Correlation Coefficients

Monday, 22 November 2021 7:59 PM

[inputs → multiple features   output → Loan-Status]

income, able-to-pay-emi,

0 1

↑ ↑

X : height of people [172, 180, 161 - - -]

Y : weight of people [75, 68, 80, - -]

Any (Relationship)?

→ if x increases y?

if X ↑ Y ↑ }

if X ↑ Y ↓ }

if X ↑ — } no effect

1. Covariance
2. Pearson Correlation Coeff
3. Spearman Rank Correlation Coeff

Covariance

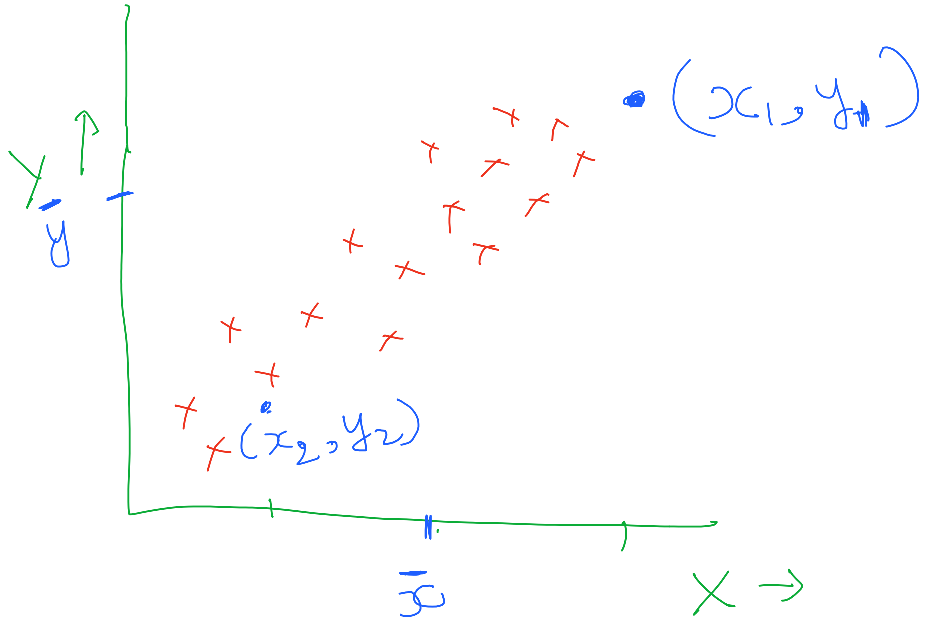
Cov (X, Y) → +ve

as X ↑, y ↑

Cov (X, Y) → -ve

as X ↑, Y ↓

Cov(X, Y) → 0 as X↑, Y (No effect)



$$\boxed{Cov(x,y) = \frac{\sum_{i=1}^{n} \underbrace{(x-\bar{x})}\cdot\underbrace{(y-\bar{y})}}{n}}$$

n → no. of points

$\bar{x}$ → mean(x)

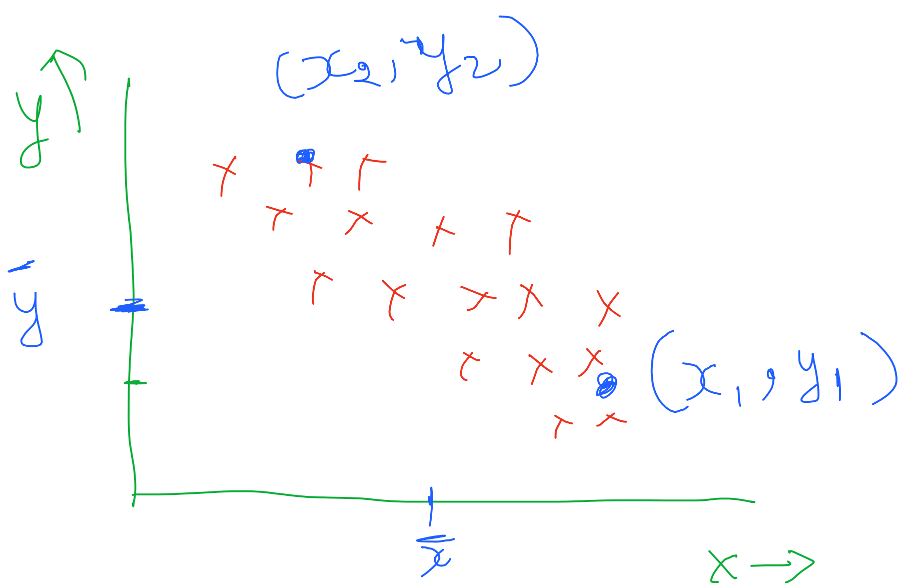
$\bar{y}$ → mean(y)

Case I : $(x_1 - \bar{x}) \cdot (y_1 - \bar{y}) \Rightarrow$ overall +ve

$(x_1 - \bar{x})$: +ve, $(y_1 - \bar{y})$: +ve

Case II : $(x_2 - \bar{x}) \cdot (y_2 - \bar{y}) \Rightarrow$ overall +ve

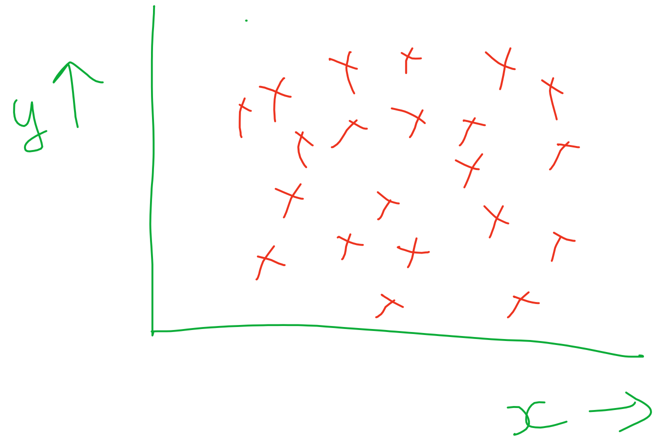
$(x_2 - \bar{x})$: -ve, $(y_2 - \bar{y})$: -ve



Cov(x, y) = -ve

$(x_1-\bar{x}) \cdot (y_1-\bar{y}) \Rightarrow$ overall $-ve$

$+ve$ $\quad\quad$ $-ve$



Covariance

↓

Variance

$$Cov(x,\boxed{y}) = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{n}$$

$$Cov(x,x) = \frac{\sum_{i=1}^{n}(x-\bar{x})\cdot(x-\bar{x})}{n} \Rightarrow \boxed{\frac{1}{n}\sum_{i=1}^{n}(x-x)^2}$$

Variance(x)

$\Downarrow$

Variance(x)

Disadv.

→ No boundation

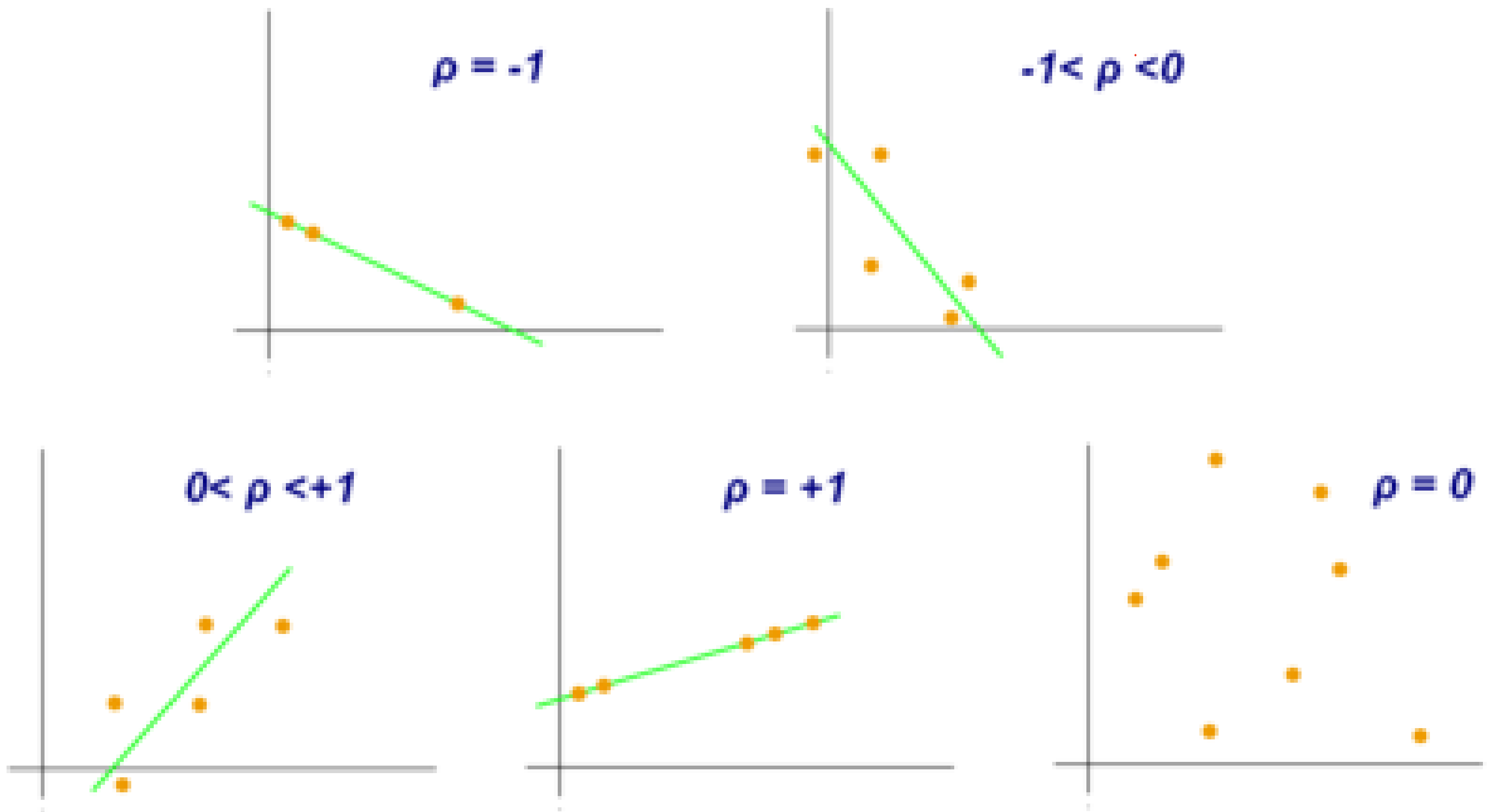
PCC [Pearson correlation Coefficient]

it tells $\boxed{\text{how strongly}}$ is the relationship

→ [-1, 1]

PCC

$$\boxed{\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}}$$



x_1 , y_1 $Corr(x_1, y_1) \Rightarrow \boxed{0.66} \rightarrow$

x_2 , y_2 Corr$(x_2, y_1) \Rightarrow$ -0.+5 $\boxed{-1}$

→ x_2 has ~~most~~ high Corr with y_1 than x_1

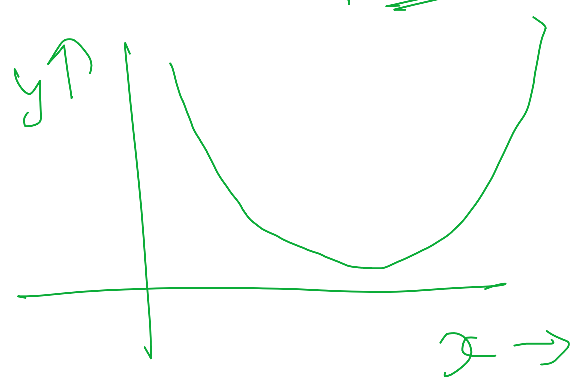
height (in cm) height (in meter)

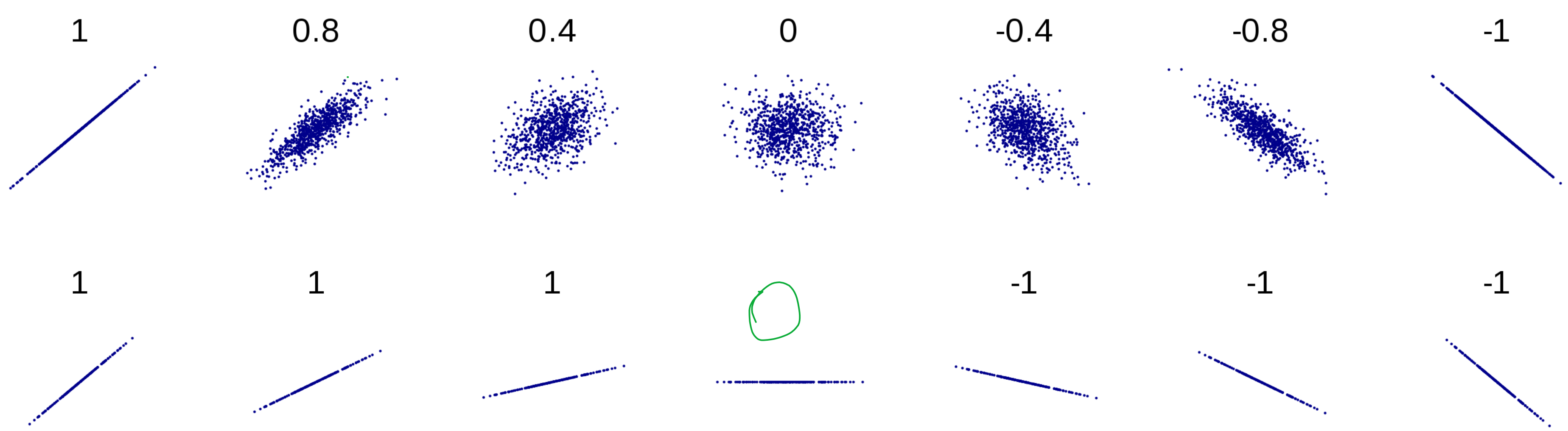
Disadv. of PCC

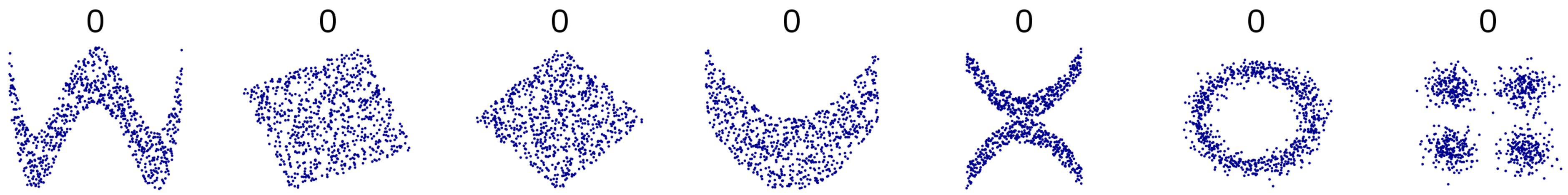
↳ biased towards "linear" relationship. $f(x) = x^2$

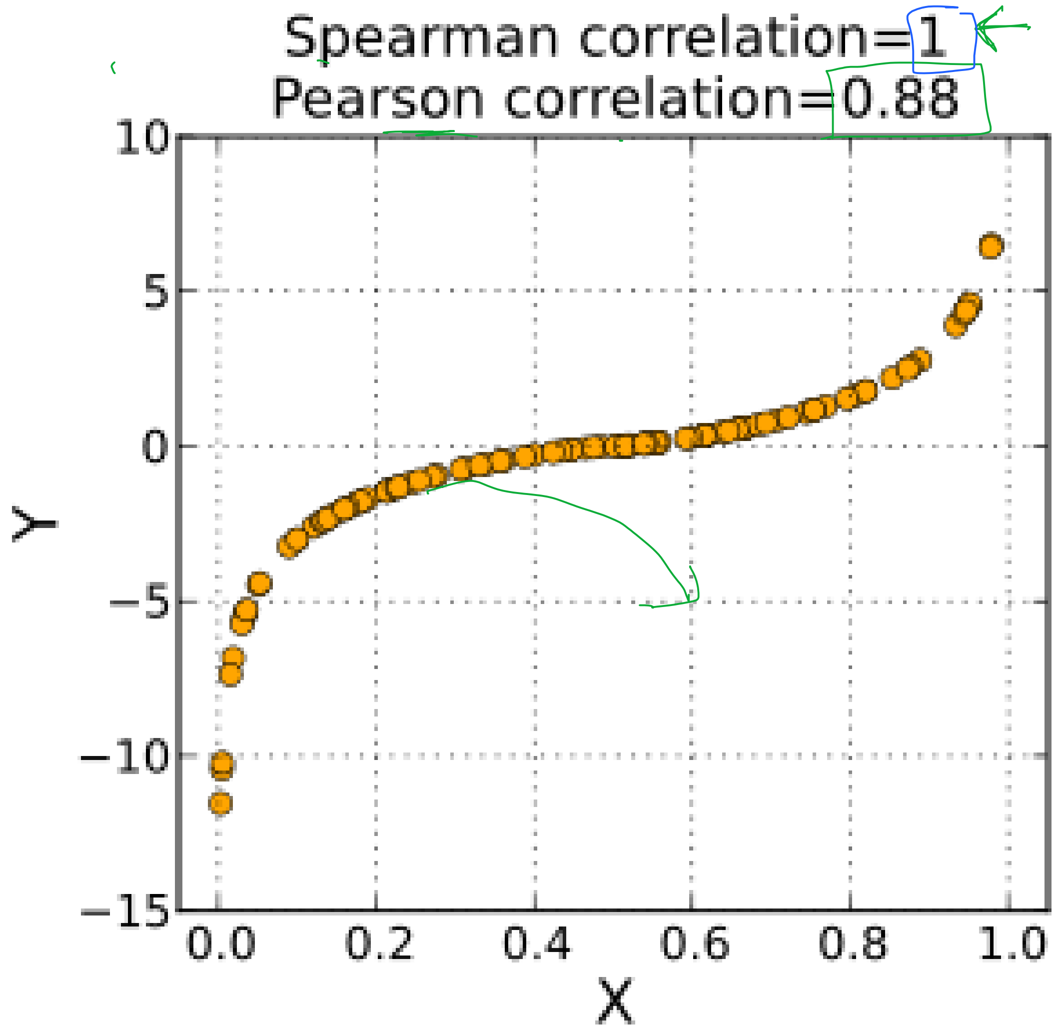
$$y_1 = m x + c$$

$$y_1 = m \cdot x_1^2$$









[Monotonic increasing]

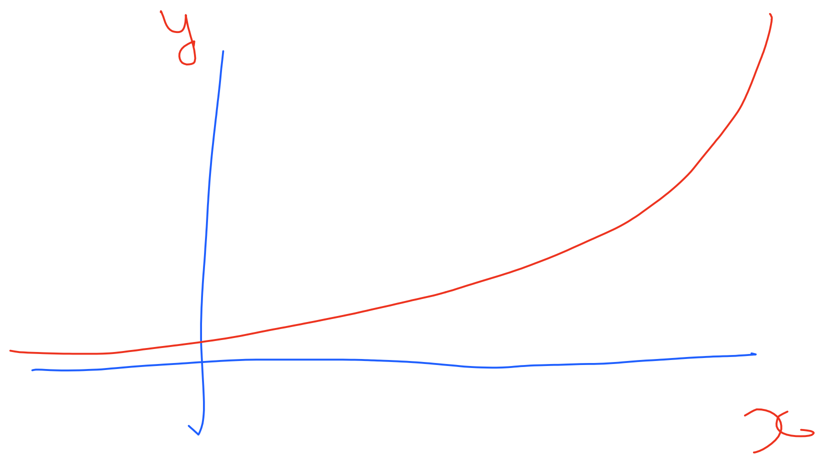
X	rank(x)	Y	rank(y)
2.5	3	1.8	1
1.3	1	2.1	2
1.4	2	5.2	3
7.2	5	7.9	5

5.1 4 6.3 4

Spearman rank corr coeff = Corr(rank(x), rank(y))

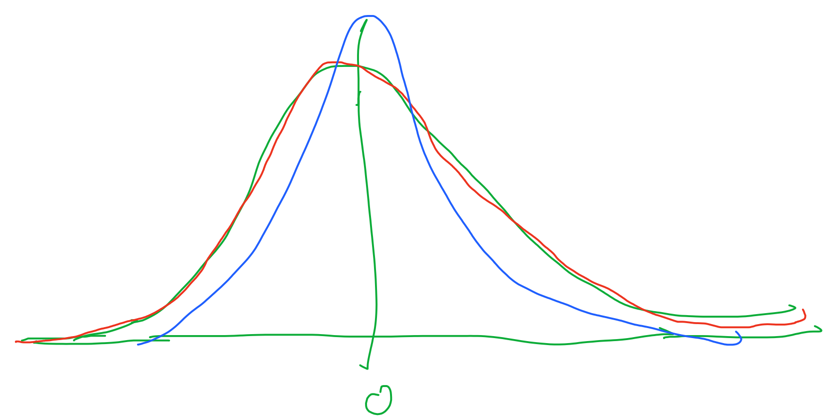
SRCC

$\boxed{y = e^{x}}$



$X = [x_1, x_2, x_3 \; ---- \; x_{500}]$

Imp. $\boxed{X}$ ~ Normal ?? or Not?



$Z = \frac{X - \bar{X}}{\sigma}$

$\mu = 0, \sigma = 1$

$X \sim N(0, 1)$

$Y \sim N(0,1)$

Q-Q Plot

$X \rightarrow x_1, x_2, x_3 - - - - , x_{5000}$

$\downarrow$ Sort

$x'_1, x'_2, x'_3 - - - - - x'_{5000}$

$\downarrow$ Percentile

Observed data = $x^{(1)}, x^{(2)}, - - - - - , x^{(100)}$

Ist percentile IInd

generated

$Y = N(0,1)$

$y_1, y_2, y_3 - - - - y_{5000}$

$\downarrow$ Sort

$y'_1, y'_2, y'_3 - - - - y'_{5000}$

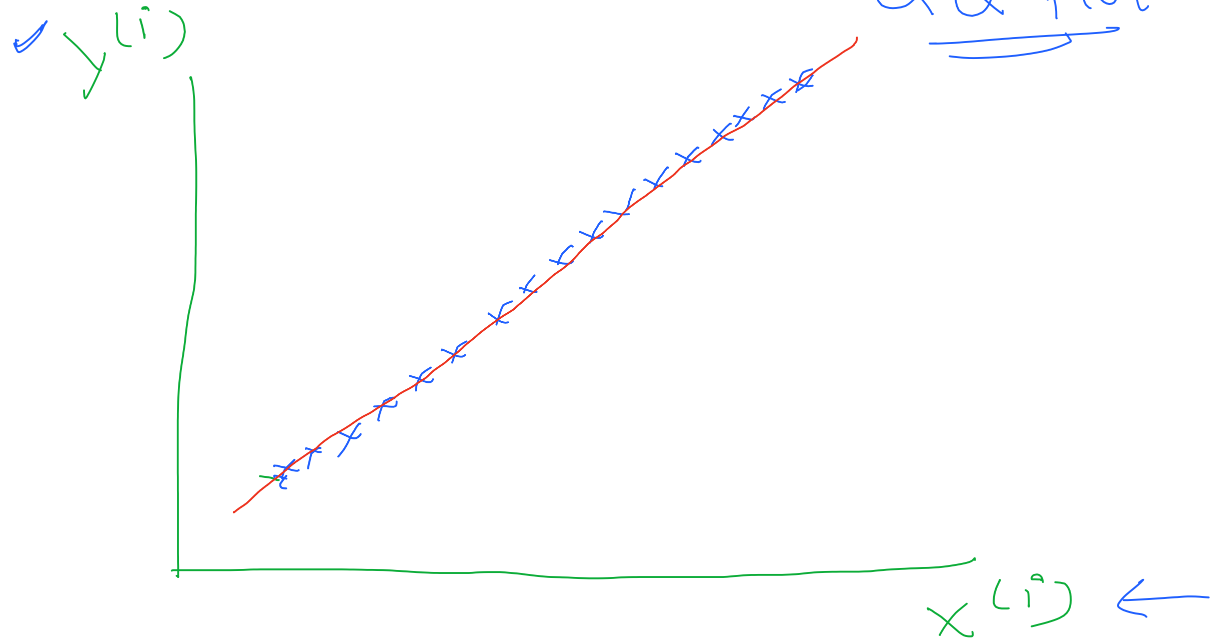
$\downarrow$ Percentile

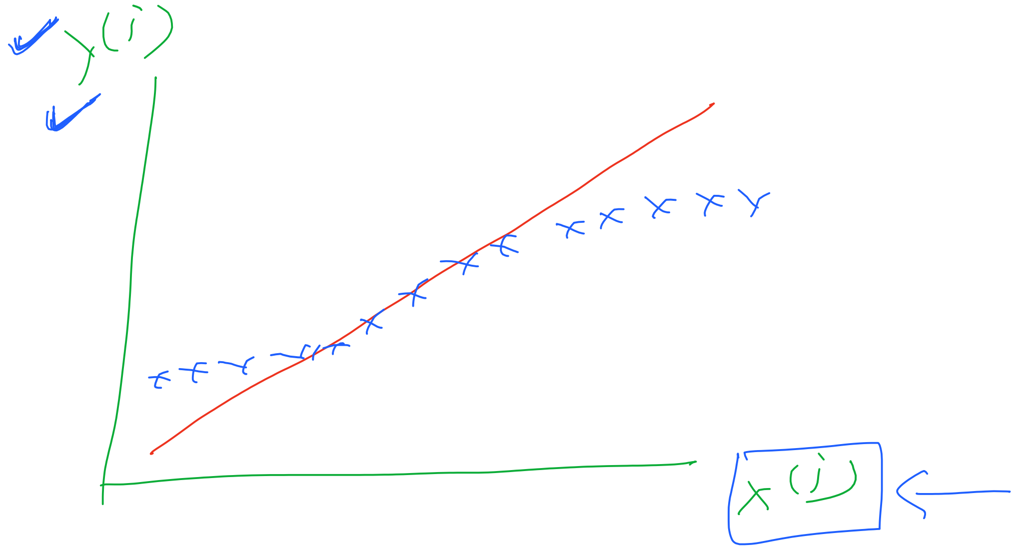
$y^{(1)}, y^{(2)}, - - - - y^{(100)}$

observed data $\qquad$ Theoretical Quantile

$$\begin{bmatrix} x^{(1)}, & y^{(1)} \\ x^{(2)}, & y^{(2)} \\ x^{(3)}, & y^{(3)} \\ \vdots & \\ x^{(100)}, & y^{(100)} \end{bmatrix}$$

QQ Plot





Power Law
Poisson Distr