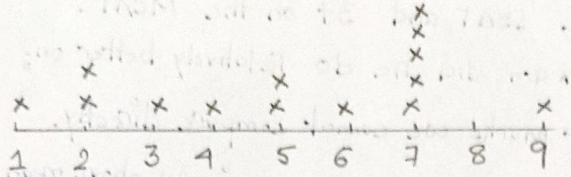
KHAN ACADEMY : STATISTICS AND PROBABILITY

02 : MODELING DATA DISTRIBUTIONS

PERCENTILE

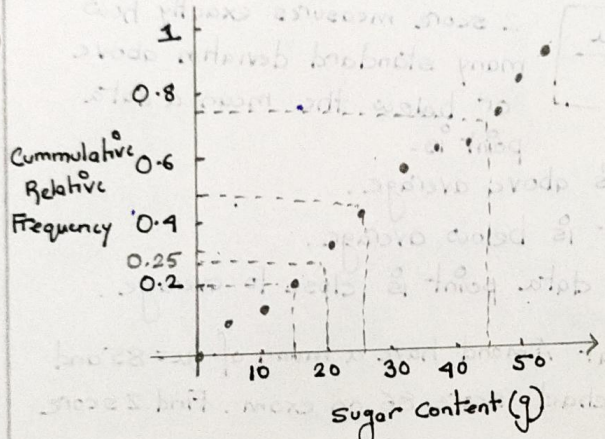
- Percentage of the data that is below the amount.



What is the percentile rank = 6?

$= \frac{7}{14} = 50\%$

Another Example – Nutritionist measures the sugar content (in grams) for 32 drinks at Starbucks. A cummulative relative frequency graph is shown below. An iced coffee has 15grams of sugar. Estimate the percentile of this drink to the nearest percentage.



So cummulative start from 0 x and y.

When sugar content = 15

It has 0.2 = 20% percentile data

Estimate the 50 percentile ≅ 25gram of sugar

So we can say 50% of coffee have 50% of or less sugar

What is the IQR of above graph = 75th percentile - 25th percentile.

$= 45 - 20 = 25$ grams

Z SCORE

Z score → How many Standard Deviation (σ) from the Mean (μ)

Q → The grade on a statistic mid-term for a high school are normally distributed with $\mu = 81$ and $\sigma = 6.3$. Calculate the z score for each of the following exam grades. Draw and label a sketch for each example.

$\mu = 81$, $\sigma = 6.3$

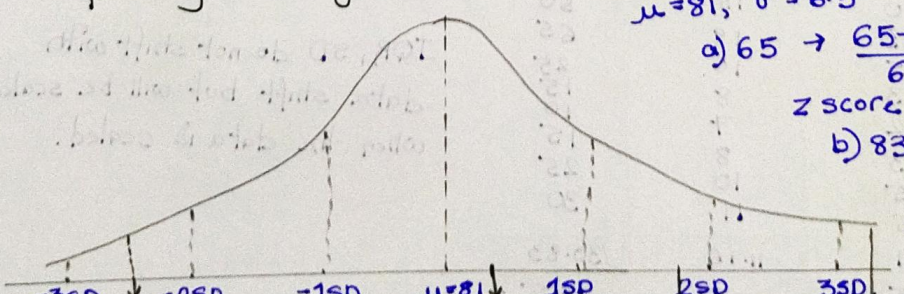
a) 65 → $\frac{65-81}{6.3} = \frac{-16}{6.3} = -2.54$

z score of 65 = -2.54

b) 83 → $\frac{83-81}{6.3} = 0.32$.

c) 93 → $\frac{93-81}{6.3} = 1.9$

d) 100 → $\frac{100-81}{6.3} = 3.02$



Example - Before applying to law school in the US, student need to take an exam called the LSAT. Before applying to medical school, students need to take an exam called the MCAT.

Here are some summary statistics of each exam:

Exam	Mean	Standard Deviation
LSAT	$\mu = 151$	$\sigma = 10$
MCAT	$\mu = 25.1$	$\sigma = 6.4$

Juwan took both exams. He scored 172 on the LSAT and 37 on the MCAT. Which exam did he do relatively better on?

- As the two exam have different scale of marks we cannot compare directly.

LSAT = $\frac{172 - 151}{10} = \frac{21}{10} = 2.1$ SD above the mean. z score is 2.1 above mean.

MCAT = $\frac{37 - 25.1}{6.4} = \frac{11.9}{6.4} \cong 1.86$. SD above the mean. z score is 1.8 above mean

So, he did relatively better in LSAT (2.1) than MCAT (1.8)

$$\text{Z Score} = \frac{\text{data point} - \text{mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$$

z score measures exactly how many standard deviation above or below the mean a data point is.

- A positive z score says data point is above average.
- A negative z score says data point is below average.
- A zscore close to 0 says that data point is close to average.

Eg - The grades on a history midterm at Almond have a mean of $\mu = 85$ and a standard deviation of $\sigma = 2$. Michael score 86 on exam. Find z score

$$\text{Z score} = \frac{\text{his grade} - \text{mean grade}}{\text{standard deviation}} = \frac{86 - 85}{2} = \frac{1}{2} = 0.5$$

His z score is 0.5. His grade was half of a standard deviation above mean.

Example - Find Mean, Median, Standard Deviation and IQR. Give inference.

	Data	Data+5	Data*5
	7	12	35
	7	12	35
	5	10	25
	8	13	40
	10	15	50
	13	18	65
	5	10	25
	3	8	15
	2	7	10
	3	8	15
	5	10	25
	6	11	30
Mean	6.11	11.16	30.83
Median	5.5	10.5	27.5
SD	2.9	2.9	14.9
IQR	2.75	2.75	13.75

Original → Data
Shift → Data+5
Scaled → Data*5

Mean, Median they shift and scale when data is shift.

IQR, SD do not shift with data shift but will be scaled when the data is scaled.

02.3

DENSITY CURVES

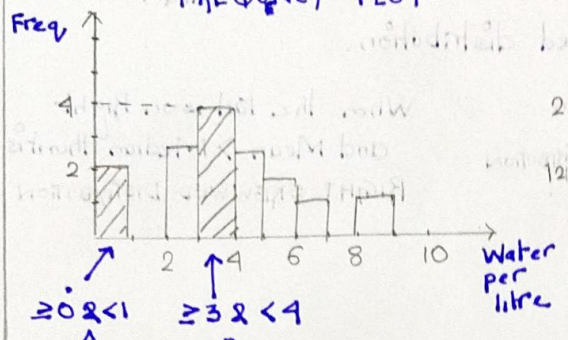
Density curve is a visualization of a distribution where the data points can take any value in a axis.

Suppose a sample of people who drink water per litre are

0.5, 0.7, 2.1, 2.2, 2.9, 3.2, 3.2, 3.3, 3.7, 4.5, 4.6, 4.8, 5.2, 5.3, 6.7, 8.1

A B

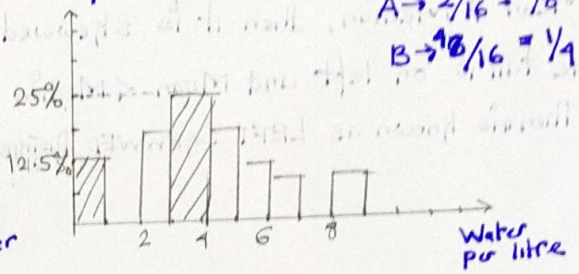
FREQUENCY PLOT



RELATIVE FREQUENCY

A → 2/16 = 1/8 = 12.5%

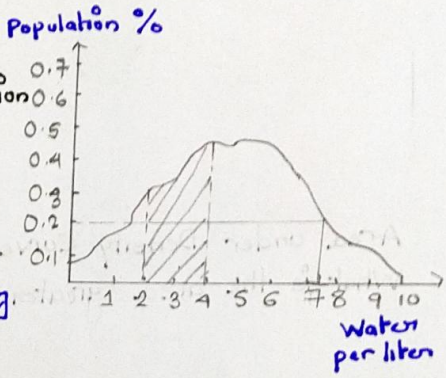
B → 4/16 = 1/4 = 25%



Suppose we have many many datapoints, then in x axis instead of 1 unit = 1 litre we need to decrease than 1 litre. So any

So take any value from the x axis, we use curve instead of frequency. Those curve plot are known as Density curve.

Suppose we have 1.2 billion India population.
And we plotted the Density curve of water consumption



- Inside the curve it represent 100% population or 1.0
- Density curve don't take negative value.
- Suppose how many people drink between 2 to 4 litre of water - Find the Area Under Curve (AUC) of ▨
  - Maybe 40% of data.

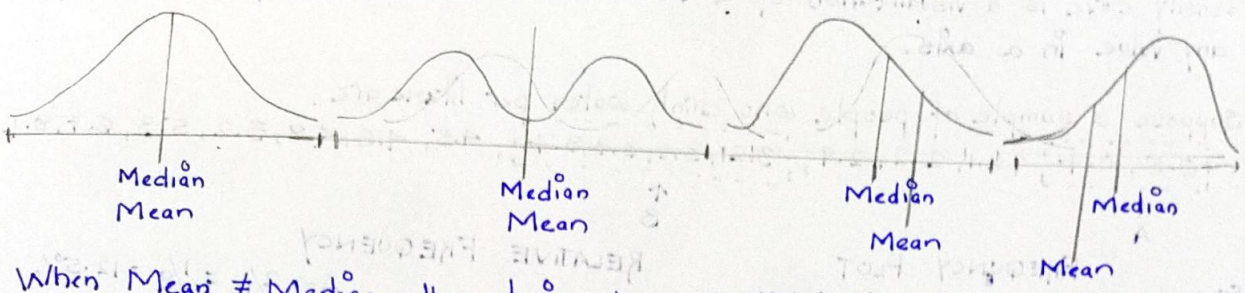
Common misconception → Suppose how many people drink exactly 7 litre of water?

- If we directly find value on y axis, it is 0.2. 20% population but that is wrong

Because, no people on averagely drink exactly 7 litre of water. that is 7.00000
They will be some less or more.

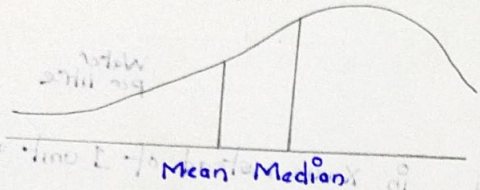
Secondly, there is no AUC under a line
So what we can do is find the area ≥ 6.9 and area < 7.1. That will give AUC of 7
Therefore, it is not about the height or corresponding ~~but~~ value but area of the curve
Vertical line don't have any width.

MEAN, MEDIAN AND SKEW FROM DENSITY CURVE

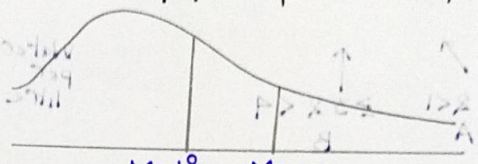


When Mean ≠ Median, then it is skewwed distribution.

When the tail is on left and Mean < Median
Then this known as LEFT SKEWWED DISTRIBUTION

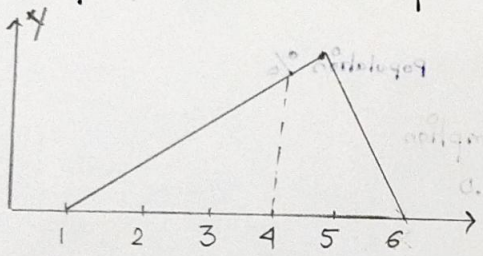


When the tail is on Right and Mean > Median then it is RIGHT SKEWWED DISTRIBUTION



Median is the middle value where left and right have equal values.

Example → Consider the following Density Curve.



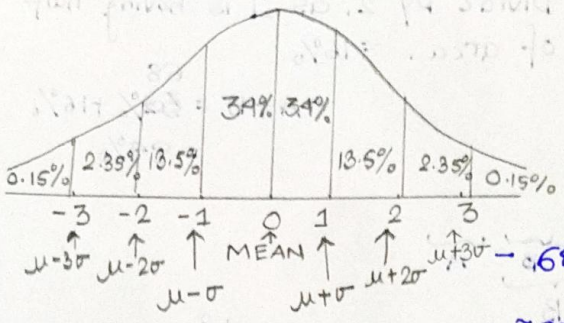
Mean of the density curve is less than Median?
Suppose, Median is half of right & left area.
Consider, Median = 4.
And we can say it is left skewe data.
So, Mean < Median. Thefore True.

Area under Density Curve is 1 ? True any area under density curve is 1.
What is the area greater than=1? It means full density area so 100%.

02.5 ~~NEURAL NETWORK~~

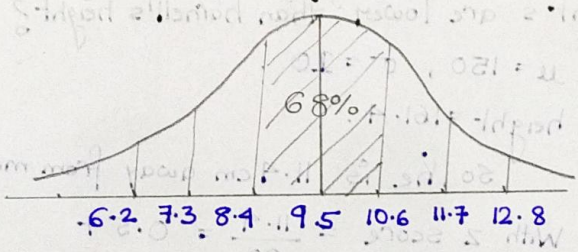
NORMAL DISTRIBUTIONS : EMPIRICAL RULE : 68-95-99.7

What is a Normal Distribution?



- Symmetric bell shape
- Mean and Median are equal, both located at the center of distribution
- 68% of the data falls within 1 standard deviation of the mean.
- 68% chance it will be in $\mu \pm 1\sigma$
- ≈95% of the data falls within 2 standard deviation of mean.
- ≈99.7% of the data falls within 3 SD of mean.

Example – Assume that the mean weight of 1 year old girls in the US is normally distributed with a mean of about 9.5kgrams with a standard deviation of approximately 1.1kgrams. Without using a a calculator estimate the percentage of 1 year-old girls in the US that meet the following conditions. Draw a sketch and shade the proper region for each problem:



$\mu = 9.5$grams, $\sigma = 1.1$grams

a) Find % less than 8.4kg.

Area less than 8.4.

Area under 1SD = 68%.

Area less than 8.4 and Area more than 10.6 = 32%.

As it is normally distributed, divide by half.

So area less than 8.4 is 16%

b) Between 7.3 and 11.7 kg → Indirectly 2SD from Mean = 95%.

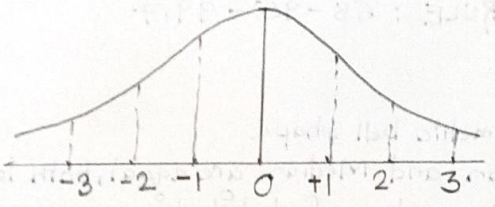
c) More than 12.8kg → 3SD → 99.7%. So area > 12.8 & area < 6.2 = 0.3%

As it is normally distributed divide by 2, Area > 12.8 = 0.15%

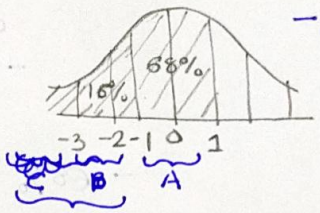
So, probability of having more than 12.8kg is 0.15%.

Standard Normal Distribution → Here mean $\mu = 0$, $\sigma = 1$.

Q) For a standard normal distribution, find the area



1) % of data below 1.
- 1SD = 68%, Remaining = 32%.
- Divide by 2, as 1 is having half of area. = 16%
- A+B = 68% + 16%
= 84%

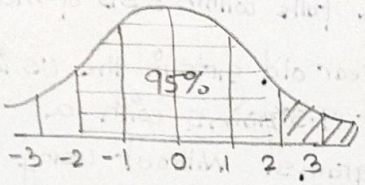


2) Data below -1.
= 16% of data

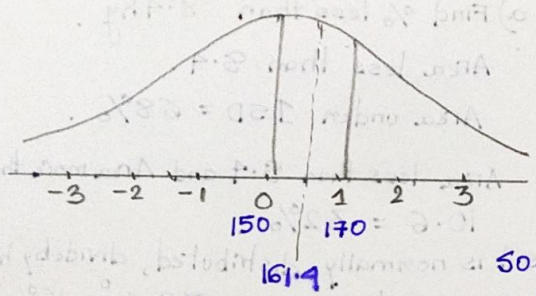
3) The Mean value = 0.

4) The standard deviation = 1 by Standard Normal Distribution

5) Data above 2. → 2SD is 95% Remaining is 5%.
So divide by 2, that is 2.5%.
So % of data above 2 is 2.5%



Q) A set of middle school student heights are normally distributed with a mean of 150 centimeters and a standard deviation of 20 cm. Darnell is a middle school student with a height of 161.4 cm What proportion of student heights are lower than Darnell's height?



$\mu = 150$, $\sigma = 20$.

height = 161.4.

So, he is 11.4 cm away from mean.

With z score = $\frac{11.4}{20}$ = 0.57

So, he is 0.57 standard deviation away from mean (z score = 0.57).

So from z table we can find proportion.

z table, sample

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0										
0.1										
0.2										
0.3										
0.4										
0.5						.7157				
0.6										
0.7										
0.8										
0.9										
1.0										
1.1										

Second decimal ↓

↑ Before second decimal

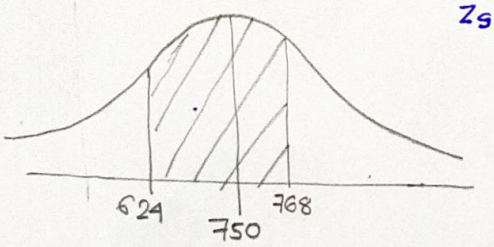
So from z table = 0.57.
71.57% population are below 161.4cm height

02.7 Q) A set of laptop prices are normally distributed with a mean of 750 dollars and a standard deviation of 60 dollars.

What is the proportion of laptop prices are between 624 dollars and 768 dollars?

Ans - $\mu = 750$, $\sigma = 60$.

% between 624 and 768. = z score (768) − z score (624)



$Z\text{score}_{768} = \frac{768-750}{60} = \frac{18}{60} = \frac{6}{20} = 0.30$.

$Z\text{score}_{768}(0.30) = 0.6179$.

$Z\text{score}_{624} = \frac{624-750}{60} = -2.1$

$Z\text{score}_{624}(-2.1) = 0.0179$.

−2.1 is 2.1 SD below the mean and 0.3 is 0.3 SD above the mean

So % between 624 and 768 = 0.6179 − 0.0179 = 0.6 ≈ 60%

Therefore 60% between 624 and 768.

~~FINDING Z SCORE FOR A PERCENTILE (OPPOSITE TO ABOVE)~~