COMPUTER LAB MANUAL
(Windows 95 Version)

STATISTICS 10

Professor Cochran's Section Only!
Winter, 2000

You only need to download this if:

- **You also purchased ActivStats**
- **You are going to install ActivStats on your own computer (you do not need this manual for the Stat Lab computers)**

Table of Contents

Introduction: Details, Details

This version of the lab manual is for use on a PC computer where ActivStats is installed. The manual for actual use with the Mac computers in the Statistics Computer Lab covers the **exact same thing--you do not need both of these**. The only difference is this manual includes some of the special things you need to use the PC version and the pictures here are what you will see in a Windows environment. The Mac manual is on the class website and it is in your reader for the course.

Each lab has several specific objectives. By the conclusion of the discussion section we hope that you will be able to accomplish each of these. However, as this is the first time using this manual it is difficult to know exactly how long each lab will take. If not all objectives are achieved in the section, we will make adjustments as the quarter progresses. We thank you for your patience--we know it's tough to be part of an experimental teaching experience, but in the long run your involvement will help to improve undergraduate education here at UCLA.

Each lab assumes you have mastered the objectives of the preceding lab. So if you miss a discussion section, you should do the missed lab work either during the hours the computer lab is open for general use or by using the ActivStats program on your own computer. Do this *before* coming to the next discussion section because the TA will focus on the current objectives during the section and you'll probably have trouble keeping up with the class.

The statistical program we are using is DataDesk. If you are using this program outside of the computer lab, you will have to access DataDesk through another program called ActivStats. We will not actually use ActivStats in this course; we will only use DataDesk. ActivStats contains many interesting materials for learning basic statistics. You can use it to supplement your learning if you desire, but you are not responsible on exams or homework for anything that comes only from ActivStats. DataDesk has lots of capabilities. We have time however to cover these only in minimal fashion. Feel free to explore the program at your leisure

In the manual there are three types of statements:

- Things you need to do, commands you actually enter are typed in this font
  - It may be an instruction to do something like point the mouse and choose an option
  - It may be a command you have to enter by the keyboard
    - Sometimes it will include the actual statement you enter
    - Sometimes a part of the command may be idiosyncratic to you: for example, if your personal file is actually called newdata, it will be written generically in the manual as **filename**. You should always edit the bolded word to the correct word on your disk.
- Screen options or words that appear in DataDesk are typed in this font
- Comments describing things or what you'll see happen are typed in this font

Indications to Click on something means to point the mouse at an option and press the mouse key once or twice (it depends on the option how often you have to do this). It is assumed that you are storing all your work on a floppy disk in your A drive.

Lab 1: Getting Up and Running

Lab 1's objectives

- Log on to the system in the computer lab
- Set up Eudora for your use
- Open DataDesk
- Save a stored data set in your personal folder
- Produce printed output (optional)

Log on to the system in the lab

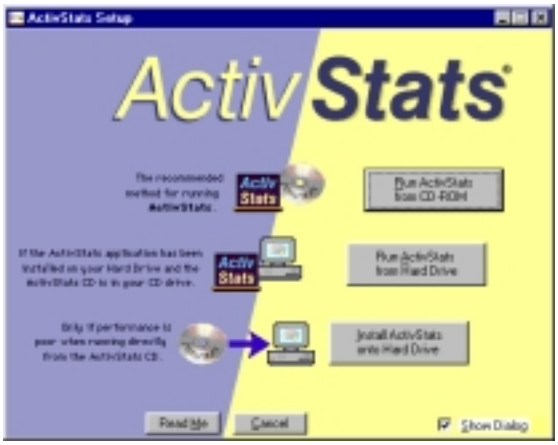
Computers in the lab are Mac computers. See the Mac Manual in your reader or on the class web site for instructions on how to do Lab 1's objectives in the lab environment. What follows here are specific instructions on how to access DataDesk on a PC after you have installed it. This lab manual version shows you how to accomplish in the PC environment the objectives of Lab 1 (with the exception of saving to a personal folder in the Lab).

First, follow the simple instructions that come with the CD-ROM to install ActivStats on your computer.

Open ActivStats at home

1. Click on the icon for ActivStats or open ActivStats via your Programs List on the Start menu.

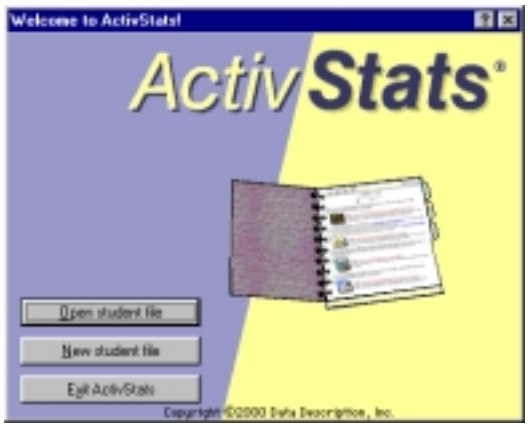
   Up will come this screen:



2. Choose **Run ActivStats from CD-ROM**

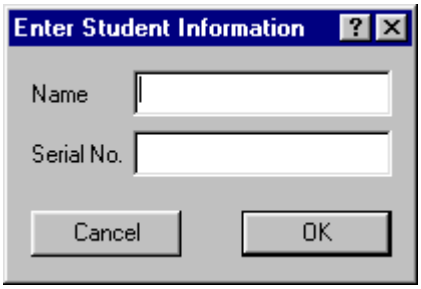
Run ActivStats for the first time

When you click the ActivStats icon, the first screen shown below will appear. The first time you use ActivStats, you will have to create a student record file for yourself.



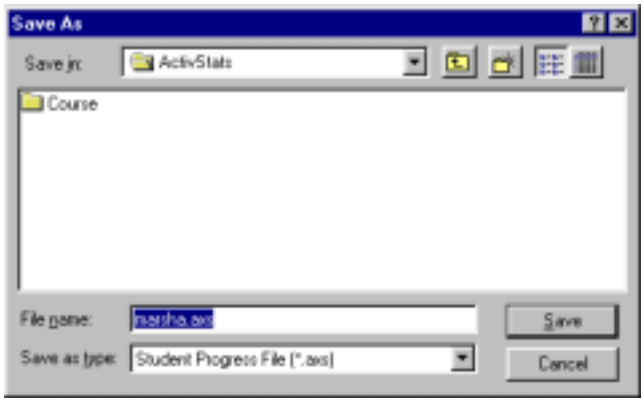
1. Choose New student file

   The following screen will appear:



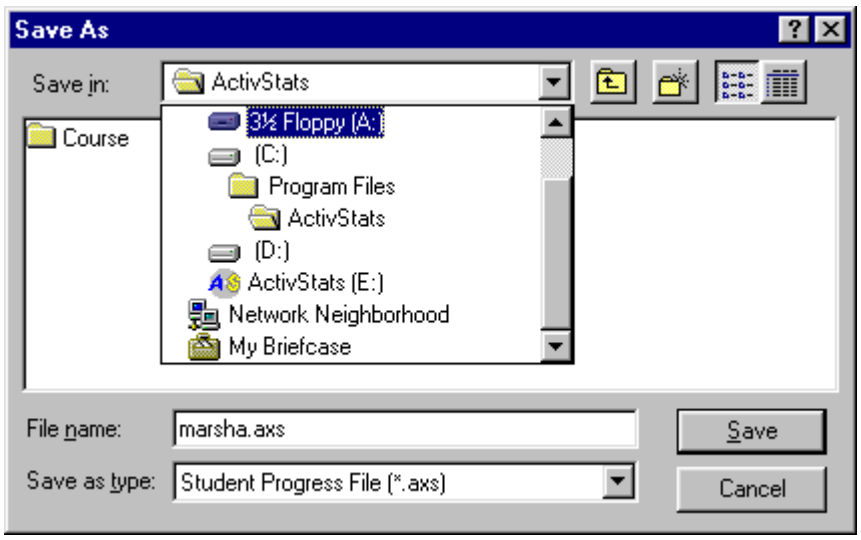
2. Enter your first name (Here "Marsha" is entered but you should use *your* name).
3. Then for the Serial Number enter the number included in your ActivStats package
4. Click OK

   The following screen will appear:



5. Save the file (the following instructions show how to save to the A drive)

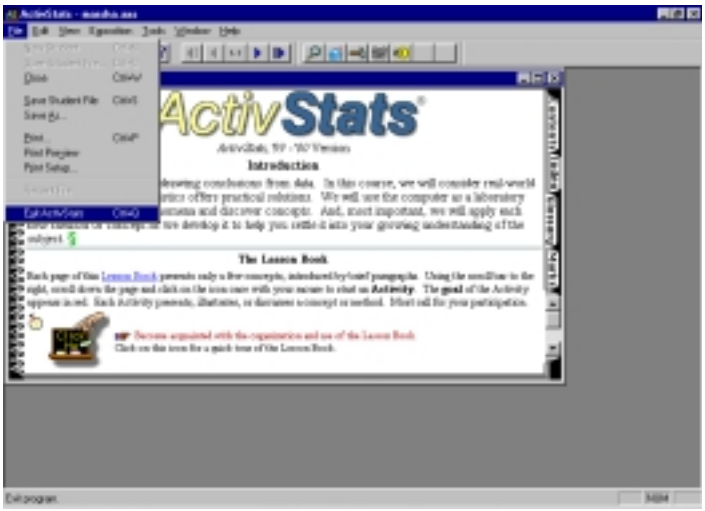
Scroll down the Save in option to highlight the A drive
Click on the Floppy A option
You'll see the computer search for the A drive (the light will go on in the drive).
Then click Save



Now you have created a student record file in which ActivStats can store your work.
Exit ActivStats so that you can do the next task.

6. Exit the program:

Highlight File at the upper left corner of the screen
Choose Exit ActivStats

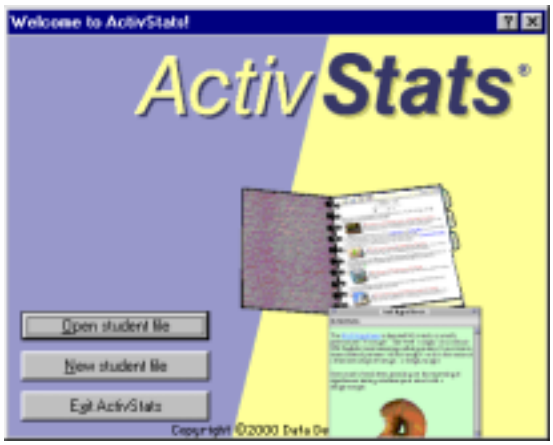


Run ActivStats when you have an existing student record file

Now, let's run ActivStats with your existing student record file. This is how you will run ActivStats from now on. (If you lose your work file, just go back to the preceding step and create a new student record file.)

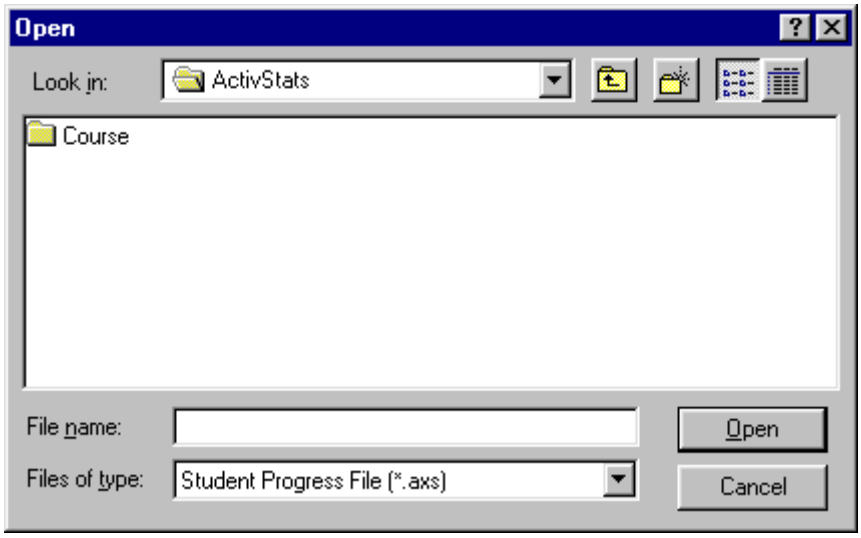
1. Choose the ActivStats icon

   Up will come this screen:



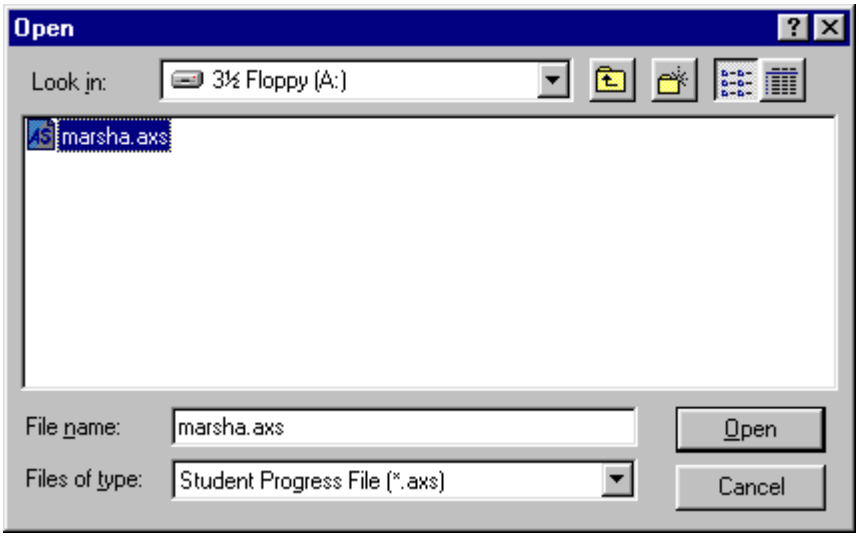
2. Choose Open student file

   The following screen will come up:



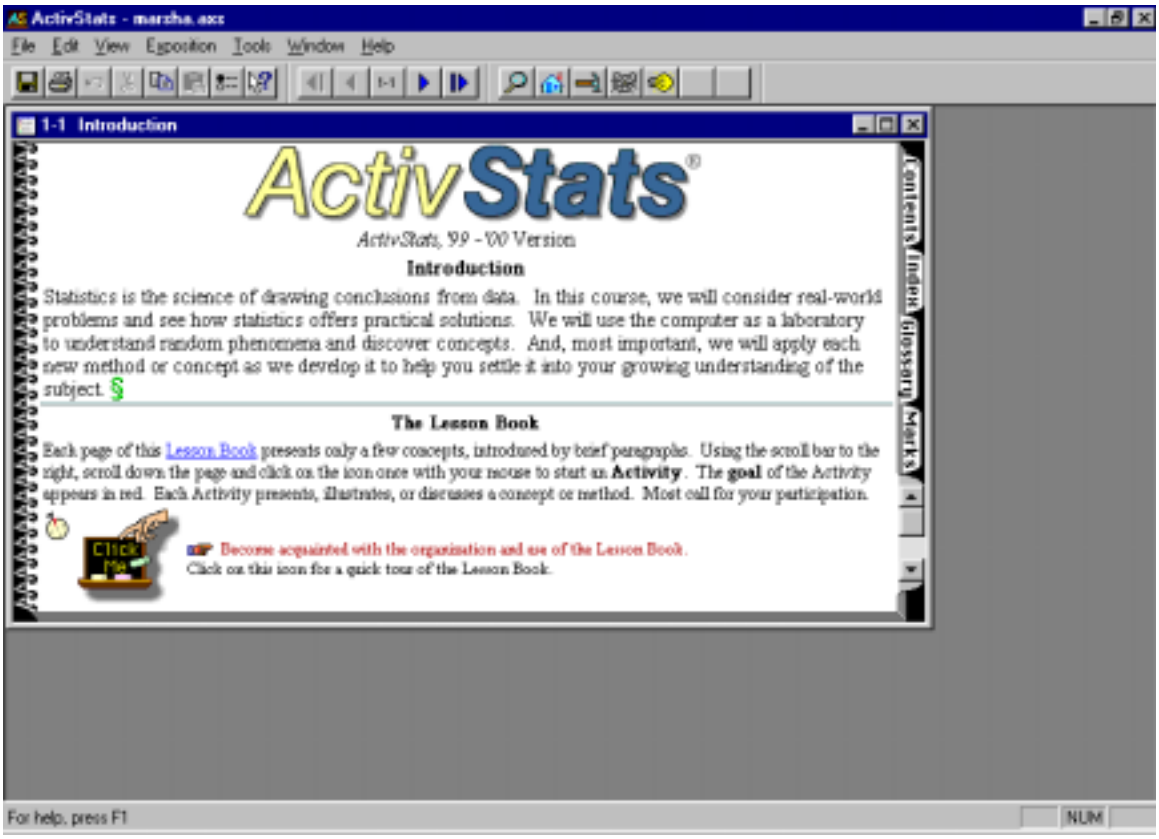
3. Tell the computer to Look in the A Drive for your file (If that is where you have stored it…)

4. Click on your file which is named **filename**.axs (**filename** = your name)



5. Click on Open to open up your file

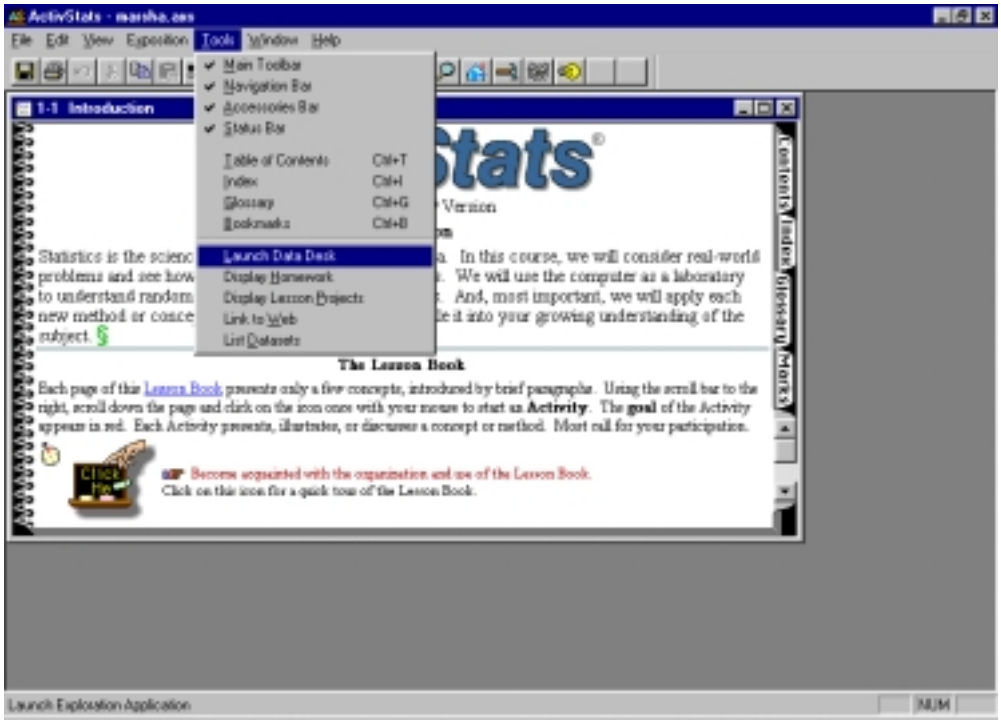
The following screen will appear:



Now, there are many things you can do with ActivStats. But we are only going to use DataDesk, a simple statistical software program contained within ActivStats

Open DataDesk

1. Under Tools, click on Launch Data Desk to start the program



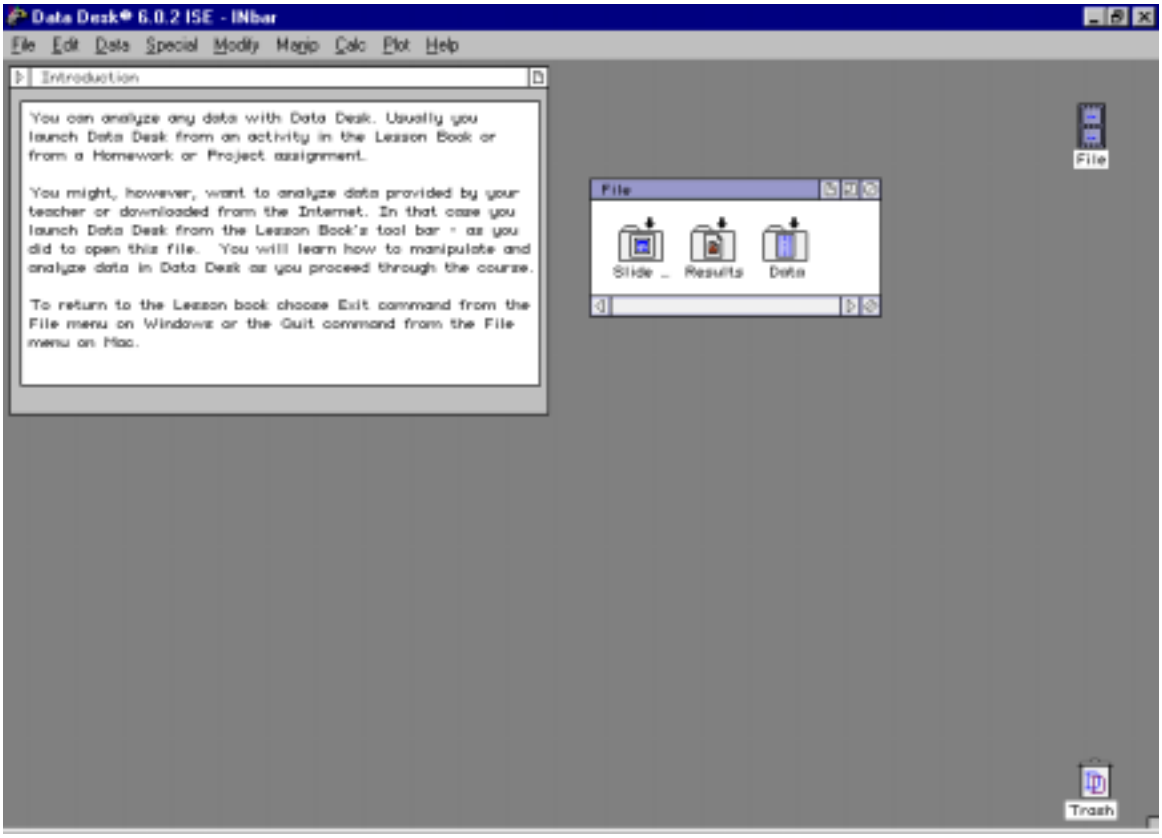
The following screen will appear:

Notice on the right there are two things. At the bottom is a **trash can** where you can throw away files and things you do not want.

At the top right is an icon called **File**.

2. Click on File

Now the following appears (Notice the new open File window):



If all three items in the File window are not shown, you can use the scroll bar on the bottom of the File window to see all three. Notice at the top of the File window on the right there are three iconic choices. The far right one minimizes or closes the window (you have to click on File to open the window again). The choice just to the left of that changes the size of the window and may move it to a new location on the screen.

Inside the window are three icons. "Slide" refers to slide show (ActivStats is a self-instructing program that contains modules to teach you statistics--we will not use these in this course). "Results" is the window that will contain your results from analyses. "Data" is a window where you can find your raw data. In data analysis, we begin with data that is put through a calculation process and output into results.

At this point, all of these three icons do not contain any information.

Retrieve an existing data set from the web to your diskette

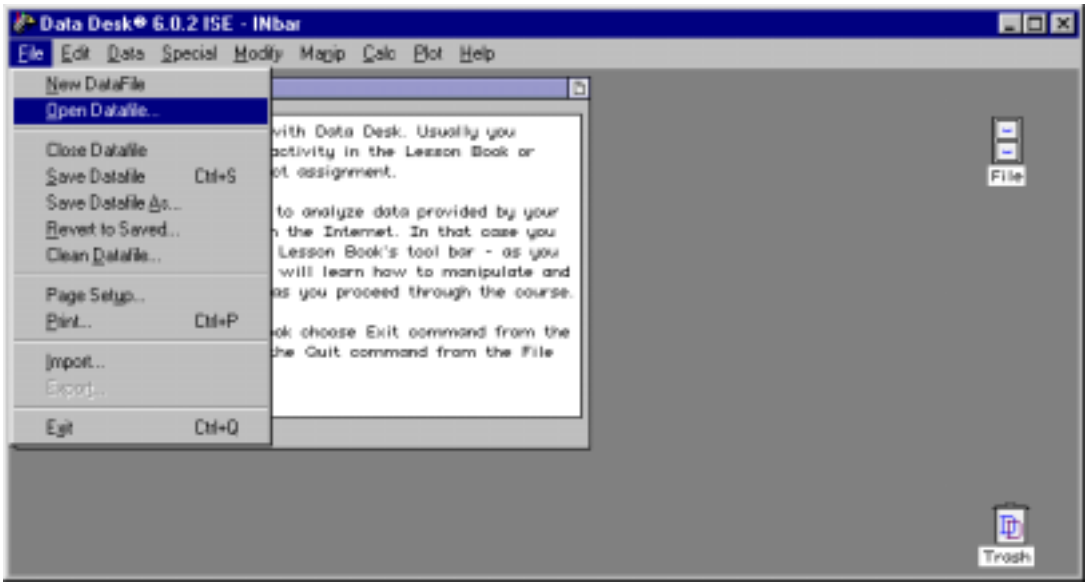
I have stored a small data set for you on the web. This data set has information about the ages of all U.S. Presidents when they took office. We'll use this data set in an upcoming lecture. The name of the file is prezages.dsk. It is stored in the class web site under Datasets. Go to the class web site and download this file onto your computer. To download the file, click on it and follow your browser's instructions to save to disk.

Open an existing data set

Now, you are going to open this file and take a look at it in DataDesk.

1. Open the data file by:

   Choosing File and clicking on Open Datafile

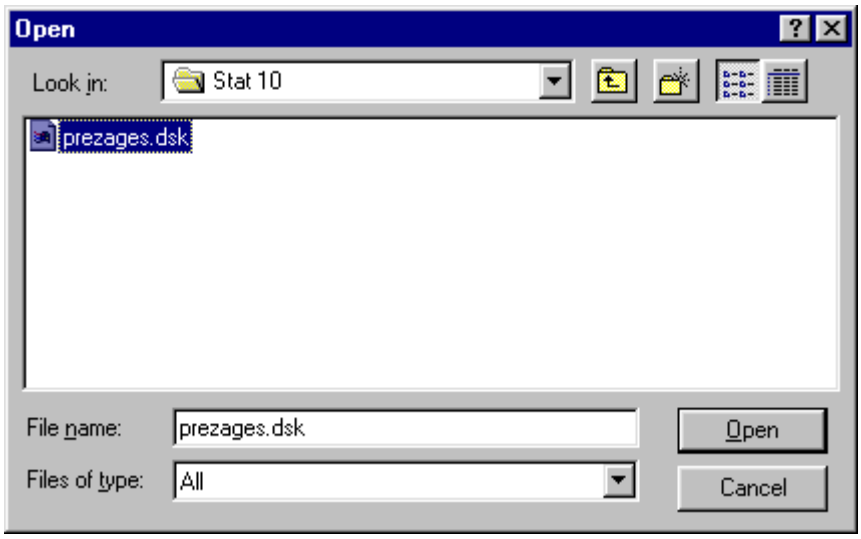


2. Click on prezages.dsk in the location you have stored it on your computer

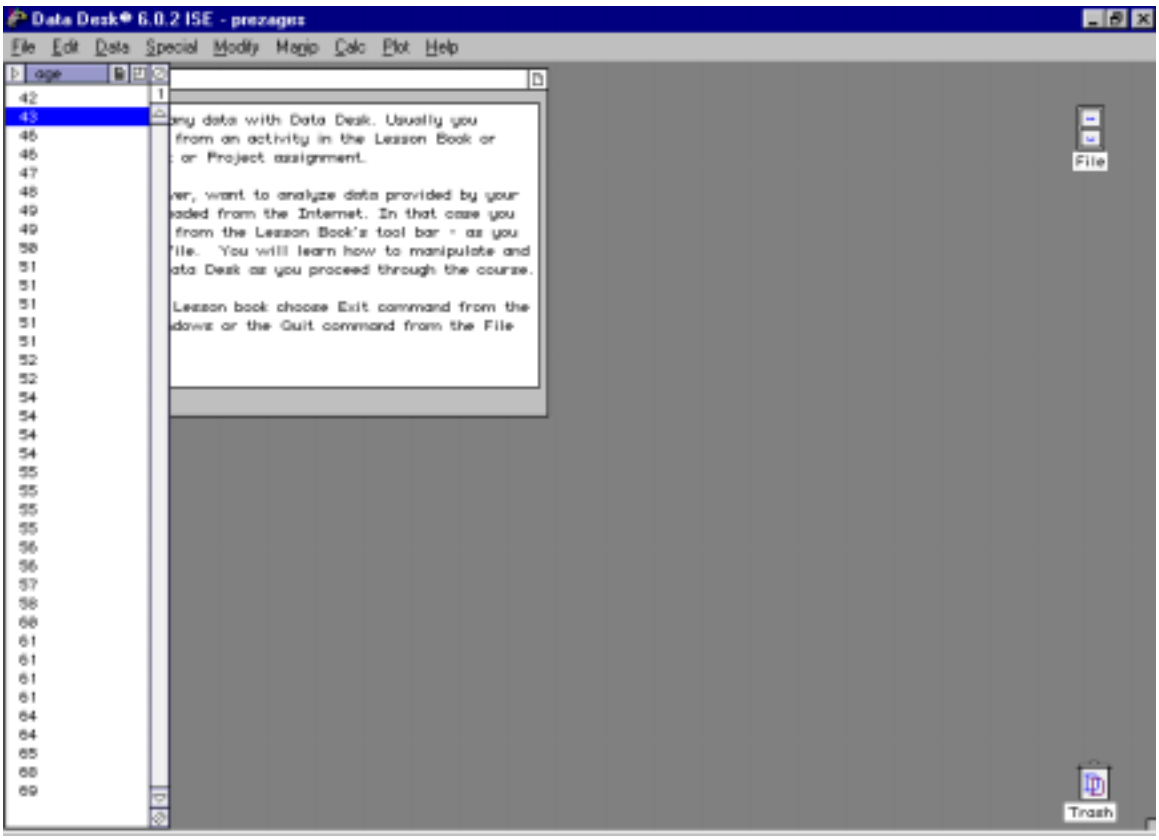
   (If you are doing this at home with no access to the web, the raw data is in the overheads for the course lectures under Lecture 4. You can enter it manually later, after Lab 2 where you will learn how to enter new data.)

2. Open the file

(Here the file is stored in the folder, Stat 10, but you need to look in the folder where your stored it.)

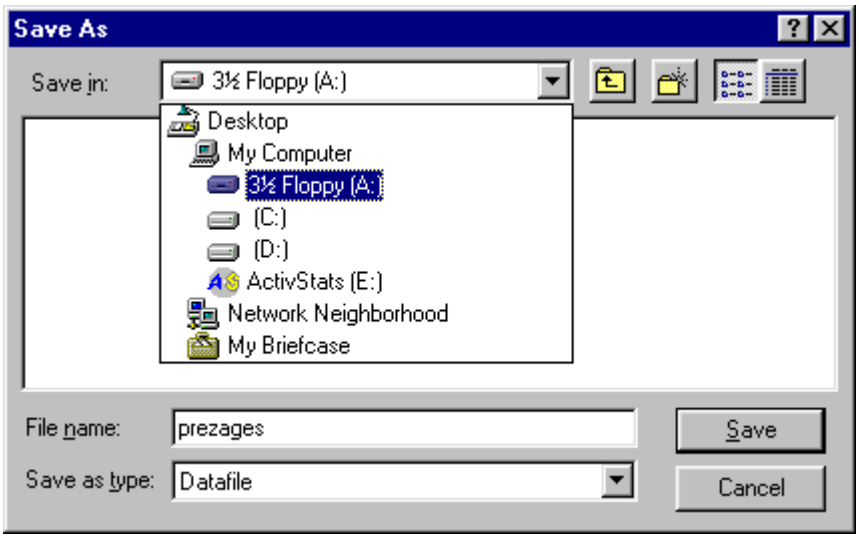


Up will come the following screen. If all the ages do not show, click the little middle icon in the upper right corner of the Age screen just to the left of the minimizing icon. That will resize the box.



Save an existing data set to your diskette

1. Save this dataset to your diskette by choosing the Save As option under File and selecting the A drive



Produce printed output

You can also print out this data set.

1. Select the Print option under File (File is in the upper left hand corner of the screen)

What prints out is the list of Presidents' ages.

Read output from the Stat Lab on your PC

When you are in the computer lab, you can mail your datasets and output to yourself as an attachment to an e-mail. Just send an e-mail to yourself and attach the file you wish to send. Then read it from your own PC.

Send files to yourself at the Stat Lab

This is just the reverse. Send yourself an e-mail with the file as an attachment. Then read it in the computer lab.

Lab 2: Creating a Variable

Lab 2's objectives

- Create a new variable
- Examine summary statistics
- Save your work to disk

Create a new variable

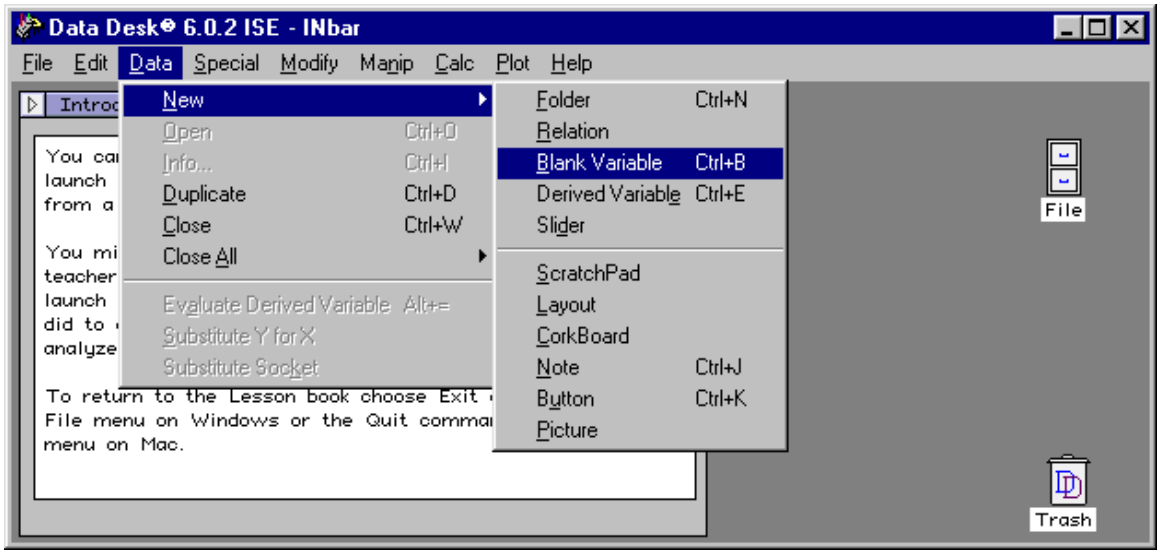
1. Open your student record file (see Lab 1, pp. 6-7)
2. Launch (or start up) DataDesk (see Lab 1, p.8)

Now you are in the DataDesk program and you are going to enter some new data that you will analyze. The data comes from Review Exercise 1 at the end of Chapter 4 in your textbook. There you are asked to:

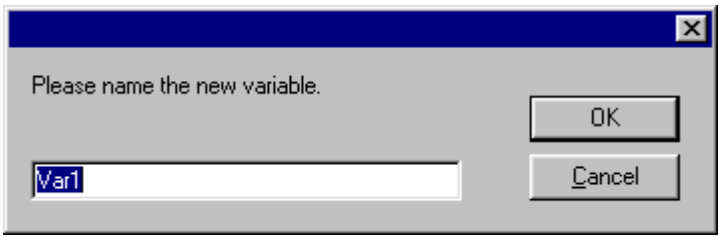
"Find the average and SD of the list 41, 48, 50, 50, 54, 57."

So, here you have one variable that contains 6 elements. The first step is to enter this information into a **Blank Variable** window.

3. To create a new variable
   Highlight Data
   Highlight New
   Choose Blank Variable



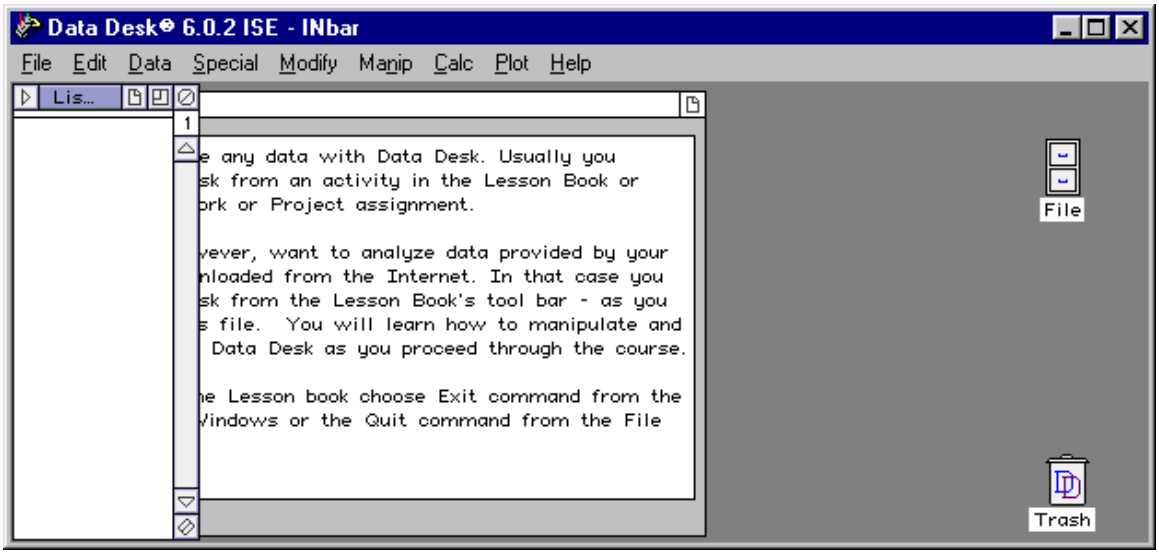
The following will appear:



Var1 is a generic name for the variable (as in Variable 1). You can use this name, but over time as you accumulate more variables, it helps to change variable names to those that carry more information for you. This will help you later on to work more efficiently with less confusion. Let's call this new variable you are creating **List1**.

4. Edit **Var1** to **List1** and click on OK

The following will appear:

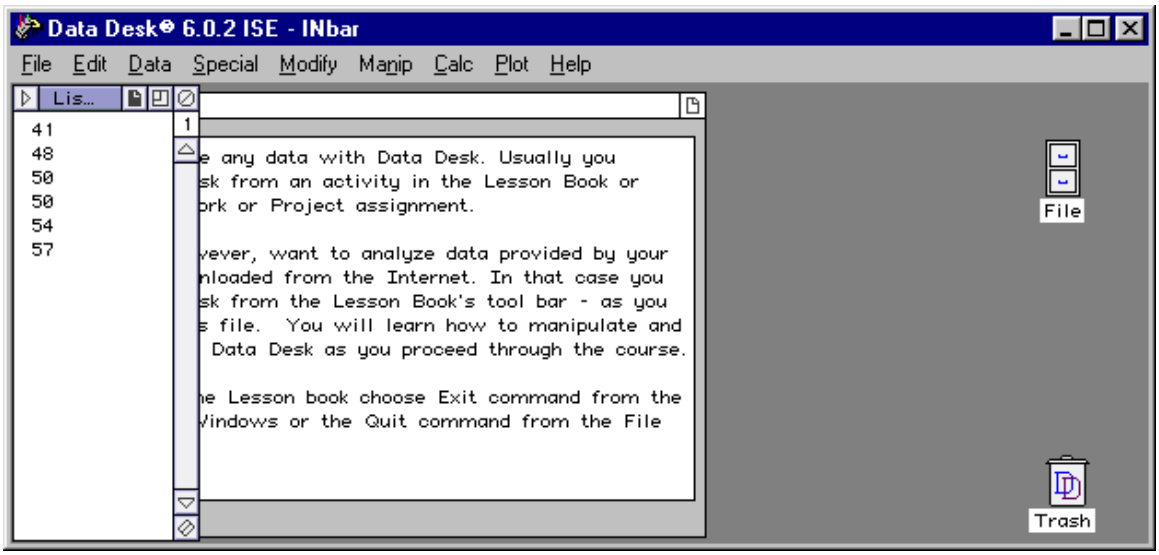


This is the window where you will store your 6 elements. At the top is the name of the variable, **List1**. The blinking cursor is waiting for you to enter the information.

5. Type in the 6 elements hitting the return key after each entry like this:

41 (Hit Enter)
48 (Hit Enter)
50 (Hit Enter)
50 (Hit Enter)
54 (Hit Enter)
57

Now the screen should look like this:



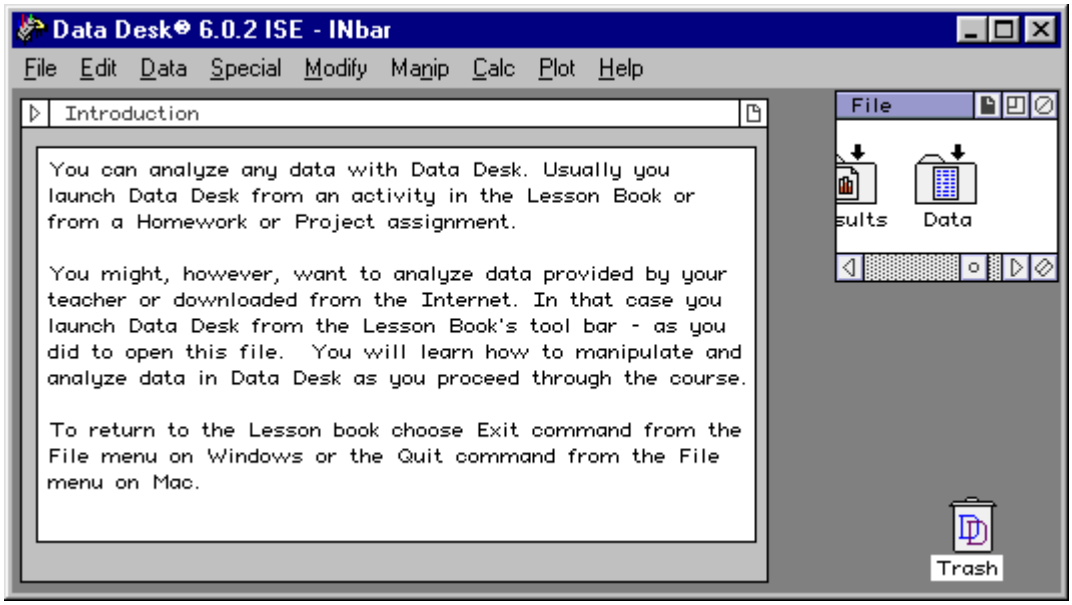
6. Close the window by clicking on the circular icon in the right side of the top bar where the variable name, List1, appears

The information has returned to the icon File (the file cabinet) on the right side of the screen.

Let's look at it.

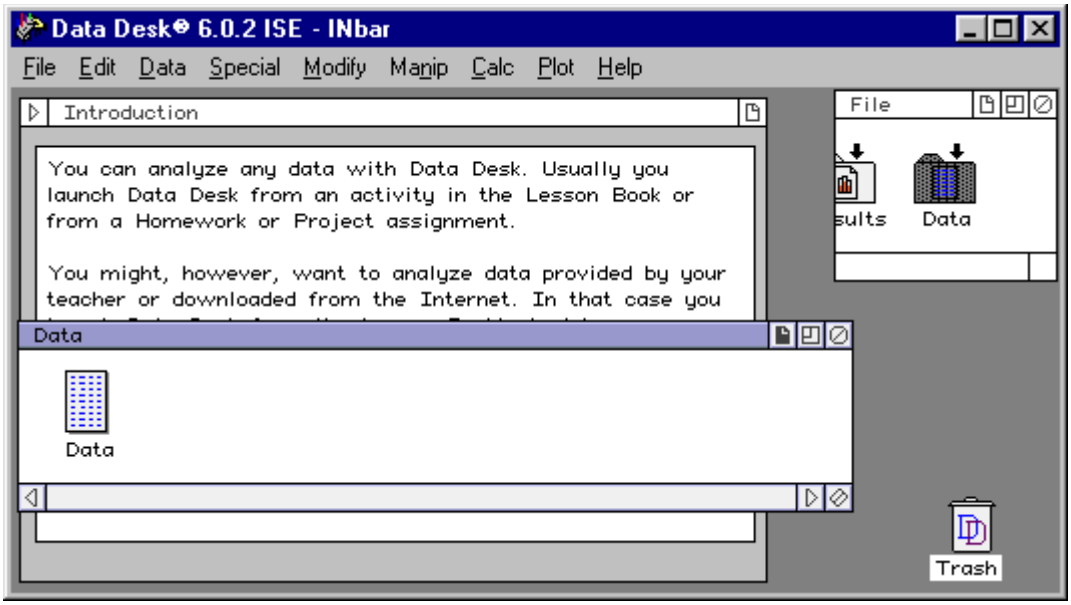
7. Click twice on File

The following appears. Use the bar at the bottom of the **File** window to slide over to expose the **Data** icon if you have to.



8. Click twice on the Data icon to open the Data window

The following will appear:

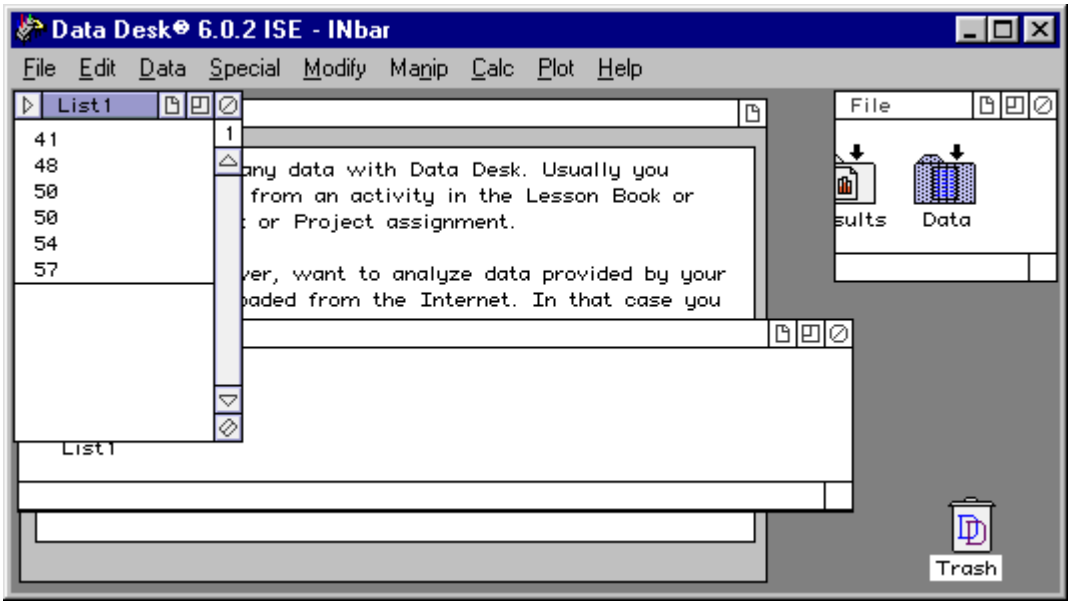


9. Click once on the Data icon in the Data window and a Y will appear

The **Y** refers to a dependent variable. You have only one variable (a list of 6 numbers) and so it is your dependent or outcome variable. If you had two or more, you could assign variables to be independent (**X**) or dependent (**Y**). This will be covered later.

10. Click twice more and the variable List1 appears as an icon

11. Click twice more on List1 and the 6 values you entered appear as shown below



12. Close all the open windows to return to showing just the File cabinet

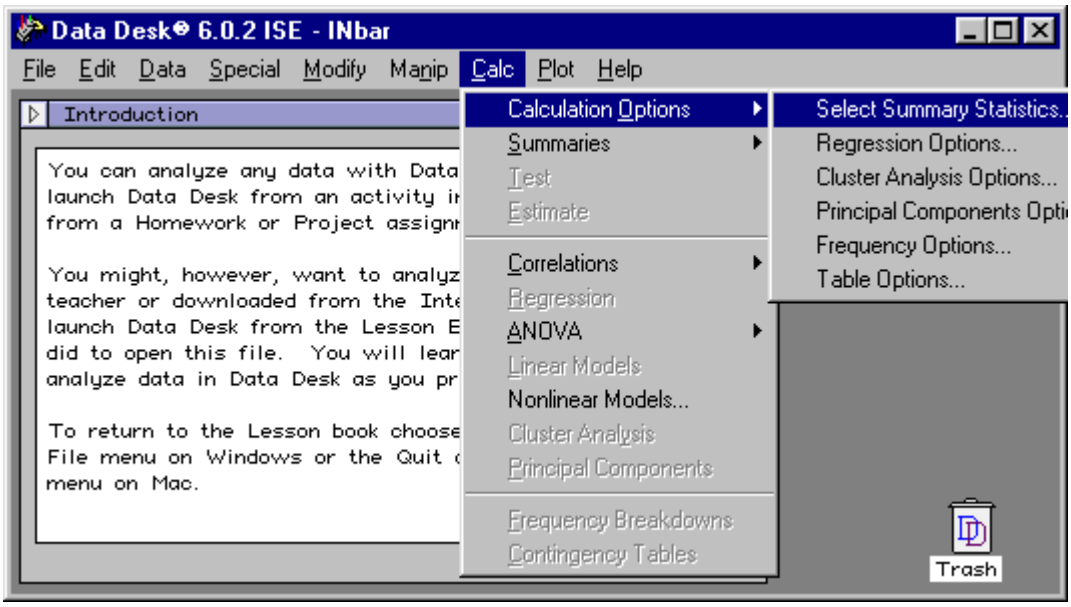
You have now created a new variable.

Examine summary statistics

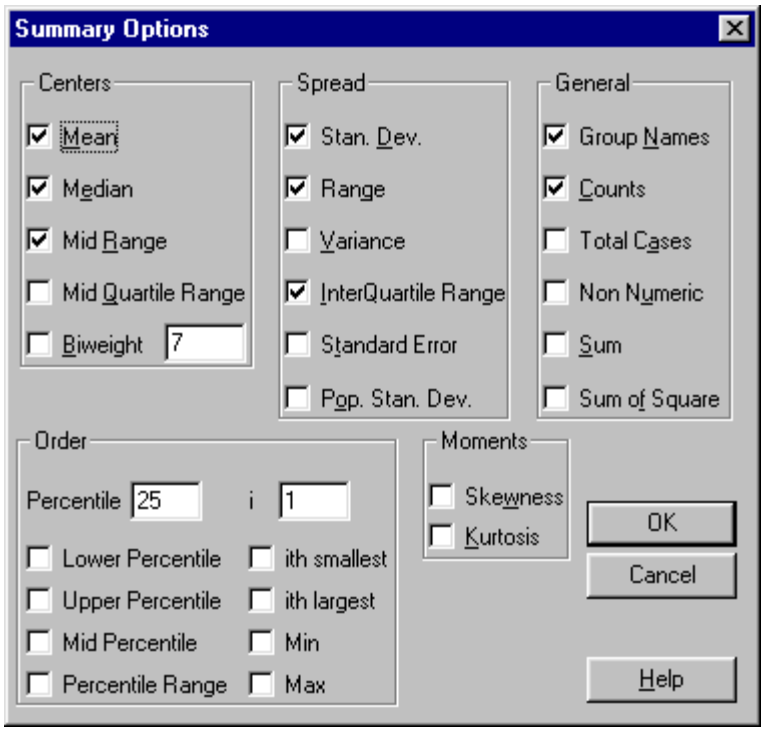
Now we are going to calculate the summary statistics. Summary statistics is another way of saying **descriptive statistics**. The first step is to tell the program what statistics you want.

1. To select the **summary** (or descriptive) **statistics** you want:

   Highlight **Calc**
   Highlight **Calculation Options**
   Click on **Select Summary Statistics**



The following window will appear:



The program is going to very quickly calculate for you whatever you want. In this instance you want the mean and Standard Deviation.

Notice that the program offers you 5 choices of summary statistics: Centers, Spread, General, Order, and Moments. Right now, all you need are two statistics, an estimate of center and an estimate of spread. You also want to be certain that the program is reading all 6 elements entered and no more (select counts or total cases--in this instance both should report back 6).

There are 5 types of center estimates to choose from. You want the mean. Make sure it is checked off.

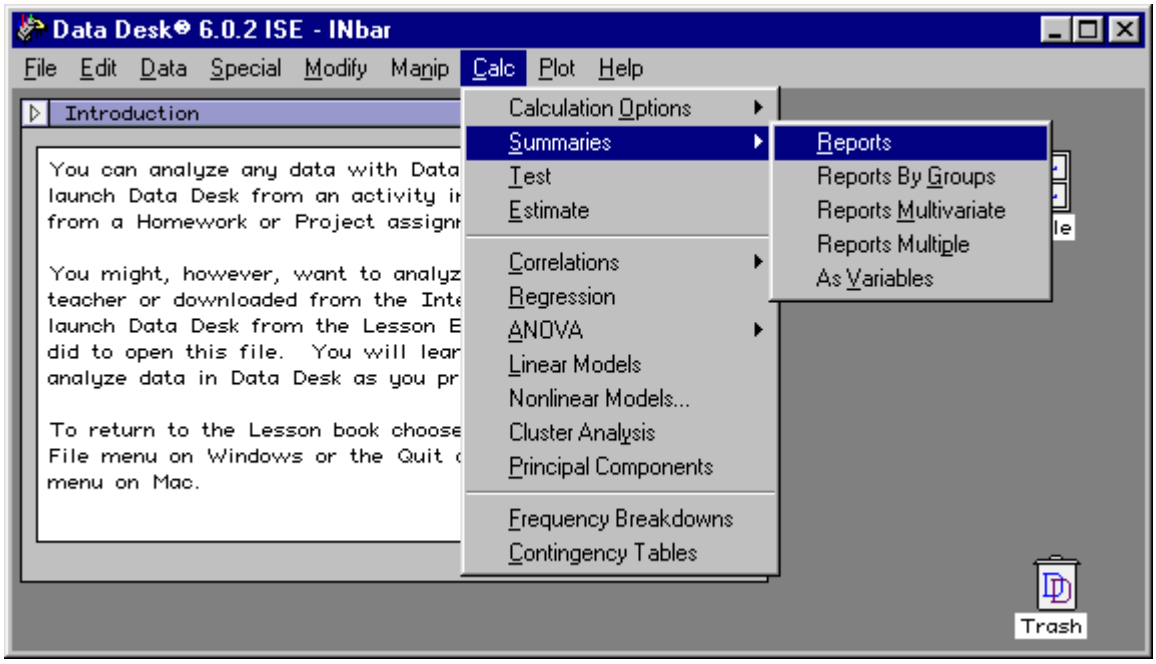
DataDesk also offers you 6 types of spread estimates. You want the SD, but the program shows you two types of SD: Standard Deviation and Population Standard Deviation. So far, in class and the textbook we have only learned how to calculate the Population Standard Deviation (the textbook calls that SD). So that is the one you want here. Make sure to check off that option. The other standard deviation is one you will eventually learn in this class (the textbook calls that SD^+).

2. Check off **Mean, Pop. Stan. Dev.,** and **Counts**

3. Click **OK**

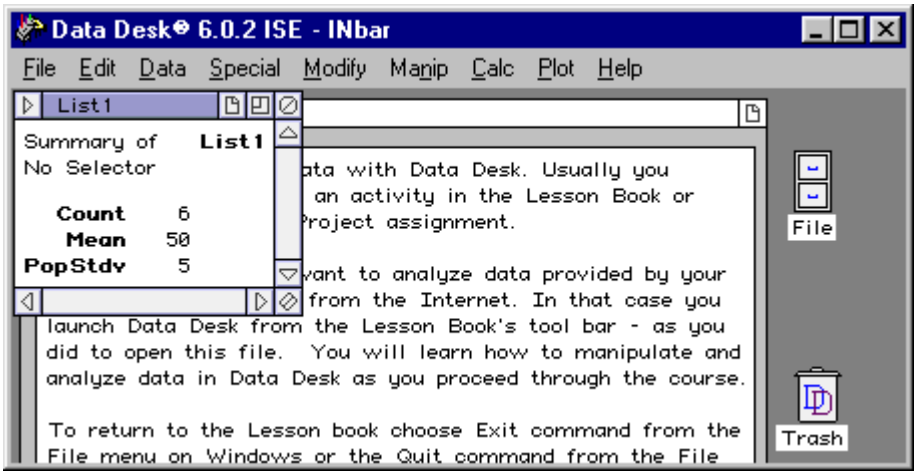
3. Now calculate the values:

Highlight **Calc**
Highlight **Summaries**
Choose **Reports**

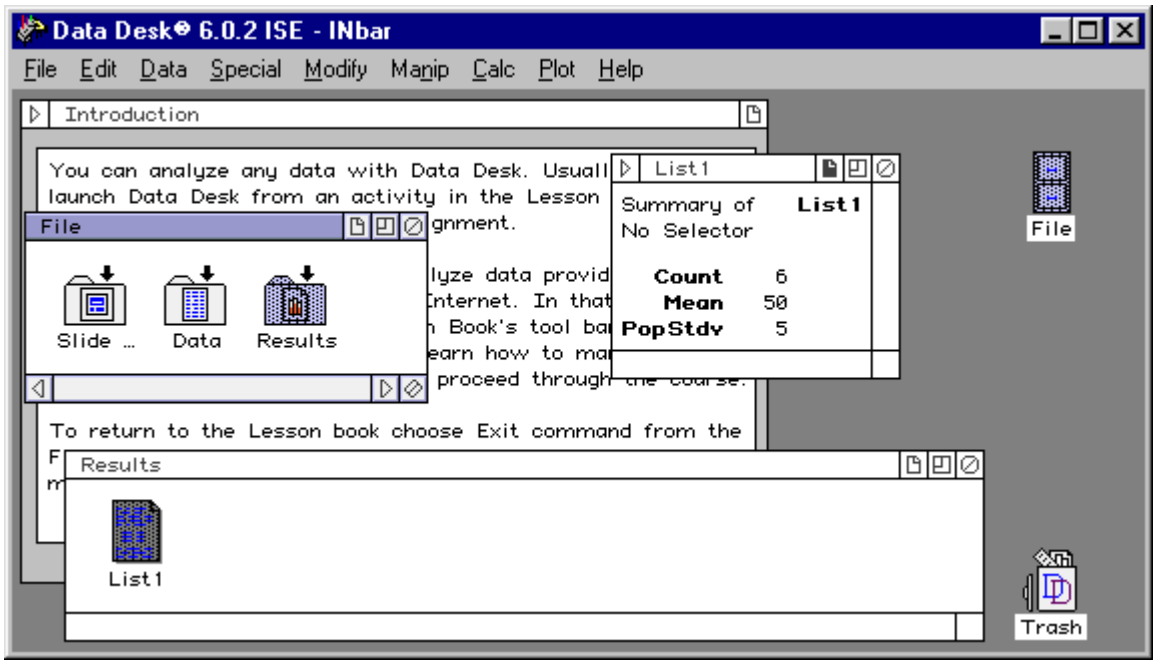


Now the statistics you desire will appear. (If nothing happens on the screen, then the variable, **List1**, has somehow not been selected for analysis. See Lab 2, p. 8 for instructions on how to select a variable for analysis.)

Your window may look somewhat different from this one if you had different options check off than what were selected here.



These results are stored in the **Results** icon in the **File** cabinet. You can click twice on the **File** cabinet. Click twice on the **Results** icon. And you will see an icon, **List1**. If you click on **List1** the same "output" will appear as below:



Now, if you close these three windows and repeat selecting **Calc**, **Summaries**, **Reports**, just like you did the first time, nothing will come up! What goes?

Well, DataDesk has lost track of which variable you want summary statistics on. You have to specify the variable again.

Specify a single variable for analysis

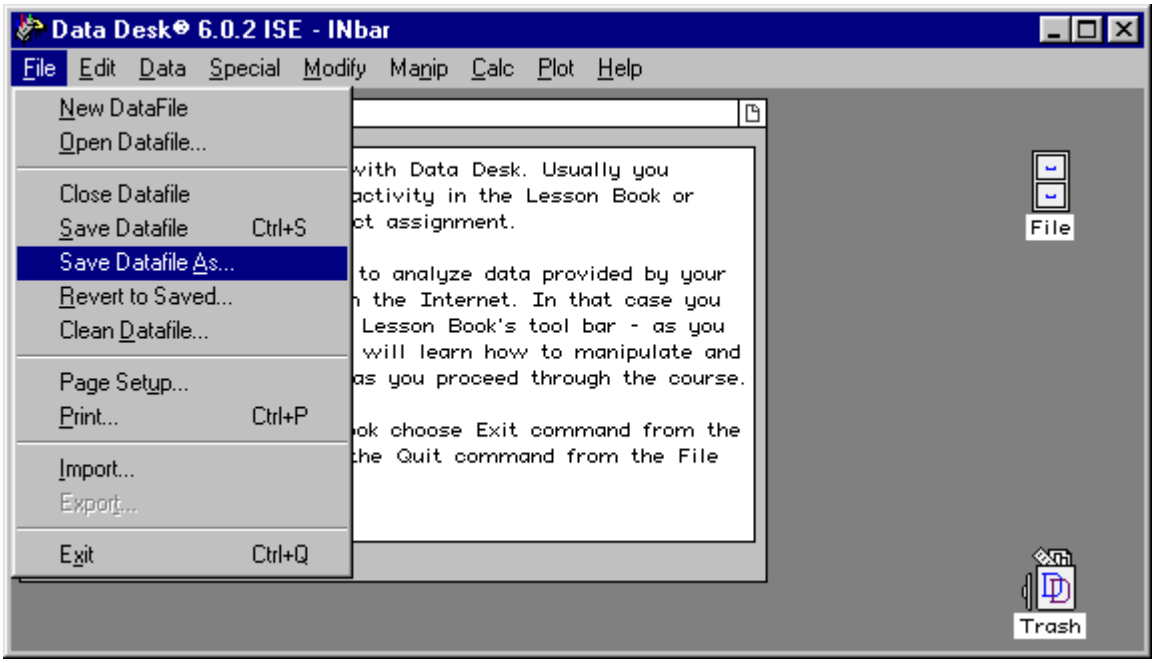
1. Click twice on **File**
2. Click twice on the **Data** icon to open the data window
3. Click once on the **Data** icon in the window so that a **Y** appears.
4. Click again until your variable comes up
5. Click the variable again so that a **Y** appears
6. Now, go back to highlight **Calc, Summaries,** and select **Reports** and up will come the "output" you want

Save your work to disk

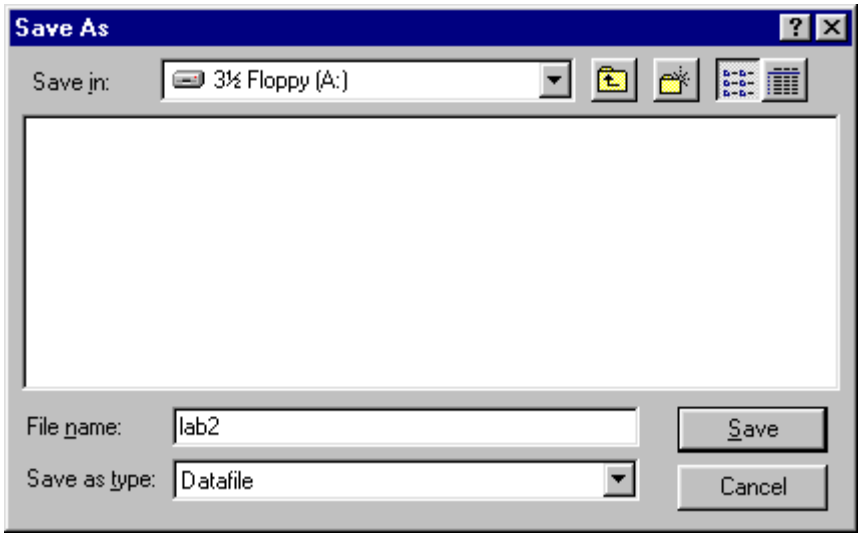
Now it's time to save your work.

1. To save a data file:

   Highlight **File**
   Select **Save File As**



2. Edit **INBar** to a new name for your file (Here the name lab2 is used).
3. Select the A drive as the location to store the file
4. Choose **Save**



Lab 3: Making Histograms

Lab 3's objectives

- Import a data file
- Create a bar graph
- Create a histogram

Import an external data file

Often data comes to us from external sources and we need to import it into statistical software programs. In this task, you will learn how to import raw data from an external source.

Data are often stored in DOS (Disk Operating System) or ASCII (pronounced ASK-EEE) text files. In Windows, these files have a .txt extension to the file name. Text files are generic and can be read easily across computer operating systems which is why data are often stored in this form.

In this instance, you will use a list of test scores given in the Review Exercise 1, Chapter 5 (p. 93) of your textbook. The data have been stored in a file called **ch5dat.txt** on the web and in the Global Shared Files Folder in the Stat lab.

The question in the textbook reads:

"The following list of test scores has an average of 50 and an SD of 10:

39	41	47	58	65	37	37	49	56	59	62	36	48
52	64	29	44	47	49	52	53	54	72	50	50	

(a) Use the normal approximation to estimate the number of scores within 1.25 SDs of the average.
(b) How many scores really were within 1.25 SDs of the average?"

In storing the data, the Prof entered the information into a file in a single column like this:

```
RE1 Data
39
41
47
58
65
37
37
49
56
59
62
36
48
52
64
29
44
47
49
52
53
54
72
50
50
```

Notice the first row is a variable name (RE1 Data). After that, each row contains 1 element. In reality, the column of numbers does not actually exist in the file as a column of numbers (that would waste too much space in a computer)--instead it looks like a string of information with each piece of information or number separated by a tab character sort of like this:

```
RE1 Data[tab]39[tab]41[tab]47[tab]58[tab]65[tab]37[tab]37[tab]49[tab]56
[tab]59[tab]62[tab]36[tab]48[tab]52[tab]64[tab]29[tab]44[tab]47[tab]49
[tab]52[tab]53[tab]54[tab]72[tab]50[tab]50[tab]
```

Computers can very easily decode this information into a single column that humans like us can read. A tab character is not always used. Sometimes a comma is used. Sometimes a blank space. Sometimes another special character is selected.

The actual character used to separate chunks of information is called a **delimiter**.

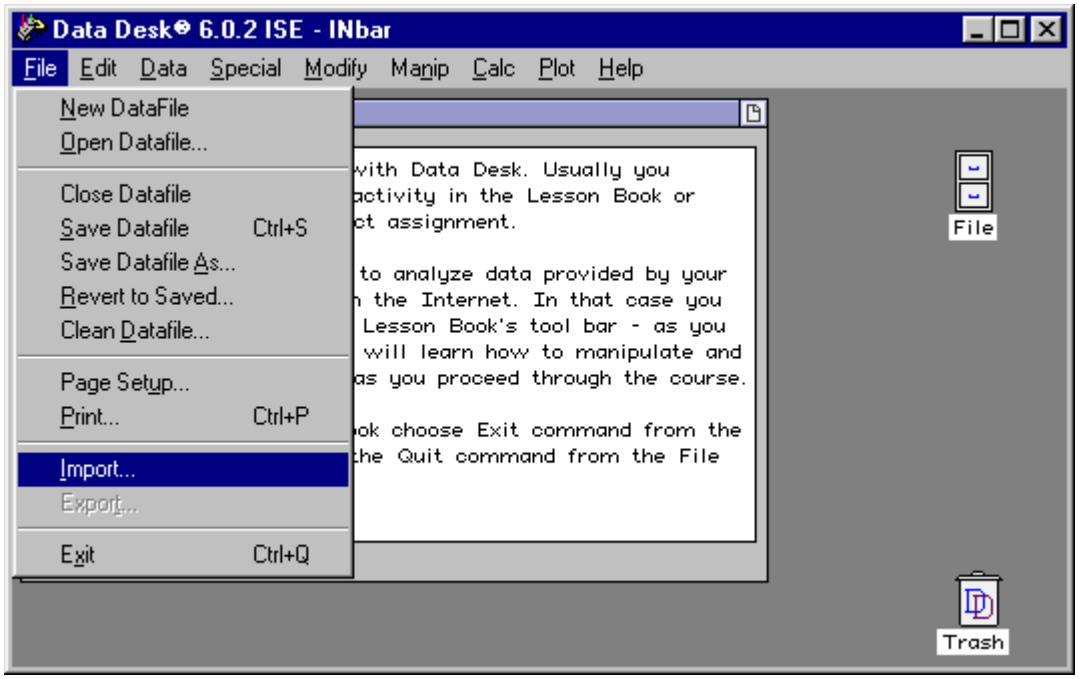
Now you are ready to process the external data.

1. Look in the datasets section of the class web site and download (save to disk) a file called **ch5dat.txt**

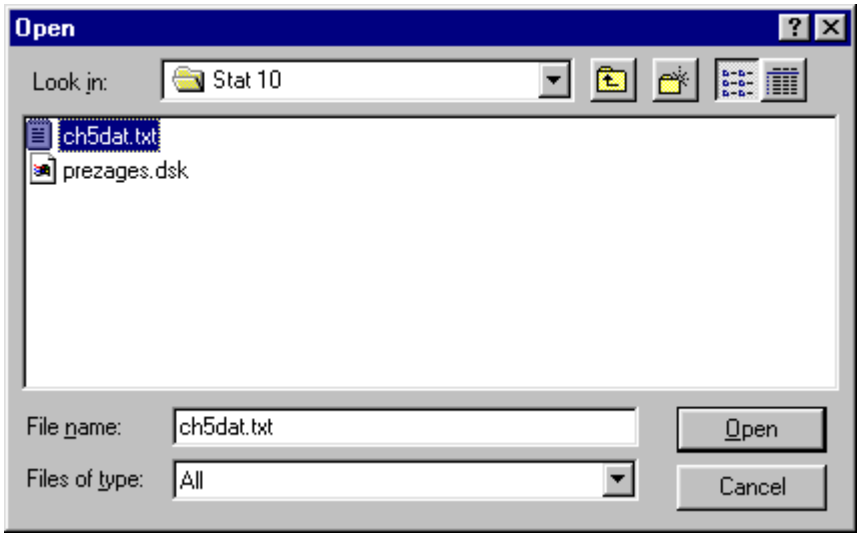
   (When you click on the file on the web, it may open to show you the numbers or it may not; it depends on how you have set up your browser. Be sure to save it as a text file to your computer.)

2. Open your student record file (see Lab 1, p. 6)
3. Launch (or start up) DataDesk (see Lab 1, p. 8)

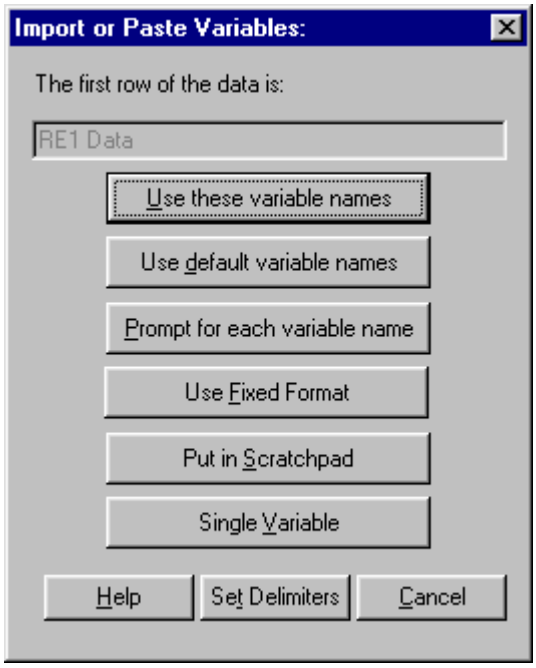
4. Begin the import procedure:
   Highlight **File**
   Select **Import**



5. Open the file (Here it was stored in a Stat 10 folder. On *your* computer it will be where you saved it.)



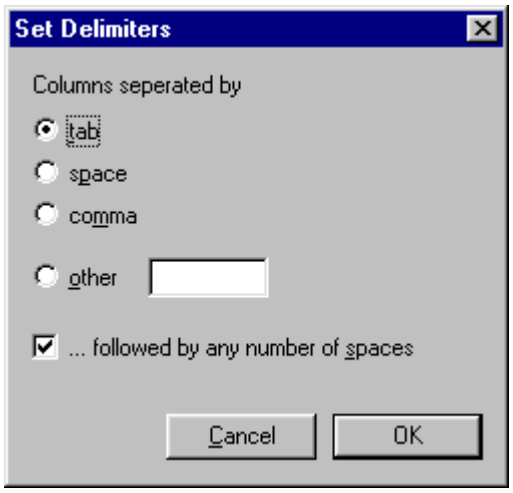
6. Open **ch5dat.txt**
   The following window will appear:



DataDesk now needs some help from you to correctly interpret the file being imported.

7. First make sure the correct delimiter is being used by clicking on **Set Delimiters**

   The following window will appear:

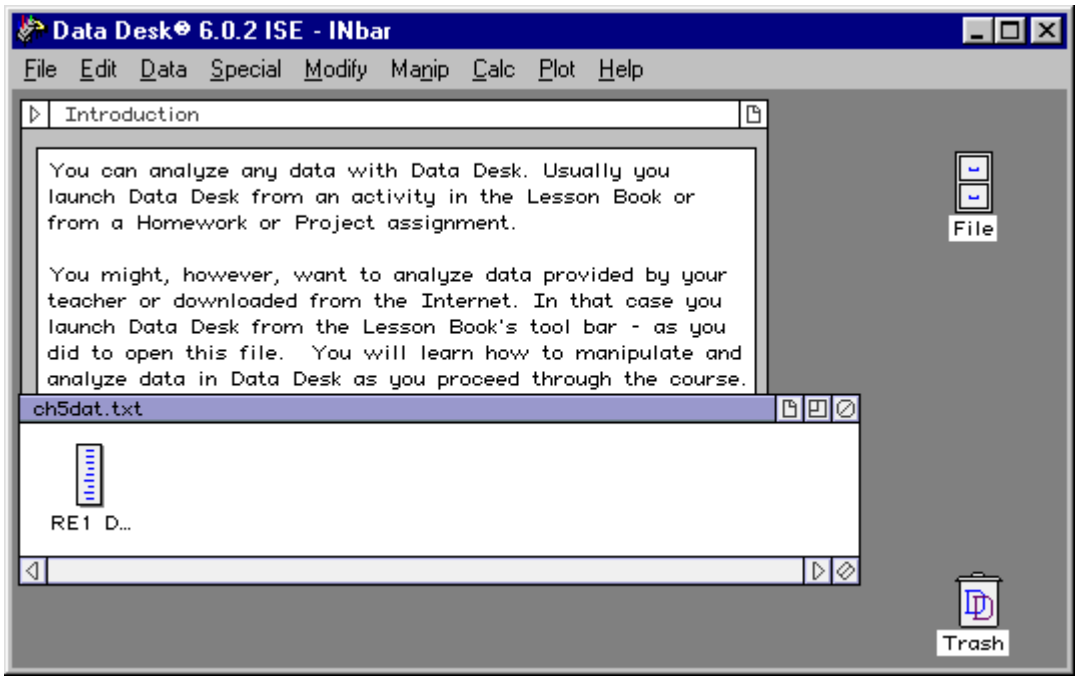


8. Be sure that **tab** is selected and click **OK**

Now the program needs to know if the first row of information is data or variable names. In this instance it is a variable name--but that will not always be true when you import files.

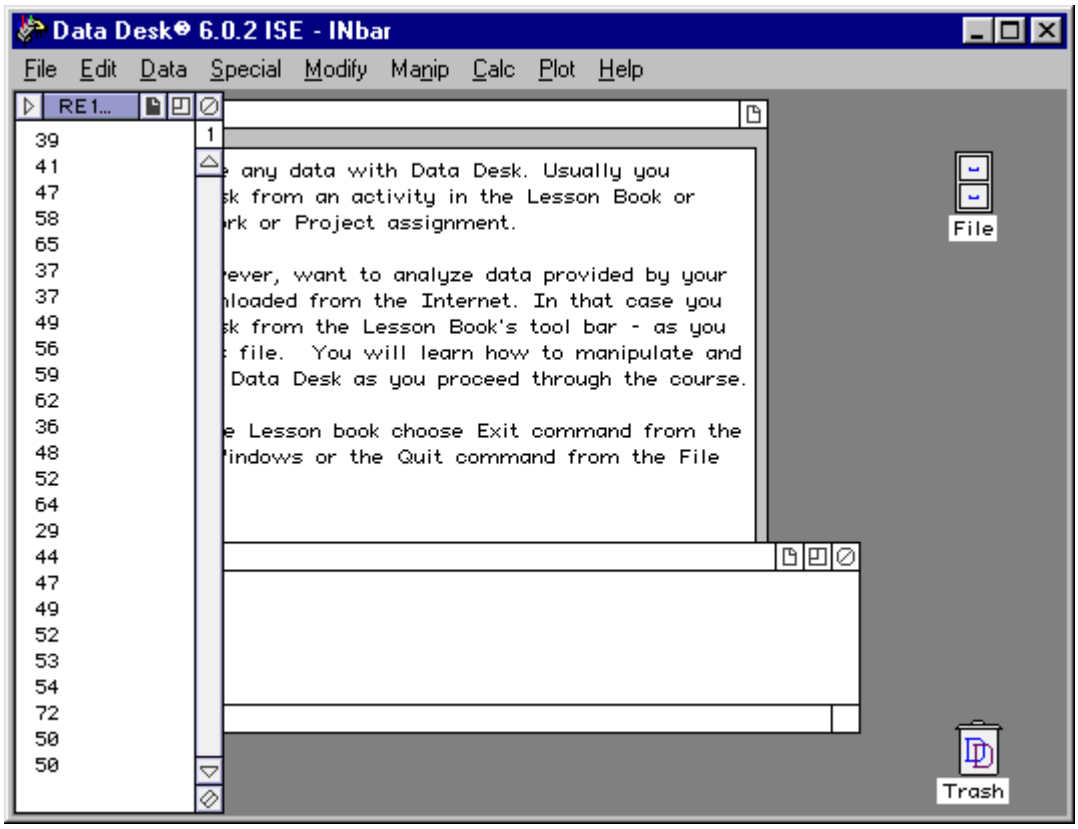
9. Select **Use these variable names**

Now the following window will appear:



10. Click on **RE1 D...** and take a look at your data to be certain it was imported correctly

    It should look like this:

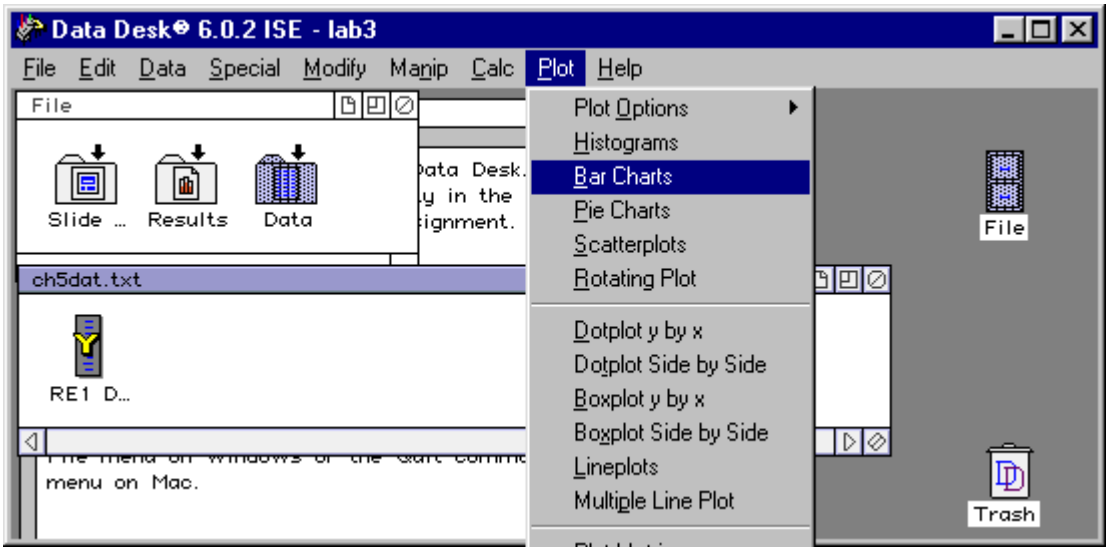


11. Save the data to a file called **lab3** on your diskette (see Lab 1, p. 12)

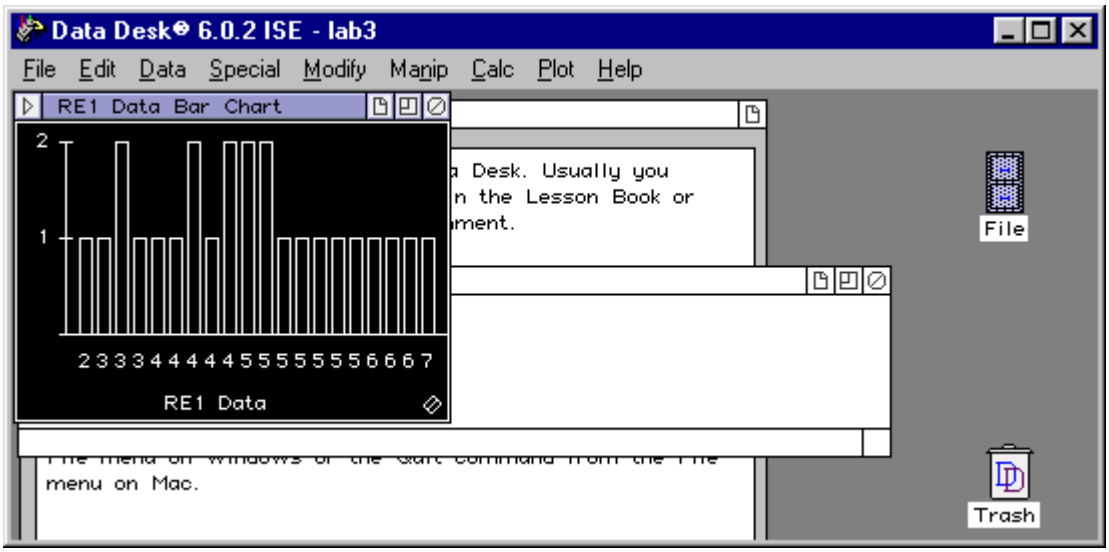
Create a bar chart

DataDesk offers many types of data plotting. The simplest is a bar chart, which is just a count along the left axis and the actual values observed along the horizontal axis.

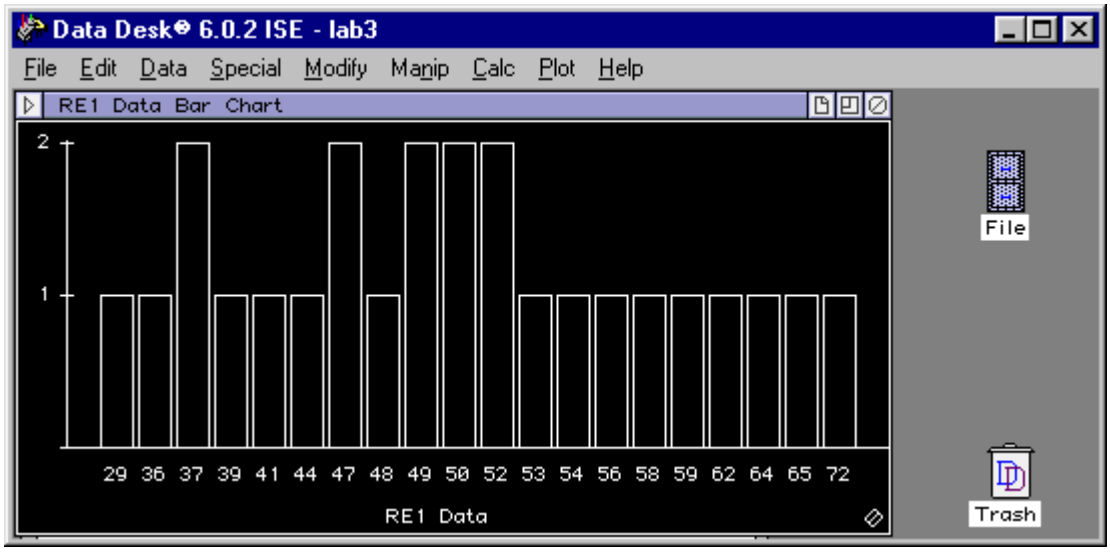
1. Highlight **Plot** and select **Bar Charts**



The following will appear--it is not very appetizing:



You can make the chart prettier by clicking the middle of the three icons located on the top right of the **RE1 Data Bar Chart** line. Now it looks like this:

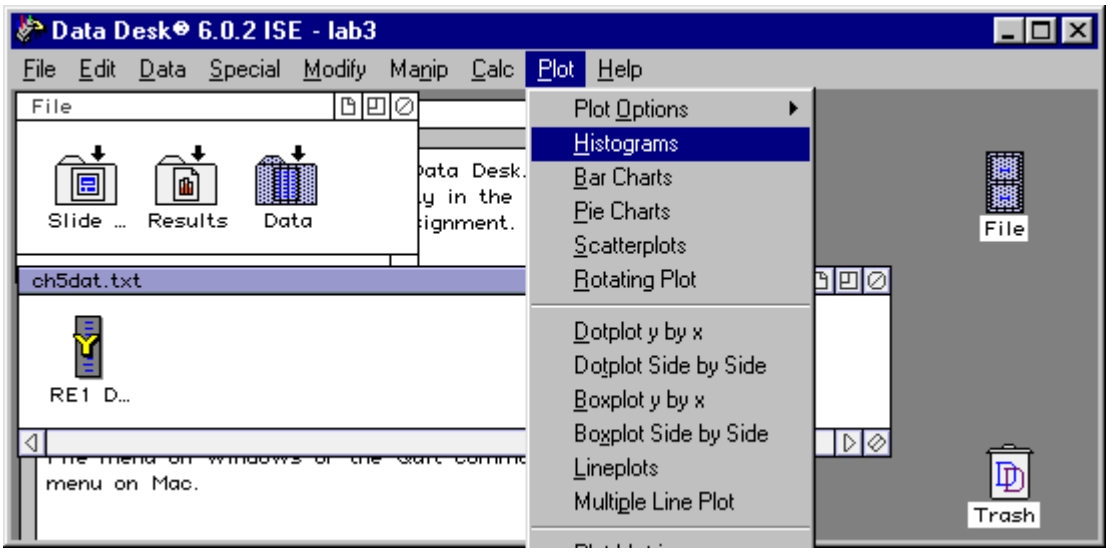


Notice the vertical axis is the count. The horizontal axis shows the values observed. Many values are missing because no one got that test score. This is **not** a histogram (why is that?).

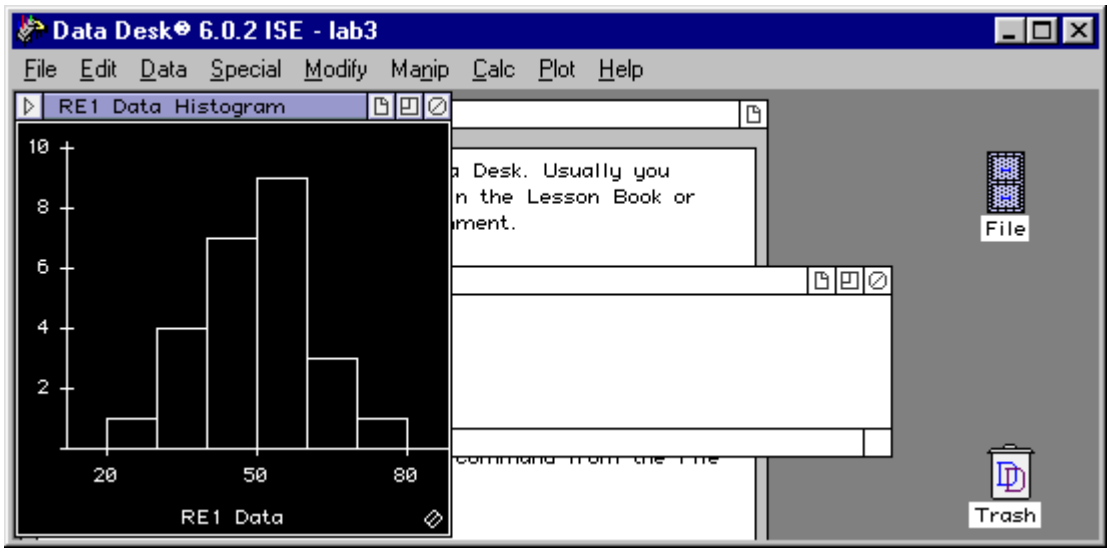
Create a histogram

You can also create a histogram.

1. Highlight **Plot** and select **Histograms**



The following will appear (you can hit the middle icon in the upper right corner to expand the graph and make it look better):

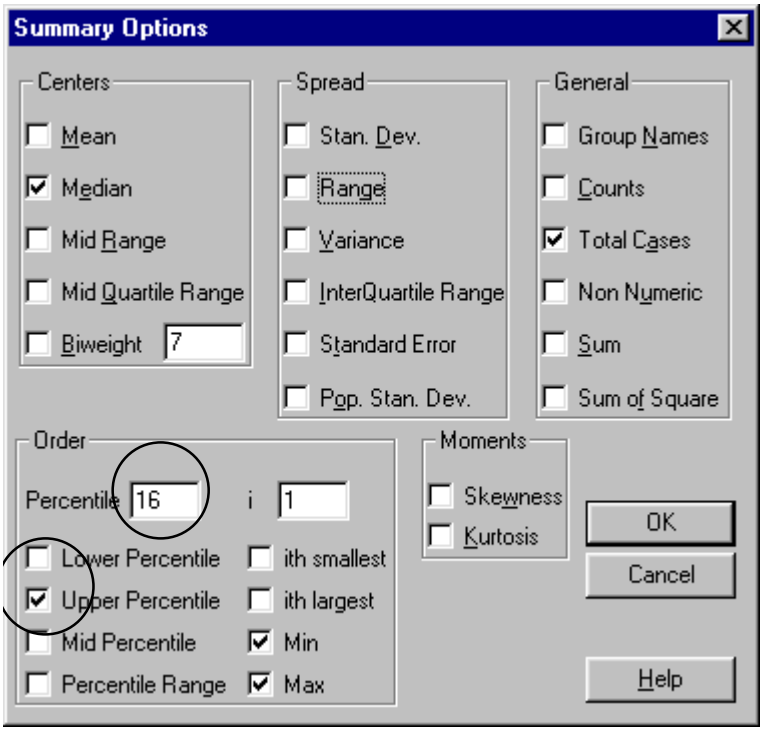


Notice that the area in the rectangles sums to (10*1 + 10*4 + 10*7 + 10*9 + 10*3 + 10*1 =) 250. So, you would estimate that 40/250 or 16% of the elements are test scores of 60 or more (if you use the left-hand convention for discrete numbers from your textbook; 60 would represent the limit. But DataDesk uses continuous numbers so within DataDesk 60.5 will be the limit as it is halfway between 60 and 61).

You can see if this is so.

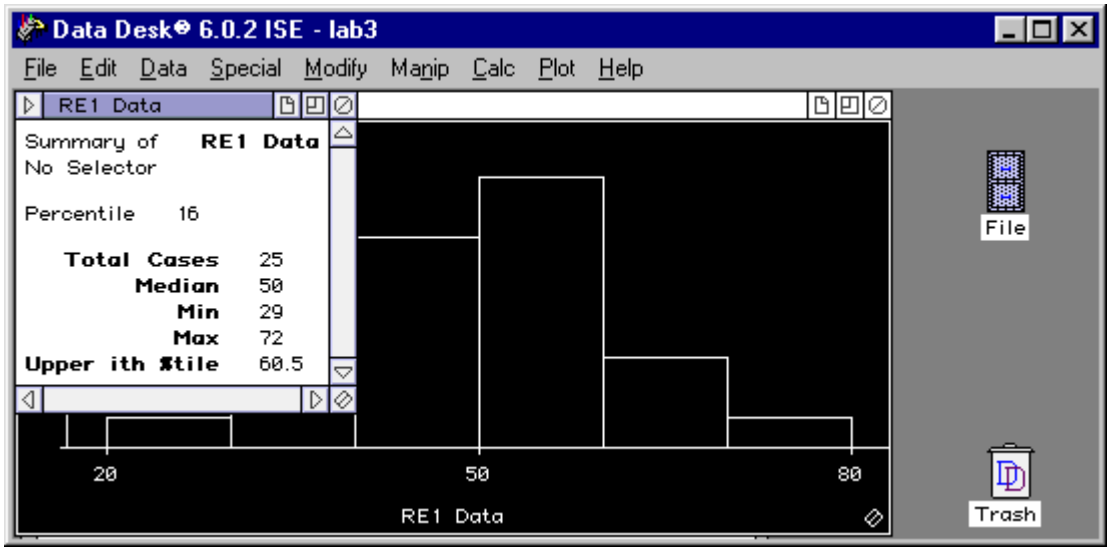
2. Use the **Calc** function (See Lab 2, p. 17)

3. Highlight **Calculation Options** and select the **Summary statistics** that will show what corresponds to the top 16% (see Lab 2, p. 18). (The request is made in the Order category. The value will be associated with the upper 16th percentile.)



4. Ask for the report of summary statistics (see Lab 2, p. 19)

The results look pretty good! Notice that scores above 60.5 are in the top 16% of the distribution, just like predicted.



Lab 4: Creating boxes

Lab 4's objectives

- Create a box
- Create a box using a relational database approach
- Generate statistics about the box
- Generate a random draw with replacement from the box
- Generate statistics describing the random draw with replacement

Create a box

Now you are going to use DataDesk to create a box like you have seen in your textbook.

The box you are going to make comes from Chapter 16, Review Exercise 7:

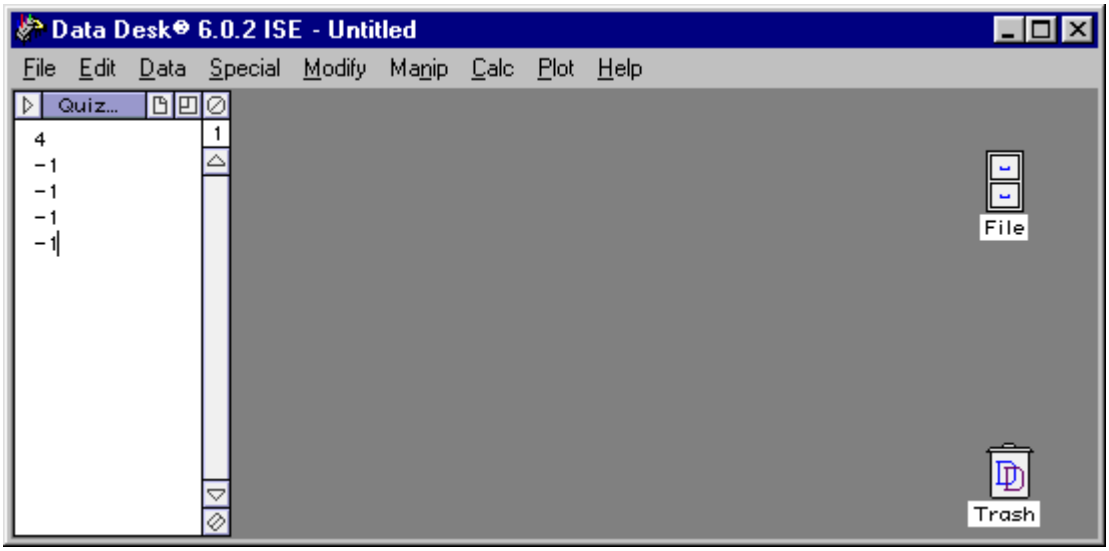
"A quiz has 25 multiple choice questions. Each question has 5 possible answers, one of which is correct. A correct answer is worth 4 points, but a point is taken off for each incorrect answer."

Let's make the box:

1. Open your student record file (see Lab 1, p. 6)

2. Launch (or start up) DataDesk (see Lab 1, p. 8)

3. Create a blank variable called **Quiz box** (or whatever name you want) with information about the box (see Lab 2, p. 13)



Notice, you entered all 5 outcomes which are {4, -1, -1, -1, -1}.

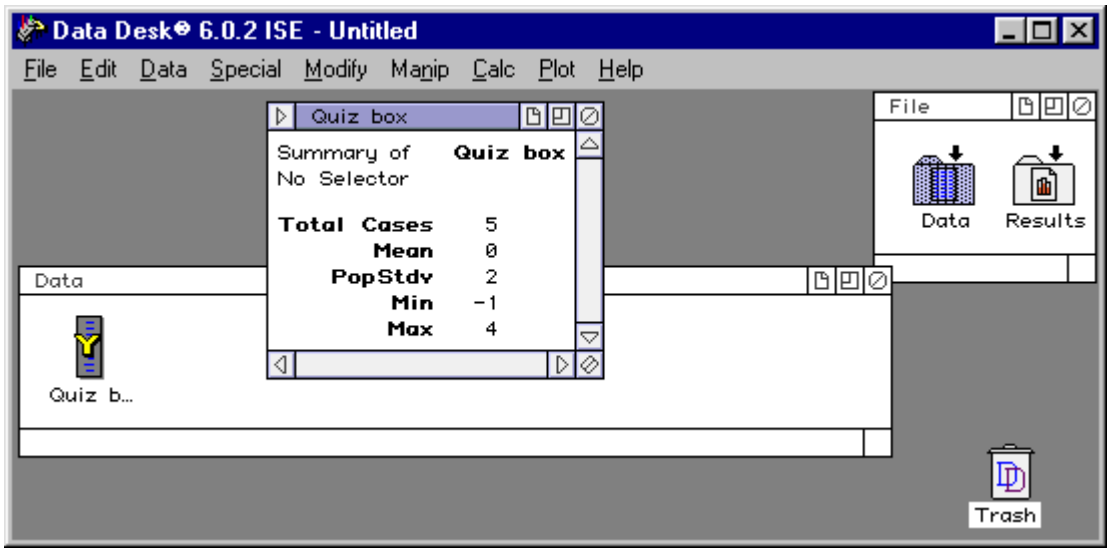
Generate statistics about the box

Now it is very simple to generate statistics about the box. What is the mean of the box and the Standard Deviation?

1. Open your data icon and click on **Quiz box** so that the **Y** appears (see Lab 2, p. 20)

2. Highlight **Calc**, and **Calculation options**, and select the **summary statistics** you want (in this instance, mean and population SD) (see Lab 2, p. 17)

The following window appears:



So the box has a mean of 0 and a SD of 2.

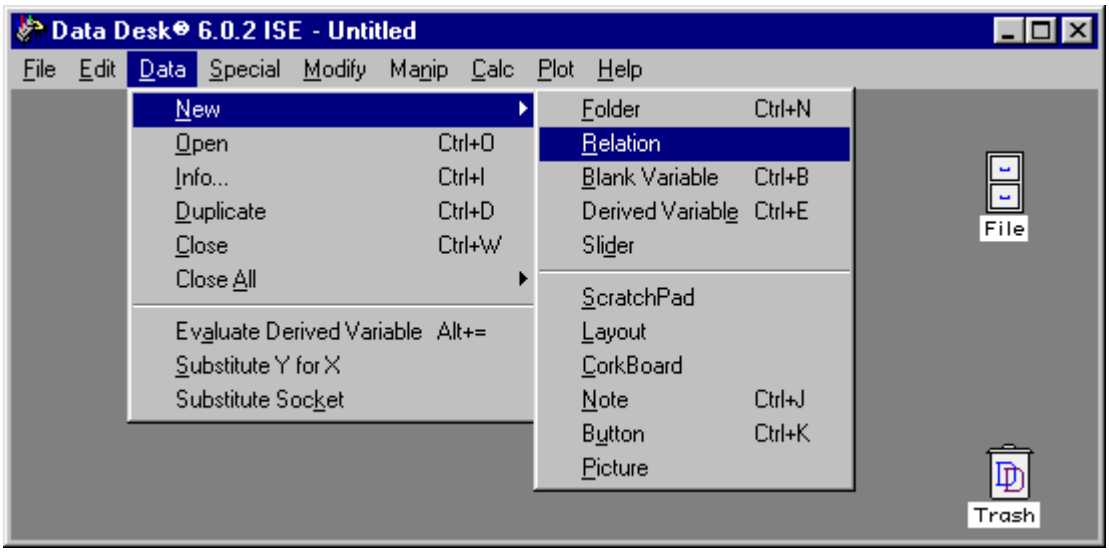
4. Save the variable to a file for use later (see Lab 2, p. 21)

Create a box using a relational database approach

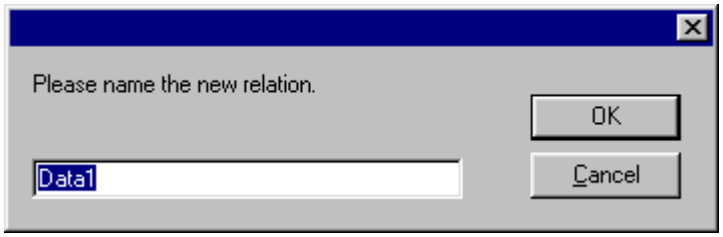
The box you just created can be thought of as resulting from two components. The first are the values (get 4 points, lose 1 point) and the second are the counts or numbers of times the values appear (1 chance of getting 4 points, 4 chances of losing 1 point). In data programming, it is often easier to keep track of things by separating the two components and then creating a relationship between them. This is called a **relational database**. The database, in this instance, would consist of two variables, each with two elements, or a total of 4 pieces of information.

Here's how to make a relational database:

1. Highlight **Data, New**, and select **Relation**



The follow window appears:



2. Edit **Data1** to a new name, **Quiz Box Relation** and click **OK**

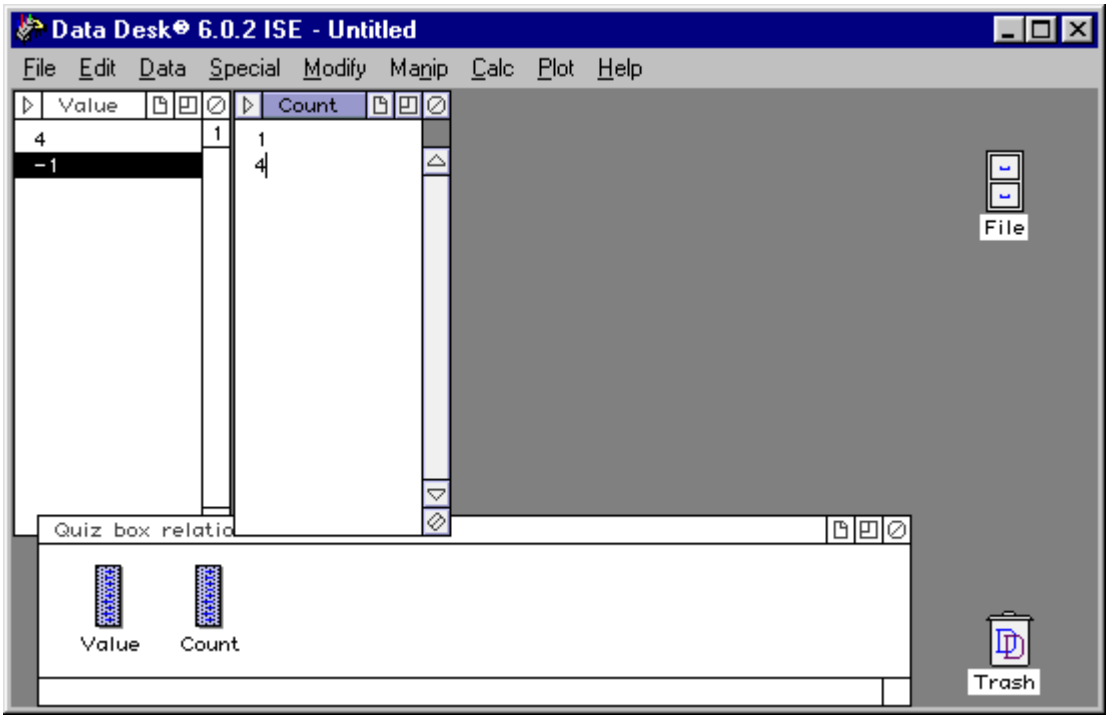
3. The **Quiz Box Relation** will open a window in which you can store the two variables containing information about the box.

4. Create a new blank variable called **value** with the two outcomes possible {4, -1} (see Lab 2, p. 13)

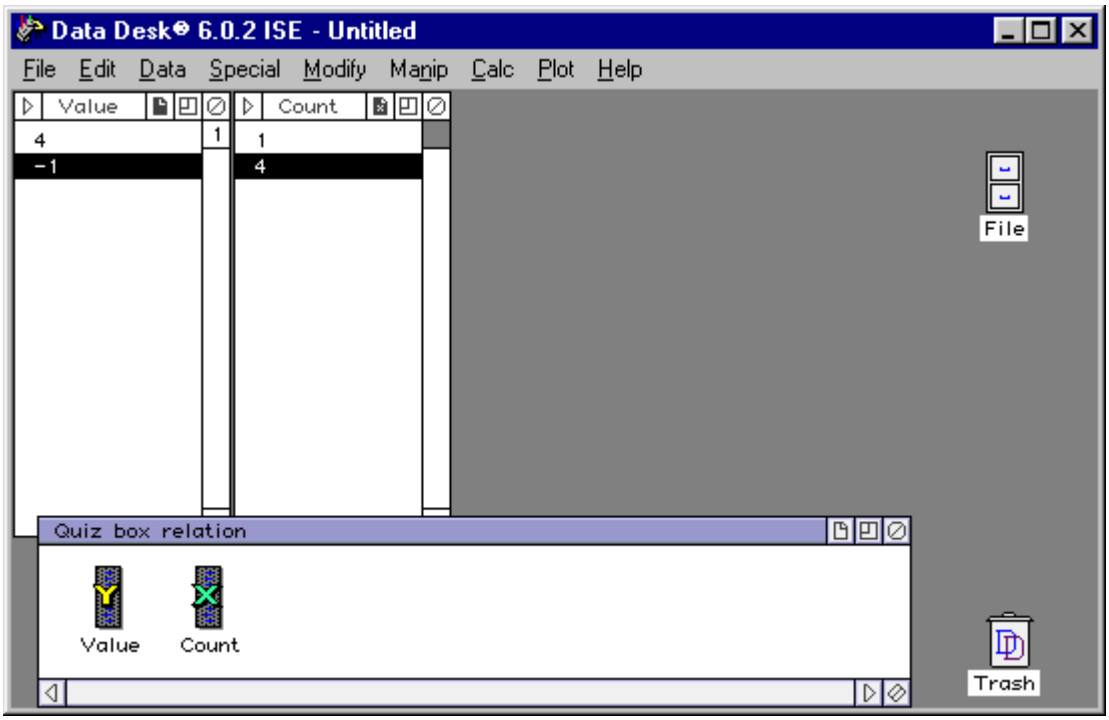
5. Create a second blank variable called **count** with the two outcomes possible {1, 4}

In doing so be very careful of two things:

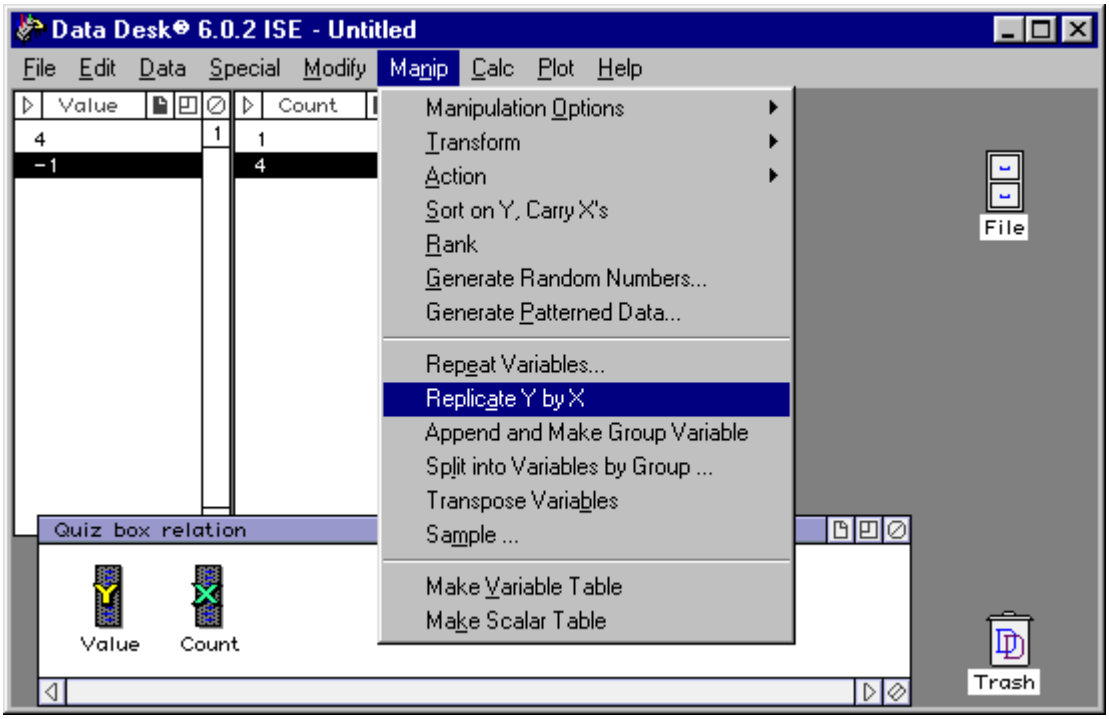
- Keep the order of entry accurate (e.g. the first row for both variables refers to 1 chance of earning 4 points, the second row to 4 chances of losing 1 point across both variables--order of input matters here!)
- Do not enter more elements than exist (for example, don't hit return after you enter the second element, or you will put a blank 3rd element in your database)



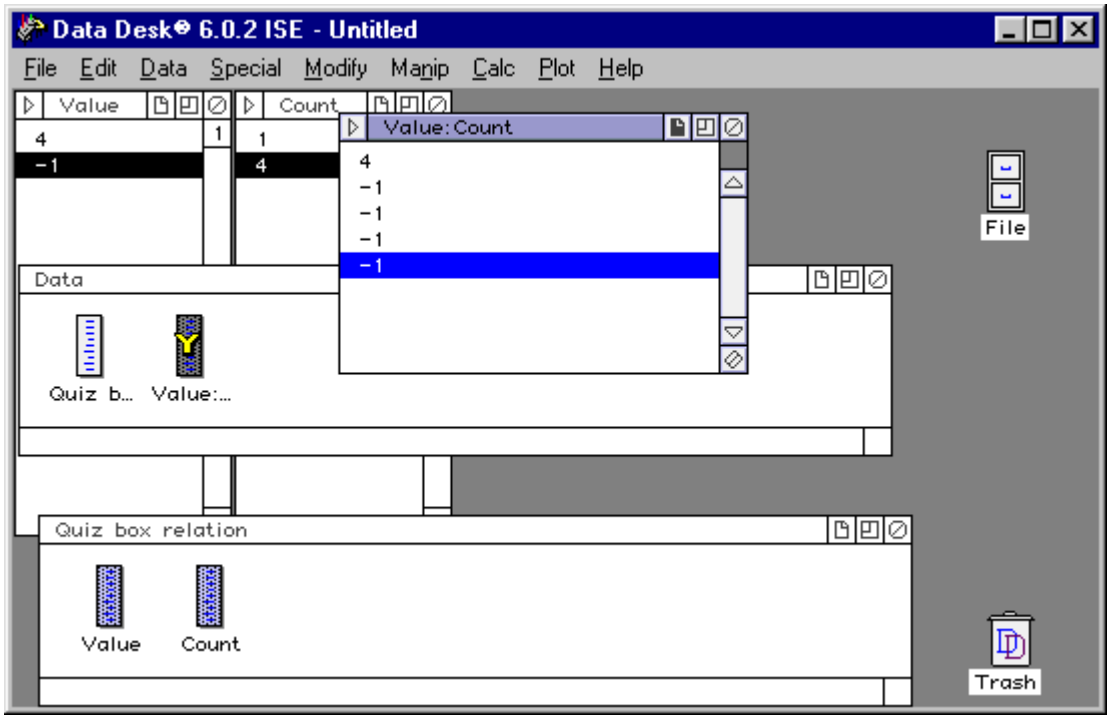
6. Designate the **value** variable as **Y** and the **count** variable as **X** by first clicking on **Value** and then holding the shift key down and clicking on **Count**



7. Now **Replicate** the elements in **Value** by the count in **Count**



The following will appear:



This is a new variable **Value:Count** that looks remarkably like our box!

For our example, it doesn't make much difference which way you create the box (5 pieces of information input vs. 4 pieces of information input plus the relational database information). But what if you had the following problem from the textbook: "A box contains 10,000 tickets: 4,000 with 0 and 6,000 with 1." It is far simpler to enter 4 pieces of information (4,000, 6,000; 0, 1) and the relational database commands than to tediously type in 10,000 elements (see problem 5 at the end of the manual).

Generate a random draw from the box

In your textbook, you learned that a student taking this exam of 25 questions and answering each one randomly would be expected to score 25 * mean of the box = 25 * 0 = 0 on the quiz. But this would vary, and we can estimate how much would be normal variation if only chance is occurring. The expected variation from 0 for a total score on the quiz when answered at random would be number of questions (25 in this instance) multiplied by the SD of the box (2 from your results shown on p.) So, you might guess a score of 0 plus or minus a chance (or SE of the sum) estimate of 10.

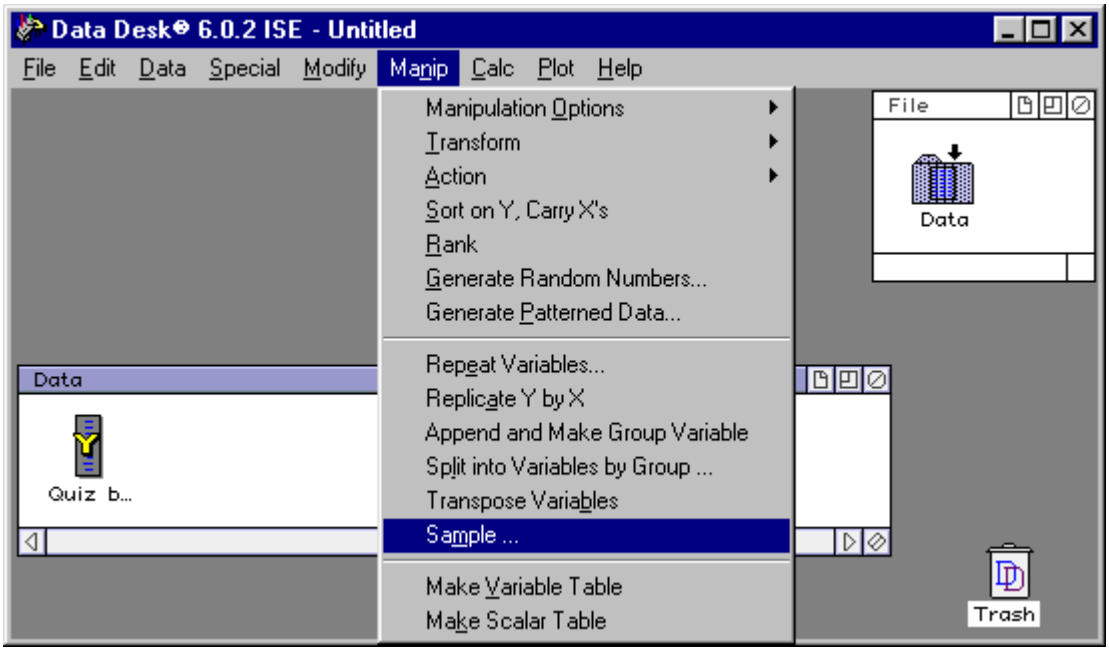
$$\sqrt{25} * SD\ of\ the\ box = 5 * 2 = 10$$

So a little more than two thirds of the time, a student answering at random should score between -10 and +10 on the quiz (versus 4*25 or 100 if all questions were answered correctly and -1*25 or -25 if all questions were answered wrong).

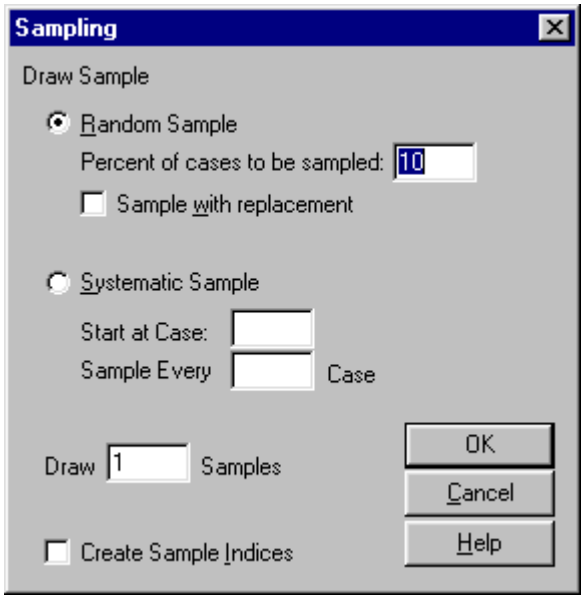
Well, let's see what happens when you actually do randomly sample 25 times. Now the output shown here will vary a little from what you actually observe because each time of random sampling results in slightly different results (Why is that?).

Steps to randomly sampling from an existing variable:

1. Open the data file containing your **quiz box** variable
2. Assign variable **Quiz box** to be **Y** (see Lab 2, p. 20)
3. Highlight **Manip**
4. Choose **Sample**



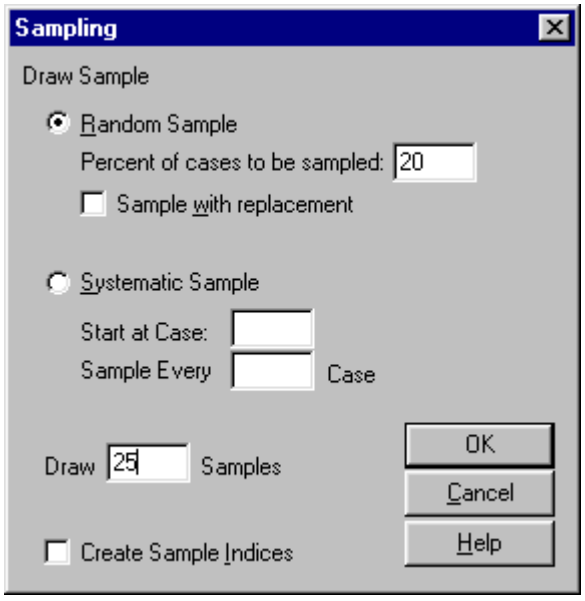
The following screen will appear:



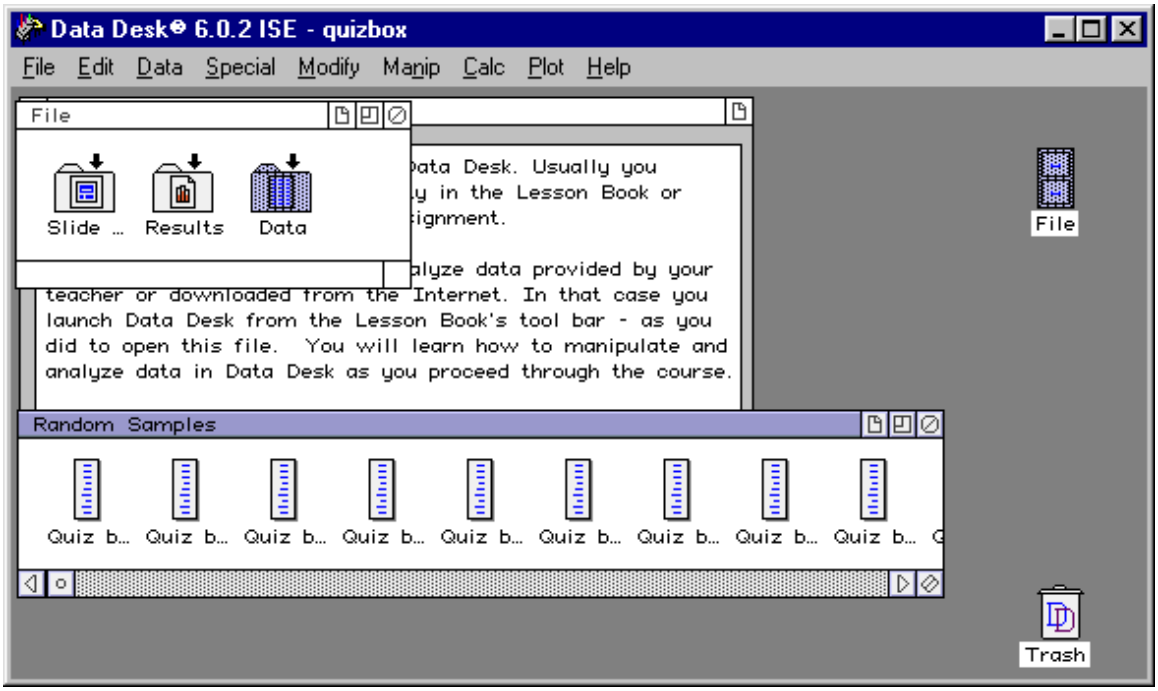
This option in DataDesk let's you decide:

- If you want a random or systematic sample (in this instance you want a random sample)
- How many elements you want to sample each time (you want 1 or 20% of the elements--called cases--to reflect a single answer (1 out of 5 choices) to each question)
- How many times you want to sample from the box (you want 25 times for the 25 questions on the quiz)
- And whether or not to creates indices (this will be skipped for this lab)

5. Edit to request a random sample of 20% of cases done 25 times and click **OK**



DataDesk will generate the following:



Open one of the samples and you'll see that it contains a single answer to a single question.

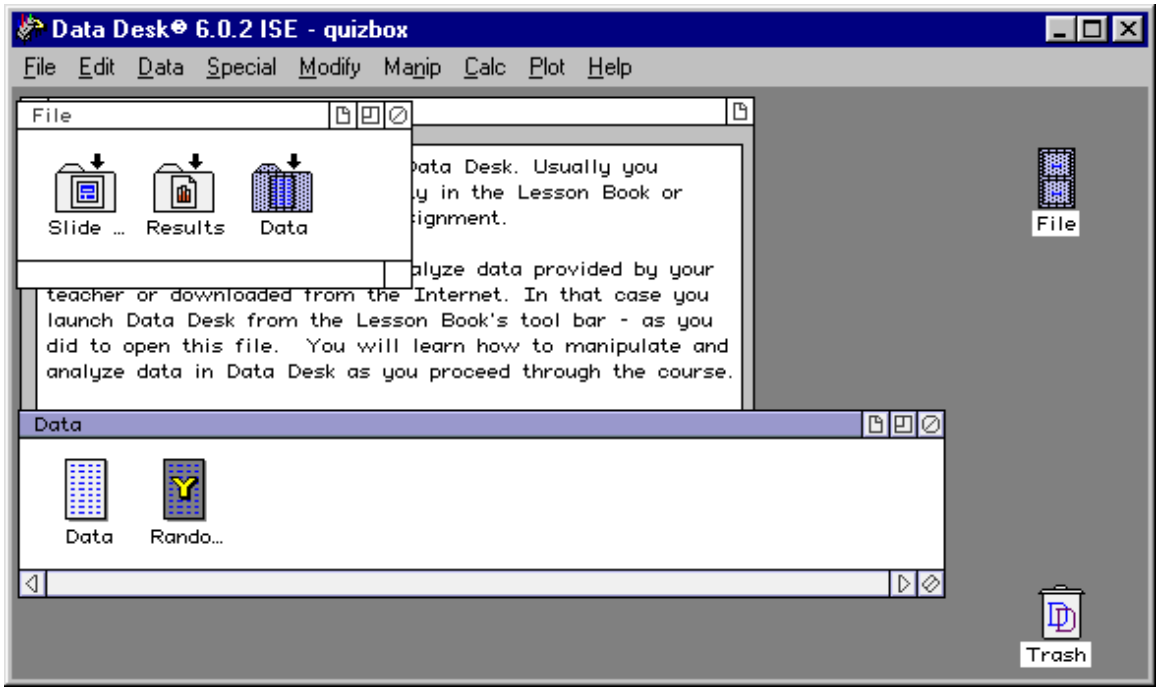
Generate statistics describing the random draw

You now have a random draw of 25 answers. But you need to collapse these 25 samples into one sample so that you can calculate summary statistics. To do this:

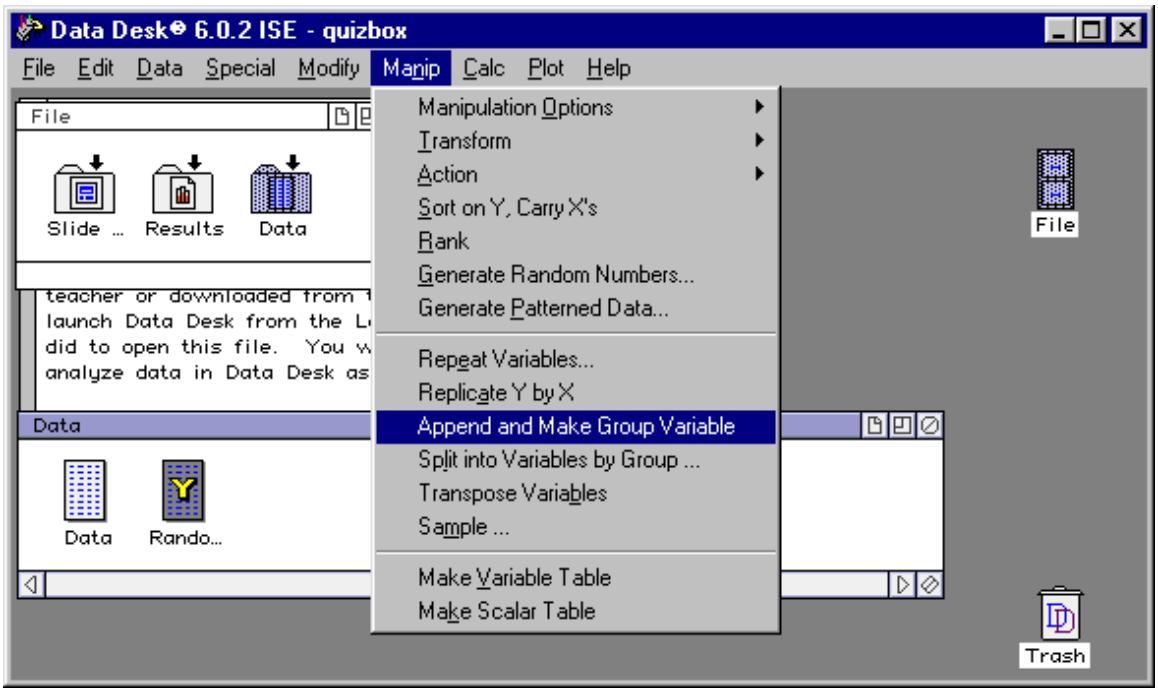
1. Click on the top right most icon of the **Random Samples** window. The data window will appear with your **Quiz box** variable labeled **Y**.

2. Click again on the top right most icon and two icons (**Data** and **Random**) will be in the window.

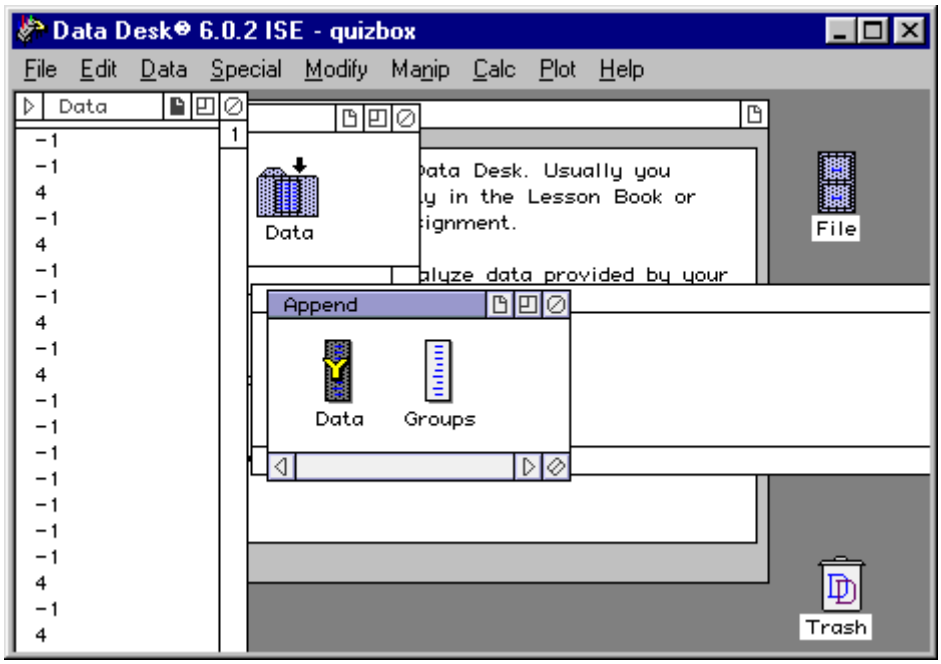
3. Click on **Random** to assign it to be **Y**



4. Now highlight **Manip** and select **Append and Make a Group Variable**

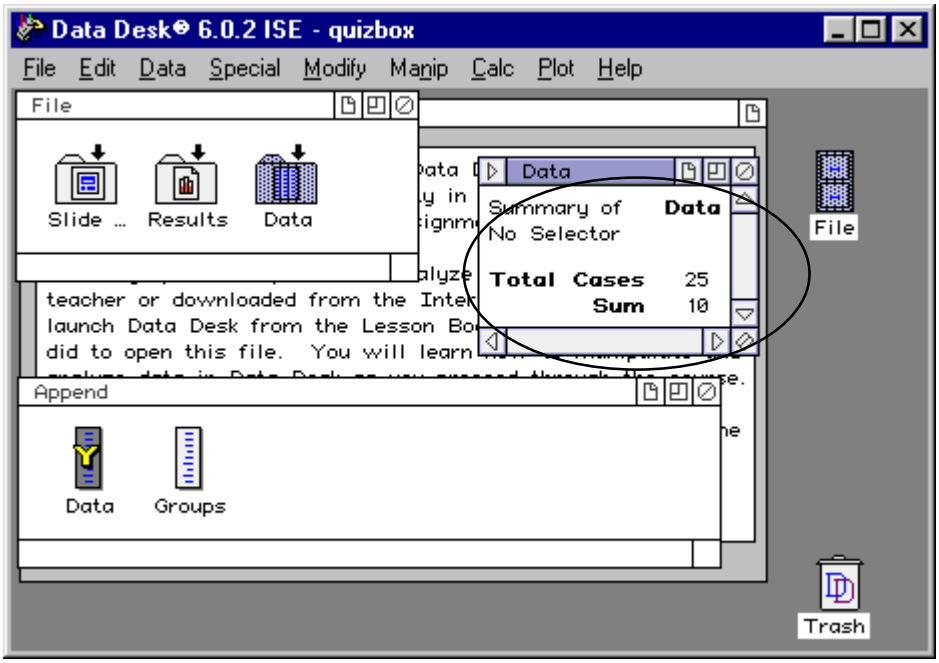


DataDesk now creates two variables. **Data** contains the 25 samples you drew. You can see this by clicking on **Data** to open it.



6. Request the sum of the elements in the box by highlighting **Calc**, making certain the report will include the sum (check **Calculation options**) and requesting from **Summaries** a **Report** (see Lab 2, p. 19)

The Prof observed the following results, well within chance expectations:



Your results will vary.

Lab 5: Correlation and Regression

Lab 5's objectives

- Create summary statistics on two variables
- Make scatterplots of two variables
- Plot regression lines
- Calculate the correlation
- Calculate regression statistics

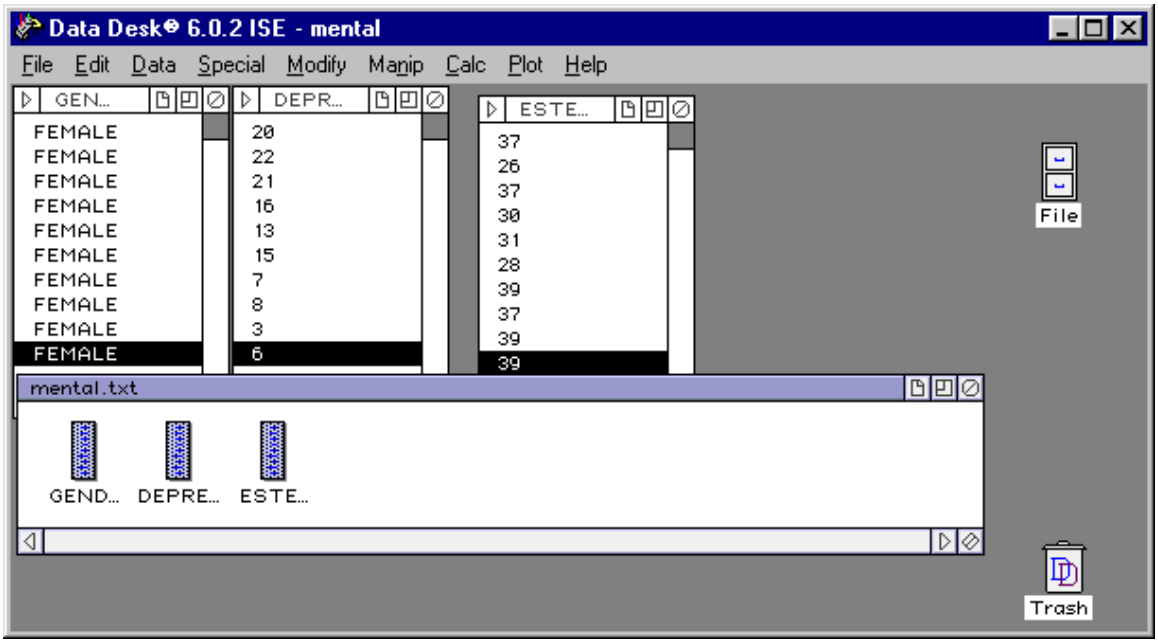
The data you are going to use comes from a real study the Prof did a few years ago. She was interested in the relationship between depression and self-esteem in college students. She expected two things to be true: 1) those with lower self-esteem should report higher levels of depressive distress, 2) women should evidence higher levels of depressive distress and lower levels of self-esteem than men.

Stored for you on the class website is a data set, **mental.dsk**, which includes responses from 200 of the subjects in the study. All were scored for current level of depressive distress (the variable is called **DEPRESSION**) and self-esteem (the variable is called **ESTEEM**). A third variable included is the subject's gender (and is called **GENDER**).

Create summary statistics on the two variables

1. Go to the web site and download **mental.dsk** to your computer (see Lab 1, p. 10)
2. Open your student record file (see Lab 1, p. 6)
3. Launch (or start up) DataDesk (see Lab 1, p. 8)
4. Open the **mental.dsk** dataset (see Lab 1, p. 10)

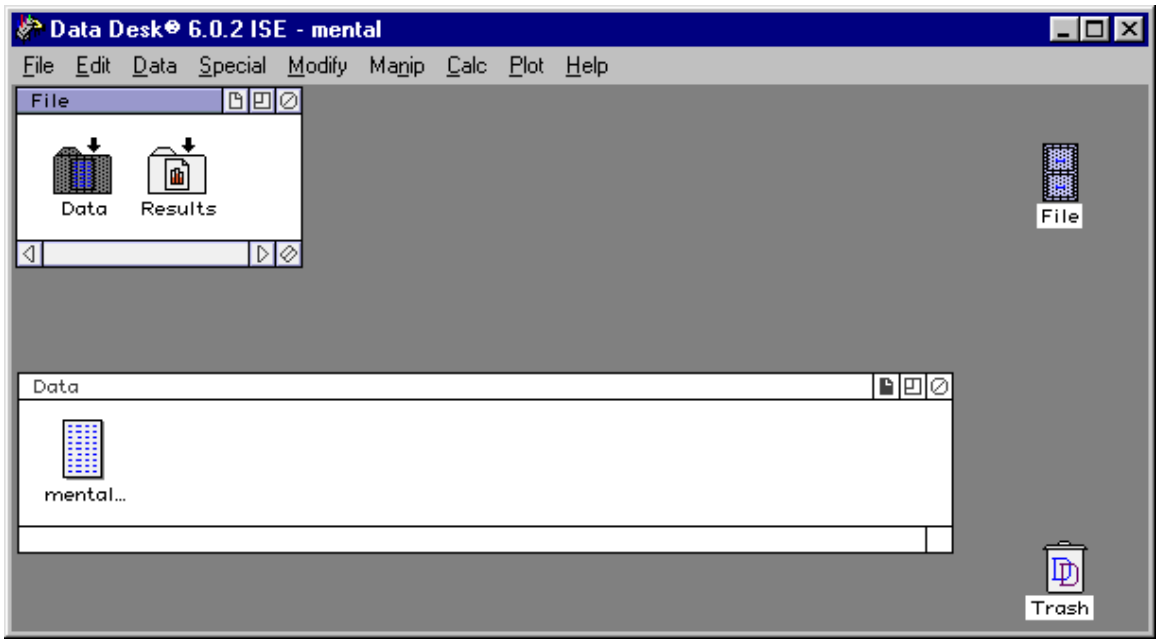
The following will appear:



Notice that the data came originally from a file, **mental.txt**, that has since been read into DataDesk. There are three variables. They are linked across the case or individual (the solid dark line shows that DataDesk reads the data as a **Female** who scored **6** on the Depression scale and **39** on the self-esteem scale).

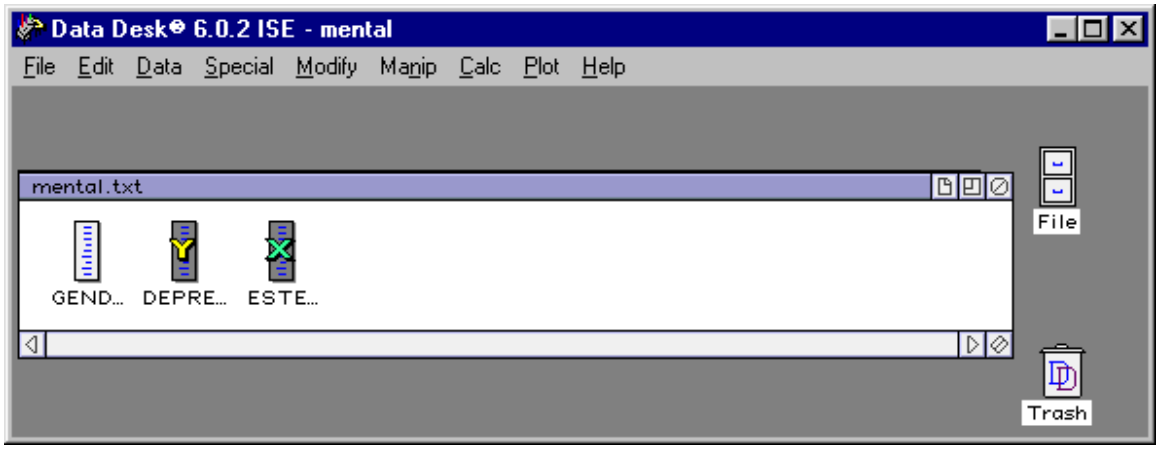
5. Try moving the cursor in one of the variable windows and you will see that it automatically moves in the others

6. Now close the data windows until only the **File** cabinet appears. Once again open the data window and you will see **mental** appear first as below. You can click on that to expose the three variables.

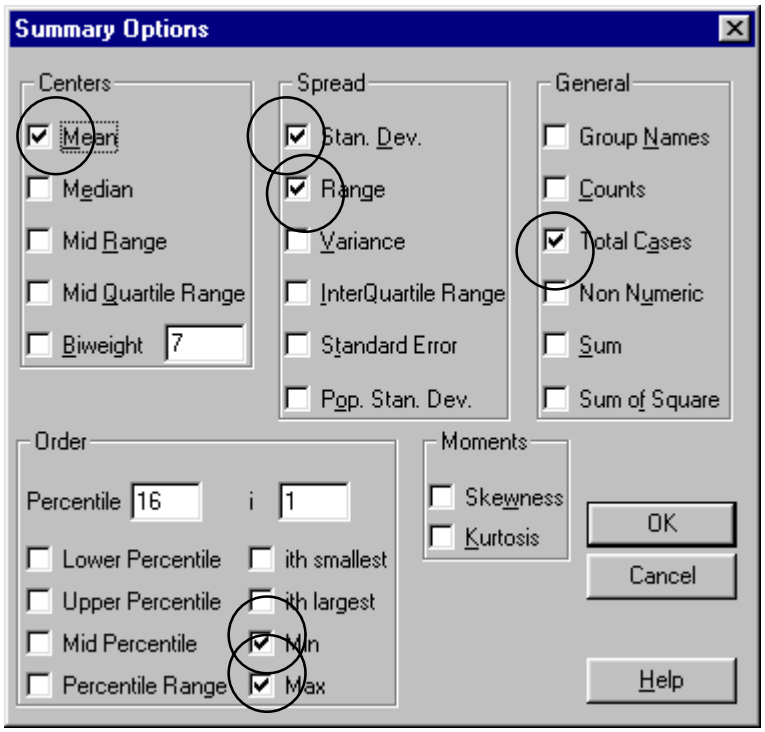


Before you ever analyze data, you should always check to see that it holds no surprises for you. That's because computers happily analyze without thinking whatever you give them and if there is a mistake in your data, it is up to you to find it. So the first step is to look at the data.

7. Assign **DEPRESSION** to be **Y** and **ESTEEM** to be **X** (see Lab 4, p. 38)

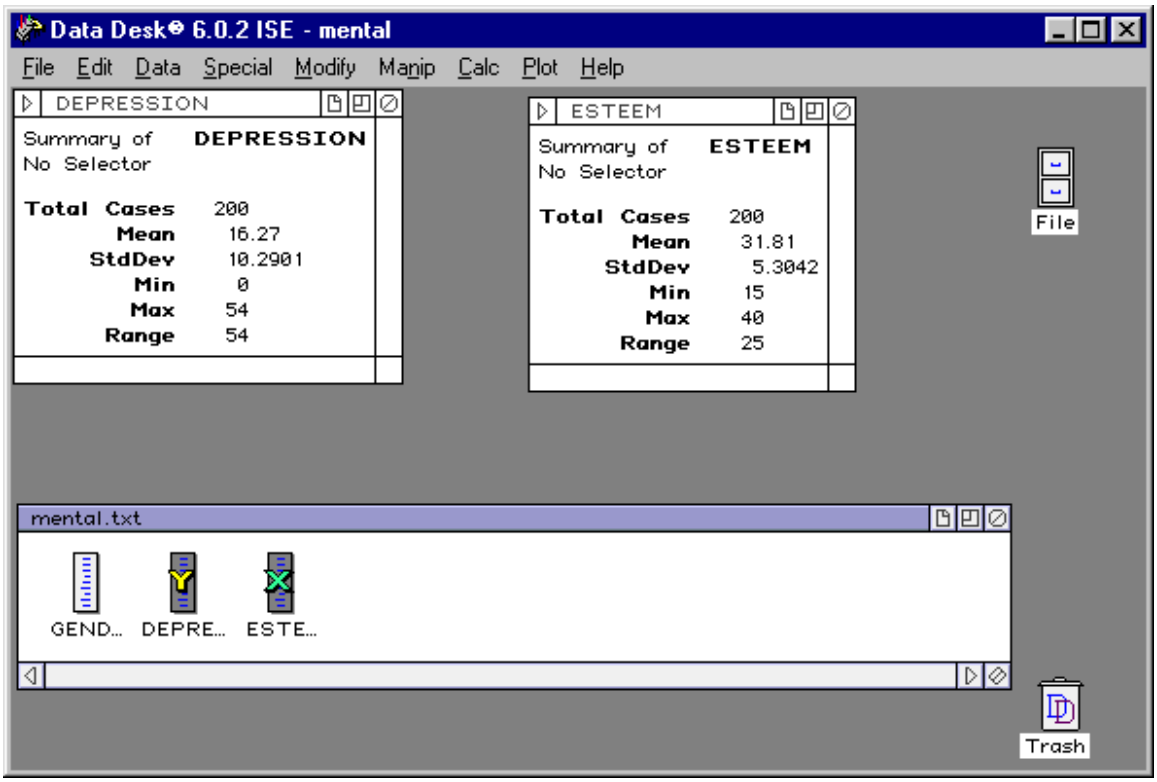


8. Highlight **Calc**, then **Calculation Options**, then **Summary Statistics** and ask for the following: (see Lab 2, p. 17)



9. Then highlight **Calc**, **Summaries**, and select **report** (see Lab 2, p. 19)

Two overlapping windows will appear. You can pull them apart to look like this:

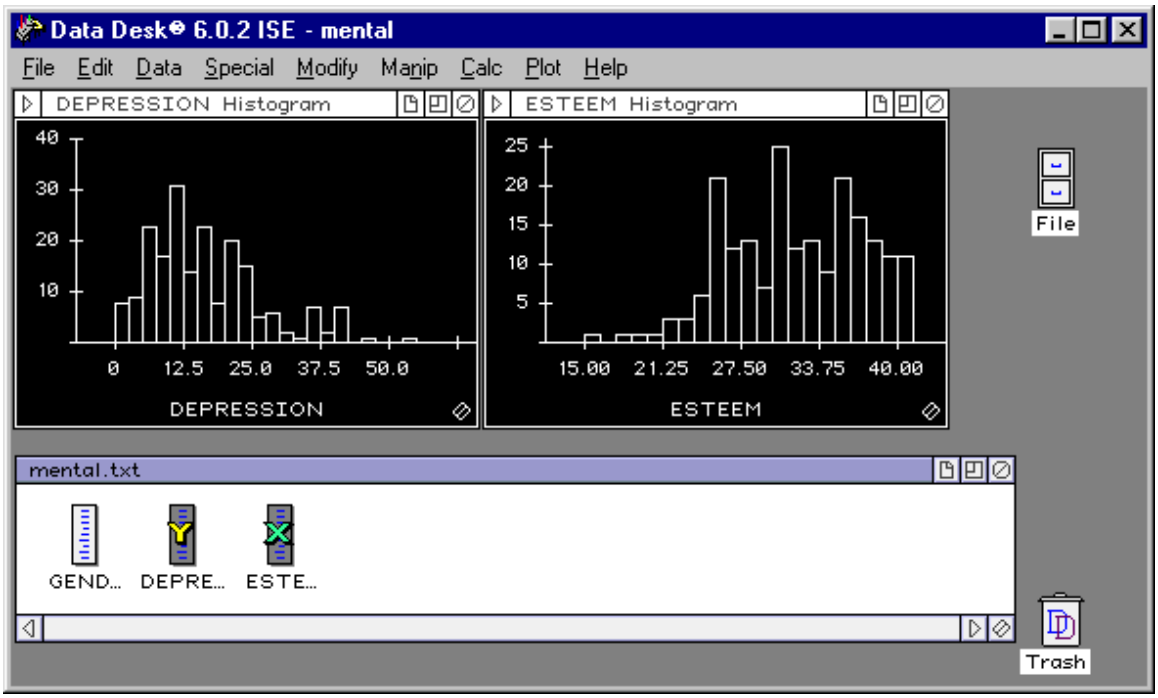


Notice, values on **DEPRESSION** range from a low of 0 to a high of 54. There are 200 cases. The average is about 16, with about 2/3's of people clustering above and below that by about 10 points. The **StdDev** is the SD^+ in your textbook, a topic not covered yet in the course. The **PopStdv** is the SD that your textbook calculates. For the moment, the thing to notice is that they are very similar though not exactly the same.

You can also plot histograms of the individual distributions.

10. Highlight **Plot**, and select **Histogram** (see Lab 3, p. 29)

The following will appear:



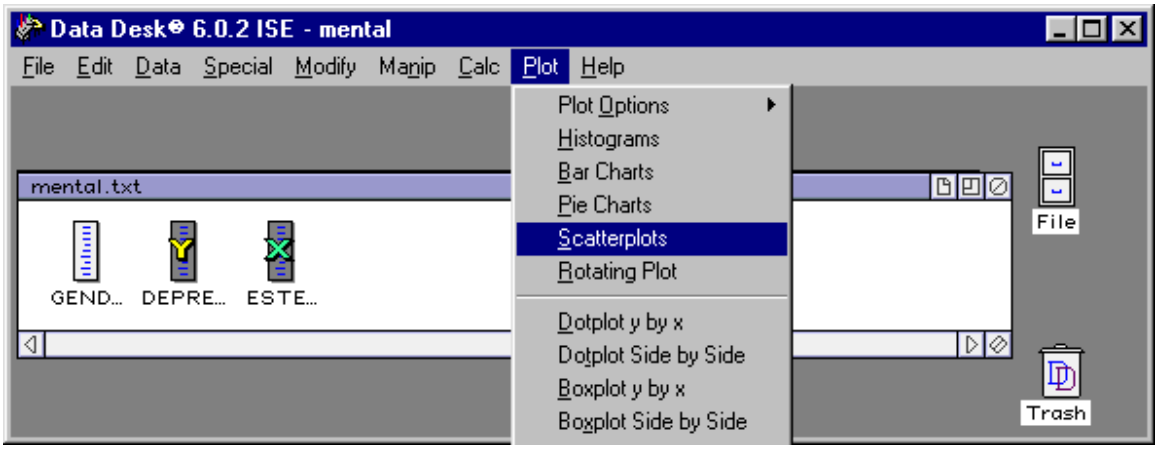
This shows you that **DEPRESSION** is skewed to the right. Most people report low levels of depression, but a few report high levels. This drags the mean to the right of the median. In contrast, **ESTEEM** is skewed to the left. Most people report high levels of self-esteem but a few report very low levels dragging the mean to the left of the median.

Make scatterplots of two variables

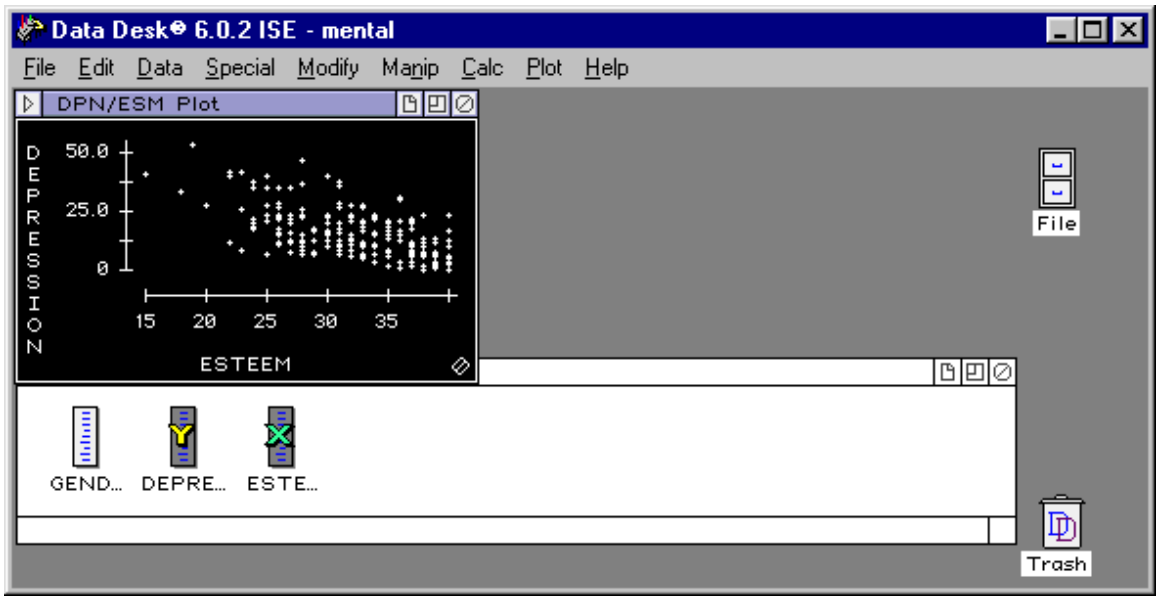
Now, you are going to see if depression is correlated with self-esteem. First, you are going to do it visually.

1. Make sure that **DEPRESSION** is still marked with a **Y** and **ESTEEM** with an **X**

2. Highlight **Plot** and choose **Scatterplots**



The following appears:



Notice on the left is the **DEPRESSION** axis and on the bottom is the **Esteem** axis. It is apparent from this scatterplot that in the sample as self-esteem goes up depression goes down (what would you predict the correlation should be, approximately? Is it positive, negative? Small, large?)

You can do several things with this scatterplot. Under **Plot Options**, you can select to reverse black and white printing. You can expand this plot. And:

3. Put the cursor over the plot. A hand will appear.

4. Hold down the mouse key, and the cursor now becomes a slider. Move it left and right, up and down.

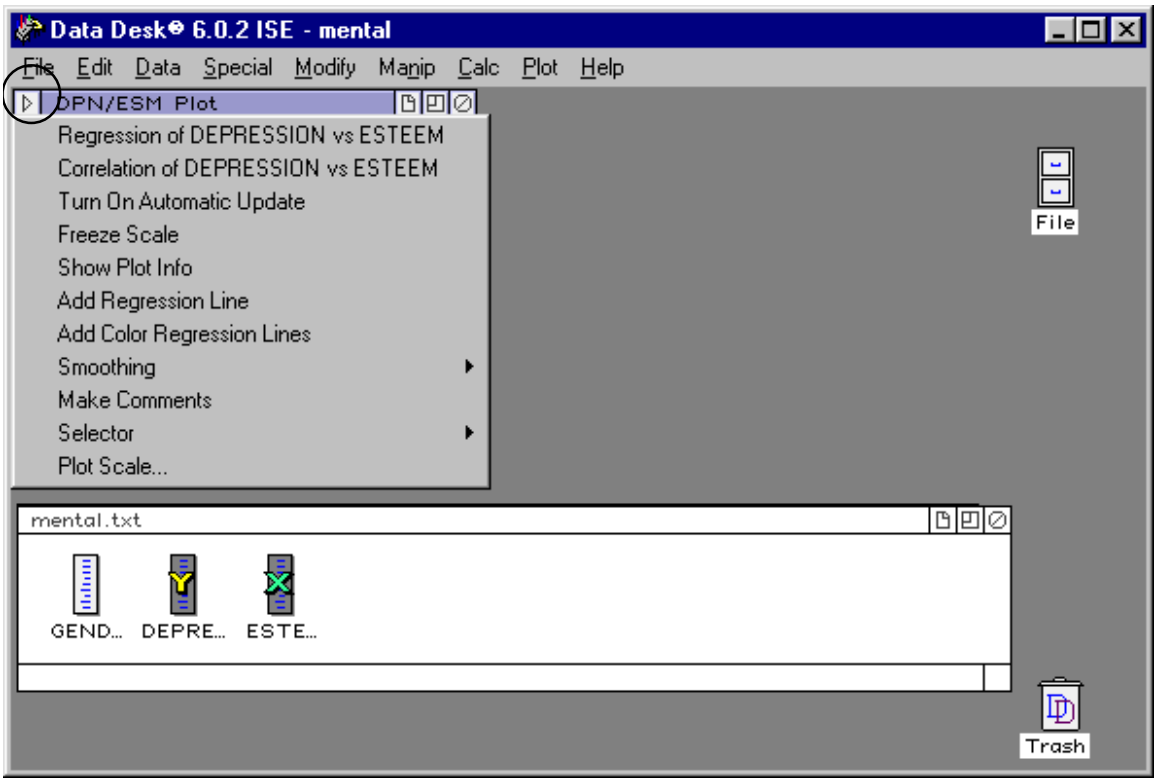
Notice as you do this, your sense of the association between depression and self-esteem may change (this means the correlation of what you observe on the screen (but not in the whole data set) is changing too!).

Plot regression lines

Now you are going to explore a new side of DataDesk. It is called the Hyperview menu.

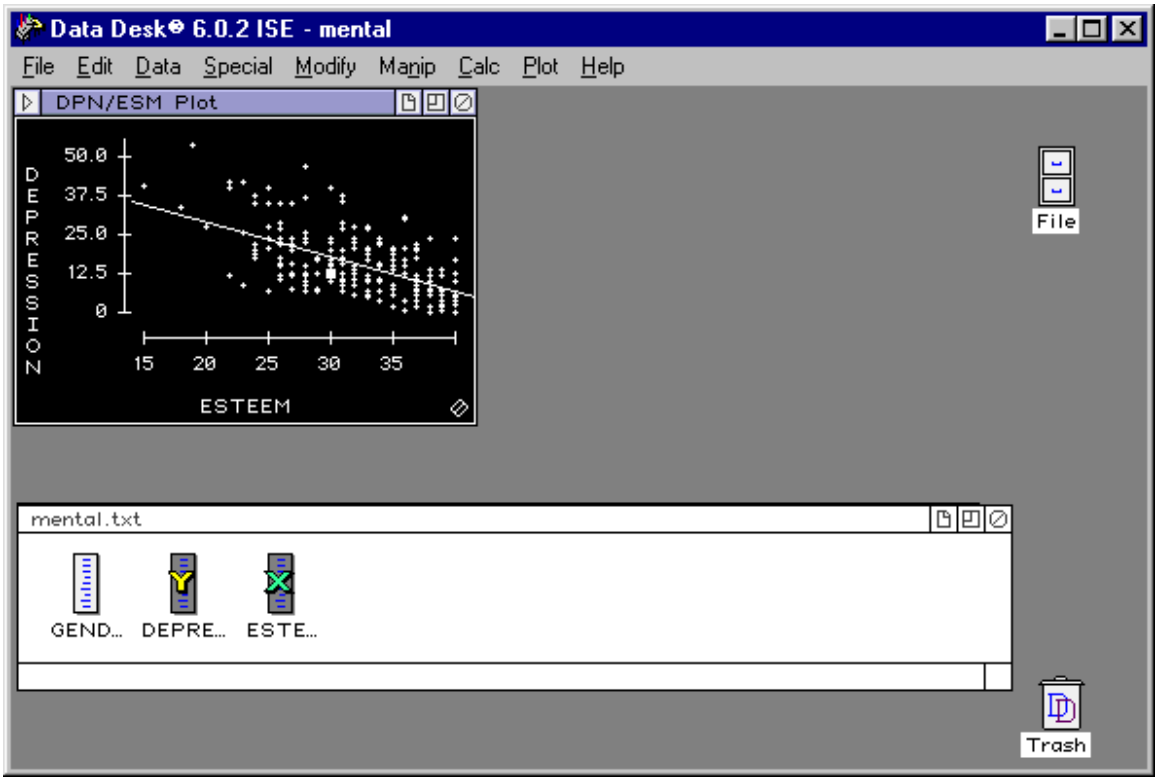
1. Click on the arrow at the upper left of the **DPN/ESM Plot** bar line

The following window will open:



2. Select **Add Regression Line**

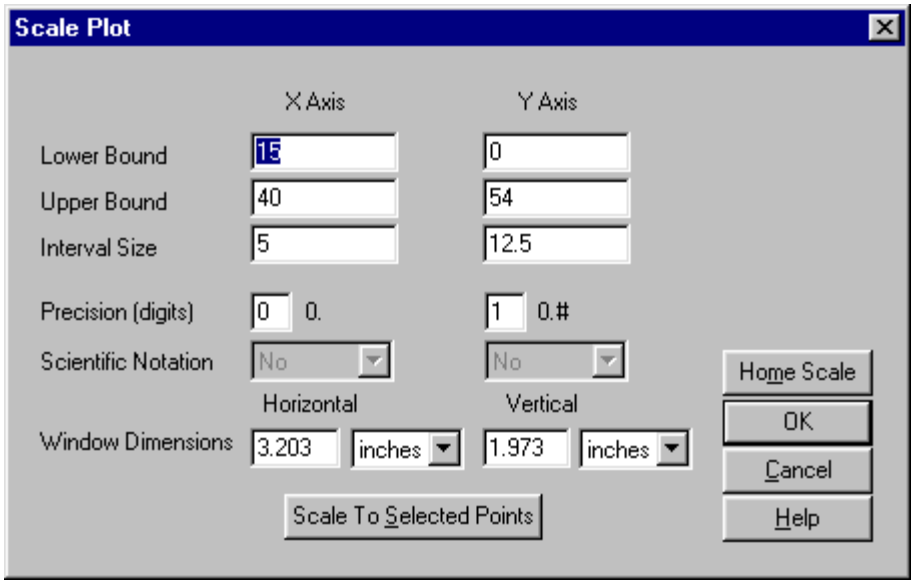
This does the following to your plot:



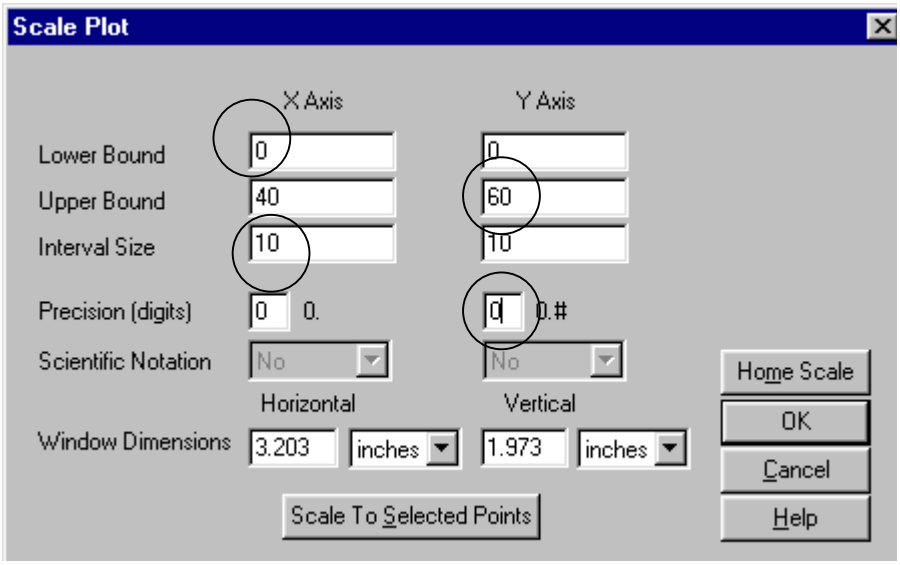
Now, try to make your plot look more spiffy.

3. In the Hyperview menu, select **Plot Scale**

The following window opens:



4. Edit this so that both variables start at zero. Set **ESTEEM** to top out at 40 and **DEPRESSION** to top out at 60. Have both variables have **interval sizes** of 10. And make both variables scale with no decimal points (**Precision digits**) as shown below:

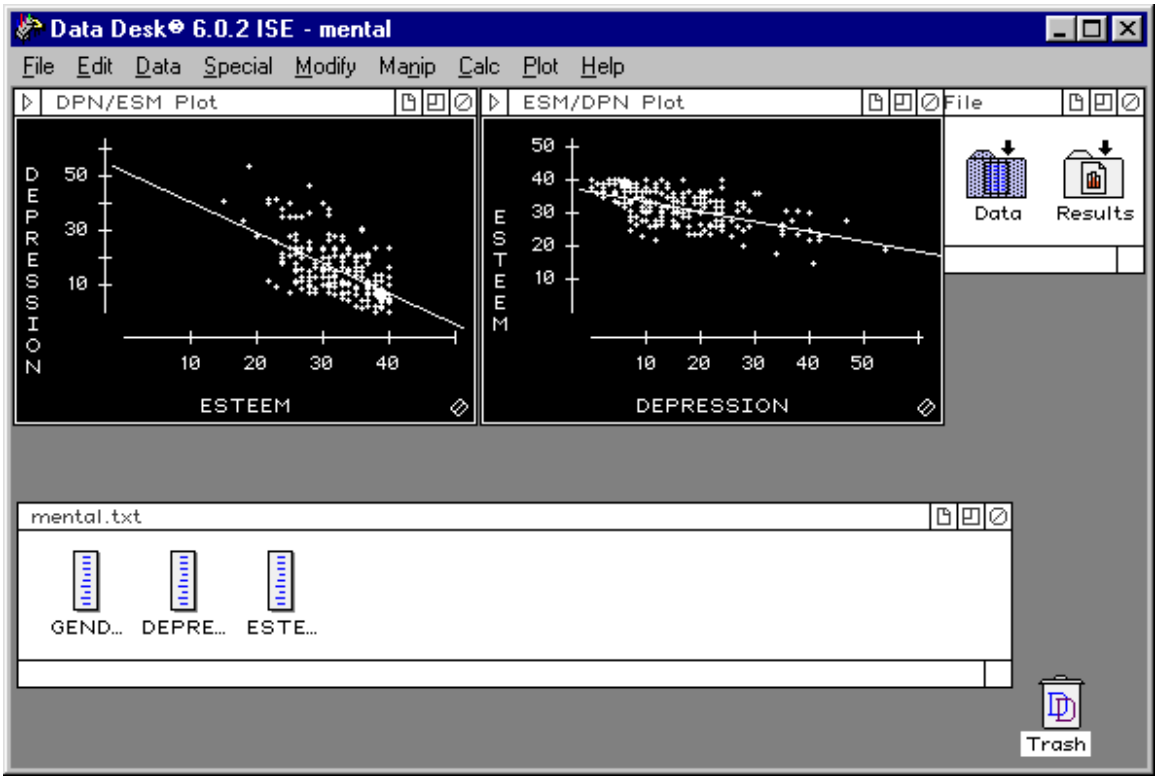


5. Click **OK**

Look how that slightly changes your plot. Leave this plot on the screen. And now you are going to plot the other regression line.

6. Click on **ESTEEM** to make it **Y**; hold the shift key and click on **DEPRESSION** to make it **X**
7. **Plot** the scatterplot relating the two variables
8. Click on the hyperview menu and adjust the **Plot scale** so that both variables start at zero, max out at 60 (for **DEPRESSION**) and 40 (for self-**ESTEEM**), go up in intervals of 10, with 0 decimal points
9. **Add the regression line**

Now you have two plots side by side:

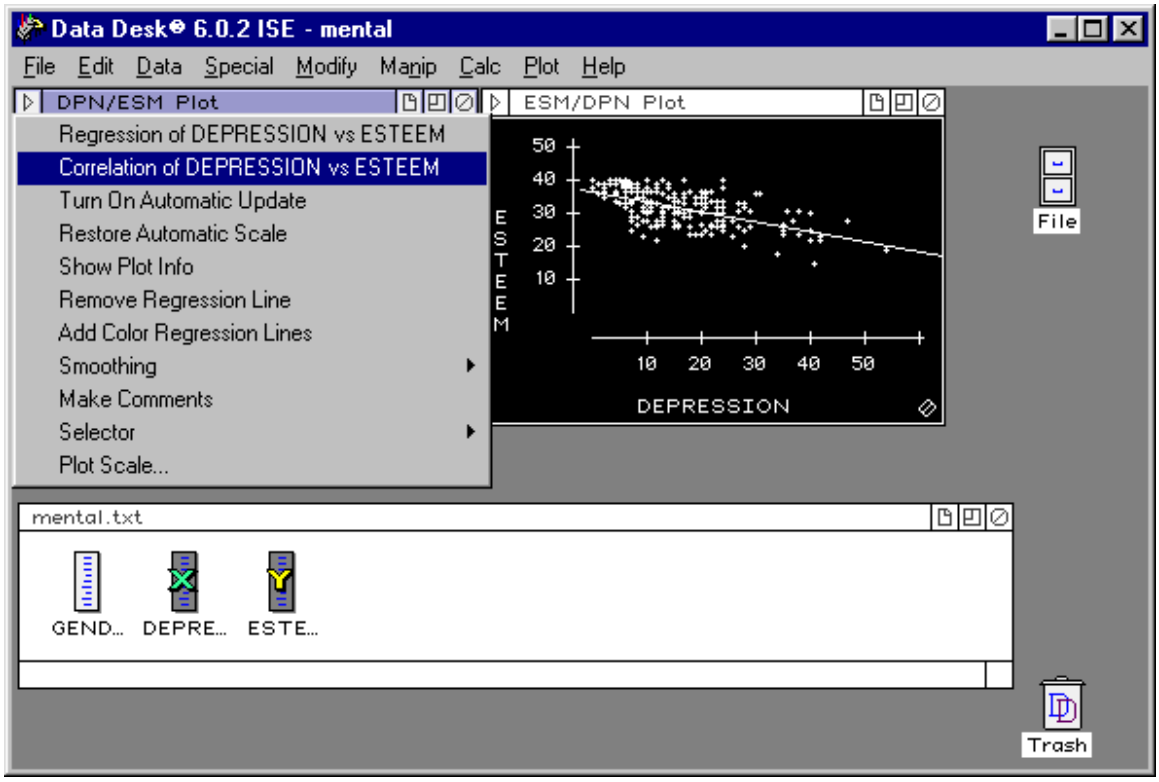


The one on the left shows depression scores predicted by self-esteem levels. The one on the right depicts self-esteem levels predicted by depression scores. Notice both the slope of the line and the intercept (the point where the line crosses the left axis) are different. Notice, too, that the plots do not look exactly the same.

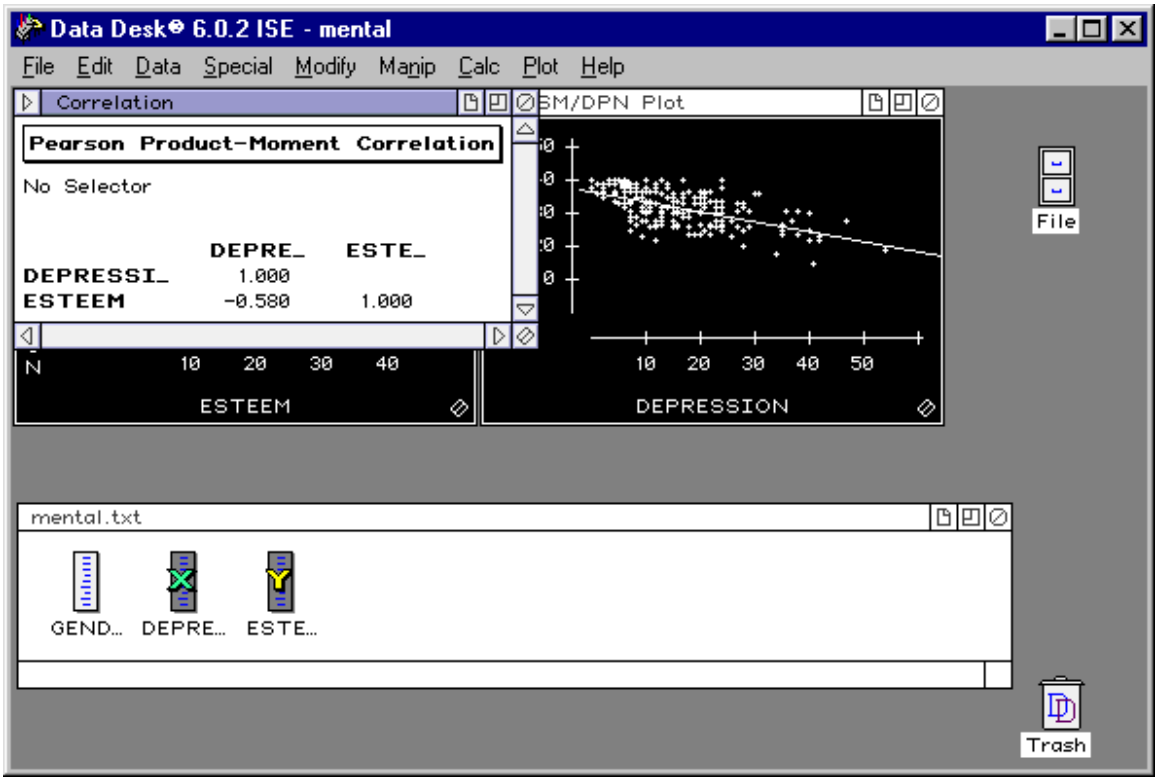
Calculate the correlation for two variables in a scattergram

To calculate a correlation

1. Click on the hyperview menu and select **Correlation of DEPRESSION vs. ESTEEM**



The following appears:



You can also do this in the other plot window with the exact same correlation result (why?). The correlation of **DEPRESSION** scores with **DEPRESSION** scores is perfect (1.00) (why?). The correlation of depression and self-esteem is moderately negative (r = -.58) as would be predicted from the visual display in the scatterplot.

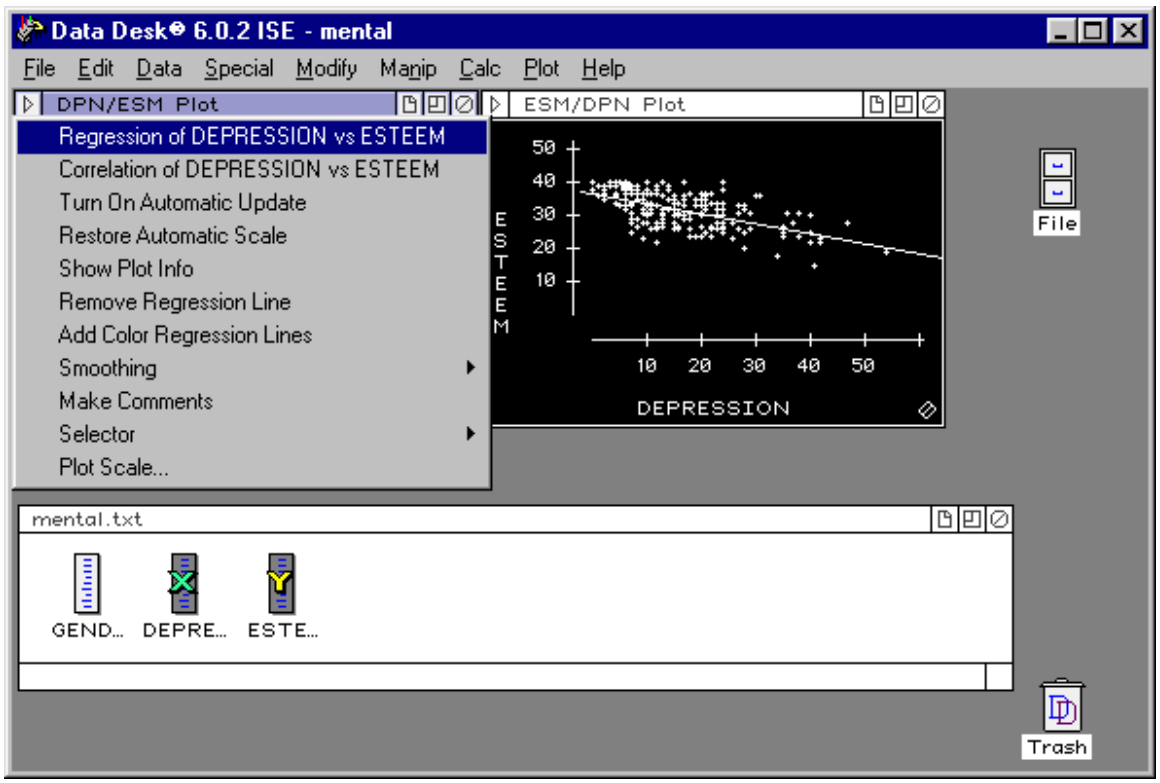
2. Now click on **Pearson Product Moment Correlation**

The window opens to show that you can choose to do two other types of correlations that we will not learn in this course. The results are different (check it out…). You can also ask for the covariance estimate, a topic not covered at length in this course.

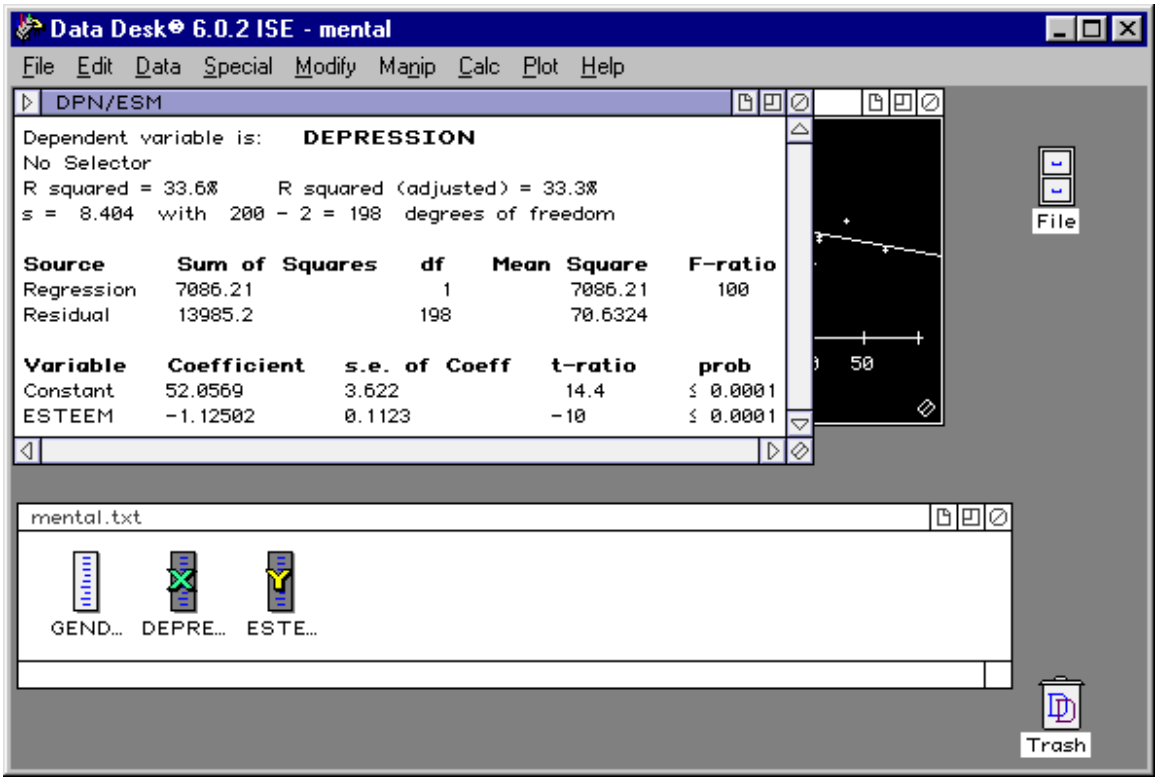
Calculate regression statistics

Now you are going to get information from DataDesk about the regression line.

1. In the hyperview menu, select **Regression of DEPRESSION vs. ESTEEM**



The following appears:



Much of this output is beyond what you have learned in this course. But some pieces you might recognize.

The **Coefficient** refers to an estimate of the linear equation, as in

DEPRESSION = -1.12502 **ESTEEM** + 52.0569

So if someone had zero self-esteem we would predict that his or her depression score would be about 52, very high in this instance.

There are also other things here we don't cover in Stat 10. Standard Errors (**s.e.**) of the coefficients are calculated (these are statements about the sampling variability that is estimated in the calculation of the coefficient). The **t-ratio** is a statistical test where the null hypothesis is that the coefficient is zero. **Prob** is the P associated with that t-value. Here it is very, very unlikely that either coefficient is actually zero.

Lab 6: Testing for Statistical Significance

Lab 6's objectives

- Perform a one-sample Z-test
- Create a grouped data set
- Perform a two-sample t-test

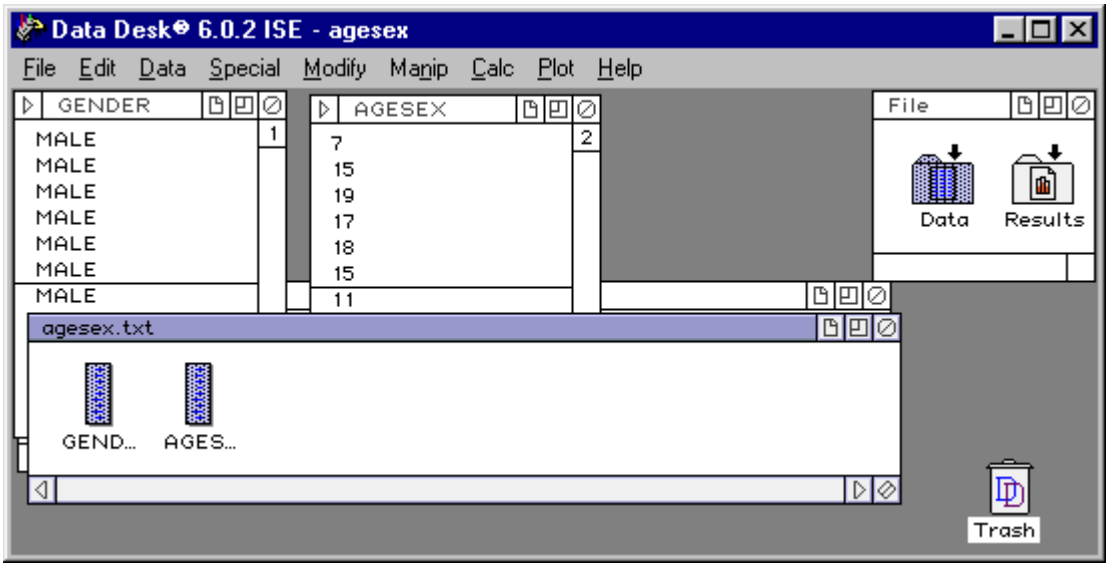
The 3rd National Health and Nutrition Examination survey estimated that among Americans between 17 and 55 years old, the age at first sexual intercourse was 17.4 years. For men, it was 16.9 years and for women it was 17.9 years. These are considered to be population-based estimates, that is, they are assumed true of Americans between 17 and 55 years of age.

In a study a few years ago, the Prof collected data from sexually experienced college undergraduates. One question that was asked was how old they were when they first had sexual intercourse. On the web, at the class data site, the Prof has stored a dataset, **agesex.dsk**, with information from 400 of these subjects, half of them male, half of them female.

Because sexually experienced college students are younger on average than sexually experienced individuals in the 3rd NHANES study, you might hypothesize that the age they report of first sexual intercourse will be significantly younger than estimates for the adult U.S. population (why would that be?). Let's test that possibility.

Perform a one-sample Z-test

1. Download the dataset, **agesex.dsk**, from the class web site (see Lab, p. 10)
2. Open your student record file (see Lab 1, pp. 6)
3. Launch (or start up) DataDesk (see Lab 1, p. 8)
4. Open the **agesex.dsk** dataset (see Lab 1, p. 10)

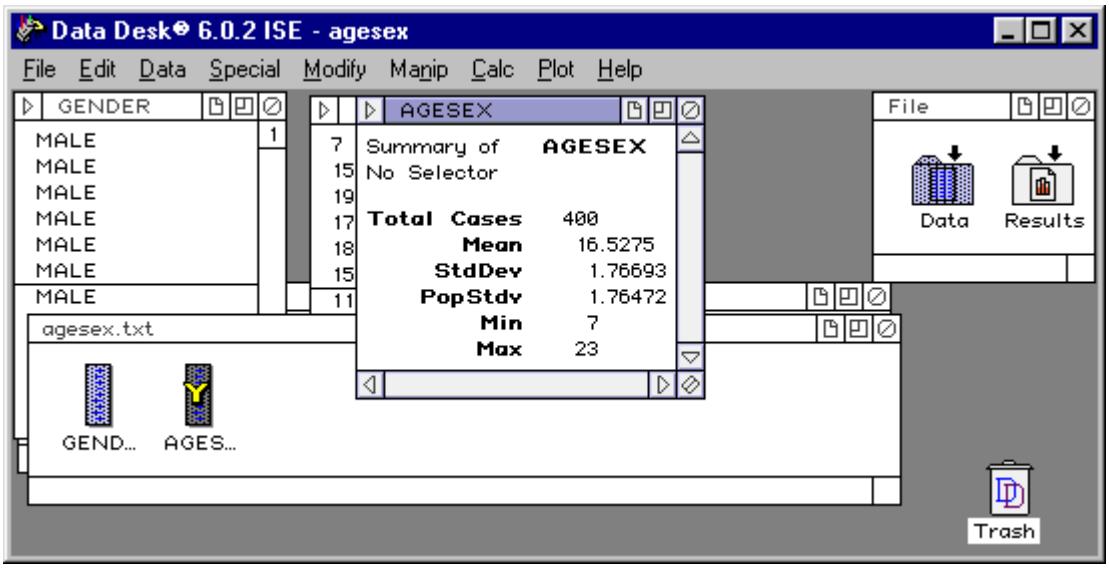


The dataset has two variables. Gender is coded as words. Age at first sexual intercourse, **agesex**, is coded as a number--the age reported by the respondent.

First take a look at the summary statistics.

5. Highlight **Calc**, adjust the options to show the mean, SD for the sample and population, and the minimum and maximum values (see Lab 1, p. 17)

6. Ask for the **summary report** (see Lab 1, p. 19)



To describe the sample: There are 400 subjects (who are all sexually experienced). They report that, on average, they were about 16 1/2 years old when they first had sexual intercourse, plus or minus about 1.8 years. The youngest age reported was 7 years and the oldest was 23. Notice that the sample SD (SD^+ in your textbook) and the estimate of the population SD (SD in your textbook) are very close. This is because the sample is relatively large. If you multiple the PopStdv by the correction factor in your textbook, you'll get the StdDev:

$$\sqrt{\frac{400}{400-1}} * 1.76472 = 1.76693$$

Now, because these are all undergraduates, it makes sense that the oldest age of first sexual intercourse is 23 years--only a handful of subjects are actually older than that. And you can't report an older age of first sexual intercourse than you currently are! That's why you would hypothesize that the age at first intercourse for this sample is younger than the population (it's not that college students are sexually precocious).

Here is the research hypothesis: The college students in this study will report a younger age at first sexual intercourse than national estimates for Americans 17 to 55 years of age.

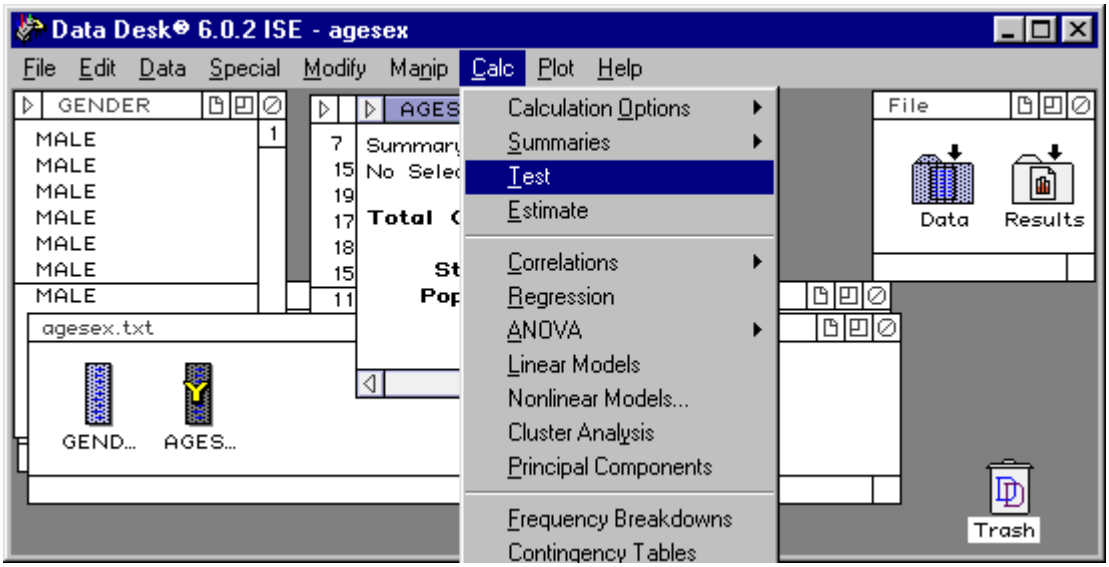
Here are the statistical hypotheses:

H_0: Mean of college students = or > 17.4 years (Null hypothesis)

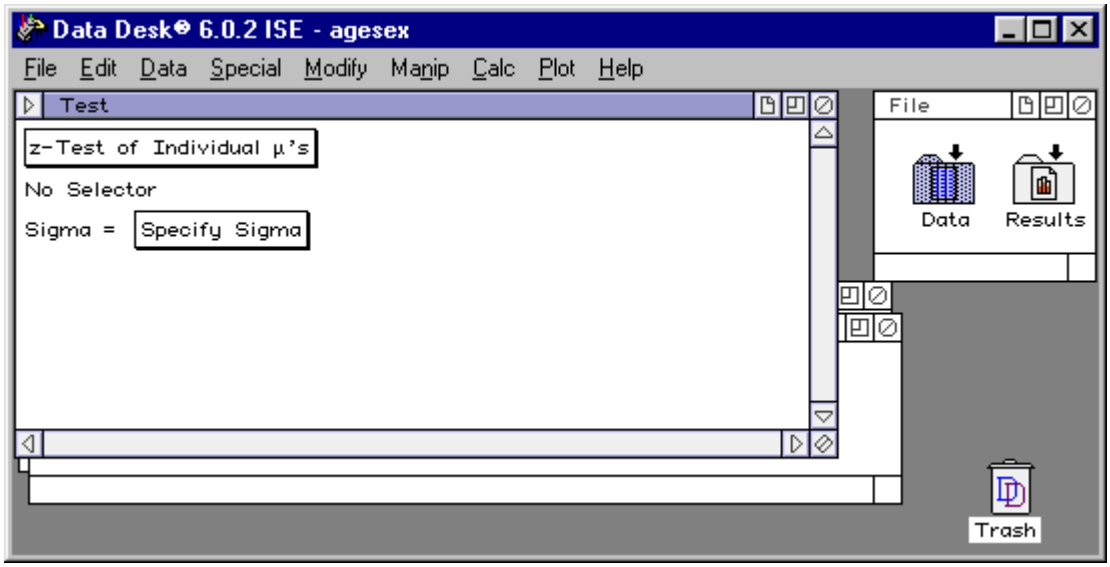
H_1: Mean of college students in study < 17.4 years (Alternate hypothesis)

Now it's time to test the hypothesis.

6. Highlight **Calc**, and select **Test**

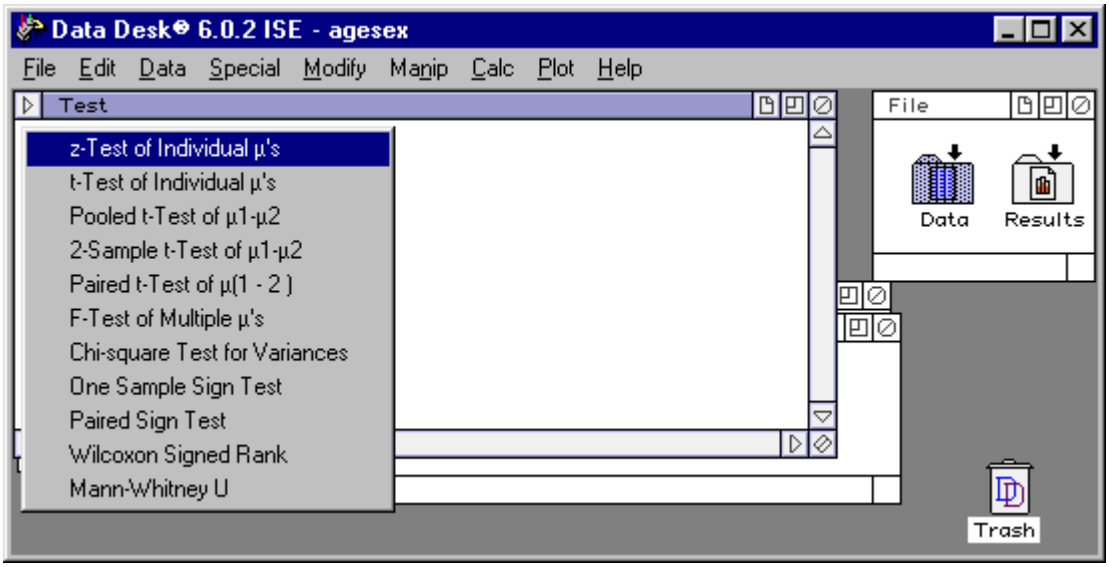


The following window opens (click on the bottom righthand corner and drag it to the right and down to make the window bigger):



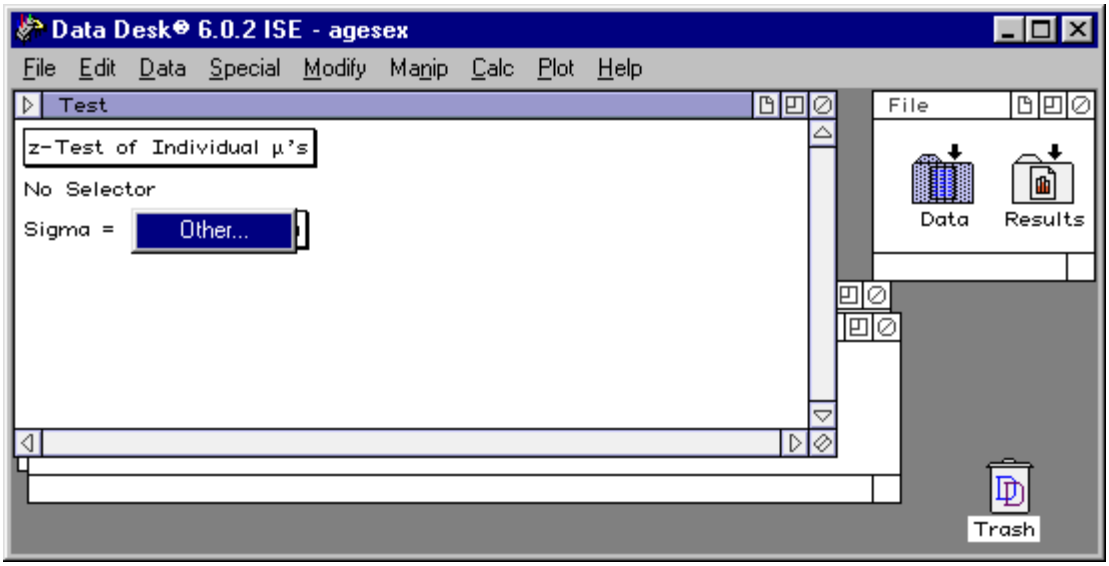
The first decision you have to make is which type of test you want DataDesk to do.

7. Click on **z-Test** to see your options. Choose **z-Test of individual μ's**

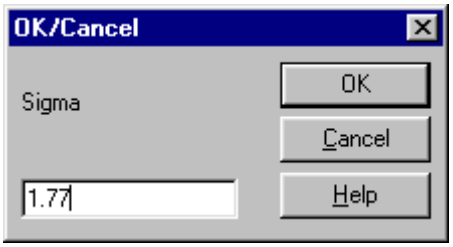


The next decision is what you estimate the SD of the U.S. population to be (the SD of the box or SD^+). Your best guess is the SD of your sample (SD^+), or 1.77.

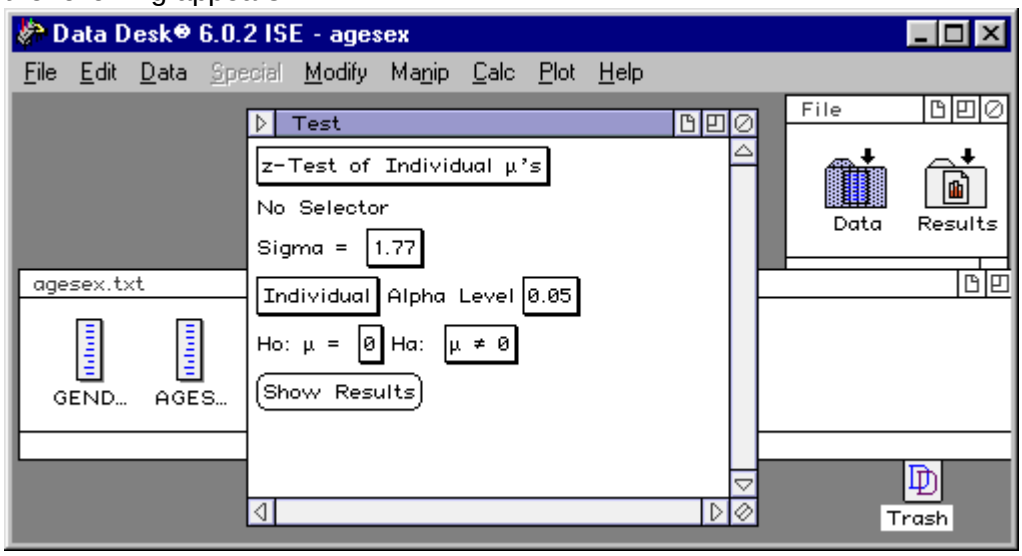
8. Click on **Specify sigma**. **Other** will appear.



9. When you let go another box will appear. Enter your estimate and click **OK**



Now the following appears:

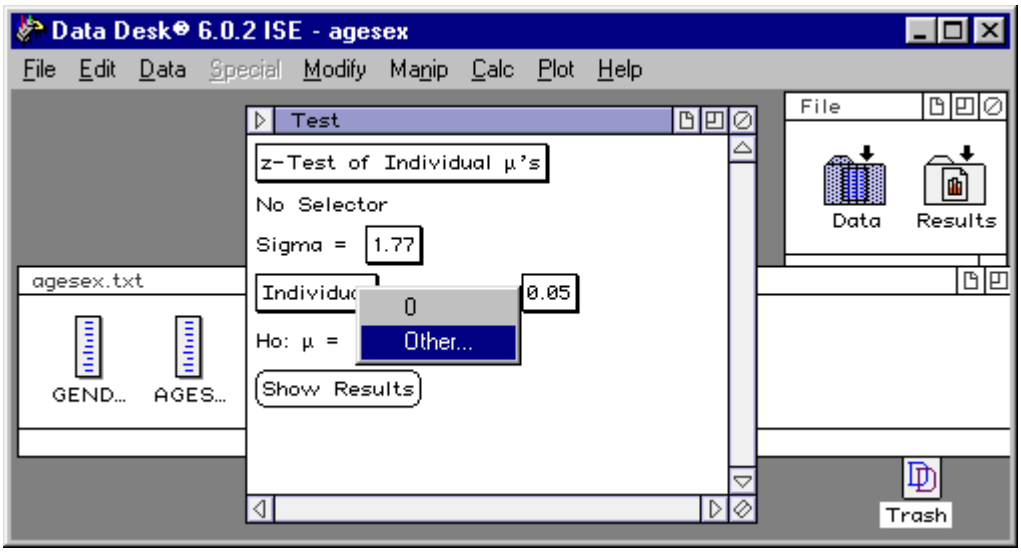


The next line down from **Sigma** indicates that you are asking DataDesk to conduct an **individual** (or single statistical) test at the .05 **alpha level**. That means you are willing to reject the Null Hypothesis if P is less than or equal to .05. Or another way of saying this, you are willing to conclude that the sample (the 400 college students) was not drawn randomly from the population (all Americans between 17 and 55 years of age) if the difference in age at first sexual intercourse between your sample and the population is consistent with what would occur 5% or less of the time *if in fact the 400 individuals had been drawn randomly* from the population. That's a brain twisting. Get it? If the difference is so rare, and so unlikely to occur naturally, you will find it just too hard to believe the Null hypothesis is a good explanation for your data. And if the Null hypothesis is very unlikely to be true--that only leaves the Alternative Hypothesis as an explanation for what you observe.

You do not need to edit this line about the **Alpha Level** in DataDesk.

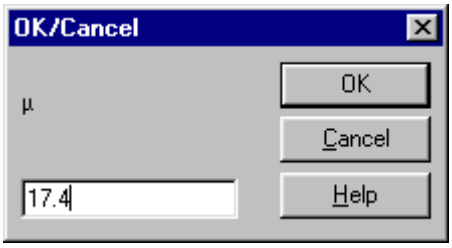
The next line you do need to edit. It indicates the population mean or $\mu = 0$, but for your hypothesis it should read 17.4.

10. Click on μ_0 and select **Other**

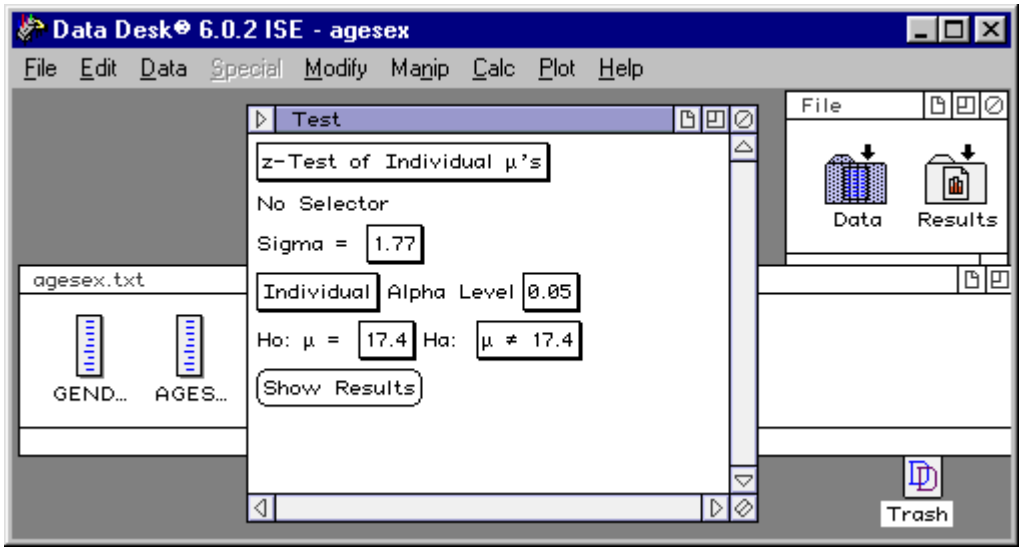


You can now enter the estimated age at first sexual intercourse for Americans

11. Enter **17.4** in the window and click **OK**



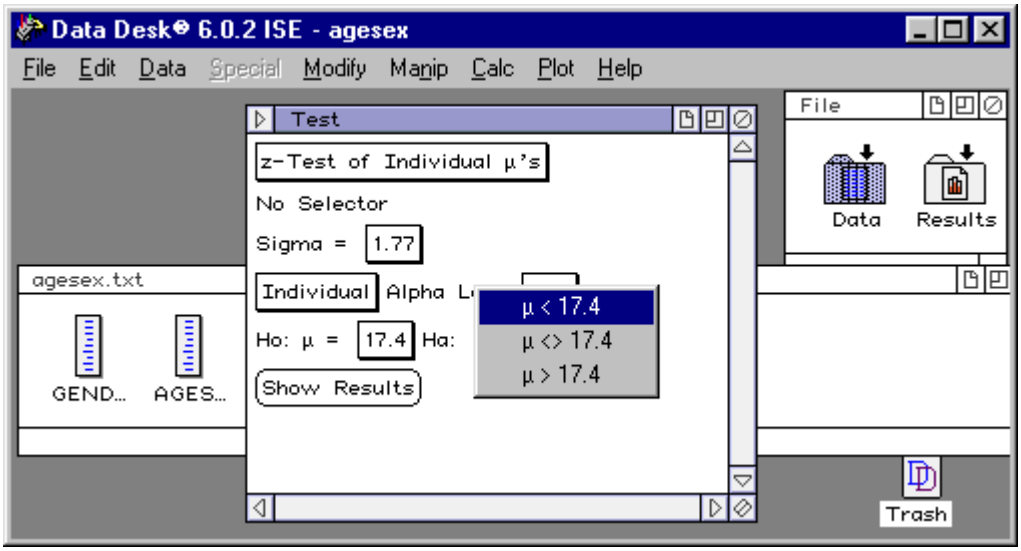
The following appears:



Now you have one last piece to edit.

Ha, or Alternative hypothesis, indicates that your alternative is simply a value not 17.4.

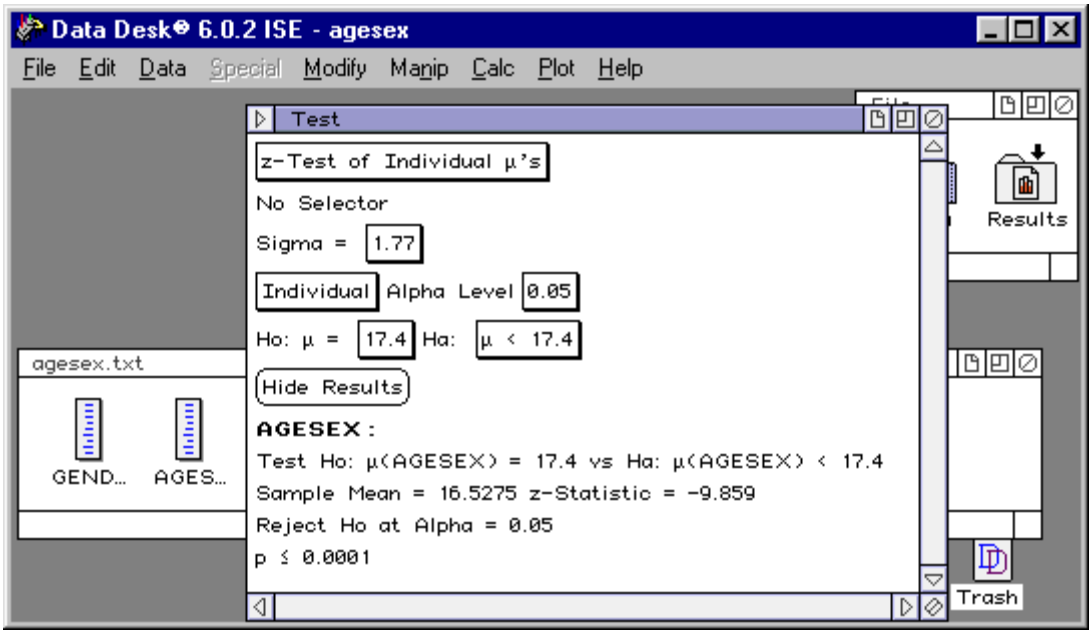
12. Change **Ha** to show that your hypothesis is $\mu < 17.4$ by clicking on $\mu \neq$ **17.4** and



select $\mu <$ **17.4**

13. Click **Show Results**

The following appears:



DataDesk now gives you all the information you need, including the Z-value, the decision (reject Null hypothesis at the P = .05 level) and the p-value associated with a z-value this big.

It seems the college students in the survey did report a younger age of first sexual intercourse than Americans age 17 to 55 years.

Create a grouped data set

Another question you can ask of this dataset is whether men and women differed in the age at which they first experienced sexual intercourse.

Here is the research hypothesis: Men and women differ in their age at first sexual intercourse.

Here are the statistical hypotheses:

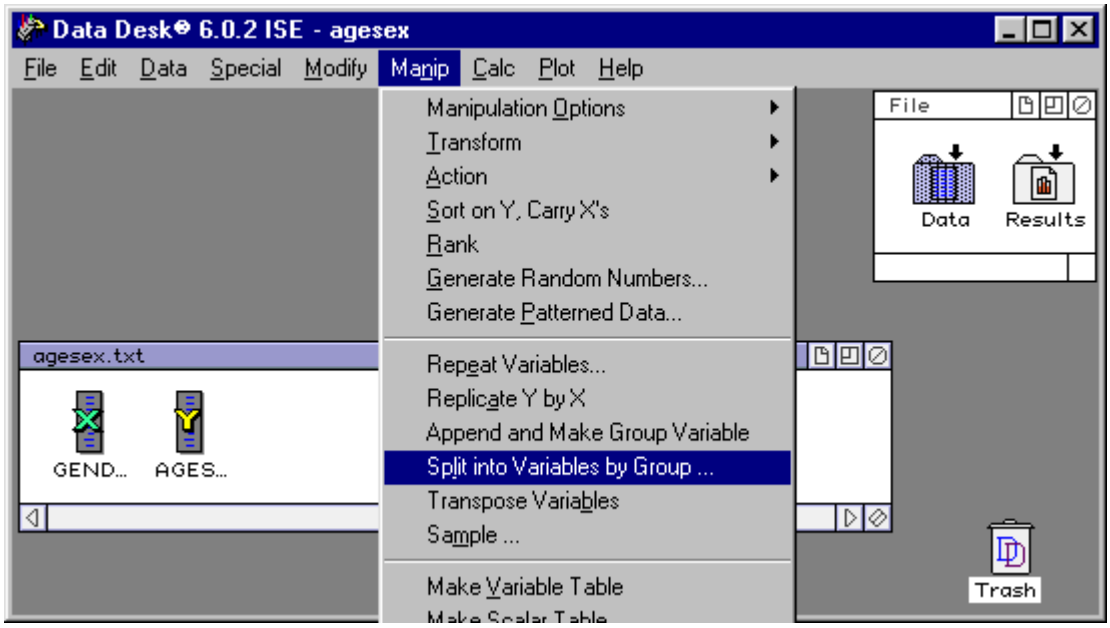
H_0: Mean age at first sex for men = mean age at first sex for women
(Null hypothesis)

H_1: Mean age at first sex for men ≠ mean age at first sex for women
(Alternate hypothesis)

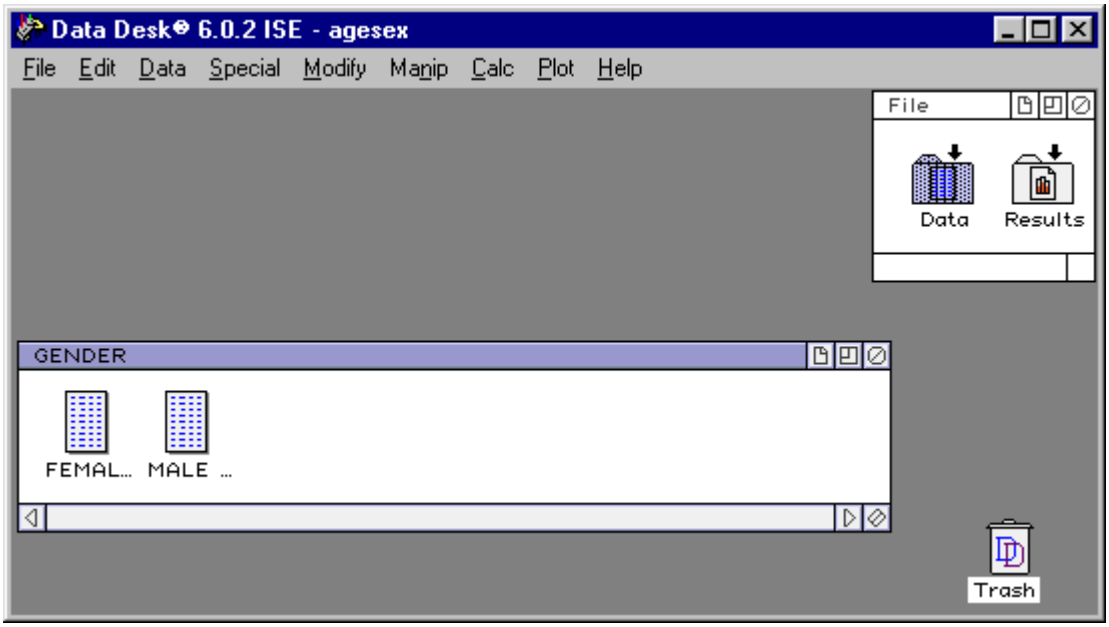
Before you can test the hypothesis, you have to divide the ages into those that are men's and those that are women's.

1. Make **agesex** the **Y** variable and **gender** the **X** variable (see Lab 4, p. 38)

2. Highlight **Manip** and select **Split into Variables by Group**



The following appears:



3. Click on the **Female** icon until you see the ages appear

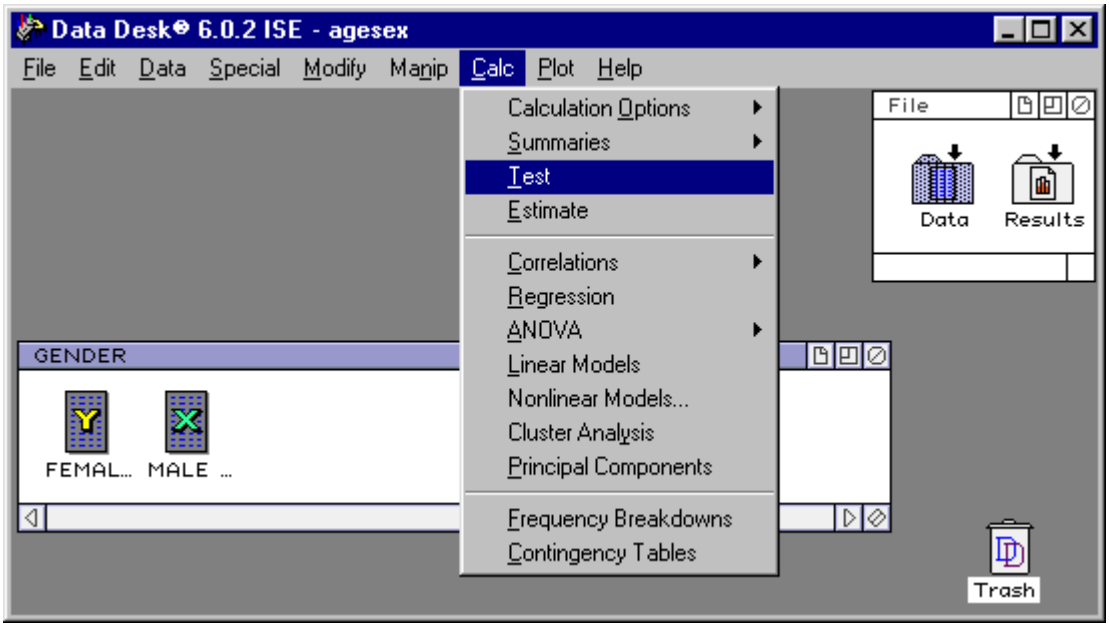
Notice that DataDesk has split the sample into two groups. You can examine the summary statistics on females to see that there are only 200 subjects.

Now you can contrast men and women.

Perform a two-sample t-test

1. Assign women (**Female**) to be **Y** and men (**Male**) to be **X**

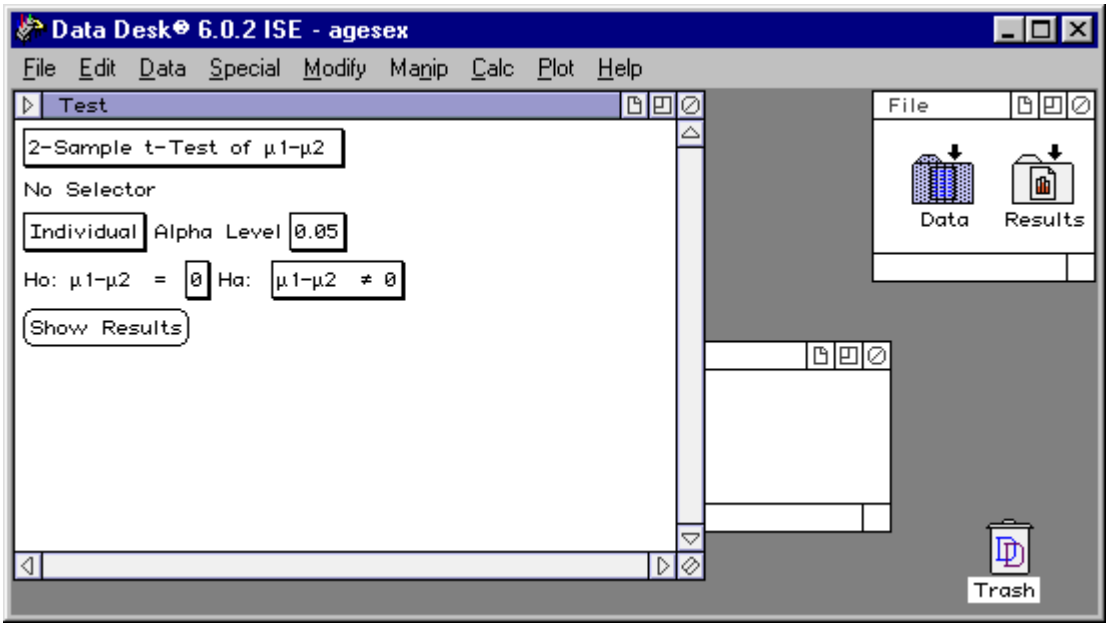
2. Highlight **Calc** and select **Test**



2. Select **2-Sample t-test of μ1- μ2** (this is nearly identical to a 2 sample Z-test when the sample is this large)

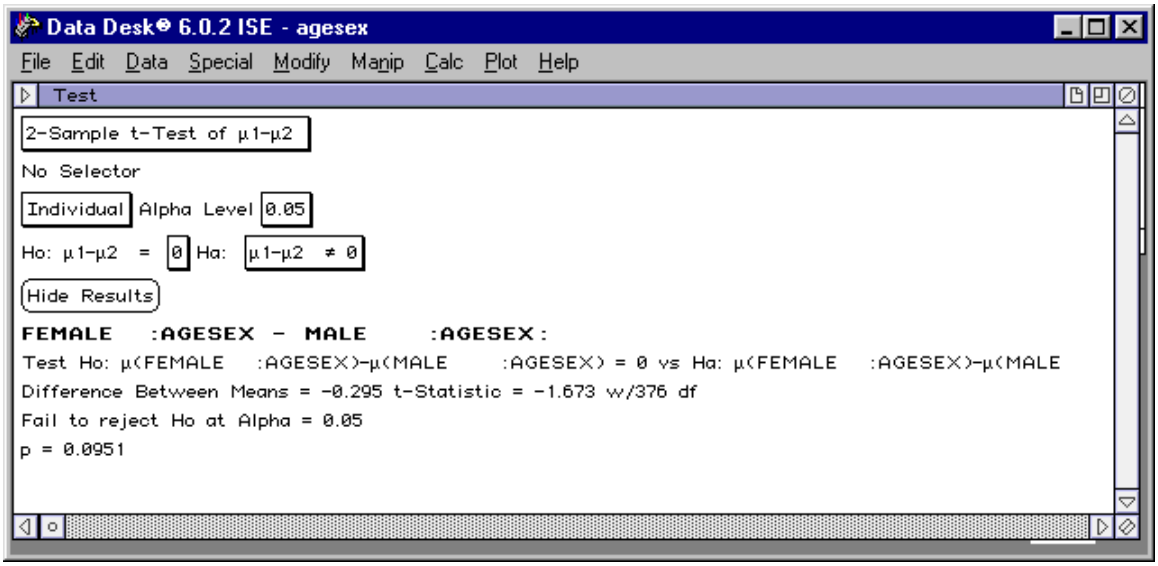
3. Edit the **Ho:** to show a difference of 0 (the μ's or population means equal) (see Lab 6, p. 66)

4. Edit the **Ha:** to show any difference (the μ's or population means are not equal) (see Lab 6, p. 67)



5. Now click **Show Results**

The following window appears:



The results tell you that in a test of the Null hypothesis, the **t-statistic** is relatively small, **p** is relatively large (p = .0951). That is about 9% of the time you would expect a difference in age at first sexual intercourse between men and women this large or larger to arise simply from our sampling of the population and that in the whole population there really is no difference between men and women. This does not meet our criteria for rejecting the null hypothesis, so we fail to reject it. We can't conclude the men and women in this survey first experienced sexual intercourse at the same age; we can't say they didn't. We simply find no evidence to believe there is a difference between men and women. The difference that we do observe is consistent with chance variation.

Lab Manual (DISC) Homework Problems

1. If you open the **Modify** menu at the top of the screen in DataDesk, what is the first or top listed option available to you?

2. If you open the **Special** menu at the top of the screen in DataDesk, what is the first or top listed option available to you?

3. Using the **List1** variable from Lab 2, calculate the skewness of this distribution of 6 elements. To do this,
   - Request skewness to be reported in the summary statistics
   - Select **List1** as a variable to be analyzed
   - Request a summary statistics report--The answer will appear on your output!

4. Using the **List1** variable from Lab 2, calculate the interquartile range of this distribution of 6 elements. To do this,
   - Request the interquartile range to be reported in the summary statistics
   - Select **List1** as a variable to be analyzed
   - Request a summary statistics report

5. A box contains 10,000 tickets: 4,000 with a 0 on it; 6,000 with a 1 on it. Estimate the skewness of the box. To do this:
   - Create a relational database--hint: there are two variables in the relation: value (0,1) and count (4,000, 6,000)
   - Replicate value by count
   - Ask for skewness in the summary statistics of the new Count:Value variable created by replication.

   (You won't be able to save this analysis to disk because we are using a student version of DataDesk that limits you to 1,000 cases)

6. Using the **prezages.dsk** dataset, what is the interquartile range for the distribution of president's ages?

7. Using the **prezages.dsk** dataset, estimate the age that cuts off the upper 20% of the distribution. To do this:
   - Follow the instruction on pages 30-32
   - Set the **Percentile** value in the **Order** section to 20

8. Suppose a student took an exam consisting of 100 questions. Each question had 5 answer options in which a correct answer was worth 4 points and an incorrect answer was not penalized, what would you expect the student to score on the exam? Also give an estimate of chance variation in this score. To do this:
   - Use the **Quiz box** variable from Lab 4
   - Edit it to contain the following elements {4,0,0,0,0}
   - Using the **Calc**ulation option, ask for the mean of this variable and the population SD
   - By hand, multiply the mean you observe by 100 to create your estimate of the student's score
   - By hand, multiply the population SD by the squareroot of the sample size to estimate the SE for the sum (10 * pop. SD)

9. Suppose a student took an exam consisting of 100 questions. Each question had 5 answer options in which a correct answer was worth 5 points and an incorrect answer was penalized 1 point, what would you expect the student to score on the exam? Also give an estimate of chance variation in this score. To do this:
   - Use the **Quiz box** variable from Lab 4
   - Edit it to contain the following elements {5,-1,-1,-1,-1}
   - Using the **Calc**ulation option, ask for the mean of this variable and the population SD
   - By hand, multiply the mean you observe by 100 to create your estimate of the student's score
   - By hand, multiply the population SD by the squareroot of the sample size to estimate the SE for the sum (10 * pop. SD)

10. How is the variable, **GENDER**, scaled in the **mental.dsk** dataset? Is it qualitative? Quantitative? Categorical? Interval? Ratio? Please give at least two descriptive terms and explain.

11. What is the range for the variable, **DEPRESSION**, in the **mental.dsk** dataset?

12. What is the estimate of covariance between **DEPRESSION** and **ESTEEM** in the **mental.dsk** dataset?

13. In the **agesex.dsk** dataset, what is the mean age at which women report their first experience with sexual intercourse?

14. In the **agesex.dsk** dataset, what is the mean age at which men report their first experience with sexual intercourse?