

R 語言問卷分析(1)

問卷資料檔處理
資料檢核與轉換
敘述統計

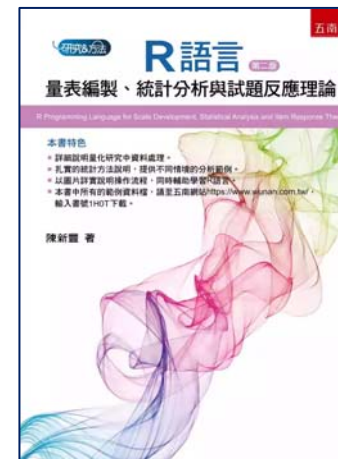
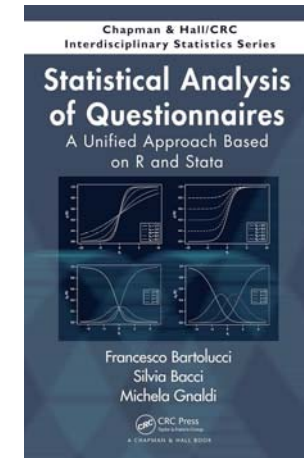
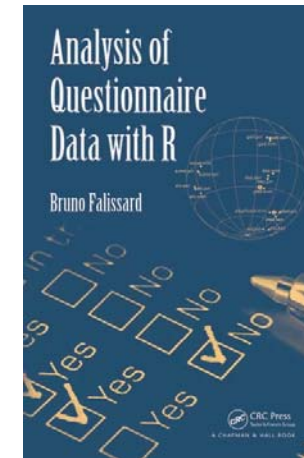
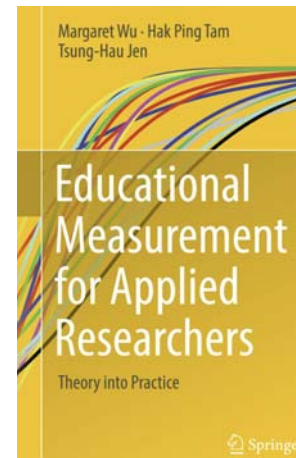
吳漢銘

國立政治大學 統計學系



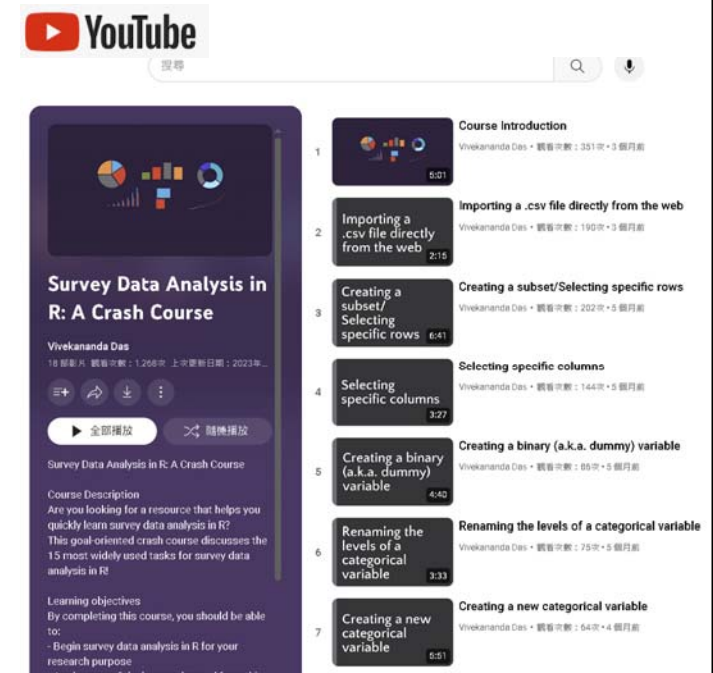
<https://hmwu.idv.tw>

- 資料檔管理與轉換
- 資料檢核與轉換
- 敘述統計量



本講義部份內容掃描自以下所列之上課用書，並無它用。
吳明隆, SPSS操作與應用：問卷統計分析實務（附光碟） 五南出版社

- Questionnaires and Surveys: Analyses with R
<https://ladal.edu.au/surveys.html>
- Survey Data Analysis with R
<https://stats.oarc.ucla.edu/r/seminars/survey-data-analysis-with-r/>
- Essentials for Analyzing Survey Data Using R
<https://rpubs.com/kiffercard/Essentials-for-Analyzing-Survey-Data-Using-R>
- Analyzing Survey Data in R
<https://rpubs.com/Onduma/surveydata>
- Getting Started with R for Survey Analysis
https://www2.hawaii.edu/~georgeha/Handouts/R_Handout_Getting_Started_with_R.html
- Analyzing Survey Data in R: A Crash Course (Part 1)
<https://vivasdas.medium.com/analyzing-survey-data-in-r-a-crash-course-part-1-9dfa4b110115>
- Survey Data Analysis in R: A Crash Course
https://www.youtube.com/playlist?list=PLLnN822fMufOrlQpfGW_7vvZF34Ppu50d





- Continuous Data 連續型資料:
 - 年收入、年資、身高、... (quantitative 計量)
- Discrete (Categorical) Data 類別資料:
 - 性別、種族、教育程度、... (qualitative 屬質)
- "Ordinal" 順序變數，次序變數:
 - 非常同意，同意，普通，不同意，非常不同意
 - 優，佳，劣
- "Nominal" 名目變數:
 - 宗教信仰、交通工具、音樂類型
- Ordinal methods cannot be used with nominal variable
- Nominal methods can be used with nominal, ordinal variables.



問卷、試卷與量表的編製

5 / 100

試卷、問卷與量表都是用於收集資料的工具，但它們各有不同的用途和特點：

1. 試卷 (Test or Exam)：

- 目的：試卷主要用於測試知識、技能或能力，常見於教育和專業認證領域。
- 特點：試卷通常有標準答案和評分標準，用以評估受試者的表現水平。

2. 問卷 (Questionnaire)：

- 目的：問卷用於廣泛收集資訊，如個人的意見、態度、感受或行為等。
- 特點：問卷可以包括開放性問題和封閉性問題，並且結構可以很靈活，適用於定性或定量研究。

3. 量表 (Scale)：

- 目的：量表專注於評估特定變數的程度或強度（如李克特量表），常用於心理測量和醫學研究，如測量焦慮、滿意度或其他心理特徵。
- 特點：量表通常包含一系列相關問題，透過統計方法來測量特定的心理或行為特徵，並有明確的計分方式。

	應用範圍	設計重點	計分系統
試卷	主要用於教育和測試	強調評分的公正和標準化	有固定答案和評分標準
問卷	調查研究，收集多樣化的資料	強調問題的多樣性和覆蓋範圍	可能沒有統一的計分方式
量表	專注於特定特徵的測	強調測量的精確性和信度	有明確的計分系統來評估特定狀態或特性



問卷/量表編製的方法及步驟

6 / 100

- **擬定編製量表的計畫:** 預算、樣本、完成時間等。
- **確定主題:** 清晰定義希望通過問卷獲得什麼資訊，有助於聚焦問卷的內容。
- **蒐集資料、確定目標群體:** 確定目標受眾。這將影響問卷的設計，包括語言的選擇和問題的難易程度。
- **擬定量表的架構、設計問題 (編製題目):**
 - 封閉式問題: 提供預設的答案選項，便於量化分析。例如，單選題或是多選題。
 - 開放式問題: 允許受訪者以自由形式回答填寫，適合收集質性資料。
 - 問題的順序: 合理安排問題的順序可以幫助提高回答率和質量。通常將簡單或不敏感的問題放在前面，逐漸過渡到更複雜或私人的問題。
- **問卷的預試:** 在問卷發放之前，進行預測調查以檢測問題的表述是否明確，答案選項是否恰當，並確保問卷能夠有效地測量預期的指標。
- **項目分析、編製正式題目**
- **問卷的執行:** 決定問卷的發放方式 (例如，紙質問卷、在線問卷等)，並設計有效的收集和追蹤回應的策略。
- **建立信度與效度:**
 - 信度的考驗: 穩定性係數(重測信度)、內部一致性係數(Cronbach α 、折半信度)
 - 效度的考驗: 效標關聯效度、團體差異的分析、因素分析
- **資料分析:** 量化分析及質性資料分析。
- **報告撰寫:** 根據問卷調查的結果撰寫報告。
報告應該包括研究的背景、方法、主要發現、結論和建議。

大約要比預定的題數多編二分之一的題目。如一個分量表若需要10題，此時就需編15題。

提出問題的標準

- 問題是否與研究目的一致
- 問題的類型是否合適(開放/封閉式)
- 問題是否令人難以回答(敏感問題)
- 問題是否涉及個人的隱私
- 問題是否有暗示作用
- 問題是否超出作答者的能力

編製題目的原則

- 用字淺顯易懂
- 每個問題只涵蓋一個觀念
- 避免主觀及情緒化的字眼
- 問題的選項應清楚界定
- 不用假設或猜測的語句
- 句子避免過長

如何擬定量表的架構

- 決定量表的因素(向度、分量表)
- 決定預編的題數、正式量表的題數
- 決定量表的量尺(五點、六點等)

分析流程：建立資料檔

7 / 100

一、基本資料

1. 我的班級：☐ 甲班 ☐ 乙班

2. 我的性別：☐ 男生 ☐ 女生

3. 第二定期考查的數學成績：_____分

二、數學學習問卷

符合程度
低 ← → 高

1. 我會努力去面對具有挑戰性的數學題目	☐	☐	☐	☐	☐
2. 同學對我數學學習上的肯定，使我更喜歡數學	☐	☐	☐	☐	☐
3. 課堂上的數學題目我大多做得出來	☐	☐	☐	☐	☐
4. 我喜愛參與數學課堂中的學習活動	☐	☐	☐	☐	☐
5. 上數學課時我的精神特別好	☐	☐	☐	☐	☐
6. 遇到較為困難的數學題目時，我不會逃避	☐	☐	☐	☐	☐
7. 我有能力幫助同學解答相關的數學問題	☐	☐	☐	☐	☐
8. 老師對我的數學學習能力與態度十分肯定	☐	☐	☐	☐	☐

Microsoft Excel - 數學效能_1.xls

檔案(E) 編輯(E) 檢視(V) 插入(I) 格式(O) 工具(T) 資料(Q) 視

Arial 10 B I U

	A	B	C	D	E	F	G	H	I	J	K
1	班級	性別	數學成就	a1	a2	a3	a4	a5	a6	a7	a8
2	1	1	60	5	1	5	2	1	5	3	3
3	1	1	42	5	2	5	2	2	4	4	2
4	1	1	78	5	2	5	2	2	3	3	1
5	1	2	65	5	2	5	2	2	4	3	4
6	1	2	68	1	1	5	2	3	5	2	5
7	1	1	57	4	2	5	2	2	4	3	6
8	1	1	55	4	2	5	2	2	5	4	4
9	1	1	97	4	2	5	2	3	4	3	3
10	1	2	87	1	2	5	3	2	3	5	3
11	1	2	92	4	2	4	3	2	4	5	6
12	1	2	75	1	2	4	3	4	3	4	2
13	1	1	55	2	2	4	3	5	5	4	5
14	1	1	64	2	2	4	3	5	4	3	5
15	1	1	71	2	2	4	2	5	3	4	5
16	1	1	78	2	2	4	2	3	4	5	5
17	1	2	84	2	2	4	2	4	5	4	5

數學效能_1/

就緒

```

> library(readxl)
> math_data <- read_excel("data/數學效能_1.xlsx")
> head(math_data)
# A tibble: 6 × 11
  班級  性別 數學成就  a1  a2  a3  a4  a5  a6  a7  a8
<dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1     1     1     60     5     1     5     2     1     5     3     3
2     1     1     42     5     2     5     2     2     4     4     2
...
> tail(math_data)
# A tibble: 6 × 11
  班級  性別 數學成就  a1  a2  a3  a4  a5  a6  a7  a8
<dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
...
4     2     1     89     5     5     5     5     2     4     4     5
5     2     2     87     5     5     5     2     2     4     5     5
6     2     2     76     5     5     4     5     3     4     4     5
> str(math_data)
tibble [50 × 11] (S3: tbl_df/tbl/data.frame)
 $ 班級      : num [1:50] 1 1 1 1 1 1 1 1 1 1 ...
 $ 性別      : num [1:50] 1 1 1 2 2 1 1 1 2 2 ...
 $ 數學成就: num [1:50] 60 42 78 65 68 57 55 97 87 92 ...
 $ a1        : num [1:50] 5 5 5 5 1 4 4 4 1 4 ...
 ...
 $ a8        : num [1:50] 3 2 1 4 5 6 4 3 3 6 ...
> summary(math_data)
  班級      性別      數學成就      a1      a2      a3
Min.   :1.00  Min.   :1.00  Min.   : 42.00  Min.   :1.00  Min.   :1.00  Min.   :1.0
1st Qu.:1.00  1st Qu.:1.00  1st Qu.: 65.50  1st Qu.:2.00  1st Qu.:1.00  1st Qu.:3.0
Median :2.00  Median :1.00  Median : 75.50  Median :3.50  Median :2.00  Median :4.0
Mean   :1.52  Mean   :1.44  Mean   : 75.78  Mean   :3.38  Mean   :1.96  Mean   :3.5
3rd Qu.:2.00  3rd Qu.:2.00  3rd Qu.: 87.00  3rd Qu.:4.00  3rd Qu.:2.00  3rd Qu.:4.0
Max.   :2.00  Max.   :2.00  Max.   :100.00  Max.   :5.00  Max.   :5.00  Max.   :5.0

```

- 兩資料檔具有共同變數: 配對變數
- 若有部份變數是另一個資料檔中所沒有的: 非配對變數
- 先開啟的檔: 作用中的資料檔
- 合併有兩種:
 - 觀察值合併 (垂直合併, **rbind**): 其變數應為「配對變數」, 將觀察值加在「作用中的資料檔」的後面
 - 變項合併 (水平合併, **cbind**): 需有一「關鍵變數」(例如: 編號、編碼值、...) , 新合併的資料檔中觀察值個數不變。
- 資料合併指令
 - Base R: **merge**
 - Tidyverse **dplyr**: **join{dplyr}**: **anti_join**, **full_join**, **inner_join**, **left_join**, **right_join**, **semi_join**

合併觀察值: rbind

10 / 100

```
> math_data1 <- read_excel("data/數學學習_1.xlsx")
> math_data1
# A tibble: 5 × 4
  編號 性別 數學成就 數學態度
  <chr> <dbl>    <dbl>    <dbl>
1 1001     1      87      38
2 1002     1      84      42
...
5 1005     2      92      40
> math_data2 <- read_excel("data/數學學習_2.xlsx")
> math_data2
# A tibble: 5 × 4
  編號 性別 數學成就 數學態度
  <chr> <dbl>    <dbl>    <dbl>
1 1006     2      74      36
2 1007     2      94      34
...
5 1010     1      70      54
> math_data_all <- rbind(math_data1, math_data2)
> math_data_all
# A tibble: 10 × 4
  編號 性別 數學成就 數學態度
  <chr> <dbl>    <dbl>    <dbl>
1 1001     1      87      38
2 1002     1      84      42
3 1003     1      75      25
4 1004     2      78      33
5 1005     2      92      40
6 1006     2      74      36
7 1007     2      94      34
8 1008     2      85      28
9 1009     1      68      30
10 1010     1      70      54
```

合併「數學學習_1.xlsx」和「數學學習_2.xlsx」

	編號	性別	數學成就	數學態度
1	1001	1	87	38
2	1002	1	84	42
3	1003	1	75	25
4	1004	2	78	33
5	1005	2	92	40

	編號	性別	數學成就	數學態度
1	1006	2	74	36
2	1007	2	94	34
3	1008	2	85	28
4	1009	1	68	30
5	1010	1	70	54

```
> math_data2_tmp <- math_data2[, sample(1:4)]
> rbind(math_data1, math_data2_tmp)
# A tibble: 10 × 4
  編號 性別 數學成就 數學態度
  <chr> <dbl>    <dbl>    <dbl>
1 1001     1      87      38
2 1002     1      84      42
3 1003     1      75      25
...
10 1010     1      70      54
>
> math_data2_tmp$TEST <- sample(1:5)
> math_data2_tmp
# A tibble: 5 × 5
  編號 性別 數學態度 數學成就 TEST
  <chr> <dbl>    <dbl>    <dbl> <int>
1 1006     2      36      74      5
2 1007     2      34      94      3
...
4 1009     1      30      68      1
5 1010     1      54      70      2
> rbind(math_data1, math_data2_tmp)
Error in rbind(deparse.level, ...) :
  numbers of columns of arguments do not match
```

合併變數: cbind

11 / 100

合併「數學學習_1.xls」、「數學學習_3.xls」。

```
> math_data1 <- read_excel("data/數學學習_1.xlsx")
> math_data1
```

```
# A tibble: 5 × 4
```

```
  編號  性別 數學成就 數學態度
  <chr> <dbl>   <dbl>   <dbl>
1 1001     1     87      38
2 1002     1     84      42
```

```
..
```

```
5 1005     2     92      40
```

```
> math_data3 <- read_excel("data/數學學習_3.xlsx")
```

```
> math_data3
```

```
# A tibble: 5 × 4
```

```
  編號 數學效能 數學焦慮 數學投入
  <chr>   <dbl>   <dbl>   <dbl>
1 1001     54     48      25
2 1002     32     38      40
```

```
..
```

```
5 1005     44     45      28
```

```
>
```

```
> cbind(math_data1, math_data3)
```

```
  編號 性別 數學成就 數學態度 編號 數學效能 數學焦慮 數學投入
1 1001     1     87      38 1001     54     48      25
2 1002     1     84      42 1002     32     38      40
3 1003     1     75      25 1003     48     46      32
4 1004     2     78      33 1004     42     41      19
5 1005     2     92      40 1005     44     45      28
```

	編號	性別	數學成就	數學態度		編號	數學效能	數學焦慮	數學投入
1	1001	1	87	38	1	1001	54	48	25
2	1002	1	84	42	2	1002	32	38	40
3	1003	1	75	25	3	1003	48	46	32
4	1004	2	78	33	4	1004	42	41	19
5	1005	2	92	40	5	1005	44	45	28

```
> math_data3_tmp <- math_data3[sample(1:5), ]
```

```
> math_data3_tmp
```

```
# A tibble: 5 × 4
```

```
  編號 數學效能 數學焦慮 數學投入
  <chr>   <dbl>   <dbl>   <dbl>
1 1005     44     45      28
```

```
...
```

```
5 1001     54     48      25
```

```
> cbind(math_data1, math_data3_tmp)
```

```
  編號 性別 數學成就 數學態度 編號 數學效能 數學焦慮 數學投入
1 1001     1     87      38 1005     44     45      28
2 1002     1     84      42 1003     48     46      32
3 1003     1     75      25 1004     42     41      19
4 1004     2     78      33 1002     32     38      40
5 1005     2     92      40 1001     54     48      25
```

```
merge(math_data1, math_data3, by = "編號")
```



rbind and cbind

12 / 100

```
> begin.experiment <- data.frame(name=c("A", "B", "C", "D", "E", "F"),  
+ weights=c(270, 263, 294, 218, 305, 261))  
> middle.experiment <- data.frame(name=c("G", "H", "I"),  
+ weights=c(169, 181, 201))  
> end.experiment <- data.frame(name=c("C", "D", "A", "H", "I"),  
+ weights=c(107, 104, 104, 102, 100))  
> # merge the data for those who started and finished the experiment  
> (common <- intersect(begin.experiment$name, end.experiment$name))  
[1] "A" "C" "D"  
> (b.at <- is.element(begin.experiment$name, common))  
[1] TRUE FALSE TRUE TRUE FALSE FALSE  
> (e.at <- is.element(end.experiment$name, common))  
[1] TRUE TRUE TRUE FALSE FALSE  
> experiment <- rbind(cbind(begin.experiment[b.at,], time="begin"),  
+ cbind(end.experiment[e.at,], time="end"))  
> experiment  
  name weights  time  
1    A    270 begin  
3    C    294 begin  
4    D    218 begin  
11   C    107  end  
2    D    104  end  
31   A    104  end
```

```
> tapply(experiment$weights, experiment$time, mean)  
      begin      end  
260.6667 105.0000
```

```
> begin.experiment  
  name weights  
1    A    270  
2    B    263  
3    C    294  
4    D    218  
5    E    305  
6    F    261  
> middle.experiment  
  name weights  
1    G    169  
2    H    181  
3    I    201  
> end.experiment  
  name weights  
1    C    107  
2    D    104  
3    A    104  
4    H    102  
5    I    100
```



merge {base}: Merge Two Data Frames

- Merge (adds variables to a dataset) two data frames horizontally by common columns or row names (key variables, either string or numeric). , or do other versions of database join operations.

```
merge(x, y, by = intersect(names(x), names(y)),  
      by.x = by, by.y = by, all = FALSE, all.x = all, all.y = all,  
      sort = TRUE, suffixes = c(".x", ".y"),  
      incomparables = NULL, ...)
```

```
# merge two data frames by ID  
total <- merge(data.frame.A, data.frame.B, by="ID")  
  
# merge two data frames by ID and Country  
total <- merge(data.frame.A, data.frame.B, by=c("ID", "Country"))
```



merge {base}, 合併的準則

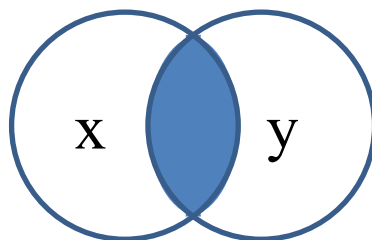
14 / 100

- **merge{base}** 合併的準則:
 - (default) the data frames are merged on the columns with names they both have.
 - The rows in the two data frames that match on the specified columns are extracted, and joined together.
 - If there is more than one match, all possible matches contribute one row each.
- **merge{base}** 重要Arguments:
 - **by, by.x, by.y**: The names of the columns that are common to both **x** and **y**. The default is to use the columns with common names between the two data frames.
 - **all.x, (all.y)**: logical;
 - if **TRUE**, then extra rows will be added to the output, one for each row in **x** that has no matching row in **y**.
 - These rows will have NAs in those columns that are usually filled with values from **y**.
 - The default is **FALSE**, so that only rows with data from both **x** and **y** are included in the output.
 - **all = TRUE (FALSE)** is shorthand for **all.x = TRUE (FALSE)** and **all.y = TRUE (FALSE)**. Logical values that specify the type of merge.
 - The default value is **all=FALSE** (meaning that only the matching rows are returned).

Different Types of Merge

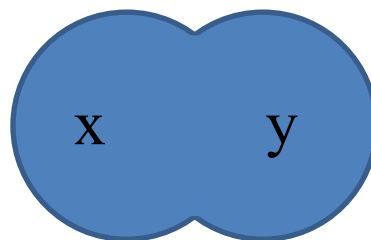
- **Natural join:** To keep only rows that match from the data frames, specify the argument **all=FALSE** (by default).
- **Full outer join:** To keep all rows from both data frames, specify **all=TRUE**. Note that this performs the complete merge and fills the columns with NA values where there is no matching data.
- **Left outer join:** To include all the rows of your data frame **x** and only those from **y** that match, specify **all.x=TRUE**.
- **Right outer join:** To include all the rows of your data frame **y** and only those from **x** that match, specify **all.y=TRUE**.

all=FALSE



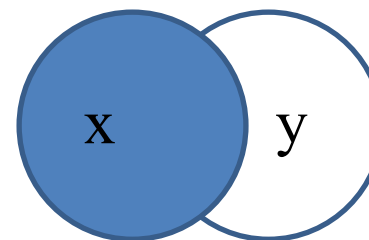
natural join

all=TRUE



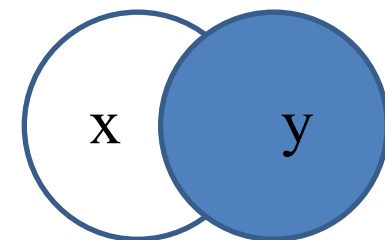
full outer join

all.x=TRUE



left outer join

all.y=TRUE



right outer join

dplyr: inner_join

full_join

left_join

right_join

Example (1)

```
> authors <- data.frame(
+   surname = I(c("Tukey", "Venables", "Tierney", "Ripley", "McNeil")),
+   nationality = c("US", "Australia", "US", "UK", "Australia"),
+   deceased = c("yes", rep("no", 4)))
> books <- data.frame(
+   name = I(c("Tukey", "Venables", "Tierney",
+             "Ripley", "Ripley", "McNeil", "R Core")),
+   title = c("Exploratory Data Analysis",
+             "Modern Applied Statistics ...",
+             "LISP-STAT",
+             "Spatial Statistics", "Stochastic Simulation",
+             "Interactive Data Analysis",
+             "An Introduction to R"),
+   other.author = c(NA, "Ripley", NA, NA, NA, NA, "Venables & Smith"))
> authors
  surname nationality deceased
1   Tukey           US      yes
2 Venables Australia      no
3 Tierney           US      no
4 Ripley           UK      no
5 McNeil    Australia      no
> books
  name          title      other.author
1  Tukey Exploratory Data Analysis      <NA>
2 Venables Modern Applied Statistics ... Ripley
3 Tierney           LISP-STAT      <NA>
4 Ripley      Spatial Statistics      <NA>
5 Ripley      Stochastic Simulation      <NA>
6 McNeil Interactive Data Analysis      <NA>
7  R Core  An Introduction to R Venables & Smith
```

<https://stat.ethz.ch/R-manual/R-devel/library/base/html/merge.html>

Example (1)

```
> (m1 <- merge(authors, books, by.x = "surname", by.y = "name"))
```

	surname	nationality	deceased	title	other.author
1	McNeil	Australia	no	Interactive Data Analysis	<NA>
2	Ripley	UK	no	Spatial Statistics	<NA>
3	Ripley	UK	no	Stochastic Simulation	<NA>
4	Tierney	US	no	LISP-STAT	<NA>
5	Tukey	US	yes	Exploratory Data Analysis	<NA>
6	Venables	Australia	no	Modern Applied Statistics ...	Ripley

```
> (m2 <- merge(books, authors, by.x = "name", by.y = "surname"))
```

	name	title	other.author	nationality	deceased
1	McNeil	Interactive Data Analysis	<NA>	Australia	no
2	Ripley	Spatial Statistics	<NA>	UK	no
3	Ripley	Stochastic Simulation	<NA>	UK	no
4	Tierney	LISP-STAT	<NA>	US	no
5	Tukey	Exploratory Data Analysis	<NA>	US	yes
6	Venables	Modern Applied Statistics ...	Ripley	Australia	no

```
> merge(authors, books, by.x = "surname", by.y = "name", all = TRUE)
```

	surname	nationality	deceased	title	other.author
1	McNeil	Australia	no	Interactive Data Analysis	<NA>
2	R Core	<NA>	<NA>	An Introduction to R	Venables & Smith
3	Ripley	UK	no	Spatial Statistics	<NA>
4	Ripley	UK	no	Stochastic Simulation	<NA>
5	Tierney	US	no	LISP-STAT	<NA>
6	Tukey	US	yes	Exploratory Data Analysis	<NA>
7	Venables	Australia	no	Modern Applied Statistics ...	Ripley

Example (2)

```
> (x <- data.frame(k1 = c(NA,NA,3,4,5), k2 = c(1,NA,NA,4,5), data = 1:5))
```

```
  k1 k2 data
```

```
1 NA  1    1
```

```
2 NA NA    2
```

```
3  3 NA    3
```

```
4  4  4    4
```

```
5  5  5    5
```

```
> (y <- data.frame(k1 = c(NA,2,NA,4,5), k2 = c(NA,NA,3,4,5), data = 1:5))
```

```
  k1 k2 data
```

```
1 NA NA    1
```

```
2  2 NA    2
```

```
3 NA  3    3
```

```
4  4  4    4
```

```
5  5  5    5
```

```
> merge(x, y, by = c("k1","k2")) # NA's match
```

```
  k1 k2 data.x data.y
```

```
1  4  4      4      4
```

```
2  5  5      5      5
```

```
3 NA NA      2      1
```

```
> merge(x, y, by = "k1") # NA's match, so 6 rows
```

```
  k1 k2.x data.x k2.y data.y
```

```
1  4  4      4      4      4
```

```
2  5  5      5      5      5
```

```
3 NA  1      1     NA      1
```

```
4 NA  1      1      3      3
```

```
5 NA NA      2     NA      1
```

```
6 NA NA      2      3      3
```

```
> merge(x, y, by = "k2", incomparables = NA) # 2 rows
```

```
  k2 k1.x data.x k1.y data.y
```

```
1  4  4      4      4      4
```

```
2  5  5      5      5
```



merge {base}, Example (3)

19 / 100

```
> stories <- read.table(header=TRUE, text='
+   storyid  title
+   1       lions
+   2       tigers
+   3       bears
+ ')
> data <- read.table(header=TRUE, text='
+   subject storyid rating
+   1       1       6.7
+   1       2       4.5
+   1       3       3.7
+   2       2       3.3
+   2       3       4.1
+   2       1       5.2
+ ')
>
> merge(stories, data, by="storyid")
  storyid title subject rating
1       1 lions       1   6.7
2       1 lions       2   5.2
3       2 tigers      1   4.5
4       2 tigers      2   3.3
5       3 bears       1   3.7
6       3 bears       2   4.1
```

```
> stories2 <- read.table(header=TRUE, text='
+   id       title
+   1       lions
+   2       tigers
+   3       bears
+ ')
>
> merge(stories2, data, by.x="id", by.y="storyid")
  id title subject rating
1  1 lions       1   6.7
2  1 lions       2   5.2
3  2 tigers      1   4.5
4  2 tigers      2   3.3
5  3 bears       1   3.7
6  3 bears       2   4.1
```

http://www.cookbook-r.com/Manipulating_data/Merging_data_frames/

Merge on Multiple Columns

```
> animals <- read.table(header=T, text='
+   size type      name
+   small  cat      lynx
+   big    cat      tiger
+   small  dog    chihuahua
+   big    dog "great dane"
+ ')
>
> observations <- read.table(header=T, text='
+   number size type
+       1   big  cat
+       2 small  dog
+       3 small  dog
+       4   big  dog
+ ')
>
> merge(observations, animals, c("size","type"))
  size type number      name
1  big  cat      1      tiger
2  big  dog      4 great dane
3 small  dog      2  chihuahua
4 small  dog      3  chihuahua
```



問卷編碼範例：基本資料

21 / 100

高中職學校行政主管時間管理現況及其策略運用調查問卷

親愛的教育先進：您好！

【說明】

研究生 ○○○ 敬上

一、基本資料

1. 性別：☐ (1) 男 ☐ (2) 女
2. 年齡：☐ (1) 30 歲以下 ☐ (2) 31-40 歲 ☐ (3) 41-50 歲 ☐ (4) 51-60 歲
☐ (5) 61 歲以上
3. 婚姻：☐ (1) 未婚 ☐ (2) 已婚 ☐ (3) 離異 ☐ (4) 喪偶
4. 最高學歷：☐ (1) 專科(含)以下 ☐ (2) 大學 ☐ (3) 研究所 40 學分班
☐ (4) 碩士 ☐ (5) 博士
5. 服務年資：☐ (1) 5 年以下 ☐ (2) 6-10 年 ☐ (3) 11-15 年
☐ (4) 16-20 年 ☐ (5) 21-25 年 ☐ (6) 26 年以上
6. 學校屬性：☐ (1) 公立 ☐ (2) 私立
7. 學校類別：☐ (1) 高中 ☐ (2) 高職
8. 學校規模(日間部)：☐ (1) 24 班以下 ☐ (2) 25-48 班 ☐ (3) 49 班以上

原始題項	變項名稱	變項標記	數值範圍	水準數值標記
一、基本資料				
1	性別		1-2	1:男生 2:女生
2	年齡		1-5	1:30 歲以下 2:31--40 歲 3:41--50 歲 4:51--60 歲 5:61 歲以上
3	婚姻		1-4	1:未婚 2:已婚 3:離異 4:喪偶
4	學歷		1-5	1:專科以下 2:大學 3:40 學分班 4:碩士 5:博士
5	服務年資		1-6	1:5 年以下 2:6--10 年 3:11--15 歲 4:16--20 年 5:21--25 歲 6:26 年以上
6	學校屬性		1-2	1:公立 2:私立
7	學校類別		1-2	1:高中 2:高職
8	學校規模		1-3	1:24 班以下 2:25--8 班 3:49 班以上

資料來源:吳明隆, SPSS操作與應用：問卷統計分析實務，五南出版社

二、時間管理認知

填答說明：請根據您的認知，在各題適當的□內打「✓」

	非常 同意	同 意	普 通	不 同 意	非 常 不 同 意
01.我覺得時間管理是每個人應具備的一種技巧。.....	☐	☐	☐	☐	☐
02.我認為時間管理是減輕壓力的一項重要因素。.....	☐	☐	☐	☐	☐
03.對時間使用的覺察與反省是改善時間管理的必要步驟。.....	☐	☐	☐	☐	☐
04.良好的時間管理者，會清楚自己的工作目標。.....	☐	☐	☐	☐	☐
05.我認為良好的時間管理，有助於提高生活品質。....	☐	☐	☐	☐	☐
06.善於時間管理的人，其能力更加使人信賴。.....	☐	☐	☐	☐	☐
07.善於時間管理的人，會更擅長於授權。.....	☐	☐	☐	☐	☐
08.善於時間管理的人，更能掌握突發事件。.....	☐	☐	☐	☐	☐
09.善於管理時間的人，會懂得運用人力資源。.....	☐	☐	☐	☐	☐
10.做好時間管理，會更能有效地去完成目標。.....	☐	☐	☐	☐	☐

原始題項	變項名稱	變項標記	數值範圍	水準數值標記
二、時間管理認知				
01	A1		1-5	
02	A2		1-5	
03	A3		1-5	
04	A4		1-5	
05	A5		1-5	

資料來源:吳明隆, SPSS操作與應用：問卷統計分析實務，五南出版社

三、時間分配

由下列項目中，排列出最能反應您平日工作時間分配的情況，請將數字依序填入□內，時間花費最多的填 1，其次填 2，以此類推……。

- ☐ 組織發展 ☐ 行政領導 ☐ 事務管理 ☐ 教學視導 ☐ 學生輔導
☐ 公共關係 ☐ 研習進修 ☐ 偶發事件 ☐ 其 他

原始題項	變項名稱	變項標記	數值範圍	水準數值標記
三、時間分配				
	B3_1	組織發展	1-9	
	B3_2	行政領導	1-9	
	B3_3	事務管理	1-9	
	B3_4	教學視導	1-9	
	B3_5	學生輔導	1-9	
	B3_6	公共關係	1-9	
	B3_7	研習進修	1-9	
	B3_8	偶發事件	1-9	
	B3_9	其他	1-9	

四、互動對象

除面對學生外，學校行政主管因為工作的關係，經常須與家長、社區民眾、長官或其他人士溝通。在下列項目中，請您依互動頻率多寡，將數字依序填入□內，互動頻率最高填1，其次填2，以此類推……。

- ☐ 上級長官 ☐ 校外夥伴 ☐ 學校同事 ☐ 學生家長 ☐ 社區民眾
☐ 民意代表 ☐ 同學朋友 ☐ 家人親戚 ☐ 其 他

原始題項	變項名稱	變項標記	數值範圍	水準數值標記
四、互動對象				
	C4_1	上級長官	1-9	
	C4_2	校外夥伴	1-9	
	C4_3	學校同事	1-9	
	C4_4	學生家長	1-9	
	C4_5	社區民眾	1-9	
	C4_6	民意代表	1-9	
	C4_7	同學朋友	1-9	
	C4_8	家人親戚	1-9	
	C4_9	其 他	1-9	

資料來源:吳明隆, SPSS操作與應用：問卷統計分析實務，五南出版社

五、困擾因素

在工作上，時常會影響您對時間管理的困擾因素有哪些？（此題為複選題，至多選五項），請在□內打「✓」

- ☐01.對許多事承諾太多無法拒絕。
- ☐02.書面資料及公文處理費時。
- ☐03.權責不清，不易做決定。
- ☐04.經常缺乏計畫，手忙腳亂。
- ☐05.工作經常拖延，無法依原訂進度執行。
- ☐06.電話干擾不斷。
- ☐07.不速之客造訪。
- ☐08.與人溝通協調，占用太多時間。
- ☐09.許多事須親自處理，授權不易。
- ☐10.經常參加會議及各項活動。
- ☐11.學校偶發事件處理。
- ☐12.上級長官臨時交辦事項。
- ☐13.同仁沒有時間管理觀念。
- ☐14.家庭問題。

原始題項	變項名稱	變項標記	數值範圍	水準數值標記	
五、困擾因素					
	D5_1		0-1	0:未勾選	1:勾選
	D5_2		0-1	0:未勾選	1:勾選
	D5_3		0-1	0:未勾選	1:勾選
	D5_4		0-1	0:未勾選	1:勾選
	D5_5		0-1	0:未勾選	1:勾選
	D5_6		0-1	0:未勾選	1:勾選
	D5_7		0-1	0:未勾選	1:勾選
	D5_8		0-1	0:未勾選	1:勾選
	D5_9		0-1	0:未勾選	1:勾選
	D5_10		0-1	0:未勾選	1:勾選
	D5_11		0-1	0:未勾選	1:勾選

資料來源:吳明隆, SPSS操作與應用：問卷統計分析實務，五南出版社

六、時間管理策略運用狀況

填答說明：請仔細閱讀下列敘述句後，根據您的意見，在各題適當的□內打「✓」

	完全符合	大部分符合	有一半符合	多數不符合	完全不符合
01.我會訂定明確的工作目標，並據此發展周詳的計畫。.....	☐	☐	☐	☐	☐
02.我會以事情的輕重緩急來編排行事優先順序。	☐	☐	☐		
04.我會先行檢查一下明日的行程並預做準備。.....	☐	☐	☐		
05.我會利用行政團隊成員的優點，合作把工作完成。	☐	☐	☐		
06.我會隨時把握機會與工作成員做良好的溝通。	☐	☐	☐		
07.我會事先做合適的時間分配，使工作都能如期完成。	☐	☐	☐		
08.發展計畫時，我會思考可能的阻礙，事先做好因應的措施。	☐	☐	☐		
09.我覺得自己是一個很會做時間管理的人。	☐	☐	☐		
10.我會善用記事本等工具，記錄每天重要的訊息和行程。	☐	☐	☐		
11.我會利用布告欄記載重要行事，讓同仁做好時間分布和管理。	☐	☐	☐		
12.我會使用電腦等工具，協助工作較有效率地完成。	☐	☐	☐		
13.我會利用電腦網頁公布工作要項，使校務運作更順暢。	☐	☐	☐		
14.我會在文件的關鍵處加標記，以便加快重讀時的速度。	☐	☐	☐		
15.我會將工作上的困難、想法與心得記錄下來，以便未來查考。	☐	☐	☐		

原始題項	變項名稱	變項標記	數值範圍	水準數值標記
六、時間策略運用				
01	E1		1-5	
02	E2		1-5	
03	E3		1-5	
04	E4		1-5	
05	E5		1-5	
06	E6		1-5	
07	E7		1-5	
08	E8		1-5	
09	E9		1-5	
10	E10		1-5	
11	E11		1-5	
12	E12		1-5	
13	E13		1-5	
14	E14		1-5	
15	E15		1-5	

資料來源:吳明隆, SPSS操作與應用：問卷統計分析實務，五南出版社

試卷之試題分析： 二元計分類型及多元計分類型

27 / 100

二元計分類型

答案

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1 ID	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	P11	P12	P13	P14	P15	
2 ANS	3	2	4	2	2	3	4	4	1	2	1	4	1	2	3	
3 ST001	3	2	4	3	4	3	4	4	1	3	2	4	3	2	4	
4 ST002	2	2	4	2	2	1	3	4	1	3	3	1	1	2	3	
5 ST003	3	2	4	4	4	3	4	4	1	4	2	1	1	2	3	
6 ST004	3	2	4	4	4	3	4	4	1	2	2	4	1	2	3	
7 ST005	1	2	4	2	2	3	4	4	1	2	1	4	1	2	4	
8 ST006	3	2	4	1	4	3	1	2	1	3	3	2	2	2	3	
9 ST007	2	2	2	1	4	3	2	4	1	4	3	2	4	2	3	
10 ST008	3	2	2	3	4	3	2	4	1	2	2	4	1	2	1	
11 ST009	4	2	4	3	4	1	3	4	1	2	1	4	1	2	4	
12 ST010																
13 ST011																
14 ST012																
15 ST013	3	1	4	2	2	3	4	1	1	2	2	4	3	4	4	
16 ST014	3	1	4	2	2	3	4	1	1	2	2	4	3	4	4	
17 ST015	3	2	4	2	2	3	4	4	1	2	1	4	1	2	3	
18 ST016	2	2	4	1	4	3	4	4	1	3	2	4	1	2	3	
19 ST017	3	2	4	2	2	3	3	4	1	3	2	4	1	2	4	
20 ST018	3	1	4	2	2	3	3	4	1	2	1	2	1	2	1	

受試者的反應資料

多元計分類型

題項B101~B106、B201~B206

	A	B	C	D	E	F	G	H	I	J	K	L	M
1 ID	B101	B102	B103	B104	B105	B106	B201	B202	B203	B204	B205	B206	
2 ST001	4	4	4	4	4	4	4	4	4	4	4	4	
3 ST002	4	4	4	4	4	4	4	4	4	4	4	4	
4 ST003	3	3	3	3	3	3	3	3	3	3	3	3	
5 ST004	3	3	3	3	3	4	3	3	3	3	3	3	
6 ST005	3	3	3	3	3	3	3	3	3	3	3	3	
7 ST006	3	3	4	4	4	4	3	3	3	3	3	3	
8 ST007	3	3	3	3	3	3	3	3	3	3	3	3	
9 ST008	3	3	3	3	3	3	3	3	3	3	3	3	
10 ST009	3	3	3	3	3	3	3	3	3	3	3	3	
11 ST010	3	3	3	3	3	3	3	4	3	3	3	3	
12 ST011	3	3	4	4	4	4	3	3	3	3	3	4	
13 ST012	3	3	3	3	3	3	4	4	4	4	4	4	
14 ST013	4	4	4	4	4	4	4	4	4	4	4	4	
15 ST014	3	3	3	3	3	3	3	3	3	4	4	3	
16 ST015	4	4	4	4	4	4	4	4	4	4	4	4	

分析要點：

- 二元計分 (答對率)
- 項目分析(itemanalysis): 以單題為單位來進行分析
- 誘答力分析 (選擇題選項分析)
 - 一道選擇題的選項分析除可了解考生在正確選項的選答情形外，更能了解「不正確選項」發揮的誘答效果，所以選項分析又可稱「誘答力分析」。所謂誘答力是指「不正確選項」吸引考生選答的程度，通常以該選項選答人數占全體到考考生人數的百分比來表示，百分比數值越大，表示誘答力越高。
 - 誘答功能判斷標準：(a) 低分組學生至少有一人選。(b) 低分組選的人數不得低於高分。
- 描述性統計
- 計算難度

二元計分的難度計算有2種方式：

 - (1) 答對人數/全部受試者
 - (2) 高低分組(前後27%、前後33%、前後25%)之平均數

計算高分組及低分組在每一題答對的人數百分比，記為 PH 及 PL 。每一題之難度： $P = (PH + PL)/2$ 。
- 計算鑑別度： $D = PH - PL$
- 高低分組平均數t檢定之t值：決斷值，criticalratio, CR值
- 計算題目與總分相關係數
- 計算總量表刪題後信度、題目信度

若試題理想難度預設值為0.5

試題難易度等級表

難易度	難易度等級
$P \geq 0.80$	極容易
$0.60 \leq P < 0.80$	容易
$0.40 \leq P < 0.60$	難易適中
$0.20 \leq P < 0.40$	困難
$P < 0.20$	極困難

美國學者伊博（1979）的評鑑標準

鑑別指數	試題評鑑
0.40以上	非常優良
0.30~0.29	優良，但可能需修改
0.20~0.29	尚可，但通常需修改
0.19以下	劣，須淘汰或修改

將學生分為高分組和低分組（例如：27%）。

- 難度就是答對率，難度值越高代表題目越簡單，越低表示越困難。 $P = (PH + PL)/2$
- 鑑別度 = 高分組答對率 - 低分組答對率。值越高表示越具鑑別度。 $(D = PH - PL)$
- 高鑑別度的題目能有效區分高低分組學生的能力。
- 鑑別度與難度有密切關係。難度越接近0.5，鑑別度越高，難度越接近極端值，鑑別度越低。

$$-1 \leq D \leq 1$$

- $D = 0$ ，無鑑別度
試題太簡單，高分組與低分組學生全部答對
試題太困難，高分組與低分組學生全部答錯
- $D = +1$
高分組學生全部答對，低分組學生全部答錯。
- $D = -1$
低分組學生全部答對，高分組學生全部答錯。

計算難度及鑑別度值的目的：整體試卷難度分配、各單元及目標的難度分配、難度值過高的題目、難度值過低的題目(尤其是鑑別度又低)。



抽樣調查的樣本數

29 / 100

- 問卷調查之抽樣樣本數多寡，並無定論。
 - 一般問卷，正式抽樣樣本數，最好為350人以上。
 - 一般而言，樣本數要占母群體10%。
 - 若母群體人數少於500，則樣本數最好占母群體20%以上。
 - 若母群體人數較少，則樣本數最好占母群體30%以上。
- 需考量研究者的時間、精力、財力等因素。
- 抽取具代表性的樣本比抽取多數但不具代表性的樣本更具外在效度。
- 如有組別，最少每組需20以上。最少不得低於15人，理想為30人以上。

N : 母群體樣本數。 k : 常數 (與信賴係數相關)。 P : 一般設為0.5。

以企業組織員工為研究對象 (5000 人), 則在隨機取樣時, 至少樣本要抽取多少, 統計推論才是可靠?

母體數: $N = 5000$, 顯著水準: $\alpha = 0.05$,

信賴係數: $1 - 0.05 = 0.95$, $k = 1.96$, $P = 0.5$ 。

若 $N = 10000$, 則 $n \approx 370$ 。若 $N = 40000$, 則 $n \approx 381$ 。

$$n \geq \frac{N}{\left(\frac{\alpha}{k}\right)^2 \frac{N-1}{P(1-P)} + 1}$$

$$\begin{aligned} n &\geq \frac{N}{\left(\frac{\alpha}{k}\right)^2 \frac{N-1}{P(1-P)} + 1} \\ &= \frac{5000}{\left(\frac{0.05}{1.96}\right)^2 \frac{5000-1}{0.5(1-0.5)} + 1} \\ &= \frac{5000}{14.0125} \\ &= 356.516 \\ &\approx 357 \end{aligned}$$

How to Choose Sample Size for a Simple Random Sample

Margin of Error (誤差範圍)

The margin of error expresses the maximum expected difference between the true population parameter and a sample estimate of that parameter. To be meaningful, the margin of error should be qualified by a probability statement (often expressed in the form of a confidence level).

For example, a pollster might report that 50% of voters will choose the Democratic candidate. To indicate the quality of the survey result, the pollster might add that the margin of error is +5%, with a confidence level of 90%. This means that if the survey were repeated many times with different samples, the true percentage of Democratic voters would fall within the margin of error 90% of the time.

Let E = the desired margin of error: $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ $E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Sample statistic	Population size	Sample size
Mean	Known	$n = \frac{z^2 \sigma^2 \left[\frac{N}{N-1} \right]}{E^2 + \left[\frac{z^2 \sigma^2}{N-1} \right]}$
Mean	Unknown	$n = \frac{z^2 \sigma^2}{E^2}$
Proportion	Known	$n = \frac{(z^2 \cdot p \cdot q) + E^2}{E^2 + \frac{z^2 \cdot p \cdot q}{N}}$
Proportion	Unknown	$n = \frac{(z^2 \cdot p \cdot q) + E^2}{E^2}$

$$n \geq \frac{N}{\left(\frac{\alpha}{k} \right)^2 \frac{N-1}{P(1-P)} + 1}$$

- Level of significance: α ($\alpha = 0.05$)
- Confidence level: $1 - \alpha$ ($1 - \alpha = 0.95$)
- Margin of error: E ($E = 0.04$)
- Critical standard score: $z_{\alpha/2}$ ($z_{\alpha/2} = 1.96$)
- Size of the population: N
- Variance of the population: σ^2
- Population proportion: $p, q = 1 - p$

Determining the sample size: interval estimation, hypothesis testing

31 / 100

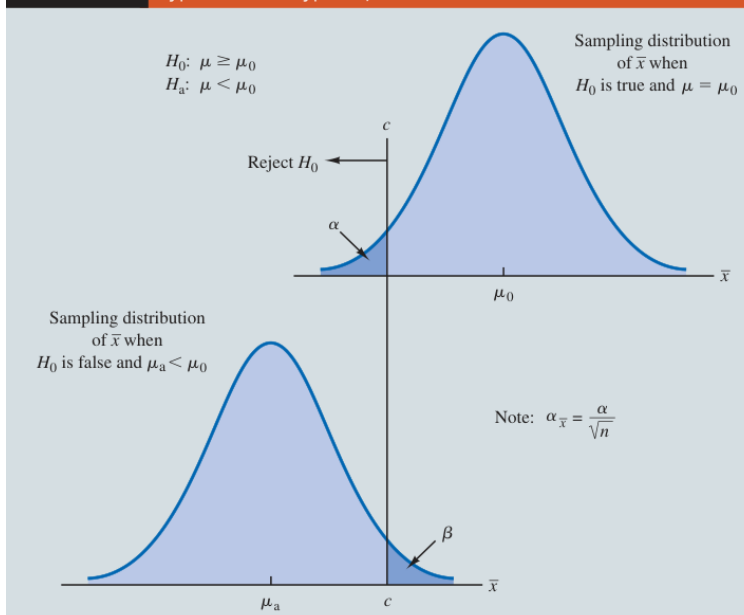
Let E = the desired margin of error:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

SAMPLE SIZE FOR AN INTERVAL ESTIMATE OF A POPULATION MEAN

$$n = \frac{(z_{\alpha/2})^2 \sigma^2}{E^2}$$

FIGURE 9.10 Determining the Sample Size for Specified Levels of the Type I (α) and Type II (β) Errors



$$H_0: \mu \geq \mu_0$$

$$H_a: \mu < \mu_0$$

$$\mu_0 - z_{\alpha} \frac{\sigma}{\sqrt{n}} = \mu_a + z_{\beta} \frac{\sigma}{\sqrt{n}}$$

SAMPLE SIZE FOR A ONE-TAILED HYPOTHESIS TEST ABOUT A POPULATION MEAN

$$n = \frac{(z_{\alpha} + z_{\beta})^2 \sigma^2}{(\mu_0 - \mu_a)^2} \quad (9.7)$$

where

z_{α} = z value providing an area of α in the upper tail of a standard normal distribution

z_{β} = z value providing an area of β in the upper tail of a standard normal distribution

σ = the population standard deviation

μ_0 = the value of the population mean in the null hypothesis

μ_a = the value of the population mean used for the Type II error

Note: In a two-tailed hypothesis test, use (9.7) with $z_{\alpha/2}$ replacing z_{α} .



- 從Excel讀取資料，並設定變數屬性

- 選擇觀察值

- 選擇部份特定條件的觀察值
- 觀察值的隨機樣本
- 使用過濾變數

資料檢核

- summary statistics
- index plot, histogram, box-plot
- heatmap

- 分割檔案 (群組資料): 以群組進行分析

- 資料轉換

- 自動重新編碼、重新編碼
- 轉換成等級觀察值、Visual Binning

- 計算變數、橫向計數

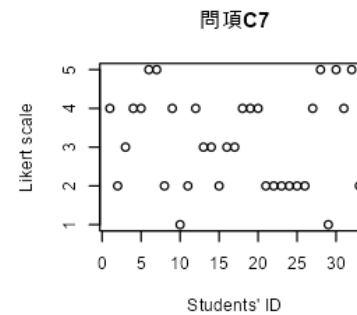
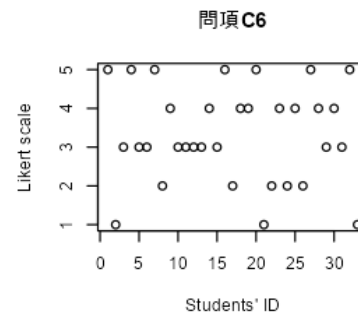
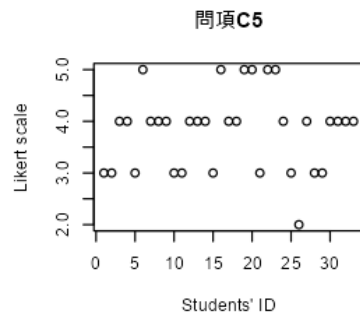
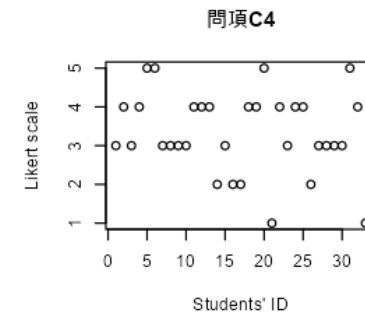
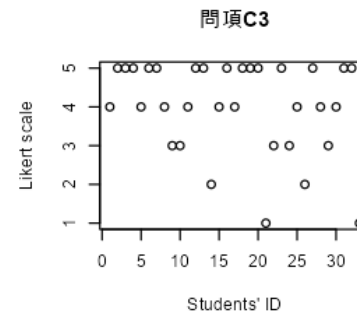
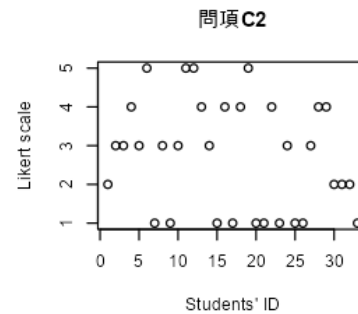
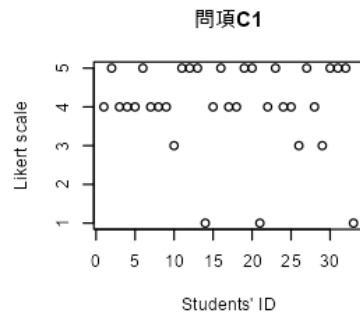
- 排序、遺漏值處理、資料整合

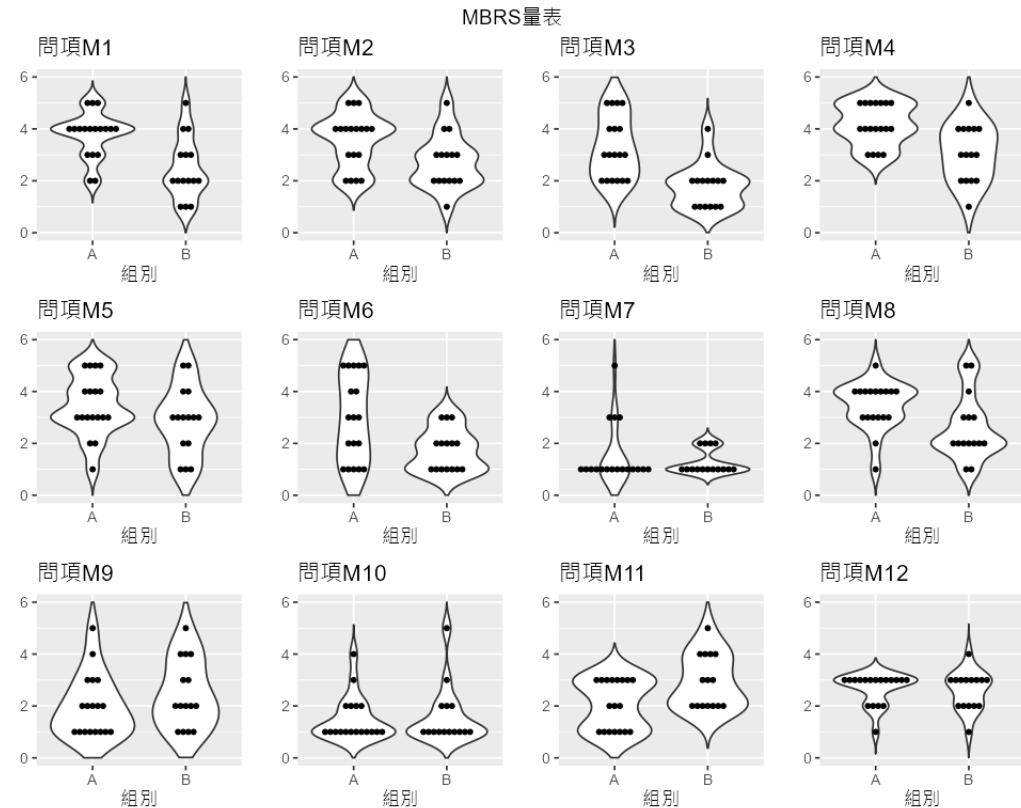
```
> CBR5 <- read.csv("data/CBR5.csv", fileEncoding = "BIG5")
> head(CBR5)
  學生編號 組別 問項C1 問項C2 問項C3 問項C4 問項C5 問項C6 問項C7
1         1   B      4      2      4      3      3      5      4
...
6        24   A      5      5      5      5      5      3      5
> dim(CBR5)
[1] 33 9
> CBR5$組別 <- as.factor(CBR5$組別)
>
> MBRS <- read.csv("data/MBRS.csv", fileEncoding = "BIG5")
> head(MBRS)
  學生編號 組別 問項M1 問項M2 問項M3 問項M4 問項M5 問項M6 問項M7 問項M8 問項M9 問項M10 問項M11 問項M12
1         1   B      3      3      2      3      2      1      2      2      2      1      2      2
...
6         7   A      5      5      4      5      5      5      3      4      2      1      1      3
> dim(MBRS)
[1] 33 14
> MBRS$組別 <- as.factor(MBRS$組別)
>
> CBR5_MBRS <- merge(CBR5, MBRS, by = "學生編號")
> CBR5_MBRS$組別.y <- NULL
> colnames(CBR5_MBRS)[2] <- "組別"
```

	A	B	C	D	E	F	G	H	I
1	學生編號	組別	問項C1	問項C2	問項C3	問項C4	問項C5	問項C6	問項C7
2	1	2	4	2	4	3	3	5	4
3	12	1	5	3	5	4	3	1	2
4	13	2	4	3	5	3	4	3	3
5	22	1	4	4	5	4	4	5	4
6	23	2	4	3	4	5	3	3	4
7	24	1	5	5	5	5	5	3	5
8	37	1	4	1	5	3	4	5	5
9	38	2	4	3	4	3	4	2	2
10	39	1	4	1	3	3	4	4	4
11	40	2	3	3	3	3	3	3	1
12	41	1	5	5	4	4	3	3	2
13	2	1	5	5	5	4	4	3	4
14	3	1	5	4	5	4	4	3	3
15	4	2	1	3	2	2	4	4	3

	A	B	C	D	E	F	G	H	I	J	K
1	學生編號	組別	問項M1	問項M2	問項M3	問項M4	問項M5	問項M6	問項M7	問項M8	問項M9
2	1	2	3	3	2	3	2	1	2	2	2
3	2	1	4	4	5	5	5	5	1	5	3
4	3	1	4	4	2	5	5	5	1	4	3
5	4	2	2	2	1	2	1	1	1	1	2
6	6	2	2	2	2	1	2	2	1	1	4
7	7	1	5	5	4	5	5	5	3	4	2
8	9	2	3	2	1	2	3	1	1	2	4
9	10	1	4	5	4	5	4	3	1	4	3
10	12	1	3	3	2	4	2	2	1	3	5
11	13	2	4	3	2	4	3	1	1	4	2
12	14	1	4	4	4	4	3	4	1	4	2
13	15	2	2	3	2	3	3	2	1	3	3
14	16	1	2	2	2	3	3	1	1	4	4
15	17	2	2	3	4	4	5	3	2	5	4

```
par(mfrow = c(2, 4))
lapply(3:9, function(x) plot(CBRS[,x],
                             xlab = "Students' ID",
                             ylab = "Likert scale",
                             main = names(CBRS)[x]))
```





```
library(ggplot2)
library(grid)
library(gridExtra)
```

```
my_violin <- function(x){
  ggplot(MBRS, aes(x = 組別, y = MBRS[, x]))
    geom_violin(trim = FALSE) +
    geom_dotplot(binaxis = 'y', stackdir = 'center') +
    scale_y_continuous(limits = c(0, 6)) +
    labs(title = names(MBRS)[x], y = "")
}
```

```
}
```

```
violin_list <- lapply(3:14, my_violin)
grid.arrange(grobs = violin_list, nrow = 3, ncol = 4, top = "MBRS量表")
```

資料檢核: 熱圖 (群集分析)

36 / 100

```
CBRS_MBRS <- merge(CBRS, MBRS, by = "學生編號")
CBRS_MBRS$組別.y <- NULL
colnames(CBRS_MBRS)[2] <- "組別"

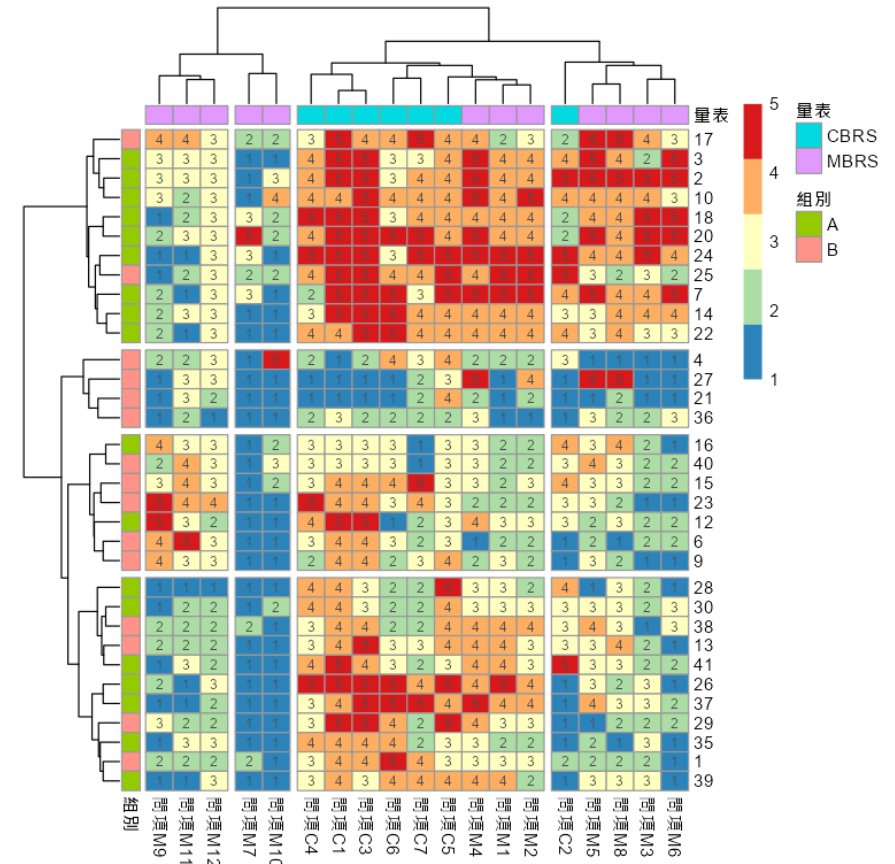
CBRS_MBRS_X <- CBRS_MBRS[, 3:ncol(CBRS_MBRS)]
rownames(CBRS_MBRS_X) <- CBRS_MBRS[, 1]
colnames(CBRS_MBRS_X)

col_groups <- data.frame(量表 = c(rep("CBRS", 7),
rep("MBRS", 12)))
row.names(col_groups) <- colnames(CBRS_MBRS_X)
col_groups

row_groups <- data.frame(組別 = CBRS_MBRS$組別)
rownames(row_groups) <- rownames(CBRS_MBRS_X)
row_groups

library(RColorBrewer)
library(pheatmap)

pheatmap(CBRS_MBRS_X,
         color = rev(brewer.pal(5, "Spectral")),
         annotation_row = row_groups,
         annotation_col = col_groups,
         cutree_rows = 4,
         cutree_cols = 4,
         display_numbers = TRUE,
         number_format = "%.0f",
         clustering_method = "ward.D2")
```



資料檢核: 熱圖 (以組別區分)

37 / 100

```
library(tidyverse)
CBRS_MBRS_means <- rowMeans(CBRS_MBRS[, 3:ncol(CBRS_MBRS)])
CBRS_MBRS_ALL <- cbind(CBRS_MBRS, means = CBRS_MBRS_means)
CBRS_MBRS_sort <- arrange(CBRS_MBRS_ALL, 組別, means)
```

```
CBRS_MBRS_X_sort <- CBRS_MBRS_sort[, 3:(ncol(CBRS_MBRS_sort)-1)]
rownames(CBRS_MBRS_X_sort) <- CBRS_MBRS_sort[,1]
colnames(CBRS_MBRS_X_sort)
```

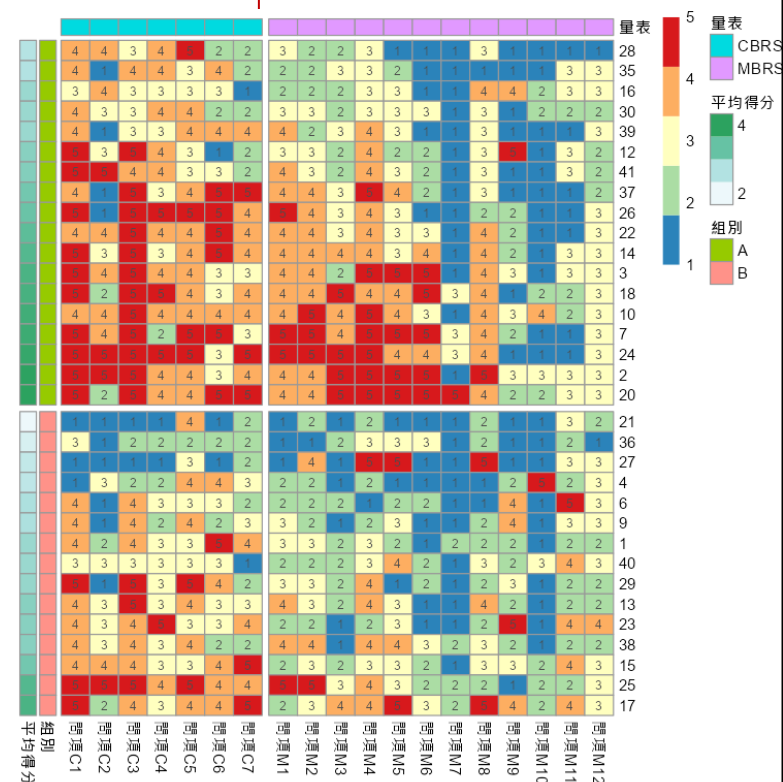
```
row_group2 <- data.frame(組別 = CBRS_MBRS_sort$組別,
                          平均得分 = round(CBRS_MBRS_sort$means, 2))
```

```
rownames(row_group2) <- rownames(CBRS_MBRS_X_sort)
colnames(row_group2)
```

```
col_group2 <- data.frame(量表 = c(rep("CBRS", 7),
                                   rep("MBRS", 12)))
rownames(col_group2) <- colnames(CBRS_MBRS_X_sort)
colnames(col_group2)
```

列位依組別及每生之平均排序，欄位依量表及每項目平均排序。

```
pheatmap(CBRS_MBRS_X_sort,
          color = rev(brewer.pal(5, "Spectral")),
          annotation_row = row_group2,
          annotation_col = col_group2,
          display_numbers = TRUE,
          number_format = "%.0f",
          cluster_rows = FALSE,
          cluster_cols = FALSE,
          gaps_row = 18,
          gaps_col = 7)
```



SPSS編碼範例： SPSS需要對變量編碼(文字轉數字)，R不用

38 / 100

student-exam.sav [資料集9] - SPSS Statistics Data Editor

檔案(E) 編輯(E) 檢視(V) 資料(D) 轉換(T) 分析(A) 統計圖(G) 公用程式(U) 增益集(O) 視窗(W) 說明(H)

顯示：6 個變數 (共有 6 個)

姓名	班級	性別	數學	英文	成績滿意度
吳云呈	甲班				66
柳芝榮	甲班				58
許靜羽	甲班				95
林妙怡	甲班				74
尤耀貞	甲班				84
黃雅德	乙班				58
陳淑惠	乙班				68
林蕙伶	乙班				80
陳韻英	甲班				93
曾雲如	甲班				
王妙鳳	甲班	女	55		
楊超菁	甲班	男	64		
陳仁欽	乙班	男	71		
李士揚	甲班	男	78		
劉燕丹	甲班	男	84		
沈水玲	乙班	女			

自動重新編碼

變數 -> 新名稱(V)

班級 -> 班級編碼

新名稱(N)：班級編碼

新增名稱(A)

重新編碼起始值

☐ 最低值(L) ☒ 最高值(H)

☐ 所有變數都使用相同重新編碼配置(U)

☐ 將空白字串值視為使用者遺漏值

範本

☐ 套用範本自(I)：檔案(E)...

☐ 另存範本(S)：檔案(I)...

確定 貼上之後(P) 重設(R) 取消 輔助說明

顯示：8 個變數 (共有 8 個)

英文	成績滿意度	班級編碼	性別編碼	var
66	1	1	2	
58	2	1	2	
95	3	1	2	
74	1	1	2	
84	1	1	2	
58	1	2	2	
68	3	2	2	
80	5	2	2	
93	3	1	2	
93	5	1	2	
85	3	1	2	
54	3	1	2	
51	4	2	1	
98	1	1	1	
100	5	1	1	
87	1	2	2	
	1	1	2	

資料檢視 變數檢視

SPSS Statistics Processor is ready

設定變數屬性，並以「數值標記」檢視

***未命名標題5 [資料集10] - SPSS Statistics Data Editor**

檔案(E) 編輯(E) 檢視(V) 資料(D) 轉換(I) 分析(A) 統計圖(G) 公用程式(U) 增益集(O) 視窗(W) 說明(H)

	名稱	類型	寬度	小數	標記	值	遺漏	欄	對齊	測量
1	姓名	字串	9	0		無	無	9	≡ 靠左	● 名義
2	班級	數字的	11	0		{1, 甲班}...	無	11	≡ 靠右	● 名義
3	性別	數字的	11	0		{1, 男生}...	無	11	≡ 靠右	● 名義
4	數學	數字的	11	0		無	無	11	≡ 靠右	▬ 尺度
5	英文	數字的	11	0		無	無	11	≡ 靠右	▬ 尺度
6	成績滿意度	數字的	11	0		{1, 非常不...	無	11	≡ 靠右	▬ 次序的

SPSS Statistics Processor is ready

53: 成績滿意度

顯示: 6 個變數 (共有 6 個)

	姓名	班級	性別	數學	英文	成績滿意度	var
1	吳云呈	甲班	女生	60	66	非常不滿意	
2	柳芝葵	甲班	女生	42	58	不滿意	
3	許靜羽	甲班	女生	78	95	普通	
4	林妙怡	甲班	女生	65	74	非常不滿意	
5	尤耀貞	甲班	女生	68	84	非常不滿意	
6	黃雅德	乙班	女生	57	58	非常不滿意	
7	陳淑惠	乙班	女生	55	68	普通	
8	林蕙伶	乙班	女生	97	80	非常滿意	
9	陳韻英	甲班	女生	87	93	普通	
10	曾雲如	甲班	女生	92	93	非常滿意	
11	王妙鳳	甲班	女生	75	85	普通	
12	楊超菁	甲班	女生	55	54	普通	
13	陳仁欽	乙班	男生	64	51	滿意	
14	李士攝	甲班	男生	71	98	非常不滿意	

資料檢視 變數檢視

數值標記

SPSS Statistics Processor is ready

數值註解

數值註解

值(A): 5

標記(L): 非常滿意

新增(A) 變更(O) 移除(R)

1 = "非常不滿意"
2 = "不滿意"
3 = "普通"
4 = "滿意"
5 = "非常滿意"

確定 取消 輔助說明

讀入資料，視情況編碼並轉成適當類別

40 / 100

```
> library(readxl)
> student_score_s1 <- read_excel("data/學生成績Data1.xlsx", sheet = 1)
> head(student_score_s1)
# A tibble: 6 × 6
  姓名    班級  性別    數學  英文  成績滿意度
  <chr>  <chr> <chr>  <dbl> <dbl>    <dbl>
1 吳沅呈  甲班  女      60    66        1
...
6 黃雅德  乙班  女      57    58        1
```

```
> student_score_s1$班級編碼 <- ifelse(student_score_s1$班級 == "甲班", 1, 2)
```

```
> head(student_score_s1)
# A tibble: 6 × 7
  姓名    班級  性別    數學  英文  成績滿意度 班級編碼
  <chr>  <chr> <chr>  <dbl> <dbl>    <dbl>    <dbl>
1 吳沅呈  甲班  女      60    66        1        1
...
6 黃雅德  乙班  女      57    58        1        2
```

```
> satisfy_scale <- c("非常不滿意", "不滿意", "普通", "滿意", "非常滿意")
> student_score_s1$成績滿意度文字因子 <- satisfy_scale[student_score_s1$成績滿意度]
> student_score_s1$成績滿意度文字因子 <- factor(student_score_s1$成績滿意度文字因子,
+                                               levels = satisfy_scale, ordered = TRUE)
```

```
> str(student_score_s1)
```

```
tibble [50 × 8] (S3: tbl_df/tbl/data.frame)
```

```
$ 姓名      : chr [1:50] "吳沅呈" "柳芝綦" "許靜羽" "林妙怡" ...
$ 班級      : chr [1:50] "甲班" "甲班" "甲班" "甲班" ...
$ 性別      : chr [1:50] "女" "女" "女" "女" ...
$ 數學      : num [1:50] 60 42 78 65 68 57 55 97 87 92 ...
$ 英文      : num [1:50] 66 58 95 74 84 58 68 80 93 93 ...
```

```
$ 成績滿意度 : num [1:50] 1 2 3 1 1 1 3 5 3 5 ...
```

```
$ 班級編碼   : num [1:50] 1 1 1 1 1 2 2 2 1 1 ...
```

```
$ 成績滿意度文字因子: Ord.factor w/ 5 levels "非常不滿意"<"不滿意"<...: 1 2 3 1 1 1 3 5 3 5 ...
```

	A	B	C	D	E	F
1	姓名	班級	性別	數學	英文	成績滿意度
2	吳沅呈	甲班	女	60	66	1
3	柳芝綦	甲班	女	42	58	2
4	許靜羽	甲班	女	78	95	3
5	林妙怡	甲班	女	65	74	1
6	尤耀貞	甲班	女	68	84	1
7	黃雅德	乙班	女	57	58	1
8	陳淑惠	乙班	女	55	68	3
9	林蕙伶	乙班	女	97	80	5
10	陳韻英	甲班	女	87	93	3
11	曾雲如	甲班	女	92	93	5
12	王妙鳳	甲班	女	75	85	3
13	楊超菁	甲班	女	55	54	3
14	陳仁欽	乙班	男	64	51	4
15	李士揚	甲班	男	71	98	1
16	劉燕丹	甲班	男	78	100	5
17	沈水玲	乙班	女	84	87	1
18	莊雅雲	甲班	女	85	95	1
19	彭景雯	甲班	女	76	94	2
20	黃千禧	甲班	女	74	88	1

保留日後要運算

```
# 選擇女性樣本
student_score_s1[student_score_s1$性別 == "女", ]

# 選擇甲班及英文大於80分之樣本
student_score_s1[(student_score_s1$班級 == "甲班") |
                  (student_score_s1$英文 >= 80), ]

# 選擇男性、數學大於等於80分及英文大於70分之樣本
student_score_s1[((student_score_s1$性別 == "男") &
                  (student_score_s1$數學 >= 80)) |
                  (student_score_s1$英文 > 70), ]

# 選擇隨機樣本
student_score_s1[sample(1:nrow(student_score_s1), 10), ]

# 選擇11號~20號的姓名、數學、英文
student_score_s1[11:20, c("姓名", "數學", "英文")]

# 選擇某一時間內的資料
student_score_s1$時間 <- seq(as.Date("2028/04/02"), by = "day",
                             length.out = nrow(student_score_s1))
selected_date <- (as.Date("2028/04/08") < student_score_s1$時間) &
                 (student_score_s1$時間 < as.Date("2028/04/20"))
student_score_s1[selected_date, ]

# 刪除成績滿意度為「非常不滿意」
student_score_s1[!(student_score_s1$成績滿意度 == 1), ]
student_score_s1[!(student_score_s1$成績滿意度文字因子 == "非常不滿意"), ]
```



比較群組：敘述統計、次數

42 / 100

```
> library(psych)
> describe(student_score_s1[, sapply(student_score_s1, is.numeric)])
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
數學	1	50	75.78	13.96	75.5	75.92	16.31	42	100	58	-0.16	-0.80	1.97
英文	2	50	79.32	16.12	85.0	80.83	16.31	38	100	62	-0.68	-0.54	2.28
成績滿意度	3	50	2.56	1.47	2.0	2.45	1.48	1	5	4	0.43	-1.25	0.21
班級編碼	4	50	1.56	0.50	2.0	1.57	0.00	1	2	1	-0.23	-1.98	0.07

```
> describeBy(student_score_s1[, sapply(student_score_s1, is.numeric)],
+ student_score_s1$班級)
```

Descriptive statistics by group

group: 乙班

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
數學	1	28	78.11	14.06	77	78.21	16.31	55	100	45	-0.09	-1.24	2.66
英文	2	28	78.29	15.82	85	79.42	16.31	38	99	61	-0.63	-0.51	2.99
成績滿意度	3	28	2.93	1.49	3	2.92	1.48	1	5	4	0.12	-1.51	0.28
班級編碼	4	28	2.00	0.00	2	2.00	0.00	2	2	0	NaN	NaN	0.00

group: 甲班

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
數學	1	22	72.82	13.57	73.5	73.28	14.83	42	94	52	-0.33	-0.68	2.89
英文	2	22	80.64	16.78	87.5	82.28	13.34	41	100	59	-0.72	-0.73	3.58
成績滿意度	3	22	2.09	1.34	1.5	1.89	0.74	1	5	4	0.86	-0.54	0.29
班級編碼	4	22	1.00	0.00	1.0	1.00	0.00	1	1	0	NaN	NaN	0.00

```
> table(student_score_s1$性別, student_score_s1$成績滿意度文字因子)
```

	非常不滿意	不滿意	普通	滿意	非常滿意
女	13	3	6	2	5
男	4	7	3	4	3

```
> table(student_score_s1$性別)
```

女 男
29 21

```
> table(student_score_s1$成績滿意度文字因子)
```

非常不滿意	不滿意	普通	滿意	非常滿意
17	10	9	6	8



two-way 列聯表

43 / 100

```
> tbl <- table(student_score_s1$性別, student_score_s1$成績滿意度文字因子)
> sum(tbl)
[1] 50
> tbl
```

	非常不滿意	不滿意	普通	滿意	非常滿意
女	13	3	6	2	5
男	4	7	3	4	3

```
>
> # overall
> prop.table(tbl)
非常不滿意 不滿意 普通 滿意 非常滿意
女 0.26 0.06 0.12 0.04 0.10
男 0.08 0.14 0.06 0.08 0.06
```

```
>
> # by row
> prop.table(tbl, margin = 1)
非常不滿意 不滿意 普通 滿意 非常滿意
女 0.44827586 0.10344828 0.20689655 0.06896552 0.17241379
男 0.19047619 0.33333333 0.14285714 0.19047619 0.14285714
```

```
>
> # by column
> prop.table(tbl, margin = 2)
非常不滿意 不滿意 普通 滿意 非常滿意
女 0.7647059 0.3000000 0.6666667 0.3333333 0.6250000
男 0.2352941 0.7000000 0.3333333 0.6666667 0.3750000
```

```
> margin.table(tbl)
[1] 50
> margin.table(tbl, margin = 1)
女 男
29 21
> margin.table(tbl, margin = 2)
非常不滿意 不滿意 普通 滿意 非常滿意
17 10 9 6 8
```

See also: **cumsum**



自動重新編碼: 算名次

44 / 100

```
> Average <- rowMeans(student_score_s1[, c("數學", "英文")])
> Rank <- rank(- Average)
> student_score_output <- data.frame(student_score_s1$姓名,
+                                     平均 = Average,
+                                     排名 = Rank)
> student_score_output
  student_score_s1.姓名 平均 排名
1          吳沄呈 63.0 40.0
2          柳芝棻 50.0 49.0
...
49        陳冠鈺 87.5 15.0
50        黃之伶 81.0 26.0
>
> student_score_output[order(student_score_output$平均, decreasing = TRUE), ]
  student_score_s1.姓名 平均 排名
46          吳宜帆 98.5  1.0
34          劉祐民 95.0  2.5
...
2          柳芝棻 50.0 49.0
45          李洋瑩 49.0 50.0
```

重新編碼使用狀況:

- 反項題重新計分。
- 連續變項數值分為數個等級。
- 背景變項水準數值重新合併。

```
> x <- c(24, 13, 26, 21, 7, 9, 2, 1, 30, 14, 20, 16, 6, 4, 12, 8,
11, 22, 18, 3)
> ifelse(x <= 10, 1, ifelse(x <= 20, 2, 3))
[1] 3 2 3 3 1 1 1 1 3 2 2 2 1 1 2 1 2 3 2 1
```

- 將年齡資料轉換為年齡群組1~20, 21~40, 41~60, 61歲以上，並編碼為A, B, C, D。

```
> set.seed(12345)
> age <- sample(1:100, 20)
> age
[1] 73 87 75 86 44 16 31 48 67 91 4 14 65 1 34 40 33 97 15 78
```

- 將" A" 與" E" 編碼為1，" C" 編碼為2，" B" 與" D" 編碼為3。

```
> set.seed(12345)
> code <- sample(LETTERS[1:5], 20, replace=T)
> code
[1] "D" "E" "D" "E" "C" "A" "B" "C" "D" "E" "A" "A" "D" "A" "B" "C"
[17] "B" "C" "A" "E"
```

See also: `cut()`, `recode{car}`

提示: %in%



重新編碼: 分數百分數轉換成等級

46 / 100

- 美國大學成績平均績點(GPA)(四分制)的計算方式如右表，請寫一R函式，將某同學之各科修課成績百分數score轉成等級及GPA。

等級 (Grade)	百分數	GPA
A	80 – 100 分	4
B	70 – 79 分	3
C	60 – 69 分	2
D	50 – 59 分	1
E	49 分以下	0

```
> set.seed(12345)
> score <- sample(0:100, 10, replace=T)
> score
[1] 72 88 76 89 46 16 32 51 73 99
```

```
gpa.table <- data.frame(grade=c("A", "B", "C", "D", "E"),
                          pscore=c("80-100", "70-79", "60-69", "50-59", "49-0"),
                          GPA=c(4, 3, 2, 1, 0))
```

```
gpa.table
set.seed(12345)
score <- sample(0:100, 10, replace=T)

score_to_gpa <- function(x){
  group.id <- ifelse(x >= 80, 1,
                    ifelse(x >= 70, 2,
                          ifelse(x >= 60, 3,
                                ifelse(x >= 50, 4, 5))))
  data.frame(score=x, gpa.table[group.id,], row.names = NULL)
}
```

```
> score_to_gpa(score)
  score grade pscore GPA
1    72     B   70-79    3
2    88     A  80-100    4
3    76     B   70-79    3
4    89     A  80-100    4
5    46     E   49-0     0
6    16     E   49-0     0
7    32     E   49-0     0
8    51     D   50-59    1
9    73     B   70-79    3
10   99     A  80-100    4
```

cut {base}: Convert Numeric to Factor

- **cut{base}** divides the range of **x** into intervals and codes the values in **x** according to which interval they fall. The leftmost interval corresponds to level one, the next leftmost to level two and so on.

```
cut(x, breaks, labels = NULL,
    include.lowest = FALSE, right = TRUE, dig.lab = 3, ordered_result = FALSE, ...)
```

```
> x <- rnorm(50)
> (x.cut1 <- cut(x, breaks = -5:5))
[1] (-1,0] (-2,-1] (-2,-1] (-1,0] (-1,0] (-2,-1] (0,1] (0,1] (-1,0] (1,2] (0,1]
...
[45] (1,2] (0,1] (-1,0] (-2,-1] (0,1] (0,1]
Levels: (-5,-4] (-4,-3] (-3,-2] (-2,-1] (-1,0] (0,1] (1,2] (2,3] (3,4] (4,5]
> table(x.cut1)
x.cut1
(-5,-4] (-4,-3] (-3,-2] (-2,-1] (-1,0] (0,1] (1,2] (2,3] (3,4] (4,5]
      0      0      1     10     18     13      8      0      0      0
> (x.cut2 <- cut(x, breaks = -5:5, labels = FALSE))
[1] 5 4 4 5 5 4 6 6 5 7 6 5 7 5 7 6 4 7 7 4 5 6 5 5 5 6 5 6 5 6 5 4 7 ...
[47] 5 4 6 6
> table(x.cut2)
x.cut2
 3  4  5  6  7
1 10 18 13  8
> hist(x, breaks = -5:5, plot = FALSE)$counts
[1] 0 0 1 10 18 13 8 0 0 0
```



cut {base} Examples

48 / 100

```
> #the outer limits are moved away by 0.1% of the range
> cut(0:10, 5)
[1] (-0.01,2] (-0.01,2] (-0.01,2] (2,4]      (2,4]      (4,6]      (4,6]      (6,8]
[9] (6,8]      (8,10]      (8,10]
Levels: (-0.01,2] (2,4] (4,6] (6,8] (8,10]
>
> age <- sample(0:80, 50, replace=T)
> summary(age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00  21.00   35.00   38.16  52.75   80.00
> cut(age, 5)
[1] (48.4,64.2] (16.8,32.6] (16.8,32.6] (48.4,64.2] (16.8,32.6] (32.6,48.4]
...
[49] (16.8,32.6] (48.4,64.2]
Levels: (0.921,16.8] (16.8,32.6] (32.6,48.4] (48.4,64.2] (64.2,80.1]
> mygroup <- c(0, 15, 20, 50, 60, 80)
> (x.cut <- cut(age, mygroup))
[1] (50,60] (20,50] (15,20] (20,50] (20,50] (20,50] (15,20] (0,15] (0,15] (60,80]
...
Levels: (0,15] (15,20] (20,50] (50,60] (60,80]
> table(x.cut)
x.cut
(0,15] (15,20] (20,50] (50,60] (60,80]
      7       5      22       8       8
```

Note: Instead of `table(cut(x, br))`, `hist(x, br, plot = FALSE)` is more efficient and less memory hungry. Instead of `cut(*, labels = FALSE)`, `findInterval()` is more efficient.

■ 變數的四則運算

```
> sqrt(student_score_s1$數學) * 10
> rowMeans(student_score_s1[, c("數學", "英文")])
```

■ 橫向計數: 單一變數 (例: 不及格)

```
> table(ifelse(student_score_s1$數學 < 60, "不及格", "及格"))
```

```
不及格    及格
      8     42
```

■ 橫向計數: 多個變數(例: 不及格科數)

```
> output <- data.frame(student_score_s1[, c("數學", "英文")],
+                       不及格科數 = apply(student_score_s1[, c("數學", "英文")], 1,
+                       function(x) sum(x < 60)))
```

```
> output
  數學 英文 不及格科數
1   60   66         0
2   42   58         2
3   78   95         0
4   65   74         0
...
```

```
> table(CBR5$問項C1)
```

```
1  3  4  5
3  3 14 13
```

```
> agree_scale <- c("非常不同意", "不同意", "普通", "同意", "非常同意")
```

```
> C1 <- factor(CBR5$問項C1, levels =1:5, labels = agree_scale, order = T)
```

```
> str(C1)
```

```
Ord.factor w/ 5 levels "非常不同意"<"不同意"<...: 4 5 4 4 4 5 4 4 4 3 ...
```

```
> table(C1)
```

```
C1
```

```
非常不同意
```

```
3
```

```
不同意
```

```
0
```

```
普通
```

```
3
```

```
同意
```

```
14
```

```
非常同意
```

```
13
```

橫向計數: 多個變數 (例:每個人在所有問題選項中「滿意」「不滿意」的個數。)



排序

50 / 100

```
> arrange(student_score_s1, 成績滿意度, 數學, 英文)
# A tibble: 50 × 8
  姓名    班級  性別  數學  英文 成績滿意度 班級編碼 成績滿意度文字因子
  <chr>  <chr> <chr> <dbl> <dbl>      <dbl>      <dbl> <ord>
1 張亞慧 甲班  女    55    68          1          1 非常不滿意
2 黃怡昕 乙班  女    56    72          1          2 非常不滿意
3 李洋瑩 甲班  女    57    41          1          1 非常不滿意
4 黃雅德 乙班  女    57    58          1          2 非常不滿意
5 吳云呈 甲班  女    60    66          1          1 非常不滿意
6 林妙怡 甲班  女    65    74          1          1 非常不滿意
7 蔡舜遠 甲班  男    67    61          1          1 非常不滿意
8 尤耀貞 甲班  女    68    84          1          1 非常不滿意
9 張鈞興 乙班  男    71    68          1          2 非常不滿意
10 黃千筠 甲班  女    71    92          1          1 非常不滿意
# ... with 40 more rows
# Use `print(n = ...)` to see more rows
>
```



具遺失(缺失)值資料 (Missing Data)

51 / 61

Missing data (missing values for **certain variables for certain cases**): **item non-response**.

When data are missing for **all variables for a given case**: **unit non-response**.

When data are missing for **a variable for all cases**: **latent or unobserved**.

	A	B	C	D	E	F	G
1	ID	C	Y	X1	X2	X3	X4
2	s1	1	78.3	69.6	74.3	NA	5.22
3	s2	2	77	69.9	72.54	NA	3.98
4	s3	3	72.2	65.7	69.74	NA	4.89
5	s4	1	33.4	NA	30.97	NA	21.54
6	s5	2	32.65	28.35	30.54	NA	9.82
7	s6	3	35.45	28.5	32.01	NA	19.81
8	s7	1	424	378	403.55	NA	12.98
9	s8	2	NA	NA	NA	NA	NA
10	s9	3	355	312.5	339.96	NA	14.14
11	s10	1	18.2	15.5	17.19	NA	13.93
12	s11	2	18.3	15.3	16.38	NA	6.92
13	s12	3	16.1	13.9	14.92	NA	10.15
14	s13	1	23.75	20.2	22.19	NA	32.81

若資料出現遺失值:

計算及演算法無法進行。

影響估計量的性質。

(e.g. means, percentages, percentiles, variances, ratios, regression parameters, etc.).

影響統計推論。

(e.g., the properties of tests and confidence intervals.)

- The missing values may give clues to **systematic aspects of the problem**.
- 如何處理遺失值:
 - 不處理，換分析演算法。
 - 刪除法。
 - 用一全域值做填補: Use a **global constant** to fill the value will misguide the mining process. (例如: 缺考給0分; 影像訊號=前景-背景)
 - 用平均或中位數等統計量做填補: Use the **attribute mean** or **median** for all samples belonging to the **same class** as the given tuple.
 - 補值法 (Missing value imputation) (most popular)

- **Neutral-value (中性值) substitution:** You can use neutral value of the likert scale. In your case neutral value would be 3.5 but if your overall average score is less than 3.5 then you could not able to use this method.
- **Mean-value substitution:** only when the number of respondents with missing data and the number of items missing were 20% or less.
- **Imputation methods:** e.g, Approximate Bayesian bootstrap with Propensity score



- Downey, R. G., & King, C. V. (1998). Missing data in Likert ratings: A comparison of replacement methods. The Journal of general psychology, 125(2), 175-191. (被引用 1009 次)
- Shrive, F. M., Stuart, H., Quan, H., & Ghali, W. A. (2006). Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. BMC medical research methodology, 6, 1-10. (被引用 832 次)
- Carpita, M., & Manisera, M. (2011). On the imputation of missing data in surveys with Likert-type scales. Journal of Classification, 28, 93-112. (被引用 68 次)
- Wu, W., Jia, F., & Enders, C. (2015). A comparison of imputation strategies for ordinal missing data on Likert scale variables. Multivariate behavioral research, 50(5), 484-503. (被引用 115 次)
- Applied Missing Data Analysis with SPSS and R
<https://bookdown.org/mwheymans/bookmi/missing-data-in-questionnaires.html>

(Missingness Mechanism)

collected data

$$X = \{X_o, X_m\}$$

observed elements

missing elements

The missingness indicator matrix R corresponds X ,
and each element of R is 1 if the corresponding element of X is missing,
and 0 otherwise.

define the missingness mechanism as

the probability of R conditional on

the values of the observed and missing elements of X :

$$Pr(R|X_o, X_m)$$



Missing Completely at Random

- 依設計產生的遺失 (Missing by Design)
 - Excluded some participants from the analysis because they are **not part of** the population under investigation.
 - **missingness codes**: (i) refused to answer; (ii) answered don't know; (iii) had a valid skip or (iv) was skipped by an enumerator error.
- 完全隨機遺失 (Missing Completely at Random, MCAR)
 - missingness is **independent** of their own unobserved values and the observed data.
$$Pr(R|X) = Pr(R)$$
 - **例**. **Miscoding or forgetting to log in answer.**
 - Imputation methods rely on the missingness being of the **MCAR** type.



Missing at Random (MAR) Missing Not at Random (MNAR)

56 / 61

- 隨機遺失 (Missing at Random, MAR) $Pr(R|X) = Pr(R|X_o)$
 - missingness does not depend on their **unobserved value** but does dependent on the **observed data**.
 - **例 1**: male participants (**observed data**) are more likely to refuse to fill out the **depression survey**, but it does not depend on the level of their depression (**unobserved value**).
 - **例 2**: if men are more likely to tell you their weight than women, **weight** is MAR.
 - We **can ignore missing data** (= omit missing observations) if we have **MAR** or **MCAR**.
- 非隨機遺失 (Missing Not at Random, MNAR)
 - Missingness that **depends on the missing** value itself.
 - **例**: question about **income**, where the high rate of missing values (usually 20%~50%) is related to the value of the income itself (**very high and very low values will not be answered**).
 - **MNAR data is a more serious issue. (not ignorable)**



一些注意事項

57 / 61

- Assuming data is **MCAR**, too much missing data can be a problem.
 - Usually a safe maximum threshold is **5%** of the total for large datasets.
 - If missing data for a certain feature or sample is more than **5%** then you probably should leave that feature or sample **out**.
- If some variable is missing almost **25%** of the data points.
 - Consider either dropping it from the analysis or gather more measurements.
 - Keep the other variables are below the **5%** threshold.
- 類別變數的補值(categorical variable): replacing categorical variables is usually **not advisable**.
 - Some common practice include replacing missing categorical variables with the **mode** of the observed ones (**questionable**).
- 我的資料有需要做補值嗎?
- 補值後的資料不可改變「原資料結構」!
- 常聽到:「資料補值後, 分類演算法的正確率提昇了」?!



R Packages for Dealing With Missing Values

58 / 100

- **Amelia (Amelia II)**: A Program for Missing Data
- **hot.deck**: Multiple Hot-Deck Imputation <https://cran.r-project.org/web/packages/package-name/>
- **HotDeckImputation**: Hot Deck Imputation Methods for Missing Data
- **impute**: (Bioconductor) Imputation for Microarray Data
- **mi**: Missing Data Imputation and Model Checking
- **mice**: **Multivariate Imputation by Chained Equations**
- **missForest**: Nonparametric Missing Value Imputation using Random Forest
- **missMDA**: Handling Missing Values with Multivariate Data Analysis (e.g., `imputePCA`, `imputeMCA`),
- **mitools**: Tools for Multiple Imputation of Missing Data
- **norm**: Analysis of Multivariate Normal Datasets with Missing Values
- **VIM**: Visualization and Imputation of Missing Values
- R packages support for missing values imputation.
 - **Hmisc**: Harrell Miscellaneous
 - **survey**: analysis of complex survey samples
 - **Zelig**: Everyone's Statistical Software
 - **rfImpute{randomForest}**: Imputations by randomForest
 - **imputation{rminer}**: Data Mining Classification and Regression Methods, Missing data imputation (e.g. substitution by value or hotdeck method).
 - **impute.svd{bcv}**: Cross-Validation for the SVD (Bi-Cross-Validation), Missing value imputation via a low-rank SVD approximation estimated by the EM algorithm.
 - **mlr**: Machine Learning in R provides several imputation methods.
<https://mlr-org.github.io/mlr-tutorial/release/html/index.html>

Package "**imputation**" was removed from the CRAN. (Archived on 2014-01-14)



列表刪除法 (List-wise Deletion)

59 / 100

- Also called the **complete case analysis**.
- All units with missing data for a variable are removed and the analysis is performed with the remaining units (complete cases).
- This is the default approach in most statistical packages.
- The use of this method is only justified if the missing data generation mechanism is **MCAR**.
- In R, using the function `na.omit()` or extract complete observations using the function `complete.cases()`.

```
> mdata <- matrix(rnorm(15), nrow=5)
> mdata[sample(1:15, 4)] <- NA
> mdata <- as.data.frame(mdata)
> mdata
```

	V1	V2	V3
1	-0.62222501	1.0807983	NA
2	0.07124865	0.5216675	-0.08334454
3	1.70707399	0.1004917	0.88197789
4	NA	-0.6595201	-0.08387860
5	NA	1.6138847	NA

```
> (x1 <- na.omit(mdata))
```

	V1	V2	V3
2	0.07124865	0.5216675	-0.08334454
3	1.70707399	0.1004917	0.88197789

```
> (x2 <- mdata[complete.cases(mdata),])
```

	V1	V2	V3
2	0.07124865	0.5216675	-0.08334454
3	1.70707399	0.1004917	0.88197789

```
> mdata[!complete.cases(mdata),]
```

	V1	V2	V3
1	-0.622225	1.0807983	NA
4	NA	-0.6595201	-0.0838786
5	NA	1.6138847	NA

快速分析一下，得知資料大概狀況



Mean/Median Substitution

60 / 100

- A very simple but popular approach is to substitute means for the missing values.
- The method preserves sample size and does not reduce the statistical power associated with sample size in comparison with list-wise or pairwise deletion.
- This method produces biased estimates and can severely distort the distribution of the variable in which missing values are substituted.
- This results in underestimates of the standard deviations and **distorts relationships between variables** (estimates of the correlation are pulled toward zero).

Due to these **distributional problems**, it is often **recommended to ignore missing values rather than impute values by mean substitution** (Little and Rubin, 1989.)

```
mean.subst <- function(x) {  
  x[is.na(x)] <- mean(x, na.rm = TRUE)  
  x  
}
```

```
median(x, na.rm = TRUE)
```

```
> mdata  
      V1      V2      V3  
1 -0.6222501  1.0807983    NA  
2  0.07124865  0.5216675 -0.08334454  
3  1.70707399  0.1004917  0.88197789  
4      NA -0.6595201 -0.08387860  
5      NA  1.6138847    NA  
> mdata.mip <- apply(mdata, 2, mean.subst)  
> mdata.mip  
      V1      V2      V3  
[1,] -0.6222501  1.0807983  0.23825158  
[2,]  0.07124865  0.5216675 -0.08334454  
[3,]  1.70707399  0.1004917  0.88197789  
[4,]  0.38536588 -0.6595201 -0.08387860  
[5,]  0.38536588  1.6138847  0.23825158
```

K-Nearest Neighbour Imputation

- The k-nearest neighbour imputation searches for the k-nearest observations (respective to the observation which has to be imputed) and replaces the missing value with the mean of the found k observations.
- It is recommended to use the (weighted) median instead of the arithmetic mean.
- **KNN minimize** data modeling assumptions and take advantage of the **correlation structure** of the data.

	C_1	C_2	$\cdots$	C_j	$\cdots$	C_n
g_1	*	✓		*		✓
g_2	■	✓		✓		✓
$\vdots$						
$\vdots$	■	✓		*		✓
$\vdots$						
g_i	■	✓				✓
$\vdots$						
$\vdots$						
g_m		*		*		

KNNimpute

Model:

$$\{g_{(k)}, k = 1, 2, \dots, K\} = \underset{k}{\operatorname{args\,max}} \operatorname{Corr}(g_1, g_i)$$

$$\{g_{(k)}, k = 1, 2, \dots, K\} = \underset{k}{\operatorname{args\,min}} \operatorname{Dist}(g_1, g_i)$$

C : Observed C_i 's without missing values

Imputation:

$$\text{Average} \quad \widehat{C_1(g_1)} = \frac{1}{K} \sum_{k=1}^K C_1(g_k)$$

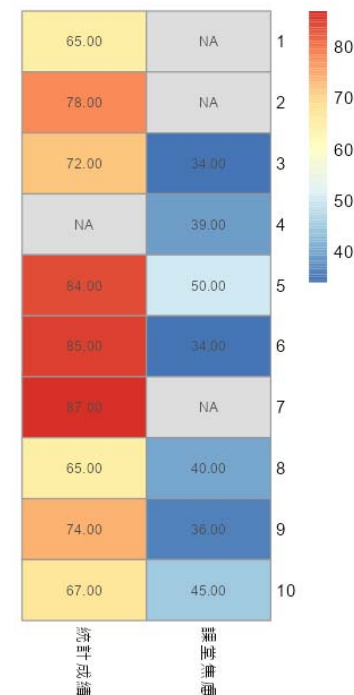
$$\text{Weighted Average} \quad \widehat{C_1(g_1)} = \frac{\sum_{k=1}^K w_k C_1(g_k)}{\sum_{k=1}^K w_k}$$

$$w_k = \frac{1}{\sum_{j \in C} [C_j(g_k) - C_1(g_1)]^2}$$

knn.impute{bnstruct}

62 / 100

```
> library(bnstruct)
>
> missing_data <- read.csv("data/置換遺漏值.csv")
> missing_data
  編號 統計成績 課堂焦慮
1     1      65      NA
2     2      78      NA
...
10    10      67      45
> str(missing_data)
'data.frame':      10 obs. of  3 variables:
 $ 編號      : int  1 2 3 4 5 6 7 8 9 10
 $ 統計成績: int  65 78 72 NA 84 85 87 65 74 67
 $ 課堂焦慮: int  NA NA 34 39 50 34 NA 40 36 45
> summary(missing_data)
      編號      統計成績      課堂焦慮
Min.   : 1.00   Min.   :65.00   Min.   :34.00
1st Qu.: 3.25   1st Qu.:67.00   1st Qu.:35.00
Median : 5.50   Median :74.00   Median :39.00
Mean   : 5.50   Mean   :75.22   Mean   :39.71
3rd Qu.: 7.75   3rd Qu.:84.00   3rd Qu.:42.50
Max.   :10.00   Max.   :87.00   Max.   :50.00
      NA's      :1      NA's      :3
> library(pheatmap)
> pheatmap(missing_data[, 2:3],
+          display_numbers = T,
+          cluster_rows = FALSE,
+          cluster_cols = FALSE)
> knn.impute(as.matrix(missing_data$統計成績), k = 5)
> knn.impute(as.matrix(missing_data), k = 5)
```



Regression Methods

- Using **fitted regression values** to replace missing values.
- The model must be chosen so that it does not yields **invalid fitted values**.
e.g., negative values.
- This technique might be more accurate than simply substituting a measure of central tendency, since the imputed value is based on other input variables.
- This technique underestimates standard errors by underestimating the variance in x.

	C_1	C_2	$\cdots$	C_j	$\cdots$	C_n
g_1	*	✓		*		✓
g_2	■	✓		✓		✓
⋮						
⋮	■	✓		*		✓
⋮						
g_i	■	✓		✓		✓
⋮						
⋮						
g_m		*		*		

Regression

Model:

$$C_1 = \beta_0 + \sum_{j \in C} \beta_j C_j$$

C: Observed C_i 's
without missing values

Imputation:

$$\widehat{C_1}(g_1) = \hat{\beta}_0 + \sum_{j \in C} \hat{\beta}_j C_j(g_1)$$



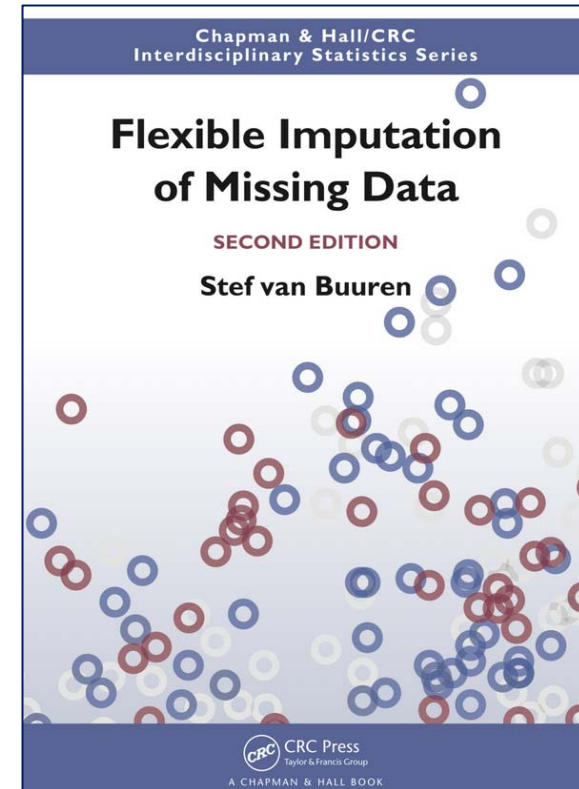
R Package: **mice**

64 / 100

- **mice**: Multivariate Imputation by **Chained Equations** in R by Stef van Buuren.
- Imputing missing values on:
 - **Continuous data**: Predictive mean matching, Bayesian linear regression, Linear regression ignoring model error, Unconditional mean imputation etc.
 - **Binary data**: Logistic Regression, Logistic regression with bootstrap
 - **Categorical data** (More than 2 categories) - Polytomous logistic regression, Proportional odds model etc.
 - **Mixed data** (Can work for both Continuous and Categorical) - CART, Random Forest, Sample (Random sample from the observed values).

電子書

Flexible Imputation of Missing Data



<https://stefvanbuuren.name/fimd>

Source: <http://www.listendata.com/2015/08/missing-imputation-with-mice-package-in.html>



Imputation using **MICE** Package^{65 / 100}

```
> mydata <- airquality
> mydata[4:10,3] <- rep(NA,7)
> mydata[1:5,4] <- NA
>
> #Use numerical variables as examples here.
> #Ozone is the variable with the most missing datapoints.
> data <- mydata[-c(5,6)]
> summary(mydata)
```

Ozone	Solar.R	Wind	Temp	Month	Day
Min. : 1.00	Min. : 7.0	Min. : 1.700	Min. :57.00	Min. :5.000	Min. : 1.0
1st Qu.: 18.00	1st Qu.:115.8	1st Qu.: 7.400	1st Qu.:73.00	1st Qu.:6.000	1st Qu.: 8.0
Median : 31.50	Median :205.0	Median : 9.700	Median :79.00	Median :7.000	Median :16.0
Mean : 42.13	Mean :185.9	Mean : 9.806	Mean :78.28	Mean :6.993	Mean :15.8
3rd Qu.: 63.25	3rd Qu.:258.8	3rd Qu.:11.500	3rd Qu.:85.00	3rd Qu.:8.000	3rd Qu.:23.0
Max. :168.00	Max. :334.0	Max. :20.700	Max. :97.00	Max. :9.000	Max. :31.0
NA's :37	NA's :7	NA's :7	NA's :5		

```
>
> #Check the missing percentages for features (columns) and samples (rows)
> pMiss <- function(x){sum(is.na(x))/length(x)*100}
> apply(mydata, 2, pMiss)
```

Ozone	Solar.R	Wind	Temp	Month	Day
24.183007	4.575163	4.575163	3.267974	0.000000	0.000000

```
> apply(mydata, 1, pMiss)
```

[1]	16.66667	16.66667	16.66667	33.33333	66.66667	33.33333	16.66667	16.66667	16.66667
33.33333	16.66667	0.00000							
...									
[145]	0.00000	0.00000	0.00000	0.00000	0.00000	16.66667	0.00000	0.00000	0.00000

Source: <http://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/>



Visualizing the Pattern of Missing Data

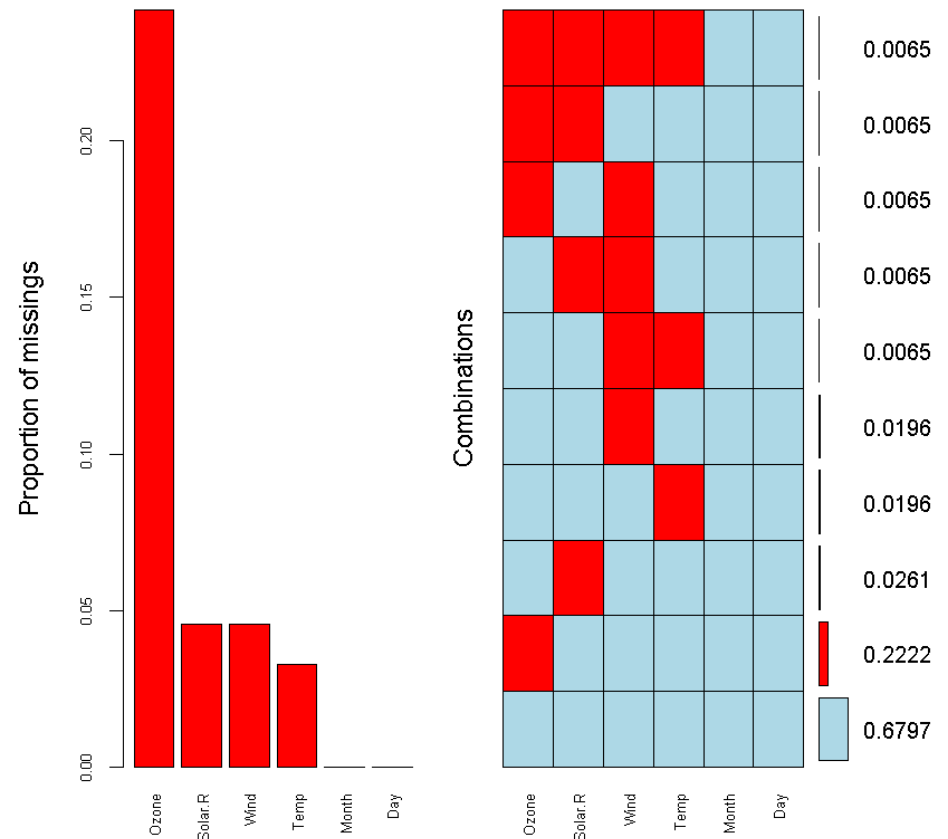
```
> library(mice)
> md.pattern(mydata)
      Month Day Temp Solar.R Wind Ozone
104      1   1   1      1     1     1   0
34      1   1   1      1     1     0   1
4       1   1   1      0     1     1   1
3       1   1   1      1     0     1   1
3       1   1   0      1     1     1   1
1       1   1   1      0     1     0   2
1       1   1   1      1     0     0   2
1       1   1   1      0     0     1   2
1       1   1   0      0     0     1   2
1       0   0   5      7     7     0   4
      37 56
```

```
> library(VIM)
> mydata.aggrplot <- aggr(mydata,
col=c('lightblue','red'), numbers=TRUE,
prop = TRUE, sortVars=TRUE,
labels=names(mydata), cex.axis=.7, gap=3)
```

Variables sorted by number of missings:

Variable	Count
Ozone	0.24183007
Solar.R	0.04575163
Wind	0.04575163
Temp	0.03267974
Month	0.00000000
Day	0.00000000

#104 samples are complete, 34 samples miss only the Ozone measurement, 4 samples miss only the Solar.R value and so on.





Number of Observations Per Patterns for All Pairs of Variables

67 / 100

```
> md.pairs(mydata)
```

\$rr

	Ozone	Solar.R	Wind	Temp	Month	Day
Ozone	116	111	111	112	116	116
Solar.R	111	146	141	142	146	146
Wind	111	141	146	143	146	146
Temp	112	142	143	148	148	148
Month	116	146	146	148	153	153
Day	116	146	146	148	153	153

\$rm

	Ozone	Solar.R	Wind	Temp	Month	Day
Ozone	0	5	5	4	0	0
Solar.R	35	0	5	4	0	0
Wind	35	5	0	3	0	0
Temp	36	6	5	0	0	0
Month	37	7	7	5	0	0
Day	37	7	7	5	0	0

- **rr**: response-response, both variables are observed
- **rm**: response-missing, row observed, column missing
- **mr**: missing-response, row missing, column observed
- **mm**: missing-missing, both variables are missing

\$mr

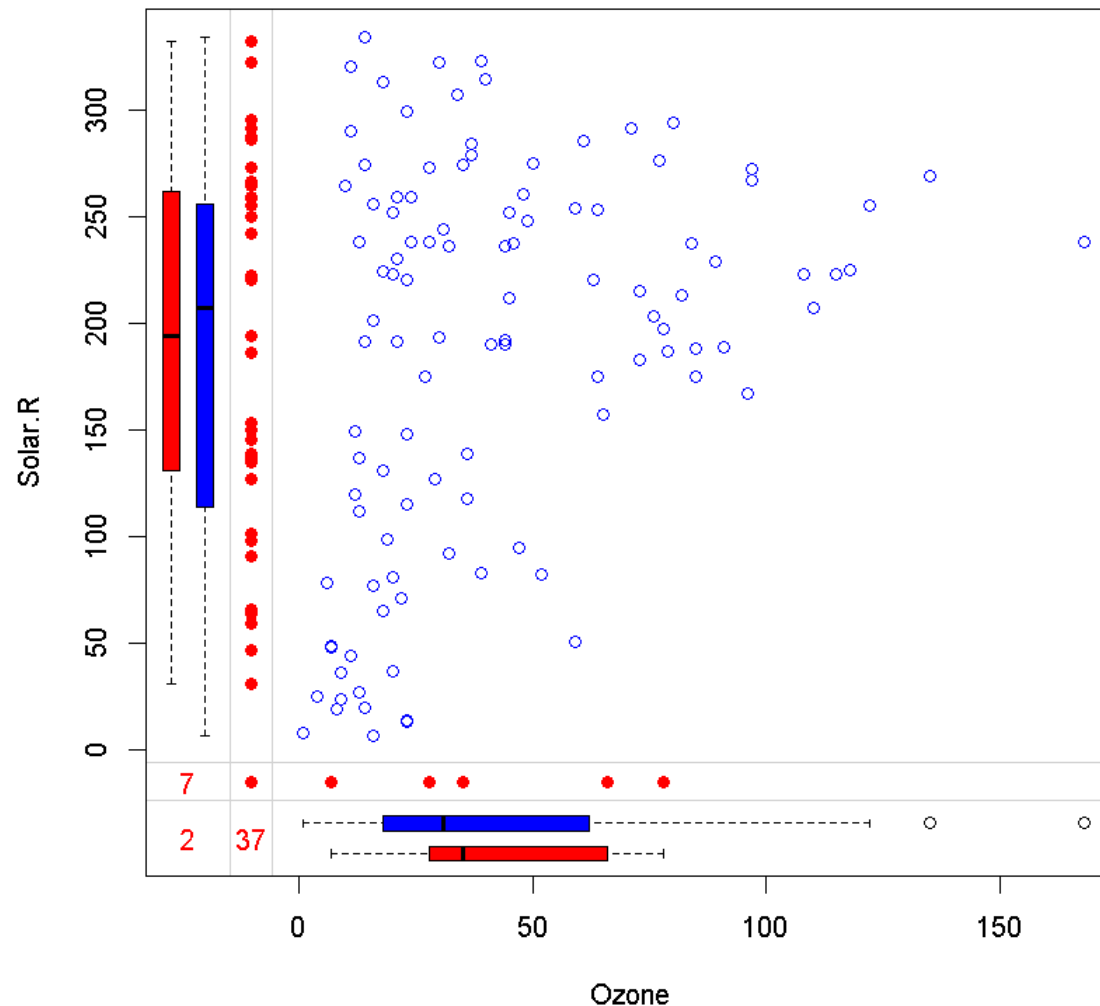
	Ozone	Solar.R	Wind	Temp	Month	Day
Ozone	0	35	35	36	37	37
Solar.R	5	0	5	6	7	7
Wind	5	5	0	5	7	7
Temp	4	4	3	0	5	5
Month	0	0	0	0	0	0
Day	0	0	0	0	0	0

\$mm

	Ozone	Solar.R	Wind	Temp	Month	Day
Ozone	37	2	2	1	0	0
Solar.R	2	7	2	1	0	0
Wind	2	2	7	2	0	0
Temp	1	1	2	5	0	0
Month	0	0	0	0	0	0
Day	0	0	0	0	0	0

Marginplot

```
> marginplot(mydata[,c("Ozone", "Solar.R")], col = c("blue", "red"))
```



- The blue box plot located on the left and bottom margins shows the distribution of the non-missing datapoints.
- The red box plot on the left shows the distribution of `Solar.R` with `Ozone` missing while
- Likewise for the `Ozone` box plots at the bottom of the graph.
- If our assumption of MCAR data is correct, then we expect the red and blue box plots to be very similar.



Generates Multivariate Imputations by Chained Equations (MICE)

```
mice(data, m = 5, method = vector("character", length = ncol(data)),  
      predictorMatrix = (1 - diag(1, ncol(data))),  
      visitSequence = (1:ncol(data))[apply(is.na(data), 2, any)],  
      form = vector("character", length = ncol(data)),  
      post = vector("character", length = ncol(data)), defaultMethod = c("pmm",  
      "logreg", "polyreg", "polr"), maxit = 5, diagnostics = TRUE,  
      printFlag = TRUE, seed = NA, imputationMethod = NULL,  
      defaultImputationMethod = NULL, data.init = NULL, ...)
```

```
> methods(mice)
```

```
[1] mice.impute.2l.norm      mice.impute.2l.pan      mice.impute.2lonly.mean  
[4] mice.impute.2lonly.norm  mice.impute.2lonly.pmm  mice.impute.cart  
[7] mice.impute.fastpmm      mice.impute.lda         mice.impute.logreg  
[10] mice.impute.logreg.boot  mice.impute.mean        mice.impute.norm  
[13] mice.impute.norm.boot    mice.impute.norm.nob    mice.impute.norm.predict  
[16] mice.impute.passive      mice.impute.pmm         mice.impute.polr  
[19] mice.impute.polyreg      mice.impute.quadratic   mice.impute.rf  
[22] mice.impute.ri           mice.impute.sample      mice.mids  
[25] mice.theme
```

see '?methods' for accessing help and source

Warning message:

In .S3methods(generic.function, class, paren
function 'mice' appears not to be S3 generic; fou

Method	Description	Scale type	Default
pmm	Predictive mean matching	numeric	Y
norm	Bayesian linear regression	numeric	
norm.nob	Linear regression, non-Bayesian	numeric	
mean	Unconditional mean imputation	numeric	
2L.norm	Two-level linear model	numeric	
logreg	Logistic regression	factor, 2 levels	Y
polyreg	Multinomial logit model	factor, >2 levels	Y
polr	Ordered logit model	ordered, >2 levels	Y
lda	Linear discriminant analysis	factor	
sample	Random sample from the observed data	any	

PMM (Predictive Mean Matching) – For numeric \

Impute Missing Values

```
> mydata.ip <- mice(mydata, m = 5, maxit = 50, meth = 'pmm', seed = 500)
```

```
iter imp variable
```

```
1 1 Ozone Solar.R Wind Temp
```

```
1 2 Ozone Solar.R Wind Temp
```

```
...
```

```
50 4 Ozone Solar.R Wind Temp
```

```
50 5 Ozone Solar.R Wind Temp
```

```
> summary(mydata.ip)
```

Multiply imputed data set

Call:

```
mice(data = mydata, m = 5, method = "pmm", maxit = 50, seed = 500)
```

Number of multiple imputations: 5

Missing cells per column:

```
Ozone Solar.R Wind Temp Month Day
```

```
37 7 7 5 0 0
```

Imputation methods:

```
Ozone Solar.R Wind Temp Month Day
```

```
"pmm" "pmm" "pmm" "pmm" "pmm" "pmm"
```

VisitSequence:

```
Ozone Solar.R Wind Temp
```

```
1 2 3 4
```

PredictorMatrix:

```
Ozone Solar.R Wind Temp Month Day
```

```
Ozone 0 1 1 1 1 1
```

```
Solar.R 1 0 1 1 1 1
```

```
Wind 1 1 0 1 1 1
```

```
Temp 1 1 1 0 1 1
```

```
Month 0 0 0 0 0 0
```

```
Day 0 0 0 0 0 0
```

Random generator seed value: 500

```
> mydata.ip$imp$Ozone
> 1 2 3 4 5
5 59 85 20 108 18
10 11 7 27 14 21
...
150 9 34 27 12 22
```

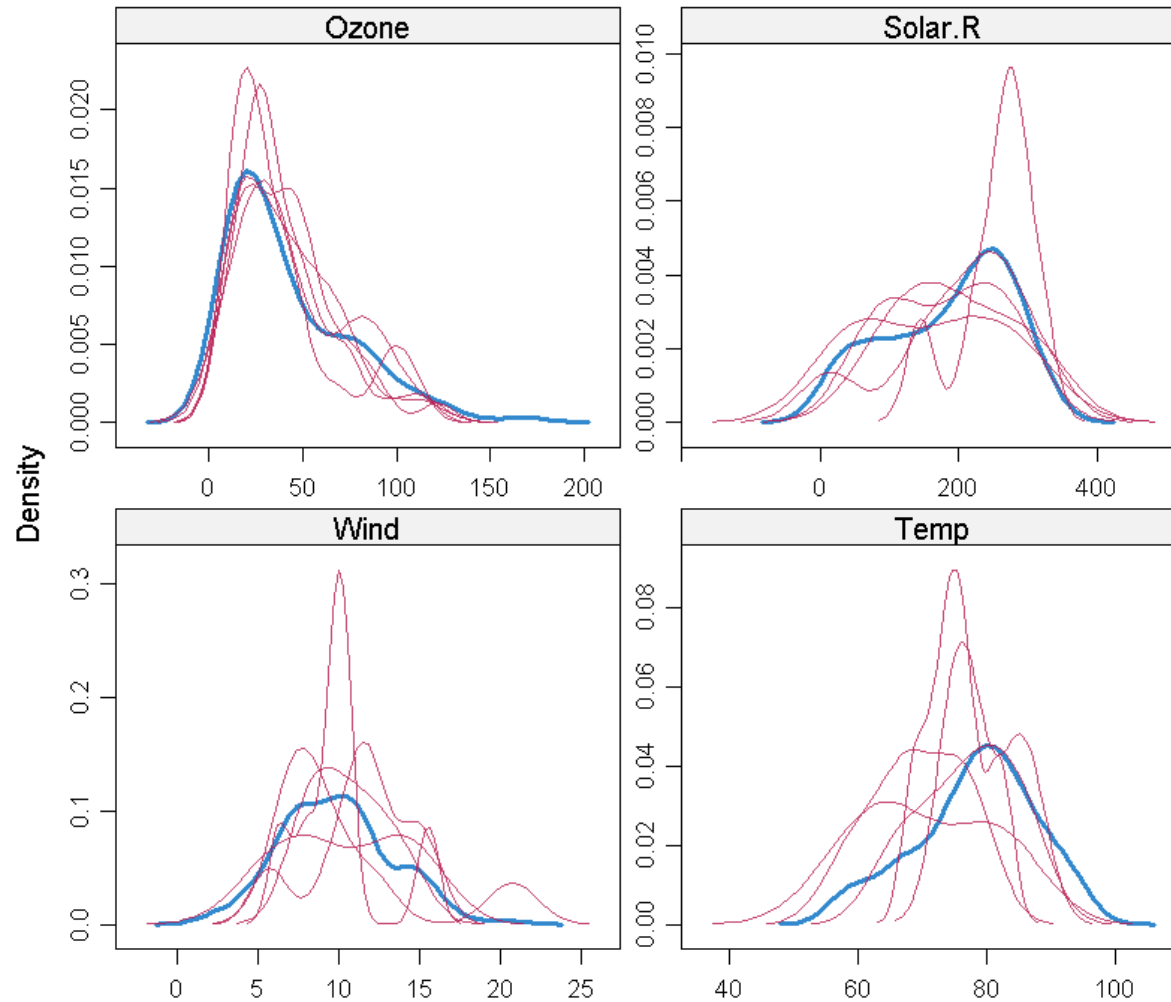
The output shows the imputed data for each observation (first column left) within each imputed dataset (first row at the top).

```
> # get back the first completed dataset out of 5
> mydata.completed <- complete(mydata.ip, 1)
```

Density Plot

71 / 100

```
> densityplot(mydata.ip)
```



The density of the imputed data for each imputed dataset is showed in magenta while the density of the observed data is showed in blue. Under MCAR, we expect the distributions to be similar.



Pooling

- Next step: fit a linear model to the data.
- **mice** fit a model to each of the imputed dataset and then pool the results together.

```
> # linear regression for each imputed data set - 5 regression are run
> modelFit1 <- with(mydata.ip, lm(Temp ~ Ozone + Solar.R + Wind))
> # pool coefficients and standard errors across all 5 regression models
> summary(pool(modelFit1))
```

	est	se	t	df	Pr(> t)	lo 95
(Intercept)	71.11418579	2.840129171	25.0390674	85.04465	0.000000e+00	65.467290906
Ozone	0.17412083	0.025108183	6.9348239	72.90551	1.383136e-09	0.124079199
Solar.R	0.01004273	0.007163085	1.4020115	87.03503	1.644683e-01	-0.004194599
Wind	-0.21504110	0.222484210	-0.9665454	61.98616	3.375274e-01	-0.659782671

	hi 95	nmis	fmi	lambda
(Intercept)	76.76108067	NA	0.1459648	0.1261138
Ozone	0.22416246	37	0.1734348	0.1510666
Solar.R	0.02428005	7	0.1418215	0.1223252
Wind	0.22970047	7	0.2026905	0.1773735

To reduce the effect of the random seed initialization, we can impute a higher number of dataset, by changing the default **m = 5** parameter in the **mice()** function.

```
mydata.ip2 <- mice(mydata, m = 50, seed = 245435)
modelFit2 <- with(mydata.ip2, lm(Temp ~ Ozone + Solar.R + Wind))
summary(pool(modelFit2))
```



Quick Tutorial on MICE Package 73 / 100

```
> # Generate 10% missing values at Random
> iris.mis <- prodNA(iris, noNA = 0.1) # library(missForest)
> # Check missing values introduced in the data
> summary(iris.mis)
> iris.mis <- subset(iris.mis, select = -c(Species))
> summary(iris.mis)
>
> # A tabular form of missing value present in each variable
> library(mice)
> md.pattern(iris.mis)
> # Visualization
> library(VIM)
> mice_plot <- aggr(iris.mis, col=c('navyblue','yellow'), numbers=TRUE, sortVars=TRUE,
                    labels=names(iris.mis), cex.axis=.7,
                    gap=3, ylab=c("Missing data","Pattern"))

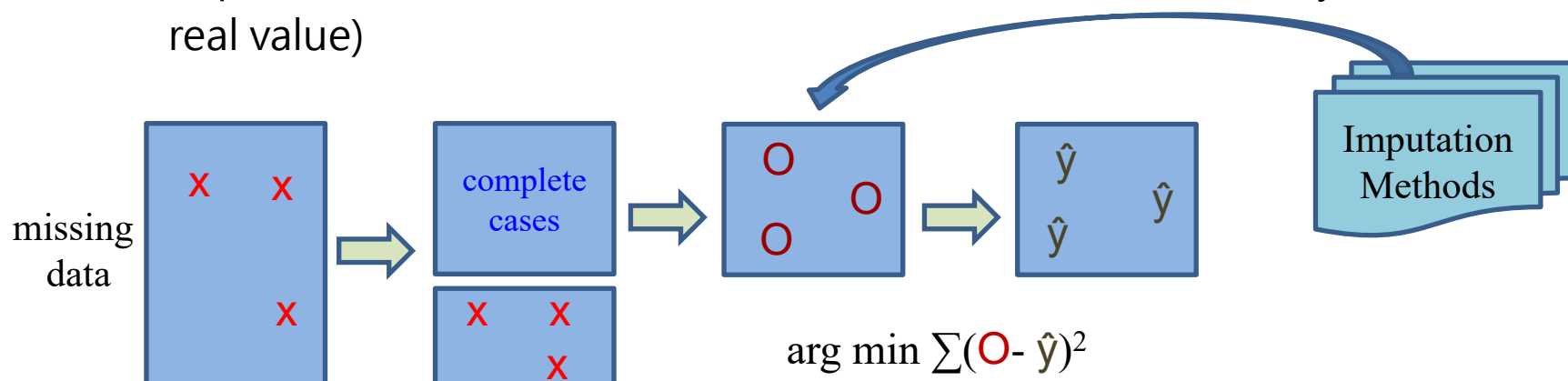
> # Imputation
> imputed_Data <- mice(iris.mis, m=5, maxit = 50, method = 'pmm', seed = 500)
> summary(imputed_Data)
> # Check imputed values
> imputed_Data$imp$Sepal.Width
> # Get complete data ( 2nd out of 5)
> completeData <- complete(imputed_Data,2)
> # Build predictive model
> fit <- with(data = imputed_Data, exp = lm(Sepal.Width ~ Sepal.Length + Petal.Width))
> # Combine results of all 5 models
> combine <- pool(fit)
> summary(combine)
```

Source: <http://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/>

哪一種補值方法較好？

74 / 100

- KNN is the most widely-used.
- Characteristics of data that may affect choice of imputation method:
 - dimensionality.
 - percentage of values missing.
 - experimental design (time series, case/control, etc.)
 - patterns of correlation in data.
- 建議:
 - add (same percentage) artificial missing values to your (complete cases) data set.
 - impute them with various methods, see which is best (since you know the real value)



```
> aggre_data <- read.csv("data/整合資料.csv")
> head(aggre_data)
  性別 學校規模 年齡 工作壓力 工作滿意 組織承諾
1    1         1   32         24         12         24
2    1         1   28         26         14         40
...
6    1         1   29         25         30         54
> str(aggre_data)
'data.frame':      90 obs. of  6 variables:
 $ 性別      : int   1 1 1 1 1 1 1 1 1 1 ...
 $ 學校規模: int   1 1 1 1 1 1 1 1 1 1 ...
 $ 年齡      : int   32 28 25 31 28 29 27 28 28 26 ...
 $ 工作壓力: int   24 26 30 31 24 25 34 40 20 24 ...
 $ 工作滿意: int   12 14 15 21 23 30 29 28 12 21 ...
 $ 組織承諾: int   24 40 54 52 53 54 52 51 32 64 ...
> aggre_data$性別 <- as.factor(aggre_data$性別)
> aggre_data$學校規模 <- as.factor(aggre_data$學校規模)
> str(aggre_data)
'data.frame':      90 obs. of  6 variables:
 $ 性別      : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
 $ 學校規模: Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
 $ 年齡      : int   32 28 25 31 28 29 27 28 28 26 ...
 $ 工作壓力: int   24 26 30 31 24 25 34 40 20 24 ...
 $ 工作滿意: int   12 14 15 21 23 30 29 28 12 21 ...
 $ 組織承諾: int   24 40 54 52 53 54 52 51 32 64 ...
>
> aggregate(aggre_data[, 3:6], by = list(性別 = aggre_data$性別,
+                                       學校規模 = aggre_data$學校規模),
+ FUN = function(x){round(mean(x), 2)})
  性別 學校規模 年齡 工作壓力 工作滿意 組織承諾
1    1         1 29.93      27.20      20.13      48.73
2    2         1 35.82      40.35      22.53      48.24
3    1         2 41.47      32.40      19.87      51.27
4    2         2 39.00      28.93      19.36      48.43
5    1         3 37.36      34.93      26.71      56.00
6    2         3 40.47      32.00      25.80      49.07
```

整合資料：整合函數

摘要統計量	特定值	觀察值個數
☐ 平均數(M)	☐ 第一個(E)	☐ 加權(E)
☐ 中位數(N)	☐ 最後一個(L)	☐ 加權遺漏(D)
☐ 總和(S)	☐ 最小值(U)	☐ 未加權(U)
☐ 標準差(R)	☐ 最大值(X)	☐ 未加權遺漏(U)

百分比

☐ 上(A) 數值： 30

☒ 下(B)

☐ 內(D) 低： 高：

☐ 外(O)

分數

☐ 上(A) 數值：

☐ 下(W)

☐ 內(I) 低： 高：

☐ 外(I)

繼續 取消 輔助說明

```
mean, median, sum, sd
x[1], x[length(x)], min, max
length(which(x > 0)) / length(x)
```

企業組織知識管理調查問卷

一、基本資料

1. 我的性別：☐男生 ☐女生

2. 我的教育程度：☐國小 ☐國中 ☐高中職 ☐專科大學 ☐研究所

3. 我的服務年資：☐5年以下 ☐6-10年 ☐11-15年 ☐16-20年 ☐21年以上

二、知識管理

	非常 不同意		←————→		非常 同意
1. 我覺得公司常請專家學者來授課或派員到外界接受訓練。	☐	☐	☐	☐	☐
2. 我覺得公司有設置各種知識庫或書面資料等供員工學習。	☐	☐	☐	☐	☐
3. 我覺得公司常透過教育訓練方式傳授工作的知能與技術。	☐	☐	☐	☐	☐
4. 我覺得公司員工常會把經驗心得用口語、書面、實做表達。	☐	☐	☐	☐	☐
5. 我覺得公司會注重資料的蒐集、分析與分類並加予儲存。	☐	☐	☐	☐	☐
6. 我覺得公司員工不善用資訊科技尋找工作相關知識。	☐	☐	☐	☐	☐
7. 我覺得公司員工常會將所獲得的知識在工作中嘗試。	☐	☐	☐	☐	☐
8. 我覺得公司員工常用電腦設備與網路系統傳遞內部資訊。	☐	☐	☐	☐	☐
9. 我覺得公司未建置多元溝通管道來與員工或外界傳遞資訊。	☐	☐	☐	☐	☐
10. 我覺得公司經常採用各種不同的方法改善工作的流程。	☐	☐	☐	☐	☐

- 知識管理量表經**預試效度分析**，建構效度包含二個層面(構念)。
 - 因素一包含題項1至題項6，命名為知識獲取。
 - 因素二包含題項7至題項10，命名為知識流通。
- 題項1至題項10所測量的特質，共同因素為「知識管理」。
- 第6題，第9題是反向題。

```
> trans_data_orig <- read.csv("data/資料轉換_1.csv")
> head(trans_data_orig)
  編號 性別 教育程度 服務年資 a1 a2 a3 a4 a5 a6 a7 a8 a9 a10
1    1    1      1      1      5 5 1 5 2 1 5 3 3 1 5
2    2    1      1      1      5 5 2 5 2 2 4 4 2 3 5
...
6    6    1      2      4 4 2 5 2 2 4 3 6 5 1
> str(trans_data_orig)
'data.frame':      55 obs. of  14 variables:
 $ 編號      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ 性別      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ 教育程度  : int  1 1 2 2 2 2 2 2 2 2 ...
 $ 服務年資  : int  5 5 4 4 4 4 4 4 4 4 ...
 $ a1        : int  5 5 5 5 1 4 4 4 1 4 ...
 ...
 $ a10       : int  5 5 4 3 2 1 2 3 5 4 ...
> summary(trans_data_orig[, c("性別", "教育程度", "服務年資")])
      性別      教育程度      服務年資
Min.   :1.000   Min.   :1.000   Min.   :1.000
1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000
Median :1.000   Median :3.000   Median :3.000
Mean   :1.436   Mean   :3.091   Mean   :2.636
3rd Qu.:2.000   3rd Qu.:4.000   3rd Qu.:4.000
Max.   :3.000   Max.   :5.000   Max.   :5.000
```

有資料輸入錯誤

要轉換成正確的類別



次數分配表

78 / 100

```
> trans_data <- trans_data_orig
> gender <- c("男生", "女生", NA)
> edu <- c("國小", "國中", "高中職", "專科大學", "研究所")
> work_year <- c("5年以下", "6-10年", "11-15年", "16-20年", "21年以上")
>
> trans_data$性別 <- factor(gender[trans_data_orig$性別], levels = gender)
> trans_data$教育程度 <- factor(edu[trans_data_orig$教育程度], levels = edu, ordered = T)
> trans_data$服務年資 <- factor(work_year[trans_data_orig$服務年資], levels = work_year, ordered = T)
> summary(trans_data[, c("性別", "教育程度", "服務年資")])
```

更正各背景變項鍵入錯誤數值。

	性別	教育程度	服務年資
男生	:33	國小 : 2	5年以下 :12
女生	:20	國中 :14	6-10年 :13
NA's	: 2	高中職 :17	11-15年 :15
		專科大學 :21	16-20年 :13
		研究所 : 1	21年以上: 2

```
>
> tbl_edu <- table(trans_data$教育程度) 需組別合併
> n <- length(trans_data$教育程度)
> freq_data <- data.frame(次數 = tbl_edu, 百分比 = round(tbl_edu/n, 2),
+ 累積次數 = cumsum(tbl_edu), 累積百分比 = round(cumsum(tbl_edu/n), 2))
> freq_data$次數.Var1 <- NULL
> freq_data$百分比.Var1 <- NULL
> freq_data
```

	次數.Freq	百分比.Freq	累積次數	累積百分比
國小	2	0.04	2	0.04
國中	14	0.25	16	0.29
高中職	17	0.31	33	0.60
專科大學	21	0.38	54	0.98
研究所	1	0.02	55	1.00

長條圖、圓餅圖

79 / 100

```
library(ggplot2)
ggplot(trans_data, aes(x = 教育程度)) +
  geom_bar() +
  labs(x = "教育程度", y = "次數")
```

```
trans_df <- data.frame(table(trans_data$教育程度))
names(trans_df) <- c("教育程度", "次數")
trans_df
```

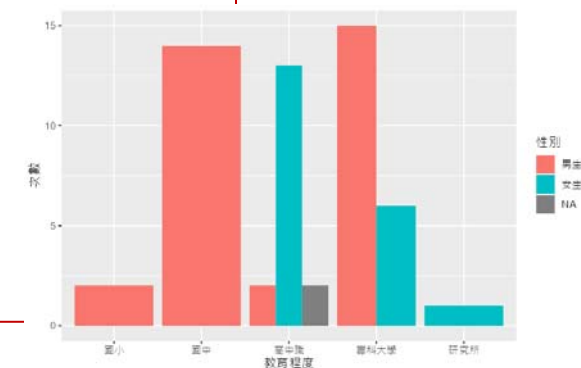
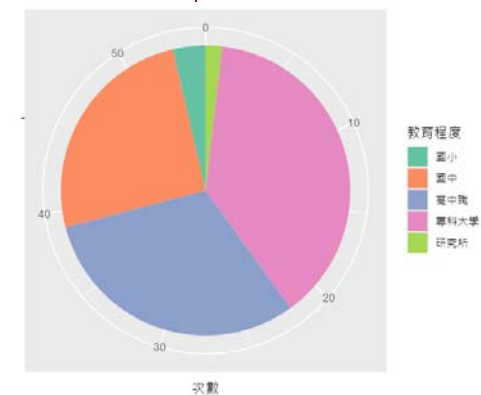
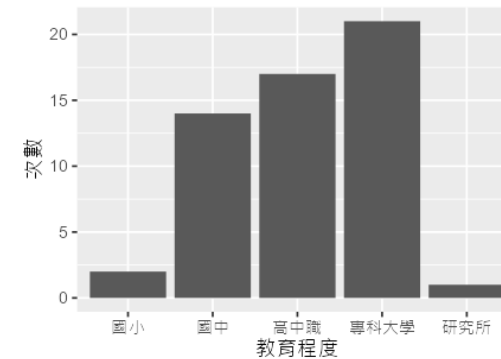
教育程度	次數
1 國小	2
2 國中	14
3 高中職	17
4 專科大學	21
5 研究所	1

```
ggplot(trans_df, aes(x = "", y = 次數, fill = 教育程度)) +
  geom_bar(width = 1, stat = "identity") +
  labs(x = "", fill = "教育程度") +
  coord_polar("y", start=0) +
  scale_fill_brewer(palette="Set2")
```

```
table(trans_data$性別, trans_data$教育程度)
```

	國小	國中	高中職	專科大學	研究所
男生	2	14	2	15	0
女生	0	0	13	6	1

```
ggplot(trans_data, aes(x = 教育程度, fill = 性別)) +
  geom_bar(position = "dodge") +
  labs(x = "教育程度", y = "次數")
```



組別合併: 重新編碼 (合併過少觀察值之組別)

80 / 100

```
> edu2 <- c("國中以下", "國中以下", "高中職", "專科大學以上", "專科大學以上")
> trans_data$教育程度 <- factor(edu2[trans_data_orig$教育程度], levels = unique(edu2),
ordered = T)
> table(trans_data$教育程度)
```

國中以下	高中職	專科大學以上
16	17	22

教育程度重新編碼

1 => 1 · 2 => 1 · 3 => 2 · 4 => 3 · 5 => 3

國小、國中 => 國中以下

專科大學、研究所 => 專科大學以上

```
>
> work_year2 <- c("5年以下", "6-10年", "11-15年", "16年以上", "16年以上")
> trans_data$服務年資 <- factor(work_year2[trans_data_orig$服務年資], levels =
unique(work_year2), ordered = T)
> table(trans_data$服務年資)
```

5年以下	6-10年	11-15年	16年以上
12	13	15	15

服務年資重新編碼

5 => 4

16-20年、21年以上 => 16年以上

```
>
> str(trans_data)
'data.frame':    55 obs. of  14 variables:
 $ 編號      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ 性別      : Factor w/ 2 levels "男生","女生": 1 1 1 1 1 1 1 1 1 1 ...
 $ 教育程度: Ord.factor w/ 3 levels "國中以下"<"高中職"<..: 1 1 1 1 1 1 1 1 1 1 ...
 $ 服務年資: Ord.factor w/ 4 levels "5年以下"<"6-10年"<..: 4 4 4 4 4 4 4 4 4 4 ...
 $ a1       : int  5 5 5 5 1 4 4 4 1 4 ...
 ...
 $ a10      : int  5 5 4 3 2 1 2 3 5 4 ...
> trans_data$服務年資
[1] 16年以上 16年以上 16年以上 16年以上 16年以上 16年以上 16年以上 ...
Levels: 5年以下 < 6-10年 < 11-15年 < 16年以上
> as.integer(trans_data$服務年資)
[1] 4 4 4 4 4 4 4 4 4 4 4 4 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2
[36] 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```



題項檢核: 描述性統計量 (針對單一題項修正錯誤值)

81 / 100

```
> summary(trans_data[, 5:14])
```

a1		a2		a3		a4		a5	
Min.	:1.000	Min.	:1.000	Min.	:1.000	Min.	:1.000	Min.	:1.000
1st Qu.:	2.000	1st Qu.:	1.000	1st Qu.:	3.000	1st Qu.:	1.000	1st Qu.:	2.000
Median	:4.000	Median	:2.000	Median	:4.000	Median	:2.000	Median	:2.000
Mean	:3.527	Mean	:2.109	Mean	:3.564	Mean	:2.255	Mean	:2.727
3rd Qu.:	5.000	3rd Qu.:	2.000	3rd Qu.:	4.000	3rd Qu.:	3.000	3rd Qu.:	3.500
Max.	:5.000	Max.	:5.000	Max.	:5.000	Max.	:5.000	Max.	:5.000

a6		a7		a8		a9		a10	
Min.	:1.000	Min.	:1.000	Min.	:1.000	Min.	:1.0	Min.	:1.000
1st Qu.:	3.000	1st Qu.:	3.000	1st Qu.:	3.000	1st Qu.:	2.0	1st Qu.:	2.000
Median	:4.000	Median	:3.000	Median	:5.000	Median	:3.0	Median	:3.000
Mean	:3.818	Mean	:3.236	Mean	:4.073	Mean	:3.2	Mean	:3.164
3rd Qu.:	5.000	3rd Qu.:	4.000	3rd Qu.:	5.000	3rd Qu.:	4.5	3rd Qu.:	5.000
Max.	:5.000	Max.	:5.000	Max.	:6.000	Max.	:5.0	Max.	:5.000

```
> sapply(trans_data[, 5:14], table)
```

\$a1

1	2	3	4	5
3	12	10	13	17

...

\$a8

1	2	3	4	5	6
2	6	9	9	27	2

...

```
> trans_data$a8[trans_data_orig$a8 == 6] <- 5 # NA
> table(trans_data$a8)
```

1	2	3	4	5
2	6	9	9	29

各題數值介於1-5。
最小值有可能大於1
最大值不能超過5



反向題的反向計分

82 / 100

- 問句範例: 「不善用」 「未建置」
- 李克特五點量表:
「非常不同意」 <= => 「非常同意」
- 進行層面加總與總分計算, 須反向計分。
- 「1 -> 5」 「2 -> 4」 ... 「5 -> 1」

```
> trans_data <- trans_data_orig
> table(trans_data$a6)
 1  2  3  4  5
 1  3 16 20 15
> trans_data$a6 <- 6 - trans_data$a6
> table(trans_data$a6)
 1  2  3  4  5
15 20 16  3  1
>
> table(trans_data$a9)
> trans_data$a9 <- 6 - trans_data$a9
> table(trans_data$a9)
```

```
> likert_5 <- c("非常不同意", "不同意", "普通", "同意", "非常同意")
> trans_data$a1 <- factor(likert_5[trans_data_orig$a1], levels = likert_5, ordered = T)
> trans_data$a1
[1] 非常同意 非常同意 非常同意 非常同意 非常不同意 同意 同意
...
Levels: 非常不同意 < 不同意 < 普通 < 同意 < 非常同意
> table(trans_data$a1)
非常不同意 不同意 普通 同意 非常同意
          3      12      10      13      17
> trans_data_a1_to_a10 <- sapply(5:14, function(x){
+   factor(likert_5[trans_data[, x]], levels = likert_5, ordered = T)
+ })
> colnames(trans_data_a1_to_a10) <- paste0("a", 1:10)
> head(trans_data_a1_to_a10)
      a1      a2      a3      a4      a5      a6      a7
[1,] "非常同意" "非常不同意" "非常同意" "不同意" "非常不同意" "非常不同意" "普通"
...
```

- 問卷調查中，某種特質，態度，行為等潛在構念，不能逐題分析。
- 單一題項所測量的不足以代表某一潛在特質或構念。
- 潛在特質或構念通常包含數個題項。
- 屬性相似的題項所要測量的共同特質稱為「建構效度」。
- 「建構效度」中的層面測量是數個題項的加總分數。
- 例如：
 - 「知識獲取」層面: 包含六個題項。
 - 「知識流通」層面: 包含四個題項。



各層面的加總及計算層面單題平均

84 / 100

```
> score_sum <- data.frame(知識獲取層面加總 = rowSums(trans_data[, 5:10]),
+                          知識流通層面加總 = rowSums(trans_data[, 11:14]),
+                          知識管理層面加總 = rowSums(trans_data[, 5:14]))
>
> head(score_sum)
      知識獲取層面加總  知識流通層面加總  知識管理層面加總
1                   15                   16                   31
2                   18                   14                   32
3                   19                   10                   29
4                   18                   11                   29
5                   13                   10                   23
6                   17                   10                   27
>
> score_average <- data.frame(知識獲取層面平均 = rowMeans(trans_data[, 5:10]),
+                             知識流通層面平均 = rowMeans(trans_data[, 11:14]),
+                             知識管理層面平均 = rowMeans(trans_data[, 5:14]))
>
> head(score_average)
      知識獲取層面平均  知識流通層面平均  知識管理層面平均
1          2.500000          4.00          3.1
2          3.000000          3.50          3.2
3          3.166667          2.50          2.9
4          3.000000          2.75          2.9
5          2.166667          2.50          2.3
6          2.833333          2.50          2.7
```

「知識獲取」各層面的加總: a1 ~ a6
「知識流通」層面的加總: a7 ~ a10
「知識管理」層面的加總: a1 ~ a10

層面單題平均得分可以看出觀察值對構念特質的知覺感受到何種程度。



層面間之比較

85 / 100

- 層面包含題項數不同，無法從層面的平均數比較。
- 如何進行層面間的比較: 求出樣本觀察值在層面或總量表得分之單題平均的描述性統計量。
 - 「知識獲取層面單題平均」的平均=2.7273
 - 「知識流通層面單題平均」的平均=3.3182
 - 樣本觀察值在「知識流通」層面單題平均得分高於「知識獲取」層面單題平均得分。
- 此差異是否顯著，須加以檢定。

```
> apply(score_average, 2, mean)
```

知識獲取層面平均
2.727273

知識流通層面平均
3.309091

知識管理層面平均
2.960000

```
> apply(score_average, 2, sd)
```

知識獲取層面平均
0.5926189

知識流通層面平均
0.5610086

知識管理層面平均
0.4724248



求測驗成績百分等級

86 / 100

- 百分等級 (PR, percentile rank): 指觀察值在某個測量變項上的測量值(分數) , 在團體中所占的等為多少。PR最高為99。
 - 例如: PR=80 , 表示100人的群體中 , 樣本觀察值的分數可以贏過80個人。
- 百分位數 (Pp, percentile point): 在群體中居某一個百分等級時的分數。
 - 例如: P80=75。百分等級為PR=80 , 數學測成績為75分。

```
> score_data <- read.csv("data/成績_1.csv")
> head(score_data)
  班級 性別 數學 英文 測驗平均
1    1    1    60   66    63.0
...
> dim(score_data)
[1] 50  5
> score_data$數學
[1] 60 42 78 65 68 57 55 97 87 92 75 55 64 71 78 84 85 76 71
...
> rank(score_data$數學) # rank(- score_data$數學)
[1] 9.0 1.0 29.5 12.5 15.0 7.0 3.0 47.0 38.5 43.5 24.0 3.0 11.0 18.5 29.5
...
> rank(score_data$數學, ties.method = "first")
[1] 9 1 28 12 15 6 2 47 37 43 23 3 11 17 29 33 35 26 18 30 32 19 16 22 5 7
...
> library(dplyr)
> round(percent_rank(score_data$數學) * 100)
[1] 16 0 55 22 29 10 2 94 73 86 45 2 20 33 55 65 69 51 33
...
> quantile(score_data$數學, probs = seq(0, 1, 0.1))
  0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
42.0 56.9 63.6 69.4 71.6 75.5 78.0 85.0 87.4 94.1 100.0
```

求出數學成績的百分等級
等級觀察值



等級觀察值、次數分配表

87 / 100

```
> scores <- c(75, 82, 90, 65, 88, 72, 95, 60, 78, 85)
> percentile_75 <- quantile(scores, 0.75)
> percentile_rank_80 <- sum(scores <= 80) / length(scores) * 100
> cat("75th Percentile:", percentile_75, "\n")
75th Percentile: 87.25
> cat("Percentile Rank of 80:", percentile_rank_80, "%\n")
Percentile Rank of 80: 50 %
```

- 求出數學成績五個等第各組人數，並以長條圖及直方圖表示。

■ 等第一: ≥ 90 ; 等第二: 80-89; 等第三: 70-79; 等第四: 60-69; 等第五: ≤ 59

```
> rank_table <- c(">= 90", "80-89", "70-79", "60-69", "<= 60")
> score_to_rank <- function(x){
+   group_id <- ifelse(x >= 90, 1,
+                     ifelse(x >= 80, 2,
+                           ifelse(x >= 70, 3,
+                                 ifelse(x >= 60, 4, 5))))
+   data.frame(score = x,
+             rank = factor(rank_table[group_id], levels = rank_table, ordered = T),
+             row.names = NULL)
+ }
>
> math_data <- score_to_rank(score_data$數學)
> table(math_data$rank)

>= 90 80-89 70-79 60-69 <= 60
    9    10    16     7     8
>
> library(psych)
> describe(math_data$score)
   vars  n  mean    sd median trimmed  mad min max range  skew kurtosis   se
X1     1 50 75.78 13.96   75.5   75.92 16.31  42 100    58 -0.16    -0.8 1.97
```



預檢資料: 檢核是否有極端值存在

88 / 100

Junk, Noisy Data, or Outlier

- As in a physics or statistics test, noise is a random error that occurs during the test process to seize the measured data. No matter what means you apply to the data gathering process, **noise inevitably exists**.
- **Deal with noisy data using smoothing:**
 - **Binning:** This is a local scope smoothing method in which the neighborhood values are used to compute the final value for the certain bin. The sorted data is distributed into a number of bins and each value in that bin will be replaced by a value depending on some certain computation of the neighboring values. The computation can be bin median, bin boundary, which is the boundary data of that bin.
 - **Regression:** The target of regression is to find the best curve or something similar to one in a multidimensional space; as a result, the other values will be used to predict the value of the target attribute or variable. In other aspects, it is a popular means for smoothing.
- **Classification or outlier:** The classifier is another inherent way to find the noise or outlier. During the process of classifying, most of the source data is grouped into couples of groups, except the outliers.

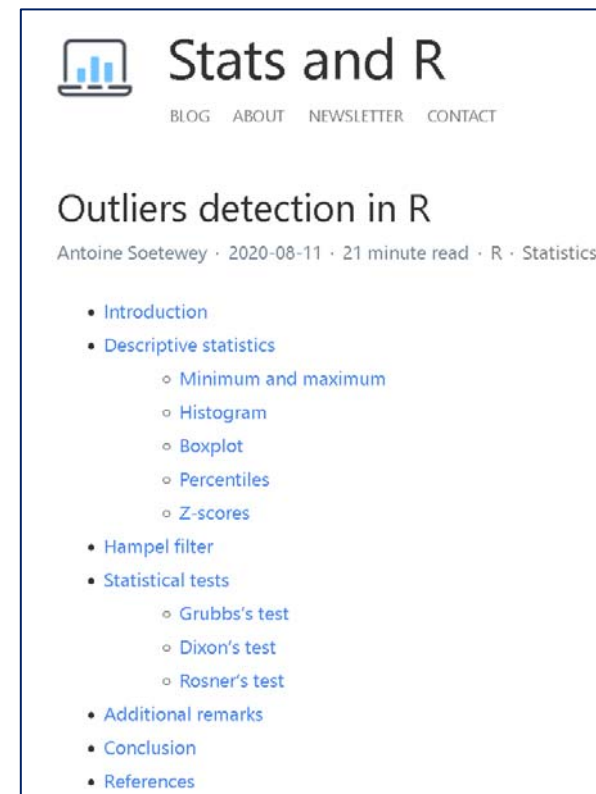
異常檢測的統計方法 (Outliers Detection in R)

89 / 61

- **Graphical techniques:** index plot, Boxplot side-by-side, scatterplot, heatmap and so on.
- **R packages:**
 - **outliers:** Tests for outliers · a collection of some tests commonly used for identifying outliers.
 - **extRemes:** Extreme Value Analysis.
 - **in2extRemes:** Into the extRemes Package, GUI to some of the functions in the package extRemes. (<http://www.assessment.ucar.edu/toolkit/>)
 - **extremevalues:** Univariate Outlier Detection
 - **Extreme Value Analysis(EVA) packages in R:** **evd**, **evdbayes**, **evir**, **fExtremes**, **lmom**, **SpatialExtremes**, **texmex**, **extRemes**, **isnev**, **texmex**, **isnev**
- **Robust approaches to data with outliers:**
Robustify the classical algorithm by replacing the sample mean vector and covariance matrix with the robust location and scatter estimators.

Outliers detection in R:

<https://statsandr.com/blog/outliers-detection-in-r/>



See also: Chapter 7, Outlier Detection, RDataMining-book-2015



Outliers Detection

- **Graphical techniques:** index plot, Boxplot side-by-side, scatterplot, heatmap and so on.
- **R packages:**
 - **outliers:** Tests for outliers
 - A collection of some tests commonly used for identifying outliers.
<https://cran.r-project.org/web/packages/outliers/index.html>
 - Grubbs' test (Grubbs 1969 and Stefansky 1972) is used to detect outliers in a univariate data set. It is based on the assumption of normality. That is, you should first verify that your data can be reasonably approximated by a normal distribution before applying the Grubbs' test.
 - **extRemes:** Extreme Value Analysis.
 - **in2extRemes:** Into the extRemes Package, GUI to some of the functions in the package extRemes. (<http://www.assessment.ucar.edu/toolkit/>)
 - **extremevalues:** Univariate Outlier Detection
 - Extreme Value Analysis(EVA) packages in R: **evd**, **evdbayes**, **evir**, **fExtremes**, **lmom**, **SpatialExtremes**, **texmex**, **extRemes**, **ismev**, **texmex**, **ismev**
- **Robust approaches to data with outliers**
 - Robustify the classical algorithm by replacing the sample mean vector and covariance matrix with the robust location and scatter estimators.

See also: Chapter 7, Outlier Detection, RDataMining-book-2015



R package: **ou1tliers**

Statistical Tests

91 / 61

- **chisq.out.test**: Chi-squared test for outlier
- **cochran.test**: Test for outlying or inlying variance
- **dixon.test**: Dixon tests for outlier
- **grubbs.test**: Grubbs tests for one or two outliers in data sample.

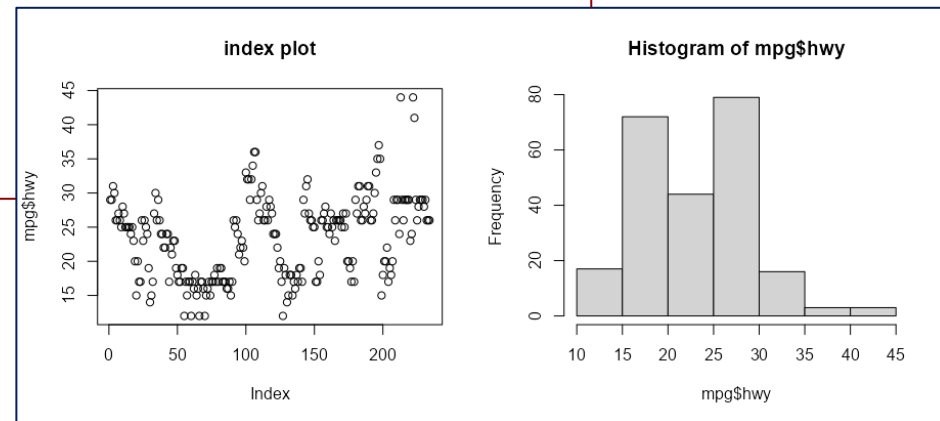
- Dixon, W.J. (1950). Analysis of extreme values. Ann. Math. Stat. 21, 4, 488-506.
- Dixon, W.J. (1951). Ratios involving extreme values. Ann. Math. Stat. 22, 1, 68-78.
- Snedecor, G.W., Cochran, W.G. (1980). Statistical Methods (seventh edition). Iowa State University Press, Ames, Iowa.
- Grubbs, F.E. (1950). Sample Criteria for testing outlying observations. Ann. Math. Stat. 21, 1, 27-58.

```
> library(outliers)
> dim(mpg)
[1] 234 11
> head(mpg, 3)
# A tibble: 3 × 11
  manufacturer model displ year   cyl trans      drv   cty   hwy fl   class
  <chr>         <chr> <dbl> <int> <int> <chr>   <chr> <int> <int> <chr> <chr>
1 audi         a4      1.8  1999     4 auto(l5) f     18    29 p    compact
2 audi         a4      1.8  1999     4 manual(m5) f     21    29 p    compact
3 audi         a4      2    2008     4 manual(m6) f     20    31 p    compact
>
> summary(mpg$hwy)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 12.00  18.00  24.00  23.44  27.00  44.00
> par(mfrow = c(1, 2))
> plot(mpg$hwy, main = "index plot")
> hist(mpg$hwy)
```

mpg {ggplot2}

Fuel economy data from 1999 to 2008 for 38 popular models of cars

hwy: highway miles per gallon



Grubbs' Test for a Single Outlier

- **Assumption:** the data (without any outliers) are approximately normally distributed.
- **Hypothesis:**
 - H_0 : There are no outliers in the data set
 - H_1 : There is exactly one outlier in the data set

- **Test Statistic:** ESD (extreme studentized deviate)

$$ESD = \max_{i=1, \dots, n} \frac{|X_i - \bar{X}|}{s}$$

- **Critical Region:** For the two-sided test, the hypothesis of no outliers is rejected if

$$ESD > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2}{n-2+t^2}} \quad \text{where } t \text{ is short for } t_{n-2,p} \text{ and } p = 1 - \alpha/(2n).$$

Grubbs, Frank (February 1969), Procedures for Detecting Outlying Observations in Samples, Technometrics, 11(1), pp. 1-21.

The Grubbs test detects one outlier at a time (highest or lowest value), so the null and alternative hypotheses are as follows:

if we want to test the highest value

- H_0 : The highest value is not an outlier
- H_1 : The highest value is an outlier

if we want to test the lowest value.

- H_0 : The lowest value is not an outlier
- H_1 : The lowest value is an outlier



R package: outliers

Statistical Tests

93 / 61

```
> test <- grubbs.test(mpg$hwy)
> test
```

Grubbs test for one outlier

```
data: mpg$hwy
G = 3.45274, U = 0.94862, p-value = 0.05555
alternative hypothesis: highest value 44 is an outlier
>
>
> test <- grubbs.test(mpg$hwy, opposite = TRUE)
> test
```

Grubbs test for one outlier

```
data: mpg$hwy
G = 1.92122, U = 0.98409, p-value = 1
alternative hypothesis: lowest value 12 is an outlier
```

```
>
>
> dixon.test(mpg$hwy)
Error in dixon.test(mpg$hwy) : Sample size must be in range 3-30
>
```

> # The p-value is 0.056. At the 5% significance level, we do not reject the hypothesis that the highest value 44 is not an outlier.

> # At the 5% significance level, we do not reject the hypothesis that the lowest value 12 is not an outlier.



Robust Statistical Methods

CRAN Task View: Robust Statistical Methods

<https://cran.r-project.org/web/views/Robust.html>

■ Robust Location and Scatter Estimators

- Median, MAD (median of the absolute deviations from the median)
- M-estimator (Huber, 1964; Maronna, 1976)
- Stahel-Donoho estimator (Stahel, 1981; Donoho, 1982)
- MVE (minimum volume ellipsoid), MCD (minimum covariance determinant) (Rousseeuw, 1983, 1984, 1985)
- S-estimator (Davis, 1987)
- Depth weighted and maximum depth estimators (Zuo, Cui and He, 2004)

MVE (minimum volume ellipsoid)

- Affine equivariant with high breakdown points.
- The existing efficient algorithm for computation.
- Readily available implementations.
- Ability to Identify extreme values.

Outlier values $\sim 2 \times \sqrt{\chi_{0.975,p}^2} + N(0, 1)$
> qchisq(0.975,5)
[1] 12.83250

Formal Tests for Normality

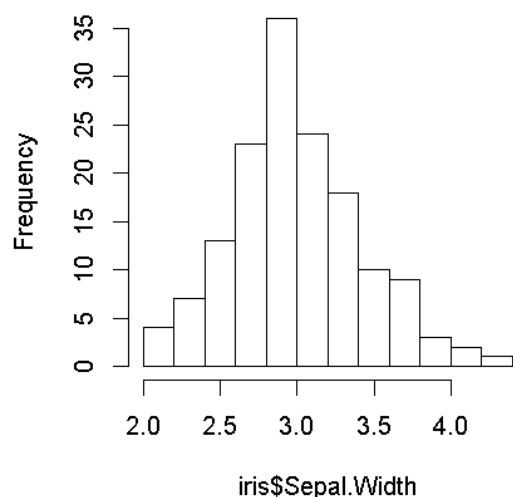
- The hypotheses used are:

H_0 : The sample data are not significantly different than a normal population.

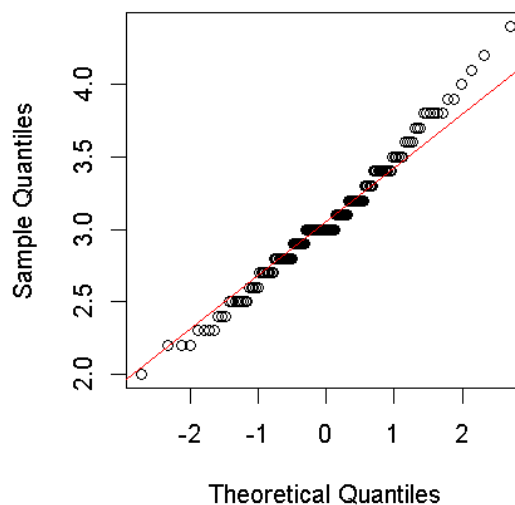
H_a : The sample data are significantly different than a normal population

```
> par(mfrow=c(1, 2))
> hist(iris$Sepal.Width)
> qqnorm(iris$Sepal.Width)
> qqline(iris$Sepal.Width, col="red")
```

Histogram of iris\$Sepal.Width



Normal Q-Q Plot



Packages: **nortest**

Five omnibus tests for testing the composite hypothesis of normality:

ad.test, **cvm.test**,
lillie.test,
pearson.test, **sf.test**



ks.test, ad.test, shapiro.test

96 / 100

- Kolmogorov-Smirnov (K-S) test (Chakravarti et al., 1967).
- The Anderson-Darling test (Stephens, 1974).
- The Shapiro-Wilk normality test (Shapiro and Wilk, 1965).
- A large p -value (larger than, say, 0.05) indicates that the sample is not different from normal with the sample's mean and standard deviation.

```
> x <- iris$Sepal.Width  
> ks.test(x, 'pnorm', mean(x), sd(x))
```

One-sample Kolmogorov-Smirnov test

```
data: x  
D = 0.10566, p-value = 0.07023  
alternative hypothesis: two-sided
```

Warning message:

```
In ks.test(x, "pnorm", mean(x), sd(x)) :  
ties should not be present for the Kolmogorov-Smirnov test
```

```
> library(nortest)  
> ad.test(iris$Sepal.Width)
```

Anderson-Darling normality test

```
data: iris$Sepal.Width  
A = 0.90796, p-value = 0.02023
```

```
> shapiro.test(iris$Sepal.Width)
```

Shapiro-Wilk normality test

```
data: iris$Sepal.Width  
W = 0.98492, p-value = 0.1012
```



Which Normality Test Should I Use?

97 / 100

- **Kolmogorov-Smirnov test:**
 - The test applies to continuous densities only.
 - It is more sensitive near the center of the density than at the tails than other tests;
 - For data sets $n > 50$.
- **The Anderson-Darling test:**
 - A-D test is a modification of the K-S test and gives more weight to the tails of the density than does the K-S test. It is generally preferable to the K-S test.
- **Shapiro-Wilks test:**
 - Doesn't work well if several values in the data set are the same.
 - Works best for data sets with $n < 50$, but can be used with larger data sets.
- **W/S test ($\text{range}(x)/\text{sd}(x)$):** simple, but effective.
- **Jarque-Bera test (`jarque.test {moments}`):** tests for skewness and kurtosis, very effective.
- **D'Agostino test (`agostino.test {moments}`):** powerful omnibus (skewness, kurtosis, centrality) test.



Which Normality Test Should I Use?

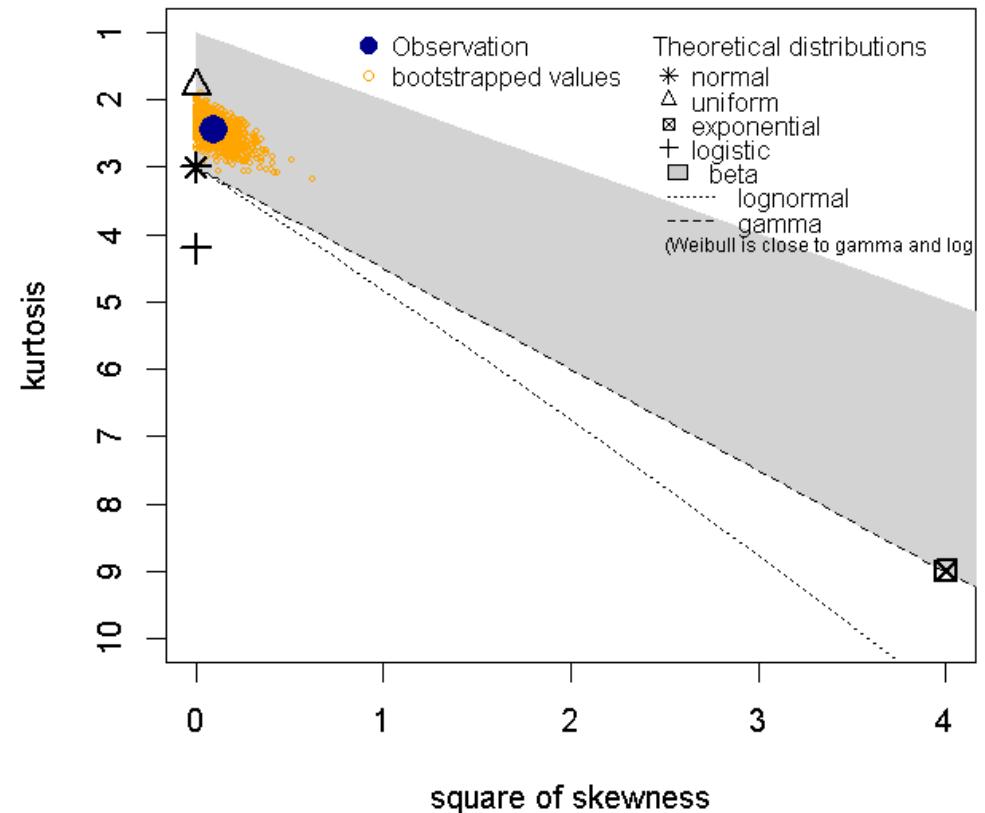
- Asghar Ghasemi and Saleh Zahediasl, [Normality Tests for Statistical Analysis: A Guide for Non-Statisticians](#), Int J Endocrinol Metab. 2012 Spring; 10(2): 486–489.
 - assessing the normality assumption should be taken into account for using [parametric statistical tests](#).
 - The K-S test, should no longer be used owing to its low power.
 - It is preferable that normality be assessed both visually and through normality tests, of which the **Shapiro-Wilk test** is highly recommended.
- **NOTE:**
 - If the data are not normal, use non-parametric tests.
 - If the data are normal, use parametric tests.
 - If you have groups of data, you MUST **test each group** for normality.
 - It's common seen that a model is built from the **training data** and is then applied to the **testing data**. Did these two data sets follow the same distribution?



An R Package for Fitting Distributions

```
> install.packages("fitdistrplus")
> library(fitdistrplus)
> x <- iris$Sepal.Length
> descdist(x, boot=1000)
summary statistics
-----
min: 4.3    max: 7.9
median: 5.8
mean: 5.843333
estimated sd: 0.8280661
estimated skewness: 0.314911
estimated kurtosis: 2.447936
> fit.n <- fitdist(x, "norm")
> summary(fit.n)
Fitting of the distribution ' norm '
by maximum likelihood
Parameters :
      estimate Std. Error
mean 5.8433333 0.06738557
sd   0.8253013 0.04764848
Loglikelihood: -184.0398
AIC: 372.0795   BIC: 378.1008
Correlation matrix:
      mean sd
mean   1  0
sd     0  1
```

Cullen and Frey graph



rapidFitFun {qAnalyst}: Function to obtain rapid fitting of multiple distributions

<https://cran.r-project.org/web/packages/qAnalyst/index.html>

Package 'qAnalyst' **was removed** from the CRAN repository.

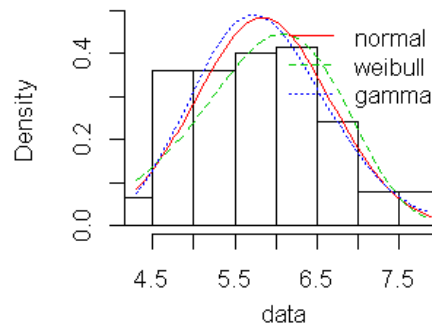
Formerly available versions can be obtained from the archive.



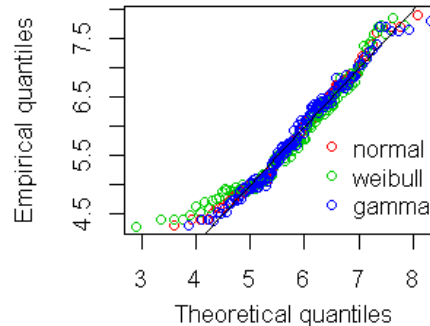
An R Package for Fitting Distributions

```
> fit.w <- fitdist(x, "weibull")
> fit.g <- fitdist(x, "gamma")
> par(mfrow=c(1,4))
> plot.legend <- c("normal", "weibull", "gamma")
> denscomp(list(fit.n, fit.w, fit.g), legendtext=plot.legend)
> qqcomp(list(fit.n, fit.w, fit.g), legendtext=plot.legend)
> cdfcomp(list(fit.n, fit.w, fit.g), legendtext=plot.legend)
> ppcomp(list(fit.n, fit.w, fit.g), legendtext=plot.legend)
```

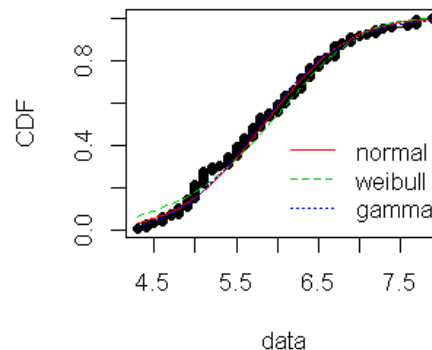
Histogram and theoretical densitie



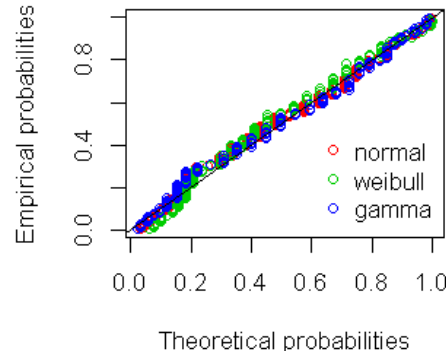
Q-Q plot



Empirical and theoretical CDFs



P-P plot



```
> summary(fit.w)
Fitting of the distribution
'weibull' by maximum likelihood
Parameters :
```

	estimate	Std. Error
shape	7.454379	0.45168136
scale	6.208005	0.07209406
Loglikelihood:	-190.7689	
AIC:	385.5377	BIC: 391.559
Correlation matrix:		
	shape	scale
shape	1.0000000	0.3323758
scale	0.3323758	1.0000000

```
> summary(fit.g)
Fitting of the distribution
'gamma' by maximum likelihood
Parameters :
```

	estimate	Std. Error
shape	50.634073	5.827566
rate	8.665336	1.002253
Loglikelihood:	-182.3061	
AIC:	368.6122	BIC: 374.6335
Correlation matrix:		
	shape	rate
shape	1.0000000	0.9950669
rate	0.9950669	1.0000000