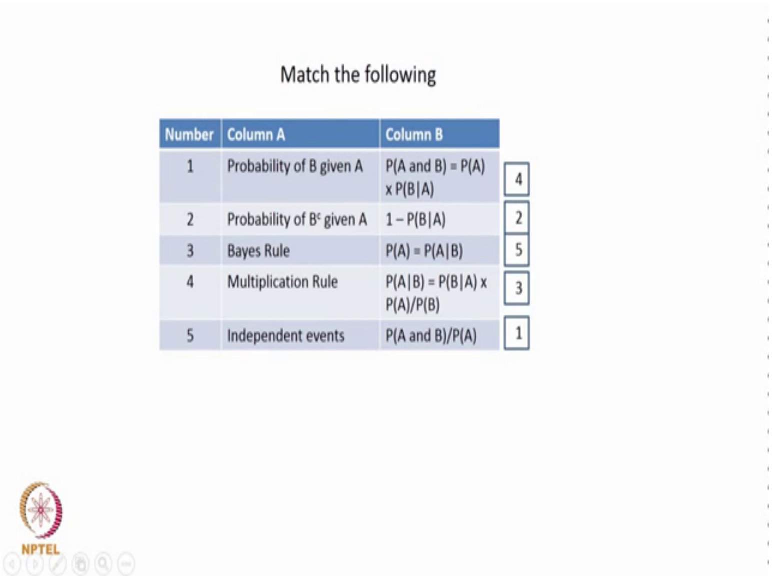
Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture - 16
Random variables

In this lecture we discussed conditional probability further. We take some simple examples to understand and illustrate the concepts that we learned in the earlier lecture.

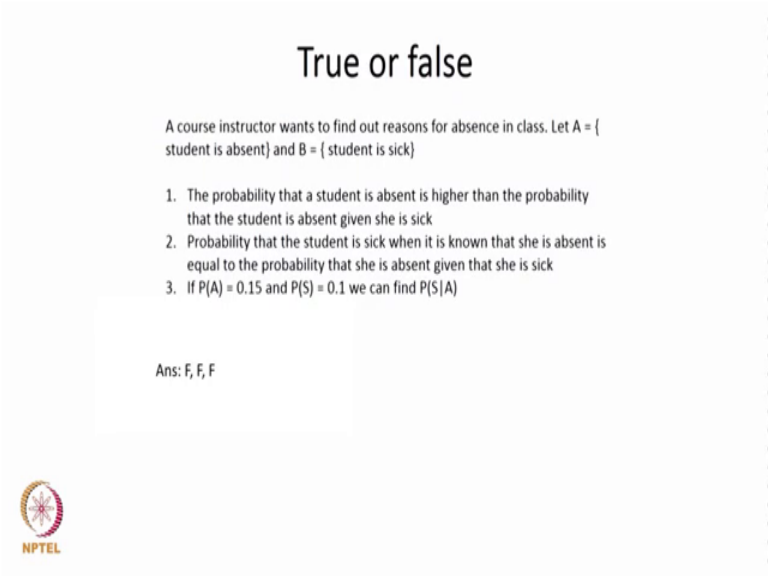
(Refer Slide Time: 00:30)



So, begin with some match the following questions so that we understand the concepts. There are 5 items given in column A and 5 items given in column B. So, probability of B given A probability of B complement given A Bayes rule multiplication rule and independent events. Probability of B given A is P of A and B divided by P of A is the equation that we saw in the earlier lecture.

So, that is the answer. Probability of B complement given A is relatively easy that is 1 minus probability of B given A which is shown as 2 here. The earlier one was shown as 1 here Baye's rule different forms. So, one of the equations is P of A given B is equal to P of B given A into P of A by P of B which is given which is Baye's rule from here. Multiplication rule is 4 P of A and B is equal to P of A into P of B given A and independent events P of A is equal to P of A given B because A and B are independent and therefore, P of A is given B.

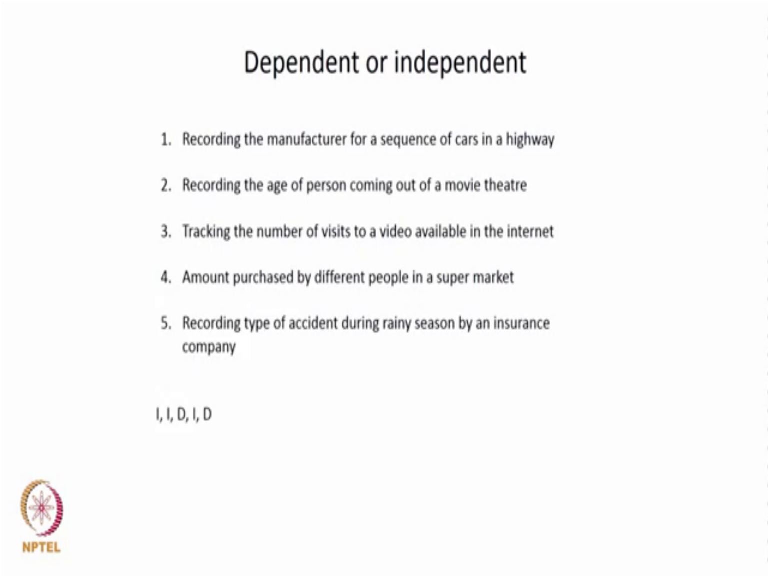
(Refer Slide Time: 01:48)



Next we look at some true or false questions. Now the course instructor wants to find out reasons for absence in class. So, let A be the event that the student is absent and B is an event the student is sick.

So, probability that a student is absent is higher than the probability that the student is absent given she is sick. So, the answer is false because generally absence could involve other reasons other than sickness and because we restricted to a reason that is sickness. So, probability of the person is absent given that the percent is sick will be much higher than the probability that the student is absent considering other things other than sickness. Therefore, the answer is false probability that the student is sick when it is known that she is absent is equal to the probability that she is absent given that she is sick this is comparing P of A given B and P of B given A and we have already seen that P of A given B is not equal to P of B given A. And therefore, the answer to this question is a false. If P of A there is P of absent is 0.15 and P of sick P of S is 0.1 then we can find P of S given A, we need to know P of A and S which is not given and therefore, we cannot find the probability of P of sick given absent with the data that is being provided.

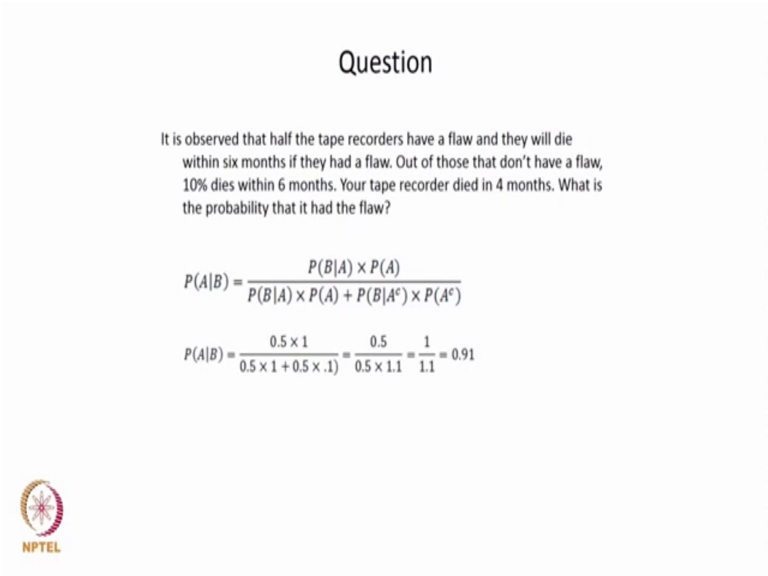
(Refer Slide Time: 03:14)



So, we also want to check whether some of these are independent or dependent. So, recording the manufacturer of A sequence of car is in a highway are not dependent so much on each other. So, they could become the sequence of cars can be independent recording the age of A person coming out of A movie theater. Again we could generalize it by saying independent though one might argue that there is a larger proportion of people belonging to a certain age bracket who would go for movies, but by and large if we assume that people of any age group, go to a movie then it is independent. Tracking the number of visits to a video available in the internet can be dependent that would depend on an earlier visit and so on. Amount purchased by different people in a supermarket does not depend for it is independent, recording the type of accident during rainy season by an insurance company.

So, moment there is a rainy season you could have more accidents that come out of skidding it would come out of other rain related things. So, there is a dependence in the whole process.

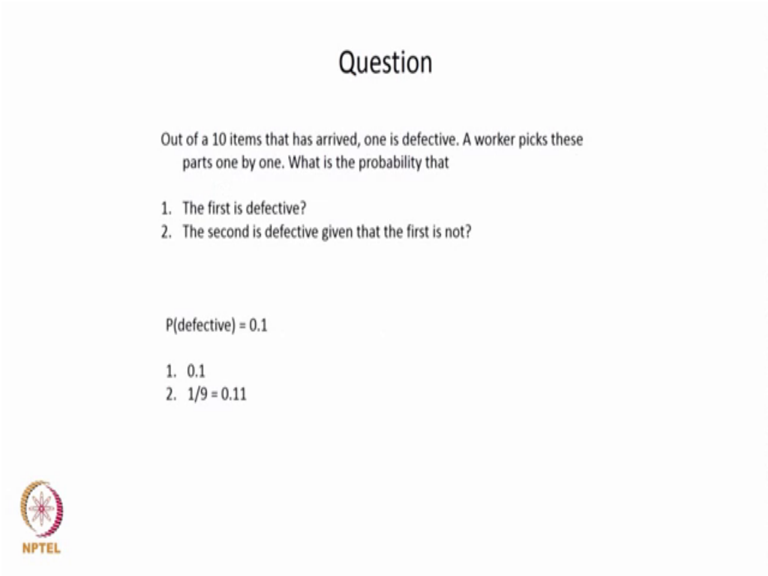
(Refer Slide Time: 04:17)



So, look at a few more questions, one question is it is observed that half the tape recorders have a flaw and such tape recorders would die within 6 months, if they had a flaw. Die meaning, they would stop working within 6 months if they had a flaw, out of those that do not have a flaw 10 percent would anyway stop working within 6 months.

Now, your tape recorder died in 4 months, which means it stopped working in 4 months. What is the probability that it had the flop. So now, you want to use the Baye's equation. So, probability that it died or it did not work given that it had a flaw is given by this equation so, 0.5 into 1 by 0.5 into 1 by because this one comes because if they have a flaw within with 0.5 it will anyway die within 6 months. So, since it died within 4 months 0.5 into 1 plus 0.5 into 1 plus 0.5 into 0.1. So, you get 0.5 divided by 0.5 into 1.1. So, 91 percent that it had the flow.

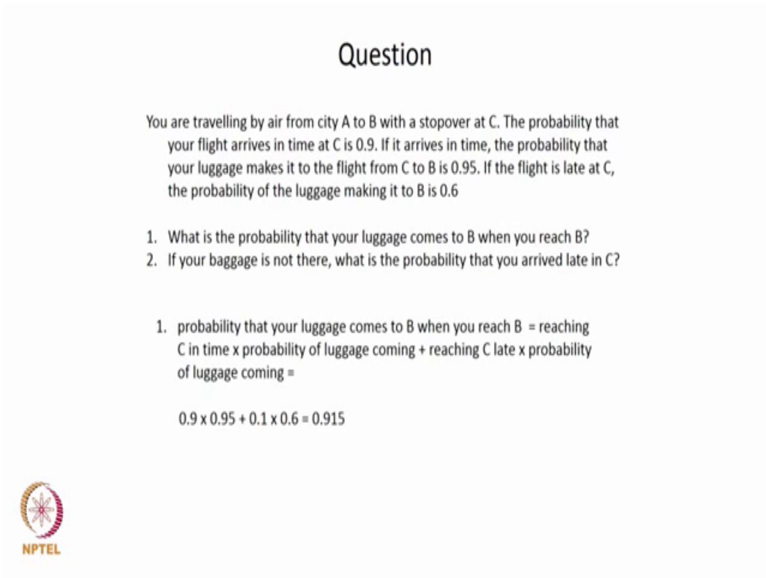
(Refer Slide Time: 05:23)



Out of 10 items that has arrived one is defective worker picks these parts one by one, what is the probability that the first is defective and the second is defective given that the first is not defective.

So, the first is defective probability of defective is 0.1. So, and there are 10. So, you pick any one randomly, and 1 by one implies it picked randomly and 0.1. Given that the first is not what is the probability that it is. So, it is 1 by 9 because the first one is not defective. So, there are 9 remaining parts out of which one is defective and therefore, it is 1 by 9 and it is 0.11.

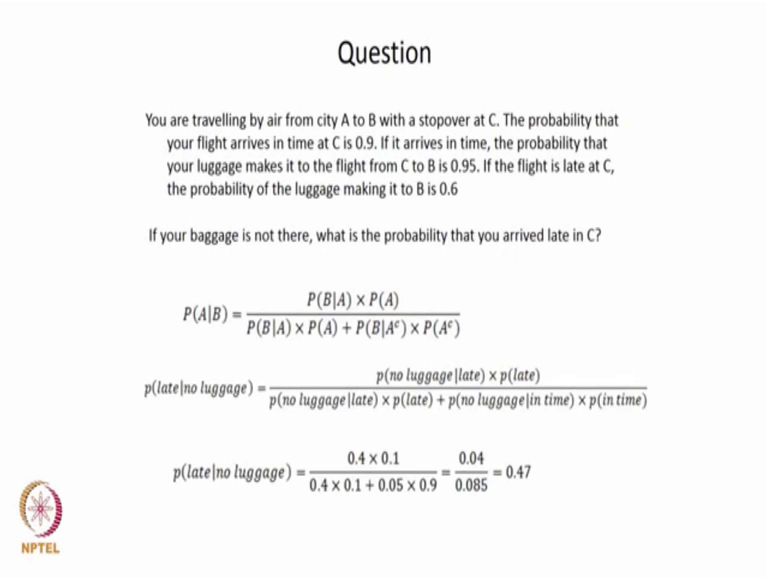
(Refer Slide Time: 06:05)



Then we look at one more question you are traveling by air from city A to city B with a stopover at C. The probability that your flight arrives in time in the intermediate airport is 0.9 and if it arrives in time the probability that your luggage makes it to the flight from C to B is 0.95. If the flight is late in the intermediate then the probability of luggage making it is 0.6. What is the probability that your luggage comes to B when you reach B. And if your baggage is not there when you reach B what is the probability that you arrived late at C which is the intermediate place.

So, question number 1, probability that your luggage comes to B when you reach B is reaching C in time and luggage going from C to B and reaching C late into probability of luggage going from C to B. So, reaching C in time is 0.9 and luggage going is 0.95. So, reaching late is 0.1 and luggage going is 0.6. So, 0.9 into 0.95 plus 0.1 into 0.6 which is 0.915.

(Refer Slide Time: 07:19)

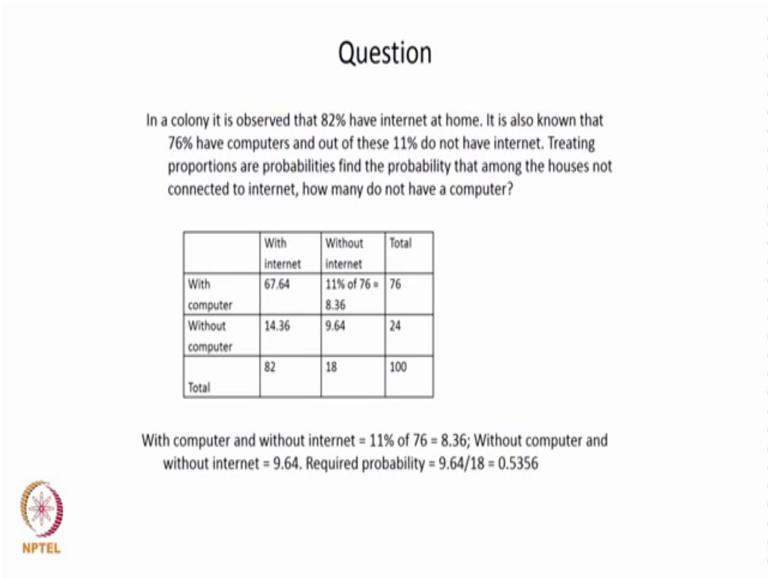


Second if your baggage is not there when you reach your destination B, what is the probability that you arrived late in the intermediate one C. So, the general equation is given here, which is now expanded to probability given that there is no luggage what is the probability that you arrived in C later.

So, probability of arriving in C late given that there is no luggage when we reached B is probability of no luggage given late into probability of late, divided by probability of no luggage given late into probability of late, plus probability of no legates given in time into probability of in time. So, probability of no luggage given late is 0.4 because probability given late no luggage 1 minus 0.6 which is 0.4 probability of late is 0.1 because arrival is 0.9.

So, 0.4 into 0.1 plus same 0.4 and 0.1 comes here plus 0.05 into 0.9. So, probability of in time is 0.9 which is here and then 0.95 percent is when it goes. So, 0.05 is the probability that it does it goes late the luggage does not come. So, 0.05 into 0.9; so, 0.04 by 0.085 which is 0.47.

(Refer Slide Time: 08:47)

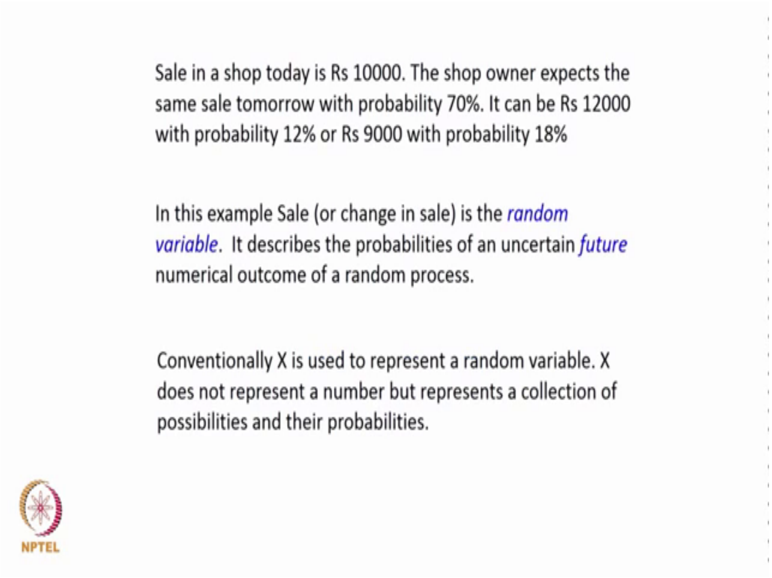


In a colony it is observed that 82 percent of the people have internet at home it is also known that 76 percent have computers and out of these 11 percent do not have internet. Treating proportions as probabilities, find the probability that among the houses not connected to the internet how many do not have a computer.

So, we could assume or save without any loss of generality that there are 100 houses. So, there are 100 houses which is shown here. Which is 100, 82 percent have internet. So, we create this table with computer without computer with internet, without internet. So, the total here with the internet is 82 therefore, without internet is 18. 76 percent have computers. So, with computers total for computer is 76 total without computer is 24. So, 11 percent do not have internet. So, with computer without internet is 11 percent of 76, which is 8.36 and therefore, we can fill the rest of the table 67.34 14.36 and so on.

So, without computer and without internet is 9.64. So, probability is 9.64 by 18 and 0.5356 is the answer. So, with this we finish our discussion on some aspects of probability as well as conditional probability. Now we move to the next topic in probability which is random variables. And will spend some time on understanding random variables. And then we will work out some examples to further our understanding.

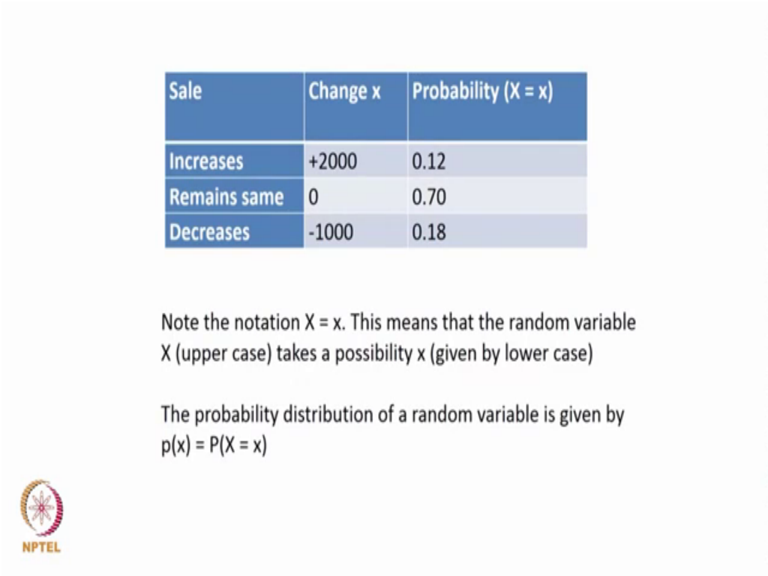
(Refer Slide Time: 10:27)



Now, let us start by saying sale in a shop today is 10000 the shop owner expects the same sale tomorrow with a probability of 70 percent. It can be 12000 with a probability of 12 percent or 9000 with a probability of 18 percent. So now, we have the sale in a shop as a variable. And this variable according to the sentence or text given can take 3 values, which is 10000, 12000 and 9000 with probabilities 70 percent 12 percent and 18 percent adding up to 100 percent. Again proportions and probabilities are used interchangeably. So, probability is expressed as a percentage.

So, in this example the sale or change in sale as the case may be is the random variable. So, here is a variable which can take multiple values with defined probabilities. So, it becomes a random variable. So, it describes the probability of an uncertain future, numerical outcome of A random process. So, conventionally capital X or uppercase X is used to represent a random variable. X does not represent a single number, but represents a collection of possibilities and the probabilities associated with them.

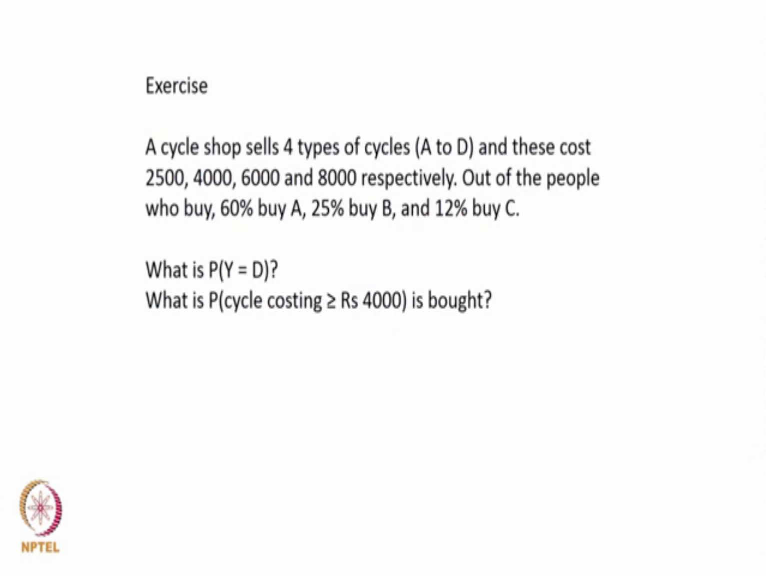
(Refer Slide Time: 11:43)



So, in this case change X is the variable we are looking at. So, increase is plus 2000 is a change with probability 0.12 if it remains the same 0 with probability of 0.7 and if it decreases it is minus 1000 with the probability of 0.18.

Note that, the notation capital X equal to small x means that the random variable X uppercase takes a possible value of small x given by the lowercase. So, probability distribution of A random variable is given by P of x equal to P of X equal to x where in X equal to x the uppercase X is the random variable and the lowercase x is the value that the random variable can take or a possibility that the random variable can take.

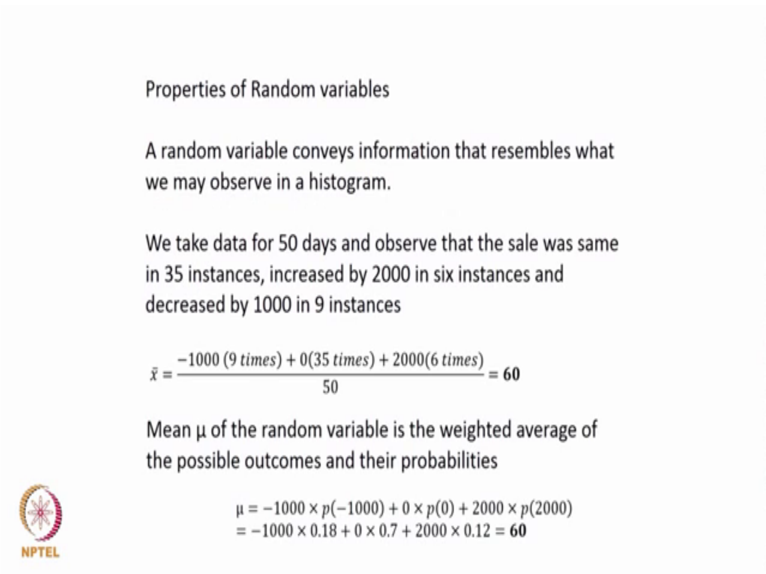
(Refer Slide Time: 12:32)



A cycle shop sells 4 types of cycles A to B and these cost 2500, 4000, 6000 and 8000 respectively. How do the people who buy 60 percent buy A 25 percent buy B, 12 percent buy C. So, 60 plus 25 is 85 plus 12 is 97.

So, what is P of Y equal to D is 3 percent because the proportions add up to 1 so, 60 plus 25 85, 85 plus 12 97 and 3 percent. What is the probability that cycle costing 4000 is bought. So, you could have 4000, 6000 and 8000 which is 25 plus 12 plus 3. So, 40 percent or 0.4 which is also equal to 1 minus less than 4000, so 1 minus 0.6.

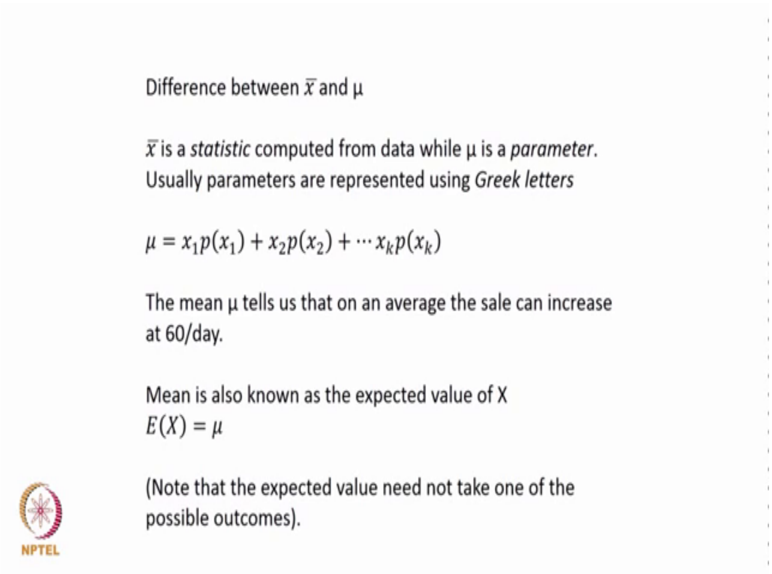
(Refer Slide Time: 13:27)



Now, what are some properties of random variables; A random variable conveys information that resembles what we may observe in a histogram. We take data for 50 days and let us say the sale was the same in 35 instances the sale increased by 2000 in 6 instances decreased by 1000 in 9 instances from the base value. So, X bar the expected value of the change in sale is minus 1000 into 9 times because it decreased by 1000 in 9 instances plus 0 into 35 times there was no change 35 out of 50 times plus 2000 increase.

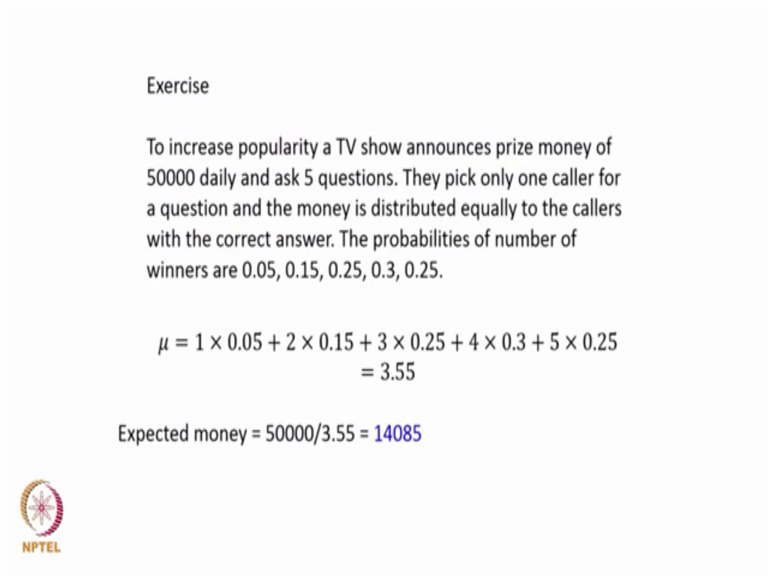
So, 2000 6 times divided by 50 which becomes 60. So, mu of the random variable is the weighted average of the possible outcomes and their probabilities. So, the expected value one could use X bar equal to 60, but then we generalize that mu equal to 60 which is the mean of the random variable.

(Refer Slide Time: 14:27)



Now, what is the difference between x bar and mu. x bar is a statistic computed from the data well mu is a parameter. Parameters are represented using Greek letters. So, mu is equal to x 1 into P of x 1 plus x 2 into P of x 2 plus etcetera. So, mu tells us that on an average the sale can increase at 60 per day. Mu is also known as expected value of x. So, E of X is mu. So, also note that the expected value need not take any of the values of the possible outcomes and it can be an entirely new number.

(Refer Slide Time: 15:02)



Another exercise to increase popularity a TV show announces a prize money of 50000 daily and asks 5 questions. They pick only one caller for a question and the money is distributed equally to the callers with the correct answer. The probabilities of number of winners are 0.05, 0.15, 0.25, 0.3 and 0.25 for the 5 questions respectively. Mu is equal to 1 into 0.05 plus 2 into 0.15 plus 3 into 0.25 plus 4 into 0.3 plus 5 into 0.25 which is 3.55. Expected money is 50000 by 3.55 which is 14000 and 85. They pick only one collar for a question and the money is distributed equally to the callers with the correct answer.

So, the probabilities of the number of winners are 0.05, 0.15, 0.25,z 0.3 and 0.25 now this means that there are 5 people who are called one for each question. Now this 0.05 means out of these 5 people only one answered it correctly. This 0.15 means out of the 5 people 2 answered it correctly and if 3 answered correctly the probability is 0.25 and so on. Therefore, the expected number of people who answered it correctly is 3.55 and therefore, expected money per person giving the correct answer is 14085.

(Refer Slide Time: 16:37)

Number of winners	Amount won	Probability (X = x)
1	50000	0.05
2	25000	0.15
3	16667	0.25
4	12500	0.3
5	10000	0.25

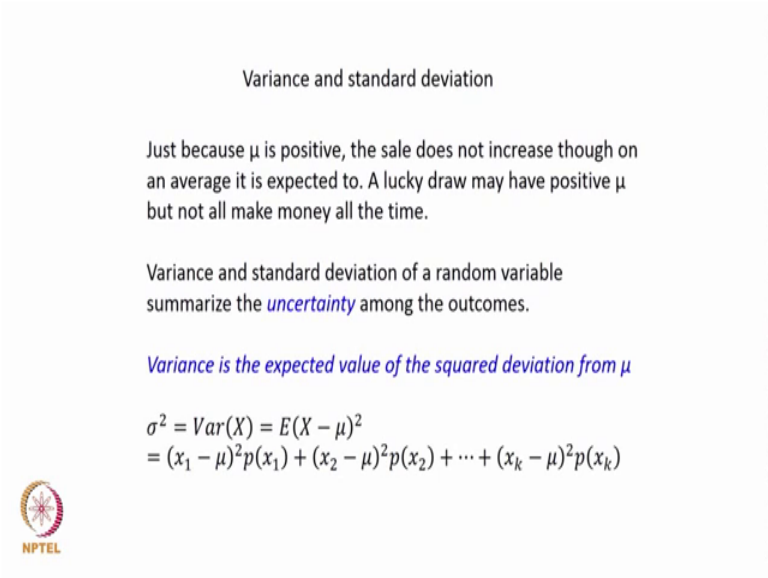
Expected value = 16666.75

The expected value of money won is higher than the Total money/expected number of winners

Now, if we say that on this day only one person gave the correct answer. So, that person gets 50000. One out of the 5 gave the correct answer. So, person gets 50000 and that happens with the probability of 0.05. 2 people gave the correct answer then the money per person is 25 percent that happens with 0.15 and so on. And if we now find the expected value it is 16600 and 66.75 versus 14000 and 85.

So, expected value of money one is higher than the total money by the expected number of winners. So, it depends on how you calculate and what is the expected value we are trying to calculate.

(Refer Slide Time: 17:21)



Now, let us also try to understand variance and standard deviation just because mu is positive the sale does not increase though on an average it is expected too. For example, a lucky draw may have a positive mu, but not all the people make money all the time. So, variance and standard deviation of a random variable summarizes the uncertainty among the outcomes.

So, variance is the expected value of the squared deviation from mu. So, sigma square is equal to variance of X is equal to expected value or X minus mu the whole square, which is x 1 minus mu the whole square into probability of x 1 plus x 2 minus mu the whole square into probability of x 2 and so on.

(Refer Slide Time: 18:04)

Sale example

Change in sale (x)	Deviation (x – μ) (μ = 60)	Squared deviation $(x - \mu)^2$	Probability (X = x)
+2000	1940	3763600	0.12
0	-60	3600	0.7
-1000	-1060	1123600	0.18

$\sigma^2 = 653880$

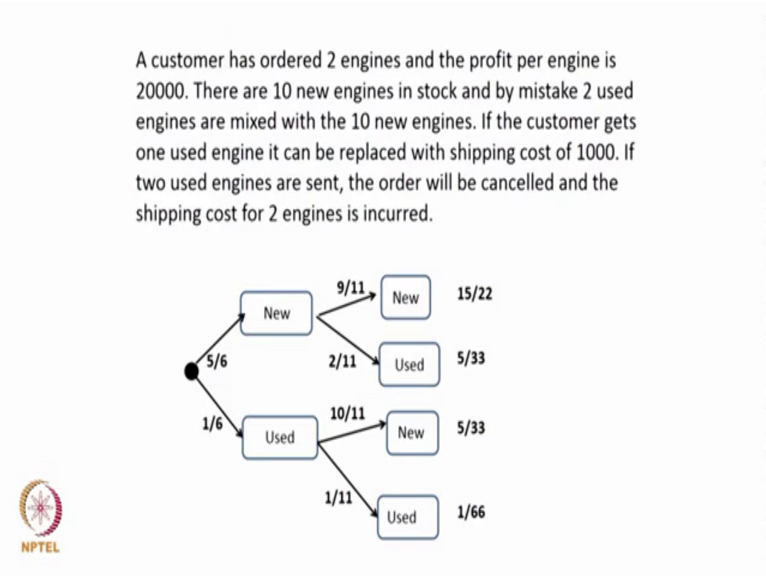
The *standard deviation* of a random variable is the square root of its variance $\sigma = \sqrt{653880} = 808.628$

To compute the variance of A random variable we revisit the example the sale example where change in sale X is the random variable and we have seen that this random variable takes 3 values 2000 plus 2000, 0 and minus thousand with probabilities 0.12, 0.7 and 0.18 we also note that these probabilities add to 1.

Now, the expected value of this random variable is 2000 into 0.12 plus 0 into 0.7 minus 1000 into 0.18. 2000 into 0.12 is 240 into 0.7 is 0, minus thousand into 0.18 is 180, minus 180. And therefore, the expected value is 240 minus 180 which is 60 which is shown here as mu equal to 60. Now to compute the variance and standard deviation we first find out the deviation which is X minus mu. So, 2000 minus 60 is 1940. 0 minus 60 is minus 60. Minus 1000 minus 60 is minus 1060. X minus mu square values are shown here for the 3. And the variance is sigma x minus mu the whole square into P of x which is 3763600 into 0.12 plus 3600 into 0.7 plus 1123600 into 0.18 which adds up to 653880.

In this variance computation, we do not have a divided by n which normally we saw when we did this in statistics. The reason being the sum of the probabilities add up to one and therefore, we do not have to divide it by n. So, sigma square is calculated 653880 which is the variance and standard deviation is the square root of it is variance we take the positive square root and we get 800 and 8.628. So, this is how we calculate the expected value or mean and the standard deviation of a random variable.

(Refer Slide Time: 20:54)



Another example customer had ordered 2 engines and the profit per engine is 20,000. There are 10 new engines and stock and by mistake 2 used engines are mixed with the 10 new engines.

If the customer gets a used engine, then it can be replaced with the shipping cost of thousands. If by for some reason both the engine sent happened to be used engines, then the order will be canceled and the shipping cost for 2 engines is incurred. Now how do we model this? Now we have new engine and used engine for the first and new engine for and used engine for the second. Now there are 2 there 10 new engines and stock and by mistake 2 used engines are mixed with the 10 new engines now there are 12 engines out of which 10 are new and 2 are used.

So, the first engine assume we is picked randomly a new engine would be picked with the probability 5 by 6. And a used engine would be picked with the probability 1 by 6. Now the second one if the first one was a new engine then out of the 10 new engine 1 has already been picked and 11 are remaining. So, another new engine would be 9 by 11 and a used engine would be 2 by 11. And in the first instance if we had by mistake picked a used engine then the second one picking a new engine is 10 by 11 and picking another used engine is 1 by 11 and therefore, the probabilities are 15 by 22, 5 by 33, 5 by 33 and 1 by 66 and let us just check if they add up to 1.

So, 15 by 22 is 45 by 66, 5 by 33 is 10 by 66 so, 45 plus 10 55 plus another 10, 65 plus 1 so, 66 by 66.

(Refer Slide Time: 22:47)

Outcome	Gain (x)	Deviation $(x - \mu)$ $(\mu = 33007)$	Squared deviation $(x - \mu)^2$	Probability $(X = x)$
Both new	40000	6993	48902049	0.682
One used	19000	-14007	196196049	0.303
Both used	-2000	-35007	1225490049	0.015

E(X) = 40000 x 0.682 + 19000 x 0.303 – 2000 x 0.015 = 33007

$$Var(X) = 111180951;$$

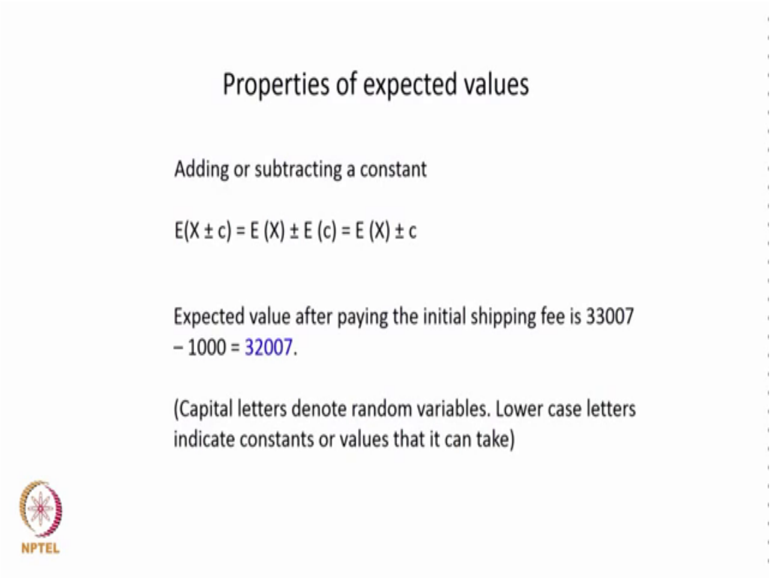
$$SD(X) = 10544.24$$

Though the expected value is high, the high SD indicates that there can be loss

Now, what happens if some reason both are both happen to be used engines, then the gain is 40000 because one used engine it can be replaced 2 used engines the order will be canceled. And the shipping cost for 2 engines is incurred. So, if both are new engines there is no issue. So, 20000 plus 20000, 40000 is the gain and that happens with the probability of 15 by 22. So, 15 by 22 is 0.682 if 1 engine is a used engine and one is a new engine the gain is 20000 the other one has to be taken back and a shipping cost of thousand is incurred.

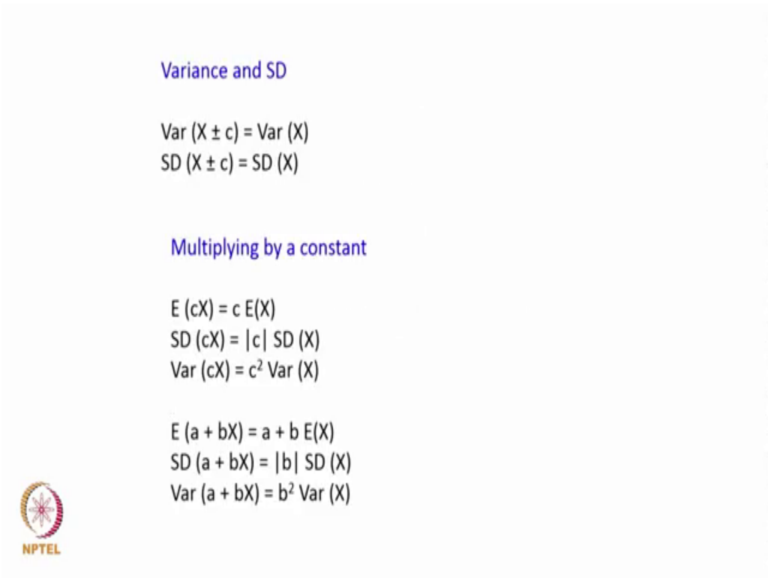
So, the actual gain is 19000 and that happens with the probability of one new one used is 5 by 33, one used one new is 5 by 33. So, it is 10 by 33, which is about 0.303 and if both happen to be used then there is no both have to be has to be recalled so, 0 and then there is a 2000 shipping cost. So, it becomes minus 2000. And now the average is 40000 and 0.682 plus 19000 and into 0.303 minus 2000 into 0.015 which is 30007. Now X minus mu is calculated 40000 minus 33007, 19000 minus 2000 minus 33 squared multiplied variance is 111180951 and standard deviation is 10544.24. So, though they expected value is high 33007 the high standard deviation indicates that in this case there can even be a loss.

(Refer Slide Time: 24:39)



However, small the probability of that loss is. So, we finished or conclude this lecture at this point where we define a random variable and then we also said it can take values we defined a notation X capital X equal to small x. Where capital X is the random variable and small x is the value. And then we define the mean or the expected value. So, we said the gain or outcome multiplied by the probability. We also defined the variance and standard deviation. So, we find the expected value and then we do X minus mu and then we do X minus mu the whole square multiplied by the probability. And then we sum it up to get the variance and then the square root of the variance is the standard deviation. So, we looked at all this. So, we define the what is a random variable and then how to calculate it is mean and standard deviation. So, will also look at some properties of random variables and answer some simple questions like what happens if for example, we add or subtract a constant and so on.

(Refer Slide Time: 25:52)



So, we look at all of these what happens to the variance when we multiply by a constant. So, all these things we will look at in the next lecture.