STATISTICAL TECHNIQUES

Data Analysis and Modelling

Data collection and presentation

Many of us probably some of the methods involved in collecting raw data. Once the data has been compiled, you will need to analyze it before presenting it.

The data collection method often involves a tradeoff between what is mathematically desirable and what is experimentally possible. It is important that when you collect your data and the way you present it should truly reflect the process you are observing.

Again, when collecting data, the primary concern is to design the data collection method in such a way to ensure that it gives the maximum possible information to the reader.

Presenting your data

Assume you have collected some weekly sales figures somewhat as follows.

	A	B
1	**Sales**	**Amount ($)**
2	Week 1	35
3	Week 2	60
4	Week 3	60
5	Week 4	35
6	Week 5	35
7	Week 6	25
8	Week 7	60
9	Week 8	170
10	Week 9	25
11	Week 10	105
12	Week 11	95
13	Week 12	80
14		

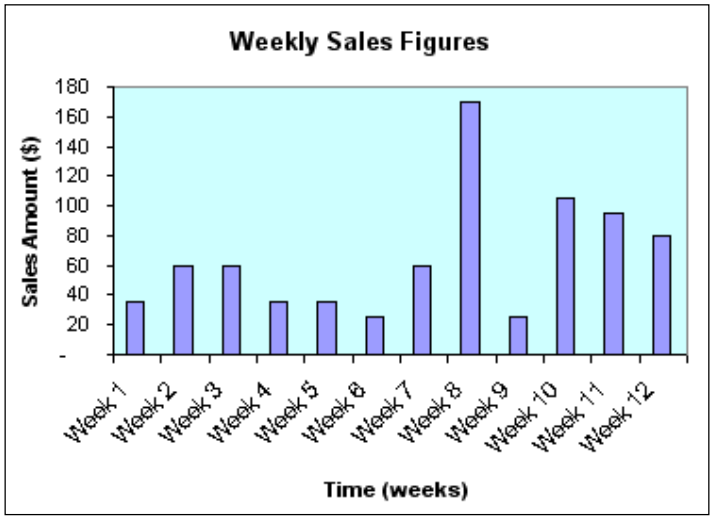
How would you represent the above weekly sales figures graphically? One can represent figures based on a discrete or continuous model. First, let us plot the above sales figures using two different graphs.

1. Select the range of data from **A2** to **B13**.

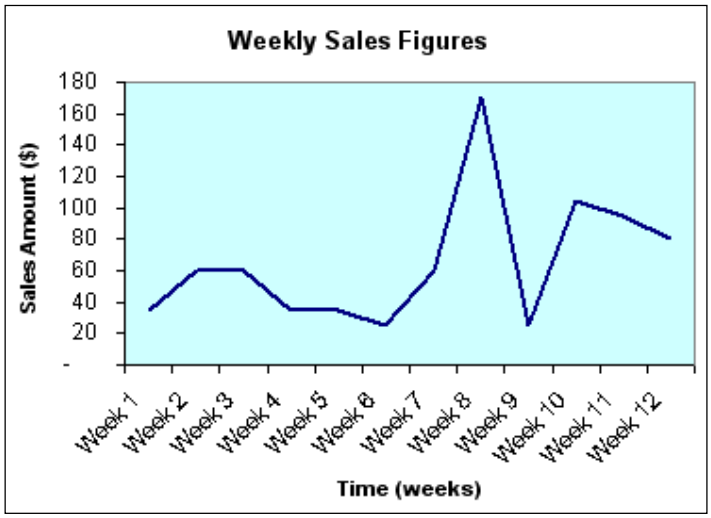
2. Click on the **Chart Wizard** () icon on the **Standard** toolbar.

3. Select the **Column Chart** type and create a chart somewhat as follows.

Authored by: Palani Murugappan



4. Similarly, select the range of data from **A2** to **B13** and create a **Line Chart** type as illustrated.



5. Note that the **Column** chart depicts a **discrete model** while the **Line** chart depicts a **continuous model**. What is the difference between the two? To answer this question, you need to know the nature of process being observed and then select the most appropriate model.

Let us analyze the weekly sales figures in detail. In the continuous model (**Line** chart), the chart is summarized by points joined by lines. However, if you intend to consider the pattern of sales variation during the 12 weeks i.e. 3 months, then this chart is quite misleading.

Firstly, we do not know if the weekly sales figures are based on average or total sales. For example, there is no mention if within a week, the sales figures fluctuate and the average is recorded.

Secondly, plotting the weekly sales figures as points joined by straight lines is incorrect. Strictly speaking, the chart implies that weekly sales are a continuous varying quantity. If you observe the continuous (**Line**) chart carefully, you will notice that between weeks 8 and 9 increases and decreases tremendously as compared to any other two continuous weeks.

Thirdly, the continuous (**Line**) chart of weekly sales does not say anything on how sales varies in the shorter term. As sales are daily based (assumption), a reader may fall into false conclusion that sales increases or decreases steadily.

Thus, a better approach would be to use a discrete model to represent the situation more clearly and accurately.

Different types of models

There are different types of models that you can use based on the data collected. The following is a summary of eight types of models.

1. A **discrete** model – one in which the independent variables must change in finite increment (or decrement) from one value to another.

2. A **continuous** model – one in which the variables accept real numbers as values and may change by arbitrarily small amounts. Continuous models deal with systems whose behavior changes continuously in time.

3. A **deterministic** model – one which yields a specific and repeatable outcome for any given set of conditions. Behavior is generally quite predictable.

4. A **stochastic** model – one which includes random elements to facilitate the same set of conditions that may yield different results. Behavior cannot be predicted entirely.

5. An **analytic** model – one that is expressed algebraically (or mathematically) by a set of equations.

6. A **simulation** model – one in which the computation (usually using a computer) resembles the structure of the process itself.

7. A **conceptual** (**theoretical**) model – one which is based on theoretical principles.

8. An **empirical** model – one which is based on outcome of experimental results rather than theory.

For the purpose of examples in this book, more emphasis will be placed on the discrete and continuous models only.

Discrete models

A discrete model of time involves measuring time in discrete units where each unit is of equal duration. In comparison, the continuous model of time records time elapsed modeled as a real number.

For most types of real time systems, the discrete model of time is considered a useful model to work with, as most computers respond to discrete time based interrupts.

Discrete event models describe the process steps. However, it may not have enough events to represent feedback loops accurately. In a discrete event simulation, activities schedule future events.

Discrete models are often used to model a manufacturing line where for example, items or "entities" move from one station to another and processing is done at each station.

Discrete models can easily represent queues and can delay processing of an activity if resources are not sufficient or available.

Discrete vs. continuous models

How do you associate time in models? For example, is time incremented step-wise in a dynamic model or is assumed to change constantly in infinitesimally small increments?

To help you answer the above question, let us consider some examples.

- Imagine a toy car rolling down a slide. The weight of the car remains the same but the speed increases as it slides further away from its initial starting point. This is an example of a physical model with continuous time.

- Generally, systems of differential equations represent continuous time models.

- An example of a discrete model is a movie or television program. In this case, the motion (movements of what you see) is achieved by viewing separate images that are taken at specified intervals.

Evaluating models

Before attempting to evaluate any models, you need to remember the following:

- **Models are only tools of a practitioner, not his/her goals**

- **The value of a model depends on how it is applied**

Thus, it is best to devise and evaluate models in terms of their intended applications.

Some steps to be considered before evaluating models are as follows:

- Need to establish the exact use for which the model is being designed for. Note that in many cases, a model is developed first and then applied in the wrong context.

- One needs to decide which assumptions are appropriate to make for the system under study. This is part of the process of fully understanding the application of the model.

- Wherever possible, create or setup a prototype model (a prototype is an attempt to predict the relationship between different variables i.e. it is somewhat a model without the carrying out the required testing) based on the assumptions made. If necessary, try several alternative approaches.

- Always evaluate the prototype model and make appropriate changes. Keep repeating this procedure until satisfied. At this stage, the model is both valid and useful for your goal.

Once you evaluate a particular model, you need to assess the merits of the model by answering the following two questions:

1. Is it valid?

- How "natural" is the model?
  - Do the variables reflect measurable biological quantities?
  - Does the form of the model realistically reflect the real processes?
- What assumptions does the model make? Are these assumptions compatible with the real world?
- How comprehensive is the model?
- Where does it breakdown?
- What explanation is not made or catered for?
- Is the model consistent with existing knowledge?
- Is it self-consistent?
- How sensitive is it to changes in particular parameters or features?

2. Is it useful?

- Does the model address the questions that people want to answer?
- What does it explain, if anything?
- Does it generate correct predictions or testable hypotheses?
- Can it be applied to real world problems?
- Is it possible to obtain the data required?
- Can the mathematics of the model be solved? If not, how difficult are the calculations?

Evaluating models using linear relationship

What is a linear relationship? In simple terms, it is a "straight line" relationship. A "straight line" is used more so for the purpose of prediction and forecasting.

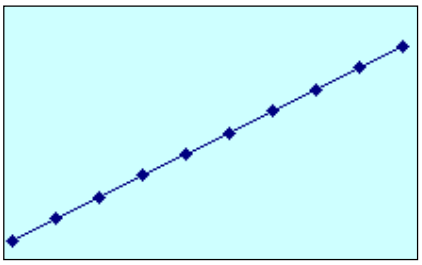
Let us consider the following example of linear relationship. Suppose you have just received your first pay check for the first month which shall be denoted by **x**. You then deposit your pay check, denoted by **p** (in dollars) every month to the same account. Assume that your existing account has a previous balance of **c** dollars before you started depositing your pay checks. Then for each month, you will have an amount based on the following equation: **f(x) = y = p*x + c**.

Assume the bank pays you interest based on your daily balance. Every month you will be receiving a little interest. However, you cannot keep all of the money in your account without spending some of it for your transportation, lodging, food, etc. Thus, the final monthly amount has a random component. After several years, you want to find out how big the random component is and to verify that the account has been increasing (or perhaps decreasing) linearly in general.

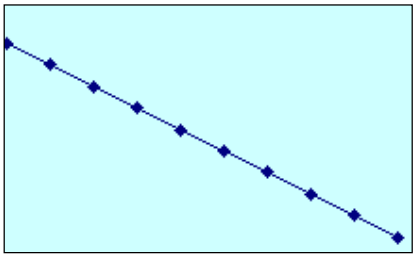
The slope of the curve (which is based on your savings) can be either one of the following:

Linearly upward i.e. if you save more than withdraw
or
Linearly downward i.e. if you withdraw more than save

The above can be depicted in the following diagrams.



Linearly upward curve



Linearly downward curve

Note that this straight line (upward or downward) relationship is what is known as a linear relationship.

Any other non-straight line relationship is called a non-linear relationship. Examples of non-linear relationship include polynomials, exponential, etc. (more of this will be covered later).

Identifying relationship

After selecting a model (based on your data), one needs to identify relationships between the variables. This involves fitting an equation to the set of data. There are three stages to the data. These are as follows:

1. Identifying if a numerical relationship exists
2. Identifying the type of numerical relationship
3. Estimating values for parameters of a known relationship

Before attempting to identify a relationship using a set of equations, you need to first collect observational data and "fit" a model to it.

When it comes to gathering data, it is desirable to find as many values as possible for the variables involved. These values should ideally cover a wide range of values to ensure that there appears to be a relationship even though on the face of it, one does not exist.

One of the easiest ways of identifying if a possible relationship exists is to plot the set of data gathered. This is usually done by taking one set of values and placing it on the x-axis and then plotting it against another value on the y-axis.

After you plot the set of data, observe if a relationship exists. Is the relationship immediately evident? If it does not, does this imply that there is no relationship between the variables in question?

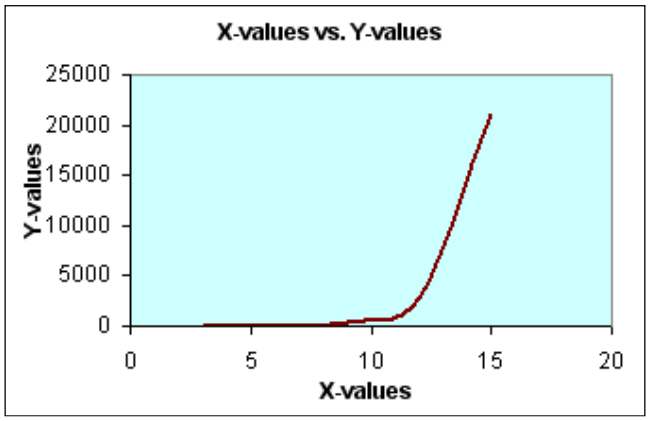
If you are performing this exercise for the first time, this could be a trap awaiting for you to fall!

Should a relationship not appear apparent based on the raw data plots, it can actually be transformed to appear as if a relationship exists. The following example will highlight this.

1. Assume you have the following set of data.

	A	B
1	**X-value**	**Y-value**
2	3	7.096
3	6	52.416
4	9	387.288
5	12	2861.696
6	15	21145.448
7		

2. Using the **Chart Wizard** () icon on the **Standard** toolbar, create a **XY (Scatter)** chart (select the **Scatter with data points connected by smoothed lines without marker**s sub-option). You should observe the following chart.

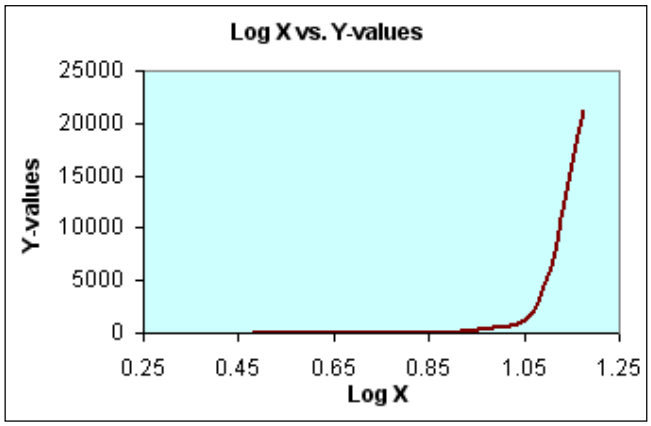


3. Note that the above plot does not indicate a straight line relationship.

4. How then would you perform the above to reveal a plot of linear relationship? To perform an attempt on identifying a linear relationship, try transforming the **X** and **Y** values to **Log X** and **Log Y** values (to the base 10). Your data should appear as follows.

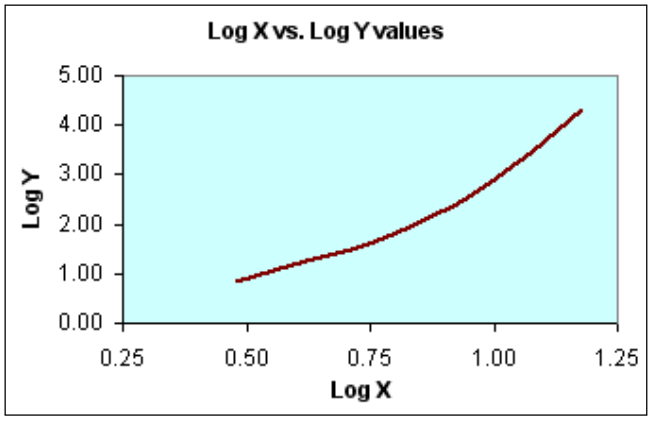
	A	B	C	D
1	**X-value**	**Log X**	**Y-value**	**Log Y**
2	3	0.477121	7.096	0.851014
3	6	0.778151	52.416	1.719464
4	9	0.954243	387.288	2.588034
5	12	1.079181	2861.696	3.456623
6	15	1.176091	21145.448	4.325217
7				

5. Next, perform a **XY (Scatter)** plot of **Log X vs. Y-values**. Your plot should appear as follows.



6. The above **Log X vs. Y-Values** plot still indicates a non-linear relationship.

7. Try plotting a **XY (Scatter)** plot of **Log X vs. Log Y** values. Observe that the plot now indicates a close linear relationship.



The above example had illustrated how you can perform the transformation of a raw set of data manually to plot a linear relationship. In the next part, you shall learn how Excel can perform much of this transformation and select the "best fit" curve for a given set of data.