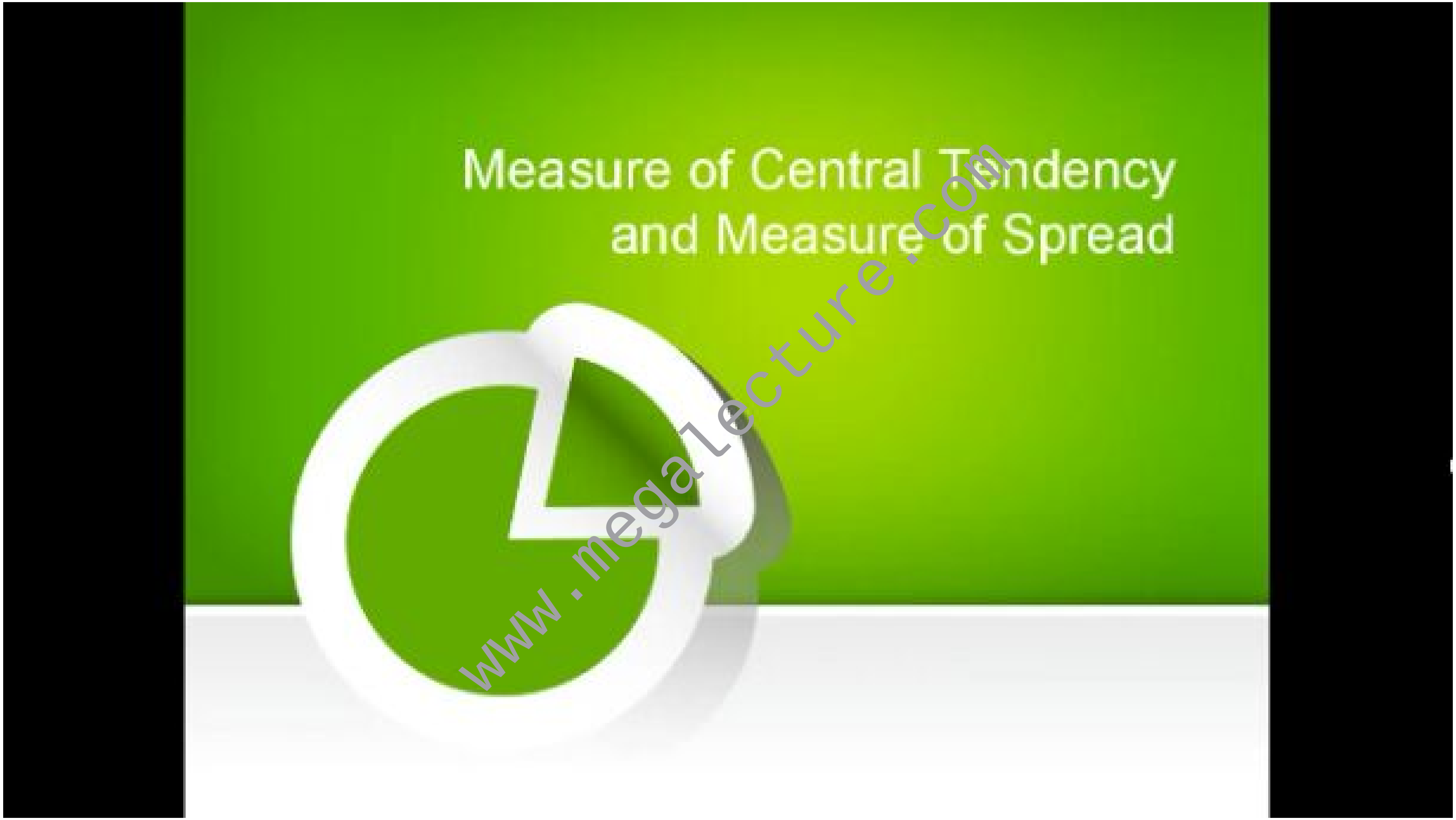
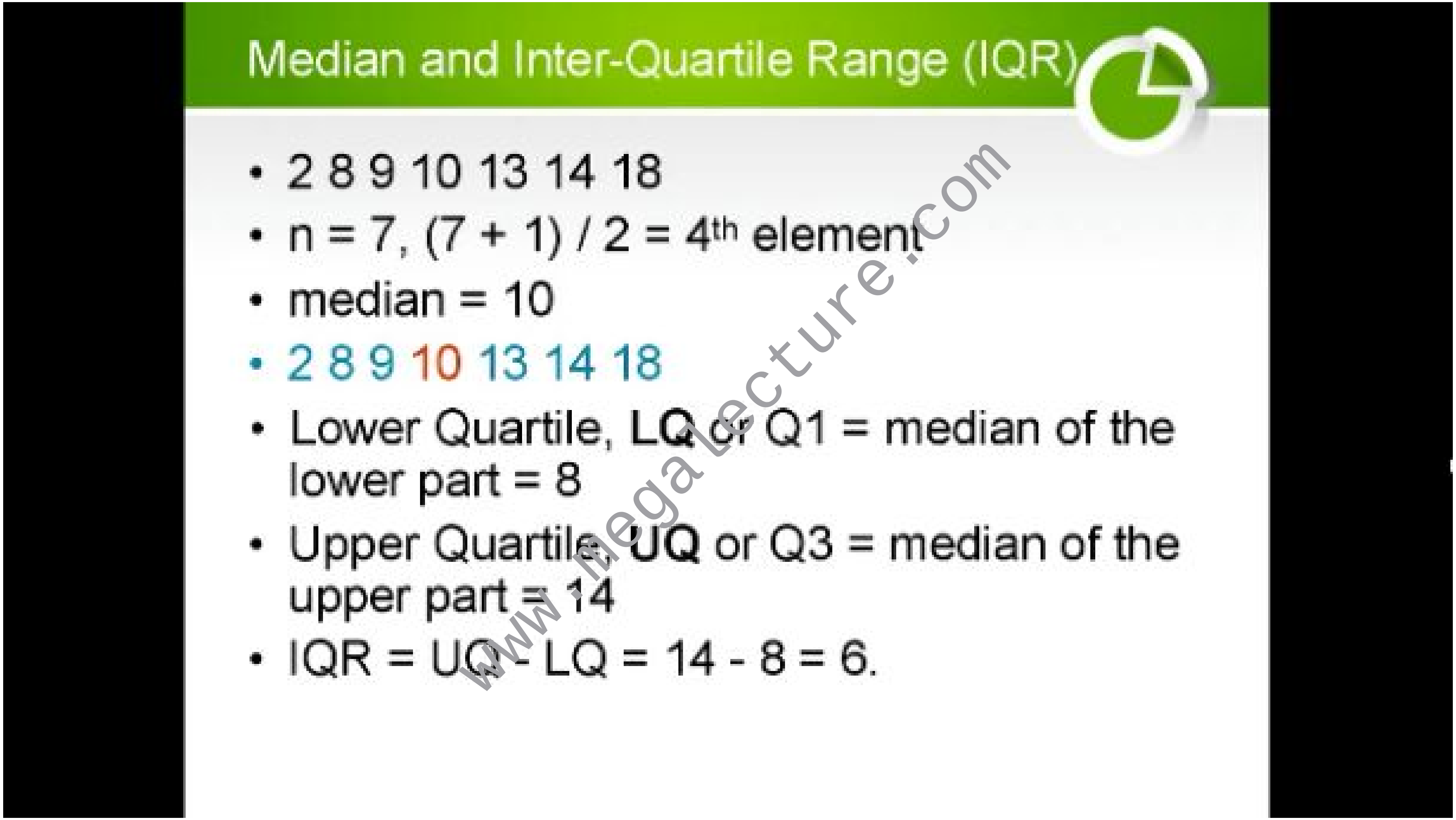
MEGA LECTURE



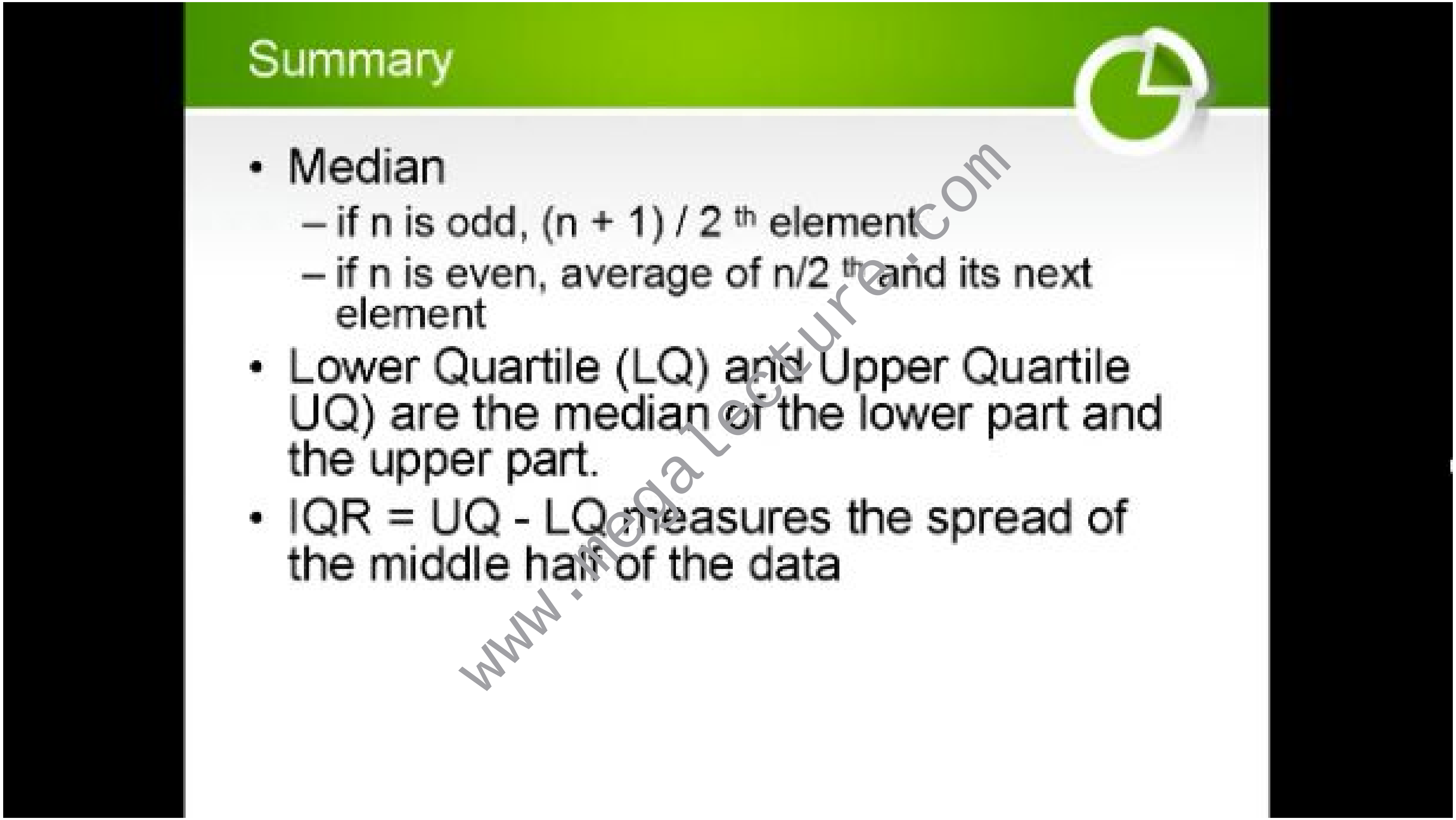
What are these two measures?

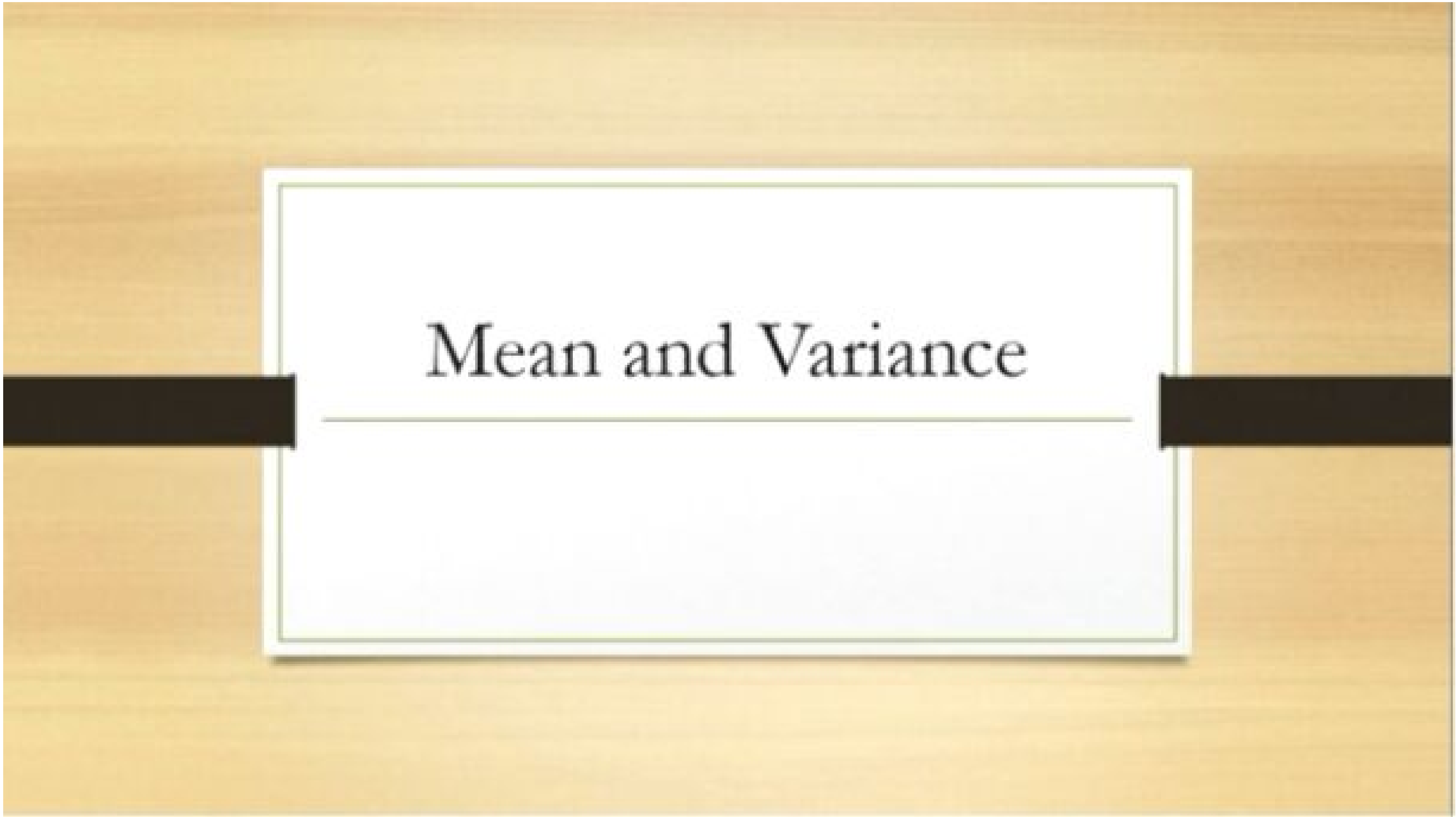
- Let's take a look at 2 groups of data
  - 1 3 5 7 9
  - 3 4 5 6 7
- The number in the middle is 5 for both groups
- The first group spans from 1 to 9.
- The second group spans from 3 to 7.


Median and Inter-Quartile Range (IQR)
• 2 8 9 10 13 14 18
• n = 7, (7 + 1) / 2 = 4th element
• median = 10
• 2 8 9 10 13 14 18
• Lower Quartile, LQ or Q1 = median of the lower part = 8
• Upper Quartile, UQ or Q3 = median of the upper part = 14
• IQR = UQ - LQ = 14 - 8 = 6.

When n is even

- 2 8 9 10 13 14 18
- median = 10, LQ = 8, UQ = 14, IQR = 6
- 2 8 9 10 13 14 18 20   n = 8, 8 / 2 = 4
- median = (10 + 13) / 2 = 11.5
- LQ = (8 + 9) / 2 = 8.5
- UQ = (14 + 18) / 2 = 16
- IQR = 16 - 8.5 = 7.5


Summary
• Median
– if n is odd, (n + 1) / 2 th element
– if n is even, average of n/2 th and its next element
• Lower Quartile (LQ) and Upper Quartile UQ) are the median of the lower part and the upper part.
• IQR = UQ - LQ measures the spread of the middle half of the data
www.megalecture.com


Mean and Variance

Mean

- Mean is what we usually call average.
- $\bar{x} = \frac{\sum x}{n}$
- 8, 2, 16, 3, 11
- $\bar{x} = (8 + 2 + 16 + 3 + 11) / 5 = 40 / 5 = 8$

Variance and Standard Deviation

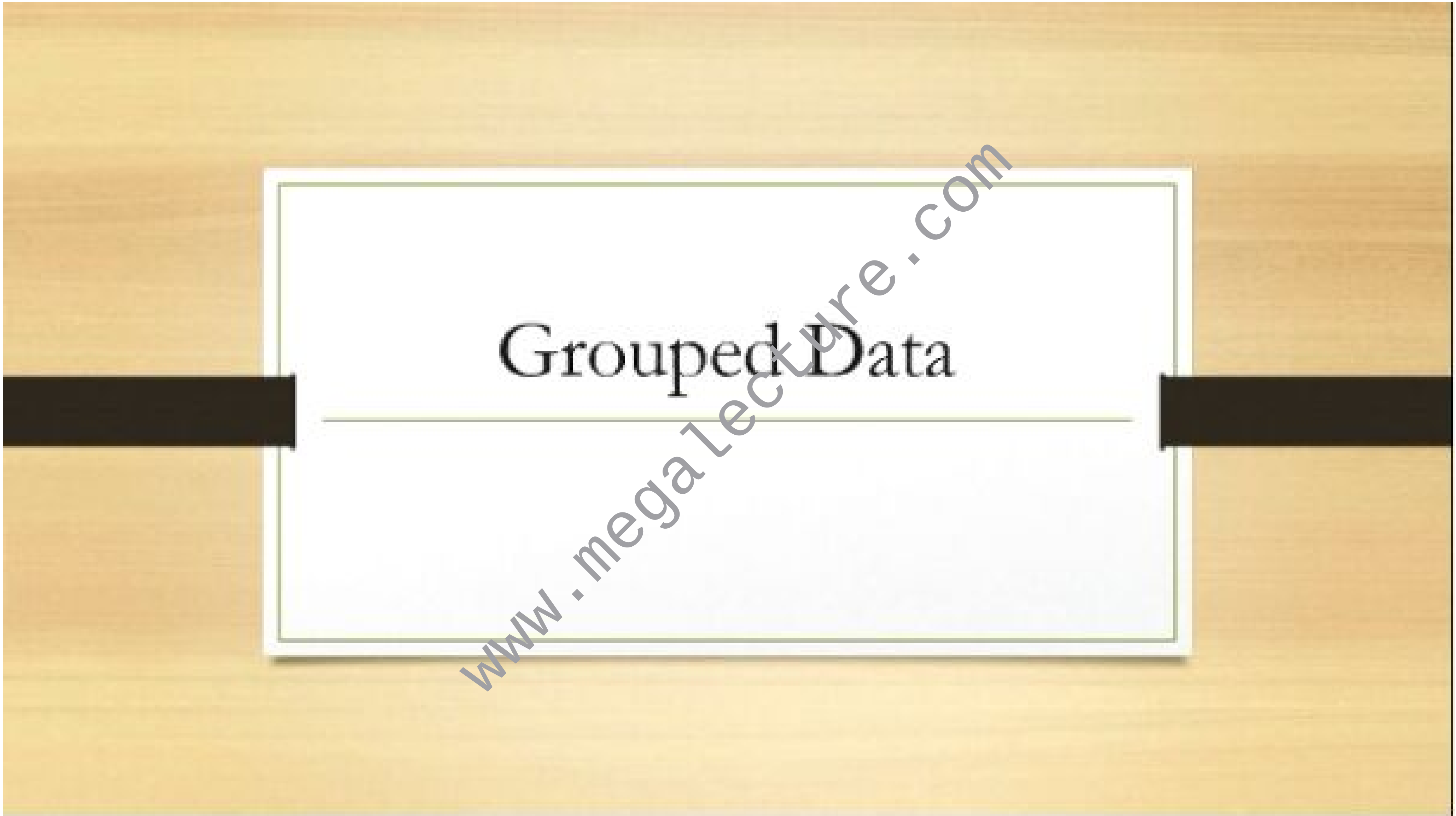
- Variance measures the average distance square to the mean.
- Standard Deviation (SD) is the square root of variance.
- $variance = \frac{\sum(x-\bar{x})^2}{n}$ distance to mean is $|x - \bar{x}|$
- 8, 2, 16, 3, 11 $\bar{x} = 8$
- $(x - \bar{x})^2$: 0, $(2-8)^2$=36, 64, 25, 9
- variance = (0 + 36 + 64 + 25 + 9) / 5 = 26.8, SD = $\sqrt{26.8}$ = 5.18

Variance and Standard Deviation

- $variance = \frac{\sum(x-\bar{x})^2}{n}$
- Variance is always non-negative. So is standard deviation.
- If variance or standard deviation is 0, all numbers are the same.
- Variance is average distance square to the mean.
- Standard deviation can be roughly thought of as average distance to the mean.

Alternative Formula for Variance

- $variance = \frac{\sum x^2}{n} - \bar{x}^2$ mean of squares minus square of mean
- 8, 2, 16, 3, 11 $\bar{x} = 8$
- x^2: 64, 4, 256, 9, 121 $\frac{\sum x^2}{n} = (64 + 4 + 256 + 9 + 121) / 5 = 90.8$
- variance $= 90.8 - 8^2 = 26.8$


Grouped Data

Frequency

- Frequency means how many appearances.

x	12	23	34	42
frequency	5	11	8	6

- 12 appears 5 times, 23 appears 11 times, ……
- 12, 12, 12, 12, 12, 23, 23, 23, 23, 23, 23, 23, 23, 23, 23, 23, 34, 34, 34, 34, …
- total $= \sum f = 5 + 11 + 8 + 6 = 30$

Frequency

x	10-20	20-30	30-50	50-80
frequency	22	8	31	9

- 22 numbers are between 10 and 20, 8 numbers are between 20 and 30, ……
- Within each class, we only know the range of those numbers, not the detail.
- total = $\sum f$ = 22 + 8 + 31 + 9 = 70.

Mean for Grouped Data

- $\bar{x} = \frac{\sum fx}{\sum f}$

x	12	23	34	42
frequency	5	11	8	6
fx	12×5=60	253	272	252

- $\sum fx = 60 + 253 + 272 + 252 = 837$, $\sum f = 5 + 11 + 8 + 6 = 30$
- $\bar{x} = \frac{\sum fx}{\sum f} = \frac{837}{30} = 27.9$

Estimate Mean

x	10-20	20-30	30-50	50-80
frequency	22	8	31	9

- Use the mid-class value to estimate the mean.

x	10-20	20-30	30-50	50-80
frequency	22	8	31	9
mid-class	(10+20)/2=15	(20+30)/2=25	40	65

- $\bar{x} = \frac{\sum fx}{\sum f} = \frac{22\times15+8\times25+31\times40+9\times65}{22+8+31+9} = \frac{471}{14} = 33.64$

Variance for Grouped Data

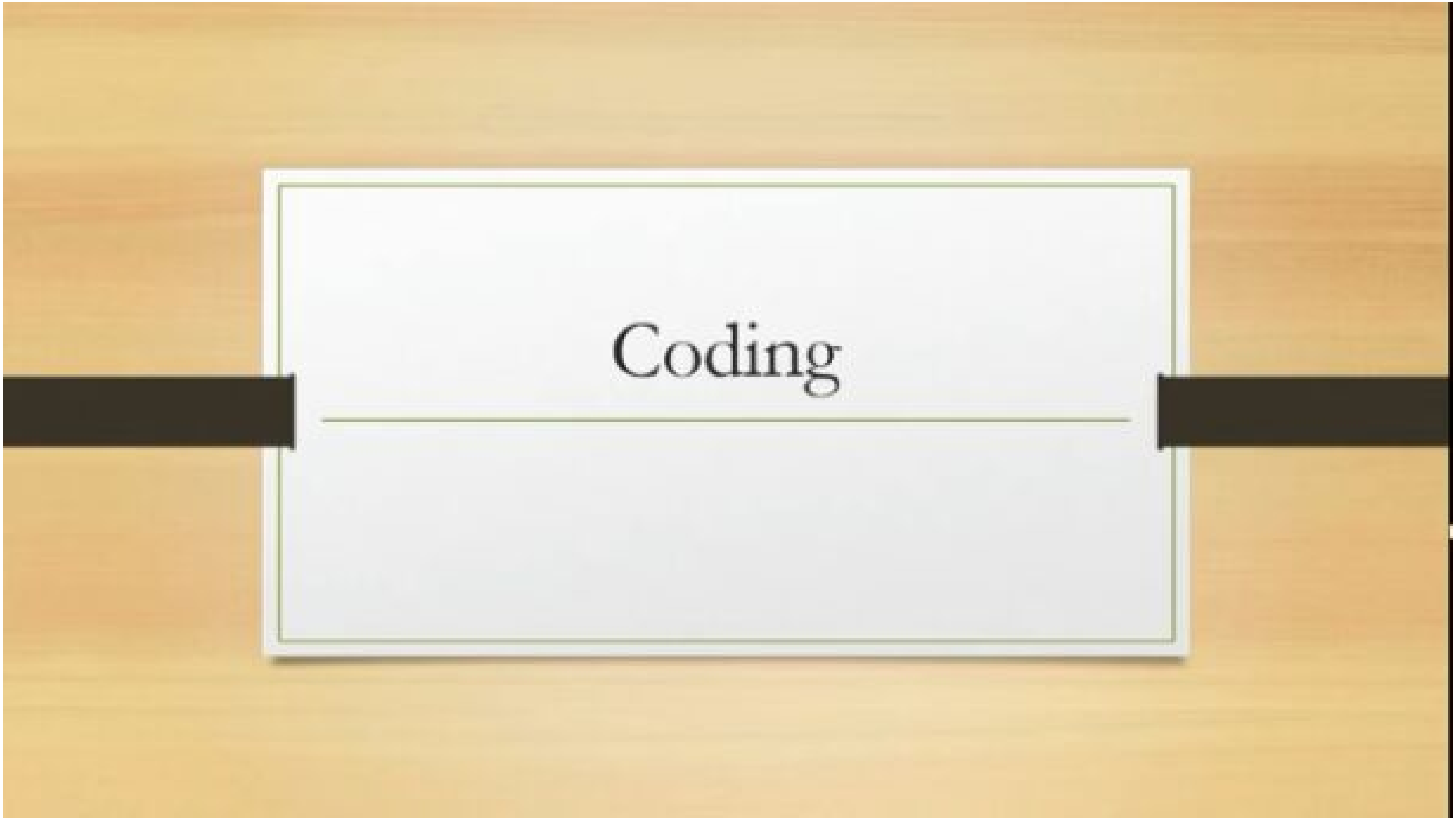
- $variance = \frac{\sum x^2 f}{\sum f} - \bar{x}^2$

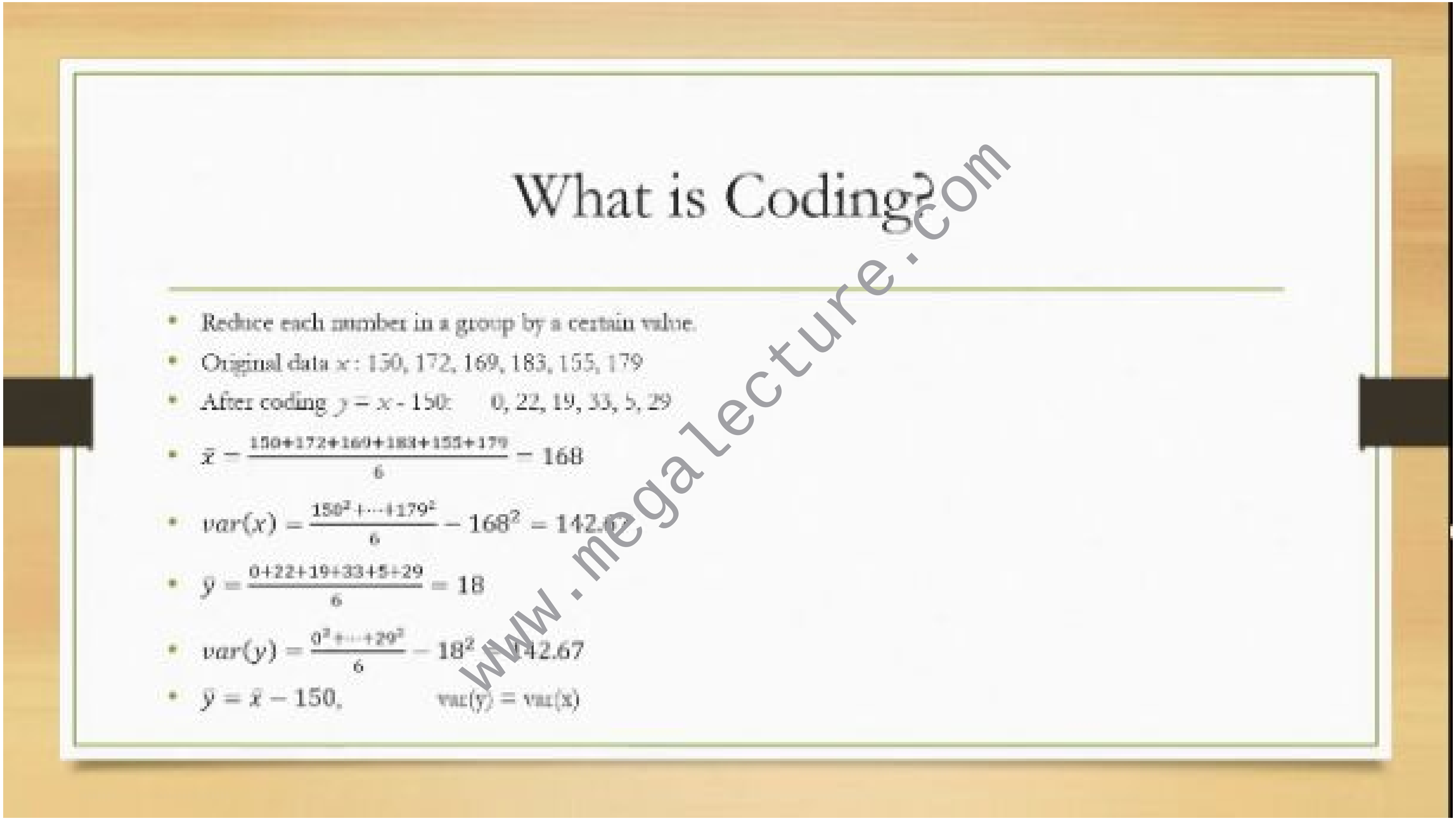
x	10-20	20-30	30-50	50-80
frequency	22	8	31	9
mid-class	15	25	40	65
x^2f	$15^2 \times 22 = 4950$	$25^2 \times 8 = 5000$	49600	38025

- $variance = \frac{4950+5000+49600+38025}{70} - \left(\frac{471}{14}\right)^2 = 261.09$

Summary

- $\bar{x} = \dfrac{\sum fx}{\sum f}$
- $variance = \dfrac{\sum x^2 f}{\sum f} - \bar{x}^2$
- Use mid-class value to estimate.


Coding


What is Coding?
Reduce each number in a group by a certain value.
Original data x : 150, 172, 169, 183, 155, 179
After coding $y = x - 150$: 0, 22, 19, 33, 5, 29
$\bar{x} = \frac{150+172+169+183+155+179}{6} = 168$
$var(x) = \frac{150^2+\cdots+179^2}{6} - 168^2 = 142.67$
$\bar{y} = \frac{0+22+19+33+5+29}{6} = 18$
$var(y) = \frac{0^2+\cdots+29^2}{6} - 18^2 = 142.67$
$\bar{y} = \bar{x} - 150,$ var(y) = var(x)

Facts about Coding

- If $y = x - a$, then $\overline{y} = \overline{x} - a$ and $\text{var}(y) = \text{var}(x)$
- Coding shifts all numbers to the left by a fixed distance, but the distance between each number is NOT changed. Therefore, the mean is shifted, but the variance and standard deviation remain the same.
- Coding problems are usually confusing and difficult to solve.
- Remember to introduce y and stick with the above two facts.

Example

The heights, x cm, of a group of 82 children are summarized as follows

$$\Sigma(x-130) = -287, \text{ standard deviation of } x = 6.9$$

(i) Find the mean height; (ii) Find $\Sigma(x-130)^2$.

n = 82. Let $y = x - 130$. Then $\Sigma y = -287$, $SD(y) = SD(x) = 6.9$

(i) $\bar{y} = \Sigma y/n = -287/82 = -3.5$. Since $\bar{y} = \bar{x} - a$, $\bar{x} = -3.5+130 = 126.5$

(ii) $var(y) = \frac{\sum y^2}{n} - \bar{y}^2 = \frac{\Sigma(x-130)^2}{82} - (-3.5)^2 = 6.9^2$ So, $\Sigma(x-130)^2 = 4908.52$


Histogram

Frequency Density

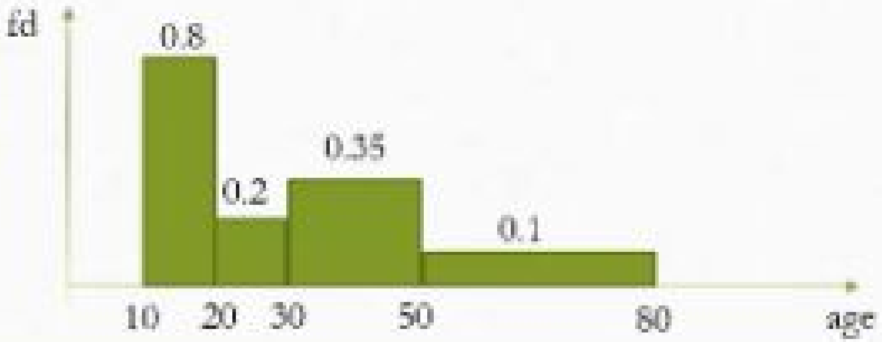
age	10-20	20-30	30-50	50-80
frequency	8	2	7	3
class width	20 10=10	10	20	30
f.d.	8/10 = 0.8	2/10 = 0.2	7/20 = 0.35	3/30 = 0.1

- Frequency Density (FD) = frequency / class width

Histogram

- Each class is represented by a bar. The height is its frequency density.
- Each bar's **area = frequency** = width * height
- The y axis is always frequency density.

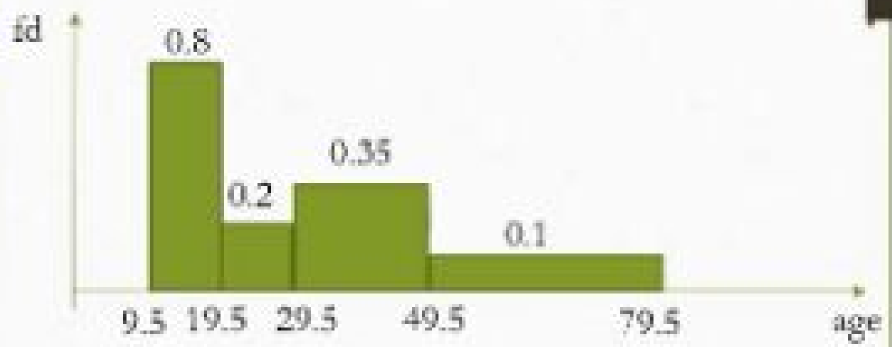
age	10-20	20-30	30-50	50-80
f.d.	0.8	0.2	0.35	0.1



No Gap

- If there are gaps between classes, we should move the boundary to the middle.

age	10-19	20-29	30-49	50-79
modified	9.5-19.5	19.5-29.5	29.5-49.5	49.5-79.5
class width	10	10	20	30
frequency	8	2	7	3
f.d.	0.8	0.2	0.35	0.1



Summary

- frequency density = frequency / class width
- The y-axis is always frequency density.
- There should be no gaps between classes. Move the boundary to the middle if original classes have gaps.
- area = frequency
- You should be able to draw the frequency table from the histogram also.


Cumulative Frequency
Diagram

Cumulative Frequency

- Think of it as a running total.

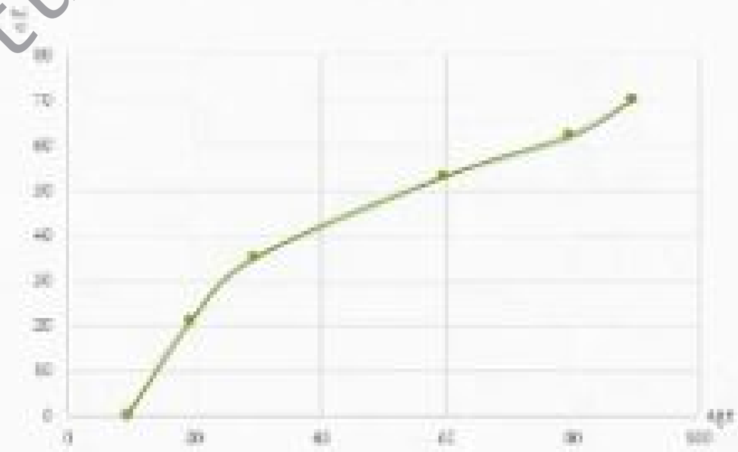
age	10-19	20-29	30-59	60-79	80-89
frequency	21	14	18	9	8
c.f.	21	35	53	62	70

- Each cumulative frequency represents the number of items less than the upper boundary of this class.

age	9.5-19.5	19.5-29.5	29.5-59.5	59.5-79.5	79.5-89.5
range	<=19.5	<=29.5	<=59.5	<=79.5	<=89.5
c.f.	21	35	53	62	70

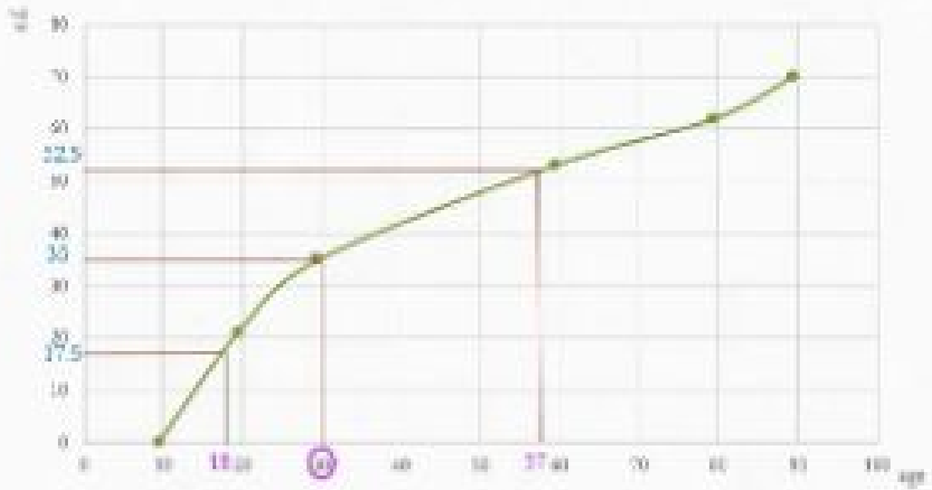
age	9.5-19.5		19.5-29.5	29.5-59.5	59.5-79.5	79.5-89.5
c.f.	21		35	53	62	70
point	(9.5, 0)	(19.5, 21)	(29.5, 35)	(59.5, 53)	(79.5, 62)	(89.5, 70)

- The y-axis is always c.f.
- The 1st point is (lowest boundary, 0).
- Other points are (upper boundary, c.f.).
- Connect all points with **a smooth curve**.
- The curve **should always increase or rise**, because c.f. never goes down.

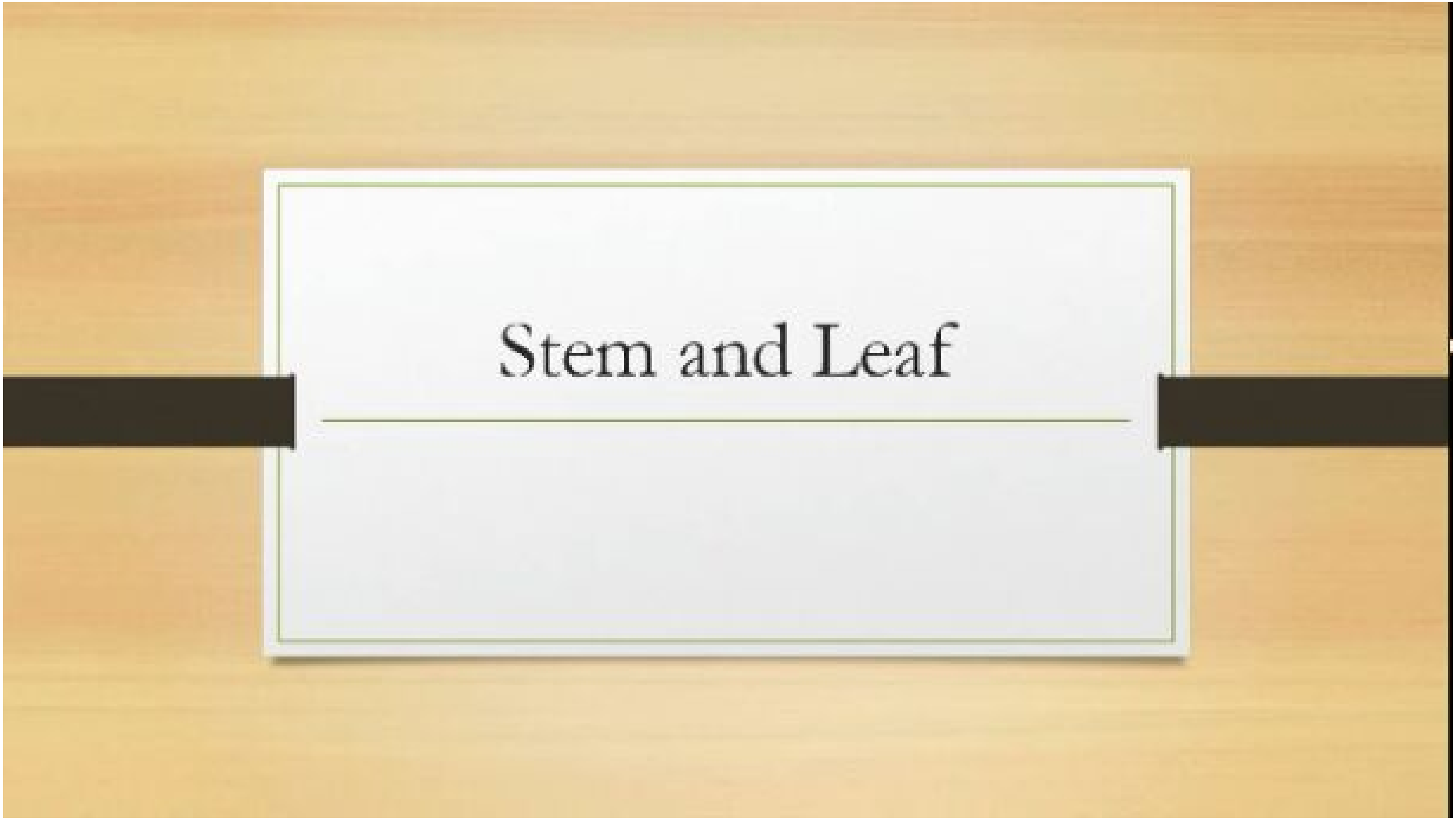


MEGA LECTURE

Estimate Median and Quartiles



Estimate Median and Quartiles


Stem and Leaf

Stem and Leaf

- Sort the numbers.
- Leaf is the last digit. Stem is the remaining.
- Remember to put in the **key**.
- Girls' height in cm: 152, 153, 153, 162, 165, 167, 167, 170, 171, 175

	girls' height
15	2 3 3
16	2 5 7 7
17	0 1 5

key: 15 | 3 means a girl's height is 153 cm

Back-to-Back Stem and Leaf

- Girls' height in cm: 152, 153, 153, 162, 165, 167, 167, 170, 171, 175
- Boys' height in cm: 159, 165, 167, 169, 173, 175, 180, 182

boys' height		girls' height
9	15	2 3 3
9 7 5	16	2 5 7 7
5 3	17	0 1 5
2 0	18	

key: 5|16|2 means boy's height is 165cm and girl's height is 162 cm.

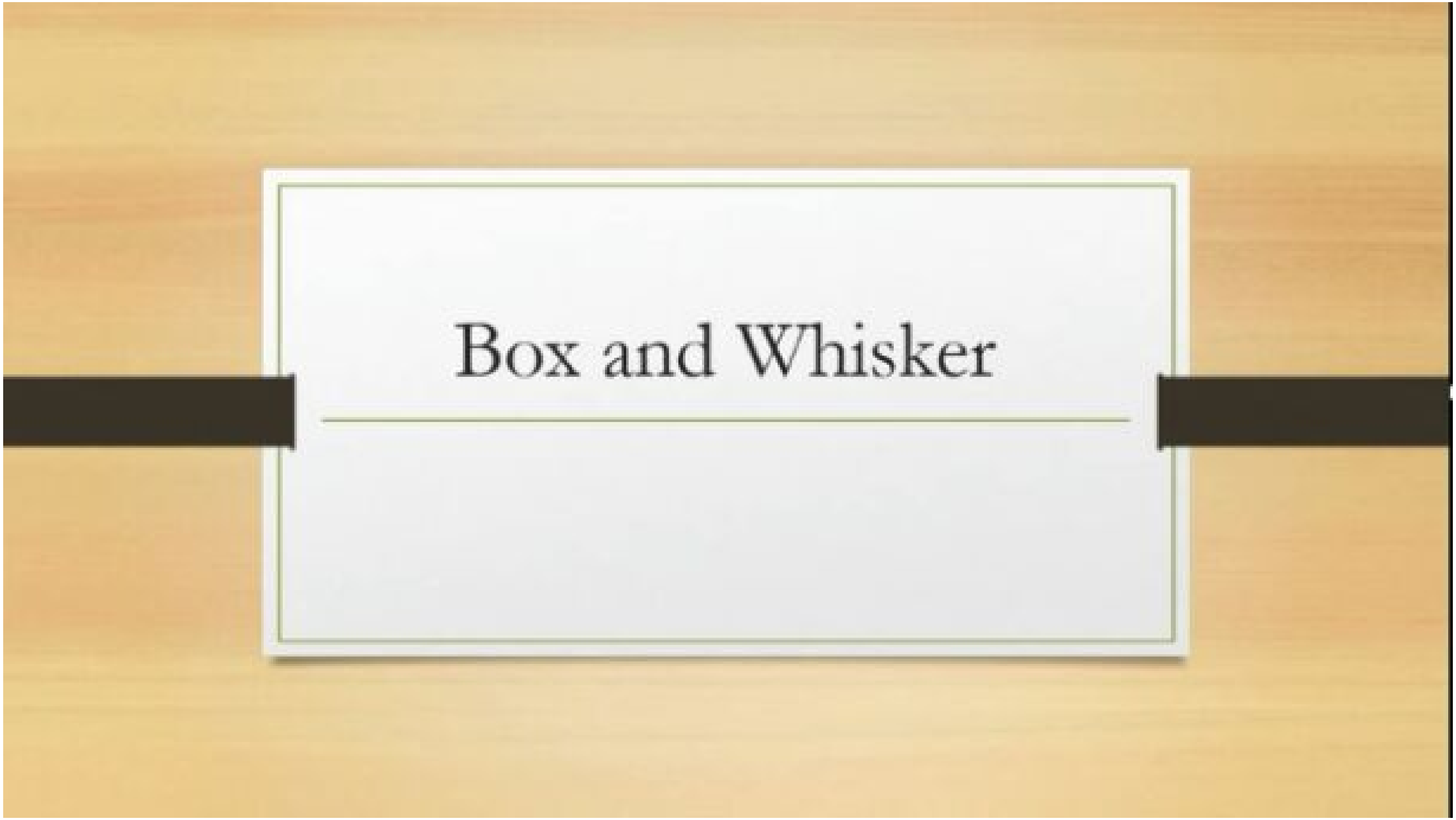
Find Median and Quartiles

- Need to consider if n is odd or even.
- Count to the correct number.

	girls' height
15	2 3 3
16	2 5 7 7
17	0 1 5

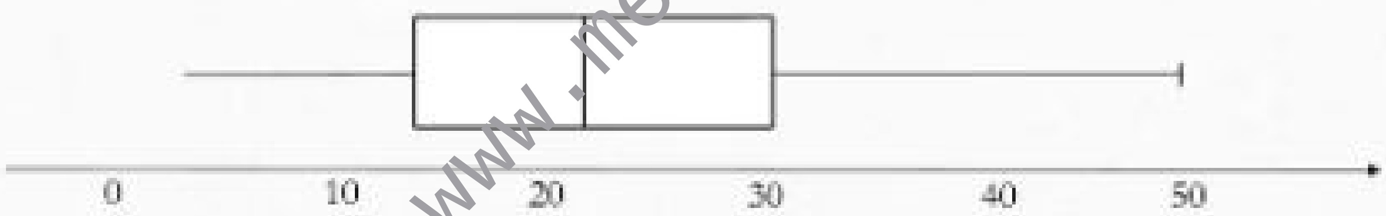
key: 15 | 3 means a girl's height is 153 cm

n =10, median = (165 + 167) / 2 = 166, LQ = 153, UQ = 170.


Box and Whisker

Box and Whisker

- 5 major points: minimum, LQ, median, UQ, maximum
- 3, 9, 13, 18, 20, 22, 29, 30, 40, 50
- min = 3, LQ = 13, median = 21, UQ = 30, max = 50



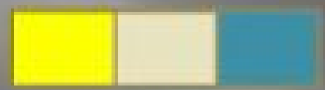

PERMUTATION
AND
COMBINATION

Introduction

- Permutation ${}^{n}P_{r}$: *arrange* r items out of n different items. *Order matters.*
- Combination ${}^{n}C_{r}$ or $\binom{n}{r}$: *select* r items out of n different items. *Order does NOT matter.*
- Example:
  - Permutation: 5 out of 10 people form a line.
  - Combination: Pick 3 out of 20 people to form a committee.

Permutation

- How many ways are there to pick 3 students out of 20 and put them in a line?



  - How many choices for ?
  - And then, how many choices for ?
  - And then, how many choices for ?
- Answer: ${}^{20}P_3 = 20 \times 19 \times 18$

Formulae for Permutation

- $^{20}P_3 = 20 \times 19 \times 18$
- $^nP_r = n(n-1)(n-2) \ldots\ldots (n-r+1)$
- $= n(n-1)\ldots(n-r+1)\dfrac{(n-r)(n-r-1)\ldots\times 2\times 1}{(n-r)(n-r-1)\ldots\times 2\times 1}$
- $= \dfrac{n!}{(n-r)!}$
- $^nP_r = n(n-1)(n-2) \ldots\ldots (n-r+1) = \dfrac{n!}{(n-r)!}$

Combination

- Take 2 numbers out of 1, 2, 3

Combination	Permutation	
1, 2	1, 2	2, 1
2, 3	2, 3	3, 2
1, 3	1, 3	3, 1

- For each case of combination, there are 2! cases in permutation.
- ${}^{3}P_{2} = 2!\ {}^{3}C_{2}$
- ${}^{n}P_{r} = r!\ {}^{n}C_{r}$

Formula for Permutation and Combination

- ${}^nP_r = r!\,{}^nC_r$
- ${}^nP_r = n(n-1)(n-2)\ \ldots\ldots\ (n-r+1) = \dfrac{n!}{(n-r)!}$
- ${}^nC_r = \dfrac{n(n-1)(n-2)\ldots(n-r+1)}{1\times 2\times \ldots \times r} = \dfrac{n!}{r!(n-r)!}$
- ${}^nP_n = n!$
- ${}^nC_0 = {}^nC_n = 1$

Permutation with Limitations

- How many numbers can you form with 1, 2, 3, 5, 6, 7, 9 (no numbers can be used more than once), if
  - they are 4-digit odd numbers?
    - Consider the last digit. Only 1, 3, 5, 7 or 9.
    - Consider the rest digits. Any limitations?
    - Answer: $5 \times {}^{6}P_{3}$
  - they are 4-digit and less than 6000?
    - First digit can only be 1, 2, 3 or 5. Answer: $4 \times {}^{6}P_{3}$
  - they are less than 6000?
    - 1-, 2- and 3-digit are all < 6000.
    - Answer: ${}^{7}P_{1} + {}^{7}P_{2} + {}^{7}P_{3} + 4 \times {}^{6}P_{3}$

Permutation with Duplicates

- Find the number of distinct permutations of the letters of the word MISSISSIPPI.
  - 11 letters in total
  - 4 Ss, 4 Is, 2 Ps, 1 M
  - If we do 11!, it treats all letters differently.
  - $\frac{11!}{4! \times 4! \times 2!}$

Permutation
All Together or All Separate

- Find the number of ways 6 women and 3 men to stand in a row so that all 3 men are standing together.
  - Treat all 3 men as **one** person. This gives 7!
  - Within the 3 men, there are 3! permutations.
  - Answer: 7! 3! = 30240
- Find the number of ways 6 women and 3 men to stand in a row so that no two men are standing next to one another
  - Arrange 6 women first. This gives 6!
  - How many spaces does this create?
  - Put each of the 3 men in one of the spaces. $^7P_3$
  - Answer: 6! $^7P_3$


COMBINATION

Combination with Limitations

- A team of 5 people, which must contain 3 men and 2 women, is chosen from 8 men and 7 women. How many different teams can be selected?
  - First, select men. How many ways?
  - Then, select women. How many ways?
  - Answer: ${}^{8}C_{3} \times {}^{7}C_{2}$

Not Both

- A team of 5 is chosen from 15 people. How many different teams can be selected if two particular people cannot be both in the team?
- Not Both = All – Both
- All: $^{15}C_5$
- Both: $^{13}C^3$
- Answer: $^{15}C_5 - {}^{13}C_3$

Combination with Duplicates

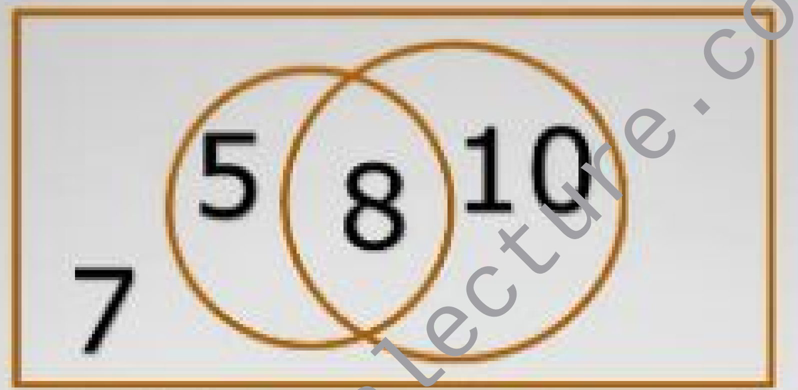
- Four letters are to be selected from the letters in the word RIGIDITY. How many different combinations are there?
  - 3 Is are all selected: ${}^5C_1$
  - Only 2 Is are selected : ${}^5C_2$
  - Only 1 I is selected : ${}^5C_3$
  - No I : ${}^5C_4$

Probability with Venn Diagram

- Probability of an event A:
  P(A) = # of possibilities in A / # of all possibilities
- A class has 14 boys and 16 girls. What's the probability of selecting a boy from this class?
- There are 14 ways to select a boy.
- There are 30 way to select a student.
- P(boy) = 14/30 = 7/15

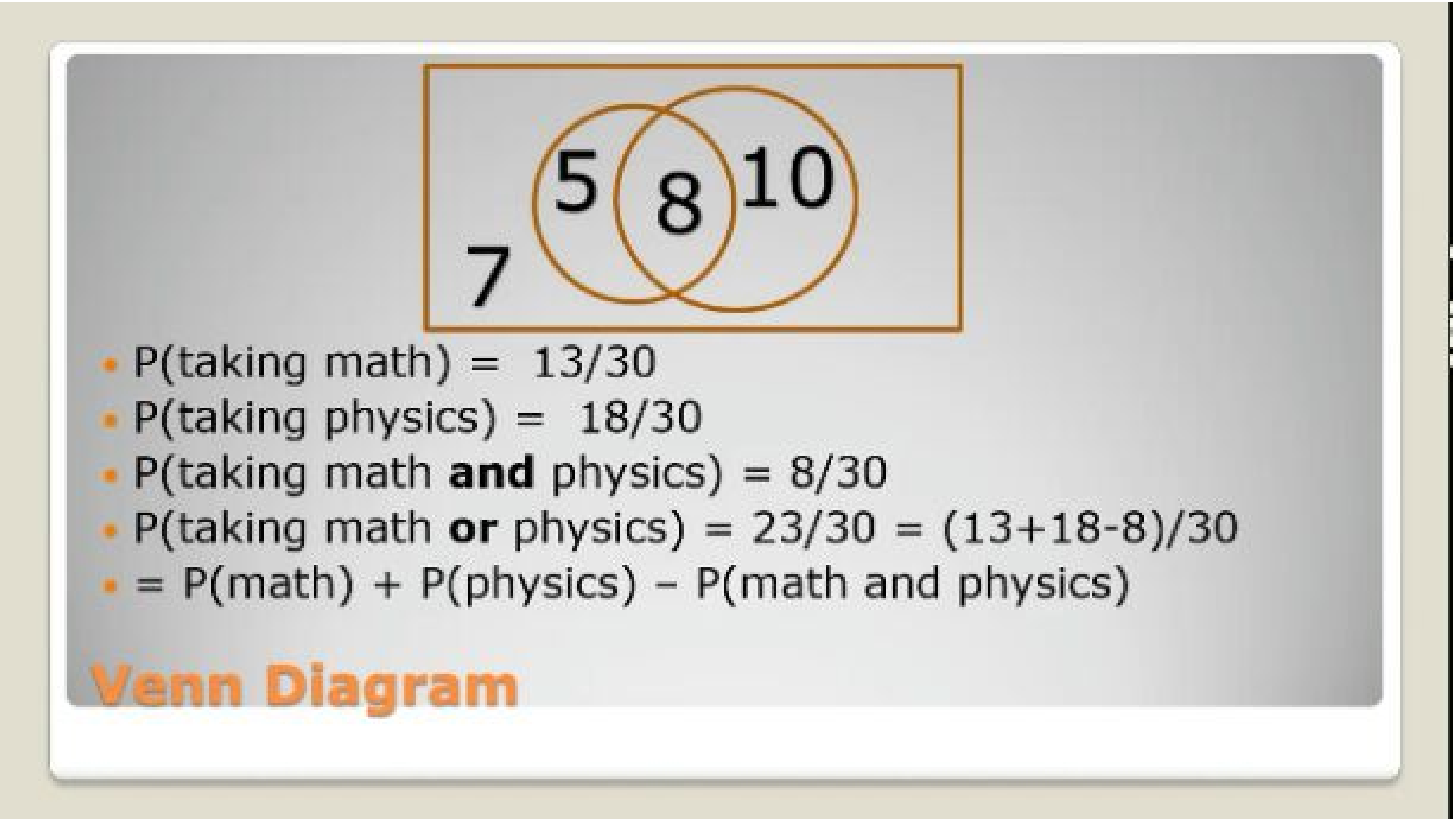
Definition

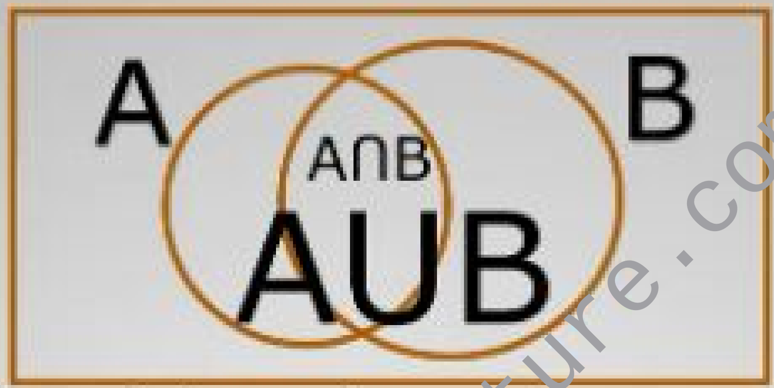
- In a class of 30, 13 students take math, 18 students take physics and 8 students take both.



- What is the probability of a student taking neither math nor physics?
- 7/30

Venn Diagram


5
8
10
7
• P(taking math) = 13/30
• P(taking physics) = 18/30
• P(taking math **and** physics) = 8/30
• P(taking math **or** physics) = 23/30 = (13+18-8)/30
• = P(math) + P(physics) – P(math and physics)
Venn Diagram



- P(A): left circle, P(B): right circle
- P(A or B), P(A U B): the area covered by both circles
- P(A and B), P(A ∩ B): the common area between two circles
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Venn Diagram

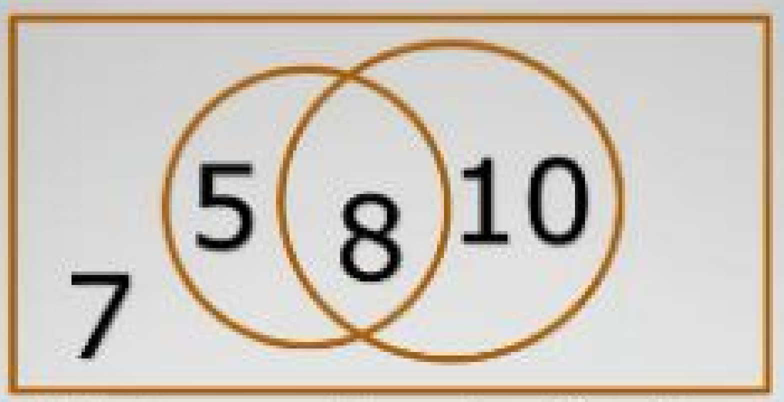
Exclusive Events

- A and B are exclusive events if A and B cannot happen at the same time.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- A and B are exclusive if
  - $P(A \cap B) = 0$

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


Conditional Probability

- In a class of 30, 13 students take math, 18 students take physics and 8 students take both.

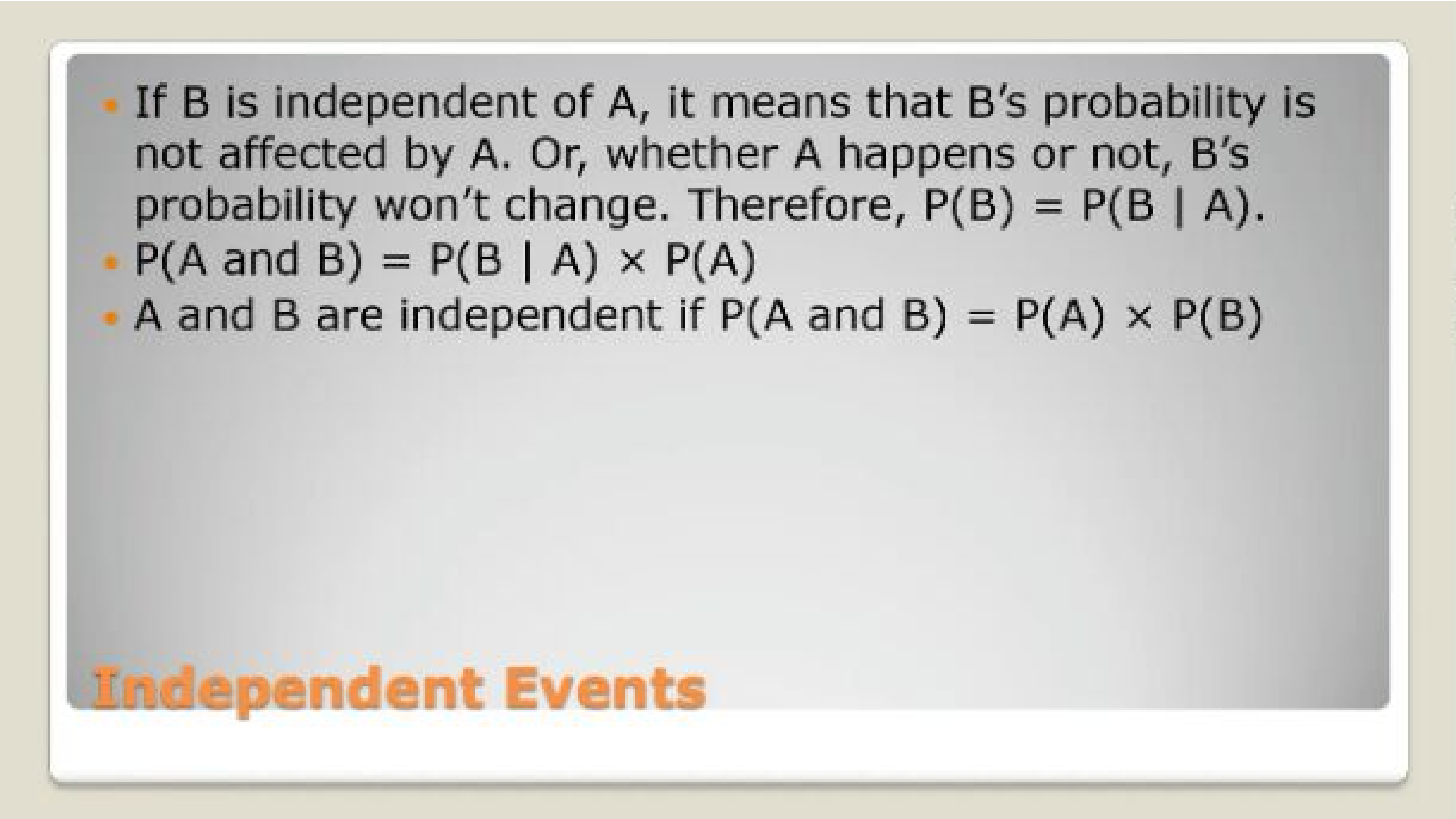


- **Given** a student takes math, what is the probability that student takes physics?
- P(physics | math) = 8/13
- = P(physics and math) / P(math)

Conditional Probability

Conditional Probability

- P(B | A) = P(A and B) / P(A)
- P(A and B) = P(B | A) × P(A)
- In Europe, 88% of all households have a television. 51% of all households have a television and a VCR. What is the probability that a household has a VCR given that it has a television?
- P(T) = 0.88, P(T and V) = 0.51, P(V | T) = ?
- P(V | T) = P(T and V) / P(T) = 0.51/0.88 = 51/88


If B is independent of A, it means that B's probability is not affected by A. Or, whether A happens or not, B's probability won't change. Therefore, P(B) = P(B | A).
P(A and B) = P(B | A) × P(A)
A and B are independent if P(A and B) = P(A) × P(B)
Independent Events

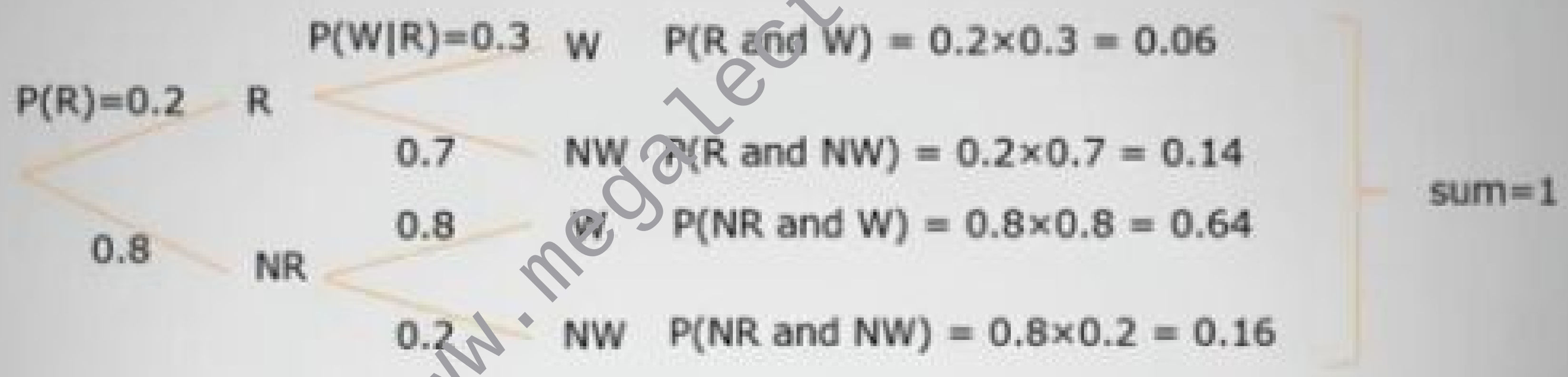
Exclusive and Independent Events

- A fair die is thrown. Event A is the number is odd. Event B is the number is less than 3. Are A and B independent? Are A and B exclusive?
- A: 1, 3, 5. P(A) = 1/2.
- B: 1, 2. P(B) = 1/3.
- A and B: 1. P(A and B)= 1/6 = P(A) × P(B).
- So A and B are independent, but not exclusive.

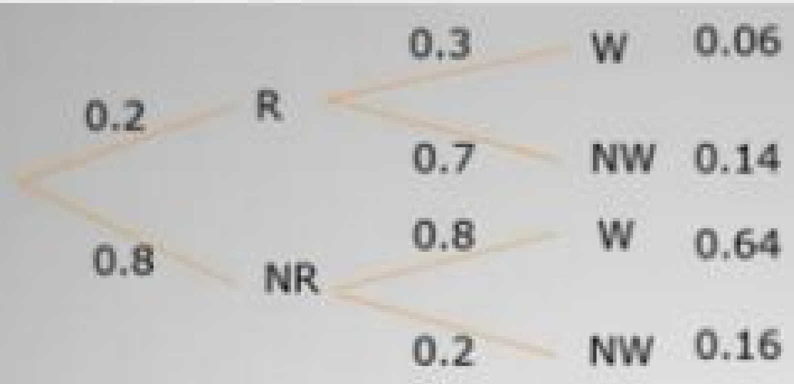

Tree Diagram

Tree Diagram

- The probability of rain is 0.2. If it rains, the probability that Tom walks his dog is 0.3. If it doesn't, 0.8. Draw a tree diagram to represent this.
- Nodes are events. Probability is written on the branch.
- Probabilities coming out of one node sum up to 1.



Conditional Probability



- What's the probability that Tom walks his dog?
- 0.06 + 0.64 = 0.7
- Given Tom walks his dog, what's the probability that it rains?
- P(R|W) = P(W and R) / P(W) = 0.06 / 0.7 = 3/35.


Object Picking Probability

- A bag contains 4 red balls, 5 green balls. Two balls are taken out. What is the probability that there are:
- 1 red and 1 green
  - Use Combination: ${}^4C_1{}^5C_1/{}^9C_2 = 5/9$
  - Take one red out: 4/9
  - Take one green out: 5/8
  - Probability = 4/9 x 5/8 = 5/18 ≠ **5/9**
  - What's wrong? R/G and G/R. Therefore 5/18×**2!** = 5/9
- 2 green balls
  - Use Combination: ${}^5C_2/{}^9C_2 = 5/18$
  - 5/9 x 4/8 = 5/18   No order any more!

Object Picking Probability

Object Picking Probability

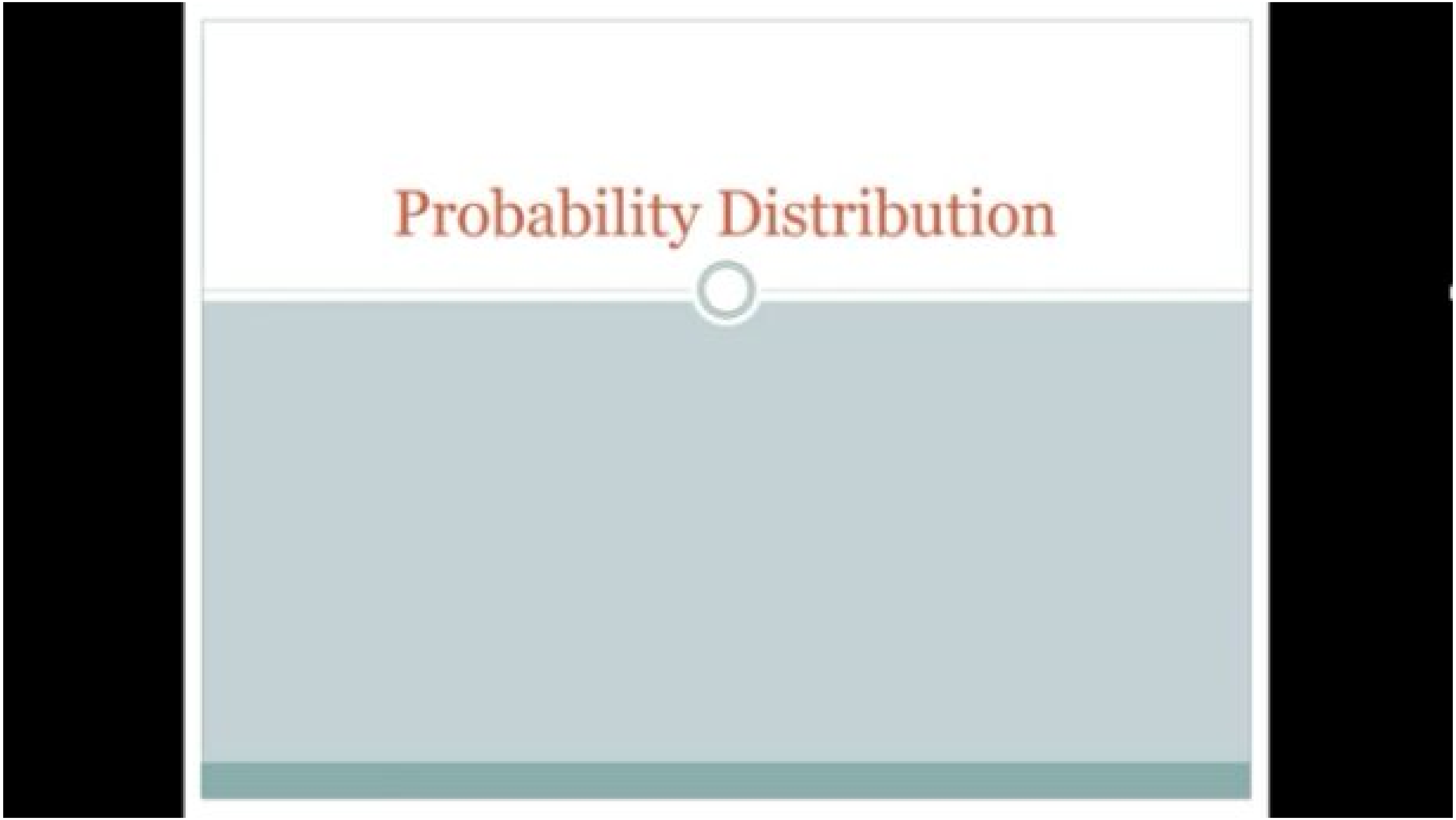
- A bag contains 4 red balls, 5 green balls. Two balls are taken out. What is the probability that there are:
- 1 red and 1 green
  - Use Combination: ${}^{4}C_{1}{}^{5}C_{1}/{}^{9}C_{2} = 5/9$
  - Take one red out: 4/9
  - Take one green out: 5/8
  - Probability = 4/9 x 5/8 = 5/18 ≠ **5/9**
  - What's wrong? R/G and G/R. Therefore 5/18×**2!** = 5/9
- 2 green balls
  - Use Combination: ${}^{5}C_{2}/{}^{9}C_{2} = 5/18$
  - 5/9 x 4/8 = 5/18 No order any more!

Object Picking Probability

- A bag contains 3 red balls, 4 green balls, 5 yellow balls. 3 balls are taken out. What's the probability that there are:
- 1 ball of each color
  - ${}^{3}C_{1}{}^{4}C_{1}{}^{5}C_{1}/{}^{12}C_{3} = 3/11$
  - $3/12 \times 4/11 \times 5/10 \times \mathbf{3!} = 3/11$
- exactly 2 green balls
  - ${}^{4}C_{2}{}^{8}C_{1}/{}^{12}C_{3} = 12/55$
  - $4/12 \times 3/11 \times 8/10 \times \mathbf{3} = 12/55$
- at least 1 yellow ball
  - $1 - \text{none} = 1 - {}^{7}C_{3}/{}^{12}C_{3} = 37/44$

Object Picking Probability

- A bag contains 3 red balls, 4 green balls, 5 yellow balls. 3 balls are taken out. What's the probability that there are:
- 1 ball of each color
  - $^3C_1{}^4C_1{}^5C_1/{}^{12}C_3 = 3/11$
  - $3/12 \times 4/11 \times 5/10 \times \mathbf{3!} = 3/11$
- exactly 2 green balls
  - $^4C_2{}^8C_1/{}^{12}C_3 = 12/55$
  - $4/12 \times 3/11 \times 8/10 \times \mathbf{3} = 12/55$
- at least 1 yellow ball
  - $1 - \text{none} = 1 - {}^7C_3/{}^{12}C_3 = 37/44$
- **Combination is recommended!**


Probability Distribution

Terminology

- A **random variable** is a quantity whose value depends on a chance.
- A random variable can be **discrete** or **continuous.**
- The **probability distribution** of a discrete random variable is a listing of the possible values of the variable and the corresponding probabilities.
- X: the number you get when throwing a die
- x: the possible numbers

x	1	2	3	4	5	6
$P(X=x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Properties

- $\sum P(X = x) = 1$
- frequency $\approx$ total frequency $\times$ probability
- Throw a die 600 times. There are about 100 times that it lands on 3.

Example

- Take 2 balls out of a bag that contains 4 red balls and 5 green balls. Draw the probability distribution table for the number of green balls that are taken out.
- X: the number of green balls. x can be 0, 1, and 2.
- $P(0) = {}^4C_2 / {}^9C_2 = 6/36 = 3/18$
- $P(1) = {}^4C_1{}^5C_1 / {}^9C_2 = 20/36 = 5/9$
- $P(2) = {}^5C_2 / {}^9C_2 = 10/36 = 5/18$
- **Make sure their sum is 1.**

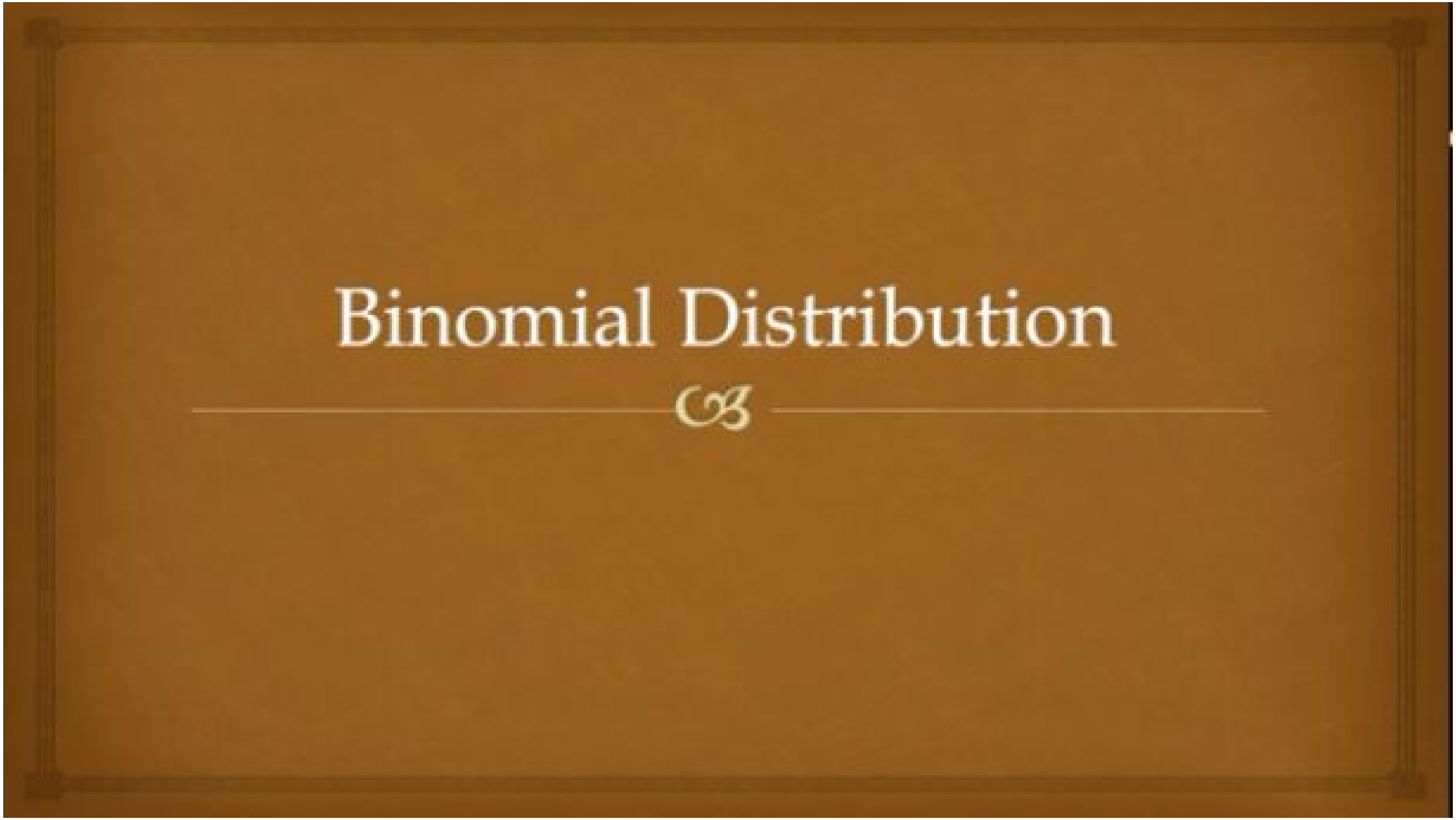
x	0	1	2
$P(X = x)$	3/18	5/9	5/18

Expectation and Variance

Expectation and Variance

x	0	1	2
$P(X = x)$	3/18	5/9	5/18

- Expectation: $\mu = E(X) = \sum xp$
- Variance: $\sigma^2 = Var(X) = \sum(x-\mu)^2 p = \sum x^2 p - \mu^2$
- $\mu = 0\times3/18 + 1\times5/9 + 2\times5/18 = 10/9$
- x^2: 0, 1, 4
- $\sum x^2 p = 0\times3/18 + 1\times5/9 + 4\times5/18 = 5/3$
- $\sigma^2 = 5/3 - (10/9)^2 = 35/81$


Binomial Distribution

A Dice Problem

- Throw 4 dice together. The random variable X is the number of 6's you get.
- X can be 0, 1, 2, 3, 4
- P(6) = 1/6, P(not 6) = 5/6
- $P(X=0) = (5/6)^4$, $P(X=4) = (1/6)^4$
- $P(X=1) = {}^4C_1\,(1/6)\,(5/6)^3$ because this one 6 can be any of the four
- $P(X=2) = {}^4C_2\,(1/6)^2\,(5/6)^2$ because these two 6's can appear anywhere in the four

A Dice Problem

- $P(X=0) = (5/6)^4$
- $P(X=1) = {}^4C_1\,(5/6)^3\,(1/6)$
- $P(X=2) = {}^4C_2\,(5/6)^2\,(1/6)^2$
- $P(X=3) = {}^4C_3\,(5/6)\,(1/6)^3$
- $P(X=4) = (1/6)^4$
- Why is the sum 1?
- Expand $(5/6 + 1/6)^4$ using the binomial theorem.
- The probability of success is $p=1/6$, $q = 1 - p = 5/6$
- $P(X=r) = {}^4C_r\,q^{4-r}\,p^r$

Binomial Distribution

- $X \sim B(n, p)$
- n: the random variable X can be from 0 to n
- p: the probability of success
- $P(X=r) = {}^{n}C_{r}\, p^{r} q^{n-r}$, $q = 1 - p$
- $P(X \leqslant r) = 1 - P(X > r)$
- $\leqslant$: at most, no more than $\geqslant$: at least, no less than

MEGA LECTURE

Example

- The fault rate of a product line is 0.02. Ten products are randomly picked from the product line. Find the probability that at least 2 are faulty.
- $p = 0.02,\ q = 1 - p = 0.98$
- $P(X \geq 2) = 1 - P(X<2) = 1 - P(0, 1) = 1 - 0.98^{10} - 10 \times 0.02 \times 0.98^{9} = 0.0162$

Four Conditions

- A single trial has exactly two possible outcomes (success p and failure q) and these are mutually exclusive ($p+q=1$).
- A fixed number of trials (n) takes place.
- The outcome of each trial is independent of the outcome of all the other trials. (*Probabilities are multiplied together.*)
- The probability of success at each trial is constant. (*p doesn't change.*)

MEGA LECTURE

Four Conditions

- Among 100 bottles, 7 are faulty. If 4 bottles are taken out, what's the probability of exactly 2 are faulty?
- Using binomial distribution, p=0.07, q=0.93
- $P(X=2) = {}^4C_2\ 0.07^2\ 0.93^2$
- **What is wrong here?**
- p is changing. First one, p = 7/100. If it's faulty, second one p = 6/99.
- Correct answer is: ${}^7C_2\ {}^{93}C_2\ /\ {}^{100}C_4$
- If it's a product line, the number of products can be considered unlimited/infinite. Therefore p is constant. We can use binomial.
- If the total number is fixed, we have to use combination.

Practical Questions

- In a certain product line, the faulty rate is 3%. Ten products are chosen at random. What is the probability that fewer than two of them are faulty?
- $X \sim B(10, 0.03)$
- $P(X < 2) = P(0, 1) = 0.97^{10} + 10 \times 0.03 \times 0.97^{9} = 0.9655$

Two-Level Questions

- The faulty rate of a bottle is 3%. Each box contains ten bottles. Eight boxes are chosen. Find the probability that exactly 7 boxes have less than 2 faulty bottles.
- X: number of faulty bottles in a box
- $X \sim B(10, 0.03)$
- $P(X<2) = 0.9655$
- Y: number of boxes that have < 2 faulty bottles
- $Y \sim B(8, 0.9655)$
- $P(Y=7) = {}^8C_7 \times 0.9655^7 \times (1 - 0.9655) = 0.216$

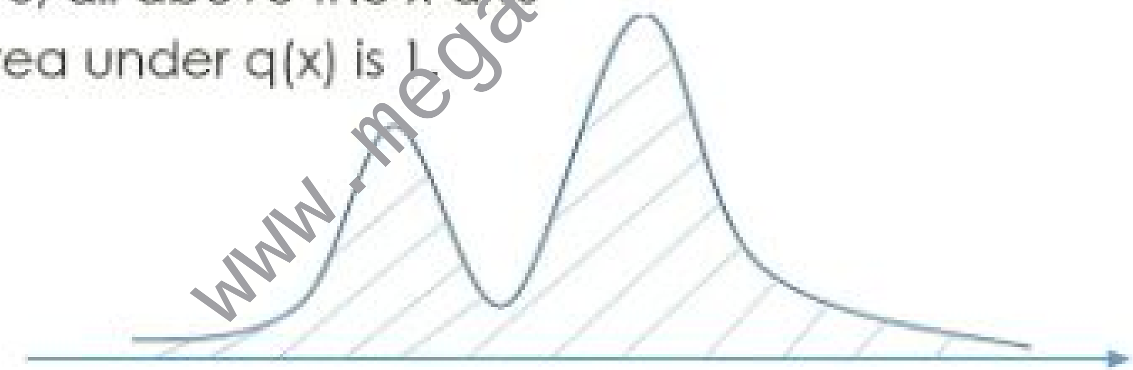
Expectation & Variance

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


Continuous
Probability
Distribution

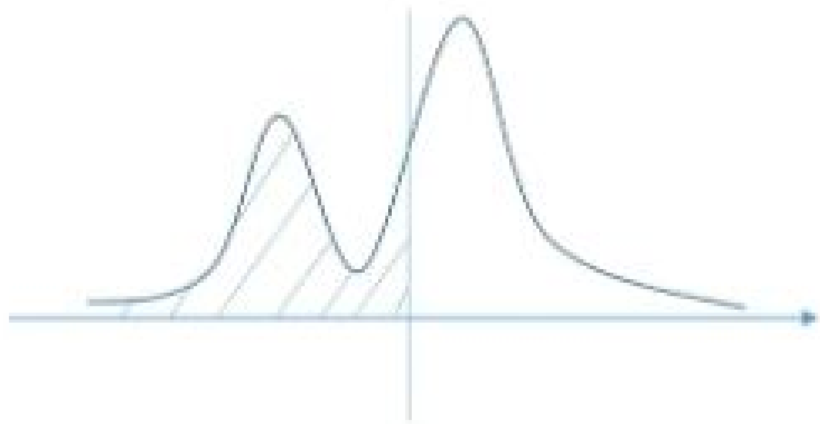
Continuous probability distribution

- In discrete probability distribution, probability is defined as P(X=x) and all probabilities add up to 1.
- Probability density function: q(x)
- q(x) > 0, all above the x-axis
- The area under q(x) is 1.



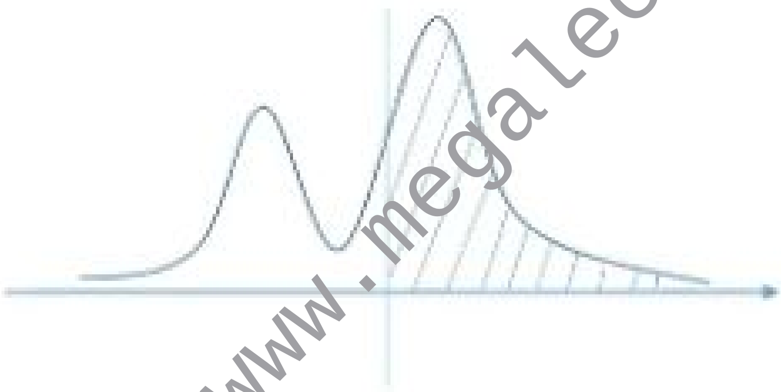
Continuous probability distribution

- P(X < x) is the area to the left of x.



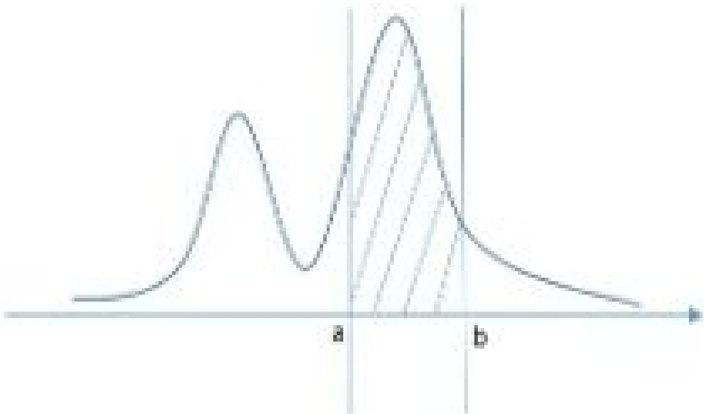
Continuous probability distribution

- $P(X > x)$ is the area to the right of x.
- $P(X > x) = 1 - P(X < x)$



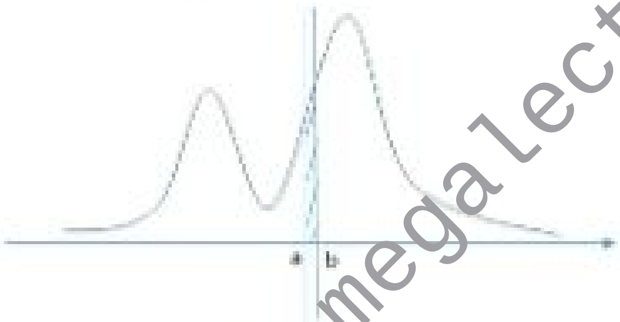
Continuous probability distribution

- P(a < X < b) is the area between a and b.
- P(a < X < b) = P(X < b) – P(X < a)



Continuous probability distribution

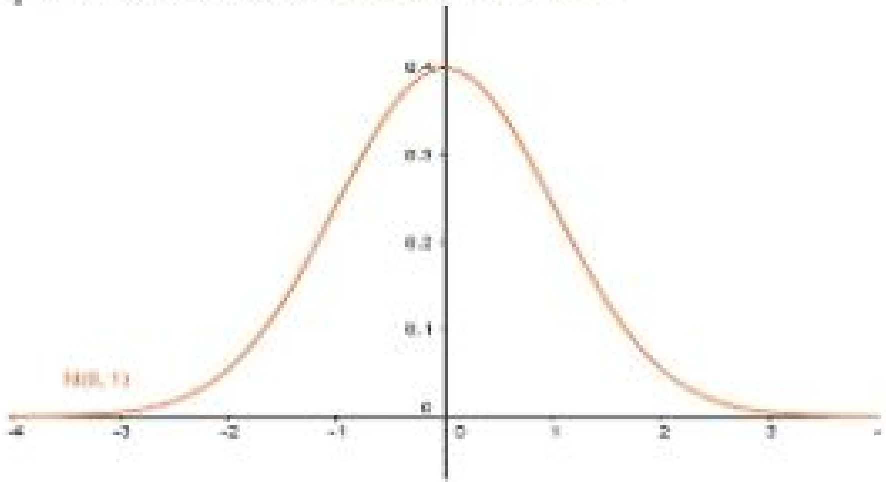
- If we draw a and b close enough, we get P(X = x) = 0



- In continuous probability distribution, P(X < x) = P(X ≤ x)
- Area = Probability

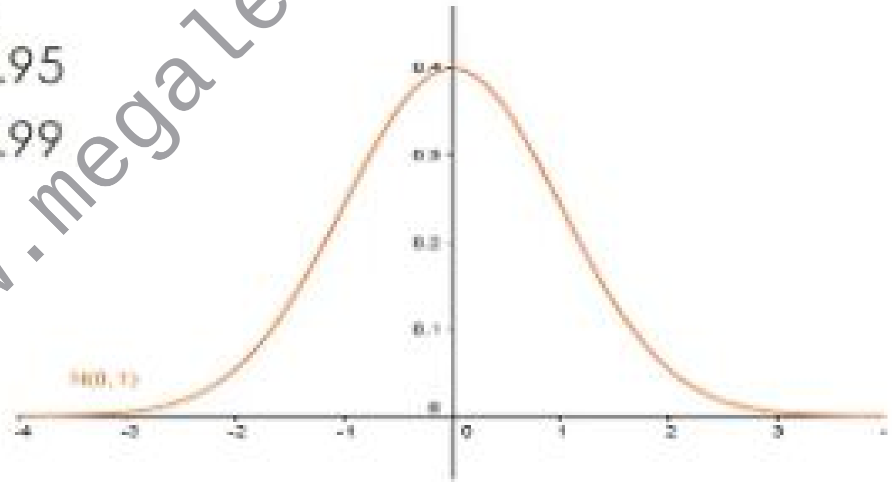
Normal Distribution

- X ~ N(μ, σ^2)
- Graph of probability density function is symmetrical about x= μ. Also called **bell curve**.



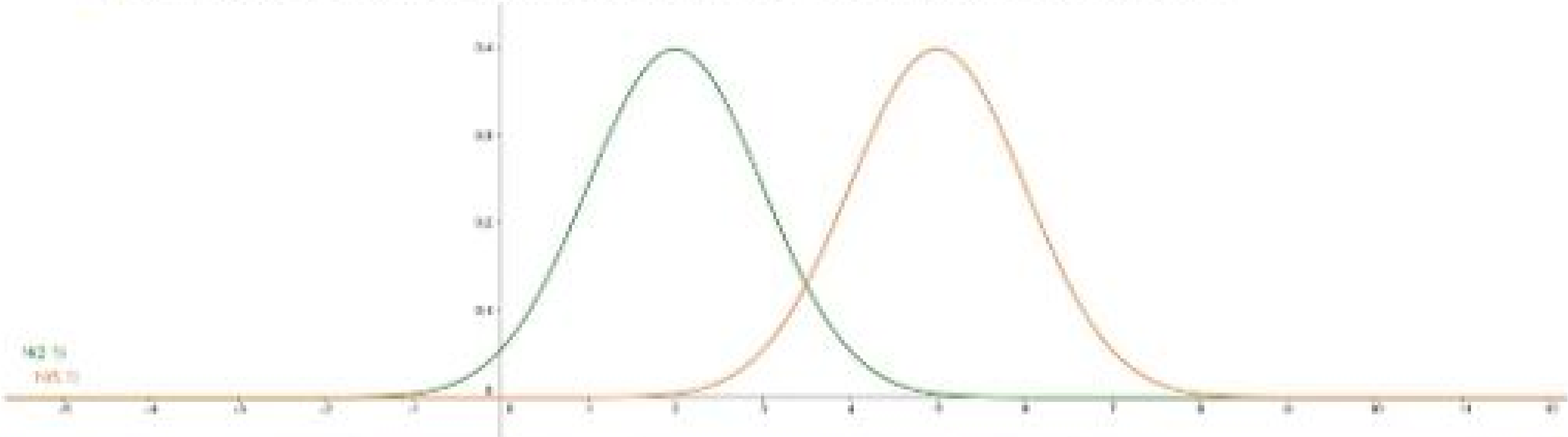
Normal Distribution

- Examples in the real world: exam scores, people's heights
- 2/3 values lie within σ, 95% lie within 2σ, 99% lie within 3σ
- $P(-\sigma < X < \sigma) = 2/3$
- $P(-2\sigma < X < 2\sigma) = 0.95$
- $P(-3\sigma < X < 3\sigma) = 0.99$



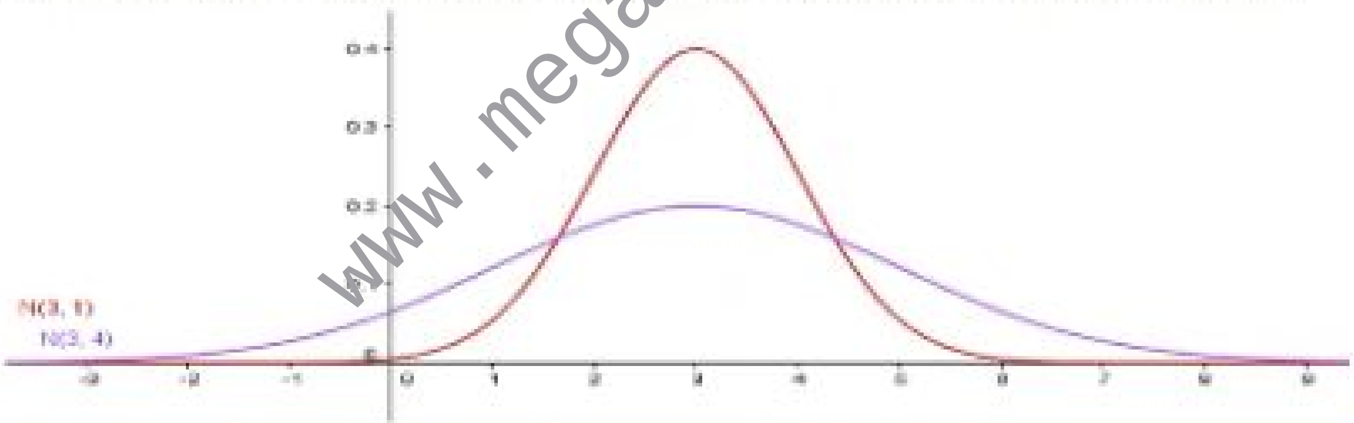
Graphs with different parameters

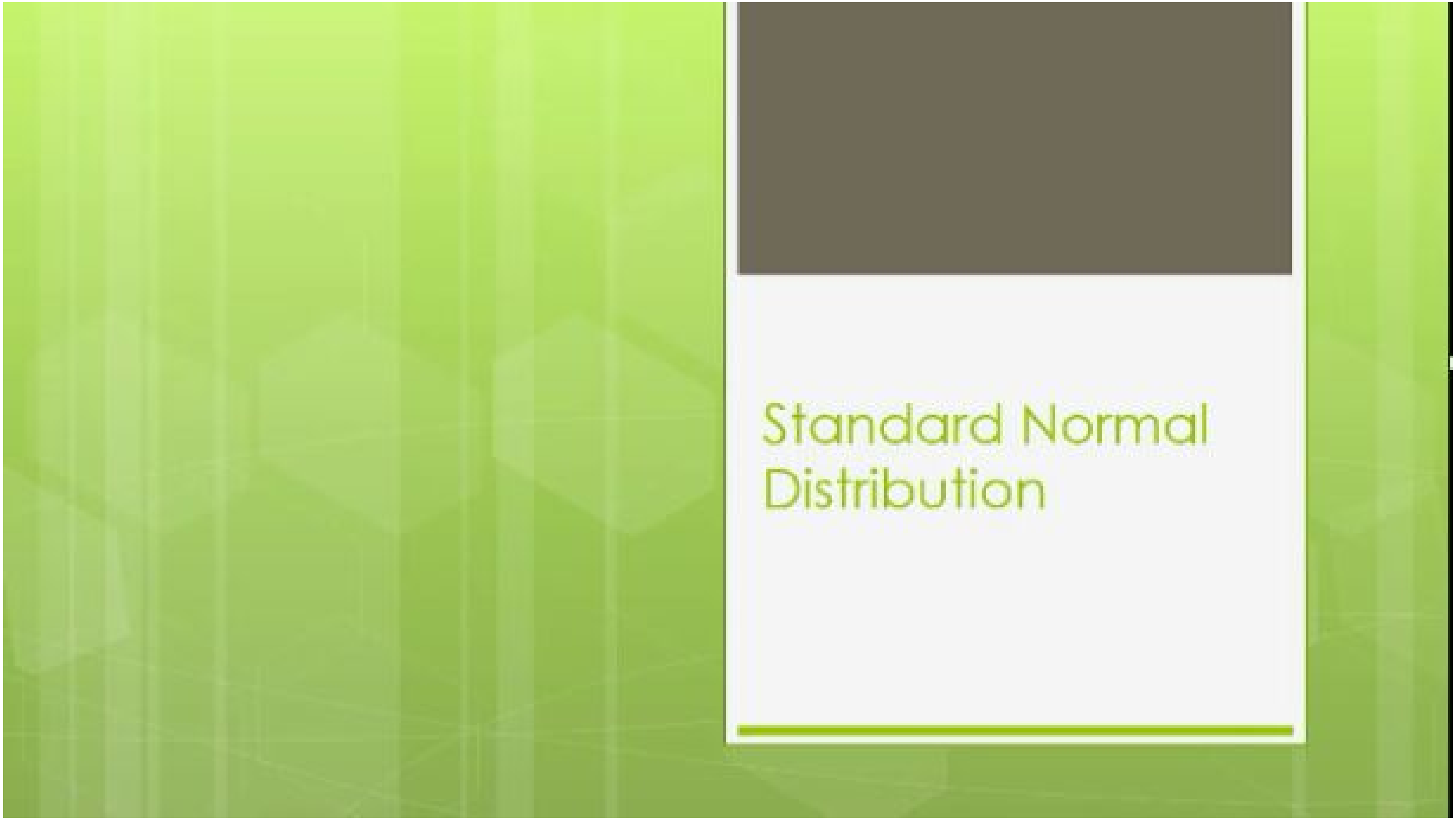
- μ defines center and σ defines the spread
- N(2, 1) and N(5, 1)
- Shapes are the same, only the center is different.



Graphs with different parameters

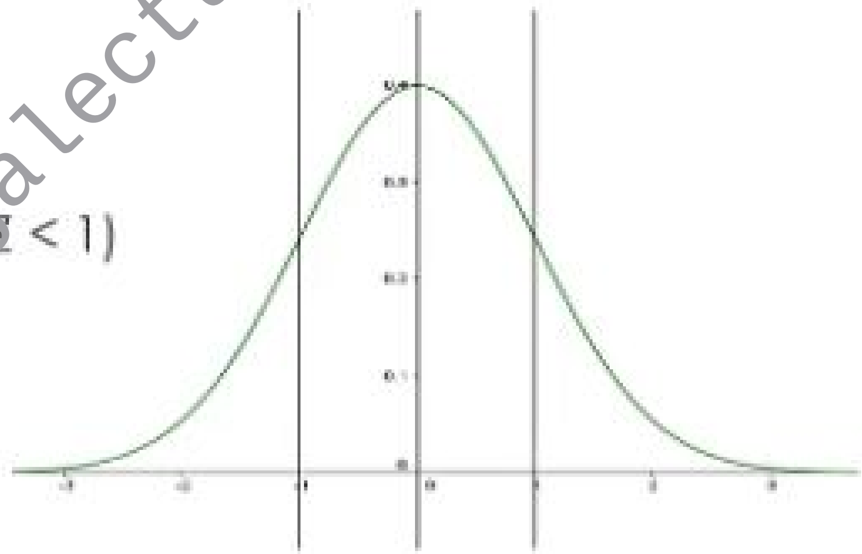
- σ defines the spread. Since the area underneath the curve is 1, the bigger σ, the lower and wider the graph.
- N(3, 1) and N(3, 4)
- Centers are the same, only N(3, 4) is lower and wider.




Standard Normal Distribution

Standard Normal Distribution

- Z: standard normal distribution
- $Z \sim N(0, 1)$
- $\Phi(z) = P(Z < z)$
- $P(Z < 1) = P(Z > -1)$
- $P(Z < -1) = P(Z > 1) = 1 - P(Z < 1)$
- $\Phi(-z) = 1 - \Phi(z)$



Look up the table

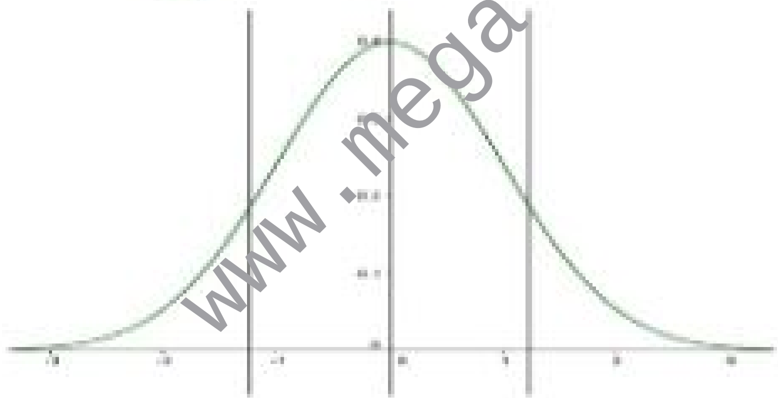
- Values in the table are $\Phi(z)$ for $z \geq 0$.
- $P(Z < 1.234) = \Phi(1.234) = 0.8907 + 0.0007 = 0.8914$

z	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
														ADD					
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359	4	8	12	16	20	24	28	32	36
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753	4	8	12	16	20	24	28	32	36
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141	4	8	12	15	19	23	27	31	35
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517	4	7	11	15	19	22	26	30	34
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879	4	7	11	14	18	22	25	29	32
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224	3	7	10	14	17	20	24	27	31
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549	3	7	10	13	16	19	23	26	29
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852	3	6	9	12	15	18	21	24	27
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133	3	5	8	11	14	16	19	22	25
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389	3	5	8	10	13	15	18	20	23
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621	2	5	7	9	12	14	16	19	21
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830	2	4	6	8	10	12	14	16	18
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015	2	4	6	7	9	11	13	15	17
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177	2	3	5	6	8	10	11	13	14
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319	1	3	4	6	7	8	10	11	13
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441	1	2	4	5	6	7	8	10	11

- $P(Z \leq 1.234) = P(Z < 1.234)$

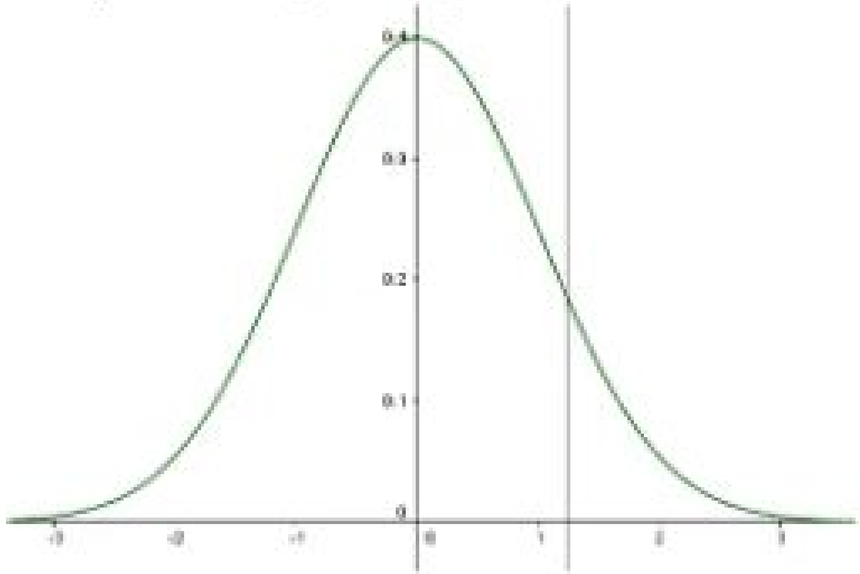
Look up the table

- The table only has non negative numbers for z.
- Use symmetry to get other values.
- $P(Z < -1.234) = \Phi(-1.234) = 1 - \Phi(1.234) = 1 - 0.8914 = 0.1086$



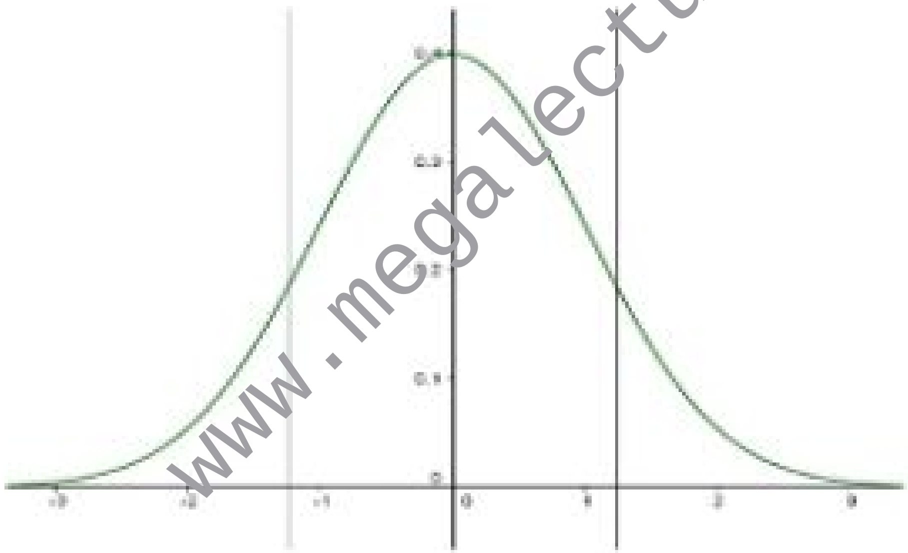
Look up the table

- $P(Z > 1.234) = 1 - \Phi(1.234) = 1 - 0.8914 = 0.1086$



Look up the table

- $P(Z > -1.234) = \Phi(1.234) = 0.8914$



MEGA LECTURE

Look up the table

- $P(Z < z)$ for $z \geq 0$ can be found in the table
- $P(Z > z) = 1 - P(Z < z)$
- $P(Z < -z) = 1 - P(Z < z)$
- $P(Z > -z) = P(Z < z)$
- Always helpful to sketch the curve.

Look up the table

- $P(0.7 < Z < 1.4)$
- $= P(Z < 1.4) - P(Z < 0.7)$
- Both values can be found directly in the table.
- $P(-1.4 < Z < 1)$
- $= P(Z < 1) - P(Z < -1.4)$
- $= P(Z < 1) - [\ 1 - P(Z < 1.4)\ ]$
- Now both values can be found in the table.


Reverse Lookup

Reverse Lookup

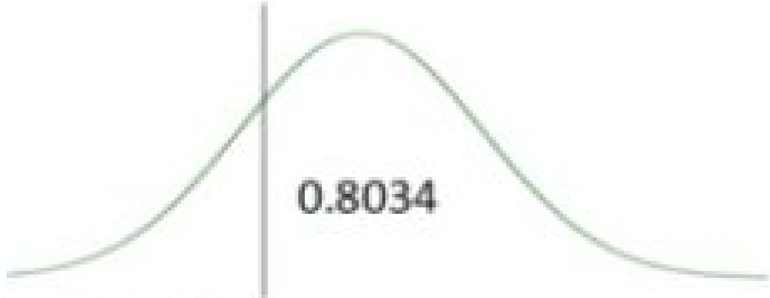
- All probability values in the table are ≥ 0.5.
- Therefore we can only do reverse lookup for probability values ≥ 0.5.
- $P(Z < z) = 0.8034,\ z = ?$

z	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
														ADD					
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359	4	8	12	16	20	24	28	32	36
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	[illegible]	0.5675	0.5714	0.5753	4	8	12	16	20	24	28	32	36
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	[illegible]	[illegible]	0.6064	0.6103	0.6141	4	8	12	15	19	23	27	31	35
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	[illegible]	0.6406	0.6443	0.6480	0.6517	4	7	11	15	19	22	26	30	34
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	[illegible]	0.6772	0.6808	0.6844	0.6879	4	7	11	14	18	22	25	29	32
0.5	0.6915	0.6950	0.6985	0.7019	[illegible]	0.7088	0.7123	0.7157	0.7190	0.7224	3	7	10	14	17	20	24	27	31
0.6	0.7257	0.7291	0.7324	0.7357	[illegible]	0.7422	0.7454	0.7486	0.7517	0.7549	3	7	10	13	16	19	23	26	29
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852	3	6	9	12	15	18	21	24	27
0.8	0.7881	0.7910	0.7939	[illegible]	0.7995	[illegible]	0.8051	0.8078	0.8106	0.8133	3	5	8	11	14	16	19	22	25
0.9	0.8159	0.8186	0.8212	[illegible]	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389	3	5	8	10	13	15	18	20	23
1.0	0.8413	0.8438	[illegible]	[illegible]	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621	2	5	7	9	12	14	16	19	21
1.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

- Z=0.854

Reverse Lookup

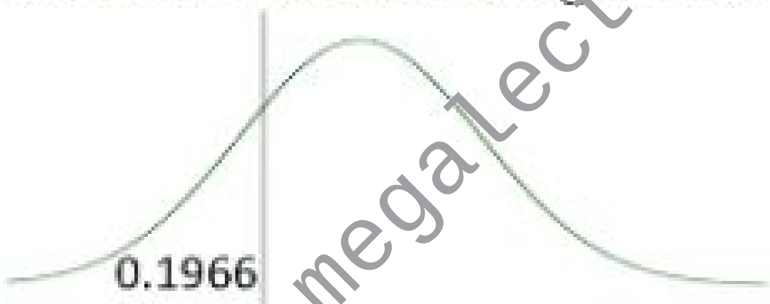
- $P(Z > z) = 0.8034$, $z = ?$
- Sketch the curve. Decide the sign of z. Negative.



- $P(Z < -z) = 0.8034$
- $z = -0.854$

Reverse Lookup

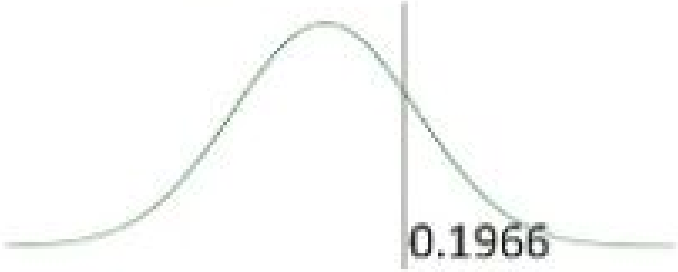
- $P(Z < z) = 0.1966$, $z = ?$
- Sketch the curve. Decide the sign of z. Negative.



- $P(Z < -z) = 1 - 0.1966 = 0.8034$
- $z = -0.854$

Reverse Lookup

- $P(Z > z) = 0.1966$, $z = ?$
- Sketch the curve. Decide the sign of z. Positive.



- $P(Z < z) = 1 - 0.1966 = 0.8034$

Reverse Lookup

- $P(Z > z) = 0.1966$, $z = ?$
- Sketch the curve. Decide the sign of z. Positive.

- $P(Z < z) = 1 - 0.1966 = 0.8034$
- $z = 0.854$

Reverse Lookup

- Sketch the curve.
- Decide the sign of z based on the probability value and the direction of the inequality sign.
- If necessary, change to $P(Z < z)$ = a value > 0.5
- Do the reverse lookup.

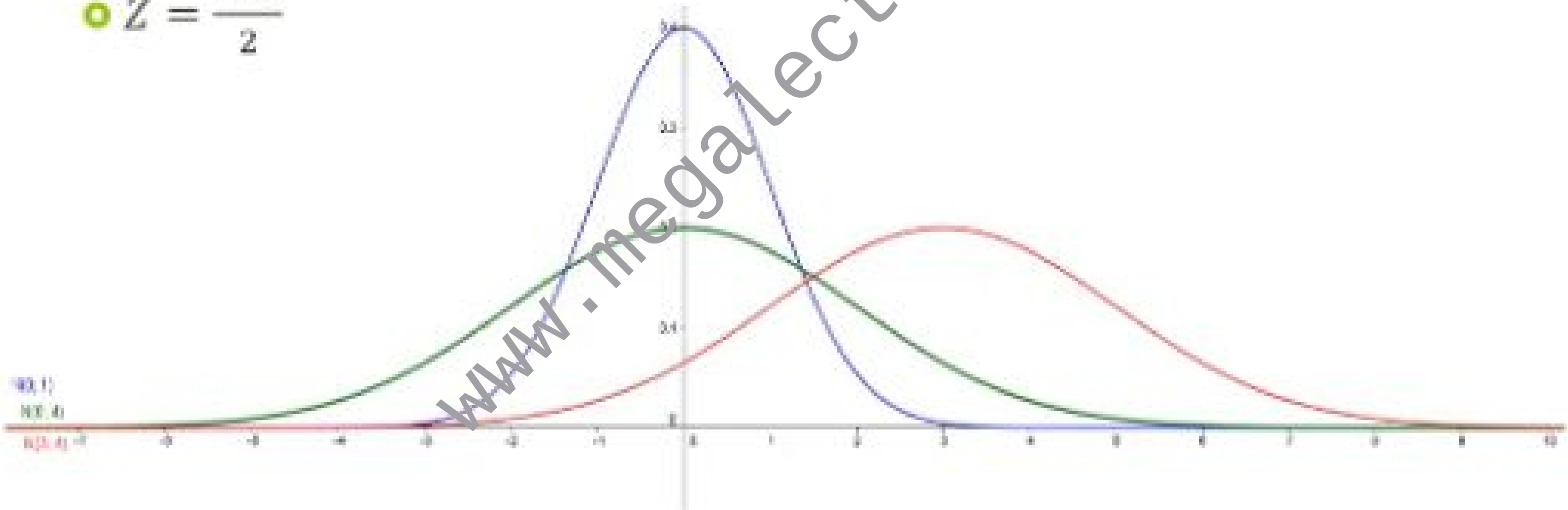

Standardize
Normal Distribution

Standardize Normal Distribution

- $X \sim N(\mu, \sigma^2)$
- $Z = \frac{X-\mu}{\sigma}$

Standardize Normal Distribution

- X ~ N(3, 4), μ = 3, σ = 2
- $Z = \frac{X-3}{2}$



Standardize Normal Distribution

- $X \sim N(3, 4)$, find $P(X < 6)$.
- $P(X < 6) = P(Z < \frac{6-3}{2}) = \Phi(1.5) = 0.9332$
- $P(X > 6) = P(Z > \frac{6-3}{2}) = P(Z > 1.5) = 1 - \Phi(1.5) = 0.0668$
- $P(X < -1) = P(Z < \frac{-1-3}{2}) = \Phi(-2) = 1 - \Phi(2) = 1 - 0.9772 = 0.0228$
- $P(X > -1) = P(Z > \frac{-1-3}{2}) = P(Z > -2) = \Phi(2) = 0.9332$

Practical Questions

- Scores from an exam are normally distributed with mean 70 points and standard deviation 10. If 5000 students took the exam, approximately how many students scored higher than 90?
- $X \sim N(70, 100)$
- $P(X>90) = P(Z>\frac{90-70}{10}) = P(Z>2) = 1 - \Phi(2) = 1 - 0.9772 = 0.0228$
- $5000 * .0228 = 114$

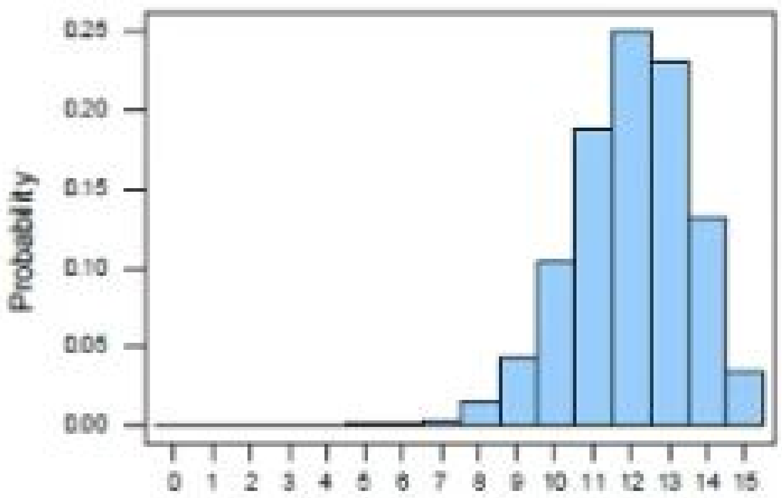
Practical Questions

- The heights of a species of cactus are normally distributed. 34.2% of the cacti are below 12 cm and 18.4% are above 16 cm. Find the mean and standard deviation of the distribution.
- P(X<12)=0.342, P(X>16)=0.184
- $\Phi\left(\frac{12-\mu}{\sigma}\right) = 0.342,\ \Phi\left(\frac{16-\mu}{\sigma}\right) = 1 - 0.184$
- Reverse look up
- $\frac{12-\mu}{\sigma} = -0.407,\ \frac{16-\mu}{\sigma} = 0.900$
- $\mu = 13.2, \sigma = 3.06$

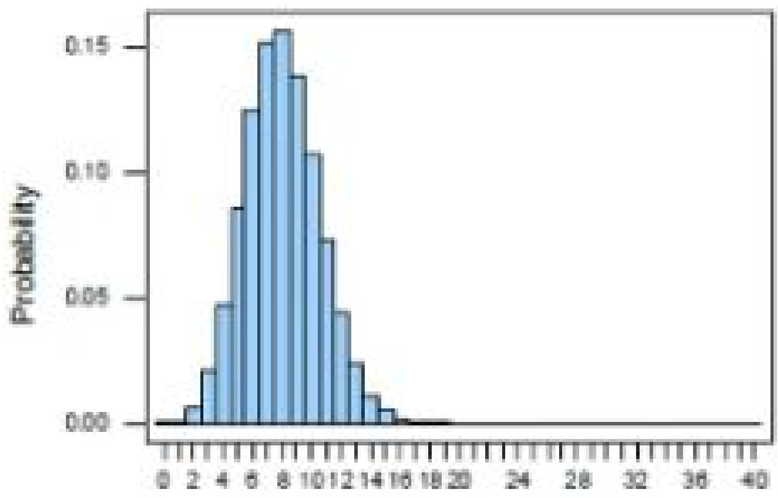

Binomial Distribution
Approximation

Binomial Distribution Graphs

Binomial distribution with n = 15 and p = 0.8



Binomial distribution with n = 40 and p = 0.2



Binomial Distribution Approximation

- Why do use normal to approximate binomial?
- X ~ B(1000, 0.02), find P(X < 200). To find out the exact value, we'll have to add 200 terms (0 – 199).
- When np>5 and nq>5, B(n, p) ➔ N(np, npq)

Continuity Correction (CC)

- Binomial is discrete. Normal is continuous.
- In continuous probability distribution, $P(X = x) = 0$.
- $X \sim B(n, p) \rightarrow Y \sim N(np, npq)$
- If $P(X \leq 5) \rightarrow P(Y \leq 5)$, since $P(Y=5)=0$, $P(5)$ is not counted in.
- $P(X \leq 5) \rightarrow P(Y < 5.5)$
- $P(X < 5) = P(X \leq 4) \rightarrow P(Y < 4.5)$
- $P(X > 5) \rightarrow P(Y > 5.5)$
- $P(X \geq 5) \rightarrow P(Y > 4.5)$

Continuity Correction (CC)

- CC is used **only** at the time of binomial approximation.
- If the distribution itself is normal, there's **no need** to do continuity correction.

Example

- The faulty rate of a product is 3%. 500 products are taken out. Find out the probability of less than 20 products are faulty.
- $X \sim B(500, 0.03)$
- $np = 15 > 5$, $nq = 500 \times 0.97 = 485 > 5$
- $\mu = np = 15$, $\sigma^2 = npq = 15 \times 0.97 = 14.55$
- $P(X<20)$
- $= P(\mathbf{Y<19.5})$
- $= \Phi(\frac{19.5-15}{\sqrt{14.55}}) = \Phi(1.180) = 0.881$


Coding –
Alternative Solution

$\Sigma(x - a)$

- $\Sigma(x - a)$
- $= (x_1 - a) + (x_2 - a) + \ldots + (x_n - a)$
- $= (x_1 + x_2 + \ldots + x_n) - na$
- $= \Sigma x - na$

$\Sigma(x-a)$

- $\Sigma(x-a)$
- $= (x_1 - a) + (x_2 - a) + \ldots + (x_n - a)$
- $= (x_1 + x_2 + \ldots + x_n) - na$
- $= \Sigma x - na$

$\Sigma(x-a)^2$

- $\Sigma(x-a)^2$
- $= (x_1-a)^2 + (x_2-a)^2 + \ldots + (x_n-a)^2$
- $= (x_1^2 - 2ax_1 + a^2) + (x_2^2 - 2ax_2 + a^2) + \ldots + (x_n^2 - 2ax_n + a^2)$
- $= (x_1^2 + x_2^2 + \ldots + x_n^2) - 2a(x_1 + x_2 + \ldots + x_n) + na^2$
- $= \Sigma x^2 - 2a\,\Sigma x + na^2$

Example

The heights, x cm, of a group of 82 children are summarized as follows

$$\Sigma(x - 130) = -287, \text{ standard deviation of } x = 6.9$$

(i) Find the mean height; (ii) Find $\Sigma(x - 130)^2$.

(i) $\Sigma(x - 130) = \Sigma x - 130n = \Sigma x - 130\times 82 = -287$. $\Sigma x = 10373$.

$\bar{x} = \Sigma x / n = 10373 / 82 = 126.5$

(ii) $6.9^2 = \Sigma x^2 / n - \bar{x}^2 = \Sigma x^2 / 82 - 126.5^2$. $\Sigma x^2 = 1316088.52$

$\Sigma(x - 130)^2 = \Sigma x^2 - 2\times 130\times \Sigma x + n\times 130^2 = 4908.52$