INDIAN INSTITUTE OF TECHNOLOGY KANPUR

Probability Models

Prof. Abhinava Tripathi

Recap: Bayes Theorem

Recap: Bayes Theorem

- Bayes theorem can be used to calculate posterior probabilities and the formula is

- P(B|A)= (P(A given B) P(B)) / P(A given B) P(B) + P(A given B’) P(B’)

- In the earlier breast cancer example, there were only two possibilities, either the female has breast cancer (Event B, P(B)) or she doesn’t have (Event B’, P(B’))

- What if there are more than two possibilities, i.e., B1, B2, and B3

Recap: Bayes Theorem

- Consider Event Bi, where a patient is treated at hospital i (1, 2, 3). Thus, the probability that a patient was treated in hospital 1 is P(B1) which is 0.6. Similarly, the probability that a patient was treated in hospital 2 is P(B2) which is 0.3. And the probability that a patient was treated in hospital 3 is P(B3) which is 0.1. So these are our prior probabilities

- Here we have B1, B2 and B3 such that the probabilities of all these 3 events sum upto 1

- In this example, B is said to be polytomous in nature while in our earlier example, B is dichotomous in nature, i.e. we only had B and B'.

Recap: Bayes Theorem

- Now assume that hospitals have shared the probability of lawsuit being filed in the past. So we have the conditional probability of a malpractice suit being filed given the hospital is known to us

P(Malpractice suit given Hospital 1)	0.002
P(Malpractice suit given Hospital 2)	0.005
P(Malpractice suit given Hospital 3)	0.007

- Let's say a malpractice suit is filed today. Then what is the probability that the suit was filed against hospital 1?

- So how are we going to solve this problem now?

Recap: Bayes Theorem

- Also let's call the event of filing a malpractice lawsuit as event A

P(A given B1)	0.002
P(A given B2)	0.005
P(A given B3)	0.007

- Remember that we also have the prior probabilities that the patient was admitted to hospital 1, 2, 3. So we have: P(B1) = 0.6; P(B2) = 0.3; P(B3) = 0.1

- Given the information that a malpractice suit is being filed against one of the hospitals, what is the probability that hospital is hospital 1 or B1? In other words, we want to calculate P(B1 given A)

Recap: Bayes Theorem

- Recall our earlier equation of Bayes theorem
- P(B1|A)= (P(A given B1) P(B1)) / P(A)
- We already know P(A given B1) and P(B1). What we don't have with us is P(A)

	B	**B'**
A	$P(A \cap B)$	$P(A \cap B')$
A'	$P(A' \cap B)$	$P(A' \cap B')$

- We have B1, B2 and B3
- We have the joint probability of A and B1 as $P(A \cap B1)$, the joint probability of A and B2 as $P(A \cap B2)$ and so on

Recap: Bayes Theorem

- We can rewrite the joint probability as follows

	B1	B2	B3
A	P(A∩B1)	P(A∩B2)	P(A∩B3)
A’	P(A’∩B1)	P(A’∩B2)	P(A’∩B3)

- We have the joint probability of A and B1 as P(A∩B1), the joint probability of A and B2 as P(A∩B2) and so on
- Hence, the marginal probability of A comes out as P(A) = P(A∩B1) + P(A∩B2) + P(A∩B3)

Recap: Bayes Theorem

- We also know how to write joint probabilities in terms of their conditional probabilities
- $P(A \cap B1) = P(A \text{ given } B1)P(B1)$; $P(A \cap B2) = P(A \text{ given } B2)P(B2)$; $P(A \cap B3) = P(A \text{ given } B3)P(B3)$
- Substituting these values back into P(A), we get
- $P(A) = P(A \text{ given } B1)P(B1) + P(A \text{ given } B2)P(B2) + P(A \text{ given } B3)P(B3)$
- Now: $P(P(B1 | A) = (P(A \text{ given } B1)\ P(B1)) / P(A)$
- Also, $P(A) = P(A \text{ given } B1)P(B1) + P(A \text{ given } B2)P(B2) + P(A \text{ given } B3)P(B3)$

Recap: Bayes Theorem

Thus we obtained the expression for conditional probabilities as shown here

- P(B1|A)= (P(A given B1) P(B1)) / P(A given B1)P(B1) + P(A given B2)P(B2) + P(A given B3)P(B3)
- P(B2|A)= (P(A given B2) P(B2)) / P(A given B1)P(B1) + P(A given B2)P(B2) + P(A given B3)P(B3)
- P(B3|A)= (P(A given B3) P(B3)) / (P(A given B1)P(B1) + P(A given B2)P(B2) + P(A given B3)P(B3))

Recap: Bayes Theorem

Let us do the numbers now

- We have Probability of A given B1 as 0.002, Probability of A given B2 as 0.005 and Probability of A given B3 as 0.007. We also have probability of B1, B2 and B3
- P(B1|A) = 0.002*0.6/(0.002*0.6 + 0.005*0.3 + 0.007*0.1) = 0.0012/(0.0012+0.0015+0.0007) = 0.3529
- Thus, if a malpractice suit takes place, then there is a 35.29 percent chance that this suit took place against hospital 1
- Now we can calculate P(B2|A) and P(B3|A) as well

Recap: Bayes Theorem

In the industry, scenarios with multiple possible events are more common

- Bayes theorem can be extended to such complex scenarios
- Consider, n mutual exclusive events, B1, B2, B3, …., Bn
- For any event 'i', P(Bi given A) = P(A given Bi) P(Bi) / P(A)
- We can generalize the expression for P(A): (P(A given B1) P(B1) + P(A given B2) P(B2) + … + P(A given Bn) P(Bn)), thus we obtain
- P(Bi given A) = P(A given Bi) P(Bi) / (P(A given B1) P(B1) + P(A given B2) P(B2) + … + P(A given Bn) P(Bn))

Introduction to Random Variables and Probability Distributions

Introduction

- In the previous discussions, we developed an understanding of basic concepts of probability and other concepts such as Bayes' theorem and its applications
- In the next set of topics, we will discuss probability distributions and their properties
- First we will learn about the concept of Random Variables and Probability Distributions

Introduction

- The next concept that we will see is that of the Expected value and Variance

- We will also discuss a discrete distribution, that is, the binomial distribution

Random Variables: I

Random Variables

- “The house always wins”: One or two people may end up winning large sums of money, but the machines are designed such that the rest of the people collectively lose more
- Consider a bag filled with 3 balls, 2 red and 1 blue; Each participant had to take out a ball, note its color, and then put it back in
- A participant who got a red ball all 4 times would receive 150 rupees; the participants who got any other result would have to pay 10 rupees

Random Variables

- We'll approach the problem in three steps
  - First we will see what all kinds of combinations can come up
  - Then we will see the probabilities of these combinations
  - We will use these probabilities to estimate the profit loss of a player playing this game once
- The one outcome all the players would want is to have all four ball being red
- But there are other possible out comes, what are they?

Random Variables

Let us look at all the possible outcomes

- We could get 4 blue balls - there's only one outcome in which this happens.
- We could get 3 blue balls and 1 red ball - this could happen in 4 ways - RBBB, BRBB, BBRB and BBBR.
- We could also get 2 blue balls and 2 red balls - this could happen in 6 ways.
- Also, we could get 1 blue ball and 3 red balls in 4 ways
- And there's only 1 way in which we could get 4 red balls
- In total, there are 16 possible outcomes

Random Variables: II

Random Variables

- Let us go back to the original question
  - How likely is the house to win?
  - What is the profit or loss you should expect while running such a business?
- In order to answer these questions, we need to analyze some probabilities
- We can quantify these possible outcomes and assign them to a variable
- In the previous example, what are the possible outcomes with 3 red balls: X=3
- In statistical terms, this ‘X’ converts possible outcomes to a number and is a random variable

Random Variables

- Such random variables can be defined in multiple ways
- For example, the number of blue balls that we drew from the bag or the number of red balls minus the number of blue balls that we have drawn from the bag
- The right way to choose the variable depends upon the nature of information you are interested in
- In this case we are interested in selection of red balls (X)
- If we want to know whether we will win or lose, we need to find different values of X and their probabilities

Probability Distributions

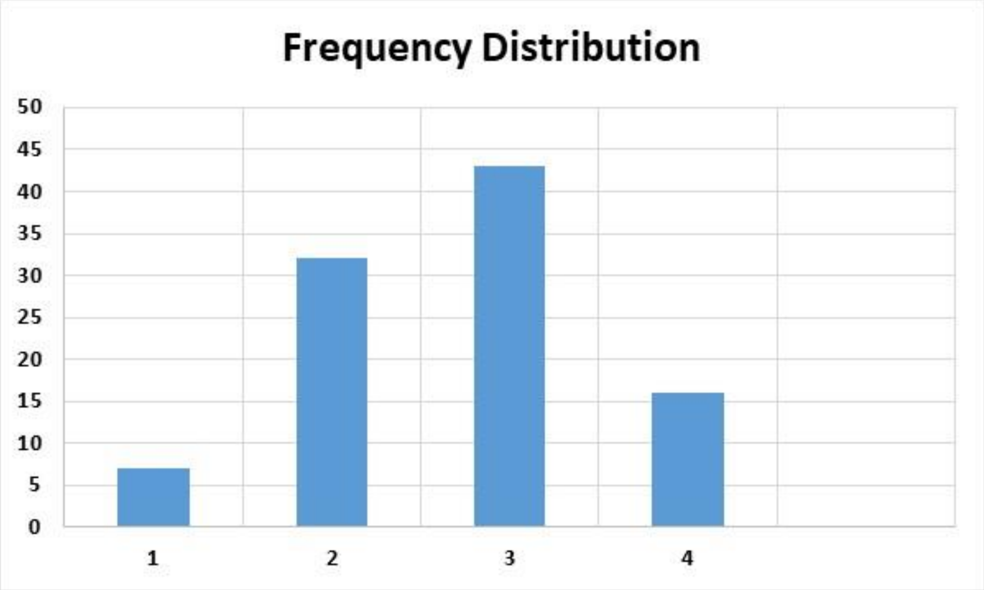
Probability Distributions

- We defined ‘X’ as the random event and number of red balls drawn
- So, for an outcome where we draw all the four blue balls only, X=0
- There are four possible outcomes if we are to draw one red ball then X=4
- In this manner, we have brought down our 16 outcomes into 5 groups where the random variable takes the value from 0 to 4
- To find out whether we loose money or win, we need to find the likelihood of each of these values

Probability Distributions

- The following results are obtained

X	Individual runs	Probabilities
0	2	0.02
1	7	0.07
2	32	0.32
3	43	0.43
4	16	0.16
Total	**100**	**1**

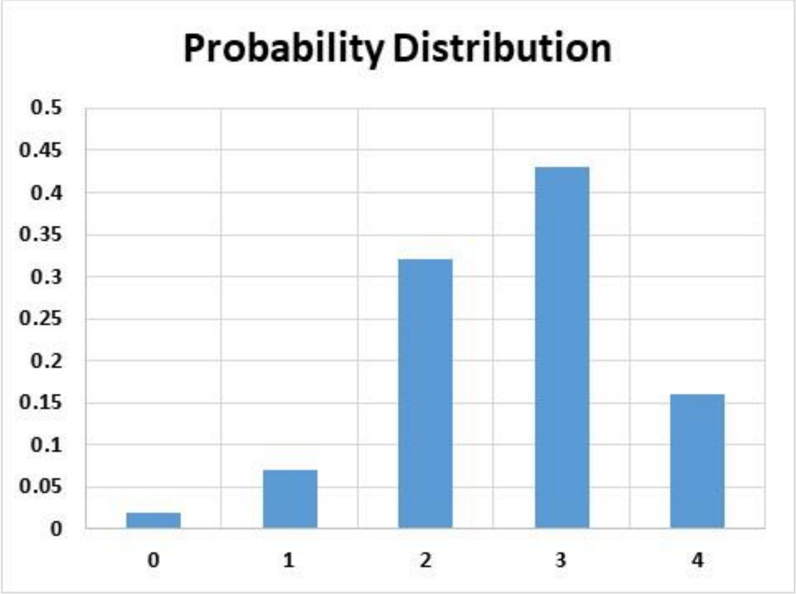


- The probability of any outcome is the number of favorable outcomes divided by the total number of outcomes
- For example, the probability of drawing two red balls is 32/100=0.32%

Probability Distributions

- These table and chart are also called probability distribution table and chart

X	Individual runs	Probabilities
0	2	0.02
1	7	0.07
2	32	0.32
3	43	0.43
4	16	0.16
Total	**100**	**1**

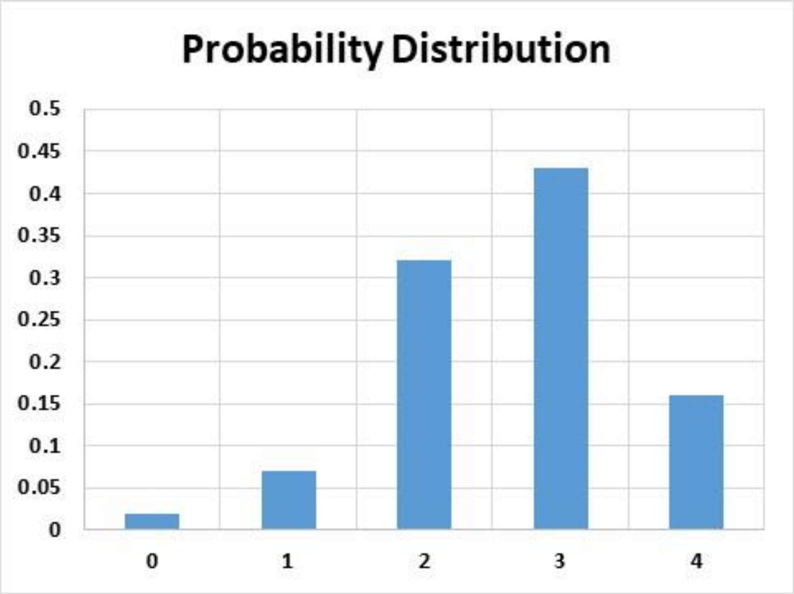


- On the X axis here, we have our 5 random variable values and on the y axis, we have our probability values

Probability Distributions

- Each value represents the probability of getting a certain number of red balls

X	Individual runs	Probabilities
0	2	0.02
1	7	0.07
2	32	0.32
3	43	0.43
4	16	0.16
Total	**100**	**1**



- Using the probability distribution, we can answer the question would we in the long run make money or lose money'

Expected Value

Expected Value

- Remember the red ball game; if somebody played it 1000 times, what will be their on average win or loss
- Consider the probability that X is equal to 1; we already know that it is 0.07
- In 1000 draws, it is expected that 1 red ball draw will occur 70 times
- Similarly, the probability that X=2, is 0.32*1000=320
- In this fashion, 160 people will draw 4 balls

Expected Value

- The total number of red balls we will get after 1000 attempts at the game will be $0*20 + 1*70 + 2*320 + 3*430 + 4*160=2640$
- Alternatively, we can say that we obtain 2.64 red balls per experiment
- The expected value of an event is defined as: X1*P(X=x1) + x2*P(X=x2) + x3*P(X=x3) + … so on, till … xn*P(X=xn)
- Expected value of the event X (X=0, 1, 2, 3, 4): EV = 0*P(X=0) + 1*P(X=1) + 2*P(X=2) + 3*P(X=3) + 4*P(X=4): 0*(0.02) + 1*(0.07) + 2*(0.32) + 3*(0.43) + 4*(0.16) = 2.64

Expected Value

- The expected value for our random variable X, that is, the number of red balls, is 2.64
- Interestingly, this number would be never obtained in any of the draw
- This expected value means that if you play this game infinite number of times, the average number of red balls per draw you expect is 2.64
- In our case, a more efficient random variable would be the expected amount won or lost in a game

Expected Value

- So we can define: X = money won after playing the game once
- For this variable, we have two values: 150 for winning (getting 4-red balls) and -10 for losing (any other outcome)
- For example, we can compute the probability of getting 0, 1, 2, 3 red balls as follows: 0.02+0.07+0.32+0.43 = 0.84
- Also, the probability of getting 4 red balls is 0.16
- Now we know that for winning probability is 0.16 and for losing the probability is 0.84

Expected Value

- We can compute the expected value as: (150*0.16) + (-10*0.84), which is equal to +15.6 dollars
- That is, a player would expect to win $15.6 dollars by playing, and therefore, the game is not profitable for the gambling house
- If the house makes money, it needs to ensure that the expected value won by the player is negative
- What to do: decrease the prize money, increase the penalty, decrease the players chances of winning, etc.

Expected Value: Example

- An insurance company estimates the probability that an accident will occur within the next year is 0.00071. Basis this information, what premium should the insurance company charge to break even on a $400,000 1-year term policy?

Event	x	P(x)	xP(x)
Live	0	0.99929	0.00
Die	400,000	0.00071	284.00
Total		1.00000	284.00

Expected Value: Example

- In case of no accident, X times P(X) becomes 0 times 0.99929 which is 0
- And in case of an accident, X times P(X) becomes 0.00071 times 400,000 which is approximately 284 dollars
- On average, the company needs to pay about 284 dollars to settle a policy

Event	x	P(x)	x*P(x)
Live	0	0.99929	0.00
Die	400,000	0.00071	284.00
Total		1.00000	284.00

Variance

Variance

- Similar to the mean or expected value of a measure, we also examine the spread or variability of the data
- This can also be computed using the probability distribution of the random variable
- Var(x) = (x - μ)^2 times P(x) summed over all the values of X
- Or alternatively $\sum_{i=1}^{n}(x_i - \mu)^2 * P(X = x_i)$, here μ (or sometimes $\bar{x}$) is the expected value or mean of the random variable (X)

Variance

- Recall our probability table

X	Probabilities	(x - μ)^2 * P(x)
0	0.02	(0 - 2.64)^2 * 0.02
1	0.07	(1 - 2.64)^2 * 0.07
2	0.32	(2 - 2.64)^2 * 0.32
3	0.43	(3- 2.64)^2 * 0.43
4	0.16	(4 - 2.64)^2 * 0.16
Total	**1**	0.8104

Binomial Distribution: I

Binomial Distribution

- Let us examine a very commonly occurring probability distribution: Binomial distribution
- Consider a simple set-up with 2 red and 1 blue balls, what is the probability of drawing 1, 2, and 3 red balls
- The probability of getting a red ball in one trial is 2/3 and the probability of getting a blue ball is 1/3
- If you are drawing 4 balls, what is the probability of drawing 4 balls from the bag

Binomial Distribution

- As per the multiplication rule, the probability of events 1 (E1) and 2 (E2) happening is P(E1)*P(E2)
- Here, we have four events, getting 4 red balls, the probability of the event is: P(Event 1)*P(Event 2)*P(Event 3)*P(Event 4)
- the probability of getting 4 red balls after 4 trials = Probability of getting a red ball in the first trial * Probability of getting a red ball in the 2nd trial * Probability of getting a red ball in the 3rd trial * Probability of getting a red ball in the 4th trial = = $\frac{2}{3}*\frac{2}{3}*\frac{2}{3}*\frac{2}{3}$ = 0.197

Binomial Distribution

- Remember that we are replacing the red ball again after drawing it
- Consider a case of drawing a blue (B) ball and 3 red (R) balls ‘BRRR’
- 1/3 is the probability of getting a blue ball in 1 trial and 2/3 is the probability of getting a red ball in 1 trial
- With multiplication rule: P(drawing a blue ball) which would be 1/3 times P(drawing a red ball) which would be 2/3 times P(drawing a red ball) which would again be 2/3 times P(drawing a red ball) : $(1/3)*(2/3)*(2/3)*(2/3) = 0.0987$

Binomial Distribution

- There are various possible combinations in which we can get 3 red balls and 1 blue ball
- There are 4 such sequences in which we get 3 red balls and 1 blue ball- RBRR, RRBR and RRRB
- The probability of these sequences is 4*0.0987=0.3948
- This is rule of addition, that is, for independent events E1 and E2, the probability of E1 or E2 is P(E1)+P(E2)

Binomial Distribution: II

Binomial Distribution

- Consider an event X=1, one red ball and three blue balls
- The probability is 4*(2/3*1/3*1/3*1/3)=0.0988
- Similarly, we can compute the probabilities of getting 0, 2 or 4 red balls
- Let us generalize this case, that is, probability of getting a red ball is p
- Thus, probability of getting 4 red balls in 4 trials would be = p^4
- Probability of getting 3 red balls= 4*p^3*(1-p)
- The probability that 0 red balls are drawn in 4 trials is (1-p)^4

Binomial Distribution

- Probability of taking 1 red and 3 blue balls is 4*p*(1-p)^3
- For X=2, it is 6 p^2 (1-p)^2; and X=3 it is 4p^3*(1-p); and X=4, it is p^4
- Now let us extend this to a more generic case of making n draws with a sucess probability of p and r favorable outcomes
- For P(X=r), e.g., r red balls and n-r blue balls, the probability of getting one such combination is p^r*(1-p)^(n-r)
- Also, there nCr combinations of getting r red balls out of total n balls
- The resulting probability is nCr *p^r*(1-p)^(n-r)

Binomial Distribution

- Using this formula for different values of r=0,1,2,3…n, we can find the probability distribution of the random variable X
- E.g., P(X=1) would be equal to $^{n}C_{1}(p)^{1}(1-p)^{n-1}$ and P(X=1) would be equal to $^{n}C_{1}(p)^{1}(1-p)^{n-1}$
- This probability distribution is called the binomial probability distribution
- Under what conditions we can use the binomial distribution?

Binomial Distribution

- Under what conditions we can use the binomial distribution?
- First, the total number of trials must be fixed at n
- The second condition is that each trial is binary in nature
- The third and final condition is that the probability of success is the same in all trials, denoted by p
- When all of these conditions are satisfied, then the random variable will follow a binomial distribution and the probability for $X = r$, that is, getting r successes in n trials, can be calculated as ${}^nC_r(p)^r(1-p)^{n-r}$, as given by the binomial distribution

Binomial Probability Distribution – Expected Value and Standard Deviation

Expected Value and Standard Deviation

- In our earlier experiment, where in we got 2.64 red balls per game as the expected value
- The expected value is nothing but the average value that we would 'expect' to get for a random variable
- E(X)= x1*P(X=x1)+x2*P(X=x2)+x3*P(X=x3)+….+xn*P(X=xn);
- And, Var(x) = $\sum_{i=1}^{n}(x_i - \mu)^2 * P(X = x_i)$
- For binomial distributions: E(x)= n*p and Var(x)= np(1-p)

Expected Value and Standard Deviation

- Let us revisit our example to calculate the number red balls to calculate the expected value and variance
- The participant can draw 4 balls, so how many red balls on average, the participant can draw
- Here n=4 and p=2/3, so E(X)= n*p=4*2/3 =8/3=2.67, i.e., the expected value of getting a red ball in a game
- Similarly, Var(x)=n*p*91-)= 4* 2/3*1/3= 8/9=0.88

Binomial Probability Distribution – Cumulative Probability

Cumulative Probability

- Using the binomial distribution, we were able to calculate the probability of getting an exact value
- For example, the probability of extracting 4 red balls: $^4C_4(2/3)^4(1/3)^0 = 0.19753$
- What if we wanted to calculate the probability of getting less than equal to 3 red balls, i.e., $P(X=0)+P(X=1)+P(X=2)+P(X=3)= 0.01235 + 0.09877 + 0.2963 + 0.39506 = 0.80247$
- That is an 80.2% chance that any randomly selected participant will have selected maximum 3 red balls while drawing 4 balls

Cumulative Probability

- Any probability where we have to determine the likelihood of X being less than a certain number is called a cumulative probability
- For example, 0.802 is the cumulative probability for X <= 3. And instead of saying P(X<=3) is 0.802, we can use F(X = 3) is 0.802, where F represents the cumulative probability
- For X=0, F(X=0) is also same as P(X<=0), since X<=0 takes only one value, X=0, then this is also same as P(X=0)

X	P(X=x)	F(X=x)
0	$=4C_0 * \left(\frac{2}{3}\right)^0 * \left(\frac{1}{3}\right)^4$ =0.01235	0.01
1	$=4C_1 * \left(\frac{2}{3}\right)^1 * \left(\frac{1}{3}\right)^3$ =0.09877	0.11
2	$=4C_2 * \left(\frac{2}{3}\right)^2 * \left(\frac{1}{3}\right)^2$ =0.2963	0.41
3	$=4C_3 * \left(\frac{2}{3}\right)^3 * \left(\frac{1}{3}\right)^1$ =0.3951	0.80
4	$=4C_4 * \left(\frac{2}{3}\right)^4 * \left(\frac{1}{3}\right)^0$ =0.1975	1.00

Cumulative Probability

- Next, let's calculate F(X = 1). Here, F(X = 1) means P(X<=1), which will be P(X = 0) + P(X = 1), which will come to 0.11111
- Similarly, we have F(X=1), which will be P(X=0) + P(X=1) and F(X=2), which will be P(X=0) + P(X=1) + P(X=2)

- Thus, we can write F(X=2) as F(X=1) + P(X=2)
- Hence, F(X=2) will be 0.111 + 0.296= 0.41
- And F(X=4) will be F(X=3) + P(X = 4) which is 0.80247 + 0.19753 =1.00

X	P(X=x)	F(X=x)
0	$=4C_0 * \left(\frac{2}{3}\right)^0 * \left(\frac{1}{3}\right)^4$ =0.01235	0.01
1	$=4C_1 * \left(\frac{2}{3}\right)^1 * \left(\frac{1}{3}\right)^3$ =0.09877	0.11
2	$=4C_2 * \left(\frac{2}{3}\right)^2 * \left(\frac{1}{3}\right)^2$ =0.2963	0.41
3	$=4C_3 * \left(\frac{2}{3}\right)^3 * \left(\frac{1}{3}\right)^1$ =0.3951	0.80
4	$=4C_4 * \left(\frac{2}{3}\right)^4 * \left(\frac{1}{3}\right)^0$ =0.1975	1.00

Continuous Probability Distributions: Continuous random variables

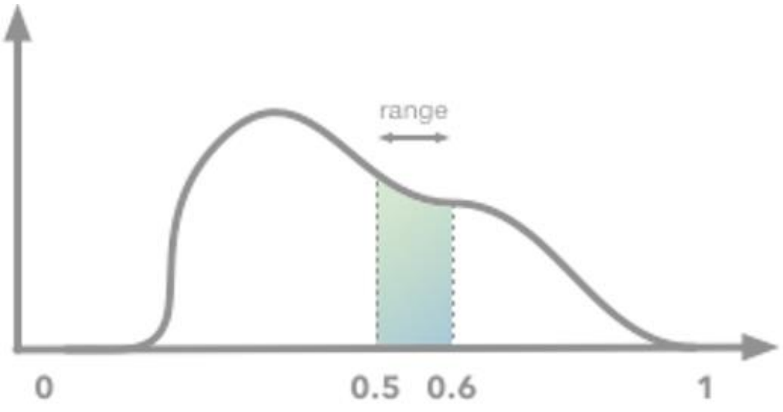
Continuous random variables

- Now we will discuss continuous probability distributions
- The random variable used to define the continuous probability distribution is called the continuous random variable
- If you are a pizza delivery manager, you are concerned with the average time it takes for a pizza delivery to reach a customer. Let us call this delivery time as random variable 'X'
- This variable can take on various values like 5-min, 30-min, 15.13-min
- Here defining the variable exact time, is defined as continuous random variable, i.e., it can be defined even up to last milliseconds

Continuous random variables

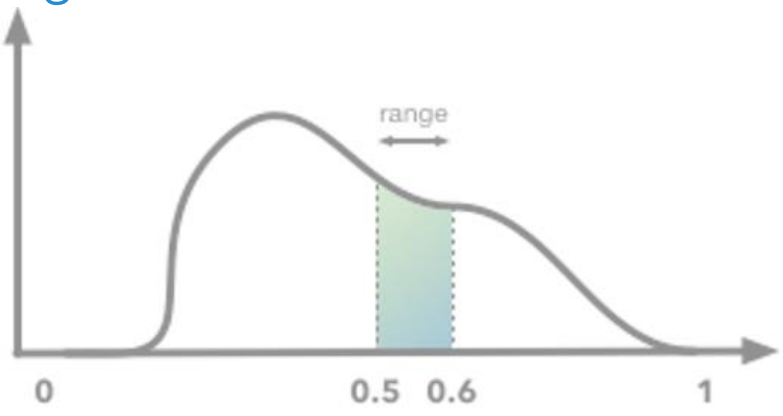
- The ‘amount of water present in a bottle’ or ‘the exact stock price at the end of a trading session’, or anything that is generally exact, this is always going to be a continuous random variable
- The values that random variable may take are infinite. For example, even between 20 to 20.1 there can be infinite values taken by a continuous random variable

- Let us draw a plot where x-axis is possible outcomes of random variable
- In case of continuous random variables, it is difficult to define/obtain the probability of a specific value of X



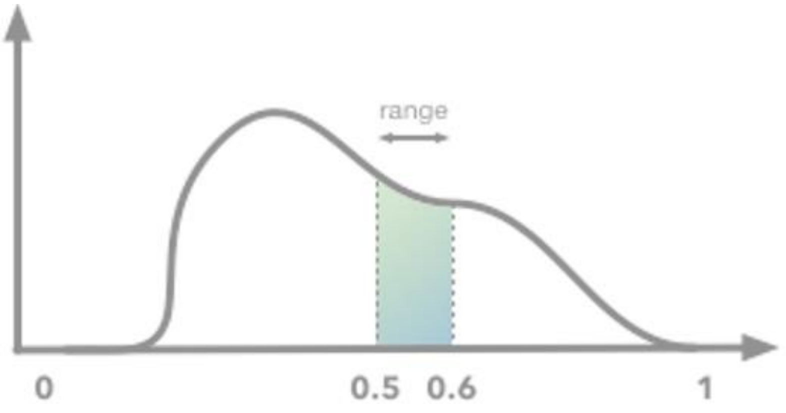
Continuous random variables

- Thus, a continuous random variable is represented by a continuous line known as probability density function
- In the case of a discrete random variable, we used to calculate the probability by saying P(X=0) or P(X=1), and plot these values
- In case of pizza delivery time, there are infinite possible outcomes for a continuous random variable; so the question that probability of observing 15-min is close to zero, and same is true for any other value
- So instead of X being any specific value, we measure the specific probability of X lying in a certain interval



Continuous random variables

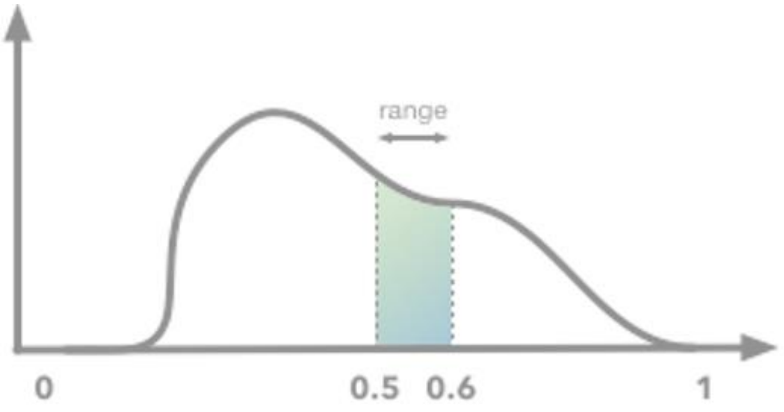
- For example, we will define the cumulative probability as X is less than equal to 30-min and greater than equal to 20-min
- This is the area under the curve between 20-min to 30-min (colored area in the plot)



- Also, if we assume that all the possible delivery times range in between 0 to 1 hour, the X will range from 0 to 1-hour, and the area of the curve in this range will be 1

Continuous random variables

- First, a continuous random variable is such that it can have an infinite number of outcomes
- Second, we represent a continuous random variable using what we call a probability density function which is essentially a continuous line drawn for all the range of values that X can take

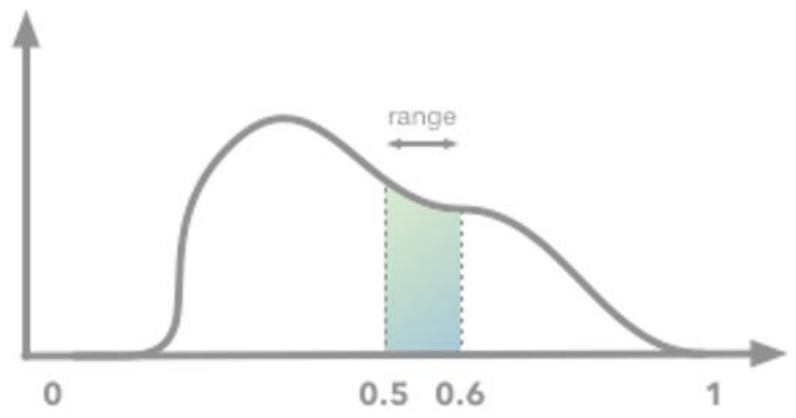


- Thirdly, the area under the curve, represents the probability that the random variable lies in that interval
- Finally, the total area under the curve will always be equal to 1

Continuous Probability Distributions: Cumulative probability for continuous Random Variables (RV)

Cumulative probability for continuous RV

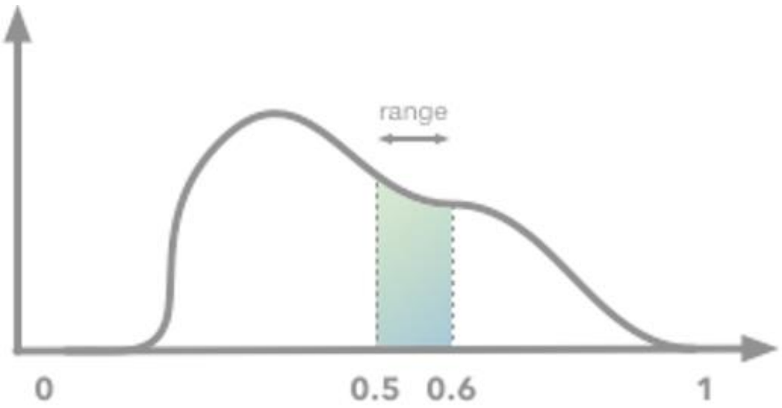
- Consider the following example. The probability density function of our continuous random variable is plotted on x-y axis. The x-axis goes from - ∞ to + ∞.



- What is the probability of observing X from -1 to +1, that is area under the cure from X=-1 to X=+1 (colored region)
- When we talk about cumulative probabilities, we are always dealing with ranges

Cumulative probability for continuous RV

- If you are given that the cumulative probability for X=1 and X=-1, as 0.6 and 0.4 respectively
- The probability of X between -1 to 1 is nothing but
- P(X < = 1) - P(X<= -1) , that is the area under the curve from -1 to 1
- In this case, P(X < = 1) is 0.6 and P(X<= -1) is 0.4, so the probability of X lying between 1 and -1 is (0.6-0.4) which is 0.2



Continuous Probability Distributions: Normal Distribution

Normal Distribution

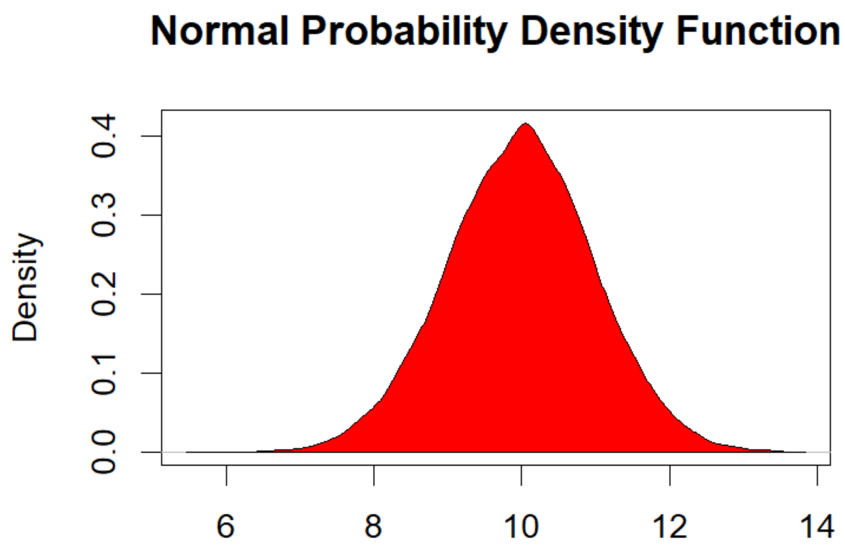
- The normal distribution is perhaps the most widely used and most important distribution when talking about distributions of continuous random variables
- There are many places where the normal distribution appears naturally
- Normal distribution has very convenient and interesting properties that makes it easy to work with them
- Also, normal distribution is an integral part of central limit theorem

Normal Distribution

- Let us go back to our earlier example of the commute time for delivering pizza to the customer’s houses
- Remember the 30-minute discount scheme to deliver the pizza. If the time is more than 30-minutes, then pizza is free or available with large discounts
- As a manager at one of the pizza outlets and given this 30-minute guarantee, you want to ensure that most of the pizzas are delivered well before these 30 minutes
- Let us consider only those commutes from the pizza outlet to the customer’s location and not include the deliveries where the delivery boy has to visit multiple locations as he is delivering multiple orders

Normal Distribution

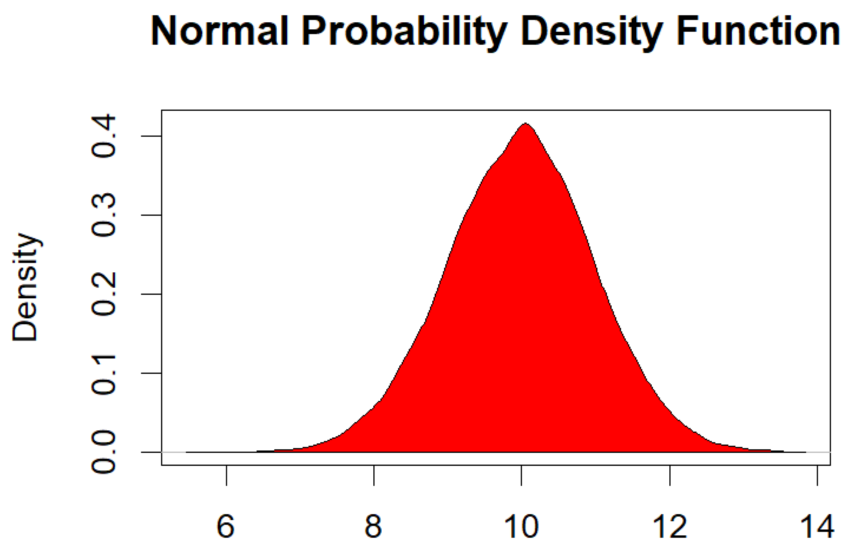
- Let us assume that it takes a maximum of 10 minutes to make the pizza. That leaves us with only 20 minutes to deliver the pizza from the outlet to the customer's location



- The commute time may vary from location to location
- The probability density function for such a scenario may appear like a normal distribution : Bell shaped curve

Normal Distribution

- This chart is known as a normal distribution
- You can see that the probability density is highest at 10 minutes
- We can also see that the probability density starts decreasing as we move towards the right or left

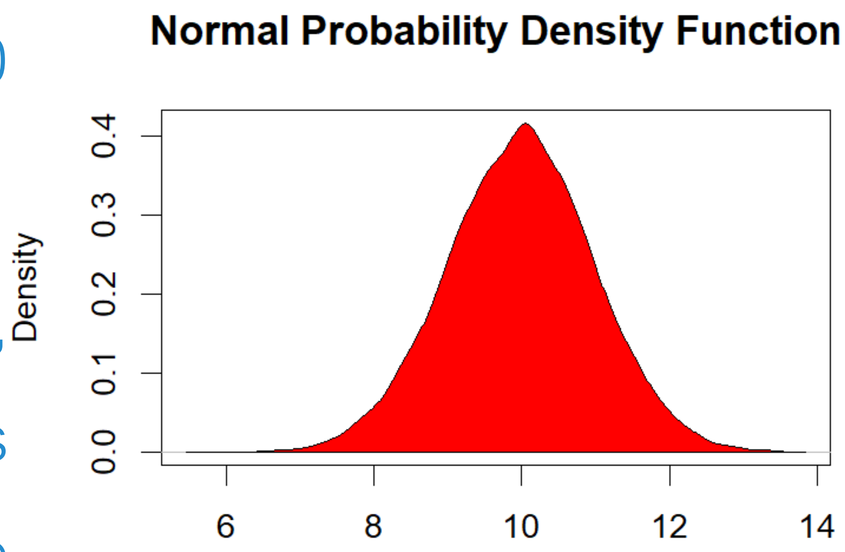


- Just by looking at this distribution, we can make several observations about the characteristics of this distribution

Normal Distribution

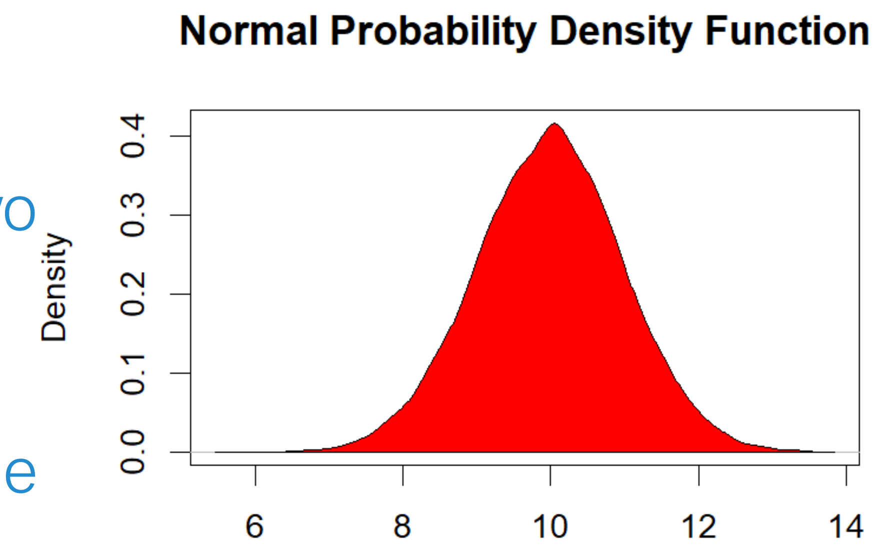
- The mean of this distribution, which in our case is 10 minutes, lies in the exact center of this distribution

- Since the distribution is symmetric around the mean, which is 10 minutes, it means that 50% of its values are less than the mean and 50% are greater than the mean
- we can see that the probability density is the highest at the mean, and decreases exponentially as we move further away from the mean



Normal Distribution

- Normal distributions are very common in nature
- Any normal distribution can be defined using only two parameters - the mean, μ and standard deviation, σ
- The mean (μ=10) is located at the center of the normal distribution, and that point also denotes the median and the mode of the distribution
- The standard deviation (σ) determines how flat and wide the normal distribution curve is
- An increased σ indicates a more dispersed and flat distribution



Probabilities for a Normal Distribution

Probabilities for a Normal Distribution

- Assume a normal distribution with a mean of ‘μ’ and standard deviation given by ‘σ’
- Using properties of normal distribution curve, we can calculate probability values
  - The probability of X lying between $\mu - \sigma$ and $\mu + \sigma$ is around 68% or 0.68
  - The probability of X lying between $\mu - 2\sigma$ and $\mu + 2\sigma$ is around 95% or 0.95
  - And the probability of X lying between $\mu - 3\sigma$ and mu + 3 sigma is around 99.7% or 0.997

Probabilities for a Normal Distribution

- Let us revisit the Pizza delivery time example, assume a normal distribution with μ =12-min average deliver time and σ =3-min standard deviation
- What is the probability of delivery time 6-18 mins
- Using our empirical rule 95% of the population lies between $\mu - 2\sigma$ and $\mu + 2\sigma$, i.e., 6-18 minutes; in other words 95% probability of observing a delivery time between 6-18 mins
- Similarly, 99.7% probability of observing a delivery time between 3-21 mins ($\mu - 3\sigma$ and $\mu + 3\sigma$)

Probabilities for a Normal Distribution

- What is the probability of observing a delivery time between 6 and 21 minutes
- Now we are looking at the interval which is $\mu - 2\sigma$ to $\mu + 3\sigma$
  - Due to symmetry of the normal distribution, we can say that 95%/2=47.5% of the population lies from $\mu - 2\sigma$ to μ
  - And 99.7%/2 of the population lies from μ to $\mu + 3\sigma$
  - Adding the two, we get 97.35% of the population lie between 6-21 mins , i.e., $\mu - 2\sigma$ to $\mu + 3\sigma$

Probabilities for a Normal Distribution

- What is the probability of observing a delivery in less than 15-mins, i.e., $P(X<=15)$ or
- Probability of X less than $(\mu + 1\sigma)$ because 15 can be written as $(\mu + 1\sigma)$
- We can divide the desired area in two parts: (a) 50% of area up to $\mu=12$, and (b) 12-15 mins $(\mu$ to $\mu + 1\sigma)$
- As noted earlier 68% of the area lies from $(\mu - 1\sigma)$ to $(\mu + 1\sigma)$ or 34% area lies from μ to $\mu + 1\sigma$
- Therefore, 50%+34% =84% of pizza deliveries are taking place in less than 15-minutes

Standard Normal Distribution: Part I

Standard Normal Distribution

- What if you want to find, ‘what is the percentage of deliveries where the commute time is between 6 and 17 minutes or between 6 and 16.95 minutes
- These values are not as easily identifiable as some of the earlier examples
- Here we can see that μ is 12 and X is 17, and the difference between these two values should be 5
- if we have to represent X in the form of (μ + some multiple times σ), then we can find this multiplication factor by using (X-μ) divided by σ
- In our case, (X-μ) / σ comes out to be (17 - 12) / 3, which equates to 5 by 3, which is around 1.67

Standard Normal Distribution

- The value $(X-\mu) / \sigma$ is denoted by z and z is called the standard normal variable
- Now, the probability of finding X lying between $(\mu - 2\sigma)$ and $(\mu + 2\sigma)$ is same as the probability of finding the new standard normal random variable z between -2 to +2, which is 95%
- Similarly, the probability of X lying between $(\mu - 1\sigma)$ and $(\mu + 1\sigma)$ is the same as the probability of Z lying between -1 and +1, which is 68%
- For a random variable X, we can find the probability of X within a certain range, in the form of a standard normal variable $z = (X-\mu) / \sigma$
- Z follows standard normal distribution

Standard Normal Distribution

- What is the difference between normal distribution and standard normal distribution
- In a normal distribution notation of interval we have $(\mu + 1\sigma)$, $(\mu + 2\sigma)$ and $(\mu + 3\sigma)$ on the right and $(\mu - 1\sigma)$, $(\mu - 2\sigma)$ and $(\mu - 3\sigma)$ on the left
- In the standard normal distribution, mean is always 0, and standard deviation is 1
- $(\mu + 1\sigma)$ will mark to 1, $(\mu + 2\sigma)$ will mark to 2 and, similarly, on the left hand side as well, $(\mu - 1\sigma)$ will mark to -1 and so on

Example

- How do we proceed to solve the problem of calculating the probability that an employee will take less than 17 minutes to commute to the office?
- First, let us convert 17 into the Z value; Remember that the μ and σ for this distribution was 12 and 3 minutes
- Thus, calculating $(X-\mu) / \sigma$, we get $(17 - 12) / 3$, which equates to 5 by 3, which is 1.67
- $P(X <= 17)$ is the same as the probability, $P(Z <= 1.67)$; we are essentially calculating the cumulative probability for $Z = 1.67$
- This value can be found from excel [=1-NORM.DIST(-1.67, 0, 1, TRUE)] or Z-table

Thanks!