

統計學習與大數據分析實務簡介 (以R語言為例)

報告人：Eddie
日期：2025-04-08

CONTENTS

目錄

- 統計學的基本概念
- 範例：智慧電錶及IoT資料
- 資料的整理與呈現

統計學的基本概念

- ◆社會科學的研究常常需要用到量化分析
- ◆研究者(或你)對那些資料感興趣?這些資料有什麼特徵?
(國家、產業、企業、特定群體....)
- ◆該如何進行資料蒐集?
- ◆該如何呈現既有的資料?
- ◆統計學可以幫助我們進行分析?

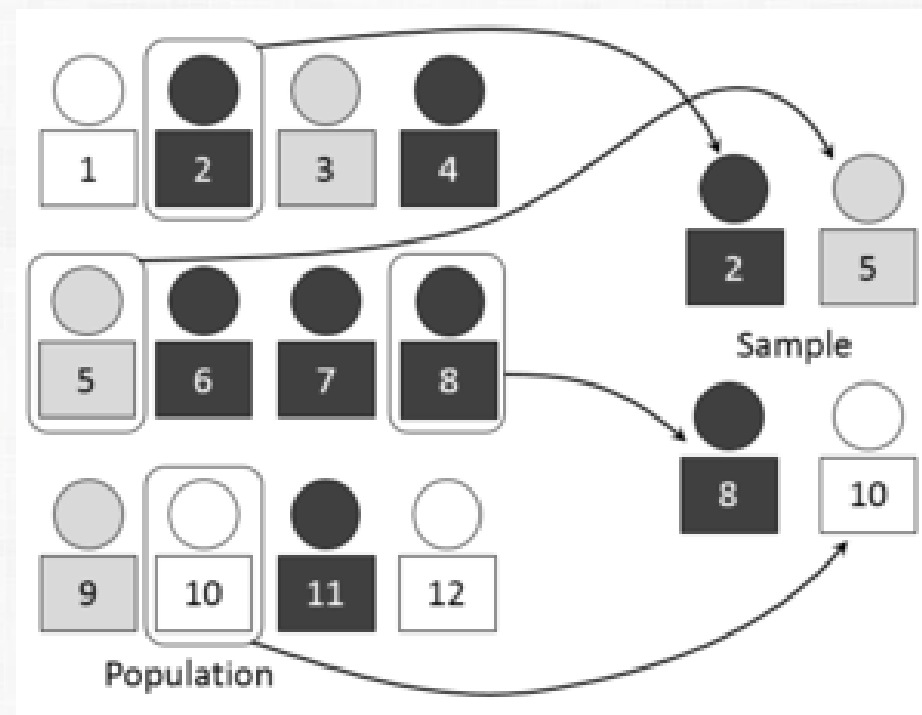
- ◆統計學是一種方法，一種工具。
- ◆狹義的統計學是指以數字表示的事實或資料 data。
- ◆廣義的統計學是指蒐集、整理、表現、分析及解釋資料，並藉科學的方法，在不確定的情況下，由樣本資料所獲得的結果，來推論母體的性質與事實，從而做出適切決策的一門學科。

◆母體 (Population)

- 母體是由具有某些共同特質 (Characteristic) 的元素 (Element) 或個體所組成的群體，是研究人員所要研究觀察的對象的全體集合。

◆樣本 (Sample)

- 樣本是由母體中抽取部份元素而組成的集合，是母體的一部份。



◆如果對母體特徵有興趣，為何要抽樣？

□ 經費有限，無法全部調查。

- COVID 19 下，台灣人消費型態的改變
- 貼政策對國人疫苗接種意願的影響

□ 毀壞性試驗。

- 食物的保鮮期限。

□ 不易取得母體。

- 不同國家居民使用每日手機時間的比較

◆母體參數

- 母體參數是描述母體資料特性的統計測量數，一般簡稱為參數或母數。參數是我們想要獲取的，是統計的核心。
- 母體平均數、變異數、標準差、母體比例。

◆樣本統計量

- 樣本統計量是描述樣本資料特性的統計測量數，一般簡稱為統計量，通常用來推論母體參數。
- 樣本平均數、變異數、標準差、樣本比例。

敘述統計學

- ◆ 敘述統計學包括蒐集、整理、表現、分析與解釋資料。意即它係討論如何蒐集資料，以及將所獲得的資料，加以整理表現解釋與分析。

推論統計學

- ◆ 推論統計學是將敘述統計中由樣本資料所獲得的結果，將之一般化推論至母體，或是由樣本統計量推論到母體參數的方法。它又稱為歸納統計學 (inductive statistics)。

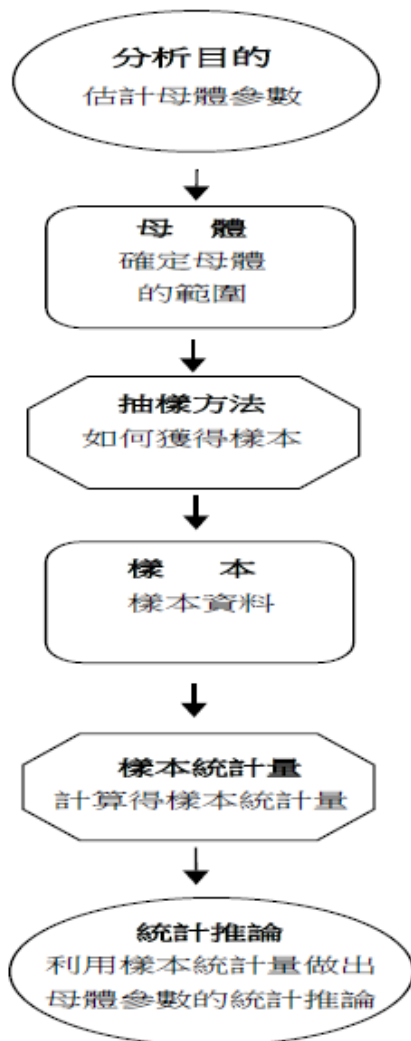
統計方法的實施步驟(1/2)




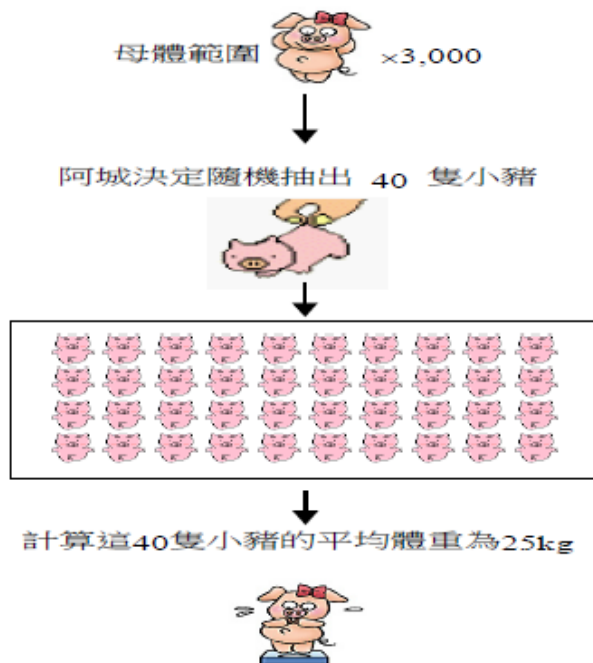
1. 確定問題：首先必須確定問題之所在 及研究分析的目的、對象與範圍。
2. 蒐集資料：針對研究對象、目的進行資料蒐集。在蒐集時應考慮是否有現成合用的資料 蒐集的成本費用如何 蒐集的方式 資料涵蓋的範圍等。
3. 審查整理呈現資料：蒐集到的資料應先審核是否完整、正確、合理與一致然後利用敘述統計學所介紹的方法進行分類整理並以文字圖形表格的方式將所獲得的結果呈現出來。
4. 分析解釋資料：根據整理的結果加以分析研究 探討各數值間的相互關係並加以比較。
5. 統計推論：根據步驟所得的結果 來推論母體參數並下結論或做建議。

統計方法的實施步驟(2/2)

圖1.5 統計方法的實施步驟



阿城是一家養豬場的主人，上一個月買進了三千頭仔豬，阿城想瞭解這一批小豬的成長情形，看看是否需要改變飼養方法與環境。



阿城認為這批小豬的成長情況合乎預期，因此不必更改飼養方法與環境。

◆依取得的方式

- **原始資料**：直接由研究人員或資料使用者依研究的目的去調查、觀察或實驗而獲得的資料。
- **次級資料**：他人所搜集、整理分析的統計資料稱為次級資料或二手級資料。

資料的種類(1/5) 次級資料的來源與資料內容

13

資料來源。	資料內容或項目。
政府機關。	各級政府機構、包括部、會、局、處、縣市政府等施政資料、調查資料等。
學術機構。	各大學、研究機構出版的論文或蒐集的資料。
企業單位。	生產、銷售、庫存、產品名稱、顧客資料等。
專業調查機關。	選舉調查、滿意度調查、收視率調查、產業調查。
產業協會。	產業會員相關生產、銷售等資料。
電視與網際網路。	報社、電視台、新聞網站、個人網站、企業網站的各種統計資料。

資料的種類(1/5) 使用次級資料(二手資料)須注意事項

- ◆資料(提供者)來源?是否客觀?
- ◆調查的對象?
- ◆抽樣方法為何?樣本大小?抽樣誤差有多大?
- ◆有效樣本數?回答率?
- ◆調查方法是甚麼?網路?郵寄?電訪?
- ◆何時做的調查?事件發生前或事件發生後?距離多久?
- ◆問卷的問題為何?

◆依資料的屬性

- **類別資料**：凡是不以數值來表示，僅以類別區分的資料，稱為類別資料，又稱為質的資料。例如畢業學校、使用的手機廠牌。
- **數量資料**：凡是可計數的資料稱為數量資料。例如 死亡率、出生率、每日上網分鐘數、TOEFL的總分...等

資料的種類(3/5)--類別資料的衡量方式

◆ **類別資料**：以類別區分的資料，它是以**文字**來描述或分類的資料。

- 就業情形 就業，未就業
- 婚姻狀況 單身，已婚，離婚
- 衡量時給予各類別一個數字 0,1,2,...。
- 各數字表示類別，沒有大小順序之關係。(0: 就業， 1: 未就業)

◆ **順序資料**：是指**資料的順序是有意義**的資料。

- 滿意程度 非常滿意，滿意，普通，不滿意，非常不滿意
- 衡量順序資料的要求是必須 保持數值高低。
- 5: 非常滿意， 4: 滿意， 3: 普通， 2: 不滿意， 1: 非常不滿意

◆依資料的對象範圍

□ **普查資料**：針對 整個母體 的每一元素進行全面性調查而得到的資料稱為普查資料。

例如：中華民國 110 年 國人主要死亡原因 。

□ **抽樣資料**：由母體中所抽取的樣本 而獲得的資料稱為抽樣資料。

例如：2014 年「九合一」選舉的民意調查

◆依資料的發生時間

□ 橫斷面資料(cross section)

- 發生於同一時點或同一期間的資料稱為橫斷面資料。
- 例如：2020 年台灣主要癌症死亡人數

□ 時間數列資料(time series)

- 發生於 不同時點或不同期間 的資料稱為時間數列資料。
- 包含日資料、月資料、季資料、年資料
- 例如：1993~2021 年各年臺灣的人口出生率

資料的種類(5/5)

◆ 橫斷面資料(cross section)



◆ 時間數列資料(time series)



出生人口溜滑梯 圖／聯合報提供

觀察

- ◆ 研究人員或其工作人員在做研究時利用觀看、查察記錄，而不與研究對象有任何接觸晤談的資料搜集方法稱為觀察。
- ◆ 有些時候只能用觀察法蒐集資料。

調查

- ◆ 調查：對影響母體特性的各種因素不做控制的情況下，進行資料搜集的方法稱為調查。
 - 普查：針對母體中每個元素進行資料之蒐集的方法稱為普查。
 - 抽樣：從母體中抽取一部份的元素進行資料蒐集的方法稱為抽樣。

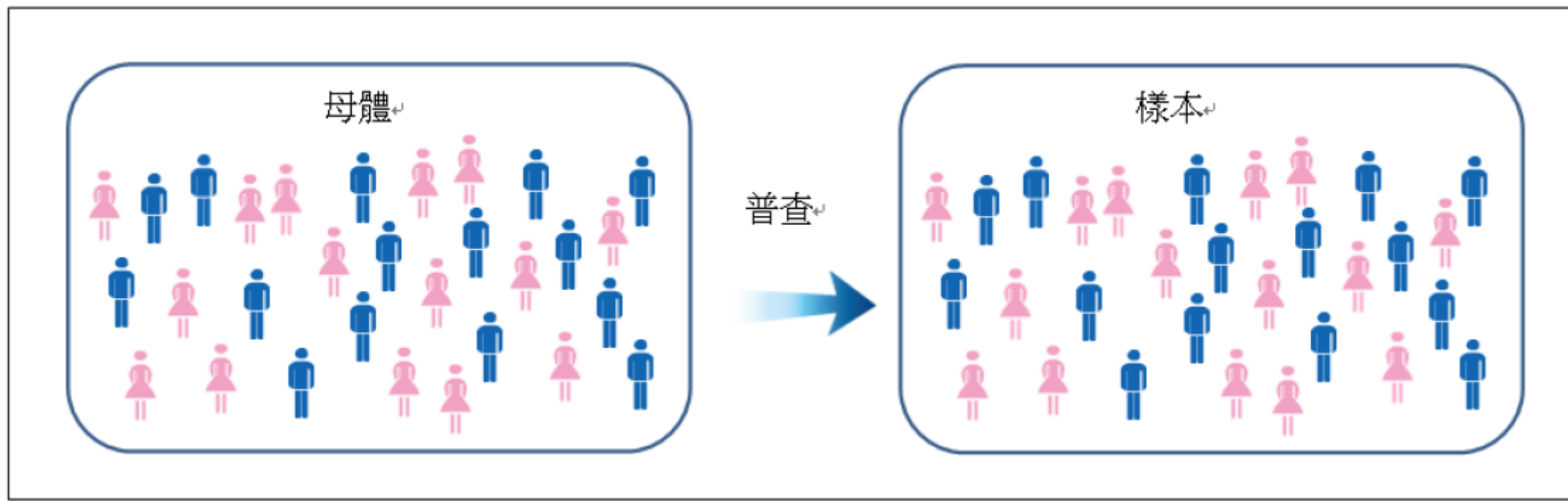
實驗

- ◆ 對影響母體特性的某些因素或其他因素加以控制的資料蒐集方法稱為實驗。

原始資料的搜集方法(2/4)—普查

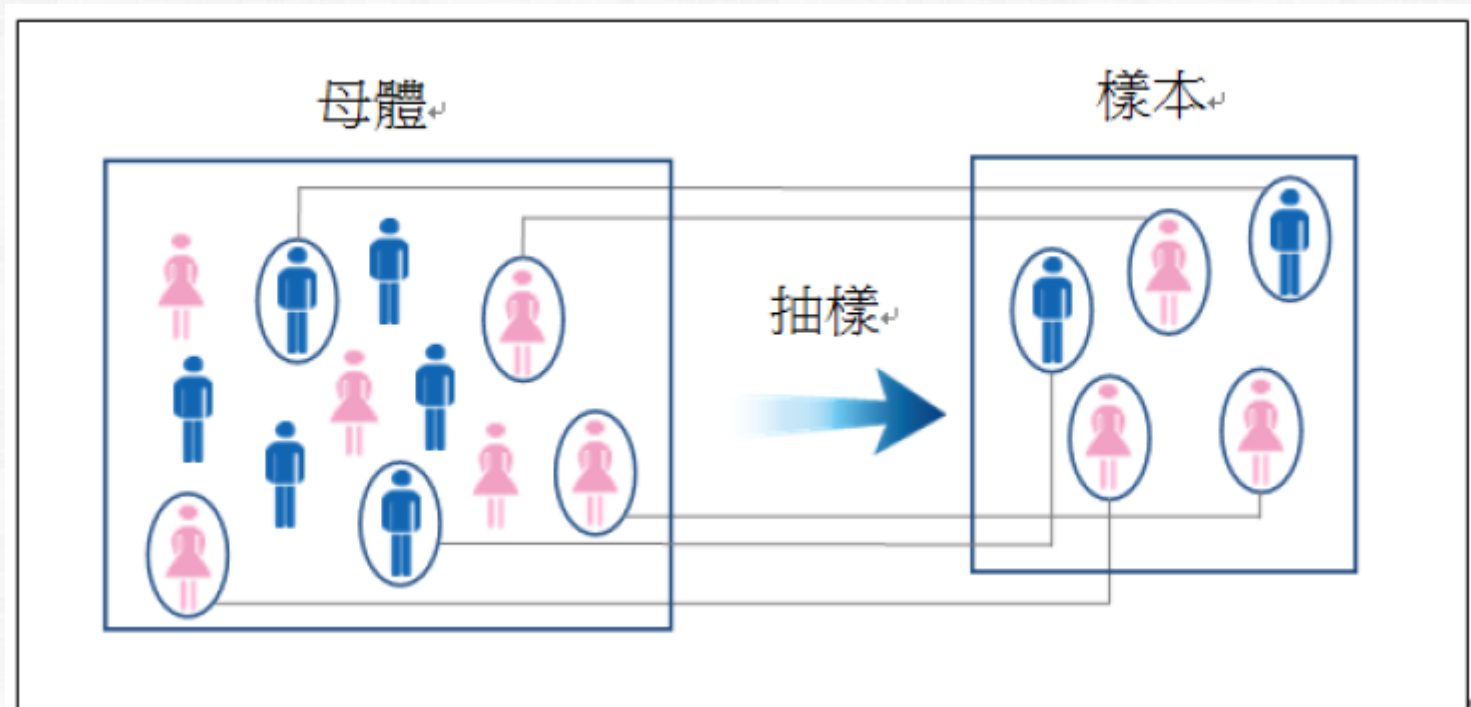
21

◆ 普查法是樣本的元素與母體的元素都一樣



原始資料的搜集方法(3/4)—抽樣

◆ 樣本的元素僅是母體元素的一部分



原始資料的搜集方法(3/4)—抽樣(續)

◆例：家庭小朋友數量的影響因素為何？(1200份已婚婦女問卷中的10份)

表 2.4 已婚婦女的相關資料

	小孩數	學歷	是否就業	薪資所得	婚姻狀況
陳玲惠	1	研究所	是	52,000	滿意
張麗真	2	大學	是	46,000	很滿意
李玉芳	1	專科	是	37,000	滿意
郭月雲	0	高職	否	0	不滿意
魏怡君	1	國中	否	0	滿意
陳昭惠	1	高中	是	26,000	不滿意
劉秀巒	0	大學	是	41,000	無意見
郭如玉	1	大學	否	0	滿意
朱淑美	0	研究所	是	47,000	很不滿意
呂淑真	1	高中	是	28,000	滿意

◆資料的衡量 將變數轉換成具有明確意義的數值

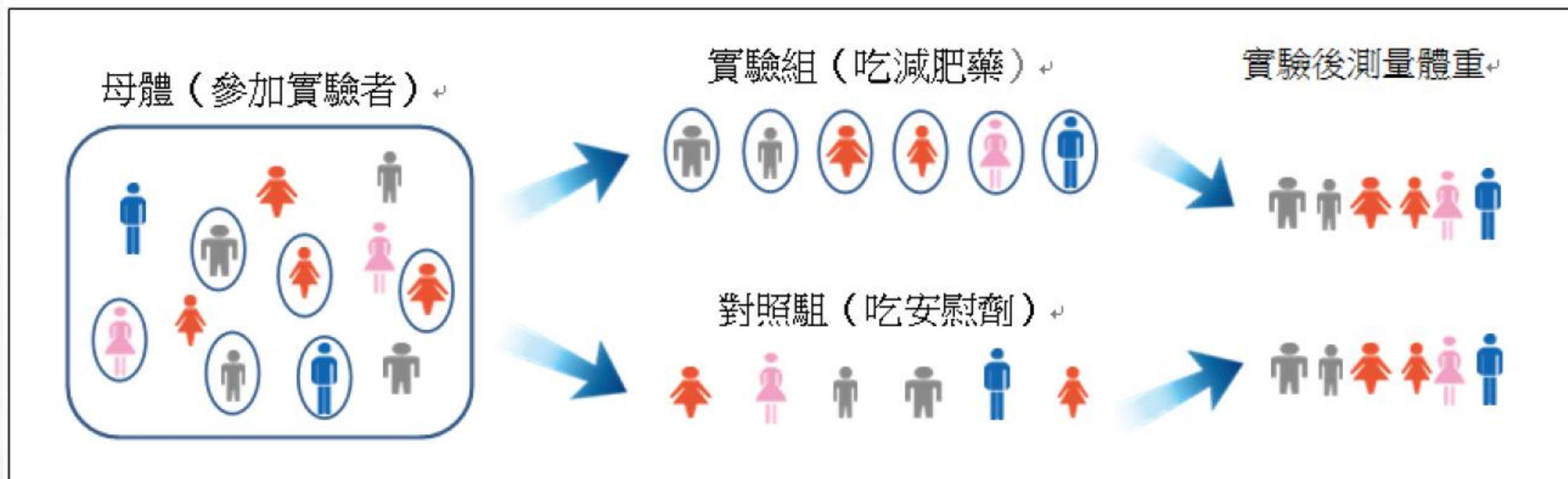
表 2.5 已婚婦女抽樣資料的衡量（數值化）

	小孩數	是否就業	學歷	結婚年齡	薪資所得	婚姻滿意
陳玲惠	1	0	5	31	52,000	4
張麗真	2	0	4	30	46,000	5
李玉芳	1	0	3	31	37,000	4
郭月雲	0	1	2	29	0	2
魏怡君	1	1	1	24	0	4
陳昭惠	1	0	2	26	26,000	2
劉秀巒	0	0	4	30	41,000	3
郭如玉	1	1	4	31	0	4
朱淑美	0	0	5	32	47,000	1
呂淑真	1	0	2	29	28,000	4

原始資料的搜集方法(4/4)—實驗

25

- ◆例：藥廠利用實驗設計，將實驗對象分成對照組與實驗組，以瞭解減肥藥對降低體重的效果。



◆因課程安排，相關內容請同學自行學習。

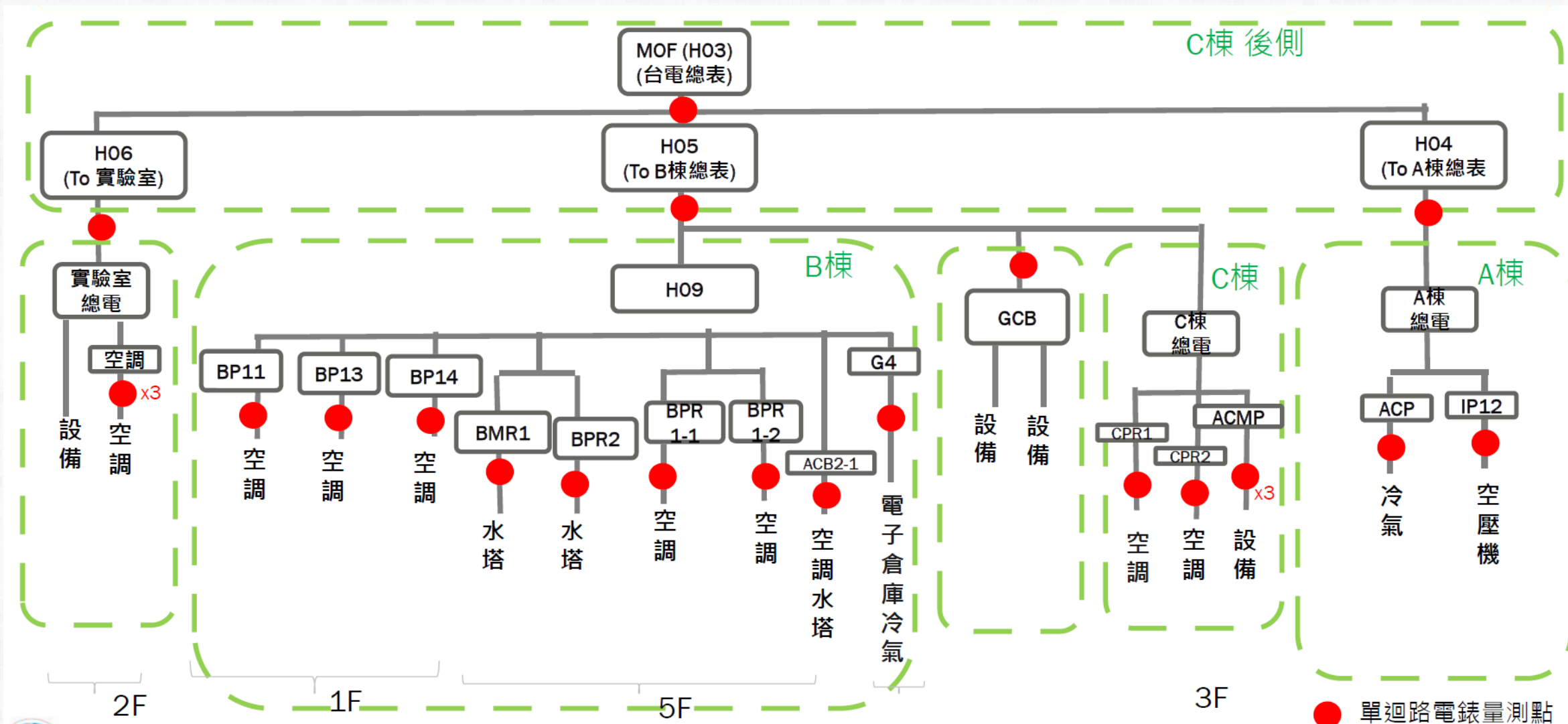
或可參考

[抽樣 - 維基百科，自由的百科全書](<https://zh.wikipedia.org/zh-tw/%E6%8A%BD%E6%A8%A3>)

範例：智慧電錶及IoT資料

公司與廠區的用電迴路關係圖

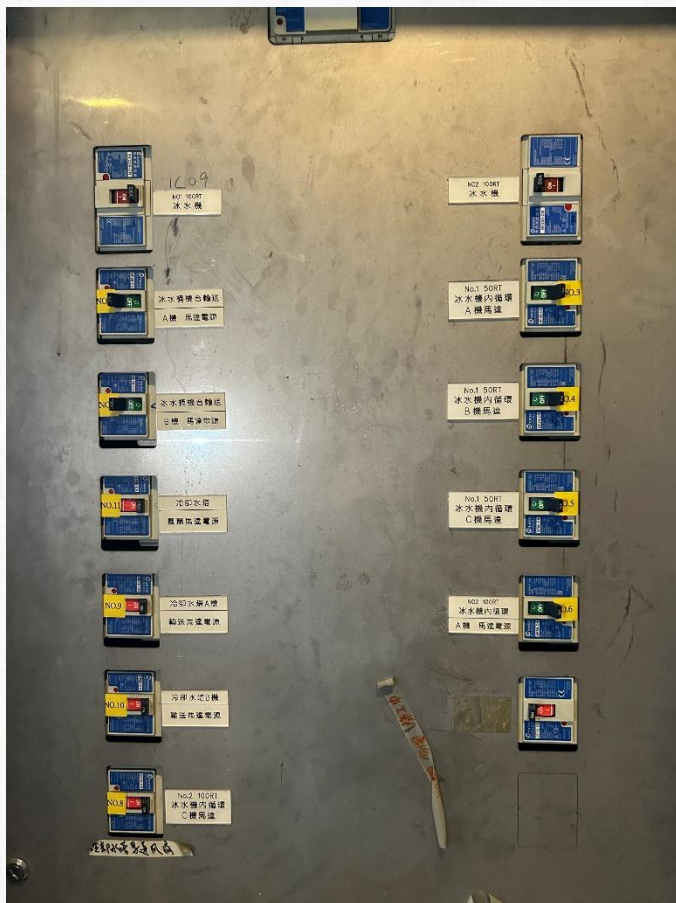
28



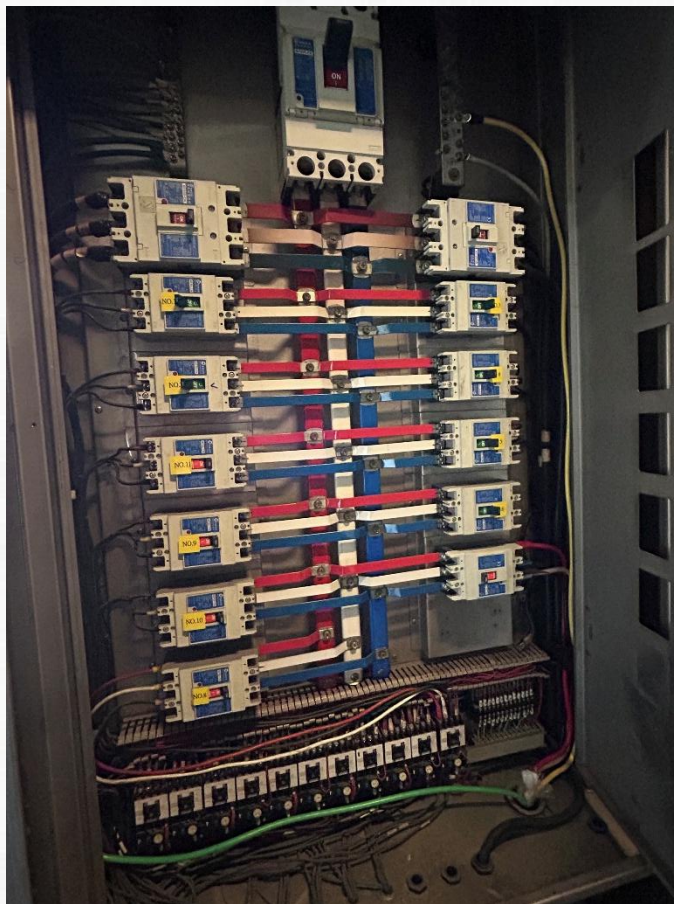
工廠或建築物的總電錶、電盤與迴路

29

◆ 總電錶(電盤)-正面



◆ 總電錶(電盤)-打開



◆ 某一個月的電費單

台灣電力公司
Taiwan Power Company

112年11月 繳費通知單(高壓電力用戶)
Nov. 2023 Electricity Bill (High Voltage)

台中市潭子區
股份有限公司

先生/女士/寶號 g07JT01 102766 通知單號碼: g011 02766

電號 Customer Number	繳費期限 Due Date	應繳總金額 Total Amount	繳費資訊 ent Info.
07-64- -2	112/11/20	*** 868,947 元	

本單僅作通知用，付款時當另給繳費憑證，其他事項請參閱背面說明。

用戶資訊 Basic Info.	計費內容 Charge Info.
電價種類：高壓電力綜合營業用	基本電費(約定) 117150.0 元
時間種類：二段式時間電價	流動電費 765030.0 元
用電地址：台中市潭子區!	功率因數調整費 -13232.7 元
用戶營業事業統一編號：22 - L2	稅前應繳總金額 827569.0 元
行業別：塑膠製品製造業(220)	營業稅 41378.0 元
代繳帳號：00192100*****	應繳總金額 868,947 元
契約容量(瓩) 經常(尖峰)契約 600	
最高需量(瓩) 經常(尖峰)需量 535	
週六半尖峰需量 305	
離峰需量 522	
計費度數(度) / Energy Consumption(kWh)	
尖峰度數 122760	
週六半尖峰度數 7260	
離峰度數 78480	
功率因數(%) 96	

其他資訊 Other Info.

輸流停電組別 H
饋線代號 H237
每度燃料成本 2.6639 元
本期被排量 103208 公升
每度繳交再生基金 0.0023 元

透過輸出台需電力

工廠常見的設備

30

◆ 冰水機(示意圖)



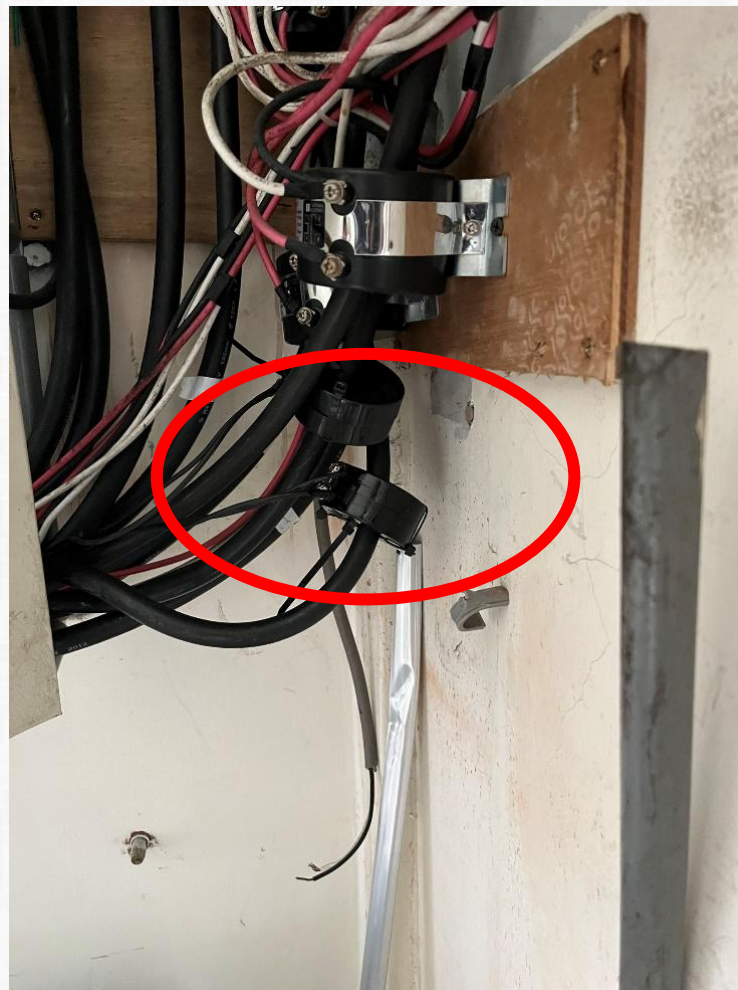
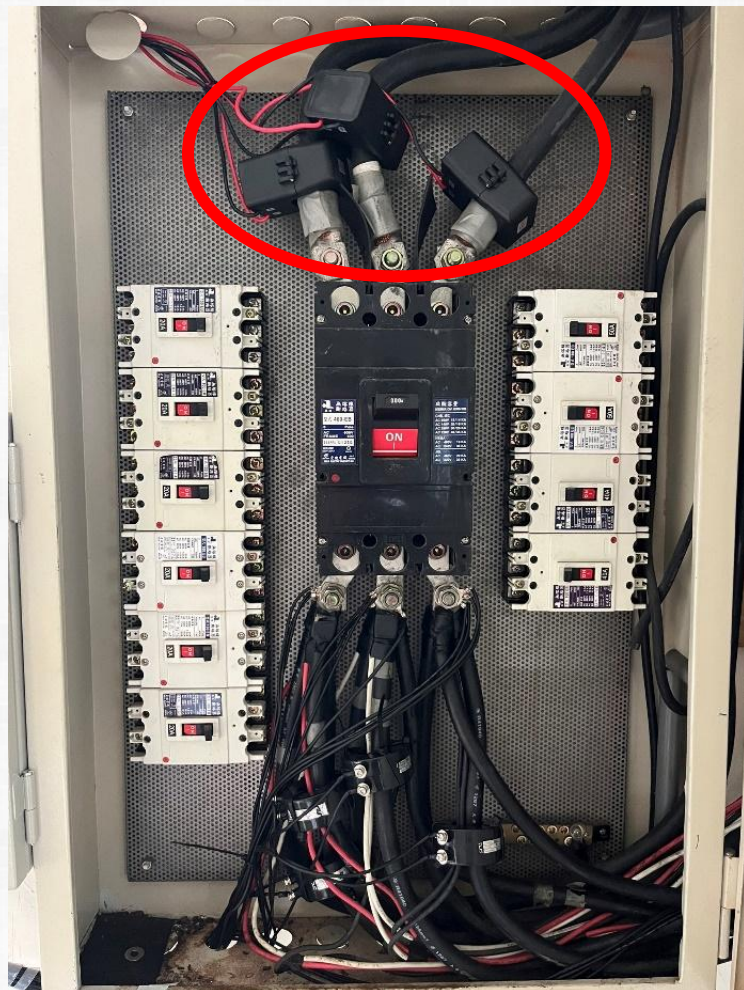
◆ 冷凍式空氣乾燥機(示意圖)



◆ CNC機台(示意圖)

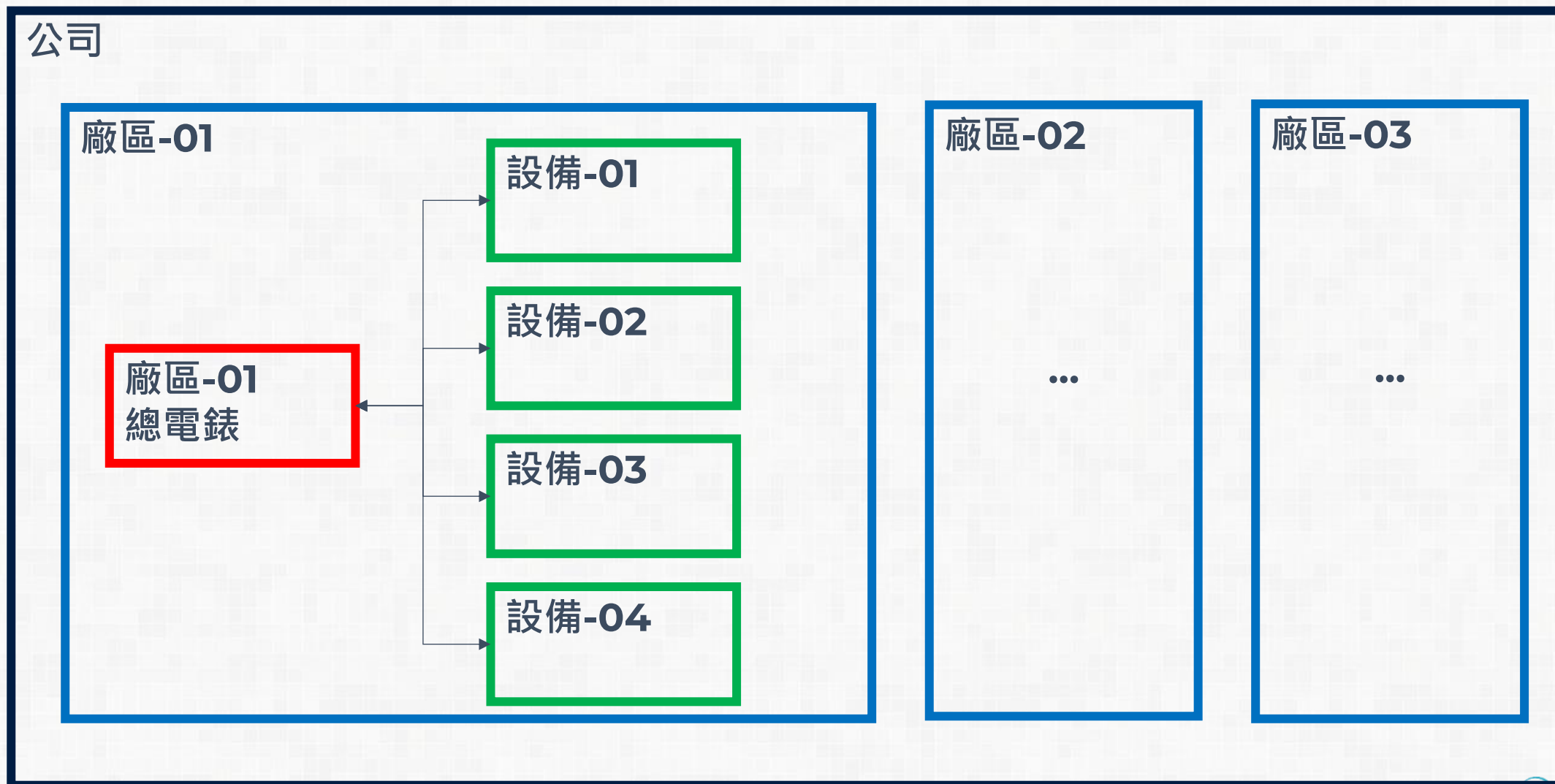


智慧電錶、比流器安裝(示意圖)



公司與廠區的用電迴路關係圖

32



在機台/設備上安裝電錶採集數據

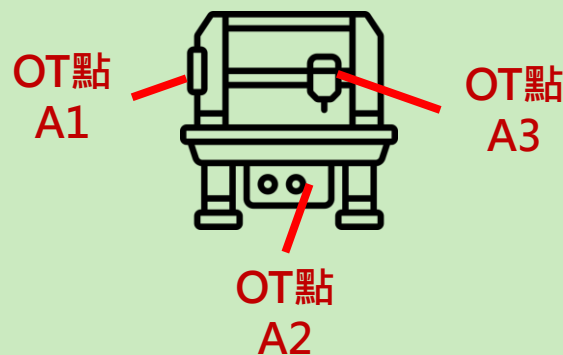
設備-01



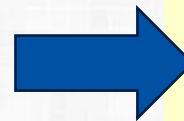
安裝電錶



智慧電錶



蒐集資料

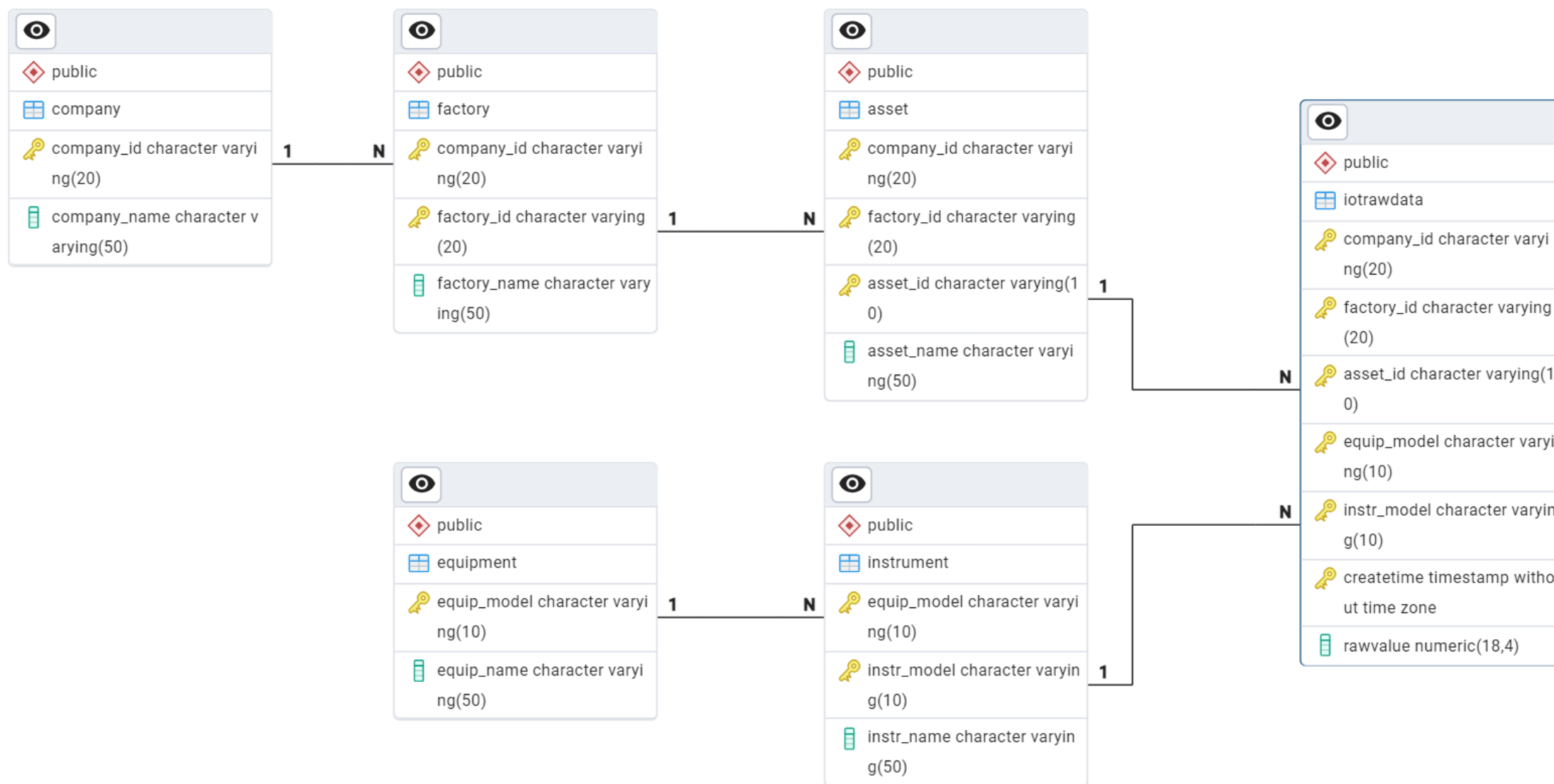


- A1 : 電壓(V)
- A2 : 電流(A)
- A3 : 能耗(F)
- A4 : 功率(P)
- A5 : 電能(W)



IoT資料庫設計與ER-Model

34



資料的整理與呈現

◆我們需要將搜集的資料彙整呈現給需求者，我們先討論如何將資料呈現其
分佈情形

◆呈現方式：

□ 圖比表好，表格比文字敘述好

◆統計表

□ 將蒐集得到的資料整理成表格的形式，並以文字或數字的形式表現出來，即是所謂的統計表。

➔我們依據資料性質(類別資料 v.s 數量資料)分別介紹常見的整理與統計表呈現方式。

◆一個完整的統計表至少應包括：

- 標題 title：包括表號 表序 與標題。
- 標目 label：標目是用來表示表身所要表示的項目或事實。
- 表身 body：表身是資料的主體 是統計表的核心。
- 資料來源及附註：應標明資料來源出處 以方便讀者查閱。

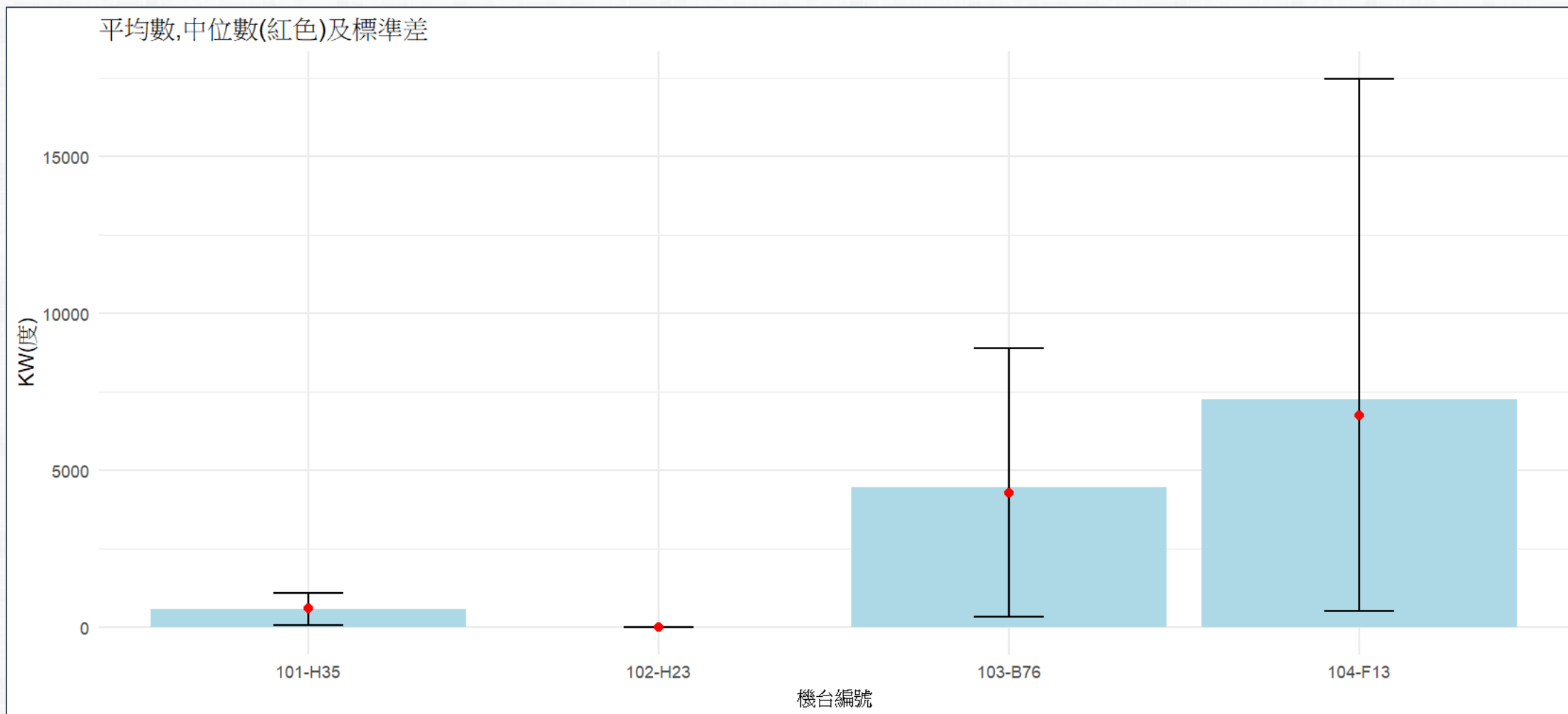
範例01：各機台用電統計資料(表格)

38

asset_id	range	max	min	counts	mean	median	Q1	Q2	Q3	variance	std_error	cv
101-H35	1037.286	1098.011	60.7250	14770	591.5968	605.669	332.8573	605.669	856.5153	9.285860e+04	2.50738380	0.51509254
102-H23	0.141	5.892	5.7510	5	5.8622	5.889	5.8880	5.889	5.8910	3.866700e-03	0.02780899	0.01060742
103-B76	8513.652	8857.328	343.6760	22992	4472.2961	4289.261	2393.7533	4289.261	6533.5781	5.384677e+06	15.30352481	0.51885893
104-F13	16305.408	17141.830	836.4221	203	7245.9960	6743.599	3503.3441	6743.599	10623.1665	1.921218e+07	307.63824138	0.60490929

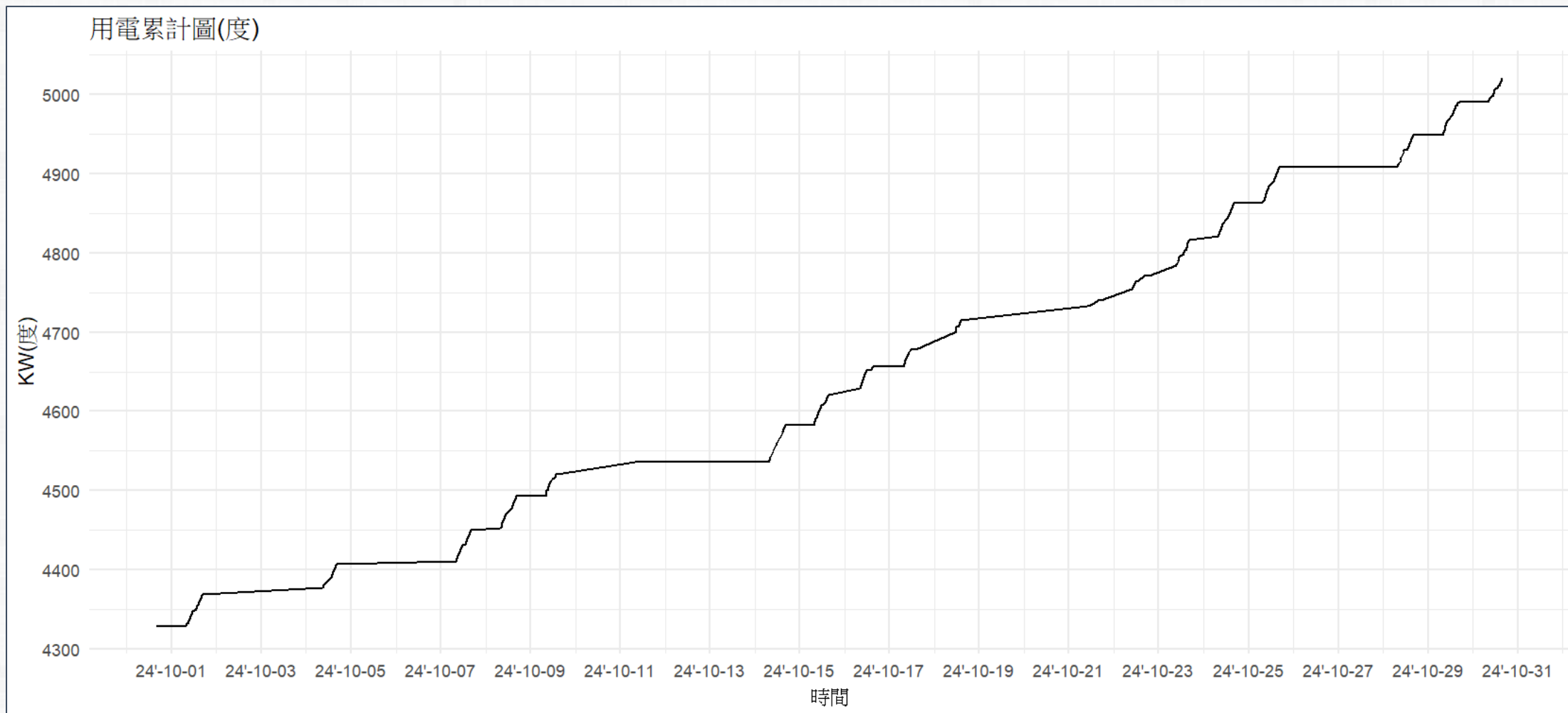
範例01：各機台用電統計資料(圖)

39



範例01：各機台用電統計資料(圖)→電錶量測累計用電(度數)

40



範例02：各機台用電統計資料/每小時(表)

41

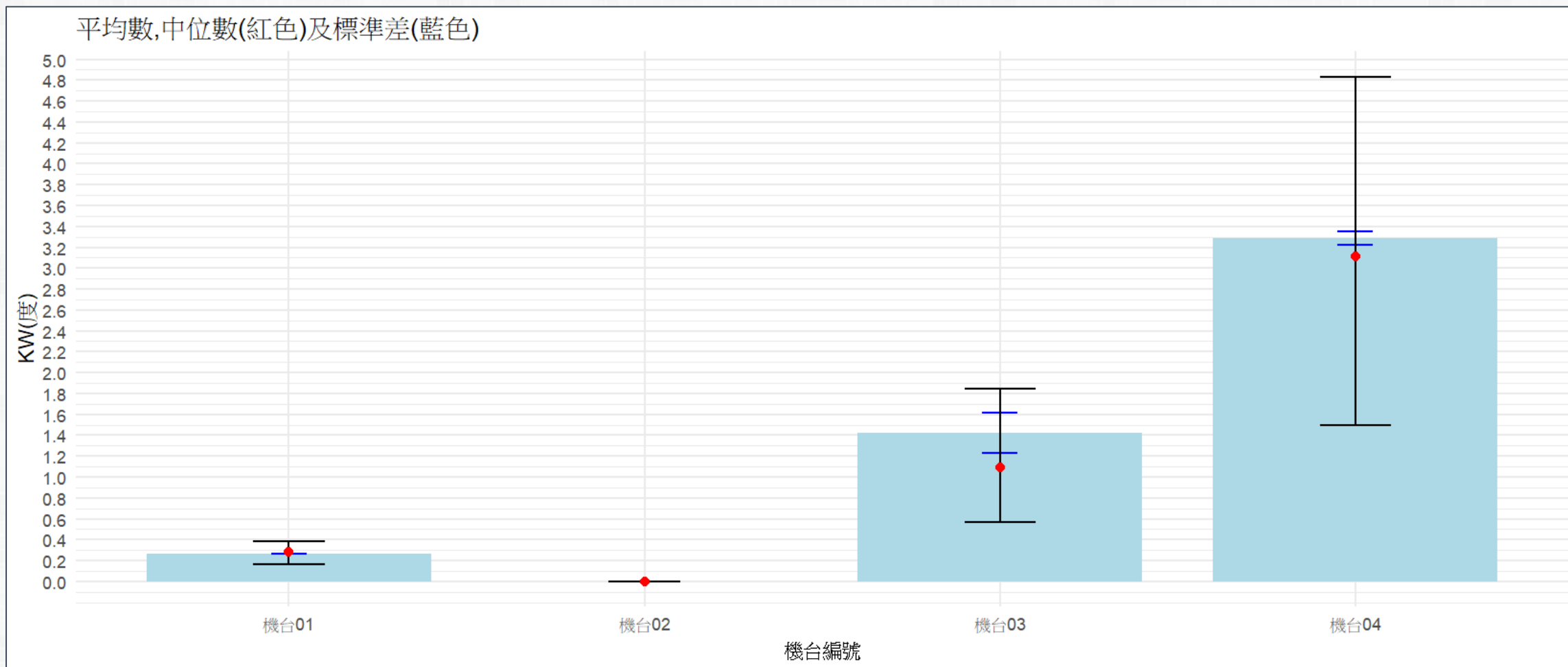
	asset_id	range	max	min	counts	mean	median	Q1	Q2	Q3	variance	std_error	cv
1	101-H35	0.4160	0.4160	0	1938	0.2491872	0.2619	0.13695	0.2619	0.380900	0.01852599	0.003091817	0.5462167
2	102-H23	0.0030	0.0030	0	3	0.0010000	0.0000	0.00000	0.0000	0.001500	0.00000300	0.001000000	1.7320508
3	103-B76	13.1724	13.1724	0	1905	2.1386367	1.3872	0.00000	1.3872	3.913100	5.45738308	0.053523528	1.0923334
4	104-F13	5.7620	5.7620	0	140	0.4069871	0.0000	0.00000	0.0000	0.500975	0.83470781	0.077215274	2.2448460

➤ 移除該小時用電度數為0

	asset_id	asset_name	range	max	min	counts	mean	median	Q1	Q2	Q3	variance	std_error	cv
1	101-H35	機台01	0.3861	0.4160	0.0299	1819	0.2654892	0.28100	0.167900	0.28100	0.385000	0.01540826	0.002910452	0.4675521
2	102-H23	機台02	0.0000	0.0030	0.0030	1	0.0030000	0.00300	0.003000	0.00300	0.003000	NA	NA	NA
3	103-B76	機台04	13.1157	13.1724	0.0567	1239	3.2882187	3.11400	1.495900	3.11400	4.833050	4.61013086	0.060998755	0.6529741
4	104-F13	機台03	5.3616	5.7620	0.4004	40	1.4244550	1.08805	0.566925	1.08805	1.841425	1.48848456	0.192904417	0.8564922

範例02：各機台用電統計資料/每小時(圖)

42



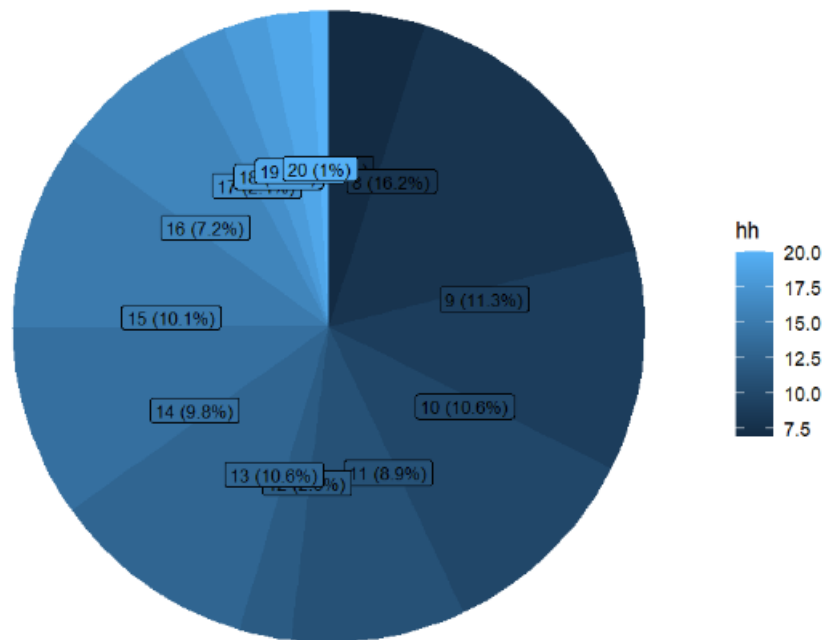
範例02：工作時用電分佈

	hh	range	max	min	counts	mean	median	Q1	Q2	Q3	variance	std_error	cv
1	7	4.4289	4.4609	0.0320	238	0.7785462	0.2335	0.146750	0.2335	1.065900	0.975498279	0.06402134	1.2686121
2	8	11.7753	11.7783	0.0030	334	2.5789608	0.4100	0.310225	0.4100	4.727550	9.632921233	0.16982662	1.2034672
3	9	13.1404	13.1724	0.0320	335	1.7902307	0.4100	0.295050	0.4100	3.142550	5.043692111	0.12270207	1.2544844
4	10	9.3104	9.3403	0.0299	327	1.6806520	0.4090	0.259400	0.4090	3.224800	3.869619346	0.10878280	1.1704593
5	11	7.5968	7.6289	0.0321	330	1.4188121	0.4080	0.260025	0.4080	2.294650	2.819524999	0.09243384	1.1834859
6	12	5.4242	5.4581	0.0339	185	0.4198578	0.3490	0.217000	0.3490	0.410000	0.309634477	0.04091088	1.3253250
7	13	7.9900	8.0220	0.0320	311	1.6763489	0.4070	0.255550	0.4070	3.331800	4.014915911	0.11362085	1.1952915
8	14	9.2591	9.2901	0.0310	310	1.5587948	0.4100	0.292500	0.4100	2.392400	3.855601272	0.11152320	1.2596709
9	15	7.4592	7.4902	0.0310	297	1.6040606	0.4100	0.271000	0.4100	3.113300	3.668657847	0.11114128	1.1940784
10	16	5.7299	5.7620	0.0321	302	1.1468868	0.3835	0.222325	0.3835	1.782825	2.042051736	0.08222993	1.2459851
11	17	4.0969	4.1309	0.0340	48	0.3769167	0.3210	0.195225	0.3210	0.393750	0.337492089	0.08385157	1.5412974
12	18	2.9762	3.0132	0.0370	47	0.3479596	0.2980	0.163500	0.2980	0.373000	0.192765375	0.06404210	1.2617861
13	19	2.1518	2.2159	0.0641	25	0.3538120	0.3139	0.192000	0.3139	0.390000	0.164351308	0.08108053	1.1458137
14	20	0.1130	0.1870	0.0740	10	0.1531800	0.1645	0.151725	0.1645	0.176725	0.001309366	0.01144275	0.2362264

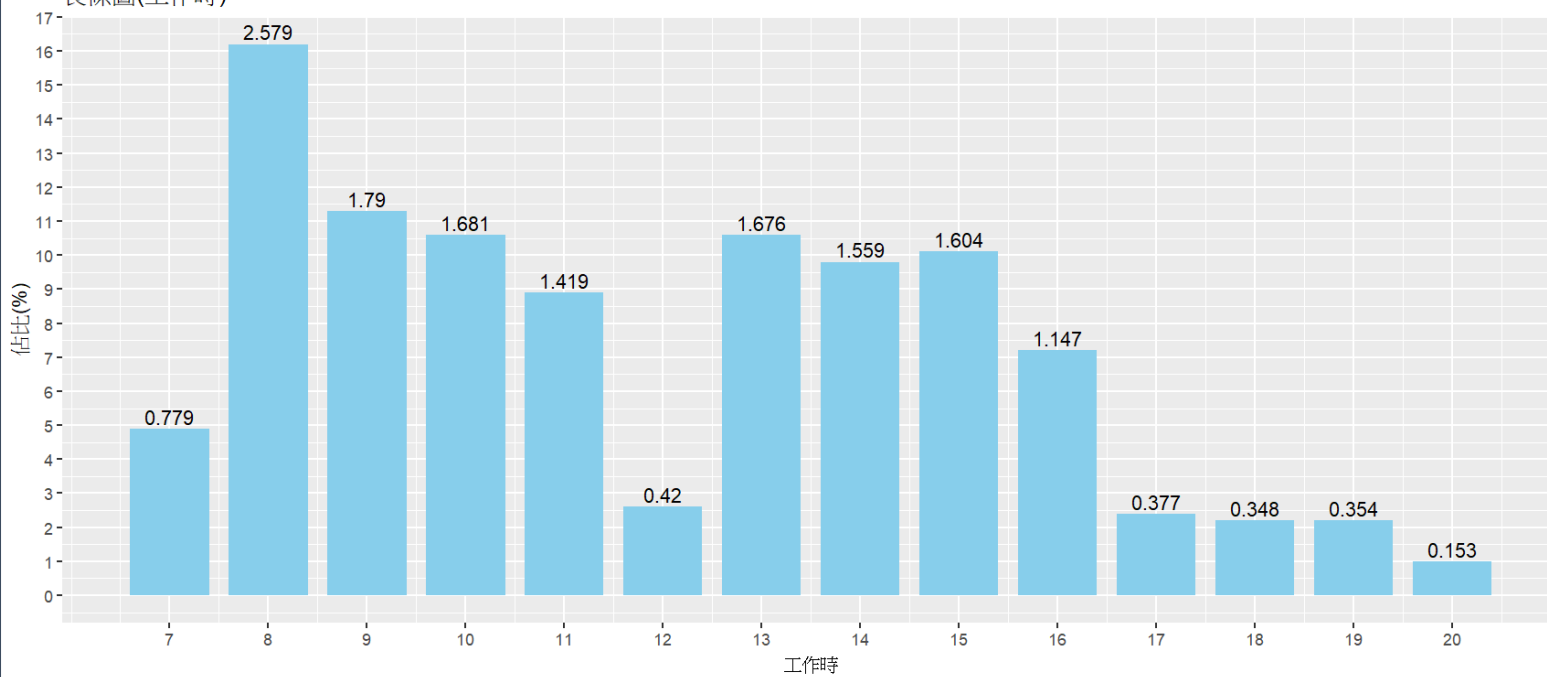
範例02：工作時用電分佈

44

圓餅圖(工作時)

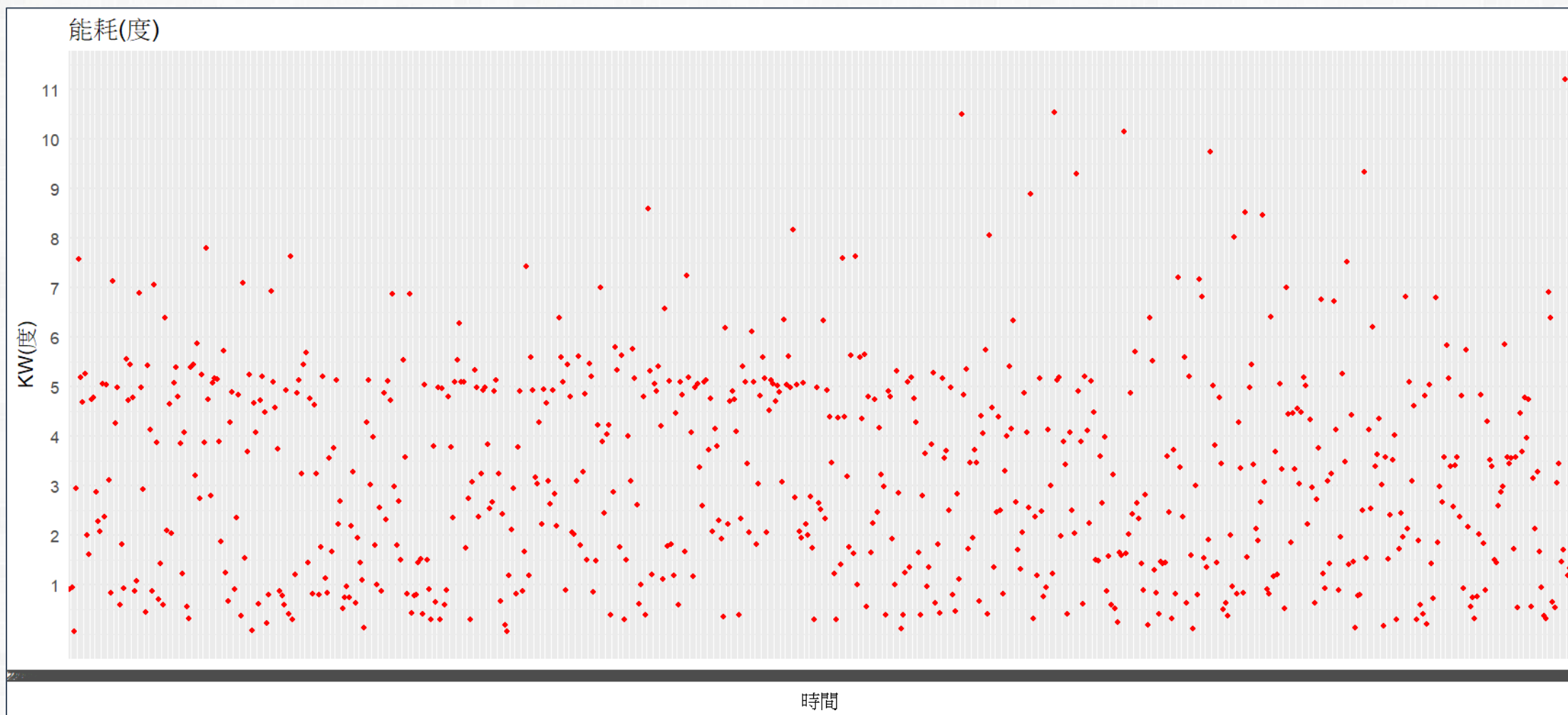


長條圖(工作時)



範例02：機台用電散佈(圖)→用電度數/小時

45



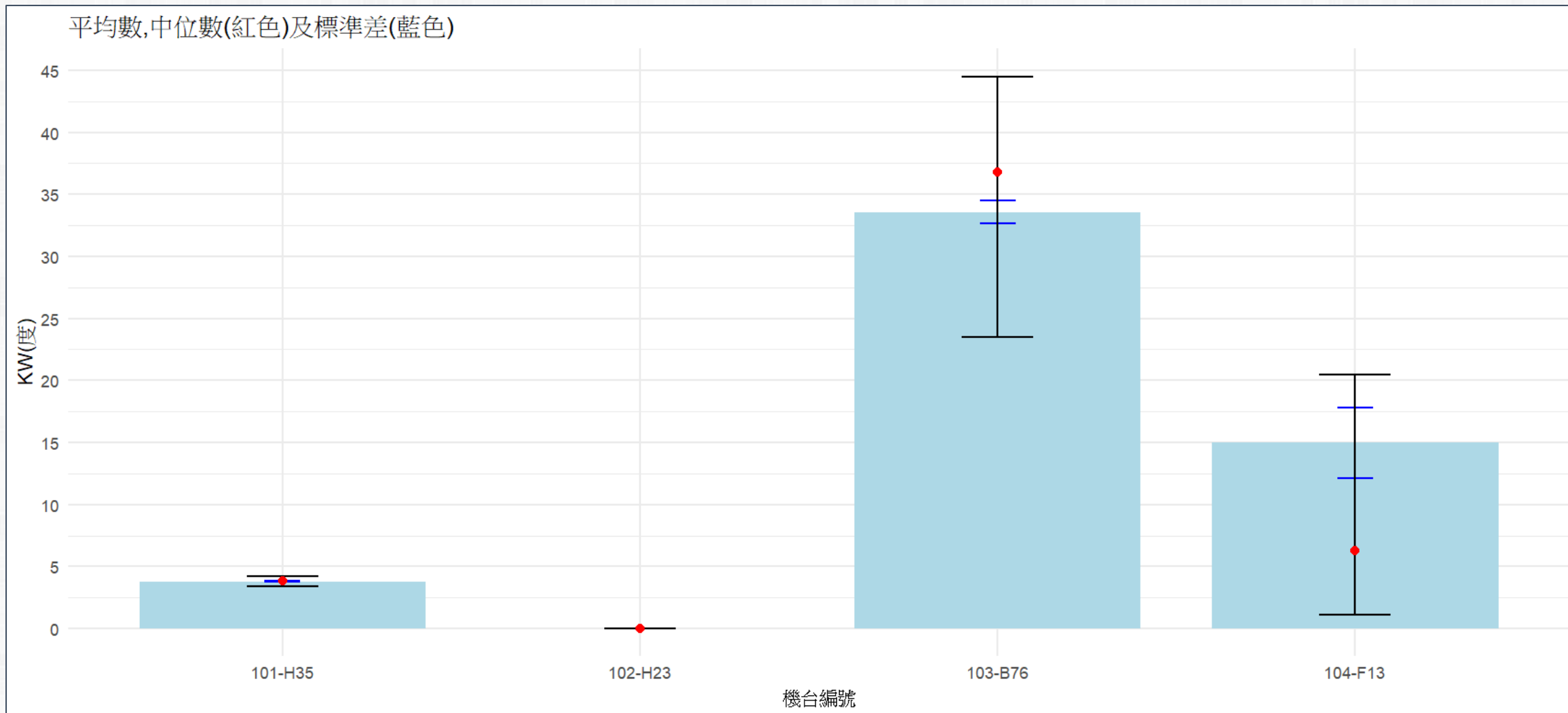
範例03：各機台用電統計資料/每日(表)

46

	asset_id	range	max	min	counts	mean	median	Q1	Q2	Q3	variance	std_error	cv
1	101-H35	5.3119	5.7780	0.4661	239	3.791631	3.8140	3.41095	3.8140	4.18095	0.6194212	0.05090895	0.2075711
2	102-H23	0.0000	0.0040	0.0040	1	0.004000	0.0040	0.00400	0.0040	0.00400	NA	NA	NA
3	103-B76	59.7481	60.4531	0.7050	217	33.561035	36.8164	23.45900	36.8164	44.50510	180.0756048	0.91095636	0.3998454
4	104-F13	79.3652	79.7656	0.4004	45	14.965684	6.2620	1.15020	6.2620	20.45800	356.6323276	2.81516657	1.2618675

範例02：各機台用電統計資料/每日(圖)

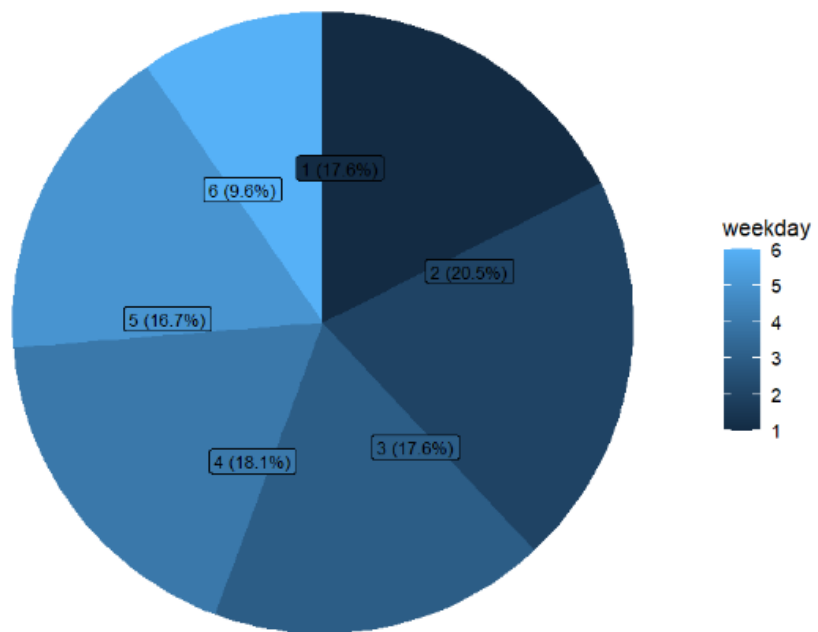
47



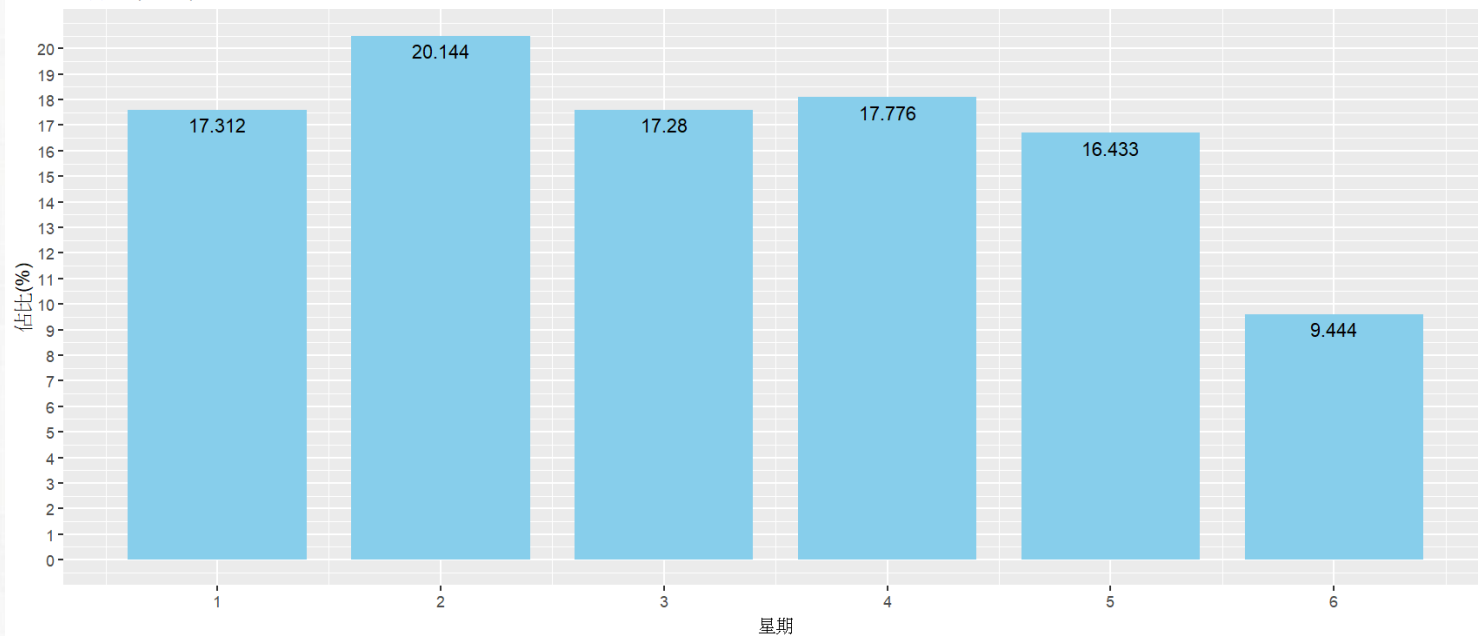
範例03：每日的用電分佈

48

圓餅圖(工作時)

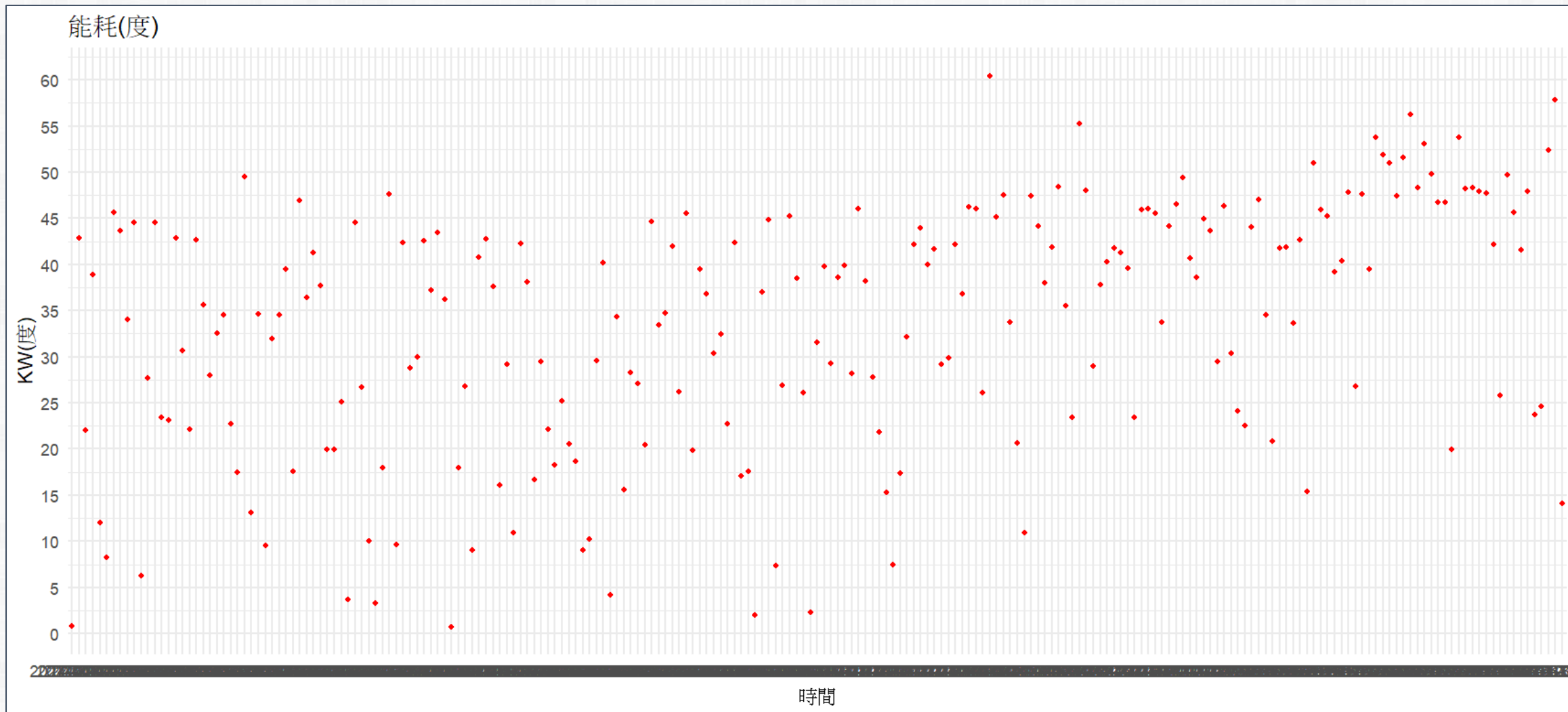


長條圖(星期)



範例03：機台用電散佈(圖)→用電度數/天

49



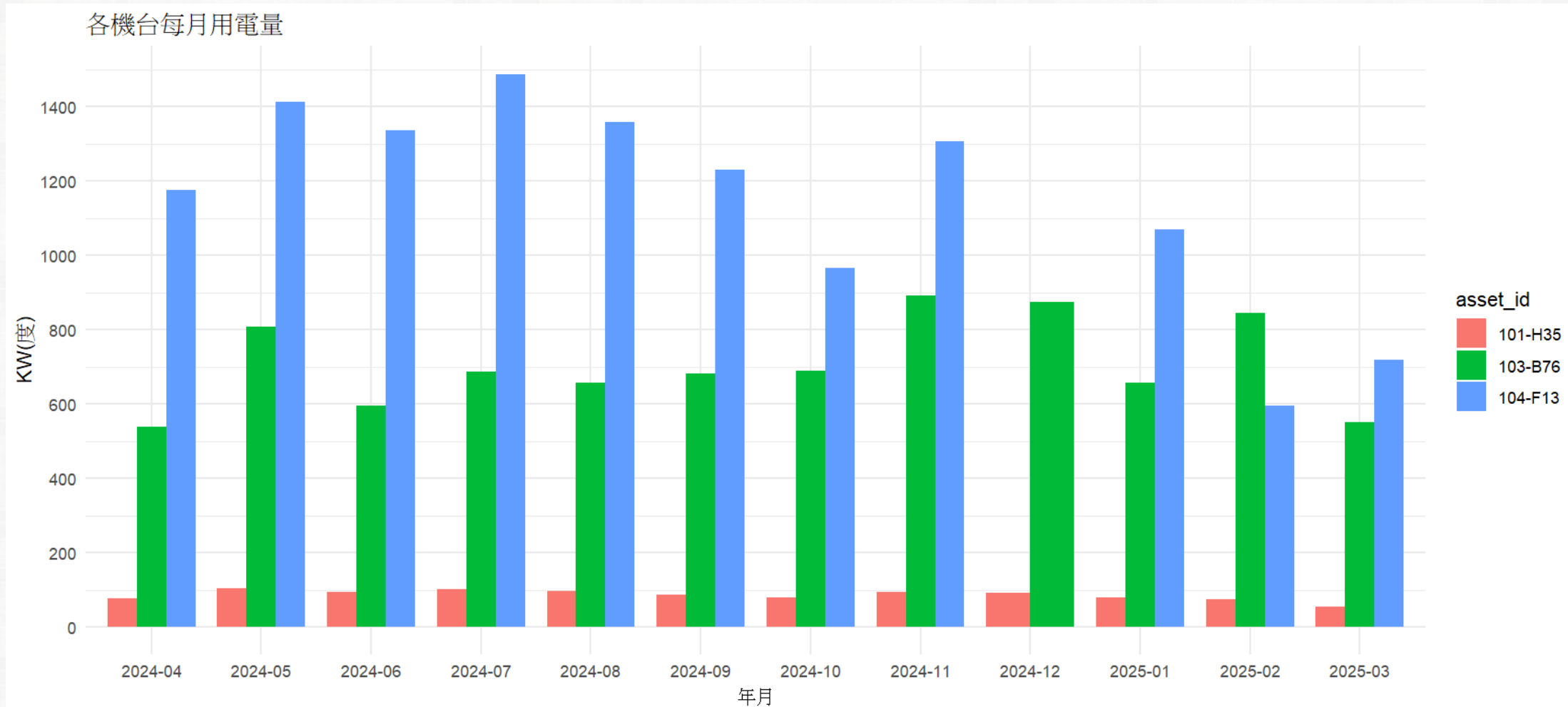
範例04：各機台用電統計資料/每月(表)

50

	asset_id	range	max	min	counts	sum_data	mean	median	Q1	Q2	Q3	variance	std_error	cv
1	101-H35	50.1469	103.884	53.7371	12	1032.269	86.02241	89.0880	78.46343	89.0880	95.12117	198.6568	4.068751	0.1638476
2	103-B76	352.9239	891.456	538.5321	12	8477.132	706.42768	685.2886	641.02557	685.2886	816.28620	14741.9194	35.049869	0.1718737
3	104-F13	892.1177	1488.446	596.3281	11	12662.325	1151.12043	1229.5503	1018.95460	1229.5503	1347.43860	82712.4961	86.714000	0.2498416

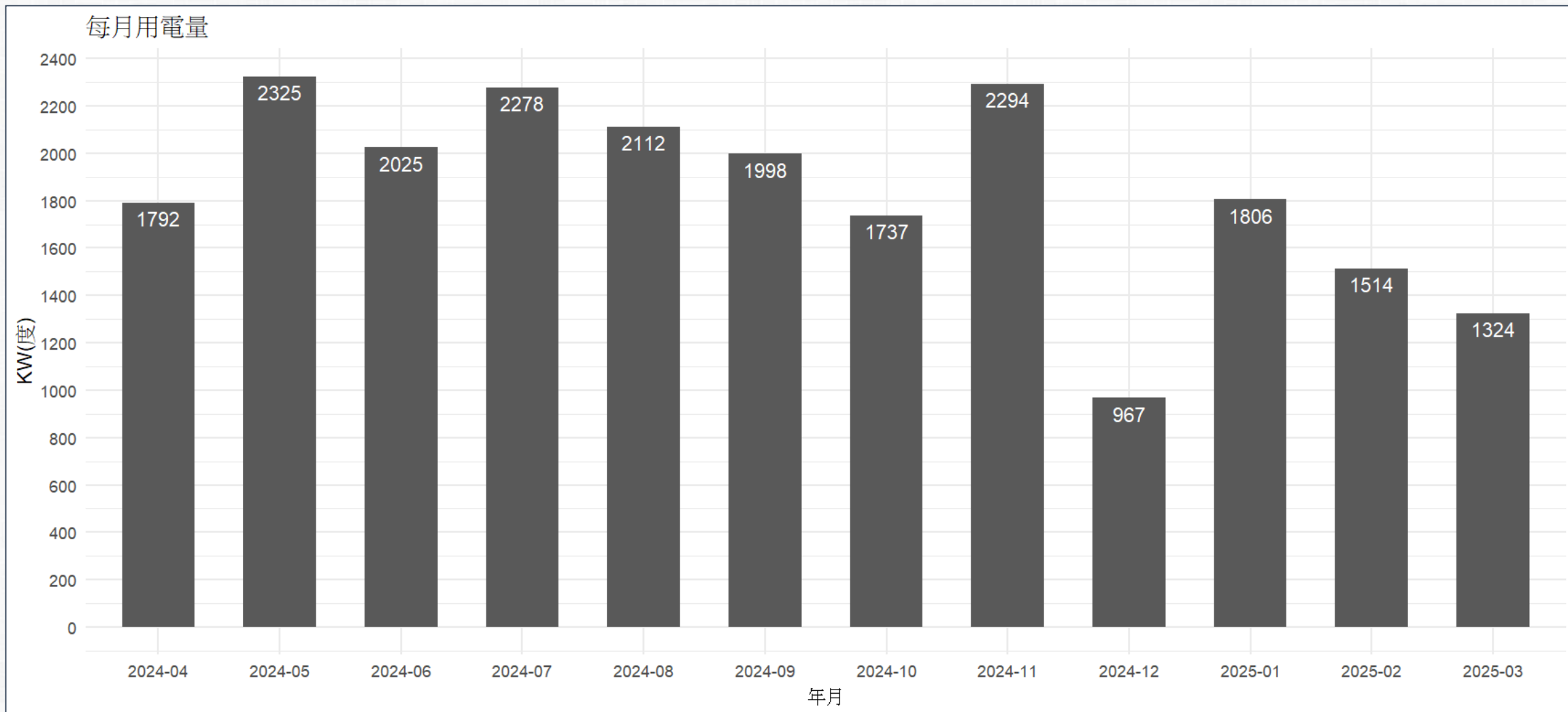
範例04：每月用電量 By 機台用電量

51



範例04：每月用電量 By 月總用電量

52



範例04：每月生產數量

53

product_month	product_amout
2024-04-01	1,742
2024-05-01	2,250
2024-06-01	1,960
2024-07-01	2,206
2024-08-01	2,052
2024-09-01	1,943
2024-10-01	1,692
2024-11-01	2,227
2024-12-01	937
2025-01-01	1,757
2025-02-01	1,469
2025-03-01	1,284

範例04：找生產數量與用電量的關係

54

$$C(X) = 8.46 * sum_energy + others$$
$$sum_energy = \alpha + \beta \times product$$

where,

sum_energy = 用電量/月

$product$ = 生產數量/月

範例04：找生產數量與用電量的關係(續)

55

$$C(X) = 8.46 * sum_energy + others$$
$$sum_energy = \alpha + \beta \times product$$

```
Residuals:
    Min       1Q   Median       3Q      Max
-6.1420 -3.1501 -0.1874  3.5160  5.8108

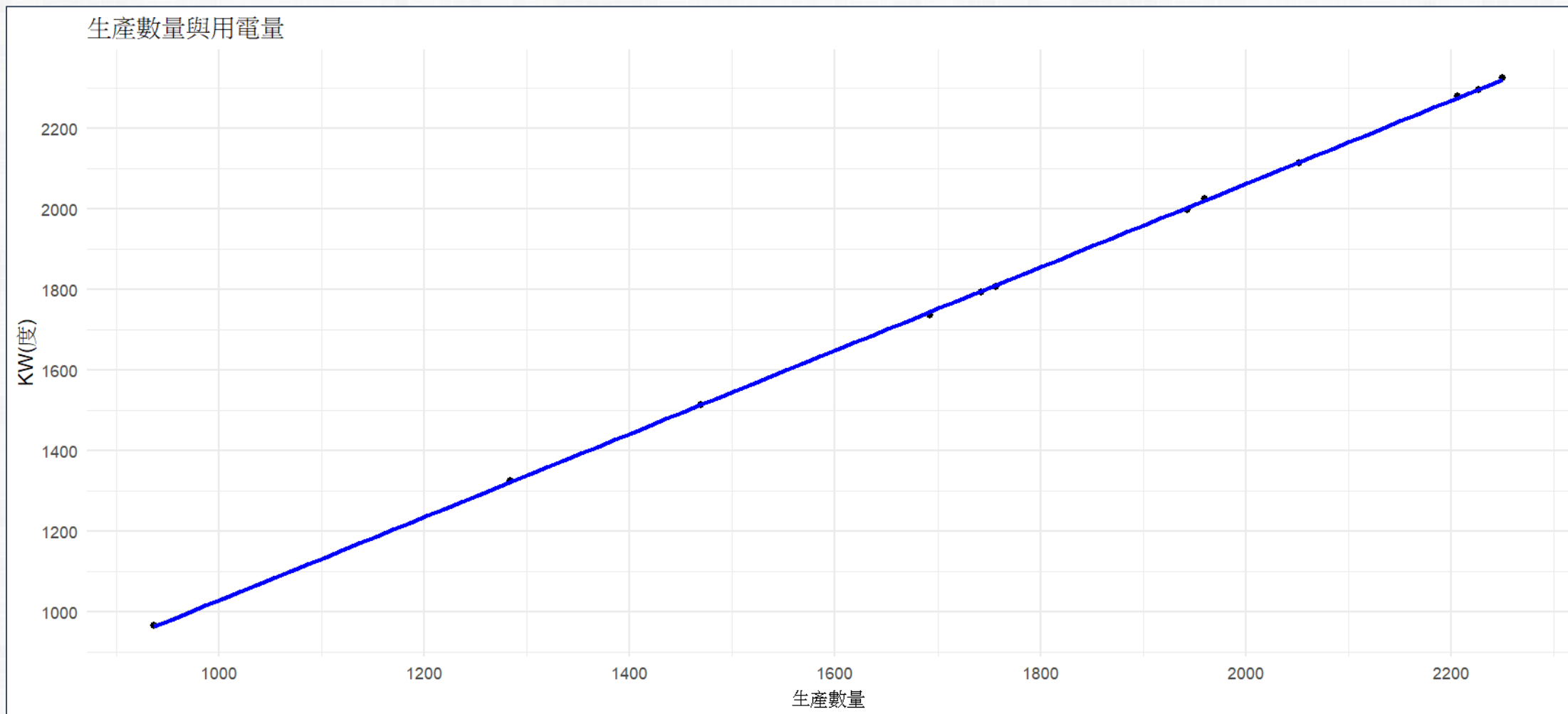
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.581799    5.904658  -0.607    0.558
product      1.032343    0.003219 320.737 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.314 on 10 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
F-statistic: 1.029e+05 on 1 and 10 DF,  p-value: < 2.2e-16
```

$$sum_energy = -3.58 + 1.032 \times product$$

範例04：找生產數量與用電量的關係(續)

56



THANK YOU.

“ Together We Are Stronger ”

ISCOM 采威國際資訊
Iscom Online International Information Inc.



ADDRESS

407臺中市西屯區西平里24鄰
漢翔東路33號



TELEPHONE

(04) 2326-5200



E-MAIL

Services@iscom.com.tw

誠信 | 熱情 | 專業 | 永續

ISCOM 采威國際資訊
Iscom Online International Information Inc.