Statistics and Probability

Hafara Firdausi, M.Kom.

Department of Information Technology
Faculty of Electrical and Intelligent Informatics
Institut Teknologi Sepuluh Nopember

01. Introduction to Statistics

Pengantar Statistika

Outline

1. Knowledge Discovery
2. Data and Variable
3. Scale of Measurements

Knowledge Discovery

Data

Data is a **collection of facts** from which conclusions may be drawn.

Data are the **raw material** from which information is derived. The information must itself be studied and its patterns analyzed further, leading to **knowledge discovery**.

```
dvdrental=# select title, release_year, length, replacement_cost from film
dvdrental-#   where length > 120 and replacement_cost > 29.50
dvdrental-#   order by title desc;
          title           | release_year | length | replacement_cost
--------------------------+--------------+--------+------------------
 West Lion                |         2006 |    159 |            29.99
 Virgin Daisy             |         2006 |    179 |            29.99
 Uncut Suicides           |         2006 |    172 |            29.99
 Tracy Cider              |         2006 |    142 |            29.99
 Song Hedwig              |         2006 |    165 |            29.99
 Slacker Liaisons         |         2006 |    179 |            29.99
 Sassy Packer             |         2006 |    154 |            29.99
 River Outlaw             |         2006 |    149 |            29.99
 Right Cranes             |         2006 |    153 |            29.99
 Quest Mussolini          |         2006 |    177 |            29.99
 Poseidon Forever         |         2006 |    159 |            29.99
 Loathing Legally         |         2006 |    140 |            29.99
 Lawless Vision           |         2006 |    181 |            29.99
 Jingle Sagebrush         |         2006 |    124 |            29.99
 Jericho Mulan            |         2006 |    171 |            29.99
 Japanese Run             |         2006 |    135 |            29.99
 Gilmore Boiled           |         2006 |    163 |            29.99
 Floats Garden            |         2006 |    145 |            29.99
 Fantasia Park            |         2006 |    131 |            29.99
 Extraordinary Conquerer  |         2006 |    122 |            29.99
 Everyone Craft           |         2006 |    163 |            29.99
 Dirty Ace                |         2006 |    147 |            29.99
 Clyde Theory             |         2006 |    139 |            29.99
 Clockwork Paradise       |         2006 |    143 |            29.99
 Ballroom Mockingbird     |         2006 |    173 |            29.99
(25 rows)
```

Knowledge discovery

From **Data,** we derive **Information**, gain **Knowledge,** and produce **Wisdom**:

D → I → K → W

The effort is sometimes described as a **DIKW pyramid** or **DIKW hierarchy**
(Rowley 2007)

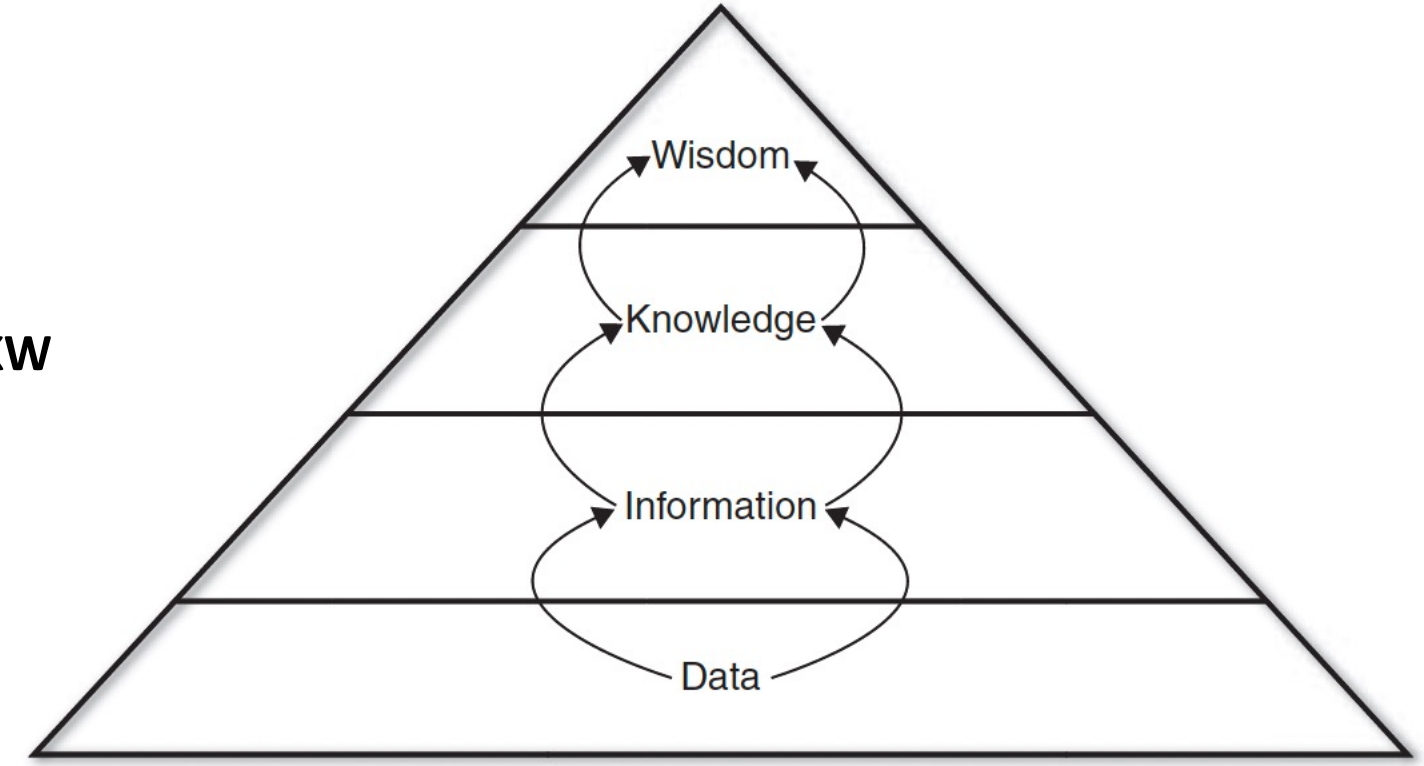
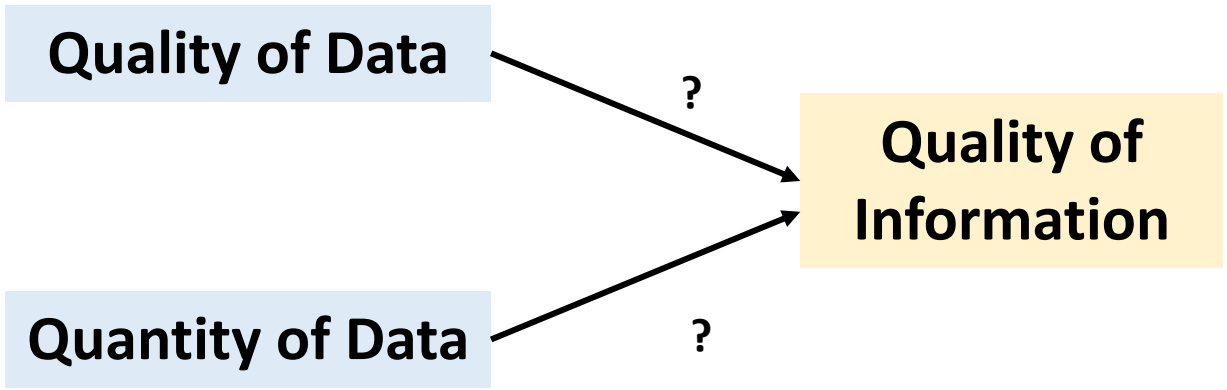


Figure 1.1 The DIKW pyramid.

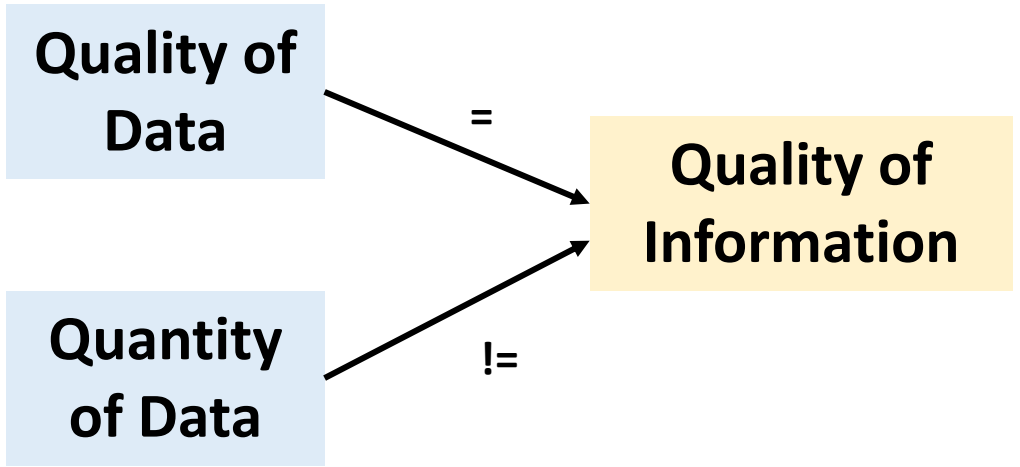
Data and Variable

Data Quality vs Quantity



Data Quality vs Quantity

- Data analytics is the need for sufficiently high quality in the data under study
- GIGO principle:
  - **“if Garbage goes In, Garbage come Out”** (Hand et al. 2001, Section 2.6)
- That is, the **quality and value of any data mining** is contingent upon the **quality of the underlying data**
- However, **the quantity of data does not always equate with the quality of information**



Two general forms of data quality distortion

Individual

- Errors in collection, entry, or some other form of disruption
  - Misplaced decimal points
  - Transposed digits
  - Measurement rounding errors
  - Missing data records
  - Impossible combinations in classification fields (example: pregnant = “yes”/sex = “male”)

Collective

- Irregularities in the selection mechanisms under which the data were identified or sampled

The Quality of Data is good, if…

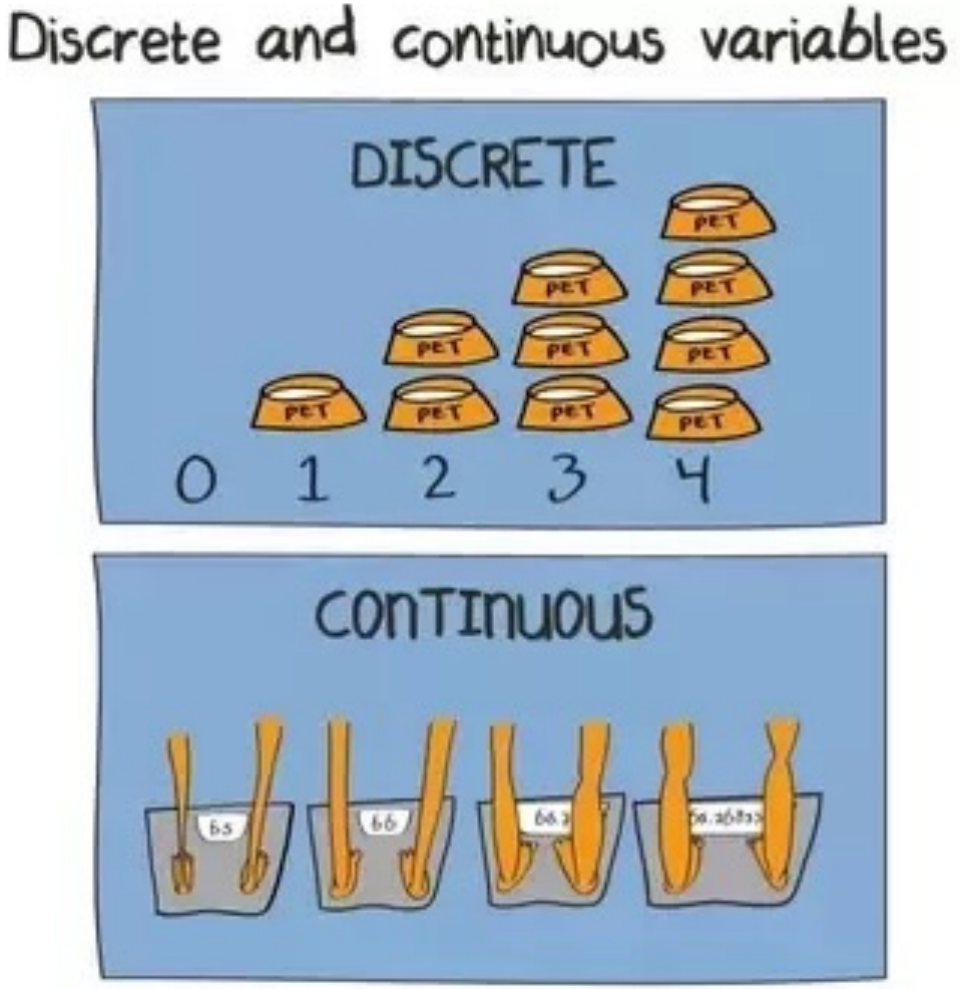
- **Objective**, reflecting the actual situation (as it is)
- **Representative**
- **The standard error** (kesalahan baku) **must be small**. An estimate is considered good (having a high degree of accuracy) if the standard error is small
- **Timely** (up to date)
- **Relevant**, meaning the data collected must be related to the problem to be solved

```
dvdrental=# select title, release_year, length, replacement_cost from film
dvdrental-#   where length > 120 and replacement_cost > 29.50
dvdrental-#   order by title desc;
          title           | release_year | length | replacement_cost
--------------------------+--------------+--------+------------------
 West Lion                |         2006 |    159 |            29.99
 Virgin Daisy             |         2006 |    179 |            29.99
 Uncut Suicides           |         2006 |    172 |            29.99
 Tracy Cider              |         2006 |    142 |            29.99
 Song Hedwig              |         2006 |    165 |            29.99
 Slacker Liaisons         |         2006 |    179 |            29.99
 Sassy Packer             |         2006 |    154 |            29.99
 River Outlaw             |         2006 |    149 |            29.99
 Right Cranes             |         2006 |    153 |            29.99
 Quest Mussolini          |         2006 |    177 |            29.99
 Poseidon Forever         |         2006 |    159 |            29.99
 Loathing Legally         |         2006 |    140 |            29.99
 Lawless Vision           |         2006 |    181 |            29.99
 Jingle Sagebrush         |         2006 |    124 |            29.99
 Jericho Mulan            |         2006 |    171 |            29.99
 Japanese Run             |         2006 |    135 |            29.99
 Gilmore Boiled           |         2006 |    163 |            29.99
 Floats Garden            |         2006 |    145 |            29.99
 Fantasia Park            |         2006 |    131 |            29.99
 Extraordinary Conquerer  |         2006 |    122 |            29.99
 Everyone Craft           |         2006 |    163 |            29.99
 Dirty Ace                |         2006 |    147 |            29.99
 Clyde Theory             |         2006 |    139 |            29.99
 Clockwork Paradise       |         2006 |    143 |            29.99
 Ballroom Mockingbird     |         2006 |    173 |            29.99
(25 rows)
```

Variable

A variable is a **characteristic** of data

- **A discrete variable** is a variable with values that can be counted or are finite.
- **A continuous variable** is a variable with unlimited values that can be measured or recorded to the required level of precision.



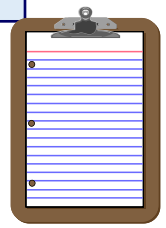
Data Sources

Primary

Data Collection

Observation

Survey

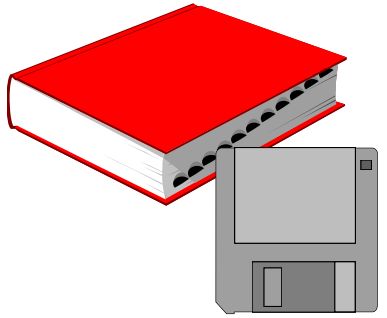


Experimentation

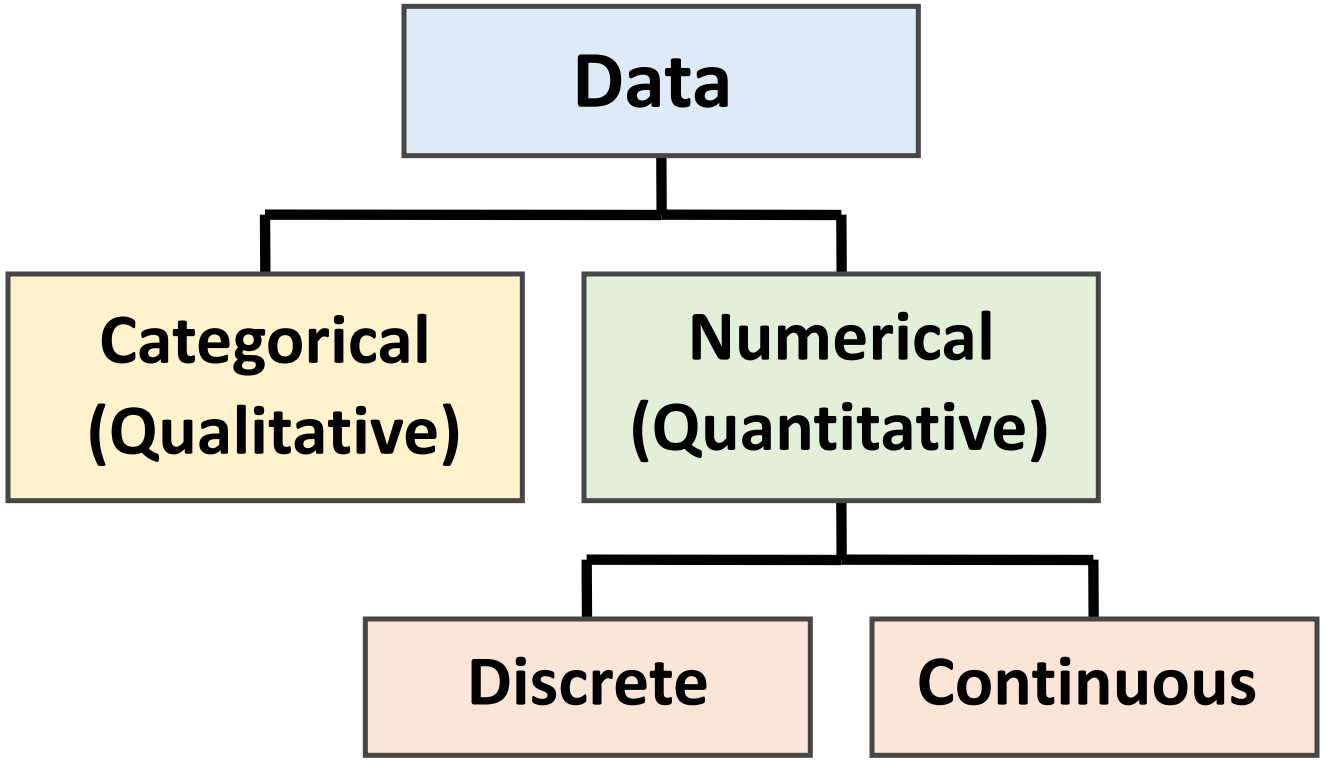
Secondary

Data Compilation

Print or Electronic



Data Types

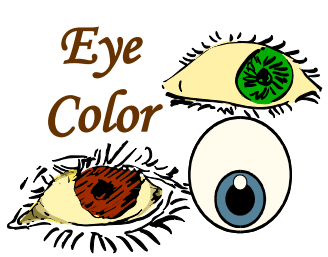

Data
Categorical (Qualitative)
Numerical (Quantitative)
Discrete
Continuous

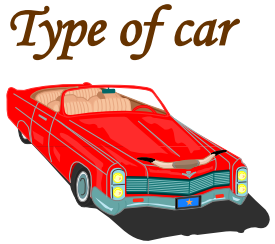
Data Types

Qualitative - Categorical or Nominal:

Examples are-

- Color
- Gender
- Nationality

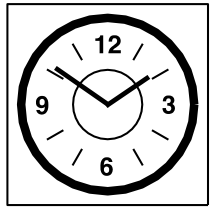


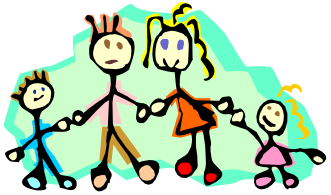


Quantitative - Measurable or Countable:

Examples are-

- Temperatures
- Salaries
- Number of points scored on a 100-point exam





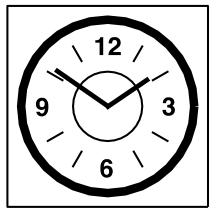
Data Quantitative

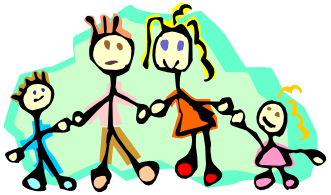
Discrete Variable

- Can only assume certain values and there are usually “**gaps**” between values.
- Discrete random variables produce numerical responses that arise from a **counting process**.
- **Example**: the number of bedrooms in a house, or the number of hammers sold at the local Home Depot (1, 2, 3, …, etc).

Continuous Variable

- A Continuous Variable can assume any value within a **specified range.**
- Continuous random variables produce numerical responses that arise from a **measuring process.**
- **Example**: The pressure in a tire, the weight of meat, the height of student





Classify this variabels as quantitative or qualitative!!

Hand phones' brand

Human Height

GPA

Kind of jobs

Numbers of computer selling

Students' score

Education level

Weight of cow

Human IQ

Number of tomato

Exercise

- Classify the following data as **quantitative** or **qualitative**. Explain your choice.
  - The state of birth of the President of the United States
  - The marital status of a corporation president
  - The price of a new textbook
  - The number of cars that enter a parking lot during a given day

Exercise

- Classify the following data as **discrete** or **continuous**. Explain your choice.
  - The number of pumps at a gas station
  - The SAT score of a randomly selected student
  - The annual income of a bank president
  - The number of elevators in a hotel lobby
  - The number of courses taken by a college freshman

- Classify the following data as **discrete** or **continuous**. Explain your choice.
  - The length of time for a long-distance phone call
  - The volume of gasoline remaining in a car's tank
  - The number of customers waiting in line at a cash register

Scales of Measurement

Scales of Measurement

There are four generally used measurement scales, listed from weakest to strongest :

1. Nominal Scale - groups or classes

- Gender

2. Ordinal Scale - order matters

- Ranks (top ten videos)

3. Interval Scale - difference or distance matters – has an arbitrary zero value

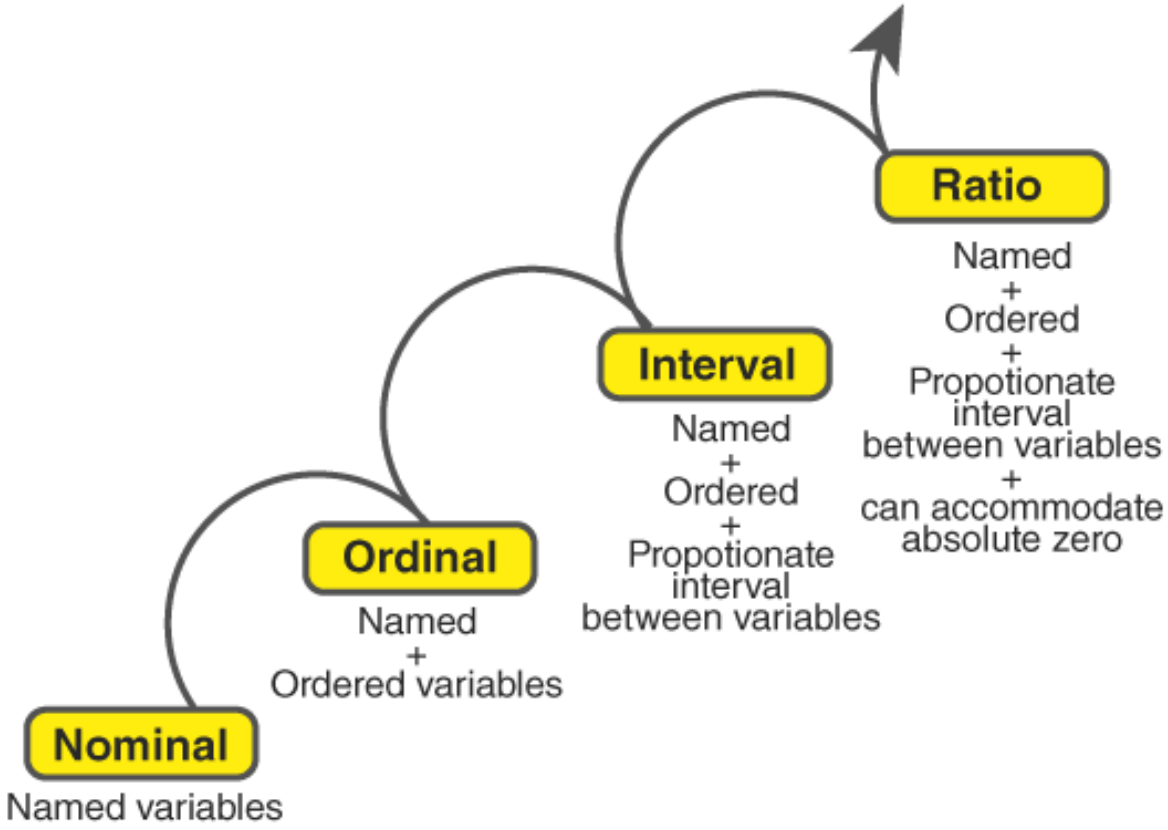
- Temperatures (^{0}F, ^{0}C)

4. Ratio Scale - Ratio matters – has a natural zero value

- Salaries

The weaker the scale of measurement, the less we can assume about relations among elements on the scale.

Scales of Measurement



The weaker the scale of measurement, the less we can assume about relations among elements on the scale.

Nominal Scale

- Values are used merely to represent the **class or category** to which an observation belongs.
- Nominal data are labels for **groups or classes.**
- Nominal data may be **verbal** or may be recorded as **numerical** codes.

Chart of number of males and females in a given class

Sex	Number of students
Male	14
Female	17

Graph of number of the number of males and females in a given class

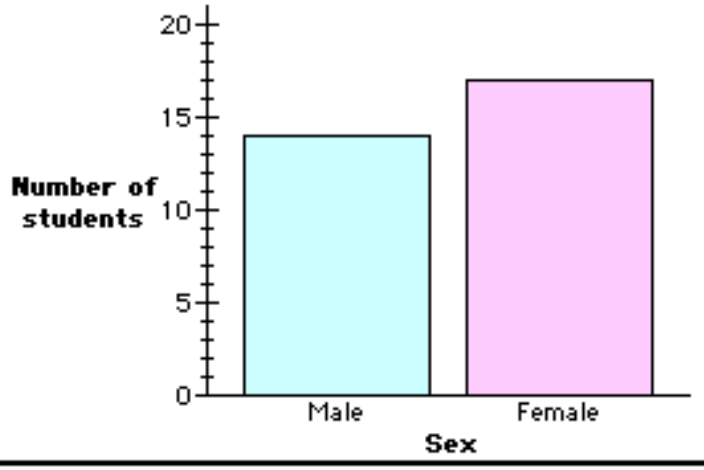
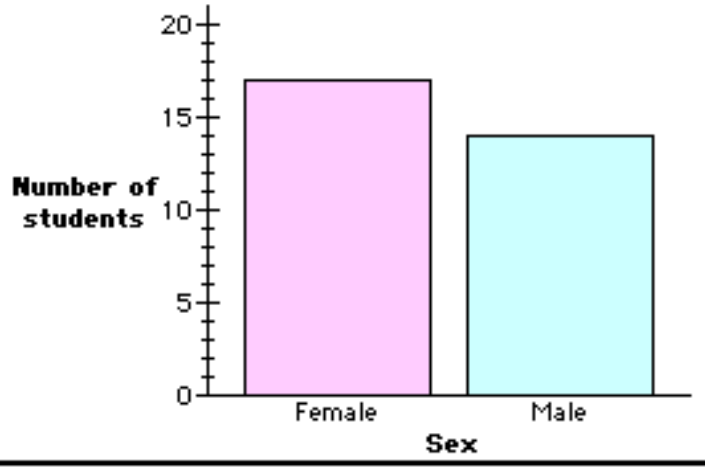


Chart of number of males and females in a given class

Sex	Number of students
Female	17
Male	14

Graph of number of the number of males and females in a given class



NOMINAL DATA

Nominal data divides variables into mutually exclusive, labeled categories.

Examples

Eye color

Blue Brown Green

Smartphone

iPhone Samsung Moto

Transport

Bus Train Car

How is nominal data analyzed?

Descriptive statistics: Frequency distribution and mode

Non-parametric statistical tests

Ordinal Scale

- The values or labels may be **ranked or ordered in some meaningful way**, for example from worst to best.

- Ordinal data may be **verbal** or may be recorded by using **numerical** codes. The differences between the numerical values are not meaningful indicators.

Chart of the number of people per star rating of the movie "Happy Math"

Number of Stars	Number of students
*	4
**	2
***	6
****	15
*****	4

Graph of the number of people per star rating of the movie "Happy Math"

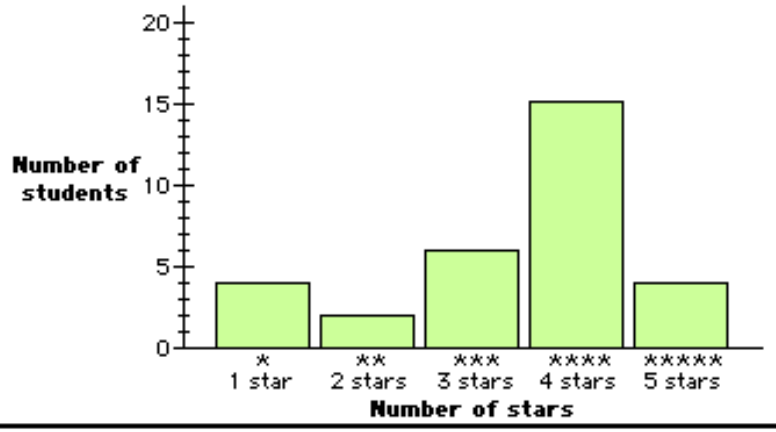
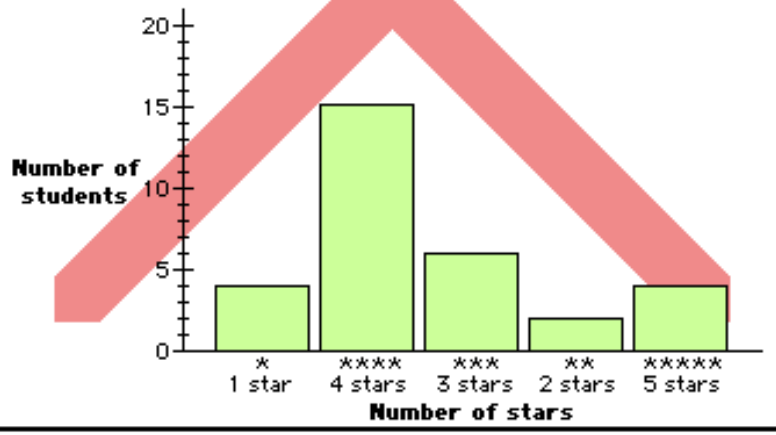


Chart of the number of people per star rating of the movie "Happy Math"

Number of Stars	Number of students
*	4
****	15
***	6
**	2
*****	4

Graph of the number of people per star rating of the movie "Happy Math"



ORDINAL DATA

Ordinal data classifies variables into categories which have a natural order or rank.

Examples

School grades

Education level

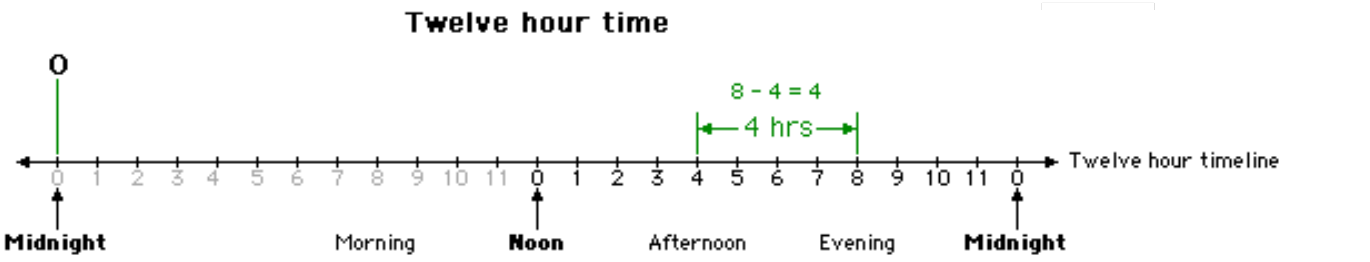
Seniority level

How is ordinal data analyzed?

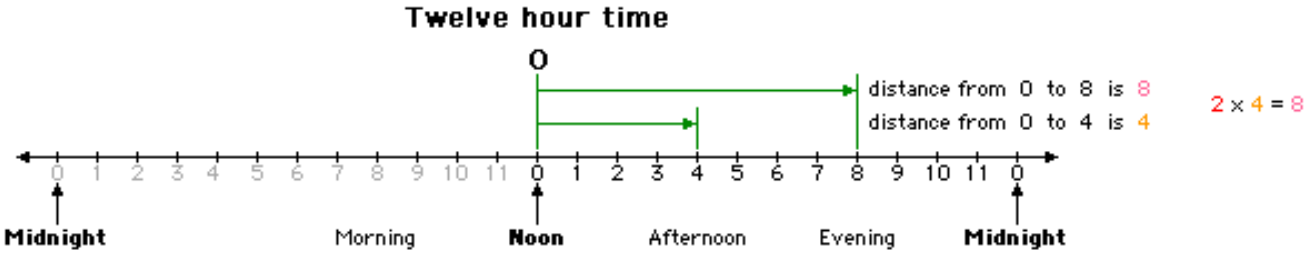
- **Descriptive statistics**: Frequency distribution, mode, median, and range
- **Non-parametric** statistical tests

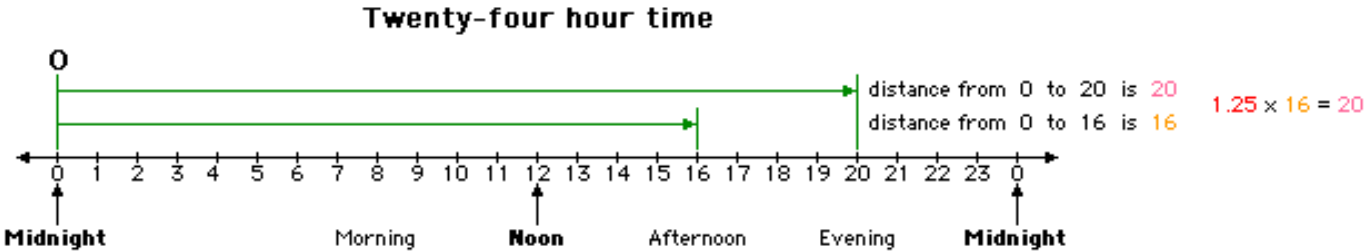
Interval Scale

- We can **assign a meaning to distances between any 2 observations**, but the ratio of 2 different measurements is not a meaningful indicator.
- Always numerical.
- Indicate the differences between the units being measured and these absolute differences are meaningful.
- Not having meaningful zero values.



Ratios have no meaning





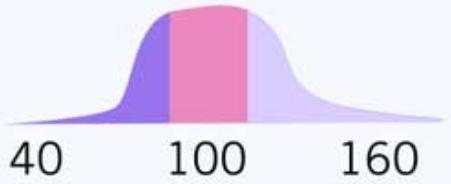
INTERVAL DATA

Interval data is measured along a numerical scale that has equal intervals between adjacent values.

Examples

Temperature

IQ score



Income ranges

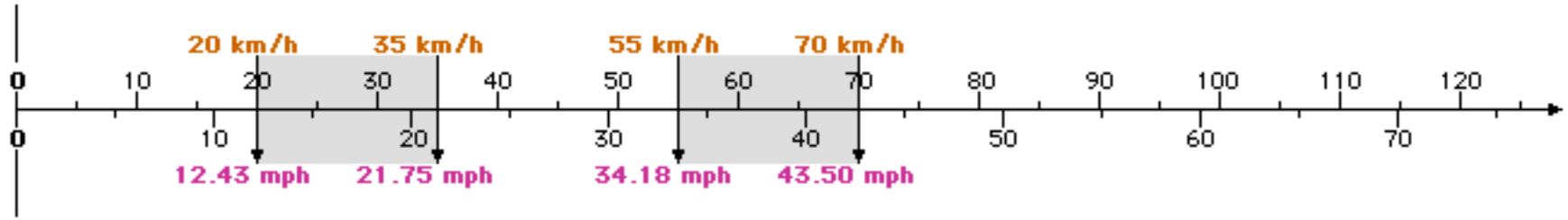
How is interval data analyzed?

Descriptive statistics: Frequency distribution; mode, median, and mean; range, standard deviation, and variance

Parametric statistical tests (e.g. t-test, linear regression)

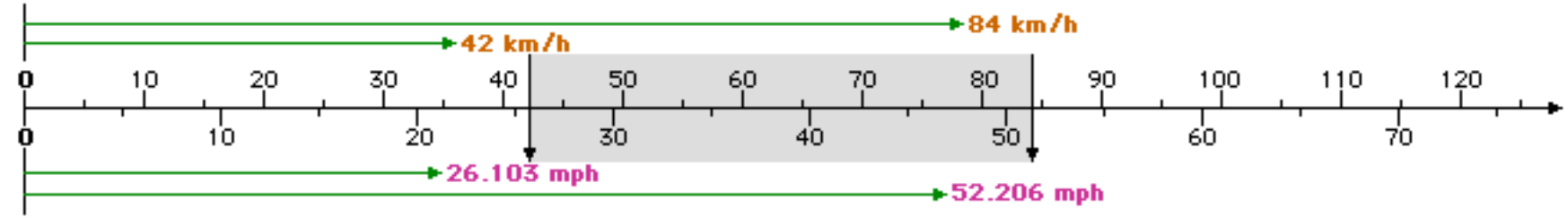
Ratio Scale

- **Distance between pairs of observations, as well as ratios of values, are meaningful.**
- Always numerical.
- Contains a meaningful zero.



$$Speed_{Metric} = \frac{1609\ km}{1000\ miles} Speed_{Imperial}$$

$$Speed_{Imperial} = \frac{1000\ miles}{1609\ km} Speed_{Metric}$$



$$Speed_{Metric} = \frac{1609\ km}{1000\ miles} Speed_{Imperial}$$

$$Speed_{Imperial} = \frac{1000\ miles}{1609\ km} Speed_{Metric}$$

RATIO DATA

Ratio data is measured along a numerical scale that has equal distances between adjacent values, and a true zero.

Examples

Weight in KG

...50 70 90...

Number of staff

...10 30 50...

Income in USD

...20k 40k 60k...

How is ratio data analyzed?

Descriptive statistics: Frequency distribution; mode, median, and mean; range, standard deviation, variance, and coefficient of variation

Parametric statistical tests (e.g. ANOVA, linear regression)

Data Types according to the time of collection

Cross-sectional Data

- Data collected at the same or nearly the same point in time.
- Example: Number of UBINUS students in the academic year 2005/2006, Number of publicly listed companies in 2006.

Time Series Data

- Data was collected over a certain period.
- Example: The movement of the rupiah exchange rate within 1 month, Indonesia's rice production from 1997 to 2006.

Categorize this variable as nominal, interval, ratio, or ordinal !!

Hand phones' brand

Human Height

GPA

Kind of jobs

Numbers of computer selling

Students' score

Education level

Weight of cow

Human IQ

Number of tomato

Exercise

- A bar owner lists the types of beer he sells and records the number of bottles of each brand he sold last week. Suppose the four brands of beer are: Budweiser, Stroh's, Miller, and Coor's. Explain how to create a numerical code to represent the brands of beer. **Does this code have any special meaning, or could the code values be assigned randomly?**
- In a taste test, 20 individuals are asked to taste four different diet colas and to rate each cola as poor, fair, good, or excellent. Are the results of this **taste test nominal data, ordinal data, interval data, or ratio data**? Explain.

1 Kelompok = 4 orang
Minggu depan

Exercise

- Categorize each of the following variables as **nominal, ordinal, interval, or ratio**:
  - For computer data entry purposes, a garden supply store classifies flowers as follows: 1 = roses, 2 = tulips, 3 = marigolds, 4 = gardenias.
  - A customer classifies her preferences in flowers from worst to best as follows: 1= roses, 2 = tulips, 3 = marigolds, 4 = gardenias.
  - A garden supply store lists its total sales of flowers as follows: 100 roses, 25 tulips, 37 marigolds, and 49 gardenias.

1 Kelompok = 4 orang
Minggu depan

Exercise

- Below you are given financial information about a sample of companies for July 11, 2001.

Company	Price ($) per share	Price/Earnings Ratio	Annual Dividend ($) per share	Sector
A	18	12.6	0.36	services
B	10	18.2	0.12	basic materials
C	13	39.5	0	technology
D	84	18.6	1.20	financial
E	14	48.2	0	healthcare
F	28	23.6	0.08	technology
G	37	18.6	0.05	healthcare
H	22	23.3	0.30	consumer-noncyclical
I	28	17.5	1.00	consumer-cyclical

1 Kelompok = 4 orang
Minggu depan

a. How many variables are in the data set?
b. Which variables are qualitative?
c. Which variables are quantitative?
d. Are the data cross-sectional or time series?
e. For each of the variables above, give the measurement scale used

Teamwork Assignment

Carilah (min 10) :

1. Variabel kualitatif
2. Variabel kuantitatif – diskrit
3. Variabel kuantitatif – kontinyu
4. Variabel dengan skala pengukuran nominal
5. Variabel dengan skala pengukuran ordinal
6. Variabel dengan skala pengukuran interval
7. Variabel dengan skala pengukuran ratio

1 Kelompok = 4 orang
Minggu depan