```
      name:  <unnamed>
       log:  /Users/eric/Work/MB/ATCR/200/Labs/lab 2.smcl
  log type:  smcl
 opened on:   8 Jul 2018, 08:54:44

 1 .
 2 . * Read in the lab1 data posted on the class web site.  These data come from a study of voluntary counseling and te
   > sting (VCT) for HIV at Mulago Hospital in Kampala, Uganda.
 3 .
 4 . use vct_revised.dta, clear

 5 .
 6 . * Briefly remind yourself of what is in the dataset.  Stata lets us look at a spreadsheet of the data:
 7 .
 8 . browse

 9 .
10 . * First see what variables are in the dataset.  Check how many observations and variables the dataset includes. */
11 .
12 . describe

Contains data from vct_revised.dta
  obs:         3,389
 vars:            15                          6 Jul 2018 08:15
 size:        77,947
-------------------------------------------------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------------------------------------------------
age             double  %12.0g                Age
age_cat         byte    %9.0g      age_cat    Age
alc_any         byte    %9.0g      alc_any    Any alcohol use
alc_unh         byte    %9.0g      alc_unh    Unhealthy alcohol use
auditc          byte    %12.0g                Audit score, last 3 months
cd4             int     %11.0g                CD4 count
cd4_cat         byte    %11.0g     cd4_cat    CD4 count
female          byte    %9.0g      yesno      female sex
hiv             byte    %12.0g     hiv        HIV status
hungry          byte    %12.0g     hungry     How often household members go hungry
lastalc         byte    %20.0g     lastalc    Last time used alcohol
n_adults        byte    %12.0g                # adults in household
n_child         byte    %12.0g                # people <18yo in household
n_support       byte    %12.0g                # people supported financially
religion        byte    %10.0g     religion   Religion
-------------------------------------------------------------------------------------------------------------------------
Sorted by: female

13 .
14 . * Now get brief summary statistics for all the variables in the dataset. The full command we'll use is summarize,
   > but we can shorten that to sum to save typing.
15 .
16 . sum

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         age |      3,387    31.72456    9.858456       17.5      80.07
     age_cat |      3,387    2.653971    .9791769          1          5
     alc_any |      3,244    .2530826    .4348449          0          1
     alc_unh |      3,244    .0786067    .2691653          0          1
      auditc |      3,244    .6390259    1.483416          0         12
-------------+---------------------------------------------------------
         cd4 |        999    329.2332    266.1177          1       1932
     cd4_cat |        999    1.618619    .9377913          0          3
      female |      3,389    .5417527    .4983272          0          1
         hiv |      3,389    .2959575    .4565393          0          1
      hungry |      3,321    3.603132    .7817434          1          4
-------------+---------------------------------------------------------
```

lastalc	3,377	.9227125	.9101892	0	2
n_adults	3,384	1.664598	1.737786	0	12
n_child	3,382	1.850089	1.924868	0	12
n_support	3,372	3.907177	3.227549	0	20
religion	3,387	3.781518	1.994997	1	8

```
17 .
18 . * Note how many people had CD4 counts.  Why is this less than the number of observations?
19 .
20 . * To get more information about continuous variables, add the detail option to the summarize command.
21 .
22 . sum age cd4, detail
```

Age

Percentiles		Smallest		
1%	17.97	17.5		
5%	19.66	17.5		
10%	21.07	17.53	Obs	3,387
25%	24.22	17.56	Sum of Wgt.	3,387
50%	29.51		Mean	31.72456
		Largest	Std. Dev.	9.858456
75%	37.66	75.22		
90%	45.91	75.36	Variance	97.18915
95%	50.39	77.84	Skewness	1.029554
99%	60.04	80.07	Kurtosis	3.972486

CD4 count

Percentiles		Smallest		
1%	5	1		
5%	14	2		
10%	36	2	Obs	999
25%	130	2	Sum of Wgt.	999
50%	283		Mean	329.2332
		Largest	Std. Dev.	266.1177
75%	463	1461		
90%	659	1601	Variance	70818.64
95%	866	1804	Skewness	1.444705
99%	1182	1932	Kurtosis	6.518639

```
23 .
24 . ********************************************************************************
25 .
26 . * Exploratory analysis using descriptive statistics
27 .
28 . * The tabstat command also gives easy summaries. Here we use the stat() option to obtain the number of observation
   > s, mean, SD, minimum, 25th percentile, median, 75th percentile, and maximum. The format option controls rounding o
   > f the statistics.
29 .
30 . tabstat age cd4, stat(n mean sd min p25 p50 p75 max) format(%8.3g)
```

stats	age	cd4
N	3387	999
mean	31.7	329
sd	9.86	266
min	17.5	1
p25	24.2	130
p50	29.5	283
p75	37.7	463
max	80.1	1932

```
31 .
32 . * Adding the option col(statistics) transposes the results. Also, we can use the line break marker /// to tell sta
```

```
   > ta that command continues after a hard return. This is handy if you want to use a relatively narrow do-file editor
   >  window. (Note that you can't use /// in the command window.)
   >
33 . tabstat age cd4, stat(n mean sd min p25 p50 p75 max) ///
   >         col(statistics) format(%8.3g)
```

variable	N	mean	sd	min	p25	p50	p75
age	3387	31.7	9.86	17.5	24.2	29.5	37.7
cd4	999	329	266	1	130	283	463

variable	max
age	80.1
cd4	1932

```
34 .
35 . * Add the by() option to see the statistics stratified by HIV status.
36 . tabstat age cd4, by(hiv) stat(n mean sd min p25 p50 p75 max) ///
   >         col(statistics) format(%8.3g)

   Summary for variables: age cd4
        by categories of: hiv (HIV status)
```

hiv	N	mean	sd	min	p25	p50	p75	max
HIV-	2384	31.3	10.5	17.5	23.3	28.3	37.5	80.1
	0	.	.	.	.	.	.	.
HIV+	1003	32.6	8.06	17.5	26.8	31.2	38	70.2
	999	329	266	1	130	283	463	1932
Total	3387	31.7	9.86	17.5	24.2	29.5	37.7	80.1
	999	329	266	1	130	283	463	1932

```
37 .
38 . * The tabstat command can only be used to stratify by one by variable at a time; tostratify by >1 variables, use t
   > he table command.
39 .
40 . table hiv female hungry, by(religion) c(mean age) format(%8.1f)
```

Religion and HIV status	How often household members go hungry and female sex							
	often		someti		seldom		never	
	no	yes	no	yes	no	yes	no	yes
Protestant								
HIV-	35.2	32.3	34.0	34.9	29.9	32.6	31.9	31.8
HIV+	37.7	32.5	33.4	31.0	39.6	34.2	35.5	31.4
Moslem								
HIV-	33.6	38.8	29.1	34.1	32.6	33.4	31.3	31.2
HIV+	33.5	36.3	33.8	29.8	33.8	31.1	34.1	30.4
5								
HIV-	24.8	34.2	30.1	28.4	26.2	32.4	29.5	30.3
HIV+	38.1	27.1	35.9	30.6	31.8	32.6	35.8	30.8
6								
HIV-	37.4	36.3	29.9	33.9	27.8	31.1	31.0	31.2
HIV+	44.5	33.3	35.7	33.0	38.6	31.6	34.9	30.1
Other								
HIV-			28.0	28.1	26.1	23.7	29.5	31.4
HIV+		40.0	38.6		46.3	37.1	34.9	32.7

```
-----------+-----------------------------------------------------
```

```
41 .
42 . * To see how categorical variables relate to each other, use the tab command.
43 . tab age_cat hiv, row
```

Key
frequency *row percentage*

Age	HIV status HIV-	HIV+	Total
<20	**191** **89.67**	**22** **10.33**	**213** **100.00**
20 to <30	**1,164** **73.53**	**419** **26.47**	**1,583** **100.00**
30 to <40	**563** **59.64**	**381** **40.36**	**944** **100.00**
40 to <50	**306** **66.96**	**151** **33.04**	**457** **100.00**
50+	**160** **84.21**	**30** **15.79**	**190** **100.00**
Total	**2,384** **70.39**	**1,003** **29.61**	**3,387** **100.00**

```
44 . tab female hiv, row
```

Key
frequency *row percentage*

female sex	HIV status HIV-	HIV+	Total
no	**1,175** **75.66**	**378** **24.34**	**1,553** **100.00**
yes	**1,211** **65.96**	**625** **34.04**	**1,836** **100.00**
Total	**2,386** **70.40**	**1,003** **29.60**	**3,389** **100.00**

```
45 . tab hungry hiv, row
```

Key
frequency *row percentage*

```
How often |
household |
```

members go hungry	HIV status HIV-	HIV+	Total
often	49	30	79
	62.03	37.97	100.00
sometimes	246	134	380
	64.74	35.26	100.00
seldom	209	112	321
	65.11	34.89	100.00
never	1,841	700	2,541
	72.45	27.55	100.00
Total	2,345	976	3,321
	70.61	29.39	100.00

```
46 . tab alc_any hiv, row
```

Key
frequency *row percentage*

Any alcohol use	HIV status HIV-	HIV+	Total
no	1,753	670	2,423
	72.35	27.65	100.00
yes	540	281	821
	65.77	34.23	100.00
Total	2,293	951	3,244
	70.68	29.32	100.00

```
47 .
48 . * Taking advantage of the 0-1 coding of the HIV indicator, we can use the table command get HIV prevalence by seve
   > ral other variables at once; here we do just age and sex
49 .
50 . table age_cat female, c(mean hiv) format(%8.2f)
```

Age	female sex no	yes
<20	0.01	0.18
20 to <30	0.16	0.35
30 to <40	0.39	0.42
40 to <50	0.38	0.29
50+	0.19	0.12

```
51 .
52 . ********************************************************************************
53 .
54 . * We can also use the Stata means command to calculate confidence intervals for the means of continuous variables,
   >  both overall and within strata.
55 .
56 . mean age, cformat(%8.2f)

   Mean estimation                 Number of obs    =      3,387
```

	Mean	Std. Err.	[95% Conf.	Interval]
age	**31.72**	**0.17**	**31.39**	**32.06**

```
57 . mean age, over(female hiv) cformat(%8.2f)

Mean estimation                   Number of obs    =      3,387

          Over: female hiv
     _subpop_1: no HIV-
     _subpop_2: no HIV+
     _subpop_3: yes HIV-
     _subpop_4: yes HIV+
```

Over	Mean	Std. Err.	[95% Conf.	Interval]
age				
_subpop_1	**31.00**	**0.31**	**30.39**	**31.60**
_subpop_2	**35.16**	**0.39**	**34.39**	**35.94**
_subpop_3	**31.67**	**0.30**	**31.08**	**32.25**
_subpop_4	**31.13**	**0.32**	**30.51**	**31.75**

```
58 .
59 . * To get, say, 90% confidence intervals, we can use the level option
60 . mean age, over(female hiv) level(90) cformat(%8.2f)

Mean estimation                   Number of obs    =      3,387

          Over: female hiv
     _subpop_1: no HIV-
     _subpop_2: no HIV+
     _subpop_3: yes HIV-
     _subpop_4: yes HIV+
```

Over	Mean	Std. Err.	[90% Conf.	Interval]
age				
_subpop_1	**31.00**	**0.31**	**30.49**	**31.50**
_subpop_2	**35.16**	**0.39**	**34.51**	**35.81**
_subpop_3	**31.67**	**0.30**	**31.17**	**32.16**
_subpop_4	**31.13**	**0.32**	**30.61**	**31.65**

```
61 .
62 . * The means command could also be used to get means and confidence intervals for binary variables coded 0-1, but a
   >  better approach is the ci prop command, with the exact option.  This avoids confidence limits less than 0 or more
   >  than 1.
63 .
64 . ci prop hiv, exact
```

				Binomial Exact	
Variable	Obs	Proportion	Std. Err.	[95% Conf.	Interval]
hiv	**3,389**	**.2959575**	**.0078411**	**.2806231**	**.3116447**

```
65 . ci prop hiv, exact level(90)
```

				Binomial Exact	
Variable	Obs	Proportion	Std. Err.	[90% Conf.	Interval]
hiv	**3,389**	**.2959575**	**.0078411**	**.2830462**	**.3091298**

```
66 .
67 . * The ci command does not allow stratification, but we can get around that using the bysort prefix:
68 .
69 . bysort female: ci prop hiv, exact

-------------------------------------------------------------------------------
-> female = no

                                                    ── Binomial Exact ──
    Variable |        Obs  Proportion    Std. Err.       [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hiv |      1,553    .2433999    .0108895        .2222362    .2655381
-------------------------------------------------------------------------------
-> female = yes

                                                    ── Binomial Exact ──
    Variable |        Obs  Proportion    Std. Err.       [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hiv |      1,836    .3404139    .0110587        .3187364    .3626026

70 .
71 . * The ci command also works for continous and Poisson-distributed count varibles
72 . ci means age

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |      3,387    31.72456    .1693953        31.39243    32.05669

73 . ci means n_adults, poisson

                                                    ── Poisson  Exact ──
    Variable |   Exposure        Mean    Std. Err.       [95% Conf. Interval]
-------------+----------------------------------------------------------------
    n_adults |        3384    1.664598    .0221789        1.621409    1.708646

74 .
75 . ********************************************************************************
76 .
77 . * Exploratory data analysis using graphical means.  First make a histogram of the age distribution.  We can use th
   > e name option so that multiple graphs can be shown in the viewer.  The replace option is needed to prevent Stata f
   > rom stopping if the graph already exists -- again, this is useful if you are rerunning a do-file to debug it. */
78 .
79 . histogram age, fcolor(blue) name(hist_age_00, replace)
   (bin=35, start=17.5, width=1.7877143)

80 .
81 . /* Now add titles to the histogram. */
82 .
83 . histogram age, fcolor(blue) title(Histogram of age) ///
   >         xtitle(Age) name(hist_age_01, replace)
   (bin=35, start=17.5, width=1.7877143)

84 .
85 . /* By default, the Y-axis shows the "density", which sums to one for all bars. Instead we can get a histogram of f
   > requencies. */
86 .
87 . histogram age, fcolor(blue) frequency ///
   >         title(Histogram of age) xtitle(Age) name(hist_age_02, replace)
   (bin=35, start=17.5, width=1.7877143)

88 .
89 . /* Now combine the two histograms into a single plot. Note that he histograms are identical in shape, and both sho
   > w that age is somewhat skewed to the right. */
90 .
91 . graph combine hist_age_01 hist_age_02, name(hist_age_combined, replace)

92 .
```

```
 93 . * To save this as a PDF file, we can use the following command.  Other options include saving as png, wmf, tif, ep
    > s (encapsulated postscript), and svg (scalable vector graphics), which some journals prefer.  See help graph expor
    > t for more information.
 94 .
 95 . graph export hist_age_combined.pdf, replace
    (file /Users/eric/Work/MB/ATCR/200/Labs/hist_age_combined.pdf written in PDF format)

 96 .
 97 . /* Now we will change the number of bars in the histogram, leaving Stata to
    > decide on their widths.
    >
    > We will do this using "loops." Specifically, the program will loop over the 4 values of nb in the first line, past
    > ing those values as so-called local variables `nb' into the histogram command in several places.  Learning to use
    > loops can make programming a lot less tedious. You will need to run the four lines of code defining the loop toget
    > her. */
 98 .
 99 . foreach nb in 5 10 25 50 {
      2.         histogram age, fcolor(blue) percent  bin(`nb') ///
    >                 title(`nb' bins) xtitle(Age) name(hanb`nb', replace)
      3.         }
    (bin=5, start=17.5, width=12.514)
    (bin=10, start=17.5, width=6.257)
    (bin=25, start=17.5, width=2.5028)
    (bin=50, start=17.5, width=1.2514)

100 . graph combine hanb5 hanb10 hanb25 hanb50, name(hist_age_bins, replace)

101 .
102 . /* As an alternative, we can vary the width of the bins, again using a loop. */
103 .
104 . foreach bw in 1 2 5 10 {
      2.         histogram age, fcolor(blue) percent width(`bw') ///
    >                 title(bin width `bw') xtitle(Age) name(habw`bw', replace)
      3.         }
    (bin=63, start=17.5, width=1)
    (bin=32, start=17.5, width=2)
    (bin=13, start=17.5, width=5)
    (bin=7, start=17.5, width=10)

105 . graph combine habw1 habw2 habw5 habw10, name(hist_age_widths, replace)

106 .
107 . /* Box plots are also useful for assessing the distribution of continuousvariables.  The box runs from the 25th to
    >  the 75th percentile, and includes a line at the median.  The so-called whiskers extend 1.5 IQRs from the bottom a
    > nd top of the box, or to the most extreme value, whichever is closer to the box. Any outliers more than 1.5 IQRs f
    > rom the box are plotted separately. */
108 .
109 . graph box age, name(box_age, replace)

110 .
111 . /* The boxplot also shows that age is somewhat right skewed, as we saw in the histograms.
    >
    > We can stratify the boxplot by up to three other factors. Note that the value labels we added to sex and hiv make
    > this plot easier to understand */
112 .
113 . graph box age, over(female) over(hiv) name(box_age_stratified, replace)

114 .
115 . /* Now we will use scatter plots to look at the association of a continuous outcome with a continuous predictor. *
    > /
116 .
117 . scatter n_support age, name(scatter_age, replace)

118 .
119 . /* Now we add a Lowess smooth through the data. Essentially this is a regression of n_support on age, without maki
    > ng the assumption that the regression line is linear. The bwidth option controls the smoothness of the line.  The
    > default is 0.8, which means that Stata uses the nearest 80% of the data to determine the vertical location of the
    > smooth at each value of the independent variable. The lineopts() option allows us to control the thickness of the
```

```
    > line, and could also be used to reset its color and whether the line is solid, dashed, etc.
    >
    > We will again using loops to explore the bandwidth. Here, we have to define a second local variable for the bandwi
    > dth, based on looping variable pct, because the plot name can't include decimals. Note that the first line of the
    > loop uses a somewhat different syntax than the previous loops.
    > */
120 .
121 . foreach pct of numlist 80(-20)20 10 1 {
      2.         local bw = `pct'/100
      3.         lowess n_support age, ///
    >                bwidth(`bw') msize(tiny) lineopts(lwidth(thick)) title("") ///
    >                name(bw`pct', replace)
      4.         }

122 . graph combine bw80 bw60 bw40 bw20 bw10 bw1, name(lowess_combined, replace)

123 .
124 . /* Reducing the bandwith imposes less smoothness, but clearly the smallest  bandwidth is going too far!
    >
    > We can also focus on a certain age range by adding an if statement. */
125 .
126 . lowess n_support age if age<60, ///
    >         msize(tiny) lineopts(lwidth(medthick)) title("") ///
    >         name(lowess_restricted, replace)

127 .
128 . /* Finally, we can use tabstat and box plots to look at the distribution of a continuous outcome when the independ
    > ent variable is categorical. */
129 .
130 . tabstat n_support, by(hungry) stat(n mean sd min p25 p50 p75 max) format(%8.3g)

    Summary for variables: n_support
         by categories of: hungry (How often household members go hungry)

       hungry |         N      mean        sd       min       p25       p50       p75
    ----------+--------------------------------------------------------------------------
        often |        77      4.38      3.36         0         1         4         6
    sometimes |       378      4.18      3.18         0         2         4         6
       seldom |       319       4.2      3.34         0         2         4         6
        never |      2530      3.81      3.21         0         1         3         5
    ----------+--------------------------------------------------------------------------
        Total |      3304      3.91      3.22         0         1         3         5
    ----------------------------------------------------------------------------------

       hungry |       max
    ----------+----------
        often |        13
    sometimes |        16
       seldom |        20
        never |        20
    ----------+----------
        Total |        20
    ------------------------

131 .
132 . graph box n_support, over(hungry) name(ins_alc, replace)

133 .
134 . /* Now close the log file.*/
135 .
136 . log close
          name:  <unnamed>
           log:  /Users/eric/Work/MB/ATCR/200/Labs/lab 2.smcl
      log type:  smcl
     closed on:   8 Jul 2018, 08:55:09
    ------------------------------------------------------------------------------------------------------------------------
```