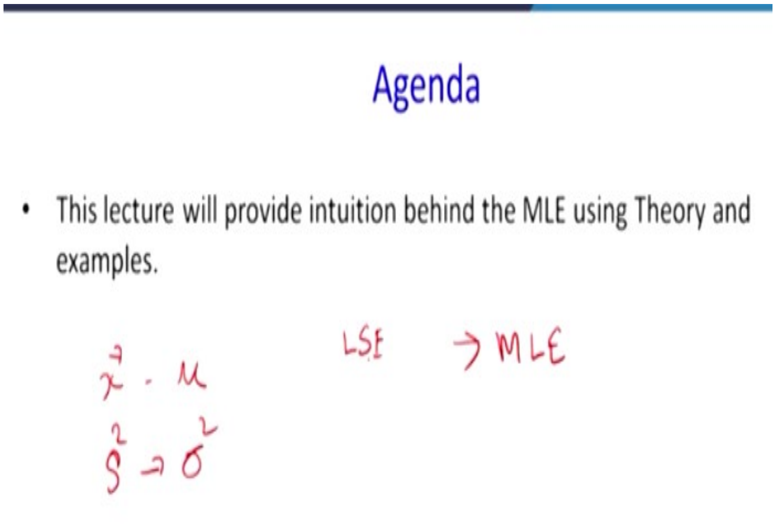
**Data Analytics with Python**
**Prof. Ramesh Anbanandam**
**Department of Management Studies**
**Indian Institute of Technology – Roorkee**

**Lecture - 36**
**Maximum Likelihood Estimation - I**

In this lecture, we will go to new way of estimating the population parameter. That method is called maximum likelihood estimation. In our previous class, we have estimated the population parameter with the help of least square or we can say with the help of method of moments. This method of estimating population parameter has lot of advantages over that two methods. That we will see in this class.

**(Refer Slide Time: 00:50)**



## Agenda

- This lecture will provide intuition behind the MLE using Theory and examples.

$\bar{x} - \mu$

$S^2 \rightarrow \sigma^2$

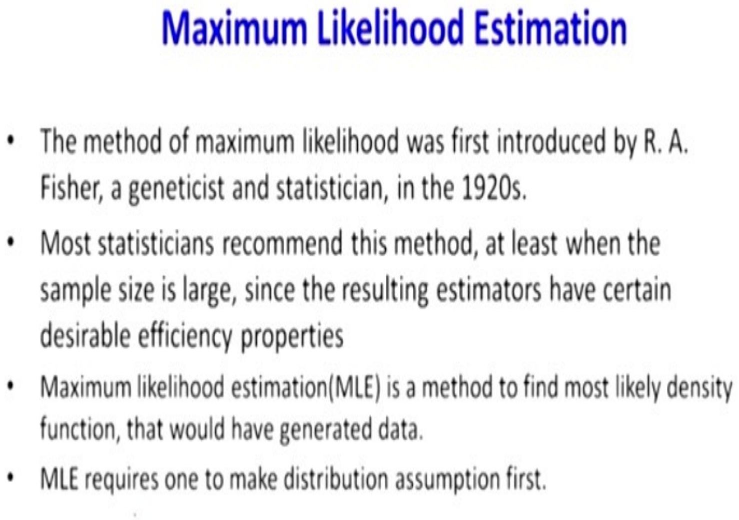
LSE $\rightarrow$ MLE

The agenda for this class is to provide an intuition behind maximum likelihood principle and theory and examples. So what we are going to do, we remember in the previous class with the help of x bar, we have predicted the mean with the help of sample variance, we have predicted the population variance with the help of moment. In the regression model, we have used least square estimate. What you have done in this?

The sum of the square of the error is minimized when we draw the best regression equation. Instead of that one, we are going to use another way of estimating population parameter with the help of maximum likelihood estimation. This is very simple. With the help of this maximum likelihood estimate, you can estimate parameter of any population, it may be any distribution. It may be binomial; it may be a Poisson. It may be an exponential.

What is the assumption? We are having in the least square estimate is that error term should follow normal distribution. Whenever the error term not following normal distribution, the maximum likelihood estimate is the best way. That we will see in this class.

**(Refer Slide Time: 02:02)**



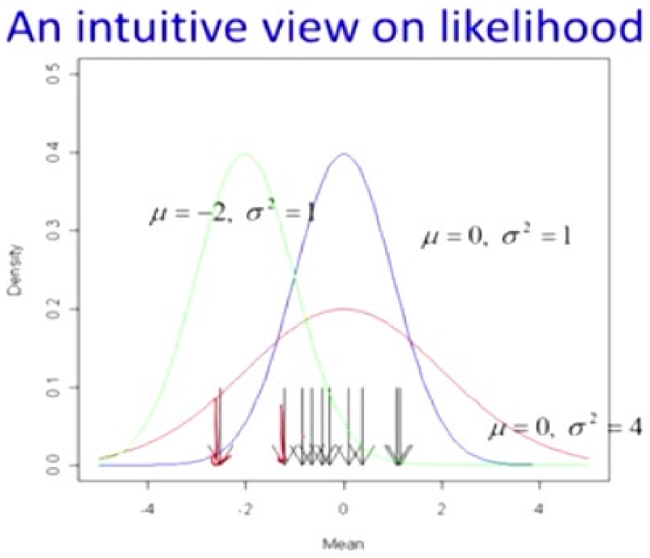
## Maximum Likelihood Estimation

- The method of maximum likelihood was first introduced by R. A. Fisher, a geneticist and statistician, in the 1920s.
- Most statisticians recommend this method, at least when the sample size is large, since the resulting estimators have certain desirable efficiency properties
- Maximum likelihood estimation(MLE) is a method to find most likely density function, that would have generated data.
- MLE requires one to make distribution assumption first.

What is maximum likelihood estimation? The method of maximum likelihood was first introduced by R. A. Fisher, a geneticist and statistician in 1920s. Most statistician recommend this method at least when the sample size is large, since the resulting estimator have certain desirable efficiency properties. Maximum likelihood estimation is a method to find most likely density function that would have generated the data.

So what we can do with the help of this MLE is that which distribution has generated the data that we can find out. Otherwise, this data set is suitable for what kind of distribution? But one assumption we have to have this maximum likelihood estimation is that it requires that one to make distribution assumption first. So in advance, we have to assume which distribution has generated that set of data.

**(Refer Slide Time: 03:01)**

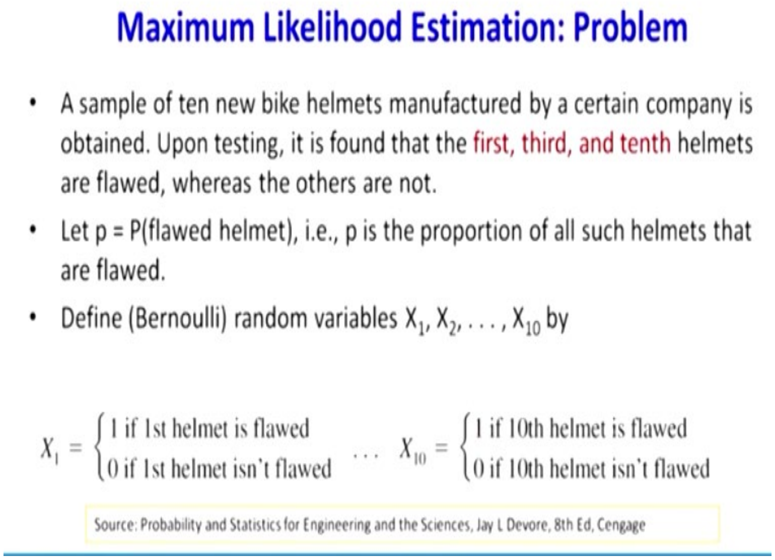


An intuitive view on likelihood

Let us see the intuitive view on likelihood. See there are some data set there, in the bottom. See there are this data set. We want to know from which normal distribution this data set might have come. There are three possibilities; one is the green line, whose mean is minus 2, variance is 1. The another one is blue, whose mean is 0 and the variance is 1. The last one is mean equal to 0, the variance is 4.

So the most suitable for this one is the blue one, because that covers all the data set. So the purpose of maximum likelihood principle is; suppose there are some data. This data has come from which distribution. So that kind of testing can be done with the help of this. Otherwise, this data set is suitable for what kind of distribution; the other way also. So this is most useful for estimating many population parameters.

**(Refer Slide Time: 04:01)**



## Maximum Likelihood Estimation: Problem

- A sample of ten new bike helmets manufactured by a certain company is obtained. Upon testing, it is found that the first, third, and tenth helmets are flawed, whereas the others are not.
- Let p = P(flawed helmet), i.e., p is the proportion of all such helmets that are flawed.
- Define (Bernoulli) random variables $X_1, X_2, \ldots, X_{10}$ by

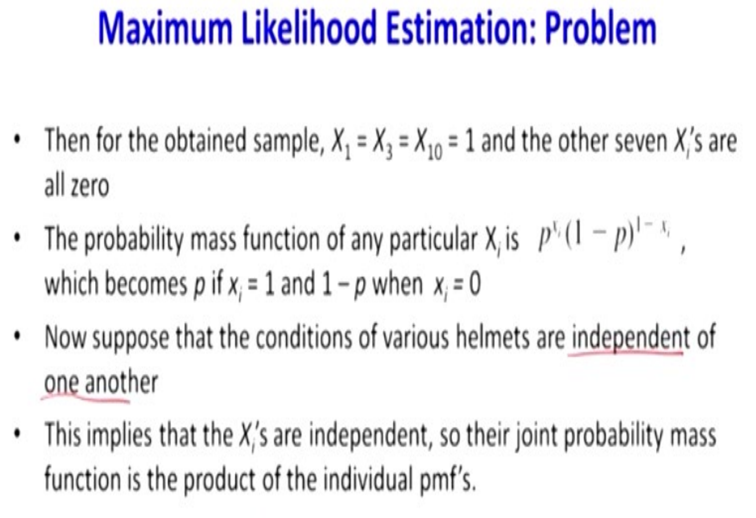
$$X_1 = \begin{cases} 1 \text{ if 1st helmet is flawed} \\ 0 \text{ if 1st helmet isn't flawed} \end{cases} \quad \ldots \quad X_{10} = \begin{cases} 1 \text{ if 10th helmet is flawed} \\ 0 \text{ if 10th helmet isn't flawed} \end{cases}$$

Source: Probability and Statistics for Engineering and the Sciences, Jay L Devore, 8th Ed, Cengage

We will take one simple example. With the help of this example, I will explain what is the application of this maximum likelihood estimation? This problem is taken from this book probability and statistics for engineering and sciences by professor Jay L. Devore 8th edition. It is Cengage publications. The problem says a sample of 10 new bike helmets manufactured by a certain company is obtained.

Upon testing, it is found that the first, third, and 10th helmets are flawed; whereas the others are not. Let p is the probability of flawed helmet that is p is the proportion of all such helmets that are flawed. Define Bernoulli random variable X1, X2, and so on up to X10 by; we are going to use X1 = 1, if the helmet is flawed; if there is a defect. X1 = 0, if the helmet is not defective. Like that, if X10 value = 1, if the 10th helmet is flawed, 0 if the 10th helmet is not flawed, defective.

**(Refer Slide Time: 05:18)**



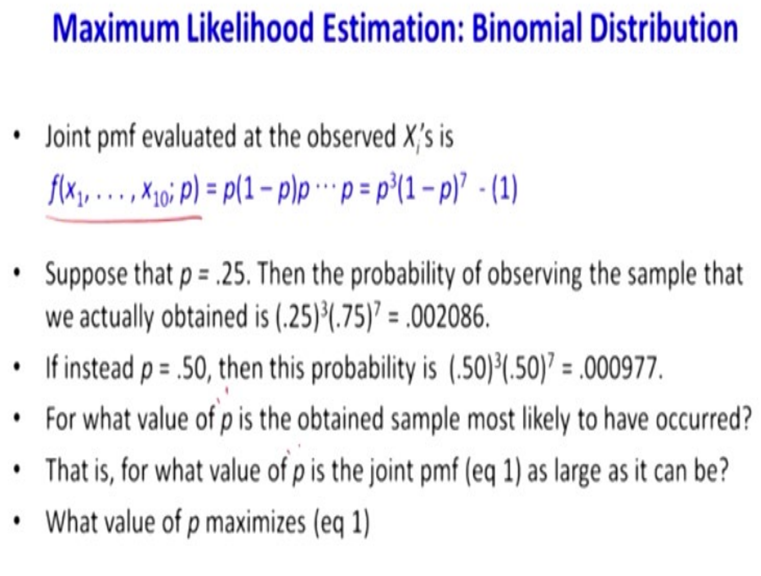
## Maximum Likelihood Estimation: Problem

- Then for the obtained sample, $X_1 = X_3 = X_{10} = 1$ and the other seven $X_i$'s are all zero
- The probability mass function of any particular $X_i$ is $p^{x_i}(1-p)^{1-x_i}$ , which becomes $p$ if $x_i = 1$ and $1 - p$ when $x_i = 0$
- Now suppose that the conditions of various helmets are independent of one another
- This implies that the $X_i$'s are independent, so their joint probability mass function is the product of the individual pmf's.

Then, for the obtained sample, say X1 = X3 = X10 = 1, because they are already given; only the first and third and 10th helmet have some defect and the other seven Xi's are all 0. The values are 0. The probability mass function of any particular Xi is $p^{Xi}$ $(1 – p)$ $^{1-Xi}$, which becomes p if Xi equal to 1 and 1 – p when Xi equal to 0. Now, suppose the conditions of various helmets are independent of one another, because this assumption is very important.

If there is independent, we can find out their joint distribution. This implies that Xi's are independent, so that their joint probability mass function is the product of their individual probability mass function.

**(Refer Slide Time: 06:19)**



## Maximum Likelihood Estimation: Binomial Distribution

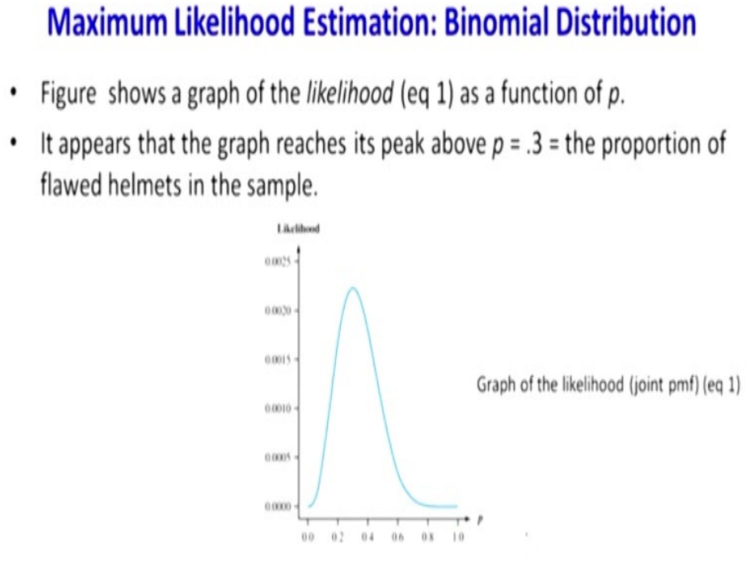
- Joint pmf evaluated at the observed $X_i$'s is
  $f(x_1, \ldots, x_{10}; p) = p(1-p)p \cdots p = p^3(1-p)^7$ - (1)
- Suppose that $p = .25$. Then the probability of observing the sample that we actually obtained is $(.25)^3(.75)^7 = .002086$.
- If instead $p = .50$, then this probability is $(.50)^3(.50)^7 = .000977$.
- For what value of $p$ is the obtained sample most likely to have occurred?
- That is, for what value of $p$ is the joint pmf (eq 1) as large as it can be?
- What value of $p$ maximizes (eq 1)

Since it is joint probability mass function, see that we have multiplied for all possibilities, pi into (1 – pi) by considering all possibilities. So when you simplify that $p^3$ into $(1 – p)^7$. This equation is 1. Suppose, in that equation, this left hand side, this value is called maximum. This is a likelihood value. Whatever is in the left hand side, I will define it later, what is the likelihood value. The left hand value is called likelihood value.

Suppose with the p = 0.25, then the probability of observing the sample that we actually obtained is 0.002086. So like that, we can supply different p values. Suppose, instead of 0.25, you supply p = 0.5, then the probability is 0.0097. You see that, when it is a 0.25, 0.002, when it is a 0.5, it has become very low. So in between this 0.25 and 0.50, we are going to get the value of p that will maximize our left hand side value.

For what value of p is obtained sample, most likely to have occurred? That was the question. What is that? For what value of p is obtained sample most likely to have occurred; that is for what value of p is the joint pmf; this one, as large as it can be; otherwise what value of p maximizes equation 1. That p value is nothing but your likelihood value.

**(Refer Slide Time: 08:07)**



## Maximum Likelihood Estimation: Binomial Distribution

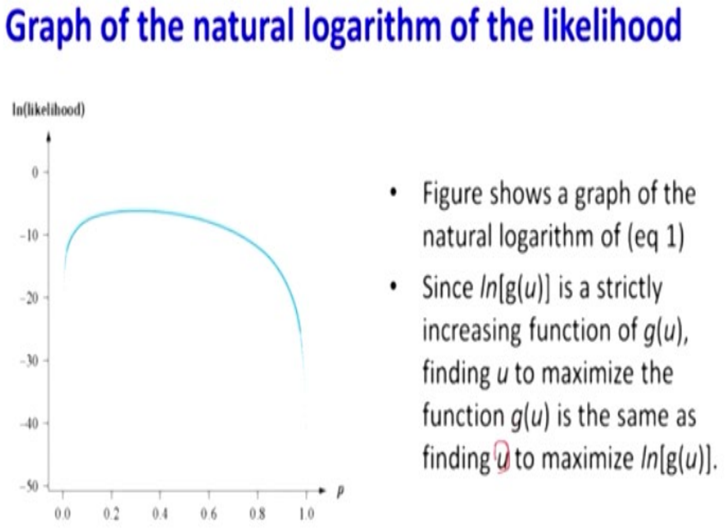
- Figure shows a graph of the *likelihood* (eq 1) as a function of *p*.
- It appears that the graph reaches its peak above *p* = .3 = the proportion of flawed helmets in the sample.

Graph of the likelihood (joint pmf) (eq 1)

Figure shows a graph of likelihood of that value of equation 1 as a function of p. So what happened in x axis, you have taken different value of p. Previously, just for one case, we have taken p value 0.25 and 0.50. See when it is 0.25, this was the likelihood value, 0.5, this was the likelihood value. So we are good to supply, draw a graph by supplying different value of p in equation 1 that we have to find out the likelihood value.

This figure shows a graph of likelihood as a function of p. It appears that the graph reaches its peak above 0.3, when the value is 0.3; the graph reaches its peak, equal to the proportion of flawed helmets in the sample. Now what you are going to do? We are going to take log of this function.

**(Refer Slide Time: 09:04)**



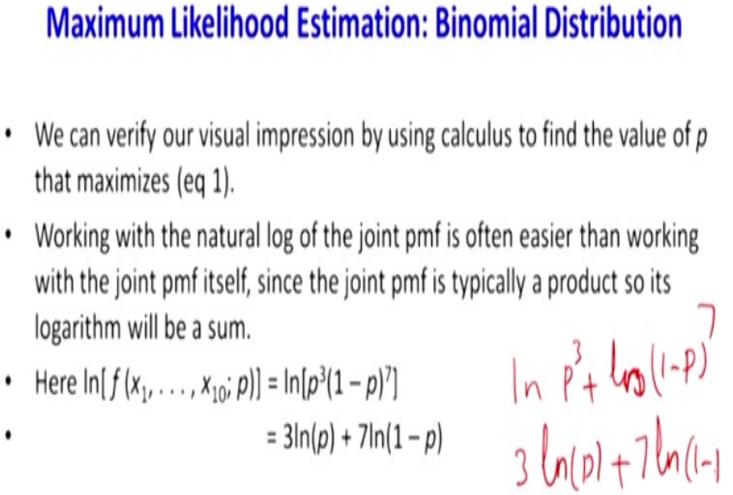
Graph of the natural logarithm of the likelihood

- Figure shows a graph of the natural logarithm of (eq 1)
- Since $ln[g(u)]$ is a strictly increasing function of $g(u)$, finding $u$ to maximize the function $g(u)$ is the same as finding $u$ to maximize $ln[g(u)]$.

I will explain; there was a reason for that. Graph of the natural logarithm of likelihood. Figure shows graph of the natural logarithm of equation 1. Since the logarithm of g[u] is strictly increasing function of g[u], finding u to maximize the function g of u is the same as finding u to maximize log of g of u. So what is happening is, whether of g of u and logarithm of g of u is the same. This figure shows a graph of the natural logarithm of equation 1.

Since log of g[ u] is strictly increasing function of g of u, finding u to maximize the function of g of u is the same as finding u to maximize log of g of u. So the u is same, whether it is g of u or log of g of u.

**(Refer Slide Time: 10:04)**



## Maximum Likelihood Estimation: Binomial Distribution

- We can verify our visual impression by using calculus to find the value of $p$ that maximizes (eq 1).
- Working with the natural log of the joint pmf is often easier than working with the joint pmf itself, since the joint pmf is typically a product so its logarithm will be a sum.
- Here $\ln[f(x_1, \dots, x_{10}; p)] = \ln[p^3(1-p)^7]$
- $= 3\ln(p) + 7\ln(1-p)$

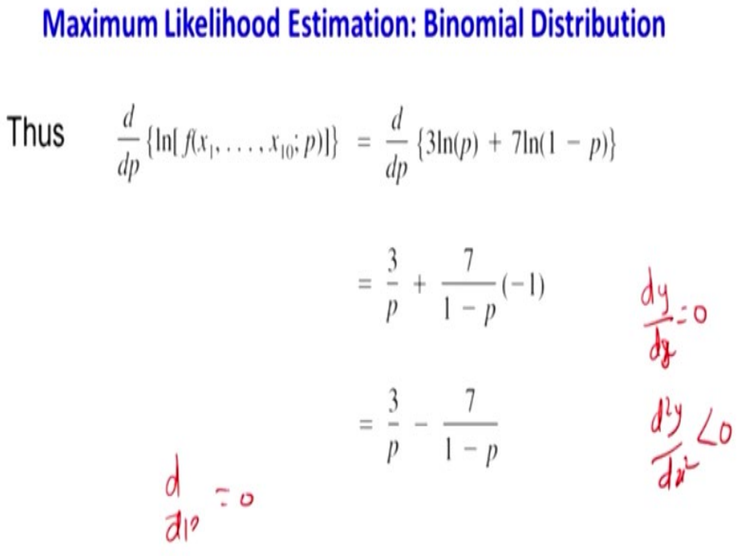
$\ln p^3 + \log(1-p)^7$

$3\ln(p) + 7\ln(1-$

We can verify our visual impression by using calculus to find out the value of p that maximizes equation 1. Working with natural logarithm of the joint probability mass function is often easier than working with the joint pmf itself. Since the joint pmf is typically a product, so the logarithm will be a sum. That is the advantage of taking log of that. So what will happen previously in equation 1, we got pq multiplied by 1 – p power 7. I am going back here; this one.

We are going to take log of this. When you take log of this, it will become, because there is a multiplication, so this will become log of $p^3$ + log of $(1 – p)^7$. So this will become 3 log (p) + 7 log (1 – p).

**(Refer Slide Time: 11:12)**



## Maximum Likelihood Estimation: Binomial Distribution

Thus $\frac{d}{dp}\{\ln[f(x_1, \dots, x_{10}; p)]\} = \frac{d}{dp}\{3\ln(p) + 7\ln(1-p)\}$

$$= \frac{3}{p} + \frac{7}{1-p}(-1)$$

$$\frac{dy}{dx} = 0$$

$$= \frac{3}{p} - \frac{7}{1-p}$$

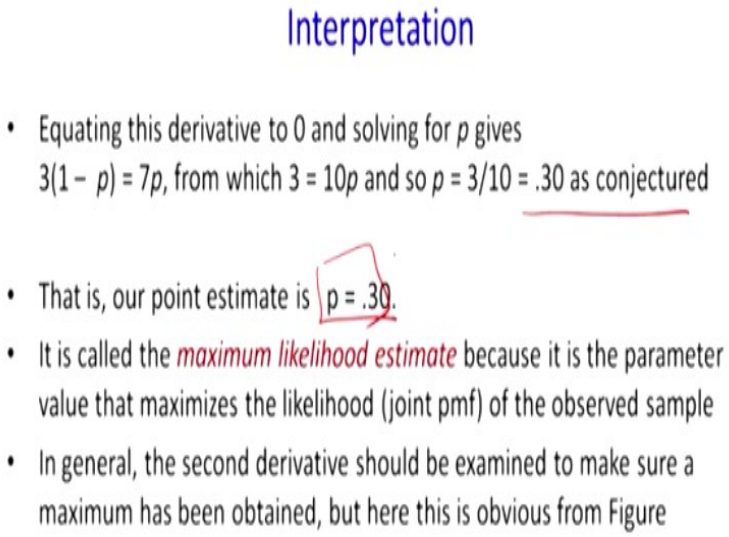
$$\frac{d^2y}{dx^2} < 0$$

$$\frac{d}{dp} = 0$$

Next one is that functions, we have to see when the value become maximum? We know that in our school, we might have studied; you see to find out the maximum value, maxima-minima. For example, the maximum value, if you say dy by dx equal to 0; then (d$^2$y/ dx $^2$ )will be less than 0 means that point will become the maximum. So this equation, this is the function of p. So we are going to differentiate that log function with respect to p.

So when you differentiate this one, so 3, log of p = (1 / p), so (3 /p), + 7 is a constant. Log of 1 – p is, this is differentiation, log of x equal to 1 by x. So 1 divided by (1 – p), again you have to differentiate this function. Differentiation of differentiation, so 0 – 1, so it will be -1. So (3/ p) – 7 divided by (1 – p). So this equations, we have to equate it to 0, because we know (d /dp) should be equal to 0. Then, we have to find out the p. So that value, the function will get maximized.

**(Refer Slide Time: 12:32)**



## Interpretation

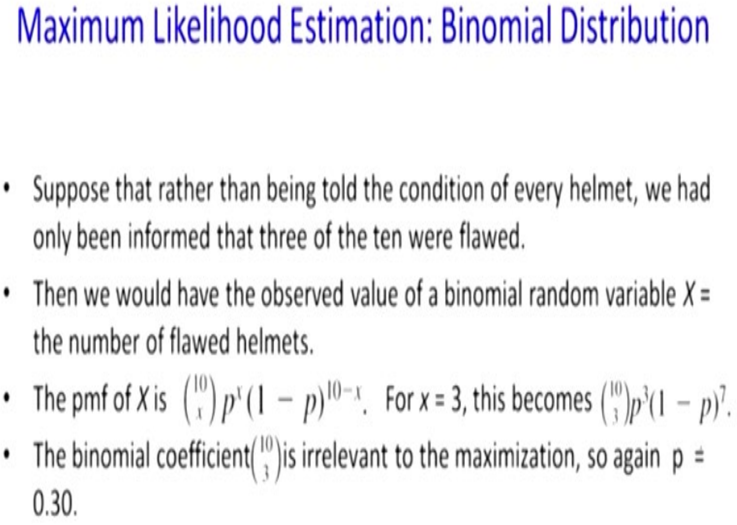
- Equating this derivative to 0 and solving for *p* gives 3(1 – *p*) = 7*p*, from which 3 = 10*p* and so *p* = 3/10 = .30 as conjectured
- That is, our point estimate is $\boxed{p = .30}$.
- It is called the *maximum likelihood estimate* because it is the parameter value that maximizes the likelihood (joint pmf) of the observed sample
- In general, the second derivative should be examined to make sure a maximum has been obtained, but here this is obvious from Figure

Equating these derivatives to 0 and solving for p, it gives 3 into 1 – p equal to 7p. So 3 equal to 7p, so p = 0.3 is conjectured. So now what is happening? Previously, we have substituted different values. Now we are using the concept of maxima, we have realized that when the p = 0.3, the function gets maximized. So it is called the maximum likelihood estimate, because it is the parameter value that maximizes the likelihood of the observed sample.

So this p = 0.3 will be nothing but the; this is an estimate for the population. In general, second derivative should be maximum to make sure the maximum has been obtained, but here this is obvious from the figure. So actually we have to differentiate one more time and we have to see whether it has become negative or not, because by looking at the figure, it seems that that point is maximum. So what is happening; this value p = 0.3 is called the maximum likelihood estimate.

So what is happening, the binomial distribution of the population parameter p, we have estimated it is 0.3. So the advantage of this maximum likelihood function is, it is helping to estimate parameter of any distribution.

**(Refer Slide Time: 14:01)**

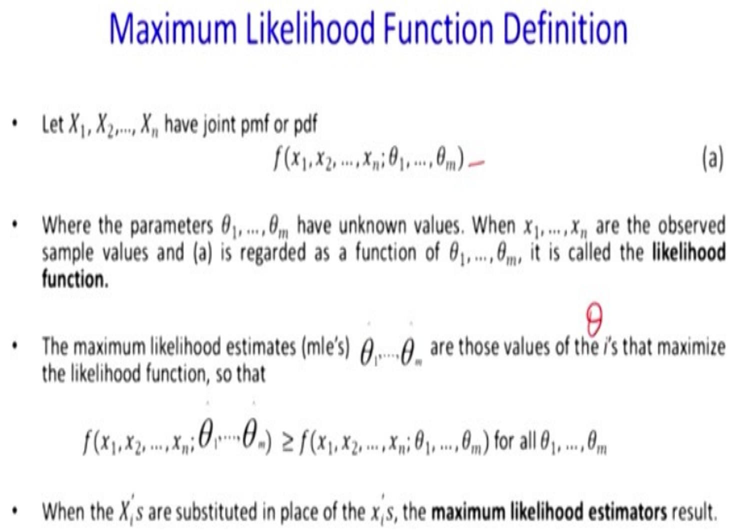


Maximum Likelihood Estimation: Binomial Distribution

- Suppose that rather than being told the condition of every helmet, we had only been informed that three of the ten were flawed.
- Then we would have the observed value of a binomial random variable $X$ = the number of flawed helmets.
- The pmf of $X$ is $\binom{10}{x} p^x (1-p)^{10-x}$. For $x = 3$, this becomes $\binom{10}{3} p^3 (1-p)^7$.
- The binomial coefficient $\binom{10}{3}$ is irrelevant to the maximization, so again $p \doteq$ 0.30.

Suppose, that rather than being told the condition of every helmet, we had only been informed that three of the 10 were flawed. Then, we would have to observe the value of binomial random variable X equal to the number of flawed helmets when you substitute 10,X; $^{10}C_x\ p^x$ into $(1 – p)^{10 – x}$. When you substitute x = 3, this is $^{10}C_3\ p^3$ into $(1 – p)^7$. We do not bother about the coefficient 10C3, because that is not a function; that is just a constant. So what they say, the binomial coefficient 10C3 is irrelevant to maximization. So again, the p = 0.3.

**(Refer Slide Time: 14:44)**



## Maximum Likelihood Function Definition

- Let $X_1, X_2, \dots, X_n$ have joint pmf or pdf

$$f(x_1, x_2, \dots, x_n; \theta_1, \dots, \theta_m) - \qquad (a)$$

- Where the parameters $\theta_1, \dots, \theta_m$ have unknown values. When $x_1, \dots, x_n$ are the observed sample values and (a) is regarded as a function of $\theta_1, \dots, \theta_m$, it is called the **likelihood function.**

- The maximum likelihood estimates (mle's) $\hat{\theta}_1, \dots, \hat{\theta}_m$ are those values of the $\theta$ $i$'s that maximize the likelihood function, so that

$$f(x_1, x_2, \dots, x_n; \hat{\theta}_1, \dots, \hat{\theta}_m) \geq f(x_1, x_2, \dots, x_n; \theta_1, \dots, \theta_m) \text{ for all } \theta_1, \dots, \theta_m$$

- When the $X_i's$ are substituted in place of the $x_i's$, the **maximum likelihood estimators** result.

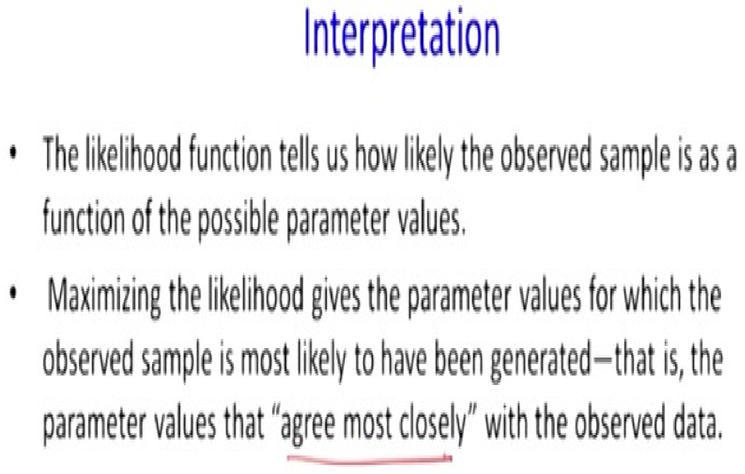
Next, we will define, what is maximum likelihood function. There are two terms there; one is likelihood function, next one is maximum likelihood function. First I will say what is likelihood function, then we will go to what is maximum likelihood function. Let X1, X2, and Xn have a joint probability mass function or probability density function; call it as f of x1, x2, up to xn ; $\Theta_1$, $\Theta_2$ and $\Theta_m$, where the parameters $\Theta_1$, $\Theta_2$ and $\Theta_m$ have unknown values.

Here the parameter is $\Theta_1$, $\Theta_2$, unknown values where x1, x2, xn are the observed sample values, then this equation a is regarded as the function of $\Theta_1$, $\Theta_2$ upto m, it is called likelihood function. So this is a likelihood function. So this function is likelihood function. The maximum likelihood estimates theta 1 hat, theta 2 hat, theta m hat are those values of theta i's that maximizes the likelihood function.

So that f ( x1, x2 up to xn ; theta 1 hat, theta 2 hat, up to theta m) is greater than or equal to f ( x1, x2, x3 up to xn ; theta 1, theta 2,…. theta m) for all values of $\Theta_1$, $\Theta_2$ and theta m. When the Xi’s are substituted in place of xi’s, the maximum likelihood estimates result. So what you have to do with that one? In the Xi’s we have to substitute xi’s that will be the maximum likelihood estimate result. So what we are doing here?

We are finding joint probability mass function, then with the help of sample values, we are predicting the population parameter.

**(Refer Slide Time: 16:48)**

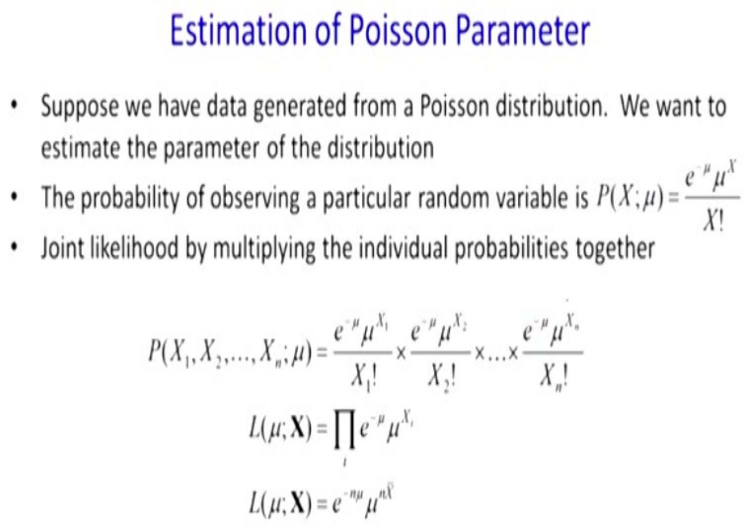


## Interpretation

- The likelihood function tells us how likely the observed sample is as a function of the possible parameter values.
- Maximizing the likelihood gives the parameter values for which the observed sample is most likely to have been generated—that is, the parameter values that "agree most closely" with the observed data.

How will you interpret that one? The likelihood function tells us how likely the observed sample is as a function of possible parameter values. So maximizing the likelihood gives the parameter values, for which the observed value is most likely to have been generated. That is, the parameter values that agree most closely with the observed data. Otherwise, we can say in other way, that this data set is more suitable for what kind of distributions or what kind of models.

**(Refer Slide Time: 17:22)**



## Estimation of Poisson Parameter

- Suppose we have data generated from a Poisson distribution. We want to estimate the parameter of the distribution
- The probability of observing a particular random variable is $P(X;\mu)=\frac{e^{-\mu}\mu^{X}}{X!}$
- Joint likelihood by multiplying the individual probabilities together

$$P(X_1,X_2,\ldots,X_n;\mu)=\frac{e^{-\mu}\mu^{X_1}}{X_1!}\times\frac{e^{-\mu}\mu^{X_2}}{X_2!}\times\ldots\times\frac{e^{-\mu}\mu^{X_n}}{X_n!}$$

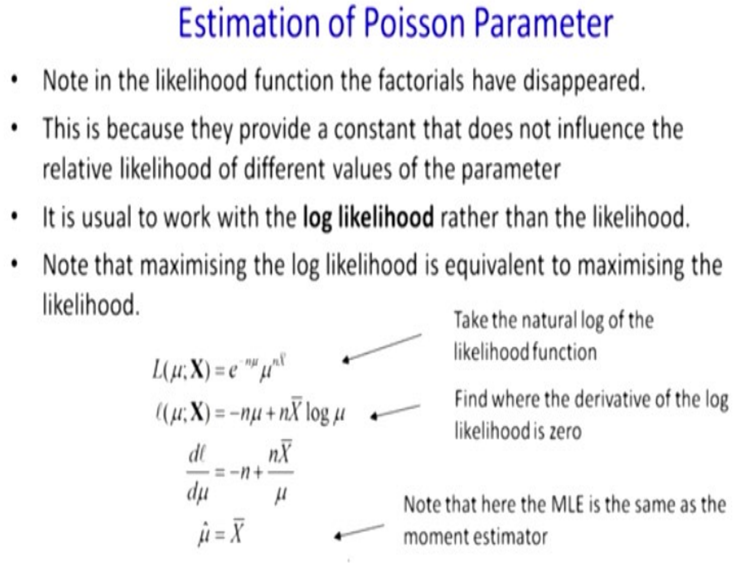
$$L(\mu;\mathbf{X})=\prod_i e^{-\mu}\mu^{X_i}$$

$$L(\mu;\mathbf{X})=e^{-n\mu}\mu^{n\bar{X}}$$

Now we will go for estimation of Poisson parameter. Suppose, we have data generated from a Poisson distribution, we want to estimate the parameter of Poisson distribution. The Poisson distribution is having only one parameter, because in Poisson distribution, it is an unique parametric distribution; it has only one parameter, that is where the mean and variance is same. The probability of observing a particular random variable $P ( X;u) = (e^{-u} u^{X} )/ X !$

So joint likelihood by multiplying the individual probabilities together, so what we will do the first step is we have to find out the joint probability function. So $(e^{-u} u^{X1} )/ X_1 !$ multiplied by $(e^{-u} u^{X2} )/ X_2 !$ and so on multiply $(e^{-u} u^{Xn} )/ Xn !$. So this can be written as product of $(e^{-u} u^{Xi} )$ because it is a product, there is an end time. When you expand it, so $e^{-nu}$, because it will become up to n times, so $u^{nX\ bar}$. Next, we have to take the log of this, we will see that.

**(Refer Slide Time: 18:45)**



## Estimation of Poisson Parameter

- Note in the likelihood function the factorials have disappeared.
- This is because they provide a constant that does not influence the relative likelihood of different values of the parameter
- It is usual to work with the **log likelihood** rather than the likelihood.
- Note that maximising the log likelihood is equivalent to maximising the likelihood.

$$L(\mu;\mathbf{X}) = e^{-n\mu}\mu^{n\bar{X}}$$

Take the natural log of the likelihood function

$$\ell(\mu;\mathbf{X}) = -n\mu + n\bar{X}\log\mu$$

Find where the derivative of the log likelihood is zero

$$\frac{d\ell}{d\mu} = -n + \frac{n\bar{X}}{\mu}$$

$$\hat{\mu} = \bar{X}$$

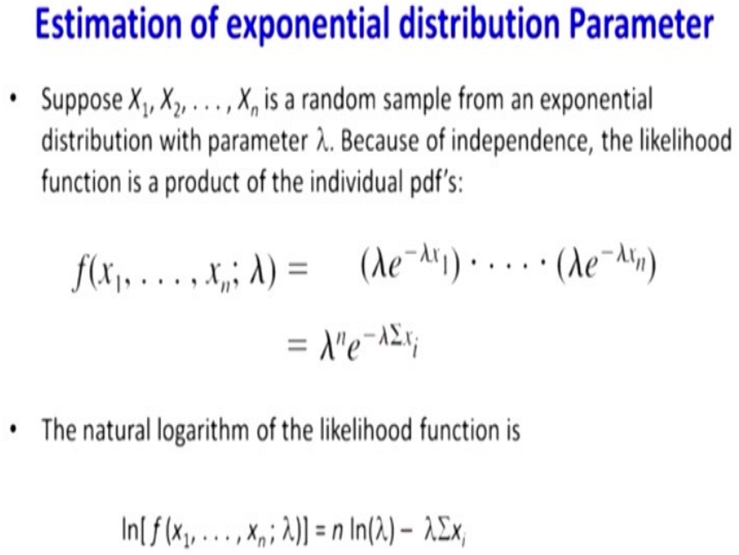
Note that here the MLE is the same as the moment estimator

Note that, the likelihood function that factorials have disappeared. We will not bother about the factorials, because that is not going to affect the result. This is because they provide a constant that does not influence the relative likelihood of different values of the parameter whether we use the constant or not, that is not required, because that will not affect our end result. It is usual to work with log likelihood rather than likelihood, because we have seen previously.

When you take log of likelihood, the differentiation is easy. Note that, maximizing the log likelihood is equivalent to the maximizing likelihood. This also, we have seen in the previous slide. So this was the likelihood function. You take log of that one. So e power, when you take log of e to the power –n mu is –mu, because it is the product, in log it will become sum, sum nX bar log of mu. Now you differentiate with respect to mu.

When you differentiate it and equate it to 0, then you are getting X bar equal to mu. So what is the result is, the sample mean is the best estimate to predict the population mean of a Poisson distribution.

**(Refer Slide Time: 20:04)**



## Estimation of exponential distribution Parameter

- Suppose $X_1, X_2, \ldots, X_n$ is a random sample from an exponential distribution with parameter λ. Because of independence, the likelihood function is a product of the individual pdf's:

$$f(x_1, \ldots, x_n; \lambda) = (\lambda e^{-\lambda x_1}) \cdot \cdots \cdot (\lambda e^{-\lambda x_n})$$
$$= \lambda^n e^{-\lambda \Sigma x_i}$$

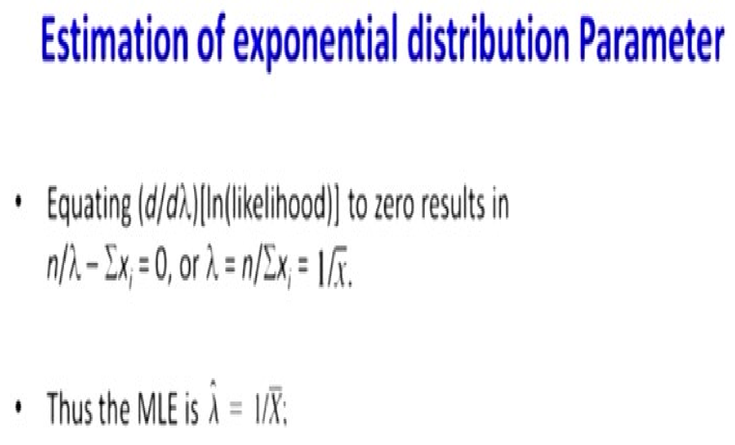
- The natural logarithm of the likelihood function is

$$\ln[f(x_1, \ldots, x_n; \lambda)] = n \ln(\lambda) - \lambda \Sigma x_i$$

Now we will go for another distribution that is estimation of exponential distribution parameter. Suppose, X1, X2, Xn is a random sample from an exponential distribution with the parameter lambda. Because of independence, the likelihood function is the product of individual pdf's. Here also $\lambda e^{-\lambda x1}$ will extend it, $\lambda e^{-\lambda x2}$ up to, you have to multiply $\lambda e^{-\lambda xn}$.

So when you simplify that, it will become $\lambda^n$ $e^{-\lambda\Sigma xi}$. When you take log of this, it will become n ln(λ) - λΣxi. Then this has to be equated to 0.

**(Refer Slide Time: 20:59)**

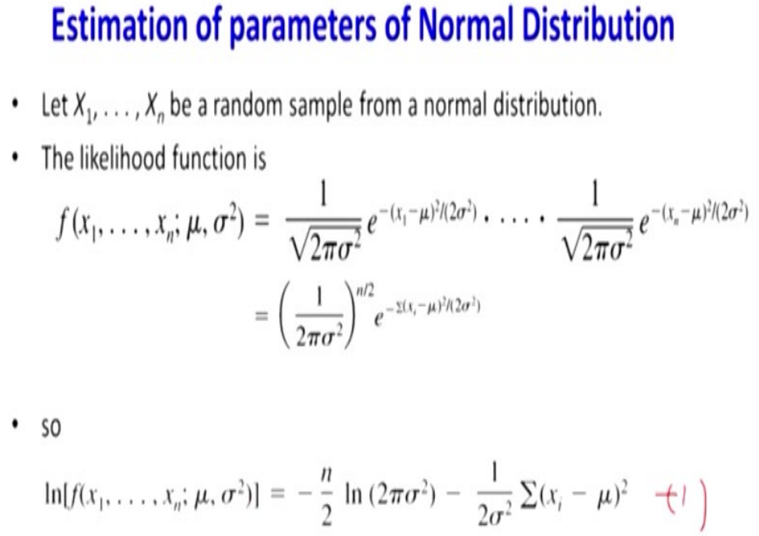


## Estimation of exponential distribution Parameter

- Equating $(d/d\lambda)$[ln(likelihood)] to zero results in $n/\lambda - \Sigma x_i = 0$, or $\lambda = n/\Sigma x_i = 1/\bar{x}$.
- Thus the MLE is $\hat{\lambda} = 1/\bar{X}$;

So when you equate it to 0, so lambda becomes lambda equal to n divided by sigma Xi that is nothing but the inverse of the sample mean. So this was the result. So what is happening is, now the inverse of sample mean is nothing but the mean of our exponential distribution.

**(Refer Slide Time: 21:24)**



## Estimation of parameters of Normal Distribution

- Let $X_1, \ldots, X_n$ be a random sample from a normal distribution.
- The likelihood function is

$$f(x_1, \ldots, x_n; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/(2\sigma^2)} \cdot \ldots \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/(2\sigma^2)}$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\Sigma(x_i-\mu)^2/(2\sigma^2)}$$

- so

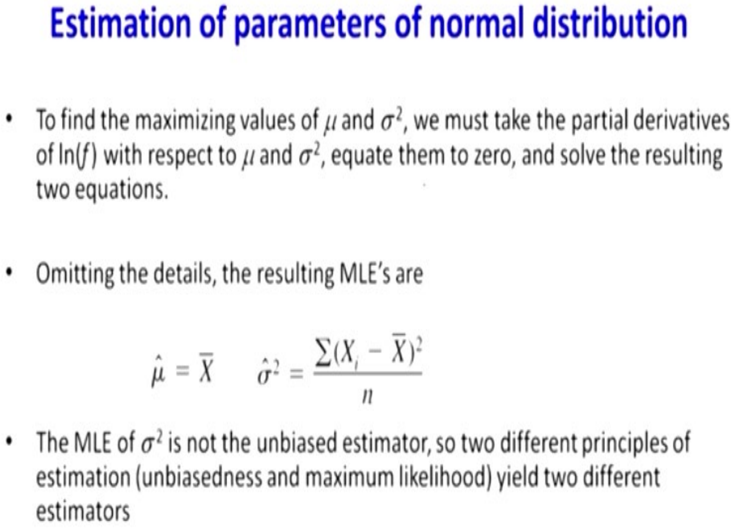
$$\ln[f(x_1, \ldots, x_n; \mu, \sigma^2)] = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \Sigma(x_i - \mu)^2 \quad (1)$$

Now we will go for estimation of parameter of a normal distribution. This was very interesting because we can say normal distribution is the father of all the distributions. Many time, if you are not knowing the nature of the distribution, you can assume that it follows normal distribution. As usual, the likelihood function for a normal distribution is, we know that the pdf, probability density function is (1 by root of 2 πσ $^2$) e to the power (–(x$_1$-u)$^2$ divided by 2σ $^2$).

So like that, this is term 1, term 2, up to nth term we can go for that. So term 1, that is when it is x1, when you substitute x2, x3, so you will get different n terms. So that is probability mass function. Joint probability function, so when you simplify that it is (1 divided by $2\pi\sigma^{2}$) to the power n/2 , into e to the power ($-\Sigma(x_1\text{-}u)^2$ divided by $2\sigma^{2}$). What will happen? When you take log of this, this is (n / 2) (ln (1) – ln ($2\pi\sigma^{2}$)).

So what will happen, log of 1 minus, because log 1 is 0, so it will become 0 – ln($2\pi\sigma^{2}$ ), because it is x power n. So –n by 2 n log 2 pi sigma square – e to the power, this one will come in that value itself, because 2 sigma square, sigma of xi – x mu whole square. Now what has to be done? This is the log value of likelihood function. This function, this equation has to be partially derivated with respect to mu and sigma square and equate it to 0.

**(Refer Slide Time: 23:19)**



## Estimation of parameters of normal distribution

- To find the maximizing values of $\mu$ and $\sigma^2$, we must take the partial derivatives of ln($f$) with respect to $\mu$ and $\sigma^2$, equate them to zero, and solve the resulting two equations.
- Omitting the details, the resulting MLE's are

$$\hat{\mu} = \overline{X} \qquad \hat{\sigma}^2 = \frac{\sum(X_i - \overline{X})^2}{n}$$

- The MLE of $\sigma^2$ is not the unbiased estimator, so two different principles of estimation (unbiasedness and maximum likelihood) yield two different estimators

Then, you will get the parameter. To find the maximizing value of mu and sigma square, we must take the partial derivatives of the previous function with respect to mu and sigma square and equate them to 0 and solve the resulting two equations. There are a lot of details, omitting the details, we will get this result. What does this result says? With the help of sample mean, we can predict the population mean.

With the help of this one, look at this one; this term is the sample variance, we can predict the population variance. So this was the outcome of, you remember, this was the result of our central limit theorem also. We can prove that central limit theorem by using this maximum likelihood estimate. But one point you should be very careful, the maximum likelihood estimate of sigma square is not the unbiased estimator.

Actually, we should look for unbiased estimator, but here it is not unbiased estimator. So, two different principles of estimation, unbiasedness and maximum likelihood yield two different estimators. In this class, I have started the intuitive meaning of maximum likelihood principle. Then, I have explained how to find out the population parameter of different distributions. First, I have seen how to predict the population parameter of binomial distribution.

Next we have seen how to predict the parameter of Poisson distribution. Then, next we have predicted the population parameter of exponential distribution. At last, we have predicted population parameter of normal distributions. In the next class, we will take one example for predicting the parameter of normal distribution. Thank you very much.