STATISTICS

1. Population → Population mean → Median, Mean
2. Sample → Sample mean
3. Random variable → ~~Differential Random variable~~
   - Discrete
   - Continuous

Population

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

Sample

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

Random Variable

- Discrete ⇒ whole no., not a floating number
  Eg: No of bank a/c a person has,
  Population of a state
- Continuous ⇒ Within a range of values, we can have any value.
  whole no., floating nos.
  E.g: Height of a person

Gaussian Distribution or Normal Distribution

$$X \sim G.D(\mu, \sigma)$$

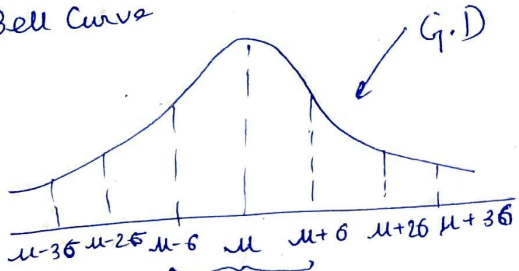
(μ: mean, σ: s.d)

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

$$Var = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$s.d = \sigma = \sqrt{Var}$$

Bell Curve



Eg: Height of a population

Empirical formula in Gaussian Distribution

$$P(\mu - \sigma \le x \le \mu + \sigma) \approx 68\%$$

$$P(\mu - 2\sigma \le x \le \mu + 2\sigma) \approx 95\%$$

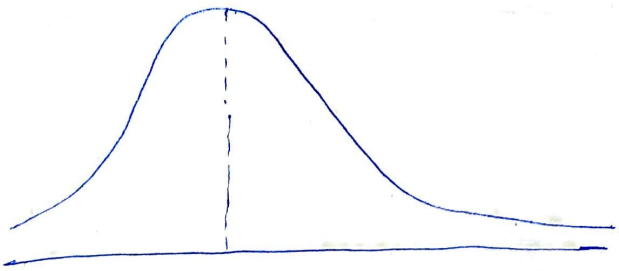
$$P(\mu - 3\sigma \le x \le \mu + 3\sigma) \approx 99.7\%$$

Log Normal Distribution

$X \sim$ Log normal distribution

if $\ln(x)$ is normally distributed

i.e $\ln(X) \sim G.D(\mu, \sigma)$



Eg: Income of the people
Product reviews

* G.D → Standard Normal Distribution (SND)
$(\mu = 0, \sigma = 1)$

$$= \frac{x_i - \mu}{\sigma}$$

Covariance

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

$x \uparrow \quad y \uparrow \quad = \quad Cov \rightarrow +ve$

$x \uparrow \quad y \downarrow \quad = \quad Cov \rightarrow -ve$

Mean, Median, Mode

Mean $\boxed{\mu = \frac{1}{n}\sum_{i=1}^{n} x_i}$

* Population $C = N$
Sample $C = n$
Population mean $= \mu$
Sample mean $= \bar{x}$

Central Limit Theorem

$X \not\approx GD(\mu, \sigma^2) \Leftrightarrow n \geq 30$

$S_1 \quad x_1, x_2 - - - , x_{30} = \bar{x}_1$

$S_2 \quad - \quad - \quad - \quad = \bar{x}_2$

$\vdots$

$S_{100} \quad - \quad - \quad - \quad = \bar{x}_{100}$

$\boxed{\bar{x} \approx GD(\mu, \frac{\sigma^2}{n})}$ $\boxed{\bar{x} \approx \mu}$

Chebyshev's Inequality

$x \approx G.D(\mu, \sigma)$

$P(\mu - \sigma \leq x \leq \mu + \sigma) \approx 68\%$

$P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 95\%$

$P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 99.7\%$

$y \not\approx G.D$

$\Downarrow$

$P(\mu - K\sigma < x < \mu + K\sigma) \geq 1 - \frac{1}{K^2}$

$K = 2$

$P(\mu - 2\sigma < x < \mu + 2\sigma) > 1 - \frac{1}{2^2} = \frac{3}{4} = 75\%$

Pearson Correlation Coefficient

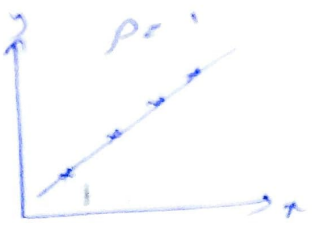
$Cov(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x) * (y_i - \mu_y)$

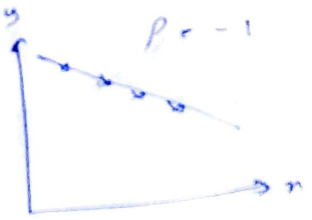
⟶ Gives the direction of relationship (x↑-y↑ or x↑ y↓)

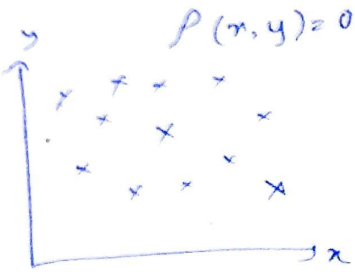
Pearson CC

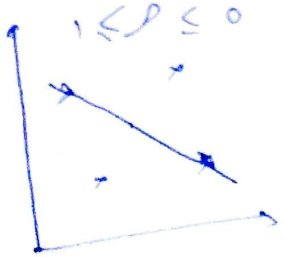
$$\rho(x, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y}$$

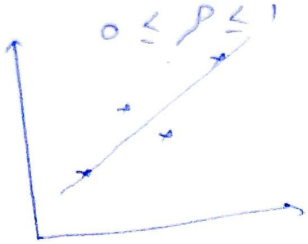
$$\rho(x, y) = -1 \leq \rho \leq 1$$











Spearman's rank correlation

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

~~$r = 1 - \frac{6 \Sigma d_i^2}{n}$~~

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

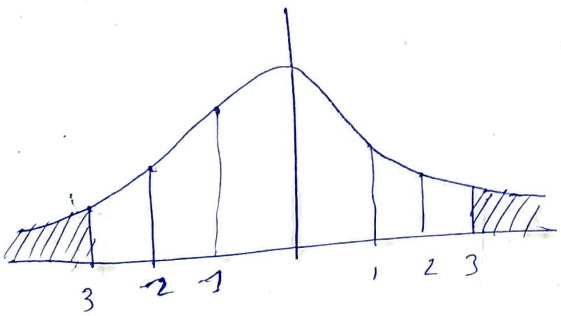
Z-score

SND $\Rightarrow \mu = 0$ & $\sigma = 1$

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

$Z < 3 \rightarrow$ Normal

$Z > 3 \rightarrow$ ~~Outlie~~ Outliers

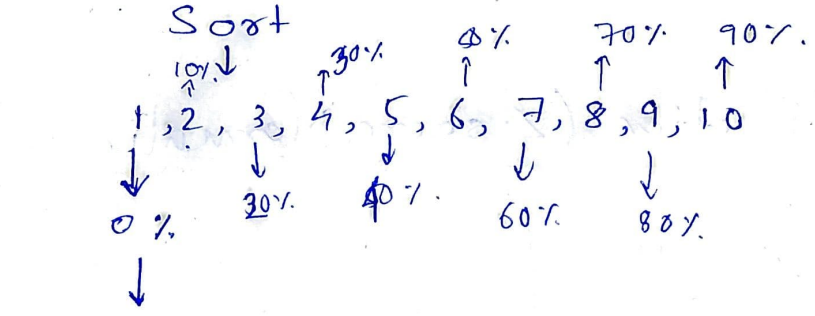


~~Outli~~ Outliers

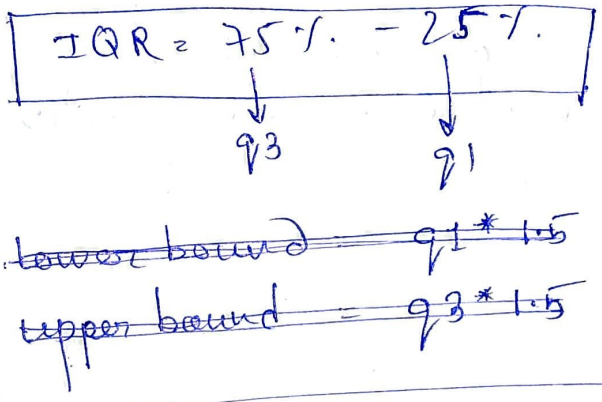
Inter Quantile Range (IQR)

* Percentiles

Let a set of numbers be 4, 9, 5 6, 7, 1, 2, 8, 10



0 percentage of nos. lesser than 1



~~lower bound = q1 * 1.5~~

~~upper bound = q3 * 1.5~~

Lower bound value = $q1 - (1.5 * IQR)$

Upper bound value = $q3 + (1.5 * IQR)$

Any values away from lower bound & upper bound values ~~is~~ are outliers.

Normalization (min-max Normalization)

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

* Scale down your values between 0 and 1.

Standardization (Z-score Normalization)

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

* Here all the features will be transformed in such a way that it will have the properties of a standard normal distribution with

$\mu = 0$ & $\sigma = 1$