Stat

Statistics

① Population → population mean
② Sample → Sample mean

mean, median, mode } measure of central tendency

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

③ Random Variable
- → Discrete Random Variable
- → Continuous Random Variable

Sample → Just taking an amount of data as a sample and create all the prediction stuff from here.

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

Random Variable

- → discrete → whole no., cannot be a floating no.
- → continuous → within a Range of Values we can have any Values

2 Gaussian dist / Normal Distribution

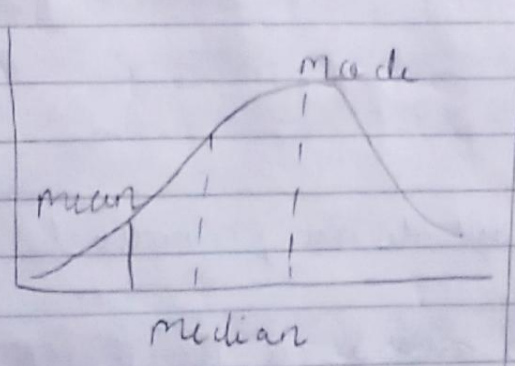
⇒ They have approximate a wide variety of Random Variables

⇒ Distributions of sample means with large enough sample sizes could be approximated to normal.

⇒ All computable statistics are elegant

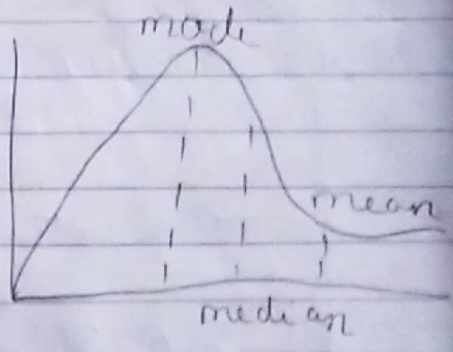
⇒ Decision based on normal distribution insights have a good track record.

Skewness



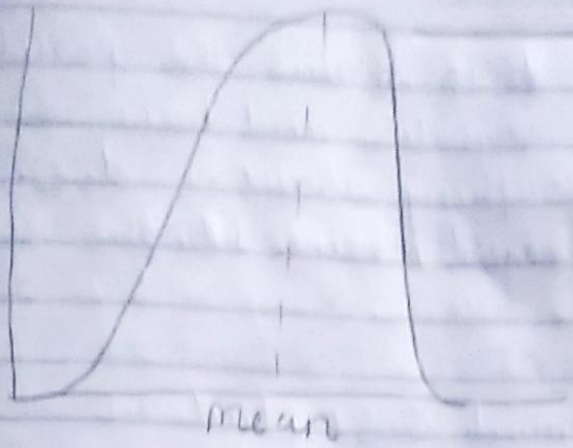
Negative direction

mean < median < mode



mode < median < mean

Positive direction



Symmetrical Normal Dist with no skewness

Mean - Median - mode

Denoted as :-

Normal → N
∽ → Distribution.
μ → mean
σ2 → variance.

N ∽ (μ, σ2).

→ A lower mean would result in the same shape of the distribution, but on the left side of the plane.

→ A higher mean would move to right.

Distribution :- A function that shows the possible values for a variable and how often they occur. Thing about a die. numbered from 1 to 6.

The Standard normal Distribution

→ $z = \frac{x - \mu}{\sigma}$ $\quad z \sim N(0, 1)$

$z \sim N(0, 1)$: mean, S.D.

z - Score

→ Adding and subtracting values from all the data Points does not change the standard deviation.

Covariance

The two variables are collected and which are correlated and the main statistic to measure this correlation is called Variance

It can be:-

> 0, The two variable moves together
= 0, The two variables are independent
< 0, " " " moves in opposite direction

Sample formula

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

population formula

$$\sigma_{xy} = \frac{\sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y)}{N}$$

Correlation

Co-relation adjusts covariance, so that the relationship between the two variables became easy and intuitive to interpret

$$\text{Correlation} = \frac{\text{cov}(x,y)}{\text{std dev}(x) * \text{std dev}(y)}$$

$$\frac{S_{xy}}{S_x S_y} \qquad \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$-1 \leq$ correlation coefficient ≤ 1

cov coeff $= 1$

The entire variability of one variable is explained by the other.

Causality = Important to understand the direction of casual relationships.

Standard Deviation Formula

$\sigma = \sqrt{\sigma^2}$

Population Standard deviation.

Sample Standard Deviation

$s = \sqrt{s^2}$.

Variance Of the Fall Data

→ 6, 8, 10, 12, 14, 16, 18, 20, 22, 24.

x_i	$d_i = x_i - 14$	Deviation from mean $(x_i - \bar{x})$	$(x_i - \bar{x})^2$
6	-4	-9	81
8	-3	-7	49
10	-2	-5	25
12	-1	-3	9
14	0	-1	1
16	1	1	1
18	2	3	9
20	3	5	25
22	4	7	49
24.	5	9	81
	5		330.

14 = is assumed mean.

Formulas

Skewness $= \frac{1}{n} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}^3}$

Variance and Std. deviation

Sample Variance formula :- $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

population Variance formula :- $\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N}$

Sample Standard deviation formula :- $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

population " " " :- $\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$

Central Limit Theorem

Sampling dist

$N\left(\mu, \frac{\sigma^2}{n}\right)$, $n > 30$ → Sample size

→ CLT allows us to perform tests, solve problems and make inferences using the normal distribution, Even when the Population is not normally distributed.

→ The mean of the samples we extract will be closer to normally distributed if we extract more samples.

→ The distribution of the sample mean is expected to have a mean equal to the mean of the orginal Dataset.

→ The distribution of the sample mean is expected to have a variance equal to the variance of the original dataset, divided by the sample size.

Standard Error

→ The Standard deviation of the dist formed by the sample means

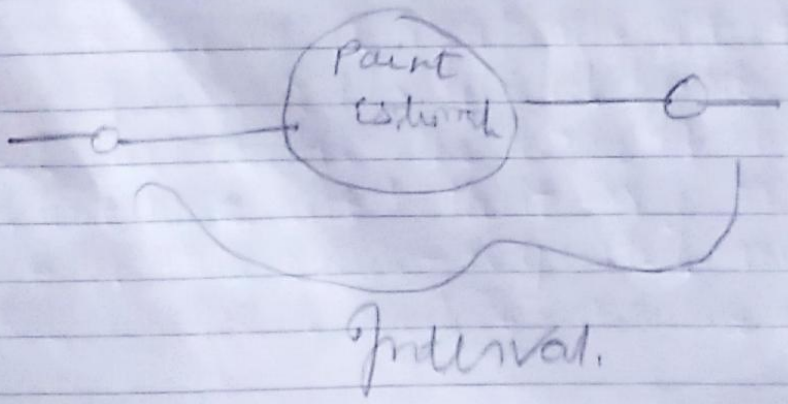
$$N\left(\mu, \frac{\sigma^2}{n}\right) \rightarrow \text{Variance}$$

$$S.D = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

→ Standard Error decreases when sample size increases.

Estimator and Estimates

① Point Estimates
② Confidence Interval Estimates.



Chebyshev's Inequality

→ $1 - \frac{1}{k^2}$

→ when $y \notin (a, b)$

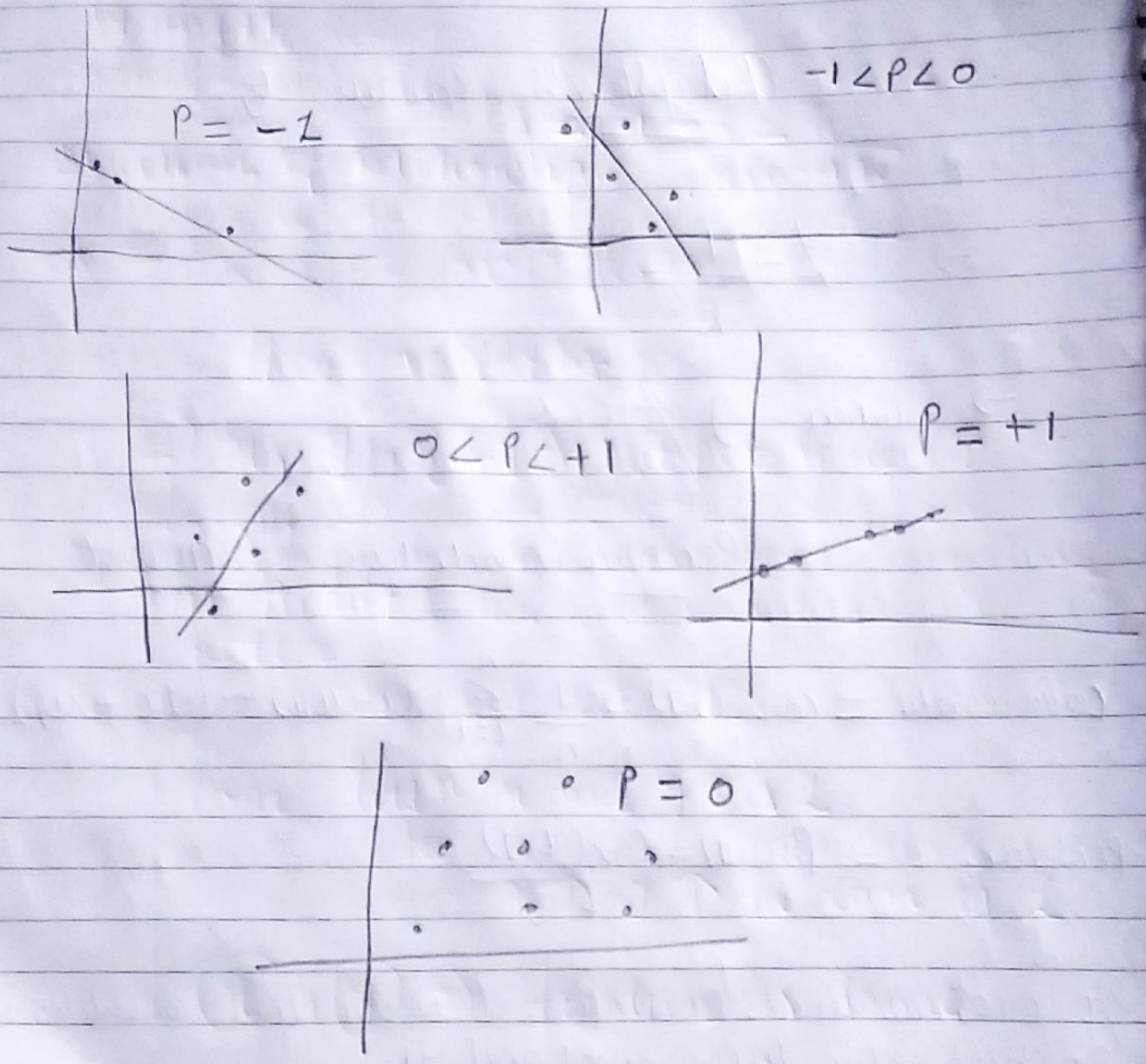
Pearson Correlation Coefficient

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

$$\text{Pearson cc} = \rho(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

→ It measures linear correlation between two variables x and y
→ It has value between +1 and −1.

+1 → Total positive linear correlation
−1 → Negative linear correlation.
0 → no linear correlation.



Defi :- Pearson Correlation Co-efficient is the covariance of the two variables, divided by the product of there standard deviations. The form of definition involves a "Product moment", that is, the mean of the product of the mean - adjusted random variables; hence the modifier product - moment in the name.

For a Population -

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

cov → covariance
σ_X → Standard deviation of X
σ_Y → Standard deviation of Y

For a sample -

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

where :

n = Sample size
x_i, y_i are the individual sample points indexed with i

$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ (The sample mean); for $\bar{y}$

1 Strength

2 Direction of Relationship

Three Variables.

{x, y}, {z} → output variable

↓

Independent feature

Corr between x and y is 1.

So

x ↑ y ↑ so corr is 1.

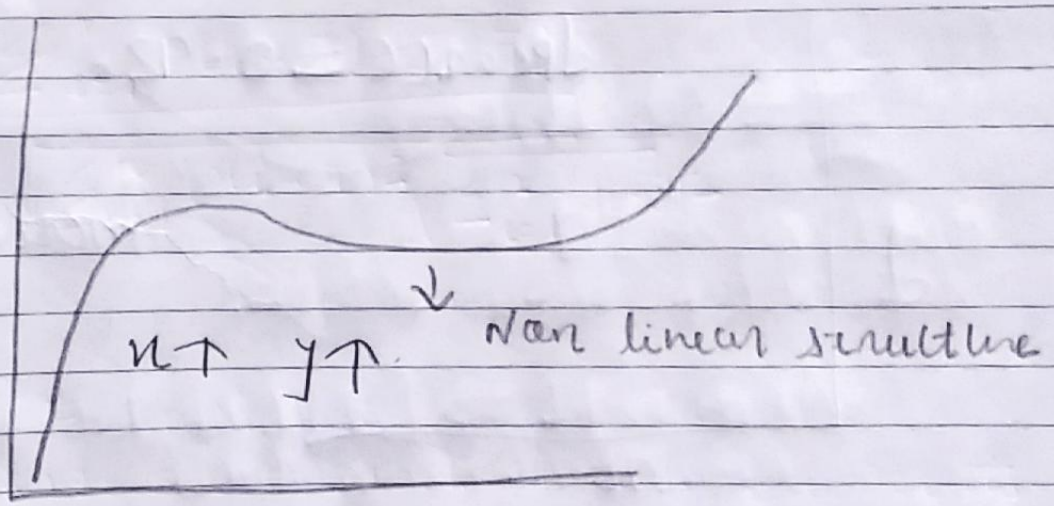
(x), (y) → feature is same.

So drop in (2) of the feature that can be x or y.

Spearman's Rank Correlation

→ The Spearman correlation coefficient is defined as the pearson correlation coefficient between the Rank variables.

→ Rank variables :→ bet^n a set of items such that for any two items) the first is Ranked higher , Rank lower than or Ranked equal to the second



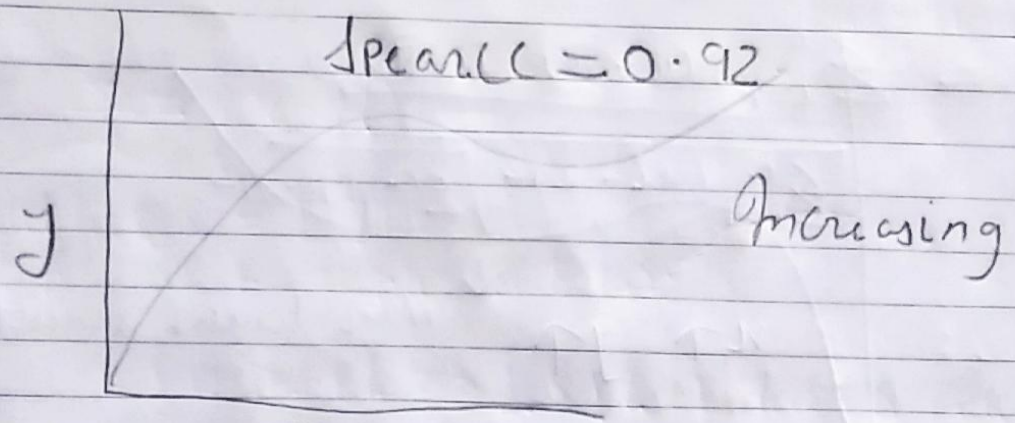
Most Important :- when the data are roughly Elliptically distributed and there are no prominent outliers, the Spearman correlation and pearson correlation give similar values

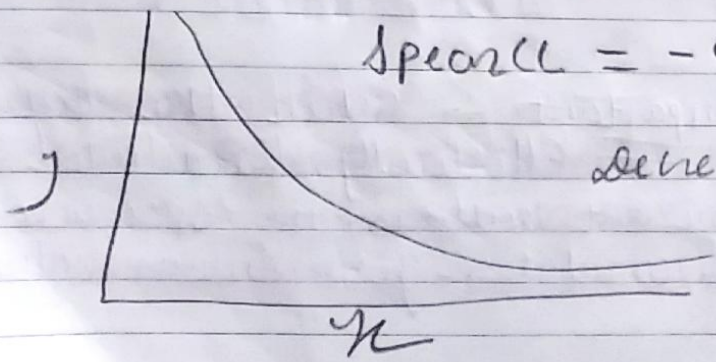
Formula

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

$$r_s = \rho_{rg_x, rg_y} = \frac{cov(rg_x, rg_y)}{\sigma_{rg_x} \sigma_{rg_y}}$$

ρ = Pearson correlation coefficient
$cov(rg_x, rg_y)$ → covariance of the Rank variables
σ_{rg_x} and σ_{rg_y} are the standard deviation of the Rank variables





Example

x_i	y_i
106	7
100	27
86	2
101	50
99	28
103	29
97	20
113	12
112	6
110	17

Rank (y_i) This is created by the highest number in y_i forest

(27) → 7th Rank highest value.

(2) → 2nd value.

(i) Arrange them in ascending order

x_i	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

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

tee Give

$$\rho = 1 - \frac{6 \times 194}{10(10^2-1)}.$$

which Evaluates tee $\rho = -29/165$

$$= -0.175.$$

with a p-Value $= 0.627$

Finding an outlier in a Dataset

→ What is an outlier?

→ An outlier is a datapoint in a dataset that is distant from all other observation. A datapoint that lies outside the overall distribution of the dataset

Techniques:
① Z-Score
② IQR → Inter quartile Range.

→ What is the Reason for an outlier to exists in a Dataset?

② Variability in the data.
② An Experimental measurement error.

→ What are the impacts of having outliers in a Dataset?

② It causes various problems during our analysis
② It may cause a significant impact on the mean and the standard deviation.

Various ways to find

1. Scatter plot
2. barplot
3. Using z score
4. Using IQR

Detecting outlier using z-score:-

Formula = (observation - mean) / standard deviation

$$Z = (x - \mu)/\sigma$$

```
outliers = []
def detect_outliers(data):

    threshold = 3 [3 std if data is falling not an outlier, if falling away from 3 then its outlier
    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
```

outlier_pt = detect_outliers(dataset)
outlier_pt

IQR

75.1. - 25.1. Values in dataset

Steps

① Arrange the data in increasing order.
② calculate first (q1) and third quantile(q3)
③ Find Interquartile range (q3-q1)
④ Find lower bond q1 * 1.5
⑤ Find upper bound q3 * 1.5

```
Sorted(dataset)
quantile2, quantile3 = np.percentile(dataset, [25, 75])
print(quantile1, quantile3)
```

→ Find the IQR.

```
iQR_Value = quantile3 - quantile2
print(iQR_value).
```

Find the lower bound value and the higher bound value.

lower-bound-val = quantile 1 - (1.5 * IQR-Value)
upper-bound-val = quantile 3 + (1.5 * IQR-val)

Print (lower-bound-val, upper-bound_val).

↓ ↓

Values Away from This bounds are considered as outliers.

Standardization vs Normalization

① Normalization - helps you to scale down your feature between 0 to 1.

② Standardization : helps you to scale down your feature based on standard normal distribution (mean = 0) & σ = 1).

∴ Normalization - (min max Norm)

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

```
df = pd.read_csv("To Read data",
                 header = None,
                 use_cols = [0,1,2]).
df.columns = ["class", "Alcohol", "malic"]
df.head().
```

from sklearn.preprocessing import minmaxscaler
scaling = minmaxscaler()
scaling.fit(df[['Allocal','male']])

② Standardization (Z-score normalization)

All the features will be transformed in such a way that it will have the properties of a standard normal distribution with mean(μ) = 0 and standard distribution (σ) = 1.

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

from sklearn.preprocessing import standard scaler

scaling = StandardScaler()

scaling.fit_transform(df[['Allocal','male']])