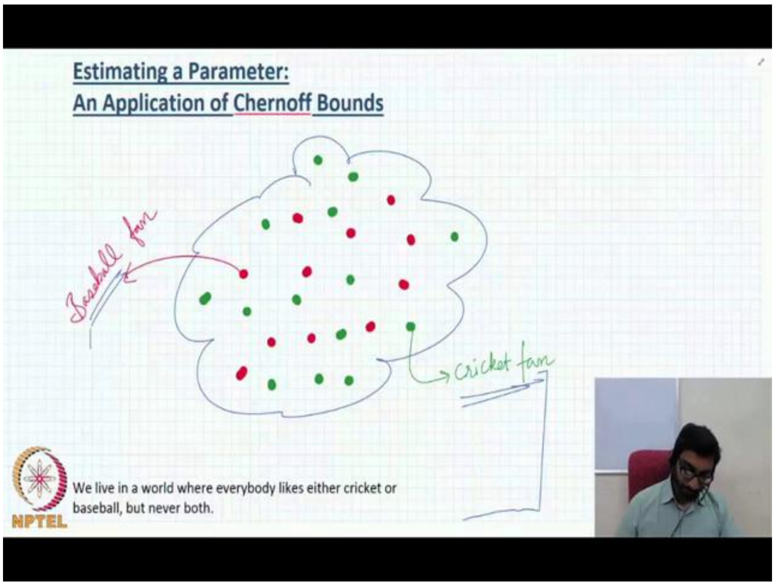
Algorithms for Big Data
Prof. John Ebenezer Augustine
Department of Computer Science and Engineering
Indian Institute of technology, Madras

Lecture – 12
Estimating a Parameter: An Application of Chernoff bound

We are now going to look at the very useful Application of Chernoff Bounds.

(Refer Slide Time: 00:15)

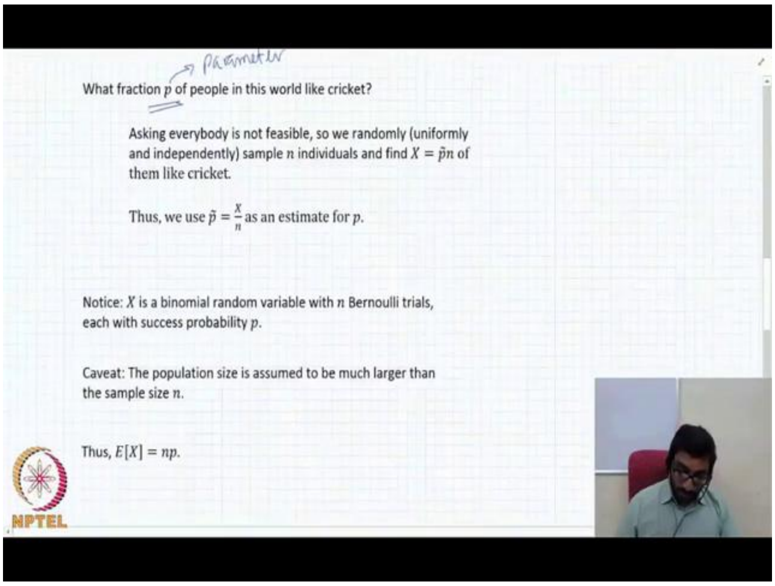


Estimating a Parameter:
An Application of Chernoff Bounds

We live in a world where everybody likes either cricket or baseball, but never both.

In this application we are asked to estimate a particular parameter. And what is this parameter? Let us consider a population, this population as a large number of people some of whom are cricket fans and the rest of them are baseball fans. And everybody likes one of the other, but not both.

(Refer Slide Time: 00:51)



parameter

What fraction p of people in this world like cricket?

Asking everybody is not feasible, so we randomly (uniformly and independently) sample n individuals and find $X = \tilde{p}n$ of them like cricket.

Thus, we use $\tilde{p} = \frac{X}{n}$ as an estimate for p.

Notice: X is a binomial random variable with n Bernoulli trials, each with success probability p.

Caveat: The population size is assumed to be much larger than the sample size n.

Thus, $E[X] = np$.

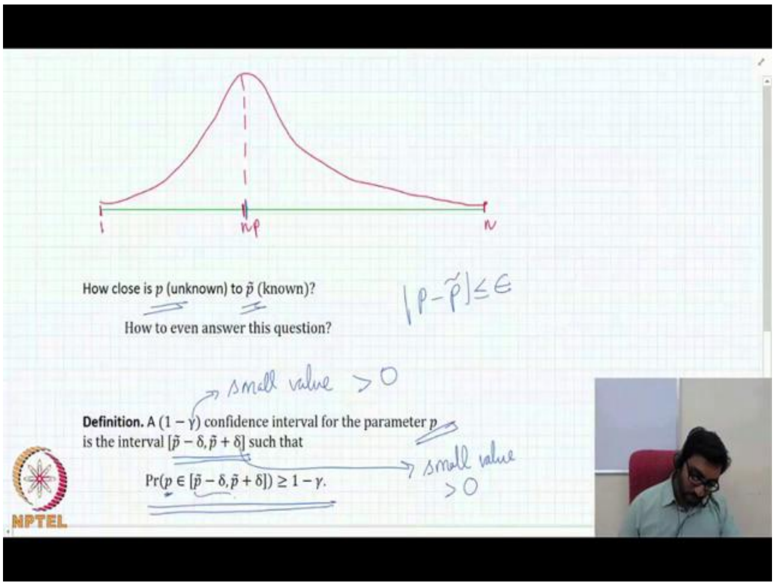
We are interested in finding out the fraction of people, the fraction P of people in this world in this population who like cricket. This is our parameter and this is the parameter that we want to estimate. Now, there is one way we could do that. We could literally go into the population as everybody whether they like cricket or baseball and find out that particular proportion P, the parameter P based on our responsive that we get. But that obviously is not feasible because the word is large.

Instead we will be resorting to sampling. So we are going to be sampling n individuals, and we will in based on the sampling that we do we will get a certain number of people who were respond saying, “yes I like cricket” and there is this random variable x. And that x is going to be some fraction of n that is that fraction is this P tilde. Now we going to use P tilde as an estimate for P, it is not going to be P perfectly but it is going to be an approximation or a good estimate of P.

Now suppose the population is very, very large in size compared to the number of samples that where going to make, then in essence every sample that we draw from the population is going to be independent of other samples that we made. So, with that sort of an assumption we can notice that the random variable X is actually a binomial random variable, because each time we sample we will be an our success is that the person we ask is a cricket fan and that success is going to happen with probability P. And we are going to count the number of such success.

This random variable X - therefore this is going to follow the binomial distribution. As a result the accepted value of E of X is going to be n times P. And if you want to draw the distribution for X its going to be looking something like that, so it mean it is going to be n times P and the distribution in a peak at n times P and it is going to have this sort of shape.

(Refer Slide Time: 03:45)



How close is p (unknown) to $\tilde{p}$ (known)?

How to even answer this question?

$|p-\tilde{p}| \leq \epsilon$

small value > 0

Definition. A $(1-\gamma)$ confidence interval for the parameter p is the interval $[\tilde{p}-\delta, \tilde{p}+\delta]$ such that

$$\Pr(p \in [\tilde{p}-\delta, \tilde{p}+\delta]) \geq 1-\gamma.$$

small value > 0

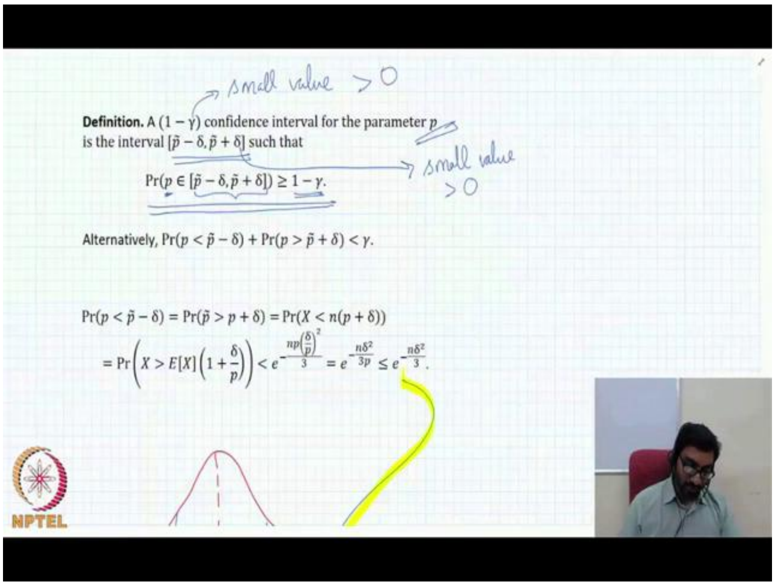
Now the question is the parameter that we want to estimate to; how close is P to P tilde which is the estimate that we have of P. Notice that this question is not as straight forward as we may like it be. For example, if we want P and P tilde to be very close to each other so the difference is at more some epsilon. So let us say we want P minus P tilde absolute value to be at most some epsilon. This is great, but there is always going to be some probability with which quality is going to be violated.

So, requiring such a guarantee is not necessarily a good thing. Instead in such cases we will have to consider a confidence interval, so let us define what a confidence interval is, so 1 minus gamma confidence interval. Gamma is going to be some small value strictly greater than 0. The 1 minus gamma confidence interval for the parameter P is an interval around P tilde. It is basically P tilde plus or minus some small delta and this delta again is a small value strictly greater than 0.

Such that, we have this following probabilistic guarantee and what is this guarantee, the probability that this parameter P lies within this interval is large in particular greater than

1 minus gamma. We are interested in a confidence interval for the parameter P with a very small gamma and a very small delta.

(Refer Slide Time: 06:09)



Definition. A $(1-\gamma)$ confidence interval for the parameter p is the interval $[\tilde{p}-\delta, \tilde{p}+\delta]$ such that

$$\Pr(p \in [\tilde{p}-\delta, \tilde{p}+\delta]) \geq 1-\gamma.$$

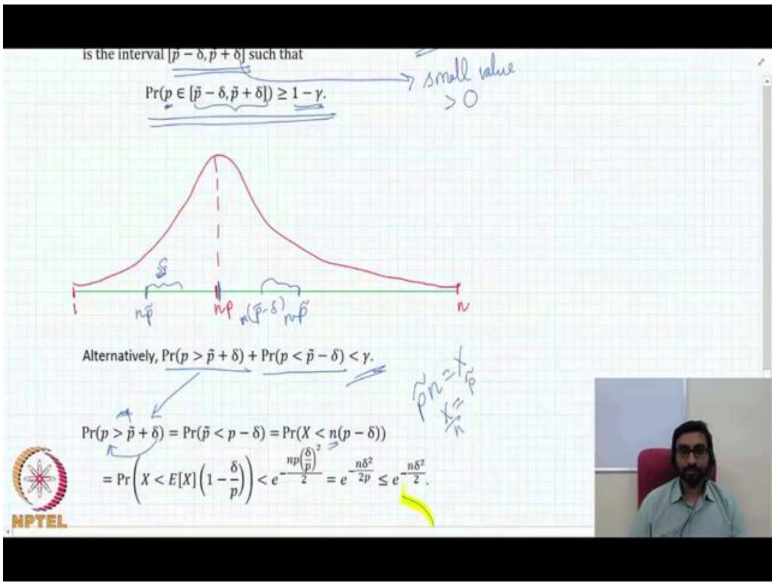
Alternatively, $\Pr(p < \tilde{p}-\delta) + \Pr(p > \tilde{p}+\delta) < \gamma$.

$$\Pr(p < \tilde{p}-\delta) = \Pr(\tilde{p} > p+\delta) = \Pr(X < n(p+\delta))$$

$$= \Pr\left(X > E[X]\left(1+\frac{\delta}{p}\right)\right) < e^{-\frac{np\left(\frac{\delta}{p}\right)^2}{3}} = e^{-\frac{n\delta^2}{3p}} \leq e^{-\frac{n\delta^2}{3}}.$$

Let us look at an alternative way to express this confidence interval. Here we are talking about the probability that the parameter P lies within an interval, that large probability. And equivalent we have stating that is the probability that it will fall outside of the interval should be small.

(Refer Slide Time: 06:51)



is the interval $[\tilde{p}-\delta, \tilde{p}+\delta]$ such that

$$\Pr(p \in [\tilde{p}-\delta, \tilde{p}+\delta]) \geq 1-\gamma.$$

Alternatively, $\Pr(p > \tilde{p}+\delta) + \Pr(p < \tilde{p}-\delta) < \gamma$.

$$\Pr(p > \tilde{p}+\delta) = \Pr(\tilde{p} < p-\delta) = \Pr(X < n(p-\delta))$$

$$= \Pr\left(X < E[X]\left(1-\frac{\delta}{p}\right)\right) < e^{-\frac{np\left(\frac{\delta}{p}\right)^2}{2}} = e^{-\frac{n\delta^2}{2p}} \leq e^{-\frac{n\delta^2}{2}}.$$

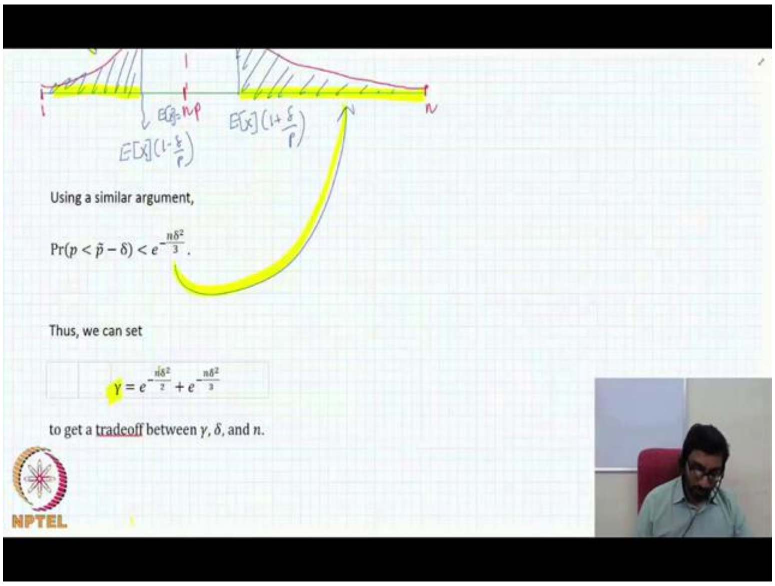
Let us get the help of this figure to understand this alternative form of the confidence interval. There are two ways in which the estimate P tilde can be far away from P. On the one hand P tilde can be so small that n times P tilde is very small, or alternatively it can be very large.

So, let us focus on the case where n time P tilde is very small. What is that mean? Well, let us ignore the n because we are trying to compare P tilde with P and the n is common so we can ignore that. Basically, even after another delta to P tilde P is still too far away that is one reason, I mean that is one way to interpret the case where P tilde is very small. So, if we want to formally state that we will say that is the probability that P is larger than even P tilde plus delta.

Similarly, the other way the P tilde can be too far away from P is, if its value is so large that n times P tilde is very large. And that can be expressed by this probability, the probability that P is smaller than P tilde minus some delta. So this will be P tilde minus some delta actually the n, n P is still smaller than that and that is capture by this delta. These are the two ways in which P tilde can be too far away, and these two events, their probability when some send up should be at most some small gamma. You can think of these two events are being the bear events and we want to upper bound their probability.

So, let us focus some one of the bear events. In particular focus on this particular bear event and bounds it is a probability, probability that P is where then P tilde plus a delta. Well, we can rewrite that event two probability that P tilde is less than P minus delta. So we are just taking the delta over to the other side. And notice that (Refer Time: 09:42) is recall that P tilde times n is your X. In other words X equals P; well, X by n is P tilde. And that is what we are going to do here, we going to replace P tilde with X by n except that n even a take it over to the other side.

(Refer Slide Time: 10:11)



Using a similar argument,

$\Pr(p < \tilde{p} - \delta) < e^{-\frac{n\delta^2}{3}}$.

Thus, we can set

$$\gamma = e^{-\frac{n\delta^2}{2}} + e^{-\frac{n\delta^2}{3}}$$

to get a tradeoff between γ, δ, and n.

Now we are close to applying chernoff bound, recall that n times P is E of X. So, if you pull out the P within here, you going to get E of X times 1 minus delta divided by P. And now we are ready to apply chernoff bound because now what we have is the probability that the binomial random variable is less than its excepted value times 1 minus some quality, And, so now we can apply a chernoff bound and the bound we will get is this one, e raise to the minus n P times delta over P square divided by 2. We can cancel out one of the P is here you get this (Refer Time: 11:16).

This P is a bit ugly we do not want that, but that easy to get read of because it is a probability and since P is always going to be upper bounded by 1, we can actually get rid of that and this becomes an n quality here, so we are finally left with this quality. And that as it turns out is the probability that the estimate P tilde will lie in this region. In a similar way, we can bound the other alternative, I mean the other bear event which is that P is less than P tilde minus delta and we will get a similar upper bound on that probability that bear event. And that will correspond to the estimate P tilde lying in this range

Finally, we have two bear events; we need to on get an upper bound of probability of the union of those two bear events. So we simple apply the union bound and that the upper bound on the probability of those two bear events should be at most gamma. Now we can get an expression for that gamma that will help us to get a trade-off between all the terms we have. This gamma let us say we set it to the sum of these two upper bound on the probability these bear events. So, gamma is equal to e to the minus blah blah blah divided by 2 plus e to the minus n delta square by 3, these are two upper bounds. And by setting it this way now we can clearly understand this trade-off. How? Well, let us look at it a little bit.

Now, let us say your boss wants you to estimate parameter P with the 1 minus gamma confidence interval P tilde plus or minus delta. Now the question is how many samples should you take? All we to do, is then apply this formula and solve for n and we will know how many times you will have to sample in order to get the appropriate confidence interval.