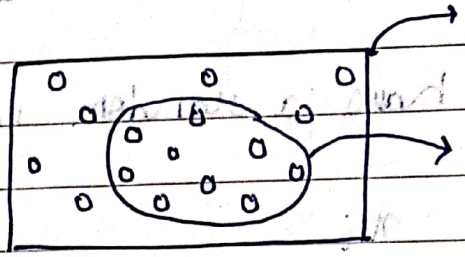
VIPUL GAURAV

Statistics For Machine Learning (Krish Naik).

Population: Consider we want to calculate the average height of all the people in a state. Total no. of people in state is said to be the population.

$$\text{Population Mean} = \frac{1}{N} \sum_{i=1}^{n} x_i$$

It refers to the total number of instances present in the data is called its population.



Sample: It refers to the subset of examples present in population.
eg. Exit polls of an election: The news reports based on sample of the entire population to determine who will win the election.

This is used to generalize and make an actual prediction on who will win the election.

$$\text{Sample Mean} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Population Mean > Sample mean

- Random Variables :
  - Discrete → whole number, not floating
  - Continous → within a range of values, we can have any value

eg. (10 - 15)

10.2, 11.3, 13.3, 14.5 ∈ (10, 15) ∈ R

↳ decimal, whole number

Height of a person (say 6 ft 2 inch or 6.2 feet)

eg. No. of bank accounts any person has.
→ 2, 3, 4, 5 ✓ 2.5 ✗
- Population of state

- Gaussian Distribution (Normal dist^n)

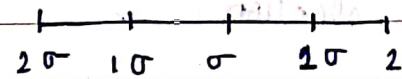
Suppose we have a random variable x.

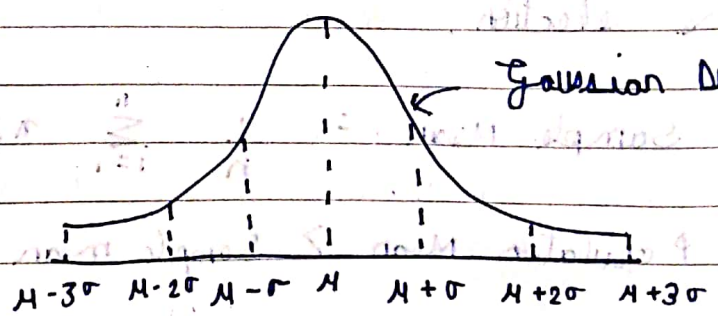
$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\text{Variance} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$\sigma = \sqrt{\text{Variance}}$$

How far is the element distributed

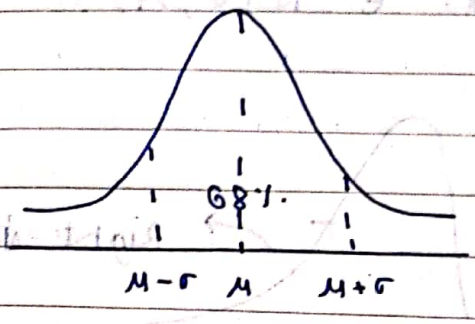




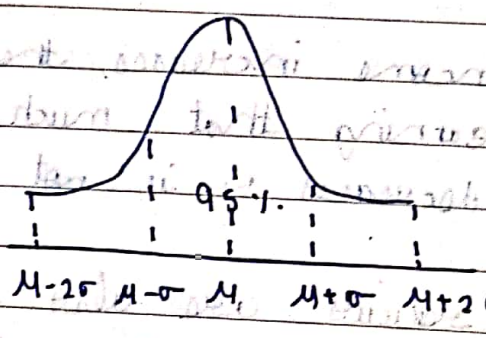
Empirical Formula:

$P(\mu - \sigma \leq x \leq \mu + \sigma) \cong 68\%$.

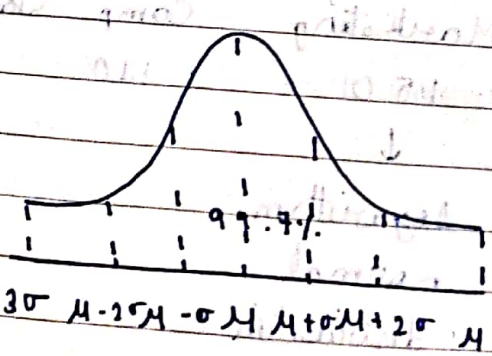
$x \to X$:



$P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \cong 95\%$.



$P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \cong 99.7\%$.

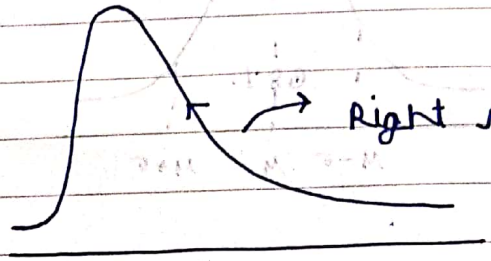


Log distribution:- It is also a form of normal distribution

$X \sim$ Log Normal dist^n if $\ln(x)$ is normally distributed

$\ln(x_1)$, $\ln(x_2)$, $\ln(x_3)$ ....
↳ normal dist^n

$\therefore$ $x \sim$ Log Normal
if $\ln(x) \in N(\sigma, \mu)$.



eg. Income of the people.

As the income increases, the number of people earning that much amount of money decreases & is not abundant.

Feedback reviews are also in form of log distribution, than the number of people receiving with long reviews will be very less.

eg.

R & D	Marketing	Comp statistics	Profit
50	150	210	300	100
↓	↓		
gaussian	logarithm		
↓	normal		
standard	distribution		
normal	↓		
($\sigma = 1, \mu = 0$)	standard		

$\therefore$ We scaled it down by: $\bar{x} = \frac{x - \mu}{\sigma}$ ← comparable

(because log(values) follows Gaussian distribution).

Covariance in Statistics.

Size	Price
1200 sqm	100
1800 sqm	200

} Can we quantify a relation between them?

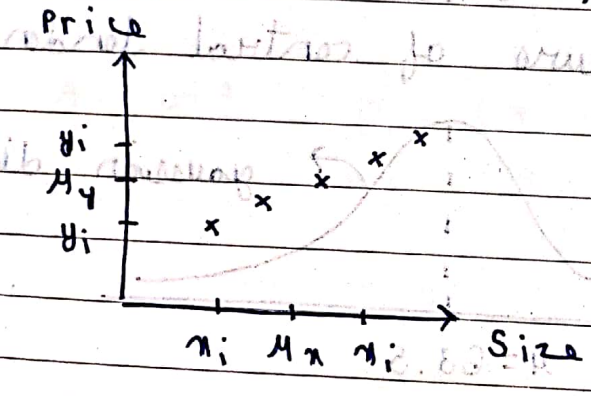
$$Cov(x, y) = \frac{1}{n} \sum_{i=1}^{n} [(x_i - \mu_x) * (y_i - \mu_y)]$$

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

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x)(x_i - \mu_x)$$

$\therefore Cov(x, x) = Variance(x)$

If $x \uparrow$ $y \uparrow$ & $x \uparrow$ $y \downarrow$



$x_i > \mu_x$

$Cov = (+ve) * (+ve)$

$= +ve.$

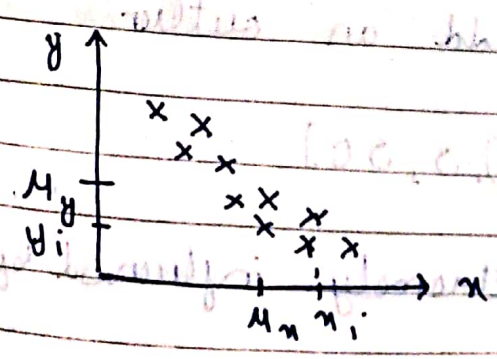
$x_i < \mu_x$

$Cov = (-ve) * (-ve)$

$= +ve$

But, Covariance will not always be positive.

Consider $x \uparrow$, $y \downarrow$:



$x_i > \mu_x$ & $y_i < \mu_y$

$Cov(x, y) = +ve * (-ve)$

$= (-ve)$

if $x_i < \mu_x$ and $y_i > \mu_y$

$Cov(x, y) = -ve * (+ve) = (-ve).$

Covariance $X \uparrow, Y \uparrow \Rightarrow +ve$ (how much?)
$X \downarrow \quad Y \uparrow \Rightarrow -ve$

We need to measure how much we need it!

Mean, Median and Mode.

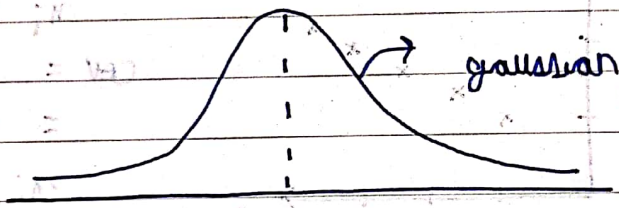
Sample of Height: {168, 170, 150, 160, 182, 142, 175}.

$$\mu = \sum_{i=1}^{n} [x_i] * \frac{1}{n}.$$

$$\mu_s = \frac{168 + 170 + 150 + 160 + 182}{7}$$

$$\mu_s = 63.5$$

It is measure of central tendency



$[1, 2, 3, 4, 5] \therefore \mu = \frac{5(5-1)}{2} = \frac{5 \times 4}{5 \times 2} = 2.$

$\frac{}{5}$

Consider I add an outlier

$[1, 2, 3, 4, 5, 50]$

$\therefore \mu = 13.$

So, it is extremely influenced by outliers

$\text{Median} = \frac{3+4}{2} = 3.5$

It lessens impact of outlier

Mode is the most frequent occuring datapoint

Age : 23, 24, 27, 32, 35, 21 ...

How to handle missing value in age ?
Mean, Median or Mode.

If there are not many outliers then we can use mean or else use median

Central Limit Theorem.

Consider a random variable X which may or may not belong to Gaussian dist^n.

$X \neq GD(\mu, \sigma^2)$ $n \geq 30$.

μ σ^2

S_1 $(x_1, x_2, \ldots, x_{30})$ $\Rightarrow \bar{x}_1$

S_2 $(x_{31}, \ldots, x_{60})$ $\Rightarrow \bar{x}_2$

$\vdots$

S_{100} $\Rightarrow (\bar{x}_{100})$

If I take all these means and try to plot it on a histogram then

$\bar{X} \approx GD\left\{\mu, \frac{\sigma^2}{n}\right\}$.

If we take multiple samples, then

$\bar{X} = GD\left(\mu, \frac{\sigma}{n^2}\right)$ in simple bell curve.

Chebyshev's Inequality

$X \approx GD(\mu, \sigma)$.

$P_r(\mu - \sigma \le x \le \mu + \sigma) \approx 68\%$.

$P_r(\mu - 2\sigma \le x \le \mu + 2\sigma) \approx 95\%$.

$P_r(\mu - 3\sigma \le x \le \mu + 3\sigma) = 99.7\%$.

$y \not\approx GD$

$$P_r(\mu - k\sigma < x < \mu + k\sigma) \ge 1 - \frac{1}{k^2}$$

$k = 2$.

k specifies range of standard deviation

$$\therefore P_r(\mu - 2\sigma < x < \mu + 2\sigma) > 1 - \frac{1}{2^2}$$

$$= 1 - \frac{1}{4}$$

$$= \frac{3}{4} = 75\%.$$

Pearson Correlation Coefficient.

1. $\text{Covariance}(x, y) = \sum_{i=1}^{n} (x_i - \mu_x) * (y_i - \mu_y)$

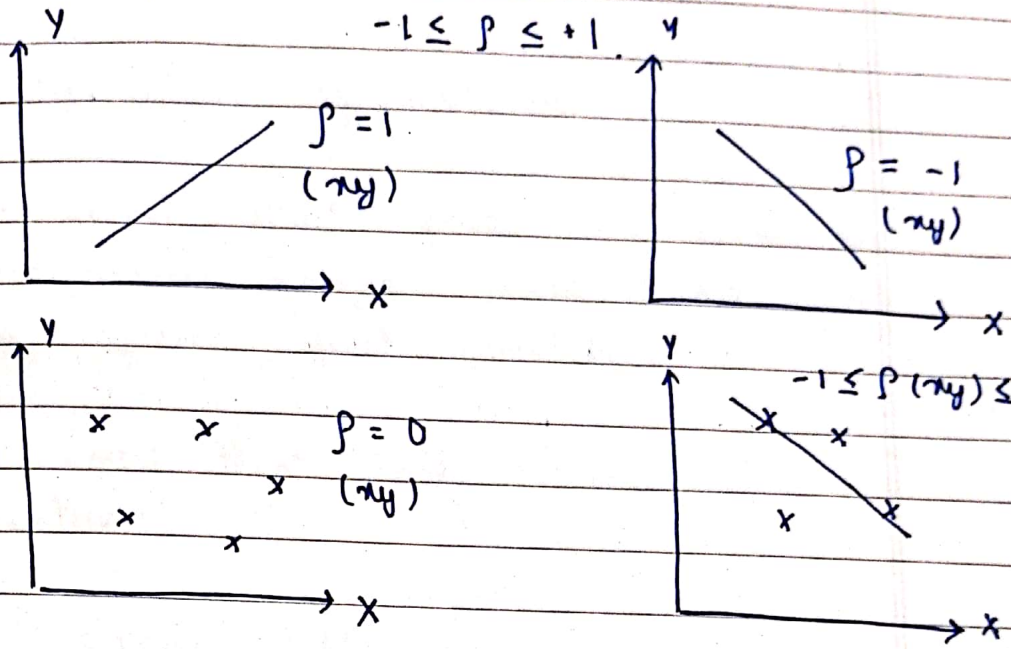
2. $\text{Pearson Coefficient} = \frac{\text{covariance}(x, y)}{\sigma_x \cdot \sigma_y}$

$$= \sum_{i=1}^{n} \frac{(x_i - \mu_x)(y_i - \mu_y)}{\sigma_x \sigma_y}$$

If $X\uparrow$ $Y\uparrow$ or $X\downarrow$ $Y\downarrow$.

+ve -ve

In covariance, we get to know the direction of relationship and do not know what is the strength of the relationship.



When two features have a very high correlation we can use either of them as a substitute for other.

Spearman's Rank Correlation

To find non-linear relationship between X & Y

We try to find Pearson correlation of ranks of X and Y.

Example:

IQ	Hours of TV / week
106	7
100	27
86	2
101	50
99	28
103	29
97	20
113	12
112	6
110	17

IQ	Hours / week (y_i)	rank x_i	rank y_i	d_i	d_i^2
86	2	1	1	0	0
97	20	2	6	-4	16
99	28	3	8	-5	25
100	27	4	7	-3	9
101	50	5	10	-5	25
103	29	6	9	-3	9
106	7	7	3	+4	16
110	17	8	5	+3	9
112	6	9	2	+7	49
113	12	10	4	+6	36

$$\rho = 1 - \frac{6 \Sigma d_i^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 194}{10(100-1)} = -0.175$$

with ρ-value = 0.62 using t-distⁿ

$\therefore$ Correlation between IQ & hours per week watching TV is very low

Finding outliers in data with Z-Scores.

An outlier is a data point in a dataset that is distant from all other observations. A data point lies outside the overall distribution of the dataset.

Identification of outliers :

1. Data point that falls 1.5 times outside interquartile range above third quartile and below first quartile.
2. Data point that falls outside of 3 standard deviations
3. If Z-Score falls outside of 2 standard deviation

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

Why should outlier exist in data?

1. Variability in data
2. An experimental measurement error

What are the impacts of having outliers?

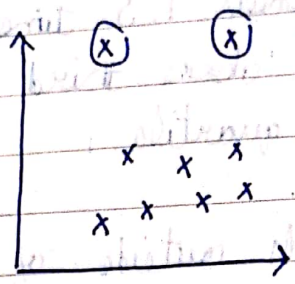
1. Causes various problems to statistical analysis
2. It impacts mean and standard deviation

IQR : 25% to 75% is $\frac{1}{4}^{th}$ to $\frac{3}{4}^{th}$

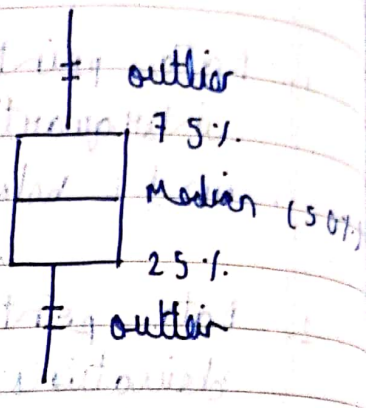
IQR = 75% - 25%

Finding outliers in dataset

1. Using scatter plots
2. Box plots
3. Using z-score
4. Using interquartile range.



Scatter



Boxplot

$$z = \frac{(x - \mu)}{\sigma}$$

```
def detect_outliers (data):

    threshold = 3.
    mean = np.mean (data)
    std = np.std (data)

    for i in data:

        z_score = (i - mean) / std
        if np.abs (z_score) > threshold:
            outliers.append (y)

    return outliers.
```

Interquartile Range

1. Arrange the data in increasing order
2. Calculate first quartile and third quartile
3. Find interquartile range = q3 − q1.
4. Find lower bound q1 * 1.5
5. Find upper bound q3 * 1.5

Anything that lies outside the interquartile lower & upper bound, we remove them.

Feature Scaling

1. Features has units and magnitude
2. Normalization helps to scale down the features between 0 to 1
3. Standardization is used to scale down the features based on standard normal dist^n

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

$$z_{std} = \frac{x - \mu}{\sigma}$$

($\mu = 0, \sigma = 1$) after standardization