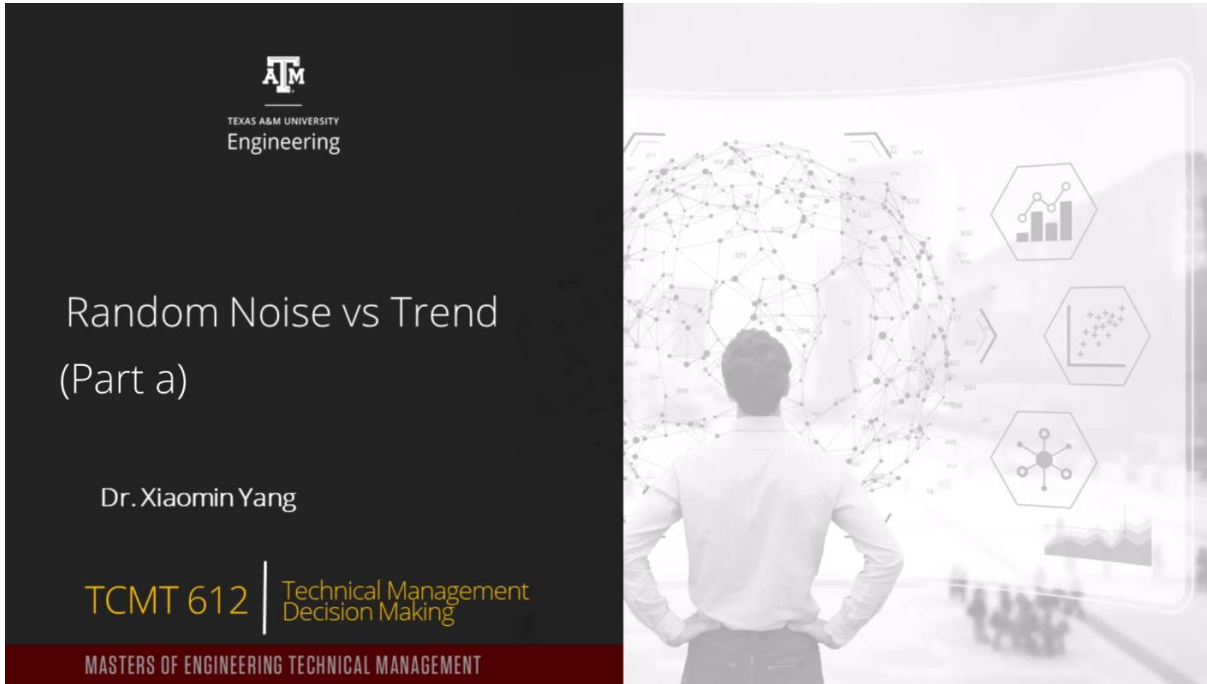


# M6L2a. Random Noise vs Trend

## Slide #1



ATM  
TEXAS A&M UNIVERSITY  
Engineering

Random Noise vs Trend  
(Part a)

Dr. Xiaomin Yang

TCMT 612 | Technical Management  
Decision Making

MASTERS OF ENGINEERING TECHNICAL MANAGEMENT

We will now discuss how to distinguish noise and trend.

How to tell the difference between random data and the business trend.

## Slide #2

### Engineering Approach to Forecasting



This table lists the monthly sale of a mature product between January 2016 and October 2017. The numbers are in thousands.

In the real world, the sales numbers always fluctuate.

Engineers may analyze the annual sales data and forecast it with a simple engineering analysis approach.

First, we plot the data in the Excel chart.

Second, we can use the Excel linear regression function to draw a line and calculate the slope of the line.

The slope represents the monthly sales trend, which is downward, so we can draw a conclusion that the sales slightly declined in the past 18 months.

Then we will communicate the sales trend to the management of your department.

The flawed analysis, however, may lead to serious consequences.

Your managers, the executives, do not have the chance to assess the details of the sales numbers, so they will likely follow your analysis. and put the product in the poor performance category.

Because the sales declined, and they may decide to make some change to improve the performance of the product.

However, if you take a closer look at the details of the sales data, you will realize that the sales numbers fluctuate in a broad range.

The monthly sale may not be a function of time, which means the monthly sale does not exhibit a trend like what you derive from the engineering analysis approach.

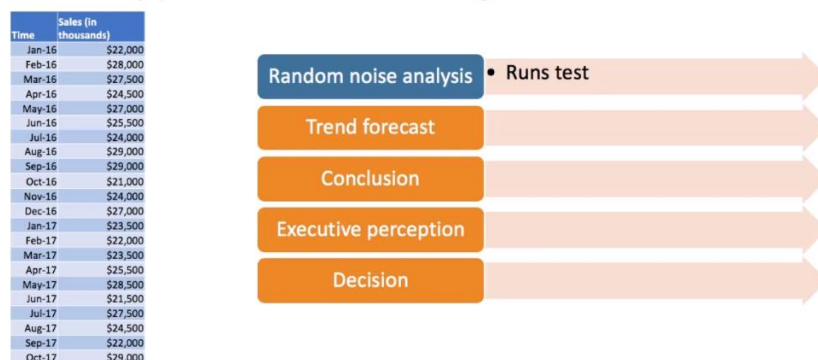
From this example, I want to show that doing business forecasting is beyond the engineering trend analysis.

You really need to consider several business factors before choosing a forecasting model.

So, forecasting is much more complicated than just doing the engineering analysis.

### **Slide #3**

#### Business Approach to Forecasting



The business approach to forecasting should include the random noise analysis step before a trend forecasting.

We need to first determine to what degree the monthly sale is correlated to time.

The method for the randomness analysis is called runs test.

## Slide #4

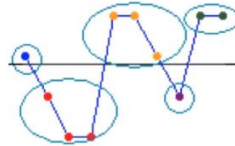
### Random Noise Analysis: Runs Test

---

The "runs test" can be used to decide if a data set is from a random process.

To what degree is the monthly sale correlated with time?  
or

To what degree is the monthly sale a function of time ?



The runs test is a statistical analysis that can be used to decide if a dataset is from a random process or presents a trend.

The runs test determines to what degree two variables, in this case monthly sales and the time, are correlated with each other, or to what degree is a monthly sale a function of time.

## Slide #5

### Runs Test Step 1: calculate a baseline value

---



Baseline value is the comparison criterion

By default, Baseline value is the mean of the sample data

Baseline value = monthly sales average = \$25,273,000

But you can also specify a different value, such as the median

To do a runs test, we first need to calculate the baseline value of the monthly sale.

The baseline level is the comparison criterion that we will use to compare the monthly sale data against the default.

The baseline value is the mean or the average of the sampled dataset.

For example, in this case, the baseline value is the average of monthly sale between January 2016 and October 2017, which is 25,273,000.

But you can also use other methods to calculate the baseline, such as the median of the dataset.

In this case, we use the average, which is default method to calculate the baseline.

We draw the baseline and also sale numbers in the chart.

The black line is the baseline, the average of monthly sale, and the orange dots are the monthly sale numbers.

But it is quite obvious the sales fluctuate around the baseline value.

## Slide #6

### Runs Test Step 2: count the number of runs

Runs: is the number of groups of observations that are above or below the baseline



Base level = monthly sale average = \$25,273,000



By definition, runs is the number of groups of observations that are above or below the baseline. Let us look at the data and I will explain to you how we come to the run numbers. We always start count with one.

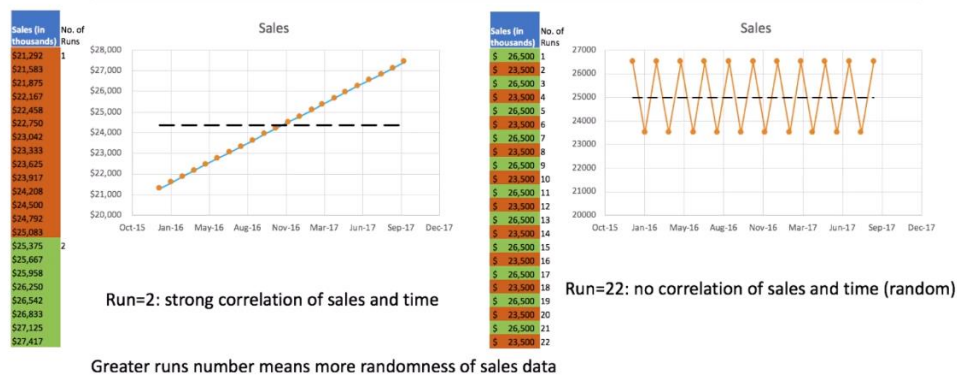
In January, the data point is below the baseline. And in February and March, the sales are above the baseline. So, we can tell from January to February, the sales numbers cross the baseline. So, we register one run number. The number of runs between January and March data is two.

In April, the monthly sales dropped below the baseline. So we register a new run. The runs number is three. The monthly sales increased in May and crossed the baseline between April and May. We register a new run. The total number of runs between January and June is four. When a sale in two adjacent months crossed the baseline, we register a new run.

Totally, the sales trend crossed the baseline 14 times between January 2016 and October 2017. The number of runs is 14. There are 14 circles which represent the number of groups that are above or below the baseline on a sales chart.

## Slide #7

### Number of Runs and Data Correlation



The number of runs represents the data correlation between two variables. For instance, the monthly sales is perfectly correlated with time, which is linear. The number of runs is 2. The first half of the sale is below the baseline, where the second half is above the baseline. So there are only two groups of data. One is below the baseline, the other is above the baseline. The number of runs is 2.

On the other extreme, the sales data is not correlated to time at all. It fluctuates above and below the baseline every month. So it perfectly fluctuates above and below the baseline. The number of runs is 22 because each data point is a run.

From the two extreme examples, you can tell that greater run numbers means less correlation between the sales and time, or more randomness of the sales data.

When the run number equals to 22, that means that sales data is totally random. It does not have any correlation with time.

But when the run number is 2, that means the sales data is perfectly correlated to time. They are in a perfect linear relationship.

## Slide #8

### Runs Test Step 3: statistics analysis (normal distribution)



The number of runs (  $R$  ) in a sequence of  $H$  data points is a random variable that follow normal distribution where:

No. of sale data points (H)	22
No. of above baseline (Ha)	11
No. of below baseline (Hb)	11
No. of runs (R)	14
Expected Run $u(R) = 1 + 2 H_a H_b / H$	12.00
Stev $d = \sqrt{((u-1) * (u-2) / (H-1))}$	2.29

For any runs number between the smallest and the greatest, we can do a statistical analysis to calculate the degree of correlation or a degree of randomness to determine to what degree the two variables are correlated or not correlated to each other.

The statistical analysis is called normal distribution analysis.

For the purpose of this course, I am not going to discuss the details of the statistical analysis.

But if you are interested in learning the details of the statistical analysis, you can read the textbook, or we can discuss that in office hours.

But I do expect you to understand for any set of data, the number of runs represents the degree of correlation between two variables.

The smaller run number represents perfect correlation between two variables and the greatest run number means the two variables do not have any correlation.

The data set is random and anywhere between you can use a normal distribution to calculate the degree of correlation.



According to the null hypothesis of statistics analysis, the number of runs,  $R$ , in a sequence of  $H$  data points is a random variable that follows normal distribution.

This table shows the formula for the normal distribution calculation.

$H$  represents the total number of data points.

$H_a$  represents the number of total data points above the baseline.

$H_b$  means the number of total data points below the baseline, and  $R$  represents the number of runs.

The expected value, or mean of runs of any kind of dataset, can be described with this formula:  $1 + 2 \times H_a + H_b$ , divided by the total number of data points,  $H$ .

And the standard deviation is the square root of the mean minus one times mean minus two divided by the total number of points minus one.

Again, I do not expect you to remember all the formulas, but I do expect you to understand that for any data set, the number of runs follow a normal distribution.

You can use the normal distribution analysis to determine to what degree the two variables are correlated to each other.