Welcome to the session on association. So,
this session's objective is to discuss association

between random variables. Now, if the random
variables are nice quantitative variables,

we can think of associations in various ways.
In those cases, we generally quantify the

association in terms of what is called us
coefficient financial world measures, so what

is called as covariance or co-movement, but
those are typically done for quantitative

variables, the interval variables, and so
on.

What if we are not talking about those interval
variables or ratio variables. We are talking

about categorical variables; how are we going
to measure the association between categorical

variables? Now, what are categorical variables?
If you recall variables are essentially categorical

or ratio or interval, remember there are four.
So, categorical variables themselves can be

classified as 2 when the categories can be
organized in a particular sequence and when

the categorical variables cannot be organized
in a particular sequence.

Gender, for example, is an example of a categorical
variable which cannot be arranged in a particular

order, whereas the ratings that we give, ratings
that a customer gives to a restaurant, for

example, the food in the restaurant was very
bad, bad, neutral, good very good. Now, this

is a data that is organized in categories,
but the categories can be ordered right. So,

you go to a restaurant, and we have this experience
very common.

At the end of the meal, the person hands over
a tablet nowadays where we have to record

our feedback about the service in the restaurant
about the quality of the food and so on and

so forth. So, there generally, the categories
are the food was very bad terrible it was

bad. I was neutral about the food; generally,
this is called something else; they polish

it and say something better the food was good
the food was very good. Sometimes, the category

names may be different. Somes they will say
they will only start with bad, average, very

good, and excellent.

So, so this is an example of a categorical
variable which can be ordered. So, this kind

of category this kind of variable is called
ordinal data ordinal data. So, what is the

variable here? The variable is food quality.
This is the variable. The variable is food

quality. And this variable food quality can
be categorized in four ways bad average very

good excellent, or the variable food quality
can take which values very bad, bad, neutral,

good, very good.

So, five possible values are four possible
values, but here there is a natural ordering

of the data. Therefore, this data is called
ordinal data. The categories can be ordered.

You cannot; you are not going to put very
good, bad, average, and excellent. So, you

are not going to put categories in a haphazard
manner. There is a natural ordering, whereas,

there is there are other categorical variables
where the data cannot be organized in a particular

sequence.

As I said, gender, or let us say that you
have a business in 18 different countries.

Now the country as a variable can take on
18 different values. I have offices in US,

France, Brazil, India, China, Australia, and
New Zealand. Now, you can say that I will

arrange the cities in alphabetical order.
Fine, you can arrange the cities in the order

in which you open your offices that is also
fine.

So, there is no natural ordering there. You
can decide the ordering, and any ordering

is okay, right. But usually, there is no ordering.
You cannot say New Zealand should always be

one and USA should always be three; you cannot
say that, and you can put any city after any

city, any city before any city. So, there
is no natural ordering there. So, those are

also categorical variables.

What we are going to discuss today is- association
between these kinds of categorical variables.

Now, in particular, we are going to split
the discussion in two parts. One is called

determining the association, and the other
one is called inferring the association. Now,

what do I mean? These are not very standard
terms. So, these, let me say that what do

I mean by determining an inferring?

Now, if a data is given to us some, sample
data is given to us, how do you determine

that the variables are associated? That is
a different question. However, we are very;

we are actually interested in a bigger question.

The data that we are going to determine association

about is going to be only a sample data; usually,
we want to infer about the population.

So, this problem is about the population.
So, once you ask, once you are determined

that there is some association, how can you
extend that to the population? How can you

generalize those insights to the entire population?
That problem is the inferring problem. So,

let us first understand how do we determine
or how do we quantify association between

the random variables from a sample data, and
then we extend that to the insights to the

population in terms of inferring about the
association.

So, let us take an example of determining
the association, and this will also determine

the association between categorical variables
will also help you recall the decision, the

discussions on conditional probability. So,
if you want to pause the video and do a refresher

on conditional probability that will be a
good idea does not matter; we are going to

discuss conditional probability in this section
also.

So, we are going to use conditional probability
to determine association between random variables.

So, let us proceed; as I said, let us proceed
with an example.

So, let us take an example from a business
school this is cooked-up data. The data that

I am going to show up is not real data. Let
me put a strong qualifier because this, the

example that I am taking, is slightly sensitive but let me tell you that this is purely cooked-up

data. So, let us say that particular B school. Let us say you shortlisted about 1200 candidates;

960 male candidates and 240 female candidates for our postgraduate management program; they

are shortlisted.

So, after that, there will be an interview and so on. And out of these 1200 candidates,

after the round of interviews and so on, 324 candidates were given the offer letter for

admission- 324 candidates were selected. Now some of these 324 may be male candidates;

some of them could be female candidates. So, let us understand the details of this data.

So, what is the data? This is data.

So, there were totally 1200 candidates who were shortlisted. Out of that, 960 were male

candidates, 240 were female candidates. Out of the 1200 candidates who were shortlisted,

they were interviewed, they were given some other tests, and after all these tests and

selection process criteria, 324 people were offered admissions okay. 324 candidates were

offered admissions, 876 candidates were not offered any admission to the postgraduate

management program.

So, once again, this is a toy example; this is a data that I created. This is not real

data. This has nothing to do with any B school. This is not real B school data. This is just

a point to drive my point across the concept across. So, generally, this kind of table

is called contingency table. So, now what happened was essentially, what happened out of these 324 candidates only 36 candidates were female candidates.

288 candidates were male candidates out of this 324, 288 were male candidates, and 36 were female candidates. So, looking at the data, the women's forum, for example, said that, yeah, this is a case of gender discrimination. So, you said that only thirty-six female candidates were offered admissions, and 288 male candidates were offered admissions.

The B school responded by saying that well, this happened purely, because there were more male candidates who were interviewed than female candidates. So, obviously, the number of selected candidates will have more male candidates than female candidates. So, yeah, as I said this kind of table is called a contingency table.

So, B school responded that there is no discrimination here. This purely happened because there were fewer female candidates who were shortlisted, and therefore, they appeared for the selection process. Now, how are we going to, as a data analysts, how are you going to analyze this data? Few things to be noted about this data; once again, this data is toy example data; this is not real data.

What else you note that this 1200 is not the population. This 1200 may be a sample from the entire list of candidates. So, you cannot generalize anything from this sample; this

is sample data; the entire population may
be much bigger, right. So, we may be looking

across the B school. We may be looking across
the ears. So, maybe the population size is

much bigger, or even for this management program,
the population size may be bigger.

We have data about 1200 candidates. Based
on the data about 1200 candidates, what are

we going to conclude, do we? Are we going
to say that the women's forum may have a point,

or are we going to say that B school's argument
may be correct? To do that as I said we are

going to look at this problem from the perspective
of conditional probability. Now, what are

the events here? What are the events here?
I can see four events.

So, let us define the event more generically.
What is the probability that a randomly selected

candidate is a male candidate? Let that event
be m. What is the event of randomly selecting

a female candidate? Let that event be w; women,
right. Now, let us say that there is an event

of an admission offer being made. Let that
event be called A. What is the event of randomly

selecting a candidate for whom an admission
offer was made.

And obviously, let us call this as N; this
is the event where a randomly selected candidate,

candidate may not be male or female. A randomly
selected candidate is not offered admission

to the program. So, these are the four events.
Let su, let them down. So, let m be the event

that a male candidate the candidate selected
is male; F be the event that a candidate is

a female. F, W you can, or you can use both.■
■
Let A be the event that the candidate is offered■
admission; the selected; the randomly selected■
■
candidate is offered admission. Let A compliment■
be the event that the admission, the candidate■
■
did not get any admission offer. Now, what■
is this AC? AC is called the complement of■
■
event A. Go back; there are only two possibilities;■
there are only two possibilities.■
■
So, either the person is offered admission,■
or the person is not offered admission. So,■
■
if this event is A, you need not call this■
event N; you can actually call this event■
■
A compliment because there are only two possible■
options either the candidate is offered admission■
■
either the candidate is not offered admission.■
So, that is why we do not waste up notation;■
■
we can simply use A compliment.■
■
We can simply use a compliment to denote the■
event that the candidate is not offered admission.■
■
Obviously, the probability of event A plus■
probability of the complement event, A complement,■
■
has to be equal to 1. This is from the axioms■
of probability. Now, what are the axioms of■
■
probability? Go back and review your session■
on the probability you must have studied axioms■
■
of before.■
■
So, these are the four events, right. Now,■
you look at the table; you look at the table.■
■
These cells, for example, 288, represent not■
a single event happening; 288 actually seems■
■
to be a combination of two events. What is■
the combination here? What is it? A combination■
■

of it is a combination that the candidate
is a male candidate and he is offered admission.

This 36 is a combination of two events. What
are the two events? The two events are; first

event is the candidate is a female candidate-
this is also a candidate who is offered admission.

So, simply defining these events may not be
enough; we actually need combinations and

other combination

So, what you are going to do is yes, so the
probability that are randomly selected candidate

is a male and offered admission. This is,
so as I said, this is a combination first

of all the candidate is a male candidate and
so this is a new event and is our for admission.

So, this is kind of a combination is called
joint event, and this probability is called

the joint probability okay joint probability.

Joint probability is the probability of the
candidate being male and offered admission.

If you recall and was this notation and was
this notation and okay, and as I have seen

as we have seen earlier how do you read the
data about it? How do you read this combination

data? This combination data is from our contingency
table this 288. So, what is the probability

of this; these two events are happening together-
the candidate being a male candidate and being

offered admission.

What is the probability of this happening
from the data? It is going to be 288 divided

by 1200. Similarly, the joint event of a candidate
being a female candidate and the candidate

being offered admission, this is the frequency-36. There are 36 candidates who are female

candidates and have been offered admission.
So, what is the probability of these two events

happening together that will be 36 divided
by 1200.

Similarly, you can define 672 divided by 1200
to be a joint probability. You can define

204 divided by 1200 to be a joint probability.
So, those are the four joint probabilities

okay those are the four joint probabilities.
Similarly, as I said, the joint probability

that a randomly selected candidate is a male
candidate but and he is not offered admission.

So, this is a combination of two events candidate
is a male candidate, which is this event m

and is always this notation and notation.

And the candidate is not of our admission
just like you use the notation A for the event

of the admission of our being made you are
going to use the notation A compliment for

A event that admission offer was not made.
And this combination has a probability of

672 divided by 1200, which is 0.56. As I said
for the other two, obviously, they should

be F, or they should be F from the previous
slide they should be F.

I can change that F, W. So, 0.03 is the joint
probability that the candidate is a female

candidate, and she is offered admission that
is probability 36 divided by 1200 the probability

that a female candidate; the candidate is
female, and she is not offered admission that

probabilities 204 divided by 1200, 0.17 this

is from the data. So, these are called joint

probabilities these are called joint probability.

Going back to the table, these are called
joint probabilities. So, these four numbers;

these four numbers 288 divided by 1200 36
divided by 1200, 672 divided by 1200, 204

divided by 1200; These are called joint probability
joint probability. Now let us read from the

column. Now for a male candidate, there are
only two possibilities either the admission

is made or the admission is not made. So,
288 male candidates are offered admission

672 male candidates are not offered admission.

Totally there will be there totally there
were 960 male candidates who were shot listed.

Now, what do you call this 960? If I if this
960 the frequency of a randomly selected candidate

being a male candidate. So, then let us go
further, let us go further and calculate this

probability as 0.8.

What is this 0.8? This 0.8 is called marginal
problem marginal probability. So, the marginal

probability is this 0.8 was obtained as 960
divided by 1200. What is this probability

of this is the probability that a randomly
selected candidate is a male. Similarly, the

probability that the randomly selected candidate
is a female candidate that probability will

be what will be that probability it will be
240 divided by 1200 and from the table, you

know that this is 0.2.

Similarly, what is this 0.27? 0.27 is the
probability that a randomly selected candidate

has been offered admission. So, this is the probability that event A has happened. This

is the probability that event A has happened. The admission offer has been made. How many

people were offered admissions? How many people were offered admission? Let us go back to

our contingency table 324. So, what is the probability that randomly selected candidates

would be offered admission?

The candidate may be male, female right now. We are not interested in that. And a randomly

selected candidate will be offered admission with a probability 324 divided by 1200 there

are 1200 people who are shortlisted, out of which 324 candidates have been offered admission.

So, what is the probability that a randomly selected candidate will be offered admission

that will be 0.27.

The remaining people will not be offered admission remaining people will not be offered admission.

So, the probability that event a compliment has happened, what is the event A compliment?

Event A compliment is the event that candidate randomly selected candidate is not of admission.

So, that is 876 times this has happened in the data.

Out of the 1200 data, out of the 1200 candidates, 876 candidates were not offered any admission,

and therefore, this will be 876 divided by 1200 times out that this is 0.76. Now, these

values 0.8, 0.2, 0.27, 0.73. They are called marginal probabilities. Easy way to remember

this is to say that they are appearing in

the margins; they are row totals and column

totals divided by the total 1200.

So, they are called marginal probabilities.
So, now we have described two kinds of probabilities

from our contingency table. These 0.8, 0.3,
0.17, they were called joint probabilities,

and these 0.8, 0.2, 0.27, 0.73 are going to
be called as marginal probability. The marginal

probability of this event happens or this
event happening or this event happening or

this event happening those are corresponding
numbers. I hope you are with me; otherwise,

you can pause and go back again.

But this is a discussion on probability; where
are we on answering the question, where are

we on answering the question. This was the
question that we need to answer. Is there

enough support for some bias happening in
the admission process? Now for that, what

we will do is we will calculate this probability.
Now, what is this probability? What is this

notation for? This is a notation for conditional
probability.

Remember, this is different from joint probability.
Joint probability was denoted by this notation

where we use and write M and A, F and A, W
and A. So, this is the notation for joint

probability, and notice the difference this
is a different notation. So, this is a notation

for what is called as conditional probability.
So, what is this probability? This is the

probability; this is a probability so, there
is a vertical line here; how do you read this

notation?

This is you read this notation as the probability of event A happening given the fact that event

M has already taken place. Let me repeat that this is the probability that event A. This

is the probability of event A happening, given the fact that event M has already taken place.

This is different from once again; let me tell you this is different from joint probability.

What is joint probability? Joint probability of event A happening and event M happening

at the same time.

Conditional probability says event M has already taken place, okay; you already know that the

candidate is a male candidate. Now, you are trying to find the probability of somebody

being offered admission when the candidate is a male candidate. The event M has already

taken place. So, how do you calculate that probability? Now, you know that you want to

understand only the admission status of the male candidates.

So, how many male candidates? 960 male candidates out of that 288 were offered admissions. So,

if you want to understand what is the probability of admissions when the candidate is a male

candidate that can b calculated by doing this 288 by 960. So, what are we saying everything

hinges on the contingency table. So, we go back to contingency table. So, we are saying

the candidate is already a male candidate.

So, we are not worried about the rest of the table. I can erase this now. Let us all; this

is only for explanation. So, let me raise
click keep the contingency table clean. Now

let me say whatever was there on the contingency.
Now we are saying that we are interested only

in this column. Why are we interested only
in this column because we are saying event

M has already taken place.

We already know that the candidate we are
talking about is a male candidate. So, candidate

can only be 1 out of 960. So, we are not talking
about the entire sample size of 1200. We are

now restricting ourselves to only one of these
960 candidates. Now how many of these 960

candidates are offered admissions? 288. So,
what is the probability that somebody will

be offered admissions given the fact that
they are a male candidate?

The very fact that they are a male candidate,
I am going to look at these 960, and then

what is the probability of somebody being
offered admissions out of this 960? 288, and

therefore, I am going to calculate that probability
of somebody being offered admissions given

the fact that the candidate is a male candidate;
that can be calculated simply by looking at

this column only 288 divided by 960.

So, this is called conditional probability.
You can do the same thing for female candidates.

Let me ask you this question what is going
to be the probability of admissions given

that the candidate is a female candidate.
Now, how many female candidates? 240 female

candidates. So, you are going to restrict
your discussion only to these 240 candidates.

Out of these 1200 and 36 for admissions, therefore, this conditional probability is going to be

calculated as 36 divided by 240.

So, this is a conditional probability which is different than the joint probability. Now,

how you can also get this 0.3 in different ways. How do you get this 0.3 in a different

way? You can get 0.3 as 288 divided by 960 okay 288 divided by 960 you will get your

0.3. You will also observe that this 0.3 can be obtained in a different way. This 0.3 can

be obtained as so you divide you are trying to get 288 divided by 960.

You divide both a numerator and divide denominator by 1200. So, you divide the numerator and

divide the denominator by 1200; you will get 0.24 and 0.8 in the numerator, and you will

get back 0.3. Where is this point to 0.4 and 0.8 in our table

that is in this table? So, how are you reading? How are you getting your 0.3? You are getting

your 0.3 as 0.24 divided by 0.8. That is what you said, right; that is what we said.

So, 0.3 is 0.24 divided by 0.8; however, what is this 0.24? What does this 0.24 represent?

You already know what this 0.24 represent. This 0.24 represents joint probability. It

is a joint probability. So, what is this? This is a joint probability. What is this

0.8? 0.8 is a marginal probability. What is this 0.3? We said this 0.3 is a conditional

probability, and that is how you get your method for calculating conditional probability.

Conditional probability is always calculated
as the ratio of joint probability divided

by marginal probability. Say that again; conditional
probability is always joint probability divided

by marginal probability. Now you can do the
same thing here. So, what will be this? What

will be 0.03 divided by 0.2? What is this
0.3 is the marginal probability. So, this

is sorry 0.03 is a joint probability divided
by 0.2, 0.2 is a marginal probability.

And obviously, 0.03 divided by 0.2 is going
to be a conditional probability. Let us go

one step further. This 0.03 this 0.24 comes
from what joint probability what is the joint

probability of? It is the joint probability
of the candidate being male and admissions

being offered. Remember, this was a notation
that we used. This was 0.24, two events happening

together, the candidate being male and the
offer of admission being made.

So, 0.24, so this is the joint probability
of M and A. What is this conditional probability?

We saw that this is the conditional probability
of A given M. So, this is the conditional

probability of A given M and what is this
08? This 0.8, it is a marginal probability

that the candidates is a male candidate sorry
is a male candidate. So, now you know the

equation.

This is the conditional probability that you
are trying to calculate. This conditional

probability, probability of A given M is probability
of A and M or M and A it does not matter.

How you say that because both events are happening

together? Either you say M and A or A and■
■
M divided by probability of M. Once again■
probability of A given M is probability of■
■
A and M divided by probability of M what is■
this? What is this probability? This probability■
■
will be probability of A given F.■
■
What is the probability that given the fact■
that the candidate is female what is the probability■
■
of event A happening? What is the probability■
of event A given that event F has already■
■
happened?. So, what is this? This is a joint■
probability admissions being offered and the■
■
candidate being a female candidate. What is■
this 0.2? 0.2 is the marginal probability■
■
that the candidate is a female candidate.■
■
So, that is how you calculate conditional■
probability. Conditional probability is calculated■
■
as the ratio of joint probability divided■
by ratio of joint probability and marginal■
■
probability. So, conditional probability is■
a ratio of joint probability and marginal■
■
probability. Similarly, you can calculate■
the other conditional probability which have■
■
already done.■
■
So, this is what we have said already; said■
this right. So, this is the conditional probability.■
■
Similarly, you can calculate the conditional■
probability of admissions offered being made■
■
given the fact that a female the candidate■
is a woman candidate that, as we said, is■
■
going to be 0.15. Now, what do you do to answer■
the question? The question of discrimination;■
■
how do you answer that?. Now you compare only■

the conditional probability.■
■
Here probability is 03, and here the probability■
is 0.15. What are these two probabilities?■
■
This? What is this 0.15? 0.15 probability■
of this is a probability that an admission■
■
offer will be made given the fact that of■
candidate is a woman candidate; that probability■
■
is 15%. However, what is this probability?■
The earlier joint probability 0.3. what is■
■
this? This is the probability that an admission■
offer will be made given the fact that the■
■
candidate is a male candidate; that probability■
is 30%.■
■
So, what is what do you conclude? Rather how■
do you conclude? You say that probability■
■
that the probability of admission offer given■
that the candidate is male is 0.3, which is■
■
twice of 0.15, which is the probability of■
admissions given the can given that the fact■
■
given the fact that the candidate is a woman■
candidate.■
■
So, once again, we are always going to say■
that conditional probability does not prove■
■
discrimination. It does not prove, but there■
may be some support to the argument there■
■
may be some support. We may have to do further■
analysis, but there may be some support to■
■
the argument, but we are very clear that conditional■
probability does not prove discrimination.■
■
As I said earlier, this was a data from a■
sample of 1200 candidates maybe if we collect■
■
more data we can be a little more sure. Therefore■
we are never going to argue that we have proved■
■
discrimination. But there is some support■

to the argument, and we are going to leave
it at that. So, this is how you use conditional probability. Now let us talk about the broader implication here.

What did we let us go back to the first table? Once again, let me remove all this; let me remove all this scribbling. So, that we can talk clearly, so here we essentially had two variables; what were the two variables? There was one variable about gender; there was one variable which was about gender. So, this was gender, and the other was admission status.

We can argue that these are categorical variables where category cannot be ordered.

How do you put male first and female; you could have put female as a first column, and the male second column does not matter which way you put right. So, this is categorical data where the data can be put only into the data has been put only in two categories.

So, we have used two values for the variable called gender. We have put two values for the variable called gender male and female.

We have put two values for the variable called admission status. Admission status is offered or not offered. These are the values that this category variable can take. So, essentially whatever we have done is towards determining any association between these two categorical variables. Gender as the categorical variable admission status as a categorical variable and we have our conclusion was that there seems to be some association between the two

random variables some association.

So, once again, what have we achieved? We have been able to determine the association

between categorical variable gender and categorical variable admission status. For the categorical

variable admission status, this variable takes on two values one was A, and one was A C those

were called A and A C. Offers being made can be one value not offered being the other value.

Gender was put up in two categories male category and the female category.

So and then we used conditional probability to determine the association between these

two categorical variables.

Hope, this is clear, and this is the very simple explanation of categorical variables

and application of that in determining the association between categorical variables;

sorry, this was a small example of conditional probability and conditional probability being

used to determine the association between categorical variables.

Actually what we have used this equation what equation did we use?

We introduce a new term called conditional probability and then we said conditional probability

is the ratio of joint probability and marginal probability..

Now this is actually called Baye's rule.

So, a Baye's rule and let us.

So, essentially what does Baye's rule talk about Baye's rule has a more generic explanation.

So, generally what happens in reality is we

have some initial guesses about events.

Initial guesses are our they are called prior
probability they are called prior probabilities.

So, using our usual concept of probability
we can translate these initial guesses into

what are called prior probability.

Those are our initial beliefs about some things.

Those initial beliefs are there to let us
keep them aside.

What do we do next?

We go and get more information.

This information may be in terms of samples,
maybe a feel test.

In general we get more information about these
events about which we had some initial guesses.

Now because in light of this data collection
in light of the sample that we have collected

in light of the field test that we have conducted,
what we may want to do is to update our initial

guess.

We may want to update our initial belief about
the events.

The updated belief is called the posterior
probability.

The posterior probability is the new probability
that we calculate for the same event but this

posterior probability considers all these
things that we have done.

We may have collected data, we may have collected
a sample, we may have done some field tests

in general, and we have some additional information.

So, what we have essentially done is that

there was a prior probability which may be■
■
because of our initial belief which is our■
initial belief and then we update our prior■
■
belief and calculate revised probabilities.■
■
Those revised probabilities are called posterior■
probabilities.■
■
So, what does Baye's rule do then?■
■
Baye's rule is essentially used to calculate■
posterior property if we have some initial■
■
belief if we have some initial property and■
we have some additional sample information.■
■
Let us recall the example that a previous■
example was about B school admissions and■
■
two types of genders male and female being■
considered.■
■
Now without any data you may say that the■
probability of admission to a particular B■
■
school may have some probability.■
■
Let us say that probability is 20 percent,■
25 percent.■
■
Now we collect data and we may have data and■
then we say that given that the candidate■
■
is a male candidate what is the probability■
that he gets admitted.■
■
Now we are talking about updated beliefs.■
■
We are still talking about the probability■
of getting admitted.■
■
We are still talking about the probability■
of getting admitted but now we have additional■
■
information.■
■
We have information that this candidate that■
we are talking about is a male candidate.■
■

Now we are saying how do you update the probability
of getting admitted?

We say that we have calculated posterior probability
because we have additional sample information.

And therefore we use conditional probability
to update our belief.

Let us take an example of this.

Let us take an example of this.

So, let us say that there are two manufacturers.

There is a manufacturer who haS2 suppliers,
two different suppliers.

Usually this is very common in the manufacturing
industry.

Usually you do not want to rely on one supplier
because if there is some disruption at one

supplier end you should not get affected because
of it.

To mitigate that risk to mitigate the supply
risk you want to spread and therefore you

want to use multiple suppliers that is a very
typical strategy commonly used in manufacturing

industries.

So, let us say that for our particular raw
material the same raw material both the suppliers

essentially supply the same raw material.

So, what has the manufacturer decided?

Manufacturer has decided that 65 percent of
the raw material 65 percent of the requirement

for the raw material will come from supplier
one S1 and the remaining 35 percent of the

raw material will come from supplier S2 okay
supplier S2.

So, there is some arrangement where the manufacturer
has told supplierS1 to supply 65 percent of

the manufacturer's requirement of the raw
material.

And the manufacturer has told S2 suppliers
to supply 35 percent of his requirement of

the raw material.

So, that is how the supply is happening.

Now you know that suppliers are not going
to supply perfect quality products.

So, let us say from historical data you know
that supplier S1 has 98 percent of the supplied

raw material in good quality.

So, there is a history that supplier S1 supplies
good quality products 98 percent of the time

and supplier S2 provides raw material of good
quality 95 percent of the time or 95 percent

of the raw material sent by supplier as to
of good quality.

This is also very common.

Nobody is going to guarantee you 100 percent
good quality raw material because of process

variations and various other factors.

There may be some rejection.

There may be some products which are not of
acceptable quality and this also happens very

typically in the manufacturing industry.

I mean you want to go, you want to get closer
and closer to 100 percent.

But you will realize that as you go closer
to 100 percent the cost actually goes up and

you may not be able to afford that cost and

therefore you typically say that 95 percent

is good enough acceptable quality for me knowing
fully well that 5 percent of the products

I may have to throw away or 5 percent of the
raw material I may have to rework before I

start using them.

This is a scenario.

Indirectly if you understand we have provided
data we have provided data 65 percent and

35 percent, 98 percent and 95 percent if you
compare this data to the contingency table

of your B school admission example you realize
that instead of providing numbers there we

said that 324 candidates were offered admissions
and so on.

Here the data is being provided in percentages
, that is the only difference.

But just like that table that data helped
us build a contingency table this kind of

data is also going to help us build a contingency
table.

This 65 percent, 35 percent, 98 percent ,95
percent should help us build our contingency

table.

But now we are going to see how to calculate
conditional probabilities without going to

those contingency tables.

I actually recommend that contingency tables
make the job a little easier.

So, I am going to ask you to build a contingency
table using this data that is currently on

the slide.

Build it, build your contingency table, it
will help you understand the data better but

we are going to look at the problem directly
. So, let us if this data is okay let me provide

the scenario.

What is the scenario?.

So, this is the scenario.

So, this is essentially what we have provided.

What does this 98 percent and 95 percent mean?

This 98 percent and 95 percent are essentially
conditional probability given the fact that

the country's supply is S1.

Given the fact that the supply is S1 what
is the probability that the raw material is

of good quality?

That is 0.98 is what this statement means.

Given the fact that the supplier is S1 what
is the probability that the raw material is

of good quality that is 98 percent which is
the conditional probability.

Similarly this 95 percent you will realize
is also conditional probability.

This is the conditional probability that raw
material will be of good quality given the

fact that it has come from supplier S2.

What is the probability that the raw material
being good quality given the fact that it

has come from S2 that probability is 0.95.

So, these are conditional probabilities.

As I said these are the numbers which can
help you build your contingency table.

Now how will you calculate the joint probability?

If I ask you this question what is the probability
that raw material is being supplied by S1

and it is of good quality the moment you say
and it is a joint probability.

So, go back you know marginal you know conditional
probability you know conditional probability

this is conditional probability.

Now from this data from the data on this slide
you want to calculate this joint probability.

So, this can be calculated using the Baye's
formula that we have discussed earlier.

So, what did Baye's formula say?

Baye's formula said conditional probability
is equal to joint probability divided by marginal

probability.

Now we are interested in calculating the joint
probability.

So, how do you calculate?

The joint probability joint probability will
be conditional probability multiplied by marginal.

So, go back to what was this conditional probability,
this was the conditional probability of raw

material being of good quality given the supplier
being a S1.

So, let us write down notation.

So, what is this?

This is the probability that the raw material
is of good quality given the fact that it

came from S1 what will this be equal to?

This will be equal to the probability that

the raw material is of good quality and it■

is supplied by S1 multiplied by the marginal■
probability that the supplier is S1.■

So, how will you calculate this?■

How will you calculate this joint probability?■

How will you calculate this joint probability?■

This joint probability is going to be calculated■
as conditional property multiplied by the■

marginal probability.■

How will you get this probability of S1 from■
this data?■

What is the probability of S1?■

What is the probability that the raw material■
actually came from a S1?■

Well, think about it . In the next slides,■
this is 0.65.■

Why is it 0.65? 0.65 because 65 percent of■
the raw material anyway comes from supplier■

S1.■

So, what is the general program that the randomly■
selected raw material came from a S1 it is■

65 percent of 65 percent and therefore the■
conditional the joint probability that the■

supplier is this S1 and the raw material is■
of good quality it should be and or some textbooks■

use it a comma does not matter which notation■
you that probability is 0.637.■

So, once again what is this?■

How do you interpret this 0.637? 0.637 is■
the probability that the raw material being■

supplied by S1 and it is of good quality is■
this joint probability.■

■

Similarly let us erase this.■

■

So, that we can read this carefully similarly■
you can calculate the joint probability that■

■

the raw material actually came from S2 and■
it is of good quality.■

■

This is the joint probability and you figure■
out that that joint probability is 0.3325.■

■

Now let us present a new scenario.■

■

Now let us say that this manufacturer has■
inspected incoming raw material.■

■

Raw material is coming in the trucks getting■
unloaded and as soon as the material gets■

■

unloaded there is an incoming quality check■
and in the incoming quality check the raw■

■

material is inspected and it is found to be■
of a bad quality.■

■

It is found to be a bad quality.■

■

Now the question is without looking at the■
papers the manufacturer can look at the papers■

■

that came in the trunk saying that this is■
the invoice and this is the challenge and■

■

this is so on and telling you who is the supplier.■

■

Without looking at those details simply by■
looking at the fact that the raw material■

■

received is of bad quality can the manufacturer■
infer that it must have come from supplier■

■

S1 or it must have come from supplier S2.■

■

Once again I will post the question again.■

■

The scenario is the manufacturer has just■
received a truck of raw material.■

■

The truck was unloaded.■

■

The raw material was inspected on the receipt
the incoming inspection is called and that

incoming inspection found out that the raw
material which was supplied was of bad quality.

Now without looking at anything further without
looking at the papers and determining or asking

anybody the manufacturer wants to know who
the supplier must have been to whom to blame.

So, what is the manufacturer interested in
finding?

The manufacturer is interested in finding
the probability that the supplier was a S1

or what is the probability that the supplier
was s but the manufacturer is not interested

in calculating the marginal probability that
the manufacturer already knows 65 percent

of the time raw material comes from S1 35
percent of the time raw material actually

comes from S2.

So, you can simply say that in this scenario
I received bad quality material.

In this scenario also 65 percent of the time
I will blame supply a S1 35 percent of the

time I will blame S2 that is not correct because
you are using marginal probabilities.

You should not be using marginal probabilities
because you have updated information you have

collected.

What is the data?

The data is the incoming inspection that you
have conducted.

The incoming raw material inspection is the
additional data that you have additional data

that you have with you.■
■
And you have found out that it is of bad quality.■
■
You have found out that it is of bad quality.■
■
So, in light of this information, now can■
the manufacturer find out what is the probability■
■
that this raw material was supplied by supplier■
S1 or this raw material was supplied by supplier■
■
S2 can we find that out?■
■
So, essentially we are asking ourselves a■
conditional probability question.■
■
So, essentially the manufacturer would like■
to know that I have a bad quality product■
■
in my hand which was revealed during the incoming■
quality inspection which needs to be blamed.■
■
Which supplier can be complained about and■
we cannot simply say 65 percent of the time■
■
blame S1 35 percent of the time blame S2 that■
I am already telling you that is not the correct■
■
answer.■
■
That is because that is the marginal probability■
and as we have explained we are not interested■
■
in marginal probability.■
■
We have additional information we have information■
that came with that was known to us only after■
■
the incoming quality inspection.■
■
We have additional information.■
■
Why do not we use that additional information■
and then decide who the supplier must have■
■
been or may have been?■
■
That is the question that we are answering.■
■
So, essentially what we are interested in■

is we are interested in the posterior probability

posterior probability that the supplier is
guilty of supplying a bad quality product,

a particular supplier given the fact that
you already have a bad quality product at

your doorstep.

So, given the fact that you already have a
bad quality product what is the probability

that it came from S1 or given the fact that
you already have a bad quality product in

your hand what is the probability that it
came from S2.

So, this is essentially Baye's rule.

So, as we said Baye's rule is conditional
probability ratio of joint probability and

marginal probability.

Now we are going to play a trick this joint
probability can now be written as marginal

probability and conditional probability again
this joint probability can be written as marginal

probability and joint property.

So, writing that once this S1 and B can be
written as probability of a S1 and probability

of B given S1 why is that true?

That is true because of what we had written
earlier.

Why is that true?

We would have said what is the probability
that given the fact that supply is S1 what

is the probability that you will get a bad
quality product?

That is calculated as the joint probability

S1 divided by the probability of S1.■
■
Now if you say this is the joint probability■
you multiply this and this and this is what■
■
you get.■
■
I hope this is okay.■
■
Now you can similarly expand the denominator■
also you can supply you can you can actually■
■
expand the denominator also.■
■
Now let us understand what the denominator■
is.■
■
What is the denominator?■
■
Denominator is the marginal probability that■
the raw material is of bad quality.■
■
It is a marginal probability that it is a■
bad quality product.■
■
Now there are only two sources for this: either■
the bad quality product could have come from■
■
S1 or it could have come from S2.■
■
So, what do you need to do if you have a bad■
quality product in your hand?■
■
So, what do you need to do?■
■
You need to calculate the probability that■
you have bad quality given the fact that it■
■
must come from S1 or you have bad quality■
product given the fact that it must have come■
■
from S2 but what is the probability that the■
supplier was S1 that is a probability of S1?■
■
What is the probability that the supplier■
was S2.■
■
Therefore what is the probability that you■
will have a bad quality product in your hand?■
■
You will have a bad quality product in your■

hand if it was applied by S1 or if it was

applied by S2.

The moment you say or you put up plus the
bad quality product must have come from either

S1 or S2 the moment you say or it is plus.

So, bad quality products could have come from
S1.

So, that is the probability of a bad quality
product given the fact that the supply was

S1 or you have a bad quality product given
the fact that it must have come from S2.

So, that is the marginal conditional probability
but what is the probability that the supplier

was S1?

That is the probability of S1 or probability
of S2.

So, that is how you expand the denominator
also.

So, which is what is explained here, what
is the probability of event B? probability

of event B is the probability of receiving
a bad quality product.

Now the bad quality product could have come
from supplier S1 or S2.

So, a bad quality product is either a joint
probability of S1 and B or S2 and B. Now you

can write S1 and S1 and B as multiplication
of marginal probability multiplied with the

conditional probability and therefore you
can expand the denominator as we had written

in the previous slide.

Now this is what you are interested in calculating.

Now essentially this is what you wanted.

What is the probability that supplier S1 will have to be blamed given the fact that I have a bad quality product at my doorstep?

This is what you are interested in.

Now we have an equation for that.

Now let us plug in the values and calculate the value of the conditional probability that we need.

How do you plug in values?

What is the probability of S1 you know the marginal probability of S1 which is 0.65 this is 0.65 this is 0.35 marginal probability that the supplier S2 35 percent of the time the raw material is applied by S2 therefore the marginal probability that the random reselected raw material was supplied by S2 is 0.35, 35 percent.

How do we get this?

How do we get this?

How do we get the conditional probability of supplier being S2 conditional probability of the raw materials of bad quality given the fact that supplier was S2.

For that we go back to the first information about this problem.

This we know this we know this.

Now what is this?

This is a conditional probability that the raw material is of good quality given the fact that it was supplied by S1 that is 0.98.

Now supplier S1 can provide or can supply either good quality product or bad quality product.

You already know the conditional probability of a good quality product . Can we not calculate the conditional probability of a bad quality product?

We should be able to know fairly straightforward 100 minus let us do that.

So, what is the probability that the raw material is of bad quality when the supplier was this S1 100 - 98 which is 2.

Similarly what is the probability that the raw material was a bad quality given the fact that supply was S2.

Now suppliers S2 supply good quality product 95 percent of the time therefore she must be supplying bad quality product 5 percent of the time which is this 0.05 which is this 0.05.

And therefore after these calculations we find that the probability of supplier S1 being a culprit given the fact that I have received a bad quality product in my factory that probability is 42 percent 42.6.

Similarly supplier S2 being a culprit given the fact that I have a bad quality raw material in my hand that probability is 0.574 or 57.4 percent.

Now let us look at the priors: what was the prior on this?

What was the prior probability for S1?

This was 0.65 what was the prior on S2?

This was zero 0.35.

Now in absence of anything in absence of anything you say that 65 percent of the raw material

is supplied by S1 therefore what is the probability that a randomly selected raw material is supplied

by us 65 percent.

What is the probability that a randomly selected raw material is supplied by S1 35 percent

35.35.

But you have additional information now you inspected the incoming raw material and you

found that the incoming raw material was of bad quality.

Now in light of this information what is the probability that the culprit supplier was

this S1?

you are calculating a posterior probability that the raw material was supplied by S1 given

the fact that the raw material was of bad quality.

You are calculating posterior probabilities using your initial belief and additional data

additional quality data inspection data.

Now you realize that profit is only 42 percent even though 65 percent of the raw material

is supplied by supplier S1.

If you find a bad quality raw material the probability that it came from S1 is only 42

percent.

On the other hand the prior belief about supplier

being S2 is 35 percent prior belief about

supplier being S2 is 35 percent however if
you find a bad quality raw material already

in your hand the probability that it actually
came from S2 goes up to 57 percent.

So, that is the difference between using prior
knowledge or coming up with posterior probabilities

with some additional information.

This is additional information that B comes
from the additional incoming inspection that

you have conducted.

You have additional data now.

In light of this data your probabilities change
and those probabilities are called posterior

probabilities of S1 and S2.

So, that is the use of conditional probability.

Once again are being used in this case for
supplier quality data.

Once again which are the two random variables
which are the two categorical variables in

this example the two categorical variables
are suppliers.

Supplier is a categorical variable because
there are only two: one is called S1 the other

one is called S2.

Now do not think that this is S1 and therefore
they should come first and this is S2 it will

come next therefore they come it can be arranged
anyway.

So, the supplier is a categorical variable
which has 2 values S1 and S2.

Quality okay quality is another random variable

which has bad quality and good quality.■
■
So, this quality of a random variable categorical■
variable has 2 values: bad and good.■
■
So, these are the two categorical variables■
and we have determined the association between■
■
the two categorical variables using Baye's■
rule using Baye's rule.■
■
I hope this is understood if you have not■
understood.■
■
We will definitely have tutorial sessions.■
■
You can definitely review the earlier discussion■
on conditional probability in other courses■
■
or you can review this video again and we■
always have help for clearing the doubts.■
■
Welcome to this session on drawing inferences■
about the association between two categorical■
■
variable.■
■
In the previous session we had seen how to■
quantify association between two categorical■
■
variables through an application of conditional■
probabilities.■
■
Let us extend that discussion and.■
■
Now focus on drawing inferences about the■
association between the variables.■
■
And right.■
■
Now we are limiting our discussion to association■
between two categorical variables.■
■
So, let us take an example let us take this■
example of brand preferences let us say that■
■
there are three brands and there is brand■
A brand B and brand C and let us say that■
■
there are different preferences among the■
brand A brand B and brand C in different cities.■

And this data let us say is collected from
a survey that was conducted in two cities

Mumbai and Chennai when we ask them what brand
do you prefer?

So, 279 people in Mumbai said that they prefer
brand A.

73 people in Mumbai said they prefer brand
B 225 people in Mumbai said that they prefer

brand C. So, totally 577 respondents and this
was the breakup of their brand preference.

403 people were surveyed in Chennai and their
preference for brand A and B and C is in the

second row all right.

So, out of the 980 people who were surveyed
444 said that they prefer brand A 120 prefer

preferred brand B and 416 preferred brand
C.

Now we saw analysis of this here you notice
that there are two categorical variables which

are the two categorical variables?

There is one variable in the columns which
is the brand Columns brand A brand B and brand

C those are the columns.

So, there is a column variable column variable
is the brand and the row variable is the city

row variable is the city.

So, the city is we can call city as our exponentially
variable and brand preference as our response

variable.

So, there are two variables and if you notice
both of them are categorical Mumbai and Chennai

are the categories of the variable called

city.■
■
Brand A brand B and brand C are the categories■
of our variable brand or brand preference.■
■
So, now we want to infer about the association■
between the two exponent two categorical variable.■
■
So, we essentially know how to summarize this■
data by calculating what is marginal probability■
■
and joint probabilities do you recall that■
what was marginal probability?■
■
What was joint probability?■
■
We saw if you recall we said whatever is in■
the margins is called marginal probability.■
■
So, 577 divided by 980, 444 divided by 980■
what does 577 divided by 980 means it is the■
■
probability of randomly selecting a respondent■
from Mumbai city.■
■
444 divided by 980 is the probability of randomly■
selecting a person who prefers brand A. So,■
■
those were marginal probability.■
■
Now do you recall what was joint probabilities?■
■
You go back to that session and understand■
the definition of joint probabilities joint■
■
probabilities is somebody are respondent being■
from Mumbai and preferring brand A. Somebody■
■
who is from Chennai an she prefers brand B■
that will be the joint probability of definition.■
■
So, essentially we wanted to ask a question■
whether brand preference is associated with■
■
the city?■
■
If the brand preference was not associated■
with city the responses would have been similar■
■
right with responses would have been similar.■
■

So, we are asking a question, are these responses similar?

Are these responses similar?

So, we want to use statistical independence statistical dependence for this.

So, So, from the conditional probability discussion you remember that we said the categorical

variables are statistically independent.

If the conditional distributions for them is identical for each category that is what

I said here if the responses are similar to each other similar to each other which means

that the conditional probabilities are similar what what what would that table look like?

That table will look like something like this.

So, if we had a third city called Delhi and we get something like this.

So, we we surveyed thousand people in Mumbai we surveyed hundred people in Chennai we surveyed

250 people in Delhi and this was their brand preference but look at the rows 44% of Mumbai

residents prefer brand 14% of Mumbai residents preferred brand B 42% preferred brand C in

Mumbai.

But the proportion was same in Chennai and Delhii. 44% from Chennai also preferred brand

A, 14% from Chennai also preferred brand B 42% of Chennai also preferred brand C. So,

this 44%, 14% and 42% was the same similarly for Delhi 110 out of 250.

So, 44% preferred brand 35 out of 250 which is about 14% preferred brand B 105 out of

250 which is about 42% preferred brand C.

So, the proportion did not change proportion did not change.

So, in such a case we can say that the two categorical variables are independent two categorical variables are independent and we can conclude that really brand preference does not seem to be dependent on cities.

You go to any cities 44% approximately 44% people prefer brand A Approximately 42% people brand prefer brand C and brand B gets the least preference out of these three brands.

Irrespective of the city that you pick but this was this was still based on the data that we have collected this was still based on the sample of 1000 people sample of 100 people sample of 250 people in each one of the cities.

So,I mean before we go there you also remember the discussion that we we said that the statistical independence is actually a symmetric property.

So, if a brand preference is independent of city, city is also independent of brand preference.

So, if you actually calculate the proportion for columns here we calculated the proportion for rows.

So, 100% out of 100% respondents 44% prefer brand A 14% prefer brand B and 42% preferred brand C this was the row proportion.

If you calculate the column proportion even that is going to be same right even that is going to be same.

So, that is going to be the same that tells

you that statistical independence is actually

a symmetric property but as I said this is
based on sample data what about the population?

Can I generalize these results?

can I generalize these results and say that
this independence will actually apply to the

entire population.

From the sample of 1000+100+250 across the
three cities we seem to think that brand preference

is independent of the cities does the conclusion
extend to the population?

Which is essentially inferencing can I infer
about the population from this sample in that

case simply looking at conditional probabilities
may not be sufficient we may have to look

at something more.

So, can we draw inferences about the population
from this single sample from this single sample

of 980 respondent this single sample of 980
respondents.

I have collected a single sample 980 respondents
577 from Mumbai 403 from Chennai.

Can I conclude about the entire population
who prefer brand A or brand B or brand C in

these two cities by looking at only this sample
that is the that is the objective of this

session.

So, how are we going to do this?

We are going to do this by testing hypothesis
obviously that is what we have been doing

for inferences.

Remember from from your BDM course.

How do we infer about the population?

We infer about the population by running a hypothesis test.

This is a special hypothesis test because it has a very different test statistics.

So, what is the hypothesis?

The hypothesis is that the categorical variables are independent it is always the no effect

null hypothesis.

Now null hypothesis is always the no effect null hypothesis.

Now hypothesis is these two categorical variables the brand preference and the cities are independent

no effect hypothesis.

Alternative hypothesis is no no they are not independent they may be dependent for that

what we are going to calculate observed frequencies and expected frequencies.

Observed frequencies come from the sample expected frequencies are going to be calculated

assuming that the null hypothesis is true.

Assuming that the variables are independent what frequencies do I expect?

What frequencies do I expect Now I am going to do this indirect validation of my null

hypothesis or not by comparing the observed frequency with expected frequency.

Now let us let us do this and calculate the test statistic.

So, what is what is the observed frequency?

Observed frequency directly comes from the sample these are the observed frequencies

279, 165, 73, 225 these are the observed frequency.

225 people really in Mumbai prefer brand C,
47 people in Chennai actually prefer brand

B this is actual observation.

So, this is observed frequency 165 is the
observed frequency.

279 is the observed frequency.

Now let us calculate the expected frequency.

Let us calculate the expected frequency.

How is the expected frequency going to be
calculated?

It is going to be calculated as the product
of row total into column total.

Row total into column total divided by the
total sample size let us calculate this.

So, how do you calculate the expected frequency?

How do you calculate this 261.4?

This 261.4 is calculated as row total which
is 577 577 multiplied by column total which

is 444 divided by 980 total sample size.

You can verify that that turns out to be 261.4
how did we get this 70.7?

This 70.7 was calculated as row total, row
total was 577 multiplied by column total column

total is 120 divided by 980.

How did we calculate this 171?

171 was calculated as row total which is 403
multiplied by column total which is 416 divided

by total sample size which is 980.

This is how I calculate the expected frequency.

■

This is a calculated this is how I calculated■
the expected frequency.■

■

279 is the observed frequency 261 is the expected■
frequency.■

■

Now I calculate my test statistic which is■
called the chi squared test statistic.■

■

The chi- squared test statistic is calculated■
as summation of the observed frequency minus■

■

expected frequency squared divided by the■
expected frequency.■

■

So, let me let me do this let me do this just■
to show you 279 and 261.4 observed frequency■

■

and expected frequency how are we going to■
calculate the Chi squared for that cell.■

■

279 minus 261.4 whole squared divided by 261.4■
261.4.■

■

So, this is the first cell right observed■
frequency 279 minus expected frequency 261.4■

■

square it divided by the expected frequency■
plus summation sign is plus the second cell■

■

what goes in the second cell?■

■

Second cell is 73 is the observed frequency■
70.7 is the expected frequency 70.7 is the■

■

expected frequency squared divide that by■
the expected frequency plus the third cell.■

■

225 is the observed frequency 224.9 224.9■
is the expected frequency.■

■

So 225 minus 244.9 square this whole numerator■
divide that by 244.9 is the third entry.■

■

Similarly for this cell this cell and this■
cell.■

■

So, +165 is the observed frequency 182.6 is■
the expected frequency 182.6 is the expected■

■

frequency divide that by 182.6.■
■
And we do that for the remaining two cells■
and the summation of all this all of this■
■
is going to be called the Chi squared test■
statistic Chi squared test statistic.■
■
Now what is going to happen if this expected■
frequency is going to be very close to the■
■
observed frequency if the expected frequency■
is going to be very close to the observed■
■
frequency what will happen?■
■
we will get very small numerator we are going■
to square it we will square it essentially■
■
to remove the negative signs.■
■
If the expected frequency is very close to■
the observed frequency the numerator is going■
■
to be fairly small and therefore the value■
of the test statistic is going to be smaller.■
■
But when am I going to get expected frequency■
close to the observed frequency when am I■
■
going to get that when am I going to get that?■
■
I am going to get that So, I am going to get■
this observed frequency close to the expected■
■
frequency only when the null hypothesis actually■
is true only when the category will be able■
■
to be independent only when the categorical■
variables are independent.■
■
So, when the null hypothesis is actually true■
the expected frequencies and observed frequencies■
■
are going to become close to each other and■
the test statistic is going to be relatively■
■
small.■
■
However if the null hypothesis is actually■
not true then for at least some of the cells■

the gap between the expected frequency and
the observed frequency will be quite large

and therefore that will result in a very large
value of the test statistics.

So, how do I decide whether the hypothesis
is true or not?

Simply by looking at the test statistics.

If the test statistic is larger in value that
gives me evidence that the null hypothesis

may not be true and therefore should be rejected
and vice versa.

So, essentially a large value of test statistic
is actually evidence against the null hypothesis.

It is actually an evidence against the null
hypothesis.

For Chi squared test statistic you actually
need degrees of freedom degrees of freedom

is actually calculated as row minus 1 multiplied
by column minus 1.

For example how many rows there are two cities.

So, the number of rows is equal to 2, r equal
to 2.

In our example therefore r minus 1 will be
one.

How many columns there are three columns for
three brands and therefore C is 3, C minus

1 will be 2 and therefore degree of freedom
for our test will actually be one multiplied

by two row r minus 1 will be 1, column C minus
1 will be 2.

So, 1 multiplied by 2 is 2 there are 2 degrees
of freedom for our Chi squared test and we

are going to get our Chi square test from

the table and we are going to calculate Chi

square test calculate Chi squared value using
this equation.

Compare the two and decide whether my Chi
square value is large or not.

So, for our for our brand preference example
it turns out that the calculated test statistic

turns out to be 7.

So, this this turns out to be 7 calculated
value calculated value turns out to be 7.

What is the tabular value for two degrees
of freedom at 95% confidence our tabular value

of test statistic turns out to be 5.99 clearly
the calculated value is bigger than the tabular

value.

Therefore we reject the null hypothesis.

We reject the null hypothesis and conclude
what was the null hypothesis?

That the null hypothesis was that the categorical
variables are independent we reject this null

hypothesis and therefore say that brand preference
does depend on the cities.

If you go to different cities people are going
to have different brand preferences that is

our conclusion at 95% confidence.

That is our conclusion at 95% confident which
means that if somebody wants to be 95 somebody

wants me to be 95% confident I will say that
cities do impact brand preference they are

not independent and I am saying this with
90% confidence 95% confidence.

What if somebody wants me to 99% confident.

What if somebody wants me to be 99% confident.

If I want to be 99% confidence the alpha value
turns out to be 0.01 and the tabular value

increases.

Tabular value goes up from 5.99 to 9.21.

Now comparing 9.21 with 7 suddenly the 7 does
not seem to be large value.

Since the calculated test statistic is smaller
than the tabular value of the test statistic

I end up not rejecting the null hypothesis.

What do I mean by not rejecting null hypothesis.

I cannot reject the null hypothesis and I
will conclude that brand preference does not

depend on cities.

Earlier when somebody wanted me to be 95%
confident I concluded that brand preference

changes with cities.

If somebody wants me to be 99% confident however
my result changes now I end up saying that

those two categorical variables are actually
independent.

So, it does not matter which city you go to
brand preference remains almost same.

Now when I am saying this when I am drawing
these conclusions when concluding about the

hypothesis I am essentially inferencing about
the entire population and not only these 980

value that we have collected from the survey.

So, now our results are more generalized.

Earlier when we looked at the conditional
probabilities we could say only about the

sample we could say from the sample it appears.■
■
Remember that conditional probability example■
was about possibility of what was about the■
■
MBA admissions given to per male and female■
candidates.■
■
Now we are talking about the entire population■
and not only the 980 values that we have collected■
■
okay not only the 980 values that we have■
collected.■
■
So, that is we drawing inferences about the■
association between two categorical variables■
■
using a Chi squared test of independence.■
■
Let us end the session here.■
■
Hello, everyone.■
■
Welcome to this lecture on the implementation■
of the chi-square test of■
■
independence in python.■
■
I am S. Srivatsa Srinivas, a co-instructor■
at the IIT Madras online B.Sc.■
■
degree program.■
■
We are going to look at the hypothetical dataset■
discussed before where we■
■
had the list of cities Mumbai and Chennai■
and the brand preferences in different cities■
■
across■
three brands A, B and C.■
■
This is what the hypothetical dataset look■
like.■
■
We have the list of■
cities and the list of brands.■
■
The objective of this exercise is to understand■
whether different cities have any impact on■
■

brand■
preferences.■
■
That is the null hypothesis in the chi-square■
test of independence is that the two■
■
categorical variables are independent and■
the alternate hypothesis is that the two categorical■
■
variables are non-independent.■
■
That is what we try to establish with this■
case.■
■
We need to perform this test by importing■
certain packages.■
■
We need to require the following■
packages NumPy, pandas, and scipy and within■
■
scipy we require the stats package.■
■
And since■
we are performing this exercise on Google■
■
Colab, I have uploaded the file to Google■
Colab and■
■
then I extract the CSV file.■
■
This is what the file looks like.■
■
A part of the file will look like this.■
■
The first step in the chi-square test of independence■
is to construct the contingency table.■
■
We■
have a nice functionality called crosstab■
■
with the pandas which will help us do.■
■
Across cities and■
brands, we construct the contingency table.■
■
The contingency table will look like this,■
where we■
■
have the cities and the brands and this is■
from the given dataset.■
■

This is called the observed
frequency.

To perform the chi-square test of independence
we need to compare this observed frequency

against the expected frequency.

That is we are trying to use a certain sample
and then

comment on whether the categorical variables
are independent or not.

We use a sample and
then comment on the population.

To do that, we need to first construct the
expected frequency.

Only when we construct the expected frequency
and we compare the observed frequency

against the expected frequency.

We have the observed frequency table here.

The next step is
to construct the expected frequency.

But before we do that, we need to know how
to access values within this table here.

Contingtab
of A will give us value corresponding to brand

A.

We have 165, 279, and 444.

Contingtab of A of
Chennai will give us 165 corresponding to

brand A and city Chennai.

And contingtab of all will
give us the total sum or the total observed

frequencies.

There is also another command which we will

look at which is contingtab dot transpose■

■
which■
will generate the transpose of the given table.■

■
Instead of city and brand, we have interchanged■
it■

■
to the brand and city.■

■
We require this command to access certain■
values which I will show you■

■
as we move along.■

■
Now the objective is to calculate the expected■
frequency.■

■
We first generate the list of cities which■
are Chennai and Mumbai.■

■
We have this data frame in■
here with the city.■

■
The unique values in the column city will■
give us the list of cities.■

■
And■
similarly, the unique values within the column■

■
brand will give us the list of brands.■

■
Once we do that, we create an empty dictionary■
called exp1 and then run a loop over cities■

■
and■
brands.■

■
Within the cities loop, we create another■
dictionary exp2.■

■
And finally, this exp2 will help■
us calculate the expected value corresponding■

■
to a city and a brand.■

■
This contingtab dot■
transpose of I will give us the value corresponding■

■
to the particular city.■

■

We have the city here and the value i is corresponding to Chennai.

The value of all here is 403 and the contingtab of j of all is corresponding

to the brand.

The brand here happens to be A is 444.

We have 403 into 444 divided by 980 which will be the value of Chennai in this case.

Let us verify for one combination that you are

convinced.

We have 403 into 444 divided by 980 which happens to be 182.58.

As we run over this loop what we store is the values of the expected frequency within

this dictionary exp1.

When we run this exp1 what we obtain is corresponding to Chennai A, Chennai

B, Chennai C, and so on.

For now, we are done with calculating the expected frequency.

The next step is to calculate the value of the chi-square.

And the chi-square value is calculated as the observed frequency minus the expected

frequency of the whole square divided by the expected frequency.

And this needs to be sum across all combinations of brands and cities.

We run a loop over cities and brands and then

we calculate the value of a particular observed value

minus the expected value of the whole square divided by the expected value.

And this needs to be summed across every combination.

And this is how the chi-square calculated value is generated.

As you can note, the chi-square calculated value turns out to be

7.009.

You can go back to the earlier lecture and refer that the chi-square calculated value

is

indeed the value shown in this case.

Now, the degrees of freedom are the length of cities minus 1 into the length of brands

minus 1.

We have cities to be 2, which is 2 minus 1, and brands to be 3, 3 minus 1.

Therefore, 2 into 1 will be 2.

Now, we use the stats package to calculate the tabulated value of the chi-squared.

We

have the level of significance to be 0.05

and we use this stats dot chi2 dot pdf to calculate the

tabulated value of chi-squared.

And when we do that, we update 5.99.

You can go back and refer that the tabulated value is

indeed 5.99.

Since the calculated value is greater than
the tabulated value, we are going to

reject the null hypothesis that the two categorical
variables are independent that is what we

conclude from this.

Instead of constructing the contingency table
and performing multiple steps, there is a

shortcut
method to this chi-squared test of independence

in python.

In that, we first create a NumPy
array as follows.

We had created the contingency tab using the
crosstab functionality of pandas

so that contingtab dot transpose of Chennai
will give us the values corresponding to Chennai.

And the contingtab dot transpose of Mumbai
will give us the values corresponding to Mumbai.

We only consider the first three values in
this case that is corresponding to brands

A, B, and C.
And we then have the values corresponding

to Mumbai brands A, B and C. And once we create
this contab, we can use the stats dot chi2

underscore contingency to generate the chi-squared
test results.

This becomes very straightforward in the fact
that you are using just one command to generate

the chi-squared test results.

As you can see here, you have directly calculated
the chi-squared

value to be 7.009 and you also update the
p-value, which is 0.0300, the degrees of freedom,

and the observed frequency.■
■
With a single command, you can update all■
these values.■
■
Now we have the level of significance to be■
0.05.■
■
Since this 0.03 is less than 0.05 this is■
a low■
■
p-value for rejecting the null hypothesis.■
■
Therefore, we can conclude that the categorical■
variables are not independent.■
■
If we can go back to the chi-squared calculated■
value, which was■
■
7.009, we can also calculate the p-value in■
the following manner.■
■
We have stats dot chi2 dot■
CDF and the chi-squared calculated value and■
■
the degrees of freedom.■
■
1 minus this value will■
end up giving us the p-value.■
■
This p-value turns out to be less than the■
level of significance■
■
which is 0.05.■
■
Therefore, we end up rejecting the null hypothesis.■
■
This is how you implement the■
chi-squared test of independence in python.■
■
Thank you.■
■
Hello everyone, welcome to the lecture on the ■
Implementation of the Chi-square Test Of ■
 ■
■
Independence in Spreadsheets. We are going to ■
look at the implementation of the chi-square ■
 ■
■
test of independence in Microsoft Excel first. ■

So, we have the example from before. We have ■
■
■
the list of cities and the brand ■
preferences corresponding to each city. ■
■
■
Note that this is a hypothetical data set, We have ■
the following data set where we have the list ■
■
■
of cities and the brand preferences corresponding ■
to each city for a particular person. We have ■
■
■
to calculate the observed frequency in this case. ■
■
The first step is to calculate the observed ■
frequency. ■
■
■
The reason is that we are trying to comment on ■
the population based on the sample. We have ■
■
■
the sample in here with a given data set and we ■
are trying to understand whether the different ■
■
■
cities have any impact on brand preferences. With ■
this sample, we are going to conclude for the ■
■
■
population on whether these two categorical ■
variables are independent or not. ■
■
■
The null hypothesis in the ■
chi-square test of independence ■
■
is that the two categorical variables ■
are independent and the alternate hypothesis is ■
■
that they are not independent. We can calculate ■
the values as follows. We have Mumbai and Chennai ■
■
and brands to A, B and C. We have the ■
observed value here. ■
■
■
We use the count ifs function to satisfy ■

certain criteria. The first case is we want

Mumbai and A.
We can do this equal to Mumbai

and brand to be equal to A.

This is how we calculate the
observed frequency for Mumbai and A. I now lock

these cells that we can extend it across the
other combinations.

We have 279 for Mumbai and A, we can drag and
drop across other cells. We require Mumbai

and B in this case, This will be Mumbai and B,

this will be Mumbai and C, this will be Chennai
and A,

this will be Chennai and B. Make
sure that you change the brand here,

and finally, this
will be

Chennai and C. We have the joint distributions
for the observed frequencies.

Now, we require the marginal
distributions corresponding to each brand

and each city. Which is
the sum of the columns here,

and the sum of the rows here.

And just use the drag and drop
feature of Excel, which will help us do this. And

also we calculate the overall sum. This is what
the observed frequency table will look like.

Next, we need to calculate the expected frequency.
As pointed out earlier, we are going to look

at the sample and then comment on the ■
population. Therefore, we need to calculate the ■

expected frequencies. And the expected ■
frequencies are calculated using the marginal ■

distributions. In the first case, where we have ■
Mumbai and A, this will be the value here ■

corresponding to A and corresponding to Mumbai ■
divided by the total which will give us the ■

expected frequency corresponding to ■
the combination Mumbai and A. ■

And as you can see, the denominator is going to ■
be common across different combinations and ■

we need to drag it across E4, then F4, and G4. ■
This E4 needs to be dragged across, I do not fix ■
that, but I fixed the other value here. ■
That is how I calculate the ■
value corresponding to Mumbai. ■

Similarly, for Chennai, can be ■
calculated as E4 into 403 which is in H3 ■

divided by the total, ■
which is H4. ■

Again, the total is not going to ■
change, I fix this and H3 is also not ■

going to change I fixed this, ■
and you can drag it across. ■

Now, we have calculated the ■
observed and expected frequencies. ■

Now, we need to calculate
the chi-squared value.

To do that, we first obtain the observed minus
expected whole square divided by expected for

each cell. What is the value corresponding to
Mumbai and A it is E2 minus Mumbai and A can

be expected, and then we have the squared term
here divided by expected. This is what is the

first value. And we drag it across the other
cells and obtain these values, the chi-square

calculated value will look something like
this. This is the sum of all these values.

As you can see, this is 7.009. You can go back to
the lecture and verify that D is 7.009. Now, we

need to look at the tabular
value of the chi-square.

The tabular value of chi-square can be
calculated using the inverse function, We have

the level of significance to be 0.05 and the
degrees of freedom to be 2.

Why the degrees of freedom is 2 is because we
have two cities Mumbai and Chennai 2 minus 1

becomes 1 and three brands A, B, and C 3 minus
1 becomes 2. So, 1 corresponding to the city

multiplied by 2 corresponding to the brand will
give us the degrees of freedom. And we obtain

the chi-square tabulated value to be 5.99.

Since the calculated value of chi-squared

is greater than the tabular
value of chi-square, we can

conclude that the null hypothesis is rejected.
Is there a way where we can perform the chi-

square test directly instead of performing these
additional calculations? We do have a chi-

square dot test which asks us to give the actual
rates. By actual rates we mean the observed

frequencies and by expected rates, we mean the

expected frequencies. And you obtain the P-
value directly in this case. You obtain the

p-value directly.
Since this P-value is

lesser than the level of significance
0.05 we can reject the null hypothesis.

This is what you conclude by directly calculating
the P-value from the observed and the

expected frequencies. This is how you implement
the chi-square test of independence in Excel.

Can you perform the same test on Google Sheets,
of course, you can do. Since these count tips

are directly following here I am going
to just copy the values in this table

directly here onto
Google Sheets.

This is just for you to illustrate
that this chi-square dot test

also works on Google Sheets. We
first have the observed frequencies and we then

have the expected frequencies and this directly
gives you the P-value.

And you can also do the other way,
the longer method where you have

chi-squared dot i and v
which is just the inverse

1 minus the level of significance which
is 0.05 degrees of freedom is 2.

We have taken 5.99. And you know already that the
chi-squared calculated value turns out to be

the sum of all these values, which is 7.009. And
again, the null hypothesis is rejected. This is

just for you to know that the same method can
be implemented on the Google Sheets, as well,

the commands corresponding to the chi-square dot
test, or the chi-square dot inverse also work

on Google Sheets. This is how you implement
the chi-square test of independence on

spreadsheets. I hope, you
like this lecture. Thank you.