# SUDAAN 說明文件

(本文件適用2005與2009NHIS資料分析用)

本次調查區域範圍為台灣地區 23 個縣市。調查對象為在台灣地區設有戶籍之常住人口（不包括居住國外者），中選樣本（人）即為受訪對象。調查母體與抽樣底冊以民國 93 年 12 月 31 日「台灣地區戶籍資料檔」之戶籍人口為抽樣母體。台灣地區 23 個縣市均可視為單一之母體，各縣市的抽樣作業程序完全獨立。

樣本的選取採多階段分層系統抽樣設計(Multi-stage stratified systematic sampling design)，各層內採用抽取率與單位大小成比例(Probability Proportional to Size, PPS)、等距抽樣法逐步抽出「鄉鎮市區」\「村里」、「鄰」、「人」。各縣市內「鄉鎮市區」均先依據其都市化程度和地理位置作分層，並依層別之不同，各層之樣本有的採二階段的方法抽出，有的採三階段的方法抽出。

一、分層

各縣市為獨立的層，各縣市內各「鄉鎮市區」之地理位置、人口分布與都市化程度有所不同，同時考量田野調查所需之人力資源配置、時間與經費下，為提升調查的效率，於是將各縣市內之「鄉鎮市區」作分層。其分層主要依據劉介宇等於 2004 年所作之台灣地區鄉鎮市區發展類型應用於大型健康調查抽樣設計之研究的結果，再參考各「鄉鎮市區」之地理位置、人口分布情形，以及過去調查訪問的經驗等因素加以微調。分層結果：一縣市最少為一層（即不分層）如新竹市所有區均屬一層，最多為四層如台北縣，總計台灣地區 23 個縣市，358 個「鄉鎮市區」共分成 53 個層別。

二、抽取率

由於各縣市內之人口數差異甚大，為取得足以代表各縣市的樣本數，控制抽樣誤差在一定範圍之內，又不讓整個樣本數過於龐大，於是各縣市採用不同之抽取率，是而未來在估算台灣地區之資料時，各縣市數值須先作加權處理。在制定抽取率時，設定縣市的基本樣本數為 800 人，以達到足夠大的推論樣本數。澎湖縣由於人口數較少，故獨立自成一群組，其抽取率為 10‰；台灣地區其餘 22 個縣市則依其人口數分成五個群組，人口數較多之縣市其抽取率較低，反之則有較高的抽取率。各群的抽取率分別設定為 1.00‰（人口數＞120 萬）、1.30‰（人口數為 70～120 萬）、2.00‰（人口數為 40～70 萬）、2.50‰（人口數為 30～40 萬）與 3.50‰（人口數＜30 萬）等五種。總計台灣地區預定抽出 30,275 個樣本，其整體抽取率約為 1.34‰，此抽取率大於 2001 年 NHIS 的抽取率（1.20‰）。

三、樣本數與各階段抽出單位數

各縣市內各層之基本樣本數再依人口數等比例分配到各層內（proportional allocation），抽出單位數則因「層」之特性而有三種不同的設計：

1.二階段抽出：運用於都市化程度較高的層別，第一階段抽出的單位為「鄰」，接著抽出「人」。該層內之「鄉鎮市區」全部涵蓋在內，如台北市之第一、二層與台北縣之第一層。

2.三階段抽出：運用於中度都市化程度、地理位置稍分散的層別，第一階段抽出的單位為「村里」，其次抽出「鄰」，最後抽出「人」。該層內之「鄉鎮市區」全部涵蓋在內，如高雄市之第二層與台北縣之第二層等。

3.三階段抽出：運用於一般「鄉鎮市區」與偏遠「鄉鎮」的層別，第一階段抽出的單位為「鄉鎮市區」，其次抽出「鄰」，最後抽出「人」。如台北縣之第三、四層與屏東縣之第二、三層等。

因考量各階段均須有基本的抽出單位數供計算抽樣誤差，所以各層之最後分配的樣本數與原分配之樣本數做了些許的調整。總計台灣地區 23 個縣市調整後之分配之樣本數為 30,680 人（原分配樣本數 30,275 人），358 個「鄉鎮市區」中共計抽出 187 個「鄉鎮市區」。

**結果**

一、抽樣執行結果

本次抽樣工作所使用之母群體資料是由衛生署資訊中心擷取自內政部戶役政資料檔之民國 93 年 12 月 31 日「台灣地區戶籍資料檔」，總人口數為 22,615,307 人。資料在經過清理修正，除去資料檔中之亂碼與錯別字後即進行分層，各縣市均按預定的抽出設計逐階段以 PPS 等距的方法進行抽樣，台灣地區 23 個縣市都依照抽樣設計抽出各層各階段的樣本。總計台灣地區抽出 187 個「鄉鎮市區」，共抽出 30,680 人，抽出之樣本資料在經衛生署資訊中心將抽樣操作過程基於個人資料保護理由加密之個案識別資料加以解密還原後，即製成樣本清冊並提供給國民健康局進行後續之面訪調查。本次調查各縣市於各類問卷所分配到的樣本數列於表五。

二、樣本分布與母體的一致性檢定

為檢視抽出樣本之代表性，利用「性別」及「年齡」等兩個變項進行抽出樣本與母體資料的一致性檢定，所有的卡方檢定結果均顯示在性別與年齡層上，各

縣市所抽出之樣本與母體資料均無顯著差異，顯示本次之樣本具有代表性。

三、加權

由於此次調查各縣市之樣本具有不同的抽取率，所以在分析時，如欲推估全國（台灣地區）的資料，各縣市的數值必須加權調整，才能縣市或全國代表性估計。權值採『事後分層』方式計算處理，各組權值的計算公式為：

$$W_i = \frac{N_i}{n_i} \times \frac{n}{N}$$

$W_i$：第 i 組的權值，組內每個案的權值均相同

$N_i$：母體第 i 組總人口數

$n_i$：第 i 組完訪樣本總個案數

$n$：完訪樣本總個案數

$N$：母體總人口數

加權的分組($i$)是以『性別』、『年齡』和『縣市內抽樣層別』等三個變數作為依據，其中『性別』分男性、女性共 2 組；『年齡』從 0 歲開始，每 5 歲取一組，80 歲以後的個案則合併成 1 組，總計分成有 17 組；『縣市內抽樣層別』則依照 2005NHIS 抽樣設計的結果，各縣市分別有 1 到 4 個不等的層別，總計台灣地區 23 個縣市，358 個「鄉鎮市區」共分成 53 個組別(組)，所以共有 1,802 個加權分組(2×17×53)。在實際計算權值時，由於有些組的樣本數太少（甚至是 0），必須做組的合併，於是便設定當該組的樣本數小於 3 時即做合併，最後共得到 1,731 組權值，受訪個案落在那一組，就用該組的權值。

計算權值使用的母體資料是由 93 年 12 月台閩地區各縣市鄉鎮市區現住人口數(按性別及年齡分),總人口數共有 22,615,307 人;樣本資料則是 2005NHIS 的完訪個案，共有 24,726 人。在資料釋出時，每個個案分別有三個權值，分別為：

1. *Wt_nation*：將全國人口的性別、年齡的比例調整到本調查之全國樣本數，此法是得到的是一個個案在台灣地區所代表的人數比例，調整回到原樣本數，用於推估台灣地區的資料時使用，所有個案的權值加總等於完訪樣本數(24,726 人)，而性別年齡之分布與台灣地區相同，其計算式為：

$$Wt\_nation_i = \frac{N_i}{n_i} \times \frac{n}{N}$$

2. *Wt_pop*：與前者方法同，不同處是加權之後得到的是全國人口數，也就是一個個案在台灣地區代表多少人，用於推估台灣地區的資料時使用，所有個案的權值加總等於全國總人口數(22,615,307 人)，其計算式為：

$$Wt\_pop_i = Wt\_nation_i \times \frac{N}{n}$$

3. *Wt_county*：縣市樣本數權值，用於推估各縣市的資料時使用，該縣市內所有個案的權值加總等於該縣市的完訪樣本數，其計算式為：

$$Wt\_county_i = \frac{N_i}{n_i} \times \frac{n_c}{N_c}$$

$N_c$：該縣市母體以性別、年齡分組總人口數

$N_c$：該縣市完訪樣本性別、年齡組總個案數

在資料釋出時，這些權數都會一併附上，使用者必須根據自己的需要使用不同的

權數。

各統計軟體均有權值的使用方法，任何軟體所加權出來的平均數都相同，而標準誤(standard error)的計算視軟體的不同而有所不同，美國所執行之大型調查，大多提供相關統計軟體使用說明，STATA 和 SUDAAN 常被使用，SAS 也有和調查有關的語法。加權方面，在 STATA 的加權分為 pweight 和 aweight，pweight 加權到母群體人數，aweight 自動將人口換成比例，也就是加總為 1，一般不建議使用，STATA 的 survey commands 也不允許此權數。下列網站有說明：http://www.cpc.unc.edu/projects/usda/help/SUDAAN_STATA.html#Top，SUDAAN 則是在程式語法中宣告使用哪組權數，SAS 亦同，權數都事先算好，在資料釋出時供使用者使用，一些大型調查會將各種權數的使用方法詳細描述，使用者則根據需求及分析內容選擇適當的權數，例如美國的營養調查（NHANES）只有約 60%接受面訪的人完成體檢，因此分析面訪資料所用權數不同於分析體檢資料的權數，在他們的網站中建議如果合併資料，以筆數少的為主選擇權數(http://www.cdc.gov/nchs/data/nhanes/nhanes_03_04/nhanes_analytic_guidelines_dec_2005.pdf)。

標準誤的估計在 STATA(SVYMEAN, SVYTAB 等)或 SUDAAN(DESCRIPT, CROSSTAB 等)均提供標準誤的估計值，STATA 和 SUDAAN 用的是相同的計算公式，因此得到的結果相同，而 SAS 中和調查相關的程序所用的公式稍有不同。要得到正確的標準誤，必須考慮抽樣設計。本報告以 SUDAAN (SUrvey DAta ANalysis)做說明，SUDAAN 顧名思義是專門分析抽樣調查資料所用，此軟體可

以在不同的抽樣設計下估計標準誤，其程式語言和 SAS 類似，有不同的程序(PROC)，和 SAS 不同的地方是要宣告抽樣設計，主要看第一抽出單位(PSU)抽出方式，本調查的 PSU 是 without replacement，也就是說鄉鎮市區（或鄰）抽出來以後就不放回去，所以本調查 DESIGN=WOR，SUDAAN 的優點是在不同階段的抽樣可以不同，以__ZERO__宣告此變項為分層變項，不計算每個層的抽出機率，以__MINUS1__來宣告此階段後全是 with replacement (WR)。SUDAAN 可以在 SAS 環境下執行，建議先以 SAS 處理好資料後，再執行 SUDAAN 指令，所需做的資料處理含

（1） SUDAAN 只接受數值型資料，即使是類別資料的分類變項也需是數值型，例如男、女必須用 1、2 來代表，字元型變項不能在 SUDAAN 下執行，而且所有變項必須從 1 開始，除了 logistic 的反應變項（Y）必須是 0、1 外；

（2） 用以分類之數值變項不能中斷，例如以 1、2、3、4 來代表縣市，不能從 1、2 跳到 10、11；

（3） 必須先按縣市、PSU 排序，也就是照分層順序先排序；

（4） 每個觀察值必須有相對應的分層變項，例如每個個案必須屬於某個層內的 PSU；

（5） 每個 PSU 內必須有多於一個觀測值以供估計標準誤。

用於本調查的 SUDAAN 程序為（假設已處理完成的資料檔是 t，分層識別碼分別為 county1、strata1、psu1，wt_p1 為權數，psu_n 為該分層內的總數）

```
proc sort data=t;
    by county1 strata1 psu1;
    /* county1：縣市改為數值型變數,
    strata1：縣市內分層,
    psu1：層內的第一抽出單位 */;
run;

proc descript data=t filetype=sas design=wor;
    /* psu 的抽出是 without replacement */;
    nest county1 strata1 psu1;
    totcnt _zero_  psu_n  _minus1_;
    /* totcnt 和 nest 是一一對應的，以_zero_宣告縣市(county1)是分層變項，縣市內之層(strata1)有不同數量的 psu，psu_n 是宣告其數量，_minus1_是宣告 psu 內每個個案抽出機率很小，此處以_minus1_宣告其抽出機率小於 1/20，因此不用宣告每個 psu 內之人數 */；
    weight wt_p1;    /*權數 */;
run;
```

以下是用 2005 年 12-64 歲的問卷資料為實例（假設已處理完成的資料檔是 d2005b1，分層識別碼分別為 county、strata、psu_id，wt_n1 為權數，psu 為該分層內 psu 的總數），利用 Sudaan 與一般的計算所做的比較 (請注意, 如果您要計算的變數有遺漏值, 請將該觀測值的加權值改為 0、負值或是 SAS 認可的 missing value , 不要直接刪除, 否則 Sudann 可能無法計算)：

連續變數:

Sudaan:

```
proc descript data=d2005b1 filetype=sas design=wor means;
    nest county strata psu_id;
    totcnt _zero_  psu  _minus1_;
    class sex county sex_county;
    var age1;
    weight wt_n1;
run;
```

一般的計算(加權):

```
proc means data=d2005b1 noprint;
    var age1;
    class sex county;
    weight wt_n1;
    output out=age1_w mean=age1_mean stderr=age1_se;
run;
```

一般的計算(不加權):

```
proc means data=d2005b1 noprint;
    var age1;
    class sex county;
    output out=age1_now mean=age1_mean stderr=age1_se;
run;
```

兩種算法的結果如下(部分結果):

年齡 (Mean± S.E.)

| 性別, 鄉鎮市區 | Sudaan | SAS 加權 | SAS 不加權 |
|---|---|---|---|
| Sex = 1 | 35.88 ± 0.15 | 35.88 ± 0.14 | 35.91 ± 0.15 |
| Sex = 2 | 36.27 ± 0.16 | 36.27 ± 0.15 | 36.60 ± 0.15 |
| County = 1 | 37.22 ± 0.38 | 37.22 ± 0.39 | 37.70 ± 0.40 |
| County = 2 | 36.27 ± 0.57 | 36.27 ± 0.57 | 34.83 ± 0.58 |
| Sex = 1, County = 1 | 36.74 ± 0.54 | 36.74 ± 0.55 | 37.41 ± 0.58 |
| Sex = 1, County = 2 | 35.99 ± 0.83 | 35.99 ± 0.80 | 33.41 ± 0.81 |
| Sex = 2, County = 1 | 37.66 ± 0.53 | 37.66 ± 0.54 | 37.98 ± 0.55 |
| Sex = 2, County = 2 | 36.56 ± 0.79 | 36.56 ± 0.81 | 36.30 ± 0.84 |

類別變數: (此例以變數 b7_n 為計算對象, 以性別來分類, 其他變數及分類請自行類推)

Sudaan:

```
proc crosstab data=d2005b1 filetype=sas design=wor;
    nest county strata psu_id;
    totcnt _zero_   psu   _minus1_;
    subgroup b7_n sex;
    level 3 2;
    tables b7_n*sex;
    weight wt_n1;
    test chisq;
    print chisq chisqdf chisqp;
run;
```

一般的計算(加權):

```
proc freq data=d2005b1 noprint;
    tables b7_n*sex/chisq;
    output out=b7 chisq;
    weight wt_n1;
run;
```

一般的計算(不加權):

```
proc freq data=d2005b1 noprint;
    tables b7_n*sex/chisq;
    output out=b7 chisq;
run;
```

兩種算法的結果如下: (The chi-square test statistic and its p-value.)

氣喘(b7_n, 共三個選項, 0 表示沒有, 1 表示有, 2 表示不知道)

| Sudaan | SAS 加權 | SAS 不加權 |
|---|---|---|
| 0.12 (0.9413) | 0.14 (0.9324) | 1.50 (0.4713) |

曾經中風(腦溢血或腦血栓)(b6, 共三個選項, 0 表示沒有, 1 表示有, 2 表示不知道)

| Sudaan | SAS 加權 | SAS 不加權 |
|---|---|---|
| 2.66 (0.2798) | 3.2537 (0.1965) | 5.0625 (0.0796) |

是否曾經吸過菸(e5, 共四個選項, 0 表示沒有吸過, 1 表示僅嘗試吸過幾次而已, 2 表示有吸過, 從以前到現在**沒有**吸超過 5 包 (100 支) 菸, 3 表示有吸過, 從以前到現在**有**吸超過 5 包 (100 支) 菸, 此例以 county 當作類別變數

| Sudaan | SAS 加權 | SAS 不加權 |
|---|---|---|
| 175.28 (0.0021) | 165.2471 (1.785951E-10) | 171.1718 (2.792645E-11) |

是否曾經嚼食過檳榔(e6, 共四個選項, 0 表示否, 從未嚼過, 1 表示是, 從過去到現在只嚼過 1-2 次, 2 表示是, 以前嚼, 現在不嚼 (最近六個月沒有嚼), 3 表示是, 現在嚼 (包括最近六個月曾嚼過)), 此例以 county 當作類別變數

| Sudaan | SAS 加權 | SAS 不加權 |
|---|---|---|
| 403.14 (0.0000) | 559.6472 (2.551535E-79) | 653.5603 (1.451394E-97) |