PubHlth 540 - Introduction to Biostatistics

1. Summarizing Data
Illustration: Stata (version 10, 11 or 12)

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 PubHlth540, 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:
Geyser.dta – This is a stata data set.

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.

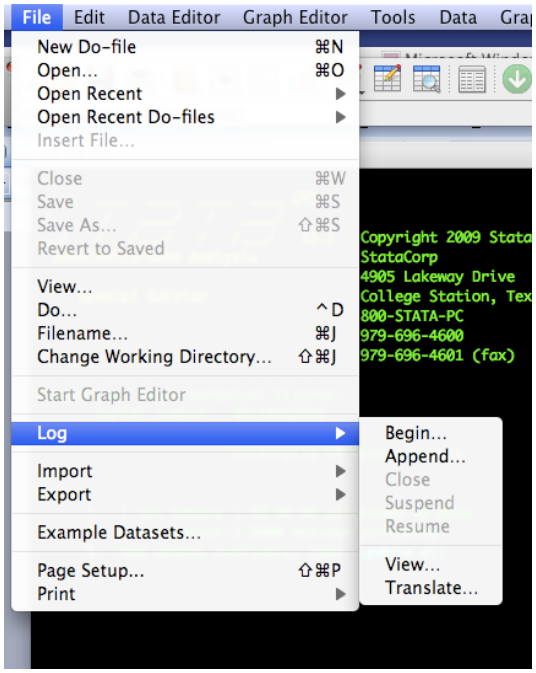
Objective:
Describe the pattern of eruptions and predict the interval of time to the next eruption.

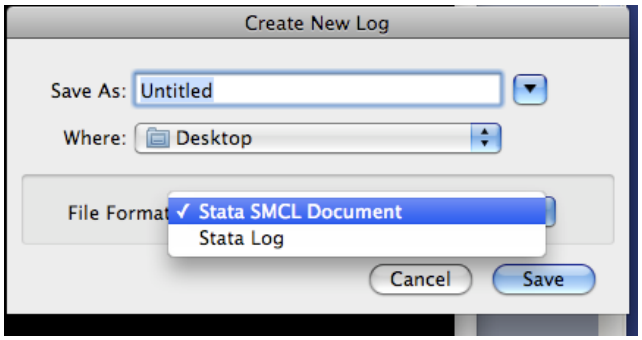
Key:

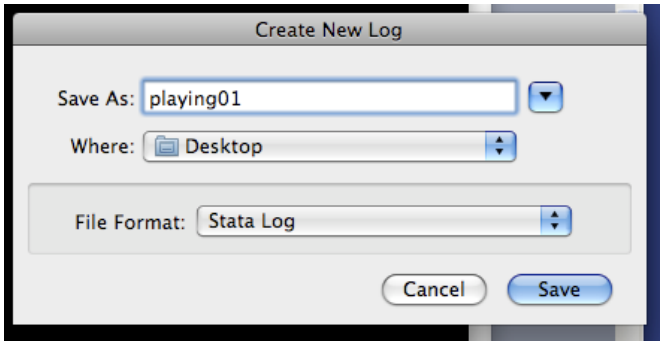
*** COMMENT – bold green**
. stata syntax – bold black
Stata output - blue

Before you launch Stata:
`Be sure that you have ALREADY downloaded geyser.dta from the course website`

1. Start a log of your session, using menu bar at top: FILE > LOG > BEGIN
Tip: At file format: From drop down menu, choose STATA LOG







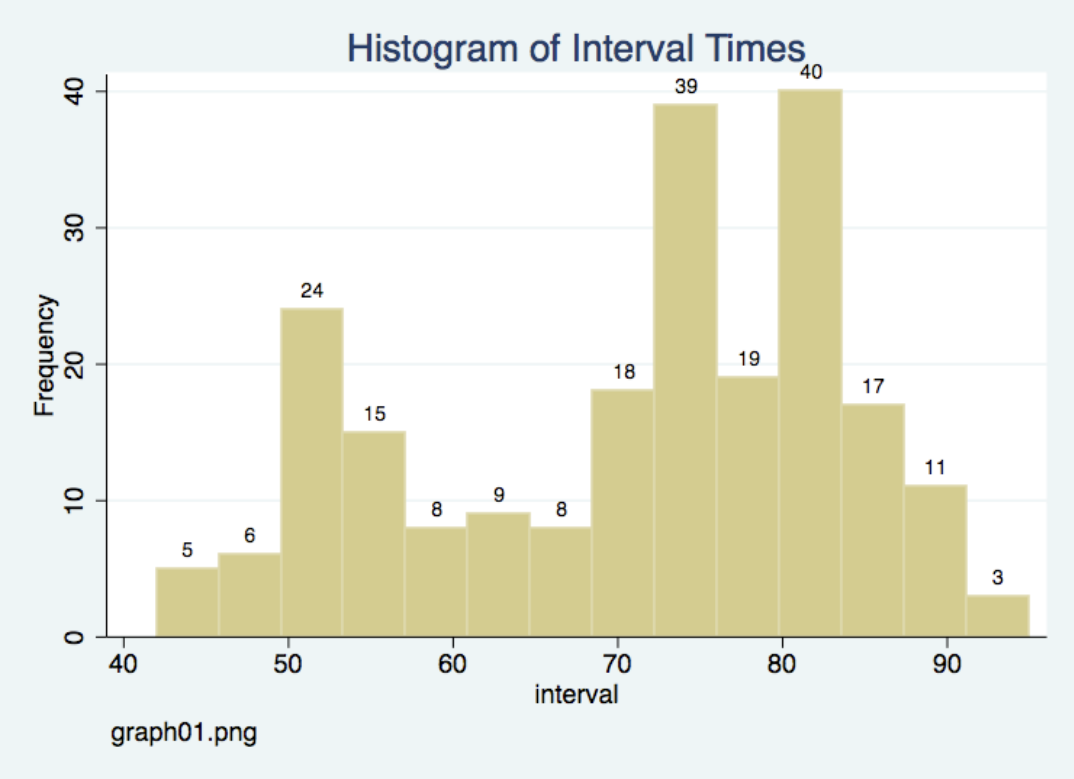
Preliminaries.

```
. *_____ Turn OFF screen by screen pause, use:  SET MORE OFF _________*
. set more off

. *_____________ Import data from desktop into Stata using menu bar at top:  FILE > OPEN ________*
. use "/Users/carolbigelow/Desktop/1.  Teaching/web540/statadata/geyser.dta"

. *___________ Compact description of variables in data set:  CODEBOOK, COMPACT
. codebook, compact

Variable   Obs Unique      Mean  Min  Max  Label
-------------------------------------------------------------------------------------------index
 222     16   12.2973    1   23
duration   222     34  3.576126  1.7  5.2
interval   222     50  71.00901   42   95
-------------------------------------------------------------------------------------------

. *_____________   Attach labels to variable names using:  LABEL VARIABLE ________*
. label variable index "Eruption Date ID"
. label variable duration "Duration eruption, minutes"
. label variable interval "Inter-Eruption Wait, minutes"
. codebook, compact

Variable   Obs Unique      Mean  Min  Max  Label
-------------------------------------------------------------------------------------------
index      222     16   12.2973    1   23  Eruption Date ID
duration   222     34  3.576126  1.7  5.2  Duration eruption, minutes
interval   222     50  71.00901   42   95  Inter-Eruption Wait, minutes
-------------------------------------------------------------------------------------------

. *_______ Detailed coding manual for one variable, using:  CODEBOOK ____*
. codebook interval
-------------------------------------------------------------------------------------------
interval                                                          Inter-Eruption Wait,
minutes
-------------------------------------------------------------------------------------------

                  type:  numeric (float)

                 range:  [42,95]                     units:  1
         unique values:  50                       missing .:  0/222

                  mean:   71.009
             std. dev:   12.7992

           percentiles:        10%       25%       50%       75%       90%
                                51        60        75        81        84
```

Obtain a Histogram of Interval Times.

```
. *_____________  Histogram of Interval Times  _________*
. histogram interval, frequency addlabels title("Histogram of Interval Times")
caption(“graph01.png”)
(bin=14, start=42, width=3.7857143)
```

You should see



Remarks
The interval times are in the range of 40 to 100 minutes, approximately.
There appears to be two groupings of interval times.
They are centered at 55 and 80 minutes, approximately.
Interestingly, there is a gap in the middle.

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

```
. * To save the graph, click on the SAVE icon in the graph itself. Save with extension .png
```

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

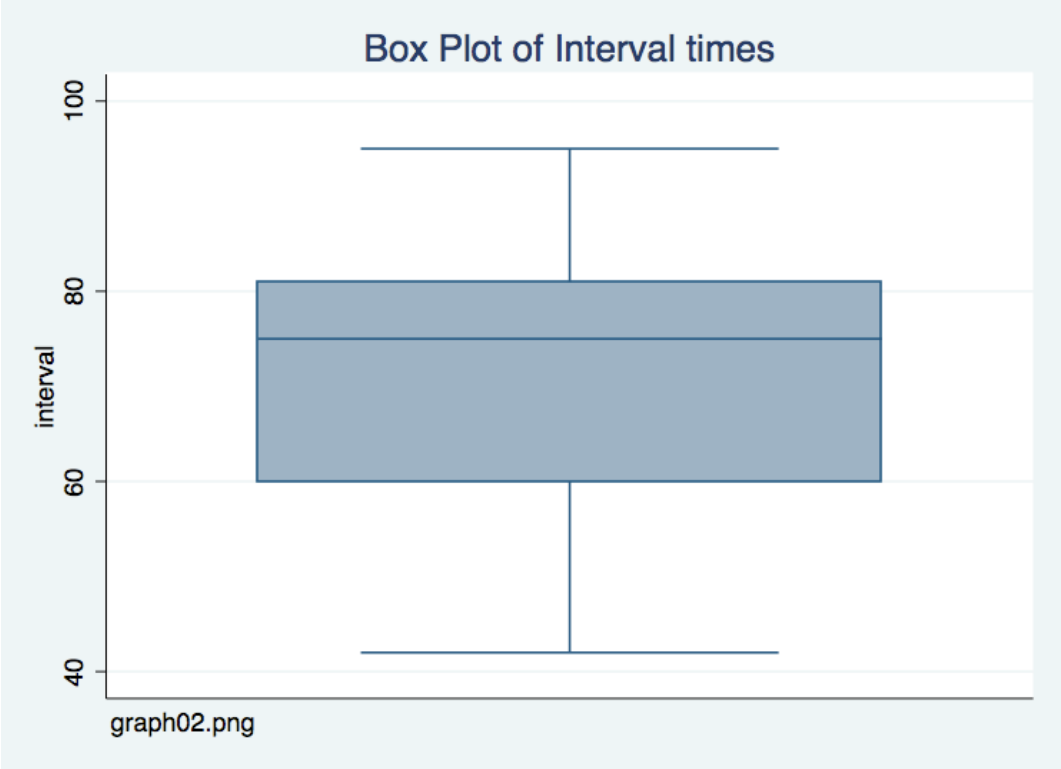
```
. *_______ Stem and Leaf of Interval times, use:  STEM with option LINES ____*
. stem interval, lines(1)
Note -  lines(1) tells stata to provide 1 row for units place values of 0, 1, 2,3, 4, 5, 6, 7,8, and 0
Stem-and-leaf plot for interval (Inter-Eruption Wait, minutes)

  4* | 23455788999
  5* | 00111111111111112223333344445555666777788889
  6* | 000011111222366677788999
  7* | 000001111122222333333333444445555555555555566666666667777777778 ... (73)
  8* | 000000000000011111111122222222223333333344444444456666667888899
  9* | 000111345

. stem interval, lines(2)
 Note -  lines(2) tells stata to provide 2 rows for units place values.
         One row will be for units values of 0 -4.  Second rol will be for units 5-9
Stem-and-leaf plot for interval (Inter-Eruption Wait, minutes)

  4* | 234
  4. | 55788999
  5* | 00111111111111112222333334444
  5. | 555566677778889
  6* | 0000111112223
  6. | 66677788999
  7* | 000001111122222333333333444444
  7. | 55555555555555566666666667777777788888889999
  8* | 000000000000011111111122222222223333333344444444
  8. | 5666666788899
  9* | 00011134
  9. | 5
```

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.

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

```
. *_____  Box Plot of Interval Times, use BOX _____*
. graph box interval, title("Box Plot of Interval Times")
 caption("graph02.png")
```



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.

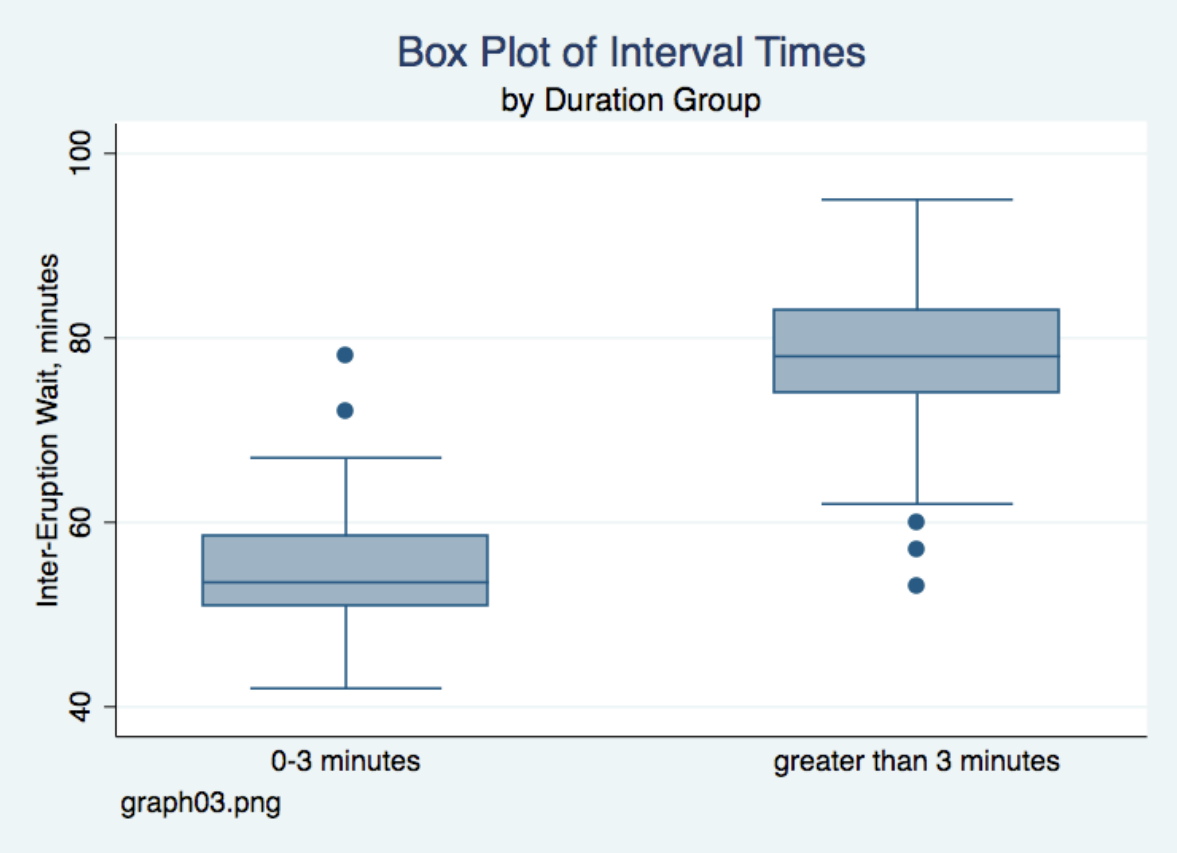
Box and Whisker plots are excellent for summarizing the distribution of ONE population. They are not informative when the sample being summarized actually represents MORE THAN ONE population.

6. 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.

```
. *________  Create grouped measure DURGRP from DURATION, using:  GENERATE and REPLACE ______*
. generate durgrp=1 if duration > 3
(68 missing values generated)
. replace durgrp=0 if duration <= 3
(68 real changes made)
```

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.

```
. * (optional) Label the variable and label the variable values *
. label variable durgrp "Duration Eruption, Grouped"
. label define durgrpf 0 "0-3 minutes" 1 "greater than 3 minutes"
. label values durgrp durgrpf

. *_____  Side by side box plot over durgrp, use:  GRAPH BOX with option OVER for sorted data__*
. *_____  To sort data, use:  SORT _______*
. sort durgrp
. graph box interval, over(durgrp) title("Box Plot of Interval Times") subtitle("by Duration Group")
caption("graph03.png")
```



10. Finally, let's look at some numerical summaries.

```
. * __  Descriptive Statistics for continuous variable INTERVAL, use SUMMARIZE w option DETAIL
. summarize interval, detail

                Inter-Eruption Wait, minutes
-------------------------------------------------------------
      Percentiles      Smallest
 1%           44              42
 5%           50              43
10%           51              44       Obs                 222
25%           60              45       Sum of Wgt.         222

50%           75                       Mean           71.00901
                      Largest          Std. Dev.      12.79918
75%           81              91
90%           84              93       Variance       163.8189
95%           88              94       Skewness      -.4822301
99%           93              95       Kurtosis       2.107263
```

```
. *_____ Descriptive statistics, nicely tabulated, by DURGRP, use: TABLE  _____*
. table durgrp, contents(n interval mean interval sd interval min interval max interval)

-------------------------------------------------------------------------------------
Duration Eruption,     |
Grouped                |  N(interval)  mean(inte~l)  sd(interval)  min(inter~l)  max(inter~l)
-----------------------+-------------------------------------------------------------
          0-3 minutes  |           68      54.72059      6.603585           42            78
greater than 3 minutes |          154       78.2013      6.895466           53            95
-------------------------------------------------------------------------------------


. *_____  Descriptive statistics for 2 continous variables, using:  TABSTAT ____*
. tabstat interval duration, stat(n mean sd min max) col(stat) format(%8.2f)

    variable |         N      mean        sd       min       max
-------------+--------------------------------------------------
    interval |    222.00     71.01     12.80     42.00     95.00
    duration |    222.00      3.58      1.08      1.70      5.20
----------------------------------------------------------------

. *_____ Frequency table for discrete variable, using: TAB _____*
. tab durgrp

    Duration Eruption, |
               Grouped |      Freq.     Percent        Cum.
-----------------------+-----------------------------------
          0-3 minutes  |         68       30.63       30.63
greater than 3 minutes |        154       69.37      100.00
-----------------------+-----------------------------------
                 Total |        222      100.00
```

So, what should you do? If you arrive to Old Faithful just after an eruption of less than 3 minutes, your waiting time to the next eruption will be between 42 and 78 minutes. Alternatively, if you arrive just after an eruption of greater than 3 minutes, your waiting time to the next eruption will be between 53 and 95 minutes.