Statistical Methods
統計方法

SEPTEMBER 5, 2023

I-CHEN LEE

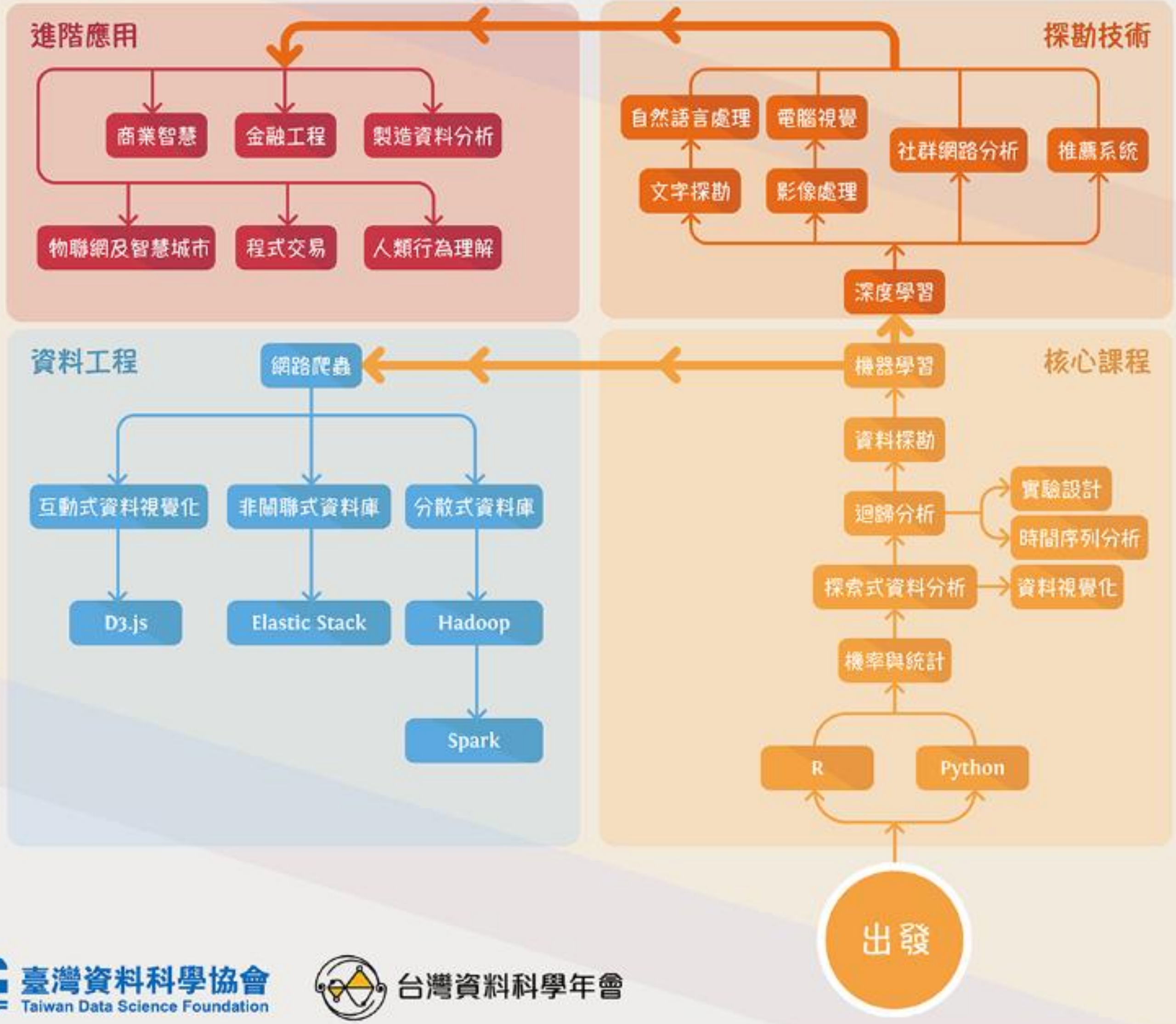
Related topics

Lots of statistical methods

- Basic statistical procedures
- Basic modeling and statistical inference
- Methods for comparison (t-test, F-test, correlation, ...)
- Methods of estimation (LSE, MLE, BE, MAP)
- Modeling (linear, nolinear, fixed/random effects, GLM)
- Multivariate analysis (PCA, FA, LDA)
- Non-regular data type (lifetime data, mixed model)
- Clustering method
- ...

Software: R, SPSS, Latex

資料科學學習地圖



進階應用

商業智慧

金融工程

製造資料分析

物聯網及智慧城市

程式交易

人類行為理解

探勘技術

自然語言處理

電腦視覺

社群網路分析

推薦系統

文字探勘

影像處理

深度學習

資料工程

網路爬蟲

互動式資料視覺化

非關聯式資料庫

分散式資料庫

D3.js

Elastic Stack

Hadoop

Spark

核心課程

機器學習

資料探勘

迴歸分析

實驗設計

時間序列分析

探索式資料分析

資料視覺化

機率與統計

R

Python

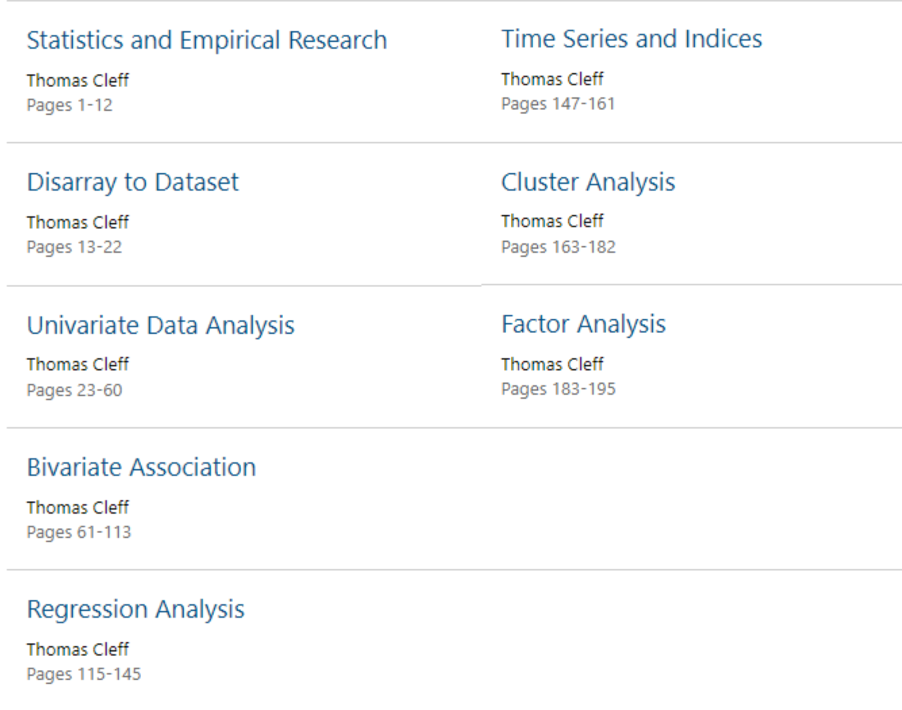
出發

臺灣資料科學協會
TDSF Taiwan Data Science Foundation

台灣資料科學年會

Reference Textbook

1. Cleff, T. (2014). ***Exploratory Data Analysis in Business and Economics***. Springer Cham
2. Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). ***An introduction to statistical learning: with applications in R***. Spinger.



R source

Akinkunmi, M. (2019). Introduction to statistics using R. *Synthesis Lectures on Mathematics and Statistics*, *11*(4), 1-235.

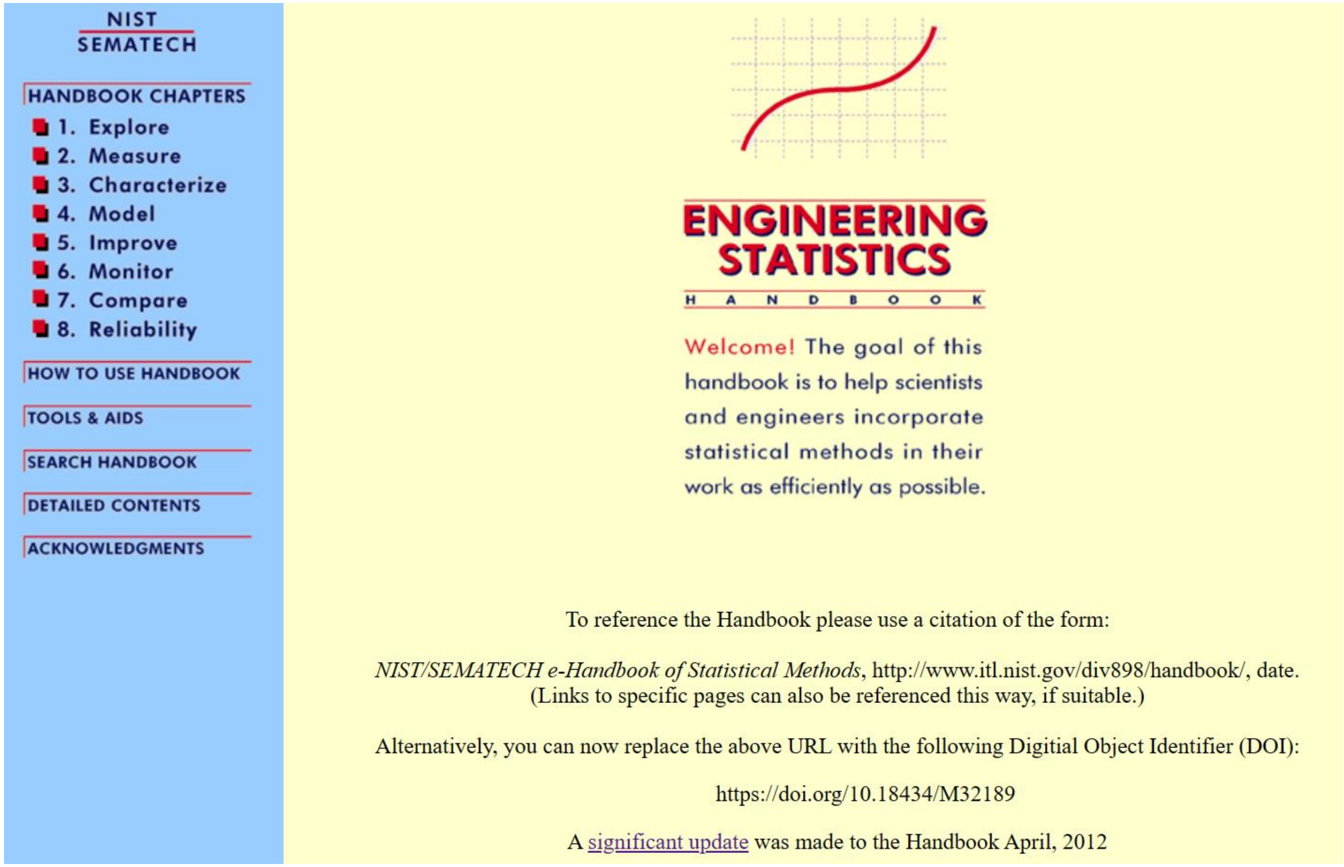
1. Download R: https://cran.csie.ntu.edu.tw/
2. Download Rstudio: https://www.rstudio.com/products/rstudio/

Additional links:

1. https://cran.r-project.org/index.html
2. https://modernstatisticswithr.com/index.html
3. https://smac-group.github.io/ds/section-data.html
4. https://cran.r-project.org/web/packages/HSAUR/vignettes/Ch_introduction_to_R.pdf

Reference

Engineering Statistics
https://www.itl.nist.gov/div898/handbook/index.htm



ENGINEERING STATISTICS HANDBOOK

Welcome! The goal of this handbook is to help scientists and engineers incorporate statistical methods in their work as efficiently as possible.

To reference the Handbook please use a citation of the form:

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, date.
(Links to specific pages can also be referenced this way, if suitable.)

Alternatively, you can now replace the above URL with the following Digitial Object Identifier (DOI):

https://doi.org/10.18434/M32189

A significant update was made to the Handbook April, 2012

Have you learned?

Basic statistical concept

- ✓ Statistical graphics
- ✓ Descriptive statistics

Regression analysis

- ✓ Response=*f*(variables)+error

Design of experiments (DOE)

- ✓ Comparative experiment

Multivariate analysis

- ✓ Principle component analysis (PCA)

Procedure for statistical analysis

1. **Recognition of & statement of problem**
2. Choice of factors, levels, and ranges
3. Selection of the response variable(s)
4. Choice of methodology
5. Statistical analysis
6. Drawing conclusions, recommendations

1.2 Two Types of Statistics

- Two terms: *descriptive statistics* and *inductive data analysis*
- The term *descriptive statistics* refers to all techniques used to obtain information based on the description of data from a population.
- The now common form of *inductive data analysis* was developed in which one attempts to draw conclusions about a total population based on a sample.

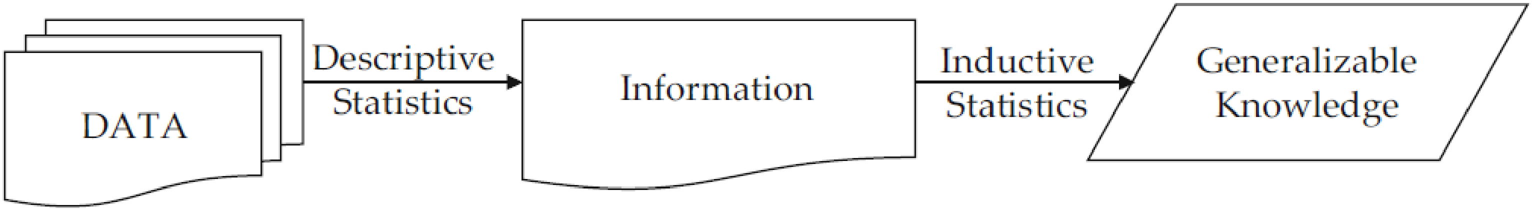


Fig. 1.1 Data begets information, which in turn begets knowledge

Cleff, T. (2014). ***Exploratory Data Analysis in Business and Economics***. Springer Cham

1.3 The Generation of Knowledge Through Statistics

- At this stage, our researcher will ask himself whether the insights obtained on the basis of this partial sample.
- Insights which he expected beforehand can be viewed as representative of the entire population.
- Generalizable information in *descriptive statistics* is always initially speculative.
- With the aid of *inductive statistical techniques*, one can estimate the *error probability* associated with applying insights obtained through descriptive statistics to an overall population.
- The researcher must decide for himself which level of error probability renders the insights insufficiently qualified and inapplicable to the overall population.

From Models to Business Intelligence

Raw data are gathered and transformed into information with strategic relevance by means of descriptive assessment methods

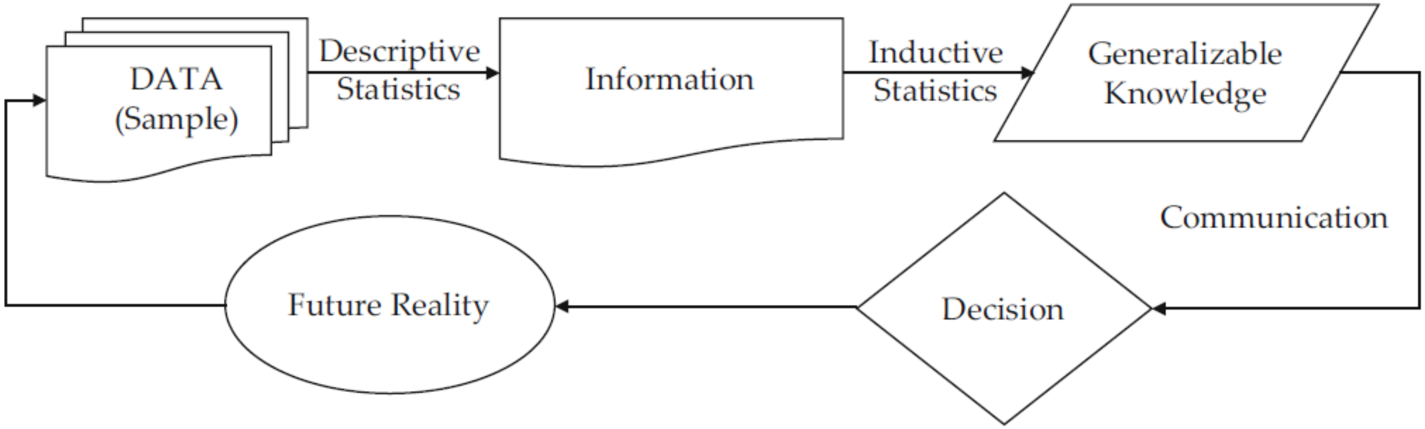
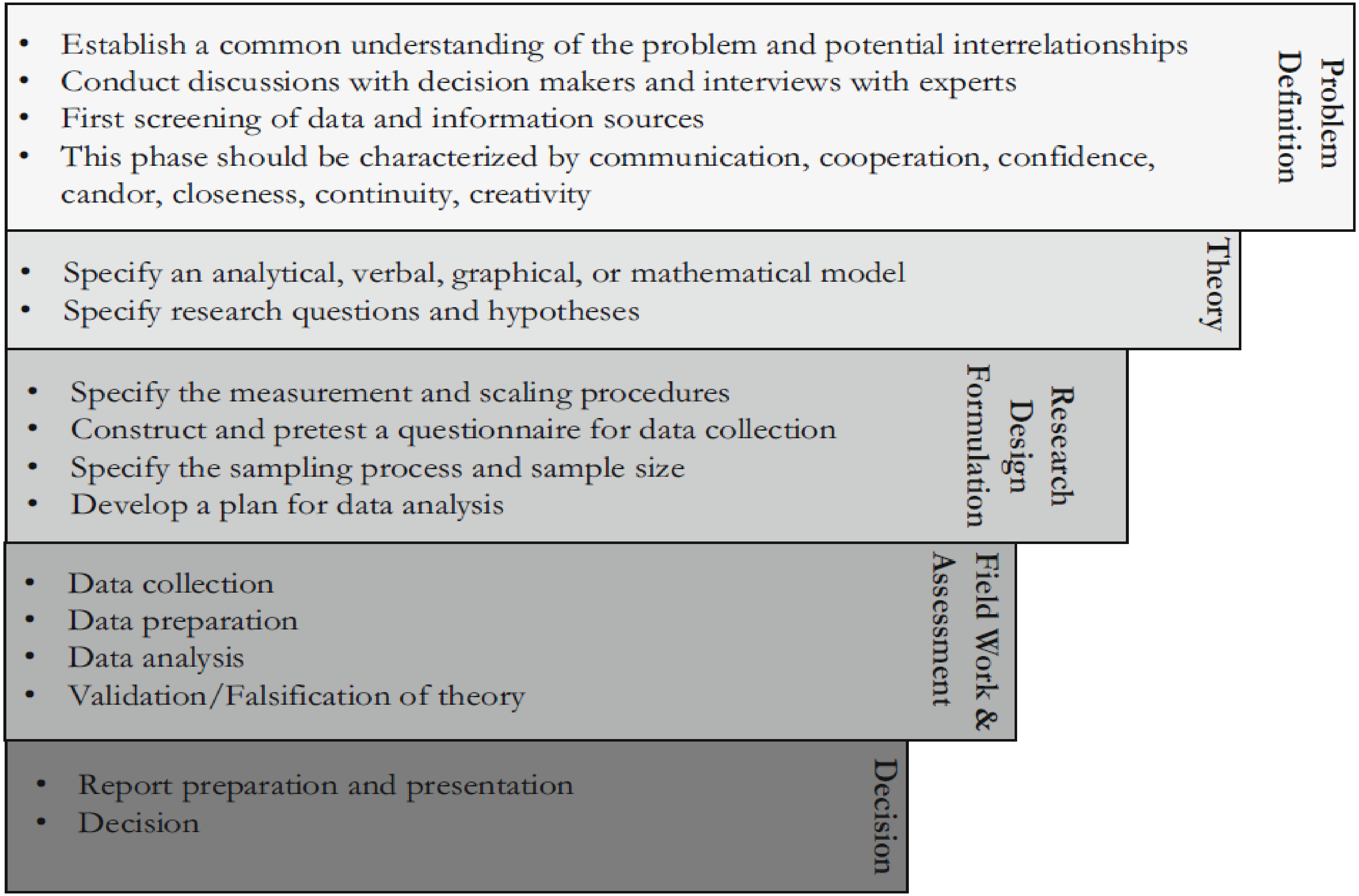


Fig. 1.6 The intelligence cycle (Source: Own graphic, adapted from Harkleroad 1996, p. 45)

Cleff, T. (2014). ***Exploratory Data Analysis in Business and Economics***. Springer Cham

Check



Problem Definition

- Establish a common understanding of the problem and potential interrelationships
- Conduct discussions with decision makers and interviews with experts
- First screening of data and information sources
- This phase should be characterized by communication, cooperation, confidence, candor, closeness, continuity, creativity

Theory

- Specify an analytical, verbal, graphical, or mathematical model
- Specify research questions and hypotheses

Research Design Formulation

- Specify the measurement and scaling procedures
- Construct and pretest a questionnaire for data collection
- Specify the sampling process and sample size
- Develop a plan for data analysis

Field Work & Assessment

- Data collection
- Data preparation
- Data analysis
- Validation/Falsification of theory

Decision

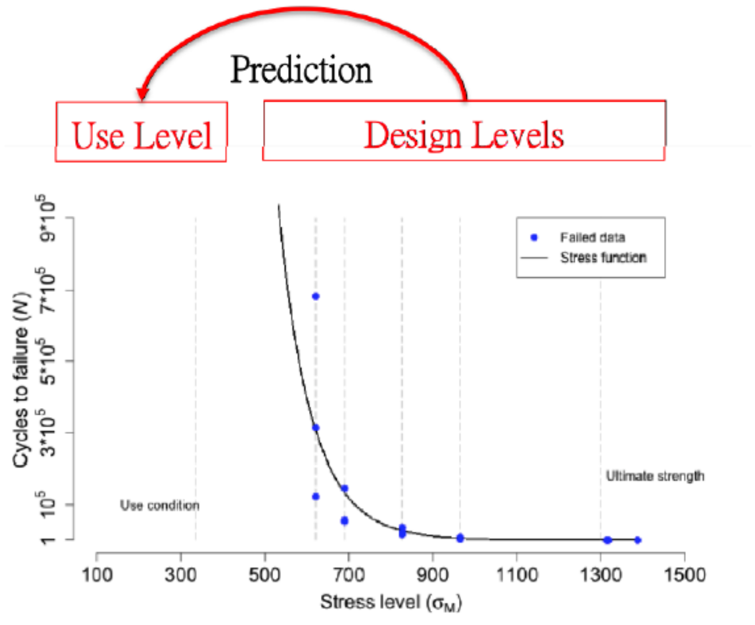
- Report preparation and presentation
- Decision

Cleff, T. (2014). ***Exploratory Data Analysis in Business and Economics.*** Springer Cham

Check: Problem Definition

- Establish a common understanding of the problem and potential interrelationships
- Conduct discussions with decision makers and interviews with experts
- First screening of data and information sources
- This phase should be characterized by communication, cooperation, confidence, candor, closeness, continuity, creativity

Problem Definition



(b) Polymer Composite Material

Problem Definition:

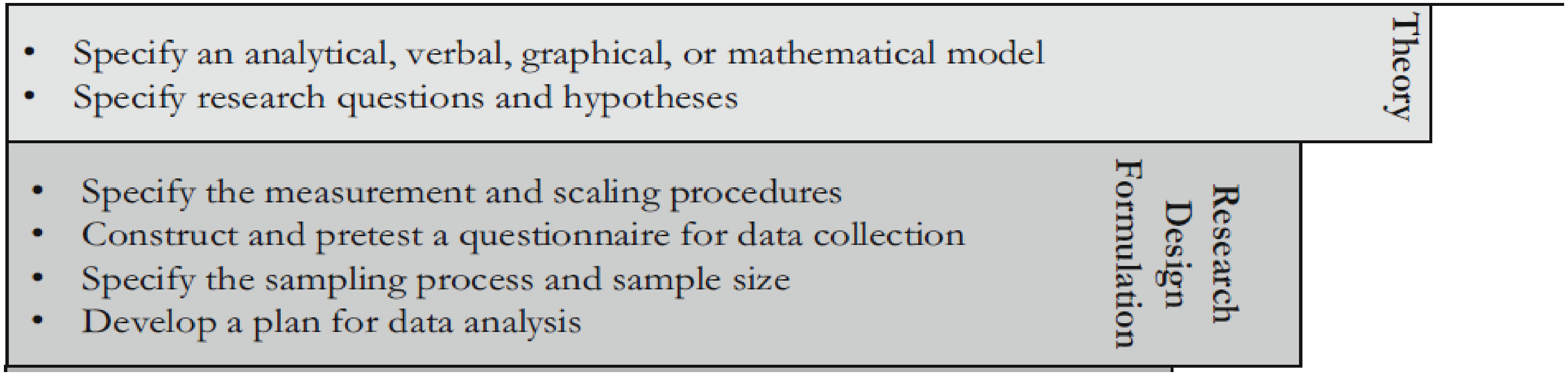
- How to design an experiment so that I can obtain a more precise S-N curve?
- The S-N curve is a non-linear stress function, which is consulted from the material experts.
- The blue points is the illustrative points.

✓ Keyword I: precise non-linear fitting

✓ Keyword II: design an experiment

Cleff, T. (2014). ***Exploratory Data Analysis in Business and Economics***. Springer Cham

Check: Theory and Design



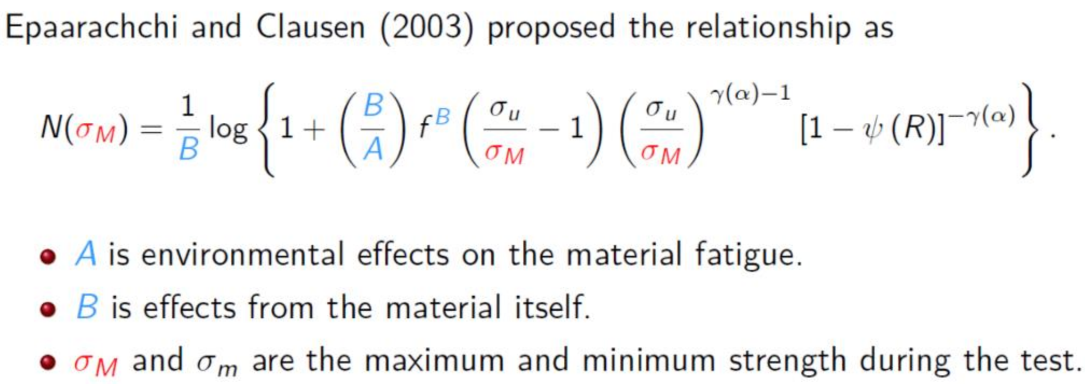
- Specify an analytical, verbal, graphical, or mathematical model
- Specify research questions and hypotheses

Theory

- Specify the measurement and scaling procedures
- Construct and pretest a questionnaire for data collection
- Specify the sampling process and sample size
- Develop a plan for data analysis

Research Design Formulation

S-N curve



Epaarachchi and Clausen (2003) proposed the relationship as

$$N(\sigma_M) = \frac{1}{B}\log\left\{1+\left(\frac{B}{A}\right) f^B \left(\frac{\sigma_u}{\sigma_M}-1\right)\left(\frac{\sigma_u}{\sigma_M}\right)^{\gamma(\alpha)-1}[1-\psi(R)]^{-\gamma(\alpha)}\right\}.$$

- A is environmental effects on the material fatigue.
- B is effects from the material itself.
- σ_M and σ_m are the maximum and minimum strength during the test.

Define an index or objective function for good fitting.

Maximize the objective function to design an experiment.

Verify the experiment is exactly the best one.

Cleff, T. (2014). ***Exploratory Data Analysis in Business and Economics.*** Springer Cham

Exploratory Data Analysis (EDA)

- EDA is an approach to data analysis that postpones the assumptions about **what kind of model the data follow with**.
- The more direct approach of **allows the data itself to reveal its underlying structure and model**.

Objectives:

- ✓ Maximize insight into a dataset
- ✓ Extract important variables
- ✓ Detect outliers
- ✓ Test underlying assumptions
- ✓ Develop models and determine optimal factor settings

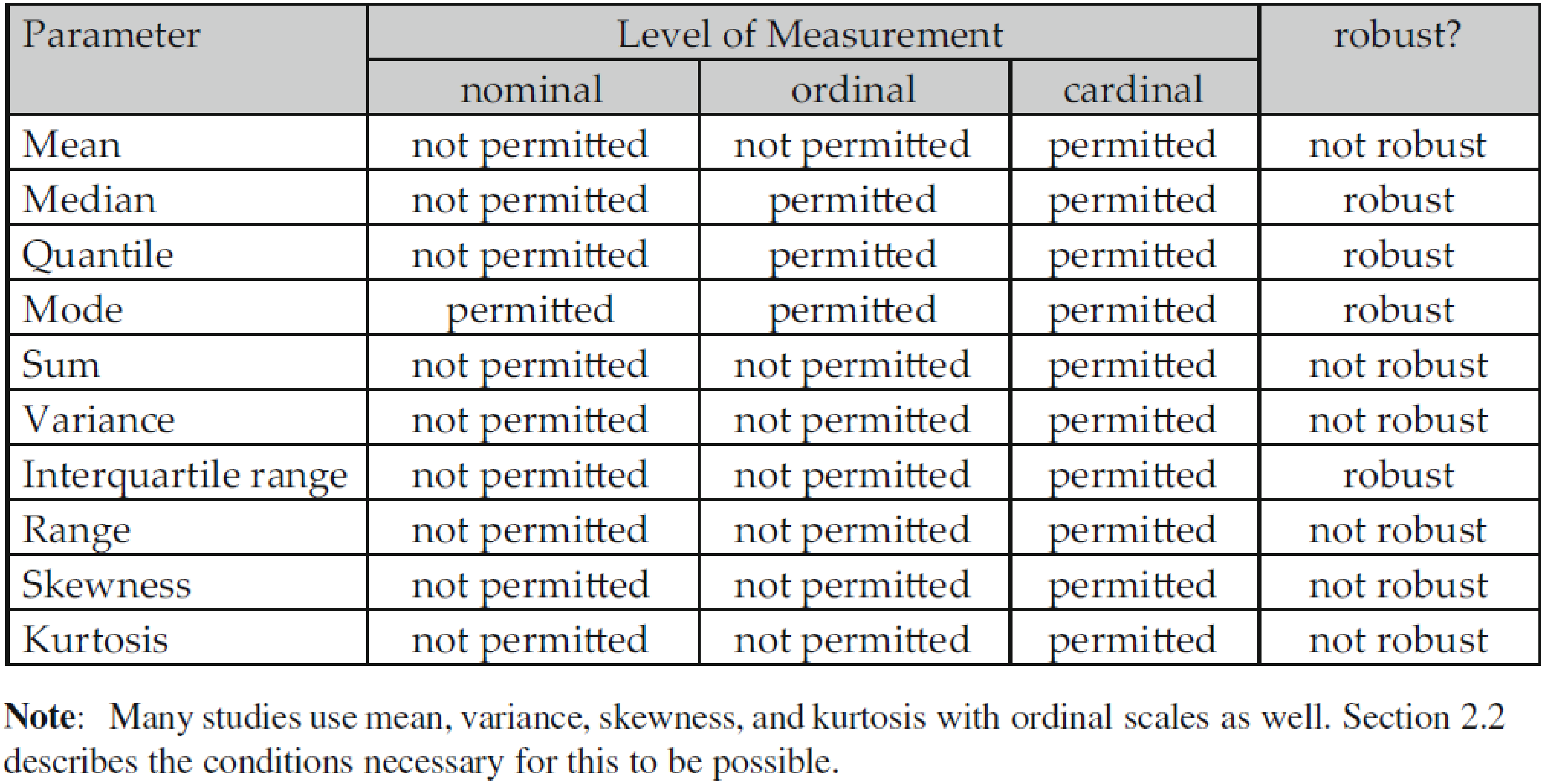
Reference:
https://www.itl.nist.gov/div898/handbook/eda/eda.htm
https://www.itl.nist.gov/div898/handbook/

Data Preprocessing : Concepts

1. **Descriptive Statistics**
   - mean
   - Standard deviation
   - Quantile
   - Frequency table
   - correlation

2. **Graphics**
   - Box plot
   - Distribution plot
   - Correlation plot
   - Comparison plot
   - Trend plot
   - …

Assignment (Section 3) Descriptive Statistics (p.52)

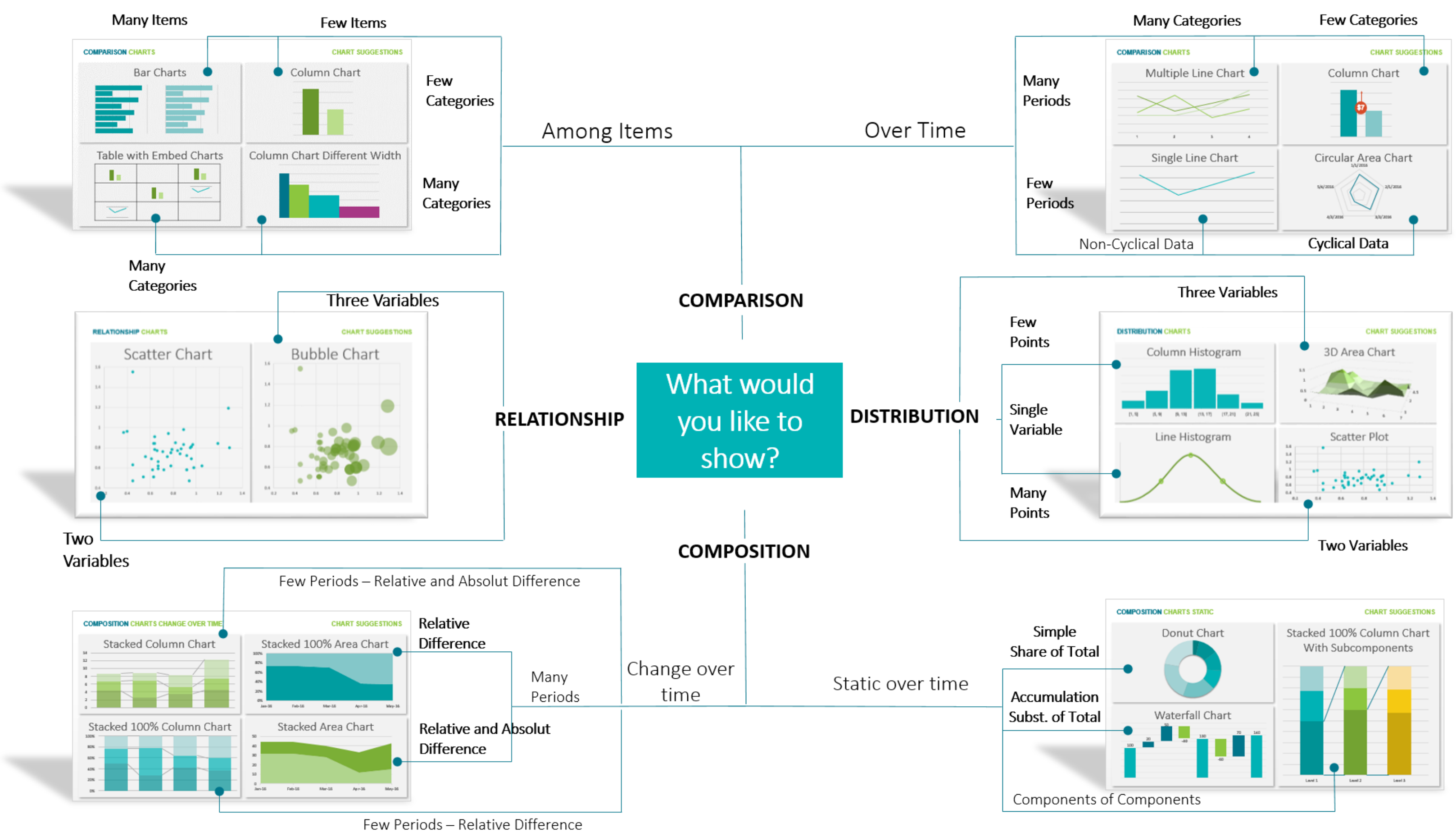


Parameter	Level of Measurement			robust?
	nominal	ordinal	cardinal	
Mean	not permitted	not permitted	permitted	not robust
Median	not permitted	permitted	permitted	robust
Quantile	not permitted	permitted	permitted	robust
Mode	permitted	permitted	permitted	robust
Sum	not permitted	not permitted	permitted	not robust
Variance	not permitted	not permitted	permitted	not robust
Interquartile range	not permitted	not permitted	permitted	robust
Range	not permitted	not permitted	permitted	not robust
Skewness	not permitted	not permitted	permitted	not robust
Kurtosis	not permitted	not permitted	permitted	not robust

Note: Many studies use mean, variance, skewness, and kurtosis with ordinal scales as well. Section 2.2 describes the conditions necessary for this to be possible.

Cleff, T. (2014). ***Exploratory Data Analysis in Business and Economics.*** Springer Cham

Graphics



What would you like to show?

COMPARISON

Among Items

Over Time

Many Items

Few Items

COMPARISON CHARTS

CHART SUGGESTIONS

Bar Charts

Column Chart

Table with Embed Charts

Column Chart Different Width

Few Categories

Many Categories

Many Categories

Many Categories

Few Categories

Many Periods

Few Periods

Multiple Line Chart

Column Chart

Single Line Chart

Circular Area Chart

Non-Cyclical Data

Cyclical Data

RELATIONSHIP

RELATIONSHIP CHARTS

CHART SUGGESTIONS

Scatter Chart

Bubble Chart

Three Variables

Two Variables

DISTRIBUTION

DISTRIBUTION CHARTS

CHART SUGGESTIONS

Three Variables

Few Points

Single Variable

Many Points

Column Histogram

3D Area Chart

Line Histogram

Scatter Plot

Two Variables

COMPOSITION

Change over time

Static over time

Few Periods – Relative and Absolut Difference

COMPOSITION CHARTS CHANGE OVER TIME

CHART SUGGESTIONS

Stacked Column Chart

Stacked 100% Area Chart

Stacked 100% Column Chart

Stacked Area Chart

Relative Difference

Many Periods

Relative and Absolut Difference

Few Periods – Relative Difference

COMPOSITION CHARTS STATIC

CHART SUGGESTIONS

Simple Share of Total

Accumulation Subst. of Total

Components of Components

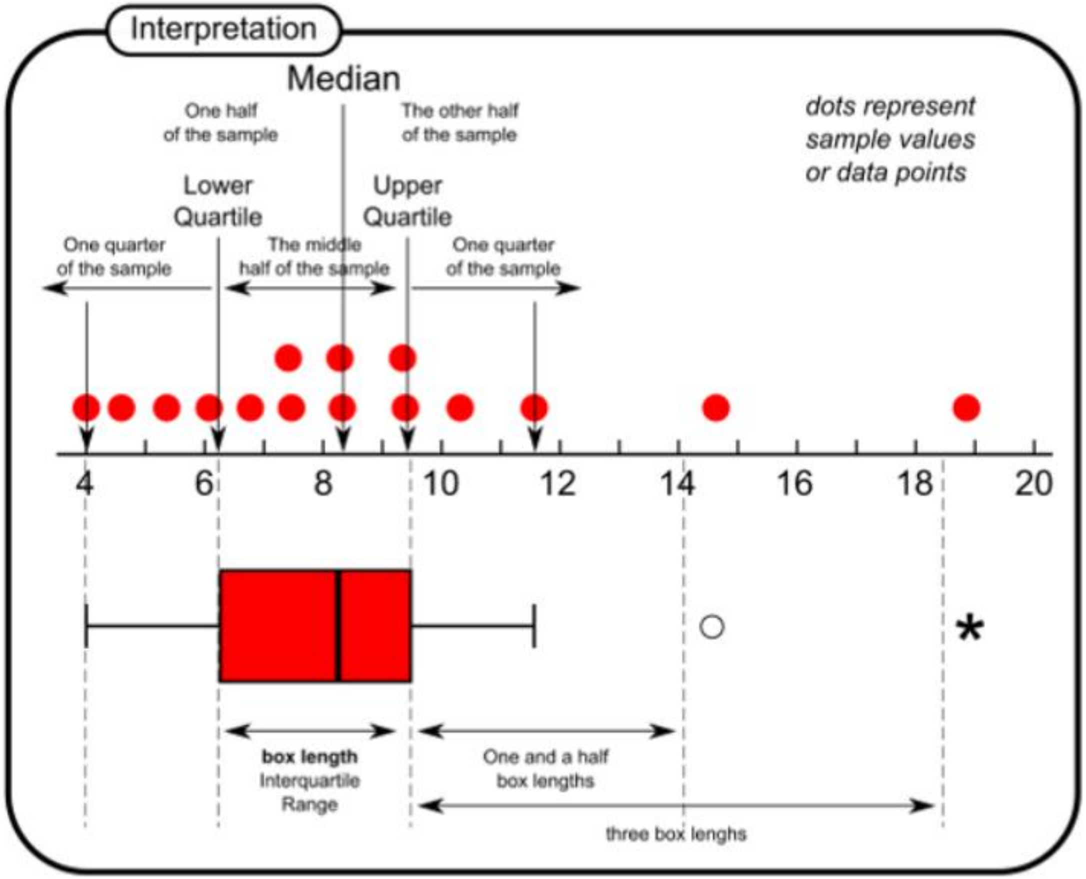
Donut Chart

Waterfall Chart

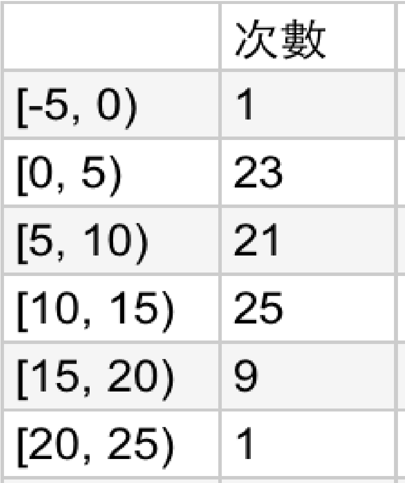
Stacked 100% Column Chart With Subcomponents

Detect outlier (abnormal point)

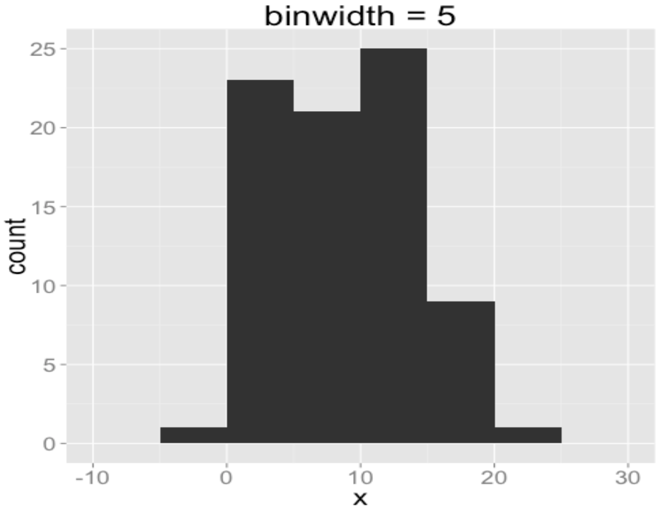
- Boxplot


Interpretation
Median
One half of the sample
The other half of the sample
dots represent sample values or data points
Lower Quartile
Upper Quartile
One quarter of the sample
The middle half of the sample
One quarter of the sample
4
6
8
10
12
14
16
18
20
*
box length Interquartile Range
One and a half box lengths
three box lenghs

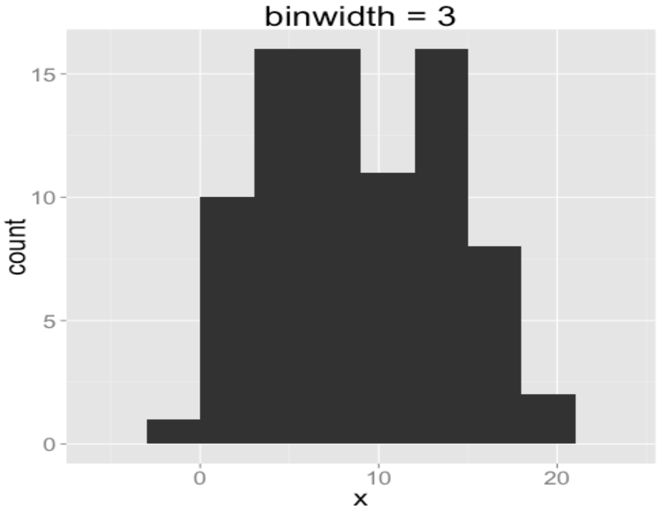
Distribution plot: histogram



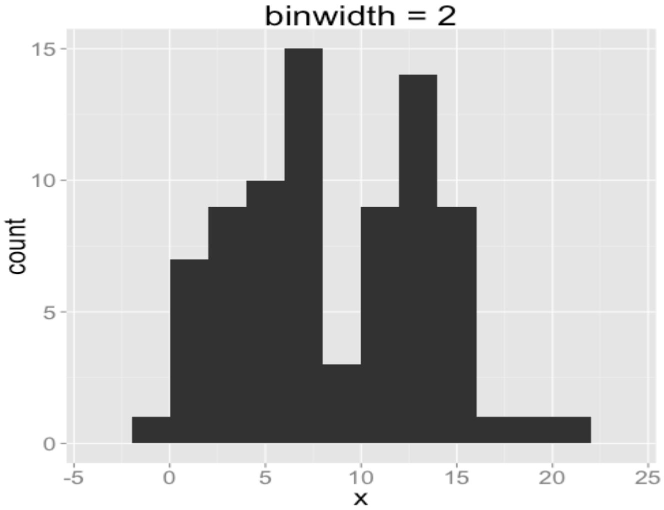
	次數
[-5, 0)	1
[0, 5)	23
[5, 10)	21
[10, 15)	25
[15, 20)	9
[20, 25)	1



binwidth = 5

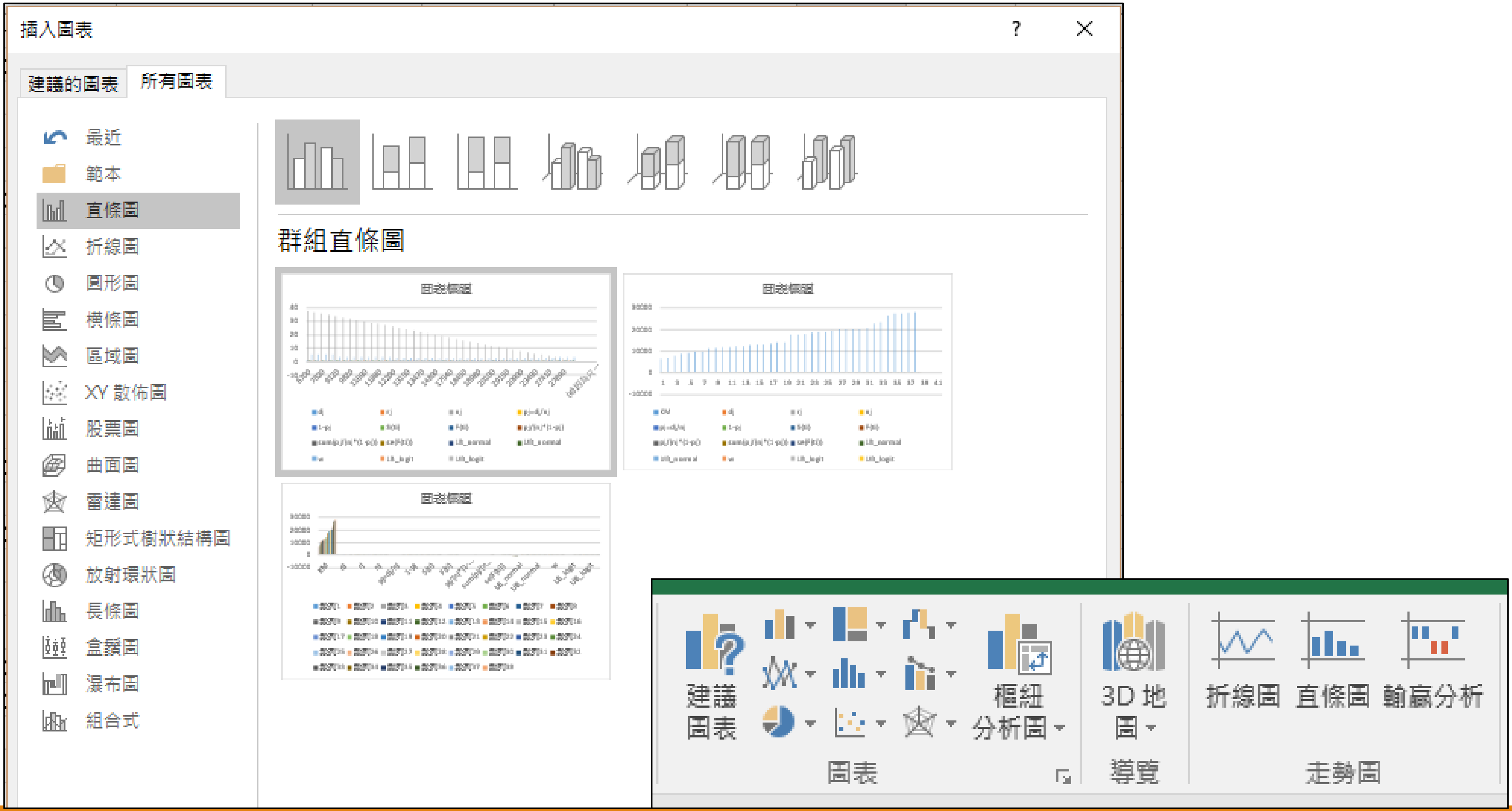


binwidth = 3

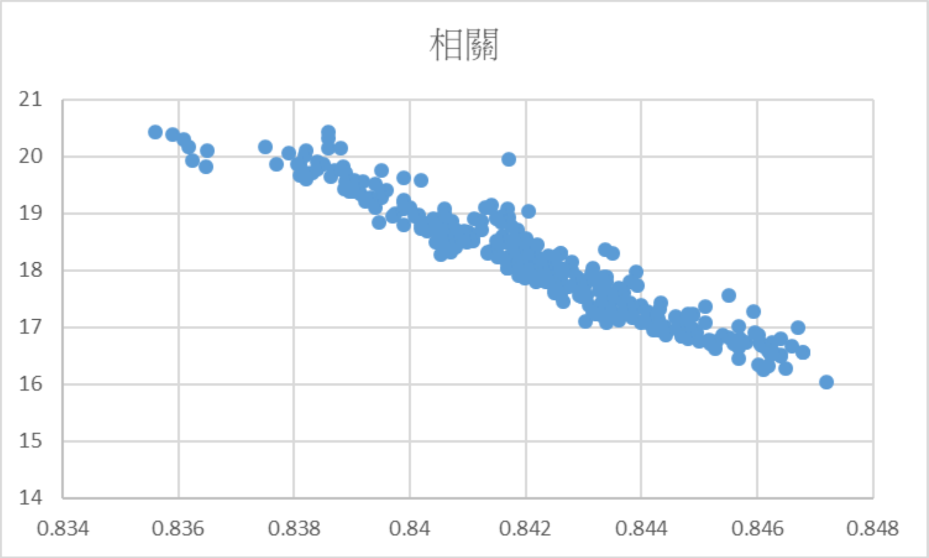


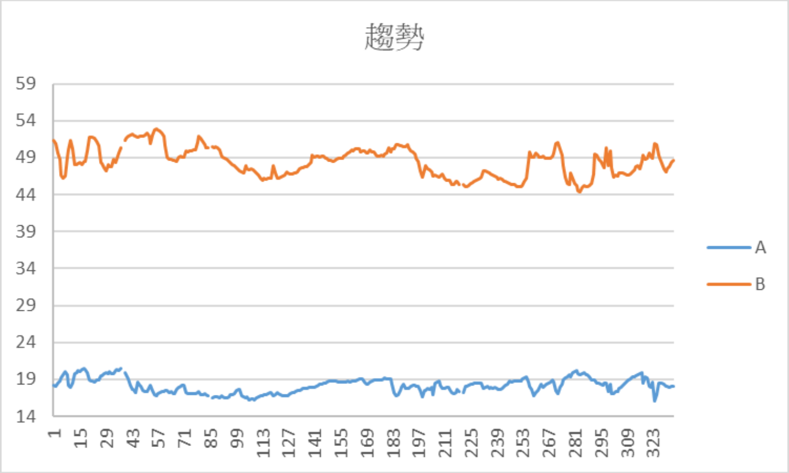
binwidth = 2

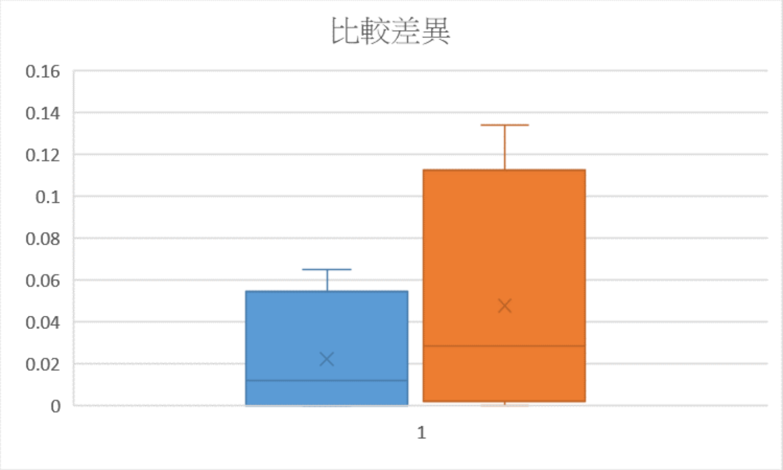
Graphics in Excel



Graphical meaning

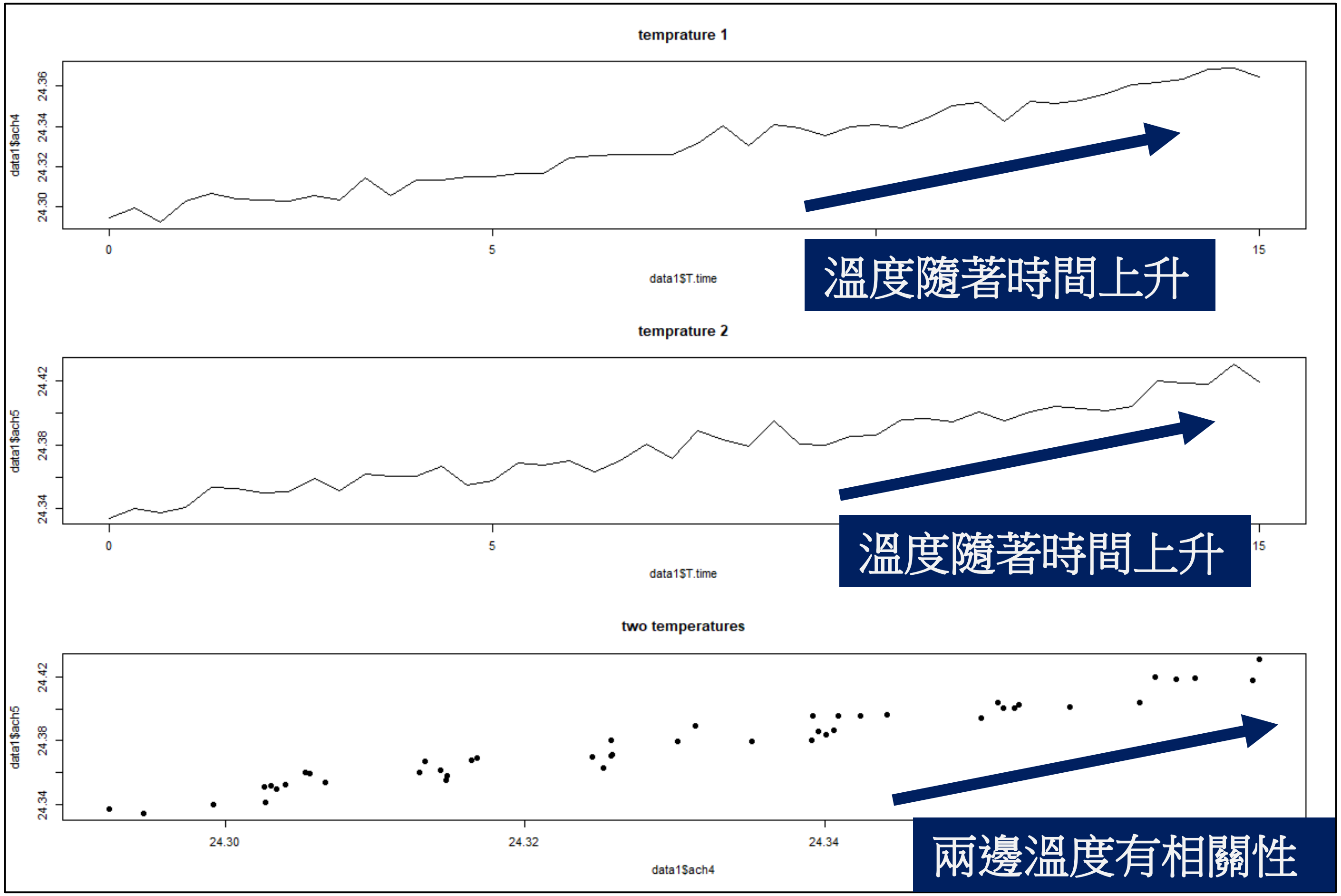






Graphical meaning

Shape / trend / period

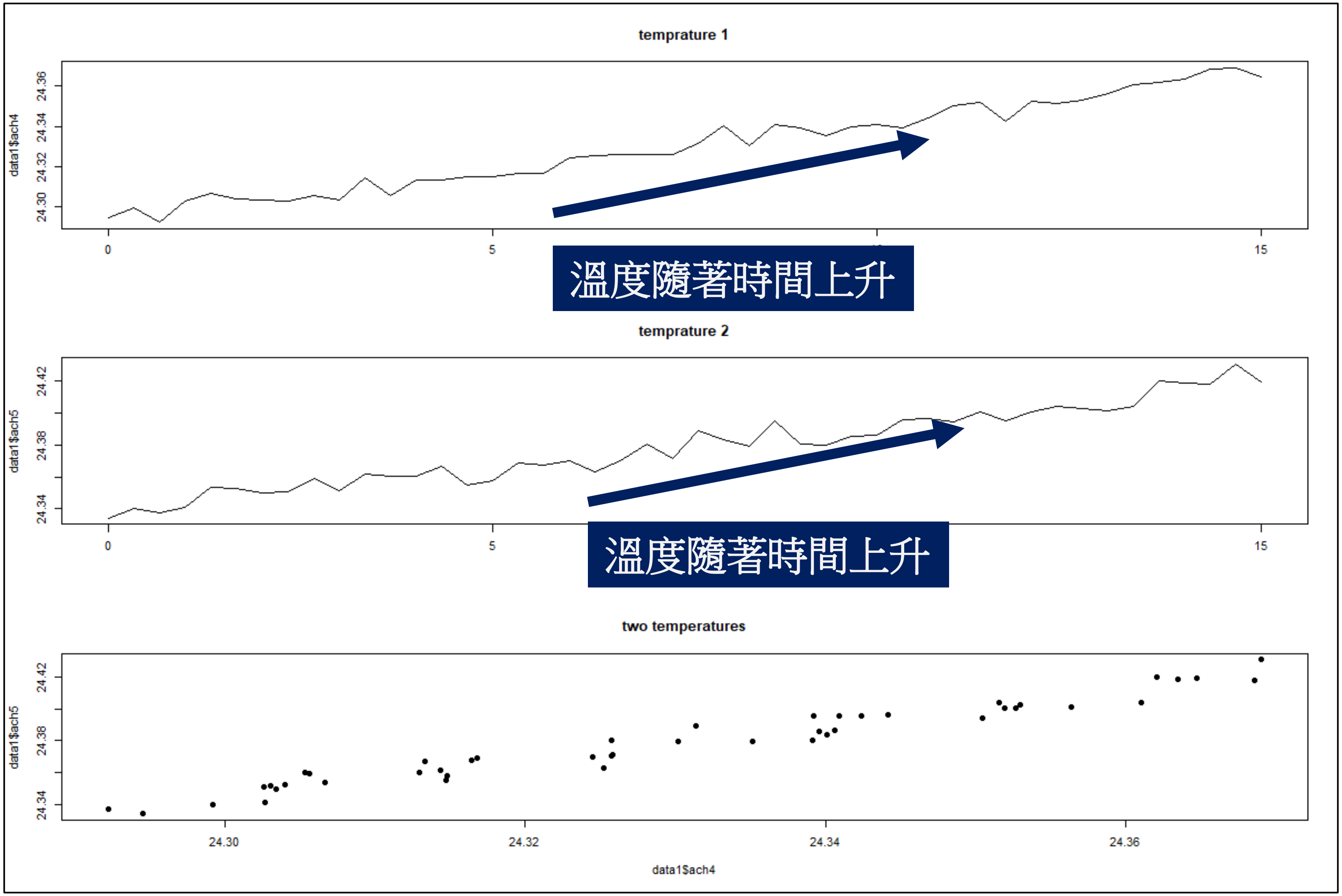


溫度隨著時間上升

溫度隨著時間上升

兩邊溫度有相關性

Relationship?

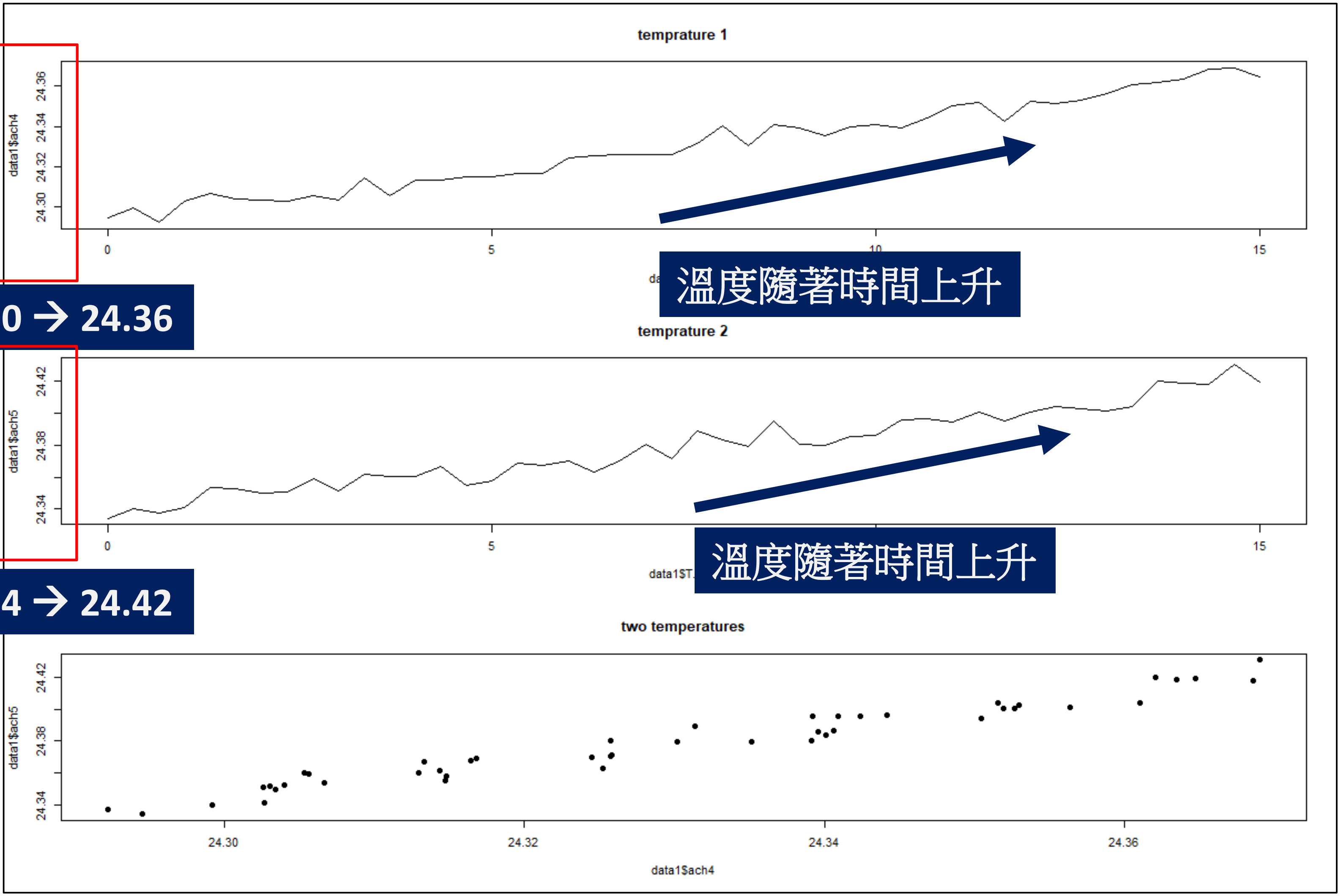


溫度隨著時間上升

溫度隨著時間上升

Q：隨著操作時間越長，溫度有明顯的增加？

Relationship?



溫度隨著時間上升

24.30 → 24.36

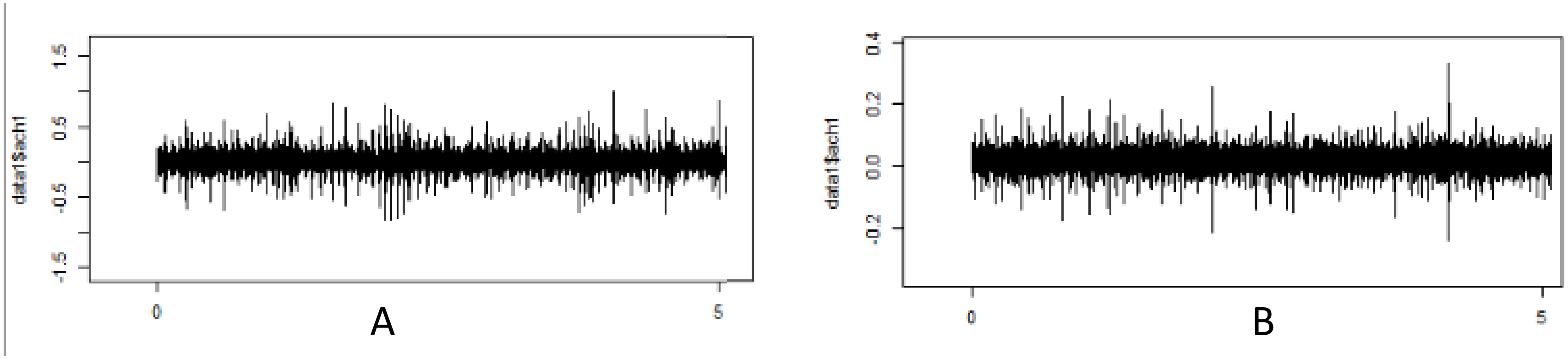
溫度隨著時間上升

24.34 → 24.42

Q：隨著操作時間越長，溫度有明顯的增加？

Difference?

已知：A為異常的機器，B為正常的機器

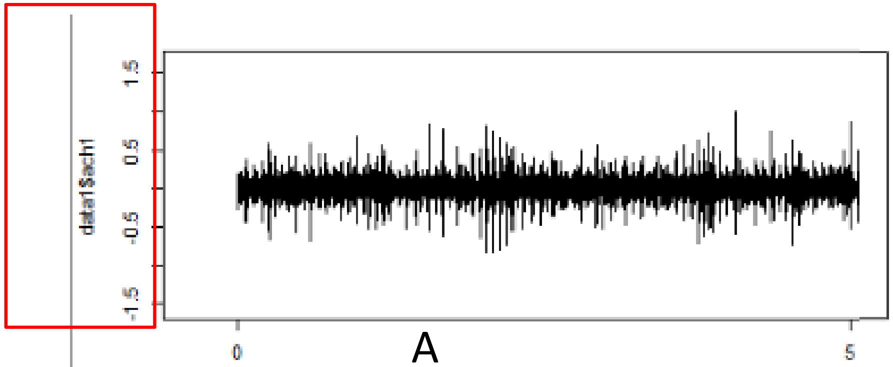


結論：A 和 B 兩者沒有差別。

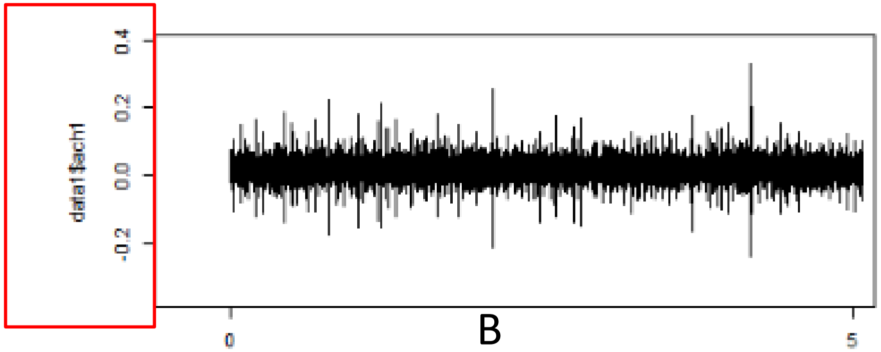
???

Difference?

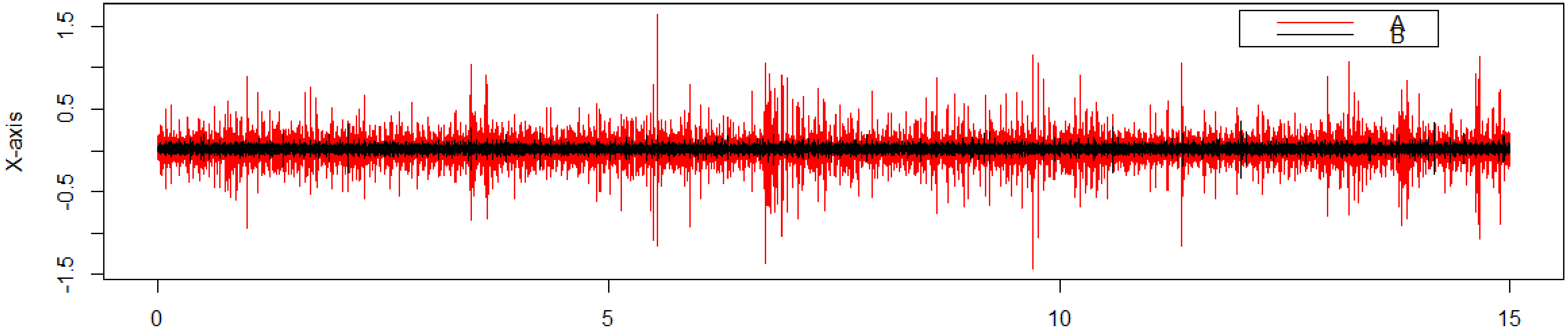
已知：A為異常的機器，B為正常的機器



-1.5 ~ 1.5



-0.2 ~ 0.4



結論：此變數在A 和 B 之間可能有差別。

Reference

Online reference

- ✓ 資料科學領域線上課程大彙整 https://taweihuang.hpd.io/2016/11/12/
- ✓ 工程統計 (*e-Handbook of Statistical Methods*) https://www.itl.nist.gov/div898/handbook/index.htm

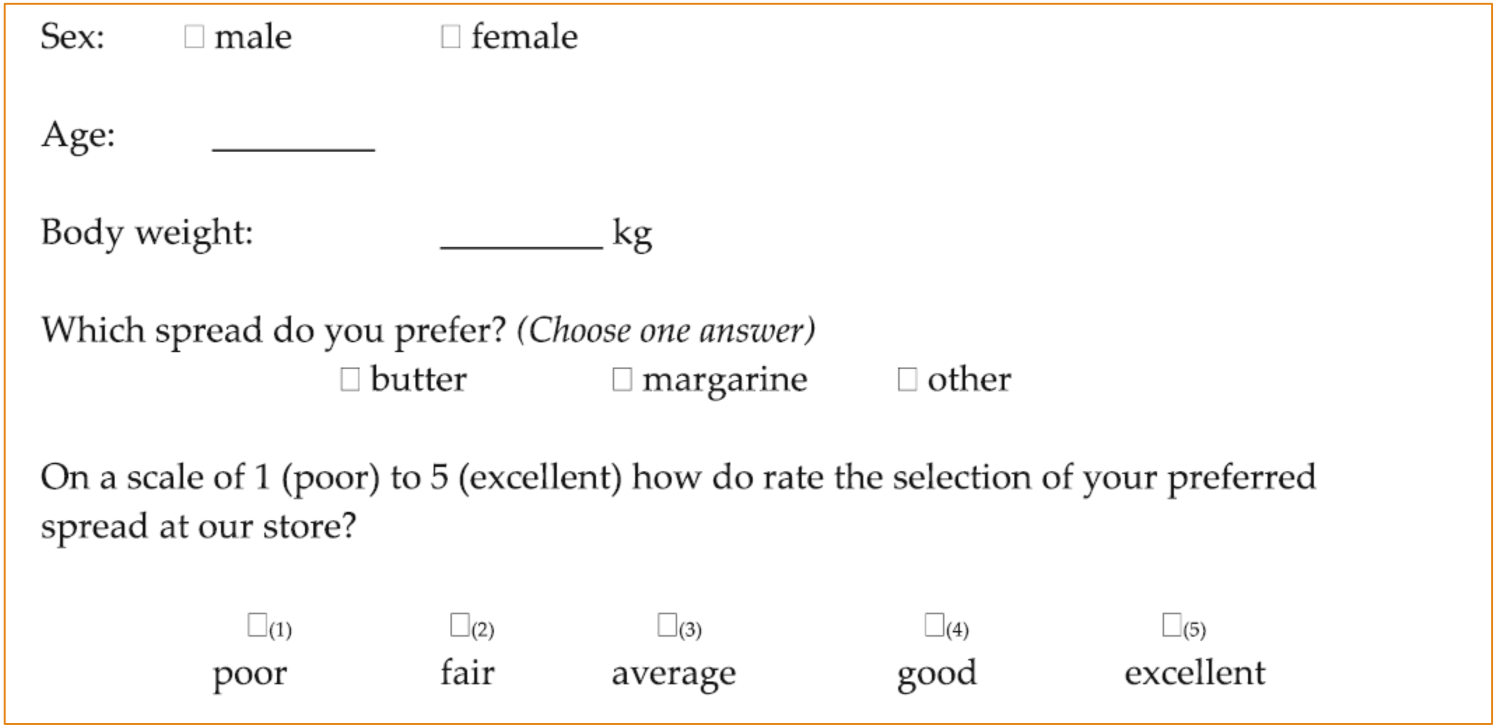
ENGINEERING STATISTICS

HANDBOOK

Welcome! The goal of this handbook is to help scientists and engineers incorporate statistical methods in their work as efficiently as possible.

Level of Measurement (Section 2)

Example: collected questionnaires from 850 customers



Sex: □ male □ female

Age: ________

Body weight: ________ kg

Which spread do you prefer? *(Choose one answer)*

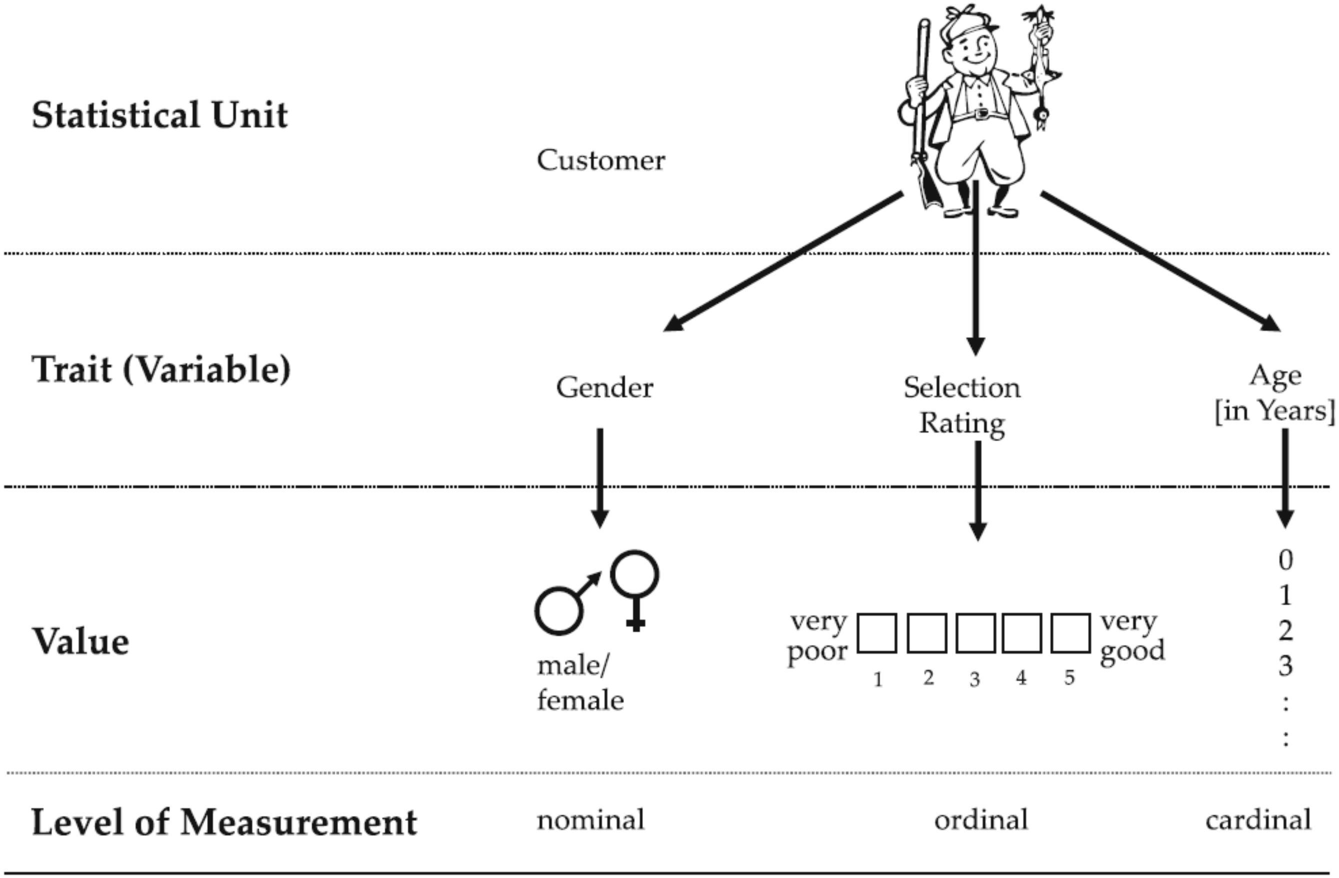
□ butter □ margarine □ other

On a scale of 1 (poor) to 5 (excellent) how do rate the selection of your preferred spread at our store?

□(1)	□(2)	□(3)	□(4)	□(5)
poor	fair	average	good	excellent

Cleff, T. (2014). ***Exploratory Data Analysis in Business and Economics***. Springer Cham

Level of Measurement (Section 2)



Statistical Unit	Customer		
Trait (Variable)	Gender	Selection Rating	Age [in Years]
Value	♂ ♀ male/ female	very poor □ □ □ □ □ very good 1 2 3 4 5	0 1 2 3 : :
Level of Measurement	nominal	ordinal	cardinal

Cleff, T. (2014). ***Exploratory Data Analysis in Business and Economics***. Springer Cham

Level of Measurement (Section 2)

- Statistical unit (who to question?)
- The relevant traits or variables (what to question?)
- The trait values (what answers can be given?)

- Variables can be classified as either discrete or continuous variables.
  - **Discrete variables** can only take on certain given numbers.
    Ex. Male/Female, size of a family (1, 2, 3, 4, …), Levels of education
  - **Continuous variables** can take on any value within an interval of numbers.
    Ex. weight or height

Cleff, T. (2014). ***Exploratory Data Analysis in Business and Economics.*** Springer Cham

Level of Measurement (Section 2)

- **Nominal scale**, which is sometimes also referred to as qualitative variable.
  - The values serve to assign each statistical unit to a specific group.
  - Every statistical unit can only be assigned to one group and all statistical units with the same trait status receive the same number.
- **Ordinal scale** means numbers are assigned and here they express a rank. With an ordinal scale, traits can be ordered

Cleff, T. (2014). ***Exploratory Data Analysis in Business and Economics***. Springer Cham

Level of Measurement (Section 2)

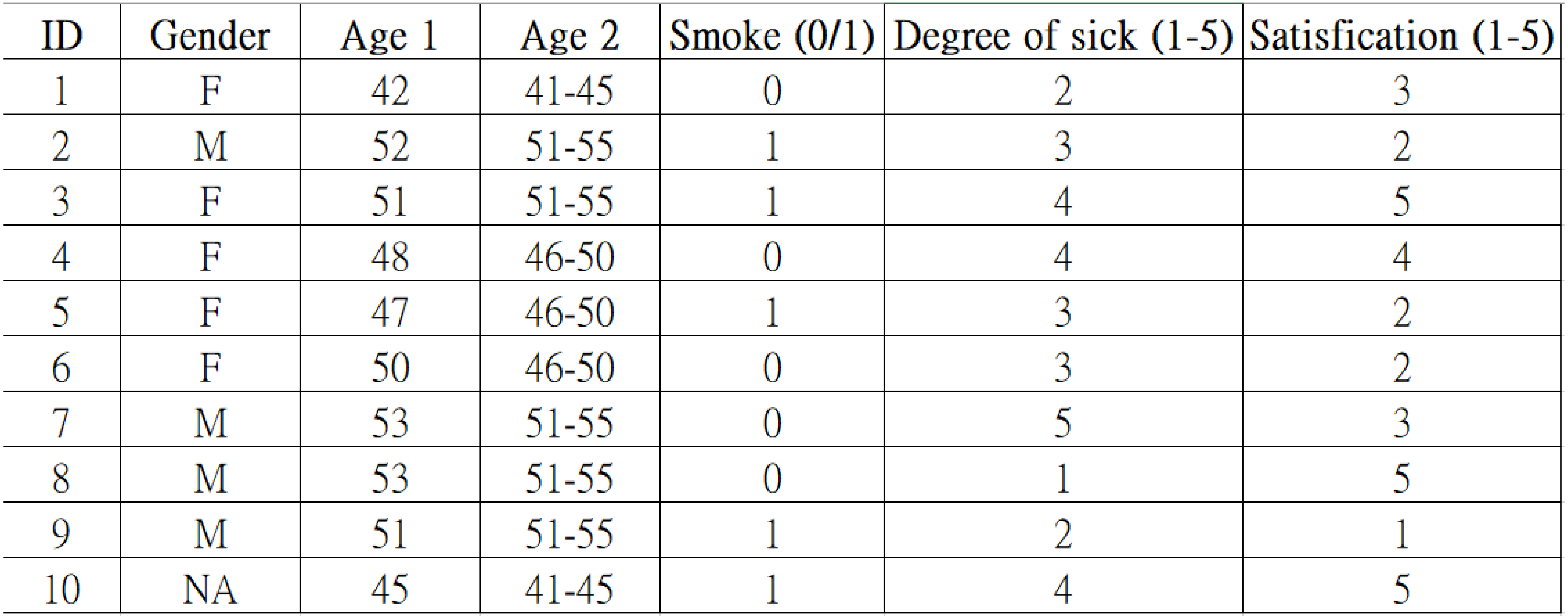
- **Cardinal scale** contains not only the information of the ordinal scales but also the distance between value traits held by two statistical units.

□ Additional perspective: the meaning of the distance between values (items).

- no meaningful
- there is meaningful and with **unequal** level of increase
- there is meaningful and with **equal** level of increase

Cleff, T. (2014). ***Exploratory Data Analysis in Business and Economics.*** Springer Cham

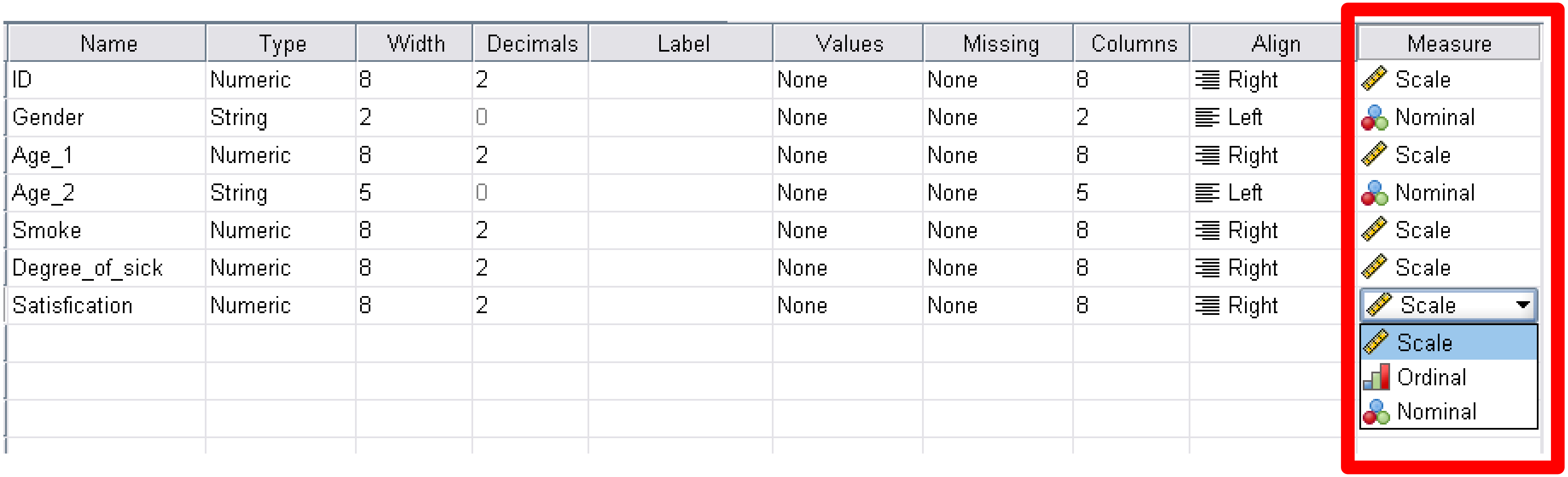
Practice (I)



ID	Gender	Age 1	Age 2	Smoke (0/1)	Degree of sick (1-5)	Satisfication (1-5)
1	F	42	41-45	0	2	3
2	M	52	51-55	1	3	2
3	F	51	51-55	1	4	5
4	F	48	46-50	0	4	4
5	F	47	46-50	1	3	2
6	F	50	46-50	0	3	2
7	M	53	51-55	0	5	3
8	M	53	51-55	0	1	5
9	M	51	51-55	1	2	1
10	NA	45	41-45	1	4	5

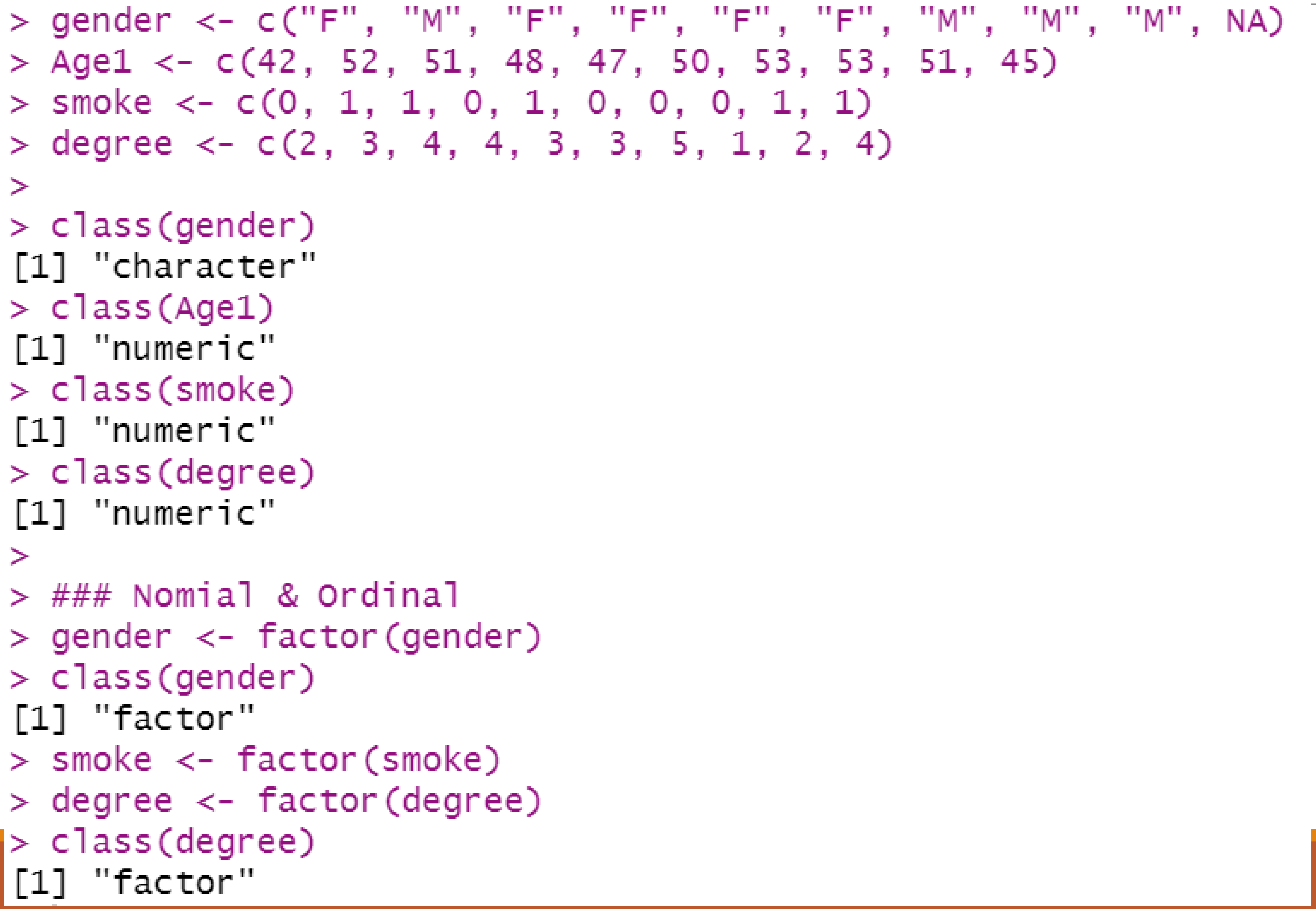
Nominal? Ordinal? Cardinal?

In SPSS



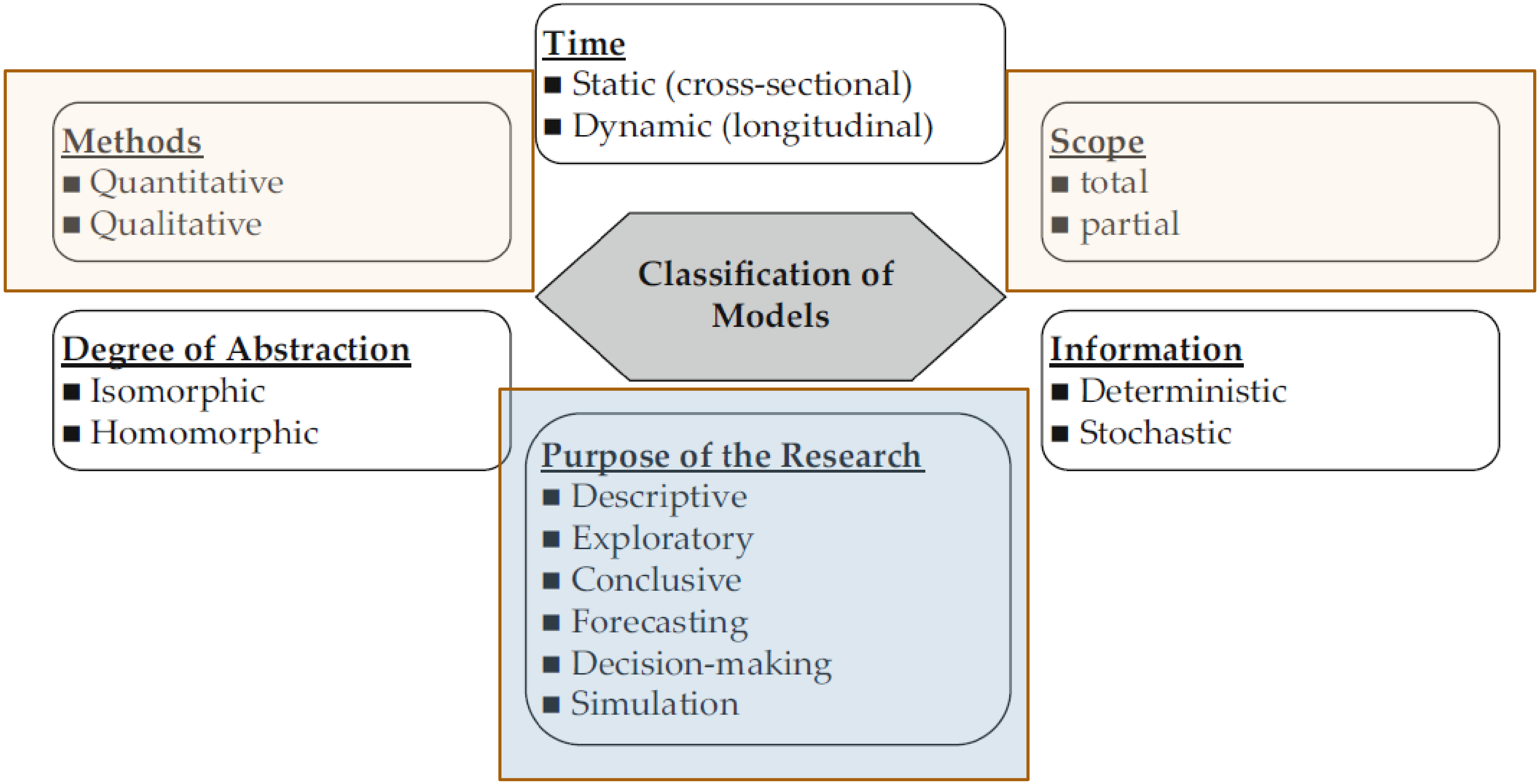
Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
ID	Numeric	8	2		None	None	8	Right	Scale
Gender	String	2	0		None	None	2	Left	Nominal
Age_1	Numeric	8	2		None	None	8	Right	Scale
Age_2	String	5	0		None	None	5	Left	Nominal
Smoke	Numeric	8	2		None	None	8	Right	Scale
Degree_of_sick	Numeric	8	2		None	None	8	Right	Scale
Satisfication	Numeric	8	2		None	None	8	Right	Scale

In R



```
> gender <- c("F", "M", "F", "F", "F", "F", "M", "M", "M", NA)
> Age1 <- c(42, 52, 51, 48, 47, 50, 53, 53, 51, 45)
> smoke <- c(0, 1, 1, 0, 1, 0, 0, 0, 1, 1)
> degree <- c(2, 3, 4, 4, 3, 3, 5, 1, 2, 4)
>
> class(gender)
[1] "character"
> class(Age1)
[1] "numeric"
> class(smoke)
[1] "numeric"
> class(degree)
[1] "numeric"
>
> ### Nomial & Ordinal
> gender <- factor(gender)
> class(gender)
[1] "factor"
> smoke <- factor(smoke)
> degree <- factor(degree)
> class(degree)
[1] "factor"
```

A systematic overview of model variants (Section 1)



Classification of Models

Time
- Static (cross-sectional)
- Dynamic (longitudinal)

Methods
- Quantitative
- Qualitative

Scope
- total
- partial

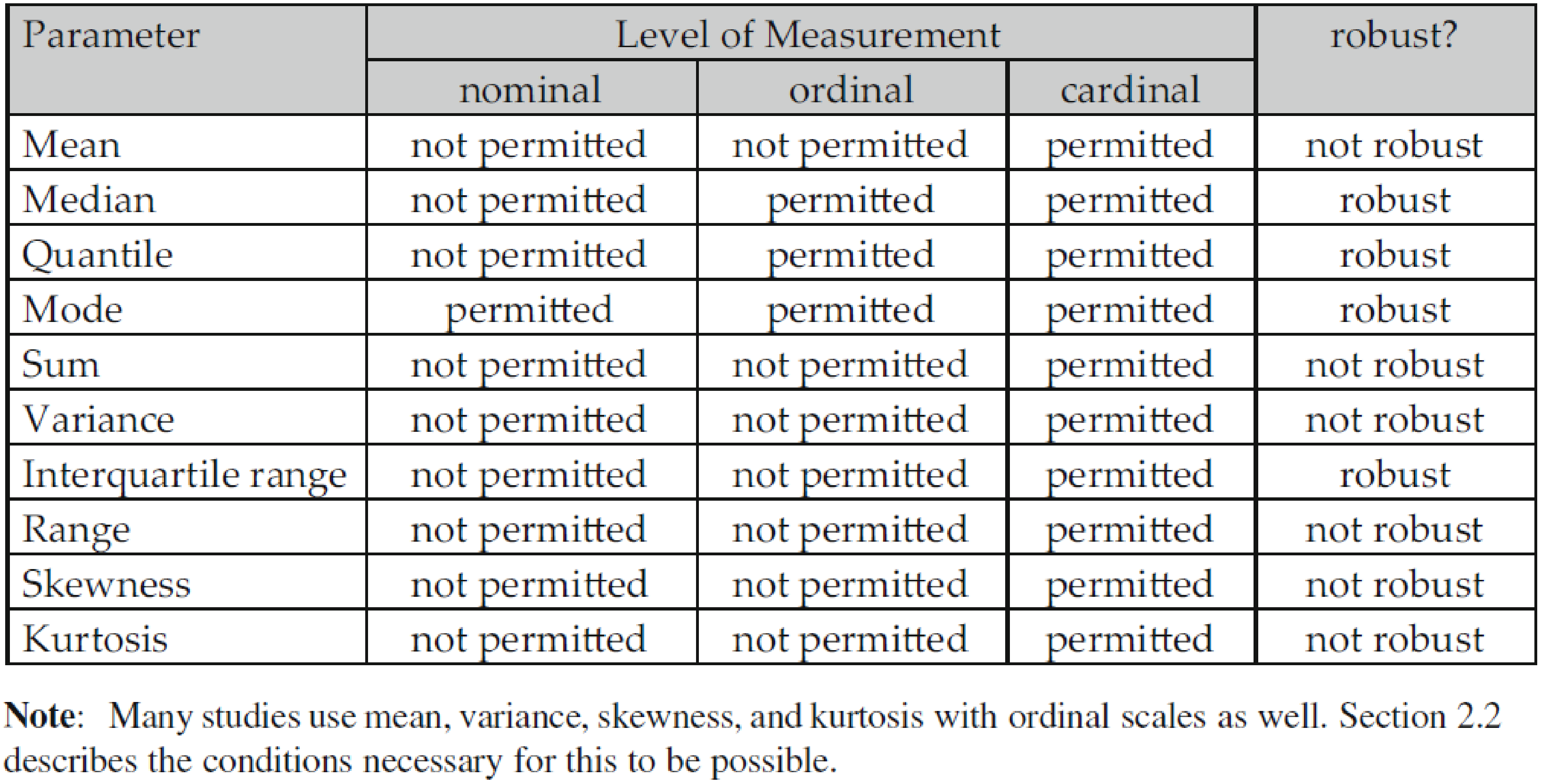
Degree of Abstraction
- Isomorphic
- Homomorphic

Information
- Deterministic
- Stochastic

Purpose of the Research
- Descriptive
- Exploratory
- Conclusive
- Forecasting
- Decision-making
- Simulation

Cleff, T. (2014). ***Exploratory Data Analysis in Business and Economics***. Springer Cham

Assignment (Section 3)
Descriptive Statistics (p.52)



Parameter	Level of Measurement			robust?
	nominal	ordinal	cardinal	
Mean	not permitted	not permitted	permitted	not robust
Median	not permitted	permitted	permitted	robust
Quantile	not permitted	permitted	permitted	robust
Mode	permitted	permitted	permitted	robust
Sum	not permitted	not permitted	permitted	not robust
Variance	not permitted	not permitted	permitted	not robust
Interquartile range	not permitted	not permitted	permitted	robust
Range	not permitted	not permitted	permitted	not robust
Skewness	not permitted	not permitted	permitted	not robust
Kurtosis	not permitted	not permitted	permitted	not robust

Note: Many studies use mean, variance, skewness, and kurtosis with ordinal scales as well. Section 2.2 describes the conditions necessary for this to be possible.

Cleff, T. (2014). ***Exploratory Data Analysis in Business and Economics.*** Springer Cham