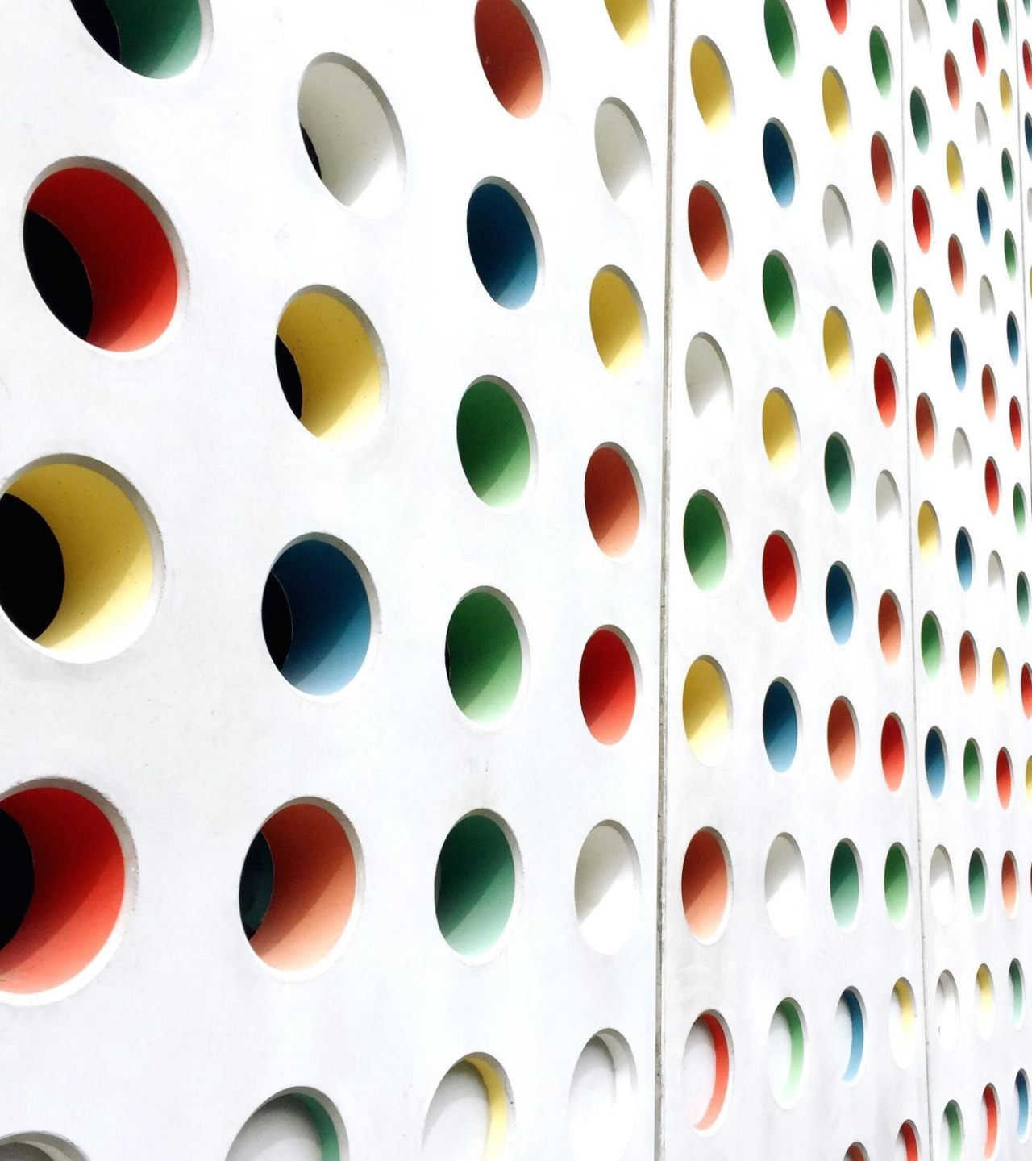


Probability Distribution and Data Modelling

BM 4419

BUSINESS ANALYTICS

Outline

- **Data integration and transformation**
- **Data reduction**
- **Summary**

May2024 _HK

Learning Outcome

At the end of this lecture, students will be able to:

To explore the appropriate technique for data pre-processing.

May2024 _HK

Data Integration

Data Integration

- Combines data from multiple sources into a coherent store
- Identify real world entities from multiple data sources

Schema

Integrate metadata from different sources.

Detecting and resolving data value conflicts :

- For the same real world entity, attribute values from different sources are different
- Possible reasons: different representations, different scales

HK

Handling Redundancy in Data Integration

Redundant attributes may be able to be detected by ***correlation analysis***

May2024 _HK

LEARNING OBJECTIVE (CORRELATION ANALYSIS)

- TO DRAW A SCATTER PLOT FOR A SET OF ORDER PAIRS.
- TO COMPUTE THE CORRELATION COEFFICIENT.

STATISTICAL METHOD

- ***CORRELATION*** IS A STATISTICAL METHOD USED TO DETERMINE WHETHER A LINEAR RELATIONSHIP BETWEEN VARIABLES EXISTS.

May2024 _HK

STATISTICAL QUESTIONS

1. ARE TWO OR MORE VARIABLE RELATED?
2. IF SO, WHAT IS THE STRENGTH OF THE RELATIONSHIP?
3. WHAT TYPE OR RELATIONSHIP EXISTS?

---

4. WHAT KIND OF PREDICTIONS CAN BE MADE FROM THE RELATIONSHIP?

May2024 _HK

SCATTER DIAGRAM

- A SCATTER PLOT IS GRAPH OF THE ORDER PAIRS (x, y) OF NUMBERS CONSISTING OF THE INDEPENDENT VARIABLE, x, AND THE DEPENDENT VARIABLE, y ARE PLOTTED BY USING CARTESIAN COORDINATES
- A SCATTER PLOT IS A VISUAL WAY TO DESCRIBE THE NATURE OF THE RELATIONSHIP BETWEEN THE INDEPENDENT AND DEPENDENT VARIABLE MEASURED ON THE SAME INDIVIDUALS.

May2024 _HK

SCATTER DIAGRAM

- IN EXAMINING A SCATTER DIAGRAM, LOOK FOR AN OVERALL PATTERN SHOWING THE

a) FORM (LINEAR RELATIONSHIP, CURVED RELATIONSHIP, CLUSTERS)

b) DIRECTION (POSITIVE OR NEGATIVE ASSOCIATION)

c) STRENGTH OF THE RELATIONSHIP

d) OUTLIERS

- WHEN THE POINTS ON THE SCATTER DIAGRAM APPEAR TO LIE NEAR A STRAIGHT LINE, KNOW AS REGRESSION LINE. THEN, THERE IS A LINEAR CORRELATION (LINEAR RELATIONSHIP) BETWEEN TWO VARIABLE.

Correlation and Cautions

- **Positive Correlation:** The correlation is said to be positive correlation if the values of two variables changing with same direction.

  Ex. Height & weight, temperature & ice cream sales.

- **Negative Correlation:** The correlation is said to be negative correlation when the values of variables change with opposite direction.

  Ex. Price & quantity demanded.

May2024 _HK

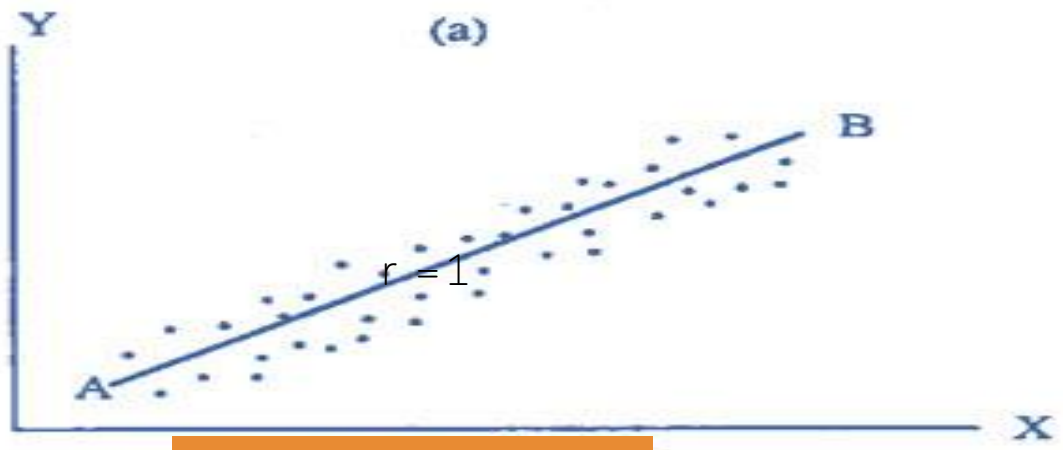
Direction of Correlation

- **Positive relationship** – Variables change in the same direction.
  - As X is increasing, Y is increasing
  - As X is decreasing, Y is decreasing
  - E.g., As height increases, so does weight.
- **Negative relationship** – Variables change in opposite directions.
  - As X is increasing, Y is decreasing
  - As X is decreasing, Y is increasing
  - E.g., As TV time increases, grades decrease

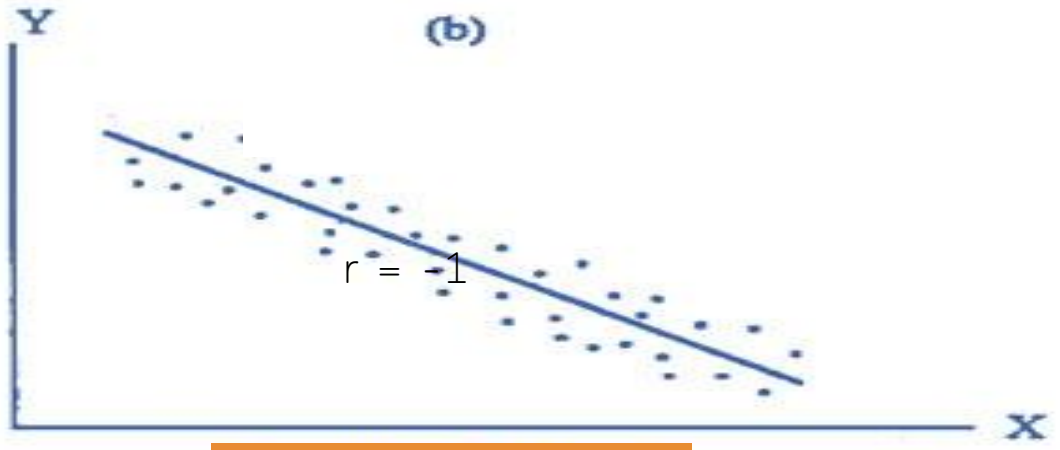
Indicated by sign; (+) or (-).

HK

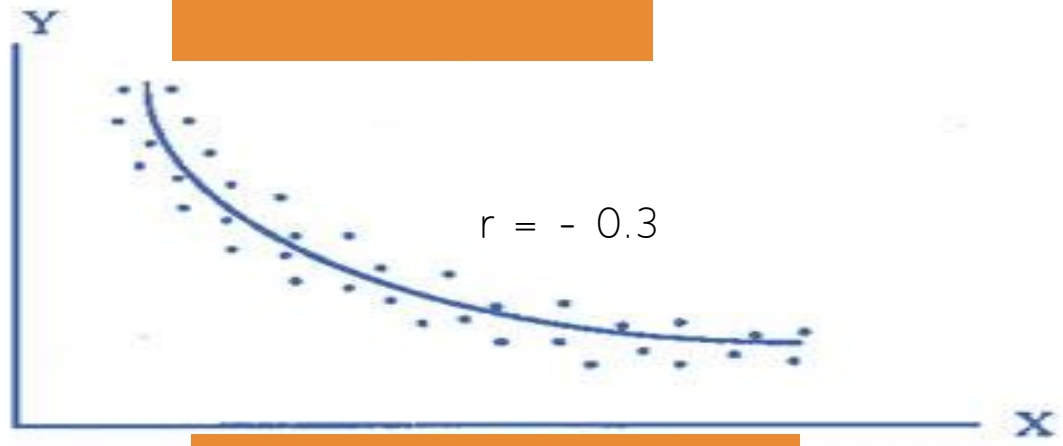
SCATTER PLOT EXAMPLE



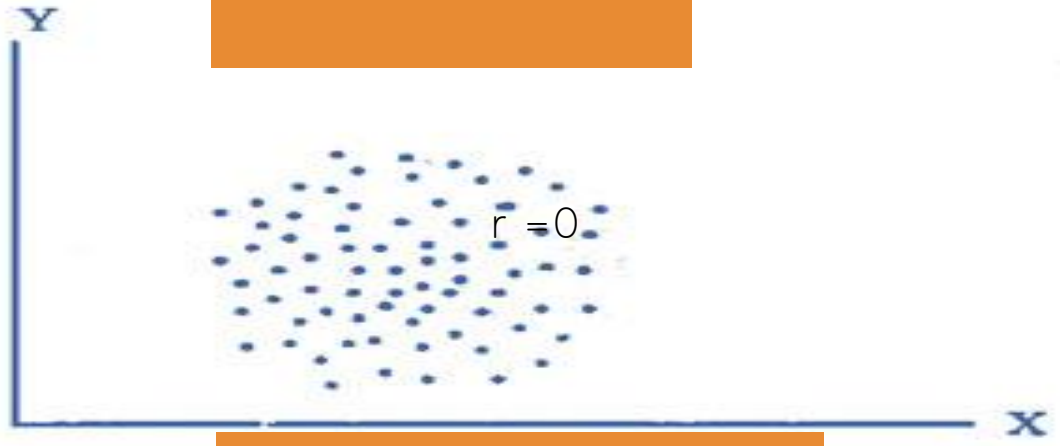
Positive and Linear



Negative and Linear



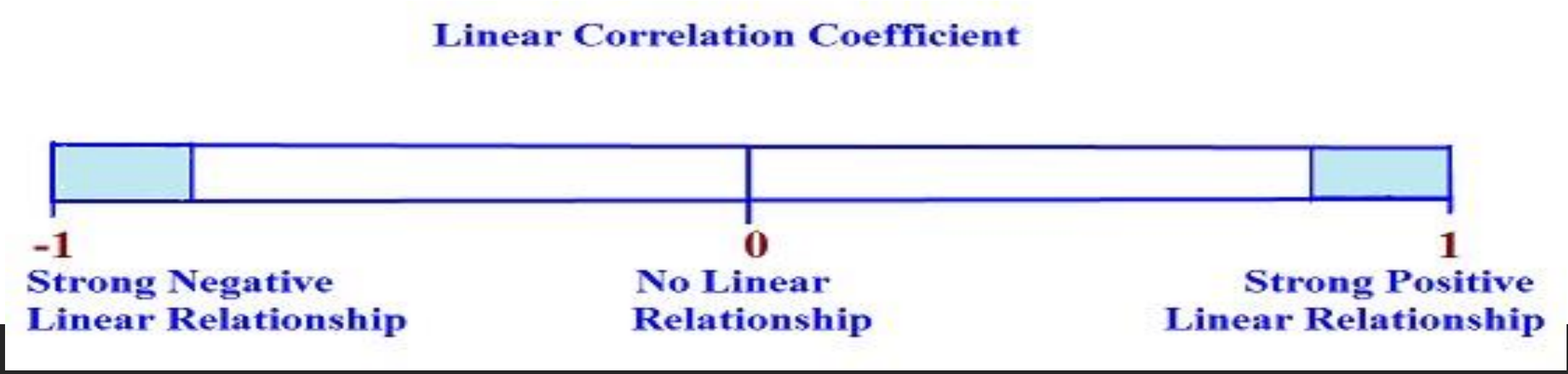
Negative and Non- Linear



No Relationship

CORRELATION COEFFICIENT

- THE STRENGTH OF A LINEAR RELATION BETWEEN x AND y VARIABLES IS DETERMINED BY THE CORRELATION COEFFICIENT DENOTED BY r.
- WHEN THERE IS NO LINEAR RELATIONSHIP BETWEEN THE VARIABLES OR ONLY A WEAK RELATIONSHIP, THE VALUE OF r WILL BE CLOSED TO 0.



FORMULA FOR THE CORRELATION COEFFICIENT, r

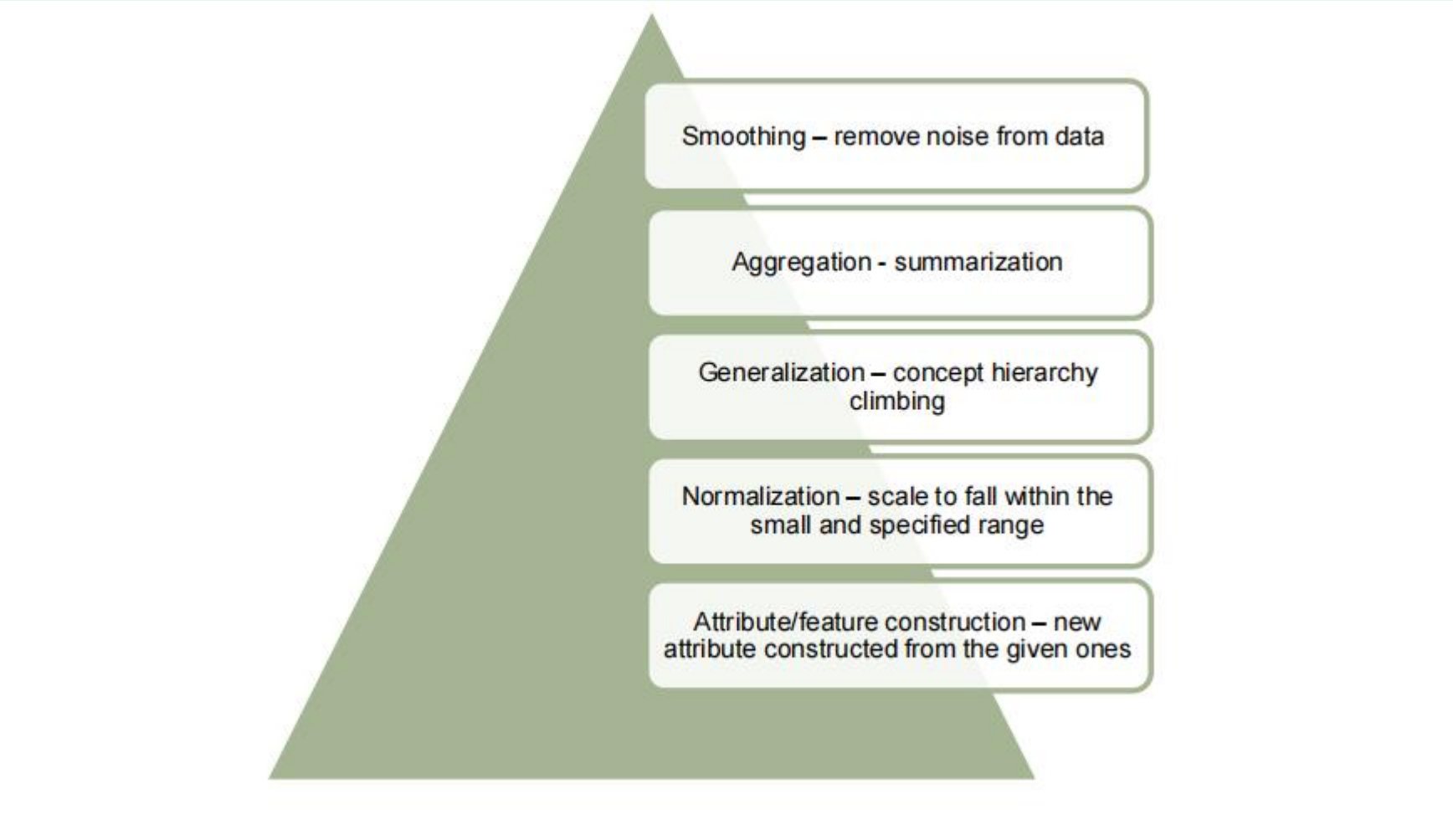
- ✓ DEVELOPED BY KARL PEARSON IN THE EARLY 1900S.
- ✓ THE PEARSON'S PRODUCT MOMENT COEFFICIENT IS A NUMERICAL MEASURE.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

May2024 _HK

WHERE n IS THE NUMBER OF DATA PAIRS.

DATA TRANSFORMATION



Data Transformation: Normalization

- Z-score normalization (μ: mean, σ: standard deviation):The standard normal distribution is a normal distribution with a mean of 0 and a standard deviation of 1.

- It is also called as z-distribution.

HK

Z-Score

- A z-value is the distance between a selected value, designated X, and the population mean μ, divided by the population standard deviation, σ.
- The formula is:

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

- z-scores make it easier to compare data values measured on different scales.

HK

Z-Score

A *z*-score reflects how many standard deviations above or below the mean a raw score is.

The *z*-score is positive if the data value lies above the mean and negative if the data value lies below the mean.

HK

Quick Test (Analyze The Data)

Suppose SAT scores among college students are normally distributed with a mean of 500 and a standard deviation of 100. If a student scores a 700, what would be her z-score?

Answer Now

May2024 _HK

Quick Test (Analyze The Data)

- A set of math test scores has a mean of 70 and a standard deviation of 8.
- A set of English test scores has a mean of 74 and a standard deviation of 16.

For which test would a score of 78 have a higher standing?

Answer Now

1 _HK

Quick Test (Analyze The Data)

What will be the miles per gallon for a Toyota Camry when the average mpg is 23, it has a z value of 1.5 and a standard deviation of 2?

Answer Now

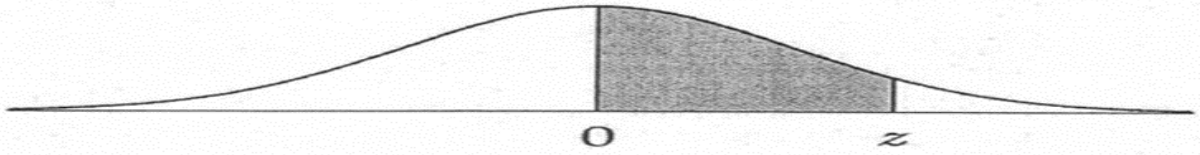
_HK

Method for obtaining probability (area) under a Standard Normal Curve

1) Find the area to the left of $Z = z_0$
2) Find the area to the right of $Z = z_0$
3) Find the area between $Z = z_0$ and $Z = z_1$

- The standard normal table can find the area (probability) under the standard normal curve.
- The table used in this topic is the "between the mean 0 and the listed value of z" standard normal table.

Table 2: The Standard Normal distribution



The tabulated values are the probabilities between the mean 0 and the listed values of z.

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
4.0	0.49997									

Example:

Find the area under the standard normal curve:

a) Between $z = 0$ and $z = 1.95$
b) Area to the right of $z = 2.32$
c) Area to the right of $z = -0.75$
d) Area to the left of $z = -1.54$
e) Between $z = -2.17$ and $z = 0$
f) Between $z = -1.56$ and $z = 2.31$

Data Reduction Strategy

❖ Why data reduction?

· A database/data warehouse may store terabytes of data

· Complex data analysis/mining may take a very long time to run on the complete data set

❖ Data reduction

· Obtain a reduced representation of the data set that is much smaller in volume but yet

produce the same (or almost the same) analytical results

❖ Data reduction strategies

· Dimensionality reduction – e.g., remove unimportant attributes

Dimensionality Reduction: Principal Component Analysis (PCA)

❖Given N data vectors from n-dimensions, find k ≤ n orthogonal vectors (principal components) that can be best used to represent data

❖Steps

· Normalize input data: Each attribute falls within the same range

· Compute k orthonormal (unit) vectors, i.e., principal components

· Each input data (vector) is a linear combination of the k principal component vectors

· The principal components are sorted in order of decreasing "significance" or strength

· Since the components are sorted, the size of the data can be reduced by eliminating the weak components, i.e., those with low variance. (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data

❖· Works for numeric data only

❖· Used when the number of dimensions is large

Summary

❖Data preparation or preprocessing is a big issue for both data warehousing and data mining

❖Descriptive data summarization is need for quality data preprocessing

❖Data preparation includes

· Data cleaning and data integration

· Data reduction and feature selection

❖ A lot a methods have been developed but data preprocessing still an active area of research