Percentile & Quartiles

Percentage = {1, 2, 4, 3, 5, 6, 7, 8}

Percentile =

$$\% \text{ of even no\%} = \frac{\text{No of even no}}{\text{Total no}} = 4/8 = 20\%$$

Percentiles :- GATE, CAT, IELS, SAT, GRE, JEE, NEET - need pe

Defⁿ - Percentile is a value below which a certain percentage of observation lies.

Eg:- 99 percentile, means person got better marks than 99% of entire students.

1) Dataset = {2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12}

What is the percentile rank of 10?

Soln $x = 10$

$$\text{Percentile rank of } x = \frac{\text{No of values below } x}{n \text{ (sample size)}} = 16/20 = 4/5 = 8$$

= 80 percentile

If $x = 8$, $\frac{9}{\cancel{20}} \times \cancel{100}^{5} = 45$ percentile

$x = 6$, $\frac{7}{\cancel{20}} \times \cancel{100}^{5} = 35$ percentile

$x = 9$, $\frac{14}{\cancel{20}} \times \cancel{100}^{5} = 70$ percentile

2) What is the value that exists @ 25 percentile?

$$\text{Value} = \frac{\text{Percentile}}{100} \times (n+1) = \frac{\cancel{25}^{5}}{\cancel{100}_{2}} \times \cancel{20} = 5^{th} \text{ index}$$

$\boxed{\text{Value} = 5.}$

@ 95 percentile, $\frac{95}{100} \times 21 = 19.25^{th}$ Index

$\boxed{\text{O/P value} = 12}$

5 No summary :-

a) Minimum
b) First quartile [25 percentile Q_1]
c) Median
d) Third quartile [75 percentile Q_3]
e) Maximum.

Used to remove outliner.

$y = \{1, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27\}$

Create a fence → Lower & upper fence.

Lower $= Q_1 - 1.5(IQR)$ → Interquatile Range $= Q_3 - Q_1$
Upper $= Q_3 + 1.5(IQR)$.

$Q_1 = 25$ percentile $= \frac{25}{100} \times 21 = 5.25^{th}$ index. = Value = 3

$\boxed{Q_1 = 3}$

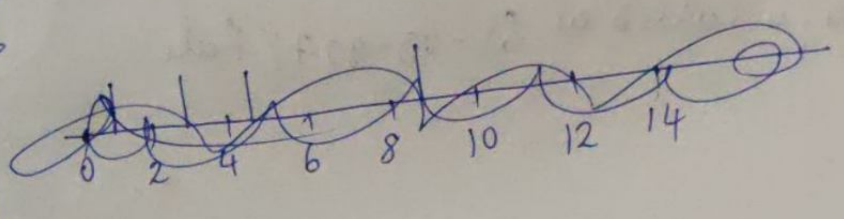
$Q_3 = 75$ percentile $= \frac{75}{100} \times 21 = 15.75^{th}$ index $= 7.5 =$ value.

$\boxed{Q_3 = 7.5}$

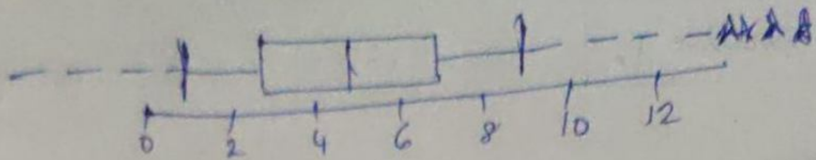
$IQR = Q_3 - Q_1 = 7.5 - 3 = 4.5$.

Lower $= Q_1 - 1.5(IQR) = 3 - 1.5(4.5) = 3 - 6.75 = -3.75$
Upper $= Q_3 + 1.5(IQR) = 4.5 + 1.5(4.5) = 4.5 + 6.75 = 11.25$

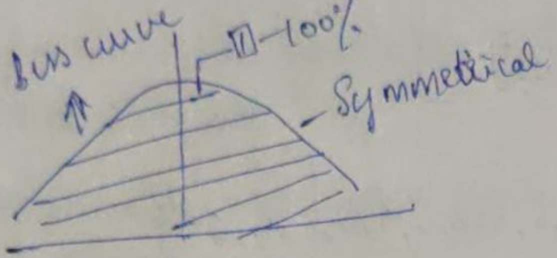
Minimum = 1
$Q_1 = 3$
$Q_3 = 7.5$
Max = 9
Median = 5



Box plot → To treat outliers



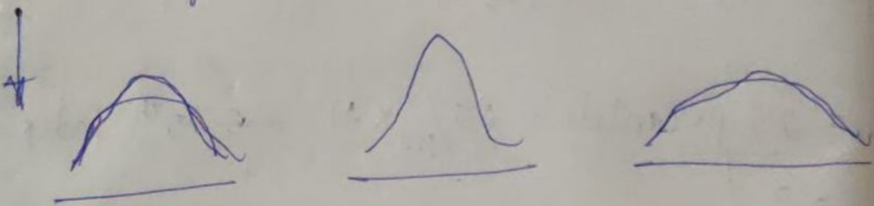
Normal Distribution :-



Kde - Kernel Density estimator.

For age, weight, height need Normal distribution

IRIS dataset ⇒ petal length, sepal length, petal width



Empirical Rule of Normal distribution / Gaussian distribution

Within the first standard deviation b/n the left & right, there are around 68% data.

Within the second standard deviation b/n the left & right, there are around 95% data.

Within the third standard deviation b/n the left & right, there are around 99% data.

This is called as 68-95-99.7% Rule.

(Q-Q) Plot :- Distribution is normal (or) not is known from (Q-Q) plot

Standard Normal distribution :- [SND]

$X \approx$ Gaussian distribution with (μ, σ)

$\Downarrow$

$Y \approx$ SND, where $(\mu = 0\ ;\ \sigma = 1)$

$X = \{1, 2, 3, 4, 5\}$

$\boxed{\mu = 3}$; $\boxed{\sigma = 1.41}$

$$Z\text{-score} = \frac{x_i - \mu}{\boxed{\sigma/\sqrt{n}}}$$

std error , where $n = 1$ [Bcz we are going consider each & every value]

$$\boxed{Z\text{-score} = \frac{x_i - \mu}{\sigma}}$$

$$= \frac{1-3}{1.414} = -2/1.414 = -1.414$$

$$= 2-3/1.414 = -1/1.414 = -0.707$$

$$= 3-3/1.414 = 0/1.414 = 0$$

$$= 4-3/1.414 = 0.707$$

$$= 5-3/1.414 = 1.414$$

Hence, $y = \{-1.414, -0.707, 0, 0.707, 1.414\}$

Why to convert Gaussian to standard ?

Age (year)	Weight (kg)	Height (cm)
24	72	150
26	78	160
32	84	165
33	92	170
34	87	150
28	83	180
29	80	175

Scaling down the value because the units of dataset is different from each other. So to graph will be spread in more width, To reduce the width we have convert Gaussian to standard distribution.

Feature scaling :-

Standardization :- $\boxed{\mu = 0 ; \sigma = 1}$

The values will be within $\{-3, 3\}$

Normalization :-

⇒ Min-Max Scaler $= \dfrac{x - x_{min}}{x_{max} - x_{min}}$

The values will be within $[0, 1]$

eg:

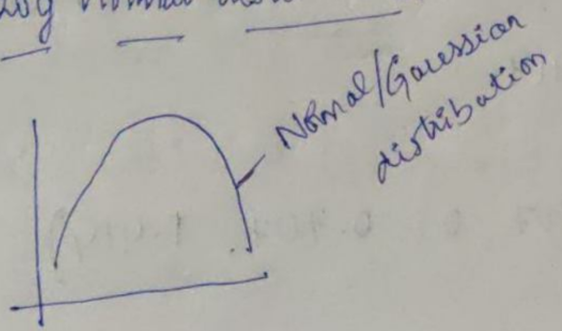
x	y
1	0
2	0.25
3	0.5
4	0.75
5	1

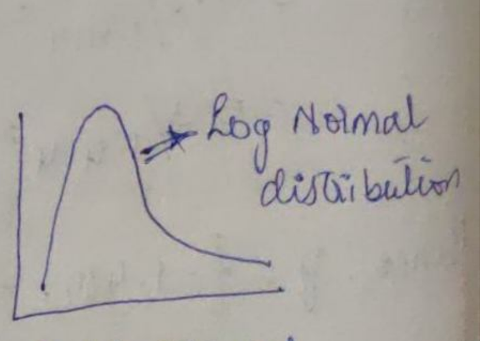
$y_1 = 1-1/5-1 = 0$

$y_2 = 2-1/5-2 = 0.25$

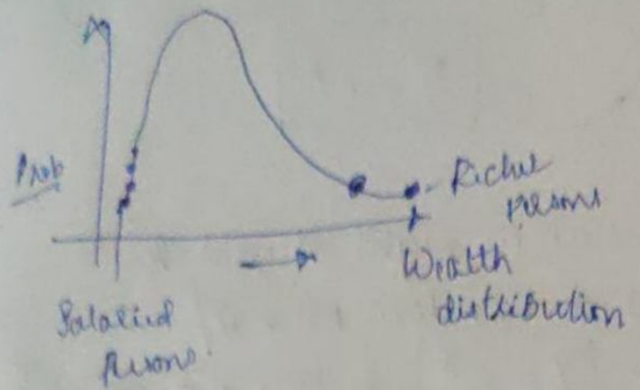
$y_3 = 3-1/5-3 = 0.5$

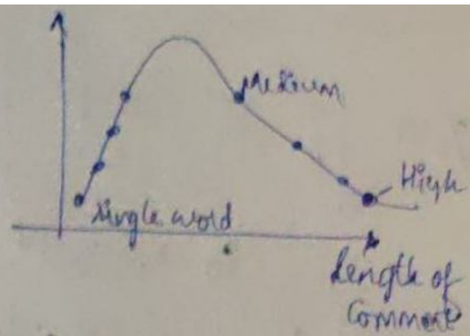
⇒ Log Normal distribution :-



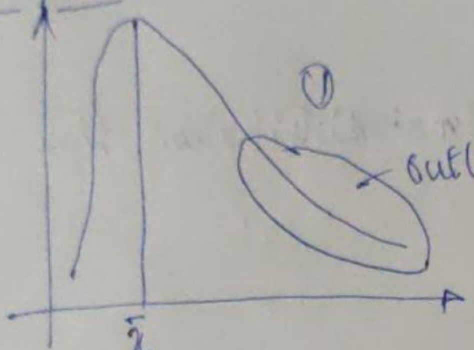


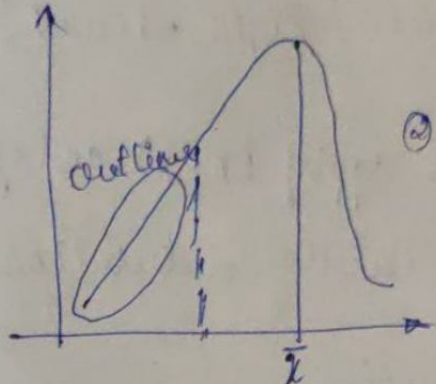
Right skewed.





Assignment



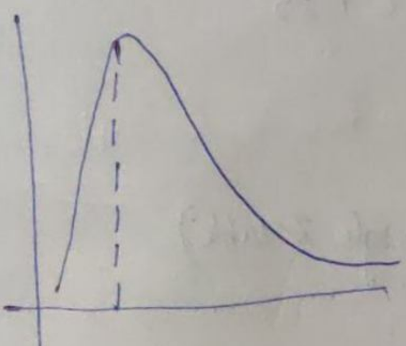


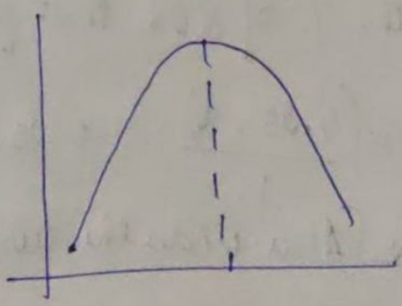
Relationship b/n median, mode, mean of ① + ②

Left skewed ⇒ mean is less than median

Right skewed ⇒ ~~mean~~ mode < median < mean.

No skewed = mean = mode = median

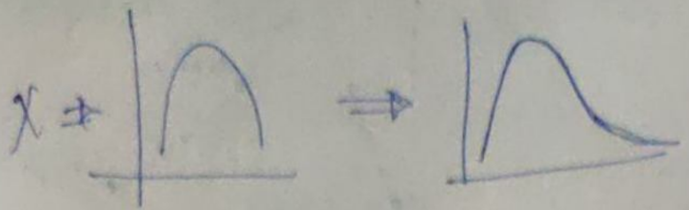




Lognormal dist

Gaussian dist

If the random variable X is a lognormally distributed then $Y = \ln(x)$ has a normal distribution.



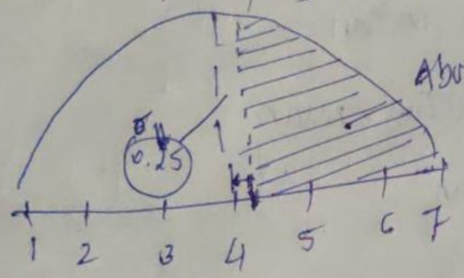
$\boxed{Y = e^{X}}$

To convert gaussian to right skewed log normal distribution, If X is a random normal/ gaussian distribution then $\boxed{Y = e^{X}}$ is log normal distribution [Right skewed]

Also If we apply ln(x) and got normal distribution, then it is log natural distribution.

Practical question:

$X = \{1, 2, 3, 4, 5, 6, 7\}$; $\mu = 4$; $\sigma = 1$



What is the % of score that falls above 4.25.

Soln:-

$Z\text{-score} = \frac{(4.25 - 4)}{1} = 0.25$.

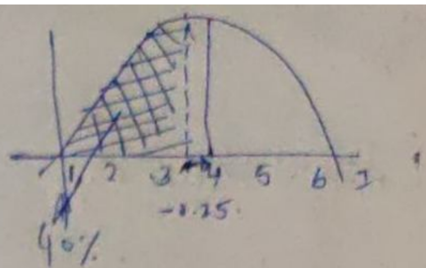
Z-table (Area under the curve) - (Google Z-table)

Remaining = 0.598 = 59.8%

So, % of score that falls above 4.25 is 41%

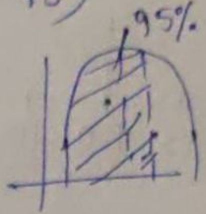
If scale falls below 3.75?

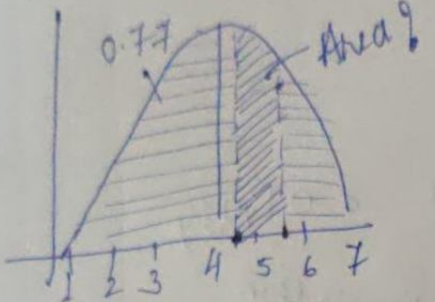
Z-score = $\frac{3.75-4}{1}$ = -0.25



40% score fall below 3.75

The score below between (4.75 & 5.75)





Z-scale

Z-scale = $\frac{4.75-4}{1}$ = 0.75 = 0.77 = 0.23

= $\frac{5.75-4}{1}$ = 1.75 = 0.9599 = 0.1899

So, the scole b/n 4.75 & 5.75 is 18.9%

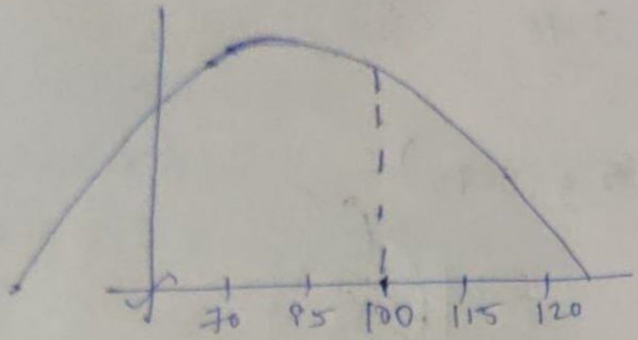
In India, the avg IQ is 100 with std deviation of 15. what is the percentage of population would you expect to have an IQ

1. Lower than 85
2. Higher than 85
3. B/n 85 to 100

Soln

1) Lower than 85

Ans
0.1587
0.8413
0.3413



1) Z-score = $\frac{x_i - \mu}{\sigma} = \frac{85-100}{15} = 1 = 0.84134$

So the % of person having IQ of lower than 85 is 0.15866

2) Z-score = $\frac{85-100}{15} = 1 = 0.84134$

So % of person having IQ of higher than 85% is 0.84134

3) Z-score = $\frac{100-100}{15} = 0 \rightarrow 0.5000$

So % of person having IQ b/n 85 to 100 is 0.34134