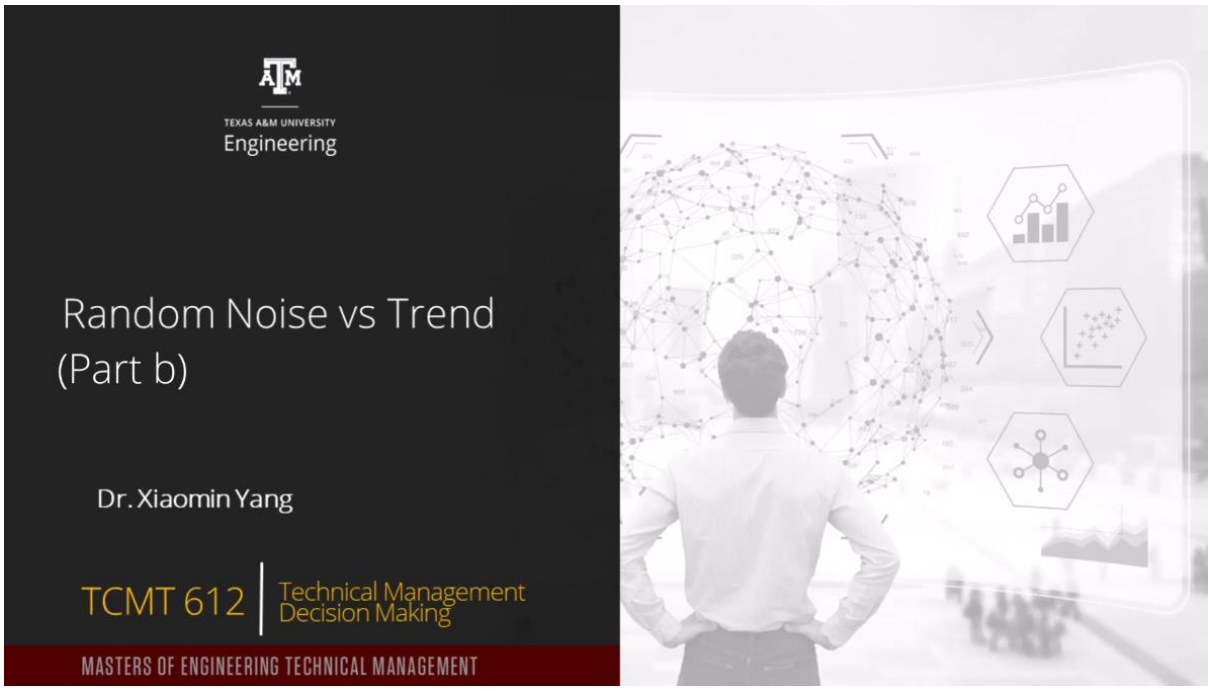
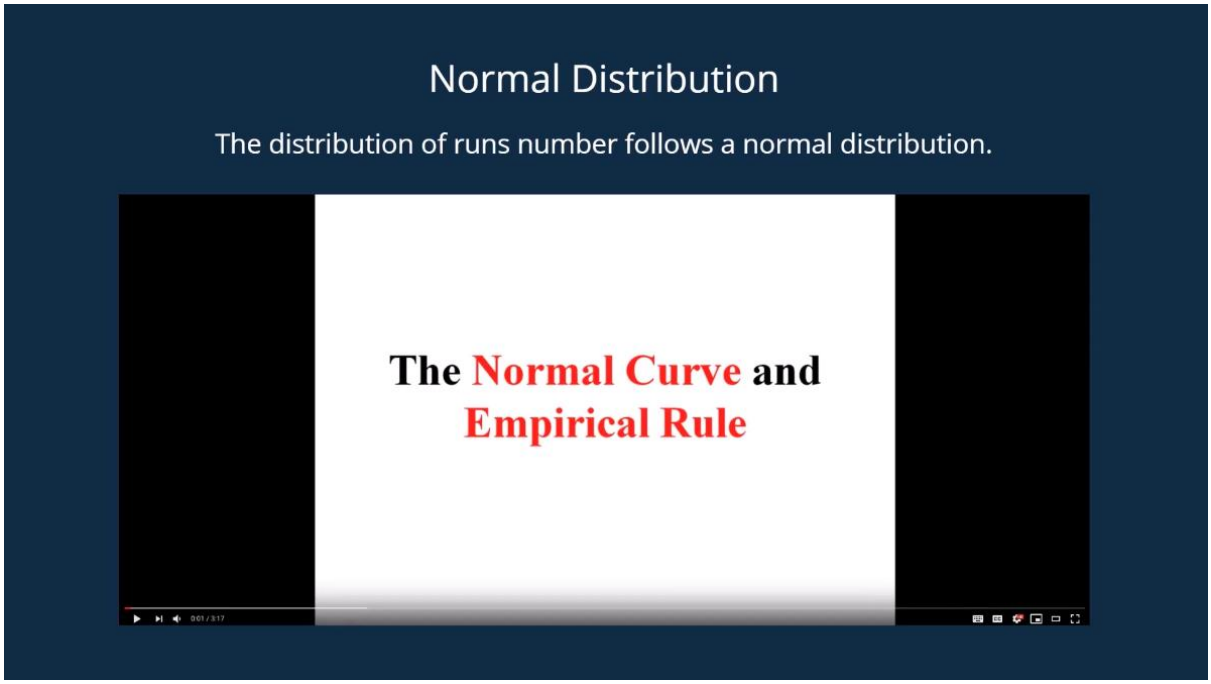
M6L2b. Random Noise vs Trend

Slide #1



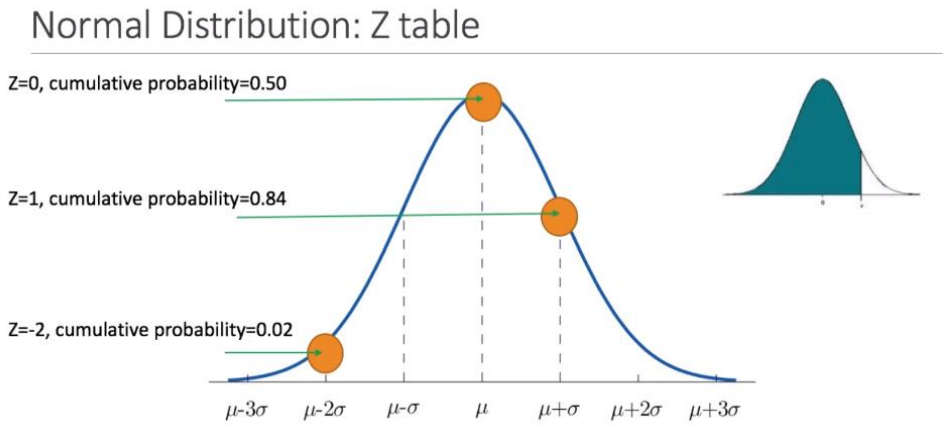
Slide #2



The distribution of runs number follows a normal distribution.

The following video clip may help refresh your memory about the normal distribution.

Slide #3



The normal distribution curve.

We can use the z table to calculate the cumulative probability given a deviation number z.

If z equals negative 2, the cumulative probability is 0.02 or 2%.

If z equals 0, the cumulative probability is 0.5 or 5%.

That is where the mean is located.

If z equals 1, which means that the data is one deviation from the mean, the cumulative probability is 0.84 or 84%.

The cumulative probability represents the light side area of the z score curve.

Since the principle of statistics analysis is the null hypothesis, the number represents the probability that the tested variable is a random variable, which means the degree of no correlation between the sale and the time in our example.

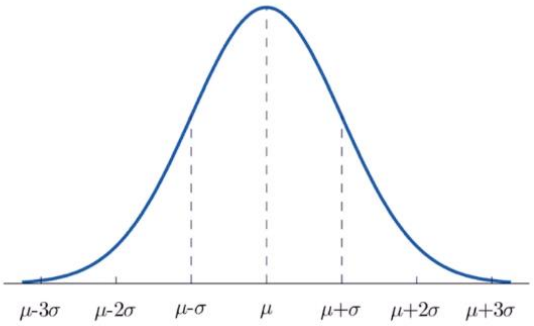
Slide #4

Normal Distribution of Runs

No. of sale data points (H)
No. of above baseline (Ha)
No. of below baseline (Hb)
No. of runs (R)

Expected Run u (R)= 1 + 2 Ha Hb / H
Stev d=sqrt ((u-1)*(u-2)/(H-1))
Z value = (R-u)/d)

MS Excel formula
Probability of random variable (no correlation):
p=NORM.S.DIST(Z, true)



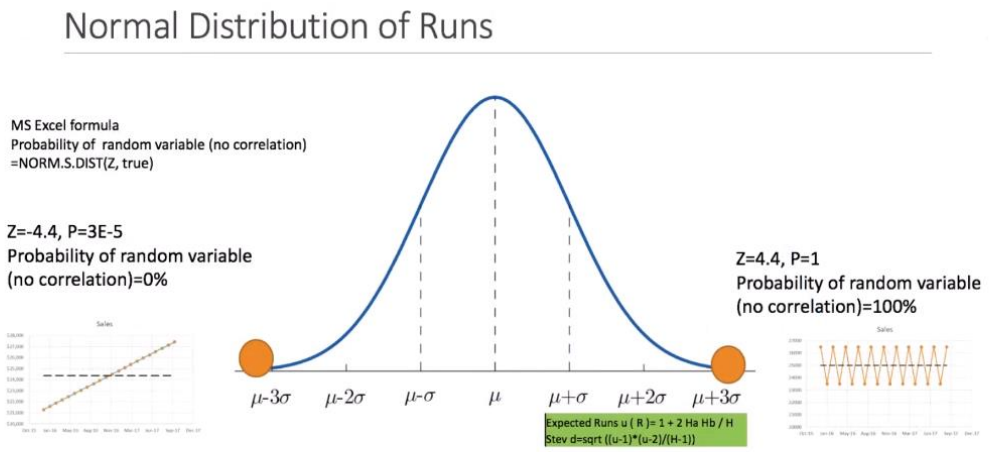
We can use the Excel normsdist function to calculate the cumulative probability of runs number.

In this formula, u is the mean of number of runs and d is the deviation.

Z is the z number, the number of deviations.

The function returns the probability that the test variable is a random variable, which means there is no correlation between the sale and the time.

Slide #5



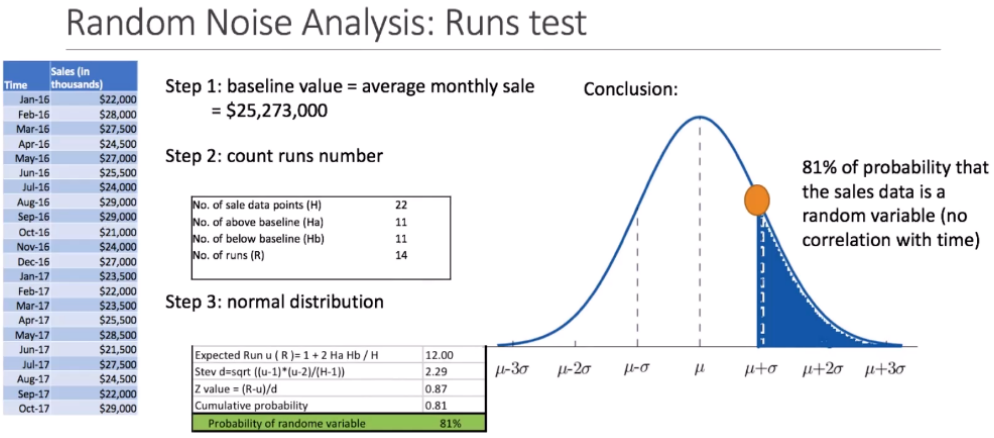
Let us explain the relationship between the cumulative probability and the probability of correlation using the two extreme cases.

The left side represents the perfect correlation between monthly sales and the time. The number of runs is two and from the normal distribution we can calculate z equals negative 4 and from the z table, or the Excel formula, normsdist, we can calculate the probability as 0, which basically means that the sale number is perfectly correlated to time, and the chance or the probability that sale number is a random variable is 0%.

At the right side, the sales data is independent of time. It is totally random. It is not a function of time. And the run number is 22. And from the calculation, we know that at z equals 4. And from the z table or the Excel function, it returns a probability of 100%. That means that the probability of the sales data being a random variable is 100%, which is very consistent with the nature of the data.

So the RUNS test result meets the characteristics of the data set very well.

Slide #6



Now, let us calculate the probability of the correlation between a monthly sale and time. Based on the runs test analysis, the left table shows the monthly sales between January 2016 and October 2017.

The first step is to calculate the baseline, the average monthly sale.

In this case, it is 25,273,000.

The second step is to count the number of runs.

The total number of sales data points is 22.

The number of sales above the baseline is 11 and the number of sales below the baseline is also 11.

The number of runs is 14.

Step 3 is to do the normal distribution analysis.

We can calculate the expected number of runs based on this formula.

The expected run or a mean of the normal distribution is 1 plus 2 times the number of points above the baseline by the number of points below the baseline divided by the total number of data points.

We can also calculate the standard deviation with this formula.

So the standard deviation is 2.29 and then we can calculate the z value which is basically the number of deviations which is the result it is 0.87 and then we can use the excel normsdist to calculate the probability.

The formula returns the number 0.81.

That means that there is 81% of possibility that the sales data is a random variable.

It does not have correlation with time.

The other way to say that is we are 20% confident that the sales data is correlated with time.

From the data, we can see it is highly possible that the sales data is just random data.

It does not have any correlation with time.

Since it is highly possible that the monthly sales are not a function of time, our previous linear regression type of analysis is flawed.

For this specific case, linear regression is the wrong forecasting model.

So, it is very important to choose the right forecasting model based on the nature of your data.

If you choose the wrong model, make the wrong assumptions, you can still do the forecasting analysis and get some results.

However, if you use your information to make business decisions, it is highly possible that you are going to make wrong decisions because your interpretation of trend is flawed.

So, before you do a forecasting analysis, if your data does not present an obvious trend, I recommend you do a simple runs test to confirm that your data is not just random data so that you can choose the right forecasting model for your analysis.