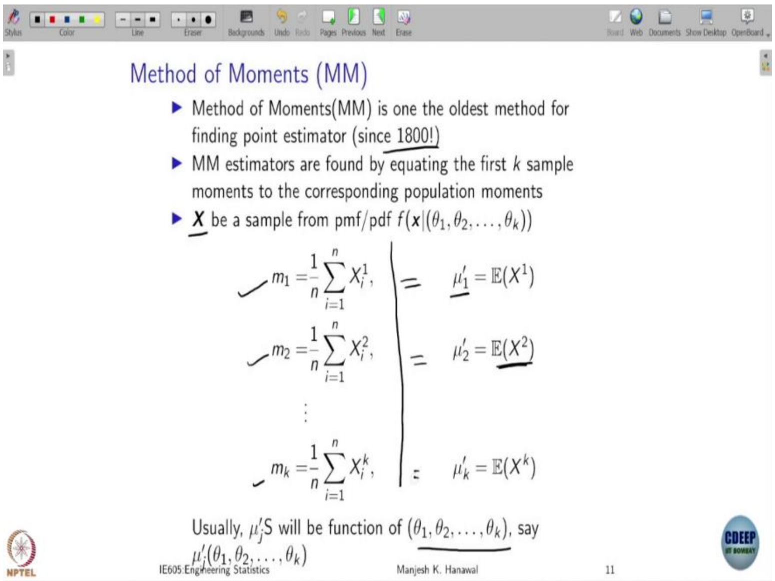
Engineering Statistics
Professor Manjesh Hanawal
Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay
Lecture 38
Methods of Moments, Baye's Estimator

(Refer Slide Time: 0:24)



Method of Moments (MM)

- Method of Moments(MM) is one the oldest method for finding point estimator (since 1800!)
- MM estimators are found by equating the first k sample moments to the corresponding population moments
- X be a sample from pmf/pdf $f(\mathbf{x}|(\theta_1, \theta_2, \ldots, \theta_k))$

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i^1, \quad = \quad \mu_1' = \mathbb{E}(X^1)$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \quad = \quad \mu_2' = \mathbb{E}(X^2)$$

$$\vdots$$

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad = \quad \mu_k' = \mathbb{E}(X^k)$$

Usually, μ_j'S will be function of $(\theta_1, \theta_2, \ldots, \theta_k)$, say $\mu_j'(\theta_1, \theta_2, \ldots, \theta_k)$

So, let us get started with the other part. Now we are going to talk about other possible way of estimating the parameters. Now there is another method called method of moments, we all studied about the moments. We know about first moment, what does first moment give, mean and we know second moment. So, second moment is related to the variance. Like that we can keep on computing the different method or moments. How you are going to compute moments? What is the easiest way to compute moment?

If you have moment generating function, you can write quickly compute your moments by differentiating those moment generating functions. So, how and this is one of the oldest methods. You will see why this is so obvious and this is like people are using it since 1800s. So, what it does is method of moment estimators, they are found by equating first k sample moments to the corresponding population moments.

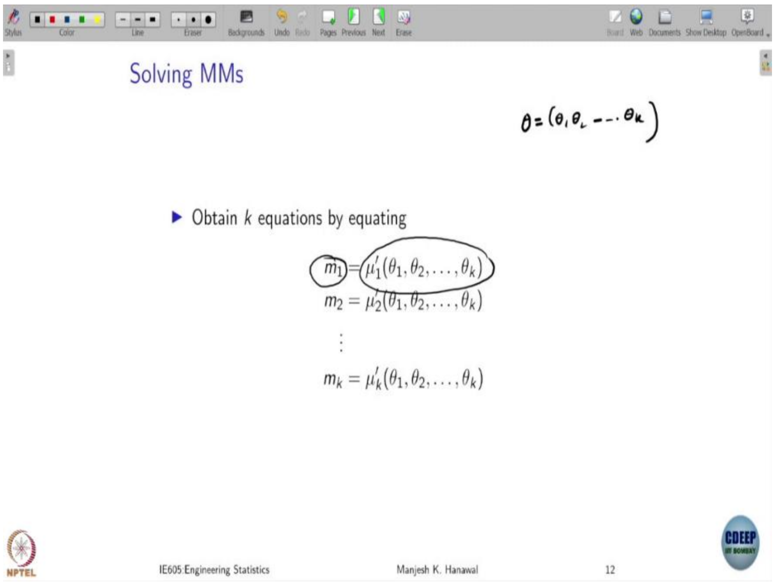
What I mean by that suppose let us say you have a sample X coming from a PDF. You can compute the sample mean which is a proxy for your first moment. Can I interpret like that? Sample mean, can I interpret sample mean to be an sample estimate of my first moment?

Now I can also go ahead take the square of the samples and average them, then I can interpret as a sample estimate of my second moment. And similarly, I can take as many exponent and of the samples and take their average. I can go on up to taking kth exponent, exponenting all the sample by k and then take the average that can I can take it as kth moment, I mean proxy for my kth moment. A sample estimate might create the moment. This is I have done with my sample part using samples only I can compute m1, m2 up to mk.

Now from my underlying population distribution it has certain parameters, the mean will be related to those parameters. Somehow, I will write that and the second moment the actual second moment will also be related to the parameter. I will find that relation and all this, exactly how the moments are truly related to the actual relationship with the parameters I will find out.

And all these actual moments, the first moment, actual second moment and the kth moment, they will be function of my parameter theta1, theta2 up to theta k, the underlying parameters. Now I can equate them. I can equate these two to find one equation, these two to find second equation like that I can equate the k, I can also equate it to get the kth equation.

(Refer Slide Time: 3:56)



Solving MMs

$\theta = (\theta_1, \theta_2 \ldots \theta_k)$

- Obtain k equations by equating

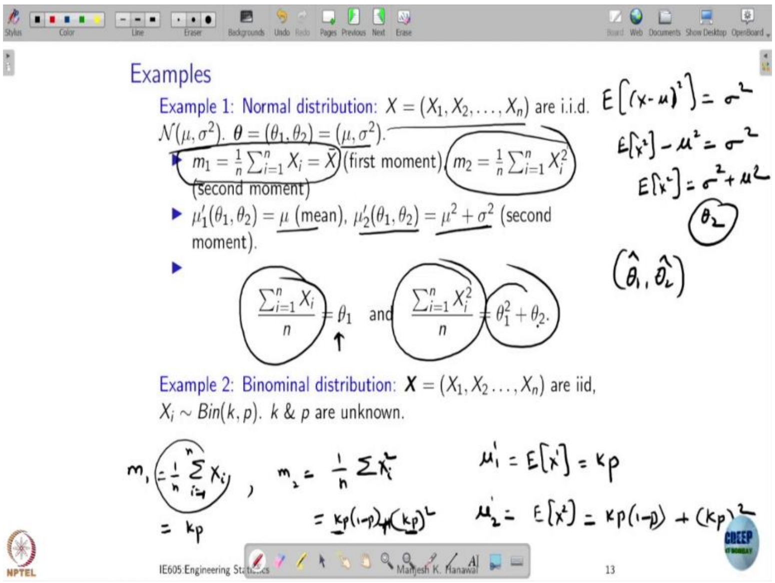
$$m_1 = \mu_1'(\theta_1, \theta_2, \ldots, \theta_k)$$
$$m_2 = \mu_2'(\theta_1, \theta_2, \ldots, \theta_k)$$
$$\vdots$$
$$m_k = \mu_k'(\theta_1, \theta_2, \ldots, \theta_k)$$

NPTEL IE605:Engineering Statistics Manjesh K. Hanawal 12 CDEEP IIT BOMBAY

So, this is what. I have this m1 computed from my sample. I know the true value from the corresponding parameters. Now I am equating them and here by equating them I get k equations and I am going to solve these k equations to get those k parameters. So, my theta here consists of k components and I need at least k equations to solve to get these k points.

So, that is what I am deriving this k equations by finding these k moments. The exact values are given by this and the actual value given from the samples are given by the left quantities.

(Refer Slide Time: 4:55)



Examples

Example 1: Normal distribution: $X = (X_1, X_2, \ldots, X_n)$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$. $\boldsymbol{\theta} = (\theta_1, \theta_2) = (\mu, \sigma^2)$.

- $m_1 = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}$ (first moment) $m_2 = \frac{1}{n}\sum_{i=1}^n X_i^2$ (second moment)
- $\mu_1'(\theta_1, \theta_2) = \mu$ (mean), $\mu_2'(\theta_1, \theta_2) = \mu^2 + \sigma^2$ (second moment).
- $$\frac{\sum_{i=1}^n X_i}{n} = \theta_1 \text{ and } \frac{\sum_{i=1}^n X_i^2}{n} = \theta_1^2 + \theta_2.$$

$E[(x-\mu)^2] = \sigma^2$

$E[x^2] - \mu^2 = \sigma^2$

$E[x^2] = \sigma^2 + \mu^2$

θ_2

$(\hat{\theta}_1, \hat{\theta}_2)$

Example 2: Binominal distribution: $\mathbf{X} = (X_1, X_2 \ldots, X_n)$ are iid, $X_i \sim Bin(k, p)$. k & p are unknown.

$m_1 = \frac{1}{n}\sum_{i=1}^n X_i$, $m_2 = \frac{1}{n}\sum X_i^2$

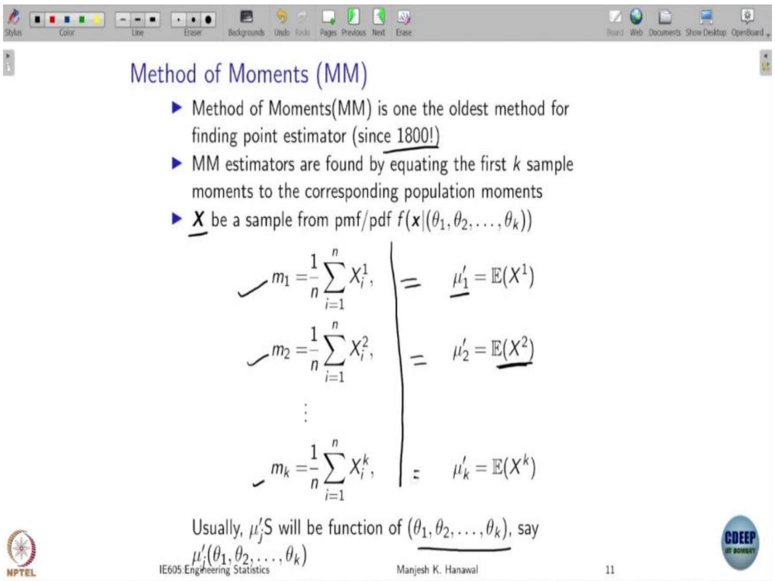
$= kp$

$= kp(1-p) + (kp)^2$

$\mu_1' = E[x^1] = kp$

$\mu_2' = E[x^2] = kp(1-p) + (kp)^2$

IE605:Engineering Statistics Manjesh K. Hanawal 13



Method of Moments (MM)

- Method of Moments(MM) is one the oldest method for finding point estimator (since 1800!)
- MM estimators are found by equating the first k sample moments to the corresponding population moments
- $\mathbf{X}$ be a sample from pmf/pdf $f(\mathbf{x}|(\theta_1, \theta_2, \ldots, \theta_k))$

$$m_1 = \frac{1}{n}\sum_{i=1}^n X_i^1, \quad = \quad \mu_1' = \mathbb{E}(X^1)$$

$$m_2 = \frac{1}{n}\sum_{i=1}^n X_i^2, \quad = \quad \mu_2' = \mathbb{E}(X^2)$$

$$\vdots$$

$$m_k = \frac{1}{n}\sum_{i=1}^n X_i^k, \quad = \quad \mu_k' = \mathbb{E}(X^k)$$

Usually, $\mu_j'S$ will be function of $(\theta_1, \theta_2, \ldots, \theta_k)$, say $\mu_j'(\theta_1, \theta_2, \ldots, \theta_k)$

IE605:Engineering Statistics Manjesh K. Hanawal 11

Now let us look into the example, how does this work. Again, let us take samples x1 to xn which are i.i.d. and which are coming from Gaussian distribution with parameter mu and sigma square. And now I am assuming that my parameters are mu and sigma square both are unknown. And I am going to represent them as theta 1 and theta 2. Actually theta 1 is mu and theta 2 is sigma square.

Now what does the method of moment say? First you compute your m1 which is basically sample mean then compute your m2. Now how many moments I need to compute here? 2 here, right? Because I have two components in my parameter. Now I have this compute my m2 also from the samples. To compute m2 to what I did? I just take the square of these samples and average them.

Now I need to compute the true values. What is mu hat in this case? So, mu hat is the first true moment, yeah true first moment? What is the true first moment in this case? It is going to be mu itself, right? Which is actually theta 1. And now the second moment, true second moment.

How it is related? It is going to be mu square plus sigma square, right? Why is that because we know that x minus mu whole square equals to sigma square but this is nothing but expectation of x square. You can simplify this; this is like this so expectation of x square is simply sigma square plus mu square. The true second moment is going to be mu square plus sigma square.

Now I got two equations. This is the first moment I got from the samples and this is the true first moment and this is the second moment I got from the sample and this is the true second moment. Notice that at theta sigma squared I have denoted as theta 2. So, that is the theta 2 and theta 1 square.

Now I have two equations and two parameters. Can I solve this and get the value for theta 1 and theta 2? So, I will get whatever the values, I will get theta 1 and theta 2 hat and that is the estimates. Whatever I get is the estimates. So, notice that in this case also theta 1 is already this sample mean and theta 2 is what is going to come from that now.

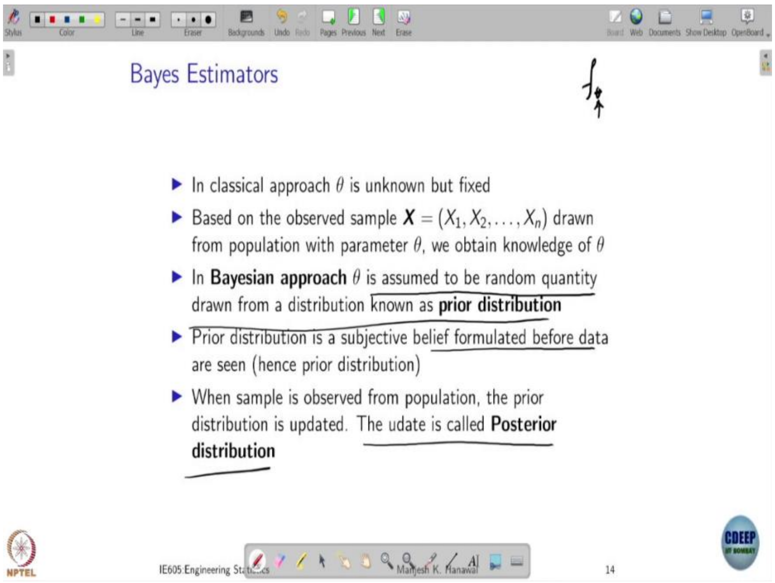
Now let us do a quick example for binomial distribution. Let us say I have this sample, first I am going to compute m1 which is summation of xi, i 1 to n and 1 by n. And now I am going to calculate m2 which is 1 by n what, xi squared. Now what is the m1 hat for me? The true first moment, I want now this quantity is here, mu 1 prime is nothing but mu 1 prime is nothing but what is my case that is expectation of 1 which is kp and mu 2 prime, I do not know why we have written prime prime here everywhere.

And this is second moment, true second moment. What is the true second moment of a binomial distribution, kp 1 minus p. So, now what I am going to do? I am going to take this to be equals to kp, okay right, kp whole square.

Now this is going to be kp and this is going to be kp1 minus p plus kp whole square. Now I have two equations to solve in two parameters. So, by the way how you are going to solve this? You can replace this kp by this quantity, you will still end up with kp into 1 minus p. That product is equals to something. How you are going to solve this? Everybody see that? So, this kp, you can replace by this value and also there is one more kp, you can again replace that by this value. Then this will become only a function in 1 minus p, that you can solve for p.

So, you see that right, method of moments is so natural. All you need to do is estimates the moments, as many moments you want and then equate it get that many questions and then solve them. Then what is the other, any question about method of moments?

(Refer Slide Time: 11:08)



Bayes Estimators

- In classical approach θ is unknown but fixed
- Based on the observed sample $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ drawn from population with parameter θ, we obtain knowledge of θ
- In **Bayesian approach** θ is assumed to be random quantity drawn from a distribution known as **prior distribution**
- Prior distribution is a subjective belief formulated before data are seen (hence prior distribution)
- When sample is observed from population, the prior distribution is updated. The udate is called **Posterior distribution**

IE605:Engineering Statistics Manjesh K. Hanawal 14

Now something will see Bayes estimators. So, in the classical approach what we did? We assumed that there is some underlying probability density function with parameter theta which is generating my samples. We assume that this function f is fixed, the structure of that is fixed. Maybe it is gaussian or binomial or something but just like the parameters are unknown. But we just said that that parameter was fixed.

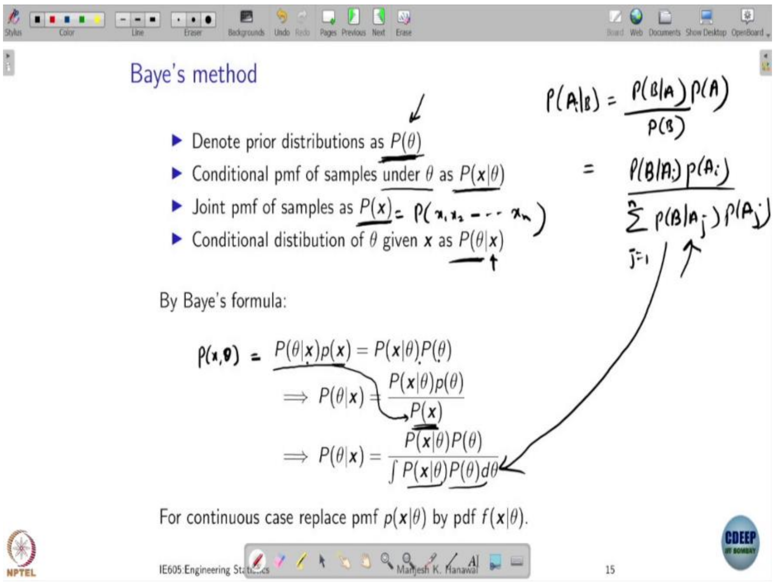
Now in the Bayesian approach what we do is, we assume that this theta itself is a random quantity drawn from some prior distribution. So, we are going to say that this theta itself is coming from some distribution which we do not know. And we are going to assume that which we do not know but we may assume that, it is coming from some fixed distribution. It is being some drawn from some distribution but what is that exact value I do not know.

In the classical example, what we did is we did not assume that theta is coming from some distribution. What we assume that that is some something fixed. It is always fixed and it is the constant value but here we are assuming that it could be coming from some, we did not put any prior distribution on this basically. But now we want to put some prior distribution saying that, this theta itself is coming from some distribution and this parallel distribution is a subjective belief about how that theta may be arriving or coming from.

But in this case we may initially assume that ok this thetas are coming from certain distribution but when you observe a sample generated, we may improve according to which distribution these thetas are coming and that is what we are going to update and call that updated distribution as posterior distributions.

So, initially we will make some assumptions saying that potentially this theta may be coming from some distribution which we call as prior distribution but when we start observing some data, we want to revisit and update that distribution and when we update the distribution that is what we are going to obtain as posterior distribution.

(Refer Slide Time: 13:56)



Baye's method

- Denote prior distributions as $P(\theta)$
- Conditional pmf of samples under θ as $P(x|\theta)$
- Joint pmf of samples as $P(x) = P(x_1, x_2 - - - x_n)$
- Conditional distibution of θ given x as $P(\theta|x)$

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)}$$

By Baye's formula:

$$P(x,\theta) = P(\theta|x)p(x) = P(x|\theta)P(\theta)$$

$$\implies P(\theta|x) = \frac{P(x|\theta)p(\theta)}{P(x)}$$

$$\implies P(\theta|x) = \frac{P(x|\theta)P(\theta)}{\int P(x|\theta)P(\theta)d\theta}$$

For continuous case replace pmf $p(x|\theta)$ by pdf $f(x|\theta)$.

IE605:Engineering Statistics Manjesh K. Hanawal 15

So, let us say my thetas are coming some from distribution and the distribution I am going to denote it as p of theta and once some theta is fixed, the samples are generated under that parameter theta according to this probability px given theta. So, theta are coming from this distribution, once you have a theta your samples are going to be distributed as per px given theta. And this is the distribution of your samples, I call it joint because this is like a probability, this is the entire sample you are going to talk about.

And now given your x, you may want to talk about the conditional distribution of your theta itself. So, notice that you will start with some distribution of theta and assuming that theta let us say I mean whatever the underlying theta, it is going to generate some samples and now using some samples we want to improve your, update your distribution about theta.

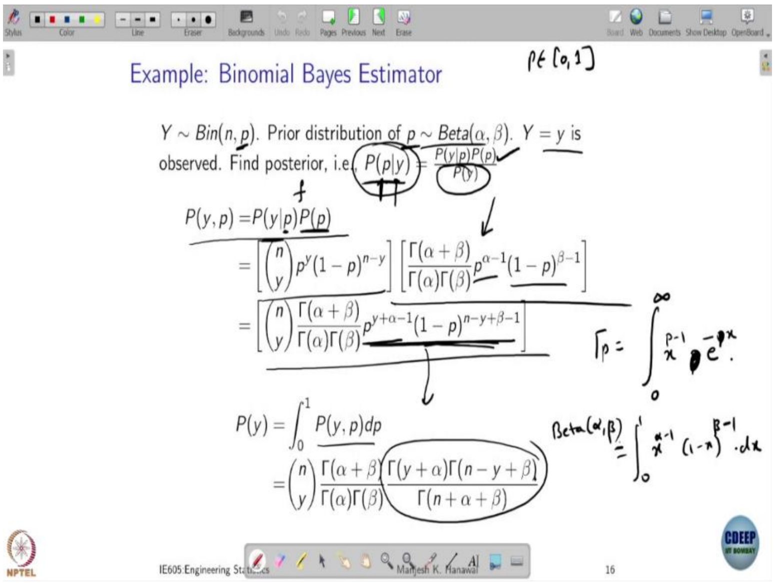
Now let us try to understand this formally. So, this is the joint PDF of x and theta. I can write it like this or like this. Now what I will do is, I will try to manipulate this P theta given x is, I have just brought this quantity at the bottom and now P theta x. Now I know that this unconditional probability of x, I can write in terms of conditional probability and then integrate it over P of theta. So, do you recall the steps that we did at the beginning when we did Bayesian formula? When we applied Bayesian formula?

So, you recall probability of A given B is equals to probability of B given A divided by, let us say, this was like probability of B, that is right, we did this, right? We are trying to do the same thing here but on this conditional on my parameter space and sample space.

And here since I am treating this theta itself is a random quantities, I can do this. Now we also said that if there is a partition here, let us say Ai are some, I want to condition on some particular thing, this I could write it as probability of B given some Aj into probability of Aj and where j equals to 1 to n.

This we did which we call it as total probability and then we use our base and formula. Exactly that is what I am doing here. And now if I am assuming this to be discrete here but if it is a continuous case replace P by fx given theta here.

(Refer Slide Time: 17:37)



Example: Binomial Bayes Estimator

$p \in [0,1]$

$Y \sim Bin(n,p)$. Prior distribution of $p \sim Beta(\alpha,\beta)$. $Y = y$ is observed. Find posterior, i.e. $P(p|y) = \frac{P(y|p)P(p)}{P(y)}$

f

$$P(y,p) = P(y|p)P(p)$$
$$= \left[\binom{n}{y} p^y (1-p)^{n-y}\right]\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\right]$$
$$= \left[\binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1}\right]$$

$$\Gamma p = \int_0^\infty x^{p-1} e^{-x} .$$

$$P(y) = \int_0^1 P(y,p)dp$$
$$= \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)}$$

$$Beta(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} . dx$$

IE605:Engineering Statistics Manjesh K. Hanawal 16

Now let us try to see how to use this binomial estimator. Let us say, I have to deal with the binomial distribution which has parameter n and p, assume that n is known p is unknown. And this p is what I want to estimate using the Bayesian Method now.

Now, if I have to use Bayesian method as I said, I need to start with the prior distribution. What could be a good prior distribution here? We know that P has to be between 0 1, right? But I do not know where it is. What could be possible prior distributions I can take? You can take uniform but more generally, I am going to start with the beta distribution. Somebody said normal, can normal be a good prior distribution? No, why? We want it to be P to be between 0 1, normal is looking into the entire range, range of real number.

So, I am going to take here instead of uniform beta with parameter alpha beta and now I want to see that. I have observed one sample and I want to see that how to compute my posterior probability. So, what is the posterior probability? Posterior probability is going to be this, probability of p given by y. So, p is my parameter and now I observed y, I want to see that initially p is assumed to be beta distribution. Now after that I assumed one sample what is a new distribution of my p, that is my posterior distribution.

Let us come to that. So, to do that I will start with my probability of observing y and p together. That probability is probability I can write it as this conditional probability and this unconditional probability. Now if you tell me I know that y is binomially distributed and you have been observed some, you have been given p. Now I know that y is a binomial and if you are already given p, this is the probability of y given p, everybody agree? This is true because y is binomial that has been assumed and now you have been already told what is the parameter p.

So, then it is simply n choose y, py, p to the power y 1 minus p and minus y. Now what is the probability of this small p? That is assumed to be beta distributed with parameters alpha and beta and notice that I should have been careful. Maybe I should have written it f, because this is a PDF now. This is a PMF because binomial is a discrete whereas p is now assumed to be continuous random variable and its PDF is this. This is a PDF of beta distribution, everybody agree?

Now, I have simply organized that I have just clubbed this p alpha minus 1 and y, I have put together and this beta minus 1 and n minus 1, I have put together. That is it, I have just simplified this.

Now I have the numerator, I need to compute my denominator p of y. Now how to compute P of y? I know I have obtained my joint probability here and I know that p is taking value between 0 1, did I miss something here in this? Nothing right, I just need to integrate it between 0 1.

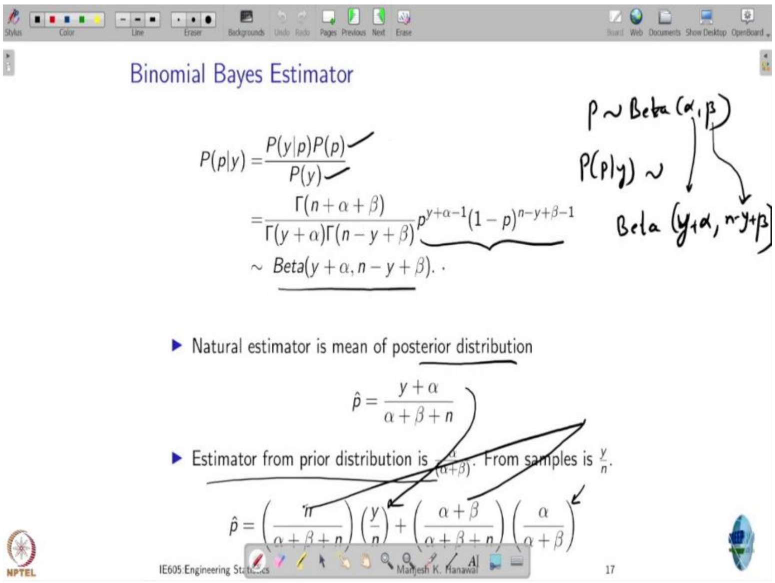
So, if I integrate this quantity over p taking 0 1 it looks like I can write it as another gamma function, if I am going to integrate, so, only p appears is in this product, if I integrate this between 0 1, it looks like this appears to be an another gamma function which I have written here. Am I correct? Is this a gamma function if I integrate it between 0 1?

I am not able to recall the gamma function definition now. But I hope this is correct. So, can you tell me what is the gamma p definition? Integration 0 to 1, x to the power?

Student: (())(23:02)

Professor: That is it? x to the power p minus 1, we have to write a gamma, so what I want here, I want a beta function. What is beta function? x to the power, this is just this, right? And what was the notation? Gamma, alpha, beta we write like this. Beta, okay fine, Beta, alpha comma beta, that is exactly in this form, right?

(Refer Slide Time: 24:33)



Binomial Bayes Estimator

$$P(p|y) = \frac{P(y|p)P(p)}{P(y)}$$

$$= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} \underbrace{p^{y+\alpha-1}(1-p)^{n-y+\beta-1}}$$

$$\sim \underline{Beta(y+\alpha, n-y+\beta)}.$$

$p \sim Beta(\alpha, \beta)$

$P(p|y) \sim$

$Beta(y+\alpha, n-y+\beta)$

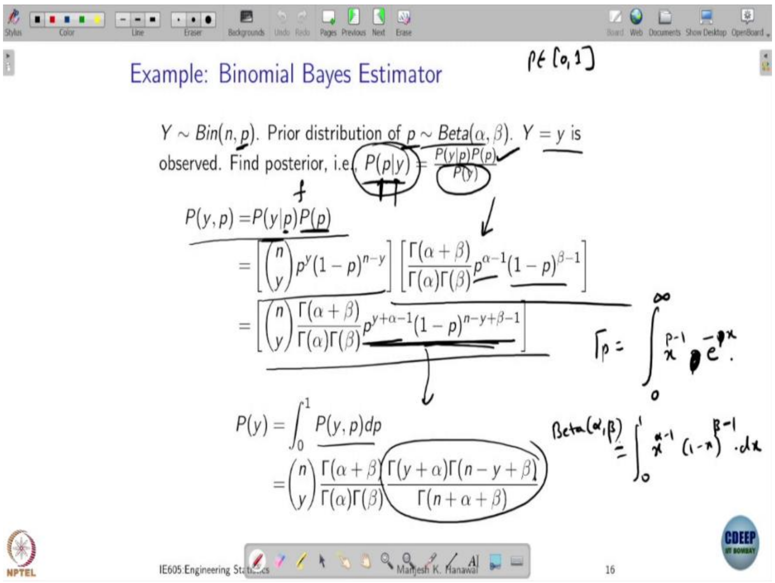
- Natural estimator is mean of posterior distribution

$$\hat{p} = \frac{y+\alpha}{\alpha+\beta+n}$$

- Estimator from prior distribution is $\frac{\alpha}{(\alpha+\beta)}$. From samples is $\frac{y}{n}$.

$$\hat{p} = \left(\frac{n}{\alpha+\beta+n}\right)\left(\frac{y}{n}\right) + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\left(\frac{\alpha}{\alpha+\beta}\right)$$

IE605:Engineering Statistics Manjesh K. Hanawal 17



Example: Binomial Bayes Estimator

$p \in [0,1]$

$Y \sim Bin(n,p)$. Prior distribution of $p \sim Beta(\alpha,\beta)$. $Y = y$ is observed. Find posterior, i.e. $P(p|y) = \frac{P(y|p)P(p)}{P(y)}$

Π f

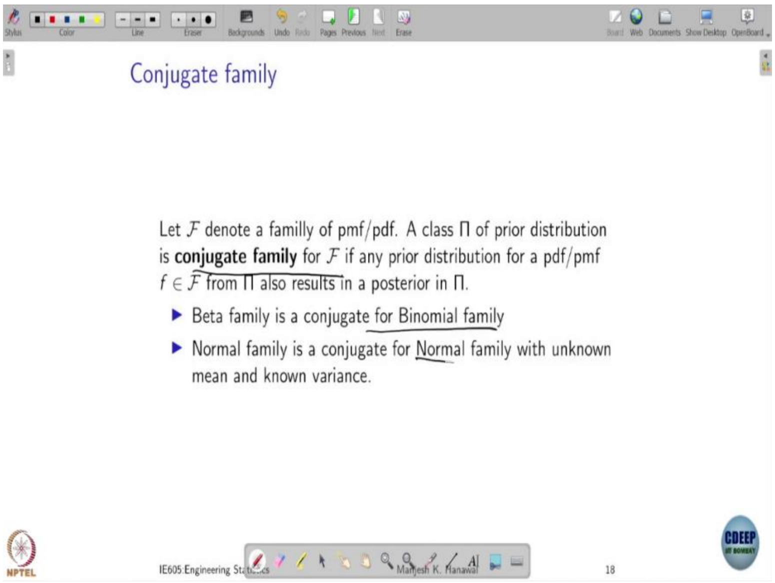
$$P(y,p) = P(y|p)P(p)$$
$$= \left[\binom{n}{y} p^y (1-p)^{n-y}\right]\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\right]$$
$$= \left[\binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1}\right]$$

$$\Gamma p = \int_0^\infty x^{p-1} e^{-x}.$$

$$P(y) = \int_0^1 P(y,p)dp$$
$$= \binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)}$$

$$Beta(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\cdot dx$$

IE605:Engineering Statistics Manjesh K. Hanawal 16



Conjugate family

Let $\mathcal{F}$ denote a familly of pmf/pdf. A class Π of prior distribution is **conjugate family** for $\mathcal{F}$ if any prior distribution for a pdf/pmf $f \in \mathcal{F}$ from Π also results in a posterior in Π.

- Beta family is a conjugate for Binomial family
- Normal family is a conjugate for Normal family with unknown mean and known variance.

IE605:Engineering Statistics Manjesh K. Hanawal 18

Now let us plug in, we have this numerator, now we computed this denominator. If I compute to this, now I will end up after this simplification. After taking these ratios, after taking the ratio of this by this, you see that this n choose y knocks off. This gamma of alpha plus beta, gamma alpha this knocks off, what remains is only this ratio which is this quantity.

Now, by definition is this, can you check, somebody can you verify beta of y plus alpha n minus 1 is exactly this? Why is that? What is this, is this right, I have forgotten all these formulas. So, if you are convinced about this fine, I do not recall this gamma beta function.

Now what we got is, we started with a, we started with the prior on alpha assuming this is beta alpha comma beta. We assume that this is a beta distribution parameter alpha and beta. Now we are saying after I observed a sample, the distribution of these p given y has now changed to beta of y plus alpha and n minus y plus beta. Now the parameter alpha has gone to alpha plus y and beta has gone to n minus y plus beta.

And now you have this new after observing this y, you have this new distribution and now what is the value of the parameter estimator you want to take? Maybe one possibility is what is that now you have this niche distribution, take the mean value of that posterior distribution as our estimation for p. And what is the mean value of this new estimator? It is y plus alpha and the sum of these two. That is alpha plus beta plus n.

Now what I have done is, I basically reorganize this. What now it is saying, this beta hat can be actually looked into the average of these two quantities. This is the prior information that I am going to basically take the weighted value of alpha plus alpha plus beta. This is my, when I have prior information, this is the mean value of the prior distribution. And the y by n is the new value after I observed.

So, the new estimator I am taking is basically the weighted average of these two and the weights are n upon alpha plus beta plus n and alpha plus beta this quantity. So, if you see that these two weights, they sum up to 1. I am basically taking a weighted average of these two quantities, I am giving this much of weight to my prior mean and whatever the observation I got new observation, the new value I am going to give this much of it.

So, one thing you have noticed is, I have started with a beta distribution as a prior but the posterior also happens to be beta distribution. And when that happens they are called conjugate family.

And you will saw that this is not just happens with a binomial, even if you would have started with a normal distribution, this would have happened. Prior you start with normal posterior would also resulted in normal but depending on what is that you want to see like in the my case here p was between 0 1, making beta made more sense rather than taking gaussian. Depending on the application you will choose either beta distribution or gaussian. Okay so let us stop here.