

Good morning everyone, welcome to this session on data visualization.

We are lucky to have Annad Srinivasan with us today, joining in from Bangalore.

Anand is an IITM alumnus B. Tech 1995 is that correct Anand, yes and he has been a very

amazing data scientist throughout his 26, 27 year old career has worked with multiple

industries Airline industries, Saber technologies, Dell.

He was with the senior management at Go Air and now he is about to start a new venture,

a new airline which is going to be launched very soon.

I can say that formally right Anand, yes you can say that formally Rahul, perfect.

So, very glad that recently Anand was listed as one of the top 10 data scientists in India.

So, very lucky to have you Anand on joining us today for the session, my pleasure Rahul

it is always good to touch base with you and happy to help in any way I can.

So, today's session is going to be on how you see data that is going to be the primary

purpose.

How do you represent data and most of the ideas are essentially Anand's idea that we

are going to discuss and therefore we have Anand in the session today.

So, let us start.

So, first of all, why do you think we need to emphasize on visualization of data?

So, you know there is a cliché right I mean a picture being worth a thousand words fundamentally

■
that still holds true.■
■
End of the day when you look at analytics■
as a whole the fundamental purpose is you■
■
know people talk about data science etcetera.■
■
The only reason that field exists is to help■
businesses make better decisions, correct.■
■
If you count that then I mean that you question■
the very we need for the analytics function■
■
as a whole right.■
■
So, I have always been an exponent of saying■
look, it does not matter what kind of decisions■
■
are based on whatever analysis.■
■
Part of that is to say why we might build■
all kinds of sophisticated models you have■
■
to understand that decision makers you have■
to be able to present them that result of■
■
that analysis not necessarily the process■
but the result of that analysis in a very■
■
precise concise and consumable format which■
will go back and say hey if I do this this■
■
will this is likely to happen is this a good■
decision or a bad decision, correct.■
■
I think so, to me that is the crux of the■
problem right I mean let us face it in the■
■
forest and nobody heard it.■
■
So, you might have the greatest analysis but■
if good decisions are not being made based■
■
on that analysis then really what is the point■
of doing that analysis.■
■
So, I think that is where the visual ability■
to communicate and communicate visually to■
■
decision makers is a very, very critical part■

of analytics.■

■
In fact, that is my pet peeve if you want■
to call it that.■

■
It is like you know a lot of people that talk■
about analytics or you know people that teach■

■
analytics they do not emphasize this component■
enough.■

■
And people who have been endless will tell■
you how long the number of times come out■

■
to me after a long meeting with senior senior■
management and then come back and say oh we■

■
had this great model we had this great analysis■
but somewhere the whole discussion got derailed■

■
into something else and people did not really■
understand the core of what it is.■

■
And we ended up wasting time discussing some■
trivial aspects which were not core to the■

■
analysis that was presented all the time.■

■
And part of that is it is not that the management■
is sitting there saying look let us have chai■

■
samosa pakora and let us just white time away■
that they are there for a reason they want■

■
to understand the issue they want to understand■
the solutions.■

■
They want to make decisions but somewhere■
there the analyst has to take responsibility■

■
to communicate that effectively and I think■
that is where visualization comes in in a■

■
very, very critical way.■

■
So, I think I cannot emphasize the importance■
of visualization enough.■

■
So, in fact I am very happy Rahul that you■
are touching upon this as part of the course■

■

because I am not seeing this exercise because■
they want to quickly jump into statistical■

■
modeling recreation models and more importantly■
somewhere you know someone seems to how quickly■

■
can we get to AI and then I think emphasizing■
the fundamentals is so, critical lovely to■

■
see that, thank you.■

■
And not only for the corporate managers right,■
I mean I think psychologically also human■

■
beings are tuned to see pictures better than■
words right, yes absolutely absolutely right.■

■
In fact that is the thing right I mean this■
is well known right vision is the most powerful■

■
sense that we have there right.■

■
In fact I think there have been studies that■
talk about cognitive studies right 70% of■

■
the information we consume is visual right■
auditory sensory touch.■

■
So, I mean you know people can sit and debate■
the 70% 30% you know like I always say right■

■
85% of statistics are made up on the fly right.■

■
But that being said, I do not think so.■

■
I think it is beyond debate that vision happens■
to be the most powerful perception that we■

■
have.■

■
So, which is why visual communication is so■
important right and that is what we will talk■

■
about as we talk through some of these things,■
perfectly.■

■
And it doesn't really matter what kind of■
data you have, data can be numerical data■

■
can be continuous, it can be discrete and■
we have visualization tools right.■

■
We have visualization tools for each data type that I mean really there is no dearth

■
of visualization techniques.

■
So, interesting that your choice of words I mean I would emphasize the techniques more

■
than the tools right.

■
You know I think people once you understand the techniques you understand the fundamentals

■
the tool is just a medium right.

■
So, when you say visualization tool maybe perspective I am thinking of software tools

■
that help you do visualization that is not the issue, correct.

■
So, I would emphasize techniques more than tools because really what we want to teach

■
people are the techniques.

■
Then the tool just becomes a means of implementing that technique correct correct that is that

■
is we wish to this that is what we wish to discuss in the session today I mean essentially

■
what are the what are the core principles on which the visualization should be based

■
irrespective of what data do you have that you want to represent.

■
And you know what the other thing is.

■
I think you know people need to understand these kinds of categorical numerical discrete

■
continuous data.

■
Because the very nature of the data right dictates how you represent them visually okay,

■
correct.

■

The number of times I have seen people show me you know trending data using categorical

data, right.

So, that is the whole point right you know if you are going to show me say a line chart

which makes sense for you know time continuous trending kind of a data.

But then you decide to draw a line chart on categorical data somewhere there you know

when you say okay between this point and this point you have drawn a line connecting these

two points which is exactly between these two, right.

So, and you know and these kinds of things the amazing thing is you know when you when

you actually talk about it you know everybody will smile and say yeah come on I mean like

obviously is not that obvious part of the funny part depending the so on your point

of view is the fact that it happens so frequently okay uh.

So, I think it is just like you know part of it is tools that is the reason why I emphasize

the difference between tools and techniques because what happens is you pull up the data

you go to excel or something and then draw click and say create chart it just it just

it is a it is a dumb tool right it just does not work.

So, you know spending that little bit of time and hopefully the points we consider today

when we discuss will get people to think for you know two seconds before creating a chart,

right.

■
And the two seconds will save you two hours■
of debate time when you are presenting your■

■
results precisely, yeah yeah yeah yeah yeah■
yeah.■

■
So, what do you think are the general benefits■
of visualizing our visual representation of■

■
the data?■

■
See benefits see the whole idea is communication■
right I mean there is only one reason why■

■
look if you think about it business has to■
improve.■

■
So, we do analysis etc to help the business■
improve and we need to communicate that to■

■
the stakeholders.■

■
So, the right decisions can be made very simply.■

■
Our objective as analysts is to communicate■
that idea.■

■
So, visualization is our language of; so I■
could almost call visualization the language■

■
of the analyst because let us face it you■
go sit in a boardroom they are not going to■

■
be interested in looking at your python code,■
yeah, correct.■

■
So, visualization is the language of analysts■
when they talk to the business stakeholder■

■
right.■

■
Now and by the way interestingly visualization■
is also the language by which the analyst■

■
understands the real world.■

■
Because the right visual representation can■
actually give you a very good view of what■

■
could be the core problem and then actually■

put you on the right path to solving the problem,■

■

yeah.■

■

If you do not have a good visual representation■
the problem itself may not necessarily be■

■

evident.■

■

So, when you say it is visual the stuff you■
talk about cognitive processes here yes that■

■

is that that is what the communication is■
visual representation visual communication.■

■

And obviously communication is to make the■
cognitive process easier.■

■

So, part of what we do visualization is to■
say look, the other person has to understand■

■

the concept and we have to make it easier■
for them to understand the concept, not harder■

■

right.■

■

There are times when the visual representation■
itself is.■

■

So complicated that you know people spend■
more time understanding the visual than what■

■

it is trying to communicate, yeah yeah yeah.■

■

So, I call that the equivalent of going to■
a foreign country and saying you know where■

■

the washroom is and the person gives you an■
answer, right.■

■

You spend more time trying to understand what■
the person was saying rather than where the■

■

washroom is, yeah.■

■

It might be just easier for the person to■
point, yeah yeah yeah.■

■

So, you know, having the wrong visualization■
is the equivalent of getting a very simple■

■

answer in a completely spend foreign language.■

■
If you take more time understanding what the person will say you know and rather than saying■

■
okay look I asked you for the washroom you pointed me there problem solved this is thank■

■
you, correct correct correct correct.■

■
And I mean I have experienced this also I mean.■

■
So, many times we take up a constraint assignment and at the end of the assignment you are supposed■

■
to make a presentation to the senior management.■

■
And uh, gently the person comes to you before the meeting and says boss do not present your■

■
math and do not present your code.■

■
Tell us what it means to us, right?"■

■
So, unless we speak the language that people understand.■

■
So that they can make better decisions about our model and our code is useless, absolutely■

■
right.■

■
And you know and you know see someone's right I mean which is kind of straying away from■

■
the core concept of what we are talking about here.■

■
I think as engineers right we also fall into the trap of confusing effort for importance■

■
right yes.■

■
So, you know you might have spent 50% of your time right cleaning up the data and lining■

■
up the data correctly and maybe only 30% actually building the model or you know or say let■

■
us say 60, 70-30 70% and you know typical■

70 or you will spend you know just cleaning■

■
up the data understanding this.■

■
That does not mean you know that your process■
of cleaning up the data has to consume 70%■

■
of your time; the time of your final presentation,■
yeah right.■

■
So, so.■

■
so we do fall into the trap of mistaking effort■
for importance, right.■

■
So, it does happen.■

■
So, that is part of the visualization is to■
say look but I think that is slightly outside■

■
the scope of what we are trying to discuss■
here but you know also be cognizant of that.■

■
So, I think when we talk about this we will■
talk about you know what is the purpose of■

■
this presentation.■

■
So, maybe some of those things will be touched■
upon as we go through, correct, correct.■

■
So, essentially by visualization we mean visually■
highlighting a few things for example through■

■
the form or through color or through spatial■
representations right, yeah.■

■
So, see when you try to build a visual representation■
right and these are all I mean if you really■

■
think about it all charts that you do use■
one of these techniques to and these are the■

■
things that the human eye spots very easily■
correct.■

■
So, for instance right I mean if you look■
at it we are trying very easily to kind of■

■
get a sense of identifying the shortest bar■
in those four lines right.■

■
I do not have to tell you that look mark 20■
is shorter right.■

■
We are I just by looking at it I know correct■
the amazing thing is we are also very good■

■
at very quickly evaluating by how much it■
is shorter I mean somebody will look at this■

■
and say I would say it is about 15 shorter■
okay it is I mean you it may be 15 18 14.25■

■
but bottom line is we without saying anything■
I look at it I get a very good idea, righ.■

■
Similarly the width right I mean you know■
when you highlight it you know when you use■

■
bold text stresses of normal text.■

■
Again we pick up on that very, very quickly■
right of course orientation size shape you■

■
know and obviously using enclosure to highlight■
which is you know when you draw a person.■

■
So, these are all attributes of visual perception■
and the whole idea is when you do.■

■
When you communicate you have to think about■
which of these can I use to draw the user's■

■
attention to the important part.■

■
How do you get them to focus on what is important■
and how do you get them to not focus on what■

■
is not important, correct uh.■

■
So, like they say right, what you say is as■
important as what you do not show or what■

■
you show is as important as what you do not■
show.■

■
You know I might have made a crack about you■
know beachwear but maybe not appropriate for■

■
the current audience.■
■

So, we leave that as it is, yeah okay.■

■

So, okay now let us get down to the core of■
what we are trying to say right.■

■

You have often spoken about the four important■
kinds of umbrella principles of visualization,■

■

Can you elaborate on that right.■

■

So, I mean something that I mean you know■
I honestly speaking unfortunately I would■

■

like to know if I could I would go back and■
say that did I get this idea from uh.■

■

So, at the risk of saying you know it was■
pleasing I was inspired right at some point■

■

I might have seen this somewhere but I kind■
of it resonated.■

■

So, I kind of made it a cornerstone of a lot■
of times when I talk to people in terms of■

■

this.■

■

So, if there is somebody out there who actually■
came up with these things you know, consider■

■

the credit date granted, right right.■

■

So, the four things I always talk about is■
when you present a chart when you build this■

■

one.■

■

These are four things that you have to absolutely■
follow and in fact write it down and go through■

■

a checklist.■

■

First is to know the purpose.■

■

What is the purpose of putting this visual■
representation together, right?■

■

It has to have a purpose it is and the purpose■
cannot be you know let me demonstrate my mastery■

■

of the charting tool, okay.■

■
So, the purpose is why does this graph have
to be here?■

■
Why does this chart have to be here?■

■
Why does this representation have to be here?■

■
I want to use the word graph chart and visual
representation interchangeably but you know■

■
sometimes in cases there could be different.■

■
Once you have a purpose correct that purpose
will automatically determine what form of■

■
representation it will do, correct.■

■
Second, I always ensure the integrity of what
you represent, right?■

■
And you know typically you will have errors
of omission and commission when you are presenting■

■
data, right.■

■
To me that is a non-negotiable right integrity
of the data because even if it is by omission■

■
or commission.■

■
Even if there is a small error in the data
correct it will and that could be in the most■

■
trivial you know sidebar after this one but
it will derail the entire analysis and essentially■

■
you know even if it was oh you know it is
actually something I made a typo and I put■

■
it into this presentation, correct.■

■
Once it gets caught the integrity of the entire
presentation will be questioned.■

■
So, integrity is absolutely critical and it
is non-negotiable.■

■
correct and you know I always talk about you
know data inc and minimizing non-data which■

■

is the equivalent of saying you know what■
spend ink on the items that you want to show■

■
do not spend ink on items that you it is not■
which is not critical to the thought process.■

■
So, maybe just moving on will show your data■
and annotate yourself with the integrity of■

■
it as well, right.■

■
So, let us take this one by one uh.■

■
So, let us understand what we mean by purpose■
right.■

■
So, by purpose we mean the actual business■
problem that we are trying to solve, right.■

■
So, no, not really necessarily Rahul.■

■
So, because see if I build a chart or a visual■
right it is a one step in the larger question■

■
that I am trying to answer.■

■
So, which means that this is because let us■
face it right, the business problem I am trying■

■
to solve is not going to be solved by one■
visual representation, right yeah.■

■
I might be going through a slide and you know■
a series of slides, this one the purple.■

■
When I put a representative graphic together■
the purpose of that should say this is going■

■
to make my ability to communicate this more■
complicated concept of this one is that much■

■
easier it sets the stage in some cases or■
this might emphasizes this, correct, okay.■

■
Another way to do it is if you look at even■
this particular graphic that we are seeing■

■
on the screen, correct.■

■
Because I always talk about the umbrella principles,■
right.■

■
What do I mean by umbrella principles in English?■

■
It covers everything, correct, yeah.■

■
And if you look at that little handle that■
I have shown there, if I go back and say what■

■
the purpose of having that handle is, it emphasizes■
the concept of the umbrella principle, yeah■

■
correct.■

■
So, it is there for a reason, correct.■

■
So, which means literally everything that■
there has to be a reason.■

■
So, if you look at it and say why is this■
there, what is the purpose of having that■

■
particular graphic on the screen?■

■
You should have a clear purpose, right.■

■
Now the purpose again the reason I say a purpose■
is not necessarily the message what we are■

■
talking about how we communicate the message,■
right.■

■
The message could be communicated over a series■
of visual representations, okay and each of■

■
those will have a purpose that contributes■
towards communicating that message, okay okay■

■
okay, right.■

■
So, understood and in fact when people are■
starting off I always encourage you to whenever■

■
you put this together right what is the purpose?■

■
Does it serve a purpose?■

■
If it does not serve a purpose remove it.■

■
Sometimes and you know there is one thing■
to say it does not remove it sometimes certain■

visual objects will actually have a counter-
purpose, correct, yeah yeah.

So, you have to be very careful and focus
on what you want to say correctly.

So, I always encourage you to know when you
are putting thought through and say that what

is the purpose of having this visual representation,
this slide, this graphic whatever it may be,

right.

But just have that in fact I used to tell
people to write it down, right saying what

is the purpose right.

So, nice, correct, correct . I understand
the difference between the message that we

want to convey and the purpose for which this
visual tool is being used.

I understand, exactly, exactly , exactly , right
, okay.

Second thing was integrity yeah and this I
can see all all too often right.

It is pretty immediate.

So, for instance right I mean you know just
to give an example you should not be presenting

in a way to destroy right.

So, if you look at the graphic on the right
that shows you, let us say typically earnings

per share is this one.

Now here is the beauty of it right.

When you look at the graph on the left on
the graph on the right without paying attention

the first thing that one would say is that
oh my god the when I see the graph on the

left it is.■

■

So, volatile, yeah right in reality it is■
not correct.■

■

I mean look if somebody pulls out a calculator■
and says okay 164, 220 what is the percentage■

■

but remember the thing is about how usually■
we see length and we are very quick to measure.■

■

So, when I look at the 1999 versus 2000 right■
on the left graph one would say we grew six■

■

x or five almost yeah yeah correct right that■
is that was not the reality right because■

■

what somebody has done is you can see the■
access has been cut off, correct, yeah.■

■

Now and in fact when I say this right somebody■
will come back and say no no I want to emphasize■

■

the variability, correct.■

■

Now and that is where I say it is being the■
integrity is lost because the reality is the■

■

variability is minuscule, correct yeah yeah.■

■

We are exaggerating on the left hand side■
yeah.■

■

So, by chopping it off; so, a lot of people■
say no I want to emphasize the difference■

■

or the variability I said in the grand scheme■
of things the variability is nothing.■

■

So, they will say I want to show that north■
is so much better than south or sales in east■

■

are so much bigger.■

■

But in reality the per capita might be you■
know off by you know 0.5% but by chopping■

■

it off I will make it up here it is off by■
you know 30 40%, correct.■

■

um This is you will be; I am amazed at how■

frequently this is used especially in media■

exactly.■

I was going to say that, yes, okay.■

And their equivalent of caviar temper is they■
will have a small little wiggly thing at the■

bottom to show that access has been cut, yeah■
yeah yeah.■

Right.■

And you know and when I see this graph that■
is the first thing I notice right the small■

wiggly squiggly line at the bottom yeah yeah.■

So, this is very common in media right and■
you know when you look at it then you and■

amazingly enough when you look at it from■
the lens of the integrity is being gone you■

actually will understand the bias of the storyteller■
in terms of what they are trying to say here,■

yeah yeah yeah that could become very evident,■
right.■

So, uh.■

So, so to me how do you eliminate as a storyteller■
as a media person I need to be making a statement■

of fact right by chopping it off you are not■
stating fact right.■

So, to me integrator integrity is so critical■
right I mean I mean and I chose the word the■

left the graph on the left is deceitful it■
is a very strong word and I chose it deliberately■

because I believe it is deceitful it is communicating■
our wrong picture, precisely, yeah.■

But the sad reality is exactly what you said■
you see every day in print media in the news■

media in the electronic media you see this■
every day yes and you know the most common■

■
technique used, right.■

■
So, again as um analytics analytics professionals■
correct and when we are presenting energy■

■
it is our responsibility to make sure these■
kinds of things do not happen, precisely yeah■

■
yeah yeah.■

■
Now moving on to the third thing about maximizing■
what we wish to highlight on the data inc,■

■
yeah.■

■
So, again, right , very simple.■

■
I am a big believer in simplicity, right?■

■
So, I have used this as an example.■

■
If you think about it right, all the gridlines■
and the yellow colors are all what I call■

■
non-data ink, correct.■

■
It does not communicate one mile of additional■
information for me.■

■
Whereas the title that I put there called■
totals by specialty is correct that actually■

■
communicates something to me, yes yes right.■

■
So, to me that is data inc it is communicating■
something to me correct the numbers are communicating■

■
something to me right.■

■
So, any amount of ink that you use to communicate■
something is useful right um anything else■

■
is styling.■

■
Now styling there are times when you have■
you know I you know it is not about oh there■

■
should be no styling understand this even■

the table that I have on the right there is■

■

some amount of styling that is being done,■
yeah.■

■

But even that styling is also very specific■
for a purpose again going back to purpose.■

■

So, if you go back and say what is the purpose■
of putting those cells in a yellow background,■

■

yeah there is nothing.■

■

Then so again it tells back to the same thing■
what is the purpose of that.■

■

So, if I go back and say what is the purpose■
of the title it has a very clear purpose it■

■

tells me what this means is what uh.■

■

So, again going back see and you know also■
if you look at it that is the difference between■

■

purpose and message, yeah, correct correct■
correct.■

■

So, the purpose is it just tells me that it■
is caused by speciality.■

■

The message might be that PCP and PT dermatologists■
are the highest, yeah yeah correct correct■

■

correct.■

■

So, the difference between a purpose and devices.■

■

So, anything that you do any color ink that■
you use that does not serve a purpose and■

■

communicate information to the end user is■
non-data ink, right and again you cannot eliminate■

■

it.■

■

So, you want to one call the lines that we■
have the horizontal and the vertical lines■

■

to be non-data right.■

■

So, that is why I said; minimize your non-data■

ink and maximize your data ink and you cannot

eliminate it correctly.

And the last point about annotating the data is essentially to help the users is that correct

yes.

So, again right, very simple, just look at the two graphs.

What the left graph does is it gives you an axis and then you know somebody has got to

hold a finger or a ruler across this to measure and read off what those bars are, yeah yeah.

It is simpler to get rid of the access tick marks but annotate the data.

Now it is much easier just visually just looks so much easier to read, correct, correct.

So, much easier on the right hand side, yes.

So, uh.

So, annotate the data right now and people will see this as the graphs get more and more

crowded and you have more and more data points.

Annotation tends to become a bit sticky because it starts looking very messy, right, yeah

yeah.

So, fair enough, fair enough right.

So, again you go back and annotate critical points you do not have to annotate every single

data point okay things like that.

So, you know, let us say you have 300 points on the x-axis and then you know, can you imagine

a graph which has 300 little labels like that?

It just becomes noise.

■
Right for annotation it is important but it
is judicious and selective on the annotation,■

■
precisely I mean yeah.■

■
Simply because it is listed as one of the
important principles, do not blindly use it,■

■
it has been our consistent message, yes, correct
okay.■

■
So, these are four principles and my submission
is that if you follow these principles right.■

■
Let us put it this way you will not be wrong
in what you're presenting, right, yeah.■

■
So, I would say this as an actual put it and
necessary but not sufficient condition for■

■
optimal representation of data, yeah, correct,■
right right yeah.■

■
In words that I will appeal to the audience
of this class right these principles are necessary■

■
but not sufficient, correct correct correct
correct.■

■
Now let us focus on the process of designing
the visualization techniques and tools right.■

■
So, you have a three-step process: first of
all trying to understand the message, then■

■
choosing a particular form and then designing
that particular tool right.■

■
So, so again this here also when you say define
the message what are you trying to communicate■

■
right and again there is a purpose what is
the purpose of this thing.■

■
So, if you are looking at it as a single visual
component right or a sub element of a graph■

■
then you can think in terms of a purpose or
it could be or what is the message I am trying■
■

to communicate with this chart correct uh.■

■
So, there is a bit of a back and forth between■
a message and a purpose, the distinction I■

■
made but the concept is the same right um■
what is the message right.■

■
And how do I ensure that when the person reading■
sees the graph he or she gets the same message■

■
that I am trying to communicate, yes that■
is very important.■

■
That is the key right that yeah what is being■
transmitted has to be received it has to be■

■
what is received right no loss in translation.■

■
Yes no loss in translation checksum should■
be there yes uh.■

■
So, that is fun and that is the key right.■

■
So, if I am trying to say something to somebody■
because you are not always going to be standing■

■
around trying to gain an ability to explain■
that graph.■

■
Somebody is going to pull up that graph or■
chart or whatever it is without the benefit■

■
of you having standing next to them, yeah■
correct.■

■
Which means that it is your responsibility■
to ensure that anybody or random let us say■

■
somebody has printed that and you see that■
graph when you buy peanuts on the beach on■

■
the paper that is that when you open it up■
you should be able to understand what the■

■
graph is, exactly, yeah yeah yeah.■

■
So, that is important right, what am I trying■
to communicate?■

■
How do I make that message clear?■

■
And at a glance you cannot come back and say ■
listen if you spend 25 minutes looking through ■

■
the graph and looking at the axis and this ■
and that and this and you will also get the ■

■
same message, yeah. ■

■
You should not have to use the magnifying ■
glass to decipher the message exactly correctly. ■

■
So, no fine print right yes. ■

■
Then I always talk about sometimes the best ■
way to communicate is just write a paragraph, ■

■
yeah yeah. ■

■
If that is the best way to get the message ■
across please do so. ■

■
There is nothing that says that you have to ■
show it graphically, correct, yeah. ■

■
Sometimes you know I use diagrams like a flow ■
diagram or a you know workflow diagram and ■

■
you know this is you know especially in factories ■
you will always know how material flows and ■

■
you know um or sometimes it would be a graph ■
sometimes it could be a tabular display. ■

■
So, you have to choose the form of the visual ■
display that is best aligned to the message ■

■
you are trying to communicate, yeah yeah. ■

■
That linkage is important: you have a message ■
and what is the visual representative form ■

■
that unambiguously conveys the same message ■
that I want to convey, yeah that is critical, ■

■
correct. ■

■
And then there is a set of design principles ■
correct. ■

■
And these design principles come into play ■

to ensure that visual cognitive cognition■

is correct.■

Yes, do not confuse you and you know I will■
spend a lot of time on that as well because■

I think that is where people also go terribly■
wrong at times.■

Choosing the right form in my opinion is actually■
much easier because once you get a clear message■

the form almost suggests itself, yeah yeah■
yeah.■

But the design principle is what makes a difference.■

So, I am equal when I say that right I mean■
you know I am just trying to draw parallel■

to see what people might understand.■

I would go to the equivalent or you drive■
it through the offside because it is a nice■

half volley outside the half stump.■

The short that you play is natural; you do■
not try to hook that ball.■

You can drive it through half the side.■

Because it is a nice half volley through the■
half side so often marked once you get the■

message the form suggests itself naturally,■
correct, yes.■

Unless you are MS Dhoni you do not try to■
helicopter shot the ball over midwicket correctly■

if it is if the ball is outside the off stream■
you will hit it through the half side through■

the offside by and large it suggests itself,■
right uh.■

So, that is what we may say you know when■
you get the right message the form kind of■

suggests itself.■

■
But the design principles make a difference■
right and that is the difference between Rahul■

■
Marathe or Anand driving the ball through■
the covers and Virat Kohli driving the ball■

■
through the covers, yeah right yeah yeah and■
that is where the design principles come,■

■
right.■

■
People will pay money to see Virat Kohli do■
this and you know you and I will.■

■
I do not think people will watch it even if■
we pay money to do it.■

■
So, I think that is where the difference comes■
correct.■

■
So, we will just go through a few of those■
things exactly.■

■
So, let us focus on these three things a little■
bit more right, yeah okay.■

■
I think we have a few examples that we will■
talk about right precisely precisely, yes.■

■
This is okay this is where we can actually■
end this session and move on to the next session.■

■
■
So, let me start the second session, ■
welcoming back the audience to this ■

■
visualization session. In the last session, ■
we had focused on the three-step process of ■

■
having a message is important. Once we have the ■
message loud and clear the form of visualization ■

■
becomes important and how do you design that ■
form was what was discussed in the last session. ■

■
Let us take examples of these three parts of ■
the visualization process in this session.■

■
■

Starting with defining the message as we have ■
said message is the most important starting point ■

■
unless the message is not clear how are you going ■
to design the visualization for the data. ■

■
So, why do you think the message is important. ■
So, again right now what is the purpose? ■

■
you know we do not sit here and create graph ■
charts for fun. we do it for a reason, now ■

■
a simple example So, I and on the right what you ■
see is a tabular this one where we have shown the ■

■
number of people like various age ■
buckets and the number of aids cases. ■

■
So, now what is it that you want to show. Now ■
my display will this one if I look at this, ■

■
I can look at several things that messages that ■
I might want to communicate with the same data. ■

■
I might want to come back and say between ■
35 to 39 that is the single largest ■

■
age bucket that we have aids cases and you can ■
see that in the data or I might come back and say ■

■
you know what between 30 to 39 that represents ■
the two largest buckets that are quite a lot.■

■
No, but I might also want to come back and ■
say do you know what there are close to 7000 ■

■
cases of children under five years ■
now to me depending on my objective ■

■
that might be what I want to emphasize. Do you ■
realize that there are kids under five with no ■

■
fault of theirs and there are 7000 and if ■
you want to say take it up to 12 years old ■

■
there are 10000 children under the age ■
of 12 who are suffering from AIDS?■

■

■
And these are children who have contracted ■
AIDS for no fault of theirs. So, in some ways, ■

■
I might be looking to emphasize that in some cases ■
I might be embarrassing just a statistically that ■

■
this is the most at-risk group, in the 25 to 39 ■
range whatever it may be. So, depending on what ■

■
is message what do I want to communicate here, and ■
accordingly, you will construct the display. ■

■
■
So, now the message is once the message is ■
understood choosing a form becomes apparent.■

■
■
And so, for instance, if I were to look at ■
it the thing that I will show is the one on ■

■
the right side which is a very simple columnar ■
display because it tells you very clearly where ■

■
the peak is. The second thing is also the ■
point I emphasized about using a line graph ■

■
for categorical data it does it. So, now ■
the data is categorical your bug indeed. ■

■
■
So, it does not make sense to make a line ■
graph. So, here is what I want to show. So, ■

■
that is what I meant by saying once I get ■
the data and I communicate this then the ■

■
form dictionary form becomes very self-evident ■
because it is a combination of the message the ■

■
type of data and when you put those two ■
together it is very self-evident that ■

■
what type of display you want to choose.■
And then, of course, I have applied all the ■

■
principles that I spoke about but because we are ■
building up those principles, for instance, I ■

■
have not annotated the data here There are things ■
like that but this is done to illustrate the point ■

■
and in fact, there are a few other things that I ■
will talk about for instance even in this display ■

■
If you look at the age buckets, I am not ■
a big fan of those vertical spacing.■

■
I mean because what you expect me to do is ■
lie down every time, I have to read this graph ■

■
tilt back or tilt the laptop up and down and ■
in fact it makes it worse if it is on a mobile ■

■
device when I tilt it, it auto rotates. So, that ■
does not help me either. So, there are others ■

■
and that will come in to address those points ■
when we talk about the design elements. ■

■
So, there are some recommended best ■
practices when we choose a form. ■

■
So, one of the things again see look at ■
this, I have sales numbers for two years ■

■
Jan to December for two weeks. Now the message I ■
want is that you know there does not seem to be ■

■
any kind of seasonality. That ■
is the message I want to show. ■

■
And the secondary message might be that the trend ■
is all wrong in 2004 the trend is not right.■

■
Now if I throw it in a tabular display like ■
this that message just does not come across, ■

■
I mean if I show the table, it is impossible to ■
show that you know maybe at some level a trend ■

■
might be visible but if my purpose message is sure ■
that there is no seasonality. The tabular display ■

■
does not show it to me it does not help. But ■
when I show it as a graphical display like that ■

■
the lack of seasonality shows up very ■
clearly that there is no seasoning.■

■
■
The trend is somewhat visible in the tabular ■
display correct but it takes some effort, 2004 ■

■
trend is probably a little bit more evident ■
than the 2003 trend because if you look at ■

■
2004 if you can top to bottom you can see it is ■
monotonically reducing. 2003 is not that way. So, ■

■
it is kind of goes up goes down and then if ■
you look towards the end, it flattens out. ■

■
■
So, the trend is not very clearly evident in the ■
2003 trend, again if I want to show that there is ■

■
no seasonality very clearly the right display is ■
big. Even for trend-right is preferred because if ■

■
one goes and looks at how many seconds it took for ■
you to understand that 2004 there is a downward ■

■
trend 2003 was an upward trend by looking at the ■
right graph and looking at the left table. ■

■
■
Then you will understand what ■
I mean by saying the form is ■

■
you know the whole idea is how do you communicate ■
that quickly without any loss in transmission. ■

■
But I mean are there some general rules about ■
when to use a table and when to use a chart ■

■
there are some general rules. ■
There are correct so for instance ■

■
when you want to display a complete ■
data set now it may not be practical ■

■
but if you want to show this one. Where we are ■
just showing the data, we may not necessarily ■

■
be inferring something from that or I might see ■
especially if you are looking at the customer ■

■
service MPS course you might just want to show ■
the kind of complete data in a table, right. ■

■
■
You might want to show if you want to highlight ■
a specific row if you want to show like all the ■
■
product categories and you just want to show them, ■
I want to factually tell you what is the sales of ■
■
every product category. I am not trying to tell ■
you that a is bigger than b or b is bigger than ■
■
c. I am just trying to factually communicate ■
that by stating these are my sales numbers.■

■
■
I am not trying to tell you that Tamil Nadu did ■
more sales in Karnataka which did more sales in ■
■
Maharashtra I am just trying to tell you that ■
Tamil Nadu hit 20 million in sales and Karnataka ■
■
24 million in sales tabular display or if ■
you wanted to say look focus here focus ■
■

there you can highlight a certain row again ■
emphasis. Now sometimes you will want to show ■
■
how the numbers are calculated. ■
So, you might say look the way it was calculated ■
■
was revenue cost margins. So, sometimes when ■
you show that it might be easier to show it than ■
■
a tabular display where you can see how it ■
is calculated. And it happens often right ■
■
you will just say revenues cost margins ■
or what you call wrong metrics right. ■
■

■
■
You know we will call it to drum matrix revenue ■
units and markets which helps this one chart ■
■
will come exactly like when you want to show ■
patterns, we want to show a change over time. ■
■
So, for some cause and effect occasionally ■
you have to be very careful about that ■
■

I only want to compare two things. Comparison is ■
probably the most common use case for charts.■

I mean essentially a pattern is effectively a meta comparison,

I can compare two things when I show that same comparison multiple objects show a pattern

or multiple, and when you show that over time it shows it becomes a time series. So, at some level,

these are all kinds of specific instances of the same concept. So, visual comparison.

So, data tables are used for two things one is just factually communicating the numbers without

drawing inference. I want to show specific items where I will show this and highlight one

correct and if you want to show visibility into the calculations

okay you know what a b c this is how the margins stack up etcetera.

So, that you can compare multiple things together right, and sometimes

just if you have equal enough you have a wide number of columns and a whole number of rows.

A graphical chart display might become overly complicated in which case the simplicity mandates

that you just show it as a table and that is also important. So, for instance, if I have

20 let us say I have sales I have something and I want to show by state but then I also

want to show unit sold revenue made margins made percentage margin what is the profit

how many sales agents were there how many outlets were there and if you want to

show 20 different metrics for those trying to show that in one graph will be impossible.

■
■
You are better off showing it as a tabular display ■
than going out and taking out individual metrics ■

■
and showing graphs to show some comparisons.■

So, there are some quick tips about when to ■

■
use a table and when to use a chart right ■

that is what we were talking about just now. ■

■
So, when you are as you are saying when ■

you have just highlighted one component. ■

■
So, if I have multiple things and I ■

just want to highlight one component ■

■
of the pie chart now I am going to put ■

a huge rider on pie charts right in ■

■
fact while I say this, I generally also say ■

friends do not let friends use pickups. ■

■
■
Pie charts are very difficult to interpret and ■

probably the most grossly misinterpreted types of ■

■
charts that you will see there. So, ■

okay the problem with pie charts is ■

■
seen when I want to show that one ■

is largely dominant over the others. ■

■
If you look at the blue light blue and dark blue, ■

can you tell me which is bigger and by how much? ■

■
Difficult little difficult yes, it is ■

exactly the reason why pie charts are;■

■
■
So perceptual right remember what I said about ■

we are very good at perceiving differences in ■

■
length and relative differences in length. ■

We are not very good at perceiving relative ■

■
differences in the area especially the shapes are ■

different if I draw a rectangle under the square ■

■
and I tell you which covers ■

are bigger and a circle and I ■

■
ask you to tell me which of these ■
covers a bigger surface area?■

■
■
Difficult very difficult and really with a ■
pie chart that is what you are asking people ■

■
to do you are asking them to compare two ■
different areas of slightly different shape ■

■
and trying to determine which is the higher ■
which has a higher area. The same pie chart ■

■
could be equivalently shown as a column chart. ■
And the message comes across very clearly. ■

■
■
So, the question is why ■
would you choose a pie chart ■

■
if the emphasis is to show the percentage ■
breakdown where it all adds up to a hundred ■

■
percent that I can you can make an argument ■
for saying pie chart. But even there ■

■
if you have two of them that are very comparable ■
right will you cannot make out the distinction ■

■
area-wise unless you annotate it.■
Exactly I was going to say annotation ■

■
might help there correct. So, that is why I said. ■
So, pie charts to me it is a chart of last resort ■

■
right. So, which is why I say right friends do ■
not let friends use my charts. So, that is a very ■

■
important right, and similarly, if you want to ■
show components of multiple items you know you use ■

■
a stacked column chart or something like that.■
Now if you look at actually an individual column ■

■
of the stack column chart that could be a pie ■
chart effectively, So, you could make a pie chart ■

■
and make it a simple column chart or a single ■
column of the stacked component. Again, ■

■

perceptually which is easier to understand, ■
yeah but here in a column chart also ■

■
comparing two orange regions and deciding which ■
one is bigger might end up becoming difficult. ■

■
So, and that is where you will also look at ■
your choice of what are the categories on the ■

■
x-axis and what are the categories that are on the ■
stacked columns correct also makes a difference. ■

■
So, and in fact, if you notice those guidelines ■
that I have drawn on those between the orange ■

■
that is precisely to address your point. ■
Because you can easily see remember when we ■

■
spoke about the components of perception where ■
we spoke about thickness length orientation.■

■
You would see the base is the orientation whether ■
it is getting thinner or thicker So, a simple ■

■
line that draws these will show you whether the ■
orange component is reducing in contribution, ■

■
So, here you will very clearly see from ■
the second column to the third column ■

■
orange has held its blue has increased ■
primarily at the expense of pink.■

■
Whereas that is very clear from ■
those lines that you have drawn ■

■
and between the first and the second one is ■
very clear that blue has gained a percentage ■

■
at the expense of orange. A very simple line ■
kind of addresses those things yeah that helps ■

■
small things. Now and you know look I mean instead ■
of not getting through individual this one.■

■
But remember when I spoke about once you ■
get the message the form is very evident ■

■
 that is the whole point if your message ■
 stresses this it is very evident. So, ■
 ■
 if your message is stressing correlation right you ■
 will do a scatter plot exactly. If your message is ■
 ■
 talking about change over time you will do ■
 a line chart you want to do an internet ■
 ■
 ■
 So that is what I meant by saying ■
 you know when you get a half volume ■
 ■
 outside the off stomp you are not ■
 going to try and do a hook shot ■
 ■
 true. So, to me just. So, this is just a quick ■
 recorder no hard and fast rule but broadly ■
 ■
 speaking that is how you think about it correct, ■
 and here are some examples for that right. ■
 ■
 ■
 So, for example, if you want to say ■
 sales of A bigger than sales of B ■
 ■
 so again this is one more thing where ■
 people might show it as a pie chart, ■
 ■
 that does not come across that ■
 you know A exceeds C by almost 2x ■
 ■
 yes whereas here it does come across yes once ■
 again yeah principle that the height is easily ■
 ■
 perceptible than the area. that is so critical So, ■
 very simple I will go to the next one right. ■
 ■
 ■
 If my message is that there is no apparent ■
 relationship between and there is no correlation ■
 ■
 correct a simple scatter plot with a regression ■
 line that tells you that you know effectively ■
 ■
 there is no correlation. But a pie chart ■
 does have its place right for example ■
 ■
 if you want to show the contribution. ■
 So, that is why I said but notice that I have ■

taken the trouble to annotate it very carefully which is critical. Without annotation
 honestly speaking it will be able to tell between C and B what is the relative strength you cannot
 difficult. So, when using if you do have to use a pie chart or if you do choose to use a pie chart
 correct make sure that annotations are there. But also remember this right the message here
 is that our competitor A has the competitor D has the smallest share. That pie chart is okay.
 If I want to do a comparison between A B C and D and show their relative strengths pie
 chart may not be appropriate but you may not the message is that that guy is the smallest
 no problem but if you want to come back and say you know if you combine D and C
 where they have the same market share as B okay. This will fail miserably,
 and the time series has always been classic. So, over some time correct and just again very
 clearly, I just want to show that look we have a problem EPS has been steadily declining you know
 it is time for a new CEO that is what this is saying correct
 I mean you can say all that if you just say look EPS has been declining the CEO's got to go
 right you are a troublemaker. You show this graph then everybody says
 you know what you are not a troublemaker you are a whistleblower.
 And the last one is essentially an example of a box plot right where clearly those dots 214
 observation number 213 observation number shows

that those are outside the tail and probably, they ■
■
represent the outliers in the data exactly. So, ■
you can talk of I mean I cannot emphasize enough ■
■
the concept of this outliers right.■
Because as you enter the workforce right ■
■
the biggest thing you will see is that people ■
will pick up an outlier and make a whole case ■
■
as if that is a regular event. So, you know and ■
they will call it oh I have anecdotal information ■
■
I mean that is the buzzword right and this is ■
what they teach you at all the MBA schools, ■
■
anecdotal information right.■
Draw me a box plot and show me ■
■
where it sticks and if your anecdote sits within ■
the big yellow bar I will talk if it is sitting ■
■
in that thing, I do not need to talk to you about ■
that exactly correct. So, it is very critical I ■
■
mean just to show that hey you know what this is ■
what it is. So, because they will come back and ■
■
say I spoke to this one customer and this one ■
customer had this kind of feedback and hence we ■
■
should change our entire product line.■
Are you bonkers, ■
■
if that were the case right ■
because when I talk to my daughter, ■
■
I will have to rebuild the entire Akasa ■
model and introduce business class seating, a ■
■
very nice example is very nice because she ■
just looked at it and said oh no business ■
■
class of this place? So, you know and if I ■
take that one example without understanding ■
■
where it sits on the outlier chart then you know ■
I will have a whole fleet of business classes.■
■
■

Fortunately, that is not going to make too much money in India probably, yes on that front Rahul

I think I have eliminated the probability we are very clear that is not the way to go into

the Indian aviation market and the data clearly shows that right. So, all right okay. So, we have

knocked down two out of the three components of the process of designing the visualization

for the particular data type.

First, we discussed the importance of messaging,

and then we spoke about various forms that immediately come from the message right.

So, we still have to discuss designing the form but let us do that in the next session. So,

let us end the session here having knocked down 3 out of 2 out of the three components

of the design. Design of visualization process. So, let us end the second session here.

Welcome back folks this is the third session on visualization.

We were speaking about the process of visualization and the two important components were first

of all trying to understand what message do we want to convey.

And once the message is clear what form of visualization should be used and Anand is

very happy to give the example of the cricket short where if the ball is outside the stumps

no chance of, we looking it on a deep square leg.

So, let us come to the third component which is designing the form that we have selected

in the second phase of the visualization part over to Anand.

■
Why designing important?■

■
Well, designing is important because you have■
got a good message you got a good form correct■

■
and the design can really completely detail■
the community message that you are trying■

■
to communicate.■

■
Almost a personal favourite is using 3D representation■
right absolutely 0 value to the message and■

■
that is being communicated correctly somebody■
decided.■

■
So, in fact, it takes it away right because■
if you look at those the bars that you see■

■
of the 25 you know which is that the two dollars■
bars which are 30 to 34 and 35 to 39, right.■

■
It is much more difficult to tell whether■
there is a difference between those two when■

■
you use 3D effects.■

■
Because of the shadow, honestly speaking my■
submission is again going back to purpose.■

■
What purpose does that 3D representation do■
whatsoever?■

■
In this case, at least right it is a harmless■
I will for your cases where you can be co■

■
counterproductive.■

■
I do not know if you have an example of a■
3D pie chart I have it on the next slide.■

■
So, I and again three same things I 3D effects■
that look this is very still same as the data■

■
is you know the same as what we saw the previous■
chart.■

■
But then we have got this 3D pie chart.■

■
If you look at it right if you look at let■

me, see the numbers a little bit right.■

■

If you look at our company which has 34 and■
the competitor here with us, 26 on the right■

■

side can pie chart on the left side it looks■
like our company and competitors have the■

■

same share.■

■

All was no, yes but look at the chart on the■
right side.■

■

yes, correct and if I wanted to do this, I■
will switch it around and rotate this pie■

■

chart.■

■

So, that competitor D which is a yellow slice■
can appear to be the same size as us.■

■

So, 3D effects have zero purposes and in some■
cases are bad this to me goes into almost■

■

deceit.■

■

Because the left side pie chart without annotations■
essentially says that our company and competitor■

■

have the same share, yes.■

■

So, that is why 3D effects are very dangerous,■
and at least you know when I used to review■

■

presentations, I would stop people whenever■
is a came in with a 3D chart I know I do not■

■

know why I do not know what reason that is■
why they got the message.■

■

I mean they are using 3D because 3D charts■
are so common everywhere today.■

■

So, I have a simple movie I will not look■
at anybody who comes into a 3D chart and it■

■

took me a while to get the message across.■

■

And I had to use multiple tools to get the■
message across but the message was and I would■

■

say you know you should never be used.■

■

And if you ask me get this particular slide■
and stick it up on your wall because this■

■

will tell you exactly why you should not use■
3D effects.■

■

And the same thing goes for even the line■
charts right we put too many things often■

■

there.■

■

Now different perspective right again you■
are showing monthly this one moreover the■

■

same we have seen this example before.■

■

We are talking about on the left side multiple■
things here right.■

■

Look at the title let us start with the title,■
right?■

■

it just says your cumulative unit sales actual■
versus plan.■

■

That is a statement of fact right.■

■

But look at the second one where I put this■
right.■

■

Sales have exceeded plan by 30000 the message■
is very clear very clear on the left side■

■

I have to calculate and not only do I have■
to calculate I have to look at the horizontal■

■

the red line show me and this one is about■
hundred extra hundred more what is the difference.■

■

So, you know what is it that it will be 300■
and 177 correct.■

■

So, you know what I said you know we said■
up for value labels wherever possible and■

■

you can delete some labels to avoid clutter.■

■

So, if you notice I have not labelled every■
single data point yes, I have to open to label■

■

every alternate one, yes.■

■

Why, it reduces clutter, and also, and since■
I have allocated that I do not need those■

■

horizontal grid lines once you annotate it■
you do not need those correct.■

■

So, again just look at aesthetically simpler■
right it is much easier more distraction right.■

■

It is the equivalent of listening to Beethoven■
with you know background white noise with■

■

some guy doing construction work in your house.■

■

I must comment that your analogies are as■
visually appealing as, I mean as they can■

■

have no I mean that is the point-like how■
to get the point across how to get the message■

■

across.■

■

So, the chart is Beethoven here on those red■
lines are the construction noise.■

■

And it is like correct but I mean in general■
by designing phase of the visualization we■

■

are trying to say that well may make it easier■
to guys exactly.■

■

But I am also calling out to see if even the■
title the key message I want to say is that■

■

you know cumulative actual sales through July■
we are ahead of the plan by 30000 I mean so,■

■

very precise and it is got the message across■
I see this.■

■

So, somebody says fantastic us a good job■
online.■

■

I mean it says I mean you want to say things■
that you want to draw attention towards right■

■

otherwise you are just keeping it open for■
discussion.■

■
So, look for this one I am going to do right?■
■
you know doctor colour that is why the director■
we should do it when I was in the department■
■
and he was ahead of the department with chemical■
legendary.■
■
He had a very interesting self-way of saying■
things right he says look if I show you an■
■
apple and you tell me it is an orange you■
are confessing to two kinds you do not know■
■
an apple and you do not know an orange understand■
principles, right?■
■
So, if you want to get something of course■
do not distract on other things because by■
■
definition people will draw their attention■
will be drawn to that and they will focus■
■
on that.■
■
So, you have to be very precise about what■
you what do you have to show.■
■
So, you know if you draw a different real■
accuracy and position when you visualize you■
■
have to be both accurate and precise.■
■
So, that completes our three-part process■
of designing visualization tools understanding■
■
the message conveying that message through■
a particular tool, and then designing that■
■
tool carefully right.■
■
I have a curiosity question for Anand.■
■
So, essentially all these charts and tables■
and these numbers get into what is called■
■
as data dashboards.■
■
So, what are our data dashboard?■
■

Dashboard right I mean and this is something
that is there in every single I mean everybody

wants to talk about dashboard.

In fact, I remember seeing I think was it
a department strip or something like that

right when you had a serial executive saying
you know and given that is coming up on Christmas

gift right.

So, an executive is writing a letter to Santa
saying dear Santa all I want for this Christmas

is a nice dashboard due from finance already
has one.

So, look I the whole idea of a dashboard is
hey how do you this photo of individual graphs

individual display self.

A dashboard in my mind the system ah a collection
of related displays with some purpose right.

And you know no problem I think this was this
I love this because it was created by, I think

Steven Fue was considered to be, you know
extremely good researcher and a well-known

authority on dashboard design right.

Dashboard definition is you know it is a visual
display and very nicely where every word is

important here.

So, a visual display of the most important
information is needed to achieve one or more

objectives.

Consolidated to a single screen, so it can
be monitored and understood at a glance, every

word here becomes critical.

At least the ones where I reduce the font

those you can encode it is a visual display■

■

end of the day correct what does it mean by■

visual displays that I see it and I comprehend■

■

correctly.■

■

Why do I say it is a visual display?■

■

Very often more often than not you will hear■

this thing called the interactive dashboard■

■

where you will see a dashboard and then people■

will say you can click here and we can show■

■

this and you can show that.■

■

No, I mean it is supposed to be visual right■

you do not necessarily have to make it interactive.■

■

I think you only show the most important information■

that is needed to achieve a particular object.■

■

I need the CEO to understand where we stand■

from a revenue perspective or a business performance■

■

perspective.■

■

It has to be on a single screen correct do■

not want to scroll up and down.■

■

And very easily more understood I look at■

it I know what happened the in fact very name■

■

dashboard was borrowed from the car automobile■

dashboard.■

■

I already said when you design a dashboard■

you think of the car dashboard.■

■

Can you imagine if you are driving a car and■

the dashboard requires you to reach in and■

■

push a button to look at the fuel consumption,■

push another button to understand what your■

■

speed is?■

■

I mean that speed display tells you that you■

know you are 32% over the other speed limit■

■

and that you know 78% of the and below your speed is that all relevant.

And also, something that says that 14% of people in the age group 18 to 22 like this

student on Facebook is that relevant.

What I need to know is how fast am I driving it right.

So, again it only shows me how fast am I driving.

It does not show me anything for me anything which is not needed correct.

I do not have to do anything to get that information I can look at it at the answer I can look

at it at the corner of my eye and I will get the information.

And it is all there on one screen.

So, to me and somebody defines a dashboard this principle has to be critical.

If you have to do all of these activities and you are throwing all this stuff onto a

dashboard and then you are going to crash the car.

The same way you put that on a business dashboard is useless and my personal favourite can be

a real-time dashboard.

A real-time cannot be a dashboard is meant where you look at it you glance at it you

get information and then you go do what you have to do.

A real-time is something that you are monitoring.

If there is a reason to monitor something real-time, we should not be a dashboard it

should be an entity.

So, so these are all important things to consider when you design a dashboard.

To me, I just absolutely love this definition right because it hits every single aspect

of what a dashboard should be and what it should not be right.

It should be at the number of times I have seen dashboard design with a tabular display

where you have to scroll.

I mean scroll some tell me this whole idea of the dashboard is so that you know I can

put multiple matters on display at the same time.

So, that I can say this is going down understood.

Now if you put tabs then by definition, they are not on the same page yes if you have to

scroll by the deficient amount on the same page and then what was the intent of food?

So, and this is how one trap people fall into when designing dashboards right essentially

is this.

You will design a dashboard and it starts off with a nice simple concept like this.

Then somebody will come back and say hey why do not we draw our social media data on top

of that?

why do not we draw our HR headcount data top on that, why do not we draw that also right

then they need to get called an executive dashboard and you know what they feel proud

of it for the fact that every carrier is this plan on one dashboard.

It does not make sense at all correct.

■
So, the director of IIT, Madras wants to see the dashboard.■

■
He wants to see how many students are enrolled how many faculty are there how many support■

■
staff what is your cost running operating cost incidents etcetera.■

■
He does not need to know what the GPA of the distribution of GPA of students was and you■

■
know how many students ah you know were absent for a particular class.■

■
He just needs to know all that.■

■
So, you have to think about those things and not try to overturn them with information.■

■
And then so they can do that correct.■

■
So, are you suggesting that before we design a dashboard there has to be some kind of a■

■
Pareto analysis saying figure out what is most?■

■
Actually, yes, I know it is not a question of Pareto analysis right because you know■

■
the problem with Pareto is earlier than you know it is a question of arbitrarily where■

■
it draws a line.■

■
You have to go back and say what is the purpose of this dashboard yes and what do you need■

■
to put there that helps the user of the dashboard achieve that precisely yes nothing else.■

■
It is not a Pareto.■

■
So, let us say the purpose of the dashboard is I just want to look at one metric and one■

■
metric only then only that does not show anything else.■

■

So, what would be the basic design principle of a dashboard?

So, which means I do not know it is very similar to the same thing right the dashboard I mean

even while I have a picture here even this is a very cluttered dashboard in my opinion.

But it helps highlight the point.

So, what we will say is that you will show the big picture.

So, there are a glad you will see certain critical things and you know there is some

color-coded red yellow green whatever you want to have that to draw attention.

You know use colours you can zoom in on specific correct and then from there, you will provide

a link to say here you know if you want to see more details you click here and you will

it will take elsewhere to look at the other details.

So, do not cramp other details into the dashboard.

That should be simple.

So, which means I see it I should see the big picture, this one gets a sense of the

metrics in some nice suitable display.

If you want to sort of revenue you may want to just show only the revenue or sometimes

you will say look let me show you the trend of the revenue as well I think because the

trend is also important.

No problem but I can zoom in I can expand it and then I can click to get into supporting

detail.

Now, why is this happening?■

■

Why is that happening?■

■

And many other dashboards support that.■

■

Now that is in this case it might be a data■
warehouse monitoring dashboard.■

■

So, you can click that I say I see something■
wrong with the data warehouse system I will■

■

click that I go to the data warehouse dashboard■
correct.■

■

I see something wrong with the you know website■
I will go up and look at the website having■

■

the dashboard.■

■

So, yes there are long as big questions asked?■

■

Whenever somebody looks at a dashboard question■
should be raised correct one is what happens.■

■

But the solution to those answering those■
questions is completely separate dashboard■

■

only those questions I have heard people say■
that right when they say I understand that■

■

a glance or you should answer all the potential■
questions that have been covered?■

■

No, it should not.■

■

You should have additional information additional■
link additional accessible information that■

■

answers those questions but that is separate■
if you called that separately.■

■

You are trying to cram all that into the dashboard■
essentially you get you would mess all over■

■

the place like no you cannot drive the car■
you are only you are spending all your time■

■

reading the dashboard.■

■

That is what happens.■

■
So, that was the basic understanding but essentially■
our dashboards are you suggesting that there■
■
should be a separate dashboard for finance■
and there should be a separate dashboard for■
■
marketing but you are saying we are not saying■
that at all we are saying find out the purpose■
■
design in dashboard correct.■
■
So, typically yes for instance since you brought■
a finance marketing, that is a CEO dashboard■
■
that we come up with which will give some■
key metrics.■
■
Now there might be one or two critical metrics■
or that will show over things that will show■
■
up for representing finance and marketing.■
■
The details so far as I might want to say■
how much do I spend on marketing this month.■
■
The CEO dashboard I might not say how much■
was spent in Tamil Nadu versus Karnataka versus■
■
Kerala it is a separate market and if you■
say look looks like your spend of monthly■
■
spend this increase in your marketing why■
is that happening?■
■
You may have a separate dashboard look over■
the last few months we will increase a lot■
■
of spending in Kerala in South India which■
is why it is increasing.■
■
We have launched a new product and that is■
why we are spending more you know one of the■
■
things I should also emphasize video talking■
about marketing and this one and more correct.■
■
So, purpose and more or less one more thing■
I should emphasize talking about marketing■
■
and this one.■

■
When you design especially when you use colours■
correct accessibility becomes a big issue.■

■
Why so, for instance typically it is very■
common to use red, green, and yellow to highlight■

■
metrics that are good, bad, or need monitoring■
correctly.■

■
What happens if the person reading the dashboard■
happens, we can?■

■
So, then your red-yellow-green and if they■
have red-green colour blindness you lost the■

■
fundamental premise of why you are highlighting■
it red in green.■

■
So, there has to be designed for accessibility■
and red, yellow, green is universally accepted■

■
for good balance right.■

■
So, but then you have to have the ability■
to be able to customize it or at least end-user■

■
to come back and say look I have a certain■
visual impairment and it should switch over■

■
to 3 Different colours that are easily distinguishable■
but at the same time even for somebody who■

■
I colour blind.■

■
And in fact, when we got a lot of websites,■
they give you colour palettes that ah while■

■
they may not appear to be the same colours■
as you see it to a colourblind person they■

■
will appear they are guaranteed to appear■
as distinct colours so for different types■

■
of colour blindness.■

■
So, sometimes you know choosing the palette■
is also important right.■

■
And then I saw him.■

■

So, when I saw this red that is the first
thing that struck me you know what if the

person here, I am looking at it when colour
blind they will not know to understand between

which system is red and which system is green.

And you know as fundamental and communication
error that could be right correct.

So, you are right.

So, to the point right that you have to have
these kinds of high level and then break down

dashboards and then you know further breakdowns
and things like that.

And you can always have them on link interlink
where if a question comes up you can always

pull up the other dashboard and then dig.

But do not try to bring everything into the
same dashboard, correct.

Thank you Annad thank you for your insights
that does help in understanding why visualization

is important and once you have understood
the purpose how to go about doing it?

So, thanks a lot, and hopefully this session
is helpful.

So, thank you, now always a pleasure for her
anytime happy to help thank you.

So, let us end the session here.

Hi this is the second session of the business
analytics course and we are going to discuss

probability distributions.

Most importantly we are going to discuss how
are we going to fit a distribution to a given

data.

■
So, first of all let us do a recap, what are
probability distributions?■

■
We already have discussed it in other courses■
but what do we think what do we recall as■
■
probability distributions.■

■
So, essentially probability distributions■
are some kind of a statistical model that■
■
shows possible outcomes of a particular event■
or course of action that event may take.■

■
So, essentially probability distributions■
for a discrete random variable may look like■

■
all the possible values of the random variable■
along with the corresponding probabilities■
■
that the random variable will take on that■
particular value.■

■
And for a continuous distributions we generally■
represent that by a density function.■

■
For example if you recall we may have said■
that the x axis represents the values of the■

■
random variable the y axis represents the■
probability and for a discrete random variable■

■
we will say that what is the probability that■
x takes on a value equal to one and then we■

■
would have said some probability what is the■
probability that x takes on a particular value■

■
2 and we would have said some probability.■

■
So, this is how the probability distribution■
looks like for a discrete random variable.■

■
Now for a continuous random variable we still■
have the same format which essentially means■

■
that x axis still represents the value of■
the random variable y axis represents some■

■
form of probability but we do not say probability■

we usually if you recall we say density function.■

■

Density function and then we would have drawn■
something like this for potential values of■

■

the random variable x .■

■

What is the difference here in the earlier■
diagram we had discrete probability masses■

■

because the random variable was discrete.■

■

Here we have continuous values of the random■
variable and therefore we can't really say■

■

that there is a probability mass sitting at■
a particular point.■

■

For example let us say that we are still talking■
about x taking on a value equal to 2.■

■

We cannot say that this is the probability,■
this is only the probability density.■

■

So, we only talk about density and for density■
we need a small interval to actually define■

■

some probability.■

■

So, you recall all of that.■

■

So, the focus of this session is not to re-describe■
density functions and probability distributions.■

■

The focus of this session is to go one step■
beyond and say that well I have data now and■

■

how do I fit some distributions to data or■
what do I do with that data.■

■

So, for example let us say that we in academic■
settings we hear this quite a lot.■

■

So, grades of a course follow a normal distribution■
what do I mean by that.■

■

So, what do I mean by that essentially grades.■

■

So, a random variable is grades here.■

■

So, the grades out of 100.■

■
Let us say so, random variable is grades and
then it follows a normal distribution which

■
essentially means that we are going to follow
assume that this is a nice well-shaped curve

■
and then some people are going to get a very
high mark some people are going to get very

■
low marks unfortunately and there are whole
bunch of people who are going to be in between.

■
So, that is what we mean by normal distribution
once again the y axis represents the density.

■
So, this is just a recall or sometimes in
the business settings we may say something

■
like this: sales next month are expected to
be uniformly distributed.

■
So, what do we mean by that.

■
So, I may say that sales can be as low as
a hundred thousand dollars, sales can be as

■
high as two hundred thousand dollars this
is sales next month, sales in the next month

■
So, it can be hundred thousand dollars or
it can be two hundred thousand dollars but

■
instead of assuming a normal distribution.

■
So, on x axis is sales here is your 100000,
here is your 200000 and we are saying that

■
it is uniformly distributed.

■
So, you know what uniform distribution is
once again y axis represents the density.

■
So, these are essentially probability distributions
normal distribution uniform distribution.

■
We have taken two examples of continuous distribution
but you get the idea.

■
So, that's how we define probability distributions
that's how we use probability distributions.

■
So, now how are we going to go about using data.■

■
So, let us say that I have business data that I have collected, the business data may be■

■
about sales volumes the business data may be about the defaulters on loans or the business■

■
data may be the salary hikes that the employees got in a particular year.■

■
It may be about any business context for this kind of a data we can directly use the data■

■
and use it in our simulations there is no need to fit any distributions.■

■
This is typically called trace driven simulation.■

■
So, let us say that we have collected sales volume over a period of time.■

■
Let us say we have a monthly sales volume for the last three years which essentially■

■
means that I have 36 values in my data-set.■

■
So, instead of first fitting a distribution to the 36 values and then using the distribution■

■
in my further analysis I can directly use these 36 values in my analysis.■

■
So, if I want to simulate I will simulate directly using these 36 values, this is generally■

■
called trace driven simulation.■

■
The second method is to actually fit a theoretical distribution.■

■
What do you mean by theoretical distribution, theoretical distribution is all these things■

■
that we spoke about earlier normal distribution, uniform distribution, binomial distribution■

■
for discrete, Poisson distribution for discrete, exponential distribution for continuous these■

■
are all theoretical distributions.■

■
So, what we may do is for the sales volume■
data that we may have, sales volume data that■

■
I may have i may try to fit, quote unquote■
fit a distribution to my data.■

■
And obviously I cannot simply say OK normal■
distribution fits very well I have to go beyond■

■
that and I have to actually check whether■
the fit that I have assumed is actually good.■

■
And I am using these terms in a very deliberate■
way because these are precisely the technical■

■
terms which are going to be helpful later■
on.■

■
So, we always are going to say we are going■
to fit a distribution we are going to check■

■
how good is this fitment.■

■
Now let us say that our business data that■
we have collected is a particularly tricky■

■
data set and it does not fit very well with■
lot of theoretical distributions or the other■

■
way around, theoretical, most of the theoretical■
distributions do not fit to our data.■

■
What are we going to do?■

■
Well it is not the end of the world instead■
of trying to fit already available distributions■

■
like a negative binomial or a double exponential.■

■
Instead of fitting those kind of already available■
distributions to the data what you can do■

■
is we can actually create our own distributions.■

■
I mean this is like making rules as we go■
along typically Kelvin category but we create■

■
our own distributions and those distributions■

are called empirical distributions.■

■

So, the sales volume data that I already spoke■
about using that data we say that well what■

■

would be the distribution where these 36 values■
could have come from.■

■

So, using these 36 values we build our own■
empirical distribution and use that distribution■

■

in our future analysis.■

■

Now what are these empirical distributions■
have you discussed empirical distributions■

■

in your earlier courses most probably you■
have.■

■

So, let us quickly recall that.■

■

So, what are these empirical distributions?■

■

Empirical distributions are essentially distributions■
built from the data that we already have collected.■

■

We are not fitting a distribution to the data■
we are actually building a distribution from■

■

the data that we have collected please notice■
the difference.■

■

So, let us go beyond.■

■

So, how does one build a distribution?■

■

First of all what are the building blocks■
when we say we are building a distribution.■

■

How do we build a distribution for example■
normal distribution let us take simplest,■

■

normal.■

■

If we were to say that I want to characterize■
a normal distribution what would we need to■

■

characterize a normal distribution well we■
will need the building blocks.■

■

So, what are these building blocks?■

■
So, essential building blocks of any distribution are the density functions, the distribution

■
functions, and we may also want to define some moments, the first moment around the

■
mean, the second moment around the mean which can be built using density also.

■
So, we have to estimate these parameters.

■
So, essentially defining a distribution means identifying a density function or a distribution

■
function from the density function you can identify the building blocks like moments

■
around the mean.

■
Mean standard deviation and so on.

■
Let us take an example of how to build a empirical distribution.

■
So, let us say the data is ungrouped.

■
So, let us say that we have collected X_1 X_2 X_3 values.

■
So, the X_1 value, X_2 value, X_3 value and let us say all the way to X_{36} .

■
These are our 36 sales volume data for 36 months in our data set.

■
Now what we are going to do is we are going to arrange them in a ascending order.

■
So, X_1 value was the first value that was recorded which was the first month but what

■
we are going to do now is we are going to arrange it in a ascending order where the

■
smallest value is called X bracket 1, OK X bracket 1, second smallest value is called

■
 X bracket 2 and the largest value is called X bracket n in our case X bracket 36, may

not be the sales volume in the 36th month■
it is actually the maximum possible sales■

■
volume that we have found in our data set.■

■
So, these are called rank order statistics■
let us not worry about rank order statistics.■

■
So, once we have arranged the data in an ascending■
order you can actually define a distribution■

■
function in this way.■

■
This is not our own creation.■

■
These are, these definitions are usually available■
in any standard statistics textbook, all right!■

■
So, this is one way.■

■
I mean by no means we are saying that this■
is the only way of defining a distribution■

■
function.■

■
Now once we get a distribution function we■
all know how to get a density function and■

■
from density function we know how to get moments■
around the mean.■

■
This is for ungroup data.■

■
So, this is for ungroup data.■

■
Now if the data were grouped meaning that■
I only know that in this interval I have ten■

■
values in the other interval I have some eight■
values in some other interval I have some■

■
five values if I have group data.■

■
So, let us say that intervals we define k ■
intervals.■

■
So, we have intervals k such intervals and■
I know that in each interval I have some n_1 ,■

■
 n_2 , n_3 values.■

■

So, in the first interval I have n_1 values
in the second interval I have n_2 values
third
interval I have n_3 values
kth interval I have n_k values and that gives me my total sample
size of n .

So, what we can do is we can create a piecewise linear function G using this definition

where each G of a_j is essentially proportion of the samples, proportion of the observations

up to that point up to that interval.

So, once again a very non unique way of defining a distribution function.

Once again notice that this is a distribution function why do I know that this is a distribution

function because the value lesser than the smallest value is 0 and the value beyond the

highest value is 1 which is a typical definition of a distribution function which goes from

zero to one.

And once again our usual methods are going to kick in where we have a distribution function

from there we get the density function and so on.

So, these are examples of how we can build empirical distribution.

Let us go back why are we saying that why did we build these empirical distributions

in the first place.

We are saying that we have data we have collected data that data may be for any context it may

be sales for our marketing data it may be financial analysis data it may be stock price

data.

■
So, let us say that a technical analyst wants
to analyze wants to invest in the stock market.■

■
Now what are technical analysts well figure
out why do not you search for it and then■

■
we will describe it in the next sessions.■

■
So, technical analysts let us say that they
want to invest and for their investment decisions■

■
they have collected stock prices for the last
three months.■

■
Let us say that I have actually tick level
data, tick level data means I get data not■

■
every hour of a trading day, I may get data
every minute or every second.■

■
So, I have huge data sets I mean that data
set will be huge.■

■
Now I want to decide whether the stock is
going to move up or move down.■

■
Now I have to predict whether I have whole
massive data set of all the stock prices up■

■
to that point for the last three months and
now I am saying tomorrow the market opens■

■
at 10 o'clock what is going to be the opening
price of this particular stock for which I■

■
have collected data.■

■
Now how are you go about doing this we said
the first option is to just use the three■

■
months of data that you have collected plain
data that you have collected use the same■

■
values.■

■
That would be called trace driven simulation.■

■
Second approach would be for the three month
data that you have collected why do not you■

■

fit a distribution and there has been enough
and more research on what is a good fit for

a stock price data.

Obviously everybody wants to crack that problem
and very clearly that I have not solved that

problem because if I had cracked that problem
I would not be sitting here it is already

11 o'clock I would be using my distribution
and playing with the market.

So, you can fit a distribution for the three
months of data that you have collected and

I have a whole bunch of candidate distributions
available.

Normal distribution, uniform distribution,
log normal distribution, weibull distribution,

the full family, not the full family, the
full forest.

And the third way is well the three months
of data that I have collected is for a fairly

weird stock, none of the distributions amicably
fit the data and therefore I want to define

my own distributions.

And therefore we got into the empirical distributions.

Therefore we got into the empirical distributions.

So, these are two examples of how to build
empirical distributions from the data that

we have.

Now let us go back and go to step number two
what if I want to fit theoretical distributions

how do I go about doing that.

So, before we do that let us quickly take
a look at how these three approaches compare

with each other.■

■

Usually approach one which is using the plane■
three months data is usually used to validate■

■

the models.■

■

We already have a model you already have the■
output and you want to validate whether that■

■

output is correct or not.■

■

So, what you do is you push these three months■
of data into your model and your model generates■

■

an output and you compare that output with■
the reality with the existing system which■

■

is what happens tomorrow check and whether■
that matches.■

■

So, essentially our trace driven simulation■
is mainly used to fit to validate a model■

■

that you already may have built using something,■
some different approach.■

■

So, you have some prior knowledge how to build■
models for stock prices you have already done■

■

that now you want to check whether that model■
is correct or not.■

■

And therefore you feed into that model these■
three months of data whatever comes out of■

■

this model should match with what happened■
in reality or so should come close to each■

■

other.■

■

The drawback of this approach is you are going■
to test your model only with the data that■

■

you have collected.■

■

So, for example going back to the sales volume■
data you only have 36 values.■

■

So, your model is going to be tested only■
using the 36 values that you have actually■

■
observed and fed into the model that may not be enough that may not be enough.■

■
Even with the three months of minute level data on the stock prices let us say the stock

■
price was fairly stable during these three months there was no turbulence in the market.■

■
So, how will you test whether your model works very well in the turbulent period.■

■
Now this data that you have collected will not give you that simulation because this

■
data was collected from a fairly stable stock period, a stock market period.■

■
So, those are some of the problems.■

■
Approaches 2 and 3 building your distributions or using a theoretical distribution kind of

■
avoid this problems.■

■
Because what you can do is once you have built a distribution you can generate values from

■
those distributions which are not restricted to the 36 values that you have actually observed

■
in your sample.■

■
So, compared to approach 1 I would say approach 2 and 3 are preferable that way.■

■
However if you can actually find a theoretical distribution that fits your data I would generally

■
avoid building empirical distributions.■

■
Therefore I would say that theoretical distributions are preferred over empirical distributions.■

■
The problem with empirical distributions is very similar to the problem that we have for

■
approach one.■

■
Now when you build an empirical distribution

from the data that you have the distribution,■

■

the shape of the distribution is completely■
governed by the data that you have used to■

■

build the distribution functions.■

■

Remember your distribution functions, your■
distribution functions are built from the■

■

data that you have.■

■

So, the shape of the distribution will be■
completely governed by your data.■

■

Now once again if the data is of a particular■
pattern then quite likely that the distribution■

■

will be biased towards that.■

■

The other problem is the distribution that■
we built usually are restricted by the smallest■

■

and the largest value.■

■

So, here the distribution is 0 for all the■
values lesser than the smallest value that■

■

you have observed.■

■

The distribution is 1 beyond or the maximum■
value that you have observed in the sample■

■

which may not be true.■

■

This is the smallest value that I have observed■
in the sample does not mean that sales cannot■

■

be lower than this.■

■

This is the maximum value that I have observed■
in the sample does not mean that my sales■

■

cannot be more than that.■

■

However, the distribution that you build using■
these data will pretty much say so.■

■

The distribution that we have built will say■
that probability of finding a sales volume■

■

lesser than the smallest value is zero.■

■

And indirectly speaking probability of finding■
a sales value sales volume bigger than the■

■

maximum value is again 0 almost 0.■

■

So, those are the problems.■

■

So, we are still not able to go beyond whatever■
we have observed in our sample.■

■

So, that's, those are the problems.■

■

So, if you want to test the validity of our■
system from an empirical, from a data that■

■

comes from an empirical distribution we may■
have problem because we cannot simulate values■

■

which are outside of the range that was fed■
into.■

■

So, those are some of the issues with empirical■
distributions.■

■

Now there may be some compelling reasons for■
using a particular theoretical distributions.■

■

For example let us say that you have data■
about reliability.■

■

Now reliability engineering has a very high■
importance for weibull distribution.■

■

So, for any data that comes about distribution■
or the reliability I would actually I would■

■

like to test whether it fits the weibull family■
is it coming close there.■

■

So, those cases also I mean theoretical distributions■
why not test it before?■

■

So, those are the, that's the difference between■
fitting a theoretical distribution and fitting■

■

an empirical distribution.■

■

■ ■

So, let us come back. Let us you know from ■

theory now. Let us take some examples. ■

■
So, for taking this example what I have done ■
is I have collected a dataset of 217 values. ■

■
Right now I will not tell you the context. ■
I will not tell you what the data is about. ■

■
Sometimes the context of the data helps you ■

■
guess a distribution. I am avoiding that because I ■
am not even telling you what the data is about.■

■
I am not telling you the units of data for ■
example I will not tell you whether it is ■

■
dollars or minutes or something some other units ■
I am not telling you that. I am just telling you ■

■
that there are 217 values we have collected ■
and for these data points we want to fit a ■

■
theoretical probability distribution. Now let ■
me tell you the properties of these 217 values. ■

■
Let me tell you a summary statistics. ■
So, you have all these things ■

■
you know the mean of these 217 values is ■
0.40 median and mode values are different, ■

■
apparently, there are multiple modes, standard ■
deviation is 0.38 skewness is 1.46 and what else, ■

■
minimum, maximum, minimum is 0.01 maximum is ■
1.96. So, what clues can you pick up from this ■

■
summary statistics because this is going to ■
help us build or guess a distribution.■

■
This is going to help us guess a distribution. Let ■
us quickly understand what this summary statistics ■

■
is telling us. If you noticed I have not named the ■
variable. I have simply called it variable one. ■

■
So, I have simply called it variable one ■
just to tell you that there is no context ■

■

to this variable. What do we notice first? ■

This is what I notice first. Look at this. ■

■

The mean is 0.4, the median is 0.28 the mode ■

is 0.05 which means that mean is not same as ■

■

mode is not same as median. ■

So, what do I understand from this for a symmetric ■

■

distribution the mean and the median and the ■

modular value coincide at the same point. Remember ■

■

normal distribution the most famous example of ■

symmetric distribution. For a normal distribution ■

■

this is where the mean is, this is where the ■

mode is, this is where the median is. So, ■

■

for a symmetric distribution the mean the median ■

and the modular value are going to coincide. ■

■

■

So, our data set seems to be not of this ■

category. Our data set has a different mean, ■

■

a significantly different median and significantly ■

different modular value. So, our data is not from ■

■

the symmetric distributions. So, that is ruled ■

out. All the symmetric distributions we can rule ■

■

out. Normal distribution is gone right now. ■

Uniform distribution is gone right now. Okay, ■

■

the mean value max value. The min ■

value is 0.01, max value is 1.96. ■

■

217 values, none of these values ■

seem to be on the negative side. ■

■

■

So, the support of our distribution does not seem ■

to be from negative infinity to positive infinity. ■

■

Why am I saying that because if the variable could ■

take on negative values from negative infinity to ■

■

positive infinity out of these 217 values at least ■

some of these values could have been negative. ■

■

Not essential but very very unlikely that if ■

the random variable can take negative values ■

■
in the 217 values that I have observed ■

none of them were negative. ■

■
■
So, the minimum value was 0.01, the maximum ■
value was 1.96 tells me that this is not the case ■

■
ok. So, the the distributions that go to the ■
negative side of the line probably are ruled out ■

■
One more important clue. This. What does this tell ■
me? What does this tell me ? Skewness. What does ■

■
skewness tell me? Skewness is what? Skewness tells ■
me about the symmetry of the distribution. ■

■
■
Now for my data set and once again I have ■
no clue what the data set represents. ■

■
The data set seems to be skewed and ■
particularly this is positive 1.46. ■

■
What does that mean? It's a positive skew. ■
Now do you recall what a positive skew means. ■

■
Okay what a positive skew means? Positive ■
skew means that it is skewed to the right. ■

■
Skewed to the right means the right ■
tail is bigger than the left tail. ■

■
■
So, this is your random variable X . This is the ■
density function. So, this is the left tail, ■

■
this is the right tail. So, we are saying that ■
a positive skew indicates that the right tail ■

■
is bigger than the left tail. This, the ■
difference in the mean mode and median ■

■
told us that ours is anyway not a symmetric ■
distribution. That was confirmed from the skewness ■

■
value. Skewness value is now telling me that ■
the right tail is bigger than the left tail.■

■
■

Right tail keeps on extending longer ■
than the extension of the left tail. ■

■
So, what are the positive skew distributions that ■
I can think of. Imagine all of those in your mind ■

■
and those are the potential candidates, those ■
are the distributions that I may want to fit ■

■
to my data. Right, so, these are the ■
clues. We will discuss these clues ■

■
in couple of slides. But these are the quick clues ■
that I can understand from the summary statistics ■

■
about the data set that I have collected.■
What more can I look at before I decide to fit ■

■
a distribution? Can you think of anything else ■
that I can present to you which will help you ■

■
guess a distribution? Obviously graphical output, ■
right? Why do not I show you box plot ■

■
This is the box plot. You ■
all know what a box plot is. ■

■
If you have not discussed box plot, well, let ■
us know and we will discuss box plot in the ■

■
subsequent sessions. This box plot is a typical ■
output from any statistical package but this box ■

■
plot does not give me a true picture. So, let me ■
tilt it by 90 degrees and show you this box plot. ■

■
Same box plot actually tilt at 90 degrees. ■
So, what does the box plot show? Box plot has ■

■
this box. This is your box. What is this box? This ■
box if you recall, this was the 25th percentile, ■

■
this was the 75th percentile, the middle line ■
is the 50th percentile. Do you know what is ■

■
50th percentile? Obviously the median and ■
then the whiskers, going back. This whisker, ■

■
the left whisker seems to be very short, the ■
right whisker seems to be going all the way. ■

Now what are these points 207 209 217 214? What are these points? These are the observation

numbers. Recall we had a data set of 217 values.

So, these are the observation numbers. So,

observation number 214, observation number 213, observation number 210, observation number 209.

So, the right tail seems to be, the right side values seem to be going. These are the values.

So, the median is somewhere here. This is somewhere the median and what was the median?

Let me go back to the previous slide. Let us go back one more, let us go back one more.

What was the median? Median was 0.28. Let us see where the median is.

0.28, yeah, somewhere there is 0.28. So, the left whisker seems to be only up to that point the

right whisker seems to be up to this point. However there are values beyond the right

whiskers. So, recall all the discussions that you may have had about the box plot and an

interpretation of the box plot. The only point that I wish to emphasize from the box plot,

the shape of the box plot is that the values on the right hand side are extending well into the

right side of the x axis which essentially means that this distribution has a right skew.

This distribution has a positive skew.

What else? Let me show you the bar chart for this data. This is the bar chart. Looks similar to

box plot. Now it tells us that there are large number of values which are close to zero,

very large number of values, about 15% ■

of the values are very very close to 0, ■

■

very close to 0 and then the kind of frequency ■

drops and there are very few values which are ■

■

more than 1.5. Very very few values, few ■

observations which are more than 1.5. ■

■

■

Recall our range was anyway 1.95 where the maximum ■

value was 1.96. So, this must have been 1.96. ■

■

So, this is the frequency, the height of the ■

bar chart obviously represents the frequency. ■

■

Now you may change the width of the bar ■

chart and try to get different shapes. ■

■

This is slightly thicker bars. The earlier one ■

was slightly thinner bars but the frequency seems ■

■

to be dropping as you go in the values. ■

So, as the values increases the frequency ■

■

seems to be dropping. Has this given you some ■

clues about what may be a distribution to fit, ■

■

what distribution may fit the data? ■

Any clues? Any ideas? keep thinking, ■

■

keep thinking. I will show you one more bar ■

chart even thicker bar this time. So, once again ■

■

as the values of the random variable increases ■

the frequency seems to be decreasing. ■

■

■

So, let us formally write down what ■

clues we get. what are the clues that ■

■

we are trying to understand from the data. ■

So, usually, for summarising whatever we have ■

■

discussed so far, for symmetric distributions ■

the mean and the median and the mode matches. ■

■

If in the data set if the mean value and ■

the median value are sufficiently close ■

■

to each other we may still think about symmetric ■

distributions but then you will ask me how close ■

■

is close enough? What is sufficiently close? ■

Well, those right now we are only making a ■

■

educated guess. We have not made any decision ■

yet. We are only guessing distributions. ■

■

■

So, if mean and median seem to ■

be close enough you can try out ■

■

symmetric distributions to fit the data but ■

if the mean and median are not close enough ■

■

probably symmetric distributions is not the ■

way to go forward. Look at the coefficient of ■

■

variation this is something that we had missed ■

out looking at the summary statistics. What is ■

■

coefficient of variation? Do you recall ■

what was coefficient of variation? ■

■

■

Coefficient of variation, CV, is indicated ■

by μ by σ sorry σ by μ . ■

■

Now what is the σ here? σ is 0.38 ■

and the mean is 0.4. This is your estimation ■

■

of σ . I mean this is not exactly σ , this ■

is actually s which is sample standard deviation ■

■

and you have a sample size of 217. In absence ■

of anything else, you are going to say that ■

■

this is my population standard deviation. ■

And this is my sample mean and I am going to ■

■

say that this is my best estimate for population ■

mean. So, the CV seems to be close to 1. 0.38 ■

■

divided by 0.4. Now CV does give me some ■

indication. If the CV is close to one ■

■

exponential distribution. CV for exponential ■

distribution is always one. If you recall ■

■

one by λ and one by λ are the mean and ■

variance. Sorry, $1/\lambda$ and $1/\lambda$ is ■

the mean and standard deviation. ■

So, for exponential distribution, CV ■

is always 1. For our data CV seems to be close to ■
one. If the histogram looks slightly right skewed ■

distribution with CV greater than 1, log normal ■
distribution may be a better approximation for ■

our data. Why is that true? Look at the shape of ■
the normal distribution log normal distribution. ■

So, for some distributions, however, this ■
CV data is not even useful. Not useful.■

Because it is not even defined. When is that? What ■
are the examples when CV may not even be defined? ■

Recall what was CV? CV was σ / μ . ■

Now do you recall standard normal ■
distribution? Standard normal distribution ■

μ is zero. If μ is zero, CV may not even be ■
defined. So, how are you going to use CV? So, ■

note that using CV may give us clues.■
However this is not the tell all kind of ■

a thing it gives us clues sometimes ■
it may not even be available. There ■

is something called a lexis ratio. Lexis ratio ■
is essentially CV for discrete distributions. ■

Similar interpretations but that is not called ■
CV that is usually called lexis ratio. ■

Now we have already discussed this skewness. ■
Skewness may give us some hints. Skewness for ■

normal distribution is zero. Because skewness ■
represents asymmetry of the data and normal ■

distribution is famously symmetric. So, if the ■
skewness value is 0 you are thinking about all the ■

symmetric distributions like normal distribution. ■

If skewness is positive, you are thinking about
right skewed distribution for example exponential
distribution which has a skewness of two.
And for skewness values which are
less than zero, negative skew,
you are talking about left skew distribution
where the left tail is bigger than the right
tail. The left tail extends longer than the
right tail. Now let me pause here and ask you.
Looking at all these discussions what seems to
be a good fit for our data? Our data has mean
mode median different, skewness is
positive, CV seems to be close to one.
CV seems to be close to one and we saw box
plot, we saw bar chart, we saw a lot of things,
what do you think? I would say let us try
fitting exponential distribution to our data.
One thing you also notice what is
the support for exponential data.
Support meaning what are the values that
an exponential random variable can take.
Exponential random variable can take
values from zero to infinity.
Can take values from zero to infinity, remember in
our data set we didn't have any negative values.
So, the support could start from 0. Obviously
our maximum value was 1.95, 1.96. Exponential
distribution can go all the way to infinity.
But why do not we check whether exponential
distribution fits very well. So, we are going to
try that at the end of the session, we are going
to share an excel sheet where you can try out

whether exponential distribution fits very well ■

■

but I have some results towards that. So, ■

let us go further, let's go beyond.■

■

■

Once you have done that, once you have ■

estimated a distribution to be fit, ■

■

let us estimate the parameters. So, parameters, ■

every probability distribution has a parameter ■

■

or set of parameters for example binomial ■

distribution you need n which is the number ■

■

of experiments to be conducted and p which is ■

the probability of success in each trial.■

■

■

For normal distribution μ and σ are the ■

two parameters. For exponential distribution ■

■

λ is the parameter. You know ■

what λ is. This is your λ ■

■

and you know how λ plays a very very ■

important role in defining the density function ■

■

and therefore define defining all the subsequent ■

properties of the exponential distribution. ■

■

■

Now the most commonly used method to ■

estimate the parameters of our distribution ■

■

happens to be MLE. What is MLE? it is the most ■

likelihood method, most likelihood estimation. ■

■

And I am assuming that MLE was also discussed ■

in some course in some sessions for you ■

■

before ,therefore, we are not going to ■

focus on how do you estimate parameters ■

■

using most likelihood estimators.■

You define a log estimate, ■

■

a log likelihood function, you take ■

the derivative of that log likelihood ■

■

function and then that's how you estimate the ■

parameters of your distribution. Let's not
go deeper into MLE but let us say that I
have guessed a distribution. For my data set,
I have guessed a distribution. Right now I
am going to try out exponential distribution.
Now exponential distribution has
a single parameter called lambda
and using maximum likelihood estimation method,
I probably have guessed the value of lambda.
Let us say that, that is also done. What's
next? Obviously guessing a distribution is
clearly not enough. We have to check how good
exponential distribution fits the data. How good
is this fit? We have, we right now have kind of
thought that exponential data will, exponential
distribution will fit the data. Well, how good is
this fit? That's the next thing to be done.
How good is this fit that is
precisely called goodness of fit
and it is called goodness of fit
because there are goodness of fit tests.
So, what are we trying to answer here through
these goodness of fit tests what are we trying
to answer we are trying to see whether the
fitted distribution is good enough for the
data what is the fitted distribution? The
distribution that we are trying to fit.
What is the distribution that we are trying
to fit? We are trying to fit an exponential
distribution what are the methods of doing that?
Well you can use frequency comparison you can
compare the frequency that comes from exponential

distribution and compare that with the frequency

that you have observed in the data set compare

that if the frequencies are matching then you say

that exponential distribution is a good fit.

You may use what are called as probability plots

and I have couple of examples in the subsequent slides those are essentially visual tools. They

tell you whether the observed probability or observed percentile or observed quartile matches the quartile then percentiles that may come from the distribution. If it fits you will

get a nice line if it does not fit you will be far away from that line a couple of slides later.

Or there are very rigorous statistical tests called goodness of fit tests many of them use a Chi square distribution there are other there are various tests those are also described.

Let us not look at frequency comparison that gets little bit technical let us look at probability plots. There are two kinds of probability plots we are going to look at.

One the first one is called Q-Q plot which is Quantile-Quantile plot. So, Quantile-Quantile plot essentially compares the the Qth quantile of the sample distribution. Sample distribution is the distribution from the sample and the correct distribution is the distribution that is fitted.

So, this indicates the distribution that we are trying to fit and this indicates the distribution that comes from the sample.

Now if this x that comes from the fitted distribution matches with the x that comes from

the sample distribution then you are going to get a nice line. So, you are going to plot all the x

that comes from the fitted distribution which is called the model distribution

and you are going to plot that against the x that comes from the sample distribution.

Now if this x matches with this x you are going to get a nice 45 degree line in your Q-Q plot.

Obviously this line is going to have an intercept of 0 and slope of 1.

So, this is going to be a 45 degree line and this is going to have an intercept

of 0 obviously as it has been drawn it is going to have a slope of 1.

Now let us very rarely I am going to get an exact 45 degree line. Most of the times

I am going to lie around this line and how far away am I from this 45 degree line tells

me how good is my model distribution how good is my fitted distribution.

If I am very far away from this line obviously the distribution that I am trying to fit

does not match with the sample. So, that is the interpretation of a Q-Q plot very similar

slightly different however interpretation is similar but the concept is different

it is called P-P plot. Probability-Probability plot.

So, essentially for a given x I am trying to plot the probability I am going to plot

the probability. P-P plot is applicable for both continuous as well as discrete data sets. So, I

am going to compare F with F . This F is the model distribution this F is the sample distribution.

Now once again if this x matches with this x I am going to have a fortified nice 45 degree line

ah once again with intercept 0 and slope 1. Generally ah why do we need 2 different plots

if the interpretation is going to be same I want a 45 degree line

or the points around a 45 degree line? Well if the interpretation is same why do I need two separate

plots P-P and Q-Q. Generally ah Q-Q plot will identify the differences between the tails of the

distribution if the fitted distribution and the sample distributions are different in their tails

that will get highlighted in the Q-Q plots. And if the difference in the distribution

is mainly in the middle portion what are the tails usually for a normal

distribution these are the tails and this is the middle portion of the distribution.

So, if the model distribution if the fitted distribution and the sample distribution

differ in the middle portion that gets highlighted in the P-P plot,

if the differences between the model distribution and sample distributions exist mainly in the tail

that gets highlighted in the Q-Q plot. Therefore we look at both the plots. Now

what was our earlier decision for the 217 values we had decided to fit

exponential distribution. Just for a reference point we are also going to fit normal distribution

and for these two distributions we are going to look at the P-P plot and the Q-Q plot.

■
So if we fit a normal distribution to our ■
data. Now our data was called variable ■

■
one and what have we tried to fit? We ■
have tried to fit normal distribution ■

■
what P-P plot does, P-P plot compares ■
probability with probability. So, what is this ■

■
this is the F from the model this ■
is the F from the sample. ■

■
So, the probability therefore this is going ■
to be from 0 to 1 this probability therefore ■

■
this is going to be from 0 to 1. So, on ■
the P-P plot we plot the probabilities ■

■
and how does the probability plot look like ■
here. So, 0 to 1 probability for the observed ■

■
I have made a mistake. So, this is not the model ■
this is the sample and this is the model. So, ■

■
and this is my 45 degree line and these are the ■
observed points these are actual observations. ■

■
Now do you think that the observed points are ■
close to the 45 degree line well not particularly ■

■
if you see the differences this is the ■
difference deviation, deviation from the normal. ■

■
So, the deviation seems to ■
be particularly high I mean ■

■
in this portion in the middle portion of the CDF ■
there is a deviation, in the left portion of the ■

■
CDF there is deviation, on the right portion ■
there is less deviation but there is deviation. ■

■
So, the deviation from the normal seems ■
to be quite high in the P-P plots.■

■
Let us look at the Q-Q plot this is ■
the P-P plot let us say this again. So, ■

in the P-P plot the deviation from the normal seems to be quite high. Now what happens if you look at the P-P plot for exponential this is where we are trying to fit exponential distribution to our data set our data set was variable one. Now look at the P-P plot once again this is the sample probability this is the fitted or the model probability 45 degree line. Look at the observed points very very close to 45 degree line. So, exponential distribution seems to be fitting better than the fit for the normal distribution. This was the fit for the normal distribution, P-P plot fit. This is the fit for the exponential distribution obviously exponential distribution seems to be fitting well. Let us observe the deviations these are the deviations notice the scale though. This deviation may look large but this deviation of the order of 0.04, the deviation from the normal was quite large of the order of 0.15 ok negative 0.15. So, there are deviations from the exponential distribution P-P plot however the deviations are of the order of 0.04 negative or positive nothing more than that. Therefore in terms of P-P plot we seem to be observing that exponential distribution fits better than the normal distribution. Let us look at the Q-Q plots let us fit a Q-Q plot let us plot a Q-Q fitting normal distribution to our data. Normal distribution I still want to Q-Q plot Q-Q plot is going to have

■
x values. So, x from the sample and x from the model which is from the fitted distribution. ■

■
Once again nowhere close to the 45 degree line. ■
Notice that this is the 45 degree line because ■

■
on the x axis this goes all the way to 2. ■
So, even though it may look like a ■

■
weird line but it is actually a 45 degree line. Now if you only take up the take up to ■

■
1.5. What if we fit we a plot a Q-Q ■
for exponential fit look at this. ■

■
Very close to the 45 degree line ■
there are some deviations here, ■

■
what are these deviations for ■
these deviations seems to be for ■

■
higher observed values higher observed values ■
close to 1.95, 1.96 towards the higher end of the ■

■
values in the sample that is where the deviation seems to be. Once again deviation. ■

■
So, even for the Q-Q plot we seem to be saying ■

■
that exponential distribution fits ■
better than the normal distribution. ■

■
So, looking at the visual tools do we conclude ■
that exponential distribution is a good fit for ■

■
the data? Well it is definitely better ■
fit compared to normal distribution ■

■
but do we conclude unfortunately not. So, what is ■
the last thing we look at we look at statistical ■

■
goodness of fit tests. So, essentially what are ■
we doing we are checking whether our data set ■

■
are IID variables what are IID variables? ■
Independent Identically Distributed random ■

■
variables with a particular distribution that ■

is our null hypothesis and there are two very

famous tests one is called chi square test the other one is called Kolmogorov Smirnov test.

Using these two tests we can actually check whether our data set has exponential distribution

fitting very well to the data. Before we do that the test of independence has to happen

we have to check whether they are independent values independent random variables.

So, there is a separate statistical test for checking the independence

and after you check the independence we come here and perform one of these two tests KS test which

is Kolmogorov Smirnov test or a simple Chi square test to check whether the data fits

or the exponential distribution with a particular value of lambda fits the data set very well.

Let us stop here if you have any questions we will definitely answer all of them but this is

essentially how we go about fitting distributions to any business data set. Obviously what will

follow is we will take examples of various business data sets in multiple contexts and

try to fit distributions to it. The last thing to be done before we close is putting the context.

Remember I told you I kept this variable as variable one. I kept this variable as variable

one without telling you the context in which the data was collected without telling you the

units of this data. Now I can tell you what the units of the data are, actually they represent

time taken to get service in a bank. This is

actually from a bank and the customers are coming ■

in and we are recording the amount of time taken ■
to get the service that they came there for.■

Obviously scale data. So, and we are ■
not going to reveal the bank we are ■

not going to reveal the branch but this ■
is essentially banking operations data ■

and now you know this is time data. Do ■
not look at the branch do not look at the ■

bank look at what the variable represents? It ■
represents time. Time can never be negative. So, ■

this variable is always going to ■
have a support over 0 to infinity.■

So, now I have told you the context and now you ■
know that the context also seems to indicate ■

that exponential distribution ■
may not be that bad a fit. And ■

you can take the support of queuing theory ■
I am trying to introduce a new idea here ■

queuing theory. Queuing theory tells you ■
a lot about the time taken in a queue and ■

therefore also you have some support to say that ■
exponential distribution fits very well. ■

Now that is a very vague statement that I made but ■

read up on the queuing theory and you will ■
understand why exponential distribution has a very ■

strong association with queuing theory. Let ■
me stop here and end the session here. ■

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. Some 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 us, 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 complement
be the event that the admission, the candidate

■
did not get any admission offer. Now, what
is this A^c ? A^c 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 complement 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 complement.

■
We can simply use a complement 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 complement 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. 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 short 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 0.73. 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 be 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 0.8? This 0.8, it is a marginal probability

that the candidate 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 0.3, 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 has 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 supplier S1 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■
Delhi. 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 brand preference example it turns out that the calculated test statistic

turns out to be 7.

So, 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 contingency table 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.■

■
■ ■
■ ■
Welcome to the ■

■
new session. In this session, we are going ■
to focus on the relationship between price ■

■
and demand through a curve called demand response ■
curve. So, what is the demand response curve? So, ■
■

generally the retailers are going to offer a price ■
for a particular product or service in the market. ■

Now, the market is going to react ■
by realizing a particular demand ■

at that particular price point. ■

So, the demand response curve essentially tells us ■

what is the realized demand at a price that is ■
offered for a particular product or service. ■

Now, why is this a business analytics topic? ■
Well, the demand response curve itself ■

is a two variable plot price on the x axis ■
and demand on the y axis price on the x axis ■

demand on the y axis. So at a particular point, ■
we keep changing the price and we keep realizing ■

the demand at that particular price. ■

So, today's session or a couple of sessions ■

are going to focus on how we estimate ■
this demand response curve. So, we are ■

going to discuss various aspects of various ■
relationship types between price and demand. ■

And in general, the basic mechanism of this ■
price and demand relationship through this ■

demand response curve. ■

So, typically, this is what a ■

demand response curve is going to look like. ■
So, we have this particular curve, a rough ■

curve which is done like this, this is what we ■
are going to call the demand response curve. ■

So, what does it say? So, at a particular price ■
let us say P_1 this is the price offered and you ■

hit the curve you look at this left and this ■
 Q_1 is the quantity demanded by the market ■

at that particular price point. ■

Now, we know that we are going to adjust ■

our demand based on the prices that are offered in the market. For example, if the price is P_3 , the quantity that is going to be demanded by the market is going to be Q_3 . So, every retailer has to decide what is the best price and let us call this best price as P^* what is the best price to be offered for that particular product or service. At that particular optimal let us say optimal price point there is going to be some optimal demand realized in the market.

Now, let us say that we make a mistake in calculating this optimal price and we end up pricing more. There is always a potential for the retailer to reduce the price which means go left to reduce the price and capture more demand because at P^* the demand is going to be Q^* , if we reduce the price if we reduce the price from P^* to P_1 , the demand is going to jump up from Q^* to Q_1 . So, this is that there is always a latent demand which is realized by this part of the region. This region represents the latent demand. How do we capture the latent demand? By reducing the price, by reducing the price we can capture more and more demand. Now, this portion of the plot represents what is called consumer surplus. Now, consumer surplus essentially is the benefit that the consumer gets by paying less and less price. So, if we are actually increasing the price if we are actually increasing the price from P^* to P_3 , let us say the consumer surplus is going to be eaten away that much. So, this light blue shaded region is going to be the reduction in the

consumer surplus that the consumer is getting, ■
 ■
 because the price is at P^* and not at P_3 . ■
 If the price is increased ■
 ■
 from P^* to P_3 , the consumer surplus is ■
 going to consumer surplus is going to go down ■
 ■
 by this much amount by the ■
 light blue region. ■
 ■
 ■
 So, this is how the region is going to be ■
 read, this is how the plot has to be read ■
 ■
 as I was saying. So, this optimal price P^* ■
 has to be very carefully chosen. So, ■
 ■
 if we reduce the price, there is always a scope to ■
 capture more demand, if we increase the price the ■
 ■
 consumer surplus may get affected we have I mean ■
 just like we are talking about consumer surplus ■
 ■
 we can talk about the producer surplus I ■
 mean I am pointing you to the basic economics ■
 ■
 course which is the slightly outside the ■
 current discussion on forecasting the curve, but ■
 ■
 nevertheless important for the discussion. ■
 Another important question then is if this is the ■
 ■
 price rate, if this is the demand response curve, ■
 how should the retailer decide this optimal price? ■
 ■
 What should be the objective of finalizing this ■
 price? Now, there are a couple of ways that we ■
 ■
 can go about. One is what is called as revenue ■
 maximizing price a revenue maximizing price ■
 ■
 every retailer wants to maximize the revenue the ■
 money that they collect from the customers of the ■
 ■
 product and service. So, that ■
 should always be a concern. ■
 ■
 ■
 However, we have to realize that maximizing ■
 revenue may not be the same as maximizing profit. ■

■
 So, there is a profit maximizing ■
 ■
 price, the price at which the profit is maximized, ■
 the price at which the revenue is maximized and ■
 ■
 we have to keep in mind that these two may be ■
 different objectives. Sorry, these are different ■
 ■
 objectives. We have to keep in mind that the ■
 optimal prices when we are maximizing revenue ■
 ■
 may be different from the optimal prices ■
 when we are maximizing the profit. ■
 ■
 ■
 How do we go about with these two objectives? Let ■
 us discuss that later. So, this is the basics of ■
 ■
 the demand response curve which is the ■
 blue curve which is pointed here. Let ■
 ■
 us understand the properties of this. ■
 So, this is essentially a function that describes ■
 ■
 how the demand varies as a function of price very ■
 similar to the demand supply curve in economics, ■
 ■
 however, this is for a single seller in a single ■
 market whereas, the demand supply curve aggregates ■
 ■
 various supply in the market and ■
 aggregates various demands in the market. ■
 ■
 So, this is slightly different from that scenario, ■
 where we are considering a single seller at ■
 ■
 a single time point in a single market. ■
 If you notice, there are four important things ■
 ■
 that you notice from this demand response curve. ■
 First of all, the demand for a product is always ■
 ■
 going to be non-negative, which means that it is ■
 always going to lie on the positive side of the ■
 ■
 y axis. You cannot have demand going negative. ■
 You cannot have demand going less than 0 does ■
 ■
 not make sense. Demand cannot be negative. ■
 So, the demand response curve is always going to ■

be on the positive side. Similarly, you cannot offer negative prices, I mean, let us not get into the details, but negative prices would mean that the retailer gives money to the customer for using the product and services, which generally does not happen in the market. Therefore, even the prices are going to be on the positive side.

So, you do not expect prices to go on the negative side. The lowest price is going to be 0. We are not going to look at a case where negative prices are even possible. So, the curve is going to be non-negative in both senses. The demand is going to be positive even if the prices offered in the market are going to be positive. It is going to be a continuous curve. So this is going to be a nice smooth continuous curve without any breakages without any breakages.

So, for example, this is the price and this is the demand. You do not expect something like this, which means that you do not know what the demand is going to be between these two price points.

So, what is going to be the demand? So, are we saying that there is no demand, but that also means that demand is 0. So, there is not going to be a discontinuity in the curve, there is not going to be a discontinuity in the curve. Very similarly, the curve is going to be very smooth and differentiable which means that tangent is always possible tangent is always possible at all the price points tangents is always possible at all the price points. So,

the curve is also going to be differentiable. ■
And unless we are talking about very specialized ■

■
goods, that curve is generally going to be ■
downward sloping. What does it mean? ■

■
■
It means that as the prices ■
increase the demand for the product ■

■
reduces. There are certain products for ■
example you can always think of Giffen goods. ■

■
Examples could be luxury items ■
for example, Rolex watches, ■

■
Rolex watches sometimes if you increase the prices ■
if you increase the prices, the demand for the ■

■
product may actually go up because the exclusivity ■
appeal for the product may go up. ■

■
■
So, but generally speaking, we ■
are going to look at goods whereas ■

■
the prices are increasing, the demand is actually ■
going to come down. So if the prices increase, ■

■
the demand is going to come down and therefore, ■
the curve is going to be a downward sloping curve, ■

■
downward sloping curve. So, these are ■
the four important properties once again, ■

■
the quantity is going to be non-negative, ■
the prices are going to be non-negative, ■

■
the prices are going to be non-negative. ■
So, it is first quadrant curve, the curve is continuous ■

■
and continuous, which means that it is continuous ■
it is differentiable and more importantly, ■

■
we are going to look at a scenario where the curve ■
is downward sloping the curve is downward sloping, ■

■
which means that as the prices go up, the ■
demand goes down, the demand goes down. ■

■
■

So, these four these four properties are ■
important. As I said, sometimes some goods ■
■
may or may not hold but we are right now going ■
to not worry about those kinds of goods. Right ■
■
we are going to look at goods where the demand ■
response curve is a downward sloping curve.■

[Music]■

so■
■
two ways of looking at how how we■
■
measure the price sensitivity uh how how■
■
sensitive is the is the demand to the■
■
price■
■
right■

so two ways of looking at it one is■
■
simply one is simple slope which is■
■
how much the demand changes with■
■
response to the change in price■
■
uh so it is essentially the change in■
■
demand demand at uh price point p_2 minus■
■
demand at a price point p_1 minus p_1 ah■
■
divided by p_2 minus p_1 so it's a it's■
■
so essentially■
■
the delta in the demand divided by delta■
■
in the price right so very very uh■
■
simple uh uh ratio of change in uh uh■
■
change in demand divided by change in■
■
price right so■

■
as we know■
■
since the curve is downward sloping■
■
downward sloping■
■
if p_1 is greater than p_2 this means that■
■
the demand at p_1 is lesser than demand■
■
at p_2 what does that mean■
■
we go back■
■
right we go back let us go back■
■
so if■
■
so let us look at■
■
let us look at what is this point here■
■
this point here is demand at p_3 ■
■
and this point here is demand at p_1 ■
■
now we know that p_3 is greater than p_1 ■
■
the price p_3 is greater than price p_1 ■
■
and we can also see that demand at■
■
 p_3 demand at p_3 is here which is q_3 ■
■
so this is q_3 which is always lesser■
■
than demand at p_1 which is q_1 right so■
■
 q_1 is greater than q_3 ■
■
whereas p_3 is greater than p_1 so we know■
■
that right■
■
which means that this slope this slope■
■
if you define the slope it is going to■
■
be a negative■
■

negative value the slope is going to be

negative value

so we can we can we can

look at slope as a local local estimator

of change in demand for a very small

change in price

so very simple ah just calculate the

calculate the slope of the slope of the

tangent at that particular point and we

will get the ah we will get the

ah sensitivity

ah

for that particular price

there is other way to calculate price

sensitivity and that is called demand

elasticity okay that is called demand

elasticity

so it is the ratio of percentage change

in demand to the percentage change in

price notice the difference notice the

difference right

notice the difference so essentially it

is this is change in demand with change

in price this is percentage change in

demand with percentage change in price

so you divide then the

■
 numerator by demand at p1 you divide the ■
 ■
 denominator by price p1 right so ■
 ■
 so essentially the numerator becomes a ■
 ■
 unitless quantity because you are ■
 ■
 dividing demand by demand so this the ■
 ■
 units cancel out ■
 ■
 the denominator becomes a unitless ■
 ■
 quantity because it is price divided by ■
 ■
 price so units cancel out so essentially ■
 ■
 unlike slope ■
 ■
 elasticity is unit less quantity it's a ■
 ■
 unit less quantity ■
 ■
 right uh so it is remember i let us ■
 ■
 define it again percentage change in ■
 ■
 demand to the percentage change in price ■
 ■
 right so percentage change in demand to ■
 ■
 the percentage change in price ■
 ■
 so for example elasticity of 2 would ■
 ■
 mean that a 10 percent reduction in ■
 ■
 price ■
 ■
 uh is essentially increases the demand ■
 ■
 by 20 percent right that that that's the ■
 ■
 meaning of this ■
 ■
 two so 20 percent increase so 10 percent ■
 ■
 reduction in price increases the demand ■
 ■

by 20■

■

right so we are we are talking■

■

percentages we are talking percentages■

■

in elasticity unlike slope where we are■

■

talking about change in demand to change■

■

in price■

■

elasticity is percentage change in■

■

demand to percentage change in price■

■

so that's the difference between■

■

elasticity and slope of the demand■

■

response curve■

■

now uh what do what what is the what is■

■

the■

■

general interpretation of elasticity■

■

there are certain goods which are■

■

supposed to be less elastic■

■

for example■

■

ah let us take the example of common■

■

salt that we use in our food■

■

now■

■

salt is required without salt the food■

■

is just not going to taste which means■

■

that even if the prices of salt go up i■

■

don't expect our consumption of salt to■

■

to change that much because salt is■

■

needed salt is essential quantity right■

■
without salt the the the food is not■
■
going to taste■
■
so i would expect the elasticity of■
■
something like salt to be much lesser■
■
right■
■
otherwise uh go to the other extreme ah■
■
think about uh think about■
■
a service like a holiday■
■
right a service like a holiday now uh uh■
■
holiday uh■
■
if the if the holiday is going to cost■
■
us too much there is a very high■
■
probability that we may change our plan■
■
right most of us may want to change our■
■
plan we may still go to our holiday but■
■
we may probably■
■
choose a different service■
■
ah reduce the number of days or do■
■
something but essentially react to that■
■
change in price■
■
so■
■
i would say that a holiday is a service■
■
where the elasticity is generally quite■
■
high right■
■
so here uh there is another■
■

thing that we have to look at is■

■

elasticity may also depend on time■

■

elasticity may also depend on time so■

■

there is a short term elasticity and■

■

there may be a long term elasticity so■

■

here here is here are few examples right■

■

so for example uh here are few examples■

■

of as i said salt■

■

for a product like salt for a product■

■

like salt uh in short term i need salt i■

■

just need salt right i mean there is■

■

absolutely no■

■

there is no alternative to that■

■

so for a for a product like salt uh the■

■

elasticity may be■

■

i mean yeah zero right uh i i just i■

■

just need salt that's it i just need■

■

salt■

■

uh for example on the other hand uh a■

■

two wheeler a two wheeler right uh a two■

■

wheeler i will say that elasticity is■

■

quite large because uh uh if the■

■

two-wheeler is going to cost me too much■

■

uh■

■

i will say ah■

■

let let me take a bus today let me take■

■
a bus today and not buy this two wheeler■
■
right so i i may i may i may look for■
■
alternate modes of transport at least in■
■
the short term at least in the short■
■
term■
■
i may i may look for alternates uh■
■
alternatives to buying a two wheeler■
■
so in that sense the the the short term■
■
elasticity of two wheeler may be much■
■
much higher than a short term elasticity■
■
for sure for salt■
■
now let us say that there is some■
■
emergency meeting that i have to attend■
■
ah there may be some emergency meeting■
■
that i have to attend and i just have to■
■
take a flight i mean there is no■
■
alternative i just have to take a flight■
■
i have to go attend a meeting■
■
i have to go attend to a personal thing■
■
uh i just need to travel■
■
so■
■
in the short term■
■
in case of emergency for example we may■
■
argue that■
■
even if the prices are higher■
■

ah if the if the demand requires i mean■
■
if the if the situation requires that i■
■
travel i will have to travel i will have■
■
to travel■
■
so pop the airline travel may have a■
■
very very uh low short-term elasticity■
■
uh movies right movies uh if the movie■
■
tickets are expensive i may postpone■
■
i may say let me watch tv at home today■
■
instead of buying a 500 rupee■
■
movie ticket■
■
or something like that let let me go to■
■
my friend's house■
■
right but■
■
i i may want to postpone■
■
that that purchase because i may want to■
■
find alternatives whereas for a salt■
■
there is actually no alternative■
■
right so■
■
in those cases i may expect a larger■
■
short-term elasticity but as i said■
■
elasticity may also have a time axis■
■
which means that the long term■
■
elasticity may be different■
■
for example for air travel■
■
now if i have emergency requirement i■

■
must reach there faster and therefore■
■
whatever is the price i may want to pay■
■
and catch that flight and reach my■
■
destination■
■
however in the long term in the long■
■
term if i have if i can plan■
■
and if the price tickets if the flight■
■
tickets are really■
■
expensive■
■
i may still want to find alternate mode■
■
of transportation and therefore in the■
■
long term airline travel■
■
may have a large value for elasticity so■
■
in the short term emergency cases i■
■
don't care if the price of the flight is■
■
uh too much i have to travel therefore i■
■
have to travel elasticity may be low in■
■
the long term if you allow me to plan my■
■
trip carefully■
■
ah if the if the■
■
price points are just not acceptable to■
■
me■
■
i may find alternate modes of■
■
transportation and therefore■
■
the elasticity goes up significantly in■
■

the long term■
■
for salt■
■
the elasticity may not change that much■
■
right because as i said salt is■
■
essential commodity salt is essential■
■
commodity■
■
so salt is essential commemoration it■
■
may go up little bit■
■
maybe i i will i will say that anyway■
■
eating up■
■
too much of salt is bad for my health i■
■
may cut down on the quantity of salt i■
■
eat■
■
however the the the margin that i have■
■
is is quite less therefore the■
■
elasticity may not change drastically■
■
as it would change for airline travel■
■
ah look at uh look at two wheeler on the■
■
other hand for a two wheeler uh today i■
■
go to the showroom and i say the prices■
■
of two wheeler is too much and therefore■
■
let today at least at least for a week■
■
time let me manage with alternate modes■
■
of transportation uh but in the long■
■
term if there is a demand for two■
■
wheeler there is a demand for two■

■
wheeler i mean really in the long term■
■
you really can't avoid■
■
so■
■
there are certain goods just wanted to■
■
show you that there are certain goods■
■
where the elasticity may come down over■
■
a period of time there are certain goods■
■
where the elasticity may go up■
■
drastically over a period of time a■
■
petrol is the example of later ah two■
■
wheeler as an example of former■
■
right so for movies movies uh uh i mean■
■
in general if the ticket prices are■
■
expensive i i may find alternate modes■
■
of■
■
alternate modes of transportation uh■
■
alternate modes of entertainment■
■
alternate modes of entertainment but uh■
■
i may still go for a movie in the long■
■
term right uh if you allow me to plan■
■
longer time maybe i'll buy subscription■
■
to one of the ott and i never have to go■
■
for uh■
■
movie theater right■
■
so the elasticity may go up drastically■
■

over a longer period of time■
■
right so that that's how elasticity■
■
changes■
■
ah depending on the goods■
■
right depending on the goods and■
■
depending on the time frame depending on■
■
the time frame right sometimes it■
■
changes drastically like 0.1 to 2.4■
■
these are by the way examples we are not■
■
saying that airline travel has a■
■
elasticity of 2.4■
■
in the long term right■
■
it may depend on■
■
it may depend on■
■
general consumer behavior right for some■
■
people airline travel■
■
may not be that elastic because even in■
■
long term uh i i i may not want to■
■
prefer alternate modes of transportation■
■
uh if i want to go for a movie i i once■
■
again 3.7 is not going to be elasticity■
■
for everyone in the long term right so■
■
these are just■
■
ah these are just■
■
representative numbers they are not■
■
elasticity values for everybody in the■

■
short term or for the long term but i■
■
hope i have conveyed the concept of■
■
elasticity■
■
depending on the product and services or■
■
depending on the time frame depending on■
■
the time access■

■
all right so■

■
just to recap we have looked at price■

■
sensitivity using two methods one is■

■
calculating the slope the other one is■

■
calculating the elasticity■

■
and we have interpreted both of them■

■
■
So, let us come to the different relationships■
between the price and demand.■

■
Different relationship.■

■
So, this this blue curve should the blue curve■
be a straight downward sloping line or should■

■
it be a nonlinear downward sloping, a downward■
sloping curve.■

■
How should the relationship be that that will■
also matter.■

■
The simplest possible example is a linear■
demand response curve.■

■
So, demand response curve.■

■
So, demand at a particular price point, demand■
at a particular price point p may have some■

■
initial value D_0 , slope is given by m and■
 p is the price.■

■

So, what is the demand at a particular price?■

■

So, this is the price and D_0 minus m into■

p will be the demand at that particular price.■

■

This D_0 is called the demand at price 0.■

■

When p is 0, what is the demand?■

■

Demand is D_0 .■

■

So, this is the demanded price equal to 0.■

■

This can also be called as the market size,■

because look at the interpretation.■

■

Price is 0, you are offering the product for■

free.■

■

So, when you are offering the product for■

free, what, how many people actually demand■

■

this product will really tell you what is■

the market size.■

■

What is the complete market size?■

■

What is the total market size?■

■

Because you really cannot do anything better.■

■

You are offering the product for free.■

■

So, the demand that, that particular price■

point which is actually 0, you cannot really■

■

use anything lesser than that.■

■

Anything less than that actually means that■

you are actually giving money to the consumers■

■

to use your product.■

■

So, as we said, we are not going to consider■

that case, we are going to look at only non-negative■

■

values.■

■

So, price equal to 0 is really the minimum■

price.■

■

At that price, what is the demand, that that is called the market size and m is called

the slope.

So, mathematically what is D_0 ?

D_0 will be the y intercept.

We are talking about a linear response curve, so we are going to draw a straight line.

Let me change the color, so, let us say that this is the demand response curve.

As I said earlier, what is on the x, always remember what is on the x axis, x axis is

always the price and y axis is always the demand.

Demand responds to a particular price.

This guy here is D_0 because the price is 0.

At price equal to 0, what is the total market size?

To be precisely correct, I should not have drawn this portion, I should simply start

here, and end here because as it is negative or non-negative.

So, I should not extend it to the negative region I should not extend it below the x

axis.

So, this is called D_0 this is called the market size.

Now the price at which demand is equal to 0 is called satiating price.

Let us represent that by a new color.

This is, this price is called satiating price.

Satiating price means, the demand is, the market is satiated.

■
That is it.■

■
If you offer, if you increase the price beyond■
PS.■

■
If you increase the price beyond PS, demand■
is expected to be 0.■

■
This is this is really the highest price that■
you can charge.■

■
This is really the highest price that you■
can charge.■

■
So, that price is given by when the demand■
is equal to 0, so set this equal to 0 and■

■
you are going to get the satiating price at■
as D_0 divided by m , so that is called the■

■
satiating price.■

■
So, we have looked at the y intercept, which■
is the market size, we have looked at the■

■
x intercept, which is called the satiating■
price.■

■
The elasticity of this curve is given by this■
expression.■

■
And you can verify from the definition of■
elasticity that this really is the elasticity■

■
of this curve.■

■
What we should notice what is important to■
notice is, this explanation you can derive■

■
using the definition that we had a couple■
of slides ago.■

■
What you should be more attentive towards■
is to realize that this elasticity depends■

■
on the price, which means that, the elasticity■
does not remain constant in this region.■

■
Elasticity keeps changing.■
■

How does it change?■

■

When price is 0, the elasticity is 0, price■
is 0 elasticity is 0.■

■

Whereas, when the price approaches P_s really■
cannot charge more than P_s because at P_s the■

■

demand becomes 0, what is the point of charging■
more than P_s , the demand has already become■

■

0.■

■

So, as the price approaches P_s , the elasticity■
goes to infinity.■

■

So, elasticity can be as low as 0, it can■
be as high as infinity when the price is approaching■

■

P_s , when the price is approaching P_s .■

■

So, for a linear response curve elasticity■
is not constant.■

■

elasticity can keep changing, alright.■

■

Take a moment to digest.■

■

So, this is ■
a linear demand response curve, and therefore,■

■

it is a straight line downward sloping, positive,■
continuous and differentiable and these are■

■

the interpretations.■

■

Interpretation of market size, interpretation■
of satiating price and the elasticity interpretation.■

■

Now, the relationship between price and demand■
need not always be linear, it may be nonlinear.■

■

So, there is another curve called constant■
elasticity curve.■

■

So, here we saw the elasticity keeps changing■
from 0 to infinity.■

■

Now, what if we want to keep the elasticity■
constant?■

■

That curve is given by this expression this
expression demand is equal to C into p to

the power of negative, ϵ , ϵ is
the elasticity. C is a constant C is the demand

when price is equal to 1, so this nonlinear
relationship will represent a curve where

the elasticity does not change.

Elasticity is constant at a particular value
of ϵ .

Now, you fix up the value of ϵ and the
curve will change.

So, essentially, this is your price x axis
is always the price y axis is always the demand,

and there may be a nonlinear relationship
once again downward sloping, non-negative,

continuous and differentiable, but the shape
of the curve is nonlinear this time.

What is the property the property is, we want
to keep the elasticity constant.

Now, you cannot guarantee that the demand
may be either finite or the demand even may

be satiated.

The market may be satiated.

Because you realize that as p approaches 0
as you reduce the price if you reduce the

price, the demand keeps spiking, demand keeps
growing.

So, demand may even approach infinity, as
the prices are reducing, so you cannot say

that demand is finite, because demand may
spike up, demand may spike up as you are closing

towards 0 from the right, and you realize,
that the demand may never touch the x axis.

The demand may never touch the x axis.■

■
It may be asymptotic on the x axis.■

■
So, for any value of p the demand may or may not even become 0.■

■
There is always some demand, there is always some customer who is trying to buy the product■

■
at a price that you may offer.■

■
So, we cannot guarantee that our demand may be finite, we cannot guarantee that demand■

■
curve will ever hit the x axis, we cannot guarantee.■

■
Revenue for the previous curve also for this curve also revenue is always given by price■

■
multiplied by the demand.■

■
So, the total demand for the product multiplied by the price charged for each product will■

■
tell you the revenue.■

■
For the constant elasticity curve the revenue will be C into p into C into p to the power■

■
of 1 minus epsilon.■

■
As I said earlier, the price may be optimized for maximizing the revenue.■

■
In that case you will maximize this curve and then find p star or we may maximize profit■

■
and therefore, get an optimal value of price.■

■
So, if we are maximizing revenue, this is the function that we are trying to maximize.■

■
Revenue is simply price multiplied by the demand.■

■
You can do the same thing even for the linear demand response curve.■

■
I hope things are clear so far.■

■
So, let us move ahead.■

■
But before we move ahead, so essentially,■
if the elasticity less than 1, what does it■

■
mean?■

■
It means that it is a product like salt, it■
is a product like salt where the product is■

■
actually quite inelastic, which means that■
the demand does not change even if the prices■

■
fluctuate.■

■
As I said, we need salt that is it, we need■
salt cannot do without salt.■

■
So, in that case, there is always a way to■
increase revenue simply by increasing the■

■
prices.■

■
Because revenues are prices multiplied by■
the demand.■

■
Even if you increase the price, the demand■
is not going to change much and therefore,■

■
revenues can be increased simply by increasing■
the prices.■

■
However, if we are dealing with a product■
which is highly elastic, which means that■

■
the elasticity is more than 1, then revenue■
making revenue probably can be increased by■

■
setting up very, very low prices, setting■
up very, very low prices.■

■
And therefore, essentially, mass selling therefore,■
we sell a lot and therefore generate more■

■
revenue.■

■
Price of each unit may be quite small, but■
that is compensated by huge demand because■

■
we saw the curve as the prices are lesser■

the demand may be shooting upwards and that

is how there is a mechanism to increase the revenue.

But there may be other mechanisms to get more profit, but right now, we are talking about

increasing the revenue because we saw the revenue functions.

So, the strategies available for inelastic product may be different from the strategies

available for a highly elastic product, highly elastic demand.

Let us pause here for a moment and recap what we have done.

So, we have looked at what is called as a demand response curve.

Demand response curve is the relationship between price and demand.

We have seen four important properties.

It is usually downward sloping, except for few goods.

Luxury goods are good example where the downward sloping may get violated.

Always non-negative, considering the case that we are offering 0 as the minimum price

and it has nice properties that it is continuous and differentiable.

We need continued continuity, we need differentiability because we should be able to take derivative.

Because we want to maximize revenue, we may want to maximize profit, therefore, finding

optimal prices, we may need these properties of continuity and differentiability.

So, these are the four nice properties of a demand response curve.

■
Demand sensitivity is measured using slope, ■
which is just change in demand to change in ■
■
prices. ■

■
Slope is generally negative or we can measure ■
price sensitivity using what is called as ■
■
elasticity. ■

■
Elasticity is percentage change in demand ■
to percentage change in price, the ratio of ■
■
these two. ■

■
So, usually unitless quantity a different ■
interpretation from slope. ■

■
And one important property of elasticity is, ■
it may have a time axis the elasticity value ■

■
in short term may be different than the elasticity ■
value in the long term. ■

■
Different relationship types between price ■
and demand we can think of a very simple linear ■

■
relationship between price and demand. ■

■
So, D_0 minus m into p where m is the slope ■
 D_0 is the y intercept, y intercept is also ■

■
called the market size, that is the demand ■
when price is 0. ■

■
When the demand is 0, that price is called ■
satiating price, elasticity of a linear demand ■

■
response curve actually is not constant. ■

■
It keeps changing, it may be as low as 0, ■
it may be very close to infinity when the ■

■
prices are fairly close to the satiating price. ■

■
So that is a very simple linear demand response ■
curves. ■

■
We can think about our nonlinear response ■
curve. ■

■
For example, we can think about a curve where the elasticity remains constant, and that

■
curve is given by demand is equal to C into p to the power of negative ϵ where ϵ

■
is the elasticity. C is a constant, because it represents demand when price is equal to

■
1.

■
In such a curve, the basic properties are not still violated, it is still a non-negative

■
curve.

■
It is still a downward sloping curve.

■
It has continuity, it is differentiable.

■
So, all those properties hold, but in this kind of a curve, we cannot guarantee that

■
the demand may be finite, we cannot guarantee that the demand is always satiated.

■
So, the revenue is simply price multiplied by demand.

■
And if we are dealing with a product, which is complete, which is very inelastic there

■
is a scope to increase revenue by simply increasing price.

■
If we are dealing with a very, very elastic product, the revenue can be increased by reducing

■
the price setting the price fairly close to 0.

■
So, let us end the session here.

■
And we will continue with this.

■
As I said the objective of this topic is to be able to estimate this demand response curve,

■
but we will come to that in the next session.

■

So, let us pause and come back for the next session of this later.

[Music]

okay welcome back to the session on

demand response curve we were looking at

we were looking at the relationship

between price and demand in the last

session we saw the relationship

to have four properties non-negativity

downward sloping continuity and

differentiability

we saw a slope and elasticity to be the

way to measure price sensitivity

and we saw two types of relationship one

was a simple linear relationship

and the other one was uh

relationship which maintains the

elasticity in the in a in the linear

demand response curve

ah the elasticity may not be constant

however in a constant elasticity

relation relationship the elasticity

remains constant

but the relationship becomes non-linear

we saw that ah we saw that the

mechanisms for increasing the revenue

■
may be different whether we are dealing■
■
with a different for■
■
ah inelastic product and it may be■
■
different for a elastic product■
■
now the now let us continue with the■
■
further question what is the what is the■
■
what is the■
■
analytics problem here■
■
the analytics problem here■
■
is to be able to estimate a demand■
■
response curve■
■
right■
■
so we can we can think of conducting an■
■
experiment in the market where we offer■
■
different prices■
■
and at those different prices we check■
■
what is the realized demand at that■
■
particular time so for example■
■
ah for for example let me change this■
■
thing■
■
so i may offer a price i may offer a■
■
price■
■
and actually measure the■
■
quantity demanded in the market so let■
■
us say that for a particular product let■
■

us not even■

■

give a specific product■

■

let us say that i offer a price of three■

■

three rupees what is the realized demand■

■

if i offer a price of six rupees what is■

■

the demand if i offer a price of two■

■

rupees what is the demand if i offer a■

■

price of 10 rupees what is the demand■

■

right now i don't expect the values to■

■

become negative because we are dealing■

■

with non-negative■

■

relationships right non-negative■

■

values■

■

so even if i say that the price is 50 i■

■

will say ah probably the demand comes■

■

down to 0 may not come down to 0 but■

■

but 0 is the■

■

the■

■

smallest value that i am going to■

■

observe■

■

so■

■

essentially we want to collect this kind■

■

of a data through an experiment■

■

so essentially we will have prices■

■

offered and corresponding realized■

■

demand values■

■
so essentially ah ■
■
demand can be considered to be the ■
■
dependent variable demand ah is a ■
■
reaction of the market demand is a ■
■
reaction of the market ■
■
prices are the triggers for that ■
■
reaction therefore prices can be ■
■
considered to be an explanatory variable ■
■
prices explain the difference in the ■
■
demand values ■
■
so prices can be considered to be ■
■
explanatory variable demand can be ■
■
considered to be dependent variable and ■
■
let us say that from some experiment ■
■
that we have conducted in the ■
■
marketplace at a particular time ■
■
i have this data available now from this ■
■
i want to estimate what may be the slope ■
■
i want to estimate what may be the ■
■
elasticity ■
■
 ah ■
■
the elasticity may not be constant ■
■
however if my relationship looks like a ■
■
relationship where ■
■
i have ■
■

i have d is equal to c into p to the
power of negative epsilon i may think of
having a constant elasticity right uh
slope
may come in only when i am talking about
a linear relationship
so pop
soap
linear relationship so essentially first
of all i may have to think about what
kind of relationship i want to fit if i
want to fit a linear relationship
between price and demand i may be
interested in finding the slope
if i am thinking of
if i am thinking of
fitting a relationship where the
elasticity may be constant
i may i may want to calculate the
elasticity i may want to estimate the
elasticity from the data that i may have
right
so the value of slope i am going to get
is only going to be an estimate of slope
from the data that i may have i
may not have the entire population

■
values i may have collected a small ah■
■
sample value of prices and the■
■
corresponding demand at those particular■
■
prices■
■
■ ■
So, what does this problem looks like? Any■
guesses. To me, this problem looks like a■
■
simple linear regression problem.■
■
So, a Simple Linear Regression, SLR, Simple■
Linear Regression, may help us calculate the■
■
y-intercept. Which are the market size and■
the slope. So, SLR will tell us if the linear■
■
relationship is a good fit for the data available■
from the market experiment.■
■
Now, what do we do when we are thinking about■
a non-linear relationship where we are, we■
■
want to keep the elasticity constant, what■
do we do with that? Can we use SLR in those■
■
cases? What is the guess? Can we use SLR in■
those cases? What is the answer? Probably,■
■
yes. And we will see how that can, how SLR■
can even help in that case. So, what I have■
■
is sample data and let us build a corresponding■
simple linear relationship, a simple linear■
■
regression model for that data.■
■
So, let me stop the presentation ■
and go ■
■
to the Excel sheet. So, let us say, let me■
increase the zoom here. So, let us say that■
■
these are the various prices offered and these■
were the corresponding demand at those particular■
■
prices, various demands at these particular■

prices. When the price was 9 rupees, the demand

realized was 2,891, when the demand was 32, the price, when the price was 32, the demand

was 370, when the price was 26, the demand was 2,946. So, these are the different combinations

of price and demand at that particular price.

So, let us see if the linear relationship fits better. So, what are we going to do?

Let us understand the plot. So, what I am going to do is draw a scatterplot. So, the

x-axis is the price, the y-axis is the demand. This is a typical demand response curve. So,

these are the realized points. So, when the price was 5 rupees, the realized demand was

6,707. When the price was 39, the realized demand was 297. When the price was 35, the

realized demand was 484. When the price was 40, the realized demand was 193.

So, now let me try and fit a linear relationship between price and demand. Let us try to fit

that. So, what I will do? I will go, add a trend line. I will add a linear trend line.

And I will ask Excel to print the equation and the r square for the relationship. Do

we know how to interpret this equation? And do we know how to interpret this r square?

Do you recall your DM course, probably it was discussed there? Anyway, does not matter.

Let us read this value.

So, what is this? This is, what is the equation? The equation is y is equal to 5842.8 minus

157.7 into x. What is y? Y is demand. What is on the y-axis? Y-axis has demand. Demand

is 5843. Let me write, let me round this up
to point 8, I will write this as 5843 minus,

I will round this up and I will say 158 into
the price because the x-axis is the price.

So, how do you interpret this 5843? Compare
that with a linear relationship which is D

is equal to D_0 minus m into p , so this is
your D_0 . This is the demand when the price

is equal to 0. Even if you offer the product
for free, what is going to be the demand?

The demand is going to be 5,800 odd.

Now, for a unit change in price, per unit
change in price how much do you expect the

demand to change, that is given by the slope.
So, for a unit change in demand, for a unit

change in price, the demand is going to go
down by 158 units, a negative slope. You increase

the price by 1 unit, the demand is going to
go down by 158 units. That is what the predicted

equation says, prediction equation says.

Now, what is this R square of 0.733? What
does that mean? Do you recall? For that let

us run a simple linear regression model. How
do you run a simple linear regression model

in Excel? So, I go to data. I have a chance
to show you the Excel add-in. There is no

data analysis toolpak here. So, let me add
a data analysis toolpak. So, we will go to

options, go to add-ins and I will ask Excel
to add analysis toolpak, data analysis toolpak.

I will say go. So, I will add an analysis
toolpak. Let me not add solver add-in right

now. Let us add that.

Now, you see one more option available here■
called data analysis. Let me use that data■
■
analysis. Let me go to the option called regression.■
What are my y values? My y values, my response■
■
variable, my dependent variable is demand.■
So, I will say this is my variable. What is■
■
the explanatory variable, x values, the x-axis■
is the price? So, these are the various price■
■
values. I will tell Excel, this is where the■
prices are. Labels, yes, I have labels in■
■
the first row, yes. Output range, let me ask■
Excel to print the output here and let us■
■
print it. So, this is how Excel runs regression.■
■
Now, let us understand, let us interpret these■
values. So, this 0.85, is multiple R. What■
■
is multiple R? If you recall, Excel had an■
option called regression. What we are running■
■
is what is called simple linear regression.■
Excel does not differentiate between simple■
■
linear regression and multiple linear regression.■
So, Excel simply says regression. Therefore,■
■
the R that is run, the R that is calculated■
is called multiple R, more suited for MLR,■
■
Multiple Linear Regression, but this R represents■
the correlation coefficient between the price■
■
and the demand.■
■
Do you recall what the correlation coefficient■
is? You must recall it from the DM course.■
■
So, how do you, independently how would you■
calculate a correlation coefficient? So, there■
■
is a function in Excel called CORREL, correlation,■
and you say that these are, this is my array■
■
number one, you do not have to have the label,■
and this is my array number two, it is negative■

0.85. Negative relationship, because as the prices are going up, the demand is coming down. So, this is the correlation coefficient between prices and demand.

Now, it should not matter whether I put the demand array first or the pricing, the prices array first. So, here I have put the demand array first and the prices array later, does

not matter. Correlation is, does not depend on which variable is entered first. So, even if I entered this first, this later, the value should not change. So, it does not matter.

It is just the relationship, it is just the correlation coefficient between two arrays of data.

So, what does this 0.855 represent? 0.855 represent a correlation coefficient between price and demand. The negative value represents the sign of correlation. And the value 0.85

represents the strength of correlation. The value is closer to 1 stronger than the association between the variables and the negative or positive value only tells us the direction

of the association. As the prices go up the demand goes down, and therefore, that is represented by this negative value.

Now, what is the R square? R square is simply the square of the R. If I square this correlation coefficient, if I do this, this square, I am going to get R square. R square is the

square of the correlation coefficient. Now, there is one more interpretation of the R square. Let us do that little later. Now,

what is that adjusted R square? Let us leave

the adjusted R square discussion also to a case when we are discussing multiple linear

regression. Adjusted R square only makes sense when we have more than one explanatory variable.

So, let us go back to PPT. Now, I have explained the important things. I have explained the

prediction equation. How do I predict demand?

So, I have predicted the y-intercept which

is 5843, the predicted value of the slope is, the estimated value of the slope is 158,

and we have explained R square which is the square of the correlation term. Let me go

back to PPT and explain a few more details about the simple linear regression.

So, what is simple linear regression? Simple linear regression describes how the conditional

mean of Y changes with different values of X. So, what is conditional mean? What do you

mean by conditional mean? You give me a value of X, I will tell you what is the value of

Y. what is the value of Y? Value of Y is going to be beta naught plus beta 1 x. What is this

beta naught? Beta naught is D0 for us. What is beta 1? Beta 1 is m for us. So, this is

called the conditional mean of y with respect to x.

So, the simple linear regression model, simple regression model, simple linear regression

model essentially tells us that it is going to be a line with an intercept of B0 and a

slope of B1. The intercept was the y-intercept which is D0, which is the market size. This

is, however, the expected value of Y. This

is the expected value of Y . So, however, the

points that we saw, the points may, the observed points may be away from them, away from this

line. So, the gap at any price point this gap between what is the expected value of

Y and what is the actual value of Y this gap is essentially the error in the regression

model.

So, the deviation from the mean is called error. Errors are usually represented by ϵ .

I should not have confused that with elasticity. But let us also denote error by ϵ and

this is not to be confused with elasticity. Elasticity is also ϵ , the error is also

ϵ , but they have no correlation, only the same notation. So, on average, we do not

expect any errors. Therefore, the error terms are expected to have a value of 0 on average.

So, what do I mean by, once again, what is this deviation? So, this deviation is, this

is the prediction line. Where is this prediction line coming from? This prediction line is

coming from the data that we have collected. So, this may be data

and this is the production line. So, what are we saying? We are saying that at this

particular price, let us say P_1 , the x-axis is always price, the y-axis is always demand.

This is my predicted value of y . And how is that given? That is calculated using this

expression, $\beta_0 + \beta_1 x$. That will tell me what is the expected value of

demand at that particular price.

However, the realized value is somewhere here, ■
the realized value was somewhere here. So, ■

■
this gap, the vertical gap between the actual ■
value of demand realized at that particular ■

■
price point and the predicted value of y coming ■
from the simple linear regression this gap ■

■
is called error. Now, on average, I do not ■
expect this error. Therefore, on average, ■

■
the expected value of error is 0. So, what ■
is, what are errors? Errors are deviations ■

■
of responses from the predicted value of y . ■
Where is the predicted value of y ? This is ■

■
the predicted value of y . This is the actual ■
value of y . At this particular price point, ■

■
let us say P_3 , the predicted value is here, ■
the actual value is here, so this is the y , ■

■
this is the error. ■

■
So, now then how am I supposed to draw this ■
line. I am supposed to draw this line in such ■

■
a way that these errors are minimized, but ■
you realize the problem with that. Sometimes ■

■
the error may be positive. Let us say if I ■
calculate this epsilon as y minus \hat{y} , y ■

■
is the predicted value of y , let us say, and ■
 \hat{y} is the observed value. So, the predicted ■

■
value comes from this β_0 plus β_1 ■
 x minus \hat{y} , \hat{y} is the actual observed ■

■
value. Now, sometimes y minus \hat{y} may be ■
positive, which is in this case, y minus \hat{y} ■

■
 \hat{y} turns out to be positive, sometimes the ■
 \hat{y} may, sometimes the epsilon may turn ■

■
out to be negative, y minus \hat{y} may be negative. ■

■
Now, if I say that, let us say this is my ■
error one and this is my error three and this ■

■
is my error seven at different values of price, ■
so if I simply add up the errors, e_1 plus ■
■
 e_2 plus e_3 all the way to e_n , I have n values ■
in my sample. And if I say that minimizing ■
■
this, it may not work out because some of ■
the positive errors may cancel out some of ■
■
the negative errors. And I may in general ■
underestimate the total error in the model. ■
■
Therefore, even though the objective is to ■
minimize the errors, we usually do not minimize ■
■
the sum of errors. ■
■
So, what do we do? What is the method of nullifying ■
the effect of positive or negative error? ■
■
Usually, we square the error terms, we usually ■
square the error terms, and only then minimize ■
■
the summation. So, in general, what are we ■
minimizing? We are minimizing the sum of square ■
■
errors, sum because it is all summed up, the ■
sum of square errors. That is the objective ■
■
with which we run the regression model. So, ■
what is this? This is a mean square error ■
■
line. This is the line where the mean sum ■
of square errors is minimized. That is how ■
■
we generate the line. ■
■
What are the properties of this error term? ■
We are making essentially three assumptions ■
■
about the error term. We want the error term ■
to be independent. What do I mean by that? ■
■
We do not want the error term even to be dependent ■
on error term e_2 , we do not want error three ■
■
to be dependent on e_1 and e_2 , we do not, in ■
general, want these error terms to be independent. ■
■
We want these error terms to have equal variances. ■

The variances of all the error terms should

be σ^2 . And we want these error terms to be normally distributed. We

want the error terms to be normally distributed.

These are the three assumptions about the

error terms.

Now, even before we interpret our simple linear regression model, we should confirm that these

assumptions hold. If the error terms are not independent, if the error terms do not have

equal variances, if the error terms do not have normality, we are going to say that the

assumptions are violated, and therefore, we should be very, very careful in interpreting

the results of our regression model.

How do we check the assumptions? We will do that little later. Right now, let us start

by saying that we assume that these three hold. We assume that the errors are independent.

We assume that the errors have equal variances.

We assume that the error terms have a normal

distribution. But these assumptions have to be checked.

So, with these error terms, what is the observed value of y ? The observed value of y is the

predicted value of y plus the error terms, where error terms have a normal distribution.

Why do I say that the error terms of normal distribution, because that is my assumption?

On average this is the mean of the normal distribution, this is the variance of the

normal distribution. On average, we had anyway said that the expected value of error terms

is 0. On average, we expect the error terms to have 0 value. So, the mean is 0 and the

variance is σ^2 . So, that is how we write this. It is a normally distributed

error term with a mean 0 and variance σ^2 .

Once again, because the error terms have these three assumptions because we want the error

terms to have these three properties, we are going to now say that the observations are

independent of each other, the observations y are independent of each other, they have

equal variances around the regression line and they are normally distributed around the

regression line. What do I mean by that?

So, essentially, if we say, these are the value of x , this is the value of y , and let

us say that this is a regression line, our regression line was downward sloping. So,

let me draw this as a downward sloping line. Now, at a particular x , at a particular value

of x , let us say x_1 , this is the predicted value of y , this is the estimated value of

y . This is what the regression line is going to give me $\beta_0 + \beta_1 x$. That

is what the line is going to give me. The observation, however, could be anywhere.

So, what is the, where is the observation? The observation we are going to say is normally

distributed. It can be anywhere on this line. It could be, the observed value of y could

be here, could be here, could be here, could be here, could be anywhere. The expected value

is going to be here. Similarly, pick a different

value of x , let us pick it sufficiently away

so that I can draw this properly, let us pick
a different value of x . This is going to be

the point on the y , a point on the green curve
is going to be the predicted value of y . So,

this is the predicted value of y .

However, where can the y , actual y be? Actual
 y , so the actual y

could be anywhere. So, it could be here, it
could be here, it could be here for that particular

value of x . So, because of this error term,
the observed value of y could be anywhere

because of the error terms. So, we can keep
drawing this normal distribution curve for

each value of x . We can keep doing this.

Let us stop and go back to the Excel sheet.
Let me stop the presentation. So, what is

the standard error now? We have seen this
is essentially your sigma epsilon, the standard

deviation of the error term. This is not a
square value, this is a standard error. So,

this is the standard deviation, not the squared
value.

So, now, we said we are going to defer the
interpretation of 0.724 later. When we take

up an example of MLR we had described this,
we had described this, now we are describing

this, which is the standard deviation of the
error terms. A number of observations, there

are 31 observations. You can notice there
are 31 observations. The first row is the

label. So, there are 31 observations. So,
 n is 31. So, this is that n 31.

So, how do you interpret this ANOVA table, first of all, degrees of freedom? Why do we

have one degree of freedom for regression? Because what is the prediction equation. The

prediction equation is y is equal to β_0 plus $\beta_1 x$. So, how many values

are we estimating? We are estimating the value of β_0 , which is the y-intercept,

we are estimating the value of slope β_1 . So, we are estimating two parameters. So,

degrees of freedom will be 2 minus 1, this is 1.

Why is the total degree of freedom 30? If you recall, remember your regression discussion

in the previous course, why is this 30? This is always $n - 1$. I had 31 observations,

31 minus 1 is 30. Why do I get 29 here? So, the total degree of freedom is 30. So, this

is 30 total degree of freedom minus the regression degree of freedom is 1, so 30 minus 1 is 29.

That is how I get my degrees of freedom. Why do I get 1 here? Recall your discussion of

regression from the previous courses. It is two parameters being estimated β_0

and β_1 , 2 minus 1 is 1. Why is this 30? Total observations are 32, sorry, 31, total

observations are 31, 31 minus 1 is 30. Why is this 29? Total degrees of freedom is 30,

regression degrees of freedom is 1, so 30 minus 1 is 29.

Now, the sum of squares, I am not going to get into details of the sum of squares. I

am going to rely on your previous course. Otherwise, I will put up a primer on how do

you interpret the sum of squares for regression and the sum of squares for residual. The Sum of squares for residual and sum of squares for regression essentially comes from how do you calculate the sum of squares for regression and the sum of squares for error. So, what is this value? What is this large value? This large value is your SSE, the sum of squares of the errors, residuals are also errors. So, this is your SSE. This is called SSR or SSM, the sum of squares of the model or the sum of squares of the regression. Usually, if you call it SSR, you have to be careful because is it SS for regression or is it SS for residuals you have to be careful. So, you can always call this SSM, the sum of squares for the model, sum of squares for error to avoid any confusion. How is mean square calculated? Mean square is calculated as the sum of squares divided by degrees of freedom. So, this 133 is calculated as the sum of squares divided by 1. How is this mean square error calculated? The mean square of error is calculated as the sum of squares of the error divided by degrees of freedom. So, if you want to verify this, how will you get this? So, this is calculated as the sum of squares divided by degrees of freedom and that is how you get your mean sum of the square. Now, Mean Sum of Squares, this is Mean Sum of Squares, this is still the sum of squares. So, this is MSE, Mean Sum of Squares, but this is still the sum of squares. Now, if you take a square root of this, what are

you supposed to get? You will get your standard

error, because we just standard deviation,
you take a square root. The square root of

MSE will give you the standard error. So,
it is that value.

So, let us get rid of these. We do not want
this. We only want to interpret the Excel

output. How do you calculate the F test statistic?
First of all, what is the F test statistic?

This is a test statistic. So, this is a test
statistic. What is the test statistic? The

test statistic is usually for a hypothesis
test. So, what hypothesis are we testing?

We are testing a hypothesis that the regression
model is significant or not significant, H_0

and H_1 . So, here we are saying the
regression model is not significant. The null

hypothesis is regression model is not significant.

Now, what is the regression model? The regression
model is $\beta_0 + \beta_1 x$. When will

the regression model not be significant? When
will we say that x has no bearing on the values

of y ? When will we say that? When this β_1
value is 0, if this β_1 value is 0, then

x will have no impact on y and then your entire
regression model will collapse. So, what is

the other way of saying that the regression
model is not significant? You can say that

this β_1 is 0. Once β_1 is 0, the regression
model will not be significant against the

alternate hypothesis that regression is significant.
So, the overall significance of the regression

is tested using this F test statistic.

Now, let us not get into the distribution.■
It has an F distribution which has two degrees■
■
of freedom, one is numerator degrees of freedom,■
one is denominator degrees of freedom. Let■
■
us not look at the shape of the F distribution.■
Let us say that this is the test statistic■
■
value. Now, you usually look at the test statistic■
value and decide about the rejection of the■
■
null hypothesis by looking at the P-value.■
This is the P-value. So, this is the P-value■
■
for this hypothesis test.■
■
Now, how do you decide about the hypothesis,■
you compare the P-value with the significant■
■
value which is alpha, alpha is usually 5 percent,■
so 0.05, and if P is less than alpha, you■
■
are generally going to reject the null hypothesis.■
What is the P-value here? P-value is, of the■
■
order of 10 to the power of negative 10 which■
is anyway going to be lesser than 0.05. Therefore,■
■
we are going to reject this null hypothesis.■
What was the null hypothesis?■
■
The null hypothesis was that the regression■
model is not significant. So, we are going■
■
to reject this null hypothesis. And therefore,■
conclude that the regression model is significant.■
■
So, we say that the alternate hypothesis is■
good, a regression model is significant. So,■
■
that is the overall significance of the regression■
model is tested using this P-value. This is■
■
the overall significance test of a regression■
model. P-value is 7.8, 10 to the power of■
■
negative 10 anything of that order is going■
to be lesser than 0.05 and therefore we can■
■
safely reject the null hypothesis. So, that■

is how you interpret the upper block, ANOVA

block of the regression output of excel.

Take a minute to digest it. Let us hope that you have understood. If you have not, you

can playback the video and go back, more importantly, go back to your regression discussion in the

previous courses that should help you understand and interpret this ANOVA table of the Excel

regression output. That is out of the way and we have confirmed that the regression

is significant, let us interpret, let us see what is the estimated value of the y-intercept

and the slope. So, y-intercept was 5842, we knew that 5842, 5842.8 was the estimate of

the y-intercept, which is the total market size. What was the slope estimate? The slope

estimate was negative 157, we knew that from the excel output earlier. So, these are your

coefficients.

So, what are we saying? We are saying that the estimate of beta naught, we cannot know

the value of beta naught, we are saying that the estimate of beta naught, let us call it

beta, b_0 , let us say that that is b_0 , that value is 5842, an estimate of beta 1, which

is b_1 , what value is negative 157.7. And anyway, this is an estimation. This is the population

value, this is the sample value, sample value from the 31 observations that you have. This

is the sample value. This is how you are going to estimate the population value from the

sample value, so these are sample values. So, there is always going to be some error.

So, this is the standard error for estimating
of intercept, this is a standard error for
estimating the slope.

Now, we can individually test whether this
beta naught is 0 or beta 1 is 0 by running

a localized hypothesis test. This was the
overall hypothesis test whether the regression

was significant, now let us run a local hypothesis
test. What is this test statistic for? This

is the test statistic for a hypothesis. What
is the null hypothesis? We are checking whether

beta 1 is 0, alternate hypothesis beta 1 not
0.

What is the P-value? P-value is anyway very,
very small. And therefore, we reject this

null hypothesis and say that beta 1 is not
0. What about this, what about this? This

is essentially checking a hypothesis that
beta naught is 0 against the alternate hypothesis

that beta naught is not 0. What is the P-value?
P-value is of the order of 10 to the power

of negative 15. And therefore, we can safely
reject even this null hypothesis, and therefore,

confirm that beta naught is not 0.

Now, can you see a similarity? Look at this
hypothesis test and look at this hypothesis

test, more particularly focusing on the P-value.
P-value here was 7.8, 10 to the power of negative

10. P-value here was the same. Why did that
happen? This is happening because we are running

essentially a simple linear regression model.
In a simple linear regression model, what

is the overall significance of regression?
Regression will not be significant if beta

naught is 0.

So, when we are checking the overall regression significance hypothesis, we are essentially

checking whether beta naught is 0. Here we are directly checking whether beta naught

is 0. So, even though the test statistic used was different, here it was a t statistic and

here it was an F statistic, the overall test is the same. We are checking whether beta

naught is 0 or not. If beta naught is 0, your regression collapses. And therefore, it is

not very surprising that the P-value is the same.

So, if you rejected the null hypothesis of overall regression insignificance and therefore

concluded that the regression was significant, you are going to pretty much conclude the

same thing here by saying that beta 1 is not 0. Excel also prints the confidence interval,

95 percent confidence interval of beta 1. It says that beta 1 is going to be anywhere

from 193 negative to 121 negatives. And what is the expected value? The expected value

is negative 157. Similarly, for beta naught, beta naught is expected to be anywhere from

5023 to 6662 and the expected value is 5842. That is how you interpret the y-intercept

and the slope reported by Excel regression output. Remember that this is only an estimate.

You never, you are going to know the actual population value beta naught and beta 1. What

you have is some estimate b_{naught} and b_1 . So, that is all you can get from regression

because you only have 31 values, 31-odd values, 31 exactly values.

Let me conclude this regression output by telling you one more way of interpreting this

R square. Earlier what did I, what did we say, R square was just the square of the correlation

coefficient. Now, you can think of R square as the ability of the regression to explain

the variability in y. So, one interpretation of R square is that the prices, these prices,

different prices are helping me explain the variability of y. There is variability in

the demand values.

So, this regression model, what are, how can we interpret this 0.73, we can say that using

price as an explanatory variable, how much explanatory power does it have, we can say

that it can explain 73 percent variability in y. Let me say that again. Using prices

as an explanatory variable can explain 73 percent variability in y. That is the interpretation

of this R square. This is called the coefficient of determination. This is

only a recall. This must have been done in the previous courses. We are only going to

recall.

How do you calculate this 0.733? So, what is the sum of the square, because of regression

it is this value. What is the total sum of squares? It is this value. So, if I simply

do this, this is the amount explained by the regression and this is the total variability

that should be 0.733. So, because of regression, I can explain this much amount of variation

■
out of the total variation in y . So, the R^2 square is called the coefficient of determination.■

■
It tells me the amount of variability explained because of the regression model. Here, it■

■
is a simple regression model. I have only used price as an explanatory variable. So,■

■
what does that do to our demand response curve?■

■
Now, going back ■
to the PPT, let me go back to PPT. So, here■

■
I estimated a linear demand response curve.■
I estimated the y-intercept. I estimated the■

■
slope. And therefore, I am going to fit a linear demand response relationship to my■

■
data. And what was the data? This was the data. And that is how we have used simple■

■
linear regression to predict our market size.■
Market size is expected to be 5842, total■

■
demand is supposed to be 5842. Even if you give away the product for free, we expect■

■
about 5842 demand value to be realized.■

■
Every increase in the prices, every unit increase in the prices is going to reduce the demand■

■
by 157 units. So, that is how we estimate the demand response curve. And let us pause■

■
here and end this session with an explanation of SLR applied to the price demand relationship.■

■
So, let us pause here, continue with this in the next session. So, I am pausing the■

■
video now.■

■

■ ■