



## 一、樣本資料型態與變項

研究的資料收集，是根據理論建構的變項而收集。

變項 **Variable** 是指「某種特徵的集合，而該特徵具有質或量上的差異，亦即可變的」。

調查與數量研究的目的係收集可以比較的變項資料以供統計分析，由於變項資料型態 (**data types**) 不同(注<sup>1</sup>)，使得計量精緻程度亦有別。統計學家一般把資料分成四種型態：類別資料 (**nominal data/ categorical data**)、等序資料 (**ordinal data/ rank data**)、等距資料 (**interval data**)、及等比資料 (**ratio data**)。應用上可簡化分為2大類：類別資料、連續資料。

### (一)類別資料

類別資料是收集所得的資料是一個「類別」而不是一個可以比較大小的「數值」，譬如調查「性別」的資料是「男」或「女」，調查「電視收視率」的資料則是「臺視」或是「中視」、「華視」。

**data types** 或稱 **data levels, data** 就是指 **variable** (變項) 的值，所以類別資料也可稱類別變項，以此類推。

#### 1、類別資料的水準

類別資料所包括的「類別」，統計上稱為「水準(**level**)」。譬如「性別」資料包括男、女2個類別，就是有2個水準。如果把「職業」分作士農工商，就是有4個水準。如果有一個量表可以把人的「動機」分作高中低3類，就是有3個水準。

「水準」這個中譯其實有點怪怪的，但如果使用「類別資料的類別」，好像也很容易混淆，所以就當成「專門定義專有詞」，倒也差強人意。

#### 2、二元資料

有些類別資料，自然只能分為兩類：是或否，有或無。譬如性別只能分為男或女，具有數學上「二項分配」 (**binomial distribution**) 的性質，或稱為「二元資料」 (**dichotomous data**) (注<sup>2</sup>)。

---

(注<sup>1</sup>)**data types** 或稱 **scale types, levels of measurement**; **data** 有時以 **variable** (變項) 表示。也有文獻將資料型態分為更多型態。

(注<sup>2</sup>)**binomial** 中譯「二項」; **dichotomous** 中譯「二元」，但從適域性來看，「二元」較易理解，且學生對「二元」的語意反應較為敏銳，故本研究者引用此概念時均稱為「二元」。

另外，與**binomial** 相對的有**polynomial, multinomial**，中譯都是「多項」，但**polynomial**是指「多**terms**或多次」，只有二元；**multinomial**則是「多元」。所以作者建議「項」和「元」的意義最好要能明確區別。



至於分類在三種以上的則稱為「多項分配 (multinomial distribution)」，或「多元資料」(polychromous data)，譬如「電視收視率」的資料等。

有些資料事實上並不只2類，但在統計上把它合併為2類，譬如在選舉調查上分為只有「藍、綠」兩種政黨，也可以相當於「二元資料」，也可以適用二項分配的性質。這種處理方法，有些學者稱為「人為二元資料」(forced dichotomous data)。

binomial 中譯「二項」；dichotomous 中譯「二元」，但從適域性來看，「二元」較易理解，且學生對「二元」的語意反應較為敏銳，故本研究者引用此概念時均稱為「二元」。

### 3、虛擬變項與資料虛擬值 Dummy Variable and Dummy Value

**Dummy Variable** 虛擬變項，是指將類別變項的「二元變項」、或「人為二元變項」的2個水準，轉化為「0，1」，從而可以虛擬為連續變項，也可以使用適用於連續資料的統計工具。

「二元變項」如「性別」：0 為女，1為男。

「人為二元變項」：在多元類別資料時，如果將各水準設定：沒有= 0，有= 1，就是將各水準建構為「人為二元變項」。

如收視率研究中：建構「臺視」變項，0 為不看臺視，1為看臺視。

有或沒有，不一定要設為「0，1」，但通常設為差距為1。

如使用網路：沒有= 1，有= 2。

**Dummy Value** 虛擬值可以指以上Dummy Variable 虛擬變項中，特定的「0，1」值；也可以指任何類別變項中，對各水準設定1個數值。

類別資料也稱為質化資料 (qualitative data)，其中的多元資料處理相當繁複，實務上經常將其轉化為人為二元資料。

#### (二)連續資料

連續資料是指可以比較大小、先後...具連續性的資料，古典觀念又分為3級。

##### 1、等序資料

等序資料是收集所得的資料，可以比較大小，但是其間的差距不明確。譬如調查樣本對某公共政策的意見是「十分贊成？很贊成？贊成？反對？很反對？十分反對？」我們收回資料後，知道「十分贊成」的態度強於「很贊成」，「很贊成」又強於「贊成」，但是「十



分贊成」、「很贊成」與「很贊成」、「贊成」之間的差距是不是一致呢？我們並無法明確測量出來。

## 2、等距資料

等距資料就是資料數值之間的距離，可以明確測量出來。譬如溫度資料、長度資料，彼此之間的高下長短，就有很具體的差距。

## 3、等比資料

等比資料就是等距資料中又具有「真正絕對零度」的特殊情形的。如溫度就沒有真正絕對零度，不能稱為等比資料；而長度有絕對零度，故為等比資料。

## 4、連續資料與準連續性資料

不過，社會科學家在實際應用資料時，都假設等序資料的階級之間，具有一種抽象相等的距離；而等距資料和等比資料的區別又不十分必要，故常把等序、等距、等比三種資料合稱為「連續資料」

（continuous data），或量化資料(quantitative data)，和「類別資料」彼此對稱，作為選擇統計方法的重要依據。

類別資料中的二元資料，具備等同「0，1」的準連續性，也可適用連續資料的分析工具。

選擇統計方法時，常以觀察變項屬於類別資料或連續性資料，作為重要依據。

## 5、等序資料與無母數統計(Non-parametric statistics)

等距、等比2種資料所適用的統計方法，稱為母數統計(Parametric statistics)。

而類別資料與等序資料，適用的統計方法，稱為無母數統計(Non-parametric statistics)

如在SPSS的[資料型態](#)中：分作Scale, Ordinal, Nominal等3型；不過，在一般研究中，即碩士階段前的研究，我們都把前兩者合併為連續（可計量）資料，與第三型的類別資料對稱，作為選擇統計工具的依據。等序資料（Ordinal），只有配合無母數統計工具時應用。

等序資料與母數統計(Parametric statistics)或無母數統計(Non-parametric statistics)工具選擇，密切相關。我們在此先埋下伏筆，待[「機率分配」](#)單元後，再一併討論。

### (三)多項資料與多元資料

與binomial相對的有polynomial, multinomial，中譯都是「多項」，但polynomial是指「多terms或多次」，只有二元，形式如下：



$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0$$

multinomial則是「多元」，形式如下：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + e$$

所以作者建議「項」和「元」的意義最好要能明確區別。

「元」是指「變項」，代數的意義則是指未知數。方程式的效標可以不是變數，而是常數，所以，這時對代數而言，可以有一元方程式。

#### (四)資料收集的策略

資料收集有以下幾項應注意的策略。

##### 1、連續資料優先於類別資料

某一變項可為連續資料或類別資料時，應以連續資料型態收集，優先於類別資料型態。

典型的範例是「年齡」變項。

在傳統、與以抄寫剪貼為主的當前教科書，以及受到這些教科書與長期「社會相信」所影響的許多研究實務，在收集「年齡」變項時，都作了事前歸類，譬如：「20歲以下、21-40歲、41-60歲、61歲以上」等。

「年齡」事前歸類包括以下各種缺點：

1. 分類主觀，欠缺客觀與邏輯必然性。
2. 無法彈性調整，如果發現不應該以20歲為區間分割，已無法復原調整。
3. 無法提供細膩分析結果，「年齡」原來是連續資料，可較類別資料有更深入的分析。

4. 不易作為歷史資料可長期使用，而多成為「一次」研究資料。

「年齡」是連續資料，就應該以連續資料之型態收集，譬如詢問：

「請問你出生年，是民國幾年？」

這樣的資料型態有以下優點：

1. 可以提供細膩分析結果，同時，可以使用「陡階檢驗」，在有分類必要時，選擇更合理、具客觀與邏輯必然性的切點。譬如，年齡對人類行為影響，可能是發生在重要歷史事件上，如臺灣光復、蔣經國總統廢除戒嚴、李登輝總統直選當選...或臺灣躍居亞洲經濟4小龍之首、開放觀光、傳統產業外移...等等，而非「幾歲」。



2. 可以自由彈性建構不同的「年齡層」，甚至比較不同分類的結果。

3. 可作為歷史資料長期使用，進而與其他資料庫整合，可作為資料探勘、大數據分析之用。

在極早期的數量研究，為何像「年齡」等連續變項時，都作事前歸類？其實有時代的工具限制性。

那時都用手工運算，連續資料的分析運算，在資料量大時，非常吃力、容易出錯，所以改用較簡易、但深度也較淺的類別資料型態。

但在當前資料與電腦時代，運算可用軟體代勞，收集資料當然應以品質為第一優先選擇。

在當前研究方法、統計分析、微積分、及其他有傳統的科目教科書，還有許多沒有跟上資訊數位時代的例子，值得我們檢討、重視與改革。

## 2、物理/行為資料優先於心理/態度資料

這個部分，移至「測量設計」的「[心理、態度、行為量表](#)」，再詳細討論。

## 3、品質資料與量化資料的認識與抉擇

這個部分，請參考「[定量方法對定性方法](#)」專文。