# BE540 - Introduction to Biostatistics
# Computer Illustration

## Topic 1 – Summarizing Data
## Software: SAS

## 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 data.

```
data temp;      /* TEMP is a temporary SAS data set */
  infile 'z:\bigelow\teaching\web540\datasets\geyser.txt';
  input index dur interval;
run;
```

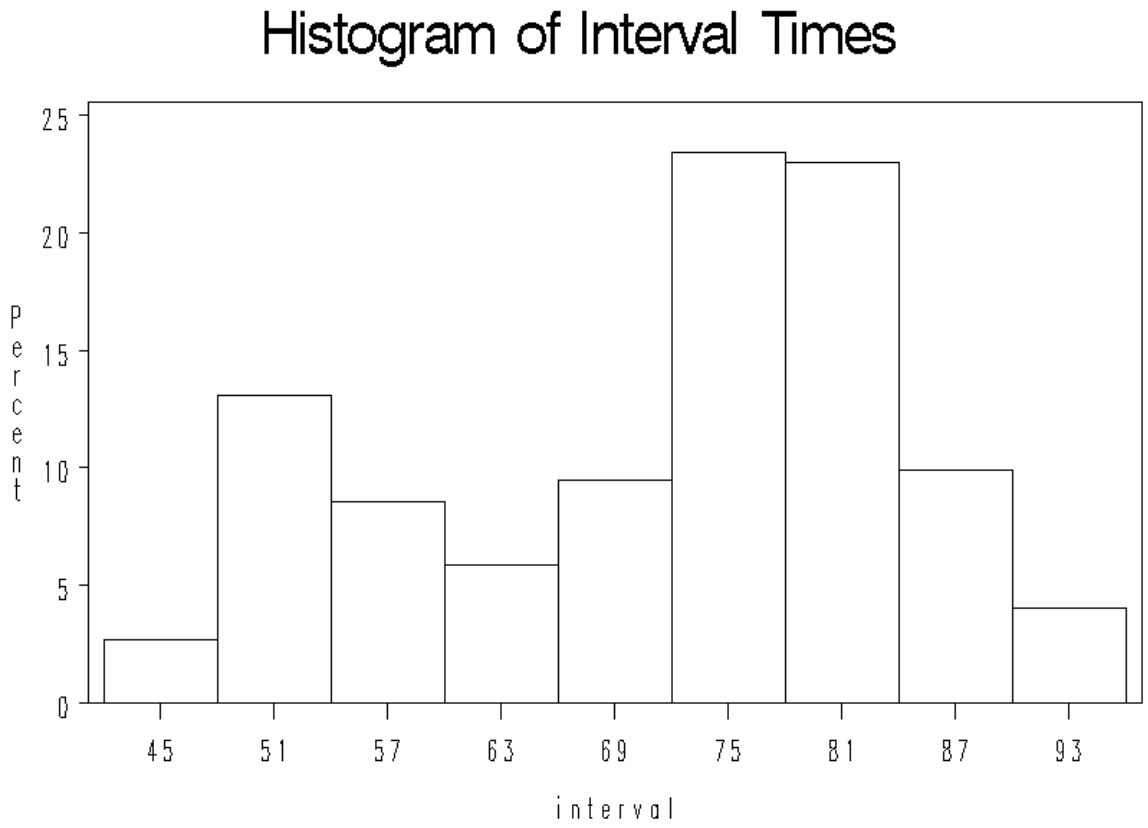
***You should see something like:***

```
NOTE: Copyright (c) 2002-2003 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) 9.1 (TS1M2)
      Licensed to UNIVERSITY OF MASSACHUSETTS COMPUTING SERVICES, Site 0011117001.
NOTE: This session is executing on the XP_PRO  platform.



NOTE: SAS initialization used:
      real time           1.03 seconds
      cpu time            0.85 seconds

1    data temp;      /* TEMP is a temporary SAS data set */
2      infile 'z:\bigelow\teaching\web540\datasets\geyser.txt';
3      input index dur interval;
4    run;

NOTE: The infile 'z:\bigelow\teaching\web540\datasets\geyser.txt' is:
      File Name=z:\bigelow\teaching\web540\datasets\geyser.txt,
      RECFM=V,LRECL=256

NOTE: 222 records were read from the infile
'z:\bigelow\teaching\web540\datasets\geyser.txt'.
      The minimum record length was 17.
      The maximum record length was 19.
NOTE: The data set WORK.TEMP has 222 observations and 3 variables.
NOTE: DATA statement used (Total process time):
      real time           0.05 seconds
      cpu time            0.04 seconds
```

## 2. Obtain a Histogram of Interval Times.

```
proc capability data=temp;
  histogram interval;
  title 'Histogram of Interval Times';
run;
```

***You should see***

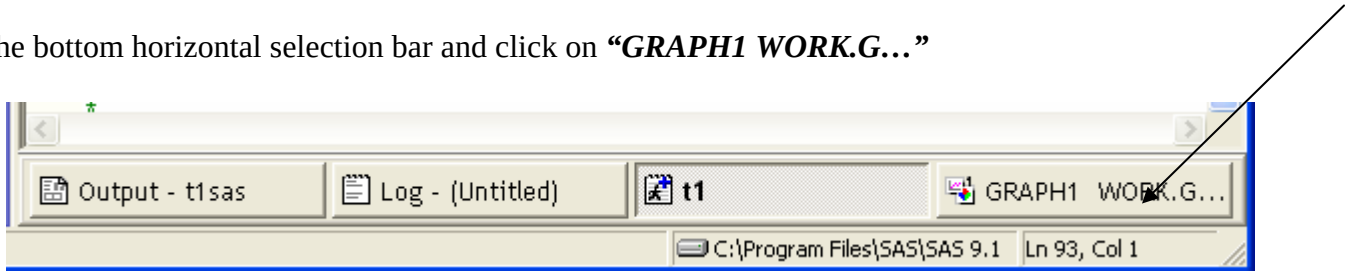



***Remarks***

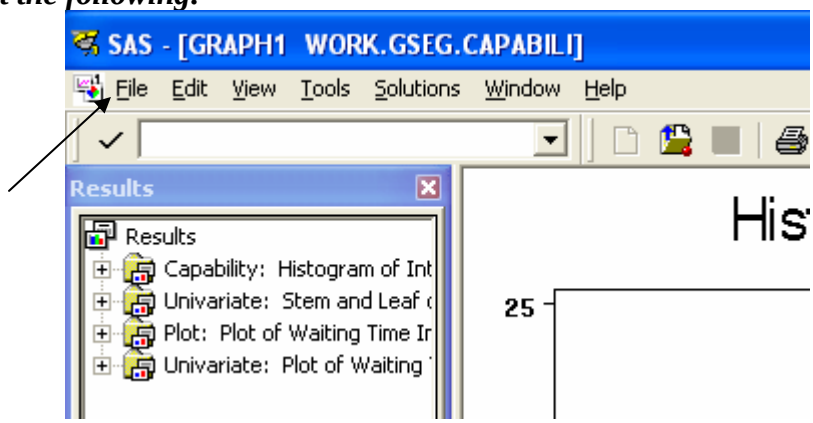
*The interval times are in the range of 40 to 100 minutes, approximately.*
*There appears to be two groupings of interval times.*
*They are centered at 55 and 80 minutes, approximately.*
*Interestingly, there is a gap in the middle.*

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

Locate the bottom horizontal selection bar and click on ***"GRAPH1 WORK.G…"***

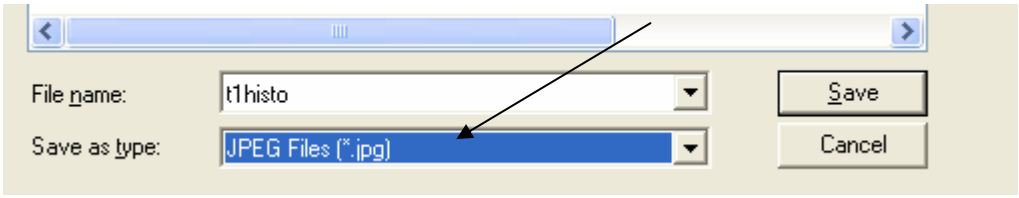



***You will then see at left the following:***




Locate the top horizontal selection bar and click on ***"FILE"***
From the drop down menu, ***"EXPORT AS IMAGE"***

From the "EXPORT AS IMAGE" screen, locate at the bottom *Save as Type.* Select ***".jpg"***
Next to *File name* provide a file name. Then click on SAVE

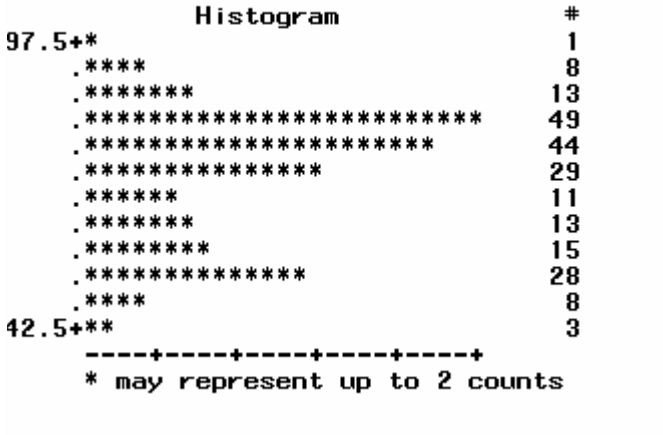



**4. Instead of a histogram, we might have constructed a stem-leaf diagram. SAS calls it "Histogram" and it is not as clear a plot as the stem and leaf that is produced in STATISTIX.**

```
proc univariate plot;
  var interval;
  title 'Stem and Leaf of Interval Times';
run;
```

This instruction produces a lot of output, in the middle of which is buried the stem and leaf plot. With a little scrolling down you will find the following in the midst of things:

***You should see***



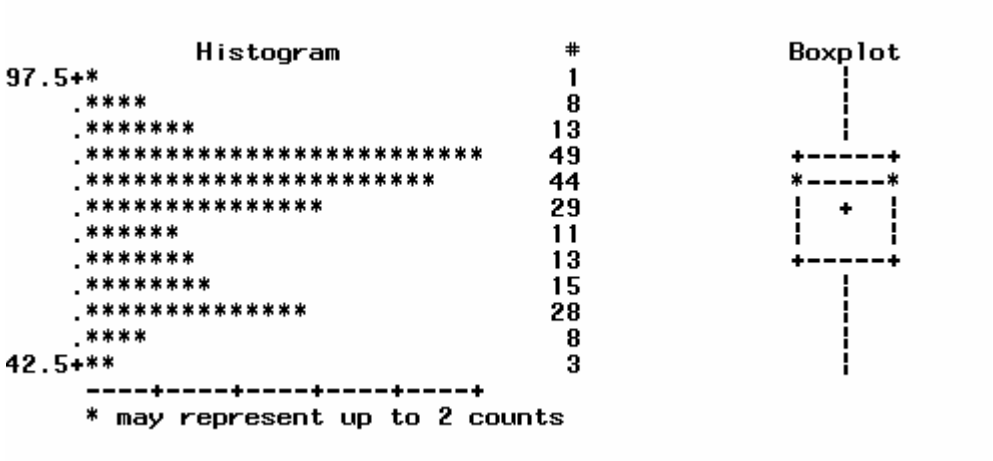

***Remarks.***

*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 rightt of the stem and leaf diagram is*
*from the bottom - a cumulative frequency from the bottom and up.*
*rom 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.**

The same SAS instruction (a PROC UNIVARIATE with the option PLOT) illustrated above also yields a Box and Whisker plot. Notice that the result in SAS is, again, not as clear as the Box and Whisker produced by STATISTIX. ***You should see***




***Remarks:***

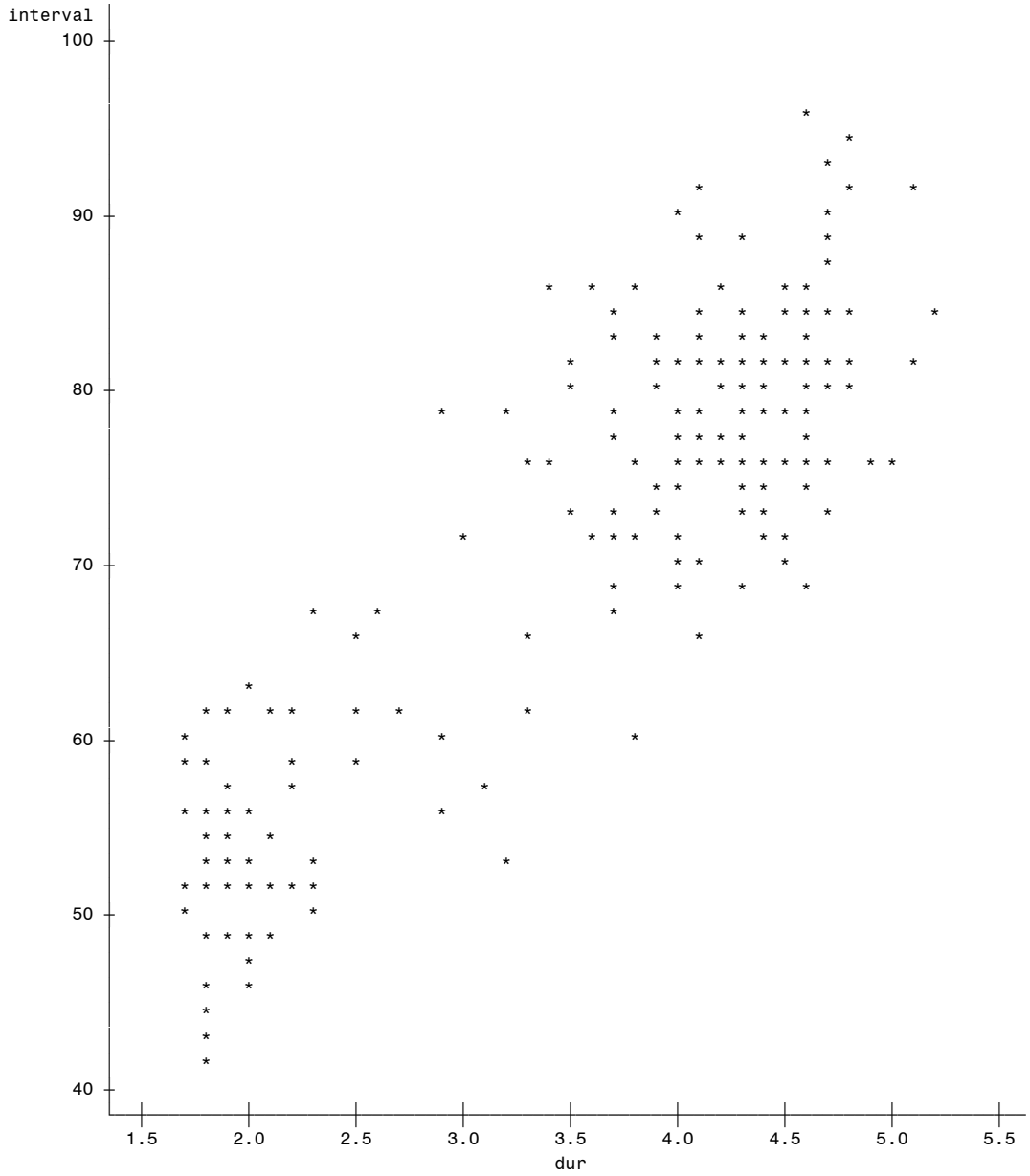
*Both the histogram and stem and leaf summaries suggested that there are two groups of interval times. This cannot be seen in a Box and Whisker plot.*

*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 DUR on the horizontal axis (X) and the outcome INTERVAL time to the next eruption on the vertical axis (Y).**

```
proc plot;
  plot interval*dur = '*';    /* use asterisk symbol in plot */
  title 'Plot of Waiting Time Intervals by Duration';
run;
```

***You should see***



***Remarks***

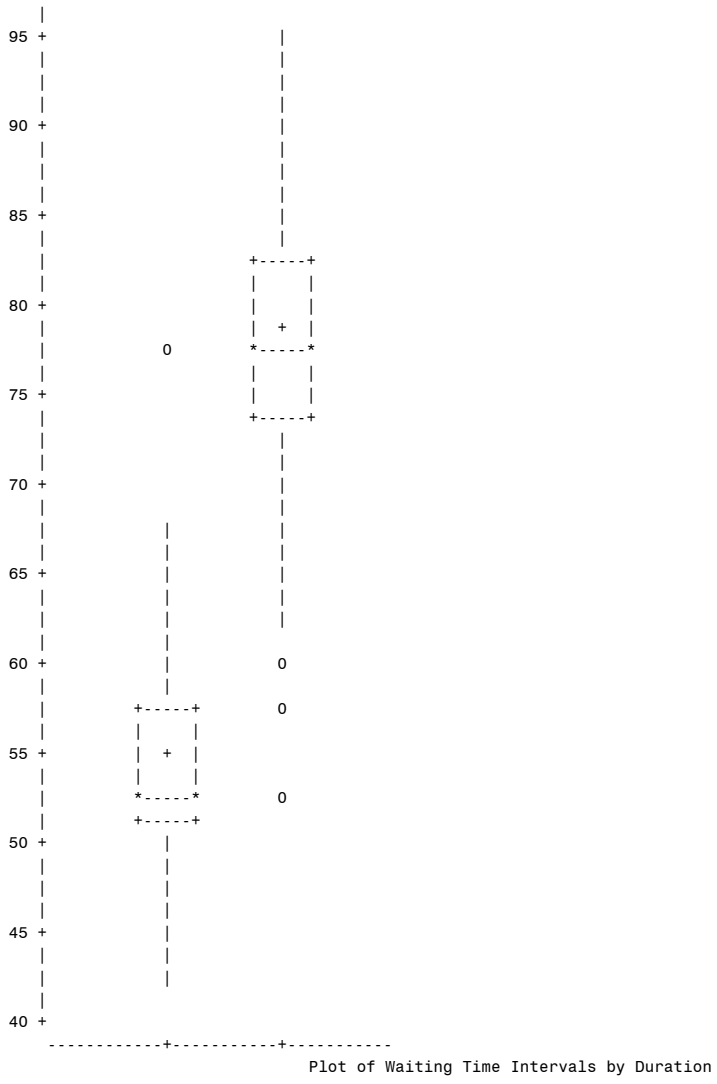
*The scatter plot confirms a suspected positive association. Longer duration times appear to predict longer intervals to the next eruption. Interestingly, the scatter plot still suggests that there are two distinct subgroups, distinguished by durations of less than versus greater than three minutes.*

**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.**

```
proc format;                    /* Define dictionary of value labels */
  value durfmt 0='0= < 3 min'
               1='1= > 3 min';
run;

run;
data temp;
  set temp;
  durgp=.;                          /*  initialize to missing  */
  if .z lt dur lt 3 then durgrp=0;
  if dur ge 3 then durgrp=1;
  label durgrp='Duration, grouped';
  format durgrp DURFMT.;     /* Apply dictionary */
run;

proc sort data=temp;            /* it is necessary to sort */
  by durgrp;
run;

proc univariate data=temp plot;
  by durgrp;      /* repeat request for each level */
  var interval;
run;
```

*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.*

This instruction produces a lot of output (not shown here because it is too much):
separate descriptive statistics for INTERVAL for each subgroup defined by DURGRP. If you scroll to the end of all the output you will find the following:



**10. Finally, let's look at some numerical summaries, separately for the two groups.**
From the same output, you will find (with a little browsing about):

**Duration Grouping = 0 (Less than 3 minutes)**

Duration, grouped=0= < 3 min ----------

The UNIVARIATE Procedure
Variable: interval

Moments

| | | | |
|---|---|---|---|
| N | 67 | Sum Weights | 67 |
| Mean | 54.4626866 | Sum Observations | 3649 |
| Std Deviation | 6.29893776 | Variance | 39.6766169 |
| Skewness | 0.85271791 | Kurtosis | 1.92262225 |
| Uncorrected SS | 201353 | Corrected SS | 2618.65672 |
| Coeff Variation | 11.5656024 | Std Error Mean | 0.76953773 |

Basic Statistical Measures

| Location | | Variability | |
|---|---|---|---|
| Mean | 54.46269 | Std Deviation | 6.29894 |
| Median | 53.00000 | Variance | 39.67662 |
| Mode | 51.00000 | Range | 36.00000 |
| | | Interquartile Range | 7.00000 |

**Plus other stuff ...**

**Duration Grouping=1 (3 minutes or more)**

Duration, grouped=1= > 3 min ----------

The UNIVARIATE Procedure
Variable: interval

Moments

| | | | |
|---|---|---|---|
| N | 155 | Sum Weights | 155 |
| Mean | 78.1612903 | Sum Observations | 12115 |
| Std Deviation | 6.89106694 | Variance | 47.4868035 |
| Skewness | -0.3361253 | Kurtosis | 1.03908644 |
| Uncorrected SS | 954237 | Corrected SS | 7312.96774 |
| Coeff Variation | 8.81647029 | Std Error Mean | 0.55350382 |

Basic Statistical Measures

| Location | | Variability | |
|---|---|---|---|
| Mean | 78.16129 | Std Deviation | 6.89107 |
| Median | 78.00000 | Variance | 47.48680 |
| Mode | 75.00000 | Range | 42.00000 |
| | | Interquartile Range | 9.00000 |

**Plus other stuff ...**

*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.*

**Here is the entire program:**

```
*_____________________________________________________________
*
*             BE540w - Introductory Biostatistics
*                   Topic 1 - Summarizing Data
*                     Computer Illustration
*
*
*    code:  t1.sas
*    prog:  carol
*    path:  z:\bigelow\.....\
*    date:  September 21, 2004
*   input:  geyser.txt
*  output:  none
* results:  t1.lst
*_____________________________________________________________;

*/  No libname statement needed for this program */

*_____________________________________________
*
*  1.  Read in the data
*_____________________________________________;
data temp;      /* TEMP is a temporary SAS data set */
  infile 'z:\bigelow\teaching\web540\datasets\geyser.txt';
  input index dur interval;
run;


*_____________________________________________________
*
*  2.  Obtain a histogram of interval times
*_____________________________________________________;
proc capability data=temp;
  histogram interval;
  title 'Histogram of Interval Times';
run;


*_____________________________________________________
*
*  4. Instead of a histogram, obtain a stem and leaf
*  5. Obtain a Box and Whisker plot.
*_____________________________________________________;
proc univariate plot;
  var interval;
  title 'Stem and Leaf of Interval Times';
run;
```

**- continued-**

```
*_____________________________________________________
*
*  6.  Produce a plot of X=duration by Y=Interval
*_____________________________________________________;
proc plot;
  plot interval*dur = '*';    /* use asterisk symbol in plot */
  title 'Plot of Waiting Time Intervals by Duration';
run;


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

proc format;       /* Define dictionary of value labels */
  value durfmt 0='0= < 3 min'
               1='1= > 3 min';
run;

run;
data temp;
  set temp;
  durgp=.;             /*  initialize to missing  */
  if .z lt dur lt 3 then durgrp=0;
  if dur ge 3 then durgrp=1;
  label durgrp='Duration, grouped';
  format durgrp DURFMT.;       /* Apply dictionary */
run;

proc sort data=temp;    /* it is necessary to sort */
  by durgrp;
run;

proc univariate data=temp plot;
  by durgrp;          /* repeat request for each level */
  var interval;
run;
```