相關分析 : SPSS에서의 適用

金 潤 實
(情報資料室)

〈차 례〉

Ⅰ. 序 論

自然科學이든 社會科學이든 한 學問의 目的은 주어진 現狀을 要約 記述할 뿐만 아니라, 나아가서 그 現狀을 보다 정확하고 광범위하게 說明하고 豫言해 주는 데 있다. 이러한 기능을 다하기 위해서는 한 現狀에 관계되는 變因들을 우선 정확하고 妥當하게 測定해야 되지만 또한 중요한 것은 이 測定된 變因들 간의 關係를 정확하게 記述해야 될 것이다. 이러한 變因들 간의 關係를 정확하게 記述하기 위해서는 不明瞭하고 非論理性이 많은 日常의 言語的 表現보다는 數學이라는 보다 정확한 言語가 必要하게 된다. 오늘날 社會科學 分野에 · 관계되는 專門雜誌나 硏究報告書에는 統計的 分析結果가 그 量과 適用方法의 多樣性이 날로

증가해가고 있다. 특히 學問의 本質에 관하여는 지금까지 여러 角度에서 說明되어 왔지만 그 中에서도 計量的 接近方式 만큼 學問의 特質을 명확히 糾明한 것은 없다. 즉 現狀을 단순화시켜 추상화한 數學的 模型을 통해 現狀의 움직임을 파악함으로써 特定主題分野를 形成하는 제반 現狀 및 變數를 糾明해 주는 것이다. 이렇듯 計量的 研究方法을 한 學問分野를 研究해 나가는데 道具學問으로 사용하는 예는 圖書館學·情報學 分野에서도 흔히 사용되는 技法이다. 특히 圖書館學·情報學 分野에서 다루어지는 대부분의 要因은 數量化 또는 變量化가 가능하므로 計量的 研究方法의 適用性이 높다고 하겠다. 圖書館學, 情報學에 있어서 이러한 計量的 接近方式으로 學問的인 基盤을 구축한 것이 바로 計量書誌學(Bibliometrics)이다. 計量書誌學은 主題文獻의 數量學的 分析을 통해 圖書館 및 情報시스템의 合理的이고 效率的인 設計를 위해 응용될 수 있다. "Bibliometrics"라는 용어는 1969년 프리챠드(Pritchard)에 의해 처음 사용되었으며 "計量書誌學은 주제문헌을 통한 커뮤니케이션의 과정을 數量化 하고자 하는 學問"이라 定義하였다. 비록 計量書誌學的 概念이 1917년 코울(Cole)과 얼스(Eales)에 의해 최초로 研究되어졌지만 근본적으로 計量書誌學은 문자화된 커뮤니케이션 過程을 밝혀주고 여러 국면을 계산 分析하는 手段에 의해 그 特性과 原則過程을 개발해 주는 것이다. 計量書誌學 研究에서 사용되는 資料源이 무엇이든지간에 資料의 加工을 위해서는 많은 양의 데이터가 要求된다. 지금까지는 이러한 處理過程을 手作業을 통해 행하여져 왔었다. 그러나 手作業 過程이 몇몇 制限點을 갖고 있다는 사실은 그만두고라도 즉시 處理할 수 있는 파라미터(parameter)가 거의 없고 파라미터의 計數가 거의 가능하지 못했다. 최근에 들어와서 計量書誌學 研究에 컴퓨터가 그 處理過程으로 사용되고 있다. 리그비(Rigby)와 튜로우니(Thuronyi)가 그들의 研究인 "지구물리학 연속간행물의 질량분석연구"라는 논문에서 컴퓨터를 사용하였고 케슬러(Kessler)가 "物理學 文獻引用의 統計的 屬性"에 관한 研究에서 컴퓨터를 사용하기 시작하였다. 本稿에서는 圖書館學 情報學 分野에서 많이 사용되는 統計的 技法 중 하나인 相關分析(correlation analysis)을 社會科學分野 컴퓨터의 소프트 웨어(software)인 SPSS (Statistical Package for the Social Science)패키지의 적용방법을 개략적으로 살펴보기로 한다.

Ⅱ. 社會科學分野 데이터 分析패키지(SPSS)

SPSS는 1965년 스탠포드大學 政治學 硏究所에서 만든 統計 處理 패키지로 니(Norman H. Nie), 훌(C. Handlai Hull), 젠킨스(Jean G. Jenkins)에 의해 이루어진 것으로 社會科學 分野의 데이터 分析에 利用할 수 있다. 이 시스템은 IBM 360과 CDC6000컴퓨터에 사용될 수 있도록 變換시켰으며 수년에 걸쳐 다른 機種에도 接續이 가능하도록 變換되어 왔다. 社會科學 分野의 硏究方法中 設問紙法에 의해 統計를 引用한 統計 分析으로는

① 適合度 分析 : x^2 統計分析

② 平均値의 差異分析 ; T檢證

③ 回歸分析(regression analysis)

④ 相關分析(correlation analysis)

⑤ 要因分析(factor analysis)

등이 있으며 이들 分析 方法을 利用하여 統計的 檢定을 한후 硏究者가 利用할 수 있도록 說明하여 이를 SPSS패키지를 사용하여 硏究·分析에 使用할 수가 있다.

Ⅲ. 相 關 分 析

1. 相關分析의 歷史的 背景

相關分析에 使用되는 相關計數는 모두 $-1 \leqq \rho \leqq +1$의 값을 갖는데 이러한 相關係數를 計算하는 統計的 方法에는 여러가지가 있으며 그 기호도 r (피어슨 r), ρ (rho), λ (lamda), r (gamma), θ (theta), n (eta) 등이 있다. 變因間의 相互 關係性을 數學的으로 表現하고자 하는 相關係數의 문제는 옛부터 여러 學者들의 主要 관심사가 되어왔는데 이 문제를 본격적으로 연구하기 시작한 것은 영국의 갤톤(Francis Galton)이다. 그러나 오늘날 우리가 使用하고 있

는 相關分析 기법으로의 발전은 1896년 피어슨(karl Pearson)에 의해 이루어졌다. 두 變因 間의 相關程度를 나타내는 相關係數를 흔히 피어슨의 相關係數 또는 피어슨의 積率 相關係數(product moment correlation coefficient)라 한다. 피어슨 이후 스피어맨(Spearman)은 順位差 相關係數(rank-different correlation coefficient)計算方法을 고안했고 가트맨(Guttman), 굿맨과 크루스칼(Goodman & Kruskal), 윌콕슨(Wilcoxon), 자스펜(Jaspen), 맥네마(McNemar), 워어커(Walker), 레브(Lev), 피터스와 반보리스(Peters & Van Voorhis)등에 의해서도 여러 공식과 연구가 이루어져 왔다.

2. 相關分析의 槪念

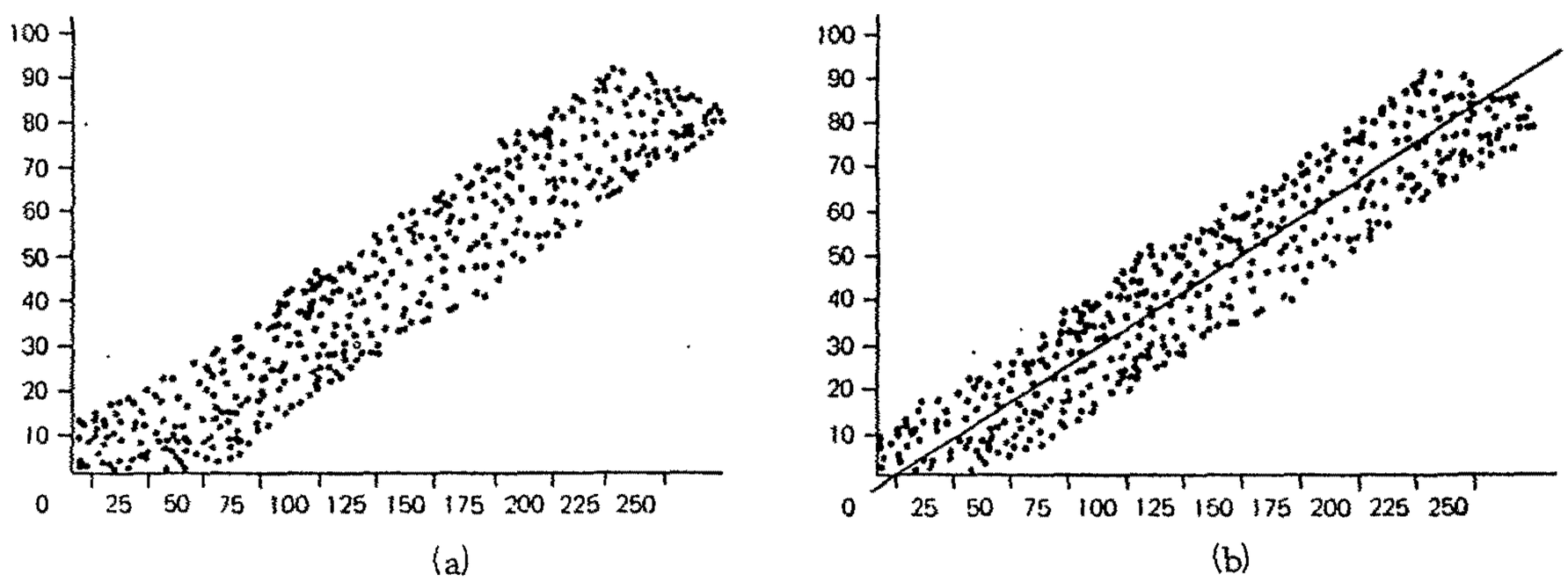
SPSS 시스템에서는 相關分析이 3개의 서브 프로그램인 PEARSON CORR, NONPAR CORR, SCATTERGRAM으로 구성되어 있다. PEARSON CORR은 序列 變因間의 相關程度를 나타내는 피어슨의 積率相關係數(product moment correlation coefficient : PMR)를 계산하는 프로그램이다. 스피어맨과 캔달의 順位相關係數는 序列 變因의 측정에 적합하게 구성되어 있어 NONPAR CORR 프로그램에 적절히 사용할 수 있으며 이 두개의 프로그램은 利用者에게 有意度 測定 및 상관 매트릭스를 제공하여 준다. SCATTERGRAM 프로그램은 두 變因間의 데이터들의 분포상태를 나타내 준다. 相關關係는 變因사이의 변화의 程度를 나타내 주는 것으로 變因사이의 結合의 程度 뿐만 아니라 관계를 비교하는 수단으로 이용되고 있다. 스피어맨의 로오(rho)나 캔달의 타우(tau)는 이중 非母數的 相關(two nonparametric correlation)으로 NONPAR CORR 서브프로그램에 의해 계산된다. 비모수적 방법(non-parametric)變因들 간의 전집분포에 대한 假定이나 조건을 요구하지 않는 통계방법으로 非分布 가정의 統計的 方法(distribution-free method)이라고도 한다. 이러한 統計的 方法은 여러 變因들 간의 序列測定이나 順位測定에 사용되므로 이러한 방법은 같은 경우에 있어서 2개의 서열이 유사한가를 測定하는데 사용된다. 물론 두 변인간의 서열은 매우 유사하지만 여전히 몇개의 차이점이 존재한다. 로오(rho)나 타우(tau)는 우리에게 그들 變因이 실제로 얼마나 類似한가를 測定해 준다. 그러나 序列의

假定은 同一한 順位를 갖는 많은 경우에는 設定하지 말아야 한다. 즉, 중복 순위가 많을 경우 실제의 비교는 2개의 서열 set 간에는 적당치가 않다. SCATTERGRAMS와 피어슨의 적률상관계수 r은 상관의 정도와 방향을 보여준다. SCATTERGRAM은 2개의 變因 즉 x와 y의 相關關係를 分布圖로 나타낸 가장 상적인 예가 〈圖 1〉이다.

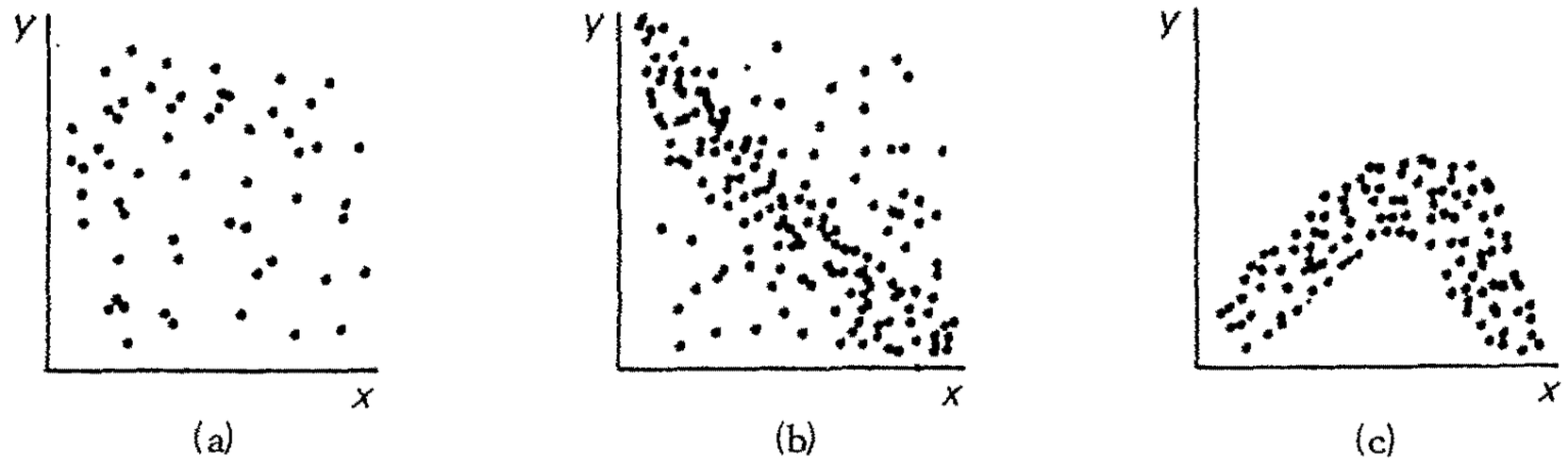
상관표에 있어서 分布圖가 매우 상세히 나타나는 것을 피하기 위하여 각각의 點들의 양상에 따라 直線이나 曲線을 그려준다. 이 경우 그 패턴이 분명하고 連續的이면 群集된 點들이 좁은 띠의 형상을 이루고 직선으로 요약이 되어 〈圖 2-a〉처럼 데이터들 간에 어떤 체계적인 상관관계가 나타나지 않는 경우가 있고 이와는 대조적으로 〈圖 2-b〉처럼 데이터의 分布狀態를 꼬집어 낼 수 있으며 〈圖 2-c〉처럼 데이터의 群集狀態가 특이한 形狀을 유지하는 경우도 있다.

이렇듯 데이터들의 군집상태(cluster)가 수학적 속성을 나타내는 직선의 패턴을 이루면 그 직선의 공식은 두 變因間의 關係를 요약해서 나타내 준다.

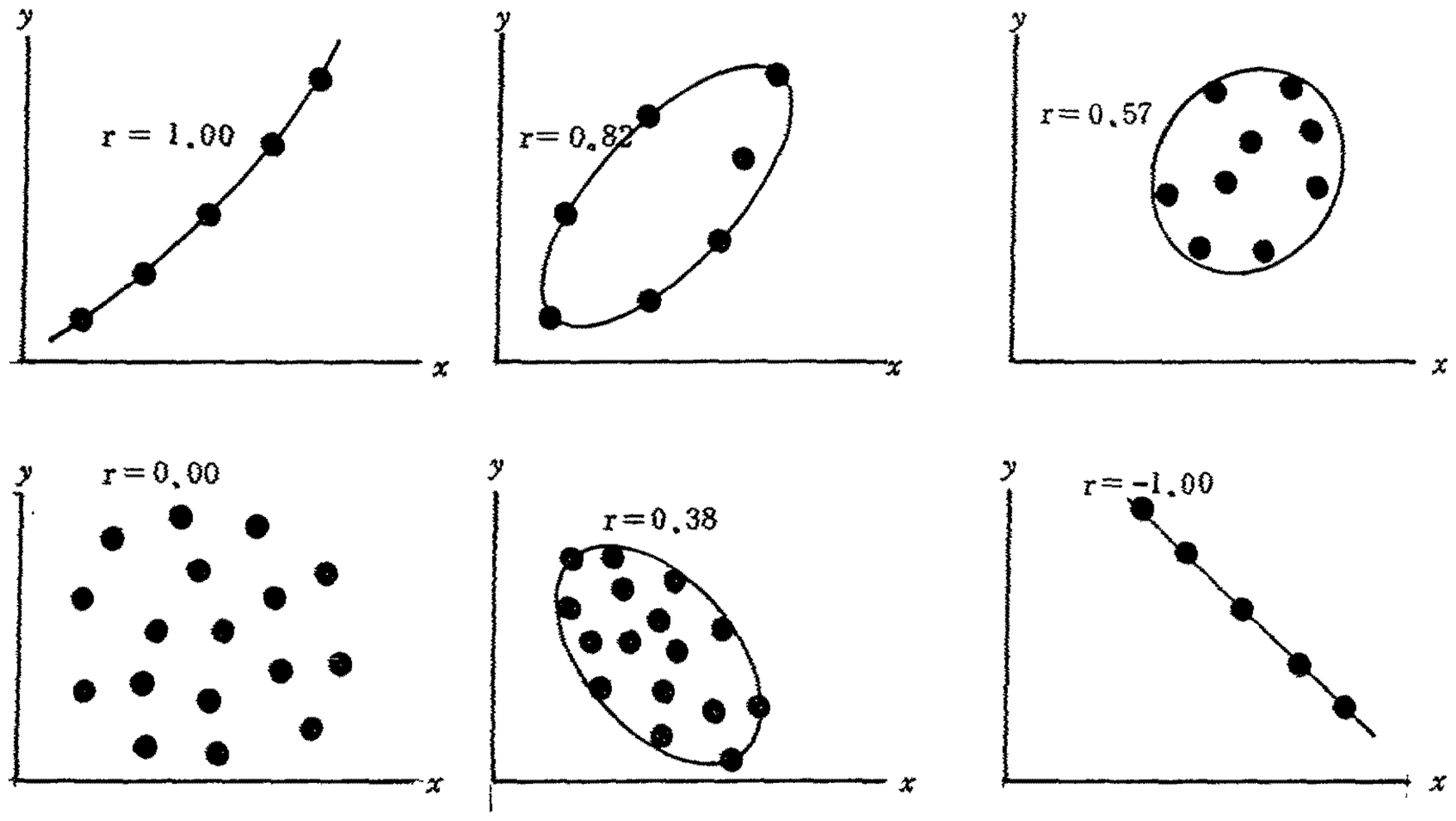
〈圖 1〉　　강한 직선관계를 나타내주는 **SCATTERGRAM**



〈圖 2〉　데이터들간에 관계를 보여주는 각기 다른 형태의 **SCATTERGRAMS**



〈圖 3〉　　두 변인간의 상관관계를 보여주는 분포도의 상관계수 r



〈圖 3〉은 두개의 變因 즉, X와 Y와의 여러 유형의 상관관계를 分布圖로 나타낸 것인데 이들 각 그림에 해당되는 피어슨의 積率 相關係數는 그 相關의 程度와 方向에 따라 $-1 \leqq r \leqq 1$의 값을 갖고 있음을 알 수 있다. 여기서의 數値는 상관의 程度를 +와 −의 符號는 상관의 方向을 나타내는데 數値가 높을 수록 상관의 程度가 높으며, 한 變因의 값이 커질때 다른 變因의 값도 커지면 r는 +의 값을, 그 반대의 경우 −값을 갖게 된다. 즉 +값은 正的相關關係(positive relationship), −값은 負的 相關關係(negative relationship)를 나타낸다. 특히 $r = \pm 1$의 값은 완전한 상관관계를, $r = 0$은 두 변인간에 아무런 상관관계가 없음을 보여준다. 한편 順位相關變因에 근거한 分布圖에서 가장 적합한 직선을 찾는 方法은 最小自乘法(least squares regression)으로 모든 점(data)들로부터 最小限의 거리를 가진 y의 직선을 구할 수 있고 이 거리 차의 제곱의 합이 최소가 되도록 직선을 나타내는 方程式을 구할 수 있다. 이러한 선형회귀(linear regression) 方式은 設問調査方法에 흔히 使用되는데 이는 간단한 相關關係를 보여주기 때문이다. 이 직선공식으로는,

$$y = a + bx$$

(a : 절편, b : 기울기)

이며, a와 b의 값이 最小自乘法에 의해 결정이 되었다면 b를 回歸係數라고 한다. SCATTERGRAM 서브 프로그램은 데이터의 構成形態를 나타내줄 뿐만 아니라 直線回歸率은 절편, 기타의 統計分析方法 계산을 할 수 있게 해준다.간혹 2원적 상관관계에 있어서는 曲線式(curvilinear)이나 多項式(nomial regression)회귀법이라 불리는 회귀분석법이 사용되는데, 이 경우 사용되는 공식으로는

$$Y = a + \hat{b}_1 x + b_2 x^2 + b_3 x^3 + \cdots\cdots\cdots b_n x^n$$

여기서 지수 n은 多項度(polynomial degree)를 나타내며, 이 方法은 두 변인간의 관계를 적절히 표현하는 방법을 찾을 때 이용할 수 있다. 대부분의 社會科學 硏究方法에서 사용되는 데이터에 적합한 線型 回歸法의 "goodness of fit"에 타당한 방법으로는 r로 표시되는 피어슨의 적률상관 계수로 이원분포의 상관계수를 하나의 최적의 선으로 나타내는 공식의 기울기를 나타내는 상수가 된다. 오차의 크기는 실제 데이터인 회귀선(regression line)과 수직의 거리를 측정하면 되는 데 모든 경우에 있어서 그 거리를 자승하여 더하여 전체중에서 1을 뺀(N-1)것으로 나누어 주면 평균이 된다. 이 통계치를 잔여변수(residual variance)라 하고 공식은 다음과 같다.

$$r^2 = \frac{\text{총변수} - \text{잔여변수}}{\text{총변수}}$$

여기서 r^2은 평균적인 결합상태를 측정한다.

3. Subprogram PEARSON CORR

SPSS 서브 프로그램인 PEARSON CORR은 2개의 변인 사이의 관계를 계산하는데 사용된다. 수학적으로 r는 X·Y를 2개의 변인으로 나타내며 그 공식은 다음과 같다.

$$r = \frac{\sum_{i=1}^{N} (X_i - \overline{X})\ (Y_i - \overline{Y})}{\left\{ \left[\sum_{i=1}^{N} (X_i - \overline{X})^2 \right] \left[\sum_{i=1}^{N} (Y_i - \overline{Y})^2 \right] \right\}^{1/2}}$$

X_i = 변인 X의 i 번째 사례
Y_i = 변인 Y의 i 번째 사례
N = 총사례 수
$\bar{X} = \sum_{i=1}^{N} X_i / N$ 변량X의 평균값
$\bar{Y} = \sum_{i=1}^{N} Y_i / N$ 변량Y의 평균값

이 공식을 보다 간단히 (N - 1)로 나누면

$$\frac{\sum_{i=1}^{n} (X_i - \bar{X})\ (Y_i - \bar{Y})}{N - 1}$$

이며, SPSS에서 사용되는 실제 相關係數公式은

$$r = \frac{\sum_{i=1}^{N} X_i Y_i - (\sum_{i=1}^{N} X_i)\ (\sum_{i=1}^{N} Y_i) / N}{\{[\sum_{i=1}^{N} X_i^2 - (\sum_{i=1}^{N} X_i)^2 / N][\sum_{i=1}^{N} Y_i^2 - (\sum_{i=1}^{N} Y_i)^2 / N]\}^{1/2}}$$

로 나타낸다. 相關係數의 有意度 測定은 student's t 를, 자유도 (n - 2)를 사용해서 $r\left[\frac{N-2}{1-r^2}\right]^{1/2}$ 로 나타낸다.

〈表 1〉은 연령(AGE), 교육(EDUC), TV시청량(TV), 신문독서량(NP) 등 각 변인간의 상관계수의 방향과 程度 그리고 그 係數의 有意度를 산출해 낸 것이다. 이 표에는 3가지 숫자가 나타나고 있는 데 맨 위의 것이 상관계수 r이고, 두번째 ( )안의 숫자는 관찰 대상자의 수 N을 나타내며, 맨 밑의 것이 零假說을 받아들일 수 있는 확률, 즉 有意度를 나타낸다. SPSS를 이용한 통계분석 산출표를 보면, 有意度가 SIGNIFICANCE 혹은 " S "의 약자 또는 "SIG" 로도 表記한다(일반 연구논문에서는 有意度를 P(Probability)로도 나타낸다). 〈表 1〉에 나타난 바와 같이 AGE와 TV간의 상관계수 r는 .4493, S= .012 이어서 有意的인 相關關係가 있다. EDUC와 NP의 관계는 r = .6853으로 이 역시 유의도가 S = .000 로 높다. 즉 교육수준이 높아짐에 따라 신문 읽는 시간이 늘어남을 보여준다.

〈表 1〉 **PEARSON CORR 산출표**

```
1                16
PEARSON CORR     AGE,EDUC,TV,NP
```

SPSS BATCH SYSTEM 84.02.13 PAGE

FILE NONAME (CREATION DATE = 84.02.13)

P E A R S O N C O R R E L A T I O N C O E F F I C I E N T S

	AGE	EDUC	TV	NP
AGE	1.0000 (0) P=*****	0.0057 (25) P=0.489	0.4493 (25) P=0.012	0.4237 (25) P=0.017
EDUC	0.0057 (25) P=0.489	1.0000 (0) P=*****	-0.5792 (25) P=0.001	0.6853 (25) P=0.000
TV	0.4493 (25) P=0.012	-0.5792 (25) P=0.001	1.0000 (0) P=*****	-0.3460 (25) P=0.045
NP	0.4237 (25) P=0.017	0.6853 (25) P=0.000	-0.3460 (25) P=0.045	1.0000 (0) P=*****

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

4. Subprogram NONPAR CORR : Spearman and/or Kendall Rank Order Correlation Coefficient

NONPARR Subprogram은 스피어맨과 켄달의 序列順位 相關係數를 계산한다. 統制語인 NONPARR은 2개의 相關係數를 비모수적방법에 근거해서 일반적인 분포나 同間尺度(interval scale)의 Metric Quality에 의존하지 않는다. 즉 그 尺度(scale)에 있어서는 序數이고 형태는 뉴메릭(numeric)이다. r_s로 표시하는 스피어맨의 로오(ρ)나 켄달의 타우(τ)는 변인간의 상관계수를 계산함에 있어서 정확한 값보다는 서열을 사용해서 측정한다. 그러므로 NONPAR CORR의 첫째일은 變因을 읽고 序列順位로 變因의 값을 바꾸어 준다. 누락된 데이터의 처리는 계속적인 수정이 요구되며 이러한 이유로 NONPAR CORR는 모든 데이터가 핵심 자리에 위치하기를 요구한다. SPSS 서브 프로그램 NONPAR CORR는 이론적으로 제한된 수의 경우를 處理할 수가 없고, 處理될 수 있는 총 變因間의 平均, 경우의 수, 계산상의 특징은 NONPAR CORR가 PEARSON

CORR 보다는 상대적으로 속도가 느리다. 출력물로는 스피어맨의 r , 켄달의 타우계수, 통계적인 유의도 검증 등이며, 계수의 메트릭스와 상관 메트릭스도 이용자의 선택에 따라 생산해 낼 수 있다. 스피어맨의 r 와 켄달의 타우는 많은 수의 데이터가 同等序列을 갖고 있을 때 보다 더 정확한 값을 산출해 낼 수 있다. 實用的인 方法으로 상당히 큰 경우의 수가 상대적으로 적은 카테고리에 分類되어 들어갈때 타우를 쓰고, r 는 보다 작은 규모의 경우에 사용된다. 그러나 각각의 節次는 結合하는 동안 修正이 되고 서로의 選擇作業에 대한 정해진 法則은 없다. 실제로 두 係數에 우선하는 기본 개념은 상당히 유사하며 각 계수의 값은 $-0.1 \sim +0.1$ 의 값을 갖지만 타우의 절대값은 피어슨맨의 r 보다 적은 값을 갖는다. 스피어맨의 r 는 모든 경우에 있어서 서열 척도에 의해 測定된 두 變因간의 相關關係를 재는 것으로 서열에 있어서 순위차 제곱의 합으로 두 서열간에 중복순위가 없어야하며 다음과 같이 계산이 된다.

$$r_s = 1 \quad \frac{6\sum_{i=1}^{N} d_i^2}{N^3 - N}$$

d_i ; 두 변인간의 순위차

N ; 표본의 수

중복순위가 발생했을 때 수정된 r_s 의 계산은

$$r_s = \frac{T_x + T_y - \sum_{i=1}^{N} d_i^2}{2(T_x T_y)^{1/2}}$$

x, y ; 서열에서 중복된 수

r_s 의 유의도 측정은 다음과 같다.

$$r_s \left(\frac{N-2}{1-r_s^2}\right)^{1/2}$$

여기서 N－2의 자유도는 Student's t 분포이다. 켄달의 τ 도 그 技法과 節次에 있어서 스피어맨의 r_s 와 유사하다. 두개의 序列 順位間의 一致의 程度에 根據한 係數의 標準化 作業으로 켄달의 τ 는 s 라 불리우는 統計値의 計算에 의해서 얻어진다. 1개의 변인을 실제의 순서에 따라 서열을 매기면 (1～

N까지) S는 실제로 얻은 점수의 총합이고 $\frac{1}{2}N(N-1)$은 가능한 最大點數의 總合을 나타내며 실제 공식은

$$\tau = \frac{S}{\frac{1}{2}N(N-1)}$$

이다.

중복순위가 있을 경우 공식은

$$\tau = \frac{S}{\sqrt{\frac{1}{2}N(N-1)-T_x}\ \sqrt{\frac{1}{2}N(N-1)-T_y}}$$

$T_x = \frac{1}{2}\sum t(t-1)$ t는 S변인의 동등그룹에 있어서 관찰된 중복수

T_x ; t는 X변인에 있어서의 중복순위수

T_y ; t는 Y변인에 있어서의 중복순위수

τ의 유의도는 평균편차를 갖는 정상분포의 τ와 비교해서 얻을 수 있다.

$$\text{유의도} : \left(\frac{4N+10}{9N(N-1)}\right)^{1/2}$$

한편 相關係數를 계산할 수 없을 때 SPSS에서는 99.0의 값을 할당하고 이는 이용자에게 계수를 계산할 수 없다는 것을 나타내준다. NONPAR CORR로

〈表 2〉 NONPAR CORR 산출표

```
1                                    16
NONPAR CORR                          TV,NP

SPSS BATCH SYSTEM                                          84.02.13        PAGE   32

FILE   NONAME   (CREATION DATE = 84.02.13)

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

VARIABLE        VARIABLE        VARIABLE        VARIABLE        VARIABLE        VARIABLE
PAIR            PAIR            PAIR            PAIR            PAIR            PAIR
--------        --------        --------        --------        --------        --------

TV      (-0.2863
WITH      N(   25)
NP        SIG .083

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

順位相關係數를 계산하며 〈表 2〉는 實例를 보여주고 있다.

〈表 2〉는 TV시청량(TV)과 신문독서량(NP)의 값을 序列尺度로 측정 했다고 가정 했을 때, 두 변인간의 관계를 스피어맨의 로오(ρ)로 산출해낸 것이다. 두 변인간의 非母數 相關係數는 −.2863 으로 負的關係(negative)이며, 그 유의도(SIG)는 .083 으로 5% 수준 내에서 零假說을 받아들이게 된다. 즉, P > .05 로 TV와 NP는 별다른 順位相關關係를 나타내지 않는다고 結論을 내릴 수 있겠다.

5. Subprogram SCATTERGRAM: Scatter Diagram of Data Point & Simple Regression

SCATTERGRAM 측정법은 노드 캐롤라이나(North Calrolina)대학의 레이놀즈(Bill Reynolds)에 의해 개발된 것으로 2개의 변인들 간의 관계를 데이터의 점으로 나타내 주는 것으로 1개의 변인이 수직축이 되고 다른 하나는 수평축이 된다. SPSS의 서브 프로그램인 SCATTERGRAM은 두 변인간의 관계를 그래프로 나타낸 것으로 그래프 상에 세로 51 컬럼, 가로 101 컬럼으로 나타낸다. 출력시에는 페이지당 55 개의 선(line)이 요구되고 페이지 사이즈에 영향을 받지 않는다. 데이터들이 인쇄된 위치는 작은 직사각형의 영역을 나타내는데 완전히 유사하지 않는 한 일치하지 않는다. 만일 하나 또는 두개의 변인이 거의 비슷한 카테고리에 속해 있다면 그 결과 데이터 점들의 중복이 되어 SCATTERGRAM의 效率性에 制限을 받게 된다. 이 경우 십자표의 작성 또는 Breakdown이 이 문제를 해결해 준다. SCATTERGRAM Subprogram에서 이용할 수 있는 통계값은 간단한 선형 회귀선과 일치한다. 그래프의 스케일은 각 변인의 最小値(lowest)와 최대치(highest)에 의해 자동적으로 결정되지만, 이용자가 任意로 결정할 수도 있다. 이 프로그램에서는 회귀선의 기울기, 적률상관계수, 통계적 유의도측정 등을 구할 수 있으며 기울기를 구하는 공식으로는

$$b = \frac{\sum_{i=1}^{N} (X_i - \overline{X})\ (Y_i - \overline{Y})^2}{\sum_{i=1}^{N} (X_i - \overline{X})^2}$$

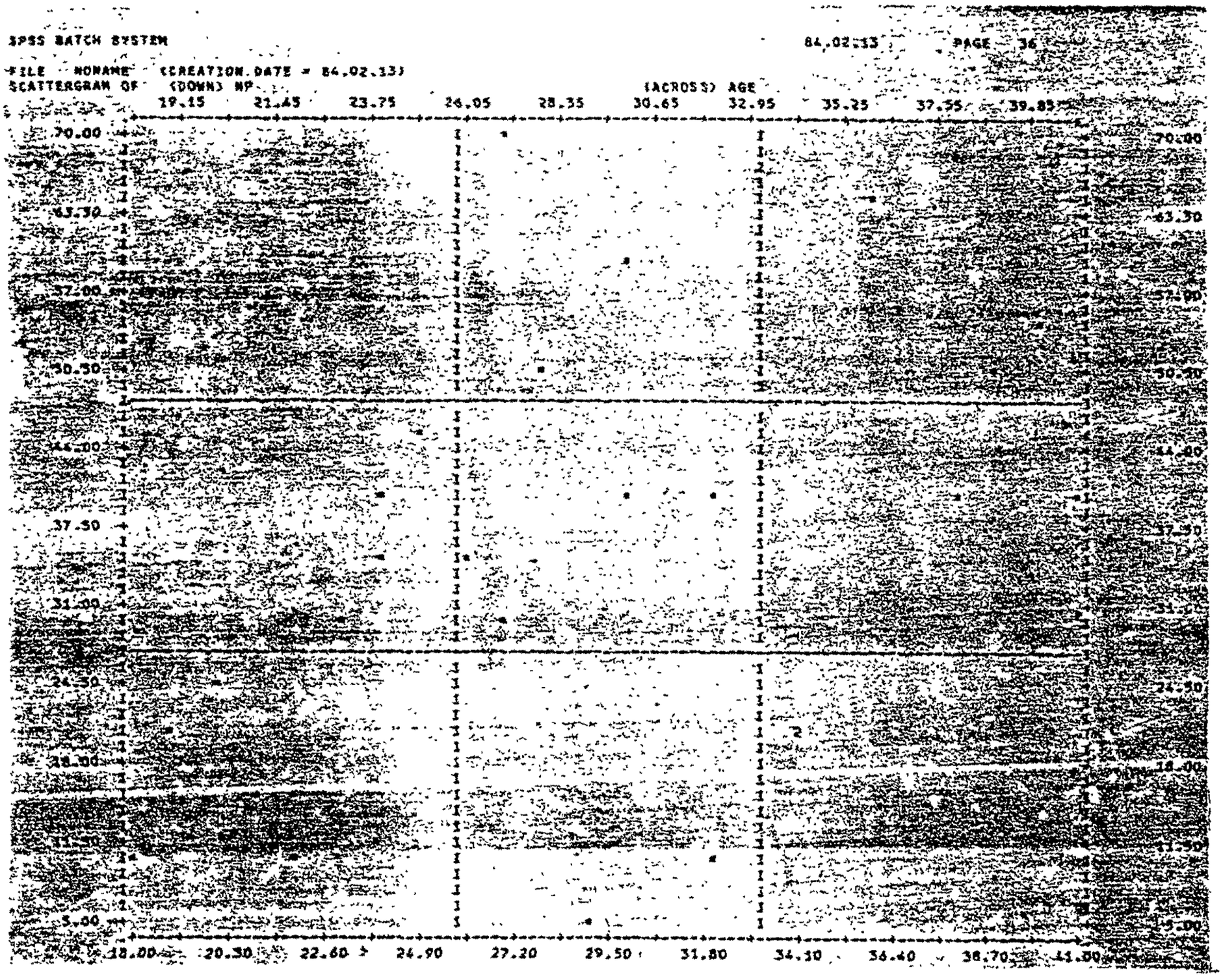
X_i = 變因 x의 i 번째 측정치(수평축)
Y_i = 變因 y의 i 번째 측정치(수직축)
N = 총 빈도수
$\overline{x}$; x의 평균값
$\overline{y}$; y의 평균값

SPSS에서 사용되는 계산공식으로는

$$b = \frac{(N\sum_{i=1}^{N} X_i Y_i) - (\sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i)}{(N\sum_{i=1}^{N} X_i^2) - (\sum_{i=1}^{N} X_i)^2}$$

SCATTERGRAM에서 얻을 수 있는 통계값으로는 피어슨의 로오(ρ) r의 有意度, 測定値의 標準誤差, 垂直軸의 常數, 기울기 등이며 SCATTERGRAM의 實際例는 〈圖 4〉와 같다. 이 표의 요약표를 보면 오른쪽 끝부분에 b係數(B)

〈圖 4〉 **SCATTERGRAM 산출표**



1.130025가 나오는데, 이것은 기울기(slope)값이고 b계수(B) 아래 常數 constant)는 1.900488이 나오는데 이는 절편(intercept)값을 나타낸다.

〈圖 4〉는 연령(AGE)을 X軸에 놓고, 신문독서량(NP)을 Y軸에 놓은 SCATTERGRAM이다. 두 變因間의 相關關係(correlation, R)는 .42365이며, 결정계수(R, squared,, r^2)는 .17948이다. 이는 연령이 신문 독서량 변량을 약 18% 설명할 수 있다는 뜻이다. Y = a + bX라는 일차방정식은 〈圖 4〉와 〈表 3〉에서 제시된 통계치들로 쉽게 산출될 수 있는데, 이 경우 절편(inter - cept) a는 1.900488, 기울기(slope) b는 1.130025이다. 따라서 주어진 통계치로부터 회귀방정식을 만들 수 있다. X와 Y 두 변인의 상관관계가 完全한 관

〈表 3〉 **REGRESSION 산출표**

```
SPSS BATCH SYSTEM                                                    84.02.27     PAGE  39
FILE   NONAME   (CREATION DATE = 84.02.27)

VARIABLE              MEAN     STANDARD DEV     CASES

NP                 34.4000          18.0462        25
AGE                28.7600           6.7636        25

SPSS BATCH SYSTEM                                                    84.02.27     PAGE  40
FILE   NONAME   (CREATION DATE = 84.02.27)

CORRELATION COEFFICIENTS

A VALUE OF 99.00000 IS PRINTED
IF A COEFFICIENT CANNOT BE COMPUTED.

             NP         AGE

NP          1.00000    0.42365
AGE         0.42365    1.00000

SPSS BATCH SYSTEM                                                    84.02.27     PAGE  41
FILE   NONAME   (CREATION DATE = 84.02.27)
* * * * * * * * * * * * * * * * * * * * 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..   NP

VARIABLE(S) ENTERED ON STEP NUMBER  1..   AGE

MULTIPLE R           0.42365              ANALYSIS OF VARIANCE   DF     SUM OF SQUARES     MEAN SQUARE          F
R SQUARE             0.17948              REGRESSION             1.         1402.81274     1402.81274     5.03099
ADJUSTED R SQUARE    0.14380              RESIDUAL              23.         6413.18726      278.83423
STANDARD ERROR      16.69833

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

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

AGE             1.130025      0.42365      0.50380       5.031
(CONSTANT)      1.900488

MAXIMUM STEP REACHED

STATISTICS WHICH CANNOT BE COMPUTED ARE PRINTED AS ALL NINES.

SPSS BATCH SYSTEM                                                    84.02.27     PAGE  42
FILE   NONAME   (CREATION DATE = 84.02.27)
* * * * * * * * * * * * * * * * * * * * 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..   NP

                                              SUMMARY TABLE

VARIABLE                              MULTIPLE R   R SQUARE   RSQ CHANGE   SIMPLE R          B            BETA

AGE                                     0.42365     0.17948     0.17948     0.42365     1.130025       0.42365
(CONSTANT)                                                                              1.900488
```

계가 아닐때($r_{xy} \neq 1.0$), 이용자는 X단위의 증가에 따른 Y단위의 증가 형태를 SCATTERGRAM으로 살펴보아 두 변인의 관계가 線型(linear)인지 曲線(curvilinear)인지, 또는 다른 函數關係인지를 육안으로 식별할 수가 있다.

Ⅳ. 相關分析의 應用

이상에서 본바와 같이 變因間의 相關度를 수치로써 표시해주는 相關分析은 처음 그 대상이 되었던 커뮤니케이션學은 물론 모든 社會科學分野, 특히 圖書館學·情報學分野의 硏究에서 매우 유용하고 중요하게 사용되고 있다. 즉, 첫째, 相關分析은 因果關係나 共通要因을 발견하려는 연구에서 많이 이용된다. 예컨대 引用文獻과 著者들 間의 相互關聯性이 무엇인가 하는 문제를 연구할 때 쓸 수가 있다. 그러나 이때 유의할 점은 두 變因 사이에 높은 相關度가 있다고 해서 반드시 이들사이에 因果關係(causal relationship)가 존재 한다는 것은아니다. 또한 相關分析은 變因들 間의 共通要因을 발견하는데도 이용이 된다. 예컨대 같은 계열의 저자들 간의 相關度에 따른 클러스터링을 통한 要因分析(factor analysis)으로 共通要因을 찾아 보는 것이 그점이다.

둘째, 相關分析은 하나의 變因을 알고 다른 變因을 豫測하려는 硏究에서 많이 사용되고 있다. 왜냐하면 相關係數가 높다는 것은 두 變因間의 共通關係가 뚜렷하다는 것을 말해, 한 變因을 알고 다른 변인을 豫測 하려는데서 더욱 많이 이용이 되고 있기 때문이다.

세째, 相關分析은 硏究 結果의 測定에도 많이 應用이 되고 있다. 測定檢査에서는 信賴度(realiability)와 妥當性(validity) 등을 檢討할 때 相關分析이 應用되고 있다.

Ⅴ. 結　　論

相關分析은 1896년 칼·피어슨(Karl Pearson)에 의해 그 技法이 學問分野

에 導入된 이래 他 學問分野에서 硏究分析 등에 많은 應用이 되고 있으나 圖書館學·情報學分野에서는 이렇다할 應用性을 보이지 못하고 있었는데 圖書館學·情報學分野의 硏究方法에 計量書誌學的 槪念이 1969년 프리차드(Pritchard)에 의해 정식으로 도입된 이래 圖書館學·情報學分野에서도 많은 應用事例를 볼 수가 있다. 컴퓨터의 등장으로 手作業에 의존하던 작업들이 電算化되어 간편해지고, 특히 1965년 스탠포드 대학 부설 政治學硏究所에서 개발한 社會科學分野 統計處理를 위한 SPSS 소프트웨어의 출현으로 그 應用性이 상당히 높아졌다 하겠다. 특히 여기에서 다룬 相關分析은 SPSS의 他 서브시스템들 보다 圖書館學, 情報學分野에서의 사용 빈도가 많고 書誌分析을 통한 硏究 즉,

① 文獻의 言語, 出版年, 出版國, 主題, 形態別 分析

② 文獻의 增加率 把握

③ 二次 서비스(索引雜誌, 抄錄誌, SDI 서비스)의 完全度 調査

④ 引用文獻分析을 통한 文獻들의 主題別 分類

⑤ 著者들의 文獻引用 패턴 및 特殊 主題 分野의 硏究前線 把握

⑥ 著者와 生産性 및 共同著作狀況 把握

⑦ 其他統計

등의 硏究時에 應用할 수 있겠다.

SPSS 패키지 외에도 계량서지학용 소프트 웨어 패키지로는 筑波大學의 대형 계산기용의 데이터 분석 프로그램인 PAB가 開發되어 가동중인데 클러스터 분석, 二元클러스터 分析, 階層分析(인용 매트릭스를 입력하여 요소를 계통적으로 배려하여 그래프 출력), 브래드포드 曲線出力, 時系列 그래프 출력을 하며, 파라미터는 SPSS 시스템과 동일한 형식으로 입력 되고 있다. 또한 퍼스날 컴퓨터(NEC PC-9800 시리즈)용 소프트 웨어가 大阪府立大에서 개발되어 序列順位, 頻度, 分布分析, 브래드포드(Bradford's distribution)분포, 로트카(Lotka's distribution) 분포, 지프(Zipf's distribution) 분포를 출력할 수 있다. 本稿에서 記述한 외에도 SPSS의 相關分析은 圖書館學·情報學分野의 書誌分析技法이 發達하고, 이를 통한 多方面의 硏究가 활발해질수록 그 應用性은 무한하다 하겠다.

〈參 考 文 獻〉

1. 김광웅, 「사회과학연구방법론」, 서울 : 박영사, 1979.
2. 권오룡, 「설문조사분석의 전산처리방법 : SPSS를 중심으로」, 「정보관리연구」, vol.16, no.2, 1983, pp.99-110.
3. 김석영, "학술잡지의 이용연구", 「정보관리연구」, vol.17, no.1, 1984, pp. 12-25.
4. 변형윤, 「통계학」, 서울 : 일조각, 1982.
5. 오택섭, 「사회과학 데이타 분석법」, 서울 : 나남, 1984.
6. 이만갑, 「사회조사방법론, 서울 : 박영사」, 1975.
7. 임현재, 「교육, 심리, 사회연구를 위한 통계방법, 서울 : 박영사」, 1976.
8. 정영미, 「계량서지학적 연구에 관한 고찰」, 「도협월보」, vol.19, no.1, 1978, pp.3-9.
9. 차배근, 「사회통계방법, 서울 : 세영사」, 1982.
10. 최희윤, 「주제문헌의 계량서지학적 분석에 관한 고찰」, 「정보관리연구」, vol,17, no.1, 1984, pp.26-58.
11. 한두완, 「요소분석의 이론과 응용」, 「정보관리연구」, vol.15, no.4, 1982, pp.179-188.
12. 上田修一, 「科學技術文獻を利用した研究」, 「ドクメンテーション研究」, vol.34, no.4, 1984, pp.175-185.
13. 日本圖書館學會, 「圖書館學の研究方法」, 東京 : 同學會, 1982.
14. Aina, L.O. "Use of SPSS for Bibliometric Study," *Program*, vol.16, 1982, pp.35-38.
15. Brooks, B.C., "Frequency-Rank Distributions," *JASIS*, vol.29, no. 1, 1978, pp.5-14.
16. Busha, C.H., *Research Methods in Librarianship*, New York : Academic Press, 1980.
17. Dixon, W.J., *Introduction to Statistical Analysis*, New York : McGraw-Hill, 1969.
18. Fisher, R.A., *Statistical Methods for Research Workers*, N.Y. : Hafner Pr., 1972.
19. Garrison, G. "Research Methods in Librarianship", *Lib. Trends*, July, 1964.
20. Hoel, P.G., *Elementary Statistics*, 4th ed., N.Y. : John Wiley, 1976.
21. Nie, N.H. SPSS : *Statistical Package for the Social Sciences*, N.Y. : McGraw-Hill, 1975.
22. Rosenverg, V., "Factor Affecting the Preferences of Industrial Personnel for Intormation Gathering Methods", *Inf. Stor. & Ret.*, vol.3, no.3, 1967, pp.95-100.
23. Walpole, R.E. *Elementary Statistical Concepts*, London : Macmillan,

1976.
24. White, H.D., "Author Cocitation ; Literature Measure", *JASIS*, vol. 32, no.3, 1981, pp.163-171.
25. Wonnacott, T. H., *Introductory Statistics*, 2nd ed., N.Y. : John Wiley, 1972.