PROBABILITY THEORY FOR DATA SCIENCE

Prof. Ishapathik Das

Department of Mathematics and Statistics

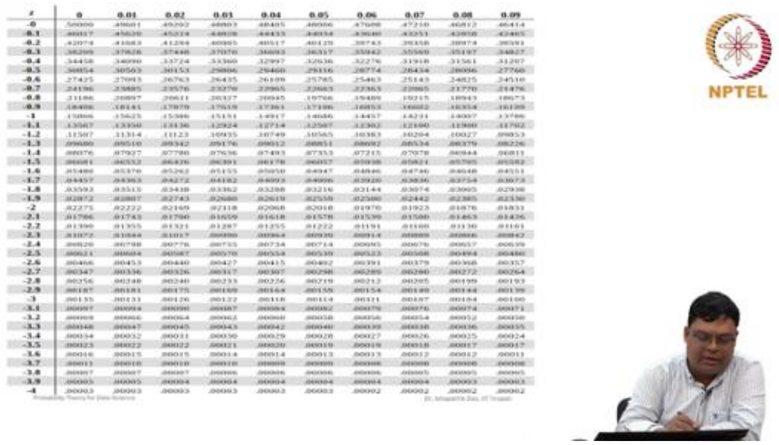
Indian Institute of Technology Tirupati

Week - 06

Lecture - 29

Applications of Normal Distributions and Conditional Distribution Function

0.28814 is the probability. Now we have to add this probability to the others. So, we can find it by first looking at P(Z ≤ -2). This probability can be found from the table. This table shows that for Z = -2, the value is 0.02275.



z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414
-0.1	.46017	.45620	.45224	.44828	.44433	.44038	.43640	.43251	.42858	.42465
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-1	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08692	.08534	.08379	.08226
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-2	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-3	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-4	.00003	.00003	.00003	.00003	.00003	.00003	.00002	.00002	.00002	.00002

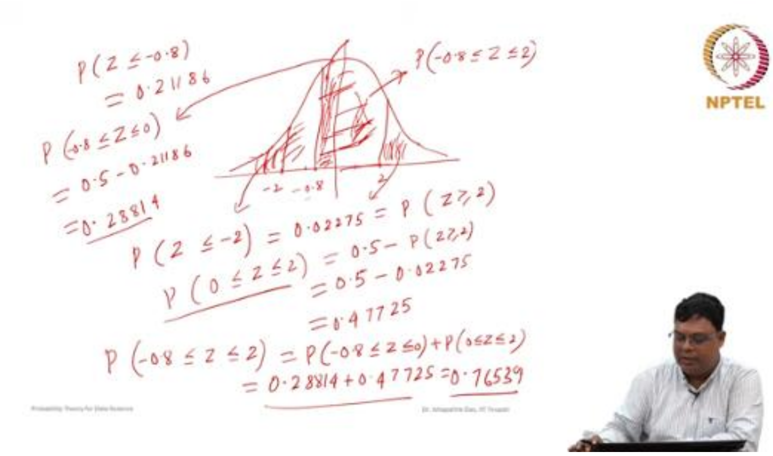
Probability Theory for Data Science

Dr. Ishapathik Das, IIT Tirupati

You can see it here: Z = -2 corresponds to 0.02275. This value is the same as the area we are considering. Now, we want to add this probability to the other probability. So, this probability will be the probability that Z ≥ 0. This is the same as the probability that Z ≥ 2. First, we need to find P(Z ≤ 2). This probability will be nothing but 0.5 - P(Z ≥ 2), because the total probability is 1/2. So, this is 0.5 - P(Z ≥ 2), which is the same as P(Z ≤ -2). This is nothing but 0.02275. So, this is how we can compute it.
We have to just subtract it: 0.5 - 0.02275 = 0.47725.

So, this probability is 0.47725. Now, finally, how will we find it? We find that -0.8 ≤ Z ≤ 2.This will be nothing but the addition of this probability. So, this is nothing but P(-0.8 ≤ Z ≤ 0) + P(0 ≤ Z ≤ 2).

So, this probability is the addition of 0.28814 + 0.47725. So, if you add these two probabilities, then we will get this: 0.28814 + 0.47725 = 0.76539. So, please check if this calculation and computation are correct. We're just doing the computation, so there may be some mistakes. You can check if there’s any numerical mistake, but the concept is how to find the probability. Now, the second question is: What is the second question? The second question is to find the probability for X ≤ 26 and X ≤ 46. So, what is the final value? It is 0.76539.



$P(Z \le -0.8)$
$= 0.21186$

$P(-0.8 \le Z \le 0)$
$= 0.5 - 0.21186$
$= 0.28814$

$P(-0.8 \le Z \le 2)$

$P(Z \le -2) = 0.02275 = P(Z \geq 2)$

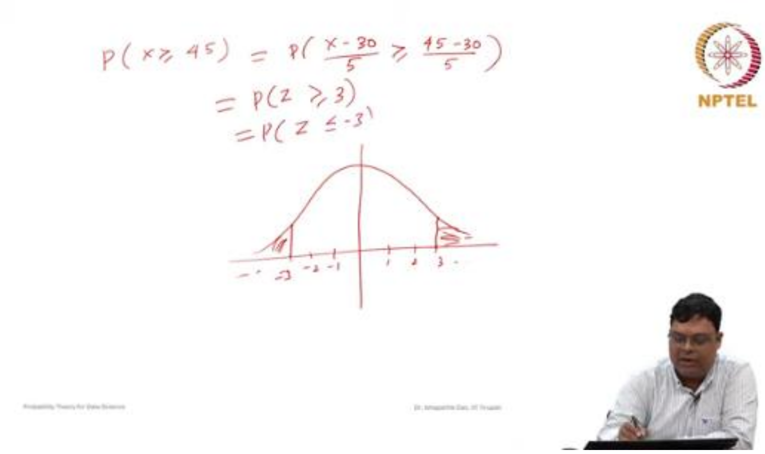
$P(0 \le Z \le 2) = 0.5 - P(Z \geq 2)$
$= 0.5 - 0.02275$
$= 0.47725$

$P(-0.8 \le Z \le 2) = P(-0.8 \le Z \le 0) + P(0 \le Z \le 2)$
$= 0.28814 + 0.47725 = 0.76539$

Now, next, we have to find the probability that X ≥ 45. So, let’s see, what is P(X ≥ 45)? So, now we have to apply this transformation, similar to that. So, X = 35, and we have to subtract 30 from X and then divide by 5. This is nothing but P((X - 30) / 5).

This is the mean that was given, and this is ≥ (45 - 30) / 5. So, this is the standard normal variable Z. This is ≥ (45 - 30), which is 15, and 15 / 5 is 3. So, this is nothing but P(Z ≥ 3). Now, in this standard normal curve, we have 1, 2, 3, and on the other side, we have -1, -2, -3.

So now, P(Z ≥ 3) is asked. This is the same as P(Z ≤ -3), by symmetry. So, P(Z ≤ -3), we will find from the table. P(Z ≤ -3) = 0.00135. So, this value will be 0.00135, as you can see here for -3.



$$P(X \geq 45) = P\left(\frac{X-30}{5} \geq \frac{45-30}{5}\right)$$
$$= P(Z \geq 3)$$
$$= P(Z \leq -3)$$

So, hopefully, you are understanding and following how we can use this table for the standard normal random variate, and how we can compute the probability for any normal distribution with any general mean and variance. So, this is how—what is this? P(Z ≤ -3). This is nothing but 0.00135. So, this value will be 0.00135.

That's the answer to the question: P(X ≥ 45), which is 0.00135. Now, the next problem we will discuss is: What is the probability? Let's see what the problem is. So now, we have to find the probability that X > 30, or |X| > 30 is greater than 5. That's what we need to find out.

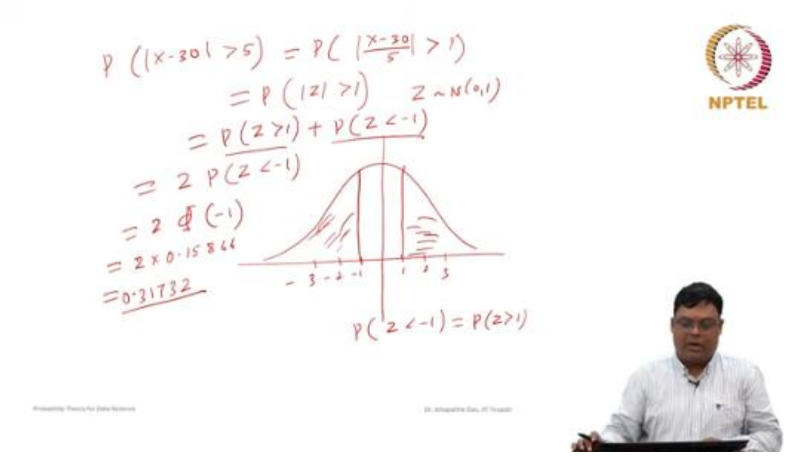
So, the probability that |X| > 30 and > 5. Now, how can we find this? The question is the probability that |X - 30| > 5. So, this is the same as saying, how can we compute it? It's essentially the absolute value of this.

So, if you divide by 5, the probability that... What is the transformation here? Z = (X - 30) / 5. So, basically, if you take (X - 30) / 5, and then take the absolute value of this, it's the same as dividing by 5, and it's > 1. Now, (X - 30) / 5 is nothing but |Z| > 1, where Z is a standard normal variable. Now, |Z| > 1 is the same as 1 - P(|Z| ≤ 1).

So, how can we find P(|Z| ≤ 1)? First, let’s check this graph. It will make it clearer to understand. Now, |Z| > 1 is essentially... one minute. So, suppose this is 1, this is 2, this is 3, and this is 3, this is 1, this is 2, and this is 3. So, when we say |Z| > 1, we’re referring to the area where Z > 1, or Z < -1.

So, this is -1, -2, -3. So, |Z| > 1 means P(Z > 1) + P(Z < -1). You can include the equality as well, since it’s a continuous random variable. Now, by symmetry, P(Z < -1) = P(Z > 1). So, we can write this as 2 * P(Z < -1), since by symmetry, P(Z < -1) = P(Z > 1). That’s why it’s 2 * P(Z < -1). This is nothing but 2 * Φ(-1), where Φ is the cumulative distribution function of the standard normal variate. So, it becomes 2 * Φ(-1). This value of Φ at -1 can be found from the table, which gives P(Z < -1). So, from the table, you can find the value for -1.

This value is 0.15866. So, 0.15866 is the probability. You can see here that the probability is 0.15866. Just remember that it is 0.15866. So, this is 2 * 0.15866. So, you just have to multiply it. This is 0.31732. So, this is the probability we need to calculate. We have found this. We just discussed two examples, which are very similar.



$$P(|X-30| > 5) = P\left(\left|\frac{X-30}{5}\right| > 1\right)$$
$$= P(|Z| > 1) \quad Z \sim N(0,1)$$
$$= \underline{P(Z > 1)} + \underline{P(Z < -1)}$$
$$= 2P(Z < -1)$$
$$= 2\Phi(-1)$$
$$= 2 \times 0.15866$$
$$= \underline{0.31732}$$

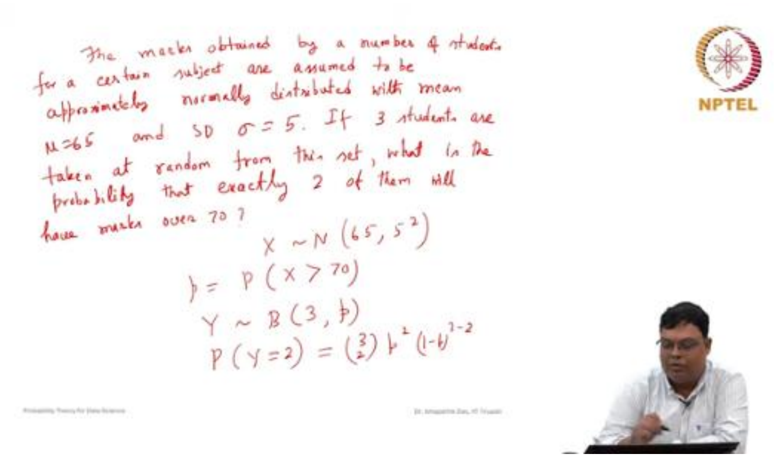
$$P(Z < -1) = P(Z > 1)$$

Through these, we've learned how to compute probabilities using the normal distribution with any mean and variance. We can discuss more examples as well. Let's move on to another numerical example. Suppose the marks obtained by a number of students in a certain subject are assumed to be normally distributed, with a mean of 65 and a standard deviation of 5. If three students are selected at random from this group, what is the probability that exactly two of them will have marks greater than 70? So, this is the question.

The marks obtained by a number of students for a certain subject are assumed to be approximately normally distributed, with a mean of 65 and a standard deviation of 5. If three students are taken at random from this set, what is the probability that exactly two of them will have marks > 70? See, here we have to use not only the normal distribution but also the binomial distribution. This is because in the last part, suppose you know that the probability, where X is the random variable representing the marks obtained by the students. It is given that $X \sim N(65, 25)$.

Now, p is the probability that any student will obtain marks > 70. Now, it is asked the question that Y, out of 3 students taken at random, Y is taken at random from this set. What is the probability that exactly 2? So, $Y \sim \text{Binomial}(n = 3, p)$. This total number of students here is 3, it is taken, and p is this probability.

This p we have to find. Then the question is exactly 2. What is the probability that out of 3, what is the probability that $Y = 2$, that 2 will be exactly over 70? So, this probability will be $C(3, 2) * p^2 * (1 - p)^1$. This is the probability we discussed in the binomial distribution. So, this probability we have to compute. So, for computing this probability, first, we have to find what is the value of p. Let us find out this value of p first. So, this p can be found as $P(X > 70)$.



The marks obtained by a number of students for a certain subject are assumed to be approximately normally distributed with mean $\mu = 65$ and SD $\sigma = 5$. If 3 students are taken at random from this set, what is the probability that exactly 2 of them will have marks over 70?

$$X \sim N(65, 5^2)$$

$$p = P(X > 70)$$

$$Y \sim B(3, p)$$

$$P(Y=2) = \binom{3}{2} p^2 (1-p)^{3-2}$$

So, this is nothing but equal to, now because X has a mean of 65. It is given that X ~ N(65, 25), where the standard deviation is 5. So, then Z = (X - 65) / 5, it will be N(0, 1). So, we will use the similar concept here. So, p = P(X > 70), which is equal to P((X - 65) / 5 > (70 - 65) / 5).

This is equal to P(Z > 1). So, now, that probability, we've done this several times. So, because this is 1, this is 2, 3, and so on. Similarly, -1, -2, -3. Now, P(Z > 1) means it is asking what is the probability here.

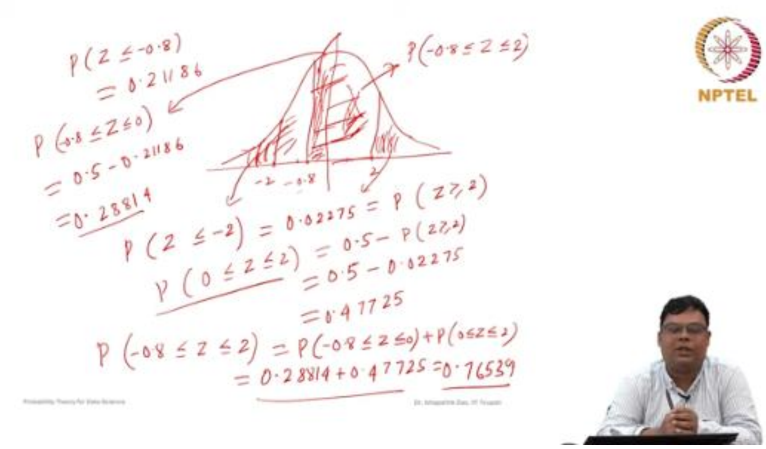
This is the same as by symmetry, P(Z < -1). So, this is the probability that P(Z < -1). We are using this negative number because from the table, we can find this value. So, P(Z < -1) is 0.15866. This value is 0.15866.

So, here it is 0.15866. Then we know this probability p. Now, we have to find the probability that Y = 2. It is asked what the probability is that exactly 2 students have marks > 70 out of 3 students. So, this is simply the probability that Y = 2. This is C(3, 2) * $(0.15866)^2$ * $(1 - 0.15866)^1$ * $(1 - 0.15866)^1$. This is 1. Now, C(3, 2) = 3. This is $(0.15866)^2$, and you need to calculate this as well. We will use a calculator to compute this.

I can tell you the value since it has already been computed. It is 0.06357. This is the value. Please check that. There may be some decimal places, and it might be approximated.

So, this is the value. So, you can check it. Now, we have already learned about some of the important discrete distribution functions. There are many other important distribution functions beyond what we have discussed here. Just some of the important distribution functions, like the binomial distribution, are frequently used.

We also discussed the Poisson distribution function. There are some other distributions, like the hypergeometric distribution, geometric distribution, and negative binomial distribution, which we did not discuss here. If we get time, we will discuss them in the later part. Now, for continuous distributions, we discussed the uniform distribution function, the exponential distribution function, the gamma distribution function, and we have also discussed the normal distribution function. We have also discussed the uses and applications of these distribution functions.



$P(Z \le -0.8)$
$= 0.21186$

$P(-0.8 \le Z \le 0)$
$= 0.5 - 0.21186$
$= 0.28814$

$P(-0.8 \le Z \le 2)$

$P(Z \le -2) = 0.02275 = P(Z \ge 2)$

$P(0 \le Z \le 2) = 0.5 - P(Z \ge 2)$
$= 0.5 - 0.02275$
$= 0.47725$

$P(-0.8 \le Z \le 2) = P(-0.8 \le Z \le 0) + P(0 \le Z \le 2)$
$= 0.28814 + 0.47725 = 0.76539$

So, this is nothing but equal to, now because X has a mean of 65. It is given that X ~ N(65, 25), where the standard deviation is 5. So, then Z = (X - 65) / 5, it will be N(0, 1). So, we will use the similar concept here. So, p = P(X > 70), which is equal to P((X - 65) / 5 > (70 - 65) / 5).

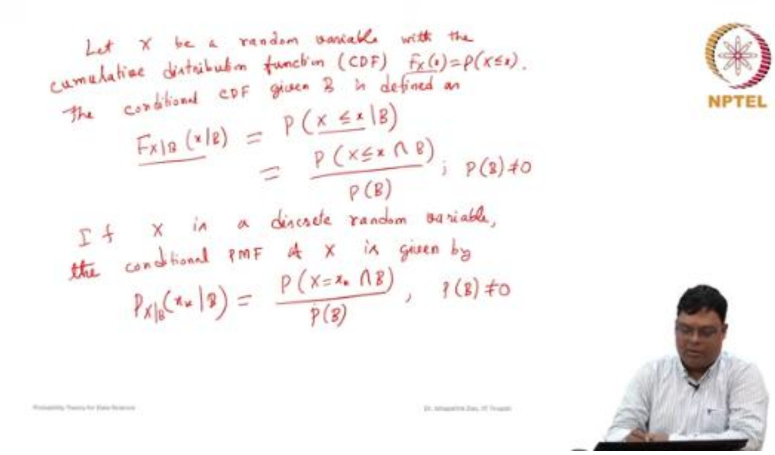
This is equal to P(Z > 1). So, now, that probability, we've done this several times. So, because this is 1, this is 2, 3, and so on. Similarly, -1, -2, -3. Now, P(Z > 1) means it is asking what is the probability here.

This is the same as by symmetry, P(Z < -1). So, this is the probability that P(Z < -1). We are using this negative number because from the table, we can find this value. So, P(Z < -1) is 0.15866. This value is 0.15866.

So, here it is 0.15866. Then we know this probability p. Now, we have to find the probability that Y = 2. It is asked what the probability is that exactly 2 students have marks > 70 out of 3 students. So, this is simply the probability that Y = 2. This is C(3, 2) * $(0.15866)^2$ * $(1 - 0.15866)^1$ * $(1 - 0.15866)^1$. This is 1. Now, C(3, 2) = 3. This is $(0.15866)^2$, and you need to calculate this as well. We will use a calculator to compute this.

I can tell you the value since it has already been computed. It is 0.06357. This is the value. Please check that. There may be some decimal places, and it might be approximated. So, this is the value. So, you can check it. Now, we have already learned about some of the important discrete distribution functions. There are many other important distribution functions beyond what we have discussed here. Just some of the important distribution functions, like the binomial distribution, are frequently used.

We also discussed the Poisson distribution function. There are some other distributions, like the hypergeometric distribution, geometric distribution, and negative binomial distribution, which we did not discuss here. If we get time, we will discuss them in the later part. Now, for continuous distributions, we discussed the uniform distribution function, the exponential distribution function, the gamma distribution function, and we have also discussed the normal distribution function. We have also discussed the uses and applications of these distribution functions.

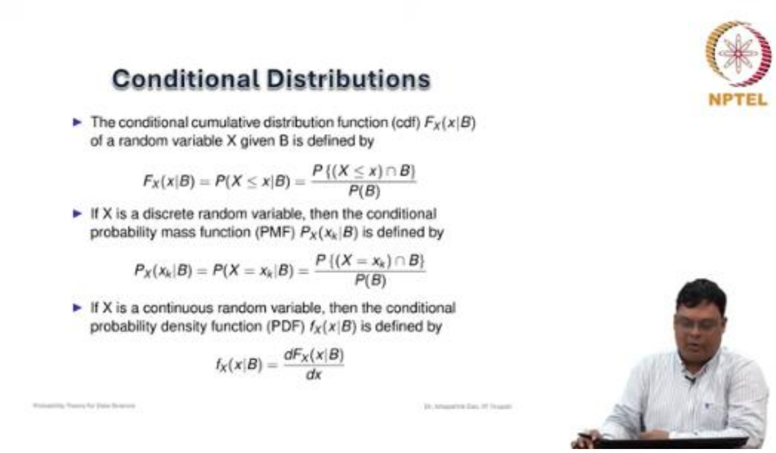


Let X be a random variable with the cumulative distribution function (CDF) $F_X(x)=P(X\le x)$.
The conditional CDF given B is defined as

$$F_{X|B}(x|B) = P(X \le x | B) = \frac{P(X\le x \cap B)}{P(B)}; \quad P(B)\neq 0$$

If X is a discrete random variable, the conditional PMF of X is given by

$$P_{X|B}(x_k|B) = \frac{P(X=x_k \cap B)}{P(B)}, \quad P(B)\neq 0$$



Conditional Distributions

- The conditional cumulative distribution function (cdf) $F_X(x|B)$ of a random variable X given B is defined by

$$F_X(x|B) = P(X \leq x|B) = \frac{P\{(X \leq x) \cap B\}}{P(B)}$$

- If X is a discrete random variable, then the conditional probability mass function (PMF) $P_X(x_k|B)$ is defined by

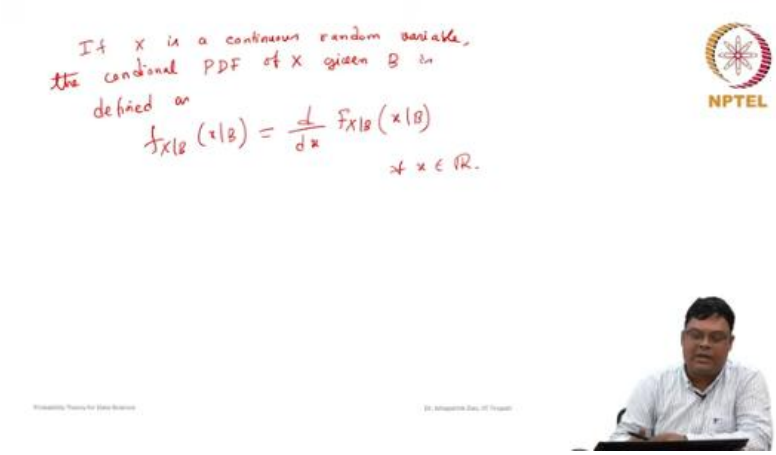
$$P_X(x_k|B) = P(X = x_k|B) = \frac{P\{(X = x_k) \cap B\}}{P(B)}$$

- If X is a continuous random variable, then the conditional probability density function (PDF) $f_X(x|B)$ is defined by

$$f_X(x|B) = \frac{dF_X(x|B)}{dx}$$

If X is a continuous random variable, then the conditional probability density function (PDF) of X given an event B is defined as f(X | B). Because it is a conditional probability density function, it is simply the derivative of the conditional cumulative distribution function (CDF). It is assumed that the conditional CDF is differentiable for continuous random variables, and this derivative defines the conditional probability density function. So, this f(X | B) is defined here.

First, you would find this, then take the derivative to find the conditional probability density function of a random variable for any X ∈ ℝ. So, how it can be utilized? Let us do a numerical example so we can get an idea of how it can be used. Let us consider a numerical example. First, we will discuss the continuous case, and then we will discuss the discrete case.



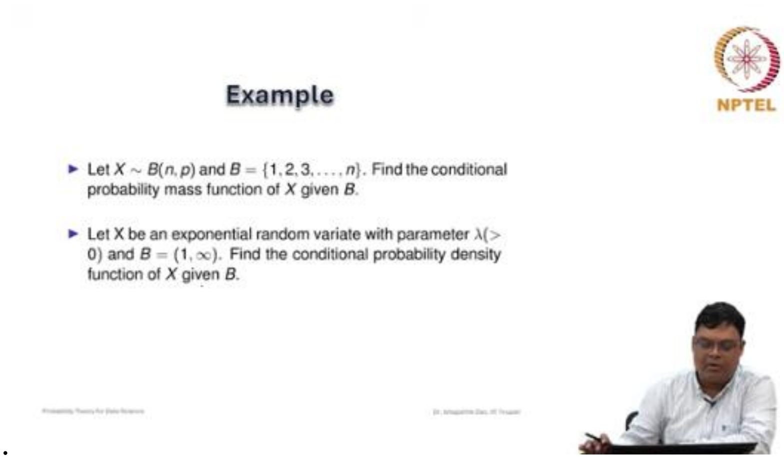
If X is a continuous random variable, the condional PDF of X given B is defined as

$$f_{X|B}(x|B) = \frac{d}{dx} F_{X|B}(x|B)$$

$$\forall x \in \mathbb{R}.$$

Let X be an exponential random variable with parameter λ > 0, and let B be given. Find the conditional probability density function of X given B. So, how can we do that? Let X be an exponential random variable. To find the conditional probability density function, you first need to find the cumulative distribution function (CDF).

For the discrete case, this is not required; we can directly find the probability mass function (PMF). However, in this case, we first need to find the CDF. In the discrete case, we will also discuss the density function. For continuous cases, we first need to find the CDF. Let's do that now.



Example

- Let $X \sim B(n, p)$ and $B = \{1, 2, 3, \ldots, n\}$. Find the conditional probability mass function of X given B.
- Let X be an exponential random variate with parameter $\lambda (> 0)$ and $B = (1, \infty)$. Find the conditional probability density function of X given B.

Probability Theory for Data Science

Dr. Ishapathik Das, IIT Tirupati

.

It is given that...