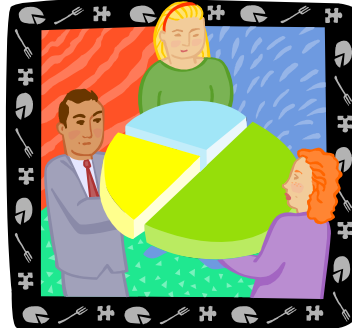
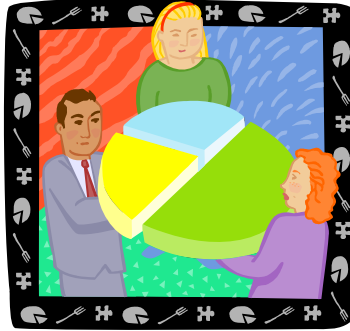


## DESCRIPTIVE STATISTICS



**Describing information with graphs and tables**  
**By Grace S. Thomson**

# DESCRIPTIVE STATISTICS



By Grace S. Thomson

This chapter contains 2 main topics related to available techniques to describe and interpret statistic data, using tables and graphs:

- How to build and interpret Frequency Distributions and Histograms
- How to build and interpret graphs and charts
  - Bar, Pie, Frequency distributions, Stem and leaf (Cross-sectional data)
  - Line and Scatter Diagrams (Time series data)

## TOPIC I

The first concept you will find here is Frequency distributions.

### **DEFINITION OF FREQUENCY DISTRIBUTION**

A frequency distribution is a summary set of data that describes information in an organized way, showing the number of observations in each class or category.

In other words, if you to describe the age composition of your class, or the income categories of your target customers, by counting them and writing these numbers in a table, you are building a Frequency distribution. It's called "frequency" because you count the number of times that a specific event occurs.

### **Elements of a frequency distribution:**

A frequency distribution table can be grouped or ungrouped.

Let's say that you have gathered data about ages for 20 students in your statistics class:

The information is as follows:

18 18 19 20 25 24 30 32 17 37 35 23 23 22 21 18 19 21 22 23

In order to build a frequency distribution table you may choose to count how many of students there are for each age, or you may group the ages by age groups.

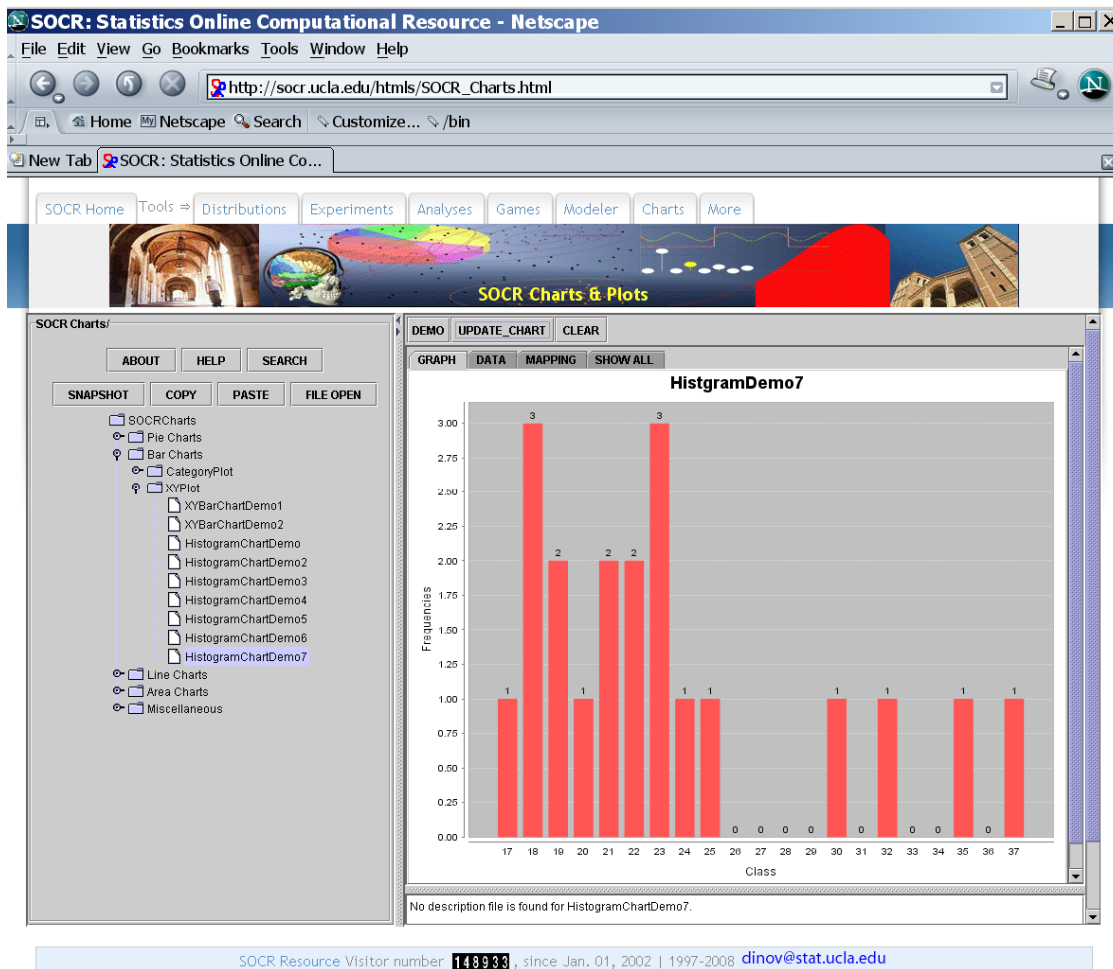
Let's take a look at the following frequency distribution table:

**Example of Frequency Distribution  
Age groups in a Statistics Class  
August 2004**

Class	fi (frequency)
17	1
18	3
19	2
20	1
21	2
22	2
23	3
24	1
25	1
26	0
27	0
28	0
29	0
30	1
31	0
32	1
33	0
34	0
35	1
36	0
37	1
Total (n)	20

This table groups the information but it's not practical for purposes of analysis. In order to be more efficient in building these tables, it is recommended to group the information in classes.

**For example:** Go to SOCR Charts ([http://socr.ucla.edu/htmls/SOCR\\_Charts.html](http://socr.ucla.edu/htmls/SOCR_Charts.html)) and select the Histogram Chart, as shown below. Copy and paste your data in SOCR, click the Update Graph button and go to the Graph Panel to see how this graph looks:



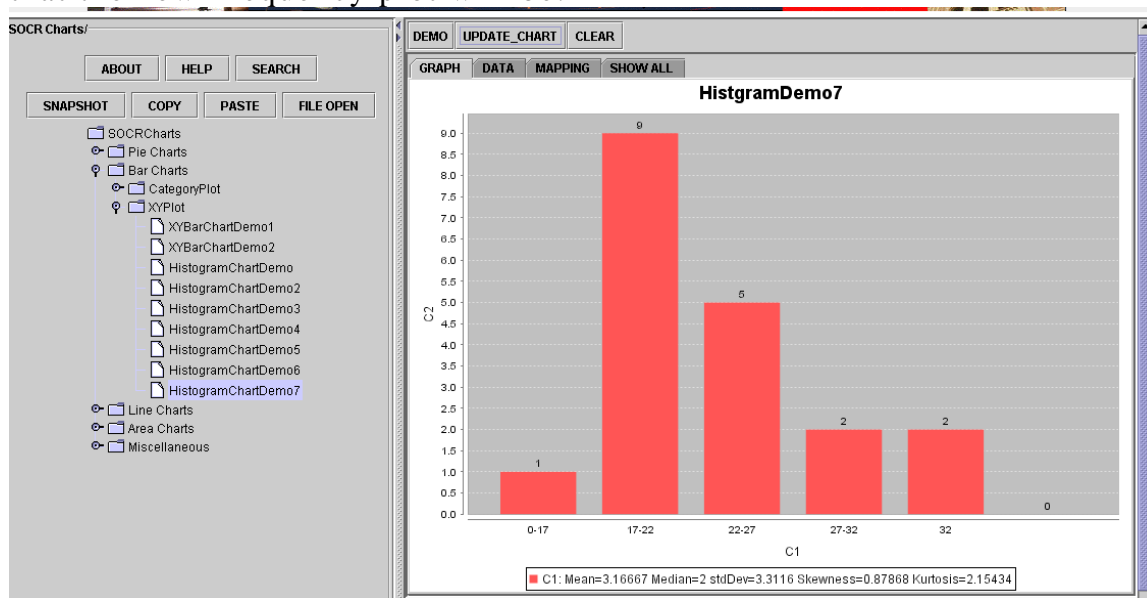
**Classes** → Categories that you will use to classify your observations. You may have as many classes as you wish, but there are some rules that will help you define the number of classes you require. e.g. 0-18 years, 19-36 years, and so on.

**Frequency (f)** → # of elements included in each class. e.g. number of students who are between 0-18 years old, 19-36 and so on. The sum of these frequencies must total your number of observations e.g. 20 students.

**Example of Frequency Distribution  
Age groups in a Statistics Class  
August 2004**

Age groups	# students
0-17	1
17-22	9
22-27	5
27-32	2
32	2
<b>Total (n)</b>	<b>20</b>

You can see again with SOCR Charts ([http://socr.ucla.edu/htmls/SOCR\\_Charts.html](http://socr.ucla.edu/htmls/SOCR_Charts.html)) that the new Frequency plot will be:



You see how simple it is? Now, you can build frequency tables for any variable or event you might be interested in.

Consider that in Statistics there are 2 types of variables according to their measurement characteristics: Discrete and Continuous.

### Type of variables

You can build frequency distributions for both Discrete and Continuous variables.

A discrete variable is any variable that you can count and measure. Good examples of a discrete variable are: number of students, # of customers etc. When count and measurement take infinite numbers, the variable is considered to be continuous. Examples of continuous variables are time, height, weight, age income, prices, etc.

### Type of Frequency distributions

Frequency distributions may be absolute or relative

**Absolute:** Frequency Calculated by counting the number of observations for each class. This is exactly what we did before.

**Example of Frequency Distribution  
Age groups in a Statistics Class  
August 2004**

<i>Classes</i> Age groups	<b>fi</b> # students
0-17	1
17-22	9
22-27	5
27-32	2
32	2
<b>Total (n)</b>	<b>20</b>

**Relative:** Sometimes it is necessary to express data in percentage, especially when you are building frequency distributions for 2 sets of data with different number of observations.

So a relative frequency is calculated as a percentage of the total observations. It's called pi and results from dividing each occurrence by the total number of observations:

$$pi = fi/n$$

**Example of Frequency Distribution  
Age groups in a Statistics Class  
(absolute and relative frequencies)  
August 2004**

<b>Class</b> Age groups	<b>fi</b> # students	<b>pi</b> %
0-17	1	5%
17-22	9	45%
22-27	5	25%
27-32	2	10%
32	2	10%
<b>Total (n)</b>	<b>20</b>	<b>100%</b>

One of the most important elements when building a distribution table is how to structure it in terms of the number of classes or categories. In our previous example,

we had 3 classes: 0-18, 18-36 and +36 year old, but in general we need to consider the following rules about classes:

1. Classes are mutually exclusive (classes can't overlap)
2. Classes are all inclusive (must include all the possible values)
3. Classes must have equal-width (between the highest and lowest number there is always the same width)
4. We should avoid empty classes (usually when classes are too narrow)

### **How to group frequency into classes?**

The process is simple:

- 1. Determine # of class you want to create.**

There is a simple rule, called Sturges' rule:

$$\text{\# of classes : } 1 + 3.322 \times \text{Log}(n)$$

For a sample of 20 elements, for example, the number of classes that you could build is:

$$\text{\# of classes: } 1 + 3.322 \times \text{Log}(20) \rightarrow 5.32 \text{ or } 5$$

Use your calculator to calculate the logarithmic base 10. Round up to the highest number.

- 2. Establish class width**

After you compute the number of groups that you may have, decide what is the width of each group, using the following formula:

$$\text{Width} = \frac{\text{largest value} - \text{smallest value}}{\text{\# of classes}}$$

If your data has 20 elements with a highest value of 37, and a smallest value of 17, the width should be:

$$W = (37 - 17) / 5 = 4$$

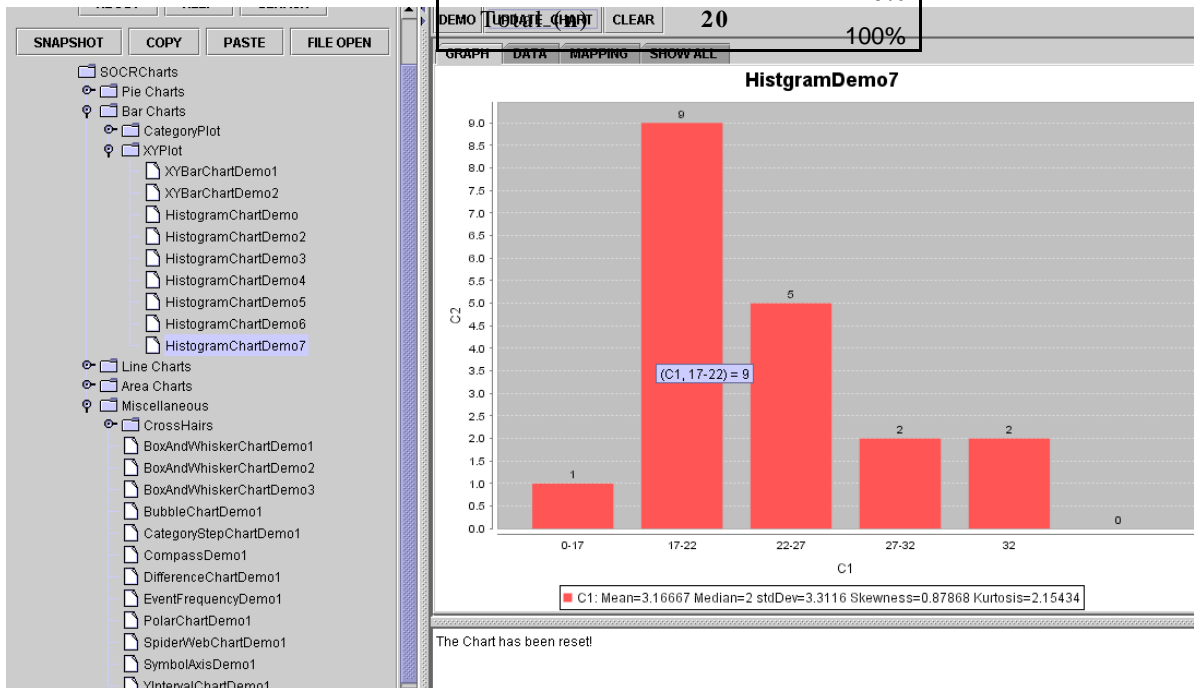
- 3. Build your table with these classes, by counting the number of elements in each category.**

We start with 17 and add 4 units to build the ranges. Notice how the next classes are built based on this.

**Frequency  
Classes were built**

Class	fi	pi
Age groups	# students	%
0-17	1	5%
17-22	9	45%
22-27	5	25%
27-32	2	10%
32	2	10%
	<b>20</b>	<b>100%</b>

**Distribution table  
with Sturges's rule**



**HISTOGRAMS**

Histograms are graphic representations of the Frequency distributions, it's easier to look at these graphs and understand the characteristics of the data. Following is the histogram of our table about age distribution used above:

**Using SOCR to build your Histogram:**

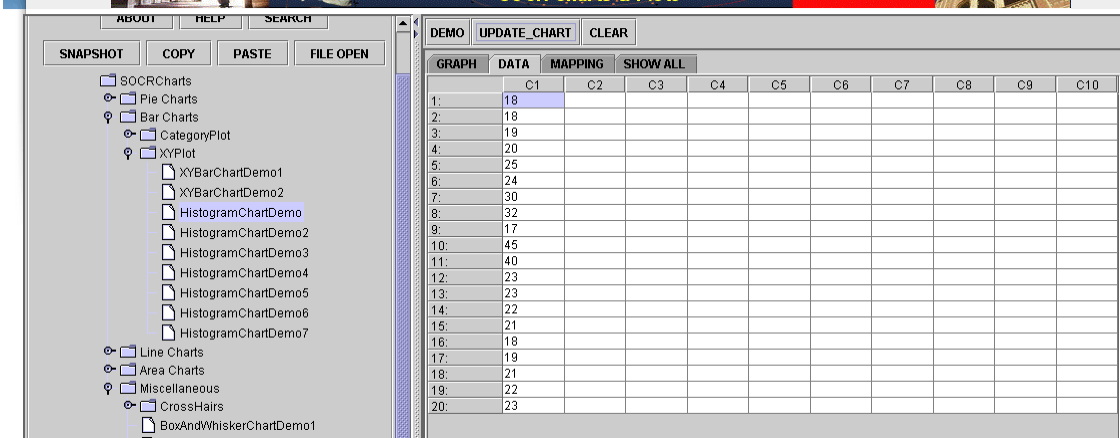
You can also use SOCR to build histograms using raw data. Let's say for example that you have the information about the ages of the 20 students of your statistics class, as follows:

18 18 19 20 25 24 30 32 17 45 40 23 23 22 21 18 19 21 22 23

- Go to SOCR Charts ([http://socr.ucla.edu/htmls/SOCR\\_Charts.html](http://socr.ucla.edu/htmls/SOCR_Charts.html))
- Choose the HistogramChartDemo ← XYPlot ← BarCharts

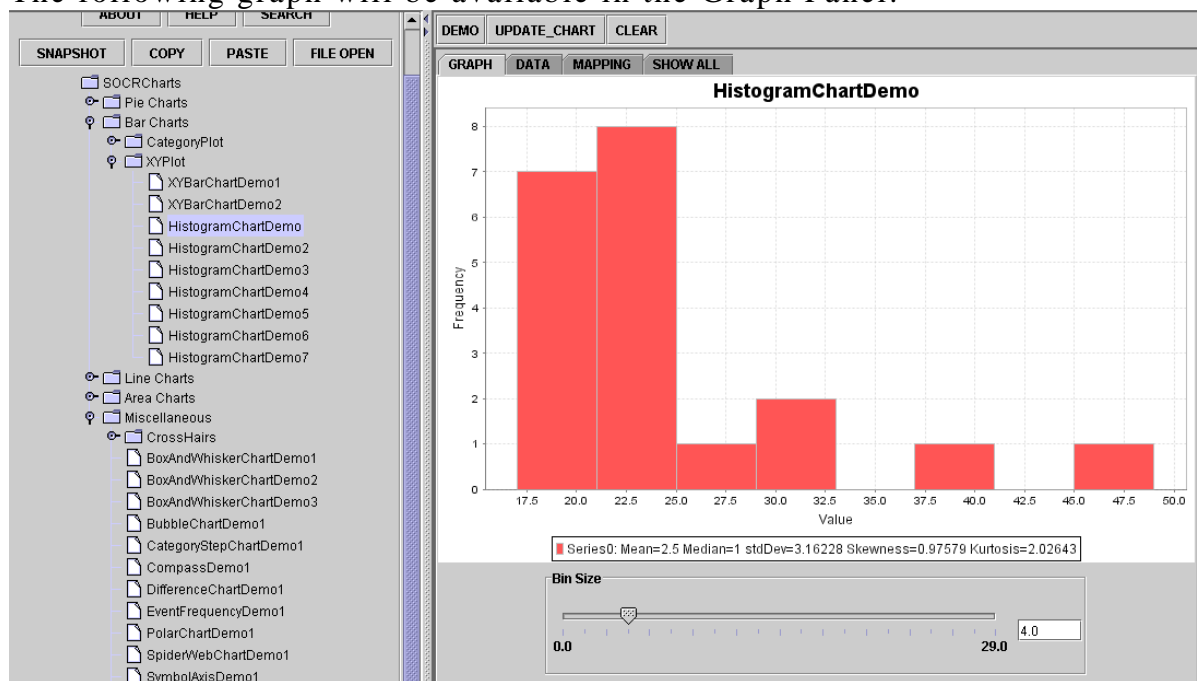


- Paste your data in the Data Tab as a column vector (to do that, you may need to save this row-data vector in a text file. Then open that with Excel make sure it shows as 20 COLUMNS, not one cell with 20 values in it. Then, copy these 20-cells and Special-Paste using “Transpose” as a vector of 20 vertical cells in 1 column in Excel. Finally, paste this one-column vector in SOCR Charts).

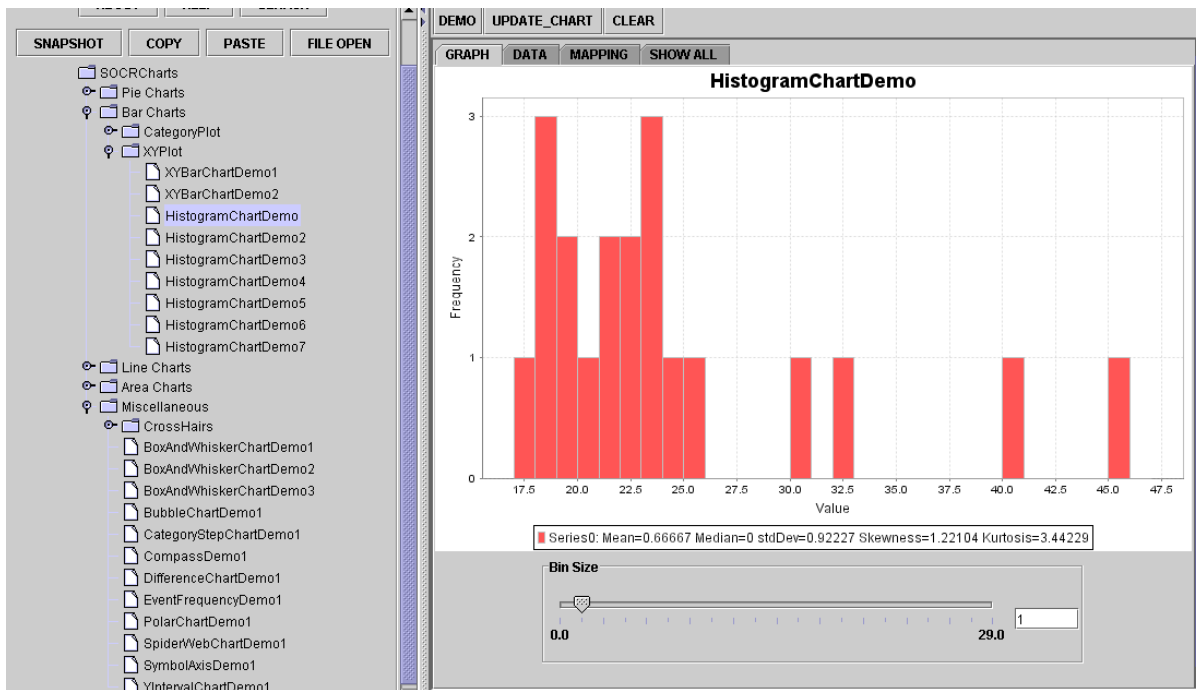


- Go to the Mapping Panel: select column “C1”
- Click **Update Graph**

The following graph will be available in the Graph Panel:



- Notice that you can change the width of the histogram bin-size. This will allow you to vary the shape of the data distribution!



## Topic 2

### **OTHER GRAPHICAL TECHNIQUES**

When information has been already processed, you may need to present it with different graphical techniques. You need to consider if data are CROSS SECTIONAL or TIME SERIES DATA.

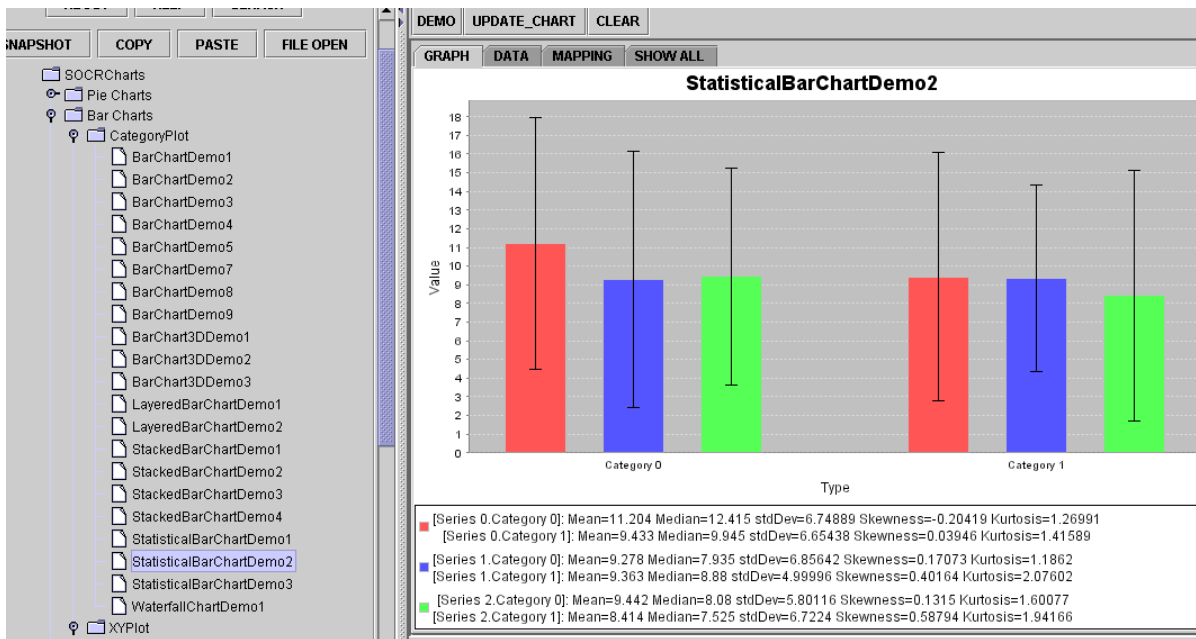
What is cross sectional data? It means that it is gathered from many observations all taken at the same time. For example number of customers who visit a store classified by age, or gender, or economic status.

What is time series data? It means that you gather and measure the information over time (e.g. monthly, quarterly or annually) For example, number of customer who visited the store every month for the last 4 years.

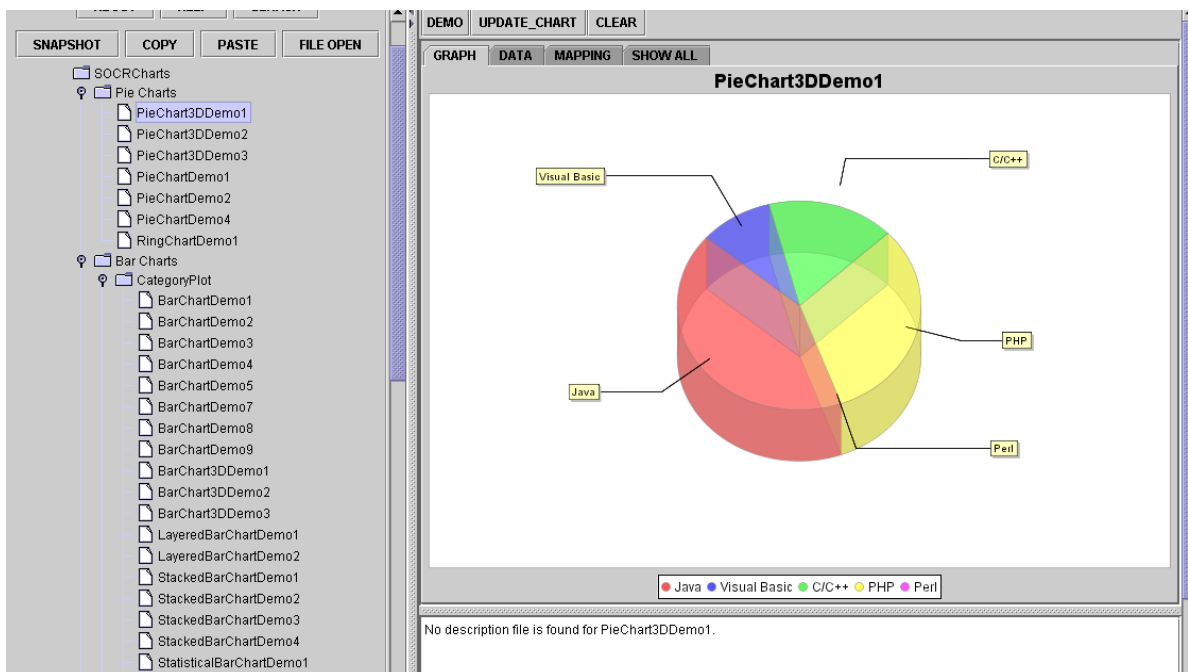
Of course you may have a combination of cross-sectional and time series data. For example: number of customers who visited the store every month for the last 4 years classified by gender.

### **When you want to describe CROSS-SECTIONAL DATA, you may use the following techniques:**

Bar Charts ([http://socr.ucla.edu/htmls/SOCR\\_Charts.html](http://socr.ucla.edu/htmls/SOCR_Charts.html)), StatisticalBarChartDemo  
 ←CategoryPlot ← BarCharts



## Pie Charts

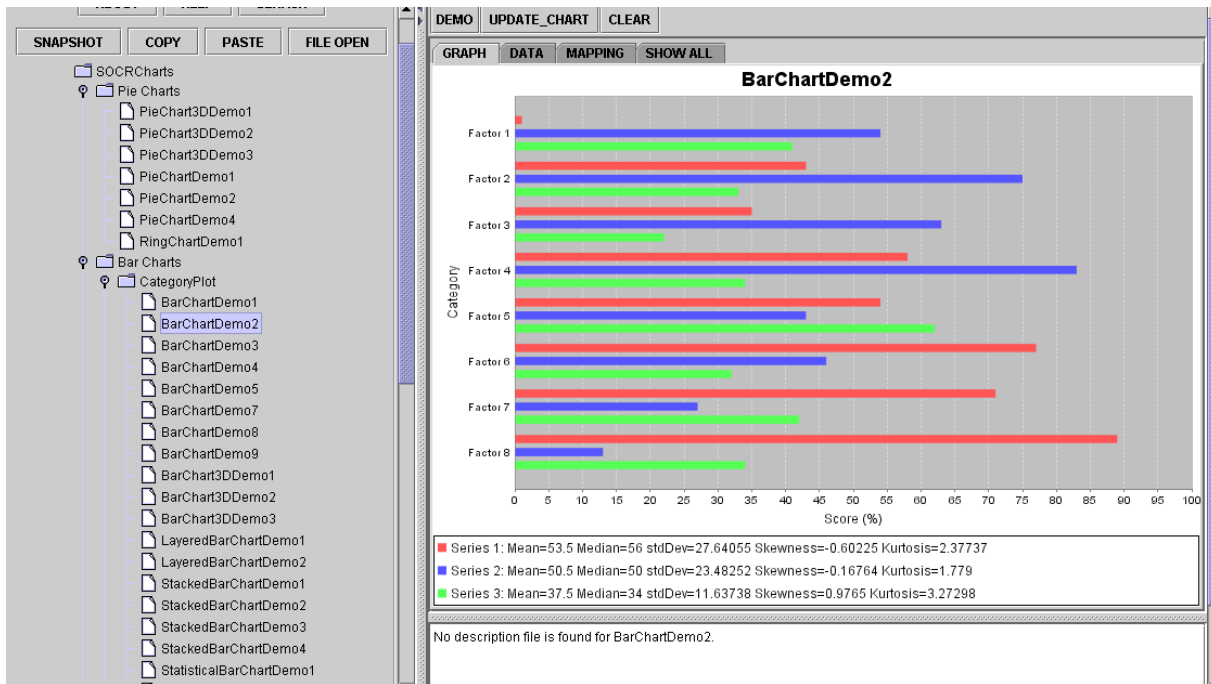


## BAR CHARTS

Graphical representation of categorical data set

- Bars may be vertical or horizontal
- All same color or different colors
- Multiple variable graphed on the same bar charts

Bar Charts are useful to graphically compare the highest bar or the longest bar and tell the importance of that specific category in the analysis. It is not recommended to represent trend, it is mostly used for comparisons.



### Differences with HISTOGRAMS:

1. Variable on X axis is categorical (country, stores, products) With histograms, the x variable was always a class (0-10, 10-20, etc)
2. There are gaps between bars. With histograms no gaps are allowed.

### PIE CHART

Graph is represented in shape of a circle:

- Divided into slices for each category or class to be displayed.
- Size of the slice is proportional to the magnitude of the category
- Slices add up to 100%

Pie charts are useful to show participation of groups within an event, but not recommended to represent trends. For example, if you want to describe the composition of your customers by age, from a 100% of population of sample.

The following example chart is based on the results of the [election for the European Parliament in 2004](#). The following table lists the number of seats allocated to each party group, along with the percentage of the total that they each make up. The values in the last column, the central angle of each sector, is found by multiplying the percentage by  $360^\circ$ .

**Group Seats Percent (%) Central angle (°)**

EUL	39	5.3	19.2
PES	200	27.3	98.4
EFA	42	5.7	20.7
EDD	15	2.0	7.4
ELDR	67	9.2	33.0
EPP	276	37.7	135.7
UEN	27	3.7	13.3
Other	66	9.0	32.5

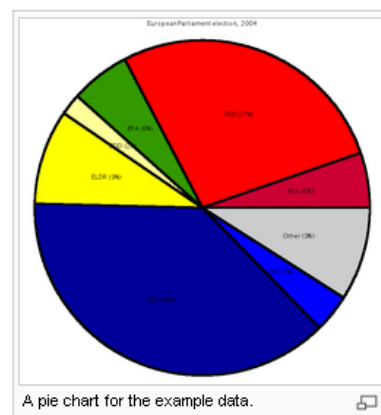


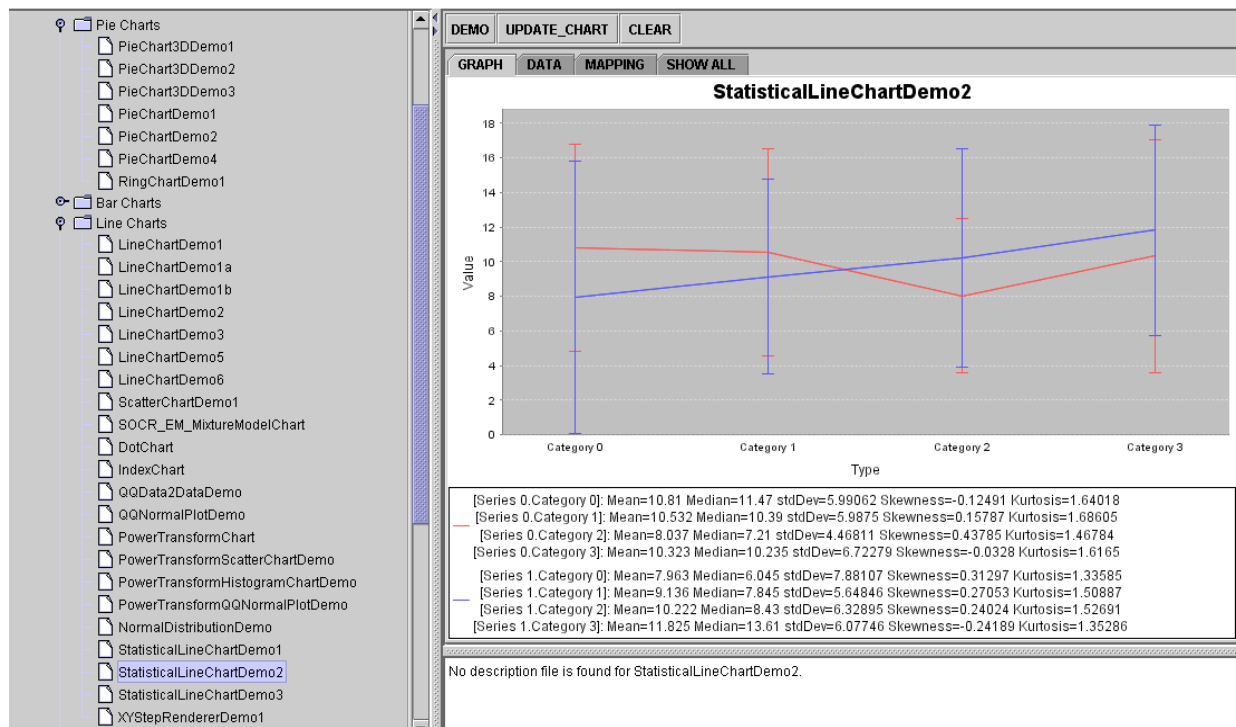
Figure 2 Example of a pie chart using information from the 2004 European Parliament election. December 29, 2006. [http://en.wikipedia.org/wiki/pie\\_chart](http://en.wikipedia.org/wiki/pie_chart)

**When you want to present data that are measured over time you may use:**

- Line Chart
- Scatter Diagrams

**LINE CHARTS**

Graph showing time on horizontal axis (x) and the variable of interest on the vertical axis (y). Data is required to be ordered chronologically, from the earliest to the most recent observation. Use SOCR options to build your Line Chart with very easy steps. It should look like this:

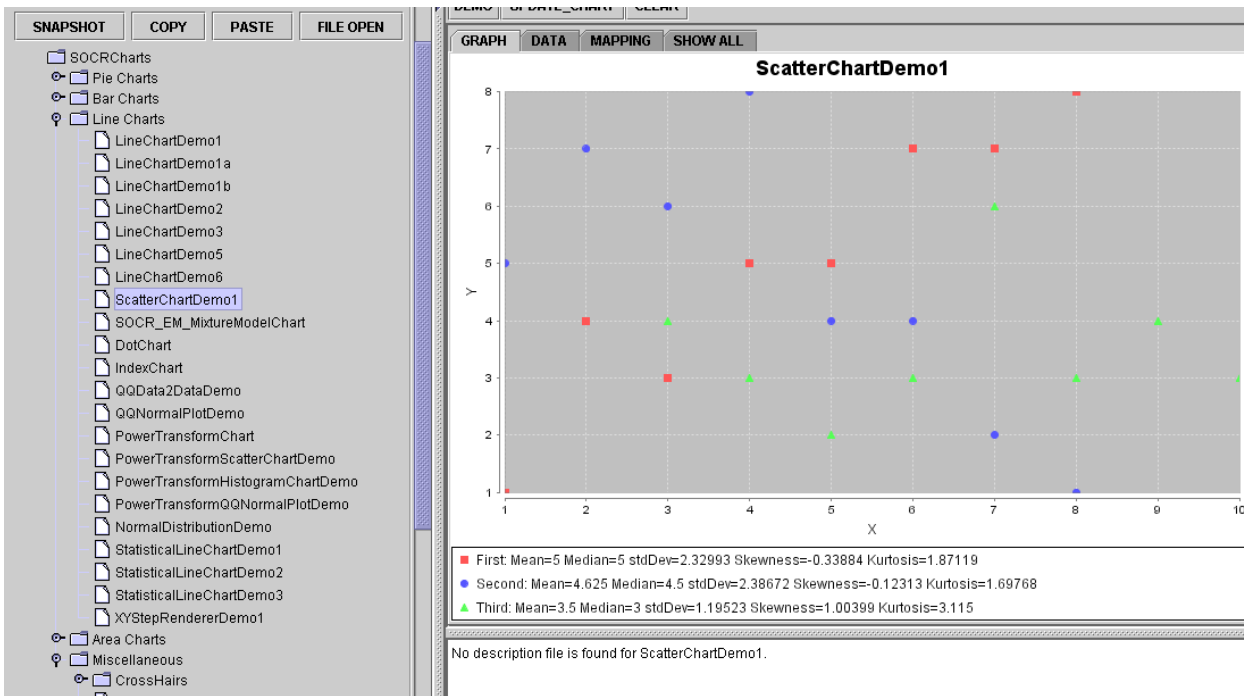


This is an application of line chart to build a Run Chart that compares information regarding a manufacturing factor over time and compares it with the standard.

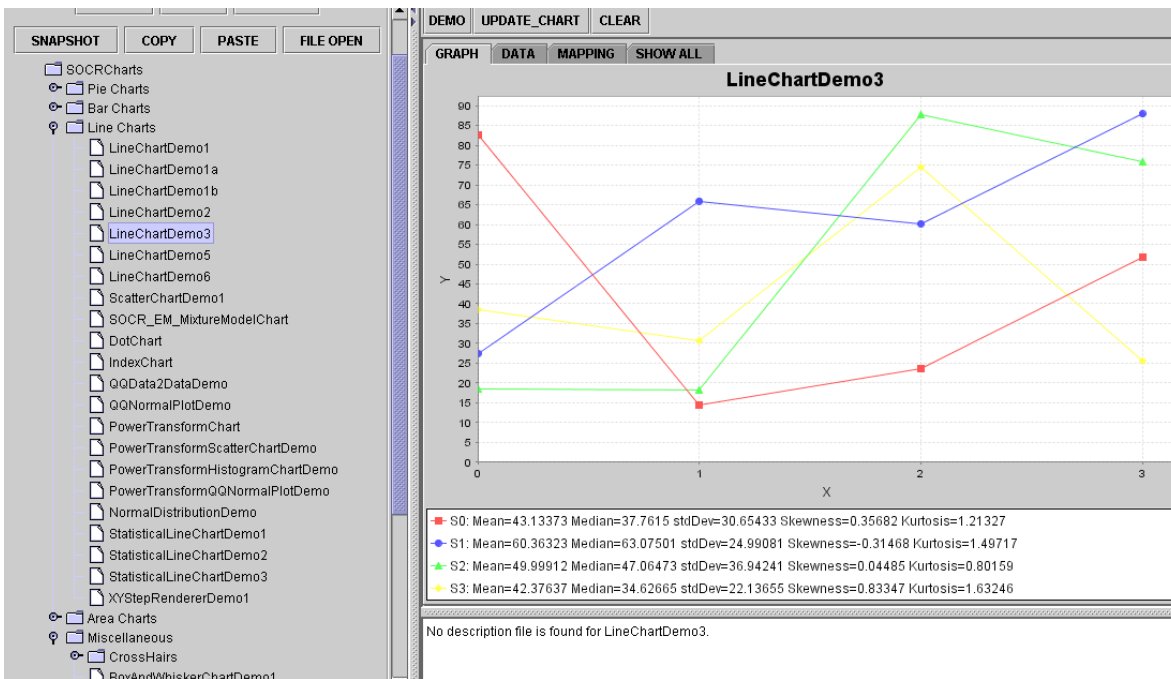
**SCATTER DIAGRAM**

It is used to describe 2 variables at the same time. Let’s say that we want to know what the level of poverty of a country is and how much education they are providing to their citizens. We may build a graph that assigns a bullet for the country and then positions this bullet to express what is the relationship between poverty and education? Or visualize the relationship between speed and time:

Scatter diagrams may have different shapes:

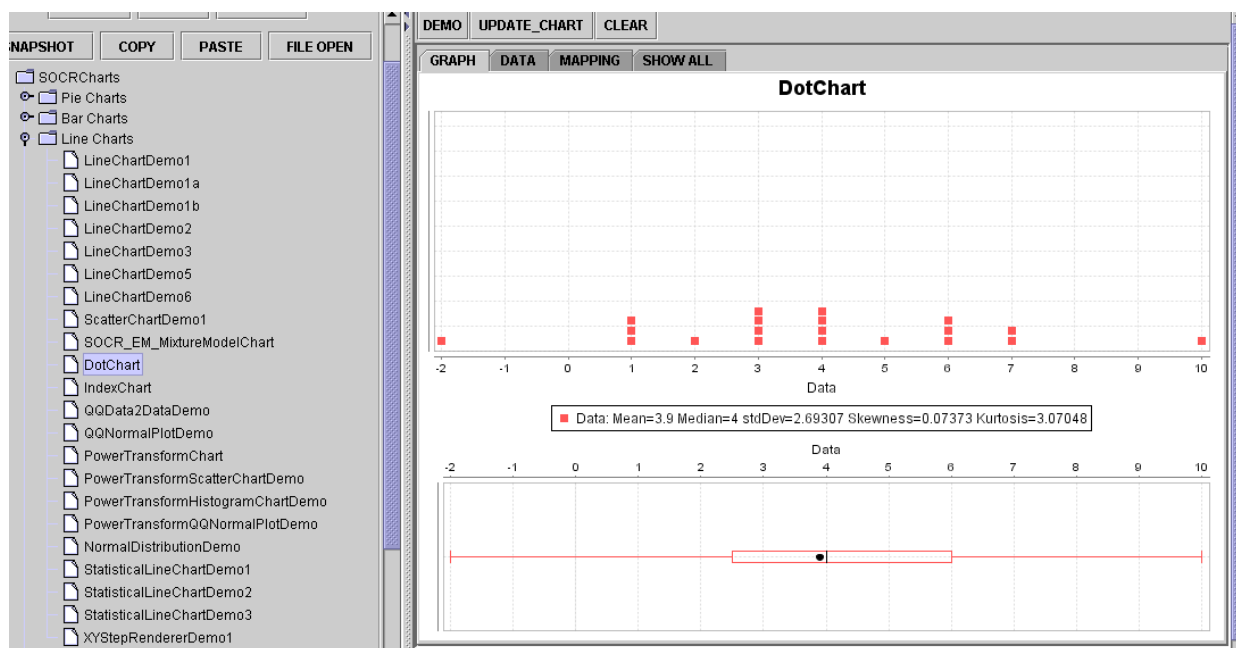


- Linear- upward
- Linear – downward
- Curvilinear upward
- Curvilinear downward
- No relationship



Scatter diagrams are very useful when we need to understand relationships between variables, and we will use them in future chapters when trying to find correlation between variables.

## Dot-Plot



**Additional SOCR Data and Descriptive statistics materials and resources:**

[wiki.stat.ucla.edu/socr/index.php/SOCR\\_Data](http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data)

[wiki.stat.ucla.edu/socr/index.php/SOCR\\_EduMaterials\\_Activities\\_CoinDieExperiment](http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_CoinDieExperiment)

Read your textbook and practice with your excel at home.

Have a nice week!