如何使用傳播調查資料庫的長期追蹤資料

中央研究院社會學研究所
吳齊殷

2021/10/15

長期資料介紹

長期資料I
時間序列研究 Time-Series Study

- 經濟學家較常用，透過蒐集**單一對象之歷年或長期資料**，觀察某一國家的長期經濟趨勢、某經濟體之特性、某一企業的發展營運狀況。
- 可以追蹤十年、二十年或甚至三十年的發展趨勢。
- 例：歐盟經濟發展狀況、台灣國民所得成長率

長期資料II
趨勢分析 Trend Analysis

- 在**不同時間點**，對**不同研究對象**，以**相同或類似的議題**進行長期資料蒐集。
- 例：社會變遷 - 每個議題以五年為周期，進行調查。

長期資料III
世代研究 Cohort Analysis

- 與趨勢分析相近，但世代分析一般會鎖定**特定年齡層（某一世代）**為研究對象，或者針對兩群或以上的年齡層群體進行長期資料蒐集。
- 例：教育改制前後的差異，則需要一群教改前的學生，以及另一群教改後的學生進行資料蒐集，並進行分析，才能釐清改制前後的差異。

長期資料IV
長期追蹤資料 Panel Study

- 固定樣本貫時追蹤研究
- 針對**同一群人 / 同一批研究對象**，在**多個時間點**進行資料蒐集。
- 例：傳播調查資料庫（二期一次與二期三次）、台灣青少年研究、臺灣教育長期追蹤資料庫

長期資料IV
長期追蹤資料 Panel Study

■ 優勢：

1. 能觀察受訪者的動態變化
2. 能研究因果次序
3. 能描繪受訪者的變化軌跡
4. 控制測量誤差和無法觀察到之其他概念變項的干擾
5. 能驗證不同理論假設與模型

■ 缺點：

1. 需要大量經費執行
2. 問題與量表設定要清晰準確
3. 研究設計要很仔細和完善
4. 調查過程很容易有樣本流失
5. 分析困難

長期追蹤資料分析方法

資料分析方法 -

- 兩個時間點的分析
  - 配對樣本t檢定
  - 配對樣本類別資料分析
  - 差異為依變數的迴歸模型
  - 差異中的差異模型(D-i-D)
- 結構方程架構
  - 存活分析
  - 事件歷史分析
  - LGM
- 多時間點線性模型：
  - 固定與隨機效果模型
  - GEE
  - autoregressive cross-lagged model
  - 成長曲線模型與混合成長模型
- 多時間點的變異數分析：
  - Repeated Measures ANOVA與MANOVA

資料分析方法I

透過多個時間點，觀察個體的變與不變，建立模型把個體特質納入模型估計或排除

- **固定效果模式Fixed-effect Models**

將不受時間變化的變項(性別、教育程度)去除後進行模型估計

常見估計方法:

1. 差異法(兩個時間點)
2. 虛擬變數(多於兩個時間點)
3. 離散法(樣本數大)

- **隨機效果模式Random-effect Models**

此模式依循固定效果模式而來，在允許每個人(樣本)的差異性下進行估計
例:不同的成績表現

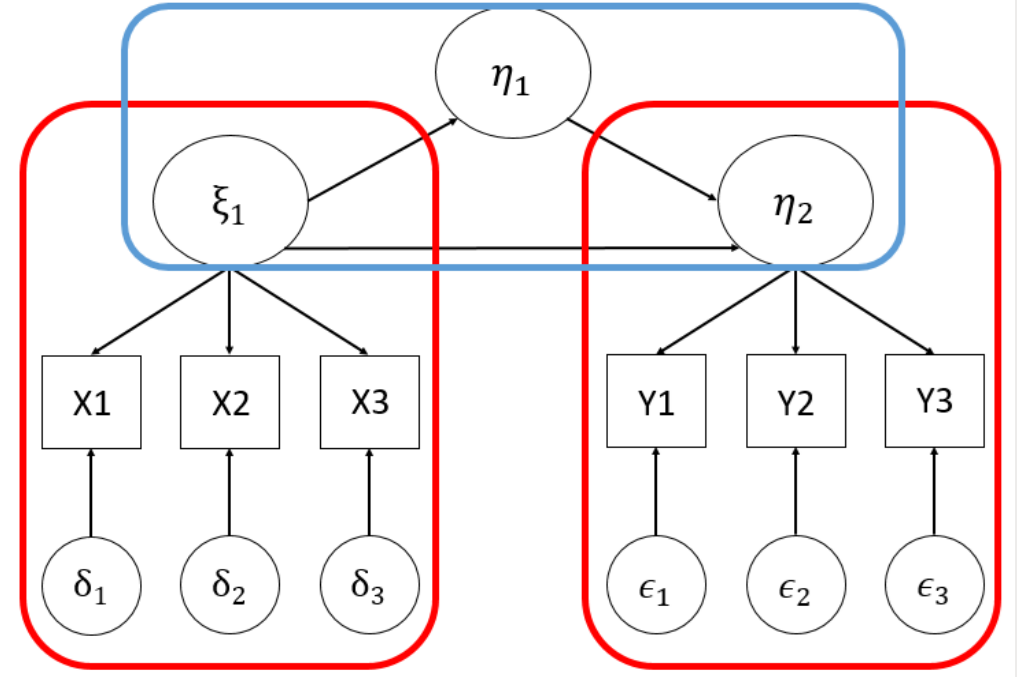
- **混合效果模式Mixed-effect Models**

同時包含固定效果與隨機效果兩種模式

資料分析方法II

結構方程模型Structural Equation Models, SEM

- 檢測多變項之間的因果關係，進而從探索性分析轉成確認性分析
- SEM可分為「測量模型」與「結構模型」兩部分
  前者主要探討潛在變數與觀察變數的關係，後者則是潛在變數之間的關係
- 優點：
  1. 考慮觀察變數的誤差
  2. 降低型一錯誤的發生
  3. 模型有多種配適指標提供參考
     (例:RMSEA、CFI、TLI)
- 缺點：
  1. 控制變項的類別(例外:社經地位)
  2. 結構方程式對於類別變項較難以分析
  3. 建立模型前的理論基礎要求



資料分析方法III

潛在成長曲線模型Latent Growth Curve Model, LGM

- 目的
  1. 增加自變項對依變項的預測關係並減少誤差
  2. 分析重複測量的變數在不同時間的變化
     例：在不同時間點下，個人成績或憂鬱情緒的成長軌跡

- 樣本變項中的改變與研究時經歷的時間長短變化有所關連

- 至少需要三波時間點的資料

- 能夠有效處理遺漏值、誤差等統計問題

- 建立模型時能依研究方向彈性調整

資料分析方法IV

階層線性模型Hierarchical Linear Modeling, HLM

- 資料的「多」層結構為重點
  1. 一筆資料有兩個或以上的抽樣單位
  2. 一個抽樣單位巢套或鑲嵌(nested)在另一個抽樣單位之下
     例:學生與班級、員工與公司

- 與OLS的抽樣單位獨立性相比，HLM則能夠分析單位之間的「相依性」

- 使用HLM的好處：
  1. 同時估計個體與總體層次的影響力
  2. 提高估計的準確性
  3. 能切割變異量，看出個體與總體層次的變異量大小
  4. 進一步探討跨層次的影響

資料分析方法 - 其他

- 多層次分析 Multi-Level Analysis
- 潛在類別模型 Latent Class Model
- 事件史分析法 Event History Analysis
- 反事實因果推論 Counterfactual Causal Inference
  - *傾向分數配對法*
- 序列分析 Sequence Analysis
- 群聚分析 Cluster Analysis
- 潛藏轉移模式 Latent Transition Model

研究方法實作

研究步驟一：

確認研究主題
查找文獻
釐清問題脈絡

- 瀏覽傳播調查資料庫的網站：link
- 相關傳播研究期刊：link

研究步驟二：

尋找資料庫
確認調查對象和問卷變項
根據問題和資料庫決定研究方法

研究步驟三：資料處理

- EDA (Exploratory Data Analysis) 探索式資料分析：了解資料、檢查有無離群或異常值、分析變間的關聯性（描述性統計、相關等）
- 多波次問卷合併：確認樣本ID/編號、變項名稱（建議可做一份自己的變項對照表）、合併方式（長型或寬型）
- 變項標準化：有些長期追蹤可能因時間橫跨幅度大，因此會有需進行物價指數調整（所得、財富等）或變項標準化的工作，像是教育成就、學科成就等，排除掉因為結構、文化差異或時間所產生的自然影響

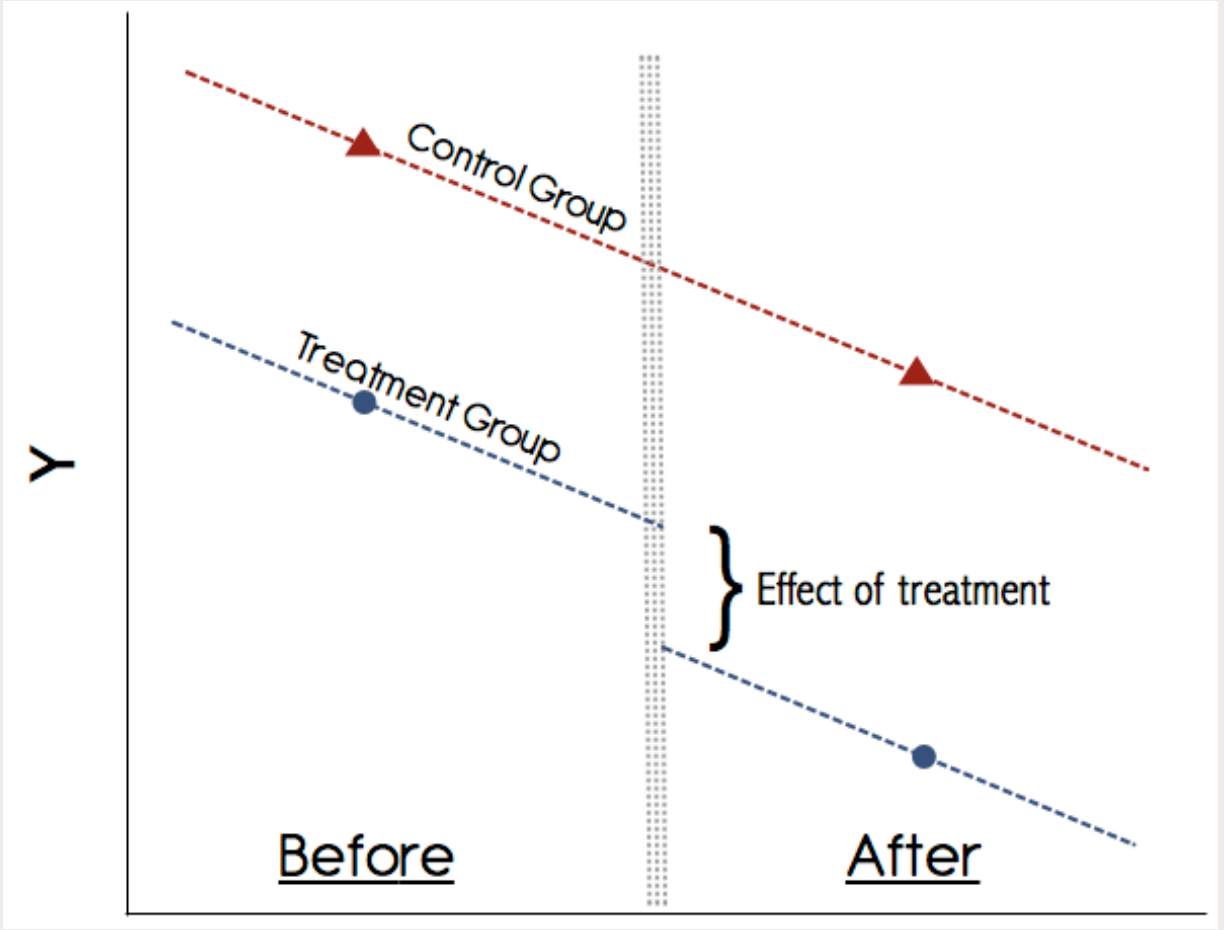
以傳播調查資料庫為例

- Panel data目前只有**兩波(兩時間點)**
- 只能選用兩時間點可用的模型
- 根據統計軟體所需，將兩波都有做的樣本整理成**長型**資料

兩時間點模型實作I - D-i-D

D-i-D
(Difference-in-Difference)

- 研究者想要知道的是介入的條件是否能夠改變受試者的情況。
- 這個模型所談的差異是指**團體**的平均差異，也就是說，依變數變化的估計是以團體為單位的。這跟我們之後會看到以個體為單位的模型有點不同



兩時間點模型實作I - D-i-D

假設研究者想瞭解性別在社會滿意度上，是否有差異，我們可以採用這個模型看看。

```
Call:
lm(formula = social_being ~ gender + time + did, data = data_3)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5643 -0.5643  0.3957  0.6054  2.6357

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.52429    0.10186  73.866   <2e-16 ***
gender      -0.09929    0.15280  -0.650    0.516
time         0.04000    0.06442   0.621    0.535
did         -0.07036    0.09664  -0.728    0.467
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.205 on 2516 degrees of freedom
Multiple R-squared:  0.007312,  Adjusted R-squared:  0.006128
F-statistic: 6.177 on 3 and 2516 DF,  p-value: 0.0003522
```

這個就是性別D-i-D的效果

兩時間點模型實作II - Fixed Effect/ Random Effect/ Mixed Effect

透過兩個時間點的比較，觀察**個體**的變與不變，而後進行檢驗決定採用何種模型。
假設研究者想看社會滿意度與虛實偏好，我們可以進行模型比較(Hausman Test)。

	Dependent variable:		
	social_being		
	OLS	panel linear	
	OLS	FE	RE
	(1)	(2)	(3)
viscon	-0.069***	-0.042***	-0.060***
	(0.008)	(0.010)	(0.008)
Constant	8.089***		8.011***
	(0.074)		(0.074)
Observations	2,520	2,520	2,520
R2	0.028	0.013	0.022
Adjusted R2	0.027	-0.975	0.022
Residual Std. Error	1.192 (df = 2518)		
F Statistic	71.237*** (df = 1; 2518)	16.330*** (df = 1; 1259)	57.523***
Note:			*p<0.1; **p<0.05; ***p<0.01

```
Hausman Test

data:  social_being ~ viscon
chisq = 7.2795, df = 1, p-value = 0.006975
alternative hypothesis: one model is inconsistent
```

由於p值小於0.05，顯示隨機效果的預設不符合資料。因此，選擇固定效果模型。

兩時間點模型實作III - GEE

GEE
(Generalized Estimating Equations)

- 非獨立的隨機樣本，由於同一個測試對象有多個資料點，如果直接用迴歸模型來分析的話，會違反研究單位獨立性的假設
- 為了在違反獨立性的狀況下，讓線性模型能夠準確被估計，研究者嘗試發展新的估計方法，Generalized estimating equation (GEE)就是應映這個需求發展出
- GEE的概念在於：由研究者設定描述這個關連性的相關矩陣，將這個關係矩陣放入估計的過程中，代表我們在參數的估計中同時控制了多波之間的關連性。最後，我們將這個關連性，連同線性模型的殘差項，一同當成新的殘差項，進而得到我們要的參數估計
- GEE所估計出來的模型一般又稱作population-averaged model，詮釋方式與subject-specific model全然不同

兩時間點模型實作III - GEE

- Liang & Zeger 曾經做過測試，發現使用GEE估計模型時，即使研究者選擇錯誤的關係矩陣，所估計出來的參數其誤差也不至於太大

- 使用Independence

$$
\begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \end{array}
\begin{array}{c}
\begin{array}{ccc} t_1 & t_2 & t_3 \end{array} \\
\begin{bmatrix} - & 0 & 0 \\ 0 & - & 0 \\ 0 & 0 & - \end{bmatrix}
\end{array}
$$

```
Call:
geeglm(formula = fm, family = gaussian, data = data_2, id = id1,
    corstr = "ind")

 Coefficients:
            Estimate Std.err    Wald Pr(>|W|)
(Intercept)   8.0886  0.1114 5271.28  < 2e-16 ***
viscon       -0.0676  0.0090   56.43  5.8e-14 ***
times         0.0253  0.0473    0.29     0.59
gender       -0.1891  0.0483   15.35  8.9e-05 ***
work          0.0528  0.0515    1.05     0.31
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = independence
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)     1.41  0.0517
Number of clusters:   2520  Maximum cluster size: 1
```

兩時間點模型實作III - GEE

- 使用Exchangeable

$$\begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \end{array} \begin{array}{c} \begin{array}{ccc} t_1 & t_2 & t_3 \end{array} \\ \begin{bmatrix} - & \rho & \rho \\ \rho & - & \rho \\ \rho & \rho & - \end{bmatrix} \end{array}$$

```
Call:
geeglm(formula = fm, family = gaussian, data = data_2, id = id1,
    corstr = "ex")

 Coefficients:
            Estimate Std.err    Wald Pr(>|W|)
(Intercept)   8.0886  0.1114 5271.28  < 2e-16 ***
viscon       -0.0676  0.0090   56.43  5.8e-14 ***
times         0.0253  0.0473    0.29     0.59
gender       -0.1891  0.0483   15.35  8.9e-05 ***
work          0.0528  0.0515    1.05     0.31
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = exchangeable
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)     1.41  0.0517
  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha        0       0
Number of clusters:   2520  Maximum cluster size: 1
```

兩時間點模型實作III - GEE

- 使用Autoregressive

$$
\begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \end{array}
\begin{array}{c}
\begin{array}{cccc} t_1 & t_2 & t_3 & t_4 \end{array} \\
\begin{bmatrix} - & \rho & \rho^2 & \rho^3 \\ \rho & - & \rho & \rho^2 \\ \rho^2 & \rho & - & \rho \\ \rho^3 & \rho^2 & \rho & - \end{bmatrix}
\end{array}
$$

```
Call:
geeglm(formula = fm, family = gaussian, data = data_2, id = id1,
    corstr = "ar1")

 Coefficients:
            Estimate Std.err    Wald Pr(>|W|)
(Intercept)   8.0886  0.1114 5271.28  < 2e-16 ***
viscon       -0.0676  0.0090   56.43  5.8e-14 ***
times         0.0253  0.0473    0.29     0.59
gender       -0.1891  0.0483   15.35  8.9e-05 ***
work          0.0528  0.0515    1.05     0.31
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = ar1
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)     1.41  0.0517
  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha        0       0
Number of clusters:   2520  Maximum cluster size: 1
```

兩時間點模型實作III - GEE

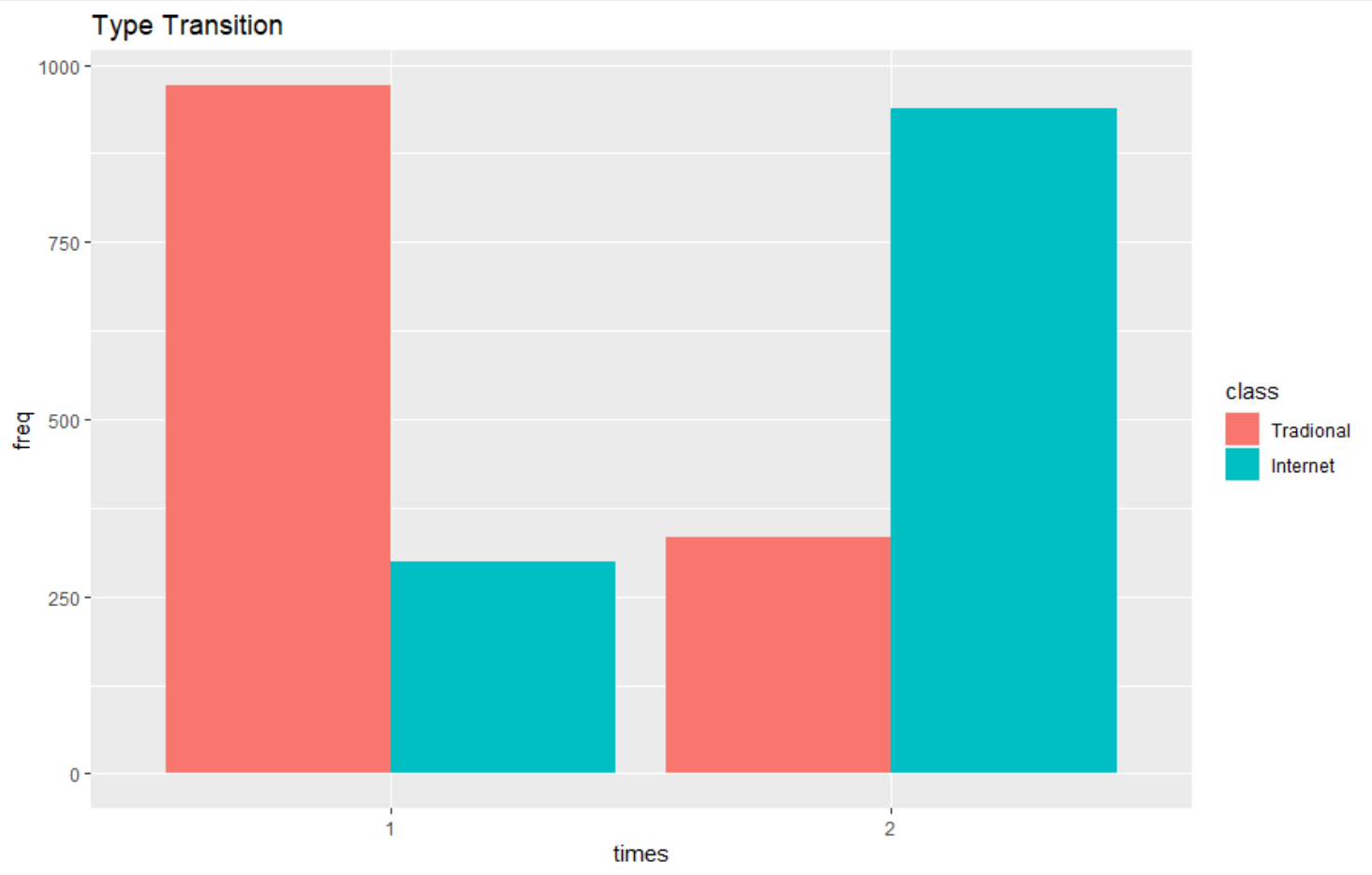
- 當研究目的僅在於描述自變數(time-varying or time-invariant variables)對依變數的平均變化的影響，而不涉及解釋個體變化的情況時，GEE會是很好的選擇
- 然而，當研究目的是想要瞭解何種因素影響個體的變化時，我們需要運用到subject-specific model

實作提點 - Latent Transition Analysis(LTA) vs Latent Growth Model(LGM)

- LTA處理**類別**類型資料(Latent Class Analysis用長期資料做)
- LGM處理**連續**類型資料

- 例如：研究者想分析不同媒介管道(報紙/電視/雜誌)使用者的變化性，就可以使用LTA來進行。

實作提點 -



- 簡單描述不同時間點的比例變化。
- 可以得知被歸類在傳統紙本媒介的使用者，在第二次調查大幅下降；反而是使用電子化媒介人口的更多了。