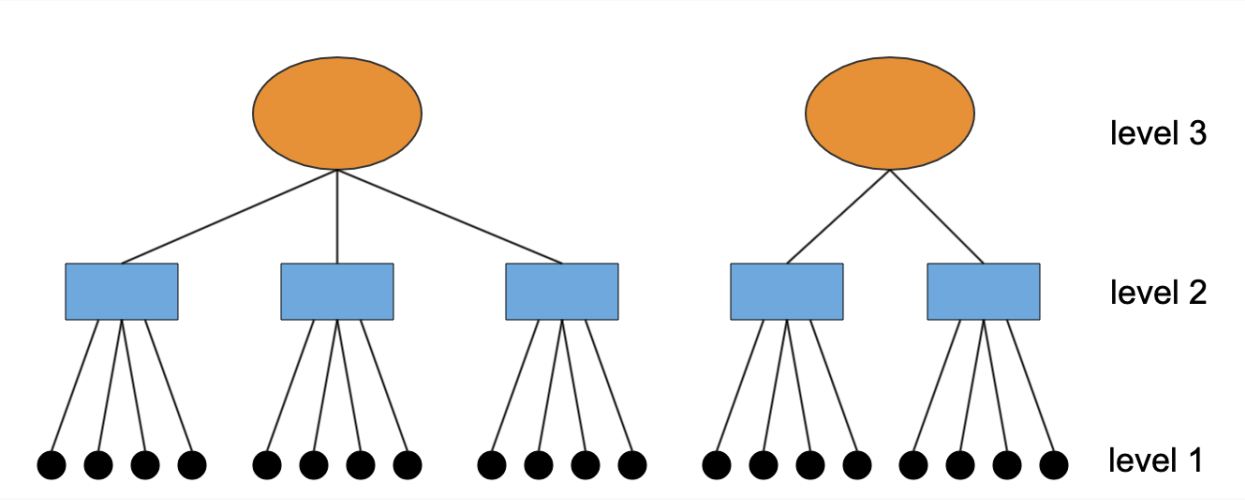
계층모형

기본 개념의 소개

서울시립대학교 통계학과 이용희
2021년 5월 11일

계층모형

- **계층모형 (hierarchical model)**은 반응변수에 영향을 미치는 요인 또는 효과가 하나 이상의 계 층(levels) 또는 군집(clusters)으로 구성된 통계적 모형이다.



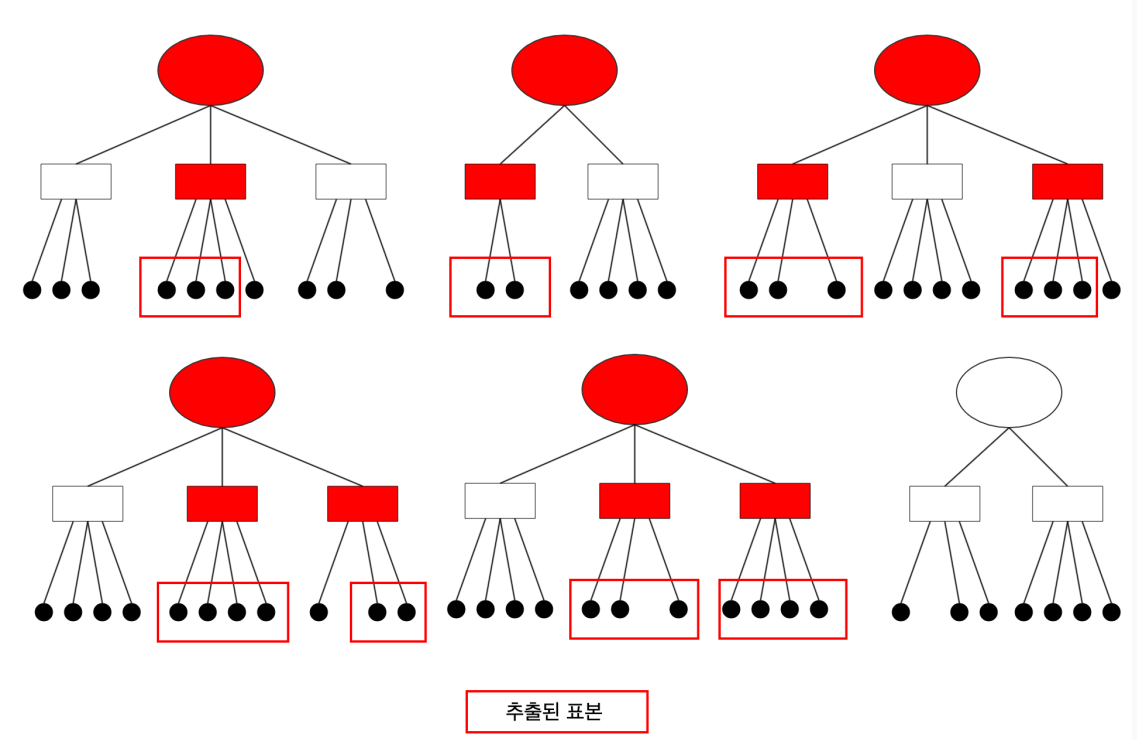
- 계층모형은 다양한 분야에서 여러 가지 다른 이름으로 사용된다.
  - 분할구 계획법 (split-plot designs)
  - 반복측정 (repated measures, longitidinal data)
  - 패널 자료 (panel data)
  - 다중수준 모형 또는 다중계층모형 (multilevel model)
  - 혼합효과모형 (mixed-effect model)

계층모형의 특징

- 계층모형의 가장 큰 특징: **같은 계층(군집)에 속하는 관측값은 독립이 아니다!**
- 예를 들어 다음과 같은 계층 구조로 모의고사 성적을 수집하였다고 가정하자.
    - 학교 (level 3)
    - 교사 (level 2)
    - 학생 (level 1)
- 동일한 교사에게 수업을 들은 학생들은 교사의 고유한 특성(수업방법, 교수법, 성실성, 능력...)에 영향을 받는다.
- 동일한 학교에 속한 교사와 학생들은 학교의 교유한 특성(학교의 형태, 물적/인적 자원, 지도자...)에 영향을 받는다.
- 따라서 같은 계층 또는 군집에 속한 학생들의 모의고사 성적은 독립이 아니다.
- 모든 관측값이 독립이라고 가정하는 전통적인 회귀모형에 대한 추론 방법과는 다른 추론 방법이 필요하다.

계층모형의 특징

- 자료는 각 계층에 속하는 추출단위를 계층적으로 임의추출한다. 예를 들어 학생들의 모의고사 성적에 대한 분석을 수행하기 위하여
  - (1) 학교를 추출하고 (2) 교사를 추출하고 (3) 학생을 추출한다



- 이러한 자료의 구조에서 **학교의 효과**, **교사의 효과** 는 전통적인 효과(예를 들어 온도, 압력, 교수법, 정책 등)와는 다르다.

고정효과와 임의효과

- **효과(effect)**는 관심 있는 반응 변수(response variable)가 변할 수 있는 원인을 제공하는 **요인(factor)**이다.
- 요인은 여러 개의 **수준(level)**을 가질 수 있다.
- **고정효과 (fixed effect)**는 수준의 수가 고정되어 있거나 유한 개로 조정할 수 있는 효과를 말한다.
  - 라면을 요리하는 경우 맛에 영향을 미치는 중요한 요인은 물의 양이며 물의 양은 550, 600, 650 ml 의 수준을 가 질 수 있다.
  - 의학 연구에서 성별은 중요한 요인이며 남자와 여자로 수준이 나타난다.
  - 재난 지원금은 사람들의 경제 활동에 중요한 영향을 미치는 요인이며 모든 국민에 대한 지원과 선별 지원 중 하나를 선택해야 한다.
- **임의효과 (random effect)** 는 무한 개에 가까운 수준들에서 관측된 또는 추출된 효과를 의미한다. 일반적으로 군집효과 또는 블럭(block, batch)효과로 볼 수 있다.
  - 학교와 교사의 효과
  - 임상실험에서의 병원 효과
  - 농업실험에서의 재배지(plot) 효과
  - 생산 공장에서 공정라인 효과 (batch effect)
  - 패널자료 또는 반복측정자료에서 개인 효과

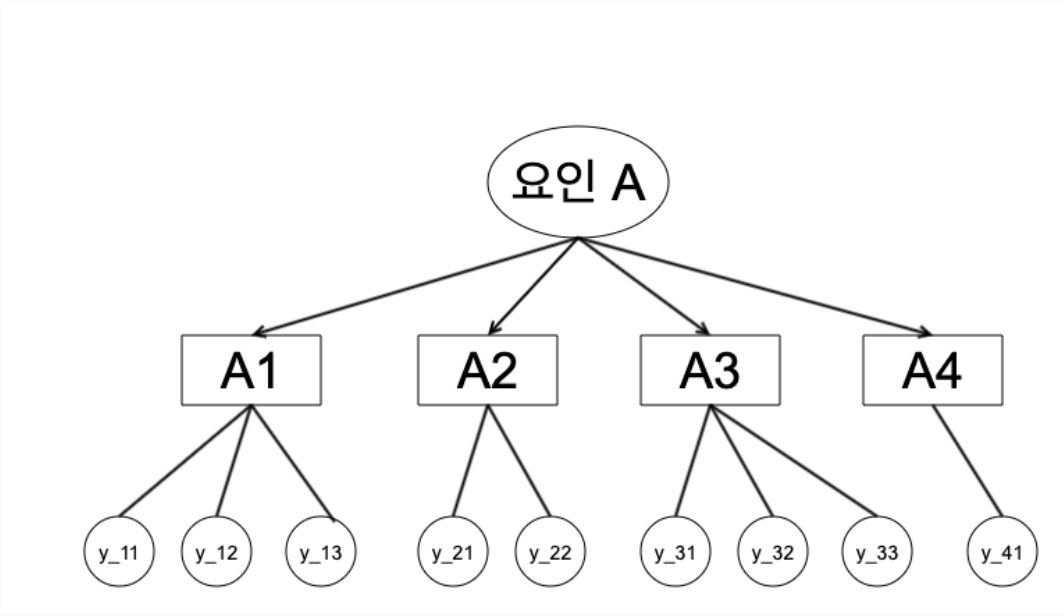
고정효과와 임의효과

- 고정효과와 임의효과를 구별할 수 있는 정확한 정의는 없으며 고정효과가 경우에 따라서 임의효과로 나타나는 경우도 많다.

> Before proceeding further with random field linear models we need to remind the reader of the adage that one modeler's random effect is another modeler's fixed effect (Schabenberger and Pierce ,2001, p. 627.)

- 일반적으로 고정효과와 임의효과를 구별하는 기준은 다음과 같다.
  - 고정효과
    - 기술적인 효과(technical effect)
    - 실험자가 기술적으로 반복하여 적용할 수 있는 효과
    - 평균 효과의 비교가 주 목적인 경우
  - 임의효과
    - 효과가 있는 것 같은데 기술적으로 명확한 설명이 어려운 효과 (Unobservable heterogeneity)
    - 숨겨진 변수 (latent variable)
    - 모집단에서 추출된 집단(group, cluster, repeated menasure)에 속하여 나타나는 효과

고정효과와 임의효과

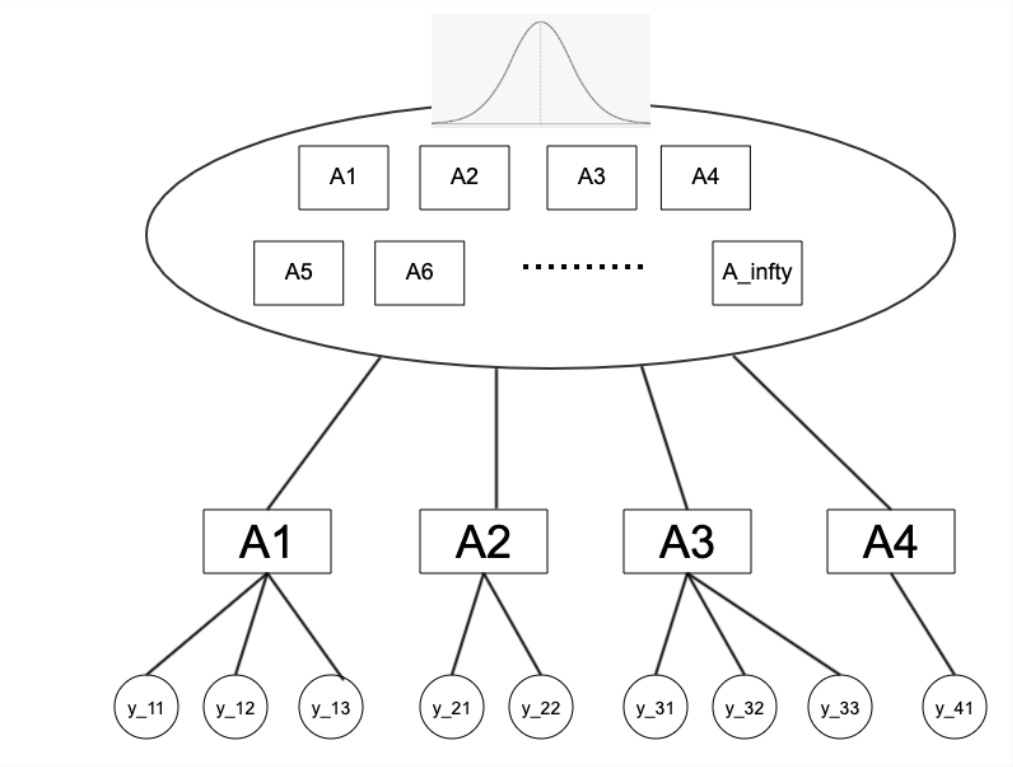


- 통계적 모형에서 고정효과
- 관측값 y_{ij}를 요인 A의 i번째 수준이 적용된 j번째 관측값이라고 하자.

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

- 위 모형에서 μ는 전체 평균을 나타내는 모수(parameter; 고정된 모르는 숫자)이다.
- $\alpha_1, \alpha_2, \cdots, \alpha_I$는 요인 A의 i번째 수준에 대한 고정효과로서 모수이다.
- 오차항 e_{ij}는 모두 독립이고 평균이 0 이며 분산이 σ_E^2인 정규분포를 따른다.
- 모형에서 관측값에 임의성(randomness)를 주는 유일한 요인은 오차항이다.

고정효과와 임의효과



- 통계적 모형에서 임의효과

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

- 위 모형에서 μ는 전체 평균을 나타내는 모수(parameter; 고정된 모르는 숫자)이다.

- $\alpha_1, \alpha_2, \cdots, \alpha_I$는 요인 A에 대한 임의효과로서 평균이 0 이며 분산이 σ_A^2인 정규분포를 따른다.

$$\alpha_i \sim N(0, \sigma_A^2)$$

- 오차항 e_{ij}는 모두 독립이고 평균이 0 이며 분산이 σ_E^2인 정규분포를 따른다.

- 모형에서 관측값에 임의성(randomness)를 주는 요인은 2개로서 임의효과와 오차항이다.

임의효과에 의한 관측값들의 종속성

- 임의효과 모형에서 같은 수준(level, cluster, block, batch, group)에 속하는 반응변수들은 공통의 임의효과(확률변수)를 공유한다.
- 예를 들어 아래 두 모형식에서 같은 그룹에 속하는 두 관측값 y_{11}과 y_{12}는 동일한 확률변수인 임의효과 α_1을 공유하므로 독립이 아니다.

$$y_{11} = \mu + \alpha_1 + e_{11}$$

$$y_{12} = \mu + \alpha_1 + e_{12}$$

- 참고로 오차항 e_{11}과 e_{12}는 서로 독립이며 더 나아가 임의효과 α_1과도 독립이다.
- 두 관측값 y_{11}과 y_{12}의 상관계수는 **그룹내 상관계수(Intra Class Correlation; ICC)**라고 부르며 다음과 같다.

$$ICC = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2}$$

- 다른 그룹에 속하는 관측값들은 서로 독립이다.

변동의 분해 - 그룹간 변동과 그룹내 변동

- 4개의 영어 학원을 추출하고 각 학원에서 20명의 학생들을 추출하여 영어 모의고사을 실시한 성적들
- 임의효과 모형

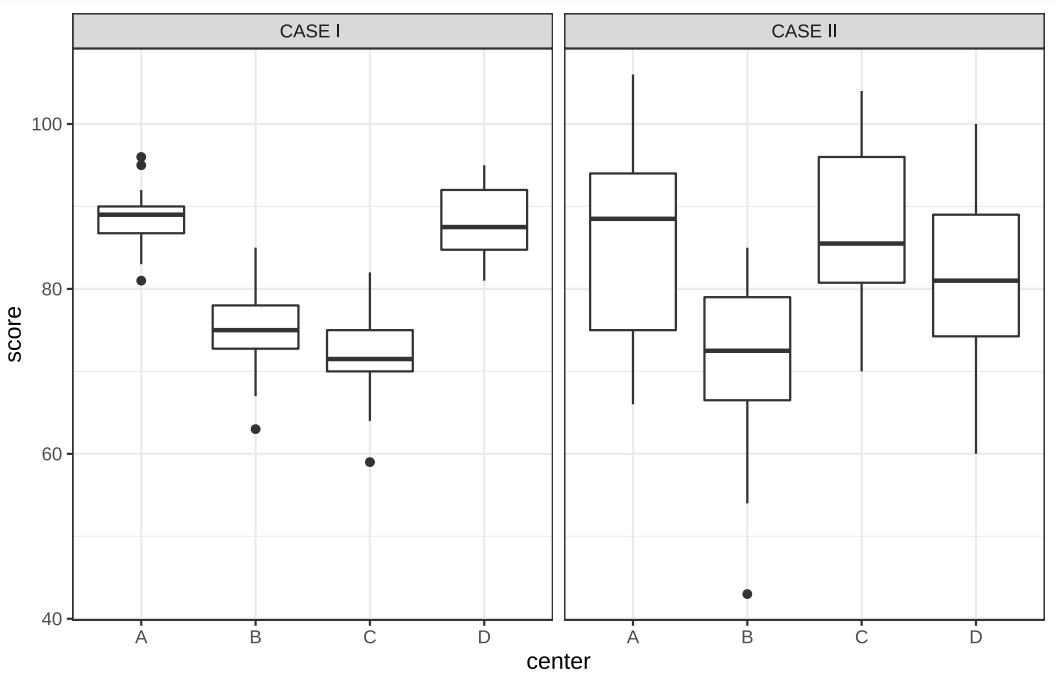
$$\text{score}_{ij} = \mu + \text{center}_i + e_{ij}$$

- 학원효과(임의효과) : 그룹간 변동

$$\text{center}_i \sim N(0, \sigma_A^2)$$

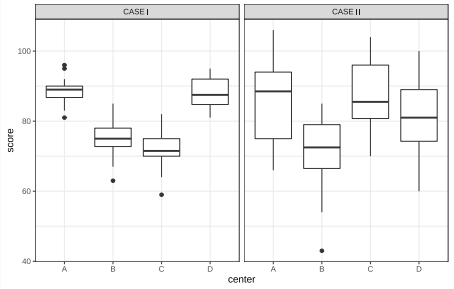
- 오차항(학생의 변동) : 그룹내 변동

$$e_{ij} \sim N(0, \sigma_E^2)$$



- 경우 I : 그룹간 변동 > 그룹내 변동
- 경우 II :그룹간 변동 < 그룹내 변동

변동의 분해



- 그룹간 변동(between-group variation): $Var(center_i) = \sigma_A^2$
- 그룹내 변동(within-group variation): $Var(e_{ij}) = \sigma_E^2$
- 총변동의 분해

$$Var(score) = Var(center) + Var(e) = \sigma_A^2 + \sigma_E^2$$

- 그룹내 상관계수

$$ICC = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2}$$

- 그룹간의 변동이 상대적으로 클수록 그룹내 상관계수가 커진다.
  - 예제 그림에서 $ICC_I > ICC_{II}$

그룹 평균의 추정

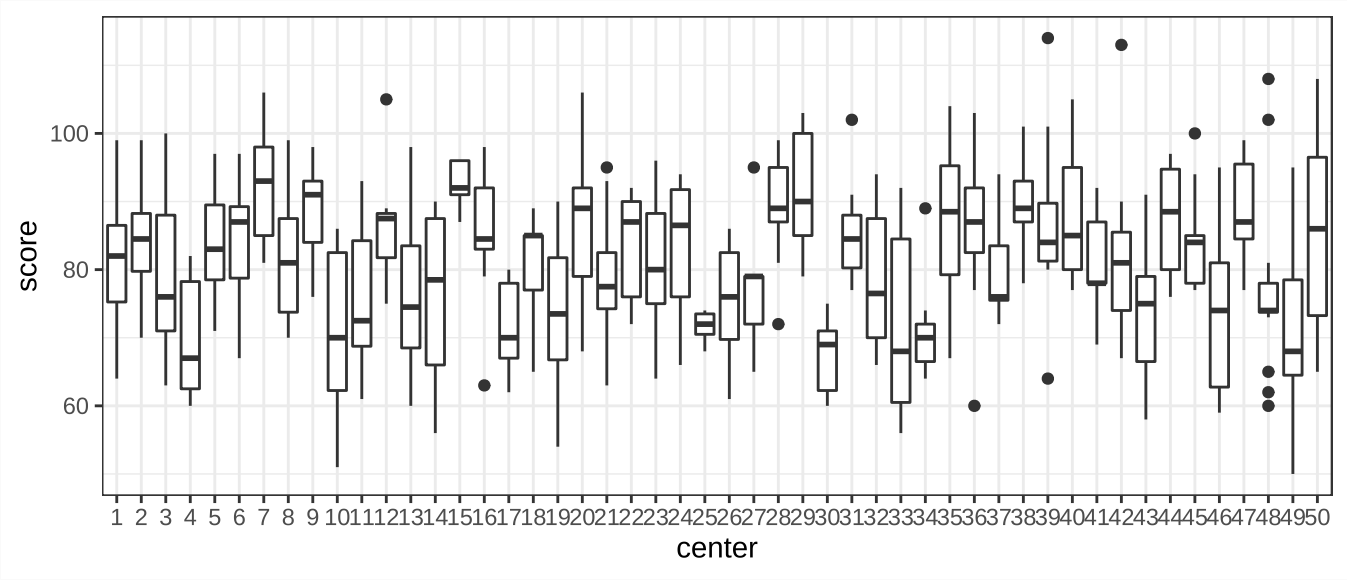
- 계층모형에서는 전체 평균에 대한 추정도 중요하지만

$$E(\text{score}_{ij}) = \mu$$

- 각 그룹 평균에 대한 추정도 중요하다.

$$E(\text{score}_{ij}|\text{center}_i) = \mu + \text{center}_i = \mu_i$$

- 예를 들어 50개의 학원를 추출하고 각 학원에서 n_i 명의 학생들을 추출했다고 가정하자 (아래 그림)
  - 각 학원에 속하는 학생들의 평균 성적은?

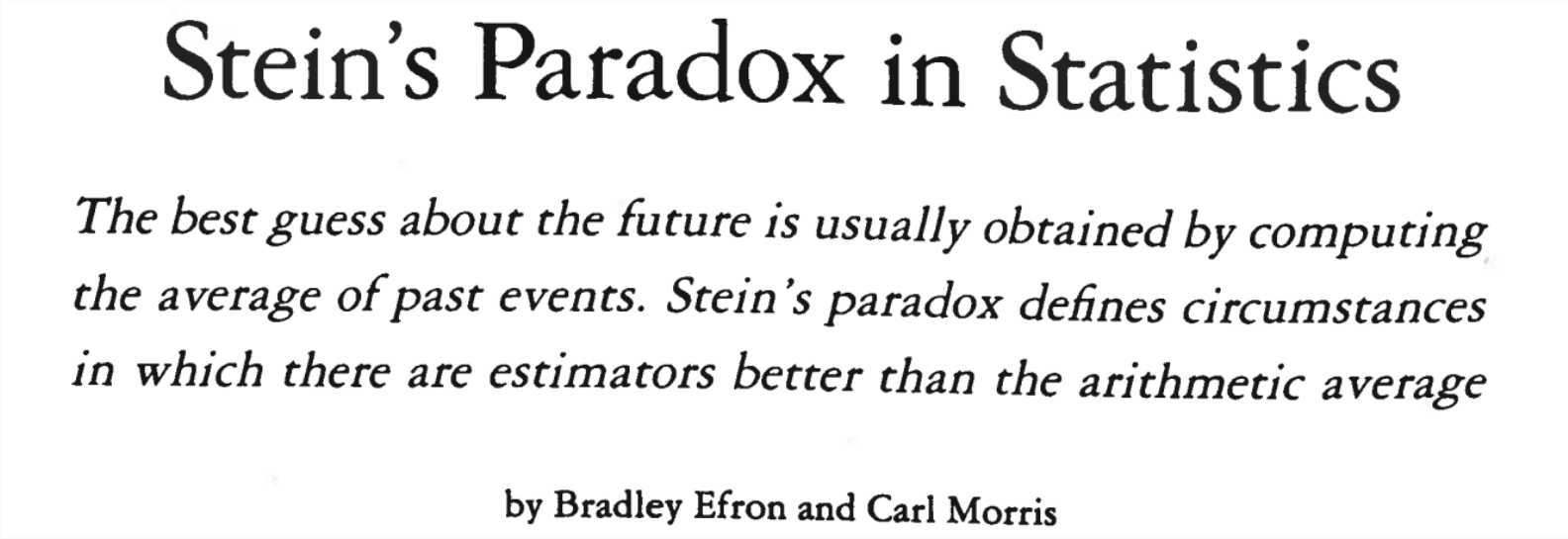


그룹 평균의 추정: 표본평균과 수축 추정량

- 그룹의 평균에 대한 추정량은 그룹에 속하는 관측값들의 표본평균이 가장 직관적이다.

$$\hat{\mu}_i = \frac{\sum_{i=1}^{n_i} y_{ij}}{n_i}$$

- James-Sterin 수축 추정량(shrinkage estimator)
- Efron and Morris, 1977



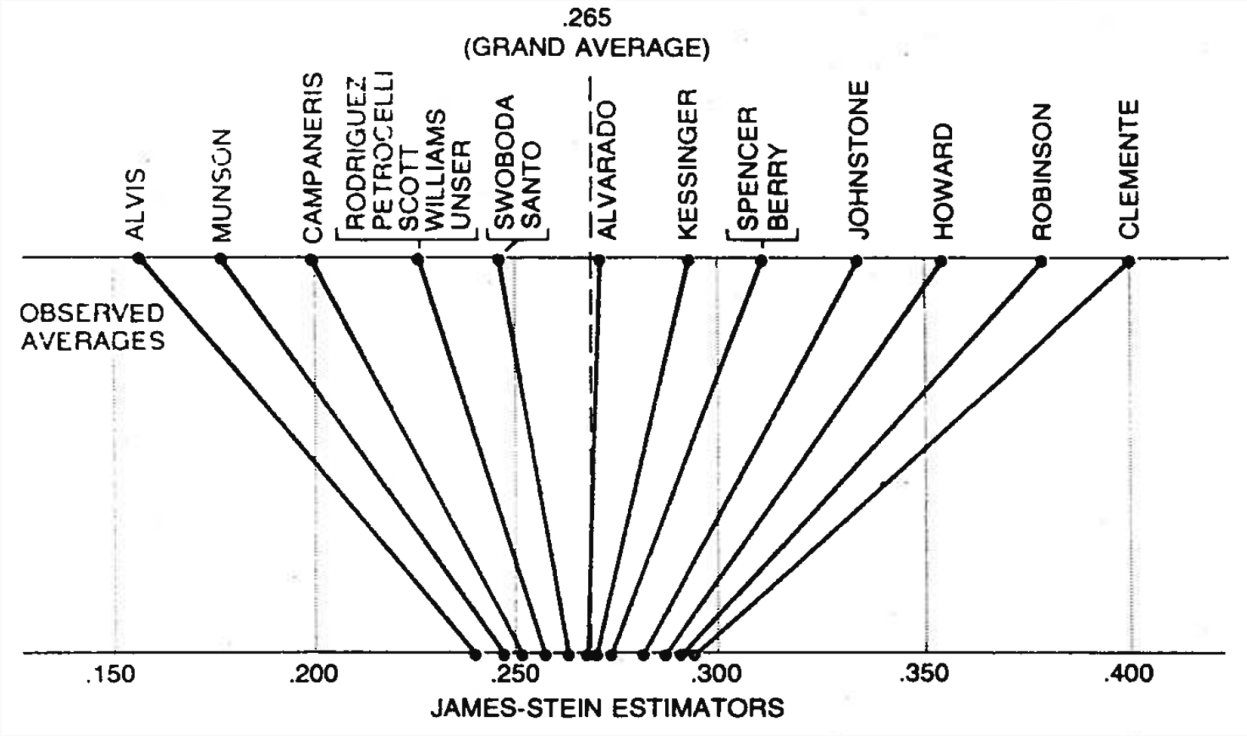
Stein's Paradox in Statistics

The best guess about the future is usually obtained by computing the average of past events. Stein's paradox defines circumstances in which there are estimators better than the arithmetic average

by Bradley Efron and Carl Morris

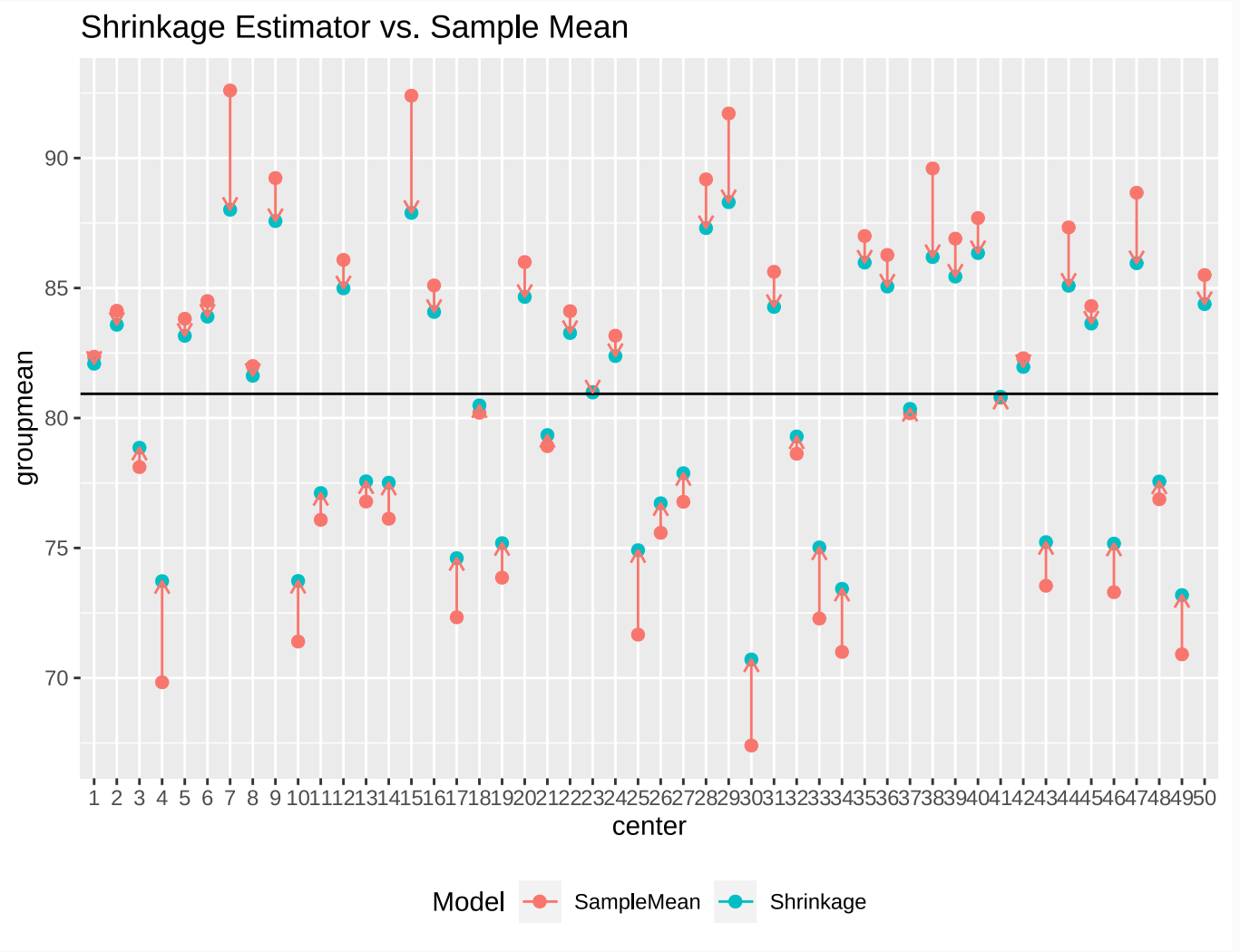
그룹 평균의 추정: 표본평균과 수축 추정량

- 1955년 Charles Stein 이 **3개 이상의 평균**을 동시에 추정하는 문제에서 표본평균보다 더 효율적인 추정량을 제시한다
  - 18 명의 미국 야구 선수
  - 시즌 초기 45번 타석에 대한 타율 자료를 이용하여 전 시즌에 대한 최종 타율을 추정
  - 전체평균 $\bar{y}$와 i 번째 그룹의 표본평균 $\bar{y}_i$의 사이에 있는 추정값이 표본 평균보다 효율적이다 (평균제곱오차가 작다)
  - $\hat{\mu}_i^{JS} = \bar{y} + c(\bar{y}_i - \bar{y})$: **전체 평균으로 수축(Shrink to grand mean)**



그룹 평균의 추정: 표본평균과 수축 추정량

- 50개 학원에 대한 표본 평균과 계층모형에서의 수축 추정량 비교



그룹 평균의 추정: 표본평균과 수축 추정량

- 계층모형에서는 **임의효과에 대한 조건부 예측값**과 **고정효과 모수의 추정값**을 결합하여 그룹 단위에 대한 평균을 추정한다.

$$\hat{\mu}_i = \hat{\mu} + \hat{\alpha}_i$$

- 임의효과에 대한 조건부 예측값을 구할 때 평균(0)으로 수축되는 현상이 나타나며 그룹내 상관계수가 작을수 수축이 크게 일어난다.
- 이러한 수축현상을 . John Tukey 는 그룹 간의 정보를 공유하여 이용하는 것이라고 표현하였다 (**"borrowing strength" between groups**)
- 다시말하면, 표본 평균은 그룹 내의 정보만을 이용하지만(**pooling no information**) 계층모형은 그룹 가 정보와 그룹 내의 정보를 동시에 이용한다(**pooling partial information**)
- 극단적으로 그룹을 무시하고 모든 자료를 합쳐서 추론하면(**pooling information completely**) 그룹 간의 변동 또는 그룹 내의 변동은 파악할 수 있다.

계층모형에 대한 추론 방법 및 확장

추정법

- 최소제곱법과 분산분석 (전통적인 방법)
- 다변량 분석법 (전통적인 방법)
- 최대 가능도 추정법 (Maxinum Likelyhood Estimation; ML)
- 제한적 최대 가능도 추정법 (Restrict Maxinum Likelyhood Estimation; REML)
  - REML 은 분산성분(variance component)에만 적용된다.
- 베이즈 추정법

확장

- 선형 혼합모형
- 일반화 선형 혼합모형(이항변수, 범주형 변수, 횟수 자료, ...)
- 비선형 혼합모형
- 다변량 혼합모형