurtosis	1.271074

```
27. su f, d                           /*Stata allows extreme abbreviation*/
```

FSIQ

Percentiles		Smallest		
1%	77	77		
5%	80	80		
10%	83	81	Obs	38
25%	90	83	Sum of Wgt.	38
50%	116.5		Mean	113.5526
		Largest	Std. Dev.	23.81539
75%	135	140		
90%	140	141	Variance	567.1728
95%	141	141	Skewness	-.1248064
99%	144	144	Kurtosis	1.271074

```
28.
29. /*find means*/
30. mean fsiq

  Mean estimation                     Number of obs    =      38
```

	Mean	Std. Err.	[95% Conf.	Interval]
fsiq	**113.5526**	**3.863367**	**105.7247**	**121.3806**

```
31.
32. /*find proportions*/
33. prop gender

  Proportion estimation               Number of obs    =      38
```

	Proportion	Std. Err.	[95% Conf.	Interval]
gender				
0	**.4736842**	**.0820856**	**.3073631**	**.6400054**
1	**.5263158**	**.0820856**	**.3599946**	**.6926369**

```
34.
35. /*make graphs*/
36. scatter f m

37. scatter f m, by(gender)

38. graph box f, over(g)

39.
40. /*make new variables as functions of old variables*/
41. gen weightkg = weight/2.2

42. gen heightm = height*0.0254

43. gen BMI = weightkg/heightm^2

44.
45. summarize w h BMI
  w ambiguous abbreviation
  r(111);

  end of do-file

  r(111);

46. do /tmp/SD12619.000000

47. summarize w* h* BMI
```

Variable	Obs	Mean	Std. Dev.	Min	Max
weight	**38**	**151.0526**	**23.47851**	**106**	**192**
weightkg	**38**	**68.66029**	**10.67205**	**48.18182**	**87.27273**
height	**38**	**68.42105**	**3.99379**	**62**	**77**
heightm	**38**	**1.737895**	**.1014423**	**1.5748**	**1.9558**
BMI	**38**	**22.67579**	**2.583314**	**18.75158**	**29.97738**

```
48.
49. /*drop extra variables*/
50. drop weightkg heightm

51.
52. /*make new categorical variables*/
53. gen IQcat = 0

54. replace IQcat = 1 if fsiq>116.5
  (19 real changes made)

55. summarize IQcat, d
```

IQcat

Percentiles		Smallest		
1%	**0**	**0**		
5%	**0**	**0**		
10%	**0**	**0**	Obs	**38**
25%	**0**	**0**	Sum of Wgt.	**38**
50%	**.5**		Mean	**.5**
		Largest	Std. Dev.	**.5067117**
75%	**1**	**1**		
90%	**1**	**1**	Variance	**.2567568**
95%	**1**	**1**	Skewness	**0**
99%	**1**	**1**	Kurtosis	**1**

```
56. graph box fsiq, over(IQcat)

57.
58. /*Using results of functions*/
59. summarize f, d
```

FSIQ

Percentiles		Smallest		
1%	**77**	**77**		
5%	**80**	**80**		
10%	**83**	**81**	Obs	**38**
25%	**90**	**83**	Sum of Wgt.	**38**
50%	**116.5**		Mean	**113.5526**
		Largest	Std. Dev.	**23.81539**
75%	**135**	**140**		
90%	**140**	**141**	Variance	**567.1728**
95%	**141**	**141**	Skewness	**−.1248064**
99%	**144**	**144**	Kurtosis	**1.271074**

```
60. return list
```

```
  scalars:
                  r(N) =  38
              r(sum_w) =  38
               r(mean) =  113.5526315789474
                r(Var) =  567.1728307254624
                 r(sd) =  23.81539062718608
           r(skewness) =  −.1248064433885799
           r(kurtosis) =  1.271073860645908
                r(sum) =  4315
                r(min) =  77
                r(max) =  144
                 r(p1) =  77
                 r(p5) =  80
                r(p10) =  83
                r(p25) =  90
                r(p50) =  116.5
                r(p75) =  135
                r(p90) =  140
                r(p95) =  141
                r(p99) =  144
```

```
61. display r(p50)
  116.5

62.
63. gen IQcat2 = 0

64. replace IQcat2 = 1 if fsiq > r(p50)
  (19 real changes made)

65.
66. tab IQcat IQcat2
```

IQcat	IQcat2 0	1	Total
0	**19**	**0**	**19**
1	**0**	**19**	**19**
Total	**19**	**19**	**38**

```
67.
68. clear

69.
  end of do-file

70. do /tmp/SD12619.000000

71. /*************************************************************/
72. /*  STATA ANALYSIS of OFHS                                   */
73. /*************************************************************/
74.
75. set memory 1g  /*increase available memory, since the data is huge*/
  (1048576k)

76. use "ofhs_virgin_file.dta"   /*read in the data*/

77.
78. /*recode some missing data*/
79. replace s10 = . if s10>9  /*maximum recorded adults = 9 */
  (0 real changes made)

80. replace s12 = . if s12>12 /*maximum number of children = 12 */
  (32 real changes made, 32 to missing)

81.
82. /*create a new variable - yes/no for children in hh*/
83. gen kids = 0

84. replace kids = 1 if s12>0
  (16375 real changes made)

85.
86.
87. /*************************************************************/
88. /* Estimation with SRS                                       */
89. /*************************************************************/
90.
91. /***************************************/
```

```
92. /*create a pseudo-SRS from Butler County*/
93.
94. keep if stratum==9  /*keep only one county - this is the coded number for Butler*/
  (38863 observations deleted)

95. keep if s10!=.      /*keep only observations with no missing values - DON'T DO THIS IN PRACTICE!*/
  (4 observations deleted)

96. keep if s12!=.
  (4 observations deleted)

97. gen N=123018

98. gen n=1082

99.
100 save "/scratch/OhioFamilyHealthSurvey/ButlerCounty.dta", replace  /*save the reduced data set*/
  file /scratch/OhioFamilyHealthSurvey/ButlerCounty.dta saved

101 clear

102
103 set memory 8000    /*don't need as much memory for this reduced data set*/
  (8000k)

104 use "/scratch/OhioFamilyHealthSurvey/ButlerCounty.dta"  /*read in the Butler County data set*/

105
106 /***************************************************/
107 /*Estimate the Mean Number of Adults per Household*/
108 /*s10 is the variable that stores the number of Adults*/
109
110 tab s10, missing  /*look at the data, the "missing" option shows any missing values*/
```

number of adults	Freq.	Percent	Cum.
1	**319**	**29.48**	**29.48**
2	**608**	**56.19**	**85.67**
3	**118**	**10.91**	**96.58**
4	**28**	**2.59**	**99.17**
5	**7**	**0.65**	**99.82**
6	**2**	**0.18**	**100.00**
Total	**1,082**	**100.00**	

```
111
112 /*****************/
113 /*First by "hand"*/
114
115 mean s10  /*calculate the mean*/

  Mean estimation                     Number of obs    =    1082
```

	Mean	Std. Err.	[95% Conf.	Interval]
s10	**1.892791**	**.0232371**	**1.847196**	**1.938386**

```
116 ereturn list  /*list the variables that the estimator produces*/

  scalars:
                  e(df_r) =  1081
               e(N_over) =  1
                    e(N) =  1082
                  e(k_eq) =  1
               e(k_eform) =  0

  macros:
               e(cmdline) : "mean s10"
                   e(cmd) : "mean"
                   e(vce) : "analytic"
                 e(title) : "Mean estimation"
             e(estat_cmd) : "estat_vce_only"
               e(varlist) : "s10"
               e(predict) : "_no_predict"
            e(properties) : "b V"

  matrices:
                     e(b) :  1 x 1
                     e(V) :  1 x 1
                    e(_N) :  1 x 1
    e(_N_strata_certain) :  1 x 1
     e(_N_strata_single) :  1 x 1
            e(_N_strata) :  1 x 1
                e(error) :  1 x 1

  functions:
                e(sample)

117
118 matrix list e(V)  /*report the variance*/

  symmetric e(V)[1,1]
               s10
  s10  .00053996

119 matrix define varmat = e(V)  /*save the variance matrix with name ‘varmat’*/

120 matrix list varmat

  symmetric varmat[1,1]
               s10
  s10  .00053996

121 scalar define myvar = varmat[1,1]  /*extract the variance into a scalar*/

122 display myvar
  .00053996

123 scalar define myvar = myvar*(1-e(N)/123082)  /*adjust via the fpc*/

124 display myvar
  .00053522

125 scalar define myse = sqrt(myvar)  /*find the standard error*/

126 display myse
  .02313475
```

```
127
128 matrix define meanmat = e(b)  /*extract the mean*/

129 matrix list meanmat

  symmetric meanmat[1,1]
            s10
  y1  1.8927911

130 scalar define mymean = meanmat[1,1]  /*save the mean as a scalar*/

131 display mymean
  1.8927911

132
133 /*Calculate CI*/
134 /* Stata uses a t statistic with n-1 df*/
135 scalar define tval = abs(invttail(e(N)-1, 0.025))  /*find the critical value*/

136 display tval
  1.9621609

137 scalar define lower = mymean - tval*myse   /*calculate the lower CI*/

138 scalar define upper = mymean + tval*myse   /*calculate the upper CI*/

139
140 scalar list mymean myse lower upper
      mymean =  1.8927911
        myse =  .02313475
       lower =   1.847397
       upper =  1.9381852

141
142 /***************************/
143 /*Next by built-in command */
144
145 capture gen N=123082            /*make a variable to hold the population size*/

146 svyset masterid, fpc(N)  /*make survey definition straight up with knowledge of N*/

        pweight: <none>
            VCE: linearized
    Single unit: missing
       Strata 1: <one>
           SU 1: masterid
          FPC 1: N

147 svy: mean s10            /*survey calculate the mean*/
  (running mean on estimation sample)

  Survey: Mean estimation

  Number of strata =       1          Number of obs    =    1082
  Number of PSUs   =    1082          Population size  =    1082
                                      Design df        =    1081

  ------------------------------------------------------------------------------
               |             Linearized
               |        Mean   Std. Err.     [95% Conf. Interval]
  -------------+----------------------------------------------------------------
           s10 |    1.892791   .0231347      1.847397    1.938185
  ------------------------------------------------------------------------------
```

```
148
149 scalar list mymean myse lower upper /*compare with hand calculations*/
      mymean =  1.8927911
        myse =  .02313475
       lower =   1.847397
       upper =  1.9381852

150
151
152 /***************************************************/
153 /*Estimate the proportion of households with kids */
154
155 svy: proportion kids
  (running proportion on estimation sample)

  Survey: Proportion estimation

  Number of strata =       1          Number of obs    =    1082
  Number of PSUs   =    1082          Population size  =    1082
                                      Design df        =    1081
```

	Proportion	Linearized Std. Err.	[95% Conf.	Interval]
kids				
0	**.5914972**	**.0148848**	**.5622909**	**.6207036**
1	**.4085028**	**.0148848**	**.3792964**	**.4377091**

```
156 svy: mean kids
  (running mean on estimation sample)

  Survey: Mean estimation

  Number of strata =       1          Number of obs    =    1082
  Number of PSUs   =    1082          Population size  =    1082
                                      Design df        =    1081
```

	Mean	Linearized Std. Err.	[95% Conf.	Interval]
kids	**.4085028**	**.0148848**	**.3792964**	**.4377091**

```
157
158 /***************************************************/
159 /*Estimate the total number of households with kids*/
160
161 svy: total kids  /*doesn't work*/
  (running total on estimation sample)

  Survey: Total estimation

  Number of strata =       1          Number of obs    =    1082
  Number of PSUs   =    1082          Population size  =    1082
                                      Design df        =    1081
```

	Total	Linearized Std. Err.	[95% Conf.	Interval]
kids	**442**	**16.10534**	**410.3987**	**473.6013**

```
162
163
164 /***************************************/
165 /* now do estimation with exact weights */
166
167 svyset, clear   /*clear the old definitions*/

168 gen wt = 123082/1082  /*define the weight equal to the inverse probability*/

169 svyset masterid [pweight=wt], fpc(N)  /*make survey definition*/

        pweight: wt
            VCE: linearized
    Single unit: missing
       Strata 1: <one>
           SU 1: masterid
         FPC 1: N

170
171 /*mean*/
172 svy: mean s10
  (running mean on estimation sample)

  Survey: Mean estimation

  Number of strata =       1          Number of obs    =    1082
  Number of PSUs   =    1082          Population size  =  123082
                                      Design df        =    1081

  ------------------------------------------------------------------------------
               |             Linearized
               |        Mean   Std. Err.     [95% Conf. Interval]
  -------------+----------------------------------------------------------------
           s10 |    1.892791   .0231347      1.847397    1.938185
  ------------------------------------------------------------------------------

173 scalar list mymean myse lower upper /*compare with hand calculations*/
      mymean =  1.8927911
        myse =  .02313475
       lower =   1.847397
       upper =  1.9381852

174
175 /*total*/
176 svy: total kids
  (running total on estimation sample)

  Survey: Total estimation

  Number of strata =       1          Number of obs    =    1082
  Number of PSUs   =    1082          Population size  =  123082
                                      Design df        =    1081

  ------------------------------------------------------------------------------
               |             Linearized
               |       Total   Std. Err.     [95% Conf. Interval]
  -------------+----------------------------------------------------------------
          kids |   50279.34   1832.049      46684.56    53874.11
  ------------------------------------------------------------------------------
```

```
177
178 /*********************************/
179 /* weights that are proportional */
180
181 svyset, clear  /*clear the old definitions*/

182 gen wtprop = 1/1082

183 svyset masterid [pweight=wtprop], fpc(N)

        pweight: wtprop
            VCE: linearized
    Single unit: missing
       Strata 1: <one>
           SU 1: masterid
         FPC 1: N

184
185 /*mean*/
186 svy: mean s10
  (running mean on estimation sample)

  Survey: Mean estimation

  Number of strata =       1          Number of obs    =    1082
  Number of PSUs   =    1082          Population size  =       1
                                      Design df        =    1081

  ------------------------------------------------------------------------------
               |             Linearized
               |        Mean   Std. Err.     [95% Conf. Interval]
  -------------+----------------------------------------------------------------
           s10 |    1.892791   .0231347      1.847397    1.938185
  ------------------------------------------------------------------------------

187 scalar list mymean myse lower upper /*compare with hand calculations*/
      mymean =  1.8927911
        myse =  .02313475
       lower =   1.847397
       upper =  1.9381852

188
189 /*total*/
190 svy: total kids  /*doesn't work*/
  (running total on estimation sample)

  Survey: Total estimation

  Number of strata =       1          Number of obs    =    1082
  Number of PSUs   =    1082          Population size  =       1
                                      Design df        =    1081

  ------------------------------------------------------------------------------
               |             Linearized
               |       Total   Std. Err.     [95% Conf. Interval]
  -------------+----------------------------------------------------------------
          kids |    .4085028   .0148848      .3792964    .4377091
  ------------------------------------------------------------------------------

191
```

```
192
193 /*********************************/
194 /* Ratios */
195
196 capture gen wt = 123082/1082  /*define the wt if it hasn't already*/

197 capture gen N = 123082

198 svyset masterid [pweight=wt], fpc(N)

        pweight: wt
            VCE: linearized
    Single unit: missing
       Strata 1: <one>
           SU 1: masterid
         FPC 1: N

199
200
201
202 /*Ratio of children to total*/
203 gen tothhmembers = s10 + s12

204 gen prophhkids = s12 / tothhmembers

205
206
207 /*Unit level*/
208 svy: mean prophhkids
  (running mean on estimation sample)

  Survey: Mean estimation

  Number of strata =       1          Number of obs    =    1082
  Number of PSUs   =    1082          Population size  =  123082
                                      Design df        =    1081

------------------------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  prophhkids |    .192587    .007504       .177863     .207311
------------------------------------------------------------------------------

209
210 /*population level*/
211 svy: ratio s12 / tothhmembers
  (running ratio on estimation sample)

  Survey: Ratio estimation

  Number of strata =       1          Number of obs    =    1082
  Number of PSUs   =    1082          Population size  =  123082
                                      Design df        =    1081

     _ratio_1: s12/tothhmembers

------------------------------------------------------------------------------
             |             Linearized
             |       Ratio   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _ratio_1 |   .2861624   .0092814      .2679508    .3043741
------------------------------------------------------------------------------
```

```
212
213
214 /*Ratio for unknown pop size*/
215 svyset, clear

216 capture gen wtprop = 1/1082

217 svyset masterid [pweight=wtprop] /*assume FPC is approx 1*/

        pweight: wtprop
            VCE: linearized
    Single unit: missing
       Strata 1: <one>
           SU 1: masterid
         FPC 1: <zero>

218
219 svy: ratio kidsratio: s12 / tothhmembers /*estimate ratio, store in kidsratio*/
  (running ratio on estimation sample)

  Survey: Ratio estimation

  Number of strata =       1          Number of obs    =    1082
  Number of PSUs   =    1082          Population size  =       1
                                      Design df        =    1081

      kidsratio: s12/tothhmembers

------------------------------------------------------------------------------
             |             Linearized
             |       Ratio   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   kidsratio |   .2861624   .0093225      .2678701    .3044547
------------------------------------------------------------------------------

220 lincom _b[kidsratio] * 332807  /*multiply estimate by total Population in Butler County*/

   ( 1)  332807 kidsratio = 0

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   95236.86     3102.6    30.70   0.000     89149.06    101324.7
------------------------------------------------------------------------------

221
222
223 /*Ratio for reduced variance*/
224 svy: ratio kidsratio: s12 / tothhmembers /*estimate ratio, store in kidsratio*/
  (running ratio on estimation sample)

  Survey: Ratio estimation

  Number of strata =       1          Number of obs    =    1082
  Number of PSUs   =    1082          Population size  =       1
                                      Design df        =    1081

      kidsratio: s12/tothhmembers

------------------------------------------------------------------------------
             |             Linearized
             |       Ratio   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   kidsratio |   .2861624   .0093225      .2678701    .3044547
------------------------------------------------------------------------------
```

```
225 lincom _b[kidsratio] * 2.61     /*ratio estimate, 2.61 is the average #members/household in Butler
```

(1) **2.61 kidsratio = 0**

	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	**.7468839**	**.0243318**	**30.70**	**0.000**	**.6991411**	**.7946268**

```
226 matrix ybarr = r(estimate)      /*lincom stores stuff in r() */

227 matrix Vybarr = r(se)^2

228 svy: mean s12                   /*usual estimate*/
  (running mean on estimation sample)

  Survey: Mean estimation

  Number of strata =       1          Number of obs    =    1082
  Number of PSUs   =    1082          Population size  =       1
                                      Design df        =    1081
```

	Mean	Linearized Std. Err.	[95% Conf.	Interval]
s12	**.75878**	**.0341748**	**.6917237**	**.8258364**

```
229 matrix ybar = e(b)

230 matrix Vybar = e(V)

231
232 /*display results*/
233 scalar define tval = abs(invttail(e(N)-1, 0.025))  /*t critical value*/

234 /*srs - consolidate results*/
235 matrix srs = (ybar, Vybar, ybar - tval*cholesky(Vybar), ybar + tval*cholesky(Vybar))

236 matrix colnames srs = estimate variance lowerCI upperCI

237 matrix rownames srs = srs

238 /*ratio - consolidate results*/
239 matrix ratio = (ybarr, Vybarr, ybarr - tval*cholesky(Vybarr), ybarr + tval*cholesky(Vybarr))

240 matrix colnames ratio = estimate variance lowerCI upperCI

241 matrix rownames ratio = ratio

242
243 matrix comparemeans = srs\ratio  /*combine the srs and ratio into one matrix*/

244
245 matrix list comparemeans

  comparemeans[2,4]
         estimate   variance    lowerCI    upperCI
    srs  .75878004  .00116791  .69172366  .82583642
  ratio  .74688393  .00059204  .69914107  .79462679
```

```
246
247
248 correlate s12 tothhmembers
  (obs=1082)
```

	s12	tothhm~s
s12	**1.0000**	
tothhmembers	**0.8425**	**1.0000**

```
249 scatter s12 tothhmembers, jitter(10) title(The correlation is ‘r(rho)’)

250
251 bysort tothhmembers: summarize s12
```

-> tothhmembers = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
s12	**227**	**0**	**0**	**0**	**0**

-> tothhmembers = 2

Variable	Obs	Mean	Std. Dev.	Min	Max
s12	**375**	**.144**	**.3515588**	**0**	**1**

-> tothhmembers = 3

Variable	Obs	Mean	Std. Dev.	Min	Max
s12	**199**	**.7688442**	**.6793001**	**0**	**2**

-> tothhmembers = 4

Variable	Obs	Mean	Std. Dev.	Min	Max
s12	**163**	**1.693252**	**.6697271**	**0**	**3**

-> tothhmembers = 5

Variable	Obs	Mean	Std. Dev.	Min	Max
s12	**88**	**2.545455**	**.9335274**	**0**	**4**

-> tothhmembers = 6

Variable	Obs	Mean	Std. Dev.	Min	Max
s12	**19**	**3.473684**	**.9048279**	**1**	**4**

-> tothhmembers = 7

Variable	Obs	Mean	Std. Dev.	Min	Max
s12	**7**	**2.857143**	**1.676163**	**1**	**5**

-> tothhmembers = 8

Variable	Obs	Mean	Std. Dev.	Min	Max
s12	**1**	**3**	**.**	**3**	**3**

-> tothhmembers = 9

Variable	Obs	Mean	Std. Dev.	Min	Max
s12	**1**	**7**	.	**7**	**7**

-> tothhmembers = 10

Variable	Obs	Mean	Std. Dev.	Min	Max
s12	**1**	**8**	.	**8**	**8**

-> tothhmembers = 13

Variable	Obs	Mean	Std. Dev.	Min	Max
s12	**1**	**10**	.	**10**	**10**

```
252
253
254
255 /*********************************/
256 /* Domain */
257 svyset, clear

258 capture gen wtprop = 1/1082

259 svyset masterid [pweight=wtprop] /*assume FPC is approx 1*/

        pweight: wtprop
            VCE: linearized
    Single unit: missing
       Strata 1: <one>
           SU 1: masterid
          FPC 1: <zero>

260
261 tab kids
```

kids	Freq.	Percent	Cum.
0	**640**	**59.15**	**59.15**
1	**442**	**40.85**	**100.00**
Total	**1,082**	**100.00**	

```
262
263 svy: mean s10, subpop(kids) /*only households with kids*/
  (running mean on estimation sample)

  Survey: Mean estimation

  Number of strata =       1          Number of obs    =    1082
  Number of PSUs   =    1082          Population size  =       1
                                      Subpop. no. obs  =     442
                                      Subpop. size     = .408503
                                      Design df        =    1081
```

	Mean	Linearized Std. Err.	[95% Conf.	Interval]
s10	**1.993213**	**.035859**	**1.922852**	**2.063574**

```
264 svy: mean s10, over(kids)   /*households with and without kids, separately*/
  (running mean on estimation sample)

  Survey: Mean estimation

  Number of strata =       1          Number of obs    =    1082
  Number of PSUs   =    1082          Population size  =       1
                                      Design df        =    1081

            0: kids = 0
            1: kids = 1
```

Over	Mean	Linearized Std. Err.	[95% Conf.	Interval]
s10				
0	**1.823438**	**.0301928**	**1.764194**	**1.882681**
1	**1.993213**	**.035859**	**1.922852**	**2.063574**

```
265
266 /*!!!THE WRONG WAY!!!*/
267 drop if kids==0
  (640 observations deleted)

268 svy: mean s10
  (running mean on estimation sample)

  Survey: Mean estimation

  Number of strata =       1          Number of obs    =     442
  Number of PSUs   =     442          Population size  = .408503
                                      Design df        =     441
```

	Mean	Linearized Std. Err.	[95% Conf.	Interval]
s10	**1.993213**	**.035883**	**1.92269**	**2.063736**

```
269
270 /*Get the right data back*/
271 clear

272 use "/scratch/OhioFamilyHealthSurvey/ButlerCounty.dta"

273
274 capture gen wtprop = 1/1082

275 svyset masterid [pweight=wtprop] /*assume FPC is approx 1*/

        pweight: wtprop
            VCE: linearized
    Single unit: missing
       Strata 1: <one>
           SU 1: masterid
          FPC 1: <zero>
```

```
276
277 /*difference between means in two domains*/
278 svy: mean s10, over(kids)
  (running mean on estimation sample)

  Survey: Mean estimation

  Number of strata =       1          Number of obs    =    1082
  Number of PSUs   =    1082          Population size  =       1
                                      Design df        =    1081

            0: kids = 0
            1: kids = 1
```

Over	Mean	Linearized Std. Err.	[95% Conf.	Interval]
s10				
0	**1.823438**	**.0301928**	**1.764194**	**1.882681**
1	**1.993213**	**.035859**	**1.922852**	**2.063574**

```
279 lincom [s10]1 - [s10]0

   ( 1) - [s10]0 + [s10]1 = 0
```

	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	**.1697752**	**.0468772**	**3.62**	**0.000**	**.0777946**	**.2617558**

```
280
281
282 /*********************************/
283 /* Regression   */
284 svyset, clear

285 capture gen wtprop = 1/1082

286 capture gen tothhmembers = s10 + s12

287 capture gen prophhkids = s12 / tothhmembers

288 svyset masterid [pweight=wtprop] /*assume FPC is approx 1*/

        pweight: wtprop
            VCE: linearized
    Single unit: missing
       Strata 1: <one>
           SU 1: masterid
          FPC 1: <zero>

289
290 svy: regress s12 tothhmembers
  (running regress on estimation sample)

  Survey: Linear regression

  Number of strata   =         1                 Number of obs      =      1082
  Number of PSUs     =      1082                 Population size    =         1
                                                  Design df          =      1081
                                                  F(   1,   1081)    =    881.00
                                                  Prob > F           =    0.0000
                                                  R-squared          =    0.7098
```

s12	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
tothhmembers	**.6700462**	**.0225744**	**29.68**	**0.000**	**.6257515**	**.7143408**
_cons	**−1.017895**	**.0522653**	**−19.48**	**0.000**	**−1.120448**	**−.9153422**

```
291 lincom _cons + tothhmembers * 2.61
```

(1) **2.61 tothhmembers + _cons = 0**

s12	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	**.7309254**	**.0179211**	**40.79**	**0.000**	**.6957614**	**.7660895**

```
292
293 /*store the results*/
294 matrix ybarreg = r(estimate)    /*lincom stores stuff in r() */

295 matrix Vybarreg = r(se)^2

296
297 /*display the results in a matrix*/
298 scalar define tval = abs(invttail(e(N)−1, 0.025))

299 matrix reg = (ybarreg, Vybarreg, ybarreg − tval*cholesky(Vybarreg), ybarreg + tval*cholesky(Vybarr

300 matrix colnames reg = estimate variance lowerCI upperCI

301 matrix rownames reg = regression

302
303 matrix comparemeans = srs\ratio\reg

304
305 matrix list comparemeans

  comparemeans[3,4]
              estimate   variance    lowerCI    upperCI
       srs  .75878004  .00116791  .69172366  .82583642
     ratio  .74688393  .00059204  .69914107  .79462679
  regression  .73092544  .00032117  .69576137   .7660895

306
307
308 graph twoway scatter s12 tothhmembers, jitter(10) || lfit s12 tothhmembers

309
310
311
312 /**DIFFERENT EXAMPLE −− income vs. hours worked**/
313 capture gen myincome = h85y

314 capture replace myincome = . if myincome>=999997 /*get rid of missing data*/

315 capture replace myincome = h85m * 12 if myincome = .
```

```
316
317 scatter myincome g73  /* income vs. hours worked */

318
319 /**DIFFERENT EXAMPLE -- premiums vs. income**/
320 capture gen mypremiums = b8

321 capture replace mypremiums = b8 * 52 if b8b==1

322 capture replace mypremiums = b8 * 52 / 2 if b8b==2

323 capture replace mypremiums = b8 * 12 if b8b==3

324 capture replace mypremiums = b8 * 12 *2 if b8b==4

325 capture replace mypremiums = b8 * 12 / 2 if b8b==5

326 capture replace mypremiums = b8 * 4 if b8b==6

327 capture replace mypremiums = b8 * 2 if b8b==7

328 capture replace mypremiums = . if mypremiums >= 99998

329 capture replace mypremiums = . if b8b >= 97

330
331 scatter mypremiums myincome if myincome!=. & mypremiums!=.

332
333 clear

334
335
  end of do-file

336 log close
      name:  /scratch/OhioFamilyHealthSurvey/analysislog.smcl
       log:  smcl
  log type:  14 Jul 2008, 15:10:21
 closed on:
```