BE540 - Introduction to Biostatistics
Computer Illustration

Topic 1 – Summarizing Data
Software: STATA

A Visit to Yellowstone National Park, USA

Source:
Chatterjee, S; Handcock MS and Simonoff JS *A Casebook for a First Course in Statistics and Data Analysis.* New York, John Wiley, 1995.

Setting:
Upon completion of BE540, you decide to take a vacation to the United States. Of particular interest is seeing an eruption of the famous "Old Faithful" geyser at Yellowstone National Park. Unfortunately, your time is limited and you do not wish to miss seeing an eruption.

This worked example illustrates descriptive analysis of a data set of 222 interval times between eruptions of the Old Faithful Geyser, measured during August 1978 and 1979.

Data File:
GEYSER1.DAT - This is a data set in ASCII format.

Description of Data:
There are three variables, in the following order:

INDEX - An index of the date of the eruption. We will not be using this variable.

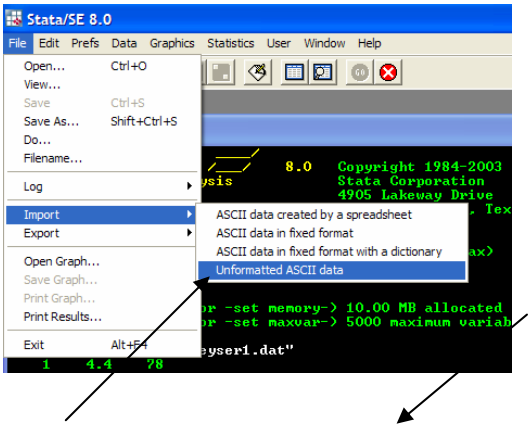
DURATION - The duration of the eruption in minutes.

INTERVAL - The length of the interval between the current eruption and the next eruption.

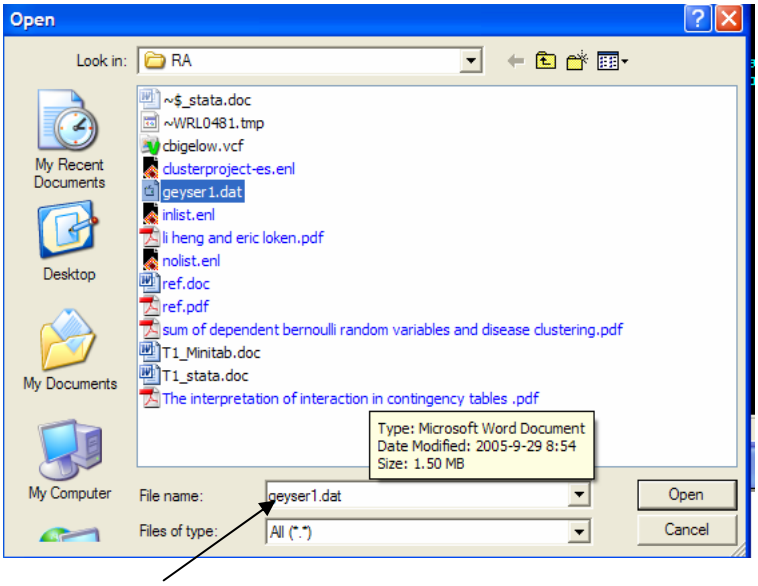
Objective:
Describe the pattern of eruptions and predict the interval of time to the next eruption.

1. Read in the ASCII format data 'GEYSER1.DAT':

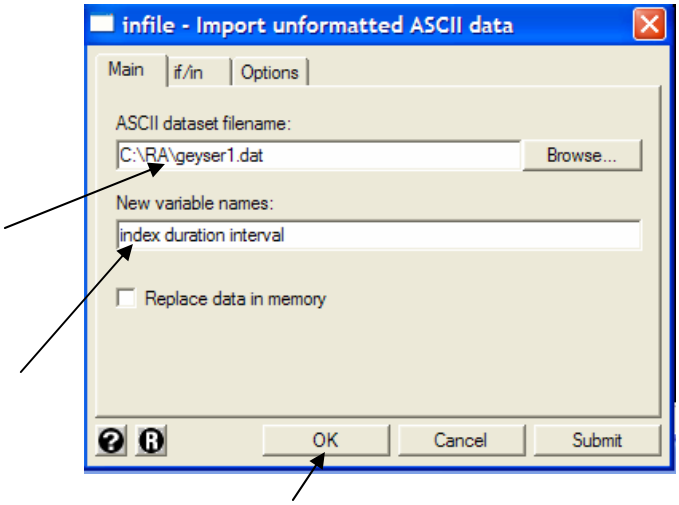
1.
File > Import Files > Unformatted ASCII data



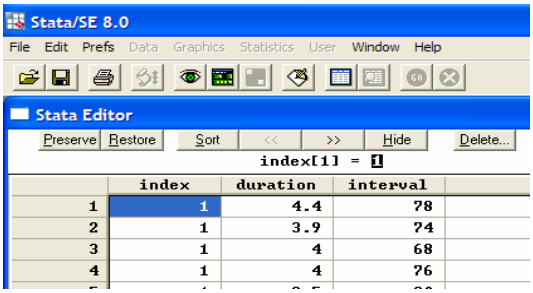
2. Choose correct directory where you save the data, select the .DAT file "GEYSER". Click OPEN.



2. Input variable names**: Index, Duration, Interval** in the "infile-Import unformatted ASCII data" frame. Click **OK**.

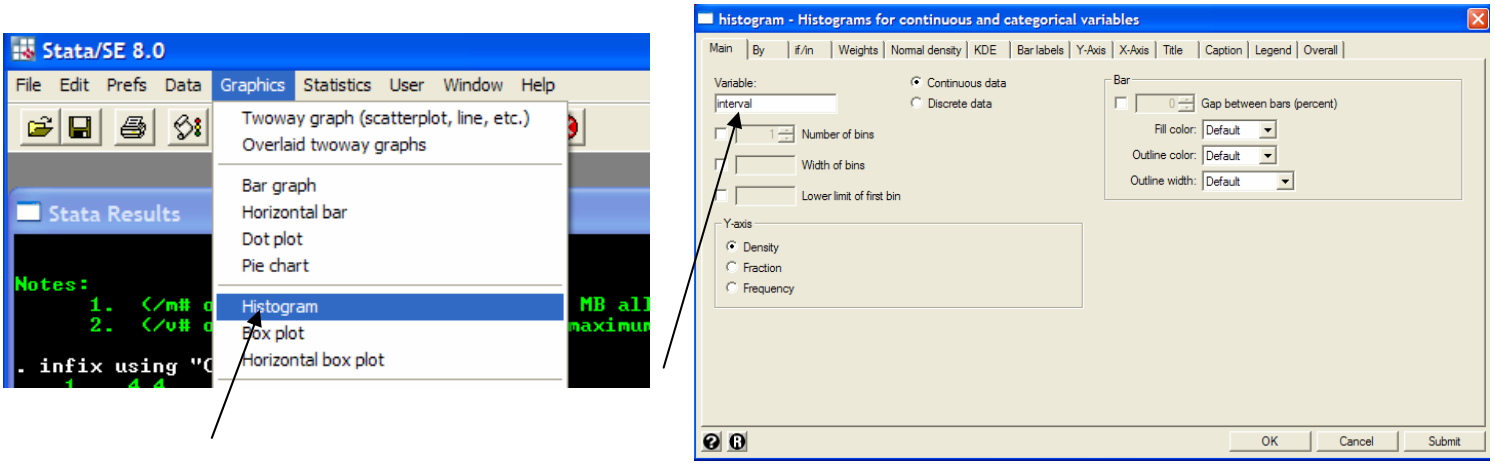


Data set (222 observations) is imported to STATA. You should see the following worksheet by click **DATA>Data eidtor:**

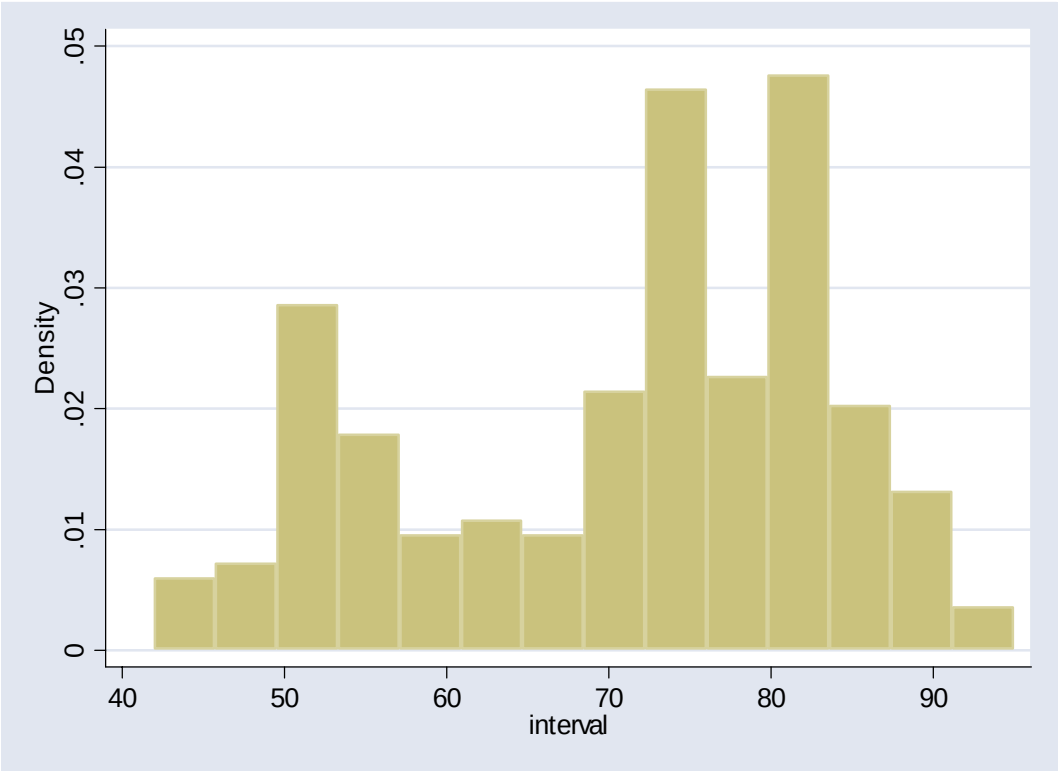


2. Obtain a Histogram of Interval Times.

Menu > **Graphics**> **HISTOGRAM**… > enter variable name, **OK** > Click **OK**

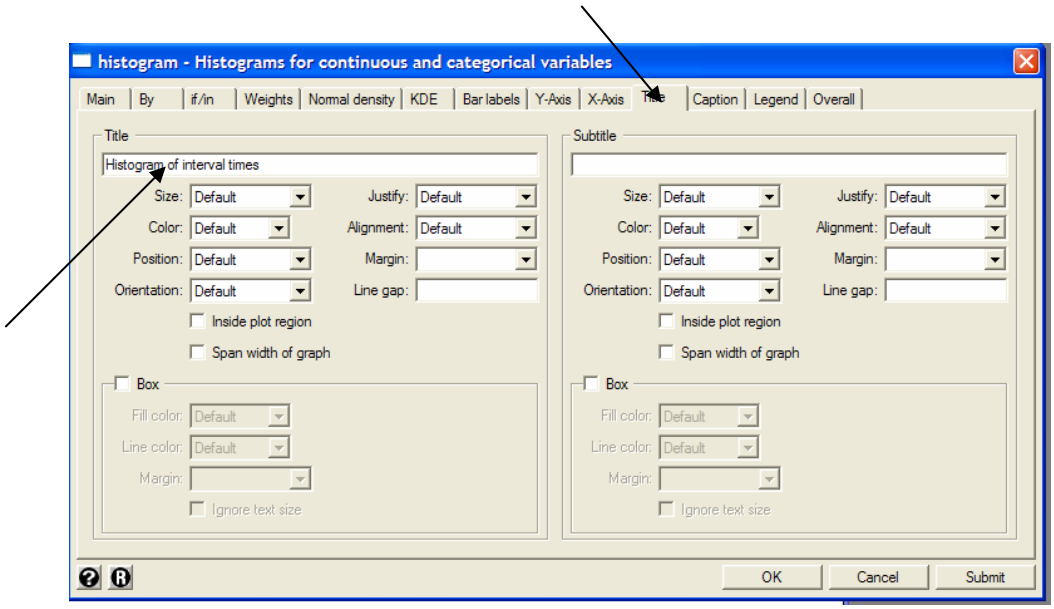


You should see:

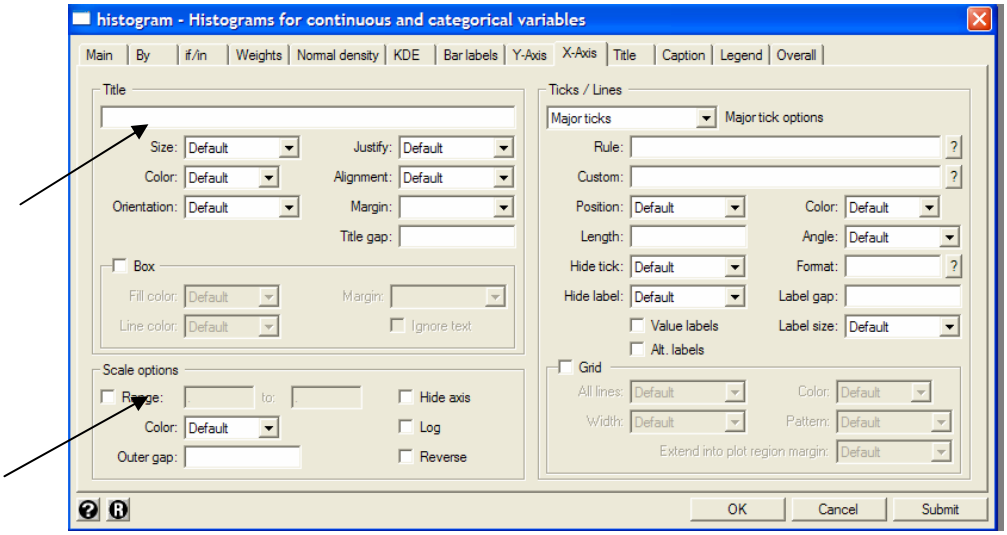


This histogram is generated with STATA default settings. If you like, you can revise the format by click **Graphics**> **HISTOGRAM**…. . The following are some examples of how to change graph options.

Click tab “title”, change plot title:

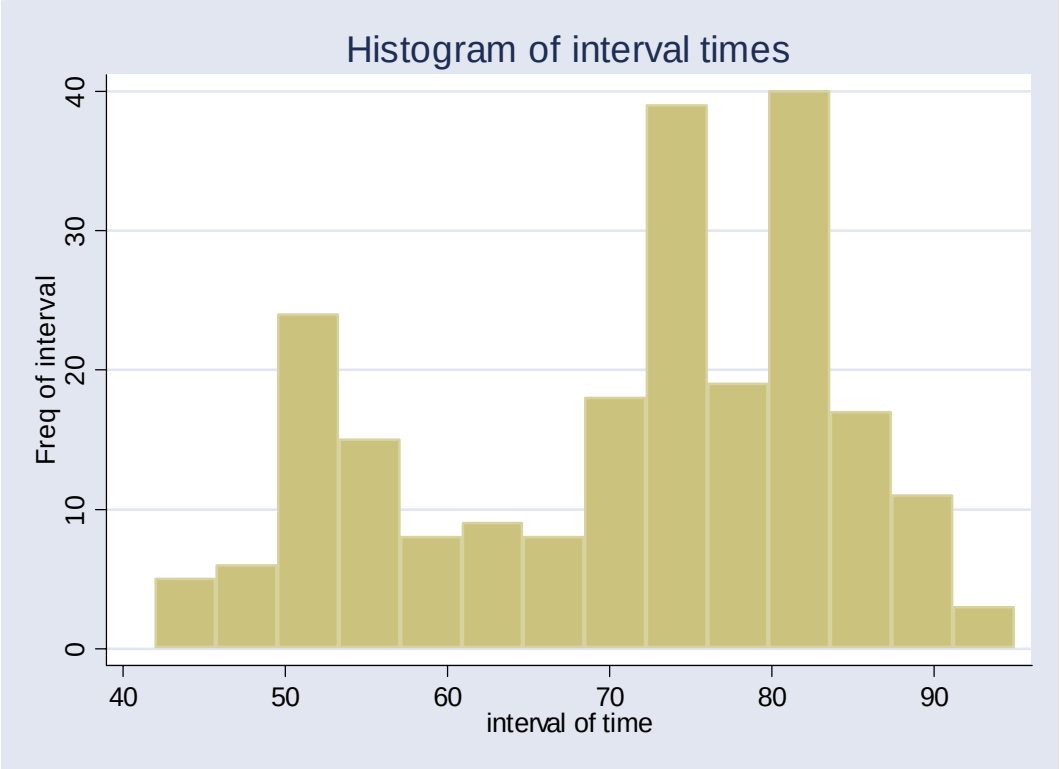


click on X-axis tab, you can change X scale such as label, range, etc.



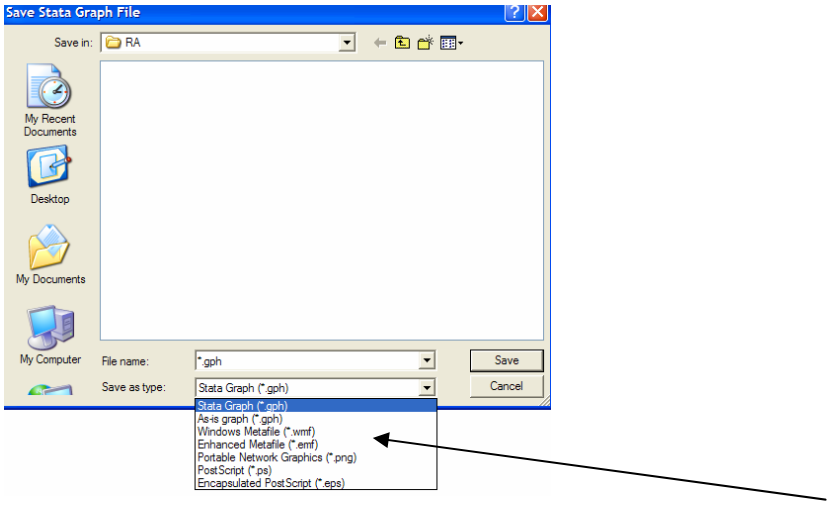
Change Y-axis setting by the same way.

Finally, the revised histogram looks like:



3. Save this histogram as a picture that you can print directly or that you can insert into a document such as this one.

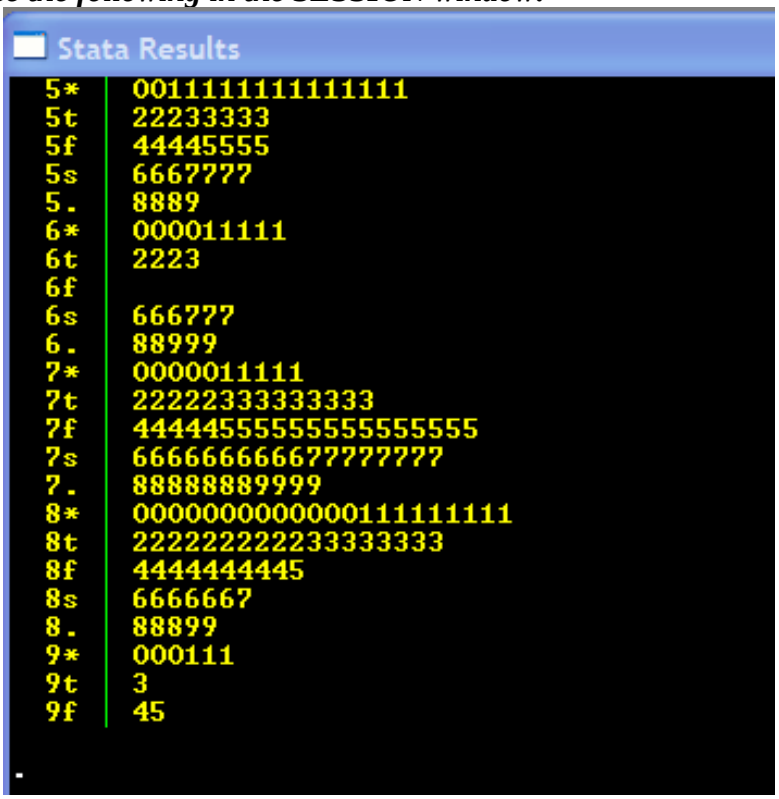
Go to **Menu** > **FILE** > **SAVE GRAPH …** > Type file name and select picture format > **SAVE**



4. Instead of a histogram, we might have constructed a stem-and-leaf diagram.

Put "stem interval" in "Stata command" window and return.

You should see the following in the SESSION window:



Remarks.

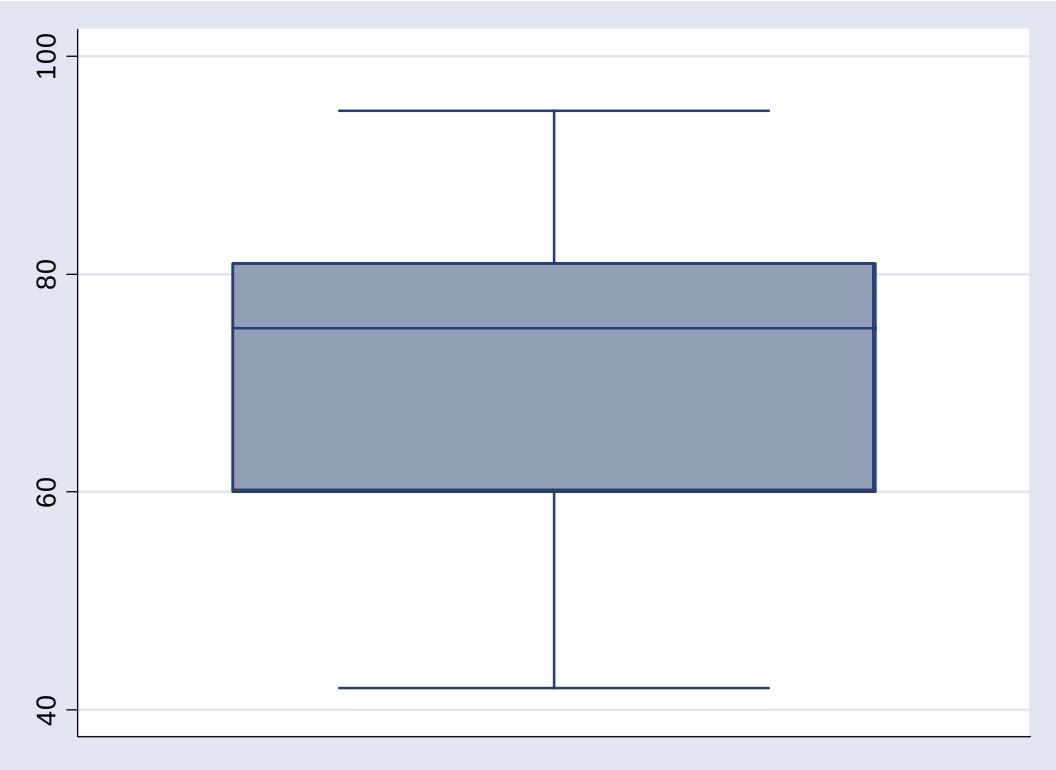
You can see that a stem and leaf diagram is very similar to a histogram. However, we can also see that the minimum and maximum interval times are 42 and 95 minutes, respectively, and that the median time is 75 minutes.

The column of numbers to the left of the stem and leaf diagram is
from the bottom - a cumulative frequency from the bottom and up.
from the top - a cumulative frequency from the top and down.

5. In this example, a Box and Whisker plot is not very informative. Let's see why.

Menu > Graphics > Boxplot …, OK > Choose variable "Interval", **SELECT, OK.**

You should see



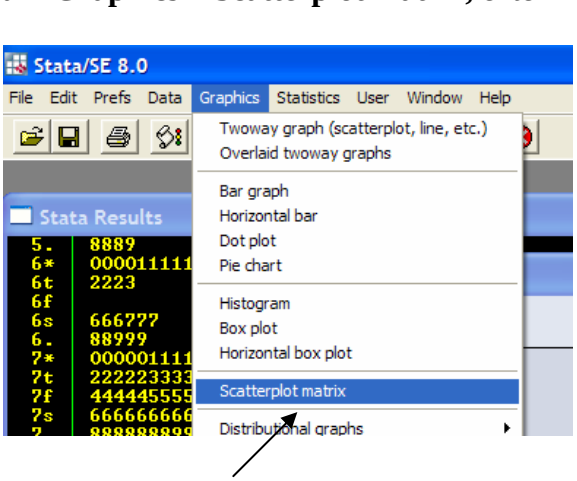
Remarks:

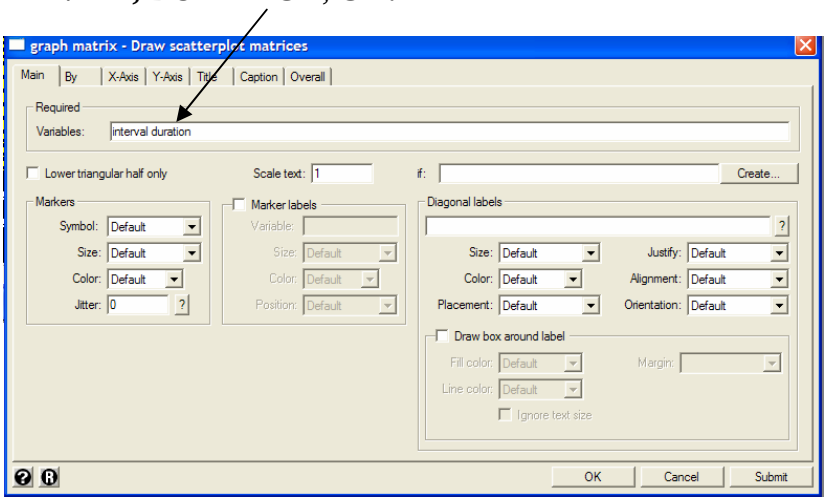
Both the histogram and stem and leaf summaries suggested that there are two groups of interval times. This cannot be seen in a Box and Whisker plot.

Box and Whisker plots are excellent for summarizing the distribution of ONE population. They are not informative when the sample being summarized actually represents MORE THAN ONE population.

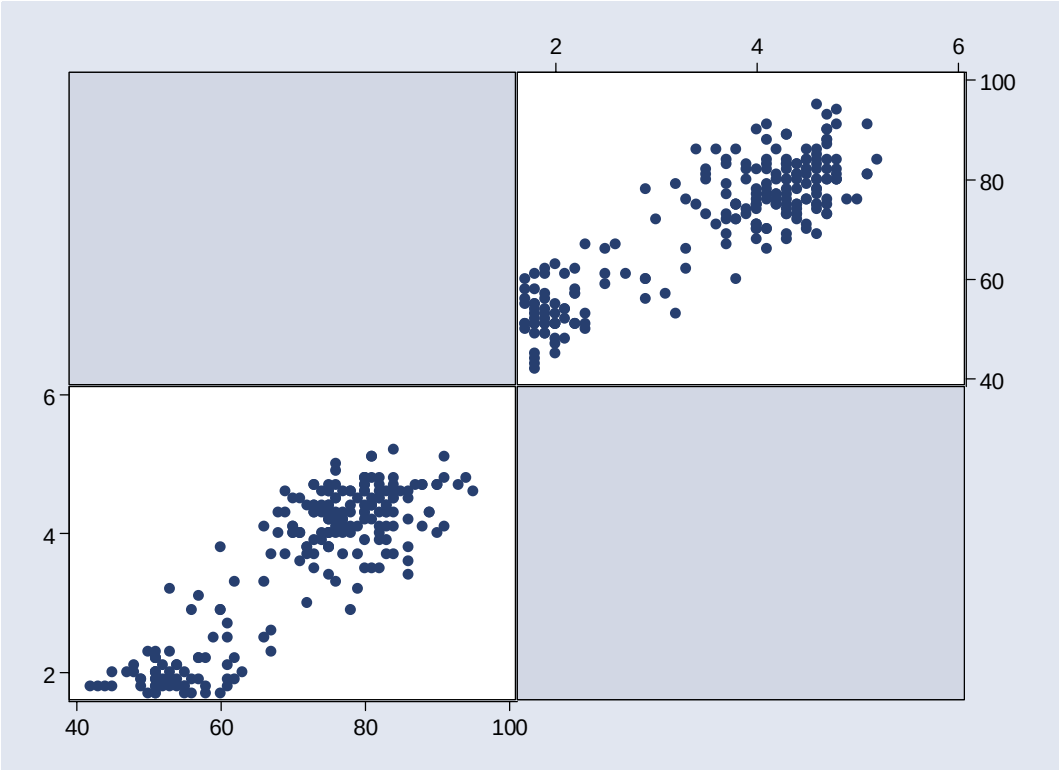
6. We have information on duration of eruption also. One possibility is that the duration of the current eruption is a predictor of the interval time to the next eruption. To investigate this possibility, construct a scatter plot of interval time versus duration. Plot the predictor DURATION on the horizontal axis (X) and the outcome INTERVAL time to the next eruption on the vertical axis (Y).

Menu > Graphics > Scatterplot matrix, enter **INTERVAL , DURATION**, **OK.**





You should see

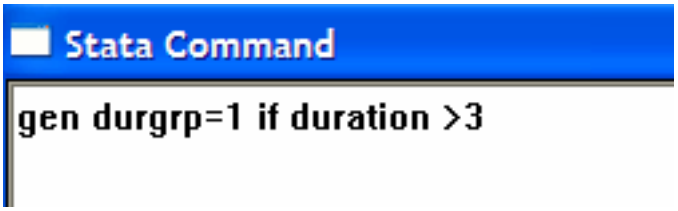


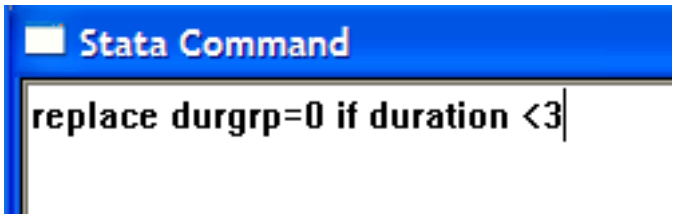
Remarks

The scatter plot confirms a suspected positive association. Longer duration times appear to predict longer intervals to the next eruption. Interestingly, the scatter plot still suggests that there are two distinct subgroups, distinguished by durations of less than versus greater than three minutes.

7. Create a grouped measure of duration and construct separate box and whisker plots of interval times for the interval times that follow eruptions less than 3 minutes in duration and the interval times that follow eruptions greater than 3 minutes in duration.

Enter "gen durgrp=1 if duration >3" and "replace durgrp=0 if duration <3".

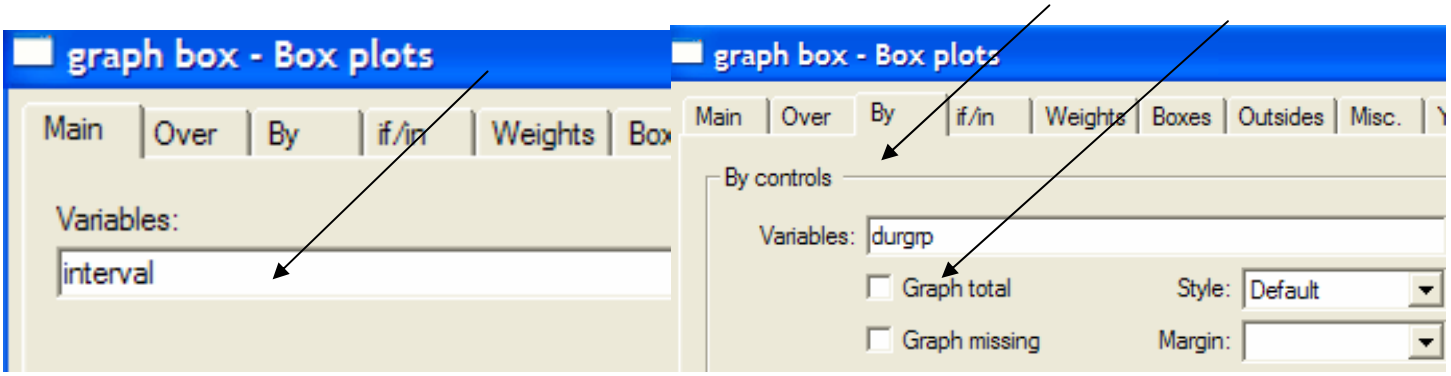




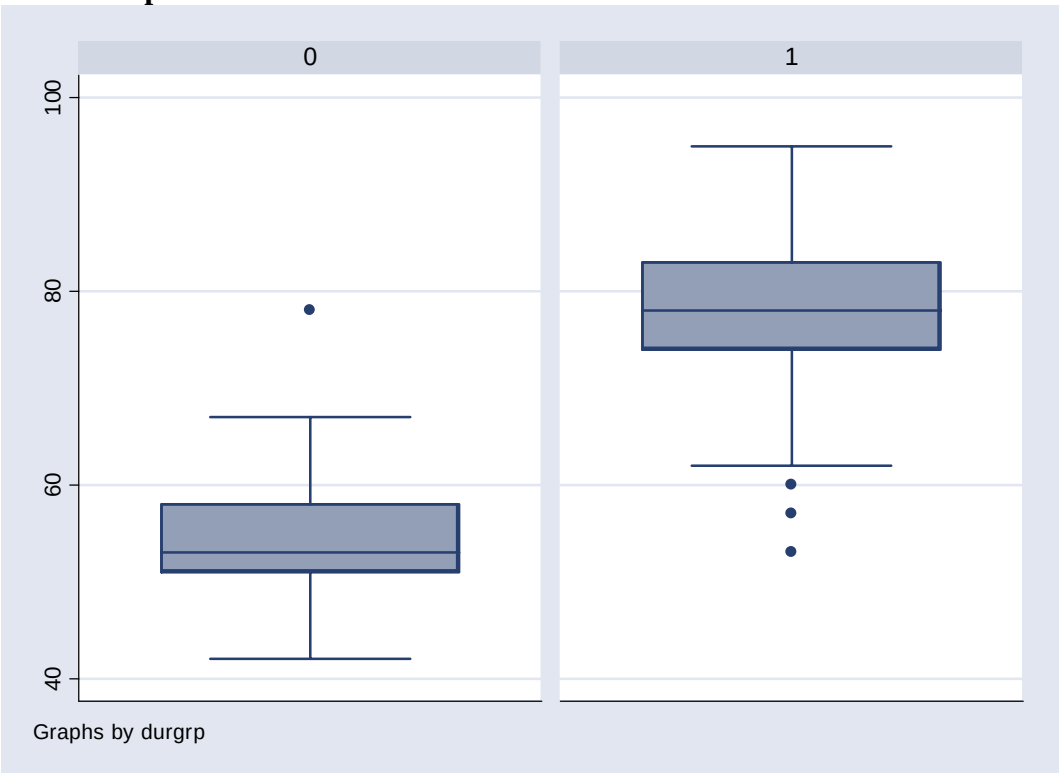
This function is to create new variables from existing variable. "Duration" is the existing, and "Durgrp" is the new variable.
If Duration > 3 minutes, Duration group is 1, otherwise, Duration group is 0.

Note: You have just created what is called an indicator variable to indicate a duration time that is greater than 3 minutes. It is equal to 0 for all durations less than 3 minutes and is equal to 1 for all durations greater than 3 minutes. Indicator variables are also called dummy variables or design variables.

Menu>Graphics > BoxPlot… >entere **Interval,** Click **by**>Input **"Durgrp", OK.**

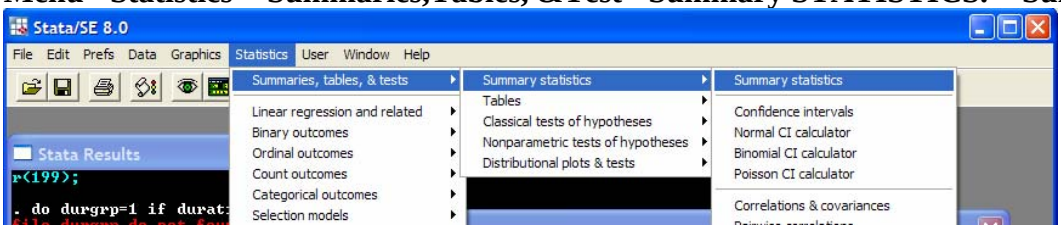


Separate box and whisker plots:



10. Finally, let's look at some numerical summaries, classified by the two groups.

Menu> Statistics > Summaries,Tables, &Test> Summary STATISTICS. > Summary STATISTICS, OK.



Stata Results

```
. bysort durgrp: summarize interval

-> durgrp = 0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    interval |         67    54.46269    6.298938        42         78

-> durgrp = 1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    interval |        154     78.2013    6.895466        53         95

-> durgrp = .

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    interval |        154     78.2013    6.895466        53         95

-> durgrp = .

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    interval |          1          72           .        72         72

--more--
```

So, what should you do? If you arrive to Old Faithful just after an eruption of less than 3 minutes, with 95% confidence, your waiting time to the next eruption will be between 53 and 56 minutes. Alternatively, if you arrive just after an eruption of greater than 3 minutes, with 95% confidence, your waiting time to the next eruption will be between 77 and 79 minutes.