

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 an ascending order where the

smallest value is called $X_{(1)}$, OK $X_{(1)}$
smallest value is called

$X_{(2)}$ and the largest value is called
 $X_{(n)}$ in our case $X_{(36)}$, 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 k th 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 sigma here? Sigma is 0.38 and the mean is 0.4. This is your estimation

of sigma. I mean this is not exactly sigma, 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 lambda and one by lambda are the mean and variance. Sorry, $1/\lambda$ and $1/\lambda$ are 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 I have 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. ■
■