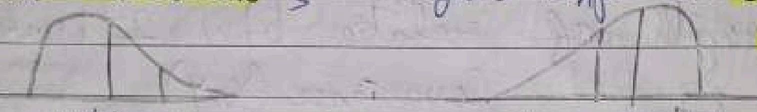
## Descriptive Statistics → to describe data

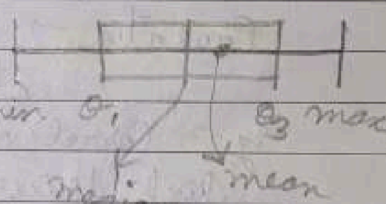
- Measures of central tendency → mean median, mode
  median more useful in data with outliers
  mode not significant for continuous data (high PDF point)
- {Mean − Median} → gives info on skewness

Right skewed

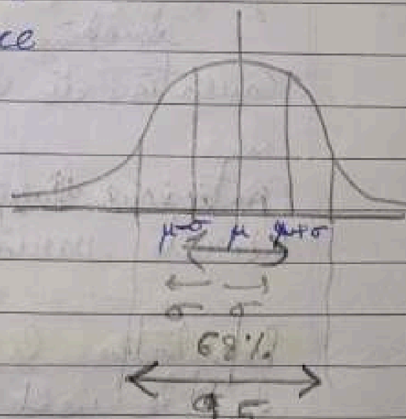



left skewed

- Measures of dispersion/spread:
  Range, IQR (Inter Quartile Range) $= Q_3 - Q_1$
  ↳ max − min
- ~~Box Plot / Whiskers~~
  Box and Whiskers Plot




- More measures of dispersion:
  Standard Deviation, variance
- 68 − 95 − 99.7 rule →
  Valid on bell shaped data
  (Perfectly apply on normal distribution)




- Chebyshev's Theorem
  Atleast $\left(1-\frac{1}{k^2}\right)^{th}$ of data lies within
  $\pm K$ standard deviations regardless of shape of distribution
  (75% data lies b/w $\mu - 2\sigma$ & $\mu + 2\sigma$)

Measures of association → Covariance, correlation, Causation

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

Covariance changes with unit of data hence we use correlation. Correlation measures strength of relation b/w 2 variables

$$\text{Correlation} = \frac{\text{Covariance}(X,Y)}{\text{Stdev}(X)\ \text{Stdev}(Y)}$$

Correlation $\in [-1, 1]$

Causation → proving that 1 variable causes other. Not the same as correlation

eg → Correl(occupancy, hotel prices) = + high but high prices $\not\leftrightarrow$ high occupancy

eg → Correl (Smoking, Cancer) = +1 & Smoking causes cancer

Causation is huge topic out of scope

Continuous variables → PDF Probability distribution fn

Discrete variables → PMF Prob Mass fn

Population & sample

Sample used as usually we can't access whole population, its cost effective & feasible

Sampling techniques such that sample is good representative of population → whole field in marketing

Random Sampling

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$\mu$ = Population mean
$\sigma$ = Population Std Dev
$\bar{x}$ : Sample Mean
s : Sample Std Deviation

**Central Limit theorem**

Sample mean is normally distributed with mean equal to population mean, irrespective of distribution type of population (be it unimodal, multimodal, symetric, skewed, discrete, continuous)

$\bar{x} \sim \text{Normal}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

**Bernoulli Process / Trial** → 2 outcomes only → win or lose

**Binomial distribution** → n trials (independent)

success = p   failure = 1 − p

Random variable X = no of successes in n trials

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{1-x} = {}^nC_x \, p^x (1-p)^{1-x}$$

**Binomial Distribution**   mean = np   Variance = npq

Eg → No of fraud reports among n tax reports, no of students passing exam

here $x \in [0, n]$ (Unlike Poisson)

**Poisson Distribution**   $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$

Mean = $\lambda$ = Variance   $x \in [0, 1, 2 \ldots \infty]$

**T / Student T distribution** → symetric, centered at 0, only 1 parameter → dof (degrees of freedom)

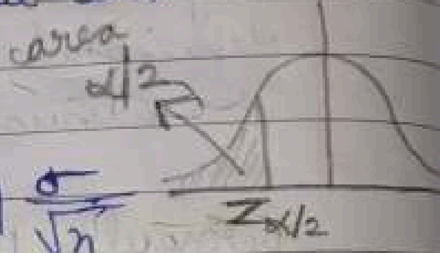
As DOF → ∞, T distribution → Standard Normal Distribution

Confidence interval

Z statistic $= \dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim$ Normal $(0,1)$

T statistic $= \dfrac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{n-1}$

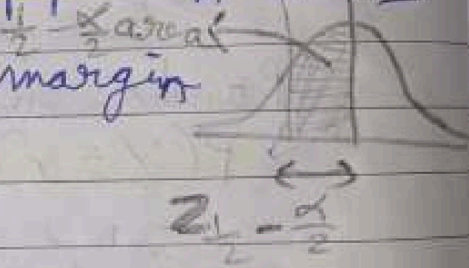
$\sigma \rightarrow$ population std dev $\quad s \Rightarrow$ Sample std dev




Constructing confidence interval $\rightarrow$

$$\bar{x} - |Z_{\alpha/2}|\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + |Z_{\alpha/2}|\frac{\sigma}{\sqrt{n}}$$

$\rightarrow$ [$(1-\alpha)$ confidence interval for population mean]




[for $\alpha$ confidence interval, $|Z_{\frac{1}{2}-\frac{\alpha}{2}}|\frac{\sigma}{\sqrt{n}}$ margin which makes intutive sense]

If $\sigma$ not known use $s$ &

$\pm|t_{\alpha/2}|\dfrac{s}{\sqrt{n}}$ = margin

~~For Pop~~ Confidence Interval for population ~~For~~ proportion

$$\hat{p} - |Z_{\alpha/2}|\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + |Z_{\alpha/2}|\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

(we don't use T distribution for population proportion)

$p \rightarrow$ population proportion

$\hat{p} \rightarrow$ sample proportion

Sample size, how big to take? $\rightarrow$

Use that industry rule of thumb

Or get % confidence needed, % tolerance (margin of error)

put in above eqn get '$n$'

Hypothesis testing → Null Hypothesis $H_0$, Alternate Hypothesis $H_A$

Steps →

i) Formulate Hypothesis $H_0$ & $H_A$

ii) Calculate T statistic $= (\bar{x} - \mu) / (s/\sqrt{n})$

iii) Cutoff value for T-statistic ($\alpha$ = significance level)

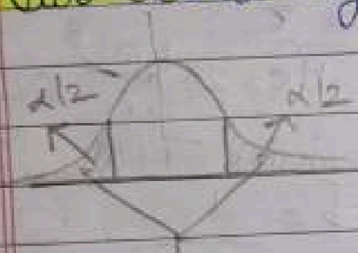
iv) Check whether T-stat falls in the rejection region

Then conclusion → accept or reject $H_0$

•) Types of $H_0$

$\mu = \dots$

Two tailed Hypothesis test




Rejection Region

$\mu \geq \dots$ → Rejection region on LHS

$\mu \leq \dots$ → Rejection region on RHS

One Tailed Hypothesis tests

Null hypothesis cannot have $\{<, >, \neq\}$

Type I errors: False +ves (Rejecting $H_0$ when it is true)

$\alpha$ = Probablity of type I error

Type II errors: False -ve (Not rejecting $H_0$ when its false)

$\beta$ = probablity of type II error

Other typical types of hypothesis tests

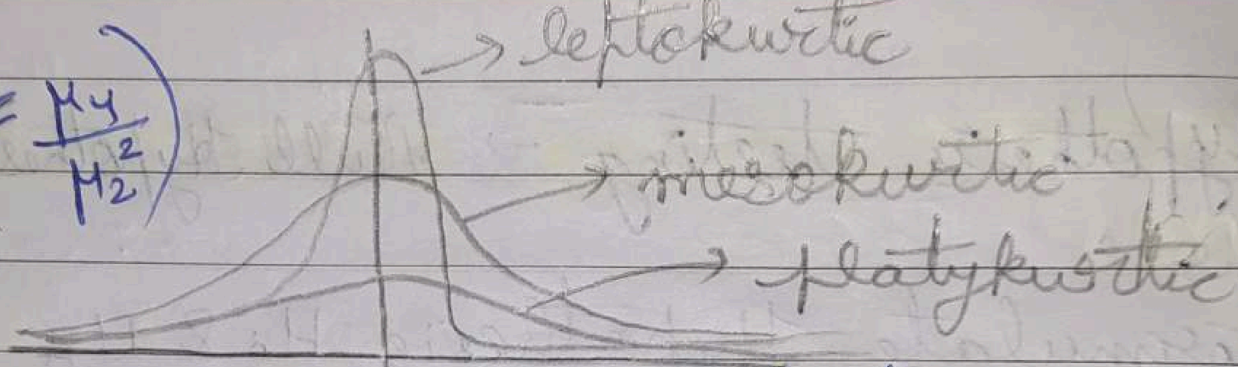
Difference in means test

↳ 2 versions → with or without equal variance assumption

Paired T test for diff in means

Diff in means tests → only for 2 populations not multiple

Kurtosis $\left(= \frac{\mu_4}{\mu_2^2}\right)$




Mesokurtic → Normal distribution → Kurtosis = 3

$\mu_4$ = 4th central moment