Interview Prep Session

[Probability & Statistics]

Recap through Questions

Q: Count the # times model will be trained, if we want to do a grid search with the following params!

→ Automated | EASY

```
params = {'depth': [3, 4, 5 3],
          'n_est': [10, 100, 1000]
          'min_samples': range(1,5)]
```

a) 10 b) 36 c) 11 d) 45

Ans: 3 × 3 × 4 = **36**

Q: If 4 customers come in, 1 will buy on avg. Find the probability of selling at-least 1 unit, if 4 customers come in.

a) 1 b) 1/4 c) $\frac{175}{256}$ d) $\frac{81}{256}$

Ans → $P(\text{Buy}) = 1/4$

$P(\text{Don't buy}) = 1 - 1/4 = 3/4$

4 cust not buying $= \left(\frac{3}{4}\right)^4 = \frac{81}{256}$

At least 1 purchase $= 1 - \frac{81}{256} = \boxed{\frac{175}{256}}$

Q: I have a binary classification model that has 90% accuracy. Find the prob that it makes exactly 4 errors out of 20 predictions

→ Ans: 1 classification

↳ Bernoulli event

$$p = 1 - 0.9 = 0.1$$

20 predictions

↳ Binomial event

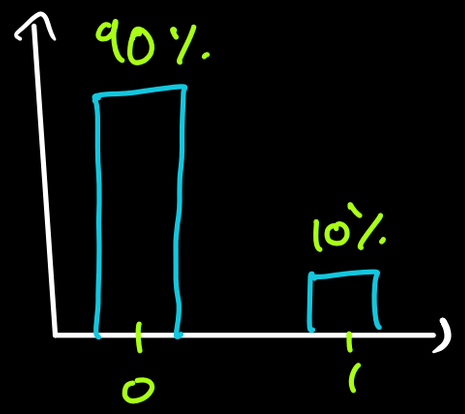
$$d = Bin(20, 0.1)$$

$$P(4) = d.\,pdf\,(4)$$

$$= {}^{20}C_4 \times (0.1)^4 \cdot (0.9)^{16}$$

Today Recap binomial

Q How to solve this??



Bernoulli dist of 1 trial.

But we have 20 trials!!

Q: What are the possible o/c in 20 trials?

→ 1 pass
→ 2, pass
} → we need atleast 4

→ 20 pass → 40 pass

Q! What is the prob of exactly 1 error?

Soln → Total possible combinations??

$\underset{1}{P/F}\ \underset{2}{P/F}\ \underset{3}{\_}\ \dots\ \underset{20}{P/F} = 2^{20}$

Number of ways in which 1 error

P F F F . . . → $^{20}C_1 = 20$ ways

$\underset{P}{(0.1)}\ \underset{F}{(0.7)}\ \underset{F}{(0.7)} \dots \underset{F}{(\ )}$

$P(1 \text{ pass}) = \frac{20}{2^{20}}$ → ✗ Events are not equally likely

Imagine getting HH with biased coin

$= p \times (1-p)^{19} = (0.1)(0.9)^{17} \times 20$

Q: Prob of getting 2 passes?

# of ways $= {}^{20}C_2 = \frac{20 \times 19}{2} = 190$

Prob $\rightarrow (0.1)^2 (0.9)^{18} \times 190$

Q: Notice a pattern yet?

Probability of getting r success in 'n' tricks

$$P(r_{succ}, n_{trial}) = {}^{n}C_r \, p^r (1-p)^{n-r}$$

Q: There are 3 servers behind a load balancer. Traffic is diverted to each server with given probability. The down time prob for each server is given. If one user received server not found error, what is the prob it was server 1.

P_ server	
S1	20%
S2	40%
S3	40%

P_ down	
S1	3%
S2	2%
S3	2%

a) 27% b) 35% c) 50% d) 10%

$P(A) = 0.2$ $\quad P(E1/A) = 0.03$

$P(B) = 0.4$ $\quad P(E2/B) = 0.02$

$P(C) = 0.4$ $\quad P(E3/C) = 0.02$

Bayes theorem

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

$$\therefore P(A/E1) = \frac{P(E1/A) \cdot P(A)}{P(E1/A) \cdot P(A) + P(E_2/B) \cdot P(B)}$$

$$= \frac{(0.03)(0.2)}{(0.03)(0.2) + (0.02)(0.7) + (0.02)(0.4)}$$

$$= \frac{6}{6 + 8 + 8} = \frac{6}{22} = \boxed{27\%}$$

All Equations

King . Truth of King . Life

$$P = \frac{\#\ \text{favourable outcomes}}{\#\ \text{total possible outcomes}}$$

$P(A') = 1 - P(A)$ # complementary

$P(A \cap B) = P(A) \cdot P(B)$ # mutually independent

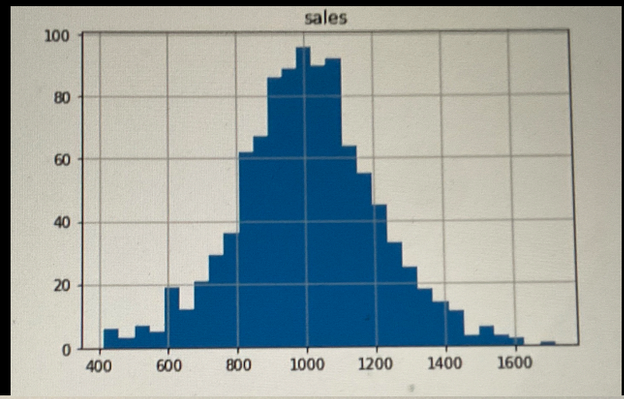
$P(A \cup B) = 1$ # exhaus $P(A \cap B) = 0$ # exclusive

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

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

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

Q. Toothpaste sales Retail →



Quiz: Which among these mostly likely represent the weekly sales data of toothpaste?

1. Bernoulli
2. Binomial
3. Gaussian
4. Geometric

• • •

Suppose the store has **beginning on-hand (BOH)** inventory of 1300. If the demand is more than the BOH, there is a need for stock **replenishment**. Which distribution best characterizes the need for stock replenishment?

Quiz: Which among these mostly likely represent need for replenishment?

1. Bernoulli
2. Binomial
3. Gaussian
4. Geometric

• • •

Suppose there are 2000 stores, each with BOH of 1300. The **distribution center (DC)** needs to calculate the number of stores which might need replenishment. Which distribution best characterizes this number?

Quiz: Suppose there are 2000 stores, each with BOH of 1300. The distribution center (DC) needs to calculate the number of stores which might need replenishment. Which distribution best characterizes this number?

1. Bernoulli
2. Binomial
3. Gaussian
4. Geometric

• • •

The manager of the DC sequentially calls a few store managers to see if they need replenishment. We need to monitor how many calls are needed till the first time a store manager asks for replenishment. Which distribution best charaterizes this?

Quiz: The manager of the DC sequentially calls a few store managers to see if they need replenishment. We need to monitor how many calls are needed till the first time a store manager asks for replenishment. Which distribution best charaterizes this?

1. Bernoulli
2. Binomial
3. Gaussian
4. Geometric

Q: What do you understand by p-value?

a) P(rejecting H_0)

b) P(rejecting H_A)

c) P(H_0 | data)

d) P(data | H_0)

Examples

→ Crime suspect

Null → H_0: Innocent (until proven guilty)

Alternate → H_A: Guilty

D: (Evidence / witness)

Data: 1) he has a knife
2) knife has blood stains
3) blood matches victims
4) No credible alibi.

$\underbrace{P(\text{Data} \mid H_0)}_{\text{p-value}} = ??$ Low

↳ Reject → Jail

→ Suppose you are the 3rd empire. On field empire has signalled "out"

H_0 → On field empire is correct

Data: Slow motion video + audio

If we have sufficient data, we can change on-field umpire decision

$P(Data \mid Out)$ =

- High → Fail to reject (out)
- Low → Reject (Not out)

Q: Which of the following is the best way to detect a normal distribution?

a) 68 – 95 – 99.7 rule

b) Q Q plot

c) KS test

d) area under curve = 1

Mathematically,

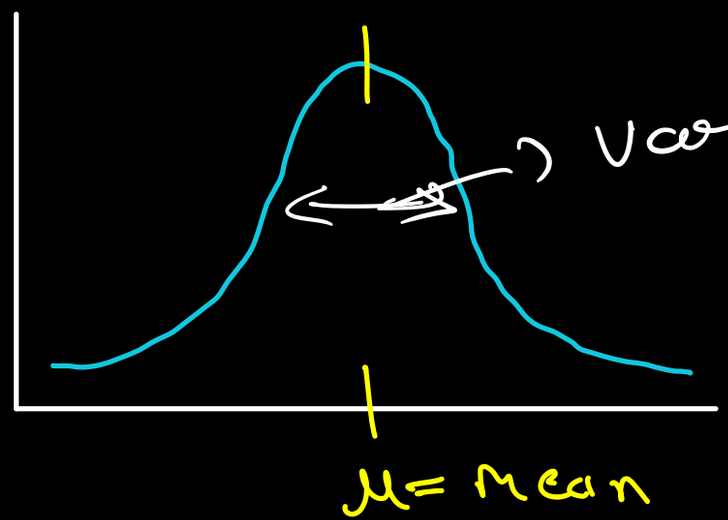
$$X \sim N(\mu, \sigma)$$

normal mean std

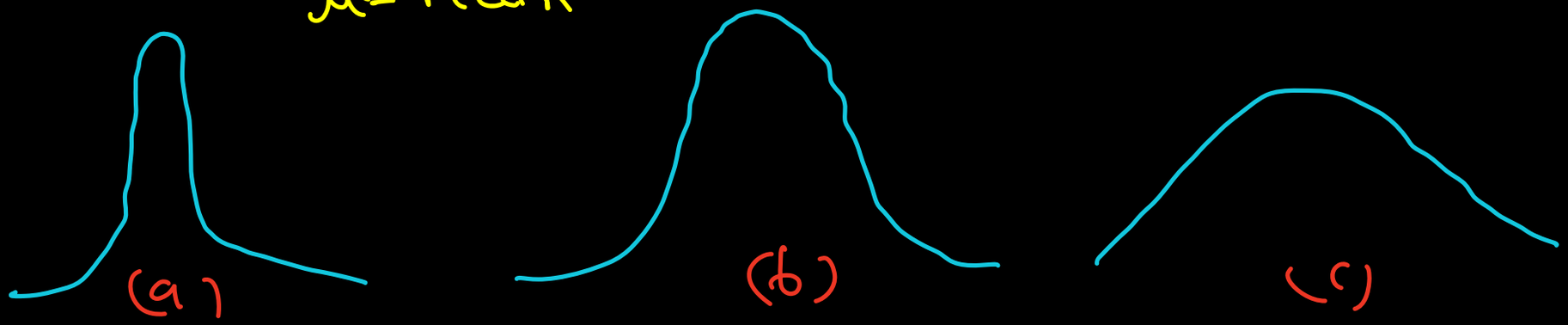
eqn

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

Not Imp



std (a) < std (b) < std (c)



Properties of Normal Dist

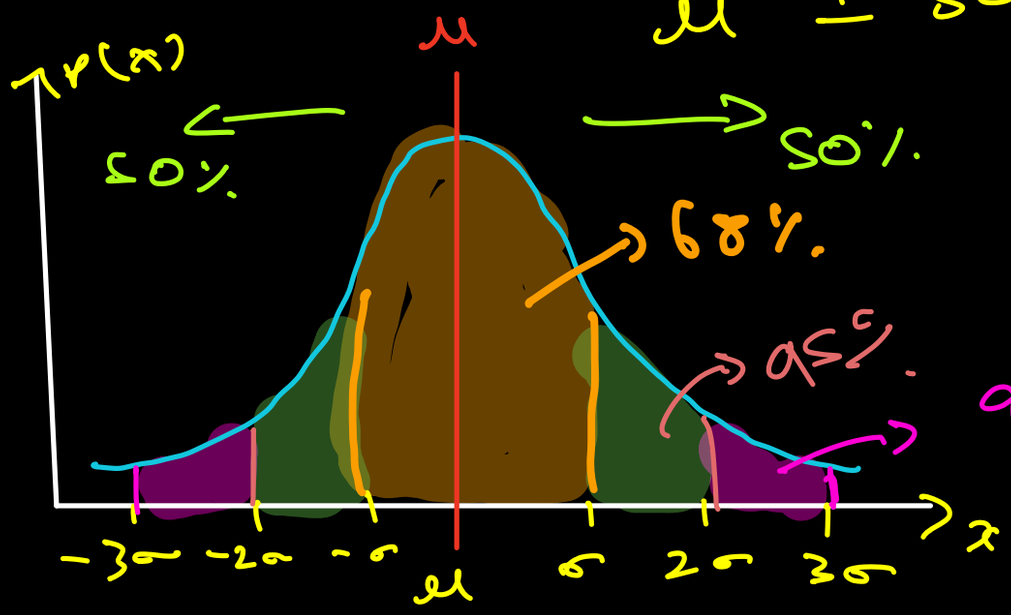
→ Symmetric

→ mean = μ , std = σ , var = σ^2

→ area with $\mu \pm \sigma = 68\%$

$\mu \pm 2\sigma = 95\%$

$\mu \pm 3\sigma = 99\%$



Q: How do calc the CI of the mean of a population if one sample is given to you?

Calc does sample ->

$\bar{x} = 12$

$s = 3$

$n = 100$

∴ 95% CI

$= 12 \pm \frac{2 \times 3}{\sqrt{100}} = \underline{11.4 \rightarrow 12.6}$

Q: Match test to application →

a) T	1 Gender affects churn
b) χ^2	2 all subjects are equally difficult
c) Z	3 Indian customers shoping profit = US
d) KS	4 $\mu_1 > c$ with 10 samples
e) F	5 $\mu_1 > \mu_2$

→ As: a) → 4
b) → 1
c) → 5
d) → 3
e → 2

Q: Diff b/w expectation and average

↳ calc the expectation of dice roll

Q: Historic avg monthly car sales are known to be 15000 units. Using last years sales data, can you design a Z-test to validate is the average has changed?

↳ No, because this sample is

not independent.

Q: Why do investors want to diversify their portfolio?

↳ the variance of the group is lower than the variance of individuals

↳ similar to what happens in bagging methods