Ch.2 Data Preprocessing.

(1) Chi-Squared Statistics.

	Observed (O)	Expected (E)	Total
Male	80	60	140
female	40	60	100
	120	120.	240

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{(80-60)^2}{60} + \frac{(40-60)^2}{60} = 6.66 + 6.66 = 13.33$$

(2) Equi-depth partitioning.

- data [0, 2, 8, 10, 20, 21, 29, 29, 30]
- divide it to k intervals of same size
- divide into 3 intervals of same size = k = 3.

$$\text{width} = \frac{\text{max} - \text{min}}{k}$$

$$= \frac{30-0}{3} = \frac{30}{3} = 10$$

So bins are (0, 10), (11, 20), (20, 30).

bin 1 → 0, 2, 8, 10

bin 2 → 20

bin 3 → 21, 29, 29, 30.

(3) Equi-depth partitioning.

- data [0, 2, 8, 10, 20, 21, 29, 29, 3
- number of values present in each bin are same
- number of values in each bin k = 3.

bin 1 → 0, 2, 8

bin 2 → 10, 20, 21

bin 3 → 29, 29, 30.

(4) Equi-depth (by means).

- replace each value in bin with mean of that bin

bin1 → 0, 2, 8, mean = 3.33 bin1 → (3.33, 3.33, 3.33)

bin2 → 10, 20, 21, mean = 17 bin2 → (17, 17, 17)

bin3 → 29, 29, 30, mean = 29.33 bin3 → (29.33, 29.33, 29.33)

(5) Equi-depth (by median) – Median is middle value of given range

bin 1 → 0, 2, 8 median = 2 bin 1 → (2, 2, 2)
bin 2 → 10, 20, 21 median = 20 bin 2 → (20, 20, 20)
bin 3 → 29, 29, 30 median = 29 bin 3 → (29, 29, 29)

(6) Equi-depth (by boundaries)
Replace each value in bean with nearest boundary value.

bin 1 → 0, 2, 8 ⟹ bin 1 → (0, 0, 8)
bin 2 → 10, 20, 21 bin 2 → (10, 21, 21)
bin 3 → 29, 29, 30 bin 3 → (29, 29, 30)

(7) Co-relation analysis –
Check whether sale of ice-creams & sunglasses are related?

ice cream sale (A)	sun-glass sale (B)
20	30
10	5
23	29
5	10

co-relation const

$$\gamma_{A,B} = \frac{\Sigma (A-\bar{A})(B-\bar{B})}{(n-1)\,\sigma_A \cdot \sigma_B}$$

$\bar{A}$ = mean of A $\quad \bar{A} = \frac{\Sigma A}{n}$

σ_A = S.D. of A.

$$\sigma_A = \sqrt{\frac{\Sigma (A-\bar{A})^2}{n-1}}$$

→

$n = 4 \quad \bar{A} = \frac{58}{4} = 14.5 \quad \bar{B} = \frac{24}{4} = 18.5$

A	B	$A-\bar{A}$	$(A-\bar{A})^2$	$B-\bar{B}$	$(B-\bar{B})^2$	$(A-\bar{A})(B-\bar{B})$
20	30	5.5	30.25	11.5	132.25	63.25
10	5	-4.5	20.25	-13.5	182.25	60.75
23	29	8.5	72.25	10.5	110.25	89.25
5	10	-9.5	90.25	-8.5	72.25	170.

$$\sigma_A = \sqrt{\frac{\Sigma (A-\bar{A})^2}{n-1}} = \sqrt{\frac{30.25 + 20.25 + 72.25 + 90.25}{3}} = \sqrt{\frac{213}{3}} = 8.426.$$

$$\sigma_B = \sqrt{\frac{\Sigma (B-\bar{B})^2}{n-1}} = \sqrt{\frac{132.25 + 182.25 + 110.25 + 72.25}{3}} = \sqrt{\frac{497}{3}} = 12.87$$

$$\gamma_{A,B} = \frac{\Sigma (A-\bar{A})(B-\bar{B})}{(n-1)\,\sigma_A \cdot \sigma_B} = \frac{63.25 + 60.75 + 89.25 + 170}{3 \times 8.426 \times 12.87} = 1.178$$

As $\gamma_{A,B} > 1$, sale of ice cream & sun glasses is positively related

(8) Min-max normalization $v' = \left\{ \frac{v - min_A}{max_A - min_A} \left[new\_max_A - new\_min_A \right] \right\} + new\_min_A$

income → max = 98000, min = 12000.

map income of 73600 to the range [0.0, 1.0]

→ $v' = \frac{v - min_A}{max_A - min_A} (new\_max_A - new\_min_A) + new\_min_A$

$= \frac{73600 - 12000}{98000 - 12000} (1.0 - 0.0) + 0.0 = 0.716.$

(9) z-score normalization (zero normalization)

$v' = \frac{v - \bar{A}}{\sigma_A}$ $\quad \bar{A} = mean$, $\sigma_A = S.D.$

mean = 54000, S.D. = 16000, map 73600.

$v' = \frac{73600 - 54000}{16000} = 1.225.$

(10) Stratified sampling:-

Group of 12 sales price records are sorted.

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215.

Use sample of size 6 & strata low, medium, high.

Use equi-width binning.

→ k = 3, min = 5, max = 215, 6 samples needs to be drawn from 3 bins.

width = $\frac{215 - 5}{3} = \frac{210}{3} = 70$, intervals [5, 75], [75, 145], [145, 215].

bin 1 → 5, 10, 11, 13, 15, 35, 50, 55

bin 2 → 72, 92

bin 3 → 204, 215

By monitoring value distribution in bin 1.

price 10/11 and 35/50 can be drawn as sample.

from bin 2 both records can be drawn as sample

from bin 3 both records can be drawn as sample

(11) Covariance — $Cov(A,B) = E(A \cdot B) - \bar{A}\,\bar{B}$.

Suppose two stocks A & B have following values in one week:
(2,5), (3,8), (5,10), (4,11), (6,14)
Check if the stocks are affected by same industry trends, will their prices rise or fall together?

↔

$$E(A) = \frac{2+3+5+4+6}{5} = \frac{20}{5} = 4$$

$$E(B) = \frac{5+8+10+11+14}{5} = \frac{48}{5} = 9.6$$

$$E(AB) = \frac{2\times5 + 3\times8 + 5\times10 + 4\times11 + 6\times14}{5} = \frac{212}{5} = 42.4$$

$$Cov(A,B) = 42.4 - 4\times9.6 = 4.$$

∴ Thus, A & B rise together since $Cov(A,B) > 0$.

(12) Dispersion of data —

following data points represent number of classes each teacher at BITS teaches. 5, 7, 5, 1, 1, 9, 4, 5, 3, 4, 6.
find out measures of dispersion of data.

↔ first arrange in ascending order

1, 1, 3, 4, 4, 5, 5, 5, 6, 7, 9

Median = middle value if number of data points are odd
= 5

Q_1 (25th percentile) = middle value of data points left side of the median
1, 1, 3, 4, 4
= 3

Q_3 (75th percentile) = middle value of data points right side of the median
5, 5, 6, 7, 9
= 6

IQR (Inter quartile range) = $Q_3 - Q_1 = 6 - 3 = 3$.

five number summary = min, Q_1, median, Q_3, max
= {1, 3, 5, 6, 9}.

~~Outlier = a value higher/lower than 1.5 × IQR~~

~~1.5 × IQR = 1.5 × 3 = 4.5,~~ $1.5 \times IQR = 1.5 \times 3 = 4.5$

Outlier = value which is located 1.5 × IQR or more below Q_1 or
value which is located 1.5 × IQR or more above Q_3.

All data values which are below $Q_1 - 4.5$ or are above $Q_3 + 4.5$ are outliers. ∴ No outlier is above dataset.

* The Type of Data
- Type of attribute - Qualitative, Quantitative
  Nominal, Ordinal, Interval, Ratio.
  Discrete, Continuous.
  -Asymetric, Symmetric.
- Type of Datasets -
  Record. data - Market Basket, Data matrix, Sparse Data matrix
  Graph based data - Data with rel° among objects
  Data with objects that are graph.
  Ordered data - Sequential data, sequence data,
  Time series, data, Spatial data.
  Characteristics of Data sets -
  - Dimensionality, Sparsity, Resolution.

* Data Quality
- Measurement & data collection issues.
- Noise & Artifact
- Precision, Bias, Accuracy.
- Outliers.
- Missing values
- Inconsistent values
- Duplicate data.

* Data Preprocessing.
- Data cleaning.
  Missing values - ignore tuple, use mean/mode, estimate using models
  Noisy data - data smoothing - binning, clustering,
  combined human & computer inspection.
  regression
  Inconsistent data - manual correction
  knowledge engineering tools for constraints
- Data Integration
  entity identification problem cust-id & cust-no
  redundancy - correlation.
  detection & resolution of data value conflicts

correlation between attribute A & B.

$$\sigma_{A,B} = \frac{\Sigma (A-\bar{A})(B-\bar{B})}{(n-1)\,\sigma_A\,\sigma_B}$$

$\bar{A}$ & $\bar{B}$ are resp mean of A & B $\quad \bar{A} = \frac{\Sigma A}{n}$

σ_A & σ_B are s.d. of A & B. $\quad \sigma_A = \sqrt{\frac{(A-\bar{A})^2}{n-1}}$

- Data Transformation
  - Smoothing
  - Aggregation
  - Generalization - concept hierarchy creation
  - Normalization - min-max, z-score, decimal scaling
  - Attribute (feature) construction
- Data Reduction
  - Data cube aggregation
  - Dimension reduction - stepwise fwd selection, backward elimination, decision tree,
  - Data compression - wavelet transform, PCA, factor analysis,
  - Numerosity reduction - regression, histogram, clustering, sampling
  - Concept Hierarchy generation
- ** Data Exploration - EDA
  - Summary stat
  - Data similarity & dissimilarity measures
  - Data visualization.
- Summary Stat
  - frequency & mode
  - Percentile
  - Measures of location - mean & median central points
  - Measures of spread - range & variation
  - Multivariate summary stat - mean, covariance, correlation
- Data similarity & dissimilarity measures.
  - Similarity, Dissimilarity, Proximity.
  - similarity, Dissimilarity bet^n simple attributes

	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases}$
ordinal	$d = \frac{\lvert x-y \rvert}{n-1}$	$s = 1 - d$
Interval or Ratio	$d = \lvert x-y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$

- Dissimilarity between data objects
  - Distances - $d(x,y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$ Euclidian
  - Minkowski - $d(x,y) = \sqrt[r]{\sum_{k=1}^{n} (x_k - y_k)^r}$
  - Manhatten - $\underset{r=1}{d(x,y)} = \sum_{k=1}^{n} |x_k - y_k|$.
  - Supremum - $\underset{r=\infty}{d(x,y)} = \lim_{r \to \infty} \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{\frac{1}{r}}$

★ Examples of Proximity Measures

Binary data. — similarity coefficients.

★ Simple Matching Coeff SMC = $\frac{\text{number of matching attribute values}}{\text{no. of attributes}}$

$$SMC = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}}$$

- Jaccard coeff. $J = \frac{\text{number of matching presences}}{\text{number of attributes not involved in } f_{00}}$

Asymetric attr.

$$= \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

- Cosine Similarity - term document matrix.

$$\cos(x,y) = \frac{x \cdot y}{\|x\| \|y\|} \qquad x \cdot y = \sum_{k=1}^{n} x_k y_k \qquad \|x\| = \sqrt{\sum_{k=1}^{n} x_k^2}$$

- Extended Jacob Coeff
  Tanimato Coeff $EJ(x,y) = \frac{x \cdot y}{\|x\|^2 + \|y\|^2 - x \cdot y}$

term documents matrix with binary attributes.

- Correlation -1 to 1

Pearsons correlation coeff.

$$corr(x,y) = \frac{\text{covariance}(x,y)}{s.d.(x) * s.d.(y)} = \frac{s_{xy}}{s_x \cdot s_y}$$

$$\text{covariance}(x,y) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})$$

$$s.d.(x) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2}$$

$$s.d.(y) = s_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \bar{y})^2}$$

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k \qquad \bar{y} = \frac{1}{n} \sum_{k=1}^{n} y_k.$$

★ Similarity for heterogeneous attributes -

① for attribute k, compute similarity $s_k(x,y)$ in range $[0,1]$

② Define indicater variable, δ_k for k^{th} attribute

$$\delta_k = \begin{cases} 0 & \text{if } k^{th} \text{ attribute is asymetric \& both objects have a value of 0 or if one of objects has missing value} \\ 1 & \text{otherwise.} \end{cases}$$

③ Compute overall similarity betn two data objects.

$$\text{similarity}(x,y) = \frac{\sum_{k=1}^{n} \delta_k \cdot s_k(x,y)}{\sum_{k=1}^{n} \delta_k}$$

If weights are used.

$$\text{similarity}(x,y) = \frac{\sum_{k=1}^{n} \delta_k \cdot w_k \cdot s_k(x,y)}{\sum_{k=1}^{n} \delta_k}$$

* Precision & Bias.

Precision – closeness of repeated measurements of same quantity to one another

Bias – systematic variation of measurements from the quantity being measured.

A device is being used to measure the speed of the vehicles. On trial, vehicles speed is 10 km/hr. five measurements are taken for the same vehicles which are {10.1, 10.3, 10.0, 10.2, 10.1} km/hr. find precision & bias for the device being used for measurement.

→ $x = \{10.1; 10.3, 10.0, 10.2, 10.1\}$

$\text{mean } \bar{x} = \frac{10.1 + 10.3 + 10 + 10.2 + 10.1}{5} = \frac{50.7}{5} = 10.14 \text{ km/hr.}$

n = no. of trials = 5.

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{4}\left[(10.1-10.14)^2 + (10.3-10.14)^2 + (10-10.14)^2 + (10.2-10.14)^2 + (10.1-10.14)^2\right]}$$

$$= \sqrt{\frac{1}{4} \times 0.052} = 0.1140$$

Precision = s.d. = 0.1140.

Bias = difference bet^n mean of values & known value of quantity measured

$= 10.14 - 10 = 0.14$

* Similarity & Dissimilarity between Simple Attributes.

Attribute Type.	Dissimilarity	Similarity.
Nominal.	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases}$
Ordinal	$d = \frac{\lvert x - y \rvert}{n-1}$	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$ $s = \frac{1}{1+d}$ $s = e^{-d}$

Consider following dataset about students studying DM subject at BTIS find out dissimilarity & similarity for each attribute in it.

Id.	Name.	Grade_Midsem	Grade_Compre	Grade_Assignment
1	Pravin	10	8	Fair
2	Palwar	8	6	Excellent
3	Pravin	10	8	Good.

→ Name is nominal attribute, & categorical,
Grade_Midsem & Grade_Compre are intervals & numeric
Grade_Assignment is categorical & ordinal.

for Name — $d(1,2) = 1$. as Pravin ≠ Palwar $s(1,2) = 0$
$d(1,3) = 0$ as Pravin = Pravin $s(1,3) = 1$
$d(2,3) = 1$. as Palwar ≠ Pravin $s(2,3) = 0$

for Grade_Midsem.
$d(1,2) = 2$
$d(1,3) = 0$
$d(2,3) = 2$

$s = -d$ then $s(1,2) = -2$, $s(1,3) = 0$, $s(2,3) = -2$

$s = \frac{1}{1+d}$
$s(1,2) = \frac{1}{1+2} = \frac{1}{3} = 0.33$
$s(1,3) = \frac{1}{1+0} = 1 = 1$
$s(2,3) = \frac{1}{1+2} = \frac{1}{3} = 0.33$

for Grade_Compre.
$d(1,2) = 2$
$d(1,3) = 0$
$d(2,3) = 2$

$s = \frac{1}{1+d}$
$s(1,2) = \frac{1}{1+2} = \frac{1}{3} = 0.33$
$s(1,3) = \frac{1}{1+0} = 1 = 1$
$s(2,3) = \frac{1}{1+2} = \frac{1}{3} = 0.33$.

for Grade_Assignment
Lets assume order as. Fair < Good < Excellent
1 2 3

$d(1,2) = \frac{|1-3|}{3-1} = \frac{2}{2} = 1$
$d(1,3) = \frac{|1-2|}{3-1} = \frac{1}{2} = 0.5$
$d(2,3) = \frac{|3-2|}{3-1} = \frac{1}{2} = 0.5$

$s(1,2) = 1-d = 1-1 = 0$
$s(1,3) = 1-0.5 = 0.5$
$s(2,3) = 1-0.5 = 0.5$

Similarity Measures

Distance between Vectors

$x = \{x_1, x_2, x_3, \ldots x_n\}$

$y = \{y_1, y_2, y_3 \ldots y_n\}$.

(1) Minkowski dist^n.

$$d(x,y) = \left\{ \sum_{k=1}^{n} |x_k - y_k|^r \right\}^{1/r}.$$

(2) Hamming dist^n (Manhattan, L1) $\longrightarrow$ $r = 1$.

$$\therefore d(x,y) = \sum_{k=1}^{n} |x_k - y_k|.$$

(3) Euclidian dist^n (L2) $\longrightarrow$ $r = 2$

$$\therefore d(x,y) = \sqrt{\sum_{k=1}^{n} |x_k - y_k|^2}.$$

(4) Supremum dist^n

$$d(x,y) = \lim_{r \to \infty} \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}.$$

eg. $x = \{1, 2, 3, 4\}$

$y = \{5, 3, 2, 1\}$

Hamming dist^n $\longrightarrow$ $r = 1$.

$\therefore d(X,Y) = |1-5| + |2-3| + |3-2| + |4-1|$

$= 4 + 1 + 1 + 3 = 9$

Euclidian dist^n $\longrightarrow$ $r = 2$

$\therefore d(X,Y) = \sqrt{|1-5|^2 + |2-3|^2 + |3-2|^2 + |4-1|^2}$

$= \sqrt{4^2 + 1^2 + 1^2 + 3^2} = \sqrt{27}$.

Proximity Measures.

(1) Simple Matching Coefficient (SMC)

x & y are two objects that consists of n binary attributes.

f_{00} = number of cases where x is 0 & y is 0

f_{01} = —— " —— x is 0 & y is 1

f_{10} = —— " —— x is 1 & y is 0

f_{11} = —— " —— x is 1 & y is 1

$$SMC = \frac{\text{number of matching attribute values}}{\text{number of attributes}}$$

$$= \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}.$$

$x = \{1, 0, 1, 0\}$ $\quad f_{00} = 1$ $\quad f_{01} = 1$

$y = \{0, 1, 1, 0\}$. $\quad f_{10} = 1$ $\quad f_{00} = 1$

$$SMC = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}} = \frac{1+1}{4} = \frac{2}{4} = 0.5$$

(2) Jaccard coefficient

$$J = \frac{\text{number of matching presences}}{\text{number of attributes not involved in 00 matches}}$$

$$= \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

$x = \{1, 0, 1, 0\}$
$y = \{0, 1, 1, 0\}$

$f_{00} = 1$ $f_{10} = 1$
$f_{01} = 1$ $f_{11} = 1$

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} = \frac{1}{1+1+1} = \frac{1}{3} = 0.33.$$

(3) Cosine similarity -

$$\cos(x,y) = \frac{x \cdot y}{\|x\| \|y\|}$$ where $\cdot$ indicates dot vector product.

$$x \cdot y = \sum_{k=1}^{n} x_k y_k.$$

$$\|x\| = \sqrt{\sum_{k=1}^{n} x_k^2}.$$

$x = \{3, 2, 0, 5\}$
$y = \{1, 0, 0, 0\}$.

$$x \cdot y = 3*1 + 2*0 + 0*0 + 5*0 = 3+0+0+0 = 3$$

$$\|x\| = \sqrt{3^2+2^2+0^2+5^2} = \sqrt{9+4+25} = \sqrt{38}$$

$$\|y\| = \sqrt{1^2+0^2+0^2+0^2} = \sqrt{1} = 1$$

$$\therefore \cos(x,y) = \frac{x \cdot y}{\|x\| \|y\|} = \frac{3}{\sqrt{38} \cdot 1} = \frac{3}{\sqrt{38}} = 0.486$$

(4) Extended Jaccard Coefficient (Tanimoto coefficient)

$$EJ(x,y) = \frac{x \cdot y}{\|x\|^2 + \|y\|^2 - x \cdot y}$$

$x = \{3, 2, 0, 5\}$
$y = \{1, 0, 0, 0\}$.

$$EJ(x,y) = \frac{3}{(\sqrt{38})^2 + (1)^2 - 3} = \frac{3}{\sqrt{38} - 2} = \frac{3}{36}$$

(5) Co-relation

$$\text{corr}(x,y) = \frac{\text{covariance}(x,y)}{\text{S.D.}(x) * \text{S.D.}(y)}.$$

$$\text{covariance}(x,y) = \frac{1}{n-1} \sum_{n=1}^{n} (x_k - \bar{x})(y_k - \bar{y})$$

$$\sigma(x) = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2}$$

$$\sigma(y) = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \bar{y})^2}$$

$$\bar{x} = \sum_{k=1}^{n} x_k$$

$$\bar{y} = \sum_{k=1}^{n} y_k$$

Q. Suppose two stocks A & B have following closing values in one week. (2,5), (3,8), (5,10), (4,11), (6,14)

(a) Check if stocks are affected by same industry trends or not.

(b) Check if one stock can be used to predict value of other stock or not.

→ (a) Co-variance – stocks vary together or not.

If its zero, they dont vary together

If its one, they vary together will be affected by same trend

$$\text{covariance}(x,y) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y})$$

$n = 5$.

$x = \{2,3,5,4,6\}$ $\quad \bar{x} = 4$

$y = \{5,8,10,11,14\}$ $\quad \bar{y} = 9.6$

$$\text{covariance}(x,y) = \frac{1}{4}\Big[(2-4)(5-9.6) + (3-4)(8-9.6) + (5-4)(10-9.6) + (4-4)(11-9.6) + (6-4)(14-9.6)\Big]$$

$$= \frac{1}{4}[9.2 + 1.6 + 0.4 + 0 + 8.8]$$

$$= \frac{20}{4} = 5$$

Its positive value, hence both stocks vary together.

(b) Co-relation – stocks are related or not

$$\text{correlation}(x,y) = \frac{\text{covariance}(x,y)}{\text{s.d.}(x)\ \text{s.d.}(y)}$$

$$\text{s.d.}(x) = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})^2}$$

$$= \sqrt{\frac{1}{4}\left[(2-4)^2 + (3-4)^2 + (5-4)^2 + (4-4)^2 + (6-4)^2\right]}$$

$$= \sqrt{\frac{1}{4}[4+1+1+0+4]} = \sqrt{2.5} = 1.58$$

$$\text{s.d.}(y) = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k-\bar{y})^2}$$

$$= \sqrt{\frac{1}{4}\left[(5-9.6)^2 + (8-9.6)^2 + (10-9.6)^2 + (11-9.6)^2 + (14-9.6)^2\right]}$$

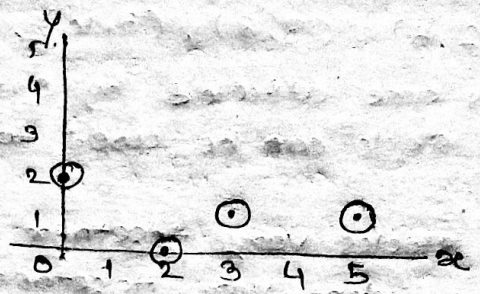
$$= \sqrt{\frac{1}{4}[21.16 + 2.56 + 0.16 + 1.96 + 19.36]} = \sqrt{11.3} = 3.36$$

$$\text{correlation}(x,y) = \frac{5}{1.58 \times 3.36} = 0.94$$

As correlation is 0.94 i.e. 94% the two stocks are strongly correlated with each other

Distance calculation

Point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1



Minkowski dist^n $d(x,y) = \left(\sum_{k=1}^{n} |x_k - y_k|^r\right)^{\frac{1}{r}}$.

$r=1$, Manhatten dist^n / L_1 norm dist^n

$d(P_1,P_2) = \sum_{k=1}^{2} |x_k - y_k| = |0-2| + |2-0| = 4$

$d(P_1,P_3) = |0-3| + |2-1| = 4$

$d(P_1,P_4) = |0-5| + |2-1| = 6$

L_1	P_1	P_2	P_3	P_4
P_1	0	4	4	6
P_2	4	0	2	4
P_3	4	2	0	2
P_4	6	4	2	0

$r=2$, Euclidean dist^n

$d(P_1,P_2) = \sqrt{\sum_{k=1}^{2} (x_k - y_k)^2} = \sqrt{(0-2)^2 + (2-0)^2} = \sqrt{4+4} = \sqrt{8} = 2.8$

$d(P_1,P_3) = \sqrt{(0-3)^2 + (2-1)^2} = \sqrt{9+1} = \sqrt{10} = 3.16$

$d(P_2,P_3) = \sqrt{(2-3)^2 + (0-1)^2} = \sqrt{1+1} = \sqrt{2} = 1.41$

$r=\infty$, Supremum dist^n

$d(P_i,P_j) = \lim_{r\to\infty} \left(\sum_{k=1}^{n} |x_k - y_k|^r\right)^{\frac{1}{r}} = \max |x_{ik} - y_{jk}|$

$d(P_1,P_2) = \max\{|0-2|, |2-0|\} = \max\{2,2\} = 2$

$d(P_1,P_3) = \max\{|0-3|, |2-1|\} = \max\{3,1\} = 3$

$d(P_2,P_3) = \max\{|2-3|, |0-1|\} = \max\{1,1\} = 1$

Euclidean Distance.

L_2	P_1	P_2	P_3	P_4
P_1	0	2.8	3.2	5.1
P_2	2.8	0	1.4	3.2
P_3	3.2	1.4	0	2.0
P_4	5.1	3.2	2.0	0

Supremum Dist^n

L_∞	P_1	P_2	P_3	P_4
P_1	0	2	3	5
P_2	2	0	1	3
P_3	3	1	0	2
P_4	5	3	2	0

★ Measures of spread : Range & Variance.

Assume that feature age contains following four values in it — {20, 25, 10, 20}. Compute all measures of spread.

→ x = age = {20, 25, 10, 20}

n = number of values = 4

$\bar{x}$ = mean = $\frac{20+25+10+20}{4} = \frac{75}{4} = 18.75$

① Range : range(x) = max(x) − min(x) = $x_n - x_1$

= 25 − 10 = 15.

② Variance : $s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

$= \frac{1}{4-1} \left[(20-18.75)^2 + (25-18.75)^2 + (10-18.75)^2 + (20-18.75)^2 \right]$

$= \frac{1}{3} (118.75) = 39.58.$

③ Standard deviation = $\sigma(x) = \sqrt{s_x^2}$

$= \sqrt{39.58} = 6.291$

④ Absolute Average Deviation (AAD) :

$AAD(x) = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$

$= \frac{1}{3} \left[|20-18.75| + |25-18.75| + |10-18.75| + |20-18.75| \right]$

$= \frac{1}{3} \left[1.25 + 6.25 + 8.75 + 1.25 \right] = 1.071$

⑤ Median Absolute Deviation (MAD) :

$MAD(x) = \text{median}(\{ |x_1 - \bar{x}|, |x_2 - \bar{x}|, \ldots, |x_n - \bar{x}| \})$

$= \text{median}(\{ |20-18.75|, |25-18.75|, |10-18.75|, |20-18.75| \})$

$= \text{median}(\{ 1.25, 6.25, 8.75, 1.25 \})$

$= \frac{6.25 + 8.75}{2} = 7.5$

⑥ Inter Quartile Range IQR :

$IQR(x) = x_{75\%} - x_{25\%}$

Arrange x in increasing order = {10, 20, 20, 25}.

$x_{25\%} = 10.5 = \frac{10+20}{2}$

$x_{75\%} = 22.5 = \frac{20+25}{2}$

$IQR(x) = 22.5 - 10.5 = 12$