COMPUTER LAB MANUAL (Statistics Lab Version)

STATISTICS 10

Winter, 2000

Table of Contents

Introduction: Details, Details

This version of the lab manual is for use in the Statistics Lab. The manual is also on the class website.

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. There is an Windows version of this manual on the web at the course web site. 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 comand 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).

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

To log on to computers in the stat lab, you **absolutely** need to have a bruin-on-line ID (BOL ID). If you do not have one, go downstairs *now* to OAC in Math Sciences 4302 and sign up for one.

If you are not officially either enrolled in the class or officially on the waiting list, you need to see Jose Garcia to get access to the computer lab. Talk to your TA about this.

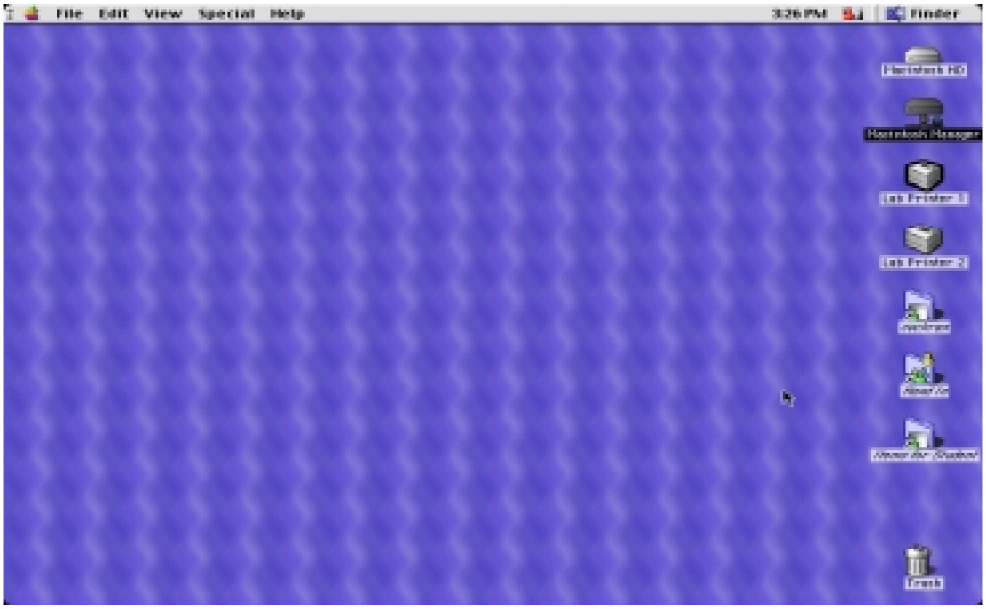
The first time you log on:

1. Enter your BOL account name (leave off the @ucla.edu extension)
2. Your password is your 9 digit student ID number
3. Follow the instructions to change your password

After this, every time you log on, simply

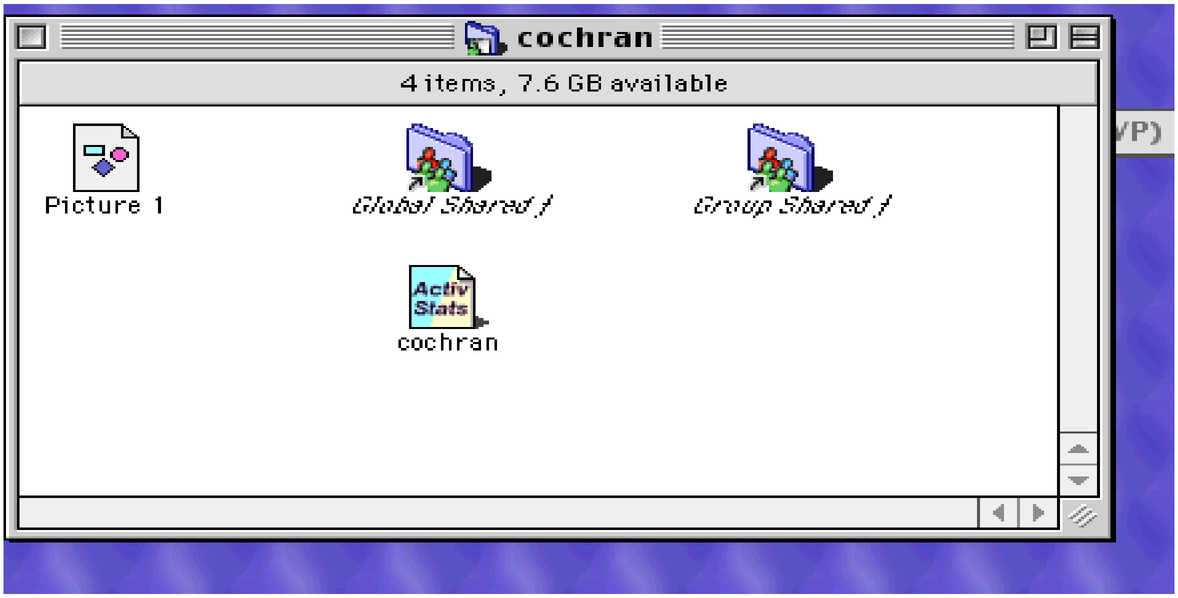
1. Enter your BOL account name
2. Enter your password

When you login you will see the following:



3. Click on the folder with your bruin-on-line name (what shows here is the Prof’s bruin-on-line name).

Then you’ll see something like the following, though it will show your BOL ID not the Prof’s:



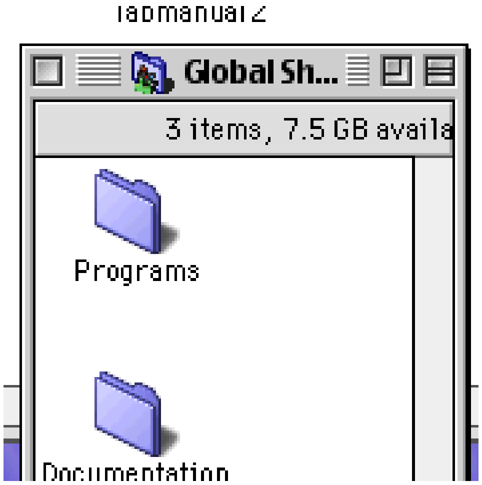
4 items, 7.6 GB available

On a Mac, each window has:

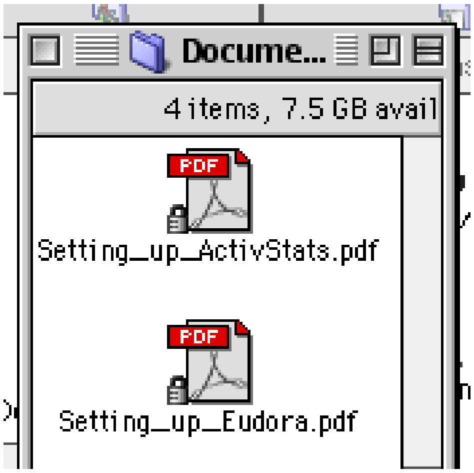
- a slot on the top left that closes the window
- a slot on the top far right that makes the window a bar (with just the the name of the window showing)
- a slot on the top right just to left of that--this resizes the window
- a touch pad at the corner of the bottom right hand side that if you put the cursor on it and hold the mouse key down allows you to resize the window however you want
- Right side and bottom sliders that allow you to change what shows on the screen if it won’t all fit within the window

4. Now click on Global Shared Files inside your BOL ID Folder

The following window opens up:

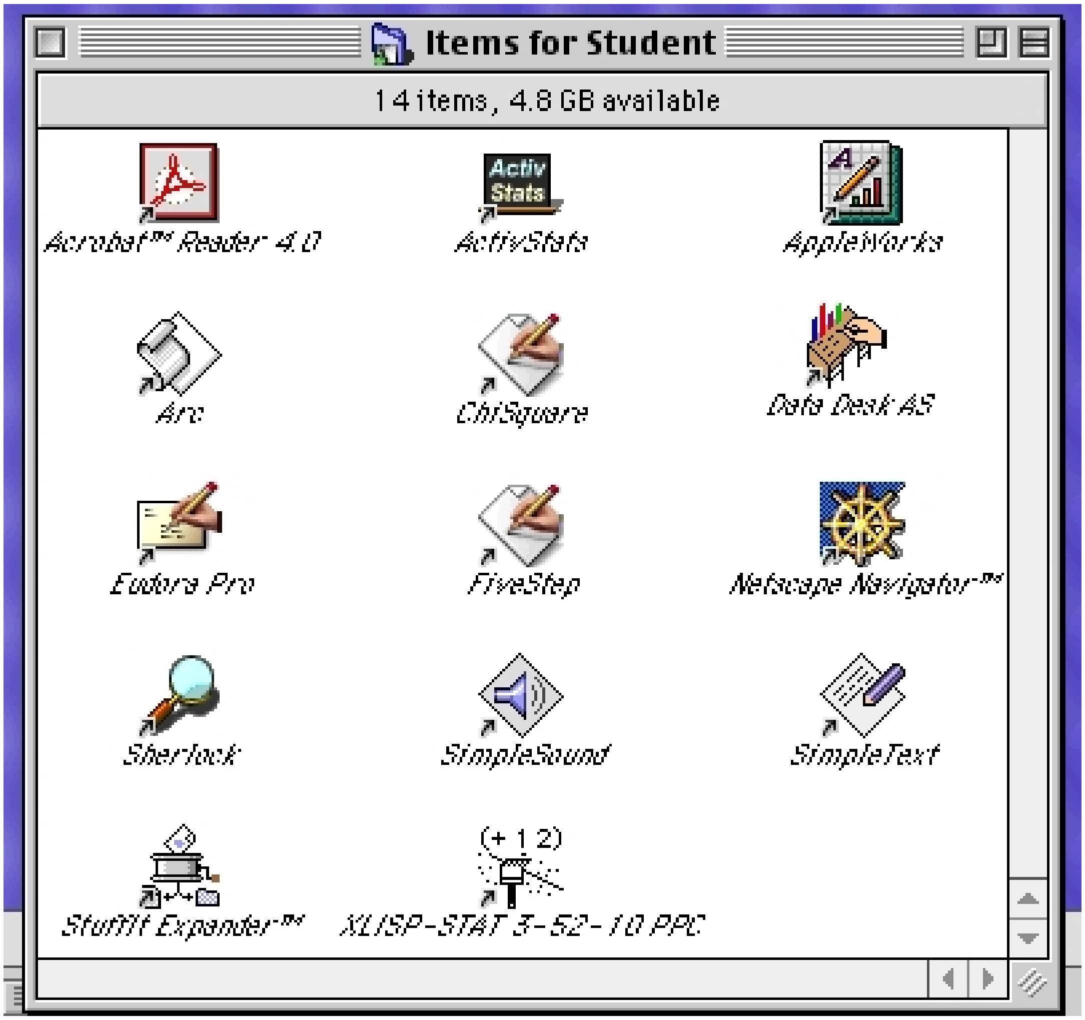


5. Click on Documentation



Now you can see How-to Instructions for use different aspects of the computer lab.

6. Now click on the folder labeled Items for Student on the right side of the screen



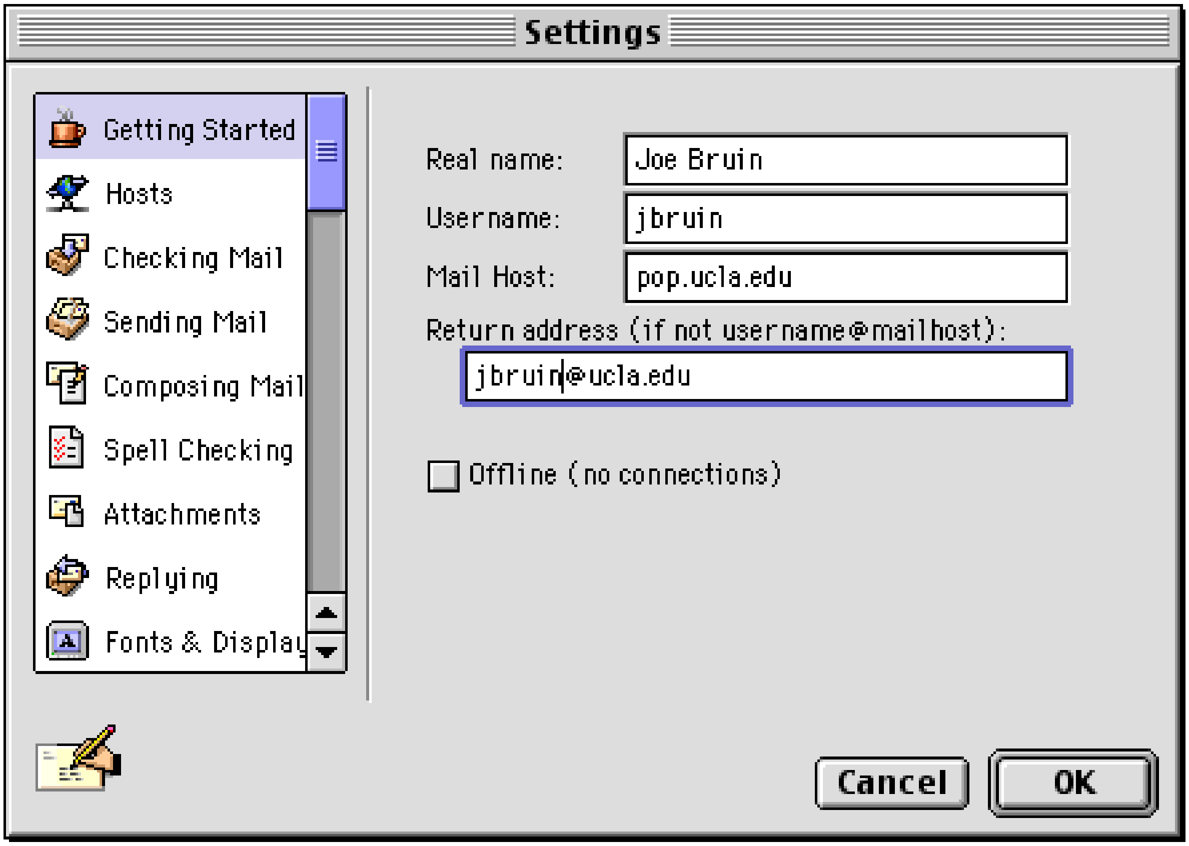
There are several programs available for you to use. The first step is to configure Eudora so that you can send and receive material from the lab at your bruin-on-line account.

Set up Eudora for your use

Eudora requires some information the first time you run it.

1. Double click on the Eudora Pro icon (it's in the Items for Student folder)

You should see the Settings window appear as shown below. If it isn't already open, open it up from the menu at the top of the screen. Select Special. Then select Settings



Real name: Joe Bruin

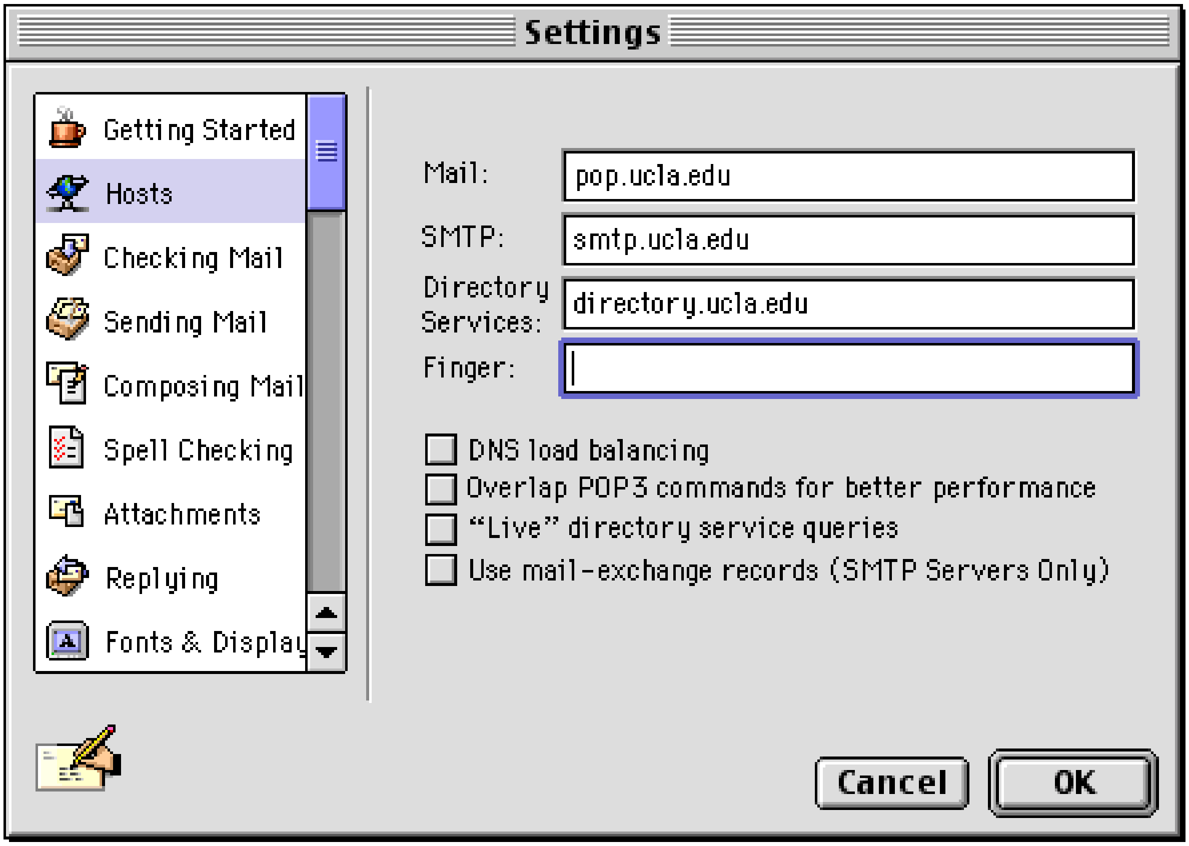
Username: jbruin

Mail Host: pop.ucla.edu

Return address (if not username@mailhost): jbruin@ucla.edu

□ Offline (no connections)

2. Click on the Getting Started icon on the left
3. In the Real name box, type your full name
4. In the Username box, type in your BOL ID
5. In the Mail Host box, type pop.ucla.edu
6. In the Return address box, type in your BOL email address
7. Click on the Hosts icon on the left



Mail: pop.ucla.edu

SMTP: smtp.ucla.edu

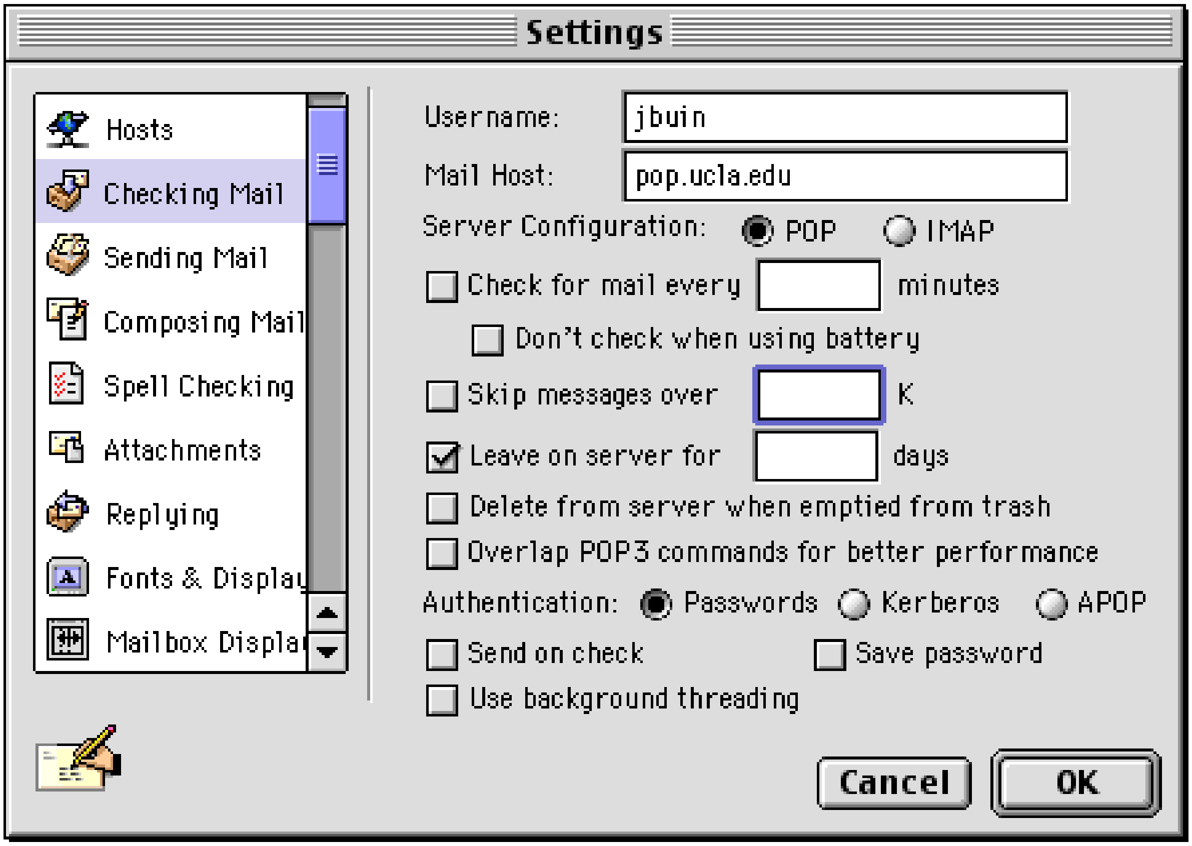
Directory Services: directory.ucla.edu

Finger:

□ DNS load balancing
□ Overlap POP3 commands for better performance
□ "Live" directory service queries
□ Use mail-exchange records (SMTP Servers Only)

8. In the SMTP box, type smtp.ucla.edu
9. In the Directory Services box, type directory.ucla.edu

10. Click on the Checking Mail icon on the left



Username: jbuin

Mail Host: pop.ucla.edu

Server Configuration: ● POP ○ IMAP

□ Check for mail every minutes

□ Don't check when using battery

□ Skip messages over K

☑ Leave on server for days

□ Delete from server when emptied from trash

□ Overlap POP3 commands for better performance

Authentication: ● Passwords ○ Kerberos ○ APOP

□ Send on check □ Save password

□ Use background threading

11. Put a check in the Leave on server box. Leave the box for days blank

If you do not do this last step, then your mail, if you check it in the lab, will be permanently delivered to the Statistics Computer Lab and deleted off of Bruin-on-Line (Ouch!). Be sure to do step 11.

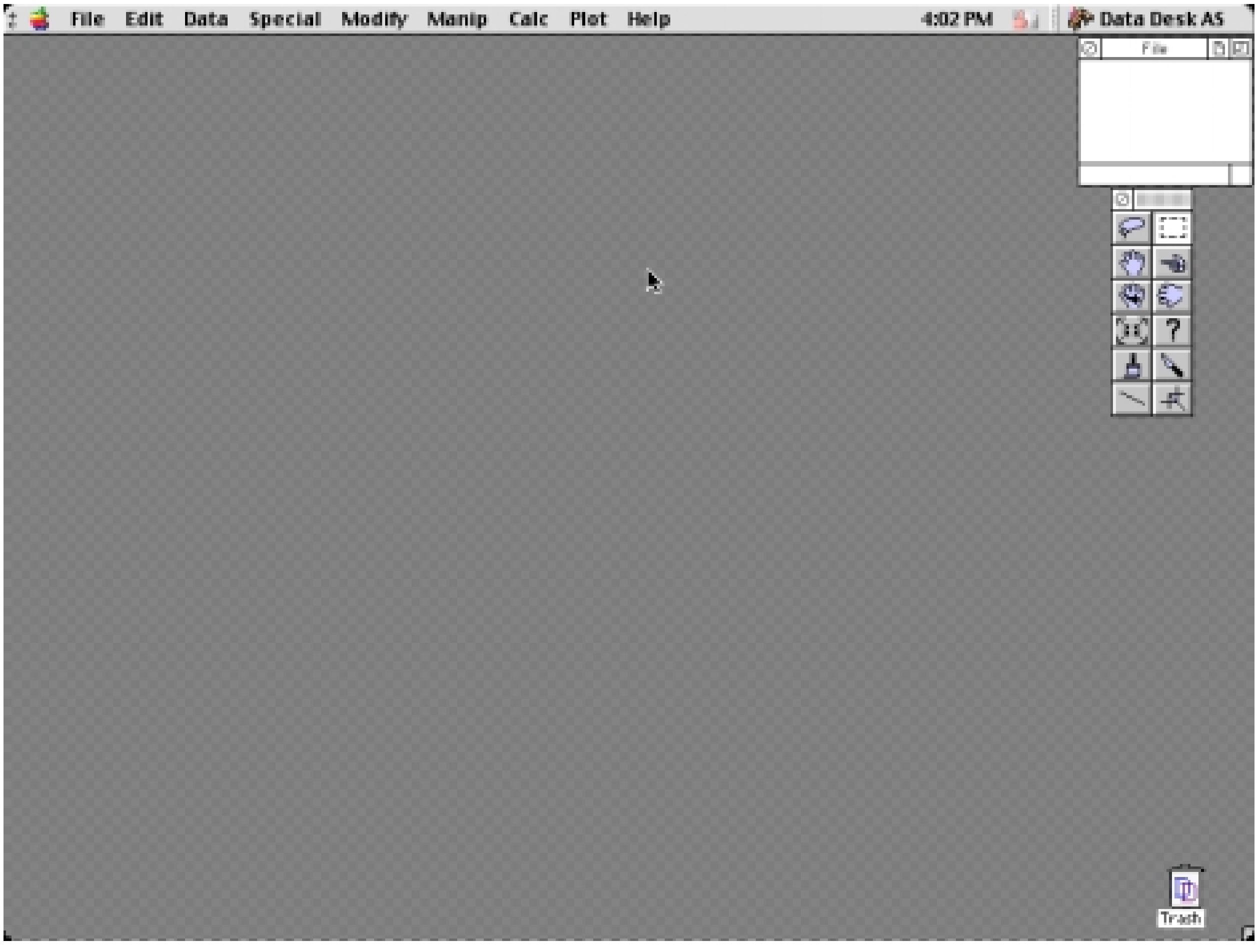
12. Click OK

Also, send yourself an e-mail to make sure that the settings are correct and that the system is working for you. Now you can send mail to yourself from the lab. If you want to mail yourself output just send it as an attachment. You can also mail yourself material to use in the lab by sending yourself an e-mail and then reading it here in the lab.

Open DataDesk

1. Click on the icon Data Desk in the Items for student folder

The following window will open:



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

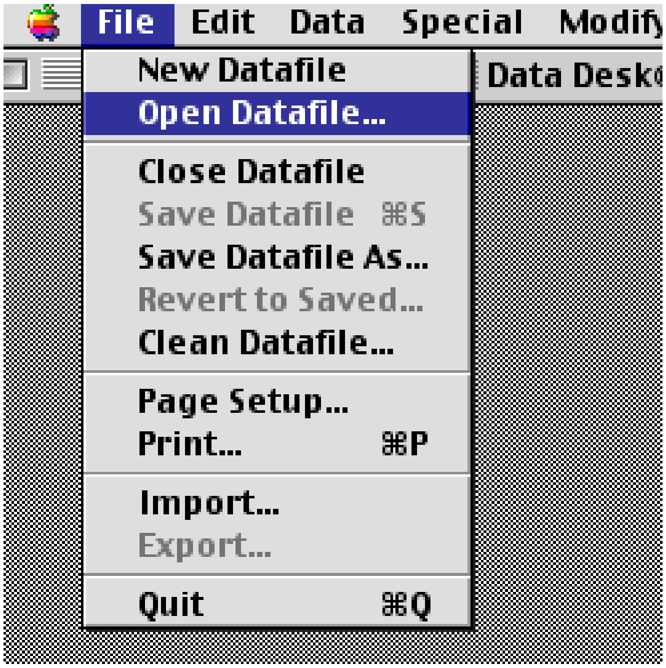
At the top right is an icon called File--it may also appear as an open window.

Open an existing data set

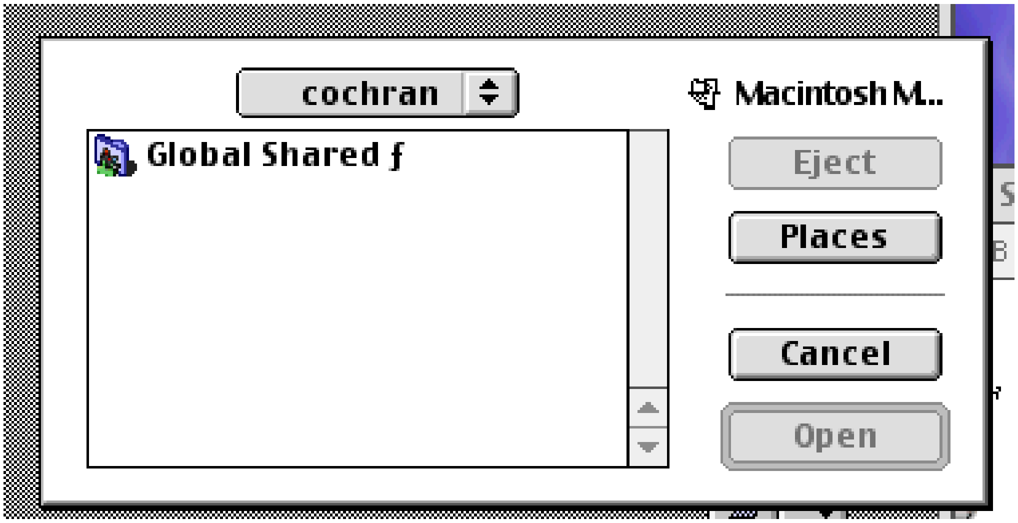
There is a small data set stored for you on the server. This data set has information about the ages of all U.S. Presidents when they took office. We'll use this data set in an upcoming lecture. The name of the file is prezages. It is stored here on the server and also on the class web site under Datasets.

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

1. Highlight File and click on Open Datafile

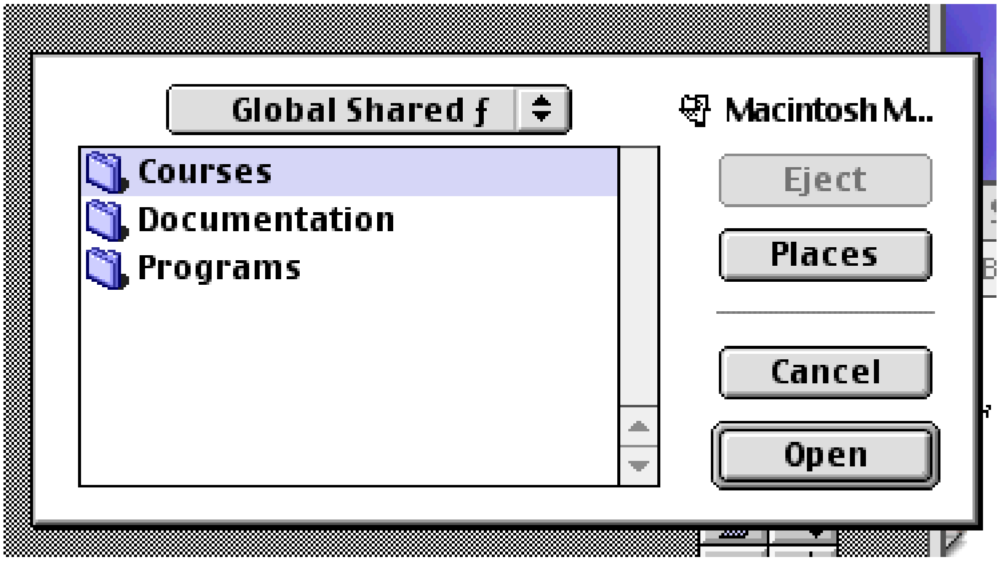


The following window opens (instead of cochran it will show your BOL ID):

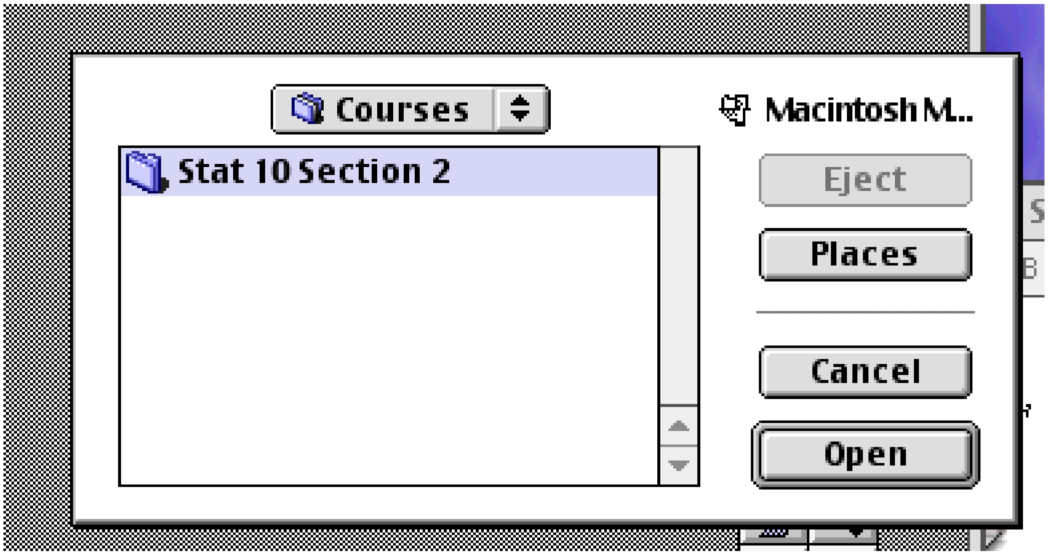


2. Click on Global Shared f

Now you can see the folders in this window:



3. Click on Courses, and select this Stat 10 course

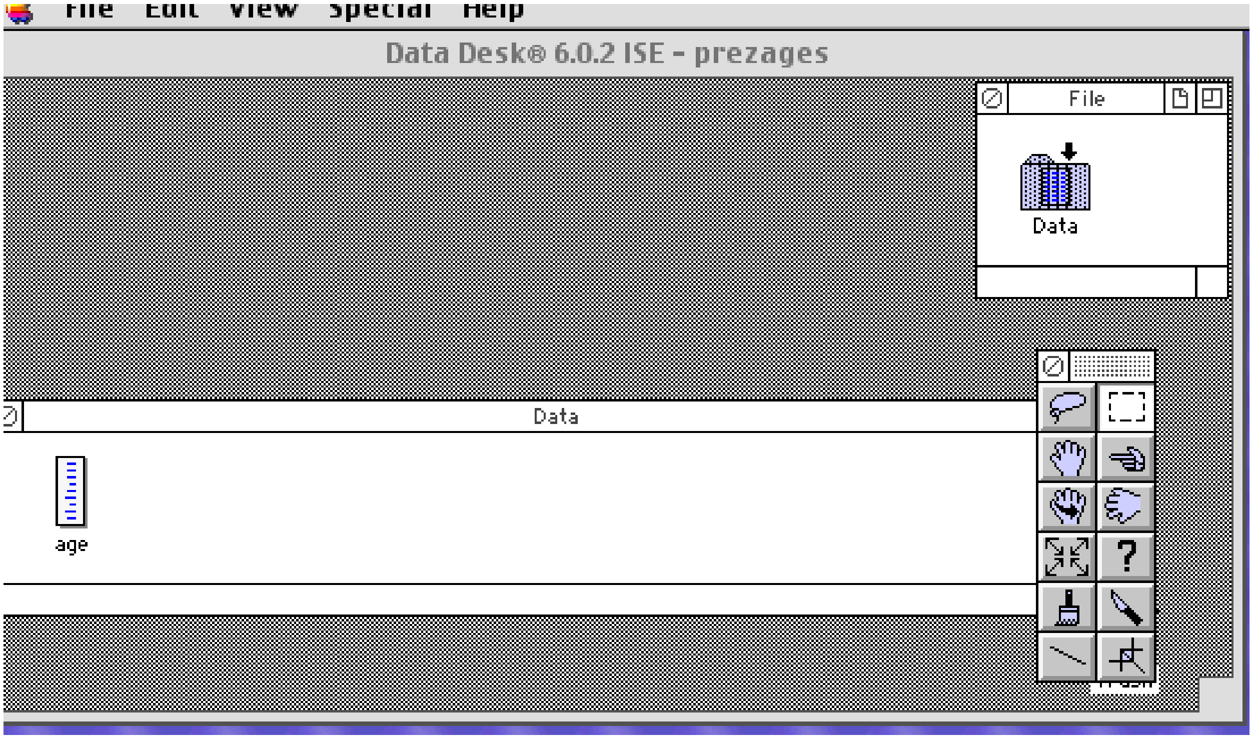


4. Click on prezages

The Mac will tell you that the file is locked (that means “read-only”). You can open it, use it, save it to your own folder but you cannot change it.

5. Click OK

Up will come the following screen:

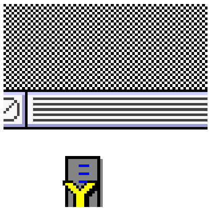


Notice several things:

- There is an icon age which is a folder that contains the data
- The icon is located in the data folder
- The data folder is also represented as an icon in the File Cabinet

6. Click on the age icon

A Y will appear on the box:



This means that DataDesk has assigned age to be a dependent variable for analyses--this will be covered later in the lab manual.

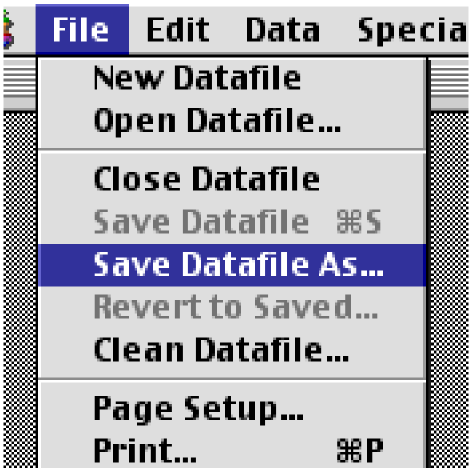
7. Click twice again on the age icon

If all the ages do not show, click the icon in the upper right corner of the Age screen. That will resize the box.

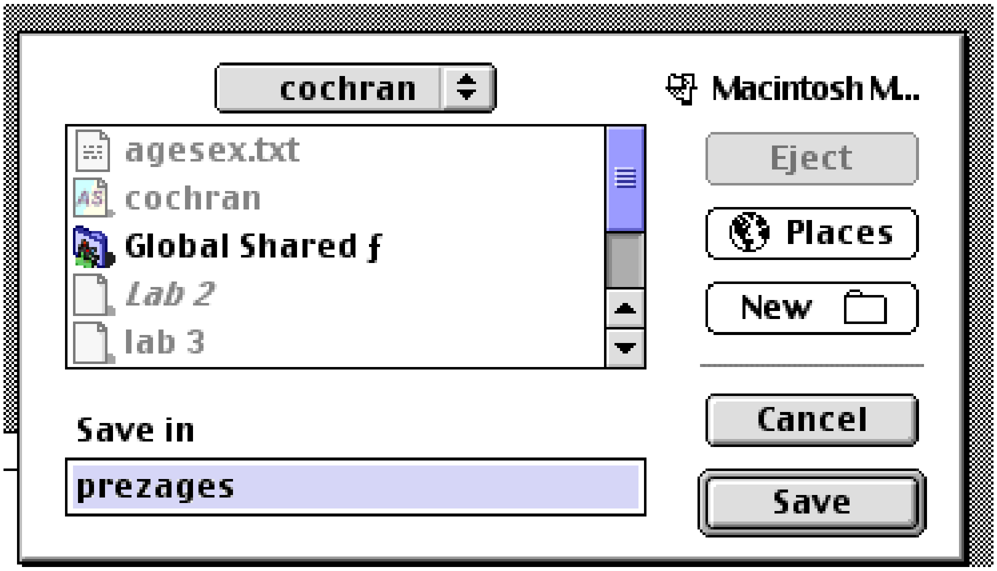
Save an existing data set to your folder

Save this dataset to your personal folder by:

1. Choosing the Save Datafile As... option under File



This window will appear (cochran shows here but on your machine it will be your BOL ID):



Save in

prezages

3. Click on Save

Now if you minimize DataDesk (click the little icon with the slit at the top right of the screen), you can see a new icon call prezages in your folder.

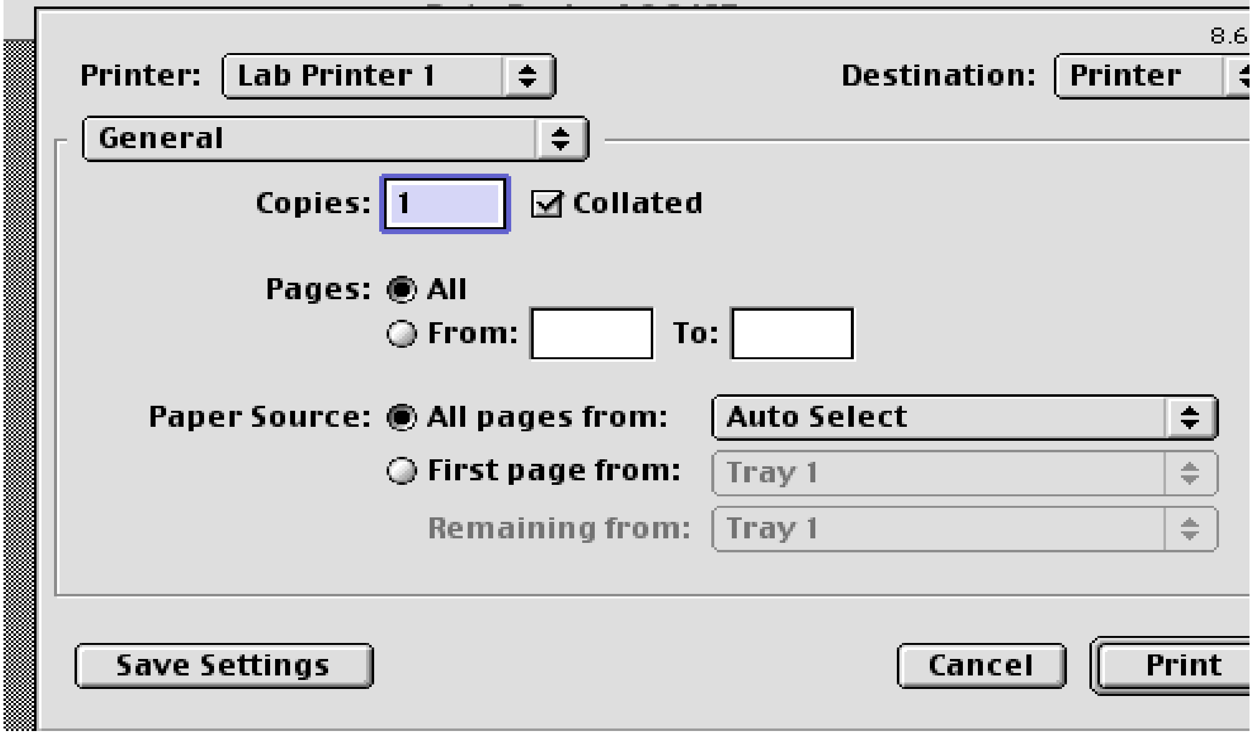
Produce printed output

You can also print out this data set.

1. Make sure that the age icon in the Data window has a Y on it (click on it if you need to)

2. Select the Print option under File in DataDesk

The following menu appears:



Printer: Lab Printer 1

Destination: Printer

General

Copies: 1 ☑ Collated

Pages: ● All

○ From: To:

Paper Source: ● All pages from: Auto Select

○ First page from: Tray 1

Remaining from: Tray 1

3. Click on Print

What prints out is the list of Presidents' ages.

Lab 2: Creating a Variable

Lab 2's objectives

* Create a new variable
* Examine summary statistics
* Save your work to your personal folder

Create a new variable

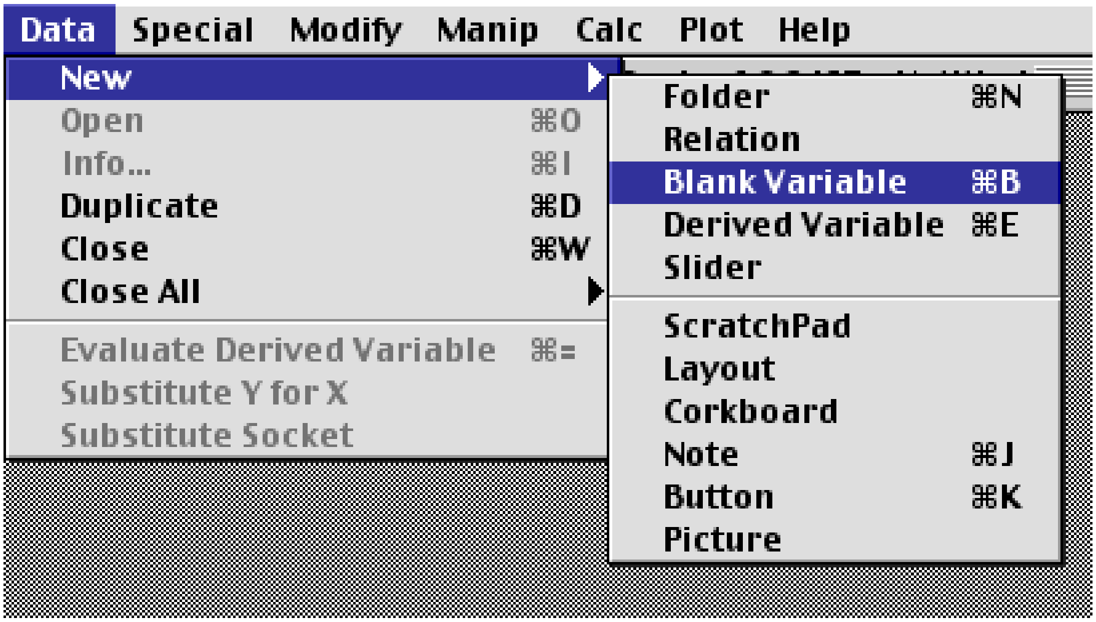
1. Open the Items for Student icon (see Lab 1, p. 7 if you don't remember)
2. Click on to the DataDesk icon

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:

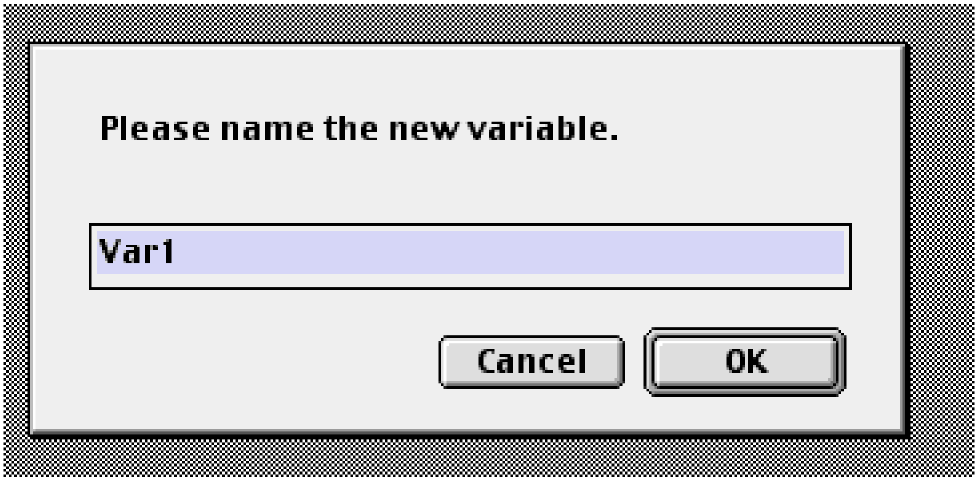
"Find the average and SD of the list 41, 48, 50, 50, 54, 57."

So, you have one variable that contains 6 elements or cases. 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:

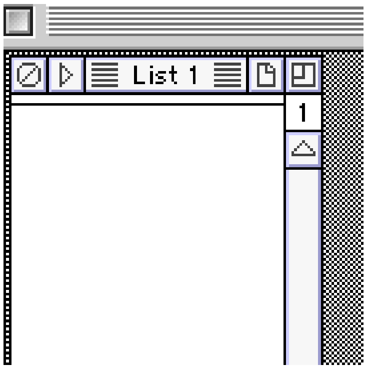

Please name the new variable.

Var1

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:

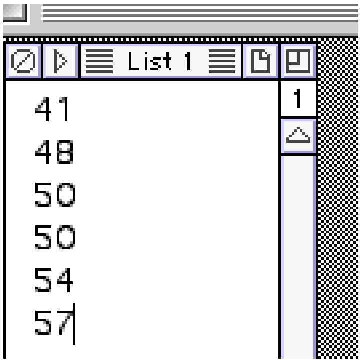


This is the window where you will store your 6 elements. At the top is the name of the variable, List 1. The blinking cursor is waiting for you to enter the information.

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:



41
48
50
50
54
57

6. Close the window by clicking on the circular icon in the left side of the top bar where the variable name, List 1, appears

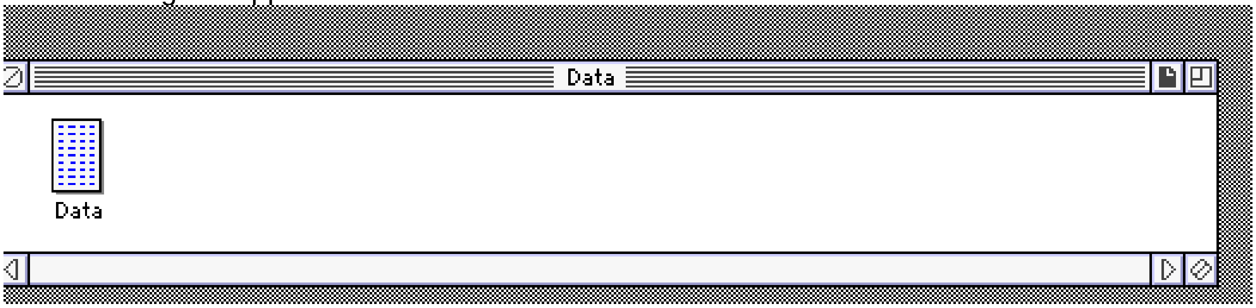
The information has returned to the icon File (the file cabinet) on the right side of the screen in the Data container.

Let's look at it.

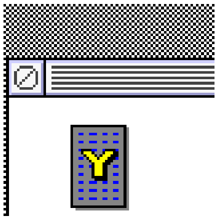
7. Click twice on File if Data isn’t showing

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

The following will appear:

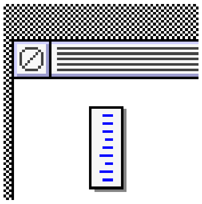


9. Click once on the Data icon in the Data window and a Y will appear

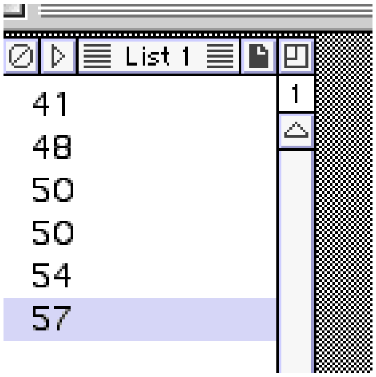


The Y refers to a dependent variable. You have only one variable (a list of 6 numbers) and so it is your dependent or outcome variable. If you had two or more, you could assign variables to be independent (X) or dependent (Y). This will be covered later in the course.

10. Click twice more and the variable List 1 appears as an icon



11. Click twice more on List 1 and the 6 values you entered appear as shown below



List 1

```
41
48
50
50
54
57
```

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

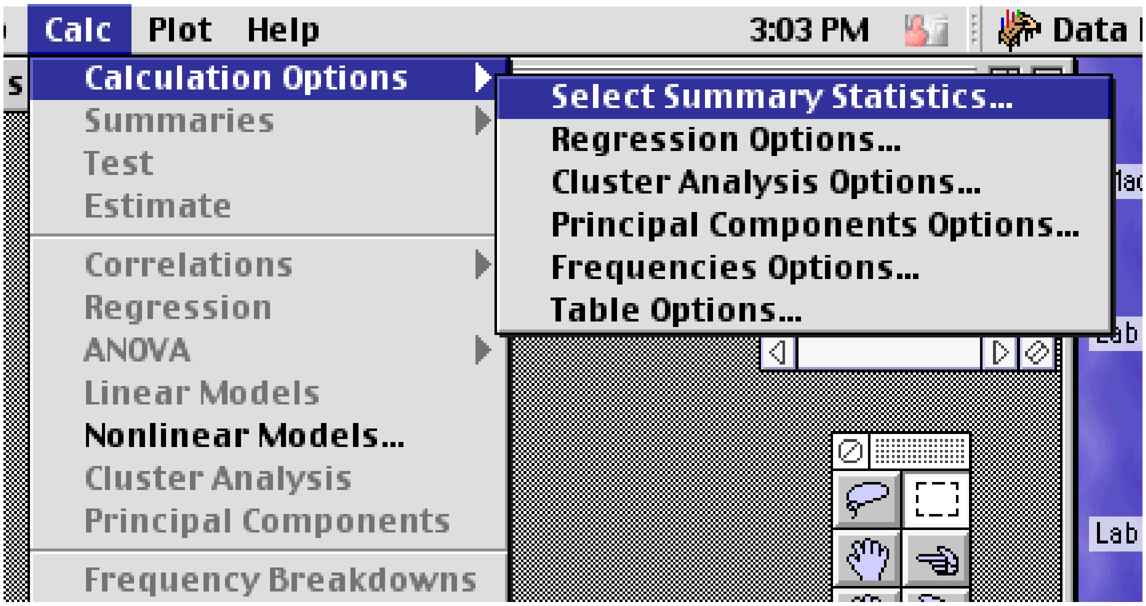
You have now created a new variable.

Examine summary statistics

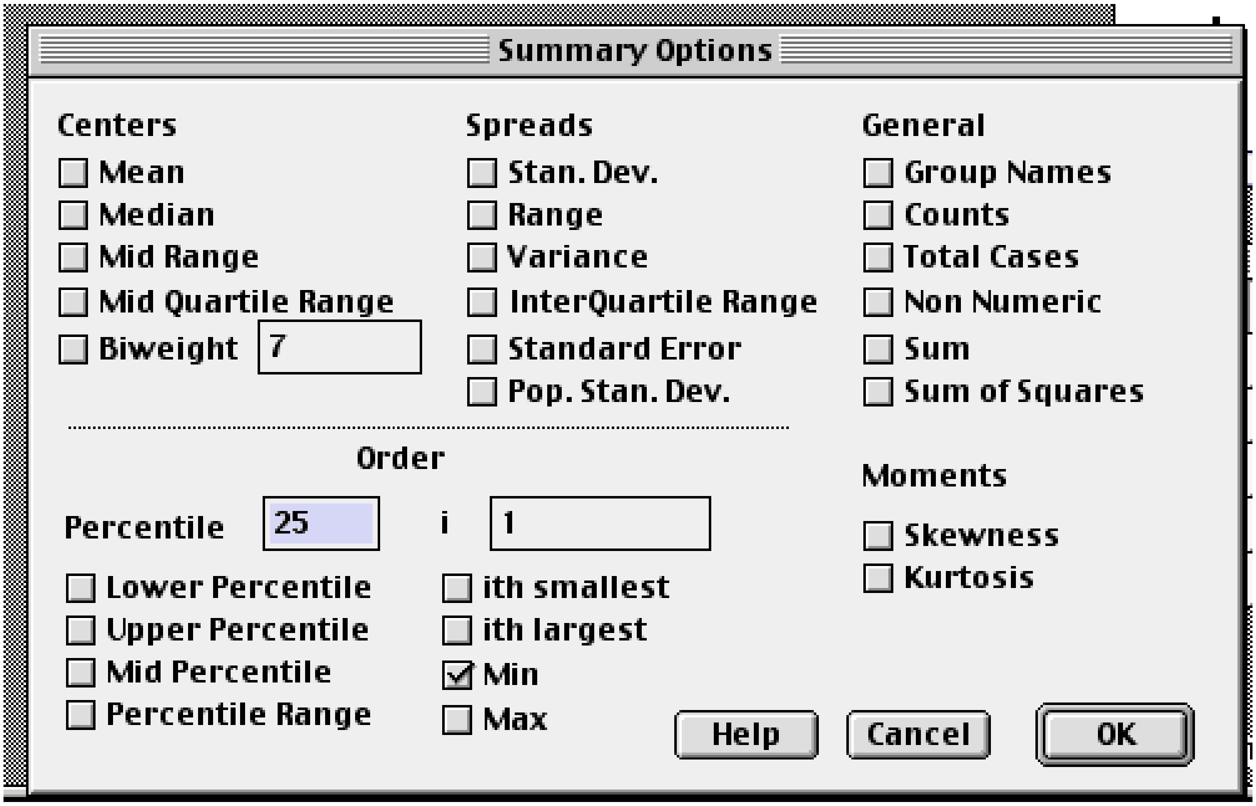
Now you are going to calculate the summary statistics. Summary statistics is another word for descriptive statistics. The first step is to tell the program what statistics you want.

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:



Summary Options

Centers

□ Mean
□ Median
□ Mid Range
□ Mid Quartile Range
□ Biweight 7

Spreads

□ Stan. Dev.
□ Range
□ Variance
□ InterQuartile Range
□ Standard Error
□ Pop. Stan. Dev.

General

□ Group Names
□ Counts
□ Total Cases
□ Non Numeric
□ Sum
□ Sum of Squares

Order

Percentile 25 i 1

□ Lower Percentile
□ Upper Percentile
□ Mid Percentile
□ Percentile Range

□ ith smallest
□ ith largest
☑ Min
□ Max

Moments

□ Skewness
□ Kurtosis

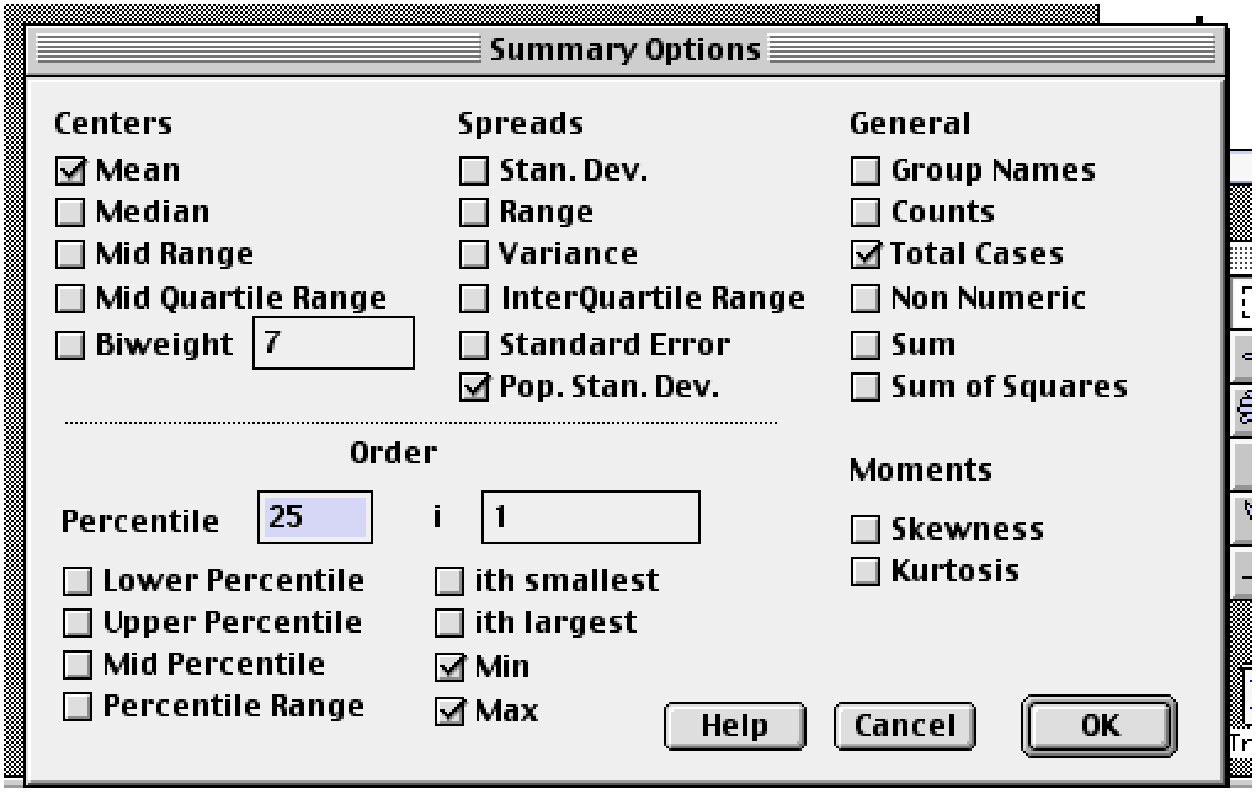
The program is going to very quickly calculate for you whatever you want. In this instance you want the mean and Standard Deviation.

Notice that the program offers you 5 choices of summary statistics: Centers, Spreads, 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 we 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



Summary Options

Centers

☑ Mean
□ Median
□ Mid Range
□ Mid Quartile Range
□ Biweight 7

Spreads

□ Stan. Dev.
□ Range
□ Variance
□ InterQuartile Range
□ Standard Error
☑ Pop. Stan. Dev.

General

□ Group Names
□ Counts
☑ Total Cases
□ Non Numeric
□ Sum
□ Sum of Squares

Order

Percentile 25 i 1

□ Lower Percentile
□ Upper Percentile
□ Mid Percentile
□ Percentile Range

□ ith smallest
□ ith largest
☑ Min
☑ Max

Moments

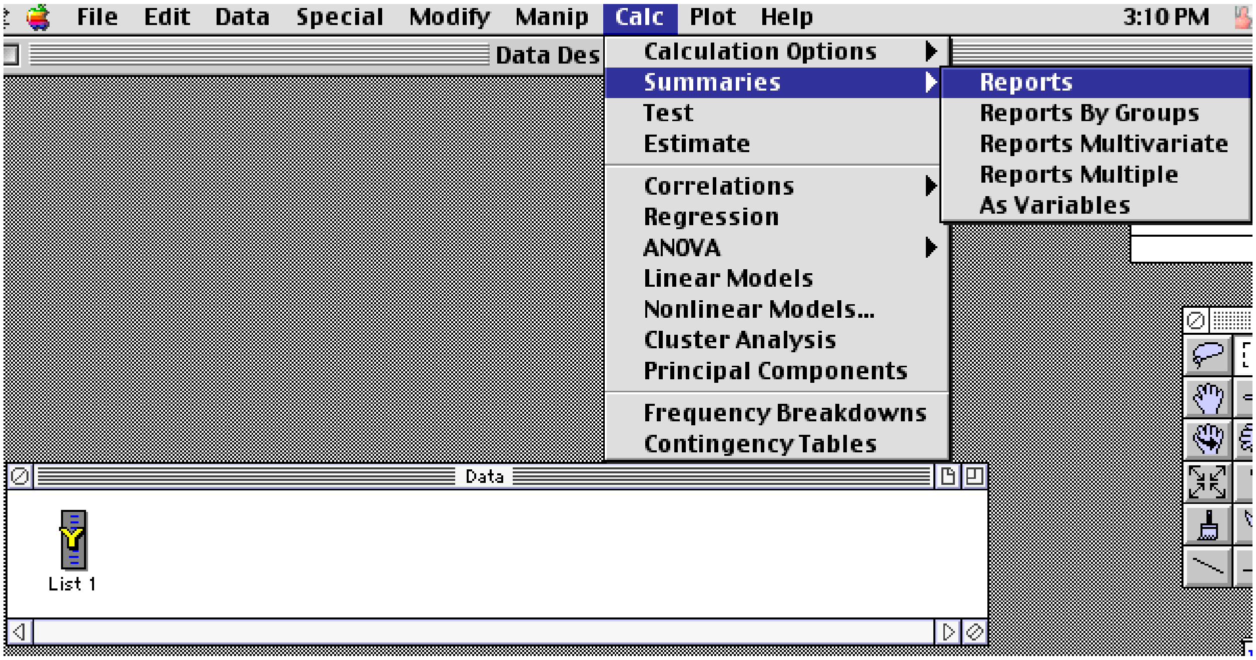
□ Skewness
□ Kurtosis

3. Click OK

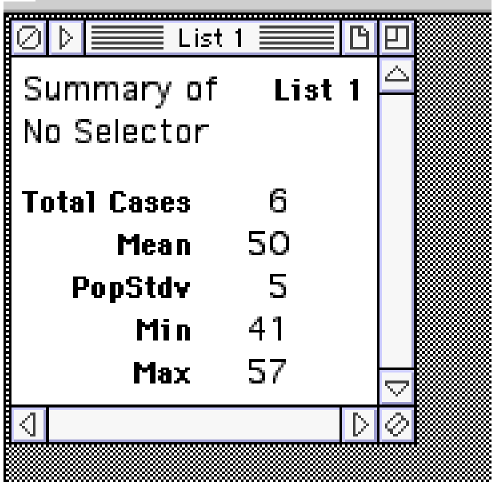
4. Assign List 1 to the be the dependent or Y variable (see Lab 2, p. 21)

5. Now calculate the values:

   Highlight Calc
   Highlight Summaries
   Choose Reports



Now the statistics you desire will appear. (If you cannot select Summaries on the screen, then the variable, List 1, has not been selected for analysis--it doesn't have a Y on it. See Lab 2, p. 21 for instructions on how to select a variable for analysis.)



```
Summary of     List 1
No Selector

Total Cases     6
       Mean    50
    PopStdv     5
        Min    41
        Max    57
```

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

These results are stored in the Results icon in the File cabinet. You can click twice on the Results icon. You will see an icon, List 1. If you close the results on your screen and then click on List 1 the same "output" will appear.

Specify a single variable for analysis

To repeat how to select a single variable for analysis:

1. Click twice on File if need be
2. Click twice on the Data icon to open the data window
3. Click once on Data in the Data 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

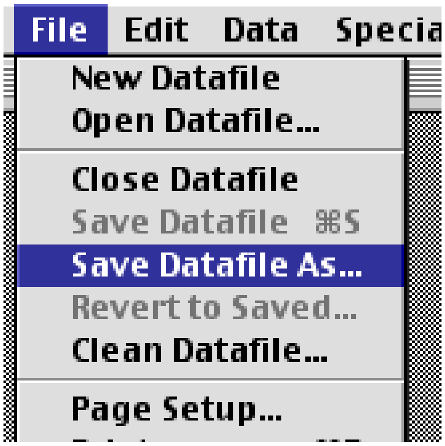
Save your work to your personal folder

Now it's time to save your work.

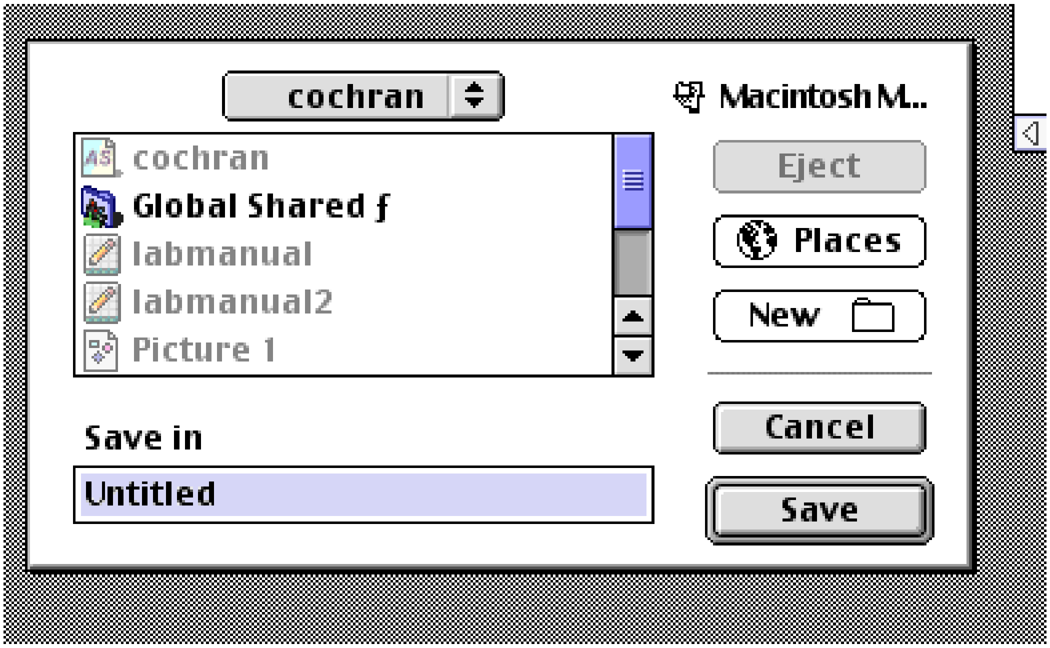
1. To save a data file:

   Highlight File

   Select Save Datafile As



2. Edit the Untitled name to a name for your file (such as, for example, Lab 2)



Save in

Untitled

3. Choose Save

Now you should see the new icon in your personal folder.

Lab 3: Making Histograms

Lab 3's objectives

* Import a data file into DataDesk
* 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. These files often have a .txt extension to the file name. Text files are generic and can be read easily across computer operating systems (like Mac and Windows) 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 in the Global Shared Files Folder on the Stat Lab server and also on the web at the class web site.

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. After that, each row contains 1 element or case. 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.

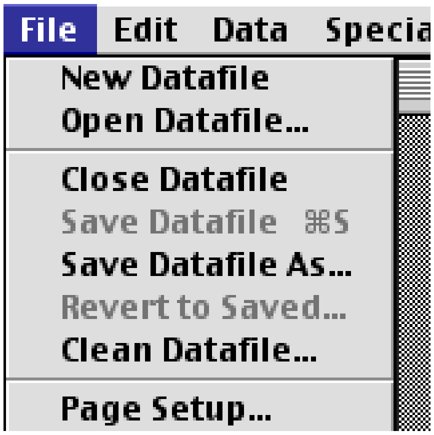
The actual character used to separate chunks of information is called a **delimiter**.

Now you are ready to process the external data.

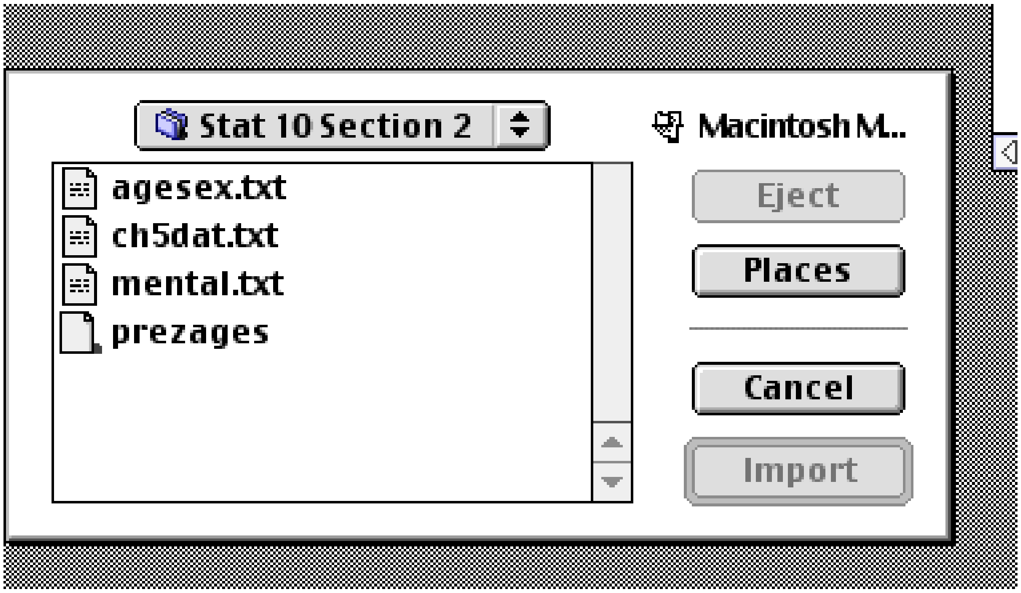
1. Open DataDesk (see Lab 1, p. 11)

2. Begin the import procedure:

   Highlight File
   Select Import



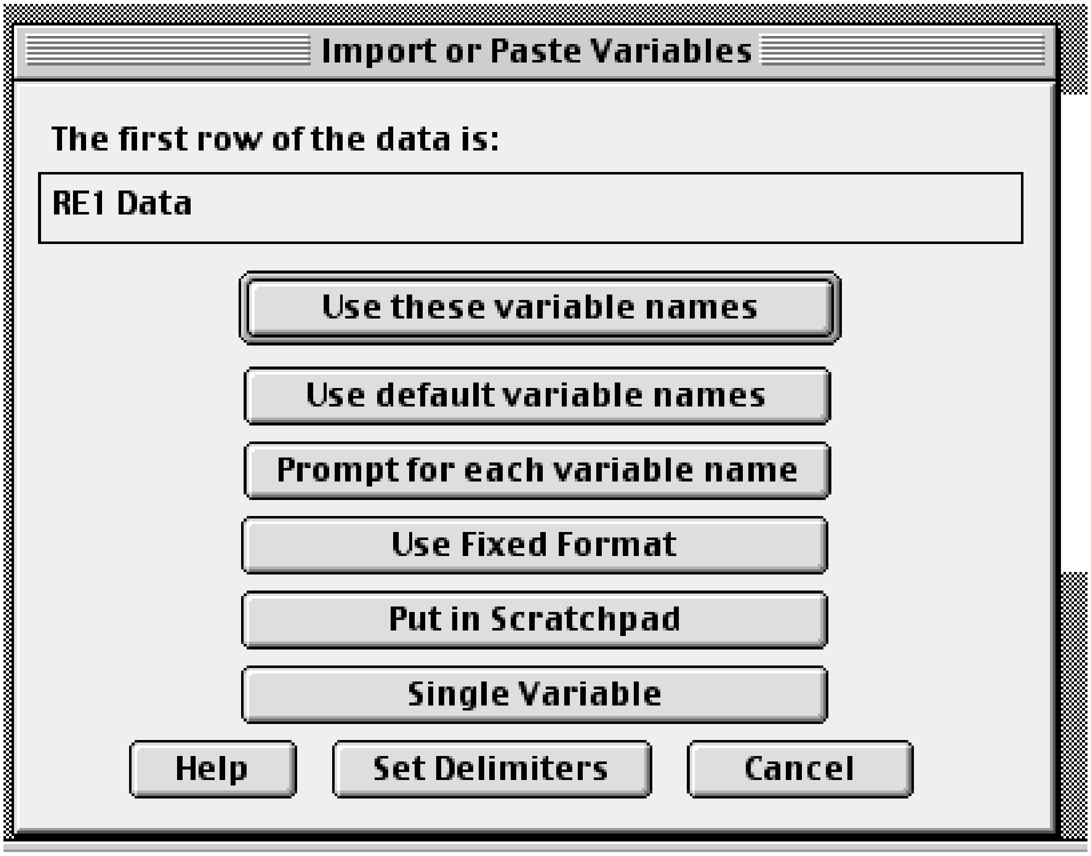
3. Select the Global Shared files folder (see Lab 1, p. 12)
4. Select Courses
5. Select this course



6. Select the dataset, ch5dat.txt and click Import

7. The computer will tell you the file is Locked. Click OK

The following window will appear:



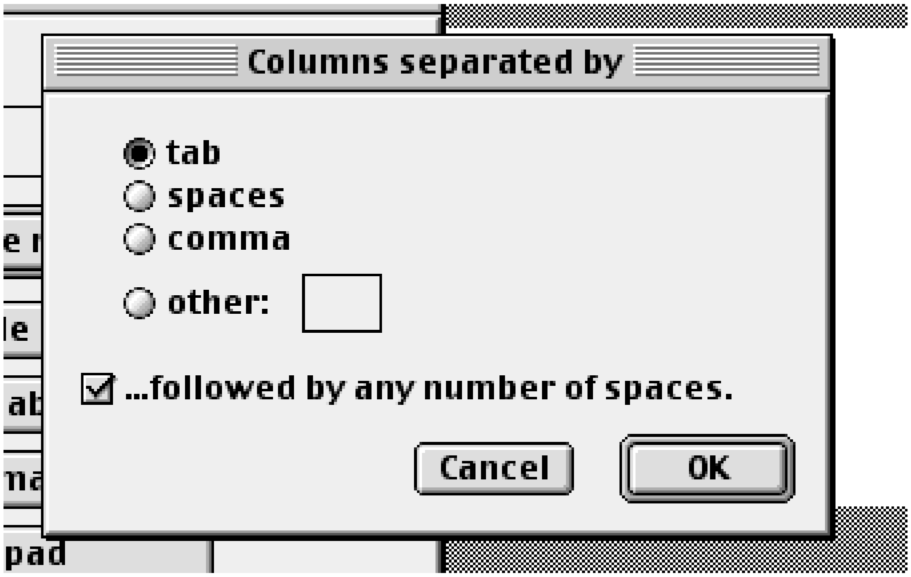
The first row of the data is:

RE1 Data

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

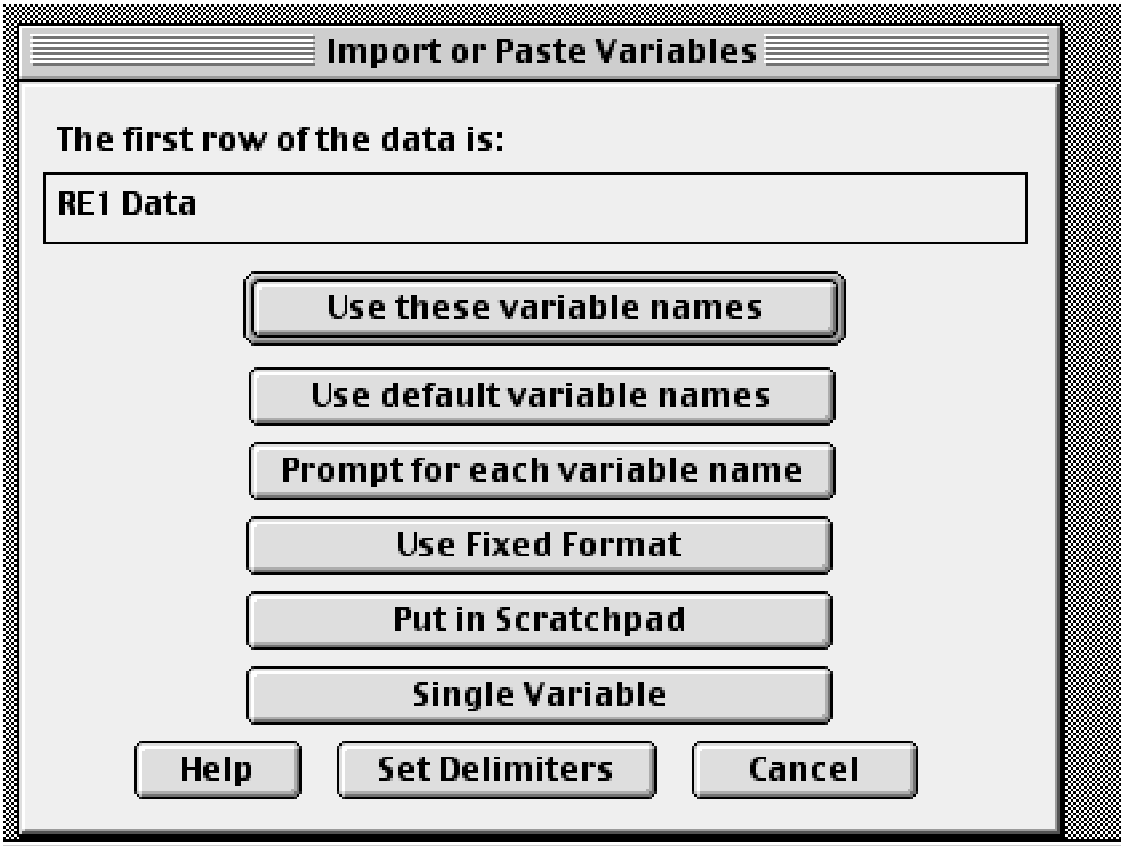
The following window will appear:



● tab
○ spaces
○ comma
○ other: □

☑ ...followed by any number of spaces.

8. Be sure that tab is selected and click OK

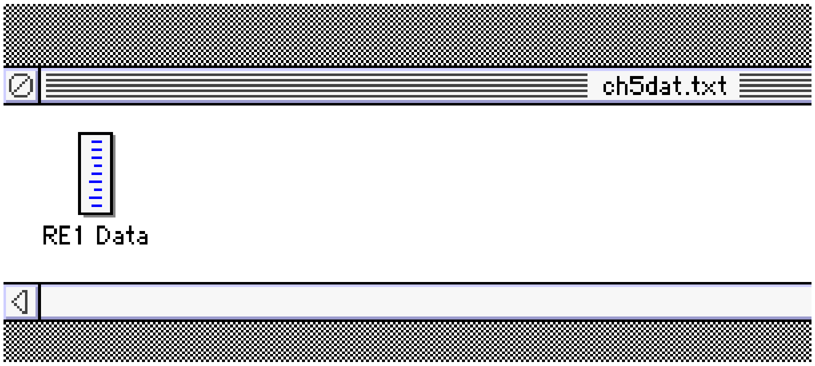


The first row of the data is: RE1 Data

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.

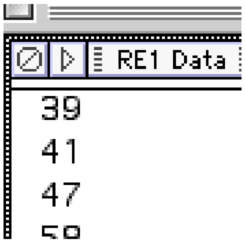
9. Select Use these variable names

Now the following window will appear:



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

It should look like this:



39
41
47

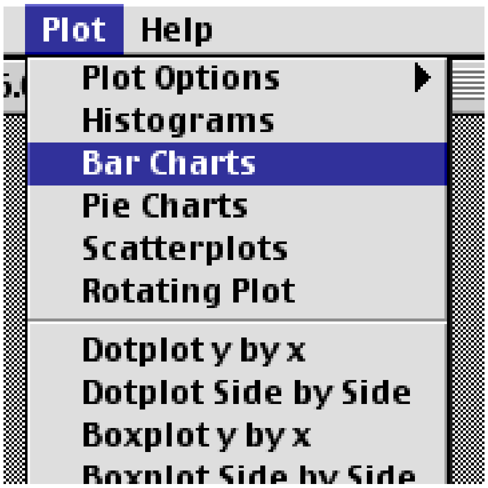
11. Save the data to a file called Lab 3 in your BOL ID folder (see Lab 1, p. 15)

Create a bar chart

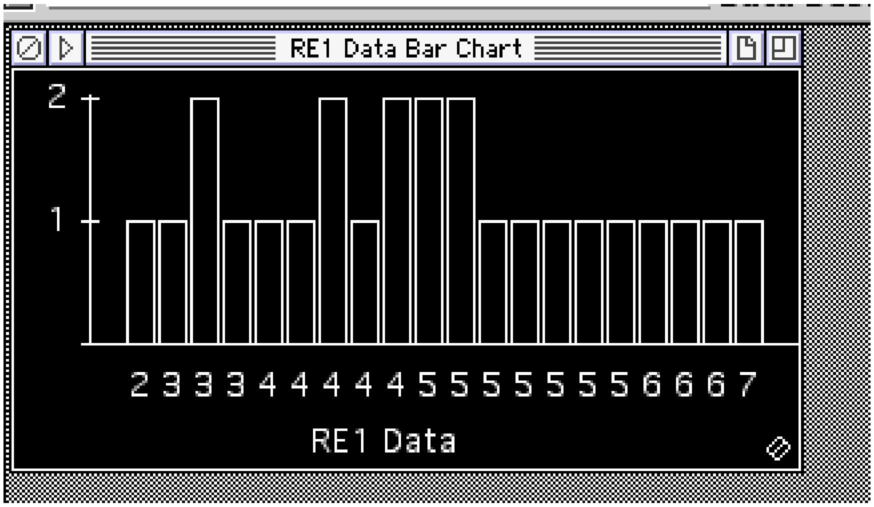
DataDesk offers many types of data plotting. The simplest is a bar chart, which is just a count along the left axis and the actual values observed along the horizontal axis.

1. Click on RE1 Data so that a Y appears on the icon (see Lab 2, p. 27)

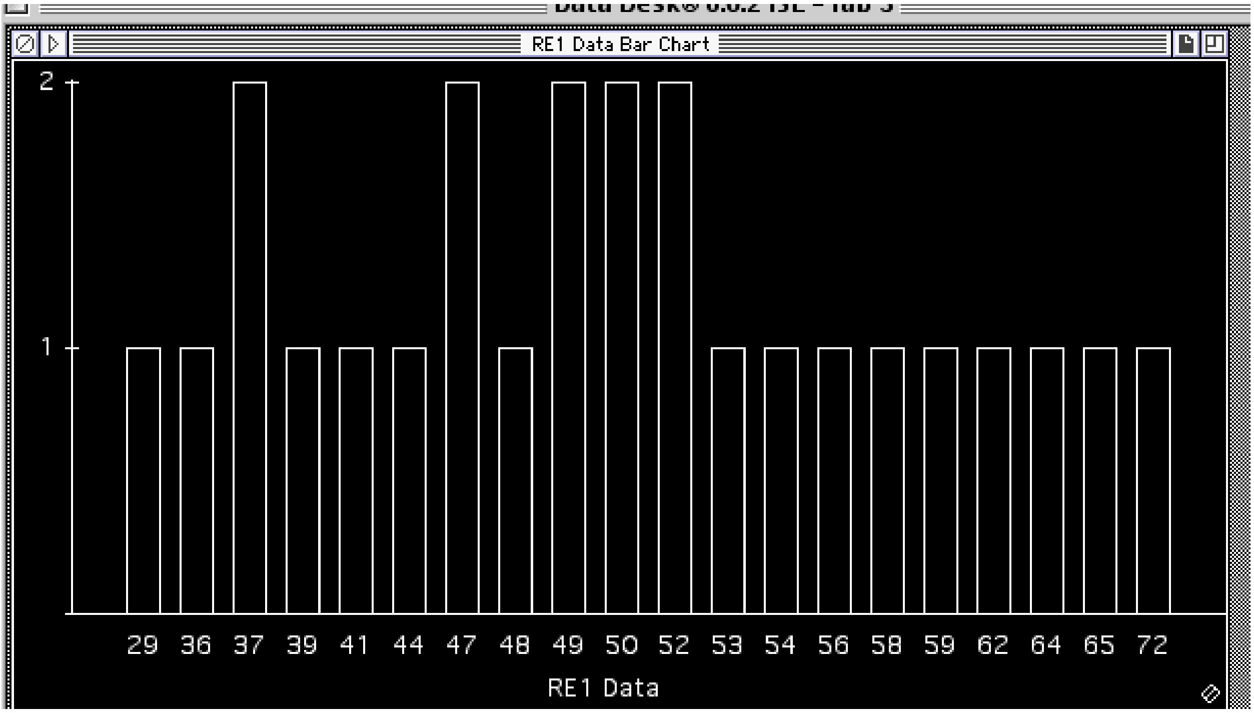
2. Highlight Plot and select Bar Charts



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



You can make the chart prettier by clicking the right most icon located on the RE1 Data Bar Chart line. Now the bar chart looks like this:



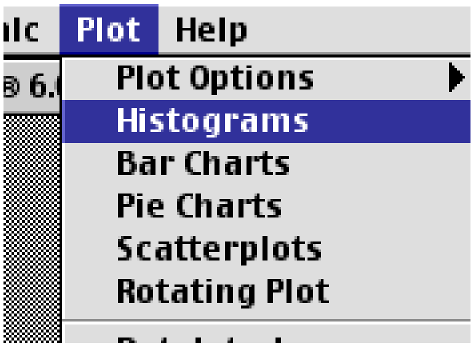
Notice the vertical axis is the count. The horizontal axis shows the values observed. Many values are missing because no one got that test score. This is not a histogram (why is that?).

Create a histogram

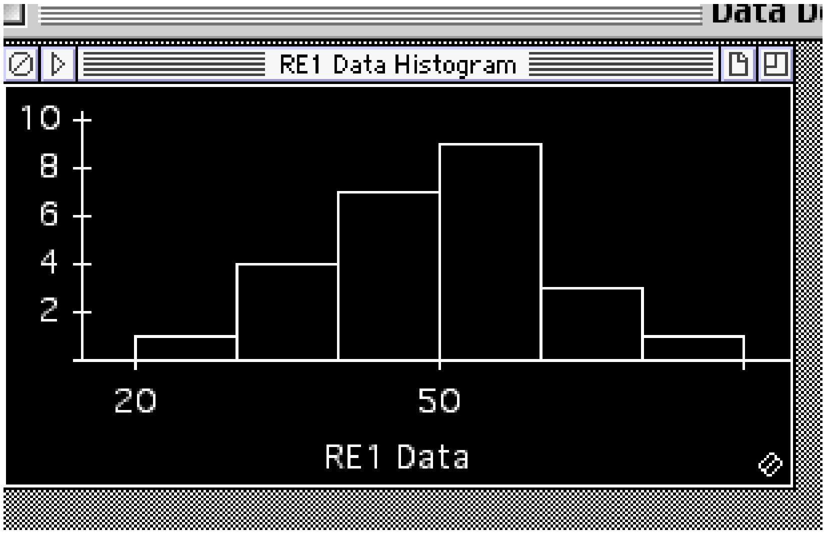
You can also create a histogram.

1. Assign RE1 Data to be the Y variable (see Lab 2, p. 27)

2. Highlight Plot and select Histograms



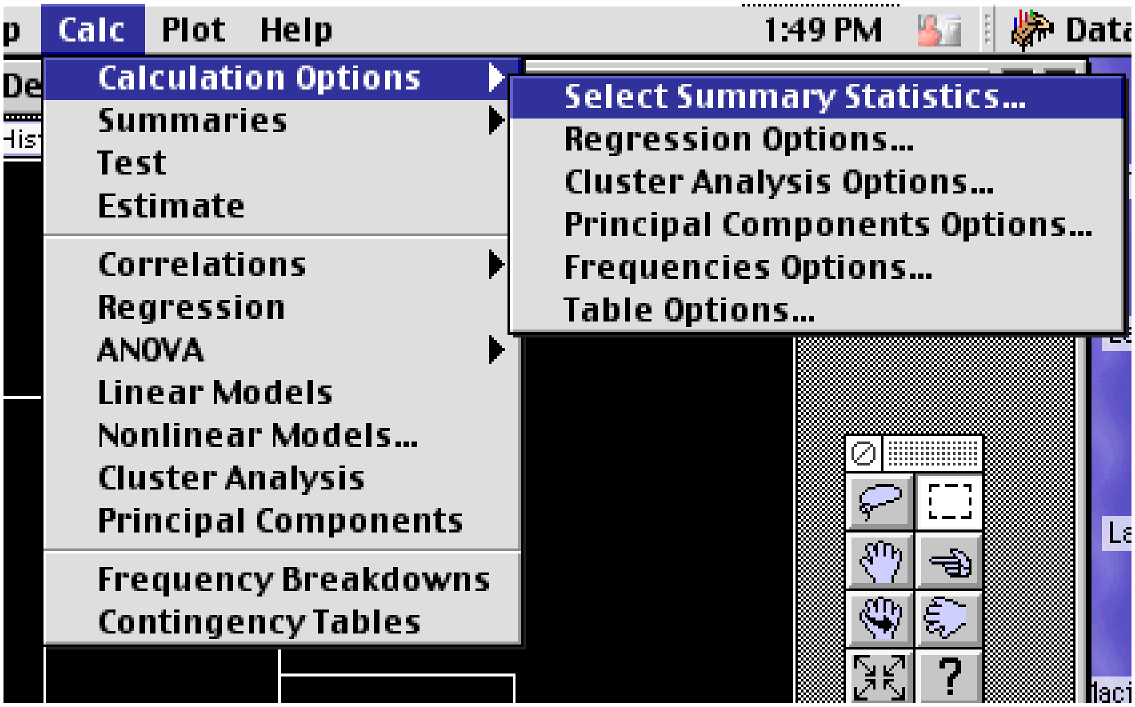
The following will appear (you can hit the icon in the upper right corner to expand the graph and make it look better):



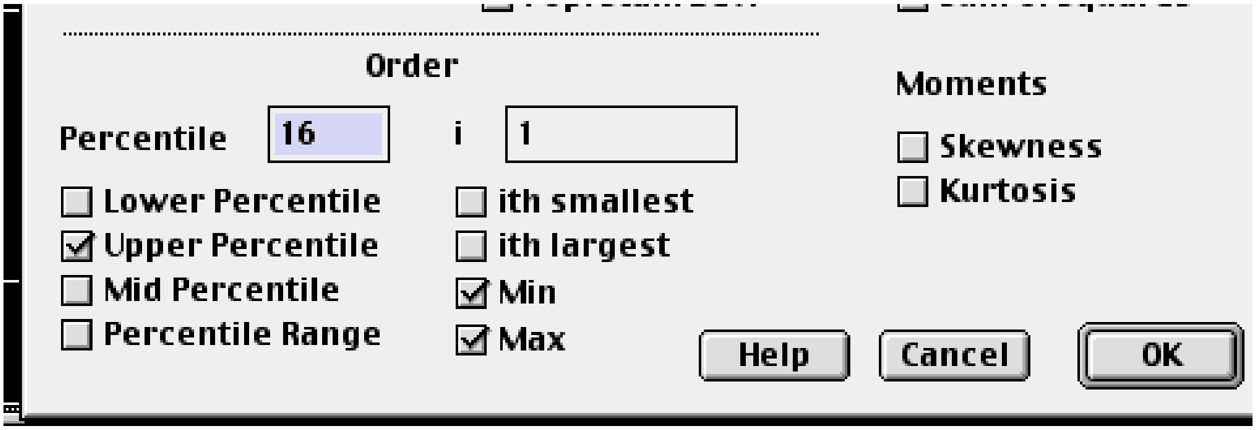
Notice that the area in the rectangles sums to (10*1 + 10*4 + 10*7 + 10*9 + 10*3 + 10*1 =) 250. So, you would estimate that 40/250 or 16% of the elements are test scores of 60 or more (if you use the left-hand convention for discrete numbers from your textbook; 60 would represent the limit. But DataDesk 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:

3. Use the Calc function (See Lab 2, p. 22)
4. Highlight Calculation Options and select the Select Summary statistics



5. Edit this to show what corresponds to the top 16% (See Lab 2, pp. 24-25). (The request is made in the Order category. The value will be associated with the upper 16th percentile.)



Order

Percentile 16 i 1

□ Lower Percentile
☑ Upper Percentile
□ Mid Percentile
□ Percentile Range

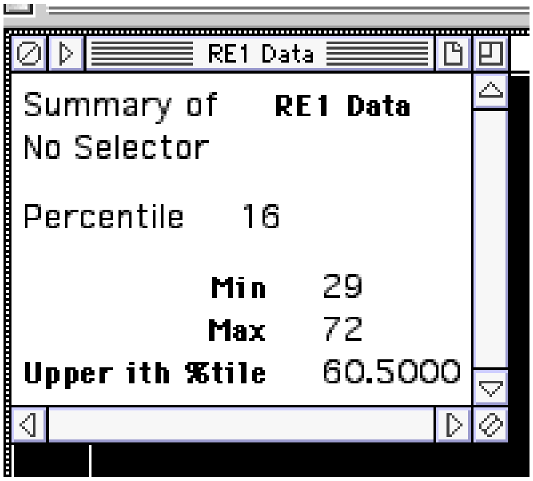
□ ith smallest
□ ith largest
☑ Min
☑ Max

Moments

□ Skewness
□ Kurtosis

6. Ask for the report of summary statistics (see Lab 2, p. 26)

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



```
Summary of      RE1 Data
No Selector

Percentile    16

               Min     29
               Max     72
Upper ith %tile        60.5000
```

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.

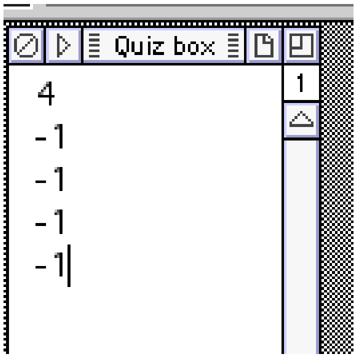
The box you are going to make comes from Chapter 16, Review Exercise 7:

"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 DataDesk (see Lab 1, p. 11)

2. Create a blank variable called Quiz box (or whatever name you want) with information about the box (see Lab 2, p. 18)



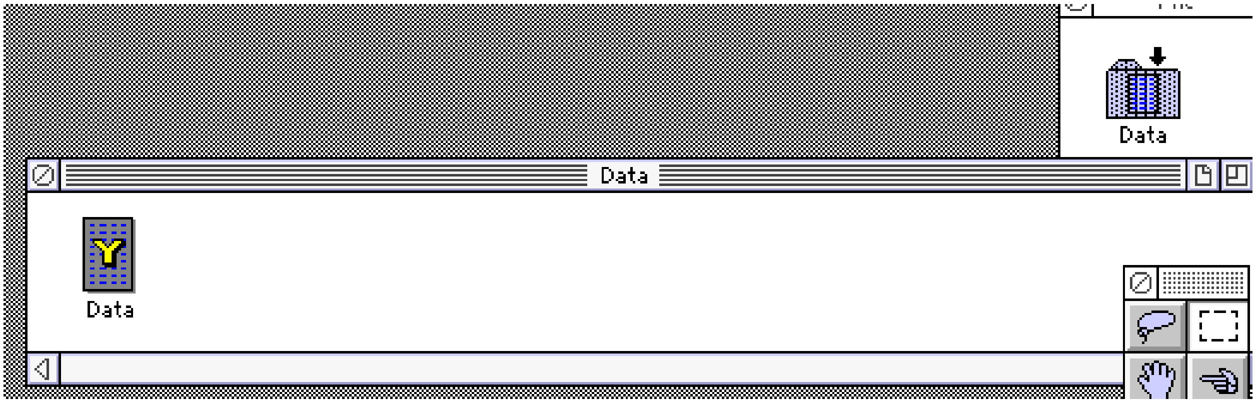
```
Quiz box
4
-1
-1
-1
-1
```

Notice, you enter all 5 outcomes which are {4, -1, -1, -1, -1}...and no more!

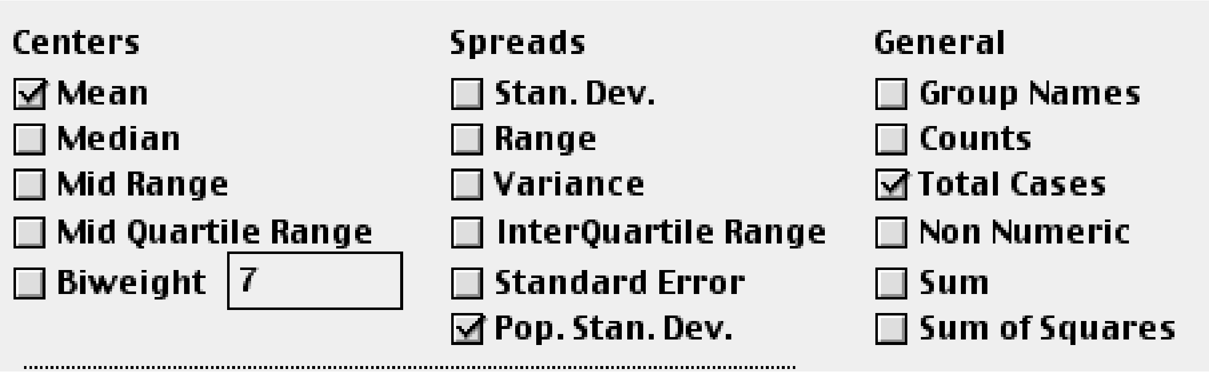
Generate statistics about the box

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

1. Click on your data icon and then click on Quiz box so that the Y appears (see Lab 2, p. 27)



2. Highlight Calc, and Calculation options, and select the Summary statistics you want (in this instance, mean and population SD) (see Lab 2, p. 23)



Centers

☑ Mean
☐ Median
☐ Mid Range
☐ Mid Quartile Range
☐ Biweight 7

Spreads

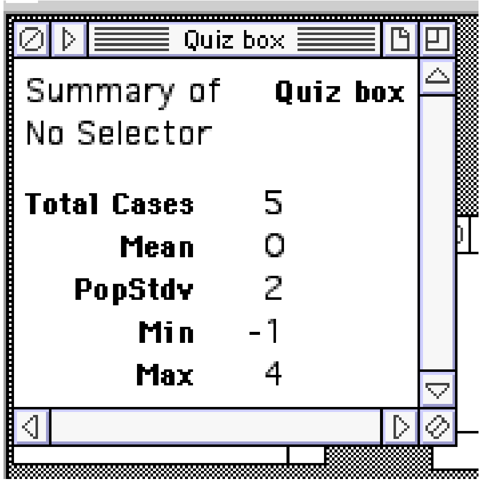
☐ Stan. Dev.
☐ Range
☐ Variance
☐ InterQuartile Range
☐ Standard Error
☑ Pop. Stan. Dev.

General

☐ Group Names
☐ Counts
☑ Total Cases
☐ Non Numeric
☐ Sum
☐ Sum of Squares

3. Highlight Calc, Summaries, and choose Reports (see Lab 2, p. 26)

The following window appears:



```
Summary of    Quiz box
No Selector

Total Cases     5
       Mean     0
    PopStdv     2
        Min    -1
        Max     4
```

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

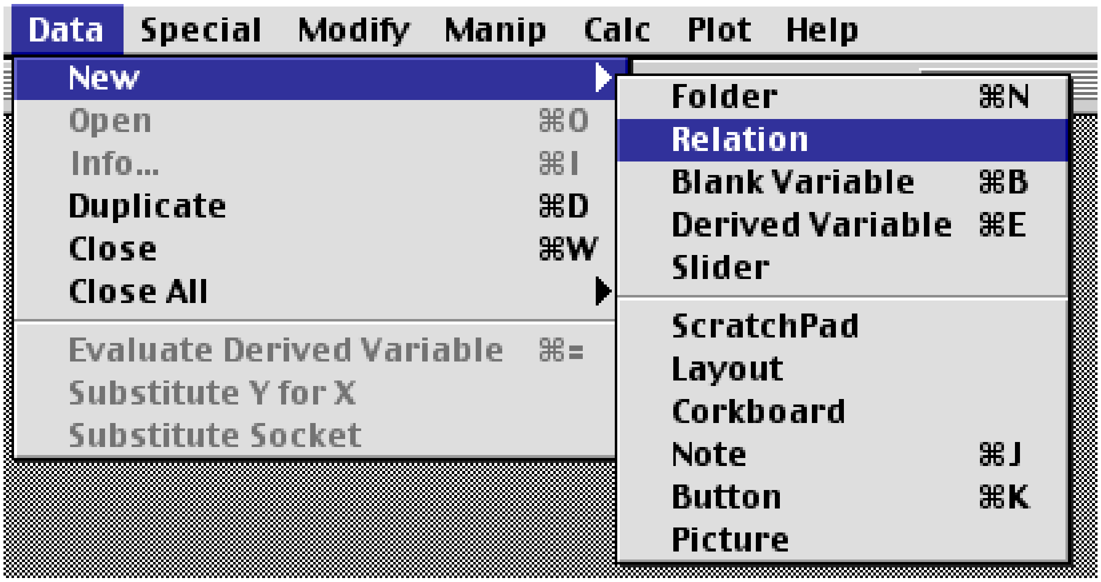
4. Save the variable to a file for later use (see Lab 2, p. 27)

Create a box using a relational database approach

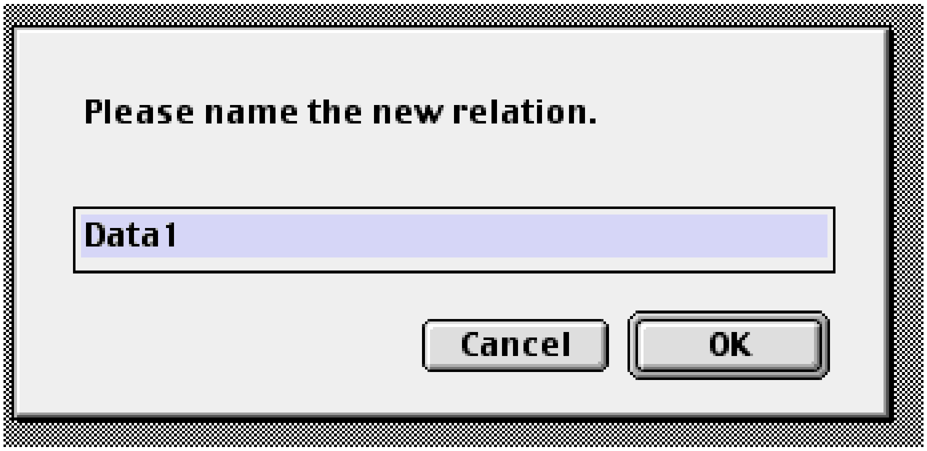
The box you just created can be thought of as resulting from two components. The first are the values (get 4 points, lose 1 point) and the second are the counts or numbers of times the values appear (1 chance of getting 4 points, 4 chances of losing 1 point). In data programming, it is often easier to 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 following window appears:

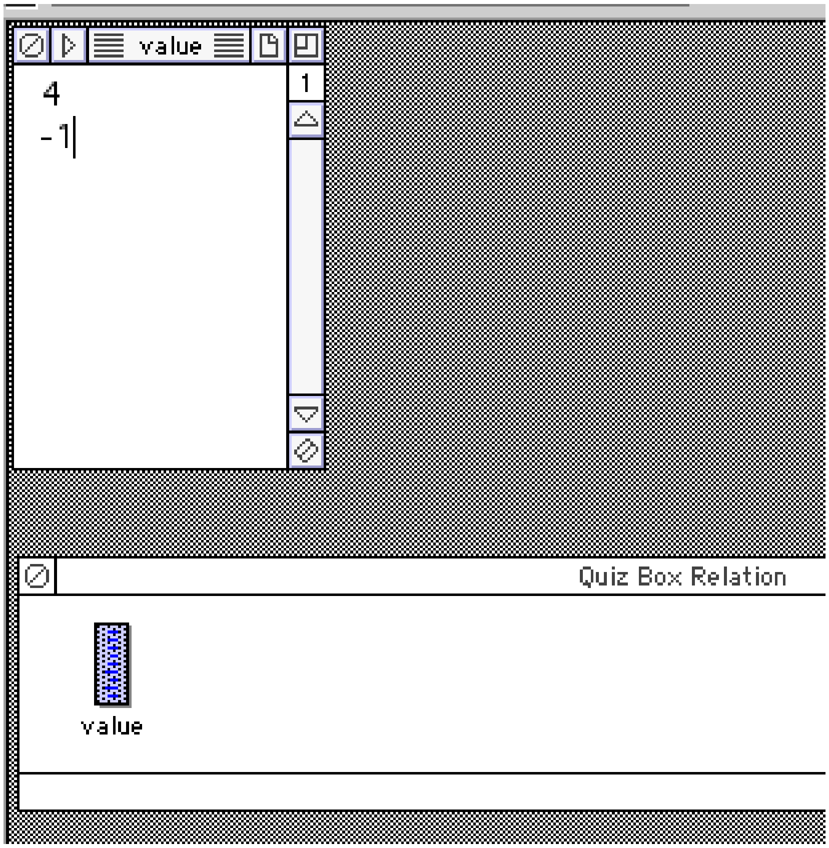

Please name the new relation.

Data1

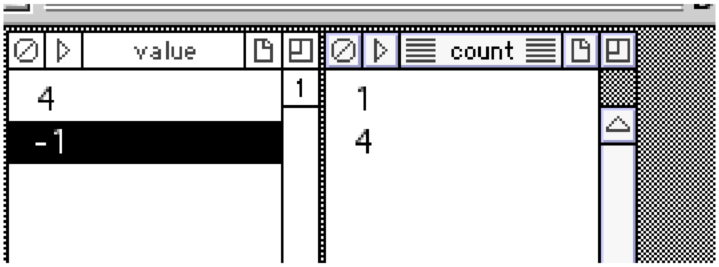
2. Edit Data1 to a new name, Quiz Box Relation, and click OK

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

3. Create a new blank variable called value with the two outcomes possible {4, -1} (see Lab 2, p. 18)



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

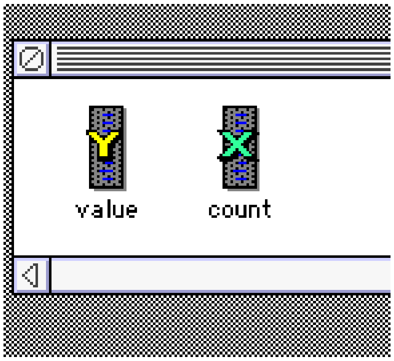


In doing so be very careful of two things:

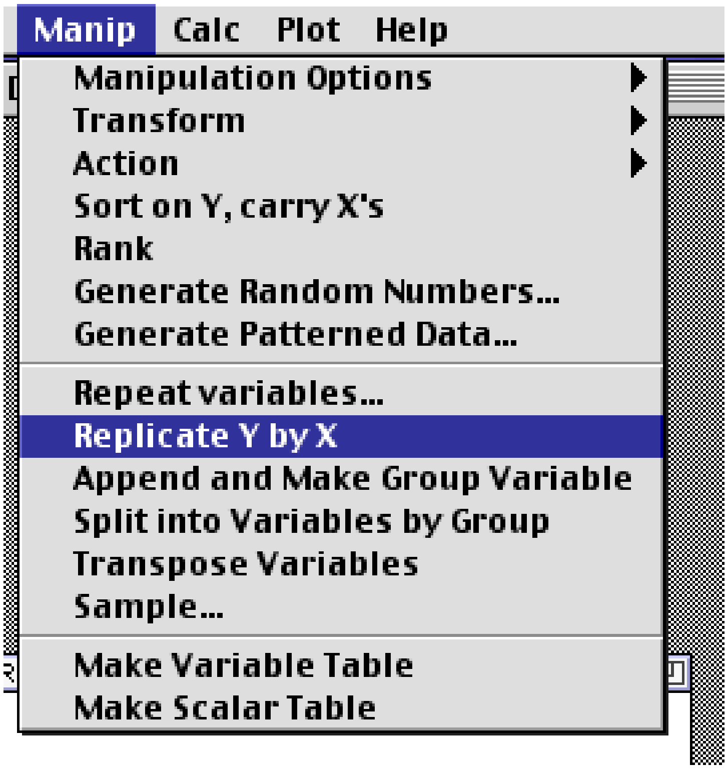
- Keep the order of entry accurate (e.g. the first row for both variables refers to 1 chance of earning 4 points, the second row to 4 chances of losing 1 point across both variables--order of input matters here!)
- Do not enter more elements than exist (for example, don't hit return after you enter the

second element or a blank 3rd element will be in your database)

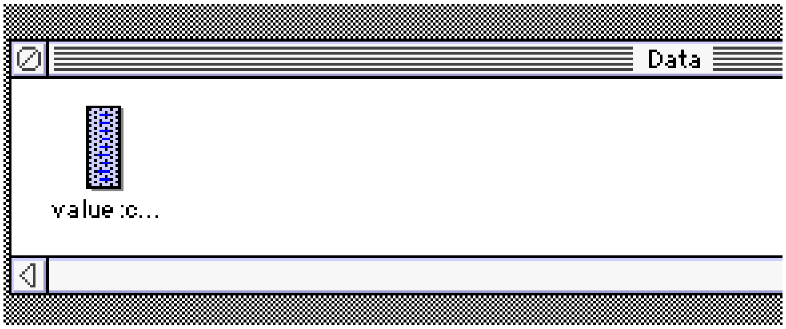
5. Designate the value variable as Y and the count variable as X by first clicking on Value and then, holding the shift key down, click on Count



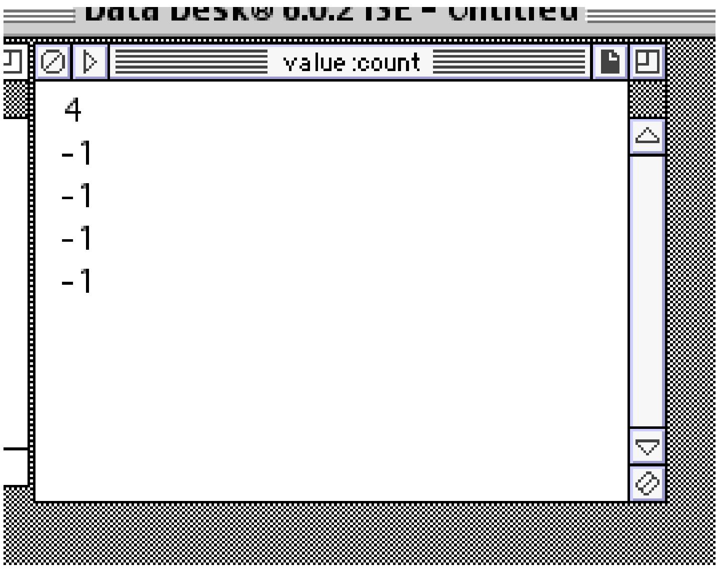
6. Now Replicate the elements in Value by the count in Count



The following will appear:



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



```
4
-1
-1
-1
-1
```

For this 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

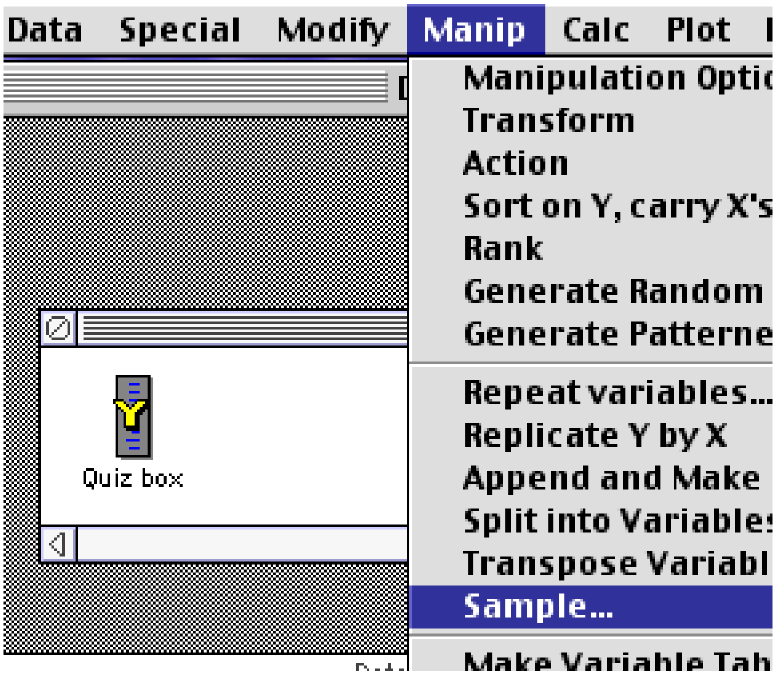
In your textbook, you learned that a student taking this exam of 25 questions and answering each one randomly would be expected to score 25 * mean of the box = 25 * 0 = 0 on the quiz. But this would vary, and you can estimate how much would be normal or chance variation if only chance is occurring. The expected variation from 0 for a total score on the quiz when answered at random would be the number of questions (25 in this instance) multiplied by the SD of the box (2 from your results shown on p. 41) So, you might guess a score of 0 plus or minus a chance (or SE of the sum) estimate of 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).

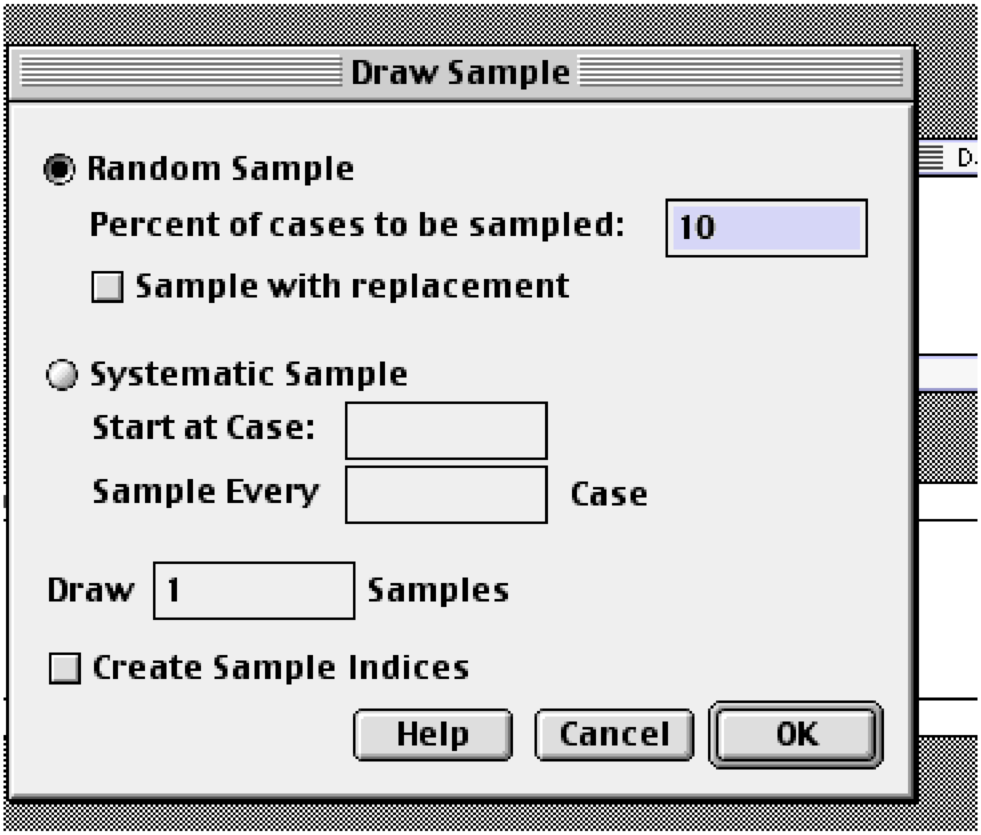
Well, let's see what happens when you actually do randomly sample 25 times. Now the output shown here will vary a little from what you observe because each time random sampling results in slightly different results (Why is that?).

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. 27)
3. Highlight Manip
4. Choose Sample



The following screen will appear:



Draw Sample

● Random Sample

Percent of cases to be sampled: 10

□ Sample with replacement

○ Systematic Sample

Start at Case:

Sample Every Case

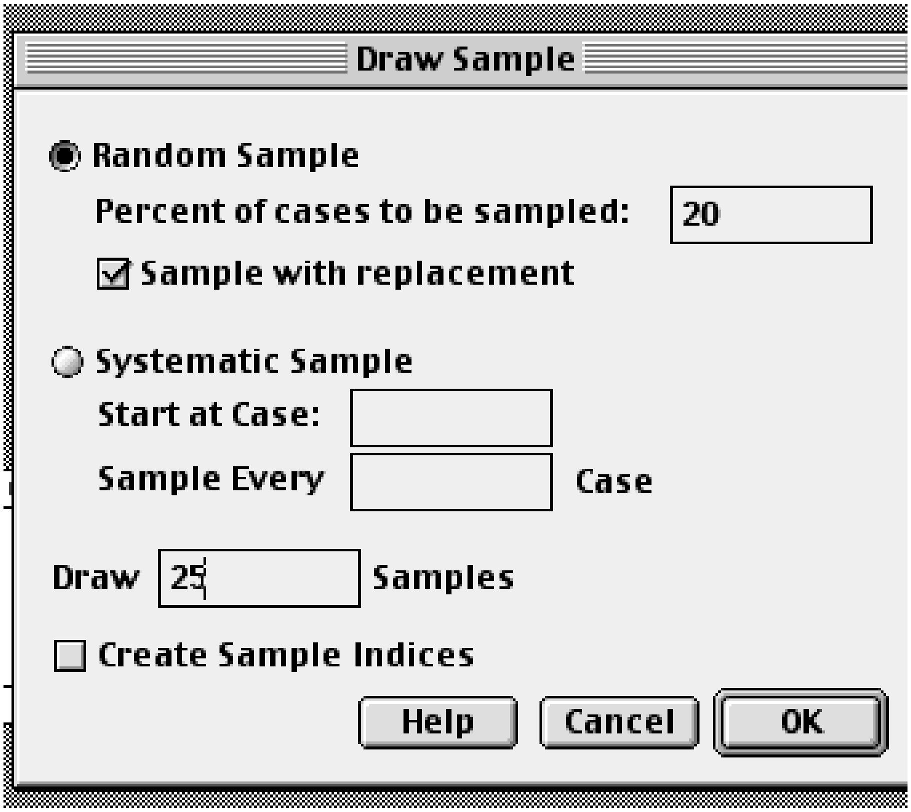
Draw 1 Samples

□ Create Sample Indices

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 create indices (this will be skipped for this lab)

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



Draw Sample

● Random Sample

Percent of cases to be sampled: 20

☑ Sample with replacement

○ Systematic Sample

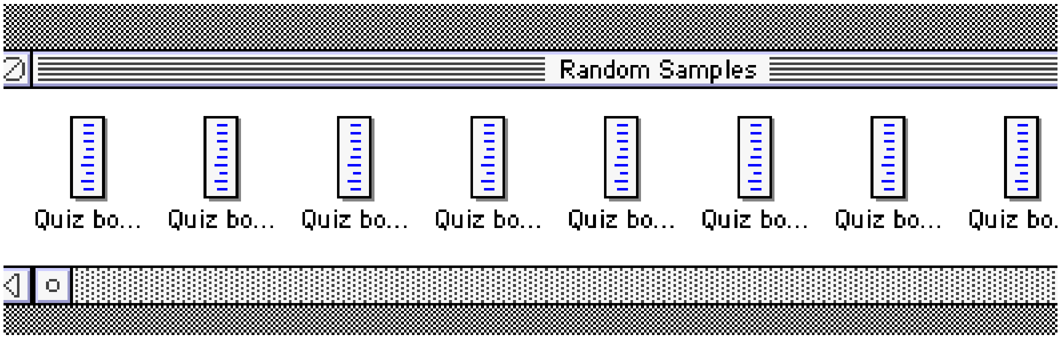
Start at Case:

Sample Every Case

Draw 25 Samples

□ Create Sample Indices

DataDesk will generate the following:

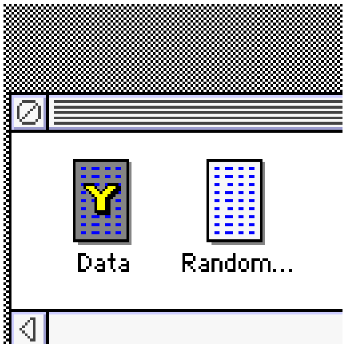


Open one of the samples and you'll see that it contains a single answer to a single question.

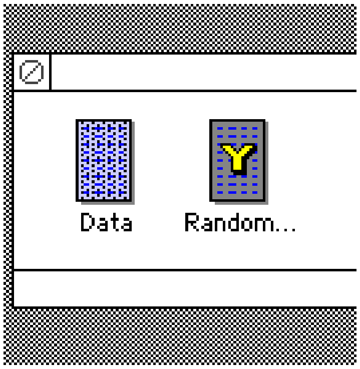
Generate statistics describing the random draw

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:

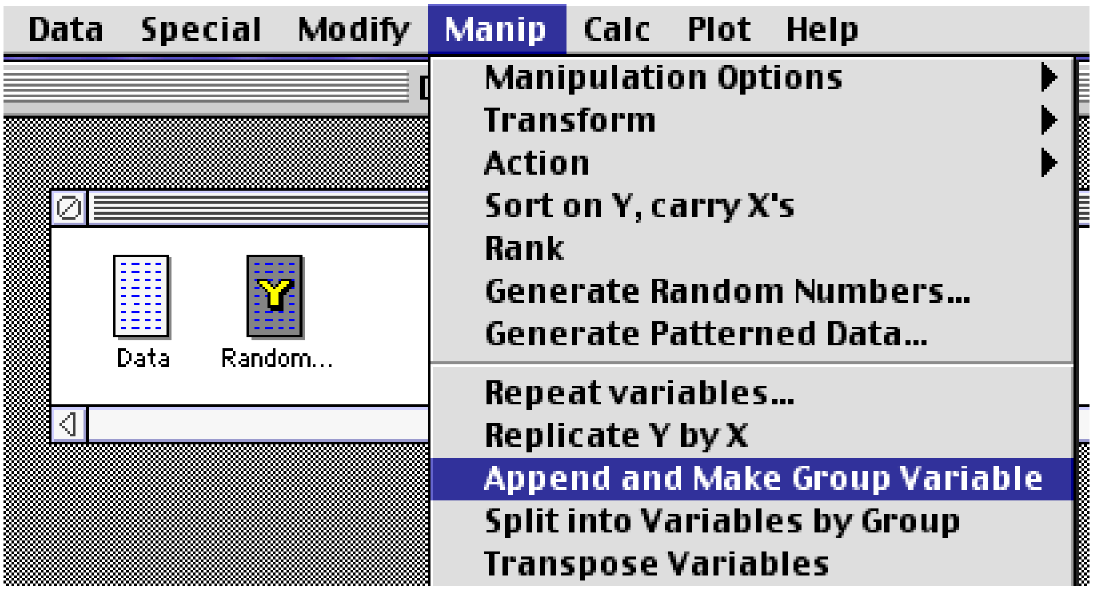
1. Click on the top left icon of the Random Samples window. The data window will appear with your Quiz box variable labeled Y



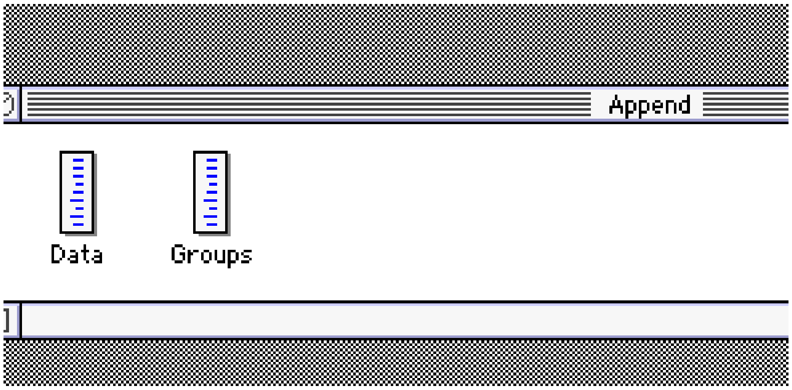
2. Click on Random to assign it to be Y



3. Now highlight Manip and select Append and Make a Group Variable

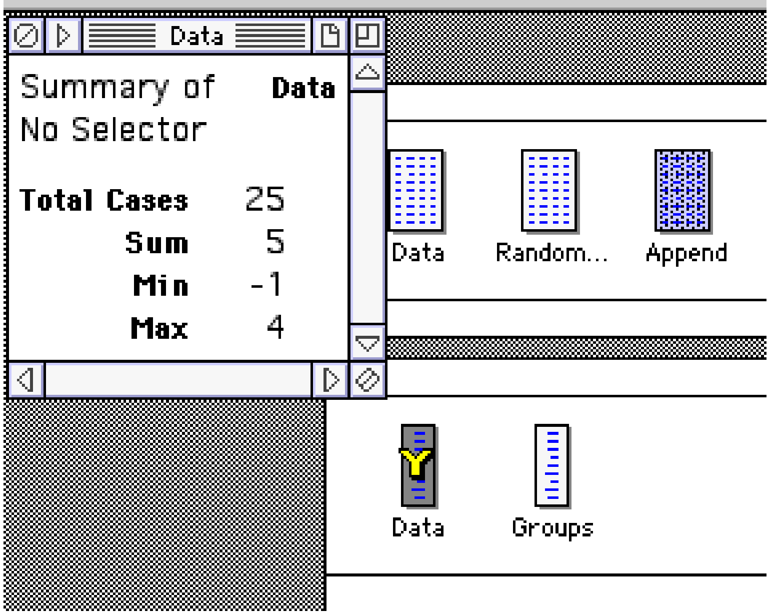


DataDesk now creates two variables. Data contains the 25 samples you drew. You can see this by clicking on Data to open it.



4. Now request the sum of the elements in the box by highlighting Calc, making certain the report will include the Sum (take a look in Calculation options) and requesting from Summaries a Report (see Lab 2, pp. 23-26)

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



Summary of **Data**
No Selector

```
Total Cases   25
        Sum    5
        Min   -1
        Max    4
```

Your results will vary.

Lab 5: Correlation and Regression

Lab 5's objectives

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

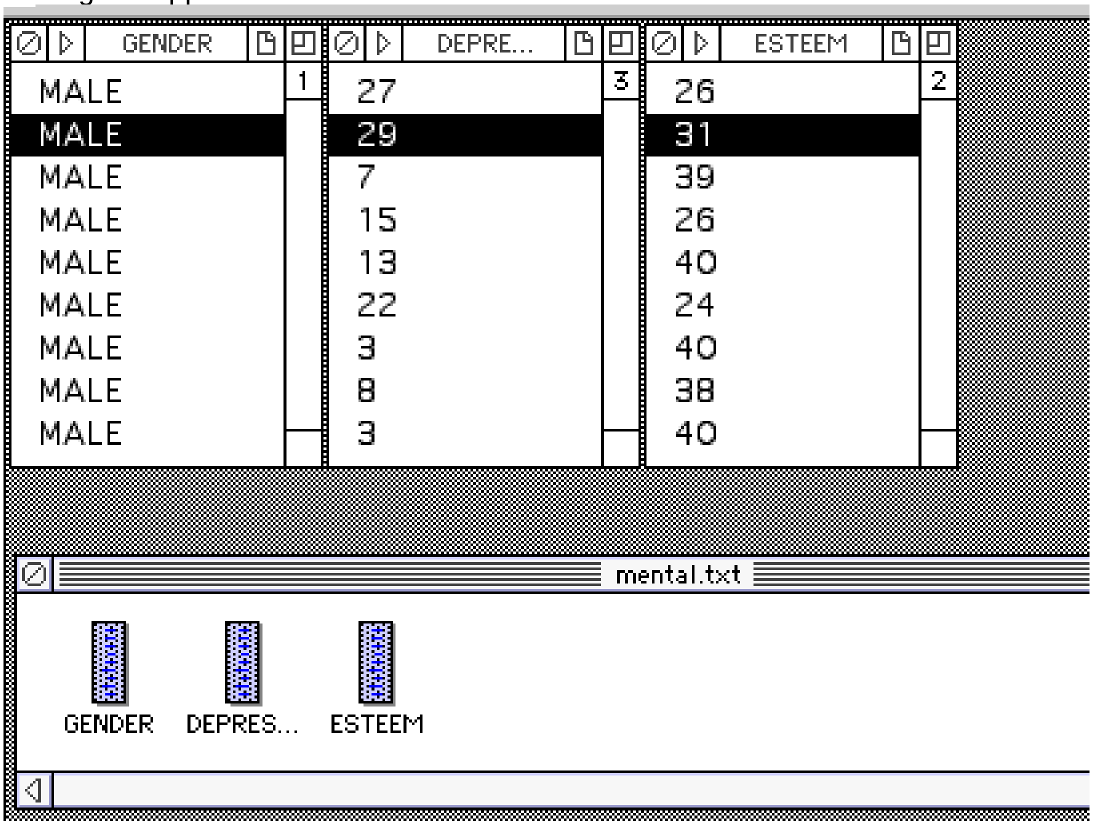
The data you are going to use comes from a real study the Prof did a few years ago. She was interested in the relationship between depression and self-esteem in college students. She expected two things to be true, based on previous research studies: 1) those with lower self-esteem should report higher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

Stored for you in the Global Shared Files folder is a data set, mental, 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. Open DataDesk (see Lab 1, p. 11)
2. Highlight File and click on Open Datafile
3. Open mental which is stored in the Global Shared Folders under this course (see Lab 1, pp 12-13)
4. Click OK to the Locked file notice

The following will appear:

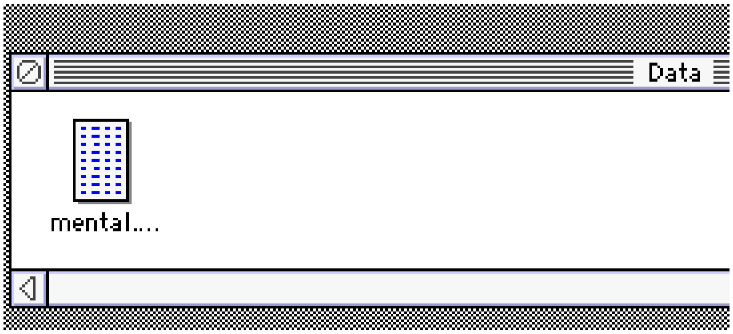


GENDER	DEPRE...	ESTEEM
MALE	27	26
MALE	29	31
MALE	7	39
MALE	15	26
MALE	13	40
MALE	22	24
MALE	3	40
MALE	8	38
MALE	3	40

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 male who scored 29 on the Depression scale and 31 on the self-esteem scale).

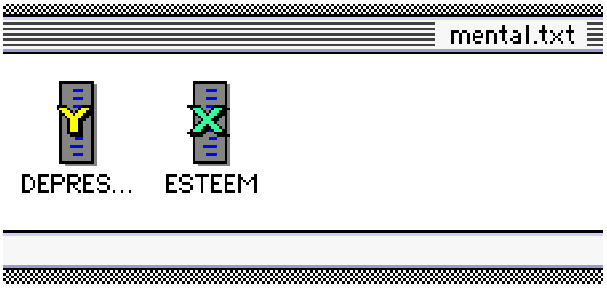
5. Try moving the cursor in one of the variable windows and you will see that it automatically moves in the others

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

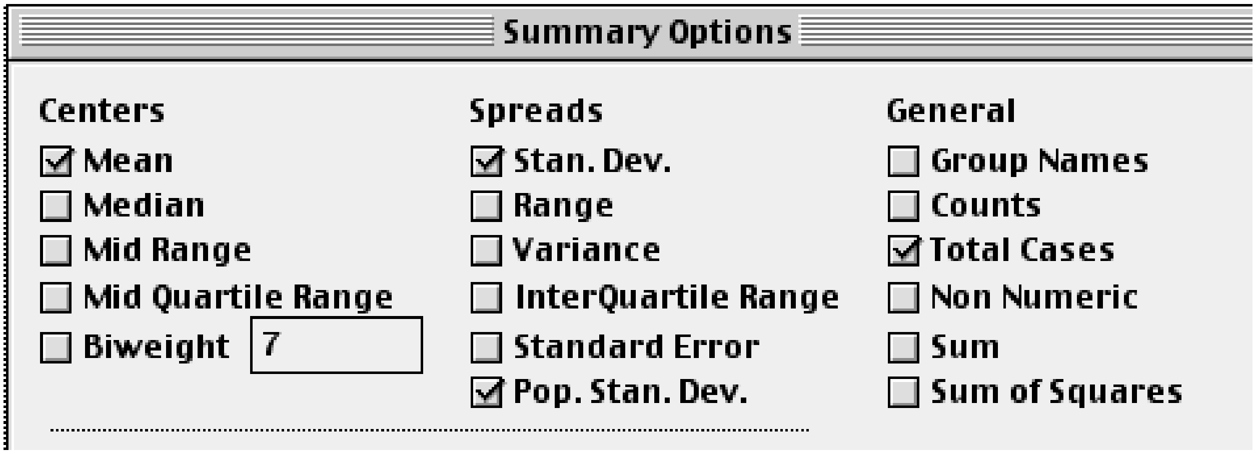


Before you ever analyze data, you should always check to see that it holds no surprises for you. That's because computers happily analyze without thinking whatever you give them and if there is a mistake in your data, it is up to you to find it. So the first step is to look at the data.

7. Assign DEPRESSION to be Y and ESTEEM to be X (see Lab 4, p. 44)



8. Highlight Calc, then Calculation Options, then Select Summary Statistics and ask for the following: (see Lab 2, p. 24)

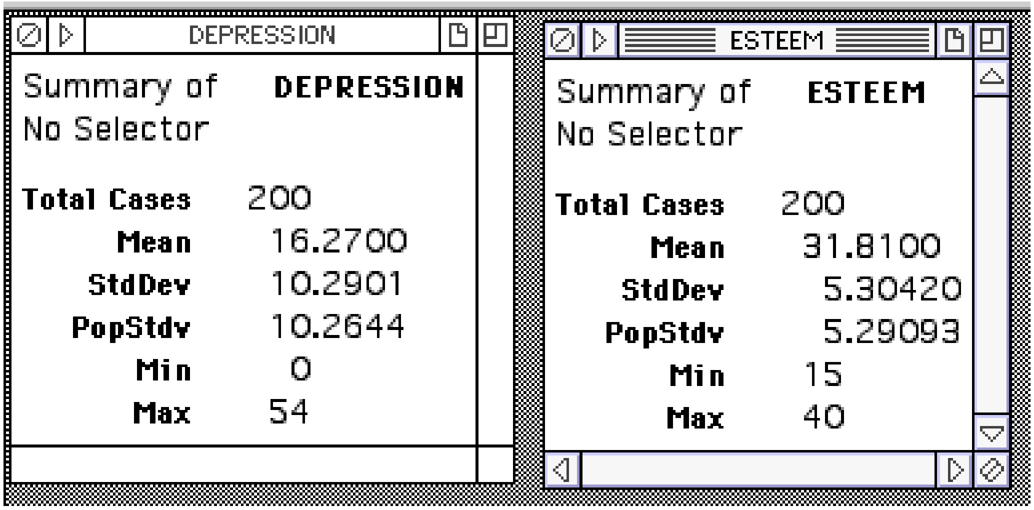


Summary Options

Centers	Spreads	General
☑ Mean	☑ Stan. Dev.	□ Group Names
□ Median	□ Range	□ Counts
□ Mid Range	□ Variance	☑ Total Cases
□ Mid Quartile Range	□ InterQuartile Range	□ Non Numeric
□ Biweight 7	□ Standard Error	□ Sum
	☑ Pop. Stan. Dev.	□ Sum of Squares

9. Then highlight Calc, Summaries, and select report (see Lab 2, p. 26)

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



```
Summary of   DEPRESSION
No Selector

Total Cases    200
       Mean     16.2700
     StdDev     10.2901
    PopStdv     10.2644
        Min      0
        Max     54
```

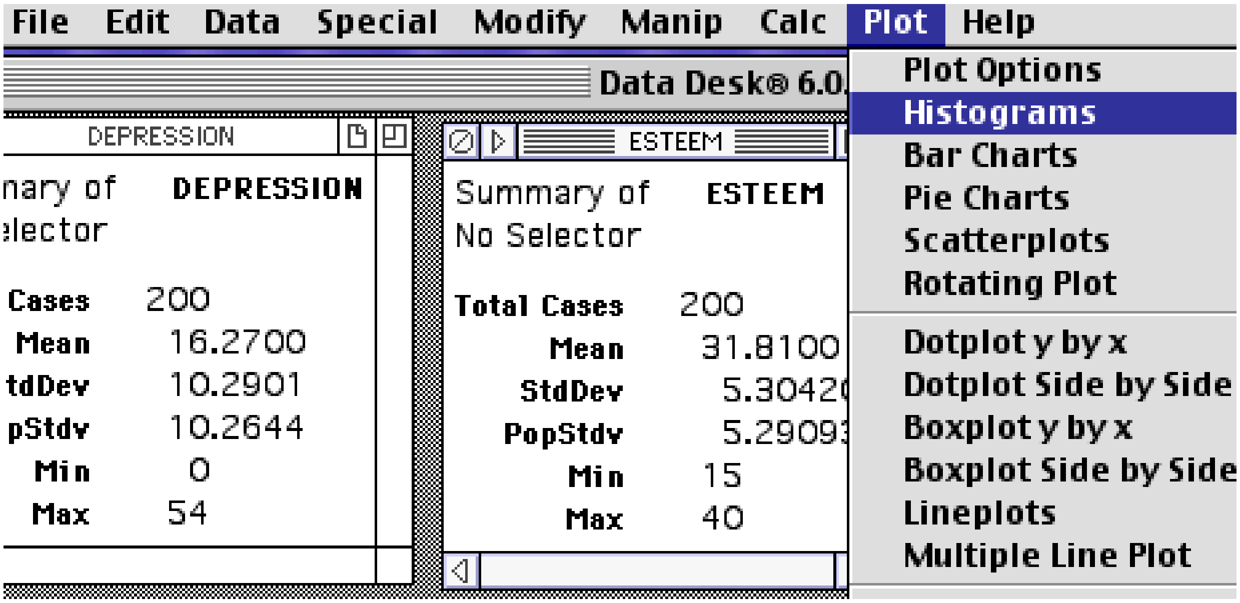
```
Summary of   ESTEEM
No Selector

Total Cases    200
       Mean     31.8100
     StdDev      5.30420
    PopStdv      5.29093
        Min     15
        Max     40
```

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. 36)



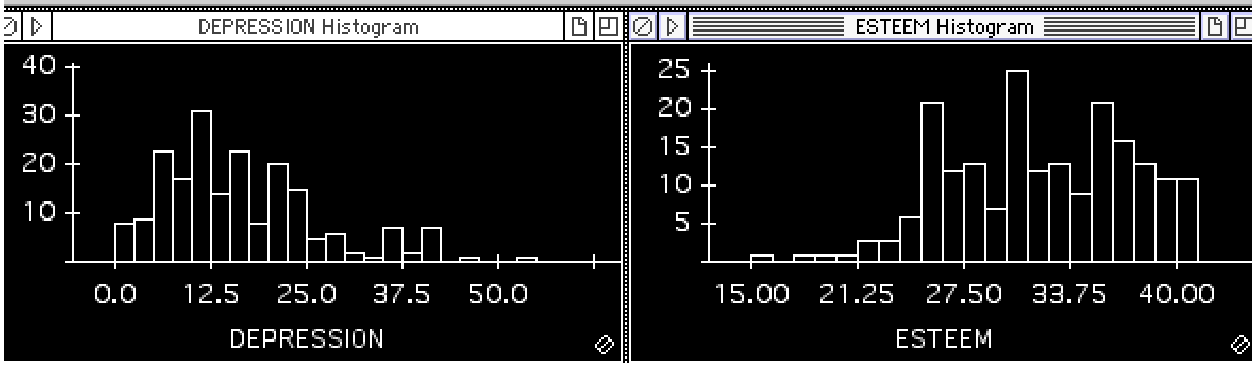
```
nary of   DEPRESSION
elector

Cases    200
 Mean     16.2700
tdDev     10.2901
pStdv     10.2644
  Min      0
  Max     54
```

```
Summary of   ESTEEM
No Selector

Total Cases    200
       Mean     31.8100
     StdDev      5.3042(
    PopStdv      5.2909:
        Min     15
        Max     40
```

The following will appear:



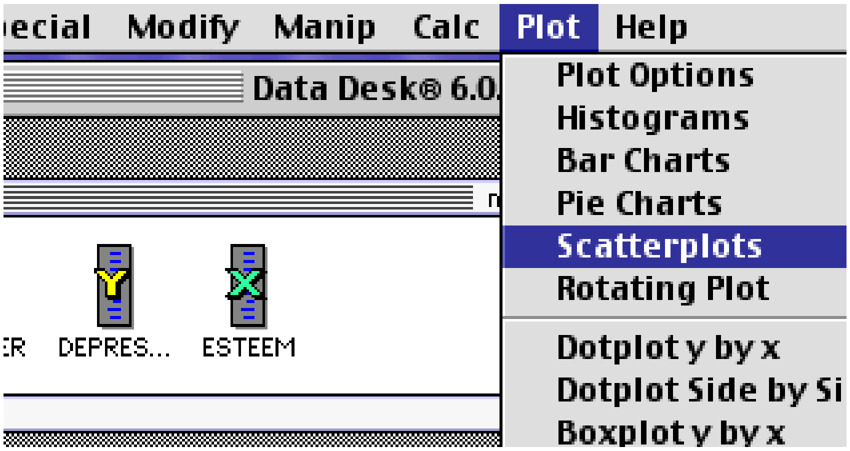
This shows you that DEPRESSION is skewed to the right. Most people report low levels of depression, but a few report high levels. The effect of this is to drag the mean to the right of the median. In contrast, ESTEEM is skewed to the left. Most people report high levels of self-esteem but a few report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

Make scatterplots of two variables

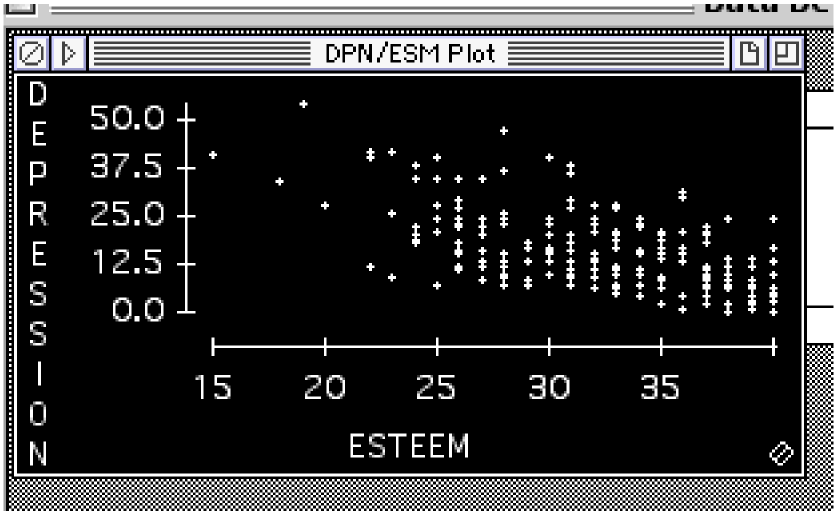
Now, you are going to see if depression is correlated with self-esteem. First, you are going to do it visually.

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?)

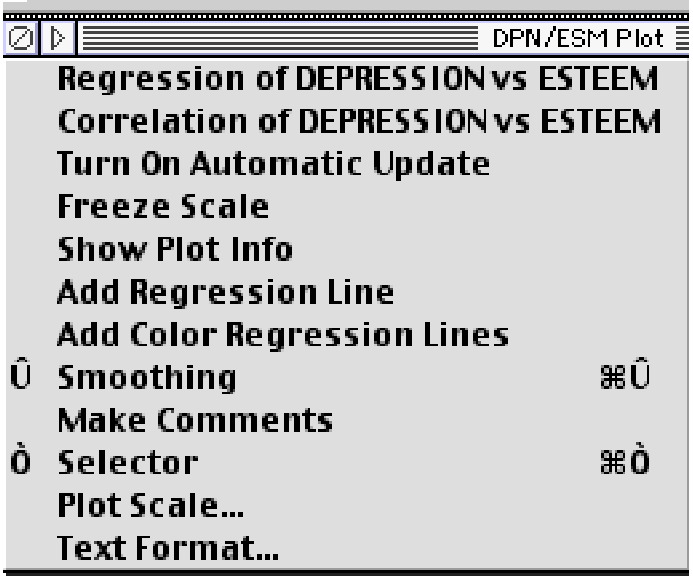
You can do several things with this scatterplot. Under Plot Options, you can select to reverse black and white printing. You can expand this plot by pulling on the bottom righthand corner of the window.

Plot regression lines

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

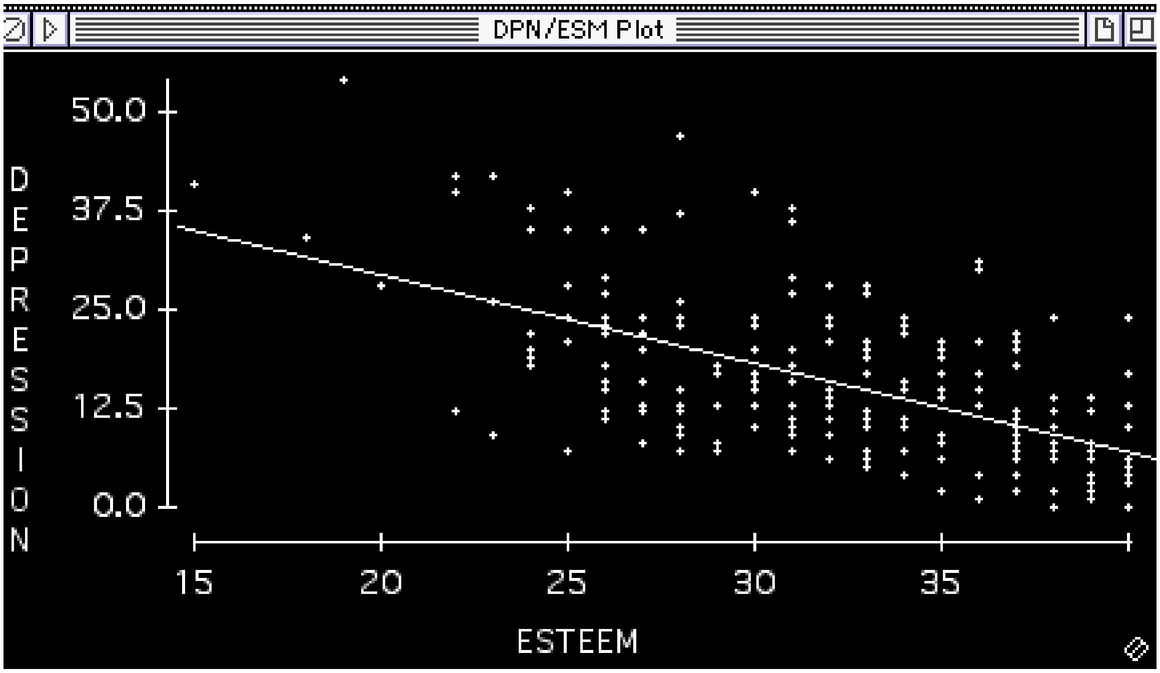
1. Click on the arrow at the upper left of the DPN/ESM Plot bar line

The following window will open:



2. Select Add Regression Line

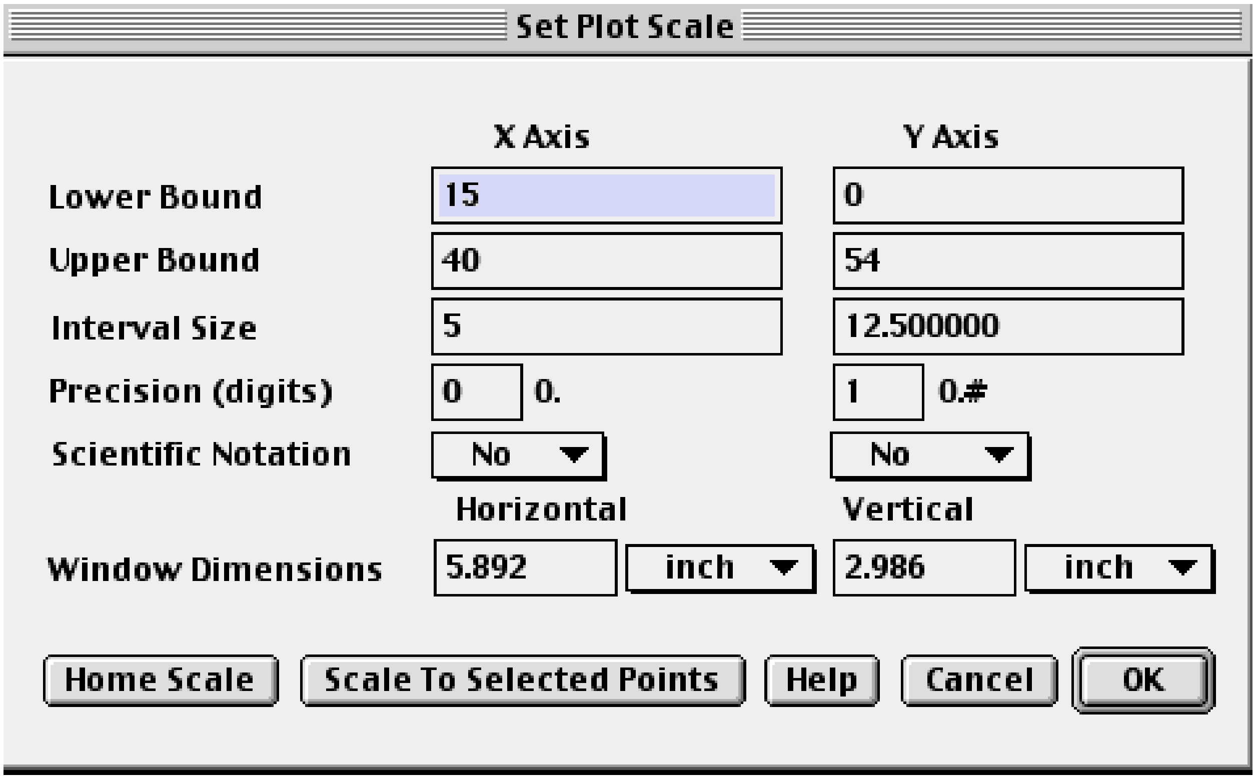
This does the following to your plot:



Now, try to make your plot look more spiffy.

3. In the Hyperview menu, select Plot Scale

The following window opens:

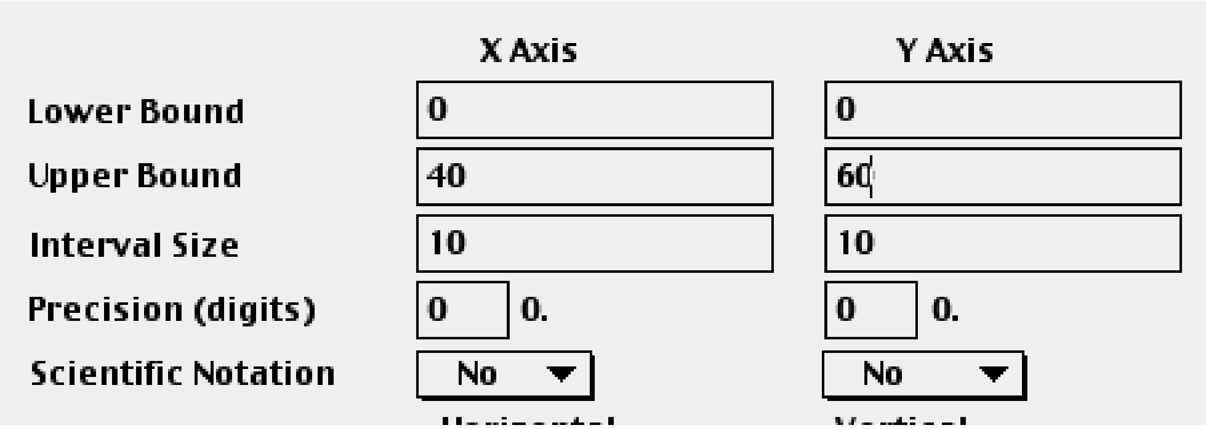


Set Plot Scale

	X Axis	Y Axis
Lower Bound	15	0
Upper Bound	40	54
Interval Size	5	12.500000
Precision (digits)	0 0.	1 0.#
Scientific Notation	No	No

	Horizontal		Vertical	
Window Dimensions	5.892	inch	2.986	inch

4. Edit this so that both variables start a 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:



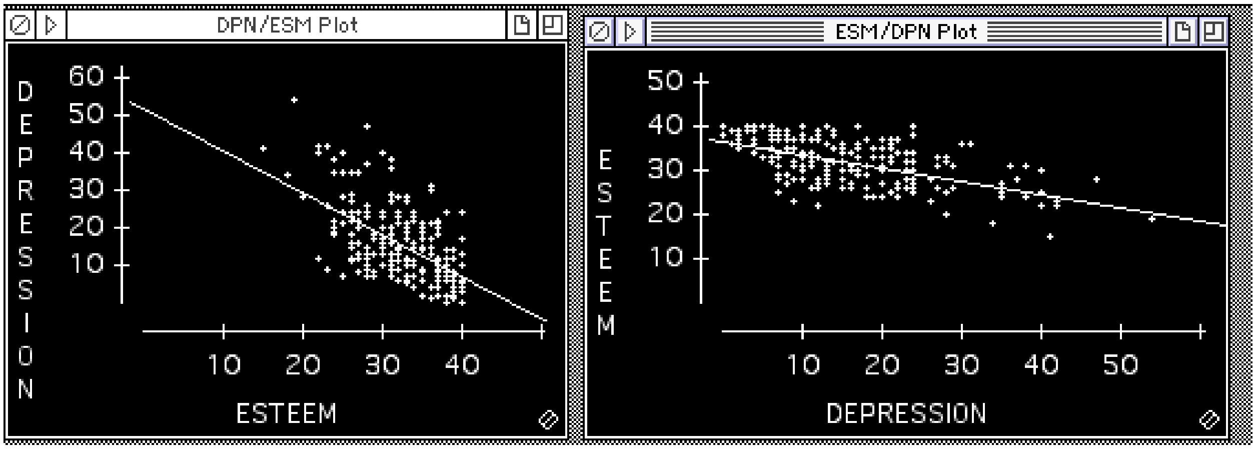
	X Axis	Y Axis
Lower Bound	0	0
Upper Bound	40	60
Interval Size	10	10
Precision (digits)	0 0.	0 0.
Scientific Notation	No	No

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 can place the two plots side by side:

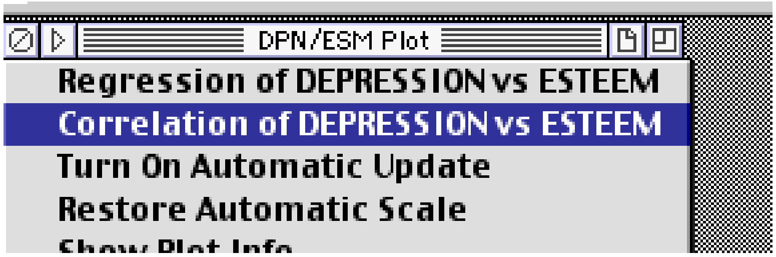


The one on the left shows depression scores predicted by self-esteem levels. The one on the right depicts self-esteem levels predicted by depression scores. Notice both the slope of the line and the intercept (the point where the line crosses the left axis) are different. Notice, too, that the plots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

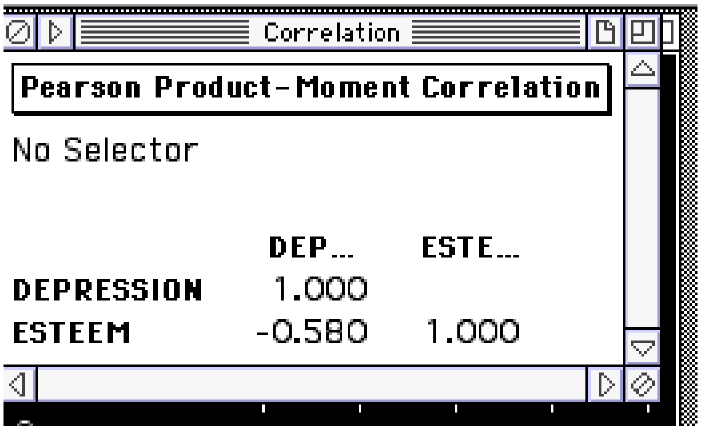
Calculate the correlation for two variables in a scatterplot

To calculate a correlation inside a scatterplot:

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



The following appears:



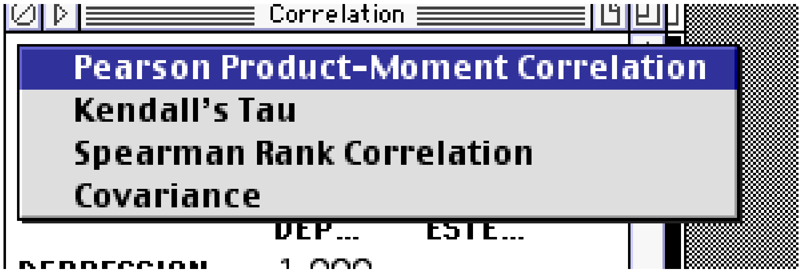
Pearson Product-Moment Correlation

No Selector

	DEP...	ESTE...
DEPRESSION	1.000	
ESTEEM	-0.580	1.000

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

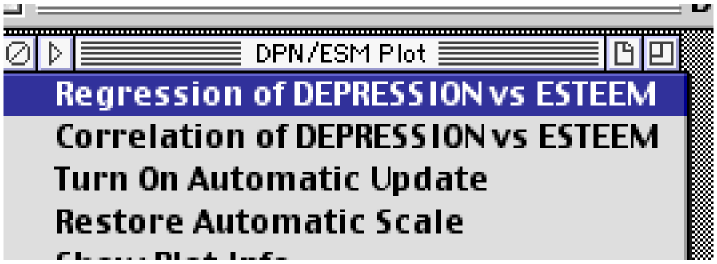


The window opens to show that you can choose to do two other types of correlations that we will not learn in this course. The results are different (check it out...). You can also ask for the covariance estimate, a topic not covered at length in this course.

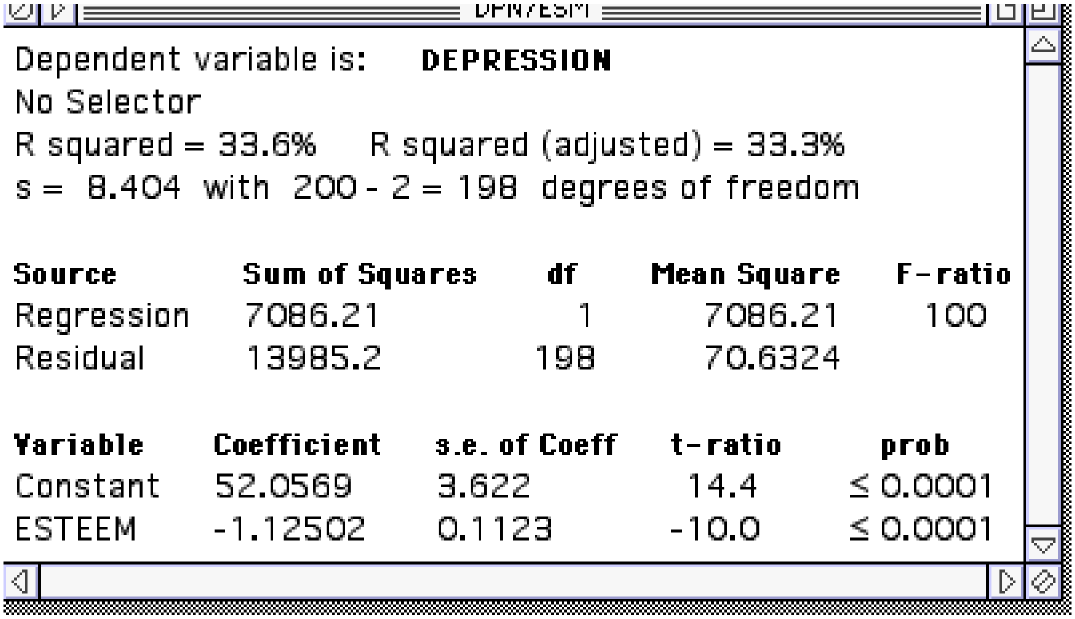
Calculate regression statistics

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:



```
Dependent variable is:    DEPRESSION
No Selector
R squared = 33.6%    R squared (adjusted) = 33.3%
s =  8.404  with  200 - 2 = 198  degrees of freedom

Source         Sum of Squares     df     Mean Square    F-ratio
Regression     7086.21             1       7086.21        100
Residual       13985.2           198       70.6324

Variable      Coefficient     s.e. of Coeff     t-ratio      prob
Constant      52.0569         3.622              14.4      ≤ 0.0001
ESTEEM        -1.12502        0.1123            -10.0      ≤ 0.0001
```

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 coeffient). The t-ratio is a statistical test where the null hypothesis is that the coeficient 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

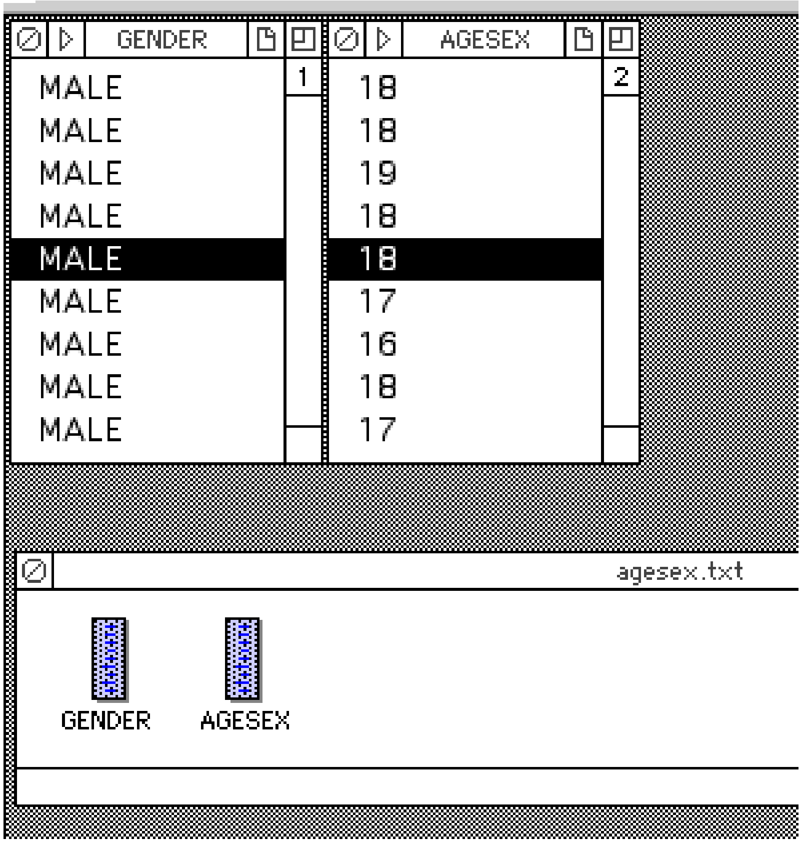
The 3rd National Health and Nutrition Examination Survey estimated that among Americans between 17 and 55 years old, the age at first sexual intercourse was 17.4 years. For men, it was 16.9 years and for women it was 17.9 years. These are considered to be population-based estimates, that is, they are 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, and in the Global Shared Files folder the Prof has stored a dataset, agesex, with information from 400 of these subjects, half of them male, half of them female.

Because sexually experienced college students are younger 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. Open DataDesk (see Lab, p. 11)
2. Open the file agesex which is stored in the Global Shared Files folder (see Lab 1, pp. 11-13)

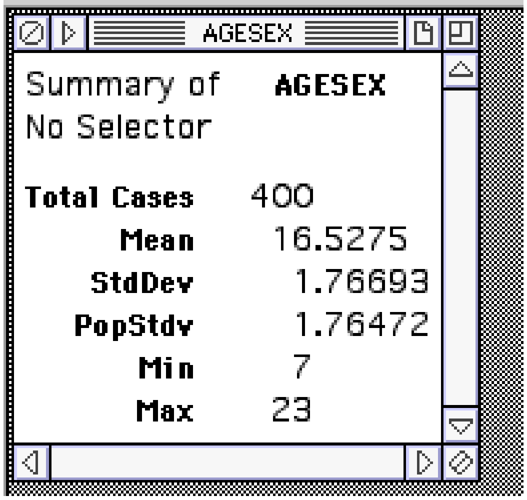


GENDER	AGESEX
MALE	18
MALE	18
MALE	19
MALE	18
MALE	18
MALE	17
MALE	16
MALE	18
MALE	17

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.

3. Assign agesex to be the Y variable (see Lab 2, p.27)
4. Highlight Calc, adjust the options to show the mean, SD for the sample and population, and the minimum and maximum values (see Lab 2, pp. 23-25)
5. Ask for the summary report (see Lab 2, p. 26)



```
Summary of     AGESEX
No Selector

Total Cases    400
       Mean     16.5275
     StdDev      1.76693
    PopStdv      1.76472
        Min      7
        Max     23
```

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 Std Dev (SD^+ in your textbook) and the estimate of the population SD (SD in your book) are very close. This is because the sample is relatively large.

Now, because these are all undergraduates, it makes sense that the oldest age of first sexual intercourse reported 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 sexual 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.

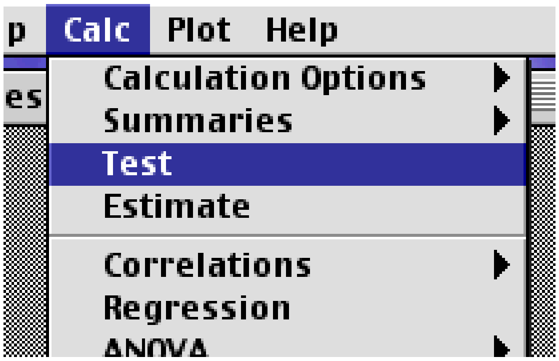
Here are the statistical hypotheses:

H_0: Mean age 1st sex of college students = or > 17.4 years (Null hypothesis)

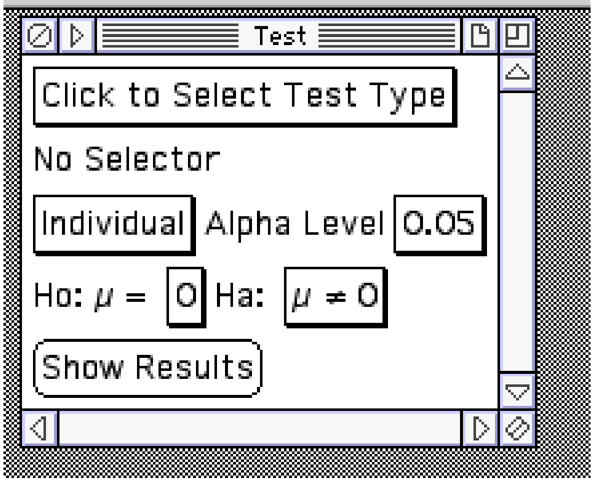
H_1: Mean of college students in study < 17.4 years (Alternate hypothesis)

Now it's time to test the hypothesis.

6. Highlight Calc, and select Test



The following window opens:



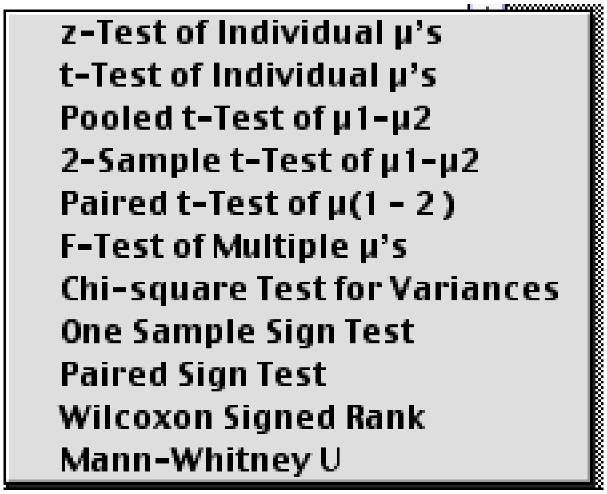
No Selector

Alpha Level 0.05

Ho: μ = 0 Ha: $\mu \neq 0$

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

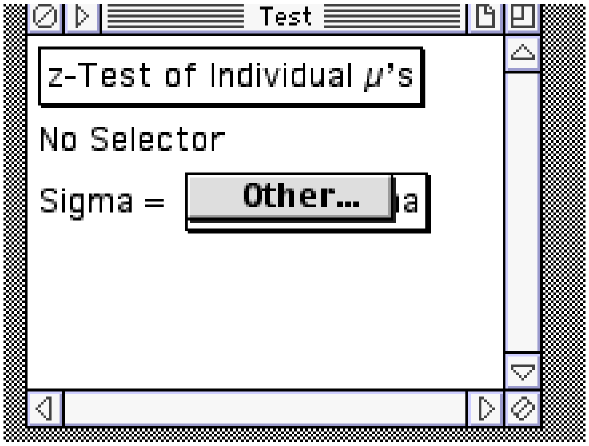
7. Click on Click to Select Test Type to see your options



8. Choose z-Test of individual μ's

This is a one-sample z-test. The next decision is what you estimate the SD of the U.S. population to be (the SD of the box). Sigma is another word for the SD of the population--it is a parameter. Your best guess, for reasons presented in your text when the t-test is described, is the SD of your sample (SD^+), or 1.77.

9. Click on Specify sigma. Other will appear

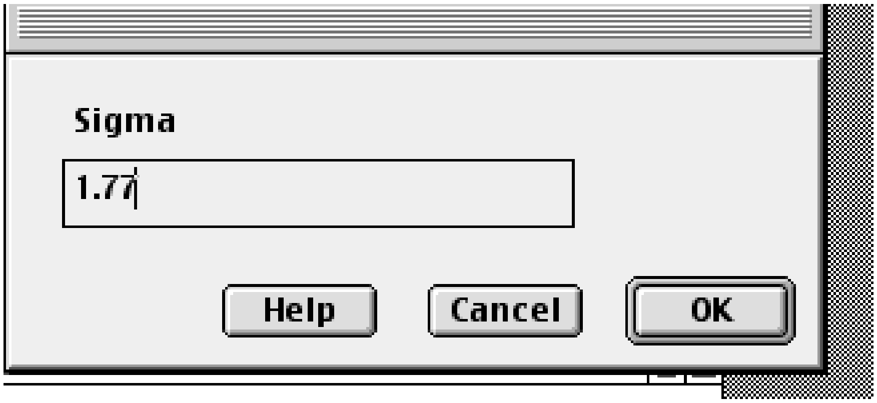


z-Test of Individual μ's

No Selector

Sigma =

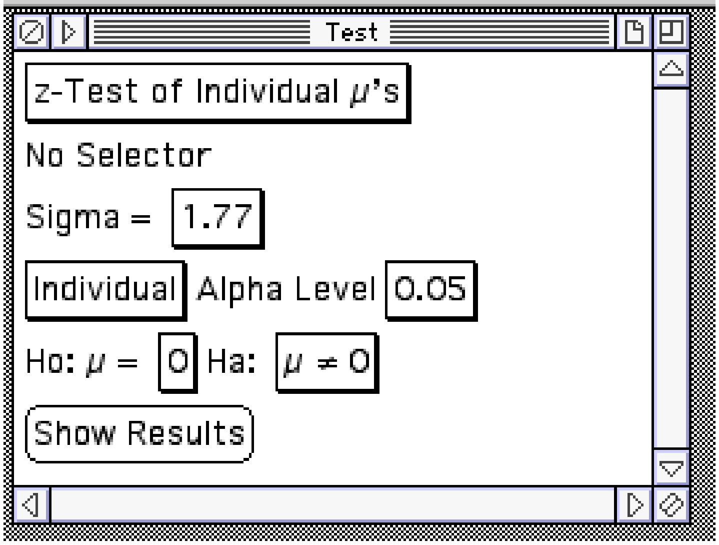
10. Click on Other and another box will appear. Enter your estimate and click OK



Sigma

1.77

Now the following appears:



z-Test of Individual μ's

No Selector

Sigma = 1.77

Individual Alpha Level 0.05

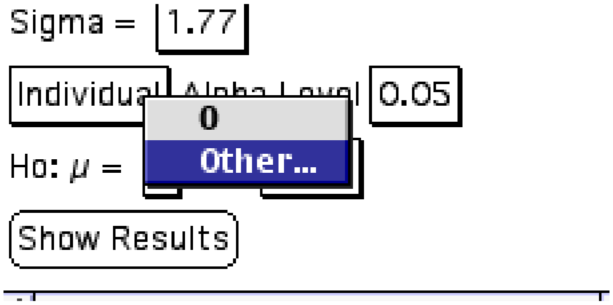
Ho: μ = 0 Ha: μ ≠ 0

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 (Americans between 17 and 55 years of age) if the difference in age at first sexual intercourse 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 μ = 0, but for your hypothesis it should read 17.4.

11. Click on μ and select Other



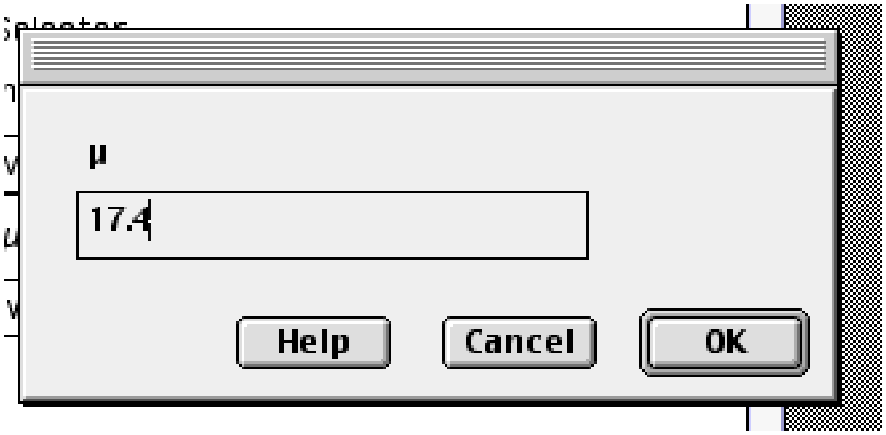
Sigma = 1.77

Individual Alpha Level 0.05

Ho: μ =

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

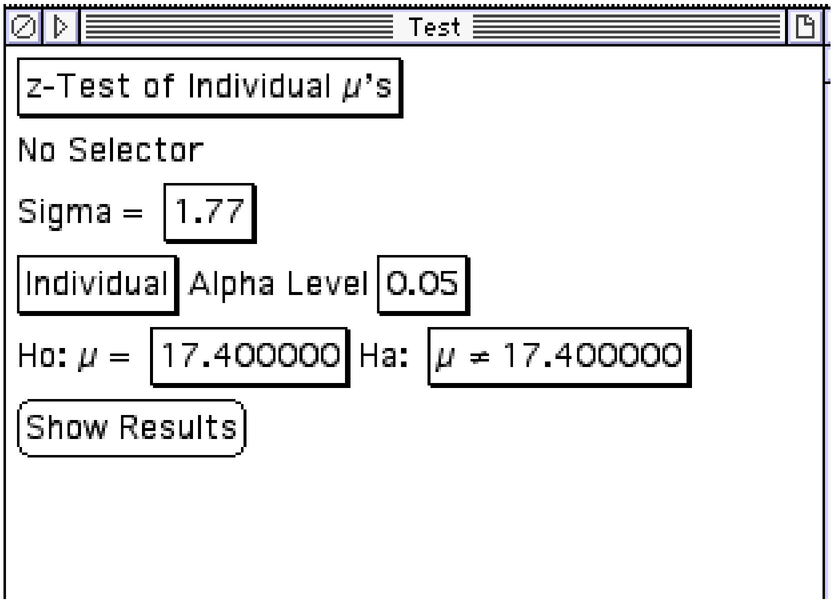
12. Enter 17.4 in the window and click OK



μ

17.4

The following appears:



z-Test of Individual μ's

No Selector

Sigma = 1.77

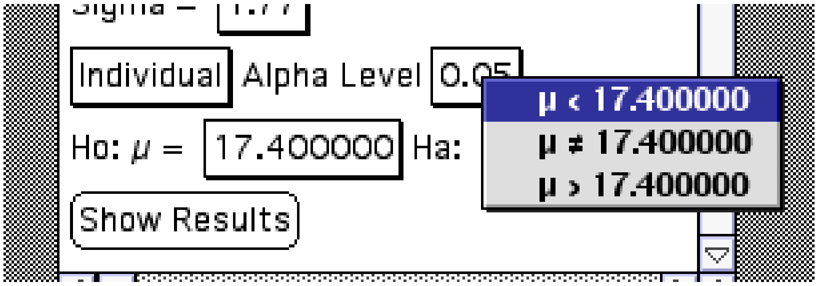
Individual Alpha Level 0.05

Ho: μ = 17.400000 Ha: μ ≠ 17.400000

Now you have one last piece to edit.

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

13. Change Ha to show that your hypothesis is μ < 17.4 by clicking on μ ≠17.4 and select μ < 17.4

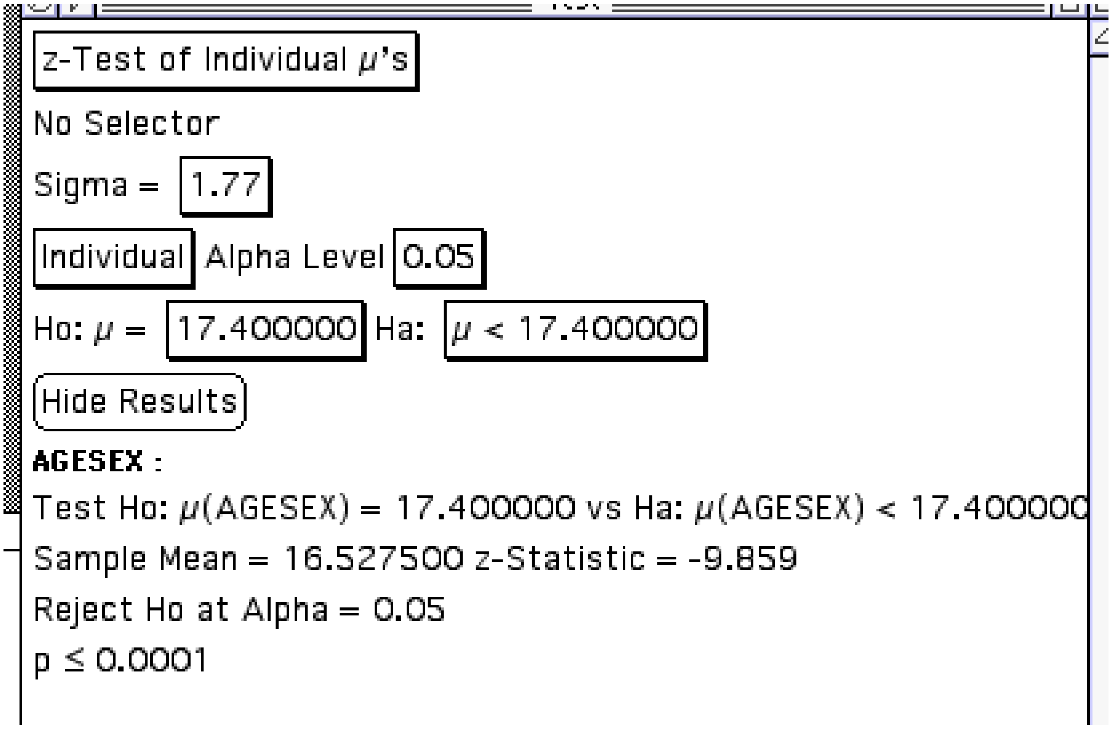


Individual Alpha Level 0.05

Ho: μ = 17.400000 Ha:

14. Click Show Results

The following appears:



z-Test of Individual μ's

No Selector

Sigma = 1.77

Individual Alpha Level 0.05

Ho: μ = 17.400000 Ha: μ < 17.400000

Hide Results

AGESEX :

```
Test Ho: μ(AGESEX) = 17.400000 vs Ha: μ(AGESEX) < 17.40000
Sample Mean = 16.527500 z-Statistic = -9.859
Reject Ho at Alpha = 0.05
p ≤ 0.0001
```

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:

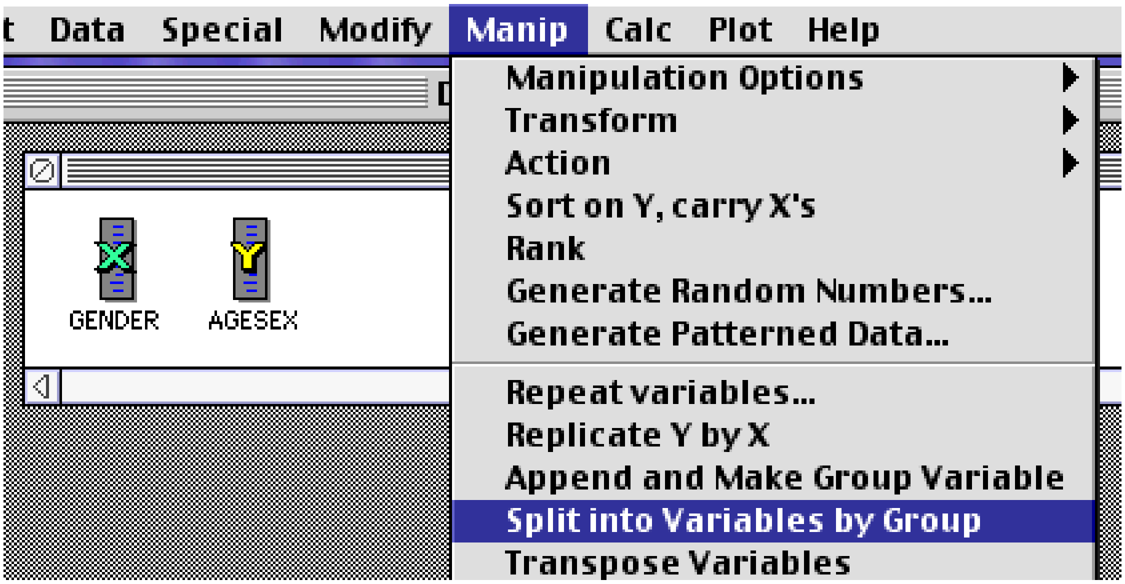
H_0: Mean age at first sex for men = mean age at first sex for women
(Null hypothesis)

H_1: Mean age at first sex for men is not equal to mean age at first sex for women
(Alternate hypothesis)

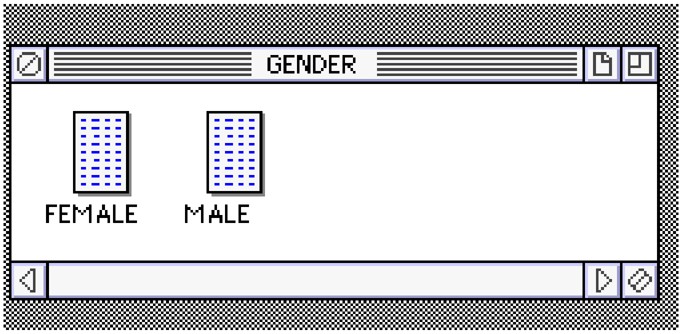
Before you can test the hypothesis, you have to divide the ages in your dataset into those that are men's and those that are women's.

1. Make agesex the Y variable and gender the X variable (see Lab 4, p. 44)

2. Highlight Manip and select Split into Variables by Group


t Data Special Modify Manip Calc Plot Help
Manipulation Options
Transform
Action
Sort on Y, carry X's
Rank
Generate Random Numbers...
Generate Patterned Data...
Repeat variables...
Replicate Y by X
Append and Make Group Variable
Split into Variables by Group
Transpose Variables
GENDER
AGESEX

The following appears:



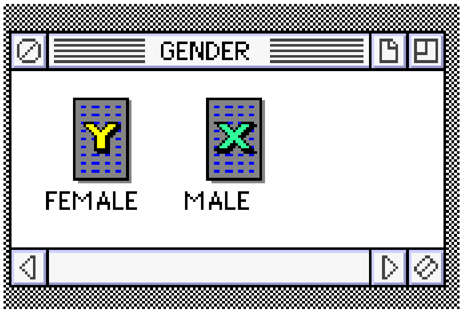
3. Click on the Female icon until you see the ages appear

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.

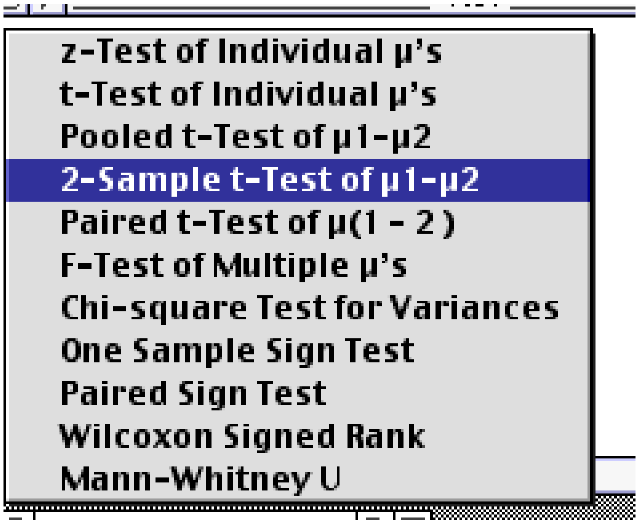
Perform a two-sample t-test

1. Assign women (Female) to be Y and men (Male) to be X (see Lab 4, p. 44)

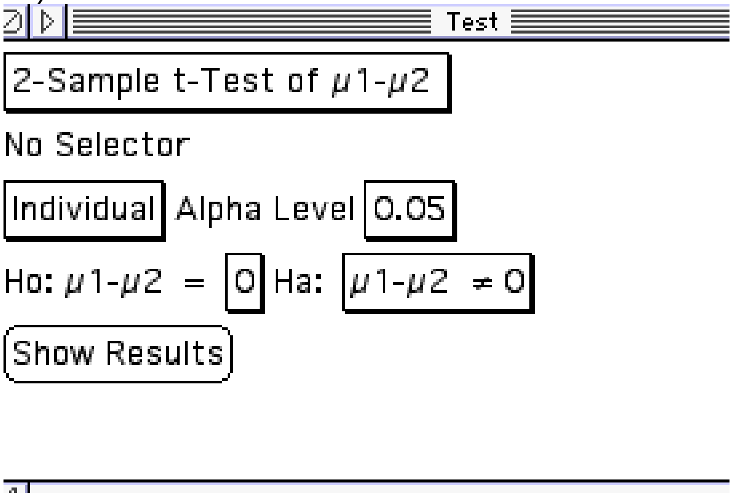


2. Highlight Calc and select Test

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



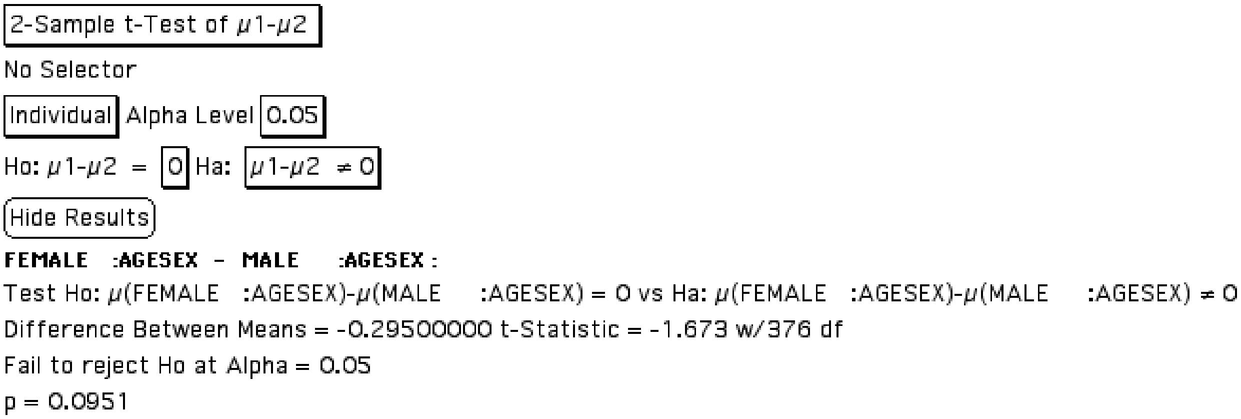
4. Edit the Ho: to show a difference of 0 (the µ's or population means are equal) (see Lab 6, p. 68)
5. Edit the Ha: to show the difference is not 0 (the µ's or population means are not equal) (see Lab 6, p. 69)



2-Sample t-Test of μ1-μ2
No Selector
Individual Alpha Level 0.05
Ho: μ1-μ2 = 0 Ha: μ1-μ2 ≠ 0

5. Now click Show Results

The following window appears:



2-Sample t-Test of μ1-μ2
No Selector
Individual Alpha Level 0.05
Ho: μ1-μ2 = 0 Ha: μ1-μ2 ≠ 0

```
FEMALE   :AGESEX  -  MALE    :AGESEX :
Test Ho: μ(FEMALE   :AGESEX)-μ(MALE    :AGESEX) = 0 vs Ha: μ(FEMALE   :AGESEX)-μ(MALE    :AGESEX) ≠ 0
Difference Between Means = -0.29500000 t-Statistic = -1.673 w/376 df
Fail to reject Ho at Alpha = 0.05
p = 0.0951
```

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 we 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 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 because you are using a student version of DataDesk that limits you to 1,000 cases)

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

7. Using the prezages dataset, estimate the age that cuts off the upper 20% of the distribution. To do this:
   - Follow the instructions on pages 37-38
   - 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 Calculation 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 Calculation 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 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 dataset?

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

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

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