

Probability and Statistics
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture – 30
Problems on Normal Distribution

(Refer Slide Time: 00:21)

The cdf of Z is denoted by

$$\Phi(z) = \int_{-\infty}^z \phi(t) dt = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

We have $1 - \Phi(z) = \Phi(-z)$ for all z

$$\Rightarrow \Phi(-z) + \Phi(z) = 1$$

So $\Phi(0) = \frac{1}{2}$

Consider $X \sim N(\mu, \sigma^2)$

$$P(a < X \leq b) = F_X(b) - F_X(a)$$
$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

One more point here is that since the values of the c d f can be evaluated using phi of minus z, plus phi of z is equal to 1 therefore, many times the tables are table head at only for either positive arguments of z or negative arguments of z.

(Refer Slide Time: 00:42)

$$F_X(x) = P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right)$$
$$= P\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

$P(X \leq b), \quad P(X > a), \quad P(a < X < b)$
 $1 - P(a < X < b)$

$P(|Z| < 0.5) = P(-0.5 < Z < 0.5)$
 $= \Phi(0.5) - \Phi(-0.5)$
 $= 1 - 2\Phi(-0.5)$
 $= 1 - 2(0.3085) = 1 - 0.617 = 0.3830$

A normal distribution curve is drawn with a vertical line at the center and a shaded area between -0.5 and 0.5 standard deviations from the mean.

(Refer Slide Time: 00:45)

Examples. 1. Assume that the time required for a distance runner to run a mile is a normal r.v. with parameters $\mu = 4 \text{ min, } 1 \text{ sec}$ and $\sigma = 2 \text{ sec}$. What is the prob. that this athlete will run the mile in less than 4 min. ? In more than 3 min, 55 sec. ?

Solⁿ: $X \sim N(241, 4)$. $X \rightarrow$ time in sec.

$P(X < 240) = P\left(\frac{X-241}{2} < \frac{240-241}{2}\right) = P(Z < -0.5)$
 $= \Phi(-0.5) = 0.3085$

$P(X > 235) = P(Z > -3) = \Phi(3) = 0.9987$

Assume that the high price of a commodity tomorrow is a normal r.v. with $\mu = 10$, the shortest interval that has prob. ... tomorrow's high price for this

So, suppose that the time required for a distance runner to run a mile is a normal random variable with parameters μ is equal to 4 minute 1 second, and a standard deviation 2 seconds. What is the probability that this athlete will run the mile in less than 4 minutes or in more than 3 minutes 55 seconds.

If we consider X as the time required and we consider it in the time measured in seconds, then X will follow normal distribution with mean 241 seconds, here 4 minutes 1 second is 241 seconds and sigma square is 4 seconds. So, what is the probability of running in

less than 4 minutes; that means, X is less than 240 what is the probability of this event? So, utilizing this relationship we can write this as X minus 241 by 2, less than 240 minus 241 by 2, which is probability Z less than minus 0.5, which is Φ of minus 0.5. So, this value we will see from the tables of the standard normal c d f, and this value turn out to be 0.3085. Most of the tables are given up to 4 or 5 decimal points, and we can also use a numerical integration rules such as Simpsons one-third rule etcetera to evaluate this.

Similarly, if we want to calculate, what is the probability that he will run in more than 3 minute 55 seconds, then probability X greater than 235, that is probability Z greater than minus 3, that is if we put 235 minus 241 divided by 2, so it becomes Z greater than minus 3. So, if we look at the shape of the distribution minus 3 suppose here then 3 is here. So, this is equivalent to Φ of 3, which is 0.9987, which is extremely high probability. So, here you can see is that mean is 4 minute 1 second; that means, almost surely he will complete the race within in more than 3 minute 55 seconds, this also brings out another important property of the normal distribution.

The normal distribution is having high concentration of probability in the center, since you are having in the density function $e^{-z^2/2}$, the terms go rapidly towards 0, the convergence towards 0 is very fast therefore, in the range around the mean μ most of the probability is concentrated. So, if we consider say probability of say modulus Z less than 0.5, that is minus 0.5 less than Z less than 0.5; here Z denotes the a standard normal distribution. So, this is equal to Φ of 0.5 minus Φ of minus 0.5. So, we can write because the value of 1 of them needs to be seen, we need not see both of them either we see Φ of 0.5 or Φ of minus 0.5, and the other 1 we can write in terms of 1 minus that. So, this we can write as 1 minus Φ of minus 0.5, so this becomes 1 minus twice, the value of Φ of minus 0.5 was 0.30. So, this is equal to 1 minus 2 into 0.3085 that is equal to 1 minus 0.6170 that is equal to point; that means, within a very short distance that is minus 0.5 to 5 itself almost 40 percent of the probability is concentrated.

(Refer Slide Time: 05:50)

$$F(x) = P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right)$$

$$= P\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

$P(X \leq b), \quad P(X > a), \quad P(a < X < b)$

$1 - P(a < X < b) \quad P(\mu - 3\sigma < X < \mu + 3\sigma) > 0.99$

$P(|Z| < 0.5) = P(-0.5 < Z < 0.5)$

$$= \Phi(0.5) - \Phi(-0.5)$$

$$= 1 - 2\Phi(-0.5)$$

$$= 1 - 2(0.3085) = 1 - 0.6170$$

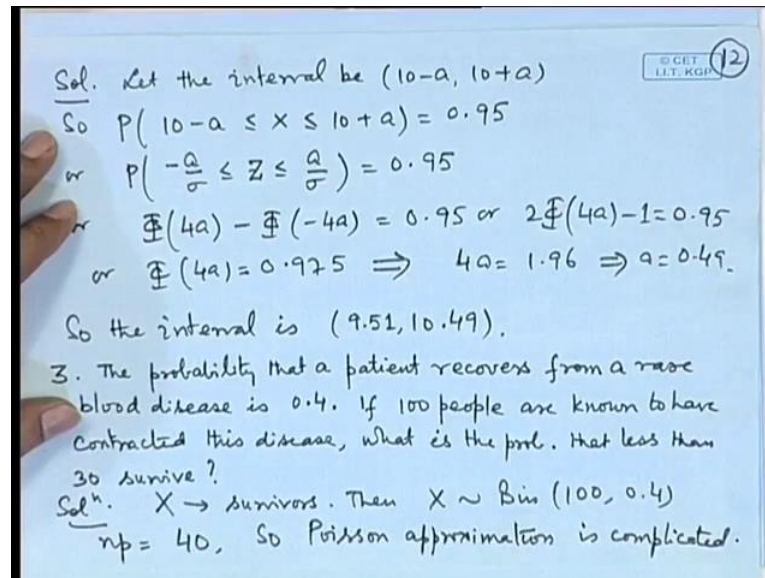
$$= 0.3830$$

A normal distribution curve is drawn with the x-axis labeled from -3 to 3. The area between -1 and 1 is shaded and labeled 60%. The area between -2 and 2 is shaded and labeled 90%. The area between -3 and 3 is shaded and labeled 99.99%.

If we consider say minus 1 to 1, this consists of all most 60 percent of the probability, if we consider minus 2 to 2, this consists of more than 90 percent of the probability, if we consider minus 3 to 3, this consists of more than 99.99 percent of the probability. So, in terms of mu sigma, this means that mu minus 3 sigma less than x, less than mu plus 3 sigma is greater than 0.99, these are known as 3 sigma limits, then minus 2 to 2 is called as 2 sigma limits. So, in the industrial applications where certain product requirements such as the width of certain bolts produced, diameters of certain nuts etcetera are various kind of quality control features which are implied in industry, if they follow the normal distribution, then the industrial standards specify that in order that the product we defined as a good item or proper item, the specification should be within 3 sigma limits. So, in industrial standards these things are quite useful.

Let us consider another application, assume that the high price of a commodity tomorrow is a normal random variable with mu is equal to 10 and sigma is equal to 1 by 4. What is the probable, what is the shortest interval that has probability 0.95 of including tomorrow's high price for this commodity? Now from the properties of the normal distribution that we discussed just now, we should consider here the interval to be a symmetrical interval around the mean, because we are requiring a shortest interval. So, shortest interval will be symmetric around these because if we take non symmetric interval then this will become slightly longer because the tail is converging faster, there is more concentration of the probabilities towards the center.

(Refer Slide Time: 08:29)



Sol. Let the interval be $(10-a, 10+a)$
So $P(10-a \leq X \leq 10+a) = 0.95$
or $P\left(-\frac{a}{\sigma} \leq Z \leq \frac{a}{\sigma}\right) = 0.95$
or $\Phi(4a) - \Phi(-4a) = 0.95$ or $2\Phi(4a) - 1 = 0.95$
or $\Phi(4a) = 0.975 \Rightarrow 4a = 1.96 \Rightarrow a = 0.49$
So the interval is $(9.51, 10.49)$.

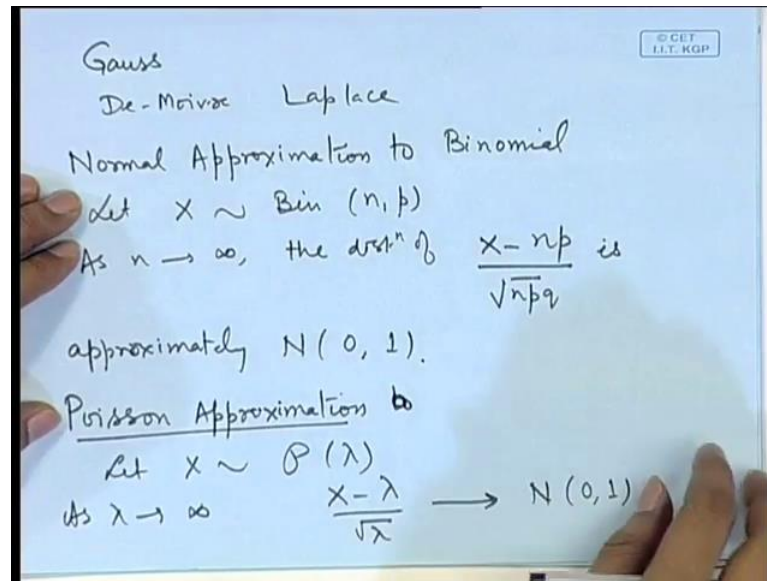
3. The probability that a patient recovers from a rare blood disease is 0.4. If 100 people are known to have contracted this disease, what is the prob. that less than 30 survive?
Solⁿ. $X \rightarrow$ survivors. Then $X \sim \text{Bin}(100, 0.4)$
 $np = 40$, so Poisson approximation is complicated.

So, we can consider since here the mean is 10, we can consider the interval of the form 10 minus a to 10 plus a. So, we want the value of a, such that the probability of X lying in this interval 10 minus a to 10 plus a is 0.95. So, consider the transformation X minus mu by sigma. So, here mu is 10 here, so X minus mu by sigma. So, it becomes minus a by sigma less than or equal to Z less than or equal to a by sigma, and sigma is 1 by 4, so this is 4 a. So, this is becoming phi of 4 a, minus phi of minus 4 a is equal to 0.95, what is the value of a for which this is satisfying?

Once again we utilize the relation between capital phi of X and capital phi of minus X. So, this becomes twice phi of 4 a minus 1 is equal to 0.95 and this gives us phi of 4 a is equal to 0.975 and from the tables of the normal distribution we can see that the point up to which the probabilities 0.975 is 1.96. So, after evaluation a, becomes 0.49 and therefore, the interval 10 minus a to 10 plus a reduces to 9.51 to 10.49. So, if the mean price is 10, and the standard deviation is 0.25 then the interval which will have the high price with probability 0.95 is 9.5 to 10.5 approximately, which is basically two sigma because here sigma is 0.25. So, two sigma becomes 0.5. So, around 10 the interval of 0.25 is. So, in two sigma limits we have more than 95 percent of the probability here.

Another important point which I mentioned was that the origin of the normal distribution, the normal distribution was derived as a distribution of the errors by gauss.

(Refer Slide Time: 10:58)



So, he was considering astronomical measurements, now he did not consider one measurements, you consider several measurements and consider the average of those measurements to consider that as the estimate of the actual measurement. So, since in each measurements some error will be concentrated and therefore, if we look at the distribution of the errors, gauss observed that it is normal distribution that is why it was called normal distribution is also called the law of errors are the error distribution etcetera, and it turns out that the sample mean or the sample sum is normally distributed; however, even apart from the gauss mathematician such as De-Moivre and Laplace etcetera they obtained the normal distribution as an approximation to binomial distribution or distribution of poisson approximated to normal distribution.

So, let us look at it normal approximation to binomial. So, let X follow binomial n, p as n tends to infinity, the distribution of X minus np divided by root npq is approximately normal 0 . Similarly poisson approximation to; so let X follow poisson λ distribution, as λ tends to infinity X minus λ by root λ this converges to normal $0,1$. Basically these were the original central limit theorems, the modern versions are for any random variable X .

So, let us look at applications of this hear, the probability that a patient recovers from a rare blood disease is 0.4 , if 100 people are known to have contracted this disease, what is the probability that less than 30 will survive? So, if we consider here X as the number of

survivors, then X follows binomial 100, 0.4 and we are interested to calculate the probability that X is less than 30.

(Refer Slide Time: 14:21)

$\mu = np = 40$ $\sigma = \sqrt{npq} = 4.899$

$P(X < 30) \approx P(X \leq 29.5) = P\left(Z \leq \frac{29.5 - 40}{4.899}\right)$

$= P(Z \leq -2.14) = \Phi(-2.14) = 0.0162.$

4. Suppose home burglaries occur in a town like events in a Poisson process with $\lambda = \frac{1}{2}$ per day. Find the prob. that no more than 10 burglaries occur in a month? Not less than 17 in a month?

Solⁿ $X \sim P(15)$ if $X \rightarrow$ no of burglaries in a month.

$P(X \leq 10) \approx P(X \leq 10.5)$ $X \approx N(15, 15)$

$= P\left(Z \leq \frac{10.5 - 15}{\sqrt{15}}\right) = \Phi(-1.16) = 0.123$ (Exact value 0.118)

$P(X \geq 17) \approx 1 - P(X \leq 16) \approx 1 - P(Z \leq 0.39) = 1 - \Phi(0.39)$

$= 1 - \Phi(0.39) = 0.3483.$ (Exact value is 0.336).

Now, if we look at the actual calculation of this using the binomial distribution, then this is reducing to $\sum_{j=0}^{29} \binom{100}{j} (0.4)^j (0.6)^{100-j}$. You can look at the difficulty of the evaluations are the complexity of the terms here, we have factorials involving $100 \text{ C } 0$, $100 \text{ C } 10$, $100 \text{ C } 20$, $100 \text{ C } 25$ etcetera, and then the powers of numbers which are a smaller than 1. So, the large powers of these numbers will yield lot of computational errors and also the terms will be complex.

However here we can see that n is large. So, if we can consider if we try to look at np is equal to 40 and we want to use poisson approximation, then that will also very complicated because that will involve the same summation e to the power minus 40, 40 to the power j by j factorial, which again involves large terms here. So, in place of that we will use the normal approximation. So, np is 40 and npq is 24. So, square root of that is 4.89. So, probability of X less than 30; now here what we do we apply so called continuity corrections, this continuity corrections is required to approximate a discrete distribution with a continuous distribution consider like this. In the binomial distribution the curve is like this, now if you are approximating it by a normal distribution and suppose this is 30.

So, it could be also less than 29; X less than or equal to 29. So, in that case if we had approximated it by normal we should have written X less than or equal to 29 whereas, here if we write straightaway we will write X less than or equal to 30, because in continuous distribution the probability of a point is negligible. So, a better thing would be to take a middle value between 29 and 30 as 29.5. So, this is called continuity corrections.

Now, we make use of the fact that this X minus μ by σ is approximately normal. So, $29.5 - 40$ divided 4.899 , this is Z less than or equal to -2.14 , that is the cdf of the standard normal distribution at the point -2.14 . So, from the tables of the normal distribution we can observe it is 0.0162 . So, the probability is quite a small that less than 30 will survive.

Let us look at another example where the Poisson distribution is approximated by normal distribution. So, consider a large town where the home burglaries occur like events in a poisson process with λ is equal to half per day that is a rate, find the probability that no more than 10 burglaries occur in a month or not less than 17 occur in a month? So, if we consider this then X follows poisson 15, if I denote by X the number of burglaries occurring in a month's time. So, since λ is half per day, in a month we assume 30 days, so the parameter λ will become 15. So, probability X less than or equal to 10, again we make use of the continuity corrections since it is less than or equal to 10 it is also same as probability X less than 11.

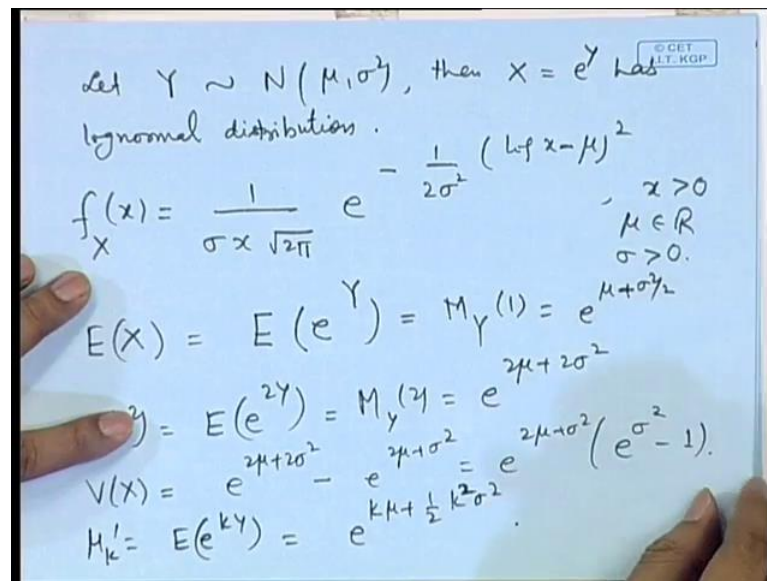
So, in normal distribution it could have been calculated as X less than or equal to 10 or X less than or equal to 11. So, as a continuity correction we take the midpoint X less than or equal to 10.5. So, if we make use of the normal approximation here, then X minus λ by $\sqrt{\lambda}$ is approximately normal as λ becomes large. So, here it is 15 here, so $10.5 - 15$ by $\sqrt{15}$, which is after simplification -1.16 . So, the value of the normal distributions cumulative distribution function at this point is 0.123 .

Similarly, not more less than 17 in a month that is probability X greater than or equal to 17, which is $1 - \text{probability } X \text{ less than or equal to } 16$ or $1 - \text{probability } X \text{ less than or equal to } 16$ by 0.5 that is the continuity correction, and after shifting by 15 and dividing by $\sqrt{15}$ it becomes 0.39 , which we can see from the tables of the normal distribution and the 0.3483 is the value. If we compare with the exact value which we

could have calculated from the e to the power minus 15, 15 to the power j by j factorial, in the first one you would have done j is equal to 0 to 10, then this value is actually 0.118. So, you can see that there is a very small margin of error even for lambda is equal to 15, in the second one the value is 0.3483 by approximation and the exact value is 0.336. So, the error margin is extremely a small.

A related distribution to normal distribution is called log normal distribution.

(Refer Slide Time: 21:01)



Let $Y \sim N(\mu, \sigma^2)$, then $X = e^Y$ has lognormal distribution.

$$f(x) = \frac{1}{\sigma x \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2} (\log x - \mu)^2} \quad \begin{array}{l} x > 0 \\ \mu \in \mathbb{R} \\ \sigma > 0. \end{array}$$

$$E(X) = E(e^Y) = M_Y(1) = e^{\mu + \frac{\sigma^2}{2}}$$

$$E(e^{2Y}) = M_Y(2) = e^{2\mu + 2\sigma^2}$$

$$V(X) = e^{2\mu + 2\sigma^2} - e^{2\mu + \sigma^2} = e^{2\mu + \sigma^2} (e^{\sigma^2} - 1)$$

$$M'_k = E(e^{kY}) = e^{k\mu + \frac{1}{2}k^2\sigma^2}$$

So, if we say Y follows normal mu sigma square then X is equal to e to the power Y has log normal distribution. So, the density function of log normal distribution will be given by 1 by sigma x root 2 pi, e to the power minus 1 by 2 sigma square log x minus mu square, here x is positive and mu and sigma is usual. The necessity of this kind of distribution can be understood like this that many times x observations may be very large and log x observations may be useful; on the other hand if Y observations are very small then e to the power Y observations may be of the reasonable size. So, if we take e to the power Y and then that will have a log normal distribution. The moments of the log normal distribution are obviously, in the form of the moment generating function of. So, this is nothing, but the moment generating function of Y at 1 that is e to the power mu plus sigma square by 2.

If we consider the second moment then it is equal to expectation of e to the power 2 y that is M y 2 that means, variance of X is equal to e to the power 2 mu plus 2 sigma

square. In general μ_k' is equal to expectation of e to the power k y that is equal to e to the power $k\mu + \frac{1}{2}k^2\sigma^2$. So, moments of all orders of the log normal distribution exists, and they can be expressed in terms of the moments of the moment generating function of the normal distribution.

(Refer Slide Time: 23:44)

5. The demand X of a certain item follows a log-normal distribution with mean 7.43 and variance 0.56. Find $P(X > 8)$.

Solⁿ $\mu_1' = e^{\mu + \frac{\sigma^2}{2}} = 7.43 \Rightarrow \mu + \frac{\sigma^2}{2} = 2.0055 \dots (1)$

$\mu_2' = e^{2\mu + 2\sigma^2} = 0.56 + (7.43)^2 = 55.7649$

$\Rightarrow 2\mu + 2\sigma^2 = 4.0211$ or $\mu + \sigma^2 = 2.0106 \dots (2)$

(1) & (2) $\Rightarrow \mu = 2.0, \sigma = 0.10$

$P(X > 8) = P(\log_e X > \log_e 8) = P(Z > \frac{\log_e 8 - 2}{0.1})$

$= P(Z > 0.79) = \Phi(-0.79) = 0.2148.$

Let us look at one application here; suppose the demand of a certain item follows a log normal distribution with mean 7.43 and variance 0.56, what is the probability that X is greater than 8? So, if we utilize this formula here, the mean of a log normal distribution is e to the power $\mu + \sigma^2$ by 2, the second moment is e to the power $2\mu + 2\sigma^2$; so we have μ_1' is equal to 7.43, which yields the equation $\mu + \sigma^2$ by 2 is approximately 2 and μ_2' is variance plus μ_1' square. So, 0.56 plus 7.43 square and we substitute here the value for this. So, after simplification this gives the equation $\mu + \sigma^2$ is equal to 2.0106. So, if we solve these two equation we get μ is approximately 2 and σ is approximately 0.1.

So, now the probability of a log normal random variable can be calculated using normal distribution. So, probability of X greater than 8 reduces to probability $\log X$ greater than $\log 8$, which is probability Z greater than $\log 8$ minus 2 divided by 0.1, which is probability Z greater than 0.79 and from the tables of the standard normal distribution we can see it is point 0.2148. So, this distribution is quite useful in various applications and

since it is directly related to the normal distribution, the calculations related to this are quite conveniently handled using the properties of the normal distribution.

Another thing that you can observe here is that this distribution will be a skew distribution, X is having a symmetric distribution, but $\log X$ having a skew distribution here, we can actually calculate the third moment the fourth moment etcetera to consider the measures of a skewness and kurtosis etcetera. We have considered almost all the important continuous distributions which arise in practice of course, one can say that if we are looking at any given phenomena, then what will be the distribution corresponding to that that can be considered by making a frequency polygon or a histogram and looking at the data you can see that what kind of distribution will be best suited to describe that data. The distribution that we have discussed so far are the once which were more important in the sense that they arise in lot of practical applications and also historically they are important as they were considered as certain phenomena like some physical phenomena, are some genetic phenomena etcetera where they arise. In the next lecture we will be considering various applications of these distributions. So, we stop here.