The Korean Communications in Statistics **Vol. 14** No. 2, 2007, pp. 301–307

A Comment for Teaching Correlation Coefficient in Elementary Statistics Course

Myongsik Oh[1)]

Abstract

A effective teaching method on correlation coefficient for elementary level statistics course is discussed in this article. The well known inequalities, such as Theorem 368 of Hardy *et al.* (1952), are used for the interpretation of concept of covariance. An Excel example is provided for the illustration of concept of correlation coefficient.

Keywords: Correlation coefficient; covariance; inequality; sum of cross products.

1. 서론

최근들어 특히 제7차 교육과정과 더불어 현재의 대학입시체제하에서는 대학신입생들의 수학에 대한 이해도가 대학에서의 수학능력에 지장을 줄 정도의 수준으로 소위 상위권대학이나 중하위권대학에서도 고르게 문제가 되고 있다. 특히 상경계를 지원한 학생들의 대부분은 고등학교에서 문과과정을 택하여 미적분은 전혀 배우지 않아 정상적인 교육이 어려운 실정에 있다. 최수일 (2006)은 제7차교육과정과 관련하여 고교에서의 통계교육의 문제점에 대하여 몇가지 관점에서 연구하였는 데 이러한 문제는 대학에서의 교육에도 어느정도 공유할 수 있는 문제점으로 판단된다.

이러한 문제점은 자연스럽게 상경계에서 저학년에서 다루는 기초통계학 교육에서도 많은 어려움을 야기시키고 있다. 따라서 자연스럽게 기초통계학교육은 예전과 다르게 주로 엑셀 혹은 Minitab등 비교적 간단하게 사용할 수 있는 프로그램을 이용하여 강의가 이루어 지게 됨으로서 상위 과목으로 올라가는 경우 통계에 대한 기본적인 개념의 이해가 부족한 편으로 지적되고 있다. 그저 자료를 입력하고 단추 몇개만 클릭하면 얻어지는 숫자들에만 관심을 갖게 되고 그 결과로 개념에 대한 확고한 이해가 없는 편향된 상태로 통계학에 대한 경험을 마치게 된다. 이는 상위과목으로 이동했을 때의 심층적인 통계문제에 대한 체계적인 분석능력을 배양하는 데 커다란 지장을 초래하게 될 것이다. 오히려 기초 통계학 수준에서는 중요한 몇 가지의 통계적방법에 대한 확고한 이해를 갖게 교육을 하는게 바람직 할 것으로 판단된다.

1) Professor, Department of Statistics, Pusan University of Foreign Studies, 55-1 Uam-Dong, Nam-Gu, Busan 608-738, Korea.
E-mail: moh@pufs.ac.kr

이러한 관점에서 볼 때 기초통계학을 수강하는 학생들에게 있어 비교적 초반에 접하는 상관관계에 대한 이해도는 회귀분석과 같은 관련분야만이 아니라 통계적추론 등 전반적인 통계학 교육의 성과에 대한 가늠자 역할을 할 수 있다. 따라서 상관관계에 대한 이해 구체제적으로는 상관계수에 대한 올바르고 확고한 이해가 중요할 것으로 생각된다. 따라서 다른 통계적방법보다는 비교적 많은 수의 교육관련논문이 나와있는 것을 알 수 있다. 대표적인 예로서 미국통계협회 (American Statistical Association)에서 발행하는 American Statistician의 Teacher's Corner를 통해 1975년 이래 15편이상의 상관계수의 교육에 관련된 눈문이 발표되고 있다. 몇 가지 대표적인 논문을 열거해 보면 Leung와 Lam (1975), Koch (1985), Rodgers와 Nicewander (1988), Rovine과 von Eye (1997) 등을 들 수 있다.

상관계수는 두 개의 변인간의 관련성을 알아보기 위한 통계적 방법 중 가장 기본적이고 중요한 척도이다. 따라서 통계학 교육 특히 입문수준의 통계학 교육에 있어 상관계수의 이해는 매우 중요한 과제 중의 하나이다. 그러나 쉬울 것 같은 상관계수의 이해는 일부의 대학을 제외하곤 대부분의 대학에서의 기초통계 수강생들에게는 그리 쉬운 일은 아니다. 이는 상관계수 하나를 계산하기 위해선 그때 까지 배운 모든 자료의 요약방법 즉 평균과 표준편차를 계산해 낼 줄 알아야 하며 여기에 더해 비교적 개념을 파악하기 쉽지 않은 공분산의 계산도 필요하다. 이러한 긴 과정은 대개 학생들에겐 좋은 기억으로 남게 되지는 않는다. 따라서 컴퓨터를 이용한 계산과 계산된 숫자에 대한 해석에만 치중을 하게 되어 실제로 성공적으로 수강을 마친 학생들 조차도 상관계수의 계산에서 각각의 세부적인 계산들이 무엇을 의미하는지 정확하게 알지 못하는 경우가 많다.

따라서 여기에서는 상관계수의 계산에서 기본적으로 사용되는 교차곱의 합에 대한 이해를 바탕으로 상관계수에 대한 기본적인 이해능력을 높이는 방안을 논하기로 한다. 이에 사용되는 널리 알려진 부등식과 엑셀을 보조적으로 사용하는 사례를 언급한다. 2절에서는 교차곱의합의 상한과 하한이 정해지는 규칙을 이해시킴으로 상관계수의 이해에 절대적인 공분산의 개념을 이해 시키는 방법을 논한다. 3절에서는 2절의 내용을 보조적 수단으로서의 엑셀을 통하여 설명하는 한 방법을 기술한다.

2. 교차곱의 합의 상한과 하한

상관계수는 두 변인의 공분산을 두 변인의 표준편차들로 나누어 준 값으로 계산되어 진다. 따라서 상관계수의 올바른 이해를 위해서는 공분산의 개념과 표준편차로 나누어 주는 이유를 정확하게 이해하는 것이 매우 중요하다. 첫째 이러한 상관계수의 계산에서 제일 중요한 부분인 공분산 즉 교차곱의 합을 왜 사용하는가 하는 점에 대한 이해를 이끌어 내야한다. 두 번째로는 표준편차로 나누어 주는 이유에 대한 명확한 설명이 필요할 것이다. 첫번째 교차곱의 합에 관한 설명은 현재 사용되고 있는 기초통계학수준의 대부분의 교재에서는 찾을 수가 없다. 한편 통계학 강의를 담당하는 대부분의 사람에게 있어서 두번째 문제는 Cauchy-Schwarz 정리를 통하여 상관계수의 절대값이 1보다 작다는 사실을 증명해 본 경험이 있을 것으로 생각한다. 이러한 관점에서 기초통계학을

수강하는 학생들에게 이 두가지 점을 쉽게 풀어 설명할 수 있는 방법을 찾아야 한다. 그러면 첫번째로 교차곱의 합을 사용하는 문제를 먼저 다루어 보자.

이 문제는 간단하나 매우 강력한 응용력을 가진 다음의 정리를 이용하여 설명할 수 있다. 이 정리는 Hardy 등 (1952)의 정리 368을 다시 쓴 것이다.

정리 2.1 두 개의 유한 수열 $a_1, a_2, \ldots, a_n$, $b_1, b_2, \ldots, b_n$가 있다 하자. 또한 이 수열을 크기순으로 재배열한 수열을 $a_{(1)}, a_{(2)}, \ldots, a_{(n)}$, $b_{(1)}, b_{(2)}, \ldots, b_{(n)}$이라 하자. 즉 $a_{(1)} \leq a_{(2)} \leq \cdots \leq a_{(n)}$, $b_{(1)} \leq b_{(2)} \leq \cdots \leq b_{(n)}$이다. 이 두 수열의 교차곱의 합은 다음과 같이 하한과 상한의 값을 갖는다.

$$\sum_{i=1}^{n} a_{(i)} b_{(n+1-i)} \leq \sum_{i=1}^{n} a_i b_i \leq \sum_{i=1}^{n} a_{(i)} b_{(i)}.$$

참고로 정리 2.1의 부등식을 이해할 수 있는 학생들에게는 다음과 같은 증명법을 소개하는 것도 유용할 것으로 생각된다. 수열 a_i는 크기순으로 나열되어 있고 b_i는 그렇지 않다고 하면 $a_j \leq a_k$인 j와 k에 대하여 $b_j > b_k$인 j와 k가 존재한다. 따라서

$$a_j b_j + a_k b_k - a_j b_k - a_k b_j = (a_j - a_k)(b_j - b_k) \leq 0$$

임을 알 수 있다. 이를 이용하면 위의 부등식을 쉽게 증명할 수 있다.

정리 2.1에서 우리는 두 수열의 교차곱의 합은 두 수열이 서로 같은 방향으로 크기순으로 나열되어 있을 때 가장 큰 값을 갖으며 반대로 서로 다른 방향으로 나열되었을 때 가장 작은 값을 갖게 됨을 알 수 있다. 학생들에게는 위 정리의 내용을 여러가지 경우의 수열조합에 대하여 실제적으로 계산한 결과를 보여줌으로써 정리 2.1의 내용을 설명한다. 예로 다음의 경우를 생각해 보자.

표 2.1: 여러가지 수열에 대한 교차곱의 합 비교

a	1	2	3	4	5	a와의 교차곱의 합	
	1	2	3	4	5	55	⇐ 최대
	3	1	4	2	5	50	
b	4	3	2	1	5	45	
	3	4	5	1	2	40	
	5	4	3	2	1	35	⇐ 최소

따라서 정리 2.1로 부터 우리는 교차곱의 합을 계산함으로서 두 수열의 배열이 어느 정도 같은 방향으로 배열되었는지를 숫자로 나타낼 수 있을 것이다. 즉 상한에 가까이 갈 수록 같은 방향으로 배열되었음을 나타내며 반대로 하한에 가까울수록 서로 다른 방향으로 배열되었을 알 수가 있다. 이점이 바로 상관계수를 이해하는데 가장 중요한 점이 되는 것이다.

한편 정리 2.1를 수열의 일부분에 축차적으로 적용하면 임의의 배열에 의한 교차곱의 합 사이의 대소 관계를 쉽게 밝힐 수 있다. 이는 정리 2.1를 증명하는 데 사용되었던

표 2.2: 두 개의 임의의 수열에 대한 교차곱의 합 비교

a	1	2	3	4	5	a와의 교차곱의 합
	5	4	3	1	2	36
		⇓				
	3	4	5	1	2	40
b			⇓			
	3	1	4	5	2	47
				⇓		
	3	1	4	2	5	50

방법을 이용하는 것이다. 예를 들어 두 수열의 (3,1,4,2,5)와 (5,4,3,1,2)의 비교를 표 2.2에서 볼 수 있다. 이 과정은 정리 2.1을 이용하여 임의의 배열간의 교차곱의 합의 대소를 밝힐 수 있다는 사실을 이해시키는 데 사용될 수 있을 것이다.

그러나 단순히 수열의 교차곱의 합을 보고 서로 같은 방향인지 아니면 반대의 방향인지를 알아내기는 어렵다. 즉 각각의 경우마다 교차곱의 합의 값이 달라 매번 상한과 하한을 계산해야만 어느쪽으로 가까운지 알 수 있게 되는 불편함이 있다. 따라서 적어도 부호의 개념을 도입하여 그 부호에 따라 방향을 알 수 있게 됨을 이해시킬 필요가 있을 것이다. 이점은 각 수열에서 자신의 평균을 뺀 수로 바꾸어 그들의 교차곱의 합을 보여줌으로써 학생들이 어렵지 않게 이해 할 것으로 생각된다. 부가적으로 공분산의 계산에서 $E(XY) - E(X)E(Y)$와 같이 두개의 평균의 곱을 교차곱의 합에서 빼내는 이유를 쉽게 설명할 수 있을 것으로 기대된다.

표 2.3: 평균으로 조정한 후 교차곱의 합의 비교

a	1	2	3	4	5	교차곱의 합
	1	2	3	4	5	55
	3	1	4	2	5	50
b	4	3	2	1	5	45
	3	4	5	1	2	40
	5	4	3	2	1	35

⇒

a	−2	−1	0	1	2	교차곱의 합
	−2	−1	0	1	2	10
	0	−2	1	−1	2	5
b	1	0	−1	−2	2	0
	0	1	2	−2	−1	−5
	2	1	0	−1	−2	−10

부호를 도입하여 즉 각 수열의 평균을 뺀 후 교차곱의 합으로 두 수열이 어느 방향으로 연관되어 있는지는 알 수 있게 되었지만 같은 방향으로 아니면 다른 방향으로 얼마나 강하게 연관되어 있는지를 아직 알기에는 쉽지 않다. 기초통계학을 수강하는 학생들보다는 높은 수준의 교재인 Hogg 등 (2005)의 예제 2.4.3 (104쪽)을 보면 이변량 균등분포를 이용하여 상관계수가 어떻게 두 변인의 직선관계의 강도를 측정하는지 보여주고 있다. 그러나 이 예제를 통한 설명은 몇 번의 시도에도 불구하고 오히려 학생들에게 혼란스러운 결과를 가져다 주고 있음을 경험할 수 있었다. 한편 상관계수가 −1과 1사이의

값을 갖음을 Hogg 등 (2005)의 연습문제 2.4.7 (107쪽) 즉 이차방정식의 판별식을 이용하면 보일 수 있다. 혹은 직접적으로 Cauchy-Schwarz 부등식을 이용하는 편이 학생들이 쉽게 이해할 수 있을 것이다.

그러나 경우에 따라 어떤 학생들은 이를 이해하는 데 조금의 어려움을 갖고 있는 것 같다. 따라서 Cauchy-Schwarz 부등식을 직접 설명하고 이에 따라 설명하는 것도 바람직하지만 그 대신 단위의 개념으로 설명하는 것도 나쁘진 않을 것이다. 즉 예를 들어 설명하자면 위의 예에서 하나의 수열쌍에다 10를 곱해서 교차곱의 합을 각각 구하고 이 두 값이 다르지만 배열된 방향의 정도는 같다는 점을 주지시키고, 다음으로 이런 단위의 역할을 표준편차가 하고 있다는 설명을 함으로써 위의 부등식을 이용한 설명과 같은 효과를 나타낼 수 있다. 이렇게 하면 교차곱의 합은 -1과 1사의의 값을 갖게 됨을 정확하지는 않지만 개념을 올바르게 이해시킬 수 있을 것이다.

3. 엑셀을 이용한 설명의 예

앞 절에서 논의한 상관계수의 개념 설명은 단순하게 수식을 이용하여 설명하는 것보다는 시각적인 설명을 곁들이는 것이 매우 효과적일 것이다. 이러한 시각적인 설명은 여러가지 도구를 이용할 수 있지만 아주 간단하게 엑셀을 사용하여 만들 수 있다. 다음의 그림 3.1은 표 2.1과 표 2.3을 엑셀을 이용하여 간단하게 보여주는 예이다. 이는 엑셀의 아주 기본 사용법만으로도 가능하기에 학생들과 같이 실습을 통하여 만들어 보는 것도 좋은 방법이 될 것으로 기대한다.

그림 3.2은 표 2.1의 내용을 산포도를 이용하여 설명하는 것이다. 교차곱의 합과 평균으로 조정한 후의 교차곱의 합을 표시하였으며 두개의 산포도를 이용한다. 화면 왼쪽의 산포도는 표 2.1에서 사용한 수열의 산포도를 나타낸다. 교차곱의 합을 설명하기 위

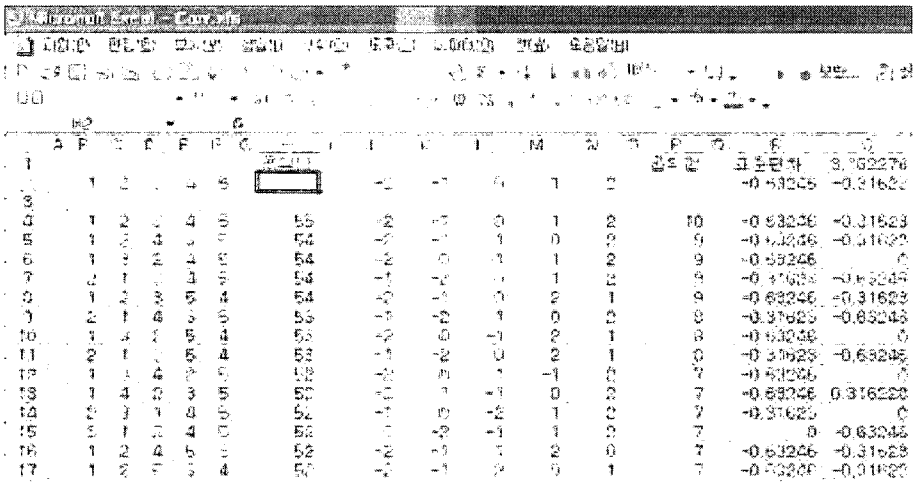

그림 3.1: 교차곱의 합의 비교

하여 수열의 크기를 5으로 작게 하였기 때문에 오른쪽의 그림은 실제적인 자료의 경우를 비교하기 위하여 해당 수열의 상관계수를 갖는 이변량정규분포의 난수를 발생시켜 그들로 산포도를 작성한 것이다. 이 과정에 있어 난수를 발생시키는 방법은 Casella와 Berger (2002)의 연습문제 4.46 (199쪽)를 이용할 수 있다. 이변량 정규분포의 평균은 각각 0으로 하고 분산은 각각 1로 하면 두 개의 독립된 표준정규분포로 부터 다음과 같이 상관계수가 ρ인 이변량정규분포를 얻을 수 있다. 단 Z_1과 Z_2는 서로 독립인 확률변수로 각각 표준정규분포를 따른다.

$$X = \sqrt{\frac{1+\rho}{2}}Z_1 + \sqrt{\frac{1-\rho}{2}}Z_2, \quad Y = \sqrt{\frac{1+\rho}{2}}Z_1 - \sqrt{\frac{1-\rho}{2}}Z_2$$

이 경우도 그림 3.1와 마찬가지로 간단한 수준의 엑셀작업이기에 학생들과 함께 작성하여 볼 수도 있을 것이다.

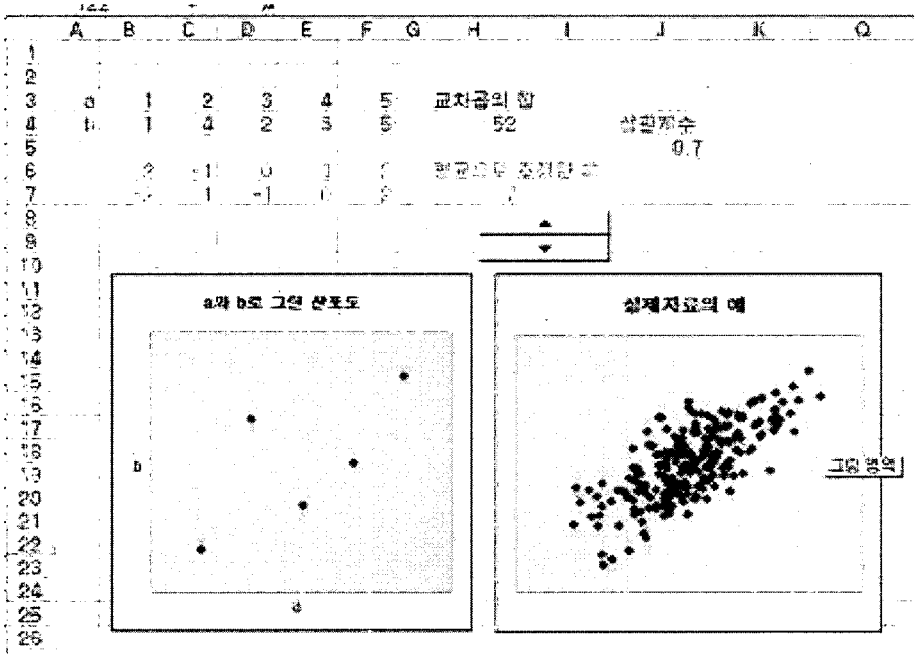

그림 3.2: 교차곱의 합과 상관계수와 실제자료의 예

4. 결론

최근의 기초통계학 교육은 앞서 지적한 바대로 수강생들의 수학 기초에 절대적으로 의존하게 되어 예전과 같은 수학을 바탕으로 한 교육은 넓은 의미로 말하자면 이론을 중심으로 한 교육은 등한시 되는 편이다. 어느 편이 효과적인 교육인지는 실상 많은 논란거리를 제공하고 있다. 그러나 현재와 같은 교육에서 통계적방법에 대한 개념교육이

효과적이라는 주장도 경험적으로 볼 때 그렇게 설득력을 가졌다고 볼 수는 없다. 특히 통계학의 상위과목으로 이동하게 되는 수강생들에게는 중요한 문제점으로 부각되고 있다. 예전과 같이 이론을 중심으로 교육으로의 회귀는 현재로서는 상당한 문제점을 갖고 있기에 이러한 문제점을 보완할 수 있는 교육방법의 적극적인 개발이 필요할 것이다. 여기에서 다룬 상관계수의 문제가 그러한 노력 가운데 하나의 일환이라고 생각한다.

참고문헌

최수일 (2006). 제7차 교육과정 고등학교 확률과 통계 교육의 문제점. <한국통계학회 2006년 춘계 학술발표회 논문집>, 91–97, 한국통계학회.

Casella, G. and Berger, R. L. (2002). *Statistical Inference.* 2nd ed., Duxbury, Pacific Grove.

Hardy, G., Littlewood, J. E. and Pólya, G. (1952). *Inequalities.* 2nd ed., Cambridge University Press, New York.

Hogg, R. V., McKean, J. W. and Craig, A. T. (2005). *Introduction to Mathematical Statistics.* 6th ed., Prentice Hall, Upper Saddle River.

Koch, G. G. (1985). A basic demonstration of the $[-1, 1]$ range for the correlation coefficient. *The American Statistician*, **39**, 201–202.

Leung, C.-K., and Lam, K. (1975). A note on the geometric representation of the correlation coefficients. *The American Statistician*, **29**, 128–130.

Rodgers, J. L. and Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, **42**, 59–66.

Rovine, M. J. and von Eye, A. (1997). A 14th way to look at a correlation coefficient: Correlation as the proportion of matches. *The American Statistician*, **51**, 42–46.

[Received February 2007, Accepted March 2007]