```
#### Biost 517: Applied Biostatistics I
#### Emerson, Fall 2005

#### Annotated Stata Log File: Homework #2
#### October 13, 2005

#### In this file I give the Stata commands I used to produce
#### the key to Homework #2. In order to properly format
#### a table useful to casual readers, I cut and pasted some
#### of the output into Excel.

#### Comments edited into the log file produced by Stata are
#### on the lines that start with the four '#' signs and are
#### printed in italics.

#### The Stata commands are put in bold face.

#### Stata output is displayed in regular typeface in blue.


#### Read in data: The infile command was typed all on one line
. infile Amales Afemales Bmales Bfemales Cmales Cfemales using hw2salary.txt
'AMales' cannot be read as a number for Amales[1]
'AFemales' cannot be read as a number for Afemales[1]
'BMales' cannot be read as a number for Bmales[1]
'BFemales' cannot be read as a number for Bfemales[1]
'CMales' cannot be read as a number for Cmales[1]
'CFemales' cannot be read as a number for Cfemales[1]
(1001 observations read)

#### The above messages told me that missing data was created for the very first
#### line of the dataset (the "[1]" tells me it is the first line) for several
#### variables. This is due to the text labels for each column in the data file.
#### To verify that I do not want this line, I ask Stata to print out the first
#### row of data.
. list in 1
        Amales   Afemales     Bmales   Bfemales     Cmales   Cfemales
  1.         .          .          .          .          .          .

#### We see that the first line has 6 missing values (indicated by ".") for the
#### numeric variables. I want to drop this row from the dataset.
. drop in 1
(1 observation deleted)

. format Amales Afemales Bmales Bfemales Cmales Cfemales %9.0f
```

```
#### I intend to just execute two commands to get all the numeric descriptive
#### statistics: I will get the number of nonmissing observations, mean, SD,
#### min, and max using "summ", and I will get the 25th, 50th, and 75th percentiles
#### using "centile". Because I can use the mean of a binary variable to get the
#### proportion, I want to create the dichotomized variables first.
. g AMslry50 = 1
. replace AMslry50 = 0 if Amales < 50000/12
(494 real changes made)

. g AMslry75 = 1
. replace AMslry75 = 0 if Amales < 75000/12
(909 real changes made)

. g AFslry50 = 1
. replace AFslry50 = 0 if Afemales < 50000/12
(623 real changes made)

. g AFslry75 = 1
. replace AFslry75 = 0 if Afemales < 75000/12
(976 real changes made)

. g BMslry50 = 1
. replace BMslry50 = 0 if Bmales < 50000/12
(544 real changes made)

. g BMslry75 = 1
. replace BMslry75 = 0 if Bmales < 75000/12
(919 real changes made)

. g BFslry50 = 1
. replace BFslry50 = 0 if Bfemales < 50000/12
(479 real changes made)

. g BFslry75 = 1
. replace BFslry75 = 0 if Bfemales < 75000/12
(953 real changes made)

. g CMslry50 = 1
. replace CMslry50 = 0 if Cmales < 50000/12
(522 real changes made)

. g CMslry75 = 1
. replace CMslry75 = 0 if Cmales < 75000/12
(898 real changes made)

. g CFslry50 = 1
. replace CFslry50 = 0 if Cfemales < 50000/12
(519 real changes made)

. g CFslry75 = 1
. replace CFslry75 = 0 if Cfemales < 75000/12
(969 real changes made)
```

```
#### Making the histograms. I saved them as "*.wmf" to insert them in the
#### MS-Word document. Cut and Paste also works.
. hist Amales, bin(20)
(bin=20, start=1600, width=377.7)

. hist Afemales, bin(20)
(bin=20, start=1418, width=406.9)

. hist Bmales, bin(20)
(bin=20, start=1392, width=450.95)

. hist Bfemales, bin(20)
(bin=20, start=1563, width=373.1)

. hist Cmales, bin(20)
(bin=20, start=1623, width=410.5)

. hist Cfemales, bin(20)
(bin=20, start=1711, width=333.5)

#### Stata does not make reading modes from the histogram particularly easy.
#### I could approximate the modes using a ruler to measure:
####       - the distance between two ticks on the x-axis
####       - the distance between the histogram peak and one of the ticks
#### and then compute the value of the histogram peak by interpolation.
#### For instance, on the histogram for Amales, I measured a distance of
#### 6.9 between the tick for 2000 and the tick for 10000. I also measured
#### a distance of 1.6 between the tick for 2000 and the midpoint of the
#### peak of the histogram. From this I could estimate the mode of the
#### histogram as
####     2000 + (1.6 / 6.9) * (10000 – 2000) = 3855
####
#### The above might (rightly) seem a colossal bother. We could use
#### the Stata facility for computing and saving density estimates (I did
#### not expect you to know how to do this) to obtain modes from the
#### densities. I show you how to do this below. (The difficulty in obtaining
#### modes from statistical packages is testimony to the infrequency with
#### which precise estimates of the mode are used descriptively.)
#### "kdensity" calculates the density, and the g() option specifies that
#### new variables a and d should store the value of the salary and the
#### density, respectively. I then sort the data in descending order of the
#### density, so the first case will have the highest density estimate.
. kdensity Amales, g(a d)
. gsort -d
. list d a in 1

     +-----------------------+
     |         d           a |
     |-----------------------|
  1. | .00030027   3962.1359 |
     +-----------------------+

. drop a d
. kdensity Afemales, g(a d)
. gsort -d
. list d a in 1

     +-----------------------+
     |         d           a |
```

```
     |-----------------------|
  1. | .00046896   4004.4127 |
     +-----------------------+

. drop a d
. kdensity Bmales, g(a d)
. gsort -d
. list d a in 1

     +-----------------------+
     |         d           a |
     |-----------------------|
  1. | .00033749   3850.1005 |
     +-----------------------+

. drop a d
. kdensity Bfemales, g(a d)
. gsort -d
. list d a in 1

     +----------------------+
     |         d          a |
     |----------------------|
  1. | .00047357   4250.612 |
     +----------------------+

. drop a d
. kdensity Cmales, g(a d)
. gsort -d
. list d a in 1

     +-----------------------+
     |         d           a |
     |-----------------------|
  1. | .00030564   3652.7092 |
     +-----------------------+

. drop a d
. kdensity Cfemales, g(a d)
. gsort -d
. list d a in 1

     +-----------------------+
     |         d           a |
     |-----------------------|
  1. | .00044875   4105.7614 |
     +-----------------------+


#### Generating the numeric descriptive statistics.
#### tabstat will get me most of them.
. tabstat Amales Afemales Bmales Bfemales Cmales Cfemales, format stat(n mean sd
min p25 p50 p75 max r iqr) col(stat)
    variable |         N      mean        sd       min
-------------+--------------------------------------------
      Amales |      1000      4366      1332      1600
    Afemales |      1000      3968       987      1418
```

```
      Bmales |      1000      4247      1348      1392
    Bfemales |      1000      4271       993      1563
      Cmales |      1000      4362      1399      1623
    Cfemales |      1000      4183      1004      1711
-------------------------------------------------------

    variable |       p25       p50       p75       max
-------------+-----------------------------------------
      Amales |      3361      4174      5160      9154
    Afemales |      3313      3933      4532      9556
      Bmales |      3298      4048      4951     10411
    Bfemales |      3608      4220      4812      9025
      Cmales |      3302      4105      5242      9833
    Cfemales |      3559      4112      4806      8381
-------------------------------------------------------

    variable |     range       iqr
-------------+--------------------
      Amales |      7554      1799
    Afemales |      8138      1219
      Bmales |      9019      1654
    Bfemales |      7462      1204
      Cmales |      8210      1940
    Cfemales |      6670      1247
-----------------------------------

#### means can be used to get the geometric means.
. means Amales Afemales Bmales Bfemales Cmales Cfemales, format
    Variable |    Type        Obs        Mean       [95% Conf. Interval]
-------------+-------------------------------------------------------------
      Amales | Arithmetic    1000    4365.683        4283.022   4448.343
             |  Geometric    1000    4169.815        4091.703   4249.418
             |   Harmonic    1000    3979.939        3904.255   4058.616
-------------+-------------------------------------------------------------
    Afemales | Arithmetic    1000    3967.616        3906.342    4028.89
             |  Geometric    1000    3844.273        3783.765   3905.749
             |   Harmonic    1000    3715.678        3652.508    3781.07
-------------+-------------------------------------------------------------
      Bmales | Arithmetic    1000    4246.866         4163.24   4330.492
             |  Geometric    1000    4049.947        3973.353   4128.017
             |   Harmonic    1000    3863.352        3789.417    3940.23
-------------+-------------------------------------------------------------
    Bfemales | Arithmetic    1000    4271.245        4209.594   4332.896
             |  Geometric    1000    4159.078        4099.598   4219.422
             |   Harmonic    1000    4046.595        3986.199   4108.848
-------------+-------------------------------------------------------------
      Cmales | Arithmetic    1000    4361.919        4275.104   4448.733
             |  Geometric    1000    4149.635        4068.993   4231.875
             |   Harmonic    1000    3948.863        3872.451   4028.351
-------------+-------------------------------------------------------------
    Cfemales | Arithmetic    1000    4182.505        4120.209   4244.801
             |  Geometric    1000    4060.544        3998.722   4123.321
             |   Harmonic    1000    3934.264         3870.62   4000.037
-------------+-------------------------------------------------------------

#### I found it easiest to use tabstat to get the proportions as a mean..
```

```
. tabstat AMslry50 AMslry75 AFslry50 AFslry75 BMslry50 BMslry75 BFslry50
BFslry75 CMslry50 CMslry75 CFslry50 CFslry75, format(%9.3f) stat(mean) col(stat)
    variable |      mean
-------------+----------
    AMslry50 |     0.506
    AMslry75 |     0.091
    AFslry50 |     0.377
    AFslry75 |     0.024
    BMslry50 |     0.456
    BMslry75 |     0.081
    BFslry50 |     0.521
    BFslry75 |     0.047
    CMslry50 |     0.478
    CMslry75 |     0.102
    CFslry50 |     0.481
    CFslry75 |     0.031
------------------------

#### Trial and error to increase mean or geometric mean salary by $50 after giving
#### one person a raise.
. list Amales in 1

     +--------+
     | Amales |
     |--------|
  1. |   5631 |
     +--------+

. means Amales

    Variable |    Type        Obs        Mean       [95% Conf. Interval]
-------------+--------------------------------------------------------------
      Amales | Arithmetic    1000    4365.683        4283.022   4448.343
             |  Geometric    1000    4169.815        4091.703   4249.418
             |   Harmonic    1000    3979.939        3904.255   4058.616
------------------------------------------------------------------------------

#### To increase the mean to $4415.683 is pretty straightforward. There are 1000
#### men that on average need an extra $50, so I give $50,000 to one of them.
. replace Amales= 5631 + 50000 in 1
(1 real change made)

. means Amales

    Variable |    Type        Obs        Mean       [95% Conf. Interval]
-------------+--------------------------------------------------------------
      Amales | Arithmetic    1000    4415.683          4285.5   4545.865
             |  Geometric    1000    4179.377        4098.368   4261.986
             |   Harmonic    1000    3982.469        3906.394   4061.566
------------------------------------------------------------------------------

#### Trial and error to increase geometric mean salary by $50. Start with giving
#### one man a $1,000,000 raise.
. replace Amales=5631 + 1000000 in 1
(1 real change made)

. means Amales
```

```
    Variable |    Type        Obs        Mean       [95% Conf. Interval]
-------------+-------------------------------------------------------------
      Amales | Arithmetic    1000    5365.683        3399.119   7332.246
             |  Geometric    1000    4191.492        4101.304   4283.663
             |   Harmonic    1000    3982.738        3906.604     4061.9
-------------------------------------------------------------------------------

#### That only got us a geometric mean $21 higher. Now try giving
#### one man a $10,000,000 raise.
. replace Amales=5631 + 10000000 in 1
(1 real change made)

. means Amales

    Variable |    Type        Obs        Mean       [95% Conf. Interval]
-------------+-------------------------------------------------------------
      Amales | Arithmetic    1000    14365.68       -5260.391   33991.76
             |  Geometric    1000    4201.133        4100.272   4304.475
             |   Harmonic    1000    3982.753        3906.615   4061.917
-------------------------------------------------------------------------------

#### That only got us a geometric mean $30 higher. Now try giving
#### one man a $100,000,000 raise.
. replace Amales=5631 + 100000000 in 1
(1 real change made)

. means Amales

    Variable |    Type        Obs        Mean       [95% Conf. Interval]
-------------+-------------------------------------------------------------
      Amales | Arithmetic    1000    104365.7       -91870.97   300602.3
             |  Geometric    1000    4210.815        4097.142   4327.643
             |   Harmonic    1000    3982.754        3906.616   4061.919
-------------------------------------------------------------------------------

#### That only got us a geometric mean $40 higher. Now try giving
#### one man a $100,000,000 raise.
. replace Amales=5631 + 1000000000 in 1
(1 real change made)

. means Amales

    Variable |    Type        Obs        Mean       [95% Conf. Interval]
-------------+-------------------------------------------------------------
      Amales | Arithmetic    1000     1004366       -957978.3    2966710
             |  Geometric    1000    4220.522        4092.541   4352.506
             |   Harmonic    1000    3982.754        3906.616   4061.919
-------------------------------------------------------------------------------

#### That did it.
```