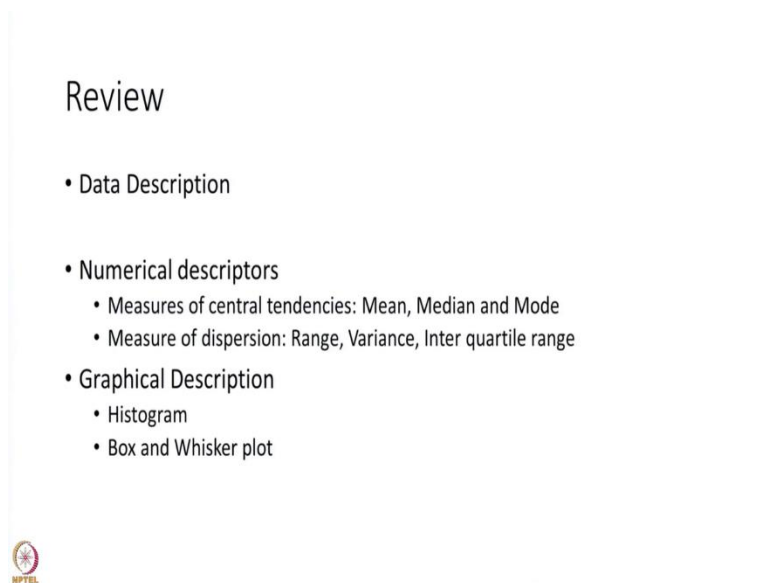


Dealing with Materials Data: Collection, Analysis and Interpretation
Professor. Hina. A Gokhale
Department of Metallurgical Engineering and Materials Science,
Indian Institute of Technology, Bombay.
Lecture 03
Probability and Distribution

Hello and welcome to Dealing with Materials Data course. In this today's session, we are going to introduce Probability and some distributions.

(Refer Slide Time: 00:35)




Review

- Data Description

- Numerical descriptors
 - Measures of central tendencies: Mean, Median and Mode
 - Measure of dispersion: Range, Variance, Inter quartile range

- Graphical Description
 - Histogram
 - Box and Whisker plot

 IITBOMBAY

Let us check what we did. We first worked out on what is called data description. So we worked out several methods of describing the data so that one can easily understand what is the data all about and what does it contain. So we presented here some of the measures of giving such a description.

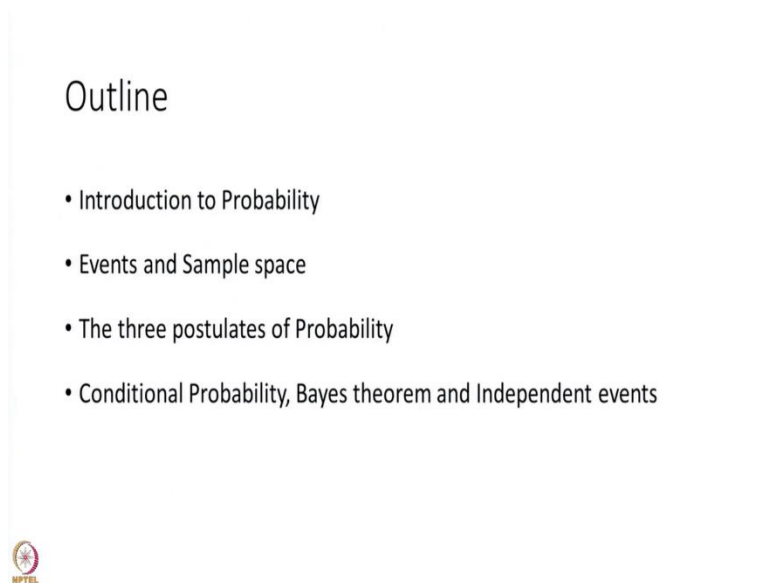
One of them is numerical descriptors such as measure of central tendency. So we have mean median mode etc. Then we also described the measure of dispersion which is range variance, inter quartile range etc. These are all numerical descriptors then we also presented some of the graphical descriptors such as histogram box and whisker plot, frequency plots etc. I would like to mention one thing here that number one this is not the exhaustive list of data descriptions.

There could be many ways and many ways can be invented to understand the relationship between within the big data sets of different variables. And also it can help us if you have a prior knowledge

what the data really represents then you can have different measures of describing data. We will see that as we went ahead as we go ahead with this course and if necessary we will have it as an appendix or an additional recording where we will show you some of this special kind of data description methods that we use.

However what is presented here is the most common methods of exploratory data analysis as it is also be called the same method is also known as storytelling in the data science parlance because in order to make people understand what is the data you have to tell a story about the data and therefore the word has come storytelling. But it means the same thing,

(Refer Slide Time: 03:04)



Today what we plan to do is we want to formally introduce probability through some definitions such as events and sample space, the three basic postulates of probability; conditional probability based theorem and independent events particularly Bayes theorem. In today's world of data analytics is playing a major role. We will talk about it when it comes to that.

(Refer Slide Time: 03:37)

Introduction to Probability

- We live in the world of chance!
- Occurrence of any event has probability associated with it
- Outcome of any experiment has an element of probability attached to it
- How to measure it? Generally,

$$Prob\{event\} = \frac{\text{no. of event specific outcomes}}{\text{Total number of all possible outcomes}}$$



So let us begin. We all live in the world of chance. What is happening now is only I know I mean this is our Indian philosophy you know what is now and you do not know what is coming up in future. So future is always a probabilistic and uncertain event and whatever happens it happens due to several circumstances and therefore we say that we live in the world of chance. So occurrence of any event that you can think of in day to day life is there is a probability associated with it, outcome of any experiment.

Let us come back to the world of material science and materials engineering and the world of experimentation. We can say that outcome of any experiment has an element of probability attached to it. Generally speaking we count in this way that we take it.

$$Prob\{event\} = \frac{\text{no. of event specific outcomes}}{\text{Total number of all possible outcomes}}$$

We will elaborate on this also briefly and in future. So then let us start with some formal definition of this event and the samples.

(Refer Slide Time: 05:02)

Probability...Some definitions

- An experiment can result in different outcome, even though it is repeated under the same manner every time is a Random Experiment
- Set of all possible outcomes of a random experiment is called the sample space of the experiment, and is denoted by S
 - S is discrete if it consists of finite or countable infinite set of outcomes.
 - S is continuous otherwise (i.e. it contains an interval of real numbers)



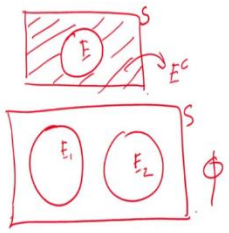
So we say that an experiment can result in different outcomes, there is a spelling error here. We have to say outcomes even though it is repeated under the same manner every time it is called a random experiment. This you must have experienced to right from the day you started doing physics experiment or chemistry experiment in the small high school laboratories. Set of all such possible outcomes of a random experiment.


You pick up a random experiment and consider a set of all possible outcomes that come out of this random experiment is called a sample space of the experiment and it is generally denoted by S . We say that S , the sample space is discrete if it contains all finite or countable or infinite set of outcomes. Suppose we are conducting an experiment of counting number of atoms that comes out of three DAPFIM then they are the number of atoms and therefore it is countable set and therefore S the even space becomes a discrete. S is continuous if we say that it, the outcome is contained in an interval of set of real numbers.

(Refer Slide Time: 06:37)

Definitions...cont..

- An Event is a subset of the sample space S
 - Union : $E_1 \cup E_2$
 - Intersection: $E_1 \cap E_2$
 - Complement: E^c
- E_1 and E_2 are mutually exclusive: $E_1 \cap E_2 = \emptyset$
- $(E^c)^c = E$





So now an event, we have defined a sample space we are defining an event. An event is a subset of a sample space S . We generally denote it by E , A , B some of the capital letters in our capital alphabets.

- Union : $E_1 \cup E_2$
- Intersection: $E_1 \cap E_2$
- Complement: E^c
- E_1 and E_2 are mutually exclusive: $E_1 \cap E_2 = \emptyset$
- $(E^c)^c = E$

Of course we know from the set theory that a compliment of a compliment is the set itself.

(Refer Slide Time: 08:46)

Probability of an event - The Three Postulates

1. Probability of event $A = P(A) \geq 0$, for any event from S $P: S \rightarrow [0,1]$
2. $P(S)=1$
3. If A_1, A_2, A_3, \dots are finite or infinite mutually exclusive events of S then,

$$P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$$

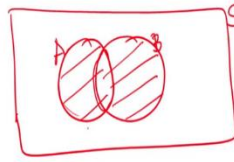
So next we come to the three postulates of Probability:

1. Probability of event $A = P(A) \geq 0$, for any event in S
2. $P(S)=1$
3. If A_1, A_2, A_3, \dots are finite or infinite mutually exclusive events of S then,

$$P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$$

(Refer Slide Time: 10:09)

thus we have....



- $P(A^c) = 1 - P(A)$
- $P(\emptyset) = 0$
- If A is discrete then, $P(A)$ is sum of probability of individual outcomes comprising A.
- If an experiment result in any one of the N equally likely outcomes, and if event A comprises of n such individual items then, $P(A) = n / N$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



Now from this we can derive a few things. It is very obvious that probability of $P(A^c) = 1 - P(A)$, because A complement union A is the full sample space S. And the probability of full sample space S is one and probability of an A complement and A are mutually exclusive obviously. Therefore probability of A complement is one minus probability of A. Probability of null set is 0. If A is discrete then probability of A is the sum of probability of individual outcomes comprising A. So if it is a discrete and it has several outcomes in it then it is some of all the... it actually follows from this particular postulate. From this third postulate.

The third point here that if A is discrete then probability of A sum of probability of all individual outcomes is derived from the third postulate. If an experimental result in any one of the N equally likely outcomes then if event A comprises of n such individual outcomes then $P(A) = n/N$.

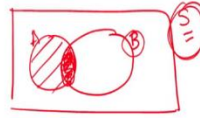
Probability of $P(A \cup B)$. Now we are not assuming that they are mutually exclusive then it is equal to $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. This you can see that $A \cap B$ actually occurs several times. So if you take your sample space S and this is A and this is B and this is your sample space A. Then it says that when you take a union of this A and B together this is $A \cup B$ then when you count the probability of $A \cup B$ it is a probability of A plus probability of B and you can see that probability of A intersection B gets counted twice so it has been removed once.

(Refer Slide Time: 13:29)

Conditional probability & Bayes Theorem

- Probability of occurrence of event A given that event B has already occurred:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

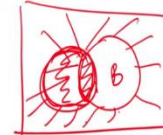


$$\therefore P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

- and for any events A and B:

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

$$\therefore P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$



Total probability Rule



Now we come to the next stage. So far we have been talking that you have a sample space that is draw a Venn diagram. There is a sample space S and within it there is event A. Okay. In real life number of times we do not know or we have very little information about the sample space S. But we are aware that actually there is an event B which has occurred also within the sample space S.

So what am I saying is that you we do not have the knowledge of sample space S as a whole but within the sample space we know that an event of B has occurred, the event B has occurred and then we want to learn something about probability of A, then of course we would not know the whole probability of A. Probability of A is this whole area. But we are aware of only that event B has occurred.

We do not know how does the whole event space S looks like, and therefore all the knowledge of A comes from this particular intersection and that I will make it dark. This intersection is the area of which we have an idea because we know B and we know that A has occurred. It means that we are aware of only A intersection B. When this happens if you want to find out the probability this kind of a probability is called a conditional probability.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

But we are interested in finding out what is happening to event A when the event B has occurred. So this situation can come up. Let us see what kind of an example could that be. Shall we say that

we are looking into the big set of certain kind of alloys. But we know that only certain allowing elements is what we are interested in and then we want to learn certain property. So this property A is talking about property of the alloy when all elements are present.

But we do not have any idea. We only know that when these some elements are present which is given by B. We know the property of A and then what is the probability of that property occurring, we cannot talk in terms of the whole sample space S. We have to talk in terms of only sample space B. This is called conditional probability and this is a simple simplification of it.

If you take this denominator to the other side you get that probability of $(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$. This is simple mathematics. Any event of A can be described as B intersection A and the B compliment is this area.

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

Sometimes this is called Total Probability rule. This is not used commonly but it is better to know that sometimes this is called total probability rule.

Now from this we come to a very important theorem which is known as Bayes theorem. It is very interesting that Bayes Theorem was discovered by Thomas Bayes. Sometime in 18th or 17th century I think it was 18th century but I might be mistaken it might be 17th century and today in the 21st century we find in extreme use of it. And it has become a very important theorem. When we were in the college and we were studying this Bayes theorem we had to create artificial problems to understand it.

But today it is the applications is available in abundance and it is in fact the most commonly used analysis which is known as Bayesian analysis will touch upon it in this course also. But this what is this Bayes theorem?

(Refer Slide Time: 20:37)

Conditional probability & Bayes Theorem

- Two events A and B are independent if any one of the following is true

$$P(A|B) = P(A) \checkmark$$

$$P(B|A) = P(B) \checkmark$$

$$P(A \cap B) = P(A)P(B) \checkmark$$

- Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(\hat{A}_k|B) = \frac{P(B|\hat{A}_k)P(\hat{A}_k)}{P(B)} \quad k=1,2,\dots$$

*B - data → reality
A - parameters estimates*



So consider that two events A and B and they are independent. Only if some of the conditions are true.

$$P(A|B) = P(A)$$

$$P(B|A) = P(B)$$

$$P(A \cap B) = P(A)P(B)$$

Any of these three conditions is met then we say that events A and B are independent events. Now we come to Bayes theorem. You recall the previous slide. Let us recall. In the previous slide we have that probability of A intersection B is equal to probability of A given B multiplied by probability of B which is same as probability of B given A multiplied by probability of A.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Looks very simple. But this gives a very important information. Let us assume or let us imagine that you have to you have certain data B.

B is your data which is given to you and A are certain parameters that describes the data. Now these parameters are called estimates. Data is the reality, parameters we obtained by what is known as estimates. We learn about this in more details in future. But then these are the parameters and these are its estimate. Data is the reality. Data is the reality because you have all observed it. So now the data keeps coming to you. If it is a real time data it keeps coming to you. It gets augmented.

So your reality becomes larger and larger. So how do you improve upon your estimates based on this larger getting reality that is what it says, A is a parameter if you estimate using B then it is same as the probability of data given the parameters in terms of probability of parameters and probability of data. This becomes an iterative treatment which can be written as probability of one iteration.

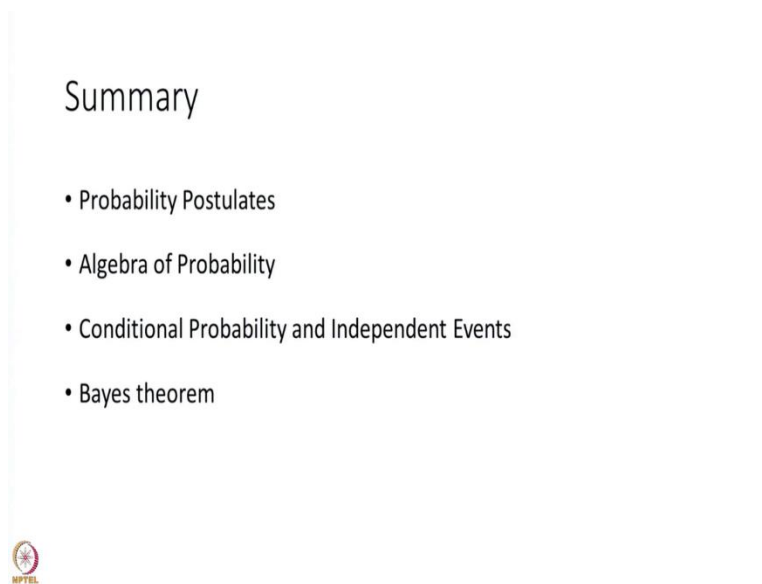
The next iteration of parameter estimation comes out from the same data given the previous estimator probability of the previous estimator divided by probability of data. Please remember that this K runs from 1, 2 onwards. So you have some initial idea. Say you are working with a data on rainfall and you have an initial idea that generally the rainfall in this area is so many percentage or it the average rainfall is so many millimeters, then you get the latest data.

From the latest data that is using the previous estimator you come up with the latest data to give you the estimator for the new data that is from the base data you can come up with the new data and this is called Bayes theorem so it means that it actually gives you an opportunity to improve upon your estimator. We will go into details when we go up on the Bayes theorem how it has gained importance today and why it was not there before.

Bayes theorem application or Bayesian analysis is extremely data intensive. You have to have lot of data in order to have application of Bayes theorem. With today's world of Facebook, Google mobile phones, etc. The real time data is increasing very fast. Even in the industry with the robotics playing its role, artificial intelligence playing its role we generate a lot of data on any processing of metallurgical industry and when this data is generated the question comes that whatever parameters which we were estimating using this data can we improve upon it.

And that improvement comes from the Bayes theorem. And therefore improves life Bayes theorem has become very important. As I said I repeat again. We will talk about Bayesian analysis in this course at a little end of the course. So with this we let us summarize what we learned today.

(Refer Slide Time: 26:56)



We first introduced probability postulates based on the fact that every experiment has is run by chance. It cannot give you the same answer all the time. There is an error factor and there is a chance factor playing with it in order to control that we must have some idea of probability and therefore we introduce probability through certain definitions like event and sample space and then we introduce probability postulates.

We did a little bit of Algebra of probability that is you know what is the probability of the null set, probability of sample space, probability of A compliment is 1 minus probability of A is what I call algebra of probability then we introduced the situation in which you may not know everything about the sample space but you may know about a particular event and you look into the whole other events occurrence in the light of this event which has already occurred which is called a conditional probability and based on conditional probability we define what are known as independent events.

It means that whether that event occurred or not has no effect on the occurrence of the other event. So if A and B are independent events then occurrence of event B has nothing to do with the

occurrence event A and finally be introduced quickly a Bayes theorem. Which as I said in today's world has gained a lot of importance and we will talk about Bayesian analysis at the end of this course. Thank you.