Probability Theory

Probability is likelyhood of an event occuring.
Mathematically it is expressed as,

$$P(\text{event}) = 0.3 = 30\%$$

Probability is given as [0, 0.5 .., 1]

0 - Impossible
1 = 100% - Definitely happen

Mathematically probability is not described in percentages but in decimal values.

$0.3 = 30\%$
$0.12 = 10.2\%$ . . . etc.

* Probability vs Proportion

Eg → Tossing of a coin 10 times.

Sample Space = [H, T]

$$P(H) = \frac{1}{2} = 0.5$$

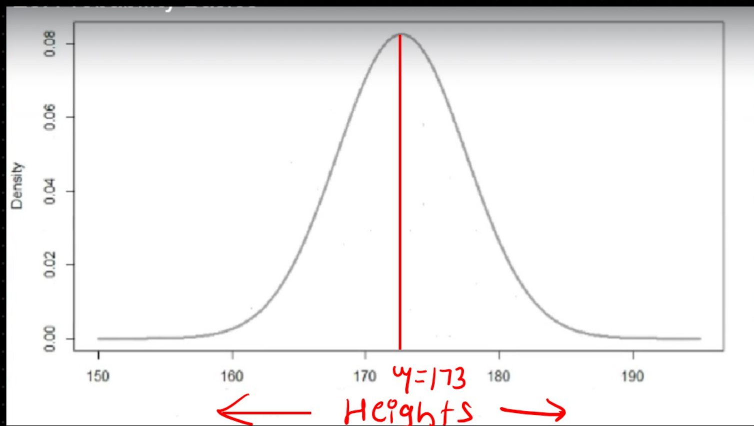
$\therefore$ $P(H) = 0.5$ ← This is probability

Now after tossing 10 times, Suppose we get H 7 times

$\therefore$ $\frac{7}{10} = 70\%$ ← This is proportion

Let us understand this with a practical eg:-

Given is a normal distⁿ of ht of people. Out of 19 people selected randomly what is the prob that their height will be greater than 173.?



Solⁿ → Since its normal distⁿ & $\mu = 173$

$\therefore \; P(x > 173) = \underline{0.5}$ ← This theoretical value of prob

But out of 19 random people, suppose 7 have height greater than 173cm.

$\therefore \; \frac{7}{19} = 37\%$ ← This is proportion of a random sample.

* Addition Rule

Suppose you want to calc prob of drawing a Queen from deck of cards.

Total cards = 52

Now there is a Queen in all 4 categories

$$\therefore P(Q) = P(Q_1) + P(Q_2) + P(Q_3) + P(Q_4)$$
$$= \frac{1}{52} + \frac{1}{52} + \frac{1}{52} + \frac{1}{52}$$
$$= \frac{4}{52}$$

Some points to consider:-

Suppose two fair coin are tossed 4 times :-
what is prob that there will be tails atleast once:-

H H
H T
T H
T T

$\} = 3$ times

$\therefore P(x) = \frac{3}{4} = 0.75 / 75\%$

Here each time two coins are tossed is an indvidual event, therefore we cannot add probabilities of individual events to know if we will get at least tails, bc we need add counts where Tails appears once.

However, we can use the add^n rule this way,

H, H → event 1

T, H → event 2 (T_1)

H, T → event 3 (T_2) } At least Once

Here T occurs twice hence an overlapping event

T, T → event 4 (T_3) ← T_1 and T_2

$P(T_1 \text{ or } T_2) = P(T_1) + P(T_2) - \underline{P(T_1 \,\&\, T_2)}$ ← overlapping event

$P(T_1 \cup T_2) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4} = 0.75 / 75\%$

$\therefore \underline{P(T_1 \text{ or } T_2) = 0.75}$

Mathematically it can be expressed as

$\underline{P(T_1 \cup T_2) = 0.75}$

{ or $\cup$ union
and $\cap$ intersection }

T_1 and T_2
This the section we substracted

Practice 2:

You work at a retail chain as an analyst.
While analyzing the sales, you found out that 30% of the customers included clothes in their purchase and 50% of the customers included food in their purchase. In addition, you found that 60% of all purchases included clothing or food.
What is the probability of food and clothing being included in a purchase?

$P(C) = 0.3$

$P(F) = 0.5$

$P(C \cup F) = 0.6$

$P(C \cap F) = ?$

$$P(C \cup F) = P(F) + P(C) - P(C \cap F)$$

$$\therefore P(C \cap F) = P(C \cup F) + P(F) + P(C)$$

$$= 0.3 + 0.5 - 0.6$$

$$\therefore P(C \cap F) = 0.2$$

* Multiplication Rule

First let's understand "Independent" & "Dependent" events
Two fair coins are tossed. and we see if there comes heads

Event A is heads comes in first coin

$P(A) = \frac{1}{2}$ ... {Independent of event B}

Event B is heads comes in second coin

$P(B) = \frac{1}{2}$ .... {Independent of event A}

Now probability of these independent events happening together is,

$$P(A \text{ and } B) = P(A) \times P(B)$$

$$= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

By old rule also,

T T $P(A \text{ and } B) = \frac{1}{4}$

T H

H T

A → H H ← B (Happens once when coin tossed 4 times)

Now Dependent events

Lets calculate probability of drawing a queen when we take 2 cards sequentially from deck of cards.

$$P(A) = \frac{1}{52}$$

Now when we successfully draw a queen one queen becomes less from 4 queens i.e 3 queen remain and no. of cards becom 51,
But if we do not draw a queen then no of queens will be 4 But no. of cards become 51.
Therefore event B will be dependent on outcome of event A

$$P(A) = \frac{4}{52}$$

if P(A) is success → ↙ ↘ ← if P(A) is failure

$$P(B) = \frac{3}{51} \qquad \frac{4}{51}$$

$$P(A \text{ and } B) = P(A) \cdot \underline{P(B|A)}$$

{ P(B|A) denotes P(B) given that P(A) is successful event. }

P(B|A) is "Conditional Probability"

In a container there are 7 black balls and 3 white balls. A person draws a ball from the container and does not put it back. Then the person draws another ball.
What is the probability that two black balls are drawn? Use the multiplication rule.

$$P(A) = \frac{7}{10} = 0.7 \qquad P(A \cap B) = 0.7 \cdot \frac{6}{9}$$

$$P(B|A) = \frac{6}{9} \qquad = 0.7 \times \frac{2}{3}$$

$$= \underline{0.66}$$

* Bayes Theorem

We know that,

$$P(A \cap B) = P(A) \cdot P(B|A)$$
$$P(B \cap A) = P(B) \cdot P(A|B)$$

According to Bayes Theorem,

$$\boxed{P(B|A) = \frac{P(B) \cdot P(A|B)}{P(A)}}$$

Let's solve a problem,

$B_1 \Rightarrow$ balls = 4, red = 2

$B_2 \Rightarrow$ balls = 6, red = 2

Experiment:
First we randomly (50% - 50%) select one bucket - then we draw a ball.

Question:
What is the probability of having selected the first bucket given we draw a red ball?

Event A $\Rightarrow$ Drawing Red Ball
Event B $\Rightarrow$ Selecting Bucket

we first select the bucket & then draw the ball, but the ques asks us the prob of selecting 1st bucket if we have red Ball.

so, we have to find $P(B|A)$

$P(A)$ = Prob of sel 1st bucket • Prob drawing red ball +
Prob of sel 2nd bucket • Prob drawing red ball

$$P(A) = \frac{1}{2} \times \frac{2}{4} + \frac{1}{2} \times \frac{2}{6} = \frac{5}{12}, \quad P(B) = \frac{1}{2}$$

$$P(A|B) = \frac{2}{4} = \frac{1}{2}$$

Acc^n to Bayes Theorem

$$P(B|A) = \frac{P(B) \cdot P(A|B)}{P(A)} = \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{5}{12}} = \underline{\underline{0.6}}$$

* Expected Value (Important)

Let's consider an experiment 'X'

experiment	outcomes	
X	$x_1 = 1$	$x_2 = 0$
	P_1	P_2

Expected Value is given by,

$$\boxed{E(x) = \sum_{i=1}^{n} P_i \cdot x_i}$$

$$E(x) = P_1 x_1 + P_2 x_2 \ldots P_n x_n$$

Question:

Maria plays a game that has two possible outcomes. If wins, she wins \$500. If she loses, she loses \$100. The chance of winning is 20%. What is the expected value of profit per game?

$x_1 = 500$, $x_2 = (-100)$ $P_1 = 0.2$ $P_2 = 100 - 0.2 = 0.8$

$$\begin{aligned} E(x) &= P_1 x_1 + P_2 x_2 + \ldots P_n x_n \\ &= 0.2 \times 500 + 0.8 \times (-100) \\ &= 100 - 80 \end{aligned}$$

$E(x) = \$20$

* Law of Large Numbers (Gamblers fallacy)

Law of Large Numbers states that if we repeat our experiment n times and we take mean of those experiments, then eventually as n reaches towards infinity we come closer to our expected value.

Suppose for X experiment expected value is $E(X)$

$$\bar{X}_n = \frac{X_1 + X_2 + \ldots X_n}{n}$$

as $n \to \infty$

$$\boxed{\bar{X}_\infty = E(X)}$$

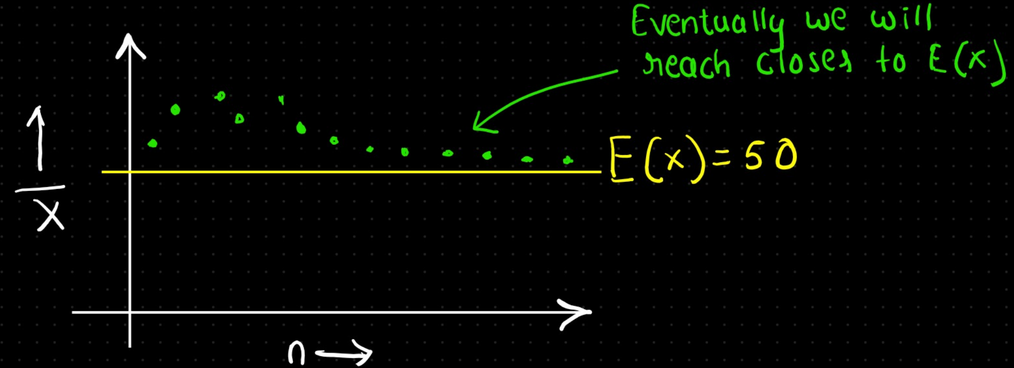
Let's toss a coin 100 times and check for heads

$$E(X) = \sum_{i=1}^{n} P(head)_i \times x_i$$

$$E(X) = 0.5 \times 100 = 50$$

Now this X is repeated infinity times, we can check this on graph

$$\bar{X} = \frac{X_1 + X_2 + X_n}{n} =$$



* Central Limit Theorem

Central limit theorem states that sample means of a given experiment approach to a normal distribution, even if the original experiment is not normally distributed.

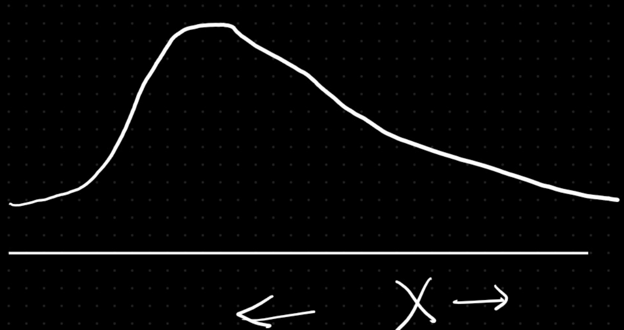
for experiment X, $S_1, S_2, S_3, \ldots S_m$ are samples of each result

$$S_1 = \{x_1, x_2 \ldots x_n\} = \overline{x_1}$$
$$S_2 = \{ \quad \} = \overline{x_2}$$
$$S_3 = \{ \quad \} = \overline{x_3}$$
$$S_m = \{ \quad \} = \overline{x_m}$$

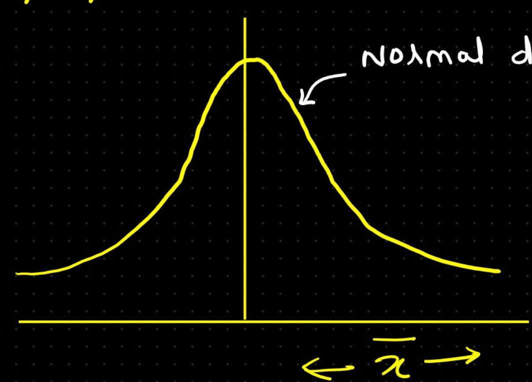
{ As m increases it gets more closer to normal distribution }

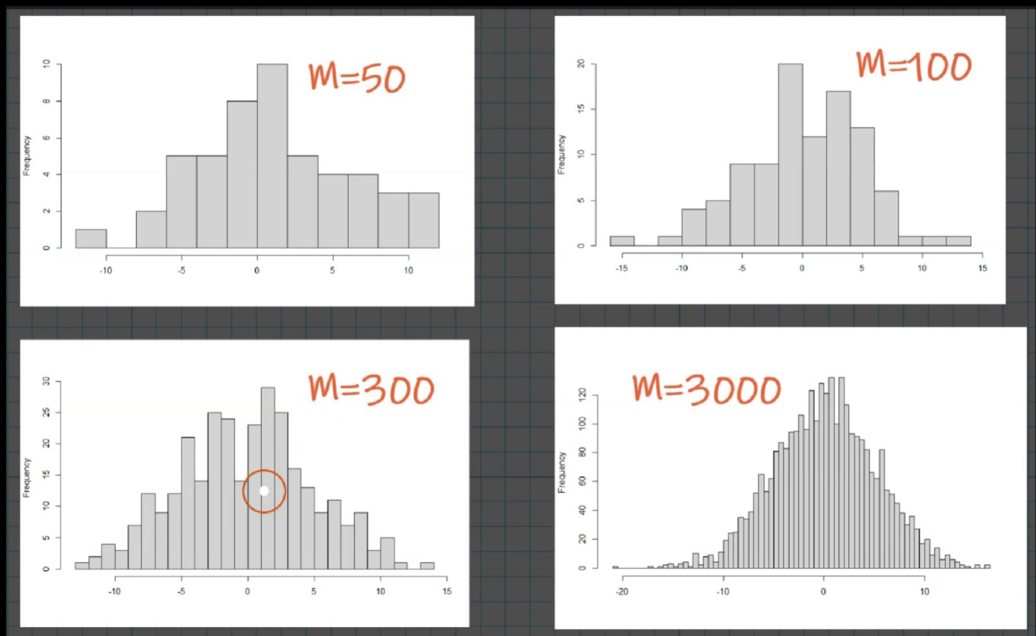
If we plot these,

for results of X



for $\overline{x}$





$\mu_s = \mu$

$\sigma_s = \dfrac{\sigma}{\sqrt{n}}$ ← sample size

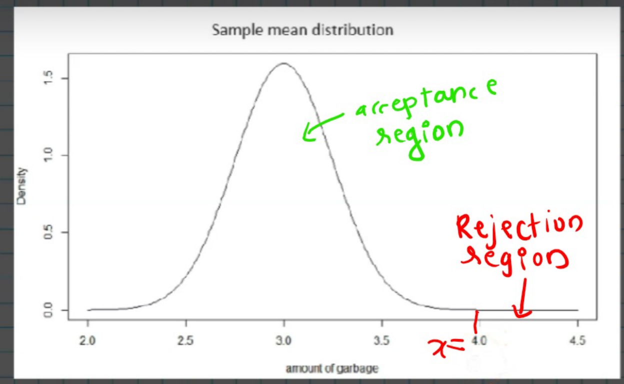
You are a Data Scientist in a waste management company. A survey found out that the an American household generates on average 3kg (6.6 pounds) of garbage per month with a standard deviation of 5 kg (11 pounds).

We have 100 customers and our business model is structured in a way that we make a profit of $5000 if we don't exceed a total of 400kg (8818.5 pounds) but make a loss of $10000 if we exceed 400kg.

What is the chance that we will make a loss and is our business model profitable?

Sol^n → $\mu_s = \mu = 3, \quad \sigma = 5, \quad n = 100,$

$$\sigma_s = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{100}} = 0.5, \quad x = \frac{400}{100} = 4$$



$$Z_{score} = \frac{x - \mu_s}{\sigma_s} = \frac{4-3}{0.5} = 2$$

from z table for Z = 2

Area under curve = $\underline{0.9772}$

$P(x<4) = 97.72\%$

Now chance of loss = $P(x>4)\, 1 - 0.9772$

$= 0.0228 = \underline{2.28\%}$

Let's calculate expected value of Profit (Y)

$E(Y) = P(x<4) \times \text{Profit} + P(x>4) \times \text{loss}$

$= 0.9772 \times 5000 + 0.0228 \times (-10000)$

$= \underline{4655}$

Since E(Y) value is +ve our business model is profitable.

Assume that you are a Data Scientist in an investment banking company. The company has 25 traders in average one trader makes a profit of $5000 per week with a s.d. of $20000.
So we want to extend our precautions against not having enough balance.
Every week our accounts are rebalanced to certain amount of money.

What is the chance that we will make a loss?

How much money does the company have to have to have at least in balance every week to not run out of cash with a chance of 99%?

$\sigma = 20000, \quad \mu = 5000, \quad n = 25, \quad x = 0$ {>0 if trader makes money, <0 then there will be loss}

$$\sigma_s = \frac{\sigma}{\sqrt{n}} = \frac{20000}{\sqrt{25}} = \frac{\cancel{20000}^{4000}}{\cancel{5}} = 4000$$

$$z = \frac{x - \mu_s}{\sigma_s} = \frac{0 - 5000}{4000} = -1.25$$

from z table area = 0.1056 = 10.56%

So there is 10.56% chance that we will loose money

Now we need to calc prob that our company doesn't make loss more than 1%

So we need to calculate x from z score and zscore from z table

from z table $z_{score} = -2.33$

$$\therefore -2.33 = \frac{x - 5000}{4000} \quad \therefore x = -4320 \text{ is avg loss of 1 trader}$$

for 25 traders $= 25 \times -4320 = \underline{-108{,}000}$

So to avoid running out of cash we must have 108,000 USD

* Binomial Distribution

Binomial Coefficient

this is pronounced as n to's K

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

Bernoulli Trial :-

Only two possible outcomes, labelled as success and failure (e.g coin flip). 'p' probability of success.

If there are number of experiments then they need to be independent.

$$P(\text{success}) = p$$
$$P(\text{failure}) = 1-p$$

$$P(X=k) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

$k = 0, 1, 2, 3, \ldots n$

X represents no of successes in n independent trials

$$\mu = E(X) = np$$

$$\sigma^2 = np(1-p)$$

E.g →

We roll a dice 3 times.

What is the probability a 2 comes up exactly twice?

soln →

Success: Rolling a 2.

Failure: Rolling any other number apart from 2.

X binomial:

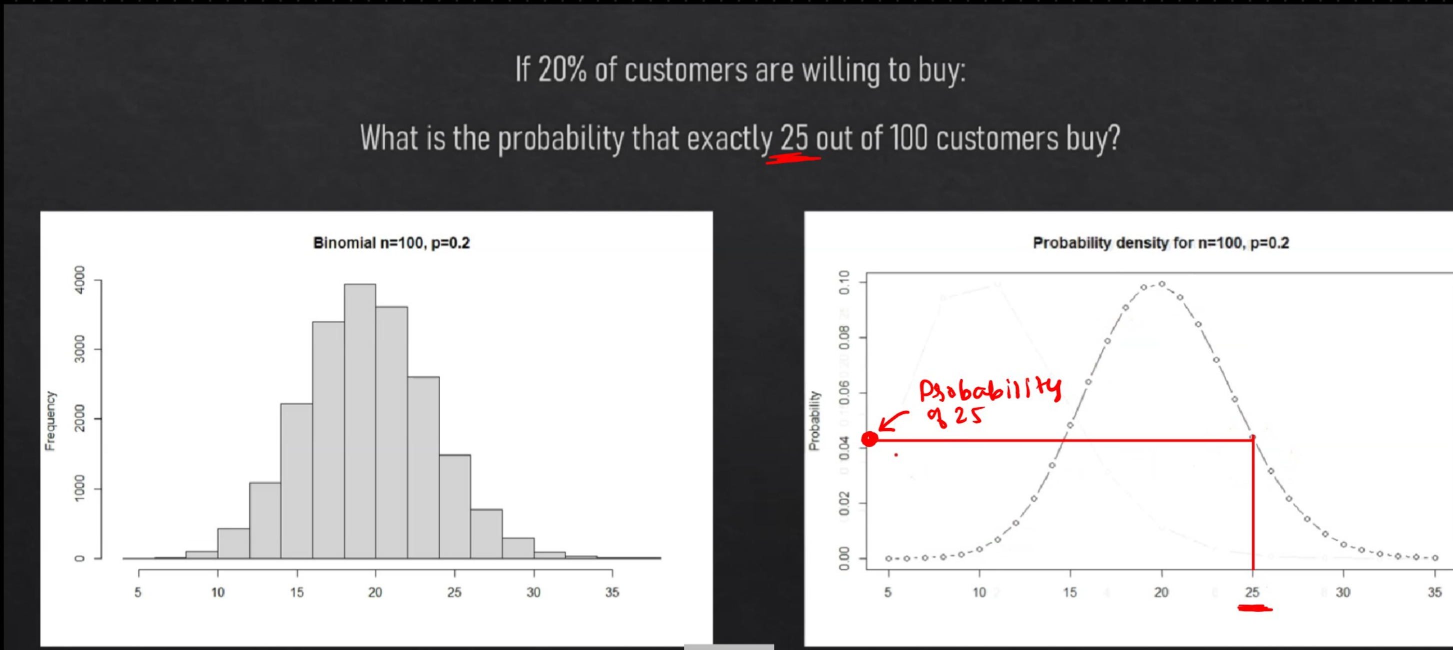
p = 1/6

n = 3

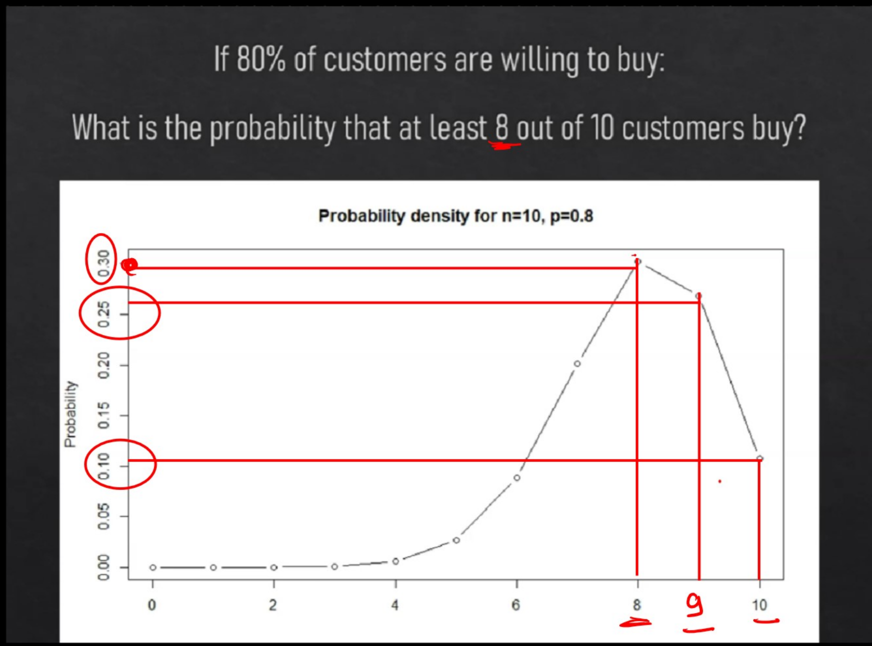
k = 2

$$P(X=2) = \binom{3}{2}\frac{1}{6}^2\left(1-\frac{1}{6}\right)^{3-2}$$

$= 0.069$



As no of experiments increase the binomial $dist^n$ looks like a 'normal distribution'.



$$P(8) + P(9) + P(10) = \underline{0.677}$$

the graph is just representing the results for all the probabilities.

So P(X=1) (probability for one success), P(X=2) (probability for two success) etc.

If we want to the probability that at least 8 customers buy we need to calculate all the single probablities, so what is the probabability that exactly 8 customers buy, what is the probability that exactly 9 customers buy and what is the probability that exactly 10 customers buy.

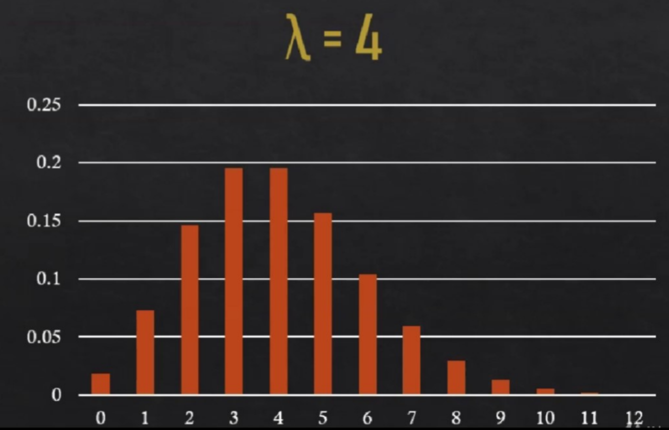
And then we need to add them up to get the probability that either 8 customers, 9 customer or 10 customers buy - in other words at least 8 customers (8, 9 or 10).

* Poissons Distribution

Poisson distribution

Discrete Distribution (whole numbers)

- Number of events occurring in a time interval or region of opportunity
- How many customers enter a store in one hour
- Only parameter λ (lambda) – expected value
- Possible values reaching from 0 to ∞
- Rate at which events occurs is constant
- The events are independent



Density function (p.m.f.) (probability mass function)

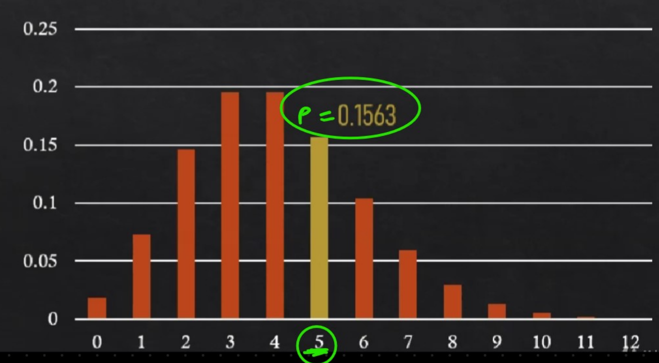
Value to check

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

$$P(X = 5) = \frac{e^{-4}4^5}{5!} = 0.1563$$

Pmf is used to calculate probability of exactly 5.

Expected val $E(X) = \lambda = 4$ Variance $V(X) = \lambda = 4$



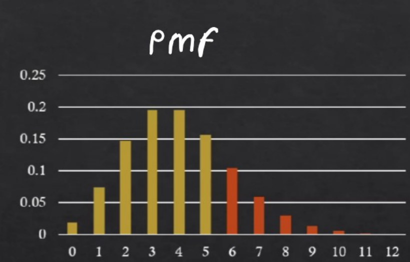
Cumulative density function (c.d.f.)

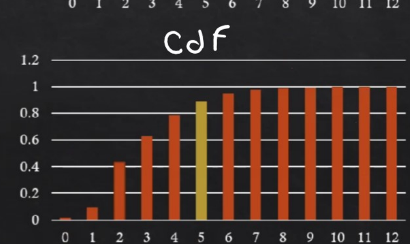
$$P(X \leq n) = e^{-\lambda} \sum_{k=0}^{n} \frac{\lambda^k}{k!}$$

cdf is use to calculate probability of not more than 5

= POISSON.DIST(5,4,TRUE) =0.785

for calculating in excel





Excel functions for Binomial & Poission Distⁿ

=BINOM.DIST(number of success, number of trials, cumulative or not)

=POISSON.DIST(number of successes, expected value, cumulative or not)

Example problems

You are working in a car company that is selling sports cars.

We know that in average there are 15 people coming into our show room per day.

Assuming the number of people entering our show room is Poisson distributed, what is the probability that exactly 20 people enter our room?

$\lambda = 15$

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

$$P(X=20) = \frac{e^{-15}\,15^{20}}{20!} = 0.0418$$

Example problems

We know 60% of our customers are men. We usually have 10 customers buying a car per day.

What is the probability that exactly 8 of our customers are men in a certain day?

$$P(X = k) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

$n = 10$

$k = 8$

$p = 0.6$

$$P(X=8) = \binom{10}{8} \cdot 0.6^8 (0.4)^2$$

$$\approx 0.1209$$