I장 | 실험 계획과 분산 분석

SAS를 이용한 실험 계획과 분산 분석 (자유아카데미)

서론

- 현대 사회의 대부분의 학문이나 생활 영역에서 실험은 여러가지 모양으로 행해지고 있다. 자연과학에서는 제품의 품질에 영향을 주는 실험조건이 무엇이며, 이에 따른 최적조건을 알아보는데 실험이 사용되고 있다. 한편 의료 보건 분야에서는 새로 개발된 약이나 치료방법이 기존의 약이나 기존 치료방법에 비해 효능면에서 비교 우위에 있는 지를 검정하는데 실험이 사용되고 있다.

서론

- 시간
- 비용
- 신뢰성
- 정확성

실험 계획의 여러가지 목적

- 두 변수 간의 인과관계를 규명한다
- 여러 설명 변수 중에 반응변수에 영향을 주는 변수를 선택한다
- 반응 변수가 최대(최소)값을 가지기 위한 설명 변수의 값을 구한다.

실험 계획의 필요성

- 실험의 횟수를 최소화하여 실험 경비를 줄인다.
- 실험 도중 예상치 못한 사고가 발생하더라도 그 영향을 최소화한다.

실험 계획을 위한 단계별 질문

- 실험을 통하여 구체적으로 무슨 정보를 얻기 원하는가?
- 실험의 반응 변수의 구체적인 단위는 무엇인가?
- 실험의 반응변수에 영향을 줄 수 있는 설명 변수는 무엇인가?
- 반응변수에 영향을 줄 수도 있는 외생 변수는 무엇인가
- 실험을 마친 후에는 어떤 분석 방법을 사용할 것인가?

참고

- Response Variable(반응변수) : 실험을 통해 개체(subject)로부터 반응을 측정하고자 하는 변수
- Explanatory Variable(설명변수) : 반응변수에 영향을 줄 것으로 예상되는 실험에 주관심이 되는 변수
- Extraneous Variable(외생변수) : 실험의 주관심은 아니지만 반응변수에 영향을 다소 미칠 것으로 우려되는 노이즈(noise)변수

- 외생변수의 영향을 최소화 시키면서,
- 반응변수에 끼치는 설명변수의 영향 조사

아래 실험에 관한 문장에는 설명변수, 반응 변수, 외생 변수가 숨어있다.

- 백화점에서는 일년 동안 새해맞이 세일, 추석 맞이 세일, 성탄 맞이 세일의 3 종류의 세일이 있다고 가정하자. 각 세일 기간 사이에 매출액의 차이가 있는지 알아보려고 신세계 백화점, 현대 백화점, 롯데 백화점을 상대로 5년간 조사하였다 (모든 세일기간은 일정하다고 가정).

실험을 마친 후 필요한 질문

- 실험계획에 따라 실험이 수행되었는가?
- 실험 계획과 일치하는 분석 방법이 수행되었는가?
- 추후 유사한 실험을 한다면 유념할 점은 무엇인가?

실험계획 용어

- Data 의 종류
  Experimental Data (실험계획법)
  Observational Data (회귀분석, 시계열, 다변량, etc)
  Survey Data (표본조사론)
- 실험단위 (Experimental Unit) : 실험조건이 행해지는 최소단위

(e.g. 기계, 논, 환자, ...)

- 실험오차 (Experimental Error): 같은 실험조건을 받은 다른 실험단위 사이에 발생하는 자연스런 오차
- 처리군 (treatment group) : 관심있는 실험조건에 노출된 집단
- 대조군 (control group) : 처리군과 달리, 관심있는 실험조건에 노출되지 않은 집단

실험계획의 기본 원리

- 반복의 원리 (replication) 같은 실험조건에 개체수를 증가시킴으로써 개체가 가지고 있는 외생변수의 영향을 상쇄시킨다
- 랜덤화의 원리 (randomization) 실험단위를 처리군/대조군에 랜덤배치함으로써 개체가 가지고 있는 외생변수의 영향을 상쇄시킨다
- 블록화의 원리 (blocking) 같은 외생변수 값을 지닌 실험단위를 블록화시킴으로써 불록내 실험단위를 비교함으로써 외생변수의 영향을 상쇄시킨다.

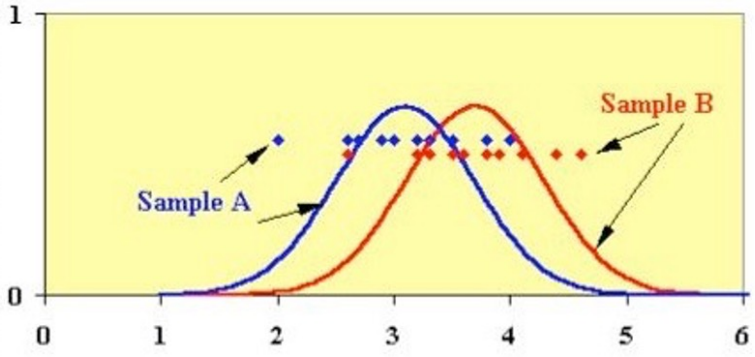
추가로 고려해야 할 '실험계획의 기본원리'는?

- 두종류 (A/B)의 다트 제품이 있다. 어느 제품이 더 명중률이 높은지 실험하기 위하여 A회사, B회사의 제품을 10개씩 구입해서, 일정 거리에서 같은 사람이 던져 보았다.A회사 제품을 던져보고, 10분후 B회사 제품을 던져봐서 그중 명중된 다트의 개수를 기록하였더니 B회사 제품이 훨씬 우수하였다.
- 알코올이 신체의 반응신경에 끼치는 영향을 조사하기 위하여, 자원봉사자 20명을 대상으로 개인 주량을 무시하고 음주 전/후의 반응신경테스트 점수를 기록하였더니, 음주 전/후의 반응 신경 테스트 점수 차이가 유의하지 않았다.

두 모집단의 모평균 비교

두 모집단 모평균 비교

$$H_0 : \mu_1 = \mu_2$$



$$Sample\ A : y_{11}, y_{12}, \cdots, y_{15} \sim N(\mu_1, \sigma^2)$$

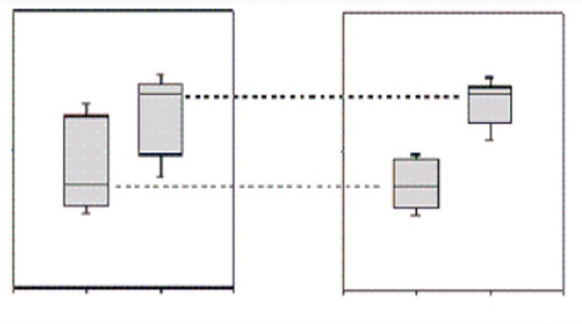
$$Sample\ B : y_{21}, y_{22}, \cdots, y_{25} \sim N(\mu_2, \sigma^2)$$

T검정 통계량

	$x_{ij} = -1$ (남자)	$x_{ij} = +1$ (여자)
관측값	$y_{11} = 327$ $y_{12} = 291$ $y_{13} = 323$ $y_{14} = 284$ $y_{15} = 305$	$y_{21} = 308$ $y_{21} = 324$ $y_{23} = 353$ $y_{24} = 344$ $y_{25} = 341$
분포가정	$y_{1j} \overset{iid}{\sim} N(\mu_1, \sigma^2)$	$y_{2j} \overset{iid}{\sim} N(\mu_2, \sigma^2)$
가설	$H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$	
검정통계량	$t_0 = \frac{\bar{y}_{1\cdot} - \bar{y}_{2\cdot}}{S_p\sqrt{\frac{1}{5}+\frac{1}{5}}} = -2.398$,	$S_p^2 = \frac{(5-1)S_1^2+(5-1)S_2^2}{(5-1)+(5-1)} = 340.75$
기각역	$\lvert t_0 \rvert > t_{\frac{0.05}{2}, 5+5-2} = 2.306$	

표 1.1: 남자와 여자의 시험성적 비교를 위한 t 검정

분산 분석



- 평균을 비교함에 있어서 **분산의 중요성**
- left : Within Variance > c (Between Variance)

>>>> no significant mean difference (평균 간 차이가 유의하지 않다)

- right : Within Variance < c (Between Variance)

>>>> significant mean difference (평균 간 차이가 유의하다)

분산분석

전체제곱합(SST)의 분할

$$\begin{aligned} \text{SST} &= \text{SS(between)} + \text{SS(within)} \\ &= (\text{처리제곱합}) + (\text{오차제곱합}) \\ &= \text{SStreat} + \text{SSE} \end{aligned}$$

분산 분석의 F검정

분산분석(ANOVA)의 F 검정

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a$ 에 대해,

Source	d.f.	S.S.	M.S.	F_0
Treatment	$a-1$	$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2$	MStreat	$\frac{\text{MStreat}}{\text{MSE}}$
Error	$\sum_{i=1}^{a}(n_i-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2$	MSE	
Total	$\sum_{i=1}^{a}n_i-1$	$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{\cdot\cdot})^2$		

표 1.2: 분산분석표(ANOVA table)의 모습

만일 $$F_0 = \frac{\text{MStreat}}{\text{MSE}} = \frac{\text{SStreat}/(a-1)}{\text{SSE}/\sum_{i=1}^{a}(n_i-1)} > F_{\alpha,\,(a-1),\sum_{i=1}^{a}(n_i-1)} \quad (1.11)$$

이면, H_0를 기각하므로 '**그룹 간 모평균의 차이는 유의하다**'고 한다.

F검정 통계량

	$x_{ij} = -1$ (남자)	$x_{ij} = +1$ (여자)
관측값	$y_{11} = 327$	$y_{21} = 308$
	$y_{12} = 291$	$y_{21} = 324$
	$y_{13} = 323$	$y_{23} = 353$
	$y_{14} = 284$	$y_{24} = 344$
	$y_{15} = 305$	$y_{25} = 341$
분포가정	$y_{1j} \overset{iid}{\sim} N(\mu_1, \sigma^2)$	$y_{2j} \overset{iid}{\sim} N(\mu_2, \sigma^2)$
가설	$H_0 : \mu_1 = \mu_2$ vs.	$H_1 : \mu_1 \neq \mu_2$
검정통계량	$F_0 = \frac{\text{MStreat}}{\text{MSE}} = \frac{\text{SStreat}/(2-1)}{\text{SSE}/\sum_{i=1}^{2}(5-1)} = 5.75$	
기각역	$F_0 > F_{0.05,(2-1),5+5-2} = 5.318$	

표 1.4: 남자와 여자의 평균 시험점수 차이에 대한 ANOVA의 F 검정

분산분석을 위한 용어 정리

- 전체제곱합 = Total Sum of Squares = SST
- 처리제곱합 = Treatment Sum of Squares = SStreat
- 오차제곱합 = Error Sum of Squares = SSE
- 처리평균제곱 = Treatment Mean Square = MStreat
- 평균제곱오차 = Mean Square Error = MSE
- 분산분석 = ANOVA = Analysis of Variance
- 자유도 = degrees of freedom = d.f.

세 집단 이상의 모평균 비교

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a$$

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a$

$$\Rightarrow \begin{cases} H_{01} & : \mu_1 = \mu_2 \\ H_{02} & : \mu_1 = \mu_3 \\ \vdots & \\ H_{0m} & : \mu_{a-1} = \mu_a \end{cases}$$

$E_i = \{H_{oi}$가 참일 때 H_{oi}를 기각하는 사건$\}$

$\Rightarrow \{H_o$ 가 참일 때 H_o를 기각하는 사건$\} = \bigcup_{i=1}^{m} E_i$

$\Rightarrow P(\bigcup_{i=1}^{m} E_i) \leqq \sum_{i=1}^{m} P(E_i)$ (본페로니 부등식)

Carlo Emilio Bonferroni

$\Rightarrow$ 실제 유의수준 $= P(reject\ \ H_0 | H_0\ is\ \ true)$

$\leqq mP(reject\ \ H_{0i} |\ \ H_{0i}\ \ is\ \ true\ \ )$

$= m \cdot (0.05)$

척도의 종류

- 연속형 척도 (Continuous Scale)

✓구간척도(interval scale) : 온도

✓비율척도(ratio scale) : 길이, 무게

- 이산형 척도 (Discrete Scale)

✓명목척도 (nominal scale): 성별

✓순서척도 (ordinal scale) : 학점

척도의 분석 방법

x	y	통계분석방법
구간/비율	구간/비율	회귀분석
구간/비율	명목/순서	범주형 자료분석
명목/순서	구간/비율	실험계획법
명목/순서	명목/순서	범주형 자료분석