設問調査分析의 電算處理方法

—SPSS를 中心으로—

權 五 龍
(KIET 電算室)

◀目 次▶

Ⅰ. 序 論

社會科學分野의 研究分析은 一般的으로 設問調査에 의하여 社會現象을 파악하는 경우가 많이 發生한다. 왜냐하면 社會現象에 대한 資料가 不足하거나 혹은 전혀 없는 상태이므로, 設問調査方法이 唯一하게 資料를 얻을 수 있는 手段이기 때문이다. 設問調査時 留意할 點은 다음과 같다.

1) 設問의 設定方法
2) 標本選定 및 크기決定
3) 調査方法
4) 統計處理方法

등이다.

위의 留意點은 모두 重要하게 다루어져야할 事項들이다. 그러나 本研究에서는 1), 2), 3)項은 正常的으로 遂行되었다는 假定下에서, 調査된 統計資料를 어떠한 方法으로 分析하여 研究者의 目的에 적합하게 利用할 수 있는가를 提示하여서 設問調査에 의한 研究分析에 도움이 되었으면 한다. 보통 앙케이트 方法에 의한 統計分析은 設問에 대한 一般事項 즉 性別, 年齡別, 宗教別, 所得別로 各問項에 대한 頻度 즉 度數와 相對度數(百分率)를 計算하여 利用하는 정도의 分析에 그친다. 이런 정도의 分析은 社會現象의 單純統計分析에 지나지 않는다. 그러므로 統計理論을 보다 많이 利用하여 深層分析하는 方法을 說明하고자 한다.

이런 方法中 ① 適合度分析 ② 平均値의 差異分析 ③ 回歸分析 ④ 相關分析 등을 利用하여 統計的 檢定을 한 후 研究者가 利用할 수 있도록 설명한 후,이를 컴퓨터의 Software인 SPSS[1] Package를

註1) SPSS는 1965년 Standford大學 政治學 研究所에서 만든 統計處理 Package이다.
(Statistical package for the social sciences)

利用하여 쉽게 處理하여 分析할 수 있도록 例題에 의하여 설명하고자 한다.

II. 処理別 差異分析

設問調査에 의한 資料分析에서 處理別이란 주로 一般事項 즉 性別, 宗教別, 年齡別 등을 말한다. 즉 한 處理에서 標本을 여러 分類로 분리할 수 있어야 한다. 宗教別 側面을 分析할 경우는 儒教, 佛教, 基督教, 天主教 등으로 標本資料를 分類한 후, 이에 준하여 各 問項別로 度數를 계산하여, 宗教的 側面에서 各問項別 有意的인 差異를 分析하여, 研究資料에 利用할 수 있다.

이와같은 경우에 必要한 統計處理는 x^2-統計分析 方法이다.

x^2-統計量이란 一般的으로 理論度數*ei라 하면

$$x^2 = \sum_{i=1}^{k} \frac{(fi - ei)^2}{ei} \cdots\cdots(1)$$

로 계산된 統計量은 自由度(k-1)의 x^2-分布를 한다.

式(1)로 定義된 x^2의 값을 檢定統計量으로 使用하는 假說檢定을 x^2-檢定法이라 하며 檢定은 다음의 順序에 의한다.

① 檢定假說을 세운다.

*) Ho : 處理別로 問項應答은 差異가 없다.

*) Ha : 處理別로 問項應答은 差異가 있다.

② 理論度數를 計算한다.

③ x^2-統計量을 計算한다.

④ *)有意水準 α에 의한 x^2-分布表에서 判定 $x^2\alpha$를 구한다.

⑤ x^2 統計量이 $x^2\alpha$값보다 크면 對立假說을 採擇한다. 즉 다시말하면 處理別로 問項에 대한 調査分析은 危險率 α 水準에서 有意的인 差異가 있다고 말할 수 있다.

x^2 統計量이 $x^2\alpha$값보다 작으면 歸無假說이 採擇된다. 즉 α의 有意水準에서 差異가 없다.

이상의 x^2-統計量에 의한 差異의 分析은 SPSS에서 CROSSTABS 命令을 使用하는데 그 形式은 다음과 같다.

*)	1 COL	16 COL
	CROSSTABS	TABLES = 變數名
		BY 變數名
	STATISTICS	ALL

위 命令에서 TABLES =에 使用하는 變數名은 二重分類表에서 行測에 表示하는 變數는 BY앞에 使用하고, 例測에 表示되는 變數는 BY뒤에 나온다.

STATISTICS命令으로 計算되는 統計量은 다음 表 I 과 같다.

위에서 살펴본 處理別 差異分析의 理論과 SPSS 命令을 使用할 實例로 살펴보자.

---

* ei : $eij = \frac{fifj}{n}$ i : 행, j = 列, n : 총수, fi : 관측도수

*) Ho : 귀무가설 - 母數의 가설에서 檢定하려는 가설
Ha : 대립가설 - 귀무가설에 대립되는 가설

*) 유의水準 : 檢定統計量이 기각역에 포함되는 確率

*) SPSS Page 218 - 245

〈表 Ⅰ〉 CROSSTABS의 統計

1) x^2 … 統計量
2) Phi – 統計量
3) Kendall – 統計量
4) 分割係數
5) Gamma 係數
6) Lamda 係數
7) ETA 係數

어떤 標本集團에서 2,429名에 대하여 人種別로 所得水準을 調査하였다.

處理別 側面에서 人種을 白人(White), 非白人(Nowhite)으로 分類하고, 設問으로 所得水準을 4등급으로 나누어 調査한 것을 電算處理한 結果는 表Ⅱ와 같다. 이런 경우 人種別로 所得水準의 分布는 差異가 存在하는지를 分析하여 보자.

먼저 SPSS의 命令語를 작성하면 다음과 같다.

```
CROSSTABS      TABLES = INCOME
                        BY  RACE

STATISTICS     ALL
```

위의 命令으로 印刷된 表Ⅱ의 結果로써 統計的 分析을 하면 아래와 같다.

二重分類表Ⅱ를 處理別 즉 人種別로 x^2 – 檢定을 하여 人種別로 所得水準에 有意的인 差가 있는지를 分析하면 다음과 같다.

1) 假說을 設定한다.

歸無假說(Ho) : 白人과 非白人의 所得水準의 差는 없다.

〈表 Ⅱ〉

CROSSTABS AND BREAKDOWN 03/11/74 PAGE 2
FIGURE 16.7
FILE ELCT72 (CREATION DATE = 03/08/74) SAMPLE FROM CHANGE STUDY
SUBFILE ELEC72

* C R O S S T A B U L A T I O N O F *
INCOME RECODED FAMILY INCOME OF RESPONDENT BY RACE RACE OF RESPONDENT
* PAGE 1 OF 1

INCOME	COUNT / COL PCT	RACE: WHITE (1)	NONWHITE (2)	ROW TOTAL
LESS THAN $4000	1	396	98	494
		18.4	36.1	20.3
$4000–$7999	2	526	70	596
		24.4	25.8	24.5
$8000–$12499	3	612	64	676
		28.4	23.4	27.8
$12500 AND OVER	4	624	40	664
		28.9	14.7	27.3
	9	79M	10M	89M
		0.0	0.0	0.0
COLUMN TOTAL		2158 88.8	272 11.2	2429 100.0

RAW CHI SQUARE = 57.41589 WITH 3 DEGREES OF FREEDOM. SIGNIFICANCE = 0.0000
CRAMER'S V = 0.15373
CONTINGENCY COEFFICIENT = 0.15195
LAMDA (ASYMMETRIC) = 0.02645 WITH INCOME DEPENDENT. = 0.0 WITH RACE DEPENDENT.
LAMDA (SYMMETRIC) = 0.02290
UNCERTAINTY COEFFICIENT (ASYMMETRIC) = 0.00814 WITH INCOME DEPENDENT. = 0.03206 WITH RACE DEPENDENT.
UNCERTAINTY COEFFICIENT (SYMMETRIC) = 0.01299
KENDALL'S TAU B = -0.13306. SIGNIFICANCE = 0.0
KENDALL'S TAU C = -0.10245. SIGNIFICANCE = 0.0000
GAMMA = -0.33859
SOMERS' D (ASYMMETRIC) = -0.25800 WITH INCOME DEPENDENT. = -0.06862 WITH RACE DEPENDENT.
SOMERS' D (SYMMETRIC) = -0.10841
ETA = 0.02179 WITH INCOME DEPENDENT. = 0.02363 WITH RACE DEPENDENT.

NUMBER OF MISSING OBSERVATIONS = 89

對立假說(Ha) : 白人과 非白人의 所得水準의 差는 있다.

2) 有意水準을 1%로 定한다.

自由度 3의 判定x^2-統計量을 [1]x^2-分布表에서 구한다.

$$x^2_{0.01}(3) = 11.34$$

3) 計算된 x^2-統計量과 判定統計量을 比較하여 判定한다.

$x^2 = 57.415 > x^2_{0.01}(3) = 13.44$ 이므로 對立假說이 採擇된다.

즉, 白人과 非白人간의 所得水準은 1%의 有意水準에서 差異가 있다는 結論을 내릴 수 있다.

III. 処理別 平均差異分析

處理別 平均差異分析이란 標本에 한 變數의 處理別로 [1]算術平均을 計算했을 경우, 各 處理別 平均은 相異하다. 이와같은 平均의 有意的 差異有無를 檢定하는 것을 말한다.

一般的으로 處理區分이 2種인 경우는 t-分布의 標準單位를 使用한다. 그러나 2種이상의 處理區分이면 分散分析方法을 利用한다.

이 分析方法은 處理對象의 分散을 여러가지 要因으로 分解하여 變化의 要因을 알아내는 方法이다.

3개이상의 處理 平均値에서 各 母平均의 有意的 差異를 찾는 것은 다음과 같이 整理한다.

Xij는 第i번째 처리조의 j번째 觀察値이며 i번째 처리조의 크기는 ni, 그 標本의 平均을 $\bar{X}i$, 標本 全體 크기를 [1]N, 전체의 平均을 $\bar{X}$라 하면

$$\sum_{i=1}^{k}\sum_{j=1}^{ni}(Xi-\bar{X})^2 = \sum_{i=1}^{k}\sum_{j=1}^{ni}(Xij-\bar{X}i)^2 + \sum_{i=1}^{k} ni\,(\bar{X}i-\bar{X})^2 \quad \cdots\cdots(3-1)$$

等式이 성립된다.

(3-1)式 左邊을 V, 右邊을 V_1, V_2로 표시하면

$$V = V_1 + V_2$$ 가 된다.

V는 全體의 變動을 표시하고, V_1은 處理內의 관찰치의 變動으로, 이를 處理內變動이라하고, V_2는 處理變動値 사이의 變動이므로 이를 處理間變動이라 한다.

全變動中에서 處理內變動은 假說에 關係없이 偶然變動을 한다고 한다.

이에 대하여 處理間變動은 假說이 眞이면 偶然變動을 표시하나, 假說이 否이면 處理間의 差異에 따른 平均値의 有意的 差異라 할 수 있다. 그러므로 處理間變動이 處理內變動보다 크면 處理間變動은 有意的이며 따라서 處理平均値의 差는 有意로 된다.

여기서 使用되는 檢定은 處理內變動과 處理間變動의 比率인 F-統計量을 利用한다. 有意水準 α라 하면 F를 檢定統計量으로 하여

$F > F\alpha$이면 檢定假說은 棄却되고 對

註 1) 統計學册의 附錄參照

1) 算術平均 $= \dfrac{\sum X}{N}$ X:變數, N:도수

1) $N = \sum Ni$ 2) $\bar{x} = \dfrac{\sum x - 1}{N}$

〈表 Ⅲ〉 分 散 分 析 表

要 因	變 動	自由度	平均平方和	F
處 理 間	$V_2 = \Sigma n_i(\bar{X}i - Xj)$	K－1	V_2 ／K－1	V_2 ／V_1
處 理 內	$V_1 = \Sigma I(Xij - \bar{X}i)^2$	N－k	V_1 ／N－k	
全 體	$V = \Sigma I(Xij - \bar{X})^2$	N－1		

〈表 A〉

1） 분산분석
2） x^2 －통계량
3） 분할계수
4） Kendall 계수
5） Gamma 계수

立假說이 採擇된다.

이상을 分散分析表로 정리하면 表Ⅲ과 같다.

이상의 處理別 平均値 差異分析은 SPSS에서 *) BREAKDOWN命令을 使用하는데 그 形式은 다음과 같다.

```
1col                16col
BREAKDOWN   VARIABLES = 變數
                      BY  處理變數
STATISTICS  ALL
```

위 命令의 Variables에 使用하는 變數名은 處理變數는 BY뒤에, 分析의 對應變數는 BY 앞에 나온다. Statistics命令으로 計算되는 統計量은 表A와 같다.

이상에서 살펴본 이론과 컴퓨터命令을 使用하는 方法을 例題로 설명한다.

어느 지역의 지원위원회 위원들의 그지역 거주연한을 조사하였다. 이 조사에 의하면 지역위원은 거주연한이 많을수록 위원이 될 가능성이 높은 지를 분석하여 보자.

지역위원회(NMEN)와 거주연한(LRES)을 變數 및 處理로 하였을 경우 SPSS 命令은

```
1 Col               16 Col
BREAKDOWN   VARIABLES = NMEN
                      BY  LRES
STATISTICS  ALL
```

로 쓸 수 있다.

이에 따른 結果表는 表Ⅲ－1과 같으며,

〈表 Ⅲ－1〉

```
CREATE SYSTEM FILE AND BREAKDOWN TABLE                          03/30/74      PAGE    3

FILE   DEMODATA (CREATION DATE = 03/30/74)   DEMONSTRATION DATA FROM A U. S. STUDY

- - - - - - - - - - - - - - - - - - - - DESCRIPTION OF SUBPOPULATIONS - - - - - - - - - - - - - - - - - -
CRITERION VARIABLE   NMEN      NUMBER OF ORGANIZATIONAL MEMBERSHIPS
    BROKEN DOWN BY   LRES      LENGTH OF RESIDENCE IN COMMUNITY
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

VARIABLE                    CODE    VALUE LABEL          SUM        MEAN      STD DEV     VARIANCE        N

FOR ENTIRE POPULATION                                  121.0000     0.5550      0.8742      0.7642    (  218)

LRES                          1.    < 3 YRS             9.000      0.180       0.482       0.232     (   50)
LRES                          2.    4-10 YRS           14.000      0.519       0.700       0.490     (   27)
LRES                          3.    > 10 YRS           98.000      0.695       0.971       0.942     (  141)

  TOTAL CASES =     220
MISSING CASES =       2 OR    0.9 PCT.
```

*) SPSS의 p 249－269.

〈表 Ⅲ-2〉

CREATE SYSTEM FILE AND BREAKDOWN TABLE 03/30/74 PAGE 4

CRITERION VARIABLE NMEM

ANALYSIS OF VARIANCE

VARIABLE	CODE	VALUE LABEL	SUM	MEAN	STD DEV	SUM OF SQ	N
LRES	1.	< 3 YRS	9.000	0.180	0.482	11.380	(50)
LRES	2.	4~10 YRS	14.000	0.519	0.700	12.741	(27)
LRES	3.	> 10 YRS	98.000	0.695	0.971	131.887	(141)
		TOTAL	121.000	0.555	0.874	156.007	(218)

ANOVA TABLE

	SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARE
BETWEEN GROUPS	9.8322	(2)	4.9161
WITHIN GROUPS	156.0073	(215)	0.7256
TOTAL	165.8395	(217)	

F = 6.7751 ETA SQRD = 0.0593

TEST OF LINEARITY

	SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARE
REGRESSION	9.6822.	(1)	9.6822
DEV FROM LINEARITY	0.1500	(1)	0.1500

F = 0.2067 R = 0.2416 R SQUARED = 0.0584

이를 분석하면 다음과 같다.

表 Ⅲ-1.은 주거연한을 處理區分으로 한 지역위원회 구성평점을 계산한 것이다. 즉 연한이 3년미만은 0.180, 4년이상 10년 미만은 0.519, 10년이상은 0.695점의 平均點을 얻었을 경우, 3개의 處理別 平均差異는 有意的인지를 檢定하기 위하여 表 Ⅲ-2의 分散分析表를 分析하여 보자.

表 Ⅲ-2의 處理內變動과 處理間變動의 比率은 F-統計量은 6.7751로 계산되었다. 유의수준 1%에서 處理別 平均差를 F檢定할 自由度(2,215)의 $F\alpha$는 4.84이다.

$F > F\alpha$이므로 歸無假說은 기각되고 대립가설이 採擇된다. 다시말하면 지역위원회 구성에 대한 평점은 그지역 주거연한이 많을수록 높은 평점을 받을 수 있다.

Ⅳ. 回歸分析

回歸分析(Regression Analysis)은 2變數가 x, y중 하나를 獨立變數(原因變數)로 하여, 이 獨立變數를 변화시킬때 생기는 效果를 다른 變數 즉 從屬變數의 變化에서 分析하는 方法이다. 그러므로 回歸分析은 獨立變數를 自由로이 管理할 수있는 것을 前提로 한다. 이 때문에 이 分析方法은 社會現象의 分析에 많이 使用한다. 社會現象은 獨立變數를 自由로이 管理할 實驗을 하는 것은 困難하므로 近似的 또는 假定的인 管理로서 社會現象을 예측관리하여 보다 이상적인 現象을 추구할 수

있는 수단을 제시할 수 있다.

○ 回歸分析의 目的

回歸分析의 目的은 2가지가 있다.

1) 獨立變數의 變數値가 주어지면, 이것을 利用하여 從屬變數를 推定한다.

2) 從屬變數를 獨立變數의 函數로서 理論的 關係를 樹立하는 것이다. 이것은 分析對象의 理論的 模型으로서 回歸關係를 樹立하여, 從屬變數가 獨立變數에 얼마나 영향을 받는가를 測定하는 데에 主目的이 있다.

○ 回歸直線

回歸分析의 目的은 두변수간의 理論的 函數關係를 樹立하는 것이다. 그러면 從屬變數(y)와 獨立變數(x)의 函數式은 다음 函數式으로 표시할 수 있다.

$$y = f(x) \cdots\cdots (4-1)$$

式(4-1)의 가장 簡單한 것은 一次式인 直線의 方程式 $y = a + bx$이다.

두變數間의 관계를 表現하는 方程式을 回歸方程式이라 하고, 이 方程式에서 그려지는 線을 回歸線이라 한다.

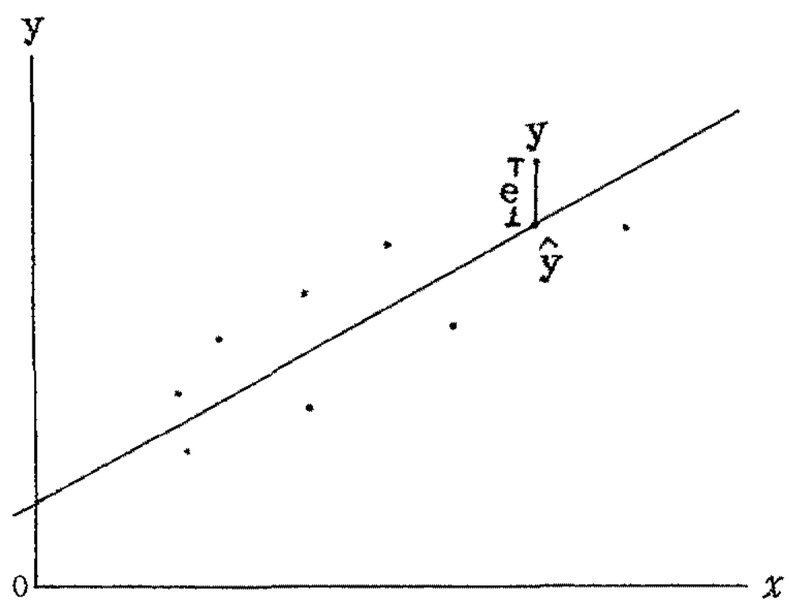
回歸線은 獨立變數의 一次式으로만 表示된다. 2次式이상의 獨立變數는 回歸分析이 不可能하다. 그러므로 回歸線은 直線으로 表示하므로 回歸直線이라 한다.

回歸直線 $y = a + bx$에서 a, b는 母數이다. 回歸方程式에서는 母數 a, b를 回歸母數라 하고, 方向係數 b를 回歸係數(Regression Coefficient)라 한다.

回歸分析에서는 母數 a, b의 값이 구해지면 獨立變數 x의 값이 주어질 경우 從屬變數 y의 값을 推定할 수 있다.

○ 最小自乘法

回歸線의 適合直線을 구하는 方法으로 가장 많이 使用되는 것이 最小自乘法이다. 最小自乘法은 適合된 直線과 觀察値와의 y축에 따라 종으로 測定한 距離 즉 偏差 e의 2乘合을 最小로 되게하는 直線을 구하는 방법이다.



$\hat{y} = a + bx$ (回歸直線)

$y = a + bx + e$

y_i ; 觀測値

$\hat{y}_i$; 適正値

e ; 편차

推定할 回歸線의 方程式을

$$\hat{y} = a + bx \cdots\cdots (4-2)$$

라 하면, 관측치들을 直線의 方程式으로 表示하면

$$y = a + bx + e \cdots\cdots (4-3)$$

와 같이 된다.

관측치와 回歸直線과의 偏差를 e라 하면

$e = y - \hat{y}$ 로 된다.

그러므로 偏差의 二乘合인

$$\sum_{i=1}^{n} e^2 = \sum (y_i - \hat{y}_i)^2$$

의 값을 最小로 하는 것이다.

$\hat{y} = a + bx_i$ 이므로

$$E = \sum_{i=1}^{n} (y_i - a - bx_i) \quad \cdots\cdots (4-4)$$

이다. 이 E를 a, b의 函數로 생각하여 E가 最小로 되는 것은 E를 각각 a, b로 微分한 偏導函數가 0이 되는 때이다.

$$\frac{\partial E}{\partial a} = \sum_{i=1}^{n} = (y_i - a - bx_i)(-1) = 0$$

$$\frac{\partial E}{\partial a} = \sum_{i=1}^{n} = (y_i - a - bx_i)(-x_i) = 0$$

이 두식을 整理하여 聯立一次方程式을 풀면

$$a = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{N\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$b = \frac{N\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{N\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

이 된다.

○ 回歸係數의 檢定

回歸係數의 推定方法으로서 偏差의 二乘合 $\sum_{i=1}^{n} e^2$ 을 最小로 하는 最小自乘法을 使用하였다. 그러므로 推定値의 標準誤差는 偏差e가 平均이 0이고 標準偏差가 δ인 正規分布를 한다고 가정한다.

回歸直線에서 平均平方偏差 $S_{y.x}^2$을 e의 推定値를 利用하면 다음과 같이 定義된다.

$$S_{y.x}^2 = \sum e^2 / n-2$$

이것이 回歸線의 分散 δ^2 의 不偏推定値로 된다.

따라서 回歸線의 推定値의 標準偏差 $S_{y.x}$는 다음과 같다.

$$S_{y.s} = \sqrt{\sum e^2 yx \ / \ n-2}$$

위의 $S_{y.x}$는 推定된 回歸線이 주어진 資料에 적합한가를 檢定하는데 使用된다. 回歸係數 b는 母回歸係數 β의 不偏推定値로서 平均이 β, 標準偏差가 $\delta/\sqrt{\sum x^2}$ 인 正規分布를 한다. δ의 推定値 $S_{y.x}$를 사용하여

$$t = \frac{b-\beta}{S_{y.s}/\sqrt{x^2}} = \frac{(b-\beta)\sqrt{x^2}}{S_{y.s}}$$

로 놓으면 t는 自由度 n-2의 t-分布를 하게 된다.

그러므로 歸無假說 $\beta = 0$는

$$t = \frac{b\sqrt{x^2}}{S_{y.s}}$$

를 계산하여 t-分布값과 비교하여 檢定할 수 있다.

○ 回歸線의 分散分析

回歸計算에서 平方合 $\sum Y^2$ 은 세개의 성분으로 分解할 수 있다. 즉

*) $Y = \bar{Y} + \hat{y} + e \quad \cdots\cdots (4-5)$

(4-5)에 의해서 $\sum Y^2$ 은

$$\sum Y^2 = N\bar{Y}^2 + \sum Y^2 + \sum e^2$$

$$= \frac{(\sum Y)^2}{n} + \frac{(\sum xy)^2}{\sum x^2} + (\sum (y-\hat{y})^2 \cdots \quad \cdots\cdots (4-6)$$

---

*) $Y = \bar{Y} + bx + e \qquad bx = \hat{y}$

$= \bar{Y} + \hat{y} + e$

〈表 4 - 1〉

원 인	자 유 도	평 방 합	평 균 평 방 합	F
회 귀 b	K − 1	$(\sum xy)^2/\sum x^2$	$\frac{(\sum xy)^2/\sum x^2}{K-1}$	$((\sum xy)^2/\sum x^2)/K-1$
편차 e	N − K	$\sum e^2$	$\sum e^2/N-K$	$\sum e^2/N-K$
전 체	N − 1	$\sum Y^2$		

로 된다. $\sum Y^2$의 분할에 따라 自由度가 이에 맞게 分割된다.

(4 − 6)式에 따라 表4 − 1의 分散分析表를 만들 수 있다.

위의 分散分析表를 利用하여 回歸母數 a, b를 同時에 有意的 檢定이 가능하다. 즉 $F > F\alpha$이면 a, b는 有意水準 α에서 採擇되어 $y = a + bx$의 回歸直線은 유용하다.

이상의 回歸分析을 SPSS *) REGRESSION命令으로 處理할 수 있으며 그 形式은 다음과 같다.

```
1 Col            16 Col
REGRESSION       VARIABLES = 變數名들
                 REGRESSION = 從屬變數 WITH
                              獨立變數
```

위의 回歸分析理論과 SPSS命令을 例題로 설명하면 다음과 같다.

지역위원들(NMEN)을 교육수준(EDUC)과 소득수준(INCOME)으로한 回歸分析을 실시하면 우선 SPSS의 命令은

```
REGRESSION      VARIABLES = NMEN, EDUC,
                            INCOME/
                REGRESSION = NMEN WITH
                             EDUC, INCOME/
```

과 같으며, 이에 대한 結果는 表4 − 6 과 같다.

表4 − 6 에서 回歸直線은 $y = 0.10062 + 0.063$ EDUC + 0.02214 INCOME이 된다.

推定된 回歸直線의 係數를 t − 檢定하여야 한다. 먼저 EDUC의 t − 統計量은 $t = 0.063/0.043$으로 $t = 7.332$, INCOME의

〈表 4 − 6〉

```
EXAMPLE OF REGRESSION WITH BLOCKWISE SELECTION                         04/03/74       PAGE    5

FILE   DEMODATA (CREATION DATE = 04/03/74)   DEMONSTRATION DATA FROM A U. S. STUDY

* * * * * * * * * * * * * * * * * * * * * * * *  M U L T I P L E   R E G R E S S I O N  * * * * * * * * * * * * * *   VARIABLE LIST  1
                                                                                                                  REGRESSION LIST  1
DEPENDENT VARIABLE..    NMEM      NUMBER OF ORGANIZATIONAL MEMBERSHIPS

VARIABLE(S) ENTERED ON STEP NUMBER  1..    EDUC       LAST YEAR OF SCHOOL COMPLETED
                                           INCOME     ANNUAL FAMILY INCOME

MULTIPLE R           0.43229                 ANALYSIS OF VARIANCE    DF      SUM OF SQUARES      MEAN SQUARE          F
R SQUARE             0.18687                 REGRESSION               2.          28.83228         14.41614         23.09671
ADJUSTED R SQUARE    0.18235                 RESIDUAL               201.         125.45699          0.62416
STANDARD ERROR       0.79004

-------------- VARIABLES IN THE EQUATION ---------------          ------------ VARIABLES NOT IN THE EQUATION ------------

VARIABLE             B          BETA      STD ERROR B       F        VARIABLE      BETA IN     PARTIAL    TOLERANCE        F

EDUC             0.06303     0.09815      0.04335       2.114        RACE          0.02418     0.02639    0.96802        0.139
INCOME           0.02214     0.38942      0.00384      33.285        LRES          0.25895     0.28280    0.96477       17.335
(CONSTANT)       0.10062

* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
```

*) SPSS의 p.320 − 360

t-統計量은 $t = 5.81$이 된다. 5%의 有意水準에서 回歸係數를 檢定하면, INCOME은 유의적인 差가 있으나, EDUC는 有意的 差가 없다. 그러나 回歸線 自體를 分散分析表에 의한 F-檢定을 하면 $F = 23.0967 > F\alpha = 4.84$이므로 有意的 差가 統計的 檢定으로 인정된다. 그러므로 回歸係數의 t-檢定과 分散分析表에 의한 F-檢定의 結果 回歸線 $y = 0.1006 + 0.063\ x_1 + 0.02234$는 적정한 推定直線이 된다.

V. 相関分析

相關分析(Correlation Coefficient)은 二變數 x, y에 關聯하는 强度의 關係를 分析하는 것이다. 이런 두변수간에 맺고 있는 關係의 强度를 측정하는 尺度를 相關係數라 한다.

x, y 두변수의 相關係數 ρxy는 다음과 같이 定義한다.

$$\rho xy = \frac{\delta xy}{\delta x \delta y} = \frac{\sum (x-\overline{x})\ (y-\overline{y})}{\sqrt{\sum (x-\overline{x})^2}\sqrt{\sum (y-\overline{y})^2}}$$

相關係數는 $-1 \leq \rho xy \leq 1$의 값을 가지며 그 性質은 다음과 같다.

① X, Y가 完全한 正의 線型關係이면 ρxy는 1이 된다.

② X, Y가 전혀 相關關係가 없으면 ρxy는 0이 된다.

③ X, Y가 正의 相關關係가 있으면 ρxy는 양의 값을 가진다.

④ X, Y가 陰의 相關關係가 있으면 ρxy는 陰의 값을 가진다.

⑤ X, Y가 完全한 負의 相關關係가 있으면 ρxy는 -1의 값을 갖는다.

相關係數는 SPSS의 *)Pearson CORR

〈表 V-1〉

RUN PEARSON CORRELATIONS WITH SYSTEM FILE AND PUNCH CORR 02/17/74 PAGE 6

FILE COMSTUDY (CREATION DATE = 02/17/74) STUDY OF AMERICAN SMALL COMMUNITIES

- - - - - - - - - - - - - PEARSON CORRELATION COEFFICIENTS - - - - - - - - - - - - -

	MEDSCH	MEDFINC	PTGOHS	PTAGRI	PTMANU	CONST
MEDSCH	1.0000	0.7162	0.5383	-0.1945	0.0809	99.0000
	(0)	(64)	(64)	(64)	(64)	(64)
	S=0.001	S=0.001	S=0.001	S=0.062	S=0.263	S=*****
MEDFINC	0.7162	1.0000	0.5267	-0.1569	0.3957	99.0000
	(64)	(0)	(64)	(64)	(64)	(64)
	S=0.001	S=0.001	S=0.001	S=0.108	S=0.001	S=*****
PTGOHS	0.5383	0.5267	1.0000	-0.3151	0.2353	99.0000
	(64)	(64)	(0)	(64)	(64)	(64)
	S=0.001	S=0.001	S=0.001	S=0.006	S=0.031	S=*****
PTAGRI	-0.1945	-0.1569	-0.3151	1.0000	-0.2148	99.0000
	(64)	(64)	(64)	(0)	(64)	(64)
	S=0.062	S=0.108	S=0.006	S=0.001	S=0.044	S=*****
PTMANU	0.0809	0.3957	0.2353	-0.2148	1.0000	99.0000
	(64)	(64)	(64)	(64)	(0)	(64)
	S=0.263	S=0.001	S=0.031	S=0.044	S=0.001	S=*****
CONST	99.0000	99.0000	99.0000	99.0000	99.0000	1.0000
	(64)	(64)	(64)	(64)	(64)	(0)
	S=*****	S=*****	S=*****	S=*****	S=*****	S=0.001

(COEFFICIENT / (CASES) / SIGNIFICANCE) (A VALUE OF 99.0000 IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED)

*) SPSS의 p. 276-290

〈表 V - 2〉

SPEARMAN CORRELATIONS WITH SYSTEM FILE INPUT 02/17/74 PAGE

FILE ORGSTUDY (CREATION DATE = 02/17/74) STUDY OF ORGANIZATIONAL MEMBERSHIP AND ACTIVITY
SUBFILE NWJERSEY PENNSYLV

- - - - - - - - - - - - - - - SPEARMAN CORRELATION COEFFICIENTS - - - - - - - - - - - - -

VARIABLE PAIR		VARIABLE PAIR		VARIABLE PAIR		VARIABLE PAIR		VARIABLE PAIR		VARIABLE PAIR	
RESDYTH	0.0847	RESDYTH	0.0202	RESDYTH	0.2022	RESDYTH	0.0119	INCOME	0.1788	INCOME	0.4432
WITH	N(250)	WITH	N(250)	WITH	N(250)	WITH	N(250)	WITH	N(250)	WITH	N(250)
INCOME	SIG .091	NACT	SIG .375	EDRESPON	SIG .001	OCLEVRES	SIG .426	NACT	SIG .002	EDRESPON	SIG .001
INCOME	0.2671	NACT	0.2126	NACT	0.1100	EDRESPON	0.2455				
WITH	N(250)	WITH	N(250)	WITH	N(250)	WITH	N(250)				
OCLEVRES	SIG .001	EDRESPON	SIG .001	OCLEVRES	SIG .041	OCLEVRES	SIG .001				

A VALUE OF 99.0000 IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED.

命令을 使用하여 구할 수 있으며 그 形式은 다음과 같다.

```
1Col                16 Col
PEARSON CORR  변수X With 변수Y
```

Pearson Corr 命令으로 X, Y 두 변수의 상관계수를 계산하며, 그 결과표는 表5-1과 같다.

表5-1에서 변수 MEDFINC와 MEDSCH의 相關係數는 $\rho = 0.7162$이다. 相關係數 ρ는 유의수준 *(S = 0.001) 0.1%에서 유의적이다. 그러므로 MEDFINC와 MEDSCH의 두 변수간의 상관의 정도는 0.7162라는 높은 상관을 갖고 있다는 結論을 내릴 수 있다.

○ 順位相關係數

두 變數 X, Y의 資料가 順位 즉 序列에 의한 경우의 觀測値일 경우 相關係數의 測定은 順位相關係數에 의거 分析할 수 있다.

順位相關係數 ρ는 $0 \leq P \leq 1$의 값을 가지며 다음과 같이 계산된다.

$$\rho = 1 - \frac{\sum d^2}{N(n^2-1)}$$

d : 두변수의 順位의 差, n : 표본의 수.

順位相關係數 $\rho = 1$이면 두변수의 順位는 높은 상관이 있으며, $\rho = 0$는 두변수의 順位는 전혀 무관하다고 말할 수 있다. 順位相關은 SPSS의 *)NONPAR CORR命令으로 계산되며 그 形式은 다음과 같다.

```
1Col                16 Col
NONPAR  CORR  변수X With 변수Y
```

NONPAR CORR로 順位相關係數를 계산하며 그 결과표는 表5-2와 같다.

表5-2에서 변수 INCOME과 RESDYTH의 순위상관계수 $\rho = 0.0847$이며, 이는 유의수준(S = 0.09) 5%下에서 유의적이 아니다.

그러므로 두변수의 順位相關은 전혀 없다는 결론을 내릴 수 있다.

즉 다시말하면 INCOME이 높을 수록

*) S : 有意水準(Significant level)

RESDYTH가 높다고는 말할 수 없다.

Ⅵ. 맺 는 말

設問調査에 의한 研究分析에 있어서 統計分析 處理方法은 가장 큰 問題點이다. 그러므로 이를 現代의 총아인 컴퓨터를 利用하여 統計處理하여 分析한다면 신빙성이 높은 統計資料를 分析·利用할 수 있다. 여기서는 設問調査分析에서 많이 使用하는 統計分析을 소개했다. 그러나 위의 統計分析이외에도 重要한 分析方法 즉 Factor Analysis, Canonical analysis, Guttman Scale, Discriminant Analysis, T-test 등이 있다. 이런 統計도 資料의 分析에 중요하므로 必要한 경우 SPSS册字를 參考하여 보다 더 면밀한 資料處理를 하면 신빙성 높은 統計分析을 할 수 있다.

〈參 考 文 獻〉

1) H. Nie Statistical Package for the Social Sciences. 1950
2) A.M Mood, Introduction to the Theory of Statistics. 1950.
3) 姜五全, 新版統計學 博英社. 1979
4) 金俊輔, 現代統計學 法文社. 1963
5) 白雲鵬, 統計學 博英社. 1979.