Structure of the Presentation

Friday, 28 June 2024 11:33 AM

1. What is statistics and why it is important for data scientists?
2. Important application of Mean and Geometric Mean
3. Introduction to probability concepts
4. Sampling method
5. Correlation and regression

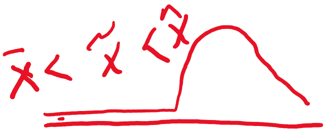
Statistics Defined

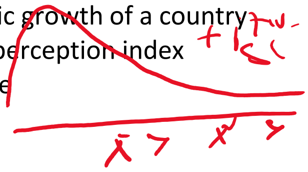
Friday, 28 June 2024 12:09 PM

STATISTICS is the science of collecting, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions.

★ Understanding statistical concepts is very important as these concepts are helpful in making informed decisions.

- Like business growth is associated with economic growth of a country
- Association of Fraud detection with corruption perception index
- Scientific productivity is related with literacy rate





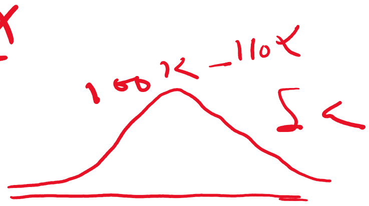
Mean
mode
Qualit.
non-num

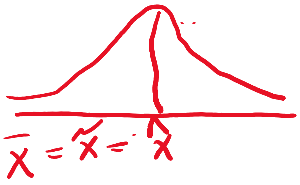
For instance

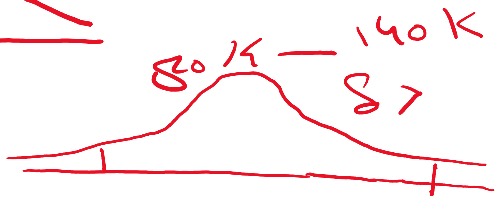
-3 0 +3

- Difference between type of variable guide the selection of statistical technique such as ~~Qualitative vs~~ quantitative variable
- Measurement scale (nominal and ordinal vs interval and ratio)
- **Population vs sample data**
- **Type of** distribution (symmetric vs asymmetric)
- Relative vs absolute analytical measurements
- Sample size
- Dispersion in the data

$$\bar{x} = \frac{\sum x}{n}$$







Special Applications of Arithmetic Mean (Mean)

What is Geometric Mean and Why it is important?
Special Application of Geometric Mean

AM Special Application Example

Friday, 28 June 2024 12:09 PM

Share price	Number of shares	
50	600	✓
40	700	✓
30	500	✓

$$\frac{120}{3} = \boxed{40}$$

40

1. You have an offer to sell all the shares at price 40/share. Will you sell the shares? Why or why not?

GM Applications

Friday, 28 June 2024 7:10 PM

10000

	Earning Growth
Year 1	10%
year 2	15%
Year 3	20%

$\frac{45}{3} = 15$

X

10000 x 10% = 1000 = 11000

comp

10%

Can you estimate the average annual growth in earning?

	bike	sugar
2008	Rs55,000	Rs32
2023	Rs160,000	Rs140
Change	105000	108

Price of which product move faster?

Survey of Probability Concepts (Beginning of Inferential Statistics)

Friday, 28 June 2024 12:09 PM

sol .5

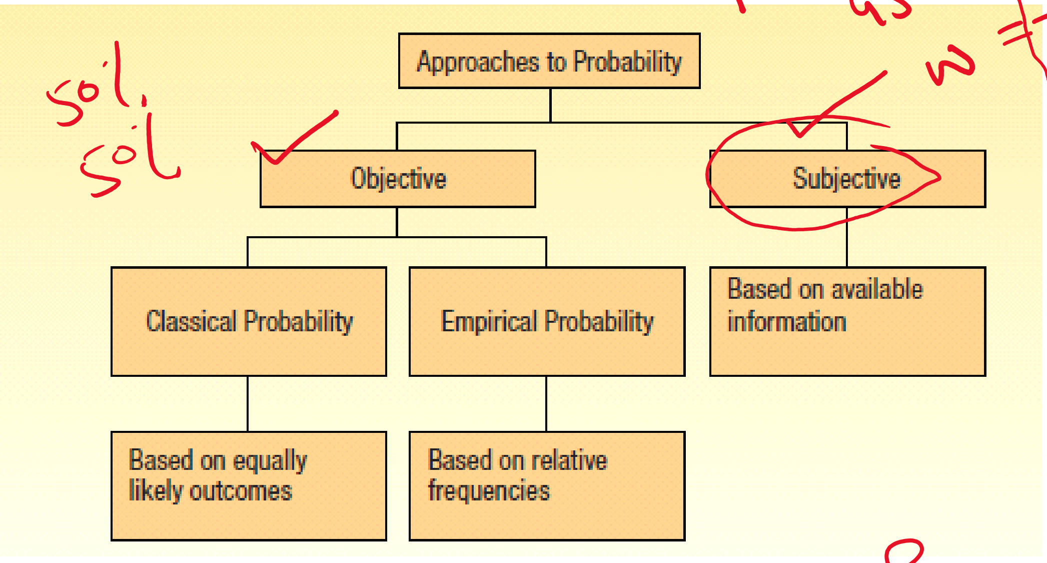
0 — 1

$F = \frac{150}{450} = 33\%$

$W = \frac{30}{50}$ 60%

PROBABILITY A value between zero and one, inclusive, describing the relative possibility (chance or likelihood) an event will occur.

sol, sol



Approaches to Probability
Objective
Subjective
Classical Probability
Empirical Probability
Based on available information
Based on equally likely outcomes
Based on relative frequencies

Important concepts in probability:

- Events
- Independent Events ✓
- Dependents Event
- Joint probability ✓

sat B R 5

2 3 4

3

1

1/4

2/5

2 4

X 0—3

45-46

45.45

Square

What is probability Distribution?

- Listing of all the possibilities along with their probabilities is known as probability distribution.

Types of probability distribution

- Discrete probability distribution
- Continuous probability distribution

Variable

X	P(x)
1	1/6
2	1/6
3	
4	
5	
6	

Disc

X	P(x)
H	1/2
T	1/2

To understand probability distribution, first understand the

- Random variables, its type and
- recap of frequency distribution

Sampling Methods

Saturday, 29 June 2024 9:34 AM

Probability sampling vs Non-probability sampling

Concept of generalizability and its relationship with sampling methods

Probability sampling techniques

- Simple random sampling
- Systematic random sampling
- Stratified random sampling
- Cluster sampling

Correlation and Regression (The Heart of Machine Learning)

Friday, 28 June 2024 8:00 PM

Same dir +

Independ X Y→ Depend Var

What is correlation analysis?

A group of techniques to measure the relationship between two variables (Independent and dependent variables).

Two approaches to access the relationship between two variables

- Graphical approach (scatter plot) using MS Excel and Python
- Mathematical approach using MS Excel and Python

Direction of relation

Strength of relation

ρ — row

r — Sample

r X-Axis

Adv Exp ↑ Sales ↑

Y +

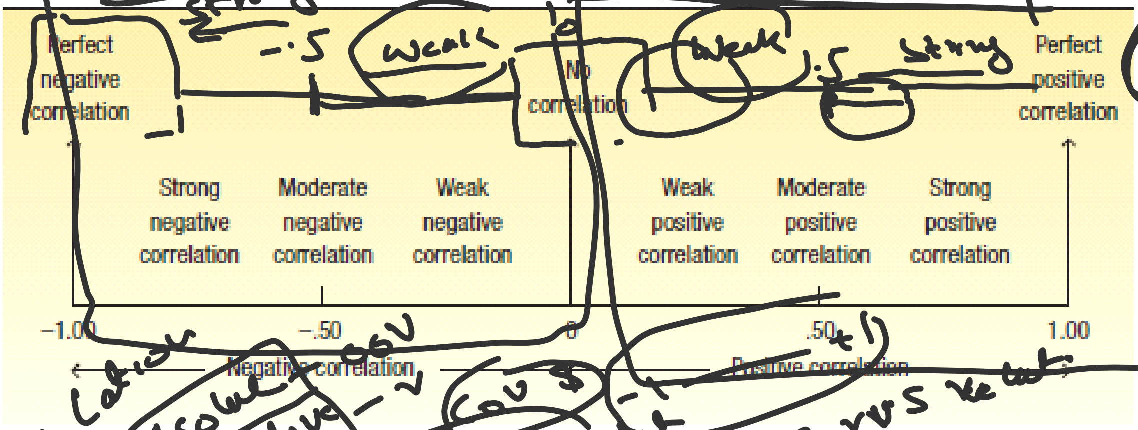
CORRELATION COEFFICIENT A measure of the strength of the linear relationship between two variables.

Quant. Var

CHARACTERISTICS OF THE CORRELATION COEFFICIENT

1. The sample correlation coefficient is identified by the lowercase letter r.
2. It shows the direction and strength of the linear relationship between two interval- or ratio-scale variables.
3. It ranges from -1 up to and including +1.
4. A value near 0 indicates there is little relationship between the variables.
5. A value near 1 indicates a direct or positive relationship between the variables.
6. A value near -1 indicates inverse or negative relationship between the variables.

numeric numeric


Perfect negative correlation
No correlation
Perfect positive correlation
Strong negative correlation
Moderate negative correlation
Weak negative correlation
Weak positive correlation
Moderate positive correlation
Strong positive correlation
−1.00
−.50
0
.50
1.00
Negative correlation
Positive correlation

r = 0 no rel

0.3

r = 0.75

r < .5

Strong -.5 Weak

Weak .5 Strong

Pop

100

relative

Absolute

relative

cov

Sports

Tips

100 Tip

corr

r = 0.8

Let's take an example of correlation analysis using:

- Graphical approach
- Mathematical approach

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

no-relation

What is R-square and how do we interpret it?

Does r represent absolute or relative measure of association?

X Adv 50

Sal

σ Stand Deviat Pop.

s Sample

What is Regression Analysis?

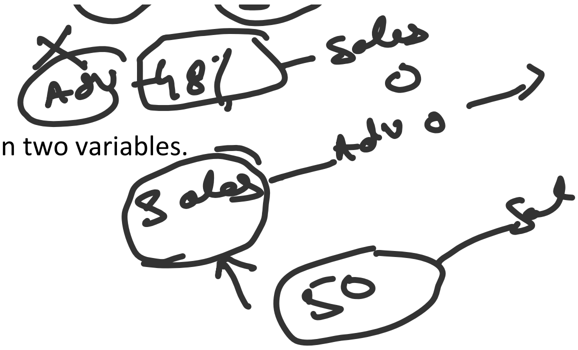
What is simple linear regression?

What is multiple linear regression?

Do we always confront with linear relationships?

Adv ↑48% Sales

Do we always confront with linear relationships?



Regression Analysis

An equation that expresses the linear relationship between two variables.

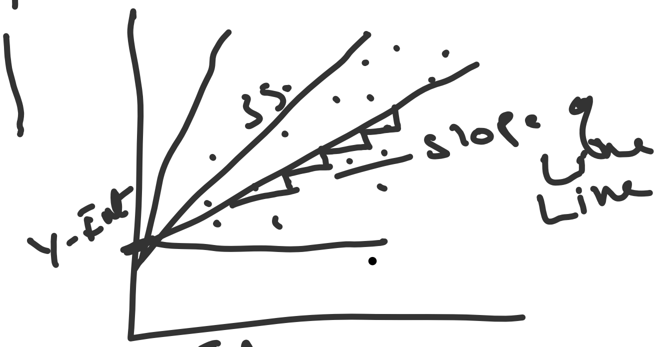
Role of Intercept and Slope coefficient

- Regression analysis is used for prediction
- What are the various applications of regression analysis specifically in data science?
- What is the role of regression analysis in building machine learning models?

L

2n

x y



Y-Int

$Y = \beta_0 + \beta_1 X_1$ Slop

pw — n

$Y = 5 + 2.5 X_2$

7.5] 2.5
10

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 - - - - \beta_n X_n$

Y-Int

$Y = a + bX$ slope of Line

$Y = a + bX$ — Simple ✓

$Y = a + b_1 X_1 + b_2 X_2$ — Mut. ✓

100 mil

$X_1 = 2$

$X_2 = ?$

$Y = a + b$

Null Hyp H_0: ____

Alternate H_1: ____

Hyp	Beta Coff	t-value Z-value 1.967	P-Value
H_1	0.75	2.5	0.001
H_2	0.02	1.58	0 09 No sig
H_3			

$H_0: \bar{X} = 0$ X

$H_1: \bar{X} \neq 0$ ✓

$H_0: \bar{X} = 0$ ✓

— $H_1: \bar{X} \neq 0$ X

$\alpha = 5\%$