Statistics Essential for Data Science.

Population: - Collection of all items of interest to our study.
- N (parameter)

Sample: - A subset of a population
- n (statistics)

Types of Data

Categorical → Nominal
→ Ordinal

Numerical → Discrete, Continuous

- Frequency distribution Charts
- Bar Charts
- Pie charts
- Pareto Diagram

- Histrogram Charts

To show relationship b/w two numerical or categorical identity → Cross Table (side by side bar Charts)
→ Scatter Plot

Frequency: Measure of occurence of variable

Relative Frequency: Measure the relative No. of occurance of variable. usually expressed in percentage.

Measurement level

Qualitative → Nominal (can't be put in any order) (eg. seasons name), Ordinal

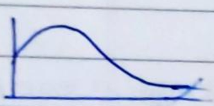
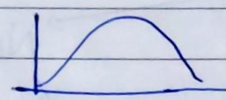
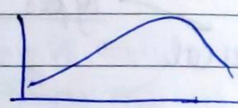
Quantitative → Interval (not have true 0) Eg temp., Ratio (have true 0)

Mean: $\frac{\sum_{i=1}^{N} x_i}{N}$ • μ (population) • $\bar{x}$ (Sample)

easily affected by outliners

measure of asymmetry

Skewness : It indicates ~~wea~~ whether the data is concentrated on one side

+ve / right	0 / symmetric	-ve / left
- mean > median - outliers are at right (tail)	- mean = median - no c	- mean < median - outliers are at to the left
		

$$\tilde{\mu}_3 = \frac{\sum_i^N (X_i - \bar{x})^3}{(N-1) * \sigma^3}$$

Variance : It measures the dispersion of a set of data points around their mean

Value obtained is parameter

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

value obtained is statistic

sample variance $= s^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{n-1}$

? Why Squaring
(i) Dispersion in non-negative
(ii) Non negative value don't cancel out
(iii) Amplifies the effect of large differences

Standard Deviation:

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

Sample SD $= s = \sqrt{s^2}$

Coefficient of Variation CV: relative standard deviation

population CV $= C_v = \frac{\sigma}{\mu}$

sample CV $= \hat{C_v} = \frac{s}{\bar{x}}$

Used to compare datasets in terms of variability

Covariance : $-\infty$ to $+\infty$

- population $\sigma_{xy} = \dfrac{\sum_{i=1}^{N}(x_i - \mu_x) * (y_i - \mu_y)}{N}$
- Sample $s_{xy} = \dfrac{\sum_{i=1}^{N}(x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$

Covariance may be : >0, the 2 move together
$=0$, the 2 are independent
<0, the 2 move opposite

Correlation Coefficient : It adjusts covariance, so that the relationship b/w 2 variables became easy and intuitive to interpret.

population $\dfrac{\sigma_{xy}}{\sigma_x * \sigma_y}$

Correlation doesn't imply causation

sample $\dfrac{s_{xy}}{s_x * s_y}$

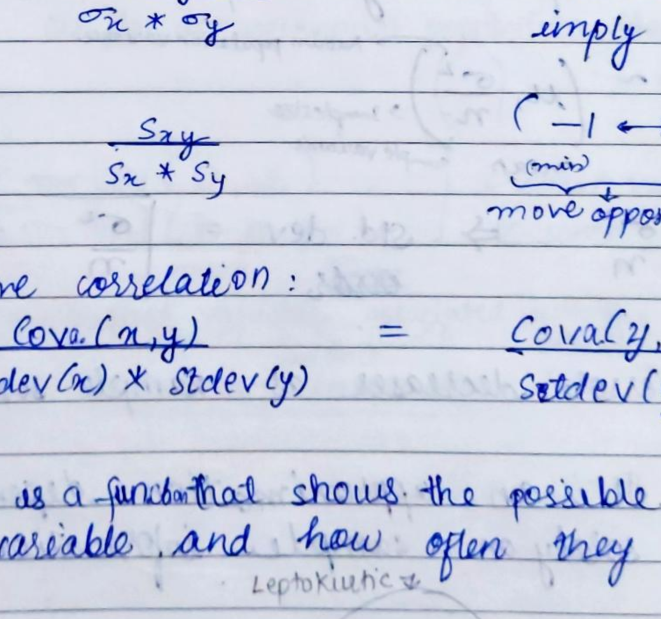
$(-1 \leftrightarrow 0 \leftrightarrow +1)$
(min) (max)
$\underbrace{-1 \leftrightarrow 0}_{\text{move opposite}}$ | 0 → independent | $\underbrace{0 \leftrightarrow +1}_{\text{move 2gether}}$

Perfect positive correlation :
$$\frac{\text{Cova}(x,y)}{\text{Stdev}(x) * \text{Stdev}(y)} = \frac{\text{Cova}(y,x)}{\text{Stdev}(y) * \text{Stdev}(x)}$$

Distribution : It is a function that shows the possible values for a variable, and how often they occur.

Normal distribution

$N \sim (\mu, \sigma^2)$ (mean, variance)



no skew
symmetrical
Bell curve
Gaussian distribution

Standard normal distribution (Z): It is a special case of normal distribution

$$N \sim (\underset{\mu}{0}, \underset{\sigma}{1})$$

⇒ How to convert Normal Distribution to standard ND?

$$\text{Z-score} = \frac{\text{Original value} - \text{Mean}}{\text{Standard deviation}} \quad \text{or} \quad \frac{x - \mu}{\sigma}$$

Central limit theorem: It states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approx normally distributed

Standard Error: The Std. deviation of the distribution formed by sample means $\bar{x}$

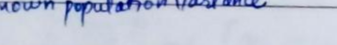
$$\text{so} \quad N_{(\text{sample})} \sim \left(\underset{\text{mean}}{\mu}, \underbrace{\frac{\overset{\text{known population variance}}{\sigma^2}}{\underset{\text{sample size}}{n}}}_{\text{sample variance}}\right)$$

$$\text{sample variance } \cancel{\text{error}} = \frac{\sigma^2}{n} \Rightarrow \text{std. dev } \cancel{\text{error}} = \sqrt{\frac{\sigma^2}{n}} \Rightarrow \frac{\sigma}{\sqrt{n}}$$

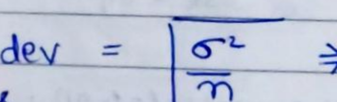
★ Std error decreases as sample size increases

Estimator It is an approximation depending solely on sample information

(i) point estimate



(ii) confidence interval estimate



interval

It is the range within which you expect the population parameter to be

Estimator of Parameter → Estimate
/how to estimate/ /what to estimate/ /concrete result/

Property $\longrightarrow$ Bias

$\downarrow$

Efficiency

An unbiased estimator has an expected value equal to the population parameter

The most efficient estimattor is the unbiased estimator with the smallest variance

Tscore $\underbrace{t_{n-1,\alpha}}_{\text{degree of freedom}} = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

oxi mately

A t-score is one form of a standardized test statistic
Used when: (i) Sample size less than 30
(ii) Has an unknown population ~~standard~~ variance.

$$\left[\begin{array}{l} Z \sim N(0, 1) \\ Z \sim N(\underset{\text{mean}}{\bar{x} - \mu_0}, \underset{\sigma}{1}) \end{array}\right]$$ critical value / Z-score

it is standardized variable associated with the test

cofidence intervals, population variance unknown, t score

$$\left[\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}, \quad \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right]$$

confidence intervals, population variance known, Z-score

$$\left[\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$ reliability factor (t, z)

Margin of Error

$\bar{x} \pm ME$

confidence level $= 1 - \alpha$

mean $\pm$ z score × std error
t-score

Steps in data-driven decision-making

(i) Formulate a hypothesis → (ii) Find the right test → (iii) Execute the test → (iv) Make a decision

Hypothesis: "A idea that can be tested"

Hypothesis	Notation	
Null	H_0	(one to be tested)
Alternative	H_1 or H_A	(everything else)

Significance level: α, It is probability of rejecting the null hypothesis, if it is true

Error Types: Types I
- False +ve
- Rejecting true null hypothesis.
- Probability of occuring = α
- More dangerous

Type II
- False -ve
- Accepting false null hypothesis
- Less dangerous
- Probability of occuring : β

(β : depends mainly upon → Sample size
→ population variance)

Power of test : $1-\beta$

Rejecting false null hypothesis or,
~~Accepting true null hypothesis~~

Testing : It is done by standardizing the variable at hand and comparing it to the z (critical value)

Reject if : Absolute values of Z-score > z

p-value : p-value is the smallest level of significance at which we can still reject the null hypothesis given the observed sample statistic

Reject if $\alpha > p$ Accept if $p > \alpha$

(evidence against null hypothesis) (smallest value = strongest evidence)

(Previous Topic) Confidence Interval Estimator ; 2 sample

(i) Dependent sample

$$\bar{d} \pm t_{n-1,\alpha/2} \frac{S_d}{\sqrt{n}}$$

(difference mean)

diff mean + std error × t-score

(ii) Independent → Population Variance known

$$(\bar{x}-\bar{y}) \pm Z_{\alpha/2} \underbrace{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}_{\text{variance of the difference}}$$

difference point estimator

mean difference

(iii) Independent → PV not known → assumed equal

std dev

$$\text{pooled sample variance } (S_p^2) = \frac{(n_x-1)S_x^2 + (n_y-1)S_y^2}{n_x+n_y-2}$$

$$\underbrace{(\bar{x}-\bar{y})}_{\text{mean difference}} \pm t_{n_x+n_y-2,\alpha/2} \sqrt{\frac{S_p^2}{n_x} + \frac{S_p^2}{n_y}}$$

{ As standard normal distribution is symmetrical around 0, the 2 statement are equivalent:
$-4.67 <$ a negative z <=> $4.67 >$ a positive z }

(iv) Independent → PV unknown → assumed different

$$(\bar{x}-\bar{y}) \pm t_{\nu,\alpha/2}\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}$$

$$\nu = \frac{\left(\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}\right)^2}{\left[\frac{s_x^2}{n_x}\right]^2/(n_x-1)+\left[\frac{s_y^2}{n_y}\right]^2/(n_y-1)}$$

- p-value (one-sided) = (1 - value in table)
- p-value (two-sided) = (1 - value in table) × 2

- The closer to 0.000 the p-value, the better. (We rejects null hypothesis at all and uncommon)
- P-value is a ~~normal~~ universal concept that works with every distribution

<u>Linear Regression</u> - A linear regression is a linear approximation of a causal relationship b/w 2 or more variable.

Process: Get sample data
↓
Design a model that works for that sample
↓
Make prediction for whole population

population formula $y = \beta_0 + \beta_1 x_1 + \varepsilon$ → Error

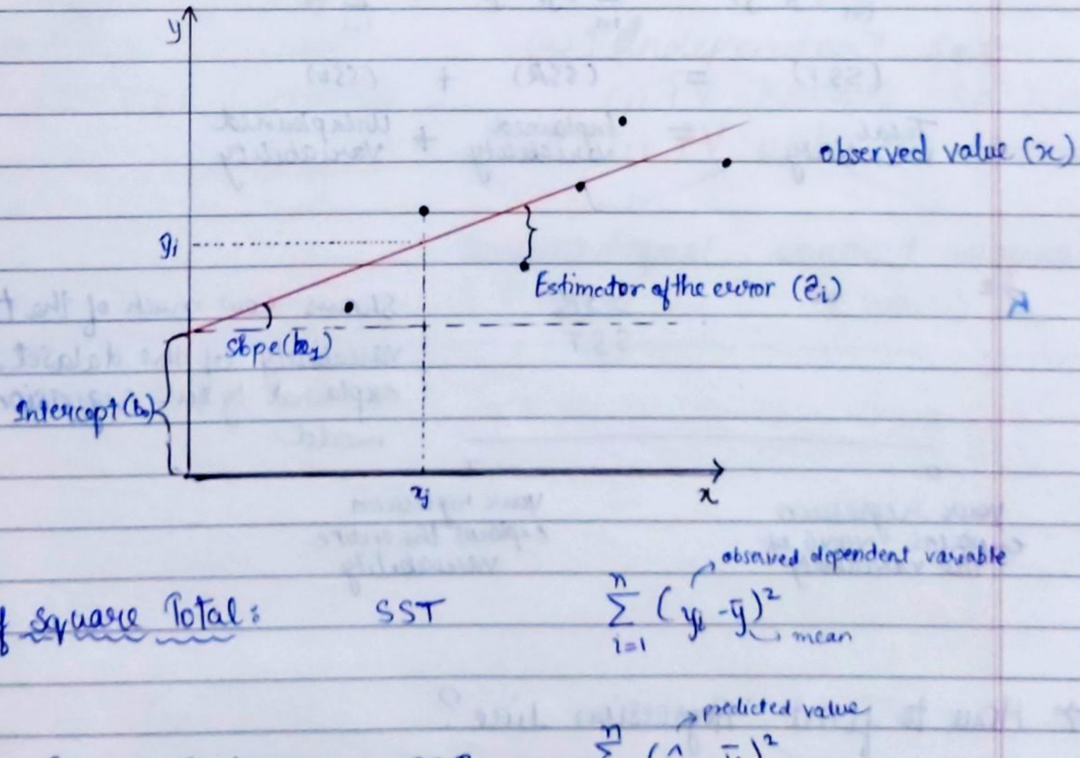
y ↓ Estimated or Predicted value

β_0 ↓ constant

x_1 ↓ Independent variable

$$\hat{y} = b_0 + b_1 x_1$$

Correlation	Regression
* Relationship	* One variable affects the other
* Not capture Causality	* Cause and effect
* $p(x,y) = p(y,x)$	* One way
* single point	* Line

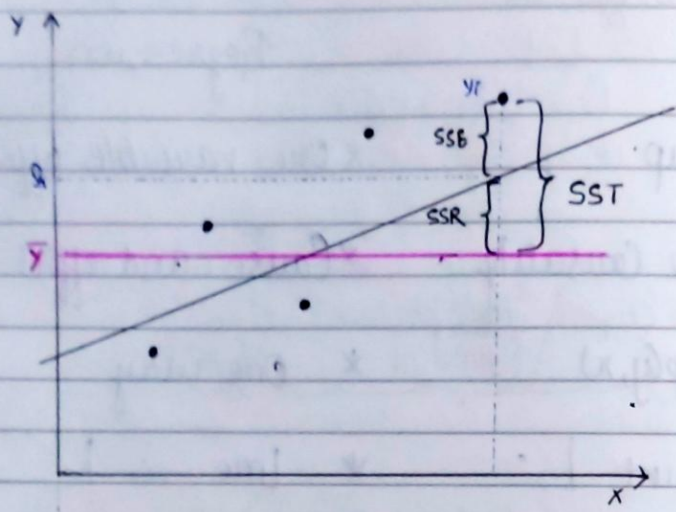


Sum of Square Total : SST $\sum_{i=1}^{n} (y_i - \bar{y})^2$ (observed dependent variable; mean)

Sum of Square Regression : SSR $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ (predicted value)

Sum of Square Error : SSE $\sum_{i=1}^{n} e_i^2$

$$SST = SSE + SSR$$



$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2$$

$$(SST) = (SSR) + (SSE)$$

Total variability = Explained Variability + Unexplained Variability

$R^2 = \dfrac{SSR}{SST}$ Shows how much of the total variability of the dataset is explained by our regression model

0 — your regression explains NONE of the variability

1 — your regression explains the entire variability

→ How to find Regression line?

Ordinary least squares (OLS)

min SSE

S(b) is the OLS estimator of β for a simple linear regression

$S(b) = \sum_{i=1}^{n} (y_i - x_i^T b)^2 = (y - Xb)^T (y - Xb)$

linear algebra

Other methods

- Bayesian method regression
- Kernal method regression
- Gaussian processe regression

Confidence Interval

1 dataset

(1) Popuplation Variance Known (z score)
(2) Population Variance unknown (t-score)

2 data set

(1) Dependent Set (t-score)
(2) ~~I~~ Independent set
(i) PV Known (z-score)
(ii) PV unknown ~~(t-score)~~

assumed equal (t-score)
assumed unequal (t score)
