```
#### Biost 517: Applied Biostatistics I
#### Emerson, Fall 2007

#### Annotated Stata Log File: Homework #2
#### October 16, 2007

#### 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 id site age male bkrace smoker estrogen prevdis diab2 bmi systBP aai
> cholest crp fib ttodth death cvddth using  inflamm.txt
'id' cannot be read as a number for id[1]
'site' cannot be read as a number for site[1]
'age' cannot be read as a number for age[1]
'male' cannot be read as a number for male[1]
'bkrace' cannot be read as a number for bkrace[1]
'smoker' cannot be read as a number for smoker[1]
'estrogen' cannot be read as a number for estrogen[1]
'prevdis' cannot be read as a number for prevdis[1]
'diab2' cannot be read as a number for diab2[1]
'bmi' cannot be read as a number for bmi[1]
'systBP' cannot be read as a number for systBP[1]
'aai' cannot be read as a number for aai[1]
'cholest' cannot be read as a number for cholest[1]
'crp' cannot be read as a number for crp[1]
'fib' cannot be read as a number for fib[1]
'ttodth' cannot be read as a number for ttodth[1]
'death' cannot be read as a number for death[1]
'cvddth' cannot be read as a number for cvddth[1]
'NA' cannot be read as a number for aai[26]
. . .
(lots of additional lines of warnings omitted)
 (5001 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
  1. | id | site | age | male | bkrace | smoker | estrogen | prevdis | diab2 |
     |  . |    . |   . |    . |      . |      . |        . |       . |     . |

     | bmi | systBP | aai | cholest | crp | fib | ttodth | death | cvddth     |
     |   . |      . |   . |       . |   . |   . |      . |     . |       . |

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

#### Now I want to set up nice formats for the variables. I first look at the
#### range of values for each variable and then choose formatting that will
#### tend to provide 3 significant digits.
. summ
    Variable |       Obs        Mean    Std. Dev.       Min        Max
          id |      5000      2500.5     1443.52          1       5000
        site |      5000      2.4702    1.131268          1          4
         age |      5000     72.8304    5.596418         65        100
        male |      5000       .4192    .4934775          0          1
      bkrace |      5000       .1546    .3615591          0          1
      smoker |      4994    .1209451    .3260962          0          1
    estrogen |      4994    .0688827    .2532799          0          1
     prevdis |      5000       .2298    .4207462          0          1
       diab2 |      4943    .1602266    .3668532          0          1
         bmi |      4987     26.6687    4.735166       14.7       58.8
      systBP |      4990    136.5549    21.86058         77        235
         aai |      4879     1.06393    .1745873      .2778     2.3846
     cholest |      4953    211.6893    39.28814         73        430
         crp |      4933    3.613825    6.152715          0        108
         fib |      4915     322.978    67.28736        109        872
      ttodth |      5000    2368.381    676.8838          5       2942
       death |      5000       .2242    .4170961          0          1
      cvddth |      5000       .0994    .2992283          0          1
```

```
#### I define appropriate formats
. format age bmi crp %9.1f
. format systBP cholest fib %9.0f
. format aai %9.2f

#### I save the file so I won't have to do this again
. save inflamm
file inflamm.dta saved

#### Problem 2a: Checking for the earliest censoring time
. bysort death: tabstat ttodth, stat(n mean sd min q max)
-> death = 0
    variable |         N      mean        sd       min       p25       p50       p75       max
-------------+----------------------------------------------------------------------------------
      ttodth |      3879  2603.711  413.5922      1480      2630      2726      2834      2942
----------------------------------------------------------------------------------------------

-> death = 1
    variable |         N      mean        sd       min       p25       p50       p75       max
-------------+----------------------------------------------------------------------------------
      ttodth |      1121  1554.069  772.7934         5       934      1609      2236      2912
----------------------------------------------------------------------------------------------

. disp 1480/365.25
4.0520192


#### Generating the variable indicating four (or more) year survival
. g surv4yr= ttodth/365.25
. recode surv4yr 0/4=0 4/max=1
(surv4yr: 5000 changes made)

. table surv4yr
----------------------
  surv4yr |      Freq.
----------+-----------
        0 |        495
        1 |      4,505
----------------------
```

```
#### Problem 2b: Graphical and tabular descriptive statistics
#### Producing histograms. Note the way I forced the xaxis to always use the same
#### scale. Also note the way to place titles on the plots.
#### After getting the plots, I used cut and paste to put them in the Word document.
. hist cholest, bin(20) title(All Patients: Cholesterol) xscale(range(0 500))
(bin=20, start=73, width=17.85)

. hist crp, bin(20) title(All Patients: CRP) xscale(range(0 120))
(bin=20, start=0, width=5.4)

. hist cholest if surv4yr==0, bin(20) title(Patients Surviving < 4 yr: Cholesterol) xscale(range(0 500))
(bin=20, start=73, width=16.15)

. hist crp if surv4yr==0, bin(20) title(Patients Surviving < 4 yr: CRP) xscale(range(0 120))
(bin=20, start=0, width=2.75)

. hist cholest if surv4yr==1, bin(20) title(Patients Surviving > 4 yr: Cholesterol) xscale(range(0 500))
(bin=20, start=78, width=17.6)

. hist crp if surv4yr==1, bin(20) title(Patients Surviving > 4 yr: CRP) xscale(range(0 120))
(bin=20, start=0, width=5.4)


#### I intend to just execute two commands to get all the numeric descriptive
#### statistics: I will use "tabstat" to get everything but the geometric mean
#### Because I can use the mean of a binary variable to get the proportion, I
#### first create the dichotomized variables.
. g hichol= cholest
(47 missing values generated)
. recode hichol min/250=0 250/max=1
(hichol: 4953 changes made)
. g hicrp= crp
(67 missing values generated)
. recode hicrp min/3=0 3/max=1
(hicrp: 4505 changes made)

. format hichol hicrp %9.3f
```

Now to get most the numeric descriptives

```
. tabstat cholest crp hichol hicrp, by(surv4yr) stat(n mean sd min q max iqr r) col(stat) format
Summary for variables: cholest crp hichol hicrp
     by categories of: surv4yr
 surv4yr |         N    mean        sd       min       p25       p50       p75       max       iqr     range
---------+------------------------------------------------------------------------------------------------------
       0 |       486     204        41        73       176       202       229       396        53       323
         |     484.0     5.4       8.1       0.0       1.0       3.0       6.0      55.0       5.0      55.0
         |   486.000   0.126     0.332     0.000     0.000     0.000     0.000     1.000     0.000     1.000
         |   484.000   0.378     0.485     0.000     0.000     0.000     1.000     1.000     1.000     1.000
---------+------------------------------------------------------------------------------------------------------
       1 |      4467     213        39        78       187       211       236       430        49       352
         |    4449.0     3.4       5.9       0.0       1.0       2.0       3.0     108.0       2.0     108.0
         |  4467.000   0.158     0.364     0.000     0.000     0.000     0.000     1.000     0.000     1.000
         |  4449.000   0.223     0.416     0.000     0.000     0.000     0.000     1.000     0.000     1.000
---------+------------------------------------------------------------------------------------------------------
   Total |      4953     212        39        73       186       210       236       430        50       357
         |    4933.0     3.6       6.2       0.0       1.0       2.0       3.0     108.0       2.0     108.0
         |  4953.000   0.154     0.361     0.000     0.000     0.000     0.000     1.000     0.000     1.000
         |  4933.000   0.238     0.426     0.000     0.000     0.000     0.000     1.000     0.000     1.000
------------------------------------------------------------------------------------------------------------------
```

Now to get the geometric means

```
. bysort surv4yr: means cholest crp
-> surv4yr = 0

    Variable |    Type         Obs        Mean       [95% Conf. Interval]
-------------+-------------------------------------------------------------
     cholest | Arithmetic      486     204.072        200.3846   207.7595
             |  Geometric      486      199.87        196.2222   203.5856
             |   Harmonic      486    195.5161        191.7199   199.4656
-------------+-------------------------------------------------------------
         crp | Arithmetic      484    5.376033        4.652803   6.099263
             |  Geometric      463    3.221088        2.949784   3.517344
             |   Harmonic      463    2.234659        2.091234   2.399206
-------------+-------------------------------------------------------------
```

```
-> surv4yr = 1
    Variable |    Type         Obs        Mean       [95% Conf. Interval]
-------------+-------------------------------------------------------------
     cholest | Arithmetic    4467     212.518        211.3749   213.6611
             |  Geometric    4467    208.9306        207.7927   210.0747
             |   Harmonic    4467    205.2662        204.1009   206.4448
-------------+-------------------------------------------------------------
         crp | Arithmetic    4449    3.422117        3.249526   3.594709
             |  Geometric    4042    2.337046         2.27629   2.399423
             |   Harmonic    4042    1.770859        1.737885   1.805108
-------------+-------------------------------------------------------------

. means cholest crp
    Variable |    Type         Obs        Mean       [95% Conf. Interval]
-------------+-------------------------------------------------------------
     cholest | Arithmetic    4953    211.6893        210.5949   212.7837
             |  Geometric    4953    208.0236        206.9329   209.1201
             |   Harmonic    4953    204.2666        203.1444   205.4013
-------------+-------------------------------------------------------------
         crp | Arithmetic    4933    3.613825        3.442087   3.785563
             |  Geometric    4505    2.415391        2.354689   2.477657
             |   Harmonic    4505    1.809456        1.776769   1.843368
-------------+-------------------------------------------------------------

#### We note in the above that the number of observations used to compute the arithmetic
#### means for CRP is different from the number used to compute the geometric and harmonic means.
#### This is due to the fact that the geometric and harmonic means are properly
#### computed only for positive (nonzero, nonnegative) numbers.
#### I look to see how many zero values there are for CRP, as well as what is the
#### lowest nonzero value.
. table surv4yr crp if crp <=5

----------------------------------------------------
          |                    crp
  surv4yr |   0.0    1.0    2.0    3.0    4.0    5.0
----------+-----------------------------------------
        0 |    21    108     94     78     34     24
        1 |   407  1,433    994    623    205    114
----------------------------------------------------
```

```
#### So, as discussed in the key, I need to do something about the zeroes. I decide to
#### "impute" values of 0.5 for all of the zeroes (not necessarily correct, but it
#### something that is commonly done.
. g crpLLD= crp
(67 missing values generated)

. replace crpLLD= 0.5 if crp==0
(428 real changes made)

. means crpLLD

    Variable |    Type         Obs        Mean       [95% Conf. Interval]
-------------+-----------------------------------------------------------
      crpLLD | Arithmetic     4933    3.657207        3.486136   3.828277
             |  Geometric     4933    2.106882        2.052126   2.163099
             |   Harmonic     4933     1.47443         1.44319   1.507054
-----------------------------------------------------------------------

. bysort surv4yr: means crpLLD

-> surv4yr = 0

    Variable |    Type         Obs        Mean       [95% Conf. Interval]
-------------+-----------------------------------------------------------
      crpLLD | Arithmetic      484    5.397727         4.67573   6.119724
             |  Geometric      484    2.970981        2.713267   3.253174
             |   Harmonic      484    1.942289        1.799804   2.109275
-----------------------------------------------------------------------

-> surv4yr = 1

    Variable |    Type         Obs        Mean       [95% Conf. Interval]
-------------+-----------------------------------------------------------
      crpLLD | Arithmetic     4449    3.467858           3.296   3.639716
             |  Geometric     4449    2.029563        1.974973   2.085662
             |   Harmonic     4449     1.43678        1.405324   1.469676
-----------------------------------------------------------------------
```

```
#### Problem 3:
#### First sort the data, because I will replace the lowest cholesterol value among the poor survivors.
. sort surv4yr cholest
. list id age male cholest crp surv4yr in 1

     +-----------------------------------------------+
     |   id    age   male   cholest   crp   surv4yr |
     |-----------------------------------------------|
  1. | 2924   84.0      1        73   1.0         0 |
     +-----------------------------------------------+

#### Problem 3a:
. replace cholest=73+20*486 if id==2924
(1 real change made)

. list id age male cholest crp surv4yr in 1
     +-----------------------------------------------+
     |   id    age   male   cholest   crp   surv4yr |
     |-----------------------------------------------|
  1. | 2924   84.0      1      9793   1.0         0 |
     +-----------------------------------------------+

. tabstat cholest hichol, by(surv4yr) stat(n mean sd min q max) col(stat) format
Summary for variables: cholest hichol
     by categories of: surv4yr

 surv4yr |         N      mean        sd       min       p25       p50       p75       max
---------+----------------------------------------------------------------------------------
       0 |       486       224       437        93       176       202       229      9793
         |   486.000     0.126     0.332     0.000     0.000     0.000     0.000     1.000
---------+----------------------------------------------------------------------------------
       1 |      4467       213        39        78       187       211       236       430
         |  4467.000     0.158     0.364     0.000     0.000     0.000     0.000     1.000
---------+----------------------------------------------------------------------------------
   Total |      4953       214       142        78       186       210       236      9793
         |  4953.000     0.154     0.361     0.000     0.000     0.000     0.000     1.000
----------------------------------------------------------------------------------------
```

```
. bysort surv4yr: means cholest
-> surv4yr = 0

    Variable |    Type         Obs        Mean       [95% Conf. Interval]
-------------+-------------------------------------------------------------
     cholest | Arithmetic     486     224.072        185.1343   263.0097
             |  Geometric     486    201.8949        197.1321   206.7727
             |   Harmonic     486    196.5914        192.9039   200.4226
-----------------------------------------------------------------------------

-> surv4yr = 1

    Variable |    Type         Obs        Mean       [95% Conf. Interval]
-------------+-------------------------------------------------------------
     cholest | Arithmetic    4467     212.518        211.3749   213.6611
             |  Geometric    4467    208.9306        207.7927   210.0747
             |   Harmonic    4467    205.2662        204.1009   206.4448
-----------------------------------------------------------------------------

. means cholest
    Variable |    Type         Obs        Mean       [95% Conf. Interval]
-------------+-------------------------------------------------------------
     cholest | Arithmetic    4953    213.6517         209.705   217.5985
             |  Geometric    4953    208.2295        207.0961   209.3691
             |   Harmonic    4953    204.3812        203.2643   205.5105
-----------------------------------------------------------------------------
```

```
#### Problem 3b: (see key for discussion of rationale)
. display 220/200
1.1
. display 1.1^486
1.309e+20

. replace cholest=73 * 1.1^486 if id==2924
(1 real change made)

. tabstat cholest hichol, by(surv4yr) stat(n mean sd min q max) col(stat) format
Summary for variables: cholest hichol
     by categories of: surv4yr

 surv4yr |         N      mean        sd       min       p25       p50       p75       max
---------+----------------------------------------------------------------------------------
       0 |       486  1.97e+19  4.33e+20        93       176       202       229  9.55e+21
         |   486.000     0.126     0.332     0.000     0.000     0.000     0.000     1.000
---------+----------------------------------------------------------------------------------
       1 |      4467       213        39        78       187       211       236       430
         |  4467.000     0.158     0.364     0.000     0.000     0.000     0.000     1.000
---------+----------------------------------------------------------------------------------
   Total |      4953  1.93e+18  1.36e+20        78       186       210       236  9.55e+21
         |  4953.000     0.154     0.361     0.000     0.000     0.000     0.000     1.000
--------------------------------------------------------------------------------------------

. bysort surv4yr: means cholest
-> surv4yr = 0
    Variable |    Type        Obs        Mean       [95% Conf. Interval]
-------------+-------------------------------------------------------------
     cholest | Arithmetic     486    1.97e+19       -1.90e+19   5.83e+19
             |  Geometric     486     219.857        182.8941     264.29
             |   Harmonic     486    196.5995        192.9085   200.4345
-----------------------------------------------------------------------
-> surv4yr = 1
    Variable |    Type        Obs        Mean       [95% Conf. Interval]
-------------+-------------------------------------------------------------
     cholest | Arithmetic    4467     212.518        211.3749   213.6611
             |  Geometric    4467    208.9306        207.7927   210.0747
             |   Harmonic    4467    205.2662        204.1009   206.4448
-----------------------------------------------------------------------
```

```
. means cholest
    Variable |    Type         Obs        Mean       [95% Conf. Interval]
-------------+-----------------------------------------------------------
     cholest | Arithmetic    4953    1.93e+18       -1.85e+18   5.71e+18
             |  Geometric    4953    209.9782        206.0942   213.9354
             |   Harmonic    4953    204.3821        203.2651   205.5115
-----------------------------------------------------------------------

#### Alternative approach yields too small a number to represent in Stata
. display 0.9 ^ 4467
3.99e-205


#### Problem 3c:
. table cholest if cholest>=201 & cholest <= 203 & surv4yr==0

----------------------
  cholest |      Freq.
----------+-----------
      201 |          4
      202 |          2
      203 |          7
----------------------

. table cholest if cholest>=210 & cholest <= 212 & surv4yr==1

----------------------
  cholest |      Freq.
----------+-----------
      210 |         44
      211 |         49
      212 |         57
----------------------
```