

Introduction to Probability – With Examples Using R
Professor. Siva Atherya
Theoretical Statistics & Mathematics Division
Indian Statistical Institute, Bangalore
Lecture No. 25
Markov and Chebyshev Inequalities

(Refer Slide Time: 00:16)

Standard Deviation Example

Definition 4.11: A discrete random variable X on the set Ω with $E(X) = \mu$ and $Var(X) = \sigma^2$.

Suppose: $X =$ Discrete random variable with values $1, 2, 3, 4$ and probabilities $\frac{1}{100}, \frac{2}{100}, \frac{3}{100}, \frac{4}{100}$.

Defn: $2 = \frac{X - E(X)}{SD(X)} \Rightarrow E(X) = 2.5$ (using symmetry)

$Var(X) = E[(X - E(X))^2]$
 $= E(X^2) - E\left[\frac{(X - E(X))^2}{100}\right]$
 $= E(X^2) - \frac{1}{100} E[(X - E(X))^2] = \frac{28.9}{100} = 0.289$

Example: $X =$ Uniform $(1, 10)$
 $\mu = \frac{1+10}{2} = 5.5$
 $\sigma^2 = \frac{(10-1)^2}{12} = \frac{81}{12} = 6.75$

$P(|X - \mu| < c) = P(|X - 5.5| < 2.9)$
 $= P(2.6 < X < 8.4)$
 $= P(X \in (2.6, 8.4)) = \frac{5.8}{10} = 0.58$

Defn: (Mean) Let X be a (μ, σ^2) random variable.
 $Var(X) = \sigma^2$.

Ex: $\mu = 50.5, \sigma^2 = 28.9$
 $\Rightarrow P(|X - \mu| <= c) = 0$

Ex: $\mu = 50.5, \sigma^2 = 28.9$
 $\Rightarrow P(|X - \mu| <= 28.9) = 0$

Let me start again. So, I hope you all can hear me. So, the idea is that we were discussing the previous lecture one issue of standardization. So, there we fixed the mean and the variance to be mean to be 0 and the variance to be 1 that is kind of standardization. The next thing I would like to do is I would like to understand where the random variable actually concentrated at so it means it close to its mean far from its mean.

And what we decided was we try in statistics a very important question to understand what how much of the random variables outside this interval of $\mu - \sigma$ to $\mu + \sigma$ or $\mu - 2\sigma$ to $\mu + 2\sigma$ this is 3 sigma or 6 sigma nowadays and so on so forth. And idea so one once one simple computing we are going to do is take uniform random variable but between 100 and compute its mean to be 50.5 that is quite easy.

The standard deviation is 28.9. Then you try and compute the basic probability of probability $X - \mu$ is less than C . So, that gives you a chance that X is between 22 and 79 that is about 58 by 100. So, there is about 58 percent chance the random variable lies in $\mu - \sigma$ to $\mu + \sigma$. And you can clearly see if if you multiply σ by 2 $\mu - 2\sigma$ negative close to negative is negative in fact 0. $\mu + \sigma$ was about near 100. So, $\mu - 2\sigma < X - \mu < 2\sigma$ is like 0. There is no chance that X can be there.

So, in some sense that this gives you an idea of how the random variable is spread out. So, in general I wanted to ask this question if I just give you a random variable X and tell you which means μ can I somehow tell you without computing the probabilities the chance that X minus μ is bigger than k series. That is what we will do now these are very important inequalities in probability which we will discuss and statistics which is quite useful everywhere gives you a first gut feel expression of what manner will do. So, here is the

(Refer Slide Time: 02:42)

Theorem 4.3.3 [Markov Inequality]

Let X be a discrete random variable which take on only non-negative values [i.e. $\text{Range}(X) \subseteq [0, \infty)$]. Suppose X has finite expected value. Then for any $c > 0$ where $\mu = E(X)$

$$P(X \geq c) \leq \frac{\mu}{c}$$

Proof:- [Sketch] $T \equiv \text{Range}(X)$ - Countable subset of $[0, \infty)$

$$\mu = \sum_{t \in T} t P(X=t)$$


So, here is the first thing this is a very powerful inequality called Markov inequality. Which is very well very very widely used some probability. So, here the idea is that let X be a discrete random variable which takes on non-negative only non-negative values which takes on only non-negative values that means the range of X is in the positive half plate. That is the range of X is contained in the non-negative probability.

And let us say again let us suppose X has a finite expected value than for any C positive we see positive probability that X is greater than equal to C is less than equal to μ by C . So, μ again as before always we just reiterate one last time μ is always E of X . So, something that one very useful so chance that μ lies about C is always given by at the most by μ by C so that is something very useful idea.

The proof is also not so hard I will just try to prove it. So, here I will add a sketch I will I will do a sketch I am going to do a precise too. And so the range is the range of X is countable so it is a countable subset countable subset of the real value of the positive real value. Because, range is counted. So, here I will write μ is the same as the sum over t in the range of X let

we call this as t is the range of X there is a confidence of t and t times the chance that X is equal to t .

(Refer Slide Time: 05:39)

Fix $C > 0$:

$$= \sum_{\substack{t < C \\ t \in T}} t \cdot P(X=t) + \sum_{\substack{t \geq C \\ t \in T}} t \cdot P(X=t)$$

$\Rightarrow \mu \geq 0 + \sum_{\substack{t \geq C \\ t \in T}} t \cdot P(X=t)$

$\Rightarrow \mu \geq C \cdot \sum_{\substack{t \geq C \\ t \in T}} P(X=t)$

So, now here this is the same as so now C is some number that I have right so let us fix this C lets fix the C positive. So, what I do is I go and divide the sum into two parts sum sum over t is less than C and t is in t plus the sum when t is bigger than C greater than equal to C and t is in t . So, this I do this and I do the same thing probability X equal to t . This is what it t is less than C and t in t . So, now once I have this again this is again a formal computation you have to make it rigorous.

So, in this in this first one in this first one everybody all this t is less than the C . In the second one all the t is at least some t right. And I know I know t is contained in 0 infinity. So, this whole first term is is non-negative so let me erase this part. This whole first term is non-negative t is non-negative and probability X square t is non-negative this first term is non-negative this is first term.

So, what I could do is I could just throw it away I could say my μ so this will imply when μ is always going to be bigger than equal to I will just throw this away I will just put this as 0 plus I will write whatever is here. The sum over t is bigger than equal to C bigger than equal to C . The chance that t times probability X so I am just throwing it away. And this t in T .

In the second step what I do is I observe that everybody in the sum you know t is non-negative but more importantly everybody is about C . So, this all these t 's are above C . So

now I can do a similar computation I can say this implies mu's are now again I forget the 0 0 is gone. I have everybody is above C so I pull the C out. And then I will leave the sum as a sum of t bigger than equal to C t is in T, t is in T probability X equals t is in T probability X equal to t. Here, I will just replace the C minus C so I remove the C outside. I can do that because every term is bigger than equal to C so our sum has to be. Now, this will in this this will imply what?

(Refer Slide Time: 08:33)

$$\begin{aligned}
 &= \sum_{\substack{t < C \\ t \in T}} \dots + \sum_{\substack{t \geq C \\ t \in T}} \dots \\
 &\Rightarrow \mu \geq 0 + \sum_{\substack{t \geq C \\ t \in T}} t \cdot P(X=t) \\
 &\Rightarrow \mu \geq C \sum_{\substack{t \geq C \\ t \in T}} P(X=t) = C P(\cup_{t \geq C} \{X=t\}) = C P(X \geq C) \\
 &\Rightarrow \mu \geq C P(X \geq C) \Rightarrow P(X \geq C) \leq \frac{\mu}{C}
 \end{aligned}$$

This will imply now that mu is bigger than equal to C times. What is this now this expression right here? Let me let me go to the next screen I will come back again. This expression right here is just everybody above C right. So, it is just it is a it is a disjoint union on everybody. So, what you do is you write this as C times the probability of the union of t bigger than equal to C t in T X equal to t and that is what that is.

But, that is just the same as C times the probability of X bigger than equal to C. There is no difference so that is your result you have shown so this will imply let me make this a little bit smaller so you can always move a little bit higher so I can do it in one page here so everything on one page. So, this whole thing will imply that mu is bigger than equal to C times probability X bigger than equal to C. That implies that what you want chance that X bigger than equal to C is less than equal to mu by C.

That is what we wanted to show that is a one page view. So, two things is crucial if we use the fact that it is non-negative because we throw the first sum away. So, that is where we use the fact that the range is non-negative that is a very crucial step. It is a very crucial step here

if the range are non-negative you cannot pull that sum out of it. And this part we use the fact that X is bigger than equal to C .

And here we use the fact that the the axis of probability the sum of the probability of a union of disjoint events is the same as the sum as the second axis of probability. So, this both this gives you a very interesting fact that the chance that X is bigger than C is at the most μ by C . it cannot be more than that. So, if C is large your μ by C is going down. That is a little better.

So, now there is sort of a this only works a long negative random variable so you might think that okay it is sort of it is going to be right Markov (())(11:27). So, you might think okay it is non-negative so it is restrictive it is not you cannot really apply it otherwise for all random variables. But, sort of a simple generation.

(Refer Slide Time: 11:49)

Chebyshev's Inequality :- X - Discrete random variable
 with finite non-zero variance. Then for any $k > 0$,

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Proof: $(|X - \mu| \geq k\sigma) \equiv$ event same as $(X - \mu)^2 \geq k^2\sigma^2$

now $Y = (X - \mu)^2$ non-negative random variable
 $E[Y] = E[(X - \mu)^2] = \sigma^2$

So, let us just do that this due to Chebyshev inequality. So, see what you do is you take X to be a random variable with finite variance actually random variable so X is random variable. So, X is a discrete random variable with finite variance with finite non-zero variance. So, now what happens is that so then you know we asked this idea right we wanted to know. Then for any k bigger than equal to 0 we wanted to know what is the chance that X minus μ is bigger than equal to k times sigma that is what we want to understand.

What is going to be there? So, there is a simple bound for this this is always going to be less than equal to 1 by k square. So, this is the Chebyshev inequality that here I do not assume anything or random variable. I just say it is non-zero variance just finite non-zero variance

implies finite mean. And this is. So, how does one prove this? Here is a proof. So, the the trick is that you want to apply Markov inequality to a non-negative random variable. X is not non-negative. So, you do not apply to X but you are almost close by you know you have X minus μ in absolute value and that is the trick.

So, what you think about is the formula so X minus μ is bigger than k times sigma right this is what you have this is the event. This event is the same as as the event so I want to make it non-negative and make it easier so I will think of it as the event X minus μ the whole square is bigger than k square sigma square. Same event if I square both sides.

Now, if I take the random variable Y it is X minus μ the whole square. That is like a non-negative random variable equal to square. And I know E of Y of the random variable is E of X minus μ whole square and that if you stay a little bit it is just the variance of X or just sigma square. That is easiest.

(Refer Slide Time: 15:19)

now $Y = (X - \mu)^2$ non-negative

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

Apply Markov Inequality to Y ,

$$\text{take } C = k^2 \sigma^2$$

$$P(Y \geq C) \leq \frac{E(Y)}{C} = \frac{\sigma^2}{k^2 \sigma^2} = \frac{1}{k^2}$$

i.e.

$$P((X - \mu)^2 \geq k^2 \sigma^2) \leq \frac{1}{k^2}$$

$$\Rightarrow P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \quad \square$$



So, now I can apply Markov inequality to Y to Y . To Y with what with with C equal to let us see so I want to know the chance that I want to know what do I want to estimate I want know X minus μ is bigger than k sigma. So, let us fix a k let us fix a k . So, k positive I guess. So, so now I I I I fix it C with just C by C as just a one over k or or case k . So, or k sigma lets say.

So, then what I do I just know that the problem takes C to be equal to is equal to k sigma k square sigma square so then I know that the probability of Y greater than equal to C is less than equal to the mean of Y divided by C . That is what it does it like. So, now what is C what

is mean of Y ? Mean of Y is $\sigma^2 C$ is $k^2 \sigma^2$ so I get 1 over k^2 .
And what is this side what is the chance that Y is bigger than C .

That is the same as what is Y ? Y is X minus μ whole square. And, what is C ? C is $k^2 \sigma^2$ that is this this is less than or equal to 1 over k^2 . But, that is the same as the chance that the absolute value of X minus μ is bigger than k times σ is less than 1 over k^2 . That is the proof. So, a fairly simple proof of the event of this problem observation. All you have to do is apply Markov inequality so it is a very useful result.

So, that means you know that the chance that X minus μ is about k times σ is always bound by 1 over k^2 . So, in decay of course this decay may not be enough for you but this is a decay. A problem. And there are there are there are other inequalities in probability which tell you how little bit if you understand more and more properties about about X you can get further descriptions about how close X is to its mean.

There is one simple inequality that tells you that X is within k units of standard deviation is like 1 over k^2 . This is an example keep in mind it is very important. So, two inequalities are very important in probability and statistics so I will just reiterate them a little bit. Let me show you the previous one. So, one is Markov inequality and the other one is Chebyshev inequality. So, both these are very very important aspects in probability.

So, one just says that if X is non-negative a brute force estimate for the tail probability is just μ by C . So, X is bigger than equal to C is kind of the tail of the distribution so you just say X between equal to C is like μ by C . And Chebyshev inequality says that you can quantify how far X from its mean within some units of standard deviation in that by k^2 . So, it is k units of a means that 1 by k^2 .

So, now these are ideas about how far random variables have mean and so on so forth now I want to also understand if I know something has happened how does my mean and variance so I just want to go back and connect the conditional problem a little bit. So, hereso now I want to do this thing.

(Refer Slide Time: 19:57)

4.4 Conditional Expectation and Conditional Variance

- X - Discrete random variable. A - event, $P(A) > 0$.

$Y = X | A$ \leftarrow Conditional distribution of $X | A$

$$P(Y=t) = P(X=t | A) = \frac{P(X=t, A)}{P(A)}$$

Example 4.4.2 Roll a die X - outcome.
 $A = \{2, 4, 6 \text{ occurs}\}$
 $k \in \{1, 2, 3, 4, 5, 6\}$
 $P(X=k) = 1/6$



Is called it is 4 point 4 my book it is called conditional variance expectations and variance. So, here I want to understand the the following idea. So, before we knew that if a random variable X was given and an event a was given we knew that the distribution of X changed with if you know that the a event a happened. So, for example let us just do a simple start a simple computation so let me do an example let me start with a simple example here. Let us say example so before that let me just let me just write down.

So, so X is discrete random variable and A was a event. The probability of A is positive. We knew that and we knew that if like Y was the random variable X condition of the event A the probability of Y equal to some number t was given by probability of X equal to t given A and that's the same as X equal to t and A are divide by probability of A . So, this is this was what we called as the conditional distribution of X given A . That is how we define this whole idea. So, now we know that let us take an example for example so we we knew that the following idea happened these things changed quite a bit so for example let us take an example 4.4.2 in my book.

Let us say so here let us say let us say you roll a die you know roll a die. And let us say X is the outcome X is the outcome. Then let us say A is the event the outcome is even so is the event that 2 4 6 occurs. So, if a the roll is even. So, now we know that probability X equal to k is 1 over 6 priory for k in in 1 2 3 4 5 and 6 that is just uniform. X is just a uniform.

(Refer Slide Time: 22:51)


$P(Y=t) = P(X=t|A) = \frac{P(X=t, A)}{P(A)}$

Example 4.4.2 Roll a die $X = \text{outcome}$.
 $A = \{2, 4, 6 \text{ occur}\}$, $P(A) = 3/6 = 1/2$
 $k \in \{1, 2, 3, 4, 5, 6\}$ $X = \text{Outcome } \{1, 2, \dots, 6\}$

- $P(X=k) = 1/6$
- $P(X=k|A) = \frac{P(X=k, A)}{P(A)} = 2 P(X=k, A)$

$k = \text{even}$
 $(X=k, A) = X=k$
 $k = \text{odd}$ $(X=k, A) = \emptyset$

$(E_x) = \begin{cases} 0 & k = \text{odd} \\ 1/3 & k \in \{2, 4, 6\} \end{cases}$



But, we also know that probability X given A X equal to k given the event A . How do we do this this is the same as probability X equal to k and the event A divided by the probability of A . Now, that is going to be equal to what probability of A is just 3 by 6 so probability of a is 3 by 6. And that is a half that we know. So, this we know is going to be twice the probability of X equal to k and A . This is what we are going to get right so let us say k and A .

So, now if you go a little bit further this is going to be what so now if k is odd X equal to k and A does not occur as a 0. That is nothing happens. If X equal to even then the chances are it is just that number k and A is a number. So, this is a simple exercise one can just check that this is going to be equal to just twice the one-sixth so you just get one-third right this is one-third if k is in is in 2, 4 and 6.

And this is because of the fact that X equal to k and A when k is even this event is same as X equal to k . So, this is the idea the operation is that k is even then the event X equal to k and A is the same as X equal to k . And if k is odd X equal to k and A is empty set. That is the idea that is the that is what you use to show this result. So, now you know that the distribution of X given in A is very different from the distribution of X alone by itself.

So, the random variable changes a lot. So similarly, the the mean also will change so for that we need to identify what do you mean by a conditional mean. So, here is a definition.

(Refer Slide Time: 25:21)

Definition 4.41: Let $X: S \rightarrow T$ be a discrete random variable. $A \subseteq S$ - event such that $P(A) > 0$.
The Conditional Expected value is defined from the conditional probabilities in the same way the (ordinary) Expected value is defined from (ordinary) probabilities. I.e.
$$E[X|A] = \sum_{t \in T} t \cdot P(X=t|A)$$



Here is the definition. Let X from S to T be a discrete random variable random variable. Let A be an event A be an event in S such that a probability of A is positive. So, then so this is something we have to be careful about the conditional expected value. So, the conditional expected value so value is defined from the conditional probability is defined from the conditional probabilities. In the same way as the ordinary problem as the expected values is.

In the same way the let us say let me write differently the ordinary expected values of ordinary expected value is found from the is defined from the ordinary probabilities. So, that is how do you do this now so like so that is E of that is E of X given the event A that is a conditional expectation of X given the event A is defined as the sum over t in T the range of t such that you multiply t and before you put just probability X equal to t that is the ordinary problem.

But, now you put the conditional problem so what you will do is so let me write this down that is in blue the conditional expectation of X given the event A is going to be this condition on the event that A happened. So, you replace it by the conditional probabilities not the original problem. So, is that clear so the idea is that you do not use the original probabilities you use the conditional probabilities to compute the condition expectation of X .

(Refer Slide Time: 28:38)

Extending this similarly,


$$\text{Var}[X|A] = E[(X - E[X|A])^2 | A]$$

Example 4.4.2:

$$E[X|A] = \sum_{t \in \{1, 2, 3, \dots, 6\}} t \cdot P(X=t|A)$$

$$= 2 \cdot \left(\frac{1}{3}\right) + 4 \cdot \left(\frac{1}{3}\right) + 6 \cdot \left(\frac{1}{3}\right) = 4$$

$$\text{Var}[X|A] = E[(X - 4)^2 | A]$$

$$= (2-4)^2 \cdot \frac{1}{3} + (4-4)^2 \cdot \frac{1}{3} + (6-4)^2 \cdot \frac{1}{3} = \frac{8}{3}$$


And similarly, similarly, so, extending this similarly we have the conditional variance of X given A X given A that is going to be the same as the you will always use the condition expectation again of X minus E of X given A whole square. So, you will use the so I will use the conditional expectation of so I will condition it right but what is inside inside I will put X minus I will again put the I will always work with the continuity of X given A whole square.

So, variance also defined as the condition expectation value of the same. Very nice. So, this is what so let us go to the example again let us continue the example let us try and do this computation. So, example what is the example the example is the previous page that s for the example. Example was the number it 4.4.2 let me continue it. So, there what we did I had to do I found out that the so if I want to find E of X given A E of X given A in that example so what I get.

So, I had a here E of X given A was the same as the sum over t in T just t in 1 to six t times the chance that so here I would have the chance that X equal to t given A that is the probability. So, X was the original variable so I will write down it in black. But, that I know it is it is we know it is it is 0 for odd and and one third for even. So, I will get 2 times one third plus 4 times one third plus 6 times one third so that is going to be equal to something like 8 from like 4. The 2 plus 4 is 6, 6 plus 6 is twelve divided by 3 is to 4.

And now you can do the same thing for the variance so the variance the random variable of X given A So, what is that going to be that is going to be again we look at this formula it is E of what is X X is X minus this random variable is going to be you have X given A we found out to be 4 so on this column is 4 this whole square given the event A. But, this again we know is

going to be the same as so what do you do you you know X takes under A X only takes even value the order is all 0.

So, you again get 2 minus 4 the whole square times 1 third plus 4 minus 4 the whole square times 1 third plus 6 minus 4 the whole square times 1 third. And that will give you 8 by 3.

So,

(Refer Slide Time: 32:22)

Example 4.4.2:

$$E[X|A] = \sum_{t \in \{1,2,3, \dots, 6\}} t \cdot P(X=t|A)$$

$$= 2 \cdot \left(\frac{1}{3}\right) + 4 \cdot \left(\frac{1}{3}\right) + 6 \cdot \left(\frac{1}{3}\right) = 4$$

$$\text{Var}_A(X|A) = E[(X-4)^2|A]$$

$$= (2-4)^2 \cdot \frac{1}{3} + (4-4)^2 \cdot \frac{1}{3} + (6-4)^2 \cdot \frac{1}{3} = \frac{8}{3}$$

- Compare $E[X]$ with $E[X|A]$ and $\text{var}(X)$ with $\text{var}(X|A)$

So, one interesting is that if you just compare E of X with E of X given A and variance of X with variance of X given A so let me just do a split view and try and understand this a little bit better global way. Let me go back here let us come down here let us just tell a little bit it is a nice competition let us go back here.

So, now if you start this off then the idea was that if you give an event A you know the random variable changes and this we have seen many times before. So, we understood that the conditional law definitely changes. For example, it was 1 third for all these and 0 for otherwise. And then I I said that the conditional mean I defined by saying that it is just over the conditional probabilities. And the conditional variance is again the same thing over the you take the conditional mean you subtract it you try and see how it differs under the conditional problem.

So, now you found out that this that the conditional mean of the of the uniform given the event A is 4. But, the original mean we also know E of X is just 1 plus the sum of the mean sum of 1 plus 2 plus 3 plus 4 5 by 6 by 6 that is just going to be 3.5. So, this is also very

known in that sense let me know so this kind of it is a it is a nice it is a nice way of understanding that, that the mean kind of gets a little bigger and the variance is 8 third.

But the variance of the original one is like let me write the original ones let me write E of X in this case was Markov back the whole screen so let me go next page in quite bigger so if you come here then you can notice that the E of X is like before was like 3.5. Here it is like 4. So, it is a little bigger. And the variance of X if you did before it is like I think the computation is I think 35 by 12.

And so you know that it is kind of close by it is not changing much but it is a little less. But the means kind of jumps up so it is kind of an interesting idea so if you give an event A you can increase the mean if you decrease the mean if $(\cdot)(35:05)$.