

**Six Sigma**  
**Prof. Dr. T. P. Bagchi**  
**Department of Management**  
**Indian Institute of Technology, Kharagpur**

**Lecture No. #05**  
**Review of Probability and Statistics – I**

Good afternoon, it is Tapan Bagchi again. I have this series of lectures prepared for you on six sigma. Now, one of the procedures or processes that is used in six sigma is the improvement process. This step comes along in that d m a i c - dmaic procedure that we had mentioned in the previous lectures. In this improvement step, what is really required is to be able to find the causes of losses or the causes of high variance, and to try do something about them, so that you end up with smaller variance or lower losses. In order to do this, we have to discover some knowledge about the process, and most of this work is imperical.

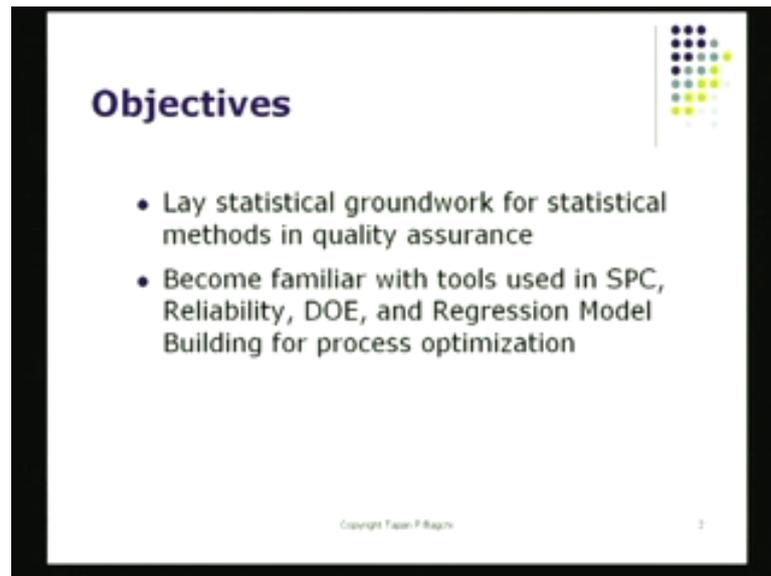
Very rarely you will be able to find a theoretical model, which if you optimize will give you the optimum result, the best result. So, most of these times, we are guided by data collection, we are guided by experiments, we are guided by data analysis and so on and so forth. This is something that we cannot escape. It is actually a key step in the process of dmaic or the six sigma approached improving quality. Now, today what we have for you, are the basic tools that are picked that have been picked up from statistics and probability. And in order to do that, what I have to do is, I have to give you some basic ideas about the concepts in probability and also the concepts in the statistics. This is obviously not a complete lecture in probability on statistics. But I will touch upon all the important concepts there, to make sure you get some practice, and also you get to see how we utilize these principles to try to tackle real data, data analysis problems.

Now, there are many books available, and a book that is fairly recent, and that are found to be quiet good is the one that is mentioned right here; you can see the display, it is called complete business statistics. And it is written by Aczel and Sounderpandian, and it is published by Tata Macgrylls, and the one that I am using is the sixth edition of the book. You could use all those any other book, but this one I found to be quiet readable, and it is got many examples which actually comes from real life. So, if you feel like you could buy a book, buy this one or you could buy some other book, it would be really

handy. If you have one of these books around to refer to it from time to time; not only for this lecture, but also when you are actually tackling six sigma project.

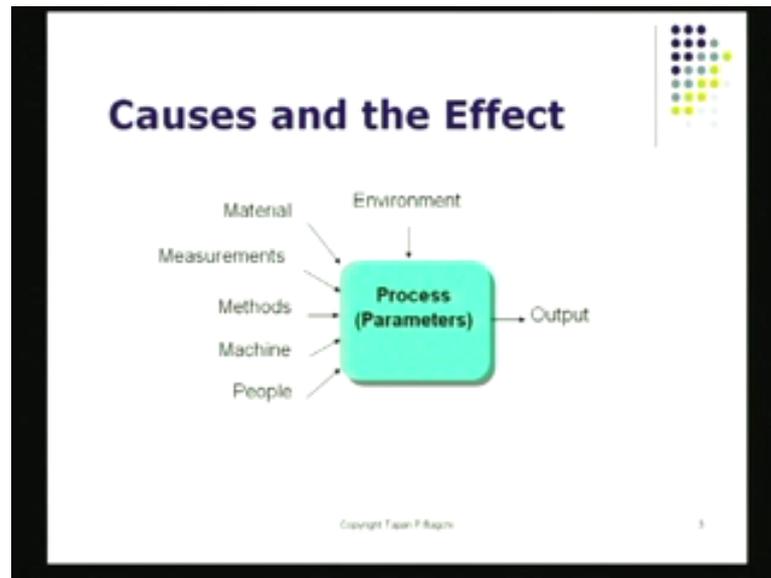
Let us begin by going to the slide number one. And the very first thing, we have is the title slide probability and statistics in six sigma a review.

(Refer Slide Time: 02:58)



The objectives of this lecture is going to be lay the statistical ground work for statistical methods and quality assurance, that I will try to do in a very broad way. And then of course, we would like to be familiar with the tools used in six sigma that include SPC reliability desire of experiments, regression model building and process optimization. These I will be touching upon and just to remind you, if you aspire to get a black belt in six sigma. You are required to be familiar with these techniques. These are techniques that actually are part and partial of the black belt training procedure. Say it is not something that I have been put here only because, he is theoretically complete, but actually these become the tools, these become the devices. That you utilize when you approach a project that is to be tackled by six sigma.

(Refer Slide Time: 03:49)

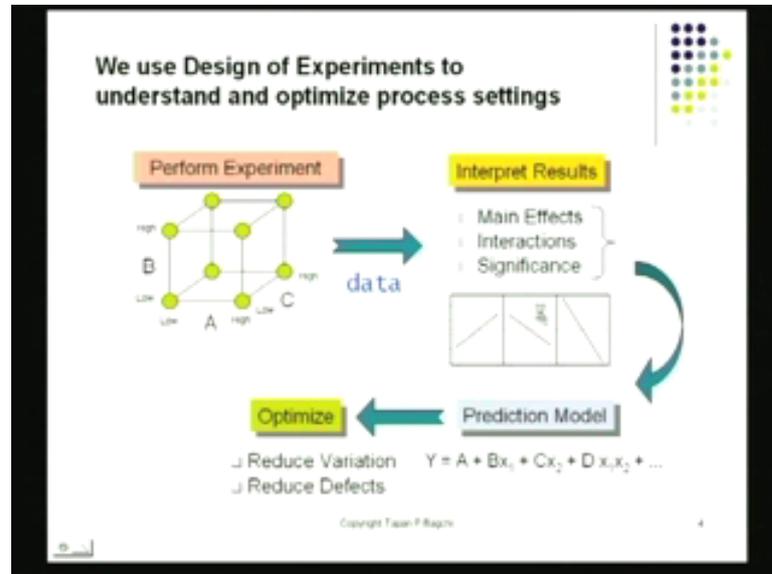


Now, why study probability and statistics, the problem is that most things around us, they have some uncertainty in them. For example, here I have got a system I have got some kind of a process, which is shown by the green box. And notice here this green box is affected by the environmental factors, it is affected by materials coming in, measurement devices and instruments and so on. The methods and technologies that is used to do it. So, machinery and also people obviously, these are all the things that are the input and they each have an impact, each of these would have an impact on the process.

Therefore, the output itself will not be really a stationary or a steady output, it would have some variation and it is this variation. When it gets beyond what is acceptable to a customer, who have a quality problem, the customer is not satisfied. In this case what we have to do is, we have to see if you can understand the cause and effect relationship, which is like start with these parameters start with these factors. And see which of these factors really is an important one, really an influential one in a impacting the output. If that happens of course, you have got at least one handle on the process, it is very possible of course, that more than one factor would be simultaneously affecting the process. In that case you have to study that affect also.

How do we study these things? We study that by looking at the output and making some deductions based on the principles of statistics and probability. And that is really the objective of today's lecture.

(Refer Slide Time: 05:18)



I have an example here and this example basically. I am going to take a couple of minutes to explain, what this example is? This optimization of a process using a experiments. Now, if you look at for example, again I will give the example of my digital watch. The theory has progressed to the point, as far as digital devices are concerned. When we can write down the exact equation to say, what this timing device will show at a particular moment in time. We can relate that to the crystal vibrations the bolts cell all the devices that are inside the, there are little tiny transistors that are inside; we can relate the output to the characteristics of those devices. And we can write down the exact equation for it.

Now, that is a pretty high level of knowledge, generally this is not true when you go and operate a plant. For example, a chemical plant or a steel plant or a mechanical plant or producing widgets and so on and so forth or a metallurgical plant many times, we are really not that sure about the running of that process, when I compared that to this state of knowledge that we have with my digital watch. Therefore, what we have to do many times, we have to do experiments usually have to run that process under different conditions. And these conditions are changed by changing the input parameters, then of

course, you look at the output and you try to see, what is the effect that turns out to be most prominent and which factor is causing it.

If you can get that cause and effect relationship, then you know that this is causing that. And therefore, by controlling the input I will be able to control the output to the desired value, this is really our goal. How do we do that? Let us take a look at this screen again I have a scheme here and I am going to explain this to you later or I am just going to show you a matrix scheme. And this matrix scheme I have got three factors involved, this is some process that has got three factors affecting it. And this is of course, at this stage speculative, we do not know exactly how does factor A impact the output or how does factor B impact the output, how does factor C impact the output that we do not know what we do know is A is a factor that can be set at two levels low and high.

B is a factor that again can be set at two levels low and high and c is a factor that can be set at two levels low and high. So, in fact, now, I have a means to set this setting of A either at low or at high, I can do the same thing for factor B, I can do the same thing for factor C. In other words now, two times two times two, I can produce eight combinations. Eight different ways I can run this process and observe the output, the result is this, A will have it is own impact, B will also have it is own impact and simultaneously C will also have it is own impact. This is going to show up with the experimental data.

So, what we have here, we have experiment number one, which is like when A is set at low level, B is set at low level and C is also set at low level. So, this is the point where I have got factor A set at low level, factor B set at low level and factor C at low level. And I observe the output I do the same thing by changing A from low setting to high setting. And again I run a trial which will be done at A at high level, B at low level and C at low level. So, I have got high, low, low, that is the point, this point and again I run the process and observe the data. In this way I will I am able to produce eight different pieces of data. Eight different conditions under which I have run the process using three factors and at two levels each.

So, I produce this data, this data is highly valuable for us. This is the data that now contains process information. What we then do is we subject this data to some statistical analysis. Statistics actually is the science for analysing data; statistics is the only science

with the help of which you can analyse data. You cannot do that with geometry, you cannot do that with geography, you cannot do that with physics, you cannot do that with english or any other subject. Statistics is actually the subject with the help of which you can analyse data that you produce, as the results how of some experiments.

So, I have got this data there, I do the analysis and the object of my analysis is to find. For example, the very first thing factor  $f_x$ , which really says does factor A have a significant effect as far as the output is concerned, does factor B have a significant effect on the output or does factor C have a significant effect on output. And is there any interaction between A and B and C, if that is also there we also have to find that. So, what we try to do is, we try to do the data analysis in such a way, that it expresses this structure of the matrix experiment. And it produces these information these pieces of information for me.

What are these little graphs here, on the Y axis I have got response plotted, on the X axis I have got the first block is for factor A, factor A at low level and then at high level. And notice here when I move, when factor A is moved from it is low setting to high setting response goes up. Which is like giving me some a clue, as to for controlling the response I can move a from low to high setting, when I look at B I find that, when I move B from high, low setting to high setting, response comes down, when I look at C again I find again when I move C from low to high response comes down.

Now, here if my goal was to try to maximize the output I will be picking the high setting for A plus, the high setting for, low setting for B plus the low setting for C. So, high at A, low at B and low at C this combination is going to give me the maximum output the maximum. The highest value for the response Y that is like something that I did pretty simple just by doing trying some experiments, doing some data analysis and plotting these little plots here. Similarly, I can also find out what interaction between these two factor effects between these three factor effects. And also I can find out by doing this, I can find the significance of these effects. By significance I mean, if I run a trial, if I run an experiment.

For example, I will be doing an experiment right now, you know this room is relatively quiet and so, there is a low level of noise. And therefore, you can hear my voice, because my voice is significant when you compare that to the background what is there in this

room? But if suppose, I start lowering my voice, which I am going to do just now; and you can hardly notice if I am saying anything, because I brought down my signal to the point it is at the same level as the background noise here. What I have really done is I have reduced the significance of my effect to the level when it is not distinguishable from the environment or the environmental noise.

Whatever we call the effect of a factor, which could be written by shown by these plots here. These are each significant when you compare them to the background noise, unless it is significant when you compare that to the background noise, we will probably just say that factor really does not have an effect. It is just like my trying to say something. So, if I have very little voice you will hardly notice anything because, my voice level or the signal is at the same level as the background noise, we got to make sure when we plot these plots. When we find these effects they are at a level that is significant, when we compare to be a noise.

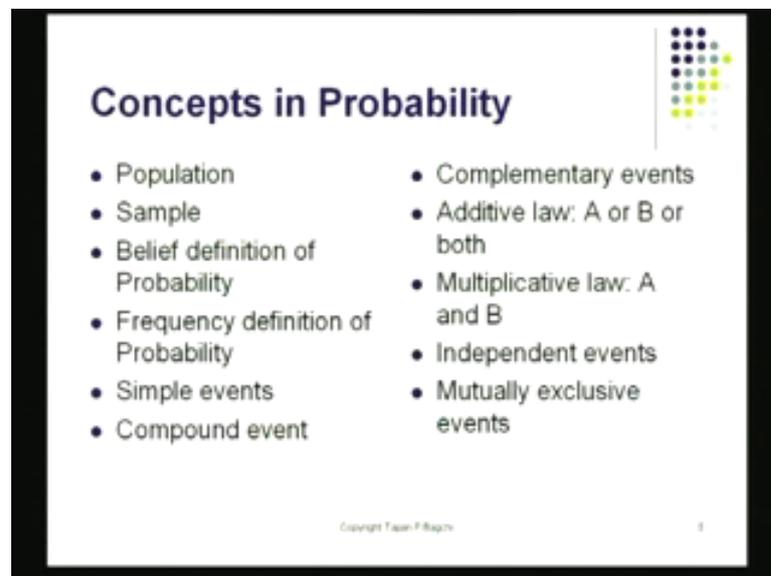
Unless you do that what is the point in controlling that factor because, the factors the same effect as background noise. So, it is not really going to be the activating control factor. Now, see I have got these things done, see I have got these plots done. And I can probably empirically optimize the process by picking the high setting for A low setting for B low setting for C. I can probably maximize Y that I can probably do, but is that really the mathematical optimum, is that the best possible optimum. Best possible optimums are producible, only when you got a mathematical equation and it can be optimized.

So, you can find the maxima or you can find the minima, it is only in under those conditions you can really optimize the process. To do that of course, it will not be good enough for me to just play with these charts here, these graphs here. I will have to build what you call an equation this is an imperial equation. And in the language of statistics this is called a regression model. Notice here, I have got the parameters A B C D, those factor effects are there I have got some of the multipliers with this. And I have got really a mathematical equation here, which is probably not as good as my, you know the equation that is used to design my digital watch. It is not as good as that because, here I have got a lot of good theory I have got Ohm's law, Kirchoff's law.

And all those things putting to this little device here, in the design of this device that will not be possible, when I am working with this little equation here. This imperical equation that you construct it, which I call the prediction model is really not going to be good enough. Is not going to be as good as the mathematical model that I have for my electrical device, but it is still pretty good. And with the help of this I can move to the next step, which is like optimized. And in optimization we do a couple of things we try to reduce variation and also we try to reduce defects, we try to cut losses and also, we try to reduce variation.

What is been my approach? what has been our approach here? It is been imperical, it is not being like working out the theory and then finding the optimum points it is not being that way. We run some experiments with the real process; this is really the approach that is utilized in six sigma. Six sigma is very empirical, you work with the real process it changes the different settings of the different parameters, different factors. And then it tries to find out which factor is most active? Which factor is not? And then you try to find the optimum settings for each of those factors and then you get either lower losses or reduced variation.

(Refer Slide Time: 15:34)



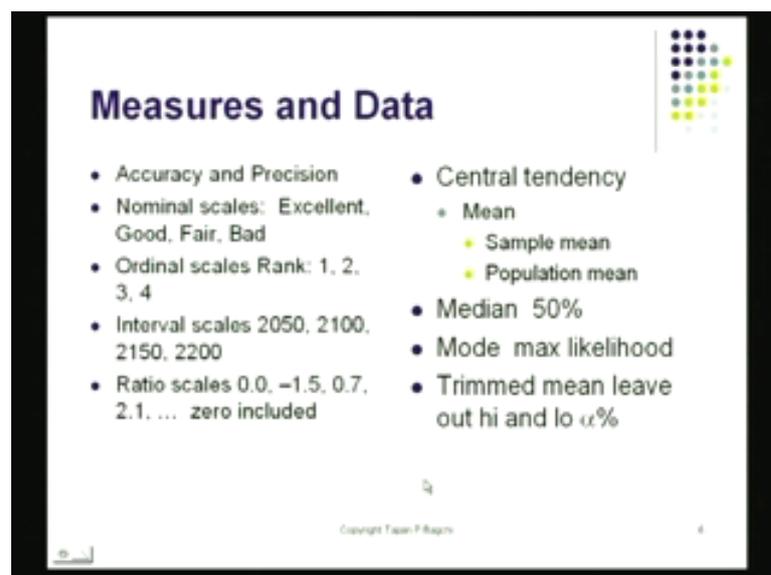
Concepts in probability there are variety of things here, I have listed out a few things there. For example, there is the concept of population, which is like any large body that consists of products or it consists of people or it consists of books or any object that is

like slightly different from the other. So, there may be a particular population or production. For example, like one days full production could be a population of widgets, like far as producing pens, then a full days production might produce 5000, might produce 5000 pens. And that those 5000 pens then they would constitute the population.

Now, if I want to do quality testing I would not be able to sample all of them, I would not be able to really test all of them. So, what all I have to do is I just have to pick a handful I will pick a handful of those pens, that they are coming out of a production I will just grab a handful of them. And then I will start looking at my sample and I will examine each of them one by one. And then I will try to then assess quality of this to try to make a make basically a statement about the quality of the full days production, that is what I will try to do. So, I have population which is the full days production and I from that I remove a sample and I have something that I called a sample.

Most of the time in statistics we play with the sample and in doing that of course, there is the chance of probability that is there. So, we study a little bit of probability, we study something called frequency of the various characteristics. And that is the frequency distribution of frequency definition of probability. Then I have got some, simple events and then in real life, we have compound events I am going to be describing all of them as we go along. So, is the list of things that we will be looking at this is our first slide.

(Refer Slide Time: 17:19)



## Measures and Data

- Accuracy and Precision
- Nominal scales: Excellent, Good, Fair, Bad
- Ordinal scales Rank: 1, 2, 3, 4
- Interval scales 2050, 2100, 2150, 2200
- Ratio scales 0.0, -1.5, 0.7, 2.1, ... zero included
- Central tendency
  - Mean
    - Sample mean
    - Population mean
  - Median 50%
  - Mode max likelihood
  - Trimmed mean leave out hi and lo  $\alpha\%$

Copyright Team P. Rajan

Then I have got some details on the way, we collect data or data may be our details basically to try to detect two things or is the process delivering an accurate performance. And is the precision of process good that is something that we will have to find out accuracy and precision. And you know in order to be able to do that I have to collect data. And the data will have to be, will have to be measured by some instrument and it might produce results like categories of data. For example, excellent good fair bad, it could be that way or I could rank them from the best one to the lowest one.

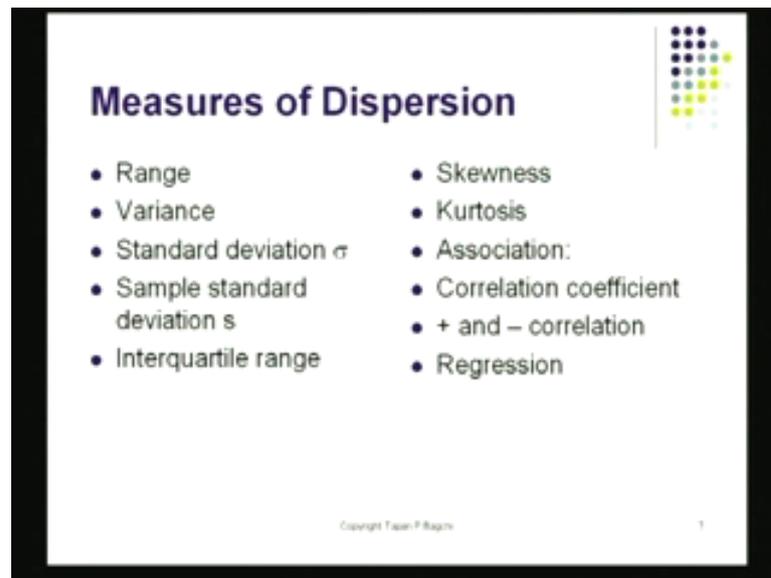
So, that also could be there ranking would be there I could also use interval scales, when I really have slots and in that, we dump data, we dump basically we put some data there here and there and so on. And I end up with what we call the use of the interval scale and of course, the most popular one is the ratio scale that produces real numbers. Then of course, when we got data collected, we would like to know, where does most of the data stand and this is something that, we will have to do by slowly then getting into the nature of the data that I have collected.

So, we will like to know for example, what is the central tendency of the data? Where is most of the data concentrated? And this is usually concentrated around the mean. And now, because I have drawn a sample from the data, from the, my population there will be something called the sample mean. And there will be something called the population mean, both are there, there is the sample mean and of course, there is the population mean. Generally speaking the population mean is unknown. And we will try to estimate that by working out what if I have calculating? What the sample mean is? So, the sample mean becomes the estimator, estimate for the population mean, there is another way to say where is most of the data concentrated.

And that is to take a look at the median of the process. And the median basically is a quantity that has equal number of in terms of frequency of probability, equal number of items below eight and above eight. So, in fact, again I am going to be giving you details on this I just want you all to know, that besides mean there are other measures. And that tend to indicate for most of the data is concentrated and this tendency is called the central tendency of the data. Then of course, we got something called mode, which is mode is the point for most of the data they seem to have the highest probability. And again I am going to be showing this to you, as we get into this.

Then sometimes of course, we trim the tail ends of the distribution, tail ends of the distribution that we end up producing, as a result of some data collection. And we end up, what we call the trimmed mean, the trimmed sample mean of the data. That is like something that we will do again once we get into this analysis.

(Refer Slide Time: 20:12)



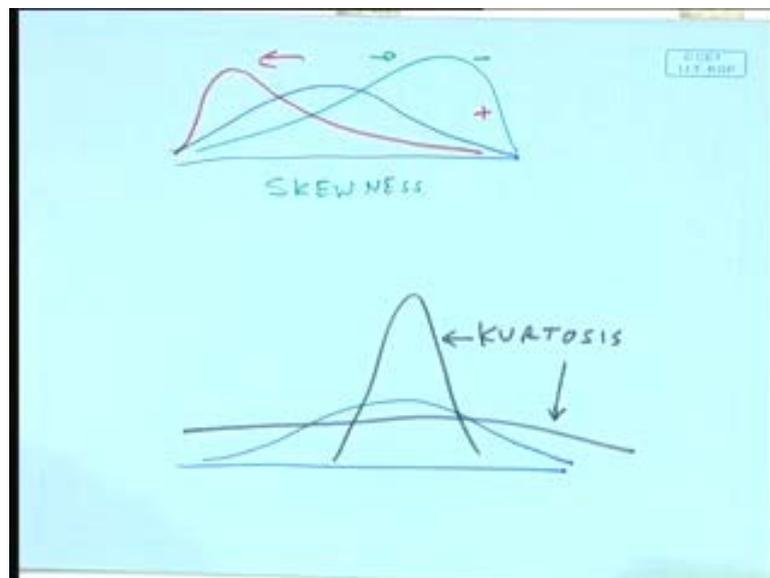
Now, that was the central tendency of the data, then there is something called the dispersion of the data. How the data distributed? How they spread around? Are they really concentrated together? When I got good precision or they spread around a lot, the measure for this is, the measure for dispersion. How we find that out? Very the, very first range is called range. Range is the difference between the max, maximum value of the observation and within the same sample the minimum value of the observation. So, if you collected five pieces of data  $x_1 \times x_2 \times x_3 \times x_4$  and you find that  $x_2$  is the highest number. So, put that aside that is the max and if you find  $x_3$  is the smallest number, then that becomes the min.

And the difference between  $x_2$  and  $x_3$  which is the difference between max and min, that becomes the range for that little sample there. Range is also a very good indicator of dispersion of the data. And of course, the data that is theoretically most useful is the standard deviation, it is the square root of the variance of the data that you have collected, that also is something that we will be looking at. Then of course, we have got a population standard deviation, which is sigma and then we have got something called the

sample standard deviation. This sample standard deviation is denoted by  $S$  and that actually is an estimated for sigma.

Sigma is the population standard deviation, then we got a quantity called inter quartile range which is the range between the third quartile and the first quartile. The difference between the first quartile and the third quartile, that space there is also an indication, how widely the data is distributed. All of these are basically dispersion indications of dispersion.

(Refer Slide Time: 21:58)



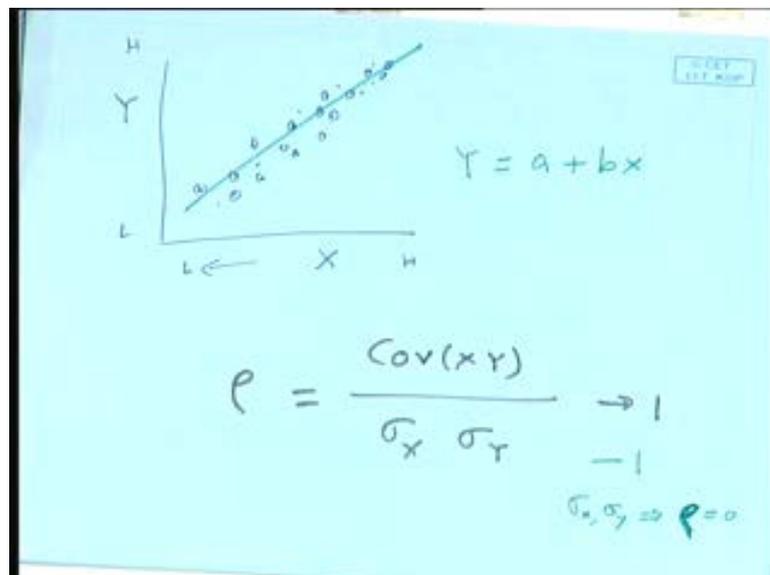
Then I may have a situation, when for example, if you look at my chart here. The normal distribution is like this symmetric, but sometimes what happens that the data tends to get skewed. So, this distribution could be like this, that is a skew to the left and this is called actually a positive skew. And it could also be that my skew turns out to be going to the right and the data could be like this, this is here. I have got most of the data going to the right this is called a negative skew. And there are formulas available that tend to tell you how this distribution is shaped? The shape of the distribution would now come from this Skewness of data. So, the word here is Skewness, there is another way I can represent the shape of the data.

That is to do it again when I start comparing that I have a normal plot, which is symmetric like this that is the standard one. It is very possible that my data is flatter than this, my data is flat like this or it is peaked like this, both of these are different from the

standard normal distribution. And these have what we call kurtosis; Kurtosis is a peculiar property which will be there in a piece of data like this or a flat piece of data like this. And this is something, we got to keep in mind when we are looking at data, we should not be using the normal distribution theory. We should be using some other theory that would be appropriate, when you got a situation like this. So, just to give you an idea I just plotted these two little graphs there.

And I gave you some indications of Skewness and of Kurtosis. Now, it is very possible, many times what happens, you have data and the data basically can be a plotted, there may be two variables from which I have collected data and my data.

(Refer Slide Time: 24:06)



My data can be put on a scattered diagram and the scattered diagram goes like this. This is x and this is y. These are two different observations and this could be for example, sale of coffee and this could be temperature. Sale of coffee at the coffee shop and this could be ambient temperature. You might find when I do this I tend to find a scattered plot which is like this which will be like this. Actually there should be probably true, if temperature is high this way and low this way. Then this would be the sale of cold drinks. As temperature becomes low, the sale of cold drinks become also become slow and as temperature goes up, the sale of cold drinks would become high. So, these two variables then they will rise together and they fall together.

This is a scattered diagram, to try to make sure you can see them. I am just going to circle a few of them and you will be able to see the scatter a bit more clearly. These are little flies and they tend to be rising together or falling together, you see that scatter there. Now, it is very possible that underneath there is some relationship. That is the relationship between  $x$  and  $y$  and that probably could be represented by a little model like that. So, first we draw a scatter diagram we take any pair of data. We took a height and weight of people. It could be anything else, we have got two data that might have some relationship and in order to discover that what you do is? Draw a scatter diagram and you end up with this little diagram.

If it shows this sort of tendency then you go for regression. Then you try to fit  $y$  as a function of  $x$  and it could just turn out to be this  $y$  is  $a$  plus  $b$   $x$ . This then becomes your predictor equation. Given some value of  $x$  you are able to predict  $y$ ,  $a$  and  $b$  are the parameters of the model. This is also used a lot when you are trying to optimize something and this is actually indicating association, but notice here not all the points are straight on that direct, directly on that straight line. There is some variation around this and that variation actually reflects the effect of not the only  $x$ , but perhaps other factors which also might be impacting  $y$ .

Therefore, what we have to do is? We have to work out something called the correlation coefficient or the co variance, covariance of  $x$  and  $y$ . If the co variance is 1, if the covariance is really high then of course, you could say that I can predict  $y$  perfectly given  $x$ . But if it is less than that then I need to have more terms in this model in order to be able to predict  $y$  as precised as I would like it to be. If that is the situation we will have to really collect more data and perhaps collect data on the third front also and fourth front also. And our regression is going to be then more complicated, but it can be done it is done all the time.

This is the empirical way to build a model. This is like using experiments to construct a model which eventually would look like my digital watches you know that fine equation that the electrical engineer worked out. Correlation coefficient of course, is a term that is used a lot and before that I have what we call co variance. The correlation coefficient it goes like this. It has got co variance  $x$   $y$  divided by sigma  $x$  and sigma  $y$ . What is this term? This term if it is like 1 then you will say  $x$  and  $y$  have a strong positive co relation.

If this is minus 1, which will be the case when I have got  $x$  and  $y$  going the other way. If it is minus 1 then you will say that  $x$  and  $y$  they will move opposite to each other.

And in many cases of course,  $x$  and  $y$  are not related in which case you will probably find  $\rho$  to be such that  $\sigma_x$  and  $\sigma_y$ . You find that  $\rho$  tends out to be pretty much close to 0 which means there is no dependency of  $x$  and  $y$ . Let me write  $\rho$  there in place of  $\sigma$ . This  $\rho$  if it turns out to be close to 0 there is no relationship between  $x$  and  $y$ . So, this is again something that is useful if am using one variable to try to control the other and this is what I indicate by saying that correlation can be plus or it can be minus.

And I have already discussed about regression which is really a model like this. Which will be a model like this; this is a simple regression model. More complicated models they will have more variables. These variables on the other side they are on the right hand side. They are called independent variables and this guy is called the dependent variable,  $a$  and  $b$  are the parameters of the model. Then of course, we got the idea of probability and probability distributions and let us quickly take a look at these things. We have something called random variables. I am going to give you some examples here. Random variables are variables that depend on events.

So, you could probably say if a girl enters the room, I will mark it as one. I will mark the value of the random variable as 1 and if a boy enters the room I will mark the value of the random variable as 0. Here by counting out numbers of 1s, I can find out how many boys have entered and by finding out how many 0s are there, I can find how many and so on. So, this can be done quite easily. If this is done I have a measure now and if the events; if they arrivals of the girls and boys is random. Then of course, the output is going to be also random and the random variable now that is indicating the kind of people who are walking in. And that is been indicated by 1 being girl and 0 being boy.

I will end up with here a random variable that will be influenced by a probability. It will be influenced by some randomness and this randomness would depend on the kind of people, who are walking in I could have a continuous random variable, many random variables like height, weight and so on. Those are continuous and I could also discrete random variables in which case I am basically countings an event. And I am looking at the count as the value of the random variable. In this case the random variable is going to

have discrete value. It will be called a discrete random variable. If you look at the tossing of a coin, that is an event that has got 2 outcomes head and tail.

When I toss it, if I toss it a large number of times then of course, I end up with an estimate of the probability of my finding the lighter of my finding a head when a random toss is made. This is good because I am collecting a lot of data, I can rely on it and so on and so forth. So, I have probably toss 500 times and I found a number to be pretty close to 0.5 maybe it is 0.501 or it is 0.497 or something which is pretty close to 0.5 and I have got an estimate there. Tossing a coin can be too complex situations also. For example, if I toss 3 times. What is the chance of my finding no heads? What is the chance of my finding one head? What is the chance of my finding two heads in the last two tosses? The first one being tail.

These are complex events and the probability of these can also be found. So, in fact what I have to do here is I have to find a method. Just like we have plus and minus and multiply and divide to play with numbers. I must have a system now to play with probabilities also. We can add probabilities we can multiply probabilities I need a system for that. For that I need some ground rules. These ground rules are called postulates of probability. These have been given about 100 years ago, more than a 100 years ago. These were worked at just like the number system it came 1000 years ago, more than a 1000 years ago. So, probabilities also were looked at an a 100 or 200 years ago people worked on the algebra for probabilities adding and subtraction probabilities.

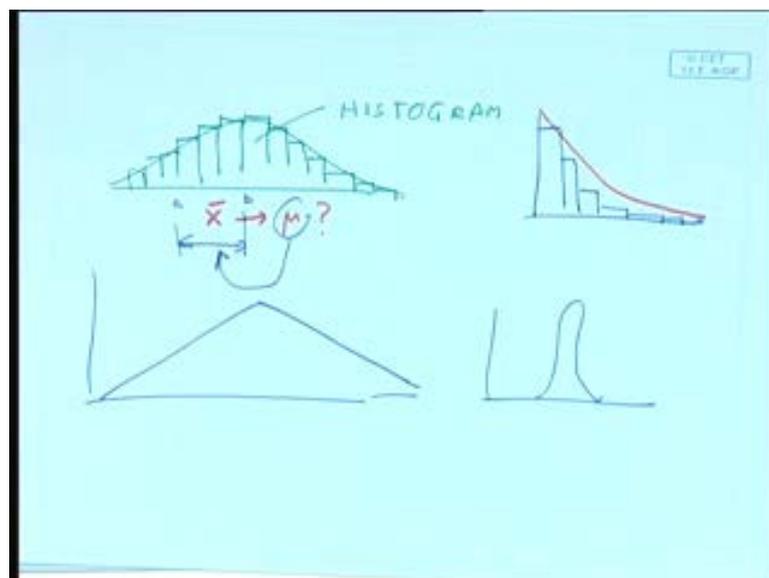
Tossing a die now a die as you know it is a thing that has got six heads. So, if I draw a little die here it would look like this. This is a die and the distinct thing about the die is it has got six faces and it will have certain number of dots on all sides. So, I could have 3 there 5 here and 2 here. These are the different faces and they are different numbers there. When I toss a die, when I throw a die the chance of the number being the number being read being 2 or 4 or 6 or 5 is going to be one-sixth because the die is symmetric. And the die when it is tossed it is thrown randomly. So, there is no likelihood of three being there all the time. It could be any number between 1 and 6.

What is the chance of my finding 0s here? The probability of my finding a 0 reading there is 0 because there is no face with a 0 with no dots there. Similarly, I cannot find a number 9 there because it is just not there in the sample space. What we have to then do

is we have to really define what we call the sample space in order that we can go out and start defining probabilities. So, we have something that we call the sample space and again I am going to give you some examples there. Then we will discuss something called the probability distribution function. We will have a probability density function, we will have a cumulative distribution function, we will have something called a expected value which is really the average.

We will also have something called as variance and these are values that are realized by looking at data that I collected when I was collecting the sample. If you collect a lot of data then it tells out that the data itself begins to take some shape. Let me give you some examples of the kind of shape data might like to pickup.

(Refer Slide Time: 34:21)

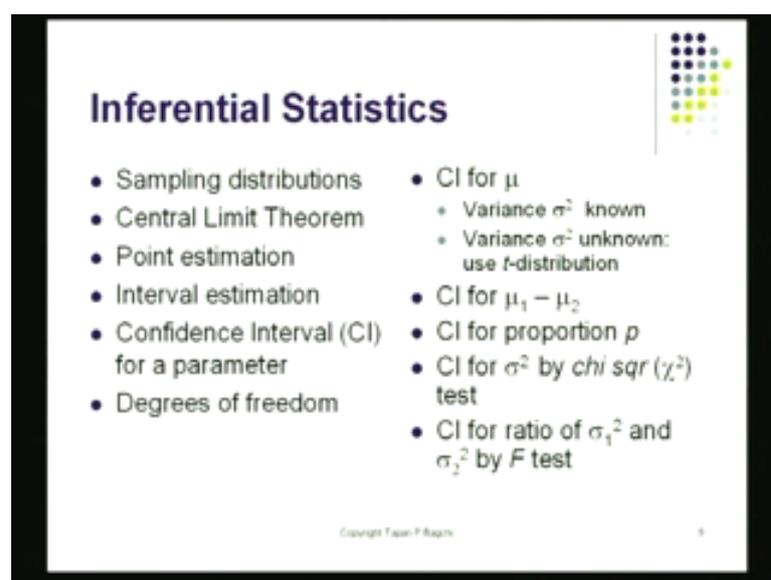


In some cases the data might take a shape like this which is the normal distribution. In some case when you do and what is this data really? Usually we will not find a continuous curve, but you will find what we call a histogram. (No Audio Time: 34:33 to 34:45) This is what you find in real. So, this is actually these bars are the histogram if you take lots and lots of data. Eventually it will begin to look like a continuous curve and that will be the distribution that we will represent. So, this histogram is an approximate representation of the real distribution that is there. And this will become a plant as you will collect more and more data.

So, some processes they lead to this sort of symmetric output. There are other processes that lead to this sort of output. Let me just give you a picture here. So, what is the shape here? The shape here would look like something like this. This is the explanation distribution. So, it is a different distribution. There are many other distributions some look like this; others have some other shape they could be like this and so on and so forth. So, there are a variety of different distributions and these are all found in nature and that is why people came out with different types of probability distribution models. Like for example, there is something called the hyper geometry distribution, something called the binomial distribution, something called the poisson distribution, normal distribution and so on and so forth.

These are all useful in studying random phenomena. Whenever you collect data it will probably come from the one of these things and this is a lot of very rich theory available with each of these distributions. And basically exploiting that we can by exploiting that we can make statements about a process then this is given by let us say the binomial distribution or the exponential distribution or the normal distribution. In quality assurance and in particular when you are going to be working in six sigma. You will probably have to know some of these distributions in order to make statements which others can also come back and reproduce. That is like that is something that we would like to be able to do.

(Refer Slide Time: 36:48)



**Inferential Statistics**

- Sampling distributions
- Central Limit Theorem
- Point estimation
- Interval estimation
- Confidence Interval (CI) for a parameter
- Degrees of freedom
- CI for  $\mu$ 
  - Variance  $\sigma^2$  known
  - Variance  $\sigma^2$  unknown: use *t*-distribution
- CI for  $\mu_1 - \mu_2$
- CI for proportion  $p$
- CI for  $\sigma^2$  by *chi sqr* ( $\chi^2$ ) test
- CI for ratio of  $\sigma_1^2$  and  $\sigma_2^2$  by *F* test

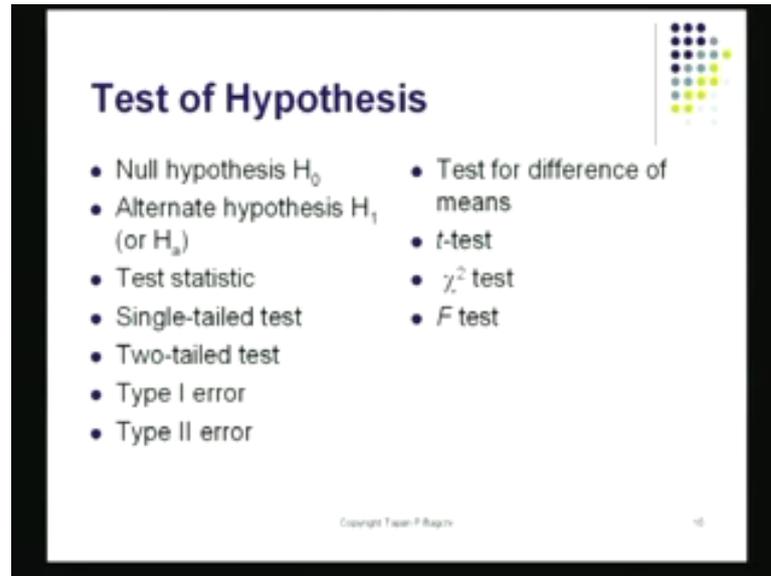
Copyright Tapan P. Bagchi 9

Then of course, many times we have collected some data, but we do not know what that true mean of the population is? In that case we will have to do something called inferencing. Inferencing basically says I have collected a certain amount of data and I am going to be using that data now to come up with some estimate of that true mean which is  $\mu$ . I could do the same thing for variance and I could do the same thing for difference between means or the proportion of parts or the defective in production. In fact we have many different measures that can give us pretty decent ideas of the ideas about the what we call defects, distributional defects or the kind of problems that we are having in a particular situation.

There are some special situations when we will need to do some testing. Testing tests are basically mathematical processes, these are mathematical tricks or methods which are utilized to make sure you really have a decent idea of what that true mean is? So, I may have a quantity. Now, for example, I had  $\bar{x}$ ,  $\bar{x}$  calculated from sample that I collected from this normal distribution. I calculated  $\bar{x}$ . How close is this  $\bar{x}$  to the true value  $\mu$ ? How close is this? This statement I can make only when I apply what we call inferential statistics. So, there is some theory there that we will be utilizing to be able to say with some confidence that I am 95 percent sure.

That the value of  $\bar{x}$  is such that around  $\bar{x}$  I can put a band and there is a 90 percent chance that this guy is in that space. There is a 90 percent chance that around  $\bar{x}$  I can put a limit between  $a$  and  $b$ . And there is a 90 percent chance that the unknown  $\mu$  is going to be residing there. There is a 90 percent chance for that. That means if I repeat this process of calculating  $\bar{x}$  a 100 times, most correctly about 90 times, I do not know 100 times. The true  $\mu$  is going to be within that. That is the idea. This is like one statement of inferencing and we do the same thing for variance and for different parameters that are also unknown.

(Refer Slide Time: 39:05)



**Test of Hypothesis**

- Null hypothesis  $H_0$
- Alternate hypothesis  $H_1$  (or  $H_a$ )
- Test statistic
- Single-tailed test
- Two-tailed test
- Type I error
- Type II error
- Test for difference of means
- $t$ -test
- $\chi^2$  test
- $F$  test

Copyright Team P. Bagchi 16

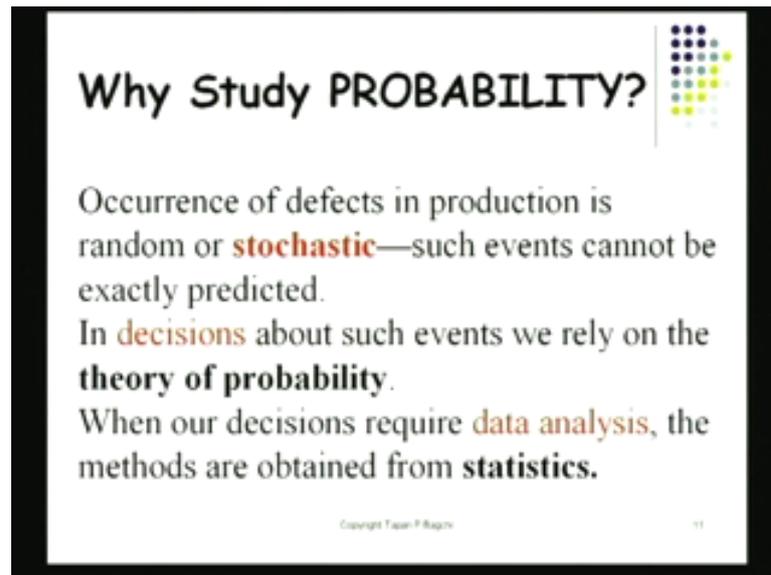
If we move along we also have this idea of test of hypothesis. For example, if you wanted to say for example, we know the first example that we had. When we had the I am just going to bring that up. This diagram here does A have an impact on the process. Does factor A have an impact on the process? This could be posed as a hypothesis or a guess. There is a way to collect data, there is a way to change the settings of A and look at the output. And there is a way to test the data and this data testing will tell us whether it is reasonable to assume that a has an impact on the process or it does not have an impact on the process. The statistical procedure here is called test of hypothesis and that is what I have here in this slide here which is the test of hypothesis.

There I set up certain conditions and I call them the null hypothesis and the alternate hypothesis. Then I carry through a process, I carry through a data analysis process. And that process ultimately comes back and tells me yes indeed, there is reason to believe that A has an impact or no I do not see the evidence that a has an impact. Because, whatever A is doing noise is doing the same thing. It is like Doctor Bagchi is trying to say something, but perhaps he is not saying anything, because what he says if there is no sound, if there is no signal that is above the background noise you will probably say Doctor Bagchi is quiet. He is not saying anything, there is no signal.

Factor A has no impact on the process and these of course, these different test they came. They have different applications and they would come as one tailed test or two tailed test

and they involve something called the type one error and type two error and so on and so forth. As we get deeper into this subject you will find I am utilizing these ideas, these concepts. And I am doing the t test sometimes, I am doing the f test sometimes, I am doing the chi-square test sometimes and sometimes I am doing a plain and simple zee test. Those we will be doing as the time comes along.

(Refer Slide Time: 41:14)



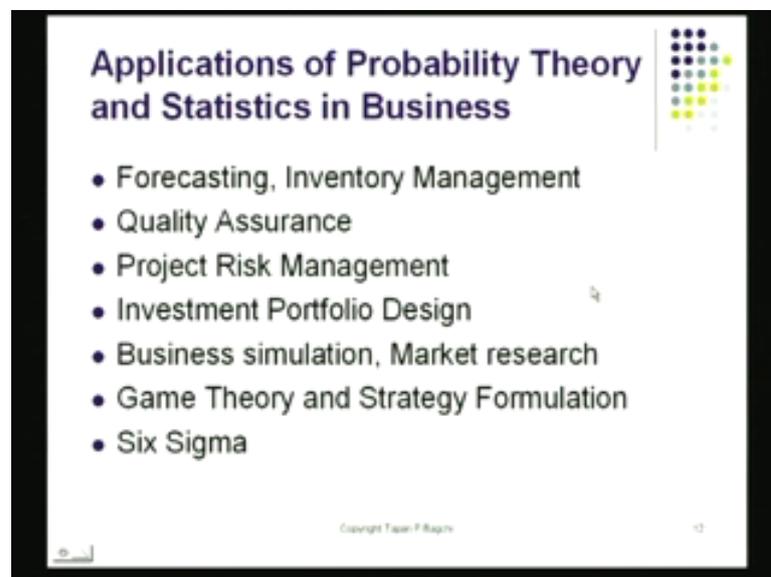
Now, what is probability? Why do we study probability? You see that phrase there called stochastic. It is a very, very important point, it is a very important phrase. Stochastic means random, it is a fancy word that the learned people they have used. Random is a word that normal people use **the** you and I would use the word random, but if I speaking as a professor I could not call it random, I would call it stochastic. So, a random variable is also a stochastic variable. So, please do not be flustered by this fancy term called stochastic. It means the same thing as random. Now, we have many, many decisions and these decisions are generally based on events that cannot be predicted exactly.

Under such conditions, we still cannot escape decision making, but we are doing this in an environment that has got lot of uncertainty in it. In that case we really have to rely on the theory of probability. We will have to rely on the theory of odds and not just simple coin toss, but also some rather complicated situations, when the events are interdependent, the events have conditionality and so on and so forth. Many other complications are there in with random events that is what we will have to do in order

first we will be able to make sound decisions and of course, when I am banking all of these things on data analysis.

I will probably observe something or I will conduct some experiments, I will collect some data. That data will be subjected through what we call data analysis. Once you have done that we will have some idea about the randomness. We will have some idea about the stochasticity of the process and then our obviously our results are going to be far more reliable.

(Refer Slide Time: 43:06)



Where do you apply probability theory? Let us take a look at this well many of you have probably done forecasting of some sort. Forecasting is one place where you apply both statistics and probability. Inventory management when you are trying to decide on the safety stock for example, there you will be using probability. And also you might be using some statistics based on past history and so on, fluctuations of past demand and that kind of thing. Quality assurance is one area that cannot escape probability and statistics because in real life when you go to a real factory producing things. There are many factors that are not in control.

Therefore, the output is variable. Once the output is variable to try to keep it in a confined two values which the customer is going to accept all I have to understand this randomness. All I will have to understand this stochasticity and of course, for this I may use a control chart to plot it. I may use what we call sampling. I may use designer

experiments; I may use multiplication modelling and so on and so forth. So, variety of techniques which are now borrowed from statistics though they utilize by people who are doing quality assurance and of course, six sigma. Six sigma uses DOE design of experts in a very big way.

Project risk management this is like one large area where things can foul up and this is by Murphy's law. Whenever you have got a project, it is very possible that certain assumptions will not hold true. For example, you are going to be building a factory and you are counting on the demand, market demand to be rising and rising and rising. If it does not happen then of course, your project is going to be a failure. But demand is not a linear function always. Demand will have some fluctuations, it will probably begin somewhere. Then it will begin to fluctuate and fluctuate and fluctuate and so on. So, to somehow catch this up to somehow understand this fluctuation and on the basis of that I will have to come up with a risk management plan.

So, whenever I have got; when I have a project going, I must do risk analysis and risk analysis strongly depends upon probability theory. There are some special methods and we may get a glimpse of this toward the end of this course when we discuss six sigma projects. We might come back and discuss a little bit about project risk management that we might do. Investment portfolio design, just this afternoon I was speaking with a senior professor and he said he is got about 50,000 rupees in his pocket. And he would like to design a stock portfolio to be able to protect his investment and at the same time generate some good returns.

Now, so his task is now going to be the market place consists of all kinds of stocks. They have different types of returns; they have different types of fluctuations, variations and so on. And what you would like to be able to do is somehow balance this risk and return. And for that you would have to optimize the composition of this portfolio. So, that is like another area where we would be using probability and also we would be using statistics. Those are also utilized there. Business simulation is one large area and market research. These are large areas where we use a lot of statistics. Game theory is a big area where strategies are found and this is actually an area that uses probability in a very big way.

And I have already mentioned six sigma uses in it is one of it is steps and I am going to write that again it is called DMAIC. DMAIC define the problem the issue that you want

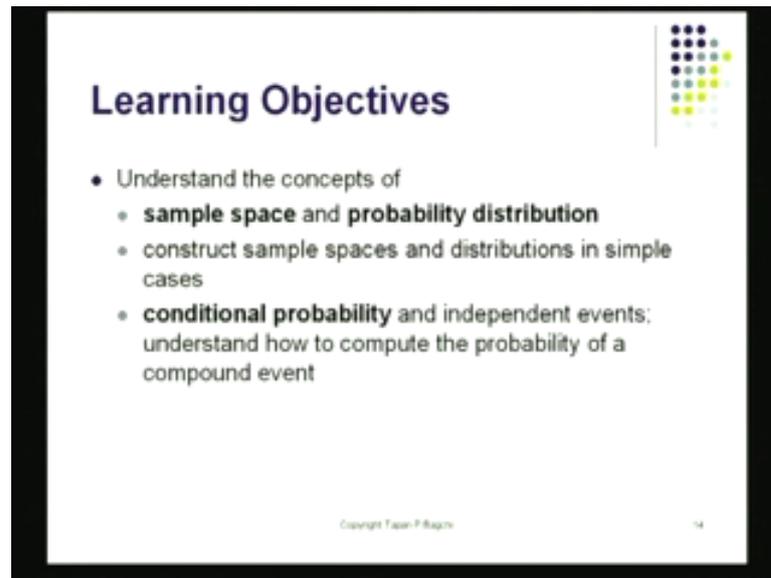
to tackle, measure, analyse, improve and control. It is the improvement step where I begin statistics and probability. I actually begin what we call design of experiments? So, D O E is used right here and D O E comes on statistics, statistics gives us D O E. So, if I am trying to use the DMAIC framework, I will not be able to escape statistics. I will have to use a weinmann form or the other.

(Refer Slide Time: 47:13)



I would be discussing some of these things. I will be discussing outcomes; I will be discussing random events and so on. I will be elaborating that as we go in the afternoon as we go deeper into this. Well let us try to get an idea of what all various things we will be discussing? I will be discussing something called what exactly is probability? What are some of the basic rules for combining probabilities for adding them, for multiplying them and so on? What are conditional events? What are compound events and how do I work out the probabilities for compound events or conditional events? We will also get a glimpse of the distributions. We will also try to get an introduction there and also we might get a glimpse of what we call hypothesis testing? That also we will try to attempt. That is a lot of things and we will see what we can get to.

(Refer Slide Time: 47:54)

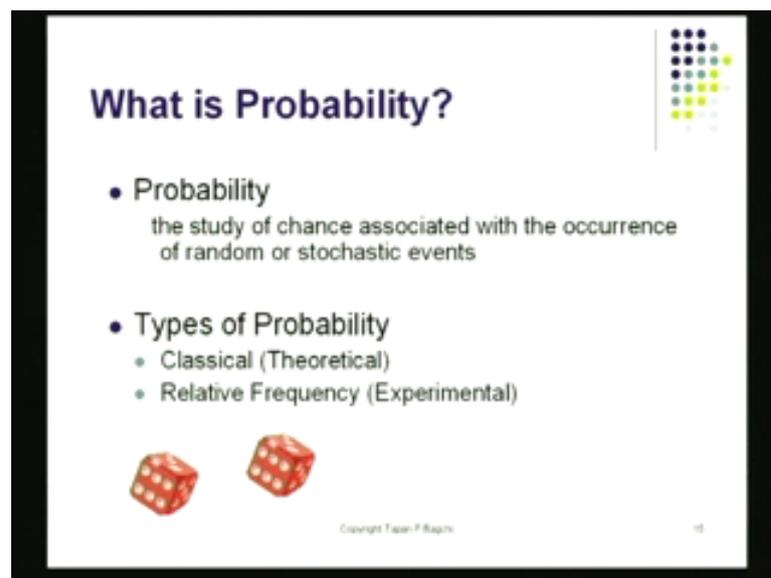


## Learning Objectives

- Understand the concepts of
  - **sample space** and **probability distribution**
  - construct sample spaces and distributions in simple cases
  - **conditional probability** and independent events; understand how to compute the probability of a compound event

Copyright Team P. Nagin 14

(Refer Slide Time: 47:59)



## What is Probability?

- **Probability**  
the study of chance associated with the occurrence of random or stochastic events
- **Types of Probability**
  - Classical (Theoretical)
  - Relative Frequency (Experimental)



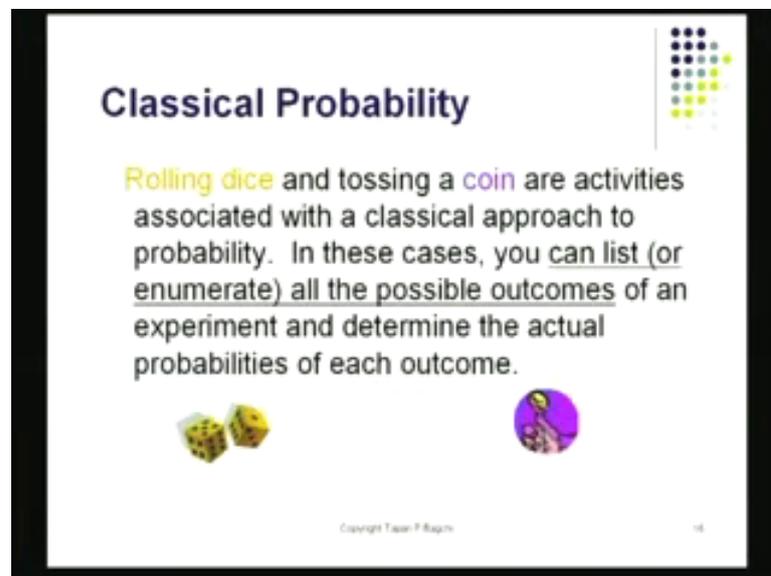
Copyright Team P. Nagin 15

To begin with what exactly is probability? This is something I would like to we are reaching the end of this hour and I would like to make sure you have some idea of what is probability before we end here? It is the study of chance associated with the occurrence of the events and these events are actually random events. It is the study of chance; that is what probability theory is. What kind of probabilities are we are talking about? We may talk about probability it is like the chance of a rain when it is cloudy and people who have some experience with monsoon clouds and so on.

They may look at the colour of the cloud and based on they may say there is a 50 percent chance, it is going to rain or say 80 percent chance, it is going to rain or there is a 5 percent chance, it is going to rain. This sort of statement is not based on really hard data, it is based on subjectivity. And this is really the basis many times this is the only basis for working out probabilities. Sometimes, what we do is? We construct a theoretical model and on the basis of the theoretical model we make predictions about the probability of complex events that we do. If am not able to do that, I will probably to sit down and collect a lot of data or run some experiments and from that figure out the probability of certain events.

So, to try to estimate probabilities, first of all what are probabilities? Probabilities basically are a study of any event that can occur by chance and we are interested in the outcome of that outcome of that experiment. There are two ways I can measure probability. I can measure probability either by theoretical consideration and I will show you how to do that? or I can measure probability by relative frequency which is like by calculating based on hard data calculating the odds that also I can do.

(Refer Slide Time: 49:49)



**Classical Probability**

Rolling dice and tossing a coin are activities associated with a classical approach to probability. In these cases, you can list (or enumerate) all the possible outcomes of an experiment and determine the actual probabilities of each outcome.

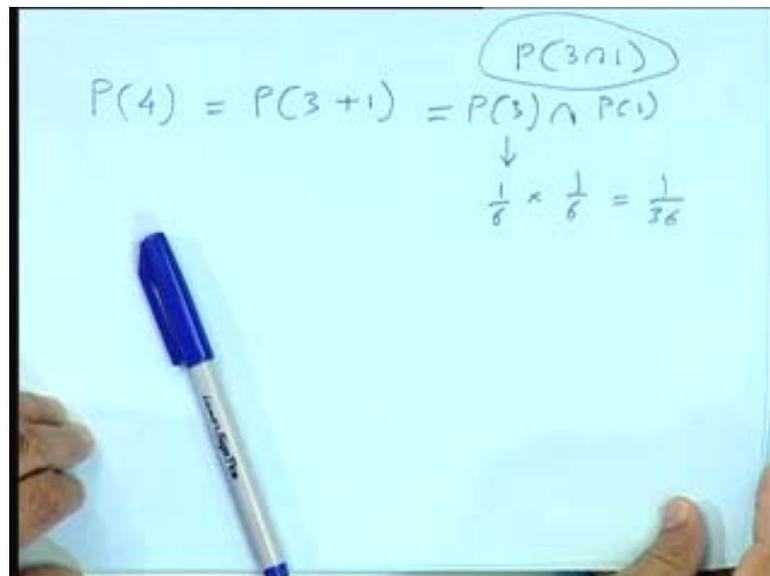
Copyright Taim F. Bagin 16

Rolling a dice, now here, I do not really need to toss a lot of coin or lot of dice. I do not need to do that. Here I can directly say that based on my experience and based on some little principles of probability theory, I can make a statement about if I roll the die two times, what is the chance that the sum of the two numbers that I see the first number and

the second number is going to be equal to 4? That I could do just based on my experience, I could do that and I am using some classical theory here. So, how could I get a number 4? I have 1 plus 3 or I have 2 plus 2 or a 3 plus 1. Of course, I cannot have any other number, any other pair of numbers that will add to a sum of 4 because none of those; none of the dice would give me 0. If I had 0 plus 4 of course, we would have another way to find 4, but that is not happening here.

I have got 1 to 6 on the first die and I have got 1 to 6 on the second die. So, I can add to get a sum of 4, I can have 1 plus 3 or I could have 2 plus 2 or I could have 3 plus 1, these are three different ways. Now, finding a 3 on a dice, finding a 3 is 1 by 6 and finding a number which is 3 on the first dice, 1 on the second dice that is also going to be 1 over 6. So, here the chance of rolling a die, 1 by 6 is the probability of rolling the first die and finding number 3 and 1 by 6 is the also the probability of finding a number 1 on the second dice. Therefore, for the two of them to occur together to give me a sum of 4 is going to be 1 by 6 multiplied by 1 by 6 that is one-sixth multiplied by 1 by 6. That is a simple calculation that I could do quite easily and it would turn out to be 1 by 6.

(Refer Slide Time: 51:52)



The image shows a whiteboard with handwritten mathematical work. At the top right, the expression  $P(3 \cap 1)$  is circled in blue. Below it, the equation  $P(4) = P(3 + 1) = P(3) \cap P(1)$  is written. A downward arrow points from the intersection symbol to the calculation  $\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$ . A blue marker is visible in the lower-left corner of the whiteboard.

So, one way to find 4 is going to be probability 3 plus 1, these are on two different dice. So, therefore, it is going to be probability of 3 and probability of 1. And the moment I do that the probability of first one is going to be 1 by 6 multiplied by 1 by 6. Notice here I have used a little bit of algebra and this algebra came from this notation. And here what I

have done is I have got really the probability of this guy 3 and 1 calculated. I have wrote it loosely like this, but really I should be writing like this. And the probability of this event 3 and 1 is going to be 1 by 6 and 1 by 6 which is 1 divided by 36. That is the probability of my finding a 3 plus 1 sum.

I could do the same thing for 1 plus 3 I could do the same thing for 2 plus 2. These are different ways by which I could construct my event 4. This is some of the numbers on the two dice to be equal to 4. So, here what I am doing is I am using some classical theory. I am using some classical theory. I could do the same thing for coin toss, I could define the coin toss in a manner. When I have a certain number of heads in a number of trials like if I throw it ten times what is the chance of my finding exactly three heads and seven tails? I could work this out using classical theory.

(Refer Slide Time: 53:20)

**Sample Space, Events and RVs**

The possible outcomes of a stochastic or random process are called **events**.

An event is a deterministic process has only one possible outcome.

The probability of a particular event is the fraction of outcomes in which the event occurs. The probability of **event A** is denoted by  $P(A)$ .

Random variables map **events** to **numbers**.

Copyright Team P. Raghav 4.3 X S A

The slide features a diagram where a circle labeled 'S' represents the sample space. Inside it, a smaller circle labeled 'A' represents an event. An arrow points from 'A' to a box containing the number '4.3', which is labeled 'X', illustrating the mapping of an event to a numerical value by a random variable.

Let us get down now, and very quickly define what we call sample space, events and random variables. The sample space actually is the set that consists of all the outcomes. For example, when I am tossing a coin, there are only two outputs possible; head and tail. So, the sample space here consists of head and tail only. So, all I really have, I have head, and I have tail, these are the only two outcomes when it comes to tossing a coin. These are by the way event, the finding a head or finding a tail, observing a head or observing a tail. These are the events or the coin appearing on head first that is going to be an event, and tail is going to be also another event. And these two events in I mean, I

will tell you later these are disjoint, and they are also mutually exclusive. There is nothing common between these two.

If I have an event which is random, I can talk about the probability of that event. For example, in coin toss probability of finding head or the probability of finding a tail. In this case is going to be the probability is going to be 0.5. What I could do is I notice here, I have got this sample space, which I have in the diagram I have marked it as  $S$ . The sample space consists of all the outcomes. So, in the head and tail case I will have a head setting there and also tail setting there inside the sample space and nothing else. Now, I could define a random variable based on that I could define it to be  $X$ , and I could map head to 1 and tail to 0.

So, I will have a little mapping done here. Head will be mapped to 1, and tail will be mapped to 0, and these are the values of the random variable. These are events and from the events, I ended up defining this random variable. And this is what I will be doing? As we move along we will be defining many different events, and we will be defining the obvious and the associated random variable with it. And we will end up with real numbers 1 or 0 or any other number that will turn out to be, that it is going to turn out to be. I am going to continue this talk as we go into the next session. Thank you very much.