Introduction to STATISTIX

Before you begin...

1) The instruction ***Click 1x*** means click one time on the mouse.
   The instruction ***Click 2x*** means click two times (together quickly) on the mouse.

2) An instruction enclosed in < > means press that key. For example ***<enter>*** is instructing you to press the enter key; you do not actually type the letters in the word "enter".

3) What you actually type is highlighted in **bold**.

To Quit STATISTIX...

You can quit at any time, but be sure you save your work first. To quit:

1) Click 1x on FILE.
2) Click 1x on EXIT.

1. Installing STATISTIX for Windows

Objective
Install STATISTIX for Windows from the 3 1/2 installation disk to the hard drive.

Reference
STATISTIX manual pages 3-4.

STEP 1. Install the STATISTIX software.

These instructions assume you are in the Windows95 main window.

1) Click 1x on START.
   Click 1x on RUN.
   Click 1x on RUN…

   You will be positioned inside a command line dialog box.

2) If your disk drive is A, type **a:install <enter>.**
   If your disk drive is B, type **b:install <enter>.**

   You will see a message that welcomes you to the student edition of STATISTIX.

3) Click 1x on CONTINUE.

   You will see a message about the distination directory. The installation will be to C:\SXW. This is fine.

4) Click 1x on CONTINUE.

   The installation will be completed for you. When complete, you will see a STATISTIX window with two icons.

1. Installing STATISTIX for Windows

STEP 2. Set the printer specifications.

1) Begin STATISTIX: From the STATISTIX window, click 2x on the STATISTIX icon.
2) Click 1x on FILE.
3) Click 1x on PRINTER SETUP.
4) Use DOWN ARROW key or UP ARROW key to locate your printer. Then press <enter>.
5) Click 1x on OK.

2. Create and Save a STATISTIX System Data Set

Objective

Enter the following data into a STATISTIX data set called LAB1.SX and save it in the directory C:\SXW.

EVENT	YEARS	SMOKE	EVENT	YEARS	SMOKE
0	18.2	0	1	12.0	0
0	15.5	0	1	15.4	0
0	10.5	1	1	10.5	0
0	10.8	1	1	19.0	1
0	10.7	1	1	10.1	1
0	12.4	0	1	9.5	1
0	11.8	0	1	10.0	0
0	10.3	0	1	9.4	1
0	12.3	1	1	14.2	0
0	11.3	1	1	13.0	1

STEP 1. Enter the data.

Begin STATISTIX:

1) From the STATISTIX window, click 2x on STATISTIX to start the program.
2) Click 1x on DATA.
3) Click 1x on INSERT
4) Click 1x on VARIABLES.
5) Click 1x inside the NEW VARIABLE NAMES dialog box.
   Type **event years smoke <enter>.**
6) Click 1x on OK

At this point you will see a spread sheet with space for only one record. You'll need to instruct STATISTIX that you actually have a data set of 20 records

7) Click 1x on DATA
8) Click 1x on INSERT
9) Click 1x on CASES
10) Click 1x inside the FIRST NEW CASE NUMBER dialog box
    Type **2 <enter>**
11) Click 1x inside the NUMBER OF CASES TO INSERT dialog box.
    Type **19 <enter>**

To enter your data by rows:

Type in the first value.
Press the RIGHT arrow.
Type in the second value.
Press <enter>.
All other values are entered the same as the second value.

To enter your data by columns:
Type in the first value.
Press the DOWN arrow.
Type in the second value.
Press <enter>.
All other values are entered the same as the second value.
Continue until all values of one column are entered.
Use the arrow keys to get to the first record in the next column and proceed like with the first column.

2. Create and Save a STATISTIX System Data Set

You should have 20 records with 3 values per record.
If you have 21 records, click 1x on the record number 21.
Click 1x on EDIT.
Click 1x on CUT.

STEP 2. Save the data.

1) Click 1x on FILE.
2) Click 1x on SAVE AS.
3) *.sx should be highlighted in the FILE NAME dialog box.
   Type **lab1.sx <enter>.**

3. Read a STATISTIX System Data Set

Objective
Read into STATISTIX the system data set C:\SXW\LAB1.SX.

Begin STATISTIX:
1) From the STATISTIX window, click 2x on STATISTIX to start the program.
2) Click 1x on FILE.
3) Click 1x on OPEN.
4) Click 2x on LAB1.SX.

4. Export a STATISTIX System Data Set to an ASCII File

Objective
Create an ASCII format data set called C:\SXW\LAB1.TXT from the STATISTIX system data set called C:\SXW\LAB1.SX .

Reference
STATISTIX manual page 80.

Remark
STATISTIX allows you to export a STATISTIX system data set to three formats: comma and " " ascii, formatted ascii, and LOTUS 1-2-3.

STEP 1. If you have not already read in LAB1.SX...
Begin STATISTIX:
1) From the Windows main window, click 2x on STATISTIX.
2) From the STATISTIX window, click 2x on STATISTIX to start the program.
3) Click 1x on FILE.
4) Click 1x on OPEN.
5) Click 2x on LAB1.SX.

STEP 2. Export to ASCII.
1) Click 1x on FILE.
2) Click 1x on EXPORT.
3) The contents of the FILE NAME dialog box should be highlighted.
   Type **lab1.txt <enter>.**

4) Go to the VARIABLES dialog box.

If you want to select variables ONE at a time:
Highlight using arrow key.
Then click 1x.
Then click RIGHT arrow next to VARIABLES box.

If you want to do the optional step of using a format statement, then select the variables in the following order:
Click 1x on EVENT.
Then click RIGHT arrow next to VARIABLES box.
Click 1x on YEARS.
Then click RIGHT arrow next to VARIABLES box.
Click 1x on SMOKE.
Then click RIGHT arrow next to VARIABLES box.

If you want to select MORE THAN ONE variables:
Click 1x on the first variable you want and keep holding the mouse button pressed while dragging the mouse pointer to the last variable you want and let go of the mouse button.
Then click RIGHT arrow next to VARIABLES box.

6) (OPTIONAL step) To export using a format statement:

6.1) Click 1x on FIXED FORMAT in the FILE FORMAT dialog box.
6.2) Click 1x inside the FORMAT STATEMENT dialog box.
6.3) Type in a format statement. We will use the following:

i1 1x d6.2 1x i1 <enter>

7) Click 1x on OK.

5. Import an ASCII Data Set to a STATISTIX System Data Set

Objective
Import the fixed format ASCII data set called C:\SXW\LAB1.TXT into a new STATISTIX system data set called C:\SXW\LAB2.SX .

Reference
STATISTIX manual page 73.

Remarks
1) Importing an ascii data set does not mean that it is automatically saved as a STATISTIX system data set. You need to do this.
2) It is possible to import an ascii file in free format.

Begin STATISTIX:
1) From the Windows main window, click 2x on STATISTIX.
2) From the STATISTIX window, click 2x on STATISTIX to start the program.

If you already ar in STATISTIX
1) Click 1x on FILE.
2) Click 1x on NEW.
3) Click 1x on FILE.
4) Click 1x on IMPORT.
5) Click 1x inside FILE NAME dialog box.
   Type **c:\sxw\lab1.txt <enter>.**

6) Go to VARIABLE NAMES box.
   Click 1x on ENTER MANUALLY.

7) Click 1x inside IMPORT VARIABLE NAMES dialog box.
   Type **event years smoke <enter>.**

6. 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 this course, 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.

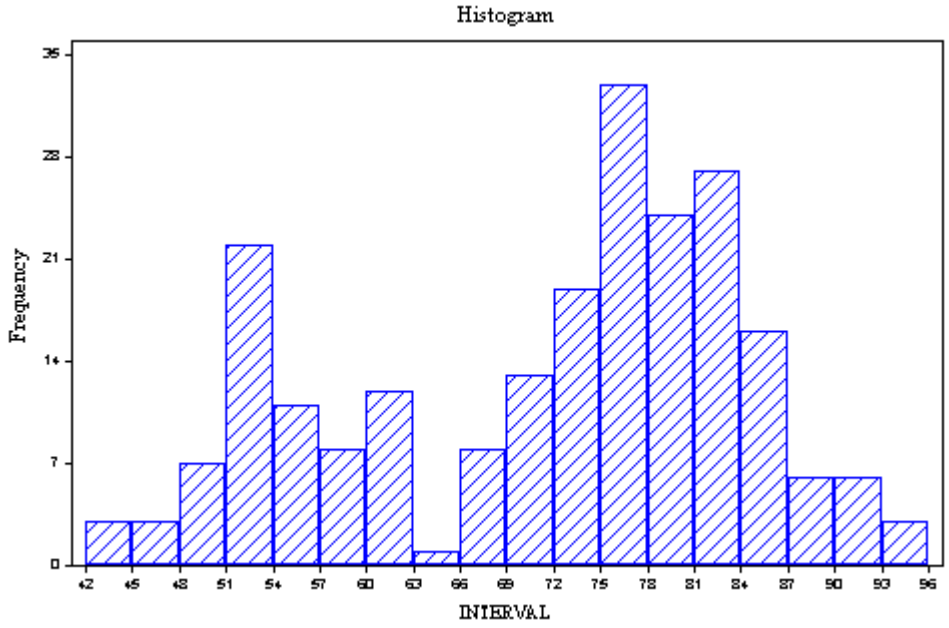
1. Read in the data.
Begin STATISTIX:
From the Windows95 main window, click 2x on STATISTIX.
From the STATISTIX main window, click 2x on STATISTIX
Click 2x on STATISTIX.
Click 1x on FILE.
Click 1x on IMPORT.
Click 1x inside FILE NAME dialog box.
Type **c:\sxw\geyser1.dat <enter>**.

Go to VARIABLE NAMES box.
Click 1x on ENTER MANUALLY.

Click 1x inside IMPORT VARIABLE NAMES dialog box.
Type **index duration interval <enter>**

Click 1x on OK.

2. Obtain a Histogram of Interval Times.
Click 1x on STATISTICS.
Click 1x on SUMMARY STATISTICS.
Click 1x on HISTOGRAM.
Use mouse to highlight the variable INTERVAL. Click 1x here.
Click 1x on RIGHT ARROW (>) next to HISTOGRAM VARIABLES dialog box.
Click 1x on OK
You should see the following figure:



Results

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 for printing.

Click 1x on FILE
Click 1x on SAVE AS
Click 1x inside FILE NAME dialog box.
Type **c:\sxw\geyser.out <enter>.**

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

Click 2x on the upper left corner [-] of the histogram window, to close.
Click 1x on STATISTICS.
Click 1x on SUMMARY STATISTICS.
Click 1x on STEM AND LEAF PLOT.
Use mouse to highlight the variable INTERVAL. Click 1x here.
Click 1x on RIGHT ARROW (>) next to PLOT VARIABLES dialog box.
Click 1x on OK.

You should see the following figure:

```
STEM AND LEAF PLOT OF INTERVAL

   LEAF DIGIT UNIT = 1                  MINIMUM  42.000
   4  2  REPRESENTS 42.                 MEDIAN   75.000
                                        MAXIMUM  95.000
          STEM  LEAVES
        3    4  234
       11    4  55788999
       39    5  00111111111111122233333444 4
       54    5  555566677778889
       67    6  0000111112223
       78    6  66677788999
      107    7  000001111122222333333334444 4
      (44)   7  55555555555555566666666667777777788888889999
       71    8  00000000000001111111122222222233333333444444444
       22    8  5666666788899
        9    9  00011134
        1    9  5

222 CASES INCLUDED   0 MISSING CASES
```

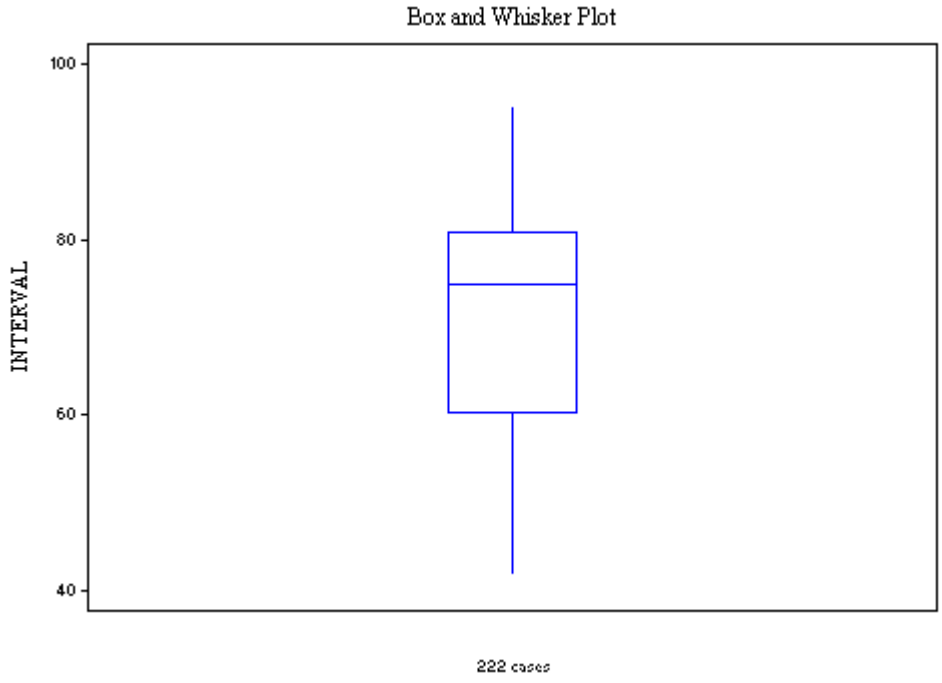
Results.

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. The row with the median is indicated by the cumulative frequency enclosed in parentheses.

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

Click 2x on the upper left corner [-] of the stem and leaf window, to close.
Click 1x on STATISTICS.
Click 1x on SUMMARY STATISTICS.
Click 1x on BOX AND WHISKER PLOT.
Use mouse to highlight the variable INTERVAL. Click 1x here.
Click 1x on RIGHT ARROW (>) next to DEPENDENT VARIABLE dialog box.
Click 1x on OK.
You should see the following picture:



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.

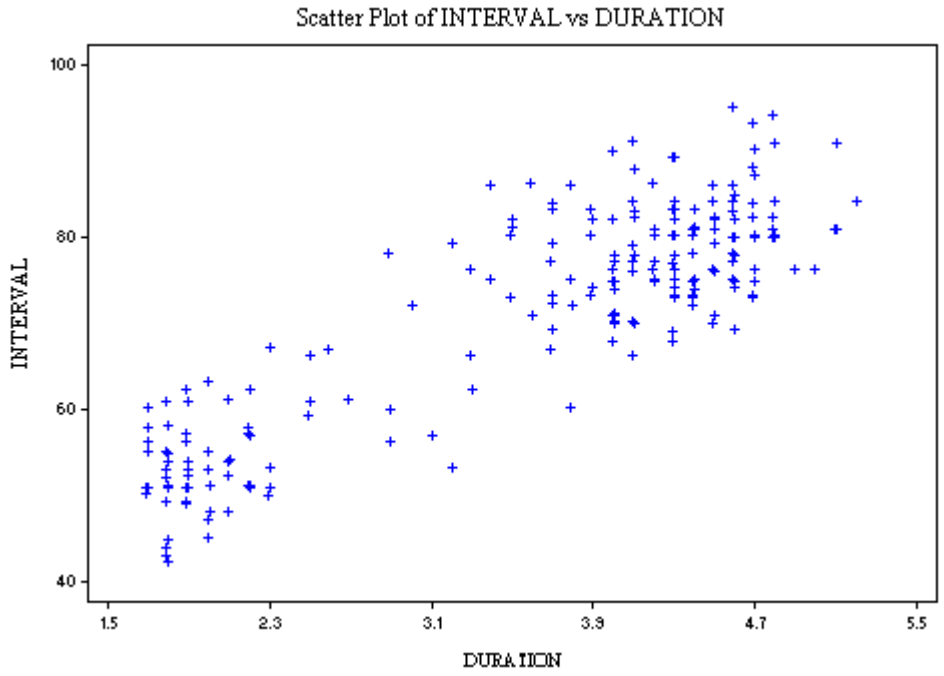
6. We have information on duration of eruption also. One possibility is that the duration of the current eruption is a predictor of the interval time to the next eruption. To investigate this possibililty, construct a scatter plot of interval time versus duration. Plot the predictor duration on the horizontal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

Click 2x on the upper left corner [-] of the box and whisker plot window, to close.
Click 1x on STATISTICS.
Click 1x on SUMMARY STATISTICS.
Click 1x on SCATTER PLOT.

Use mouse to highlight the variable DURATION. Click 1x here.
Click 1x on RIGHT ARROW (>) next to X-AXIS dialog box.

Use mouse to highlight the variable INTERVAL. Click 1x here.
Click 1x on RIGHT ARROW (>) next to Y-AXIS dialog box.

Click 1x on OK.
You should see the following picture:



Results

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.

Click 2x on the upper left corner [-] of the scatter plot window, to close.
Click 1x on DATA.
Click 1x on TRANSFORMATIONS.
Click 1x inside the TRANSFORMATIONS dialog box.
Type **If duration < 3**
Then durgrp=0
Else durgrp=1
Click 1x on GO
Click 1x on OK

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.

8. Label the values of the new variable DURGRP, so that output of analyses using this variable will be easier to read.

Click 2x in upper left corner [-] of transformations, to close.
Click 1x on DATA.
Click 1x on LABELS.
Click 1x on VALUE LABELS.
Use mouse to highlight DURGRP. Click 1x on RIGHT ARROW (>) next to SOURCE VARIABLE.

Go to the DEFINE LABEL BOX:

Click 1x inside the VALUE box. Type **0** ***DO NOT type <enter>***.
Click 1x inside LABEL box. Type **<3 min**
Click 1x on RIGHT ARROW (>) next to the VALUE LABELS box.

Click 1x inside the VALUE box. Delete the 0. Type **1** ***DO NOT type <enter>***.
Click 1x inside LABEL box. Delete the <3 min. Type **>3 min**
Click 1x on RIGHT ARROW (>) next to the VALUE LABELS box.

Click 1x on CLOSE.

9. Now we can look at the separate distributions of the interval times using the Box and Whisker Plot procedure. First, we'll look at the Box and Whisker plots.

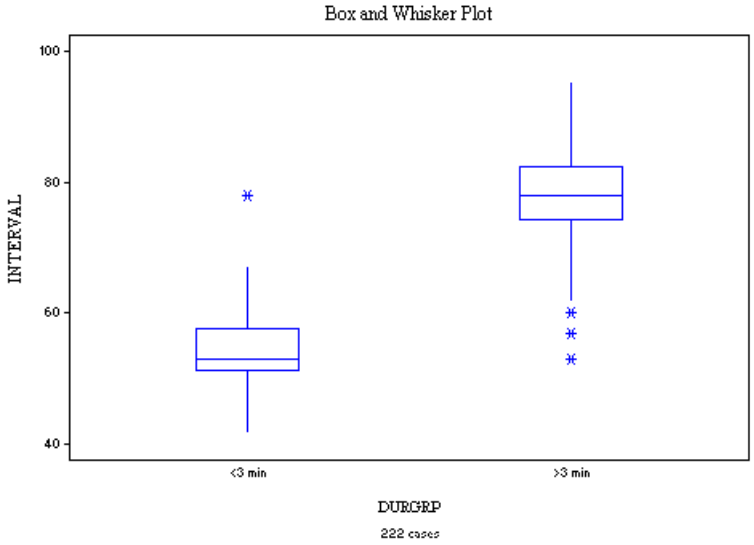
Click 1x on STATISTICS.
Click 1x on SUMMARY STATISTICS.
Click 1x on BOX AND WHISKER PLOT.

Notice that STATISTIX has remembered your choice of INTERVAL as the dependent variable.

Use mouse to highlight the variable DURGRP. Click 1x here.
Click 1x on RIGHT ARROW (>) next to CATEGORICAL. Click 1x here.

Click 1x on OK.

You should see the following picture:



10. Finally, let's look at some numerical summaries, separately for the two groups.

Click 2x on the upper left corner [-] of the box and whisker plot window, to close.
Click 1x on STATISTICS.
Click 1x on DESCRIPTIVES.

Use mouse to highlight the variable INTERVAL. Click 1x here.
Click 1x on RIGHT ARROW (>) next to DESCRIPTIVES.

Use mouse to highlight the variable DURGRP. Click 1x here.
Click 1x on RIGHT ARROW (>) next to GROUPING VARIABLE.

Go to STATISTICS TO REPORT box. Click 1x on each of the statistics you want.
Here, we choose the mean, sd, conf. int., median, min, and max.

Click 1x on OK.
You should get the following:

```
DESCRIPTIVE STATISTICS FOR DURGRP = <3 min

                INTERVAL
LO 95% CI         52.926
MEAN              54.463
UP 95% CI         55.999
SD                6.2989
MINIMUM           42.000
MEDIAN            53.000
MAXIMUM           78.000


DESCRIPTIVE STATISTICS FOR DURGRP = >3 min

                INTERVAL
LO 95% CI         77.068
MEAN              78.161
UP 95% CI         79.255
SD                6.8911
MINIMUM           53.000
MEDIAN            78.000
MAXIMUM           95.000
```

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.

7. How to Print Results

There are two ways to do this:

1. Send results to printer directly
2. Send results to a file which you edit and print later using some other software (e.g. Word, WordPerfect, etc)

It's a good idea to send your results to a file, beginning with the first set of results. Additional results can be added to this same file as your analysis proceeds.

1. How to Send Results to Printer

Click 1x on FILE
Click 1x on PRINT
Release mouse

2. How to Send Results to a File for the First Time

Suppose we wish to send our results to C:\SXW\EXER1.OUT

Click 1x on FILE.
Click 1x on SAVE AS.

Click 1x inside the FILE NAME dialog box.
Type **c:\sxw\exer1.out <enter>**
Click 1x on OK

3. How to Send Additional Results to the Same File

Suppose we wish to add new results to C:\SXW\EXER1.OUT

Click 1x on FILE.
Click 1x on SAVE AS.

The contents of the FILE NAME dialog box should be highlighted.
Type **c:\sxw\exer1.out <enter>**
Click 1x on APPEND
Click 1x on OK

8. EXERCISE: Video Display Terminals and Pregnancy

One of the case studies that we will be discussing this week is an investigation of the relationship between spontaneous abortion and the use of video display terminals (Schnorr, Grajewski, Hornung et al, 1991). This study is also described in Steenland K, Case Studies in Occupational Epidemiology (New York: Oxford Press, 1993, pp 7-20.)

This exercise asks you to begin considering the data used in this study. Your subdirectory C:\SXW contains a data set called VDT.DAT with 882 observations, formatted as follows:

Columns	Variable	Label/Codes	Format
1-6	AVGVDT	average hours vdt in 1st trimester	d6.3
7	-		1x
8-9	NUMCIGS	# cigarettes/day	i2
10	-		1x
11	PRIORSAB	prior spontaneous abortion (1=yes,0=no)	i1
12	-		1x
13	SAB	spontaneous abortion (1=yes,0=no)	i1
14	-		1x
15	SMOKSTAT	smoker (1=yes,0=no)	i1
16	-		1x
17	PRTHYR	prior thyroid condition (1=yes,0=no)	i1
18	-		1x
19	VDTEXPOS	VDT exposure (1=yes,0=no)	i1

8.1 Perform a chi square test of association between PRISAB and SAB. You should get a p-value of 0.0416.

8.2 Construct a histogram for of AVGVDT for workers with VDT exposure only. *Hint:* You will need to use the data management menu to select the subset of the analysis sample with VDT exposure. See page 37 of your STATISTIX manual.

8.3 Construct side by side Box and Whisker plots of NUMCIGS for the subgroups of exposed and non-exposed workers.

9. SOLUTION: Video Display Terminals and Pregnancy

Read in data set

From the STATISTIX main window, click 2x on STATISTIX
Click 1x on FILE
Click 1x on IMPORT

The file name dialog box should be highlighted.
Type **c:\sxw\vdt.dat <enter>**

Go to the VARIABLE NAMES box.
Click 1x on ENTER MANUALLY.
Click 1x inside IMPORT VARIABLE NAMES dialog box.
Type **avgvdt numcigs priorsab sab smokstat prthyr vdtexpos**.

Go to the FORMAT STATEMENT box.
Type **d6.3 1x i2 1x i1 1x i1 1x i1 1x i1 1x i1 <enter>**.
Click 1x on OK.

SOLUTION TO 8.1: Obtain chi-square test statistic
Click 1x on STATISTICS
Click 1x on ASSOCIATION TESTS
Click 1x on CHI-SQUARE TEST...

In the VARIABLES dialog box
Click 1x on PRIORSAB
Click 1x on RIGHT ARROW (>) next to ROW VARIABLE

In the VARIABLES dialog box
Click 1x on SAB
Click 1x on RIGHT ARROW (>) next to COLUMN VARIABLE
Click 1x on OK

You should see the following results:

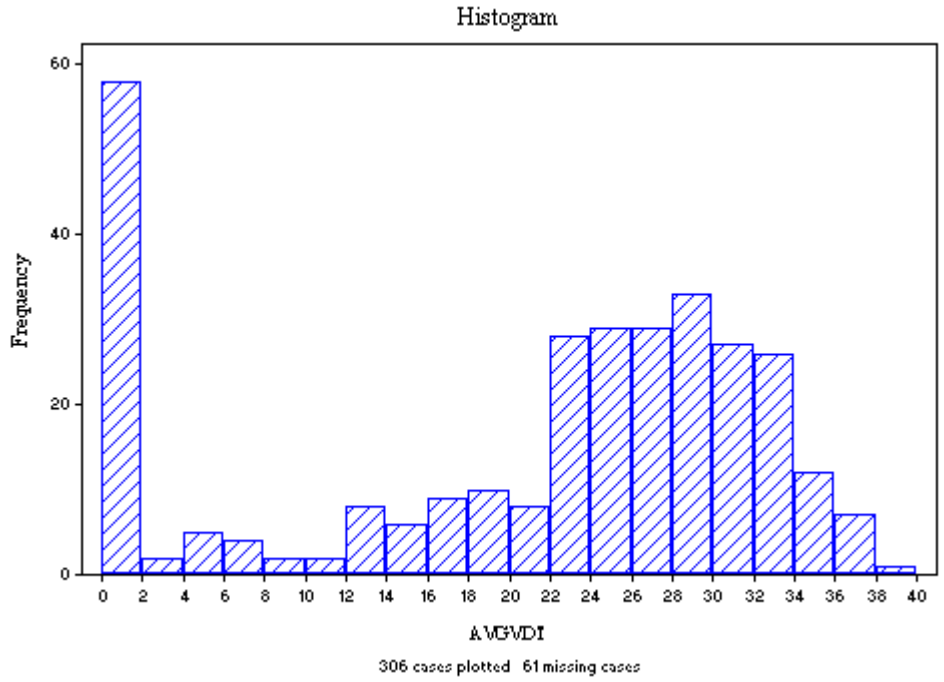
```
CHI-SQUARE TEST FOR HETEROGENEITY OR INDEPENDENCE
FOR 1 = PRIORSAB SAB

                          SAB
PRIORSAB              0           1
                +-----------+-----------+
0       OBSERVED |   661     |   112     |    773
        EXPECTED |   653.81  |   119.19  |
    CELL CHI-SQ  |     0.08  |     0.43  |
                +-----------+-----------+
1       OBSERVED |    85     |    24     |    109
        EXPECTED |    92.19  |    16.81  |
    CELL CHI-SQ  |     0.56  |     3.08  |
                +-----------+-----------+
                    746         136          882

OVERALL CHI-SQUARE     4.15
P-VALUE               0.0416
DEGREES OF FREEDOM         1

CASES INCLUDED 882    MISSING CASES 1
```

Results: The chi-square test tests whether the incidence of having had a prior spontaneaous abortion is independent of having a spontaneous abortion. If in truth the two incidences are independent the chances of seeing this result or one more extreme is approximately 4 in 100 or 1 in 25.

SOLUTION TO 8.2: Contruct histogram for AVGVDT for workers with VDT exposure only

Click 2x in upper left corner of CHI-SQUARE TEST window, to close.
Click 1x on DATA
Click 1x on OMIT/SELECT/RESTORE CASES...
Click 1x in the OMIT/SELECT/RESTORE EXPRESSION dialog box.
Type **omit vdtexpos=0**
Click 1x on GO
Click 1x on OK
Click 1x on CLOSE
Click 1x on STATISTICS
Click 1x on SUMMARY STATISTICS
Click 1x on HISTOGRAM
Click 1x on the variable AVGVDT in the VARIABLES dialog box
Click 1x on RIGHT ARROW (>) next to the HISTOGRAM VARIABLES dialog box
Click 1x on OK

You should see the following graph:



Results: The histogram shows that a large group of the exposed workers used video display terminals for a number of hours that ranged typically between 22 and 34 hours. It also shows that among the exposed workers another group used video display terminals for a number of hours was typically less than 2 hours.

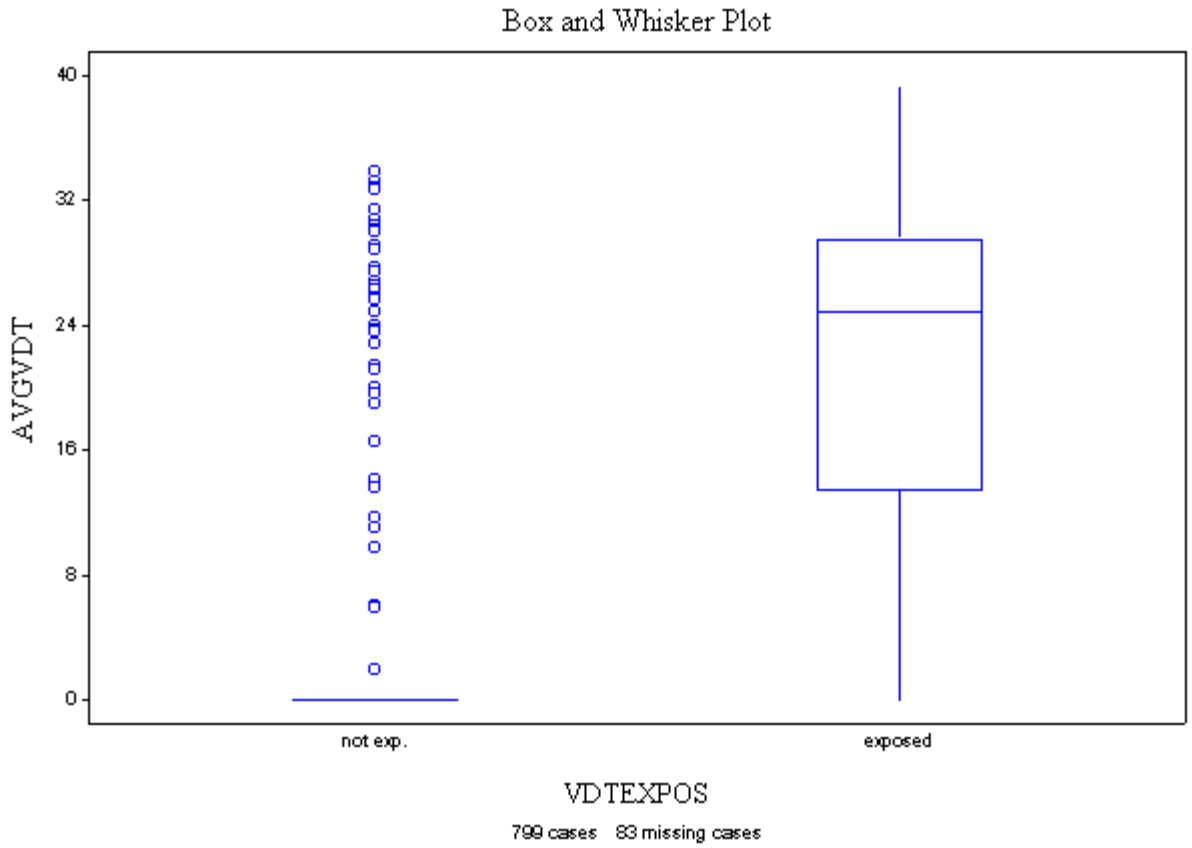
SOLUTION TO 8.3: Contruct side by side Box and Whisker plots of NUMCIGS for the subgroups of exposed and non-exposed workers

Click 2x in upper left corner of HISTOGRAM window, to close.
Click 1x on DATA
Click 1x on OMIT/SELECT/RESTORE CASES...
Type **restore** in the OMIT/SELECT/RESTORE EXPRESSION dialog box
Click 1x on GO
Click 1x on OK
Click 1x on CLOSE

(optional)
Click 1x on DATA.
Click 1x on LABELS
Click 1x on VALUE LABELS
Click 1x on VDTEXPOS in the VARIABLES dialog box
Click 1x on RIGHT ARROW (>) next to the SOURCE VARIABLE dialog box
Go to the DEFINE LABEL BOX
Click 1x inside the VALUE box. Type **0** ***DO NOT type <enter>***
Click 1x inside LABEL box. Type **not exp.** ***DO NOT type <enter>***
Click 1x on RIGHT ARROW (>) next to the VALUE LABELS box
Click 1x inside the VALUE box. Delete the 0. Type **1** ***DO NOT type <enter>***
Click 1x inside LABEL box. Delete the "not exp.". Type **exposed** ***DO NOT type <enter>***
Click 1x on RIGHT ARROW (>) next to the VALUE LABELS box
Click 1x on CLOSE
(End of optional)

Click 1x on STATISTICS
Click 1x on SUMMARY STATISTICS
Click 1x on BOX AND WISKER PLOTS...
In the VARIABLES dialog box
Click 1x on NUMCIGS
Click 1x on RIGHT ARROW (>) next to DEPENDENT VARIABLE
In the VARIABLES dialog box
Click 1x on VDTEXPOS
Click 1x on RIGHT ARROW (>) next to CATEGORICAL VARIABLE
Click 1x on OK

You should see the following graph:



Click 2x in upper left corner of BOX AND WISKER PLOTS window, to close.