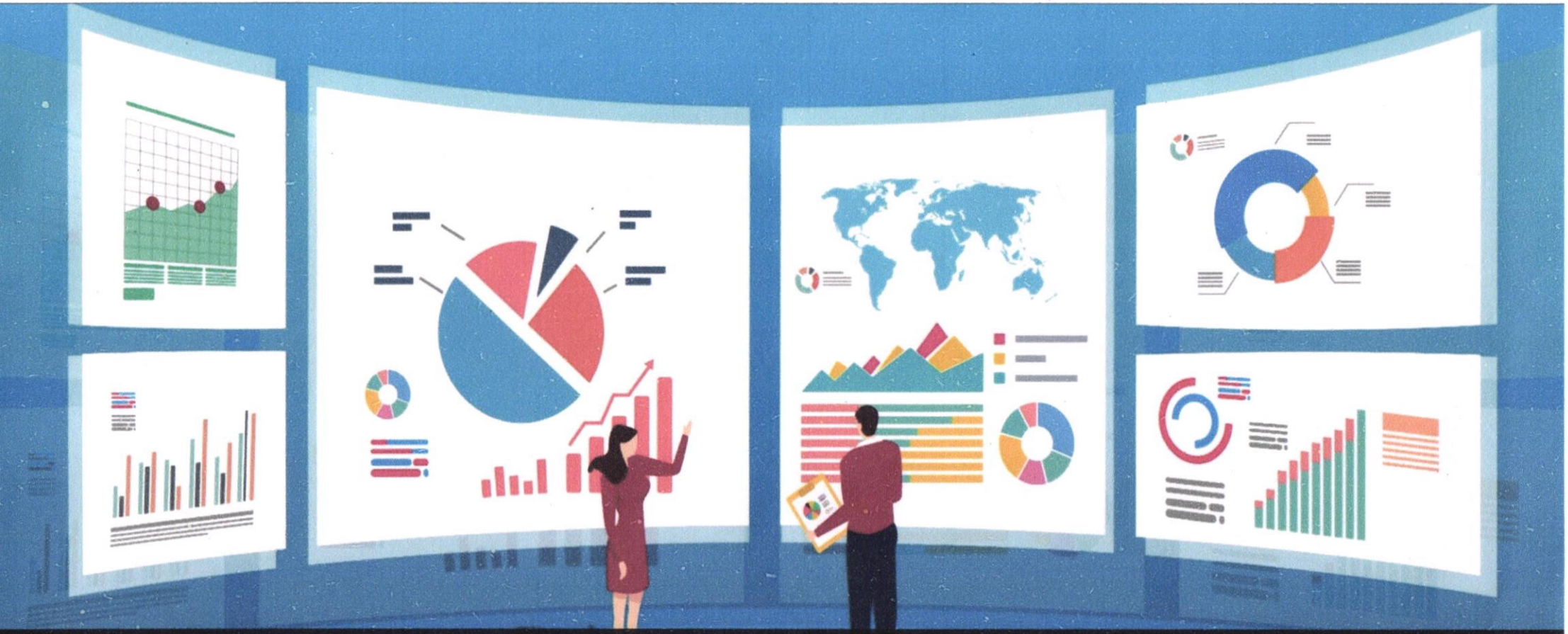
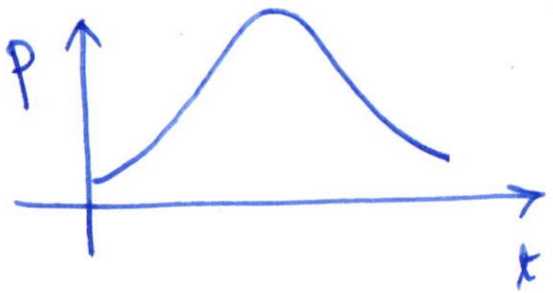


Inferential Statistics

INSAID

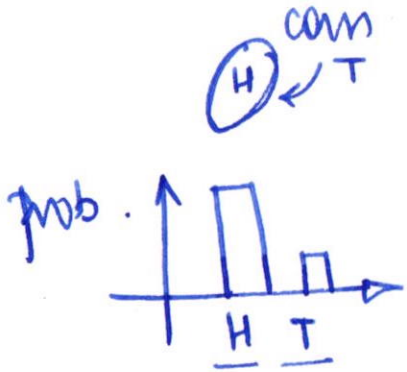


- amt of sales
- production o/p
- TRP — television rating points

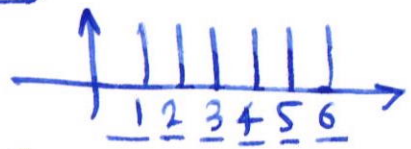
Probabilistic

Random variable

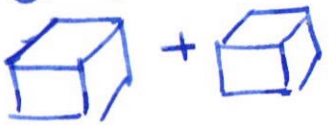
That variable whose value is unknown until outcome. But we know it's Probability distribution

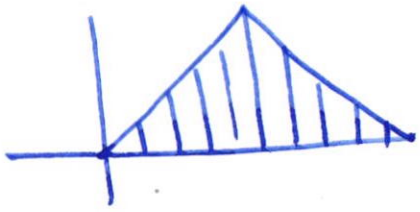


playing cards



category

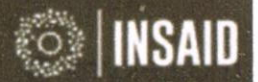


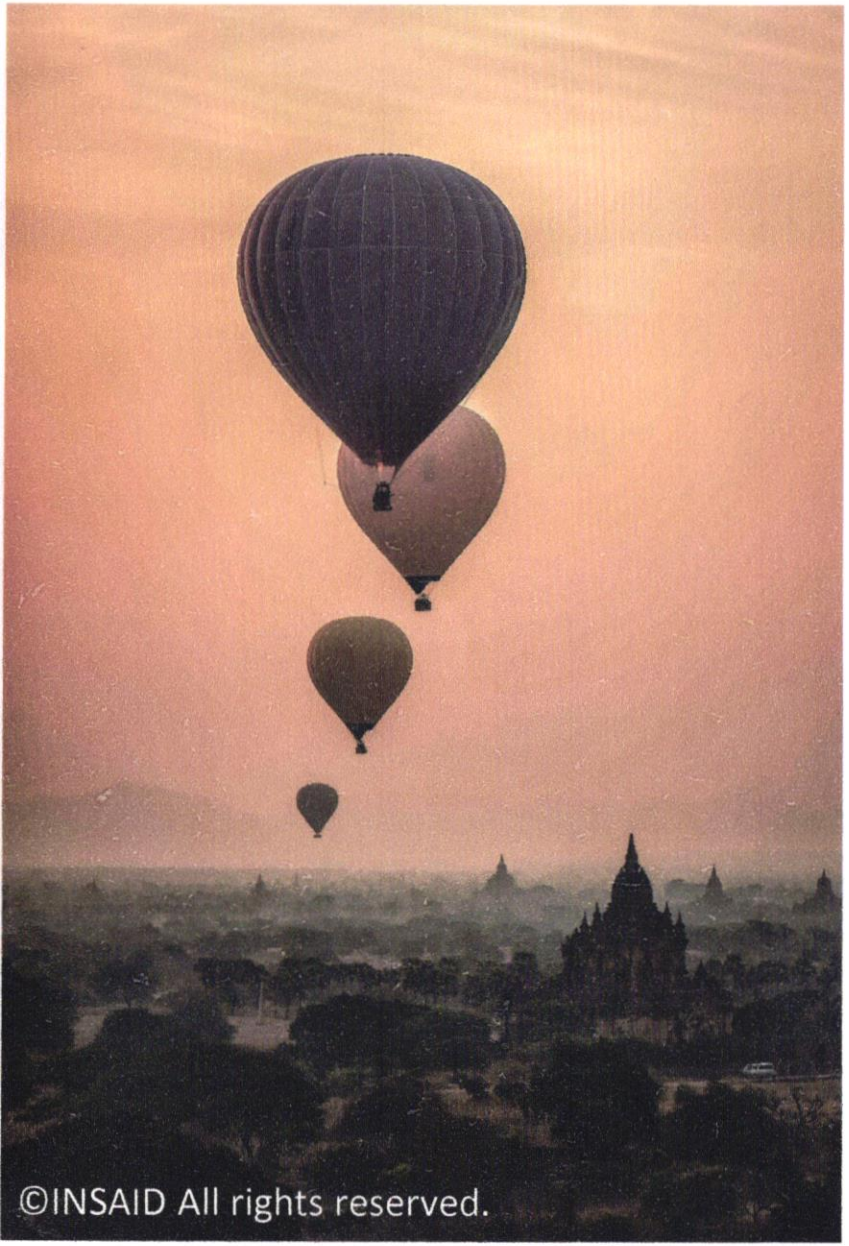


1/0

Agenda

- **Introduction to Probability**
- Basics of Probability
- Conditional Probability
- Random Variables
- Normal Distribution
- Hypothesis Testing





Contribution of Statistics to Data Science

- Infer population insights from sample statistics.
- Understand the underlying data.
- Draw inferences from data.
- Make predictions.

INSAID

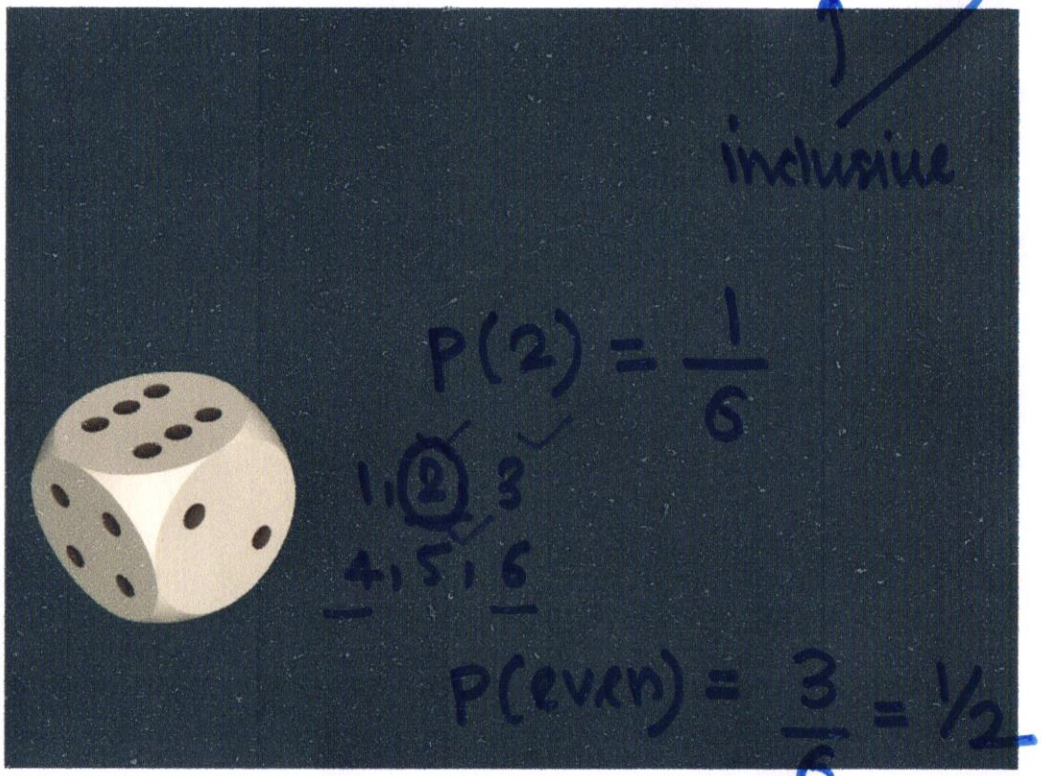
What is Probability?

- An **estimation** of how likely something is to happen.
- The **analysis of events** governed by probability is called statistics.
- **Example:** Chances to get a 6 on every roll of dice is always 1 / 6.

chances of occurring

$$\frac{\#\text{fav outcomes}}{\text{Total } \#\text{outcomes}}$$

[0, 1] inclusive

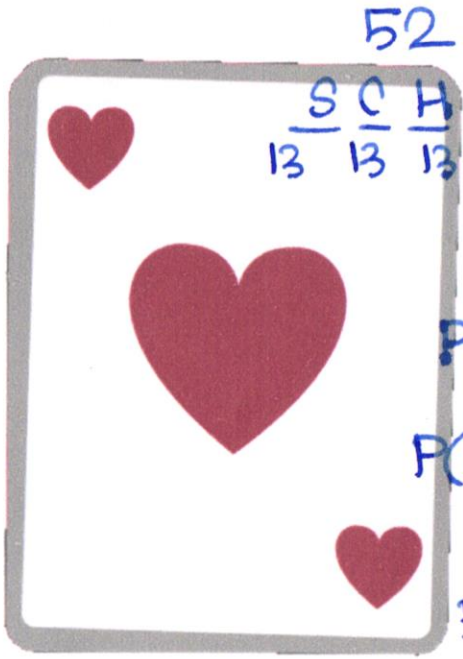


$$P(\text{prime}) = \frac{3}{6} = \frac{1}{2}$$

INSAID

Another Example

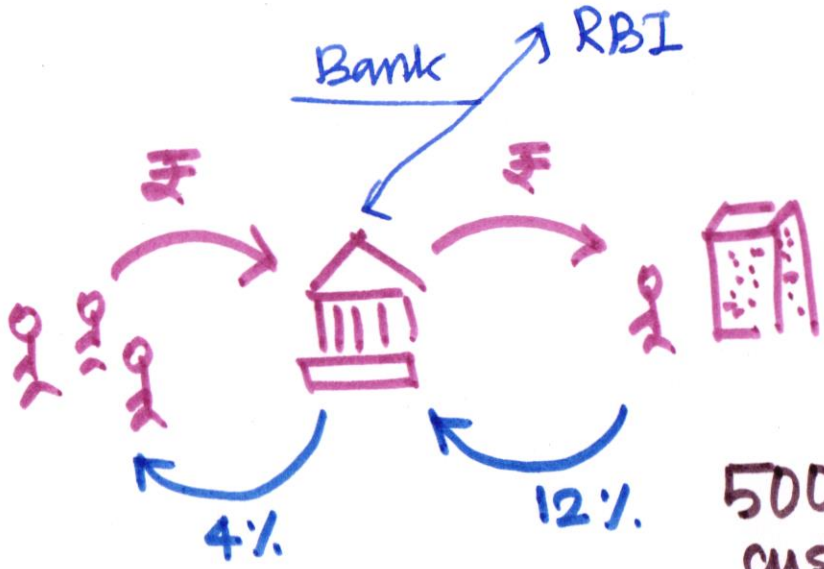
- What are the chances that you would get a King of Spade from a deck of 52 cards?
- **Solution:** 1 / 52.



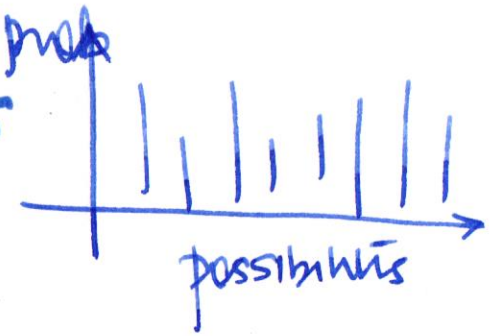
INSAID

Necessity of Probability

- To understand the nature of chance and variation in life.
- To be a well-informed or effective citizen.
- To understand the risk and relative risk.
- **Example:** If credit score is good there is 70% probability of providing a loan.



Random variables



P (covid +) ⟹ Average for the population

Conditional probability.

P (covid + | IT. with)

prob of being covid + given his/her IT profess. and with

P(default) ⟹ Aggregate prob.

P (default | CIBIL, income, ITR) ⟹ conditional prob for the person.

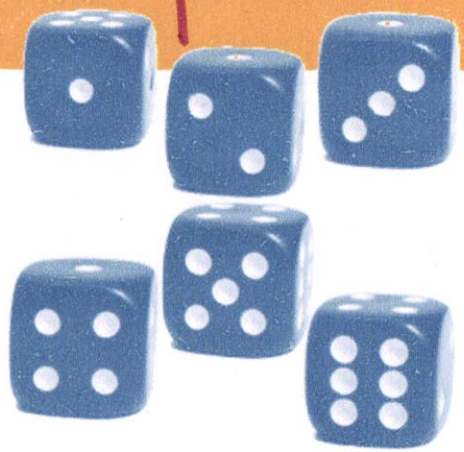
X axis

out-come

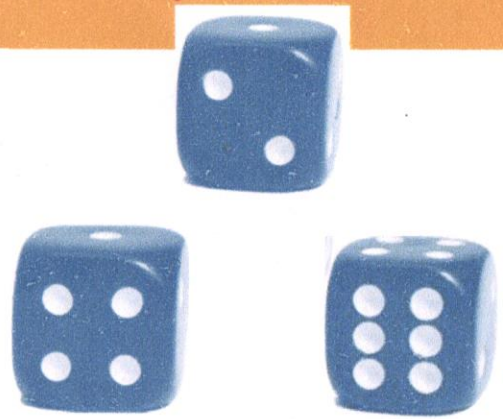
possibilities

favourable events.

(Sample Point)



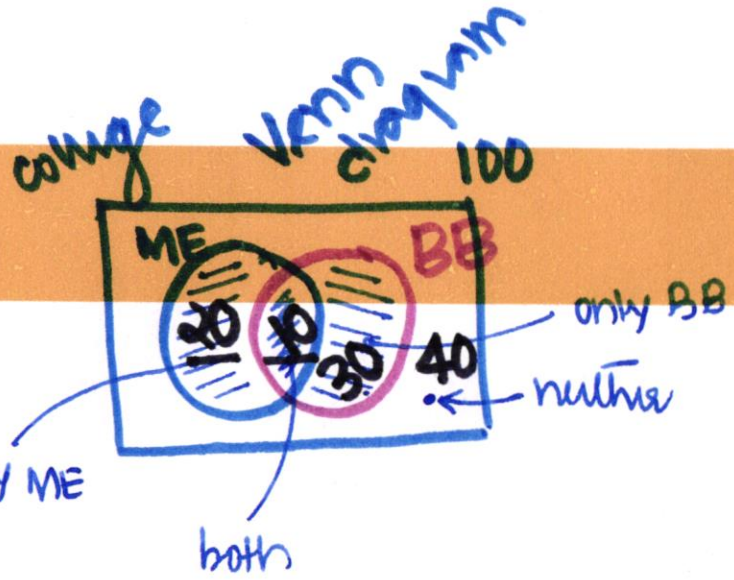
(Sample Space)



(Event)

Experimental Components

- **Sample Point:** Random value of sample space.
- **Sample Space:** Range of values of a random variable.
- **Event:** A set of outcomes of an experiment to which a probability is assigned.



'or'
'union' ∪

'and'
'intersection'
∩

$P(ME) = \frac{30}{100}$

$P(BB) = \frac{40}{100}$.

$P(\text{not } ME) \quad \frac{70}{100}$

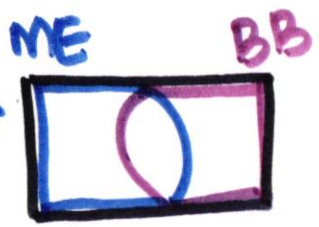
$P(ME \text{ and } BB) = \frac{10}{100}$

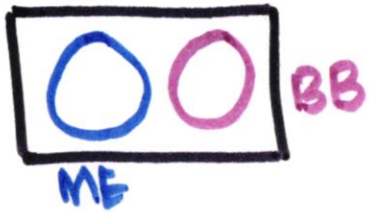
$P(ME \cap BB)$

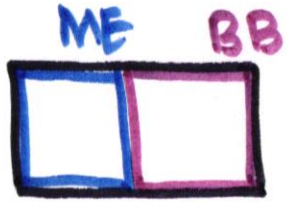
$P(ME \text{ or } BB) = \frac{60}{100}$

$P(ME \cup BB)$

Mutually Exclusive and Cumulatively Exhaustive

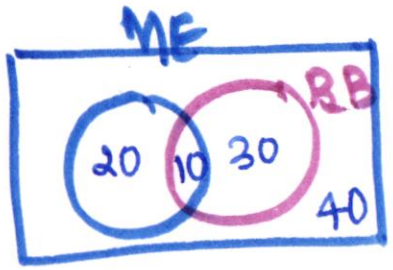






Both.
Mutually Exclusive
and
Cum. Exhaustive

INSAID

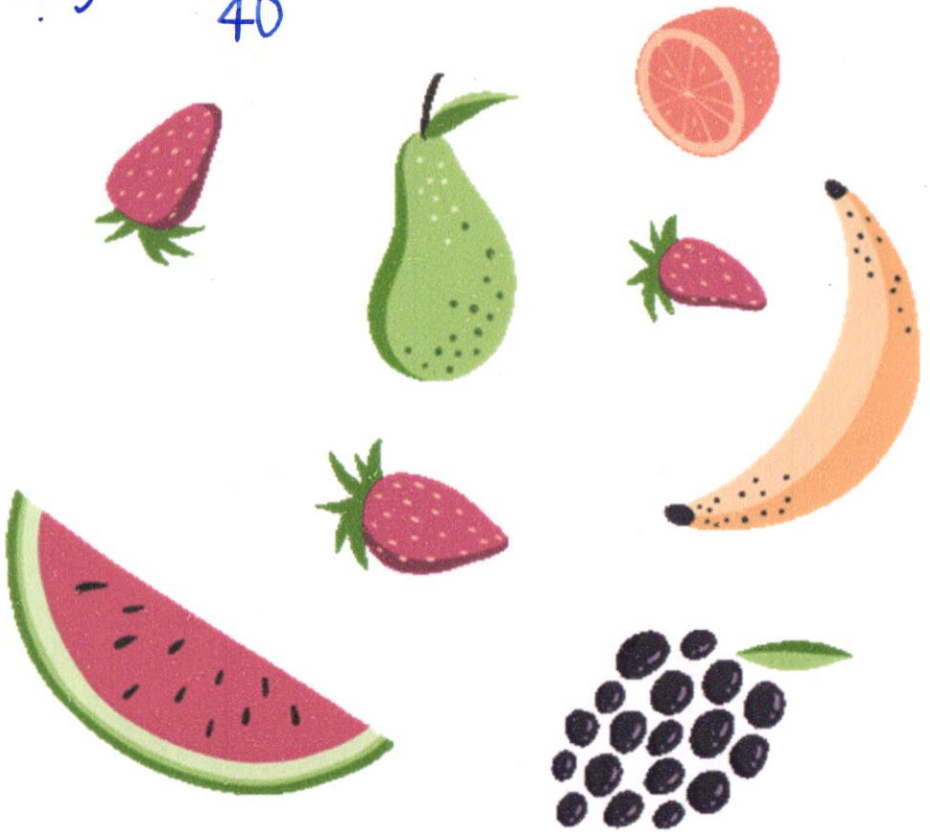


$$P(BB|ME) = \frac{10}{30}$$

$$P(ME|BB) = \frac{10}{40}$$

$$\frac{\#\,fav}{\text{Total }\#\,events}$$

Conditional Probability



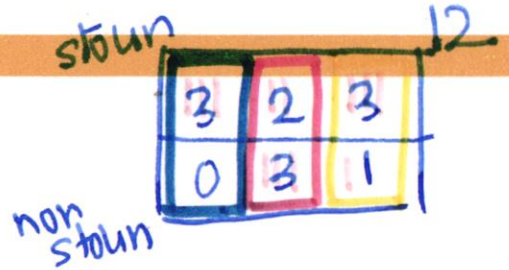
- A measure of the probability of an event occurring given that another event has occurred.
- **Example:** What is the probability of **buying strawberry** considering its condition (**Fresh** or **Not Fresh**)?

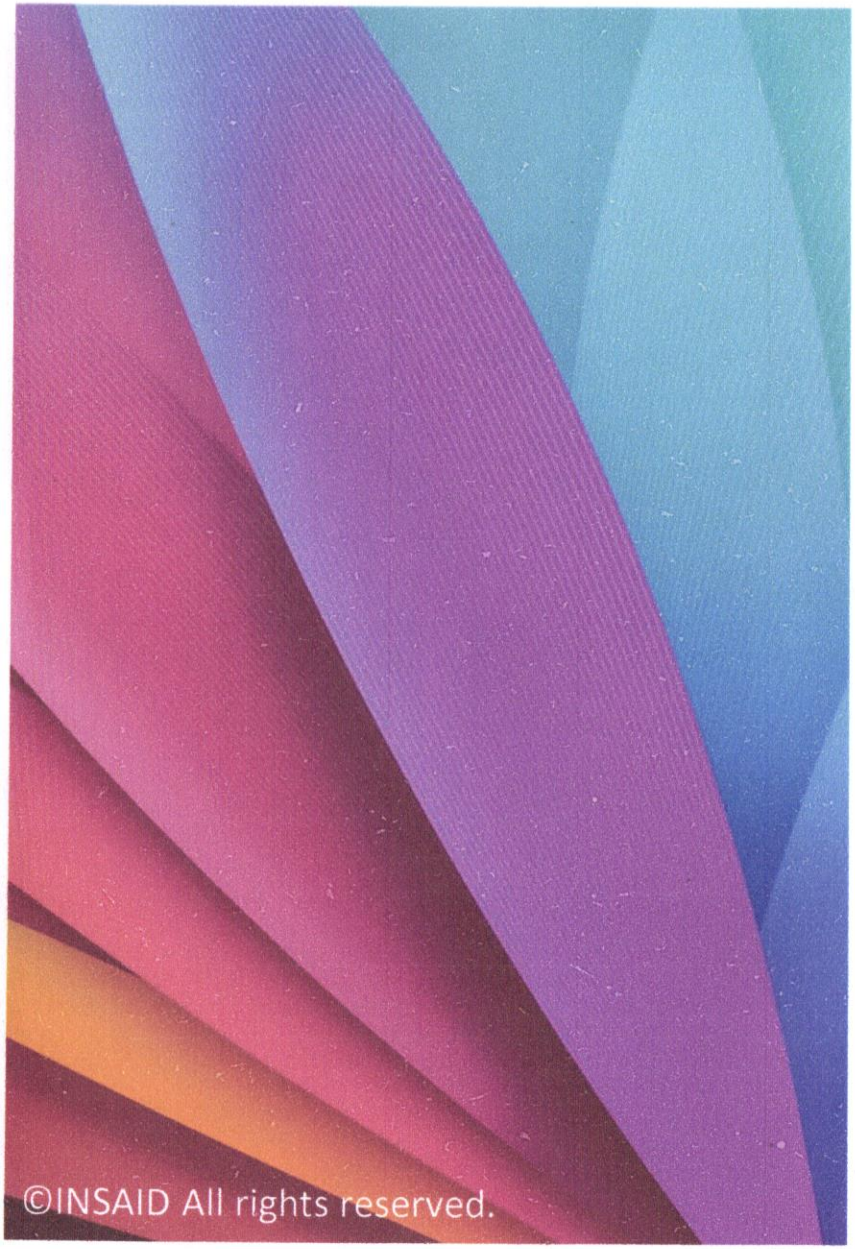
INSAID

Additional Example: Conditional Probability

- What is the probability of the Red Car being stolen?
- The answer is:
  - P(Red | Stolen = Yes)
  - 2 ÷ 8 = 0.25

Index	Color	Type	Origin	Stolen
1	Red	Sports	Domestic	No
2	Red	Sports	Imported	No
3	Blue	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	Yes
5	Yellow	SUV	Domestic	Yes
6	Yellow	SUV	Imported	No
7	Blue	SUV	Imported	Yes
8	Red	Sports	Domestic	No
9	Red	SUV	Imported	Yes
10	Yellow	SUV	Imported	Yes
11	Red	SUV	Domestic	Yes
12	Blue	Sports	Domestic	Yes





Applications of Conditional Probabilities

- Categorizing News
- Email Spam Detection
- Sentiment Analysis
- Facial Recognition
- Fraud Detection
- & Many More...

P(spam | 10 kg of gold) =

P(khichdi | raining in London)

INSAID

if A and B are independent

$P(A|B) = P(A)$

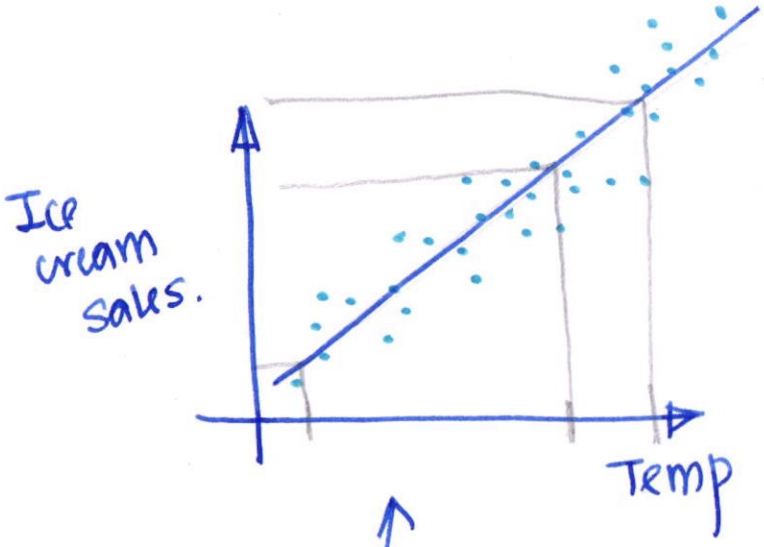
Group A (in India)

Group A (in USA)

Independent Events

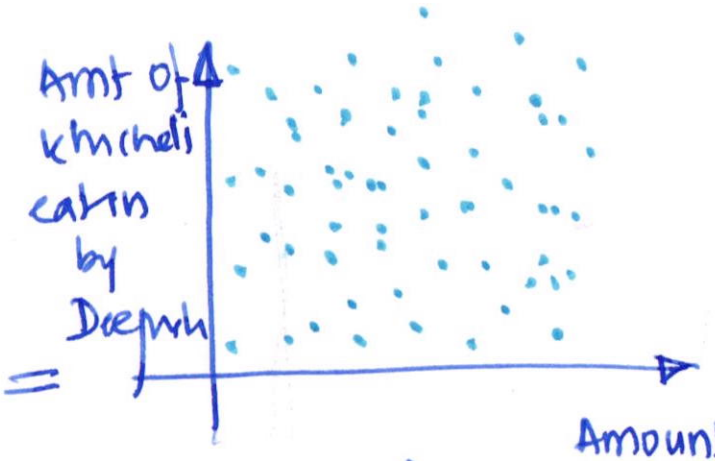
- Friends Group A (in India) and Friends Group A (in USA) walking on road are two independent events.

$P(\text{khichdi} \mid \text{rain in london}) = P(\text{khichdi})$



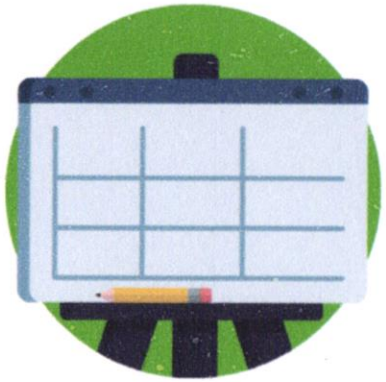
shows dependency

scatter plots



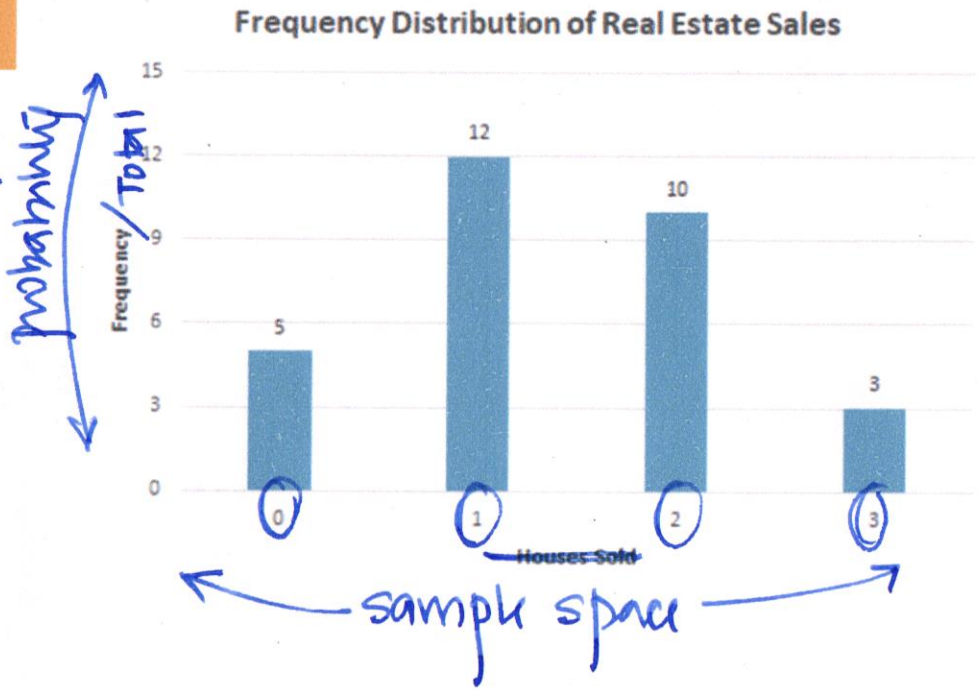
independent

Dependent Events



Walmart Sales

- The spike in sales of umbrellas in Walmart is dependent on rains and thunderstorm.

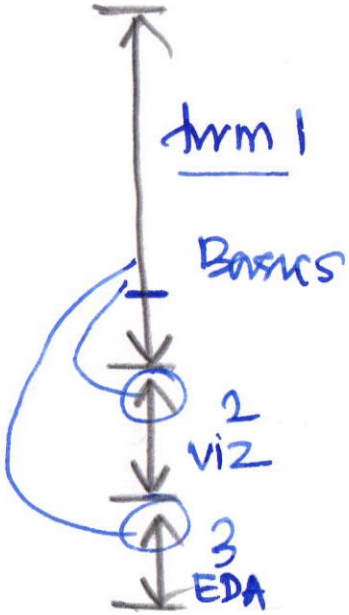


#	Houses Sold	Days	Probability
1	0	5	5 ÷ 30 = 0.17
2	1	12	12 ÷ 30 = 0.4
3	2	10	10 ÷ 30 = 0.3
4	3	3	3 ÷ 30 = 0.1
Total	6	30	30 ÷ 30 = 1

Random Variables

- A random variable X, is a variable whose possible values are numerical outcomes of a random phenomenon.
- **Example:** No. of houses sold by a real estate agent in a month.

INSAID



# hours sold per day	# day	prob
0	5	5/30 = .17
1	12	12/30 = .4
2	10	10/30 = .33
3	3	3/30 = .1
	30	

Avg # hours sold per day = Total # / # days = $\frac{0\times5}{30} + \frac{1\times12}{30} + \frac{2\times10}{30} + \frac{3\times3}{30}$

Expected No.

$$E = 0\times p(0) + 1\times p(1) + 2\times p(2) + 3\times p(3)$$

$$\boxed{E = \sum x\, p(x)}$$

Expected Probability Distribution

Average

- $E(x) = \sum(x*P(x))$
- $E(x) = (0*0.17) + (1*0.40) + (2*0.30) + (3*0.10)$
- $E(x) = 1.3$

Standard Deviation

- $\sigma(x) = \sqrt{(\sum[x - E(x)]^2 P(x))}$
- $\sigma(x) = \sqrt{1.360}$
- $\sigma(x) = 1.16$

#Houses sell

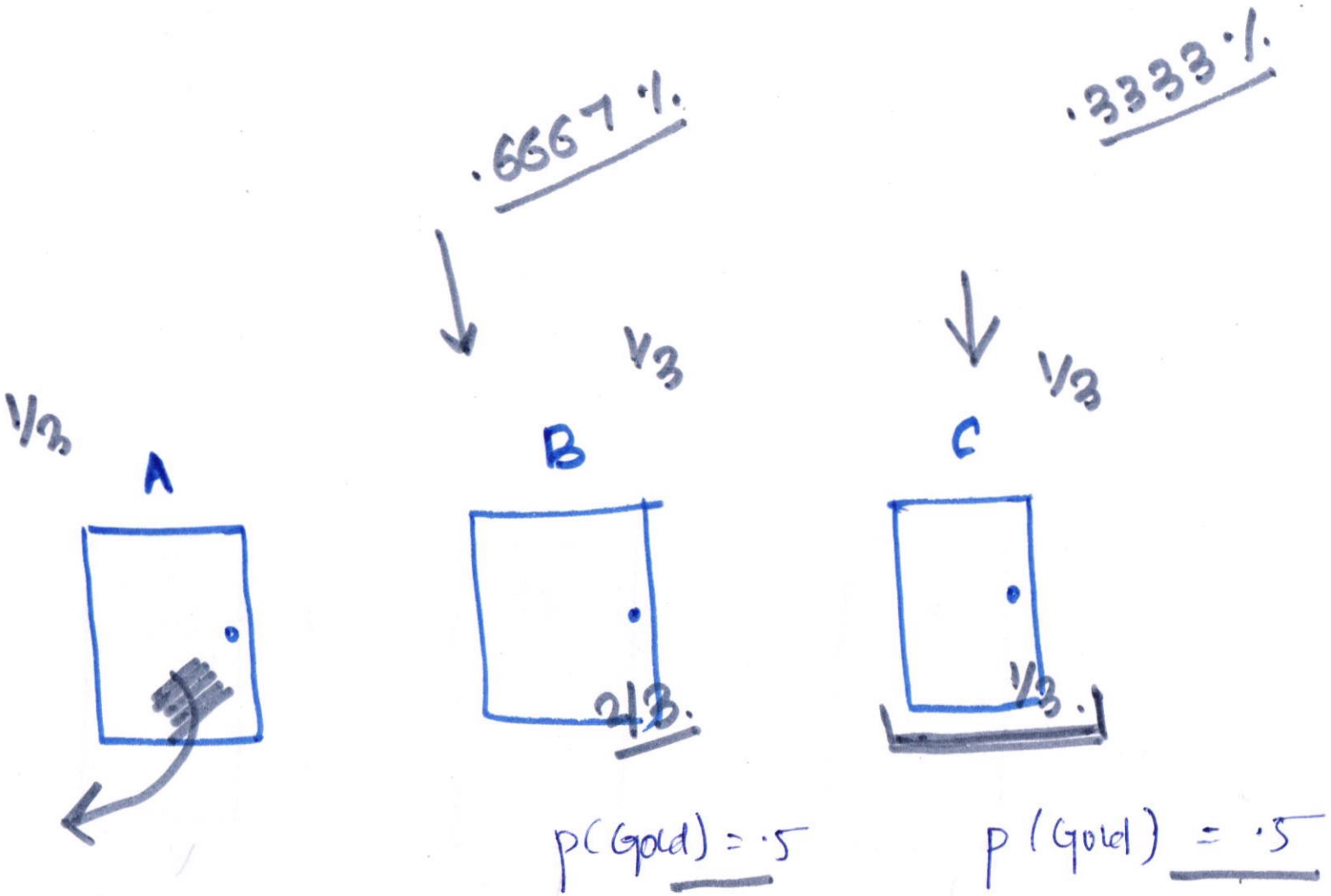
#	Houses Sold	Days	Probability
1	0	5	5 ÷ 30 = 0.17
2	1	12	12 ÷ 30 = 0.4
3	2	10	10 ÷ 30 = 0.3
4	3	3	3 ÷ 30 = 0.1
Total	6	30	30 ÷ 30 = 1

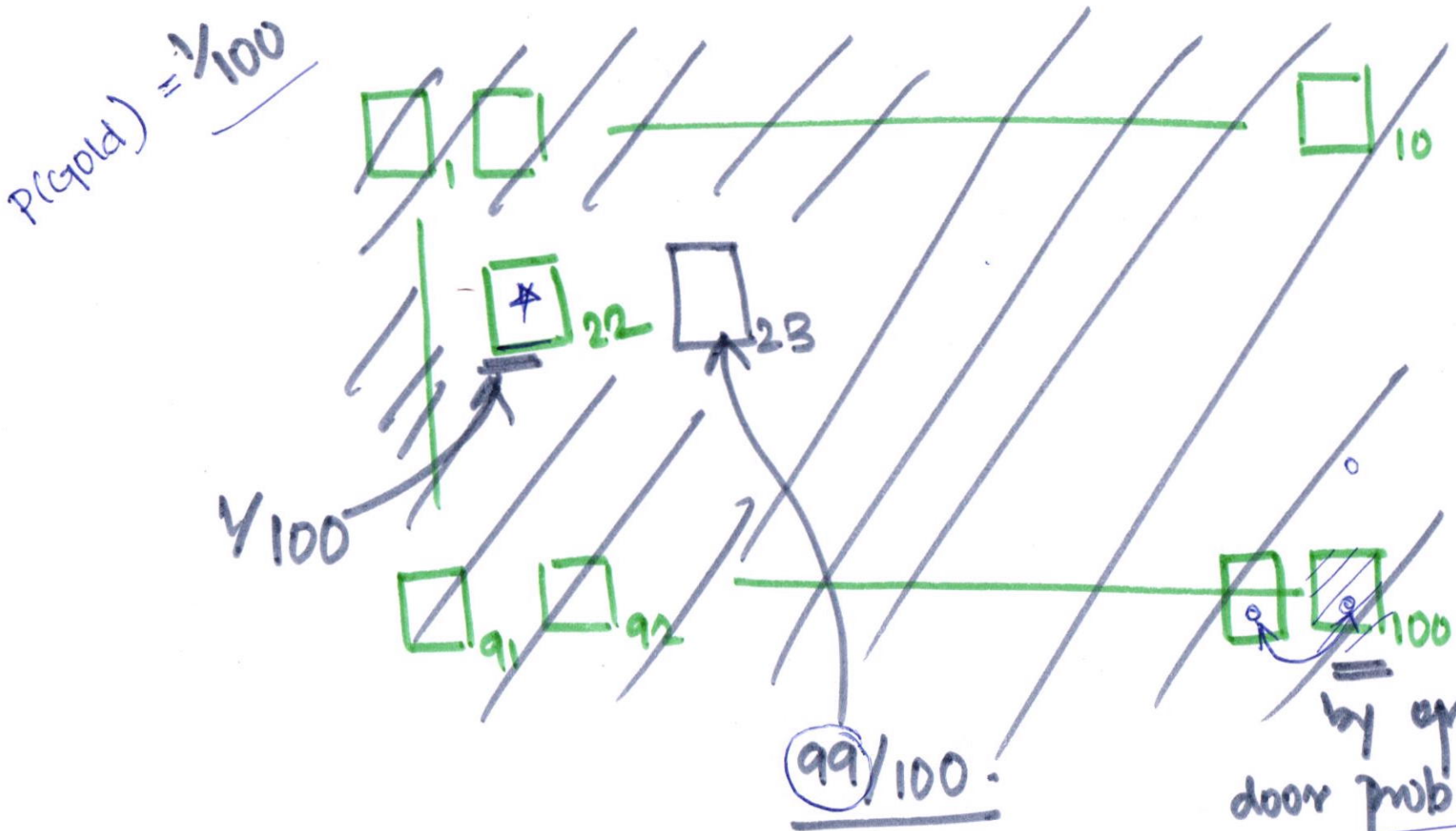
$P(\text{sell 2 houses}) = \frac{10}{30}$

Total # houses = 0×5
+ 1×12
+ 2×10
+ 3×3
41 houses.

- The average number of sales is 1.3 in a day.
- The standard deviation (σ) is 1.16 houses sold per day.

Avg # hous/day = $\frac{41}{30}$ ≈ 1.36


.6667 %.
.3333 %.
1/3
A
1/3
B
1/3
C
2/3.
1/3.
p(Gold) = .5
p(Gold) = .5

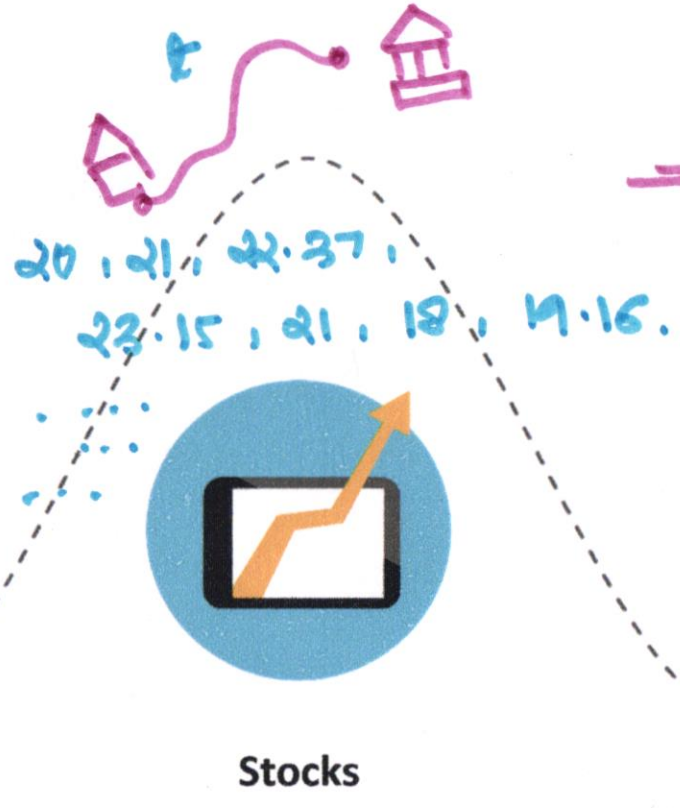


by opening the 100th door prob of 22nd door having gold is unchanged and still is 1/100

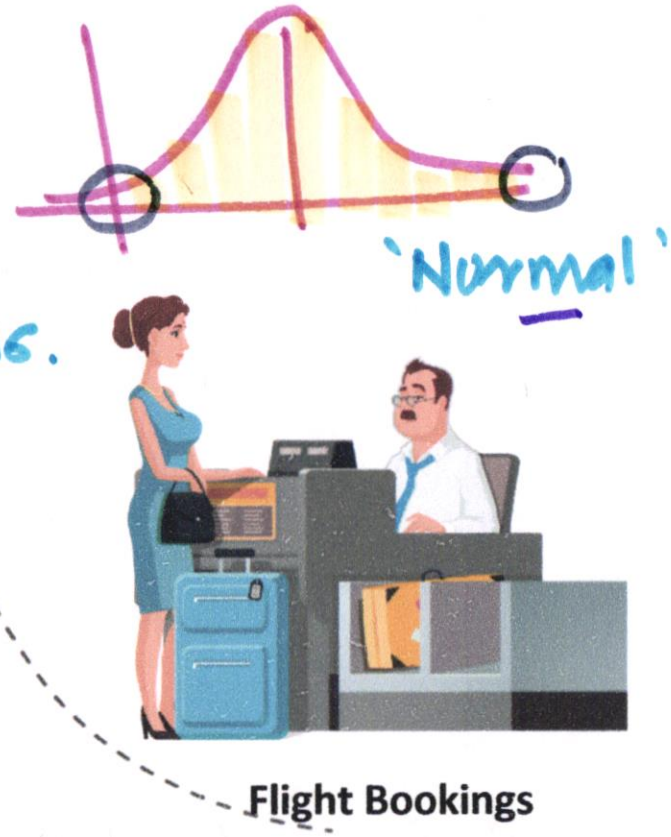
switching is good !

* I trust deepseek and he says its true

IQ Scores



Stocks



Flight Bookings

Normal Probability Distribution

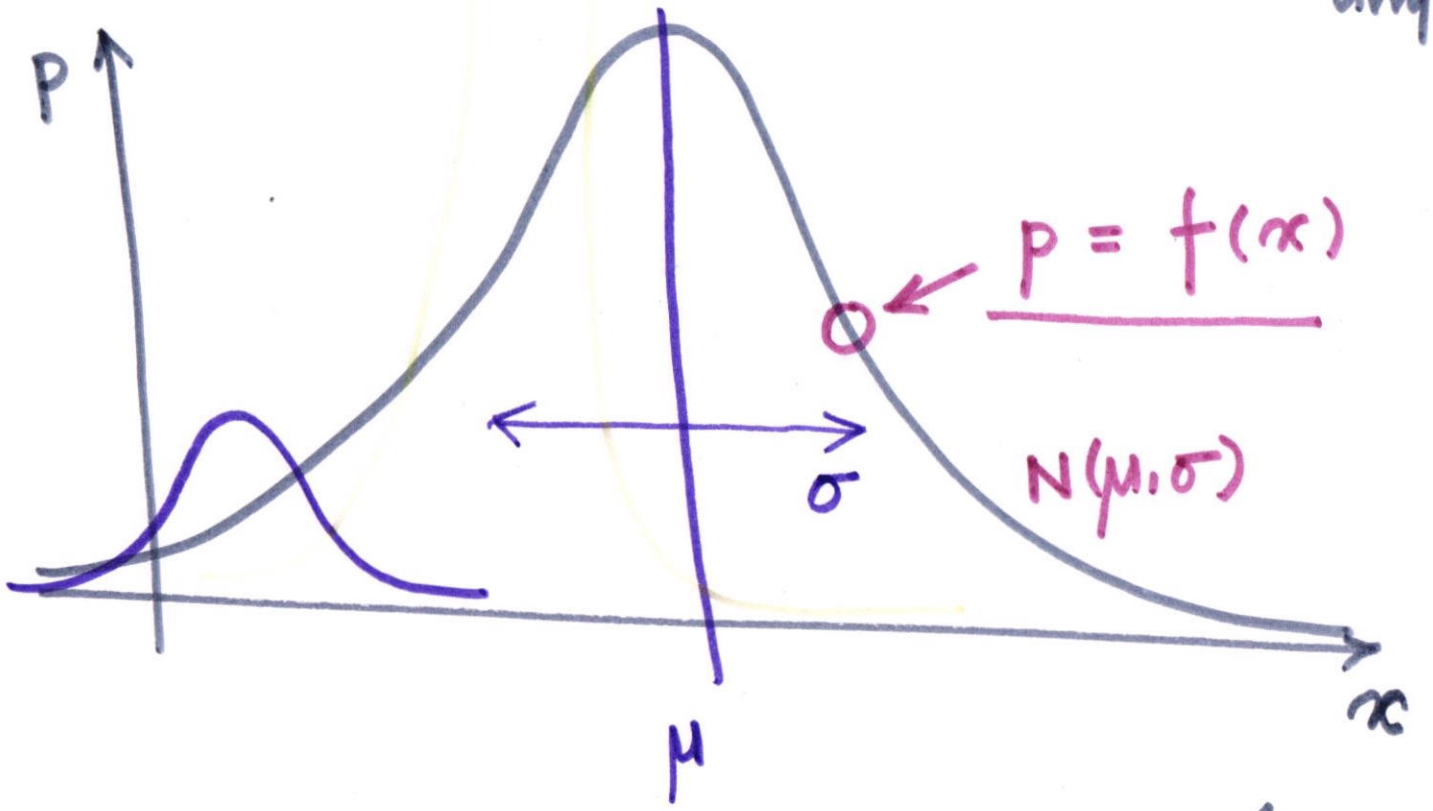
Bell curve

continuous variable

- Event Normal probability distribution models many natural processes, manufacturing processes and human endeavors.

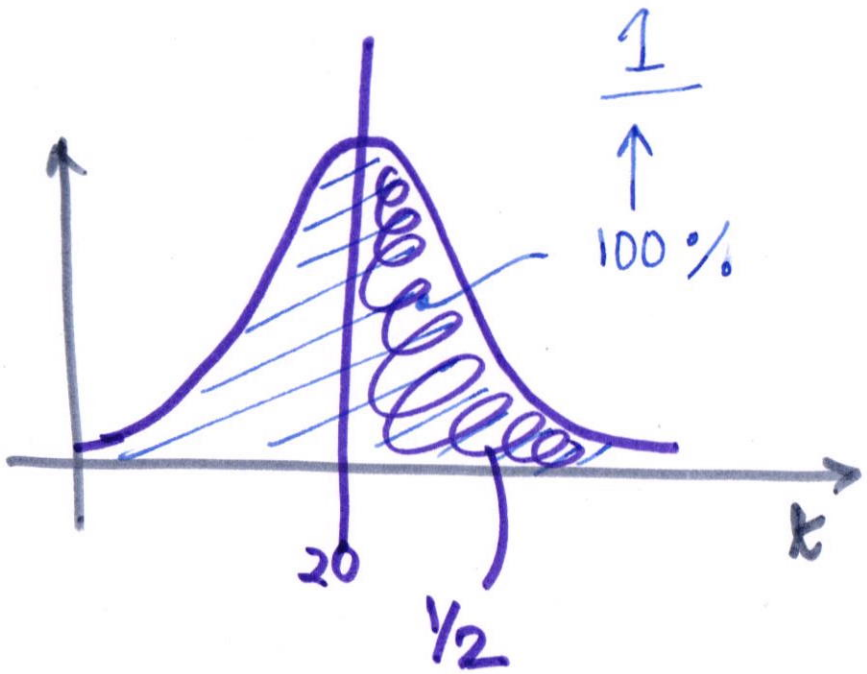
INSAID

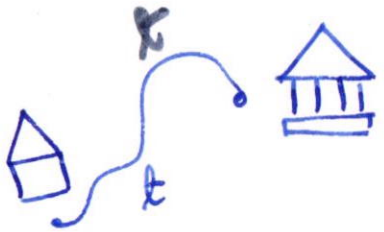
Height
Length - tree leaves



Height = $\underline{\underline{N}}$ (μ = 5.5 feet, σ = .5 feet)

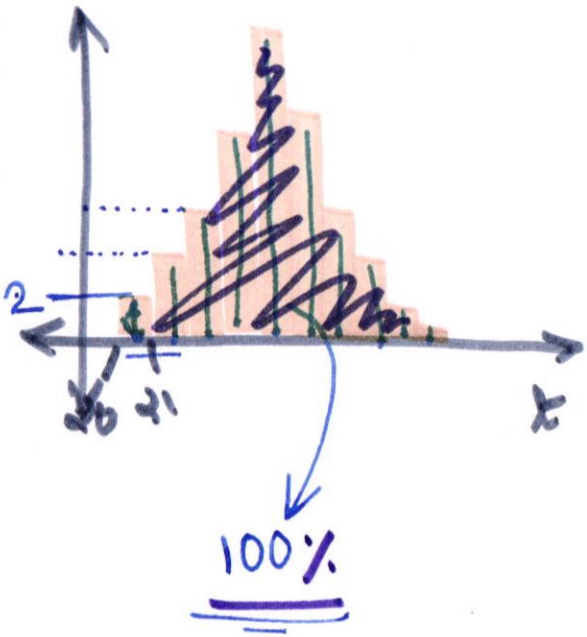
Length of Leaf = $\underline{\underline{N}}$ (μ = 5 cm, σ = 1 cm)

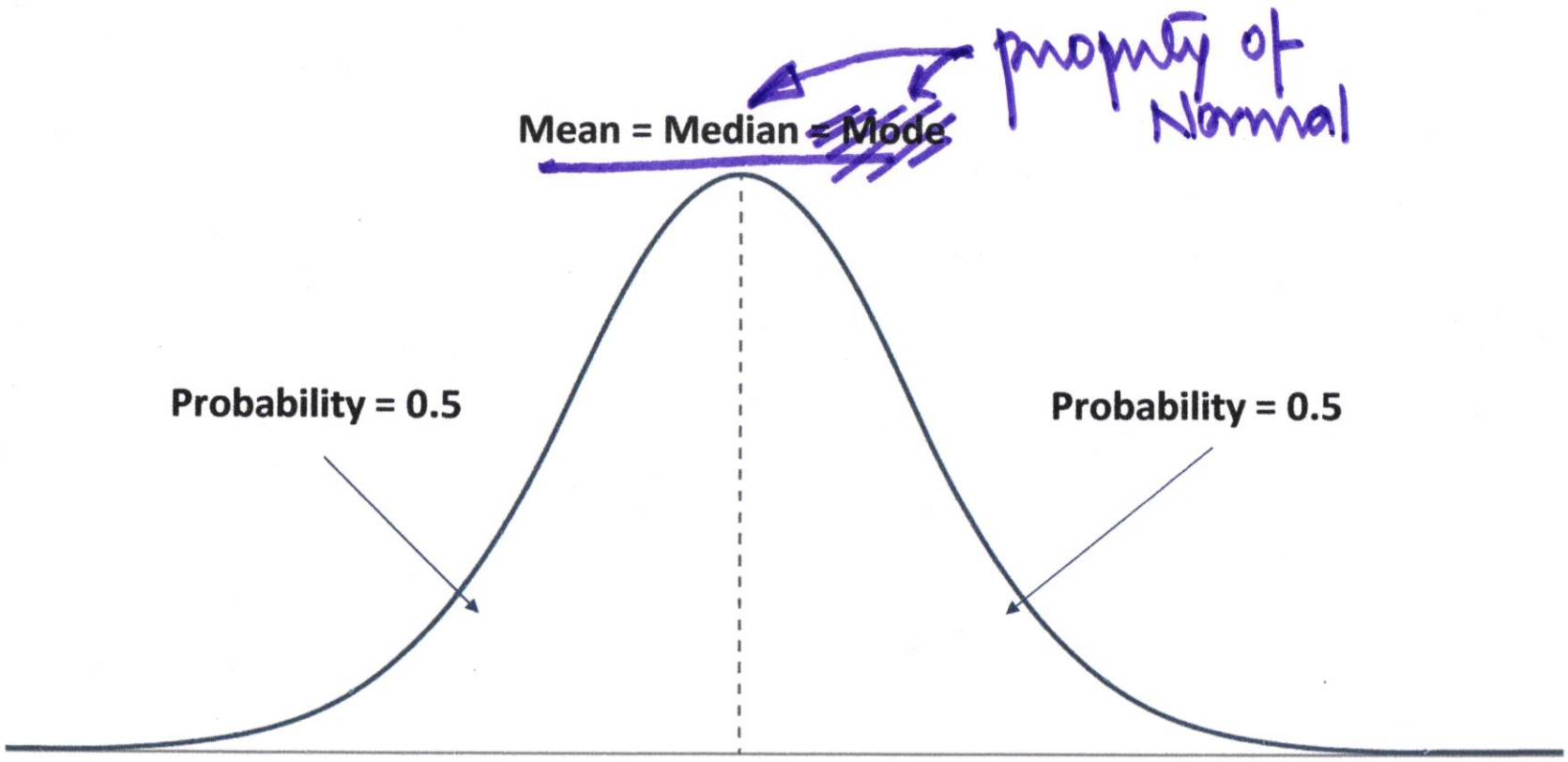




t	#
20 - 21	2
21 - 22	
22 - 23	
23 - 24	
⋮	

100

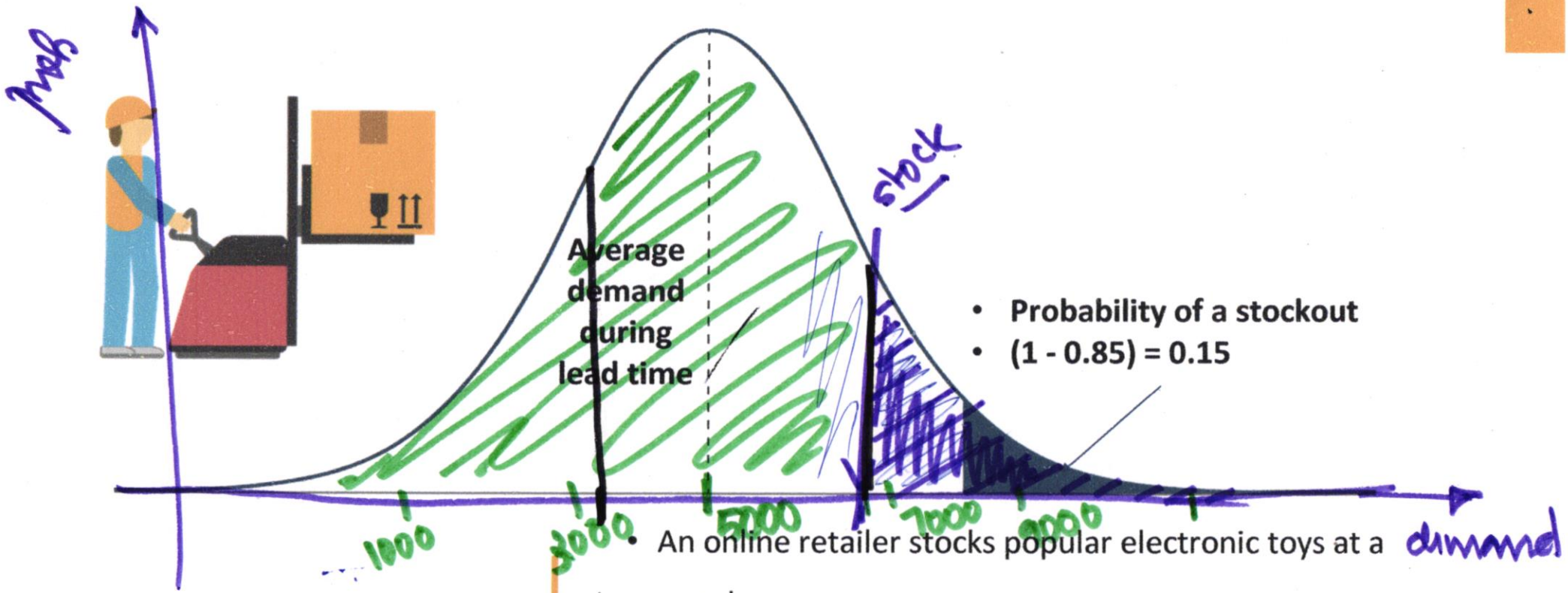




Normal Distribution (Visual Appearance)

- Has a peak.
- Has symmetricity in nature.
- Centered around.

INSAID



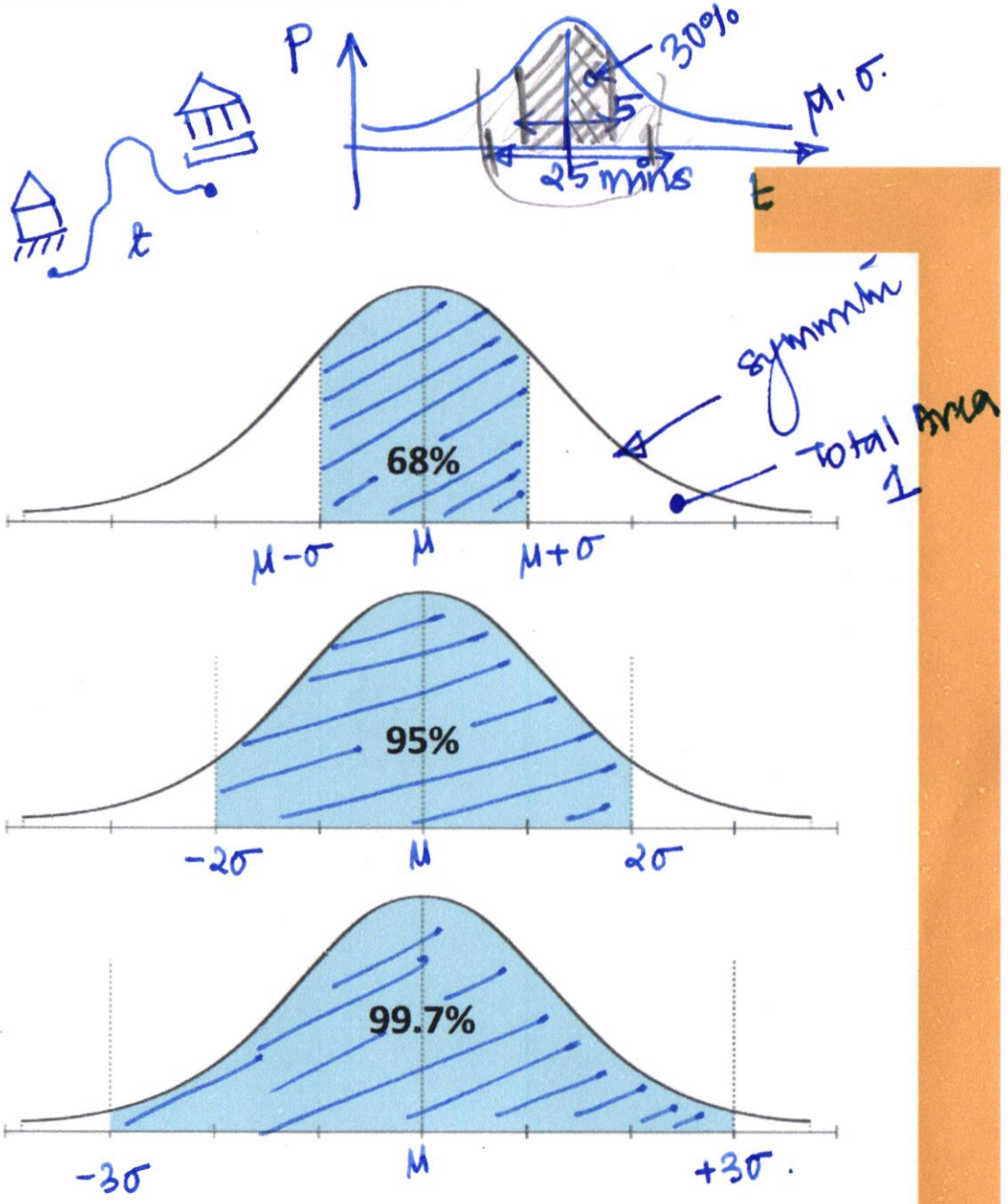
Stock Inventory (Example)

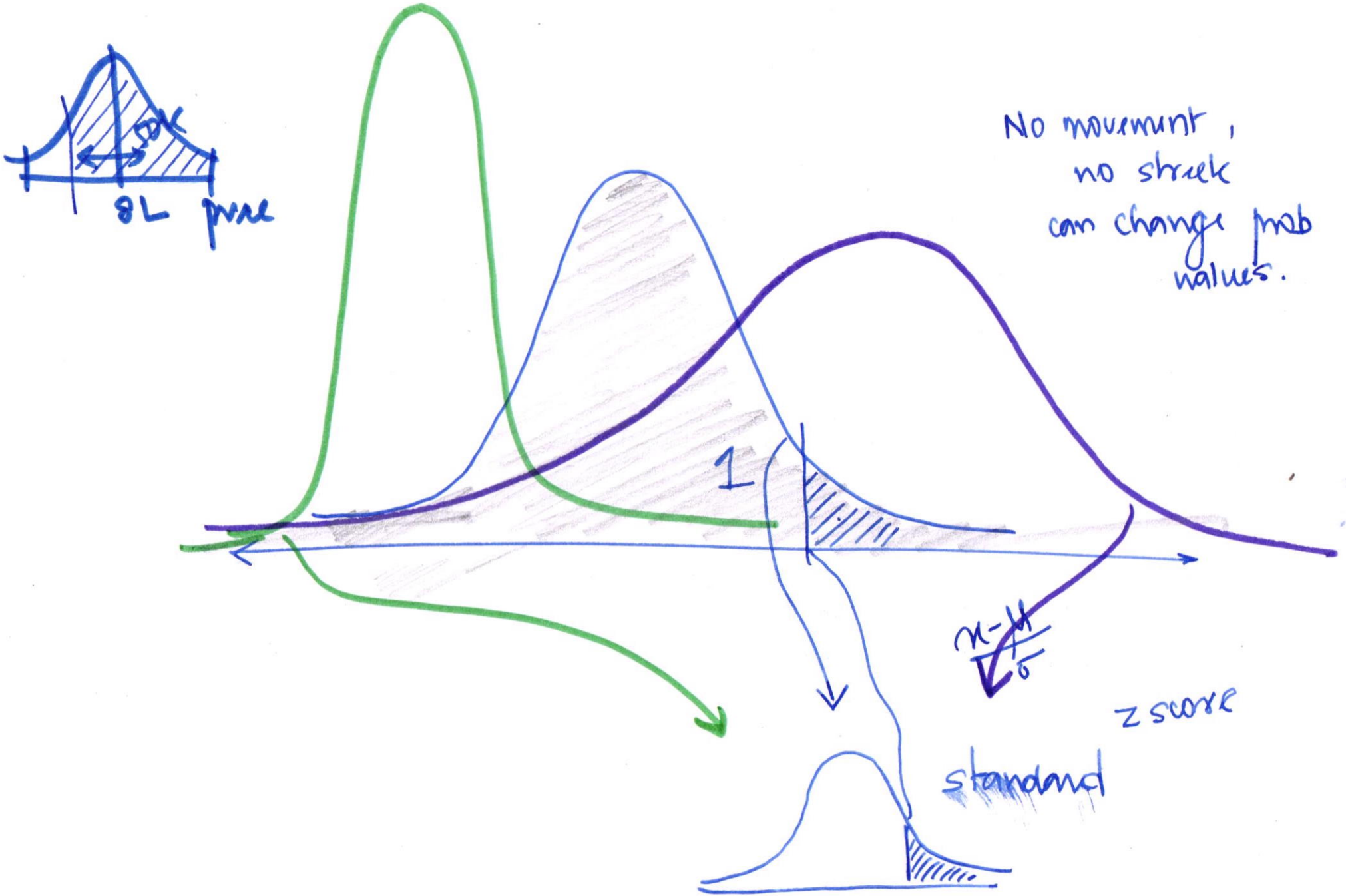
- An online retailer stocks popular electronic toys at a warehouse.
- Every week he has to decide the no. of units per toy to stock.
- He observes the weekly demand of toys follows a normal distribution.

INSAID

Empirical Rule

- 68% of values lies within 1 standard deviation from the mean.
- 95% of values lies within 2 standard deviations from the mean.
- 99.7% of values lies within 3 standard deviations from the mean.

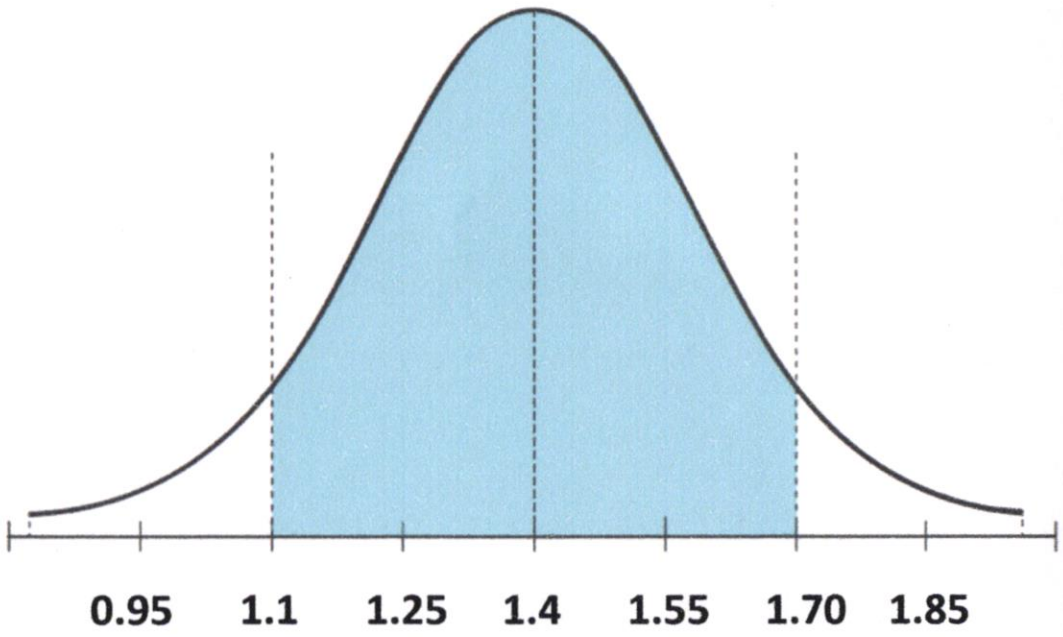



SD
SL price
No movement,
no streak
can change prob
values.
1
$\frac{x-\mu}{\sigma}$
z score
standard

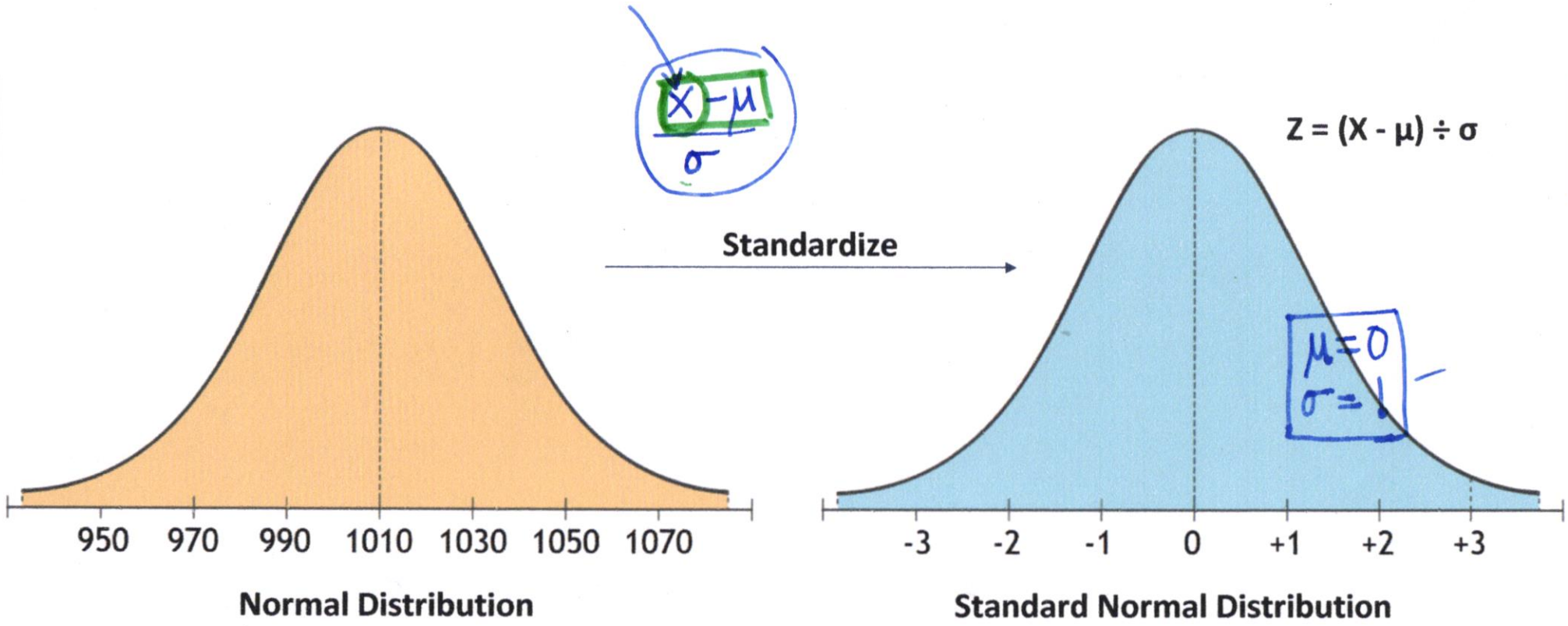
Empirical Rule (Example)

- Normal distribution of height of students.
- Average height is 1.4m.
- 95% of students at school are between **1.1m and 1.7m** tall.

- **Mean = (1.1 + 1.7) ÷ 2 = 1.4**
- **1 S.D. = (1.7 - 1.1) ÷ 4 = 0.15**



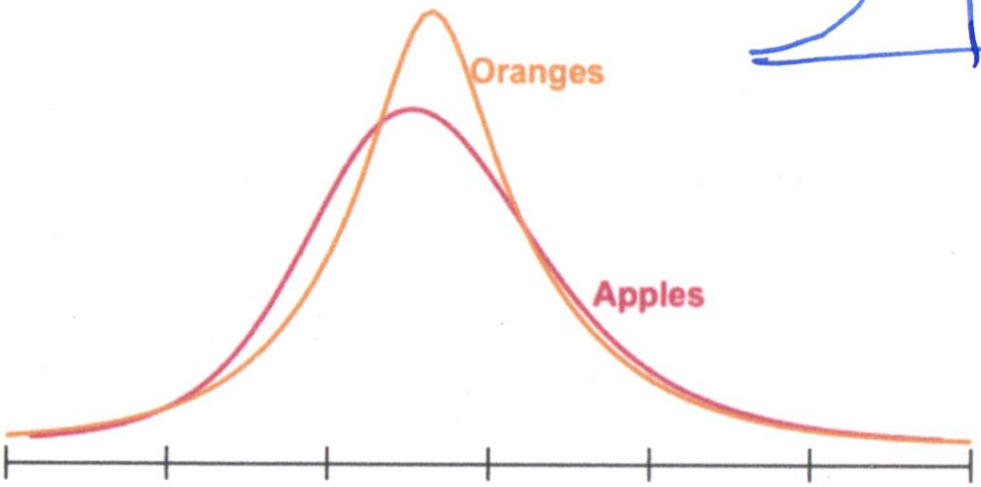
INSAID



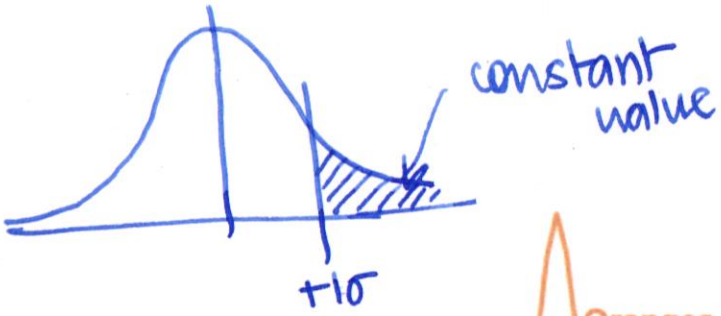
Standardizing Data Values

- The number of **standard deviations from the mean** is also called the "Standard Score", "sigma" or "z-score".

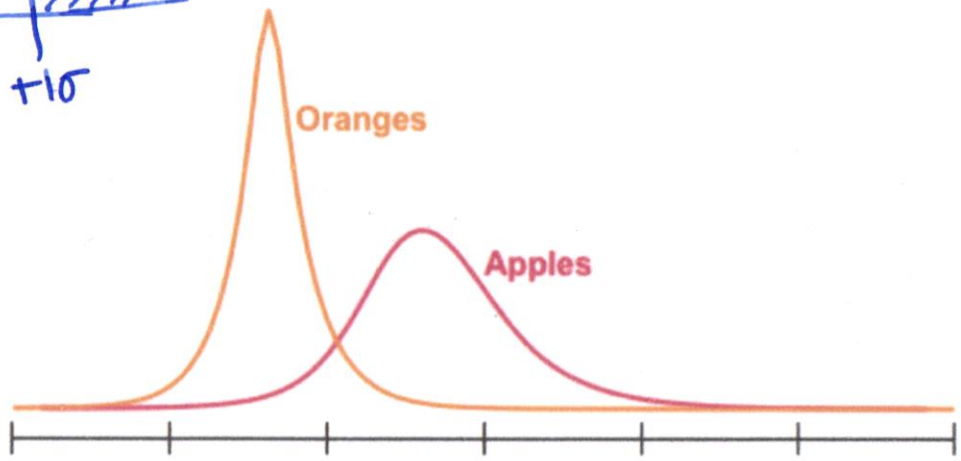
INSAID



Price in Currency



$$Z = (X - \mu) \div \sigma$$



Standardize Price

Reason of Standardizing?

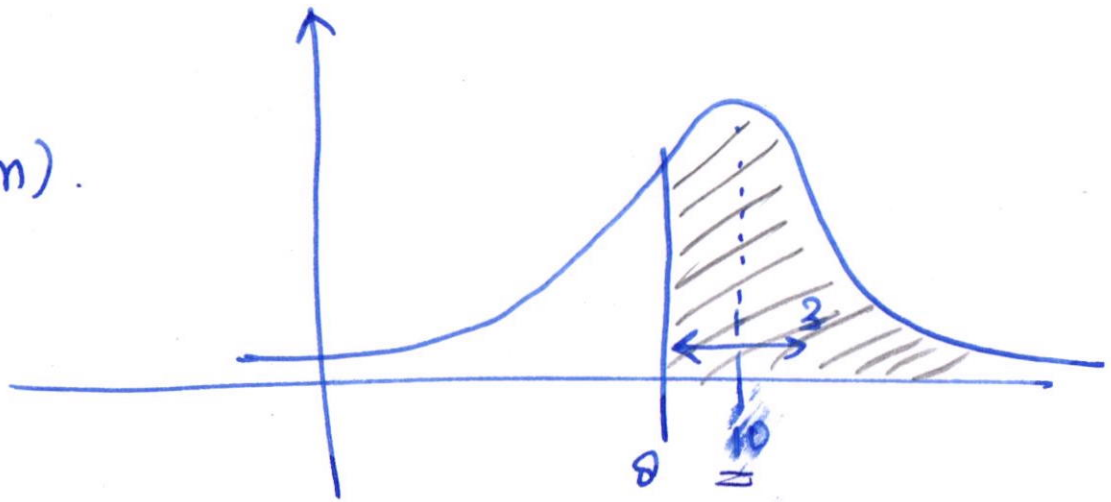
- To analyze data, a common ground is required.
- It helps in achieving consistent results.
- **For Example:** Price Currency in US and India is different.

INSAID

P (length > 8 cm).

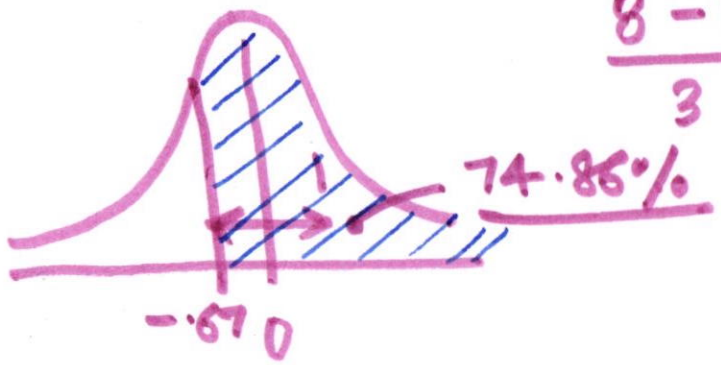
$\mu = 10$ cm

$\sigma = 3$ cm



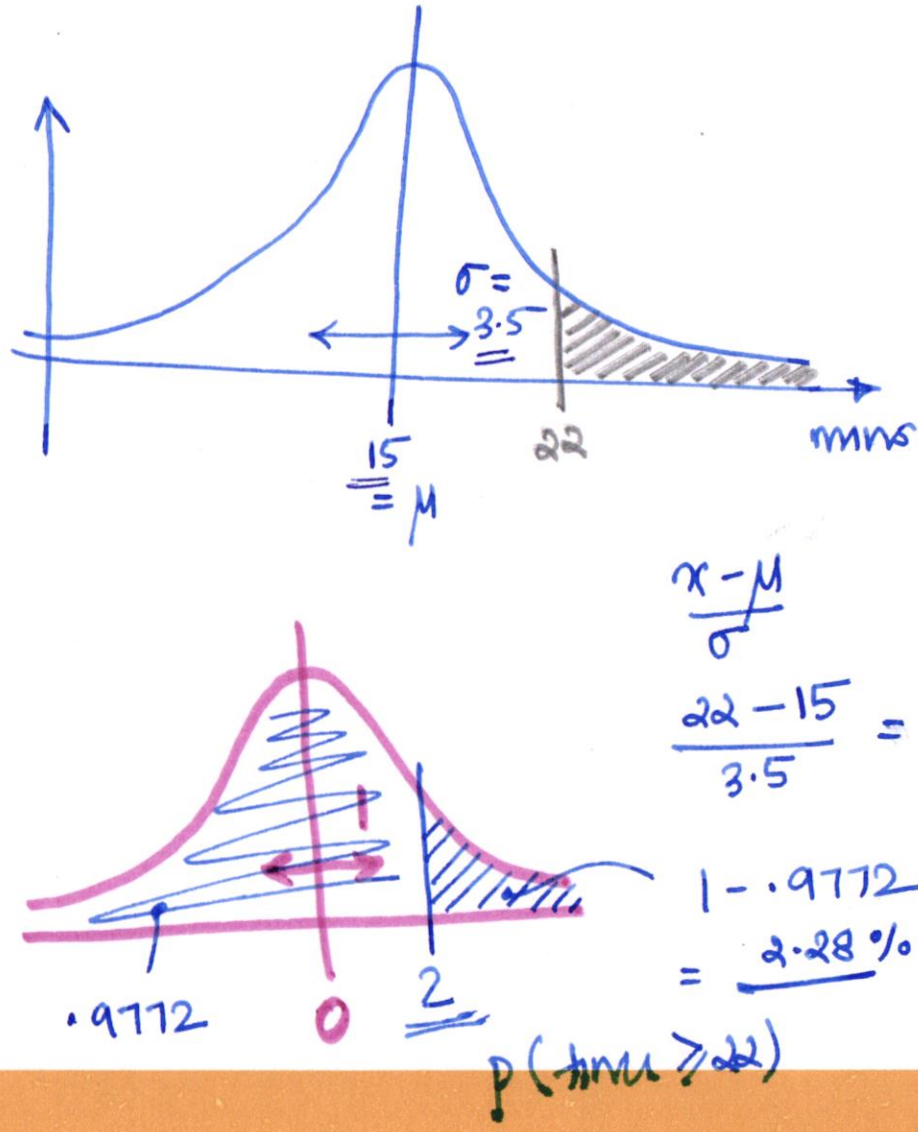
$$\frac{x - \mu}{\sigma}$$

$$\frac{8 - 10}{3} = -0.67$$

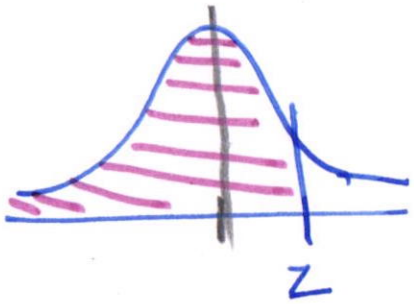


Problem: Flight Delay

- IGI Airport Managers found that average time to load passengers on a plane is equal to 15 minutes with standard deviation of 3.5 minutes. Their next takeoff is ready to Bangalore.
- What is the probability that flight will take 22 minutes or more?



INSAID



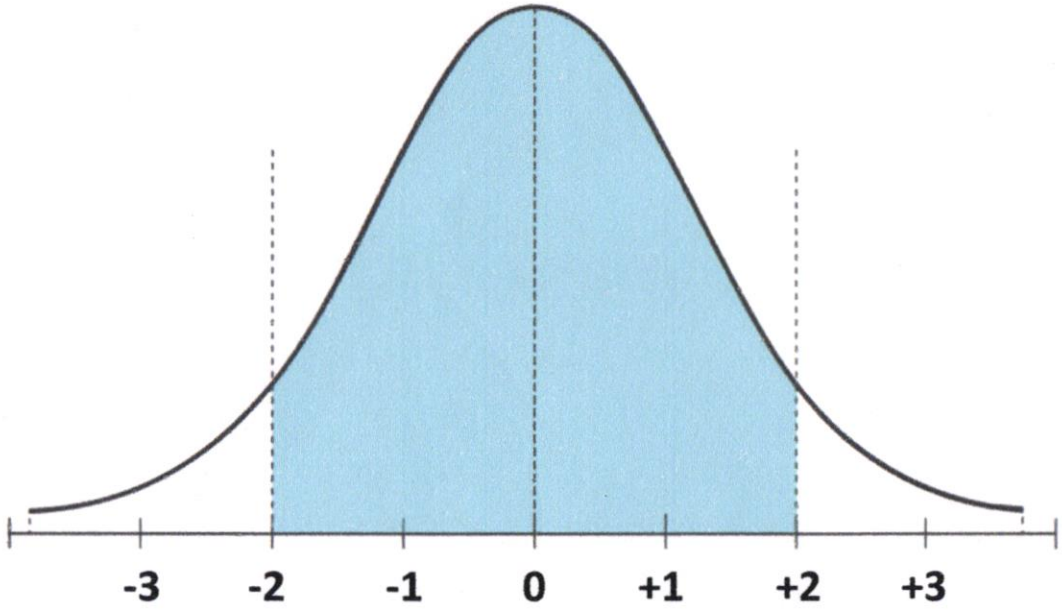
Area	
0	·5
·01	
·02	
·03	
·04	
⋮	
2	
⋮	
3	
⋮	
∞	1

	·01	·02	·03	·04	——	·1
0	·5	⋮	·	·		

Solution: Flight Delay

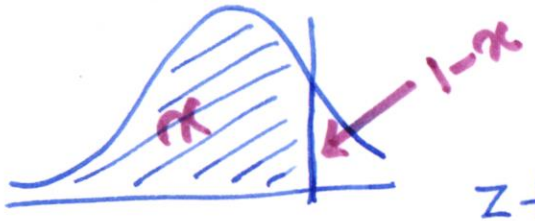
Given data:

$\mu = 15$ minutes , $\sigma = 3.5$ minutes , $P(X \geq 22) = ?$



Z – Table

- Given table estimates area percentage for positive z – score.
- There also exist negative z - score table.
- Shows the area percentage (to either left or right) of a given z-score on a standard normal distribution.



Z table

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9958	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

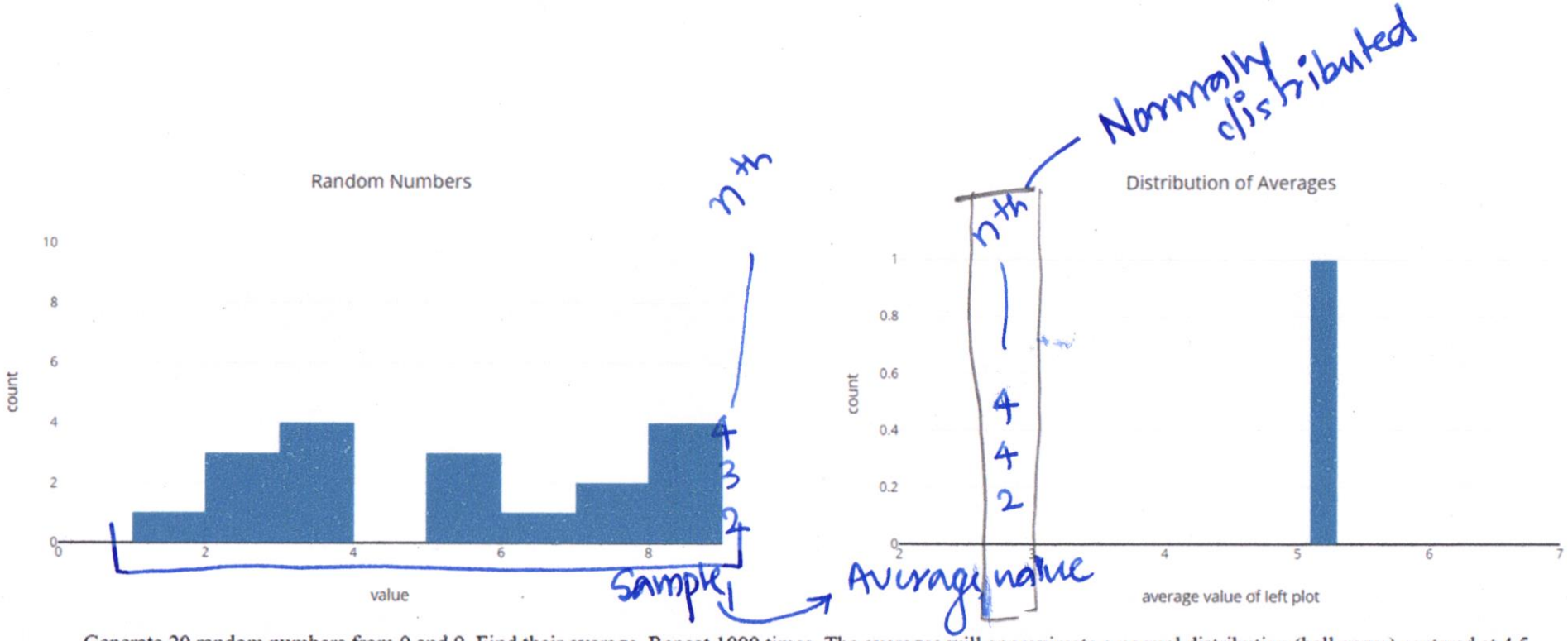
SD

$\mu + z\sigma =$ travel time

99.98%

.2%

INSAID



Generate 20 random numbers from 0 and 9. Find their average. Repeat 1000 times. The averages will approximate a normal distribution (bell curve) centered at 4.5.

Central Limit Theorem

- If you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement , then the distribution of the sample means will be approximately normally distributed. Experiment Here.

INSAID

guess *

* can be checked / verified.

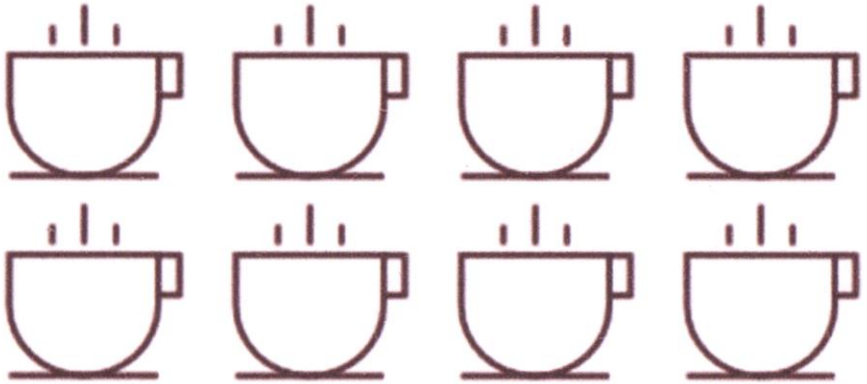
Accept / Reject.

Hypothesis Testing

- An **educated guess** about something in the world around you.
- It should be testable, either by experiment or observation.
- **Example:** A new medicine you think might work.

Ronald Fisher

Test Of Distinction



- **Null (H_0):** She's unable to distinguish teas.
- **Alternate (H_A):** She's an expert in distinguishing.

Lady Tasting Tea

Origin Of Hypothesis Testing

- Ronald thought it is impossible to distinguish whether a tea is made with water first or milk first.
- Tried an experiment with a random lady to whom he lost.

INSAID

- **Null (H_0):** There's no Home Field Advantage.
- **Alternate (H_A):** There's Home Field Advantage.

Example: Baseball Match

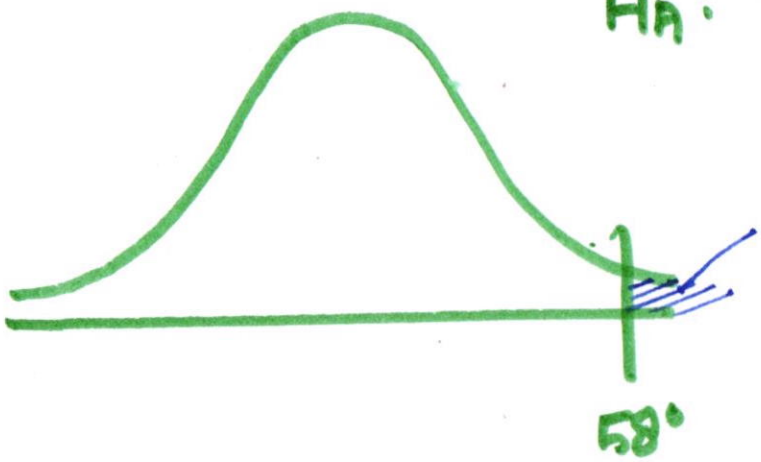
- According to the player, there's no home field advantage I.e. there's no bias (higher chances of winning) playing at home.
- A reporter tried a test to find out whether it is true or not.

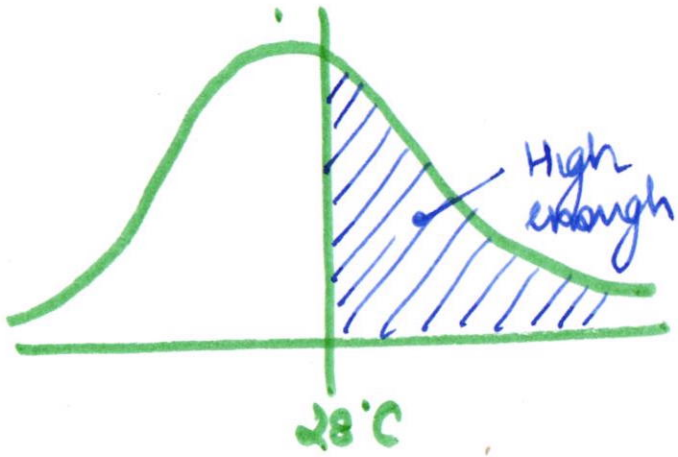
INSAID

H_0: Temp in Indore is $\geq 58\,^{\circ}C$. ← Reject the Null

H_A: Temp in Indore is not $\geq 58^{\circ}C$

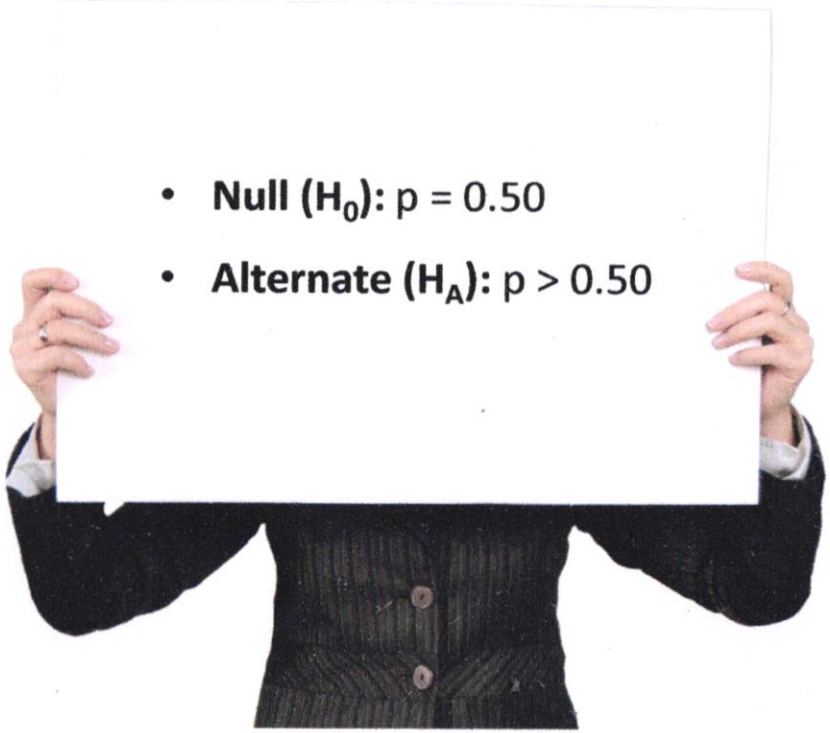
95% confidence





H_0: Temp in Indore right now $\geq 28\,^{\circ}C$ ← Fail to reject the Null Hypothesis

H_A: Tem is $< 28\,^{\circ}C$



- o = Observed Winning Percentage
- e = Expected Winning Percentage
- p = Chances of Winning
- q = Chances of Loosing
- n = Total Number of Matches Played

- $Z = (o - e) \div \sqrt{(p * q \div n)}$
- $Z = (0.53 - 0.50) \div \sqrt{((0.50 * 0.50) \div 2431)}$
- $Z = 3.76$

This results in a p-value < 0.0001

Reporter's Conclusion

- The null hypothesis turn out to be rejected as the threshold found to be more than 50% evident by p-value.
- The data seems to support that the home field team wins **more than 50%** of the time.

INSAID