Agenda

1. Percentiles and Quartiles ✓
2. 5 Number Summary {Outliers} ✓
3. Box plot ✓
4. Covariance And Correlation }.
5. Probability distribution function
6. Different types of distributions

Studying these just to determine and visualize outliers in the data set.

All these work in sorted data

1) Percentiles And Quartiles [GATE, CAT]

Percentage : 1, 2, 3, 4, 5, 6

% of numbers that are odd = $\frac{3}{6}$ = $\frac{\text{No. of odd numbers}}{\text{No. of total no}}$

$= \frac{1}{2} = 50\%$

Percentiles :

Defn : A percentile is a value below which a certain percentage of data points lie.

Eg: The CAT cutoff for IIM is typically from 99 to 100 percentile. So only top 1 percentage of population is able to make into IIMs.

$n = 15$

$X : \{2, 3, 3, 4, 6, 6, 6, 7, 8, 8, 9, 9, \boxed{10}, 11, 12\}$

Percentile Rank of (value) 10 = $\frac{\#\text{ of values below } 10}{n} * 100$

$= \frac{12}{15} \times 100 = 80$ percentile

What 80 percentile means

80 percentile = 80% of the distribution fall below the value of 10 //

(2) What value exists at 25 percentile ?

Formula $\text{value} = \frac{\text{Percentile}}{100} * (n+1)$

$= \frac{25}{100} * 16 = 4^{th}$ Element $= 4$

Fourth element is 4 after arranging number in ascending order.

1st 2nd 3rd 4th

$X = \{2, 3, 3, 4, 6, 6, 6, 7, 8, 8, 9, 9, 10, 11, 12\}$

$\frac{4+4}{2} = 5$

If 25 percentile value had came in decimal EG:4.5 then simply take average of 4th and 5th element (after arranging data in ascending order) and the result will be value at 25 percentile.

4.5

Why we are learning this?
Ans: Because later these will be used to calculate the outliers.

Quartiles

(1) Q1 → 25 percentile

Q2 → Median → 50 percentile

Q3 → 75 percentile

(2) 5 Number Summary

5 number summary is represented using Box Plot

(1) Minimum
(2) First Quartile (25 percentile) (Q1)
(3) Median (Q2)
(4) Third Quartile (75 percentile) (Q3)
(5) Maximum

Remove the Outliers

Outlier is a value that is completely unique or not following a general trend.

$X = \{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 29\}$ (29: Outlier)

Through observation we can tell in above 29 is outlier. But it's important to mathematically prove the same. Below we have tried to obtain outliers by drawing higher and lower fence, above or below which data points will be treated as outliers.

[Lower Fence ⟷ Higher Fence]

Lower Fence = Q1 − 1.5(IQR)

Inter Quartile Range = Q3 − Q1

Higher Fence = Q3 + 1.5(IQR)

$X = \{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 29\}$. (29 is outlier)

Q1 = 25 percentile $= \frac{25}{100} * 20 = 5^{th}$ value $= 3$

Q3 = 75 percentile $= \frac{75}{100} * 20 = 15^{th}$ value $= 7$

IQR = 7 − 3 = 4

Lower Fence = $Q_1 - 1.5(IQR)$
$= 3 - 1.5(4)$
$= 3 - 6$
$= -3$

Higher Fence = $Q_3 + 1.5(IQR)$
$= 7 + 1.5(4)$
$= 7 + 6$
$= 13$

Note upper and lower fence value are inclusive and should not be considered outlier. Point above and below should be considered outliers.

$[-3, 13]$.

Since 29 is above higher fence, it is considered as outlier.

There is not data point below lower fence.

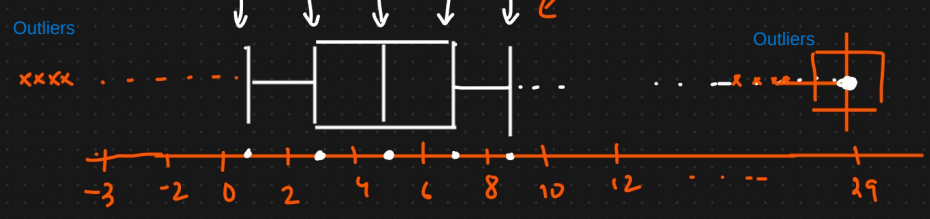
$X = \{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 29\}$. (29 is outlier)

Box plot [To visualize Outliers]

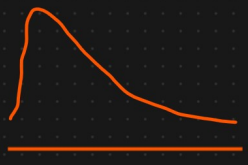
1. Minimum value = 1
2. Q1 = 3
3. Median Q2 = 5
4. Q3 = 7
5. Maximum = 9

Box plot

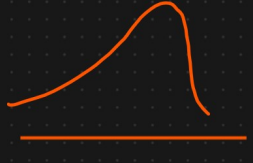
This belongs to gaussian/normal distribution where Q3-Q2=Q2-Q1(when this difference is not same then left or right skewed graph is formed which is discussed below)



Interview question: Box plot for left and right skewed



mean > median > mode



Case 2 : for left skewed graph box plot will be shifted right(ie; median =Q2 will shift right)

Means: Q3-Q2 < Q2-Q1

Case 1 : for right skewed graph box plot will be shifted left (ie; median =Q2 will shift left)

$Q_3 - Q_2 > Q_2 - Q_1$ Means: Q3-Q2 > Q2-Q1

Internal Assignment

$\frac{-5+3}{2} = -1$

Y is a random variable.

$Y = \{-13, -12, -6, -5, 3, 4, 5, 6, 7, 7, 8, 10, 10, 11, 24, 55\}$.

[Lower Fence $\longleftrightarrow$ Higher Fence]

$Q1 = \frac{25}{100} \times 17^{4} = 4.25$

$Q3 = \frac{75}{100} \times 17 = 12.75$

Lower Fence $= -1 - 1.5(10+1)$

$= -1 - 1.5(11)$

$= -17.5$

$[-17.5, 26.5]$.

Higher Fence $= 10 + 1.5(10+1)$

$= 10 + 16.5$

$= 26.5$

55 is an outlier

Different question where Z is random variable

$Z = \{1, 2, 4, 6, 7, 12, 18, 34, 77, 66, 108, 99, 14\}$.

$\{1, 2, 4, 6, 7, 12, 14, 18, 34, 66, 77, 99, 108\}$

$Q1 = 5$

$Q3 = 71.5$

$[-94.75, 171.25]$

$Q3 = \frac{75}{100} \times 14 = \frac{42}{4} \; 2 + 10.5$

$Q3 = \frac{66+77}{2} = 71.5$

Covariance And Correlation

[Relationship between X and Y]

X	Y
2	3
4	5
6	7

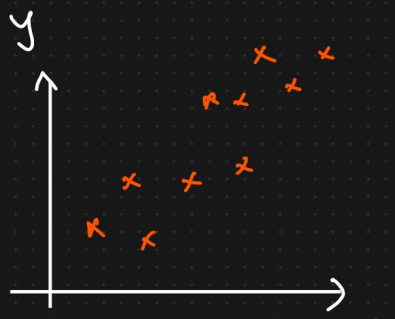
→	X↑	Y↑
→	X↓	Y↑
→	X↑	Y↓

IS Haunted

Size of House → Price

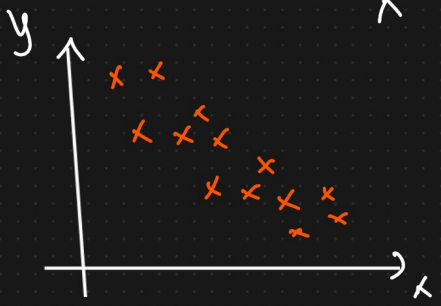
All possible relationships between 2 variables. I more can be added that is by changing x y does not change(x = y)

8 9 → x↓ y↓



Case 1: Direct relationship
By increasing x, y increases and by decreasing x, y decreases.

x↑	y↑
x↓	y↓



Case 2: Indirect/inverse relationship
By increasing x, y decreases and by decreasing x, y increases.

x↑	y↓
x↓	y↑

Our aim is to get some values out of these relationships OR we just want to establish this whole thing mathematically. So, the first technique for doing this is Covariance.

① Covariance

$$Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

How to remember. This formula is similar to the formula for variance just remove square(power) and add y - ybar

$$Var(x) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

This is interview question: What is relationship b/w variance and covariance.

Ans: Variance of any random variable X is equal to Covariance of that random variable X with itself.

ie; Var(X) = Cov(X,X)

Covariance of x with itself = Cov (x,x)

$$= \sum_{i=1}^{n} \frac{(x_i - \bar{x}) * (x_i - \bar{x})}{n-1}$$

$$\Downarrow$$

$$Var(x) \Leftarrow Cov(x,x)$$

Covariance simply means relationship of one with another random variable. It can be +ve or -ve.

Cov(x,y)

Case 1:
For direct relationship +ve Cov

x↑	y↑
x↓	y↓

⇒ +ve Covariance

Case 2:
For indirect/inverse relationship -ve Cov

x↑	y↓
x↓	y↑

⇒ -ve Covariance.

Var(X) means how well X is spread.

Cov(X,X) means relationship of X with itself

Below is numerical example of Cov(x,y)

x	y
→ 2	3
→ 4	5
6	7
$\bar{x} = 4$	$\bar{y} = 5$

$$Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

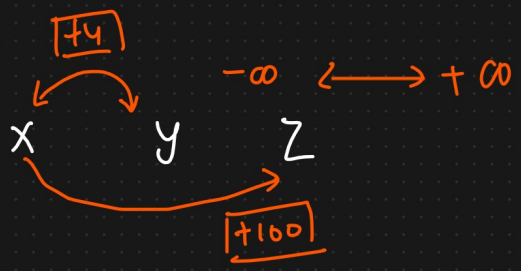
$$= \frac{[(2-4)(3-5) + (4-4)(5-5) + (6-4)*(7-5)]}{2}$$

$$= \frac{4+0+4}{2} = \frac{8}{2} = 4 \text{ +ve covariance}$$

X and y are having a positive covariance

Advantages

① Relationship between X & y



Disadvantage

① Covariance does not have a specific limit value

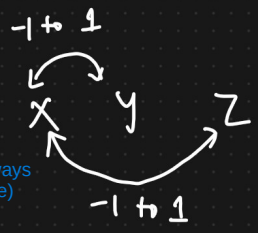
Advantage of covariance is that we are able to define relationship b/w elements but only disadvantage is one cannot define the strength of relationship(bonding). Pearson here comes into rescue.

Please note that the numerical value obtained just gives us direction whether it will be +ve cov or a -ve cov. There is no limit on this values and can go upto +ve/-ve infinity. These values cannot be used for determining strength (eg: x is less correlated with y as compared to correlation of x with z since, cov just give direction of cov)

① Pearson Correlation Coefficient

[-1 to 1]

O/P of pearson corr coeff will always range from -1 to +1(both inclusive)



$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}$$

Read as row of x and y

Cov(x,y) = Covarriance of x and y
sigma x = standard deviation of x
sigma y = standard deviation of y

④ The more the value towards +1 the more +ve correlated it is

② The more the " " -1 " " -ve " " "

Pearson corr coeff of x and y = 0.6
Pearson corr coeff of x and z = 0.7

Now, we can say that x is highly correlated with z as compared to correlation of x with y. This is something which we were not able to determine using Covariance alone.

Imp

-->Pearson can only capture linear relationship(See Pearson_wiki.png in the same folder)

-->For, non linear we use Spearman(See Spearman_wiki.png in the same folder)

X y 0.6

X Z 0.7

② Spearman Rank correlation

$$r_s = \frac{Cov(R(x), R(y))}{\sigma_{R(x)} * \sigma_{R(y)}}$$

Denoted by r suffix s

The one which is largest is given rank of 1, second largest is given rank of 2 .. so on. and then is used in Spearman equation.

X	y	R(x)	R(y)
5	6	3	1
7	4	2	2
8	3	1	3
1	1	5	5
2	2	4	4

R(x)=rank of x
R(y)=rank of y
Cov = covariance
sigma = standard deviation

Note that using covariance we were not able to determine strength of correlation (only direction of correlation were able to determine). So, the concept of correlation was introduced to determine strength of correlation (Pearson and Spearman). Question is that why we are doing this?

During EDA/feature engineering we need to select features that are highly correlated with output features and drop less correlated ~0. Here these concepts will help in determining the same.

Feature Selection

+ve	+ve	+ve	≈0	-ve	O/P
Size of house	No. of Rooms	Location	~~No. of people Staying~~	Haunted	Price ↑
+ve correlated with o/p since, by increasing size of house price will increase	+ve correlated with o/p	+ve correlated with o/p	this will have negligible impact over price of the house hence correlation is near to 0	-ve correlated with o/p	

Probability Distribution Function And Probability Density Function

Probability Mass function

Probability Distribution Function

(Below are the types of Probability Distribution Function)

1. Probability density fn
2. Probability mass fn ✓
3. Cummulative distribute fn.

Never consider probability distribution fun^ similar to probability density fun^.

Probability density fun^ is of the type of Probability distribution function.

When we want to draw histogram to see the distribution of data consisting of discrete random variables, we will use probability mass fun^.

1. PMF

PMF characterizes the distribution of a discrete random variable

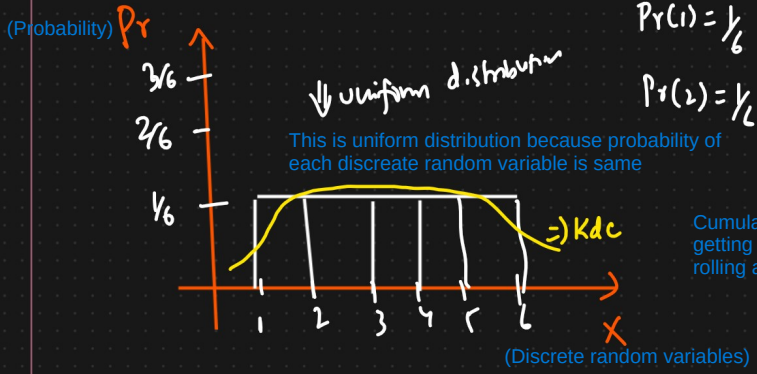
1. Discrete Random Variable

Discrete random variable:
Whole numbers with some range of values. Note that such variable holds outcomes of random experiment for eg: rolling a dice.

Eg: Rolling a dice {1,2,3,4,5,6}

$Pr(1) = 1/6$

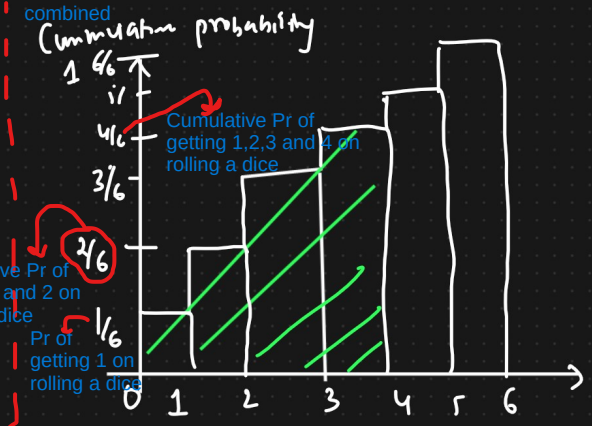
$Pr(2) = 1/6$



This is uniform distribution because probability of each discreate random variable is same

Cumulative distribution function (cdf)

Cumulative means combined



$Pr(1 \text{ or } 2) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$

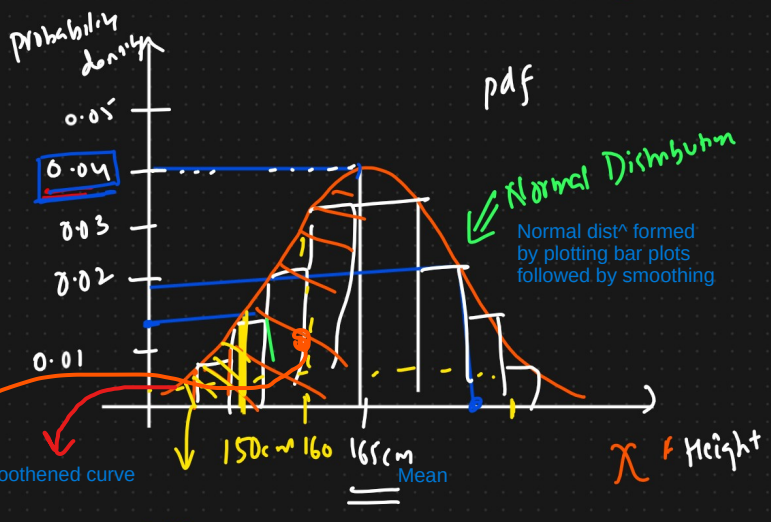
$Pr(X \leq 4) = Pr(X=1) + Pr(X=2) + Pr(X=3) + Pr(X=4)$

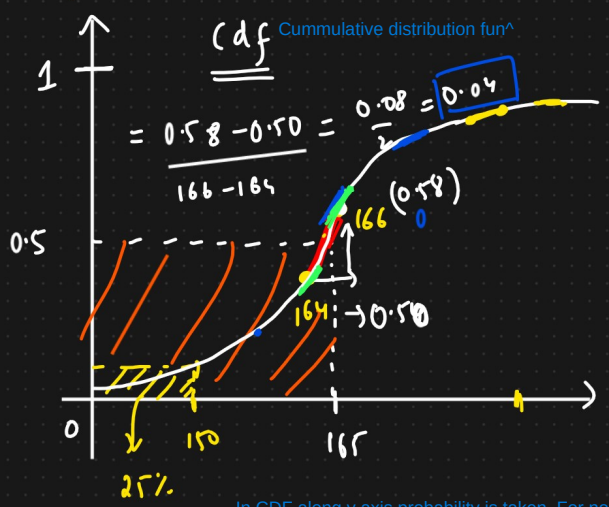
$= O/P$

② Probability Density Function

Probability density function deals with the distribution of Continuous random variable. For eg: I am 175.5 cm long.

① Distribution of Continuous Random Variable





In CDF along y axis probability is taken. For normally distributed PDF the CDF thus formed is S shaped ranging from 0 to 1 on y axis (range of probability is [0,1]). Mean height is 165 that's why it's projection on axis is 0.5

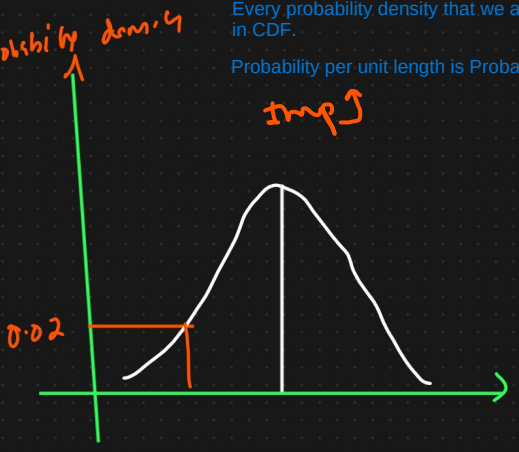
Area under probability density function gives Probability of that continuous random variable. For Eg; normal distribution with mean height 165 cm, since this is mean represent 50% of distribution. So, area under this point will be 0.5 which is also the probability of occurrence of mean data point 165cm(This example is represented in above PDF and CDF using orange shaded area/section/region).

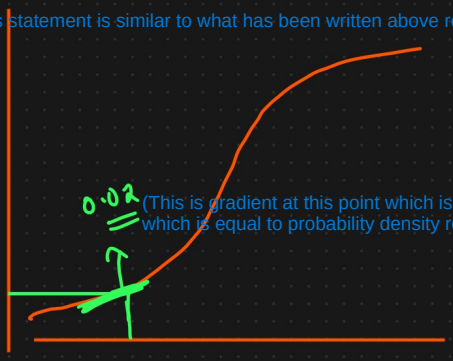
$\int_{-\infty}^{\infty} = 0.25$

Probability Density ⇒ Gradient of Cummulative Curve

Every probability density that we are getting in PDF is coming from gradient (which is basically slope) of respective data points in CDF.

Probability per unit length is Probability density(This statement is similar to what has been written above related to gradient)





Different types of Distribution

Why are we learning about these distributions?
Ans: because when we will be plotting our data it will follow one of these distributions. So knowing about these distributions in advance will help us understand our data better.

1. Normal / Gaussian Distribution → pdf
2. Standard Normal Distribution → pdf
3. Log Normal Distribution → pdf
4. Power Law Distribution → pdf
5. Bernoulli Distribution → pmf
6. Binomial Distribution → pmf
7. Poisson Distribution → pmf

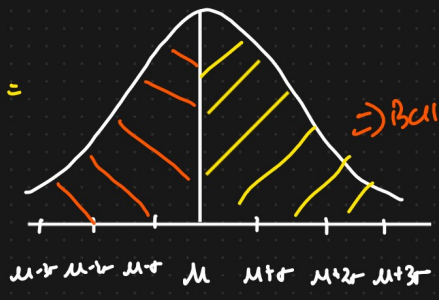
⑧ Uniform Distribution → Discrete → pmf
→ Continuous → pdf

⑨ Exponential Distribution. → pdf

⑩ CHI SQUARE Distribution → pdf

⑪ F Distribution → pdf.

① Normal/Gaussian Distribution



sigma is standard deviation

X :- Continuous Random Variable.

Symmetrical curve ie; LHS=RHS

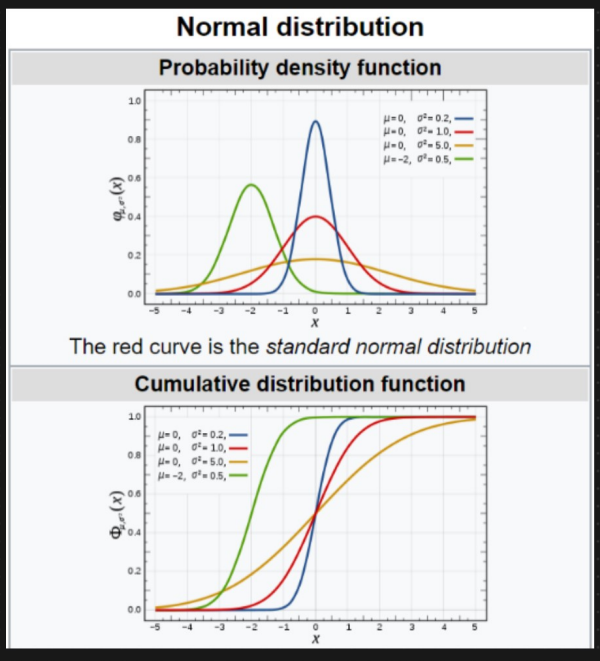
For normal distribution mean = median = mode.

This is plotted by plotting histogram(continuous random variable in x axis and frequency along the y axis) followed by smoothing.

Most of the data distributions =in this universe follows the normal distribution.

IRIS data set follows normal distribution which is obtained after plotting data collected by botanists.

Eg:- Height, weight, age, IRIS dataset



$X \sim N(\mu, \sigma^2)$

Denoting X random variable distributed normally(N) with some mean and variance.

Support parameters (Mean and variance is support parameters)

μ = mean

σ^2 = variance

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

For every distribution use wikipedia. See fig, CDF and PDF formula. Just understand and not by heart remember these.

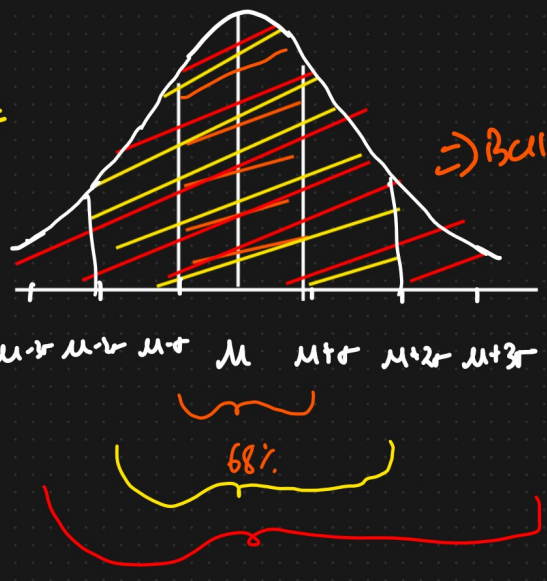
All distributions following normal distribution is assumed to follow the Empirical Rule that states that:
--> 68% of data points will lie in the 1st standard deviation of distribution
-->95% of data points will lie in the 2nd standard deviation of distribution
-->99.7% of data points will lie in the 3rd standard deviation of distribution.

Empirical Rule — 68 - 95 - 99.7% Rule

100 datapoint

$X = \{ \quad \}$

mean : mode : median



⇒ Bell Curve

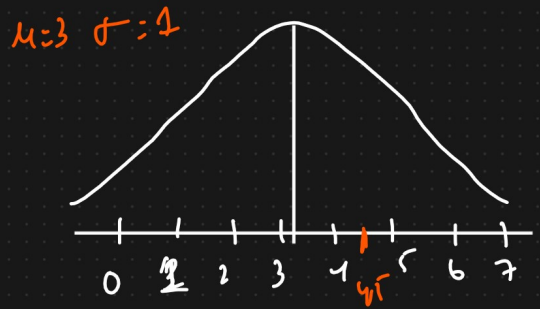
Standard Normal Distribution

Suppose we have a random variable X distributed normally with mean=3 and standard deviation=1. If we transform it into normal distribution with mean=0 and standard deviation=1, the transformed distribution thus formed is called standard normal distribution.

$X \quad \mu = 3 \quad \sigma = 1$

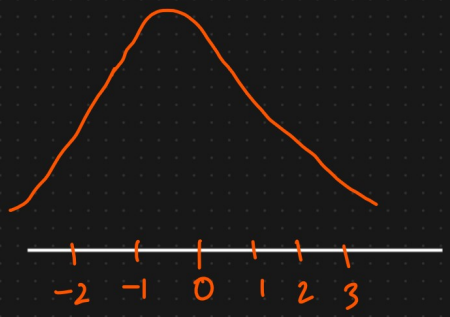
This transformation is done with Zscore

↓ Standard Normal Distribution



Transformation ⇒

$\mu = 0, \sigma = 1$



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

Z-score tells you about a value how many standard deviation away from the mean

$\frac{3-3}{1} = 0$

1 in normal distri^ get transformed to -2 for standard normal distri^ $= \frac{1-3}{1} = -2$

Similarly, 2 in normal distri^ get transformed to -1 for standard normal distri^ $= \frac{2-3}{1} = -1$

Similarly, mean 3 in normal distribution get transformed into 0 in standard normal distribution.

$\boxed{4.5} \Rightarrow \frac{4.5-3}{1} = 1.5$

Numerical like this are there on later lectures. Using Z table to find area from left

X $\mu = 4$ $\sigma = 1$

What is the percentage of scores lies above 4.5?

$\Rightarrow$ z table

$$z\text{-score} = \frac{4.5 - 4}{1} = 0.5 \Leftarrow$$

Area under curve $= 1 - 0.6914$

$= 0.3086 = 30.86\%$