Journal of the Korean Data & Information Science Society
2010, **21**(1), 1–9

한국데이터정보과학회지

반복측정의 분할구 자료에 대한 혼합모형†

최재성[1]

[1]계명대학교 통계학과

접수 2009년 9월 1일, 수정 2009년 11월 21일, 게재확정 2009년 12월 30일

요 약

본 논문은 분할구 실험에서 반복측정 요인이 처치의 한 요인으로 고려될 때, 실험자료의 분석을 위한 혼합모형과 모형내 미지모수의 추론을 위한 방법을 논의한다. 반복측정 요인으로 공간요인을 고려하고 공간요인의 수준은 분할구에 할당되나 연구자가 임의로 배정할 수 없는 실험환경이 가정된다. 이러한 실험의 특성을 갖는 자료벡터의 확률분포로 복합대칭의 공분산 구조를 갖는 다변량 정규분포를 논의하고 있다. 또한, 가정된 실험환경에 부합하는 적합한 자료의 예를 통하여 제시된 모형의 타당성과 관련모수들의 추론방법을 다루고 있다.

주요용어: 반복측정 요인, 복합대칭의 공분산 구조, 분할구, 혼합효과.

1. 서론

분할구 실험계획의 한 특성은 처리 또는 처치로 주어지는 요인들의 수준이 서로 다른 크기의 실험단위를 필요로 할 때, 행해지는 실험의 한 설계방법이다. 상이한 크기의 실험단위를 요하는 실험계획들로는 반복측정 계획, 실험단위들의 구조가 지분관계에 있는 지분계획 등을 들 수 있다. 분할구 실험계획과 관련된 문헌들은 Milliken과 Johnson (1984), Cochran과 Cox (1957), 그리고 Mead (1988)등에서 다양한 유형의 자료와 관련된 모형과 분석방법들을 살펴볼 수 있다. 분할구 실험계획의 또 다른 특성은 요인들의 수준결합으로 주어지는 처치가 동일 실험구내 분할구에서 모두 배정되지 않는 불완비 불록의 설계구조를 갖는 점이다. 또한, 요인들의 수준은 해당하는 실험단위에서 임의로 배정되는 것을 가정하고 있다.

그러나 분할구 실험계획하에서 처리로 주어지는 요인들의 수준이 해당하는 실험단위에 임의로 배정될 수 없다면, 분할구 실험계획과 관련된 자료분석 방법은 타당하지 않게 된다. 따라서, 본 논문에서는 분할구 실험과 관련된 가정이 성립하지 않는 경우에 요인들의 수준효과를 추론하기 위한 문제를 논의하고자 한다. 서로 다른 크기의 실험단위를 필요로 하는 분할구 실험에서 연구자가 임의로 처치를 해당하는 실험단위에 배정할 수 없는 경우를 생각해 볼 때, 반복측정 요인이 처치의 구성요인으로 존재하는 경우이다.

처치로 반복측정 요인이 고려될 때 유의할 점은 이 요인의 수준이 해당하는 실험단위에 임의로 배정될 수 없다는 점에서 실험단위에서 관측되는 관측값들 간에 독립성을 가정할 수 없다는 점이다. 즉, 반복측정 요인의 수준들에서 관측되는 값들 간에는 어떤 상관성을 갖는 상관성의 구조하에 자료가 분석되어야 함을 의미하고 있다. 관측값들 간의 상관성을 나타내는 상관성의 구조는 공분산 구조로 표현될 수 있다.

† 본 연구는 2009년도 계명대학교 비사연구기금으로 이루어졌음.

[1] (704-701) 대구광역시 달서구 신당동 1000번지, 계명대학교 자연과학대학 통계학과, 교수.
E-mail: jschoi@kmu.ac.kr

본 논문은 분할구 유형의 실험에서 반복측정 요인으로 공간요인을 고려하고 있다. 처치의 한 요인인 공간요인은 실험구나 분할구에서 나타날 수 있으나 분할구에 배정됨을 가정한다. 또한, 실험구에서 행해지는 처치는 두 고정요인의 수준결합으로 실험이 행해지는 것을 가정한다. 그리고 실험구의 블록화가 요구되는 실험을 계획하여 블록요인인 확률요인과 처치의 구성요인들로는 고정요인만을 고려한다. 실험의 형태는 분할구 실험과 유사하나 분할구 실험계획에서 요구되는 가정이 만족되지 않는 실험의 유형은 다양하다. 여기서는 그러한 실험의 한 유형을 생각하고 분석에 유용한 혼합모형을 개발하고 적용가능한 예를 통하여 구체적으로 분석의 타당성을 살펴보고자 한다.

2. 혼합모형에 대한 가정

분할구 유형의 실험자료를 분석하기 위한 선형모형의 논의를 위해 필요한 가정들을 살펴보기로 한다. 실험을 위한 실험단위들은 동질적인 실험단위들로 구분되는 어떤 분류기준에 따라 동수의 실험단위를 갖는 몇 개의 그룹으로 나누어진다고 가정한다. 따라서, 블록요인은 확률요인이다. 블록내 실험구에 해당하는 실험단위들에 임의로 배정되는 처리로 세 요인들을 가정한다. 세 요인중 요인 A는 a개의 수준 $(i = 1, 2, \cdots, a)$을 갖는 고정요인이고, 요인 B는 b개의 수준 $(j = 1, 2, \cdots, b)$을 갖는 고정요인이라 둔다. 요인 A의 a개 수준과 요인 B의 b개 수준결합이 처리로 실험의 특성상 큰 단위의 실험단위에 배정되고, 반복측정요인 T의 g개 수준 $(k = 1, 2, \cdots, g)$은 작은 단위의 실험단위에 배정된다고 가정한다.

동일 개체의 반응이 연구자가 임의로 배정할 수 없는 반복측정 요인의 수준들에서 반복측정될 때, 관측반응 간에 일정한 구조적 상관성을 배제할 수 없게 된다. 개체 또는 실험단위의 반응변수 Y가 반복측정 요인, T의 g개 시점 $(k = 1, 2, \cdots, g)$에서 주기적으로 반복측정 된다고 가정한다. 따라서, 반복측정으로 인한 반복측정값들 간의 공분산구조를 가정한다. 요인들의 수준결합으로 주어지는 처치효과들의 올바른 추론을 위한 공분산 구조의 유형은 실험의 특성에 따라 다양하게 주어진다.

분할구 실험계획에서는 처치로 이용되는 요인들의 수준이 해당하는 실험단위들에 임의로 배정되므로 해당하는 실험단위들에서 관측되는 오차성분들은 독립임을 가정하게 된다. 그러나, 시간 또는 공간요인이 반복측정 요인으로 처치의 한 구성요인으로 간주될 때, 개체의 반응간에는 어떤 상관성을 갖는 관측값들로 나타날 수 있다. 왜냐하면, 시간요인이 처리의 한 고정요인으로 이용될 때, 시간요인의 수준들은 연구자가 임의로 해당하는 실험단위에 배정할 수 없게 되고 이로 인해서 해당하는 실험단위에서 관측되는 값들은 일정한 상관관계를 띄면서 주어지는 경향이 있다. 또한, 반복측정요인으로 공간요인을 고려하고 공간요인의 정해진 지점에서 측정이 행해질 때, 이들 수준들은 연구자가 임의로 실험단위에 배정할 수 있는 수준이 아니므로 분할구 실험의 가정을 만족시키지 못하게 된다. 공간요인으로 인해 분할구 실험의 자료분석에서 요구되는 가정이 만족되지 않는 경우에 관측값 간의 상관성이 존재하게 된다. 이러한 상관성을 고려하지 않은채 요인효과에 대한 추론을 위한 자료의 분석은 정확하지 않게 된다. 따라서, 본 연구에서는 반복측정 요인이 고려된 분할구 실험자료를 분석하기 위해 개체내 반응값 간의 상관성을 나타내는 어떤 공분산 구조의 가정하에서 혼합모형을 다루고 있다.

3. 분할구 실험자료에 대한 혼합모형

두 개의 상이한 크기의 실험단위를 갖게 되는 분할구 실험으로 부터 예상되는 분산성분들은 적어도 두 개 존재하게 된다. 대단위의 실험단위인 실험구의 변동에 따른 분산성분과 소단위의 실험단위인 분할구의 변동에서 부터 관측되는 오차분산이다. 또한 대단위의 실험단위들인 실험구가 이질적인 현상을 나타내고 있을 때, 적절한 분류기준에 의해 동질적인 실험구들의 집단인 블록으로 구획될 때, 블록간의 변동 또한 분산성분을 제공하게 된다. 실험의 계획에서 블록요인이 고려될 때, 이를 고려한 모형은 혼합모형

이 됨을 인지할 수 있게 된다.

혼합모형에 대한 인식은 실험단위들의 이질성으로 인한 동질적 실험단위들의 설계구조를 위한 블록요인의 고려때문에 발생할 수도 있지만 처리구조를 나타내는 요인들 중 일부가 확률요인일 때도 고려되어야 함을 알 수 있다. 처리구조내 확률요인에 따른 혼합모형에 관한 논의는 Milliken과 Johnson (1984)에서 볼 수 있다.

2절에서 논의된 실험환경에서 분할구 실험자료를 분석하기 위한 혼합모형을 논의하기로 한다. 처리구조에서 주어지는 세 요인 A, B와 T는 고정요인들이다. 실험의 설계구조에서 동질적 실험단위들로 구성하게 되는 블록요인을 D라 두고 D는 $l = 1, 2, \cdots, d$ 의 d개 수준을 갖는다 하자. 블록요인 D의 수준 l에서의 블록효과를 δ_l로 나타낸다. 실험단위에서 측정되는 변수를 y라 둔다.

처리요인 A의 수준 i에서의 수준효과를 α_i라 두면, a개의 α_i들은 고정효과들 이다. 처리요인 B의 수준 j에서의 수준효과를 β_j라 두면, b개의 β_j들은 고정효과들을 나타낸다. 처리요인 T의 수준 k에서의 수준효과를 τ_k라 두면, g개의 τ_k 수준효과들도 고정효과들을 나타낸다. 블록 l에서 요인 A의 수준 i, 요인 B의 수준 j, 그리고 요인 T의 수준 k가 처리로 주어지는 개체 또는 실험단위의 반응변수 y에 대한 반응을 y_{ijkl}라 두자. 분할구 실험계획을 고려하고 있으므로 요인 A의 a개 수준과 요인 B의 b개 수준의 결합수준이 처리로 큰 실험단위인 실험구에서 측정된다고 가정하고 반복측정 요인 T의 g개 수준이 작은 실험단위인 분할구에서 측정된다고 가정한다. 두 개의 서로 다른 크기의 실험단위의 분산성분들과 확률요인인 블록요인의 블록변동에 따른 분산성분을 포함하는 두 유형의 효과들을 고려한 혼합모형은 다음과 같다.

$$y_{ijkl} = \mu + \delta_l + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_{ijl} + \tau_k + (\alpha\tau)_{ik} + (\beta\tau)_{jk} + (\alpha\beta\tau)_{ijk} + \epsilon_{ijkl}. \tag{3.1}$$

식 (3.1)에서 μ는 전체의 평균을 나타내고 α_i는 고정요인 A의 i번째 수준의 고정효과를 나타낸다. β_j는 고정요인 B의 수준 j의 고정효과이다. $(\alpha\beta)_{ij}$, $(\alpha\tau)_{ik}$, $(\beta\tau)_{jk}$는 요인 A와 B, A와 T, 그리고 B와 T의 두 요인 교호작용을 나타낸다. $(\alpha\beta\tau)_{ijk}$는 세 요인 교호작용을 나타낸다. 그리고 δ_l은 블록 l에서의 분산성분과 관련된 블록효과, γ_{ijl}는 실험구의 변동을 나타내는 분산성분과 관련된 확률효과이고 ϵ_{ijkl}은 분할구에서의 변동성분을 나타내는 오차이다.

두개의 서로 다른 실험단위와 블록을 갖는 분할구 실험에서의 일반적인 가정은 분산성분을 나타내는 확률효과들인 δ_l, γ_{ijl}과 오차를 나타내는 ϵ_{ijkl}은 서로 독립이고 각기 $N(0, \sigma_\delta^2)$, $N(0, \sigma_\gamma^2)$ 그리고 $N(0, \sigma_\epsilon^2)$인 분포를 따른다는 가정하에 해당하는 분산성분을 추론하게 된다. 그러나, 분할구에서 처리로 간주되는 T의 g개 수준들은 확률화에 의해 배정될 수 없게 된다. 왜냐하면, 반복측정 요인인 T의 시간 또는 공간상의 제한성으로 인해 g개 수준들은 연구자가 임의로 배정할 수 없다. 처리들이 비확률화에 의해 소단위의 실험단위에 행해짐으로 인해서 분할구에서 관측되는 g개 오차들은 오차간에 일정한 상관성을 갖는 오차들로 관측됨을 예상할 수 있다. 따라서, 이러한 현상들을 고려할 때, 분할구 실험의 설계구조와 관련된 모형의 일반적인 가정을 할 수 없게 된다. 즉, 분할구로 간주되는 작은 실험단위에서 관측값간의 오차들은 어떤 상관성을 갖게 되므로 이를 고려한 공분산 구조하에서 모형이 제안되어야 함을 의미하고 있다.

따라서, 오차벡터를 $\boldsymbol{\epsilon}$라 두고, $\boldsymbol{\epsilon}=[\epsilon_1, \epsilon_2, \cdots, \epsilon_n]'$으로 정의한다. 여기서, n은 실험에서 관측된 반응값들의 수를 나타낸다. 따라서 n은 $dabg$개 임을 알 수 있다. 모형 (3.1)에서 분산성분과 오차벡터의 공분산 구조를 고려한 가정으로 인해 벡타 $\boldsymbol{\delta}= [\delta_1, \delta_2, \cdots, \delta_d]'$는 블록효과 벡터이다. 블록효과 벡터에 대한 분포로 다변량 정규분포 $MVN(\mathbf{0}, \sigma_\delta^2 \boldsymbol{I})$를 가정한다. $\boldsymbol{\gamma}_l=[\gamma_{11l}, \gamma_{12l}, \cdots, \gamma_{abl}]'$로 나타내면, $\boldsymbol{\gamma}=[\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \cdots, \boldsymbol{\gamma}_d]'$인 벡터이다. 따라서, $\boldsymbol{\gamma}$는 $MVN(\mathbf{0}, \sigma_\gamma^2 \boldsymbol{I})$인 다변량 정규분포를 가정한다. 그리고 T의 g개 수준에서 측정되는 관측값들의 오차벡터에 대한 분포로 $MVN(\mathbf{0}, \boldsymbol{\Sigma}_g)$를 가정한다.

두개의 상이한 실험단위를 필요로 하는 분할구 실험에서 개체단위의 실험단위에 행해지는 두 요인중 요인 A의 a개 수준이 양적수준이라면 이에 해당하는 모형은 다음과 같이 표현된다.

$$y_{ijkl} = \mu + \delta_l + \alpha u_i + \beta_j + \beta_j u_i + \gamma_{(ijl)} + \tau_k + \tau_k u_i + (\beta\tau)_{jk} + (\beta\tau)_{jk} u_i + \epsilon_{(ijkl)}. \quad (3.2)$$

모형 (3.2)에서의 변화는 고정요인 A의 a개 수준들이 $u_1, u_2, \cdots, u_a$로 주어지는 양들로 주어질 때, 요인 A의 양적변화에 따른 회귀계수로 표현됨을 알 수 있다. 이는 분할구 자료를 분석하기 위한 기본 모형으로 부터 모형에 포함되는 독립변수의 유형과 특성에 따라 다양하게 변형될 수 있음을 나타내고 있다. 달리말하면, 모형 3.1에서의 요인 A의 a개 수준들이 수준들의 한 모집단에서 임의로 추출된 수준들로 간주되면 요인 A는 확률요인으로 취급된다. 이 경우에 수준들의 효과는 확률효과를 나타내므로 이와 관련된 분산성분을 고려하게 된다. 개체의 실험단위에 적용되는 요인 A가 확률요인으로 취급될 때, 확률요인과 고정요인과의 교호작용은 그 효과들이 확률효과로 주어지므로 이에 따른 분산성분들의 분포를 고려한 혼합모형이 구축되어야 한다. 확률요인의 수준이 독립성을 가정할 수 있는 확률표본으로 간주되면 효과들의 분포로 일반적으로 평균이 0이고 상수분산을 갖는 정규분포를 가정하게 된다. 확률수준의 표본추출에 있어서 어떤 종속성이 예상되는 경우 확률효과에 대한 분포로 이를 고려한 공분산 구조의 확률분포를 가정하게 된다. 오차성분에 대한 공분산 구조로 복합대칭구조, 1차 자기회귀 구조등을 생각할 수 있다.

모형 (3.1)과 관련된 공분산 구조로 복합대칭구조를 가정할 때, 그 구조형태를 행렬로 표현해 보기로 한다. 모형 (3.1)의 행렬표현은 다음과 같이 주어진다.

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z_1\delta} + \boldsymbol{Z_2\gamma} + \boldsymbol{\epsilon}. \quad (3.3)$$

식 (3.3)에서 $\boldsymbol{X}$는 $n \times p$인 모형행렬을 나타낸다. 여기서, p는 $(1+a+b+ab+g+ag+bg+abg)$개의 고정효과들의 수이다. $\boldsymbol{\beta}$는 $p \times 1$개의 모수들을 나타내는 모수벡터이다. $\boldsymbol{Z_1}$은 블록요인 D의 d개 수준들의 블록효과와 관련된 $n \times d$ 계수행렬이다. $\boldsymbol{\delta}$는 블록요인 D의 d개 수준효과들을 성분으로 갖는 확률효과 벡터이다. $\boldsymbol{Z_2}$는 큰 실험단위에서의 오차성분을 나타내는 $n \times abd$ 계수행렬이다. $\boldsymbol{\gamma}$는 abd개의 성분을 갖는 확률오차들의 벡터이다. 모형 (3.1)에서의 ϵ_{ijkl}들로 구성되는 $\boldsymbol{\epsilon}$은 n개의 오차성분을 나타내는 오차벡터이다. 블록효과와 개체단위가 실험단위로 이용되는 실험구에서 오차들에 의한 효과들은 확률효과로 간주되므로 관련된 분산성분의 추정을 위한 일반적인 가정을 따른다. 즉, 오차간의 독립성과 정규분포의 가정하에 분산성분들에 대한 추론이 행해지게 된다. 그러나, 개체의 소단위인 분할구에서의 오차들은 처리가 확률화에 의해 배정되지 않으므로 오차간에 상관성을 가정한 공분산 구조를 가정한다. 식 (3.3)에서 관측벡타 $\boldsymbol{y}$의 공분산 행렬을 살펴보면,

$$Var(\boldsymbol{y}) = \boldsymbol{Z_1}(\sigma_\delta^2 \boldsymbol{I}_{d\times d})\boldsymbol{Z_1'} + \boldsymbol{Z_2}(\sigma_\gamma^2 \boldsymbol{I}_{abd\times abd})\boldsymbol{Z_2'} + \boldsymbol{R} \quad (3.4)$$

로 표현된다. 모형 (3.1)에서의 가정들을 행렬식으로 표현한 식 (3.3)에서의 가정들로 표현하면, 블록효과 벡터인 $\boldsymbol{\delta}$는 $MVN(\boldsymbol{0}, \sigma_\delta^2 \boldsymbol{I}_{d\times d})$, 실험구에서의 오차벡터를 나타내는 $\boldsymbol{\gamma}$는 $MVN(\boldsymbol{0}, \sigma_\gamma^2 \boldsymbol{I}_{abd\times abd})$ 이고 $\boldsymbol{\epsilon}$은 $MVN(\boldsymbol{0}, \boldsymbol{R})$로 나타낼 수 있게 된다. 단, $\boldsymbol{R}$은 오차벡터 $\boldsymbol{\epsilon}$의 공분산행렬을 의미한다. 오차간의 독립성이 간주될 때, 공분산 행렬 $\boldsymbol{R}$은 $\sigma_\epsilon^2 \boldsymbol{I_n}$이 되고 오차벡터 $\boldsymbol{\epsilon}$의 분포로 $MVN(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I_n})$인 다변량 정규분포를 가정한다. 이러한 분포의 가정은 반복측정 요인 T의 g개 수준들이 분할구의 실험단위들에 임의로 배정될 수 있을 때 성립하게 된다. 그러나 동일 개체 또는 동일 실험단위에서 반복측정된 관측값들은 일반적으로 상관성을 갖게 되고 이러한 상관성을 고려한 공분산 구조하에 자료를 분석하게 된다. 동일 실험단위내 관측값들 간에 상관성을 갖게 하는 요인은 확률화의 원리에 의해 배정될 수 없는 반복측정 요인이 존재하기 때문이다.

공분산 구조를 고려하게 하는 반복측정 요인은 분할구 실험의 실험구에서, 또는 분할구에서 또는 실험구와 분할구 모두에서 발생할 수 있는 요인임을 생각할 때 분할구 실험자료의 분석에서 간과할 수 없는 한 중요한 요인으로 인식될 필요가 있다.

다시말하면, 분할구 실험에 의해 자료를 분석할 때, 단순히 상이한 크기의 실험단위로 부터 발생하는 분산성분들을 고려한 분석이 행해지게 되나, 처치구조에 있어 반복측정 요인의 존재는 공분산 구조하의 좀더 복잡한 분석을 필요로 하게 된다. 이는 단순히 혼합모형에 내포된 확률효과들의 분산성분이나 상이한 크기의 실험단위들의 분산성분들에 대한 추론보다는 오차간에 내재한 상관성을 고려한 공분산 구조하에 처치효과들을 비교할 때, 추론에 정확성을 더할 수 있음을 예상할 수 있다. 그러므로, 반복측정 요인을 감안한 분할구 실험 계획으로 행해질 때, 야기될 수 있는 공분산 구조를 감안한 모형제시가 필요하게 된다.

오차벡터 $\boldsymbol{\epsilon}$의 공분산 구조로 복합대칭 (compound symmetry)의 구조를 가정한다. 상관성을 갖는 반복측정 요인 T의 g개 오차들에 대한 복합대칭 구조의 공분산 행렬을 $\boldsymbol{\Sigma}_g$로 둘 때, $\boldsymbol{R}$은 $\boldsymbol{\Sigma}_g$를 대각원소로 갖는 블럭 대각행렬로 주어진다. 즉, 개체간의 변이는 독립이고 개체내의 오차는 가정된 상관성을 갖게 됨을 나타내고 있다. 행렬식 (3.3)에서 추정되어야 할 모수들은 고정효과 벡터 $\boldsymbol{\beta}$, 확률효과 벡터들의 분산성분들 σ_δ^2, σ_γ^2 그리고 공분산 구조의 σ_ϵ^2이다. 이들 모수들을 추론하기 위한 방법으로 Corbeil 과 Searle (1976)에 의한 REML (Restricted Maximum Likelihood)방법을 이용하기로 한다. REML 방법은 관측값들의 우도함수를 고정효과를 나타내는 부분과 고정효과를 제외한 확률효과들 만을 포함하는 두 부분으로 나누고 있다. 고정효과가 제외된 확률효과만의 주변우도는 모형의 고정효과들에 종속되지 않는 제한된 우도함수 (restricted likelihood function)를 나타낸다. 이 우도함수를 최대로 하는 모수추정량들이 분산성분들의 REML 추정량들이다.

4. 방수천 자료의 예

본 논문에서 제시된 혼합모형의 적용 예로써 다양한 색으로 염색된 원단의 방수가공과 관련된 자료를 생각해 보기로 한다. 일상생활에서 필요한 의류제품, 가죽제품등은 다양하게 염색된 가공원단을 이용해 방수처리한 제품을 생산하게 된다. 필요제품을 생산하기 위해 염색된 원단을 이용하여 방수처리된 가공원단을 제조하기 위해 두 가지 방수방법 (A)과 세 종류의 방수액 (B)을 이용하여 방수처리를 할 수 있다고 하자. 제조된 방수원단의 효능을 알아보기 위해 일정회수의 세탁후 방수처리된 천의 겉과 속에서 방수율 (y)이 측정된다고 하자. 처리에 따른 방수율을 측정하기 위해 방수방법의 두 수준, a_1, a_2와 방수액의 세 수준 b_1, b_2, b_3의 결합수준으로 구성되는 6개의 처리 (a_1, b_1), (a_1, b_2), (a_1, b_3), (a_2, b_1), (a_2, b_2), (a_2, b_3)에서 일정크기의 염색된 원단의 표면을 방수처리 한다 하자. 실험을 위해 일정크기의 염색된 원단중 임의로 세 가지 색의 원단을 추출하여 이용한다고 가정한다. 추출된 색의 원단을 이용하여 6개의 방수 가공처리를 위해 동일크기의 실험단위로 나눈다고 가정한다. 이 경우 실험에 이용되는 동일 크기의 세 가지 색의 원단은 블록으로 간주되고 처리에 이용되는 동일크기의 천 조각은 실험구로 취급된다. 방수처리가 겉감에 행해져야 할 때, 겉감과 속감에 대한 지정은 임의로 행해지지 않게 된다. 즉, 방수처리된 천의 측정면은 겉과 안의 공간적으로 정해진 지역이므로 연구자가 임의로 배정할 수 없는 특성을 갖게 된다. 방수율의 측정은 방수처리된 표면과 이면에서 측정한다고 가정한다. 왜냐하면, 방수원단의 가공방법과 방수액의 종류에 따라 일정회수의 세탁후에 방수원단의 표면과 이면에서 측정되는 방수율은 차이가 있을 수 있기 때문이다. 이 경우 방수율 측정면은 분할구로 간주되고 각 면에서 측정된 방수율은 동일 실험단위에서 반복측정됨으로 인해서 동일 실험단위내 측정된 두 방수율은 상관성을 갖는 측정값으로 인식된다. 그러나 실험단위 간의 관측값들은 독립적으로 측정되기 때문에 상관성을 갖지 않는다고 가정한다. 원단의 효과를 δ_l $(l = 1, 2, 3)$라 둘 때, 동일 크기의 세 개 원단

은 동일 크기의 원단중 임의로 추출된 세 개이므로 확률효과이다. 처리가 행해지는 일정크기의 천들도 동일 크기의 모집단에서 추출된 실험단위들로 간주되므로 이들의 효과는 확률효과이고 γ_{ijl} ($i = 1, 2$, $j = 1, 2, 3$, $l = 1, 2, 3$)으로 표시된다. 두 가지 방수방법은 고정요인의 두 수준 a_i ($i = 1, 2$)를 나타내므로 이들 수준효과 α_i ($i = 1, 2$)는 고정효과를 나타낸다. 방수액 종류의 세 수준은 고정요인의 세 수준 b_j ($j = 1, 2, 3$)을 나타내므로 수준효과 β_j ($j = 1, 2, 3$)는 고정효과를 나타낸다. 반복측정 요인인 측정공간 T의 두 수준 t_k ($k = 1, 2$)의 효과를 나타내는 τ_k ($k = 1, 2$)도 고정효과들이다. 반복측정 요인 공간 T의 두 수준이 비확률화에 의해 소단위의 소재에 배정됨으로 인해서 동일 처리의 특성을 갖는 소 단위의 소재간에 일정한 상관성을 띄는 공분산 구조를 갖는다고 예상할 수 있다. 확률요인 A와 고정요인 B의 수준결합으로 주어지는 처치가 임의로 배정된 일정 크기의 소재에 행해지고 고정요인 T의 두 수준들이 비확률화에 의해 배정된다고 하자. 다음은 위의 실험으로 부터 수집된 가상의 생성자료표를 나타내고 있다.

표 4.1 방수천 자료의 생성표

원단	요인A	요인B	요인T	
			t_1	t_2
1	a1	b1	82.0	75.0
1	a1	b2	70.8	52.5
1	a1	b3	45.0	35.8
1	a2	b1	63.0	59.0
1	a2	b2	72.8	68.8
1	a2	b3	54.0	45.8
2	a1	b1	85.0	78.0
2	a1	b2	79.8	55.8
2	a1	b3	52.0	40.8
2	a2	b1	65.0	62.0
2	a2	b2	74.8	69.8
2	a2	b3	52.0	42.8
3	a1	b1	79.0	68.0
3	a1	b2	69.8	65.8
3	a1	b3	82.0	70.8
3	a2	b1	60.0	58.0
3	a2	b2	76.8	73.8
3	a2	b3	62.0	62.0

위 자료를 분석하기 위한 모형으로 식 (3.1)에서 세 요인 교호작용이 없다고 가정할 때, 자료에 해당하는 모형은 다음과 같이 주어진다.

$$y_{ijkl} = \mu + \delta_l + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_{ijl} + \tau_k + (\alpha\tau)_{ik} + (\beta\tau)_{jk} + \epsilon_{ijkl}. \tag{4.1}$$

단, $i = 1, 2$, $j = 1, 2, 3$, $l = 1, 2, 3$ 이고 $k = 1, 2$이다. μ, α_i, β_j, τ_k, $(\alpha\beta)_{ij}$, $(\alpha\tau)_{jk}$, 그리고 $(\beta\tau)_{jk}$는 고정효과들이고, δ_l은 원단 l과 연관된 확률효과이다. γ_{ijl}은 원단 l내 처리 ij가 행해진 실험단위 (실험구)의 오차를 나타내는 확률효과이고 ϵ_{ijkl}은 원단블록 l내 처리 ij가 행해진 실험단위 ijl의 k번째 측정면 (분할구)의 오차를 나타낸다.

모형내 모수들을 추정하기 위해 REML 방법을 이용한다. 관측벡터 $\mathbf{y}$의 우도함수는 적절한 변수변환을 이용할 때, 고정효과를 포함하는 부분과 고정효과와 상관없는 두 부분의 결합우도함수로 주어진다. 고정효과를 포함하지 않는 변환벡터의 주변우도함수는 고정효과와 무관한 분산성분들만의 제한우도함

수 (restricted likelihood function)이고 모형내 분산성분들의 REML 추정값은 제한우도함수를 이용하여 구해진다. 모형 (4.1)을 자료에 적합시킨 후, 모형내 유의성을 나타내지 않는 모수를 제외한다. 다음의 축약된 모형을 적합시켜 본다.

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_{ijl} + \tau_k + (\alpha\tau)_{ik} + \epsilon_{ijkl}. \tag{4.2}$$

SAS 프로그램을 이용한 축약모형내 분산성분들의 REML 추정값은 $\hat{\sigma}^2_\gamma$=92.66, 이고 $\hat{\sigma}^2_\epsilon$=11.62이다. 분산성분들에 대한 Wald Z 검정통계량의 관측값들은 유의수준 0.05에서 모두 유의함을 보이고 있다. 모형내 고정효과를 나타내는 모수들의 추정값과 추정오차는 $\hat{\mu}$=49.80(4.7465), $\hat{\alpha}_1$=0.06(4.8140), $\hat{\beta}_1$=15.75(5.7293), $\hat{\beta}_2$=15.53(5.7293), $(\hat{\tau})_1$=4.27(1.6072)이고 $(\widehat{\alpha\tau})_{11}$=7.17(2.2729)이다. 고정효과와 관련된 모수들의 t검정통계량의 관측값들은 유의수준 0.05에서 모두 유의함을 나타내고 있다. 모형의 적합성에 대한 우도비 검정통계량의 관측값은 ${\chi^2}_{(1)} = 24.37$로 관측되고 p값은 0.0001 미만이므로 모형의 유의성을 나타내고 있다. 즉, 방수천 자료에 대한 분석모형으로 방수율이 측정되는 측정면이 분할구로 비확률화에 의해 주어질 때, 측정값들의 상관성을 고려한 혼합모형이 인식되어야 함을 보여주고 있다.

5. 결론

본 논문은 분할구 실험계획의 특성과 자료분석에 요구되는 가정의 일부가 성립하지 않는 경우에 실험자료를 분석하기 위한 혼합모형을 다루고 있다. 즉, 분할구 실험의 특성이 성립되지 않는 공간요인을 가정할 때, 모형의 변화와 자료분석에 필요한 가정은 어떻게 주어져야 하는 문제를 논의하고 있다. 제기된 문제를 다루기 위해 구체적인 실험환경 조건하에서 실험이 행해지고 자료가 수집됨을 가정하여, 자료분석을 위한 혼합모형을 제시하고 있다. 이러한 혼합모형은 분할구 실험의 유형에서 고려되는 요인들의 수, 요인의 유형, 그리고 공간요인의 수준이 실험구에 배정되는 지 또는 분할구에 배정되는 가에 따라 다양하게 제시된다. 공간요인의 수준이 어디에 배정되는 가에 따라 수집된 자료의 분석을 위한 다변량 정규분포에 대한 가정도 달라질 수 있음을 알 수 있다.

본 논문에서의 주안점은 반복측정 요인이 처리의 한 구성요인인 경우를 가정하고 있다. 특히, 연구자가 임의로 배정할 수 없는 일반적인 시간요인 보다는 공간요인이 처리의 한 구성요인으로 작용할 때, 분할구 실험자료를 분석하기 위한 혼합모형을 제시하고 있다. 제시된 혼합모형내 모수를 추론하기 위해 동일 실험구내 관측의 소단위인 분할구에서의 측정값들 간의 상관성을 고려한 공분산 구조하에서 모수의 추정방법과 모형의 적합성을 자료의 예를 통하여 논의하고 있다.

참고문헌

최재성 (2007). 순서형 자료에 대한 비례승산 혼합효과 모형. <한국데이터정보과학회지>, **18**, 471−479.

최재성 (2008). 반복측정의 다가 반응자료에 대한 주변확률 모형. <한국데이터정보과학회지>, **19**, 577−585.

Cochran, W. G. and Cox, G. M.(1957). *Experimental designs*, 2nd edition, John Wiley and Sons, Inc., New York.

Corbeil, R. R. and Searle, S. R. (1976). A comparison of variance component estimators. *Biometrics*, **32**, 779−791.

Graybill, F. A. (1976). *Theory and application of the linear model*, Wadsworth, Inc. California.

Huynh, H., and Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measures designs have exact F-distributions. *Journal of the American Statistical Association*, **65**, 1582−1589.

Mead, R. (1988). *The design of experiments*, Cambridge University Press, Cambridge.

Milliken, G. A. and Johnson, D. E. (1984). *Analysis of messy data*, Van Nostrand Reinhold, New York.

SAS Institute, Inc. (1996). *SAS system for mixed models*, SAS Institute, Inc., Cary, North Carolina.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance components*, John Wiley and Sons, Inc., New York.

Journal of the Korean Data &
Information Science Society
2010, **21**(1), 1–9

한국데이터정보과학회지

A mixed model for repeated split-plot data †

Jaesung Choi[1]

[1]Department of Statistics, Keimyung University

Abstract

This paper suggests a mixed-effects model for analyzing split-plot data when there is a repeated measures factor that affects on the response variable. Covariance structures are discussed among the observations because of the assumption of a repeated measures factor as one of explanatory variables. As a plausible covariance structure, compound symmetric covariance structure is assumed for analyzing data. The restricted maximum likelihood (REML)method is used for estimating fixed effects in the model.

Keywords: Covariance structure, repeated measures factor, split-plot, treatment.

† The present research has been conducted by the Bisa Research Grant of Keimyung University in 2009.
[1] Professor, Department of Statistics, Keimyung University, 1000 Shindang-Dong, Dalseo-Gu, Daegu 704-701, Korea. E-mail: jschoi@kmu.ac.kr