



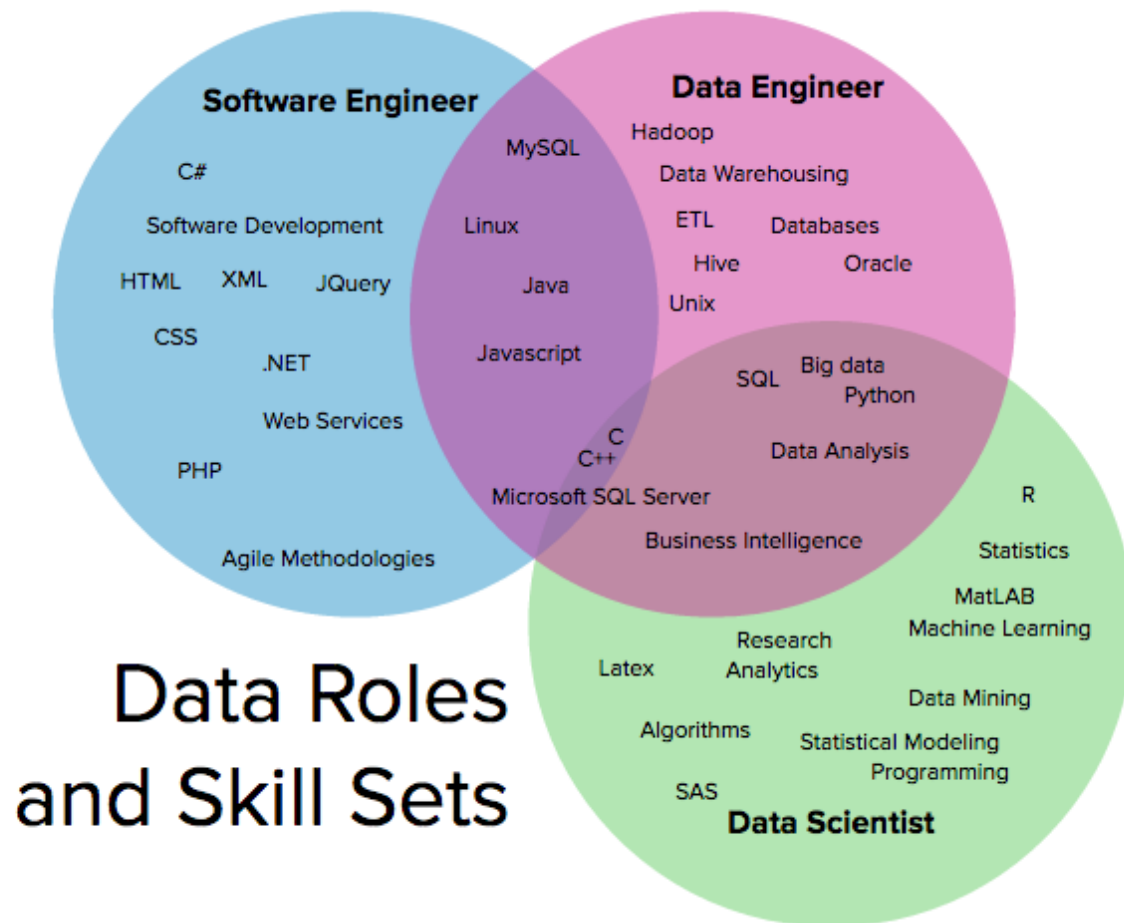
國立臺灣大學
National Taiwan University

資料科學

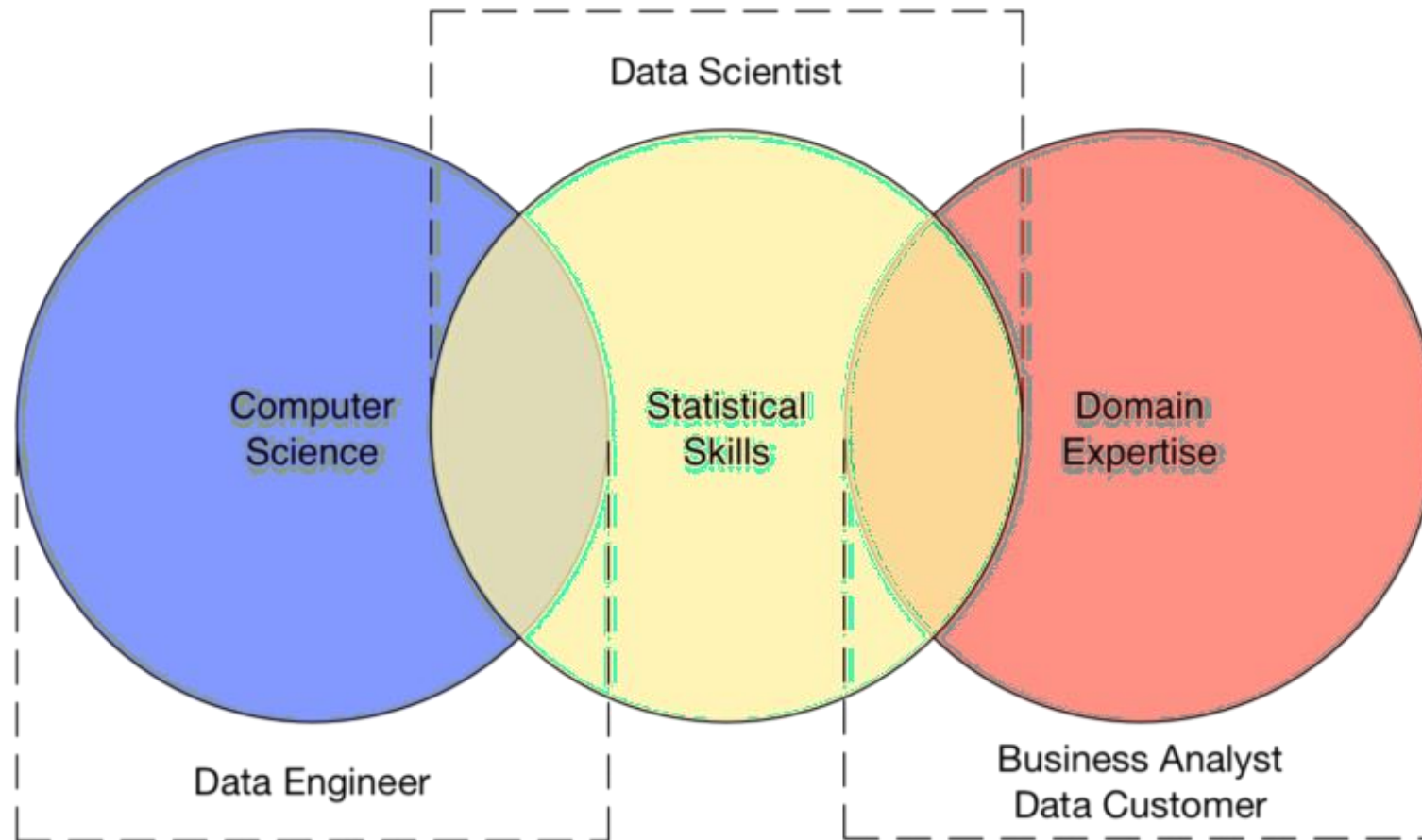
國立臺灣大學共同教育中心

蔡芸琤

資料科學



資料科學



分析方法

1. 敘述統計
2. 相關係數分析
3. 迴歸分析
4. 變異數分析
5. 時間序列分析
6. 機器學習
7. 深度學習

敘述統計

隨機抽樣的樣本集合，進行某些計算與繪圖，忠實呈現出樣本的某些特性。這些計算出的數值，以及呈現出來的圖形，反映出樣本的某些統計特性，讓統計者能透過數值或圖形，大致了解樣本的統計特徵。

基本函數	意義	基本函數	意義
mean()	平均數	summary()	描述統計摘要
median()	中位數	sd()	標準差
range()	全距	var()	變異數
quantile()	四分位數	skewness()	偏度
IQR()	四分位差	kurtosis()	峰度

相關係數分析

變項間的相關程度高或低，得到的相關係數只能說明這兩個變項間是正相關、負相關，或者是無關。

相關程度之高低，在正負0.3之間（即0.3至-0.3之間）稱為低度相關；在正負0.3-0.6之間（即指介於0.3至0.6，-0.3至-0.6之間）稱為中度相關；而在正負0.6至0.9之間（即指在0.6至0.9，-0.6至-0.9之間）則稱為高度相關；若是為正負1，即表示完全相關；若是為0，即表示無關。

基本函數	意義
<code>cor()</code>	計算相關係數
<code>cor.test()</code>	相關係數分析

導引|問題

http://www.r-web.com.tw/guider/1/section_A.php



雲端資料分析暨導引系統 x

← → ↻ www.r-web.com.tw/guider/1/section_A.php ☆

 **雲端資料分析暨導引系統**
Data Analysis & Guiding System - Cloud

首頁 | 網站導覽 | 會員登入 | 初階使用者 ▾

讚 分享

G+1

初階導引 資料處理 分析方法 圖表繪製 機率分配 輸出結果 自創巨集

首頁 » 初階導引 » 導引問題區

單變數分析導引問題區

你選擇的區域為單變數分析，請依序回答下列問題，我們將由你所選的答案中，提供您最適合於分析你的資料的統計方法。若此處方法非你所需，或你覺得不適合可點選最下方回導引分析首頁重新選擇。

Q1.你欲分析的資料中包含多少樣本組數? ● A1.一組 ● A2.二組 ● A3.三組(含)以上

[參閱說明及範例](#)

進行下一個問題

以處理政府公開的資料為例

政府資料開放平台：<http://data.gov.tw>

行政院環境保護署：<http://data.gov.tw/node/6350>

空氣品質監測日污染指標欄位名稱

1. 測站編號(SiteId)
2. 測站名稱(SiteName)
3. 監測日期(MonitorDate)
4. 空氣污染指標PSI(PSI)
5. 二氧化硫PSI副指標(SO2SubIndex)
6. 一氧化碳PSI副指標(COSubIndex)
7. 臭氧PSI副指標(O3SubIndex)
8. 懸浮微粒PSI副指標(PM10SubIndex)
9. 二氧化氮PSI副指標(NO2SubIndex)

透過 R 整理資料

<https://dsp.im/2015/02/r-packages/>

https://rpubs.com/bradleyboehmke/data_wrangling



```
1 rm(list = ls(all=TRUE))
2 library(dplyr)
3 rawdata = read.csv(
4   'AQXDaily_20170409211519.csv',
5   encoding = 'UTF-8')
6
7 df <- rawdata %>% group_by(X.U.FEFF.SiteId)
8 result = summarise(df, mean(SO2SubIndex), mean(COSubIndex))
9
```

8:10 (Top Level)

R Script

Console C:/Users/pecu6/Desktop/

```
encoding = 'UTF-8')
>
```

```
> df <- rawdata %>% group_by(X.U.FEFF.SiteId)
```

```
> result = summarise(df, mean(SO2SubIndex), mean(COSubIndex))
```

```
> view(result)
```

```
>
```

Data

df	1000 obs...
ra...	1000 obs...
re...	76 obs. ...

	Name	Size
	~\MICROSOFT WORD 文件.docx	162 B
	SHPaper	
	SCI Paper.Ink	1 KB
	python教科書	
	oCam.Ink	986 B
	NTU-CSX	
	Google 雲端硬碟.Ink	1.7 KB
	FXpaper	
	FinTech	
	Azure	
	AQXDaily_20170409211519.csv	34.9 KB
	Amazon Chime.Ink	934 B
	airPollution.R	227 B

台大1996年至2015年專任師資人數

1. 資料準備
2. 探索資料
3. 項目分析
4. 輸出圖表
5. 觀察與說明



資料準備

Data -> CSV File

高達 80% 的過程花在準備資料與整理資料階段，包括表格整合以及大量的資料格式轉換。資料清理與轉換是資料準備的重點。

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

payment.R

```
1 rm(list = ls(all=TRUE))
2 library(plotly)
3 teachers = read.csv('teachers.csv', header = TRUE)
4 positions = c("Prof.", "Associate Prof.", "Assistant Prof.",
5              "Insturctor", "TA")
6 years = length(teachers[,1])
7 allStaffs = data.frame(teachers[,1],
8                         rowSums(teachers[,c(2,7)]),
9                         rep(positions[1], years))
10 names(allStaffs) = c("years", "people", "position")
11 for(tid in seq(3,6))
12 {
13   if(tid < 5)
14   {
15     temp = data.frame(teachers[,1],
16                       rowSums(teachers[,c(tid,tid+5)]),
17                       rep(positions[tid-1], years))
18   }
19   else
20   {
21     temp = data.frame(teachers[,1],
22                       teachers[,tid],
23                       rep(positions[tid-1], years))
24   }
25   names(temp) = c("years", "people", "position")
26   allStaffs = rbind(allStaffs, temp)
27 }
28 p<-plot_ly(allStaffs, x=~years, y=~people, color=~position) %>%
29   add_lines() %>%
30   layout(title = "國立臺灣大學1996年至2015年專任師資人數",
31          xaxis = list(title = "年度"),
32          yaxis = list(title = "人數"))
33 p
34 htmlwidgets::saveWidget(p, "index.html")
35
```

Environment History

Global Environment

Environment is empty

35:1 (Top Level) 2 R Script 2

Console Files Plots Packages Help Viewer

HW3 請自行挑選一張表格進行整理

例如原始資料：

<http://acct2013.cc.ntu.edu.tw/acct2013/acct6/6.doc>

透過 R 整理資料

完成基本敘述統計分析

Microsoft Word [相容模式] - Word

檔案 常用 插入 設計 版面配置 參考資料 郵件 校閱 檢視 告訴我想要執行的動作... YC Tsai 共用

表6：歷年專任教師人數,1950-2012

年度	總計 (不含助教)	教授	副教授	助理教授	講師	助教	與校外合聘 不在校支薪教師			
							小計	教授	副教授	助理教授
1950	313	178	71	---	64	190	---	---	---	---
1951	336	186	74	---	76	186	---	---	---	---
1952	353	187	77	---	89	178	---	---	---	---
1953	389	204	81	---	104	168	---	---	---	---
1954	438	224	93	---	121	196	---	---	---	---
1955	459	229	98	---	132	211	---	---	---	---
1956	521	254	102	---	165	188	---	---	---	---
1957	518	255	98	---	165	206	---	---	---	---
1958	556	279	108	---	169	193	---	---	---	---
1959	576	281	122	---	173	167	---	---	---	---
1960	582	283	127	---	172	199	---	---	---	---
1961	629	295	163	---	171	226	---	---	---	---
1962	650	304	170	---	176	233	---	---	---	---
1963	696	322	179	---	195	224	---	---	---	---
1964	723	335	179	---	209	228	7	4	3	---
1965	731	338	189	---	204	259	7	5	2	---
1966	768	354	195	---	219	235	7	5	2	---
1967	814	368	220	---	226	256	6	5	1	---
1968	821	369	220	---	232	268	9	5	4	---
1969	912	390	284	---	238	295	10	5	5	---

第 1 頁，共 2 頁 889 個字 英文 (美國) 100%

期末報告規範

1. 資料來源
2. 透過整合的資料，問一個：「原本單一的資料無法回答的問題」
3. 利用資料與 R 的繪圖功能，回答上述的問題
4. 在跑完程式之後，執行 `sessionInfo()` 並且把結果貼到 Rmarkdown 輸出。
5. 在上述的過程中，在處理資料與繪圖的過程中遇到什麼困難？如何活用 R 解決問題？額外自學了什麼？