1장. 자료의 생성

박명현

서울대학교 학부대학

Overview

1. 통계학의 어원
2. 통계의 사용
3. 통계학이란?
4. 자료의 수집
5. 표본조사
6. 통계적 실험

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' 뽑야야 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

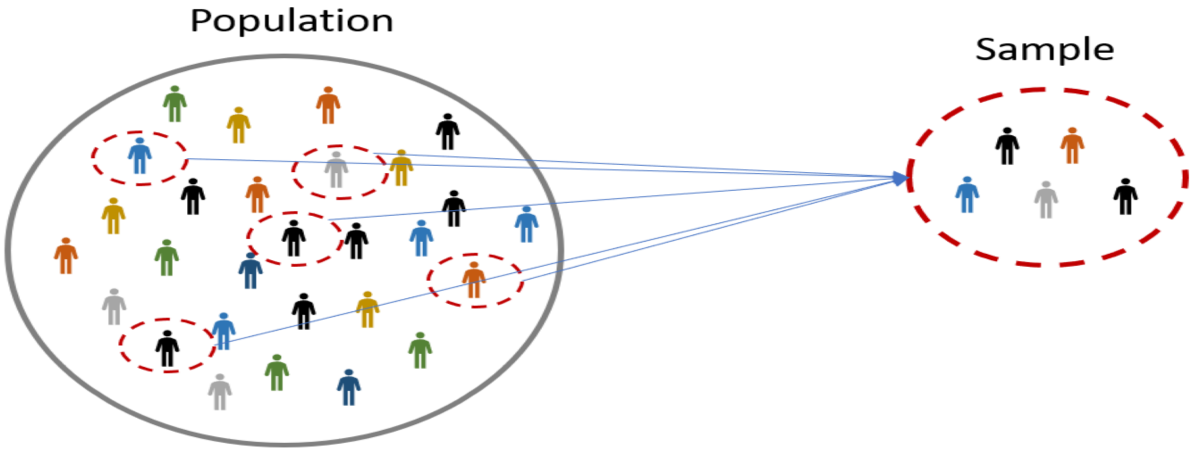
3. 조사한 내용으로부터 전체 유권자의 성향에 대하여 어떻게 결론을 내릴 것인가?
   - 표본에서 구한 특정 후보의 지지율이 모집단에서의 지지율을 얼마나 잘 반영할까?

     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

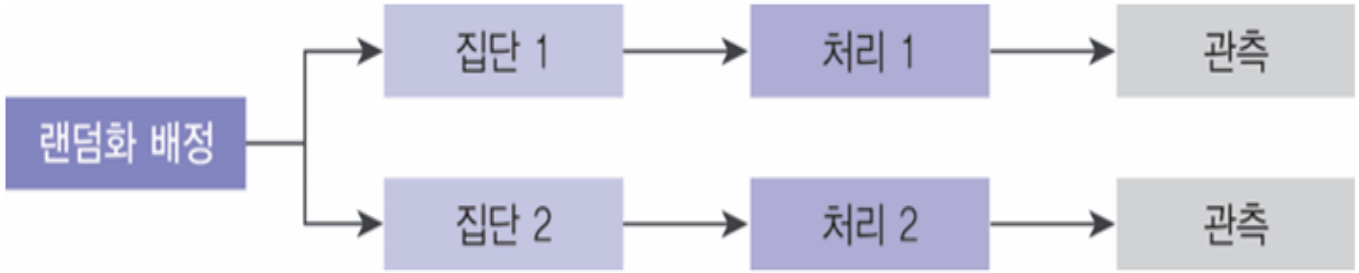


그림 1-2 랜덤화 비교실험

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

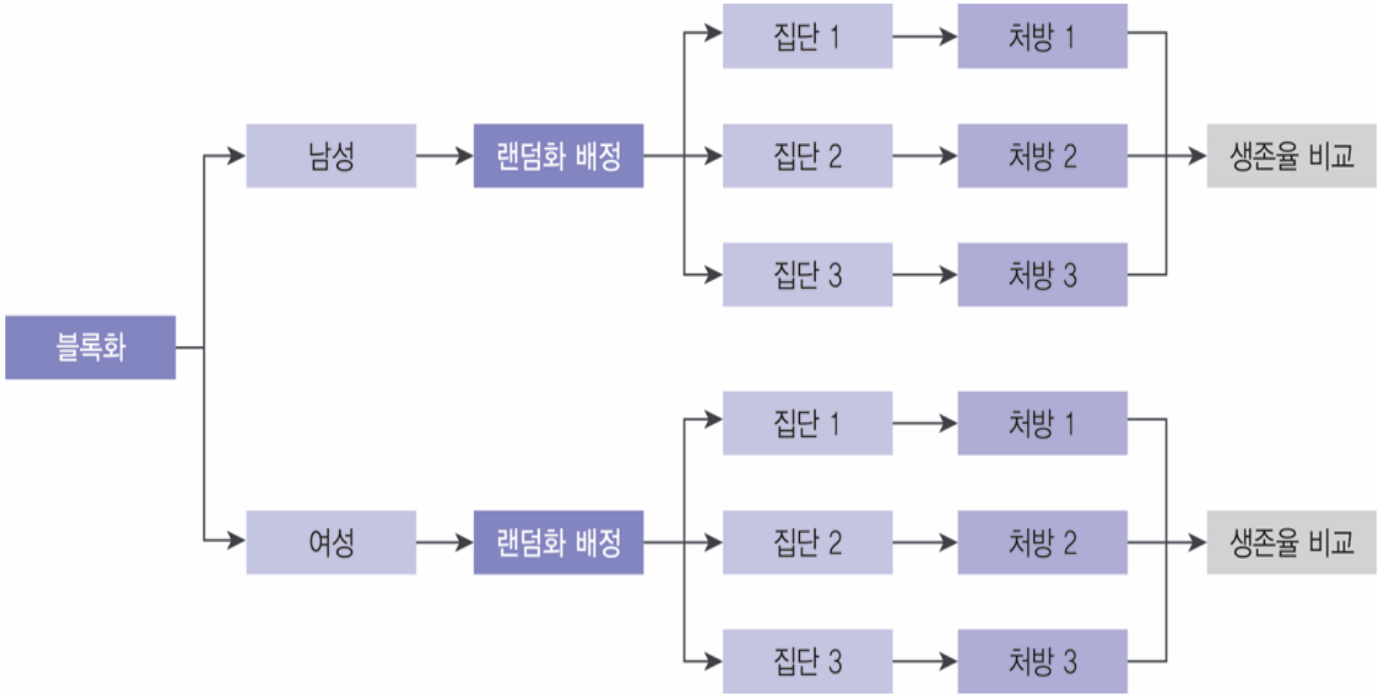


그림 1-4 [예 1-10]의 랜덤화블록계획

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.

    → 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.

2장. 모집단과 표본

박명현

서울대학교 학부대학

기술통계(descriptive statistics)

- 기술통계: 자료를 정리, 요약하여 기술하는 것
  - 자료의 대략적인 모습을 보여주고(summary), 섬세한 분석을 위한 기초단계로 활용된다.

1. 그래프나 표를 이용한 기술통계
   - 표본을 도표화함으로써 모집단 분포의 개형을 파악
   - 도수분포표, 분할표, 히스토그램, 산점도 등
2. 수치에 의한 기술통계
   - 중심위치 척도: 평균, 중앙값 사분위수 등
   - 변동성 척도: 분산, 범위 등
   - 연관성 척도: 공분산, 상관계수

자료의 구분(사칙연산 가능 여부에 따른)

- 질적자료(qualitative)
  - 사칙연산 불가능
  - 관측결과가 몇개의 범주 또는 항목의 형태로 나타남
  - 예) 직업, 종교, 성별 등
  - 범주형 자료(categorical data)라 부르기도 함
- 양적자료(quantitative)
  - 사칙연산 가능
  - 관측된 값이 수치로 측정됨
  - 합계, 평균, 분산 등으로 자료를 요약 정리할 수 있으며 비교도 가능
  - 예) 몸무게, 용돈, 가격 등
  - 숫자형 자료(numerical data)라 부르기도 함

자료의 구분(측정척도에 따른)

- 질적자료
  - 명목형 자료(nominal data)
    - 이름이나 문자로 나타내어지는 자료.
    - 범주간 순서가 없음
      예: 혈액형(A,B,O,AB), 성별(남, 여)
      예: 직업종류(전문직, 회사원, 공무원)
  - 순서형 자료(ordinal data)
    - 이름이나 문자로 나타내어지는 자료이나 범주들 간에 순서가 있는 자료
    - 사칙연산을 할 수 없음
      예: 선호도(아주좋아함, 좋아함, 보통, 싫어함, 아주싫어함)
      예: 학점(A, B, C, D, F), 옷사이즈(S,M,L,XL)

자료의 구분(측정척도에 따른)

- 양적자료
  - 연속형 자료
    - 연속적인 값을 가지는 자료들
      예: 키, 몸무게
    - 많은 경우에 정확한 수치보다는 반올림 등을 통해 나타냄
      예: 키170cm 는 169.5에서 170.5사이의 값을 의미함
  - 이산형 자료
    - 관측가능한 값이 셀수 있음
      예: 주사위 결과, 교통사고 건수

자료 종류의 예

자료	종류
전문직업분류(변호사, 의사, 회계사등)	질적자료, 범주형자료
몸무게	양적자료, 숫자형자료
학력(초, 중, 고, 대)	질적자료, 순서형자료, 범주형자료
학력(교육연한)	양적자료, 숫자형자료

- 자료의 계층에 따라 선택할 수 있는 통계분석 도구가 달라지게 됨
  - 나이는 연수로 표시되면 숫자형자료. 평균과 분산을 구할 수 있고, 히스토그램 등의 그래프로 표시가능.
  - 나이를 10년 단위로 그룹화 하여 도수분포표와 바차트 등으로 표시할 수 있음.
  - 나이를 소년, 청년, 중년, 장년 등으로 순서형 자료로 표시하였다면 도수분포표와 바차트는 가능. 그러나 평균이나 분산을 구할 수 없음.

그래프나 표를 이용한 기술통계
(graphical and tabular description)

도수분포표(frequency table)

- 표본자료를 하나의 표로 요약한 것.
- 서로 다른 특성값에 대한 자료의 개수, 즉 도수(frequency)나 상대도수(relative frequency)를 구하여 특성값과 함께 나열한 것
- 상대도수=도수/전체자료의 갯수

표 1–1 어느 시의 시장 선거 여론조사 결과

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

- 수치형 자료도 전체 범위를 몇 개의 계급(구간)으로 나누어 도수분포표 작성 가능

도수분포표 예제

Example

예제 2.6) 다음은 어떤 집단에서 40명을 뽑아 키를 측정한 결과이다. 이 표본자료를 이용하여 도수분포표를 작성하여라(단위 : cm)

183	168	165	158	180	176	168	160	160	150
180	183	170	170	185	170	165	170	168	154
152	173	173	165	163	163	160	178	138	171
163	168	165	165	178	170	168	183	145	174

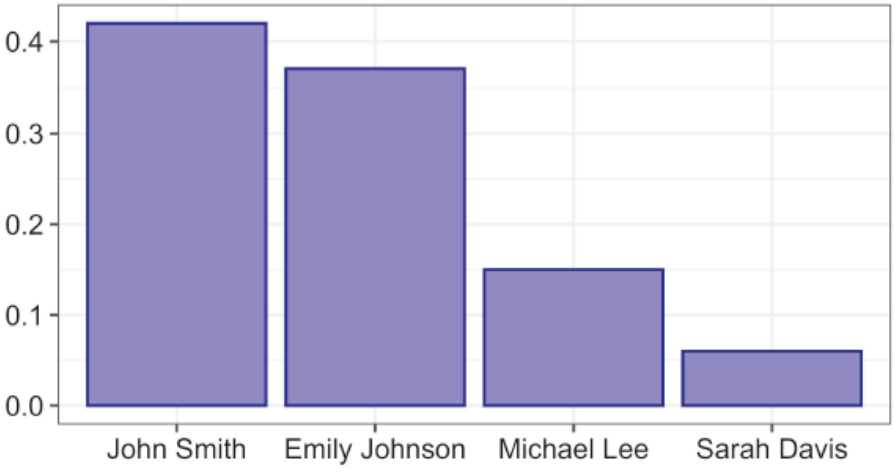
sol) 먼저 최대값=185, 최소값=138이므로 범위=47이다.
계급의 개수를 6으로 하면 $(47/6 = 7.83)$이므로 폭은 8로 한다.
시작점= 138-1/2× 1=137.5

표 2-6 [예제 2-6]의 도수분포표

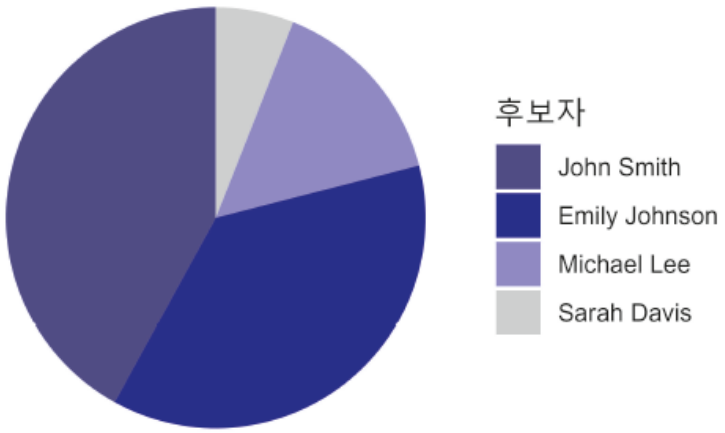
계급(cm)	도수	상대도수
137.5 이상~145.5 미만	2	0.050
145.5 이상~153.5 미만	2	0.050
153.5 이상~161.5 미만	5	0.125
161.5 이상~169.5 미만	13	0.325
169.5 이상~177.5 미만	10	0.250
177.5 이상~185.5 미만	8	0.200
합계	40	1

막대그래프와 원형그래프

- 막대그래프(bar graph): 각 특성값의 막대의 높이가 상대도수에 비례하게 그린 것
- 원형그래프(pie graph): 원을 부채꼴 모양으로 나누는데, 각 부채꼴의 넓이(또는 중심각의 크기)가 상대도수에 비례하게 나눈 것



(a) 막대그래프



(b) 원형그래프

그림 2–11 [예제 2–5]의 막대그래프와 원형그래프

히스토그램(histogram)

- 도수분포표를 이용하여 표본자료의 분포를 그래프로 나타낸 것

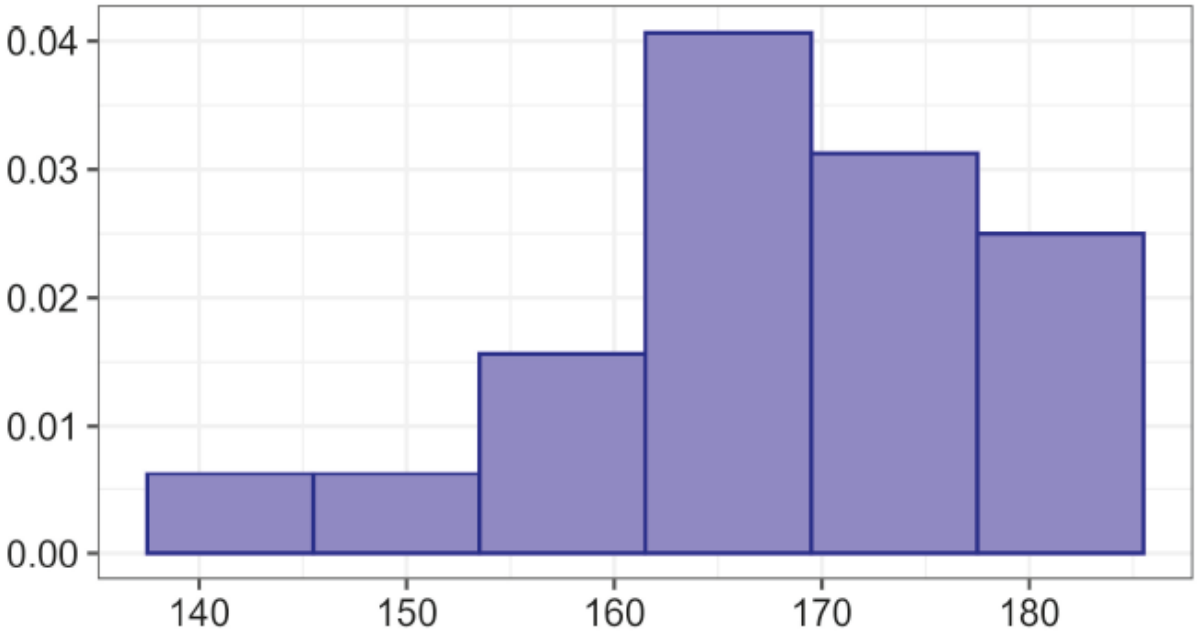


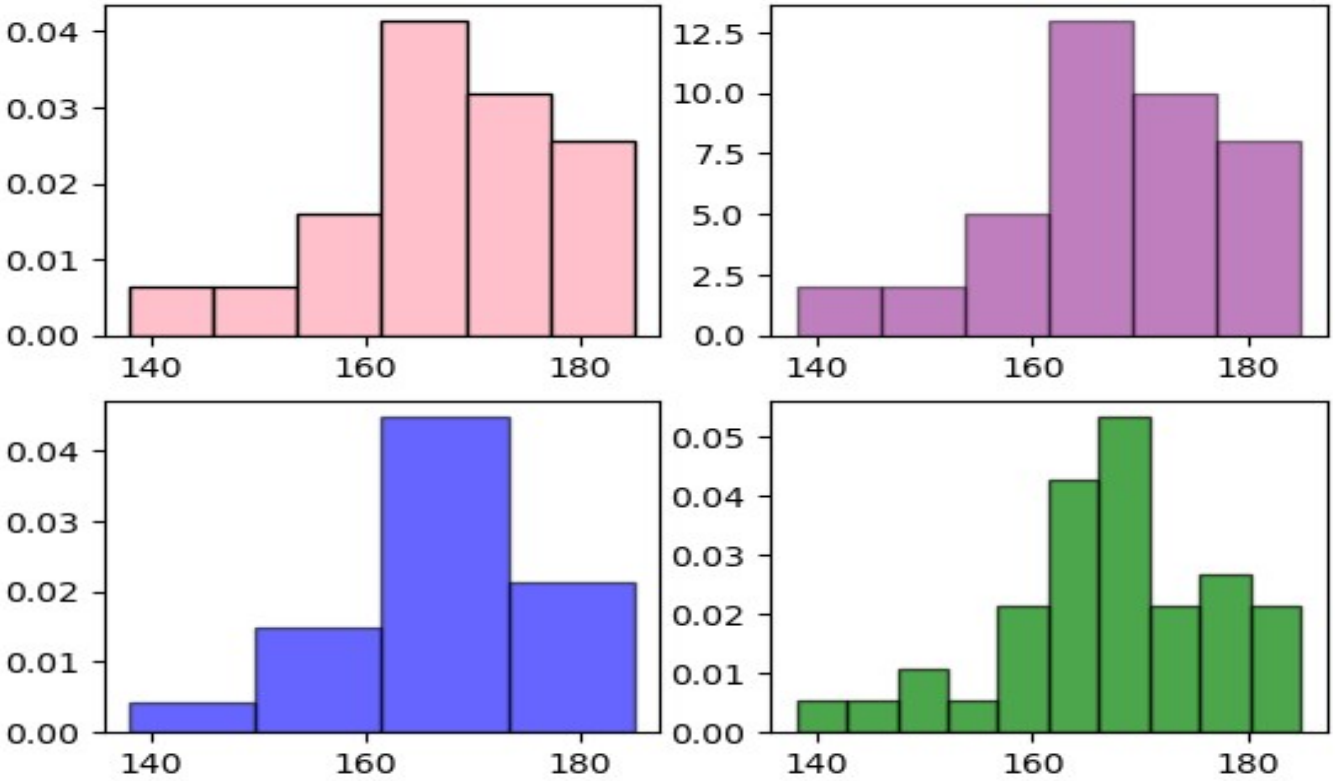
그림 2-12 [예제 2-6]의 히스토그램 : 모든 계급의 폭이 일정한 경우

히스토그램(histogram)

- 수평축위에 계급구간을 표시하고 그 위로 각 계급의 상대도수에 비례하는 넓이의 직사각형을 그린다.
- 직사각형의 넓이가 상대도수에 비례
- 직사각형의 높이=상대도수/계급의 폭
- 각 계급의 폭은 일정
- 전체 직사각형 넓이의 합은 1이 되어야 함
- 폭의 너비에 따라 분포의 형태가 달라짐

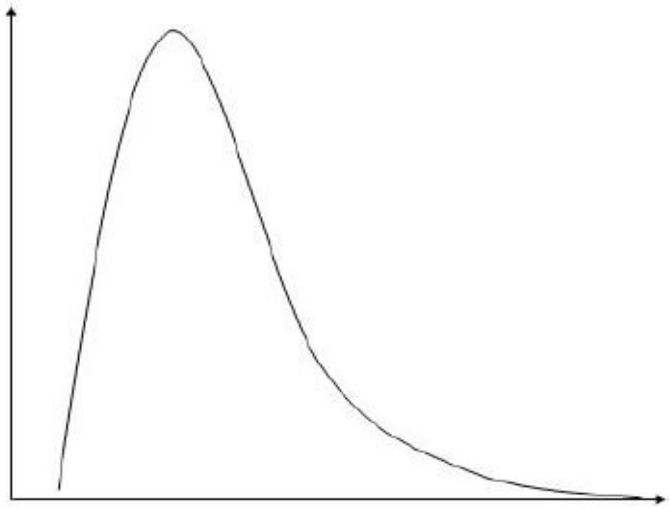
히스토그램(histogram)

- 높이를 도수, 계급의 폭을 4개, 10개로 바꾸어 그린 히스토그램
  - hist(data,bins=6,color ='purple',edgecolor='black',alpha = 0.5,density=True)

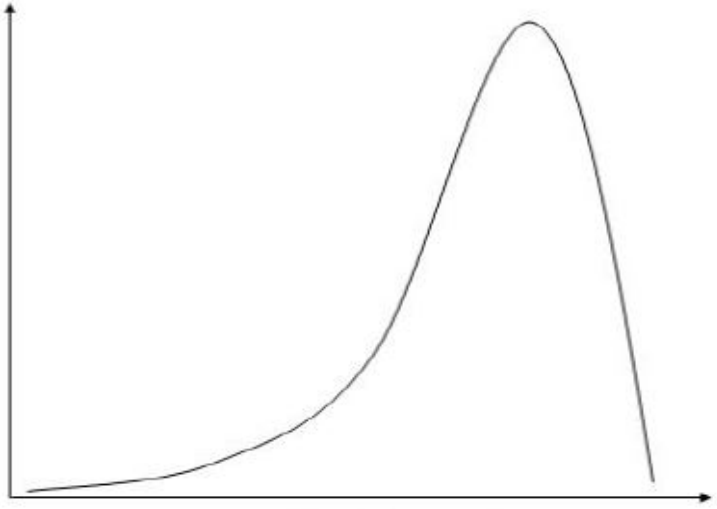


히스토그램(histogram)

- 자료에 대한 정보를 특징적 형태로 제공함
- 대칭(Symmetry)
- 왜도(Skewness): 좌우로 늘어진 정도
  - 양의 왜도(Right skewed) : 오른쪽으로 길게 늘어짐
  - 음의 왜도(Left skewed) : 왼쪽으로 길게 늘어짐



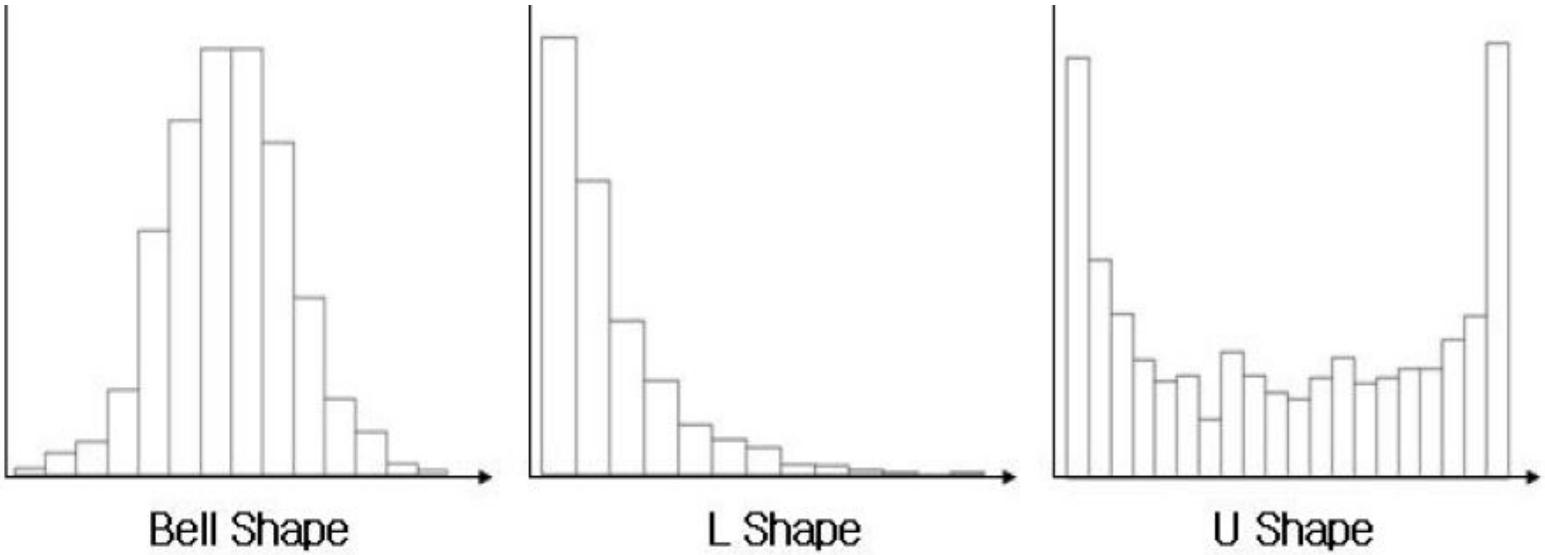
Right Skewed



Left Skewed

히스토그램(histogram)

- 모양(Shape): 종 모양(Bell Shape), L자 모양, U자 모양



- 봉우리 개수: 단봉(Unimodal), 2봉(Bimodal)
  - 상이한 집단의 자료들이 섞여 있을때
  - 남녀 구별하지 않은 몸무게 자료
  - 서울 강남과 강원도 삼척의 Apartment가격 자료

줄기-잎 그림(stem-and-leaf display)

- 히스토그램을 옆으로 돌려놓은 것과 같은 모양
- 세로 열을 줄기(stem), 오른쪽으로 나열된 가로 행을 잎(leaf)
- 자료의 값을 그대로 유지하고 있다는 장점
- 방대한 자료의 경우에는 그리기 어려움

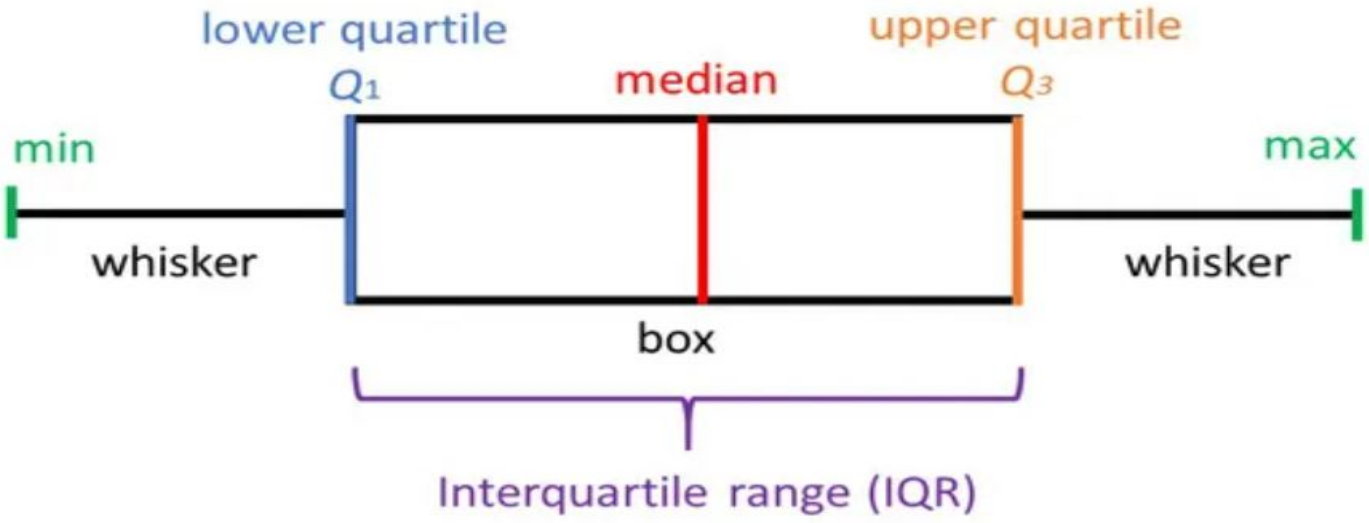
1단계	2단계	
18	18	3, 0, 0, 3, 5, 3
17	17	6, 0, 0, 0, 0, 3, 3, 8, 1, 8, 0, 4
16	16	8, 5, 8, 0, 0, 5, 8, 5, 3, 3, 0, 3, 8, 5, 5, 8
15	15	8, 0, 4, 2
14	14	5
13	13	8

3단계	
18	0, 0, 3, 3, 3, 5
17	0, 0, 0, 0, 0, 1, 3, 3, 4, 6, 8, 8
16	0, 0, 0, 3, 3, 3, 5, 5, 5, 5, 5, 8, 8, 8, 8, 8
15	0, 2, 4, 8
14	5
13	8

그림 2-15 [예제 2-6]의 줄기-잎 그림을 그리는 과정

상자그림(boxplot)

- 다섯숫자요약(five number summary)을 그래프로 표현한 것
  - 최솟값, Q_1(first quartile), 중위수, Q_3(third quartile), 최댓값
- 상자(box): Q_1, 중위수, Q_3를 사용하여 만든다.
- 상자의 길이: 사분위범위(IQR)
- 상자로부터 최솟값, 최댓값까지 선을 그어 수염(whiskers)을 만든다.

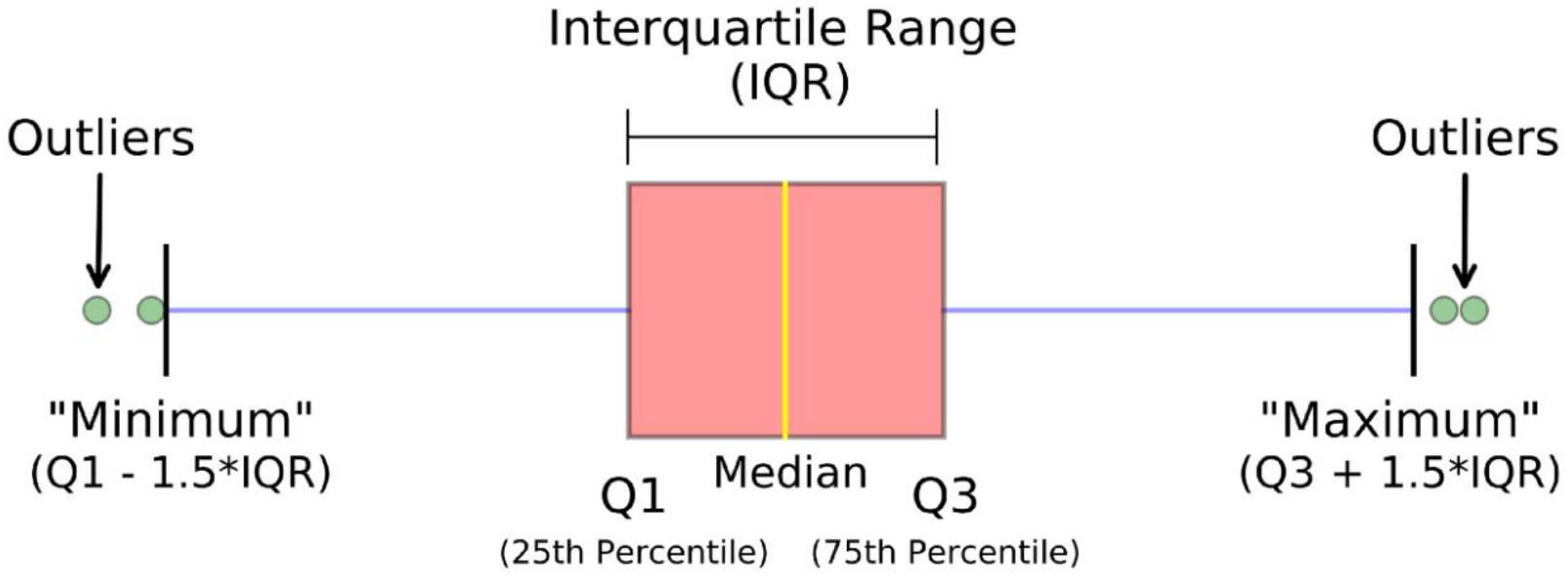


- 이상치(outlier)가 있으면 수염이 과도하게 길어진다.

상자그림(boxplot)_투키(John W. Tukey)

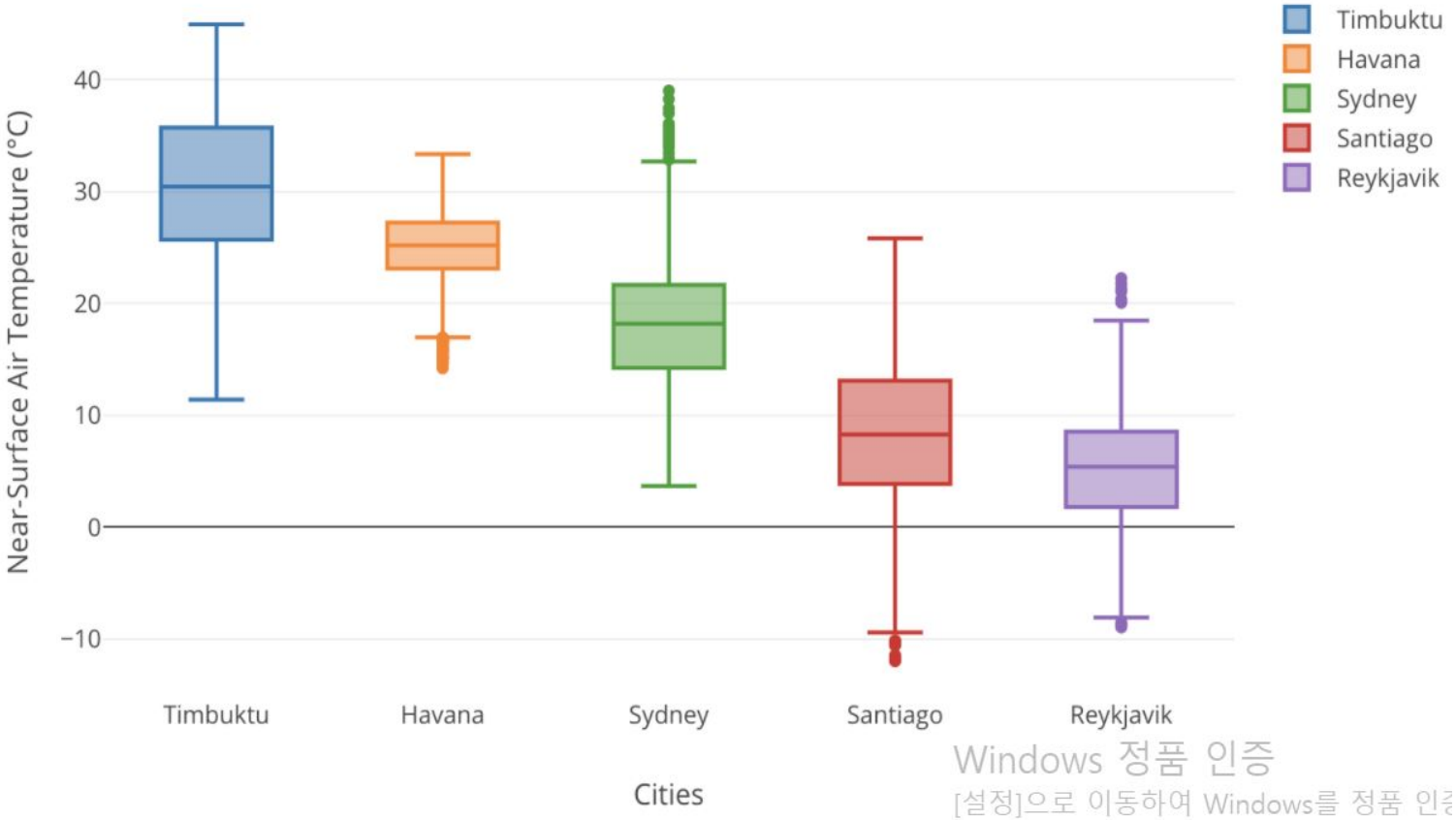
- 투키가 제안한 방법:

- 상자(box)를 만든다.
- 박스의 각 모서리 (Q1, Q3)로 부터 IQR의 1.5배 내에 있는 가장 멀리 떨어진 데이터 점까지 수염을 그린다.
- 수염보다 바깥쪽에 데이터가 존재한다면 이상치(outlier)로 간주 (*등의 기호로 표시)



상자그림(boxplot)

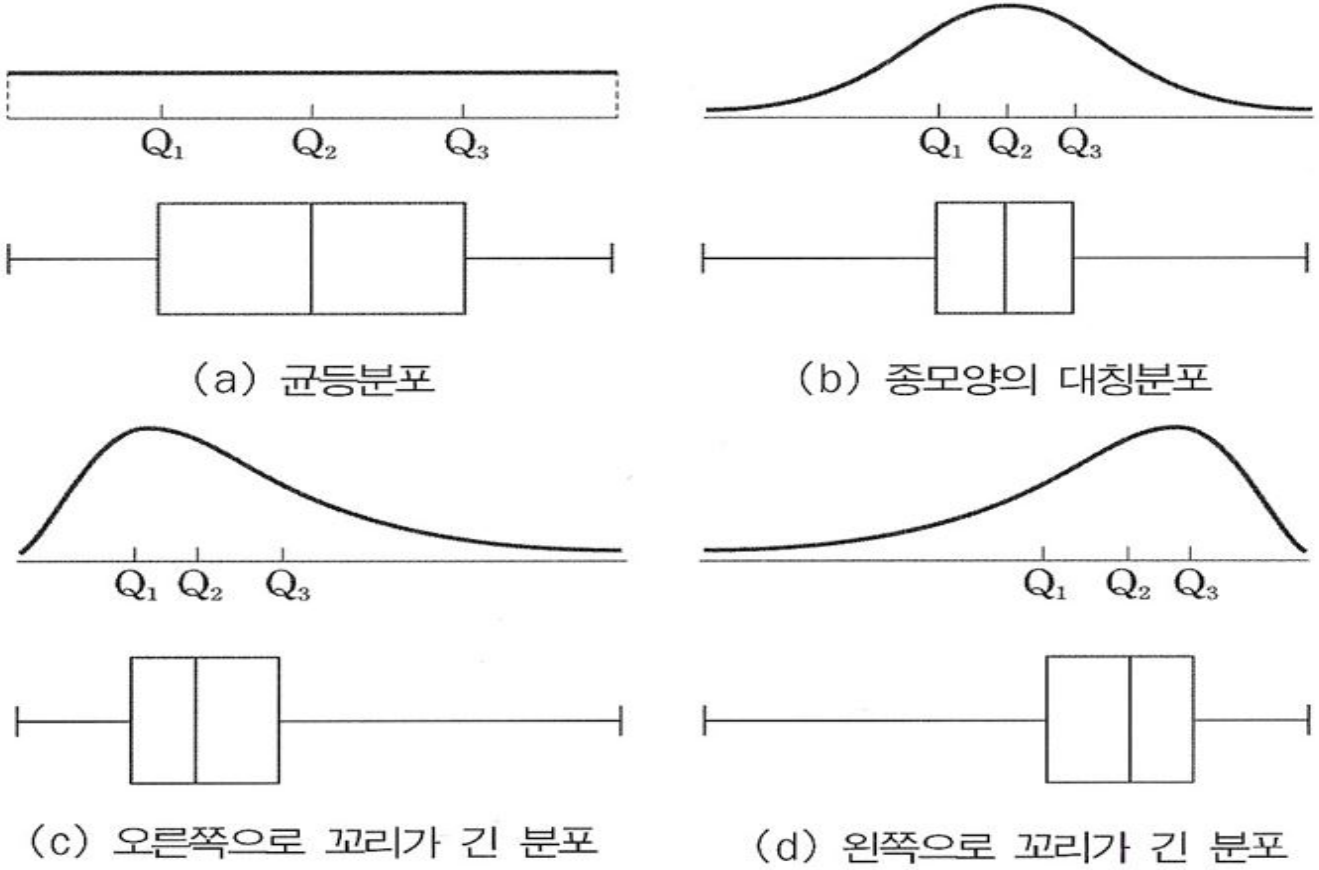
- 자료를 개괄적으로 알아보기 위함
- 여러 개의 분포를 한눈에 비교할 때 유용



Windows 정품 인증
[설정]으로 이동하여 Windows를 정품 인증

상자그림(boxplot)

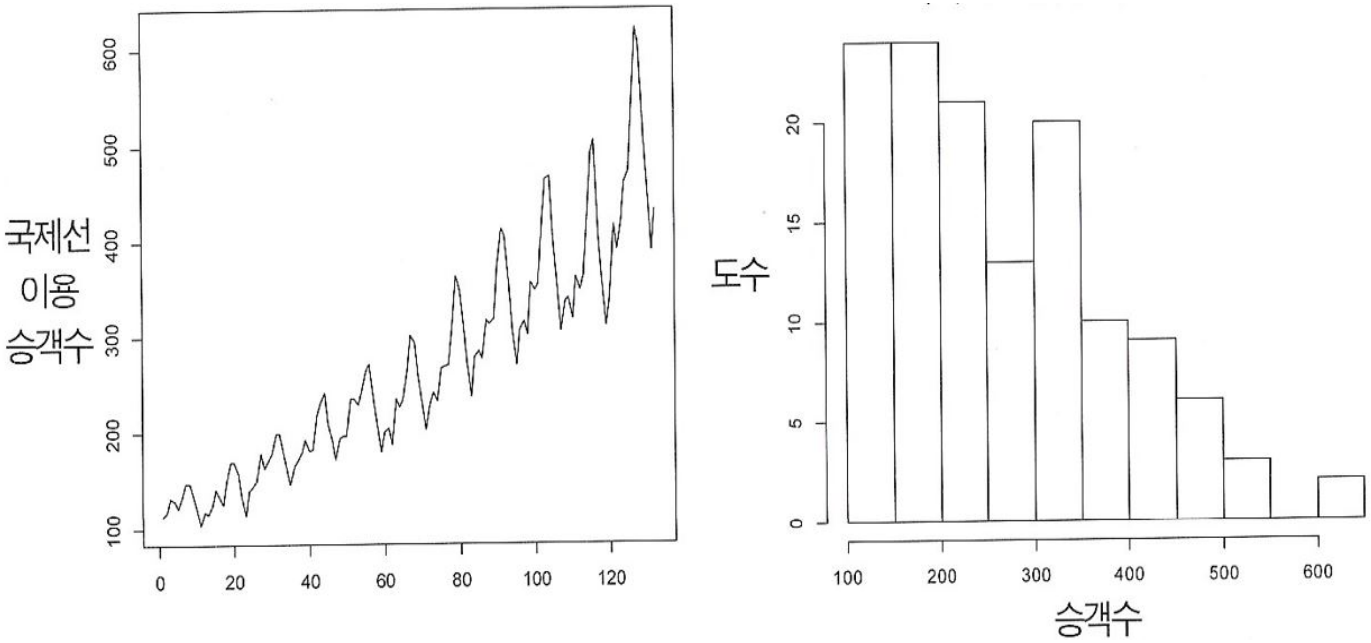
- 분포에 따른 상자그림의 형태를 나타낸 것



시계열그림(time series plot)

- 시간의 변화에 따라 얻게 되는 시계열 자료를 그래프로 표현
- X축: 시간, Y축: 해당자료
- 추세(trend)와 계절변동(seasonal variation) 등을 찾을 수 있음

Figure: 12년간의 월별 국제선 항공 승객수의 시계열그림과 히스토그램



이차원 자료의 도표화

1. 분할표(contingency table)

- 이차원 표본자료를 하나의 표로 요약한 것
- 각 특성의 값과 함께 각 이차원 특성값의 상대도수(또는 도수)를 이차원 표에 나열 (일차원인 경우의 도수분포표표)

예) 다음의 표는 400명의 학생에 대하여 국어, 영어, 수학 세 과목의 선호도를 조사한 결과를 분할표로 나타낸 것이다.

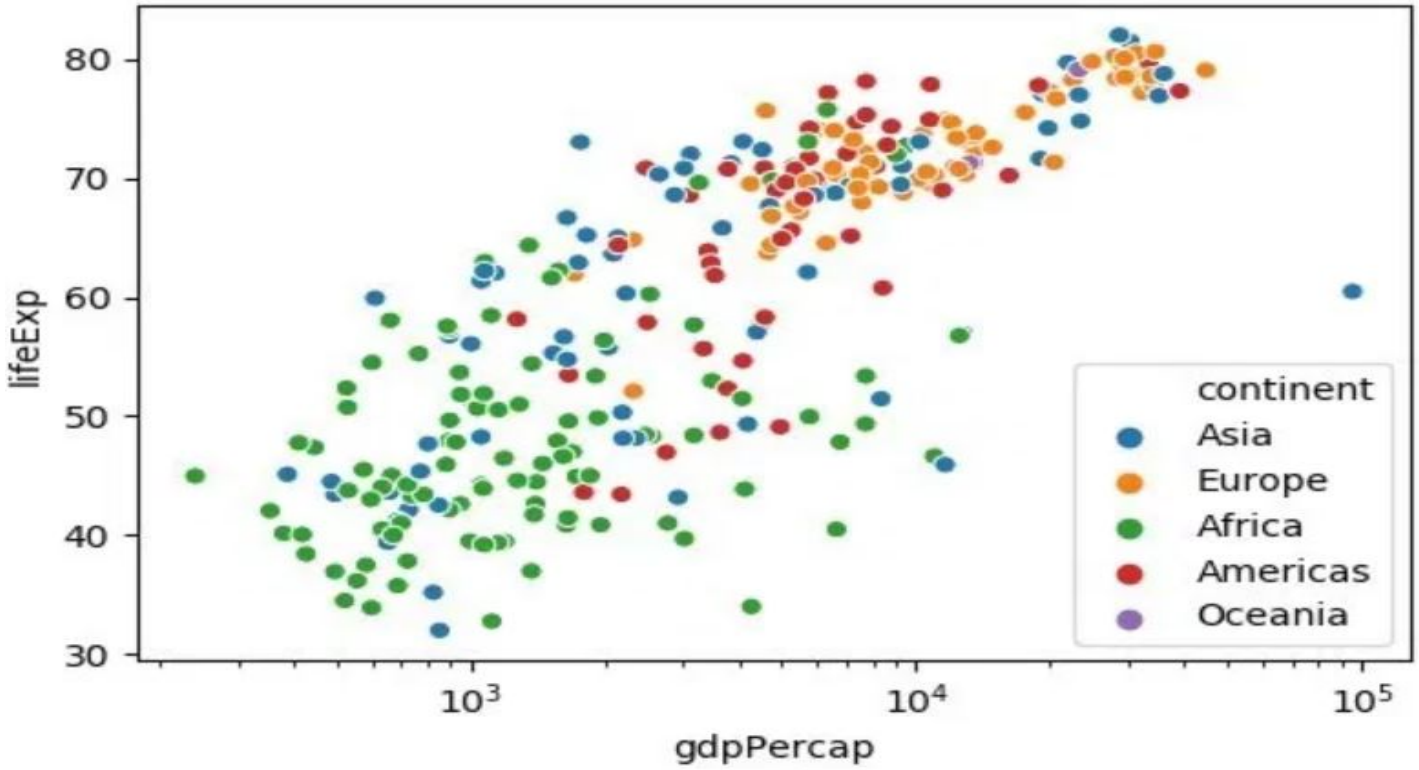
표 2-7 이차원 표본자료의 분할표

	국어	영어	수학	합계
남자	0.18	0.24	0.18	0.60
여자	0.14	0.16	0.10	0.40
합계	0.32	0.40	0.28	1.00

- 수치형 자료도 범위를 몇 개의 계급으로 나누어 분할표 작성 가능

이차원 자료의 도표화

2. 산점도(scatter plot): 각 이차원 자료값의 (x축:특성1의 값, y축:특성2의 값)의 점을 좌표평면 위에 찍은 것.
   - 두 변수의 연관성과 변화 관계를 쉽게 알아볼 수 있음



이차원 자료

Example

예제 2.7) 다음의 첫번째 값은 키(cm)이고, 두번째 값은 몸무게(kg)이다. 이 표본자료를 이용하여 분할표와 이차원 히스토그램, 산점도를 작성하여라.

(183,82)	(168,52)	(165,52)	(158,55)	(180,93)
(176,57)	(168,60)	(160,48)	(160,50)	(150,49)
(180,79)	(183,73)	(170,60)	(170,66)	(185,74)
(170,82)	(165,57)	(170,54)	(168,57)	(154,55)
(152,50)	(173,63)	(173,60)	(165,49)	(163,63)
(163,57)	(160,57)	(178,77)	(138,54)	(171,69)
(163,50)	(168,57)	(165,54)	(165,59)	(178,79)
(170,54)	(168,79)	(183,70)	(145,37)	(174,65)

sol) 키의 계급의 갯수는 6개이고(예제 2.6), 몸무게도 같은 방법으로 구해보면

- 최대값=93, 최소값=37, 범위=56이고, 계급의 개수를 6으로 하면 $56/6 = 9.33$이므로 폭은 10이다.

Figure: 예제 2.7의 분할표

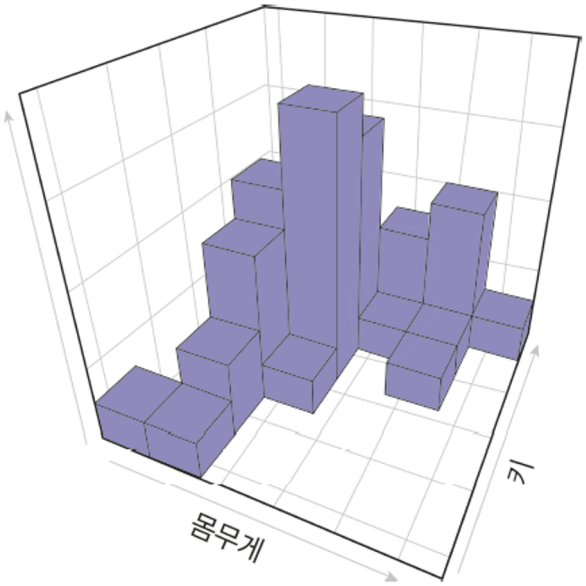
표 2-8 [예제 2-7]의 분할표

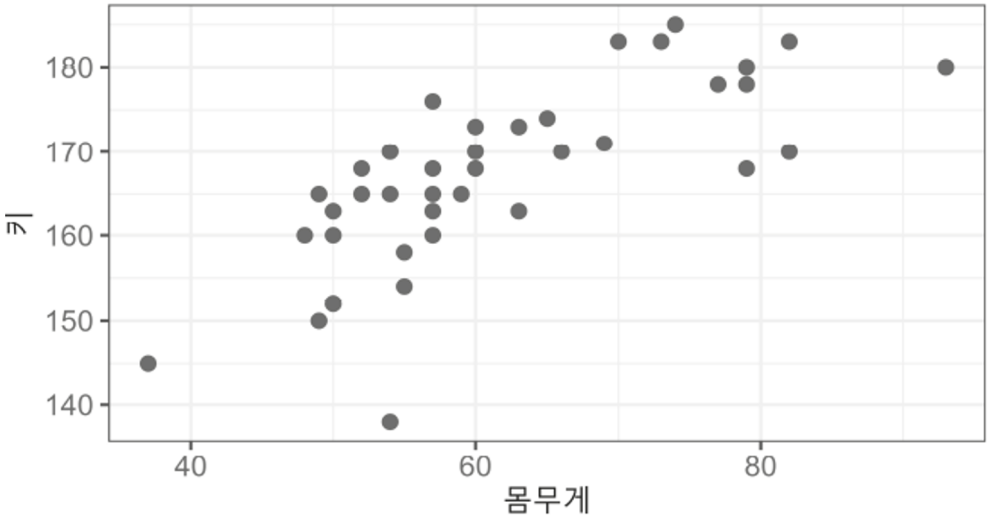
키 \ 몸무게	36.5 이상~46.5 미만	46.5 이상~56.5 미만	56.5 이상~66.5 미만	66.5 이상~76.5 미만	76.5 이상~86.5 미만	86.5 이상~96.5 미만	합계
137.5 이상~145.5 미만	0.025	0.025	0.000	0.000	0.000	0.000	0.050
145.5 이상~153.5 미만	0.000	0.050	0.000	0.000	0.000	0.000	0.050
153.5 이상~161.5 미만	0.000	0.100	0.025	0.000	0.000	0.000	0.125
161.5 이상~169.5 미만	0.000	0.125	0.175	0.000	0.025	0.000	0.325
169.5 이상~177.5 미만	0.000	0.050	0.150	0.025	0.025	0.000	0.025
177.5 이상~185.5 미만	0.000	0.000	0.000	0.075	0.100	0.025	0.200
합계	0.025	0.350	0.350	0.100	0.150	0.025	1.000

이차원 자료의 도표화

3. 히스토그램이차원 계급을 평면 위에 표시하고 그 위에 각 이차원 계급의 상대도수를 부피로 갖는 직육면체를 그린 것.

- 예제 2.7의 이차원 히스토그램과 산점도





수치에 의한 기술통계
(numerical description)

용어정리

<u>모집단</u>

- 모수(parameter): 모집단의 특징을 나타내는 대표값
- 모집단의 대표값: 모평균, 모분산, 모표준편차 등

  예) 2021년도 대학교 신입생들의 한달 용돈에 관심. 모평균$= \mu$

<u>표본</u>

- 표본의 대표값: 표본평균, 표본분산, 표본표준편차 등

  예) 전체 대학교 신입생 중 100명을 랜덤하게 추출하여 한달 용돈의 평균값을 조사. 표본평균$= \bar{x}$

- 이러한 표본의 대표값들은 모수를 추론하는 데 사용됨. $(\bar{x} = \hat{\mu})$

중심위치 척도

- 중심위치 척도: 자료들이 대략 어떠한 값을 갖는지를 알아보는 것 어느 위치를 중심으로 자료들이 모여 있는지를 나타내는 척도
- 위치를 나타내는 대표값:
  1. 평균
  2. 절사평균
  3. 중앙값
  4. 최빈값
  5. 백분위수
  6. 사분위수
- 모집단 $\{c_1, c_2, \cdots, c_N\}$에서 c_i는 i번째 추출단위의 특성값을 의미
- 표본의 특성값: $\{x_1, x_2, \cdots, x_n\}$

중심위치 척도

1. 평균(mean): 자료를 모두 더하여 자료의 수만큼으로 나누어준 값. 산술평균을 의미

- 모평균(population mean):

$$\mu = \frac{1}{N}\sum_{i=1}^{N} c_i$$

- 특성값 중에서 서로 다른 것들을 $c_1^*, c_2^*, \cdots, c_k^*$, 각각의 도수를 $f_1, f_2, \cdots, f_k$라고 하면,

$$\mu = \sum_{i=1}^{k} c_i^* \frac{f_i}{N}$$

- 표본평균(sample mean): μ을 추측하는 데 사용됨: $\bar{x} = \hat{\mu}$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

중심위치 척도

Example (2.1)

다음은 2020년 기준 우리나라 15세 이상 기혼 여성 16,549천 명의 출생아 수를 조사한 결과이다. 이 유한모집단의 모평균을 구하라.

특성값(출생아 수)	0	1	2	3	4	5*
상대도수	0.083	0.189	0.477	0.150	0.054	0.047

단, 표의 5*는 출생아 수가 5명 초과이면, 5명으로 간주했음을 의미한다.

sol) $\mu = 0 \times .083 + 1 \times .189 + \cdots + 5 \times .047 = 2.044$

- 평균은 물리적으로 무게중심의 의미를 가짐.

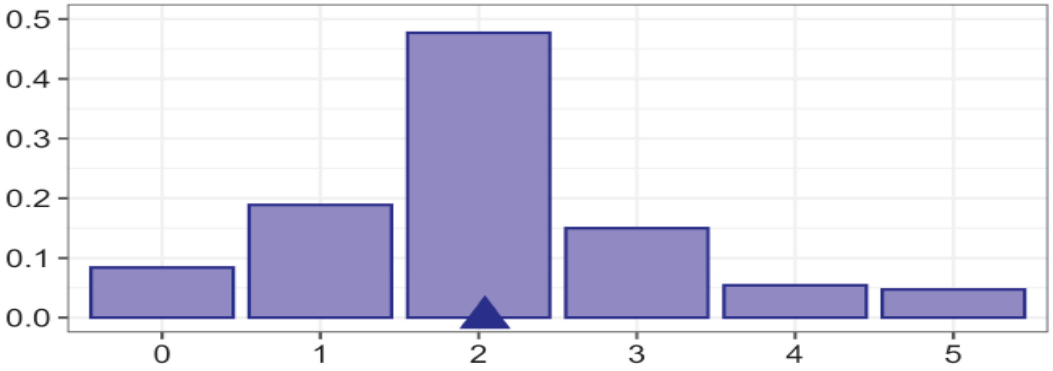


그림 2–7 균형점으로서의 평균(유한모집단)

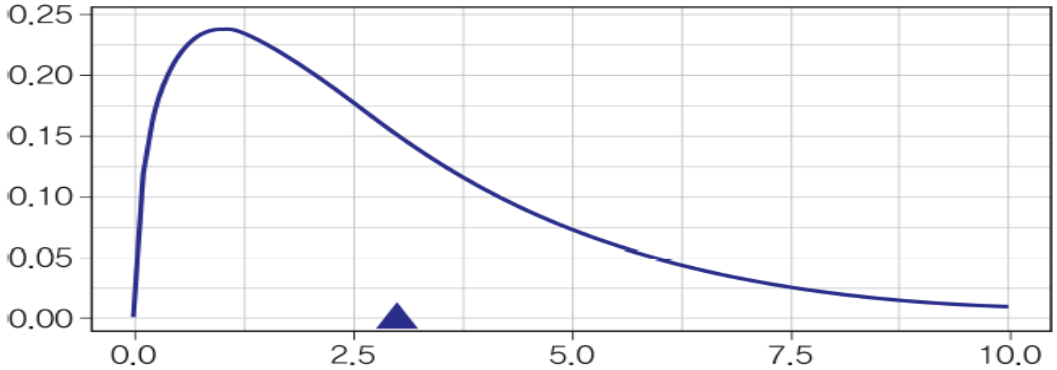


그림 2–8 균형점으로서의 평균(무한모집단)

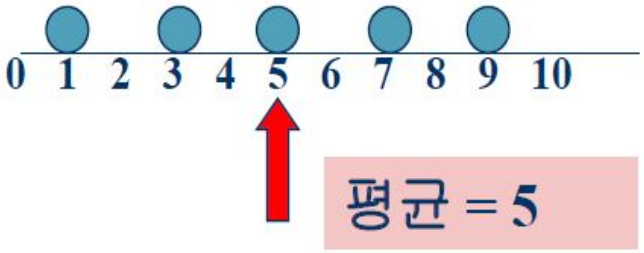
- 막대 위의 각 특성값 위치에 상대도수에 비례하는 질량을 놓았을 때의 균형점 위치
- 즉, 분포의 균형점으로서의 중심위치를 나타낸다.

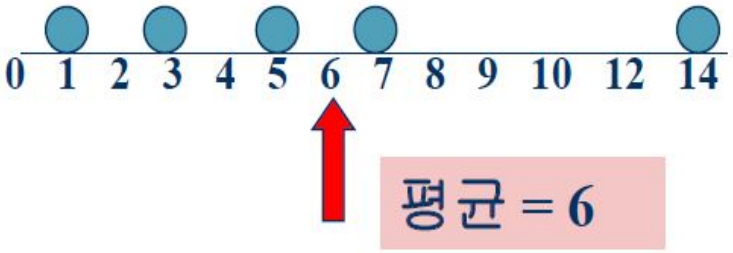
중심위치 척도

- 자료에 특이하게 작거나 큰 값, 즉 특잇값(outlier)이 있으면 표본평균은 이들 값에 크게 영향을 받음

예) 1, 3, 5, 7, 9 -> 평균=25/5=5

1, 3, 5, 7, 14 -> 평균=30/5=6





중심위치 척도

2. $100p\%$ 절사평균(trimmed mean):

   순서대로 나열된 표본의 특성값 중에서 양쪽 $100p\%$를 버린 후 가운데 $100(1-2p)\%$ 특성값의 평균으로 정의.

$$\bar{x}_p$$

- 특이하게 작거나 큰 값이 자료에 있어도 영향을 받지 않게 됨

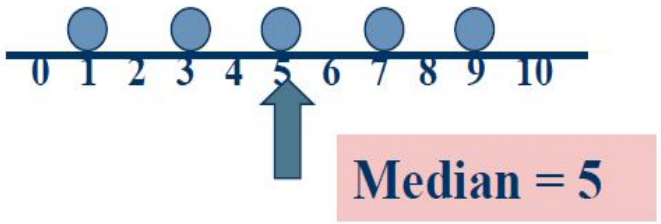
중심위치 척도

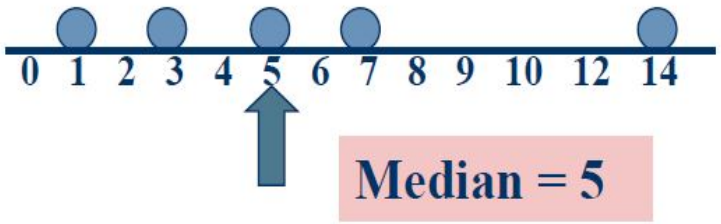
3. 중앙값(median): 특성값을 크기 순으로 늘어 놓았을 때, 가운데에 해당하는 값. 연속적인 경우에는 밀도곡선의 전체 넓이를 이등분하는 점

- 자료가 홀수인 경우: 가운데 값, $\frac{n+1}{2}$번째 자료
- 자료가 짝수인 경우: 가운데 두 개($\frac{n}{2}$, $\frac{n}{2}+1$번째) 자료의 평균
- 특이하게 작거나 큰 값이 자료에 있어도 영향을 받지 않게 됨

예) 1, 3, 5, 7, 9 -> 중앙값= 5

1, 3, 5, 7, 14 -> 중앙값= 5





중심위치 척도

4. 최빈값(mode): 자료 중 그 빈도수가 최대인 값

- 명목자료인 경우 평균과 중앙값은 의미가 없으므로 최빈값을 사용
- 자료가 적은 경우 최빈값은 무의미
- 최빈값은 여러 개가 나올 수 있음

예) 올해의 유행 옷 색깔

예) 한 의류매장에서 어느날 판매된 남성복 바지들의 허리Size는 다음과 같다. 최빈값은34이고 중앙값은 33.5이다. 이 경우 자료의 중심위치가 33.5 Inch라고 하는 것보다 34 Inch라고 하는 것이 더 상식적이다.

(31, 34, 36, 33, 28, 34, 30, 34, 32, 40)

중심위치 척도

5. 제p백분위수(pth percentile):
   특성값을 작은 것부터 순서대로 나열하였을 때 p%의 특성값이 그 값보다 작고, 또한 $(100-p)$%의 특성값이 그 값보다 크게 되는 값.

   - 연속형 자료의 경우

   - 이산형 자료의 경우

$$(n+1)\frac{p}{100}, \quad or \quad 1+(n-1)\frac{p}{100}$$

중심위치 척도

6. 사분위수:
   - 자료를 크기 순으로 나열한 후 똑같은 크기의 네 덩어리로 만들 때 그 경계에 해당하는 값
   - 자료의 $\frac{1}{4}$, $\frac{2}{4}$, $\frac{3}{4}$, 혹은 25%,50%, 75%에 해당하는 값

- 제1사분위수(first quartile): 제25백분위수, Q_1

$$Q_1 = (n+1)\frac{25}{100}\text{번째 순위 값}$$

- 중앙값(median): 제50백분위수, Q_2
- 제3사분위수(third quartile): 제75백분위수, Q_3

$$Q_3 = (n+1)\frac{75}{100}\text{번째 순위 값}$$

중심위치 척도

- 예) 다음과 같은 자료를 얻었다고하자.

  1, 0, 7, 5, 3, 2, 0, 1, 8, 4

  선형보간법을 이용하여 4분위수를 구하여라.

sol) 자료를 크기순으로 나열 -> 0,0,1,1,2,3,4,5,7,8

Q_1 : $(10 + 1) \times 25 / 100 = 2.75$ 번째에해당

따라서 $0 + (1 - 0) \times 0.75 = 0.75$.

Q_2 : $(10 + 1) \times 50 / 100 = 5.5$ 번째에해당

따라서 $2 + (3 - 2) \times 0.5 = 2.5$.

Q_3 : $(10 + 1) \times 75 / 100 = 8.25$ 번째에해당

따라서$5 + (7 - 5) \times 0.25 = 5.5$.

중심위치 척도 예제

Example

예제 2.6) 다음은 어떤 집단에서 40명을 뽑아 키를 측정한 결과이다. 이 표본자료를 이용하여 평균, 사분위수, 10% 절사평균을 구하여라.

183	168	165	158	180	176	168	160	160	150
180	183	170	170	185	170	165	170	168	154
152	173	173	165	163	163	160	178	138	171
163	168	165	165	178	170	168	183	145	174

sol) $\bar{x} = \frac{1}{40}(183 + 168 + \cdots + 174) = 167.45$

$Q_1 = (40+1) \times \frac{25}{100}$번째 값=10.25번째 값=163

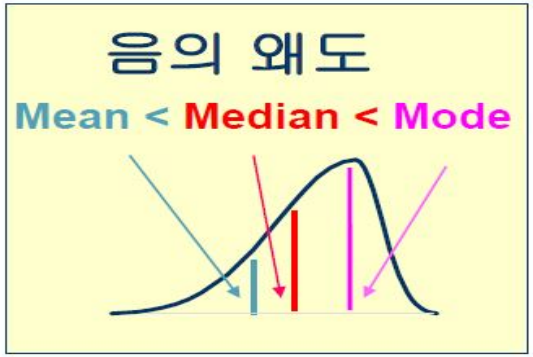
$Q_2 = (40+1) \times \frac{50}{100}$번째 값=20.5번째 값=168

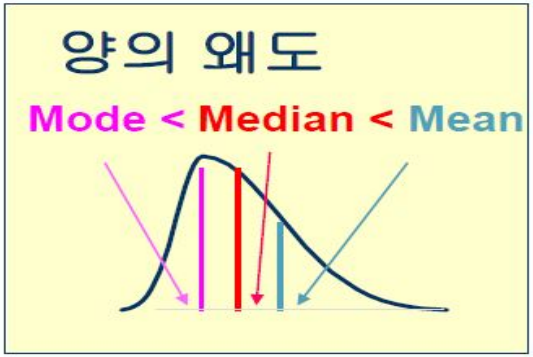
$Q_3 = (40+1) \times \frac{75}{100}$번째 값=30.75번째 값=173.75

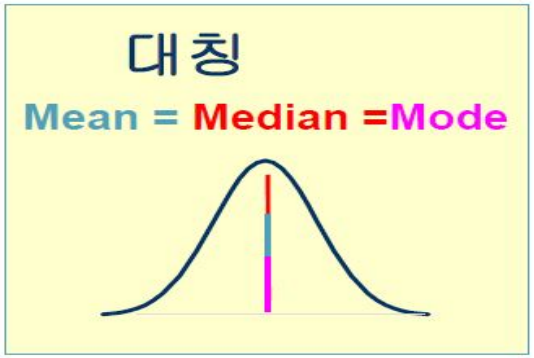
$\bar{x}_{0.1} = \frac{1}{32}(180 + 180 + \cdots + 154) = 168.09$

(양끝 4개의 값을 버린 후 가운데32개 값의 평균)

중심위치 척도의 비교







- 대칭: mean=median=mode
- 양의 왜도: mode< median<mean
- 음의 왜도: mean< median< mode

변동성 척도

- 변동성 척도: 자료들이 얼마나 변동하거나 퍼져있는지를 표시.
- 분포의 산포(특성값이 흩어져 있는 상태)를 나타내는 대표값:
  1. 평균절대편차
  2. 범위
  3. 사분위수범위
  4. 분산, 표준편차

1. 평균절대편차(mean absolute deviation(MAD)):
   각 특성값이 중앙값으로부터 떨어진 평균거리를 나타냄

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - Q_2|$$

변동성 척도

2. 범위 (Range) : 최댓값과 최솟값의 거리
   - 쉽고 빠르게 구할 수 있음
   - 특이하게 크거나 작은 값이 있을 경우 자료의 범위가 왜곡됨
   - 자료의 개수와 상관없이 같게 나올 수 있음

   ⇒ 자료의 변동성을 대표하지 못하는 경우가 많음

3. 사분위수범위(interquartile range(IQR)):
   제3사분위수와 제1사분위수의 차로 정의

$$IQR = Q_3 - Q_1$$

   - 가운데 50%특성값의 범위를 나타냄
   - 양쪽 극단 값에서 자료의 25%씩 안쪽으로 들어와 있는 값의 거리 므로 특잇값의 영향을 거의 받지 않음

변동성 척도

4. 분산(variance), 표준편차(standard deviation):

- 분산: 자료 각각이 그 평균으로부터 떨어져 있는 거리를 자승한 것의 평균값.
- 표준편차: 항목들의 자승효과를 없애기 위해 분산에 자승근을 한 것
- 분포의 산포를 나타내는 대표값.

- 모분산과 모표준편차: σ^2, σ 로 표시

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(c_i - \mu)^2, \qquad \sigma^2 = \sum_{i=1}^{k}(c_i^* - \mu)^2 \frac{f_i}{N}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(c_i - \mu)^2}$$

변동성 척도

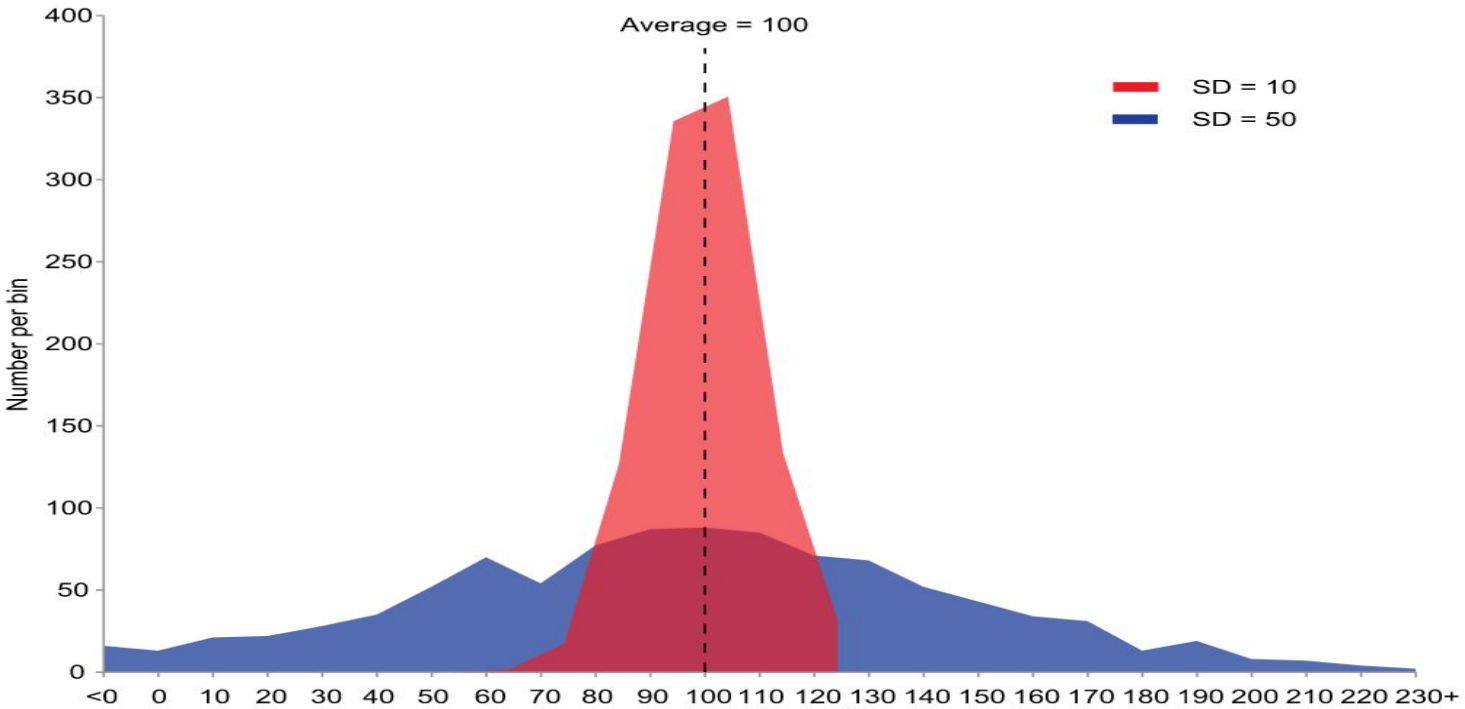
- 표본분산(sample variance)과 표본표준편차(sample standard deviation): s^2, s로 표시

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- 자료 하나하나의 값이 전부 고려되어 구해진 변동성 척도
- 값이 클수록 변동성이 큼, 넓게퍼짐
- 음수가 될 수 없음

변동성 척도

Figure: 평균이 같은 두 모집단의 분산 비교



- The red population has mean 100 and variance 100 (SD=10)
- while the blue population has mean 100 and variance 2500 (SD=50).

간편 계산식

- 간편 계산식을 통해서 다음과 같이 계산할 수 있다.

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$$

$$pf) \quad \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2$$

$$= \sum x_i^2 - 2\bar{x}^2 n + n\bar{x}^2 = \sum x_i^2 - n\bar{x}^2$$

- 따라서 표본분산을 다시 표현하면 다음과 같다.

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

변동성 척도

Example (2.3)

example 2.2의 모집단에 대하여 모분산과 모표준편차를 구하여라.

sol) $\sigma^2 = (0-1.9)^2 \times 1/10 + (1-1.9)^2 \times 3/10 + \cdots + (4-1.9)^2 \times 1/10 = 1.29$

$\sigma = \sqrt{1.29} \simeq 1.14$

Example

예제 2.6의 표본자료를 이용하여 표본사분위수범위와 표본표준편차를 구하여라.

sol) $Q_1 = 163,\ Q_3 = 173.75$이므로 $IQR = 173.75 - 163 = 10.75$

$s^2 = \frac{1}{40-1}\{(183 - 167.45)^2 + \cdots + (174 - 167.45)^2\} = 108.61$

$s = 10.42$

연관성 척도

- 연관성(association): 이차원 또는 그 이상의 항목에 대한 자료들이 상호 관련되어 있는 성질
- 이차원 특성값의 분포에서 선형관계(linear relationship)의 연관성을 나타내는 대표값들:
  1. 상관계수
  2. 공분산
- (특성1, 특성2)의 이차원 모집단 $(c_{11}, c_{21}), (c_{12}, c_{22}), \cdots, (c_{1N}, c_{2N})$가 있을 때, 특성1과 2의 모평균, 표준편차를 $\mu_1, \sigma_1, \mu_2, \sigma_2$라고 하자.
- 이차원 표본자료 $(x_1, y_1), \cdots, (x_n, y_n)$가 있을때, 특성1과 2의 표본평균, 표본표준편차를 $\bar{x}$, s_1, $\bar{y}$, s_2라 하자.

두 이차원 모집단의 결합분포

모집단 A					모집단 B				
	특성1					특성1			
특성2	0	1	2	합	특성2	0	1	2	합
0	.1	.2	.1	.4	0	.0	.0	.4	.4
1	.0	.2	.1	.3	1	.0	.2	.1	.3
2	.0	.0	.3	.3	2	.1	.2	.0	.3
합	.1	.4	.5	1.0	합	.1	.4	.5	1.0

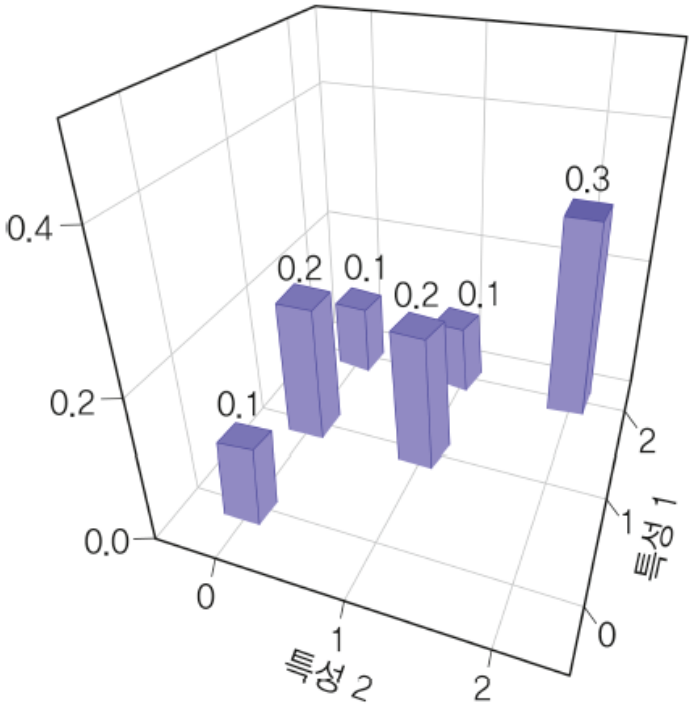
- 두 모집단 A, B의 특성별 모평균과 모표준편차는 같다.

모집단 A:

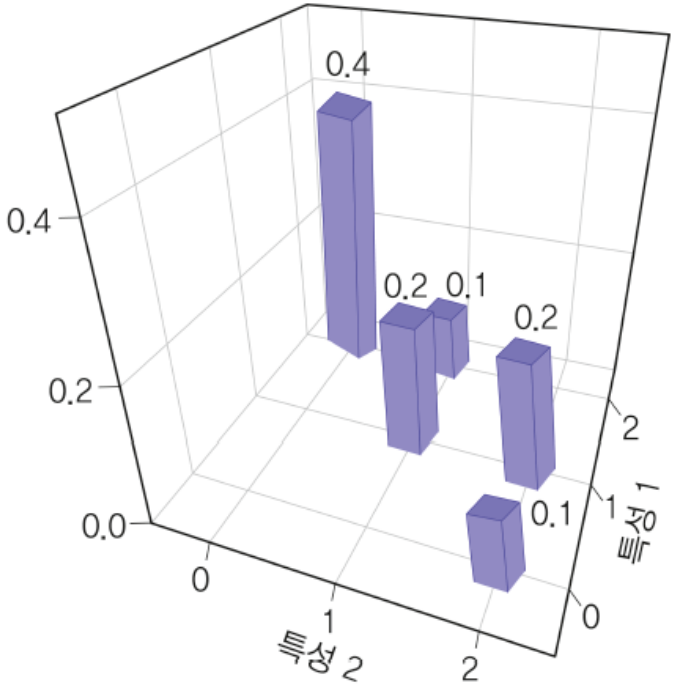
- $\mu_1 = 1.4$, $\sigma_1 = .6633$
- $\mu_2 = .9$, $\sigma_2 = .8307$

모집단 B:

- $\mu_1 = 1.4$, $\sigma_1 = .6633$
- $\mu_2 = .9$, $\sigma_2 = .8307$



(a) 모집단 A



(b) 모집단 B

그림 2–10 두 이차원 모집단의 결합분포

- 그러나 모집단 A의 분포는 좌표평면에서 양의 방향, 모집단 B는 음의 방향으로 놓여있다.
- 두 특성의 변화 관계를 나타내는 추가적인 대표값이 필요.

연관성 척도

1. 공분산(covariance): 두 변수 사이의 선형관계의 연관성을 측정
   - 두 변수가 각기 평균으로부터 떨어진 값을 서로 곱한 후, 더하여 표본에서는 $n-1$, 모집단에서는 N으로 나눈 값.

- 모집단 공분산(population covariance):

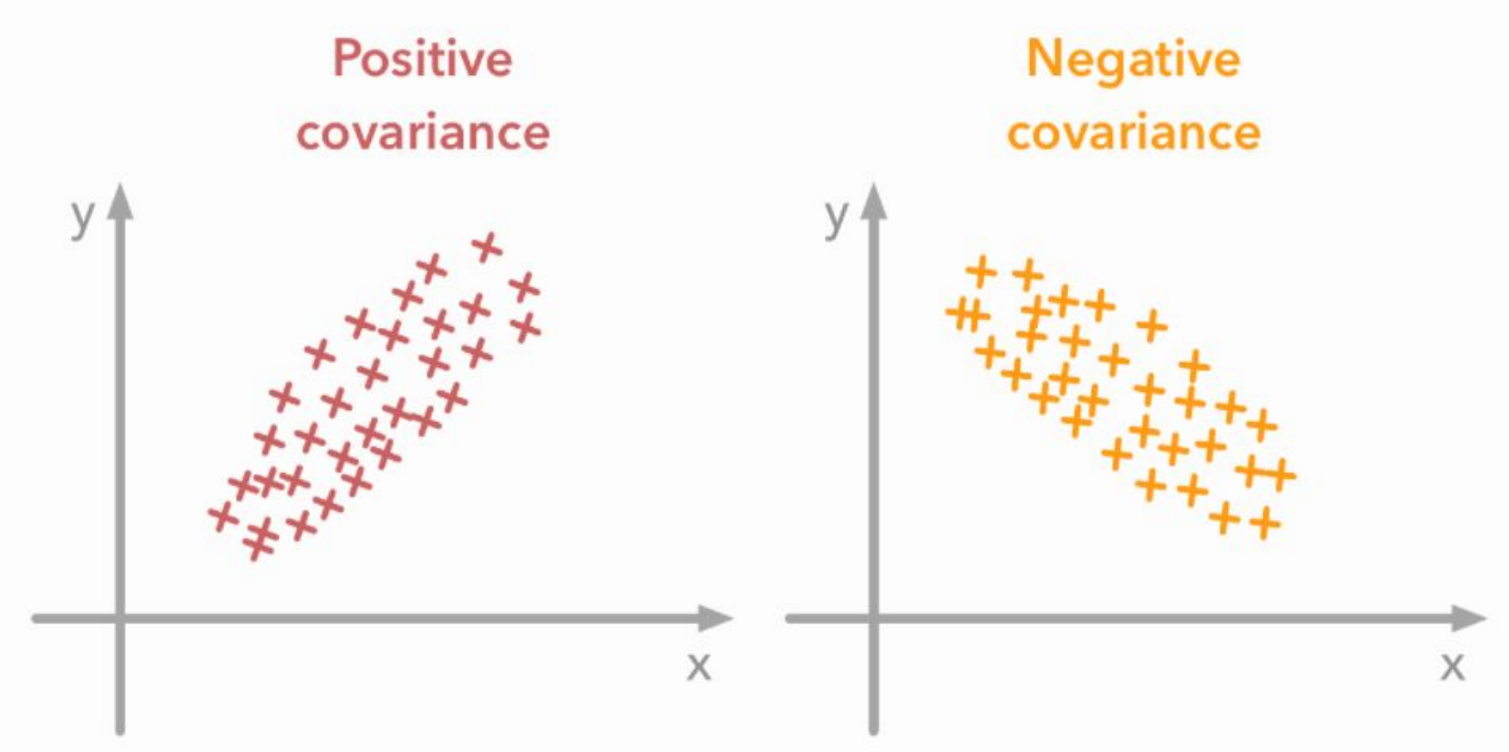
$$\rho\sigma_1\sigma_2 = \frac{1}{N}\sum_{i=1}^{N}(c_{1i}-\mu_1)(c_{2i}-\mu_2)$$

- 표본공분산(sample covariance):

$$rs_1s_2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

연관성 척도

- 공분산의 부호
  - 양의 관계(Positive Relationship)
    : x와 y가 평균에 대하여 서로 같은 값을 가질 때
  - 음의 관계(Negative Relationship)
    : x와 y가 평균에 대하여 서로 반대 값을 가질 때



연관성 척도

- 공분산의 성질: (rs_1s_2에서도 동일하게 성립)

1. $0 < \rho\sigma_1\sigma_2$: x와 y는 양의 선형관계를 의미한다.
2. $\rho\sigma_1\sigma_2 < 0$: x와 y는 음의 선형관계를 의미한다.
3. $\rho\sigma_1\sigma_2 = 0$ x와 y는 선형 관계가 존재하지 않는다.

- 각 변수간의 다른 척도 기준이 그대로 값에 반영되어 공분산 값의 크기에 대한 기준이 없음

예) (키,몸무게)

연관성 척도

2. 상관계수(correlation coefficient):
   공분산을 각자의 표준편차로 나눈 값, 공분산을 표준화 한 것.

- 모집단의 모상관계수: ρ로 표시.

$$\rho = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{c_{1i}-\mu_1}{\sigma_1}\right)\left(\frac{c_{2i}-\mu_2}{\sigma_2}\right)$$

- 표본상관계수(sample correlation coefficient): r로 표시

$$\begin{aligned} r &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_1}\right)\left(\frac{y_i-\bar{y}}{s_2}\right) \\ &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \end{aligned}$$

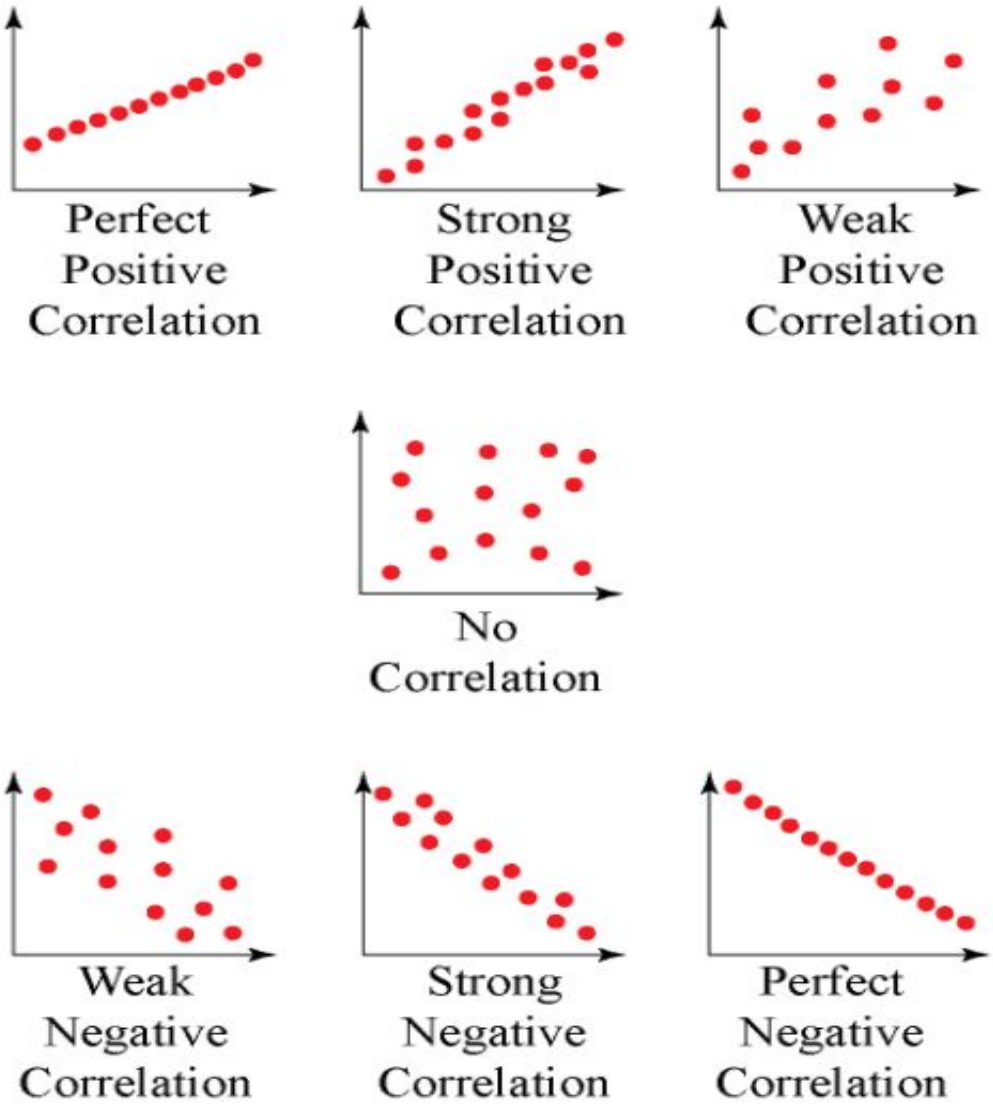
연관성 척도

- 두 변수가 어느 방향으로 얼마나 변동하는지 또는 퍼져있는지를 나타냄(부호(sign)와 크기(magnitude))

- 상관계수의 성질:

1. 항상 $-1 \leq \rho \leq 1$ (CH3, pf)
2. $0 < \rho \leq 1$: x와 y는 양의 선형관계를 의미
3. $-1 \leq \rho < 0$: x와 y는 음의 선형관계를 의미
4. $\rho = 0$: 선형 관계가 존재하지 않음
5. 기울기가 한 직선 주위에 모집단의 분포가 집중될수록 상관계수의 값은 그 직선의 방향에 따라 양 극단값-1, +1에 가까워진다

연관성 척도


Perfect Positive Correlation
Strong Positive Correlation
Weak Positive Correlation
No Correlation
Weak Negative Correlation
Strong Negative Correlation
Perfect Negative Correlation

간편 계산식

- $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$, $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2$

- $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$

$$
\begin{aligned}
pf)\ \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) &= \sum (x_i - \bar{x})y_i - \bar{y}\sum (x_i - \bar{x}) \\
&= \sum x_i y_i - \bar{x}\sum y_i \\
&= \sum x_i y_i - n\bar{x}\bar{y}
\end{aligned}
$$

- $$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_1}\right)\left(\frac{y_i - \bar{y}}{s_2}\right) = \frac{1}{n-1}\frac{\sum x_i y_i - n\bar{x}\bar{y}}{s_1 s_2}$$

연관성 척도 예제

Example

다음 표의 광고비와 매출액의 선형적 연관성을 측정하여라.

	광고비(x_i)	매출액(y_i)	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$
	5	16	-5	-9	45
	6	19	-4	-6	24
	7	18	-3	-7	21
	8	20	-2	-5	10
	9	24	-1	-1	1
	11	26	1	1	1
	12	30	2	5	10
	13	32	3	7	21
	14	31	4	6	24
	15	34	5	9	45
합	100	250	0	0	202

sol) $rs_1s_2 = \frac{202}{10-1} = 22.44,\ r = \frac{22.44}{(3.496)(6.532)} = 0.98$

연관성 척도 예제

Example

예제 2.7의 표본자료를 이용하여 표본상관계수를 구하여라.

sol) $\bar{x} = 167.45$, $s_1 = 10.42$, $\bar{y} = 61.45$, $s_2 = 11.94$이므로

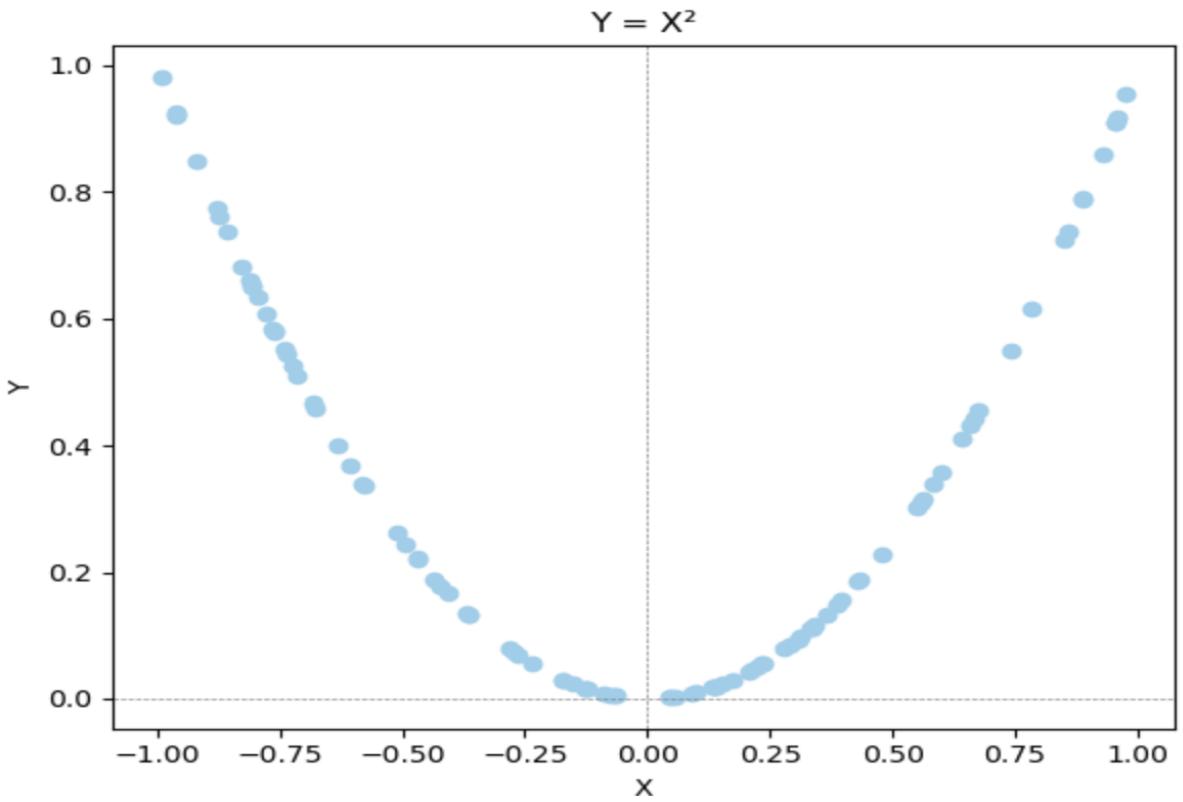
$$\begin{aligned} rs_1s_2 &= \frac{1}{40-1}\{(183-167.45)(82-61.45) \\ &\quad + \cdots + (174-167.45)(65-61.45)\} \\ &= 90.57 \end{aligned}$$

- 따라서 표본상관계수는 다음과 같다.

$$r = \frac{90.57}{10.42 \times 11.94} = 0.728$$

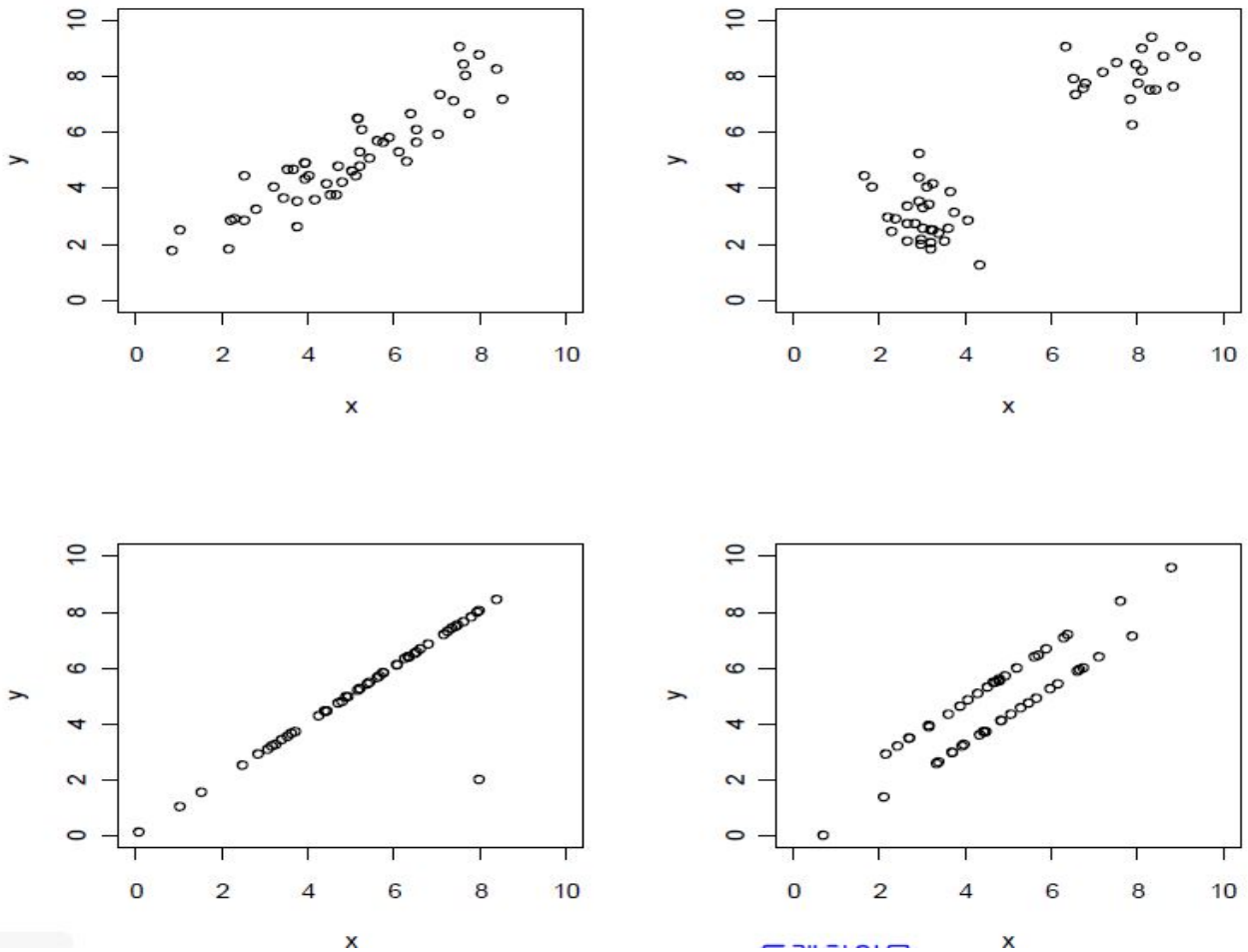
연관성 척도

Figure: 비선형 관계인 경우. $\rho = 0$



연관성 척도

Figure: $\rho = 0.9$ 값을 갖는 여러 경우



3.1장. 확률분포

박명현

서울대학교 학부대학

확률분포

- 확률변수의 대표적인 확률분포들를 소개하고 그의 성질들을 파악해보자.
- <u>이산형 확률분포</u>

  - 베르누이분포
  - 이항분포
  - 초기하분포

- <u>연속형 확률분포</u>

  - 균일분포
  - 정규분포

이산형 확률분포

베르누이분포

- 어떤 확률실험이 두 개의 결과만을 가질 수 있다고 가정하자.
  예)입시에서 합격과 불합격,
  스포츠 경기에서 승리와 패배,
  수술 후 환자의 치유 여부 등

- 베르누이시행(Bernoulli trial)

(1) 시행의 결과가 '성공(s)' 또는 '실패(f)' 두 가지뿐이고, 따라서 표본공간은 $S = \{s, f\}$로 원소가 두 개인 집합이다.

(2) 성공의 확률을 $p = P\{s\}$, 실패의 확률을 $q = P\{f\}$로 나타내며, 따라서 $0 \leq p \leq 1$이고 $q = 1 - p$가 된다.

베르누이분포

- 베르누이확률변수(Bernoulli random variable):
  베르누이시행의 표본공간 $S = \{s, f\}$에서 시행의 결과에 따라 0 또는 1의 값을 대응시키는 확률변수

$$X = \begin{cases} 1, & if\ Success \\ 0, & if\ Fail \end{cases}$$

- 성공의 확률이 p인 베르누이분포는 다음과 같다.

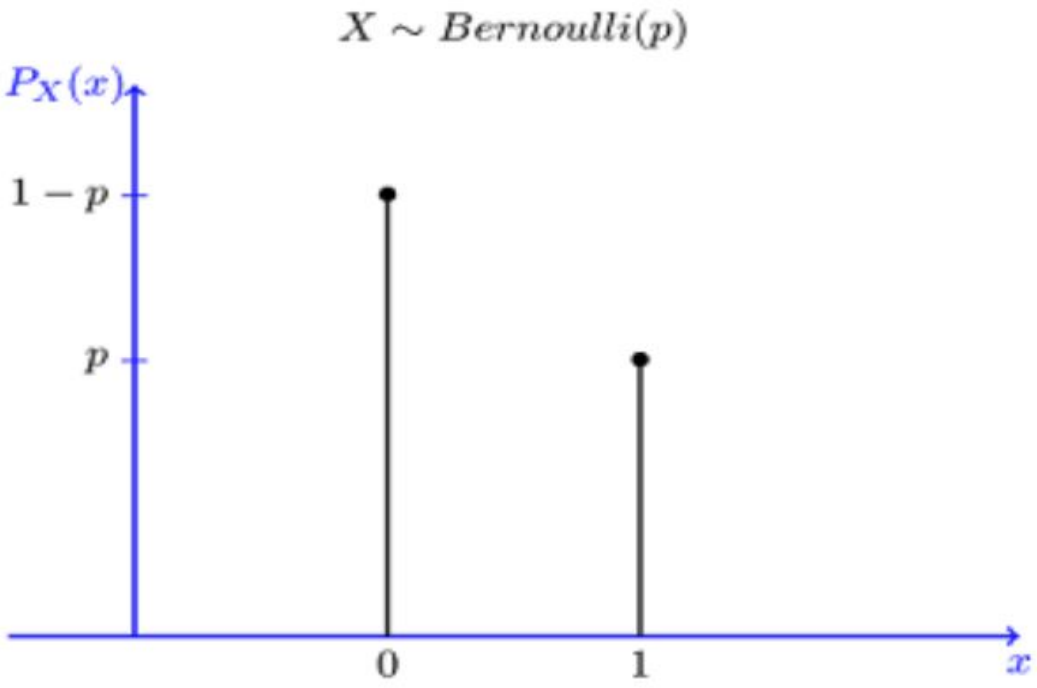
$$p(x) = p^x(1-p)^{1-x} = \begin{cases} p, & if\ (X=1) \\ 1-p, & if\ (X=0) \end{cases}$$

- 베르누이분포의 평균과 분산
  $E(X) = 1 \cdot p + 0 \cdot (1-p) = p$
  $Var(X) = E(X^2) - E(X)^2 = 1^2 \cdot p + 0^2 \cdot (1-p) - p^2 = p(1-p)$

베르누이분포

표 4-1 베르누이 분포

x	0	1
$p(x)$	$1-p$	p



이항분포(binomial distribution)

- 시행의 결과가 "1"(성공) 또는 "0"(실패) 와 같이 이원적으로 분류되는 (무한)모집단에서 랜덤표본 $X_1, X_2, \cdots, X_n$을 추출하였다.
- 추출된 확률변수 $X_1, X_2, \cdots, X_n$은 서로 독립이며 성공의 확률이 p인 베르누이분포 $B(1, p)$를 따른다:

$$X_i = \begin{cases} 1, & (success) \\ 0, & (fail) \end{cases} \qquad for\ i = 1, \cdots, n$$

- 이때 총 n번의 시행 중 성공의 횟수(표본에서 "1"의 개수)는 다음과 같이 X_i들의 합인 X로 표현되고, 이때 확률변수 X는 이항분포를 따른다.

$$X = X_1 + \cdots + X_n \sim B(n, p)$$

이항분포

- 이항분포: 성공률 p인 베르누이시행을 n번 독립적으로 반복시행할 때 성공 횟수 X의 분포.

  - 이항분포의 확률질량함수는 다음과 같다.

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x};\ x = 0, 1, \cdots, n$$

  - x번의 실험에서 성공, $n-x$번의 실험에서 실패할 확률: $p^x(1-p)^{n-x}$
  - x번의 성공과 $n-x$번의 실패를 한 줄로 늘어 놓는 경우의 수: $\binom{n}{x}$
- 이항계수(binomial coefficient): $\binom{n}{x} = \frac{n!}{(n-x)!x!}$
- 확률변수 X는 총 n번의 시행 중 성공한 횟수로 정의한다.
- 확률변수 X가 모수 (n, p)를 따를 때, $X \sim B(n, p)$라 표현한다.
- $n = 1$일 때 $B(n, p)$는 "1"의 확률이 p인 베르누이분포가 된다.

이항분포의 평균과 분산

- $X = \sum_{i=1}^{n} X_i, \quad X_i \sim B(1,p)$의 성질을 이용하여 유도가능.
- 이항분포의 평균: $E(X) = np$

$$
\begin{aligned}
pf) \quad E(X) &= E(X_1) + \cdots + E(X_n) \\
&= p + \cdots + p \\
&= np
\end{aligned}
$$

- 이항분포의 분산: $Var(X) = np(1-p)$

$$
\begin{aligned}
pf) \quad Var(X) &= Var(X_1 + \cdots + X_n) \\
&= \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j) \\
&= \sum_{i=1}^{n} Var(X_i) \\
&= np(1-p)
\end{aligned}
$$

이항분포

- $n = 6$일 때, $p = 0.5, 0.3, 0.7$값의 변화에 따른 이항분포의 모양변화

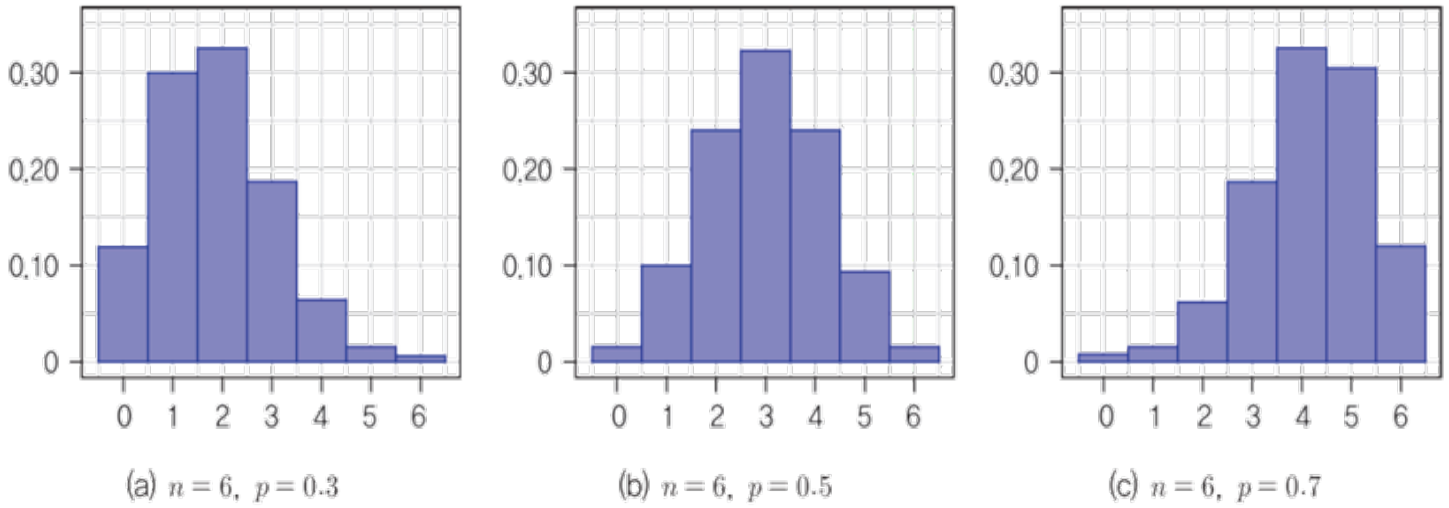


그림 4-4 $n = 6$일 때 이항분포

- $p = 0.5$: 분포의 모양이 대칭,
  $p < 0.5$: skewed to the right, $p > 0.5$: skewed to the left.

이항분포의 확률계산

- 이항분포의 누적확률분포표를 이용하여 계산.
  => $P(X \leq c)$의 값들이 여러 가지 n, p, c값에 대하여 주어져 있음.

Figure: 이항분포의 누적확률분포표

$$P\{X \leq c\} = \sum_{x=0}^{c} \binom{n}{x} p^x (1-p)^{n-x}$$

		p										
		.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
	c											
$n=1$	0	.950	.900	.800	.700	.600	.500	.400	.300	.200	.100	.050
	1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n=2$	0	.902	.810	.640	.490	.360	.250	.160	.090	.040	.010	.002
	1	.997	.990	.960	.910	.840	.750	.640	.510	.360	.190	.097
	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n=3$	0	.857	.729	.512	.343	.216	.125	.064	.027	.008	.001	.000
	1	.993	.972	.896	.784	.648	.500	.352	.216	.104	.028	.007
	2	1.000	.999	.992	.973	.936	.875	.784	.657	.488	.271	.143
	3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n=4$	0	.815	.656	.410	.240	.130	.063	.026	.008	.002	.000	.000
	1	.986	.948	.819	.652	.475	.313	.179	.084	.027	.004	.000
	2	1.000	.996	.973	.916	.821	.688	.525	.348	.181	.052	.014
	3	1.000	1.000	.998	.992	.974	.938	.870	.760	.590	.344	.185
	4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

이항분포의 확률계산

- 다양한 형태의 확률도 $P(X \leq c)$꼴로 바꾸어 계산할 수 있다.
- a, b가 정수일 때,

$$P(X = a) = P(X \leq a) - P(X \leq a - 1)$$

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a - 1)$$

$$P(a < X < b) = P(X \leq b - 1) - P(X \leq a)$$

$$P(X \geq a) = 1 - P(X \leq a - 1)$$

- binomial distribution calculator
  https://homepage.divms.uiowa.edu/ mbognar/applets/bin.html

이항분포

Example

5지 선다형 문제로 구성된 20문항의 시험에서 무작위로 답을 선택할 때, 다음 확률을 구하라.
(a) 정답을 하나도 맞히지 못할 확률
(b) 정답을 2개 이상 맞힐 확률
(c) 정답을 4개 이상 6개 이하 맞힐 확률

sol) 20문항 중 정답의 수를 X라고 하면, $X \sim B(n = 20, p = 0.2)$이고, 따라서 이항분포표에서 다음의 확률을 얻을 수 있다.

(a) $P(X = 0) = \binom{20}{0} 0.2^0 0.8^{20} = 0.012$

(b) $P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 0.931$

(c) $P(4 \leq X \leq 6) = P(X \leq 6) - P(X \leq 3) = 0.913 - 0.411 = 0.502$

이항분포

Example

한 택배 회사의 정시 배달 성공률은 95%로 알려져 있다. 배송 20건을 단순랜덤추출하여 조사했을 때, 표본에서 정시 배달률이 90% 이하일 확률을 구하라

sol) 배송 20건 중 정시 배달 실패 건수를 확률변수 X라고 하면, 근사적으로 $X \sim B(n = 20, p = 0.05)$

- 표본에서 정시 배달 성공률이 90% 이하라는 것은 정시 배달 실패율 $\frac{X}{20}$가 10% 이상이라는 뜻이다. 이를 식으로 나타내어 계산하면 다음과 같다.

$$P(\frac{X}{20} \geq 0.1) = P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 0.264$$

초기하분포(hypergeometric distribution)

예) 어떤 주머니에 D개의 빨간 공과 N-D개의 하얀 공이 들어 있을 때, n개의 공을 랜덤하게 비복원으로 뽑았을 때, X를 빨간 공의 개수라고 하자.

- 특성값이 두 가지로 분류되는 경우
- 특성값 "1"의 개수가 D, "0"의 개수가 $N-D$인 크기 N의 (유한)모집단에서 크기 n인 랜덤표본을 비복원으로 뽑음
- $X = \sum_{i=1}^{n} X_i, \quad X_i \sim B(1,p)$
- n개의 표본에서 "1"의 개수를 통계량 X라고 하자.
- X는 초기하분포를 따르며, 그 확률질량함수는 다음과 같다.

$$p(x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}; \; x = 0, 1, \cdots, n \; (n \le D, n \le N-D)$$

초기하분포(hypergeometric distribution)

- 모비율 $p = \frac{D}{N}$의 추정은 표본비율 $\hat{p}$을 이용

  $\hat{p}$=(표본에서 "1"의 개수)$/n = \frac{X}{n}$

- 초기하분포의 평균과 분산

  - $E(X) = np$ 단, $p = D/N$

  - $Var(X) = np(1-p) \cdot \frac{N-n}{N-1}$

- 초기하분포의 이항분포 근사

: 초기하분포(비복원추출)에서 추출방법을 복원추출로 하는 경우 다음의 조건을 만족하면, 초기하 분포는 이항분포 $B(n, p)$에 수렴한다.

$$N \longrightarrow \infty,\ D \longrightarrow \infty,\ \frac{D}{N} \longrightarrow p$$

초기하분포

Example

신세대 의식 구조 조사를 실시하고자 찬성자의 수가 6명, 반대자의 수가 4명인 크기가 10인 집단에서 3명의 랜덤표본을 뽑았다. 이 중 찬성자수의 평균과 분산을 구하여라.

sol) 찬성자 수 X는 아래의 초기하분포를 따른다.

$$p(x) = \frac{\binom{6}{x}\binom{4}{3-x}}{\binom{10}{3}}; \ x = 0, 1, 2, 3$$

- 따라서, X의 평균과 분산은 다음과 같이 계산된다.

$$E(X) = 0 \cdot \frac{\binom{6}{0}\binom{4}{3}}{\binom{10}{3}} + 1 \cdot \frac{\binom{6}{1}\binom{4}{2}}{\binom{10}{3}} + 2 \cdot \frac{\binom{6}{2}\binom{4}{1}}{\binom{10}{3}} + 3 \cdot \frac{\binom{6}{3}\binom{4}{0}}{\binom{10}{3}} = \frac{9}{5}$$

- 마찬가지 방법으로 $E(X^2) = \frac{19}{5}$. 따라서

$$Var(X) = E(X^2) - E(X)^2 = \frac{14}{25}$$

초기하분포의 이항분포 근사

Example

어떤 제품을 생산하는 공정의 불량률은 5%로 알려져 있다. 오늘 생산한 10,000개의 제품 중에서 20개를 단순랜덤추출하여 조사할 때 표본의 불량률이 10%이상일 확률을 구하여라.

sol) 20개 중 불량품의 수를 X라고 하면 X는 초기하분포를 따른다.

$$\begin{aligned} P(X/20 \geq 0.1) &= P(X \geq 2) \\ &= 1 - P(X=0) - P(X=1) \\ &= 1 - \frac{\binom{9500}{19}\binom{500}{1}}{\binom{10000}{20}} - \frac{\binom{9500}{20}\binom{500}{0}}{\binom{10000}{20}} = \cdots \end{aligned}$$

- 그런데 X는 근사적으로 이항분포 $B(20, 0.05)$를 따르므로

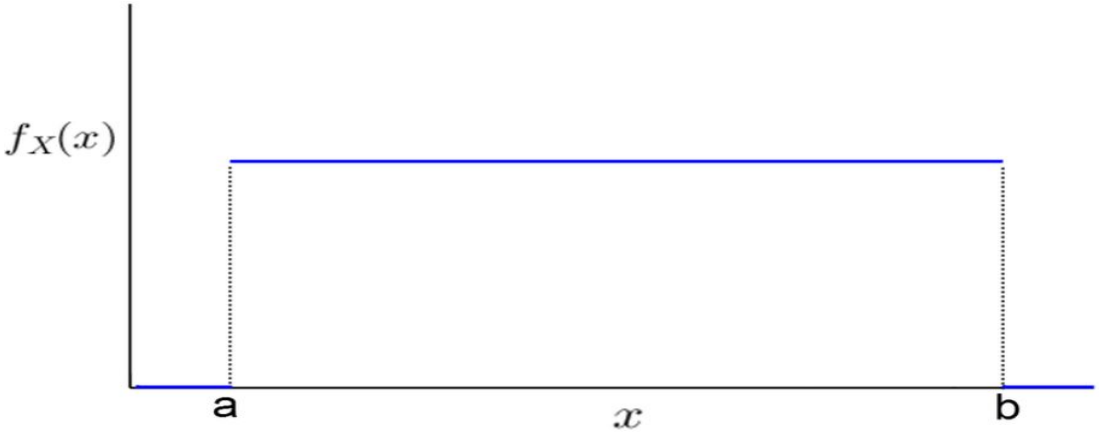
$$P(X \geq 2) = 1 - P(X \leq 1) \approx 1 - 0.736 = 0.264$$

연속형 확률분포

균일분포(uniform distribution)

- 균일분포: 확률변수 X가 어느 구간 (a,b)에서 정의되고, 확률밀도 함수의 크기(높이)가 동일한 확률분포

Figure: U(a,b)의 확률밀도함수



예) 약속시간에 늦는 시간이 특별한 경향이 없이 0 ∼ 60분 사이에 랜덤하게 골고루 퍼져 있을 때의 늦는 시간 X

- 컴퓨터로 0과 1사이에서 난수(random number)를 생성하는 경우, 생성된 난수를 X라고 하면 그 값들은 모두 동일한 확률로 나타남

균일분포(uniform distribution)

- 균일분포의 확률밀도함수:

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & o.w \end{cases}$$

- $E(X) = \int_a^b x \cdot f(x)dx = \int_a^b x \cdot \frac{1}{b-a} dx = \frac{a+b}{2}$
- $E(X^2) = \int_a^b x^2 \cdot f(x)dx = \int_a^b x^2 \cdot \frac{1}{b-a} dx = \frac{a^2+ab+b^2}{3}$
- $V(X) = E(X^2) - E(X)^2 = \frac{(b-a)^2}{12}$
- $a \le c \le d \le b$를 만족하는 c와 d에 대하여

$$P(c \le X \le d) = \int_c^d f(x)dx = \frac{1}{b-a} \int_c^d 1dx = \frac{d-c}{b-a}$$

정규분포(normal distribution)

- 정규분포: 연속형 분포 가운데 가장 널리 쓰이는 확률분포
- 자연현상에서 비롯된 수많은 결과들은 그 값이 평균에 집중되어 있고 평균에서 멀어질 수록 도수가 작아지는 현상을 보인다.
- 정규분포곡선은 평균값을 중심으로 하여 좌우대칭의 형태를 띄고, 평균에서 좌우로 멀어질수록 수평축에 무한히 가까워지는 종 모양을 이룬다.

정규분포

- 정규분포를 결정하는 중요한 두 개의 값: 평균 μ, 표준편차 σ
- 확률변수 X가 평균이 μ, 표준편차가 σ인 정규분포를 따를 때, 이를 $X \sim N(\mu, \sigma^2)$으로 나타낸다. 이의 밀도함수는 다음과 같다.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad -\infty < x < \infty \tag{1}$$

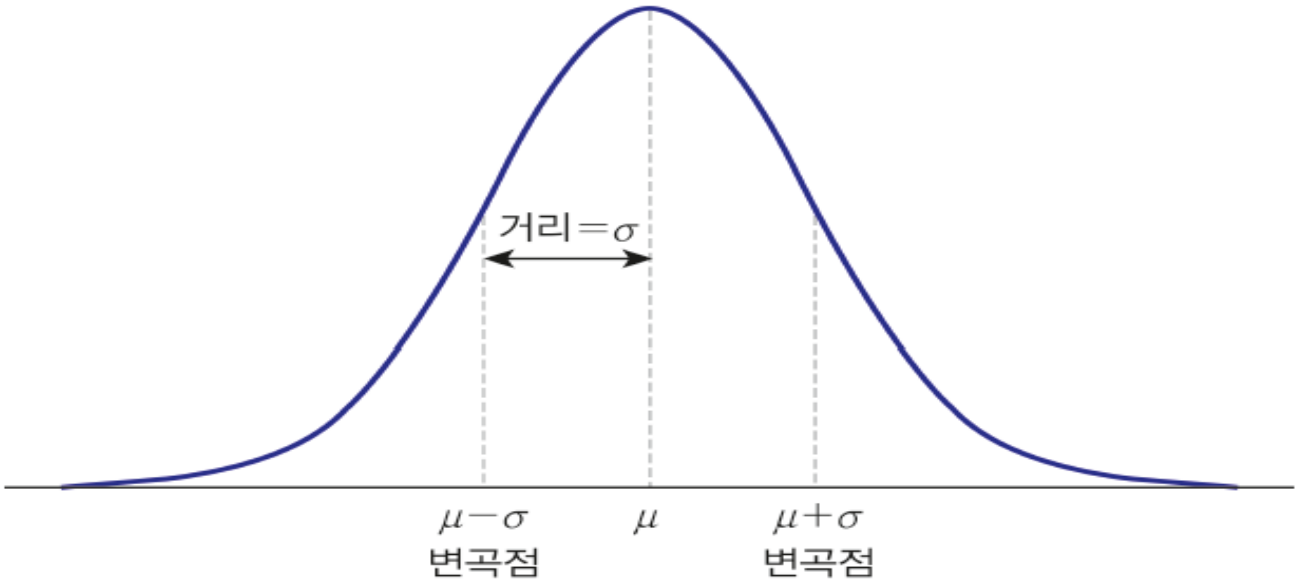


그림 4–1 정규분포의 밀도함수

정규분포의 특성

- 정규분포의 밀도곡선의 특징

  (1) 모평균 μ를 중심으로 좌우 대칭

  (2) 곡선은 대칭점 μ에서 가장 높음

  (3) 대칭점에서 멀어질 수록 곡선의 높이가 점점 낮아짐

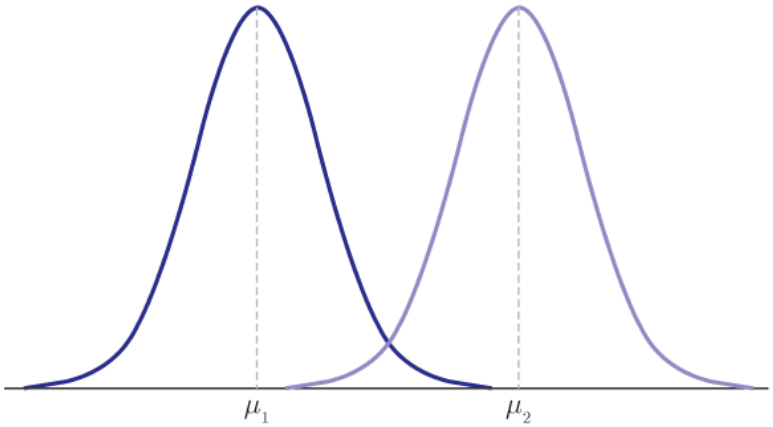
  (4) 곡선의 변곡점은 μ을 기준으로 좌우 각각 σ만큼 떨어져 있음

- $X \sim N(\mu, \sigma^2)$일 때, $P(a \leq X \leq b)$는 정규곡선의 (a,b)구간의 아래 부분의 영역의 넓이에 해당

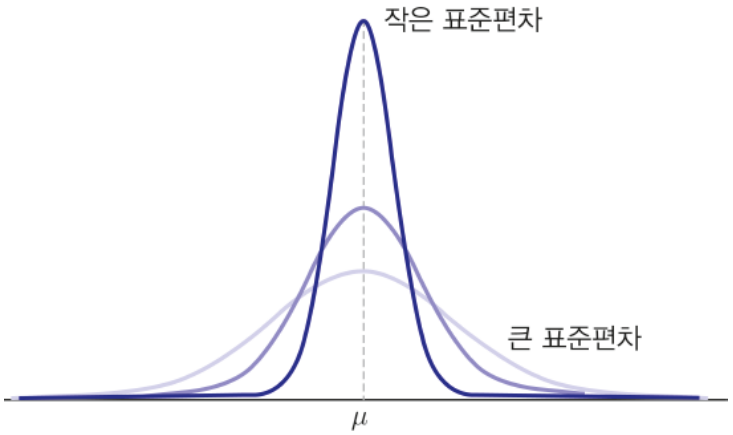
  - 정규곡선 아래 전체의 넓이는 1

정규분포의 특성

- 평균값의 변화는 곡선의 중심위치만 이동시키고, 곡선의 모양에는 영향을 미치지 않음
- σ가 작아질수록 μ 근처의 곡선의 넓이가 커지고, 반대로 σ가 커질수록 곡선은 점차 평평해짐



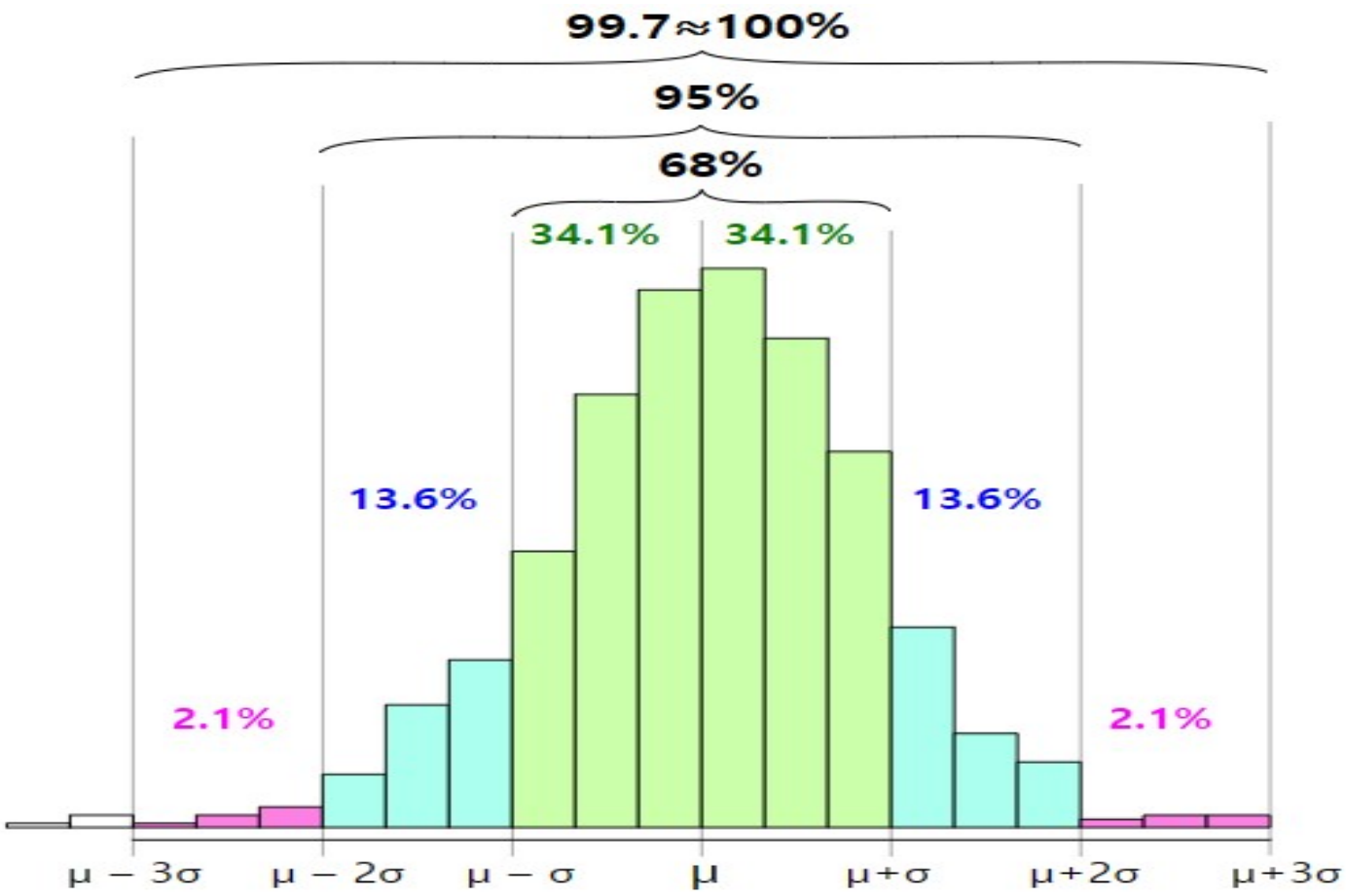
(a) 표준편차는 같고 평균이 다른 경우



(b) 평균은 같고 표준편차가 다른 경우

그림 4–2 평균과 표준편차의 변화에 따른 정규분포의 변화

변동성 척도(normal distribution case)



- $P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 68.27\%$
- $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 95.45\%$

정규분포의 성질

- 정규분포의 성질 1:
  $X \sim N(\mu, \sigma^2)$일 때 임의의 상수 a,b에 대하여 다음이 성립한다.

$$aX + b \sim N(a\mu + b, a^2\sigma^2)$$

- 정규분포의 성질 2:
  $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$이고 X_1과 X_2가 서로 독립이면

$$a_1X_1 + a_2X_2 \sim N(a_1\mu_1 + a_2\mu_2, a_1^2\sigma^2 + a_2^2\sigma_2^2)$$

표준정규분포(standard normal distribution)

- 정규분포 $N(\mu, \sigma^2)$를 따르는 확률변수 X를 표준화하면, 표준화된 확률변수는 평균이 0, 표준편차가 1인 정규분포를 따른다. 즉, $X \sim N(\mu, \sigma^2)$일 때, 다음이 성립

$$\frac{X-\mu}{\sigma} \sim N(0,1)$$

- 표준정규분포: 평균이 0, 표준편차가 1인 정규분포, 확률변수 Z로 표기

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

- 표준정규분포의 확률밀도함수:

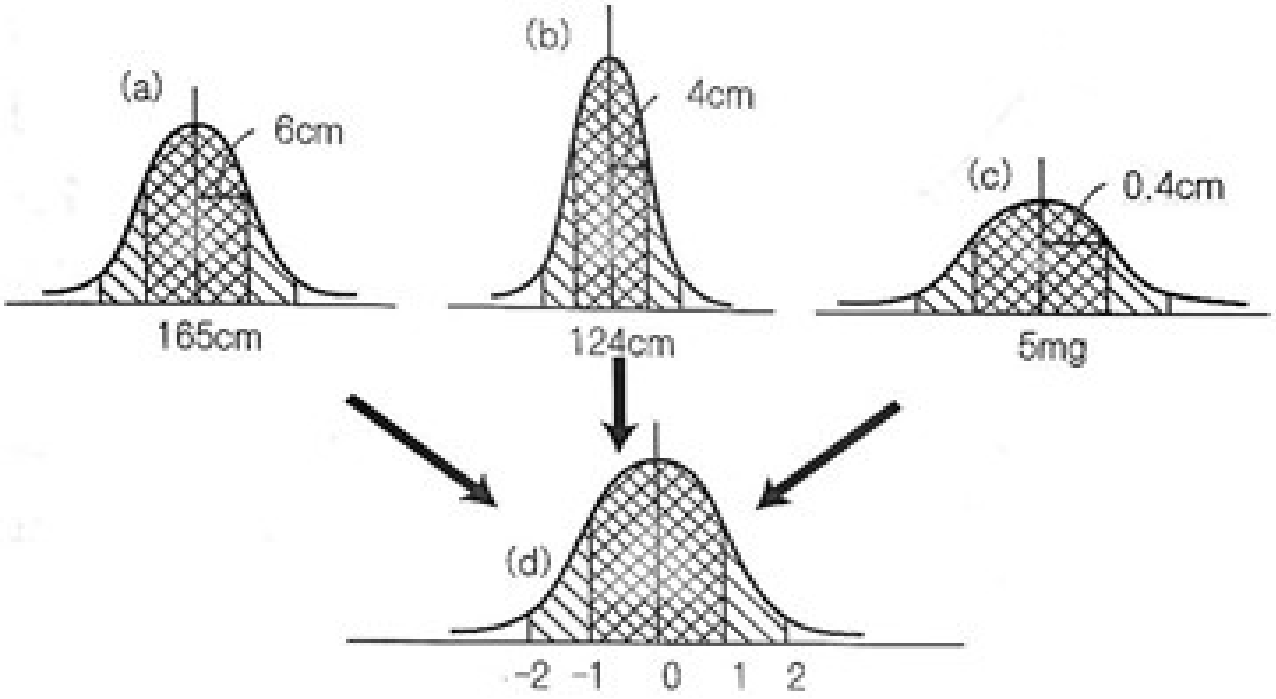
$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \ -\infty < z < \infty$$

정규분포의 성질

- 정규분포의 성질 3:
  $X \sim N(\mu, \sigma^2)$이면
  $$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

- 정규분포의 성질 4:
  $Z \sim N(0, 1)$이면
  $$X = \sigma Z + \mu \sim N(\mu, \sigma^2)$$

정규분포의 표준화



(a) 성인 남자의 신장(cm), (b) 초등학교 3학년의 신장(cm),
(c) 한 알의 위장약에 포함되어있는 판토텐산칼슘의 양(mg)

정규분포의 확률계산

- 정규분포 $N(\mu, \sigma^2)$에 관한 여러가지 확률은 표준정규분포에 관한 확률로 바꾸어 계산할 수 있다.
- 정규분포의 확률 계산: 확률변수 X의 분포가 $N(\mu, \sigma^2)$일 때, $(X-\mu)/\sigma$의 분포는 $N(0,1)$이므로, 다음의 사실이 성립함

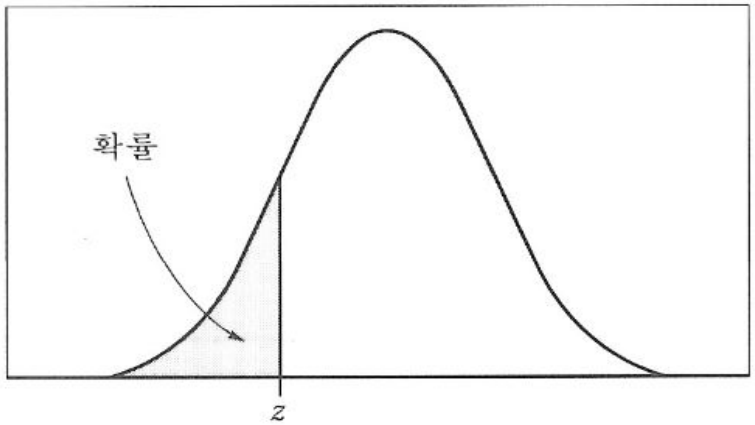
$$P(X \le b) = P\left(\frac{X-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right) = P\left(Z \le \frac{b-\mu}{\sigma}\right)$$

$$\begin{aligned} P(a \le X \le b) &= P\left(\frac{a-\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right) \\ &= P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right) \\ &= P\left(Z \le \frac{b-\mu}{\sigma}\right) - P\left(Z \le \frac{a-\mu}{\sigma}\right) \end{aligned}$$

표준정규분포표

- 표준정규분포의 확률: 표준정규분포표에 여러 가지 z값의 왼쪽 부분의 넓이, 즉 $P(Z \leq z)$의 값이 계산되어 있다.
- 예) $P(Z \leq -3.42) = 0.0003$

Figure: 표준정규분포표



표의 숫자는 z보다 같거나 작은 확률을 나타낸다.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014

표준정규분포의 확률계산

- 표준정규분포표를 이용한 확률의 계산

(1) $P(a \leq Z \leq b) = P(Z \leq b) - P(Z < a)$

(2) $P(Z \leq -z) = P(Z \geq z)$

(3) $P(Z \geq z) = 1 - P(Z < z)$

(4) $P(Z \leq -z) = 1 - P(Z < z)$

(5) $P(Z \leq 0) = P(Z \geq 0) = \frac{1}{2}$

(6) $P(|Z| \leq z) = 2 \times (P(Z \leq z) - 0.5)$

표준정규분포의 확률계산

Example

$P(Z < -1.9$ 또는 $Z > 2.1)$을 구하여라.

sol) $P(Z < -1.9$ 또는 $Z > 2.1) = P(Z < -1.9) + P(Z > 2.1)$

- 정규분포표에서

  $P(Z < -1.9) = 0.0287$

  $P(Z > 2.1) = 1 - P(Z \leq 2.1) = 1 - 0.9821 = 0.0179$
- 따라서 $P(Z < -1.9$ 또는 $Z > 2.1) = 0.0466$

Example

$P(Z > z) = 0.025$가 되는 z의 값을 구하여라.

sol) $P(Z \leq z) = 1 - P(Z > z) = 1 - 0.025 = 0.975$이므로

- 표준정규분포표에서 z의 값은 1.96임을 알 수 있다.

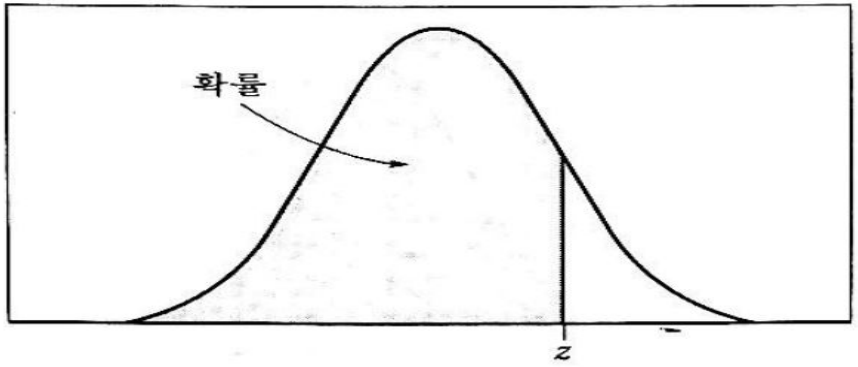
표준정규분포의 확률계산

Example

$P(Z \le c) = 0.7$이 되는 c의 값을 구하여라.

sol) $P(Z \le 0.52) = 0.6985 < 0.7 < 0.7019 = P(Z \le 0.53)$

1. $c = (0.52 + 0.53)/2 = 0.525$
2. 선형보간법을 이용.



표의 숫자는 z보다 같거나 작은 확률을 나타낸다.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549

정규분포의 백분율

- $Z \sim N(0,1)$일 때, 주어진 α값에 대하여 $P(Z > z) = \alpha$를 만족하는 z의 값을 z_α로 표기하고, 이를 표준정규분포의 제 $100(1-\alpha)$백분위수 또는 $(1-\alpha)$분위수(quantile)라고 한다.
- 예) $z_{0.005} = 2.58$, $z_{0.025} = 1.96$, $z_{0.05} = 1.645$

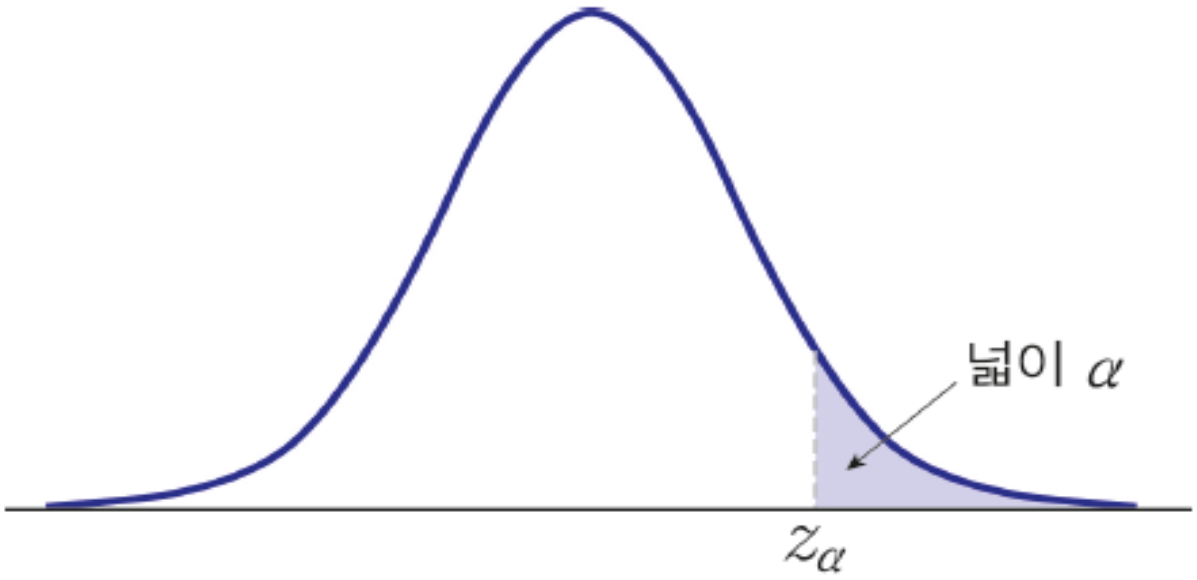


그림 4–3 z_α의 위치

정규분포의 확률계산

Example

통계학 수업을 듣는 학생들의 성적 분포가 근사적으로 $N(60, 10^2)$을 따른다고 한다. 75점 이상을 받은 학생에게 A학점을 부여할 때, A학점을 받을 학생의 비율을 근사적으로 구하라

sol) 통계학 성적을 나타내는 확률변수 $X \sim N(60, 10^2)$라 하면, 75점 이상을 받게 될 학생의 비율은 근사적으로 다음과 같다.

$$\begin{aligned} P(X \geq 75) &= P\left(\frac{X-60}{10} \geq \frac{75-60}{10}\right) \\ &= P(Z \geq 1.5) \\ &= 0.0668 \end{aligned}$$

정규분포의 확률계산

Example

매일 아침 운동으로 달리기와 자전거 타기를 한다고 하자. 달리기로 소모하는 칼로리는 평균 200kcal, 표준편차 15kcal인 정규분포를 근사적으로 따르고, 자전거 타기로 소모하는 칼로리는 평균 80kcal, 표준편차 5kcal인 정규분포를 근사적으로 따른다고 한다. 이때 하루 운동하여 300kcal 이상을 소모할 확률을 구하라.

sol) X_1: 달리기로 소모하는 칼로리

X_2: 자전거 타기로 소모하는 칼로리

$S = X_1 + X_2$: 운동으로 소모하는 총 칼로리

$$X_1 \sim N(200, 15^2), X_2 \sim N(80, 5^2)$$

- 이때, 운동으로 소모하는 총 칼로리 S도 정규분포를 따르고 이때의 평균과 분산은 각각 다음과 같다.

$$E(S) = E(X_1) + E(X_2) = 280$$

$$Var(S) = Var(X_1) + Var(X_2) = 15^2 + 5^2 = 250$$

- 따라서 S는 다음과 같은 정규분포를 따른다.

$$S \sim N(280, 250)$$

- 하루 운동으로 300kcal 이상을 소모할 확률은

$$\begin{aligned} P(S > 300) &= P\left(\frac{S - 280}{\sqrt{250}} > \frac{300 - 280}{\sqrt{250}}\right) \\ &= P(Z > 1.26) = 1 - P(Z \leq 1.26) \\ &= 1 - 0.8962 = 0.1038 \end{aligned}$$

이상치 판단 기준 (Outlier Detection)

1. IQR (Interquartile Range) 기준
   - Q1 = 제1사분위수, Q3 = 제3사분위수
   - IQR = Q3 - Q1
   - 이상치: $x < Q1 - 1.5 \times IQR$ 또는 $x > Q3 + 1.5 \times IQR$
2. Z-score 기준 (정규분포 가정)
   - $z = \frac{x-\mu}{\sigma}$ (μ: 평균, σ: 표준편차)
   - 일반적으로 $|z| > 2$ 또는 $|z| > 3$이면 이상치
3. 시각적 방법
   - Boxplot: 사분위 기반 이상치 확인
   - Scatter plot: 변수 관계에서 튀는 점 확인

3장. 확률과 확률분포

박명현

서울대학교 학부대학

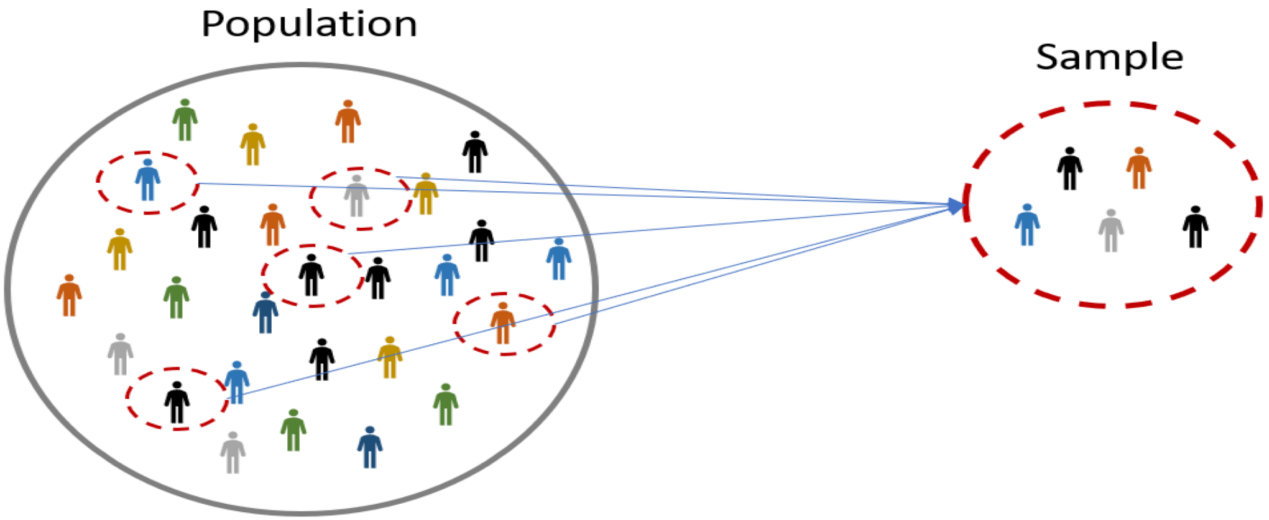
Review

- 확률의 정의
- 확률의 성질
  - 조건부/곱셈/전확률/베이즈/독립
- 확률변수
  - 이산형/연속형
- 확률분포의 대표값
  - 중심/변동/연관성

확률(probability)의 정의

- 표본이 뽑히는 과정
- 통계자료를 분석하기 위해서는 통계이론이 필요
- 통계적 조사에서는 표본을 바탕으로 모집단 전체에 대한 결론을 이끌어 낸다.

  - 이에 대한 논리적 근거가 되는 것이 확률의 개념이다.



확률의 정의

- 확률을 정의하기 위해 다음의 용어를 알아보자.

- 사건(event): 발생 가능한 결과들의 집합
- 단순사건(simple event) 또는 근원사건(elementary event): 오직 한 개의 원소로 이루어진 사건
- 표본공간(sample space): 일어날 수 있는 모든 가능한 단순사건을 모아 집합으로 표시한 것. S로 표시.
  - 모든 원소를 포함(exhaustive)
  - 상호배반(mutually exclusive)

사건과 표본공간의 예

- 주사위를 던지는 경우 표본공간 $S = \{1, 2, 3, 7\text{보다 작은 짝수}\}$?

- 동전과 주사위를 동시에 던지는 실험에 대한 표본공간
  sol) $S = \{H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6\}$

- 어느 공정에서 10개의 불량품이 제조될 때까지 제조된 제품의 개수를 조사할 때,
  sol) $S = \{10, 11, 12, \cdots\}$
- 한 공장에서 생산된 전구의 수명시간을 조사하기 위해 한 개의 전구를 추출하여 수명시간을 관측할 때,
  sol) $S = \{t | t \geq 0\}$

확률의 정의

- 확률의 고전적 정의(라플라스, Laplace, 1749-1827)

: N개의 원소로 구성된 표본공간 $S = \{e_1, \cdots, e_N\}$에서 각각의 근원 사건이 일어날 가능성이 같은 경우에 m개의 원소로 구성된 사건 A의 확률은 다음과 같이 정의한다.

$$P(A) = \frac{m}{N}$$

예) 주사위를 던졌을 때 홀수의 눈이 나올 확률:

$A = \{1, 3, 5\}$, $P(A) = \frac{3}{6}$

- 확률을 가능성에 대한 수리적 측도로 위와 같이 정의
- 표본공간이 유한개의 원소만을 가질 때 적용됨
- 각 결과가 나타날 가능성이 같지 않은 경우에는 적용이 안됨

확률 예제

Example (3.1)

흰 탁구공 3개와 노란 탁구공 1개가 들어 있는 상자에서 탁구공 2개를 단순랜덤추출할 때, 흰 탁구공 1개와 노란 탁구공 1개가 나올 확률을 구하라.

sol) - 흰 탁구공 3객를 각각 a,b,c로, 노란 탁구공을 d로 표기하면 표본공간은 다음과 같다.

$$S = \{\{a,b\},\{a,c\},\{a,d\},\{b,c\},\{b,d\},\{c,d\}\}$$

- 흰 탁구공 1개와 노란 탁구공 1개가 나올 사건은

$$A = \{\{a,d\},\{b,d\},\{c,d\}\}$$

- 색상과 상관없이 탁구공 2개가 뽑힐 가능성은 모두 같다고 할 수 있으므로

$$P(A) = \frac{3}{6} = \frac{1}{2}$$

확률의 정의

- 일반적인 확률의 개념은 동일한 통계적 조사를 반복적으로 시행하면서 누적되는 경험을 바탕으로 이루어짐
- 한 공장에서 생산되는 제품의 불량품에 관심

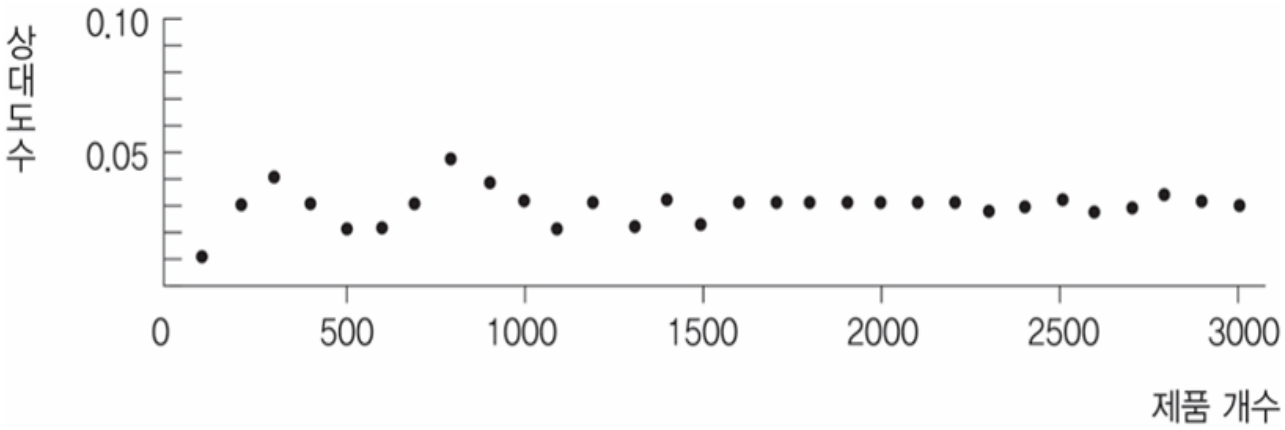


그림 3-1 불량품의 상대도수 변화

- 상대적 비율 접근법:
  불량품의 비율은 제품의 수가 증가함에 따라 어떤 상수에 수렴
  -> 이 상수를 불량품이 생산될 확률로 정의

확률의 정의

- 안드레이 콜모고로프(Andrey Kolmogorov, 1903 1987)는 확률을 상대도수의 극한 개념으로 위와 같이 확률의 공리적 정의를 정립

- 확률의 공리적 정의:

  (1)표본공간 S에서 임의의 사건 A에 대하여 $0 \leq P(A) \leq 1$

  (2) $P(S) = 1$

  (3) 서로 배반인 사건 $A_1,\ A_2, \cdots$에 대하여

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$$

  를 만족할 때, $P(A)$를 사건 A의 확률이라 한다.

- 상대도수의 극한과 비슷하게 반복 시행에서의 경험을 바탕으로 한 경험적 확률뿐만 아니라 개인의 주관에 따라 확률 값을 정하는 주관적 확률 개념도 포함

확률에 관한 성질

- $A \cup B$: 사건 A, B의 합집합, $A \cap B$: 사건 A, B의 곱집합
- A^c: 사건 A의 여집합

- 확률에 관한 성질

(1) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(2) $P(A) = 1 - P(A^c)$

(3) $A \subset B$이면 $P(A) \leq P(B)$

Figure: $A \cup B$의 분할

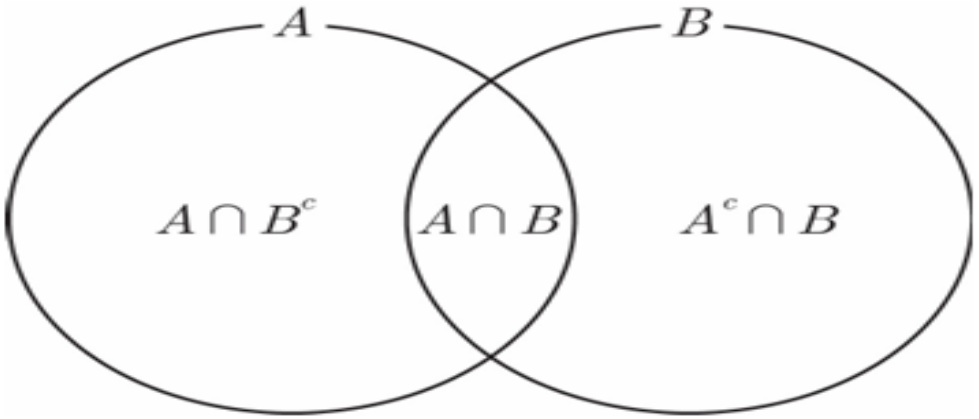


그림 3–2 $A \cup B$의 분할

확률에 관한 성질

Example (3.2)

세 개의 주머니에서 각각 1부터 5까지의 번호가 매겨진 5개의 구슬이 들어 있다. 각 주머니에서 1개씩의 구슬을 꺼낼 때, 꺼낸 구슬의 숫자의 합이 5 이상일 확률을 구하여라.

sol) A: 꺼낸 숫자의 합이 5 이상일 사건

B: 꺼낸 숫자의 합이 3일 사건, C: 꺼낸 숫자의 합이 4일 사건

$$P(A^c) = P(B) + P(C)$$

- 가능한 모든 결과는 $5 \times 5 \times 5 = 125$가지이고
  $B = \{(1,1,1)\}$, $C = \{(1,1,2).(1,2,1).(2,1,1)\}$이므로

$$P(B) = 1/125, \quad P(C) = 3/125$$

따라서 $P(A) = 1 - \{P(B) + P(C)\} = 1 - \frac{4}{125} = \frac{121}{125}$.

조건부확률

- 조건부확률(conditional probability): 한 사건이 일어났다는 조건하에 다른 사건이 일어날 확률
- 사건 A가 주어졌을 때, 사건 B가 일어날 조건부확률은 다음과 같이 정의된다.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad (P(A) > 0)$$

- 사건 A를 새로운 표본공간으로 간주

- 예) 학점이 좋은 학생의 취업 확률(조건='학점이 좋다')

조건부확률

Example (3.3)

두 개의 주사위를 던지는 경우 첫번째 던진 주사위의 눈이 두번째 주사위의 눈보다 클 때, 두 주사위의 눈의 합이 7일 확률을 구하여라.

sol) A: 첫번째 던진 주사위의 눈이 두번째 주사위의 눈보다 클 사건
B: 두 주사위의 눈의 합이 7일 사건

- 표본공간은 36개, 사건 A는 15개의 원소를 가지므로

$$P(A) = 15/36$$

- $B = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$이고, $A \cap B = \{(4,3),(5,2),(6,1)\}$이므로,

$$P(B) = 6/36, \quad P(A \cap B) = 3/36$$

- 따라서 $P(B|A) = \frac{P(A \cap B)}{P(A)} = 1/5$

곱셈법칙

- 곱셈법칙:
  $P(A) > 0,\ P(B) > 0$이면 다음이 성립한다.

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

Example (3.4)

불량품 20개와 양호품 80개로 구성된 로트에서 2개의 제품을 단순랜덤 추출할 때, 2개 모두 불량품일 확률을 구하여라.

sol) A: 첫번째 꺼낸 제품이 불량품일 사건
B: 두번째 꺼낸 제품이 불량품일 사건

- 구하는 확률은 $P(A \cap B) = P(B|A)P(A) = \frac{19}{99} \times \frac{20}{100} = \frac{19}{495}$.

베이즈 정리(Bayes' theorem)

- 조건확률을 구할 때, 조건 상황이 역으로 되어있는 확률을 이용하는 것

Theorem (베이즈 정리)

$A_i \cap A_j = \emptyset (i \neq i)$이고, $A_1 \cup A_2 \cup \cdots \cup A_n = S$일 때 $A_1, A_2, \cdots, A_n$으로 표본공간 S를 분할하면

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \\ = \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1) + \cdots + P(A_n)P(B|A_n)}$$

- 사전확률(prior probability):
  사건 B가 일어나기 전의 확률 $P(A_i)$
- 사후확률(posterior probability):
  사건 B가 일어난 후의 확률 $P(A_i|B)$

베이즈 정리 예제

Example

어떤 대학의 통계학과 학생의 30%는 1학년, 25%는 2학년, 25%는 3학년, 20%는 4학년 학생이라 하자. 그런데 1학년의 50%, 2학년의 30%, 3학년의 10%, 4학년의 2%가 수학 과목의 수강생이라 한다. 수학 과목을 수강하고 있는 학생 중 통계학과 1학년 학생의 비율을 구하여라.

sol) A_1: 뽑힌 학생이 1학년일 사건, A_2: 뽑힌 학생이 2학년일 사건
A_3: 뽑힌 학생이 3학년일 사건, A_4: 뽑힌 학생이 4학년일 사건
B: 뽑힌 학생이 수학 과목 수강생일 사건

- $P(A_1) = 0.3$, $P(A_2) = 0.25$, $P(A_3) = 0.25$, $P(A_4) = 0.2$
- $P(B|A_1) = 0.5$, $P(B|A_2) = 0.3$, $P(B|A_3) = 0.1$, $P(B|A_4) = 0.02$

- 사건 A_1, A_2, A_3, A_4는 서로 배반이므로,

$$\begin{aligned} P(B) &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) \\ &\quad + P(B|A_3)P(A_3) + P(B|A_4)P(A_4) \\ &= 0.5 \times 0.3 + 0.3 \times 025 + 0.1 \times 0.25 + 0.02 \times 0.2 \\ &= 0.254. \end{aligned}$$

- $P(B) = 0.254$이므로

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{P(A_1)P(B|A_1)}{P(B)} = \frac{0.3 \times 0.5}{0.254} = 0.59$$

- $P(A_1|B) > P(A_1)$

의료진단과 베이즈 규칙

Example

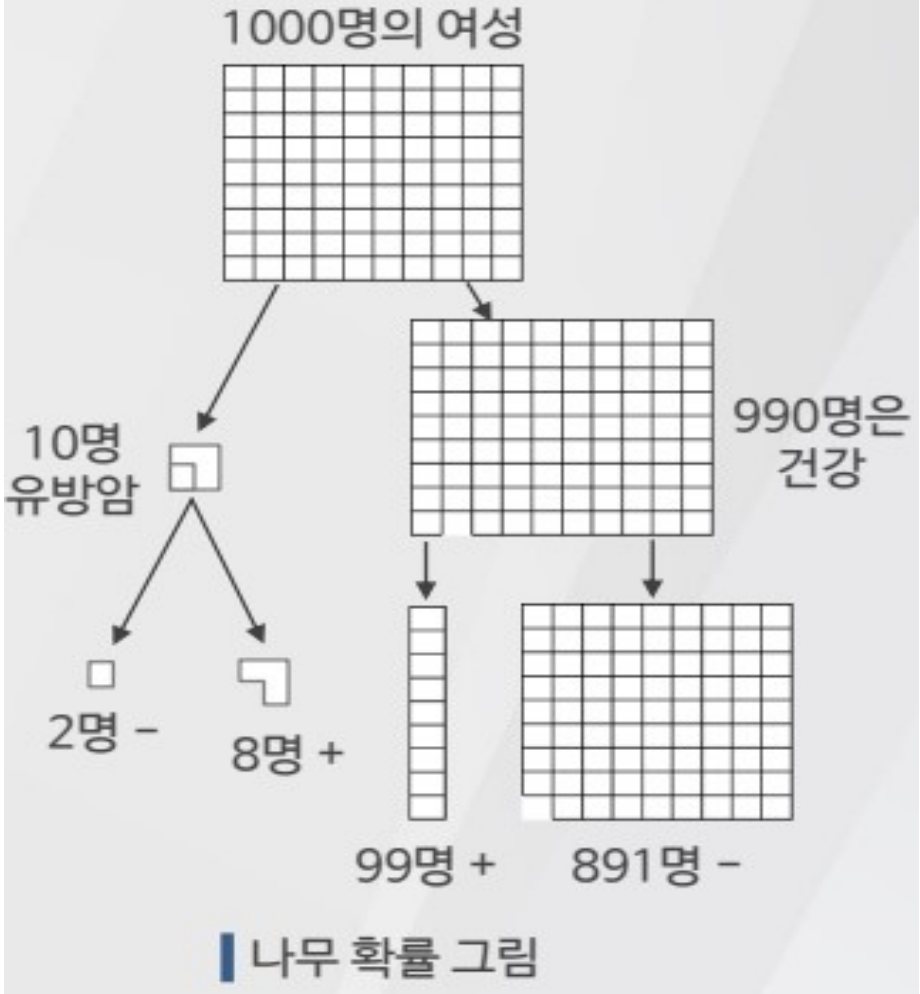
한 여성이 유방조영술 촬영 결과 양성판정을 받았다. 그 여성이 진짜로 유방암에 걸렸을 확률은 어떻게 될까? 유방암이 있는 환자가 유방조영술 검사를 통해 유방암 판정을 받을 확률(양성 반응)은 80%이고, 유방암이 걸리지 않은 여성이 음성 반응을 보일 확률은 90%, 그리고 검사를 받은 여성 중 1%에게 실제 암이 있다고 가정하자.

sol) C: 유방암에 걸린 경우, $+$: 검사 결과 양성반응, $-$: 음성반응

$P(C) = 0.01$, $P(+|C) = 0.8$, $P(-|C^C) = 0.9$

$$\begin{aligned} P(+) &= P(+|C)P(C) + P(+|C^C)P(C^C) \\ &= 0.8 \times 0.01 + 0.1 \times 0.99 = 0.107. \end{aligned}$$

$$P(C|+) = \frac{P(C \cap +)}{P(+)} = \frac{P(C)P(+|C)}{P(+)} = 0.075$$



- $P(C) = 0.01 \Rightarrow P(C|+) = 0.075$
- 암 발생률이 낮음(검사를 받는 여성 대다수가 건강함)
- $P(+|C)$와 $P(C|+)$ 개념 혼용

독립사건

- 독립사건: 사건 A가 일어날 사건이 사건 B가 일어날 확률에 아무런 영향을 미치지 않음

예) 동전이 앞, 뒤가 나오는 것과 주사위가 어떤 값이 나오는지는 서로 무관.

$$P("H" \cap "1") = P("H")P("1")$$

예) 주사위가 무엇이 나오는지 조건을 걸어도 동전이 앞, 뒤 나오는 것은 전혀 영향을 받지 않는다.

$$P("H"|"1") = P("H")$$

독립사건

Theorem (독립사건)

사건 A와 사건 B가 독립이면

$$P(A \cap B) = P(A)P(B),$$
$$P(A|B) = P(A), (P(B) > 0 \text{인 경우})$$
$$P(B|A) = P(B), (P(A) > 0 \text{인 경우})$$

이 성립한다.

- 사건 A와 사건 B는 서로독립(mutually independent)이라 한다.
- 독립이 아닌 두 사건을 서로종속(mutually dependent)이라고 한다.

- 상호배반과 독립은 다른 개념
  - 상호배반: $P(A \cap B) = 0$
  - 독립 : $P(A \cap B) = P(A)P(B)$

독립사건

Example (3.6)

Example 3.4에서 두 사건 A와 B는 서로 독립인가? 또한 단순랜덤복원추출할 경우, A와 B는 서로 독립인가?

sol) 전확률공식을 이용하면

$$\begin{aligned} P(B) &= P(B|A)P(A) + P(B|A^c)P(A^c) \\ &= \frac{19}{99} \times \frac{20}{100} + \frac{20}{99} \times \frac{80}{100} = \frac{1}{5}. \end{aligned}$$

따라서 $P(B) \neq P(B|A)$이므로 서로 독립이 아니다.

- 단순랜덤복원추출의 경우에는

$$P(B|A) = \frac{20}{100} = \frac{1}{5}$$

이므로 $P(B) = P(B|A)$가 되어 서로 독립이다.

확률변수

- 표본공간을 수의 집합에 대응시켜 생각해보자.

Example

동전 1개를 던져 나오는 결과를 관측하는 실험이 있다고 하자.

- 표본공간은 $S = \{'H', 'T'\}$
- 확률은 $P('H') = P('T') = \frac{1}{2}$
- 표본공간 S에서 다음과 같이 함수 X를 정의해보자.

$$X('H') = 1, \ X('T') = 0$$

- 확률을 대응시키면,

$$P(X = 1) = \frac{1}{2}, \ P(X = 0) = \frac{1}{2}$$

- 표본공간 S와 그 위에서 주어진 확률이 함수 X에 의해 수의 집합 $\{0, 1\}$과 그 위에서 주어진 확률에 대응되는 것.

확률변수

- 확률변수(random variable): 표본공간에서 정의된 실수값 함수

- 확률변수 -> X, Y 등 대문자로 표현
- 확률변수가 취하는 값 -> x,y 등 소문자로 표현

$$P(X = x)$$

- 확률변수의 의미: 모집단에서 하나의 특성값을 관측할때, 그 관측값은 어떤것을 관측하느냐에 따라 값이 변하므로 확률변수임.

확률분포

- 확률분포(probability distribution): 확률변수 X의 값에 따라 확률이 어떻게 흩어져 있는지를 합이 1인 양수로 나타낸 것을 X의 확률분포라 한다.
- 동전 1개를 던지는 실험에서 확률변수 X의 확률분포

표 3-1 X의 확률분포표

x	0	1
$P(X=x)$	$\frac{1}{2}$	$\frac{1}{2}$

이산확률변수와 확률질량함수

- 이산확률변수(random variable of discrete type): 취할 수 있는 모든 값이 $x_1, x_2, \cdots$로 셀 수 있는(countable) 값을 갖는 확률변수
- 확률변수 X가 x의 값을 취할 확률은 다음과 같다.

$$p(x) = \begin{cases} P(X = x_i), & when\ x = x_i (i = 1, 2, \cdots) \\ 0, & when\ x \neq x_i \end{cases} \tag{1}$$

- 이때, p(x)를 확률변수 X의 확률질량함수(probability mass function, p.m.f)라고 한다.
- 확률질량함수의 성질:

1. $0 \leq p(x) \leq 1,\ \sum_{all\ x} p(x) = 1$
2. $P(a < X \leq b) = \sum_{a<x\leq b} p(x)$

이산확률변수와 확률질량함수

Example

어느 인플루언서가 올린 게시물 12개 중 4개가 '좋아요' 1만 개를 넘었다고 하자. 게시물 중 3개를 단순랜덤추출했을 때, '좋아요'가 1만 개를 넘는 게시물의 수를 X라고 하자. 확률변수 X의 확률분포를 구하라.

sol) 게시물 12개 중 3개를 택하는 방법의 수는 $\binom{12}{3} = 220$이므로,

$$p(0) = P(X=0) = \binom{8}{3}\binom{4}{0}/\binom{12}{3} = \frac{56\times 1}{220} = \frac{14}{55}$$

$$p(1) = P(X=1) = \binom{8}{2}\binom{4}{1}/\binom{12}{3} = \frac{28\times 4}{220} = \frac{28}{55}$$

$$p(2) = P(X=2) = \binom{8}{1}\binom{2}{2}/\binom{12}{3} = \frac{8\times 6}{220} = \frac{12}{55}$$

$$p(3) = P(X=3) = \binom{8}{0}\binom{4}{3}/\binom{12}{3} = \frac{1\times 4}{220} = \frac{1}{55}$$

이를 정리하면 다음과 같은 X의 확률분포표를 얻는다.

x	0	1	2	3
p(x)	$\frac{14}{55}$	$\frac{28}{55}$	$\frac{12}{55}$	$\frac{1}{55}$

연속확률변수와 확률밀도함수

- 연속확률변수(random variable of continuous type): 취할 수 있는 모든 값이 하나씩 셀 수 없는 경우(uncountable), 즉 어떤 구간 내의 모든 값을 취할 수 있는 확률변수

- 연속확률변수의 성질

  1. 임의의 c에 대해 P(X=c)=0
  2. $P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b)$

- 연속확률변수 X에 대하여 함수 f(x)가 다음을 만족시킬 때,

  1. $f(x) \geq 0,\ \int_{-\infty}^{\infty} f(x)dx = 1$
  2. $P(a \leq X \leq b) = \int_a^b f(x)dx$ (단, $-\infty \leq a < b \leq \infty$)

  => f(x)를 X의 확률밀도함수(probability density function, p.d.f)라고 한다.

연속확률변수와 확률밀도함수

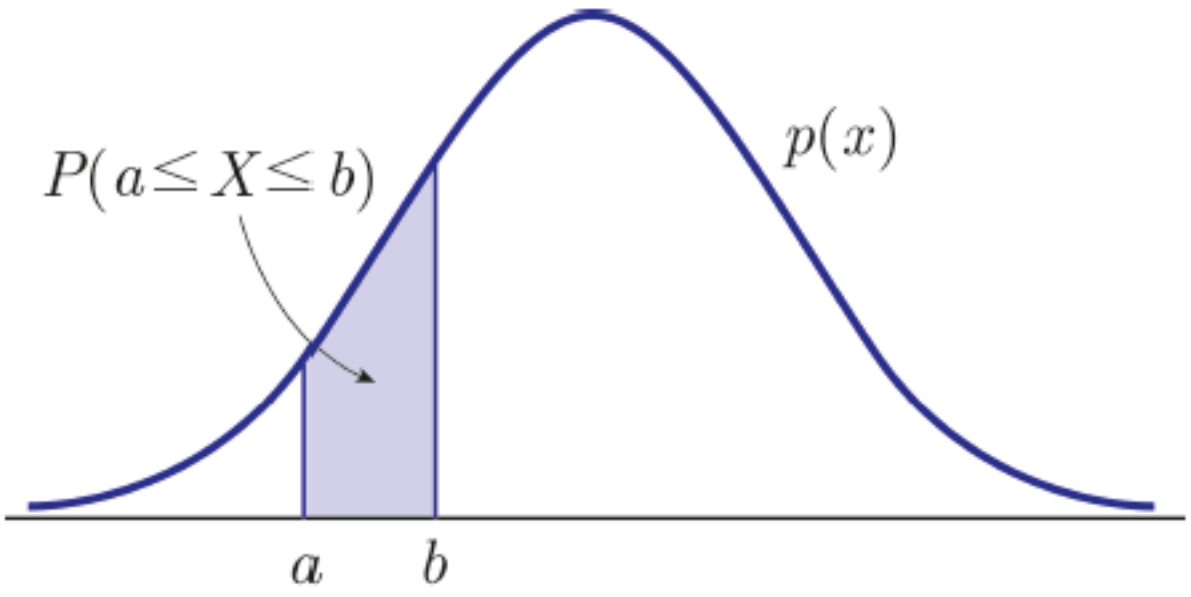


그림 3–5 연속확률변수의 확률밀도함수와 확률

- $P(a \leq X \leq b)$의 확률은:
  f(x)를 구간 [a,b]에서 적분한 것(빗금친 부분의 넓이)

연속확률변수와 확률밀도함수

Example

아래 그림과 같이 숫자판 중심에 한쪽 끝이 고정된 화살표가 자유롭게 회전할 수 있다고 하자. 이 화살표를 회전시켰을 때, 화살표가 저절로 멈추면서 가리킨 눈금을 X라고 하자. X는 0에서 1까지의 모든 값을 취할 수 있는 연속확률변수이다.

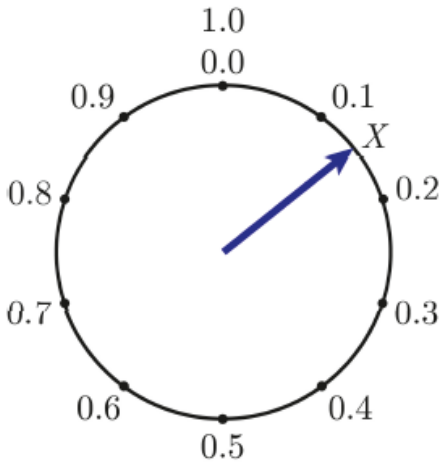


그림 3–3 연속확률변수의 예

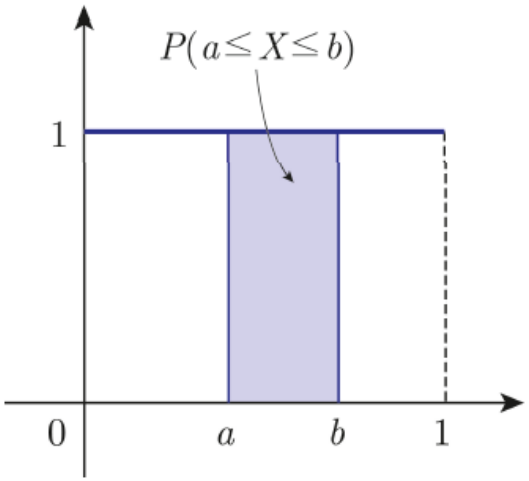


그림 3–4 확률밀도함수의 예

연속확률변수와 확률밀도함수

Example

연속확률변수 X의 확률밀도함수가

$$f(x) = \begin{cases} bx(1-x), & (when\ 0 \le x \le 1) \\ 0, & (when\ x < 0\ or\ x > 1) \end{cases}$$

로 주어질 때, 상수 b와 확률 $P(0 \le X \le 2/3)$를 구하여라.

sol) 확률밀도함수의 정의로부터 $\int_0^1 f(x)dx = 1$이여야 하므로,

$$b\int_0^1 x(1-x)dx = b\left[-\frac{1}{3}x^3 + \frac{1}{2}x^2\right]_0^1 = b\left[-\frac{1}{3} + \frac{1}{2}\right] = \frac{b}{6} = 1$$

따라서 b=6.

- $P(0 \le X \le 2/3) = \int_0^{2/3} 6x(1-x)dx = 20/27.$

확률분포의 대표값

- 중심위치 척도
  - 확률변수의 평균 또는 기대값
- 변동성(산포) 척도
  - 표준편차, 분산
- 연관성 척도
  - 공분산, 상관계수

- 이산확률변수 X와 Y가 각각 취하는 값이 $x_1, x_2, \cdots$와 $y_1, y_2, \cdots$이고 확률질량함수는 $p(x)$, $p(y)$라고 가정하자.
- 연속확률변수 X, Y에 대하여도 확률밀도함수를 각각 $f(x)$, $f(y)$라고 가정하자.

기대값(expected value)

- 확률변수 X의 평균(mean) 또는 기대값(expected value)은 다음과 같다.

$$E(X) = \mu = \begin{cases} \sum_{all\ x} xp(x), & (discrete\ case) \\ \int_{-\infty}^{\infty} xf(x)dx, & (continuous\ case) \end{cases}$$

- 확률의 가중값으로 계산되는 모든 가능한 결과들의 평균값.

Example

확률변수 X의 확률분포가 다음 표와 같을 때 X의 평균을 구하여라.

x	1	2	3	4
$p(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

sol) $E(X) = \sum_{x=1}^{4} xp(x) = 1 \times \frac{1}{8} + 2 \times \frac{3}{8} + 3 \times \frac{3}{8} + 4 \times \frac{1}{8} = 2.5$

기대값(expected value)

- 앞의 예제에서 평균 2.5는 X의 확률분포의 그래프에서 중심과 같은 의미를 가진다.

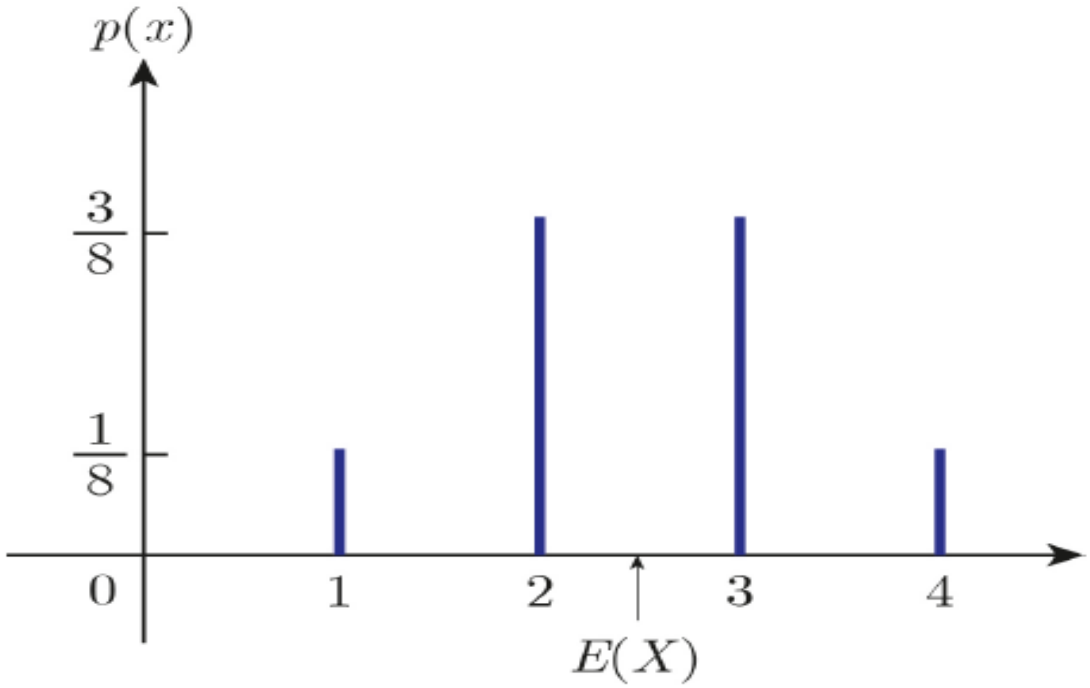


그림 3–6 평균의 역학적 의미

기대값(expected value)

Example

동전을 3회 던지는 실험에서 앞면의 개수를 X라고 할 때, 확률변수 X의 확률분포를 구하고, X의 기대값을 구하여라.

- sol) X의 확률질량함수, $p_X(x)$는 다음 표와 같다.

Table: X의 확률분포

x	0	1	2	3
$p_X(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

- 따라서 X의 기대값은

$$E(X) = \sum_{x=0}^{3} x p_X(x) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = \frac{3}{2}$$

기대값(expected value)

- 확률변수의 함수, $g(X)$도 확률 분포를 가지는 확률변수이다.
- $g(X)$의 기대값:

$$E[g(X)] = \begin{cases} \sum_{all\ x} g(x)p(x), & (discrete\ case) \\ \int_{-\infty}^{\infty} g(x)f(x)dx, & (continuous\ case) \end{cases}$$

- 확률변수 X의 함수, $g_1(X)$, $g_2(X)$에 대하여 다음이 성립한다. (c_1, c_2는 상수)

  1. $E[c_1 g_1(X)] = c_1 E[g_1(X)]$
  2. $E[c_1 g_1(X) + c_2 g_2(X)] = c_1 E[g_1(X)] + c_2 E[g_2(X)]$

기대값(expected value)

Example

위 예제에서(동전을 3회 던지는 실험) 확률변수 $Y = X(3-X)$의 확률분포를 구하고, Y의 기대값을 구하여라.

sol) Y의 확률질량함수, $p_Y(y)$는 다음 표와 같이 주어진다.

Table: $Y = X(3-X)$의 확률분포

y	0	2
$p_Y(y)$	$\frac{1}{4}$	$\frac{3}{4}$

따라서 $E(Y) = \sum y p_Y(y) = 0 \times \frac{1}{4} + 2 \times \frac{3}{4} = \frac{3}{2}$

- 또는 다음과 같이 직접 계산도 가능

$$E(Y) = E[X(3-X)] = \sum_{x=0}^{3} X(3-X)p_X(x)$$
$$= 0(3-0)p_X(0) + \cdots + 3(3-3)p_X(3) = \frac{3}{2}$$

기대값(expected value)

Example

어느 버스 정류장에서 매시 0분, 15분, 35분에 버스가 각각 1회씩 출발한다. 한 사람이 우연히 이 정거장에 도착했을 때, 버스 출발까지 기다려야 하는 시간의 기댓값을 구하라.

sol) 정류장에 도착한 시각을 X분이라 하면, X의 확률밀도함수는

$$f(x) = \frac{1}{60},\ 0 \le x \le 60$$

- 이 사람이 x분에 왔을 때 기다리는 시간 g(x)는

$$g(x) = \begin{cases} 15 - x, & 0 < x \le 15 \\ 35 - x, & 15 < x \le 35 \\ 60 - x, & 35 < x \le 60 \end{cases} \tag{2}$$

- 따라서 g(X)의 기대값, E(g(X))은

$$\begin{aligned}
&\int g(x)f(x)dx \\
&= \int_0^{15}(15-x)\frac{1}{60}dx + \int_{15}^{35}(35-x)\frac{1}{60}dx + \int_{35}^{60}(60-x)\frac{1}{60}dx \\
&= \frac{125}{10}
\end{aligned}$$

- 즉, 기다리는 시간의 기대값은 10분 25초이다.

분산과 표준편차

- 확률변수의 분산(variance)과 표준편차(standard deviation):
  - 산포를 나타내는 대표값
  - X의 평균이 μ이고 확률질량/밀도함수가 p(x), f(x)일 때,

$$(1)\ Var(X) = E[(X-\mu)^2] = \begin{cases} \sum_{all\ x}(x-\mu)^2 p(x), & (discrete\ r.v) \\ \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx, & (continuous\ r.v) \end{cases}$$

$$(2)\ sd(X) = \sqrt{Var(X)} = \sqrt{E[(X-\mu)^2]}$$

- Var(X)=σ^2, sd(X)=σ

분산과 표준편차

- 분산의 간편계산법: $Var(X) = E(X^2) - [E(X)]^2$

$$
\begin{aligned}
pf)\ Var(X) &= E[(X-\mu)^2] \\
&= E(X^2 - 2\mu X + \mu^2) \\
&= E(X^2) - 2\mu E(X) + \mu^2 \\
&= E(X^2) - \mu^2
\end{aligned}
$$

분산과 표준편차

Example

동전을 2회 던질 때 나오는 표면의 개수 X의 평균과 표준편차를 구하여라.

sol) X의 확률분포는 다음 표와 같다.

x	0	1	2
$p_X(x)$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

- 따라서 X의 기대값과 표준편차는

$$E(X) = \sum_{x=0}^{2} x p_X(x) = 0 \times \frac{1}{4} + 1 \times \frac{2}{4} + 2 \times \frac{1}{4} = 1$$

$$E(X^2) = \sum_{x=0}^{2} x^2 p_X(x) = 0^2 \times \frac{1}{4} + 1^2 \times \frac{2}{4} + 2^2 \times \frac{1}{4} = \frac{3}{2}$$

$$Var(X) = \frac{3}{2} - 1^2, \ sd(X) = \sqrt{Var(X)} = 1/\sqrt{2}$$

기대값과 분산의 성질

- *for any* $a, b \in \mathbb{R}$,
  aX+b의 기대값과 분산은 다음과 같다.

- E(aX+b)=aE(X)+b

  pf) $E(aX+b) = \sum(ax+b)p(x) = a\sum xp(x) + b\sum p(x) = a\mu + b$

- $Var(aX+b) = a^2 Var(X)$, $sd(aX+b) = |a|sd(X)$

$$
\begin{aligned}
pf)\ Var(aX+b) &= E[(aX+b-E(aX+b))^2] \\
&= E(aX+b-a\mu-b)^2 \\
&= a^2 E(X-\mu)^2 \\
&= a^2 Var(X)
\end{aligned}
$$

확률변수의 표준화(standardization)

- 표준화: 확률변수 X에 대하여 그 평균 μ를 뺀 다음 표준편차로 나누는 것.

  - 확률변수 X를 표준화시킨 확률변수 Z는 다음과 같다.

$$Z = \frac{X - E(X)}{sd(X)}$$

  - 평균이나 분산에 의존하지 않고도 데이터의 상대적인 위치 관계를 알 수 있는 지표이다.

- 이때 Z의 평균은 0, 분산은 1이다.

  - $E(Z) = E\left(\frac{X-E(X)}{sd(X)}\right) = E\left(\frac{1}{\sigma}X - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma}E(X) - \frac{\mu}{\sigma} = 0$

  - $Var(Z) = Var\left(\frac{1}{\sigma}X - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma^2}Var(X) = 1$

두 확률변수의 결합분포

- 두 개 이상의 확률변수를 동시에 고려할 때 이들 사이의 확률적 관계를 나타내는 방법에 대해 학습해 보자.

예) - 아버지의 키와 아들의 키

- 환자의 건강기록을 구성하는 변수로 나이, 혈압, 몸무게 등 고려

Table: 사건 A와 B가 발생할 확률

	B_1	B_2	total
A_1	$P(A_1 \cap B_1)$		$P(A_1)$
A_2		$P(A_2 \cap B_2)$	
	$P(B_1)$		1

- 결합확률(joint probability): 교집합(intersection)의 확률
- 주변확률(marginal probability): 어느 한쪽만의 변수를 생각할 때의 확률

두 확률변수의 결합분포

- 결합확률분포(joint probability distribution):

  두 확률변수 X와 Y에 대해, 이들이 취할 수 있는 값들의 모든 순서쌍에 확률이 어떻게 분포되어 있는지를 합이 1인 양수로 나타낸 것.

Table: 동전(X)와 주사위(Y)를 동시에 던질 때의 예

	1	2	3	4	5	6	합계
앞면	1/12	1/12	1/12	1/12	1/12	1/12	1/2
뒷면	1/12	1/12	1/12	1/12	1/12	1/12	1/2
	1/6	1/6	1/6	1/6	1/6	1/6	1

- P(X=1, Y=2)=
- P(Y=1)=

두 이산형 확률변수의 결합분포

- 이산확률변수 (X,Y)가 순서짝 (x,y)를 취할 확률은 다음과 같고, 이때 $p(x,y)$를 X, Y의 결합확률질량함수(joint p.m.f)라고 한다.

$$p(x,y) = P(X = x_i, Y = y_j), \ for \ (x,y) = (x_i, y_j), (i,j = 1,2,\cdots)$$

- 주변확률질량함수(marginal p.m.f)

  다음 식은 각각 X와 Y의 주변확률질량함수이고, X, Y의 분포를 결정해 준다.

$$p_1(x) = \sum_{all \ y}^{\infty} p(x,y), \quad p_2(y) = \sum_{all \ x}^{\infty} p(x,y)$$

- 결합확률질량함수의 성질:

  (1) $0 \le p(x,y) \le 1, \ \sum_{all \ x} \sum_{all \ y} p(x,y) = 1$

  (2) $P(a < X \le b, \ c < Y \le d) = \sum_{a<X\le b} \sum_{c<Y\le d} p(x,y)$

두 연속형 확률변수의 결합분포

- 연속형확률변수 X,Y의 결합확률밀도함수(joint p.d.f)는 어떤 2차원 집합 A에 대하여 다음을 만족한다.

$$P\{(X,Y) \in A\} = \int\int_A f(x,y)dxdy$$

- X와 Y 각각의 분포를 나타내는 주변확률밀도함수(marginal p.d.f)는 다음과 같다.

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)dy, \quad -\infty < x < \infty$$

$$f_2(y) = \int_{-\infty}^{\infty} f(x,y)dx, \quad -\infty < y < \infty$$

- 결합확률밀도함수의 성질:
  (1) $0 \le f(x,y)$, for all x and y, $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$
  (2) A가 $(x,y) : a \le x \le b,\ c \le y \le d$라면,
  $P\{(X,Y) \in A\} = \int_c^d \int_a^b f(x,y)dxdy$

두 확률변수의 결합분포 예제

Example

서로 다른 동전 A, B, C를 동시에 던지는 실험에서 확률변수 X, Y, Z를 다음과 같이 정의하자.

$$X = \begin{cases} 1, & \textit{when } A \textit{ is Heads} \\ 0, & \textit{otherwise} \end{cases}$$

$$Y = \begin{cases} 1, & \textit{when } A, B \textit{ are Heads} \\ 0, & \textit{otherwise} \end{cases}$$

$$Z = \begin{cases} 1, & \textit{when } B, C \textit{ are Heads} \\ 0, & \textit{otherwise} \end{cases}$$

X와 Y, X와 Z의 결합확률분포를 구하여라.

sol) 표본공간과 그 원소에 대응하는 확률변수 X, Y, Z의 값들은 다음과 같다.

표본공간	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
X	1	1	1	1	0	0	0	0
Y	1	1	0	0	0	0	0	0
Z	1	0	0	0	1	0	0	0

- 위의 표로부터 X와 Y, X와 Z의 결합확률분포는 다음과 같다.

(1) X와 Y

	x=0	x=1	Sum
y=0	2/4	1/4	3/4
y=1	0	1/4	1/4
Sum	2/4	2/4	1

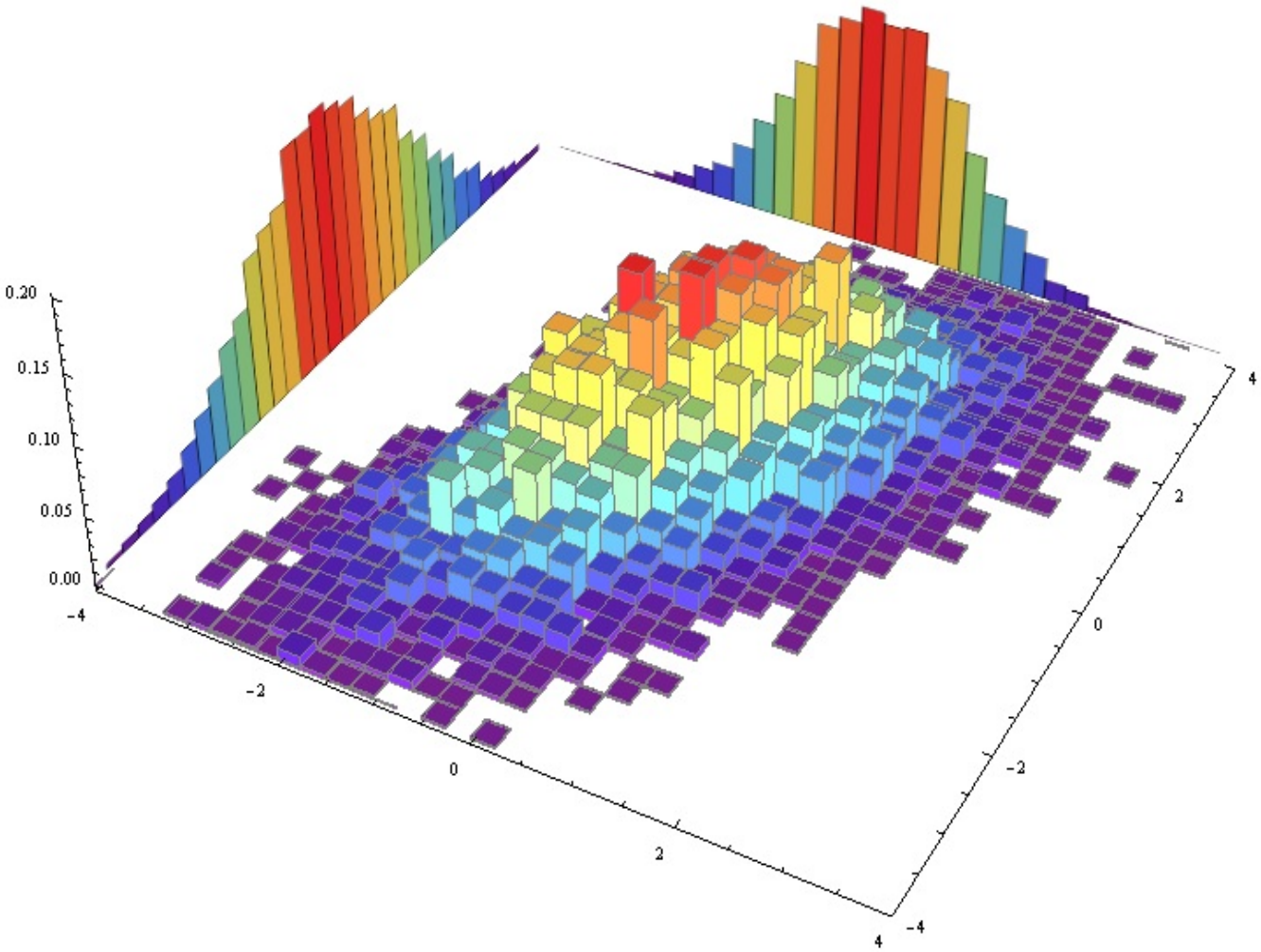
(2) X와 Z

	x=0	x=1	Sum
z=0	3/8	3/8	6/8
z=1	1/8	1/8	2/8
Sum	4/8	4/8	1

- 각 표에서 중간부분은 결합확률질량함수
열의 합, 행의 합은 주변확률질량함수를 나타낸다.

두 확률변수의 결합분포

- Joint probability distribution of two random variables.



두 확률변수의 결합분포

- 두 확률변수의 함수도 자신의 확률분포를 가지는 확률변수이다.
- 두 확률변수의 함수, $g(X, Y)$의 기대값:

$$E[g(X,Y)] = \begin{cases} \sum_{all\ x}\sum_{all\ y} g(x,y)p(x,y), & (discrete\ case) \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy, & (continuous\ case) \end{cases}$$

- X와 Y의 함수 g_1, g_2에 대하여 다음의 사실들이 성립한다.
  (c_1, c_2는 상수)
  $E[c_1 g_1(X,Y) + c_2 g_2(X,Y)] = c_1 E[g_1(X,Y)] + c_2 E[g_2(X,Y)]$
- E(X+Y)=E(X)+E(Y)

두 확률변수의 결합분포

Example

하나의 주사위를 던져서 나오는 눈의 종류에 따라 상금이 걸린 게임이 있다. A, B 두 사람이 각각 다음과 같은 게임을 택했다고 하자.

A: 1 또는 2 가 나오면 100원
3 또는 4 가 나오면 200원
5 또는 6 이 나오면 300원
B: 짝수가 나오면 100원
홀수가 나오면 (눈의 수X100)원

이 경우에 A와 B의 수입의 합 Z에 대한 분포를 구하여라.

sol) A의 수입과 B의 수입을 각각 X와 Y로 나타내면, X와 Y의 결합확률분포는 다음과 같다.

Table: X와 Y의 결합확률분포

x	y: 100	300	500	행의 합
100	2/6	0	0	2/6
200	1/6	1/6	0	2/6
300	1/6	0	1/6	2/6
열의 합	4/6	1/6	1/6	1

- $P(X = 100, Y = 100) = P(\text{"1"} \ or \ \text{"2"}) = 2/6$
- $E(X) = 100 \times 2/6 + \cdots = 200$, $E(Y) = 100 \times 4/6 + \cdots = 200$

- $Z = X + Y$라 하면 Z의 확률질량함수 $p(z)$는 다음과 같다.

Table: $Z = X + Y$의 확률분포

z	200	300	400	500	800	합계
$p(z)$	2/6	1/6	1/6	1/6	1/6	1

- $E(Z) = 200 \times 2/6 + \cdots = 400 = E(X) + E(Y)$

독립성

- 두 확률변수 X와 Y가 서로 독립이라는 것은 모든 x와 y에대하여 $\{X = x\}$인 사건과 $\{Y = y\}$인 사건이 서로 독립(mutually independent)임을 뜻한다.

Theorem

두 이산형 확률변수 X와 Y의 결합확률함수를 $p(x, y)$라 하고, X와 Y의 주변확률함수를 각각 $p_1(x)$, $p_2(y)$라 하자. 이 때,

$$p(x, y) = p_1(x)p_2(y)$$

가 성립할 때, X와 Y는 서로 독립이라고 한다.

- 두 연속형 확률변수 X와 Y의 결합확률질량함수를 $f(x, y)$, X와 Y의 주변확률질량함수를 각각 $f_1(x)$, $f_2(y)$라 하면 다음이 성립할 때 서로 독립이라고 한다. $f(x, y) = f_1(x)f_2(y)$

독립성 예제

Example

하나의 동전을 세 번 던질 때, X=처음 두 번에서 나오는 앞면의 개수, Y=세 번째에 나오는 앞면의 개수 라고 하면 X와 Y는 서로 독립인가?

sol) 먼저 X와 Y의 결합확률분포를 구해보자.

	y=0	y=1	행의 합
x=0	1/8	1/8	1/4
x=1	2/8	2/8	1/2
x=2	1/8	1/8	1/4
열의 합	1/2	1/2	1

- 위의 표에서 모든 (x,y)에 대해서 $p(x,y) = p_1(x)p_2(y)$가 성립함을 알 수 있다. 따라서 X와 Y는 서로 독립이다.

$$p_1(0)p_2(0) = \frac{1}{4} \times \frac{1}{2} = \frac{1}{8} = p(0,0), \cdots, p_1(2)p_2(1) = \frac{1}{4} \times \frac{1}{2} = p(2,1)$$

공분산과 상관계수

- 공분산(covariance): 두 확률변수 X와 Y의 연관관계의 척도

$$\mathrm{Cov}(X, Y) = E[(X - \mu_1)(Y - \mu_2)]$$

- 공분산의 간편계산법: $\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y)$

$$\begin{aligned} pf)\ \mathrm{Cov}(X, Y) &= E[(X - \mu_1)(Y - \mu_2)] \\ &= E(XY - \mu_2 X - \mu_1 Y + \mu_1\mu_2) \\ &= E(XY) - \mu_2 E(X) - \mu_1 E(Y) + \mu_1\mu_2 \\ &= E(XY) - \mu_1\mu_2. \end{aligned}$$

- Var(X)=Cov(X,X)

공분산과 상관계수

- 상관계수(correlation coefficient): 두 확률변수의 직선관계가 얼마나 강하고 어떤 방향인지를 나타냄 (방향과 정도)
- 확률변수 X, Y를 표준화한 것의 곱의 기대값
  (or) 공분산을 σ_1, σ_2로 나눈 것

$$\mathrm{Corr}(X, Y) = E\left[\left(\frac{X-\mu_1}{\sigma_1}\right)\left(\frac{Y-\mu_2}{\sigma_2}\right)\right] = \frac{\mathrm{Cov}(X, Y)}{\mathrm{sd}(X)\mathrm{sd}(Y)}$$

- $-1 \leq \mathrm{Corr}(X.Y) \leq 1$
- X와 Y가 완전한 직선 관계를 가지는 경우: $Y = aX + b$,
  ($a \neq 0, b$는 상수)

$$\begin{cases} \mathrm{Corr}(X, Y) = 1, & (a > 0) \\ \mathrm{Corr}(X, Y) = -1, & (a < 0) \end{cases} \tag{3}$$

공분산과 상관계수

Example

X와 Y의 결합분포가 다음과 같이 주어질 때, X와 Y의 공분산과 상관계수를 구하여라.

표 3-14 X와 Y의 결합확률분포표

x \ y	0	1	2	3	행의 합
0	0.05	0.05	0.10	0.00	0.20
1	0.05	0.10	0.25	0.10	0.50
2	0.00	0.15	0.10	0.05	0.30
열의 합	0.10	0.30	0.45	0.15	1.00

sol) X와 Y의 평균과 분산을 먼저 구해보자.

$$\mu_1 = E(X) = \sum x p_1(x) = 0 \times 0.2 + 1 \times 0.5 + 2 \times 0.3 = 1.1$$

$$\mu_2 = E(Y) = \sum y p_2(y) = 0 \times 0.1 + 1 \times 0.3 + 2 \times 0.45 + 3 \times 0.15 = 1.65$$

$E(X^2) = \sum x^2 p_1(x) = 0^2 \times 0.2 + 1^2 \times 0.5 + 2^2 \times 0.3 = 1.7$

$E(Y^2) = \sum y^2 p_2(y) = 0^2 \times 0.1 + 1^2 \times 0.3 + 2^2 \times 0.45 + 3^2 \times 0.15 = 3.45$

$\sigma_1^2 = 1.7 - 1.1^2 = 0.49, \quad \sigma = 0.7$

$\sigma_2^2 = 3.45 - 1.65^2 = 0.7275, \quad \sigma = 0.85$

- 공분산과 상관계수를 구해보자.

Table: E(XY)의 계산

	y=0	y=1	y=2	y=3
x=0	0.05	0.05	0.1	0
	(0)	(0)	(0)	(0)
x=1	0.05	0.1	0.25	0.1
	(0)	(1)	(2)	(3)
x=2	0	0.15	0.1	0.05
	(0)	(2)	(4)	(6)

- $E(XY) = \sum xyp(x, y) = 0 \times 0.05 + \cdots + 6 \times 0.05 = 1.9$
- $\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y) = 1.9 - 1.65 \times 1.1 = 0.085$
- $\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_1 \sigma_2} = \frac{0.085}{0.7 \times 0.85} = 0.14$

공분산과 상관계수의 성질

- $\mathrm{Cov}(aX+b,\ cY+d) = ac\mathrm{Cov}(X,Y)$

$$
\begin{aligned}
pf)\ \ \mathrm{Cov}(aX+b,\ cY+d) & \\
&= E[(aX+b-a\mu_1-b)(cY+d-c\mu_2-d)] \\
&= E[ac(X-\mu_1)(Y-\mu_2)] \\
&= ac\mathrm{Cov}(X,Y).
\end{aligned}
$$

- $\mathrm{Corr}(aX+b,\ cY+d) =$

$$
\begin{aligned}
pf)\ \ \mathrm{Corr}(aX+b,\ cY+d) & \\
&= \frac{\mathrm{Cov}(aX+b,\ cY+d)}{\sqrt{\mathrm{Var}(aX+b)}\sqrt{\mathrm{Var}(cY+d)}} \\
&= \frac{ac\mathrm{Cov}(X,\ Y)}{\sqrt{a^2\mathrm{Var}(X)}\sqrt{c^2\mathrm{Var}(Y)}} \\
&= \frac{ac}{|ac|}\mathrm{Corr}(X,Y).
\end{aligned}
$$

확률변수의 결합분포

- 확률변수 X와 Y의 선형결합: $aX \pm bY$

$$E(aX \pm bY) = aE(X) \pm bE(Y)$$

$$Var(aX \pm bY) = a^2 Var(X) + b^2 Var(Y) \pm 2abCov(X, Y)$$

- 일반적으로, 확률변수 $X_1, \cdots, X_n$와 상수 $c_1, \cdots, c_n$에 대하여,

$$E\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} [c_i \cdot E(X_i)]$$

$$Var\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)$$

확률변수의 결합분포

- Var(X+Y)=Var(X)+Var(Y)+2Cov(X,Y)

$$
\begin{aligned}
pf)\quad Var(X+Y) &= E[X+Y-E(X+Y)]^2 \\
&= E[X+Y-E(X)-E(Y)]^2 \\
&= E[(X-E(X))+(Y-E(Y))]^2 \\
&= E[(X-E(X))]^2 + E[(Y-E(Y))]^2 \\
&\qquad + 2E[(X-E(X))(Y-E(Y))] \\
&= Var(X) + Var(Y) + 2\mathrm{Cov}(X,Y).
\end{aligned}
$$

상관계수 범위 증명

- $-1 \leq \mathrm{Corr}(X.Y) \leq 1$

pf) $\left(\frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y}\right)$의 분산을 고려해 보자.

$$
\begin{aligned}
Var\left(\frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y}\right) &= Var\left(\frac{X}{\sigma_X}\right) + Var\left(\frac{Y}{\sigma_Y}\right) \pm 2\mathrm{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) \\
&= \frac{1}{\sigma_X^2} Var(X) + \frac{1}{\sigma_Y^2} Var(Y) \pm 2\frac{1}{\sigma_X \sigma_Y}\mathrm{Cov}(X, Y) \\
&= \frac{1}{\sigma_X^2}\sigma_X^2 + \frac{1}{\sigma_Y^2}\sigma_Y^2 \pm 2\mathrm{Corr}(X, Y)
\end{aligned}
$$

이때, $0 \leq 1 + 1 \pm 2\mathrm{Corr}(X, Y)$ 이므로 다음이 성립한다.

$$-1 \leq \mathrm{Corr}(X.Y) \leq 1$$

두 확률변수의 독립성

- 두 확률변수가 서로 독립인 경우:
  (1) E(XY)=E(X)E(Y)

$$
\begin{aligned}
pf)\quad E(XY) &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} x_i y_j p(x_i, y_j) \\
&= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} x_i y_j p_1(x_i) p_2(y_j) \\
&= \sum_{i=1}^{\infty} x_i p_1(x_i) \sum_{j=1}^{\infty} y_j p_2(y_j) \\
&= E(X)E(Y).
\end{aligned}
$$

  (2) Cov(X,Y)=0, Corr(X,Y)=0

$$pf)\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y) = 0$$

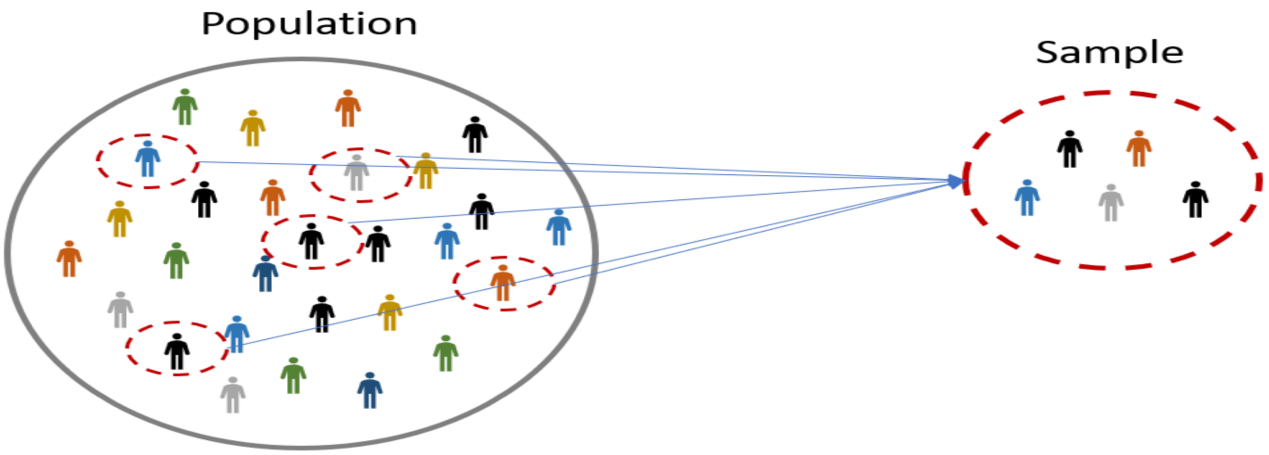
  (3) $Var(X \pm Y) = Var(X) + Var(Y)$

4장. 표본분포

박명현

서울대학교 학부대학

표본분포(Sampling distribution)



- 표본의 추출과정을 알아야 모집단 추론 가능
  - 시간과 공간의 제약, 사용 가능한 비용, 조사원 등 경제적 제약
- 모집단을 추론하기 위해서 표본평균이나 표본분산 등의 값과 오차를 계산하려면 표본통계량의 분포가 필요(표본분포)

모집단과 표본

모집단(population)

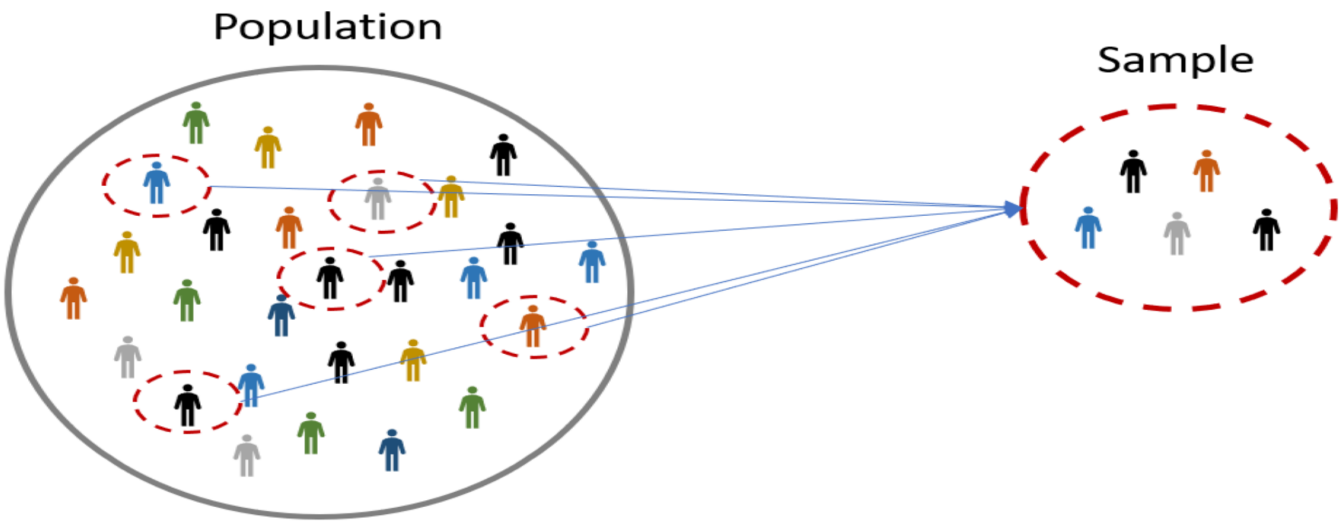
- 모수(parameter): 모집단의 특징을 나타내는 대표값
  예) 모평균 μ, 모비율 p, 모분산 σ^2

표본(sample)

- 통계량(statistic): 표본의 관측값들로부터 계산되는 수치함수
- 추정량(estimator): 모수의 추정을 위해 사용되는 통계량
  - 주로 대문자로 표시. 예) 표본평균 $\bar{X}$, 표본분산 S^2
  - 'hat'을 붙여 표기. 예) $\hat{\mu} = \bar{X}$, $\hat{\sigma}^2 = S^2$, $\hat{p}$
- 추정치(estimate): 관측치로부터 계산된 모수의 예측값(관찰값).
  - 소문자로 표시. 예) $\bar{x}$, s^2, $\hat{p}$

예) 모평균 μ 추정시 표본 $X_1, \cdots, X_n$의 표본평균 $\bar{X}$는 추정량이고, 실제 자료 $x_1, \cdots, x_n$의 표본평균 $\bar{x}$는 추정값이 된다.

모집단과 표본



- 추정량을 하나의 [ ]로 취급
  - 어떤 표본이 추출되느냐에 따라서 하나의 통계량이 여러 가지 다른 값으로 실현될 수 있다.
  - 하지만 하나의 표본이 추출되고 나면 이제 통계량은 하나의 값으로 실현된다.(추정치)

단순랜덤추출법(simple random sampling)

- 단순랜덤복원추출(sampling with replacement):
  - 한 번 선택한 추출단위는 다시 모집단에 되돌려놓고 다음 단위를 추출
  - 크기가 N인 유한모집단에서 크기 n인 표본을 복원추출하는 경우에는 매번 동일하게 $1/N$의 학률로 1개씩 선택
  - 무한모집단의 랜덤표본을 정의하기 위한 개념적 도구로 이용됨.
- 단순랜덤비복원추출(sampling without replacement):
  - 선택한 추출단위를 다시 모집단에 되돌려 놓지 않고 다음 단위를 뽑는 방식
  - 각 항목은 $1/N, 1/(N-1), \cdots, 1/(N-n+1)$의 확률로 1개씩 선택
  - 일반적으로, 모집단의 크기가 표본의 크기에 비하여 충분히 클 때 유한모집단에서 표본을 단순랜덤비복원추출하는 것은 단순랜덤복원추출하는 것과 별 차이가 없다.

단순랜덤복원추출

Example

N개의 제비 중 M개의 당첨제비(Success)가 있을 때 2개의 제비를 단순랜덤복원추출하는 경우를 생각해 보자.

- X_i: $i$th번째 시행의 결과를 나타내는 베르누이확률변수, $i = 1, 2$

$$X_i = \begin{cases} 1, & if\ Success \\ 0, & if\ Fail \end{cases}$$

- '독립적':

$$P(X_2 = 1) = P(X_2 = 1|X_1 = 1) = P(X_2 = 1|X_1 = 0) = M/N$$

$$P(X_2 = 0) = P(X_2 = 0|X_1 = 1) = P(X_2 = 0|X_1 = 0) = (N - M)/N$$

- '동일한 분포': $P(X_1 = 1) = P(X_2 = 1) = M/N$

=> 두 확률변수 X_1, X_2는 서로 독립이며 동일한 분포를 가진다.

단순랜덤비복원추출

Example

N개의 제비 중 M개의 당첨제비(Success)가 있을 때 2개의 제비를 단순랜덤비복원추출하는 경우를 생각해 보자.

- '독립적':

  $P(X_1 = 1) = M/N$, $P(X_2 = 1|X_1 = 1) = (M-1)/(N-1)$,
  $P(X_2 = 1|X_1 = 0) = M/(N-1)$ 이므로

$$\begin{aligned} P(X_2 = 1) &= P(X_2 = 1 \cap X_1 = 1) + P(X_2 = 1 \cap X_1 = 0) \\ &= P(X_2 = 1|X_1 = 1)P(X_1 = 1) + \\ &\qquad P(X_2 = 1|X_1 = 0)P(X_1 = 0) \\ &= \frac{M-1}{N-1}\frac{M}{N} + \frac{M}{N-1}\frac{N-M}{N} = \frac{M}{N}. \end{aligned}$$

  $P(X_2 = 1) \neq P(X_2 = 1|X_1 = 1)$이므로 두 시행은 독립적이지 않음

- '동일한 분포':

  $P(X_1 = 1) = P(X_2 = 1) = M/N$

  => 두 확률변수 X_1, X_2는 동일한 분포를 가지지만 서로 독립은 아니다.

- if $N \gg 2$ 이면,

  (모집단의 크기가 표본의 크기에 비하여 충분히 클 때)

  $P(X_2 = 1) \simeq P(X_2 = 1|X_1 = 1)$이 되어, X_1, X_2는 독립에 가깝게 됨.

- 즉, 유한모집단에서의 랜덤표본 $\{X_1, X_2, \cdots, X_n\}$의 성질은 단순랜덤복원추출에 의한 표본의 성질과 유사함을 알 수 있다.

랜덤표본(random sample)

- 랜덤표본

  - X_i: i번째로 뽑힌 추출단위의 특성값을 나타내는 확률변수

  $$i = 1, \cdots, n$$

  (1) 유한모집단인 경우 단순랜덤비복원추출로 뽑은 표본을 랜덤표본이라고 한다. 이 때, 크기 n인 랜덤표본을 $\{X_1, X_2, \cdots, X_n\}$로 나타낸다.

  (2) 무한모집단인 경우 서로 독립이며, 모집단 분포와 동일한 분포를 갖는 확률변수들의 집합 $\{X_1, X_2, \cdots, X_n\}$를 랜덤표본이라 한다. 즉, 모집단의 분포를 $f(x)$라 할 때 다음과 같이 표기한다.

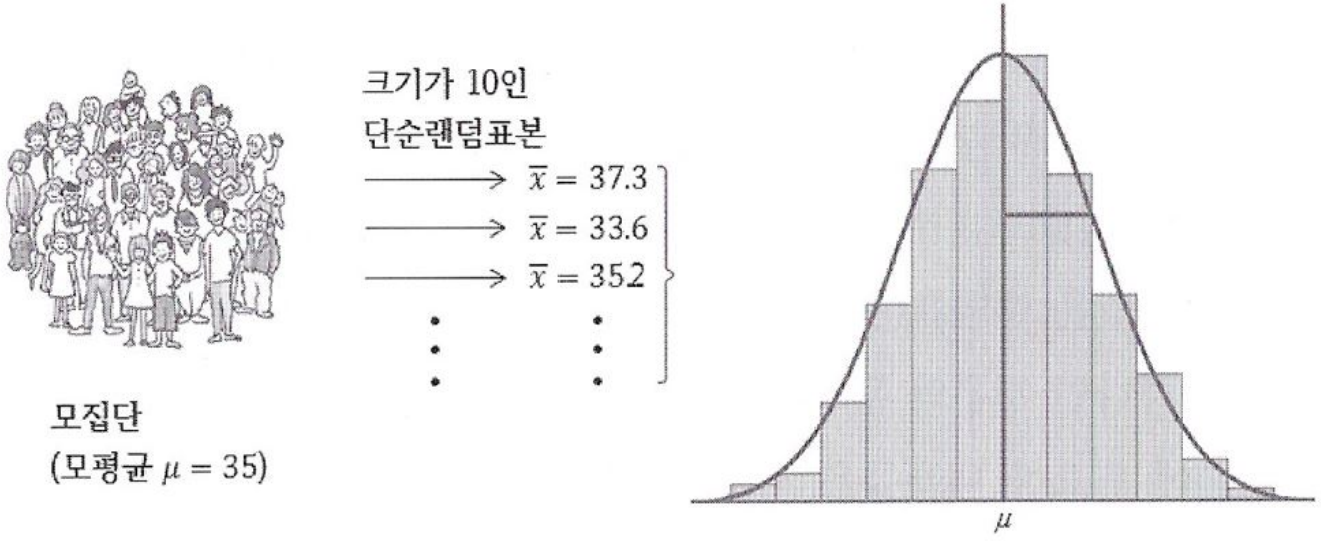
  $$X_i \sim_{iid} f(x), \ (i.i.d : \textit{independent and identically distributed})$$

- 5장 이후에서는 무한모집단에서의 랜덤표본을 가정

표본분포(sampling distribution)

- 표본분포: 통계량의 확률분포

Figure: 모집단으로 구한 표본평균 $\bar{X}$의 표본분포



- 평균이 $\mu = 35$인 모집단에서 $n = 10$인 단순랜덤표본을 추출
- 각 표본에서 사람들 나이의 평균 $\bar{X}$을 계산
- 표본에 따라 그 평균값이 다르게 나타남
- 계산된 표본평균들로 히스토그램을 그려보면 $\bar{X}$은 μ를 중심으로 분포하고 있음을 알 수 있음.

표본분포

Example

모집단 $\{2, 2, 4, 5\}$에서 크기 2인 표본을 단순랜덤비복원 추출하였을 때, 표본평균의 표본분포를 구하여라.

Table: 가능한 표본과 표본평균의 값

가능한 표본	$\{2,2\}$	$\{2,4\}$	$\{2,5\}$	$\{2,4\}$	$\{2,5\}$	$\{4,5\}$
표본평균	2	3	3.5	3	3.5	4.5
표본이 뽑힐 확률	1/6	1/6	1/6	1/6	1/6	1/6

Table: 표본평균 $\bar{X}$의 표본분포

표본 평균의 값	2	3	3.5	4.5	합
확률	1/6	1/3	1/3	1/6	1

- 모수: $\mu = 3.25$, 추정량: $\hat{\mu} = \bar{X}$, 추정치: $\bar{x} = 2, 3, 3.5, 4.5$

표본분포

Example

{찬성,반대,찬성,찬성,반대}와 같은 모집단에서 크기 3인 표본을 단순 랜덤비복원추출로 뽑아 찬성의 모비율 p(이 경우, $p = 0.6$)를 추측하는 경우 표본비율의 표본분포를 구하여라.

- 가능한 표본: {찬성,찬성,찬성},{찬성,찬성,반대},{찬성,반대,반대}

Table: 표본비율 $\hat{p}$의 분포표

표본비율($\hat{p}$)	1	2/3	1/3
확률($p(\hat{p})$)	1/10	6/10	3/10

- 표본비율 $\hat{p}$은 표본추출의 결과에 따라 하나의 수값을 대응시키는 확률변수임과 동시에 통계량이며
- 위의 표에서 보이는 것처럼 표본비율 $\hat{p}$의 확률분포를 표본분포라 한다.

표본평균의 분포(무한모집단)

- 일반적인 모집단(주로 무한모집단)에서 표본평균의 표본분포에 대한 특징에 대해서 알아보자.

Theorem

모평균이 μ, 모분산이 σ^2인 무한모집단에서 크기 n인 랜덤표본 $X_1, \cdots, X_n$을 추출하였을 때, 랜덤표본의 표본평균 $\bar{X}$에 대하여 다음이 성립한다.

$$(1)\ E(\bar{X}) = \mu$$

$$(2)\ Var(\bar{X}) = \frac{\sigma^2}{n}, \quad sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

표본평균의 분포(무한모집단)

pf) (1) 랜덤표본의 정의로부터 $E(X_i) \equiv \mu$이므로,

$$\begin{aligned} E(\bar{X}) &= \frac{1}{n} E(X_1 + \cdots + X_n) \\ &= \frac{1}{n}\{E(X_1) + \cdots + E(X_n)\} \\ &= \mu \end{aligned}$$

pf) (2) $X_1, \cdots, X_n$이 서로 독립이고 $Var(X_i) \equiv \sigma^2$이므로,

$$\begin{aligned} Var(\bar{X}) &= \frac{1}{n^2} Var(X_1 + \cdots + X_n) \\ &= \frac{1}{n^2}\{Var(X_1) + \cdots + Var(X_n)\} \\ &= \frac{\sigma^2}{n} \end{aligned}$$

표본평균의 분포(무한모집단)

- 추출한 표본에 따라 표본평균의 값이 달라지므로 표본평균의 분포가 생긴다.
- 표본평균 $\bar{X}$의 평균은 모집단의 평균과 같으며($E(\bar{X}) = \mu$) 표본의 크기 n이 클수록 그 분산이 0에 가까워져서

$$Var(\bar{X}) = \sigma^2/n$$

결국 $\bar{X}$는 모집단의 평균인 μ근처에 밀집되어 분포한다.

표본평균의 분포(정규분포 모집단)

- 모집단에 대한 통계적 추론시 모분포에 대한 정규성 가정이 많음

Theorem

모집단의 분포가 정규분포 $N(\mu, \sigma^2)$일 때 표본평균 $\bar{X}$는 정규분포 $N(\mu, \frac{\sigma^2}{n})$을 따른다.

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

- 정규모집단에서 단순임의추출한 표본평균의 분포는 표본의 개수와 상관없이 항상 정규분포를 따른다.
- 표준화:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \iff \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

표본평균의 분포(임의의 모집단)

Theorem (중심극한정리(central limit theorem))

평균이 μ, 분산이 σ^2인 임의의 모집단에서 표본의 크기 n이 충분히 크면, 랜덤표본의 표본평균 $\bar{X}$는 근사적으로 정규분포 $N(\mu, \frac{\sigma^2}{n})$을 따른다. 즉, n이 충분히 클 때 다음이 성립한다.

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \dot{\sim} N(0, 1)$$

- 표본의 크기 n이 충분히 클 때에는 모집단의 분포가 어떤 형태이든 간에, 표본평균 $\bar{X}$의 분포는 근사적으로 정규분포를 따른다.
- 경험적으로 $n > 30$이면 근사정도가 만족할 만하다고 알려져 있다.

중심극한정리 예시

Example

모집단 분포가 $0, 1, \cdots, 9$의 각 정수를 택할 확률이 동일하게 0.1인 이산확률분포라 하자. 이 모집단에서 크기 5인 랜덤표본을 100개 추출한 결과와 각 표본평균 $\bar{X}$의 값이 아래표에 주어져 있다.

표 4-5 [예 4-4]의 모집단 분포에서 추출한 크기 5의 랜덤표본

표본 번호	관측값	합계	평균	표본 번호	관측값	합계	평균
1	4, 7, 9, 0, 6	26	5.2	51	4, 7, 3, 8, 8	30	6.0
2	7, 3, 7, 7, 4	28	5.6	52	2, 0, 3, 3, 2	10	2.0
3	0, 4, 6, 9, 2	21	4.2	53	4, 4, 2, 6, 3	19	3.8
4	7, 6, 1, 9, 1	24	4.8	54	1, 6, 4, 0, 6	17	3.4
5	9, 0, 2, 9, 4	24	4.8	55	2, 4, 5, 8, 9	28	5.6
6	9, 4, 9, 4, 2	28	5.6	56	1, 5, 5, 4, 0	15	3.0
7	7, 4, 2, 1, 6	20	4.0	57	3, 7, 5, 4, 3	22	4.4
8	4, 4, 7, 7, 9	31	6.2	58	3, 7, 0, 7, 6	23	4.6
9	8, 7, 6, 0, 5	26	5.2	59	4, 8, 9, 5, 9	35	7.0

- 이러한 100개의 $\bar{X}$값들에 대한 히스토그램을 그려보면 다음과 같다.

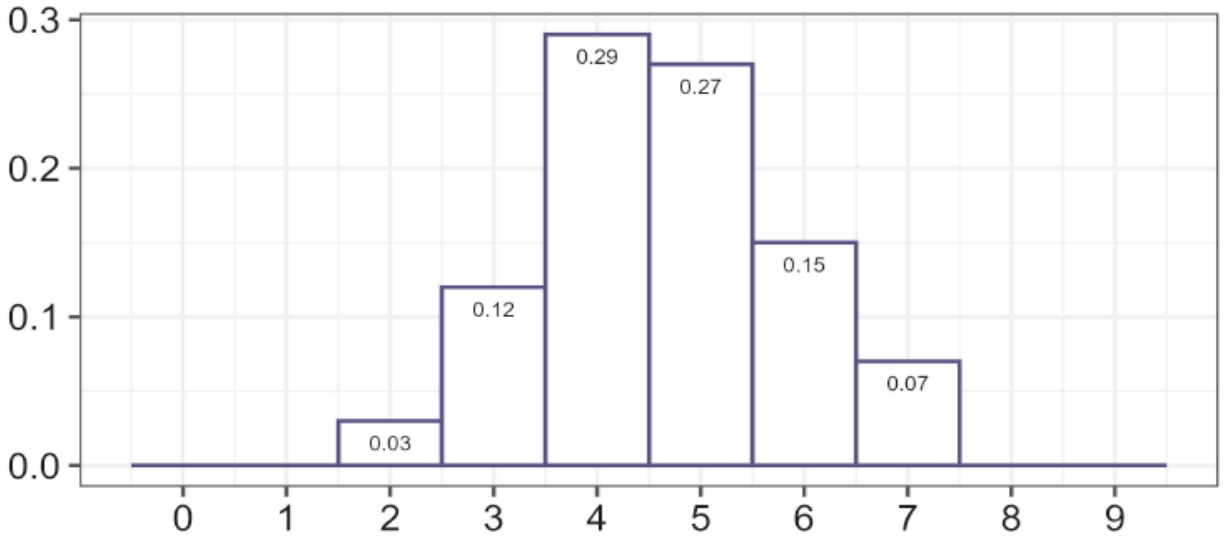
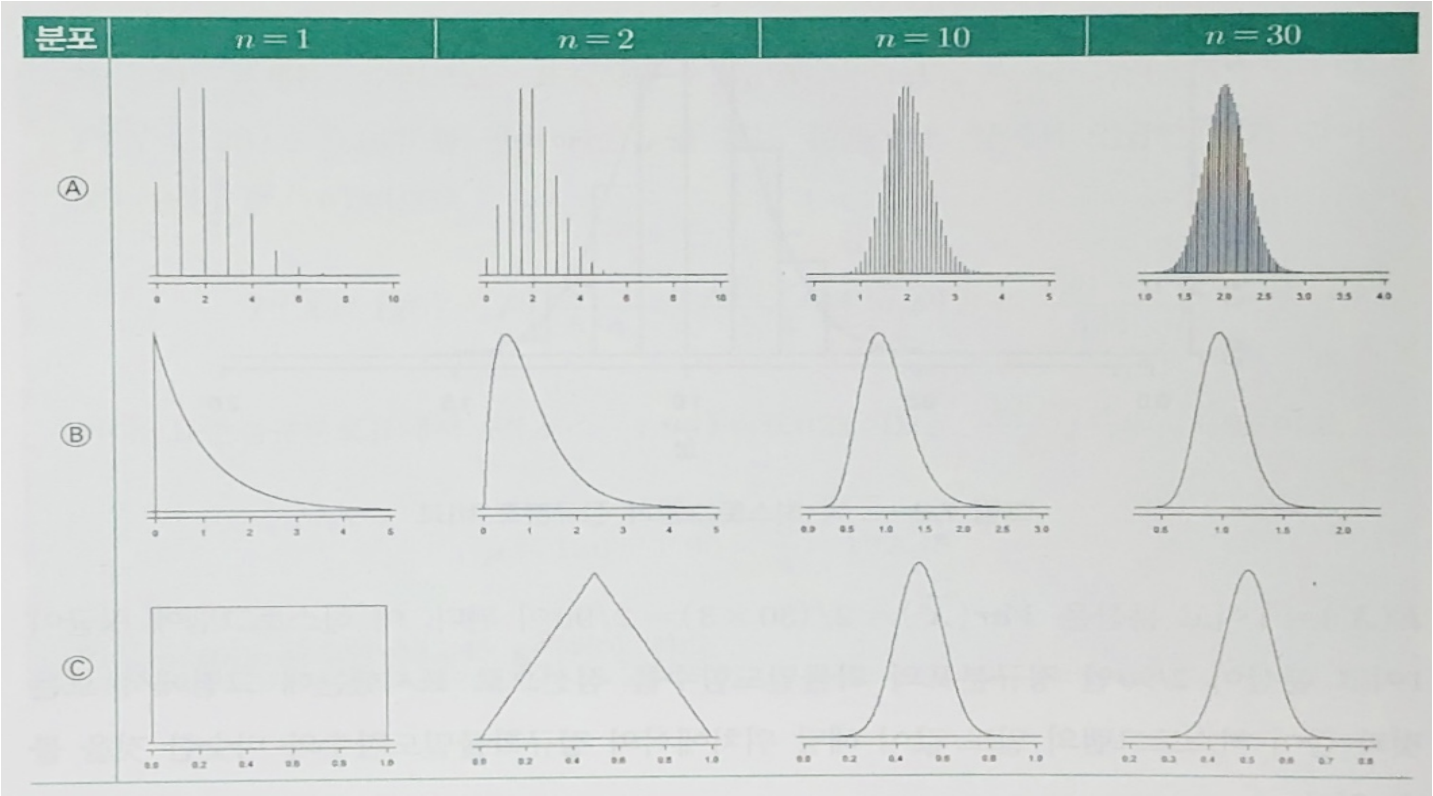


그림 4–5 [표 4–5]에 주어진 $\overline{X}$ 값들의 히스토그램

- 모집단의 형태가 정규분포가 아니더라도 표본의 크기가 5인 경우에 조차 히스토그램의 형태가 정규분포의 형태에 가까움을 알 수 있다.

중심극한정리(central limit theorem)

- 표본의 크기 n에 따른 표본평균 $\bar{X}$의 분포
  - 다양한 형태의 모집단에서 확률표본을 뽑았을 때
  - 표본크기가 작은 경우에는 모집단의 형태에 영향을 받으나
  - n이 커지면서 표본평균의 분포가 정규분포와 비슷해짐



표본평균의 분포 - 정리

- 표본평균 $\bar{X}$의 분포 (단순임의추출에 의한)

1. 모집단이 $N(\mu, \sigma^2)$ 일 때:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

=> 정확한 분포. 정규분포의 성질임(수리적으로 정확한 결과)

2. 모평균이 μ, 모분산이 σ^2, 표본크기 $n \geq 30$일 때:

$$\bar{X} \dot{\sim} N(\mu, \frac{\sigma^2}{n})$$

=> 중심극한정리에 의한 근사분포.

표본평균의 분포(유한모집단)

- 모평균이 μ, 모분산이 σ^2, 크기가 N인 유한모집단에서 비복원으로 크기가 n인 랜덤표본 $X_1, \cdots, X_n$을 추출한 경우에는 다음이 성립한다.

$$E(\bar{X}) = \mu, \;\; Var(\bar{X}) = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$$

- n 고정시키고 N을 무한대로 늘리면

$$Var(\bar{X}) = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n} \quad \xrightarrow{N->\infty} \quad \frac{\sigma^2}{n}$$

표본평균의 분포(유한모집단)

- N을 고정시키고 표본크기 n을 늘리면, 결국 전수조사가 된다. 이때 분산은 점점 줄어들어 0으로 수렴하고 표준오차 또한 감소한다. 즉, 표본추출의 통계적 의미가 없어진다.

$$Var(\bar{X}) = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n} \xrightarrow{n->N} 0$$

- 유한모집단인 경우에도 N과 n이 충분히 크면 다음이 성립한다. (중심극한정리 적용)

$$\frac{\bar{X}-\mu}{\sigma\sqrt{\frac{N-n}{n(N-1)}}} \dot{\sim} N(0,1)$$

중심극한정리 예시

Example

한 식품 회사에서 새로 개발한 음료의 평균 칼로리가 168kcal이고 표준편차가 6kcal로 알려져 있다. 음료 100개를 단순랜덤추출했을 때, 표본평균이 167kcal 이상 169kcal 이하일 확률을 구하라.

sol) $n = 100(> 30)$이므로 $\bar{X}$는 근사적으로 정규분포를 따르고 평균은 $\mu = 168$, 표준편차는 $\sigma/\sqrt{n} = 6/\sqrt{100} = 0.6$이다.

$$\bar{X} \dot{\sim} N(168, 0.6^2)$$

- $Z \sim N(0,1)$로 표준화하면

$$\begin{aligned} P(167 \le \bar{X} \le 169) &= P\left(\frac{167-168}{0.6} \le \frac{\bar{X}-168}{0.6} \le \frac{169-168}{0.6}\right) \\ &\doteq P(-1.67 \le Z \le 1.67) \\ &= 0.9525 - 0.0475 = 0.9050 \end{aligned}$$

t분포

Definition (t분포:t distribution)

표준정규분포 $N(0,1)$을 따르는 확률변수를 Z라 하고 이와는 독립이며 자유도 k인 카이제곱분포 χ^2를 따르는 확률변수를 V라 할 때,

$$T = \frac{Z}{\sqrt{V/k}}$$

의 분포를 자유도 k인 t분포라고 한다. 이 때 기호로는 다음과 같이 표기한다.

$$T \sim t(k)$$

- 자유도는 분포가 퍼져있는 정도를 나타내는 것으로 자유도가 클수록 가운데로 모이고 표준정규분포로 수렴한다.

t분포

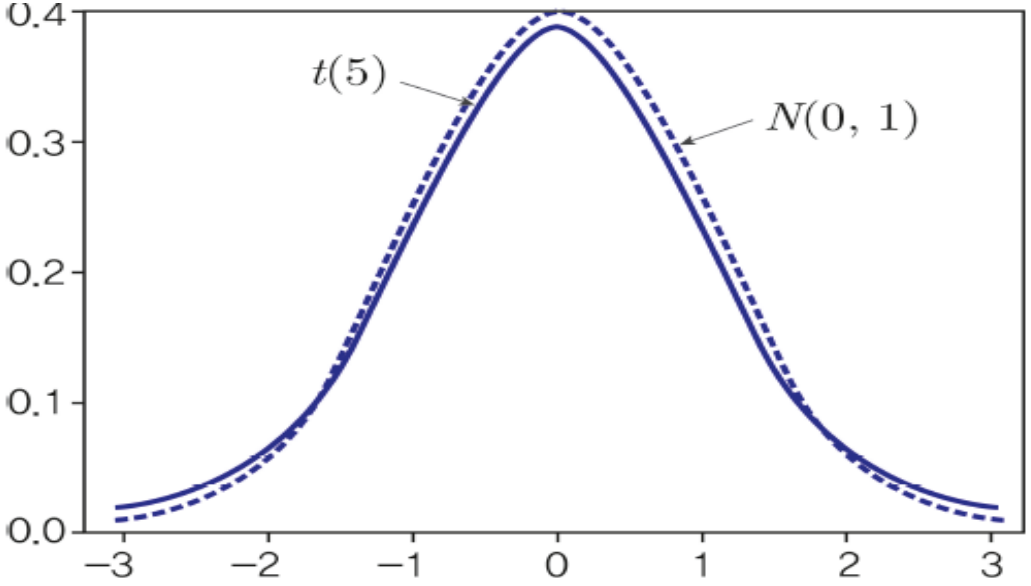


그림 4–8 표준정규분포 $N(0,\ 1)$과 자유도 5 인 t 분포의 비교

- t분포는 표준정규분포와 유사한 모양.
- 0을 중심으로 좌우 대칭형.
- 표준정규분포에 비하여 두터운 꼬리를 갖고 있음.

t분포의 분위수

- $T \sim t(k)$일 때 다음의 식을 만족하는 $t_\alpha(k)$를 자유도가 k인 t분포의 $(1-\alpha)$분위수라고 한다.

$$P\{T \geq t_\alpha(k)\} = \alpha$$

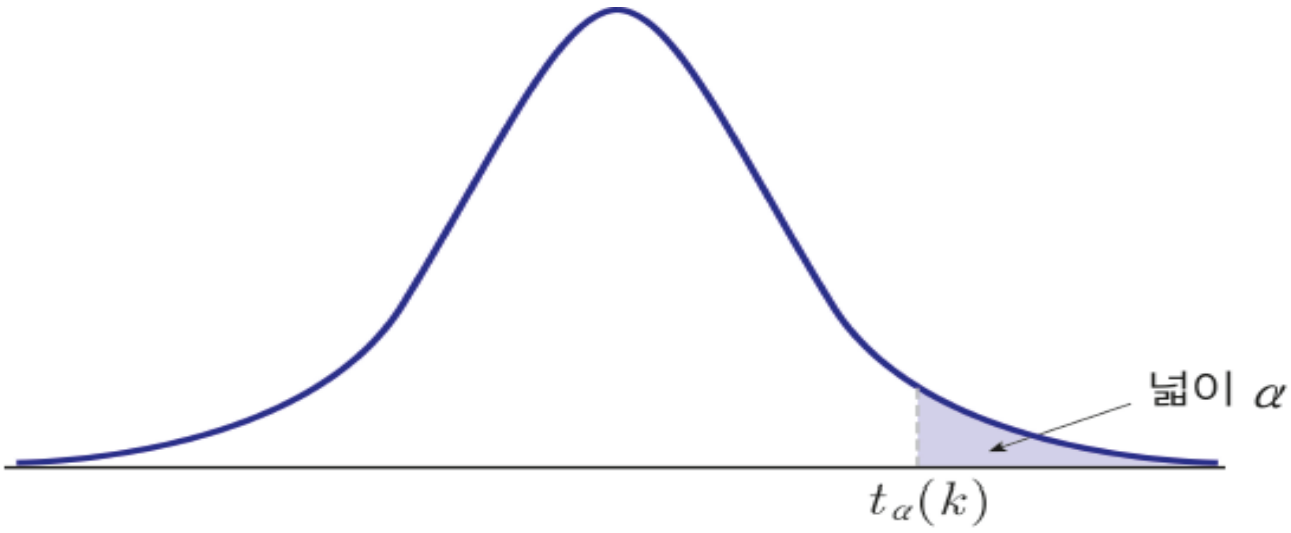
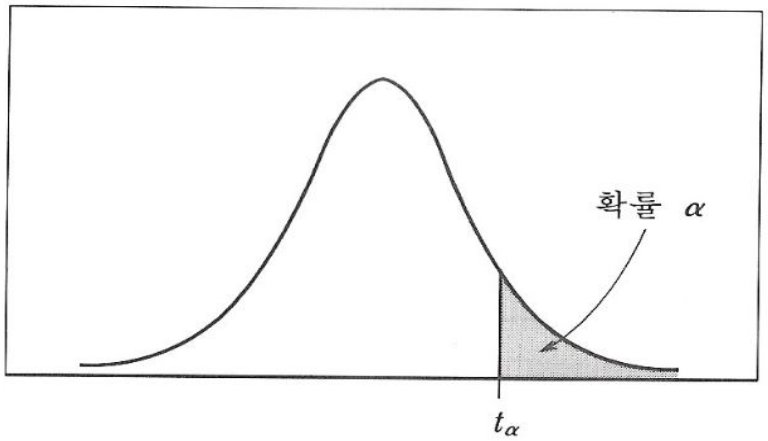


그림 4–9 $t_\alpha(k)$의 위치

t분포표

- t분포표: 자유도별로 각 수준에 대한 여러가지 분위수가 나와있다.
  예) $t_{0.1}(4) = 1.5333$, $t_{0.05}(4) = 2.132$, $t_{0.01}(4) = 3.747$

Figure: t분포표



표의 숫자는 주어진 α 값에 대한 t_α의 값을 나타낸다.

df	α											
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	[illegible]	[illegible]	[illegible]

Windows 정품 인증
[설정]을 클릭하여 Windows를

스튜던트화된 표본평균의 분포

- $X_1, \cdots, X_n \sim_{i.i.d} N(\mu, \sigma^2)$일 때, 다음이 성립함을 다룬 바 있다.

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \quad or \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

그런데 σ가 미지인 경우에는, σ대신에 표본표준편차 S를 대입하여 스튜던트화(studentization)된 표본평균의 분포를 이용할 수 있다.

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

Theorem

$X_1, \cdots, X_n$이 정규모집단 $N(\mu, \sigma^2)$으로부터의 랜덤표본일 때, 표본평균 $\bar{X}$에 대하여 다음이 성립한다.

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

pf) $\frac{\bar{X}-\mu}{S/\sqrt{n}}$의 분포를 유도해 보자.

- $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$이며 $\bar{X}$와 S^2는 서로 독립임이 알려져 있으므로
- t분포의 정의에 따라 $\frac{\bar{X}-\mu}{S/\sqrt{n}}$는 다음과 같이 자유도 $(n-1)$인 t분포를 따르게 된다.

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{(\bar{X} - \mu)/\left(\frac{\sigma}{\sqrt{n}}\right)}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} \sim t(n-1)$$

표본비율의 분포

- 비율의 표본분포
  - 정부 정책에 대한 찬성률
  - 벤처기업의 파산 비율
  - 생산제품의 불량률

- 비율은 평균의 특수한 경우임
  예) 다섯 번 시도 중 두 번 성공, 세 번 실패면 성공 비율은 0.4

$$\hat{p} = 0.4 = \frac{S+S+F+F+F}{5}$$
$$= \frac{1+1+0+0+0}{5} = \frac{\sum x_i}{n} = \bar{x}$$

표본비율의 분포

- $X_1, \cdots, X_n$이 성공확률이 p인 베르누이 분포 $B(1,p)$를 따르는 무한모집단으로부터의 랜덤표본일때 다음과 같이 정의된다.

$$X_i = \begin{cases} 1, & (success) \\ 0, & (fail) \end{cases} \quad \sim_{iid} B(1,p), for\ i = 1, \cdots, n$$

- n개 표본에서 "성공"의 갯수 X는 베르누이 시행 X_i들의 합으로 표현되며 성공확률을 p로하는 이항분포를 따르게 된다.

$$X = \sum_{i=1}^{n} X_i \sim B(n,p)$$

표본비율의 분포

- 이때 표본비율(sample proportion)을 다음과 같이 정의하자.

$$\hat{p} = \frac{X}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- 표본비율 $\hat{p}$의 분포는 베르누이시행의 평균분포 또는 이항분포 확률변수를 n으로 나눈 확률변수의 분포가 되며 평균과 분산은 다음과 같다.

$$E(\hat{p}) = \frac{E(X)}{n} = \frac{np}{n} = p$$

$$Var(\hat{p}) = \frac{Var(X)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

표본비율의 분포 예시

Example

남자 6명, 여자 4명인 모집단에서 $n = 2$인 표본을 뽑아 모집단에서의 남자의 비율을 추정하려 한다.

sol) 모집단 남성비율은 0.6, 모분산은 0.6(1-0.6)=0.24

- 추출 가능한 모든 표본의 경우의 수: $\binom{10}{2} = 45$

표본구성	(남,남)	(남,여)	(여,여)
경우의 수	15회	24회	6회
확률	0.3333	0.5333	0.1333
남성의 비율($\hat{p}$)	1	0.5	0

- $E(\hat{p}) = 1 \times 0.3333 + 0.5 \times 0.5333 + 0 \times 0.1333 = 0.6$
- 따라서 $E(\hat{p}) = p$이다. 즉, 표본비율 $\hat{p}$의 기대값은 모집단의 모비율 즉, 모수 p와 같다.

표본비율의 분포

- 표본비율 $\hat{p}$의 근사분포:

  중심극한정리로부터 n이 클 때 다음과 같이 근사적으로 정규분포를 따르게 된다. (이항분포의 정규근사)

$$\hat{p} \dot{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

$$or$$

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1) \tag{1}$$

이항분포의 정규근사

- 이항분포에서 n이 크면 확률의 계산이 힘듬
- 이산형(discrete)인 이항분포의 여러 확률을 중심극한정리를 적용시켜 연속형(continuous)의 정규분포로 근사적으로 구할 수 있다.
- $E(\hat{p}) = p$, $Var(\hat{p}) = \frac{p(1-p)}{n}$이므로 중심극한정리에 의하여 다음이 성립하고,(by,(1))

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1)$$

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{X - np}{\sqrt{np(1-p)}}$$

이므로 다음의 정리가 성립한다.

이항분포의 정규근사

Theorem (이항분포의 정규근사)

이항분포 $B(n,p)$를 따르는 확률변수 X는 n이 클 때 근사적으로 정규분포 $N(np, np(1-p))$를 따른다. 즉, 다음이 성립한다.

$$\frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1)$$

- 근사의 정밀도는 $np > 5$이고 $n(1-p) > 5$일 때 상당히 좋은 것으로 알려져 있다.

이항분포의 정규근사

- $B(15, 0.4)$의 히스토그램과 이를 근사하는 $N(9, 3.6)$의 곡선

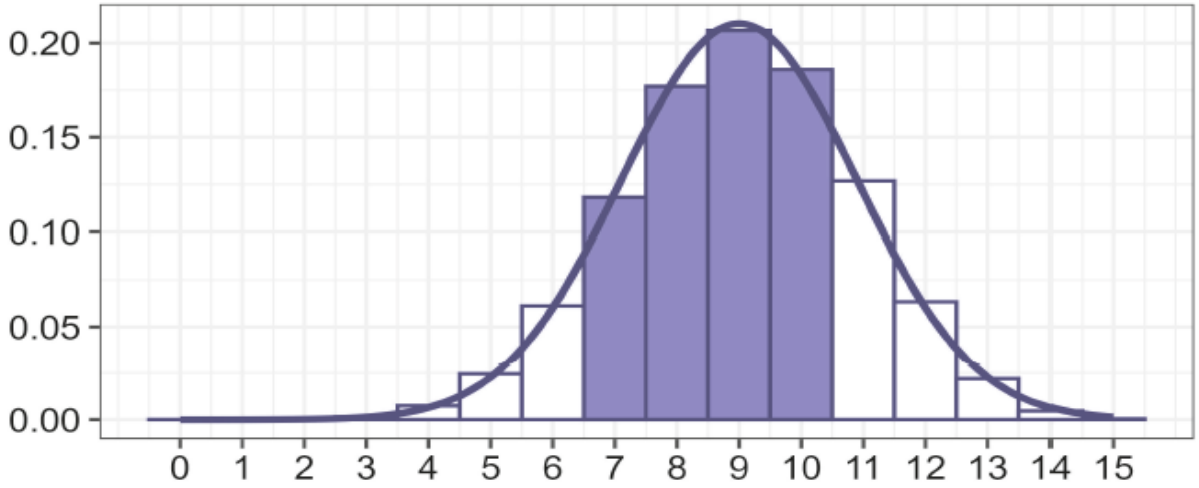


그림 4–6 이항분포 $B(15,\ 0.6)$의 히스토그램과 이를 근사하는 정규분포곡선

- n은 15로서 그리 크지 않더라도 근사정도가 좋음.
- 일반적으로 이항분포가 대칭형일 때는 근접 속도가 매우 빠르다.
- $n = 30$ 정도면 비대칭형일 때도 상당히 근접한다.
- 극단적인 비대칭($p \approx 0$ *or* $p \approx 1$)인 경우에는 n이 아주 커야한다.

이항분포의 정규근사 - 연속성 수정

- 연속성수정(continuity correction):

  [그림 4-6]에서 이항분포의 확률 $P(7 \le X \le 10)$은 히스토그램의 빗금친 부분의 넓이가 되므로, 이를 $P(6.5 \le X \le 10.5)$의 정규분포 $N(9, 3.6)$의 넓이로써 근사시켜 근사의 정밀도를 높일 수 있다.

- $\mu = np,\ \sigma = \sqrt{np(1-p)}$일 때,

  $$P(a \le X \le b) \approx P\left(\frac{a-0.5-\mu}{\sigma} \le Z \le \frac{b+0.5-\mu}{\sigma}\right)$$

  $$P(a \le X < b) \approx P\left(\frac{a-0.5-\mu}{\sigma} \le Z \le \frac{b-0.5-\mu}{\sigma}\right)$$

  $$P(a < X \le b) \approx P\left(\frac{a+0.5-\mu}{\sigma} \le Z \le \frac{b+0.5-\mu}{\sigma}\right)$$

  $$P(a < X < b) \approx P\left(\frac{a+0.5-\mu}{\sigma} \le Z \le \frac{b-0.5-\mu}{\sigma}\right)$$

이항분포의 정규근사 예시

Example

X가 이항분포 $B(100, \frac{1}{5})$일 때, 다음 확률의 근사값을 구하여라.
(a) $P(X \leq 21)$
(b) $P(15 \leq X \leq 19)$

sol) $np = 100 \times \frac{1}{5} = 20$, $\sqrt{np(1-p)} = \sqrt{100 \times \frac{1}{5} \times \frac{4}{5}} = 4$이므로,

$$\begin{aligned}(a)\ P(X \leq 21.5) &= P(\frac{X-20}{4} \leq \frac{21.5-20}{4}) \\ &= P(Z \leq 0.375) \\ &= 0.6462\end{aligned}$$

$$\begin{aligned}(b)\ P(14.5 \leq X \leq 19.5) &= P(\frac{14.5-20}{4} \leq \frac{X-20}{4} \leq \frac{19.5-20}{4}) \\ &= P(-1.375 \leq Z \leq -0.125) \\ &= 0.4503 - 0.0846\end{aligned}$$

정규분포 분위수대조도

- z_α : 표준정규분포의 제 $100(1-\alpha)$백분위수로, $Z \sim N(0,1)$일 때, 주어진 α값에 대하여 $P(Z > z) = \alpha$를 만족하는 분위수의 값.
- 정규분포 분위수대조도(or quantile-quantile plot; Q-Q plot):

  정규모집단의 가정을 검토하는 방법으로 표준정규분포의 분위수와 이에 대응하는 자료 분포의 분위수를 좌표로 하는 점들을 그림으로 나타낸 것.
- 점들이 직선 주위에 밀접하여 나타나면 모집단의 분포가 정규분포라는 가정이 타당하다고 할 수 있다.

정규분포 분위수대조도

Example

$x_1, x_2, \cdots, x_{100}$이 정규분포 $N(47, 10^2)$를 따르는 모집단으로부터의 표본인지를 검토하여 보자.

- 표준정규분포의 분위수를 $\{z_{0.99}, z_{0.98}, \cdots\}$라고 하고

  자료의 값들을 크기순으로 정열한 $x_{(1)} < x_{(2)} < \cdots < x_{(100)}$에서 $x_{(i)}$를 대략 이 자료 분포의 $(i/100)$ 분위수라고 하자.

정규분포 분위수대조도

- 이 때, 다음과 같은 관계식에 의해서

$$\frac{X-\mu}{\sigma} = Z \iff X = \sigma Z + \mu$$

$x_{(1)}$, $x_{(2)}, \cdots$에 대응하는 정규분포 $N(47, 10^2)$의 분위수는 각각 다음과 같이 주어진다.

$$10 \times z_{0.99} + 47,\ 10 \times z_{0.98} + 47, \cdots$$

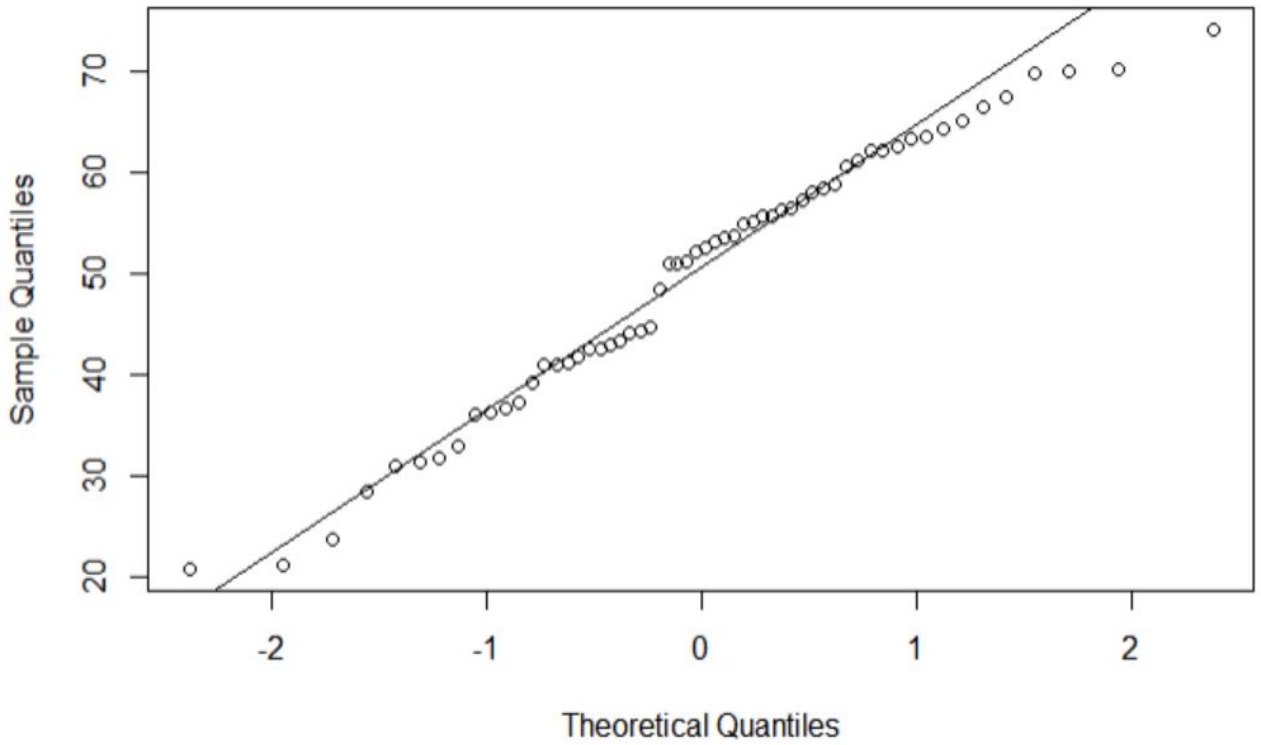
- 실제 모집단의 분포가 $N(47, 10^2)$인 경우에는 위의 서로 대응하는 분위수들은 서로 가깝게 나타날 것이다.

정규분포 분위수대조도

- 그러면 자료분포의 분위수와 이에 대응하는 표준정규분포의 분위수를 좌표로 하는 점들을 그림으로 그려보자.

$$(z_{0.99}, x_{(1)}),\ (z_{0.98}, x_{(2)}), \cdots$$

Figure: 정규분포 분위수대조도 출력 결과

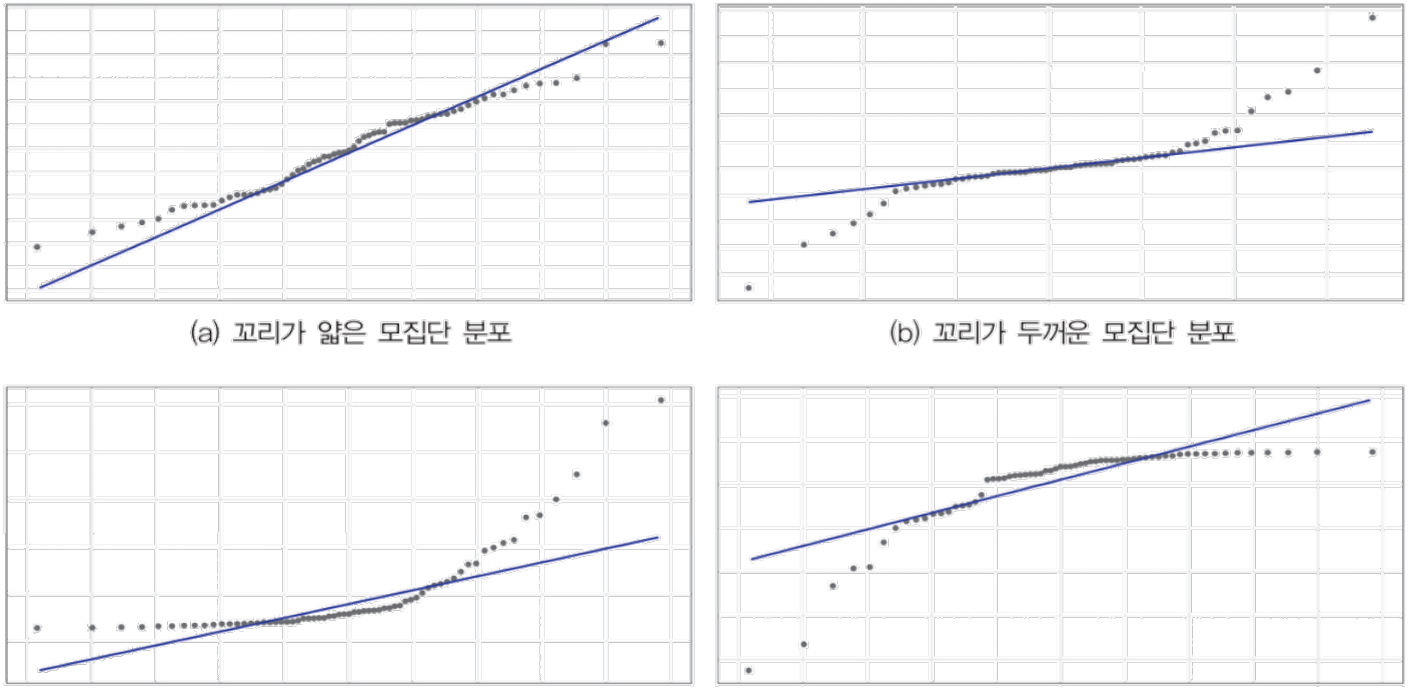


정규분포 분위수대조도

- 자료분포의 분위수와 이에 대응하는 표준정규분포의 분위수를 좌표로 하는 점들은 직선 주위에 밀집하여 나타날 것이다.
- 위 그림의 경우에는 정규모집단의 가정이 타당하다고 할 수 있다.
- 통계패키지에서는 표본의 크기가 n일 때 자료 분포의 분위수 $x_{(i)}$에 대하여 표준정규분포의 (i/n) 분위수 대신 흔히 $\left(\frac{i-3/8}{n+1/4}\right)$ 분위수를 대응시켜 그린다.

정규분포 분위수대조도

- 정규분포 분위수대조도에서 점들이 직선 주위에 가까이 나타나지 않는 경우, 모집단이 정규분포가 아닐 가능성이 큼
- 아래 그림은 이러한 경우에 추측할 수 있는 전형적인 모집단 분포의 형태



(a) 꼬리가 얇은 모집단 분포

(b) 꼬리가 두꺼운 모집단 분포

(c) 오른쪽 꼬리가 긴 모집단 분포

(d) 왼쪽 꼬리가 긴 모집단 분포

그림 4–12 정규분포 분위수대조도와 모집단 분포의 형태

5장. 통계적 추론

박명현

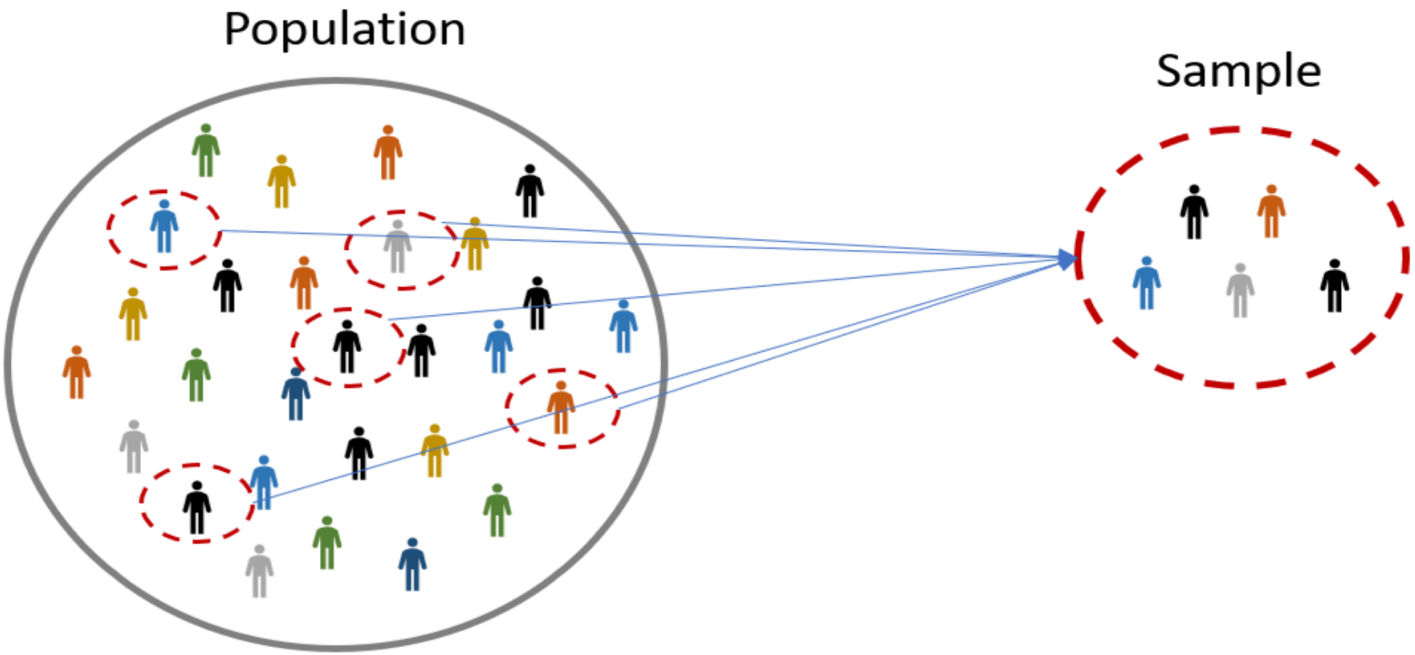
서울대학교 학부대학

Overview

1. 통계적 추론(statistical inference)
2. 모수의 추정
   - 점추정
   - 구간추정
   - 가설검정
   - 검정오류
3. 모평균에 관한 추론
4. 모비율에 관한 추론
5. 모분산에 관한 추론

통계적 추론(statistical inference)

- 통계적 추론: 표본에서 얻은 자료의 정보를 이용하여 모집단에 관한 추측이나 결론을 이끌어 내는 과정
- 추측의 대상: 모집단의 평균, 표준편차, 비율 등과 같은 모집단의 특성치(모수)



통계적 추론(statistical inference)

- 추정(estimation): N개 또는 무한의 원소로 된 모집단에서 n개의 표본을 추출한 후, 이를 이용하여 모집단의 모수의 값을 추측하고 그 오차범위를 제시하는 과정
  - 점추정(point estimation): 모수를 어떤 하나의 값으로 추측하는 것
    예) 청소년들의 하루 TV시청시간(μ)에 대한 점 추정값은 2.75시간이다. $\hat{\mu} = \bar{x} = 2.75$
  - 구간추정(interval estimation): 모수를 추정에 수반된 오차의 크기가 고려된 어떤 구간(신뢰구간)으로 추측하는 것
    예) '청소년들의 하루 TV 시청시간은 2시간에서 3시간 사이이다.
    '$\hat{\mu} \in (2, 3)$
- 가설검정(hypothesis testing): 모집단의 어떤 현상에 대한 예상 또는 주장이 타당한지 표본자료를 이용하여 판단하는 것

모수의 추정

- 모수(population parameter): 모집단의 특성을 나타내는 수치적 측도, θ로 표시. 예) μ, σ^2
- 추정량(estimator): 모수의 추정을 위해 사용되는 통계량, $\hat{\theta}$ 또는 $\hat{\theta}(X_1, X_2, \cdots, X_n)$. 예) $\hat{\mu} = \bar{X}$, $\hat{\sigma}^2 = S^2$
- 추정값(estimate): 관측치로부터 계산된 모수의 예측값. 예) $\bar{x}$, s^2
- 표집오차(sampling error): 모집단 전체를 관측하지 않고, 부분 집합인 표본에서만 관측함으로써 생기는 오차
- 표준오차(standard error): 추정량이 갖는 분산의 제곱근으로 즉, 추정량의 표준편차를 의미(변동성)

$$SE(\hat{\theta}) = \sqrt{Var(\hat{\theta})}$$

점추정
(Point Estimation)

점추정(point estimation)

- 모평균 μ에 대한 추정

  - 추정량: $\hat{\mu} = \bar{X}$

  - 표준오차(모표준편차 σ를 알 때):

$$SE(\hat{\mu}) = \sqrt{Var(\hat{\mu})} = \sqrt{Var(\bar{X})} = \sigma/\sqrt{n}$$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

  - 표준오차(모표준편차 σ를 모를 때):

$$SE(\hat{\mu}) = s/\sqrt{n}$$

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

점추정(point estimation)

- 모비율 p에 대한 추정
  (n개의 랜덤표본에서 그 속성을 갖는 것의 개수: X)
  - 추정량: 표본비율을 사용 $\hat{p} = \frac{X}{n}$
  - 표준오차:$SE(\hat{p}) = \sqrt{Var(\hat{p})} = \sqrt{\frac{p(1-p)}{n}}$

- 모표준편차 σ의 보편적인 추정량은 표본표준편차 S이고,
  모분산 σ^2의 추정량으로는 표본분산인 S^2을 사용한다.

$$\hat{\sigma} = \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)} = S, \quad \hat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1) = S^2$$

점추정의 예

Example (5.1)

한 타이어 제조 회사는 새로운 공정을 개발하며, 기존 공정에 비해 타이어 수명 평균이 증가할 것으로 기대하고 있다. 그러나 표준편차는 기존 공정의 표준편차 $\sigma = 5,000$km를 그대로 유지한다고 가정한다. 새로운 공정으로 제작된 시제품 $n = 100$개를 실험한 결과, 표본의 수명 평균이 $\bar{x} = 38,000$km로 나타났다. 이 결과를 바탕으로 새로운 공정으로 제작한 타이어의 모평균 수명 μ를 얼마로 추정할 수 있을까?

- 모평균 μ에 대한 추정량(estimator)은 $\hat{\mu} = \bar{X}$
- 점 추정값(estimate)은 표본평균 $\bar{x}$을 사용: $\hat{\mu} = \bar{x} = 38,000$km
- $\hat{\mu} = \bar{X}$의 표준오차(standard error):

$$SE(\hat{\mu}) = \sqrt{Var(\hat{\mu})} = \sqrt{Var(\bar{X})} = \frac{\sigma}{\sqrt{n}}$$

점추정의 예

Example

한 공장에서 생산하는 플라스틱판의 두께에서 변이 정도는 중요한 특성치로 간주된다. 이 공장에서 생산된 플라스틱판 중 표본 10개를 추출하여 두께를 측정한 결과는 다음과 같다. 이 자료를 기반으로 플라스틱 두께의 모표준편차 σ에 대한 추정값을 구하라.

$$226, 228, 226, 225, 232, 228, 227, 229, 225, 230$$

- sol) 표본평균 $\bar{x} = (226 + 228 + \cdots + 230)/10 = 227.6$이므로
- 표본분산은
  $s^2 = \{(226 - 227.6)^2 + \cdots + (230 - 227.6)^2\}/(10 - 1) = 5.1556$
- 따라서, 모표준편차 σ의 추정값은

$$\hat{\sigma} = s = \sqrt{5.1556} = 2.27$$

좋은 추정량

- 하나의 모수에 대해서 수많은 추정량이 존재한다.

$$\hat{\mu} = \bar{X} \quad or \quad \hat{\mu} = Median$$

이 중 좋은 추정량은 어떤 것인가?

$$\hat{\theta} - \theta = \hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta$$

- 추정량 $\hat{\theta}$의 표집오차 = 변동 + 편향
- 변동 = $\hat{\theta} - E(\hat{\theta})$: $\hat{\theta}$가 평균값에서 떨어진 거리
- 편향 = $E(\hat{\theta}) - \theta$: $\hat{\theta}$의 평균 표집오차 $E(\hat{\theta} - \theta)$
- 예)

$$\bar{X} - \mu = \bar{X} - E(\bar{X}) + E(\bar{X}) - \mu$$

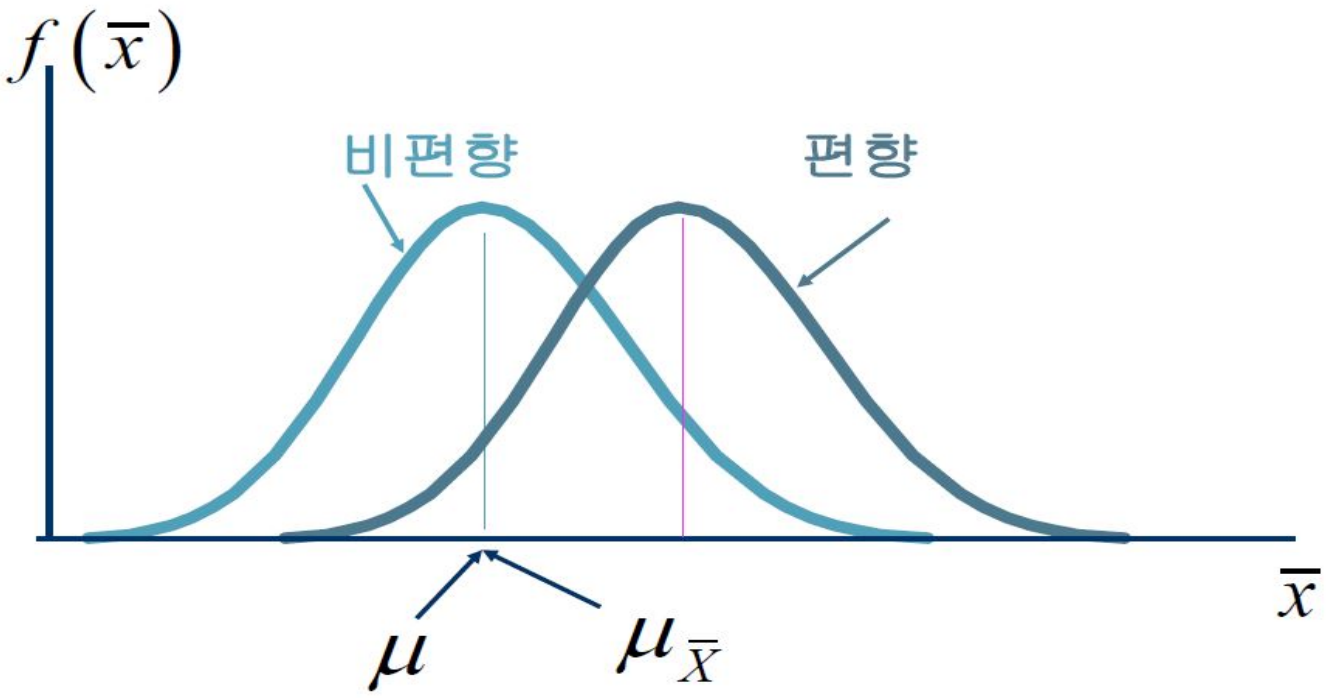
1. 비편향성(Unbiasedness)

- 추정량의 기대값이 모수와 같음
  - $E(\hat{\theta}) = \theta$이면, $\hat{\theta}$은 θ에 대한 비편향추정량(unbiased estimator)
  - $E(\hat{\theta}) \neq \theta$이면, $\hat{\theta}$은 편향추정량(biased estimator)
  - 편향(biased)=$E(\hat{\theta}) - \theta$
  - 편향이 없는 추정량. 편향 크기(bias)=0
- 비편향추정향을 사용하면 그 추정량이 모수와 일치한다는 보장은 없지만, 표본으로부터 얻어지는 추정량을 계산하는 작업을 무한히 반복할 경우, 추정량 값들이 평균적으로는 모수와 일치하게 됨을 의미.

  "추정값은 항상 틀린 값이지만, 평균적으로는 틀리지 않으면 좋겠다"

비편향성(Unbiasedness)

- $E(\hat{\theta}_1) = \theta$ vs $E(\hat{\theta}_2) \neq \theta$



비편향성 예제

Example

모수 μ에 대한 점추정량으로서 표본평균 $\bar{X}$에 대한 비편향성 여부를 판단하여라.

sol) 표본평균 $\bar{X}$의 기대값을 구하면 모수 μ와 일치한다.

$$E(\bar{X}) = \frac{1}{n}E(X_1 + \cdots + X_n) = \frac{1}{n}\{E(X_1) + \cdots + E(X_n)\} = \mu$$

따라서 표본평균 $\bar{X}$는 모평균 μ에 대한 비편향추정량이 된다.

- 표본중위수(Median) : 대칭분포 모평균 μ에 대한 비편향추정량
- 표본비율 : 모비율 p의 비편향추정량
- 표본분산 : 모분산 σ^2의 비편향추정량 (n-1로 나누었던 이유)

2. 효율성(Efficiency)

- 추정량의 분산: 추정량의 변동 $\hat{\theta} - E(\hat{\theta})$의 제곱의 기대값

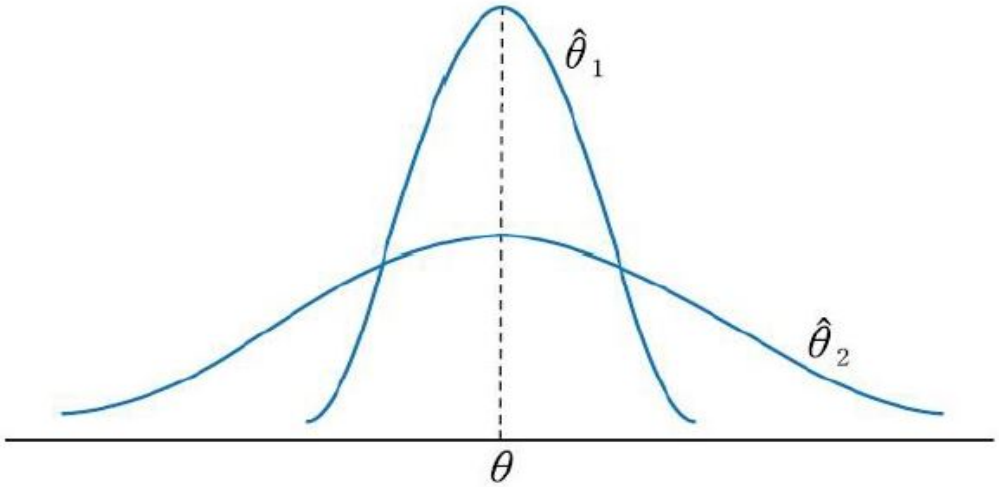
$$E[(\hat{\theta} - E(\hat{\theta}))^2]$$

- 추정량의 변동이 작으려면 추정량의 분산이 작아야 하고, 분산이 작은 통계량은 상대적으로 더 효율적(relative efficient)이다.
- 추정량의 분산이 작다는 것은 표본을 여러 번 추출하여 추정량들을 각각 구해 보았을 때 그 변동폭이 작다는 뜻이다.

"추정값이 표본에 따라 그 때 그 때 많이 다르지 않으면 좋겠다"

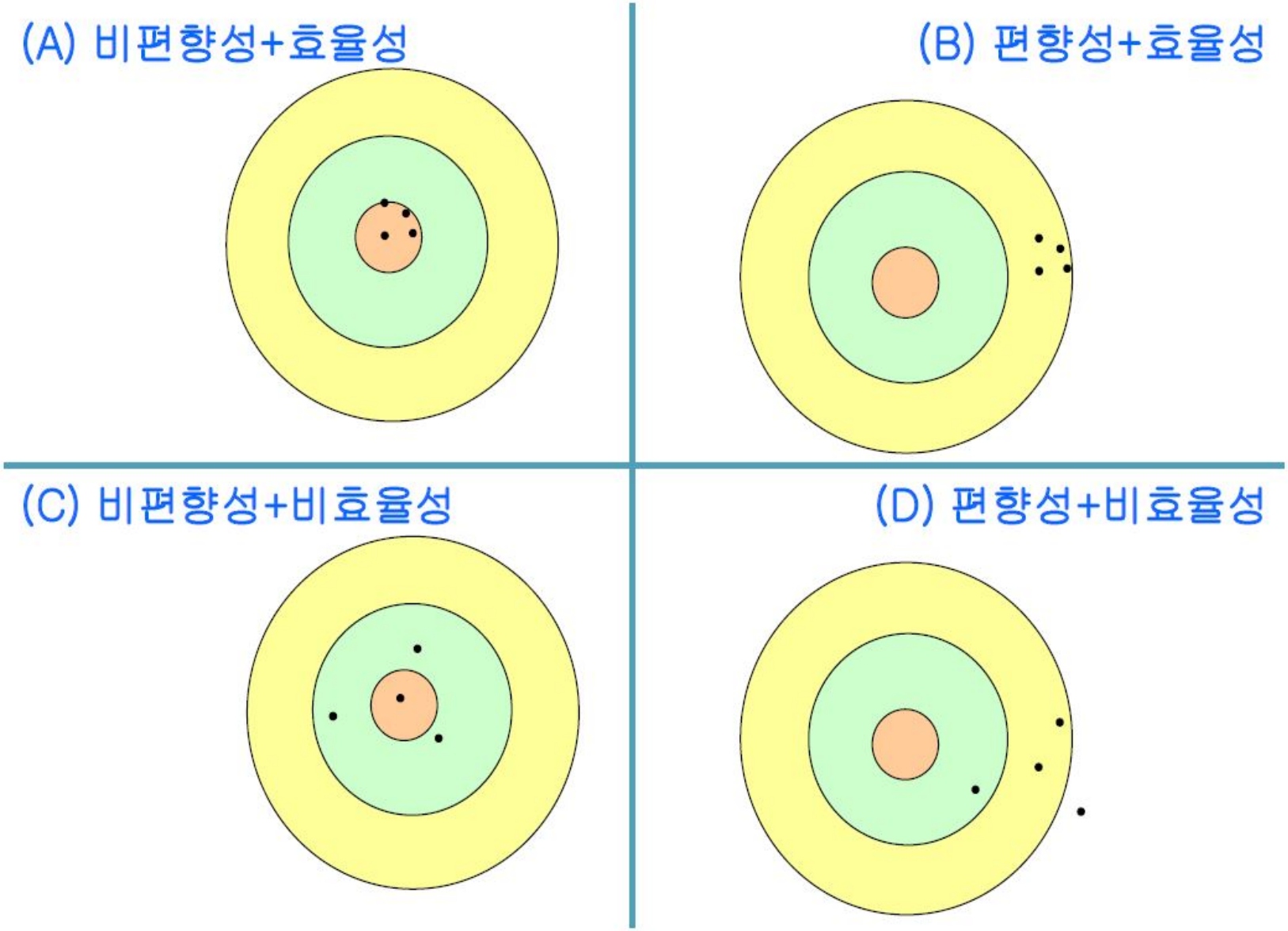
효율성(Efficiency)

예) 두 추정량 $\hat{\theta}_1$과 $\hat{\theta}_2$의 확률밀도함수가 다음과 같다고 가정하자.



- 두 추정량 모두 비편향추정량이지만 $\hat{\theta}_1$의 분산이 $\hat{\theta}_2$의 분산보다 작다.
- $\hat{\theta}_1$의 분포가 $\hat{\theta}_2$보다 모수 θ에 가깝게 몰려있으므로 모수에 가까운 값이 나올 가능성이 높다.
- 이때, $\hat{\theta}_1$이 $\hat{\theta}_2$보다 더 효율적인 추정량이라고 한다.

비편향성과 효율성(화살의 과녁 비유)



효율성(Efficiency)

- 앞의 그림에서 (A)는 알고자 하는 모수를 정확히 맞추면서, 표준오차도 작다. 즉, 추정량이 비편향성과 효율성을 모두 만족시키고 있다.

- 편향과 분산을 동시에 줄이는 것이 이상적이나, 기술적인 한계가 있어, 비편향추정량 중에서 분산을 최소로 하는 추정량을 찾는 전략을 쓰게 된다.

- 최소분산비편향추정량(minimum variance unbiased estimator):
  같은 표본에서 도출된 비편향추정량 중에서 분산이 최소가 되는 추정량
  예) 정규확률표본에서

$$Var(\bar{X}) = \frac{\sigma^2}{n}, \qquad Var(Median) = \frac{\pi}{2} \cdot \frac{\sigma^2}{n}$$

효율성 예제

Example

어떤 회사에서 생산하는 베어링의 평균 직경 μ를 추정하기 위해 10개의 표본 $X_1, X_2, \cdots, X_{10}$을 추출하였다. 평균 직경에 대한 다음의 두 추정량을 비교하시오.

$$\hat{\theta}_1 = \frac{1}{10}\sum_{i=1}^{10} X_i, \qquad \hat{\theta}_2 = \frac{1}{5}\sum_{i=1}^{5} X_i$$

sol) 모평균이 μ이고 모분산이 σ^2일 때, 두 추정량의 평균과 분산은 다음과 같다.

$$E[\hat{\theta}_1] = \mu, \quad Var[\hat{\theta}_1] = \frac{1}{10}\sigma^2, \qquad E[\hat{\theta}_2] = \mu, \quad Var[\hat{\theta}_2] = \frac{1}{5}\sigma^2$$

- 두 추정량이 모두 비편향추정량이나, $\hat{\theta}_1$의 분산이 더 작으므로 μ에 대한 추정량으로 $\hat{\theta}_1$을 선호.
- 즉, 표본크기가 클수록 그에 따른 표본평균은 더 효율적인 추정량

3. 일치성(Consistency)

- 표본크기가 점차 커짐에 따라 점 추정량의 값이 모수에 근접해지는 경우
- 일치성은 추정량 $\hat{\theta}$가 모수 θ로 확률적으로 수렴(convergence in probability)한다는 것을 뜻하고 이는 다음과 같이 정의된다.

  - 임의의 양의 상수 ϵ에 대하여

$$\lim_{n\to\infty} \Pr(|\hat{\theta} - \theta| > \epsilon) = 0$$

  - $\Pr(\hat{\theta} = \theta) \xrightarrow[n\to\infty]{} 1$이 되는 $\hat{\theta}$은 일치추정량이다.

" 현실적으로 표본을 무한으로 할 수는 없다. 하지만 표본이 커지면 추정값이 모수와 같아질 것이라는 보장이 있으면 좋겠다."

일치성(Consistency)

예) 표본평균 $\bar{X}$

- 표본평균의 분산은 $\frac{\sigma^2}{n}$으로 표본수가 늘어나면 0으로 수렴한다.
- 추정량의 분산이 0으로 수렴한다는 것은 그 추정량(표본평균)의 표본추출로 인한 변동이 점차 줄어들고, 결국 그 추정량의 평균(모평균)으로 수렴함을 의미한다.
- 즉, 임의의 양의 상수 ϵ에 대하여

$$\lim_{n\to\infty} \Pr(|\bar{X}_n - \mu| > \epsilon) = 0$$

일치성 예제

Example

모평균 μ에 대한 추정량으로 아래의 $\hat{\mu}$를 고려할 때, 이에 대한 비편향성 및 일치성 여부를 판단하여라.

$$\hat{\mu} = \frac{1}{n-1}\sum_{i=1}^{n} X_i$$

sol) 추정량 $\hat{\mu}$의 기대값은 다음과 같이 계산된다.

$$E(\hat{\mu}) = E\left(\frac{X_1 + \cdots + X_n}{n-1}\right) = \frac{n}{n-1}\mu \neq \mu$$

여기서 $E(\hat{\mu})$이 모수 μ와 같지 않으므로 $\hat{\mu}$은 편향추정량이다.

- 그리고 $\hat{\mu}$은 $\frac{n}{n-1}\bar{X}$이므로 $Var(\hat{\mu}) = \left(\frac{n}{n-1}\right)^2 \frac{\sigma^2}{n}$이다.
- n이 무한히 커지면 다음과 같이 $\hat{\mu}$의 평균은 μ, 분산은 0이 되어

$$E(\hat{\mu}) = \frac{n}{n-1}\mu \xrightarrow[n\to\infty]{} \mu$$

$$Var(\hat{\mu}) = \left(\frac{n}{n-1}\right)^2 \frac{\sigma^2}{n} = \frac{n}{(n-1)^2}\sigma^2 \xrightarrow[n\to\infty]{} 0$$

$\hat{\mu}$은 확률적으로 μ가 된다. 즉,

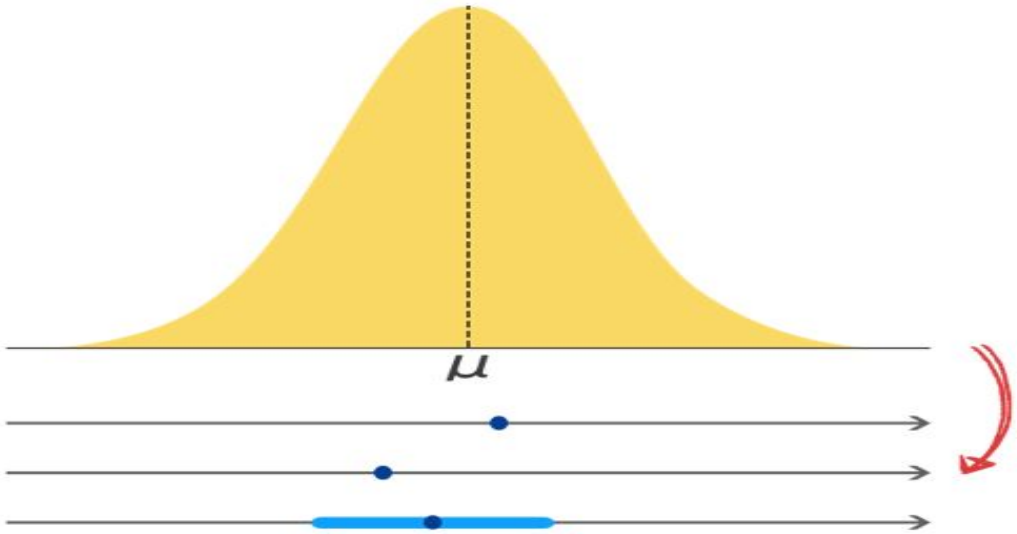
$$\Pr(\hat{\mu} = \mu) \xrightarrow[n\to\infty]{} 1$$

이므로 $\hat{\mu}$은 일치추정량이다.

구간추정
(Interval Estimation)

구간추정

- 모평균 μ는 하나의 상수값인 반면 표본평균 $\bar{X}$는 표본에 따라 많은 다른 값을 취할 수 있는 확률변수이다.
- 모수 μ를 하나의 값으로 추정하는 것이 아니라 추정에 수반된 오차의 크기가 고려된 구간으로 나타내자.



- 구간추정: 관심의 대상인 모수 θ가 포함되어지리라 여겨지는 구간을 표본으로부터 구해내는 것

신뢰구간(confidence interval, CI)

- 아래 식에서 관심모수 θ를 포함할 확률이 $1-\alpha$인 구간 (L,U)를 $100(1-\alpha)\%$ 신뢰수준(confidence level)에서의 신뢰구간이라고 한다.

$$1-\alpha = P(L \leq \theta \leq U)$$

  - 신뢰구간의 하한값 L 또는 상한값 U는 확률변수이다.

- 모수 θ의 $100(1-\alpha)\%$ 신뢰구간:

$$\text{point estimate} \pm \ z_{\alpha/2} \times SE(\hat{\theta})$$

- 신뢰수준: $100(1-\alpha)\%$

모평균 μ의 신뢰구간 - σ를 알 때

- μ를 구간추정하려면 표본평균 $\bar{X}$의 분포에 대한 정보가 필요하다.
- $\bar{X}$의 분포:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \quad \Longrightarrow \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

- 정규분포의 성질로부터 다음이 성립한다.

$$\begin{aligned}
1 - \alpha &= \Pr(-z_{\alpha/2} < Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) \\
&= \Pr(-z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}) \\
&= \Pr(\mu - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}) \\
&= \Pr(\bar{X} - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}})
\end{aligned}$$

모평균 μ의 신뢰구간 - σ를 알 때

- 모평균 μ의 $100(1-\alpha)\%$ 신뢰구간:

$$(\bar{x} - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}) \tag{1}$$

- 오차범위와 신뢰구간의 길이

$$z_{\alpha/2} SE(\hat{\mu}) = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \qquad 2 \times z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- 주의: 식 (1)을 μ의 확률구간으로 읽으면 안 된다.

신뢰수준 90%의 신뢰구간의 의미

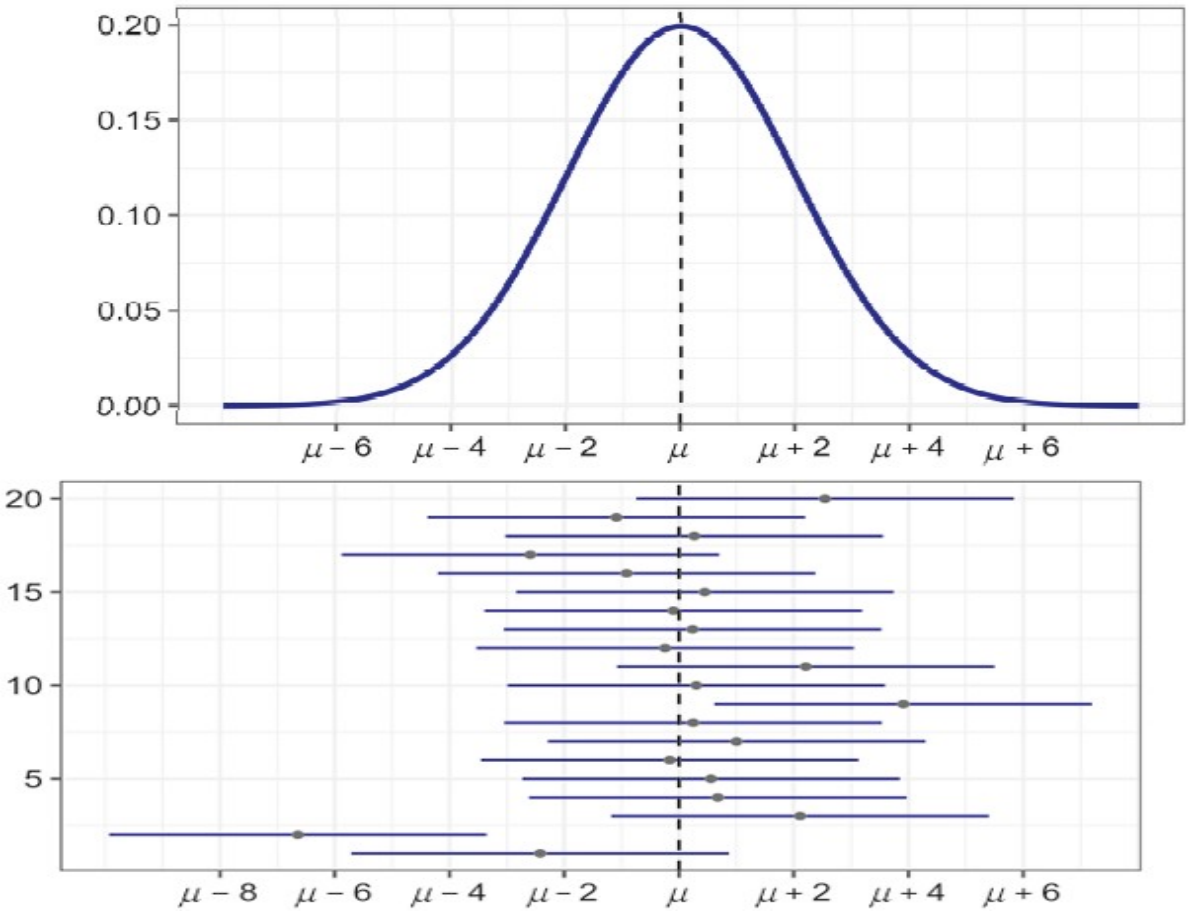
- 모수는 우리가 모르긴 하지만 불변의 값, 상수임.
  따라서 '모수μ가 신뢰구간에 포함될 확률'이라고 해석하면 안된다.
  ex) 원주율 π

예) 다음 그림에서 위의 분포는 정규모집단에서의 $\bar{X}$의 분포이고 아래 쪽에 주어진 구간들은 $n = 25$개의 표본을 20번 추출하여 구한 신뢰수준 90%의 신뢰구간 20개이다.

$$(\bar{x} - 1.645\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.645\frac{\sigma}{\sqrt{n}})$$

- 신뢰구간은 $\bar{x}$를 중심으로 대칭

신뢰수준 90%의 신뢰구간의 의미



- 20개의 구간 중에서 18개의 구간은 모평균 μ를 포함하고 있음
- 즉, 신뢰수준 90%의 의미는 신뢰구간의 가상적인 반복 사용에서 90%가 적중하리라는 뜻

신뢰구간

- 여러 번 표본을 뽑아 신뢰구간을 구하였을때 '신뢰구간이 모수μ를 포함할 비율'이 0.90이다. $\Rightarrow$ 그 비율이 신뢰수준 90%

- 신뢰수준($100(1-\alpha)\%$):
  - 여러 번 구한 신뢰구간 중 추정하고자 하는 모수를 포함하는 신뢰구간의 비율.
  - 흔히 90%, 95%, 99%를 사용하며 이는 각각 $\alpha = 0.1, 0.05, 0.01$에 대응되고 각각의 경우에 $z_{\alpha/2}$ 값은 다음과 같다.

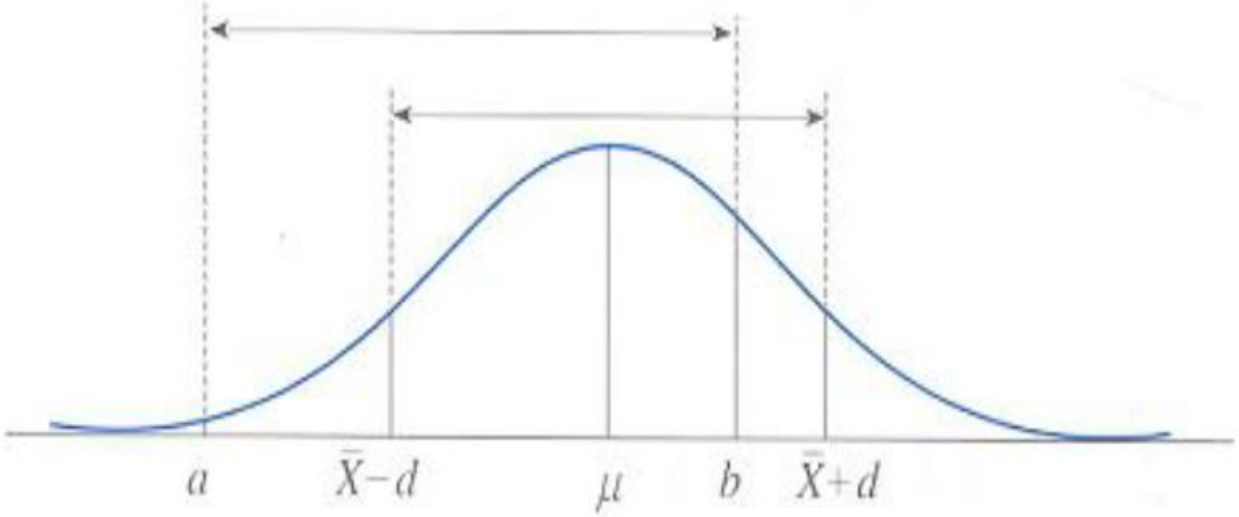
$$z_{0.05} = 1.645, \ z_{0.025} = 1.96, \ z_{0.005} = 2.576$$

- 신뢰수준 $100(1-\alpha)\%$ 신뢰구간의 의미:

  n개의 표본으로 표본평균 $\bar{x}$를 구하는 작업을 무한히 반복한다면 그 들 중 약 $100(1-\alpha)\%$의 표본에 대하여 모평균 μ가 신뢰구간에 속할 것이다.

신뢰구간

- 통계량의 분포가 대칭인 경우
  모평균에 대한 신뢰구간은 $(\bar{X} \pm d)$의 형태로 표본평균을 중심으로 같은 거리만큼 가감하는 것이 효율적이다.



- 신뢰구간 (a, b)도 같은 신뢰수준은 만족하나, 구간의 크기가 $(\bar{X} \pm d)$의 경우보다 길다.
  $\Longrightarrow$ 같은 신뢰수준에서 구간의 크기를 최소로 하는 것은 $(\bar{X} \pm d)$

모평균 μ의 신뢰구간 - σ가 미지인 정규모집단

- 모표준편차 σ가 미지인 정규모집단에서 크기 n인 랜덤표본을 추출:

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

- 이러한 표본분포로부터 다음을 알 수 있다.

$$P\{-t_{\alpha/2}(n-1) \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq t_{\alpha/2}(n-1)\} = 1-\alpha$$

- 위의 식을 정리하면 모평균 μ에 대한 $100(1-\alpha)\%$신뢰구간은 다음과 같다.

$$(\bar{x} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}})$$

모평균 μ에 대한 신뢰구간 정리

- (정규모집단) σ를 모르더라도 표본수 n이 커지면 s가 σ에 근접하고, t분포는 z분포에 수렴한다. 따라서 n이 크면, σ를 모르는 경우의 신뢰구간은 σ를 아는 경우의 신뢰구간과 비슷해진다.
- (X_i가 정규분포를 따르지 않더라도) 표본수가 크면, 중심극한정리에 의하여 $\bar{X}$의 분포는 정규분포에 근접한다. 표본수가 작을 때에는 모집단 분포가 최소한 대칭에 가깝거나, 모집단 분포가 정규분포와 유사해야 한다.
- σ가 미지인 정규모집단에서 표본크기 n이 크면 정규모집단의 가정이 완화될 수 있다.

모평균 μ에 대한 신뢰구간 정리

- 가정사항:
  1. 정규모집단 $N(\mu, \sigma^2)$에서 크기 n인 랜덤표본을 추출
  2. 만약 정규분포가 아니라면 표본의 수가 충분히 크다. $n \geq 30$

- 모평균 μ에 대한 $100(1-\alpha)\%$신뢰구간

  - σ를 알 경우:

$$(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

  - σ를 모를 경우:

$$(\bar{x} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}},\ \bar{x} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}})$$

  - σ가 미지이지만 $n \geq 30$인 경우, by CLT:

$$(\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}},\ \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}})$$

신뢰구간 예제

Example (타이어 제조 회사 continue..)

표준편차 $\sigma = 5,000$km, $n = 100$, $\bar{x} = 38,000$km일 때 95%신뢰구간을 구해보자. ($\alpha = 0.05$)

sol) $\bar{X}$의 분포는 근사적으로 정규분포를 따름: $\bar{X} \sim N(\mu, \frac{5{,}000^2}{100})$

- $\alpha = 0.05$이므로 $z_{\alpha/2} = z_{0.025} = 1.96$이고 $1.96\frac{\sigma}{\sqrt{n}} = 980$이다.
- $\bar{x} = 38,000$이므로 95%신뢰구간은 다음과 같다.

$$(\bar{x} - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}) = (37,020, 38,980)$$

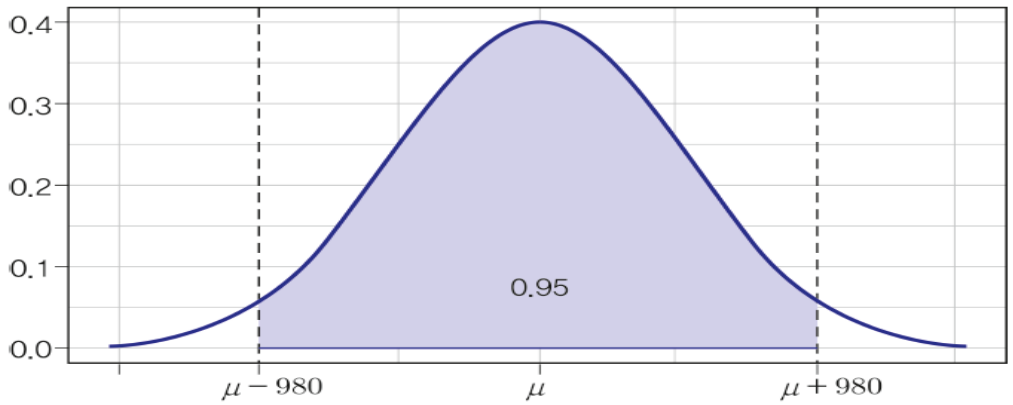


그림 5–1 표본평균 $\overline{X}$ 의 분포

신뢰구간 예제

Example

모표준편차가 $\sigma = 10$으로 알려진 정규모집단에서 $n = 25$개의 표본을 택한 결과 표본평균이 $\bar{x} = 40$이었다면, 모평균 μ의 추정값 $\hat{\mu} = \bar{x} = 40$의 90%오차범위와 신뢰구간을 구하여라.

sol) $\alpha = 0.1$이므로 $z_{\alpha/2} = 1.645$, 오차범위는

$$1.645\sigma/\sqrt{n} = 3.29$$

- μ의 90% 신뢰구간은 다음과 같다.

$$(\bar{x} - 1.645\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.645\frac{\sigma}{\sqrt{n}}) = (36.71, 43.29)$$

- μ의 95% 신뢰구간은 다음과 같다.($z_{\alpha/2} = 1.96$)

$$(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}) = (36.08, 43.92)$$

신뢰구간 예제

Example

물리학자인 뉴컴과 미켈슨은 실험실에서 일정한 거리만큼 떨어진 곳의 거울을 이용하여, 빛이 7,400m의 거리를 움직이는 시간을 측정하였다. 그리고 이 측정시간(초)을 (소요시간-24800$\times 10^{-9}$) $\times$ 10^9으로 변환하여 얻은 $n = 64$개의 자료의 표본평균과 표본표준편차값은 각각 $\bar{x} = 27.750$, $s = 5.083$이다. 이 자료로부터 빛이 7,400m의 거리를 움직이는데 걸리는 시간 μ의 99% 신뢰구간을 구하여라. (정규모집단 가정)

sol) 99% 신뢰구간을 구하려면 $t(63)$ 백분위수를 필요로 한다. 그러나 t분포표에 이 값이 없으므로 가장 가까운 경우인 $t_{0.005}(60) = 2.660$ 값을 사용한다.

$$\begin{aligned}
\bar{x} \pm t_{0.005}(63)s/\sqrt{n} \doteq \bar{x} \pm t_{0.005}(60)s/\sqrt{n} \\
&= 27.75 \pm 2.660 \times 5.083/\sqrt{64} \\
&= 27.750 \pm 1.690 \\
&= (27.06, 29.44)
\end{aligned}$$

신뢰수준과 신뢰구간의 크기

- 신뢰수준은 높을수록, 구간의 크기는 작을수록 바람직.
- 하지만 신뢰수준과 구간의 크기는 서로 상충됨

  => 신뢰수준을 높이면 더 넓은신뢰구간이 얻어지고, 반대로 정교한 구간추정을 위하여 구간의 크기를 작게 하면 신뢰수준이 떨어지는 문제 발생
- 일반적으로 문제에 따라 신뢰수준을 고정, 신뢰수준을 만족하면서 구간의 크기를 가장 작게 하는 신뢰구간을 도출

  ex) $2 \cdot z_{\alpha/2} \times SE(\hat{\theta})$

  1. 분산이 작은 추정량을 사용
  2. 표본수를 늘임

표본크기의 결정 - 모평균 μ의 추정

- μ의 구간추정에서 신뢰수준을 높이면 더넓은 신뢰구간이 얻어지고, 추측되는 μ의 범위가 너무 넓어지게 되므로, 신뢰수준의 선택에서 이를 고려하여야 한다.
- 표본크기 n을 적절히 선택함으로써 오차범위나 신뢰구간의 길이를 조절할 수 있다.
- 모평균의 추정에서 표본크기의 결정:

  $100(1-\alpha)\%$ 신뢰수준으로 오차범위를 d 이하로, 또는 신뢰구간의 길이를 $2d$ 이하로 하기 위한 표본의 크기 (σ를 알 때)

$$n \geq (z_{\alpha/2} \cdot \sigma/d)^2$$

표본크기의 결정 예제

Example

중앙아메리카의 저소득층 원주민을 대상으로 혈청 내 콜레스테롤 양을 조사하려고 한다. 콜레스테롤 양은 근사적으로 정규분포를 따르고, 표준편차가 $\sigma = 30$(mg/L)라고 가정 하자. 모평균 μ에 대한 95% 오차범위를 5(mg/L)이하로 제한하려면, 몇 명의 저소득층 원주민을 표본조사해야 하는가?

sol) 필요한 표본의 크기를 n이라고 하면

$$n \geq (1.96\sigma/5)^2 = 138.2976$$

따라서, 필요한 표본의 크기는 139명 이상이어야 한다.

유의성검증(test of significance)과 가설검정(hypothesis testing)

가설검정(hypothesis testing)

- 가설검정: 모집단의 어떤 현상에 대한 예상 또는 주장이 옳은지 틀린지 표본자료를 이용하여 판단하는 것
- 통계적가설(hypothesis): 모집단의 특성 또는 모수에 대한 대립되는 두 가지 주장에 대하여 통계적으로 다루기 편리하도록 정리한 것
- 귀무가설(null hypothesis, H_0):
  모집단의 모수를 하나의 값 또는 구간 등으로 표시
  예) 한국 청소년들의 하루 평균 TV 시청 시간은 하루 3시간이다.

$$H_0 : \mu = 3$$

- 대립가설(alternative hypothesis, H_1):
  귀무가설에서 제시하는 모수의 값을 제외한 나머지 영역에서 모수의 값을 정의
  예) 청소년들의 하루 평균 TV 시청 시간은 3시간이 아니다.

$$H_1 : \mu \neq 3$$

가설설정 예시

예) 무죄추정의 원칙: 피고인 또는 피의자는 유죄판결이 확정 될 때 까지는 무죄로 추정한다는 원칙

- 귀무가설(null hypothesis, H_0): 피고는 무죄이다.
- 대립가설(alternative hypothesis, H_1): 피고는 유죄이다.

- 가설 설정
  - 특별한 증거 없이는 현재로서는 참이라고 여겨지는 가설을 귀무가설로 하여야 한다.
  - 연구자가 통계적으로 증명하고 싶은 것. 즉, 자료에서 뚜렷한 증거로 입증하고자 하는 가설을 대립가설로 상정한다.
- 귀무가설과 대립가설이 정해진 후 표본이 귀무가설의 주장을 뒷받침하지 못하는 상황이 되면 귀무가설을 기각하게 된다.

가설 설정 규칙

1. 귀무가설은 모수를 특정한 값으로 표현한다.($H_0 : \mu = \mu_0$)
2. 귀무가설에는 반드시 " = "이 포함된다.
3. 대립가설은 비교하는 값의 양쪽을 다 고려하는 양측검정(two-sided hypothesis)과 한쪽만 고려하는 단측검정(one-sided hypothesis)이 있다.

$$H_1 : \mu \neq \mu_0 \text{ -> 양측검정}$$
$$H_1 : \mu > \mu_0 \text{ -> 단측검정}$$
$$H_1 : \mu < \mu_0 \text{ -> 단측검정}$$

- 문제에서 양측검정과 단측검정 중 어느 것을 사용해야 하는지에 대한 판단은 어느 가설이 문제의 내용을 더 잘 반영하는 지를 검토하여 분석자가 결정

가설검정 예제

Example

한 페인트 제조회사에서 생산하는 특수 페인트의 건조시간은 평균 75분, 표준편차 9.4분인 것으로 알려져 있다. 연구소에서는 새로운 첨가제를 사용한 시제품을 사용하면 평균 건조시간이 줄어들 것이라고 주장한다. 실제 시제품 100개를 생산하여 건조시간을 조사한 결과 $\bar{x} = 73.5$가 나왔다. 새로운 첨가제를 사용한 시제품을 사용시 건조시간이 단축된다는 뚜렷한 증거가 있으면 이를 전체 공정에 적용하려고 한다.

- 새로운 첨가제를 사용한 특수 페인트의 평균 건조시간을 μ라고 하자.
- 귀무가설과 대립가설은 각각

$$H_0 : \mu = 75 \qquad H_1 : \mu < 75$$

- $n = 100$으로 충분히 크므로 $\bar{X}$의 분포는 $N(\mu, 9.4^2/100)$이다.

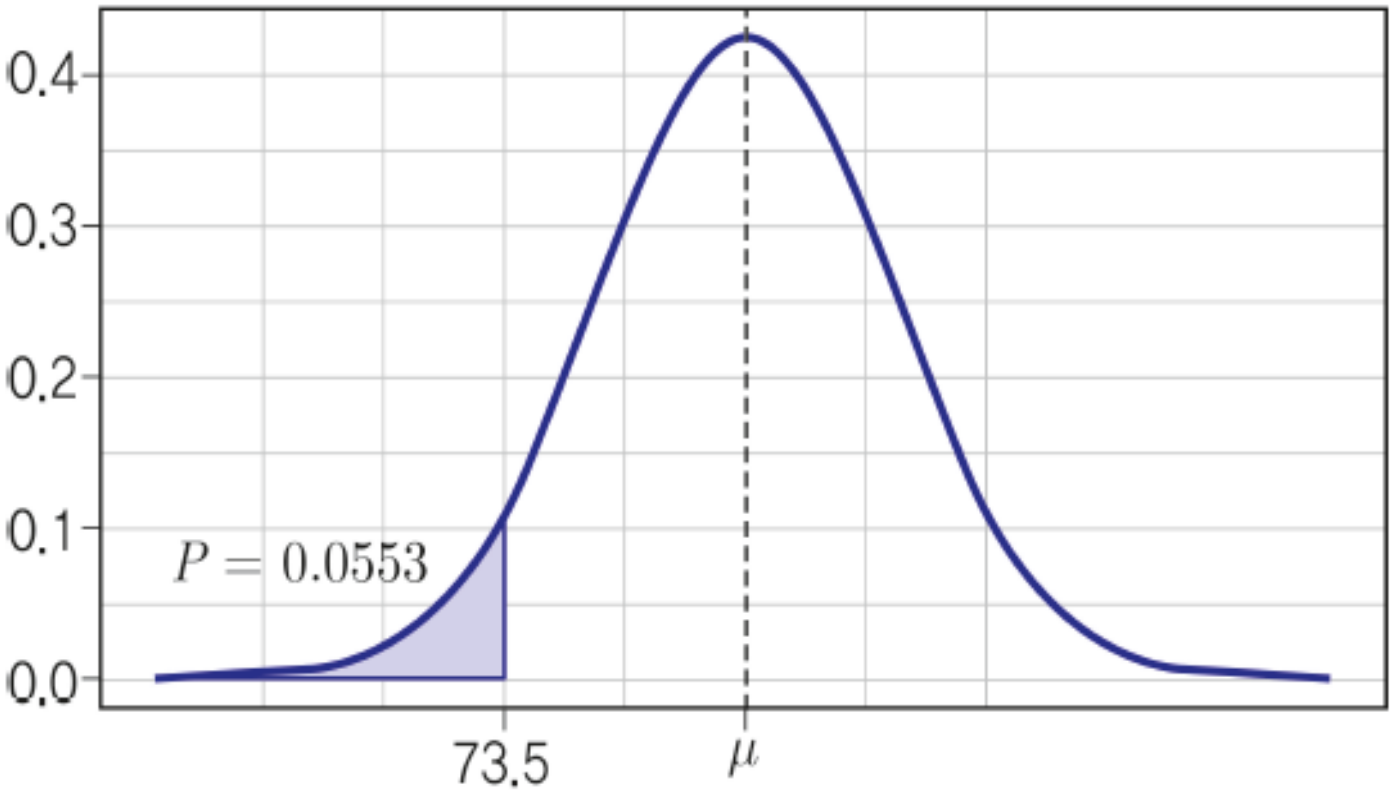


그림 5-3 왼쪽 단측 대립가설일 때 유의확률

- 표본평균 $\bar{X}$가 75보다 작은 쪽으로 멀어질수록 이는 대립가설을 지지하는 증거가 강해진다.

- 귀무가설하에서 관측값 73.5보다 더 대립가설을 지지하는 방향으로 표본평균이 나올 확률은 다음과 같다.

$$
\begin{aligned}
P &= P\{\bar{X} \le 73.5 | \mu = 75\} \\
&= P\left\{\frac{\bar{X} - 75}{9.4/\sqrt{100}} \le \frac{73.5 - 75}{0.94} | \mu = 75\right\} \\
&= P\left\{\frac{\bar{X} - 75}{9.4/\sqrt{100}} \le -1.5956 | \mu = 75\right\} \doteq 0.0548
\end{aligned}
$$

- 확률값이 매우 작으므로 실제 관측된 증거($\bar{x} = 73.5$)보다 더욱 극단적으로 사건이 발생할 가능성은 적다. 즉 더 적은 건조시간이 나올 확률은 매우 적으며, $\bar{x} = 73.5$의 값으로도 충분히 건조시간이 줄어들었다는 것을 강하게 시사한다고 볼 수 있다.
- 즉, 관측값 $\bar{x} = 73.5$의 값은 대립가설 $\mu = 75$보다 충분히 작아 귀무가설을 기각하게 된다.

검증통계량(test statistics)

- 검증통계량(test statistics): 검증에 사용하기 위하여 표본자료에서 구한 통계량. 일반적으로 검정하려는 모수의 점추정량이 되기도 하고, 이 점추정량을 표준화 한 것을 사용하기도 함.
- 관심 모수 μ: 검정통계량(Z)과 그것의 관측값(z_0)은 다음과 같다.

1. σ를 알 때

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1), \quad z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

2. σ를 모를 때

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1), \quad t_0 = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$$

P값과 기각역

- 유의확률 or P값(significance probability, or P-value):

  귀무가설이 사실일 때, 검증통계량이 실제 관측된 값보다 대립가설을 지지하는 방향으로 더욱 치우칠 확률.

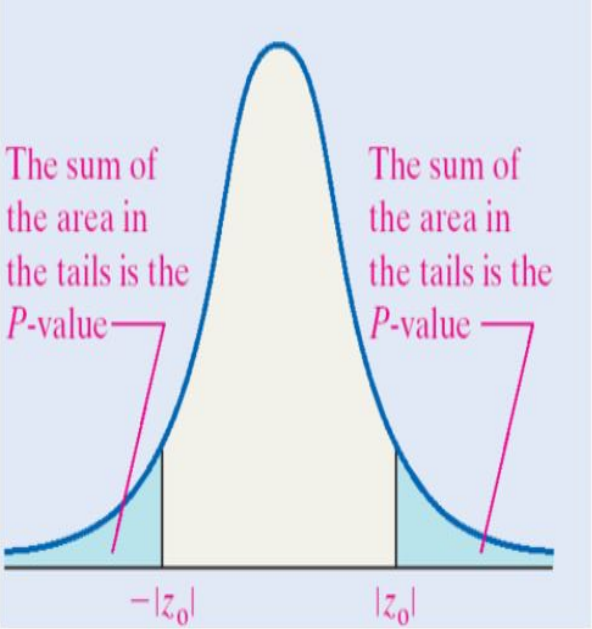
- 유의수준(significance level):

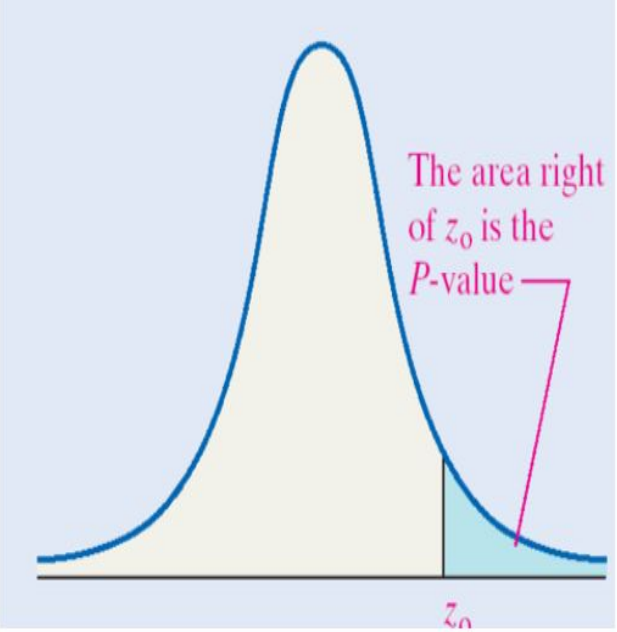
  귀무가설 H_0에 대한 반증의 강도에 대하여 유의수준을 미리 정해놓고 유의확률(P값)을 유의수준값과 비교함. α로 표시. 흔히 $\alpha = 0.05$, $\alpha = 0.01$을 사용.

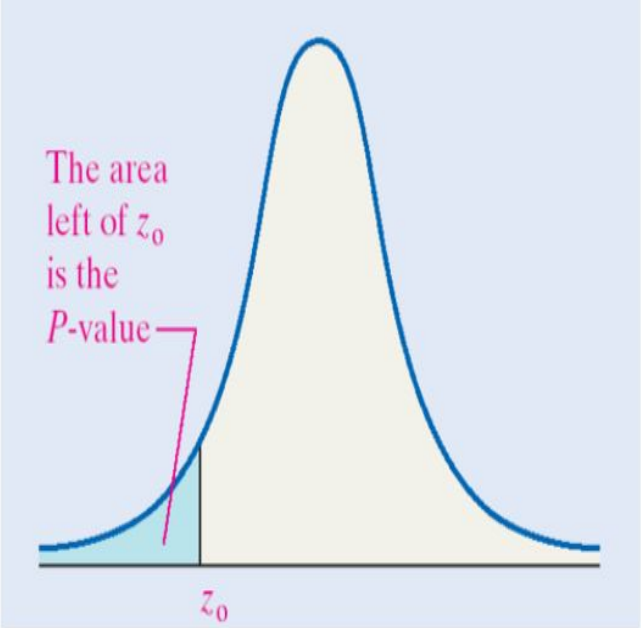
P값과 기각역

- 대립가설과 P값
  (a) $H_1 : \mu > \mu_0$의 경우에는 $P = P(Z \geq z_0)$
  (b) $H_1 : \mu < \mu_0$의 경우에는 $P = P(Z \leq z_0)$
  (c) $H_1 : \mu \neq \mu_0$의 경우에는 $P = P(|Z| \geq |z_0|)$

(c) (a) (b)







P값과 기각역

- p값에 의한 판정

  - p값 $> \alpha$이면, H_0기각 못함.

  - p값 $\leq \alpha$이면, H_0기각.

  "조사 결과가 통계적으로 유의하다"

- 통계적 유의성

: 유의확률(p-value)이 유의수준 α이하이면, 조사 결과가 유의수준 α에서 통계적으로 유의하다고 하며 이는 귀무가설에 대한 반증의 강도가 지정된 수준보다 강함을 뜻한다.

P값과 기각역

- 기각역(rejection region): 지정된 수준에서 H_0가 사실이 아니라고 기각하게 되는 즉, 통계적 유의성을 주장할 수 있는 검증통계량의 영역
- 유의수준, α는 보통 $\alpha = 0.05$, $\alpha = 0.01$을 사용.
- 대립가설과 기각역

  (a) $H_1 : \mu > \mu_0$의 경우에는 $z_0 \geq z_\alpha$

  (b) $H_1 : \mu < \mu_0$의 경우에는 $z_0 \leq -z_\alpha$

  (c) $H_1 : \mu \neq \mu_0$의 경우에는 $|z_0| \geq z_{\alpha/2}$

P값과 기각역

$H_0: \mu = 3$

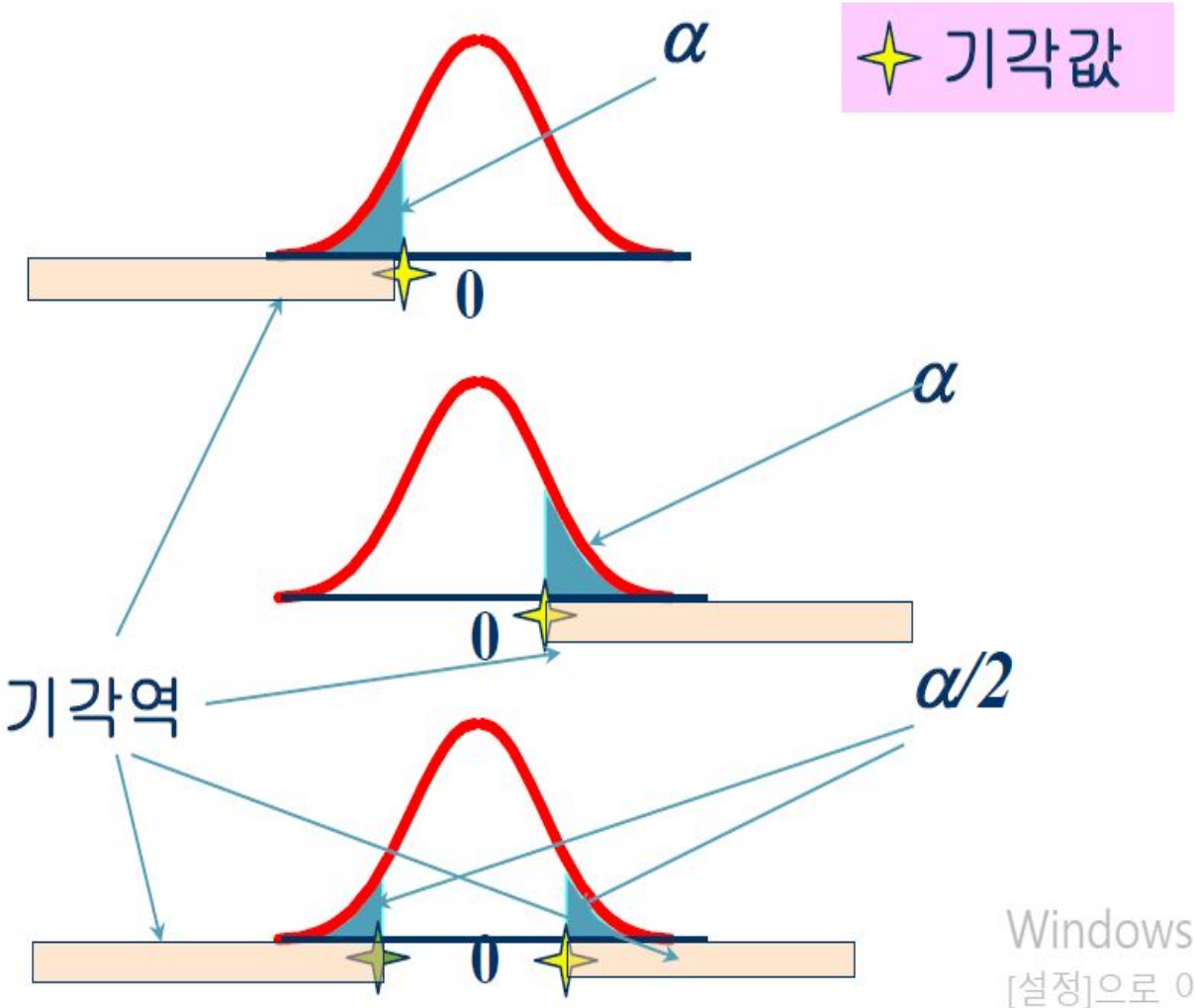
$H_a: \mu < 3$

$H_0: \mu = 3$

$H_a: \mu > 3$

$H_0: \mu = 3$

$H_a: \mu \neq 3$



Windows
[설정]으로 0

P값과 기각역

- 기각역에 의한 판정
  - 유의수준 α에 따라 기각역을 먼저 정하고,
  - 검증통계량의 값이 이러한 기각역에 속하면:
    "유의수준 α에서 귀무가설 H_0를 기각한다."
  - 검증통계량의 값이 기각역에 속하지 않으면:
    "유의수준 α에서 귀무가설 H_0를 기각할 수 없다."

가설검정의 단계 요약

1. 모집단의 특성에 대한 주장, 즉 가설을 설정
   (귀무가설 H_0와 대립가설 H_1)
2. 검정의 유의수준 α를 정함
3. 표본을 추출하여 귀무가설 하에서 검정통계량을 계산
4. 검정통계량에 대한 P값을 구하여 가설에 대한 결론 도출
   or 기각역에 속하는지 검토하여 가설에 대한 기각 여부 결정

모평균에 대한 가설검정; σ를 알 경우

- 모평균의 유의성검증(모표준편차 σ를 알고 유의수준이 α일 때)
  귀무가설이 $H_0 : \mu = \mu_0$인 경우에 검정통계량은

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

  이고, 이의 관측값을 z_0라고 할 때, 다음이 성립한다.

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \mu > \mu_0$	$P = P(Z \geq z_0)$	$z_0 \geq z_\alpha$
(b) $H_1 : \mu < \mu_0$	$P = P(Z \leq z_0)$	$z_0 \leq -z_\alpha$
(c) $H_1 : \mu \neq \mu_0$	$P = P(\lvert Z\rvert \geq \lvert z_0\rvert)$	$\lvert z_0\rvert \geq z_{\alpha/2}$

- - P값에 의한 판정: $P \leq \alpha$이면, H_0기각
  - 기각역에 의한 판정: 위의 기각역 조건이면 H_0기각

모평균에 대한 가설검정; σ를 알 경우

Example

한 타이어 제조 회사에서 생산중인 타이어의 수명시간은 평균이 37,000km이고, 표준편차는 5,000km인 것으로 알려져 있다. 이제, 타이어의 수명을 증가시키는 공정을 개발하고 있는 이 회사의 연구소에서 개발중인 새공정에 의해 시제품을 100개 생산하여 조사한 결과 평균 수명이 $\bar{x} = 38,000$km이었다. 새 공정에 의한 타이어 수명시간의 표준편차 $\sigma = 5,000$km로 유지된다고 할 때, 유의수준 $\alpha = 0.05$하에서 이 연구소의 조사결과는 새 공정이 성공적임을 뜻하는가?

sol) 새 공정에 의한 타이어의 평균수명을 μ라고 하자. 타이어의 수명을 증가시키는 공정을 개발하고자 하므로 귀무가설과 대립가설은 다음과 같다.

$$H_0 : \mu = 37,000 \quad H_1 : \mu > 37,000$$

- 표본크기 $n = 100$이므로 Z의 분포는 근사적으로 표준정규분포를 따른다.
- $\bar{x} = 38,000$이므로 검증통계량의 관측값을 계산하면

$$z_0 = \frac{\bar{x} - 37,000}{5,000/\sqrt{100}} = 2$$

- P값으로 판정:

  유의확률은 다음과 같고, 유의수준 $\alpha = 0.05$보다 작아 귀무가설을 기각한다.

$$P = P\{Z \geq 2\} \doteq 1 - 0.9772 = 0.0228 < 0.05$$

- 기각역으로 판정: 표준정규분포표에서 상방 5%백분위수인 $z_{0.05} = 1.645$와 관측값 z_0를 비교한다.
- 기각역은 $z_0 \geq 1.645$이므로 $z_0 = 2$값이 다음 그림과 같이 기각역안에 속하므로 관측결과가 유의수준 5%에서 통계적으로 유의하다.

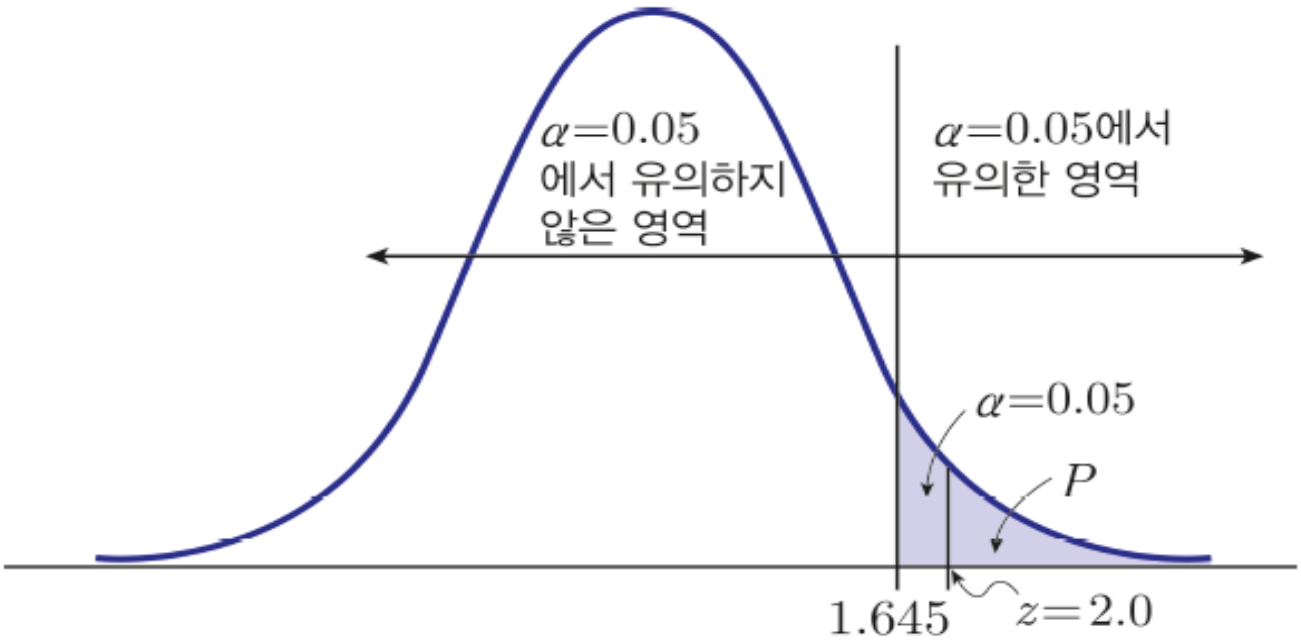


그림 5–5 유의수준 5%에서 유의한 영역

- 즉 새 공정이 상당히 성공적이라는 증거라고 할 수 있다.

모평균에 대한 가설검정; σ를 알 경우

Example

한 제약 회사에서 약품의 주성분 농도가 0.85(gr/L)로 유지되는 제조법의 개발을 목표로 연구중이다. 3번의 반복 측정에서 주성분 농도의 평균이 $\bar{x} = 0.8404$(gr/L)이었을 때 이 제조법에 의한 주성분의 실제 농도 μ가 0.85(gr/L)가 아니라고 의심할만한 증거가 있는가? 농도의 측정값들은 실제의 농도 μ를 평균으로 하고 표준편차가 $\sigma = 0.0123(gr/L)$인 정규분포를 따른다고 한다. 유의수준은 $\alpha = 0.05$라고 설정하자.

sol) 귀무가설과 대립가설은 다음과 같다.

$$H_0 : \mu = 0.85 \quad H_1 : \mu \neq 0.85$$

- 검증통계량의 관측값을 계산하면

$$z_0 = \frac{\bar{x} - 0.85}{0.0123/\sqrt{3}} = -1.35$$

- P값으로 판정: 양측가설이고, 모집단이 정규분포를 따르므로

$$P = P\{|Z| \geq |-1.35|\} = P(Z \leq -1.35) + P(Z \geq 1.35)$$
$$= 2 \times 0.0885 = 0.1770 > 0.05 = \alpha$$

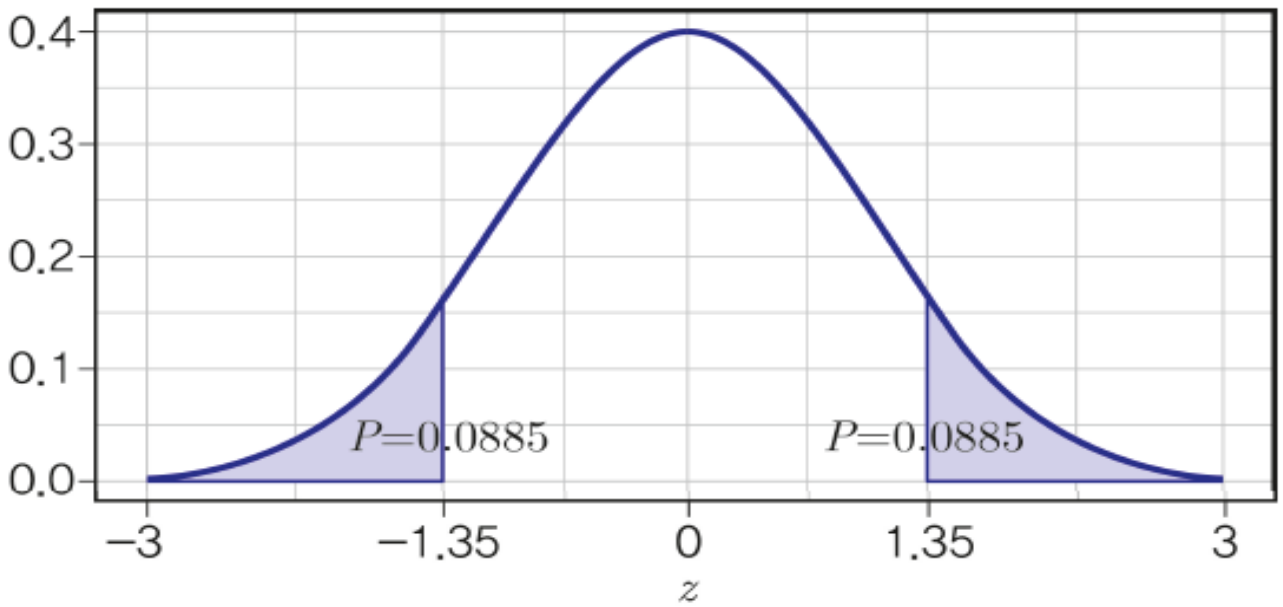


그림 5−4 양측 대립가설일 때 유의확률

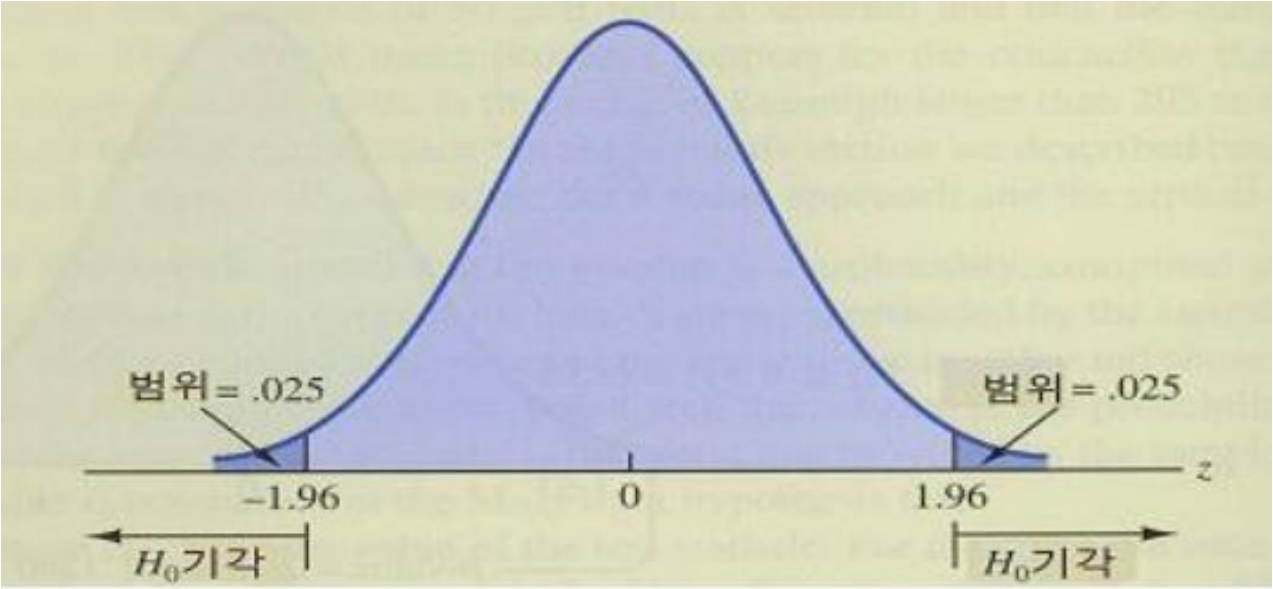
- P값이 유의수준 $\alpha = 0.05$보다 크기 때문에 귀무가설을 기각할 수 없다.

- 기각역으로 판정:

  $\alpha = 0.05$이고 양측검증이므로 기각역은 다음과 같다.

$$z_0 \leq -1.96 \ or \ z_0 \geq 1.96$$

Figure: 유의수준 5%하에서의 기각역



$z_0 = -1.35$이므로 기각역에 속하지 않는다. 즉 실제 농도가 0.85가 아니라고 의심할만한 뚜렷한 증거가 없다고 할 수 있다.

모평균에 대한 가설검정; σ를 알 경우

Example (5.6)

한 제약회사에서 생산하고 있는 진통제는 진통효과가 나타나는 시간이 평균 30분인 것으로 알려져 있다. 새로 개발한 진통제의 진통효과가 더 빠른가를 확인하기 위하여, 50명의 환자를 랜덤추출하여 시간을 관측한 결과 평균이 $\bar{x} = 28.5$(분)이었다. 새로운 진통제에 의한 진통효과가 나타나는 시간이 표준편차가 $\sigma = 5$(분)인 정규분포를 따른다고 하고, 적절한 가설을 유의수준 5%에서 검정하여라.

sol) 새로운 진통제의 효과가 나타나는 시간의 평균을 μ(분)라고 하면, 귀무가설과 대립가설은 다음과 같다.

$$H_0 : \mu = 30 \quad H_1 : \mu < 30$$

- 검증통계량의 관측값을 계산하면

$$z_0 = \frac{\bar{x} - 30}{5/\sqrt{50}} = \frac{28.5 - 30}{5/\sqrt{50}} = -2.12$$

- 기각역으로 판정:

  $\alpha = 0.05$이고 왼쪽 단측검증이므로 기각역은 다음과 같다.

$$z_0 \leq -z_{0.05} = -1.645$$

- $z_0 = -2.12$이므로 기각역에 속한다. 따라서 유의수준 5%에서 귀무가설을 기각한다.
- 즉 새로운 진통제의 효과가 더 빠르다는 증거가 지정된 수준 5%보다 강하게 나타났다.

모평균에 대한 가설검정; σ를 모를 때(t 검정)

- 모평균의 유의성검증(모표준편차 σ가 미지인 정규모집단)
  귀무가설이 $H_0 : \mu = \mu_0$인 경우에 검정통계량은

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

  이고, 이의 관측값을 t_0라고 할 때, 다음이 성립한다.

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \mu > \mu_0$	$P = P(T \geq t_0)$	$t_0 \geq t_\alpha(n-1)$
(b) $H_1 : \mu < \mu_0$	$P = P(T \leq t_0)$	$t_0 \leq -t_\alpha(n-1)$
(c) $H_1 : \mu \neq \mu_0$	$P = P(\lvert T \rvert \geq \lvert t_0 \rvert)$	$\lvert t_0 \rvert \geq t_{\alpha/2}(n-1)$

- - p값에 의한 판정: $P \leq \alpha$이면, H_0기각
  - 기각역에 의한 판정: 위의 기각역 조건이면 H_0기각

모평균에 대한 가설검정; σ를 모를 때(t 검정)

- σ를 모를 경우 μ에 대한 가설검정 유의사항

- σ를 모를 때의 가설검정은 모집단이 정규분포를 따라야 한다.
- 표본수 n이 커지면 자유도 $n-1$도 커지면서 t분포는 정규분포에 수렴한다. 따라서 표본수가 크면($n \geq 30$), σ를 모를 때라도 σ를 s로 대치하여 정규 분포(Z 통계량)를 사용하고 σ를 알 경우의 검정을 사용해도 큰 무리가 없다.

Table: 모평균에 대한 가설검정 정리

분포 모집단	σ 아는 경우	σ 모르는 경우(s사용)
정규분포	z 통계량	t 통계량[a]
대표본 비정규분포	근사 z 통계량	근사 z 통계량[b]

a: n이 크면 근사 Z 사용 가능
b: n이 크면 근사 T 사용 가능

모평균에 대한 가설검정; σ를 모를 때

Example

전구회사에서 생산하는 전구의 평균수명은 1950시간으로 알려져 있다. 새로이 개발중인 전구의 평균수명 μ가 기존의 전구보다 수명이 더 길다고 할 수 있는가를 확인하기 위하여, 9개의 시제품을 생산하여 그 수명시간을 조사한 결과가 다음과 같다.

2000, 1975, 1900, 2000, 1950, 1850, 1950, 2100, 1975

적절한 가설을 세우고, 수명의 분포가 정규분포라는 전제하에서 가설에 대한 유의수준5%의 검정을 하여라. 또한 유의확률은 얼마인가?

sol) 문제의 뜻에 적절한 귀무가설과 대립가설은 다음과 같다.

$$H_0 : \mu = 1950 \quad H_1 : \mu > 1950$$

- 주어진 자료에서 표본평균과 표본표준편차를 계산하면

$$\bar{x} = (2000 + 1975 + \cdots + 2100 + 1975)/9 = 1966.7$$

$$s = \sqrt{\{(2000 - 1966.7)^2 + \cdots + (1975 - 1966.7)^2\}/8} = 69.6$$

- 이로부터 t검정통계량의 관측값을 계산하면

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1966.7 - 1950}{69.6/\sqrt{9}} = 0.718$$

- 기각역으로 판정:
  오른쪽 단측검증이므로 기각역은 $t_0 \geq t_{0.05}(8) = 1.86$ 이고
  $t_0 = 0.718 < 1.86$이므로 귀무가설을 기각할 수 없다.

- P값으로 판정:

  유의확률은 다음의 확률을 계산하여야 한다.

$$P = P(T \geq 0.718), \;\; T(8)$$

- 그런데 t분포표에는 그 값이 나와있지 않고, 다음과 같이 0.718은 $t(8)$분포의 상방 25%백분위수(0.706)와 20%백분위수(0.889) 사이의 값임을 알 수 있다.

$$(0.25 =)P(T \geq 0.706) > P(T \geq 0.718) > P(T \geq 0.889)(= 0.2)$$

  따라서 유의확률이 항상 유의수준보다 큰 것을 알 수 있고, 실제로 계산해보면 유의확률은 $P = 0.247$이다.

모평균에 대한 가설검정; σ를 모를 때

Example

반도체의 특정한 용도에 사용되는 실리콘 다이오드는 0.6볼트의 가동전압이 요구되며, 이러한 목표에서 벗어나면 조정이 필요하다. 이러한 실리콘 다이오드를 제조하는 공장에서 120개를 랜덤추출하여 조사한 결과 가동전압(볼트)의 평균과 표준편차가 각각 다음과 같았다.

$$\bar{x} = 0.62, \quad s = 0.11$$

적절한 가설을 세우고, 유의수준 5%에서 가설을 검정하여라. 모평균에 대한 95% 신뢰구간도 구해 보자.

sol) 제조 공정에 의한 가동전압의 모평균을 μ볼트라고 하면, 귀무가설과 대립가설은 다음과 같다.

$$H_0 : \mu = 0.6 \quad H_1 : \mu \neq 0.6$$

- 이 경우에 표본크기가 $n = 120 > 30$이므로 정규모집단의 가정이 없어도 대표본근사에 의하여 t검정을 할 수 있다.
- t검정통계량의 관측값을 계산하면

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{0.62 - 0.6}{0.11/\sqrt{120}} = 1.992$$

- 또한, 표본크기 $n = 120 > 30$이므로 $\bar{X}$의 분포도 정규분포에 근접한다고 할 수 있으므로 $t_{0.025}(n-1)$대신 $z_{0.025} = 1.96$를 사용할 수 있다.

- 기각역으로 판정:

  $\alpha = 0.05$이므로 기각역은 $|t_0| \geq z_{0.025} = 1.96$이고,
  $|t_0| = 1.992 > 1.96$이므로 유의수준 5%에서 귀무가설을 기각한다.

- p값으로 판정:

  유의확률은 $P = P(|T| \geq 1.992)$이고 정규분포에 근사하므로 표준정규분포표로부터 그 값은 다음과 같고 귀무가설을 기각한다.

$$P \doteq 2 \times 0.0233 = 0.0466 < \alpha$$

- 95% 신뢰구간에서 $t_0.025(119) = 1.98$이므로,

$$\begin{aligned} \bar{x} \pm t_{0.025}(119)s/\sqrt{n} &= 0.62 \pm 1.98 \times 0.11/\sqrt{120} \\ &= (0.6001, 0.6399) \end{aligned}$$

신뢰구간과 가설검정

- 신뢰수준이 $100(1-\alpha)\%$인 신뢰구간을 이용하여, 유의수준 α, 즉 $100\alpha\%$의 양측검정 결과를 예측할 수 있다.
  - 모평균이 μ이고 모분산이 σ^2인 정규모집단에서 크기 n인 랜덤표본을 뽑았다고 하자.
  - μ에 대한 $100(1-\alpha)\%$ 신뢰구간은 다음과 같다.

$$\bar{x} \pm z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$$

어떤 b라는 값이 이 구간 안에 들어있으면 다음의 식이 성립한다.

$$|\bar{X}-b| < z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}, \qquad \frac{|\bar{X}-b|}{\sqrt{\frac{\sigma^2}{n}}} < z_{\alpha/2}$$

신뢰구간과 가설검정

- 이제 $H_0 : \mu = b$인 양측검정을 생각해 보자.
- 검정통계량:

$$Z = \frac{\bar{X} - b}{\sqrt{\frac{\sigma^2}{n}}}$$

(2)에 의해서 귀무가설은 유의수준 α에서 기각되지 못한다.

- 신뢰구간은 모수의 값으로서 기각되지 못하는 귀무가설의 집합이라고 볼 수 있다.

검정오류

- 검증오류(test error): 가설을 채택하거나 기각할 때 확률적으로 틀릴 가능성

- 제1종의 오류(type I error:α):

  귀무가설이 맞는데도 잘못하여 이를 기각할 확률
  (=검정의 유의수준 significance level)

$$P[reject\ H_0\ |H_0\ is\ true] = \alpha$$

- 제2종의 오류(type II error:β):

  대립가설이 사실임에도 불구하고 귀무가설을 기각하지 못하게 되는 확률

$$P[fail\ to\ reject\ H_0\ |H_0\ is\ false] = \beta$$

검정오류 예시

예) H_0: 도둑이 아니다. H_1: 도둑이다.

- α: 선량한 사람을 도둑으로 몰아 잡아들임
  $P[reject\ H_0\ |H_0\ is\ true]$
- β: 도둑을 도둑이 아니라고 놓아줌
  $P[fail\ to\ reject\ H_0\ |H_0\ is\ false]$

예) H_0: 임신이 아니다. H_1: 임신이다.

검정오류

- 귀무가설이 사실일 때, 이를 기각하지 않을 확률($=1-\alpha$)
- 검정력(power of the test): 대립가설 하의 평균값 $\mu = \mu_1$이 사실일 때, 귀무가설 H_0을 기각할 확률($=1-\beta$)

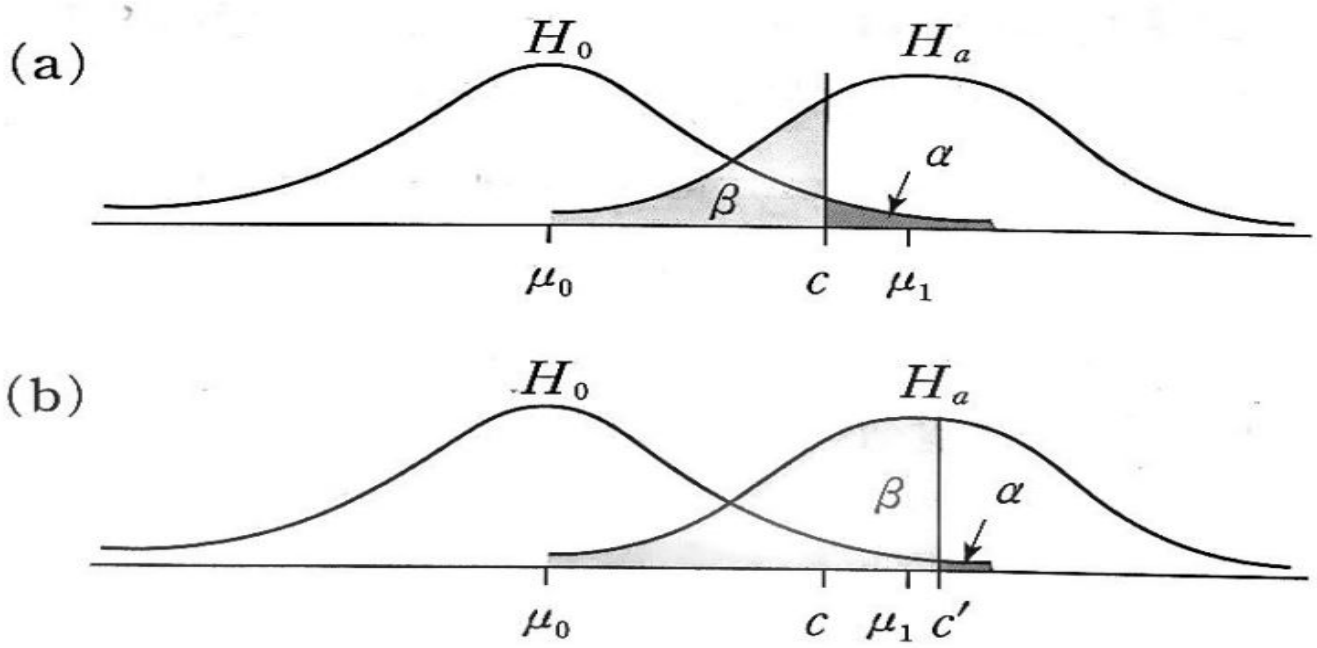
Table: 귀무가설의 진위와 의사결정

검정결과	실제상황	
	H_0가 사실	H_1이 사실
H_0를 기각못함	옳은 결정($1-\alpha$)	제2종의 오류(β)
H_0를 기각(H_1을 택함)	제1종의 오류(α)	옳은 결정($1-\beta$)

검정오류의 최소화

- 두 가지 검정오차인 α와 β를 최소로 하는 기각역을 구하는 것이 최선 => 그러나 α를 너무 작게 하려다 보면 β가 너무 커지는 모순 관계가 있음

  예) $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$이고, H_1의 대표값이 μ_1일 때, 임계점 c에 의하여 α와 β가 결정된다.

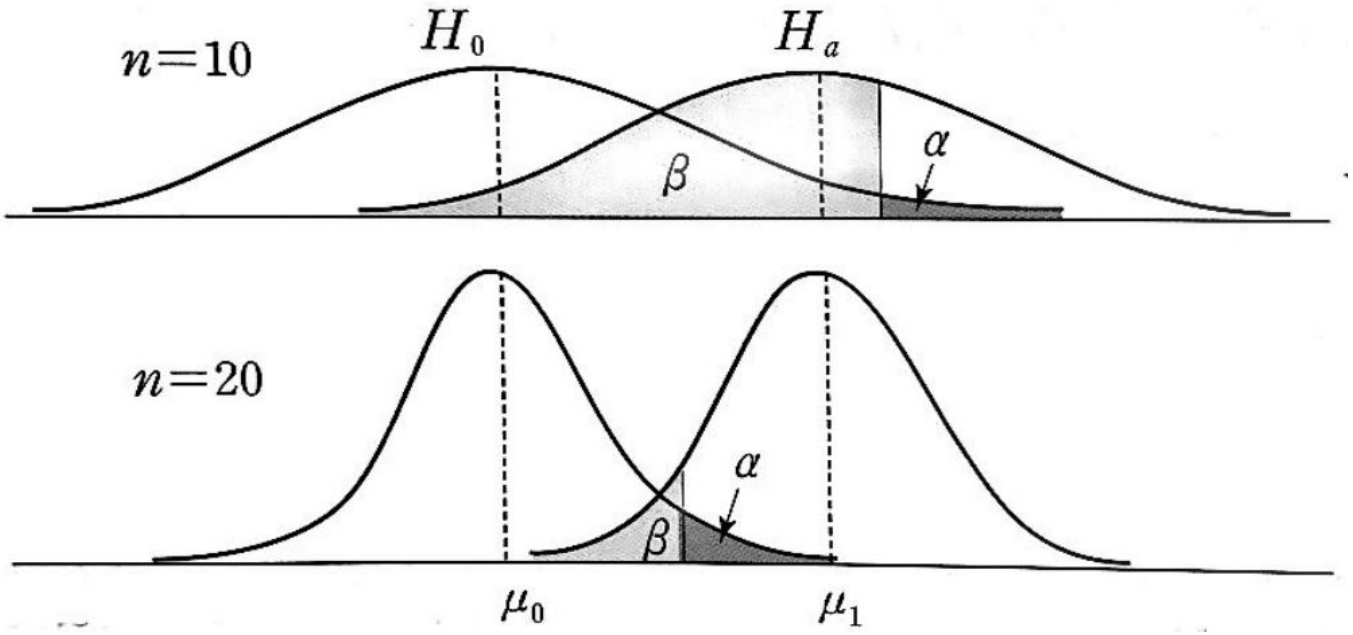


검정오류의 최소화

- 실제문제에선 주어진 α를 만족시키는 기각역중에서 β를 최소로 하는 기각역을 선택

  $\Longrightarrow (1-\beta)$가 클수록 검정력이 큼
- 표본수가 커지면 일반적으로 모수추정량의 분산이 작아지고

  고정된 α에 대하여 표본수 n이 증가함에 따라임계점이 왼쪽으로 이동하므로 β가 작아지고 검정력$(1-\beta)$이 커짐



검정오류 예제

Example

example 5.6(64p)에서 제1종의 오류를 범할 확률을 구하시오.

sol) 가설은 $H_0 : \mu = 30 \quad H_1 : \mu < 30$이고 왼쪽 단측검정이다.

- 유의수준 α의 기각역은 다음과 같다.

$$Z = \frac{\bar{X} - 30}{5/\sqrt{50}} \leq -z_\alpha$$

- 따라서 제1종의 오류를 범할 확률은 다음과 같고 이는 유의수준 α와 같음을 알 수 있다.

$$P\left\{\frac{\bar{X} - 30}{5/\sqrt{50}} \leq -z_\alpha | \mu = 30\right\} = \alpha$$

- 유의수준을 α로 지정한다는 것은 제1종의 오류를 범할 확률의 허용한계를 α로 제시한다는 것으로 해석할 수 있다.

검정오류 예제

Example

example 5.6에서 새로운 진통제의 효과가 나타나는 평균시간 μ가 28분이면 이는 괄목할 만한 개선이라고 한다. 이제 실제로 $\mu = 28$일 때 유의수준 5%의 기각역에 의한 제2종 오류를 범할 확률을 구하여 보자.

sol) 유의수준 5%의 기각역: $\frac{\bar{x}-30}{5/\sqrt{50}} \leq -1.645 = -z_{0.05}$

- 제2종오류 β는 $P[(fail\ to\ reject\ H_0\ |H_0\ is\ false]$이므로

$$\begin{aligned}
\beta &= P\left\{\frac{\bar{X}-30}{5/\sqrt{50}} > -1.645|\mu = 28\right\} \\
&= P\left\{\frac{\bar{X}-28}{5/\sqrt{50}} + \frac{28-30}{5/\sqrt{50}} > -1.645|\mu = 28\right\} \\
&= P\left\{\frac{\bar{X}-28}{5/\sqrt{50}} > 2\sqrt{2} - 1.645|\mu = 28\right\} \\
&= P\{Z > 1.183\} \doteq 0.1190
\end{aligned}$$

검정오류 예제

Example

example 5.6에서 실제로 $\mu = 28$일 때 제2종의 오류를 범할 확률이 $\beta = 0.10$이하가 되도록 하며, 유의수준이 $\alpha = 0.05$인 검정법을 사용하려고 한다. 이 때, 요구되는 표본의 크기를 구하여라.

sol) 요구되는 표본의 크기를 n이라고 하면, $\mu = 28$일 때 제2종의 오류를 범할 확률은 다음과 같이 주어진다.

$$\begin{aligned}
\beta &= P\left\{\frac{\bar{X} - 30}{5/\sqrt{n}} > -1.645 | \mu = 28\right\} \\
&= P\left\{\frac{\bar{X} - 28}{5/\sqrt{n}} + \frac{28 - 30}{5/\sqrt{n}} > -1.645 | \mu = 28\right\} \\
&= P\left\{Z > \frac{30 - 28}{5/\sqrt{n}} - 1.645\right\}
\end{aligned}$$

- 따라서, 이러한 확률을 0.10이하로 하는 다음의 식을 만족하는 n을 계산해보자.

$$P\left\{Z > \frac{30-28}{5/\sqrt{n}} - 1.645\right\} \leq 0.10$$

- 표준정규분포표로부터 $z_{0.10} = 1.282$이므로

$$\frac{30-28}{5/\sqrt{n}} - 1.645 \geq 1.282$$

$$n \geq \left(\frac{1.645+1.282}{(30-28)/5}\right)^2 = 53.7\cdots$$

- 그러므로 요구되는 표본의 크기는 $n = 54$이다.

검정오류 예제

Example

example 5.6에서 유의수준 5%의 검정법에 대하여 $\mu = 28$에서의 검정력을 구하여라.

sol) 검정력은 $1 - \beta$이므로 앞의 예제의 결과로부터

$$P\left\{\frac{\bar{x} - 30}{5/\sqrt{50}} \leq -1.645 | \mu = 28\right\} = 1 - 0.1190 = 0.8810$$

임을 알 수 있다.

모비율의 추정

- 불량률, 실업률, 지지율 등과 같이 모집단에서 특정한 속성을 갖는 개체의 비율을 뜻하는 모비율의 추정과 검정에 대하여 알아보자.
- 모비율이 p이고, 크기 n인 랜덤표본에서 그 속성을 갖는 것의 개수를 X라고 하면 $X \sim B(n, p)$.
- 모비율 p의 추정량: 표본비율 $\hat{p} = \frac{X}{n}$

$$E(\hat{p}) = \frac{E(X)}{n} = p, \quad Var(\hat{p}) = \frac{p(1-p)}{n}, \quad SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

- $\hat{p}$의 표준오차의 추정:

$$\widehat{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

모비율 p의 신뢰구간(n이 클 때)

- 이항분포의 정규분포 근사로부터 표본크기 n이 충분히 크면 X의 분포는 근사적으로 정규분포 $N(np, np(1-p))$에 가까워 지고, 이를 표준화 하면 다음과 같다.

$$X \sim N(np, np(1-p)) \quad \Longrightarrow \quad \frac{X - np}{\sqrt{np(1-p)}} \dot{\sim} N(0,1)$$

$$\frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \dot{\sim} N(0,1)$$

- 따라서 표본비율 $\hat{p}$은 근사적으로 다음과 같은 정규분포를 따른다.

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

모비율 p의 신뢰구간(n이 클 때)

- 정규분포의 성질에 의하여

$$P\left\{-z_{\alpha/2} \le \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \le z_{\alpha/2}\right\} = 1-\alpha$$

이므로 이를 이용하여 모비율 p의 신뢰구간을 구할 수 있다.

- 위의 식을 p에 관하여 정리하면 다음과 같이 쓸 수 있고,

$$p \in \{\hat{p} - z_{\alpha/2}\sqrt{p(1-p)/n},\ \ \hat{p} + z_{\alpha/2}\sqrt{p(1-p)/n}\}$$

- 이 때, 모수인 p의 값을 모르므로 $\hat{p}$으로 대신 사용한다.

모비율 p의 신뢰구간(n이 클 때)

- 모비율 p에 대한 $100(1-\alpha)\%$신뢰구간

$$(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n},\ \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n})$$

- $100(1-\alpha)\%$ 오차범위

$$z_{\alpha/2}\widehat{SE}(\hat{p}) = z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

- 표본크기 n이 크다는 것의 기준

$$n\hat{p} \geq 5,\ \text{and}\ n(1-\hat{p}) \geq 5$$

모비율의 추정 예제

Example (7.1)

서울에 거주하는 가구 중 돌아오는 추석에 귀향할 가구의 비율을 추정하기 위해 500가구를 단순랜덤추출하여 조사한 결과, 79가구가 귀향할 것이라고 응답했다. 서울에 거주하는 가구의 귀향비율 p에 대한 95% 신뢰구간을 구하라

sol) 귀향률 p의 추정량은 $\hat{p}$이고 추정값은 다음과 같다.

$$\hat{p} = 79/500 = 0.158$$

- 표준오차의 추정값:

$$\widehat{SE}(\hat{p}) = \sqrt{0.158(1-0.158)/500} = 0.016$$

- $n\hat{p} \geq 5,\ n(1-\hat{p}) \geq 5$가 만족되므로 귀향률 p의 95% 신뢰구간은

$$(0.158 - 1.96 \times 0.016, 0.158 + 1.96 \times 0.016) = (0.126, 0.190)$$

표본크기의 결정 - 모비율 p의 추정

- 모비율의 추정에서 $100(1-\alpha)\%$오차범위를 d이하로 하는 것이 요구된다면

$$z_{\alpha/2}SE(\hat{p}) \leq d, \ i.e., \ z_{\alpha/2}\sqrt{p(1-p)}/\sqrt{n} \leq d$$

가 만족되도록 표본크기 n이 정해져야 한다.

(a) 모비율의 사전 추정값이 p^*로주어진 경우

$$n \geq p^*(1-p^*)(z_{\alpha/2}/d)^2$$

(b) 모비율에 대한 사전 정보가 없는 경우

$$n \geq \frac{1}{4}(z_{\alpha/2}/d)^2$$

표본크기의 결정 예제

Example

어느 지역의 구청장 선거에 갑, 을 두 명의 후보가 입후보했을 때, 이들의 지지율을 알아보고자 한다. 아무런 사전 정보가 없는 상태에서 갑 후보의 지지율에 대한 추정의 95% 오차범위를 3%이내로 유지할 때, 필요한 표본의 크기를 구하라.

sol) $z_{0.025} = 1.96$이고, 모비율에 대한 사전 정보가 없는 경우이므로 표본크기 n은 다음을 만족해야 한다.

$$n \geq \frac{1}{4}(1.96/0.03)^2 = 1067.1$$

- 따라서 $n = 1068$명의 표본조사를 하여야 한다.

모비율에 대한 가설검정(표본크기가 큰 경우)

- $\hat{p}$과 가설의 비율 p_0이 얼마나 차이가 있느냐에 따름
- 귀무가설이 $H_0 : p = p_0$인 경우에, 검정통계량은 (87 page)

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \dot{\sim} N(0,1)$$

이고 이의 관측값을 z_0라고 할 때, 다음이 성립한다.

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : p > p_0$	$P = P(Z \geq z_0)$	$z_0 \geq z_\alpha$
(b) $H_1 : p < p_0$	$P = P(Z \leq z_0)$	$z_0 \leq -z_\alpha$
(c) $H_1 : p \neq p_0$	$P = P(\lvert Z\rvert \geq \lvert z_0\rvert)$	$\lvert z_0\rvert \geq z_{\alpha/2}$

- 표본크기 n이 크다는 것의 기준:$np_0 \geq 5,\ and\ n(1-p_0) \geq 5$

모비율에 대한 가설검정

Example

Example 7.1에서 과거 추석 때의 귀향률이 20%였다고 하자. 표본 조사 결과에 따르면 돌아오는 추석에 서울에 거주하는 가구의 귀향률(500가구 중 79가구)이 과거에 비해 줄어들었다고 할 수 있는지를 유의수준 1%수준에서 검정하여라.

sol) 돌아오는 추석 때의 귀향률을 p라고 하면 가설은 다음과 같다.

$$H_0 : p = 0.2, \quad H_1 : p < 0.2$$

- $np_0 = 100 \geq 5$, $n(1-p_0) = 400 \geq 5$이므로 표본크기는 충분히 크다고 할 수 있고
- 표본크기 $n = 500$ 중 귀향하려는 가구의 수는 $X = 79$이므로, 검정통계량의 관측값은

$$z_0 = \frac{79/500 - 0.2}{\sqrt{0.2 \times 0.8/500}} = -2.348$$

- 왼쪽 단측검증이므로 유의수준 1%의 기각역은 다음과 같고 관측된 z_0는 이 기각역에 속한다.

$$z_0 \leq -z_{0.01} = -2.327$$

- 따라서, 유의수준 1%에서 귀무가설을 기각할 수 있다.
- 즉, 표본조사 결과에 의하면 귀향률이 과거보다 낮아질 증거가 뚜렷하다고 할 수 있다.

모분산에 관한 추론

- 모집단의 변동성(variability) 또는 퍼짐(spread)의 정도에 관심이 있는 경우, 모분산이 추론의 대상이 됨.

- 예) 볼베어링을 일정한 규격으로 생산하는 공정에서 지름의 변동성이 크다면, 평균적으로 규격에 맞더라도 불량품이 많아지는 문제가 발생한다.

  => 이때 베어링 지름의 변동성을 확인하는 방법으로 모분산에 대한 추론을 할 수 있다.

- 이러한 추론은 모분산 σ^2의 추정량인 다음의 표본분산을 이용하여 할 수 있다. ($E(S^2) = \sigma^2$임)

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)}$$

표본분산의 분포

Theorem (정규모집단에서의 표본분산의 분포)

$X_1, \cdots, X_n$을 정규분포 $N(\mu, \sigma^2)$으로부터의 랜덤표본이라 할 때, 표본분산 $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$에 대하여 다음이 성립한다.

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

- 가정:
  - 모집단이 정규분포를 따름
  - 단순임의추출을 해야 함
- 위의 표본분산의 분포를 이용하여 σ^2에 관한 추정이나 유의성검증을 할 수 있다.

정규모집단에서의 표본분산의 분포

pf) $X_1, \cdots, X_n$이 정규분포 $N(\mu, \sigma^2)$의 랜덤표본일 때,

$$\frac{X_i - \mu}{\sigma} \sim N(0,1)$$

이고 서로 독립$(i = 1, \cdots, n)$이므로
카이제곱분포의 정의로부터 다음이 성립한다.

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n)$$

그런데 다음의 식에서 $S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1)$이므로

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 + n \left(\frac{\bar{X} - \mu}{\sigma} \right)^2$$

정규모집단에서의 표본분산의 분포

다음과 같이 바꿔 쓸 수 있다.

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2=\frac{(n-1)S^2}{\sigma^2}+\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$$

여기에서 다음의 정의에 의해

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\sim\chi^2(n),\quad\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2\sim\chi^2(1)$$

카이제곱분포의 가법성을 이용하면 다음이 성립함을 예상해 볼 수 있다.

$$\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$$

카이제곱 분포

Definition (카이제곱분포:chi-squared distribution)

확률변수 $Z_1, \cdots, Z_k$이 표준정규분포 $N(0,1)$의 랜덤표본일 때,

$$Z_1^2 + \cdots + Z_k^2$$

의 분포를 자유도(degrees of freedom)가 k인 카이제곱분포라고 한다.
이 때 기호로서

$$Z_1^2 + \cdots + Z_k^2 \sim \chi^2(k)$$

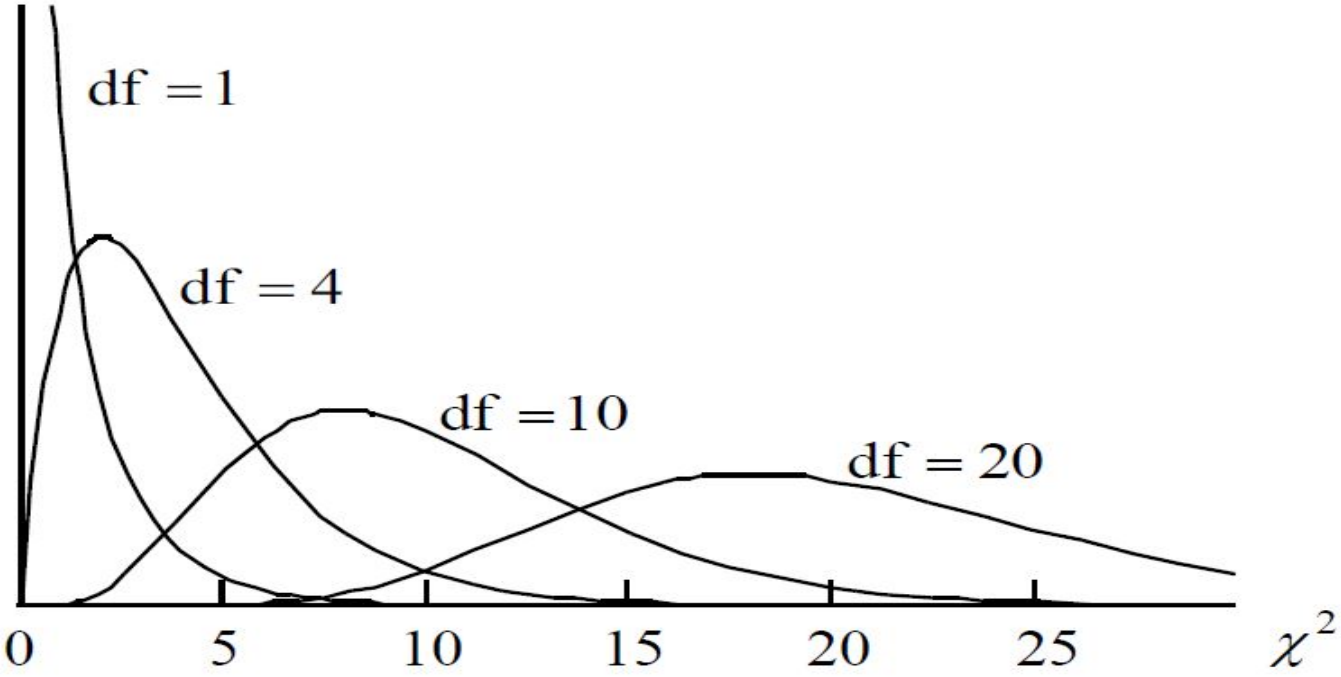
로 나타낸다.

- 카이제곱분포는 다음의 표본분산 S^2에 관계되는 분포이다.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

카이제곱 분포

- 자유도의 변화에 따른 여러가지 카이제곱분포의 모양은 다음과 같다.



카이제곱 분포의 분위수

- 자유도가 k인 카이제곱분포의 $(1-\alpha)$ 분위수를 흔히 $\chi^2_\alpha(k)$로 나타낸다. 즉, $V \sim \chi^2(k)$일 때 다음과 같다.

$$P\{V \geq \chi^2_\alpha(k)\} = \alpha$$

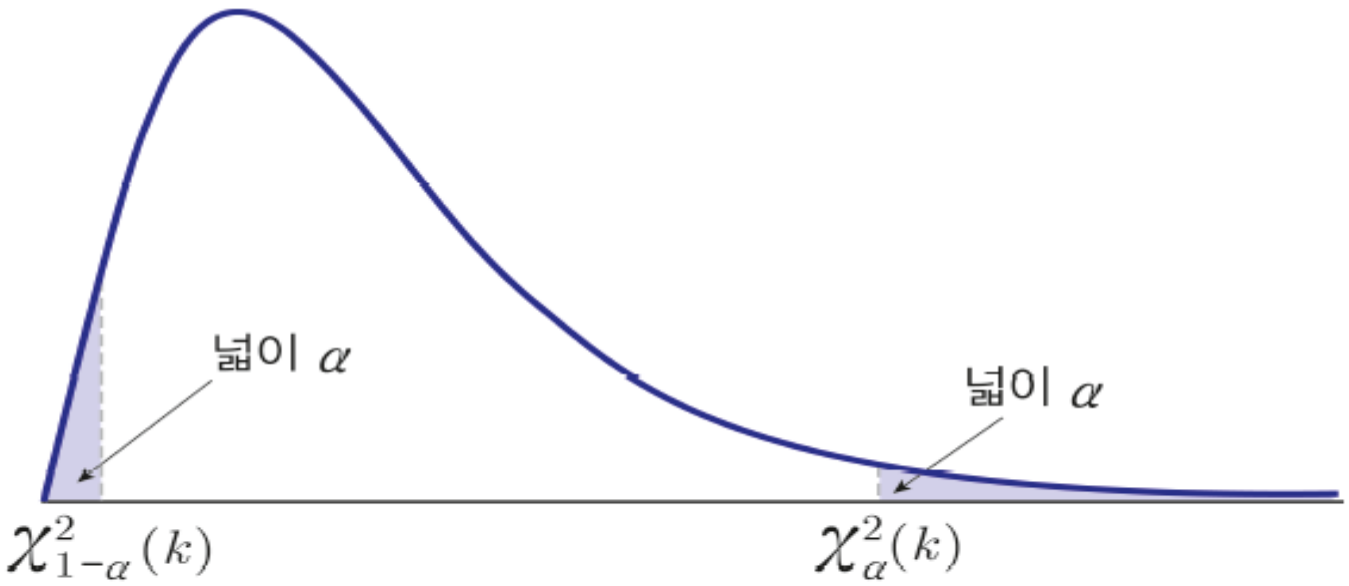
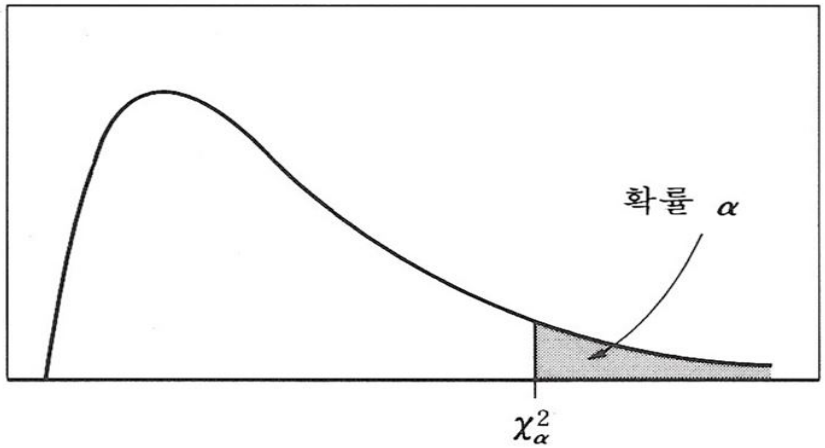


그림 4–7 카이제곱분포의 형태

카이제곱분포표

- 아래 여러가지 자유도에 대한 카이제곱분포표를 이용하여 분위수를 구할 수 있다.

  예) $\chi^2_{0.995}(4) = 0.207$, $\chi^2_{0.975}(4) = 0.484$, $\chi^2_{0.05}(4) = 9.49$



표의 숫자는 주어진 α값에 대한 χ^2_α의 값을 나타낸다.

df	α							
	.995	.990	.975	.950	.050	.025	.010	.005
1	392704×10^{-10}	$157088\times^{-9}$	$982069\times^{-9}$	$393214\times^{-8}$	3.84146	5.02389	6.63490	7.87944
2	.0100251	.0201007	.0506356	.102587	.99147	7.37776	9.21034	10.5966
3	.0717212	.114832	.215795	.351846	7.81473	9.34840	11.3449	12.8381
4	.206990	.297110	.484419	.710721	9.48773	11.14433	13.2767	14.8602
5	.411740	.554300	.831211	1.145476	11.0705	12.8325	15.0863	16.7496

카이제곱분포의 성질 1

Theorem (카이제곱분포의 가법성)

$V_1 \sim \chi^2(k_1)$, $V_2 \sim \chi^2(k_2)$이고, V_1과 V_2가 서로 독립이면 다음이 성립한다.

$$V_1 + V_2 \sim \chi^2(k_1 + k_2)$$

- 즉, 두 확률변수가 서로 독립이고 각각 카이제곱분포를 따를 때 이들의 합도 카이제곱분포를 따른다.

카이제곱분포의 성질 2

- 자유도가 k인 카이제곱분포를 따르는 확률변수 V의 기댓값과 분산은 각각 다음과 같다.

$$V \sim \chi^2(k) \implies E(V) = k, \ Var(V) = 2k$$

- 표본분산 S^2은 모분산 σ^2의 불편추정량이다. $(E(S^2) = \sigma^2)$

pf.1) $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ 이므로 $E\left(\frac{(n-1)S^2}{\sigma^2}\right) = n-1$이다.

따라서 $E(S^2) = \sigma^2$.

pf.2) $\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2$ 이므로

$$\begin{aligned} E((n-1)S^2) &= E(\textstyle\sum_{i=1}^n (X_i - \bar{X})^2) \\ &= E(\textstyle\sum_{i=1}^n (X_i - \mu)^2) - E(n(\bar{X} - \mu)^2) \\ &= n\sigma^2 - n\frac{\sigma^2}{n} = (n-1)\sigma^2 \end{aligned}$$

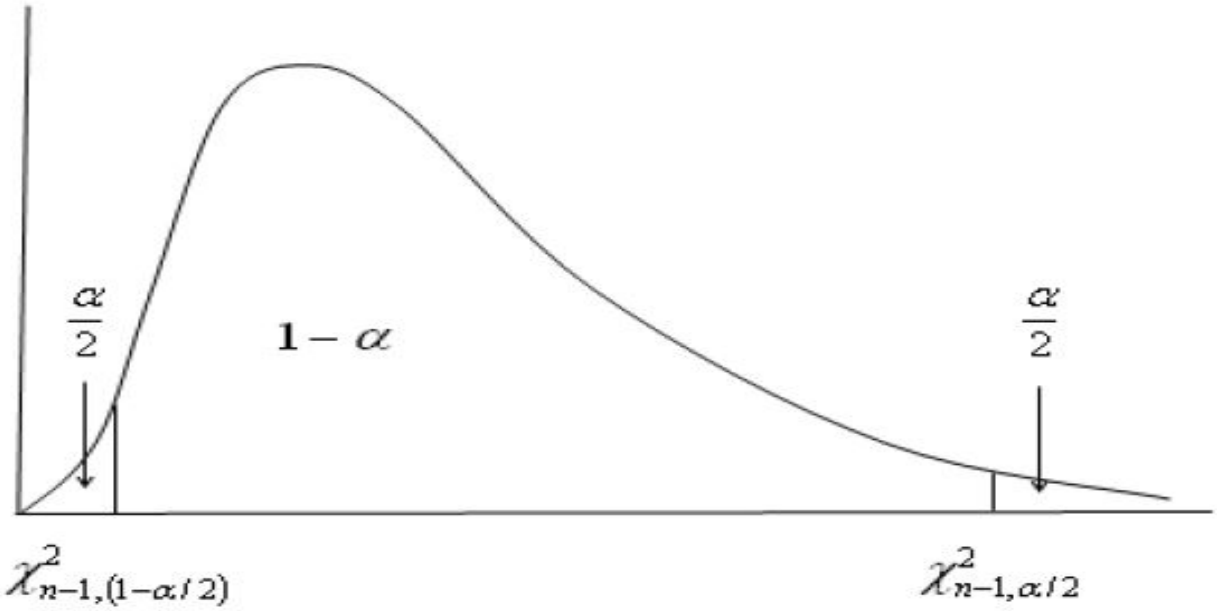
따라서 $E(S^2) = \sigma^2$.

모분산의 신뢰구간

- 표본분산 S^2의 표본분포는 $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$이므로
- $\chi^2(n-1)$의 백분위수인 $\chi^2_{1-\alpha/2}(n-1)$, $\chi^2_{\alpha/2}(n-1)$에 대하여

$$1-\alpha = P\{\chi^2_{1-\alpha/2}(n-1) \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{\alpha/2}(n-1)\}$$

Figure: $\chi^2(n-1)$분포의 유의수준 α의 신뢰구간



모분산의 신뢰구간

- σ^2으로 정리하면

$$1-\alpha = P\left\{\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right\}$$

이므로 모분산 σ^2에 대한 $100(1-\alpha)\%$신뢰구간은 다음과 같이 주어진다.(정규모집단의 경우)

$$\left\{\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}\right\}$$

모분산의 가설검정(정규모집단의 경우)

- 귀무가설 $H_0 : \sigma^2 = \sigma_0^2$인 경우에 검정통계량은

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

이고, 이의 관측값을 χ_0^2라고 할 때 다음이 성립한다.

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \sigma^2 > \sigma_0^2$	$P = P(\chi^2 \geq \chi_0^2)$	$\chi_0^2 \geq \chi_\alpha^2(n-1)$
(b) $H_1 : \sigma^2 < \sigma_0^2$	$P = P(\chi^2 \leq \chi_0^2)$	$\chi_0^2 \leq \chi_{1-\alpha}^2(n-1)$
(c) $H_1 : \sigma^2 \neq \sigma_0^2$	$P = 2P(\chi^2 \geq \chi_0^2)$ or	$\chi_0^2 \geq \chi_{\alpha/2}^2(n-1)$ or
	$P = 2P(\chi^2 \leq \chi_0^2)$	$\chi_0^2 \leq \chi_{1-\alpha/2}^2(n-1)$

- 양측가설인 경우 $2P(\chi^2 \geq \chi_0^2)$ or $2P(\chi^2 \leq \chi_0^2)$ 중 1보다 작은 값을 유의확률로 계산

모분산의 가설검정(정규모집단의 경우)

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \sigma^2 > \sigma_0^2$	$P = P(\chi^2 \geq \chi_0^2)$	$\chi_0^2 \geq \chi_\alpha^2(n-1)$
(b) $H_1 : \sigma^2 < \sigma_0^2$	$P = P(\chi^2 \leq \chi_0^2)$	$\chi_0^2 \leq \chi_{1-\alpha}^2(n-1)$
(c) $H_1 : \sigma^2 \neq \sigma_0^2$	$P = 2P(\chi^2 \geq \chi_0^2)$ or $P = 2P(\chi^2 \leq \chi_0^2)$	$\chi_0^2 \geq \chi_{\alpha/2}^2(n-1)$ or $\chi_0^2 \leq \chi_{1-\alpha/2}^2(n-1)$

모분산의 가설검정

Example

한 카페에서는 아메리카노 한 잔에 원두를 10g을 사용하고 있다. 그러나 입력한 중량만큼 원두를 갈아내는 기계는 항상 정확하지 않으며, 약간의 변동이 발생할 수 밖에 없다. 이러한 변동이 너무 크면 커피 맛과 품질이 고객의 기대에 부합하지 않을 수 있으므로 중량의 표준편차 1.5g을 초과하면 기계에 이상이 있다고 간주한다. 어느 날 점검에서 무작위로 기계를 10번 작동하여 측정한 간 원두의 무게가 다음과 같이 주어졌다.

(단위 : g)

11.15	9.62	9.83	10.85	12.66	9.51	11.23	9.38	9.11	8.56

기계에 이상이 있는지를 유의수준 5%에서 검정하고, 원두 무게의 표준편차에 대한 90% 신뢰구간을 구하라. (과거 관리 기록에 따르면 원두 무게의 분포는 정규분포라고 가정할 수 있다.)

sol) 아메리카노 한 잔에 들어가는 원두 중량의 표준편차를 σ라고 하면 검정하려는 가설은 다음과 같다.

$$H_0 : \sigma^2 = 1.5^2 \quad H_1 : \sigma^2 > 1.5^2$$

- 주어진 자료에서 평균과 표준편차를 구하면 각각 다음과 같다.

$$\bar{x} = 10.19, \quad s = 1.24$$

- 이로부터 카이제곱검정통계량의 관측값을 계산하면

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{9(1.24)^2}{1.5^2} = 6.15$$

- p값 유의확률은 다음과 같다.

$$P = P\{\chi^2 \geq 6.15\}, \quad \chi^2 \sim \chi^2(9)$$

- 통계 소프트웨어를 이용하면 실제 유의확률은 0.72이므로 귀무가설을 기각할 수 없다.

- 기각역으로 판정:

  기각역은 $\chi_0^2 \geq \chi_{0.05}^2(9) = 16.92$이고, $\chi_0^2 = 6.15 \leq 16.92$이므로 기각역에 속하지 않아 유의수준 5%에서 귀무가설을 기각할 수 없다.
- 즉, 원두 추출량의 변동에 이상이 있다는 증거가 뚜렷하지 않다.

- σ^2에 대한 90% 신뢰구간은 $n = 10$, $\chi_{0.05}^2(9) = 16.92$, $\chi_{0.95}^2(9) = 3.33$이므로

$$\left(\frac{9(1.24)^2}{16.92} \leq \sigma^2 \leq \frac{9(1.24)^2}{3.33}\right) = (0.818 \leq \sigma^2 \leq 4.156)$$

따라서 모표준편차 σ에 대한 90% 신뢰구간은 다음과 같다.

$$(\sqrt{0.818}, \sqrt{4.156}) = (0.9,\ 2.04)$$

6장. 두 모집단의 비교에 관한 추론

박명현

서울대학교 학부대학

두 모집단의 비교에 관한 추론

- 두 모집단의 평균에 대한 추론
  - 대응비교에 의한 모평균의 비교
- 두 모집단의 비율에 대한 추론
- 두 모집단의 분산에 대한 추론
  - F분포

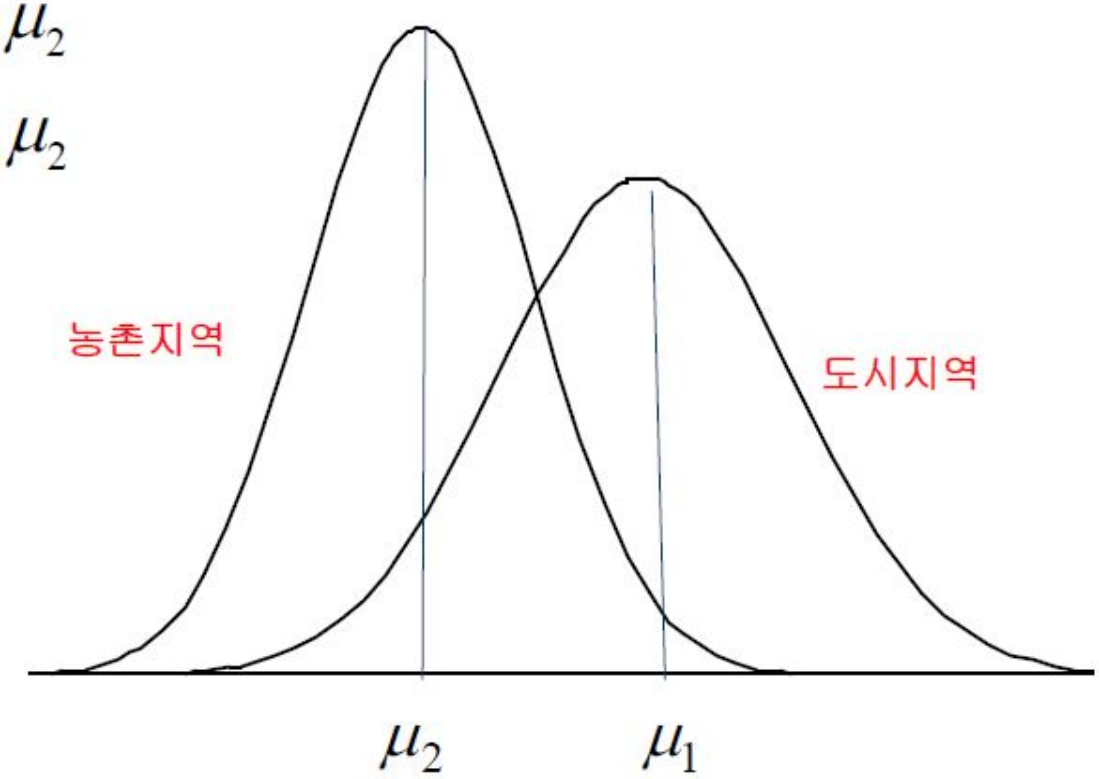
두 모집단의 모평균의 비교

예) 국가에서 국민의 삶의 질을 향상시키기 위한 정책을 내놓기 위해 현재 계층 간, 지역 간 삶의 질의 차이에 대한 관찰을 하고자 함.

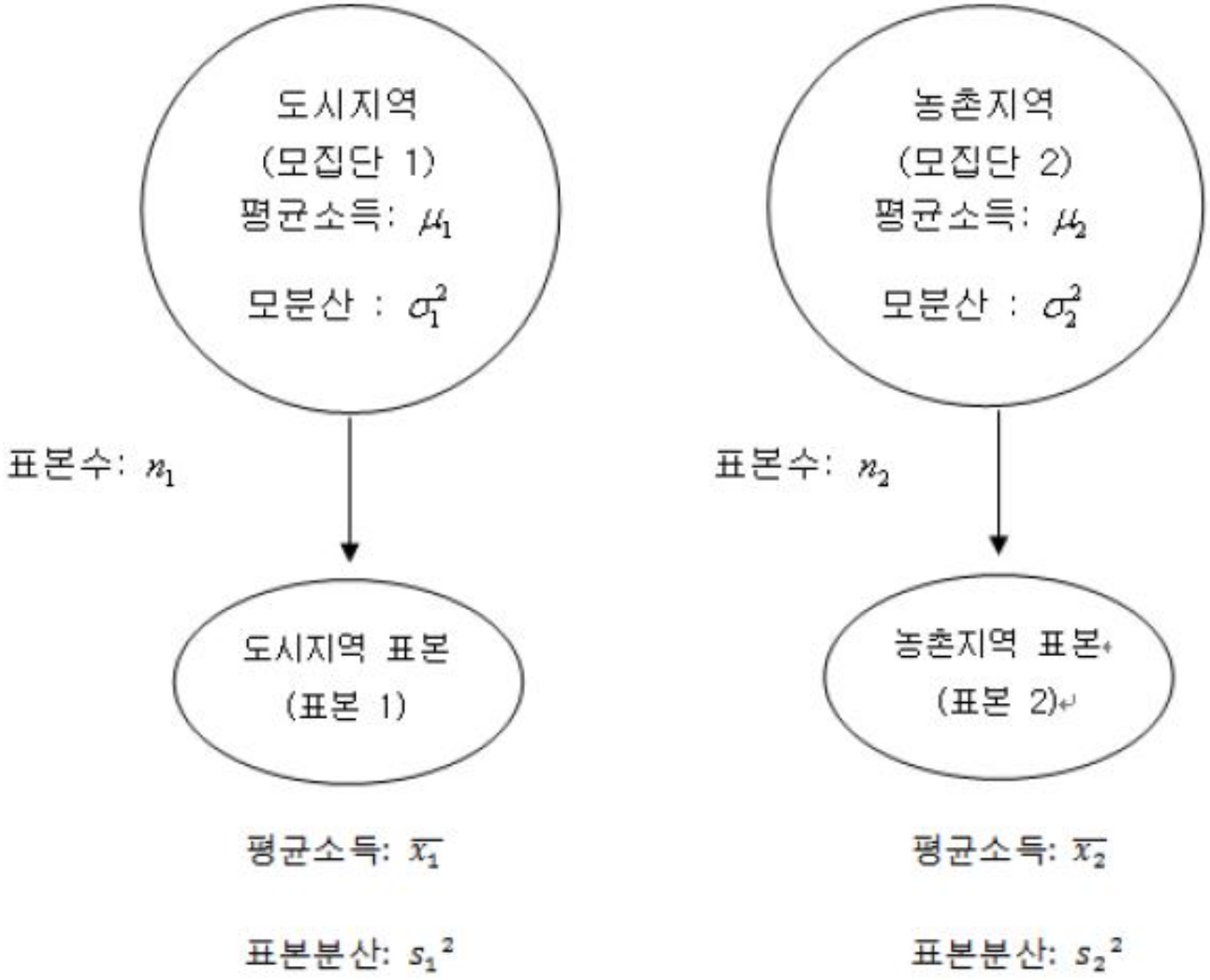
- 도시/농촌지역의 소득에 차이가 있는지 평균비교를 할 수 있다.

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$



두 모집단의 모평균의 비교



- 표본1: $\{X_{11}, \cdots, X_{1n_1}\}$, 표본2: $\{X_{21}, \cdots, X_{2n_2}\}$

두 모집단의 모평균의 비교

- 두 모집단의 평균 비교에 필요한 가정
  (1) 각 그룹에서의 관측값들은 각 모집단에서의 랜덤표본이다.
  (2) 서로 다른 그룹에서의 관측값들은 독립적으로 관측된 것이다.
  (3) 두 모집단은 각각 정규분포를 따른다.
  : 정규분포를 따르지 않더라도 표본이 클 경우 CLT 적용 가능.
- 서로 다른 두 모집단에서 각각 랜덤추출된 두 표본 => 위의 가정사항을 만족
- 주어진 실험단위를 대상으로 두 가지의 처리를 적용하여 비교하는 경우 => 전체 실험단위를 랜덤하게 두 그룹으로 나누어 각 그룹에 별로의 처리를 적용해야 위의 가정사항을 만족할 수 있다.
- 랜덤화(randomization): 각 처리를 적용할 실험단위를 랜덤하게 정하는 과정

$\bar{X}_1 - \bar{X}_2$의 표본분포

- $\mu_1 - \mu_2$를 표본평균의 차인 $\bar{X}_1 - \bar{X}_2$로 추정한다.

$$\widehat{\mu_1 - \mu_2} = \bar{x}_1 - \bar{x}_2$$

- $\bar{X}_1 - \bar{X}_2$의 기대값

$$E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$$

- $\bar{X}_1 - \bar{X}_2$의 분산과 표준오차

$$\begin{aligned} Var(\bar{X}_1 - \bar{X}_2) &= Var(\bar{X}_1) + Var(\bar{X}_2) - 2Cov(\bar{X}_1, \bar{X}_2) \\ &= Var(\bar{X}_1) + Var(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \end{aligned}$$

$$\sqrt{Var(\bar{X}_1 - \bar{X}_2)} = \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$\bar{X}_1 - \bar{X}_2$의 표본분포

- 정규 모집단이면 $(\bar{X}_1 - \bar{X}_2)$도 정규분포를 따르고,
  비정규 모집단이더라도 표본수가 크면 근사적으로 $(\bar{X}_1 - \bar{X}_2)$는 정규분포에 가까워진다.(by, CLT)
- 따라서 $\bar{X}_1 - \bar{X}_2$의 표본분포는 (근사적으로) 다음과 같고,

$$(\bar{X}_1 - \bar{X}_2) \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

이를 표준화 하면 다음과 같이 (근사적으로) 표준정규분포를 따르게 된다.

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1) \tag{1}$$

두 모평균의 차이에 대한 추론(σ_1^2, σ_2^2를 아는 경우)

I. σ_1^2, σ_2^2를 아는 경우

1. 두 모집단의 모분산 값이 알려진 경우
2. 정규분포 모집단 가정. 아닌 경우에도 대표본 근사 가능한 경우

- $\mu_1 - \mu_2$에 대한 $100(1-\alpha)\%$신뢰구간:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

두 모평균의 차이에 대한 추론(σ_1^2, σ_2^2를 아는 경우)

- 귀무가설: $H_0 : \mu_1 = \mu_2 + a \quad or \quad H_0 : \mu_1 - \mu_2 = a$
- 검정통계량:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

- 검정통계량의 관측값을 z_0라고 할 때

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \mu_1 - \mu_2 > a$	$P = P(Z \geq z)$	$z_0 \geq z_\alpha$
(b) $H_1 : \mu_1 - \mu_2 < a$	$P = P(Z \leq z)$	$z_0 \leq -z_\alpha$
(c) $H_1 : \mu_1 - \mu_2 \neq a$	$P = P(\lvert Z \rvert \geq \lvert z \rvert)$	$\lvert z_0 \rvert \geq z_{\alpha/2}$

- - p값에 의한 판정: $P \leq \alpha$이면, H_0기각
  - 기각역에 의한 판정: 위의 기각역 조건이면 H_0기각

두 모평균의 차이에 대한 추론(등분산, $\sigma_1^2 = \sigma_2^2$)

II. σ_1^2, σ_2^2를 모르는 경우

1. 두 모분산이 같다고 가정할 때: $\sigma_1^2 = \sigma_2^2 = \sigma^2$(등분산 가정)
2. 정규분포 모집단 가정

- 합동표본분산(pooled sample variance): S_p^2

  S_p^2은 공통 모분산 σ^2의 추정량으로, S_1^2, S_2^2의 자유도인 $n_1 - 1$과 $n_2 - 1$의 가중평균의 형태로서 다음과 같이 정의된다.

$$
\begin{aligned}
S_p^2 &= \frac{(n_1 - 1)}{(n_1 - 1) + (n_2 - 1)} S_1^2 + \frac{(n_2 - 1)}{(n_1 - 1) + (n_2 - 1)} S_2^2 \\
&= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}
\end{aligned}
$$

두 모평균의 차이에 대한 추론(등분산, $\sigma_1^2 = \sigma_2^2$)

- $\bar{X}_1 - \bar{X}_2$의 분산 추정

$$Var(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$$

$$\widehat{Var}(\bar{X}_1 - \bar{X}_2) = \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2} = \hat{\sigma}^2(\frac{1}{n_1} + \frac{1}{n_2}) = S_p^2(\frac{1}{n_1} + \frac{1}{n_2})$$

- $\bar{X}_1 - \bar{X}_2$를 추정된 분산으로 표준화한 통계량은 자유도 $(df = n_1 + n_2 - 2)$인 t분포를 따르게 됨(증명)

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

두 모평균의 차이에 대한 추론(등분산, $\sigma_1^2 = \sigma_2^2$)

- 등분산 정규모집단 $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$에서의 t분포 증명
- $[(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)] / \left(\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) \sim N(0,1)$,
  $\frac{(n_1+n_2-2)S_p^2}{\sigma^2} \sim \chi^2(n_1+n_2-2)$이며, 이들은 서로 독립이다.(p.72)
- 따라서, t분포의 정의로부터 다음이 성립하고

$$\frac{[(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)] / \left(\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)}{\sqrt{\frac{(n_1+n_2-2)S_p^2}{\sigma^2} / (n_1+n_2-2)}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- 자유도가 $n_1 + n_2 - 2$인 다음의 t분포를 따른다고 할 수 있다.

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

두 모평균의 차이에 대한 추론(등분산, $\sigma_1^2 = \sigma_2^2$)

- $\mu_1 - \mu_2$에 대한 $100(1-\alpha)\%$신뢰구간

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}(n_1 + n_2 - 2) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- 귀무가설이 $H_0 : \mu_1 - \mu_2 = a$인 경우에 검정통계량은

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

- 검정통계량의 관측값을 t_0라고 할 때

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \mu_1 - \mu_2 > a$	$P = P(T \geq t)$	$t_0 \geq t_\alpha(n_1 + n_2 - 2)$
(b) $H_1 : \mu_1 - \mu_2 < a$	$P = P(T \leq t)$	$t_0 \leq -t_\alpha(n_1 + n_2 - 2)$
(c) $H_1 : \mu_1 - \mu_2 \neq a$	$P = P(\lvert T \rvert \geq \lvert t \rvert)$	$\lvert t_0 \rvert \geq t_{\alpha/2}(n_1 + n_2 - 2)$

두 모평균의 차이에 대한 추론 예제

Example

한 의학 연구에서 비만 치료를 위한 GLP-1(Glucagon-Like Peptide-1) 작용제인 세마글루타이드(semaglutide)의 체중 감소 효과를 다음과 같이 분석했다. 건강 상태가 비슷한 지원자 12명을 무작위로 6명씩 두 그룹으로 나누고 한 그룹에는 세마글루타이드를, 다른 그룹에는 위약을 투여했다. 각 그룹은 12주 동안 주 1회 피하주사를 맞았으며, 식습관 운동 등 생활습관은 비슷하게 유지하도록 통제했다. 12주가 지난 시점에서 각 지원자의 체중 감소량을 측정했고, 결과는 다음 표와 같다.

(단위 : kg)

세마글루타이드군	5.12	4.87	5.64	5.01	5.49	5.25
위약군	3.62	3.98	4.14	3.25	4.42	3.60

두 집단의 평균 체중 감소량에 차이가 있는지를 유의수준 5%에서 검정하고, 평균 체중 감소량 대한 95% 신뢰구간을 구하라. 단, 두 집단의 분산은 같고 정규모집단을 가정한다.

sol) 두 집단의 평균 체중 감소량을 각각 μ_1, μ_2라 하면, 검정하려는 가설은 다음과 같다.

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$

- 주어진 자료에서 평균과 표준편차를 구하면 다음과 같다.

	세마글루타이드군	위약군
표본크기(n_i)	6	6
평균($\bar{x}_i$)	5.23	3.84
표준편차(si)	0.29	0.42

- 이를 이용하여 합동표본분산을 구하면

$$S_p^2 = \frac{(6-1)0.29^2 + (6-1)0.42^2}{6+6-2} = 0.13025, \quad S_p = 0.361$$

- 따라서 t검정통계량의 관측값을 계산하면

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(5.32 - 3.84) - 0}{0.361\sqrt{\frac{1}{6} + \frac{1}{6}}} = 6.67$$

- t분포표에서 $t_{0.025}(10) = 2.228$
- 기각역으로 판정: 기각역은 $|t| \geq t_{0.025}(10) = 2.228$이고, $|t_0| > 2.228$이므로 유의수준 5%에서 귀무가설을 기각할 수 있다.
- p값으로 판정: $P = P(|t| \geq 6.67)$이고 t분포표에서 $P < 0.001$임을 알 수 있다.
- 지금까지의 내용을 종합하면 세마글루타이드와 위약군 간의 평균 체중 감소량이 다르다는 증거가 충분하다.
- $\mu_1 - \mu_2$에 대한 95%신뢰구간을 구하면

$$
\begin{aligned}
&(\bar{x}_1 - \bar{x}_2) \pm t_{0.025}(10) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\
&= (5.23 - 3.84) \pm t_{0.025}(10) 0.361 \sqrt{\frac{1}{6} + \frac{1}{6}} \\
&= 1.39 \pm 0.46 = (0.93, 1.85)
\end{aligned}
$$

두 모평균의 차이에 대한 추론(등분산, 대표본의 경우)

II. σ_1^2, σ_2^2를 모르는 경우

1. $\sigma_1^2 = \sigma_2^2 = \sigma^2$(등분산 가정)
2. but, 모집단의 분포가 정규분포가 아닌 경우

- 모집단의 분포에 관계없이 대표본의 경우 다음의 통계량은 중심극한정리에 따라 표준정규분포를 따르게 된다.

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \dot{\sim} N(0,1)$$

- 대표본의 조건: $n_1 \geq 30, n_2 \geq 30$

두 모평균의 차이에 대한 추론(등분산, 대표본의 경우)

- $\mu_1 - \mu_2$에 대한 $100(1-\alpha)\%$신뢰구간

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- 귀무가설이 $H_0 : \mu_1 - \mu_2 = a$인 경우에 검정통계량은

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- 검정통계량의 관측값을 z_0라고 할 때

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \mu_1 - \mu_2 > a$	$P = P(Z \geq z)$	$z_0 \geq z_\alpha$
(b) $H_1 : \mu_1 - \mu_2 < a$	$P = P(Z \leq z)$	$z_0 \leq -z_\alpha$
(c) $H_1 : \mu_1 - \mu_2 \neq a$	$P = P(\lvert Z \rvert \geq \lvert z \rvert)$	$\lvert z_0 \rvert \geq z_{\alpha/2}$

두 모평균의 차이에 대한 추론(이분산, $\sigma_1^2 \neq \sigma_2^2$)

II. σ_1^2, σ_2^2를 모르는 경우

1. 두 모분산이 다르다고 가정할 때: $\sigma_1^2 \neq \sigma_2^2$(이분산 가정)
2. 정규분포 모집단 가정

- $\bar{X}_1 - \bar{X}_2$의 분산: 이분산의 경우 두 표본평균간의 독립성에 의해

$$Var(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

- 분산의 추정치:

$$\widehat{Var}(\bar{X}_1 - \bar{X}_2) = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

- 모집단 정규분포이고 표본크기 n_1, n_2가 5 이상이면 $\bar{X}_1 - \bar{X}_2$를 추정된 분산으로 표준화한 통계량은 자유도를 적절히 선택함으로써 t분포에 가까운 것이 알려져 있다.

두 모평균의 차이에 대한 추론(이분산, $\sigma_1^2 \neq \sigma_2^2$)

- 귀무가설이 $H_0 : \mu_1 - \mu_2 = a$인 경우에 검정통계량은

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \dot{\sim} t(df), \qquad df = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}$$

- $\mu_1 - \mu_2$에 대한 $100(1-\alpha)\%$신뢰구간

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}(df)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- 검정통계량의 관측값을 t_0라고 할 때

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \mu_1 - \mu_2 > a$	$P = P(T \geq t)$	$t_0 \geq t_\alpha(df)$
(b) $H_1 : \mu_1 - \mu_2 < a$	$P = P(T \leq t)$	$t_0 \leq -t_\alpha(df)$
(c) $H_1 : \mu_1 - \mu_2 \neq a$	$P = P(\lvert T\rvert \geq \lvert t\rvert)$	$\lvert t_0\rvert \geq t_{\alpha/2}(df)$

두 모평균의 차이에 대한 추론 예제

Example

최근에는 하루 만에 배송이 가능한 서비스가 활성화되면서, 식료품을 온라인으로 구매하는 가정이 증가하고 있다. 한 경제심리학자는 소비자가 오프라인 매장에서 식료품을 구매할 때의 지출이 온라인 구매할 때의 지출보다 적다고 주장한다. 이를 확인하기 위해 무작위로 선정한 18명과 20명에게 각각 오프라인과 온라인에서 4인 가구의 2주일 치 식량을 쇼핑하도록 하고, 지출액을 조사한 결과는 다음과 같다.

(단위 : 천 원)

오프라인	169	206	257	294	252	283	240	207	230	183
	298	269	256	277	300	126	318	184		
온라인	289	334	278	268	336	438	388	388	394	394
	425	386	356	342	305	365	355	312	209	458

오프라인 쇼핑의 지출액이 온라인 쇼핑보다 적다는 증거가 있는지 유의수준 5%에서 검정하라.

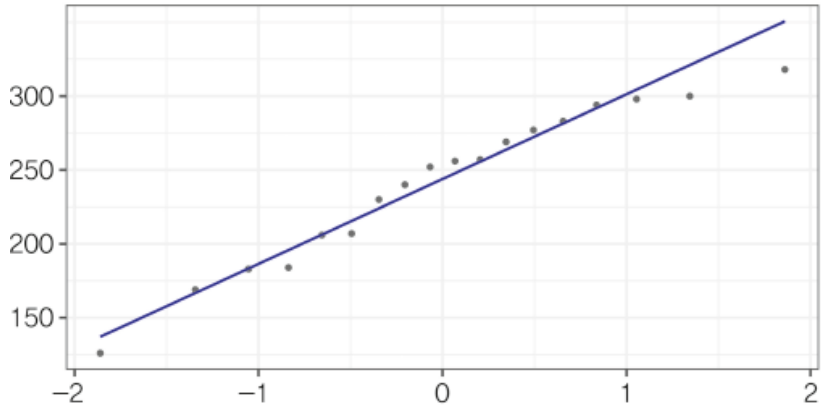
sol) 오프라인 쇼핑의 평균 지출액을 μ_1, 온라인 쇼핑의 평균 지출액을 μ_2라고 하면 검정하고자 하는 가설은 다음과 같다.

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 < \mu_2$$

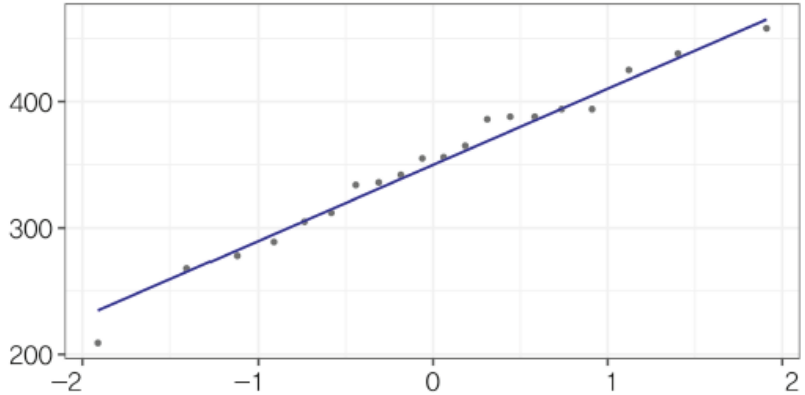
- 주어진 자료에서 평균과 표준편차를 구하면 다음과 같다.

	오프라인	온라인
표본크기(n_i)	18	20
평균($\bar{x}_i$)	241.6	351.0
표준편차(si)	52.82	61.92

- 두 집단의 정규분포 분위수대조도를 그려보면 다음과 같이 전반적으로 정규모집단을 전제로 하는 방법의 적용에 큰 무리가 없는 것으로 판단된다.



(a) 오프라인 쇼핑



(b) 온라인 쇼핑

- 따라서 t검정통계량의 관측값을 계산하면

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{(241.6 - 351.0) - 0}{\sqrt{\frac{52.82^2}{18} + \frac{61.92^2}{20}}} = -5.875$$

- 자유도를 계산하면 다음과 같다.

$$df = \frac{(52.82^2/18 + 61.92^2/20)^2}{\frac{(52.82^2/18)^2}{17} + \frac{(61.92^2/20)^2}{19}} = 35.91$$

- $t_{0.05}(35.91)$의 값을 t분포표에서 찾아보면

$$t_{0.05}(30) = 1.697 > t_{0.05}(35.91) > t_{0.05}(40) = 1.684$$

이므로 보간법을 이용하면 다음과 같다.

$$t_{0.05}(35.91) = 1.684 + (1.697 - 1.684) \times 0.409 = 1.689$$

- 실제 문제에서 자유도 df는 t분포표에서 df에 제일 가까운 자연수의 자유도를 이용한다. (위의 경우 $t_{0.05}(40) = 1.684$이용)

- 기각역으로 판정:

  기각역은 $t_0 \leq -t_{0.05}(35.91)$이고, $t_0 = -5.875 < -1.684$이므로 유의수준 5%에서 귀무가설을 기각할 수 있다.

  * 통계 소프트웨어를 이용해서 계산하면 $t_{0.05}(35.91) = 1.68841$

- p값으로 판정:

  $P = P(T \leq -5.875)$이고 t 분포표에서 $P < 0.0005$이므로 귀무가설을 기각할 수 있다.

- 이상을 종합하면, 유의수준 5%에서 오프라인 매장에서 식료품을 구매할 때 온라인 구매보다 평균 지출액이 적다는 증거가 뚜렷하다.

두 모평균의 차이에 대한 추론(이분산, 대표본의 경우)

II. σ_1^2, σ_2^2를 모르는 경우

1. $\sigma_1^2 \neq \sigma_2^2$(이분산 가정)
2. but, 모집단의 분포가 정규분포가 아닌 경우

- 모집단의 분포에 상관없이 n_1, n_2가 충분히 크면 중심극한정리에 의해 대략적으로 표준정규분포를 따르게 된다.

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \dot{\sim} N(0,1)$$

- 대표본의 조건: $n_1 \geq 30, n_2 \geq 30$

두 모평균의 차이에 대한 추론(이분산, 대표본의 경우)

- $\mu_1 - \mu_2$에 대한 $100(1-\alpha)\%$신뢰구간

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- 귀무가설이 $H_0 : \mu_1 - \mu_2 = a$인 경우에 검정통계량은

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \dot{\sim} N(0,1)$$

- 검정통계량의 관측값을 z_0라고 할 때

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \mu_1 - \mu_2 > a$	$P = P(Z \geq z)$	$z_0 \geq z_\alpha$
(b) $H_1 : \mu_1 - \mu_2 < a$	$P = P(Z \leq z)$	$z_0 \leq -z_\alpha$
(c) $H_1 : \mu_1 - \mu_2 \neq a$	$P = P(\lvert Z\rvert \geq \lvert z\rvert)$	$\lvert z_0\rvert \geq z_{\alpha/2}$

두 모평균의 차이에 대한 추론 예제

Example

지역 환경에 따라 학력에 차이가 있는가를 알아보기 위하여, 두 도시의 중학교 1학년 학생 중에서 각각 90명과 100명을 랜덤추출하여 동일한 시험을 시행한 결과가 다음가 같았다.

	도시 1	도시 2
표본크기(n_i)	90	100
평균($\bar{x}_i$)	76.4	81.2
표준편차(s_i)	10.2	7.6

두 도시의 중학교 1학년 학생 전체의 평균 성적에 차이가 있는가 유의수준 5%에서 검정하고, 유의확률도 구하여라. 또한, 평균성적에 차이에 대한 신뢰수준 95%의 신뢰구간도 구하여라.

sol) 도시 1, 2의 중학교 1학년 학생 전체의 평균성적을 각각 μ_1, μ_2라고 하자.

- 검정하려는 가설은 다음과 같다.

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$

- 조사한 표본의 크기가 각각 $n_1 = 90$, $n_2 = 100$으로 충분히 크므로, 이 경우에 검정통계량의 관측값을 계산하면

$$z_0 = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{(76.4 - 81.2) - 0}{\sqrt{\frac{10.2^2}{90} + \frac{7.6^2}{100}}} = -3.65$$

- $z_{0.025} = 1.96$이므로, $\mu_1 - \mu_2$에 대한 95%신뢰구간은 다음과 같다.

$$\begin{aligned}(\bar{x}_1 - \bar{x}_2) \pm z_{0.025}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} &= (76.4 - 81.2) \pm 1.96\sqrt{\frac{10.2^2}{90} + \frac{7.6^2}{100}} \\ &= (-7.38, -2.22)\end{aligned}$$

- 기각역으로 판정:

  기각역은 $|z_0| \geq 1.96$이고, $|z_0| = 3.65 \geq 1.96$이므로 기각역에 속하여 유의수준 5%에서 귀무가설을 기각할 수 있다.

- p값으로 판정:

  유의확률은 $P = P(|Z| \geq)$이고 표준정규분포표에서 그 값을 찾아보면

$$\begin{aligned} P(|Z| \geq 3.65) &= P(Z < -3.65) + P(Z > 3.65) \\ &= 0.00013 + 0.00013 = 0.00026 \end{aligned}$$

- P값은 $\alpha = 0.05$보다 작아 귀무가설을 기각한다.
- 이상을 종합하면, 두 도시의 중학교 1학년 학생의 학력에는 차이가 있다는 뚜렷한 증거가 있다.

대응비교에 의한 모평균의 비교

예) 다이어트하기 전과 다이어트 한 달 후의 몸무게처럼 동일한 개체에 대해 실험 전과 실험 후의 측정값의 차이를 비교하는 경우

- 앞 절에서 다룬 바와 같이 측정 결과를 두 모집단의 표본으로 간주하고 평균 비교를 할 수도 있지만 대응비교(paired comparison)를 이용하면 더욱 정확한 실험 효과를 측정할 수 있다.

- 대응비교: 두 모집단의 평균을 비교할 때 동질적인 비교 대상들로 쌍을 이루어 각 쌍내에서의 차를 이용하여 비교하는 방법

대응비교에 의한 모평균의 비교

예) 두 종류의 진통제가 수면에 미치는 효과를 비교하고자 한다. 실험에 참여하기로 한 환자들 중 6명의 환자를 랜덤추출하고 각 환자에게 두 종류의 진통제를 랜덤한 순서로 각각 1회씩 복용하게 하여 숙면시간에 차이가 있는지 조사하였다.

환자 1: (진통제 A, 진통제 B)

환자 2: (진통제 A, 진통제 B)

... ...

환자 6: (진통제 A, 진통제 B)

- 실험에 참여하기로 한 환자들의 건강상태에 상당한 차이가 있다. 이런경우 두 집단이 독립이라는 가정을 만족 못하므로, 환자들의 이질성을 감안하여 대응비교를 하고자 한다.

  => 두 집단의 측정값의 차이를 이용하여 추론하면 된다.

대응비교에 의한 모평균의 비교

- 주어진 자료: $(X_1, Y_1), \cdots, (X_n, Y_n)$
- 처리효과를 나타내는 모평균 $\mu_1 = E(X_i)$, $\mu_2 = E(Y_i)$를 비교하려면, 자료에서 얻은 측정값의 차이인 d_i를 구하여 그 차이가 0인가에 대한 문제로 바꾸어 추론한다.

$$d_i = X_i - Y_i \ (i = 1, 2, \cdots, n)$$

- 가정사항

1. $d_i = X_i - Y_i$들은 각 모집단에서의 랜덤표본이라 가정.
2. 모집단은 정규분포를 따른다고 가정.
   (표본크기 n이 충분히 클 때에는 근사적으로 성립함)
3. 실험단위를 랜덤하게 추출하고, 처리의 적용순서도 랜덤하게 설정.

대응비교에 의한 모평균의 비교

- d_i의 기대값: $E(d_i) = \mu_1 - \mu_2 = \mu_D$

- 모평균 μ_D의 추정량: 표본평균 $\bar{D} = \sum_{i=1}^{n} \frac{d_i}{n}$
  모분산의 추정량: 표본분산 S_D^2

$$S_D^2 = \frac{1}{n-1} \sum (d_i - \bar{d})^2$$

- $\bar{D}$의 기대값: $E(\bar{D}) = \mu_D$
  $\bar{D}$의 분산:

$$Var(\bar{D}) = \frac{s_D^2}{n}$$

대응비교에 의한 모평균의 비교

- $H_0 : \mu_1 - \mu_2 = 0$ $(\mu_D = 0)$ 에 대한 검정통계량:

$$T = \frac{\bar{D} - \mu_D}{\sqrt{\frac{S_D^2}{n}}} \sim t_{n-1}, \quad df = n - 1$$

- $\mu_1 - \mu_2$에 대한 $100(1-\alpha)\%$신뢰구간

$$\bar{d} \pm t_{\alpha/2}(n-1)\sqrt{\frac{s_D^2}{n}}$$

- 검정통계량의 관측값을 t_0라고 할 때

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \mu_D > 0$	$P = P(T \geq t)$	$t_0 \geq t_\alpha(n-1)$
(b) $H_1 : \mu_D < 0$	$P = P(T \leq t)$	$t_0 \leq -t_\alpha(n-1)$
(c) $H_1 : \mu_D \neq 0$	$P = P(\lvert T \rvert \geq \lvert t \rvert)$	$\lvert t_0 \rvert \geq t_{\alpha/2}(n-1)$

대응비교 예제

Example

두 종류의 진통제가 수면에 미치는 효과를 비교하기 위해 실험을 진행하려 한다. 그러나 실험에 참여한 환자들의 건강상태가 서로 간에 크게 달라, 이를 고려한 실험설계가 필요하다. 따라서 환자 6명을 랜덤추출하고, 각 환자에게 진통제 두 종류를 각각 1회씩 복용시킨 후 수면 시간의 차이를 측정하여 효과를 비교하기로 했다. 이때 복용 순서에 따라 효과가 달라질 수 있으므로 환자마다 어떤 진통제를 먼저 복용할지를 동전을 던져 무작위로 결정했다. 두 진통제 A, B의 수면시간을 각각 μ_A, μ_B라 하자. 아래 표에 제시된 실제 조사 결과를 바탕으로 수면 시간의 차 $\mu_A - \mu_B$에 대한 95% 신뢰구간을 구하여라.

표 6-1 진통제에 따른 수면 시간의 측정값

(단위 : 시간)

환자	1	2	3	4	5	6
진통제 A	4.8	4.0	5.8	4.9	5.3	7.4
진통제 B	4.0	4.2	5.2	4.9	5.6	7.1

sol) 대응비교에 의한 자료이므로 그 차이 d_i를 구하면 다음과 같다.

Table: 진통제에 따른 숙면시간의 측정값(시간)

환자	1	2	3	4	5	6
진통제A	4.8	4.0	5.8	4.9	5.3	7.4
진통제B	4.0	4.2	5.2	4.9	5.6	7.1
차이	0.8	-0.2	0.6	0.0	-0.3	0.3

- 차이에 대한 평균과 표준편차를 구하면 각각 다음과 같다.

$$\bar{d} = 0.20, \quad s_D = 0.443$$

- 자유도는 $n - 1 = 5$이므로 $t_{0.025}(5) = 2.571$이다.
- 따라서, $\mu_A - \mu_B$의 95% 신뢰구간은 다음과 같다.

$$(\bar{d} \pm t_{0.025}(5)\sqrt{\frac{s_D^2}{n}} = 0.20 \pm 2.571 \times \frac{0.443}{\sqrt{6}} = (-0.26, 0.66)$$

대응비교 예제

Example

재활치료를 위해 개발된 보행 보조 웨어러블 로봇이 실제로 운동 효과를 증가시키는지 확인하고자 한다. 이를 위해 한 재활 센터에서 무작위로 선정한 15명을 대상으로 실험을 진행했다. 모든 환자가 웨어러블 로봇을 착용하지 않은 상태와 착용한 상태에서 각각 10분간 가벼운 걷기 운동을 수행했을 때의 소모 칼로리를 비교한 결과는 아래 표와 같다. 재활치료 시 웨어러블 로봇을 착용하면 칼로리 소모가 더 커진다는 증거가 있는지 유의수준 1%에서 검정하라.

표 6-2 보행 보조 웨어러블 로봇 착용 전과 후의 소모 칼로리

(단위 : kcal)

로봇 착용 여부	1	2	3	4	5	6	7	8
미착용	23.9	23.9	20.0	19.2	22.2	11.8	21.7	27.7
착용	25.2	28.2	24.9	20.2	19.7	16.9	22.6	20.8
로봇 착용 여부	9	10	11	12	13	14	15	
미착용	15.6	22.5	23.2	19.3	15.2	10.0	11.2	
착용	23.8	18.3	23.7	22.5	16.2	12.1	22.5	

sol) 웨어러블 로봇 착용 전 평균 소모 칼로리를 μ_1, 착용 후 평균 소모 칼로리를 μ_2라 하면,

- 가설은 다음과 같다.

$$H_0 : \mu_1 = \mu_2 \ (\text{or } \mu_D = 0) \quad H_1 : \mu_1 < \mu_2 \ (\text{or } \mu_D < 0)$$

- 관측값의 차이 d_i를 구하면 다음과 같다.

$$-1.3 \quad -4.3 \quad -4.9 \quad -1.0 \quad 2.5 \quad -5.1 \quad -0.9 \quad 6.9$$
$$-8.2 \quad 4.2 \quad -0.5 \quad -3.2 \quad -1.0 \quad -2.1 \quad -11.3$$

- 이들의 평균과 표준편차를 구하면 각각 다음과 같고

$$\bar{d} = -2.01, \quad s_D = 4.58$$

- 이로부터 t검정통계량의 관측값을 계산하면

$$t_0 = \frac{\bar{d} - 0}{s_D/\sqrt{n}} = \frac{-2.01 - 0}{4.58/\sqrt{15}} = -1.70$$

- 자유도는 $n-1=14$이므로 $t_{0.01}(14)=2.624$이다.
- 기각역으로 판정:

  기각역은 $t_0 \le -2.624$이고, $t_0=-1.7 \ge -2.624$이므로 유의수준 1%에서 귀무가설을 기각할 수 없다.

- 유의확률은 다음과 같다.

$$P = P\{T \le -1.7\}, \quad T \sim t(14)$$

  t분포표로부터 $0.05 < P < 0.1$이다. 통계 소프트웨어를 이용하면 실제 유의확률은 $P=0.056$으로 계산된다.
- 종합하면 1% 유의수준에서는 재활치료 웨어러블 로봇이 소모 칼로리를 증가시킨다는 뚜렷한 증거가 없다.

등분산 이표본 비교 VS 대응비교

- 주어진 자료: $(X_1, \cdots, X_n), (Y_1, \cdots, Y_n)$
- 모평균 $\mu_X = E(X_i)$, $\mu_Y = E(Y_i)$를 비교하고자 한다.
  총 자료의 갯수: n or $2n$, σ_X^2, σ_Y^2의 값은 모른다고 가정.

- 가설: $H_0 : \mu_X = \mu_Y$ $(or\ \mu_D = 0)$

- 각각의 검정 통계량과 표본 분포는 다음과 같다.

[등분산 이표본의 비교]

$$T = \frac{(\bar{X} - \bar{Y}) - 0}{S_p\sqrt{\frac{1}{n} + \frac{1}{n}}} \sim t(2n-2)$$

[대응비교]

$$T = \frac{\bar{D} - 0}{\sqrt{\frac{S_D^2}{n}}} \sim t(n-1)$$

등분산 이표본 비교 VS 대응비교

1. 검정 통계량의 분자는 동일하다.

$$\bar{D} = \bar{X} - \bar{Y}$$

2. $(\bar{X} - \bar{Y})$의 분산을 생각해 보자.
   - If $\bar{X}$ and $\bar{Y}$ are independent:

$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y})$$

   - If $\bar{X}$ and $\bar{Y}$ are dependent:

$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) - 2Cov(\bar{X}, \bar{Y})$$

$\bar{X}$와 $\bar{Y}$가 양의 상관관계를 보인다면($Cov(\bar{X}, \bar{Y}) > 0$) 대응비교의 검정통계량 값이 더 커져서 H_0를 기각하기 쉬워진다.

$$S_p\sqrt{\frac{1}{n} + \frac{1}{n}} > \sqrt{\frac{S_D^2}{n}}$$

등분산 이표본 비교 VS 대응비교

3. t분포의 자유도를 생각해보자.

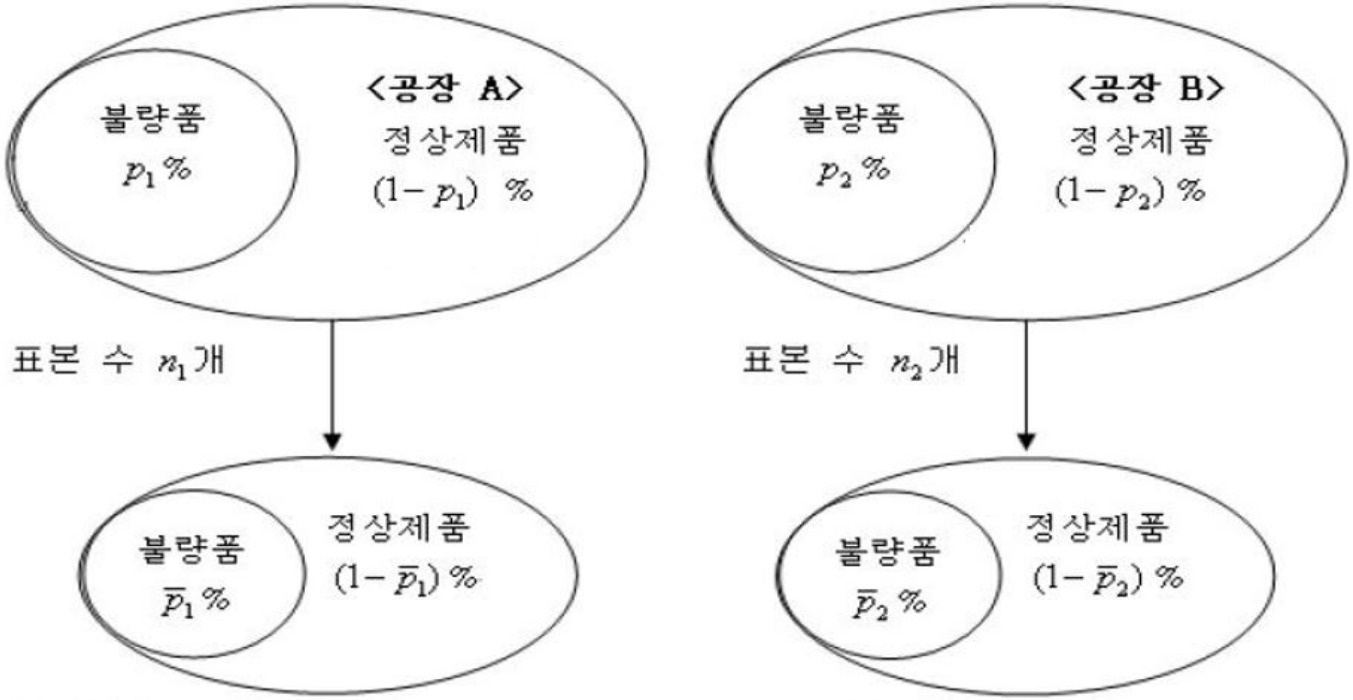
- 등분산 이표본의 비교에선 자유도가 $2n-2$이고 대응비교에서는 $n-1$이다. 따라서 같은 조건일 때 대응비교의 귀무가설 기각이 더 힘들어 진다. (대응비교는 자유도에서 penalty가 있음)

- 따라서 실험단위들이 이질적이고 각 쌍 내에서 값들이 양의 상관관계를 가진다면 대응비교가 적합하고,
- 위의 조건이 아니라면, 자유도의 penalty때문에 대응비교를 사용하는 장점이 사라지게 된다.

두 모비율의 차이에 대한 추론

예) 서로 다른 두 치료방법에 의한 치유율의 비교,
두 제조 과정에서의 불량률의 비교

- 두 모집단에서 각각 n_1, n_2개의 랜덤표본을 서로 독립적으로 추출한 경우에 관심의 대상인 속성이 두 표본에서 나타나는 도수를 각각 X_1, X_2라고 하자.



두 모비율의 차이에 대한 추론

- X_1과 X_2는 각각 이항분포 $B(n_1, p_1)$, $B(n_2, p_2)$를 따르며, 서로 독립이고, p_1과 p_2의 추정량은 $\hat{p}_1 = X_1/n_1$, $\hat{p}_2 = X_2/n_2$이라고 하자.
- 두 모비율의 차인 $p_1 - p_2$의 추정량:

$$\widehat{p_1 - p_2} = \hat{p}_1 - \hat{p}_2 = X_1/n_1 - X_2/n_2$$

- $\hat{p}_1 - \hat{p}_2$의 기대값:

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

- $\hat{p}_1 - \hat{p}_2$의 분산:

$$Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

- $\hat{p}_1 - \hat{p}_2$ 표준오차 추정량:

$$\widehat{SE}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

두 모비율의 차이에 대한 추론

- 표본크기 n_1, n_2가 충분히 큰 경우

$$\hat{p}_1 \dot{\sim} N(p_1, p_1(1-p_1)/n_1), \quad \hat{p}_2 \dot{\sim} N(p_2, p_2(1-p_2)/n_2)$$

- $\hat{p}_1$과 $\hat{p}_2$은 서로 독립이므로

$$(\hat{p}_1 - \hat{p}_2) \dot{\sim} N(p_1 - p_2, p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2)$$

- 따라서 n_1, n_2가 충분히 큰 경우 다음이 성립하고

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}$$

- 모비율을 표본비율로 대체하면 표본분포는 다음과 같다.

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}} \dot{\sim} N(0,1) \tag{2}$$

두 모비율의 차이에 대한 신뢰구간

- 다음을 이용하여 $p_1 - p_2$에 대한 신뢰구간을 구할 수 있다.

$$P\left\{-z_{\alpha/2} \le \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}} \le z_{\alpha/2}\right\} \doteq 1 - \alpha$$

- $p_1 - p_2$에 대한 $100(1-\alpha)\%$신뢰구간:

$$((\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\widehat{SE}(\hat{p}_1 - \hat{p}_2),\ (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\widehat{SE}(\hat{p}_1 - \hat{p}_2))$$

- 표본크기가 크다는 것의 기준:
  $n_1\hat{p}_1,\ n_1(1-\hat{p}_1),\ n_2\hat{p}_2,\ n_2(1-\hat{p}_2)$가 모두 5이상

두 모비율의 차이에 대한 신뢰구간 예제

Example (7.5)

현재 누구나 그 효과를 인정하는 소아마비 백신이지만, 1954년 미국에서는 그 효능을 검증하기 위한 대규모 임상시험이 시행되었다. 백신의 효과를 올바르게 평가하기 위해, 백신이 아닌 위약을 접종받는 어린이들로 구성된 대조집단과 백신을 접종받는 처리집단을 대상으로 이중 눈가림 실험이 실시되었다. 접종 후 마비 증세를 보인 어린이들의 수를 관측간 결과는 다음과 같았다.

표 7-1 1954년 소아마비 백신의 실험 결과

(단위 : 명)

	표본 크기	마비 증세가 나타난 어린이
대조집단	201,229	115
처리집단	200,745	33

대조집단과 처리집단에서 마비 증세를 보인 비율은 p_1, p_2라 할 때, $p_1 - p_2$의 추정값과 95%신뢰구간을 구하여라.

sol) 모비율 p_1과 p_2의 추정값이 각각

$$\hat{p}_1 = 115/201,229 = 5.71 \times 10^{-4}, \quad \hat{p}_2 = 33/200,745 = 1.64 \times 10^{-4}$$

이므로, $p_1 - p_2$의 추정값은 $\widehat{p_1 - p_2} = \hat{p}_1 - \hat{p}_2 = 4.07 \times 10^{-4}$이다.

- 이 추정값의 표준오차는 다음과 같으므로

$$\widehat{SE}(\widehat{p_1 - p_2}) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{201,229} + \frac{\hat{p}_2(1-\hat{p}_2)}{200,745}} = 0.60 \times 10^{-4}$$

- $p_1 - p_2$의 95%신뢰구간은 다음과 같다.

$$(4.07 \times 10^{-4} - 1.96 \times 0.60 \times 10^{-4},\ 4.07 \times 10^{-4} + 1.96 \times 0.60 \times 10^{-4})$$

- 즉 신뢰수준 95%에서 $p_1 - p_2$ 값의 범위를 다음과 같이 추정할 수 있다.

$$2.89 \times 10^{-4} < p_1 - p_2 < 5.25 \times 10^{-4}$$

두 모비율의 비교를 위한 검정

- 귀무가설($H_0 : p_1 = p_2$)이 사실일 때,
  두 모집단에서의 공통인 모비율 p를 다음과 같이 생각해보자.

$$p_1 = p_2 = p$$

- p의 추정량: $\hat{p}$ 합동표본비율(pooled sample proportion)

  -두 표본을 합하여 모두 $n_1 + n_2$개의 표본에서 $X_1 + X_2$개가 대응하는 속성을 가진 것으로 생각하자.

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

두 모비율의 비교를 위한 검정

- $\hat{p}_1 - \hat{p}_2$의 분산:

$$Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

- $\hat{p}_1 - \hat{p}_2$의 분산의 추정량:

$$\widehat{Var}(\hat{p}_1 - \hat{p}_2) = \hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

- $(\hat{p}_1 - \hat{p}_2)$의 표본분포는 귀무가설이 $H_0 : p_1 = p_2 (= p)$이 참일 때 다음과 같다.

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{p(1-p)(n_1^{-1} + n_2^{-1})}}$$

두 모비율의 비교를 위한 검정

- 귀무가설: $H_0 : p_1 = p_2$
- 검정통계량은 공통 모비율이 $\hat{p} = \frac{X_1+X_2}{n_1+n_2}$ 일 때 다음과 같다.

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})(n_1^{-1} + n_2^{-1})}} \dot{\sim} N(0,1)$$

검정통계량의 관측값을 z_0라고 할 때, 다음이 성립한다.

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : p_1 > p_2$	$P = P(Z \geq z)$	$z_0 \geq z_\alpha$
(b) $H_1 : p_1 < p_2$	$P = P(Z \leq z)$	$z_0 \leq -z_\alpha$
(c) $H_1 : p_1 \neq p_2$	$P = P(\lvert Z \rvert \geq \lvert z \rvert)$	$\lvert z_0 \rvert \geq z_{\alpha/2}$

두 모비율의 비교를 위한 검정 예제

Example

Example 7.5의 자료를 이용하여, 소아마비 백신이 효과가 있다고 할 수 있는가를 유의수준 1%에서 검정 하여라.

sol) p_1과 p_2는 각각 가짜약과 백신 접종 후의 마비율을 나타낸다고 하자.

- 검증하려는 가설은 다음과 같다.

$$H_0 : p_1 = p_2, \qquad H_1 : p_1 > p_2$$

- 귀무가설 하에서 공통 모비율의 추정값을 구하면

$$\hat{p} = \frac{115 + 33}{201229 + 200745} = 3.68 \times 10^{-4}$$

- 검정통계량의 관측값은

$$z_0 = \frac{5.71 \times 10^{-4} - 1.64 \times 10^{-4}}{\sqrt{\hat{p}(1-\hat{p})(n_1^{-1} + n_2^{-1})}} = 6.727$$

- 오른쪽 단측인 대립가설인 경우의 유의확률을 구하면 다음과 같이 매우 작은 값이 도출됨을 알 수 있고, 따라서 귀무가설을 기각 할 수 있다.

$$P = P(Z \geq 6.727) = 8.66 \times 10^{-12} < 0.01$$

- 즉 임상실험 결과는 소아마비 백신이 예방 효과가 뚜렷함을 시사하고 있다.

두 모비율의 비교를 위한 검정 예제

Example (7.7)

[표 7-2]는 한 범죄학 학술지에 실린 연구 결과로, 충동적 살인범과 계획적 살인법의 교정 효과에 차이가 있는지 조사하기 위한 것이다. 일정 기간 복역 후 가석방된 충동적 살인범과 계획적 살인범 중 각각 42명과 40명을 랜덤추출하여 가석방이 성공인 경우(재범 없음)와 실패인 경우(재범 발생)를 관측한 결과가 다음과 같다. 살인범의 유형에 따라 가석방 성공률, 즉 재범하지 않을 확률에 차이가 있는지를 유의수준 5%에서 검정하라.

표 7-2 살인범의 유형에 따른 가석방 성공 여부

(단위 : 명)

	성공	실패	표본 크기
충동적 살인범	13	29	$n_1 = 42$
계획적 살인범	22	18	$n_2 = 40$
합계	35	47	$n = 82$

sol) 충동적 살인범과 계획적 살인범에 대한 가석방의 성공률을 각각 p_1, p_2라고 하면 검정하려는 가설은 다음과 같다.

$$H_0 : p_1 = p_2, \qquad H_1 : p_1 \neq p_2$$

- p_1, p_2와 귀무가설 하에서의 공통 모비율의 추정값은

$$\hat{p}_1 = 13/42 = 0.31, \qquad \hat{p}_2 = 22/40 = 0.55$$

$$\hat{p} = \frac{13 + 22}{42 + 40} = 0.427$$

- 이로부터 검정통계량의 관측값을 구하면

$$z_0 = \frac{0.31 - 0.55}{\sqrt{\hat{p}(1 - \hat{p})(n_1^{-1} + n_2^{-1})}} = -2.20$$

- 유의수준 5%에서 검정통계량의 관측값이 기각역 $|z_0| \geq z_{0.025} = 1.96$에 포함되므로, 귀무가설을 기각 할 수 있다.

두 모집단의 분산에 대한 추론

- 두 모집단의 분산 비교의 예
  - 서로 다른 투자 포트폴리오에 관련된 위험 비교
  - 두 모평균의 비교 시 등분산 가정이 적합한지에 대한 분석

- 두 모집단의 표준편차 σ_1, σ_2를 비교

$$\sigma_1^2 = \sigma_2^2 \iff \sigma_1^2/\sigma_2^2 = 1$$

$$\sigma_1^2 > \sigma_2^2 \iff \sigma_1^2/\sigma_2^2 > 1$$

$$\sigma_1^2 < \sigma_2^2 \iff \sigma_1^2/\sigma_2^2 < 1$$

두 모집단의 분산에 대한 추론

- 추정량
  - 정규모집단 1의 랜덤표본: $X_{11}, \cdots, X_{1n_1}$
  - 정규모집단 2의 랜덤표본: $X_{21}, \cdots, X_{2n_2}$
  - 모분산 σ_1^2, σ_2^2의 점추정량: 표본분산 s_1^2, s_2^2
  - 모분산 비율 $\frac{\sigma_1^2}{\sigma_2^2}$의 점추정량: $\frac{s_1^2}{s_2^2}$

- 모분산 비율에 대한 추론시 유용한 분포:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

이를 이용하여 σ_1^2/σ_2^2에 대한 추정이나 유의성검증을 할 수 있음

F 분포

Definition (F 분포:F distribution)

V_1과 V_2가 서로 독립이고 각각 자유도 k_1, k_2인 카이제곱분포를 따를 때,

$$F = \frac{V_1/k_1}{V_2/k_2}$$

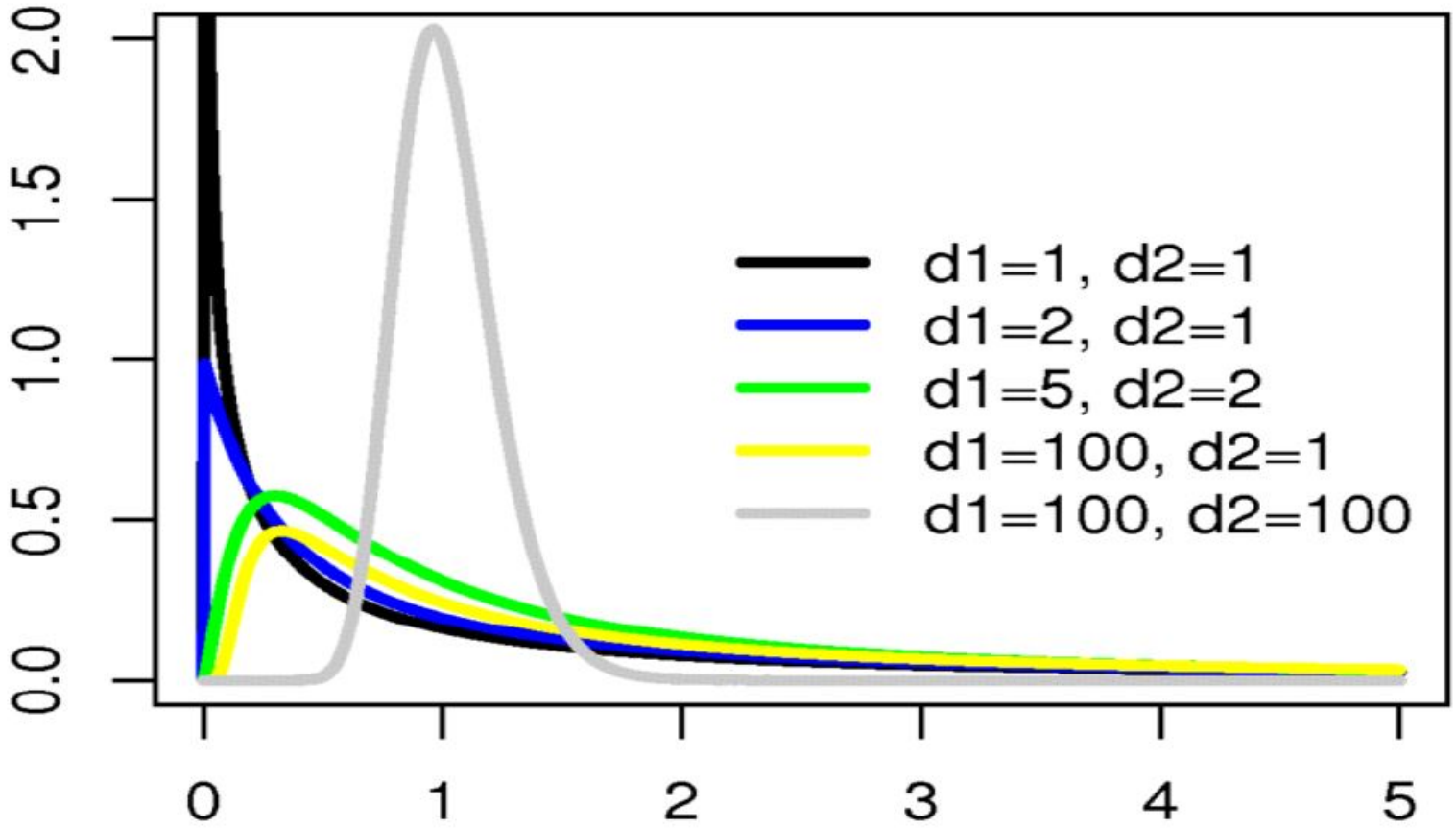
의 분포를 자유도 (k_1, k_2)인 F 분포라고 한다. 이 때 기호로는 다음과 같이 나타낸다.

$$F \sim F(k_1, k_2)$$

- 분자 자유도 k_1과 분모 자유도 k_2에 따라 그 모양이 달라진다.
- 두 정규모집단의 분산 비교시 분산비에 관한 추론에 이용됨

F 분포

Figure: 자유도의 변화에 따른 여러가지 F분포의 모양



F 분포

- $F \sim F(k_1, k_2)$일 때 다음의 식에서 $F_\alpha(k_1, k_2)$를 자유도가 (k_1, k_2)인 F 분포의 $(1-\alpha)$분위수라고 한다.

$$P\{F \geq F_\alpha(k_1, k_2)\} = \alpha$$

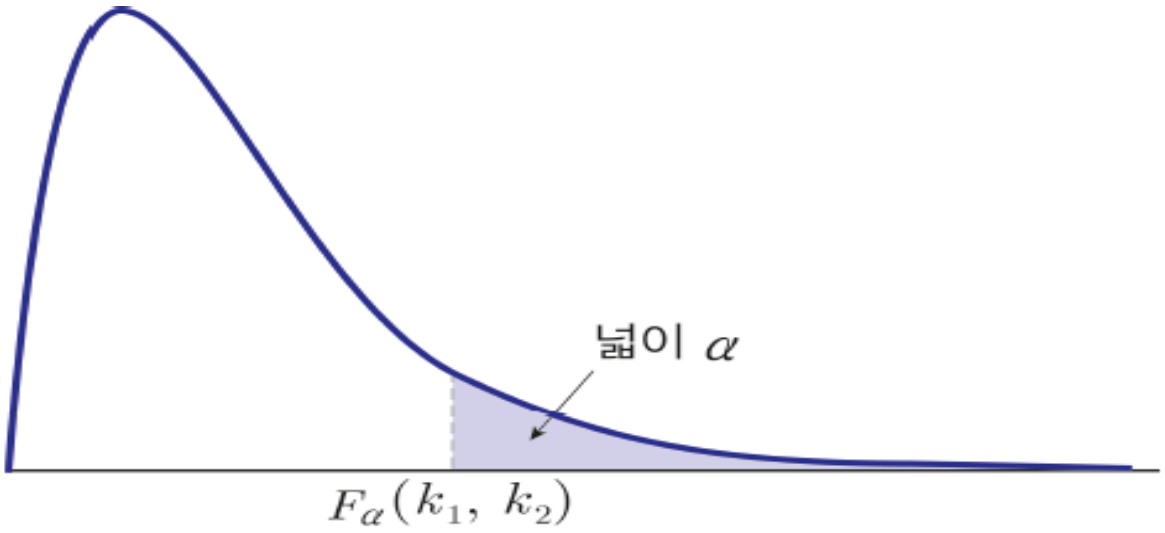
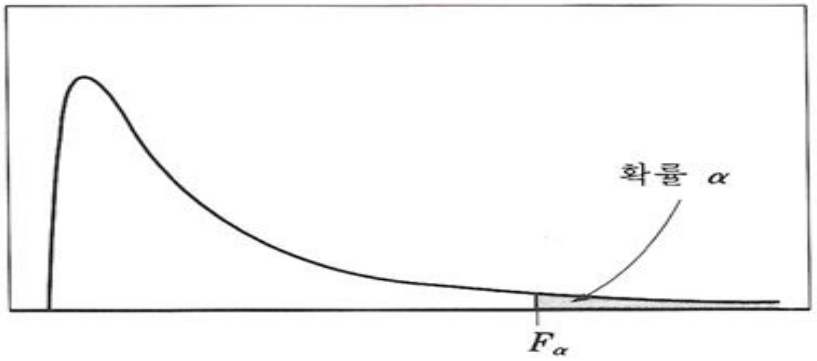


그림 4–10 F **분포의 형태**

F 분포표

- F 분포표에서 자유도별로 여러 가지 분위수를 구할 수 있다.
  예) $F_{0.10}(4,2) = 9.24$, $F_{0.05}(4,2) = 19.25$



표의 숫자는 주어진 α값에 대한 F_α의 값을 나타낸다.

		분자의 자유도								
분모의 자유도	α	1	2	3	4	5	6	7	8	9
	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
	.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
1	.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28
	.010	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
	.001	405284	5000000	540379	562500	576405	585937	592873	598144	602284
	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
2	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
	.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
	.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39

F 분포의 성질

Theorem (F 분포의 성질)

확률변수 F가 자유도 (k_1, k_2)인 F 분포를 따를 때, 다음이 성립한다.

$$\frac{1}{F} \sim F(k_2, k_1)$$

- F 분포의 성질을 이용하면 다음과 같은 사실을 알 수 있다.

$$F_{1-\alpha}(k_2, k_1) = \frac{1}{F_\alpha(k_1, k_2)}$$

예) $F_{0.05}(3, 4) = 6.59$이므로 $F_{0.95}(4, 3) = 1/6.59 \approx 0.15$

F분포와 t분포와의 관계

Theorem (F분포와 t분포와의 관계)

확률변수 T가 자유도 k인 t분포를 따를 때 다음이 성립한다.

$$T^2 \sim F(1, k)$$

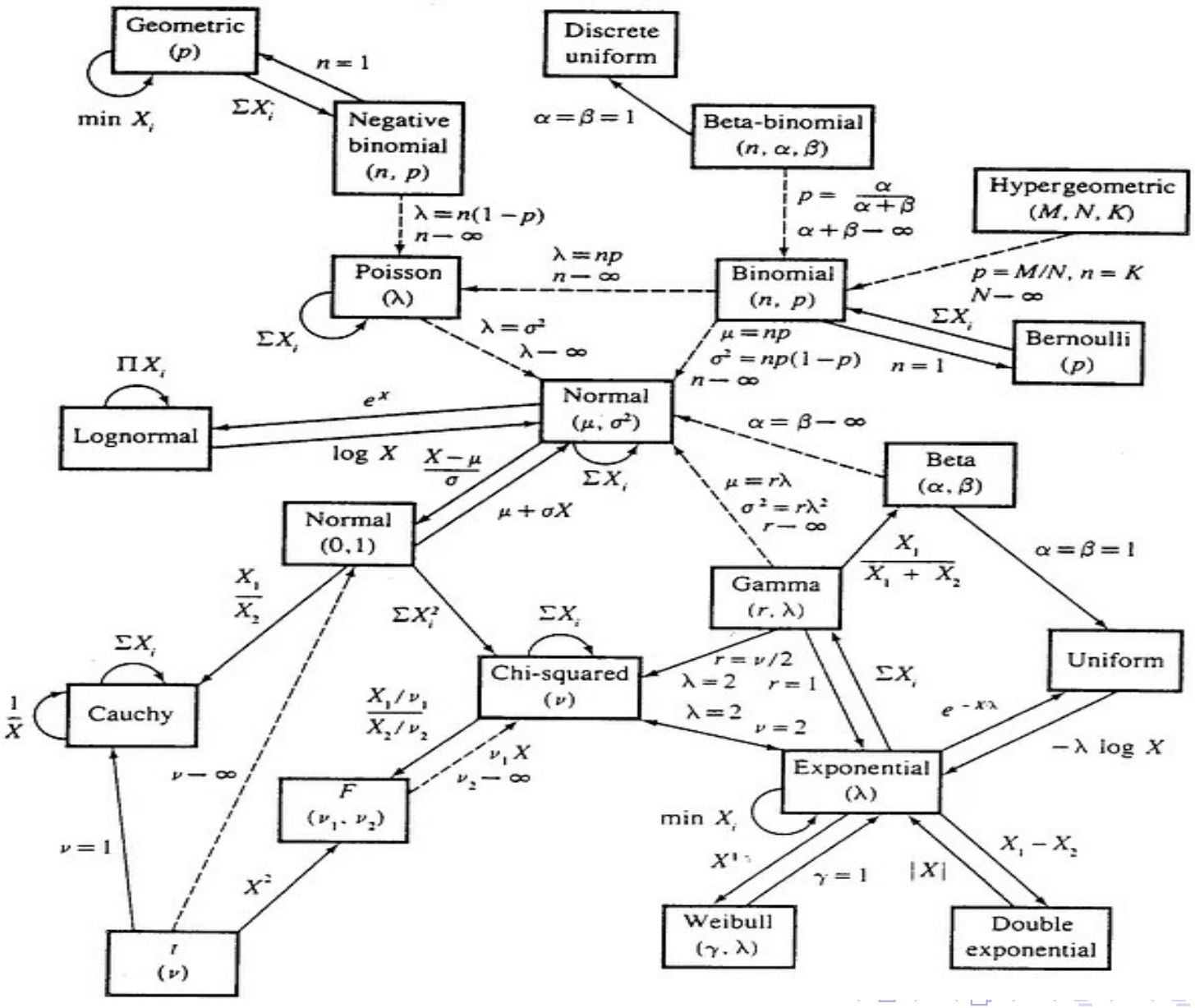
pf) 확률변수 T가 자유도 k인 t분포를 따를 때 다음과 같이 가정하자.
$T = \frac{Z}{\sqrt{V/k}}$ (단, $Z \sim N(0,1)$, $V \sim \chi^2(k)$이며 서로 독립)
양변을 제곱하면 다음과 같고

$$T^2 = \frac{Z^2/1}{V/k}$$

여기에서 $Z^2 \sim \chi^2(1)$이고, Z와 V는 서로 독립이므로 다음의 사실이 성립한다.

$$T^2 \sim F(1, k)$$

Table of Common Distributions



표본분산의 비에 관한 분포

Theorem

$(X_{11}, \cdots, X_{1n_1})$, $(X_{21}, \cdots, X_{2n_2})$가 서로 독립이고 각각 $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$에서의 랜덤표본이라 하자. 이 때, S_1^2과 S_2^2을 각각 두 랜덤표본의 표본분산이라고 하면 다음이 성립하고, 이들은 서로 독립이다.

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1), \quad \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1)$$

따라서 F분포의 정의에 따라 이들을 각각 자신의 자유도로 나눈 후의 비는 다음과 같이 자유도 (n_1-1, n_2-1)인 F분포를 따르게 된다.

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$$

- 모분산 비교시 가정:
  1. 두 모집단이 정규분포를 따름
  2. 표본추출시 독립적인 단순 임의추출을 해야 한다.

모분산의 비에 대한 신뢰구간(정규모집단의 경우)

- 다음을 이용하여 σ_1^2/σ_2^2에 대한 신뢰구간을 구할 수 있다.

$$P\left(F_{1-\alpha/2}(n_1-1,n_2-1) \le \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \le F_{\alpha/2}(n_1-1,n_2-1)\right) = 1-\alpha$$

$$P\left(\frac{S_1^2}{S_2^2}/F_{\alpha/2,n_1-1,n_2-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}/F_{1-\alpha/2,n_1-1,n_2-1}\right) = 1-\alpha$$

한편, F분포의 성질: $F_{1-\alpha/2}(n_1-1,n_2-1) = 1/F_{\alpha/2}(n_2-1,n_1-1)$

- 모분산의 비 σ_1^2/σ_2^2에 대한 $100(1-\alpha)\%$ 신뢰구간 :

$$\left(\frac{s_1^2}{s_2^2}/F_{\alpha/2}(n_1-1,n_2-1),\ \frac{s_1^2}{s_2^2}\cdot F_{\alpha/2}(n_2-1,n_1-1)\right)$$

모분산의 비에 대한 유의성검증(정규모집단의경우)

- 귀무가설 $H_0 : \sigma_1^2 = \sigma_2^2$인 경우에 검정통계량은

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2}$$

이고, 이의 관측값을 f라고 할 때 다음이 성립한다.

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \sigma_1^2 > \sigma_2^2$	$P = P(F \geq f)$	$f \geq F_\alpha(n_1 - 1, n_2 - 1)$
(b) $H_1 : \sigma_1^2 < \sigma_2^2$	$P = P(F \leq f)$	$f \leq 1/F_\alpha(n_2 - 1, n_1 - 1)$
(c) $H_1 : \sigma_1^2 \neq \sigma_2^2$	$P = 2P(F \geq f)$ or	$f \geq F_{\alpha/2}(n_1 - 1, n_2 - 1)$ or
	$P = 2P(F \leq f)$	$f \leq 1/F_{\alpha/2}(n_2 - 1, n_1 - 1)$

- 양측가설인 경우 $2P(F \geq f)$ or $2P(F \leq f)$ 중 1보다 작은 값을 유의확률로 계산

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \sigma_1^2 > \sigma_2^2$	$P = P(F \geq f)$	$f \geq F_\alpha(n_1 - 1, n_2 - 1)$
(b) $H_1 : \sigma_1^2 < \sigma_2^2$	$P = P(F \leq f)$	$f \leq \frac{1}{F_\alpha(n_2-1,n_1-1)}$
(c) $H_1 : \sigma_1^2 \neq \sigma_2^2$	$P = 2P(F \geq f)$ or $P = 2P(F \leq f)$	$f \geq F_{\alpha/2}(n_1 - 1, n_2 - 1)$ or $f \leq \frac{1}{F_{\alpha/2}(n_2-1,n_1-1)}$

두 모집단의 분산에 대한 추론 예제

Example

콘크리트에 균열이 발생하면 이를 보수하기 위해 주로 폴리머(polymer)를 주입한다. 다음은 한 연구보고서에서 폴리머 두 종류를 사용했을 때 측정한 주입 압축률에 대한 자료이다.

(단위 : 10^6psi)

에폭시(epoxy)	1.75	2.12	2.05	1.97
MMA 프리폴리머(prepolymer)	1.77	1.59	1.70	1.69

폴리머 두 종류의 주입 압축률 산포가 다르다는 증거가 있는지를 유의수준 5%에서 검정하고, 이들의 표준편차 비에 대한 95% 신뢰구간을 구하여라. (정규모집단이라고 가정)

sol) 두 폴리머의 주입 압축률의 표준편차를 각각 σ_1, σ_2라고 하자.

- 검정하고자 하는 가설은 다음과 같다.

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

- 주어진 자료에서 평균과 표준편차를 구하면 각각 다음과 같다.

	Epoxy	MMA Prepolymer
표본크기(n_i)	4	4
평균($\bar{x}_i$)	1.97	1.69
표준편차(s_i)	0.160	0.074

- F검정통계량의 관측값을 계산하면

$$f = S_1^2/S_2^2 = 0.160^2/0.074^2 = 4.675$$

- 기각역으로 판정:

  기각역은 $f \le 1/F_{0.025}(3,3) = 1/15.44$ 또는 $f \ge F_{0.025}(3,3) = 15.44$

  검정통계량의 관측값 $f = 4.675$이 기각역에 속하지 않아 유의수준 5%에서 귀무가설을 기각할 수 없다.

- 모분산의 비인 σ_1^2/σ_2^2에 대한 95% 신뢰구간은 다음과 같고

$$\left(\frac{0.160^2}{0.074^2}/15.44,\ \frac{0.160^2}{0.074^2}\cdot 15.44\right) = (0.301, 72.182)$$

- 모표준편차의 비인 σ_1/σ_2에 대한 95% 신뢰구간은 다음과 같다.

$$(\sqrt{0.301}, \sqrt{72.182}) = (0.549, 8.496)$$

두 정규모집단에서의 표본분산의 분포(등분산 가정)

- $X_1, \cdots, X_{n_1}$과 $Y_1, \cdots, Y_{n_2}$은 서로 독립이고 각각 $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$에서의 랜덤표본이라 가정하자.
- 각 표본의 표본분산은 다음과 같고,

$$S_1^2 = \sum_{i=1}^{n_1}(X_i - \bar{X})^2/(n_1 - 1), \quad S_2^2 = \sum_{i=1}^{n_2}(Y_i - \bar{Y})^2/(n_2 - 1)$$

- 각각의 분포는 서로 독립이고 다음과 같다.

$$\frac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi^2(n_1 - 1), \quad \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2(n_2 - 1)$$

- 따라서 카이제곱분포의 가법성에 의하여 다음이 성립한다.

$$\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2) \tag{3}$$

두 정규모집단에서의 표본분산의 분포(등분산 가정)

- 이때, S_p^2을 다음과 같이 정의하면

$$
\begin{aligned}
S_p^2 &= \frac{(n_1-1)}{(n_1-1)+(n_2-1)}S_1^2 + \frac{(n_2-1)}{(n_1-1)+(n_2-1)}S_2^2 \\
&= \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}
\end{aligned}
$$

- (3)의 식을 다음과 같이 나타낼 수 있다.

$$
\frac{(n_1+n_2-2)S_p^2}{\sigma^2} \sim \chi^2(n_1+n_2-2)
$$

- S_p^2은 공통분산 σ^2의 추정량으로서 모분산이 σ^2인 두 랜덤표본을 모두 이용하여 만든 것이다. 이를 두 랜덤표본의 합동표본분산(pooled sample variance)이라고 한다.

7장. 범주형자료의 분석

박명현

서울대학교 학부대학

카이제곱검정(Chi-Squared Tests)

- Three statistical techniques are presented, to analyze categorical data.
- Those tests use the χ^2 as the sampling distribution of the test statistic.

1. 적합도 검정(goodness of fit test)
   : 표본으로 얻어진 범주의 분포가 주어진 분포에 부합하는지 검정
2. 동질성 검정(test of homogeneity)
   : 두 개 이상의 집단에서의 분포가 같은지 검정
3. 독립성 검정(test of independence)
   : 범주형 변수들끼리의 독립성 여부를 검정

범주형 변수와 분할표

- 범주형 자료(categorical data)

  - 측정값이 몇 가지 범주(category) 중 하나로 기록되는 자료

  - 명목형, 순서형 자료

  예) 사람의 혈액형을 측정하는 경우 (A, B, AB, O)

  예) 직업에 대한 만족도에 대해 측정하는 경우 (매우 불만족, 다소 불만족, 다소 만족, 매우 만족)

- 분할표(contingency table)

  - 무작위로 추출한 20명에 대해 혈액형(4개의 범주)을 조사하여 표로 나타낸 자료

  - 범주형 자료의 도수분포표

	A	B	O	AB	합
인원수	6	8	4	2	20
비율	0.3	0.4	0.2	0.1	1

범주형 변수와 분할표

- 분할표(contingency table)

  - 무작위로 추출한 20명에 대해 혈액형(4개의 범주)과 성별(2개의 범주)을 조사하여 표로 나타낸 자료

	A	B	O	AB	합
F	4 (0.2)	4 (0.2)	1 (0.05)	1 (0.05)	10 [0.5]
M	2 (0.1)	4 (0.2)	3 (0.15)	1 (0.05)	10 [0.5]
합	6 [0.3]	8 [0.4]	4 [0.2]	2 [0.1]	20

  - ( )안의 숫자는 결합확률(joint probability)
  - [ ]안의 숫자는 주변확률(marginal probability)

- 위 표본결합분포로 모집단의 결합분포를 추정할 수 있다.
- 표본에서의 연관성 여부가 완전한 모집단에서의 연관성에 대한 증거가 될 수 있는가?

다항분포(multinomial distribution)

Definition (Multinomial Distribution)

하나의 시행에서 여러 범주 중 하나가 일어날 수 있을 때, 이를 독립적으로 n번 반복했을 때 각 범주의 발생 횟수를 나타내는 확률분포를 다항분포, $(X_1, X_2, \ldots, X_k) \sim Multinomial(n; p_1, p_2, \ldots, p_k)$라고 하고, 이의 확률질량함수는 다음과 같다.

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_k = x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k},$$

- k: 범주의 개수, n: 전체 시행 횟수
- p_i: i번째 범주의 발생 확률, $p_i \geq 0$ 이며 $\sum_{i=1}^{k} p_i = 1$
- X_i: i번째 범주의 관측 횟수, $\sum_{i=1}^{k} X_i = n$
- $E(X_i) = np_i$, $\quad Var(X_i) = np_i(1 - p_i)$

χ^2 Tests

Theorem

만약 $X_1, X_2, \ldots, X_k$ 가 다항분포 $M(n, p_1, p_2, \ldots, p_k)$를 따르고, 표본의 크기 n이 충분히 커서 $np_i \geq 5$ $(i = 1, 2, \ldots, k)$를 만족한다면,

$$\sum_{i=1}^{k} \left(\frac{(X_i - E(X_i))^2}{E(X_i)} \right) = \sum_{i=1}^{k} \left(\frac{(X_i - np_i)^2}{np_i} \right) \sim \chi^2(k-1)$$

근사적으로 자유도 $k-1$인 카이제곱분포를 따른다.

적합도 검정(goodness of fit test) -모집단의 분포가 알려진 경우-

- 주거형태, 주식의 등급, 사고발생 요일 등 하나의 범주형 변수의 경우 각 범주에 속하는 분포에 관심이 있을 수 있다.
- 과거자료 또는 추측에 의해 가정할 수 있는 분포를 실제 표본과 비교하여 가정된 분포가 맞는지 검정한다.
- 적합도 검정(GOF, goodness of fit test):

H_0: 표본은 가정된 분포를 따른다.

H_1 : 표본은 가정된 분포를 따르지 않는다.

- 가정된 분포가 적합한지 검정하는 것.

- 각 범주의 관측도수(observed frequency)와 H_0하에서의 추정 기대도수(estimated expected frequency)의 차이를 비교한다.

적합도 검정(goodness of fit test) -모집단의 분포가 알려진 경우-

예) 다음의 표는 제품 4개에 대해 알려진 시장점유율과 실제 500명을 대상으로 사용하고 있는 제품의 사용자수를 나타낸 것이다.

Table: 제품의 시장점유율

제품	A	B	C	D	합계
시장점유율	20%	40%	10%	30%	100%
표본수	110	195	47	148	500

- k: 전체 범주의 수
- p_i: 전체 k개의 범주 중 i번째 범주의 기대비율
- O_i: 실제 표본으로부터 얻어진 i번째 범주의 관측(observed)도수
- E_i: 표본수 n에 대해 i번째 범주에 기대되는(expected) 갯수
- i번째 범주의 기댓값: $E_i = n \times p_i$

적합도 검정(goodness of fit test)

- 적합도 검정의 가설:

$$H_0 : p_1 = 0.2,\ p_2 = 0.4,\ p_3 = 0.1,\ p_4 = 0.3$$

H_1: At least one p_i is not equal to its specified value

- 분포에 대한 가정: $X_1, X_2, \cdots, X_k \sim M(n, p_1, p_2, \cdots, p_k)$
- Under H_0,

  - 적합도 검정 통계량의 근사분포:

$$\chi^2 = \sum_{i=1}^{k} \left(\frac{(O_i - E_i)^2}{E_i} \right) \dot{\sim} \chi^2(k-1)$$

- 기댓값(E_i)과 관측값(O_i)이 차이가 많이나면 $(O_i - E_i)^2$ 값과 검정 통계량의 관측값이 커지게 되고, H_0을 벗어난다고 할 수 있다.

적합도 검정(goodness of fit test)

- 검정통계량:

$$\chi^2 = \sum_{i=1}^{k} \left(\frac{(O_i - E_i)^2}{E_i} \right) = \sum_{i=1}^{k} \left(\frac{(O_i - np_i)^2}{np_i} \right) \dot{\sim} \chi^2(k-1)$$

이의 관측값을 χ_0^2이라고 하면 다음이 성립한다.

유의확률	유의수준 α의 기각역
$P = P(\chi^2 \geq \chi_0^2),\ \chi^2 \sim \chi^2(k-1)$	$\chi_0^2 \geq \chi_\alpha^2(k-1)$

- 이 때, 각 범주에서의 기대값 E_i이 5 이상으로 가정하고, 이를 만족하지 않으면 범주를 통합하여 검정한다.

적합도 검정(goodness of fit test)

예) (continue..) Under H_0, 기댓값과 실제 관측값의 차이를 계산하면

제품	A	B	C	D	합계
시장점유율(p_i)	20%	40%	10%	30%	100%
제품수(O_i)	110	195	47	148	500
기댓값(E_i)	100	200	50	150	500
$\frac{(O_i-E_i)^2}{E_i}$	1	0.125	0.18	0.027	1.33

- 검정통계량의 관측값 $\chi_0^2 = \sum_{i=1}^{k}\left(\frac{(O_i-np_i)^2}{np_i}\right) = 1.33$이고
- $0.5 < P(\chi^2 \geq \chi_0^2) < 0.75$이므로 H_0를 기각할 수 없다.
- 따라서 표본자료는 알려진 시장점유율을 통계적으로 부정할 만한 증거가 되지 못한다.

동질성 검정(test of homogeneity)

- 동질성 검정: 두 개 이상의 집단에서 분포가 같은지 여러 모집단을 비교, 분석하는 방법.
- r개의 모집단과 c개의 범주(category) 가정
- 동질성검정의 통계적 가설

  H_0: 여러 모집단의 분포가 동일하다.

  H_1 : 이들이 서로 다르다.

- 위 가설에 대한 검정은 카이제곱분포에 근사하여 수행.
- 각 범주의 관측도수(observed)가 귀무가설 H_0하에서의 추정 기대도수(expected)와 차이가 많이 나면 H_0를 기각.

동질성 검정(test of homogeneity)

- c개의 범주로 나뉘어진 r개의 다항모집단을 관측한 결과가 다음과 같이 $r \times c$ 분할표(contingency table)로 주어져 있다.

Table: r개의 다항모집단에 대한 관측도수(O_{ij})

	범주 1	범주 2	$\cdot$	범주 j	$\cdot$	범주 c	표본크기
모집단 1의 표본	O_{11}	O_{12}	$\cdot$	O_{1j}	$\cdot$	O_{1c}	n_1
$\cdots$	$\cdots$	$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdots$	$\cdots$
모집단 i의 표본	O_{i1}	O_{i2}	$\cdot$	O_{ij}	$\cdot$	O_{ic}	n_i
$\cdots$	$\cdots$	$\cdots$	$\cdot$	$\cdots$	$\cdot$	$\cdots$	$\cdots$
모집단 r의 표본	O_{r1}	O_{r2}	$\cdot$	O_{rj}	$\cdot$	O_{rc}	n_r
합계	$O_{\cdot 1}$	$O_{\cdot 2}$	$\cdot$	$O_{\cdot j}$	$\cdot$	$O_{\cdot c}$	n

- O_{ij} = 표에서 (i,j)칸의 관측도수(observed frequency)(or 표본수)
- $O_{\cdot j} = \sum_{i=1}^{r} O_{ij}$ (j번째 열의 합)
- $O_{i\cdot} = \sum_{j=1}^{c} O_{ij} = n_i$ (i번째 행의 합)

동질성 검정(test of homogeneity)

Table: r개의 다항모집단에 대한 분할표 (i,j)칸에서의 모비율 p_{ij}

	범주 1	범주 2	$\cdot$	범주 j	$\cdot$	범주 c
모집단 1의 표본	p_{11}	p_{12}	$\cdot$	p_{1j}	$\cdot$	p_{1c}
모집단 2의 표본	p_{21}	p_{22}	$\cdot$	p_{2j}	$\cdot$	p_{2c}
$\cdots$	$\cdots$	$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdots$
모집단 i의 표본	p_{i1}	p_{i2}	$\cdot$	p_{ij}	$\cdot$	p_{ic}
$\cdots$	$\cdots$	$\cdots$	$\cdot$	$\cdots$	$\cdot$	$\cdots$
모집단 r의 표본	p_{r1}	p_{r2}	$\cdot$	p_{rj}	$\cdot$	p_{rc}
추정 모비율	$\hat{p}_1 = \frac{O_{\cdot 1}}{n}$	$\hat{p}_2 = \frac{O_{\cdot 2}}{n}$	$\cdot$	$\hat{p}_j = \frac{O_{\cdot j}}{n}$	$\cdot$	$\hat{p}_c = \frac{O_{\cdot c}}{n}$

- $p_j (= p_{1j} = \cdots = p_{rj})$ (Under H_0, 각 범주의 공통 모비율)
- $\hat{p}_j = \frac{O_{\cdot j}}{n}$ (범주 j에서 공통 모비율의 추정값)
- $\sum_{j=1}^{c} p_{ij} = 1$ (i^{th} 다항모집단의 각 범주에 대한 모비율의 합은 1)

동질성 검정(test of homogeneity)

Table: r개의 다항모집단에 대한 기대도수($\hat{E}_{ij}$)

	범주 1	범주 2	$\cdot$	범주 j	$\cdot$	범주 c	표본크기
모집단 1의 표본	$\hat{E}_{11}$	$\hat{E}_{12}$	$\cdot$	$\hat{E}_{1j}$	$\cdot$	$\hat{E}_{1c}$	n_1
모집단 2의 표본	$\hat{E}_{21}$	$\hat{E}_{22}$	$\cdot$	$\hat{E}_{2j}$	$\cdot$	$\hat{E}_{2c}$	n_2
$\cdots$	$\cdots$	$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdots$	$\cdots$
모집단 i의 표본	$\hat{E}_{i1}$	$\hat{E}_{i2}$	$\cdot$	$\hat{E}_{ij}$	$\cdot$	$\hat{E}_{ic}$	n_i
$\cdots$	$\cdots$	$\cdots$	$\cdot$	$\cdots$	$\cdot$	$\cdots$	$\cdots$
모집단 r의 표본	$\hat{E}_{r1}$	$\hat{E}_{r2}$	$\cdot$	$\hat{E}_{rj}$	$\cdot$	$\hat{E}_{rc}$	n_r
추정 모비율	$\hat{p}_1$	$\hat{p}_2$	$\cdot$	$\hat{p}_j$	$\cdot$	$\hat{p}_c$	

- $\hat{E}_{ij}$: H_0하에서의 추정 기대도수(estimated expected frequency)
- $\hat{E}_{ij} = n_i \times \hat{p}_j$($H_0$하에서의 추정 확률)

동질성 검정(test of homogeneity)

- 동질성 검정의 가설

$$H_0 : p_{1j} = p_{2j} = \cdots = p_{rj} \ (j = 1, 2, \cdots, c), \quad H_1 : not \ H_0$$

- 분포에 대한 가정:

$$X_{i1}, X_{i2}, \cdots, X_{ic} \ \sim \ M(n_i, p_{i1}, p_{i2}, \cdots, p_{ic}), \quad i = 1, \cdots, r$$

- Under H_0, p_j의 추정값은 다음과 같이 합동표본비율로 주어지고

$$\hat{p}_j = \frac{O_{1j} + O_{2j} + \cdots + O_{rj}}{n_1 + n_2 + \cdots + n_r} = \frac{O_{\cdot j}}{n}$$

- i번째 집단에서의 추정 기대도수는 다음과 같다.$(i = 1, \cdots, r)$

$$\hat{E}_{i1} = n_i \hat{p}_1 = n_i \frac{O_{\cdot 1}}{n}, \quad \hat{E}_{i2} = n_i \hat{p}_2 = n_i \frac{O_{\cdot 2}}{n}, \quad \cdots, \quad \hat{E}_{ic} = n_i \hat{p}_c = n_i \frac{O_{\cdot c}}{n}$$

동질성 검정(test of homogeneity)

- 관측도수 O_{ij}와 추정 기대도수 $\hat{E}_{ij}$의 차이를 나타내는 다음의 카이제곱통계량은

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}}$$

  Under H_0, 표본크기가 충분히 크면 근사적으로 다음과 같이 카이제곱분포를 따른다고 알려져 있다.

$$\chi^2 \dot{\sim} \chi^2((r-1)(c-1))$$

- 각 범주의 관측도수 O_{ij}가 귀무가설 H_0하에서의 추정 기대도수 $\hat{E}_{ij}$와 차이가 많이 나면 χ^2 관측값도 커지게 되고 H_0를 기각하게 된다.

동질성 검정(표본크기가 큰 경우: $\hat{E}_{ij} \geq 5$)

- c개의 범주로 나뉘어지는 다항모집단 r개를 비교하기 위한 가설

$$H_0 : p_{1j} = p_{2j} = \cdots = p_{rj} \ (j = 1, 2, \cdots, c) \quad H_1 : not\ H_0$$

- 검정통계량:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \quad (\hat{E}_{ij} = n_i O_{\cdot j}/n)$$

이의 관측값을 χ_0^2이라고 하면 다음이 성립한다.

유의확률	유의수준 α의 기각역
$P = P(\chi^2 \geq \chi_0^2),\ \chi^2 \sim \chi^2((r-1)(c-1))$	$\chi_0^2 \geq \chi_\alpha^2((r-1)(c-1))$

동질성 검정(test of homogeneity)

예) [예제 7-7]에서 충동적 살인범과 계획적 살인범의 가석방 성공률에 차이가 있는지 유의수준 5%에서 검정하였다.

$$H_0 : p_1 = p_2, \quad H_1 : p_1 \neq p_2$$

- 동질성 검정: 충동적 살인범과 계획적 살인범의 모집단이 동일하다.

$$H_0 : p_{11} = p_{21},\ p_{12} = p_{22} \quad H_1 : not\ H_0$$

	성공(j=1)	실패(j=2)	표본크기
충동적 살인범(i=1)	O_{11} =13	O_{12} =29	$n_1 = 42$
계획적 살인범(i=2)	O_{21} =22	O_{22} =18	$n_2 = 40$
합계	35	47	$n = 82$

동질성 검정(test of homogeneity)

- Under H_0, $\hat{p}_1 = 35/82 = 0.427$, $\hat{p}_2 = 47/82 = 0.573$이므로,

  - 충동적 살인범에 대한 가석방이 성공할 $\hat{E}_{11}$와 실패할 $\hat{E}_{12}$는

$$\hat{E}_{11} = n_1\hat{p}_1 = 42 \times 35/82 = 17.93$$
$$\hat{E}_{12} = n_1\hat{p}_2 = 42 \times 47/82 = 24.07$$

  - 계획적 살인범에 대한 가석방이 성공할 $\hat{E}_{21}$와 실패할 $\hat{E}_{22}$는

$$\hat{E}_{21} = n_2\hat{p}_1 = 40 \times 35/82 = 17.08$$
$$\hat{E}_{22} = n_2\hat{p}_2 = 40 \times 47/82 = 22.92$$

- 추정 기대도수와 관측도수를 표로 정리하면 다음과 같다.

	성공	실패
충동적 살인범	$O_{11} = 13(\hat{E}_{11} = 17.93)$	$O_{12} = 29(\hat{E}_{12} = 24.07)$
계획적 살인범	$O_{21} = 22(\hat{E}_{21} = 17.08)$	$O_{22} = 18(\hat{E}_{22} = 22.92)$
합계	35	47

동질성 검정(test of homogeneity)

- 카이제곱통계량 χ^2의 관측값을 구해보자.

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} = \frac{(13-17.93)^2}{17.93} + \cdots + \frac{(18-22.92)^2}{22.92} = 4.84$$

- Under H_0, 검정통계량 χ^2는 근사적으로 $\chi^2(1)$을 따르고, 유의수준 5%에서 기각역은 다음과 같아서 귀무가설을 기각하게 된다.

$$\chi^2 \geq \chi^2_{0.05}(1) = 3.84$$

Z 검정통계량과의 비교($r = c = 2$)

- 그런데 카이제곱통계량 χ^2의 관측값은 다음의 예제 7.7에서의 Z 검정통계량의 제곱과 같다.

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(n_1^{-1} + n_2^{-1})}} = -2.20$$

- 즉, 다음의 관계가 일반적으로 성립하며,

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} = z^2 \ \sim \ \chi^2(1)$$

H_0하에서 Z가 근사적으로 표준정규분포를 따르므로 그것의 제곱꼴은 자유도가 1인 카이제곱분포를 따르게 됨을 알 수 있다.

다항모집단의 동질성검정 예제

Example

액화천연가스(LNG) 저장 기지의 후보로 고려되는 세 지역의 여론을 조사하기 위해 각 지역에서 각각 400명, 350명, 350명을 랜덤추출하여 기지건설에 대한 찬반을 조사했다. 조사결과가 [표 7-5]와 같을 때, 지역에 따라 찬성률에 차이가 있다고 할 수 있는지 5% 유의수준으로 검정하여라.

표 7-5 LNG 저장 기지 건설에 대한 찬반 조사 결과

(단위 : 명)

	찬성	반대	표본 크기
지역 1	198	202	$n_1 = 400$
지역 2	140	210	$n_2 = 350$
지역 3	133	217	$n_3 = 350$
합계	471	629	$n = 1100$

sol) 세 지역의 찬성률 p_{11}, p_{21}, p_{31}에 차이가 있는지 검정하기 위해 다음과 같은 가설을 세운다.

$$H_0 : p_{11} = p_{21} = p_{31}, \qquad H_1 : not\ H_0$$

- H_0하에서 $\hat{p}_1 = 471/1100,\ \hat{p}_2 = 629/1100$이고,
- 추정 기대도수를 구해보면

$$\hat{E}_{11} = 400 \times (471/1100) = 171.27, \quad \hat{E}_{12} = 400 \times (629/1100) = 228.73$$

$$\hat{E}_{21} = 350 \times (471/1100) = 149.86, \quad \hat{E}_{22} = 350 \times (629/1100) = 200.14$$

$$\hat{E}_{31} = 350 \times (471/1100) = 149.86, \quad \hat{E}_{32} = 350 \times (629/1100) = 200.14$$

- 이로부터 카이제곱통계량의 관측값을 계산해보자.

$$\sum_{i=1}^{3}\sum_{j=1}^{2}\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} = \frac{(198-171.27)^2}{171.27}+\cdots+\frac{(217-200.14)^2}{200.14} = 11.75$$

- P값은 다음과 같고 카이제곱분포표로부터 유의수준 5% 이하인 것을 알 수 있다.

$$0.0025 < P = P(\chi^2 \geq 11.75) < 0.005, \quad \chi^2 \sim \chi^2(2)$$

- 즉 지역에 따라 LNG기지 건설에 대한 찬성률에 차이가 있음을 뚜렷이 나타낸다고 할 수 있다.

다항모집단의 동질성검정 예제

Example

공단에 인접한 세 지역에서 공해를 느끼는 정도가 지역에 따라 차이가 있는지를 확인하기 위해, 각 지역에서 97명, 95명, 99명을 랜덤추출하여 산업공해로 인한 악취를 느끼는 빈도를 조사한 결과가 다음 표와 같다. 표에서 범주는 다음과 같이 정의된다.

* 범주 1: 매일 * 범주 2: 적어도 일주일에 한 번 * 범주 3: 적어도 한달에 한 번

* 범주 4: 한 달에 한 번보다 적게 * 범주 5: 전혀 악취를 느끼지 않음

이 자료를 이용하여 지역에 따라 공해를 느끼는 정도가 다르다고 할 수 있는지를 유의수준 1%에서 검정하여라.

표 7-7 지역별 공해 인식 빈도

(단위 : 명)

	범주 1	범주 2	범주 3	범주 4	범주 5	표본 크기
지역 1	20	28	23	14	12	$n_1 = 97$
지역 2	14	34	21	14	12	$n_2 = 95$
지역 3	4	12	10	20	53	$n_3 = 99$
합계	38	74	54	48	77	$n = 291$

- 지역 i의 각 범주 j에 대한 모비율을 p_{ij} $(i = 1, 2, 3,\ j = 1, 2, 3, 4, 5)$이라고 하면, 가설은 다음과 같다.

$$H_0 : p_{1j} = p_{2j} = p_{3j} (j = 1, 2, \cdots, 5) \quad H_1 : not\ H_0$$

- H_0하에서 공통인 각 범주의 모비율에 대한 추정값은

$$\hat{p}_1 = 38/291, \hat{p}_2 = 74/291, \hat{p}_3 = 54/291, \hat{p}_4 = 48/291, \hat{p}_5 = 77/291$$

- 이를 이용하여, H_0하에서의 추정 기대도수는 다음과 같고, 아래 표에 그 값이 정리가 되어 있다.

$$\hat{E}_{1j} = 97\hat{p}_j, \hat{E}_{2j} = 95\hat{p}_j, \hat{E}_{3j} = 99\hat{p}_j \quad (j = 1, 2, \cdots, 5)$$

	범주 1	범주 2	범주 3	범주 4	범주 5
지역 1	12.67	24.67	18.00	16.00	25.67
지역 2	12.41	24.16	17.63	15.67	25.14
지역 3	12.93	25.18	18.37	16.33	26.20

- 카이제곱통계량의 관측값을 계산해보자.

$$\chi_0^2 = \sum_{i=1}^{3}\sum_{j=1}^{5}\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} = \frac{(20-12.67)^2}{12.67} + \cdots + \frac{(53-26.20)^2}{26.20}$$

$$= 71.36$$

- 모든 기대도수가 5 이상이므로 $\chi^2 \sim \chi^2((3-1)(5-1))$

- 유의수준 1%의 기각역은 카이제곱분포표로부터 다음과 같고

$$\chi_0^2 \geq \chi_{0.01}^2(8) = 20.09$$

관측값이 기각역에 속하므로, 지역에 따라 공해를 느끼는 정도에 차이가 있다고 유의수준 1%에서 말할 수 있다.

범주형 자료에 의한 독립성 검정

- 한 모집단의 각 개체에 대하여 두가지 특성 A, B를 관측
  예) 혈액형과 성별, 학년과 학점 등 두 개의 범주형 변수를 관측
  $\Longrightarrow$ 이들 특성 간의 관련성이 주된 관심.
- 독립성 검정: 한개의 모집단에서 각 개체에 대해 두 특성 A, B를 관측한 경우 이들 특성이 서로 독립(or 관련성)인지 검정하는 방법
  - 사건 A와 사건 B가 독립이면 다음이 성립

$$P(A \cap B) = P(A)P(B)$$

- 동질성 검정: r개의 모집단에서 하나의 범주형 특성을 관측한 경우, 여러 모집단 별 범주형 특성의 분포가 같은지 검정하는 방법

범주형 자료에 의한 독립성 검정

- 특성 A, B가 각각 r개와 c개의 범주로 나뉘어지는 경우에 두 특성에 대한 관측도수는 다음의 $r \times c$분할표로 주어진다.

Table: 두 특성 A, B의 관측도수(O_{ij})

	B_1	B_2	$\cdots$	B_j	$\cdots$	B_c	합계
A_1	O_{11}	O_{12}	$\cdots$	O_{1j}	$\cdots$	O_{1c}	$O_{1\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_i	O_{i1}	O_{i2}	$\cdots$	O_{ij}	$\cdots$	O_{ic}	$O_{i\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_r	O_{r1}	O_{r2}	$\cdots$	O_{rj}	$\cdots$	O_{rc}	$O_{r\cdot}$
합계	$O_{\cdot 1}$	$O_{\cdot 2}$	$\cdots$	$O_{\cdot j}$	$\cdots$	$O_{\cdot c}$	n

독립성 검정(test of independence)

Table: 두 특성 A, B의 모비율(p_{ij})

	B_1	B_2	$\cdots$	B_j	$\cdots$	B_c	합계
A_1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$	p_{1c}	$p_{1\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$	p_{ic}	$p_{i\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_r	p_{r1}	p_{r2}	$\cdots$	p_{rj}	$\cdots$	p_{rc}	$p_{r\cdot}$
합계	$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot j}$	$\cdots$	$p_{\cdot c}$	1

- 두 특성 A, B가 서로 독립인 경우: $p_{ij} = p_{i\cdot}p_{\cdot j}$
- 분할표 (i, j)셀에서의 추정된 모비율

$$\hat{p}_{ij} = \hat{p}_{i\cdot}\hat{p}_{\cdot j} = (O_{i\cdot}/n)(O_{\cdot j}/n)$$

독립성 검정(test of independence)

Table: 두 특성 A, B의 기대도수($\hat{E}_{ij}$)

	B_1	B_2	$\cdots$	B_j	$\cdots$	B_c	합계
A_1	$\hat{E}_{11}$	$\hat{E}_{12}$	$\cdots$	$\hat{E}_{1j}$	$\cdots$	$\hat{E}_{1c}$	$O_{1\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_i	$\hat{E}_{i1}$	$\hat{E}_{i2}$	$\cdots$	$\hat{E}_{ij}$	$\cdots$	$\hat{E}_{ic}$	$O_{i\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_r	$\hat{E}_{r1}$	$\hat{E}_{r2}$	$\cdots$	$\hat{E}_{rj}$	$\cdots$	$\hat{E}_{rc}$	$O_{r\cdot}$
합계	$O_{\cdot 1}$	$O_{\cdot 2}$	$\cdots$	$O_{\cdot j}$	$\cdots$	$O_{\cdot c}$	n

- 독립일 때 (i,j)셀에서의 추정 기대도수

$$\hat{E}_{ij} = n\hat{p}_{ij} = O_{i\cdot}O_{\cdot j}/n$$

범주형 자료에 의한 독립성 검정

- O_{ij}와 $\hat{E}_{ij}$를 이용하여 다음의 카이제곱통계량을 구할 수 있고

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}}$$

- H_0가 사실일 때 표본크기가 충분히 크면 근사적으로 다음과 같이 자유도가 $(r-1)(c-1)$인 카이제곱분포를 따른다고 알려져 있다.

$$\chi^2 \dot{\sim} \chi^2((r-1)(c-1))$$

- 동질성 검정과 독립성 검정

  : 두 검정은 검정통계량의 형태와 검정방법은 동일하지만, 가설 및 표본의 추출 형태가 서로 다르므로 문제의 뜻에 맞게 표본을 추출하고 그에 맞는 추론을 하여야 한다.

두 특성의 독립성 검정(표본크기가 큰 경우: $\hat{E}_{ij} \geq 5$)

- 독립성 검정에 대한 가설:

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j} \quad (i = 1, \cdots, r)(j = 1, \cdots, c), \quad H_1 : not\ H_0$$

- 분포에 대한 가정:

$$X_{11}, \cdots, X_{rc} \ \sim \ M(n, p_{11}, \cdots, p_{rc})$$

- 검정통계량:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \quad (\hat{E}_{ij} = O_{i\cdot}O_{\cdot j}/n)$$

- 검정통계량의 관측값을 χ_0^2이라고 하면 다음이 성립한다.

유의확률	유의수준 α의 기각역
$P = P(\chi^2 \geq \chi_0^2),\ \chi^2 \sim \chi^2((r-1)(c-1))$	$\chi_0^2 \geq \chi_\alpha^2((r-1)(c-1))$

두 특성의 독립성 검정 예제

Example

대도시 근교에서 출퇴근하며 혼자 승용차를 이용하는 사람들 중 250명을 랜덤추출하여 이들을 승용차 크기와 통근 거리로 분류한 결과가 [표 7-10]과 같다. 이 자료를 이용하여 승용차 크기와 통근 거리 사이 관계가 있는지를 유의수준 5%에서 검정하여라.

표 7-10 승용차 크기와 통근 거리

(단위 : 대)

	15 km 미만	15 km 이상 30 km 미만	30 km 이상	합계
경승용차	6	27	19	52
소형승용차	8	36	17	61
중형승용차	21	45	33	99
대형승용차	14	18	6	38
합계	49	126	75	250

sol) 표에서 (i,j)칸에 대응하는 모비율을 p_{ij}라고 하면, 가설은 다음과 같다.

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j} \quad (i = 1, \cdots, 4)(j = 1, 2, 3), \quad H_1 : not\ H_0$$

- H_0하에서 각 칸의 추정 기대도수를 구하면

$$\hat{E}_{11} = \frac{52 \cdot 49}{250} = 10.19,\ \hat{E}_{12} = \frac{52 \cdot 126}{250} = 26.21,\ \hat{E}_{13} = \frac{52 \cdot 75}{250} = 15.60$$

$$\cdots \qquad \cdots \qquad \cdots$$

$$\hat{E}_{41} = \frac{38 \cdot 49}{250} = 7.45,\quad \hat{E}_{42} = \frac{38 \cdot 126}{250} = 19.15,\ \hat{E}_{43} = \frac{38 \cdot 75}{250} = 11.40$$

- 다음으로 카이제곱통계량의 관측값을 계산해보자.

$$\chi_0^2 = \sum_{i=1}^{4}\sum_{j=1}^{3}\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} = \frac{(6-10.19)^2}{10.19} + \cdots + \frac{(6-11.40)^2}{11.40}$$
$$= 14.15$$

- 모든 기대도수가 5 이상이므로 $\chi^2 \sim \chi^2((4-1)(3-1))$

- 유의수준 5%의 기각역은 카이제곱분포표로부터 다음과 같고

$$\chi_0^2 \geq \chi_{0.05}^2(6) = 12.59$$

관측값이 기각역에 속하므로, 유의수준 5%에서 승용차의 크기와 통근 거리 사이에는 관계가 있다고 말할 수 있다.

8장. 상관분석과 회귀분석

박명현

서울대학교 학부대학

서론

- 자료를 이용하여 변수와 변수 사이의 관계에 대한 통계적 추론을 하고자 한다.

  예) 지능지수와 학업성적, 흡연량과 수명 등

- 상관분석(correlation analysis)

  - 두 변수 사이의 관계 유무 또는 관계의 강도에 대한 연관성을 분석하는 방법.

  - 표본상관계수 r을 이용하여 모상관계수 ρ를 추론.

- 단순선형회귀분석(simple linear regression analysis):

  두 변수 사이의 함수 관계로서 가장 기본적인 형태인 일차함수관계, 즉 직선 관계를 모형으로 하여 분석

- 다중회귀분석(multiple regression analysis):

  두 개 이상의 변수가 한 변수에 영향을 줄 때, 한 변수를 여러 변수의 함수로 나타내어 분석하는 방법

상관분석(correlation analysis)

상관분석

- 모상관계수 ρ: 각 개체의 두 가지 특성을 변수 X, Y로 나타낼 때, 두 변수 사이에 직선관계가 어느 정도 강하게 있는가를 나타냄

$$\rho = Corr(X, Y)$$

- 표본상관계수 r: 모상관계수 ρ의 추정값으로 사용.

  - 두 특성에 대한 자료가 $(x_1, y_1), \cdots, (x_n, y_n)$ 으로 주어지면

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

1. 항상 $-1 \leq r \leq 1$
2. $r \approx 1$: 두 변수 사이에 기울기가 양수인 직선관계가 강함
3. $r \approx -1$: 두 변수 사이에 기울기가 음수인 직선관계가 강함
4. $r \approx 0$ 두 변수 사이에 직선관계가 없음

상관분석

- 두 변수의 관계에 대한 대략적인 파악은 산점도로 가능

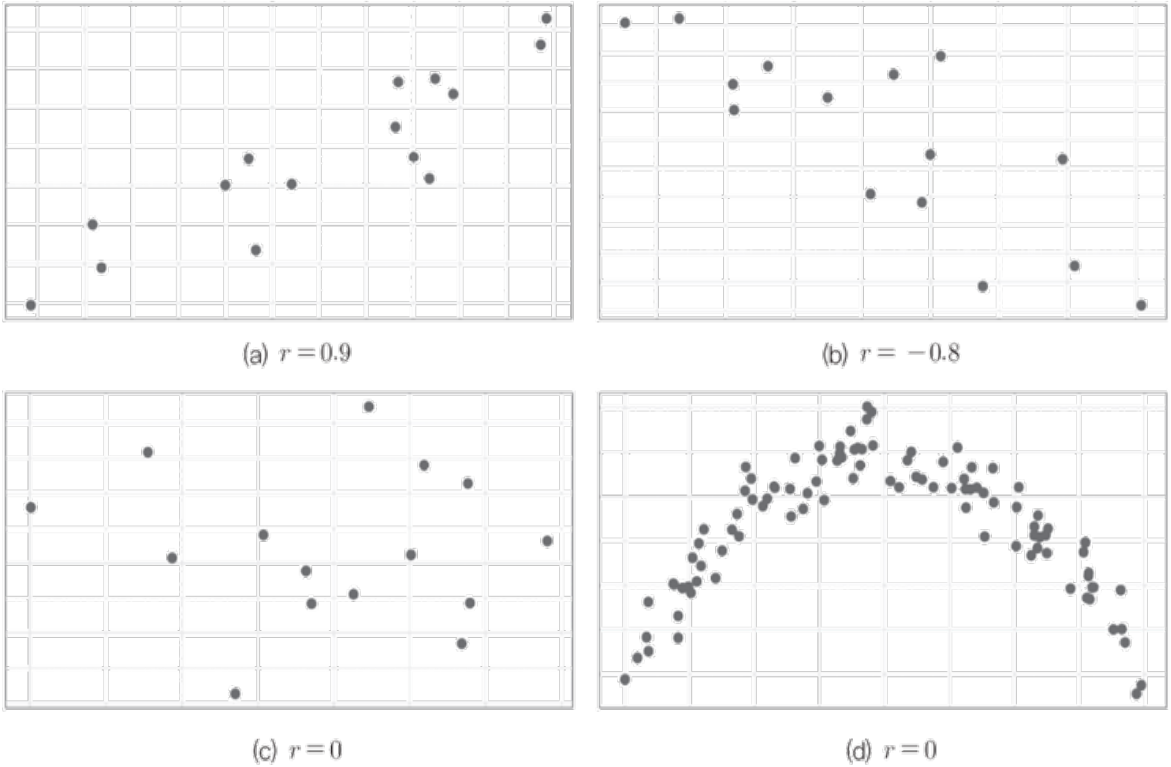


그림 8-1 여러 가지 경우의 산점도와 표본상관계수

- r의 값이 0에 가까운 것은 두 변수 사이에 직선관계가 약한 것을 시사하는 것이지 아무런 관계가 없음을 뜻하는 것은 아니다.

표본상관계수 예제

Example (8.1)

다음 자료는 어느 고등학교 학생 중에서 랜덤하게 추출된 20명의 수학능력 모의시험에서 언어영역과 외국어영역의 점수이다.

학생번호	1	2	3	4	5	6	7	8	9	10
언어 영역	42	38	51	53	40	37	41	29	52	39
외국어 영역	30	25	34	35	31	29	33	23	36	30

학생번호	11	12	13	14	15	16	17	18	19	20
언어 영역	45	34	47	35	44	48	47	30	29	34
외국어 영역	32	29	34	30	28	29	33	24	30	30

이 자료에 대한 산점도를 그리고, 언어영역(x)과 외국어영역(y) 성적 사이의 표본상관계수를 구하고 해석하여라.

sol) 자료로부터 다음의 값을 계산하였다.

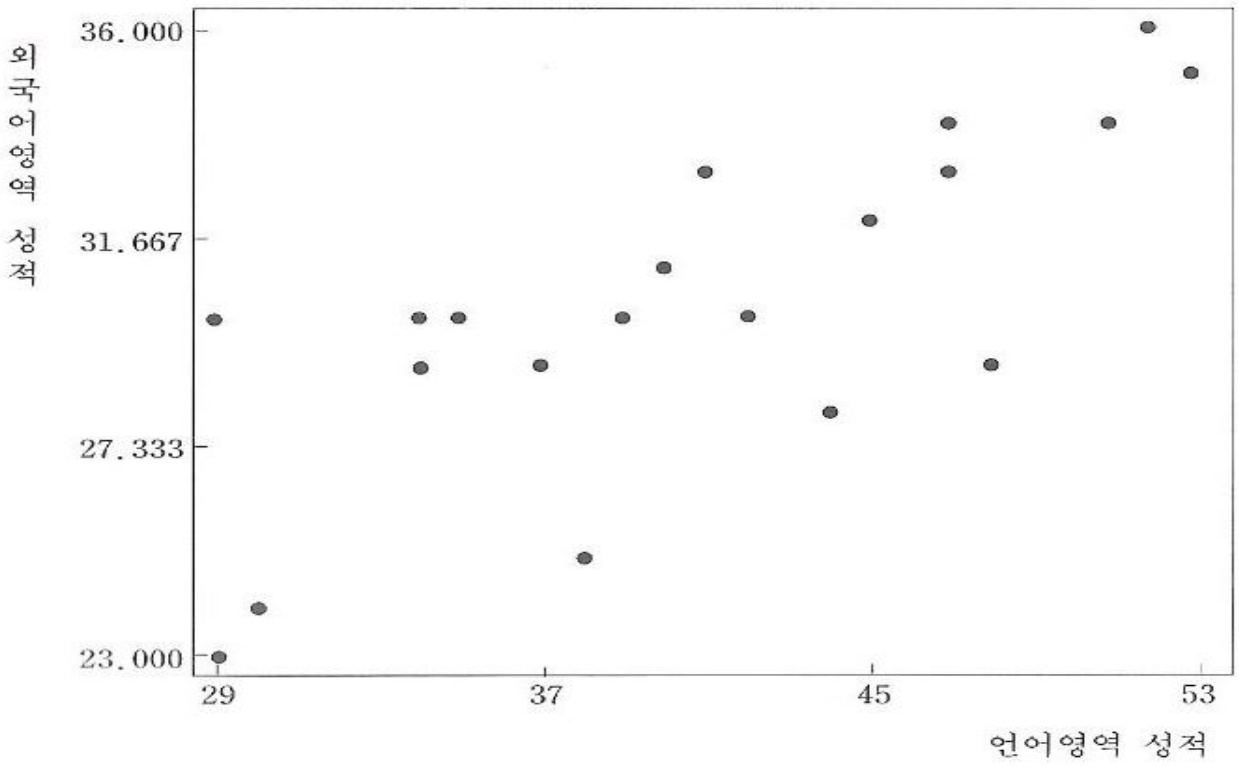
$\bar{x} = 815/20$, $\bar{y} = 605/20$,

$\sum x_i y_i = 25033$, $\sum x_i^2 = 34295$, $\sum y_i^2 = 18533$

- 간편계산식을 이용하여 표본상관계수를 구해보자.

$$
\begin{aligned}
r &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \\
&= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2}\sqrt{\sum y_i^2 - n\bar{y}^2}} \\
&= \frac{25033 - n\bar{x}\bar{y}}{\sqrt{34295 - n\bar{x}^2}\sqrt{18533 - n\bar{y}^2}} = 0.757
\end{aligned}
$$

Figure: 언어영역과 외국어영역 성적간의 산점도



- 두 점수 사이에 양의 선형관계가 있고 그 정도는 76% 이다.
- 이는 두 점수간에 존재하는 상관의 성질과 정도만을 의미하며, 두 변수간의 관계를 인과적으로 설명하지는 않는다.

상관관계의 유무에 관한 검정

- 표본상관계수 r을 이용하여 모상관계수 ρ를 추론
- 가정사항: 이변량정규모집단
- 귀무가설 $H_0 : \rho = 0$에 대한 검정통계량은

$$T = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}$$

이고, 이의 관측값을 t라고 할 때, 다음이 성립한다.

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \rho > 0$	$P = P(T \geq t),\ T \sim t(n-2)$	$t \geq t_\alpha(n-2)$
(b) $H_1 : \rho < 0$	$P = P(T \leq t),\ T \sim t(n-2)$	$t \leq -t_\alpha(n-2)$
(c) $H_1 : \rho \neq 0$	$P = P(\lvert T\rvert \geq \lvert t\rvert),\ T \sim t(n-2)$	$\lvert t\rvert \geq t_{\alpha/2}(n-2)$

상관분석 예제

Example

예제 8.1의 경우에 고등학교 학생 전체의 언어영역과 외국어영역 성적이 이변량정규분포를 따른다고 할 때, 이들 성적 사이의 상관관계가 있는가를 유의수준 5%에서 검정하여라.

sol) 검정하고자 하는 가설은

$$H_0 : \rho = 0 \qquad H_1 : \rho \neq 0$$

이고, $n = 20$, $r = 0.757$이므로 검정통계량의 관측값은

$$t = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}} = \sqrt{20-2}\,\frac{0.757}{\sqrt{1-0.757^2}} = 4.915$$

- t분포표로부터 $t_{0.025}(18) = 2.101$이고

$$|t| = |4.915| \geq 2.101$$

이므로, 유의수준 5%에서 $H_0 : \rho = 0$을 기각한다. 즉, 상관관계가 있다는 뚜렷한 증거가 있다.

회귀분석(regression analysis)

회귀분석

- 회귀분석(regression analysis): 두 변수 사이의 함수 관계를 파악하여 한 변수의 값으로부터 다른 변수의 값에 대한 예측.

$$Y = f(x) + \epsilon$$

- 설명변수(explanatory variable): x, 다른 변수에 영향을 주는 변수
  - 예측변수(predictor variable), 독립변수(independent variable)
- 반응변수(response variable): Y, 설명변수에 영향을 받는 변수
  - 종속변수(dependent variable)
- 회귀분석의 목적: 주어진 자료를 이용하여 설명변수와 반응변수의 관계를 구체적인 함수 $f(x)$ 형태로 나타내고, 설명변수의 값으로부터 반응변수의 값을 예측 $\hat{f}(x)$ 하고자 함

단순선형회귀분석 예시

Example (8.2)

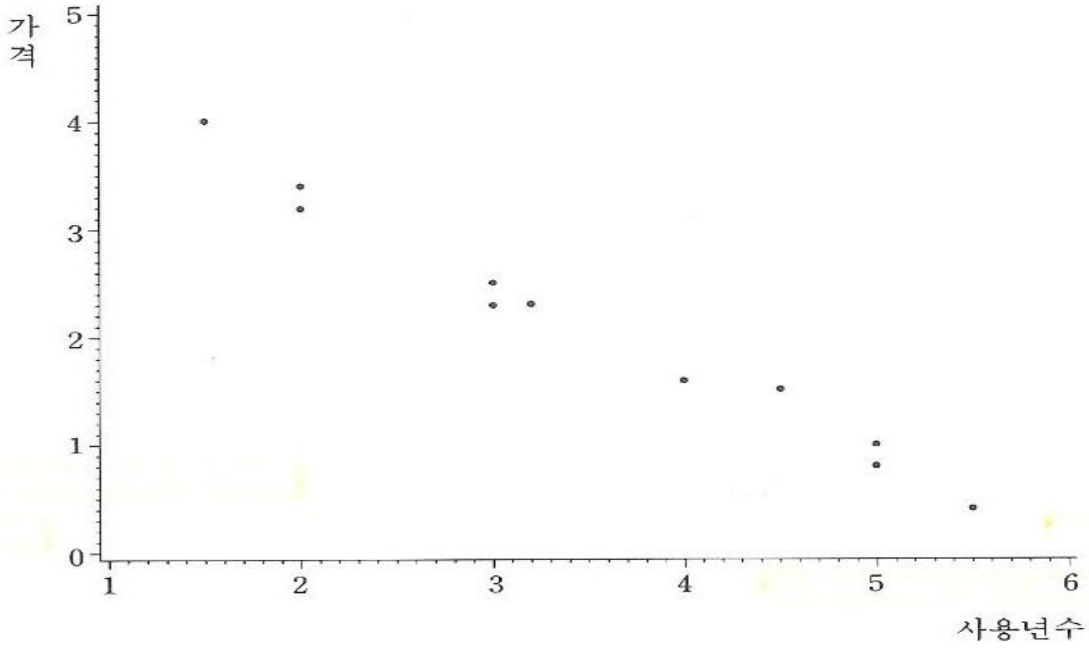
어떤 승용차의 가격이 출고 연도가 지남에 따라 얼마나 떨어지는가를 알아보기 위하여, 이 승용차에 대한 중고판매가격에 대한 표본조사를 한 결과 다음의 자료를 얻었다.

Table: [8.1] 사용년수에 따른 중고차 가격 (단위: 백만 원)

사용년수(x)	1.0	1.5	2.0	2.0	3.0	3.0	3.2	4.0	4.5	5.0	5.0	5.5
중고차가격(y)	4.5	4.0	3.2	3.4	2.5	2.3	2.3	1.6	1.5	1.0	0.8	0.4

- 이 자료로부터 사용년수(x)와 중고차가격(y)의 대략적인 관계의 파악을 위하여 산점도를 그려보았다.
- 설명변수: $(x_1, \cdots, x_{12})$, 반응변수: $(y_1, \cdots, y_{12})$

Figure: 사용년수에 따른 중고차 가격의 변화



- 사용년수가 증가함에 따라 중고차 가격은 점차 감소: $5.1 - 0.8x$
- Y 변수에 있는 변동은 오차항 ϵ으로 설명 가능
- 단순선형회귀모형의 가정: $Y = \beta_0 + \beta_1 x + \epsilon$

단순선형회귀모형(simple linear regression model)

- 설명변수의 값에 대응하는 반응변수의 값이 $\beta_0 + \beta_1 x$주위에 나타난다고 생각될 때, 다음의 선형식을 고려해 보자.

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \cdots, n$$

- Y_i: 종속변수, 반응변수
- x_i: 독립변수, 설명변수, 주어진 상수(constant)값
- 절편 β_0과 기울기 β_1: 미지의 모수(unknown parameters)
- ϵ_i: 오차. Y와 x의 선형관계로 설명할 수 없는 나머지부분

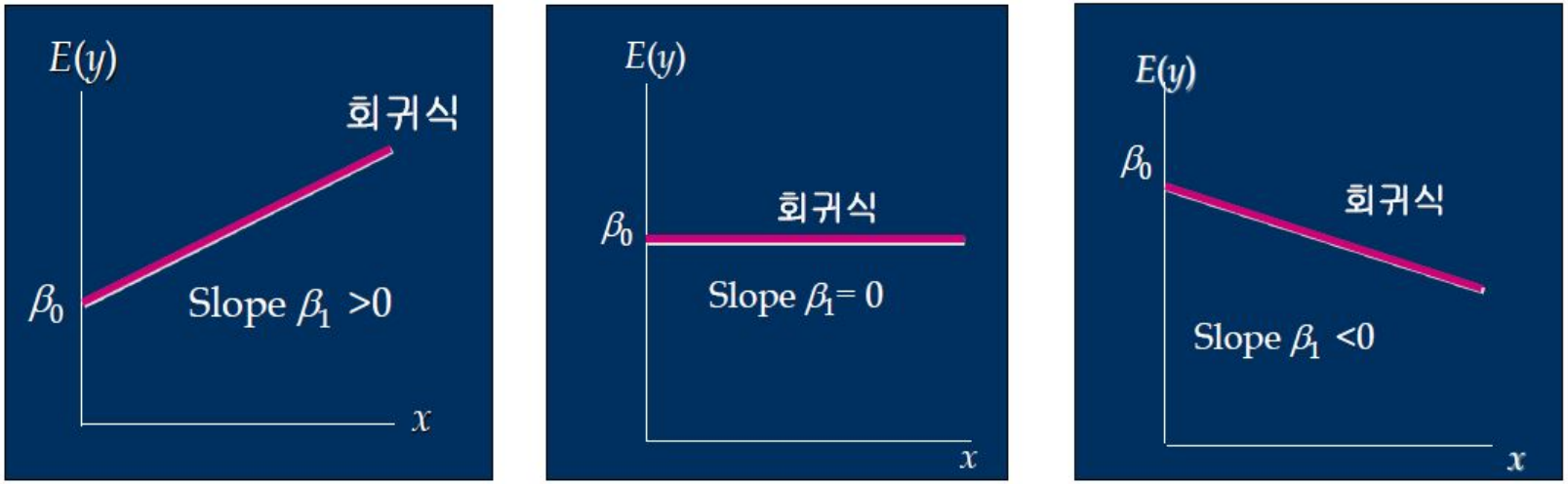
$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

- 모회귀직선: $E(Y_i|x_i) = \beta_0 + \beta_1 x_i$; a function of x_i
  - 설명변수에 따른 반응변수의 평균값을 나타냄

모회귀계수(절편 β_0, 기울기 β_1)

- β_0: 미지의 모수이며 y축의 절편으로 $x = 0$일 때 y의 값이다.
- β_1 : 미지의 모수이며 기울기로 x가 1단위 증가할 때 y의 변동량으로 해석한다.

- 양의 선형관계 (β_1>0)
- 연관성 없음(β_1 =0)
- 음의 선형관계(β_1<0)



오차항 ϵ_i

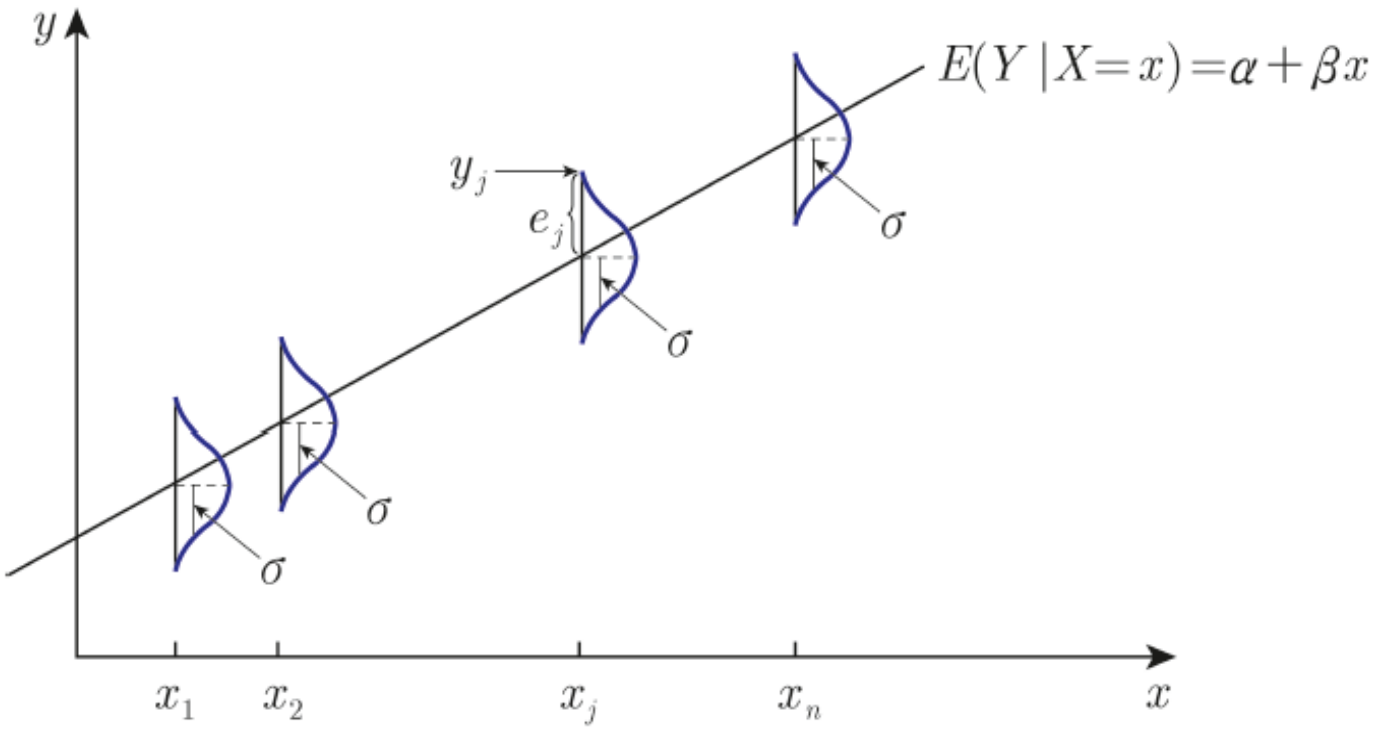


그림 8–4 단순선형회귀모형

- $\epsilon_i \sim_{i.i.d} N(0, \sigma^2)$
- $\epsilon_1, \cdots, \epsilon_n$은 서로 독립적으로 관측된다고 가정

오차항 ϵ_i의 가정사항

- 오차항 ϵ_i의 가정사항 $(i = 1, 2, \cdots, n)$

$$\epsilon_i \sim_{i.i.d} N(0, \sigma^2) \ (\sigma^2 unknown)$$

1) the mean of the error variable ϵ_i is 0 $(E(\epsilon_i) = 0)$
2) the variance of ϵ_i is $\sigma^2 (Var(\epsilon_i) = \sigma^2 > 0)$
3) the errors ϵ_i and ϵ_j are independent (for all $i, j; i \neq j$)
4) ϵ_i is normally distributed

단순선형회귀모형의 가정사항

- $\epsilon_i \sim_{i.i.d} N(0, \sigma^2) \longrightarrow Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ ($\sigma^2 unknown$)

  1) $E(Y_i|x_i) = \alpha + \beta_1 x_i$; a function of x_i: 선형성
     - $E(Y_i|x_i) = E(\beta_0 + \beta_1 x_i + \epsilon_i) = E(\beta_0 + \beta_1 x_i) + E(\epsilon_i) = \beta_0 + \beta_1 x_i$

  2) $Var(Y_i|x_i) = \sigma^2$; the same for all levels of X: 등분산성
     - $Var(Y_i|x_i) = Var(\beta_0 + \beta_1 x_i + \epsilon_i) = Var(\epsilon_i) = \sigma^2$

  3) Y_i and Y_j are independent (for all $i, j; i \neq j$) : 독립성
     - $Cov(Y_i|x_i, Y_j|x_j) = Cov(\beta_0 + \beta_1 x_i + \epsilon_i, \beta_0 + \beta_1 x_j + \epsilon_j) = Cov(\epsilon_i, \epsilon_j) = 0$

  4) Y_i is normally distributed: 정규성

- Y, ϵ: 확률변수($r.v$)
- x_i는 자료에서 주어진 값. non-random, fixed values
- β_0, β_1, σ^2: unknown parameters to be estimated

모회귀계수(절편 β_0, 기울기 β_1)의 추정

Theorem (최소제곱법(method of least squares))

단순선형회귀모형 $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i (i = 1, \cdots, n)$*에서 오차의 제곱합인*

$$SS(\beta_0, \beta_1) = \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_i)\}^2 = \sum_{i=1}^{n} \epsilon_i^2$$

을 최소로 하는 β_0, β_1*의 값을 그 추정값으로 한다.*

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{(\beta_0, \beta_1)} \sum_{i=1}^{n} \epsilon_i^2$$

- 최소제곱법에 의한 β_0, β_1의 추정량 $\hat{\beta}_0$, $\hat{\beta}_1$을 구하려면, 오차의 제곱합 $SS(\beta_0, \beta_1)$을 β_0, β_1에 대하여 각각 편미분하여 얻은 식을 으로 놓은 연립방정식을 풀면된다.

최소제곱법

- 이러한 연립방정식을 정규방정식(normal equation)이라고 하며 (1), (2)와 같이 주어진다.

$$\begin{cases} \frac{\partial SS}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial SS}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0, \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0 & (1) \\ \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 & (2) \end{cases}$$

- 이 때, (1), (2)의식을 다음의 계산 공식을 사용하여 계산하면 편리하다.

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i = \sum_{i=1}^{n} x_i(y_i - \bar{y})$$

최소제곱법

- (1),(2)에 각각 $\sum x_i$, n을 곱하면 다음과 같고

$$\begin{cases} \sum x_i \sum y_i - n\beta_0 \sum x_i - \beta_1 (\sum x_i)^2 = 0 & (3) \\ n\sum x_i y_i - n\beta_0 \sum x_i - n\beta_1 \sum x_i^2 = 0 & (4) \end{cases}$$

(3)식에서 (4)식을 빼주면 다음과 같은 방정식을 얻을 수 있다.

$$\sum x_i \sum y_i - n\sum x_i y_i - \beta_1 (\sum x_i)^2 + n\beta_1 \sum x_i^2 = 0$$

- 마지막으로 β_1로 묶어준 다음에

$$\beta_1 (n\sum x_i^2 - (\sum x_i)^2) = n\sum x_i y_i - \sum x_i \sum y_i$$

최소제곱법

- 정리하면 다음과 같이 $\hat{\beta}_1$을 구할 수 있다.

$$\hat{\beta}_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{\sum x_i y_i - \bar{x}\sum y_i}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$
$$= \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

- $\hat{\beta}_0$은 $\sum_{i=1}^n y_i - n\beta_0 - \beta_1 \sum_{i=1}^n x_i = 0$의 (1)번식을 n으로 나누면

$$\bar{y} - \beta_0 - \beta_1 \bar{x} = 0$$

따라서 $\hat{\beta}_0$은 다음과 같이 구한다.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

최소제곱법

- 이렇게 얻게 되는 추정량 $\hat{\beta}_0$, $\hat{\beta}_1$를 단순선형회귀모형에서 β_0와 β_1의 최소제곱추정량(least squares estimator)이라 부르고, 다음과 같이 정리된다.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

- 이로부터 구해지는 직선은 최소제곱회귀직선(least squares regression line)이라 부르고 다음과 같다.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

(The fitted or estimated regression line)

최소제곱법 예제

Example

예제 8.2 (13p)에서, 사용년수를 x로 하고 중고차 가격을 반응변수 Y로 하여 단순선형회귀모형을 가정하자. 이 때, 표의 자료로부터 β_0와 β_1의 최소제곱추정값을 구하고 산점도에 최소제곱회귀직선을 그려 보아라.

sol) 표의 자료로부터 필요한 통계량의 값들을 계산하면

$$\bar{x} = 39.7/12 = 3.308, \qquad \bar{y} = 27.5/12 = 2.292$$

$$S_{xy} = \sum x_i y_i - n\bar{x}\bar{y} = 69.81 - 12 \cdot 3.308 \cdot 2.292 = -21.169$$

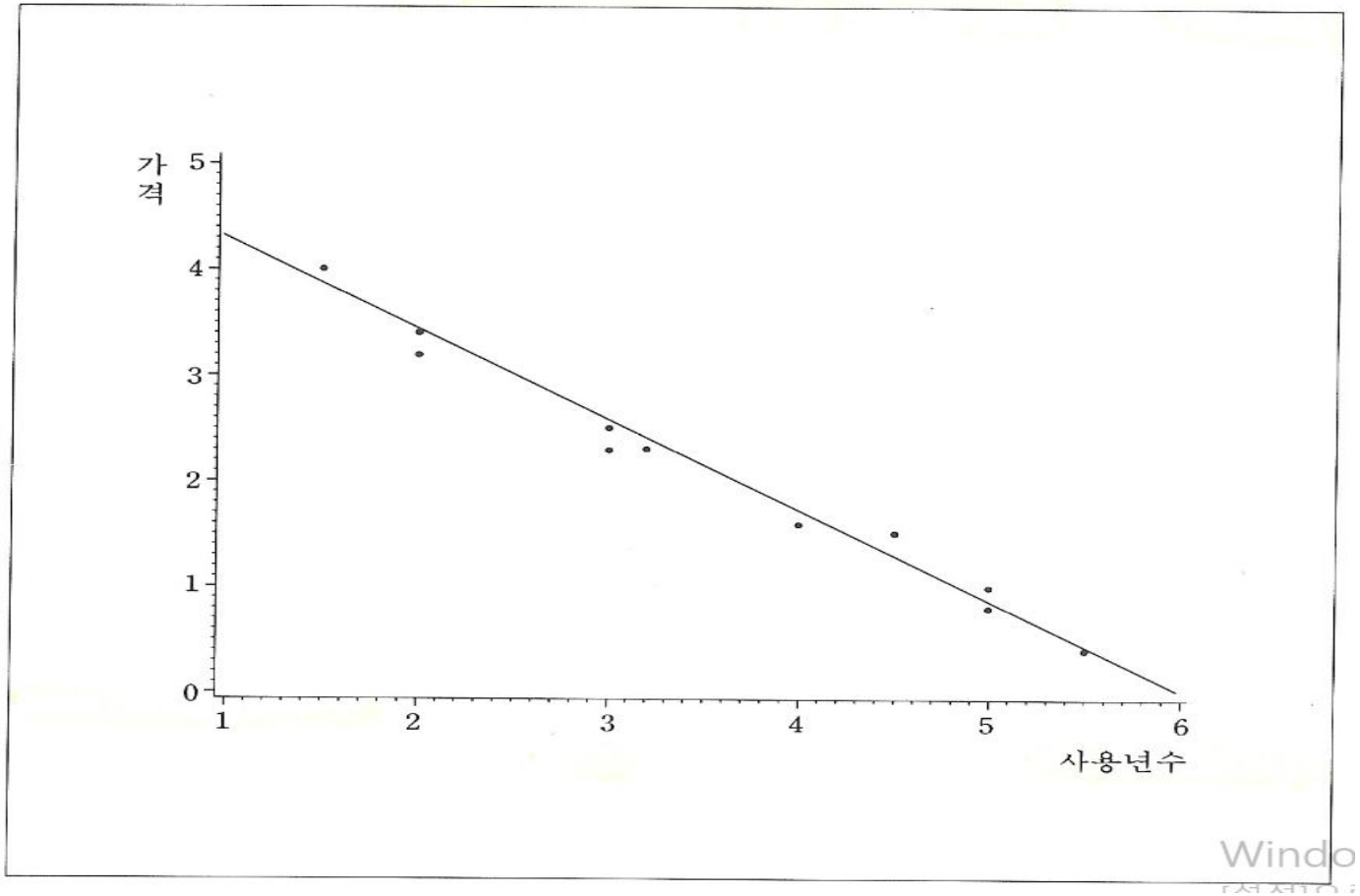
$$S_{xx} = \sum x_i^2 - n\bar{x}^2 = 155.99 - 12 \cdot 3.308^2 = 24.649$$

- 이로부터 모회귀계수의 최소제곱추정값을 계산하면 다음과 같고

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = -0.859, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 5.133$$

- 최소제곱회귀직선은 $\hat{y} = 5.133 - 0.859x$ 이다.

Figure: 표 8.1의 자료에 대한 산점도와 최소제곱회귀직선



- 산점도를 통해 단순선형회귀모형에 의한 적합이 어느 정도 사용년수에 따른 중고차 가격의 변화를 잘 설명해주고 있음을 알 수 있다.
- $x_2 = 1.5$

$E(y_i|x_i)$와 $\hat{y}_i$

- 종속변수 Y를 다음의 두 가지 방식으로 설명할 수 있다.

  $y_i = \beta_0 + \beta_1 x_i + \epsilon_i = E(y_i|x_i) + \epsilon_i$ (ideal model)

  $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i = \hat{y}_i + e_i$ (based on data estimates)

- 예제 8.2에서 $x_2 = 1.5$일 때, 종속변수의 기대값을 구해보자.

$$E(Y_2|x_2) = 5.1 - 0.8 \times 1.5 = 3.9 \neq (y_2 = 4), \quad \epsilon_4 = 0.1$$

  $x_2 = 1.5$일 때, 종속변수의 추측값을 구해보자.

$$\hat{y}_i = 5.133 - 0.859 \times 1.5 = 3.8485 \neq (y_2 = 4), \quad e_2 = 0.1555$$

- $\hat{y}_i$은 반응변수의 평균인 $E(Y_i|x_i) = \beta_0 + \beta_1 x_i$의 추정값이다.

$$\hat{y}_i = \widehat{E(Y_i|x_i)}$$

잔차(residual) e_i

- 잔차(residual): 설명변수 x_i에 대응되는 추측값 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$과 실제 관측값 y_i의 차이.

  - 추정된 회귀식이 반응 변수 관찰치를 설명하지 못하고 있는 부분.

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad (i = 1, 2, \cdots, n)$$

  - e_i is small $\rightarrow$ Our fitted model is good
  - e_i is large $\rightarrow$ We need another model

- 잔차 e_i 들은 오차항 ϵ_i 의 관측값인 것처럼 생각할 수 있고, 오차분산 σ^2의 크기에 따라 크거나 작게 나타난다.
- 따라서, 오차분산 $Var(\epsilon_i) = \sigma^2$에 대한 정보를 제공해주는 잔차들의 제곱합을 이용하여 오차분산을 추정하자.

오차분산 σ^2의 추정

- 오차제곱합(error sum of squares, SSE)

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- 오차제곱합(SSE)을 이것의 자유도 $(n-2)$로 나눈 평균제곱오차(mean squared error, MSE)을 이용하여, σ^2을 추정할 수 있다.

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

즉, $E(MSE) = \sigma^2$이다.

제곱합의 분해(decomposition)

- 최소제곱회귀직선에 의한 단순선형회귀모형의 적합이 관측 결과를 얼마나 잘 설명하고 있는가에 대하여 생각해 보자.

- 반응변수의 관측값 $y_1, \cdots, y_n$이 그 평균인 $\bar{y}$와 떨어진 정도인 $y_i - \bar{y}$를 다음과 같이 두 편차의 합으로 나타낼 수 있다.

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

- $y_i - \bar{y}$: 총편차(total deviation)
- $y_i - \hat{y}_i$: 잔차로서 오차항에 기인하는 편차
- $\hat{y}_i - \bar{y}$: 설명변수 x를 고려한 회귀직선에 의하여 설명되는 편차

제곱합의 분해(decomposition)

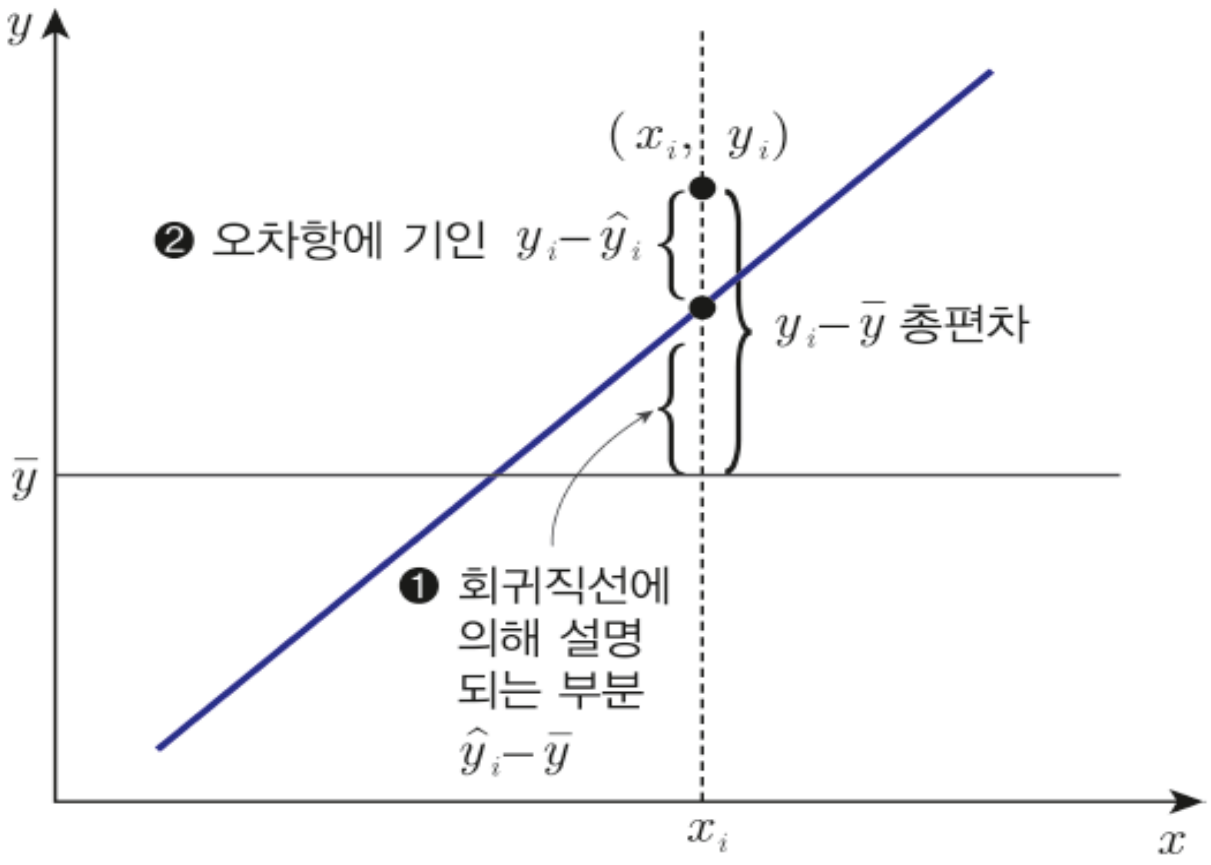


그림 8–6 총편차 $y_i - \bar{y}$의 분해

제곱합의 분해(decomposition)

- 총편차의 분해식의 양변을 제곱하여 합하면 다음이 성립함을 알 수 있다.

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

- 총편차의 제곱합=잔차의 제곱합+회귀로 설명되는 편차의 제곱합
- 자료의 변동= 오차에 기인하는 부분+회귀직선에 기인하는 부분

- SST: 총 제곱합(total sum of squares)
- SSE: 오차 제곱합(error sum of squares)
- SSR: 회귀 제곱합(regression sum of squares)

총제곱합의 분해

- 제곱합: $SST = SSE + SSR$

  $$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$
- 자유도: $n - 1 = (n - 2) + 1$

- d.f = number of information - limiting conditions
- $n - 1$ d.f:

  (n observed values $y_i - \bar{y}$) - (1 constraint $\sum(y_i - \bar{y}) = 0$)
- $n - 2$ d.f:

  (n observed values $y_i - \hat{y}_i$) - (2 d.f are lost $\hat{\beta}_0, \hat{\beta}_1$ to obtain $\hat{y}_i$)
- 1 d.f:

  (We know 2 parameters $\hat{\beta}_0, \hat{\beta}_1$) - (1 d.f is lost, $\sum(\hat{y}_i - \bar{y}) = 0$)

평균제곱(mean square)과 제곱합(sum of squares)

- 평균제곱(mean square): 대응하는 제곱합을 d.f 로 나누어 준 것.
  - 오차 평균제곱: $MSE = \frac{SSE}{n-2}$
  - 회귀 평균제곱: $MSR = \frac{SSR}{1}$

- Useful calculation formula

$$SST: \quad \sum_{i=1}^{n}(y_i - \bar{y})^2 = S_{yy}$$

$$SSR: \quad \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}\{\hat{\beta}_1(x_i - \bar{x})\}^2 = \hat{\beta}_1^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}}$$

$$SSE: \quad \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = S_{yy} - \hat{\beta}_1^2 S_{xx}$$

결정계수(coefficient of determination) R^2

- 결정계수: 총제곱합 SST 중에서 회귀제곱합 SSR이 차지하는 비율

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- 자료 전체의 변동을 나타내는 SST 중에서 회귀에 의하여 설명되는 SSR이 크면, 회귀모형이 관측결과를 잘 설명

- 결정계수 R^2은 표본상관계수의 제곱이다.

$$R^2 = \frac{SSR}{SST} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \left\{ \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \right\}^2 = r^2$$

- $0 \leq R^2 \leq 1$
  - $R^2 \approx 1$: 추정된 회귀식이 총 변동량의 많은 부분을 설명
  - $R^2 \approx 0$: 추정된 회귀식이 총 변동량을 적절하게 설명하지 못함

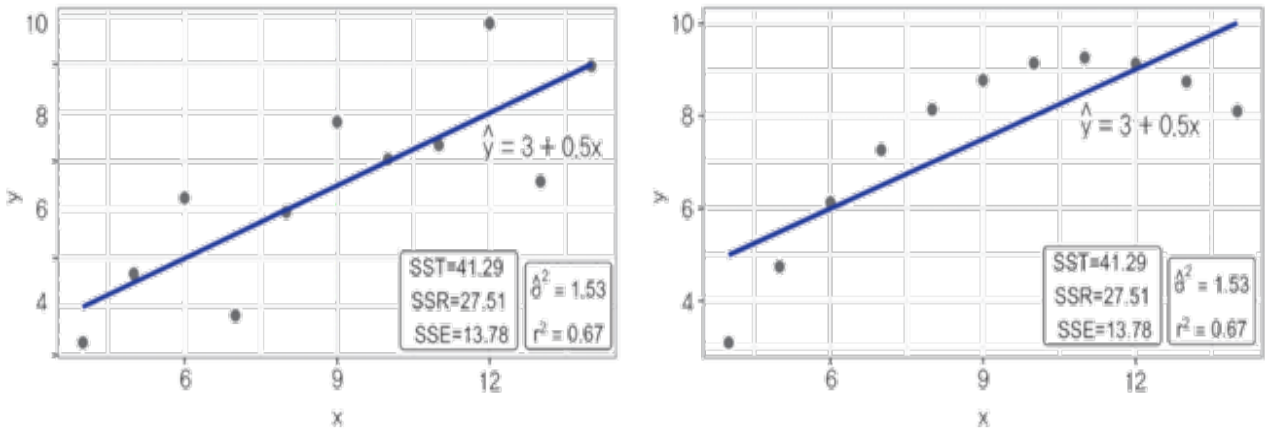
회귀분석시 주의점

- 다음의 표의 자료를 이용하여, 회귀분석을 한 결과를 비교해 보자.

표 8-3 모든 통계량이 거의 일치하는 두 조의 자료[10]

자료 A	x	10	8	13	9	11	14	6	4	12	7	5
	y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68
자료 B	x	10	8	13	9	11	14	6	4	12	7	5
	y	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74

- (a), (b)는 SST, SSE, SSR, R^2 등의 통계량 값이 거의 일치한다.



(a) 자료 A의 최소제곱회귀직선과 통계량

(b) 자료 B의 최소제곱회귀직선과 통계량

그림 8-7 통계량은 같지만 관계가 다른 [표 8-3]의 자료

- 그러나 두 자료는 전혀 다른 관계를 나타내고 있다.(산점도)
- 단순히 통계량만에 의한 해석보다는 산점도 or 잔차의 검토가 필요하다.

단순회귀분석에서의 추론

- 다음의 단순선형회귀모형의 가정이 타당할 때

$$\begin{cases} Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \\ \epsilon_i \ \sim_{i.i.d} \ N(0, \sigma^2) \ \ (i = 1, \cdots, n) \end{cases}$$

- 회귀직선이 의미가 있는지 검정하기 위해서 다음의 가설을 고려해 보자.

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

- 만일 $\beta_1 = 0$이라면 설명변수와 반응변수사이에 선형적인 관계가 없다는 것이 된다.
- 회귀직선이 유의하여 선형적인 연관성이 있다면, $\beta_1 \neq 0$일 것이다.

회귀직선의 유의성 F 검정

- 귀무가설 H_0가 사실일 때 다음의 F 검정통계량은 자유도가 1, $n-2$인 F분포를 따른다고 알려져 있다.

$$F = \frac{SSR/1}{SSE/(n-2)} \sim F(1, n-2)$$

pf) by Cochran's Theorem

$\frac{SSR}{\sigma^2}$, $\frac{SSE}{\sigma^2}$ are independently distributed chi-square random variables with 1, $(n-2)$ degrees of freedom

$$\frac{\frac{SSR}{\sigma^2}/1}{\frac{SSE}{\sigma^2}/(n-2)} = \frac{SSR/1}{SSE/(n-2)} \sim F(1, n-2)$$

회귀직선의 유의성 F 검정

- 제곱합: $SST = SSE + SSR$
- 회귀직선이 유의하다면 자료의 산포를 나타내는 SST 중에서 SSR이 차지하는 비중이 커지고
  SSE가 차지하는 비중은 작아질 것이다.

  즉, SSR/SSE이 커질수록 증거가 강해진다.

- 이때, F검정의 대립가설은 양측방향($H_1 : \beta_1 \neq 0$)이나, β_1의 값이 음수나 양수로 큰 값 모두 F값이 큰 값으로 나오므로 오른쪽단측 검정을 한다.

회귀직선의 유의성 F 검정

- 가설: $H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$
- 검정통계량(under H_0):

$$F = \frac{SSR/1}{SSE/(n-2)} \sim F(1, n-2)$$

이러한 F 검정통계량의 관측값이 f이면 다음이 성립한다.

유의확률	유의수준 α의 기각역
$P = P\{F \geq f\}, \quad F \sim F(1, n-2)$	$f \geq F_\alpha(1, n-2)$

- 분산분석표(analysis of variance table)

요인	제곱합	자유도	평균제곱	F값	유의확률
회귀	SSR	1	MSR=SSR/1	f=MSR/MSE	$P\{F \geq f\}$
잔차	SSE	$n-2$	MSE=SSE/$(n-2)$		
계	SST	$n-1$			

단순선형회귀분석 예제

Example

예제 8.2 (13p)에서 단순선형회귀모형을 적용할 때, 오차분산 σ^2의 추정값, 결정계수를 구하고, 또한 회귀직선의 유의성 F 검정을 하여라.

sol) 표 8.1의 자료로부터 $S_{xy} = -21.169$, $S_{xx} = 24.649$, $S_{yy} = 18.469$
따라서, 잔차제곱합 SSE, 평균제곱오차 MSE는 다음과 같다.

$$SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 18.469 - (-21.169)^2/24.649 = 0.289$$

$$MSE = \hat{\sigma}^2 = 0.289/10 = 0.0289$$

- 결정계수의 값은

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{(-21.169)^2}{24.649 \cdot 18.469} = 0.9844$$

회귀직선이 전체 자료의 산포 중 98%를 설명해 준다는 뜻이고, 단순선형회귀모형에 의한 적합이 관측 결과를 잘 설명해주고 있다.

- 검정하고자 하는 가설은 다음과 같다.

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

각 제곱합을 계산하면

$$SST = 18.469, \;\; SSR = S_{xy}^2/S_{xx} = (-21.169)^2/24.649 = 18.180$$

- 이로부터 평균제곱과 F 검정통계량의 관측값을 구하면

$$MSR = SSR/1 = 18.180, \;\; MSE = SSE/(12-2) = 0.0289$$

$$f = MSR/MSE = 18.180/0.0289 = 629.76$$

- F분포표에서 $F \sim F(1,10)$일 때 $P(F \geq 629.76) < 0.001$ 이므로 회귀직선의 유의성에 대한 증거가 뚜렷하다고 할 수 있으며, 이러한 검정 과정을 분산분석표로 나타내면 다음과 같다.

요인	제곱합	자유도	평균제곱	F값	유의확률
회귀	$SSR = 18.180$	1	MSR=18.180	f=629.76	< 0.001
잔차	SSE=0.289	10	MSE=0.0289		
계	$SST = 18.469$	11			

8장. 상관분석과 회귀분석

박명현

서울대학교 학부대학

기울기 β_1에 대한 추론

- 모회귀계수 β_1의 최소제곱추정량 $\hat{\beta}_1$를 다음과 같이 표현할 수 있다.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- $\hat{\beta}_1$의 기대값과 분산

$$\begin{aligned} E(\hat{\beta}_1) &= \frac{1}{S_{xx}} \sum (x_i - \bar{x}) \cdot E(Y_i | x_i) = \frac{1}{S_{xx}} \sum (x_i - \bar{x})(\beta_0 + \beta_1 x_i) \\ &= \frac{\beta_0}{S_{xx}} \sum (x_i - \bar{x}) + \frac{\beta_1}{S_{xx}} \sum (x_i - \bar{x}) x_i = \beta_1 \end{aligned}$$

$$Var(\hat{\beta}_1) = \frac{1}{S_{xx}^2} \sum (x_i - \bar{x})^2 \cdot Var(Y_i | x_i) = \frac{\sigma^2}{S_{xx}^2} \sum (x_i - \bar{x})^2 = \frac{\sigma^2}{S_{xx}}$$

기울기 β_1에 대한 추론

- $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} \cdot y_i$ 이므로
  독립적으로 정규분포를 따르는 y_i $(i = 1, \cdots, n)$들의 선형결합인 $\hat{\beta}_1$ 또한 다음의 정규분포를 따르게 된다. (σ^2 unknown)

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}}) \tag{1}$$

- Sampling distribution of $\hat{\beta}_1$

$$\frac{\hat{\beta}_1 - \beta}{s(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t(n-2)$$

- $\hat{\beta}_1$의 추정된 분산과 표준오차:

$$s^2(\hat{\beta}_1) = \widehat{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{S_{xx}}, \qquad s(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

β_1에 대한 추론

- β_1에 대한 $100(1-\alpha)\%$ 신뢰구간 (σ^2 unknown)

$$\hat{\beta}_1 \pm t_{\alpha/2}(n-2)\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

- β_1에 대한 유의성 t 검정

  - 귀무가설: $H_0 : \beta_1 = b$

  - Under H_0, 검정통계량:

$$T = \frac{\hat{\beta}_1 - b}{s(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - b}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \sim t(n-2)$$

여기서 검정통계량의 관측값을 t라고 하면 다음이 성립한다.

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \beta_1 > b$	$P = P(T \geq t)$	$t \geq t_\alpha(n-2)$
(b) $H_1 : \beta_1 < b$	$P = P(T \leq t)$	$t \leq -t_\alpha(n-2)$
(c) $H_1 : \beta_1 \neq b$	$P = P(\lvert T\rvert \geq \lvert t\rvert)$	$\lvert t\rvert \geq t_{\alpha/2}(n-2)$

β_1에 관한 추론 예제

Example

예 8.2에서 단순선형회귀모형을 적용할 때, 다음의 추론을 하여라.
(a) 회귀직선의 유의성을 t검정으로 유의수준 1%에서 검정하고, 분산분석표에 의한 F검정과의 관계를 설명하여라.

sol) 검정하고자 하는 가설은 다음과 같다.

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

- 예제 8.2의 풀이 과정에서

$$n = 12, \quad \hat{\beta}_1 = -0.859, \quad S_{xx} = 24.649, \quad \hat{\sigma}^2 = MSE = 0.0289$$

이므로 t 검정통계량의 관측값은

$$t = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} = \frac{-0.859}{\sqrt{0.0289/24.649}} = 25.087$$

- t분포표에서 $t_{0.005}(10) = 3.169$이므로, 다음과 같이 기각역에 속한다.

$$|25.087| = |t| \geq t_{0.005}(10) = 3.169$$

즉, 유의수준 1%에서 회귀직선이 유의하다고 할 수 있다.

- 이러한 회귀직선의 유의성 t검정은 분산분석표를 이용한 F검정으로도 할 수 있으며,

$$F = \frac{SSR/1}{SSE/(n-2)} = \frac{\hat{\beta}_1^2 S_{xx}}{MSE} = \left(\frac{\hat{\beta}_1 - 0}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \right)^2 = T^2$$

단순선형회귀분석에서는 F검정과 t검정이 다음의 관계에 있음을 알 수 있다.

$$t_{n-2}^2 = F(1, n-2)$$

유의성 F검정과 T검정비교

- F검정은 회귀모형의 모든 기울기 모수의 유의성에 대한 총제적 검정(overall test)을 수행한다.

  - 일반적으로 절편 β_0는 F검정에서 제외한다.

- 단순회귀모형에서는 β_0를 제외한 회귀계수가 β_1 기울기 하나이므로 $H_0 : \beta_1 = 0$를 검정하는 것이 되어, 개별 검정인 t검정과 총체적 검정인 F 검정의 귀무가설과 대립가설이 동일한 형태가 되며 검정의 결과도 완벽하게 일치한다. 그 관계는 정확히 $t^2 = F$이다.

- 다중회귀에서는 기울기가 여러 개 이므로 t검정과 F 검정이 같지 않다.

Example

(b) 자동차의 가격이 사용년수에 따라 연평균 감소액이 80만원을 초과하는지 유의수준 5%에서 검정하여라.

sol) 자료의 측정 단위가 백만원이므로, 검정하고자 하는 가설은 다음과 같다.

$$H_0 : \beta_1 = -0.8 \quad H_1 : \beta_1 < -0.8$$

- t 검정통계량의 관측값은

$$t = \frac{\hat{\beta}_1 - (-0.8)}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} = \frac{-0.859 + 0.8}{\sqrt{0.0289/24.649}} = -1.723$$

- t분포표에서 $t_{0.05}(10) = 1.812$이므로, 다음의 기각역에 속하지 않는다.

$$t \leq -t_{0.05}(10)$$

- 따라서, 유의수준 5%에서 H_0을 기각할 수 없고, 중고가격의 연평균 감소액이 80만원을 초과한다는 증거는 없다고 할 수 있다.

Example

(c) 모회귀계수 β_1의 95% 신뢰구간을 구하고 그 의미를 해석하여라.

sol) t분포표에서 $t_{0.025}(10) = 2.228$이므로, β_1의 95%신뢰구간은

$$\hat{\beta}_1 \pm t_{0.025}(10)\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = -0.859 \pm (2.228)\sqrt{0.0289/24.649}$$
$$= -0.859 \pm 0.076$$

- 즉, 95% 신뢰구간은

$$-0.935 \leq \beta_1 \leq -0.783$$

이고, 중고차 가격은 연평균 783,000원에서 935,000원 사이로 감소한다고 신뢰수준 95%에서 결론지을 수 있다.

절편 β_0에 관한 추론

- 모회귀계수 β_0의 최소제곱추정량 $\hat{\beta}_0$은 다음과 같이 표현 가능

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \sum_{i=1}^{n} \frac{y_i}{n} - \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \bar{x}$$

- $\hat{\beta}_0$의 기대값과 분산:

$$\begin{aligned} E(\hat{\beta}_0) &= E(\bar{Y} - \hat{\beta}_1 \bar{x}) = E(\frac{1}{n} \sum Y_i) - E(\hat{\beta}_1 \bar{x}) \\ &= \frac{1}{n} \sum E(Y_i | x_i) - E(\hat{\beta}_1)\bar{x} = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0 \end{aligned}$$

$$\begin{aligned} Var(\hat{\beta}_0) &= \sum \left\{ \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} \right\}^2 Var(Y_i | x_i) \\ &= \sigma^2 \left\{ \sum \frac{1}{n^2} - 2 \sum \frac{(x_i - \bar{x})}{nS_{xx}} \bar{x} + \sum \frac{(x_i - \bar{x})^2 \bar{x}^2}{S_{xx}^2} \right\} \\ &= \sigma^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right\} \end{aligned}$$

절편 β_0에 관한 추론

- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \sum \frac{y_i}{n} - \frac{\sum(x_i-\bar{x})\cdot y_i}{\sum(x_i-\bar{x})^2}\bar{x} = \sum \left\{ \frac{1}{n} - \frac{(x_i-\bar{x})\bar{x}}{S_{xx}} \right\} y_i$ 이므로 $\hat{\beta}_0$ 또한 다음의 정규분포를 따르게 된다.

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right))$$

- σ^2을 $\hat{\sigma}^2$으로 대체한 표본분포는 다음과 같다.

$$\frac{\hat{\beta}_0 - \beta_0}{s(\hat{\beta}_0)} = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t(n-2)$$

β_0에 대한 추론

- β_0에 대한 $100(1-\alpha)\%$ 신뢰구간 (σ^2 unknown)

$$\hat{\beta}_0 \pm t_{\alpha/2}(n-2)\ \hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}$$

- β_0에 대한 유의성 t 검정

  - 귀무가설: $H_0 : \beta_0 = a$

  - Under H_0, 검정통계량:

$$T = \frac{\hat{\beta}_0 - a}{s(\hat{\beta}_0)} = \frac{\hat{\beta}_0 - a}{\hat{\sigma}\sqrt{n^{-1}+\bar{x}^2/S_{xx}}}$$

이고, 이의 관측값을 t라고 하면 다음이 성립한다.

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \beta_0 > a$	$P = P(T \geq t)$	$t \geq t_\alpha(n-2)$
(b) $H_1 : \beta_0 < a$	$P = P(T \leq t)$	$t \leq -t_\alpha(n-2)$
(c) $H_1 : \beta_0 \neq a$	$P = P(\lvert T\rvert \geq \lvert t\rvert)$	$\lvert t\rvert \geq t_{\alpha/2}(n-2)$

절편 β_0에 관한 추론의 주의점

- 실제 응용에서는 특수한 상황을 제외하고는 절편 β_0에 대하여 가설검정이나 신뢰구간 추정을 하지 않는다.

  예) 키를 설명변수, 몸무게를 종속변수로 한 회귀식

  - 키가 0일 때의 몸무게 추정값인 β_0의 추정은 실질적인 의미가 없음
  - 관측된 설명변수의 범위를 많이 벗어나게 되면 모형의 타당성이 담보되지 못함

- 그런데 특별한 경우가 아니면 β_0이 통계적으로 비유의적이라 하더라도 회귀모형에 포함시켜 분석을한다.

  - β_0를 회귀모형에서 제외하면 추정된 회귀식이 자료를 잘 설명하지 못함

β_0에 관한 추론 예제

Example

예 8.2에서 단순선형회귀모형을 적용할 때, 모회귀직선의 절편 β_0에 관한 95% 신뢰구간을 구하여라.

sol) 예제 8.2-3의 풀이 과정에서

$$n = 12, \quad \bar{x} = 3.308, \quad S_{xx} = 24.649, \quad \hat{\beta}_0 = 5.133, \quad \hat{\sigma}^2 = 0.0289$$

이고 $t_{0.025}(10) = 2.228$이므로, β_0의 95% 신뢰구간은

$$\hat{\beta}_0 \pm t_{0.025}(10)\ \hat{\sigma}\sqrt{n^{-1} + \bar{x}^2/S_{xx}} = 5.133 \pm 2.526$$

회귀모형을 이용한 예측(prediction)

- 추정된 회귀식을 이용하여 특정한 x_0 값에 대응하는 반응변수 Y의 평균값을 추정 또는 예측해야 하는 경우가 있다.

  - x_0를 우리가 관심있는 설명변수의 특정한 값이라고 하자.

  - 회귀모형에서 $x = x_0$일 때의 예측 가능한 Y의 값은 $\beta_0 + \beta_1 x_0$이고,

    이 값은 $x = x_0$일 때 반응변수 Y값들의 평균값이다.

  - $E(Y_0|x_0)$를 $x = x_0$일 때의 평균반응값이라 하자.

- $E(Y_0|x_0) = \beta_0 + \beta_1 x_0$의 점추정:

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

평균반응 $E(Y_0|x = x_0) = \beta_0 + \beta_1 x_0$에 관한 추론

- $E(Y_0|x_0)$의 구간추정을 위해 $\hat{Y}_0$의 분포에 대해서 알아보자.

$$\begin{aligned}
\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 &= (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x_0 \\
&= \sum_{i=1}^{n} \frac{y_i}{n} - \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \bar{x} + \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} x_0 \\
&= \sum_{i=1}^{n} \left[\frac{1}{n} + \frac{(x_i - \bar{x})}{S_{xx}} (x_0 - \bar{x}) \right] y_i
\end{aligned}$$

- $\hat{Y}_0$은 독립적으로 정규분포를 따르는 y_i들의 선형결합이므로, $\hat{Y}_0$도 정규분포를 따른다고 할 수 있다.

평균반응 $E(Y_0|x = x_0) = \beta_0 + \beta_1 x_0$에 관한 추론

- $\hat{Y}_0$의 기대값과 분산은 다음과 같다.

$$E(\hat{Y}_0) = E(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \beta_0 + \beta_1 x_0$$

$$\begin{aligned} Var(\hat{Y}_0) &= Var(\hat{\beta}_0 + \hat{\beta}_1 x_0) = Var(\bar{y} + \hat{\beta}_1(x_0 - \bar{x})) \\ &= Var(\bar{y}) + Var((x_0 - \bar{x})\hat{\beta}_1) - 2Cov(\bar{y}, (x_0 - \bar{x})\hat{\beta}_1) \\ &= \frac{\sigma^2}{n} + \frac{(x_0 - \bar{x})^2 \sigma^2}{S_{xx}} = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \end{aligned}$$

- $\hat{\beta}_0 + \hat{\beta}_1 x_0$를 표준화하고 분모의 σ^2을 그 추정량인 $\hat{\sigma}^2$로 바꾸어 주면 자유도 $(n-2)$인 t분포를 따른다고 알려져 있다.

$$\frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x_0)}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}} \sim t(n-2)$$

평균반응 $E(Y_0|x = x_0) = \beta_0 + \beta_1 x_0$에 관한 추론

- $\beta_0 + \beta_1 x_0$에 관한 $100(1-\alpha)\%$ 신뢰구간:

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\alpha/2}(n-2)\ \hat{\sigma}\sqrt{n^{-1} + (x_0 - \bar{x})^2/S_{xx}}$$

- $\beta_0 + \beta_1 x_0$에 대한 유의성 검정:

  귀무가설 $H_0 : \beta_0 + \beta_1 x_0 = \mu_0$에 관한 검정통계량은

$$T = \frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - \mu_0}{\hat{\sigma}\sqrt{n^{-1} + (x_0 - \bar{x})^2/S_{xx}}}$$

  이고, 이의 관측값을 t라고 하면 다음이 성립한다.

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \beta_0 + \beta_1 x_0 > \mu_0$	$P = P(T \geq t)$	$t \geq t_\alpha(n-2)$
(b) $H_1 : \beta_0 + \beta_1 x_0 < \mu_0$	$P = P(T \leq t)$	$t \leq -t_\alpha(n-2)$
(c) $H_1 : \beta_0 + \beta_1 x_0 \neq \mu_0$	$P = P(\lvert T\rvert \geq \lvert t\rvert)$	$\lvert t\rvert \geq t_{\alpha/2}(n-2)$

평균반응에 관한 추론 예제[예 8.2]

Example

(a) 사용년수가 $x = 2.5$년인 중고차의 평균 가격이 2,800,000원보다 높은지 유의수준 1%에서 검정하여라.

sol) 검정하고자 하는 가설은 다음과 같다.

$$H_0 : \beta_0 + \beta_1 2.5 = 2.8 \quad H_1 : \beta_0 + \beta_1 2.5 > 2.8$$

- t 검정통계량의 관측값은

$$t = \frac{(\hat{\beta}_0 + \hat{\beta}_1 2.5) - 2.8}{\hat{\sigma}\sqrt{n^{-1} + (2.5 - \bar{x})^2 / S_{xx}}} = 3.293$$

이고, 부록의 t분포표에서 $t_{0.01}(10) = 2.764$이므로, 검정통계량의 값이 기각역에 속하는 것을 알 수 있다.

$$3.293 = t \geq t_{0.01}(10)$$

따라서 $\alpha = 0.01$에서 귀무가설을 기각할 수 있고, 2.5년 지난 중고차의 평균 가격이 2,800,000원보다 높다는 뚜렷한 증거가 있다.

Example

(b) 사용년수가 $x = 1,\ 3,\ 5$일 때에 중고차의 평균 가격의 95% 신뢰구간을 각각 구하고 이를 그래프로 그려 보아라.

- 앞에서의 요약 통계량으로부터, $x = 1$ 일 때 평균 가격의 95% 신뢰구간은 다음과 같이 주어진다.

$$(\hat{\beta}_0 + \hat{\beta}_1 1) \pm t_{0.025}(10)\ \hat{\sigma}\sqrt{n^{-1} + (1-\bar{x})^2/S_{xx}}$$

$$= (5.133 - 0.859) \pm (2.228)\sqrt{0.0289}\sqrt{1/12 + (1-3.308)^2/24.649}$$

$$= 4.274 \pm 0.207$$

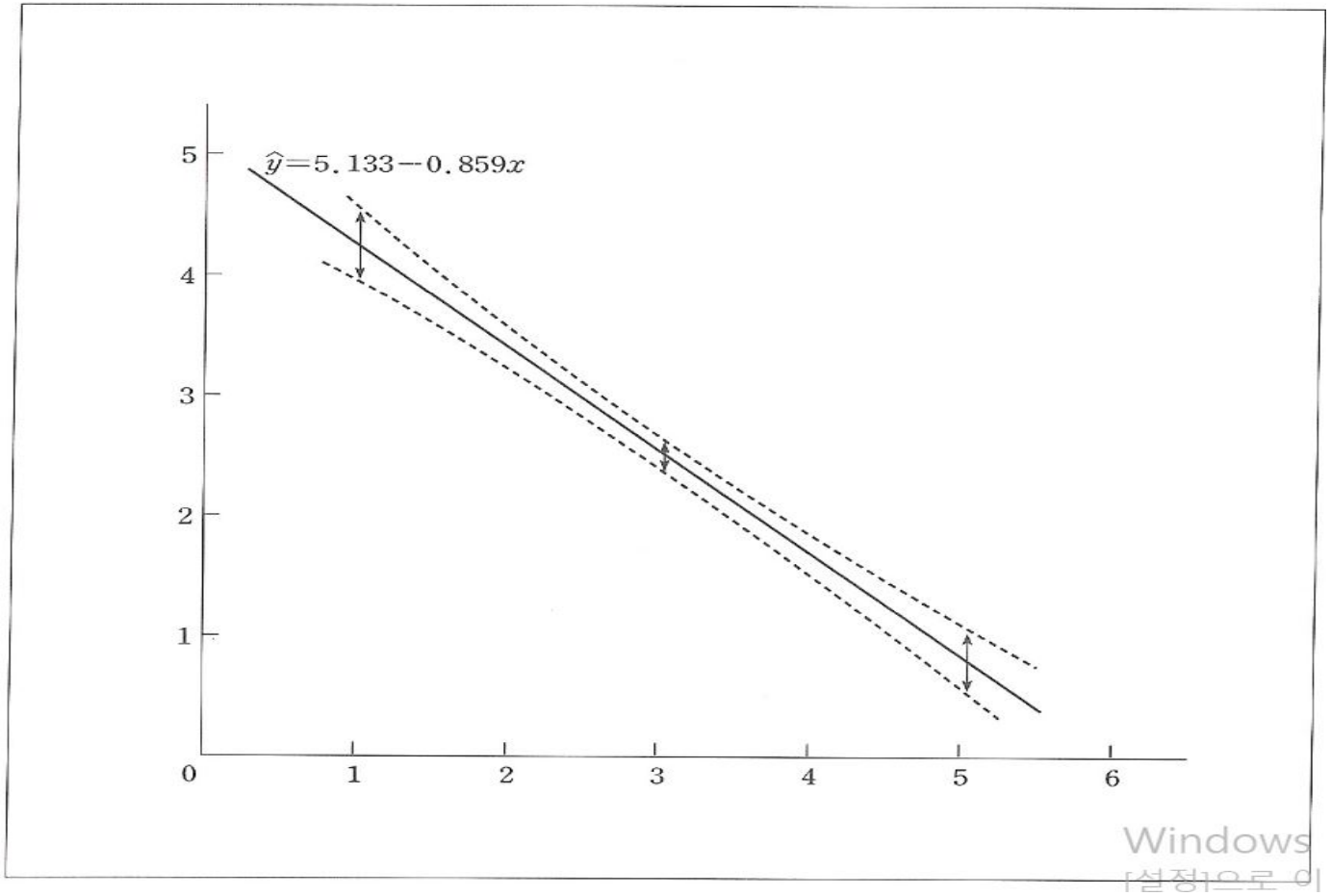
- 같은 방법으로 $x = 3,\ 5$일 때 평균 가격 $\beta_0 + \beta_1 x$에 대한 95% 신뢰구간은 각각 다음과 같이 구할 수 있다.

  $x = 3$ 일 때 2.556 ± 0.112

  $x = 5$ 일 때 0.838 ± 0.169

평균반응의 95% 신뢰구간

- 앞 예제에서 평균반응의 95% 신뢰구간과 상한, 하한을 그려보자.



평균반응의 95% 신뢰구간

- 신뢰구간에 $(x_0 - \bar{x})^2$의 항이 들어 있으므로 신뢰구간의 너비는 $x_0 = \bar{x}$일 때 가장 짧고, $\bar{x}$에서 멀어질수록 폭이 커짐을 알 수 있다.
- 따라서 설명변수 x의 값이 원자료의 범위를 많이 이탈하는 경우 주의를 해야 한다.
- $\bar{x}$에서 너무 멀리 그리고 추정된 회귀식에 이용된 x의 범위를 많이 이탈하는 경우, 그 예측력에는 신뢰성이 떨어진다는 것을 의미한다.

모형의 진단(diagnostic) 및 처방(remedy)

- 최소제곱법으로 추정된 회귀식을 어떻게 평가할 수 있을까?
- 단순회귀추정식 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$이 설명변수와 반응변수 사이의 연관성을 얼마나 잘 설명하고 있는지, 회귀모형 적용의 타당성을 검토하여 보자.

- 우선적으로 산점도(scatter plot)를 그려서 선형회귀모형이 적합한지 판단할 수 있다.

  $\Longrightarrow$통계적 유의성을 점검할 수 있는 수치를 제공하진 못하지만, 직관적인 판단이 가능하게 해준다.

모형의 진단(diagnostic) 및 처방(remedy)

- 회귀진단 항목

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (i = 1, 2, \cdots, n) \text{ 에서}$$

(a) $E(\epsilon_i) = 0$, i.e, $E(Y_i|x_i) = \beta_0 + \beta_1 x_i$ (선형성)

(b) $Var(\epsilon_i) = \sigma^2 > 0$ (등분산성)

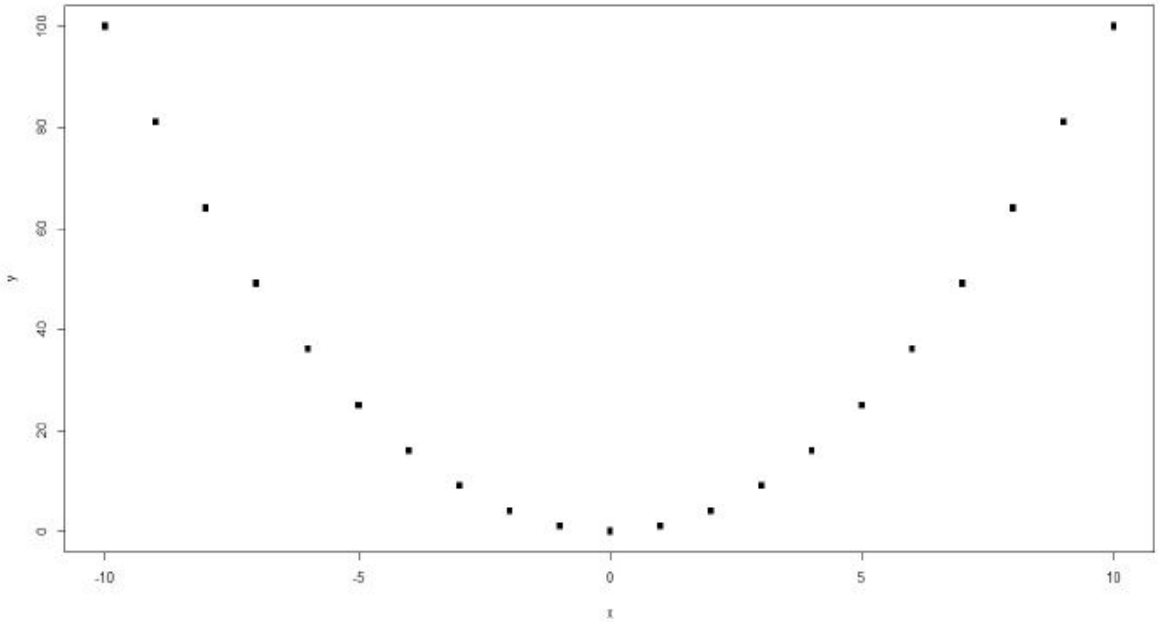
(c) $\epsilon_1, \cdots, \epsilon_n$은 서로 독립 (독립성)

(d) $\epsilon_i \sim N(0, \sigma^2)$ (정규성)

(e) 이상치(outlier)나

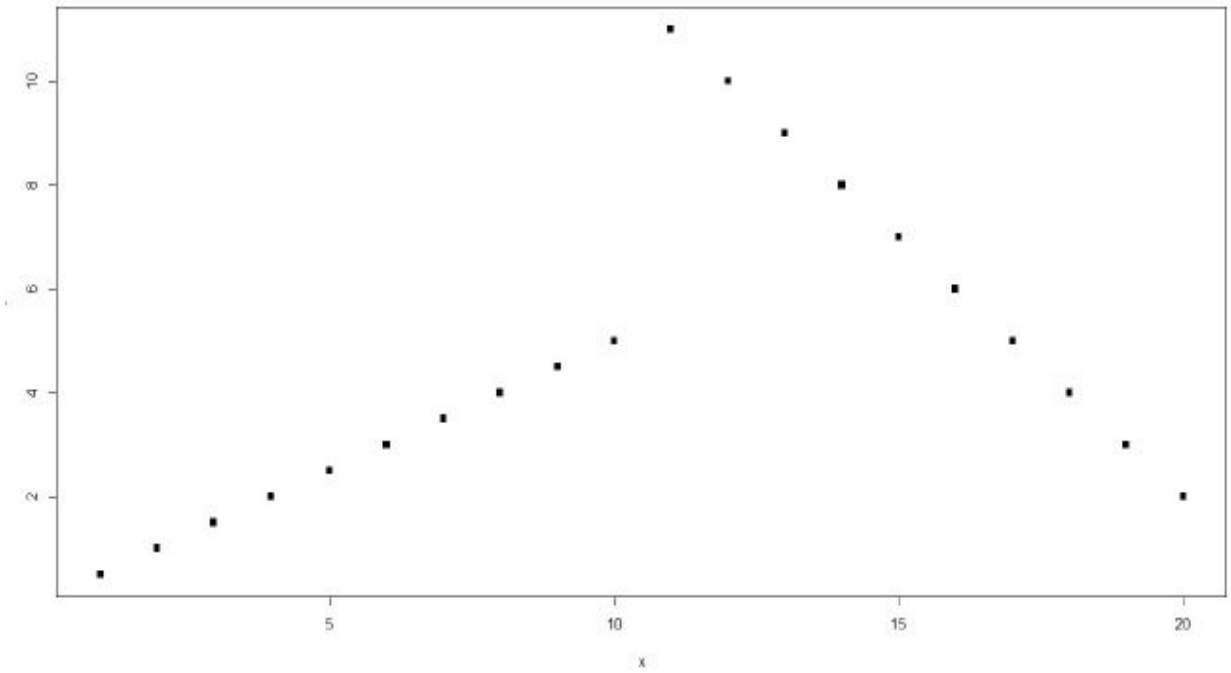
(f) 영향력이 큰 관찰치(influential observation)의 존재

산점도를 이용한 선형성 진단



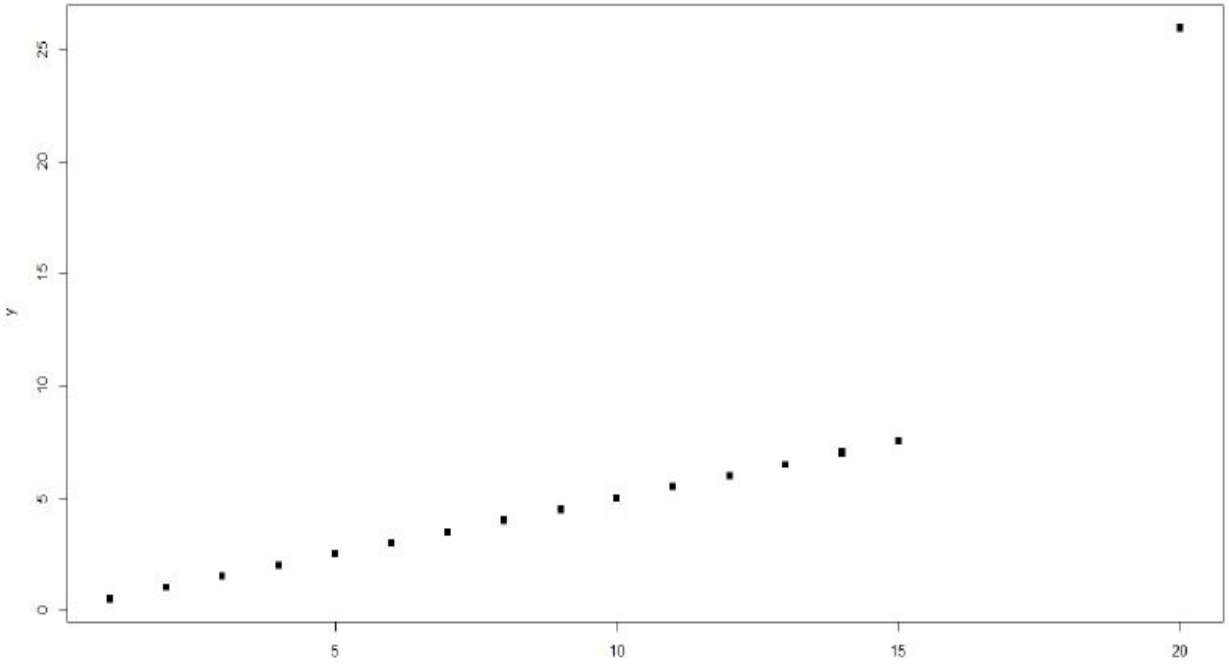
- 처방: 두 변수가 비선형 관계를 보이면 변수변환 등의 기법을 사용하거나 2차 이상의 다항식 또는 로그 함수 등의 비선형 모형(nonlinear model)을 적용

산점도를 이용한 선형성 진단



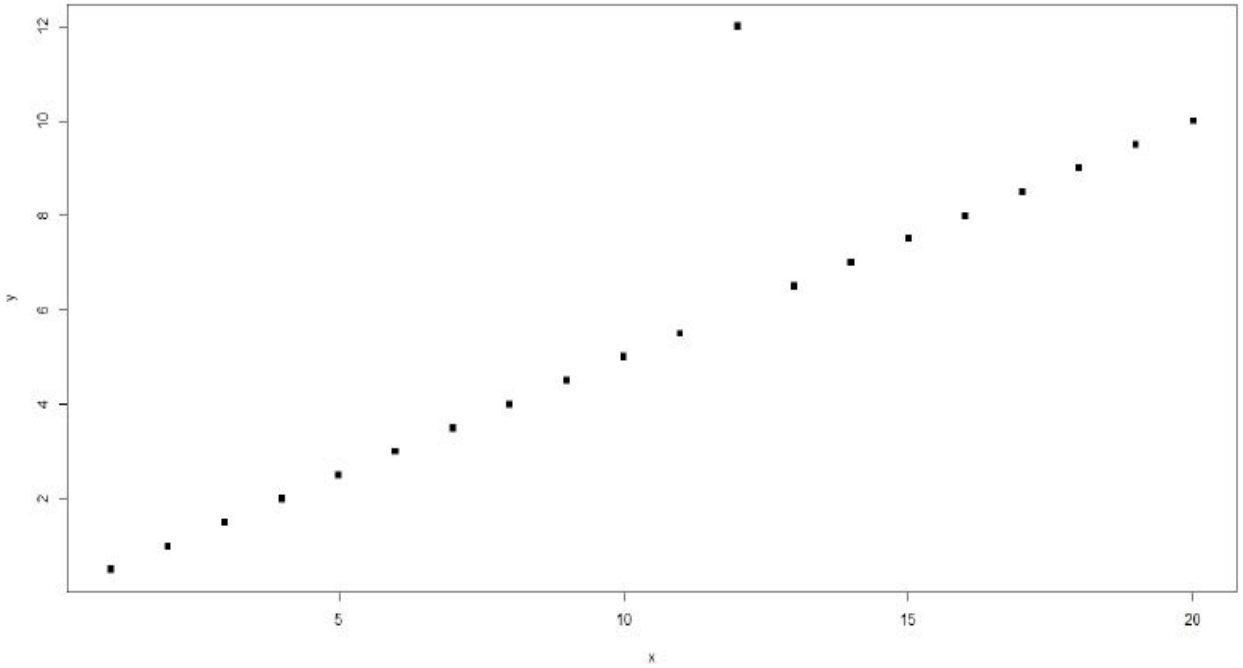
- 처방: 두 개의 그룹으로 나누어 각각 선형 모형을 시도

산점도를 이용한 이상치 진단



- 처방: 이상치가 단순 오기에 의한 실수인지 점검하고, 단순 오기가 아니라면 이상치를 그대로 포함할지, 제거한 후 선형모형을 적용할지 결정

산점도를 이용한 영향관찰치 진단



- 처방: 입력 실수인지 확인을 먼저 하고, 영향력을 판단하기 위하여 관측값을 제거한 후 모형을 추정하고 추정치 값들, R^2 등이 얼마나 차이 나는지 비교.

잔차분석(analysis of residual)

- 모형의 검토에서 흔히 다음의 잔차를 이용한다.

$$e_i = y_i - \hat{y}_i$$

where $y_i = i$번째 반응변수 관찰값, $\hat{y}_i = i$번째 반응변수 추정값

- 오차에 대한 가정사항: $\epsilon_i \sim_{i.i.d} N(0, \sigma^2)$
- 회귀모형의 가정을 표준화된 오차항을 이용하여 나타내보자.

$$\frac{\epsilon_i - 0}{sd(\epsilon_i)} = \frac{\epsilon_i (= Y_i - (\beta_0 + \beta_1 x_i))}{\sigma} \sim_{i.i.d} N(0,1), \quad i = 1, \cdots, n$$

- (반)스튜던트화 잔차(semi-studentized residual): 잔차 e_i를 그 표준편차의 추정값인 $\widehat{sd}(e_i)$으로 나눈 값

$$\hat{e}_{st,i} = \frac{e_i}{\widehat{sd}(e_i)} = \frac{e_i (= y_i - \hat{y}_i)}{\hat{\sigma}}, \quad i = 1, \cdots, n$$

잔차분석(analysis of residual)

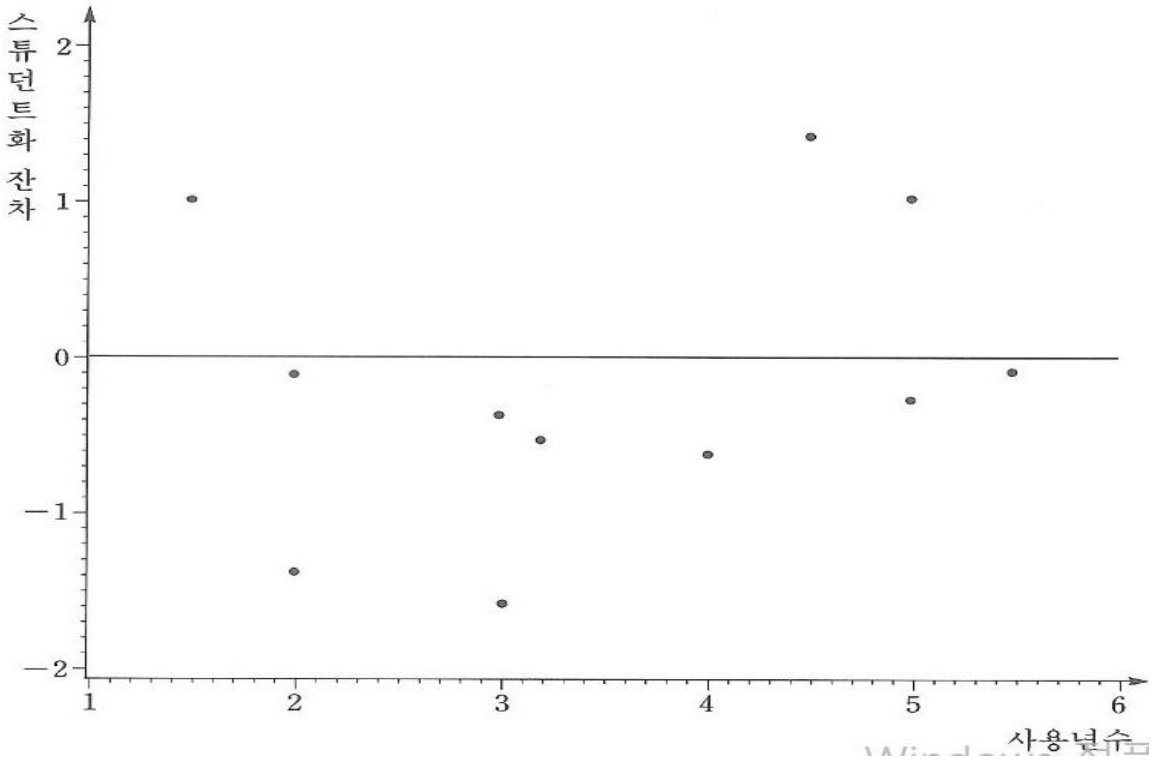
- 잔차분석:
  - 스튜던트화 잔차를 표준화된 오차항의 관측값처럼 생각하여 회귀모형의 가정을 검토할 수 있다.
  - 스튜던트화 잔차인

$$\hat{e}_{st,i} = \frac{e_i}{\hat{\sigma}}, \ i = 1, \cdots, n, \quad \longrightarrow \quad \left\{ \frac{y_1 - \hat{y}_1}{\widehat{sd}(e_1)}, \ \frac{y_2 - \hat{y}_2}{\widehat{sd}(e_2)}, \cdots, \frac{y_n - \hat{y}_n}{\widehat{sd}(e_n)} \right\}$$

    들이 표준정규분포에서의 서로 독립인 n개의 관측값과 유사하게 나타나는가를 검토하여, 단순선형회귀모형의 적용 타당성을 알아본다.

- 흔히 좌표평면에 점 $(x_1, \hat{e}_{st,1}), \cdots, (x_n, \hat{e}_{st,n})$을 나타내는 잔차도(residual plot)를 이용하여 분석을 하게 된다.

- 잔차도에서 스튜던트화 잔차들이 다음과 같이 나타나면



1. 선형성: 대략 0에 관하여 대칭적으로 나타남
2. 등분산성: 설명변수의 값에 따른 잔차의 산포가 크게 다르지 않음
3. 독립성: 점들이 특정한 형식을 가지고 나타나지 않음
4. 정규성: 대부분의 점이 ±2의 범위 내에 있음

=> 단순선형회귀모형을 적용하는 것이 타당해 보인다.

단순선형회귀모형의 가정에 어긋나는 경우

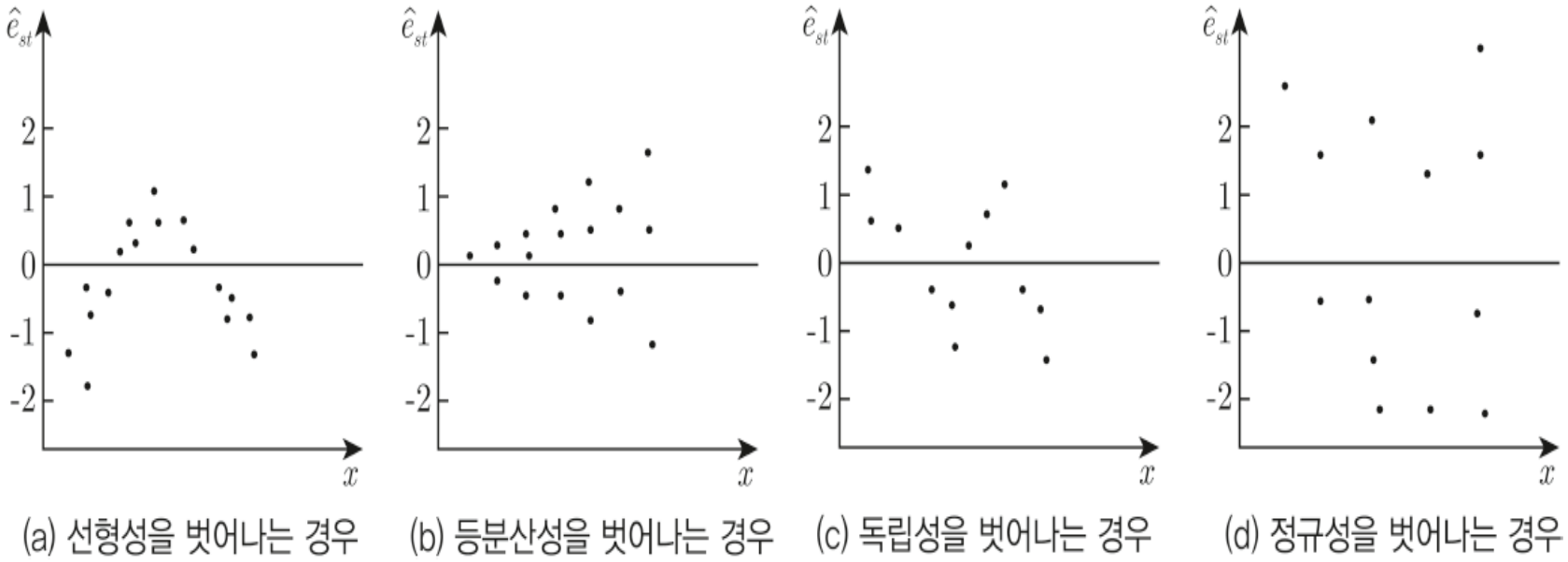


그림 8-10 단순선형회귀모형의 가정에 어긋나는 경우

통계적 추론의 순서

- 단순선형회귀모형을 적용한 통계적 추론시 분석 과정

1. 산점도를 그려서 대략적인 직선 관계를 파악한다.
2. 최소제곱회귀직선과 평균제곱오차를 구하여 잔차분석을 행한다. 이 때, 스튜던트화 잔차들이 대체로 0 에 관해 대칭적으로 나타나고 ± 2이내의 범위에 들어가며 특별한 패턴을 가지지 않음을 확인한다.
3. 잔차분석을 통과하면 신뢰구간이나 검정과 같은 추론을 행한다.

단순선형회귀분석 예제

Example

공동주택의 가구당 월간 전기요금이 월평균 기온에 따라 어떤 변화하는지 알아보기 위해 조사한 결과가 아래 표와 같다.

표 8-6 월평균 기온에 따른 월간 전기요금

(단위 : ℃, 천 원)

기온	전기요금	기온	전기요금	기온	전기요금	기온	전기요금	기온	전기요금
−2.2	14.0	0.8	12.2	8.2	12.9	15.2	11.9	21.8	10.3
−1.9	13.1	1.8	13.0	9.2	11.3	16.3	10.9	22.3	9.1
−1.7	14.4	3.9	11.3	14.2	10.8	21.1	9.6	23.6	7.5
−1.3	13.1	7.0	10.5	14.5	10.0	21.1	8.1	23.6	10.5
−0.7	14.8	8.0	9.7	14.9	9.9	21.5	9.2	24.8	10.0

(a) 단순선형회귀모형 적용이 가능한지 잔차분석을 하여라.

sol) [표 8-6]의 자료에 대하여 산점도를 그려보면 두 변수 사이에 대략 직선관계가 성립함을 알 수 있다.

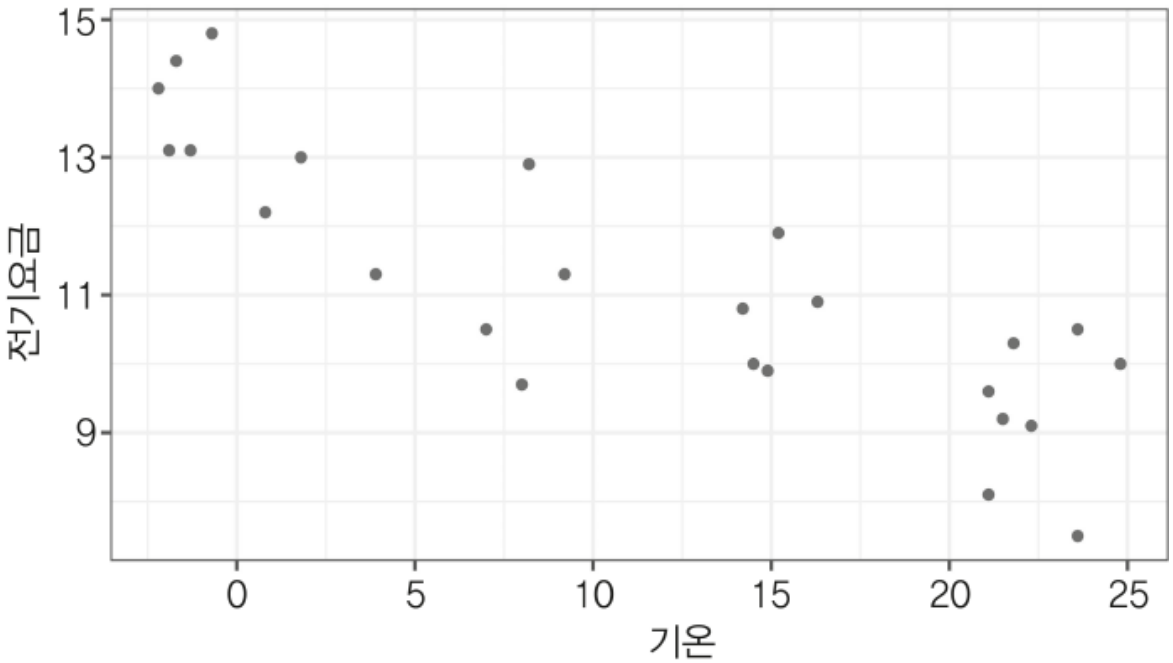


그림 8-11 [표 8-6]에 대한 산점도

- 이 자료에 최소제곱회귀직선을 적합하기 위한 통계량을 계산하면

$$n = 25, \quad \bar{x} = 11.44, \quad \bar{y} = 11.124$$

$$S_{xx} = 2210.44, \quad S_{yy} = 88.8856, \quad S_{xy} = -374.9945$$

- 이로부터 최소제곱회귀직선과 오차분산의 추정값을 구하면 다음과 같고

$$\hat{\beta}_1 = S_{xy}/S_{xx} = -0.17, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 13.06$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 13.06 - 0.17x$$

$$\hat{\sigma}^2 = SSE/(n-2) = \{S_{yy} - S_{xy}^2/S_{xx}\}/(n-2) = 1.0987$$

- 최소제곱회귀직선을 산점도에 그려보면 다음과 같다.

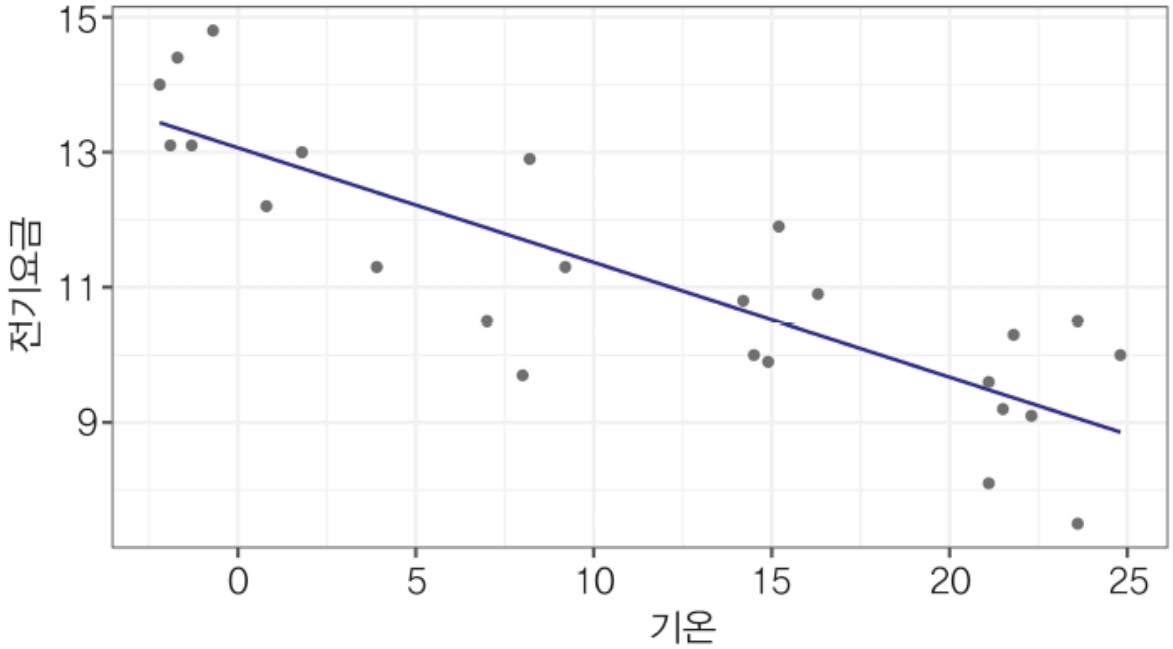


그림 8-12 [표 8-6]에 대한 산점도와 최소제곱회귀직선

- 최소제곱회귀직선 $\hat{y}_i = 13.06 - 0.17x_i$와 $\hat{\sigma} = \sqrt{\hat{\sigma}^2} = 1.0482$를 이용하여 스튜던트화 잔차인 $\hat{e}_{st,i} = \frac{y_i - \hat{y}_i}{\widehat{sd}(Y_i - \hat{Y}_i)}$을 계산하여 잔차도를 그려보았다.

- 잔차도에서 잔차들은 대략 0에 관해 대칭적이며 설명변수의 값에 따른 산포도 비슷하고, 특별한 패턴을 갖고 있지 않으며 모두 ± 2의 범위 이내에 있음을 알 수 있다.

- 이러한 회귀모형을 가정할 때, 회귀직선이 자료의 전체 산포 중에서 얼마나 설명하는가를 알기 위하여 결정계수를 계산하면

$$R^2 = \frac{SSR}{SST} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = 0.7157$$

로서, 전체 산포의 약 72%가 회귀직선에 의하여 설명된다.

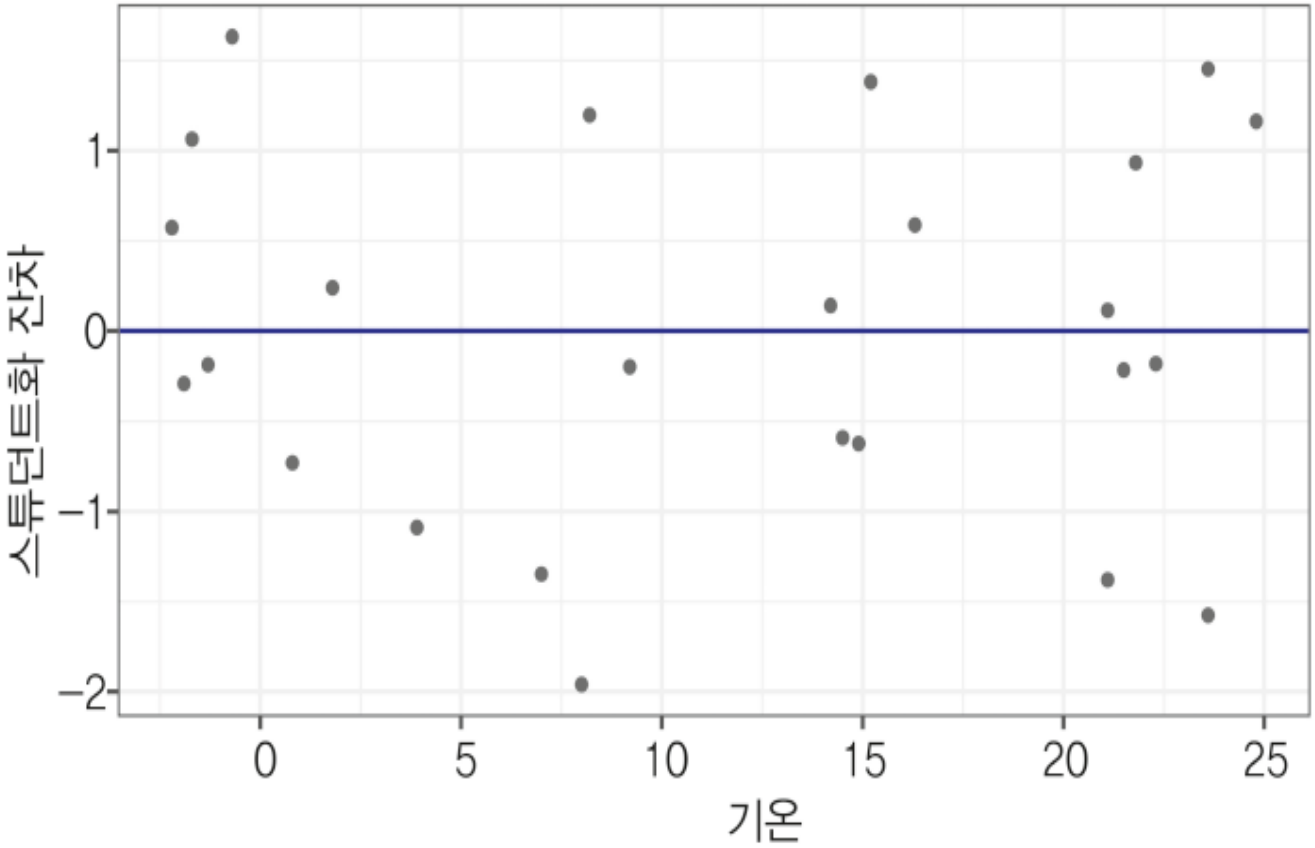


그림 8-13 [표 8-6]에 대한 잔차도

Example

(b) 유의수준 5%에서 회귀직선의 유의성검정을 하여라.

sol) 회귀직선의 유의성을 판단하기 위한 가설은 다음과 같다.

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

- 제곱합을 계산하면

  $SST = S_{yy} = 88.8856$

  $SSR = S_{xy}^2 / S_{xx} = 63.6165$

  $SSE = SST - SSR = 25.2691$

  이고 각각의 자유도는 25-1=24, 1, 25-2=23 이다.

- F검정통계량의 관측값은 평균제곱의 비인

$$f = \frac{SSR/1}{SSE/23} = 57.904$$

이고 F분포표에서 $F_{0.05}(1, 23) = 4.96$ 이므로,

$$f \geq F_{0.05}(1, 23)$$

이 성립한다. 즉, 유의수준 5%에서 회귀직선은 유의하다.

- 분산분석표로 과정을 정리하면 다음과 같다.

요인	제곱합	자유도	평균제곱	F값	유의확률
회귀	$SSR = 63.6165$	1	MSR=63.6165	f=57.904	0.0001
잔차	SSE=25.02691	23	MSE=1.0987		
계	$SST = 88.8856$	24			

Example

(c) 월평균기온이 10일 때, 월간 평균 전기요금에 대한 95%의 신뢰구간을 구하여라.

sol) 월평균기온이 $x = 10$일 때 $E(Y|x) = \beta_0 + \beta_1 10$의 추정값은

$$\hat{\beta}_0 + \hat{\beta}_1 10 = 13.06 + (-0.17)10 = 11.36$$

이 추정값의 표준오차를 계산하면

$$\hat{\sigma}\sqrt{n^{-1} + (x_0 - \bar{x})^2/S_{xx}} = 0.2121$$

- 한편, t분포표에서 $t_{0.025}(23) = 2.069$이므로 $\beta_0 + \beta_1 10$의 95%신뢰구간은 (단위:1000원) 다음과 같다.

$$(\hat{\beta}_0 + \hat{\beta}_1 10) \pm t_{0.025}(23)\ \hat{\sigma}\sqrt{n^{-1} + (x_0 - \bar{x})^2/S_{xx}} = 11.36 \pm 0.44$$

8.2장 다중회귀분석

박명현

서울대학교 학부대학

다중회귀분석(multiple regression analysis)

- 일반적으로 반응변수 Y의 변화는 두 개 이상의 설명변수 $(x_1, x_2, \cdots, x_k)$에 의하여 영향을 받는 경우가 많이 있다.

예) 광고비와 설비투자가 매출액에 미치는 영향

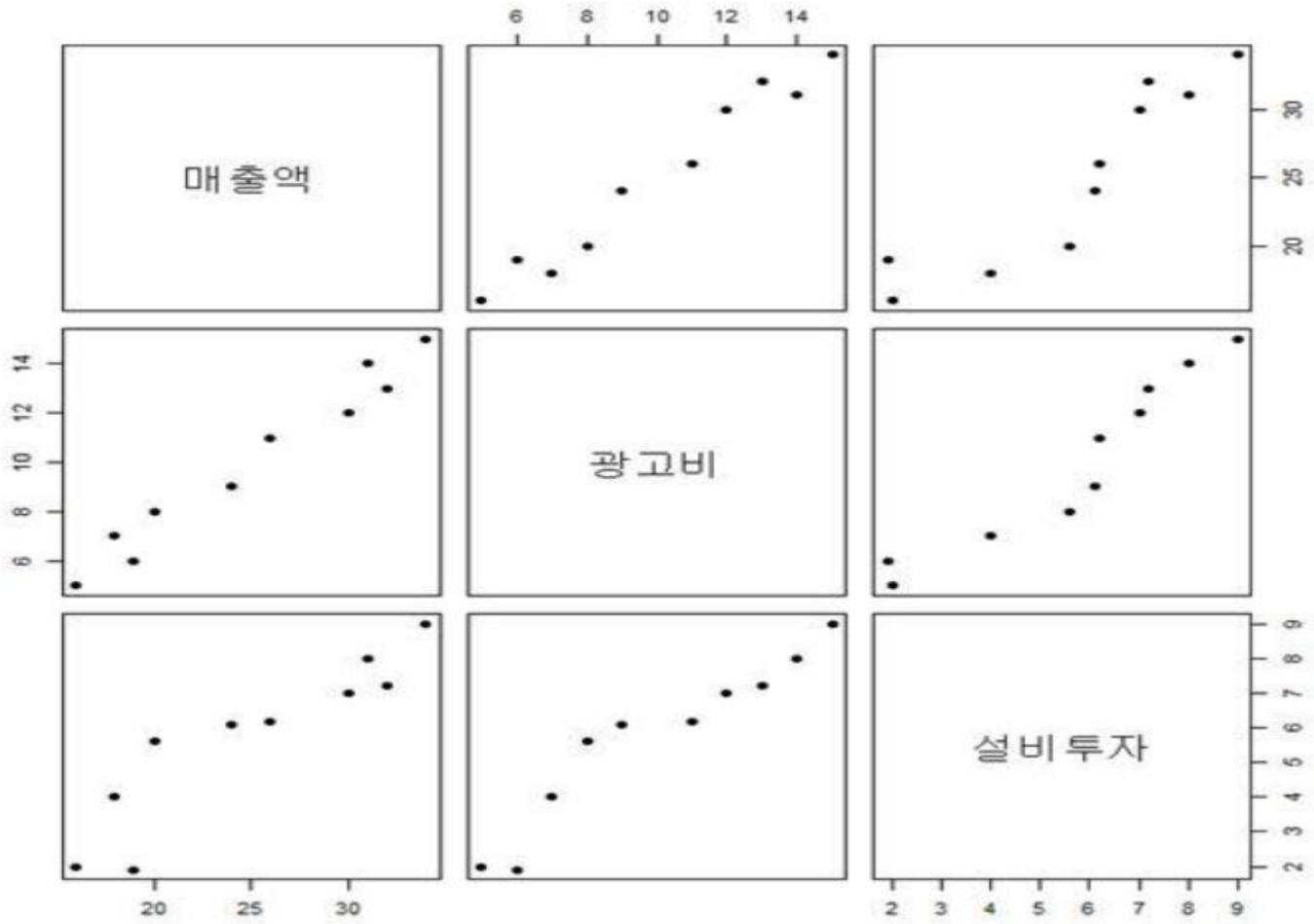
- 반응변수 Y: 매출액
- 첫 번째 설명변수 x_1: 광고비
- 두 번째 설명변수 x_2: 설비투자

$\Longrightarrow$ 하나의 설명변수만을 고려하는 단순회귀모형과 달리 두 개 이상(multiple) 의 설명변수로 반응변수를 설명

예제

- 광고비 및 설비투자와 매출액의 연관성 비교 자료(단위: 억원)

Figure: Scatter plot matrix



다중선형회귀모형(multiple linear regression model)

- 다음의 다중선형회귀모형을 고려해보자.

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i \quad (i = 1, 2, \cdots, n)$$

예) 광고비(x_{1i})와 설비투자(x_{2i})로 기업의 매출액(y_i) 설명하기

	y_매출액	x1_광고비	x2_설비투자
0	16	5	2.0
1	19	6	1.9
2	18	7	4.0
3	20	8	5.6
4	24	9	6.1
5	26	11	6.2
6	30	12	7.0
7	32	13	7.2
8	31	14	8.0
9	34	15	9.0

다중선형회귀모형

- 다중선형회귀모형

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i \quad (i = 1, 2, \cdots, n)$$

- 반응변수, 종속변수 $Y_1, \cdots, Y_n$
- k개의 설명변수, 독립변수 $x_{1i}, \cdots, x_{ki}$
  (확률변수가 아닌 상수로 취급)
- $\beta_0, \beta_1, \cdots, \beta_k$: 미지의 모회귀계수(실수)
- ϵ_i: 기댓값이 0, 분산이 σ^2인 오차항을 나타내는 확률변수
- $\sigma^2 > 0$: 미지의 오차분산

- 다중선형회귀식

$$E(Y_i | x_{i1}, \cdots, x_{ki}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} \quad (i = 1, 2, \cdots, n)$$

다중선형회귀모형의 가정사항

- 오차항의 가정사항: $\epsilon_i \sim_{i.i.d} N(0, \sigma^2)$ $(\sigma^2 unknown)$ $(i = 1, \cdots, n)$

  1) 오차항(ϵ_i)의 평균은 0이다: $(E(\epsilon_i) = 0)$
  2) 오차항의 분산 σ^2는 x_{ji}의 모든 값에 대하여 항상 같다.
     (등분산 가정: $Var(\epsilon_i) = \sigma^2 > 0$)
  3) 오차항 ϵ_i와 ϵ_j는 서로 독립이다 (for all $i, j; i \neq j$: 독립성)
  4) 오차항은 정규분포를 따른다. $\epsilon_i \sim N(0, \sigma^2)$

- Y변수는 각각 평균 $\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}$, 분산 σ^2인 정규분포를 하고 서로 독립이다.

$$Y_i \sim N(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}, \sigma^2) \ (\sigma^2 unknown)$$

최소제곱법(Least squares method)

- 최소제곱법(Least squares method)
  다중선형회귀모형 $Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} + \epsilon_i$ $(i = 1, 2, \cdots, n)$에서 오차의 제곱합 $\sum_{i=1}^{n} \epsilon_i^2$을 최소로 하는 $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$을 그 추정값(estimated value)으로 찾는다.

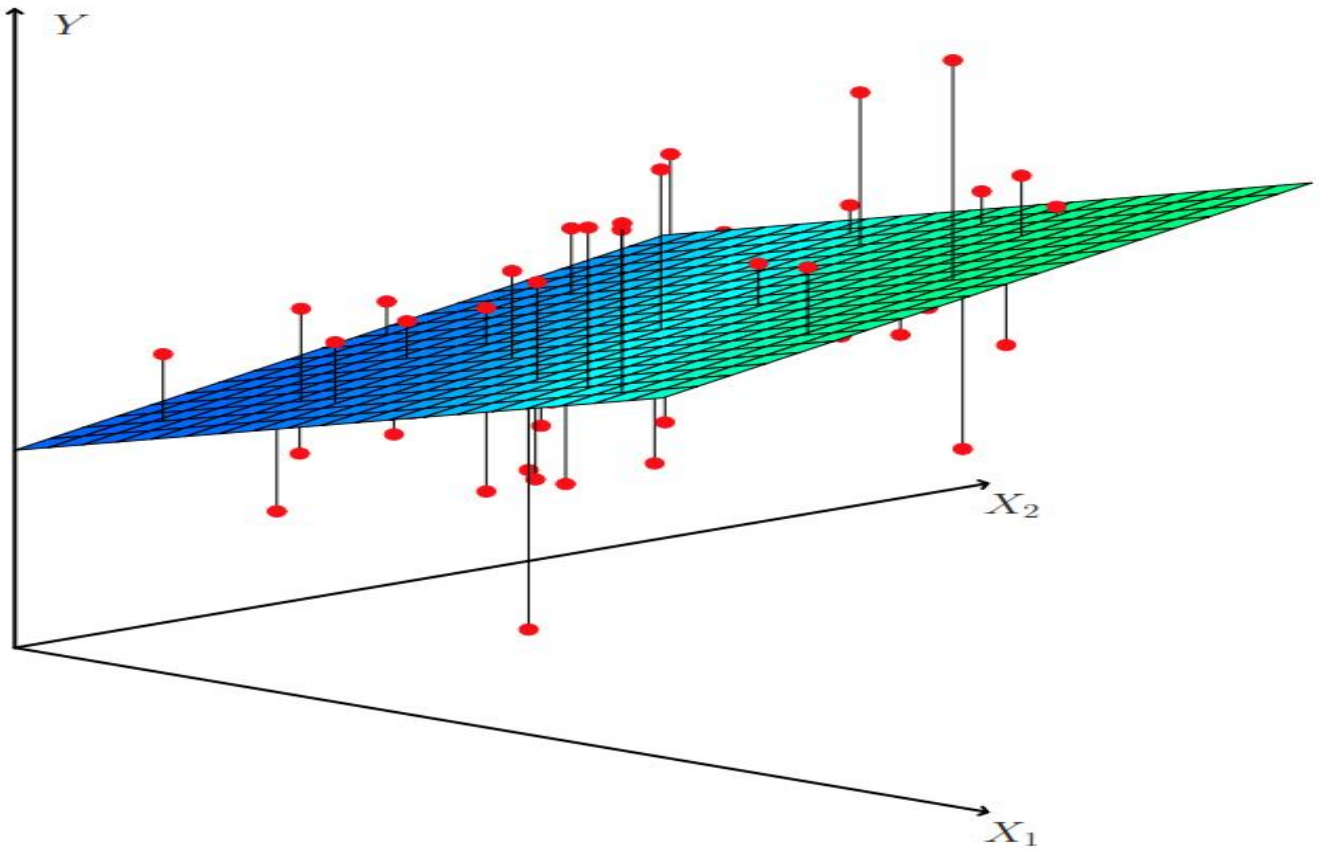
$$(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k) = \arg\min_{(\beta_0, \beta_1, \cdots, \beta_k)} \sum_{i=1}^{n} \epsilon_i^2$$

- 최소제곱추정값 $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$는 오차의 제곱합(SS)을 $\beta_j, (j = 0, 1, \cdots, k)$에 대하여 각각 편미분하여 얻은 식을 0으로 놓은 다음의 연립방정식을 풀면된다.

$$\frac{\partial SS}{\partial \beta_j} = 0, \quad (j = 0, 1, \cdots, k)$$

$*$ 계산은 실제로 컴퓨터로 수행하게 된다.

최소제곱법(Least squares method)



- In a three-dimensional setting, with two predictors and one response, the least squares regression line becomes a plane.
- The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane.

$E(Y_i)$와 $\hat{y}_i$

- 추정된 다중선형회귀모형의 회귀식

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki} \quad (i = 1, 2, \cdots, n)$$

- $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$: 모회귀계수 $\beta_0, \beta_1, \cdots, \beta_k$의 최소제곱추정값.

- $y_i = E(y_i|x) + \epsilon_i$

  $y_i = \hat{y}_i + e_i$

- 회귀분석의 목적:
  다중선형회귀식(반응변수 Y의 기대값)을 추정하는 것.

추정계수 $\hat{\beta}_j$의 해석

- 추정계수 $\hat{\beta}_j$ $(j = 1, \cdots, k)$의 해석:
  j번째 설명변수 x_j를 제외한 나머지 설명변수 $(x_1, \cdots, x_{j-1}, x_{j+1}, \cdots, x_p)$들이 고정되어 있다는 가정 하에 x_j가 1단위 변화할 때 y의 변화량이다.

- 추정된 $\hat{\beta}_0$은 $x_1 = x_2 = \cdots = x_k = 0$일 때 Y절편으로 설명변수 값들의 범위가 0을 포함하지 않을 경우 큰 의미를 두지 않는다. 하지만 회귀식의 적합성을 높이기 위하여 모형에 포함한다.

추정계수 $\hat{\beta}_j$의 해석 예제

예) 매출액 예제에서 최소제곱법에 의한 추정회귀식은 다음과 같다. (x_1: 광고비 x_2 : 설비투자)

$$\hat{y} = 6.0985 + 2.2567x_1 - 0.643x_2$$

- 추정계수 2.2567[-0.643]의 의미: $x_2[x_1]$가 고정되어 있을 때 $x_1[x_2]$가 1단위(1억원) 증가할 때 매출액이 2.2567단위[0.643단위] 증가[감소]함을 의미한다.
- 추정계수 6.0985의 의미: $x_1, x_2 = 0$일 때의 매출액을 나타냄. 설명변수들의 범위가 0을 포함하지 않는다면 큰 의미를 두지 않는다.

오차분산 σ^2의 추정

- 잔차(residual): $e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ji}, \quad i = 1, 2, \cdots, n$
- 잔차의 값들이 작다는 것의 의미는 최소제곱회귀직선에 의한 적합이 실제 관측 결과를 잘 설명해주는 것이다.
- 다중선형회귀모형에서의 오차분산 σ^2도 오차제곱평균(MSE)으로 σ^2을 추정한다.

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-k-1} = \frac{1}{n-k-1} \sum_{i=1}^{n} e_i^2$$

이 때, $E(MSE) = \sigma^2$이다.

제곱합의 분해(Decomposition)

- 총제곱합(SST)은 잔차제곱합(SSE)과 회귀제곱합(SSR)으로 분해된다.

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$SST \quad = \quad SSE \quad + \quad SSR$$

- 설명변수가 k개일 때, 이들의 자유도는

$$n-1 \quad = \quad (n-k-1) \quad + \quad k$$

- 오차제곱평균(MSE), 회귀제곱평균(MSR), 총제곱평균(MST)은

$$MSE = \frac{SSE}{(n-k-1)}, \qquad MSR = \frac{SSR}{k}, \qquad MST = \frac{SST}{n-1}$$

결정계수 R^2

- 결정계수 R^2

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

  - 다중회귀분석에서는 독립변수가 추가되면 R^2가 증가함
    => 과적합(overfitting)의 문제 발생

- 수정된 결정계수 *Adjusted* R^2

$$Adj\ R^2 = 1 - \frac{MSE}{MST} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

  - 회귀모형에 변수를 추가하는것에 패널티를 줌. 따라서 유용한 변수를 추가해야 *Adjusted* R^2가 증가함

- $Adj\ R^2 \leq R^2$

다중선형회귀모형에 대한 유의성 F검정

- 회귀직선이 의미가 있는지를 검정하기 위해서 모든 설명변수들과 반응변수의 선형적 관계를 검토해보자.
- 귀무가설 및 대립가설

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad H_1 : \textit{At least one } \beta_j \neq 0 \ \ (j = 1, \cdots, k)$$

- F검정은 회귀모형의 모든 모수의 유의성에 대한 총체적검정(overall test for significance of all parameters)을 수행한다.

(일반적으로 β_0는 제외)

- 검정통계량

$$F = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)}$$

(SSR: 회귀식에 의한 변동량, SSE: 오차에 의한 변동량)

다중선형회귀모형에 대한 유의성 F검정

- 귀무가설 H_0가 사실일 때 F 검정통계량은 다음의 F분포를 따른다.

$$F \sim F(k, n-k-1)$$

- F 검정통계량의 관측값이 f이면 유의확률은

$$P = P\{F \geq f\}, \quad F \sim F(k, n-k-1)$$

이고 유의수준 α의 기각역은 다음과 같다.

$$f \geq F_\alpha(k, n-k-1)$$

Table: 회귀직선의 유의성검정을 위한 분산분석(ANOVA)표

요인	제곱합	자유도	평균제곱	F값	유의확률
회귀	SSR	k	MSR	f=MSR/MSE	$P\{F \geq f\}$
잔차	SSE	$n-k-1$	MSE		
계	SST	$n-1$			

다중선형회귀모형 예제

Example (예제 8.10)

어떤 플라스틱 제품의 강도(kg/cm^2)가 그 제품을 생산하는 사출 과정 중의 사출온도와 사출시간(sec)에 따라서 어떻게 변화하는가를 알아보기 위하여 실험을 한 결과 다음과 같은 자료를 얻는다.

사출온도(x_1)	195	179	205	204	201	184	210	209
사출시간(x_2)	57	61	60	62	61	54	58	61
강도(y)	81.4	122.2	101.7	145.2	135.9	64.8	92.1	113.8

사출온도와 사출시간을 각각 설명변수 x_1, x_2로 하고 강도를 반응변수 Y로 하는 다중선형회귀모형을 적용할 때, 최소제곱법에 의하여 추정된 회귀직선은 $\hat{y} = -428.8850 - 0.2338x_1 + 9.8295x_2$ 이고, 제곱합이 다음과 같이 주어졌을 때, 이 계산 결과를 이용하여 모회귀함수의 유의성 검정을 유의수준 5%에서 하고, 결정계수 R^2의 값을 구하여라.

$$SST = 5257.8788, \quad SSE = 591.1514$$

sol) 검정하고자 하는 가설은 다음과 같다.

$$H_0 : \beta_1 = \beta_2 = 0 \quad H_1 : At\ least\ one\ \beta_j \neq 0 \ \ (j = 1, 2)$$

- 회귀제곱합을 구하면

$$SSR = SST - SSE = 4666.7274$$

이고, 설명변수가 $k = 2$ 이므로 SST, SSE, SSR의 자유도는 각각

$$n - 1 = 7, \quad n - k - 1 = 5, \quad k = 2$$

- 따라서, 회귀평균제곱, 잔차평균제곱과 F 검정통계량을 계산하면

$$MSR = SSR/2 = 2333.3637, \quad MSE = SSE/5 = 118.2303$$

$$f = MSR/MSE = 19.736$$

- F분포표에서 $F_{0.05}(2,5) = 5.79$이므로 기각역은 다음과 같다.

$$f \geq F_{0.05}(2,5) = 5.79$$

- 따라서, 유의수준 5%에서 모회귀함수가 유의하다고 할 수 있고, 통계패키지를 이용하면 P=0.0042임을 알 수 있다.

- 이러한 검정결과를 분산분석표로 정리하면 다음과 같다.

요인	제곱합	자유도	평균제곱	F값	유의확률
회귀	4666.7274	2	2333.3637	f=19.736	$P = 0.0042$
잔차	591.1514	5	118.2303		
계	5257.8788	7			

- 결정계수의 값은 $R^2 = SSR/SST = 0.8876$으로서, 자료 전체의 산포 중에서 약 89%가 회귀직선에 의하여 설명됨을 알 수 있다.

다중선형회귀모형 예제

Example

광고비와 설비투자가 매출액에 어떤 영향을 미치는가를 알아보고자 한다. 매출액은 반응변수 Y이고, 광고비와 설비투자는 각각 x_1, x_2이다.

광고비(x_1)	5	6	7	8	9	11	12	13	14	15
설비투자(x_2)	2	1.9	4	5.6	6.1	6.2	7	7.2	8	9
매출액(y)	16	19	18	20	24	26	30	32	31	34

최소제곱법에 의한 회귀분석 결과의 분산분석표를 이용하여 모회귀함수의 유의성검정을 유의수준 1%에서 하고, R^2의 값을 구하여라.

```
Response: y
          Df Sum Sq Mean Sq  F value    Pr(>F)
x1         1 370.95  370.95 232.5153 1.256e-06 ***
x2         1   1.89    1.89   1.1828    0.3128
Residuals  7  11.17    1.60
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

sol) 검정하고자 하는 가설은 다음과 같다.

$$H_0 : \beta_1 = \beta_2 = 0 \quad H_1 : At\ least\ one\ \beta_j \neq 0 \ \ (j = 1, 2)$$

- 회귀제곱합 SSR은 370.95+1.89=372.84이고, MSR은 186.42이다.
- 유의수준 $\alpha = 0.01$하에서 검정통계량은 $f = 186.42/1.6 = 116.51 > F_{0.01}(2, 7)$ 이므로 귀무가설이 기각된다.
- 즉, 총체적 검정인 F검정 결과 광고비 또는 설비투자비 중 적어도 하나는 매출액에 선형적으로 유의적인 의미를 갖는 것으로 해석된다.
- 결정계수의 값은 $R^2 = SSR/SST = 372.84/384.01 = 0.9709$로서 광고비 및 설비투자비가 매출액을 선형적으로 많은 부분, 즉 전체 총변동량의 약 97%정도를 설명하고 있다.

개별 계수 β_j에 대한 추론

- 회귀모형의 F검정결과 귀무가설 $H_0 : \beta_1 = \cdots = \beta_k = 0$이 기각되면, 최소한 하나 이상의 β_j가 0이 아닌 것이므로 개별적 변수에 대한 t유의성 검정을 실시한다.
- 귀무가설 및 대립가설

$$H_0 : \beta_j = 0 \quad H_1 : \beta_j \neq 0$$

- 표본분포

$$\frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \sim t(n-k-1)$$

여기서 $s(\hat{\beta}_j)$는 $\hat{\beta}_j$의 추정된 분산의 제곱근인 표준오차

- β_j에 관한 $100(1-\alpha)\%$ 신뢰구간:

$$\hat{\beta}_j \pm t_{\alpha/2}(n-k-1) \times s(\hat{\beta}_j)$$

개별 계수 β_j에 대한 유의성 t검정

- β_j에 대한 유의성 검정:
  귀무가설 $H_0 : \beta_j = 0$ 에 관한 검정통계량은 다음과 같다.

$$T = \frac{\hat{\beta}_j - 0}{s(\hat{\beta}_j)}$$

- P값 방법: 검정통계량의 관측값을 t라고 하면 유의확률은 다음과 같고, 만약 $P \leq \alpha$이면, H_0를 기각한다.

$$P = P(|T| \geq |t|)$$

- 기각역 방법: 유의수준 α에서 다음을 만족하면 H_0를 기각한다.

$$|t| \geq t_{\alpha/2}(n-k-1)$$

유의성 t검정 예제

Example

아래 표는 광고비와 설비투자가 매출액에 미치는 영향을 조사한 자료를 이용하여 추정한 회귀모형의 결과이다.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.0985     1.3618   4.478 0.002872 **
x1            2.2567     0.4048   5.575 0.000838 ***
x2           -0.6430     0.5913  -1.088 0.312805
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.263 on 7 degrees of freedom
Multiple R-squared:  0.9709,    Adjusted R-squared:  0.9626
F-statistic: 116.8 on 2 and 7 DF,  p-value: 4.195e-06
```

광고비와 설비투자의 회귀계수들이 0과 유의적으로 다른지 유의수준 5%하에서 개별 검정을 하고, 각 계수의 95% 신뢰구간을 구하여라.

sol) 각 회귀 계수에 대한 가설은 다음과 같다.

$$H_0 : \beta_j = 0 \quad H_1 : \beta_j \neq 0$$

- 양측검정이고, $(n-k-1) = 7$이므로 t분포표로부터 $t_{0.025}(7) = 2.365$임을 알 수 있다.
- β_1에 관한 95% 신뢰구간:

$$\hat{\beta}_1 \pm t_{0.025}(7) \times s(\hat{\beta}_j) = 2.2567 \pm 2.365 \times 0.4048$$

- β_2에 관한 95% 신뢰구간:

$$\hat{\beta}_2 \pm t_{0.025}(7) \times s(\hat{\beta}_j) = -0.643 \pm 2.365 \times 0.5913$$

- β_1에 대한 검정: 검정통계량은 다음과 같고, 기각역에 속하여 유의수준 5%하에서 귀무가설 $H_0 : \beta_1 = 0$을 기각한다.

$$t = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} = \frac{2.2567}{0.4048} = 5.575 \geq t_{0.025}(7) = 2.3652$$

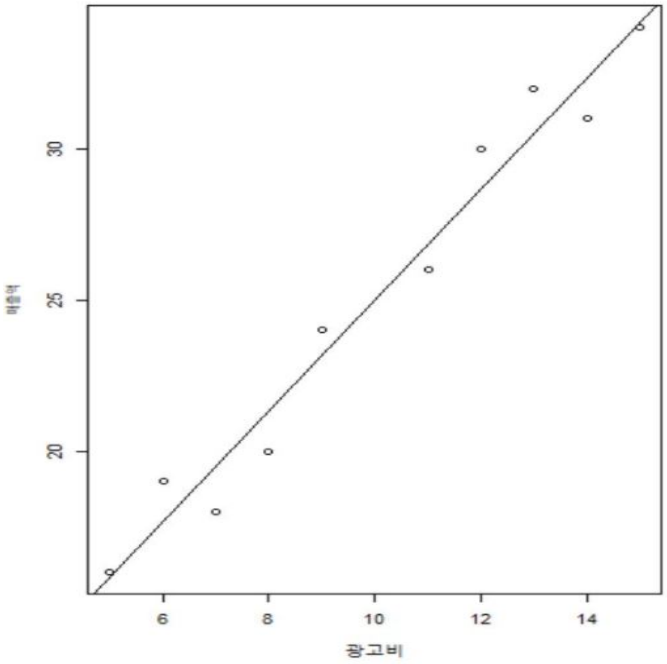
P값도 $P < \alpha = 0.05$로 귀무가설을 기각한다.

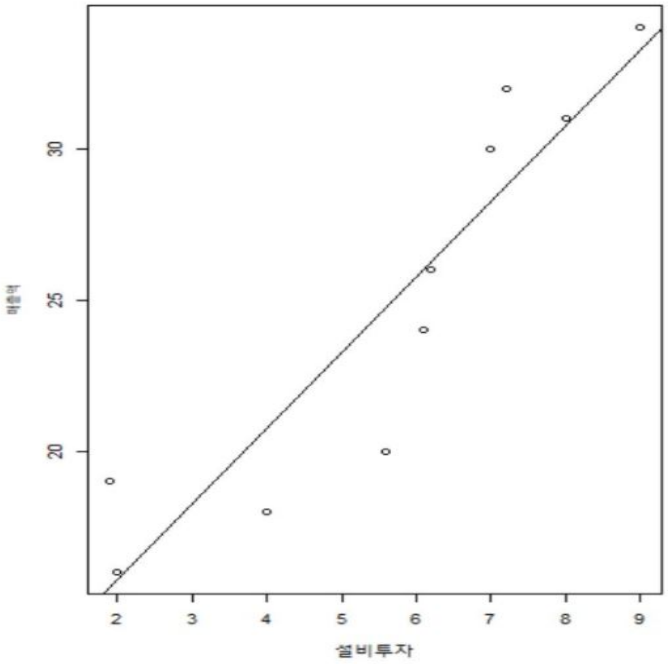
- β_2에 대한 검정: 검정통계량은 다음과 같고, 기각역에 속하지 않으므로 유의수준 5%하에서 귀무가설 $H_0 : \beta_2 = 0$을 기각할 수 없다.

$$t = \frac{\hat{\beta}_2}{s(\hat{\beta}_2)} = \frac{-0.643}{0.5913} = -1.088 \geq -t_{0.025}(7) = -2.365$$

P값도 $P > \alpha = 0.05$로 귀무가설을 기각할 수 없다. 즉, 설비투자비는 매출액에 통계적으로 비유의적인 관계로 판단된다.

다중선형회귀분석시 주의점

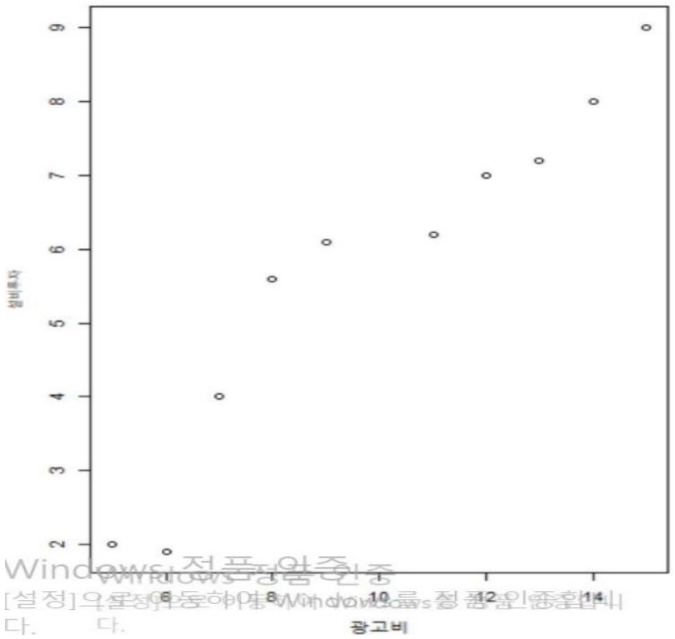




- 광고비와 매출액의 단순선형회귀모형에서 광고비는 F검정에서 통계적으로 유의적이다. ($\hat{y} = 6.63 + 1.83x_1$, $R^2 = 0.966$)
- 설비투자와 매출액도 F검정에서도 설비투자가 매우 유의적인 변수로 설명이 된다. ($\hat{y} = 10.72 + 2.50x_2$)

다중선형회귀분석시 주의점

Figure: 광고비와 설비투자의 산점도



- 광고비와 설비투자의 선형관계를 나타내는 계수인 상관계수가 0.95로 높음에 주의하자.
- 광고비와 설비투자의 높은 선형관계로 인하여 매출액을 설명하는 변수로서의 설비투자의 역할은 광고비가 전부 가지고 가서 설비투자의 역할은 더 이상 없게 된 것이다.

다중공선성(multicollinearity)

- 다중공선성: 설명변수 중 하나와 다른 설명변수들의 선형결합(linear combination)이 매우 높은 상관관계에 있을 때

  $\Longrightarrow$ 설명변수들 사이의 선형독립성(linear independence)이 성립하지 못한 상태

  예) 설명변수에 키, 앉은 키, 다리 길이가 다 있다고 가정하자. 그런데 키는 앉은 키와 다리 길이를 더한 값과 거의 일치한다. 이 경우, 다중공선성 문제가 발생한다.

- 다중공선성이 존재하면 설명변수들의 기울기를 정확히 계산하지 못하거나 아예 계산이 되지 않는다.($Var(\hat{\beta}_j) \uparrow$)

다중공선성(multicollinearity)

- 다중공선성의 진단

  1. 개별 설명변수끼리의 산점도에서 거의 직선에 가까운 선형관계를 보이는 경우
  2. 설명변수 하나를 뺄 때와 추가할 때 다른 설명변수들의 기울기 추정값들이 많이 변동하는 경우
  3. 분산분석표에 의해서 유의적으로 기울기들이 모두 0이 아니지만 문제가 되는 설명변수의 기울기는 유의적이지 않은 경우
  4. 유의성 F 검정에서는 회귀식이 유의하지만, 개별 유의성 t 검정에서는 그렇지 않은 경우 의심해 볼 수 있다.
  5. 분산확대인수(variance inflation factor)를 계산

- 처방 : 설명변수들 끼리 선형 독립성을 유지해야 한다. 이를 위해, 문제가 되는 설명변수를 제거한다.

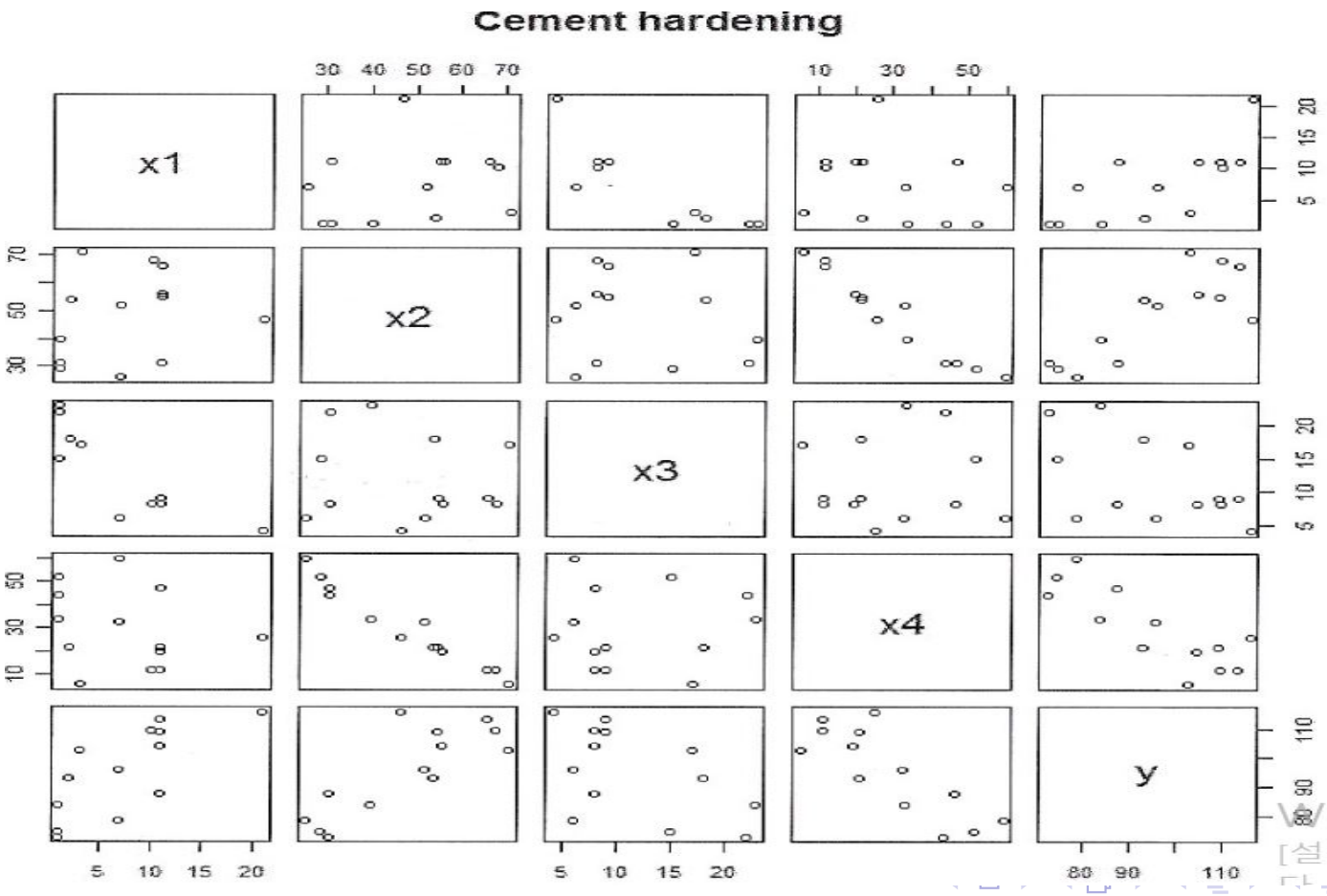
회귀진단(regression diagnostics)

- 단순선형회귀모형에서의 진단 및 처방과 비슷
- 회귀진단 항목:
  (a) 모형의 선형성(linearity)
  (b) 오차의 등분산성(homoscedasticity), 독립성(independence), 정규성(normality)
  (c) 이상값(outlier)의 존재
  (d) 영향력이 큰 관찰치(influential observation)의 존재
  (e) 다중공선성(multicollinearity)
- 설명변수들이 여러개이므로 산점도행렬(scatter plot matrix)을 이용할 수 있다.

산점도행렬(scatter plot matrix) 예

- 선형모형 적합성, 이상치나 영향력이 큰 관찰치 존재 여부는 단순 선형회귀분석과 같이 산점도로 확인 할 수 있다.

Figure: Y와 x_1, x_2, x_3, x_4의 산점도행렬



잔차분석(analysis of residual)

- 오차에 대한 가정사항을 진단하려면 잔차도를 그린다.

- 잔차그림은 단순회귀분석에서와 같이 (semi)스튜던트화 잔차들을 이용한다.

$$\hat{e}_{st,i} = \frac{e_i}{\widehat{sd}(e_i)} = \frac{(y_i - \hat{y}_i)}{\widehat{sd}(Y_i - \hat{Y}_i)} \quad (i = 1, \cdots, n)$$

- 그런데 다중회귀에서는 설명변수가 여러 개이므로 설명변수의 값을 가로축으로 할 수 없다. 따라서 추측값 $\hat{y}$을 가로축으로, 잔차를 세로축으로 하여 잔차도를 그린다.

$$(\hat{y}_1, \hat{e}_{st,1}), (\hat{y}_2, \hat{e}_{st,2}), \cdots, (\hat{y}_n, \hat{e}_{st,n})$$

잔차분석

- 그림으로부터 다중선형회귀모형의 적용이 타당함을 알 수 있다.

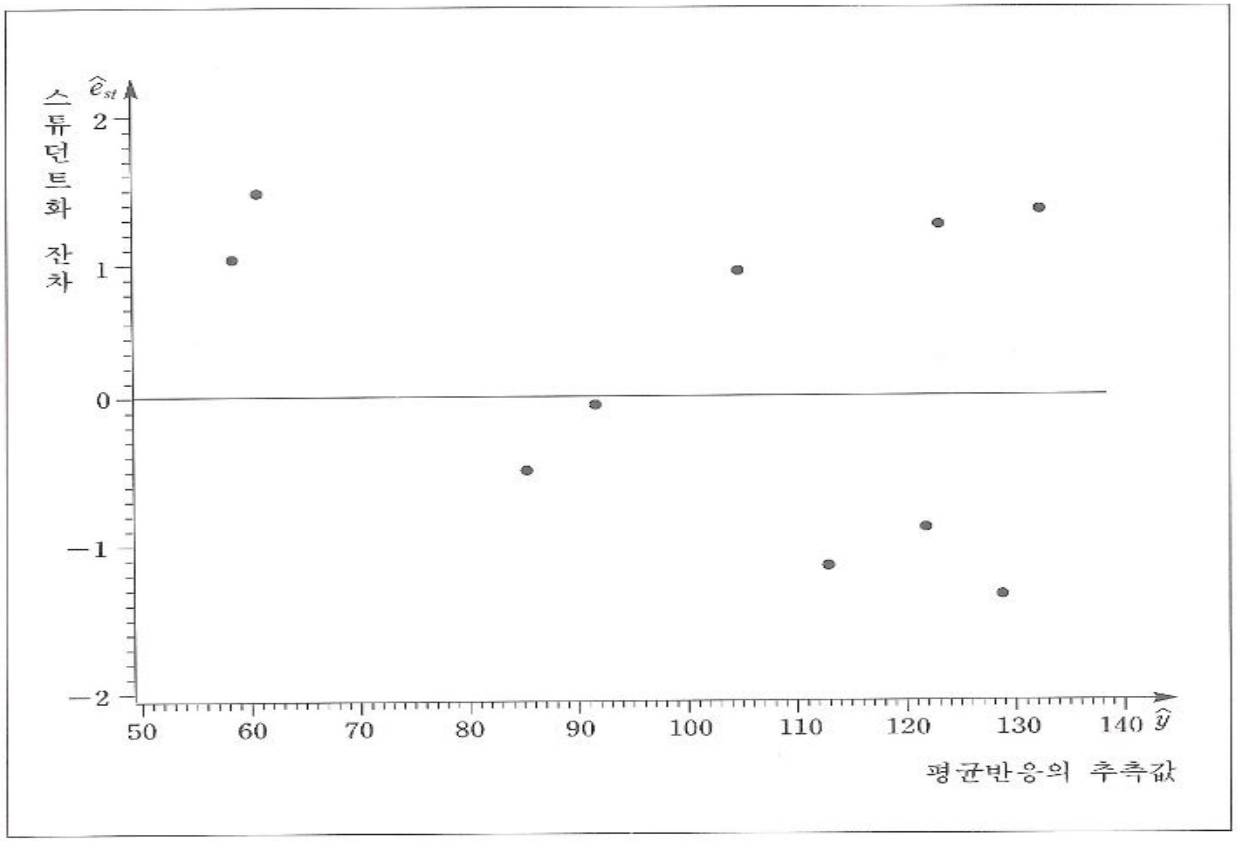



Figure: 예제 8.10 자료에 대한 잔차도