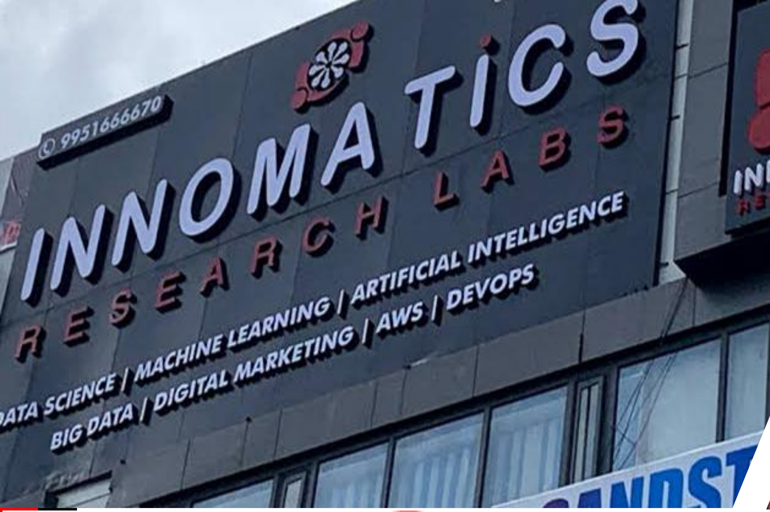


INNOMATICS RESEARCH LABS

Topics:

- Introduction to Probability
- Various Probability and Bayes Theorem
- Random Variable
- Probability Distribution – Bernoulli Binomial and Poisson
- Normal Distribution

Introduction to Probability

Intuitive understanding…..

- You and a friend are at a cricket match, and out of the blue he offers you a bet that neither player will hit a century in that game. Should you take the bet?
- Your company is launching personalized marketing campaign to millions of potential customers. To which customer should you offer what type of product.
- A widget maker in your factory that normally breaks 4 widgets for every 100 it produces has recently started breaking 5 widgets for every 100. When is it time to buy a new widget maker?
- You are conducting poll on national election for a big media house. How many people do you have to poll? How do you ensure that your poll is free of bias? How do you interpret your results?

What is Probability?

- Measure of likeliness of something happening
  - Strength of belief that something is true
  - Mathematical way of expressing uncertainty

- Given n observations of an event, it denotes the proportion of observations where a given event occurs
  - $P(E) = \frac{\#\ of\ outcomes\ in\ which\ the\ even\ occurs}{total\ possible\ \#\ of\ outcomes}$

  - Probability of a single event is always between 0 and 1

Examples:

- In a coin toss, probability of a head or tails appearing?
- In a roll of dice, probability of 3 appearing?
  - Six possible outcomes: {1,2,3,4,5,6}
  - Each outcome equally likely, therefore probability of an outcome:1/6
  - Probability of an odd number appearing?
- Probability of amount of rain in Mumbai in August?
- Probability of RCB winning IPL 2021?

Introduction to Probability Theory

- Analytics applications involve tasks such as prediction of probability of occurrence of an event, testing a hypothesis, building models to explain variation of importance to the business such as Profitability, market share, demand etc…
- Many important tasks in analytics deal with uncertain events & it is essential to understand probability theory that can be used to predict & measure them
- We don’t know the outcomes of a particular situation until it happens. Will it rain today? Will I pass the next math test? Will my favorite team win the toss? Will I get a promotion in next 6 months? All these questions are examples of uncertain situations we live in.

Probability Theory - Terminology

Experiment – It can be either deterministic or random

- Deterministic: Outcome always same and determined
- Random: Many possible outcomes from a range of value.

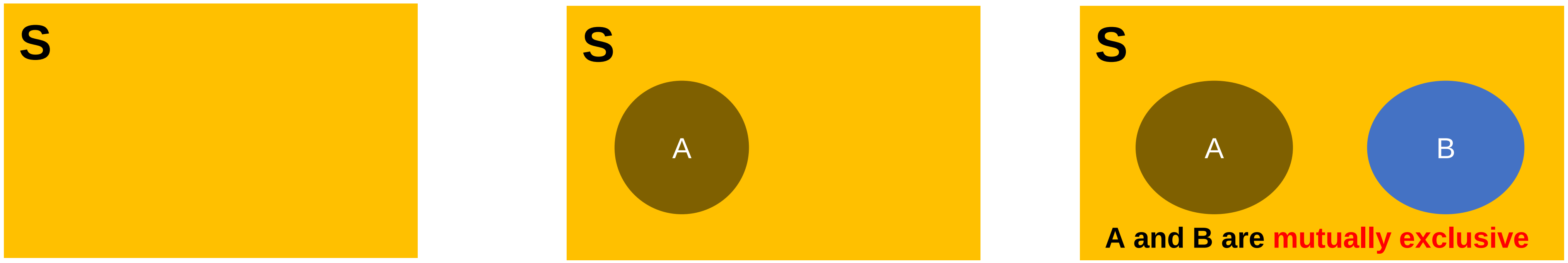
An experiment in which the outcome is not known with certainty .i.e. the output of this experiment cannot be predicted with certainty. Whether it rains on a daily basis is an experiment.

Outcome is the result of a single trial. So, if it rains today, the outcome of today's trial from the experiment is "It rained"

Event is one or more outcome from an experiment. "It rained" is one of the possible event for this experiment. Subset of sample space.

Probability is a measure of how likely an event is. So, if it is 60% chance that it will rain tomorrow, the probability of Outcome "it rained" for tomorrow is 0.6

- **Sample Space-** Given a random experiment K, set of all possible outcomes for K is denoted by S as sample space. For example: Rain experiment K, has a sample space S = {“It rained”, “It did not rain”}



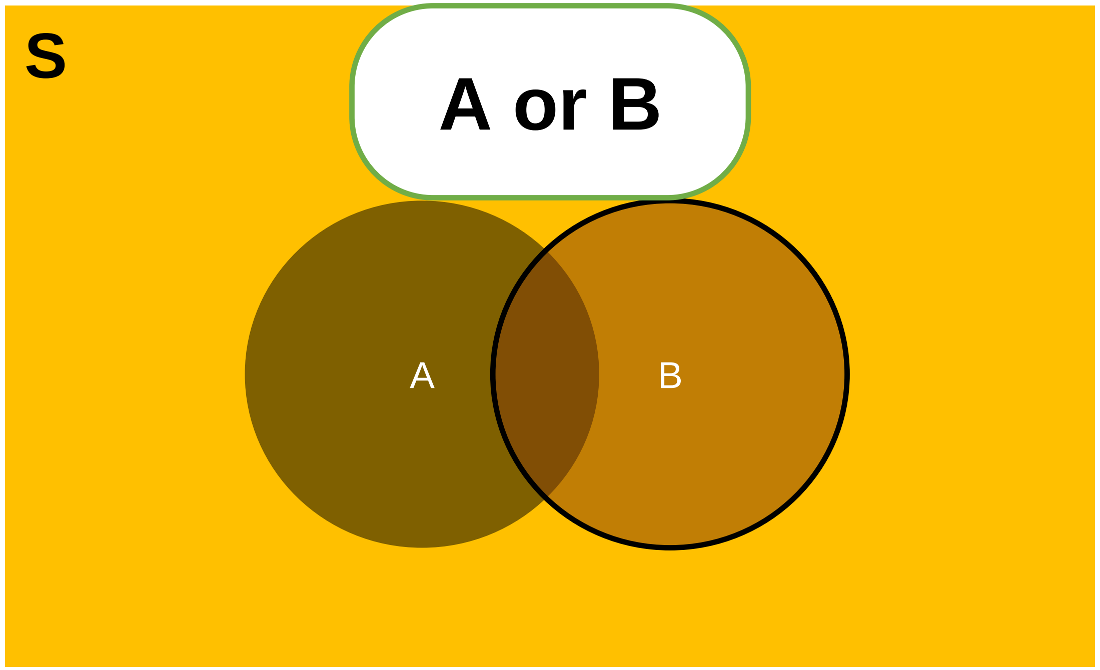
Area of the rectangle denotes sample space, and since probability is associated with area, it cannot be negative.

Mutually Exclusive – If event A happens, event B cannot.

Algebra of events:

Consider two events A and B

- **Intersection (A∩B):** Set of all elements common to A and B
- **Union (A U B):** Set of all elements belonging to either A or B
- **Difference (A-B):** Elements belonging to A but not to B
- **A and B are mutually exclusive** or disjoint if A∩B = Ø

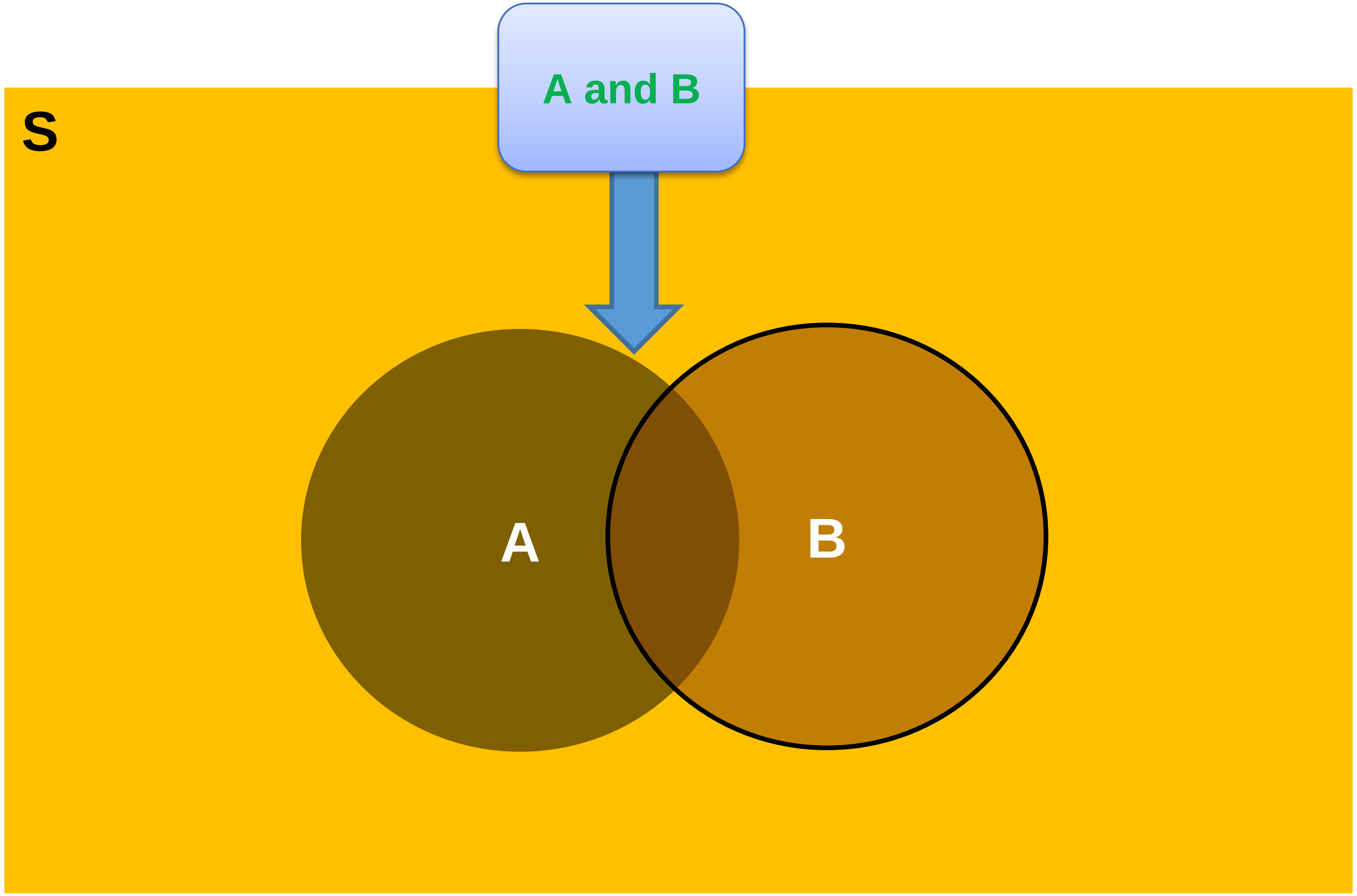


$$P(A\ or\ B) = P(A) + P(B)\ - P(A\ and\ B)$$

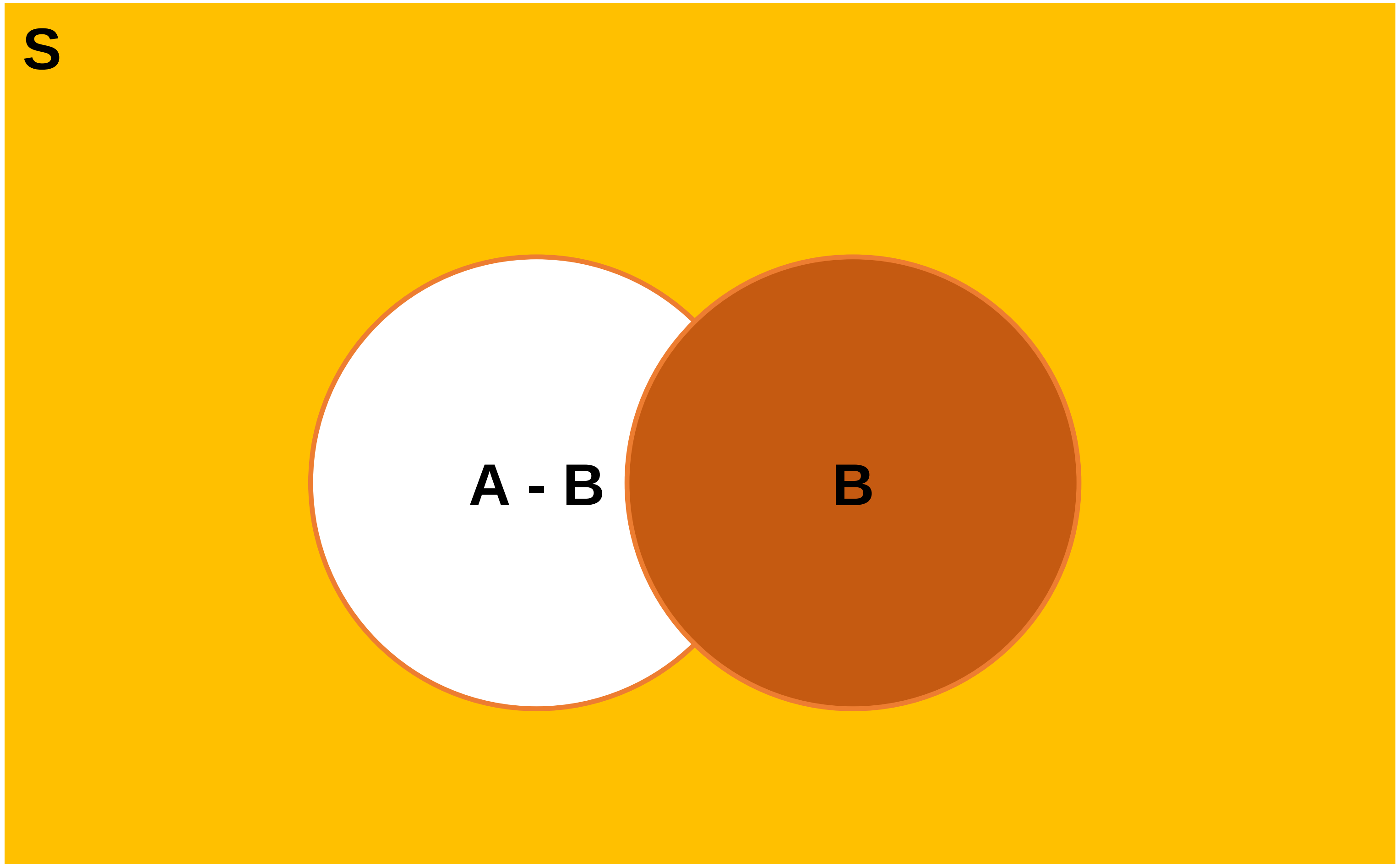
Example

- Event A – Customers who default on loans
- Event B – Customers who are High Net Worth Individuals

Intersection (A and B):



Difference (A – B) :



Axioms of Probability:

According to axiomatic theory of probability, the probability of an event E satisfies the following axioms:

For an experiment K, and event A:

- The probability of an event E always lies between 0 and 1
  - $0 <= P(A) <= 1$
- The probability of universal set = 1
  - $P(S) = 1$
- If A1, A2, A3…are mutually exclusive then P(A1 U A2 UA3…) = P(A1)+P(A2)+P(A3)…

Probability – Types:

Contingency table summarizing 2 variables, *Loan Default* and *Age:*

		Age			
		Young	Middle-aged	Old	Total
Loan Defaults	No	10,503	27,368	259	38,130
	Yes	3,586	4,851	120	8,557
	Total	14,089	32,219	379	46,687

Probability – Types:

Convert it into probabilities

		Age			
		Young	Middle-aged	Old	Total
Loan Defaults	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

Probability - Types

Marginal Probability:

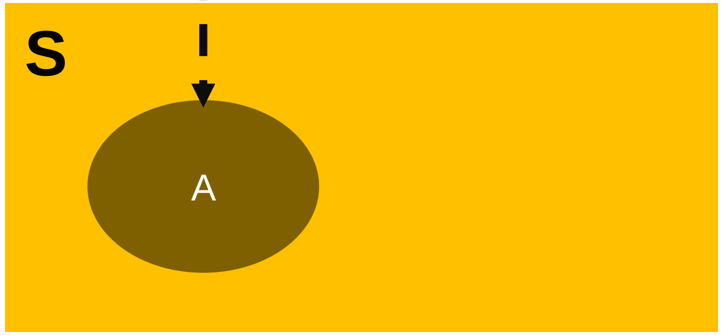
		Age			
		Young	Middle-aged	Old	Total
Loan Defaults	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

Marginal Probability

Probability describing a single attribute

P(Middle-aged) = 0.690

P(old) = 0.008



Probability - Types

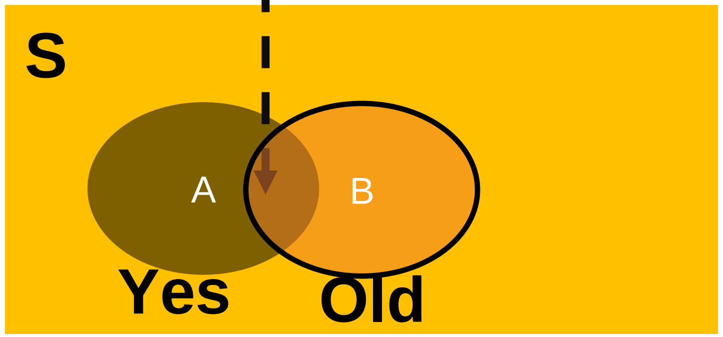
Joint Probability:

		Age			
		Young	Middle-aged	Old	Total
Loan Defaults	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

Joint Probability

Probability describing a combination of attribute

P(Yes and old) = 0.003



Probability - Types

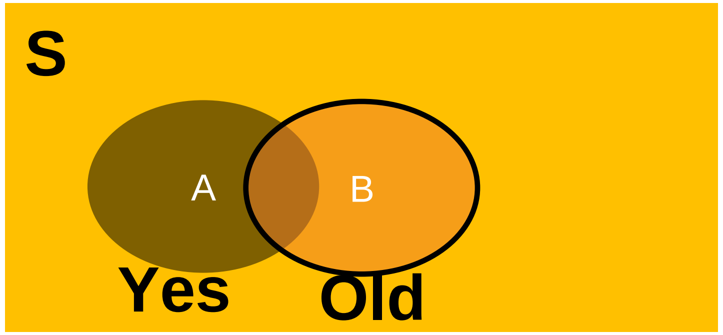
Union Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Defaults	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

P(Yes or old) = P(Yes) + P(old) – P(Yes and old)

= 0.184 + 0.008 – 0.003

= 0.189

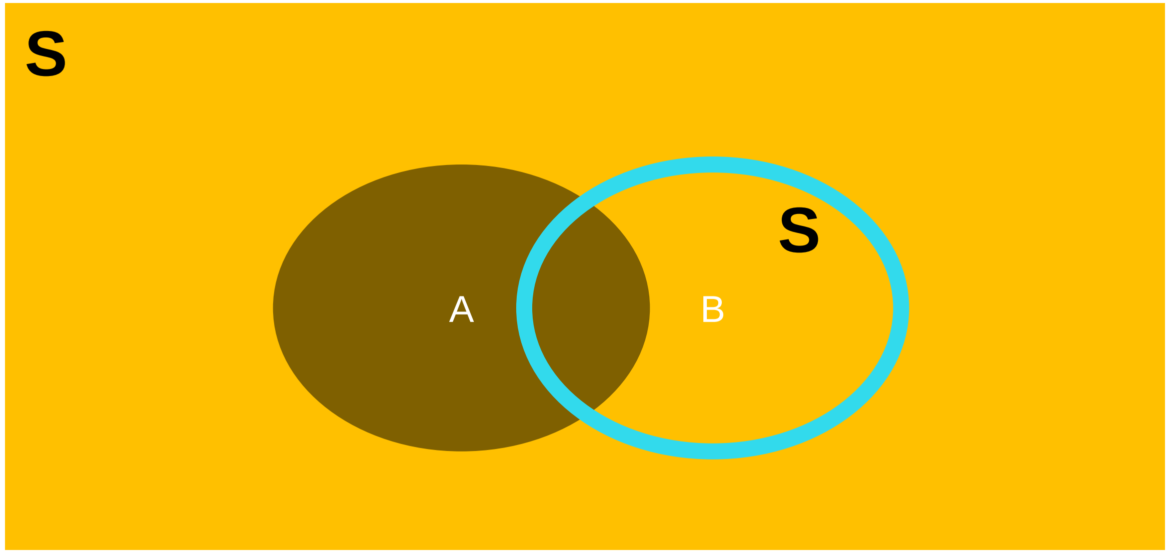


Probability - Types

Conditional Probability

- Probability of *A occurring* ***given that*** B has occurred.
- The sample space is restricted to a single row or column.
- This makes rest of the sample space irrelevant.

Probability, i.e., $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$

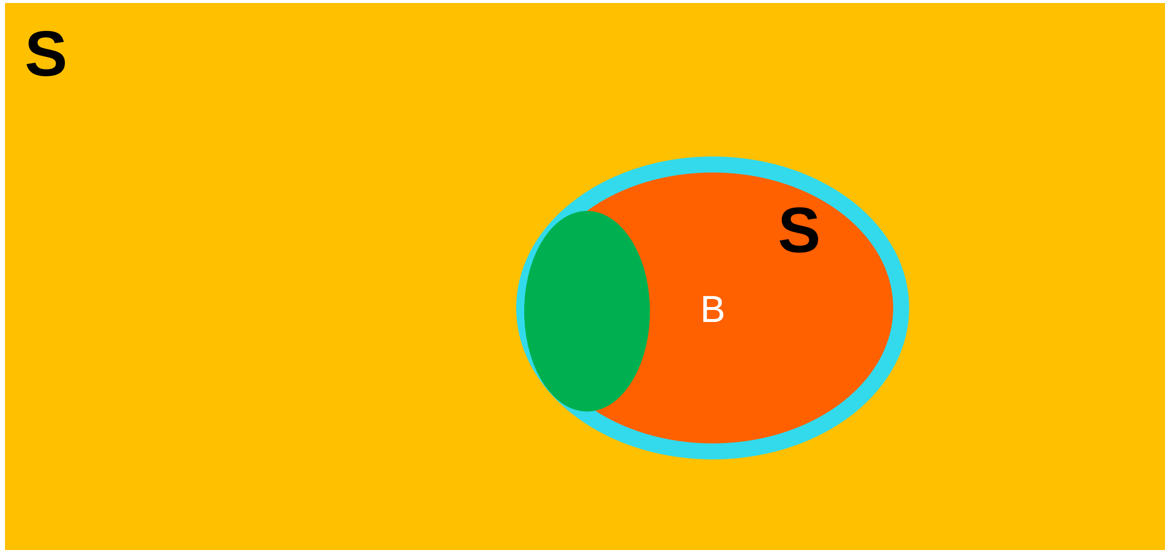


Probability - Types

Conditional Probability

- Probability of *A occurring **given that*** B has occurred.
- The sample space is restricted to a single row or column.
- This makes rest of the sample space irrelevant.

Probability, i.e., $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$

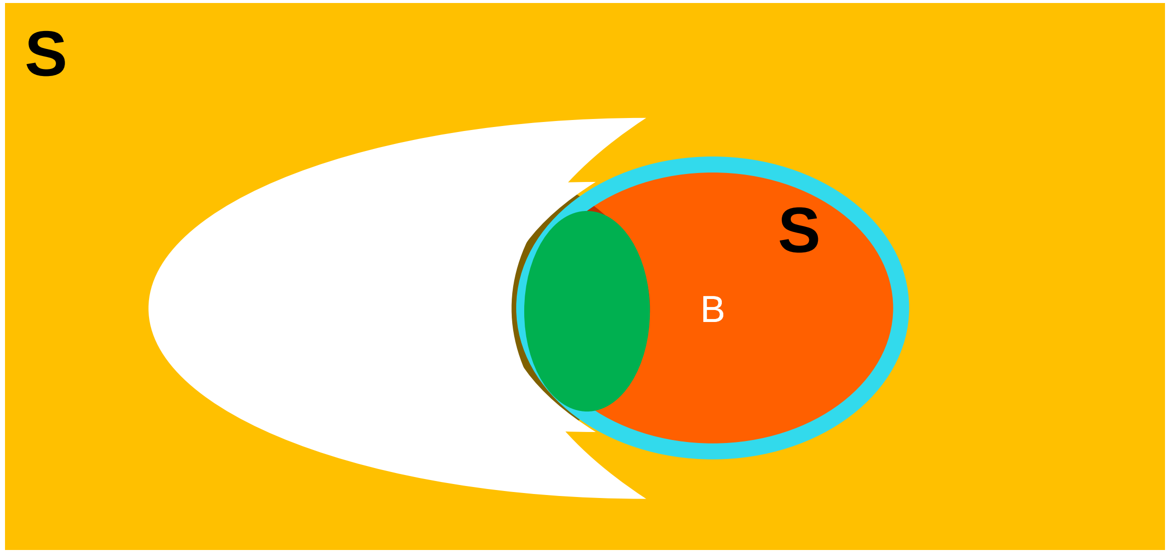


Probability - Types

Conditional Probability

- Probability of *A occurring* ***given that*** B has occurred.
- The sample space is restricted to a single row or column.
- This makes rest of the sample space irrelevant.

Probability, i.e., $P(A|B) = \dfrac{P(A \text{ and } B)}{P(B)}$



Probability – Types:

Conditional Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Defaults	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

What is the probability that a person will not default on the loan payment **given she is middle-aged?**

Probability, i.e., $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$

P(No | Middle-Aged) = 0.586/0.690 = 0.85

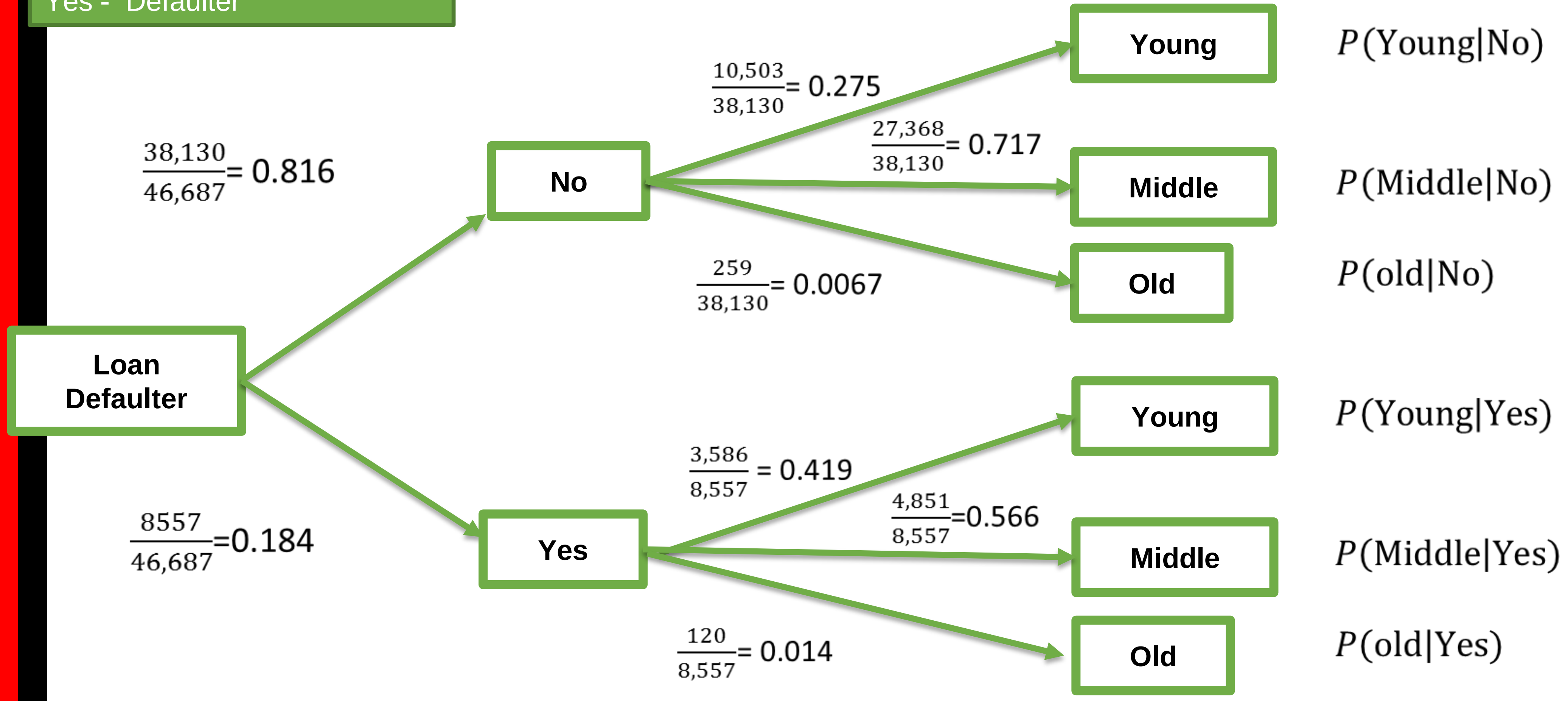
Note that this is the ratio of **Joint Probability to Marginal**

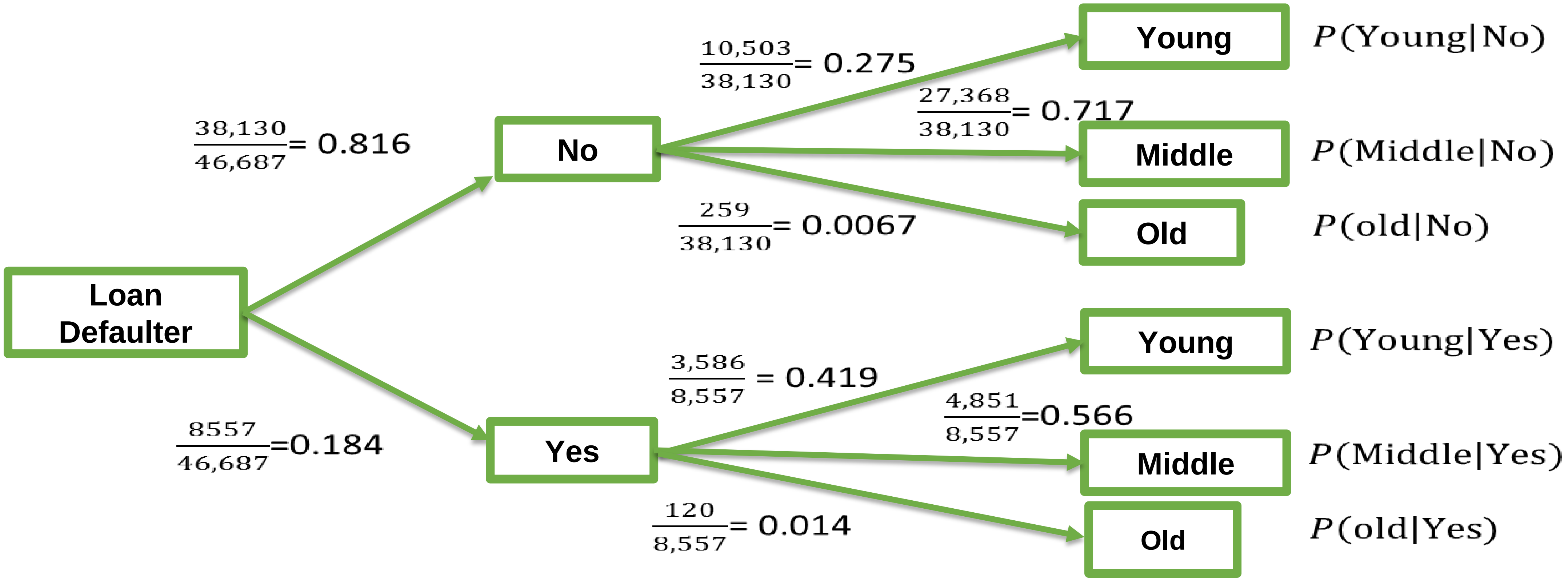
P(Middle-Aged | No) = $\frac{0.586}{0.816}$ = 0.72 (Order Matters)

		Age			
		Young	Middle-aged	Old	Total
Loan Defaults	No	10,503	27,368	259	38,130
	Yes	3,586	4,851	120	8,557
	Total	14,089	32,219	379	46,687

Age			
Young	Middle-aged	Old	Total
0.225	0.586	0.005	0.816
0.077	0.104	0.003	0.184
0.302	0.690	0.008	1.000

No – Non-defaulter
Yes - Defaulter




Loan Defaulter
$\frac{38,130}{46,687} = 0.816$
No
$\frac{8557}{46,687} = 0.184$
Yes
$\frac{10,503}{38,130} = 0.275$
Young
$P(\text{Young}|\text{No})$
$\frac{27,368}{38,130} = 0.717$
Middle
$P(\text{Middle}|\text{No})$
$\frac{259}{38,130} = 0.0067$
Old
$P(\text{old}|\text{No})$
$\frac{3,586}{8,557} = 0.419$
Young
$P(\text{Young}|\text{Yes})$
$\frac{4,851}{8,557} = 0.566$
Middle
$P(\text{Middle}|\text{Yes})$
$\frac{120}{8,557} = 0.014$
Old
$P(\text{old}|\text{Yes})$

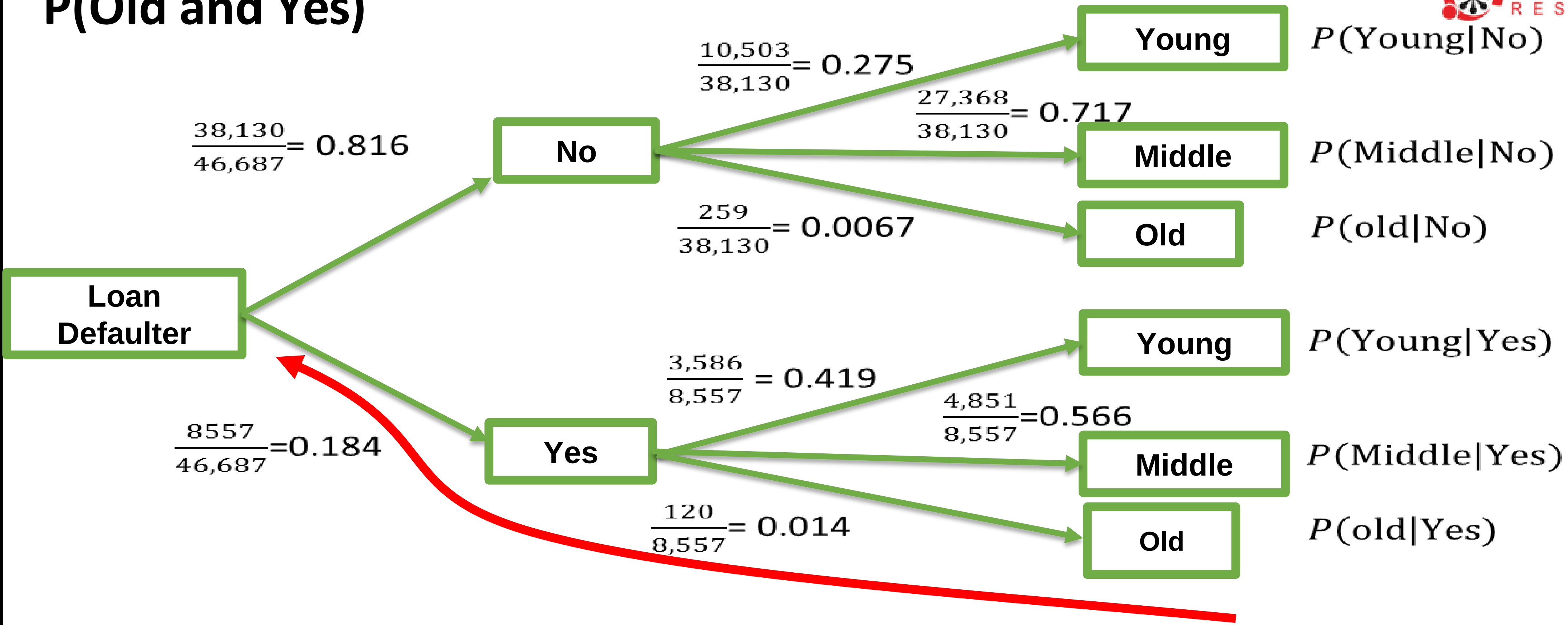
Check Time on Probability Types:

- $P(\text{Old and Yes}) = ?$
- $P(\text{Yes and Old}) = ?$
- $P(\text{Old}) = ?$
- $P(\text{Yes}) = ?$
- $P(\text{Old} \mid \text{Yes}) = ?$
- $P(\text{Yes} \mid \text{Old}) = ?$
- $P(\text{Young} \mid \text{No}) = ?$

Which Type of Probability:

- P(Old and Yes) = Joint Probability
- P(Yes and Old) = Joint Probability
- P(Old) = Marginal Probability
- P(Yes) = Marginal Probability
- P(Old | Yes) = Conditional Probability
- P(Yes | Old) = Conditional Probability
- P(Young | No) = Conditional Probability

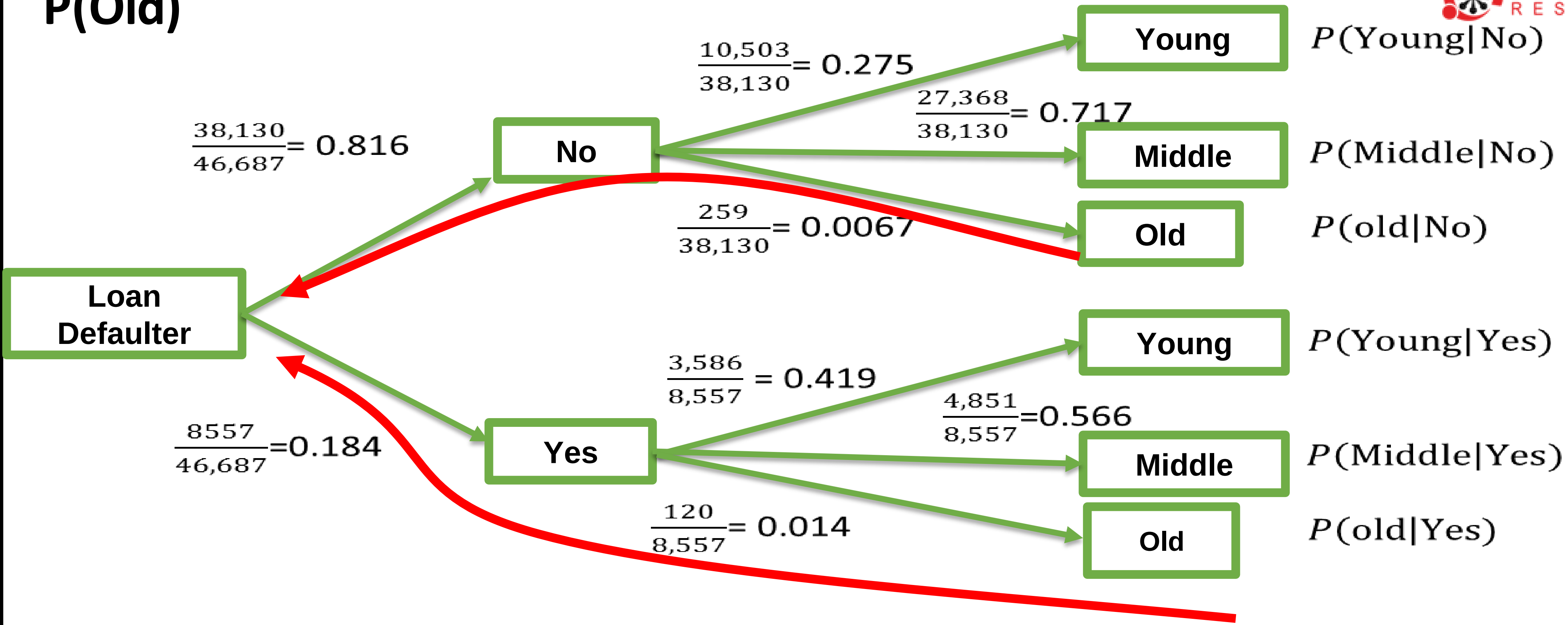
P(Old and Yes)



$$P(\text{Old} \mid \text{Yes}) = \frac{P(\text{Old and Yes})}{P(\text{Yes})}$$

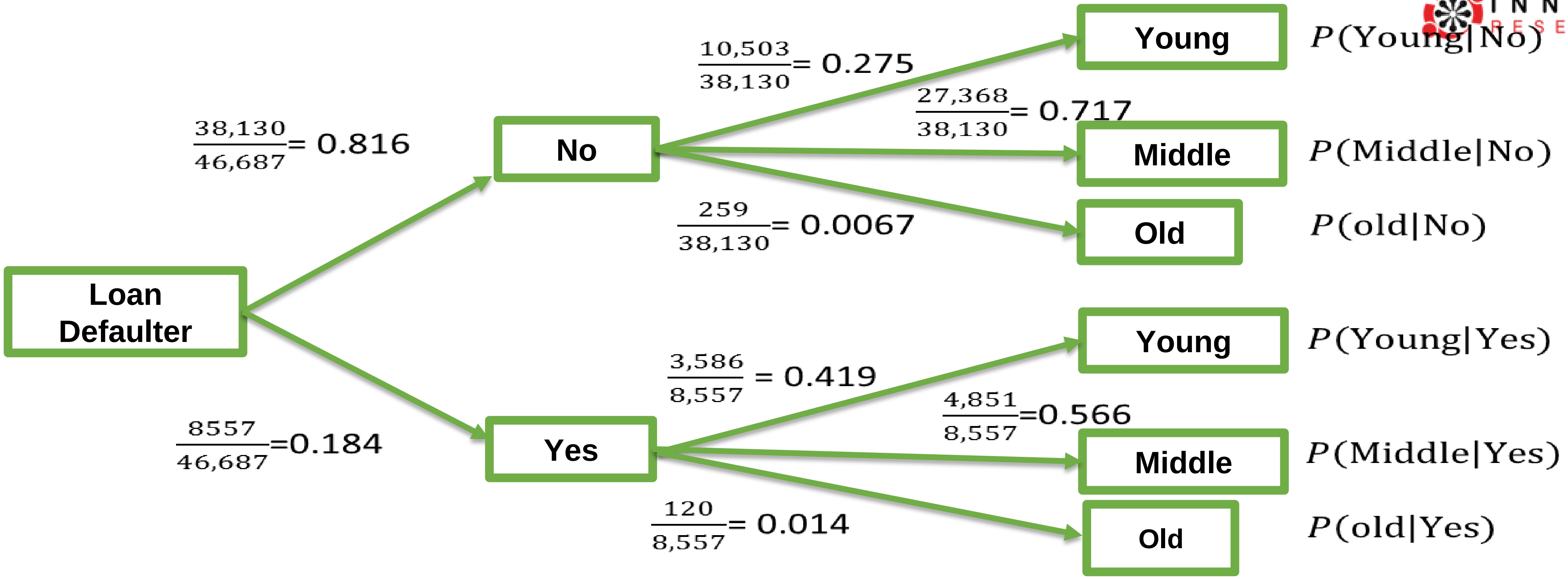
$$P(\text{Old and Yes}) = \text{P(Old} \mid \text{Yes)} * \text{P(Yes)} = 0.014 * 0.184$$

P(Old)



P(Old) = $P(\text{Old and Yes}) + P(\text{Old and No})$

$$P(Old) = \; 0.014 * 0.184 + 0.0067 * 0.816$$



- P(Old and Yes)
- P(Yes and Old)
- P(Old)
- P(Yes)
- P(Old | Yes)
- P(Yes | Old)
- P(Young | No)

Probability - Types

Conditional Probability

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \Rightarrow P(A \text{ and } B) = P(B) \, * \, P(A|B)$$

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} \Rightarrow P(A \text{ and } B) = P(A) \, * \, P(B|A)$$

Equating, we get

$$P(B) \, * \, P(A|B) = P(A) \, * \, P(B|A)$$

$$\therefore \; P(A|B) = \frac{P(A) \, * \, P(B|A)}{P(B)}$$

Probability - Types

$$\therefore\ P(A|B) = \frac{P(A)\ *\ P(B|A)}{P(B)}$$

In loan defaulters older people make up only 1.4%. Now the probability that someone defaults on a loan is 0.184, Find the probability default on loan knowing that he is old person. Older people make up only 0.8%.

Ans:

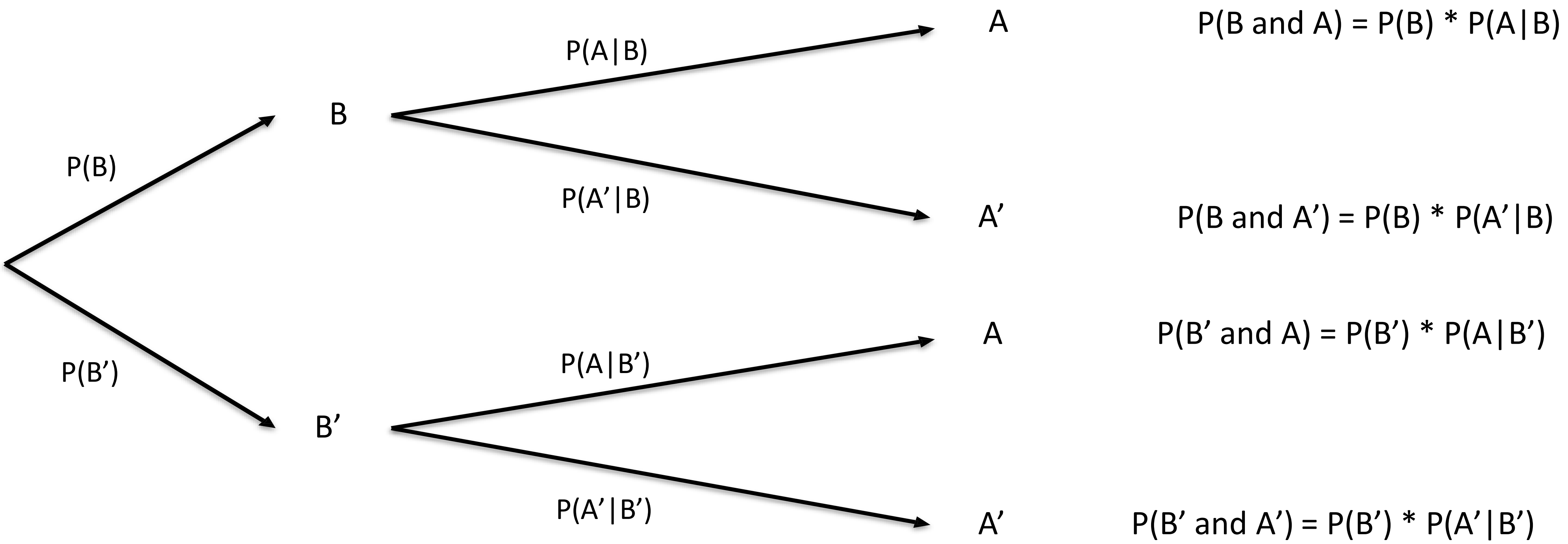
$$P(\mathrm{Old}|\mathrm{Yes}) = 0.014$$

$$P(Old) = 0.008$$

$$P(Yes) = 0.184$$

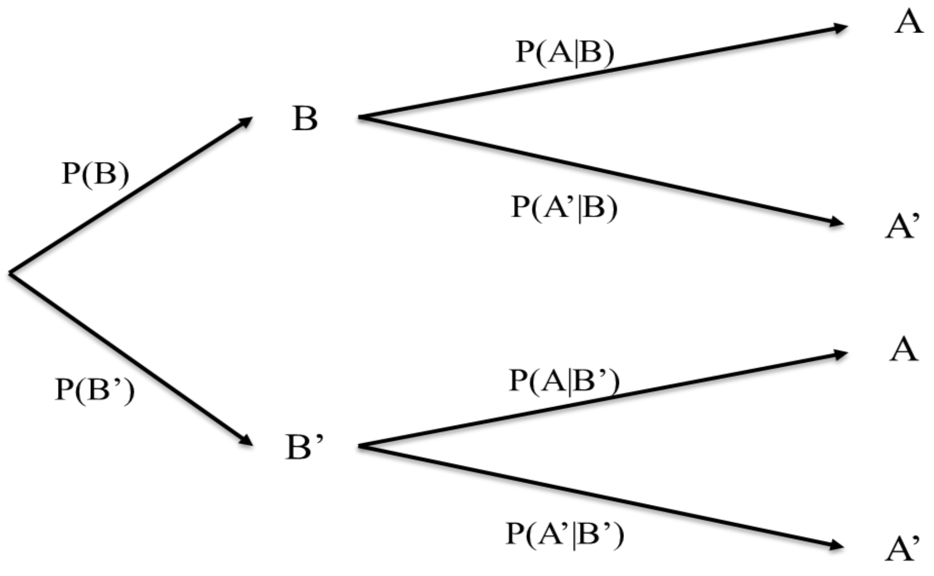
$$P(\mathrm{Yes}|\mathrm{Old}) = \frac{P(\mathrm{Yes}) * P(\mathrm{Old}|\mathrm{Yes})}{P(\mathrm{Old})} = \frac{0.184 * 0.014}{0.008} = 0.32$$

Generalized Probability Tree



State each probability in English; note B' means "not B".

Conditional Probability → Bayes Theorem



P(B and A) = P(B) * P(A|B)

P(B and A’) = P(B) * P(A’|B)

P(B’ and A) = P(B’) * P(A|B’)

P(B’ and A’) = P(B’) * P(A’|B’)

$$P(B|A) = \frac{p(B)*P(A|B)}{P(A)} = \frac{P(A|B)*P(B)}{P(A|B)*P(B)+P(A|Not\ B)*p(Not\ B)}$$

Bayes' Rule

- P(A) : Probability of event A occurring
- P(A|B): Probability of A occurring given B occurred
- P(B|A): Probability of B occurring given A occurred
- P(A n B): Probability of A and B occurring simultaneously **(Joint probability of A and B)**

  Joint probability of A and B

  P(A n B) = P(A|B)*P(B) = P(B|A)*P(A)

- Bayes' rule allows to find P(A|B) from P(B|A), i.e. to 'invert' conditional probabilities.
  - P(B|A) = $\frac{P(A|B)*P(B)}{P(A)}$
  - Also, P(A) = P(A|B)*P(B) + P(A|B')*P(B')

$$\frac{38{,}130}{46{,}687} = 0.816$$

Example

- A builder markets houses through brokers and online. Builder has data that 60% of prospective buyers who know about his houses know it through brokers, and 40% know through online website. He also knows that out of prospective buyers contacted through brokers, 25% actually end up buying a house; and out of those through who knew through online medium 20% end up buying. Which one is more effective medium for the builder (broker or online)?

Solution:

- Let us setup the problem.
- B: Marketing done through broker
- O: Marketing done online
- H: House actually bought
- $P(B) = .6$, $P(O) = .4$, $P(H|B) = .25$, $P(H|O)=.2$
- We need to know $P(B|H)$ and $P(O|H)$ to make decision.
- We know that

$$P(B|H) = P(B \cap H)/P(H)$$

$$\Rightarrow (P(H|B)\ P(B))/P(H) = .25 * .6/P(H) = .15/P(H)$$

- P(H) is not given to us. But we know that
  - $P(H) = P(H|B)P(B) + P(H|O)\,P(O)$
  - => .25*.6 + .2*.4 = .15 + .08 = .23
- Thus, $P(B|H) = .15/.23 = .6$
- Now, do the same for $P(O|H)$
- $P(O|H) = (P(H|O)\,P(O))/P(H)$
  - => .2*.4/.23 = .34
- Thus, 60% of potential buyers marketed through brokers buy house, but only 34% buy who are marketed through online.
- Therefore brokers is more effective medium

Bayes' Theorem ⇒Spam filtering

Apache SpamAssassin™

Spam Assassin works by having users train the system. It looks for patterns in the words in emails marked as spam by the user. For example, it may have learned that the word "free" appears in 30% of the mails marked as spam, i.e., P(Free | Spam) = 0.30. Assuming 1% of non-spam mail includes the word "free" and 50% of all mails received by the user are spam, find the probability that a mail is spam if the word "free" appears in it.

Draw the probability tree diagram

Bayes' Theorem:

P(Spam) = 0.50

P(Free | Spam) = 0.30 (*aka* Prior Probability)

P(Free | No spam) = 0.01

P(Spam | Free) = ? (*aka* Posterior or Revised Probability)

$$P(\text{ Spam } | \text{ Free }) = \frac{p(Spam) * p(free|spam)}{p(Free\ |Spam) * p(Spam) + p(Free|No\ Spam) * p(No\ Spam)}$$

$$= \frac{0.5 * 0.3}{0.3 * 0.5 + 0.01 * 0.5} = \frac{0.15}{0.155} = 0.9677$$

This helps the spam filter automatically classify the messages as spam.

Random Variable

What is a Random Variable?

- A random variable assigns a **numerical** value to each possible outcome of a **random experiment**.
  - Describes probability for an uncertain outcome of a random event
- Random because for any run of the experiment, one of several outcomes can be realized
  - With each outcome there is a chance (probability) associated
- To calculate the likelihood of occurrence of an event, we need to put a framework to express the outcome in numbers.
- We can do this by mapping the outcome of an **experiment to numbers**.

Examples for Random Variable:

Example 1:	Example 2:
Let's define X to be the outcome of a coin toss. X = outcome of a coin toss Possible Outcomes: 1 if heads 0 if tails	Suppose, I win the game if I get a sum of 8 while rolling two fair dice. I can define my random variable Y to be (the sum of the upward face of two fair dice) Y can take values = (2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12) **Intuitive thinking:** Each value of the random variable may or may not be equally likely. There is only 1 combination of dice, with sum 2{(1,1)}, while a sum of 5 can be achieved by {(1,4), (2,3), (3,2), (4,1)}. So, 5 is more likely to occur as compared to 2. On the contrary, the likelihood of a head or a tail in a coin toss is equal and 50-50.

Types of Random Variable?

Discrete:

- Finite number of outcomes
- The number of chocolates in a box

Continuous:

- All possible values in some range
- The amount of dessert I will eat today
- Large number of outcomes -> continuous distribution

Some more examples:

- **Discrete Random Variables:**
  - Credit Rating (usually classified into different categories such as low, medium and high or using labels such as AAA, AA, A, BBB etc.)
  - Number of orders received at an e-commerce retailer
  - Customer Churn (Random variables take binary values: a) Churn and b) Do not churn)
  - Fraud (Random variables take binary values a) Fraudulent transaction and b) Genuine Transaction)
  - Any experiment that involves counting (for example: number of returns in a day from customers of e-commerce portals such as Amazon, Flipkart, number of customers not accepting job offers from an organization)

Note: In analytics, classification problems, an important class of problems, is an example of discrete random variable

Some more examples:

Continuous Random Variables

1) Market share of a company (which takes any value from an infinite set of values b/w 0% and 100%)
2) Percentage of attrition among employees of an organization
3) Time to failure of engineering systems
4) Time taken to complete an order placed at an e-commerce portal
5) Time taken to resolve a customer compliant at call & service centres

Discrete or Continuous?

- Today's weather:
- The color of a car chosen at random:
- Height of a student in class:
- Number of insurance policies issued by a company on a day:
- Sometimes, it is good to approximate discrete as continuous

Discrete or Continuous?

- Today's weather: **Continuous**
- The color of a car chosen at random: **Discrete**
- Height of a student in class: **Continuous**
- Number of insurance policies issued by a company on a day: **Discrete**
- Sometimes, it is good to approximate discrete as continuous

Example for Probability Distribution

- Consider the simple experiment of tossing a coin three times.
- Let X = number of times the coin comes up heads. The 8 possible elementary events, and the corresponding values for X, are:

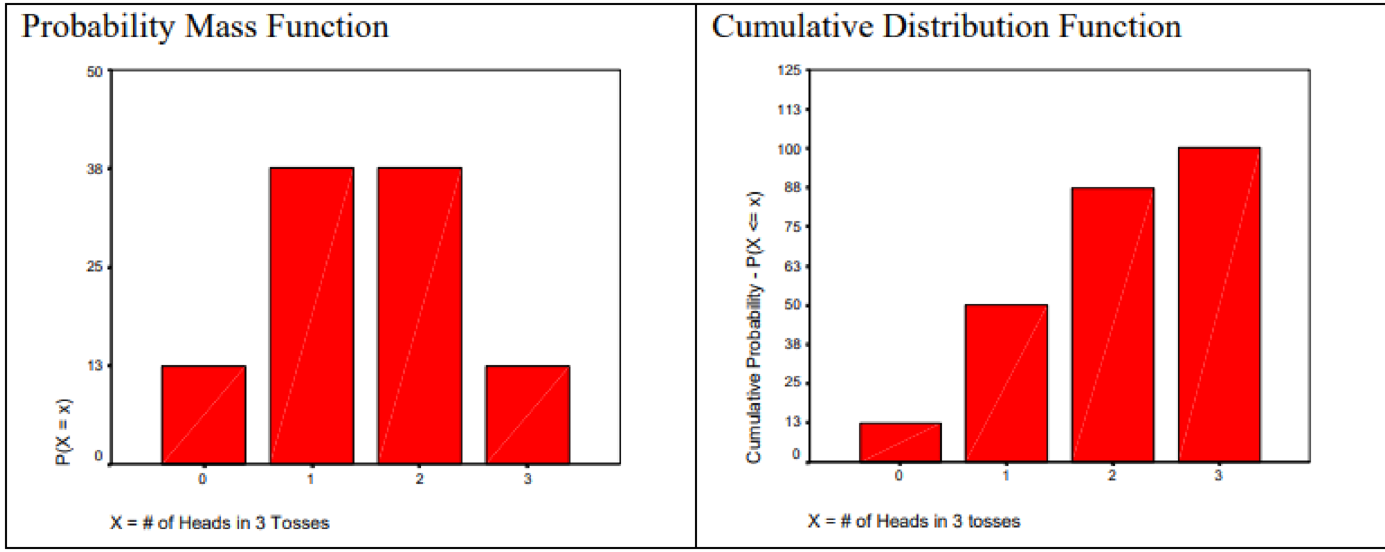
Elementary event	Value of X
TTT	0
TTH	1
THT	1
HTT	1
THH	2
HTH	2
HHT	2
HHH	3

Example for Probability Distribution

- Therefore, the probability distribution for the number of heads occurring in three coin tosses is:

x	p(x)	F(x)
0	1/8	1/8
1	3/8	4/8
2	3/8	7/8
3	1/8	1

Graphically, we might depict this as



Summarizing a Random Variable

- Random variables need to be summarized

- Most times, summarization is done using two numbers:
  - Central tendency or mean or expected value
  - Dispersion or spread

There are others as well, but for now we will focus on these two

Expected Value or Mean

- Measure to find out expected(average) value
- Mean is probability weighted average value that is ‘expected’ to occur
- For discrete RV: μ = E(X) = $\Sigma x * p(x)$

Outcome x	Probability p
10	1/6
20	1/6
30	1/6
40	1/6
50	1/6
60	1/6

Mean = 10*(1/6) + 20*(1/6) + 30*(1/6) + ….. + 60*1/6 = 35

$$\mu = E(X) = \Sigma x * p\ x$$

Computing the spread:

- Sum of squared differences from the mean is the variance
- $\Sigma pi\ (xi - \mu\ )2$ = Var(X)

Outcome x	Probability p
0	0.03
1	0.14
2	0.35
3	0.32
4	0.15
5	0.01

Var (X)

=0.03 * (0 – 2.45) ^ 2 + 0.14 * (1 – 2.45) ^ 2+ …..+ 0.01 * (5 – 2.45) ^ 2

=1.0675

σ =STD(X)= square root of Var(X)

= (1.0675) ^ (1/2) = 1.0332

Probability Distribution

Probability Distribution:

- The probability distribution of random variable describes how the probability is distributed over the outcomes of random variable.
- There are lots of distribution functions. Both discrete and continuous in nature
- We will study commonly used distributions:
  - **Discrete** : Binomial, Poisson
  - **Continuous** : Uniform, Normal
- Note that in case of discrete random variable, the probability distribution is defined by **probability mass function.**
- In case of continuous random variable, the probability distribution is defined by **probability density function**.

Bernoulli:

- Let us flip a coin. X is a Random variable with Bernoulli distribution

  - **X ~ Bernoulli(p)**

Outcome x	Probability p
1	p
0	1-p

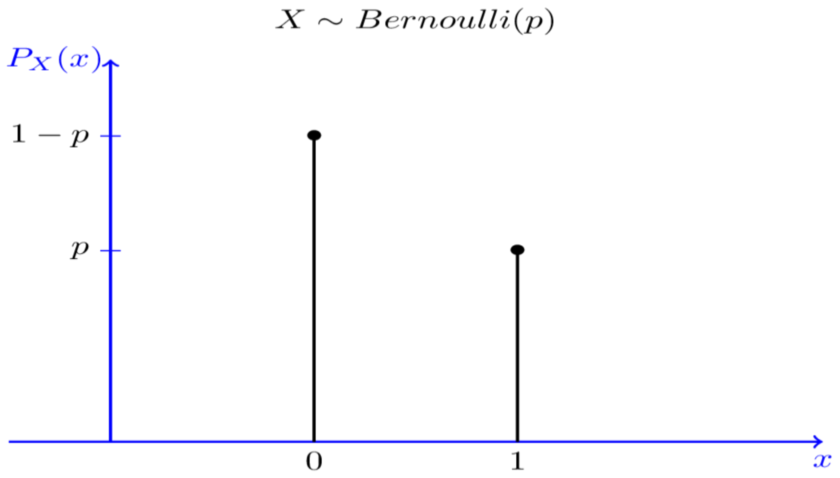
For example: In Hyderabad, every person has 60% chance of having a vehicle.

- 1 if they have vehicle, 0 otherwise
  - **X ~ Bernoulli(.60)**

Bernouli

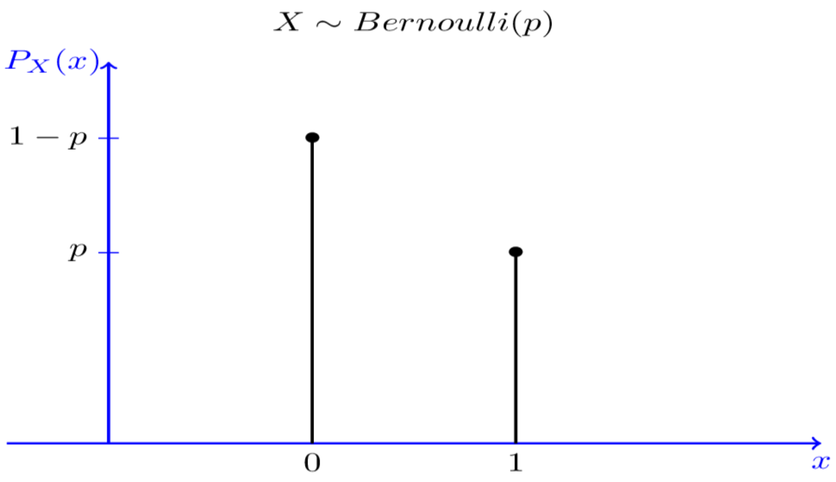
- There are two possibilities (pass or fail) with probability *p* of

  success and *q = 1-p* of failure..



Expectation : p
Variance : pq

Bernouli



$$Expectation, E(x) = \sum x_i \, P(x_i)$$

$$= 1 * p + 0 * q$$

$$= p$$

$$Variance, Va(x) = \sum (x_i - \mu)^2 \; P(x_i)$$

$$= (1 - p)^2 * p + (0 - p)^2 * (1 - p)$$

$$= p\,(1 - p)$$

$$= pq$$

Bernoulli Distribution

- Extension of Bernoulli trials to n times, each with the same probability of success.
- Select 10 people in Hyderabad at random.
- Let X = number of them having a vehicle
- Of 10 randomly chosen people, what is the probability that first 8 will have vehicle and the next 2 not?
  - P(YYYYYYYYNN) = (.60) ^ 8 * (.40) ^ 2
- Probability of first 3 having a vehicle and next 7 not?
  - P(YYYNNNNNNN) = (.60) ^ 3* (.40) ^ 7

Bernoulli Distribution Contd….

- P(X=3)?
  - P(YYYNNNNNNN) = (.60)3*(.40)7
  - P(YYNYNNNNNN) = (.60)3*(.40)7
  - P(NNNYNNNYNY) = (.60)3*(.40)7

And so on….

- In order to arrive at final probability, we would have to add up all possible cases. How many such arrangements are there?

Detour on Permutations & Combinations

Imagine 10 Innomatics students are sitting in front row & I have to pick 3 of them, in how many ways can I do it?

Answer:

How many ways to arrange 10?

- 10! = 10*9*8*7*6*5*4*3*2*1

Let us say I pick 3 people. In how many ways I can arrange them?

- 3!

For the 7 I did not pick, how many ways?

- 7!

Therefore, total number of ways to pick them are: 10! / (3! * 7!) = 120 ways

To generalize: (nCr), where n is the number of observations .i.e 10C3 = 120

Getting back to Binomial

- Number of such cases are 10C3 = 120 ways

- P(X=3) = 120 * (0.60)^3 * (0.40) ^ 7

- Generally, it can be written as:

$$P(X = x) = \frac{n!}{x!n - x!}(p)^x(1 - p)^{n-x}$$

Summarizing Binomial Distribution:

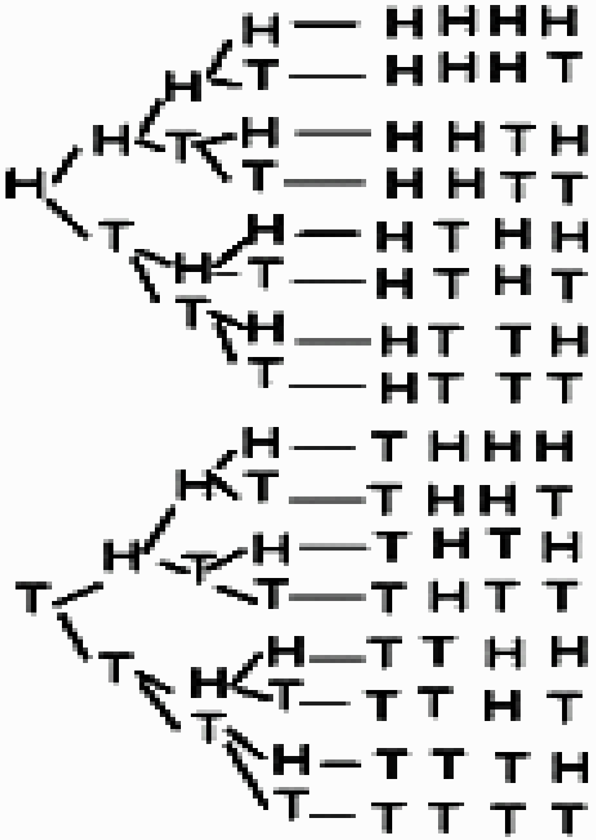
- There are n independent trials.
- Each trial has two outcomes only: 1 or 0 with probability p & 1-p respectively
- Parameters of a binomially distributed random variable are **Number of trials** (n) and **Probability of success** (p).
- It is denoted as X ~ Bin(n,p)
- **PMF** of X is given by

$$P(X=x)=\frac{n!}{x!n-x!}(p)^x(1-p)^{n-x}$$

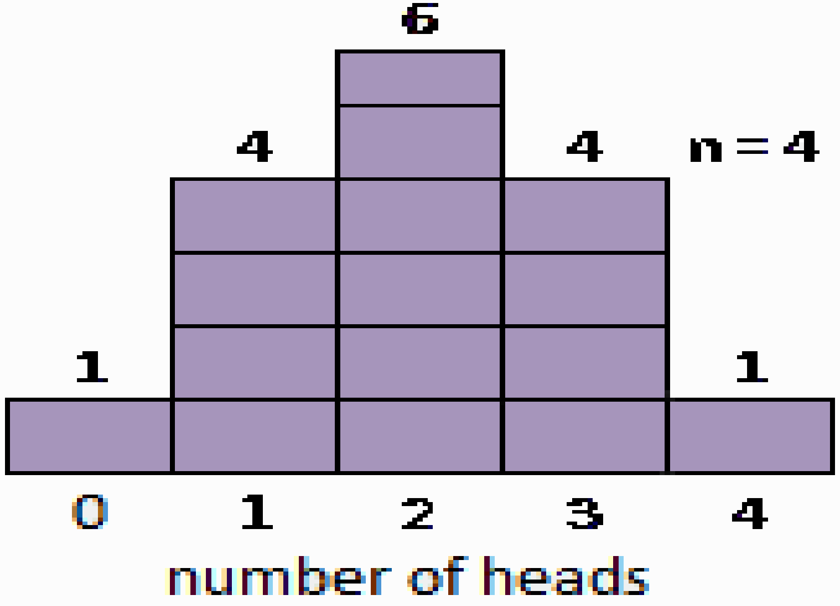
- For X, E(X) = np and V(X) = n*p*(1-p)

Binomial distribution

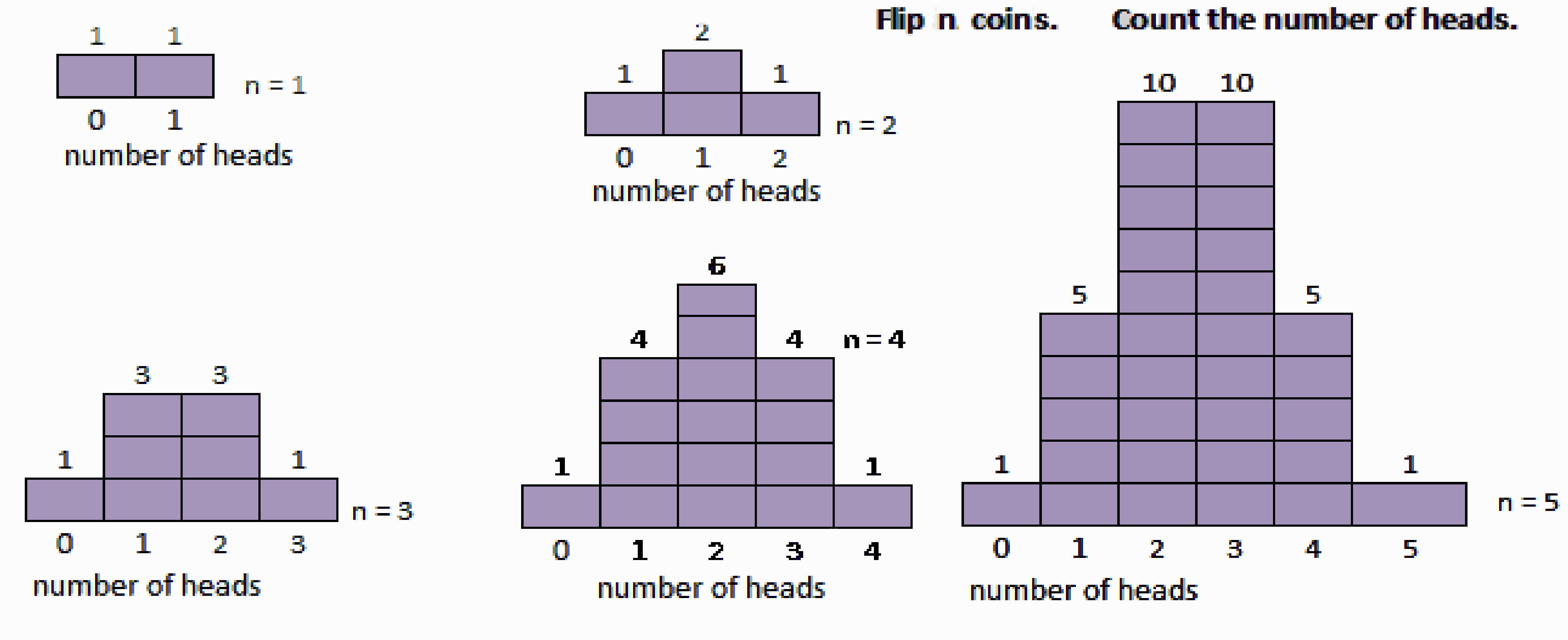
Flip a coin four times. Count the number of heads.



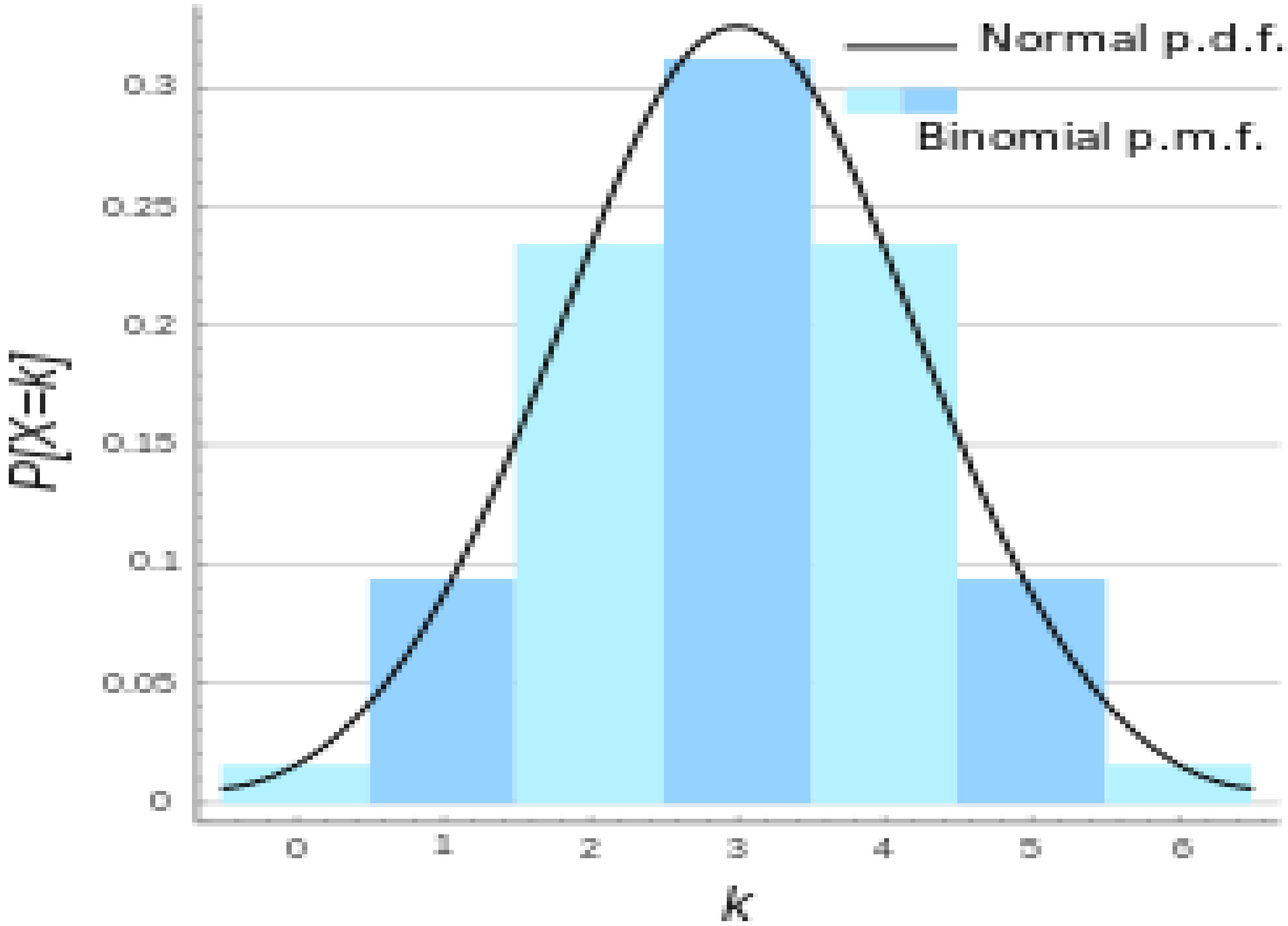
x	P(x) = p	
0	0.0625	1 /16
1	0.25	4 /16
2	0.375	6 / 16
3	0.25	4 /16
4	0.0625	1/16



Binomial distribution



Binomial distribution



Applications of Binomial

- If an insurance company knows the probability of a claim **being fraudulent (or not)**, binomial distribution can help determine the probability that there would be more than ‘k’ fraudulent claims in the next 1000 claims
- If an advertising company knows the probability of a customer **buying a certain product**, if he/she receives an advertisement on their Facebook account, binomial distribution can help them determine whether ‘k’ purchases would be made from a total of 10000 Facebook ads sent. Effectively, they can determine how many Facebook advertisements to send, if they want to achieve a particular sales target
- Suppose that a hardware store manager has 3 different suppliers. If he knows the probability of a newly purchased bolt **being defective**, based on the supplier it is bought from, binomial distribution can help him/her check whether there will be less than ‘k’ defective bolts in a box of 2000 bolts. This would help his decide which supplier to buy the bolts from.

Properties of Binomial

- As the value of ‘n’ increases, if n is very large, p is very small and **n*p** is finite:
  - Binomial Distribution => Poisson Distribution
    - Poisson distribution allows to calculate the probability of ‘k’ successes without specifying the number of trials ‘n’

- As the value of ‘n’ becomes very large:
  - Binomial Distribution => Normal Distribution

Poisson Distribution

- For situations, if n is large, p is small, and np is a finite number, then we can use the following PMF function:

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \qquad x = 0,1,2,\ldots$$

Where λ = np

- E(X) = λ and V(X) = λ

- It is essentially an extension of Binomial distribution

Poisson Distribution: Example

There are more than 50 Million active phone connections in India used daily. On average, 1000 phones break because of falling while traveling.

Question: Compute the probability that my phone will break while traveling today.

It can be approached using Binomial distribution

- p = P(phone breaks while traveling) = 1000/50 Million
- n = 50 Million
- Thus, Probability =

Computationally very intensive !

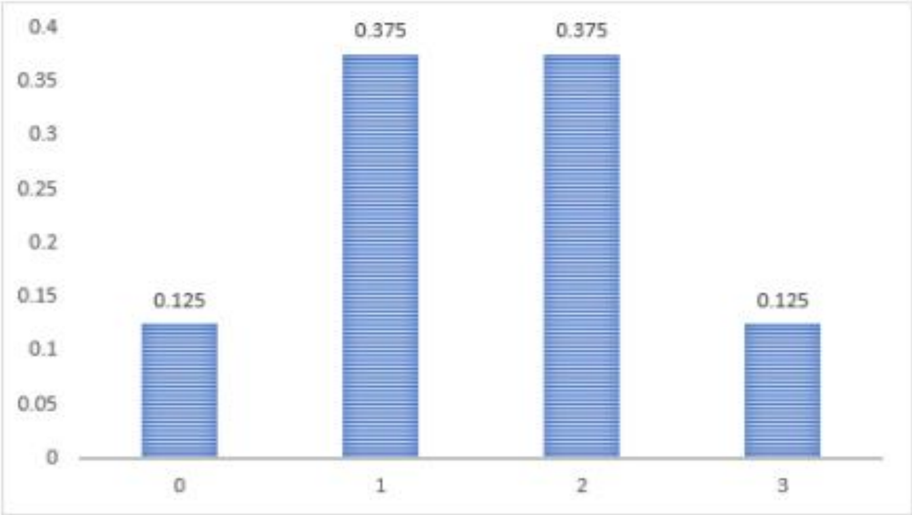
Discrete RV Distribution example

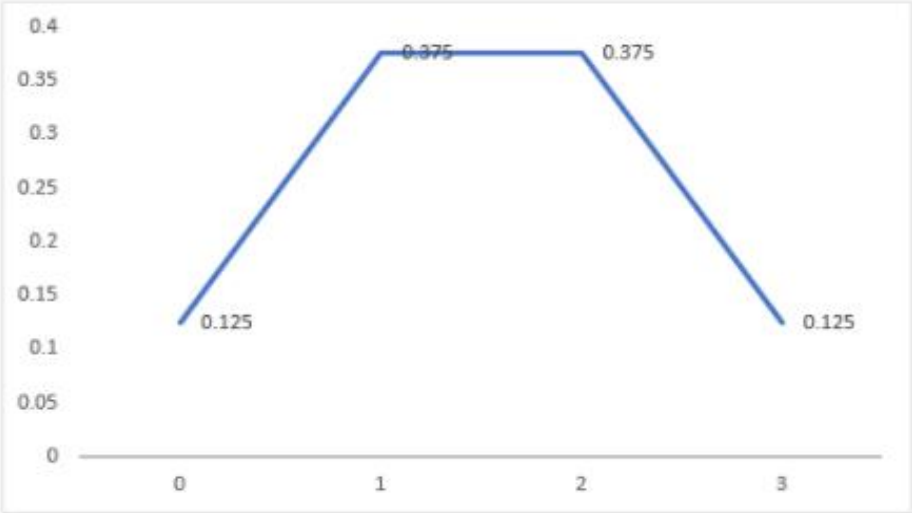
- Let's assume that you love playing badminton. You have been practicing it quite often.
- Below given the data of your practice sheet for 8 days. Your practice is segregated in 3 times in a day. "No" signifies that badminton is not played and "Yes" signifies that badminton is played.
- **X is a random variable** which signifies the number of times badminton is played.

Day	Shift 1	Shift 2	Shift 3
1	No	No	No
2	No	No	Yes
3	Yes	No	Yes
4	Yes	No	No
5	No	Yes	Yes
6	No	Yes	No
7	Yes	Yes	No
8	Yes	Yes	Yes

Example contd.

- Plotting the probability for X (random variable) in Y-axis and the X value or outcomes in X-axis we get the probability distribution of the discrete random variable X.
- The probability distribution of discrete random variable is known as **Probability mass function**.
- Some example of **discrete probability distribution** is Bernoulli distribution, Binomial distribution, Poisson distribution etc.





Another example

- Consider the number of daily fraudulent transactions at a bank branch & the corresponding probabilities given in table below.
- The values denote possible values of the random variable & corresponding probability .i.e. describes how probability is distributed across different values of the random variable

Random Variable X (X= no of fraudulent transactions)	xi=0	xi=1	xi=2	xi=3	xi=4
P(X=xi)	0.20	0.15	0.25	0.25	0.15

Information from the given table

P(X=0, no fraudulent transaction on any given day) = 0.20

Similarly, P(X=1) =0.15, P(X=2) =0.25, P(X=3) =0.25 and P(X=4)=0.15

- **Probability Mass Function (PMF),** P(xi) satisfies the following conditions:
  - P(xi) >=0
  - ΣP(xi)=1
- **Cumulative Distribution Function (CDF)**, F(xi) is the probability that the random variable X takes values <=xi
  - F(xi) = P(X<=xi)

F(2) = P(X<=2) = P(X=0) + P(X=1) + P(X=2) = 0.60

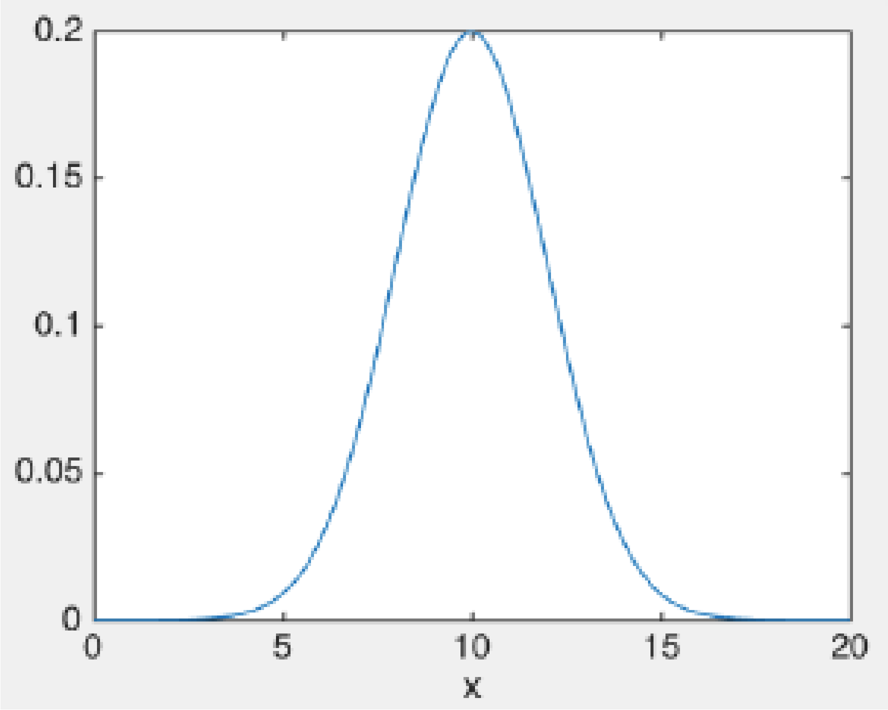
Continuous Distributions

- Amount of rainfall in Hyderabad in any month: **Continuous Random Variable**
- We have a range but it can take any values in the range (say 0-20 cm)
- Probability that rainfall in a month would be 10 cm?
  - Zero
  - Infinite possibilities and probability of any single event is zero.

PMF is no longer applicable. Replaced by Probability Density Function (PDF)

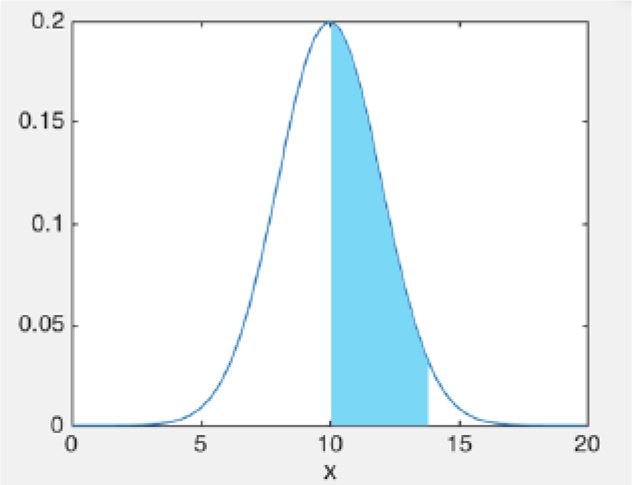
- Probability of rainfall between 9.9 – 10.1 cm?
- We can find probability in a particular area

PDF example

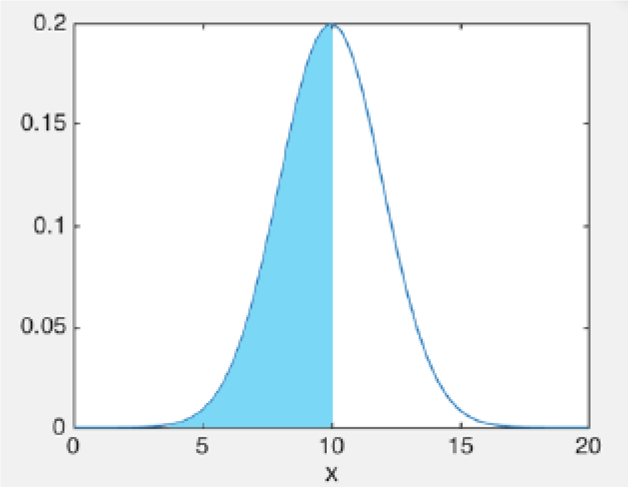


- Amount of rainfall (in cm) to fall in Hyderabad during month of August.

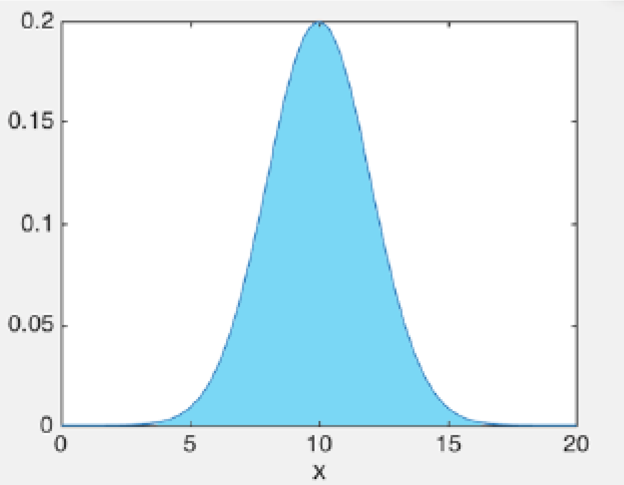
These figures below depict a Probability Density Function (PDF)



$P(10 \leq X \leq 14)$



$P(X \leq 10)$



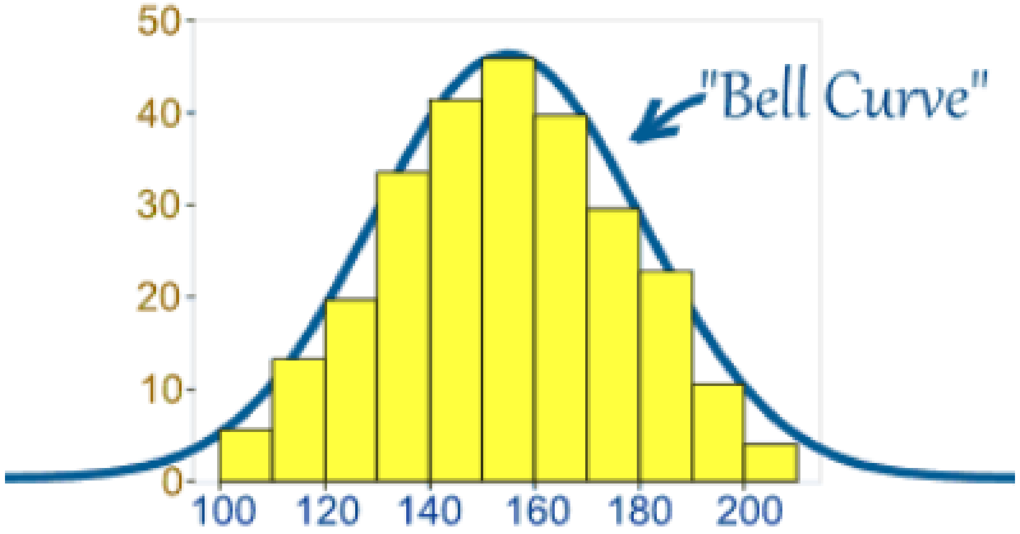
$P(X \leq 20)$

Normal Distribution

Normal Distribution

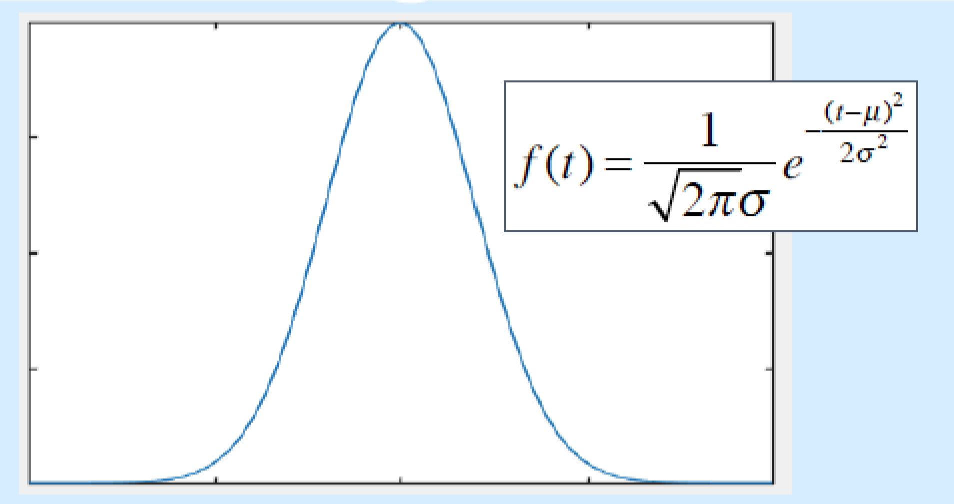
- One of the most important continuous distributions most of the real life phenomena can be modeled using a normal distribution
    - Height, weight, grades, salary ….

- A normal distribution can graphically be represented as a bell shaped curve



Normal Distribution Equation

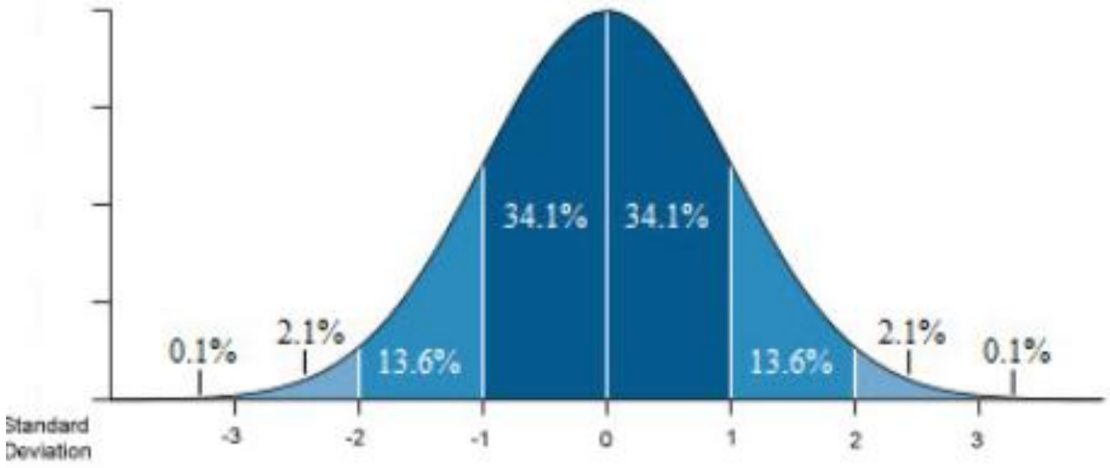
- Just like we can specify a Binomial distribution by n and p, we can specify a normal distribution by μ **and** σ
- (μ is mean of normal random variable and σ is standard deviation)
- Formally, **X ~ N(μ ,σ2)**
- The curve is **pdf** of a normal random variable

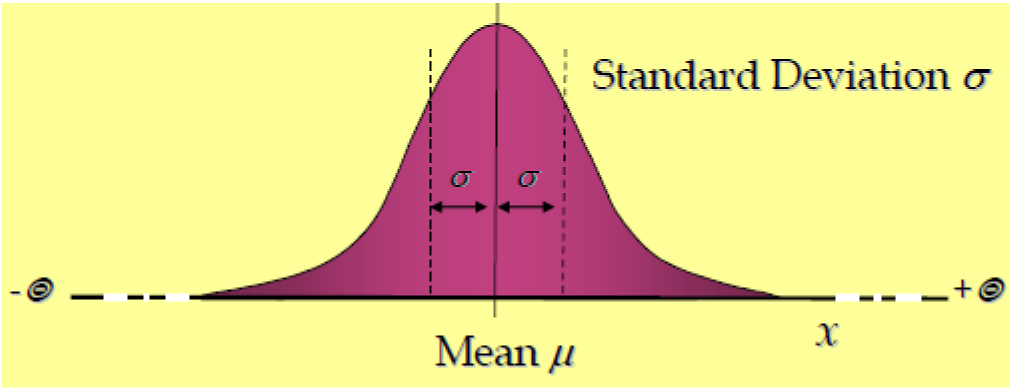


$$f(t) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$$

Normal Distribution contd.

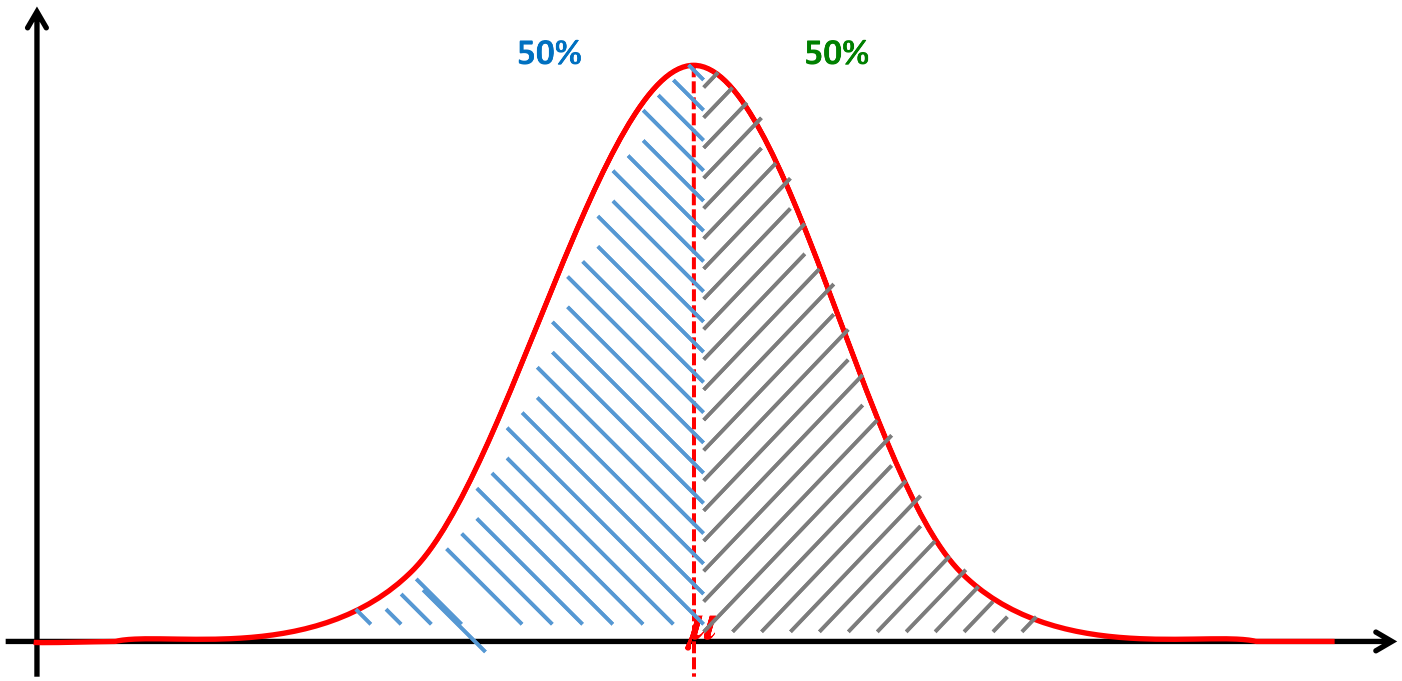
- **X ~ N(μ ,σ2)**
- A normal random variable can be completely specified by μ ,σ
  - Takes values from - ∞ to +∞
  - μ specifies location of centrality, and σ specifies width of curve
  - Symmetric around mean (mean=median=mode)
  - 68.2% area is within one s.d (σ) away from mean (μ) and 95% within 2 s.d away from mean





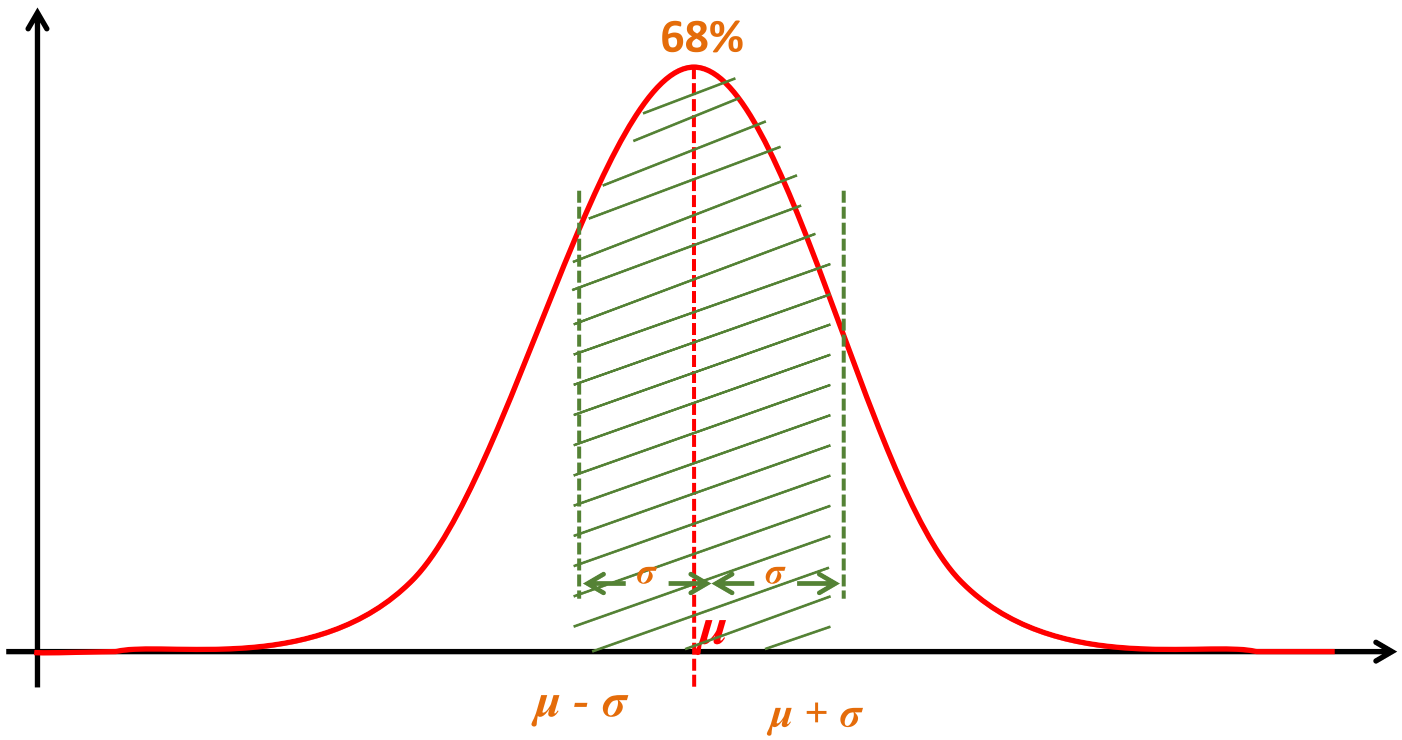
The properties of a normal distribution:

- It is a bell-shaped curve.
- It is symmetrical about the mean, μ. (The mean, the mode and the median all have the same value).
- The total area under the curve is 1 (or 100%).
- 50% of the area is to the left of the mean, and 50% to the right.



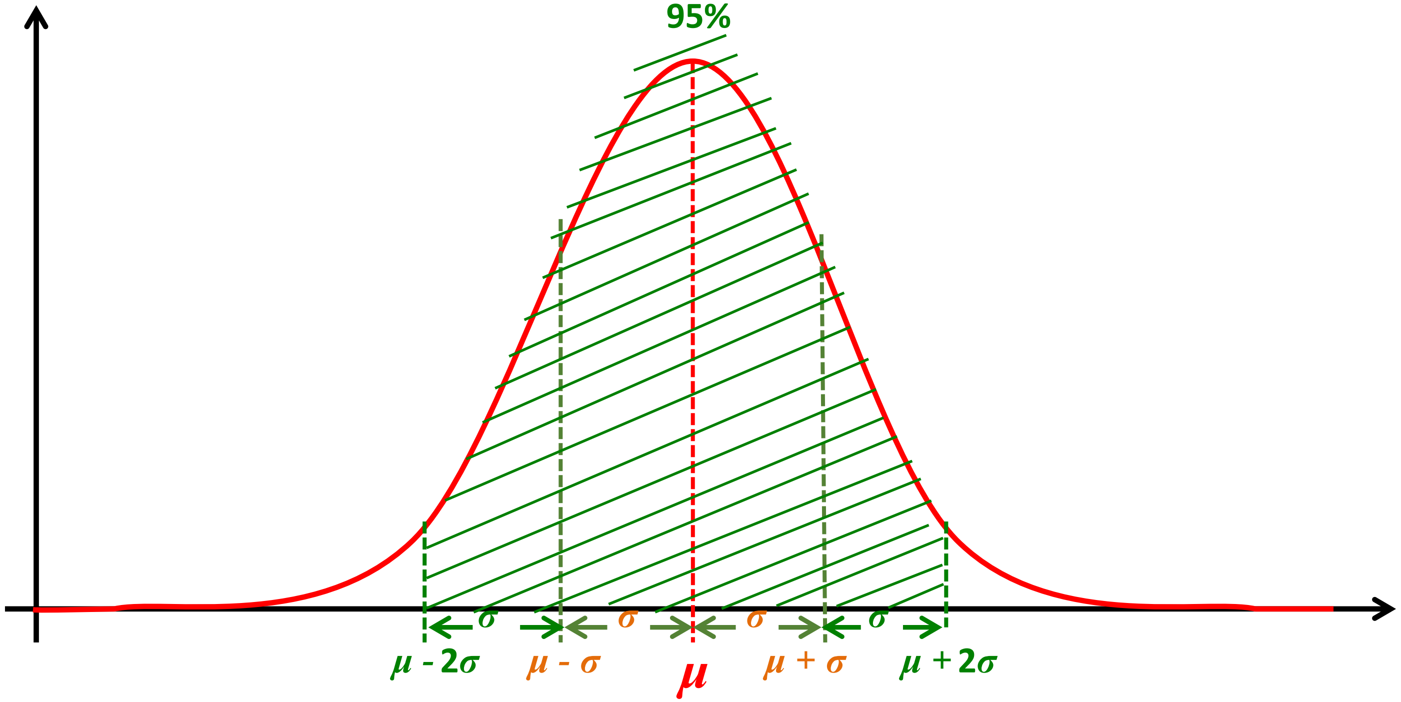
The properties of a normal distribution:

- It is a bell-shaped curve.
- It is symmetrical about the mean, μ. (The mean, the mode and the median all have the same value).
- The total area under the curve is 1 (or 100%).
- 50% of the area is to the left of the mean, and 50% to the right.
- Approximately 68% of the area is within 1 standard deviation, σ, of the mean.



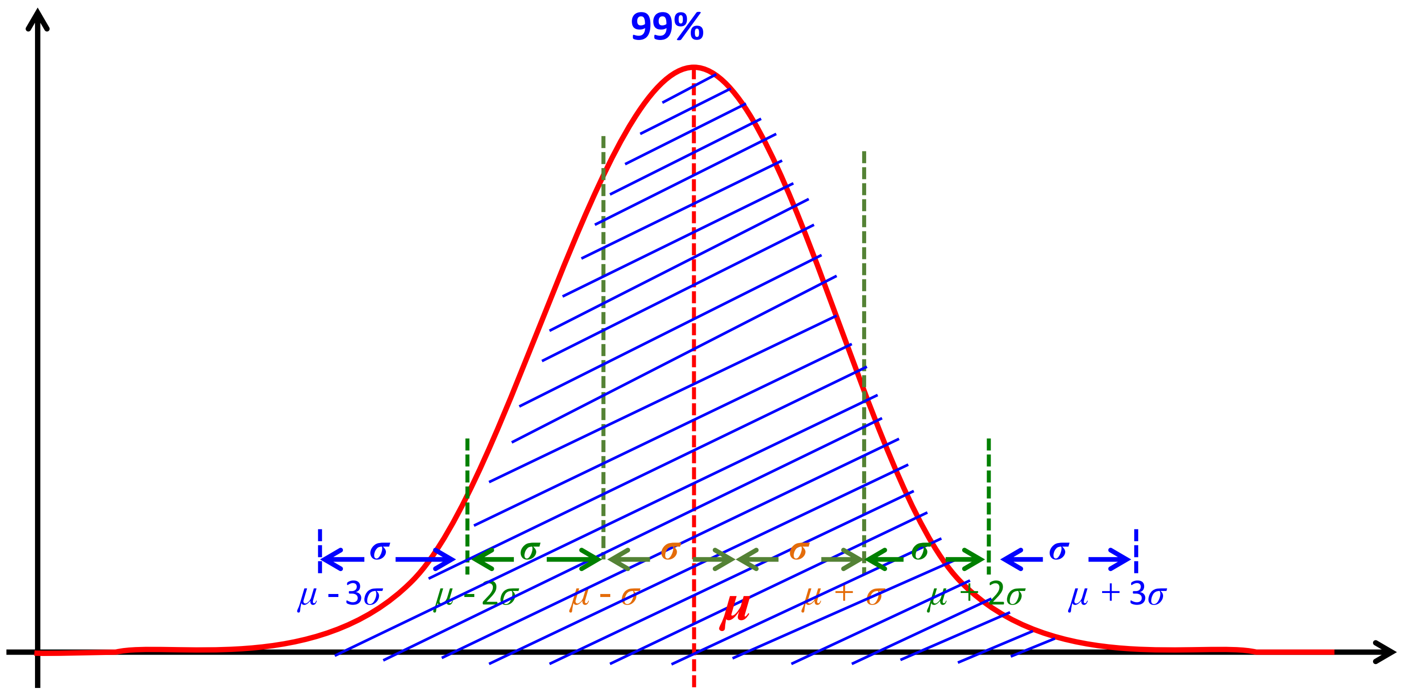
The properties of a normal distribution:

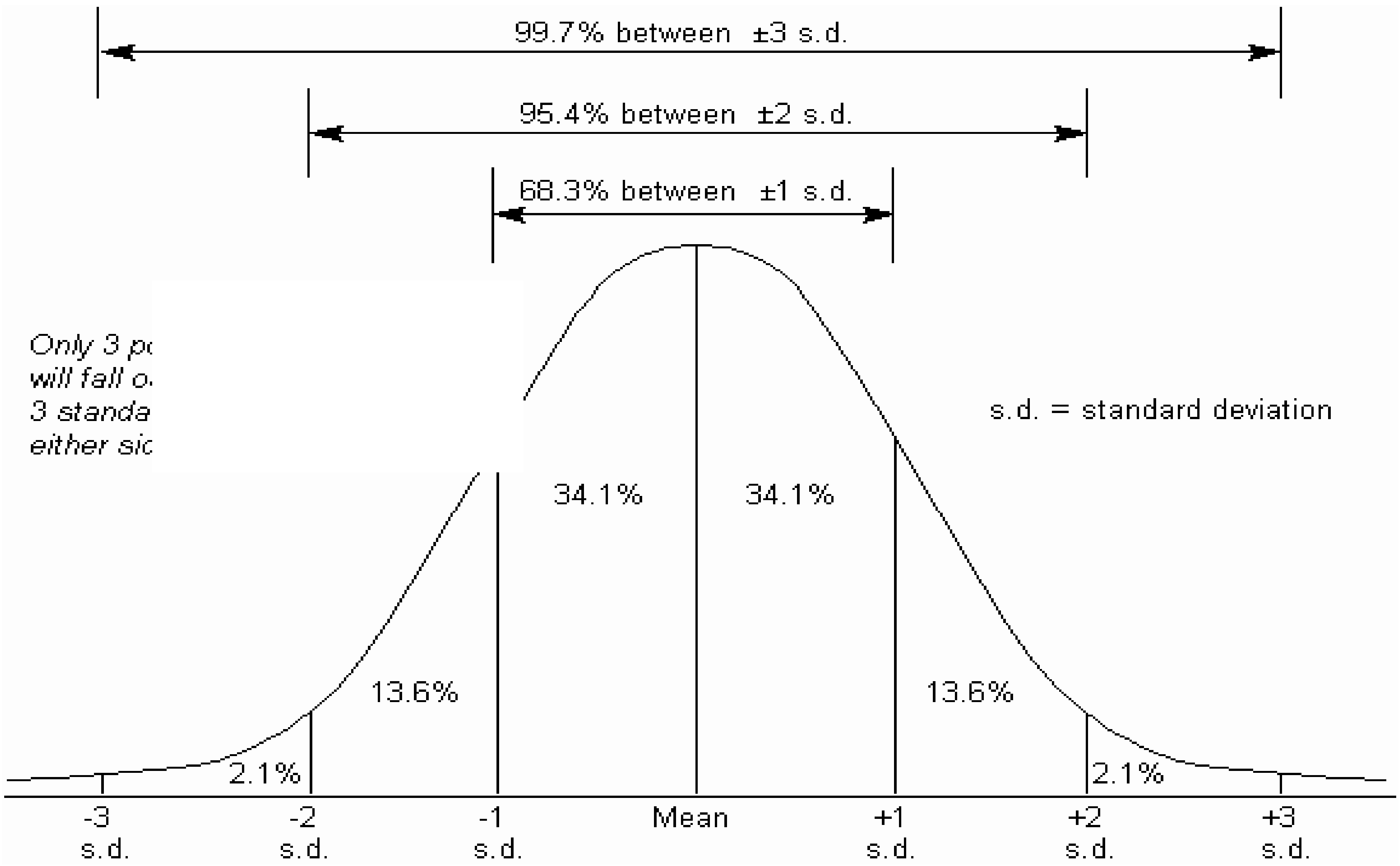
- It is a bell-shaped curve.
- It is symmetrical about the mean, μ. (The mean, the mode and the median all have the same value).
- The total area under the curve is 1 (or 100%).
- 50% of the area is to the left of the mean, and 50% to the right.
- Approximately 68% of the area is within 1 standard deviation, σ, of the mean.
- Approximately 95% of the area is within 2 standard deviations of the mean.



The properties of a normal distribution:

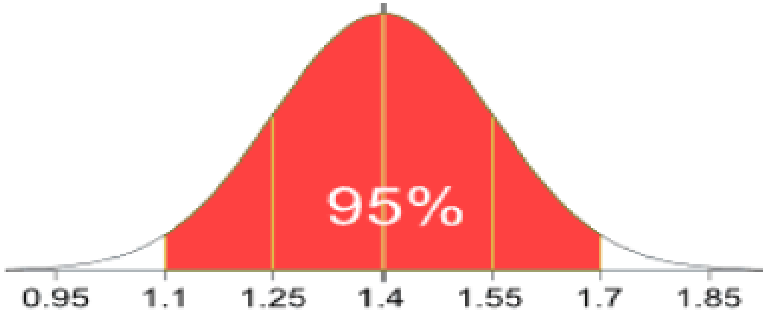
- It is a bell-shaped curve.
- It is symmetrical about the mean, μ. (The mean, the mode and the median all have the same value).
- The total area under the curve is 1 (or 100%).
- 50% of the area is to the left of the mean, and 50% to the right.
- Approximately 68% of the area is within 1 standard deviation, σ, of the mean.
- Approximately 95% of the area is within 2 standard deviations of the mean.
- Approximately 99% of the area is within 3 standard deviations of the mean.




99.7% between ±3 s.d.
95.4% between ±2 s.d.
68.3% between ±1 s.d.
Only 3 pc
will fall o.
3 standa
either sic
s.d. = standard deviation
34.1%
34.1%
13.6%
13.6%
2.1%
2.1%
-3 s.d.
-2 s.d.
-1 s.d.
Mean
+1 s.d.
+2 s.d.
+3 s.d.

Normal Distribution: Example

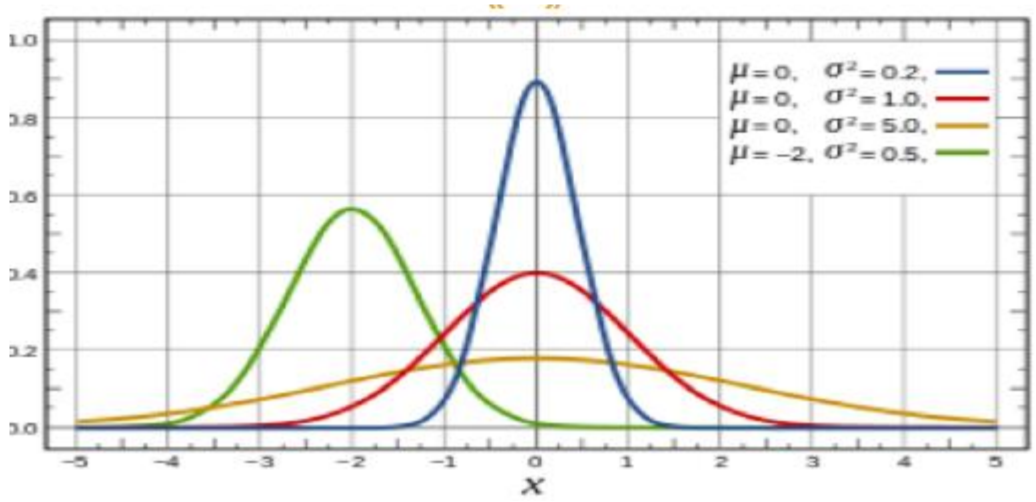
Height of 95% of Innomatics students varies between 1 m and 1.6 m. Assume that height is normally distributed.



- **What is the mean?**
  - Mean = (1+1.6)/2 = 1.3 m
- **What is standard deviation?**
  - We know 95% represents 2 sd away from mean (on both sides)
    - (1.6-1)/4 = 0.15 m is standard deviation
- **What does it mean?**
  - It says that out of every 1000 values 680 should be within 1 sd (likely)
  - For every 1000, 950 should be within two sd (very likely)
  - For ever 1000, 997 within 3 sd (almost certainly)

Normal Distribution contd.

- There can be many different shapes (either because of change in mean or standard deviation).

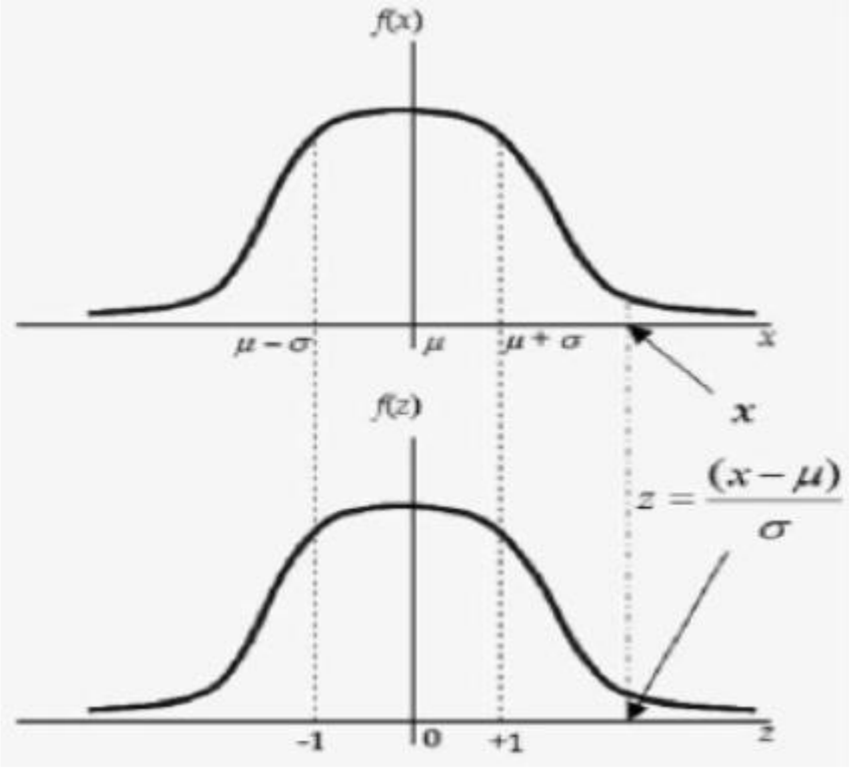


Standardizing the Normal Distribution

- Assume X1 ~ N(10 ,106), and X ~ N(7 ,56)
- If you have to compare these two, how would you do that?
  - **Standardizing** comes to rescue

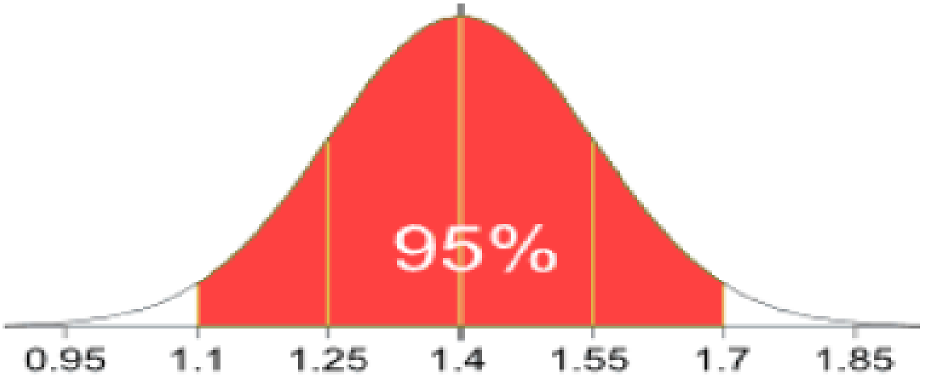
We define standard normal variable (Z) with mean 0 and sd 1

- $Z=(X-\mu)/\sigma$
- Z also has a normal distribution
- Z score tells position of a point relative to other points in distribution
- Alternatively, it tells how many s.d. away from mean a point is



Recalling the example

- Assume following distribution for height of students in Innomatics class

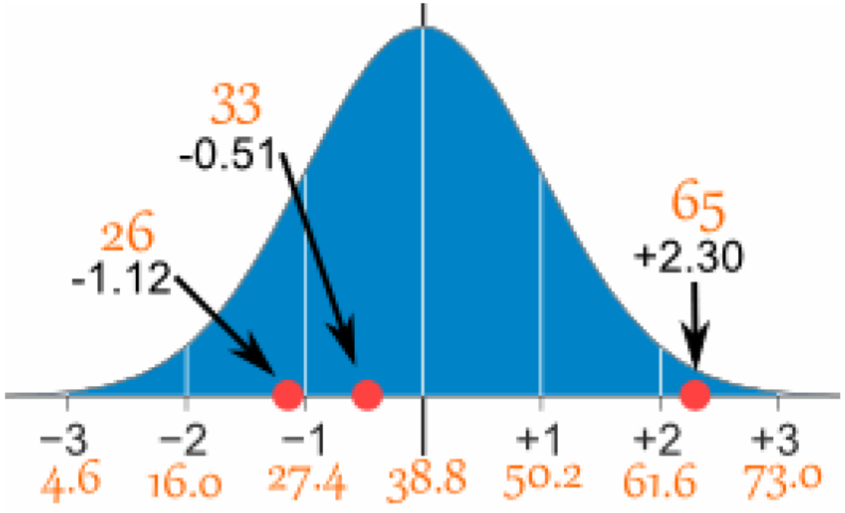


- Say, height of a student is 1.85m.
  - Student is 3 s.d. away from mean
  - Height of student has **z-score of 3**

Another Example

Assume following daily temperatures in **Hawaii**

- 26, 33, 65, 28, 34, 55, 25, 44, 50, 36, 26, 37, 43, 62, 35, 38, 45, 32, 28, 34
- Mean?
  - **38.8**
- S.D?
  - **11.4**
- Convert them to z-scores $Z=(X-\mu)/\sigma$
  - 26 -> **(26-38.8)/11.4** = -1.12
  - 33 -> **(33-38.8)/11.4** = -0.51
  - 65 -> **(65-38.8)/11.4** = 2.3 and so on



Z Score & Probability

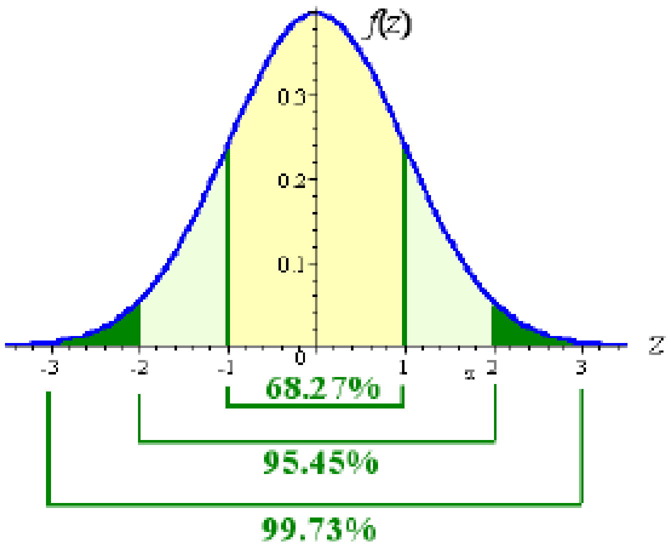
- $X \sim N(\mu, \sigma 6) \Rightarrow Z = (X - \mu)/\sigma$

  - for x1 and x2 belonging to X, we have corresponding z1 and z2 (belonging to Z)

  - Area under curve between X=x1 and X=x2 is same as that under Z for Z=z1 and Z=z2

  - P(x1<=X<=x2) = P(z1<=Z<=z2)

  - -1 <= Z <= 1 => 68%
  - -2 <= Z <= 2 => 95%
  - -3 <= Z <= 3 => 99.7%



PMF (Probability Mass Function) vs PDF (Probability Density Function)

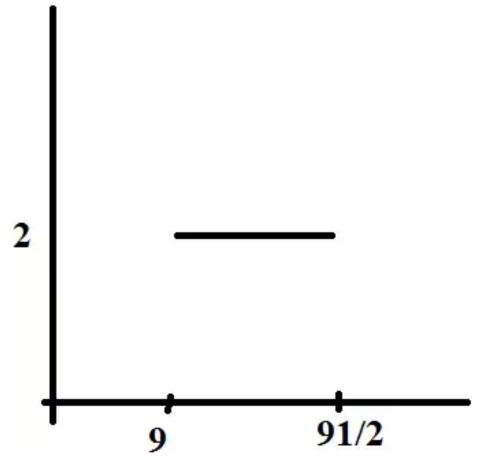
Consider **rolling a dice.** First let's define the Sample space: Sample space is the set of all possible outcomes. So in the case of rolling a dice once, the sample space is {1,2,3,4,5,6}.

Now what is the **probability** of getting 1? well, 1/6 (assuming unbiased dice).

List all the outcomes/events and their corresponding probabilities. That's **PMF** (Probability Mass Function). 1-1/6, 2-1/6, 3-1/6, 4-1/6, 5-1/6, 6-1/6. This is a function (mapping) from the events to the probability values.

PDF is analogous to **PMF**, but it's defined for continuous random variables. For example, if I say that I will come between 9pm and 9.30pm today, but expect me at any time in between with equal probability, then the pdf corresponding to that random variable is:

Here the range of pdf of any value between 9 and 91/2 is 2. But we are not interested in that value, rather we want to answer questions like what is the probability that I will come between 9 and 9.15. Then you can find the area under the curve between 9 and 91/4, which is 1/2.



Gaussian For Blind date

Julie is a student, and her best friend keeps trying to get her fixed up on blind dates in the hope that she'll find that special someone.

- She want to date only tall guys

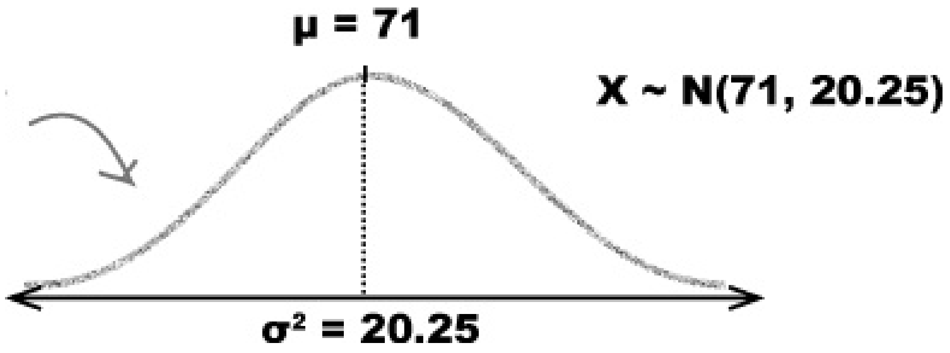
Oh! By the way Julie is 64" tall.

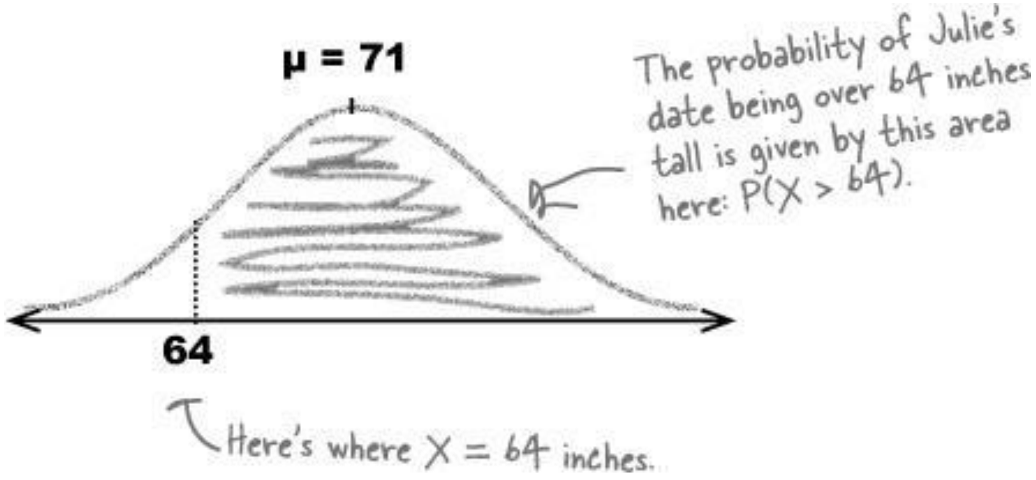
Probability Distribution

- Step 1: Determine the distribution

Julie wants to marry a person taller than her and is going on blind dates. The mean height of the 'available' guys is 71" and the variance is 20.25 inch^2.

Julie is 64" tall.

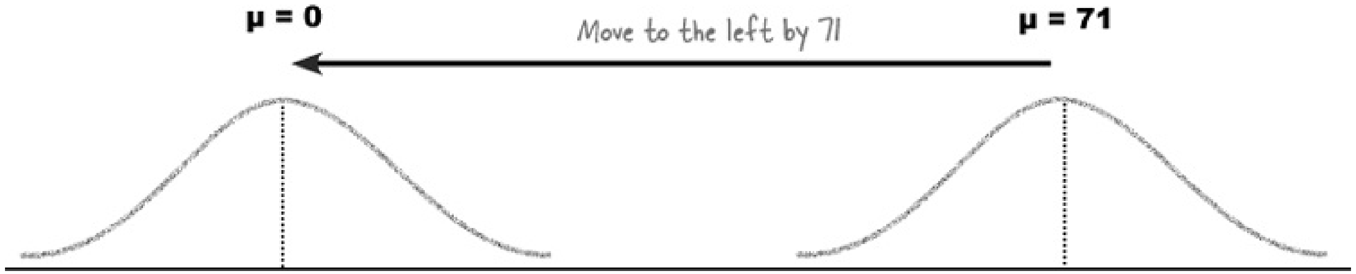


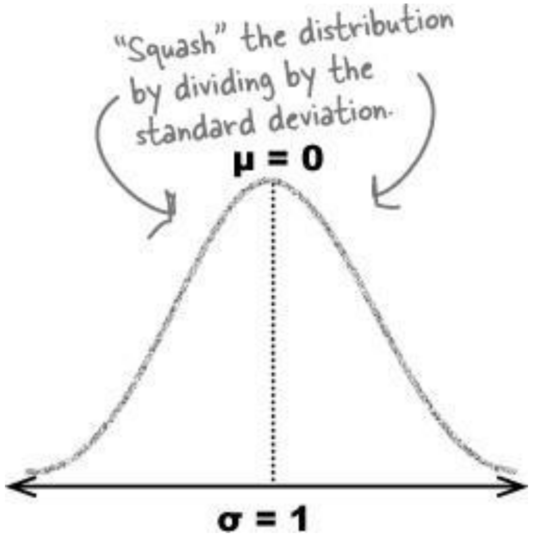


Calculating Normal Probabilities

- Step 2: Standardize to Z ~ N(0,1)

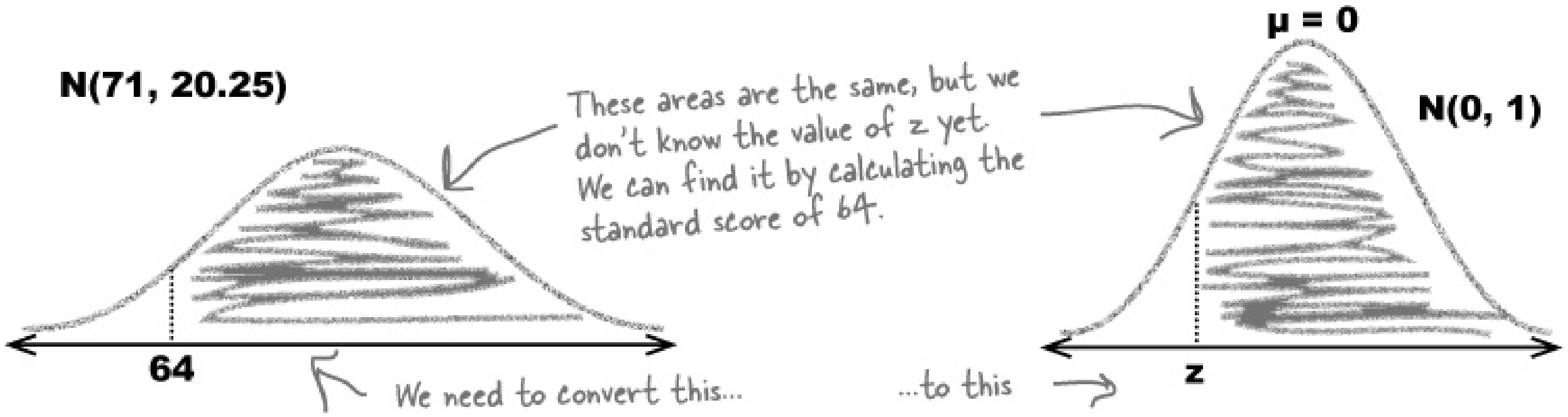
1. Move the mean This gives a new distribution X-71 ~ N(0,20.25)





Calculating Normal Probabilities

- Step 2: Standardize to Z ~ N(0,1)



Calculating Normal Probabilities

- Step 3: Look up the
- probability in the tables

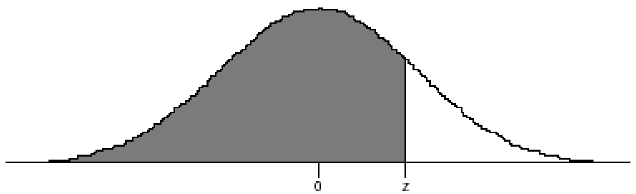
Note the tables give P(Z<z)

$$Z = \frac{X - 71}{4.5} = -1.56$$

for height 64”

$$P(Z > -1.56) = 1 - P(Z < -1.56)$$
$$= 1 - 0.0594$$
$$= 0.9406$$

Z-table



Normal Deviate z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-4.0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.9	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.8	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.7	.0001	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379

Attention Check

Find the probability **Julie** finds a man in the between 66” to 76”.

Z-Score @ 66” = (66 – 71)/4.5 = - 1.11

P(Z = -1.11) = 0.1131

Z-Score @ 76” = (76 - 71)/4.5 = 1.11

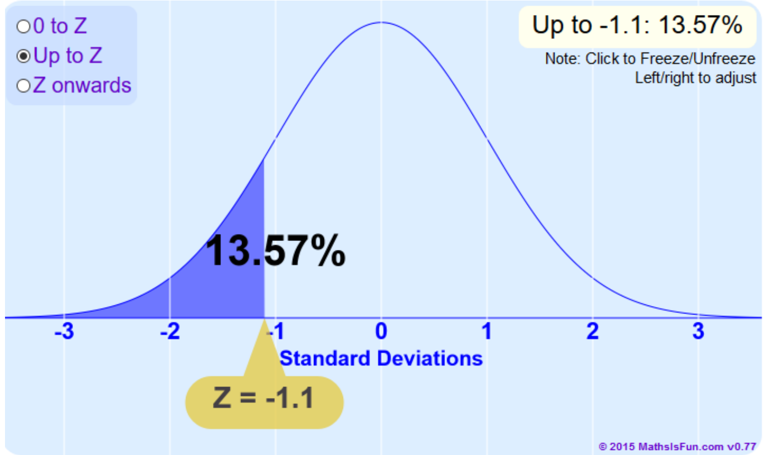
P(Z = 1.11) = 0.8665

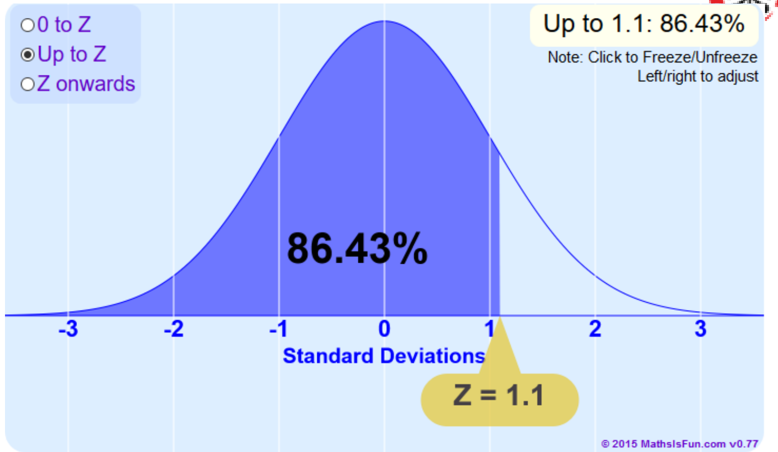
P(66” < X < 76”) = P(X = 76”) – P(X = 66”)

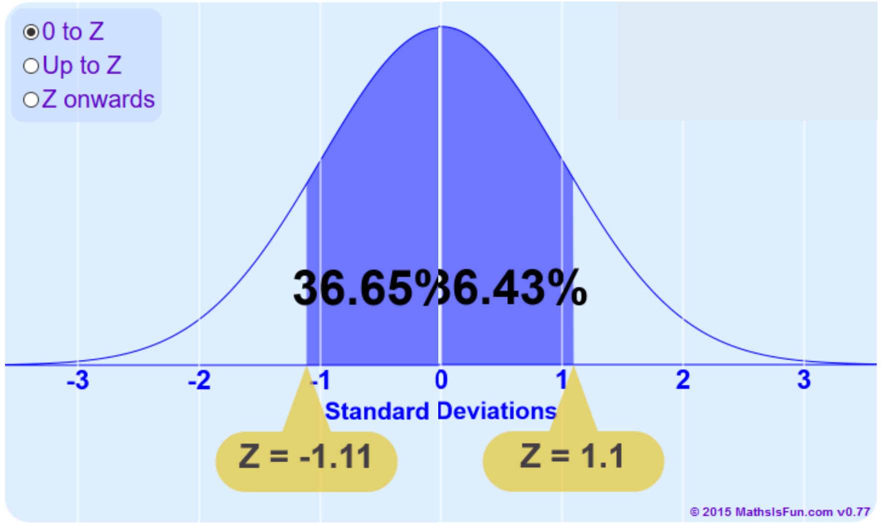
= 0.8665 – 0.1131

= 0.7534

Z-table







$P(-1.11 < Z < 1.11)$

$= P(Z = 1.11) – P(Z = -1.11)$

$= 0.8665 – 0.1131$

$= 0.7534$

THANK YOU