Using CrunchIt (http://bcs.whfreeman.com/crunchit/bps4e) or StatCrunch (www.calvin.edu/go/statcrunch)

1. In general, this package is far easier to use than many statistical packages. Every so often, however, there are inexplicable problems. If you have problems with printing or anything else, feel free to call me or e-mail me. Please do not spend 2 hours, e.g., trying to get something to print. Ask for help.

2. CrunchIt will sometimes not work with Internet Explorer. It will ask for Java to be downloaded, or some such message. Unfortunately, downloading Java does not help. However, if you use Mozilla Firefox, CrunchIt works like a charm.

3. To enter or get data:

a. Homework problems: After arriving at the CrunchIt website, on the left you will find a list of chapters. Select the one which contains the desired data. Then select the table or problem number on which you are working. (If you don't find the problem #, then the data are probably in a table. Read the problem again to find out for sure. Occasionally you need to enter it by hand.)

b. Entering data by hand: You can simply type the data into the rows and columns. Data are generally entered by putting the information for each individual in one row with one column per variable.

e.g.

Name	Age	Height
Mary	4	36
Gord	5	42

You can change the column headings from "var1" "var2" etc. to Name, Age by clicking on the appropriate column heading, backspacing to get rid of var1, and typing in the new column heading.

c. Getting data from an excel or other file: Go to the *Data* menu. Select *Load Data*, then *From File*. Browse to find your data file. Check to see if the first line should be the column headings, and then click *okay*. (Note that if the column headings are more than one word, the headings may be split up and one word will head each column. You will need to do some editing in this case.) Alternatively, you can use the *Data/Load Data/From Paste* option, after copying your data onto your clipboard.

4. To print:

Select *Options*, then *Print*. Alternatively, you can copy the output and paste it into a Word document. To do that, select *Options*, then *Copy*.

If the labels on graphs are too small to read, first expand the window and then select *Options* and print again. (The labels should automatically re-size.)

4. Bar graphs: to see the distribution of a categorical variable (e.g. car color)

Under *graphics*, choose *bar plot*. You then have a choice between *with data* and *with summary*. If your data are in the form of categories and counts/percents (i.e. you have the number or percent in each category already), then you should choose *with summary*. Then tell the computer which column the categories are in, and which column contains the counts/percents. Click *next*, then *next* again, and type in some labels (e.g. x: Car Color. Y: # (or % of) cars) and a title (Distribution of car color). Create the graph.

If your data are simply all in a long column, and the computer needs to count the number falling into each category, you should choose *with data*. Fill in the requested information: which column contains the data, (and after selecting *next* twice) axis labels (the x-axis label will be the variable you are doing a bar graph of, such as Car Color. The y-axis label will be Count, or Frequency, or #Cars, . . .), and title ("Distribution of car color"). Create the graph.

We passed by a couple of options: *Where* and *Group by*. Let's say we looked at car colors for Fords and Hondas. If you wanted a bar graph of car color for only the Fords, then you would use the "where" option, and type in something along the lines of "cartype=Ford". If you wanted a separate bar chart for Fords and Hondas, you would use the "Group by" option, and you would choose the variable "cartype" as the *Group by* variable.

5. Pie charts: to see the distribution of a categorical variable (e.g. color of car)

Under *graphics*, choose *pie chart.* You then have a choice between *with data* and *with summary*. If your data are in the form of categories and counts, then you should choose *with summary*. If you have a column of data, and the computer needs to count the number falling into each category, you should choose *with data*. Fill in the requested information: which columns contain the data, title, . . . and create the graph.

6. Histograms: to see the distribution of a quantitative variable (e.g. #doctors, duration of Old Faithful eruptions)

Under *graphics*, choose *histogram*. Select the column you want a histogram of, select *next* three times, give the histogram axis labels (the x-axis should be the variable name e.g. #doctors, and the y-axis should be Count, or Frequency, or #States, . .) and a Title ("Distribution of the #doctors/100,000 in the 50 states). Occasionally you will want to use the *group by* option (found on the first histogram page—below where you select the variable you want a histogram of). This would be if, e.g. you wanted to see if the distribution of #doctors varied by parts of the U.S. (south, west, etc.).

If you want to specify where the bins start and the bin widths, you can do that after the first *Next*.

7. Stemplot: to see the distribution of a quantitative variable

Under *graphics*, choose *stem and leaf*. Select the column that contains the quantitative data. Occasionally you will want to use the *group by* option, if you want to see how the distribution changes for a few groups.

8. Descriptive statistics: mean, sd, 5#summary

Under *Stat*, choose *summary stats.* You probably have your data in columns, so choose *Columns.* Select the column that contains the data for which you want summary statistics, and click *Calculate*. If you want to compute the summary stats separately for a few groups (e.g. males and females), then next to *group by*, specify which variable contains the groups.

9. Scatter plot: to see the relationship between two quantitative variables

Under *graphics*, choose *scatter plot*. Select the x-variable (predictor) and the y-variable (response). Occasionally you will want to have different color points for some groups; then specify which variable contains the groups in *group by*. As you hit *Next*, you will have various choices (e.g. points or lines. Choose points). Add axis labels and a title. Create the graph.

10. Correlation: a numerical measure of the strength and direction of the linear relationship between two quantitative variables.

Under *stat*, choose *summary stats*, then *correlation*. Specify which variables you want to correlate, and click on *calculate*.

11. Simple linear regression: y=a+bx

Under *stat*, choose *regression*, then *simple linear*. Specify the x-variable (predictor) and the y-variable (response). Click *next*. Predict y for a certain x, if you want. Click *next*. Select *plot the fitted line*. This will show your regression line and the data points. Also select *histogram of residuals* and *residuals vs x-values*. This will allow you to check the residuals for a normal distribution, patterns, outliers, etc., so we can see if the simple linear model is appropriate. Click *Calculate*.

12. Tables (counts, %, chi-square statistic): to see if there is a relationship between two categorical variables

If your data are entered in long columns, such as

<u>Gender</u>	<u>Smoker</u>
M	Yes
M	No
M	No
F	No
F	Yes
.	.

then select *Stat*, *Tables*, *Contingency*, *with data* to get going.

Select your row variable and your column variable (e.g. Gender and Smoker). Click *next*. Display *row percents*, *column percents,* and *expected count.* Click *Calculate*. This will give you counts, row and column percents, expected counts if the null hypothesis is true, and the chi-square statistic.

If your data have already been summarized (the counts in each cell determined), type them into CrunchIt as follows:

Row	Gender	Yes	No
1	M	50	40
2	F	40	30

Select *Stat*, *Tables*, *Contingency*, *with summary* to get going. Specify the columns for the table (**Yes** and **No**, from the above data). Tell which column the row labels are in (**Gender**, from the above data). Type in a name for the column variable (e.g. **Smoker**). Click *Next*. Display *row percents*, *column percents,* and *expected count.* Click *Calculate*. This will give you counts, row and column percents, expected counts if the null hypothesis is true, and the chi-square statistic.

13. Generating random numbers (for sampling or assigning treatments)

Select *Data*, then *Simulate Data*, then *Uniform*. By Rows:, type **100** (actually, any number larger than your population size). By columns:, type **1**. For a, enter **1**, and for b, enter the number of people or objects in your population. Select *simulate*. A new column of data will appear, with numbers in random order. You can ignore the decimal part, or round, whichever you prefer, and use the numbers to choose which people will be in your sample, or to assign treatments to people.

14. t-procedures (drawing inferences about means)

a. <u>one-sample</u>

1. confidence interval: Under *Stat*, select *T statistics*, then *one sample*. Select the column (variable) for which you want a confidence interval (CI). Click on *next*. Select *Confidence Interval*, and then specify the confidence *level* (.90, .95, .99). Click on *calculate*.

2. hypothesis testing: Under *Stat*, select *T statistics*, then *one sample*. Select the column (variable) for which you want to test a hypothesis about the population mean. Click on *next*. Select *Hypothesis test*. Specify the null hypothesis (mean=_____) and the alternate hypothesis (≠, >, <). Click on *calculate*.

b. <u>paired samples</u>

1. confidence interval: Under *Stat*, select *T statistics*, then *paired*. Select the column that contains sample 1, and the column that contains sample 2. Click on *Next*. Select *Confidence Interval*, and then specify the confidence *level* (.90, .95, .99). Click on *calculate*.

2. hypothesis testing: Under *Stat*, select *T statistics*, then *paired*. Select the column that contains sample 1, and the column that contains sample 2. Click on *Next*. Select *Hypothesis test*. Specify the null hypothesis (mean diff=_____) and the alternate hypothesis (≠, >, <). Click on *calculate*.

c. <u>two-sample</u>

1. If the two samples are in separate columns:

a. confidence interval: Under *Stat*, select *T statistics*, then *two samples.* Tell the computer which columns (variables) contain the two samples. Unselect *pool variances* (we are using the unequal variances approach, to be safe). Click on *next*. Select *Confidence Interval*, and then specify the confidence *level* (.90, .95, .99). Click on *calculate*.

b. hypothesis testing: Under *Stat*, select *T statistics*, then *two samples.* Tell the computer which columns (variables) contain the two samples. Unselect *pool variances* (we are using the unequal variances approach, to be safe). Click on *next*. Select *Hypothesis test*. Specify the null hypothesis (mean diff=_____) and the alternate hypothesis (≠, >, <). Click on *calculate*.

2. If the two samples are in one column, and there is another column which specifies which group they are in:

a. confidence interval: Under *Stat*, select *T statistics*, then *two samples.* For both sample 1 and sample 2, give the column (variable) that contains the quantitative variable. By *where*, type the variable that contains the groups, and give the name of one of the groups (e.g. **group=1**) by sample 1, and the name of the other group (e.g. **group=2**) by sample 2. Unselect *pool variances* (we are using the unequal variances approach, to be safe). Click on *next*. Select *Confidence Interval*, and then specify the confidence *level* (.90, .95, .99). Click on *calculate*.

b. hypothesis testing: Under *Stat*, select *T statistics*, then *two samples.* For both sample 1 and sample 2, give the column (variable) that contains the quantitative variable. By *where*, type the variable that contains the groups, and give the name of one of the groups (e.g. **group=1**) by sample 1, and the name of the other group (e.g. **group=2**) by sample 2. Unselect *pool variances* (we are using the unequal variances approach, to be safe). Click on *next*. Select *Hypothesis test*. Specify the null hypothesis (mean diff=_____) and the alternate hypothesis (≠, >, <). Click on *calculate*.

15. z-procedures FOR PROPORTIONS (Note that the z-statistics option under STAT is for inferences for means when the population standard deviation is known. We will not be using this.)

   a. <u>one-sample</u>

      1. confidence interval: Under *Stat*, select *Proportions*, then *one sample*. If you already know the #successes and n (the sample size), select *with summary*. Type in the #successes+2 and the total #observations+4 (for the Agresti-Coull +4 method), and click *Next*. Select *confidence interval* and the level of confidence you want *(.90, .95, .99*). Click on *Calculate*.

   With StatCrunch (but not CrunchIt), you can select the Agresti-Coull method, so you don't need to add +2 to the successes and failures.

   If your data are in a column, and you want the computer to add up the #successes and the sample size for you, select *with data*. By *Outcomes in:*, specify which variable contains your data, and then by *Success*, specify which group you consider the "success." Click *Next*. Select *confidence interval* and the level of confidence you want *(.90, .95, .99*). Select *Next*, and use the *Agresti-Coull* procedure (this is the +4 procedure). Click on *Calculate*.

      2. hypothesis testing: Under *Stat*, select *Proportions*, then *one sample.* If you already know the #successes and n (the sample size), select *with summary*. Type in the #successes and the total #observations, and click *Next*. Select *hypothesis testing.* Specify the null hypothesis (prop=_____) and the alternate hypothesis (≠, >, <). Click on *calculate*.

   If your data are in a column, and you want the computer to add up the #successes and the sample size for you, select *with data*. By *Outcomes in:*, specify which variable contains your data, and then by *Success*, specify which group you consider the "success." Click *Next*. Select *hypothesis testing.* Specify the null hypothesis (prop=_____) and the alternate hypothesis (≠, >, <). Click on *calculate*.

   b. <u>two-sample</u>

      1. confidence interval: Under *Stat*, select *Proportions*, then *two samples*. If you already know the #successes and n (the sample size) for both groups, select *with summary*. Type in the #successes and the total #observations for both groups, and click *Next*. Select *confidence interval* and the level of confidence you want (*.90, .95, .99*). Click on *Calculate*.

      If the data from your groups are in two columns, and you want the computer to add up the #successes and the sample size, select *with data*. Specify which columns contain the two samples, and what you consider to be the "success." Click *Next*. Select *confidence interval* and the level of confidence you want (*.90, .95, .99*). Click on *Calculate*.

      If the success/failure data are in one column, and your groups are in a second column, then again select *with data.* Specify the column which contains the success/failure data for both Sample 1 and Sample 2. Specify what you consider to be the "success." By *Where* for Sample 1, specify your first group ("gender=M"), and by *Where* for sample 2, specify your second group ("gender=F"). Click *Next*. Select *confidence interval* and the level of confidence you want (*.90, .95, .99*). Click on *Calculate*.

      2. hypothesis testing: (See the information by confidence intervals to get the data ready for analysis.) Select *hypothesis testing* instead of confidence interval. Specify the null hypothesis (prop diff=_____) and the alternate hypothesis (≠, >, <). Click on *calculate*.

16. goodness-of-fit (the chi-square test for goodness-of-fit: checking to see if a categorical variable has a specific distribution)

Type in two columns of data: first the observed count in each of the cells (i.e. groups), and then the expected count in each of the cells if the categorical variable had the specified distribution. The expected counts may be the total sample size divided by the #cells (e.g. to see if the total # crimes was evenly distributed over the course of a week), or they may be the total sample size * the percent expected in that cell.

Once your data are in two columns, select *Stat*, then *Goodness-of-fit*, then *Chi-square test*. Specify which column holds your observed counts, and which column holds your expected counts. Then select *Calculate*.

17. Multiple regression

When we have multiple predictors for one response, we use multiple regression. To do this in StatCrunch or CrunchIt, select *Stat*, then *Regression*, then *Multiple Linear*. Select your X variables (predictors). Specify your Y variable (response). Click *Next*. Select *Save residuals*. Click on *Calculate*.

Do scatter plots of the residuals (saved in a new column) vs your predictors variables.

18. Analysis of Variance

When we have two or more groups (generally 3+), and want to compare them on the mean of a quantitative variable, then we use ANOVA. To do ANOVA in CrunchIt or StatCrunch, select *Stat*, then *ANOVA*, then *One Way*. If the data from your groups are in separate columns, select those columns, and click on Calculate. If the quantitative variable (responses) are in one column, and the grouping information is in a different column, select *Compare values in a single column*. Specify which column contains the responses, and which column contains the groups. Click on Calculate.

If $p \leq \alpha$, we reject the null hypothesis of equal means, and we need to figure out which means differ significantly from each other. In CrunchIt, we would do this with several two-sample t-tests, comparing the means of the groups two at a time. First, choose a new (lower) α, since we're doing multiple comparisons (according to Bonferroni, the new α= old α / #tests. E.g., if you have 3 groups, you will be doing 3 pairwise comparisons (A vs B, A vs C, and B vs C). So divide $\alpha/3$, and this is your new α. Now, carry out the two-sample t-tests, following the directions given previously. In StatCrunch, you can select Tukey HSD with confidence level .95. This will give you a confidence interval for the difference between each pair of means.