Statistical Methods
統計方法

SEPTEMBER 12, 2023

I-CHEN LEE

Q&A with TA

TA: 徐瑋柔

Regular TA office hour at 62225:

- Thursday 9：00~10：00
- Friday 16：00~17：00

In order to answer your questions more effectively and accurately, please send her an email to describe your questions one day before the TA hour.

r26114092@gs.ncku.edu.tw

Problem Definition

Purpose of statistical analysis:

- Knowing distributions of data
- Comparison (Two groups / multi-groups)
- Fit a model to investigate the relationship
- Classification
- Clustering
- Dimension reduction

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

Univariate data analysis (Section 3)

- ✓Distribution function (pie chart, a horizontal bar chart, or a vertical bar chart, histogram)

Bivariate data analysis (Section 4)

- ✓Relationship (Scatter plot)

Multivariate data analysis (Sections 5-8)

- ✓Relationship (Scatter plot with labels)
- ✓Clustering (Scatter plot with labels)

R source

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

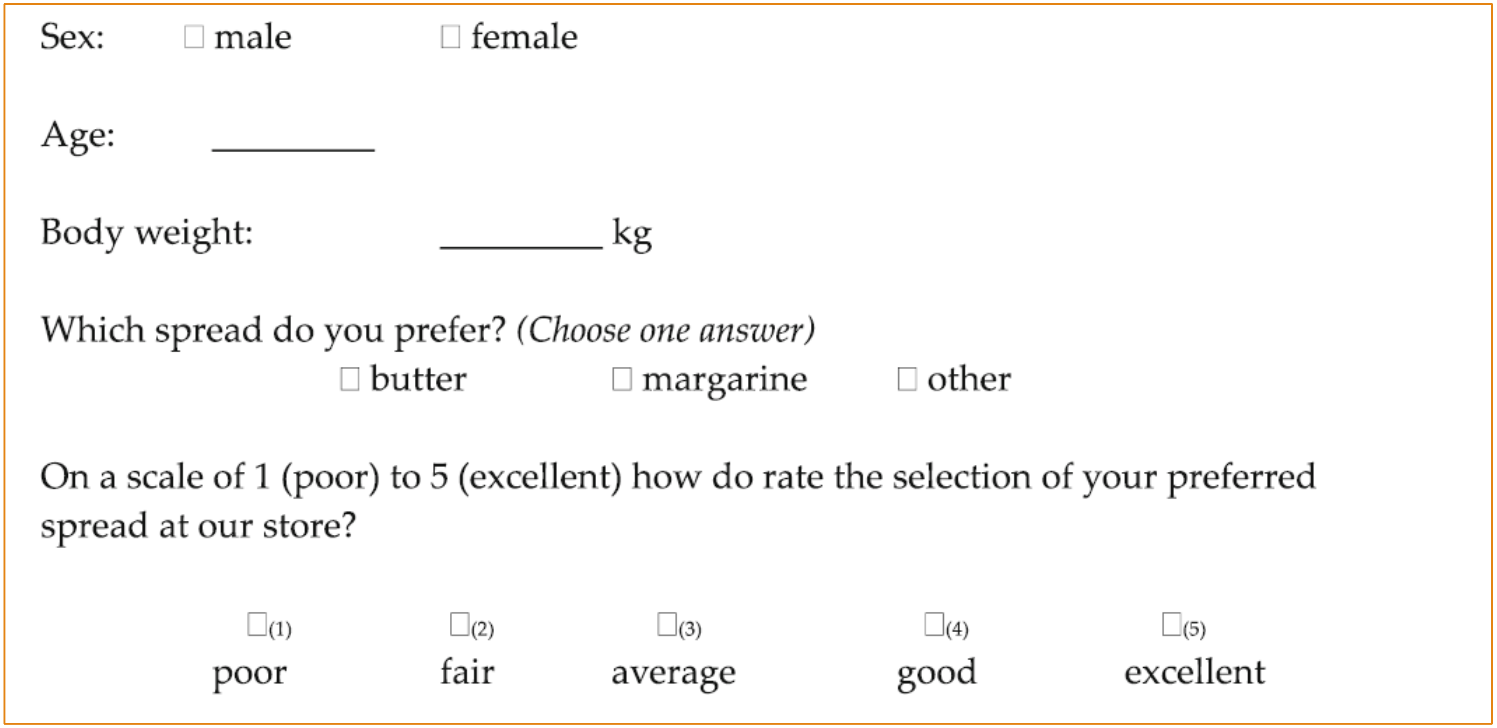
1. Download R: https://cran.csie.ntu.edu.tw/
2. Download Rstudio: https://www.rstudio.com/products/rstudio/

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

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)*

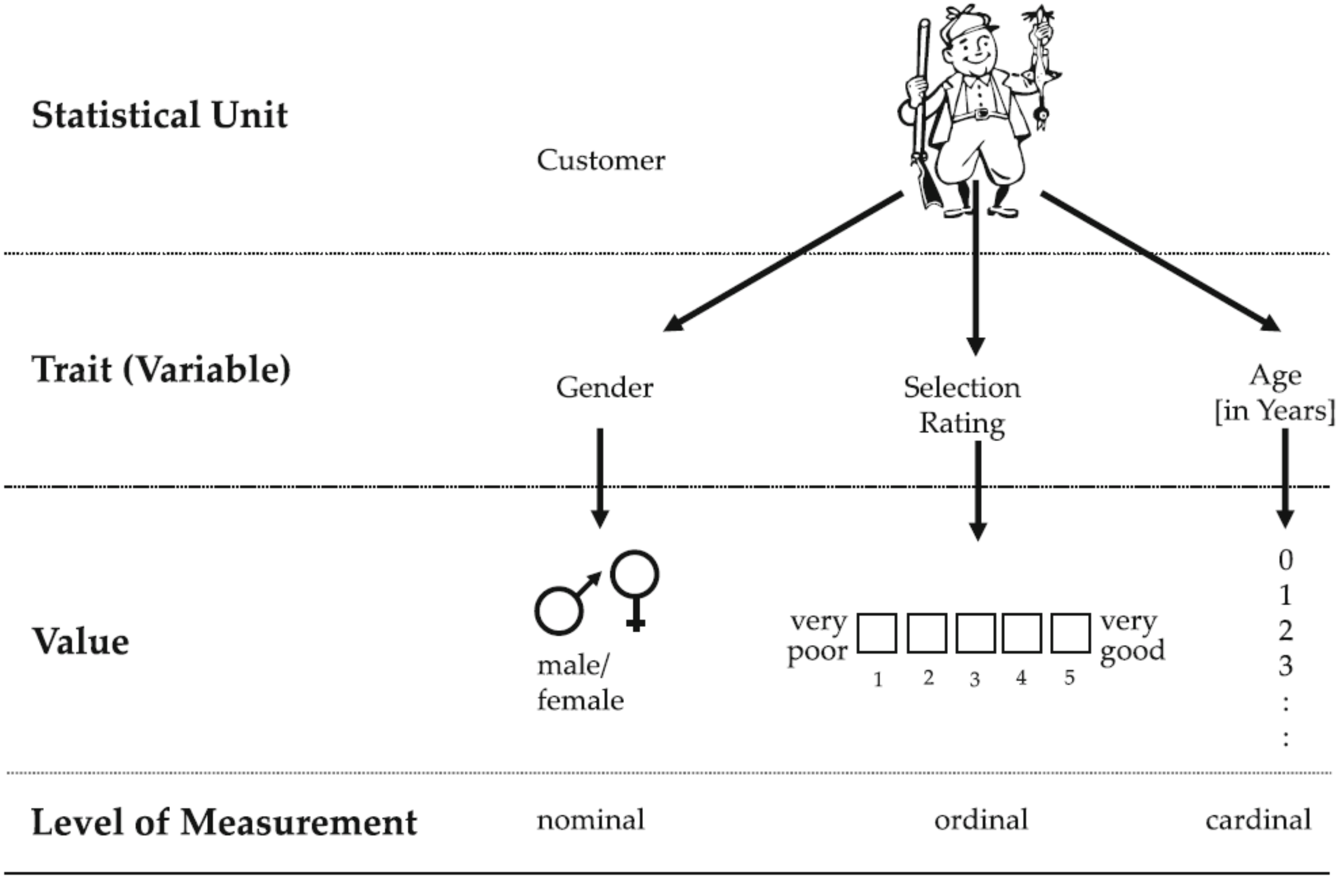
□ butter □ margarine □ other

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

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

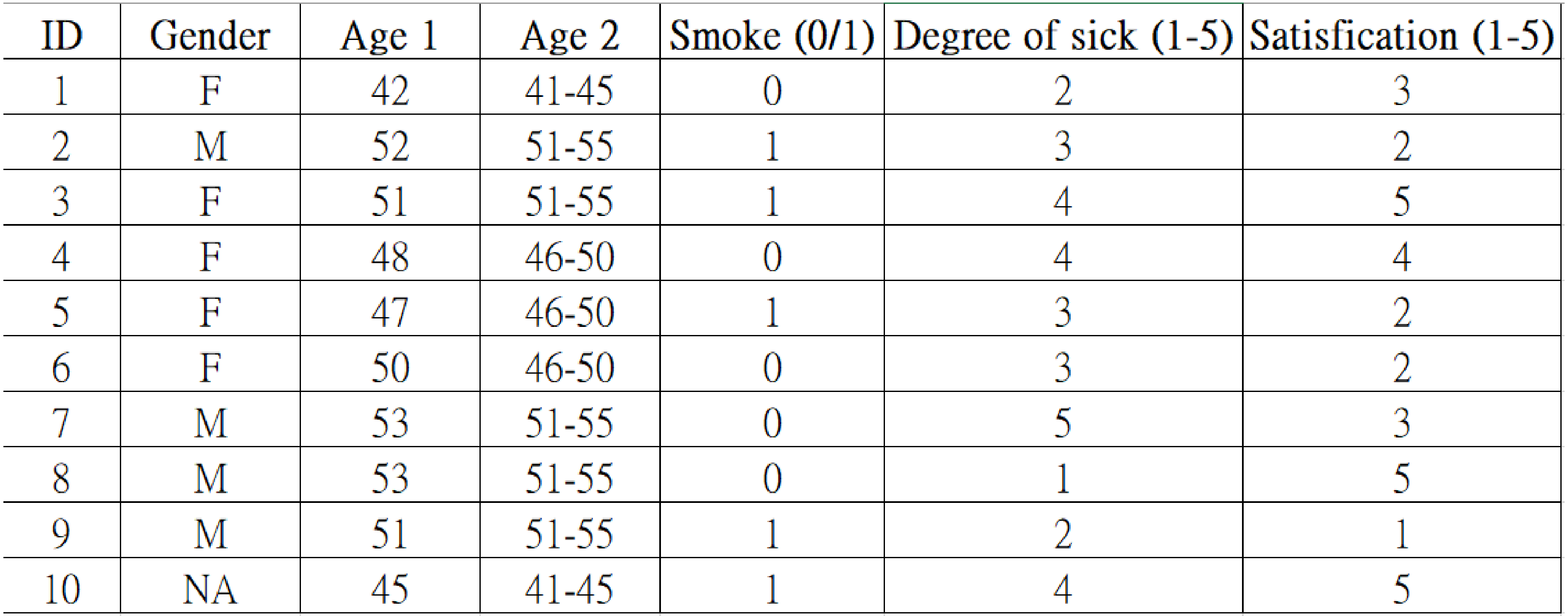
Level of Measurement (Section 2)

- 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

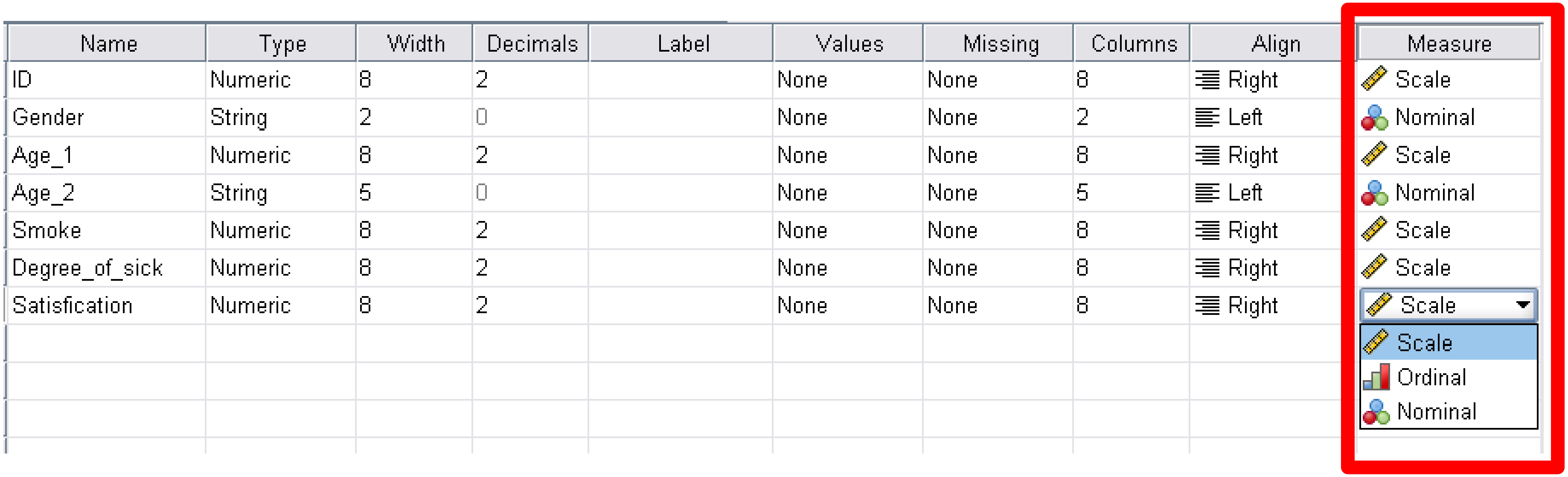
Practice (I)



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

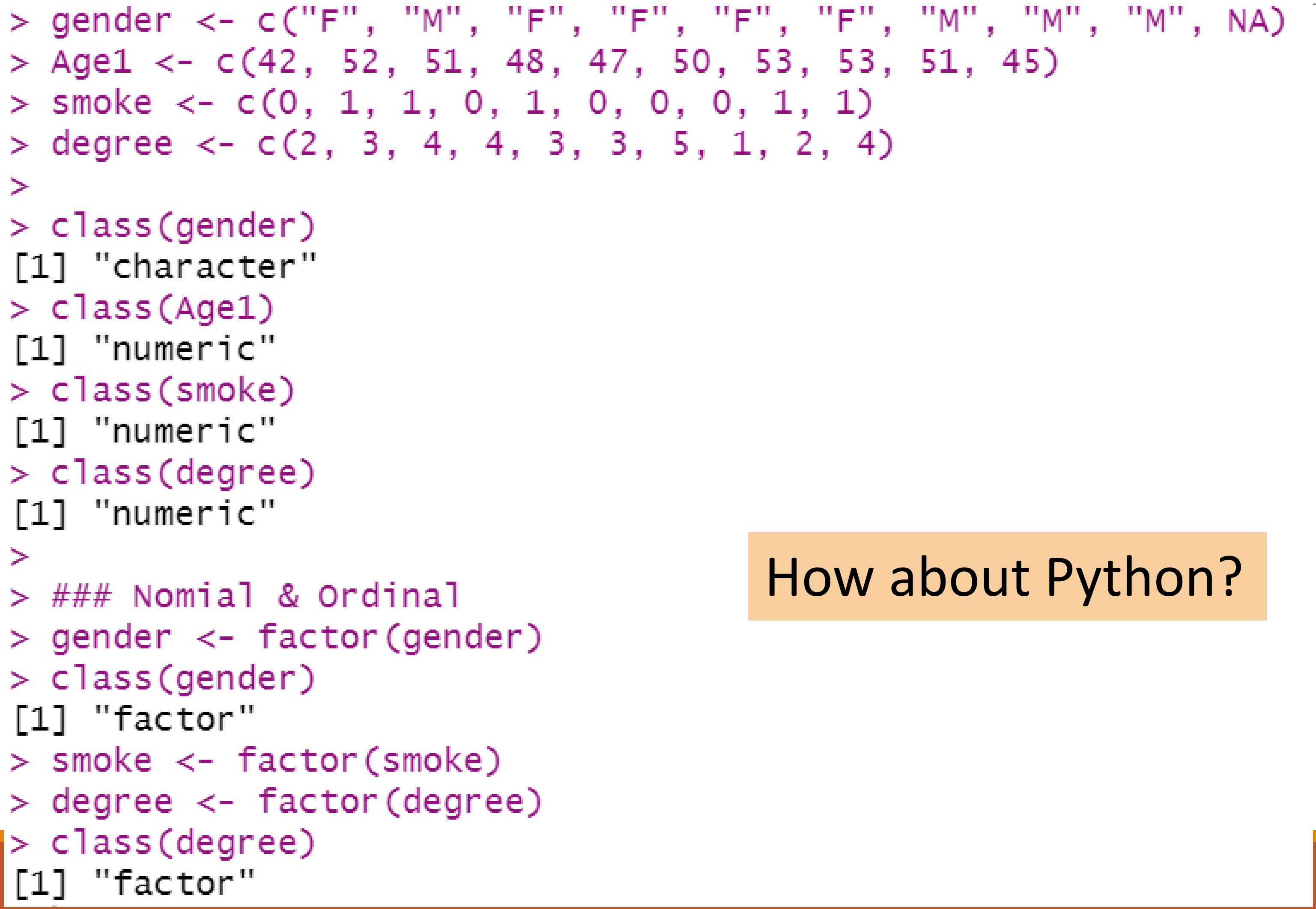
👍 : Nomina 👏 : Ordinal 🙌 : Cardinal

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

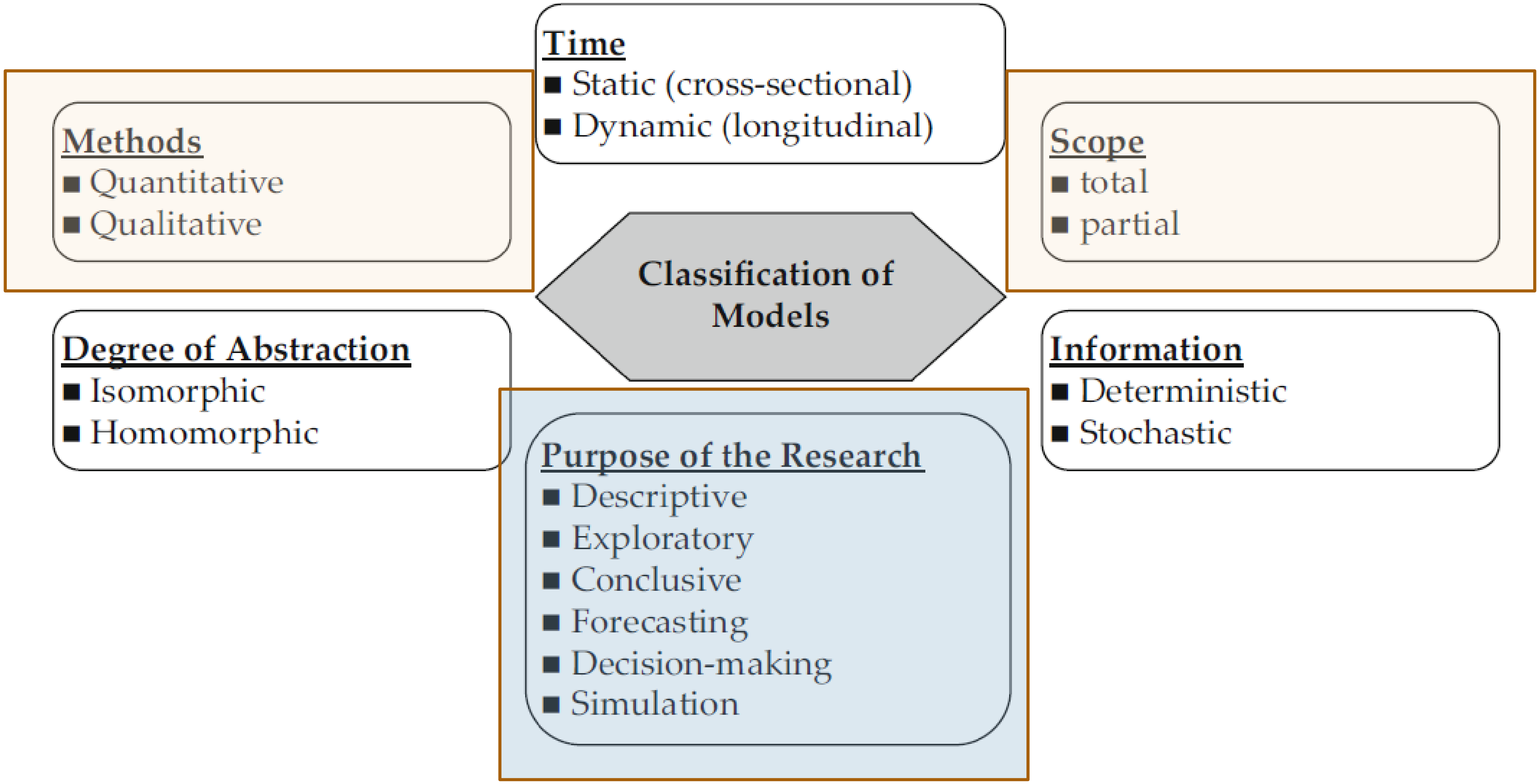
In R (week02.R)



```
> 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"
```

How about Python?

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

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

From Models to Business Intelligence

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

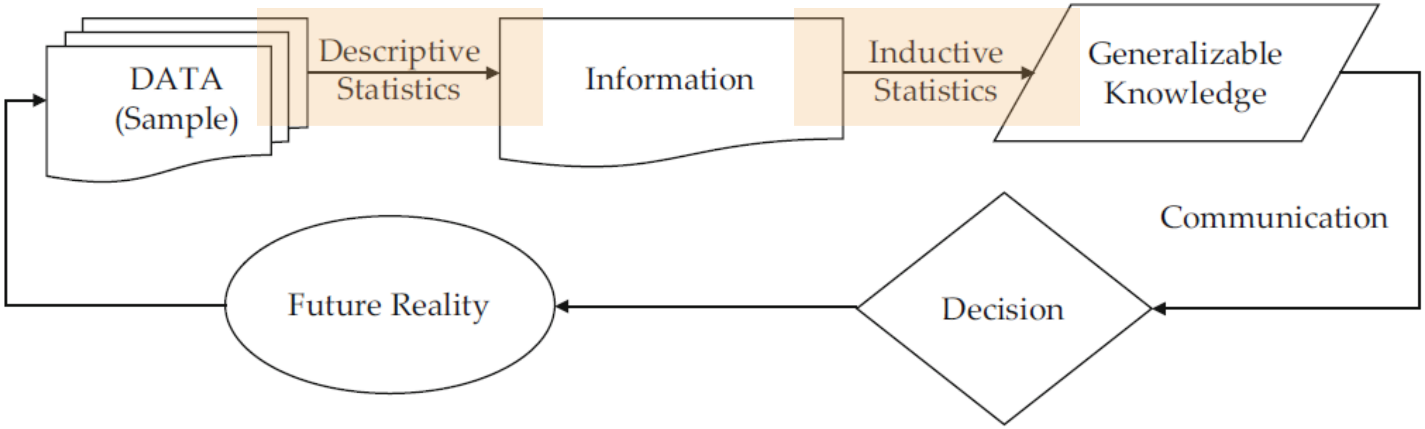


Fig. 1.6 The intelligence cycle (Source: Own graphic, adapted from Harkleroad 1996, p. 45)

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

Exploratory Data Analysis (EDA) 探索性資料分析

Most EDA techniques are graphical in nature with a few quantitative techniques.

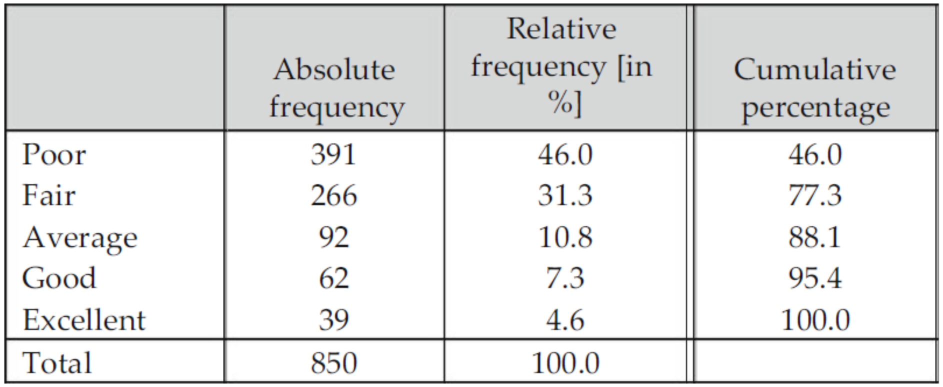
- **Plotting: to identify and understand the patterns of data**
- Basic statistical concepts and assumptions with the corresponding the plots
- Pattern recognition and conceptual models (linear and non-linear)
- Outlier detection

Univariate Analysis

Analysis of **only** one variable

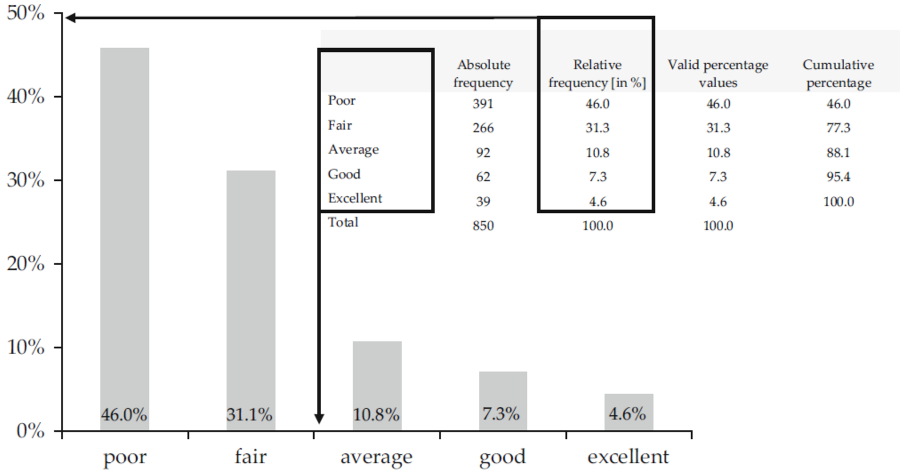
Nominal? Ordinal? Cardinal variables?

- Graphical representations for distributions (Bar chart/Histogram)
  - Frequency table (Bar chart)
  - Cumulative percentage



	Absolute frequency	Relative frequency [in %]	Cumulative percentage
Poor	391	46.0	46.0
Fair	266	31.3	77.3
Average	92	10.8	88.1
Good	62	7.3	95.4
Excellent	39	4.6	100.0
Total	850	100.0	

Fig. 3.2 Frequency table for selection ratings



	Absolute frequency	Relative frequency [in %]	Valid percentage values	Cumulative percentage
Poor	391	46.0	46.0	46.0
Fair	266	31.3	31.3	77.3
Average	92	10.8	10.8	88.1
Good	62	7.3	7.3	95.4
Excellent	39	4.6	4.6	100.0
Total	850	100.0	100.0	

Fig. 3.3 Bar chart/Frequency distribution for the selection variable

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

Univariate Analysis
Histogram

Cardinal variables

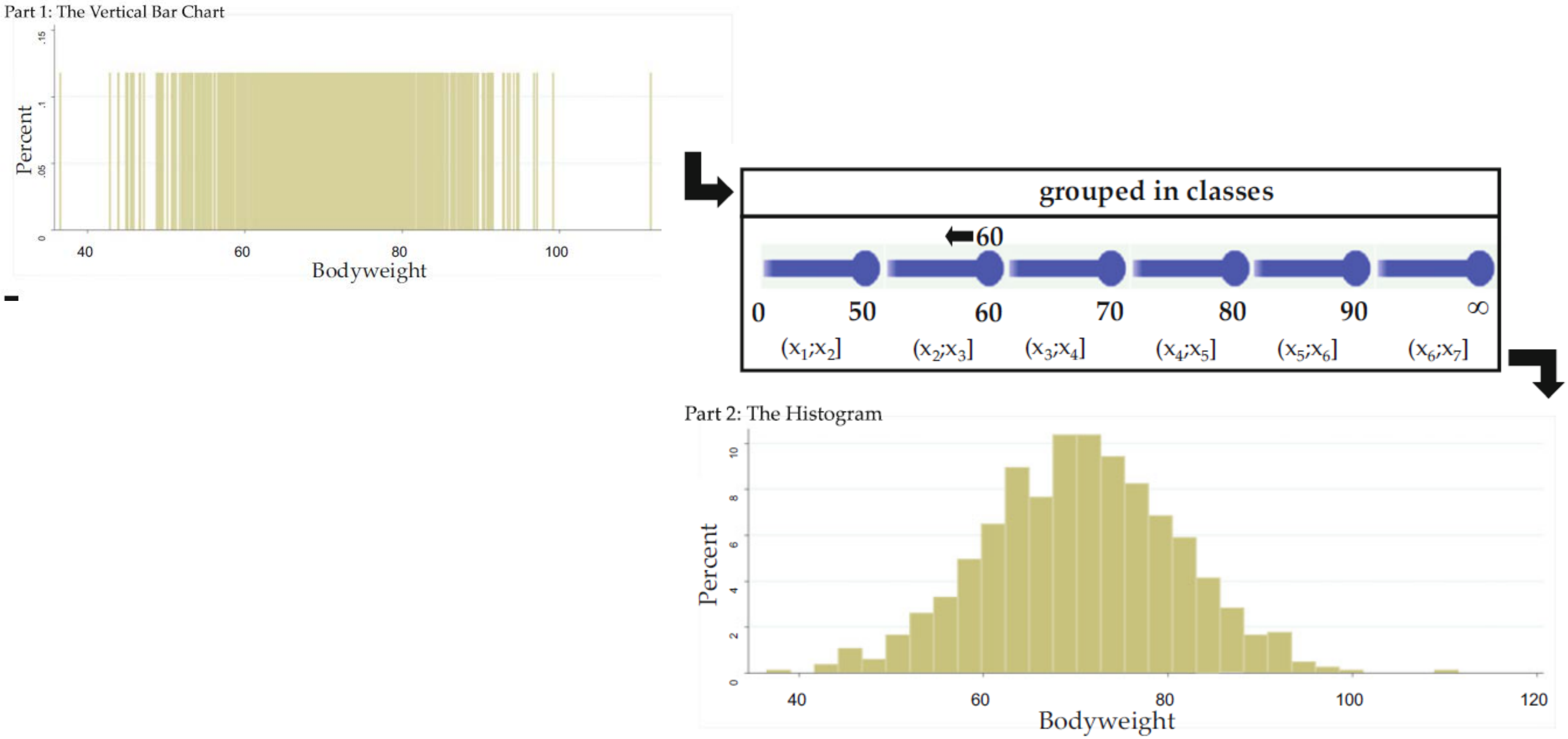


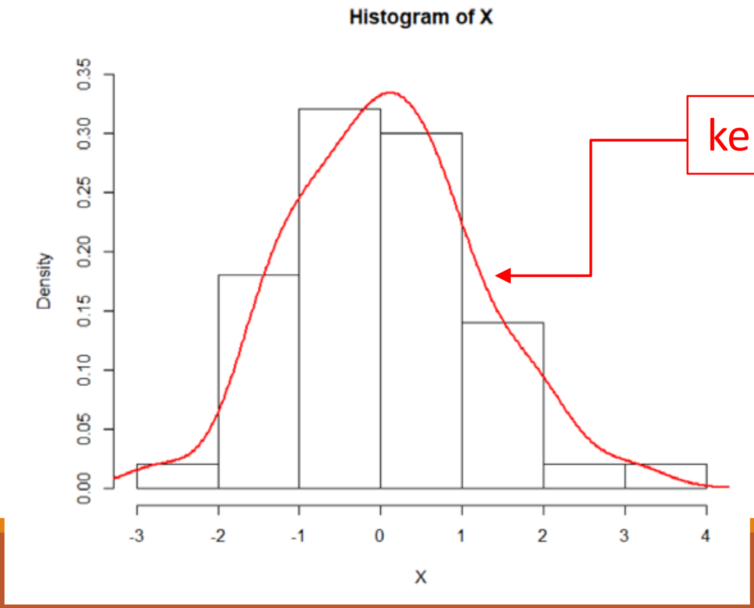
Fig. 3.7 Using a histogram to classify data

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

Histogram (直方圖)

Histogram: it is an "approximate" representation of the distribution of numerical data.

- bin size: the range of values
- frequency: the number of values in the specific bin
- it could be thought of as a kernel density estimation



kernel density function

Histogram (直方圖)

Let n denote the number of observations, and

$$n = \sum_{i=1}^{k} m_i,$$

where m_i is the number of values in the ith bin, and k is the bin size. That is,

$$m_i = \left|\{x | x \in [a_i, b_i]\}\right|, \ \ i = 1, \ldots, k,$$

where a_i and b_i are the lower and upper bounds of the ith bin. Intuitively, $k = \left[\dfrac{\max X - \min X}{h}\right]$.

Density estimation

The kernel density estimation is a nonparametric way to estimate the probability density function.

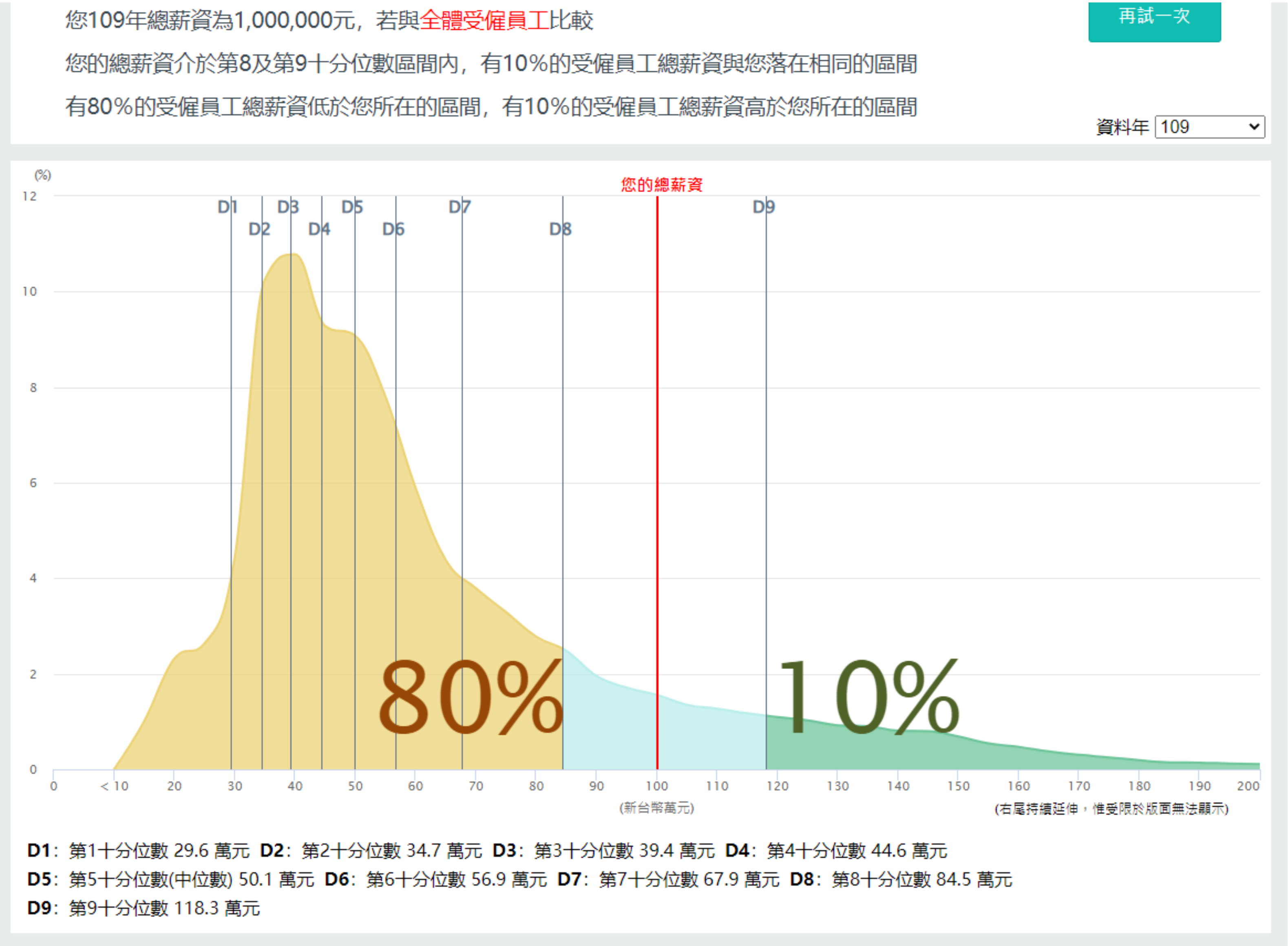
$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

where h is the bandwidth and $K(\cdot)$ is the kernel function.

- Gaussian kernel:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{u^2}{2}\right\}.$$

$$h = 1.06\hat{\sigma}n^{-1/5} \text{ or } h = 0.9\min\left\{\hat{\sigma}, \frac{\text{IQR}}{1.34}\right\} n^{-1/5}.$$



再試一次

您109年總薪資為1,000,000元，若與全體受僱員工比較

您的總薪資介於第8及第9十分位數區間內，有10%的受僱員工總薪資與您落在相同的區間

有80%的受僱員工總薪資低於您所在的區間，有10%的受僱員工總薪資高於您所在的區間

資料年 109

D1：第1十分位數 29.6 萬元 **D2**：第2十分位數 34.7 萬元 **D3**：第3十分位數 39.4 萬元 **D4**：第4十分位數 44.6 萬元
D5：第5十分位數(中位數) 50.1 萬元 **D6**：第6十分位數 56.9 萬元 **D7**：第7十分位數 67.9 萬元 **D8**：第8十分位數 84.5 萬元
D9：第9十分位數 118.3 萬元

https://earnings.dgbas.gov.tw/experience_sub_01.aspx

However,...

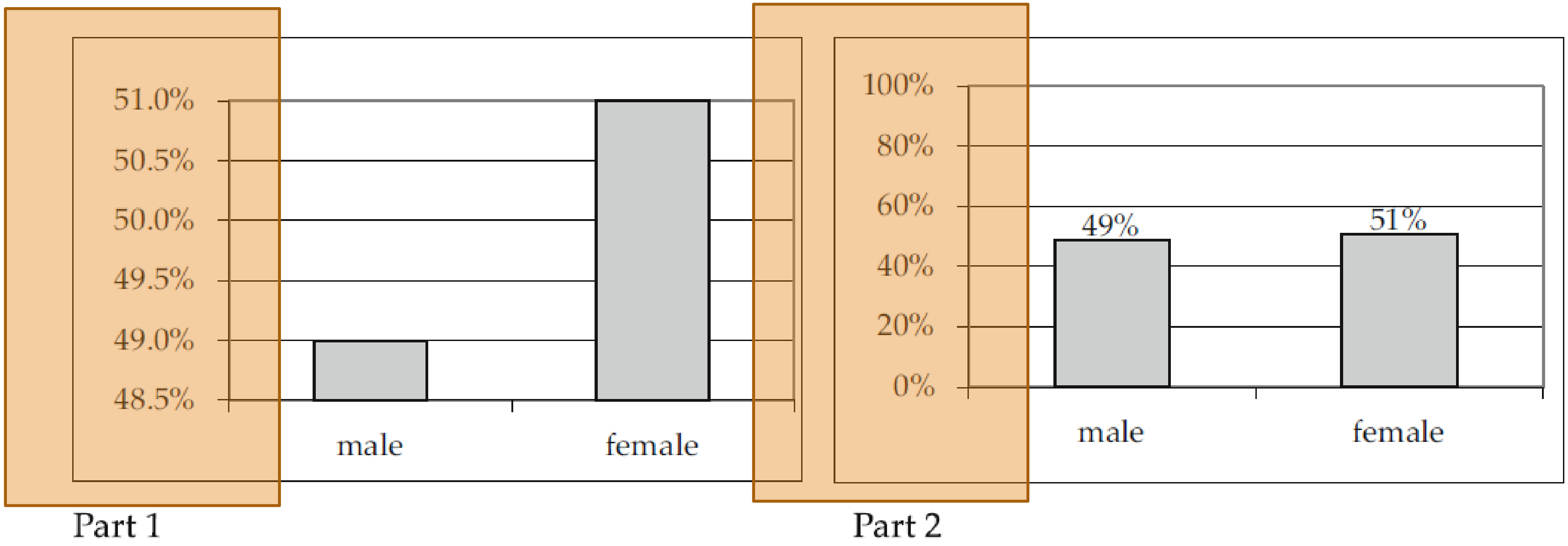


Fig. 3.5 Different representations of the same data (1)...

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

$$k = \left\lceil \frac{\max X - \min X}{\boxed{h}} \right\rceil .$$

However,...

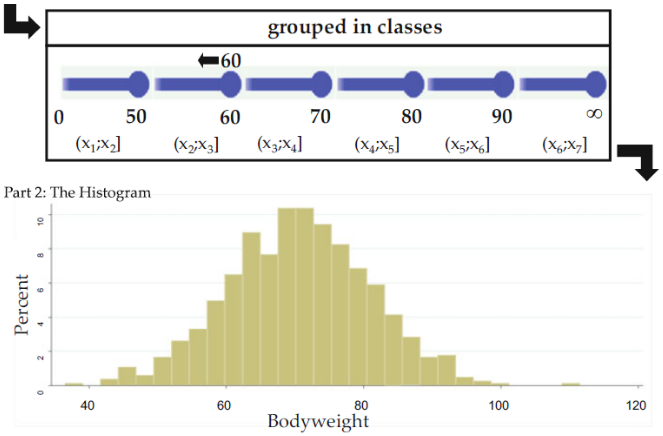
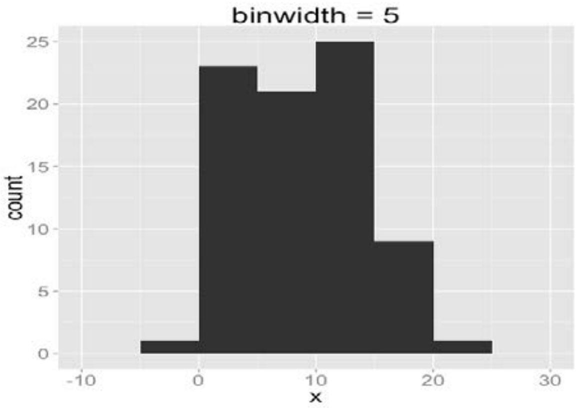
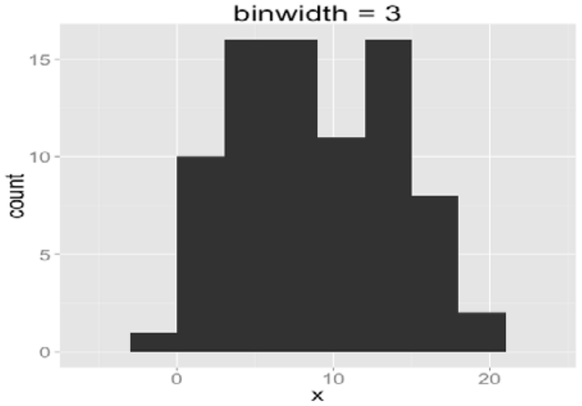
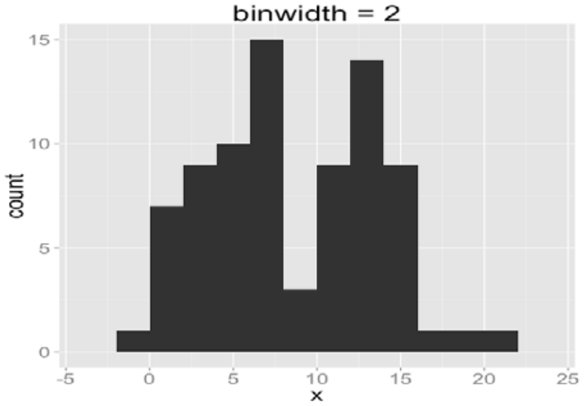


Fig. 3.7 Using a histogram to classify data







Summary of Section 3 (I)

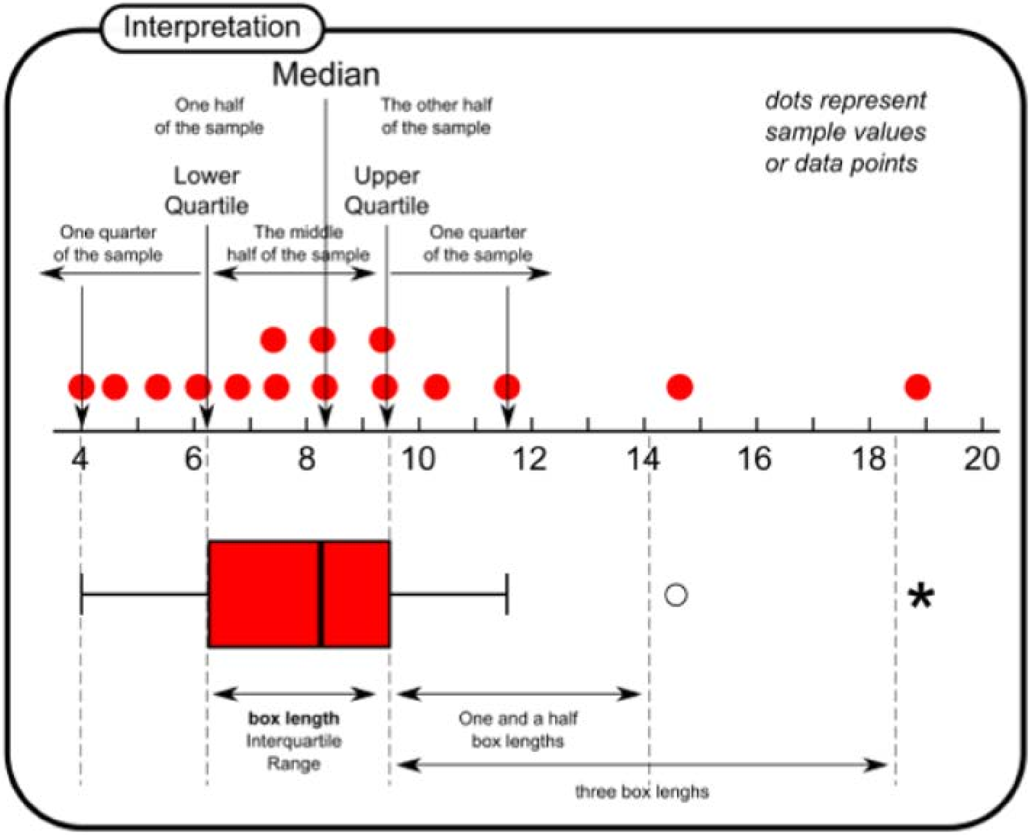
- Measures of Central Tendency (集中趨勢)
  - Mode (眾數)
  - Mean (平均數)/Geometric Mean/Harmonic Mean
  - Median (中位數)
  - Quartile (百分位數) and Percentile

- Dispersion Parameters (分散程度)
  - IQR (interquartile range, 四分位間距)
  - Range (全距)
  - Standard Deviation and Variance (標準差及變異數)
  - MAD (median absolute deviation)

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

Boxplot

Graphics with descriptive statistics


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

Histogram and Boxplot

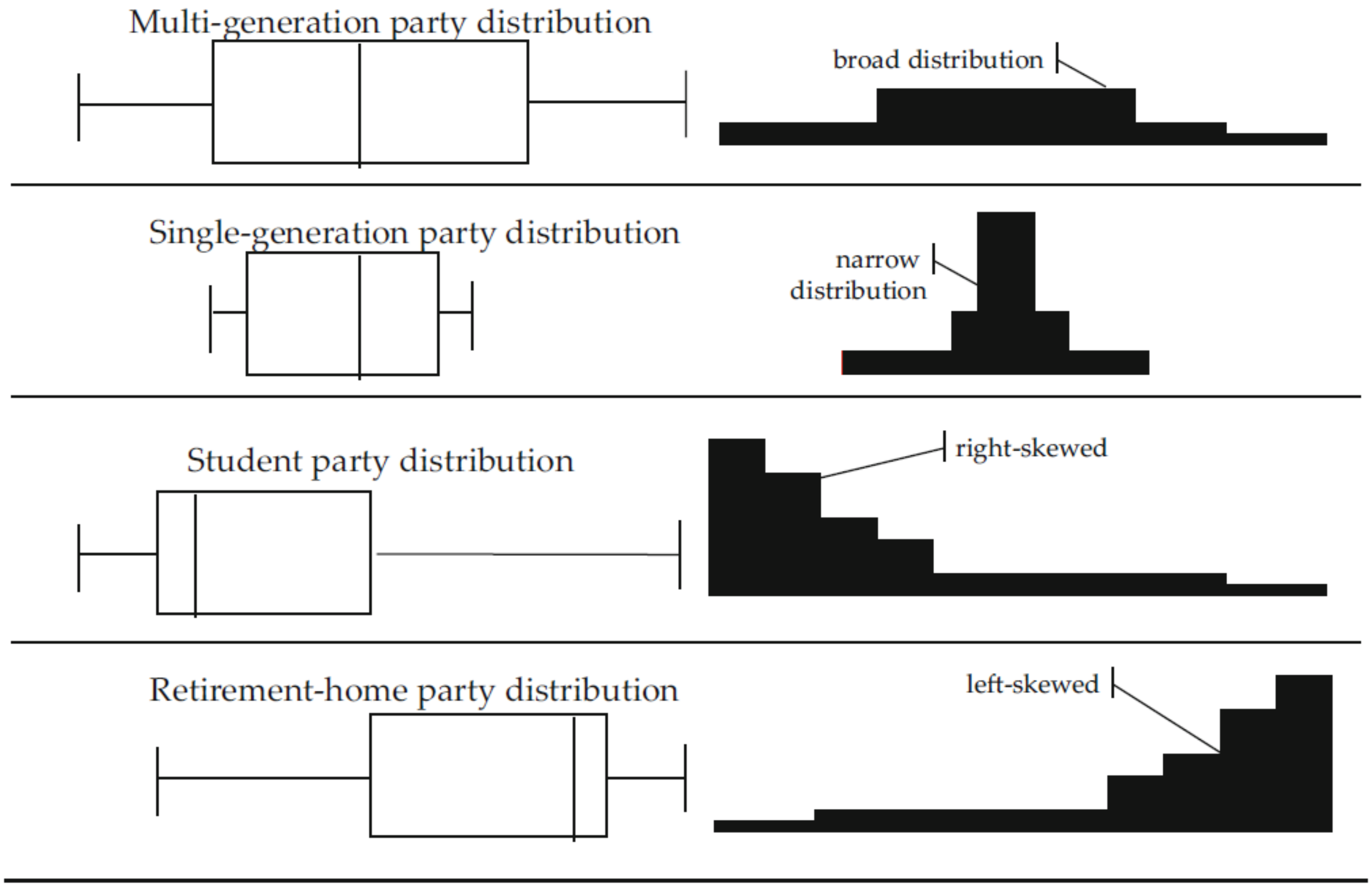
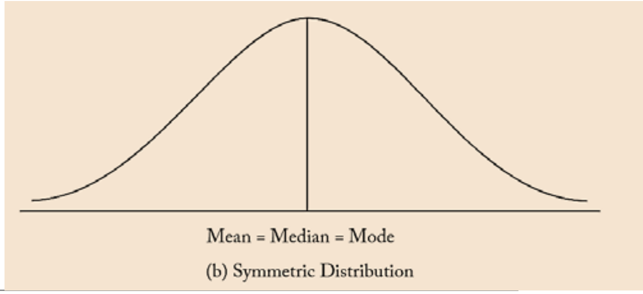

Multi-generation party distribution
broad distribution
Single-generation party distribution
narrow distribution
Student party distribution
right-skewed
Retirement-home party distribution
left-skewed

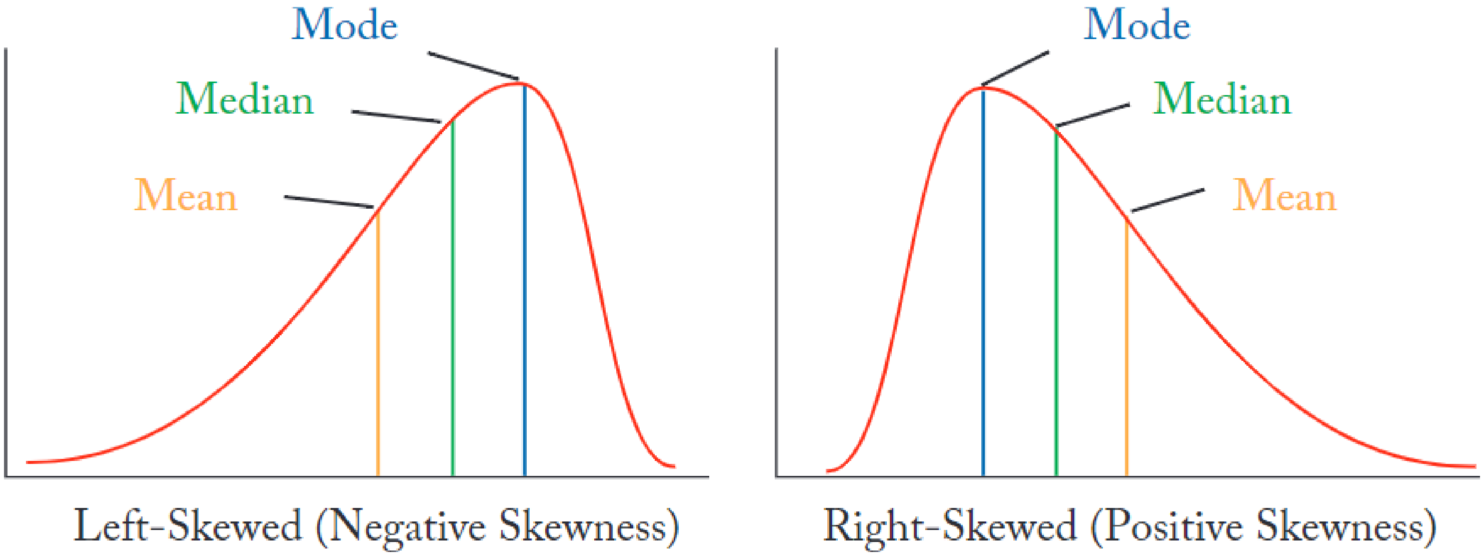
Fig. 3.18 Interpretation of different boxplot types

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

Descriptive Statistics (II)



- Skewness (gives the shape compared to the symmetric one)


Mode
Median
Mean
Left-Skewed (Negative Skewness)
Mode
Median
Mean
Right-Skewed (Positive Skewness)

Akinkunmi, M. (2019). Introduction to statistics using R.

Descriptive Statistics (III)

- Kurtosis (gives the shape compared to the normal distribution)
  - ➢ Leptokurtic: the peak of the distribution is steeper (Kurtosis > 3)
  - ➢ Mesokurtic: normal distribution (Kurtosis = 3)
  - ➢ Platykurtic: a flat peak (Kurtosis <3)

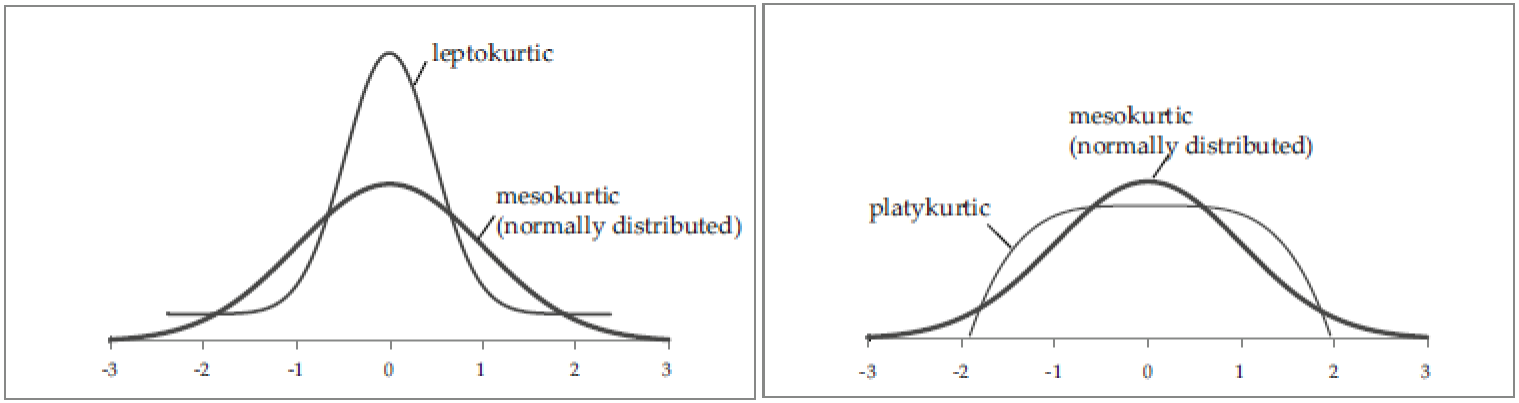
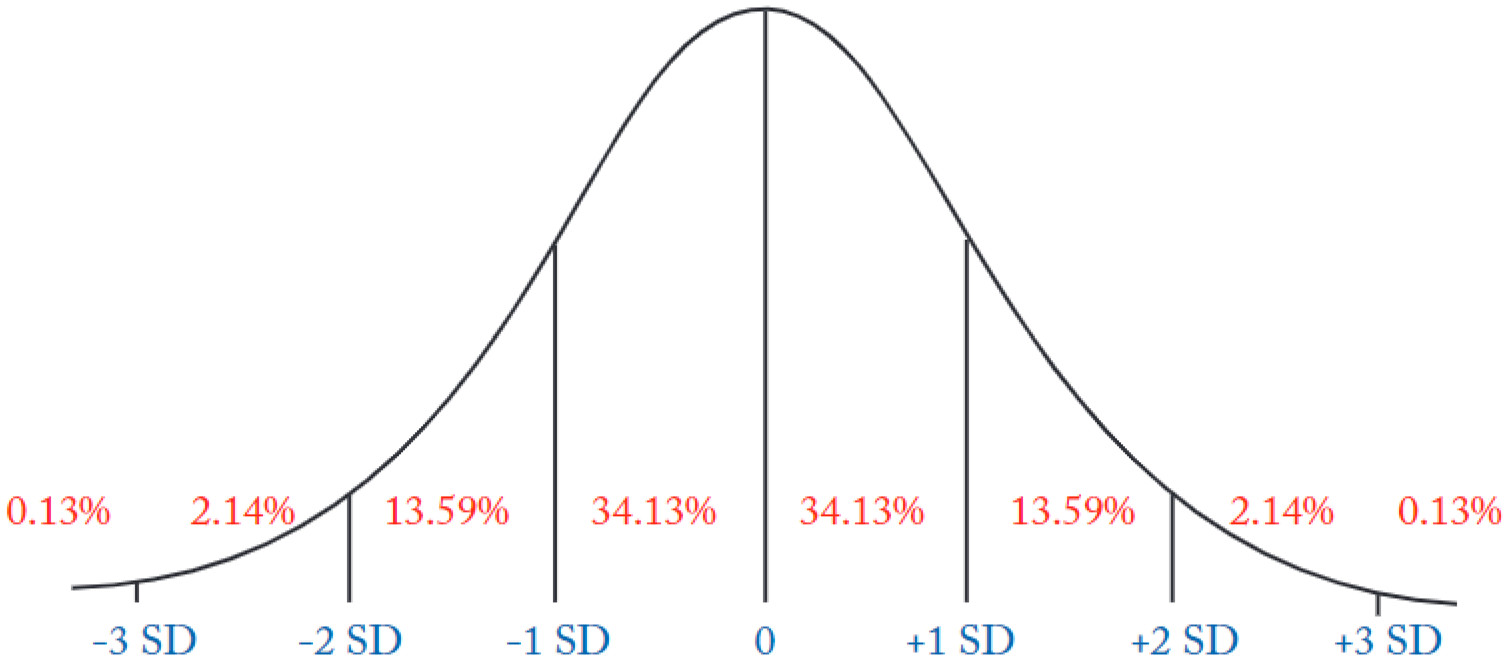


Fig. 3.22 Kurtosis distributions

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

The normal distribution



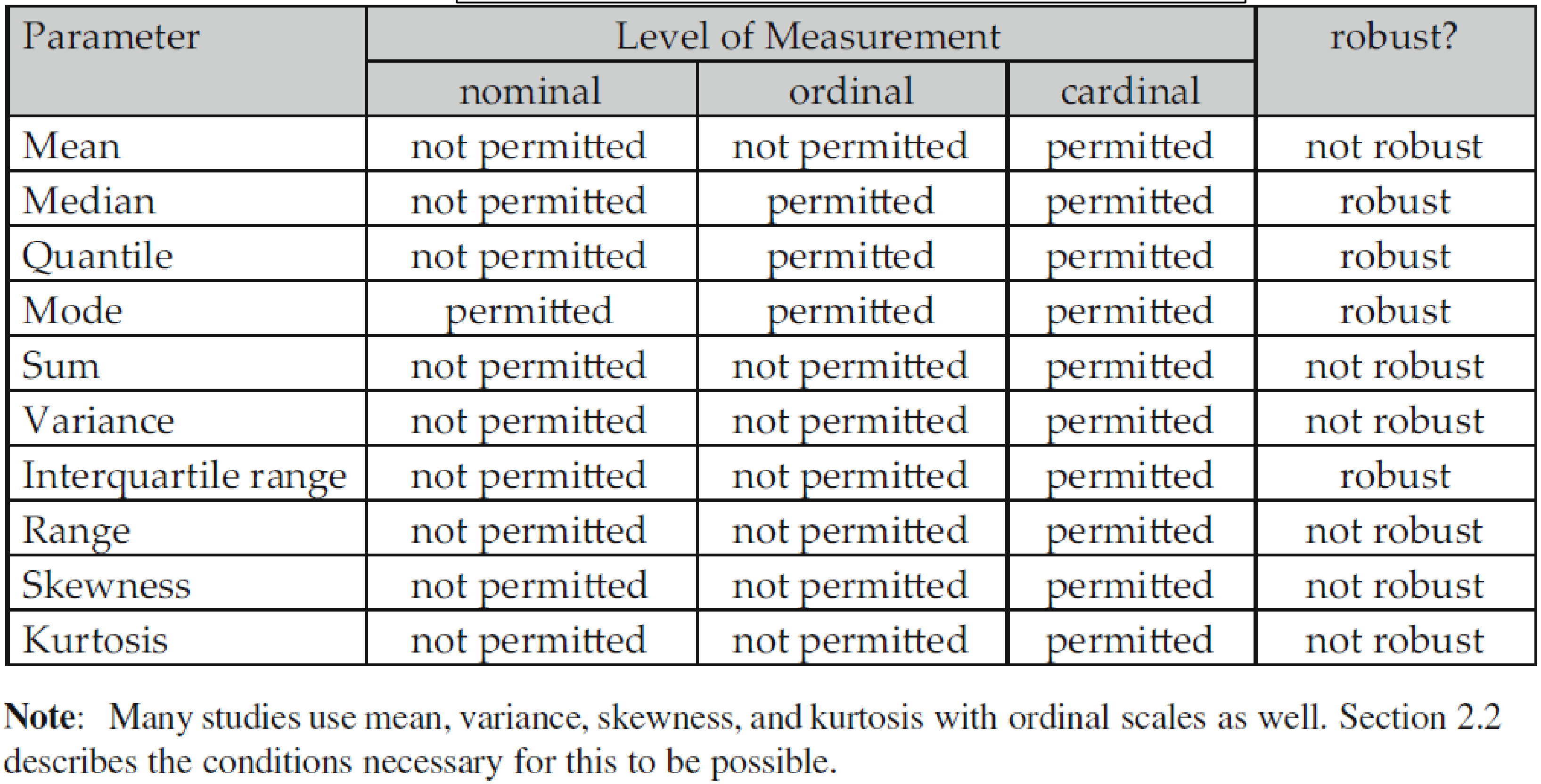
The normal distribution

a) The mean, median, and mode have the same value.
b) The curve is symmetric.
c) The total area under the curve is 1. (Why?)
d) The curve is denser in the center and less dense in the tails.
e) Normal distribution has two parameters: **mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma^2}$**.
f) The formulation of the curve is called the **probability density function** (pdf):

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, -\infty < x < \infty.$$

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

Nominal, Ordinal, Cardinal variables



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.

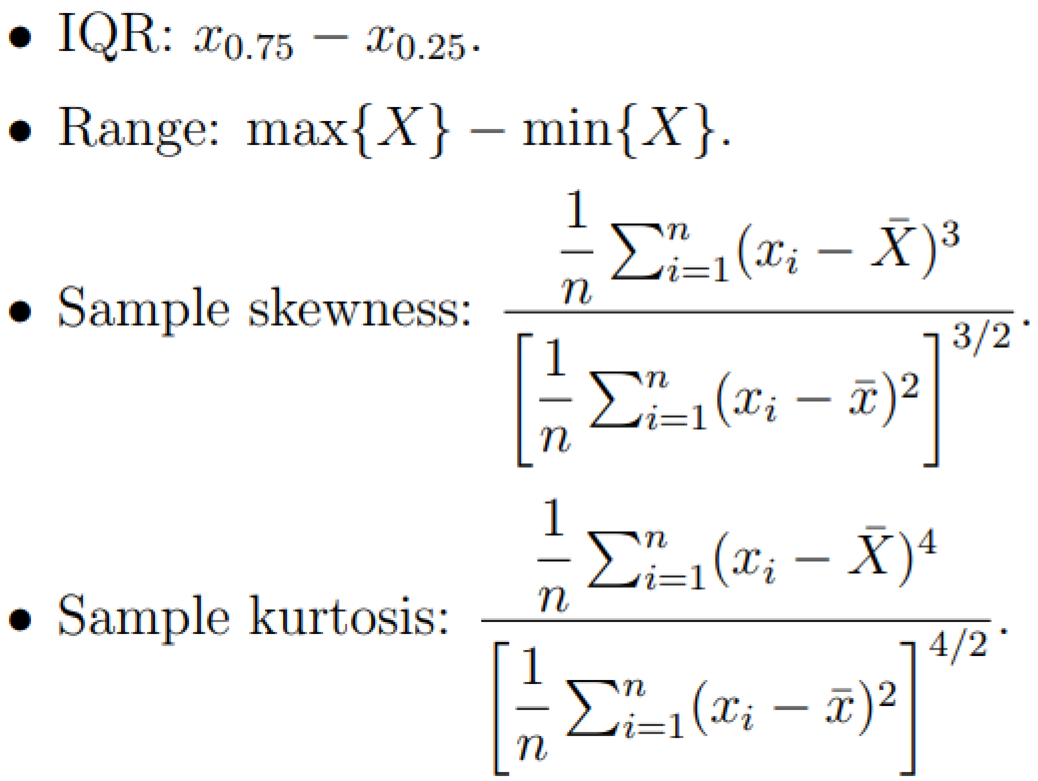
Cleff, T. (2014). ***Exploratory Data Analysis in Business and Economics.*** Springer Cham

The dataset is $X = \{x_1, x_2, \ldots, x_n\}$.
Let the ordered dataset is $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, where $x_{(j)} \leq x_{(j+1)},\ j = 1, \ldots, n-1$.

- Mean: $\bar{X} = \dfrac{1}{n} \sum_{i=1}^{n} x_i$.
- Median:

$$x_{0.5} = \tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right] & \text{if } n \text{ is even.} \end{cases}$$

- Quantile x_q: $Pr\{X \leq x_q\} \leq q$.
- Mode: The value with the largest frequency.
- Sum: $\sum_{i=1}^{n} x_i$.
- Sample variance: $Var(X) = \dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$.
- Standard deviation: $S = \sqrt{Var(X)}$.



- IQR: $x_{0.75} - x_{0.25}$.
- Range: $\max\{X\} - \min\{X\}$.
- Sample skewness: $\dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}}$.
- Sample kurtosis: $\dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{4/2}}$.

Try them on Latex!

Robustness of Parameters

Q: Do outliers affect the quantities?

If the quantity is not affected by outliers, then it is a robust quantity.

Example:

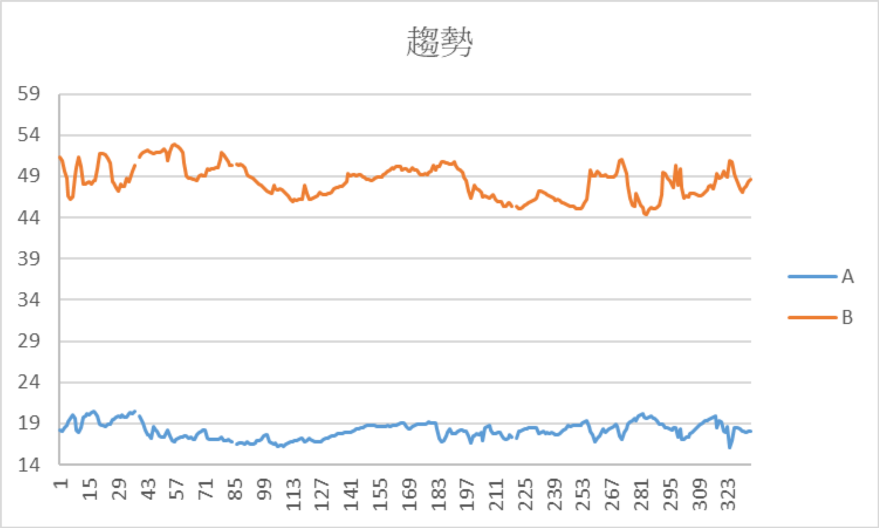
Set 1: 4, 4.5, 5, 6, 6.5, 7, 7.5, 8, 8.5, 9, 9.5, 10, 12, 15, **19**

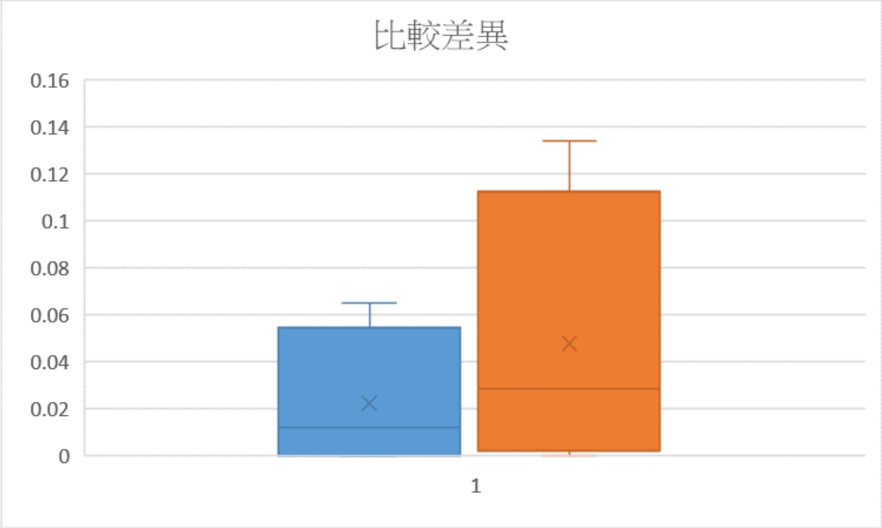
Set 2: 4, 4.5, 5, 6, 6.5, 7, 7.5, 8, 8.5, 9, 9.5, 10, 12, 15, **30**

	Set 1	Set 2	Robust?
Mean	8.77	9.5	
Median	8	8	V
Range	15	26	
IQR	3.5	3.5	V
Variance	16.39	40.54	

Purpose: Comparison

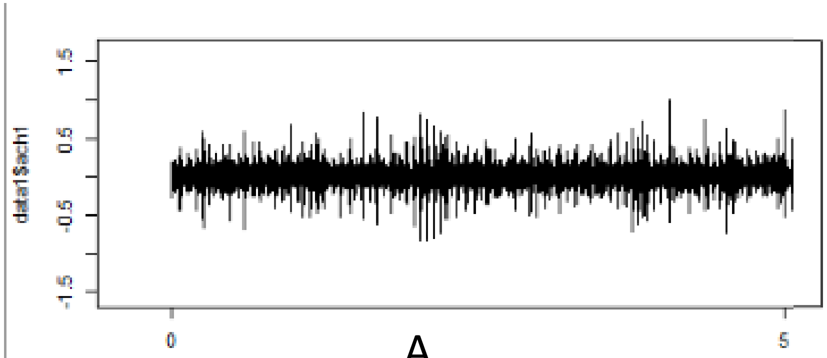
Purpose: Comparison



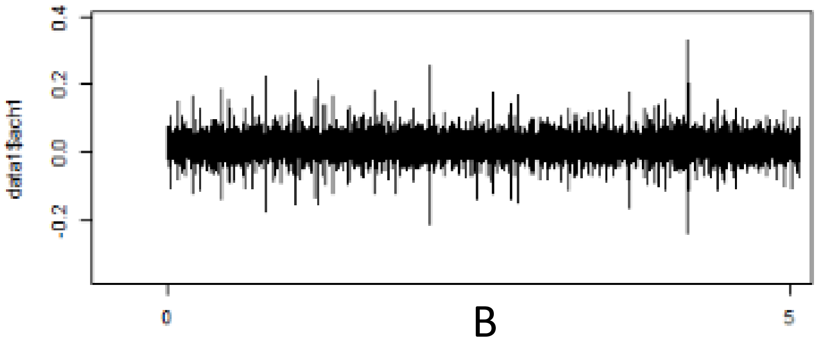


Comparison via figures

Given: A is the abnormal machine，B is the normal machine



A



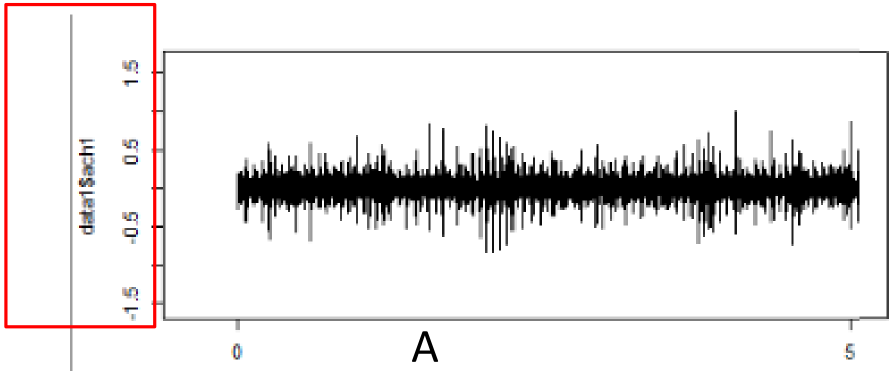
B

Conclusion：There are is difference between A and B.

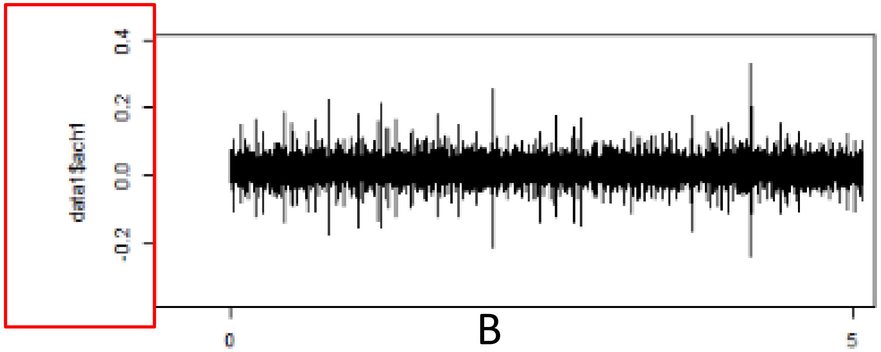
???

Misleading via figures

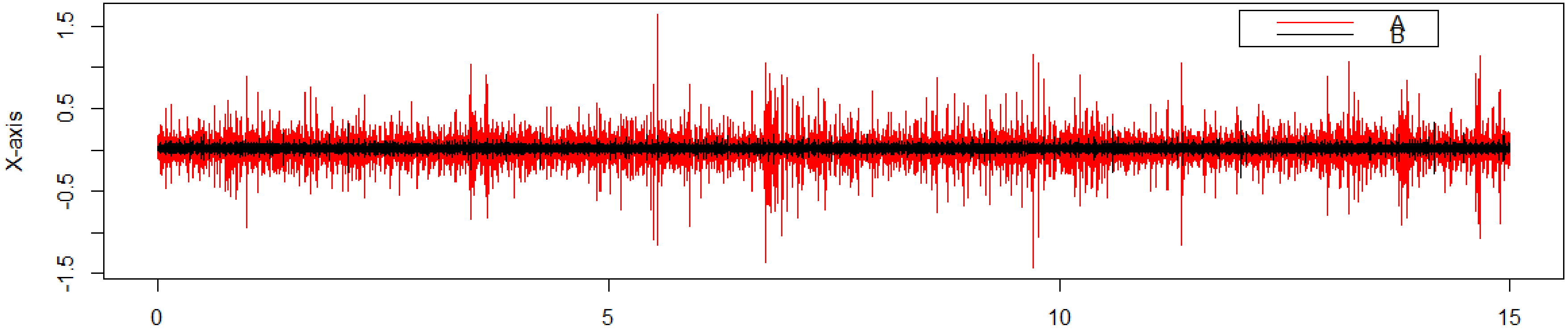
Given: A is the abnormal machine， B is the normal machine



-1.5 ~ 1.5

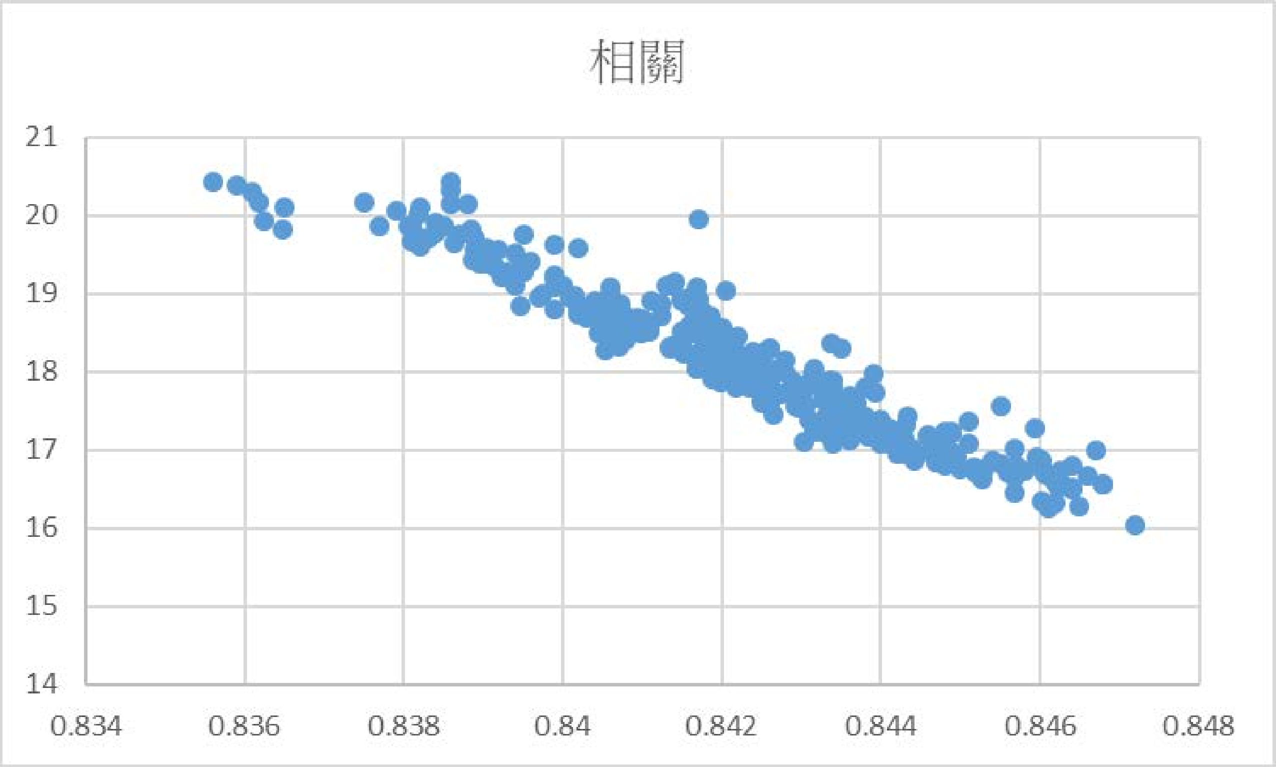


-0.2 ~ 0.4



Conclusion：There is obvious difference between A and B from figures.

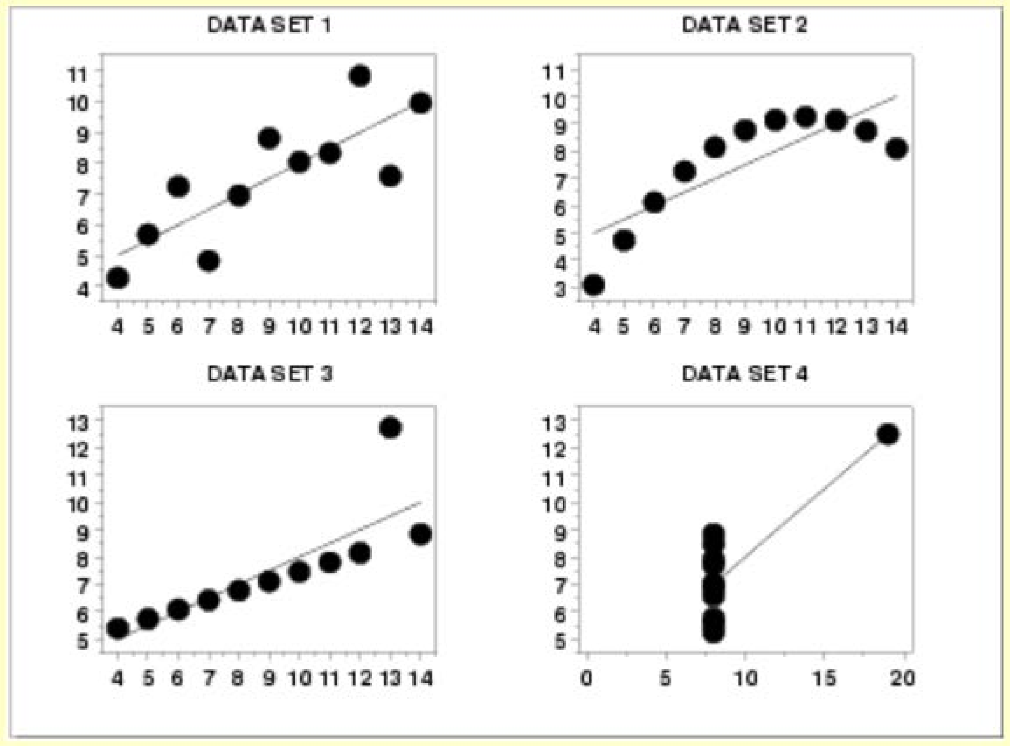
Relationship
Scatter plot I



Relationship
Scatter plot II

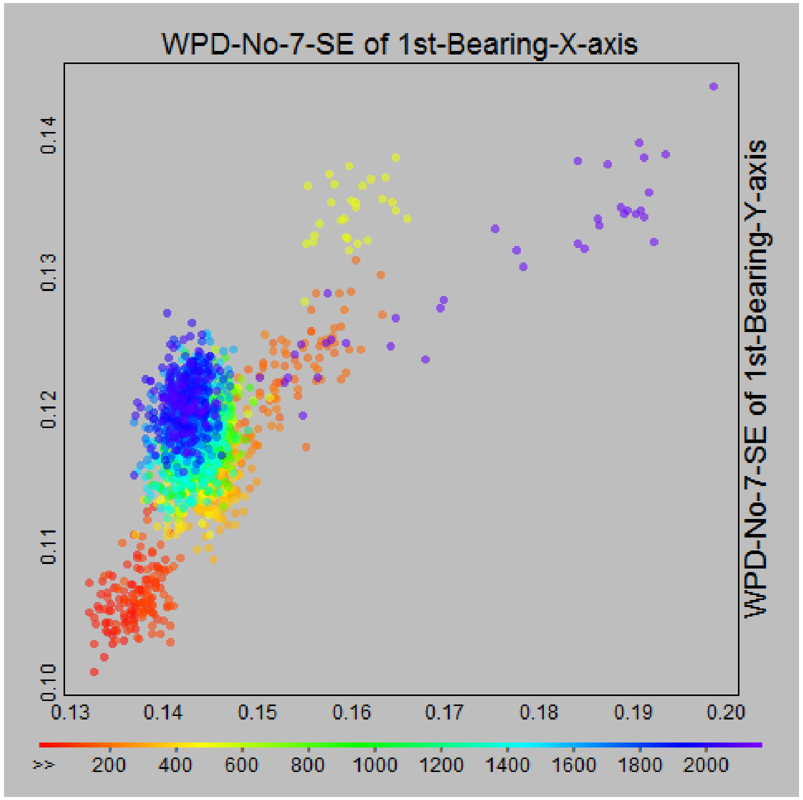
1.1.6. An EDA/Graphics Example (nist.gov)

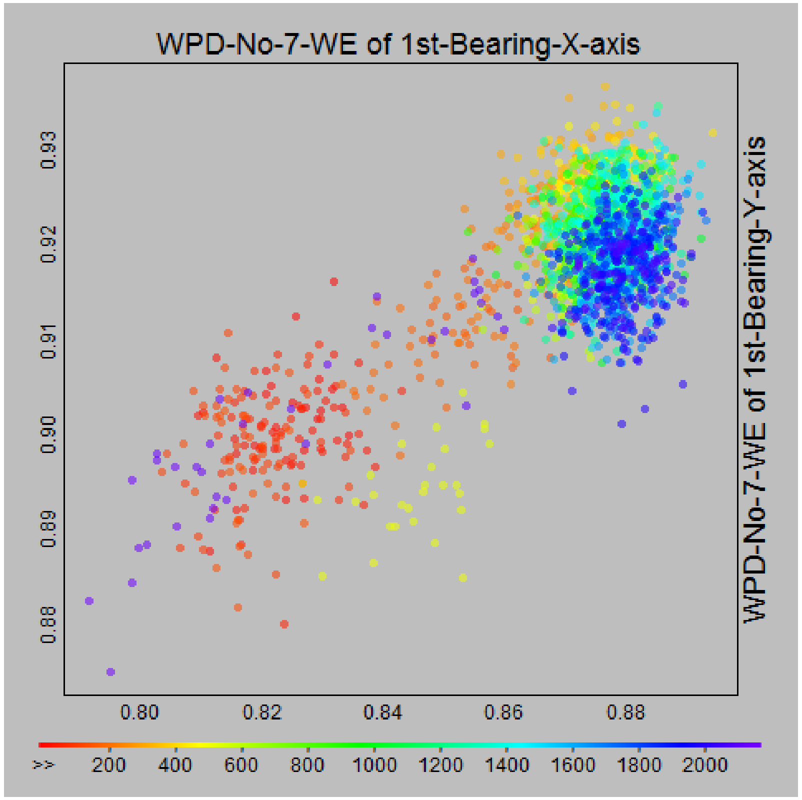
- Summary
- EDA



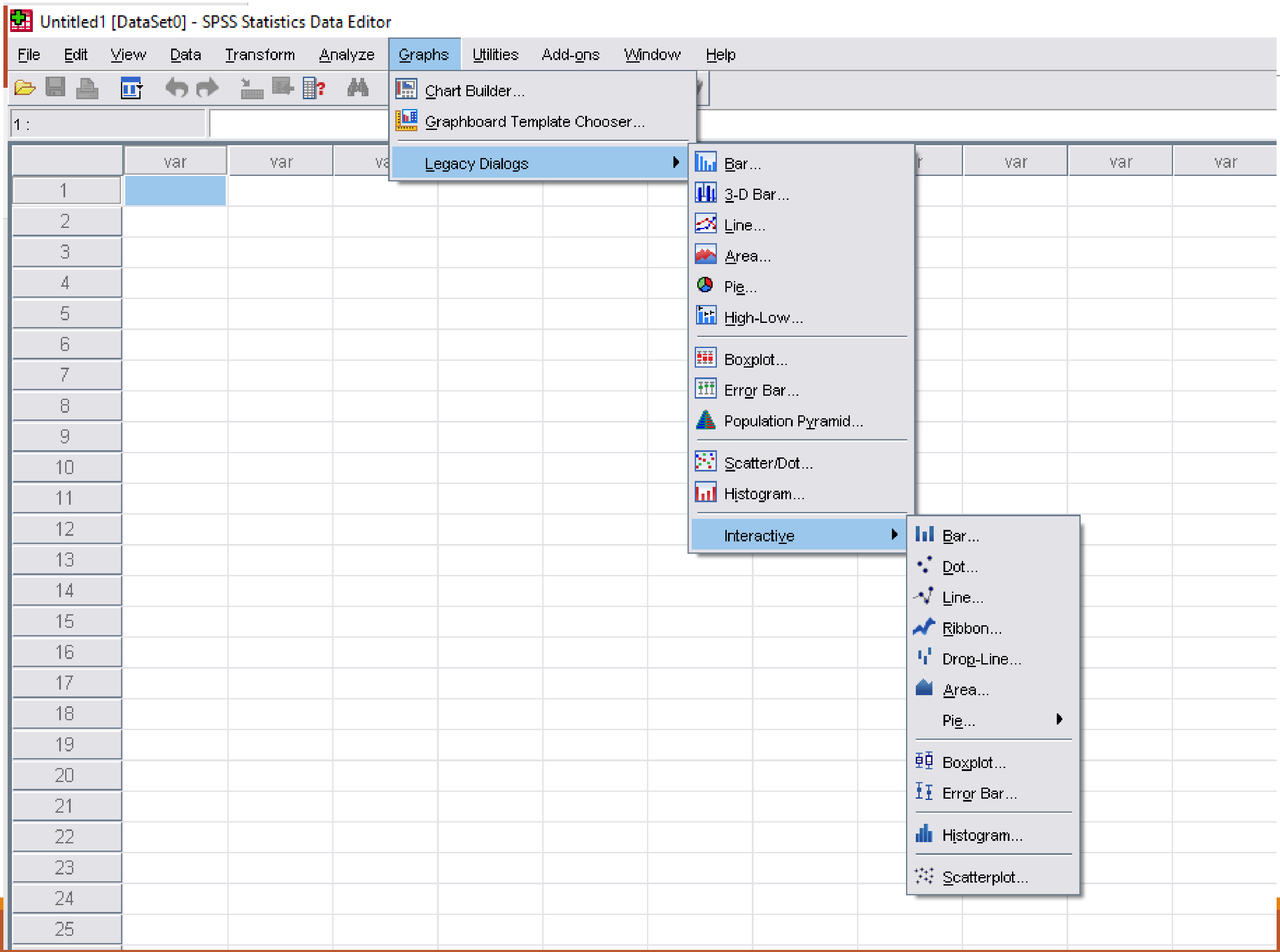
Relationship and classification Scatter plot III

Use multiple colors or symbols to identify groups/labels.



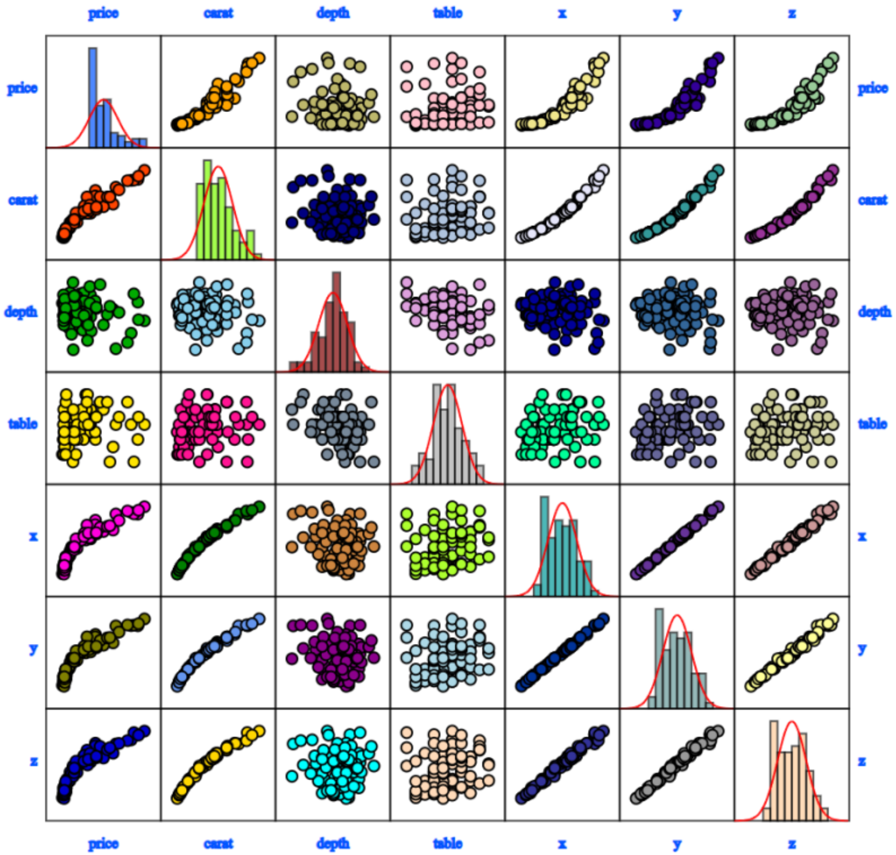


In SPSS


Untitled1 [DataSet0] - SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help
Chart Builder...
Graphboard Template Chooser...
Legacy Dialogs
Bar...
3-D Bar...
Line...
Area...
Pie...
High-Low...
Boxplot...
Error Bar...
Population Pyramid...
Scatter/Dot...
Histogram...
Interactive
Bar...
Dot...
Line...
Ribbon...
Drop-Line...
Area...
Pie...
Boxplot...
Error Bar...
Histogram...
Scatterplot...
1 :
var
var
var
var
var
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25

Diamond Dataset

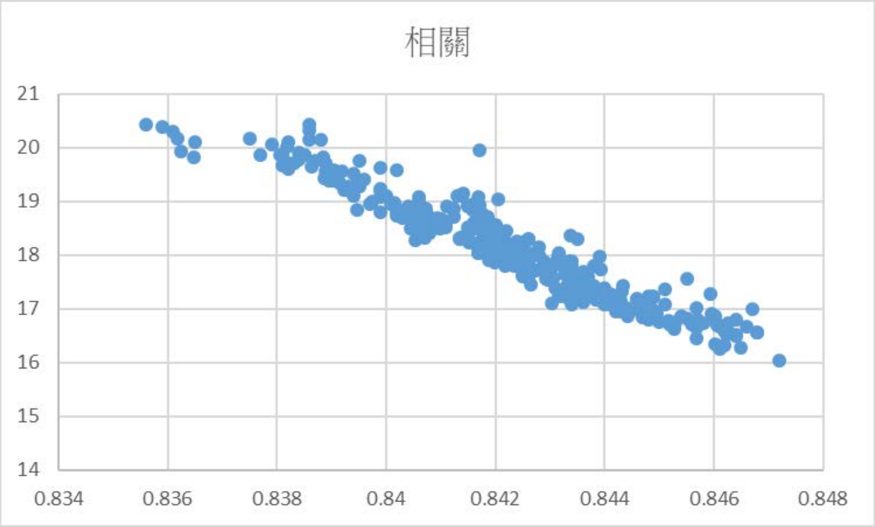
- Distribution
- Relationship
- Pattern
- Outliers

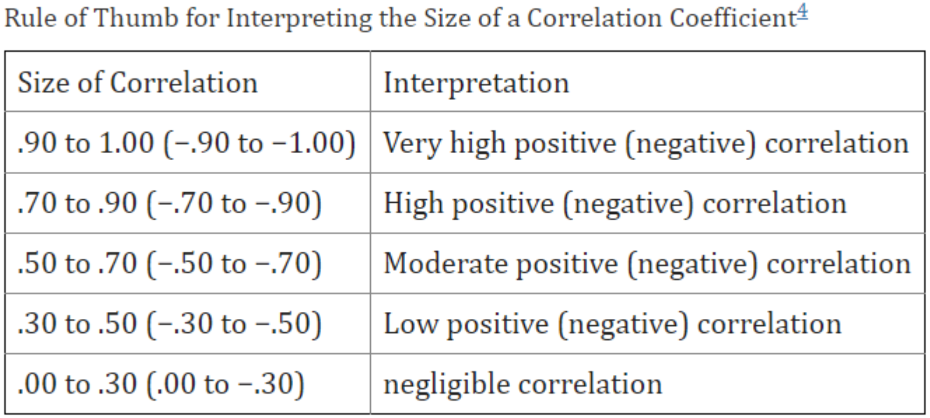


More about correlation

Relationship
Correlation coefficient

- $r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$
- $r \in [-1, 1]$

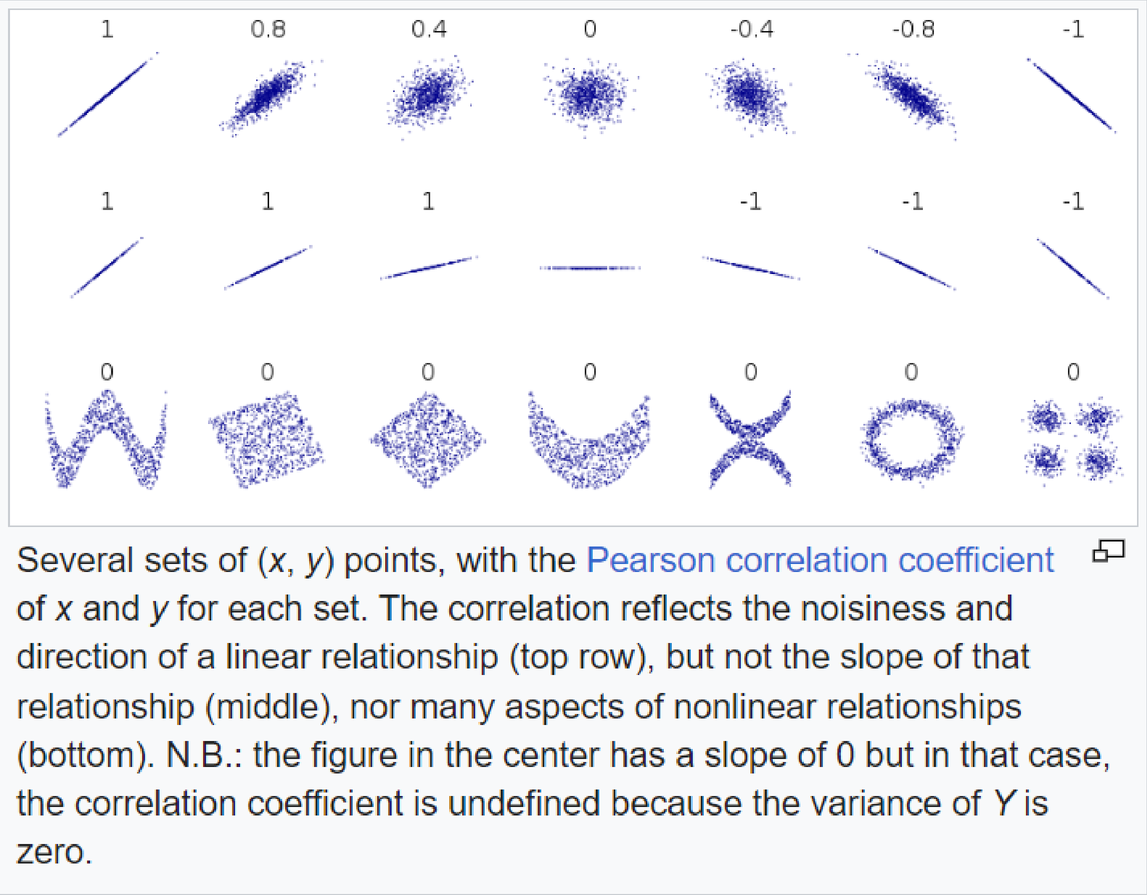



Rule of Thumb for Interpreting the Size of a Correlation Coefficient[4]

Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. Malawi medical journal, 24(3), 69-71.

相關係數與散佈圖



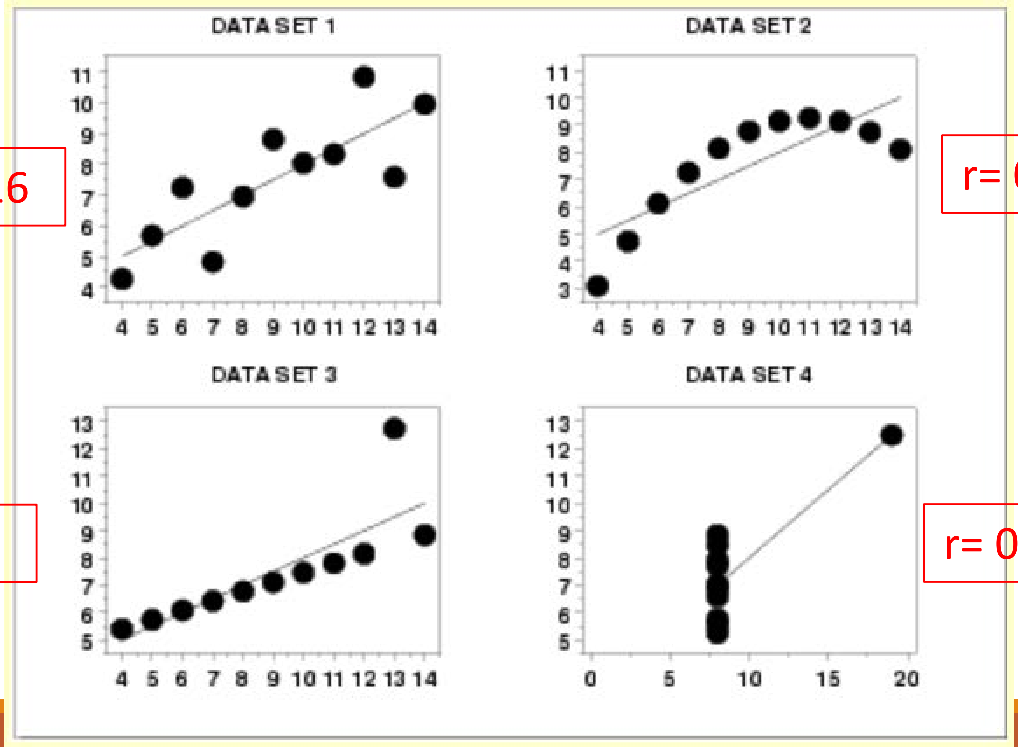
Several sets of (*x*, *y*) points, with the Pearson correlation coefficient of *x* and *y* for each set. The correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case, the correlation coefficient is undefined because the variance of *Y* is zero.

https://en.wikipedia.org/wiki/Correlation

散佈圖
Correlation (相關係數)

1.1.6. An EDA/Graphics Example (nist.gov)

- Summary
- EDA



r= 0.816

r= 0.816

r= 0.816

r= 0.817

Assignment

Use data “iris” in R to show

1. What are the types of variables?
2. For each variables, give the histogram/barplot and boxplot.
3. Use a table to summarize variables with the descriptive statistics including mean, median, variance, standard deviation, range, IQR, skewness, and kurtosis.
4. Are all cardinal variables symmetric?
5. Is there any outlier?