Statistix® at Clemson University

Marie Coffin Herman Senter

Contents

Chapter 1

Introduction – Read Me First!

Statistix is a statistical analysis package. It is installed on 20 machines in the E-4 Martin Hall lab. Lab hours vary from one semester to another; lab hours will be posted on the lab door, and announced in class. Statistix is also available on the networked PC's around campus. These PC's are located in different labs with different hours. Individuals with an IBM-type PC may purchase Statistix at the bookstore (approximately $50). Statistix is *not* available for Macintosh computers.

This manual is intended to provide a brief introduction to the package for new users. The intended audience is students in MthSc 301, 302 and 405. However, we chose Statistix as a package because of its broad applicability to a variety of data analysis problems. We hope that many other people will use the package, and that this manual will be of use to them also. The manual is organized so that the reader can turn immediately to the topic of interest (maybe you're doing histograms in class today), and need not read the entire manual from cover to cover. However, this introductory chapter contains some general information, so you should finish reading it before turning to the topic of interest.

1.1 Starting the program

In the lab, select a computer and, if necessary, turn it on. It should display the network login screen. (If not, it is probably faulty and you should try another computer.) You will need to type in your userid and password. (If

you do not know these, please go to the Help Desk in the basement of Martin Hall.) Click twice (with the left mouse button) on the "CES" icon. This will open a new window. Click twice on the "Statistix" icon to start the program running. The first panel you see is the "base menu", a listing of the many procedures available. You are ready to begin work.

1.2 Selecting a dataset

From the base menu, the first thing you need to do is tell Statistix what dataset will be analyzed. If you try to choose a topic like "Linear models" or "Histogram", but you haven't chosen a dataset, you will get an error message. We expect that students will first use Statistix to analyze sets of data from the textbook. These datasets are already stored in the lab computers, and are also available on diskettes in the E-4 lab, so you can access them easily. If you want to analyze a set of data that's *not* already on the computer (or is on the computer, but not stored in Statistix), you will first need to enter your data. Please turn to Chapter 11 to find out how to do this.

To select a stored dataset, choose **File management** on the main menu and **Retrieve** on the next menu. You will see a blue option window. In the middle of the window is a list of all the available files. Type in the name of the one you want to use and press **F1** to retrieve it. If you get an error message at this point, you probably spelled the filename wrong. Just type it again, and press **F1** again. (Alternatively, press **F2**. Then use the arrow keys to highlight the filename you want, and press **Enter** to select it.) If you are retrieving a file from a diskette, you may need to change the *drive designator* to the floppy drive (usually this is the A: drive); if you are having trouble with this, ask the lab monitor for help. Once you have retrieved a file, there will be a little gray message at the bottom of the screen saying that your file has been read. Press any key on the keyboard to get back to the main menu.

A listing of data files appears in Appendix A starting on p. 52.

1.3 Finding your way around

Statistix is a menu-driven program, which makes it easy for new users to sit down and do a data analysis without having to "study" the program beforehand. When you start running it, you will get a screen with nothing on it but the base menu. This is like a table of contents: it lists the general topics of data analysis. You can highlight any topic on the base menu by using the arrow keys ($\rightarrow\uparrow\downarrow\leftarrow$), and then explore that topic in more depth by pressing the **Enter** key, or by clicking with the left-hand button of your mouse. This will give you a secondary menu, with more detailed topics on it. Move around the secondary menu until you find the specific topic you want, and use **Enter** or the mouse to select that topic.

1.4 Changing your mind

Part of the convenience of a menu-driven program is that you can always change your mind. If you choose "histograms", for example, and after looking at your histogram you decide a boxplot would look better, simply press **ESC** (the escape key, probably on the upper left-hand of your keyboard) to leave the "histogram section" of the program. Pressing **ESC** often enough will always get you back to the base menu, so if worst comes to worst, you can always start over. **Pressing ESC at the base menu will exit the program.**

1.5 Printing

Once you get the results you want from Statistix, you will probably want to print them. Look at the very bottom your results screen. There is usually a list of options there, with the first letter of each highlighted like this:

Print **S**ave **F**ile

and so forth. This means that by pressing "P" on your keyboard, you will send your results to the printer.

1.6 Exiting the Program

Exit Statistix by pressing the **ESC** key, (you may need to do this several times) until you are back at the base menu. Press **ESC** once more to exit the program.

1.7 Note

If you find errors in this manual or have suggestions on how it could be improved, please tell Marie Coffin or Herman Senter in the Math Sciences Dept.

Chapter 2

Graphs

Whenever you have a new set of data, it's a good idea to make graphs to explore the data. Statistix makes it easy to produce interesting graphs quickly. Before drawing any of the graphs in this chapter, you must tell the computer what data to use. If you are using a set of data given in class, **please turn to p. 2 and follow the directions under "Selecting a dataset".** To enter your own data and use it, please read Chapter 11.

To demonstrate the graphs in this chapter, we will use a set of data called "gessel.sx". This dataset records the case number, age in months (when first word was spoken), sex (m or f) and gesell score (a measure of mental ability), for 21 children. You can load this dataset yourself, following the directions on p. 2.

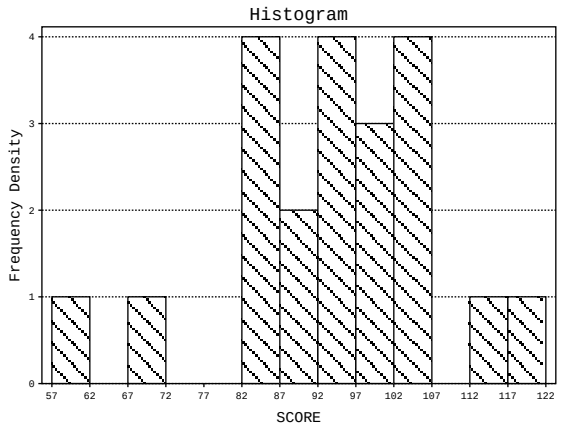
2.1 Histograms and Bar Charts

A histogram is a horizontal "bar chart", showing the distribution of a set of data. To get a histogram, choose **Summary Statistics** on the base menu, and **Histogram** on the next menu. The screen will become an **option screen**, where you give the computer some information.

The first section of the option screen lists all the variables in your dataset. The second section asks you which variable should be used in drawing a histogram. Type in a variable name and press the **Enter** key. If a red box appears at the bottom of the screen with the words **unknown variable**, you have made a mistake in typing the name. Type it again.

When you have typed the variable correctly, press the **F1** function key (function keys are at the top of the keyboard), and your histogram will appear. The variable you select can be either a numeric variable (in which case you will get a histogram) or a character variable (in which case you will get a bar chart).

A histogram of the *score* variable looks like this:



To print your histogram, press **P**.

2.1.1 Modifications

You can modify the histogram of a numeric variable by choosing the width of the bars, instead of letting the computer choose them. Go to the third options section. To get a horizontal scale running from 40 to 100 with widths of 5, type 40 100 5 and press the **F1** key.

2.1.2 A normal curve

To see if your data are reasonably normal, you can tell the computer to draw the best-fitting normal curve on top of your histogram. Go to the bottom options secton, and change the line that reads
Do want normal curve superimposed over the histogram? ... *No*
to **Yes**.

2.2 Stem-and-leaf Diagrams (Stemplots)

To draw a stemplot, choose **Summary Statistics** on the Base Menu, and **Stem and Leaf Plots** on the next menu. You will get an "options screen". At the top, you will see all the variables in your dataset listed. In the next section, you need to type the names of the variable (or variables for which you want stemplots. *Or* you can type ALL; this will give you a stemplot for each variable in your dataset. Press **F1** to see your stemplot. A stemplot for the *scores* variable in the "gesell.sx" dataset looks like this:

```
STATISTIX 4.0                GESELL, 08/11/94, 16:21

STEM AND LEAF PLOT OF SCORE

   LEAF DIGIT UNIT = 1
   1  2  REPRESENTS 12.

          STEM  LEAVES
        1    5  7
        1    6
        2    7  1
        7    8  33467
       (5)   9  13456
        9   10  0002245
        2   11  3
        1   12  1

21 CASES INCLUDED     0 MISSING CASES
```

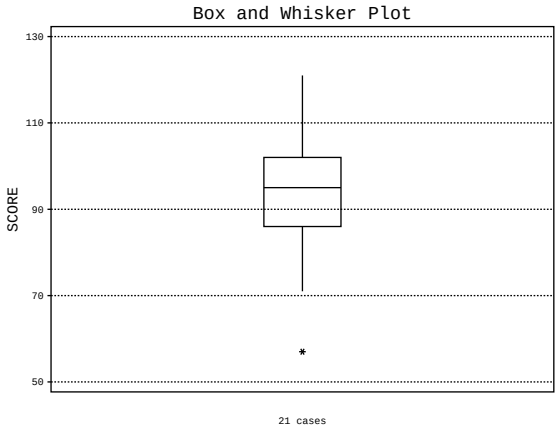
If there are a great many leaves to each stem, the computer may break each stem in half (an upper half and a lower half). The lower half of the stem (leaves 0 through 4) will have a "*" beside it, and the upper half (leaves 5 through 9) will have a "."

2.3 Box-and-Whisker Plot (Boxplot)

Choose **Summary Statistics** on the base menu and **Box and Whisker Plot** on the next menu.

On the options screen, move to the third section, and enter the variable name. Press **F1** to see your boxplot.

A boxplot for the "gesell.sx" dataset looks like this:



2.3.1 Side-by-side boxplots

If your data consist of several groups, you can use side-by-side boxplots to compare the groups. To do this, you must tell the computer how the data are grouped. Either each group will be in a separate column of the dataset (**Table**), or the groups depend on the value of some variable (**Categorical**). For

example, the "gesell.sx" data could be divided into groups of males (sex=m) and females (sex=f).

Table

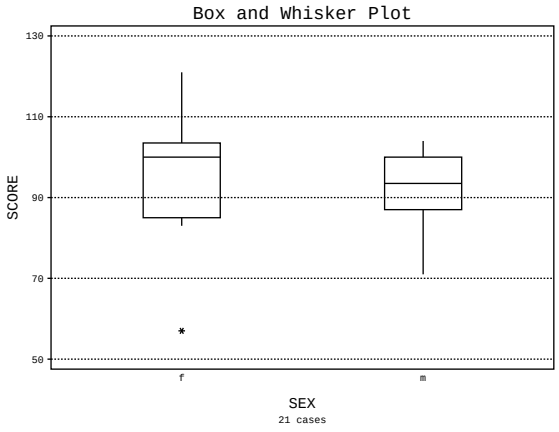
Male	Female
95	113
71	96
⋮	⋮
94	86
	100

Categorical

Score	Sex
95	M
71	M
⋮	⋮
94	M
113	F
⋮	⋮
100	F

Table Format

If each group is in a separate column of the dataset, move to the second section of the options screen and type **T**. Press the **Enter** key. Then type the variable name for each group. Press **F1** to see your boxplots. For example, side-by-side boxplots for males and females in the "gesell.sx" data look like this:

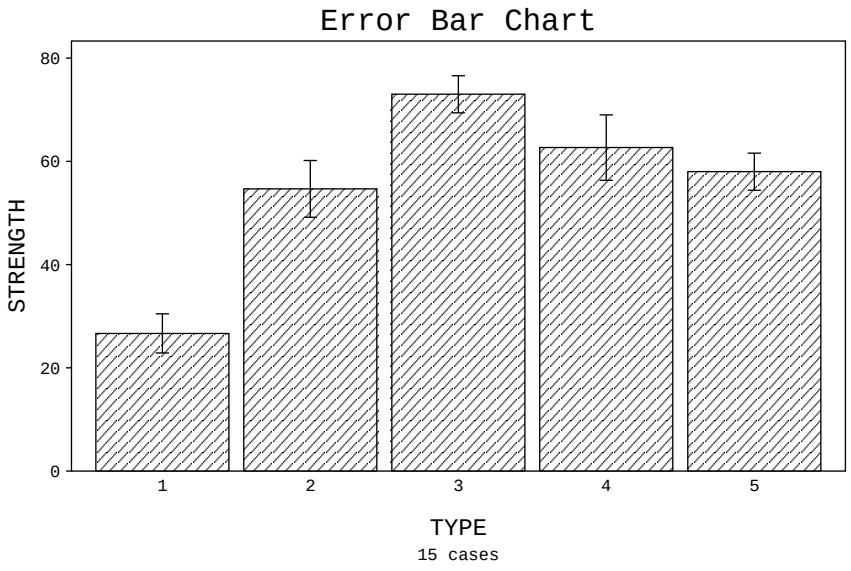


Categorical Format

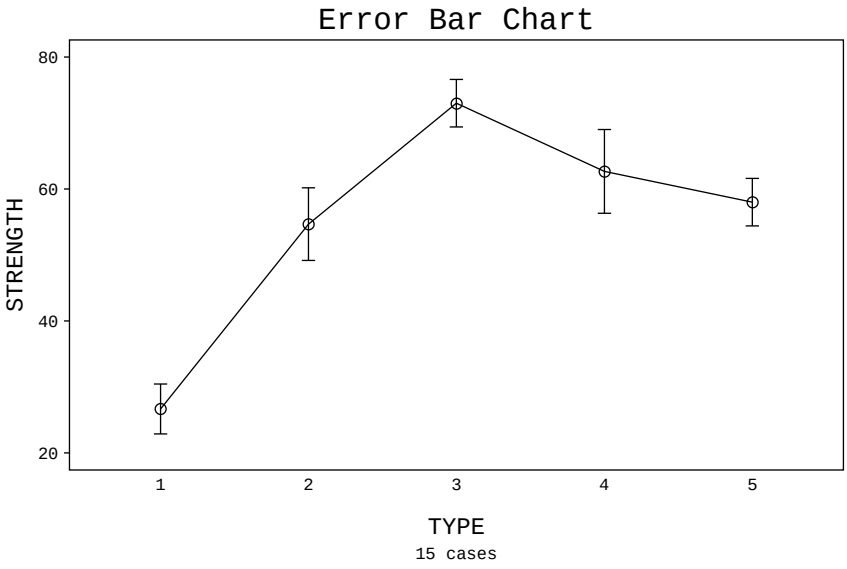
If groups depend on the value of some variable, move to the second section of the option screen and type **C**. In the third section, type the variable you want boxplots for, and in the fourth section, type the variable that divides the data into groups. Then press **F1** to see your boxplots.

2.4 Error Bar Charts

An error bar chart is used to compare several groups of data. The data may be broken into groups by a classifying variable (**Categorical** format), or the groups may be separate variables (**Table** format). There are two forms of the graph:



Bar form: the height of the bar is the mean, and the vertical line is the standard deviation.



Line form: the dot represents the mean, and the vertical line is the standard deviation.

In both forms, the standard deviation is optional.

To draw an error-bar chart, choose **Summary Statistics** on the base menu, and **Error bar chart** on the next menu. All the variables will be displayed at the top of the screen. In the next section of the screen, you must choose either **Table** or **Categorical**, depending on how the data are grouped. See page 8 of this manual for an example of each.

Table Format

If each group is in a separate column of the dataset, move to the next section of the options screen and type **T**. Press the **Enter** key. Then type the variable name for each group. At the bottom of the screen, you can choose the form of chart (Bar or Line), and choose whether or not to display the standard deviations (Yes or No). Then press **F1** to see your chart.

Categorical Format

If the data consists of observations in one variable and categories in another variable, move to the next section of the options screen and type **C**. Press the

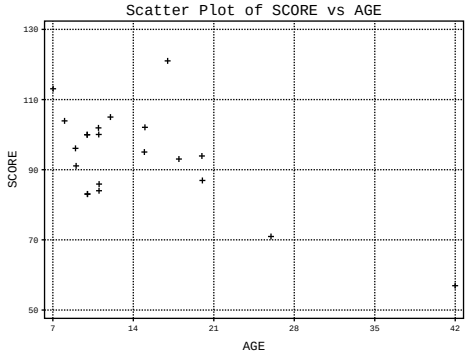
Enter key. In the next section of the screen, type the name of the variable that contains the observations. Press the **Enter** key again. In the next section, type the names of one or two variables that specify the categories. At the bottom of the screen, you can choose the chart form (Bar or Line) and whether or not to display the error bars (Yes or No). Press **F1** to see your chart.

2.5 Scatterplots (X-Y plots)

Scatterplots are used to display the relationship between **two** variables. To draw a scatterplot, choose **Summary Statistics** on the base menu and **Scatter plots** on the next menu. All the variables will be listed at the top of the option screen. In the next section of the screen, you need to type in two variable names, like this:

xvariable,yvariable

The x-variable (the one you type first) will be the horizontal axis, and the y-variable will be the vertical axis. Once you have typed in both variable names, press **F1** to see your scatterplot. The scatterplot of *age* vs. *score* for gesell.sx looks like this:



2.5.1 Plotting a line

To plot the best fitting straight line (regression line) on your scatterplot, go back to the options screen. In the bottom section, change the
Do you want to display a regression line (Y/N) **No**
to **Yes**. Then press **F1** to see your scatterplot and regression line.

2.6 Control Charts

Control charts such as $\bar{x}$-charts, R-charts and s-charts are used in quality control to track the behavior of important production variables. To draw these charts, choose **Quality Control** on the base menu. Read Chapter 10 to find out more about control charts.

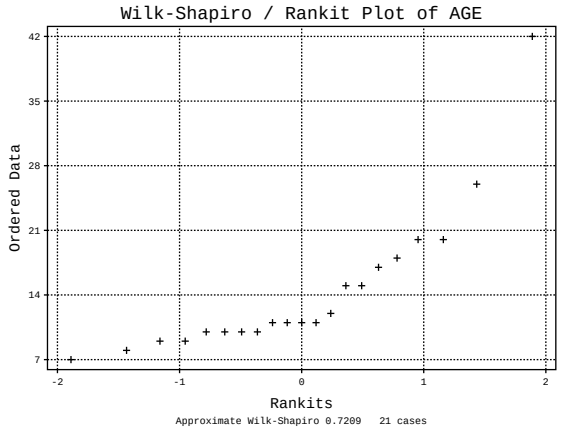
2.7 Time plots

Time plots are used to track the behavior of a variable across time. Time plots have the variable of interest on the vertical (or Y) axis, and time on the horizontal (or X) axis. To draw such a plot, choose **Time series** on the base menu, and **Time series plot** on the next menu. The option screen will show you all the variables in your dataset. In the second section, type the name of the variable to be plotted over time. This will be the vertical axis of the plot. In the third section, type the name of the variable that contains the time points. This will be the horizontal axis. Then press **F1** to see your plot.

2.8 Normal probability plots (Normal quantile plots)

A normal probability plot is a plot of standardized variables against the quantiles of the normal (or Gaussian) density function. If the plot is reasonably close to a straight line, that indicates that the data might very well be normally distributed. If the plot deviates from a straight line in some **systematic** way, that indicates the data are not normally distributed.

To draw a normal probability plot, choose **Randomness, Tests of Normality** on the base menu and **Wilk-Shapiro/Rankit plot** on the next menu. On the second section of the options screen, type in the variable for which you want a normal probability plot. Press **F1** to see your plot. A normal probability plot of *age* in the "gesell.sx" dataset looks like this:



The plot shows definite curvature, so the ages are probably not normally distributed.

Chapter 3

Descriptive Statistics

Descriptive statistics are numbers that are used to describe a dataset. Typically they will be things like $\bar{x}$ (the mean) or s (the standard deviation), but a number of other descriptive statistics are also available.

To find descriptive statistics, choose **Summary statistics** on the base menu, and **Descriptive statistics** on the next menu.

Below is a complete list of the descriptive statistics available on Statistix. The ones marked with a "*" are probably the most useful.

* N = # of observations

  Missing = # of missing observations (i.e. observations in the data set for which no value was recorded)

  Sum = $\sum x_i$

* Mean = $\bar{x} = \frac{1}{n} \sum x_i$

* SD = s = standard deviation

* SE Mean = $s/\sqrt{n}$ = standard error of the mean

* Conf. int = confidence interval for the mean

  C.V = $s/\bar{x} \times 100$ = coefficient of variation (%)

* Median

* Min/max = largest and smallest observations

* Quartiles = Q_L and Q_U = the upper and lower quartiles

  MAD = median absolute deviation

  Biased var = $\frac{1}{n}\sum(x_i - \bar{x})^2$

  Skew = coefficient of skewness

  Kurtosis = coefficient of kurtosis

You must tell the computer which variable (or variables) to use to calculate the statistics. You can calculate summary statistics on as many numeric variables as you want. Type in the variable name (or names) in the second section of the option screen.

At the bottom of the screen, all the available statistics are listed. A few of them are marked with X's beside them. If you press **F1** you will see these marked statistics calculated for each variable you chose. To calculate other statistics, put an X beside the ones you want. Use the space bar or delete key to erase an X. When you have chosen all the statistics you want, press **F1** to see the results. They will be printed in a little table like the one below:

```
STATISTIX 4.0          GESELL, 08/11/94, 17:16

DESCRIPTIVE STATISTICS

                    SCORE
N                      21
LO 95% CI          87.300
MEAN               93.667
UP 95% CI          100.03
SD                 13.987
MINIMUM            57.000
1ST QUARTI         85.000
MEDIAN             95.000
3RD QUARTI         102.00
MAXIMUM            121.00
```

3.1 Grouping variables

If your data are divided into groups according to the value of some variable (like sex), you can calculate statistics separately for each group. In the third section of the option screen, type in the name of the group variable. Here are some statistics on the "gesell.sx" dataset, calculated separately for males and females.

```
STATISTIX 4.0  GESELL, 08/11/94, 17:14

DESCRIPTIVE STATISTICS FOR SEX = f

                      AGE       SCORE
N                      11          11
LO 95% CI         7.1072      83.566
MEAN              13.636      95.182
UP 95% CI         20.166      106.80
SD                9.7188      17.291
MINIMUM           7.0000      57.000
1ST QUARTI        10.000      84.000
MEDIAN            11.000      100.00
3RD QUARTI        12.000      105.00
MAXIMUM           42.000      121.00

DESCRIPTIVE STATISTICS FOR SEX = m

                      AGE       SCORE
N                      10          10
LO 95% CI         11.032      84.967
MEAN              15.200      92.000
UP 95% CI         19.368      99.033
SD                5.8271      9.8319
MINIMUM           8.0000      71.000
1ST QUARTI        9.7500      86.000
MEDIAN            15.000      93.500
3RD QUARTI        20.000      100.50
MAXIMUM           26.000      104.00
```

3.2 Changing the level of the confidence interval

If you want to change the confidence level of the interval, return to the options screen for Descriptive Statistics, and type in the new confidence level in the fourth section. For a 90% confidence interval, type **90**, for a 99% interval, type **99**, etc.

Chapter 4

Simple Linear Regression I (includes correlation)

Simple linear regression is used to fit a model of the form

$$y = \beta_0 + \beta_1 x$$

(Sometimes this model is written as $y = \alpha + \beta x$ instead.) In this chapter we will discuss scatterplots, finding the regression equation, and finding the correlation (r). The other aspects of regression analysis (estimation, testing, residuals analysis) are covered in Chapter 5.

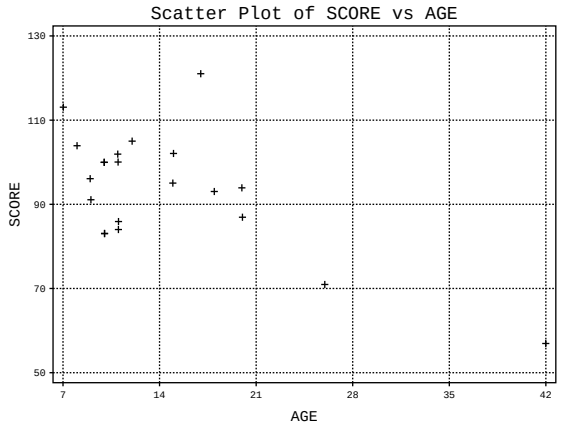
4.1 Scatterplots (X-Y plots)

Your data must contain at least two variables. Choose **Summary statistics** on the base menu and **Scatterplot** on the next menu. At the top of the option screen will be a list of all the variables in your dataset. In the next section, you must type in two variable names, one to go on the horizontal (X) axis, and one to go on the vertical (Y) axis. The x-variable should be first, then the y-variable, like this:

age,score

(There is room on the screen to type in more than one pair of x-y variables, but this is not used very often.) Press **F1** to see your scatterplot.

The scatterplot showing the relationship between age and score in the "gesell" dataset is below.



4.2 Correlation

The correlation r measures the strength of the linear relationship between two variables. To find it, choose **Linear models** on the base menu, and **Correlation (Pearson)** on the next menu. On the second section of the option screen, type the names of your two variables. You can type in more than two: if you do, it will calculate the correlation between each pair of variables. Press **F1** to see the results.

4.3 The regression equation

The regression line is also called the least-squares line. To find the equation of the regression line that best fits your data, choose **Linear models** on the base menu, and **linear regression** on the next menu.

The first section of the option screen will list all your variables. In the second section, type in the *dependent* or Y variable. In the next section, type

in the *independent* or X variable. Press **F1** to get the equation.

The output for this step is confusing, because it gives you a lot more information than just the regression line. Look at the sample output below to see how to find the equation of the regression line.

```
STATISTIX 4.0                              GESELL, 08/11/94, 13:58

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF SCORE

PREDICTOR
VARIABLES    COEFFICIENT    STD ERROR     STUDENT'S T        P
---------    -----------    ---------     -----------     ------
CONSTANT         109.874       5.06780         21.68       0.0000
AGE             -1.12699       0.31017         -3.63       0.0018

R-SQUARED            0.4100   RESID. MEAN SQUARE (MSE)   121.505
ADJUSTED R-SQUARED   0.3789   STANDARD DEVIATION          11.0229

SOURCE         DF       SS           MS          F         P
----------    ---   ----------   ----------   -----    ------
REGRESSION      1     1604.08      1604.08    13.20    0.0018
RESIDUAL       19     2308.59      121.505
TOTAL          20     3912.67

CASES INCLUDED 21   MISSING CASES 0
```

(1) The constant coefficient is b_0 (or a, depending on the notation your book uses). So, in this example, $b_0 = 109.874$.

(2) The "age" coefficient is b_1 (or just b). In this example, $b = -1.2699$.

So, in this example, the regression equation is

$$\hat{y} = 109.874 - 1.2699x$$

(Notice how the "-" sign gets carried along.)

Chapter 5

Tests and Intervals

This chapter explains how to find confidence intervals and perform hypothesis tests.

5.1 A confidence interval for μ

To find a confidence interval for the mean of a single sample, choose **Summary statistics** on the base menu and **Descriptive statistics** on the next menu. In the second section of the options screen, type the variable name.

In the fourth section, type the level of confidence interval you want. 95% is the default. If you want this, you don't need to type anything here. For a 90% interval type **90**, for a 99% interval type **99**, etc.

At the bottom of the screen, place an X beside "Conf. int." and erase the X's (use the space bar or delete key) beside any other statistics. Press **F1** to see your confidence interval.

5.2 A confidence interval for $\mu_1 - \mu_2$.

This confidence interval is not available in Statistix.

5.3 A one-sample t-test

A one-sample t-test is used to test $H_0 : \mu = \#$ against one of the alternatives $H_a : \mu < \#$, $H_a : \mu > \#$, or $H_a : \mu \neq \#$. The test is appropriate if the data are normally distributed, or if the sample size is large. For a small sample, you should first check to see if the data look reasonably normal. See Section 2.8 to do this.

Choose **One, two & multisample tests** on the base menu and **One-sample t test** on the next menu. The variable names for your dataset will be listed at the top of the options menu. Type in the name of the variable you want to test. In the next section of the screen, type in the actual value of # in your hypothesis test. Finally, choose the correct alternative hypothesis (Not equal, Less than, Greater than) and press **F1**.

For example, we could use the "gesell.sx" dataset to test $H_0 : \mu = 100$ vs. $H_a : \mu \neq 100$, where μ is the mean gesell score for the entire population. The output is:

```
STATISTIX 4.1                    GESELL, 08/24/95, 13:59

ONE-SAMPLE T TEST FOR SCORE

NULL HYPOTHESIS: MU =  100.0
ALTERNATIVE HYP: MU <> 100.0

MEAN          93.667
STD ERROR     3.0522
T              -2.08
DF                20
P             0.0511

CASES INCLUDED 21    MISSING CASES 0
```

We see that the value of the test statistic is $t = -2.08$, with 20 degrees of freedom. The p-value of this test is 0.0511, so there is at least some evidence that the average gesell score is not 100. Remember, small p-values are evidence *against* the null hypothesis.

5.4 Paired sample t-test

The paired sample t-test of $H_0 : \mu_d = 0$ is appropriate when the data consist of pairs of observations. For the test to work, the data must be in two columns like this:

Before	After
10	9
15	13
9	7
⋮	⋮
11	8

Choose **One, two & multi-sample tests** on the base menu, and **Paired t test** on the next menu. Type in the variable names for the two columns. Press **F1** to see the results. The output for the "thread.sx" dataset looks like this:

```
STATISTIX 4.0   THREAD, 08/11/94, 17:37

PAIRED T TEST FOR LEFT - RIGHT

MEAN          9.5600
STD ERROR     6.1000
T               1.57
DF                24
P             0.1302

CASES INCLUDED 25    MISSING CASES 0
```

(1) $\bar{x}$

(2) $s/\sqrt{n}$

(3) t-statistic

(4) DF = degrees of freedom

(5) P = p-value of the hypothesis test $H_0 : \mu_d = 0$. Recall that small values of P are evidence *against* H_o.

5.5 Two sample t-test

A two sample t-test is used to test

$$H_0 : \mu_1 = \mu_2$$

The test assumes the two samples are independent.

To run this test, choose **One, two and multi-sample tests** on the base menu, and **Two-sample t test** on the next menu.

5.5.1 How are the data arranged?

On the options screen, you must specify how the data are arranged. Either the two samples are in two columns of the dataset (**Table**), or the values are all in one column, but there is another column (another variable) that separates the two groups (**Categorical**). Type in the one that applies to your data. Example:

Table

Male	Female
95	113
71	96
⋮	⋮
94	86
	100

Categorical

Score	Sex
95	M
71	M
⋮	⋮
94	M
113	F
⋮	⋮
100	F

5.5.2 Table option

If the data are arranged in a table, there are two columns or variables in the dataset that represent the two groups. Type in the two variable names in the next section of the option screen, and press **F1** to see the results.

5.5.3 Categorical option

The data values are all in one column, and another variable is used to distinguish between the two groups. Type the variable name that contains the data

in the next section of the options menu. In the last section of the screen, type the variable name that separates the two groups. The results of comparing gesell scores for males and females is given below.

```
STATISTIX 4.0                              GESELL, 08/11/94, 13:59

TWO-SAMPLE T TESTS FOR SCORE BY SEX

                          SAMPLE
   SEX           MEAN      SIZE       S.D.        S.E.
----------   ----------   ------   ---------   ---------
f               95.182       11      17.291      5.2133
m               92.000       10      9.8319      3.1091

                        T       DF        P
                     ------   ------   ------
EQUAL VARIANCES        0.51     19     0.6153
UNEQUAL VARIANCES      0.52     16.1   0.6073

                          F      NUM DF   DEN DF      P
TESTS FOR EQUALITY    -------   ------   ------   ------
      OF VARIANCES      3.09       10        9     0.0521

CASES INCLUDED 21    MISSING CASES 0
```

5.5.4 Output

The first part of the output gives summary statistics for the two groups. The next part gives the actual results of the t-test. *Two* t-tests are done. The first one assumes $\sigma_1^2 = \sigma_2^2$, and the other assumes $\sigma_1^2 \neq \sigma_2^2$. For each test, the value of t, the degrees of freedom, and the p-value are given. Recall that small values of p are evidence *against* H_0.

The third section of output shows an F-test of

$$H_0 : \sigma_1^2 = \sigma_2^2$$

.

5.6 Comparing many samples

To perform a test of

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$$

you can use an analysis of variance (ANOVA or AOV) F-test. Choose **One, two and multi-sample tests** on the base menu, and **One-way AOV** on the next menu.

5.6.1 How are the data arranged?

Either the dataset is arranged so that each group is in a separate column (**Table**), or all the data is in one column, and another column (or variable) contains values of some grouping variable (**Categorical**).

Example:

Table

Basal	DRTA	Strat
4	7	11
6	7	7
	⋮	
9	10	8

Categorical

Score	Group
4	Basal
6	Basal
⋮	⋮
7	DRTA
⋮	⋮
8	Strat

Table arrangement

The data are arranged in several columns. Type the variable name for each column (e.g. basal drta strat). Press **F1** to see the results.

Categorical arrangement

All the observations are in one column. Type in the variable name for this column. In the next section of the options page, type in the variable name of the variable that specifies the groups. Press **F1** to see the results.

5.6.2 The output

When you press **F1**, you will see the results of the analysis of variance. The analysis of variance table (ANOVA table) is at the top of the screen. The first source of variation is labled **between** (which means between groups), and is sometimes called "treatment" or "model". The second source of variation is labled **within** (which means within groups), and is sometimes called "error" or "residual". The F-ratio given in the table is the test statistic for testing

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$$

and the p-value is next to it. Recall that small values of p are evidence *against* H_0.

5.7 Other tests

Most of the other tests in the "One, two and multi-sample tests" option are nonparametric tests. These are alternatives to the t-tests and F-test discussed here, and are most useful when the data do not seem to conform to the normal distribution.

Chapter 6

Linear Regression II

This chapter gives the details on doing a regression analysis. Throughout the chapter, we assume the model is

$$y = \beta_0 + \beta_1 x$$

To perform a regression analysis, you must have a dataset with at least two numeric variables (x and y). If you are using a set of data given in class, **please turn to p. 2 and follow the directions under "Selecting a dataset"** To enter your own data at the keyboard, read Chapter 11.

We will demonstrate linear regression using the dataset "gessel.sx", described at the beginning of Chapter 1. The y variable will be the gessel score (a measure of mental ability), and the x variable will be the child's age in months when he or she first spoke. Briefly, we want to predict a child's gessel score from knowing the age of first speech.

6.1 Scatterplots and Linearity

For a given set of data, x and y may or may not be linearly related. To see if a linear relationship is plausible for your data, you may want to start your analysis by looking at a scatterplot of y vs. x. Scatterplots are described in more detail in Section 4.1, p. 19; what follows is a brief description. Choose **Summary Statistics** on the base menu and **Scatter plots** on the next menu. Type the variable names: xvariable,yvariable in the option screen, and press **F1**.

6.2 The regression (least squares) equation

Choose **Linear models** on the base menu, and Linear regression on the next menu. This brings up an Options Screen where you describe the analysis you want. On the second section of the screen, type the name of the *dependent* or Y variable. (You can also choose the Y variable from a list: use **F2**, **arrow keys** and **Enter** to select a variable, and use **ESC** to exit the list.) In the next section, type the name of the *independent* or X variable. (You can choose this from a list also – see above.) Then press **F1** to see the results.

The first part of the output is the **Coefficients Table**. This is a page of regression statistics, including the regression coeffieients, some measures of precision (including r^2) and an ANOVA table. The calculated value of **intercept** is listed under **Coefficients** beside "constant". The calculated value of **slope** is also under **Coefficients** beside the x-variable name. For example, the Coefficients Table for gessel.sx is shown below.

```
STATISTIX 4.0                                GESELL, 08/11/94, 13:58

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF SCORE

PREDICTOR
VARIABLES    COEFFICIENT    STD ERROR     STUDENT'S T       P
---------    -----------    ---------     -----------    ------
CONSTANT         109.874       5.06780         21.68       0.0000
AGE             -1.12699       0.31017         -3.63       0.0018

R-SQUARED            0.4100   RESID. MEAN SQUARE (MSE)   121.505
ADJUSTED R-SQUARED   0.3789   STANDARD DEVIATION         11.0229

SOURCE         DF        SS           MS          F         P
----------    ---    ----------   ----------   -----    ------
REGRESSION      1      1604.08      1604.08    13.20    0.0018
RESIDUAL       19      2308.59      121.505
TOTAL          20      3912.67

CASES INCLUDED 21   MISSING CASES 0
```

From it, we see the regression equation is $\hat{y} = 109.874 - 1.12699x$ or

$score = 109.874 - 1.12699 age.$

6.3 Testing $\beta_1 = 0$

The Coefficients Table gives two statistics for testing $H_0 : \beta_1 = 0$ (i.e. no significant regression between Y and X) vs $H_a : \beta_1 \neq 0$.

T-test: The value of **Student's t** next to the x-variable name is the calculated value of $t = b_1/s_{b_1}$, and the p-value for testing $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$ is next to it. A small value of p is evidence against H_0, so a small p-value suggests that y and x are significantly related.

F test: Equivalently, we can test $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$ by using the F statistic in the ANOVA table at the bottom of the screen. In a simple linear regression, the t-test and F-test are equivalent.

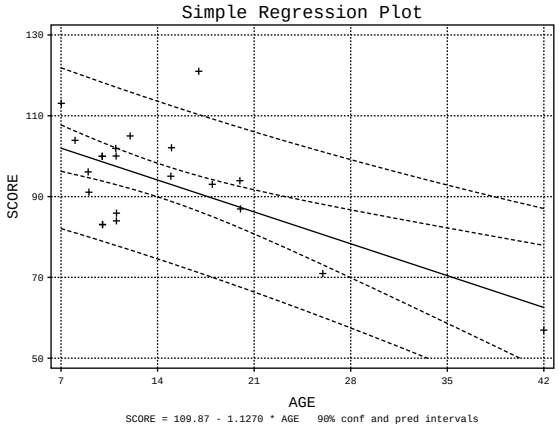
For the gessel.sx data, we have t=-3.63 (or F= 13.20) and p=0.0018.

Note: At this point in the regression analysis, a lot of additional information is available from the **Results menu**. You invoke this by hitting the **ESC** key. Some of these features are mentioned below.

6.4 Plotting the fitted line

From the Coefficients Table, press **ESC** to bring up the **Regression Results** menu. Then select **Plots**, and choose **Simple regression plot** and examine the resulting graph. It displays the data points, the regression line and 95% confidence bounds on the regression line, and 95% prediction bounds on the individuals.[1] The prediction bands are always wider than the confidence bands, so the outside bounds are the prediction bounds. An example regression plot is on the next page.

[1] If you want confidence bands for a confidence level other than the default 95% , return to the **Regression Results** menu (by hitting **ESC**) and select **Interval Coverage**. Type in the confidence level (anything from 50 to 99) and press **F1**. The new confidence level will be used in all subsequent reports.



6.5 Residuals Plots

Residuals plots are useful for assessing the appropriateness of the linear model, for assessing the reasonableness of the error assumptions, and for spotting outliers.

6.5.1 Residuals vs. Fitted Values

A plot of (standardized) residuals versus fitted values ($\hat{y}_i$) is produced by selecting "Std. resids by fitted values" on the **Plots** submenu. **ESC** returns you to the Plots submenu.

6.5.2 Normal plot of residuals

A normal probability plot of the residuals checks to see if the errors are normally distributed. Choose "Rankit plot" on the **Plots** submenu. Is the plot linear? Random scatter about a line is no cause for alarm, but systematic curvature indicates non-normal errors, and suggests that the linear model may not be valid.

Note: Other residuals plots can be produced by using the **Save residuals** option. You must then return to the base menu and choose "Summary statistics" to get the desired scatterplots. See Section 2.4 for details of these scatterplots.

6.6 Predictions

You can use your regression equation to predict y. On the **Regression Results** menu (hit **ESC** if you are in a plotting screen), choose the "Predictions" option. If you want to compute $\hat{y}$ for a point already in the dataset, use the **case method** by entering **C** followed by the case number of the observation. If you want to compute $\hat{y}$ for a new point (not in the dataset), type **V** for **value method** and enter the x value.

The resulting output will list $\hat{y}$, 95% confidence bounds (for the average value of y at this level of X), and 95% prediction bounds (for an individual Y at this level of X). If you want a confidence interval other than the default 95% value, see the footnote on p. 31.

Chapter 7

Multiple Regression

The multiple regression model is of the form

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

and the model can also contain terms of the form $\beta_i X_i^2$ or $\beta_i X_i^3$ or $\beta_i X_i X_j$. If you wish to use a model with these more complicated forms in it, please see Section 7.4, p. 37.

Most of the procedures in multiple regression are the same as in simple linear regression. If you have not read Chapter 6 yet, please take a minute to go back and do so.

7.1 Specifying the variables

Start Multiple Regression by choosing **Linear models** on the main menu, and **Linear regression** on the submenu, just as you did for simple linear regression. In the option screen, you will need to type in the Y (dependent) variable and the X (independent) variables. You may specify as many X variables as you want. Simply type in the variable names one after another, leaving blank spaces in between. (Or use **F2**, **arrow keys**, and **Enter** to select variables from the list, and **ESC** to exit the list).

Generally, the last two sections of the options screen should be ignored. When you have typed in all your X variables, press **F1** to see the results.

The Regression Results Screen looks about the same as it did in simple linear regression. At the top is a list of all the parameter estimates (one

for each independent variable, plus the constant term $\hat{\beta}_0$), the estimated standard deviation of each, and test statistics for performing hypothesis tests on the parameters (see Section 7.2 below). R^2 is below this. The ANOVA table is at the bottom of the screen.

```
STATISTIX 4.0                                        CSDATA, 09/08/94, 16:25

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF GPA

PREDICTOR
VARIABLES    COEFFICIENT    STD ERROR     STUDENT'S T       P       VIF
---------    -----------    ---------     -----------    ------    -----
CONSTANT         2.58988       0.29424          8.80    0.0000
HSM              0.16857       0.03549          4.75    0.0000      1.5
HSE              0.04510       0.03870          1.17    0.2451      1.6
HSS              0.03432       0.03756          0.91    0.3619      1.9

R-SQUARED           0.2046      RESID. MEAN SQUARE (MSE)    0.48977
ADJUSTED R-SQUARED  0.1937      STANDARD DEVIATION          0.69984

SOURCE         DF       SS           MS          F        P
----------    ---   ----------   ----------   -----   ------
REGRESSION      3     27.7123      9.23744    18.86   0.0000
RESIDUAL      220     107.750      0.48977
TOTAL         223     135.463

CASES INCLUDED 224   MISSING CASES 0
```

For example, using the csdata.sx dataset, the regression equation is

$$\hat{y} = 2.58988 + .16857HSM + .04510HSE + .03432HSS$$

which could also be written as

$$\hat{y} = 2.58988 + .16857X_1 + .04510X_3 + .03432X_2$$

7.2 Testing $\beta_i = 0$

Sometimes in a multiple regression analysis, we need to decide whether or not a particular X variable, say X_i, is actually related to Y (i.e. actually belongs in the model). To test the hypothesis $H_0 : \beta_i = 0$ vs $H_0 : \beta_i \neq 0$, the test statistic $t = b_i/s_{b_i}$ is listed in the *Student's T* column of the Results table, across from the name of X_i. The p-value of the test is listed next to the t-stastic. In general, small values of p are evidence *against* H_0.

For example, to test $H_0 : \beta_3 = 0$, the test statistic is t = 0.914, which has a p-value of p=.3619. This is a fairly large p-value, so we have no evidence against H_0. Under these circumstances, it might be best to remove X_3 from the model.

7.3 Testing $\beta_1 = \beta_2 = \cdots = \beta_p = 0$

Before testing individual parameters in the model, it is a good idea to perform a *global test* to see if the model as a whole is useful at all in predicting Y. This test is usually written as $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ vs H_a: not H_0. This is an F-test, and the F-statistic to use is given in the ANOVA table at the bottom of the Results screen, with the p-value of the test next to the F-statistic.

In our example, F = 18.86 and p= 0.0000. This extremely small p-value indicates that the model is indeed a useful predictor of Y.

7.3.1 Residuals Plots

Residuals plots in multiple regression are done just as in linear regression. See Section 6.5, p. 32 for details. Briefly, press **ESC** at the Results screen to get a menu of Regression results options. Choose **Plots** to draw residuals plots.

7.3.2 Predictions

Predictions in multiple regression are also just the same as in linear regression. See Section 6.6, p. 33 for details. Briefly, press **ESC** to get the Regression Results menu. Choose **Predictions**. Choose either **C** for case

method (to find $\hat{y}$ for an observation in the dataset), or **V** for value method. If you choose the value method, you must type in a value for each independent variable. The output consists of two columns: in the first column is the predicted value ($\hat{y}$) and the upper and lower 95% prediction bounds for an individual, and the standard error of the predicted value. In the second column is the fitted value (E(y)), and the upper and lower 95% confidence bounds for the mean of y, and the standard error of the fitted value. To change the confidence level from 95% to some other value, see the footnote on p. 31.

7.4 Adding X_i^2 or X_iX_j to the model

To add terms to the model of the form X_i^2 (or X_i^3 or X_i^5) or terms like X_1X_2 or $X_1^2X_3$, first go back to the base menu (you may have to hit **ESC** several times). Choose **Data management** on the base menu and **Transformations** on the next menu. Since the terms listed above are simple transformations of variables in the dataset, you can define these terms very easily by typing in

x12 = x1 * x2

This will create a new variable called x12, which is just x1 multiplied by x2. Now, to include this variable in your model, go back to **Linear Models** and specify x12 as one of the independent variables.

Other variables can also be created this way. For example:

timesq = time^2 will add a new term timesq to the dataset that is really $time^2$.

mscore3 = mscore^3 adds $mscore^3$

newx = x1^2 * x3 adds newx, which is really $X_1^2X_3$.

7.5 Warning

In a multiple regression analysis, it is possible that you will accidentally include several X variables that are highly correlated with each other. For example, you might try to predict the amount of money a person spends on travel, using age, income, and gender as independent variables. In this

instance, age and income may be highly correlated. This situation, called *multicollinearity*, causes computational problems for the computer. When this occurs, the program will automatically remove one of the highly correlated variables from the model, so that it can proceed to do the calculations.

Chapter 8

One Way Analysis of Variance (ANOVA)

Analysis of Variance (also called ANOVA or AOV) is used in general to decide if the varying levels of a treatment (or several treatments) had any effect on the outcome of the experiment. If there is no effect of varying the level of treatment, the treatment means $(\mu_1, \mu_2, \ldots, \mu_p)$ should all be the same, or (equivalently) the treatment effects (differences between treatment means and overall mean) should all be zero. The treatment effects are denoted $\alpha_1, \alpha_2, \ldots, \alpha_p$, and we test the null hypothesis of no treatment effect by testing $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_p = 0$ vs H_a : not H_0. To perform a one-way ANOVA test (i.e. just one treatment at several levels), choose **Linear Models** on the base menu, and **One-Way AOV** on the next menu.

8.1 Specifying the variables

The data in your dataset must be arranged in either **Table** or **Categorical** format. In tabular format, each level of treatment has its own column of data. In Categorical format, all the treatments are in one column, with another column of values that indicate which category a particular treatment represents:

Table	
Male	Female
95	113
71	96
⋮	⋮
94	86
	100

Categorical	
Score	Sex
95	M
71	M
⋮	⋮
94	M
113	F
⋮	⋮
100	F

In the first section of the options screen, all the variables in the dataset will be listed. In the second section, type in **T** for Table or **C** for Categorical.

Table arrangement: if the data are arranged so that each treatment level is in a different column, this is called **T**able. In the last section of the options screen, type in the column names (two or more) that represent the treatments. Press **F1** to see the ANOVA table.

Categorical arrangement: if the data are all in one column, with a second column of categories, this is called **C**ategorical. In the next section of the options screen, type in the name of the dependent variable. This is the name of the column that has the results in it. In the last section of the options screen, type in the name of the categorical variable. This is the name of the column that names the treatment levels. Press **F1** to see the ANOVA table.

8.2 The ANOVA table

The ANOVA table is a summary of the calculations that are performed to test the hypothesis above. The philosophy behind ANOVA is that if the levels of treatment make a difference, then the distance between one treatment mean and another will be larger than the distance between observations within a treatment group. So, the test statistic is

$$F = MS(Between)/MS(Within)$$

which is sometimes written as

$$F = MS(Treatment)/MSE$$

This F-statistic and its p-value are given in the ANOVA table at the top of the screen. (The numerator and denominator degrees of freedom — ν_1 and ν_2 — are also in the table.) A small p-value is evidence against H_0, so a small p-value indicates that there are significant differences between the levels of treatment.

The rest of the screen contains other information for doing more complicated analyses of the data. At the bottom of the screen is a summary, containing the mean and standard deviation for each treatment group.

Chapter 9

General ANOVA

The one-way ANOVA procedure described in the previous chapter assumes there is just one treatment at several levels. The general ANOVA procedure is designed for more complicated ANOVA situations, where there may be several treatments, treatments may be crossed or nested, etc. This procedure is flexible and powerful, and only a small portion of its capabilities will be shown here. The Statistix 4.1 user's manual, pp. 196–224 contains a complete description of the procedure.

Choose **Linear Models** in the base menu and **General AOV/AOCV** in the next menu.

9.1 Specifying the Variables

In this procedure, the data must be entered "categorically": this means that all the outcomes of the experiment appear in one column of the data, and there are other columns (at least two) that specify the levels of treatment associated with that observation. If there are multiple observations (repeated measures) at some (or all) treatment combinations, there must be a column that counts the repeated measures. If your data are not in this form, you will have to fix this before you can perform the ANOVA. Please see Sections 9.3 and 9.4 (pp. 44–45).

In the second section of the options screen, type in the name of the dependent variable, or select it from a list, using **F2**, **arrow keys** and **Enter**, and then **ESC** to exit the list.

In the third section, type in the names of the treatment variables, one after the other, with blanks in between. You can specify interaction terms of the form t1*t2 (for the t_1t_2 interaction) or t1*t2*t3 for a three-way interaction. When you have entered all the treatment variables and combinations, press **F1** to see the results.

9.2 The ANOVA table

The resulting ANOVA table is similar to the output of a one-way ANOVA. If you are not familiar with this, please take a minute to read Section 8.2, p. 40.

The ANOVA table contains one F-ratio for each treatment variable in the model. This F-ratio is appropriate for testing whether or not the levels of that treatment have an effect. For example, in the dataset chromium.sx, the two treatments are CHROMIUM (at two levels) and EAT (at two levels). The output is given below.

```
STATISTIX 4.0                          CHROMIUM, 09/07/94, 15:56

ANALYSIS OF VARIANCE TABLE FOR ENZYME

SOURCE          DF       SS          MS          F        P
-------------  ----  ----------  ----------  -------  ------
CHROMIUM (A)      1    0.00152     0.00152     0.05  0.8649
EAT (B)           1    0.50552     0.50552    15.43  0.1587
A*B               1    0.03276     0.03276
-------------  ----  ----------
TOTAL             3    0.53980
GRAND AVERAGE     1    95.6680
```

The F-ratio for testing whether or not the two levels of chromium are significantly different is F = 0.05, with a p-value of p=.8649. Since this is a large p-value, we would suspect that chromium has little effect on the outcome of the experiment.

Notice that each treatment is assigned a letter of the alphabet (A, B, etc). This is for convenience in specifying interaction terms. Notice also that the last term A*B is the error or residual.

More complicated analyses (e.g. pairwise comparison of means) are available from the Results menu. Access this by pressing **ESC** at the Anova table.

9.3 Data in the wrong format

The general ANOVA procedure requires that the data be in a particular format: one variable for the outcome, and other variables indicating the level of each treatment for that observation. So, your data should look like this:

Temp.	Pressure	Strength
100	1	1.55
100	2	1.78
100	3	1.95
150	1	1.67
150	2	1.82
150	2	2.03

where "Strength" is the observed outcome of varying Temperature and Pressure.

Your data may have been stored differently. For example, the same dataset might have been stored as:

Temp	P1	P2	P3
100	1.55	1.78	1.95
150	1.67	1.82	2.03

To reformat the data, you need to use the "stack" procedure. This does exactly what it sounds like: it will take all the variables (in this case, they are all strengths) and stack them in one column. At the base menu, choose **Data management**, and pick **Stack/unstack** on the next menu. Choose "stack" in the second section of the options window. In the third section, list the variables you want to combine into one variable. (In our example, we would list P1 P2 P3). In the next section, give a name for the new "destination variable". (In our example, we might call this Strength). In the last section,

enter a name for a new variable, which will record which level of treatment is represented by each stacked observation. (In our example, we would call this Pressure). Then press **F1** to reformat the data. Now you are ready to perform the ANOVA.

9.4 Adding a variable for repeated measures

Statistix is different from most other statistical procedures in that if you have two observations at the same treatment combination, they must be distinguished from each other by a new variable (i.e. REP = 1 or 2). You can easily add this new variable to your dataset as follows. On the base menu choose **Data management** and on the next menu **Transformation**. The "cat" transformation is the one you want. In the bottom section of the options window, type

REP = cat(2,1)

All this function does is create a new variable REP which will be a sequence 12121212... long enough to fill the entire dataset. So, if you had three repeated measures at each treatment combination, use

REP = cat(3,1)

to create a sequence 123123123123... that will fill the dataset.

The first argument to cat controls the length of the sequence, and the second number controls how many times each number will appear in the sequence. So

REP = cat(5,2) creates a sequence 11223344551122334455... long enough to fill the dataset.

Chapter 10

Quality Control Charts

Statistix offers a variety of quality control charting options, including X bar, P, R, s, Pareto, U and MR charts. Only a few of these will be described here.

10.1 The X Bar chart

This is a plot of the sample mean over time. It is used to control the mean of the process (i.e. keep it "on target"), and may also indicate when the process variability is increasing. Choose **Quality control charts** on the base menu and X bar chart on the next menu. The program assumes that at each sample point, you have the individual values for that sample stored in separate columns. (If you want to make an X-bar chart and your data consist of $\bar{x}_1, \ldots, \bar{x}_n$, do not use the X bar procedure. Instead, see 10.1.2 below.) If all your data are in a single column, see 10.1.1 below to fix this. To make an X bar chart, enter the names of the columns containing the observations.

In the third section of the options window, you must specify **S** (standards given) or **C** (compute $\bar{x}$ and s for a retrospective chart).

Standards given: If you specify standards given, you must then enter μ and σ in the last section of the options window. Then press **F1** to see your chart.

Retrospective: If you specify **C** for a retrospective chart, you must then choose $\bar{R}$ or $\bar{s}$ as the estimate of standard deviation. Type "R bar" or "s bar" in the fourth section of the options window. Then type in the numbers of the first and last observations to be used in calculating the control limits.

Usually, in problems given in class, you will use the entire dataset to do these calculations. In practice, however, if you suspected that the process was going out of control at some point, you would *not* use data beyond that point to calculate the control limits. Press **F1** to see your chart. The estimated mean and standard deviation are printed at the bottom of the chart.

10.1.1 Data in the wrong format

Control charts are picky: Statistix assumes that your data will be stored in a particular format, with one column for each member of the sample, and one row for each time point. It may be that your data were all stored in a single column, with categories in another column that indicate which observation is which. Before you can make a control chart, you need to *unstack* your data. This does exactly what it sounds like: it will take your single column of data and make it into several columns, one for each member of the sample.

To unstack your dataset, return to the base menu and choose **Data management**, then choose **Stack/unstack** on the next menu. Type "unstack" in the second section of the options window. In the third section, type the name of the variable that has the values you want to unstack. In the fourth section, type the name of the categorical variable that tells which observation is which. In the last section, type the names for the new "unstacked" variables. Now you are ready to draw a control chart.

10.1.2 Making an X bar chart from $\bar{x}_1, \ldots, \bar{x}_n$

The usual procedure for making an X bar chart expects that you have the raw data — several observations at each time point. Sometimes the data have already been summarized, and all you have is $\bar{x}$ at each time point. You can still make an X bar chart, but you must use a different procedure. At the base menu choose **Quality control charts** and on the next menu choose **I chart**. **I** stands for individual. In the second section of the option window, type in the name of the variable that contains the sample means. In the third section, type **S** for a "standards given" chart. In the next section, type in the mean and standard deviation. **Be sure to divide the standard deviation by $\sqrt{n}$.** Press **F1** to see your chart.

10.2 R charts

These charts of the sample range are used to control the process variability. Choose **Quality control charts** on the base menu and **R chart** on the next menu. As in the X bar chart, the program assumes that at each sample point, you have the individual values for that sample stored in separate columns. If all your data are in a single column, see Section 10.1.1, p. 47 to fix this. Type the names of the observation columns in section 2 of the options window. Enter **S** (standards given) or **C** (retrospective) in section 3.

Standards given: If you specify standards given, you must then enter μ and σ in the last section of options window. Then press **F1** to see your chart.

Retrospective: Type in the numbers of the first and last observations to be used in calculating the control limits. Usually, in problems given in class, you will use the entire dataset to do these calculations. In practice, however, if you suspected that the process was going out of control at some point, you would *not* use data beyond that point to calculate the control limits. Press **F1** to see your chart. The estimated σ used to calculate the control limits is printed at the bottom of the chart.

Chapter 11

Entering Data

Most of the analysis done in class will involve datasets that are already on the computer. To use a dataset already on the computer, please see p. 2. However, in a project to analyze your own data, you will need to type it in and store it yourself. To enter data in Statistix, go back to the base menu. Choose **Data management** on the base menu and **Data entry** on the next menu.

11.1 Naming the variables

The computer is going to set up a spreadsheet-like area in which you can enter your data. Before it can do this, you must name the variables. In the second section of the options window, type in all the variable names, separated by spaces. If you want a character (string) variable, type it in like this: name(S20) — this makes "name" a character variable 20 characters long. If you want an integer variable, type it in like this: caseno(I) — this makes "caseno" an integer variable. When you have typed in all your variable names, press F1 to set up the spreadsheet.

11.2 Entering the Data

When you have typed in the variable names and pressed F1, you should get a screen that has observation numbers down the side and your variable names across the top. The cursor will be positioned in the first variable field. If this

field is numeric, it will contain an "M" (M stands for "missing", because you haven't entered a value for that variable yet). Simply type in the value, and use the **Enter** key to move from one field to the next. The arrow keys will also move the cursor around the screen. When all the data has been entered, press **F2** to access the top menu, and move the cursor to **Save**. A window will appear, where you can type in the name of the file in which to save your data. If you are using a public computer, the data will be saved in Drive C. If you would like to save the data on your own disk instead, move the cursor to the "drive designator" (near the bottom of the screen), and change the drive designator to **a:**. Put your disk in the floppy drive before pressing **F1** to save the data.

11.3 Using Data from a Floppy Disk

To use data that you have already saved on your own disk, choose **File management** on the base menu and **Retrieve** on the next menu. Change the "drive designator" near the bottom of the screen from **C:\ SX*.SX** to **a:**, put your data disk in the floppy drive and press **F1**. Statistix will read your data disk, and print a listing of the data files on the screen. Choose the one you want. If your data file was not created in Statistix, it is probably in Ascii format. Please see Section 11.5, p. 51, on importing ascii files.

11.4 Transferring Data

If you have entered your data on a public computer, you should save it on your own disk. This is very important: **the disk space of the public computers is erased frequently!** To copy your data onto a floppy disk, first leave Statistix. This is done by pressing the **ESC** key (perhaps several times) until you are returned to the base menu. Pressing **ESC** once more will exit the program. Then insert your disk into the floppy drive. At the C:\ > prompt, type **copy** *filename*.**sx a:*** where "filename" should be the name of the file where you stored your data.

11.5 Importing and Exporting Files

Statistix stores data files in its own format. A data file that you entered in the Statistix spreadsheet cannot be read by any other program, nor can it simply be printed as it stands. You must *export* the file before you can do this. Statistix has the capability to export data files in ascii, comma and quote ascii, formatted ascii, Excel, Quattro Pro, and Lotus 1-2-3 formats. To do this, choose **File management** in the base menu, and **Export** on the next menu.

Similarly, files created in some other program must be *imported* to Statistix. To do this, choose **File management** in the main menu and **Import** on the next menu. For details of importing and exporting files, see Chapter 3 of the Statistix® 4.1 manual[1].

[1]The Statistix® 4.1 manual is published by Analytical Software.

Appendix A

List of data files

The following is a list of data files available on the computer. This includes all the data sets referenced in **Introduction to the Practice of Statistics**, 2^{nd} ed. by Moore and McCabe. Descriptions of the data can be found on the indicated pages of that text.

Many datasets from **Statistics for Engineering Problem Solving** by Stephen Vardeman are also available. They are not listed here, because they are stored under their respective numbers. For example, the dataset used in Problem 7-38, p. 441 of the text is stored as **7_38.sx** on the computer.

Page no.	Filename	size	Page no.	Filename	size
3	NEWCOMB	470	33	AIRLINE	814
22	CRANK	186	36	HOTDOG	833
26	ELDERLY	567	88	BASEBALL	559
27	GROWING	406	91	FOOTBALL	927
27	RAIN	437	91	NZWINE	410
28	EARTH	242	105	LOTTERY	6239
30	CORN	257	112	CARS	1536
31	RIOTS	319	112	BANKS	402
32	DEATHS	208	113	LEAN	342
32	RUTH	277	114	BOTULISM	311

Page no.	Filename	size	Page no.	Filename	size
116	COLOR	1025	419	JOE	222
116	TOMATO	256	420	CIRCUIT	105
118	KALAMA	158	420	MOORE	728
132	GESSEL	244	444	DRP	272
137	CONSUME	129	470	RADON	119
140	FLOW	156	488	STAM	12016
142	POLLUTIO	446	492	DIALYSIS	50
146	GAS	1304	492	WINE	66
147	OIL	328	523	HAND	421
159	POP	331	524	MLA	307
160	VEHICLES	164	528	PIGS	512
164	FOSSIL	66	534	DIRECTED	246
177	DATES	72	543	SYSTOLIC	1394
182	SEAFOOD	514	552	SSHA	252
192	POULTRY	2190	569	POTENCY	119
195	WABASH	13219	586	SYCAMORE	1409
196	PENALTY	3612	594	TREES	14901
207	BATTING	389	596	TWINS	3427
212	REGRESS	394	620	STOCK	766
214	EXHAUST	1166	622	BUSINESS	4816
215	TEACHER	1513	626	SEVENTH	2508
244	MIGRANE	697	632	LOAN	179
246	WEIGHT	339	632	PEOPLE	223
254	BOTTLES	210	634	WOMEN	131
255	STAT	494	651	DENSITY	1853
256	CLUB	739	674	MANATEE	260
269	SCORES	106	675	PISA	155
289	CANDY	98	678	NBA	809
317	SIZE	99	682	OHM	72
318	FIFTH	130	683	OXYGEN	277
417	DIAMETER	181	725	PRETEST	328
418	CERAMIC	208	786	TENSION	257

Page no.	Filename	size
791	CHEESE	898
792	CSDATA	8551
793	WOOD	912
794	READING	1686
795	MAJORS	12924
	HOUSE	658
	ACTIVE	410
	JOBS	1337
	CRAPS	263