Chapter 2

1.what are the types of Data Sets

■ **Record** ■ Relational records ■ Data matrix, :numerical matrix ■ Document data: text documents: term-frequency vector ■ Transaction data	■ **Ordered** ■ Video data: sequence of images ■ Temporal data: time-series ■ Sequential Data: transaction sequences ■ Genetic sequence data
■ **Graph and network** ■ World Wide Web ■ Social or information networks ■ Molecular Structures	■ **Spatial, image and multimedia**: ■ Spatial data: maps ■ Image data: ■ Video data:

>>>2. what are Important Characteristics of Structured Data

- **Dimensionality** : Curse of dimensionality
- **Sparsity** : Only presence counts
- **Resolution**: Patterns depend on the scale
- **Distribution**:Centrality and dispersion

3 **Data Objects**

- Data sets are made up of data objects. A **data object** represents an entity.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses

>>> Also called ***samples , examples, instances, data points, objects, tuples***.

>>>Data objects are described by **attributes**.

>>>Database rows → data objects; columns → attributes.

>>> Attribute (or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object. *E.g., customer _ID, name, address*

4. Discrete vs. Continuous Attributes

Discrete Attribute	Continuous Attribute
■ Has only a finite or countably infinite set of values ■ E.g., **zip codes, profession**, or the set of words in a collection of documents ■ Sometimes, represented as **integer** variables ■ Note: **Binary attributes are a special case** of discrete attributes	■ Has **real** numbers as attribute values ■ **E.g., temperature, height, or weight** ■ Practically, real values can only be measured and represented using a finite number of digits ■ Continuous attributes are typically represented as **floating-point** variables

>>>5.Attribute Types

- **Nominal:** categories, states, or "**names of things**"
  - *Hair_color = { black, blond, brown, grey, red, white}*
  - marital status, occupation, ID numbers, zip codes
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., **gender**
  - Asymmetric binary: outcomes not equally important.
    - e.g., **medical test** (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal**
  - Values have a **meaningful order (ranking)** but magnitude between successive values is not known.
  - *Size = {small, medium, large}*, grades, army rankings
- **Numeric**
  - **Quantity** (integer or real-valued)
  - **Ratio**
    - Inherent **zero-point**
    - e.g**.**, ***temperature in Kelvin****, length, counts, monetary quantities*
  - **Interval**
    - No **zero-point** . Scale of **equal-sized units**. Values have order
    - E.g., ***temperature in C°or F°****, calendar dates*

6. Basic Statistical Descriptions of Data

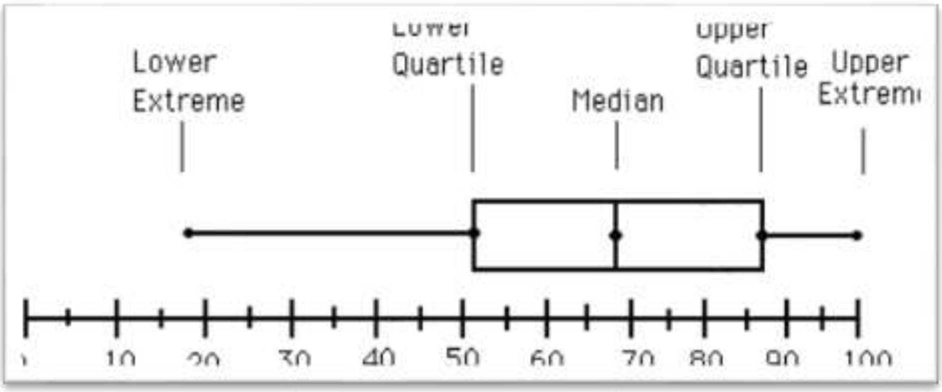
- Motivation :To better understand the data: central tendency, variation and spread
- Data dispersion characteristics : median, max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
- Dispersion analysis on computed measures correspond to transformed cube

>>>6. Graphic Displays of Basic Statistical Descriptions

- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis repres. frequencies
- **Quantile plot**: each value x_i is paired with f_i indicating that approximately 100 f_i% of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

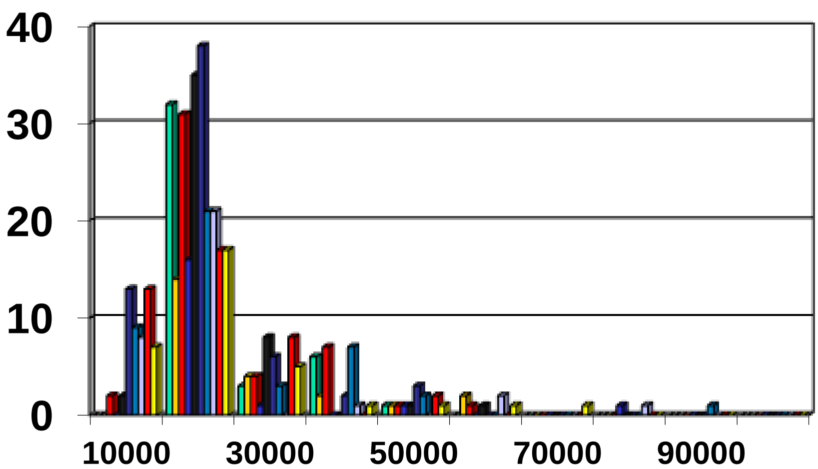
7 . Boxplot Analysis

- Five-number summary of a distribution: **Minimum, Q1, Median, Q3, Maximum**
- Data is represented with a **box**
- The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
- The median is marked by a line within the box
- **Whiskers**: two lines outside the box extended to Minimum and Maximum
- **Outliers**: points beyond a specified outlier threshold, plotted individually



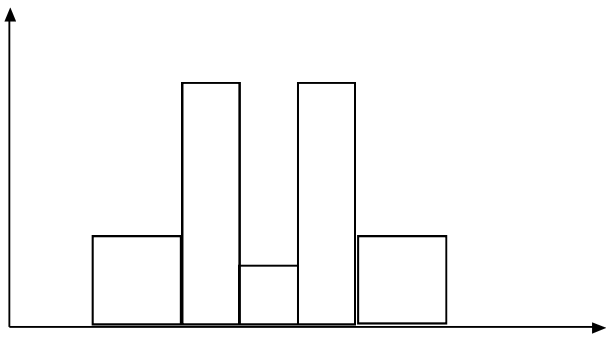
8. Histogram Analysis

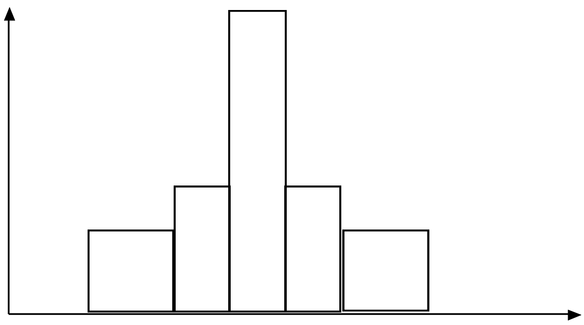
- Histogram: Graph display of tabulated **frequencies, shown as bars**
- It shows what proportion of cases fall into each of several categories
- **>>>Differs from a bar chart in that it is the *area* of the bar that denotes the value, not the height** as in bar charts, a crucial distinction when the categories are not of uniform width
- The categories are usually specified as **non-overlapping** intervals of some variable. The categories (bars) must be adjacent



9.why Histograms Often Tell More than Boxplots

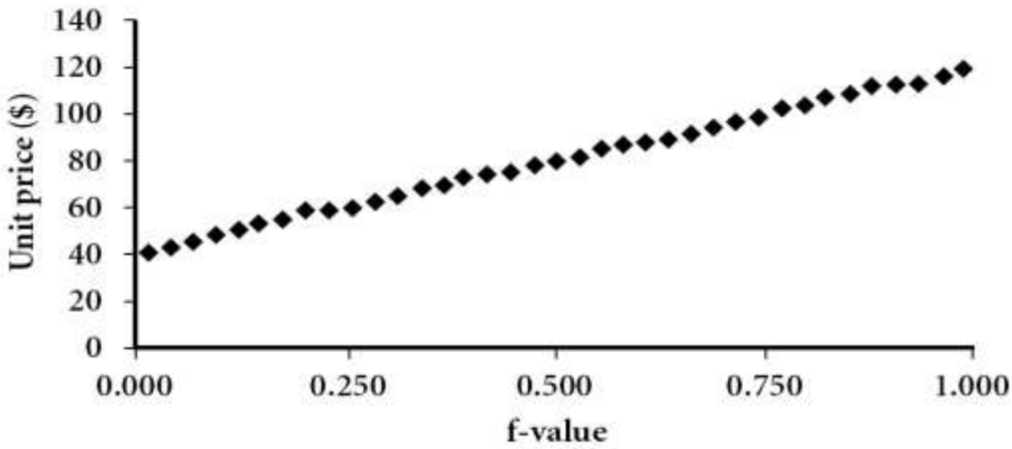
- The two histograms may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions





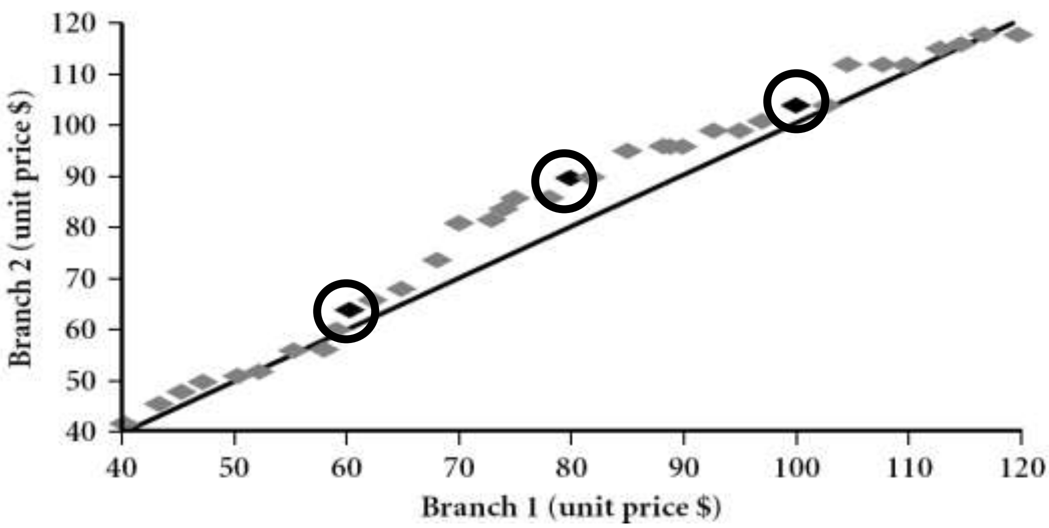
10. Quantile Plot

- **Displays** all of the data (allowing the user to assess both the **overall behavior and unusual** occurrences)
- For a data x_i **data sorted in increasing order,** f_i **indicates that approximately 100** f_i**%** of the data are below or equal to the value x_i



11. Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one **univariate** distribution **against** the corresponding quantiles of **another**
- View: Is there is a shift in going from one distribution to another? **yes**

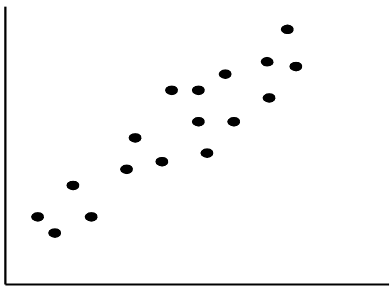


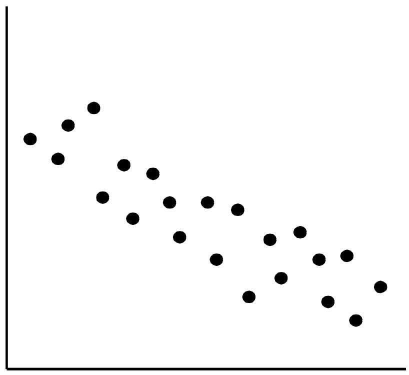
We need to **label** the dark plotted points as **Q1, Median, Q3** – that would help in understanding this graph.

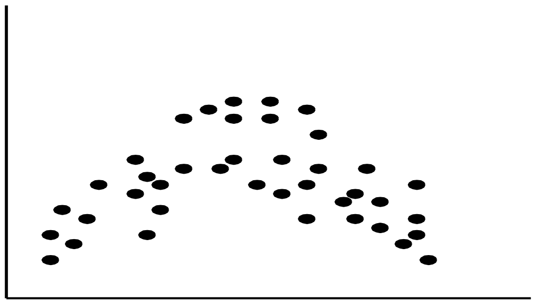
12. Scatter plot

- Provides a first look at **bivariate** data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points
- Determine Positively and Negatively **Correlated Data**

Positive **negative** left half fragment is positively correlated
right half is negative correlated



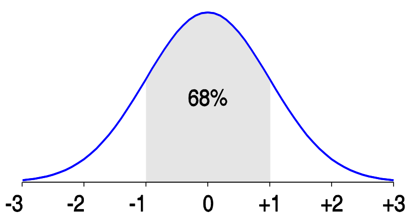


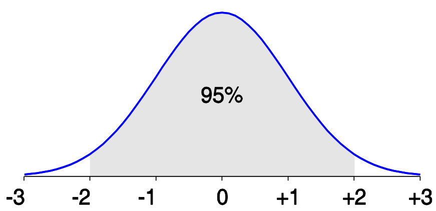


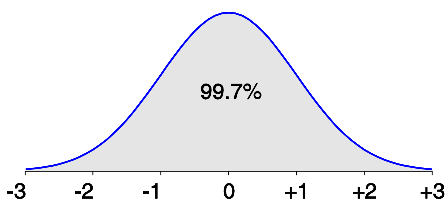
Uncorrelated Data

13. Properties of Normal Distribution Curve(μ: mean, σ: standard deviation)

- From **μ–σ to μ+σ: contains about 68%** of measurements
- From **μ–2σ to μ+2σ: contains about 95%** of it
- From **μ–3σ to μ+3σ: contains about 99.7%** of it







14. Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):

  Note: *n* is sample size and *N* is population size.

$$\mu = \frac{\sum x}{N} \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - Weighted arithmetic mean:

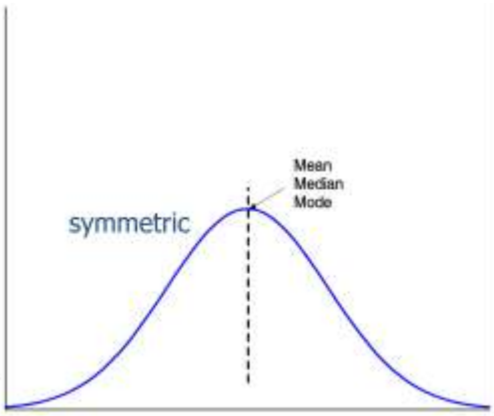
$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

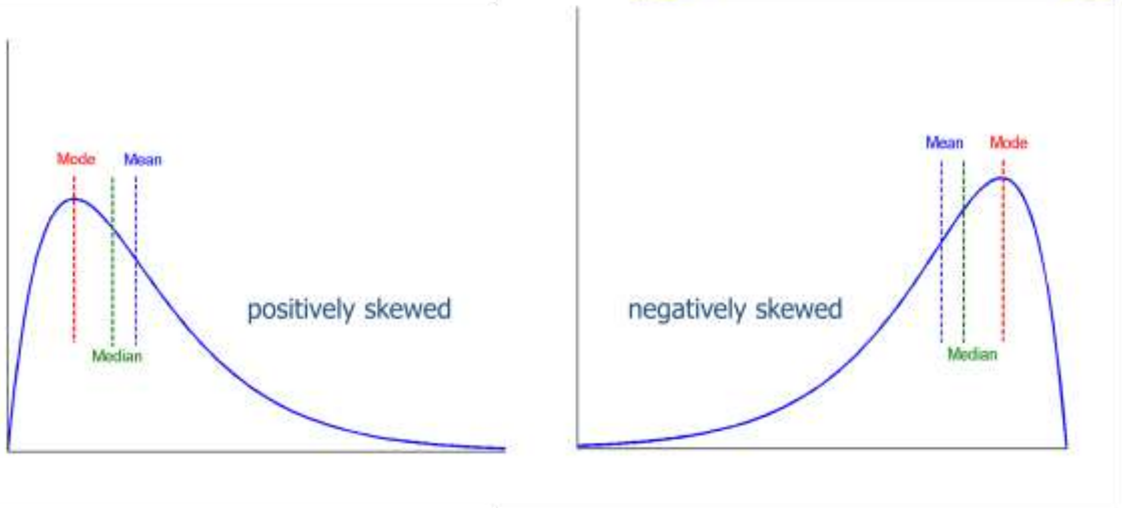
  - Trimmed mean: chopping extreme values
- Median:
  - Middle value if odd number of values, or average of the middle two values
  - Estimated by interpolation (for *grouped data*):

$$median = L_1 + \left(\frac{n/2 - (\sum freq)l}{freq_{median}}\right)width$$

- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula: $mean - mode = 3 \times (mean - median)$

Symmetric vs. Skewed Data





15. Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
  - **Inter-quartile range**: IQR = $Q_3 - Q_1$
  - **Five number summary**: min, Q_1, median, Q_3, max
  - **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - **Outlier**: usually, a value higher/lower than 1.5 x IQR
- Variance and standard deviation (*sample: s, population: σ)*
  - **Variance**: (algebraic, scalable computation)

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

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2]$$

  - **Standard deviation** *s (or σ)* is the square root of variance s^2 *(or σ^2)*

16. Proximity refers to a similarity or dissimilarity

Similarity	**Dissimilarity** (e.g., distance)
■ Numerical measure of **how alike** two data objects are ■ Value is higher when objects are more alike ■ Often falls in the **range [0,1]**	■ Numerical measure of **how different** two data objects are ■ Lower when objects are more alike ■ Minimum dissimilarity is often 0 ■ **Upper limit varies**

17. Data Matrix and Dissimilarity Matrix

Data matrix	Dissimilarity matrix
■ n data points with p dimensions ■ **Two modes** $\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$	■ n data points, but registers only the distance ■ A **triangular** matrix ■ **Single mode** $\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$

اللي جاي قوانين ومسائل
صحصح لو سمحت

➔ Proximity Measure for Nominal Attributes

- Can take 2 or more states, e.g., red, blue, green (generalization of a binary attribute)
- Method 1: Simple matching : ***m*: # of matches, *p*: total # of variables**

$$d(i,j)=\frac{p-m}{p}$$

- Method 2: Use a large number of binary attributes
  - creating a new binary attribute for each of the *M* nominal states

➔ Proximity Measure for Binary Attributes

- A contingency table for binary data

		Object *j*		
		1	0	sum
Object *i*	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

- **Distance** measure for **symmetric** binary variables:

$$d(i,j)=\frac{r+s}{q+r+s+t}$$

- **Distance** measure for **asymmetric** binary variables:

$$d(i,j)=\frac{r+s}{q+r+s}$$

- Jaccard coefficient (***similarity*** measure for ***asymmetric*** binary variables):

$$sim_{Jaccard}(i,j)=\frac{q}{q+r+s}$$

Note: Jaccard coefficient is the same as "**coherence**":

$$coherence(i,j)=\frac{sup(i,j)}{sup(i)+sup(j)-sup(i,j)}=\frac{q}{(q+r)+(q+s)-q}$$

>>>18.determine Dissimilarity between Binary Variables

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is a symmetric attribute
- The remaining attributes are asymmetric binary

➔ Let the values Y and P be 1, and the value N 0

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

Standardizing Numeric Data

- Z-score: $z = \frac{x-\mu}{\sigma}$
  - X: raw score to be standardized, μ: mean of the population, σ: standard deviation
  - negative when the raw score is below the mean, "+" when above
- An alternative way: Calculate the **mean absolute deviation**

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|)$$

Where $m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{nf})$.

- >> Using mean absolute deviation is more robust than using standard deviation

Distance on Numeric Data: Minkowski Distance

- *Minkowski distance*: A popular distance measure

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

where *i* = (x_{i1}, x_{i2}, …, x_{ip}) and *j* = (x_{j1}, x_{j2}, …, x_{jp}) are two *p*-dimensional data objects, and *h* is the order (the distance so defined is also called L-*h* norm)

- **Properties**
  - **d(i, j) > 0 if i ≠ j, and d(i, i) = 0 (Positive definiteness)**
  - **d(i, j) = d(j, i) (Symmetry)**
  - **d(i, j) ≤ d(i, k) + d(k, j) (Triangle Inequality)**
- A distance that satisfies these properties is a metric

Special Cases of Minkowski Distance

- *h* = 1: <u>**Manhattan**</u> (city block, L_1 norm) distance
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i, j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_p} - x_{j_p}|$$

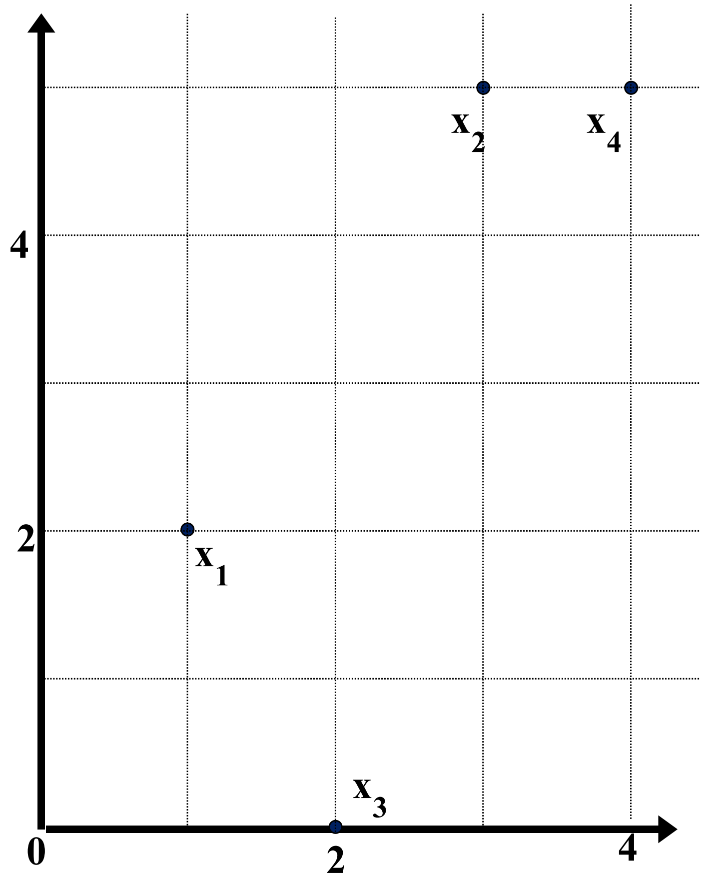
- $h = 2$: (L_2 norm) **Euclidean** distance

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

- $h \to \infty$. "**supremum**" (L_{max} norm, L_∞ norm) distance.
- This is the maximum difference between any component (attribute) of the vectors

$$d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|$$

19. Calculate Data Matrix and Dissimilarity Matrix (Minkowski Distances)



Data Matrix

point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Minkowski Distance

Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Ordinal Variables

- Order is important, e.g., rank. An ordinal variable can be discrete or continuous
- Can be treated like interval-scaled
  - replace x_{if} by their rank $r_{if} \in \{1, ..., M_f\}$
  - map the range of each variable onto [0, 1] by replacing *i*-th object in the *f*-th variable by $z_{if} = \frac{r_{if} - 1}{M_f - 1}$
  - compute the dissimilarity using methods for interval-scaled variables

Attributes of Mixed Type

- A database may contain all attribute types
  - Nominal, symmetric binary, asymmetric binary, numeric, ordinal
- One may use a weighted formula to combine their effects

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

  - *f* is binary or nominal:
    $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$ otherwise

- f is numeric: use the normalized distance
- f is ordinal
  - Compute ranks r_{if} and
  - Treat z_{if} as interval-scaled

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

Cosine Similarity

- A **document** can be represented by thousands of attributes, each recording the *frequency* of a particular word (such as keywords) or phrase in the document.

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- Other vector objects: gene features in micro-arrays, …
- Applications: information retrieval, biologic taxonomy, gene feature mapping, ...
- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \; ||d_2|| ,$$

where • indicates vector dot product, $||d||$: the length of vector d

<u>20 Find the similarity between documents 1 and 2.</u>

d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)

d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

$d_1 \bullet d_2$ = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25

$||d_1||$= $(5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5}$=$(42)^{0.5}$ = 6.481

$||d_2||$= $(3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5}$=$(17)^{0.5}$ = 4.12

$\cos(d_1, d_2)$ = 0.94