기초통계학

Basic Statistics

기초통계학 I	기초통계학 II
○ 통계학이란? ○ 기술통계 – 표와 그래프, 수치적 해석 ○ 확률 ○ 확률변수와 확률분포 ○ 표집분포	○ 통계적 추론 – 추정, 검정 ○ 분산분석 ○ 회귀분석 ○ 범주형 자료분석

숙명여자대학교 SOOKMYUNG WOMEN'S UNIVERSITY

■ 통계학이란?

○ 포털사이트 검색창에서 "그냥도전 동전 돌리기"라고 검색

- ◦ 500원짜리 동전을 돌렸을 때 학이 나올 확률이 70% 정도
- ◦ 실제로 500원짜리 동전을 1000번 돌리는 실험
- ◦ 실험 결과 1000번 중 학이 679번이 나옴
- ◦ 이 결과를 토대로 학이 나올 가능성이 70%정도 된다는 것이 얼추 맞다고 주장

■ 모집단과 표본

◉ 조선(대한제국 포함)시대 **임금의 수명**은?

○ 자료 수집 ⇨ 전체 자료 수집이 가능

【표 1】 조선 임금의 수명

임금	수명	임금	수명	임금	수명	임금	수명	임금	수명
태조	74	세조	52	명종	32	숙종	60	철종	33
정종	63	예종	20	선조	57	경종	37	고종	68
태종	56	성종	38	광해군	67	영조	83	순종	53
세종	54	연산군	31	인조	55	정조	49		
문종	39	중종	57	효종	41	순조	45		
단종	17	인종	31	현종	34	헌종	23		

숙명여자대학교 SOOKMYUNG WOMEN'S UNIVERSITY

◉ 궁에서 임금과 같이 생활했던 **내시의 수명**은?

- 자료 수집 ⇨ 전체 자료 수집이 가능?
- 이윤묵(1741~1816)의 양세계보(養世係譜)에는 81명의 내시 기록
- 평균수명은 70세, 81명 중 100세 이상 생존한 사람은 3명
- Q/P. **81명이 전체 내시를 대표할 수 있는가?**

● 모집단(population)

- ○ 통계학에서는 잘 정의된 연구목적과 이에 대한 명확한 연구대상을 설정
  - 예) 임금의 평균수명, 내시의 평균수명
- ○ 연구대상이 되는 모든 개체의 집합
  - 예) 전체 임금, 전체 내시
- ○ 실제로 관심을 가지는 것은 대상 자체보다는 그 대상의 속성에 관심을 가지기 때문에 전체 대상의 속성이 모집단이 되기도 함
  - 예) 전체 임금의 수명, 전체 내시의 수명

- 대부분의 모집단은 매우 커 전체를 조사하기 힘듬
  - 예) 대통령 선거에서 어떤 후보자의 지지율?
  - 대한민국 18대 대통령선거에서 유권자인 선거인 수는 40,507,842명
- 모집단을 명확하게 정의할 수 있는 경우도 있지만 애매한 경우도 있음
  - 예) "그냥도전 동전 돌리기"에서의 모집단은?
  - 통계분석을 할 때 이런 문제를 심심치 않게 만나는데 이 경우 모집단을 동전 돌리기 실험을 무한히 많이 반복수행하여 결과를 모아 놓은 것으로 이해

○ 개념적으로 규정한 조사 대상 전체를 **목표모집단(target popuation)** 또는 대상모집단이라고 하고 실제로 표본을 추출하기 위해 규정한 조사 대상 전체를 **조사모집단(survey population)**

- 예) 경제활동인구조사
- 목표모집단: 군인 및 수감자 등을 제외한 대한민국 영토 내에 거주하는 15세 이상 모든 국민
- 조사모집단: 조사의 편의나 여건을 고려해 도서지역, 기술시설 및 특수시설 거주자는 조사 대상에서 제외

○ 전수조사(census)

- ◦ 모집단 전체를 대상으로 조사하는 경우
- ◦ 센서스(census)는 추정하다(to estimate)라는 뜻의 라틴어 “censere”에서 유래 되었으며 센서스, 공공치안, 국가재정 등의 일을 담당하던 고대로마 관료를 censor라고 함
  - 아우구스투스(Augustus)가 센서스를 위해 출생 도시로 가서 호적 등록하라는 명령을 내렸으며 이 명령에 따라 예수의 부모는 고향을 가는 과정에서 예수를 출산
- ◦ 144년에 중국 한나라에서 실시한 조사에서는 994만 가구에 4973만 명이 사는 것으로 기록

○ 우리나라의 경우

- 1949년에 정부수립 후 대규모 조사를 시작하였으나 한국전쟁 중 자료가 모두 소실
- 1955년에 간이총인구조사를 거쳐 1960년까지 국세(國勢)조사라는 이름으로 실시
- 1963년 통계위원회에서 일본식 용어인 국세조사를 사용금지 하였으며 이후 '센서스'와 '총조사'라는 용어를 번갈아 사용하다가 1990년 이후 총조사로 통일
- 2010년 11월에 실시한 '인구주택총조사'에서는 인터넷을 이용한 조사가 본격적으로 병행 실시했으며 2015년에는 전수조사 대신 행정자료 이용과 표본조사를 병행

● 표본(Sample)

- ○ 모집단으로부터 선택된 일부의 개체
  - 예) 양세계보에 수록된 81명의 내시
  - 예) "그냥도전 동전 돌리기"에서 나온 1000번의 동전 결과
- ○ Q/P. **추출된 표본이 모집단 특성을 대표할 수 있는가**?
  - 예) 양세계보에 기록된 내시가 모두 특정 시기의 내시라고 한다면 그 시기의 정치, 사회적 상황과 의학기술에 영향을 받음 ⇨ 조선시대의 전체 내시를 대표한다고 보기 어려움

○ 1936년 미국대통령 선거

- 공화당의 랜던과 민주당의 루즈벨트
- 'Literary Digest'는 **전화기 및 자동차 보유자 대상**으로 엽서를 보내 회송된 236만여 명의 의견을 분석한 결과 랜던 57%, 루즈벨트 43%
- Gallup은 수천 명의 표본조사를 토대로 루즈벨트 56%, 랜던 44%
- 선거결과에서 루즈벨트 63%, 랜던 37%
- Gallup이 루즈벨트의 당선을 예측했지만 예측한 득표율과 실제 득표율 간에 차이가 현재의 조사 결과들에 비해 큼

숙명여자대학교 SOOKMYUNG WOMEN'S UNIVERSITY

○ 확률추출(probability sampling)

◦ 연구목적에 필요한 자료와 정보를 여건이나 상황, 정확성 등을 고려하여 표본 수집 ⇨ 표본론, 실험계획법

◦ 모집단을 대표 할 수 있는 표본은 네이만이 제안한 **확률표본추출법**을 기반으로 얻을 수 있음

- 단순확률추출
- 계통표본추출
- 층화확률추출
- 집락표본추출

- 어떤 표본이 선택되는가에 따라 결과에 차이가 발생 ⇨ 변동성이 있는 표본의 정보를 이용하여 전체 모집단의 특성을 완벽하게 파악하는 것은 불가능
- 제한된 표본의 정보에 확률을 이용하여 모집단의 특성에 대해 추론 ⇦ **통계적 추론(statistical inference)**
- 확률추출방법에 의해 얻어진 표본을 이용하여 모집단에 대한 통계적 추론이 가능
- 많은 경우의 통계적 추론에서는 확률추출이 아닌 방법을 통해 얻어진 표본 또는 자료를 이용하는데 이런 추론은 모집단에 대해 심각하게 왜곡된 결론을 도출할 수 있음

통계학이란

- 관심 또는 연구의 대상이 되는 모집단의 특성을 파악하기 위해
- 모집단부터 일부의 자료(표본)를 수집하고
- 수집된 표본을 정리, 요약, 분석하여 표본의 특성을 파악한 후
- 표본의 특성을 이용하여 모집단의 특성에 대해 추론하는 원리와 방법을 제공하는 학문

기술통계

Descriptive Statistics

■ 자료의 종류와 구조

○ 통계분석 방법은 자료의 속성과 분석 목적에 따라 달라짐

○ 분석하고자 하는 자료가 분석방법에서 가정한 조건을 얼마나 만족하는지에 따라 분석방법의 적절성이 결정

⇨ 자료의 속성에 따른 분류 필요

◉ 임의로 선택된 아이돌 그룹 멤버들의 프로필 자료

변수(변량)

관측개체	번호	성별	연령	신장	체중	비만도	혈액형	멤버수
	001	남	25	181	65	정상	B	5
	002	남	23	175	55	저체중	A	4
	003	여	19	161	44	저체중	A	4
	004	남	23	178	67	정상	A	6
	005	여	21	165	45	저체중	A	5
	006	여	20	165	47	저체중	O	4
	007	남	23	170	58	정상	A	6
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

○ **변수(variable)** 또는 변량(variate)

- 일변량 자료(univariate data): 하나의 변수만 있는 자료
- 다변량 자료(multivariate data) : 여러 개의 변수로 이루어짐

○ **관측개체(observation)**

□ 자료의 분류

【그림 2.1】 속성에 따른 자료의 분류

○ 범주형 자료(categorial data)

◦ 명목형 자료(nominal data)

- 숫자로 바꾸어도 그 값이 크고 작음을 나타내는 것이 아니라 단순히 범주를 표시
- 예) 성별(주민번호), 혈액형

◦ 순서형 자료(ordinal data)

- 범주의 순서가 상대적으로 비교 가능
- 예) 비만도(저체중, 정상, 과체중, 비만, 고도비만), 학점, 선호도
- 대부분 수치형 자료를 그룹화 하여 순서형 자료로 바꿈

◉ 아이돌 그룹 프로필 자료

- BMI지수는 체중(Kg)을 신장(m)의 제곱으로 나눈 값(Kg/m^2)
- 번호 001에서 007까지의 BMI

  (19.8, 17.6, 17.0, 21.1, 16.5, 17.3, 19.9)
- 20대 이하의 경우 비만도 평가 기준

BMI	비만도 판정
18미만	저체중
18이상 ~ 23미만	정상(표준)
23이상 ~ 25미만	과체중
25이상 ~ 30미만	비만
30이상	고도비만

○ 관측개체의 비만도는

(정상, 저체중, 저체중, 정상, 저체중, 저체중, 정상)

○ 수치형 자료(numerical data)

- ○ **이산자료(discrete data)**
  - 셀 수 있는 형태의 자료(countable data)
  - 예) 멤버의 수
- ○ **연속자료(continuous data)**
  - 연속적인 속성을 가지는 자료
  - 예) 신장, 체중
  - 연속자료는 이산화를 통해 자연수 형태로 표시되는 경우가 많음

- 변수는 해당하는 자료의 형태에 따라 크게 범주형 변수와 수치형 변수로 나뉘고 세부적으로 명목형 변수, 순서형 변수, 이산변수, 연속변수로 나뉨
- 자료를 측정하는 척도에 따라 명목척도, 순서척도, 구간척도, 비율척도로 나누기도 함

■ 표를 이용한 자료정리

□ 도수분포표(Frequency Table)

- ○ 일변량 범주형자료를 정리하는데 있어 기본이 되는 표
- ○ **도수(frequency)**는 임의의 범주에 속한 관측개체의 수
- ○ 각각의 범주에 몇 개의 관측개체가 있는지를 정리한 표
- ○ 각 범주의 도수가 상대적으로 얼마나 많이 있는지 비교할 수 있도록 전체 관측개체에서 해당 범주의 도수가 차지하는 비율인 **상대도수(relative frequency)**를 추가

$$\text{상대도수} = \frac{\text{해당 범주 관측개체의 수}}{\text{전체 관측개체의 수}}$$

- 상대도수에 100을 곱해 %로 표시하기도 함

◉ 파이가게에서 지난 1주일 동안 판매된 파이의 종류와 도수

【표 2.1】파이 판매량과 비율

파이종류	판매량	판매비율(%)
애플	59	25.2
딸기	52	22.2
블루베리	47	20.1
초코	32	13.7
고구마	27	11.5
바나나	17	7.3
합계	234	100.0

○ 이 표에서는 판매량이 큰 순으로 정렬하여 표시하였는데 때로는 범주를 가나다순으로 정렬하여 표시하기도 함

○ 수치형 자료에 대한 도수분포표

- ◦ 나올 수 있는 값이 몇 개로 한정되어 있는 경우, 해당 숫자와 일치하는 자료의 개수로 도수분포표를 작성
- ◦ 수치형 자료를 포함한 순서형 자료에 대한 도수분포표에는 범주의 순서에 따라 정렬하며 작은 값으로부터 도수나 상대도수를 누적시킨 값을 추가하여 사용

◉ 1889년 영국 Saxony Geissler 지역의 병원기록으로 12명의 아이를 가진 6115 가구를 대상으로 조사한 자료

【표 2.2】 Saxony Geissler 지역의 아들 수 분포

아들 수	도수	상대도수(%)	누적상대도수(%)
0	3	0.05	0.05
1	24	0.39	0.44
2	104	1.70	2.14
3	286	4.68	6.82
4	670	10.96	17.78
5	1,033	16.89	34.67
6	1,343	21.96	56.63
7	1,112	18.18	74.82
8	829	13.56	88.37
9	478	7.82	96.19
10	181	2.96	99.15
11	45	0.74	99.89
12	7	0.11	100.00
합계	6,115	100.00	100.00

- 모두 딸이거나 아들인 가구는 10(0.16%)가구
- 딸과 아들의 수가 같은 가구는 가장 많은 1,343(21.96%)가구
- 딸의 수가 많은 가구는 2120(34.67%)
- 아들의 수가 많은 가구는 2518(100%-56.63% = 43.37%)로 아들이 많은 가족이 딸이 많은 가구보다 많음

- 다양한 값으로 구성되어 있는 수치형자료는 관측된 값들을 몇 개의 구간으로 그룹화 하여 해당 그룹에 속한 관측개체의 개수로 도수분포표를 만듦
  - 그룹을 **계급(class)**이라고 함
- 수치형 자료를 그룹화 할 땐 **계급의 수와 경계값**을 정함
  - 자료의 특성을 고려해 작성자가 임의로 결정할 수 있음
  - 수치형 자료의 그룹화는 히스토그램과 관련되어 있음
    ⇨ 히스토그램을 작성하는데 사용되는 방법을 소개
- 계급의 수는 자료의 수에 비례
  - n: 전체 자료의 수
  - k: 그룹의 수

제곱근 방법	$k = \lceil \sqrt{n} \rceil$
Sturges 공식	$k = \lceil \log_2 n + 1 \rceil$
Rice 공식	$k = \lceil 2n^{1/3} \rceil$

- $\lceil x \rceil$는 x보다 크거나 같은 값 중에 가장 작은 정수
- 적절한 k는 자료의 개수나 형태에 영향을 받기 때문에 어떤 방법이 더 좋다고 할 수 없음

○ 계급의 경계값

- 동일간격으로 나눌 수 있으나 그룹의 간격이나 폭을 설명하기 편한 값 또는 자료의 구조에 맞춰 그룹화를 선택적으로 하는 것이 좋음

- 예) 최소값이 7이고 최대값이 34이고 3개의 계급으로 나누고자 할 때 [7, 16), [16, 25), [25, 34]로 할 수 있으나 [5, 15), [15, 25), [25,35)으로 하면 설명이 보다 용이할 수 있음
- 예) 연간 소득에 대한 분석을 하고자 할 때 대부분의 자료가 1억 원 이하이므로 동일한 간격으로 나누는 것보다 1천만미만, [1천만, 2천만), [2천만, 3천만), [3천만, 5천만), [5천만, 7천만), [7천만, 1억), [1억, 1억5천만), 1억5천 이상과 같이 자료의 수가 많은 구간을 세분화 하여 나누는 것이 소득의 분포를 파악하는데 더 적절함

◉ 대학정보공시 취업률자료

○ 2011년 통계학 관련 42개 학과의 취업률

○ 2011년 취업률은 2010년 8월, 2011년 2월 졸업자 중 취업대상 중에서 2011년 6월 1일에 건강보험DB연계취업자와 해외취업자로 등록된 사람의 비율

55.6	83.3	43.4	58.1	31.6	55.6	60.7	64.6	73.3	55.6	64.3
52.8	22.7	46.3	71.4	53.8	64.5	67.9	71.4	80.0	59.5	40.5
77.1	58.6	65.4	52.4	66.7	91.3	41.3	72.1	61.9	78.4	63.6
41.0	65.2	81.3	54.8	19.6	50.0	53.1	41.2	56.5		

○ 우선 적정한 구간으로 자료를 그룹화

- 최소취업률은 19.6%, 최대취업률은 91.3%
- 최소점을 10%, 최대점을 100%, 계급 길이를 10%
- 도수가 적은 계급은 통합

【표 2.3】2011년 통계학 관련학과 취업률

취업률	도수	상대도수	누적상대도수
10%이상~40%미만	3	0.071	0.071
40%이상~50%미만	6	0.143	0.214
50%이상~60%미만	13	0.310	0.524
60%이상~70%미만	10	0.238	0.762
70%이상~80%미만	6	0.143	0.905
80%이상~100%	4	0.095	1.000
합계	42	1.000	1.000

- 취업률이 50%이상~60%미만 사이에 있는 학과가 31.0%인 13개로 가장 많고 60%이상~70%미만인 학과가 두 번째로 많음
- 80%이상의 취업률을 보인 학과가 4개인 반면 40%미만인 학과가 3개
- 누적상대도수를 이용하면 취업률이 60% 미만인 학과는 전체 42개 학교 중 52.4%임

□ 분할표

○ 도수분포표는 일변량 자료를 정리한 표

○ 두 개 이상의 변수를 동시에 고려하여 관측개체의 빈도를 정리할 필요가 있음

- 예) 성별과 혈액형 간에 관계

성별	혈액형			
	A	B	AB	O
남자	Ⓐ			
여자				

- 칸(cell) : 각 범주에 교차되는 부분

○ **분할표(contingency table), 교차표(cross tabulation)** : 2개 이상의 변수에 대해 교차시켜 빈도를 표시한 표

- 행과 열의 수를 같이 표시(2×4 분할표)

○ 비율을 표시할 때에는 분석 목적 또는 자료가 어떻게 수집되었는지에 따라 다르게 표시될 수 있음

◉ 세 가지 스마트폰모델에 대한 남녀별로 선호도 비교

- ○ 남자 76명과 여자 70명을 대상으로 세 가지 모델 중 가장 마음에 드는 모델을 선택
- ○ 남자 중 모델 A는 35, B는 23, C는 18명이 선택하고 여자 중 A는 17명, B는 33, C는 20명이 선택

【표 2.4】 성별에 따른 스마트폰모델 선호도 결과

성별	스마트폰모델			합계
	A	B	C	
남자	35	23	18	76
여자	17	33	20	70
합계	52	56	38	146

○ 분석목적은 선호도에서 남녀 간 차이여부 ⇨ 남자 중 각각의 모델을 선호한 비율과 여자 중 각각의 모델을 선호한 비율을 따로 표시

【표 2.5】 성별에 따른 스마트폰모델 선호도 결과 (비율)

성별	스마트폰모델			합계
	A	B	C	
남자	35 (46.0%)	23 (30.3%)	18 (23.7%)	76 (100%)
여자	17 (24.3%)	33 (47.1%)	20 (28.6%)	70 (100%)
합계	52 (35.6%)	56 (38.4%)	38 (26.0%)	146 (100%)

◉ 코골이 정도와 심장질환의 연관성

◦ Norton과 Dunn (1985)는 2484명을 대상으로 코를 얼마나 고는가와 심장질환을 앓고 있는지를 조사

【표 2.6】 코골이 정도와 심장질환 조사

심장질환	코골이 정도				합계
	골지 않음	가끔 골음	거의 매일	매일	
예	24	35	21	30	110
아니오	1355	603	192	224	2374
합계	1379	638	213	254	2484

○ 조사대상 2484명 중 심장질환을 앓고 있는 사람은 4.4%인 110명이고 아닌 사람은 95.6%인 2374명

○ 단순히 조사대상 2484명을 조사하여 나온 결과이므로 각 칸의 비율은 2484명을 기준으로 계산

【표 2.7】 코골이 정도와 심장질환 관계 조사 (비율)

심장질환	코골이 정도				합계
	골지 않음	가끔 골음	거의 매일	매일	
예	24 (1.0%)	35 (1.4%)	21 (0.8%)	30 (1.2%)	110 (4.4%)
아니오	1355 (54.6%)	603 (24.3%)	192 (7.7%)	224 (9.0%)	2374 (95.6%)
합계	1379 (55.6%)	638 (25.7%)	213 (8.5%)	254 (10.2%)	2484 (100.0%)

○ k차원 분할표(k-dimensional contingency table)

◦ 3개 이상의 범주형 변수에 대해 분할표

- 변수의 개수가 k라고 하면 k차원 분할표 또는 k원 분할표(k-way contingency table)

◦ 타이타닉(RMS Titanic)호 승객과 승무원 생존자와 사망자수

- 1912년 4월 15일 북대서양에서 침몰
- 당시 승선한 사람의 이름이 중복 기재되거나 출항 직전에 취소한 사람들이 있어 구조기록에 대한 정확한 통계는 알 수 없음

⇨ 위키피디아(http://en.wikipedia.org/wiki/RMS_Titanic)

- 그룹(성인남녀와 어린이), 생존여부, 등급(객실등급과 승무원)으로 분류하여 교차 정리한 3원 분할표

【표 2.8】 타이타닉호 생존자와 사망자 수

그룹		남자		여자		어린이		전체		
생존여부		생존	사망	생존	사망	생존	사망	생존	사망	합
등급	1등실	57	118	140	4	5	1	202	123	325
	2등실	14	154	80	13	24	0	118	167	285
	3등실	75	387	76	89	27	52	178	528	706
	승무원	192	693	20	3	-	-	212	696	908
전체		338	1352	316	109	56	53	710	1514	2224

■ 그래프를 이용한 자료정리

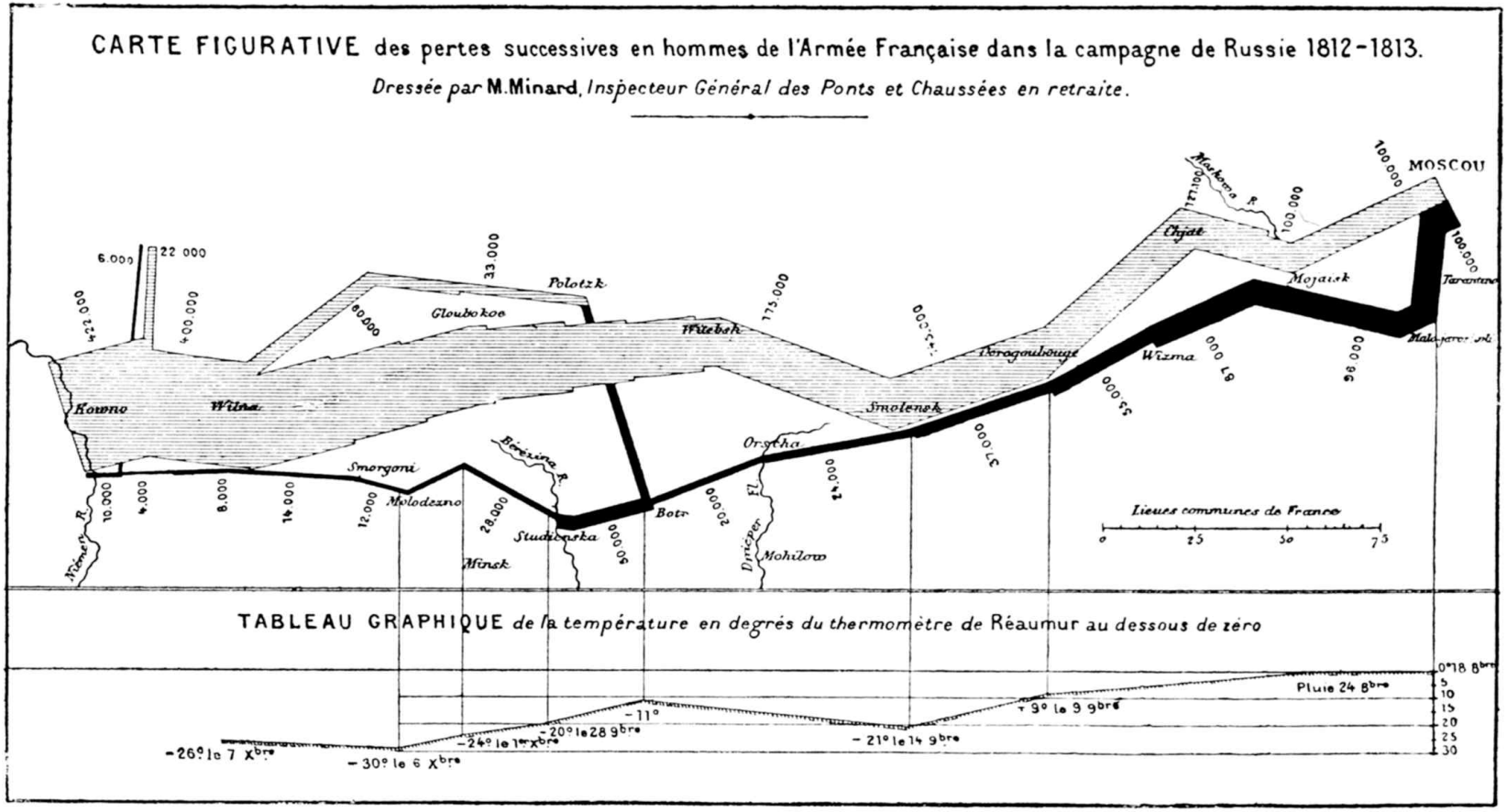
○ 대부분의 사람들은 숫자나 수식에 대해 두려움을 가짐

⇨ 어떤 현상을 숫자나 수식으로 설명하는 것보다 그림과 같은 시각적인 방법을 이용하여 설명하면 이해를 잘 하는 경향이 있음

○ 프레이페어(W. Playfair, 1759~1823)

- 1786년: 경제자료에 대한 선그래프(line graph)와 막대차트
- 1801년: 파이차트와 원그래프(circle graph)를 이용

◉ 나폴레옹의 러시아원정



【그림 2.2】 나폴레옹 군대의 진군과 후퇴

○ 1812년 6월부터 1813년 1월까지 원정 과정에서의 병력과 후퇴할 때 해당지역의 기온을 표시한 1861년 그림

- 진군할 때는 엷은 색으로, 후퇴할 때는 진한 색으로 표시
- 해당 지역에서 생존했던 병사의 수를 선의 두께로 표시
- 원정초기 442,000여명의 병력이 출정하였으나 모스코바에 도착한 병사는 100,000여명이었고 끝까지 생존해 돌아온 병사는 10,000여명
- 후퇴하는 동안의 기온의 변화를 같이 볼 수 있는데 최저 기온이 -30℃까지 떨어짐

○ 톨스토이는 '전쟁과 평화'를 통해 1000여 쪽에 걸쳐 기술

□ 파이차트(Pie chart)

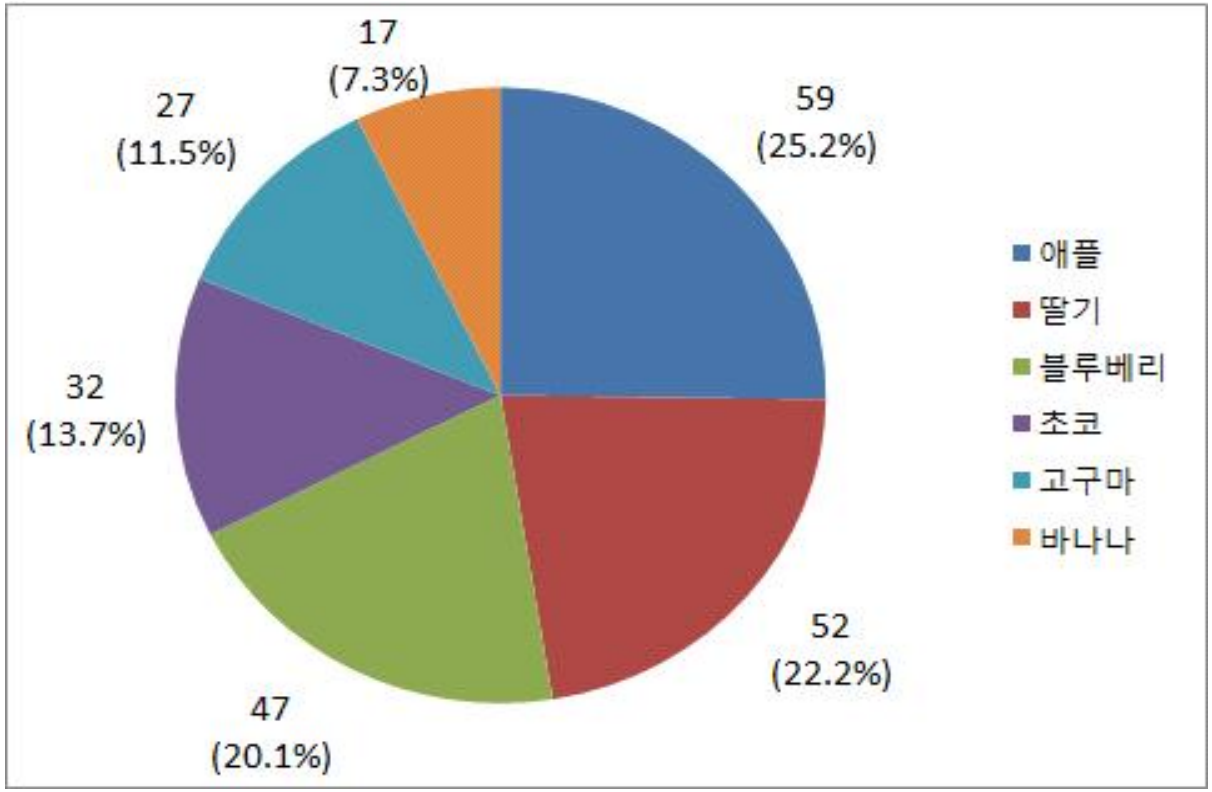
- ○ 원을 먼저 그리고 원점을 기준으로 각 범주에 해당되는 비율만큼 각도를 분할하여 표시
- ○ 해당 범주의 각도 = 비율 × 360°
- ○ 원을 사용하는 이유는 각 범주의 각도와 면적의 비가 항상 동일하기 때문

◉ 파이판매량

○ 비율에 대해 파이차트의 각도를 계산

【표 2.9】파이차트 각도

파이종류	각도	파이종류	각도	파이종류	각도
애플	90.8	딸기	80.0	블루베리	72.3
초코	49.2	고구마	41.5	바나나	26.2



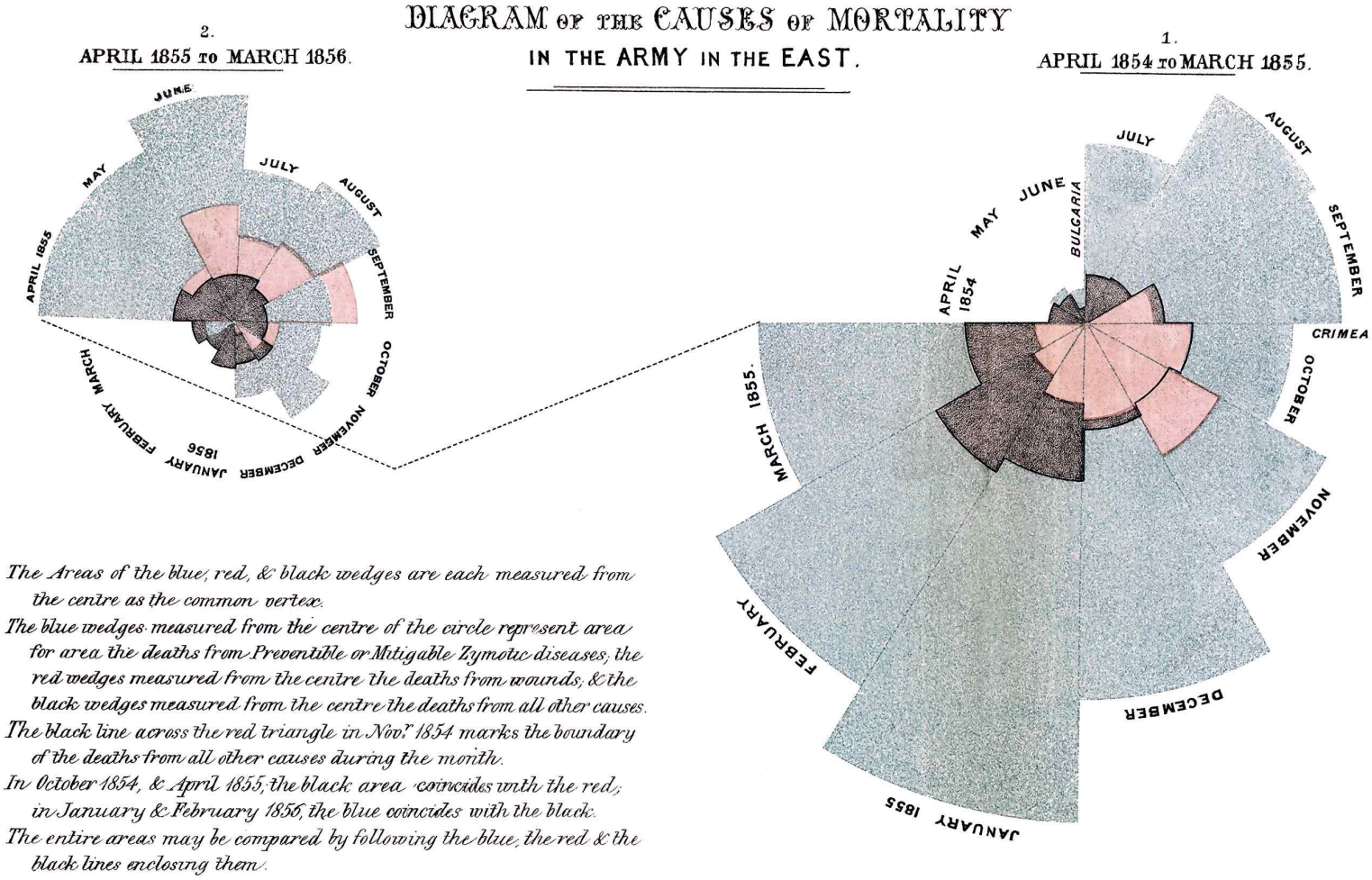
【그림 2.3】파이판매량에 대한 파이차트

○ 애플파이와 딸기파이의 각도가 상대적으로 큼

○ 바나나파이의 각도가 다른 것에 비해 작음

○ 나이팅게일 로즈 다이어그램(Nightingale rose diagram)

- ○ 나이팅게일(F. Nightingale, 1820~1910)
  - 1859년 여성최초로 영국 왕립통계학회 회원
  - 미국통계학회 명예회원
- ○ 1954년 4월부터 1856년 3월까지 크림전쟁기간 중 이스트지역 부대에서 사망한 사병들의 사인을 부상, 질병, 기타원인으로 분류하여 정리
- ○ 각 원인별로 사망한 병사들의 수를 면적으로 표시
- ○ 1954년 7월부터 1955년 8월까지 사망한 병사의 압도적인 사인은 부상이 아닌 장티푸스, 콜레라, 이질과 같은 질병에 의한 것 ⇨ 병원의 위생을 개선하게 한 근거자료로 사용



【그림 2.4】 크림전쟁에서 이스티지역 부대의 원인별 사망자 수

□ 막대그래프(Bar chart)

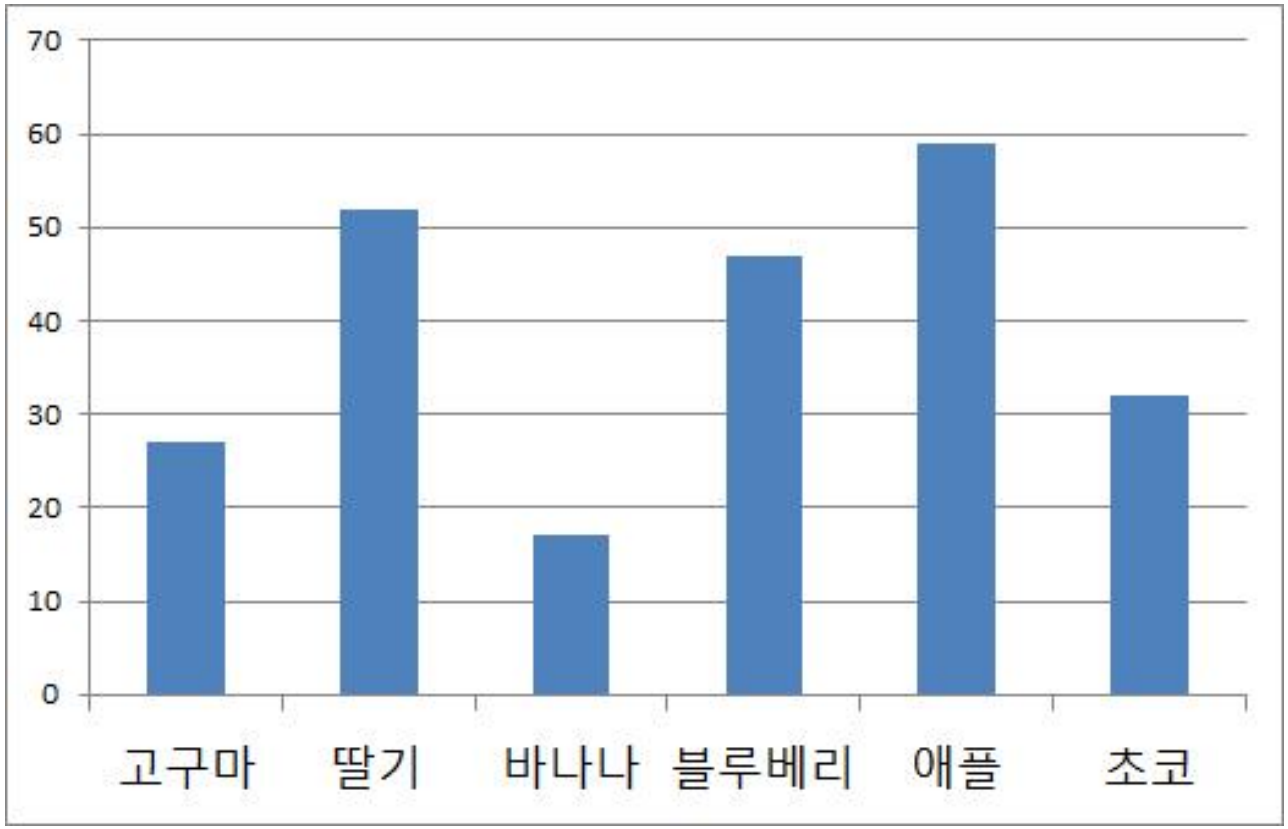
◦ Cleveland(1985)

- 동일한 척도에서의 위치, 길이, 각도와 기울기, 면적, 부피, 색상과 밀도 순으로 인지 ⇨ 파이차트에서는 비슷한 비율을 가지는 범주들 간의 비교가 쉽지 않음
- 예) 딸기파이와 블루베리파이의 판매량의 비교?

◦ 각 범주의 빈도를 비교하고자 한다면 가능한 동일한 척도에서의 위치나 길이로 표시

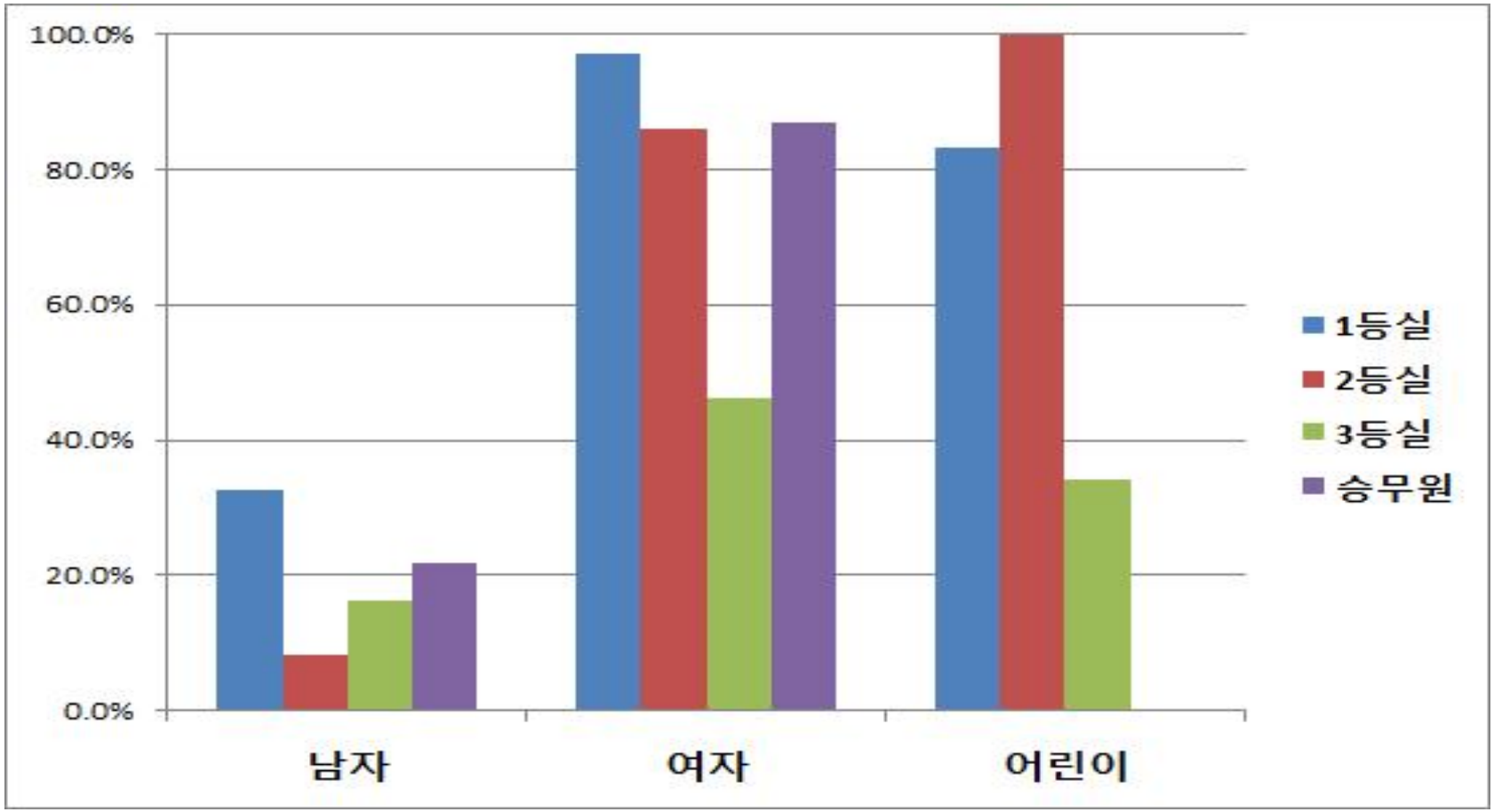
◦ 막대그래프는 각 범주의 도수나 상대도수를 막대의 길이로 표시한 그림

◉ 파이판매량



【그림 2.5】 파이판매량의 막대그래프

◉ 타이타닉호 생존자의 비율



【그림 2.6】 타이타닉호 승객 및 승무원 생존율 비교

- 남자의 경우 전반적으로 사망자가 생존자보다 많았으며 특히 2등실 승객의 생존율이 낮음
- 여자의 경우 3등실 승객의 생존율이 상대적으로 매우 낮았으며 3등실 어린이의 생존율도 다른 등실의 어린이보다 낮음
- 승무원의 생존율이 1등실을 제외한 나머지 등실의 승객보다 생존율이 높음
- 남녀, 어린이 구분하지 않은 상태에서 생존율을 비교하면 1등실 62.2%, 2등실 41.2%, 3등실 25.2%, 승무원 23.3%로 승무원의 생존율이 가장 낮은 것을 집계되어 남녀 및 어린이로 나누었을 때와 다른 결과를 보여줌

◉ 인구추계교실

- ○ 통계청 국가통계포털 사이트인 KOSIS(http://kosis.kr)
- ○ 출생, 사망, 국제이동 등과 같은 인구변동요인을 고려하여 우리나라 미래 인구에 대한 표와 그래프로 제시
- ○ 2010년 현재 40~50세 연령의 인구가 많음
- ○ 2040년 인구구조의 특징
  - 2010년 40세였던 70세 근처의 연령인구가 가장 많고 80세 이상 인구도 현재보다는 현저하게 많아짐
  - 출산연령대가 되는 현재 10세 이하의 인구감소로 2040년 5세 이하의 어린이가 다시 감소하는 추세를 보일 것으로 예상

【그림 2.7】 우리나라 2010년과 2040년(예상) 인구추계비교

□ 히스토그램(Histogram)

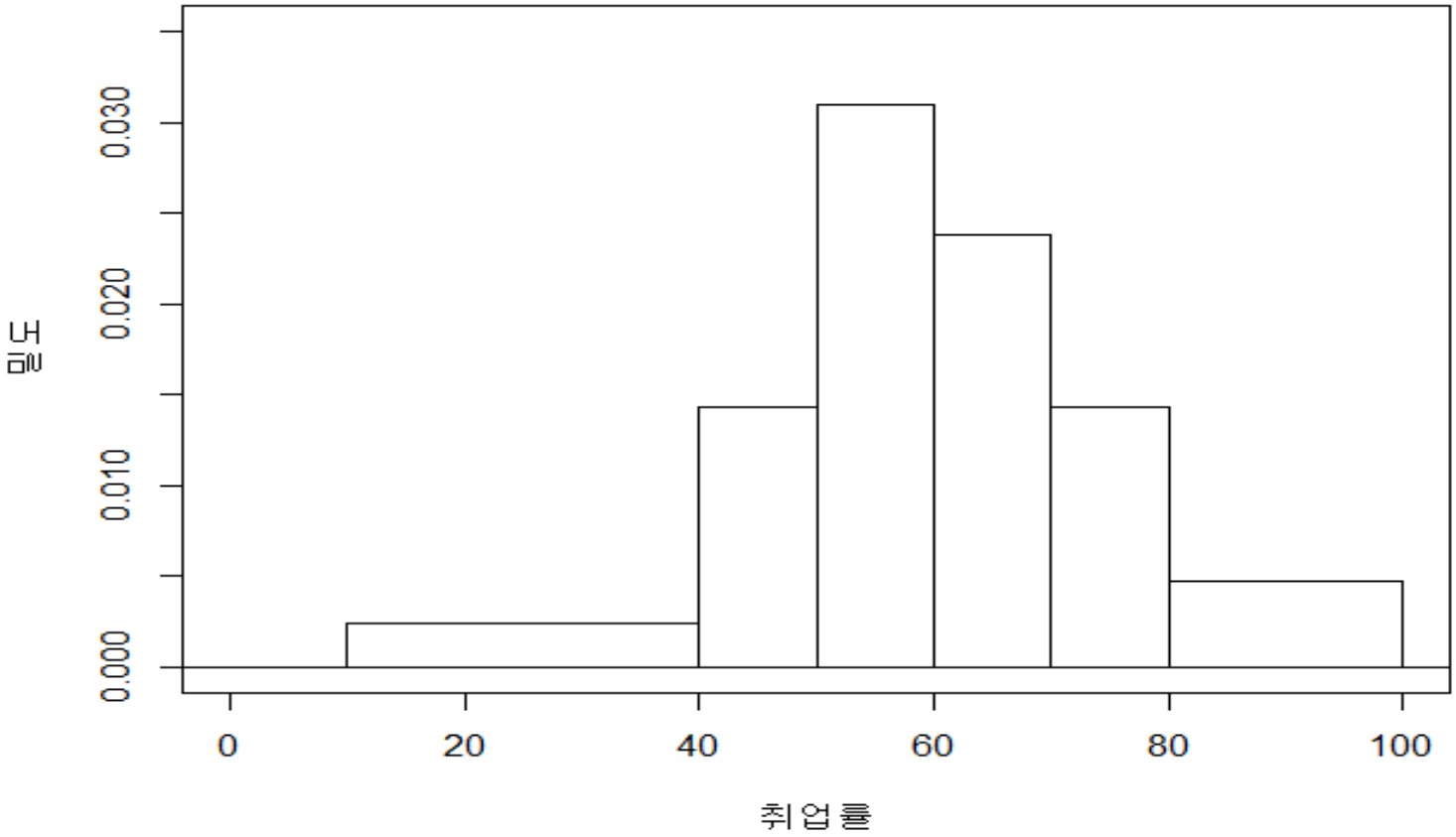
- ○ 통계학에서는 히스토그램과 막대그래프를 엄격히 구분
- ○ 히스토그램(histogram)은 수치형 자료 특히 연속형 자료의 분포형태를 표시
- ○ 해당 구간의 상대도수를 직사각형의 면적으로 표시하여 **전체 면적이 1**

$$\text{높이} = \frac{\text{상대도수}}{\text{계급폭}}$$

- ○ 이 높이는 해당 구간에 자료들이 얼마나 모여 있는지를 나타내는 측도로 **밀도(density)**라고 함

◉ 통계학 관련 학과 취업률

구간	밀도	구간	밀도
10%이상~40%미만	0.0024	60%이상~70%미만	0.0238
40%이상~50%미만	0.0143	70%이상~80%미만	0.0143
50%이상~60%미만	0.0310	80%이상~100%	0.0047



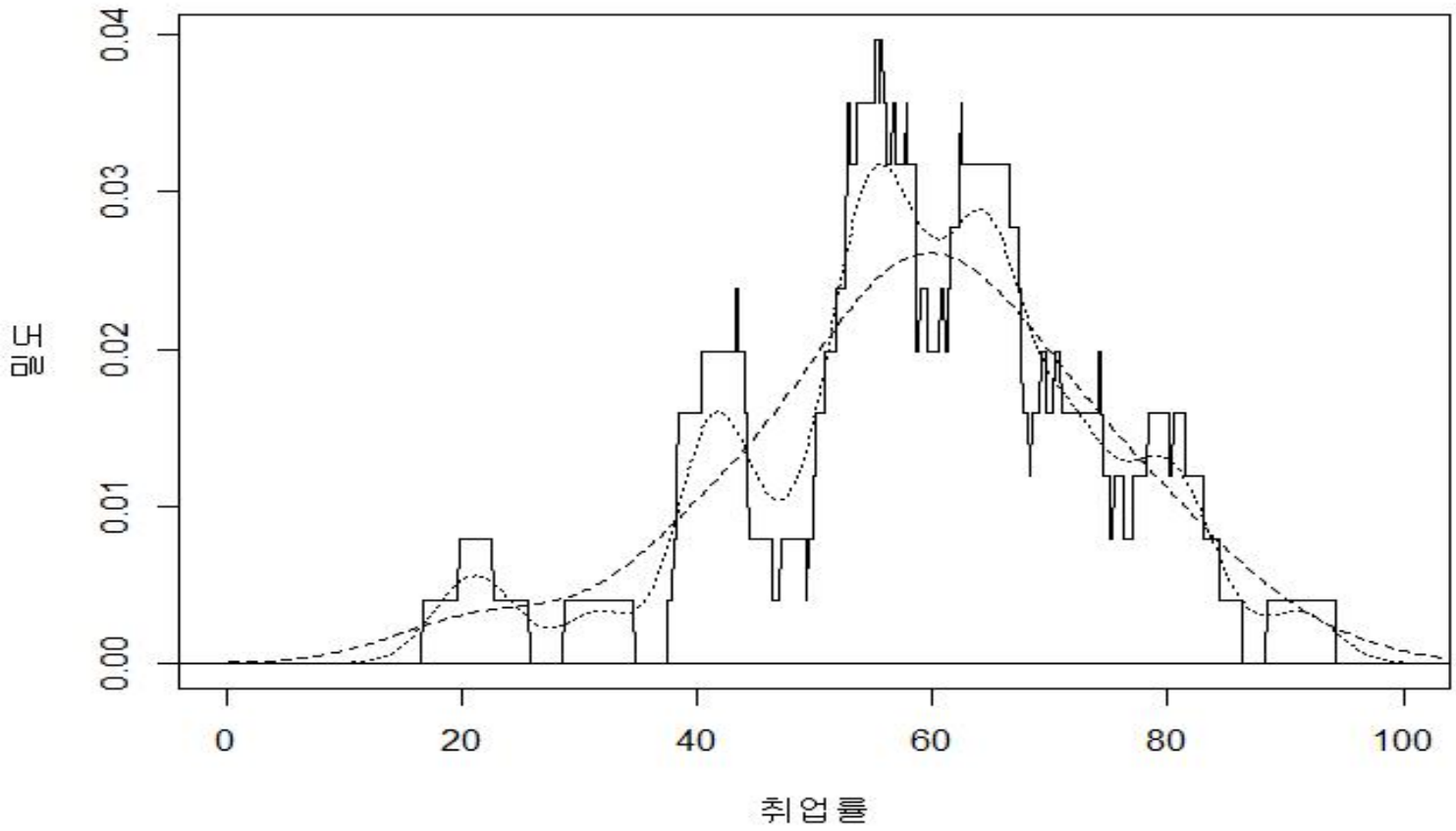
【그림 2.8】 통계학 관련전공 취업률 히스토그램

○ 50%에서 60%까지 취업률을 보인 학과가 가장 많음

○ 대부분이 40%에서 80%에 있음

○ 구간을 어떻게 정하는가에 따라 모양이 조금씩 달라짐

- 첫 번째 취업률 구간 10%이상~40%미만을 15%이상~40%미만으로 바꾸면 도수분포표는 변화가 없으나 히스토그램에서 해당 구간의 길이가 30에서 25로 줄어들어 높이는 0.0024에서 0.0029로 높아짐



【그림 2.9】취업률에 대한 밀도함수추정

□ 줄기-잎 그림(stem-and-leaf plot)

- ○ 히스토그램은 계급을 어떻게 정하는가에 따라 형태가 조금씩 달라질 수 있고 그룹화 된 자료들의 상대도수만을 사용하기 때문에 개개의 관측값이 얼마인지는 알 수 없음
- ○ 줄기-잎 그림은 관측값의 정보를 그대로 간직하면서 자료가 어떻게 분포되어 있는지를 알려주는 그림
- ○ 자료를 순서대로 정렬한 후 줄기에는 기본단위의 10배의 값을 표시하고 잎에는 관측값의 기본단위에 해당되는 값을 표시

◉ 통계학 관련 학과 취업률

○ 자료를 반올림하여 정수형태로 표시

20	23	32	40	41	41	41	43	46	50	52	53	53	54
55	56	56	56	56	58	59	60	61	62	64	64	64	65
65	65	67	68	71	71	72	73	77	78	80	81	83	91

○ 취업률의 단위는 1이고 20에서 91까지 분포되어 있으므로 줄기에는 2부터 9까지 표시하고 각 줄기 왼쪽에 해당하는 단위 값을 오름차순으로 잎을 표시

2 | 03
3 | 2
4 | 111136
5 | 023345666789
6 | 01244555578
7 | 112376
8 | 013
9 | 1

○ 자료가 많은 경우에는 10 단위를 2개로 나눠 잎에 위쪽은 0~4, 아래쪽은 5~9로 표시하거나 5개로 나눠 (0,1), (2,3), (4,5), (6,7), (8,9)로 표시할 수도 있음

□ 산점도(scatter plot)

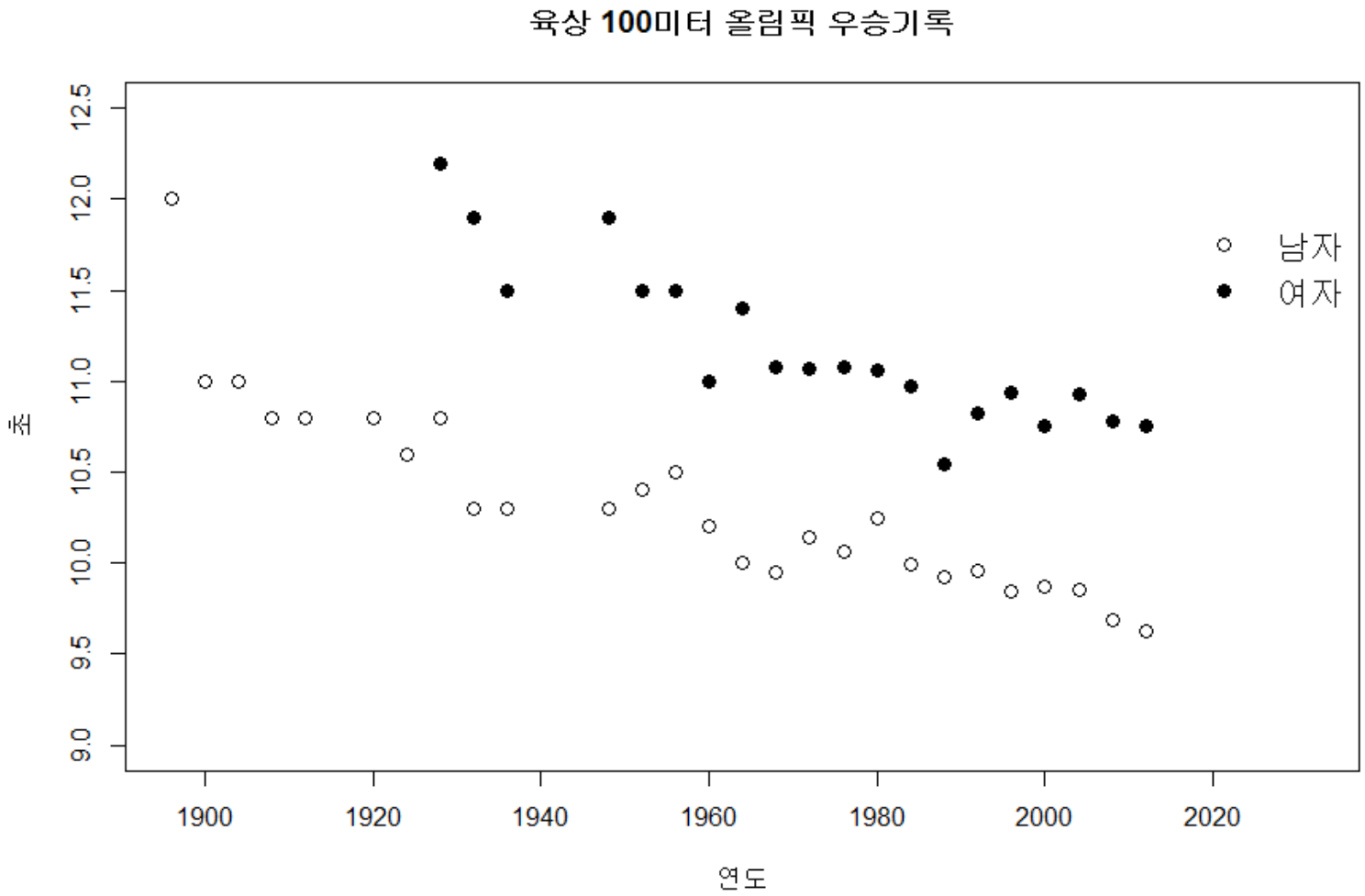
- ○ 다변량 자료에 대한 분석에서 주요 문제 중 하나는 수치적 변수들 간의 관계를 유도
- ○ 각각의 관측개체에 대해 두 변수의 값은 순서쌍 (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)으로 표시
- ○ 산점도는 순서쌍 자료를 2차원 평면상에 점으로 표시

◉ 올림픽 100미터 우승기록

○ 1896년 아테네올림픽부터 2012년 런던올림픽까지 자료

【표 2.10】 올림픽 육상 100미터 우승기록

연도	우승기록		연도	우승기록		연도	우승기록		연도	우승기록	
	남자	여자		남자	여자		남자	여자		남자	여자
1896	12	-	1928	10.8	12.2	1964	10	11.4	1992	9.96	10.82
1900	11	-	1932	10.3	11.9	1968	9.95	11.08	1996	9.84	10.94
1904	11	-	1936	10.3	11.5	1972	10.14	11.07	2000	9.87	10.75
1908	10.8	-	1948	10.3	11.9	1976	10.06	11.08	2004	9.85	10.93
1912	10.8	-	1952	10.4	11.5	1980	10.25	11.06	2008	9.69	10.78
1920	10.8	-	1956	10.5	11.5	1984	9.99	10.97	2012	9.63	10.75
1924	10.6	-	1960	10.2	11.0	1988	9.92	10.54	2016	?	?



【그림 2.10】 올림픽 육상 100미터 우승기록

- 연도가 증가함에 따라 남녀 모두 기록이 전반적으로 빨라지는 추세
- 여성의 기록이 좀 더 빨리 감소하는 추세

○ 산점도 행렬(scatter matrix)

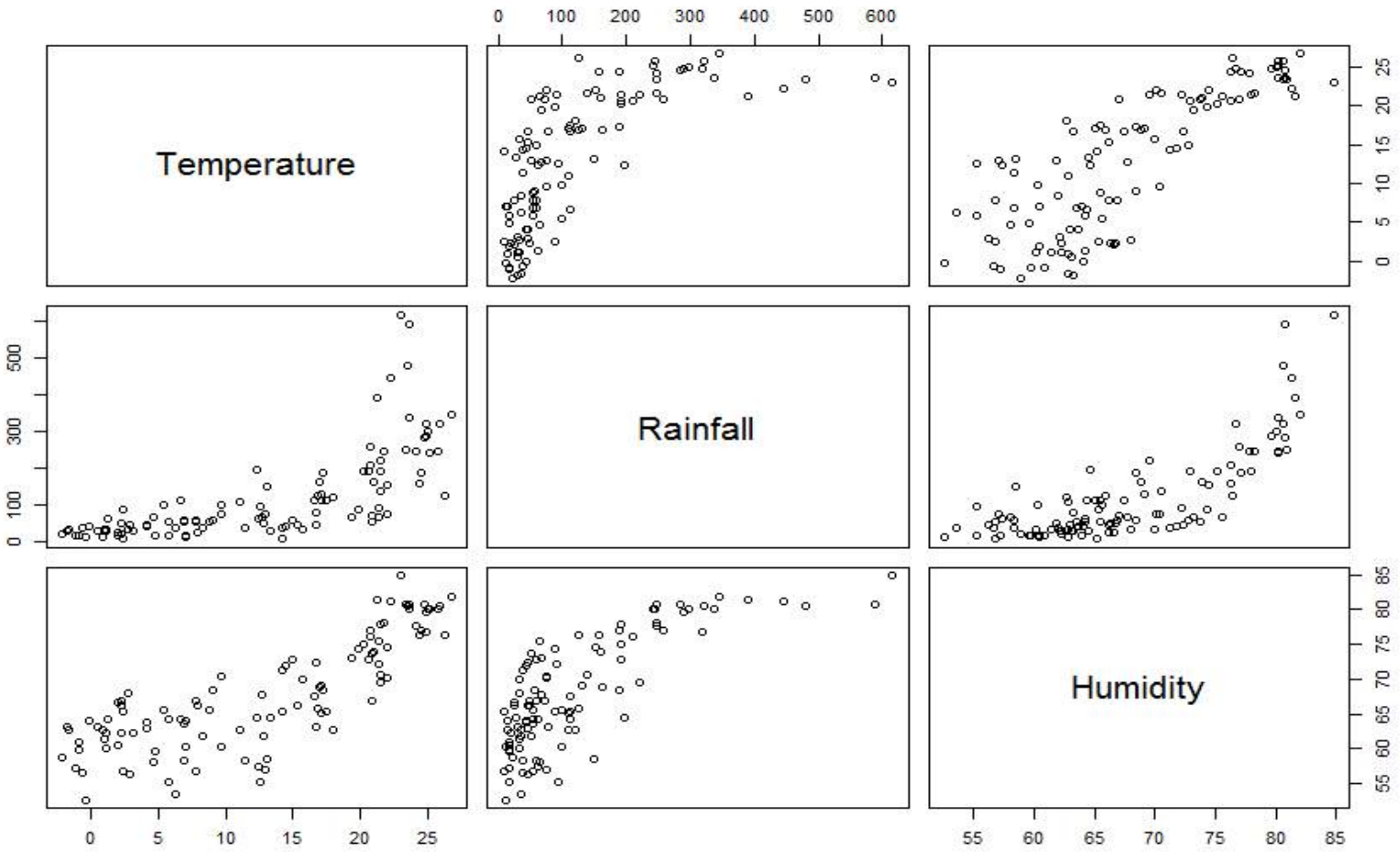
◦ 3개 이상의 수치형 변수에 대해 두 변수를 쌍으로 조합하여 산점도를 행렬형태로 표시

◉ 기상자료들 간의 관계

◦ 2002년 1월부터 2010년 12월까지 우리나라 전체의 월간 평균기온, 강수량, 평균습도

◦ 어떤 변수값이 크면 다른 변수도 값이 커지는 경향

◦ 평균기온과 평균습도는 직선, 강수량과 나머지 변수는 곡선의 형태로 증가하는 형태를 가짐



【그림 2.11】 월간 기상자료 간의 관계

○ 시계열그림(time series plot)

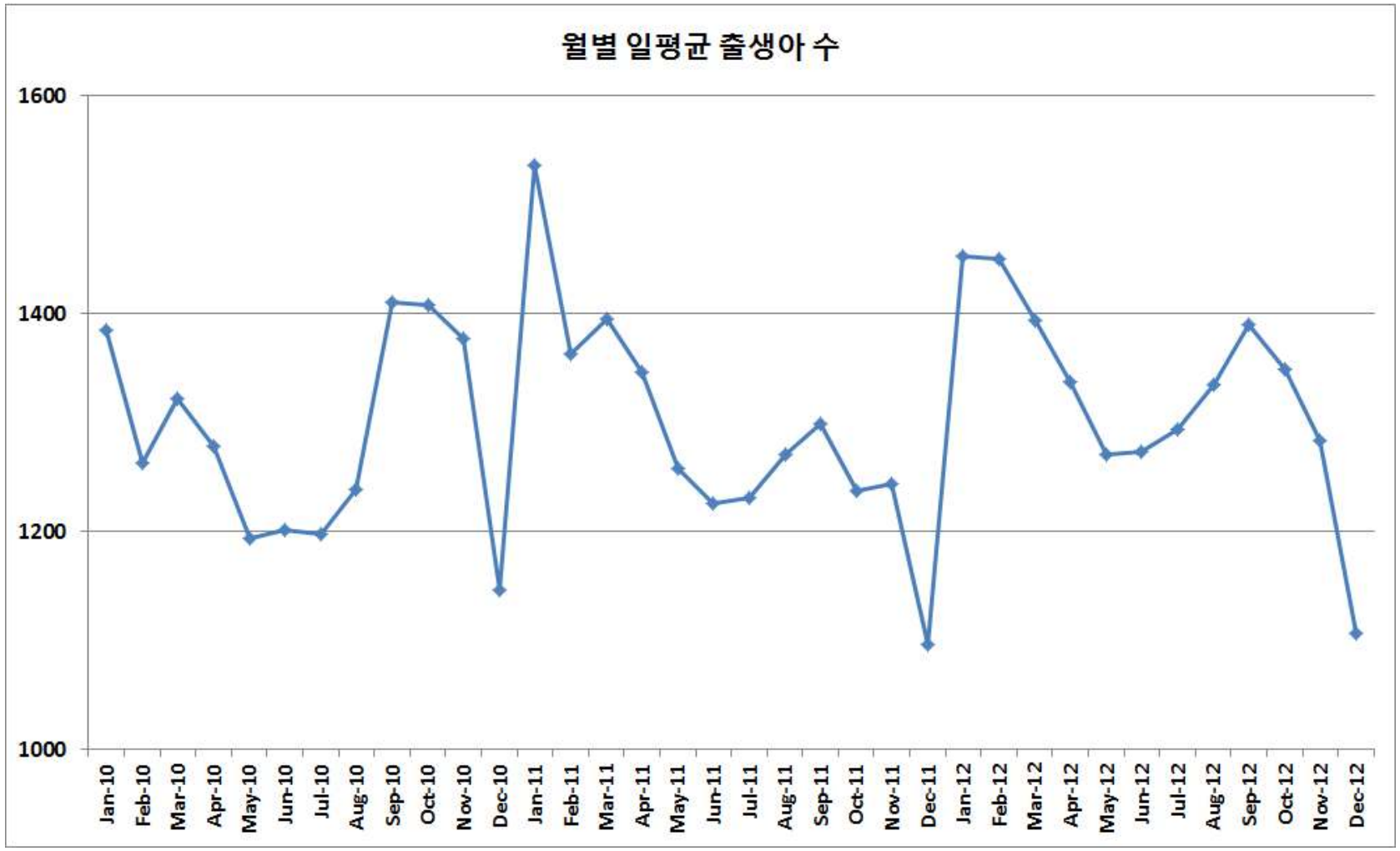
- ◦ 자료가 시간에 따라 관측된 경우 x 축에 관측시점, y 축에 관측된 값을 표시한 산점도
- ◦ 【그림 2.10】도 일종의 시계열 그림
- ◦ 시계열그림은 시간에 따라 순서적으로 자료가 얻어지기 때문에 관측값의 표시할 때 순서가 중요
- ◦ 자료가 많은 경우 점으로 표시하면 순서 파악이 어려울 수 있어 시계열 그림을 그릴 땐 일반적으로 관측값을 선으로 연결

◉ 월별 하루 평균출생아수

○ 통계청에서 발표하는『월간 인구동향』통계 중 2010년 1월부터 2012년 12월까지 자료

【표 2.11】 월별 하루 평균출생아수 (단위: 명)

월 / 연도	1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
2010년	1385	1263	1322	1278	1194	1201	1197	1238	1410	1407	1377	1146
2011년	1535	1363	1395	1346	1257	1226	1231	1270	1299	1237	1244	1096
2012년	1452	1450	1394	1337	1271	1273	1294	1335	1390	1348	1283	1106

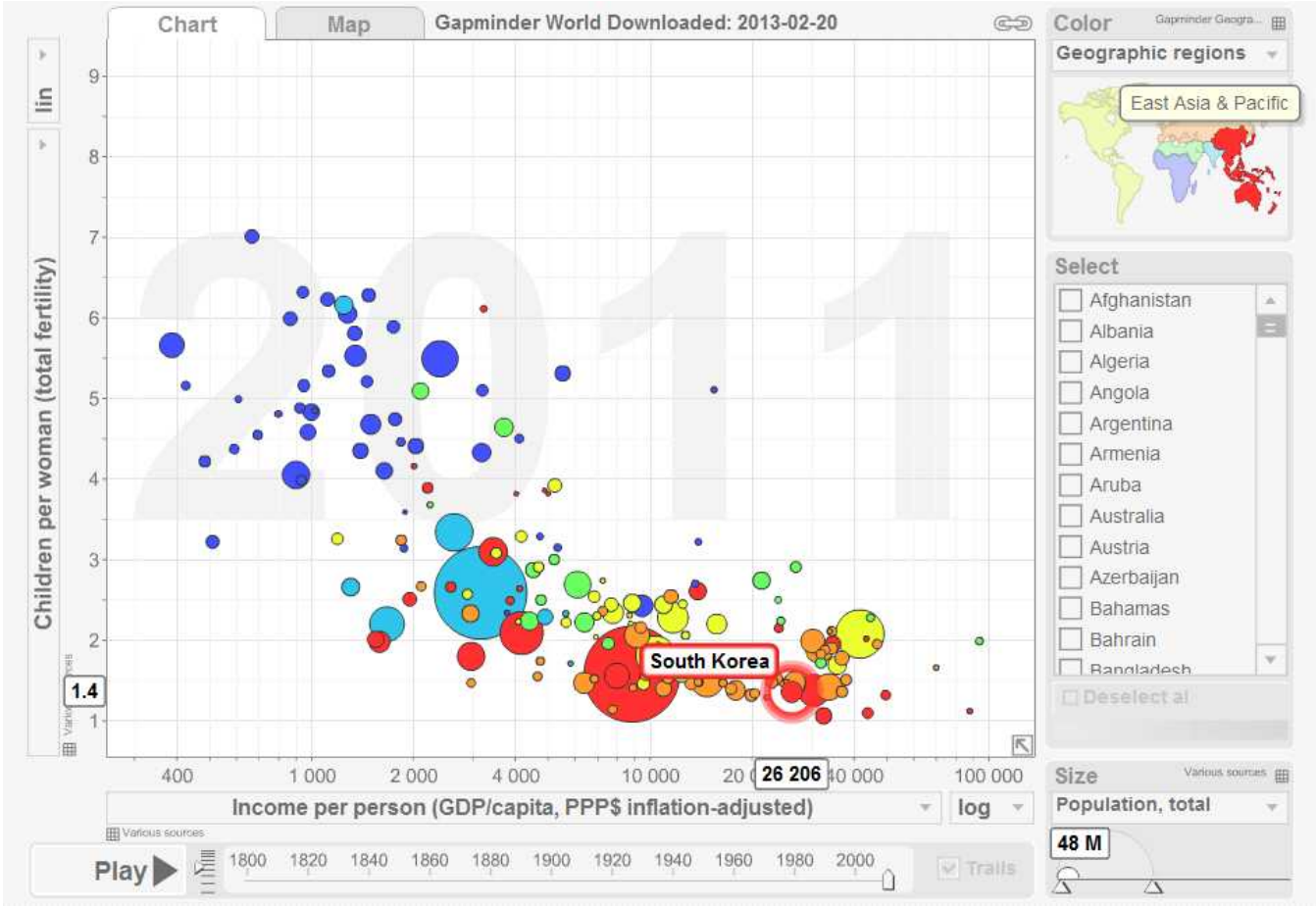


【그림 2.12】 월별 하루 평균출생아수 추이

○ 모든 연도에서 1월의 출생아 수가 상대적으로 많은 반면 12월의 수 현저히 적음 ⇦ 12월에 태어났지만 1월로 출생신고를 한 사례가 적지 않음

○ 1월, 12월을 제외한 나머지 달에서는 각 연도별로 5월부터 7월까지의 출생아수가 적음

◉ Gapminder World



【그림 2.13】 Gapminder World 프로그램

- 1인당 소득과 여성 1인당 자녀의 수의 관계
  - 소득이 높을수록 자녀의 수가 줄어드는 경향
- 해당 나라가 어느 지역에 표시되어 있는지를 색깔로 표시
- 인구의 수를 원으로 크기로 표시
- 연도에 따라 변화의 추이를 볼 수 있음
- 화면에 있는 점을 클릭하면 그 점이 어느 나라이고 해당국가의 x과 y값 뿐만 아니라 인구수와 어느 지역에 있는지를 표시해 줌
- 2011년 우리나라의 인구가 48백만이고 우리나라는 East Asia & Pacific에 있으며 1인당 출산율은 1.4명, 1인당 국내총생산(GDP)은 $26206 정도 됨

■ 수치를 이용한 자료정리

- ○ 그래프 같은 시각적 기법은 자료의 특성을 파악하는데 있어 중요한 정보를 제공하지만 그것을 보는 사람에 따라 주관적으로 해석될 수 있음
- ○ 자료분석의 최종 결과는 객관적으로 그 자료의 특성을 나타내는 수치로 제시

□ 중심위치

- ○ n개의 수치형 자료: $x_1, x_2, \ldots, x_n$
  - x_i는 i번째 표본의 값
  - n을 **표본크기(sample size)**
- ○ 중심위치로 가장 많이 사용되는 통계값은 표본평균이며 대체 통계값으로 중앙값, 절사평균, 최빈값 등이 있음

① 표본평균(sample mean)

◦ 표본평균은 표본의 합을 표본크기로 나눈 값

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i}^{n} x_i$$

- $\overline{x}$는 x bar라고 읽는데 통계학에서 bar 표시는 해당 자료의 평균을 의미

◦ 평균을 중심으로 좌우의 무게가 같은 무게중심

◉ 오름차순으로 정렬된 자료 $x_1, \ldots, x_n$의 무게중심이 $\bar{x}$이고 $\bar{x}$의 좌측에 m개의 자료가, 나머지가 우측에 있다면

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

○ $\sum_{i=1}^{n}(x_i-\bar{x})=\sum_{i=1}^{n}x_i-n\bar{x}=0$가 되어 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$

○ $x_i-\bar{x}$: i번째 표본의 **편차(deviation)**이며 편차의 합은 0

◉ 통계학 관련 학과 취업률

○ 통계학 관련 42개학과의 취업률의 합은 2486.4

$$\bar{x} = \frac{55.6 + 83.3 + \cdots + 41.2 + 56.3}{42} = \frac{2486.4}{42} = 58.77$$

○ 표본비율(sample proportion)

◦ 관측값이 어떤 범주에 속하면 x_i의 값을 1, 속하지 않으면 0으로 표시

◦ 전체 표본 중에서 이 범주에 포함된 표본의 수는 $y = x_1 + \cdots + x_n$ 이며 이 범주에 포함된 표본비율은

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

◦ **표본비율 또한 일종의 표본평균**으로 이해할 수 있음

◉ 통계학전공 학생의 취업률

○ 임의로 선택된 42개 통계학 관련 학과를 2010년 8월과 2011년 2월에 졸업한 1568명(남자 715명, 여자 853명) 중 취업대상자 1371명(남자 628명, 여자 760명)의 자료

○ 건강보험DB직장가입자와 해외취업자는 791명(남자 367명, 여자 423명)인 것으로 조사

【표 2.12】 통계학과 졸업생의 취업률

취업률	취업자	취업대상자	취업률
전체	791	1371	55.70%
남자	367	628	58.44%
여자	423	760	55.66%

○ 이상점(outlier)

- ○ 대부분의 관측값들에서 멀리 떨어져 있는 일부 관측값
- ○ 표본을 수집하는 과정에서 이상점이 자료에 포함되는 경우와 아닌 경우 표본평균을 비교해 보면 값이 차이가 크게 나는 경향이 있음 ⇨ 이상점에 로버스트(robust)하지 않음

◉ 8명의 졸업생의 초임월급 실수령액(단위 만원) 자료

200, 225, 210, 205, 205, 220, 350, 205

- ○ 8명의 수령액 합은 1820만원이고 평균은 $\bar{x} = \frac{1820}{8} = 227.5$
- ○ 문제는 8명 중 7명의 수령액이 평균보다 낮아 평균이 중심위치로 적절한가에 대한 의문
- ○ 이와 같은 결과는 자료 중 350만원이라는 값이 다른 자료와 너무 동떨어져 있어 평균의 값을 크게 만들었기 때문에 발생

- 표본추출과정에서 이상점이 포함될 수도 있고 안 될 수도 있는데 포함여부에 따라 결과에 차이가 크게 발생한다면 대푯값으로써 적절하지 않음
- 표본평균은 이상점에 영향을 많이 받기 때문에 자료에 이상점이 있는 경우 안정적인 중심위치로 적절하지 않음
- 중심위치로 표본평균을 사용하려면 계산 전에 자료에 이상점이 있는지를 먼저 확인

○ 표로 정리된 자료의 표본평균

◦ 원자료를 알 수 없기 때문에 정확한 표본평균을 계산할 수 없지만 근사적인 값은 구할 수 있음

【표 2.13】 수치형자료의 도수분포표

계급	도수	상대도수	누적 상대도수	계급 중간값	밀도
$[L_1,\ U_1)$	f_1	r_1	c_1	m_1	d_1
$[L_2,\ U_2)$	f_2	r_2	c_2	m_2	d_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[L_k,\ U_k]$	f_k	r_k	c_k	m_k	d_k

- 계급중간값 $m_j = (L_j + U_j)/2$을 구함
- m_j는 j번째 계급에 속하는 관측값들의 대표하는 값이고 $m_j f_j$는 해당 계급의 관측값의 합에 대한 근사값
- k개의 계급이 있는 경우 표본평균은

$$\overline{x}_g = \frac{1}{n}\sum_{j=1}^{k} m_j f_j = \sum_{j=1}^{k} m_j \left(\frac{f_j}{n}\right) = \sum_{j=1}^{k} m_j r_j$$

◉ 통계학 관련학과 취업률

【표 2.14】 2011년 통계학 관련학과 취업률

취업률	도수	상대도수	누적 상대도수	계급 중간값
10%이상~40%미만	3	0.071	0.071	25
40%이상~50%미만	6	0.143	0.214	45
50%이상~60%미만	13	0.310	0.524	55
60%이상~70%미만	10	0.238	0.762	65
70%이상~80%미만	6	0.143	0.905	75
80%이상~100%	4	0.095	1.000	90

$$\overline{x_g} = \frac{1}{42}(25\times 3 + 45\times 6 + \cdots + 90\times 4)$$
$$= 25\times 0.071 + 45\times 0.143 + \cdots + 90\times 0.095 = 60.01$$

○ 원자료를 이용하여 나온 표본평균 58.77와 비슷한 것을 볼 수 있다.

② 표본중앙값(sample median)

- 자료를 크기순서대로 나열했을 때 가운데 위치에 있는 값으로 표본중위수라고도 함
- **순서통계량(order statistics)** : 표본을 오름차순으로 정렬했을 때 i 번째로 작은 값을 $x_{(i)}$ 라고 하면 $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ 가 성립
  - 예) 만약 $n=5$ 이면 3번째 순서통계량 $x_{(3)}$, $n=6$ 이면 3번째와 4번째 순서통계량 사이인 $(x_{(3)}+x_{(4)})/2$ 을 표본중앙값이라고 정의

○ 표본중앙값의 일반식

$$\tilde{x} = \begin{cases} x_{(k_1)}, & n = \text{홀수} \\ \frac{1}{2}\left(x_{(k_2)} + x_{(k_2+1)}\right), & n = \text{짝수} \end{cases}$$

- $k_1 = (n+1)/2$이고 $k_2 = n/2$이다.

◉ 통계학 관련학과 취업률

○ 오름차순으로 정렬

19.6	22.7	31.6	40.5	41.0	41.2	41.3	43.4	46.3	50.0	52.4	52.8
53.1	53.8	54.8	55.6	55.6	55.6	56.5	58.1	58.6	59.5	60.7	61.9
63.6	64.3	64.5	64.6	65.2	65.4	66.7	67.9	71.4	71.4	72.1	73.3
77.1	78.4	80.0	81.3	83.3	91.3						

○ $n = 42$ 이므로 취업률의 표본중앙값은 21번째와 22번째 순서통계값 58.6과 59.5의 평균

$$\tilde{x} = \frac{58.6 + 59.5}{2} = 59.05$$

○ 표본중앙값은 극단적인 값에 영향을 받지 않음

- 예) 취업률 자료에서 19.6이 0으로 가거나 91.3이 100으로 가도 표본중앙값의 변화는 없음

⇨ 이상점의 유무에 관계없이 안정적은 중심위치를 제공한다는 것을 의미하며 이를 이상점에 로버스트(robust)하다고 함

○ 표본중앙값을 계산하는데 있어 자료의 값들은 순서통계량을 구하는데 이용될 뿐이고 중앙에 있는 하나 또는 두 개의 관측값만 직접 사용 ⇨ 자료가 가지고 있는 정보를 다 활용하지 못함

○ 어떤 값을 중심위치로 사용해야 하는가?

- 두 통계값을 계산하여 차이가 크지 않으면 표본평균을 차이가 크면 중앙값을 사용하는 방법을 제안 ⇦ 두 값의 차이가 크다는 것은 자료 중에 이상점이 있을 가능성이 높기 때문
- 일반적으로 임금이나 소득에 관련된 자료에는 이러한 이상점들이 종종 발생하기 때문에 대푯값으로 평균을 사용하면 체감하는 것보다 높게 느껴지는 경우가 있음

◉ 미국 실질임금

- 2013년 5월 1일 The New York Times에 F. Norris가 쓴 'Can Every Group Be Worse Than Average? Yes.'
- 물가를 보정한 미국인 실질임금의 중앙값은 13년 전보다 0.9% 증가
- 고용된 사람들을 교육수준에 따라 그룹은 나누어 2000년 대비 2013년 실질임금의 중앙값을 비교하면 모두 감소

그룹	고교 중태	고교 졸업	대학 중태	대졸 이상
증가율	-7.9%	-4.7%	-7.6%	-1.2%

⇨ 심슨의 역설(Simpson's paradox)

○ 13년 동안 각 그룹에 해당되는 인원에 변동으로 발생

- 그룹 내에서는 실질소득이 줄어 듬
- 상대적으로 보수가 높은 대졸이상 인원의 채용은 늘어나고 고졸 이하의 채용은 줄어 듬
- 전체적으로 임금이 상승한 것처럼 보임

③ 표본절사평균(sample trimmed mean)

- 표본평균은 모든 자료의 정보를 사용하지만 이상점에 로버스트 하지 않은 반면 표본중앙값은 로버스트하지만 자료의 정보를 다 활용하지 못한다는 장단점
- 두 통계값이 가지고 있는 장점을 살리면서 단점을 줄여주는 통계값
- α% 표본절사평균은 순서통계량의 하위 α%에서 상위 α%까지의 자료를 이용하여 표본평균을 계산
  - α **백분위수(percentile)** : 순서통계량에서 하위 α%의 값
  - $p = \alpha/100$이면 p**분위수(quantile)** = α백분위수

- 상위 α%의 값은 (100-α)백분위수 또는 $(1-p)$분위수
- 하위 α%에 해당되는 순서를 계산하면 상위 α%는 반대방향에서 동일한 순서에 해당되는 값
- 예) 30개의 자료가 있고 하위 α%가 4번째 순서통계량이었다면 상위 α%는 위에서 4번째인 27번째 순서통계량 ⇨ α% 표본절사평균은 4번째 순서통계량부터 27번째 순서통계량까지 24개 자료의 평균

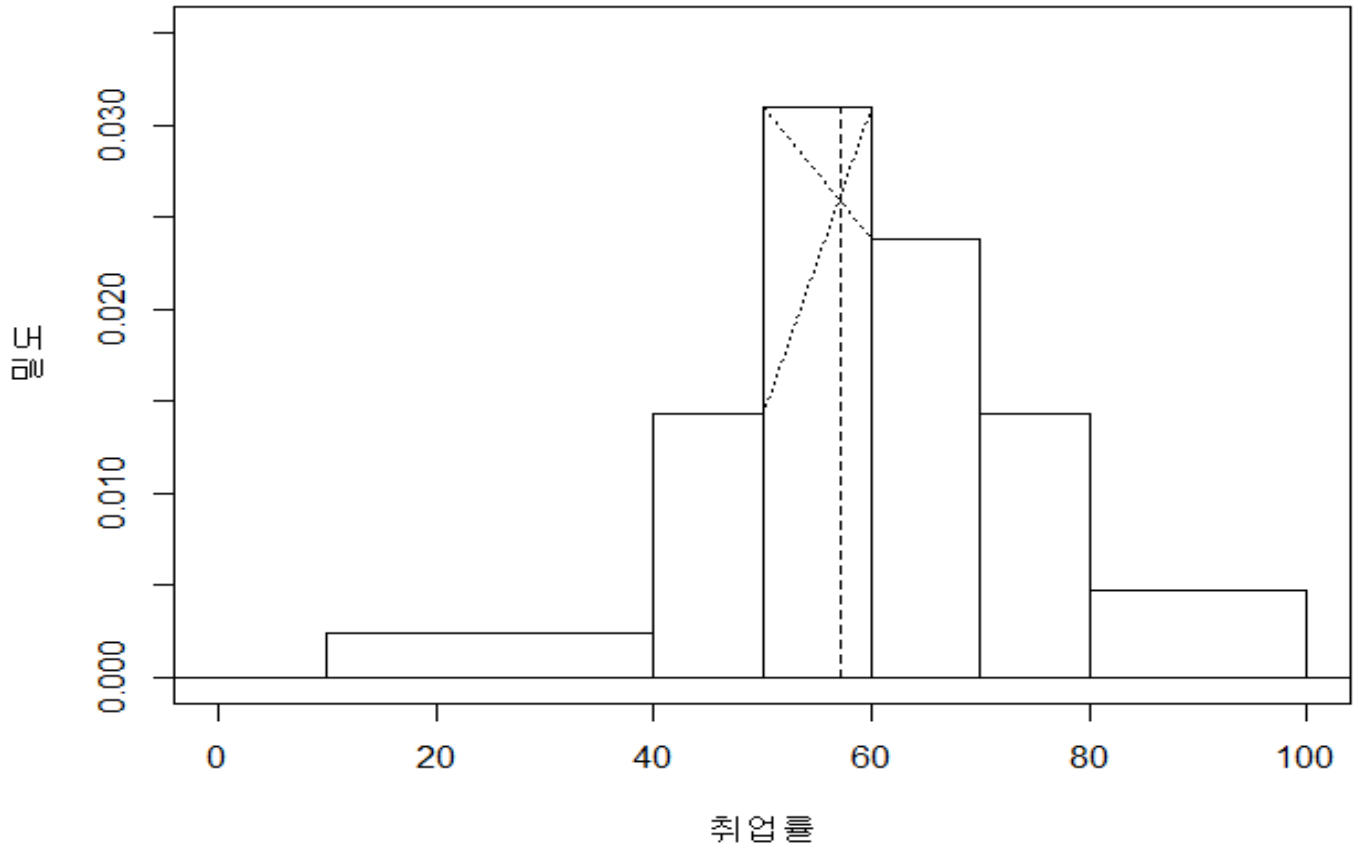
○ 적절한 크기의 α를 정하면 자료에 포함된 이상점이 제외시키면서 (100-2α)%만큼의 관측값을 사용 ⇨ 많은 자료정보를 사용하면서 로버스트한 중심위치를 제공

【표 2.15】 2013 ISU 여자피겨스케이트 점수표

선수	요소	심판									평균	절사 평균
		1	2	3	4	**5**	6	7	8	9		
김연아	Skating Skills	9.50	9.25	9.25	9.00	**8.75**	9.50	9.25	9.25	9.00	9.19	**9.21**
	Transition/Linking Footwork	9.00	9.25	8.75	8.75	**8.50**	9.25	9.00	8.75	8.75	8.89	8.89
	Performance/ Execution	10.0	10.0	9.00	9.25	**8.50**	9.50	9.75	9.00	9.00	9.33	**9.36**
	Choreography/ Composition	9.50	9.75	9.00	9.25	**8.50**	9.00	10.0	8.75	9.00	9.19	9.18
	Interpretation	10.0	10.0	9.25	9.00	**8.50**	9.25	10.0	9.00	9.00	9.33	**9.36**
아시다 마오	Skating Skills	8.50	8.75	8.50	8.25	**8.75**	8.50	8.50	8.75	9.00	8.61	8.61
	Transition/Linking Footwork	8.25	8.50	8.25	8.00	**8.50**	8.25	8.00	8.25	8.00	8.22	8.21
	Performance/ Execution	8.50	9.00	8.50	8.25	**8.75**	8.75	8.75	8.50	8.50	8.61	8.61
	Choreography/ Composition	9.00	9.00	8.50	8.50	**8.75**	8.50	8.75	8.50	8.50	**8.67**	8.64
	Interpretation	8.50	8.75	8.50	8.50	**9.00**	8.75	8.75	8.25	9.00	8.67	8.68

④ 표본최빈값(sample mode)

- 자료 중 빈도가 가장 많은 값
- 연속형 자료의 경우에는 자료의 값을 직접 사용하기보다는 그룹화하여 히스토그램을 그리고 간단하게 가장 높은 밀도를 가지는 구간의 중간값을 최빈값으로 사용하거나 내사법을 이용하여 가장 높은 밀도의 위치를 추정
- 여러 개 나올 수 있어 자주 사용하는 통계값은 아니지만 일봉 형태의 히스토그램에서는 가장 높은 밀도를 가지는 부분으로 중요한 위치



【그림 2.14】 통계학 관련전공 취업률 최빈값

□ 퍼짐의 측도

○ 중심위치만큼 중요한 통계값이 **산포(dispersion)**

○ 자료들이 얼마나 퍼져 있는가를 나타낼 뿐만 아니라 중심위치가 얼마나 안정적인지에 대한 중요한 정보를 제공

- 자료가 조밀하게 모여 있는 경우 중심위치에 대한 정확도는 높아지지만 넓게 퍼져 있는 경우 중심위치의 변동성이 커지기 때문에 신뢰도가 떨어짐

① 범위(range)

◦ 자료 중 가장 큰 값과 작은 값의 차이

$$\text{범위} = x_{(n)} - x_{(1)}$$

- 예) 취업률 자료에서 최고 취업률은 91.3%이고 최저 취업률은 19.6%

  ⇨ 취업률 자료의 범위: 91.3% - 19.6% = 71.7%

◦ 표본은 최대값 $x_{(n)}$과 최소값 $x_{(1)}$을 계산하는데만 이용하기 때문에 많은 정보를 활용하지 못함

◦ 이상점이 있으면 전체 형태와 관계없이 범위가 클 수 있어 범위를 통해 퍼진 정도를 평가하기에는 무리가 있음

② 사분위범위(Interquartile-Range)

- **사분위수(quartile)** : 자료를 동일한 비율로 4등분 할 때의 세 위치
- 자료를 오름차순으로 정렬했을 때
  - 25% 지점: 제1사분위수(Q_1)
  - 50% 지점: 제2사분위수(Q_2) = 표본중앙값
  - 75% 지점: 제3사분위수(Q_3)
- 사분위(간)범위는 제3사분위수와 제1사분위수의 차이

$$\text{IQR} = Q_3 - Q_1$$

○ 사분위수 계산 I

- $k = np$, $p = 0.25, 0.5, 0.75$ 계산
- k가 정수이면 $\left(x_{(k)} + x_{(k+1)}\right)/2$, 아니면 $x_{(k'+1)}$, k'는 k의 정수부분

◉ 취업률 자료

- $n = 42$이므로 $p = 0.25$일 때 k=42×0.25=10.5 ⇨ $Q_1 = x_{(11)} = 52.4$
- $p = 0.75$ 일 때 k=42×0.75=31.5 ⇨ $Q_3 = x_{(32)} = 67.9$
- $IQR = 67.9 - 52.4 = 15.5$

○ 사분위수 계산 II

- ◦ $k=(n-1)p+1, \quad p=0.25, 0.5, 0.75$ 계산
- ◦ k가 정수이면 $x_{(k)}$가 해당 사분위수, 아니면 비례에 의한 내사법을 적용

◉ 취업률 자료

- ◦ $n=42$이므로 Q_1에 해당되는 위치는 41×0.25+1=11.25
- ◦ 11번째와 12번째 순서통계값 사이에 있으며 비례식에 의해

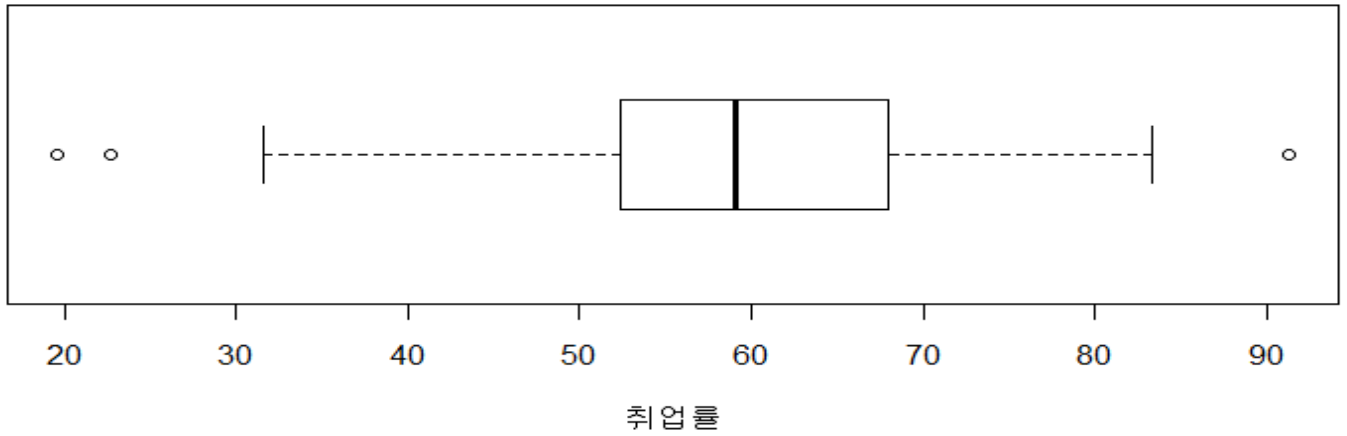
$$\begin{aligned} Q_1 &= 0.75\times x_{(11)} + 0.25\times x_{(12)} \\ &= 0.75\times 52.4 + 0.25\times 52.8 = 52.5 \end{aligned}$$

- ◦ Q_3는 31.75번째 위치로 0.25×66.7+0.75×67.9=67.6
- ◦ $IQR = 67.6 - 52.5 = 15.1$

○ 상자그림(box plot)

◦ Tukey라는 통계학자에 의해 제안된 그림

◦ 그룹 간의 비교나 이상점 검출 등에 사용되는 그림



【그림 2.15】 통계학 관련학과 취업률의 상자그림

◦ Q_1, Q_2, Q_3을 계산하여 직사각형의 상자를 표시

- $Q_1 = 52.5,\ Q_2 = 59.05,\ Q_3 = 67.6$
- Q_1과 Q_2의 거리가 Q_2와 Q_3의 거리보다 짧은 것은 Q_1과 Q_2 사이에 있는 자료가 좀 더 조밀하게 모여 있음

◦ $IQR,\ L = Q_1 - 1.5 \times IQR,\ U = Q_3 + 1.5 \times IQR$를 계산

- $L = 52.5 - 1.5 \times 15.1 = 29.85,\ U = 67.6 + 1.5 \times 15.1 = 90.25$

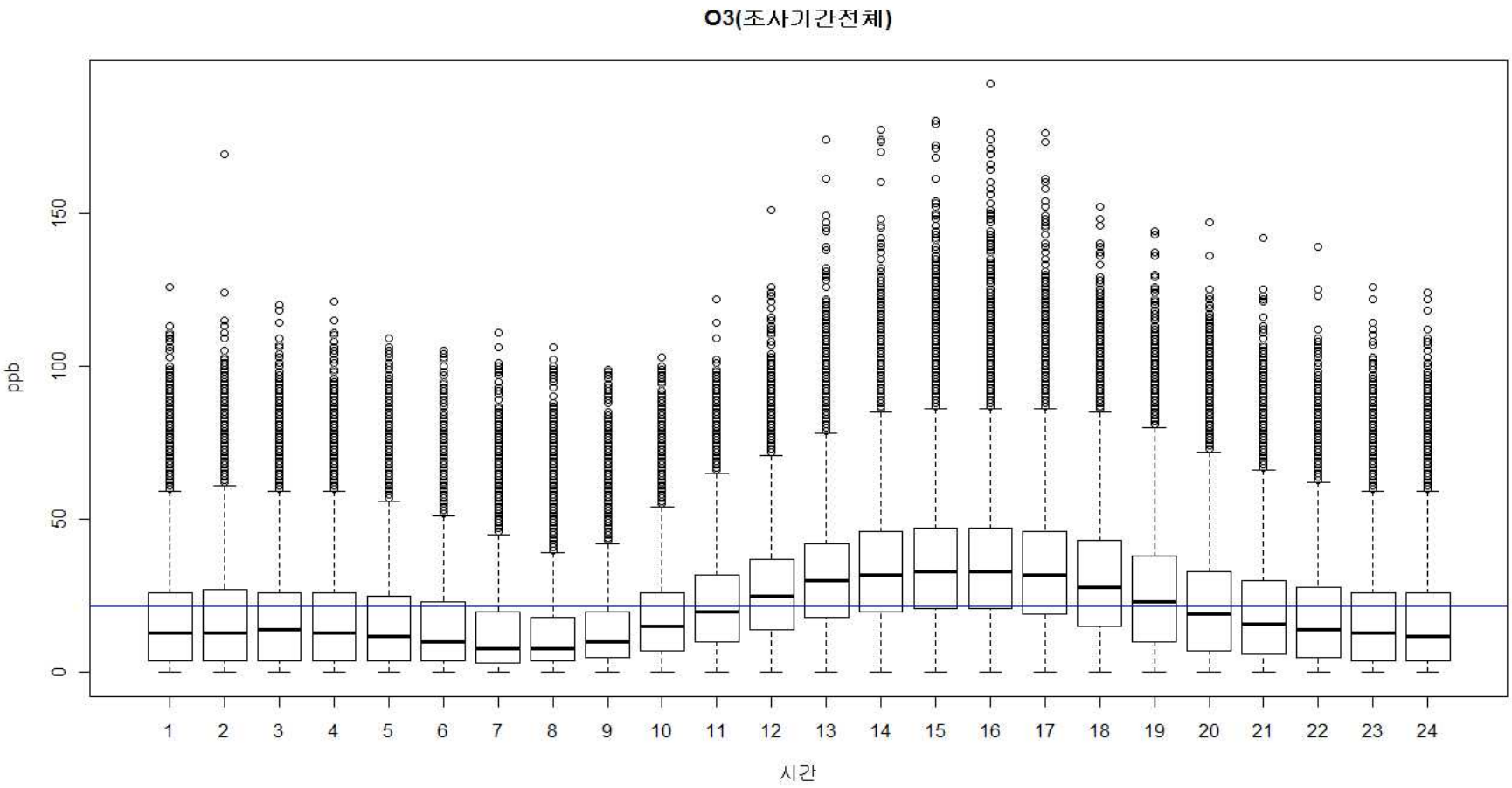
◦ L보다 큰 관측값 중 가장 작은 값, U보다 작은 관측값 중에 가장 큰 값에 직선에 직선을 표시하고 상자와 연결

- 31.6과 83.3에 직선표시

◦ 직선 밖의 관측값은 이상점으로 ○로 표시: 19.6, 22.7, 91.3

◉ 오존(O_3)자료

- 대기오염측정소의 주요 오염원에 대한 기초자료를 확보하기 위해 주요 100개 대기오염측정소를 선정
- 2005년 1월 1일 1시부터 2008년 12월 31일 24시까지 매 시간별로 측정된 오존(O_3)자료
- 모든 시간대에서 많은 이상점이 발견
- 오존의 오염도는 7시부터 9시까지 낮아졌다가 11시부터 급격히 증가하여 15시부터 17시경에 가장 높은 오염도를 가지다 서서히 감소하는 형태
- 중간의 실선은 전체 오존오염도 평균을 표시한 것



【그림 2.16】 시간대별 오존 오염도

③ 표본분산과 표본표준편차

- 범위나 사분위수범위의 경우 특정 위치의 두 값을 이용
- 모든 자료들 간의 거리의 합을 이용하는 방법은?
- **거리(distance)**: 임의의 점 a, b, c에 대해,
  - $a = b$이면 $D(a, b) = 0$이고 그 역도 성립
  - $D(a, b) = D(b, a)$
  - $D(a, b) \leq D(a, c) + D(c, b)$
  - 예) $D(a, b) = |a - b|$, $D(a, b) = (a - b)^2$

○ 모든 관측값들 간 거리의 합

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|, \qquad \sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_i - x_j\right)^2$$

- 자료들이 넓게 퍼져 있으면 이 합들은 커질 것이고 모여 있으면 작아짐
- n^2 개의 거리 합을 계산

○ 임의의 중심위치 a에서 자료들이 떨어져 있는 거리의 합

$$L_1(a) = \left|x_1 - a\right| + \left|x_2 - a\right| + \cdots + \left|x_n - a\right| = \sum_{i=1}^{n}\left|x_i - a\right|$$

$$L_2(a) = (x_1 - a)^2 + (x_2 - a)^2 + \cdots + (x_n - a)^2 = \sum_{i=1}^{n}(x_i - a)^2$$

○ **이 측도를 사용하기 위해서는 a값을 정해야 하는데 어떤 값으로 선택해야 할까?**

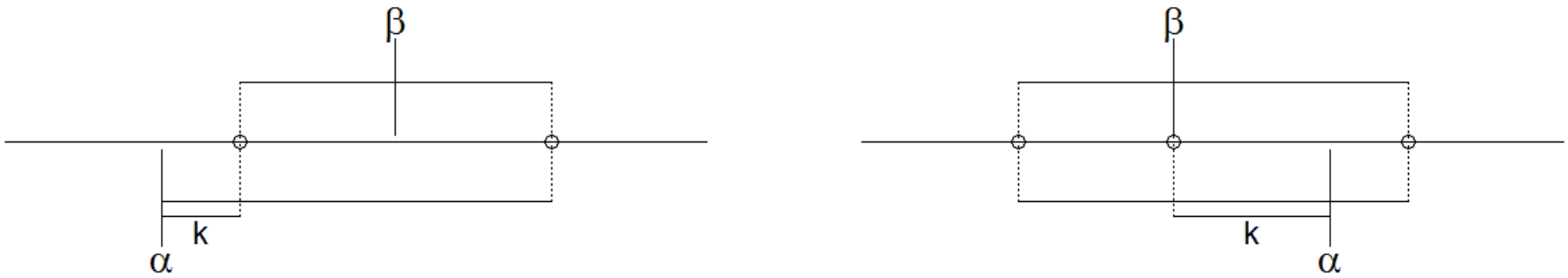
※ a가 좋은 중심위치가 되려면 자료들과의 거리가 가능한 짧아야 하며 결국 **거리의 합을 최소로 만드는 값**

○ $L_2(a)$를 a에 대해 미분한 식이 0이 되는 값

$$\frac{dL_2(a)}{da} = -2\sum_{i=1}^{n}(x_i - a) = -2\left\{\sum_{i=1}^{n}x_i - na\right\} = 0$$

$$\Rightarrow\ a = \frac{1}{n}\sum_{i=1}^{n}x_i = \overline{x}$$

○ $L_1(a)$의 경우 a로 미분불가능



【그림 2.17】 절대편차합의 비교

○ 자료가 2개 있을 때, $L_1(\beta)$가 $L(\alpha)$보다 밑에 있는 선분의 길이 k의 두 배 작음

○ 자료가 3개인 경우에는 $L(\beta)$가 $L(\alpha)$보다 k만큼 작음

⇨ $L_1(a)$를 최소로 만드는 a는 표본중앙값

○ 퍼져있는 정도를 나타내는 통계값

- $L_1(\tilde{x}) = \sum_{i=1}^{n} \left| x_i - \tilde{x} \right|$

- $L_2(\overline{x}) = \sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2$ ⇦ 편차의 제곱합

○ 편차의 합이 0 ⇦ $\frac{dL_2(a)}{da} = 0$ 를 만족하는 $a = \overline{x}$

○ 표본분산(sample variance)

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

- 표본분산은 n개의 편차를 사용하는 것 같지만

  $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$이라는 제약조건 때문에 $n-1$개의 편차 정보를 사용

- $n-1$: 자유롭게 가질 수 있는 편차의 개수라고 해 **자유도(degree of freedom)**이라고 함

○ 표본분산 간이식

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

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

○ 표본표준편차(sample standard deviation)

- $x^2+y^2=r^2 \Rightarrow \sqrt{x^2+y^2}=r$
- 표본분산은 편차의 제곱합을 이용하기 때문에 분산의 단위는 관측값 단위의 제곱
- 눈으로 이해하는 산포와 일치하기 위해서는 자료를 측정할 때의 단위로 표시

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

◉ 취업률 자료

○ 표본의 합과 제곱합

$$\sum_{i=1}^{42} x_i = 2468.4, \quad \sum_{i=1}^{42} x_i^2 = 154975.4$$

○ 편차의 제곱합

$$\sum_{i=1}^{42} (x_i - \overline{x})^2 = 154975.4 - \frac{2468.4^2}{42} = 9904.006$$

○ 표본분산과 표본표준편차

$$s^2 = \frac{9904.006}{41} = 241.56, \quad s = \sqrt{241.56} = 15.54$$

○ 표준화

◦ 수능시험은 과목별로 난이도가 다를 수 있기 때문에 원점수로 과목 간 성적을 비교 X ⇨ 표준화점수

$$z_i = \frac{x_i - \bar{x}}{s}$$

◦ $\sum_{i=1}^{n} z_i = \frac{1}{s}\sum_{i=1}^{n}(x_i - \bar{x}) = 0 \Rightarrow \bar{z} = 0$

◦ $\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2 = \frac{1}{n-1}\sum_{i=1}^{n} z_i^2 = \frac{1}{(n-1)s^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 1$

○ 표준화는 평균을 0, 표준편차를 1이 되도록 만들어 측정 단위에 영향을 받지 않게 중심위치와 척도(scale)를 조정하고 상대적 비교가능

○ 원자료 대신 도수분포표 형태로 주어진 경우

- m_i: i번째 계급의 중간값
- f_i: i번째 계급의 도수

$$\sum_{i=1}^{n}(x_i-\overline{x})^2 \simeq \sum_{i=1}^{k}\left(m_i-\overline{x}_g\right)^2 f_i$$

- 표본분산과 표본표준편차

$$s_g^2=\frac{1}{n-1}\sum_{j=1}\left(m_j-\overline{x}_g\right)^2 f_i\,, \quad s_g=\sqrt{s_g^2}$$

◉ 취업률 자료

$$
\begin{aligned}
s_g^2 &= \frac{1}{42-1}\left\{(25-60.01)^2\,3+(45-60.1)^2\,6+\ \cdots\ +(90-60.1)^2 4\right\} \\
&= \frac{10550}{41} = 257.32 \\
s_g &= \sqrt{257.32} = 16.04
\end{aligned}
$$

④ 변동계수(coefficient of variation)

○ 표준편차가 평균에 영향을 받는 경우

- 예) 후진국의 소득분포와 선진국의 소득분포를 비교
- 예) 유아와 성인의 신장이나 체중의 분포를 비교

⇨ 비교 그룹간의 평균이 큰 차이가 있고 자료의 특성이 평균이 커지면 산포도 커지는 경향이 있기 때문

○ 실제로 0을 하한으로 가지는 많은 자료들이 이러한 성질을 가지고 있음

○ 표준편차만 이용하여 산포를 비교하는 것은 적절하지 않을 수 있어 평균으로 표준편차를 보정

$$CV = \frac{s}{\bar{x}} \times 100$$

- 100을 곱하는 이유는 표본평균에 비해 표본표준편차가 얼마나 큰지를 % 개념으로 표시하기 위한 것으로 100을 생략할 수도 있음

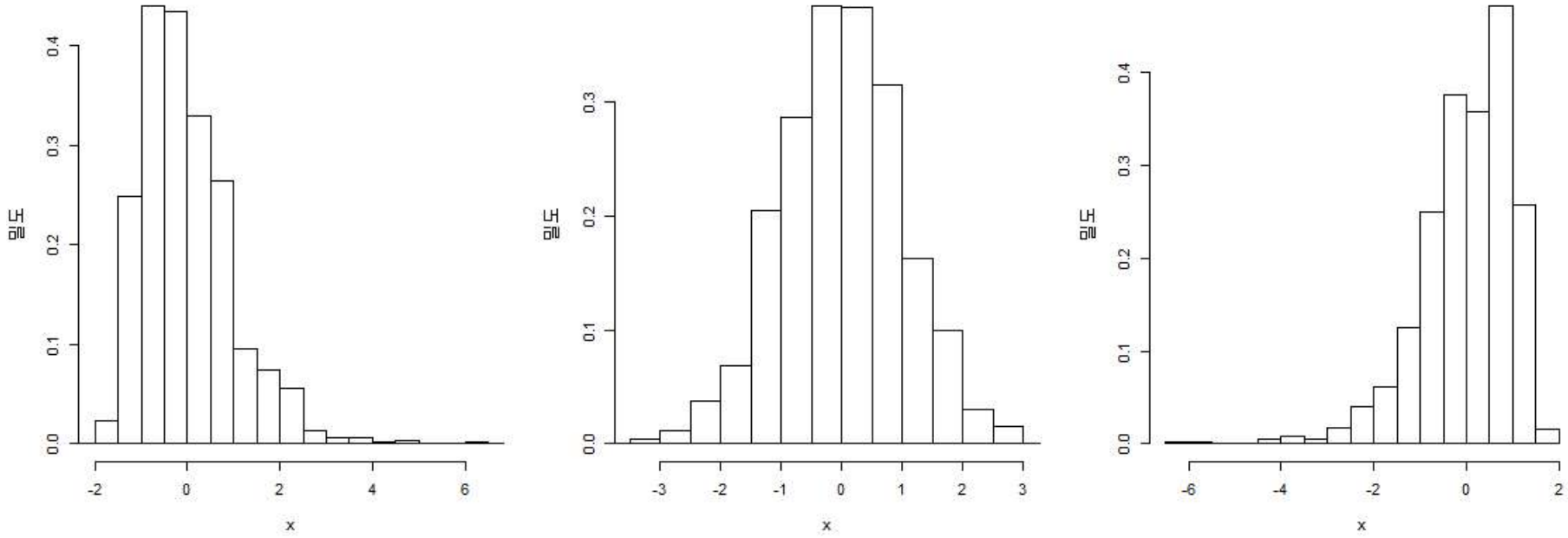
○ 신장과 체중과 같이 단위가 전혀 다른 자료들의 퍼져있는 정도를 비교할 때에도 사용

□ 분포의 형태

- ○ 수치형 자료에 대한 통계분석 방법은 대부분 모집단이 중심위치를 기준으로 좌우대칭인 형태를 가진다고 가정
- ○ 통계분석의 적절성은 분석방법에서 가정한 조건을 자료가 얼마나 만족하고 있는지에 영향을 받음
- ○ 자료의 분포 형태에 대한 측도이면서 자료가 모집단의 가정을 얼마나 만족하는지에 대한 측도로 사용

○ 왜도(skewness)

- ○ 자료가 중심위치를 기준으로 대칭적으로 분포되어 있는지는 히스토그램이나 상자그림을 통해 확인



【그림 2.18】 히스토그램

◦ 【그림 2.18】은 모두 평균이 0이고 표준편차가 1이지만 형태가 다른 자료의 히스토그램

◦ 왜도: 피어슨(Karl Pearson) 제안

$$\sqrt{b_1} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^3$$

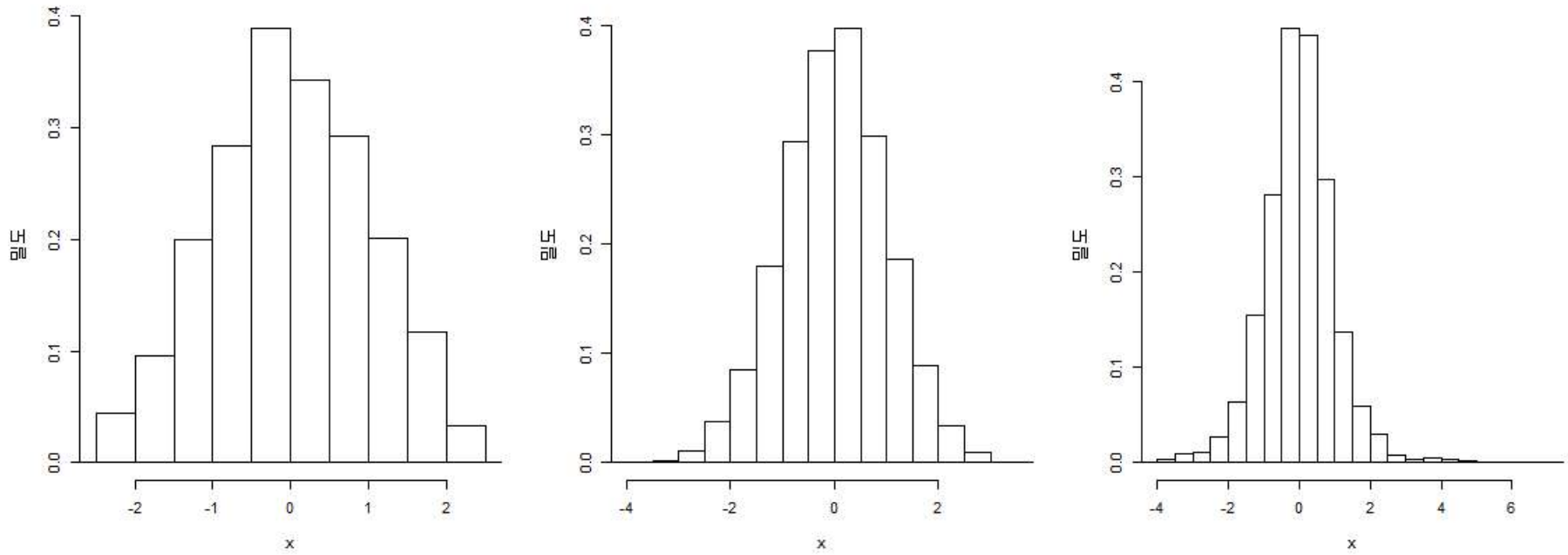
- $(x_i - \overline{x})^3$: 평균을 중심으로 왼쪽은 음수, 오른쪽은 양수
- 자료가 평균에서 멀어질수록 큰 음수나 큰 양수가 됨

◦ 좌우가 비슷한 형태를 가진다면 음수와 양수가 상쇄되어 $\sqrt{b_1}$은 0 근처 ⇨ 대칭적(symmetric)

- 왼쪽 그림: 오른쪽의 꼬리부분이 길게 부분되어 있어 큰 양수값을 가지는 자료가 있어 $\sqrt{b_1}$은 대칭일 때 보다 큰 값을 가짐 ⇨ 양의 왜도(positive skewness)를 가짐 또는 오른쪽으로 왜도(skewed to the right)됨
- 오른쪽 그림: 대칭일 때 보다 작은 값 ⇨ 음의 왜도(negative skewness)를 가짐 또는 왼쪽으로 왜도(skewed to the left)됨
- 두터운 꼬리(heavy tail) : 꼬리가 길게 분포된 것
- 수정된 왜도: $\sqrt{b_1} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s} \right)^3$

○ 첨도(kurtosis)

◦ 양쪽꼬리가 얼마나 두터운지를 나타내는 값



【그림 2.19】 히스토그램

○ 【그림 2.19】은 평균이 0이고 표준편차가 1인 자료의 히스토그램

○ 모두 대칭적인 형태를 가지나 꼬리가 왼쪽은 짧고 오른쪽은 길며 중간은 중간정도

○ 첨도: 피어슨(Karl Pearson) 제안

$$b_2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^4$$

- $(x_i - \overline{x})^4$: 평균을 중심으로 자료가 멀리 있으면 큰 값
- 항상 양수가 되며 분포의 중심보다는 꼬리부분이 얼마나 두터운지에 따라 영향을 많이 받음

○ 정규분포의 경우 이론적으로 첨도는 3

$$b_2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^4 - 3$$

○ 수정된 첨도

$$b_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

○ 심한 왜도를 가지거나 양쪽 꼬리가 두터운 경우에는 자료 중 이상점이 있을 가능성이 높아짐

○ 왜도나 첨도는 자료의 분포 형태를 나타내는 측도뿐만 아니라 분석 방법의 적절성을 확인하기 위한 측도로 사용

◉ 취업률 자료

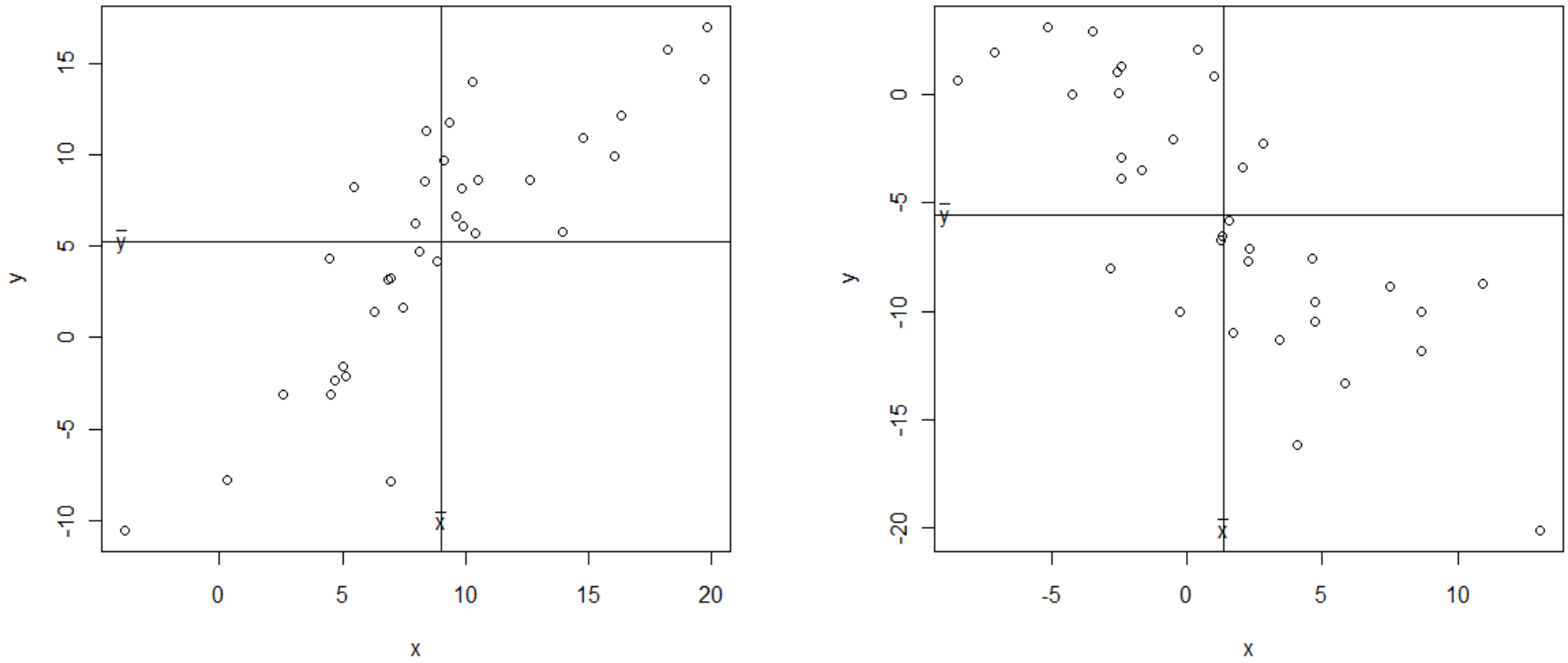
- $\sum x_i = 2468.4,\ \sum x_i^2 = 154975.4,\ \sum x_i^3 = 10211388,$
  $\sum x_i^4 = 699463185$
- 피어슨 왜도와 첨도

$$\sqrt{b_1} = \frac{1}{n-1}\sum\left(\frac{x_i - \overline{x}}{s}\right)^3 = -\frac{16.22}{41} = -0.396$$

$$b_2 = \frac{1}{n-1}\sum\left(\frac{x_i - \overline{x}}{s}\right)^4 = \frac{127.37}{41} = 3.107$$

■ 공분산과 상관계수

- 산점도를 통해 두 수치형 변수 간에 관계가 있는지를 시각적으로 확인
- 두 수치형 변수 간에 **직선관계**가 어느 정도인지를 나타내는 통계값
- 자료: $(x_1, y_1), (x_2, y_2), \ \ldots\ , (x_n, y_n)$



【그림 2.20】 양의 기울기와 음의 기울기를 가지는 산점도

- 양의 기울기를 가지는 경우 $(\overline{x}, \overline{y})$를 중심으로 1과 3사분면에 자료들이 많고 길게 분포
- 음의 기울기를 가지는 경우 대부분의 자료들이 2와 4사분면에서 길게 분포
- 자료의 직선관계를 표시하고자 할 때 $(\overline{x}, \overline{y})$을 중심으로 1과 3, 2와 4사분면의 자료가 동일한 성질을 가짐

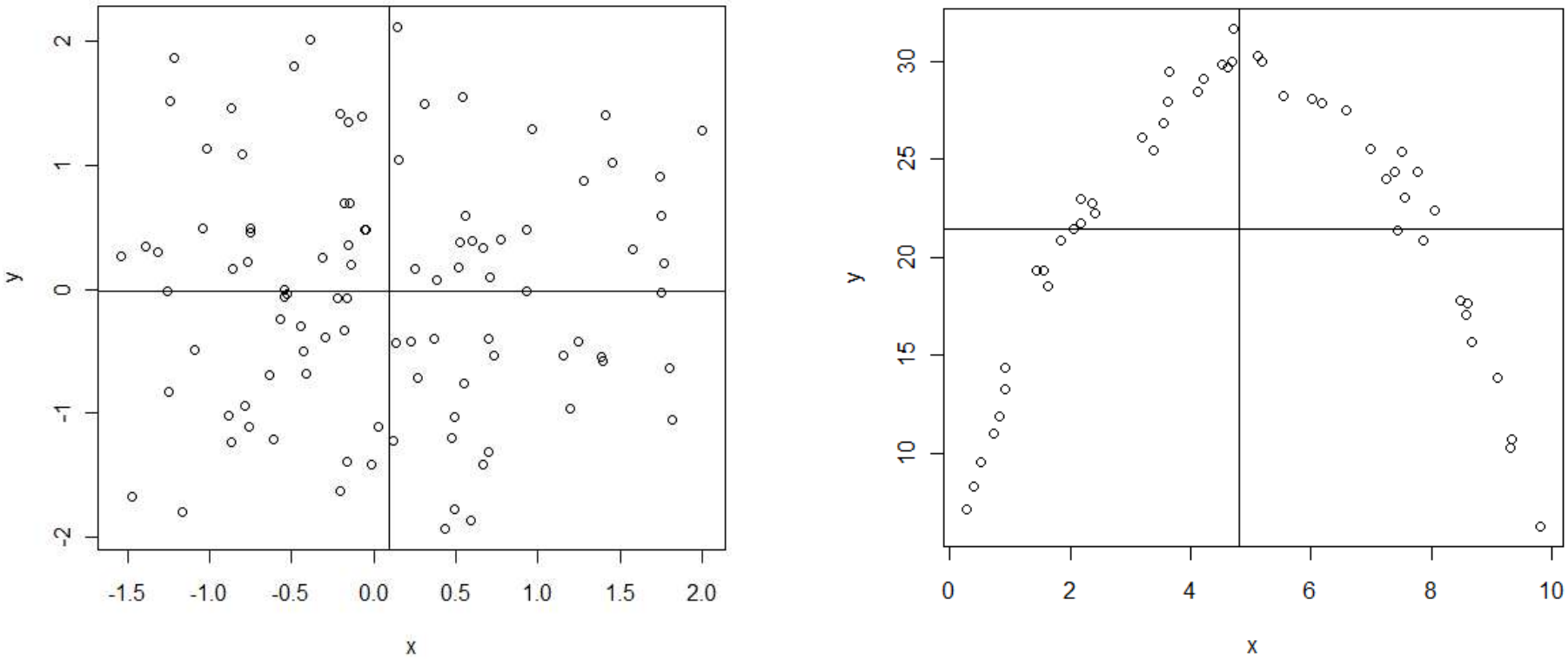
$$(x_i - \overline{x})(y_i - \overline{y})$$

⇨ 1과 3사분면은 양수, 2과 4분면의 값은 음수

○ 표본공분산(sample covariance)

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

- ◦ 왼쪽 산점도와 같이 양의 기울기를 가지는 선분에 자료가 모여 있으며 c는 양의 값
- ◦ 오른쪽 산점도와 같이 음의 기울기를 가지는 선분에 모여 있으며 음의 값
- ◦ 표본분산은 $\frac{1}{n-1}\sum(x_i - \overline{x})^2 = \frac{1}{n-1}\sum(x_i - \overline{x})(x_i - \overline{x})$가 되는데 뒤에 있는 항에서 x_i를 y_i로 바꾸면 c



【그림 2.21】 직선관계가 없는 산점도의 예

○ 표본공분산의 간편식

$$
\begin{aligned}
c &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n}x_i\,y_i - n\overline{x}\,\overline{y}\right\} \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n}x_i\,y_i - \frac{1}{n}\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i\right\}
\end{aligned}
$$

◉ 올림픽 개최 연도와 육상 100미터 우승기록

【표 2.16】 올림픽 100미터 자료 정리

	번호	x	y	x^2	y^2	xy
남자	1	1900	11	3610000	121	20900
	2	1904	11	3625216	121	20944
	⋮	⋮	⋮	⋮	⋮	⋮
	26	2012	9.63	4048144	92.7369	19375.56
	합	50924	266.95	99770832	2745.034	522514.2
	번호	x	z	x^2	z^2	xz
여자	1	1928	12.2	3717184	148.84	23521.6
	2	1932	11.9	3732624	141.61	22990.8
	⋮	⋮	⋮	⋮	⋮	⋮
	20	2012	10.75	4048144	115.5625	21629
	합	39456	223.67	77851232	2505.158	441064.2

○ 연도와 남자 우승기록의 표본공분산

$$c = \frac{1}{26-1}\left(522514.2 - \frac{1}{26}(50924)(266.95)\right)$$
$$= \frac{-338.177}{25} = -13.527$$

○ 연도와 여자 우승기록의 표본공분산

$$c = \frac{1}{20-1}\left(441064.2 - \frac{1}{20}(39456)(223.67)\right)$$
$$= \frac{-191.976}{19} = -10.104$$

○ 두 자료 모두 음의 기울기를 가지는 직선관계

○ 표본상관계수(coefficient of correlation)

◦ 공분산의 문제점은 측정 단위에 영향을 받기 때문에 그 값 자체로 선형관계의 정도를 알 수는 없음

- 예) 연도와 우승기록의 관계에서 우승기록을 초 단위가 아닌 분 단위로 표시하면 남자의 표본공분산은 –13.527/60=-0.225

◦ 피어슨의 표본상관계수: 표준화된 표본공분산

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)$$

○ 표본상관계수 간편식

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x}\,)(y_i - \overline{y}\,) = \sum_{i=1}^{n} x_i\, y_i - n\,\overline{x}\,\overline{y}$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x}\,)^2 = \sum_{i=1}^{n} x_i^2 - n\,\overline{x}^{2}$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y}\,)^2 = \sum_{i=1}^{n} y_i^2 - n\,\overline{y}^{2}$$

$$\Rightarrow\ r = \frac{\displaystyle\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\displaystyle\sum_{i=1}^{n} (x_i - \overline{x})^2}\ \sqrt{\displaystyle\sum_{i=1}^{n} (y_i - \overline{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}}\ \sqrt{S_{yy}}}$$

○ 표본상관계수의 성질

- $-1 \leq r \leq 1$
- 자료들이 어떤 기울기를 가지는 직선에 조밀하게 모일수록 $|r|$는 1에 근접
- 음의 기울기인 경우 r는 음수이며 음의 상관관계가 존재한다고 하고 양의 기울기인 경우 양수가 되며 양의 상관관계가 존재한다고 함
- 모든 관측값들이 직선 위에 위치하면 $|r| = 1$이 된다.
- $|r| \simeq 0$이면 상관관계가 없다고 함

◉ 올림픽 개최 연도와 우승기록

○ 남자의 상관계수

$$s_{xx} = 99770832 - \frac{50924^2}{26} = 30302.15$$

$$s_{yy} = 2475.034 - \frac{266.95^2}{26} = 4.177$$

$$r_{xy} = \frac{-338.177}{\sqrt{30302.15}\,\sqrt{4.177}} = -0.951$$

○ 여자의 상관계수

$$s_{xx} = 77851232 - \frac{39456^2}{20} = 12435.2$$

$$s_{zz} = 2505.158 - \frac{223.67^2}{20} = 3.745$$

$$r_{xz} = \frac{-191.976}{\sqrt{12435.2}\,\sqrt{3.745}} = -0.890$$

○ 두 표본상관계수 모두 –1에 가까운 값을 가지는 것으로 나타났으며 이는 연도와 우승기록 간에는 확실한 음의 상관관계가 있음

○ 표본상관계수는 두 변수 간에 직선관계가 있는지를 나타낼 뿐 인과관계를 나타내는 것은 아님

- 예) 휴대전화 보급률과 기대수명에 대한 상관계수를 구해보면 매우 높은 양의 상관관계를 가짐 ⇨ 기대수명을 늘리기 위해 휴대전화 보급을 늘려야 한다?

○ 잠복변수(lurking varaible): 두 변수에 영향을 주거나 관계가 있는 변수

○ 제3의 변수에 의해 나타나는 상관관계를 허위상관(spurious correlation) 또는 가짜상관 ⇨ 각각의 변수에서 잠복변수의 영향을 제거하고 표본상관계수를 계산하여 관련성을 파악

확률
(Probability)

■ 기본개념

◉ 확률이 발생하는 상황에서의 공통적인 특징

① 주사위 던지기
② 앞면이 나올 때까지 동전 던지기
③ 구매 한 스마트폰의 수명

- 실험을 시행하기 전에 발생할 수 있는 모든 결과는 알 수 있음
  ① $\{1, 2, 3, 4, 5, 6\}$
  ② 앞면[1]을 H, 뒷면을 T이라고 하면, $\{H, TH, TTH, \cdots\}$
  ③ x를 수명(단위 일)이라고 하면, $\{x \mid 0 \leq x\}$

- 실험을 하기 전까지 이들 결과 중 어떤 것이 발생할 것인지에 대해 확실하게 예측할 수 없음

- **확률실험(random experiment)**: 위의 두 성질을 가지는 실험

1) 일반적으로 그림이 있는 부분을 앞면 숫자가 있는 부분을 뒷면이라고 함

- **표본공간(sample space, Ω)**: 확률실험에서 발생 가능한 모든 결과들의 집합

- **사건(event)**: 표본공간 내에서 우리가 관심을 가지는 부분집합
  ① 홀수가 나오는 경우
  ② 3번 이상 던지는 경우
  ③ 365일 이전에 수명을 다 하는 경우

- 어떤 사건 A가 발생한다는 것은 실험결과가 A에 속하는 원소 중 하나이거나 A에 포함되어 있다는 것을 의미

■ **확률(probability)**: 이러한 사건이 발생할 가능성이 얼마나 되는지를 나타내는 수치적 측도

  ○ 확률을 언급하기 위해서는 해당하는 확률실험이 전제되어야 하고 이에 따른 표본공간과 사건이 정해져야 함

■ 표본공간과 사건은 수학적 관점에서 보면 일종의 집합

  ⇨ 확률을 정의하고 계산하기 위해서는 집합에 대한 기본 정의와 연산을 알아야 함

【표 3.1】 집합의 정의와 연산

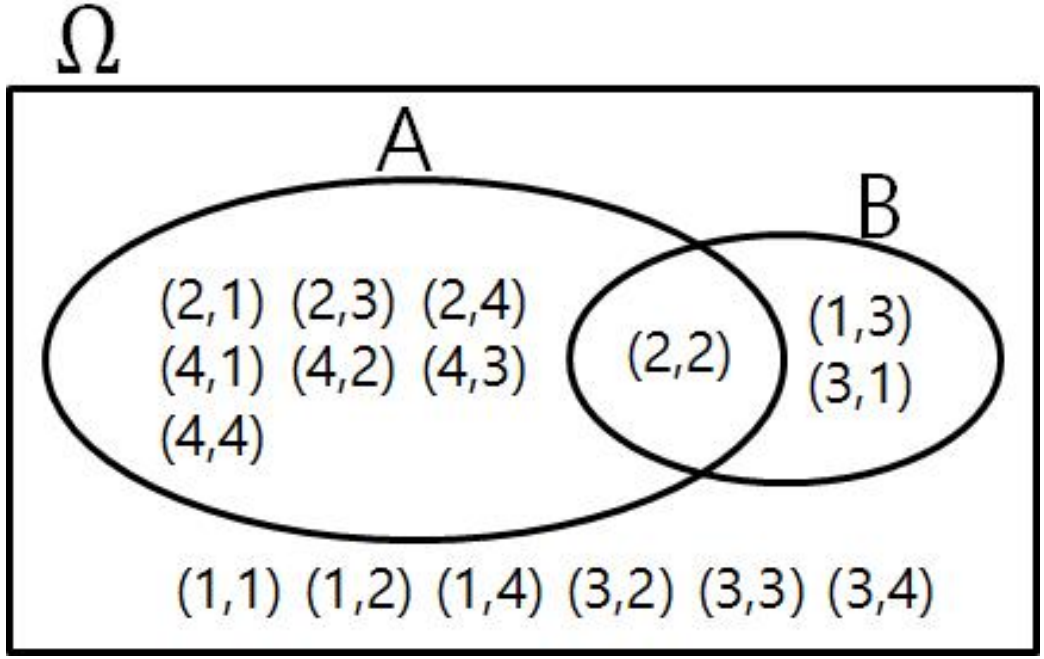
정의 및 법칙	표시 및 내용
• A와 B의 합사건(union)	$A \cup B = \{\omega \mid \omega \in A$ 또는 $\omega \in B\}$
• A와 B의 곱사건(intersection)	$A \cap B = \{\omega \mid \omega \in A$ 그리고 $\omega \in B\}$
• A의 여사건(complement)	$A^c = \{\omega \mid \omega \notin A$ 그리고 $\omega \in \Omega\}$
• 교환법칙(commutative law)	$A \cup B = B \cup A,\ A \cap B = B \cap A$
• 결합법칙(associative law)	$(A \cup B) \cup C = A \cup (B \cup C)$ $(A \cap B) \cap C = A \cap (B \cap C)$
• 분배법칙(distributive law)	$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$
• 드모르간(De Morgan)의 법칙	$(A \cup B)^c = A^c \cap B^c,\ (A \cap B)^c = A^c \cup B^c$
• 무한개의 사건이 존재하는 경우 $(A_1, A_2, \ldots)$	$\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \cdots \cup A_n$ $\bigcap_{i=1}^{n} A_i = A_1 \cap A_2 \cap \cdots \cap A_n$

- **서로배반사건(disjoint, mutually exclusive)**: 임의의 두 사건 A와 B가 공통부분이 없는 경우, 즉 $A \cap B = \varnothing$
- 밴 다이어그램(Venn diagram) 이용

◉ 정사면체 주사위 두 개를 던지기

- A: 첫 번째 주사위가 짝수인 사건
- B: 두 주사위의 합이 4인 사건

$$\Omega = \{(i,j) \mid i = 1,2,3,4,\ \ j = 1,2,3,4\}$$
$$A = \{(i,j) \mid i = 2,4,\quad j = 1,2,3,4\}$$
$$B = \{(i,j) \mid i+j = 4\} = \{(1,3),\ (2,2),\ (3,1)\}$$



- $A^c \cap B = \{(1,3), (3,1)\}$
- $A^c \cap B^c = (A \cup B)^c = \{(1,1), (1,2), (1,4), (3,2), (3,3), (3,4)\}$

■ 확률의 이해

□ 고전적 확률

○ 17세기 중반 파스칼이나 페르마 등이 도박문제에 대해 의견을 교환

- 카드게임에서 어떤 패가 더 높은 패인지를 결정하기 위해 각 패의 발생할 수 있는 빈도를 계산

○ 표본공간에서 사건이 해당되는 원소가 차지하는 비율

◦ 표본공간이 n개의 원소로 이루어져 있고 각 근원사건의 **발생가능성이 동일(equally likely)**한 경우, k개의 원소를 가지는 사건 A의 확률

$$P(A) = \frac{k}{n}$$

◉ 정사면체 주사위

◦ 각 눈이 나올 가능성은 동일

$$P(A) = \frac{8}{16} = \frac{1}{2}, \quad P(B) = \frac{3}{16}$$

$$P(A^c \cap B) = \frac{2}{16} = \frac{1}{8}, \quad P(A^c \cap B^c) = \frac{6}{16} = \frac{3}{8}$$

○ 경우의 수(the number of cases)

◦ 확률을 계산하기 위해서는 표본공간과 사건에 있는 원소의 개수를 효율적으로 계산하는 것이 중요

◦ 경우의 수의 기본 법칙은 **곱의 법칙(multiplication rule)**

- 어떤 실험이 m개의 연속된 단계로 이루어져 있고 i-번째 단계에서 발생 가능한 결과의 수가 n_i개이면 전체 실험에서 발생 가능한 경우의 수는

$$n = n_1 \times n_2 \times \cdots \times n_m$$

◉ 세트메뉴의 경우의 수

- 세트메뉴에는 4가지 음료수, 2가지 샐러드, 5가지 메인, 4가지의 디저트 중에서 각각 하나씩을 선택
- 선택할 수 있는 세트의 종류는 $4 \times 2 \times 5 \times 4 = 160$

- 경우의 수의 일반적인 문제 : 1번부터 n번까지 적혀있는 공이 들어 있는 주머니에서 k개를 무작위로 선택
- k개를 어떻게 추출하고 나열할 것인지에 따라 달라짐
- 추출방법:
  - **복원(with replacement)추출**
  - **비복원(without replacement)추출**
  - Q? 만약 한꺼번에 k개의 공을 뽑으면?
- 뽑힌 순서를 고려 여부
  - 순서 고려 O: (1, 2)와 (2, 1)을 다른 것
  - 순서 고려 X: (1, 2)와 (2, 1)을 같은 것
  - Q? 뽑은 것을 크기 순서대로 정렬(sorting)한다고 하면?

추출 / 배열	복원	비복원
순서고려	Ⓐ	Ⓑ
순서무시	Ⓒ	Ⓓ

Ⓑ 순열(permutation)

Ⓐ 중복순열

Ⓓ 조합(combination)

Ⓒ 중복조합

◦ 순열은 각 단계에서 선택할 수 있는 수가 하나씩 줄어듬

$$n\times(n-1)\times\ \cdots\ \times(n-k+1)=\frac{n!}{(n-k)!}$$

◦ 중복순열의 수는 매번 n개 선택 가능

$$n\times\ \cdots\ \times n=n^k$$

◦ 순서를 고려하지 않는다면 같은 번호로 이루어진 순열이 같은 것으로 취급 ⇨ 선택된 k개의 공의 순서열

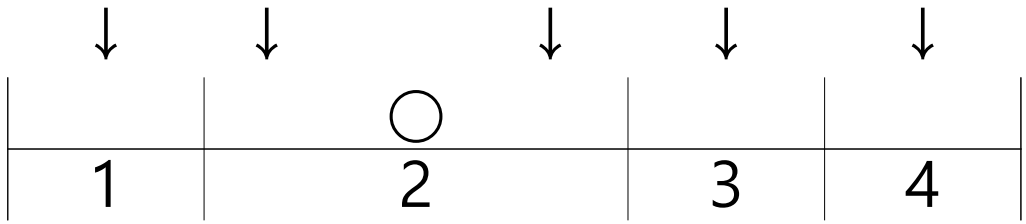
$$k\times(k-1)\times\ \cdots\ \times1=k!$$

순서를 고려하지 않는 조합의 수는

$$\frac{n\times(n-1)\times\ \cdots\ \times(n-k+1)}{1\times2\times\ \cdots\ \times k}=\frac{n!}{k!\,(n-k)!}=\binom{n}{k}$$

○ 중복조합의 경우, $n^k/k!$?

- $\{1,2,3,4\}$ 에서 2개를 선택하는 중복조합은 10개: (1,1), (1,2), (1,3), (1,4), (2,2), (2,3), (2,4), (3,3), (3,4), (4,4)
- $4^2/2! = 8 \neq 10$
- n 개의 선택 가능한 공간에 k 개의 공을 넣는 실험



- 첫 번째는 4가지 두 번째는 5가지 $4 \times 5 = 20$
- 순서열의 수 $2!$ ⇨ $\dfrac{5 \times 4}{2!} = \dfrac{5!}{2!\,3!} = \binom{5}{2} = 10$

- 만약 3개를 선택한다면 세 번째 공의 화살표는 6개가 되어 $6\times5\times4$가 되고 선택된 3개의 순서열의 수는 $3!$

$$\frac{6\times5\times4}{3!}=\frac{6!}{3!3!}=\binom{4+2}{3}=20$$

- 중복조합의 경우의 수

$$\frac{n(n+1)\times\cdots\times(n+k-1)}{k!}=\frac{(n+k-1)!}{k!(n-1)!}=\binom{n+k-1}{k}$$

【표 3.2】 경우의 수

배열 \ 추출	복원	비복원
순서고려	n^k	$\frac{n!}{(n-k)!}$
순서무시	$\binom{n+k-1}{k}$	$\binom{n}{k}$

◉ 나눔Lotto 6/45

- 1에서 45까지의 숫자에서 6개의 번호를 비복원 추출하고 크기순서대로 정렬한 당첨번호를 제공
- 1등: 선택한 6개의 번호가 당첨번호와 모두 일치
- 5개만 일치하면 2등 또는 3등
- 전체 가능한 경우의 수

$$\#(\Omega) = \binom{45}{6} = \frac{45\times44\times43\times42\times41\times40}{6\times5\times4\times3\times2\times1} = 8,145,060$$

$$\Rightarrow\ P(\text{1등}) = 1/8145060$$

- 2등과 3등: 순서와 관계없이 6개 당첨번호 중 5개를 선택하고 나머지 하나는 다른 39개 중 하나를 선택

$$\#(\text{2등 또는 3등}) = \binom{6}{5} \times 39 = 6 \times 39 = 234$$

$$\Rightarrow P(\text{2등 또는 3등}) = 234/8145060 = 1/34807.95$$

◉ Birthday problem

- 1년을 365일이고 **365일 동안 태어날 가능성**이 동일
- A: k명의 사람이 모두 다른 생일을 가지는 사건
- k명의 사람이 각각 365일 중 한 날을 선택할 수 있으므로

$$\#(\Omega) = 365^k$$

- 모든 사람이 생일이 다르다는 것은 각기 다른 날짜를 선택하는 것으로 365일을 k번 비복원 추출하는 방법

$$\#(A) = 356 \times 364 \times \cdots \times (365-k+1) = \frac{365!}{(365-k)!}$$

$$\Rightarrow P(A) = \frac{365!/(365-k)!}{365^k} = \frac{365!}{365^k(365-k)!} = \prod_{j=0}^{k-1}\left(1-\frac{j}{365}\right)$$

k	5	10	15	20	30	40	50
$P(A)$	0.9729	0.8831	0.7471	0.5886	0.2937	0.1088	0.0296

○ 연속표본공간

◦ 발생가능성이 동일한 상황을 선이나 평면 등을 이용

◦ 사건 A가 발생한다는 것은, 영역 Ω내에서 임의의 한 점을 무작위로 택할 때 이 점이 영역 A에 있다는 의미

◦ 사건 A의 확률은 전체 영역에서 A가 차지하는 비율

$$P(A) = \frac{\|A\|}{\|\Omega\|}$$

- $\|\cdot\|$는 길이, 면적, 부피 등을 의미

□ 상대도수의 극한개념

- ○ 동전의 앞면이 나올 확률은 1/2?
- ○ 고전적 확률: "앞면과 뒷면의 발생가능성이 동일하고 $\Omega = \{H, T\}$, $A = \{H\}$ 이므로 $P(A) = 1/2$ 이다"라고 해석
- ○ 동전던지기 실험

실험자	던진 횟수	앞면	상대도수
Buffon	4040	2048	0.5080
Pearson	12000	6019	0.5016
Pearson	24000	12012	0.5005

- 실험을 계속 수행하면 상대도수가 0.5로 수렴

○ 윷을 하나 던졌을 때 평면이 위로 나타날 확률은?

- 윷을 n번 던졌을 때 평면이 나온 횟수를 $n(A)$라면

$$P(A) \simeq \frac{n(A)}{n}$$

○ 만약 실험을 무한히 반복한다면 $n(A)/n$은 어떤 값으로 수렴하는데 이 극한값을 사건 A가 일어날 확률을 정의

$$P(A) = \lim_{n \to \infty} \frac{n(A)}{n}$$

○ "그냥도전 동전 돌리기"에서 각각의 실험에서 발생하는 결과는 표본이고 **실험을 무한히 반복**한다는 것은 **표본이 결국에는 모집단**이 됨

- **확률은 모집단**이 어떻게 형태로 이루어져 있는지를 보여줌
- 상대도수의 극한으로써의 확률은 많은 표본을 통해 모집단의 특성을 파악하는 방식을 따른다고 해서 **통계적 확률(statistical probability)**이라고 함
- 일기예보나 전쟁게임과 같이 모의실험(simulation)을 통해 결과를 도출하는 분야에서 많이 사용되고 있으며 고속 컴퓨터의 보급으로 인해 더욱더 활용도가 높아지고 있음

◉ 생일문제

○ 365일 각각의 날에 태어날 가능성이 동일?

상황	1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
Ⓐ	1452	1450	1394	1337	1271	1273	1294	1335	1390	1348	1283	1106
Ⓑ	2000	2000	2000	2000	2000	2000	1000	1000	1000	1000	1000	1000

○ Ⓐ: 2012년에 태어난 484,300명의 아이들의 월별 하루 평균출생아수

○ Ⓑ: 1월에서 6월까지의 일별 출생아가 7월에서 12월까지 일별 출생아의 두 배 가정

○ 각각의 k에 대해 일억 번 실시한 비율

【표 3.3】 상황에 따른 생일문제의 확률 추이

k	5	10	15	20	30	40	50
$P(A)$	0.9729	0.8831	0.7471	0.5886	0.2937	0.1088	0.0296
Ⓐ	0.9728	0.8825	0.7464	0.5870	0.2922	0.1076	0.0292
Ⓑ	0.9699	0.8709	0.7230	0.5551	0.2569	0.0853	0.0202

□ 공리적 확률

- ○ 콜모고로프(A. N. Kolmogorov, 1903-1987)
- ○ 확률의 공리는 확률이론의 기반
- ○ $P(\cdot)$: **확률측도(probability measure)**

[공리 1] $P(\Omega) = 1$

[공리 2] $0 \le P(A) \le 1, \quad A \subset \Omega$

[공리 3] 서로배반인 사건 $A_1, A_2, \ldots, A_n$에 대해,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$$

■ 확률의 기본정리

① $P(A^c) = 1 - P(A)$

$$1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c)$$

- ○ $P(\varnothing) = 0$
- ○ 생일문제: k명 중 적어도 두 사람 이상이 같은 생일을 가지는 사건은 A^c

$$P(A^c) = 1 - P(A) = 1 - \frac{365!}{365^k (365-k)!} = 1 - \prod_{j=0}^{k-1} \left(1 - \frac{j}{365}\right)$$

◉ 1000장 중 4장이 당첨복권

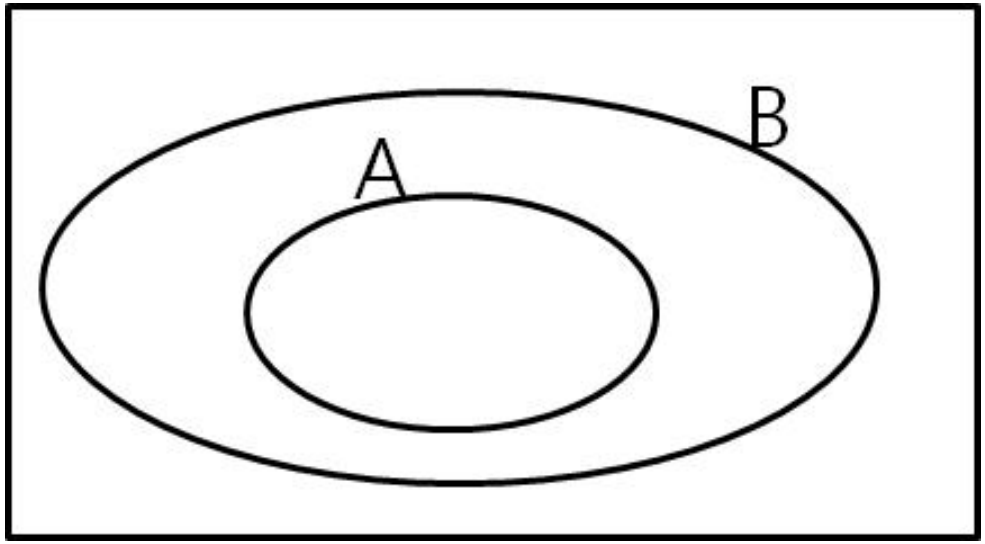
- 4장의 복권을 구입한다면 적어도 한 장 이상의 당첨복권을 구입하게 될 확률은?
- A: 한 장 이상의 당첨복권을 구입할 사건 = 당첨복권이 한 장, 두 장, 세 장, 네 장인 경우

  ⇨ A^c: 구입한 4장 모두 당첨되지 않을 사건

$$P(A^c) = \frac{996 \times 995 \times 994 \times 993}{1000 \times 999 \times 998 \times 997} = 0.9841$$

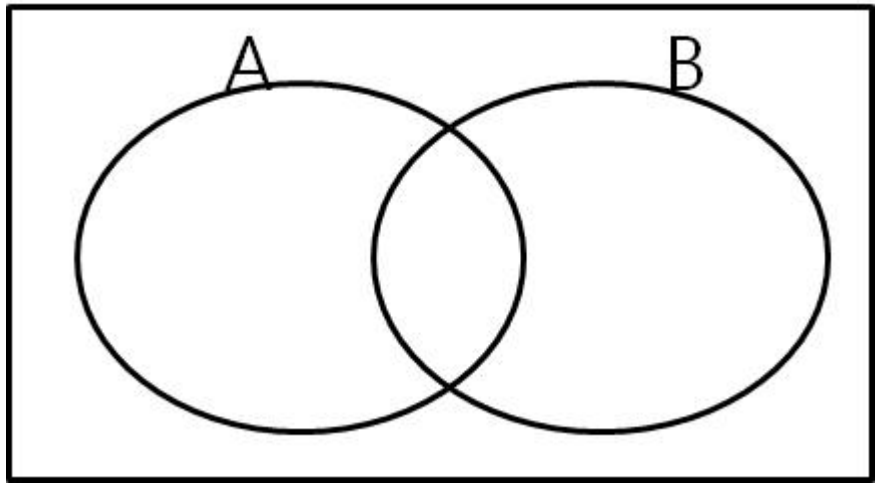
$$\Rightarrow\ P(A) = 1 - P(A^c) = 0.0159$$

② $A \subset B$이면 $P(A) \leq P(B)$



◦ $B = A \cup (B \cap A^c)$

③ $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



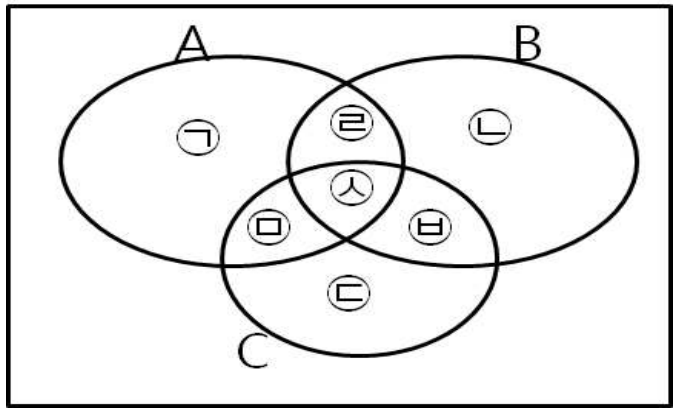
◦ $A = (A \cap B) \cup (A \cap B^c)$

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

$$P(B) = P(A \cap B) + P(A^c \cap B)$$

$$P(A \cup B) = P(A \cap B^c) + P(A \cap B) + P(A^c \cap B)$$

◦ 사건 $A,\ B,\ C$



$$P(A\cup B\cup C)=P(㉠)+P(㉢)+P(㉦)+P(㉡)+P(㉤)+P(㉥)+P(㉣)$$

- $P(㉠)=P(A)-P(A\cap B)-P(A\cap C)+P(A\cap B\cap C)$
- $P(㉡)=P(A\cap B)-P(A\cap B\cap C)$
- $P(㉣)=P(A\cap B\cap C)$

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

○ n 개의 사건 $A_1, A_2, \ldots, A_n$ 에 대해

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{1 \le i < j \le n} P(A_i \cap A_j)$$

$$+ \cdots + (-1)^{n-1} P(A_1 \cap A_2 \cap \cdots \cap A_n)$$

④ $P(A \cup B) \leq P(A) + P(B)$

◦ 부울의 부등식(Boole's inequality): $P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)$

◦ $P(A^c \cup B^c) \leq P(A^c) + P(B^c)$

⇨ $P(A^c \cup B^c) = 1 - P(A \cap B) \leq 1 - \{P(A) + P(B)\}$

◦ 본페로니 부등식(Bonferroni's inequality)

$$P\left(\bigcap_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n} P(A_i) - (n-1)$$

◉ 1000장 중 4장이 당첨복권

- A_i: i번째 복권이 당첨될 사건 ⇨ $P(A_i) = 0.004$
- $A = \bigcup_{i=1}^{4} A_i$로 표시

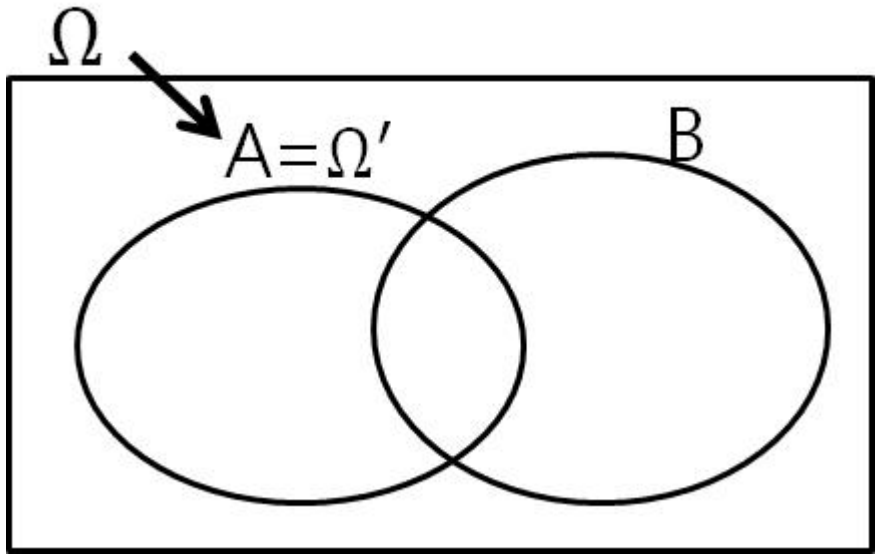
$$P(A) = P\left(\bigcup_{i=1}^{4} A_i\right) = 0.0159 \leq \sum_{i=1}^{4} P(A_i) = 0.016$$

■ 조건부확률

◉ 동전 두 개를 던지는 실험에서 어떤 한 동전이 앞면이라는 것을 알았을 때, 두 동전 모두 앞면일 사건의 확률은?

- ○ $\Omega = \{HH,\ TH,\ HT,\ TT\}$
- ○ 어떤 한 동전이 앞면이라는 정보가 추가로 주어지면 표본공간에서 $\{TT\}$가 발생할 수 없음
  - ⇨ 표본공간은 $\{HH, TH, HT\}$로 축소

○ **조건부 확률(conditional probability)**: 확률실험에서 새로운 정보 또는 조건이 추가되었을 때 사건의 확률



○ 사건 A가 발생했다면 A 이외의 것은 일어날 수 없기 때문에 A가 새로운 표본공간 Ω'이 되고, B가 발생한다는 것은 A 안에서 $A \cap B$에 있는 원소가 발생하는 것을 의미

◦ A하에서 B의 조건부확률은 A에서 $A \cap B$가 차지하는 비율

◦ 사건 A가 주어졌을 때 사건 B의 조건부 확률은 $P(B|A)$로 표시

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

○ 사망률(mortality rate)

◦ 어느 해의 40대 사망률: 그해 **40대 이상인 사람들 중**에서 40대에 사망한 사람의 비율

- 표본공간이 전체 연령대에서 40대 이상으로 축소

◦ 생존율(survival rate): 40대 이상인 사람 중 그 해 생존한 사람의 비율로 (1-사망률)로 계산

◉ 완전생명표

○ 통계청에서 발표한 2012년 자료의 일부분

○ 인구 10만 명에서 시작하여 각 연령까지 생존한 사람의 수

【표 3.4】 완전생명표

연령	생존자		
	전체	남자	여자
0세	100,000	100,000	100,000
1세	99,709	99,686	99,733
40세	98,158	97,727	98,619
41세	98,048	97,581	98,546
60세	92,679	89,823	95,657
61세	92,146	89,046	95,379
80세	64,812	53,265	75,732
81세	61,712	49,691	72,910

◦ 0세 사망자는 10만 중 100000-99709=291명

⇨ 영아 사망률= 291/100000=0.00291(0.29%)

◦ 40세의 남성사망률은 40세 이상 남자 생존자 97727명 중 40세에 사망한 97727-97581=146명

⇨ 40세 남성사망률 $= \frac{146}{97727} = 0.00149 = 0.15\%$

◦ 80세 여성 생존율은 1-2822/75732=1-0.0373=0.967로 96.7%

【표 3.5】 연령별 사망자와 사망률

연령	구분	전체	남자	여자
0세	사망자	291	314	267
	사망률	0.29%	0.31%	0.27%
20세	사망자	33	44	21
	사망률	0.03%	0.04%	0.02%
40세	사망자	110	146	73
	사망률	0.11%	0.15%	0.07%
60세	사망자	533	777	278
	사망률	0.58%	0.87%	0.29%
80세	사망자	3,100	3,574	2,822
	사망률	4.78%	6.71%	3.73%

□ 조건부확률의 활용

① $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$

- ○ 곱사건이 순차적인 사건들의 조건부확률의 곱으로 표시

◉ 정상 제품 90개와 불량품 10개가 들어있는 상자에서 무작위로 2개를 비복원으로 추출

- ○ $\Omega = \{(정상_1, 정상_2), (정상_1, 불량_2), (불량_1, 정상_2), (불량_1, 불량_2)\}$
- ○ (정상1, 정상2)의 확률은?
  - 첫 번째가 정상일 확률은 90/100
  - 두 번째**도** 정상일 확률은 89/99

- $P(\text{정상}_1, \text{정상}_2) = \frac{90}{100} \times \frac{89}{99} = \frac{89}{110}$

$P(\text{정상}_1)$ ↵ ↳ $P(\text{정상}_2 | \text{정상}_1)$

$$P(\text{정상}_1, \text{불량}_2) = P(\text{정상}_1)P(\text{불량}_2 | \text{정상}_1) = \frac{90}{100} \times \frac{10}{99} = \frac{10}{110}$$

$$P(\text{불량}_1, \text{정상}_2) = P(\text{불량}_1)P(\text{정상}_2 | \text{불량}_1) = \frac{10}{100} \times \frac{90}{99} = \frac{10}{110}$$

$$P(\text{불량}_1, \text{불량}_2) = P(\text{불량}_1)P(\text{불량}_2 | \text{불량}_1) = \frac{10}{100} \times \frac{9}{99} = \frac{1}{110}$$

○ A_1, A_2, A_3에 대해 $P(A_1 \cap A_2) > 0$이면

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)\frac{P(A_1 \cap A_2)}{P(A_1)}\frac{P(A_1 \cap A_2 \cap A_3)}{P(A_1 \cap A_2)}$$

$$\Rightarrow\ P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)$$

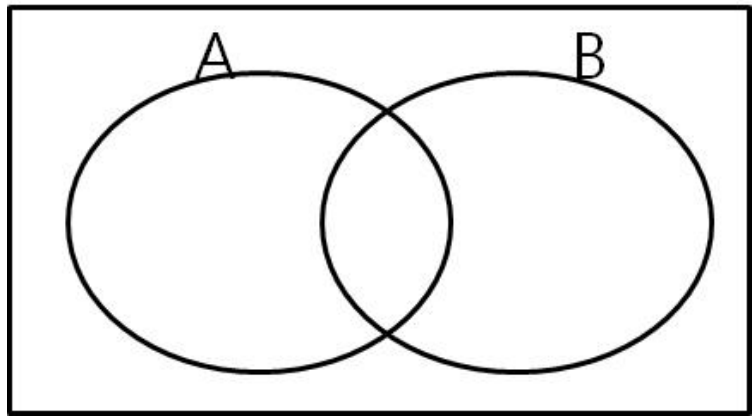
○ 수학적 귀납법을 이용하면, $P(A_1 \cap \cdots \cap A_{n-1}) > 0$

$$P(A_1 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)\cdots \times P(A_n|A_1 \cap \cdots \cap A_{n-1})$$

◉ 당첨복권이 4장인 복권 1000장 발매

- ○ A_i: 구입한 4장의 복권 중에서 i번째 복권이 당첨될 사건
- ○ $P(A_i) =$?
  - $P(A_1) = 0.004$

② $P(B) = P(A \cap B) + P(A^c \cap B)$
$= P(A)P(B|A) + P(A^c)P(B|A)$



○ $P(A_2) = P(A_1)P(A_2|A_1) + P(A_1^c)P(A_2|A_1^c)$
$= \frac{4}{1000}\frac{3}{999} + \frac{996}{1000}\frac{4}{999} = \frac{4}{1000} = 0.004$

○ $$\begin{aligned} P(A_3) = & \; P(A_1 \cap A_2 \cap A_3) + P(A_1^c \cap A_2 \cap A_3) \\ & + P(A_1 \cap A_2^c \cap A_3) + P(A_1^c \cap A_2^c \cap A_3) \\ = & \; P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \\ & + P(A_1^c)P(A_2|A_1^c)P(A_3|A_1^c \cap A_2) \\ & + P(A_1)P(A_2^c|A_1)P(A_3|A_1 \cap A_2^c) \\ & + P(A_1^c)P(A_2^c|A_1^c)P(A_3|A_1^c \cap A_2^c) = 0.004 \end{aligned}$$

○ 어떤 일련의 사건들이 순차적 또는 결합된 형태로 되어 있는 상황에서 특정 시점이나 위치에 해당되는 사건의 확률은 앞에서 발생할 수 있는 상황이나 연결된 상황들의 확률을 모두 더하여 구할 수 있음

◉ 스팸메일 필터

- 어떤 메일시스템의 수신메일 중 40%가 스팸메일(S)이고 나머지는 정상메일(N)
- 스팸메일 중 내용에 "A"라는 단어가 있는 메일은 25%이고 정상메일 중 이 단어가 있는 경우는 2%
- 전체 메일 중 "A" 단어를 포함한 메일의 비율은?

$$P(A) = P(S \cap A) + P(N \cap A) = P(S)\,P(A|S) + P(N)P(A|N)$$

- $P(S) = 0.4,\ P(N) = 0.6,\ P(A|S) = 0.25,\ P(A|N) = 0.02$

$$P(A) = 0.4 \times 0.25 + 0.6 \times 0.02 = 0.1 + 0.012 = 0.112$$

○ 확률수형도(probability tree)

○ 표본공간의 분할(partition)

◦ 사건 $A_1, \ldots, A_n$가

Ⓐ 서로배반사건, 즉 모든 $i \neq j$에 대해 $A_i \cap A_j = \varnothing$

Ⓑ 전체를 이루는 사건(exhaustive), 즉 $A_1 \cup \cdots \cup A_n = \Omega$

이면, 사건 $A_1, \ldots, A_n$을 표본공간 Ω의 분할이라고 함

◦ 사건 $A_1, \ldots, A_n$이 표본공간 Ω의 분할이면

$$P(B) = P(B \cap \Omega) = P((B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_n))$$

$$= P(B \cap A_1) + \cdots + P(B \cap A_n)$$

$$= P(A_1)P(B|A_1) + \cdots + P(A_n)P(B|A_n)$$

○ 베이즈정리(Bayes' theorem)

◦ $P(B|A)$은 순서적으로 볼 때, 대부분 사건 A가 먼저 발생하고 B가 이어 발생하는 상황에 대한 확률

- A는 원인, B는 결과의 형태를 가짐
- 원인의 가능성인 $P(A)$ 또는 $P(A^c)$는 사건 B가 관측되기 이전의 확률

⇨ **사전확률(prior probability)**

○ 어떤 문제에서는 결과를 얻은 상태에서 그 결과가 발생하게 된 원인을 역으로 추정 ⇨

사례-대조연구(case-control study)

- 결과 B의 관측했을 때 그 원인이 A일 사건의 확률은?

$$P(A|B)$$

- 이 확률을 사건 B가 관측된 후의 A의 확률이라고 해서 **사후확률(posterior probability)**이라고 함

◉ 암진단

- 암에 대한 간이진단 검사를 실시하는데 암에 걸렸을 때 양성반응이 나올 확률은 0.96이고 암에 걸리지 않았을 때 양성반응이 나올 확률이 0.05
- 만약 이 검사에서 양성반응이 나왔다면?
- 확률적 표현: A를 암에 걸린 사건
  - $P(+|A) = 0.96, \quad P(+|A^c) = 0.05$
  - 양성반응이 나왔을 때 암에 걸렸을 확률은 $P(A|+)$

$$P(+|A) \neq P(A|+)$$

◦ 베이즈(Thomas Bayes, 1701-1973)

◦ 만약 $P(B) > 0$, $P(A|B) = P(A \cap B)/P(B)$

- $P(A) > 0$, $P(A^c) > 0$이면, ①과 ②에 의해

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)\,P(B \mid A)}{P(A)\,P(B \mid A) + P(A^c)\,P(B \mid A^c)}$$

◦ $P(A|+)$를 계산하기 위해서는 $P(A)$를 알아야 함

- 만약 $P(A) = 0.01$이라고 하면

$$P(A|+) = \frac{P(A)P(+|A)}{P(A)P(+|A) + P(A^c)P(+|A^c)}$$

$$= \frac{0.01 \times 0.96}{0.01 \times 0.96 + 0.99 \times 0.05} = \frac{0.0096}{0.0591} = 0.1624$$

- $P(+|A) = 0.96$과는 상당한 차이

○ 베이즈 정리의 일반식

- 사건 $A_1, ..., A_n$은 표본공간 Ω의 분할
- 모든 i에 대해 $P(A_i) > 0$이면

$$P(A_k|B) = \frac{P(A_k)P(B|A_k)}{P(B)} = \frac{P(A_k)\,P(B|A_k)}{\sum_{i=1}^{n} P(A_i)\,P(B|A_i)}$$

◉ 스팸메일 필터

○ 수신메일 내용에 "A"라는 단어가 있을 때 이 메일이 스팸메일일 확률은?

- 확률식: $P(S|A)$

$$P(S|A) = \frac{P(S \cap A)}{P(A)}$$

- $P(A) = 0.112$
- $P(S \cap A) = P(S)P(A|S) = 0.4 \times 0.25 = 0.1$

$$\Rightarrow \; P(S|A) = \frac{P(S \cap A)}{P(A)} = \frac{0.1}{0.112} = 0.8929$$

■ 독립사건(independent events)

- $P(A) > 0$이고 $P(B) > 0$이면

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

- 만약 사건 A가 사건 B의 발생에 영향을 주지 않고 B가 사건 A에 영향을 주지 않는다면,

$$P(B|A) = P(B), \quad P(A|B) = P(A)$$

- 사건 A와 B가 서로 영향을 주고받지 않는 경우, "사건 A와 B는 **독립사건(independent events)**이다."라고 함

$$\rightleftarrows \; P(A \cap B) = P(A)\,P(B)$$

- 표본공간과 공집합은 임의의 사건 A와 독립

$$P(\Omega \cap A) = P(A) = P(\Omega)P(A)$$

$$P(\varnothing \cap A) = P(\varnothing) = P(\varnothing)P(A)$$

◉ 두 개의 정육면체 주사위

- A: 두 주사위의 합이 6인 사건
- B: 두 주사위의 합이 7인 사건
- C: 첫 번째 주사위의 눈이 3인 사건

$$A = \{(1,5),(2,4),(3,3),(4,2),(5,1)\}$$

$$B = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$$

$$C = \{(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)\}$$

- A와 C는 독립인가? B와 C는 독립인가?

◦ 셋 사건 A, B, C에 대해 다음의 네 등식이 모두 성립하면 셋 사건 A, B, C는 독립사건 또는 서로 독립적(mutually independent)이라 함

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

$$P(A \cap C) = P(A)\,P(C)$$

$$P(B \cap C) = P(B)\,P(C)$$

$$P(A \cap B \cap C) = P(A)\,P(B)\,P(C)$$

◉ 주사위 3개를 던지기

- 6이 최소한 한번 이상 나올 확률은?
- A: 주사위 눈이 6인 경우가 최소한 한번 이상 나올 사건
- A_i: i번째 주사위 눈이 6인 사건

$$P(A) = 1 - P(A^c) = 1 - P(A_1^c \cap A_2^c \cap A_3^c)$$

◉ 전기전달시스템

- ○ 세 개의 ON/OFF 스위치로 구성
- ○ 스위치 A, B, C가 ON일 확률은 각각 0.7, 0.8, 0.6
- ○ 각각의 스위치는 독립적으로 세팅
- ○ A와 B는 직렬, C와 A는 병렬로 구성

◦ 시스템이 전기를 전달할 사건은 $C \cup (A \cap B)$

$$\begin{aligned} P(\text{전기전달}) &= P(C \cup (A \cap B)) \\ &= P(C) + P(A \cap B) - P(A \cap B \cap C) \end{aligned}$$

- 각각의 스위치가 독립적으로 세팅

$$\begin{aligned} P(\text{전기전달}) &= P(C) + P(A)\,P(B) - P(A)\,P(B)\,P(C) \\ &= 0.6 + 0.7 \times 0.8 - 0.7 \times 0.8 \times 0.6 = 0.824 \end{aligned}$$

○ 용어 및 정의

- 확률실험(random experiment)
- 표본공간(sample space)
- 사건(event)
- 서로배반사건(disjoint, mutually exclusive)
- 발생가능성 동일(equally likely)
- 복원(with replacement) & 비복원(without replacement)
- 통계적 확률(statistical probability)
- 조건부 확률(conditional probability)
- 분할(partition) - exhaustive
- 독립사건(independent events)

○ 정리(theorem)

- 경우의 수: 곱의 법칙
- 확률의 공리
- $P(A^c) = 1 - P(A)$
- $A \subset B$이면 $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$
- $P(A \cup B) \leq P(A) + P(B)$
- $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$

- $P(B) = P(A \cap B) + P(A^c \cap B)$
  $= P(A)P(B|A) + P(A^c)P(B|A)$
- 베이즈정리

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)\,P(B \mid A)}{P(A)\,P(B \mid A) + P(A^c)\,P(B \mid A^c)}$$

확률변수와 확률분포
(Random Variable
& Probability Distribution)

■ 확률변수(random variable)

- **표본공간에서 정의된 실수 함수**
- 불확실성을 가지는 사회적·자연적 현상을 일종의 확률실험으로 이해
- 여기서 얻어진 표본공간을 숫자로 표시하여 불확실한 현상을 수학적으로 모형화 함
- 이를 통해 구체적으로 계량화된 분석을 할 수 있음

◉ 동전 3개 던지기

◦ X: 앞면의 수, Y: 앞면과 뒷면의 수의 차이

$$\Omega = \{\ HHH,\ HHT,\ HTH,\ THH,\ HTT,\ THT,\ TTH\ \ TTT\ \}$$

	↓	↓	↓	↓	↓	↓	↓	↓
$X=$	3	2	2	2	1	1	1	0
$Y=$	3	1	1	1	1	1	1	3

- 확률변수는 정의역이 표본공간 Ω이고 공역이 실수인 함수

- 표본공간의 임의의 원소 ω에 대해 원칙적으로 $X(\omega)$와 같이 표시해야 하지만 편의상 (ω) 표시를 생략
- 통계학에서는 일반적으로 확률변수를 대문자 $X,\ Y,\ Z$ 등으로 표시하며 확률변수가 취하는 값을 소문자 $x,\ y,\ z$ 등으로 표시

◉ 동전을 앞면이 나올 때까지 던지는 확률실험

◦ X: 동전을 던지 횟수, Y: 앞면이 나올 때까지의 뒷면의 수

$$\Omega = \{\ H,\ TH,\ TTH,\ TTTH,\ TTTTH,\ \cdots\ \}$$

	↓	↓	↓	↓	↓	
$X=$	1	2	3	4	5	$\cdots$
$Y=$	0	1	2	3	4	$\cdots$

- 확률변수는 변수가 취하는 값에 따라 이산확률변수와 연속확률변수로 나눔
- **이산확률변수(discrete random variable)**: 확률변수가 가질 수 있는 값들이 가산(countable) 또는 셀 수 있는 경우
  - '가산' 또는 '셀 수 있다'는 말은 확률변수의 값들이 자연수 1, 2, 3, ...과 대응관계를 가진다는 뜻
  - 예) 불량품의 개수, 사고건수,...
- **연속확률변수(continuous random variable)**: 가질 수 있는 값이 셀 수 없을 정도로 많은 경우
  - 예) 수명, 신장, 체중
- 이산형과 연속형의 구분이 명확하지 않는 경우, 가정의 적절성이나 분석의 난이도 등을 고려하여 적절하게 선택

■ 확률분포(Probability Distribution)

○ 확률변수는 표본공간의 값을 숫자로 바꾼 함수이기 때문에 확률변수가 어떤 값을 가진다는 것은 표본공간 내에 대응하는 원소들이 존재

- $X = x$ 이면 표본공간에 $\{\omega \mid X(\omega) = x,\ \omega \in \Omega\}$ 를 만족하는 사건이 존재
- 임의의 상수 a, b 에 대해 $a \le X \le b$ 이면 이에 해당하는 사건 $\{\omega \mid a \le X(\omega) \le b,\ \omega \in \Omega\}$ 이 존재

⇨ 이는 확률변수에 대해 $X = x$ 또는 $a \le X \le b$ 에 대응하는 확률을 계산할 수 있음

◉ 동전을 세 번 던지기

$$P(X=0) = P(\{TTT\}) = \frac{1}{8}$$

$$P(X=1) = P(\{HTT, THT, TTH\}) = \frac{3}{8}$$

$$P(X=2) = P(\{HHT, HTH, THH\}) = \frac{3}{8}$$

$$P(X=3) = P(\{HHH\}) = \frac{1}{8}$$

- 표본공간에서 사건의 확률은 단순히 확률
- 확률변수는 숫자로 표시되어 특정 지점이나 영역에서의 확률을 표시할 수 있어 확률이 어떤 형태로 분포되었다는 말을 할 수 있음 ⇨ 그림으로 표시가능
- 확률변수가 가질 수 있는 값에 대해 확률을 표시한 것을 **확률분포(probability distribution)**라고 함
- 확률분포표(probability distribution table): 확률변수의 확률을 표로 표시한 것
  - 예) 동전 세 번 던지기: 앞면의 수 X

x	0	1	2	3
$P(X=x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

○ 확률은 모집단이 어떤 형태로 이루어져 있는지를 보여줌

⇨ 확률분포 또한 모집단을 숫자로 표시했을 때의 형태를 표시한 것 = 모집단의 확률구조

○ 모집단의 확률구조를 표시하는 방법

- 이산확률변수: **확률질량함수**, 누적분포함수, …
- 연속확률변수: **확률밀도함수**, 누적분포함수, …

□ 확률질량함수(probability mass function, pmf)

○ 이산확률변수 X가 임의의 값 x일 확률 $P(X=x)$를 x에 대한 함수로 생각

$$f(x)=P(X=x)$$

○ 경우에 따라 확률변수 X를 강조하기 위해 $f_X(x)$로 표시

◉ 동전 세 번 던지기

◦ X: 앞면의 수 ⇨ X가 가질 수 있는 값은 $x = 0,\ 1,\ 2,\ 3$

$$f(0) = \frac{1}{8},\quad f(1) = \frac{3}{8},\quad f(2) = \frac{3}{8},\quad f(3) = \frac{1}{8}$$

◦ Y: 앞면과 뒷면의 수의 차이 ⇨ $y = 1,\ 3$

$$f_Y(1) = \frac{6}{8} = \frac{3}{4},\quad f_Y(3) = \frac{2}{8} = \frac{1}{4}$$

◉ 앞면이 나올 때까지 동전을 던지기

◦ X: 던진 횟수

$$f(1) = P(X=1) = P(\{H\}) = \frac{1}{2}$$

$$f(2) = P(X=2) = P(\{TH\}) = \frac{1}{4} = \left(\frac{1}{2}\right)^2$$

$$f(3) = P(X=3) = P(\{TTH\}) = \frac{1}{8} = \left(\frac{1}{2}\right)^3$$

$$\vdots$$

$$\Rightarrow \quad f(x) = \left(\frac{1}{2}\right)^x, \qquad x = 1, 2, 3, \ \ldots$$

- **기하분포(geometric distribution)**

○ Y: 뒷면의 수

- $Y = X - 1$의 관계를 가지며 해당 확률은 동일

$$f(0) = \frac{1}{2}, \quad f(1) = \frac{1}{4} = \left(\frac{1}{2}\right)^2, \quad f(2) = \frac{1}{8} = \left(\frac{1}{2}\right)^3, \ \cdots$$

$$\Rightarrow \ f_Y(y) = \left(\frac{1}{2}\right)^{y+1}, \quad y = 0, 1, 2, \ \ldots$$

○ 확률질량함수의 성질

◦ $f(x)$는 $X=x$일 때의 확률이기 때문에, X가 가질 수 있는 값이 $x_1, x_2, x_3, \ldots$이면

① 모든 $i=1,2,\ldots$에 대해 $0 \le f(x_i) \le 1$

② $\sum_{i=1}^{\infty} f(x_i) = 1$

③ $P(a \le X \le b) = \sum_{x_i \in [a,b]} f(x_i)$

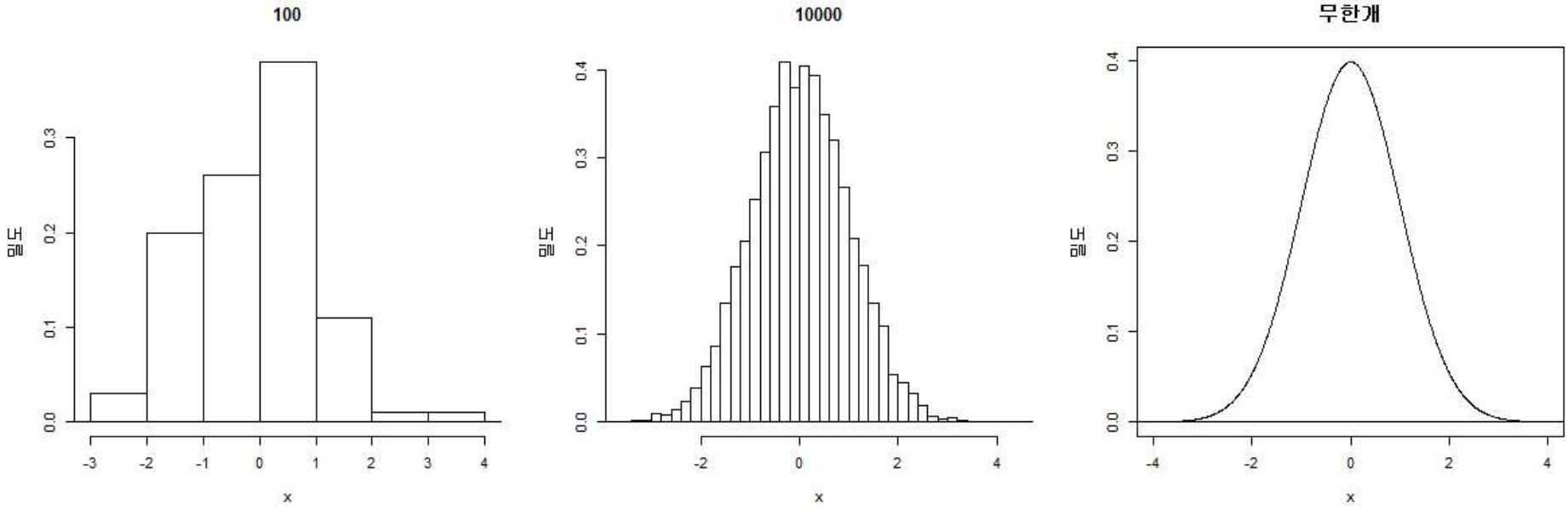
○ 확률변수의 변환(transformation)

- ○ 확률변수의 함수
  - 예) $Z = X^2$, $W = (X - 1.5)^2$
- ○ 함수의 함수도 함수 ⇨ 확률변수의 함수도 확률변수
  - 예) 동전 세 번 던지기

x	0	1	2	3
z	0	1	4	9
w	2.25	0.25	0.25	2.25
$P(Z=z)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

$$P(W = 0.25) = 3/4, \quad P(W = 2.25) = 1/4$$

□ 확률밀도함수(probability density function)

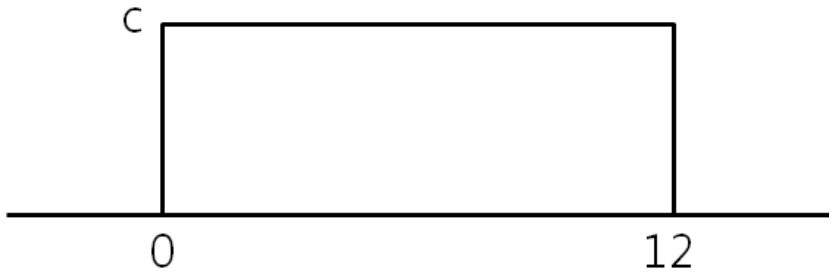


○ 세 번째 그림은 연속확률변수 X의 분포형태 모집단의 형태를 나타낸 것으로 임의의 지점 x에서의 밀도를 $f(x)$라고 표시하면 $f(x)$를 확률밀도함수라고 함

◉ 0~12까지의 숫자가 표시된 돌림판

- 표본공간: $\Omega = \{x : 0 < x \leq 12\}$
- X: 바늘이 지적하는 위치
- 0에서 12사이에서 발생가능성이 동일

  ⇨ 밀도는 이 구간에서 동일 : $f(x) = c$



- 전체 면적은 1이 되어야 하므로 $c = 1/12$

$$f(x) = \frac{1}{12}, \qquad 0 < x \leq 12$$

○ 확률밀도함수에서의 확률

- ○ 히스토그램에서 면적이 해당 구간에서의 비율(상대도수)
- ○ 확률밀도함수에서의 면적이 해당 구간에서의 확률
- ○ X가 구간 $[a, b]$에 속할 확률

$$P(a \le X \le b) = \int_a^b f(x)\,dx$$

- 예) X가 3에서 6사이에 있을 확률

$$P(3 \le X \le 6) = \frac{3}{12} = \frac{1}{4}$$

- ○ $X = 3$일 확률은?

- 어떤 점에서는 면적은 $f(x)$의 크기와 관계없이 항상 0
- X가 연속확률변수일 때에는 모든 x에 대해 $P(X=x)=0$
- 확률밀도함수 $f(x)$는 x에서의 확률이 아니라 상대적인 밀도를 나타내는 것
- X가 연속확률변수이면

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

○ 확률밀도함수의 성질

◦ 임의의 연속확률변수 X의 확률밀도함수 $f(x)$는

① 모든 x에 대해 $f(x) \geq 0$

② $\int_{-\infty}^{\infty} f(x)\,dx = 1$

③ $P(a < X \leq b) = \int_{a}^{b} f(x)\,dx$

◦ **누적분포함수(cumulative distribution function)**: ③의 특별한 형태

$$P(X \leq x) = \int_{-\infty}^{x} f(u)\,du = F(x)$$

■ 기댓값(expectation, expected value)

◉ 표본평균

- ○ 임의로 5개의 표본을 선택: 1, 1, 2, 5, 6
- ○ 표본평균

$$\bar{x} = \frac{1+1+2+5+6}{5}$$

$$= 1\times\frac{2}{5}+2\times\frac{1}{5}+5\times\frac{1}{5}+6\times\frac{1}{5}=3$$

⇨ 관측된 값에 자료 중 그 값이 차지하는 비율을 곱하여 더한 것으로 표시

○ 전체 표본이 n개 있고 자료 중 서로 다른 값이 k개가 있어 이들 값을 $x_1, \ldots, x_k$라고 하고 x_i의 값을 가지는 자료의 개수를 n_i라고 하면

$$\overline{x} = \frac{x_1 n_1 + \cdots + x_k n_k}{n} = \sum_{i=1}^{k} x_i \frac{n_i}{n} = \sum_{i=1}^{k} x_i p_i$$

○ $p_i = n_i/n$는 x_i의 자료가 차지하는 비율

○ 통계적 확률의 관점에서 볼 때, n을 계속 크게 하면 표본들은 모집단으로, 표본비율 p_i는 확률질량함수 $f(x_i)$로, 표본평균은 **모평균(population mean)**으로

$$\overline{x} = \sum_{x_i} x_i p_i \quad \rightarrow \quad \sum_{x_i} x_i f(x_i) = \mu$$

- 표본평균이 자료들의 무게중심이듯이 평균은 확률분포(또는 모집단)의 무게중심
- 확률변수의 기댓값: 확률변수에 대하여 평균적으로 기대하는 값이라는 의미 = 모평균
  - 확률변수 X의 기대값

$$E(X) = \sum_x x\, f(x) = \mu$$

○ 연속확률변수의 기댓값

◦ 이산형의 기댓값에서

- $\sum$을 $\int$으로
- 확률질량함수 $f(x) = P(X = x)$를 확률밀도함수에 단위길이를 곱한 $f(x)\,dx$로 바꾸어 계산

$$\mu = E(X) = \int x f(x)\,dx$$

◉ 동전 세 번 던지기

◦ X : 앞면의 수

$$E(X) = 1.5$$

◉ 돌림판

◦ $f(x) = 1/12, \quad 0 < x \leq 12$

$$E(X) = 6$$

○ 변환된 변수의 기댓값

◉ 동전 세 번 던지기

◦ $W=(X-1.5)^2$의 기댓값은? $E(W)=\sum_w w f_W(w)$

◦ W의 확률질량함수 $f_W(w)$를 유도

$$f_W(0.25)=3/4, \quad f_W(2.25)=1/4$$

$$\Rightarrow \ E(W)=0.25\times\frac{3}{4}+2.25\times\frac{1}{4}=0.75$$

○ W의 확률질량함수

$$f_W(0.25) = \frac{3}{4} = f_X(1) + f_X(2) = \frac{3}{8} + \frac{3}{8}$$

$$f_W(2.25) = \frac{1}{4} = f_X(0) + f_X(3) = \frac{1}{8} + \frac{1}{8}$$

○ $E(W)$의 첫 번째 항과 두 번째 항

$$0.25 \times \frac{3}{4} = (1-1.5)^2 f_X(1) + (2-1.5)^2 f_X(2)$$

$$2.25 \times \frac{1}{4} = (0-1.5)^2 f_X(0) + (3-1.5)^2 f_X(3)$$

$$\Rightarrow E(W) = E((X-1.5)^2) = \sum_{x=0}^{3} (x-1.5)^2 f_X(x)$$

○ 확률변수 X의 함수인 $Y=g(X)$의 기댓값

$$E(Y)=E(g(X))=\begin{cases}\sum_x g(x)f_X(x), & \text{이산확률변수}\\ \int g(x)f_X(x)\,dx, & \text{연속확률변수}\end{cases}$$

- 이산확률변수: $E(X^2)=\sum_x x^2 f(x)$
- 연속확률변수: $E(X^2)=\int x^2 f(x)\,dx$

○ 기댓값의 성질

① 임의의 상수 a의 기댓값은 $E(a)=\sum_x af(x)=a\sum_x f(x)=a$

② $E(aX+b)=\sum_x (ax+b)f(x)=a\sum_x xf(x)+b=aE(X)+b$

③ $$\begin{aligned} E(g_1(X)+g_2(X)) &= \sum_x \{g_1(x)+g_2(x)\}f(x) \\ &= \sum_x g_1(x)f(x)+\sum_x g_2(x)f(x) \\ &= E(g_1(X))+E(g_2(X)) \end{aligned}$$

◉ W의 기댓값

$$\begin{aligned} E((X-1.5)^2) &= \sum_{x=0}^{3}(x-1.5)^2 f_X(x) \\ &= \sum_{x=0}^{3} x^2 f_X(x) - \sum_{x=0}^{3} 3x f_X(x) + \sum_{x=0}^{3} 1.5^2 f_X(x) \\ &= E(X^2) - 3E(X) + 1.5^2 \end{aligned}$$

○ $E(X) = 1.5$

○ $E(X^2) = \sum_{x=0}^{3} x^2 f_X(x) = 3$

○ $E((X-1.5)^2) = 0.75$

□ 모분산(population variance)

○ 표본분산

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i (x_i - \overline{x})^2 = \frac{n}{n-1}\sum_{i=1}^{k} (x_i - \overline{x})^2 p_i$$

○ n을 계속 크게 하면

- 표본분산은 모분산으로
- p_i는 $f(x_i)$로
- $\overline{x}$는 μ로
- $n/(n-1)$은 1로

○ 모분산을 σ^2로 표시

$$s^2 = \frac{n}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 p_i \quad \rightarrow \quad \sigma^2 = \sum_{i=1}^{k}(x_i - \mu)^2 f(x_i)$$

○ 확률변수 X의 분산을 $Var(X)$로 표시

$$Var(X) = \sum_{x}(x-\mu)^2 f(x) = E((X-\mu)^2)$$

- 분산은 $g(X) = (X-\mu)^2$의 기댓값

○ $Var(X) = \sum_{x}(x-\mu)^2 f(x) = E(X^2) - \mu^2 = E(X^2) - E(X)^2$

○ 연속확률변수

$$Var(X) = \int (x-\mu)^2 f(x)\,dx = \int x^2 f(x)\,dx - \left(\int x f(x)\,dx\right)^2$$

- 표준편차: $\sigma = \sqrt{\sigma^2} = SD(X)$

◉ 동전 세 개를 던지기: 앞면의 수 X의

- 평균: $\mu = 1.5$
- 분산: $Var(X) = E((X-\mu)^2) = E((X-1.5)^2) = 0.75$
- 표준편차: $\sigma = \sqrt{0.75} = 0.866$

◉ 이산균일분포

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

- 2.5를 중심으로 대칭이므로 $E(X) = 5/2 = 2.5$
- $E(X^2) = \frac{30}{4}$
- $\sigma^2 = E(X^2) - E(X)^2 = \frac{30}{4} - \left(\frac{5}{2}\right)^2 = \frac{5}{4} = 1.25$
- $\sigma = \sqrt{\sigma^2} = \sqrt{1.25} = 1.118$

◉ 돌림판

$$f(x) = 1/12, \quad 0 < x \leq 12$$

- $E(X) = 6$
- $E(X^2) = 48$
- $Var(X) = 12$
- $SD(X) = \sqrt{12} = 3.464$

○ 기대값의 성질

⑤ $Var(aX+b) = E((aX+b)^2) - E(aX+b)^2$

$= a^2E(X^2) + 2abE(X) + b^2 - (a^2E(X)^2 + 2abE(X) + b^2)$

$= a^2(E(X^2) - E(X)^2) = a^2\,Var(X)$

- 위치의 변화를 주는 상수 b는 분산에 영향을 주지 않음
- 분산은 측정단위 척도의 제곱으로 표시되기 때문에 a의 제곱을 곱함

⑥ $SD(aX+b) = \sqrt{Var(aX+b)} = \sqrt{a^2\,Var(X)} = |a|SD(X)$

■ 결합분포와 주변분포

◉ 동전 세 번 던지기

○ X: 앞면의 수, Y: 앞면과 뒷면의 수의 차이

$$\Omega = \{\ HHH,\ HHT,\ HTH,\ THH,\ HTT,\ THT,\ TTH\ \ TTT\ \}$$

	↓	↓	↓	↓	↓	↓	↓	↓
$X=$	3	2	2	2	1	1	1	0
$Y=$	3	1	1	1	1	1	1	3

○ 두 변수를 동시에 고려한 확률분포?

Y \ X	0	1	2	3
1				
3				

□ 결합확률질량함수(joint p.m.f.)

○ 두 개 이상의 확률변수들을 동시에 고려한 확률분포

○ 두 이산확률변수 X와 Y에 대해

$$f(x, y) = P(X = x,\ Y = y)$$

- 수식에서 ,는 $\cap$를 의미
- 예) 동전 세 번 던지기

$$f(0,3) = \frac{1}{8},\ f(1,1) = \frac{3}{8},\ f(2,1) = \frac{3}{8},\ f(3,3) = \frac{1}{8}$$

○ n개의 이산확률변수 $X_1, ..., X_n$에 대해

$$f(x_1,...,x_n) = P(X_1 = x_1,...,X_n = x_n)$$

□ 주변확률질량함수(marginal p.m.f.)

○ 표본공간이 사건 $B_1, \ldots, B_n$로 분할될 때 사건 A의 확률은

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$$

○ 사건 A가 $X = x$, B_i가 $Y = y_i$라고 하면

$$P(A \cap B_i) = P(X = x,\ Y = y_i)$$

$$P(X = x) = P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(X = x, Y = y_i)$$

$$\Rightarrow\ f_X(x) = \sum_{y} f(x,y), \quad f_Y(y) = \sum_{x} f(x,y)$$

○ $f_X(x)$를 X의 주변확률질량함수, $f_Y(y)$를 Y의 주변확률질량함수라고 함

◉ 동전 세 번 던지기

Y \ X		x				합
		0	1	2	3	
y	1	0	3/8	3/8	0	3/4
	3	1/8	0	0	1/8	1/4
합		1/8	3/8	3/8	1/8	1

○ 독립 확률변수

◦ 사건 A와 B는 독립 $\leftrightarrows$ $P(A\cap B)=P(A)P(B)$

◦ 두 확률변수 X와 Y는 독립 $\leftrightarrows$ 모든 x, y에 대하여

$$f(x,\, y)=f_X(x)\,f_Y(y)$$

◦ n개의 이산확률변수 $X_1,\, ...,\, X_n$이 서로독립(상호독립) $\leftrightarrows$ 모든 $x_1,...,x_n$에 대해

$$f(x_1,\, ...,\, x_n)\;=\;f_{X_1}(x_1)\cdots f_{X_n}(x_n)=\prod_{i=1}^{n}f_{X_i}(x_i)$$

◉ 동전 세 번 던지기

$$f(1,1) = 3/8 \neq f_X(1) f_Y(1) = (3/8)(3/4) = 9/32$$

⇨ X와 Y는 독립이 아님

◉ $f(x,y) = \dfrac{xy}{36}, \quad x = 1,2,3, \ y = 1,2,3$

Y \ X	1	2	3	f_Y
1	1/36	2/36	3/36	1/6
2	2/36	4/36	6/36	2/6
3	3/36	6/36	9/36	3/6
f_X	1/6	2/6	3/6	1

○ 모든 x, y에 대해 $f(x,y) = f_X(x) f_Y(y)$ 성립

□ 기댓값

○ $E(X)=\sum_x xf_X(x),\ E(Y)=\sum_y yf_Y(y)$

○ 확률변수 X와 Y에 대해, $X+Y$의 기댓값? XY의 기댓값?

○ 결합확률질량함수나 결합확률밀도함수를 이용

$E(X+Y)=$

$E(XY)=$

○ 기댓값 정리

◦ $E(X+Y)=E(X)+E(Y)$

◦ X와 Y가 독립이면 $E(XY)=E(X)E(Y)$

○ 공분산(Covariance)

◦ 표본공분산 $c = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$

◦ 두 확률변수 X와 Y의 공분산

$$Cov(X, Y) = \sum_x \sum_y (x-\mu_X)(y-\mu_Y) f(x,y) = E((X-\mu_X)(Y-\mu_Y))$$
$$= \sum_x \sum_y xy f(x,y) - \mu_X \mu_Y = E(XY) - E(X)E(Y)$$

◦ 두 확률변수의 직선관계의 정도를 나타내는 측도

◦ X와 Y가 독립이면 $E(XY) = E(X)E(Y)$ 이므로
$Cov(X, Y) = 0$ 하지만 그 역은 일반적으로 성립하지 않음

◉ 결합확률분포표

x \ y	-1	0	1	$f_X(x)$
0	$\frac{1}{3}$	0	$\frac{1}{3}$	$\frac{2}{3}$
1	0	$\frac{1}{3}$	0	$\frac{1}{3}$
$f_Y(y)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1

- $E(X) = 1/3,\ E(Y) = 0$
- $E(XY) =$
- $f(0,-1) = \frac{1}{3} \neq \frac{2}{3} \times \frac{1}{3} = f_X(0) f_Y(-1)$ ⇨ 독립 아님

○ $Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(x, Y)$

○ X와 Y가 독립이면 $Var(X \pm Y) = Var(X) + Var(Y)$

○ 상관계수(coefficient of correlation)

◦ 공분산은 척도에 영향을 받음 ⇨ 표준화 필요

◦ 두 확률변수 X와 Y의 상관계수

$$\rho = Cor(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

◦ $\sigma_{XY} = Cov(X, Y)$라고 표시하면

$$\rho_{XY} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]$$

◦ $-1 \leq \rho \leq 1$

■ 이항분포와 그에 관련된 분포들

○ 베르누이(Bernoulli) 시행

① 각 실험에서 발생 가능한 결과는 단 2가지
- 예: (성공,실패), (앞면,뒷면)

② 각 실험이 독립적으로 수행

③ 모든 실험에서 결과의 확률은 항상 동일
- $P(S)=p,\ P(F)=1-p=q$

Jacob Bernoulli

Johann Bernoulli

◉ 불량품검사

◦ 10개의 제품중 3개가 불량품

- 2개를 복원추출하는 경우 ⇨ 베르누이 시행

$$P(S_{1,}S_2) = P(S_1)P(S_2|S_1) = \frac{3}{10} \times \frac{3}{10}$$

- 2개를 비복원추출하는 경우 ⇨ 독립? 확률고정?

$$P(S_1, S_2) = P(S_1)P(S_2|S_1) = \frac{3}{10} \times \frac{2}{9}$$

○ 10000개의 제품중 3000개가 불량품

- 2개를 복원추출하는 경우 ⇨ 베르누이 시행

$$P(S_1, S_2) = P(S_1)P(S_2|S_1) = \frac{3000}{10000} \times \frac{3000}{10000}$$

- 2개를 비복원추출하는 경우

$$P(S_1, S_2) = P(S_1)P(S_2|S_1) = \frac{3000}{10000} \times \frac{2999}{9999}$$

$$P(S_2) = 0.3 \neq 0.29993 = P(S_2|S_1) \simeq P(S_2)$$

※ 모집단의 크기가 크고 표본의 크기가 상대적으로 크지 않는 경우, 독립적으로 반복되는 베르누이 실험으로 간주해도 큰 차이가 없음 ⇨ 근사모형으로 사용가능

◯ 베르누이 확률변수

◦ 성공할 확률 = p인 경우 $X \sim B(p)$로 표시하며

- $X = \begin{cases} 1, & \text{성공} \\ 0, & \text{실패} \end{cases}$
- $P(X=1) = P(\text{성공}) = p,\ P(X=0) = P(\text{실패}) = 1-p$

$$\Rightarrow\ f(x) = P(X=x) = p^x(1-p)^{1-x}, \quad x = 0,1$$

◦ 기댓값

- $E(X) = p = P(\text{성공})$
- $E(X^2) = p = P(\text{성공})$
- $Var(X) = p(1-p) = P(\text{성공})P(\text{실패})$

□ 이항분포 (Binomial distribution)

○ 성공할 확률이 p인 베르누이 실험을 n번 반복했을 때, 성공횟수(X)의 분포

○ 성공횟수 X는 n**개의 베르누이 확률변수를 합**한 것

$$\begin{array}{lccccc} & X_1 + & X_2 + & \cdots & + X_n = & X \\ S & 1 & 1 & \cdots & 1 & \downarrow \\ F & 0 & 0 & \cdots & 0 & \text{성공횟수} \end{array}$$

- $X_i \sim B(p)$

◦ 베르누이 시행은 독립을 의미

- $E(X) = np$
- $Var(X) = np(1-p)$

◉ 주사위 세 번 던지기

◦ X: 1이 나온 횟수 (1이면 S, 아니면 F)

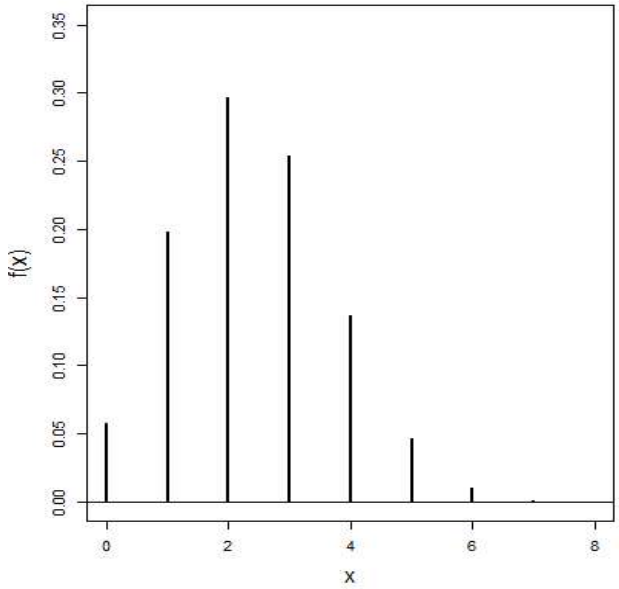
0	1	2	3
FFF	SFF FSF FFS	SSF SFS FSS	SSS
$\binom{3}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^3$	$\binom{3}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^2$	$\binom{3}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^1$	$\binom{3}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^0$

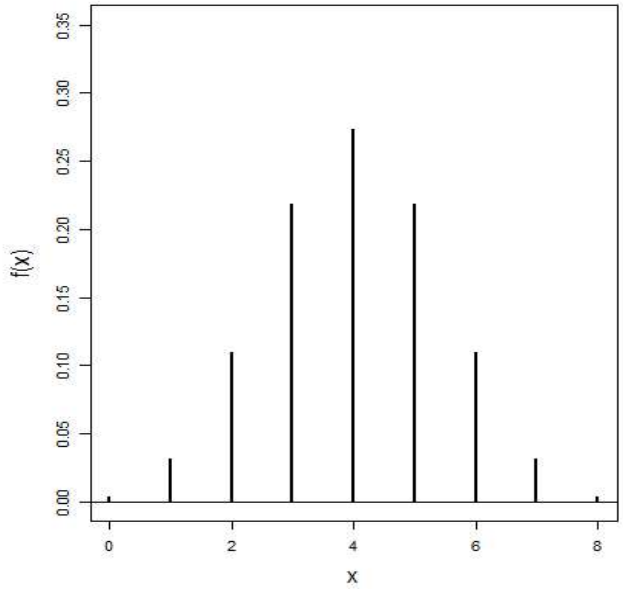
○ 일반식: $f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ ..., n$

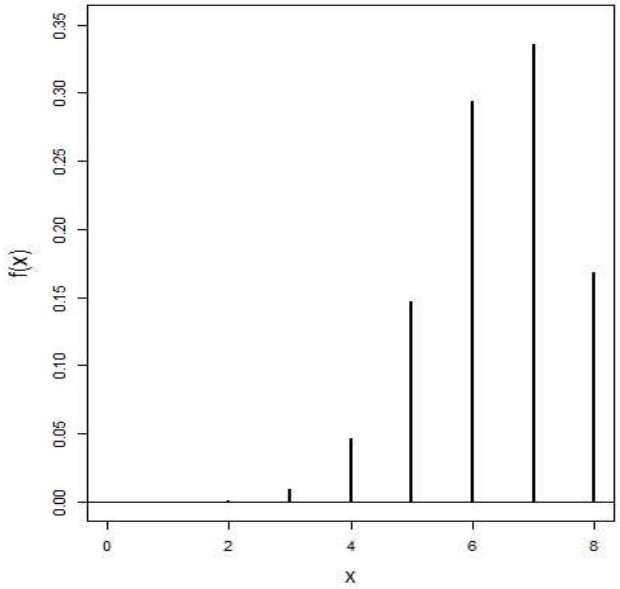
○ 표시 $X \sim B(n,p)$

- n은 시행횟수이고 p는 성공할 확률

$p = 0.3$ $p = 0.5$ $p = 0.8$







○ n과 p 값은 이항분포의 모양을 결정

⇨ 분포의 특성(모양, 기댓값 등)을 완전히 결정하는 값을 **모수(parameter)**라고 함

○ 분포의 모수를 알면 해당 분포의 모든 것을 알 수 있음

⇨ **통계학 문제: 모수는?**

◉ 항암제 완치율

- 어떤 암에 대한 기존 항암제의 완치율은 50%
- 어느 제약회사에서 새로운 항암제를 개발하여 항암제의 효과를 확인하기 위해 15명의 환자를 대상으로 실험
- 만약 새로운 항암제의 완치율이 기존과 같다면
  ① 8명이 완치될 확률은?
  ② 적어도 10명까지 치유될 확률은? 0.941
- **통계문제:** 환자 중 12명의 환자가 치유되었다면, 새로운 항암제의 효과가 기존의 것보다 있다고 할 수 있는가? **0.018**

□ 초기하분포 (Hypergeometric Dist.)

- ○ 크기가 N인 모집단이 크기가 M과 $N-M$인 두 개의 부모집단으로 나누어진 경우 ⇨ **유한모집단**
- ○ n개의 표본을 **비복원으로 추출**할 때, 크기가 M인 부모집단(A)에서 추출될 표본 수의 분포
- ○ 확률질량함수 :

$$f(x)=\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

- $x=\max(0,n-N+M),\ \ldots,\min(n,M)$

○ 초기하분포도 각 시행에서 A 집단에서 추출되면 1 다른 집단에서 추출되면 0으로 표시한 확률변수의 합

$$\begin{array}{lccccc} & X_1 + & X_2 + & \cdots & + X_n = & X \\ A & 1 & 1 & \cdots & 1 & \downarrow \\ B & 0 & 0 & \cdots & 0 & A\text{에서 추출된 표본의 수} \end{array}$$

- $P(X_i = 1) = P(A) = M/N,\ P(X_i = 0) = 1 - M/N$
- $E(X_i) = M/N \Rightarrow E(X) = nM/N$

○ 다른 점은 추출이 비복원으로 각각의 시행이 독립이 아님

⇨ $Var(X) = ?$

○ n에 비해 N이 상대적으로 큰 경우

- 비복원의 효과가 적기 때문에 베르누이 실험으로 근사
- 초기하 분포은 $p = M/N$인 이항분포로 근사

◉ 품질관리 – Operating Characteristic(OC) curve

- 50개의 전구들이 들어 있는 상자에서 10개의 전구를 무작위로 선택하여 검사
- 불량전구의 개수가 1개 이하이면 이 회사의 전구를 구매
- 만약 이 상자에 10개의 불량품이 있을 때, 구매할 확률은?

- 만약 k개 불량품이 있을 때, 구매할 확률은?

◉ 본관 앞 연못에 사는 물고기는 몇 마리?

○ 꼬리표를 붙인 20마리의 물고기를 연못에 넣고 어느 정도 지난 후 물고기 15마리를 잡았을 때 꼬리표가 있는 물고기의 분포는?

○ 15마리 중 4마리가 꼬리표가 있는 물고기라면?

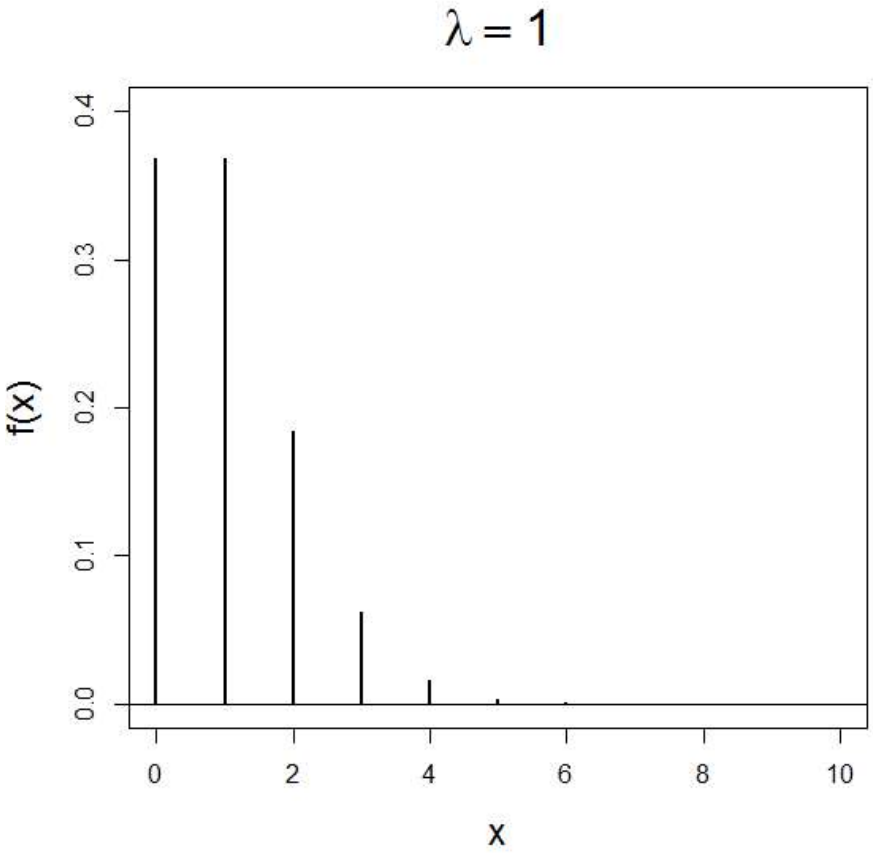
□ 포아송분포 (Poisson distribution)

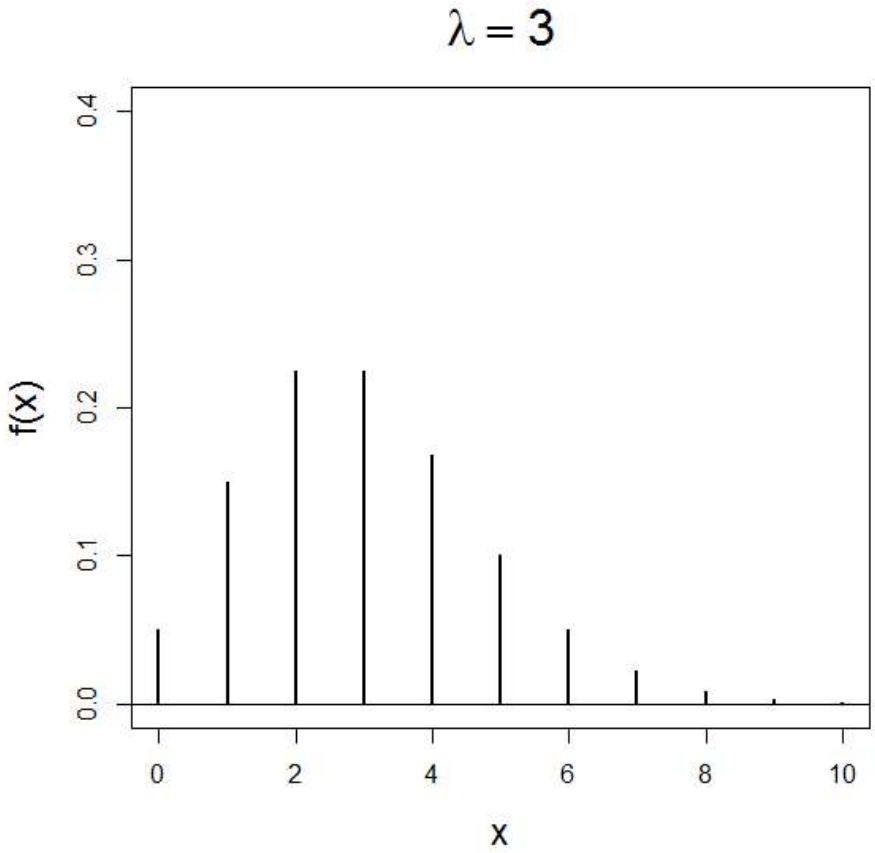
○ 발생 가능성이 희박한 사건이 임의의 구간 안에서
평균적으로 λ번 발생할 때, 이 사건이 일어날 횟수의 분포

○ 확률질량함수

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, \ldots .$$

- 표시: $X \sim P(\lambda)$


λ = 1
f(x)
0.0
0.1
0.2
0.3
0.4
0
2
4
6
8
10
x


λ = 3
f(x)
0.0
0.1
0.2
0.3
0.4
0
2
4
6
8
10
x

○ 이항분포의 포아송 근사

◦ p가 작고 n이 큰 경우, 이항분포의 확률을 포아송분포를 사용해서 근사할 수 있음

- $\lambda = np$라고 하면, $p = \lambda/n$

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \simeq \frac{e^{-\lambda}\lambda^x}{x!}$$

◉ 컴퓨터 프로그램 버그

◦ 500개 모듈 당 평균 한 개의 버그 발생

◦ 독립적으로 제작된 1500개 다른 모듈로 이루어진 프로그램 패키지에서 버그가 2개 이하일 확률은?

- 모듈 당 버그가 발생할 확률 $p = 1/500$
- X: 패키지에서의 버그 수, $X \sim B(1500, 1/500)$

  ⇨ $P(X \le 2) = 0.4230$

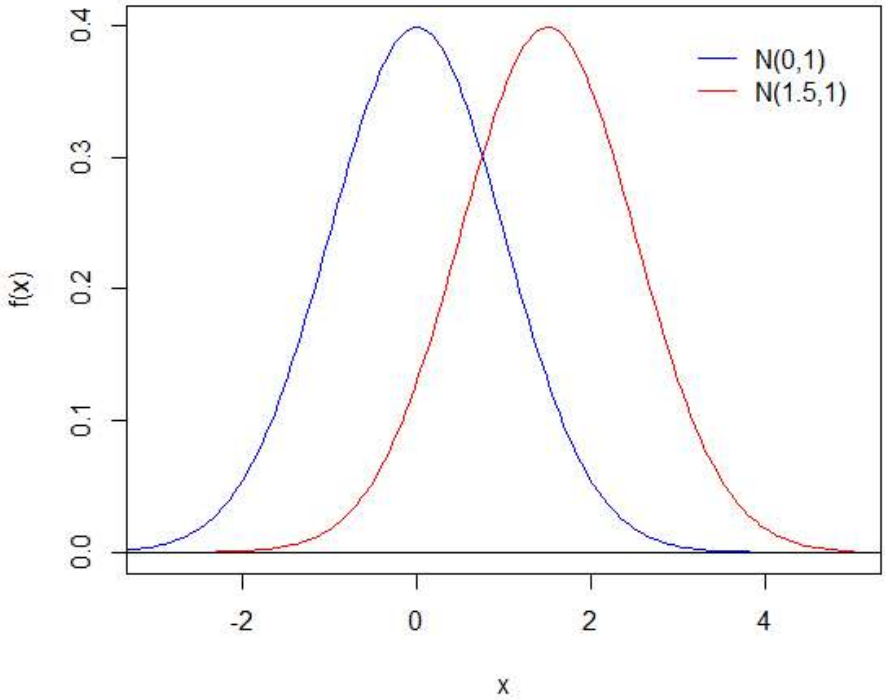
- $\lambda = 1500/500 = 3$ ⇨ $P(X \le 2) \simeq 0.4232$

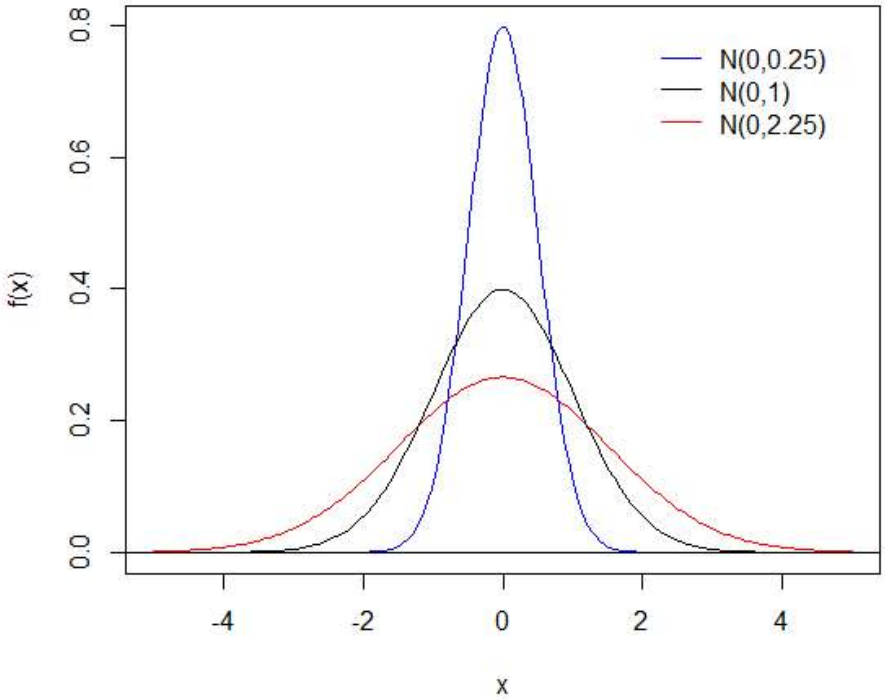
■ 정규분포(Normal Distribution)

- Gauss가 각종 물리실험을 수행할 때 발생하는 측정오차를 설명하기 위해 적용한 분포
- 모든 학문 분야에서 확률모형 또는 근사모형으로 사용
- 평균은 중심위치를 종모양(bell-shaped)의 **대칭형태**를 가짐
- 평균이 μ이고 분산이 σ^2인 정규분포의 확률밀도함수

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

  - μ: 분포의 중심
  - σ^2: 퍼져있는 정도
  - 표시: $X\sim N(\mu,\sigma^2)$


N(0,1)
N(1.5,1)
f(x)
x
0.0
0.1
0.2
0.3
0.4
-2
0
2
4


N(0,0.25)
N(0,1)
N(0,2.25)
f(x)
x
0.0
0.2
0.4
0.6
0.8
-4
-2
0
2
4

○ **확률계산:** $P(a < X < b) = ?$

$$P(a < X < b) = \int_a^b f(x)\,dx = \int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx \quad = ?$$

○ 표준정규분포(standard normal distribution)

- $\mu = 0$ 이고 $\sigma^2 = 1$ 인 경우

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

  - 0을 중심으로 대칭
- 일반적으로 Z로 표시: $Z \sim N(0,1)$
- 확률계산:

$$P(a < X < b) = \int_a^b f(x)\,dx = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \;\; =?$$

  - 수치해석학적으로 계산
  - 표로 제시

◉ 표준정규분포의 확률계산

- ○ 표의 종류 : $P(Z \le z)$, $P(Z > z)$, $P(0 < Z < z)$
- ○ 그림과 0을 중심으로 대칭이라는 사실을 이용
- ○ $Z \sim N(0,1)$ 이면
  - $P(Z \le 1.37) = 0.9147$
  - $P(0.5 < Z \le 1.2) = 0.1934$
  - $P(Z \le -1.96) = 0.0250$
  - $P(-0.15 < Z < 1.60) = 0.5048$
- ○ α가 주어지고 $P(Z < z) = \alpha$를 만족하는 z를 계산
  - $P(Z < z) = 0.975$를 만족시키는 z는?
  - $P(-z < Z < z) = 0.90$를 만족시키는 z는?

○ 정규분포의 표준화

- 확률변수 X의 평균이 μ이고 표준편차가 $\sigma(\sigma>0)$인 경우

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

- Z : 표준화된 확률변수
- $E(Z)=0,\ \ Var(Z)=1 \Rightarrow\ SD(Z)=1$
- **선형변환된 정규확률변수도 정규분포를 따름**

$$X\sim N(\mu,\sigma^2)\ \Rightarrow\ Z=\frac{X-\mu}{\sigma}\sim N(0,1)$$

$$Z\sim N(0,1)\ \Rightarrow\ X=\sigma Z+\mu\sim N(\mu,\sigma)$$

◉ $X \sim N(60, 16)$ 일 때

- $P(55 \leq X < 63) = 0.6678$

- $P(X \leq x) = 0.025$ 를 만족하는 x 는?

◉ 시험 점수의 분포

○ 평균이 490이고 표준편차가 50인 정규분포를 따른다면

- 600점 이상 받을 확률은? 0.0139

- 상위 5%인 사람의 점수는? 572.25

○ 정규분포의 정리

- ◦ $X \sim N(\mu, \sigma^2)$ 이고 $a \neq 0$ 이면, $aX+b \sim N(a\mu+b, a^2\sigma^2)$
- ◦ **두 정규확률변수의 선형결합도 정규분포를 따름**

  ⇨ 모수인 **평균과 분산은**?

  - $X_1 \sim N(\mu_1, \sigma_1^2)$ 이고 $X_2 \sim N(\mu_2, \sigma_2^2)$ 이면,

$$X_1 \pm X_2 \sim N(\mu_1 \pm \mu_2, \sigma_1^2 + \sigma_2^2 \pm 2\sigma_{12})$$

  - 추가 가정: X_1과 X_2가 독립이면, $\sigma_{12} = 0$

$$X_1 \pm X_2 \sim N(\mu_1 \pm \mu_2, \sigma_1^2 + \sigma_2^2)$$

- ◦ $\sigma_{12} = 0$ 이면, X_1과 X_2는 독립

◉ 모의실험

◉ 아침식사: 빵과 우유를 먹는다고 가정

- 빵의 열량은 평균 200kcal, 표준편차 15kcal인 정규분포
- 우유의 열량은 평균 80kcal, 표준편차 5kcal인 정규분포
- 아침식사에서 300 칼로리이상 섭취할 확률은? 0.1030

- 동일한 식사를 일주일 했을 때, 300kcal 이상 섭취할 날이 하루일 확률은? 0.376

■ 확률표본과 표집분포

○ 확률표본(random sample)

- ◦ 모집단에서 랜덤하게 선택되어진 관측값
- ◦ 서로 독립이고 동일한 분포를 따른다고 가정

  ⇨ independent and identically distributed (iid)

  - $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$

  - $X_1, \ldots, X_n \overset{\text{iid}}{\sim} B(p)$

◯ 표집분포(sampling distribution)

◉ 확률분포가 다음과 같을 때

x	0	1	2
$P(X=x)$	2/5	2/5	1/5

◦ $\mu = E(X) = \frac{4}{5}$

◦ $\sigma^2 = Var(X) = \frac{14}{25}$

○ 위의 분포에서 두 개의 확률표본을 추출한 경우

- 두 표본의 평균 $\overline{X}$의 분포는?
- 두 표본의 최대값 Y의 분포는?

		X_2		
		0	1	2
X_1	0	4/25	4/25	2/25
	1	4/25	4/25	2/25
	2	2/25	2/25	1/25

- 표본평균 $\overline{X}$의 분포는?

$\overline{x}$	0	1/2	1	3/2	2
$P(\overline{X}=\overline{x})$	4/25	8/25	8/25	4/25	1/25

- $E(\overline{X})=\frac{4}{5}$, $Var(\overline{X})=\frac{14}{50}=\frac{1}{2}\times\frac{14}{25}$

- 표본 최대값 Y의 분포는?

y	0	1	2
$P(Y=y)$			

○ 평균이 μ이고 분산이 σ^2인 분포에서 n**개의 확률표본**을 추출했을 경우, **표본평균** $\overline{X}$**의 분포**는?

- $E(\overline{X}) = ?$

- $Var(\overline{X}) = ?$

- $SD(\overline{X}) = ?$ ⇨ 표준오차(standard error, SE)
- 분포의 형태는?

○ 정규분포

- $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$ 이고 X_1과 X_2가 독립

$$\Rightarrow X_1 \pm X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

- $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ 이면

  - $X_1 + \cdots + X_n \sim N(n\mu, n\sigma^2)$
  - $\overline{X} = \dfrac{X_1 + \cdots + X_n}{n} \sim N(\mu, \sigma^2/n)$
  - $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

○ 중심극한정리 (Central limit theorem, CLT)

- 평균이 μ이고 분산이 σ^2인 모집단에서
- n개의 확률표본을 추출했을 경우
- n을 크게 할수록 모집단의 형태와 관계없이
- **표본평균의 분포는** 평균이 μ이고 분산이 $\frac{\sigma^2}{n}$인 **정규분포에 근사**

$$\overline{X} \simeq N(\mu, \frac{\sigma^2}{n}) \Rightarrow Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \simeq N(0,1)$$

◉ R을 이용한 모의실험

- ○ 균일분포: $f(x) = 1, \quad 0 < x < 1$
  - $E(X) = 0.5, \quad Var(X) = 1/12$
- ○ 감마분포(2,1): $f(x) = xe^{-x}, \quad 0 < x < \infty$
  - $E(X) = 2, \quad Var(X) = 2$
- ○ 이산분포

x	0	1	2	3
$P(X = x)$	0.4	0.1	0.1	0.4

  - $E(X) = 1.5, \quad Var(X) = 1.85$

◉ 어느 모집단의 평균은 82이고 표준편차가 12

① 이 모집단에서 64개의 확률표본을 추출하였을 때,

$P(80.8 \leq \overline{X} \leq 83.2)$의 근사확률은?

② 100개의 확률표본을 추출하였다면,

◯ 이항분포의 정규근사

◦ $X \sim B(n,p)$

- X_i : i번째 베르누이 확률변수

$$\Rightarrow \; E(X_i)=p, \;\; Var(X_i)=p(1-p)$$

- $X = X_1 + X_2 + \cdots + X_n$
- 표본비율 : $\hat{p} = X/n = \overline{X}$

◦ n이 큰 경우, 중심극한정리에 의해

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

$$\Rightarrow \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \simeq N(0,1) \Rightarrow \frac{X-np}{\sqrt{np(1-p)}} \simeq N(0,1)$$

○ 연속성 수정

- 이항분포는 이산형이고 정규분포는 연속형
- 연속확률변수 X는 $P(X \le x) = P(X < x)$이지만 이산확률변수는 아님

$$\begin{array}{ccccc} P(X \le x-1) & = & P(X < x) & \neq & P(X \leqq x) \\ \updownarrow & & \updownarrow & & \updownarrow \\ P\left(Z \le \dfrac{x-1-np}{\sqrt{np(1-p)}}\right) & \neq & P\left(Z < \dfrac{x-np}{\sqrt{np(1-p)}}\right) & = & P\left(Z \le \dfrac{x-np}{\sqrt{np(1-p)}}\right) \end{array}$$

- 0 또는 자연수 x에 대해 이항확률변수 X는

$$P(X<x)=P(X\le x-1),\ P(X\le x)=P(X<x+1)$$

$$P(X>x)=P(X\ge x+1),\ P(X\ge x)=P(X>x-1)$$

$$\Rightarrow\ P(X<x)\ \simeq\ P\left(Z<\frac{x-1/2-np}{\sqrt{np(1-p)}}\right)\ \simeq\ P(X\le x-1)$$

$$\Rightarrow\ P(X>x)\ \simeq\ P\left(Z<\frac{x+1/2-np}{\sqrt{np(1-p)}}\right)\ \simeq\ P(X\ge x+1)$$

◉ 여론조사

◦ 전체 국민 중 60%가 A 정책에 대해 적극 찬성하다고 가정

◦ 150명을 무작위로 뽑아 찬성하는 사람의 비율을 알아보려고 할 때 적극 찬성하는 사람이 78명 이하일 확률은?

- $X \sim B(150, 0.6)$ 일 때 $P(X \leq 78)$?
- 정확한 확률 = 0.0284
- $X \simeq N(90, 36)$

$$P(X \leq 78) \simeq P(Z \leq -1.917) = 0.0276$$

◯ 표본평균 차의 분포

◦ 두 독립적인 정규분포에서 각각 m, n개의 확률표본 추출

- $X_1, \ldots, X_m \overset{\text{iid}}{\sim} N(\mu_1, \sigma^2)$, $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu_2, \sigma^2)$

- $\overline{X} - \overline{Y}$의 표집분포는?

○ 평균이 μ_1, μ_2이고 분산이 σ_1^2, σ_2^2인 두 독립적인 모집단에서 각각 m, n개의 확률표본 추출했을 때, $\overline{X}-\overline{Y}$의 표집분포는?

○ 성공할 확률이 p_1과 p_2인 베르누이 시행을 독립적으로 각각 m, n씩 했을 때 $\hat{p}_1 - \hat{p}_2$의 표집분포는?