# GAISE Workshop

## Session 3
## Nov. 8, 2005
## 3:30 – 5:30 pm

**Brian Smith**

# Using technology for developing concepts and analyzing data

# Uses of Technology

- **Super calculator/Number cruncher**
  - In this mode the use of technology allows the student to move from solving problems with small samples to analyzing large, realistic data sets.
  - Low level of conceptual understanding.

- **Grapher**
  - Assists in visualization of data, exploration of concepts of central tendency and variability, shape of distribution (symmetric? unimodal? normal?)
  - Medium level of conceptual understanding

- **Interactive Explorer**
  - Integrates numerical measures and graphs into an interactive package that permits exploration, visualization, and active learning.
  - High level of conceptual understanding

In this session we will investigate five technologies:

1) Excel
2) Minitab
3) Fathom
4) TI-83/84
5) Java Applets

A brief word about the advantages and limitations of each technology:

**Excel** is a superb number cruncher but its statistical capabilities are limited. Doubt has been cast on its random number generator for large scale commercial and scientific applications , but it is adequate for educational purposes. Several add-ins that add extra statistical features and improve user interface are available, and are often included with text books. Major disadvantage in the classroom is that many statistical procedures are not supported and it takes time and effort to program one's own applications.

**Minitab** is a true statistical package with a large number of built-in statistical procedures. Many applications that are very limited in Excel are automatically available in Minitab, and in particular the ability to generate graphs to accompany procedures is useful. Solves large scale problems easily, has an intuitive Windows based user-friendly interface, and a relatively short learning curve. It also has the advantage that it is a serious statistical package that can be used for commercial applications as well as being an effective educational tool.

**Fathom** is an original and highly effective package for interactive exploration. It is an educational, pedagogical package and encourages independent investigation of the relationships between variables.  It includes a superb collection of real data, including U.S. census data, for use by students, and also incorporates a number of useful interactive demonstrations that can be employed by the instructor for illustrating statistical concepts. Fathom's interface is different from the point-and-click environment that we have come to expect from the Window's based applications, but when you get used to it, Fathom is easy to use and requires little time and effort for the instructor who wishes to demonstrate statistical concepts.

**The TI-83/84** has the major advantage of being inexpensive and portable – essentially it puts major computing power into the hands of every student. Numerical output and graphical are limited by the screen size and the quality of the graphics, but nevertheless the universality of the calculators makes them a useful and effective tool for both numerical computation and graphical explorations.

**Applets** are small applications that are intended to demonstrate one concept e.g. Central Limit Theorem. There are many applets online and the user should carefully check them out before using them in class – they are not all of uniform quality. But there are some sites which are highly reliable and have effective applets. It is often useful to pause for a few minutes in a class to pull up an applet to illustrate a particular concept – and then tell students to continue to explore the applet, and the underlying concept, in out-of-class time.

# Excel

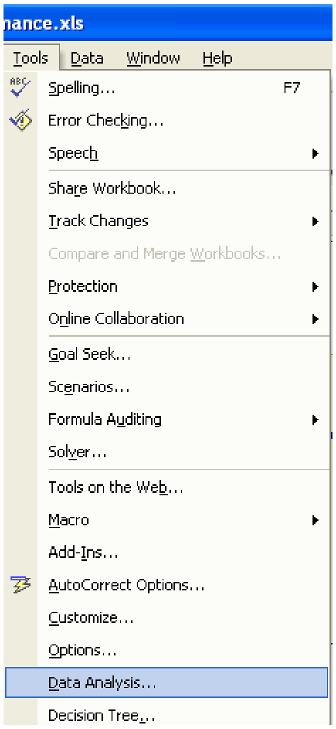
The following data ,obtained from the DASL website, shows long jump records for the Olympic games from 1900 (Year = 0) to 1984 (Year = 84).

| **Long jump** | **year** |
|---:|---:|
| 282.88 | 0 |
| 289.00 | 4 |
| 294.50 | 8 |
| 299.25 | 12 |
| 281.50 | 20 |
| 293.13 | 24 |
| 304.75 | 28 |
| 300.75 | 32 |
| 317.31 | 36 |
| 308.00 | 48 |
| 298.00 | 52 |
| 308.25 | 56 |
| 319.75 | 60 |
| 317.75 | 64 |
| 350.50 | 68 |
| 324.50 | 72 |
| 328.50 | 76 |
| 336.25 | 80 |
| 336.25 | 84 |

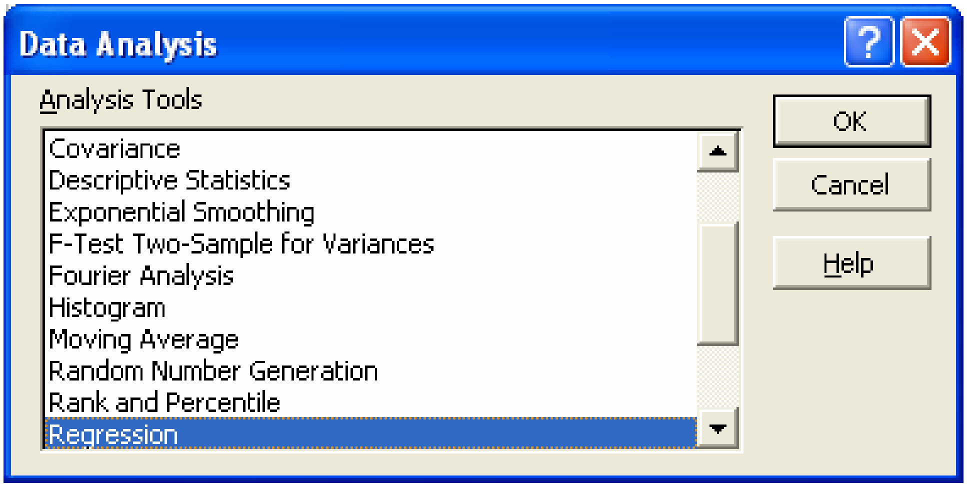
## Level 1 use of Excel: supercalculator

To perform a **Simple Linear Regression** in Excel:

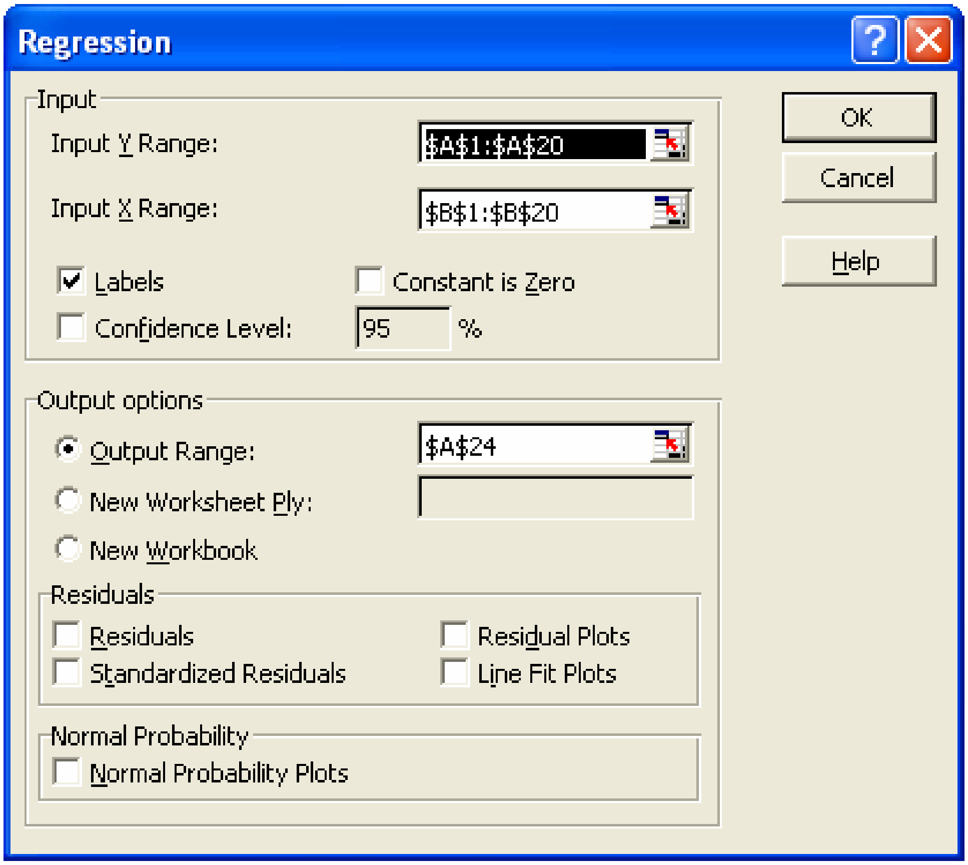
Click on Tools □ Data Analysis



Select Regression from the menu



Complete the dialog box (be sure to check the "labels" box!):



Input

Input Y Range: $A$1:$A$20

Input X Range: $B$1:$B$20

☑ Labels　□ Constant is Zero

□ Confidence Level: 95 %

Output options

● Output Range: $A$24

○ New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals　□ Residual Plots

□ Standardized Residuals　□ Line Fit Plots

Normal Probability

□ Normal Probability Plots

Click on OK to obtain the following output:

**SUMMARY OUTPUT**

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.8703 |
| R Square | 0.7575 |
| Adjusted R Square | 0.7432 |
| Standard Error | 9.7568 |
| Observations | 19 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 5054.89 | 5054.89 | 53.10 | 1.27001E-06 |
| Residual | 17 | 1618.33 | 95.20 | | |
| Total | 18 | 6673.22 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|---|
| Intercept | 283.45 | 4.28 | 66.22 | 0.00000000 | 274.42 | 292.49 |
| Year | 0.6131 | 0.0841 | 7.2870 | 1.2700E-06 | 0.4356 | 0.7906 |

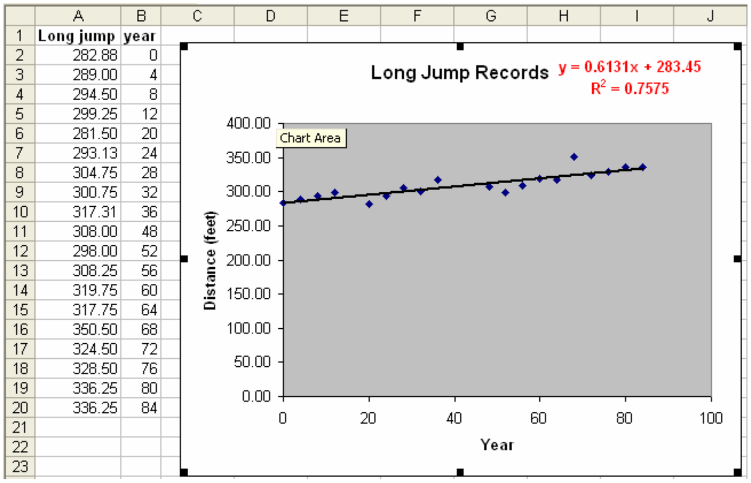
The regression equation is $\hat{Y} = \beta_0 + \beta_1 X = 283.45 + 0.6131X$ .

We see that in the test of hypothesis $H_O$: $\beta_1 = 0$ the *p*-value is 1.2700E-06 (compare with TI-84) and we conclude *Reject $H_O$*, i.e. conclude that there is a linear relationship between length and year.

Further, we note that a 95% confidence interval for $\beta_1$ is $0.4356 \leq \square_1 \leq 0.7906$.

## Level 2 use of Excel: grapher

Use the Chart Wizard feature to draw a scatter diagram and superimpose a linear regression on the diagram, specifying the linear regression equation and the $R^2$ value.

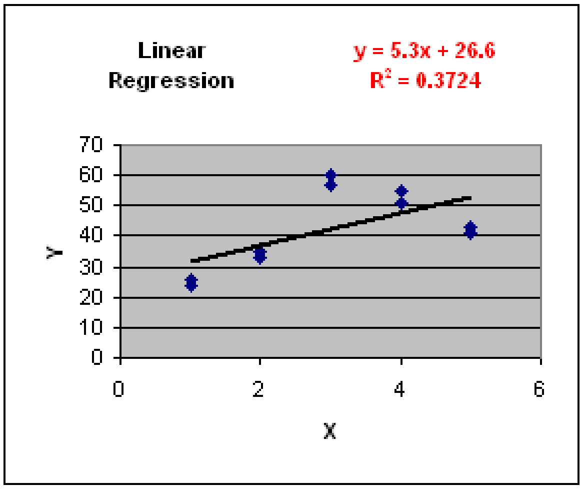


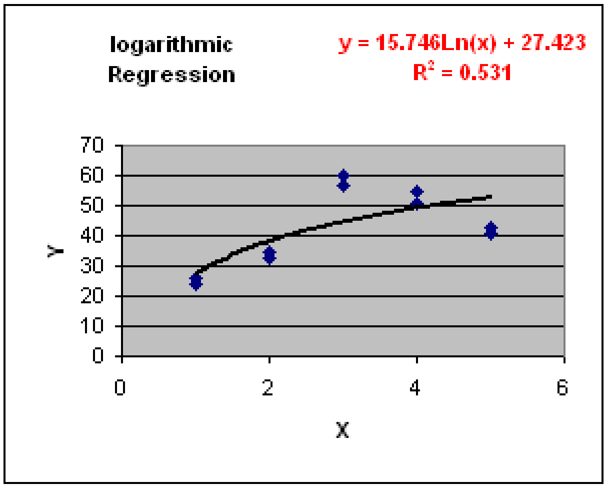
| | A | B |
|---|---|---|
| 1 | Long jump | year |
| 2 | 282.88 | 0 |
| 3 | 289.00 | 4 |
| 4 | 294.50 | 8 |
| 5 | 299.25 | 12 |
| 6 | 281.50 | 20 |
| 7 | 293.13 | 24 |
| 8 | 304.75 | 28 |
| 9 | 300.75 | 32 |
| 10 | 317.31 | 36 |
| 11 | 308.00 | 48 |
| 12 | 298.00 | 52 |
| 13 | 308.25 | 56 |
| 14 | 319.75 | 60 |
| 15 | 317.75 | 64 |
| 16 | 350.50 | 68 |
| 17 | 324.50 | 72 |
| 18 | 328.50 | 76 |
| 19 | 336.25 | 80 |
| 20 | 336.25 | 84 |
| 21 | | |
| 22 | | |
| 23 | | |

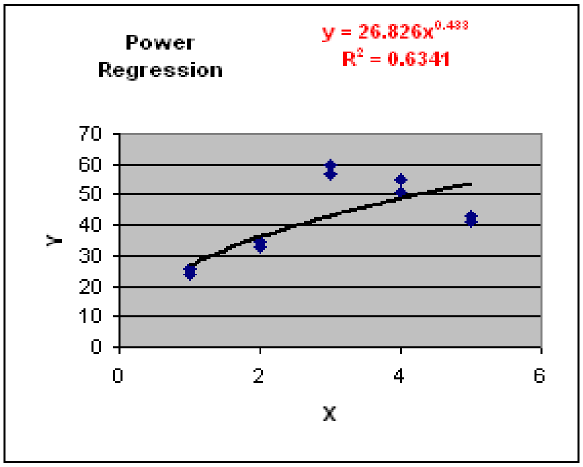
## Level 3 use of Excel: Interactive

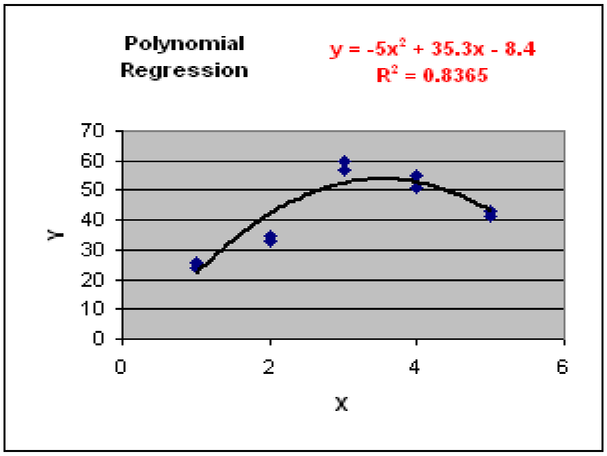
**Explore the relationship between X and Y for the following data**

| X | Y |
|---|---|
| 1 | 24 |
| 1 | 26 |
| 2 | 35 |
| 2 | 33 |
| 3 | 57 |
| 3 | 60 |
| 4 | 55 |
| 4 | 51 |
| 5 | 43 |
| 5 | 41 |







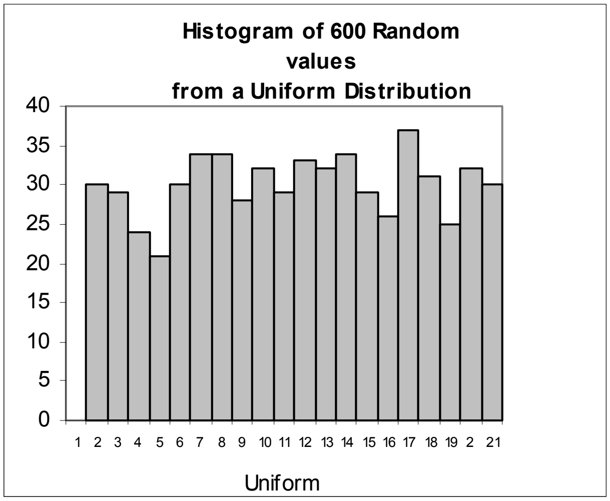


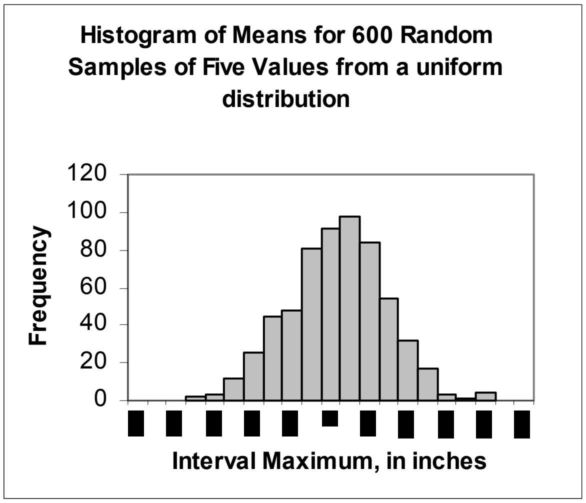
Experiment: Change one of the Y-values in the table and see the effect on the four graphs! For example, change the first Y-value from 24 to 200. Note: The graph is updated but the regression equations and $R^2$ values are not! It is necessary to delete these values and recompute

**Simulating the Central Limit Theorem in Excel**

## The Scenario

Consider a production process where the actual lengths of metal pipes produced vary uniformly Between 19.5 and 20.5 cm. At regular intervals a random sample of five pipes is selected and the mean length is recorded. After 600 samples have been selected, the distribution of the sample means is plotted. Even though the population values are uniformly distributed and the sample size is small, we see that the distribution of sample means is approximately normal.



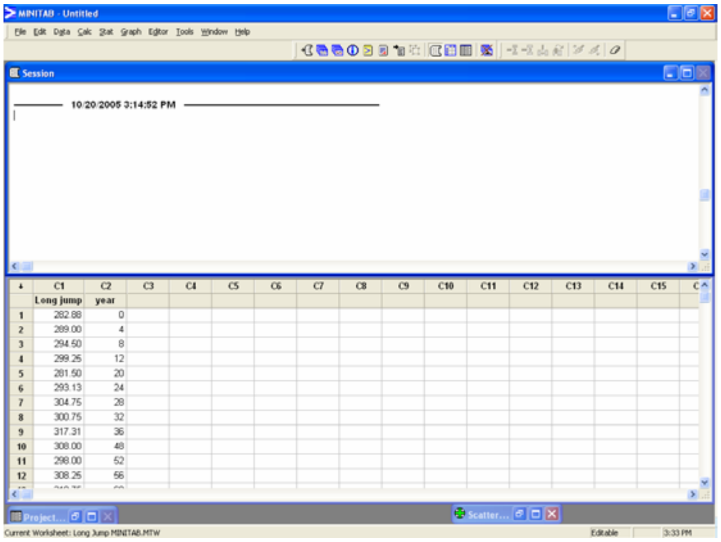


The ability to produce plots and to regenerate them interactively is based on the fact that the Rand() function in Excel, which generates a random number from a uniform distribution between 0 and 1, is automatically renewed every time the *recalculate* key (F9) is pressed.

The formula {=FREQUENCY(G3:G602,H2:H22)}is used to create the histogram. Every time the F9 function key is pressed the two graphs are regenerated. It is a nice demonstration of bthe fact that while 600 values of the original variable (length of metal pipe) is uniformly distributed in the range 19.5 to 20.5, the means of 600 samples of 5 metal pipes is approximately normally distributed.

# Minitab

We will show how Minitab can be used to analyze the long jump data:



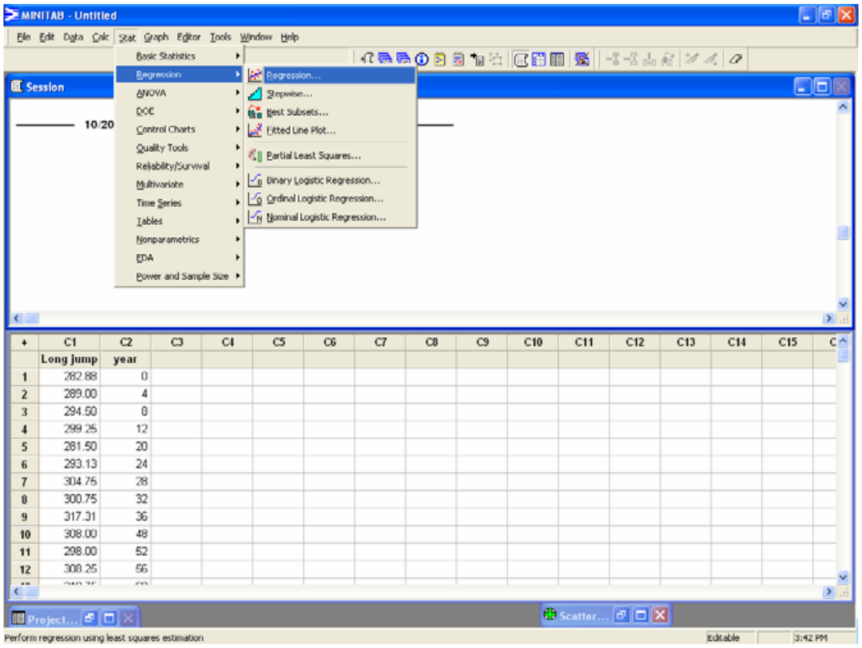
10/20/2005 3:14:52 PM

|  | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | C11 | C12 | C13 | C14 | C15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | Long jump | year |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1 | 282.88 | 0 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2 | 289.00 | 4 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 3 | 294.50 | 8 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 4 | 299.25 | 12 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 5 | 281.50 | 20 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 6 | 293.13 | 24 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 7 | 304.75 | 28 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 8 | 300.75 | 32 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 9 | 317.31 | 36 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 10 | 308.00 | 48 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 11 | 298.00 | 52 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 12 | 308.25 | 56 |  |  |  |  |  |  |  |  |  |  |  |  |  |

Current Worksheet: Long Jump MINITAB.MTW

Notice that the Minitab screen is divided into a *session window* (top half) and the *data window* (lower half). The data window is basically a spreadsheet and data can be exchanged with Excel by cutting and pasting. The long jump data is already stored in the data window.

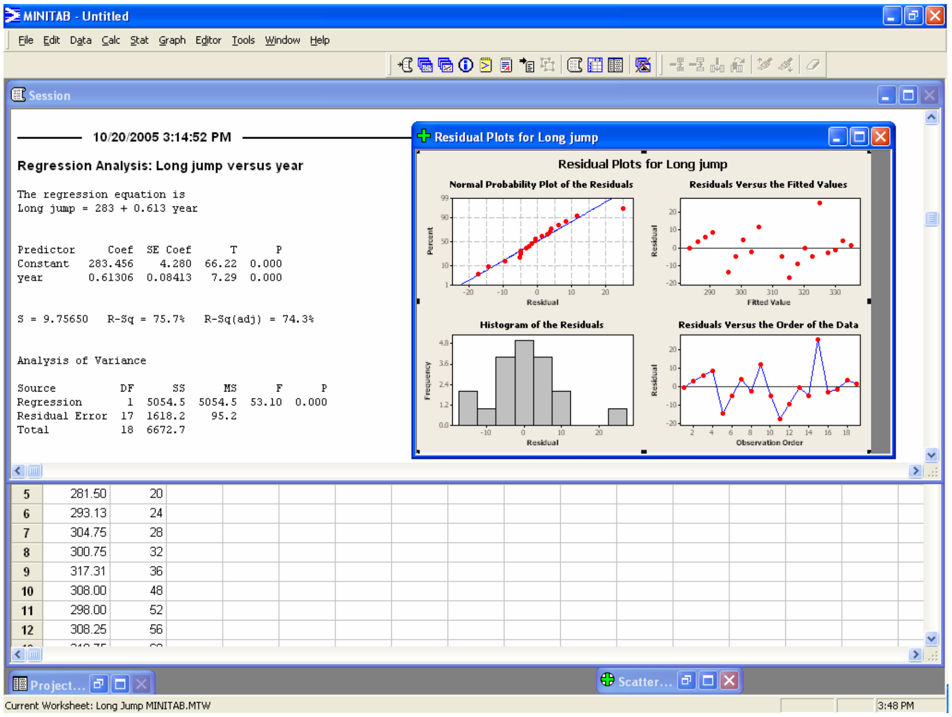
To perform a regression analysis select *Stat* > *Regression* > *Regression* as shown below:



|  | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | C11 | C12 | C13 | C14 | C15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | Long jump | year |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1 | 282.88 | 0 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2 | 289.00 | 4 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 3 | 294.50 | 8 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 4 | 299.25 | 12 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 5 | 281.50 | 20 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 6 | 293.13 | 24 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 7 | 304.75 | 28 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 8 | 300.75 | 32 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 9 | 317.31 | 36 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 10 | 308.00 | 48 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 11 | 298.00 | 52 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 12 | 308.25 | 56 |  |  |  |  |  |  |  |  |  |  |  |  |  |

Perform regression using least squares estimation

Running the regression analysis and selecting the *"Four in One"* option  from the *Graphs* menu, we get the following results:


MINITAB - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
Residual Plots for Long jump
Residual Plots for Long jump
Normal Probability Plot of the Residuals
Percent
99
90
50
10
1
-20 -10 0 10 20
Residual
Residuals Versus the Fitted Values
Residual
20
10
0
-10
-20
290 300 310 320 330
Fitted Value
Histogram of the Residuals
Frequency
4.8
3.6
2.4
1.2
0.0
-10 0 10 20
Residual
Residuals Versus the Order of the Data
Residual
20
10
0
-10
-20
2 4 6 8 10 12 14 16 18
Observation Order
Project...
Scatter...
3:48 PM

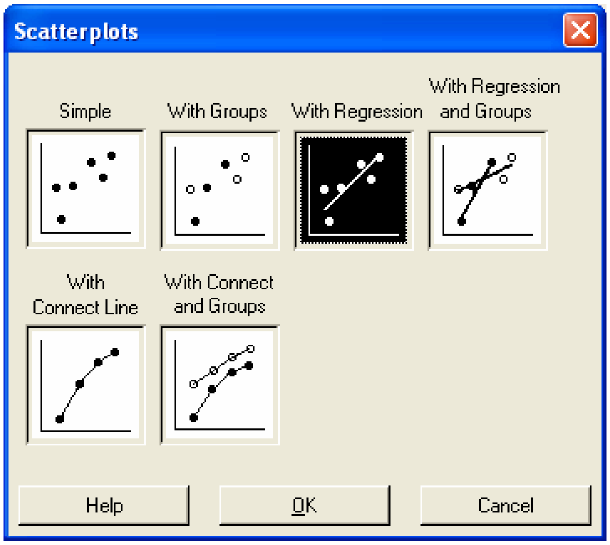
```
—————— 10/20/2005 3:14:52 PM ——————

Regression Analysis: Long jump versus year

The regression equation is
Long jump = 283 + 0.613 year


Predictor      Coef  SE Coef      T      P
Constant    283.456    4.280  66.22  0.000
year        0.61306  0.08413   7.29  0.000


S = 9.75650   R-Sq = 75.7%   R-Sq(adj) = 74.3%


Analysis of Variance

Source          DF      SS      MS      F      P
Regression       1  5054.5  5054.5  53.10  0.000
Residual Error  17  1618.2    95.2
Total           18  6672.7
```

| | | |
|---|---|---|
| 5 | 281.50 | 20 |
| 6 | 293.13 | 24 |
| 7 | 304.75 | 28 |
| 8 | 300.75 | 32 |
| 9 | 317.31 | 36 |
| 10 | 308.00 | 48 |
| 11 | 298.00 | 52 |
| 12 | 308.25 | 56 |

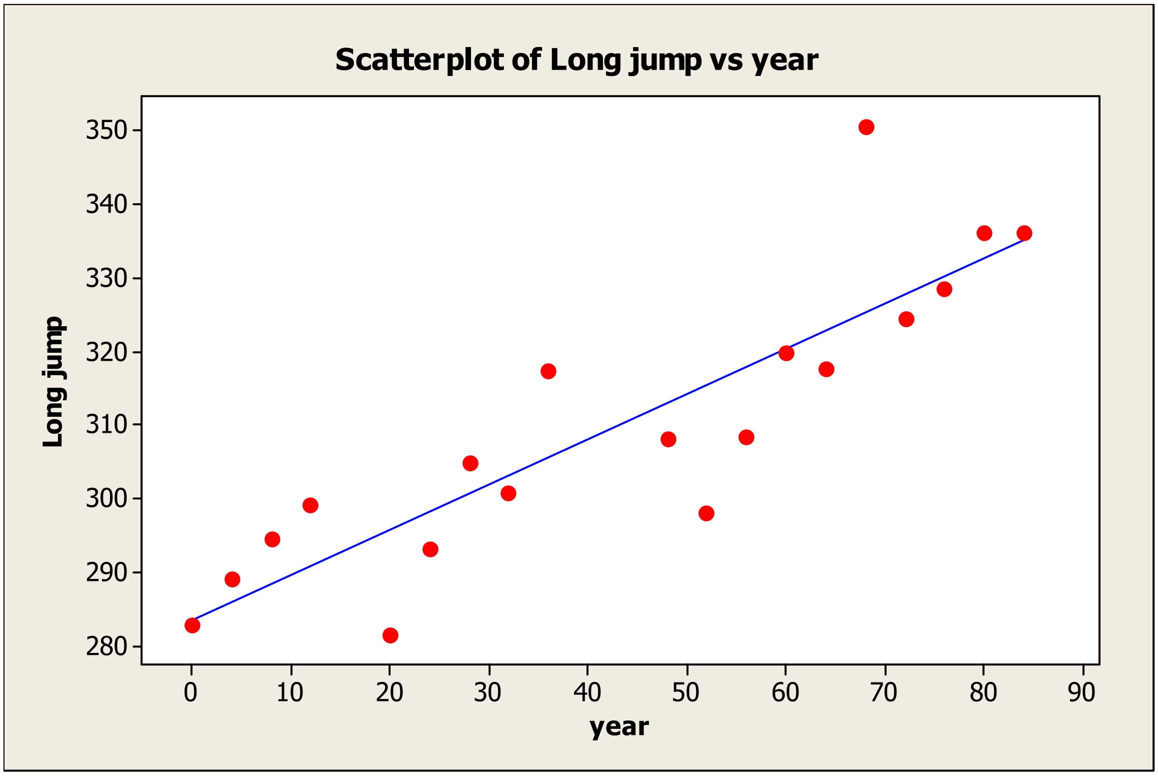
Current Worksheet: Long Jump MINITAB.MTW

Notice that this output includes not only the regression equation and the ANOVA table for the regression, but also four graphs which can be useful for testing regression assumptions (e.g. normal distribution of residuals, equal variances, etc.)

To generate a scatter diagram select *Graph > Scatterplot*



and choose the "*With Regression*" option from the dialog box above.
The scatterplot appears below:

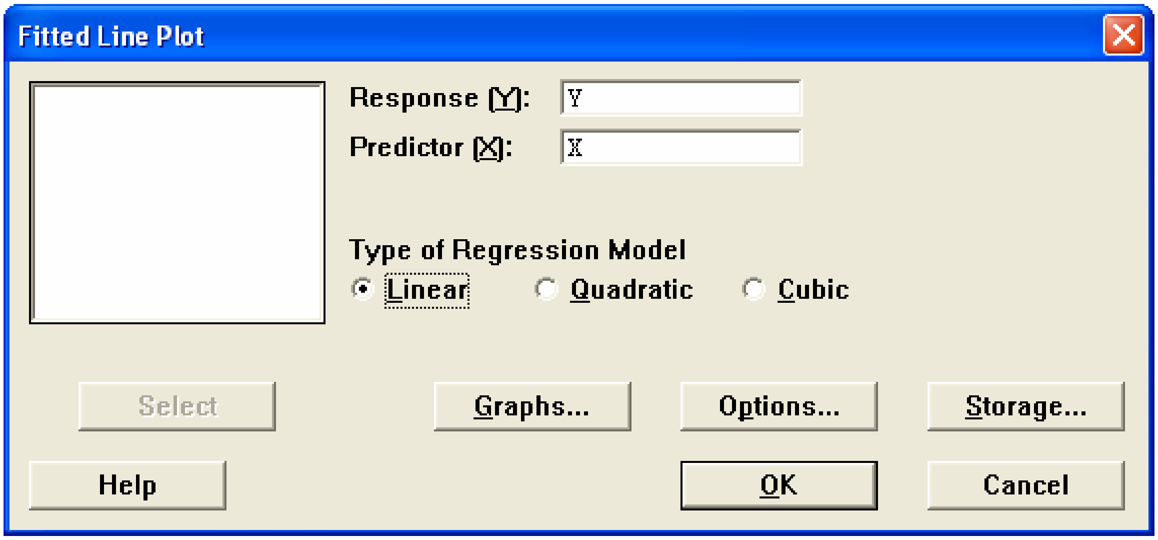


Repeating analysis we have already performed in Excel:

**Explore the relationship between X and Y for the following data**

| X | Y |
|---|---|
| 1 | 24 |
| 1 | 26 |
| 2 | 35 |
| 2 | 33 |
| 3 | 57 |
| 3 | 60 |
| 4 | 55 |
| 4 | 51 |
| 5 | 43 |
| 5 | 41 |

Select *Stat* > *Regression* > *Fitted Line Plot*.
Select the Type of Regression Model from the screen below:



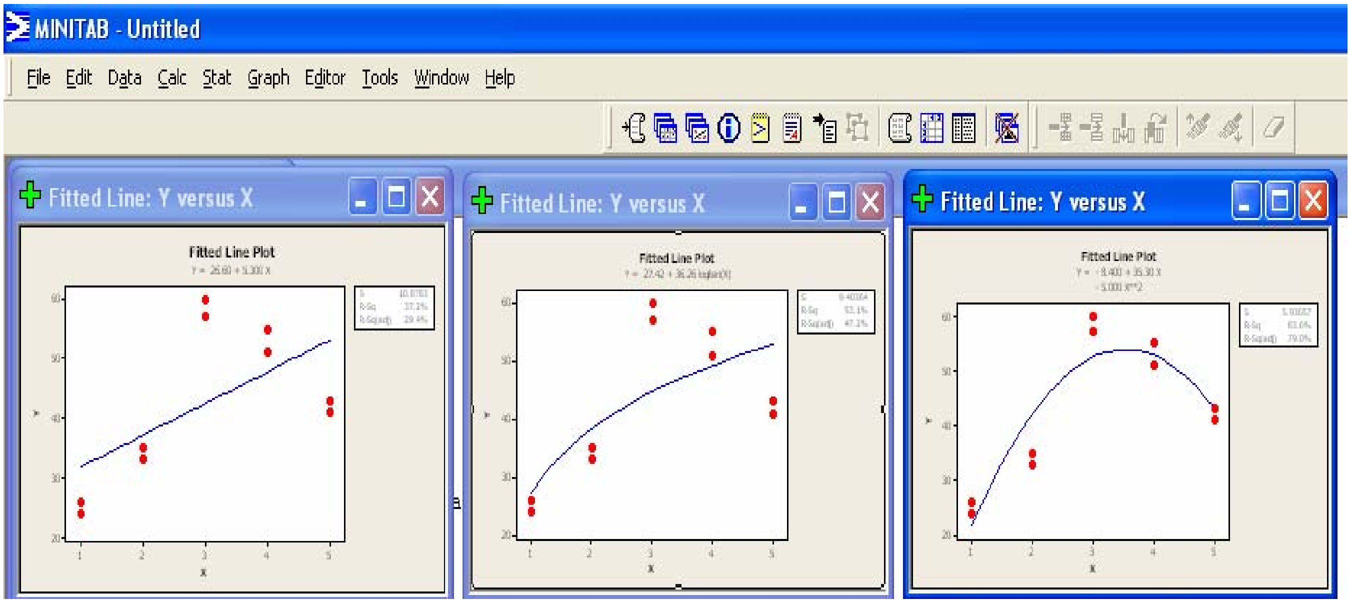
Response (Y): Y

Predictor (X): X

Type of Regression Model

● Linear ○ Quadratic ○ Cubic

We can produce the following graphs:



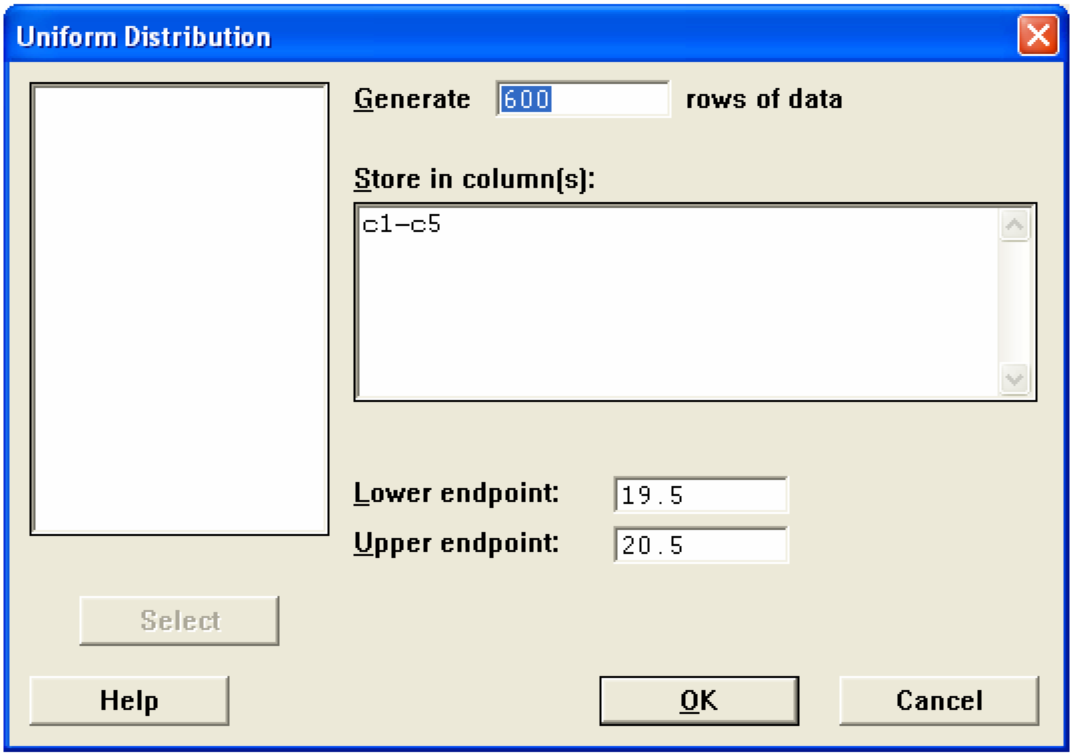
These graphs are the same as the ones produced using Excel.

**Central Limit Theorem Simulation in Minitab**

Repeat the process of generating 600 samples of 5 random values uniformly distributed between 19.5 and 20.5.

Select *Calc > Random Data > Uniform*

Complete the following screen:


Uniform Distribution

Generate 600 rows of data

Store in column(s):

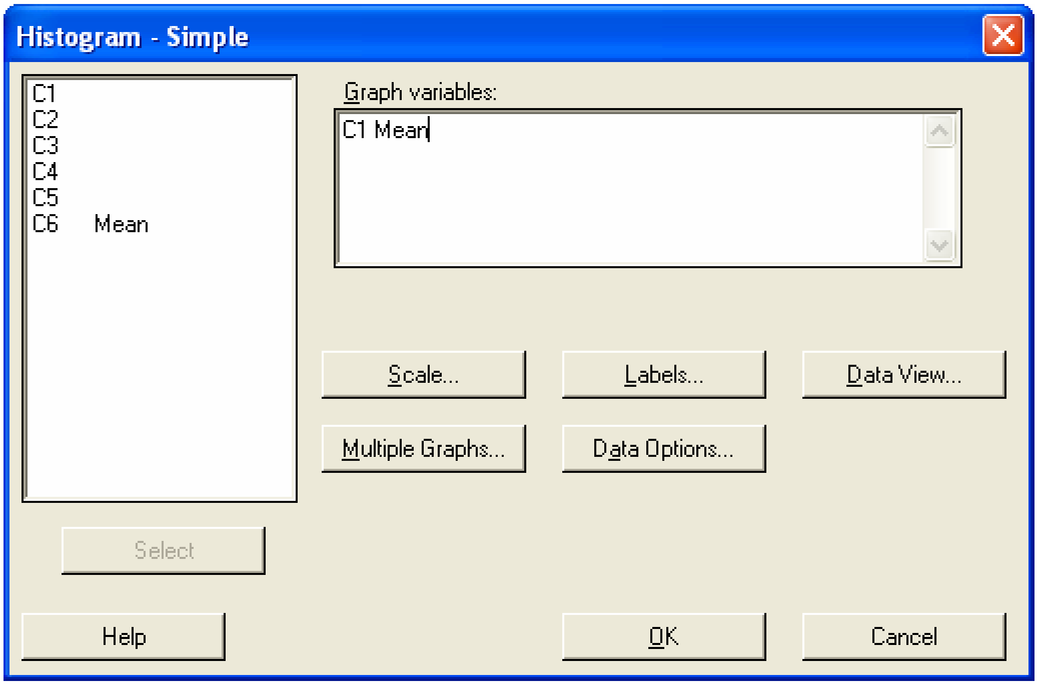
c1-c5

Lower endpoint: 19.5

Upper endpoint: 20.5

Next, graph one column of 600 values of the uniform distribution, and also graph the 600 values of the means of samples of size 5:

Select *Graph* > *Histogram* and complete the dialog box:

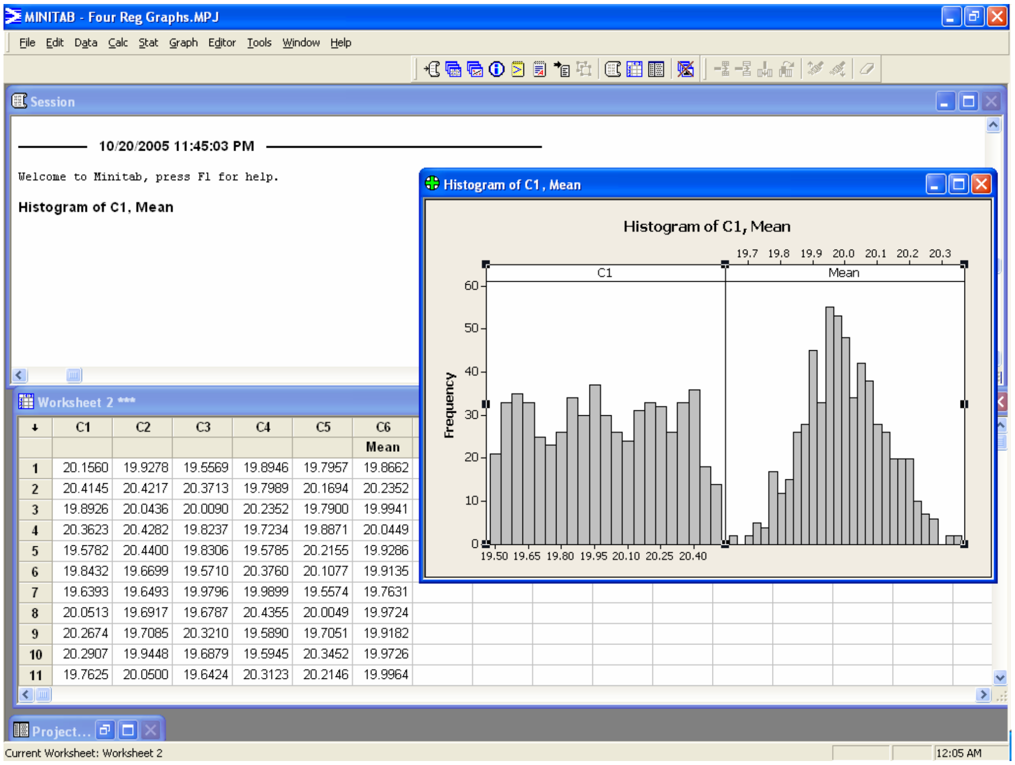

Histogram - Simple

C1
C2
C3
C4
C5
C6 Mean

Graph variables:

C1 Mean

You will see the two graphs below:



10/20/2005 11:45:03 PM

```
Welcome to Minitab, press F1 for help.
```

**Histogram of C1, Mean**

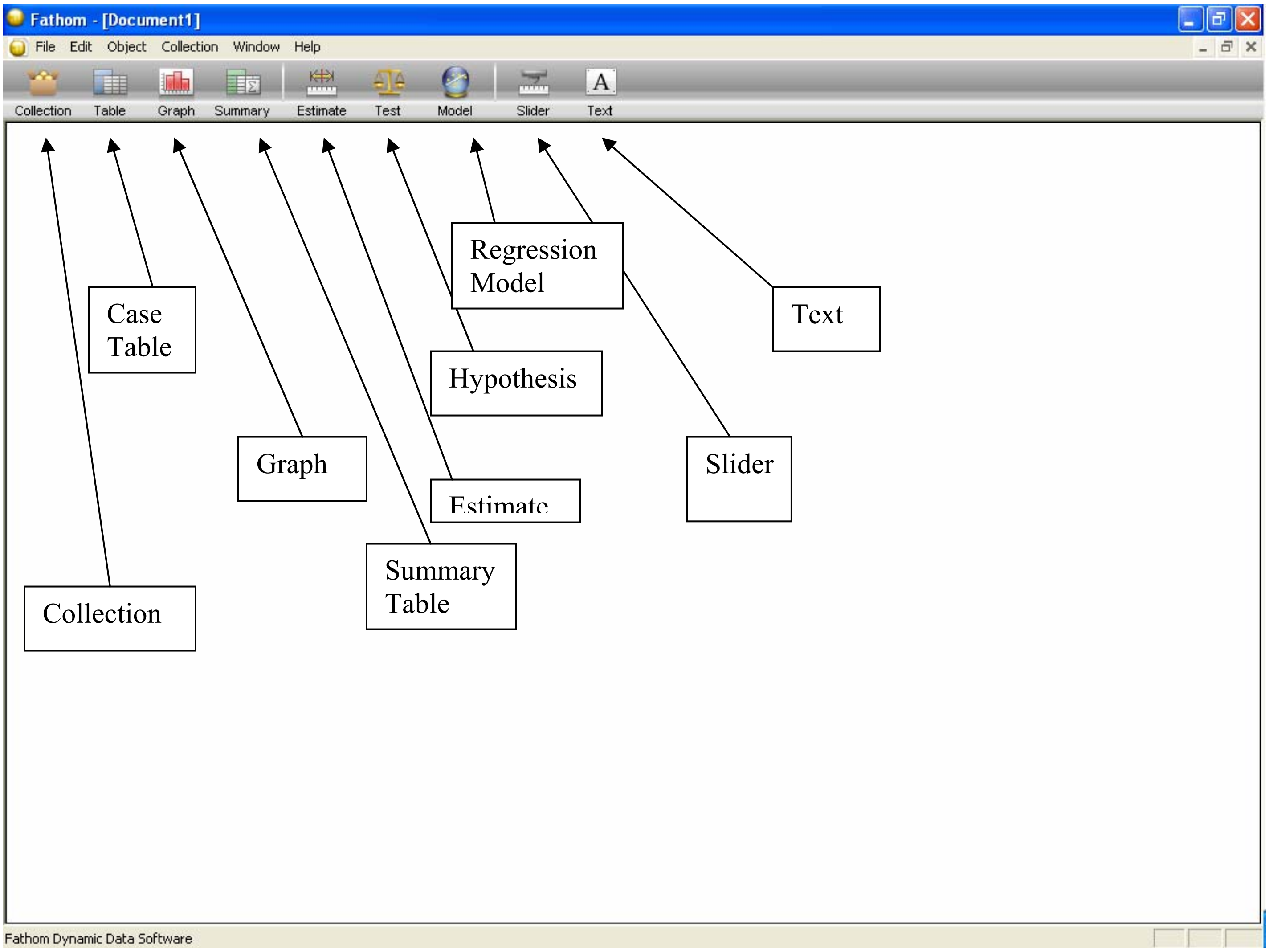
| ↓ | C1 | C2 | C3 | C4 | C5 | C6 |
|---|---|---|---|---|---|---|
| | | | | | | Mean |
| 1 | 20.1560 | 19.9278 | 19.5569 | 19.8946 | 19.7957 | 19.8662 |
| 2 | 20.4145 | 20.4217 | 20.3713 | 19.7989 | 20.1694 | 20.2352 |
| 3 | 19.8926 | 20.0436 | 20.0090 | 20.2352 | 19.7900 | 19.9941 |
| 4 | 20.3623 | 20.4282 | 19.8237 | 19.7234 | 19.8871 | 20.0449 |
| 5 | 19.5782 | 20.4400 | 19.8306 | 19.5785 | 20.2155 | 19.9286 |
| 6 | 19.8432 | 19.6699 | 19.5710 | 20.3760 | 20.1077 | 19.9135 |
| 7 | 19.6393 | 19.6493 | 19.9796 | 19.9899 | 19.5574 | 19.7631 |
| 8 | 20.0513 | 19.6917 | 19.6787 | 20.4355 | 20.0049 | 19.9724 |
| 9 | 20.2674 | 19.7085 | 20.3210 | 19.5890 | 19.7051 | 19.9182 |
| 10 | 20.2907 | 19.9448 | 19.6879 | 19.5945 | 20.3452 | 19.9726 |
| 11 | 19.7625 | 20.0500 | 19.6424 | 20.3123 | 20.2146 | 19.9964 |

Current Worksheet: Worksheet 2

Once again we see that the histogram of the original values is uniformly distributed while the mean values are approximately normally distributed.

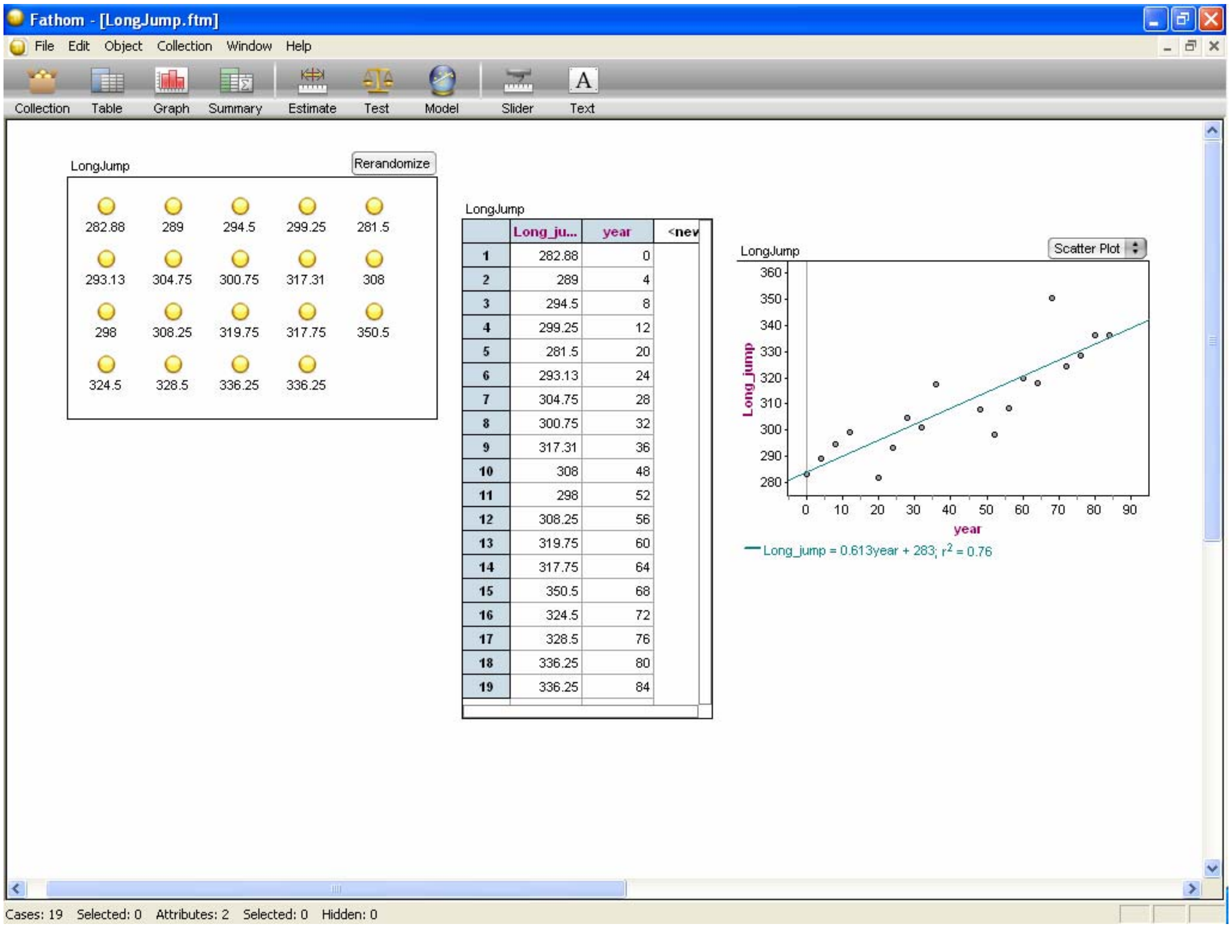
# **Fathom**

## Anatomy of a screen in Fathom 2™<br>Dynamic Data Software



The following screen shows three elements of a Fathom analysis for the long jump data:

(1) the collection<br>
(2) the case table<br>
(3) a graph (scatter plot with least squares regression line).



LongJump

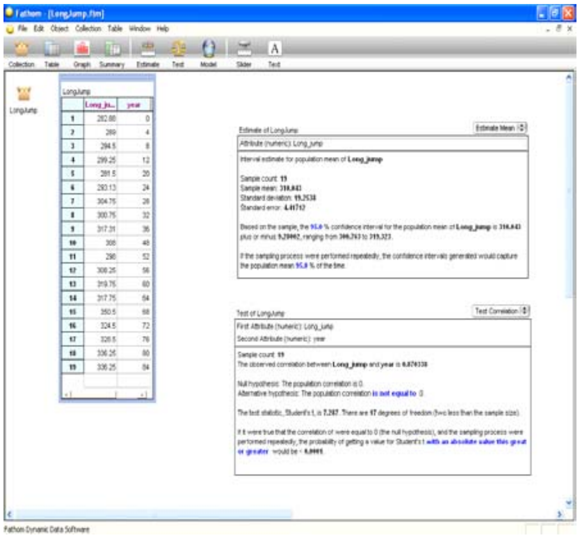
|  | Long_ju... | year | <new |
|---|---|---|---|
| 1 | 282.88 | 0 |  |
| 2 | 289 | 4 |  |
| 3 | 294.5 | 8 |  |
| 4 | 299.25 | 12 |  |
| 5 | 281.5 | 20 |  |
| 6 | 293.13 | 24 |  |
| 7 | 304.75 | 28 |  |
| 8 | 300.75 | 32 |  |
| 9 | 317.31 | 36 |  |
| 10 | 308 | 48 |  |
| 11 | 298 | 52 |  |
| 12 | 308.25 | 56 |  |
| 13 | 319.75 | 60 |  |
| 14 | 317.75 | 64 |  |
| 15 | 350.5 | 68 |  |
| 16 | 324.5 | 72 |  |
| 17 | 328.5 | 76 |  |
| 18 | 336.25 | 80 |  |
| 19 | 336.25 | 84 |  |

— Long_jump = 0.613year + 283; $r^2 = 0.76$

Cases: 19 Selected: 0 Attributes: 2 Selected: 0 Hidden: 0

Each of the elements above can be dragged into the document.

The next screen shows (1) a confidence interval for the mean distance jumped and (2) a test of hypothesis for the correlation between distance and year.


Fathom - [LongJump.ftm]
File Edit Object Collection Table Window Help
Collection Table Graph Summary Estimate Test Model Slider Text
LongJump
Estimate Mean
Test Correlation
Fathom Dynamic Data Software

Estimate of LongJump

Attribute (numeric): Long_jump

Interval estimate for population mean of **Long_jump**

Sample count: **19**
Sample mean: **310.043**
Standard deviation: **19.2534**
Standard error: **4.41702**

Based on the sample, the 95.0 % confidence interval for the population mean of **Long_jump** is **310.043** plus or minus **9.27962**, ranging from **300.763** to **319.323**.

If the sampling process were performed repeatedly, the confidence intervals generated would capture the population mean 95.0 % of the time.

Test of LongJump

First Attribute (numeric): Long_jump
Second Attribute (numeric): year

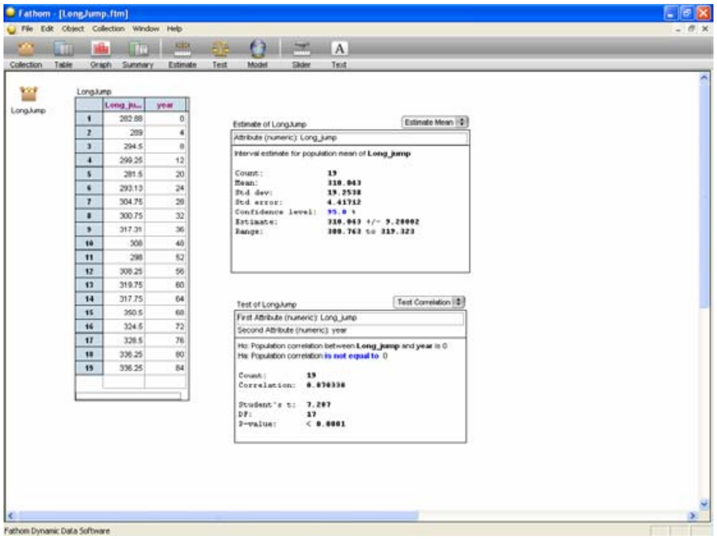
Sample count: **19**
The observed correlation between **Long_jump** and **year** is **0.870338**

Null hypothesis: The population correlation is 0.
Alternative hypothesis: The population correlation **is not equal to** 0

The test statistic, Student's t, is **7.287**. There are **17** degrees of freedom (two less than the sample size).

If it were true that the correlation of were equal to 0 (the null hypothesis), and the sampling process were performed repeatedly, the probability of getting a value for Student's t **with an absolute value this great or greater** would be < **0.0001**.

Exercise: repeat the confidence interval and test of hypothesis with *Verbose* toggled off.



LongJump

|  | Long_ju... | year |
|---|---|---|
| 1 | 282.88 | 0 |
| 2 | 289 | 4 |
| 3 | 294.5 | 8 |
| 4 | 299.25 | 12 |
| 5 | 281.5 | 20 |
| 6 | 293.13 | 24 |
| 7 | 304.75 | 28 |
| 8 | 300.75 | 32 |
| 9 | 317.31 | 36 |
| 10 | 308 | 48 |
| 11 | 298 | 52 |
| 12 | 308.25 | 56 |
| 13 | 319.75 | 60 |
| 14 | 317.75 | 64 |
| 15 | 350.5 | 68 |
| 16 | 324.5 | 72 |
| 17 | 328.5 | 76 |
| 18 | 336.25 | 80 |
| 19 | 336.25 | 84 |

Estimate of LongJump

Attribute (numeric): Long_jump

Interval estimate for population mean of **Long_jump**

```
Count:              19
Mean:               310.043
Std dev:            19.2538
Std error:          4.41712
Confidence level:   95.0 %
Estimate:           310.043 +/- 9.28002
Range:              300.763 to 319.323
```

Test of LongJump

First Attribute (numeric): Long_jump

Second Attribute (numeric): year

Ho: Population correlation between **Long_jump** and **year** is 0

Ha: Population correlation **is not equal to** 0

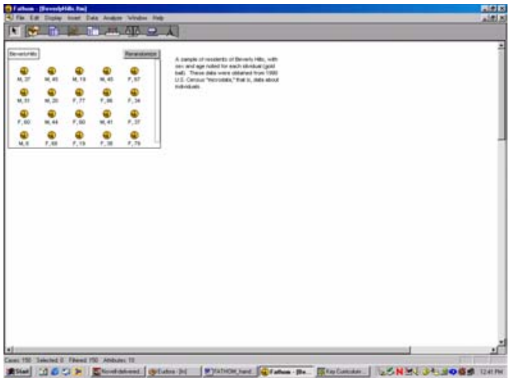
```
Count:          19
Correlation:    0.870338

Student's t:    7.287
DF:             17
P-value:        < 0.0001
```

## Exercise 1: Exploring Data

We use the file BeverlyHills to explore the data graphically.

- Investigate various graphs and see the impact of selecting an observation on all open graphs and the inspector.
- Investigate a scatter diagram of Income on Age
- Restrict age range (filter) to 18 – 80.
- See impact of the filter on the regression line.
- Then try age range 18 - 55

File → Open → Sample Documents → Learning Starter Guides → BeverlyHills


Fathom - [BeverlyHills.ftm]
File Edit Display Insert Data Analyze Window Help
BeverlyHills
Rerandomize

A sample of residents of Beverly Hills, with sex and age noted for each individual (gold ball). These data were obtained from 1990 U.S. Census "microdata," that is, data about individuals.

The screen displays two entities:

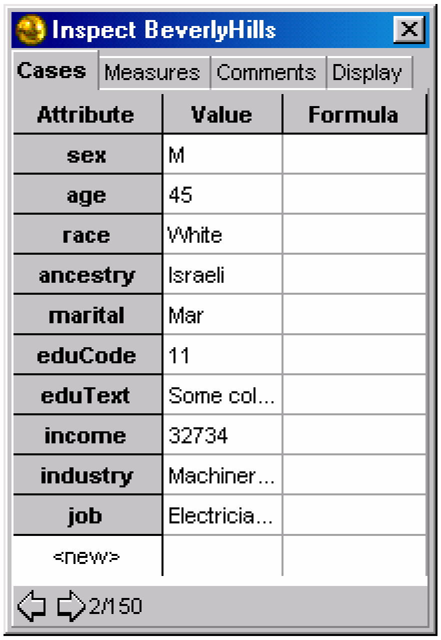
1. A "collection"
2. A text box describing the collection

Each gold ball in the collection represents one case.

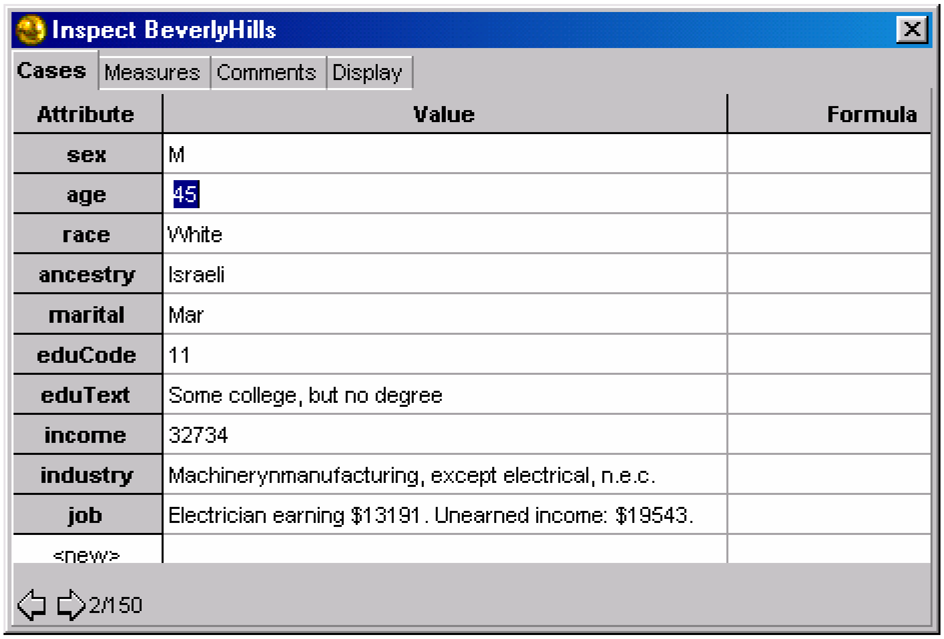
When the cursor is placed over the collection the status bar in the bottom left-hand corner of the screen indicates that the collection has 150 cases and 10 attributes (variables).

To explore the attributes double click the collection. You will see the Inspector screen as below:

By clicking on a gold ball in the collection you can inspect the attributes for that case in the Inspector. The screen below shows the attributes for the 2nd case in the collection. Note that it is necessary to resize the Inspector to see the full description of some of the attributes.



| Attribute | Value | Formula |
|---|---|---|
| sex | M | |
| age | 45 | |
| race | White | |
| ancestry | Israeli | |
| marital | Mar | |
| eduCode | 11 | |
| eduText | Some col... | |
| income | 32734 | |
| industry | Machiner... | |
| job | Electricia... | |
| <new> | | |



| Attribute | Value | Formula |
|---|---|---|
| sex | M | |
| age | 45 | |
| race | White | |
| ancestry | Israeli | |
| marital | Mar | |
| eduCode | 11 | |
| eduText | Some college, but no degree | |
| income | 32734 | |
| industry | Machinerynmanufacturing, except electrical, n.e.c. | |
| job | Electrician earning $13191. Unearned income: $19543. | |
| <new> | | |

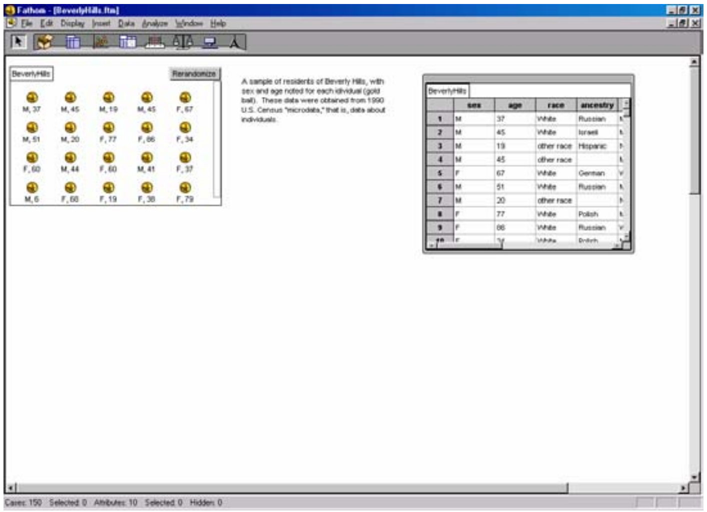
Click on the **Comments** tab and you will see the description of the data set. This is a good place to store documentation and notes about the data.

Close the Inspector by clicking its close box.

## For spreadsheet fans!

Click on the collection to select it (you will see a border around it).

To create a **Case Table** (spreadsheet), choose **Case Table** from the **Insert** menu or drag a new **Case Table** into the document.



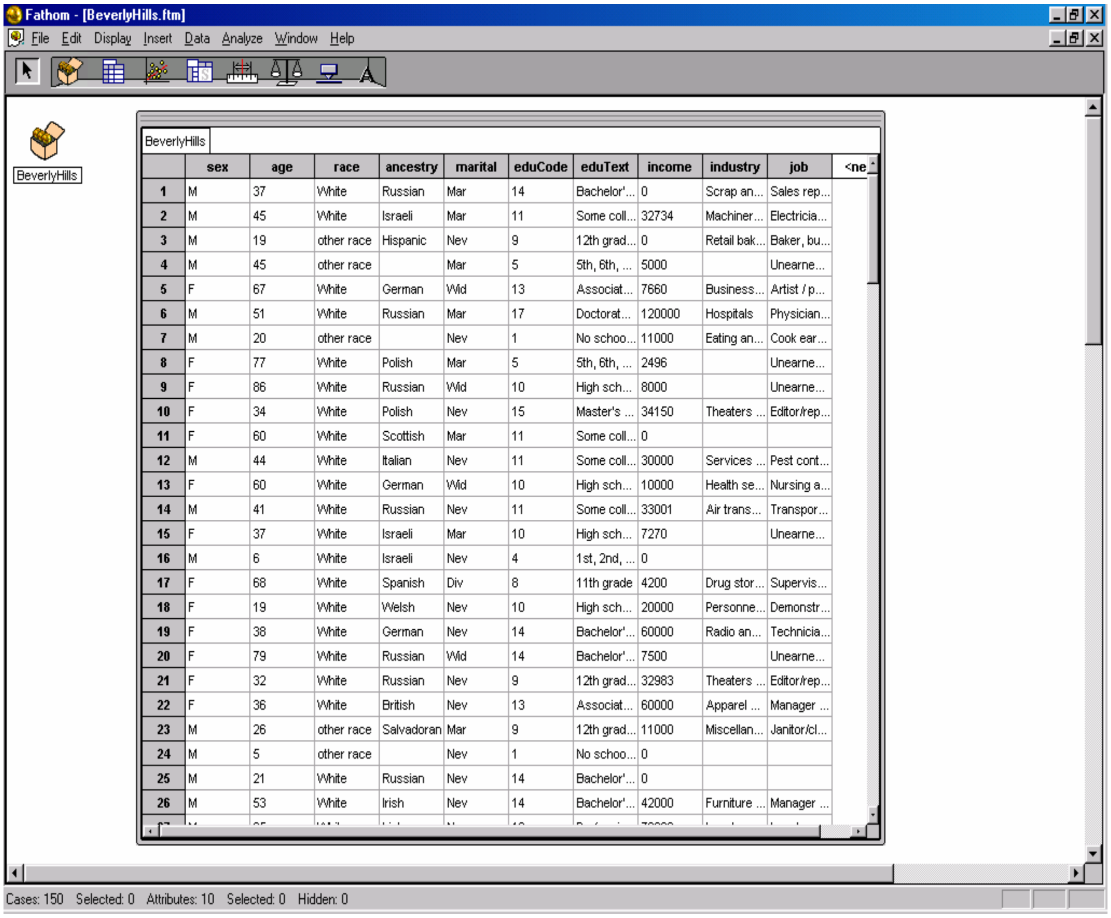
A sample of residents of Beverly Hills, with sex and age noted for each idividual (gold ball). These data were obtained from 1990 U.S. Census "microdata," that is, data about individuals.

| | sex | age | race | ancestry |
|---|---|---|---|---|
| 1 | M | 37 | White | Russian |
| 2 | M | 45 | White | Israeli |
| 3 | M | 19 | other race | Hispanic |
| 4 | M | 45 | other race | |
| 5 | F | 67 | White | German |
| 6 | M | 51 | White | Russian |
| 7 | M | 20 | other race | |
| 8 | F | 77 | White | Polish |
| 9 | F | 86 | White | Russian |

Cases: 150 Selected: 0 Attributes: 10 Selected: 0 Hidden: 0

The **Case Table** can be enlarged by dragging on a bottom corner of the table.

To see the full table we close the text box, shrink the collection until it is "iconized" and enlarge the **Case Table** by dragging on a bottom corner of the table.


Fathom - [BeverlyHills.ftm]
File Edit Display Insert Data Analyze Window Help
BeverlyHills
BeverlyHills

| | sex | age | race | ancestry | marital | eduCode | eduText | income | industry | job | <ne |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | M | 37 | White | Russian | Mar | 14 | Bachelor'... | 0 | Scrap an... | Sales rep... | |
| 2 | M | 45 | White | Israeli | Mar | 11 | Some coll... | 32734 | Machiner... | Electricia... | |
| 3 | M | 19 | other race | Hispanic | Nev | 9 | 12th grad... | 0 | Retail bak... | Baker, bu... | |
| 4 | M | 45 | other race | | Mar | 5 | 5th, 6th, ... | 5000 | | Unearne... | |
| 5 | F | 67 | White | German | Wid | 13 | Associat... | 7660 | Business... | Artist / p... | |
| 6 | M | 51 | White | Russian | Mar | 17 | Doctorat... | 120000 | Hospitals | Physician... | |
| 7 | M | 20 | other race | | Nev | 1 | No schoo... | 11000 | Eating an... | Cook ear... | |
| 8 | F | 77 | White | Polish | Mar | 5 | 5th, 6th, ... | 2496 | | Unearne... | |
| 9 | F | 86 | White | Russian | Wid | 10 | High sch... | 8000 | | Unearne... | |
| 10 | F | 34 | White | Polish | Nev | 15 | Master's ... | 34150 | Theaters ... | Editor/rep... | |
| 11 | F | 60 | White | Scottish | Mar | 11 | Some coll... | 0 | | | |
| 12 | M | 44 | White | Italian | Nev | 11 | Some coll... | 30000 | Services ... | Pest cont... | |
| 13 | F | 60 | White | German | Wid | 10 | High sch... | 10000 | Health se... | Nursing a... | |
| 14 | M | 41 | White | Russian | Nev | 11 | Some coll... | 33001 | Air trans... | Transpor... | |
| 15 | F | 37 | White | Israeli | Mar | 10 | High sch... | 7270 | | Unearne... | |
| 16 | M | 6 | White | Israeli | Nev | 4 | 1st, 2nd, ... | 0 | | | |
| 17 | F | 68 | White | Spanish | Div | 8 | 11th grade | 4200 | Drug stor... | Supervis... | |
| 18 | F | 19 | White | Welsh | Nev | 10 | High sch... | 20000 | Personne... | Demonstr... | |
| 19 | F | 38 | White | German | Nev | 14 | Bachelor'... | 60000 | Radio an... | Technicia... | |
| 20 | F | 79 | White | Russian | Wid | 14 | Bachelor'... | 7500 | | Unearne... | |
| 21 | F | 32 | White | Russian | Nev | 9 | 12th grad... | 32983 | Theaters ... | Editor/rep... | |
| 22 | F | 36 | White | British | Nev | 13 | Associat... | 60000 | Apparel ... | Manager ... | |
| 23 | M | 26 | other race | Salvadoran | Mar | 9 | 12th grad... | 11000 | Miscellan... | Janitor/cl... | |
| 24 | M | 5 | other race | | Nev | 1 | No schoo... | 0 | | | |
| 25 | M | 21 | White | Russian | Nev | 14 | Bachelor'... | 0 | | | |
| 26 | M | 53 | White | Irish | Nev | 14 | Bachelor'... | 42000 | Furniture ... | Manager ... | |

Cases: 150 Selected: 0 Attributes: 10 Selected: 0 Hidden: 0

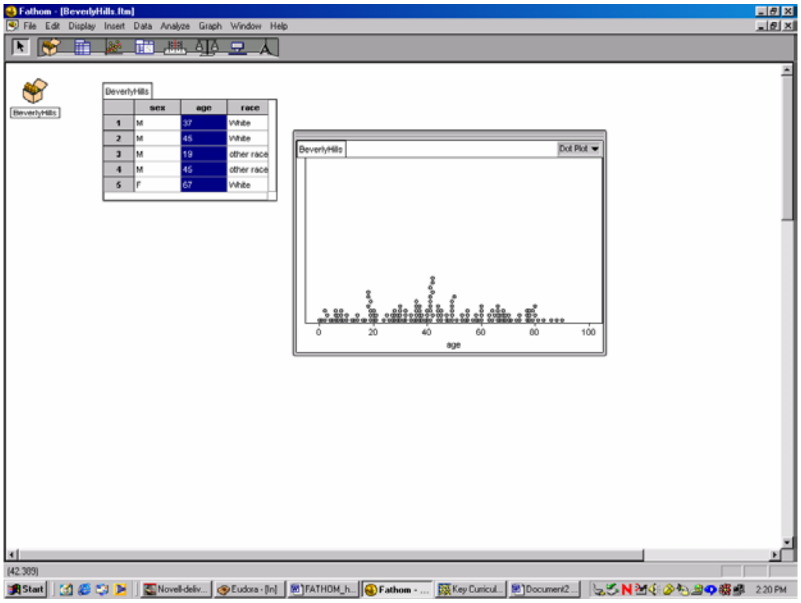
# Creating a Graph

Insert a graph into the document by (1) clicking on **Insert** → **Graph** (2) pressing CTRL-G or (3) dragging a new graph into the document

We will make a graph of the attribute **age**.

Drag the **age** attribute from the **Case Table** and drop it onto the horizontal axis of the graph, where you see the message “drop an attribute here”.

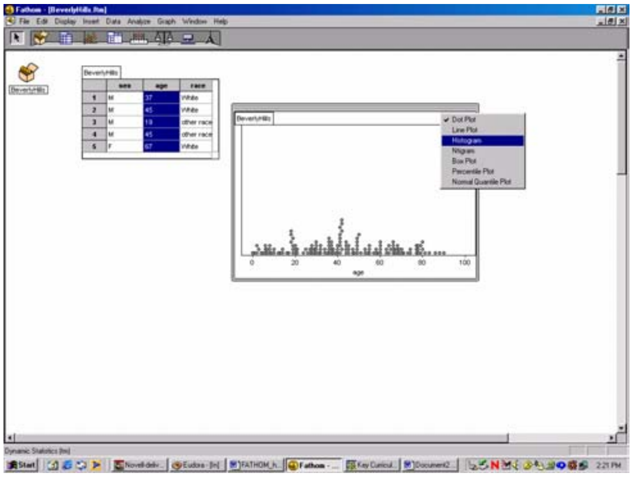
A Dot Plot of **age** will now be displayed.

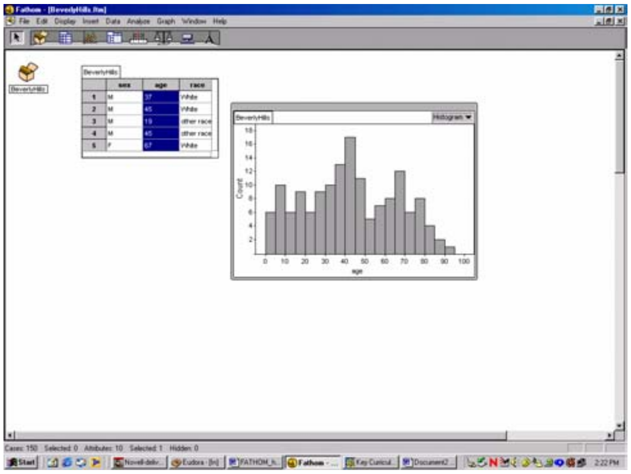


BeverlyHills

|   | sex | age | race |
|---|---|---|---|
| 1 | M | 37 | White |
| 2 | M | 45 | White |
| 3 | M | 19 | other race |
| 4 | M | 45 | other race |
| 5 | F | 67 | White |

This graph can be changed to a histogram by choosing Histogram from the popup menu as seen below:



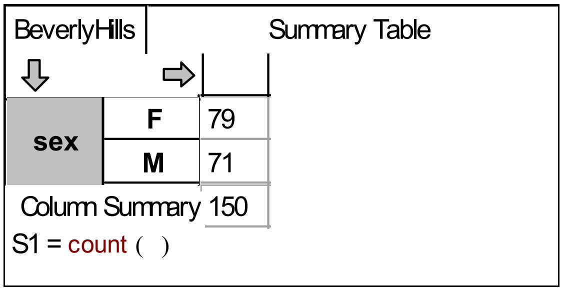


Fathom - [BeverlyHills.ftm]
File Edit Display Insert Data Analyze Graph Window Help
BeverlyHills
Histogram
Count
18
16
14
12
10
8
6
4
2
0
10
20
30
40
50
60
70
80
90
100
age
Cases: 150 Selected: 0 Attributes: 10 Selected: 1 Hidden: 0
2:22 PM

Experiment with dragging other variables onto the horizontal axes of the graph e.g. sex (how many of the respondents are male? What percentage is female?), **marital** (what percentage is divorced? Never married?)

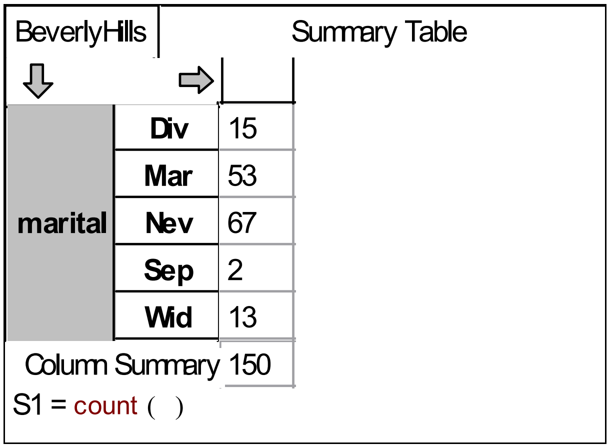
## Making a Summary Table

Drag a **Summary Table** into the document. Drag the attribute **sex** onto the down-pointing arrow to obtain the first graph below. The second graph is generated by dragging the attribute **marital** onto the down arrow.



| | | |
|---|---|---|
| sex | F | 79 |
| | M | 71 |
| Column Summary | | 150 |

S1 = count ( )

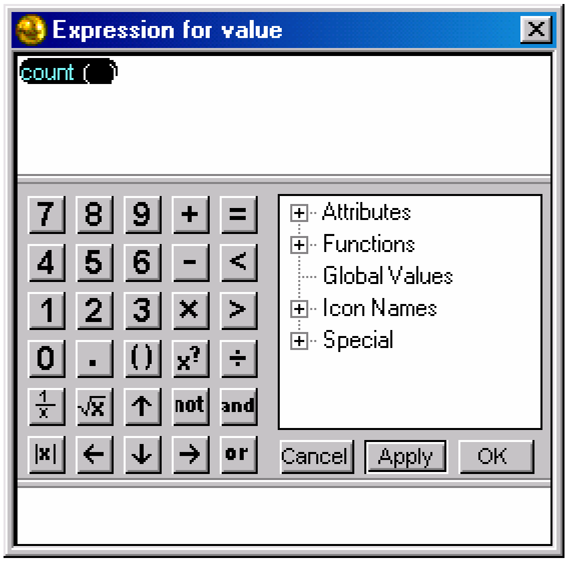


| | | |
|---|---|---|
| marital | Div | 15 |
| | Mar | 53 |
| | Nev | 67 |
| | Sep | 2 |
| | Wid | 13 |
| Column Summary | | 150 |

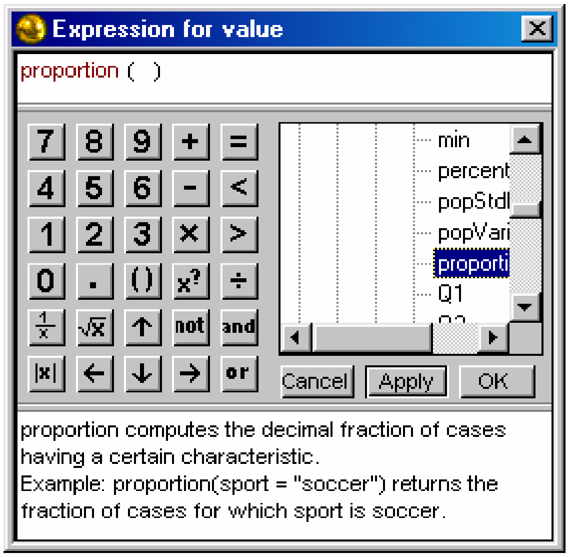
S1 = count ( )

When showing a categorical variable the default display is the number of cases in each group.

Other statistics can be displayed by changing the formula. For example to display the proportion of cases in each category, double-click on the formula S1 = count( ) to open the formula editing box. You can then edit the formula to display the proportion.



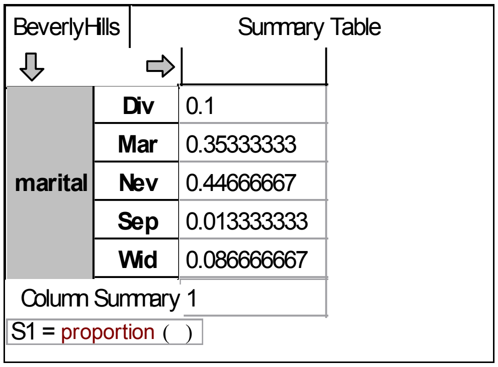
count ( )



proportion ( )

proportion computes the decimal fraction of cases having a certain characteristic.
Example: proportion(sport = "soccer") returns the fraction of cases for which sport is soccer.

Click on OK to close the formula editor and the following results are displayed:

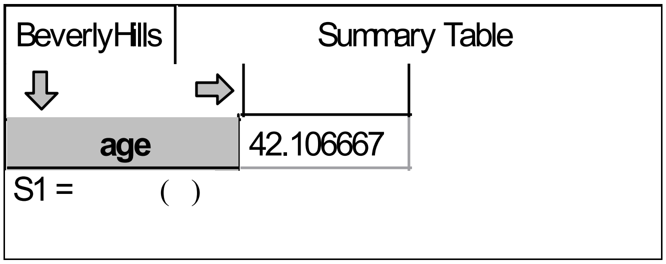


BeverlyHills | Summary Table

| | | |
|---|---|---|
| marital | Div | 0.1 |
| | Mar | 0.35333333 |
| | Nev | 0.44666667 |
| | Sep | 0.013333333 |
| | Wid | 0.086666667 |
| Column Summary | | 1 |

S1 = proportion ( )

Close the **Summary Table** and open a new one. Drag the attribute **age** to the down arrow to obtain the following table.



BeverlyHills | Summary Table

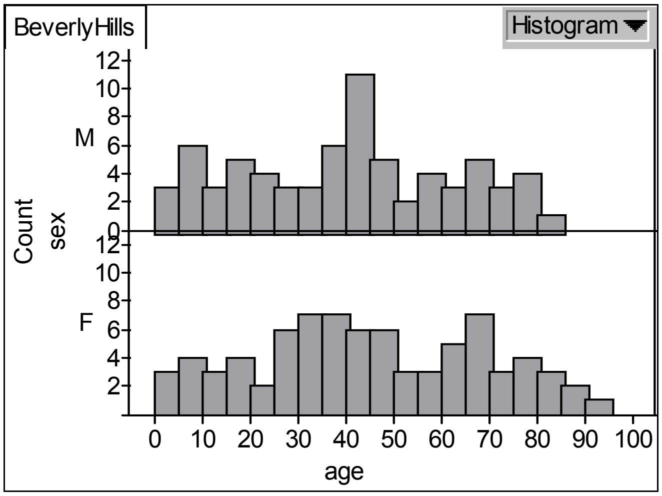
| | |
|---|---|
| age | 42.106667 |

S1 = ( )

Note that for a numerical attribute the default display is the mean. We see that the mean age of the people surveyed is approximately years 42.

## Splitting a graph

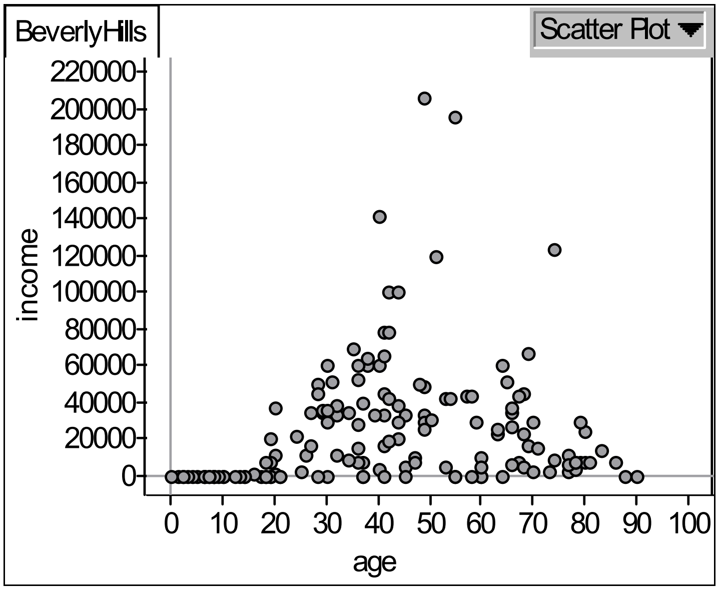
Drag a new graph into the document and drop the **age** attribute onto the horizontal axis. Change the graph to a histogram. Now drag the **sex** attribute onto the vertical axis. You will now see that the histogram of **age** has been split on the attribute **sex**.



We can now compare the distribution of age for male and female respondents.

## Graphing Two Continuous Numeric Attributes

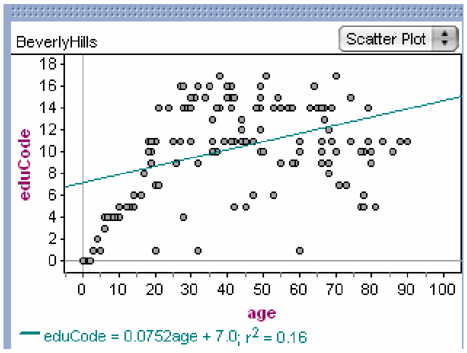
To create a graph of **income** vs **age**, drag the attribute income onto the vertical axes of the graph age is already on the horizontal axis.)  The following scatter diagram will be displayed.



We note that the under 20 year olds have almost no income, the highest incomes are earned by people in the 40 – 60 age group, and incomes decrease for older people.

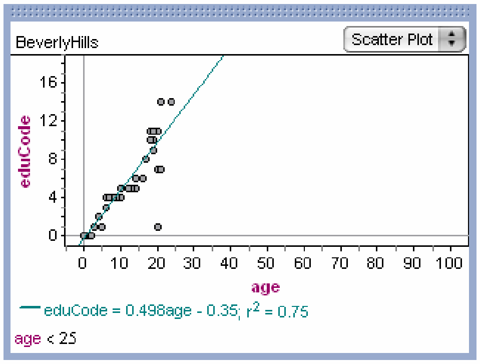
Now replace *income* on the vertical axis with *educode*. We will use the new graph to illustrate the concepts of **Outliers and Influential Observations.** A point which lies far from the line (and thus has a large residual value) is known as an ***outlier***. Such points may represent erroneous data, or may indicate a poorly fitting regression line. If a point lies far from the other data in the horizontal direction, it is known as an ***influential observation***. The reason for this distinction is that these points may have a significant impact on the slope of the regression line.

The next graph shows a scatter diagram of Educode on Age.



Note that the $r^2$ value is 0.16, implying that 16% of variation in Educode (Educational level) is explained by Age.

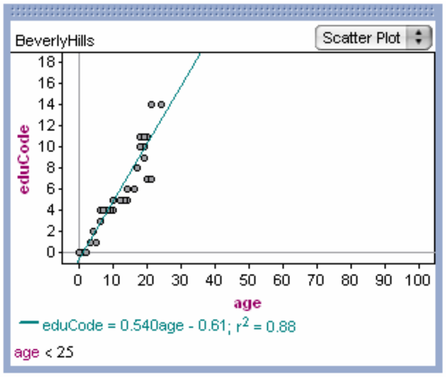
Now apply the filter ***Age < 25*** to produce the following graph:



We see that in the age range ***0 - 25*** the $r^2$ value increases to $r^2 = 0.75$ (for obvious reasons!).

Notice that there is one "unusual" observation corresponding to a twenty year old male with an Educode of 1 ("no school completed"). He is working as a cook and earning $11,000. This observation is an outlier.

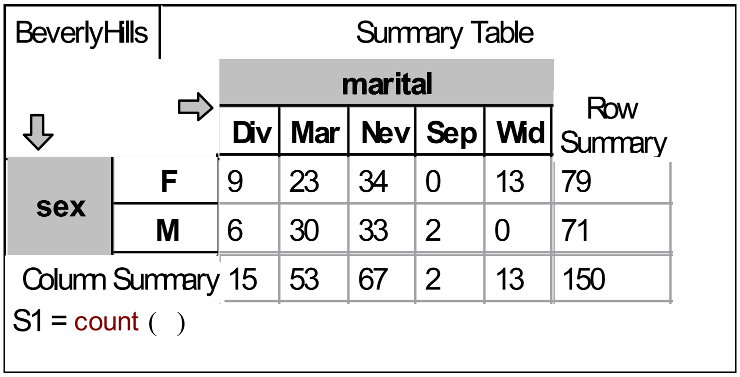
When this observation is removed we see that the $r^2$ value increases to 0.88.



Question: How would you write up a report on this data set? Would you include the outlier and report a coefficient of determination of 75%, or analyze the data set with the outlier removed, and include an exception report in your final statement?

**Contingency Tables (Cross Tabulation)**.

To create a contingency table of two categorical variable we drag a **Summary Table** into the document and then drop one of the attributes onto the horizontal axis, and the other onto the vertical axis. For example, the following table shows a cross-tabulation of **marital** status by **sex**.



BeverlyHills Summary Table

| | | marital | | | | | |
|---|---|---|---|---|---|---|---|
| | | Div | Mar | Nev | Sep | Wid | Row Summary |
| sex | F | 9 | 23 | 34 | 0 | 13 | 79 |
| | M | 6 | 30 | 33 | 2 | 0 | 71 |
| Column Summary | | 15 | 53 | 67 | 2 | 13 | 150 |

S1 = count ( )

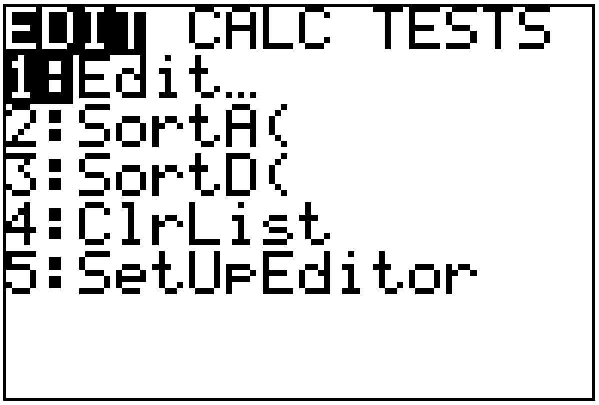
This example has introduced the basics - but there is lots more!! It is worthwhile to linger over this example for a while and experiment with other graphs and tables. For example we may wish to explore box plots and Ntigrams (for which cases are grouped into bins of equal sample size). Remember when you draw graph you can see the number (percentage ) of observation in each category by moving the cursor to the specified class and looking at the values in the status bar. Observe the ability to rescale the axes by dragging.

**<u>Other things to explore</u>**

- Test hypothesis that mean age = 40
    - Toggle **Verbose** on and off
- Establish A 95% confidence interval for mean age
- Investigate a multiple regression of Income on Age and Educode
    - Change order of independent variables and see effect on $r^2$

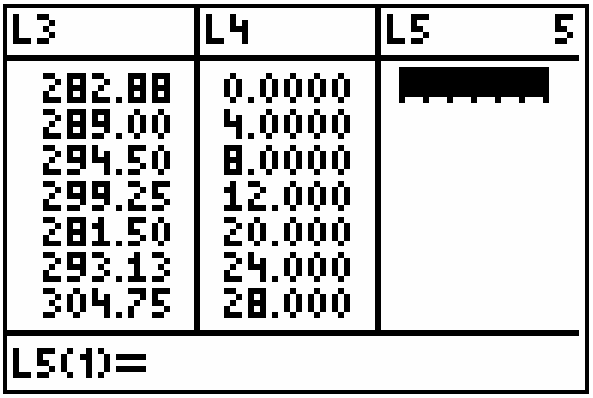
**TI-84 Plus SE**

## Press <Stat>



**Press <Enter> and scroll to the right.**

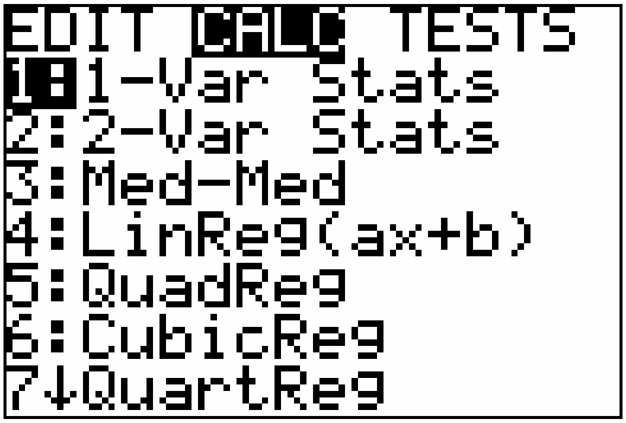
**Long Jump data is stored in L3, Year is stored in L4**



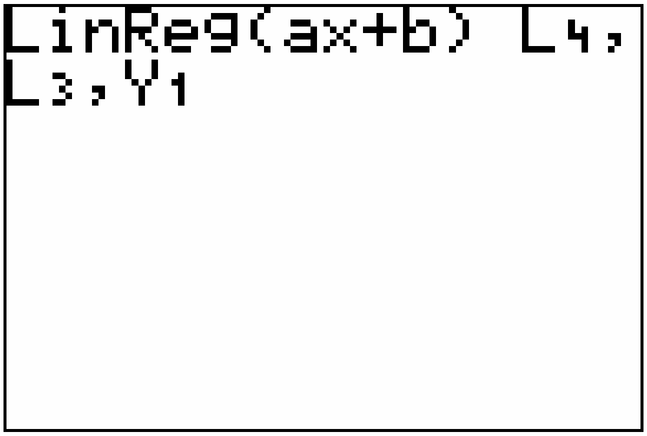
| L3 | L4 | L5 5 |
|---|---|---|
| 282.88 | 0.0000 | ------ |
| 289.00 | 4.0000 | |
| 294.50 | 8.0000 | |
| 299.25 | 12.000 | |
| 281.50 | 20.000 | |
| 293.13 | 24.000 | |
| 304.75 | 28.000 | |

L5(1)=

**Press <Stat>, Select CALC**

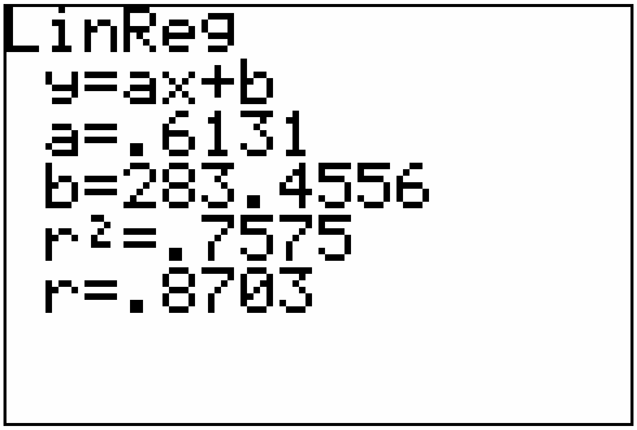


**Select 4: LinReg(ax+b)**



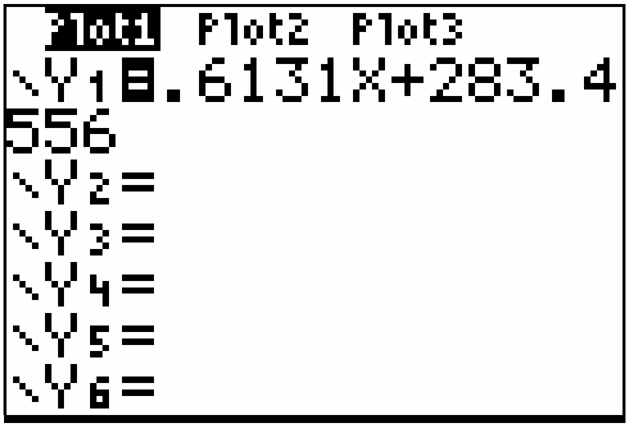
```
LinReg(ax+b) L₄,
L₃,Y₁
```

**Press <Enter>**



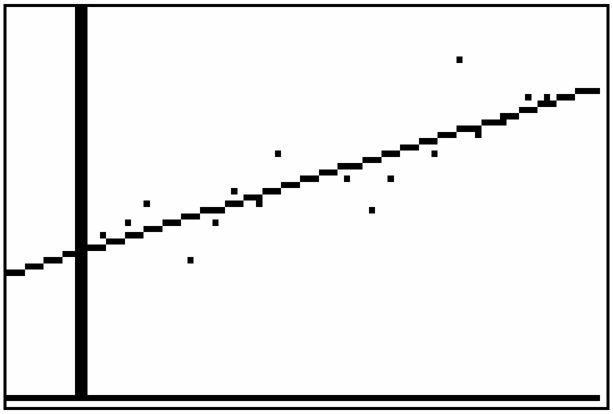
```
LinReg
 y=ax+b
 a=.6131
 b=283.4556
 r²=.7575
 r=.8703
```

**The regression equation is now stored in Var $Y_1$**



```
 Plot1  Plot2  Plot3
\Y1=.6131X+283.4
556
\Y2=
\Y3=
\Y4=
\Y5=
\Y6=
```

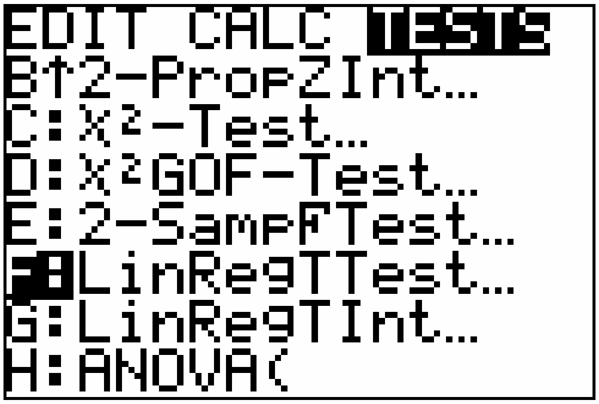
**Press <GRAPH> to see the regression line:**



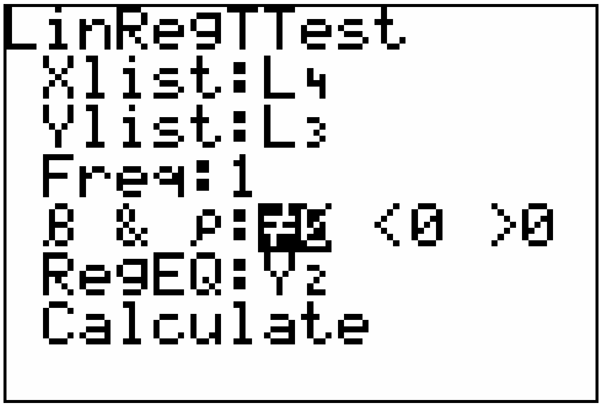
**To perform a test of hypothesis:**

**$H_O$ : □ = 0 vs $H_A$ : □ □ 0      or      $H_O$ : □ = 0 vs $H_A$ : □ □ 0**

**Press <STAT>, highlight TESTS, and scroll down to F:LinRegTTest…**

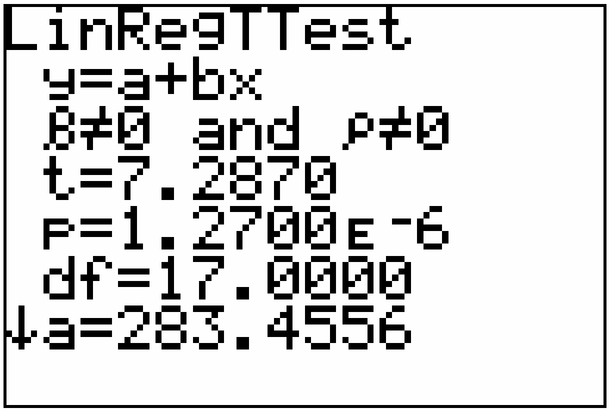


**Complete the screen as shown below:**



```
LinRegTTest
 Xlist:L4
 Ylist:L3
 Freq:1
 β & ρ:≠0 <0 >0
 RegEQ:Y2
 Calculate
```

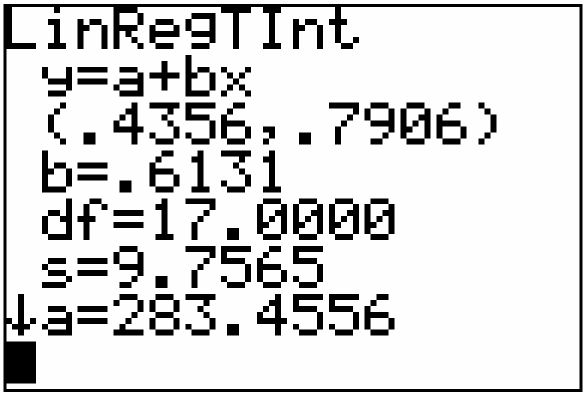
Select **Calculate and press <ENTER>**



```
LinRegTTest
 y=a+bx
 β≠0 and ρ≠0
 t=7.2870
 p=1.2700E-6
 df=17.0000
↓a=283.4556
```

We see that the computed t-value is 7.2870 and the p-value is 1.2700E$^{-}$6, therefore we reject $H_o$ and conclude that there is a significant linear relationship between Y (Distance) and X (Year).

To construct a 95% confidence interval for □ we select the test **G:LinRegTInt** and we get the following result:

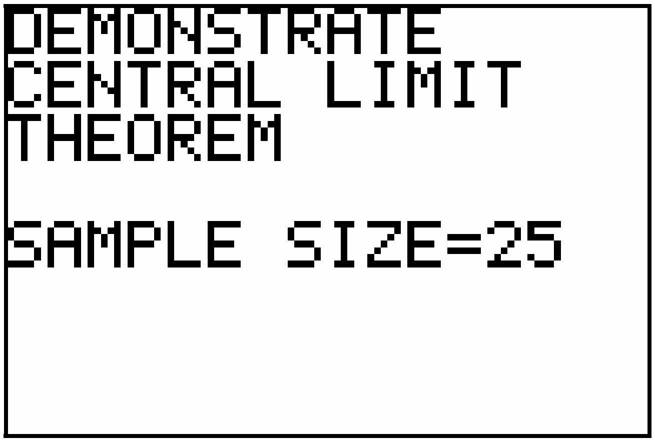


```
LinRegTInt
 y=a+bx
 (.4356,.7906)
 b=.6131
 df=17.0000
 s=9.7565
↓a=283.4556
```

The 95% confidence interval is 0.4356 ≤ □ ≤ 0.7906

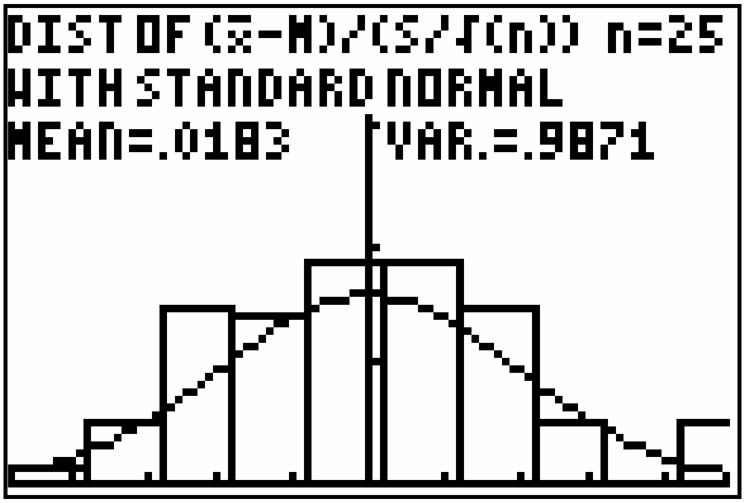
**<u>Demonstration of Central Limit Theorem</u>**

## Program clt.83p (https://www.ticalc.org/archives/files/fileinfo/99/9919.html



```
DEMONSTRATE
CENTRAL LIMIT
THEOREM

SAMPLE SIZE=25
```

$\overline{x}$ is being sampled 99 times
from the uniform distrib.
Sample Size = 25
Men =.5 Variance=.0833



DIST OF $(\overline{x}-M)/(S/\sqrt{(n)})$ n=25
WITH STANDARD NORMAL
MEAN=.0183 VAR.=.9871