# 1. Intro to Data & Variables

* {Data Analysis} → Story telling with data → information

Journal : Noise v Signals

* Anatomy of Dataset

→ 'Tidy' data
- Columns → data type
- Rows → data / Quantity ...

* Types of Data
  - * Nominal, Ordinal — Catagorical
    - String = "TEXT" — Catagories
    - Dates / Time
    - Numeric → Currency
  - Types of numeric data
    - → Number
    - → Temprature | Absolute Zero
    - → Currency
    - → Percentage

4 TYPES of Measurements

Nominal - Catagorical, no order
Ordinal - Catagorical, Ordered
Interval - numeric w/ 0 arbitary
Ratio - numeric w/ 0 absolute

$X$ - set of number

$X_{row, column}$

$x_i$ - data element ($i^{th}$)

$\Sigma ( )$ $\Sigma_i$ → summation

→ This first

PEMDAS

1st 2nd

$E[X]$ = expectation / Expected Value

---

= Sum ( B# : B* )

= Avg ( D# : D* )

$$\bar{x} = \frac{\sum_1^n x_i}{n}$$

Mean

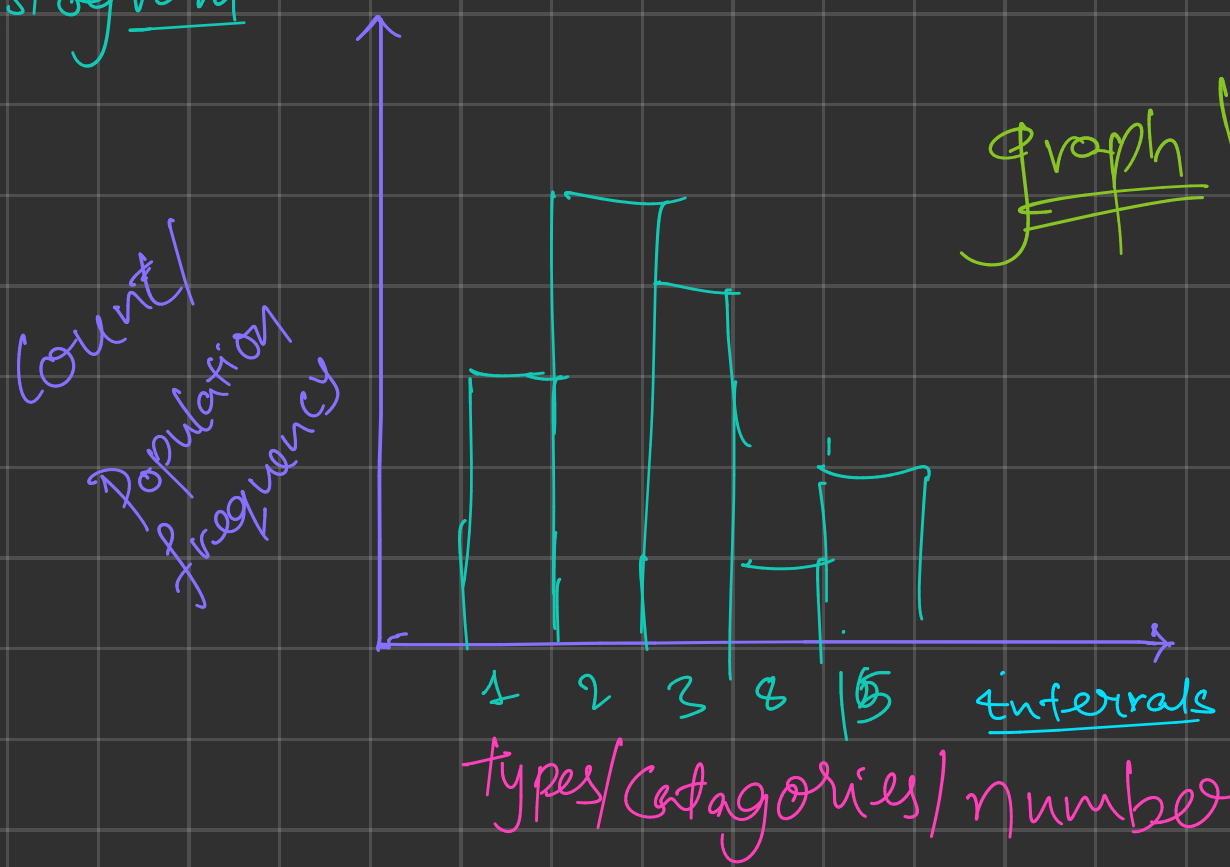
$$\sigma = \sqrt{\frac{\Sigma ([x_i - \bar{x}])}{(n-1)}}$$

ST DEV

Standard Deviation $\sigma^2 \Leftrightarrow$ Variance

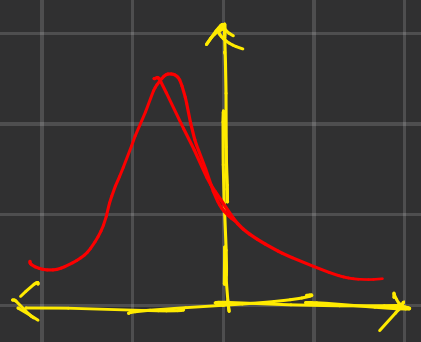
# 2. Summarization Data w/ Distribution Graphs

* Histogram



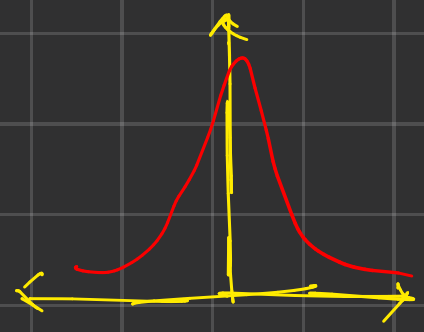

Graph!

* Skewness = skew(num....)



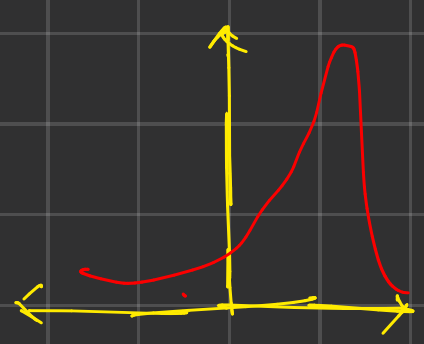
Pos. Skew

SK = +



No Skew

SK = 0



Neg. Skew

SK = -

## * Count distribution

- Bound at zero
- Often many zeros
- Generally positive skewed.

→ Other distribution

- Uniform Distribution
- Normal (Gaussian)
- Log-normal
- Poisson
- Geometric
- Weibull
- Binomial
- Bernoulli
- Cauchy
- Exponential
- Erlang (Gamma)

## * Bar Graphs & Pie Charts

→ Histogram : Continuous x-range

→ Bar g. : Discrete / categories

✓ Bar vs Pie ×

$$x\% = \frac{x}{100} \times 360^\circ$$

# 3. Summary Statistics

## Mode

Most frequent score

→ categorical data

Sampling stability

## Mean ($\bar{x}$)(M)

Pro & Cons

Rep. all data

outliers can bias!

$$\bar{x} \pm (n * \sigma)$$

$$\sigma = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}}$$

dist. from mean

$$\hat{\sigma} = \sqrt{\frac{\sum_i ((x_i - \bar{x})^2)}{n-1}}$$

Always $\frac{\sum_i (x_i - \bar{x})}{n} = 0$

## Median

Half data above

Half data below

→ Not biased by skew.

## SD vs IQR

Standard Deviation

→ Mean is not enough info

→ Alternative to $\sigma$. (w/ outliers)

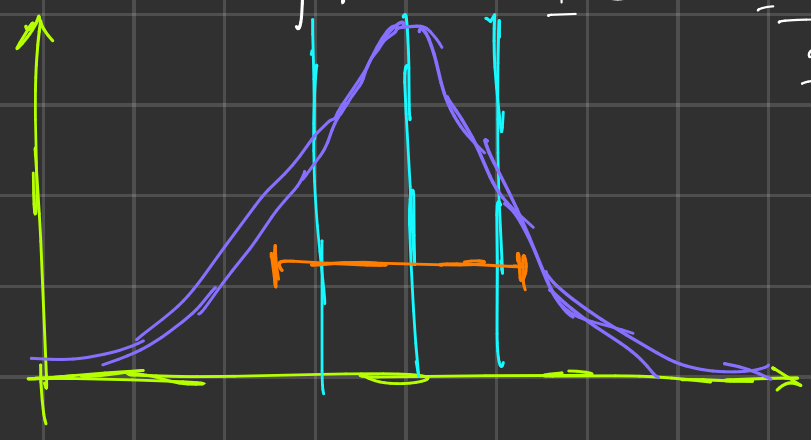
Inter Quartile Range

(4 parts)

→ Semi Inter-Quartile Range

→ 25th percentile - 75th p.

↳ Half point $\pm \sigma$ $\tilde{x} \pm \frac{IQR}{2}$



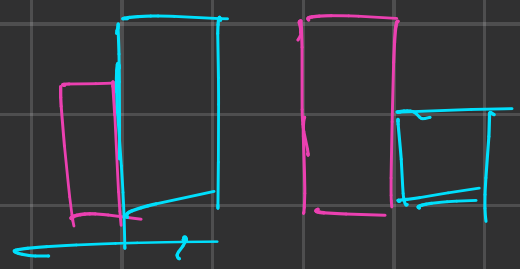
= Quartile ( A# : A* , (#of quartile^th percentile)

= quartile.Exc ( " " " " " )

Proportions ⟶ are x not averages

Cross tabs
↳ categorical data

| | sugar | no sug |
|---|---|---|
| Decaf | x | y |
| caff | z | w |



# 4. Business Statistics

* Business Math + KPI Fundamentals
  ↳ key Performance Indicators

Display + Context +

* KPI Example

- Revenue - Cost/operations - Incident Resol. Time

Note: These are industry & company dependent

* KPI Exercises

Business Analytics
Financial Analytics
Marketing Analytics
IT Analytics

# 5. Introduction to Inferential Math and Forecasting

* Forecasting Concepts

→ Future → Null Hypothesis Testing

Yes/No ← "Real trend" ←

→ Time Series Analysis

Time dependent ←

Multiple Points

One point

* T- testing → Comparing 2 groups

Null Hypothesis Testing

$diff. = \bar{x}_1 - \bar{x}_2$

Ho Null → Real − Signal

Idiosyncrasy − Noise

Contradict

- p- Value → Minimal!

To reject N.Hypo. needs to be next to impossible

$P < 0.5$ → general but Lower

Prefer

- T-test

→ $\bar{x}$ $\sigma$ $n(x)$

$$t = \frac{\bar{x}_1 \quad \bar{x}_2}{\sqrt{\left[\frac{\hat{\sigma}_1 (n_1 - 1) + \hat{\sigma}_2 (n_2 - 1)}{(n_1 + n_2 - 2)}\right]\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

$(n_1 - 1) + (n_2 - 1)$ → $df$ : degrees of freedom

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left[\frac{\sigma_1 n_1 + \sigma_2 n_2}{(n_1 + n_2)}\right]\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

PValue

↳ = TDIST( ABS( ↓ ), df, 2 )

= TValue

Ref : Crash Course Statistics !

# * Time Series Forecasting -

→ Annual → Semi annual → Quarterly
→ monthly → Weekly → Daily → Hourly

decomposing w/ diff scales