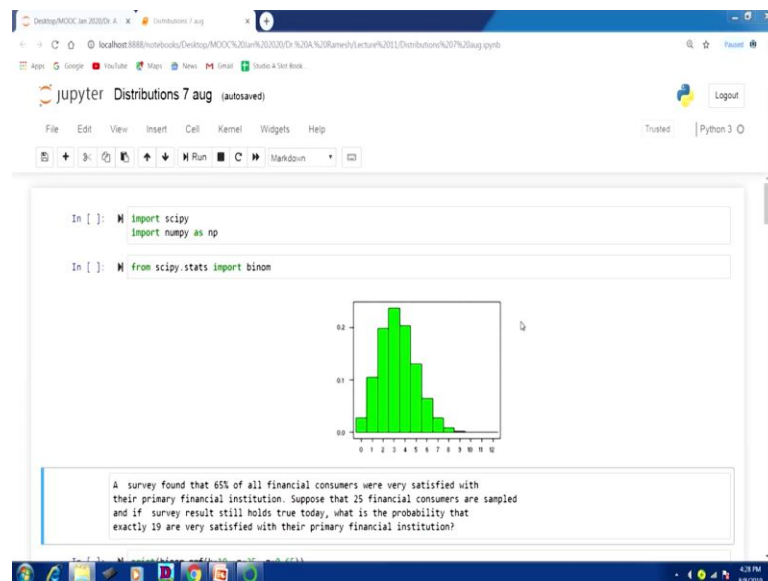


Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 11
Python Demo for Distribution

Dear students in the last lecture we have seen probability distributions in this lecture with the help of Python we solve some problems from probability distributions.

(Refer Slide Time: 00:44)



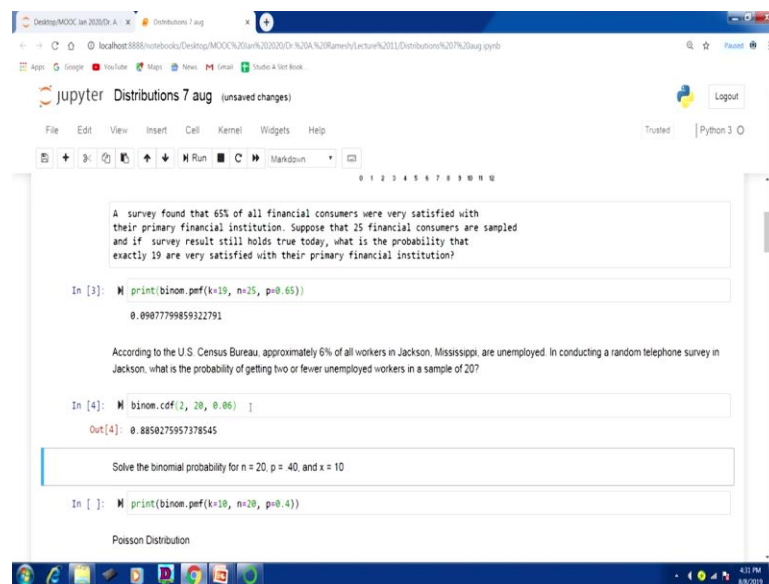
The problem is taken from a book written by Ken Black the title of the book is applied statistics so we will see the problems now. Now I am importing scipy importing numpy as np from scipy dot stats import binom, binom is for doing binomial operations and one more thing you can import picture also for example this picture that is that empirical distribution picture I have taken from this source that is exclamation symbol square bracket that link and that link should not be in square bracket then you will when you do that link we can get that picture directly I am executing this.

Yes now we will see the problem a survey found that 65% of all financial consumers were very satisfied with their primary financial institution. Suppose that 25 financial consumers are sampled and if the survey result still hold the true today what is the probability that exactly 90 are very satisfied with their primary finance institutions. By looking at the problem we have to

see what kind of distribution we are going to use here there are 2 possibilities satisfied or not satisfied.

So, there are 2 possibilities there so then we can go for binomial distributions print binom dot pmf, probability mass function k equal to 19 that is we say in the probability distribution current context, n equal to number of sample p is probability of success. Now we can see that the answer is 0.09 so there is 0.09 the probability that exactly nine are satisfied with a primary financial institutions.

(Refer Slide Time: 02:43)



The screenshot shows a Jupyter Notebook interface with the following content:

- Cell 1:** A text block describing a survey: "A survey found that 65% of all financial consumers were very satisfied with their primary financial institution. Suppose that 25 financial consumers are sampled and if survey result still holds true today, what is the probability that exactly 19 are very satisfied with their primary financial institution?"
- Cell 2:** A code cell with the command `print(binom.pmf(k=19, n=25, p=0.65))` and the output `0.09077799859322791`.
- Cell 3:** A text block describing a survey: "According to the U.S. Census Bureau, approximately 6% of all workers in Jackson, Mississippi, are unemployed. In conducting a random telephone survey in Jackson, what is the probability of getting two or fewer unemployed workers in a sample of 20?"
- Cell 4:** A code cell with the command `binom.cdf(2, 20, 0.06)` and the output `0.8850275957378545`.
- Cell 5:** A text block asking to "Solve the binomial probability for n = 20, p = 40, and x = 10".
- Cell 6:** A code cell with the command `print(binom.pmf(k=10, n=20, p=0.4))`.
- Cell 7:** A text block titled "Poisson Distribution".

We go to the next problem this Book this problem also taken from that book, according to US Census Bureau approximately 6 percentage of all workers in Jackson Mississippi are unemployed in conducting a random telephone survey in Jackson what is the probability of getting 2 or fewer unemployed workers in a sample of 20. Here we want to know 2 are less so say the probability of 0 plus probability of 1 plus probability of 2.

So we were to do the cumulative density function. So, here for doing that one you have to enter type `binom.cdf(2, 20, 0.6)` the 2 represents less than or equal to 2, 20 represents the sample, the p represents the probability. So, when you run this you are getting community probability of 88.5%.

We will take another problem solve the binomial probability n equal to 20 sample size is 20 p equal to 0.4 x equal to 10 so `binom dot pmf` you will get the answer for 0.117.

We will go to the next distribution Poisson distribution. So, in the Poisson distribution for doing Poisson distribution you have to import the library Poisson distribution from `scipy dot stats` import Poisson first we will find out Poisson probability mass function so `poisson.pmf (3, 2)`. 3 represents the x and 2 represent the mean. We will see another problem, suppose bank customers arrives randomly on any weekday afternoon at an average of 3.2 customers every 4 minutes, what is the probability of exactly 5 customers arriving in a 4 minute interval on a weekday afternoon.

By looking at the problem you say that we know that the arrival pattern follow Poisson distribution and you have to be very careful on the unit of mean and the unit of x both are in 4 minutes then no problems they simply can `poisson.pmf(5, 3.2)` is your x value Poisson dot `pmf` so 3.2 is the arrival rate, so 5,3 that is the 11.39%. You will see one more problem, bank customers arrive randomly on weekday afternoon at an average of 3.2 customers every 4 minutes what is the probability of having more than 7 customers in you 4 minute interval on a week day afternoon.

So, here we have to find out the probability of x greater than 7 so what we will do first we will find up to 7 with the help of this world that we will save any object called `prob`, equal to `poisson dot cdf 7 and lambda 3.2` so when you subtract 1 minus of this 1 then we will get probability of more than that, yes so I am finding up to 7, when you substrate 1 minus that up to 7 we will get to more than 7. we will see another problem on Poisson.

On a bank has an average random arrival rate of 3.2 customers every 4 minutes what is the probability of getting exactly 10 customers during 8 minutes interval. now it should be very careful here the unit of x and unit of λ are different, because it's a 4 minutes it is 8 minutes so you have to convert into same units so multiply by 32 by 2 you will get 6.4, so λ equal to 10 so Poisson dot `pmf` of 10, 6.4 will give you the answer for 0.0527.

(Refer Slide Time: 06:47)

The screenshot shows a Jupyter Notebook interface. At the top, the title is 'Distributions 7 aug'. Below the title bar, there's a menu with 'File', 'Edit', 'View', 'Insert', 'Cell', 'Kernel', 'Widgets', and 'Help'. A toolbar contains icons for running, saving, and other actions. The notebook content includes:

```
In [13]: poisson.pmf(10, 6.4)
Out[13]: 0.052790043854115495
```

Below this is a section titled 'Uniform Distribution' with a text box containing a problem statement:

Suppose the amount of time it takes to assemble a plastic module ranges from 27 to 39 seconds and that assembly times are uniformly distributed. Describe the distribution. What is the probability that a given assembly will take between 30 and 35 seconds?

Below the text box, there are several code cells:

```
In [ ]: U= np.arange(27, 40, 1)
U
```

```
In [ ]: from scipy.stats import uniform
uniform.mean(loc=27, scale=12)
```

```
In [ ]: uniform.cdf(np.arange(30, 36, 1), loc=27, scale=12)
```

```
In [ ]: Prob = 0.6666667 - 0.25
Prob
```

At the bottom of the notebook, there is a small text box with a reference to the National Association of Insurance Commissioners and a problem about automobile insurance costs.

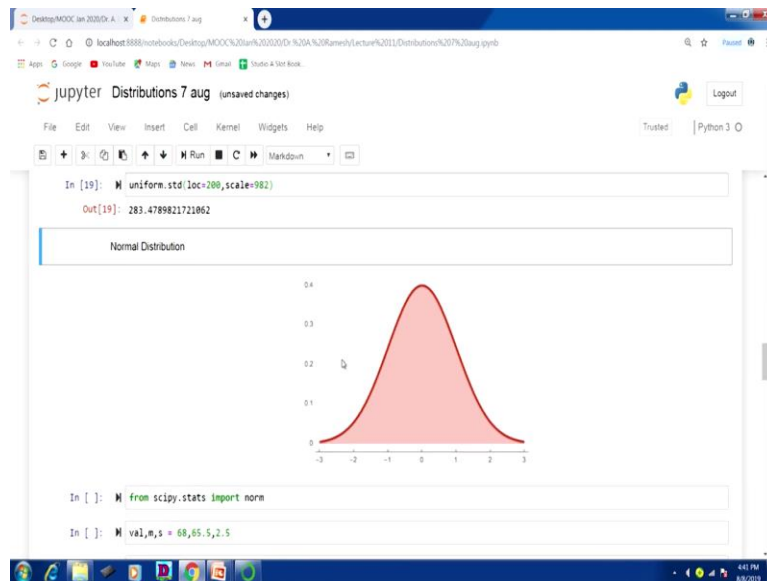
We will go to uniform distribution next suppose the amount of time it takes you see that the amount of time it takes to assembly a plastic module ranges from 27 to 39 seconds and the assembly time are uniformly distributed describe the distribution what is the probability that a given assembly will take between 30 to 35. first we will develop that array, so here u equal to np dot arrange, there are 2 functions one is range another one is arrange, arrange means its array, array function if you type if you type simply range that is a list.

So 27 is the starting value say '40-1' because it is $n - 1$, not 40, 39 will be the last value in the array, the increment is by 1, we got this one, now this is our uniform distribution, now from scipy dot start import uniform so we will find out the mean of this distribution so for that purpose uniform dot mean loc is the starting point 27, scale is how much plus 12 so it is a $27 + 12$ is 39. So, that is a syntax for, so the mean is 33 otherwise in the uniform distribution finding the mean is not a very complicated formula simply you have to find out the $(a + b) / 2$.

Then we will do the cumulative distribution cdf, cumulative function, so uniform dot cdf np dot arrange 30 what the question was asked is 30 to 35. So, np dot array 30 because it is a 35 you have to go out to 36 the increment is 1, starting point is 27 this scale is 12. so this will give you the probability between 30 to 35 so, when you run this so the probability of 30 is 0.25, 31 is 0.33, 32 is 0.41 and so on. Suppose we want between 30 and 35 for 30 the probability is 0.25 for 35 it

is 0.66, so if you subtract $0.66 - 0.25$ you will get the; and so far that probability that the given assembly will take between 30 to 35 seconds.

(Refer Slide Time: 09:33)



You will see one more problem according to the National Association of Insurance Commissioners the average annual cost of automobile insurance in the United States in a recent year was 691 dollar. Suppose the automobile insurance cost are uniformly distributed in the United States with an average of, from \$200 to 1182 dollar what is the standard deviation of this uniform distribution. So, we have to find out standard deviation of this distribution. Before that will check the mean the mean is given 691 dollar.

So, we will verify this uniform dot mean loc starting point is 200 the difference is the scale is 982 that is 1182 minus sorry \$200, this is 1 so it is the extra 691 dollar if you want to know the standard deviation of the uniform distribution because this formula is different it is not simple standard deviation so uniform dot std loc is a 200, scale is 982 you will get the standard deviation of 283.47.

Next we will move to the normal distribution here also I have inserted a picture of normal distribution you see that, the picture is taken from this link actual exclamation square break here that link okay when you execute this you will get here picture of probability distribution, that picture. First we will have to import a library norm that is imported from scipy so from scipy dot

stats import norm that is the value, mean, standard deviation: 68, 65.5, 2.5 suppose, if x equal to 68 the mean of that normal distribution is 65.5 standard deviation is 2.5 what is the probability? So, we will run that, first you have to run this also, yes the probability is 0.8413.

If you want to x less than that value suppose, if you want to know cumulative distribution of x greater than value you have to substrate from 1. Suppose if we want to move the value 68 and above. So, already known we know up to 68 this much value so, the remaining area is because we know the area of the normal distribution is 1. So, 1 minus remaining that value will give you the right side area. Suppose if you want to know the value between x1 and x2.

For example value 1 less than or equal to x less than or equal to value 2 so it is a very simple printout norm dot cdf you will find out the upper limit and the lower limit. Simply you type the lower limit because the value ms already I have declared. Now suppose the between 68 and 63, x values 60 and 63 if we want to know the area that it plays a very simple reason to receive a lot of our time. Suppose what is the probability of obtaining a score greater than 700 on your GMAT test that has mean 494 and standard deviation of 100 assume GMAT score are normally distributed.

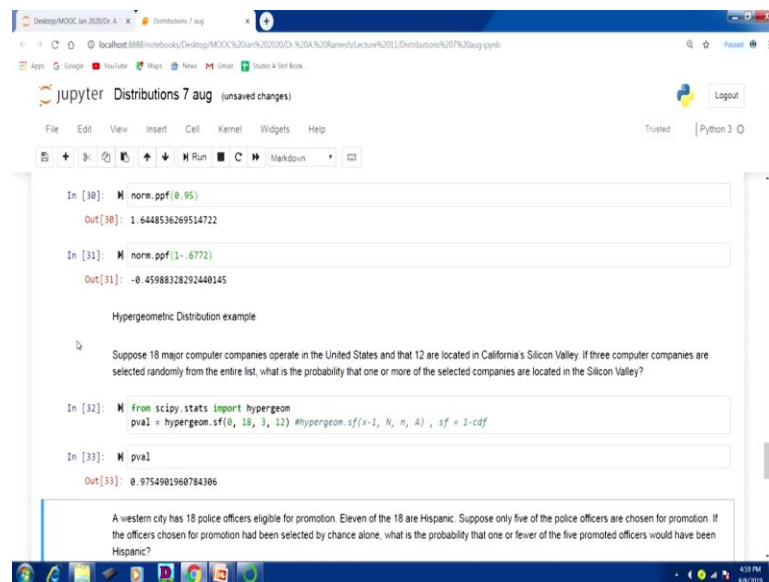
There is another example what is the probability of x greater than 700 when mean equal to 494 and standard deviation is 100. So, because we want to know x greater than equal to 700 so we have to find out x equal to 700 then subtract from 1 so, print 1 minus norm dot cdf 700 the 494, 100 will give you the answer for. what is the probability that randomly drawing his score the 550 or less so we have to need x value less than or equal to 550 so 515, 494, 100 will be the answer.

What is the probability of randomly obtaining a score between 300 to 600 the GMAT examination actually this problem is taken from statistics for management Levin and Reuben. Now you see that the upper limit is 600 lower limit is 300 between 600 and 300 what is the probability so print norm.cdf 600, 494, 100) minus this was upper limit, minus norm.cdf(300 ,494, 100) is the lower limit.

What is the probability of getting a score between 350 and 450 on the same GMAT exam, 450 350 there is another example, similar to previous one into this one. Now we are going to do the reverse of that. Now if they are so far be able to find the cdf cumulative probability. Now suppose the area is given if the area is given we want to know the x value, if it is a standard normal distribution we want to know the z value because the default function is the standard normal distribution where the mean equal to 0 standard deviation 1.

So area under 0.95 the corresponding z value is 1.645 this value you can read it from the table the same way which I have explained in the in my theory lecture. Suppose if you want to know most importantly here norm dot ppf that is a probability function. So, now I am norm ppf 1 - 0.672 will give you the left side area. So, we will see what is the corresponding let us say it is z value is, yeah here we are going in the left hand side so the z value is minus 0.459.

(Refer Slide Time: 15:31)



```

In [30]: norm.ppf(0.95)
Out[30]: 1.6448536269514722

In [31]: norm.ppf(1 - .672)
Out[31]: -0.45988328292440145

Hypergeometric Distribution example

Suppose 18 major computer companies operate in the United States and that 12 are located in California's Silicon Valley. If three computer companies are selected randomly from the entire list, what is the probability that one or more of the selected companies are located in the Silicon Valley?

In [32]: from scipy.stats import hypergeom
pval = hypergeom.sf(0, 18, 3, 12) #hypergeom.sf(k-1, N, n, A), sf = 1-cdf

In [33]: pval
Out[33]: 0.9754901960784306

A western city has 18 police officers eligible for promotion. Eleven of the 18 are Hispanic. Suppose only five of the police officers are chosen for promotion. If the officers chosen for promotion had been selected by chance alone, what is the probability that one or fewer of the five promoted officers would have been Hispanic?

```

Now we will see an example of hyper geometric distribution. the example says suppose 18 major computer companies operate in the United States and that 12 are located in California's Silicon Valley. If 3 computer companies are selected randomly from their entire list what is the probability that one or more of the selected companies are located in the Silicon Valley. What things you have to notice here is 1 or more. So, for that means we have to see what is the probability of getting 1 or more selected companies.

So from `scipy dot stats import hypergeom` p value equal to `hypergeom dot sf`, `sf` means survival function. So, here if it is one or more means that $1 - 1$ so 0.0 , 18 represents the population size 3 means we are 3 we are choosing that is the number of sample 3 , 12 means the number of success in the population that is a capital A , the same notation what you are used in our theory. So, here the p value is 0.9754 .

(Refer Slide Time: 16:45)

```

In [38]: from scipy.stats import hypergeom
pval = hypergeom.sf(0, 18, 3, 12) #hypergeom.sf(x-1, N, n, A) , sf = 1-cdf

In [39]: pval
Out[39]: 0.9754981968784306

A western city has 18 police officers eligible for promotion. Eleven of the 18 are Hispanic. Suppose only five of the police officers are chosen for promotion. If the officers chosen for promotion had been selected by chance alone, what is the probability that one or fewer of the five promoted officers would have been Hispanic?

In [35]: pval = hypergeom.cdf(1, 18, 5, 11)
In [36]: pval
Out[36]: 0.84738562891583275

In [ ]: cdf = 1 - pval

```

We will see another example a western city has 18 police officers eligible for promotion 11 of 18 are Hispanic, suppose only 5 of the police officers are chosen for promotion if the officer chosen for promotion had been selected by chance alone what is the probability that one or fewer of the 5 promoted officers would have been his Hispanic. So, what we need to know here 1 or fewer, so here we have to find out the cumulative probability. So, the formula for finding the cumulative probability for a hyper geometric function is the p value.

I am going to save in the name of p value equal to `hypergeom dot cdf 1`, so 18 represents the population size 5 represents, because choosing 5 , 11 represents the number of success in the population. So, when you run this getting 0.04738 .

(Refer Slide Time: 17:43)

The screenshot shows a Jupyter Notebook interface with the following content:

```
In [40]: pval = hypergeom.cdf(1, 10, 5, 11)

In [36]: pval

Out[36]: 0.04738562091503275
```

Exponential Distribution Example

A manufacturing firm has been involved in statistical quality control for several years. As part of the production process, parts are randomly selected and tested. From the records of these tests, it has been established that a defective part occurs in a pattern that is Poisson distributed on the average of 1.38 defects every 20 minutes during production runs. Use this information to determine the probability that less than 15 minutes will elapse between any two defects?

```
In [ ]: mu1 = 1/1.38 # for 20 mins
mu1

In [ ]: from scipy.stats import expon
expon.cdf(0.75, 0, (1/1.38)) # 15/20 = 0.75, loc=0 because y = (x - loc) / scale, and y= x/scale,
```

We can also define the function manually

```
In [ ]: def CDFExponential(lamb,x): #lamb = Lambda
```

Now we will go for next example on exponential distribution we will take a sample problem. A manufacturing firm has involved in statistical quality control for several years. As part of the production process parts are randomly selected and tested from the records of these tests it has been established that the defective part occur in a pattern that is a Poisson distributed on the average of 1.38 defects every 20 minutes during production run. Use the information to determine the probability of less than 15 minutes will elapse between any 2 defects.

Here how to look at the 2 things in this problem one is the mean of your Poisson distributions given μ and second thing is the between any 2 defects. Now when as I told you in theory itself whenever the between 2 things you have to go for exponential distribution. Now first we do find the mean of your exponential distribution. So, the mean of your exponential distribution is 1 by mean of the Poisson distribution. So, here is Poisson distribution mean is 1.38 so the lambda we can call it as μ_1 that is μ_1 is for the mean of here exponential distribution μ_1 equal to 1 divided by 1.38.

So that value is this much suppose, what was asked probability that is less than 15 minutes from scipy we have to import exponential function. So, we have to find out the cumulative probability further to `expon.cdf` so the 0.75 represents because we got 0.75 from 15 divided by 20 because that is the mean was in the Poisson distribution mean was for 20 minutes. Now the problem for the exponential distribution is asked for 15 minutes.

So, we are dividing 15 by 20 so that the units are matching. So, we need to find out the cumulative function of exponential distribution. The lower limit of that x is 0 the upper limit is 0.75 so `exponent dot cdf` upper limit 0.75, lower limit and the lambda value so you will get the 0.644. Students, is so far we have seen we have seen binomial distribution, how to use Python. Then you have seen Poisson distribution, we have seen uniform distribution.

We have seen normal distribution and exponential distribution and hypergeometric distribution also. So, we will continue in the next class with a new topic that is on sampling and sampling distribution, thank you.