Unit 4 – Introduction to Stata *version 16*
Homework 2 of 2
SOLUTIONS

Before you begin
Download from the course website the two Stata data sets introstata1.dta and introstata2.dta
Place these on your desktop

__1. Data Screening Plan

Variable	Type	Coding	Data Screens (Stata command)
ALL			1. Review data structure (describe) 2. Review data codebook (codebook) 3. Review label code names (label list)
studyid	Character		1. Identify missing values (list) 2. Identify duplicates (duplicates)
city	Numeric	1=yes, 0=no	1. Identify missing values (list) 2. Identify out of range values (list) 3. Screen using tabulation (tabulate) 4. Check for inconsistency of value with other variables (list): Does not live in metro but does live in city center?
dues	Continuous		1. Perform range check (tabstat) 2. Identify missing values (list) 3 Screen using inspect (inspect) 4. Check for inconsistency of value with other variables (list): Paid union dues last week but is not a member of the union?

```
. *  Read in the data introstata1.dta from the toolbar at top
FILE > OPEN followed by browse to find and select


. *  ----- 2) Variable labels
. label variable studyid "Study ID"
. label variable city "Lives in Metro Area"
. label variable dues "Dues paid last week"



. * -----  3) Discrete variable value labels
. label define cityf 0 "No" 1 "Yes"
. label values city cityf
```

```
. * -----  4) Data Screening
. * Variables:  ALL
. describe

Contains data from /Users/cbigelow/Desktop/1.  Teaching/web690c/data/introstata1.dta
  obs:         1,000                          Working Women Survey
 vars:             3                          19 Oct 2015 12:14
 size:        12,000                          (_dta has notes)
-------------------------------------------------------------------------------
storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
studyid         float   %9.0g                 Study ID
city            float   %9.0g      cityf      Lives in Metro Area
dues            float   %9.0g                 Dues paid last week
-------------------------------------------------------------------------------
Sorted by:

. codebook

-------------------------------------------------------------------------------
studyid
Study ID
-------------------------------------------------------------------------------

                  type:  numeric (float)

                 range:  [1,5159]                     units:  1
         unique values:  1,000                    missing .:  0/1,000

                  mean:   2590.68
             std. dev:   1510.65

           percentiles:        10%       25%       50%       75%       90%
                             464.5      1258      2606      3932      4643

-------------------------------------------------------------------------------
city
Lives in Metro Area
-------------------------------------------------------------------------------

                  type:  numeric (float)
                 label:  cityf

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  0/1,000

            tabulation:  Freq.   Numeric  Label
                           296         0  No
                           704         1  Yes

-------------------------------------------------------------------------------
```

```
dues
Dues paid last week
-------------------------------------------------------------------------------

                 type:  numeric (float)

                range:  [0,29]                       units:  1
        unique values:  30                       missing .:  3/1,000

                 mean:   5.47342
            std. dev:   8.95317

          percentiles:        10%       25%       50%       75%       90%
                               0         0         0        10        22
```

```
. * ---  Variable:  studyid
. * ---  a)  Check for missing values
. list if studyid==.
```

Note: Stata returns nothing here because there are no missing values.

```
. * ---  b) Check for duplicate study ids
. duplicates report studyid

Duplicates in terms of studyid

--------------------------------------
   copies | observations       surplus
----------+---------------------------
        1 |         1000             0
--------------------------------------

. duplicates tag, generate(dup)

Duplicates in terms of all variables

. list studyid if dup>0
```

Note: Stata returns nothing here because there are no duplicates.

```
. * --- Variable:  city
. * ---  a) Check for missing values
. list studyid if city==.
```

Note: Stata returns nothing here because there are no missing values.

```
. * ---  b)  Check for out of range values
. list studyid city if (city<0 | city>1)
```

Note: Stata returns nothing here because there is nothing out of range.

```
. * --- c) screen
. tabulate city, missing
```

```
   Lives in |
 Metro Area |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |        296       29.60       29.60
        Yes |        704       70.40      100.00
------------+-----------------------------------
      Total |      1,000      100.00
```

```
. * --- Variable:  dues
. * ---  a) Check for missing values
. list studyid if dues==.

     +---------+
     | studyid |
     |---------|
113. |    5007 |
394. |    4632 |
765. |    4532 |
     +---------+
```

```
. * --- b) check distribution overall
. summarize dues, detail

                    Dues paid last week
-------------------------------------------------------------
      Percentiles      Smallest
 1%             0              0
 5%             0              0
10%             0              0       Obs                 997
25%             0              0       Sum of Wgt.         997

50%             0                      Mean            5.47342
                        Largest        Std. Dev.      8.953166
75%            10             29
90%            22             29       Variance       80.15918
95%            26             29       Skewness       1.383192
99%            29             29       Kurtosis       3.435988
```

```
. * --- b)  Check distribution separately for groups defined by city
. sort city
. tabstat dues, by(city) statistics(n mean sd min q max) missing

Summary for variables: dues
     by categories of: city (Lives in Metro Area)

  city |         N      mean        sd       min       p25       p50       p75       max
-------+--------------------------------------------------------------------------------
    No |       295   4.99661  8.663591         0         0         0         9        29
   Yes |       702  5.673789  9.070677         0         0         0        11        29
-------+--------------------------------------------------------------------------------
 Total |       997   5.47342  8.953166         0         0         0        10        29
--------------------------------------------------------------------------------------
```

Part A
DO FILE

```
*  -----   Input data
use "/Users/cbigelow/Desktop/1.  Teaching/web690c/data/introstata1.dta"


*  ----- 2) Variable labels
label variable studyid "Study ID"
label variable city "Lives in Metro Area"
label variable dues "Dues paid last week"


* -----  3) Discrete variable value labels
label define cityf 0 "No" 1 "Yes"
label values city cityf


* -----  4) Data Screening
* Variables:  ALL
describe
codebook


* ---  Variable:  studyid
* ---  a)  Check for missing values
list if studyid==.
* ---  b) Check for duplicate study ids
duplicates report studyid
duplicates tag, generate(dup)
list studyid if dup>0


* --- Variable:  city
* ---  a) Check for missing values
list studyid if city==.
* ---  b)  Check for out of range values
list studyid city if (city<0 | city>1)
* --- c) screen
tabulate city, missing


* --- Variable:  dues
* ---  a) Check for missing values
list studyid if dues==.
* --- b) check distribution overall
summarize dues, detail missing
summarize dues, detail
* --- b)  Check distribution separately for groups defined by city
sort city
tabstat dues, by(city) statistics(n mean sd min q max) missing
```

```
. *****  -------------  PART B   ---------------------  *****

. *  Clear the current workspace and read in the data introstata2.dta
. clear
FILE > OPEN followed by browse to find and select



. * -----  8)  Create lnlead
. generate lnlead=ln(lead)if !missing(lead)
(286 missing values generated)
. label variable lnlead "Natural Logarithm (lead)"



. * -----  9)  Create bmi
. generate bmi=(100^2)*(weight/(height*height))
. label variable bmi "Body Mass Index (kg/m^2)"
. replace bmi=. if weight==.
(0 real changes made)
. replace bmi=. If height==.
(0 real changes made)


. * -----  10)  Create bmi_group
. generate bmi_group=bmi
. recode bmi_group (min/18.49=1) (18.5/24.99=2) (25.0/29.99=3) (30/max=4)
(bmi_group: 500 changes made)

. replace bmi_group=. if bmi==.
(0 real changes made)


. * -----  11) Discrete variable value labels and variable label
. label define bmif 1 "Underweight" 2 "Normal" 3 "Overweight" 4 "Obese"
. label values bmi_group bmif
. label variable bmi_group "BMI, Grouped"



. * -----  12) Create racer, a reverse coding of race
. * -----   a) Preliminary:  Look at codings of source variable race
. numlabel, add
. tabulate race

   1=white, |
   2=black, |
    3=other |      Freq.     Percent        Cum.
------------+-----------------------------------
   1. White |        444       88.80       88.80
   2. Black |         48        9.60       98.40
   3. Other |          8        1.60      100.00
------------+-----------------------------------
      Total |        500      100.00
```

```
. generate racer=race
. recode racer (1=3) (2=2) (3=1)
(racer: 452 changes made)
. replace racer=. if race==.
(0 real changes made)

. label define racerf 1 "Other" 2 "Black" 3 "White"
. label values racer racerf
. label variable racer "Race, reverse coded"
. numlabel, add
. tab2 race racer

-> tabulation of race by racer

  1=white, |
  2=black, |       Race, reverse coded
   3=other |  1. Other   2. Black   3. White |     Total
-----------+---------------------------------+----------
  1. White |         0          0       444 |       444
  2. Black |         0         48         0 |        48
  3. Other |         8          0         0 |         8
-----------+---------------------------------+----------
     Total |         8         48       444 |       500


. save "/Users/cbigelow/Desktop/1.  Teaching/web690c/data/mystatadata2.dta"
file /Users/cbigelow/Desktop/1.  Teaching/web690c/data/mystatadata2.dta saved

. log close
```