1장. 자료의 생성

박명현

서울대학교 학부대학

Overview

통계학의 어원

- 통계학(statistics)의 어원: 국가(state)와 산술(arithmetic)의 합성어
- 과거에 통계가 국가정치 및 경영과 밀접한 관련이 있었음

예) - 고대 이집트에서는 BC 3000년경 피라미드 건설을 위한 인구조사를 실시한 기록

- BC 1500년경의 출이집트 후 이스라엘의 인구조사를 담고 있는 구약성경의 민수기(Numbers)는 '사람의 수'를 의미
- 인도에서는 16세기에 행정통계조사에 대한 기록이 정리되어있음

통계의 사용

- 현대에는 자연과학, 공학, 의학, 사회, 심리, 역사, 경제, 경영 등 모든 학문 분야 및 실생활에서 통계가 사용됨

- 약품 A가 약품 B보다 더 효과가 있는가?
- 어린이들에게 읽기를 가르치는 새로운 방법이 효과적인가?
- 어떤 방파제가 다음 100년 동안 파고를 견뎌낼 것인가?
- 금융회사의 현재 포지션의 리스크는 어느 정도인가?
- 30년 후 우리나라 인구분포는 어떻게 될 것인가?

"통계학은 모든 과학의 하인이다"
- 네이만(J.Neyman) -

통계의 사용

- 최근에는 컴퓨터의 발달과 더불어 대용량 정보를 신속하게 처리하는 것이 가능해져서 위성사진의 판독, 카드 부정사용의 적발, 유전자 정보분석 및 영상인식 등에서도 통계적인 이론과 방법론이 다양하게 적용되고 있다.
- 또한 오랜기간 축적된 대용량의 데이터 및 인터넷의 활용 등으로 인해 생겨나는 방대한 실시간 데이터인 빅데이터를 활용하여 의미있는 정보를 추출해 내는것이 매우 중요한 작업이 되었다.
- 즉 정보화 시대인 현대사회에서는 단순히 통계 수치 자료를 제시하는 것만이 아니라 방대한 데이터베이스로부터 사용자가 필요로 하는 정보를 취합하고 추출하여 새로운 정보 가치를 창출해 내어 의사결정에 반영하는 통계학의 역활이 더욱 강조되고 있다.

여론조사 예시

예) 어느 시의 시장 선거 결과 예측

- 유권자 지지 성향에 대한 여론조사를 실시
- 목적: 각 후보에 대한 전체 유권자의 지지율을 파악하는 것
- 모집단: 투표 당일 투표장에 갈 유권자 전체

표 1–1 어느 시의 시장 선거 여론조사 결과

후보명	지지율
John Smith	42%
Emily Johnson	37%
Michael Lee	15%
Sarah Davis	6%

여론조사가 실시되는 과정

1. 전체 유권자 중에서 몇 명의 유권자를 어떻게 뽑을까?
   - 시간적, 경제적인 이유로 전체가 아닌 일부만 조사
   - 뽑힌 유권자가 전체 유권자의 성향을 잘 대표할 수 있어야 한다. 즉, '골고루' 뽑아야 한다.

     예) 1936년 미국 대통령 선거:

     클럽회원 명단에서 파악한 주소로 우편조사, 전화조사의 단점

2. 수집한 자료를 어떻게 정리 요약할 것인가?
   - 표본에서 얻은 자료를 정리, 분석하여 모집단의 특성을 알아내고자 한다.

여론조사가 실시되는 과정

3. 조사한 내용으로부터 전체 유권자의 성향에 대하여 어떻게 결론을 내릴 것인가?
   - 표본에서 구한 특정 후보의 지지율이 모집단에서의 지지율을 얼마나 잘 반영할까?

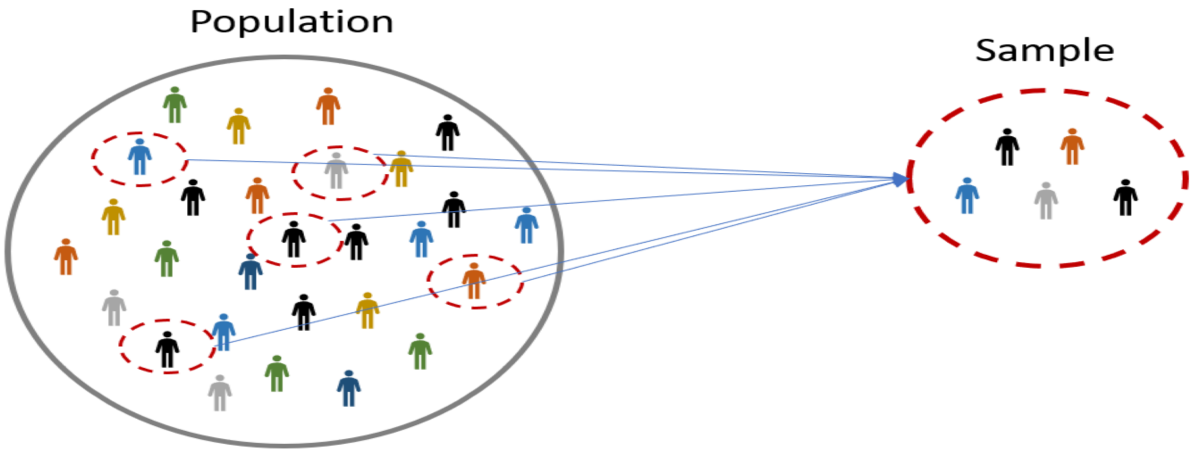
     예) 전체 유권자의 지지율에서도 John Smith 후보가 Emily Johnson 후보를 앞선다고 결론 내릴 수 있는가.

4. 결론에 대한 신뢰성을 어떻게 측정할 것인가?
   - 일부만으로 조사를 하여 전체를 알아보려고 하니 오차가 수반될 수밖에 없음

용어 정리

- 모집단(population): 정보를 얻고자 하는 대상이 되는 전체 집단
  - 유한모집단finite population
    - 유한한 개수의 추출단위로 구성된 모집단
    - 예시) 특정 년도의 시장 선거에서 유권자 전체
  - 무한모집단infinite population
    - 무한한 개수의 추출단위로 구성된 모집단
    - 실제로 존재하는 것이 아닌 가상의 개념
    - 예시) 앞으로 모든 세대의 스마트폰 사용자와 그들의 수면 자료
- 표본(sample): 모집단에 관한 정보를 얻기 위하여 실제로 관측한 것들의 모임. 모집단의 일부.
- 추론(inference): 표본에서 얻은 정보를 이용하여 모집단에 대한 정보를 예측 또는 추론

통계학이란?



- 관심 또는 연구의 대상인 모집단의 특성을 파악하기 위해 모집단으로부터 일부분에 대한 자료(표본)를 수집하고
- 이를 정리, 요약, 분석하여 표본의 특성을 파악한 후
- 표본의 특성을 이용하여 모집단의 특성에 대해 추론하는 원리와 방법을 제공하는 과학적인 학문

- 통계학이 적용되는 절차:

  데이터 수집 $\longrightarrow$ 통계적 분석 $\longrightarrow$ 의사결정, 미래예측

자료의 수집

- 자료 수집의 방법: 관측(observation)과 실험(experiment)

1. 관측에 의한 방법: 대상이 되는 개체들에게 아무런 조치도 취하지 않고 단지 관측을 통해 결과를 구하는 것

   예) 여론조사

2. 실험에 의한 방법: 반응을 관찰하기 위해 연구의 목적에 부합하는 조건에서 의도적으로 개체에 어떤 처리를 부가하여 어떤 변화를 일으키는가를 연구하는 것

   예) 연구실 실험

   → 둘다 통계적 분석을 위해 이용되는 표본(sample)이 된다.

자료의 수집 예시

- 유아기 영어 조기교육이 중학생 영어성취도에 미치는 영향을 조사

- 관측연구: 100명의 중학생을 뽑아 그들의 유아기 영어교육 여부와 현재의 영어성취도를 조사한다. 연구자가 이 중학생들에게 아무런 개입도 하지 않았으므로 이 연구는 관측자료에 근거한 것이다.
- 실험연구: 같은 목적으로 100명의 유아를 뽑아 그 중 50명에게는 영어 조기교육을 시키고 50명에게는 영어 조기교육을 시키지 않았다고 하자. 이들이 중학생이 되었을 때 영어 성취도를 조사하여 유아기 영어교육 여부와 연관시킨다면 이는 실험자료에 근거한 연구이다.

용어정리

- 개체(unit): 모집단의 개별 요소.
  예) 사람이나 가구, 학교, 사업체, 동물 등
  - 추출단위(sampling unit): 전체를 구성하는 각 개체.
  - 실험단위(experiment unit): 실험이 수행되는 개체
- 변수(variable): 개체의 특성을 나타내 주는 것.
  - 표본 내 개체들에 대하여 변수가 측정된다.
  예) 키, 몸무게, 성별, 흡연 유무 등

- 설명변수(explanatory variable): 실험이나 관측연구의 결과에 영향을 줄 수 있는 변수

  예) 영어 조기교육의 여부
- 반응변수(response variable): 설명변수의 변화에 따라 값이 결정되는 변수

  예) 중학생의 영어성취도
- 특성값(characteristic): 각 추출단위의 특성을 나타내는 값.

  예)지지하는 후보자의 이름, 키의 수치, 수확량의 값 등
- 모수(parameter): 모집단의 특성을 나타내는 특성치

예제

Example

전국의 남자 고등학생의 키에 대한 연구를 하고자 10,000명을 뽑아서 그들의 키를 측정한다고 하자. 키 외에도 나이나 몸무게, 허리둘레 등의 자료들도 같이 조사를 하였다.

- 모집단: 전국의 남자 고등학생 모두를 모아놓은 집단
- 개체: 집단 안의 학생 한 명 한 명
- 표본: 10,000명의 학생들
- 변수: 키, 나이, 몸무게, 허리둘레

  - 개체의 특성을 나타내며 개체별로 다른 값을 가지게 됨
- 모수: 전국의 남자 고등학생의 키의 평균
- 관측된 표본으로 구한 평균키로 모수에 대한 추론을 한다.

표본조사와 전수조사

- 전수조사(census): 모집단의 모든 구성원에 대하여 조사하여 얻은 자료를 바탕으로 모집단의 특성을 분석하는 방법

  ex) 인구주택총조사
  - 한국 내의 모든 가구들을 대상으로 나이, 혼인상태, 주택건평 등의 기초적인 인구 및 가구에 관한 사항을 조사
  - 매 5년마다 통계청에서 인구주택총조사를 실시
  - 전국의 인구 규모와 분포를 제시하여 각종 정책 입안을 위한 기초 자료로 사용.

- 표본조사(sample survey): 모집단의 일부인 표본을 추출하여 이 표본에서 얻은 정보를 기초로 모집단의 특성을 추론하는 방법. 관측 연구의 가장 대표적인 예.

  ex) 경제활동인구조사
  - 국민의 취업, 실업 등의 경제적 특성을 조사
  - 통계청에서 매월 1차례씩 실시

- 정확성, 비용, 시간의 비교

표본조사와 전수조사

- 표본조사의 장점:
  1. 비용절감
  2. 시간단축
  3. 조사의 적합성
     - 전수조사 불가능
  4. 오차관리 가능
     - 추정값의 오차를 계산함으로써 현재 추정값의 불확실성에 대한 정보 제공

- 표본조사의 단점:
  1. 표집오차 발생
  2. 희소항목에서는 표본조사 불가능
  3. 시간의 흐름에 따라 잘못된 추정가능
     - 모집단은 해마다 변동하기 때문에 시간이 지남에 따라 표본의 추정값은 실제를 반영 못할 수 있음

표본조사와 전수조사

- 표본조사는 실제와는 다른 예측을 하게 될 위험성을 항상 가지고 있음
- 표본에서 얻은 추정값은 반드시 실젯값과 차이를 갖게 됨

- 오차(error): 주어진 추정량과 모수와의 차이를 그 추정량의 오차라고 함
- 표집오차(sampling error): 모집단의 일부분인 표본에 의하여 모집단을 추론하기 때문에 일어나는 오차. 표본의 개수와 역의관계
- 비표집오차(nonsampling error): 모집단을 다 조사하여도 발생할 수 있는 오차. 예) 무응답, 거짓 응답, 자료입력오류 등

표본조사(sample survey)

- 표본조사

- 표본이 모집단의 속성을 잘 대표해야 함
- 표본의 대표성을 확보하기 위해서는 모집단의 모든 추출단위가 표본으로 추출될 기회를 같게 해 주어야 한다.
- 통계적 표본설계는 모집단 각 추출단위가 표본으로 추출될 확률을 미리 부여하는 확률적 추출법(probability sampling)을 근거로 설계

단순랜덤추출법(simple random sampling)

Definition (단순랜덤추출법(단순임의추출법))

모집단에서 n개의 추출단위로 구성된 모든 부분집합(단순랜덤표본)들이 표본으로 선택될 확률이 같도록 설계된 표본추출방법

- 개념적으로 간단하며 통계이론 전개의 기본틀로 사용
- 크기가 n인 모든 가능한 "표본"에 동등한 선출 기회를 부여
  - 모집단의 크기가 N일 때, 크기가 n인 표본을 뽑는다고 하면, 가능한 모든 표본의 수는 $\binom{N}{n}$이고, 특정 표본이 뽑힐 확률은 $1/\binom{N}{n}$이다.
- 모집단의 각 추출단위에 동등한 선출 기회를 부여 ("simple = equal probability")
- or 무작위 추출방법

- 실제 문제에서는 난수표(random number table)나 컴퓨터의 난수 발생 프로그램을 이용하여 표본을 추출
- 이와 같은 랜덤화(randomization)는 수리적으로 정의되지는 않지만 불확실성을 가장 잘 나타냄 (임의, 무작위, 랜덤은 모두 같은 의미)

Example (표본으로 선택될 확률 구하기)

어떤 모집단이 5개의 개체 {1, 2, 3, 4, 5}로 이루어져 있다고 하자. 모집단에서 한꺼번에 2개를 무작위로 뽑는다고 하면 다음과 같이 10가지 표본을 얻을 수 있다.

(1,2), (1,3), (1,4), (1,5), (2,3), (2,4), (2,5), (3,4), (3,5), (4,5)

- 각 개체가 선택될 가능성이 같다면 각 쌍이 표본으로 뽑힐 확률은 1/10
- 각 개체가 매 추출에서 무작위로 선택될 확률은 1/5
- 어떤 개체가 표본으로 뽑힐 확률은 2/5

다른 표본추출방법의 종류

- 층화추출(stratified sampling)

  : 모집단을 몇 개의 중복되지 않는 층으로 나누어 각 층에서 단순 랜덤표본을 뽑는 방법

- 군집추출(cluster sampling)

  : 개체 대신 군집(cluster)을 무작위로 선택한 후, 그 내에 있는 모든 개체들을 표본으로 추출하는 방법

- 계통추출(systematic sampling)

  : 모집단 목록에서 추출 간격 k를 정하여 매 k번째가 되는 단위를 표본으로 선정하는 추출법

표본조사시 유의사항

- 표본 추출시 랜덤표본을 사용하지 않으면 모집단의 일정 부분에 치우치는 편향(bias)이 발생

  예) TV 뉴스 등에서의 인터뷰, 인터넷 포탈사이트에서의 의견조사

  - 사용자 편의에 따라 표본을 추출하는 경우

- 표본조사 시 신뢰성 있는 결과를 얻기 위해서는 아래 항목들의 관리가 필수적

1. 통계적 표본설계
2. 완전한 모집단의 리스트 준비

- 리스트에서 누락된 추출단위가 모집단의 나머지 추출단위와 특성이 다르면 편향 또는 치우침 현상이 발생할 수 있음

표본조사시 유의사항

3. 정확한 설문지 작성
   - 설문지의 문항 구성 방식이나 사용된 어휘에 따라 응답에 큰 영향
   - 예시) "환경 보호 규제는 중소기업에 과도한 부담을 주어 많은 일자리를 잃게 하고, 경제 성장을 저해할 수 있습니다. 당신은 이 규제를 완화하는 것에 찬성하십니까, 반대하십니까?"
   - 특정 이해관계자가 환경 보호 규제에 반대하면서, 설문조사로 유리한 여론을 조성하고자 질문을 구체적인 규제로 한정하고 부정적인 효과를 강조하는 방식으로 응답을 유도하고 있음

표본조사시 유의사항

4. 철저한 조사자의 교육 및 감독
   - 불법적이거나 사회적으로 민감한 행동에 대한 질문은 해당 행동의 발생 비율이 실제보다 낮게 추정될 가능성이 높음
   - 조사자의 말투, 태도 혹은 성별 또한 응답자의 답변에 영향을 미칠 수 있음
   - 예시) "일반인들은 일주일에 3번은 운동합니다. 당신은 얼마나 자주 하십니까?" => 응답자는 실제보다 운동을 더 자주 한다고 답할 가능성이 큼
5. 무응답의 관리
   - 표본으로 선정된 사람이 조사에 응하지 않거나 협조하지 않을 때 발생

통계적 실험

- 실험: 사람이나 동물 또는 사물에 어떤조작을 가하면 어떻게 반응 또는 변하는가를 연구하는 것
- 실험의 목적: 실험환경이나 실험조건이 변할 때 어떤 반응의 변화가 있는지를 보는 것
- 실험환경을 통제하고 조정한다.

- 용어정리
  - 처리(treatment): 실험단위에 특정한 실험 환경 또는 실험 조건을 적용하는 것
  - 인자(factor): 통계적 실험에서 실험환경이나 실험조건을 나타내는 변수
  - 인자의 수준(level): 인자가 취하는 값

통계적 실험

Example (교과서상 문제의 위치)

백신 연구를 생각해보자. 특정 백신을 25명에게 동일한 용량으로 주사한다고 가정 할 때, 주사 후 30분이 경과했을 때 측정된 혈액 내 백신 농도가 반응변수가 된다. 이 실험에서는 주어진 백신의 일정량 주사가 한 가지 수준에서 한 가지 인자만을 포함한다.
만약 세 가지 다른 용량의 백신을 주사한다면, 이 실험은 세 가지 수준에서 한 가지 인자를 포함한다.

통계적 실험

Example (교과서상 문제의 위치)

고등학교 수학 교과서에서 문제의 위치와 질문의 종류가 교육에 어떤 영향을 미치는가를 연구하고자 한다. 교과 내용은 같으나 6가지의 서로 다른 시범 교과서를 만들어 시험 성적을 비교한다.

	질문의 종류(유형)		
문제의 위치	간단한 사실	계산	개념
후반부	1	2	3
전반부	4	5	6

- 실험단위: 학생들
- 인자: 위치와 유형 2개
- 반응변수: 시험점수
- 수준: 위치는 2가지 수준, 문제의 유형은 3가지 수준
- 처리: 6가지의 처리를 실험

통계적 실험

Example (교과서상 문제의 위치)

연구자들은 새로운 당뇨병 치료제가 혈당 조절에 미치는 영향을 알아보기 위해 실험을 설계했다. 실험에 참여한 당뇨병 환자 18,000명 중 절반은 새로운 치료제를, 나머지 절반은 기존 치료제를 매일 복용했다. 6개월 후 두 집단의 혈당 수치를 측정한 결과, 새로운 치료 제를 복용한 집단의 평균 혈당 수치가 기존 치료제를 복용한 집단보다 낮았다. 이 실험에서 실험단위, 인자, 인자의 수준, 반응변수를 찾아라.

- 실험단위
- 인자
- 인자의 수준
- 반응변수

실험계획의 원리

- 실험계획의 원리: 분석의 목적에 맞는 올바른 자료를 구하기 위한 원리로 다음의 것들을 고려할 수 있다.

1. 비교실험: 처리집단(treatment group)과 대조집단(control group)을 비교하여 실험의 결과가 처리의 효과에 의한 것이라는 것을 보일 수 있다.

   - 예) 진통제 효과 비교

   - 위약 효과, 플라시보 효과(placebo effect): 의사에 대한 신뢰와 치료에 대한 기대감에서 비롯, 환자는 어떤 처방이라도 긍정적으로 반응할 수 있다.

   - 통제집단을 통해 외부인자의 영향을 막아 관심 인자의 실제 효과를 측정할 수 있다.

   - 관심 인자 이외의 다른 외부 인자의 효과를 극소화 한다.

실험계획의 원리

2. 랜덤화(randomization): 우연에 의한 실험단위 배정은 각 실험단위가 처리집단이나 대조집단에 배정될 기회를 동등하게 부여

   - 이는 두 집단을 동질적으로 만들어, 있을 수 있는 잠재변수의 영향을 줄이기 위한 것이다.

   - 우연적(random) vs 우발적(haphazard)

   - 우연으로 설명하기에는 차이가 너무 큰 경우, 그 차이는 처리 효과에 의한 것이라고 결론

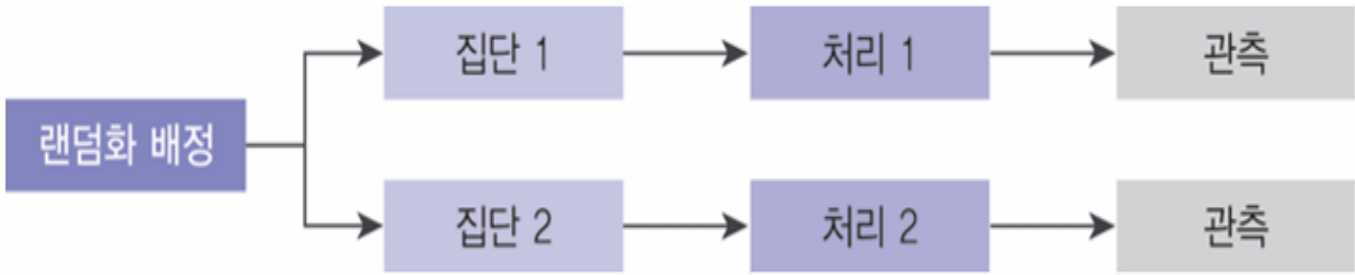


그림 1-2 랜덤화 비교실험

실험계획의 원리

3. 블록화: 실험 전에 동일한 처리에 대해 유사한 반응을 보일 것으로 예상되는 실험단위들을 집단으로 형성하는 것을 의미한다. 이때 각 집단을 블록block이라 한다.

Example

어떤 암은 성별에 따라 진전 속도가 다르다고 한다. 이 암에 대한 세 가지 처방을 비교하는 임상실험을 진행한다고 하자. 성별 차이가 실험 결과에 영향을 미칠 수 있으므로 남성 환자 30명과 여성 환자 30명을 각각 블록으로 설정하고, 각 블록에서 랜덤화를 통해 10명씩 세 집단으로 나눈 뒤 세 가지 처방을 각 집단에 배정한다.

실험계획의 원리

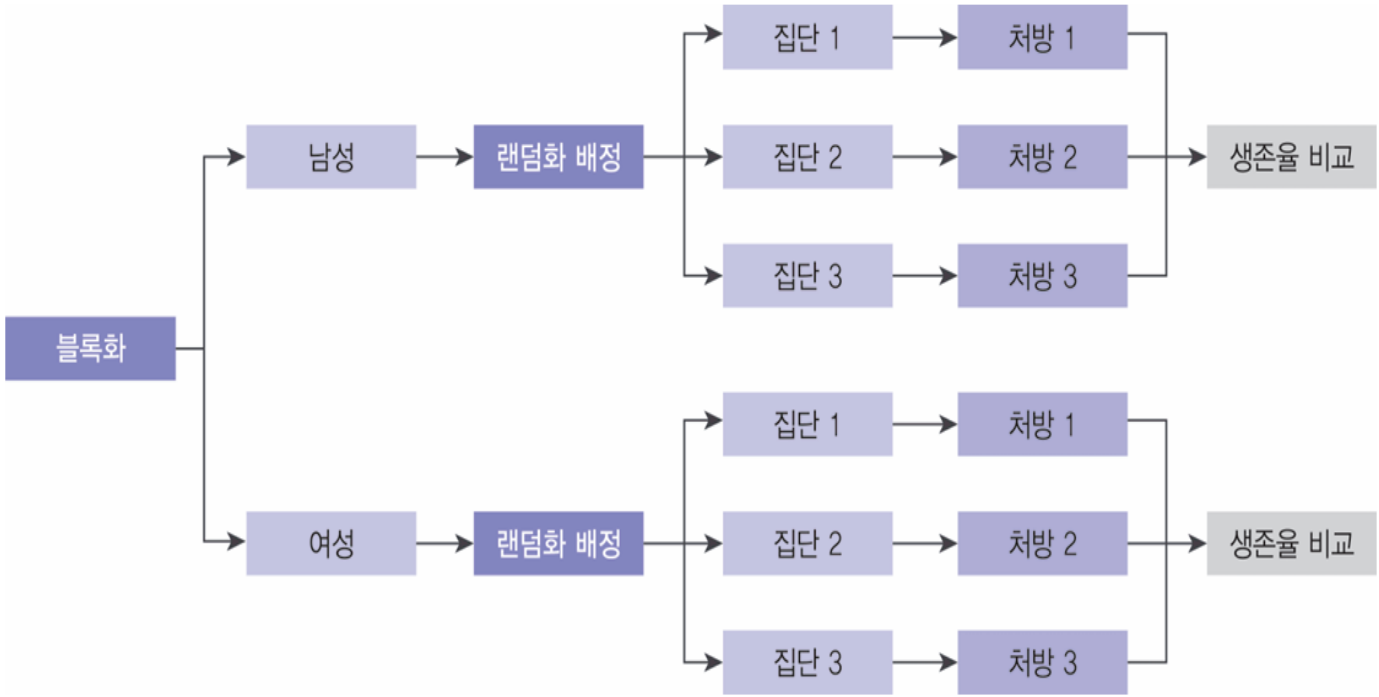


그림 1-4 [예 1-10]의 랜덤화블록계획

4. 반복: 실험의 결과가 우연히 나온 것이 아니라는 것을 반복을 통해 보일 수 있다. 반복해서 동일하게 나오는 결과는 실험에 대한 신뢰성을 높인다.

표본조사와 실험계획의 차이점

- 실험계획
  - 통계적 실험계획에 의해 자료를 생성하는 경우에는 각 실험단위가 어떤 처리를 받을지 실험자가 결정.

    예) 백신 연구를 세 가지의 다른 종류의 피임약을 세 그룹에 주사. 30분 후 피 속의 약의 농도를 측정.

    $\rightarrow$ 이때, 관심 인자 이외의 다른 외부 인자의 효과는 극소화하여 실험을 설계

  - 설명변수와 반응변수 사이의 관계를 알아보고자 한다면 실험을 통해 수집된 자료를 분석하는 것이 더 타당하다.
  - 윤리적으로 실험계획을 계획할 수 없는 조사도 있다.

- 표본조사

- 단순히 표본으로 뽑힌 각 추출단위가 어떤 처리를 받았는지 관측만 하게 됨

  예) 흡연이 폐암의 발생에 미치는 영향을 조사하기 위해 흡연 그룹과 비흡연 그룹의 폐의 건강상태를 조사.

  $\rightarrow$ 외부인자의 효과를 통제하기 어려움

- 관측연구를 통해서는 연구에 포함되지 않은 다른 변수들의 영향을 제거하기가 매우 어렵다.
- 통계적 실험계획이 불가능한 경우에는 표본조사에 의한 관측연구(observation study)가 불가피하며, 조사결과의 해석시 외부 인자에 대한 격별한 주의가 필요.