한국심리학회지: 임상
The Korean Journal of Clinical Psychology
2001. Vol. 20, No. 1, 179-194

변량분석사용의 잘못된 관행: F값만을 보고하는 경우, 사후비교에 의해 상호작용효과를 해석하는 경우, 그리고 표본의 크기가 작다는 이유로 비모수검증을 하는 경우*

박 광 배† 엄 진 섭

충북대학교 심리학과

본 논문은 한국에서 발표되는 논문들에서 자주 발견되는 변량분석과 실험자료 분석의 비바람직한 관행 세 가지에 대하여 주의를 환기하고, 그러한 관행이 시정되어야할 이유를 설명하기 위한 것이다. 한국에서 변량분석을 잘못 사용하는 가장 흔한 경우는 논문에 오직 F값만을 보고하는 것이다. F 값만을 보고한 연구들은 학문의 의사소통 기능을 제한하는 결과를 초래한다. 두번째로 자주 발견되는 오류는 Scheffé 검증이나 Tukey 검증 등의 사후비교(post-hoc comparison) 방법들에 의해 상호작용효과에 대한 해석을 시도하는 잘못된 관행이다. 셀평균간의 차이검증이나 혹은 소위 "단순주효과검증(test of simple main effect)"은 상호작용효과를 이해하기 위한 적절한 절차가 아니다. 보다 적절한 절차는 직교대비(orthogonal contrast)에 의한 계획비교(planned comparison)를 활용하는 것이다. 마지막으로, 표본의 관찰단위수가 작다는 이유로 t 검증이나 F 검증을 하는 대신 비모수검증(nonparametric test)을 하는 것은 비모수검증 뿐만 아니라 t 혹은 F 검증에 대한 이해부족에서 초래되는 잘못된 관행이다.

주요어 : 변량분석, 상호작용, 사후비교, 비모수검증

† 교신저자(Corresponding author) : 박 광 배 / 충북 청주시 개신동 산 48 충북대학교 심리학과 / E-mail : kwangbai@cbucc.chungbuk.ac.kr.

* 본 논고에서 지적하는 문제들은 전혀 새로운 문제들이 아니고 변량분석과 관련된 초보적이고 구태의연한 내용과 관련된 문제들이다. 필자들의 관점에서 볼 때, 가장 기초적이고 초보적인 내용들에서 오류가 자주 발견되므로 학계의 주의를 환기할 필요가 있다고 판단되어 본 논고를 집필하게 되었음을 밝힌다.

한국에서 많은 심리학 연구들이 변량분석 (Analysis of Variance) 을 활용하여 가설검증 및 자료분석을 하고 있다. 본 논문은 한국에서 발표되는 논문들에서 자주 발견되는 변량분석의 비바람직한 관행 세 가지에 대하여 주의를 환기하고, 그러한 관행이 시정되어야할 이유를 설명하기 위한 것이다. 변량분석을 잘못 사용하는 가장 흔한 경우는 논문에 오직 F 값만을 보고하는 것이다. 두번째 경우는 Scheffé 검증이나 Tukey 검증 등의 사후비교 (post-hoc comparison) 방법들에 의해 상호작용효과에 대한 해석을 시도하는 잘못된 관행이다. 세번째 경우는 표본의 관찰단위수가 작다는 이유로 t 검증이나 F 검증을 하는 대신 비모수검증 (nonparametric test)을 하는 것이다.

1995년부터 1999년까지 5년 동안 한국심리학회 산하의 분과학회에서 발간하는 11종의 학술지들에 총 746편의 논문들이 발표되었다. 이 논문들중 변량분석에 의한 자료분석결과를 보고한 논문의 수와 F값만 보고한 논문의 수, 사후비교 방법들에 의해 상호작용효과에 대한 해석을 시도한 논문의 수, 그리고 표본의 관찰단위수가 작다는 이유로 비모수검증을 한 논문의 수를 집계해본 것이 표 1부터 표 3까지 제시되었다.

먼저, F 값만을 보고하는 것이 바람직하지 않은 이유를 설명하고, 사후비교의 의미와 상호작용효과의 의미를 각각 설명한 후, 상호작용효과를 해석하기 위하여 사후비교를 이용해서는 안되는 이유를 밝히고,

표 1. 심리학회 발간 11 종 학회지에 1995년부터 1999년까지 발표된 전체논문과 변량분석에 의한 자료분석 결과를 보고한 논문의 수.

전체논문 편수	변량분석을 사용한 논문편수	변량분석표를 제시한 논문편수	MSE 혹은 ω^2만 제시한 논문편수	F와 p만 제시한 논문편수
746	362	56	31	275

주) MSE 혹은 ω^2만 제시한 논문은 대부분 유의미한 결과가 있는 효과만 제시하였음.

표 2. 심리학회 발간 11 종 학회지에 1995 년부터 1999 년까지 발표된 전체논문과 상호작용효과를 보고한 논문의 수.

전체 논문 편수	다원변량 분석을 사용한 논문편수	상호작용 효과가 유의미한 논문편수	그림이나 평균으로 해석한 논문편수	해석을 하지 않은 논문편수	단순주효과 혹은 단순 상호작용 효과를 검증한 논문편수	쌍비교 혹은 분리된 변량분석을 수행한 논문편수	상호작용효과가 없는데도 사후비교를 한 논문편수
746	213	156	51	6	49	47	18

주) 단순주효과 혹은 단순상호작용효과라고 명시하지 않고 추가분석을 수행한 경우는 쌍비교 혹은 분리된 변량분석으로 포함시켰음.

표 3. 심리학회 발간 11 종 학회지에 1995 년부터 1999 년까지 발표된 전체논문과 비모수검증에 의한 자료분석 결과를 보고한 논문의 수.

전체논문 편수	비모수검증을 사용한 논문편수	타당한 이유를 제시한 논문편수	사례수가 적어서 비모수검증을 사용한 논문편수	이유를 제시하지 않은 논문편수
746	4	1	1	2

직교대비(orthogonal contrast)에 의한 계획비교(planned comparison) 를 활용할 수 있는 상황과 방법을 기술하고자 한다. 마지막으로, 표본의 관찰단위수가 작다는 이유로 t 검증이나 F 검증을 하는 대신 비모수검증 (nonparametric test) 을 하는 것이 옳지 않은 이유를 간단히 설명하고자 한다.

F 값만을 보고하는 관행의 폐단

F 값만을 보고하는 연구관행의 폐단은 단적으로 말해서 학문의 커뮤니케이션 기능을 제한시킨다는 것이다. 그 이유는 최소한 3 가지로 요약할 수 있다. 첫째는 연구결과에 대한 해석이 극심하게 추상화된다는 것이고, 둘째는 서로 다른 연구들의 결과를 서로 비교할 수 없다는 것이며, 세번째로는 연구자가 자신의 실험적 조작에 대하여 적용하는 묵시적 가정을 다른 연구자가 파악할 수 없다는 것이다.[1]

연구결과에 대한 궁극적 해석은 모수추정치에 의해 이루어진다.

상관연구들은 상관계수를 보고하고, 그 상관계수의 통계적 의미를 판단할 수 있는 t, 혹은 F 값을 보고하는 것이 일반적 관행이다. 회귀분석을 이용하는 연구들은 회귀계수 혹은 결정계수를 보고하고, 그 계수들의 t 값 혹은 F 값을 보고하는 것이 일반적 관행이다. 이 연구들에서 상관계수 혹은 회귀계수들을 보고하는 첫째 이유는 그 계수들이 바로 연구자가 검증한 이론적 모형의 내용을 기술하는 모수추정치들이기 때문이다. 관계의 정도를 파악하기 위하여 상관분석 혹은 회귀분석을 하였다면 파악된 관계의 정도를 최우선적으로 보고하는 것은 너무도 당연한 일이다. 상관계수, 회귀계수, 결정계수들은 변인들 사이의 관계의 정도를 나타낸다. 관계의 정도가 강한지 약한지의 여부는 그 계수들의 유의도 판단을 위한 t 값이나 F 값에 의해 판단할 수 없다. 왜냐하면 t 값이나 F 값 등의 통계적 지수들은 관계의 정도뿐만 아니라 자료의 양 즉, 오차의 자유도에 의해서 좌우되기 때문이다. 상관분석과 회귀분석에서 계수들을 보고하지 않는 것은 연구자가 가설모형을 애매하게 제시하고 끝내 그 모형의 구체적 내용을 밝히지 않는 것과 같다.

변량분석도 상관분석이나 회귀분석과 마찬가지로 일반선형모형의 일종이다. 즉, 독립변인과 종속변인의 관계를 파악하기 위한 통계적 모형이다. 변량분석에서 독립변인과 종속변인 사이의 관계는 소위 "효과크기(effect size)"라고 불리우는 η 혹은 η^2로 표현되며, 이 지수들이 바로 변량분석모형의 모수추정치이다(Cohen (1969)은 표준화된 평균차이를 효과크기로 정의하고 있는데, η는 바로 표준화된 평균차이다. 따라서 엄밀한 의미에서 η만을 효과크기로 불러야하지만 η^2는 단순히 η를 자승한 것이므로 역시 효과크기로 간주된다). η^2는 전체자승합(SST)에 대한 집단간자승합(SSB)의 비율이다. 즉, 종속변인의 전체 변산 중 특정한 독립변인에 의해, 혹은 독립변인들 사이의 상호작용에 의해 설명되는 부분의 비율이다. 따라서 변량분석표가 논문에 제시되어 있으면 그 표에 제시된 자승합들을 이용하여 독자들은 매우 쉽게 η 혹은 η^2를 산출할 수 있다.

상관계수, 회귀계수, 결정계수, η^2 등의 모수추정치들은 자료를 수학적으로 요약한 것이며, 주어진 자료를 초월하여 존재하는 통계적 이론이나 가정 혹은 추상적 개념에 기초한 것이 아니다. R. A. Fisher의 유일한 박사학위 제자였던 Rao(1992)에 의하면, Fisher는 모수추정을 "자료축약(reduction of data)"의 한 방

1) F-검증과 관련하여 여기에서 논의하는 문제와는 촛점이 약간 다르긴 하지만, F-검증을 비롯하여 모든 통계검증에서의 유의도 검증에 관련된 이념적 논의가 1997 년에 Psychological Science 에서 심층적으로 다루어진 바 있다(Shrout, 1997; Hunter, 1997; Harris, 1997; Abelson, 1997; Estes, 1997; Scarr, 1997). 논의의 골자는 심리학 연구들에서 유의도 검증을 전면 폐지해야하는가의 문제이다. 이러한 주장이 제기되는 마당에서 F 값과 p 값만을 보고하는 관행은 시급히 시정되어야할 것으로 사료된다. 본 논고에서는 유의도 검증에 관련된 이념적 논의와는 별도로, F 값과 p 값만을 보고하는 관행의 보다 더 근본적인 문제점을 지적하고자 한다.

법으로 간주하였다고 한다. 예를 들어 η^2는 SSB와 SST의 단순비율이고, SSB와 SST는 각각 자료수치들과 그들의 평균에 의해 계산되는 숫자들에 불과하다. η^2를 산출하는 과정에는 어떠한 통계적 이론이나 가정도 전혀 개입하지 않는다. 이러한 모수추정치들을 보고하는 것은 연구의 특정한 목적에 부합하도록 자료 그 자체를 일목요연하게 정리하고 요약하여 제시하는 것이라고 볼 수 있다. 말하자면 상관계수, 회귀계수, 결정계수, η^2 등의 모수추정치들은 연구의 특정한 목적에 부합하도록 요약된 자료 그 자체이고, 따라서 이 모수추정치들에 대한 해석도 철저하게 자료의 속성에 준하여 이루어지면 된다. 예를 들어, 회귀계수의 의미는 독립변인에서의 한단위 변화에 수반되는 종속변인값의 평균적 단위변화이다. 연구자가 회귀계수를 이해하고 해석하기 위해서는 자신이 사용한 변인들의 단위 이외에 어떠한 다른 개념도 필요하지 않다. η^2의 의미는 더 간단하다: 종속변인의 변산중 독립변인과 공변하는 부분의 비율이다.

반면에 t 혹은 F의 의미는 고도로 관념화된 통계적 이론들에 기초하고 있고, 그 지수들은 자료 자체가 아니라 현실성이 불확실한 여러 가지 가정들의 전제하에 모집단의 추상적인 상태를 반영한다. 그런데 대부분의 연구에서 모집단의 정체가 불분명하므로, 결과적으로 t 혹은 F의 의미는 극도로 추상적이고 유보적일 수 밖에 없다. 변량분석의 결과를 보고하면서 자승합들에 대한 기술없이 오로지 F값만 보고하는 경우는 말하자면 자료를 보고하지 않는 것과 같고, 따라서 F값의 의미를 구체적으로 해석할 수 없게 된다.

F 값만을 보고하는 연구들은 서로 비교될 수 없다.

서로 다른 연구의 결과들이 서로 비교될 수 있는 것은 모수추정치들에 의해 가능한 것이며, 그들의 통계적 유의미도에 의한 것이 아니다. 예를 들어 어떤 연구자가 발견한 두변인 사이의 상관계수가 0.80 이었고, 기존의 다른 연구에서 발견된 동일한 변인들 사이의 상관계수가 0.75 이었다면, 이 연구자는 자신의 연구결과가 기존의 연구결과와 대략 일치한다는 것을 알 수 있다. 반면에 통계적 유의미도를 판단하기 위한 t 값이나 F 값은 관계의 정도 뿐만 아니라 자료의 양에 의해서도 달라지는 지수들이므로 t 값이나 F 값에 의해서는 두개의 서로 다른 연구들에서 발견된 결과들이 서로 유사한 것들인지, 다른 것들인지를 판단할 수 없다.

η^2는 전체자승합(SST)에 대한 집단간자승합(SSB)의 비율인데, 분모가 특정한 연구에서 채택된 독특한 실험조작과는 무관한 전체자승합이므로, 동일한 변인들을 사용한 서로 다른 연구들에서 계산된 η^2는 서로 비교될 수 있다. 이 이유로 인하여 F 값만을 보고하는 실험연구는 메타분석(meta-analysis)에 포함될 수 없게 된다. 근자에 들어 메타분석의 중요성이 점차로 광범위하게 인식되고 있다. 그 이유는 한 개의 연구에 의해 어떤 이론적 쟁점에 관하여 결정적인 결론을 내릴 수 있는 경우가 거의 없기 때문이다. 어떤 연구에서 특정한 결과가 도출되었더라도 그 결과가 신뢰받기 위해서는 유사한 연구들이 반복적으로 비슷한 결과들을 보여주어야 한다. 이것은 소위 "복제가능성(replicability)"이라고 불리우는 과학의 중요한 기본조건중의 하나이다. 복제가능성이란 특정한 연구가 모든 절차 및 속성에서 동일하게 반복되면 동일한 결과가 나와야한다는 원칙이다. 그러나 현실적으로 어떤 연구를 모든 절차 및 속성에서 동일하도록 반복한다는 것은 불가능한 일이다. 복제가능성의 이념을 현실적으로 구현하기 위한 방법이 메타분석이다. 메타분석은 각기 다른 분석방법에 의하여 수행된 다양한 연구들의 결과를 모두 효과크기로 변환한 후, 그 효과크기들을 취합하여 많은 연구들의 결과들에서 공통적으로 발견되는 경향성을 도출하는 기법이다. 그런데 변량분석을 이용한 연구가 F 값만을 보고하는 경우에는 그것을 효과크기로 변환할 수 없고, 결과적으로 메타분석에 포함될 수 없게 된다. 변량분석의 결과를 효과크기로 전환하기 위해서는 자승합들이 보고되거나 η 혹은 η^2가 반드시 보고되어야 한다. 메타분석에 포함될 수 없다는 것은 다른 연구들과의 비교가 어렵게 된다는 것을 의미하고, 궁극적

으로는 학문발전에 대한 기여의 폭이 극히 제한된다는 것을 의미한다.

실험조건들에 대한 묵시적 가정에 따라 F 값은 달라진다.

실험자료의 경우, 연구자가 독립변인의 수준들에 대하여 어떤 묵시적 가정을 하느냐에 따라서 F 값이 달라지고, 독립변인 효과의 통계적 유의미도가 뒤바뀔 수 있다. 실험에서 조작된 독립변인의 수준들이 무수히 많은 수준들 가운데에서 무작위로 추출된 것이라고 가정하는 경우를 무선모형 (random model) 이라고 부르고, 이때에는 독립변인의 무수히 많은 수준들중 실험에 포함되지 않은 다른 수준들에도 분석결과가 일반화되는 것으로 해석한다. 반면에 실험에서 조작된 독립변인의 수준들이 그 독립변인이 가질 수 있는 수준들을 모두 망라한 것으로 가정하는 경우를 고정모형 (fixed model) 이라고 부르고, 이때에는 분석결과가 그 수준들에만 적용되는 것으로 해석하여야 한다. 그런데 문제는 어떤 독립변인에 대하여 무선모형을 가정해야 하는지, 아니면 고정모형을 가정해야 하는지의 여부는 분석결과를 어떻게 일반화할 것인지를 결정하는 연구자에게 전적으로 달려있다는 것이다 (Neter & Wasserman, 1974, p. 525). 예를 들어 거의 동일한 연구를 수행하는 2 명의 연구자들중 한명은 자신의 연구에서 이용된 독립변인에 대하여 고정모형을 가정하는 반면, 또 다른 연구자는 동일한 독립변인에 대하여 무선모형을 가정할 수 있다. 이렇게 독립변인의 수준들에 대한 연구자들의 묵시적 가정이 다르면, 거의 동일한 자료에 있어서도 독립변인의 효과를 검증하기 위한 F 값은 거의 언제나 달라질 뿐만 아니라, 한 연구에서는 통계적으로 유의미하고, 다른 연구에서는 유의미하지 않을 수 있다. 연구자들이 만약 최종보고서 혹은 논문에서 F 값만을 보고한다면, 이러한 혼란을 교정할 수 없게 된다. 그 이유를 보다 명확히 이해하기 위하여 평균자승의 기대값을 우선 살펴보자.

두개의 독립변인을 가지는 교차요인설계에서 요인 A가 c개의 수준을 가지고, 요인 B가 r개의 수준을 가진다고 가정하자. 요인 A에 의한 편차 ($\mu_{i.} - \mu_{..}$), 요인 B에 의한 편차 ($\mu_{.j.} - \mu_{..}$), 상호작용에 의한 편차 ($\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..}$), 집단내 편차 ($Y - \mu_{ij}$)를 전체 모집단에 있는 모든 구성원에 대해서 산출하는 것을 상상해보자. 이 편차들의 모집단 변량을 각각 σ_A^2, σ_B^2, σ_{AB}^2, σ_e^2 라고 표기하자. 마지막으로, (c)x(r) 개의 모집단 각각에서 n 명씩의 피험자를 표집하여 종속변인을 측정하고 변량분석을 하면 네 개의 평균자승들(MSB_A, MSB_B, MSB_{AB}, MSW) 이 구해진다. 이 표집과정과 변량분석을 무수히 여러번 반복하여 각 변량분석에서 산출된 평균자승들의 평균을 구하면 그들이 평균자승의 기대값들인바, 다음과 같은 구성요소들로 이루어진 것이다.

$$E(MSB_A) = \sigma_e^2 + \frac{R-r}{R} n\sigma_{AB}^2 + nR\sigma_A^2$$

$$E(MSB_B) = \sigma_e^2 + \frac{C-c}{C} n\sigma_{AB}^2 + nC\sigma_B^2$$

$$E(MSB_{AB}) = \sigma_e^2 + n\sigma_{AB}^2$$

$$E(MSW) = \sigma_e^2$$

위의 수식에서 대문자 C와 R은 각각 독립변인 A와 B의 모든 가능한 수준들의 갯수이다. 위의 기대값들에서 우리가 이해해야할 사항은 다음과 같다. ① 요인 A의 집단평균들의 변산은 요인 A 자체의 효과(σ_A^2), 상호작용효과(σ_{AB}^2), 그리고 우연적인 개인차(σ_e^2)에 의해 유발될 것으로 기대된다; ② 요인 B의 집단평균들의 변산은 요인 B 자체의 효과(σ_B^2), 상호작용효과(σ_{AB}^2), 그리고 우연적인 개인차(σ_e^2)에 의해 유발될 것으로 기대된다; ③ 각 집단평균에서 A-효과 및 B-효과를 제거한 나머지 부분의 변산 즉, 상호작용에 의한 평균의 변산은 상호작용효과(σ_{AB}^2), 그리고 우연적인 개인차(σ_e^2)에 의해 유발될 것으로 기대된다; ④ 각 집단내의 개별적인 종속측정치들의 변산

즉, 집단내 변산은 우연적인 개인차(σ_e^2)에 의해 유발될 것으로 기대된다.

위에서 설명한 평균자승의 기대값은 각 효과의 통계적 유의미도를 결정하는데 매우 중요하다. 이 유의미도 결정은 연구자가 자신의 실험을 어떻게 개념화하느냐에 따라 달라진다. 연구자가 만약 자신의 실험에 이용된 각 독립변인의 조건들이 모든 가능한 조건들을 망라하고 있거나, 연구의 관심과 연구결과의 일반화가 오직 실험에 이용된 조건들에만 국한된다고 개념화하는 경우에는, (C=c)이고 (R=r)이라고 가정하는 것과 같다. 이것을 고정모형(fixed model)이라고 부른다. 즉, 독립변인의 조건들이 고정적이고 같은 가설을 다른 사람이 혹은 다른 상황에서 검증하는 경우에도 동일한 조건들이 설정되는 모형이다.

반면에 만약 연구자가 자신의 독립변인이 실제로는 매우 많은 수준들을 가지는데, 다만 그 모든 수준들을 한 개의 실험에서 모두 조작하여 실험집단을 구성할 수 없으므로, 그중 일부의 조건들만 무작위로 추출하여 이용했다고 가정하는 경우에는 (C > c)이고 (R > r)이다. 이러한 경우를 무선모형(random model)이라고 부른다. 만약 한 개의 독립변인은 수준이 고정되었다고 가정되고, 다른 한 개의 독립변인은 수준들이 무작위로 추출되었다고 가정하는 경우에는 혼합모형 (mixed model) 이라고 부른다.

자, 이제 연구자가 자신의 실험이 어떤 모형에 준하는지를 결정하는데 따라서 평균자승의 기대값이 어떻게 달라지는지 살펴보자. 우선 고정모형을 가정하는 경우에는 C=c이고 R=r이므로 C-c=0이고 R-r=0이 된다. 따라서 평균자승의 기대값들은 다음과 같이 변한다.

$$E(MSB_A) = \sigma_e^2 + nR\sigma_A^2$$

$$E(MSB_B) = \sigma_e^2 + nC\sigma_B^2$$

$$E(MSB_{AB}) = \sigma_e^2 + n\sigma_{AB}^2$$

$$E(MSW) = \sigma_e^2$$

반면에 무선모형을 가정하는 경우에는 C에 비하여 c가 매우 작고, R에 비하여 r이 매우 작으므로 (C-c)/C와 (R-r)/R은 모두 1에 근접한다. 따라서 평균자승의 기대값들은 다음과 같이 기술된다.

$$E(MSB_A) = \sigma_e^2 + n\sigma_{AB}^2 + nR\sigma_A^2$$

$$E(MSB_B) = \sigma_e^2 + n\sigma_{AB}^2 + nC\sigma_B^2$$

$$E(MSB_{AB}) = \sigma_e^2 + n\sigma_{AB}^2$$

$$E(MSW) = \sigma_e^2$$

만약 요인 A의 수준들은 고정되었고, 요인 B 의 수준들은 무작위로 추출되었다고 가정하는 혼합모형이 있다면, C=c이고, (R-r)/R=1이므로, 그 혼합모형의 평균자승들은 다음과 같이 기대된다.

$$E(MSB_A) = \sigma_e^2 + n\sigma_{AB}^2 + nR\sigma_A^2$$

$$E(MSB_B) = \sigma_e^2 + nC\sigma_B^2$$

$$E(MSB_{AB}) = \sigma_e^2 + n\sigma_{AB}^2$$

$$E(MSW) = \sigma_e^2$$

평균자승의 기대값에 대한 논의에서 우리는 이원변량분석에서 각 효과의 평균자승 기대값은 그 효과 자체와 다른 요소들의 합으로 이루어진다는 것을 알 수 있다. 따라서 유의미도 검증은 특정한 효과가 "다른 요소들" 에 비해서 큰 정도에 의해서 판단된다. 예를 들어 고정모형의 경우, 요인 A의 평균자승 MSB_A는 $\sigma_e^2 + nR\sigma_A^2$로 구성될 것이 기대된다. 그런데 $nR\sigma_A^2$는 요인 A 자체의 효과이고, 반면에 σ_e^2는 오차변산이다. 다행히도 σ_e^2의 추정치를 계산할 수 있는데, 바로 MSW가 σ_e^2일 것으로 기대되기 때문이다. 따라서 MSB_A와 MSW를 서로 비교하여 후자에 비해 전자가 많이 크면, 우리는 요인 A의 효과가 유의미하다는 판단을 할 수 있다. 이 비교를 하는 한 방식이 바로 F 를 산출하는 것이다.

각 효과를 위한 F 는 연구자가 자신의 독립변인들이 가지는 수준을 어떻게 간주하느냐에 따라서 다르게 산출된다. 고정모형의 경우에는 A-효과, B-효과, AB-효과를 위한 F 는 모두 그들의 평균자승을 MSW로 나누어주어 산출한다. 반면에, 무선모형의 경우에는 A-효과와 B-효과를 위한 F는 그들의 평균자승을 MSB_{AB}로 나누어주어 산출하여야 적절한 비교가 되고, 상호작용효과를 위한 F는 MSB_{AB}를 MSW로 나누어주어 산출한다. 동일한 논리적 과정에 의하여 혼합모형의 효과들에 대한 유의도 판단을 해야한다.

A와 B의 독립변인을 가지는 어떤 실험자료에 대하여 독립변인의 수준에 대한 가정을 각기 달리하여 변량분석을 실시하고 F 값을 구하면 어떤 결과가 초래되는지를 예시하기 위하여 다음의 예를 살펴보자.

표 4를 보면, 독립변인의 수준들에 대하여 고정모형을 가정하건 무선모형을 가정하건 자승합과 평균자승에는 전혀 차이가 발생하지 않는다. 그러나 F 값은 크게 달라질 수 있고, 또한 유의미도 결정도 정반대로 달라질 수 있다.

동일한 독립변인에 대하여 고정모형을 가정하느냐 아니면 무선모형을 가정하느냐의 여부는 결과의 일반화를 어떻게 할 것인가에 의해 좌우된다. 그런데 어떤 연구자가 고정모형을 가정하고 변량분석을 하고 F 값만을 보고하였다면, 무선모형을 가정하는 경우는 어떤 결과가 산출될지에 관심을 가지는 다른 연구자가 그 결과에 대하여 전혀 아무런 추정을 할 수 없게 된다. 또한 만약 어떤 독립변인에 대하여 일부의 연구들은 무선모형을 가정하고 다른 연구들은 고정모형을 가정하는 상황에서 그 연구들의 결과를 일관성있게 통합하기 위해서는 변량분석표가 완벽하게 보고되는 것이 필수적이다. F 값만을 보고한 연구들은 말하자면 학문의 커뮤니케이션 기능을 심각하게 제한하는 결과를 초래한다. 특히, 연구자가 독립변인에 대하여 고정모형을 가정한 것인지, 아니면 무선모형을 가정한 것인지의 여부를 연구보고에 명시하지 않고 F 값만을 보고하는 경우에는 그 F 값의 의미가 전혀 파악될 수 없다는 점을 명심할 필요가 있다.

표 4. 모형의 가정에 따른 변량분석 결과의 차이.

고정모형을 가정하는 경우

변산원	자승합	자유도	평균자승	F
(A)	4482.6	2	2241.3	20.49**
(B)	2764.8	1	2764.8	25.27**
(A)x(B)	8145.0	2	4072.5	37.23**
집단내	2626.4	24	109.4	
전체	18018.8	29	621.3	

무선모형을 가정하는 경우

변산원	자승합	자유도	평균자승	F
(A)	4482.6	2	2241.3	0.55
(B)	2764.8	1	2764.8	0.68
(A)x(B)	8145.0	2	4072.5	37.23**
집단내	2626.4	24	109.4	
전체	18018.8	29	621.3	

혼합모형을 가정하는 경우 (요인 A 는 무선, 요인 B 는 고정)

변산원	자승합	자유도	평균자승	F
(A)	4482.6	2	2241.3	20.49**
(B)	2764.8	1	2764.8	0.68
(A)x(B)	8145.0	2	4072.5	37.23**
집단내	2626.4	24	109.4	
전체	18018.8	29	621.3	

사후비교에 의해 상호작용효과를 해석하는 오류

사후비교

일원변량분석(one-way analysis of variance) 혹은 특정한 요인의 주효과(main effect)에 대한 전반적 F 검증(omnibus F test)이 유의미하면 모집단 평균들이 모두 동일하다는 영가설이 기각된다. 평균차이들에 대한 사전가설 없이, 전반적 검증 결과 집단간 차이가 유의미한 것으로 파악된 후, 어떤 평균들이 서로 차이가 나는지를 규명하기 위하여 평균들을 쌍으로 혹은 묶어서 비교하는 경우에는, 그 차이에 대한 판단에서 제1종 오류를 범할 확률이 매우 높아질 수 있다. 예를 들어서 a_1, a_2, a_3의 세 집단을 가진 실험에

서 집단차이에 대한 일원변량분석을 수행한 결과 집단간 차이의 F 값 (전반적 F) 이 유의미한 것으로 나타났다면, 그러한 결과는 a_1-a_2의 평균차이, a_1-a_3의 평균차이, a_2-a_3의 평균차이, (a_1+a_2)-a_3의 평균차이, (a_1+a_3)-a_2의 평균차이, (a_2+a_3)-a_1 의 평균차이 등 모두 6 개의 평균차이 중 어떤 것에 의해서도 발생할 수 있고, 따라서 그들 중 어떤 평균차이가 유의미한 것인지를 파악하게 된다.

한 세트의 실험자료를 이용하는 여러 개의 검증에서 적어도 한 번 제1종 오류를 범할 확률을 Experiment-wise α라고 부른다. 평균차이를 파악하기 위한 각각의 검증에서 제1종 오류를 범할 확률 α를 0.05로 정해놓으면, 6개의 평균차이 검증에서 적어도 한번 제1종 오류를 범할 확률은 다음과 같다.

$$\text{Experiment-wise } \alpha = 1-(0.95)^6 = 0.26$$

위의 수식에서 $(0.95)^6$는 6개의 검증에서 판단오류를 한 번도 범하지 않을 확률이다. 이 확률을 1에서 빼면 적어도 1번의 판단오류를 범할 확률이 되는데, 검증을 6번 하는 경우 이 Experiment-wise α는 0.05 보다 훨씬 높은 0.26 이다. 사후비교를 위한 여러 가지 방법들은 이 Experiment-wise α가 0.05를 넘지 않도록 통제하면서 평균차이를 검증하기 위한 것들이다. 가장 널리 쓰이는 방법들은 Scheffè 검증과 Tukey 검증이다. 이 방법들은 각기 다른 비교상황을 가정한다.

Scheffè 검증의 특징은 두 개의 집단평균들 사이의 차이에 대한 검증 뿐만 아니라, 모든 유형의 대비에 의해서 이루어지는 비교들에도 적용된다는 것이다. 예를 들어, 실험집단이 5 개인 경우 그중 3 개의 집단을 나머지 2 개의 집단과 비교하는 대비에 대해서도 Scheffè 검증을 활용할 수 있다. 반면에 다른 사후비교 방법들은 일반적으로 두 개의 집단평균들 사이의 차이검증에 적용된다. Scheffè 검증은 사전과 사후를 막론하고 모든 다중비교 방법들중에 가장 보수적인 방법이다. 즉, 특정한 대비가 유의미하다는 판단을 하게될 가능성이 가장 낮은 검증방법이다. Scheffè 검증이 보수적인 이유는 g개의 실험집단에 대하여 모든 가능한 비교와 대비분석을 다 수행하는 경우에 Experiment-wise α가 0.05 혹은 0.01을 초가하지 않도록 하려는데 목적이 있기 때문이다.

집단평균들의 사후비교를 위한 Tukey 검증은 HSD (honestly significant difference) 검증이라고 불리우기도 하는데, 집단평균들을 두 개씩 비교하는데 적용되는 방법이다. 따라서 이 방법은 Scheffè 검증에 비하여 적용범위가 제한적이다. 그러나 평균쌍들의 비교를 위해서는 Scheffè 검증보다 검증력 (power)이 높은 방법이다 (그러나 사전검증보다는 검증력이 낮다). 만약 실험집단이 g개 있다면, g(g-1)/2개의 평균쌍에 대한 비교를 할 수 있다. Tukey 검증은 모든 가능한 평균쌍의 비교를 하는 경우 Experiment-wise α가 0.05 혹은 0.01이 넘지않도록 통제하는 방법이다.

Scheffè 검증과 Tukey 검증 이외에 Newman-Keuls 검증과 Duncan 검증이 간혹 사후비교를 위하여 이용된다. 이 두 가지 검증은 Tukey 검증의 변형이며, Tukey 검증과 마찬가지로 평균쌍들의 차이에 대한 통계적 판단을 한다. 그런데 Tukey 검증은 집단평균들 사이의 모든 가능한 쌍비교를 가정하고, 모든 가능한 쌍비교를 하는 경우에 그 비교들의 Experiment-wise α를 적정수준에 고정시키기 위하여 한 개의 임계치를 설정한 후, 모든 평균차이를 이 임계치와 비교하여 유의도 판단을 한다. 반면에 Newman-Keuls 검증과 Duncan 검증은 비교되는 평균쌍이 실험내에서 어떤 서열관계에 있는가에 따라서 각기 다른 임계치를 적용하는 방법이다. 따라서 이 두 검증에서는 우선 집단평균들이 가장 작은 것부터 가장 큰 것 순서로 배열되고, 비교되는 평균쌍들이 서로 몇 개의 평균들을 사이에 두고 떨어져 있는지에 따라서 임계치가 결정된다. 따라서 이 두 검증을 하는 경우에는 Tukey 검증을 하는 경우에 비하여 일반적으로 임계치가 낮게 설정되고, 결과적으로 유의미한 평균차이가 더 많이 나타난다. Newman-Keuls 검증과 Duncan 검증은 Experiment-wise α 를 정하는데에서 서로 차이가 난다. 만약 Newman-Keuls 검증이 Experiment-wise α = 0.01 수준에서 이루어진다면, 같은 검증을 Duncan

검증에 의해서 하면 Experiment-wise $\alpha = 0.02$ 혹은 Experiment-wise $\alpha = 0.03$ 수준에서(비교되는 평균쌍들이 몇 개의 평균들을 사이에 두고 떨어져 있는가에 따라서 다름)하게 된다. 결론적으로 특정한 평균쌍 사이의 차이가 유의미한 것으로 결론지어질 가능성은 Scheffè 검증이 가장 낮고, 그 다음이 Tukey 검증, Newman-Keuls 검증, Duncan 검증의 순서를 가진다.

상호작용효과

상호작용효과의 수학적 의미는 실험설계행렬에서 대각선평균의 차이이다. 즉, 상호작용효과는 수학적으로 다음과 같이 정의된다.

요인 A

요인 B	a_1	a_2	
b_1	$\overline{Y}_{11}$	$\overline{Y}_{21}$	$\overline{Y}_{.1}$
b_2	$\overline{Y}_{12}$	$\overline{Y}_{22}$	$\overline{Y}_{.2}$
	$\overline{Y}_{1.}$	$\overline{Y}_{2.}$	$\overline{Y}_{..}$

요인 A 와 B 의 상호작용효과:

$$\frac{(\overline{Y}_{11} + \overline{Y}_{22})}{2} - \frac{(\overline{Y}_{12} + \overline{Y}_{21})}{2}$$

그런데 상호작용효과의 이 수학적 의미는, "상호작용효과"라는 개념이 독립변인들이 서로 결합하여 이루어내는 효과인데, 그 결합하는 방식이 합산적(additive)이 아니라 곱셈적 (multiplicative) 이라는 것을 뜻한다. 곱셈적인 결합효과를 이해하기 위하여, 요인 A와 요인 B의 주효과, 그리고 상호작용효과를 대비로 표현하면 표 5와 같다.

위에서 사용된 대비가중치들을 표로 제시하면 표 6과 같다.

AB-효과를 위한 대비의 가중치를 보면 A-효과를 위한 대비가중치와 B-효과를 위한 대비가중치의 곱이라는 것을 알 수 있다. 따라서 대비 Ψ_3로 표현되는 상호작용효과는 두 개의 독립변인들의 곱셈적인 결합에 의해 생성되는 효과이다.

상호작용효과의 해석적 의미는 "한 요인의 효과가 다른 요인의 각 수준에서 동일하지 않은 것"을 의미한다. 앞서 상호작용효과의 수학적 의미는 실험설계행렬에서 대각선평균의 차이로 정의하였고, 그 수학적 정의가 위에서 설명되었다. 그런데 상호작용효과의 이 수학적 정의는 다음과 같이 재정의할 수 있다.

$$\frac{(\overline{Y}_{11} + \overline{Y}_{22})}{2} - \frac{(\overline{Y}_{12} + \overline{Y}_{21})}{2}$$

$$= \frac{(\overline{Y}_{11} - \overline{Y}_{12})}{2} - \frac{(\overline{Y}_{21} - \overline{Y}_{22})}{2}$$

↙ A의 첫 번째 수준 (a_1) 에서의 B-효과

↘ A의 두 번째 수준 (a_2) 에서의 B-효과

표 6. 대비가중치로 표현된 주효과와 상호작용효과.

	대비가중치(contrast weights)		
집단	A-효과	B-효과	AB-효과
a_1b_1	1	1	1
a_1b_2	1	-1	-1
a_2b_1	-1	1	-1
a_2b_2	-1	-1	1

표 5. 대비로 표현된 주효과와 상호작용효과.

효 과	대 비
A	$\Psi_1 = (1)\ \overline{Y}_{11} + (1)\ \overline{Y}_{12} + (-1)\ \overline{Y}_{21} + (-1)\ \overline{Y}_{22}$
B	$\Psi_2 = (1)\ \overline{Y}_{11} + (1)\ \overline{Y}_{21} + (-1)\ \overline{Y}_{12} + (-1)\ \overline{Y}_{22}$
AB	$\Psi_3 = (1)\ \overline{Y}_{11} + (1)\ \overline{Y}_{22} + (-1)\ \overline{Y}_{12} + (-1)\ \overline{Y}_{21}$

따라서 상호작용효과의 해석적 의미는 한 독립변인과 종속변인의 관계가 다른 독립변인의 각 수준에서 동일하지 않다는 것이다. 만약 각 독립변인이 2개씩의 수준을 가진다면, 상호작용효과의 해석적 의미는 "한 독립변인과 종속변인의 상관관계가 다른 독립변인의 수준에 따라 달라진다" 라고 표현될 수 있다. 즉, 상호작용효과란 평균차이를 의미하는 것이 아니고, 평균차이의 차이 혹은 관계의 차이를 의미한다.

사후검증에 의하여 상호작용효과를 해석하는 오류

Scheffè 검증과 Tukey 검증 등의 사후검증방법들은 실험설계행열에서 외곽평균들의 차이를 검증하기 위한 방법들이며, 실험설계행열의 내부평균 혹은 쎌평균들의 차이를 검증하기 위한 것이 아니다. 반면에 상호작용효과는 쎌평균들의 양상에 의해 결정된다. 그러나 앞서 설명한 바와 같이 상호작용효과는 쎌평균들의 단순한 차이를 의미하는 것이 아니고, 쎌평균들의 차이의 차이를 의미한다. 평균들의 차이의 차이에 대한 통계적 유의미도를 판단하는 사후검증방법은 존재하지 않는다.

흔히 상호작용효과가 유의미한 경우, 쎌평균들의 차이를 통계적으로 사후검증하고, 그 검증결과를 주관적으로 해석하여 상호작용효과를 파악하고자 시도하는 경우가 많다. 예를 들어, 상호작용효과가 유의미한 경우 Scheffè 검증이나 Tukey 검증을 이용하여 $\overline{Y}_{21} - \overline{Y}_{22}$ 의 차이를 검증하고, $\overline{Y}_{11} - \overline{Y}_{12}$의 차이를 검증한 후, 그 두개의 검증결과의 차이를 주관적으로 파악하여 상호작용효과의 양태를 해석하는 것이다. 그런데 이러한 임의적인 관행은 3 가지의 중요한 사실을 간과하는 잘못된 관행이다.

첫째로, 앞서 언급한 바와 같이 Scheffè 검증과 Tukey 검증 등의 사후검증방법들은 실험설계행열에서 외곽평균들의 차이를 검증하기 위한 방법들이며, 실험설계행열의 내부평균 혹은 쎌평균들의 차이를 검증하기 위한 것이 아니다 (일원변량분석의 경우에는 외곽평균과 쎌평균이 동일하므로 마치 쎌평균 사이의 차이를 검증하는 듯이 보인다. 그러나 일원변량분석의 경우에도 사후검증은 원칙적으로 외곽평균들 사이의 차이에 대한 검증이다). 독립변인이 2 개 이상일 때 쎌평균들의 차이를 검증하기 위하여 Scheffè 검증 혹은 Tukey 검증을 사용하는 경우에 제1 종오류를 범할 확률이 어떻게 설정되는지에 대한 확고한 근거가 설정된 바 없다.

둘째로, 이원변량분석에서 $\overline{Y}_{21} - \overline{Y}_{22}$, $\overline{Y}_{11} - \overline{Y}_{12}$, $\overline{Y}_{11} - \overline{Y}_{21}$, $\overline{Y}_{12} - \overline{Y}_{22}$의 차이들을 파악하는 것을 소위 "단순주효과검증 (test of simple main effect)" 이라고 부른다. 그런데 이 단순주효과들은 상호작용효과의 일부분이 아니고, 주효과와 상호작용효과가 혼합된 것이다(이것에 대한 증명은 Kirk, 1982, p. 365-371을 참고하라). 따라서 이 쎌평균 차이들을 통계적으로 검증한다고 하여도 그것이 반영하는 것은 요인의 주효과일 수도 있고, 상호작용효과일 수도 있으며, 둘다일 수도 있다.

이원변량분석에서 쎌평균들의 차이를 Scheffè 검증 혹은 Tukey 검증하는 것은 단순주효과에 대한 검증을 Scheffè 검증 혹은 Tukey 검증이 설정하는 제1종 오류를 범할 확률에 의해 판단하는 것이다. 그런데 이 단순주효과들의 차이를 다시 주관적으로 파악하여 상호작용효과를 해석할 때는 판단오류의 확률에 대한 어떠한 논리적 근거도 존재하지 않는다 (Betz & Gabriel, 1978). 다시 말해서 Scheffè 검증 혹은 Tukey 검증에 의해 단순주효과들을 파악한 후, 그 단순주효과들의 차이를 주관적으로 파악하여 상호작용효과를 해석하는 것은 얼핏 매우 과학적이고 엄격한 통계분석인 듯이 보이지만 사실은 쎌평균들의 그래프를 시각적으로 해석하는 것과 다를 것이 없는 비통계적인 방법이다.

셋째로, 쎌평균 차이(단순주효과)가 모두 유의미하지 않더라도 상호작용효과는 유의미할 수 있고, 그 반대도 성립한다. 예를 들어 다음과 같은 실험자료가 존재한다고 가정하자.

	a_1	a_2
b_1	$\overline{Y}_{11}$=6.82 s_{11}=1.72 n_{11}=5	$\overline{Y}_{21}$=11.44 s_{21}=2.02 n_{21}=5
b_2	$\overline{Y}_{12}$=9.62 s_{12}=2.42 n_{12}=5	$\overline{Y}_{22}$=9.61 s_{22}=2.69 n_{22}=5

표 7. 변량분석표.

변산원	자승합	df	평균자승	F
A	26.556	1	26.556	5.282*
B	1.153	1	1.153	.229
AxB	26.772	1	26.772	5.325*
집단내	80.438	16	5.027	
전체	134.918	19	7.101	

위의 경우는 상호작용효과는 유의미하지만, $\overline{Y}_{21} - \overline{Y}_{22}$의 차이와 $\overline{Y}_{11} - \overline{Y}_{12}$의 차이는 모두 유의미하지 않다 (이 쎌평균들의 차이는 사후검증방법들에 비하여 검증력이 월등히 높은 단순 t 검증을 하여도 유의미하지 않다). 상호작용효과는 유의미하지 않지만, $\overline{Y}_{21} - \overline{Y}_{22}$의 차이와 $\overline{Y}_{11} - \overline{Y}_{12}$의 차이는 모두 유의미한 예는 훨씬 쉽게 만들어낼 수 있다. 이러한 경우들이 존재하는 이유는 상호작용효과가 단순한 쎌평균들의 차이를 의미하는 것이 아니고, 관계의 차이 혹은 차이의 차이를 의미하는 것이기 때문이다. 따라서 소위 단순주효과검증이라고 흔히 불리우기도 하는 쎌평균간의 차이검증은 상호작용효과를 이해하는데 전혀 도움이 되지 않는다(Kirk, 1982, p. 371).

상호작용효과의 해석을 위한 대비분석

모든 독립변인들이 2개씩의 수준들로 이루어진 경우에는 상호작용효과를 해석하기 위한 추가적인 분석이 필요하지 않다. 그런 경우에는 쎌평균들을 이용하여 실험설계행열을 작성하거나 그래프를 만들어서 상호작용효과를 해석하면 된다.

적어도 한개의 독립변인이 세 개 이상의 수준을 가질 때 상호작용효과가 유의미한 경우, 그 결과를 정확히 파악하기 위해서는 대비분석이 수행될 필요가 있다. 그런데 대비분석은 연구자의 사전 의도 혹은 가설에 의해 이루어진다. 연구자의 의도란 연구자가 비교하기를 원하는 평균들을 의미한다. 예를 들어 한 독립변인 A는 a_1, a_2, a_3의 수준을 가지고, 다른 독립변인 B는 b_1과 b_2의 수준을 가지는데, 연구자는 우선 a_1과 a_2의 합해진 집단과 a_3 집단의 차이에 관심이 있고, 그 다음으로는 a_1 과 a_2의 차이에 관심을 가진다고 가정하자. 이 두 가지 관심을 각각 대비에 반영하면, 상호작용효과에 대해서도 자연히 각각의 대비 중 어떤 것이 독립변인 B와 상호작용하는 것인지의 여부가 관심의 대상이 된다.

전체 집단수가 6개 이므로, 이러한 연구자의 의도를 반영하는 직교대비(orthogonal contrast)들은 최대 5개가 만들어질 수 있다. 요인 A는 3개의 수준을 가지므로, 2개의 직교대비에 의해 그 주효과가 분리된다. 요인 B는 2개의 수준을 가지므로, 1개의 직교대비에 의해 그 주효과가 분리된다. 상호작용효과는 요인 A의 대비들과 요인 B의 대비를 서로 곱해준 2 개의 직교대비들로 분리된다. 따라서 여기서 예로 든 연구자의 의도를 반영하는 5개의 직교대비들의 가중치를 표 8과 같이 설정할 수 있다.

표 8. 주효과와 상호작용효과의 대비가중치.

			A-효과		B-효과	AB-효과	
A	B	쎌평균	대비-1	대비-2	대비-3	대비-4	대비-5
a_1	b_1	17.8	1	1	1	1	1
a_1	b_2	4.0	1	1	-1	-1	-1
a_2	b_1	1.0	1	-1	1	1	-1
a_2	b_2	65.2	1	-1	-1	-1	1
a_3	b_1	35.8	-2	0	1	-2	0
a_3	b_2	43.0	-2	0	-1	2	0

Ψ_1 = (1)17.8+(1)4.0+(1)1.0+(1)65.2+(-2)35.8+(-2)43.0= -69.6

Ψ_2 = (1)17.8+(1)4.0+(-1)1.0+(-1)65.2+(0)35.8+(0)43.0= -44.4

Ψ_3 = (1)17.8+(-1)4.0+(1)1.0+(-1)65.2+(1)35.8+(-1)43.0= -57.6

Ψ_4 = (1)17.8+(-1)4.0+(1)1.0+(-1)65.2+(-2)35.8+(2)43.0= -36.0

Ψ_5 = (1)17.8+(-1)4.0+(-1)1.0+(1)65.2+(0)35.8+(0)43.0= 78.0

위에서 계산된 대비들의 자승합을 구하면 다음과 같다.

SSB_{Ψ_1} = $5(-69.6)^2/[(1)^2+(1)^2+(1)^2+(-2)^2+(-2)^2]$ = 2018.4

SSB_{Ψ_2} = $5(-44.4)^2/[(1)^2+(1)^2+(-1)^2+(-1)^2+(0)^2+(0)^2]$ = 2464.2

SSB_{Ψ_3} = $5(-57.6)^2/[(1)^2+(-1)^2+(1)^2+(-1)^2+(1)^2+(-1)^2]$ = 2764.8

SSB_{Ψ_4} = $5(-36.0)^2/[(1)^2+(-1)^2+(1)^2+(-1)^2+(-2)^2+(2)^2]$ = 540

SSB_{Ψ_5} = $5(78.0)^2/[(1)^2+(-1)^2+(-1)^2+(1)^2+(0)^2+(0)^2]$ = 7605

이와 같은 대비자승합들을 이용하여 표 9와 같은 변량분석표를 작성할 수 있다.

표 9를 보면 다음과 같은 사실들이 확인된다. ① 특정한 효과를 분리한 직교대비들의 자승합을 합하면 그 효과의 자승합이다; 예를 들어 대비-1과 대비-2는 요인 A의 효과를 분리한 것인데, 이 직교대비들의 자승합의 합은 요인 A의 자승합이다. ② 특정한 효과를 분리한 직교대비들의 자유도를 합하면 그 효과의 자유도이다; 예를 들어 대비-4와 대비-5는 상호작용의 효과를 분리한 것인데, 이 대비들의 자유도의 합은 상호작용효과의 자승합이다. ③ 모든 대비들의 F 값을 산출하기 위한 오차항은 그 대비들의 모체가 되는 효과의 오차항과 같다; 예를 들어 대비-1과 대비-2는 요인 A효과의 일부이므로 이 대비들의 F 값을 산출하기 위한 오차항은 A-효과의 오차항과 같은 MSW이다.

	a_1+a_2	a_3	
b_1	9.4	35.8	18.2
b_2	34.6	43.0	37.4
	22.0	39.4	

	a_1	a_2	
b_1	17.8	1.0	18.2
b_2	4.0	65.2	37.4
	10.9	33.1	

표 9. 대비분석표.

변산원		자승합	자유도	평균자승	F
(A)		4482.6	2	2241.3	20.49**
	대비-1	2018.4	1	2018.4	18.45**
	대비-2	2464.2	1	2464.2	22.52**
(B)		2764.8	1	2764.8	25.27**
	대비-3	2764.8	1	2764.8	25.27**
(A)x(B)		8145.0	2	4072.5	37.23**
	대비-4	540.0	1	540.0	4.94*
	대비-5	7605.0	1	7605.0	69.52**
집단내		2626.4	24	109.4	
전 체		18018.8	29	621.3	

대비분석 결과, 모든 대비들이 또한 유의미하였다. 이 대비들의 의미를 해석하기 위하여 다음과 같은 두개의 평균표를 작성할 수 있다. 숫자는 모두 평균들이다.

대비-1의 유의미한 효과는 a_1과 a_2를 합한 집단의 평균 (22.0)과 a_3 집단의 평균(39.4) 사이의 차이가 유의미하다는 것을 의미한다. 대비-2의 유의미한 효과는 a_1 집단과 a_2 집단의 평균차이(10.9와 33.1)가 유의미하다는 것을 의미한다. 대비-3의 유의미한 효과는 요인 B에 따른 차이(18.2 와 37.4)이다. 대비-4의 유의미한 효과는 [(9.4+43.0)-(35.8+34.6)] 이 유의미하다는 것을 의미한다. 대비-5의 유의미한 효과는 a_1의 실험처치와 a_2의 실험처치가 유발하는 반응의 차이가 b_1과 b_2 집단에서 다른데, b_1 집단에서는 a_1 (17.8)이 a_2 (1.0)보다 큰 반응을 유발하는 반면, b_2 집단에서는 그 정반대이다.

대비의 자승합을 구하기 위해서는 각 쎌의 n이 동일하다는 가정이 필요하다. 그런데 만약 각 쎌의 피험자수가 동일하지 않으면, 그 피험자수의 조화평균 (harmonic mean) 을 계산하여 n으로 간주한다.

실험의 사례수가 작으면 비모수검증을 해야한다는 오해

Gasset(Student, 1908)은 작은 크기의 표본들로 이루어진 표집분포(sampling distribution)는 중심한계법칙 (central limit theorem)이 상정하는 정규분포에서 벗어난다는 사실을 발견하였고, 작은 크기의 표본들로 이루는 표집분포의 정확한 속성을 규명하여 t-분포라고 명명하였다. R. A. Fisher(1925)는 작은 크기의 표본들로 이루는 표집분포는 실제로 Gasset이 규명한 t-분포를 한다는 사실을 수학적으로 증명하였고, 이 새로운 지식에 기초하여 변량분석과 F-검증을 발전시켰다 (Fisher, 1931; Yates & Mather, 1963)[2]. 즉, t-검증과 변량분석은 실험의 사례수가 작은 경우에도 모집단의 상태에 관하여 정확한 확률추정을 하기 위해 개발된 통계검증 방법이다.

변량분석은 경우에 따라서 실험집단간 사례수가 동일하다는 가정을 요구하기는 하지만, 사례수의 절대크기에 대한 어떠한 가정에도 기초하지 않는다. 따라서 실험조건에 대한 사례들의 무선할당이 이루어진 조건하에서는 사례수가 큰 경우나 작은 경우에 공히 변량분석이 사용될 수 있다. 심지어 반복측정 설계자료의 분석에서 "피험자 요인"이 하나의 구획요인(blocking factor)으로 설정되는 경우에는, 각 구획조건당 단 1 개의 사례가 존재하는 경우에도 변량분석은 훌륭한 통계적 도구로서 활용될 수 있다.

다만, 실험자료의 분석에서 사례수가 작은 경우에는 두 가지 잠재적인 문제가 예상될 수 있다. 첫째는 모집단에 대한 일반화의 타당성 문제이고, 둘째는 Fisher 가 "지역통제 (local control)" 라고 명명한 가외변인의 통제가 효과적으로 달성되지 못할 가능성이 증가한다는 문제이다. 그러나 이 잠재적 문제들은 변량분석의 문제가 아니라 모든 자료분석(비모수검증을 포함)이 공히 가지는 문제이다. 작은 크기의 표본을 사용하면서 이러한 잠재적 문제를 회피할 수 있는 자료분석방법은 존재하지 않는다.

실험자료에 대한 비모수검증은 t-검증과 변량분석이 가지는 모집단의 정규분포성 가정이나 변량의 동질성가정이 타당하지 않은 경우에 채택할 수 있는 대안이다. 그러나 이 가정들은 모두 모집단의 상태에 대한 가정들이라는 사실을 주목해야 한다. 모집단으로부터 표집을 조금만 하면 모집단의 상태가 바뀌어 버리는 불가능한 상황을 제외하면, 표본의 크기가 작다고 해서 변량분석의 가정이 충족되지 않고, 표본의 크기가 크면 그것이 충족되는 경우는 존재하지 않는다. 모집단의 정규분포가정이나 변량의 동질성 가정이 타당하지 않은 경우에는 표본의 사례수가 십수만이 되더라도 비모수검증을 하는 것이 바람직하고, 그러한 가정이 타당한 경우에는 표본의 사례수가 작더라도 변량분석을 사용할 수 있다.

2) 두집단 비교를 위하여 변량분석을 하는 경우 Gasset 의 t 와 Fisher 의 F 는 $t^2=F$ 의 관계를 가진다는 사실을 상기하라.

결 론

량분석을 이용한 연구에서 F 값만을 보고하는 것은 평균차이가 "통계적으로 유의미하다"는 것만을 판단하는 것이다. 그런데 통계적으로 유의미하다는 것은 모집단에서 평균차이가 零이라는 가정을 기각하는 것에 불과하다. 다시 말해서 F 값은 연구에서 구해진 자료에 관한 극히 애매하고 제한적인 정보에 지나지 않는다. 더군다나 F 값에 대한 해석은 매우 추상적이고, 때로는 비현실적인 가정들을 전제로 한다. 이렇게 제한적이고 추상적인 정보만 가지고는 그 연구에 관한 의미있는 학문적 의사소통이 제한될 수 밖에 없다. 또한 연구자가 자신의 실험적 조작에 대하여 적용하는 묵시적 가정을 다른 연구자가 파악할 수 있으려면 변량분석표 전체를 보고하는 것이 바람직하다.

상호작용효과는 셀평균들의 단순한 차이를 의미하는 것이 아니고, 셀평균들의 차이의 차이를 의미한다. 평균들의 차이의 차이에 대한 통계적 유의미도를 판단하는 사후검증방법은 상호작용효과 이외에 따로 존재하지 않는다. 셀평균 차이는 상호작용효과의 일부분이 아니고, 주효과와 상호작용효과가 혼합된 것이다. 따라서 셀평균 차이를 통계적으로 검증한다고 하여도 그것이 반영하는 것은 주효과일 수도 있고, 상호작용효과일 수도 있으며, 둘 모두일 수도 있다. 상호작용효과의 해석은 계획비교(planned comparison)를 통하여 이루어지는 것이 바람직하다.

모집단의 정규분포가정이나 변량의 동질성 가정이 타당하지 않은 경우에는 표본의 사례수가 십수만이 되더라도 비모수검증을 하는 것이 바람직하고, 그러한 가정이 타당한 경우에는 표본의 사례수가 작더라도 변량분석을 사용할 수 있다.

참고문헌

Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8*, 1, 12-15

Betz, M. A. & Gabriel, K. R. (1978). Type IV errors and analysis of simple effects. *Journal of Educational Statistics, 3*, 121-143.

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences.* New York: Academic Press.

Estes, W. K. (1997). Significance testing in psychological research: Some persisting issues. *Psychological Science, 8*, 18-20

Fisher, R. A. (1925). Theory of statistical estimation. *Cambridge Philosophical Society, 22*, 700-725.

Fisher, R. A. (1931). Principles of plot experimentation in relation to the statistical interpretation of the results. In *Rothamsted Conferences, 13*, 11-13.

Harris, R. J. (1997). Significance tests have their place. *Psychological Science, 8*, 8-11

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8*, 3-7

Kirk, R. E. (1982). *Experimental Design: Procedures for the Behavioral Sciences. Belmont:* Brooks/ Cole Publishing Company.

Neter, J. & Wasserman, W. (1974). *Applied Linear Statistical Models.* Homewood: Richard D. Irwin, Inc.

Rao, C. R. (1992). R. A. Fisher: The founder of modern statistics. *Statistical Science, 7*, 34-48.

Scarr, S. (1997). Rules of evidence: A larger context for the statistical debate. *Psychological Science, 8*, 16-17

Shrout, P. E. (1997). Should significance tests be banned?: Introduction to a special section exploring the pros and cons. *Psychological Science, 8*, 1-2

Student (1908). The probable error of a mean. *Biometrika, 6*, 1-25.

Yates, F. & Mather, K. (1963). Ronald Aylmer Fisher, 1890-1962. *Bibliographic Memoirs of Fellows of the Royal Society of London, 9*, 91-129.

원고 접수일 2000. 8. 14.
수정원고접수일 2000. 11. 15.
게재결정일 2000. 12. 21.

Undesirable Uses of Analysis of Variance: Reporting F-values Only, Comparing Cell Means in a Post-hoc Manner to Interpret Interaction Effects, and Performing Nonparametric Tests for the Reason of Small Sample Size.

Kwang B. Park **Jinsup Eom**

Department of Psychology, Chungbuk National University

This paper is to remind cases in which analysis of variance is misused in psychological literature and explain why such misuses of ANOVA should be avoided. The most common misuse of ANOVA is that authors report only F-values without any other results from the ANOVA. The secondly common misuse is that some authors perform unplanned comparisons of cell means to interpret significant interaction effects. The thirdly common misuse is that some authors perform nonparametric tests instead of standard parametric ANOVA for the reason of small sample sizes.

Keywords : ANOVA, interaction, post-hoc comparison, nonparametric test