# STATA 002

January 16, 2002

The datafile being used here is wtlosspilotwide.dta. The study from which this data was generated is described on the class website under datasets. This data is described (in STATA terms) as being in wide format because although there are measurements at multiple time points for each individual, there is only one record per individual.

```
. log using kay.log,replace

. des

Contains data from
C:\Stata\MyFiles\BiometryClass2\problems\classdemo\wtlosspilotwide.dta
  obs:            79
 vars:            21                          16 Jan 2002 07:53
 size:         5,135 (100.0% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
id              float  %9.0g
group           byte   %8.0g                  Treatment Group
age             byte   %8.0g                  Age in years
heightin        float  %4.1f                  Height in inches
heightm         float  %9.0g                  Height in meters
wtlbs0          float  %5.1f                  Weight at baseline in pounds
wtlbs6          float  %5.1f                  Weight at six months in pounds
wtkg0           float  %6.2f                  Weight at baseline in kilograms
wtkg6           float  %6.2f                  Weight at six months in
                                                kilograms
bmi0            float  %4.1f                  Body Mass Index (wt in kgs/ht
                                                in meters squared) at baseline
bmi6            float  %4.1f                  Body Mass Index (wt in kgs/ht
                                                in meters squared) at six
                                                months
chol0           int    %8.0g                  Total cholesterol at baseline
chol6           int    %8.0g                  Total cholesterol at six months
hdl0            byte   %8.0g                  HDL at baseline
hdl6            byte   %8.0g                  HDL at six months
ldl0            int    %8.0g                  LDL at baseline
ldl6            int    %8.0g                  LDL at six months
sbp0            float  %5.1f                  Systolic Blood Pressure at
                                                baseline
sbp6            float  %5.1f                  Systolic Blood Pressure at six
                                                months
dbp0            float  %5.1f                  Diastolic Blood Pressure at
                                                baseline
dbp6            byte   %8.0g                  Diastolic Blood Pressure at six
                                                months
-------------------------------------------------------------------------------
Sorted by:

. sort id
```

The listing on the next page gives you an idea of how STATA stores the data.

```
. list id group age sbp0 sbp6
```

| | id | group | age | sbp0 | sbp6 |
|---|---|---|---|---|---|
| 1. | 1 | 2 | 44 | 119.0 | 109.0 |
| 2. | 2 | 2 | 29 | 123.0 | 127.0 |
| 3. | 3 | 2 | 26 | 118.0 | 120.0 |
| 4. | 4 | 2 | 41 | 108.0 | 102.0 |
| 5. | 5 | 2 | 29 | 89.0 | 96.0 |
| 6. | 6 | 2 | 40 | 134.0 | 121.0 |
| 7. | 7 | 2 | 43 | 96.0 | 105.0 |
| 8. | 8 | 2 | 28 | 120.0 | 110.0 |
| 9. | 9 | 2 | 35 | 110.0 | 121.0 |
| 10. | 10 | 2 | 35 | 98.0 | 94.0 |
| 11. | 11 | 2 | 38 | 117.0 | 114.0 |
| 12. | 12 | 2 | 30 | 109.0 | 106.0 |
| 13. | 13 | 2 | 26 | 133.0 | 129.0 |
| 14. | 14 | 2 | 37 | 113.0 | 107.0 |
| 15. | 15 | 2 | 34 | 110.0 | 110.0 |
| 16. | 16 | 2 | 37 | 100.0 | 104.0 |
| 17. | 17 | 2 | 31 | 110.0 | 115.0 |
| 18. | 18 | 2 | 33 | 115.0 | 118.0 |
| 19. | 19 | 2 | 33 | 104.0 | 120.0 |
| 20. | 20 | 2 | 39 | 120.0 | 117.0 |
| 21. | 21 | 2 | 34 | 102.0 | 120.0 |
| 22. | 22 | 2 | 37 | 119.0 | 124.5 |
| 23. | 23 | 2 | 38 | 90.0 | 90.0 |
| 24. | 24 | 2 | 38 | 118.0 | 120.0 |
| 25. | 25 | 2 | 26 | 110.0 | 114.0 |
| 26. | 26 | 2 | 26 | 104.0 | 109.0 |
| 27. | 27 | 2 | 36 | 100.0 | 99.0 |
| 28. | 28 | 2 | 37 | 120.0 | 121.0 |
| 29. | 29 | 2 | 18 | 119.0 | 103.0 |
| 30. | 30 | 2 | 25 | 112.0 | 109.0 |
| 31. | 31 | 2 | 34 | 97.0 | 92.0 |
| 32. | 32 | 2 | 41 | 135.0 | 125.0 |
| 33. | 33 | 2 | 41 | 108.0 | 111.0 |
| 34. | 34 | 2 | 36 | 98.0 | 110.0 |
| 35. | 35 | 2 | 18 | 105.0 | 107.0 |
| 36. | 36 | 2 | 27 | 110.0 | 115.0 |
| 37. | 37 | 2 | 29 | 102.0 | 107.0 |
| 38. | 38 | 2 | 33 | 110.0 | 123.0 |
| 39. | 39 | 2 | 40 | 107.0 | 101.0 |
| 40. | 40 | 2 | 32 | 90.0 | 90.0 |
| 41. | 41 | 2 | 36 | 104.0 | 94.0 |
| 42. | 42 | 2 | 35 | 101.0 | 104.0 |
| 43. | 43 | 2 | 34 | 131.0 | 120.0 |
| 44. | 44 | 2 | 33 | 97.0 | 100.0 |
| 45. | 45 | 2 | 39 | 111.0 | 111.0 |
| 46. | 46 | 3 | 37 | 109.0 | 118.0 |
| 47. | 47 | 3 | 30 | 108.0 | 95.0 |
| 48. | 48 | 3 | 37 | 96.0 | 89.0 |
| 49. | 49 | 3 | 28 | 102.0 | 125.0 |
| 50. | 50 | 3 | 27 | 124.0 | 110.0 |
| 51. | 51 | 3 | 40 | 129.0 | 128.0 |
| 52. | 52 | 3 | 21 | 104.0 | 103.0 |
| 53. | 53 | 3 | 45 | 112.0 | 115.0 |
| 54. | 54 | 3 | 28 | 121.0 | 111.0 |
| 55. | 55 | 3 | 25 | 108.0 | 93.0 |
| 56. | 56 | 3 | 37 | 114.0 | 95.0 |
| 57. | 57 | 3 | 29 | 120.0 | 118.0 |
| 58. | 58 | 3 | 33 | 105.0 | 110.0 |
| 59. | 59 | 3 | 44 | 107.0 | 112.0 |
| 60. | 60 | 3 | 32 | 110.0 | 126.0 |
| 61. | 61 | 3 | 37 | 107.0 | 121.0 |
| 62. | 62 | 3 | 30 | 106.0 | 126.0 |
| 63. | 63 | 3 | 26 | 100.0 | 92.0 |
| 64. | 64 | 3 | 24 | 95.0 | 109.0 |
| 65. | 65 | 3 | 42 | 128.0 | 128.0 |
| 66. | 66 | 3 | 27 | 127.0 | 132.0 |
| 67. | 67 | 3 | 30 | 91.0 | 97.0 |
| 68. | 68 | 3 | 41 | 146.0 | 135.0 |

```
 69.        69         3        29  110.0   98.0
 70.        70         3        35  120.0  120.0
 71.        71         3        36  130.0  125.0
 72.        72         3        35   89.0  115.0
 73.        73         3        38  119.0   97.0
 74.        74         3        26   94.0   95.0
 75.        75         3        38  132.0  121.0
 76.        76         3        36   95.0   87.0
 77.        77         3        36  100.0   97.0
 78.        78         3        26  114.0  114.0
 79.        79         3        38  121.0  117.0
```

```
. tab group

  Treatment |
      Group |      Freq.     Percent        Cum.
------------+-----------------------------------
          2 |         45       56.96       56.96
          3 |         34       43.04      100.00
------------+-----------------------------------
      Total |         79      100.00
```

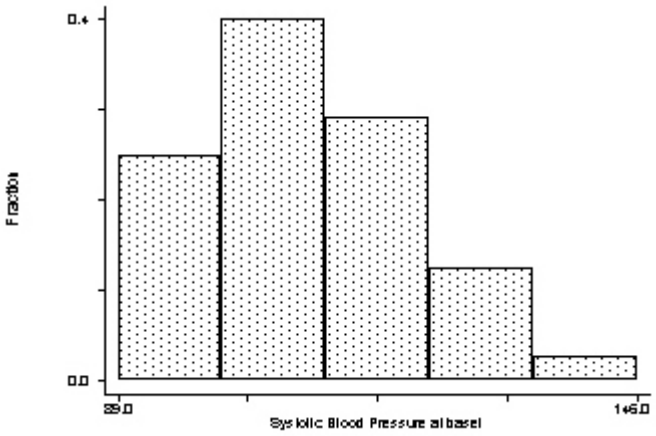
```
. sum sbp0        [This includes all participants.]

    Variable |     Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        sbp0 |      79    110.6203     12.2331         89        146
```

```
. sum sbp0 if group == 2    [This selects only the 45 participants in Group 2]

    Variable |     Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        sbp0 |      45    109.9111    11.40339         89        135
```

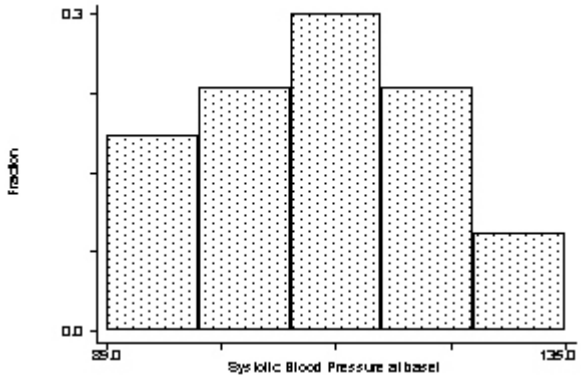
```
. sum sbp0 if group == 2,det     [det requests more detail]

             Systolic Blood Pressure at baseline
-------------------------------------------------------------
      Percentiles      Smallest
 1%           89             89
 5%           90             90
10%           97             90       Obs                  45
25%          102             96       Sum of Wgt.          45

50%          110                      Mean           109.9111
                        Largest       Std. Dev.      11.40339
75%          118            131
90%          123            133       Variance       130.0374
95%          133            134       Skewness       .3237869
99%          135            135       Kurtosis       2.727598
```

```
. tab group,sum(sbp0)

           | Summary of Systolic Blood Pressure
 Treatment |              at baseline
     Group |        Mean   Std. Dev.       Freq.
-----------+------------------------------------
         2 |       109.9        11.4          45
         3 |       111.6        13.4          34
-----------+------------------------------------
     Total |       110.6        12.2          79
```

```
. graph sbp0      [histogram for all participants]
```



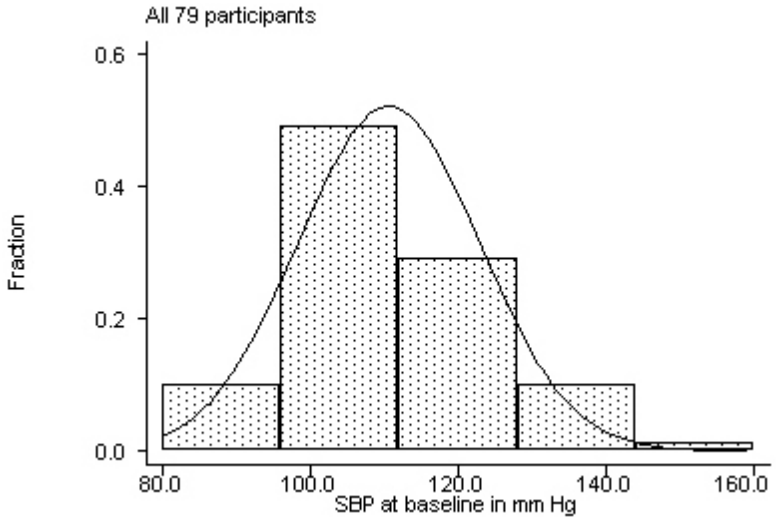

```
. graph sbp0 if group == 2    [histogram for only Group 2 participants]
```



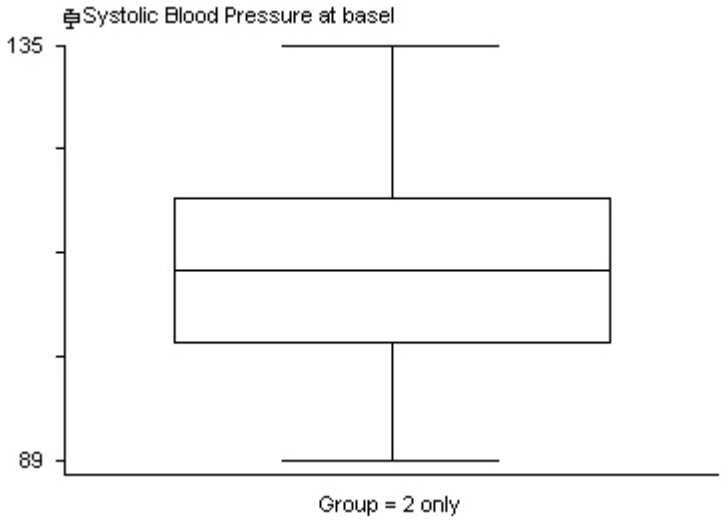

**[Lack of labeling on the above two graphs is frowned upon.]**

```
. set textsize 120        [This increases the size of the text on graphs.  The
                          default is 100.]

. graph sbp0,t1(All 79 participants) xlab ylab normal b2(SBP at baseline in mm
Hg)
```

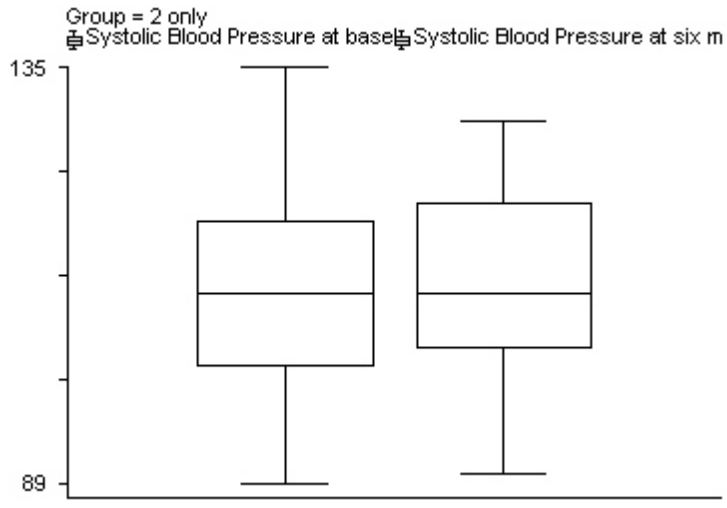
[The normal curve overlaid on the histogram below has mean = 110.6 and SD = 12.2 since those are the values of the mean and SD for SBP at baseline for all 79 participants]



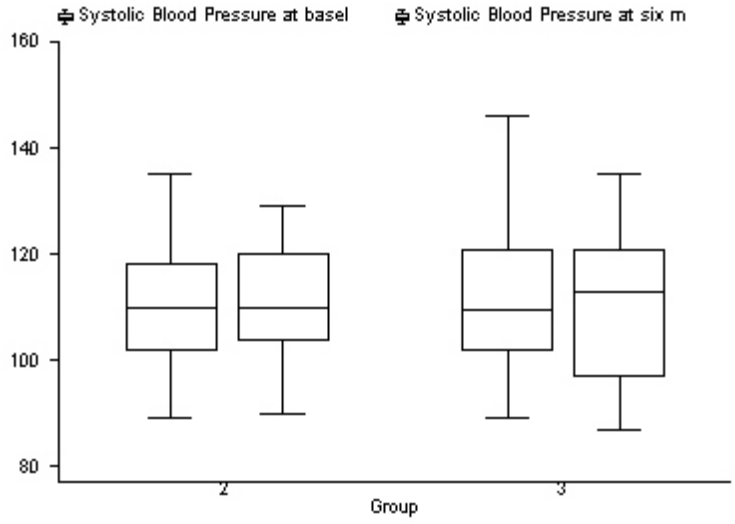

```
. graph sbp0 if group == 2,box b2(Group = 2 only) [box-and-whisker plot]
```



```
. graph sbp0 sbp6 if group == 2,box t1(Group = 2 only)
```




```
. graph sbp0 sbp6,box by(group) b2(Group) ylab
```



This is the same data as used on the earlier pages but it is in the long format. This means there is one record for each measurement time point. Since the participants were measured at baseline and six months, there are two records per participant.

```
. des

Contains data from
C:\Stata\MyFiles\BiometryClass2\problems\classdemo\wtlosspilot.dta
  obs:           158
 vars:            14                          16 Jan 2002 08:40
 size:         6,952 (100.0% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
id              float  %9.0g
visit           byte   %10.0g      visit      Measurement time point
group           byte   %8.0g                  Treatment Group
age             byte   %8.0g                  Age in years
heightin        float  %4.1f                  Height in inches
heightm         float  %9.0g                  Height in meters
wtlbs           float  %5.1f                  Weight in pounds
wtkg            float  %6.2f                  Weight in kilograms
bmi             float  %4.1f                  Body Mass Index - wt in kg/ht
                                                in m^2
chol            int    %8.0g                  Total cholesterol in mg/dL
hdl             byte   %8.0g                  HDL in mg/dL
ldl             int    %8.0g                  LDL in mg/dL
sbp             float  %5.1f                  Systolic Blood Pressure in mm Hg
dbp             float  %8.0g                  Diastolic Blood Pressure in mm
                                                Hg
-------------------------------------------------------------------------------
Sorted by:  id  visit

. list id visit group sbp

             id       visit     group    sbp
  1.          1    baseline         2  119.0
  2.          1  six months         2  109.0
  3.          2    baseline         2  123.0
  4.          2  six months         2  127.0
  5.          3    baseline         2  118.0
  6.          3  six months         2  120.0
  7.          4    baseline         2  108.0
  8.          4  six months         2  102.0
  9.          5    baseline         2   89.0
 10.          5  six months         2   96.0
 11.          6    baseline         2  134.0
 12.          6  six months         2  121.0
 13.          7    baseline         2   96.0
 14.          7  six months         2  105.0
 15.          8    baseline         2  120.0
 16.          8  six months         2  110.0
 17.          9    baseline         2  110.0
 18.          9  six months         2  121.0
--Break--
r(1);
```

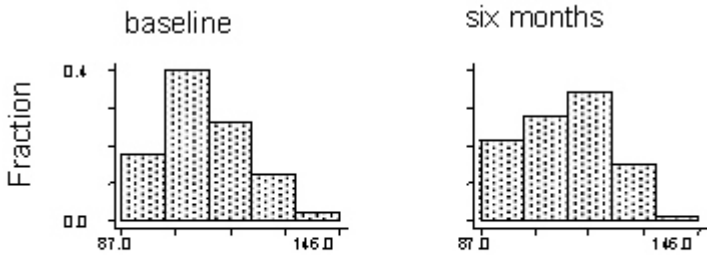
Notice that the table below counts everyone twice.

```
. tab group

  Treatment |
      Group |      Freq.     Percent        Cum.
------------+-----------------------------------
          2 |         90       56.96       56.96
          3 |         68       43.04      100.00
------------+-----------------------------------
      Total |        158      100.00
```

```
. tab group visit

           |   Measurement time
 Treatment |         point
     Group |  baseline  six month |     Total
-----------+----------------------+----------
         2 |        45         45 |        90
         3 |        34         34 |        68
-----------+----------------------+----------
     Total |        79         79 |       158
```

```
. graph sbp,by(visit)
not sorted [STATA wants the "by" variable to be sorted.]
r(5);

. sort visit

. graph sbp,by(visit)
```

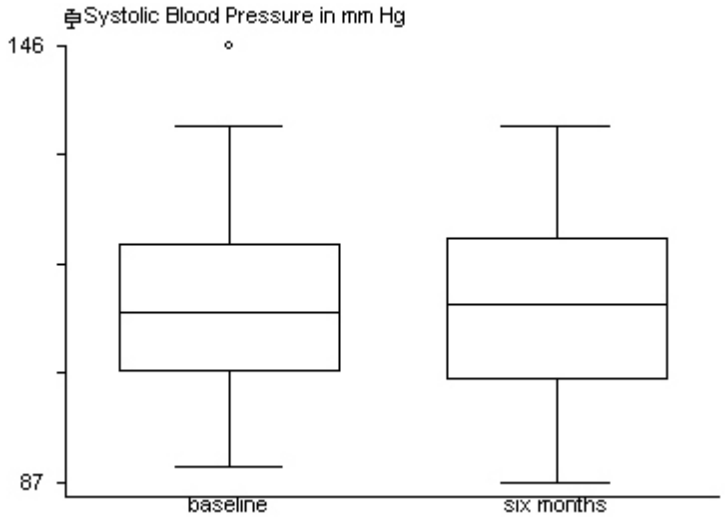
[I moved the labels at the bottom closer to the histograms using PowerPoint]



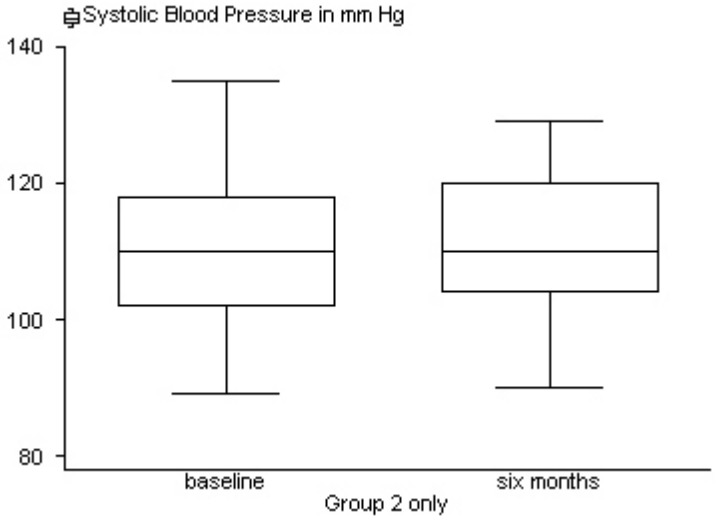

Systolic Blood Pressure in mm Hg

Histograms by Measurement time point

```
. graph sbp,by(visit) box
```



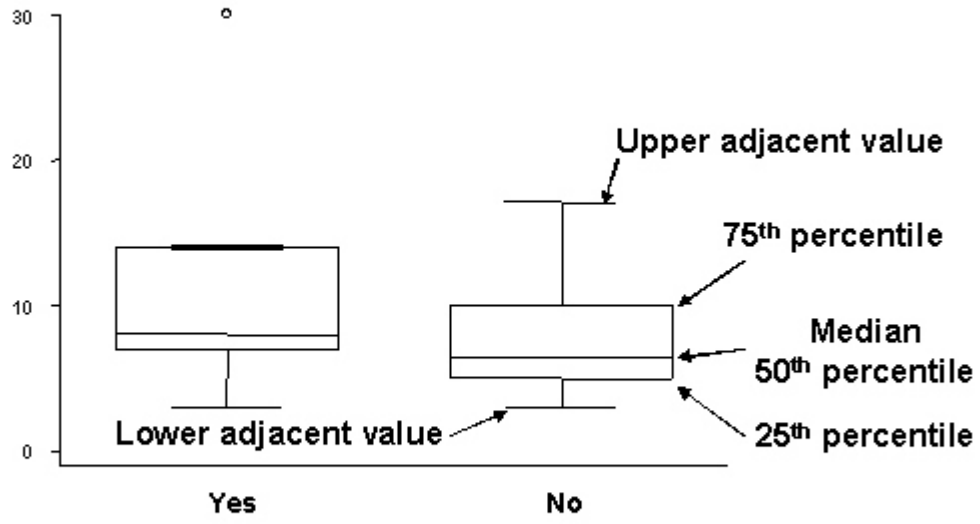

```
. graph sbp if group == 2,by(visit) box ylab b2(Group 2 only)
```



The hospital stay data is a sample taken from a larger data set collected on persons discharged from a selected Pennsylvania hospital as part of a retrospective chart review of antibiotic usage. See the Infectious Disease Problem on page 39 of Rosner. The STATA data file hospital.dta is on the course website. The list of variables is given below.

```
.log using kay.log,replace

. des

Contains data from C:\Wp51\MyFiles\BiometryCourse2\problemset\hospital.dta
  obs:            25
 vars:            11                          11 Jan 2002 16:44
 size:         1,200 (100.0% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
id              float  %12.0g
stay            float  %4.2f                  Length of hospital stay in days
age             float  %12.0g                 Age in years
sex             float  %12.0g      sexmf      Gender
temp            float  %4.3f                  First Temperature Following
                                                Admission in degrees F
wbc             float  %12.0g                 First White Blood Count (x10^3)
                                                Following Admission
antibio         float  %12.0g      yesno      Received Antibiotic - yes/no
bactcul         float  %12.0g      yesno      Received Bacterial Culture -
                                                yes/no
service         float  %12.0g      serv       Hospital Service
catage          float  %11.0g      ageqtile2
                                              Age in Quartiles from lowest to
                                                highest
logstay         float  %9.0g                  Log transform of length of stay
-------------------------------------------------------------------------------
Sorted by:  antibio

. label list
ageqtile2:
           1 <= 25
           2 >25&<=41
           3 >41&<=56
           4 >56
sexmf:
           1 Male
           2 Female
yesno:
           1 Yes
           2 No
serv:
           1 Medical
           2 Surgical

. log close
```

**Duration of hospital stay in days by antibiotic use - yes/no**




$U = 75^{th}\ Ptile + 1.5(75^{th}\ Ptile - 25^{th}\ Ptile)$
$L = 25^{th}\ Ptile - 1.5(75^{th}\ Ptile - 25^{th}\ Ptile)$
U and L are due to John Tukey

STATA command to get plot above: .graph stay,box by(antibio)
Dataset on website is: hospital.dta

Let x denote the length of hospital stay.
Let $x_{(i)}$ denote the x's in ascending order (i.e. $x_{(i)} \leq x_{(i+1)}$). Then the upper adjacent value is defined as the $x_{(i)}$ such that $x_{(i)} \leq U$ and $x_{(i+1)} > U$. The lower adjacent value is the $x_{(i)}$ such that $x_{(i)} \geq L$ and $x_{(i-1)} < L$.

```
-> antibio = Yes

                  Duration of Stay in Days
-------------------------------------------------------------
      Percentiles      Smallest
 1%             3             3
 5%             3             7
10%             3             8       Obs                   7
25%             7             8       Sum of Wgt.           7

50%             8                     Mean           11.57143
                        Largest       Std. Dev.      8.810167
75%            14             8
90%            30            11       Variance       77.61905
95%            30            14       Skewness       1.436382
99%            30            30       Kurtosis       3.922354
```

For the antibiotic = yes group, U = 14 + 1.5(14 - 7) = 24.5 and L = 7 - 1.5(14 - 7) = -3.5 The upper adjacent value is 14 for the antibiotic = yes group since 14 ≤ U and 30 (the next largest value) > U. But 14 is also the 75th percentile. So the upper whisker falls on top of the 75th percentile line (or falls on top of the upper part of the box). The lower adjacent whisker is 3.

## Test of the equality of the variances using STATA sdtest.

```
. sdtest stay,by(antibio)

Variance ratio test

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     Yes |       7    11.57143     3.32993    8.810167    3.423383    19.71947
      No |      18    7.444444    .8715632    3.697729    5.605607    9.283282
---------+--------------------------------------------------------------------
combined |      25         8.6    1.143095    5.715476    6.240767    10.95923
------------------------------------------------------------------------------

                           Ho: sd(Yes) = sd(No)

               F(6,17) observed   = F_obs             =     5.677
               F(6,17) lower tail = F_L   = 1/F_obs   =     0.176
               F(6,17) upper tail = F_U   = F_obs     =     5.677

 Ha: sd(Yes) < sd(No)        Ha: sd(Yes) ~= sd(No)        Ha: sd(Yes) > sd(No)
  P < F_obs = 0.9979      P < F_L + P > F_U = 0.0225       P > F_obs = 0.0021
```

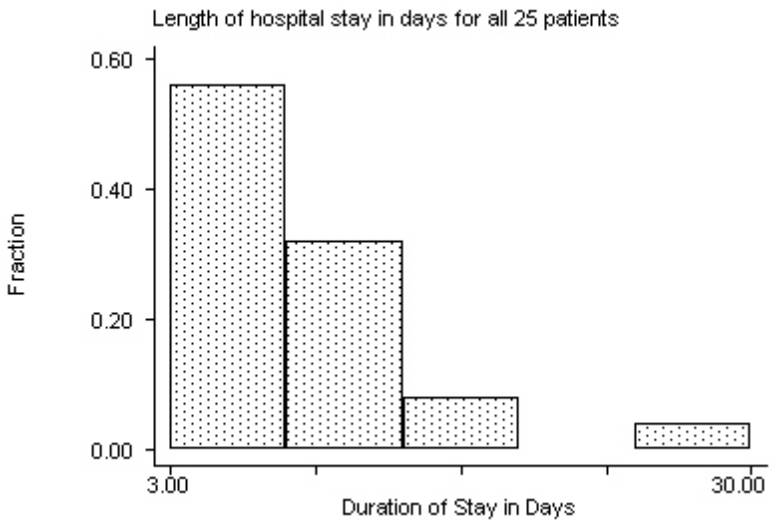
Note that F_obs = 5.677 = $8.810167^2/3.697729^2$ = (variance for yes)/(variance for no)

If you graph the variable length of stay (something we should have done first), you'll see that it is not very normal looking. I would transform the variable by taking the natural log of the stay.

The figure below is the graph of the length of stay in days for all 25 patients. You can see it is not very normal looking.

The STATA command to get the figure below is:

```
.graph stay, t1(Length of hospital stay in days for all 25 patients)
```

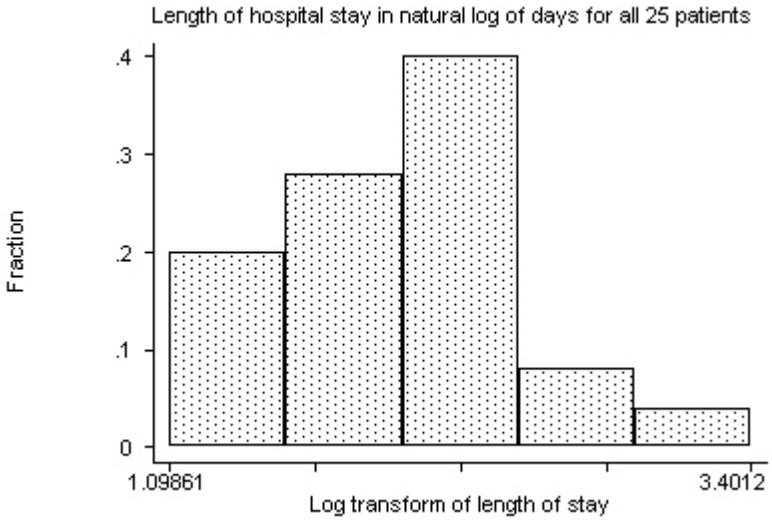


The figure below is the graph of the length of stay transformed using the natural log. It is much more normal looking.

The get the log transformation of the variable “stay”, I defined a new variable called logstay as follows: .gen logstay = log(stay)

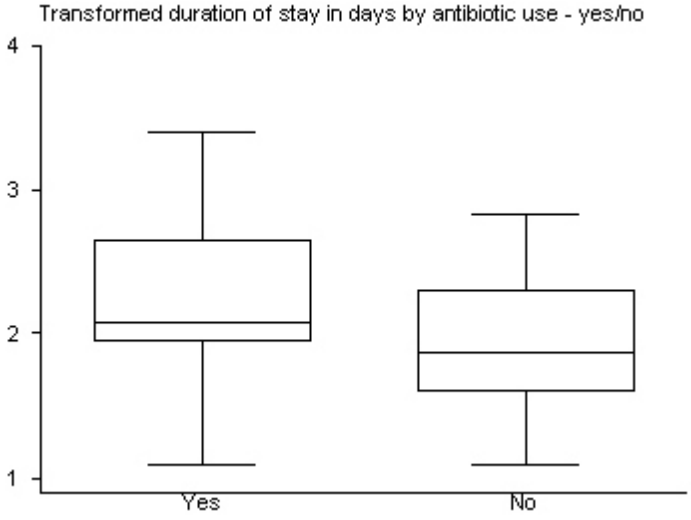
The graph below was obtained with the following command:

graph logstay, t1(Length of hospital stay in natural log of days for all 25 patients) ylab




The box-and-whisker plot of the transformed data shows variances that are much more similar than were the variances for the untransformed data.

.graph logstay,box t1(Transformed duration of stay in days by antibiotic use - yes/no)



## Test of equality of variances for the transformed data.

```
. sdtest logstay,by(antibio)

Variance ratio test

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     Yes |       7    2.234508     .266063    .7039364    1.583475    2.885541
      No |      18    1.892577     .117218    .4973138    1.645269    2.139886
---------+--------------------------------------------------------------------
combined |      25    1.988318    .1137752    .5688759    1.753498    2.223138
------------------------------------------------------------------------------

                         Ho: sd(Yes) = sd(No)

             F(6,17) observed   = F_obs           =     2.004
             F(6,17) lower tail = F_L   = 1/F_obs =     0.499
             F(6,17) upper tail = F_U   = F_obs   =     2.004

 Ha: sd(Yes) < sd(No)        Ha: sd(Yes) ~= sd(No)        Ha: sd(Yes) > sd(No)
  P < F_obs = 0.8785      P < F_L + P > F_U = 0.3211       P > F_obs = 0.1215
```

Below I have compared patients on antibiotics to those not taking antibiotics with respect to the length of stay in days and the transformed length of stay assuming first that the variances are equal and then that they are not.

Notice that for the untransformed data the answers you get assuming equality and inequality of variances are more different than those with the transformed data.

## Assumes variances are equal - data in original form

```
. ttest stay,by(antibio)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     Yes |       7    11.57143     3.32993    8.810167    3.423383    19.71947
      No |      18    7.444444    .8715632    3.697729    5.605607    9.283282
---------+--------------------------------------------------------------------
combined |      25         8.6    1.143095    5.715476    6.240767    10.95923
---------+--------------------------------------------------------------------
    diff |            4.126984    2.454132               -.9497745    9.203743
------------------------------------------------------------------------------
Degrees of freedom: 23

                    Ho: mean(Yes) - mean(No) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =   1.6816                 t =   1.6816                  t =   1.6816
   P < t =   0.9469             P > |t| =   0.1062              P > t =   0.0531
```

## Assumes variances are unequal - data in original form

```
. ttest stay,by(antibio) unequal

Two-sample t test with unequal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     Yes |       7    11.57143     3.32993    8.810167    3.423383    19.71947
      No |      18    7.444444    .8715632    3.697729    5.605607    9.283282
---------+--------------------------------------------------------------------
combined |      25         8.6    1.143095    5.715476    6.240767    10.95923
---------+--------------------------------------------------------------------
    diff |            4.126984    3.442101               -4.05132    12.30529
------------------------------------------------------------------------------
Satterthwaite's degrees of freedom:   6.8389

                    Ho: mean(Yes) - mean(No) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =   1.1990                 t =   1.1990                  t =   1.1990
   P < t =   0.8648             P > |t| =   0.2704              P > t =   0.1352
```

## Using the natural log of the variable length of stay and assuming variances equal

```
. ttest logstay,by(antibio)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     Yes |       7    2.234508     .266063    .7039364    1.583475    2.885541
      No |      18    1.892577     .117218    .4973138    1.645269    2.139886
---------+--------------------------------------------------------------------
combined |      25    1.988318    .1137752    .5688759    1.753498    2.223138
---------+--------------------------------------------------------------------
    diff |            .3419306    .2488347               -.1728232    .8566843
------------------------------------------------------------------------------
Degrees of freedom: 23

                    Ho: mean(Yes) - mean(No) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =   1.3741                 t =   1.3741                 t =   1.3741
   P < t =   0.9087           P > |t| =   0.1826           P > t =   0.0913
```

## Using the natural log of the variable length of stay and assuming variances are unequal

```
. ttest logstay,by(antibio) unequal

Two-sample t test with unequal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     Yes |       7    2.234508     .266063    .7039364    1.583475    2.885541
      No |      18    1.892577     .117218    .4973138    1.645269    2.139886
---------+--------------------------------------------------------------------
combined |      25    1.988318    .1137752    .5688759    1.753498    2.223138
---------+--------------------------------------------------------------------
    diff |            .3419306    .2907397               -.3224428    1.006304
------------------------------------------------------------------------------
Satterthwaite's degrees of freedom:  8.44295

                    Ho: mean(Yes) - mean(No) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =   1.1761                 t =   1.1761                 t =   1.1761
   P < t =   0.8642           P > |t| =   0.2717           P > t =   0.1358
```