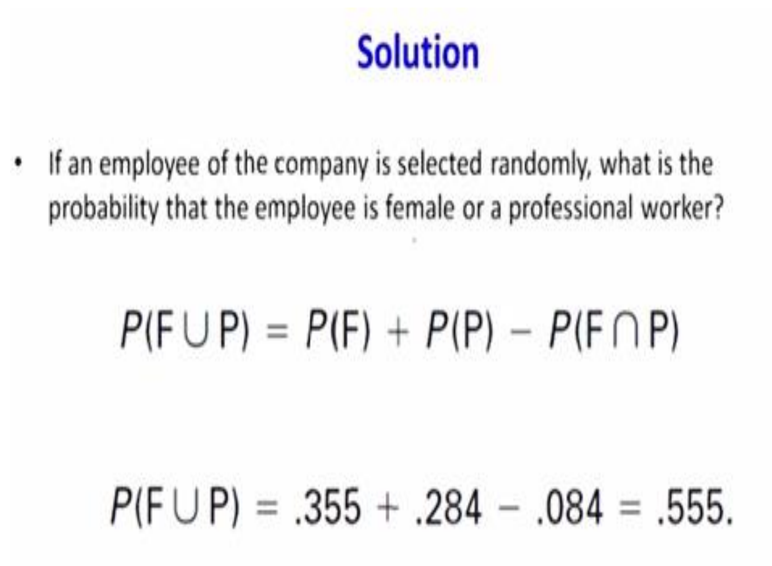
Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 07
Introduction to Probability - II

Dear Students, we will continue that the concept of probability in lecture number 7 will take one example then we will try to understand the concept of marginal probability, joint probability, and conditional probability in this problem. The problem is a company data reveal that 155 employees worked in a 4 types of positions. Shown here again is the raw values matrix also called the contingency table with the frequency count for each category.

And subtotals and totals containing a breakdown of these employees by type of position and the sex. Look at this contingency table in the row it is given what kind of position they are holding whether managerial professional technical clerical,

(Refer Slide Time: 01:20)



In column what is their sex, suppose when an employee of the company is selected randomly, what is the probability that the employee is female, in the contingency table row represents the type of position, column represents sex, the type of position they may hold whether they can have the they can work as a managerial position, professional position, technical clerical, we asked their sex also in the datasets, the question is if an employee of the company is selected randomly.

What is the probability that the employee is female or professional worker that is what is the P(F U P)? F is the female, P is the professional. So, as per the law of addition of probability

P(F U P) = P (F) + P (P) - P (F ∩ P),

P(F) we can find out and going to the previous slide.

(Refer Slide Time: 02:22)

Contingency Table

COMPANY HUMAN RESOURCE DATA

		Sex		
		Male	Female	
Type of Position	Managerial	8	3	11
	Professional	31	13	44
	Technical	52	17	69
	Clerical	9	22	31
		100	55	155

So, the P (F) is there are 55 females total there are 155 So, when = 55/155, you will get 0.355 then P(P), going to previous slide P(P), is probability of professionals, there are 44 professionals, 44 /155 that will give you 0.284 then, minus P (F ∩ P), that is for female related same time their working in a professional type of position that is 13. So, 13 /155 will give you 0.084

P(F U P) = 0.555.

(Refer Slide Time: 03:10)

Problem

- Shown here are the raw values matrix and corresponding probability matrix for the results of a national survey of 200 executives who were asked to identify the geographic locale of their company and their company's industry type.
- The executives were only allowed to select one locale and one industry type.

Shown here are the raw value matrix and corresponding probability matrix of the result of a national survey of 200 executives who are asked to identify the geographical location of their company and their company's industry type. So, there asking 2 question what is their company's geographic location and what kind of industry they are working. The executives were only allowed to select one location and one industry because they can work only one industry in one location.

(Refer Slide Time: 03:45)

RAW VALUES MATRIX

		Geographic Location				
		Northeast D	Southeast E	Midwest F	West G	
Industry Type	Finance A	24	10	8	14	56
	Manufacturing B	30	6	22	12	70
	Communications C	28	18	12	16	74
		82	34	42	42	200

This table shows 0 there is a industry type maybe finance manufacturing communications, calling it ABC. The geographic location may be Northeast, Southeast, Midwest and West, So for example, in the finance A, there are 56 people are working manufacturing B there are 70 people are working, in the communications 74 in the Northeast location, there are 82 people, in Southeast location 34 people and Midwest location 42 and West 42.

(Refer Slide Time: 04:21)

Questions

a. What is the probability that the respondent is from the Midwest (F)?

b. What is the probability that the respondent is from the communications industry (C) or from the Northeast (D)?

c. What is the probability that the respondent is from the Southeast (E) or from the finance industry (A)?

The question is, what is the probability that the respondent is from the Midwest? Directly we can read this answer from the table, second question is what is the probability that the respondent is from the communications industry or from the northeast? So here the addition of the 2 probability that is a P(C) and P(D). The third question is what is the probability that the respondent is from the Southeast or from the finance industry?

(Refer Slide Time: 04:55)

PROBABILITY MATRIX

		Geographic Location				
		Northeast D	Southeast E	Midwest F	West G	
Industry Type	Finance A	.12	.05	.04	.07	.28
	Manufacturing B	.15	.03	.11	.06	.35
	Communications C	.14	.09	.06	.08	.37
		.41	.17	.21	.21	1.00

So from the given table, we find out the probability that means we have divided each element in the cell divided by the gross total. After dividing that we got 0.12, 0.05, 0.04, 0.04 Now this is a matrix conditional probability. From this table, we can pick up whatever answer we wanted to get for answering that question

(Refer Slide Time: 05:20)

Mutually Exclusive Events

Type of Position	Gender: Male	Female	Total
Managerial	8	3	11
Professional	31	13	44
Technical	52	17	69
Clerical	9	22	31
Total	100	55	155

$$P(T \cup C) = P(T) + P(C)$$
$$= \frac{69}{155} + \frac{31}{155}$$
$$= .645$$

For example. I am going back, suppose what is the probability that the respondent is from the Midwest? So I am going to next slide Midwest is F. So, it is a 0.21 going back, what is the probability that the respondent is from the communication industry or Northeast? So you have to see P (C U D) = = P (C) + P (D) - P (C ∩ D) you can find out then what is the probability that respondent is from southeast or the same thing? P (E union A) = P (E) + P (A) - P (E ∩ A) These values you can directly pick up from the previous table from, this table now we will go for mutually exclusive event suppose okay suppose if you want to find out P (T union C) here T is the those who are technical professional P of T union C is P (T)+ P (C) because generally it will be minus P (T intersection C), but that is not possible, because a person cannot work in 2 industry at a time.

So, it is mutually exclusive event in the mutually exclusive event in the intersection will become 0 that is a P (T union C) = P(T) + P (C) that another term that is minus P (T ∩ C) will become 0. So, P(T) is a 69 /155, + P (C) is 31/155 when you simplify it is 0.645.This is example of mutually exclusive event.

(Refer Slide Time: 07:00)

Mutually Exclusive Events

Type of Position	Gender: Male	Female	Total
Managerial	8	3	11
Professional	31	13	44
Technical	52	17	69
Clerical	9	22	31
Total	100	55	155

$$P(P \cup C) = P(P) + P(C)$$
$$= \frac{44}{155} + \frac{31}{155}$$
$$= .484$$

There is another P(P U C) = P (P) + P (C) there would not be any intersections even that formula which I told the 2 slides before also that the intersection component will become 0 because the person cannot work in 2 industries, there is an example of mutually exclusive event.

(Refer Slide Time: 07:18)

Law of Multiplication

$$\boxed{P(X \cap Y) = P(X) \cdot P(Y|X) = P(Y) \cdot P(X|Y)}$$

In the law of multiplication P (X ∩ Y) = P (X) x P(Y\X) = P (Y) x P (X \ Y) is a law of multiplication. What will happen if event X and Y are independent event? Simply we can multiply P (X) into P (Y), that you will see later.

(Refer Slide Time: 07:42)

Problem

- A company has 140 employees, of which 30 are supervisors.
- Eighty of the employees are married, and 20% of the married employees are supervisors.
- If a company employee is randomly selected, what is the probability that the employee is married and is a supervisor?

You will see another problem. A company has a 140 employees of which 30 are supervisors 80 of the employees are married, and 20% of the married employees are supervisors. If a company employee is randomly selected, what is the probability that the employee is married and is a supervisor? Wherever this kind of problem comes, if you are able to construct the contingency table whatever question being asked you can pick up from there directly. So, from the given data.

(Refer Slide Time: 08:15)

		Married		
		Y	N	Sub total
Supervisor	Y	0.1143		30
	N			110
	Sub total	80	60	140

First we will construct a contingency table in the contingency table you see there are 140 employees out of which 30 are supervisors out of 140, 80 people are married and for example, those who are married at the same time supervisors, what do you do that is you have to multiply that how will you that if you multiply 80 into 0.2 divided by 140 you will get this answer.

(Refer Slide Time: 08:43)

$$P(M) = \frac{80}{140} = 0.5714$$

$$P(S|M) = 0.20$$

$$P(M \cap S) = P(M) \cdot P(S|M)$$

$$= (0.5714)(0.20) = 0.1143$$

So, for example, you see that how we are getting that we are the previously they will get 0.1143 we will see how we are getting so, we know the probability of married people 80 /140. This is given those who are supervisors at the same time married. 0.2 that is a 20% is given. If you want to know those who are married at the same time they are supervisors. So, we have to use conditional probability P (M) x P (S \ M). So, P (S\M) is known, 0.20 is given. So P(M∩S) = 0.1143.

(Refer Slide Time: 09:32)

Law of Multiplication

Probability Matrix of Employees

	Married		
Supervisor	Yes	No	Total
Yes	.1143	.1000	.2143
No	.4571	.3286	.7857
Total	.5714	.4286	1.00

$$P(\bar{S}) = 1 - P(S)$$
$$= 1 - 0.2143 = 0.7857$$
$$P(\bar{M} \cap \bar{S}) = P(\bar{S}) - P(M \cap \bar{S})$$
$$= 0.7857 - 0.4571 = 0.3286$$

$$P(M \cap \bar{S}) = P(M) - P(M \cap S)$$
$$= 0.5714 - 0.1143 = 0.4571$$
$$P(\bar{M} \cap S) = P(S) - P(M \cap S)$$
$$= 0.2143 - 0.1143 = 0.1000$$
$$P(\bar{M}) = 1 - P(M)$$
$$= 1 - 0.5714 = 0.4286$$

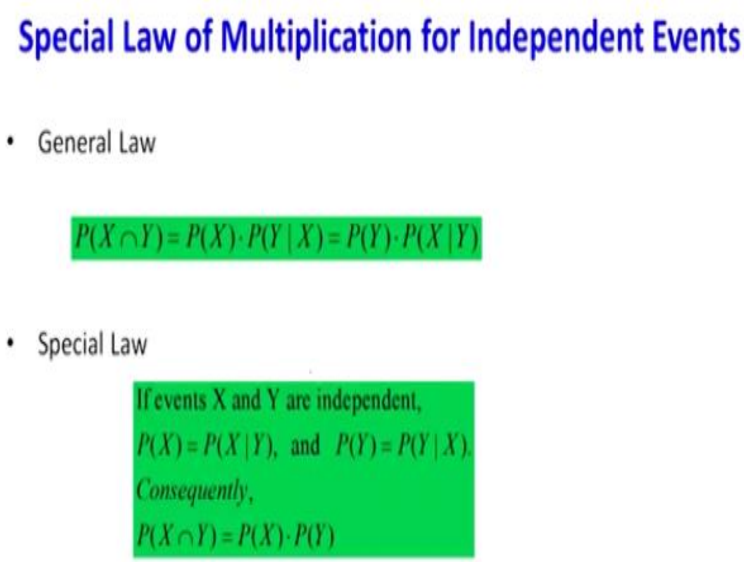
You see that once you know that one cell in the contingency table filling the remaining cell is so easy. And whatever value you want to pick up we can pick up for example, I have filled the first

0.1143 I know what is the row total and column total from that I can subtract it I can get the remaining rows that is a application of contingency table. Suppose P (S) = 1 – P(Sbar), that we know that P (S bar), that is 0.7857.

I am saying this one, this location, this location 0.7857 if you want to know P (M bar intersection S bar) that means, those who are not M those are not married. At the same time, who are not supervisors, so, that is nothing but this location 0.326 those are not married the same time they are not supervisor this locations, P(M intersection S bar) those are married, but not the supervisors.

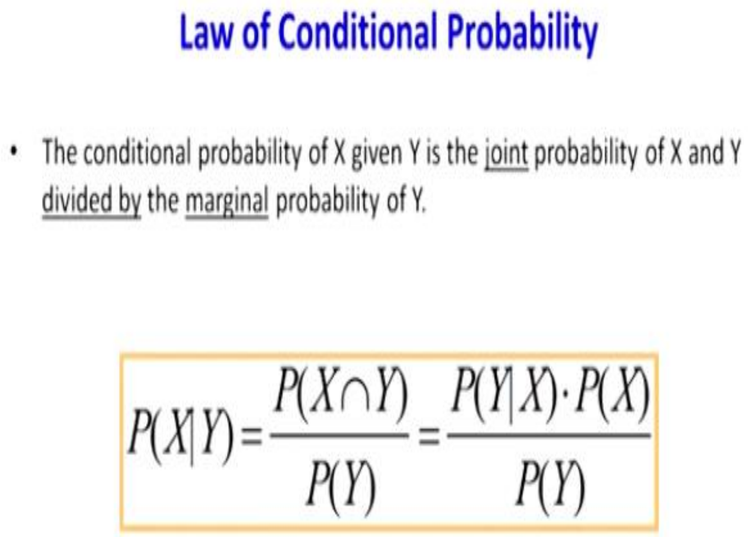
So, that is a P(M)- P (M∩S). So, this value is 0.5714 minus this one. So, we will get this point, nothing but in the contingency table if you know one cell the remaining this can be found out

(Refer Slide Time: 10:54)



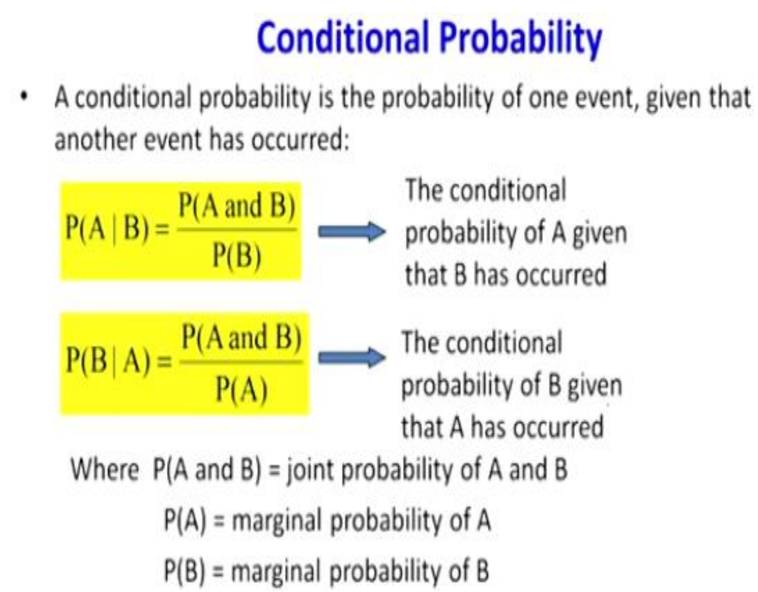
The special law multiplication for independent events general law is if P (X intersection Y)= P (X) x P(Y \ X) = P (Y) x P (X \ Y). special law, that is if X and Y are independent. So, P (X) = P (X \ Y), because, when the event X and Y are independent, the outcome of X is not depending on outcome of Y. So, P (X\ Y) will become P (X) itself. So, similarly, P (Y) and P(X) if not independent P(Y \ X) will become P(Y) itself and you substitute the there. So, P(X ∩ Y) will be P(X) into P (Y), only when the both even are independent.

(Refer Slide Time: 11:45)



This also law of conditional probability, this also we have seen previously also the conditional probability of X given Y is joint probability of X and Y divided by the marginal probability of Y So joint properties intersection the P (X intersection Y) / P (Y). So this can be just by readjusting (P(Y / X) x P (X))/ P (Y).

(Refer Slide Time: 12:14)



Little detailed explanation on conditional probability, A conditional probability is the probability of one event, given that another event has occurred suppose if I say P(A\ B). So, first you have to find the intersection of P (A∩ B) then divide by P (B). So the conditional probability of A given that B has occurred. This is an explanation for this. Suppose, if you want to know P of B given A has occurred, so P (A∩ B) /P (A) where P of A and B equal to join probability of A and B. So P (A) is marginal property of event A, P (B) is marginal property event B

(Refer Slide Time: 12:54)

Computing Conditional Probability

- Of the cars on a used car lot, 70% have air conditioning (AC) and 40% have a CD player (CD). 20% of the cars have both.
- What is the probability that a car has a CD player, given that it has AC ?
- We want to find P(CD | AC).

We will take an example how to find out the conditional probability of the cars on a used car lot, 70% have air conditioning Air Conditioning and 40% have CD player. 20% of the cars have both. So, what is the probability that a car has a CD player, given that it has AC that means, we want to find out P of CD given that AC is there.

(Refer Slide Time: 13:19)

Computing Conditional Probability

	CD	No CD	Total
AC	0.2	0.5	0.7
No AC	0.2	0.1	0.3
Total	0.4	0.6	1.0

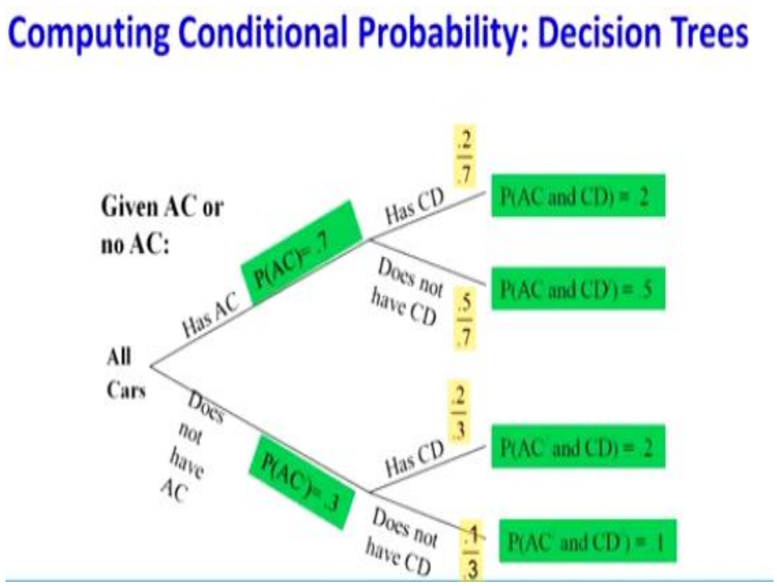
$$P(CD \mid AC) = \frac{P(CD \text{ and } AC)}{P(AC)} = \frac{.2}{.7} = .2857$$

Given AC, we only consider the top row (70% of the cars). Of these, 20% have a CD player. 20% of 70% is about 28.57%.

As I told you just to you draw the contingency table because all the values are given? So, what is the value I am going back see for example 70% of the cars having AC So, this value I am going back 40% of the cars having CD player So, this value and you subtract minus 1 will get that one another information is given 20% of the car have both like by see that 0.2 this value. So, once you know these values other cells can be find out.

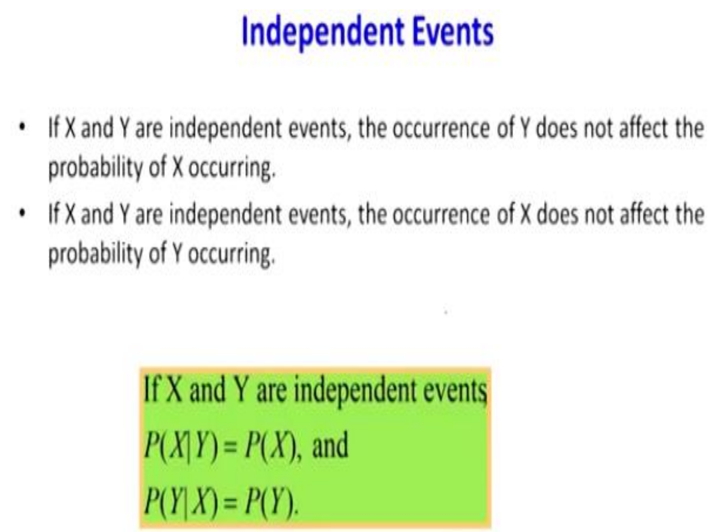
So, if you want to know P (CD\AC) as per the definition, P of CD and AC divided by P (AC) so, this is a 0.2 this is by P (AC) is by 0.7. So, this is 0.2857. So, given the AC we only consider the top row 70% of the cars of these and 20% CD player, so 20% of % is 28.57% okay there so, we are getting the conditional probability.

(Refer Slide Time: 14:25)



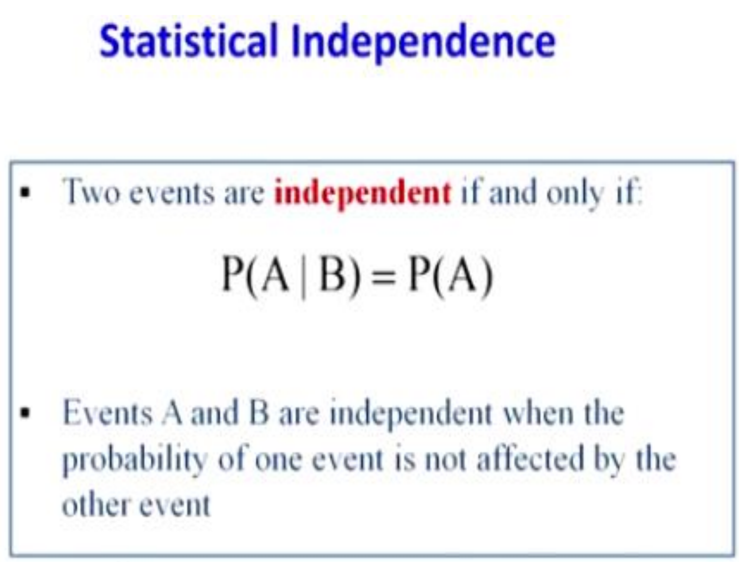
So, this conditional probability can be explained with the help of a tree diagram, because the tree diagram is easy to visualize. So, having AC, having not AC, having CD, having not CD having CD having not CD. So, 0.7 0.2 0.5 0.2 0.1 so, if you want to know having CD, so you have to divide 0.2 divided by 0.7. For example, if you want to know this this arc, so, this is a 0.5 divided by 0.7 and so on because a tree diagram is very easy to understand.

(Refer Slide Time: 15:12)



Then we will see the definition of independent event, if an X and Y are independent events, the occurrence of Y does not affect the probability of X occurring, so, X and Y are not connected. Similarly, if X and Y are independent events, the occurrence of Y does not affect the probability of X occurring, you see that P of if X and Y are independent P (X \ Y) = P(X), P(Y\ X) = P(Y) this we have seen the previous also.

(Refer Slide Time: 15:44)



This is another example 2 events are independent This is the condition P(A\ B) = P(A). So, this condition is for testing independent, even A and B are independent. When the probability of one event is not affected by other event.

(Refer Slide Time: 16:05)

Independent Events Demonstration

	Geographic Location				
	Northeast D	Southeast E	Midwest F	West G	
Finance A	.12	.05	.04	.07	.28
Manufacturing B	.15	.03	.11	.06	.35
Communications C	.14	.09	.06	.08	.37
	.41	.17	.21	.21	1.00

Test the matrix for the 200 executive responses to determine whether industry type is independent of geographic location.

So, we will take one example will check the practical application of this concept of independent events. This also this data previously given. So, we have asked what kind of industry are working, whether you will finance manufacturing communications that we asked to the geographical locations. Now, you see that the question is tested the matrix for the 200 executive respondents to determine whether the industry type is independent of geographical location that means, we were to find out is there any dependency between the geographical location and what kind of industry.

For example, in India, if you look at there, most of you know, software companies are in south. So, is there any connection between the geographical location and kind of industry which are located. We will take this Example finance and the best region, so, when you go this,

(Refer Slide Time: 17:05)

Independent Events Demonstration Contd...

$$P(A|G)=\frac{P(A\cap G)}{P(G)}=\frac{0.07}{0.21}=0.33 \quad P(A)=0.28$$

$$P(A|G)=0.33\neq P(A)=0.28$$

If you want to know P(A\G) = P(A∩G)/ (G), P(A∩G) we can directly read from the table 0.07 this one this value then P (G) 0.21 directly we can read from the table. So, what you do that value is 0.33, but P(A) when you look at P(A), so, P(A) is 0.28. So, now what is happening, the P(A\G) is not equal to P(A). If it is equal, both are independent, since it is not equal, there is a kind of dependency between the geographical location and the type of industry which are located there. So, this is a one way to test the independency.

(Refer Slide Time: 18:06)

Independent Events

	D	E	
A	8	12	20
B	20	30	50
C	6	9	15
	34	51	85

$$P(A|D)=\frac{8}{34}=.2353$$

$$P(A)=\frac{20}{85}=.2353$$

$$P(A|D)=P(A)=0.2353$$

For example, you take for another example, because A given D, this is another example. So, P(A∩ D) is 8 here the actual count is given. Any way you can you can do both the way also P(A∩ D) is 8 divided by P(D) , P (D) is 34. So, we are getting this value, but you see the P (A) is 20 divided by 85, 20 divided by 85. So, both P (A \ D) and P(A) are same. So, these are independent events. Then example, if at the same P (A / D) equal to P(A) both events are independent this is the way to test the independent.

(Refer Slide Time: 18:55)

Revision of Probabilities: Bayes' Rule

- An extension to the conditional law of probabilities
- Enables revision of original probabilities with new information

$$P(X_i|Y) = \frac{P(Y|X_i)P(X_i)}{P(Y|X_1)P(X_1) + P(Y|X_2)P(X_2) + \cdots P(Y|X_n)P(X_n)}$$

Next we are going to an important application that is a Bayes ruler Bayes theorem, it is used to revise the probabilities, it has lot of applications in higher level of probability theory. And extension to the conditional law of probabilities enables revision of original probability with the new information's. So, P (X \ Y) equal to P(Y \ X) x P (X) divided by the summation of this one I will tell you the net the next slides

(Refer Slide Time: 19:29)

$$P(x/y) = \frac{P(x \cap y)}{P(y)}$$

$$P(y/x) = \frac{P(x \cap y)}{P(x)}$$

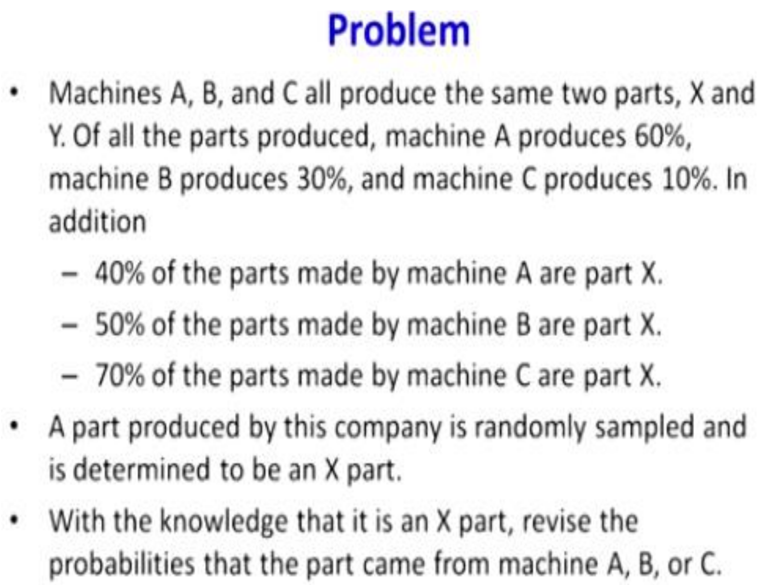
$$P(x/y)\,P(y) = P(y/x) \cdot P(x)$$

$$P(y/x) = \frac{P(x/y) \cdot P(y)}{P(x)}$$

For example, supposed see that P (X\Y) So, this can be written as P (X∩Y)/P(Y) this can be written as P of x intersection y divided by P of x you look at this here also P of x intersection y this also be x intersection y. So, this can be written as P (x\y) multiplied by P (y) equal to P (y\ x) multiplied by P(x), you see that if I look at this suppose, I know P (x\ y).

Suppose, if i want to know the reverse of this that is P of y by x, you see that I know P of x by y, I am getting reverse of that that is P of y by x. So, from this you can write it P of y by x is nothing but P of x by y multiplied by P of y divided by P of x. Here the P of x is only because here only 2 outcome there are if there are more outcomes here the sigma of P(x) will come, the sigma of P(x) is nothing but different combination of joint probabilities that we will see with the help of an example

(Refer Slide Time: 21:12)



This is a very typical example machine A and B, and C all produce the same 2 parts X and Y. Of all of the parts produced, machine A produces 60% machine B produces 30%, and machine C produces 10%. In addition 40% of the parts made by machine A are part X 50% of the parts made by machine B are part X 70% of the parts made by machine C is part X. A party produced by this company is randomly sampled and determined to be an X part with the knowledge that it is an X part. revise the probabilities that the part came from machine A, B, and C First, we will solve this with the help of a tabular format.

(Refer Slide Time: 22:01)

Event	Prior $P(E_i)$	Conditional $P(X \mid E_i)$	Joint $P(X \cap E_i)$	Posterior
A	.60	.40	$(.60)(.40) = .24$	$\frac{.24}{.46} = .52$
B	.30	.50	.15	$\frac{.15}{.46} = .33$
C	.10	.70	.07	$\frac{.07}{.46} = .15$
			$P(X) = .46$	

For example, there are 3 mission is there mission A and BC that 60% of that part was produced by machine A, 30% was produced by machine B, 10% is by C previously we have seen how that formula for conditional probability has come now, I will tell you an application of Bayes theorem.

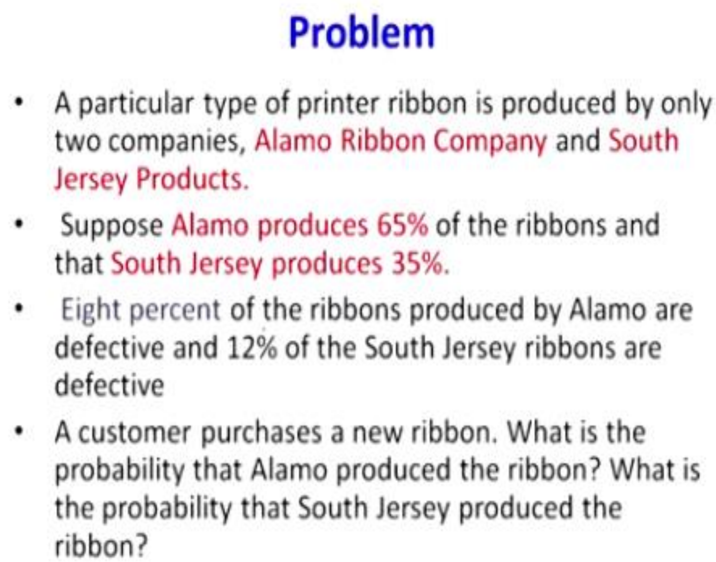
(Refer Slide Time: 22:29)

A	B
40%	60%
2%	3%

Suppose there are 2 say there are 2 supplier, supplier A supplier B, I know that, say the 40% of the product supplied by supplier A, remaining 60% supplied by supplier B from my past experience. I know that the 2% out of 40% is 2% will be defective product which are supplied by supplier A from supplier B, I know from my past experience he used to supply 3% of defective products out of 60.

By using their products that I have assembled a new machine now the machine is not working, the machine is not working, but I want to know what is the probability that product was supplied by supplier A, If the machine is not working, what is the probability that the product was supplied by supplier B? So this is the application of your Bayes theorem, we will see with the help of an example.

(Refer Slide Time: 23:36)



Different options there. The problem is a particular type of printer ribbon is produced by only 2 companies that company names is are Alamo Ribbon Company and South Jersey Products. Suppose Alamo produces 60% of the ribbons and South Jersey produces 35% of the ribbons from over experience. Look at this 8% of the ribbon produced by Alamo or defective and 12% of the South Jersey Ribbons are defective from our past experience, A customer purchases a new ribbon.

What is the probability that Alamo produced the ribbon? Otherwise, what is the probability that South Jersey produced the ribbon, like in the previous example, the machine is not working, what is the probability that product was supplied by supplier A what is the probability that product was supplied by supplier B the same example.

(Refer Slide Time: 24:39)

Revision of Probabilities with Bayes' Rule: Ribbon Problem

$$P(Alamo) = 0.65$$
$$P(SouthJersey) = 0.35$$
$$P(d|Alamo) = 0.08$$
$$P(d|SouthJersey) = 0.12$$

$$P(Alamo|d) = \frac{P(d|Alamo)\cdot P(Alamo)}{P(d|Alamo)\cdot P(Alamo) + P(d|SouthJersey)\cdot P(SouthJersey)}$$
$$= \frac{(0.08)(0.65)}{(0.08)(0.65)+(0.12)(0.35)} = 0.553$$

$$P(SouthJersey|d) = \frac{P(d|SouthJersey)\cdot P(SouthJersey)}{P(d|Alamo)\cdot P(Alamo) + P(d|SouthJersey)\cdot P(SouthJersey)}$$
$$= \frac{(0.12)(0.35)}{(0.08)(0.65)+(0.12)(0.35)} = 0.447$$

Now, first you will find out the marginal probability and conditional probability. So, P of Alamo that is 65% of the product was supplied by Alamo South Jersey 35% from the past experience I know the defective parts which was supplied by Alamo supplier 0.08. Similarly, the defective products which was supplied by South Jersey is 0.12, you see that this some would not be 100, but this some will be 100 because this is a total they supply in 8% is in the 65 out of 65, 8% is defective products are supplied by Alamo person.

Now, as for the formula now, we look at this V now it is reverse Now, we know it is defective. Then we want to find out what is the probability that was supplied by Alamo. So, we look at this P (D) given by Alamo multiple by P of Alamo, look at this this component, this is the sum of all possibilities P (D) given a Alamo multiplied by P (P Alamo)+ P (D given by South Jersey) multiplied by P (South jersey) So, this 0.08 is given 0.65 is given.

So, this was combination of 0.08 and 0.65 this and this was combination of 0.35 and 0.12. So, but because we have to add these two. So, when you divide by this is 0.553 that is, if the product the ribbon is defective, then there is a 50% chance it was supplied by supplier alone. Similarly, the product is defective what is the probability that it was supplied by South Jersey, the same thing, 0.12 to multiplied by 0.35 because P of D Alamo is given then 0.08 This is the all combination this denominator same, so, 0.447 that is there is of 44.7% chance that defective product was surprised by South Jersey.

(Refer Slide Time: 27:01)

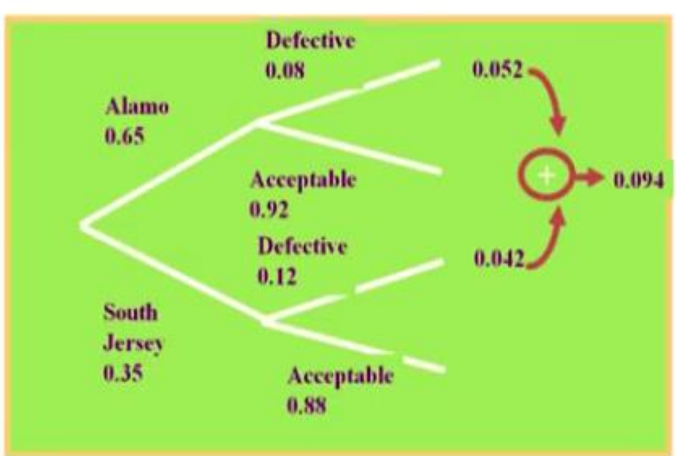
Revision of Probabilities with Bayes' Rule: Ribbon Problem

Event	Prior Probability $P(E_i)$	Conditional Probability $P(d\|E_i)$	Joint Probability $P(E_i \cap d)$	Revised Probability $P(E_i\|d)$
Alamo	0.65	0.08	0.052	$\frac{0.052}{0.094}$ =0.553
South Jersey	0.35	0.12	0.042 0.094	$\frac{0.042}{0.094}$ =0.447

If you look at in the in the tabular form it is very easy. So, this is the first. So, this is the event Alamo South Jersey this fellow supplying 65% this is 35%, this is the conditional probability, we know that this supplier Alamo will supply 8% of defective products, this fellow will supply 12% of 2 products. So, first we have to find out the joint property 0.0 to 0.08 this one we have to add it. Then this joint property has to be divided by this 0.04. So, that will give you use see that, here we know the details of P of D given the; we are finding the reverse of that P of E given by D that was the advantage of this byes theorem. Now, it is 0.094, this was 0.447.

(Refer Slide Time: 27:54)

Revision of Probabilities with Bayes' Rule: Ribbon Problem



This can be shown in the pictorial form. There is a Alamo South Jersey defective not defective. Defective 12% remaining this percentage when we multiply by 0.02 when you multiple this 0.042 when you added, we are getting 0.094. In this lecture, I have explained the example of mutually exclusive events then I have explained the multiplication, then explain the independent events then I have explained the concept of Bayes theorem, then I have explained with the help of a problem, the application of Bayes theorem with that will conclude this lecture. Thank you very much.