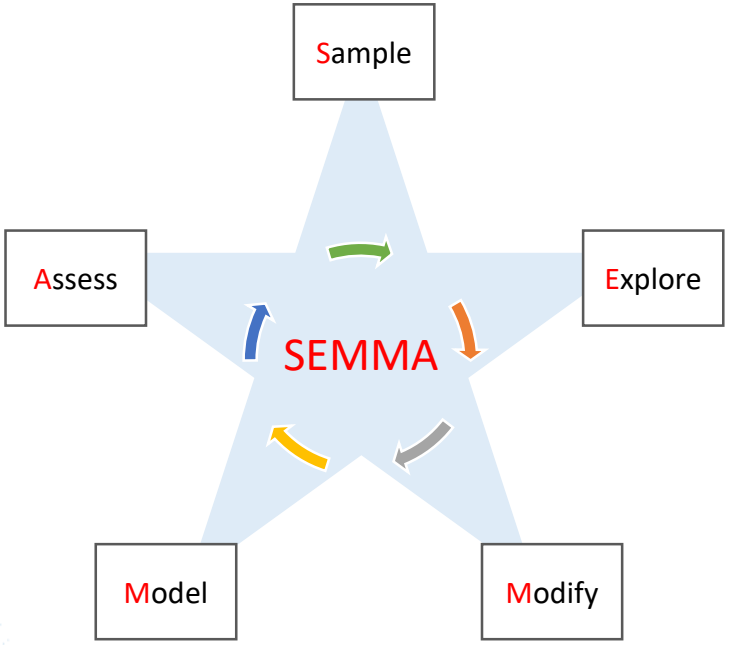
For Big Data

- 기초 통계학 & 데이터 분석 방법론 -

Daniel Rex Park
(박영식)

Intro.

:: Data Analytics - Methodology



KDD 방법론	CRISP-DM 방법론
분석대상 비즈니스 이해	Business Understanding
Select	Data Understanding
Preprocess	
Transform	Data Preparation
Data Mining	Modeling
Interpret / Evaluate	Evaluation
데이터 마이닝 활용	Deployment

If you want to be a Data Scientist::

Step 1	Step 2	Step 3	Step 4	Step 5
Basic Statistic (Hypothesis Test/ Correlation/ Regression etc)	**R-Basic and Intermediate** (Structure of R Data :: Scalar, Vector, Factor/ Matrix, Array/ Data.frame, List)	**Basic Python, Code Understanding** (int, float, str / List, Tuple, Dic, Set, Bool / If, for while/ def, class)	**Concept Of Machine Learning** (Classification, Regressor/ Dim Reduction, Association Rule Cluster)	**Concept Of Deep Learning** (Feed Forward Neural Network, Recurrent Neural Network, Convolutional Neural Network etc)

Basic Concept of Statistics

Daniel Rex Park
(박영식)

Chapter 1. Definition of Statistics

Daniel Rex Park
(박영식)

Chapter 1 통계학의 정의

1)통계학이란?

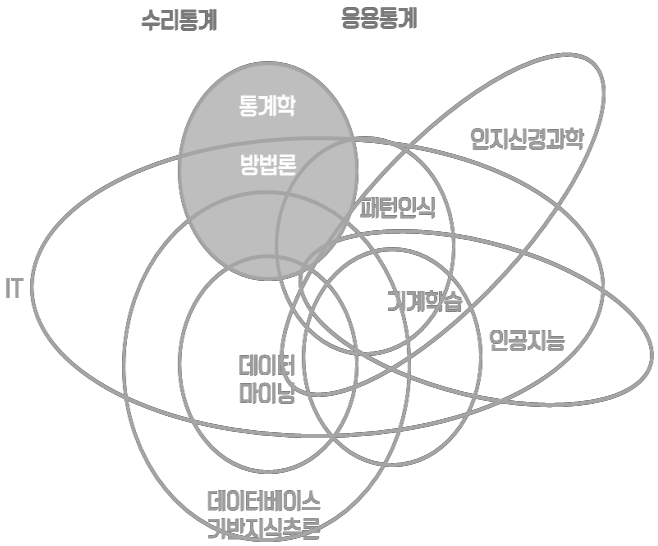
1)-1 **불확실한 상황**에서 **현명한 의사결정**을 하기 위한 **이론**과 **방법**의 체계

1)-2 통계학은 **자료의 수집, 분류, 분석**과 **해석**의 체계를 갖음

2) 통계학 활용 사례

- 야구시합의 승부를 예측하기 위한 승률 조사
- 대학 진학시 수능성적의 분포와 특정학과의 합격선 예측
- 복권의 당첨확률 계산
- 사회현상을 과학적으로 분석 및 예측
  :: 심리통계, 교육통계, 경영통계, 경제통계 등

※ 통계학의 적용사례가 달라지는 것이지
통계학 내용이 달라지는 것은 아님.



Chapter 1 통계학의 정의

3)모집단과 표본

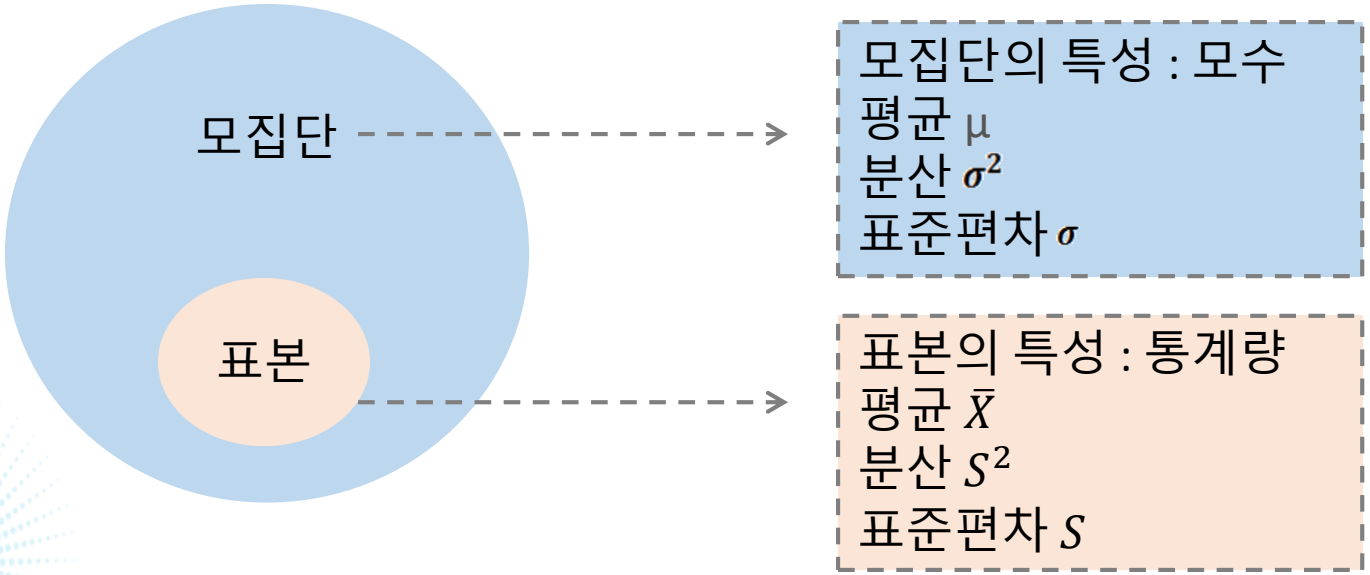
3)-1 모집단(population) : **연구자의 관심의 대상**이 되는 **모든 개체**의 집합

3)-2 표본집단(sample) : **모집단**에서 **조사 대상**으로 **채택된 일부**

4)모수와 통계량

4)-1 모수(parameter) : **모집단**의 **특성**을 **수치**로 나타낸 것

4)-2 통계량(statistic) : **표본**의 **특성**을 **수치**로 나타낸 것

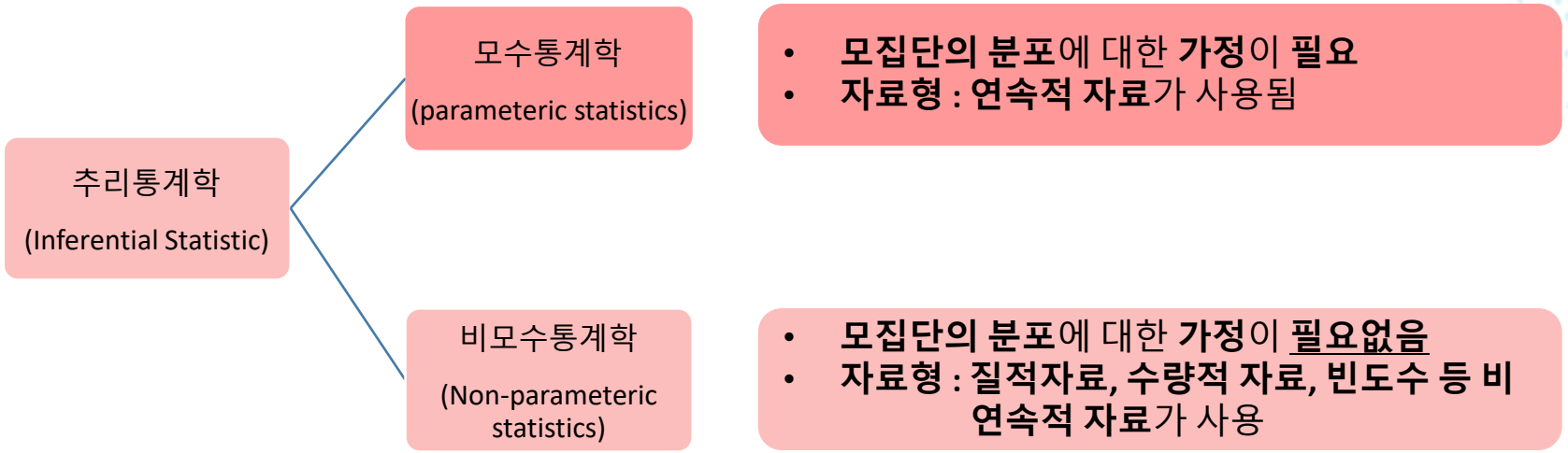


Chapter 1 통계학의 정의

Ex 1) 우리 학급의 평균 시험 점수를 알기 위해 자료를 구하여 평균을 구하는 것은?
A. 기술 통계학

Ex 2) 몇 백명의 표본을 통해 그들의 점수를 바탕으로 전국 고교생들의 모의고사 성적을 추론하는 방법에 대한 통계학은?
A. 추리 통계학 or 추측 통계학

Chapter 1 통계학의 정의



- 모수통계학에서는 **모집단의 분포모양이 정규분포**라는 가정과 활용될 자료가 **연속형**이어야 한다는 가정을 지닌다
- 비모수통계학에서는 **모집단의 정규성 가정이 필요가 없으며** 심지어 표본의 크기가 충분하지 않아도 활용 가능하며 **질적 자료 및 비연속적 자료(빈도수 등)**을 활용하여 사용한다.

편리성 : 모수통계학 < **비모수통계학**
신뢰성: **모수통계학** > 비모수통계학
활용성: **모수통계학** > 비모수통계학

※ 향후 학습 방향 :: 자료의 종류 → 모수통계학(추리통계학) → 비모수통계학(추리통계학)

Exercise Chapter 1

(1) 통계학이란 무엇인가요?

(2) 모집단과 표본은 무엇인가요?

(3) 모수와 통계량이 무엇인가요?

(4) 기술통계학과 추리통계학의 차이는 무엇인가요?

(5) 모수통계학과 비모수통계학은 무엇인가요?

Chapter 2. 통계학의 자료

Daniel Rex Park
(박영식)

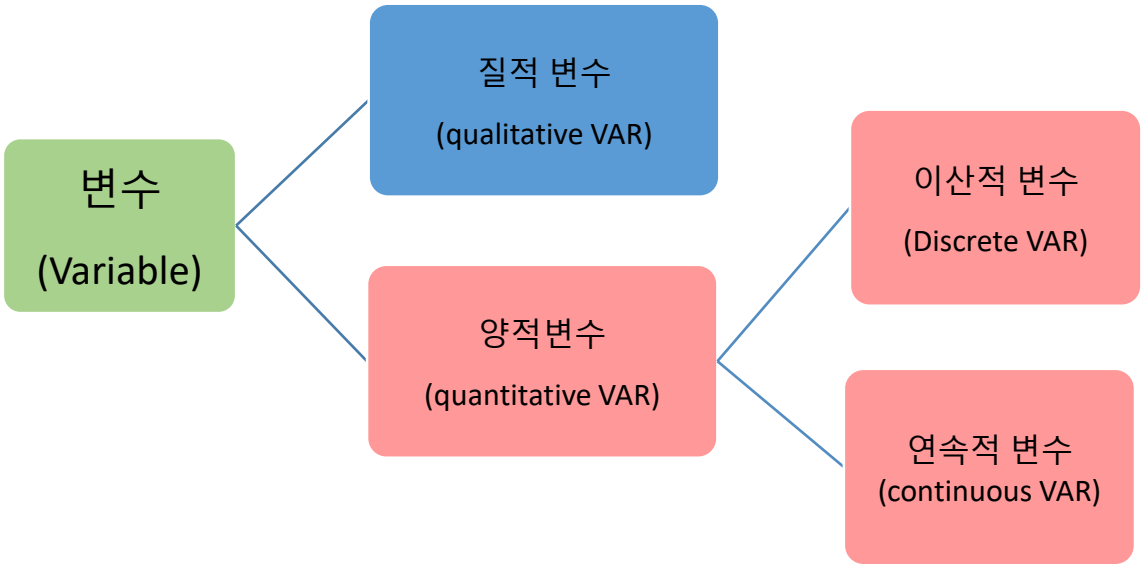
Chapter 2 Intro

- 변수란 무엇인가?
- 자료의 종류
- 변수의 수준
- 자료의 정리 및 해석
- 도수분포표 작성법

Daniel Rex Park
(박영식)

Chapter 2 통계학의 자료

- 변수 및 자료의 종류



- 질적변수(Qualitative Variable) : 종교/성별/직업 등 속성을 수치로 나타내기 어려운 변수
- 양적변수(Qualitative Variable) : 점수/통조림 용량/기업 매출 수치로 나타낼 수 있는 변수
- 이산적 변수(DiscreteVariable) : 세대 수, 학생 수, 핸드폰 판매 수와 같이 정수값인 변수
- 연속적 변수(Continuous Variable) : 길이, 키, 몸무게, 온도 등 연속적인 실수값인 변수

Chapter 2 통계학의 자료

- 변수의 수준

명목척도 (nominal)

- 가장 낮은 수준의 변수, 단순한 구분 기호
- Ex) 성별, 종교, 출생지, 자녀유무, 운동선수들 번호 등 …

서열척도 (ordinal)

- 측정대상 간 순서를 매기기 위한 변수
- Ex) 석차, 선호도 등 …

등간척도 (interval)

- 측정대상의 순서, 순서 사이의 간격을 알 수 있는 변수
- Ex) 온도, 지능지수, 대학학년 등 …

비율척도 (ratio)

- 등간변수의 특성 + 측정자료 간 비율 계산
- Ex) 연봉, 월급, 거리 등 …

Exercise Chapter 2

(1) 다음 자료를 이산적 자료와 연속적 자료로 구별하세요.
①증권시장에서의 하루 동안 거래되는 주식수
②기상청에서 한 시간 간격으로 보도되는 각 지역별 기온
③어떤 공장에서 생산되는 TV의 수명
④한 대학교 앞 모 커피전문점의 하루 평균 방문자 수
⑤0000년 00월 국내흥행 상위 5개 영화에 대한 평균 주말 총수입

A. 이산, 연속, 연속, 이산, 이산

Exercise Chapter 2

(1) 어느 도시의 1일부터 31일까지의 낮 최고기온을 재보았더니 아래와 같습니다.

15.2	11.0	16.8	23.2	14.3	21.9	22.4
20.5	15.0	17.0	12.8	21.0	27.7	28.0
18.8	16.4	14.9	20.0	23.5	23.9	24.0
13.2	13.6	24.1	25.9	30.8	26.3	32.1
29.2	31.5	28.5				

- 등급의 수를 5개로 하여 도수분포표를 작성해보세요.

※ 해당 예제는 현대 통계학 교재를 참고하였습니다.

Exercise Chapter 2

① 최고측정값과 최저측정값은 각각 32.1과 11이다. 따라서 등급구간은 $\frac{(32.1-11)}{5}$ = 4.22임.

따라서 가까운 정수 5로 구간을 나누면 됨.

기온(X_i)	정확한계	빈도수(f)	상대빈도	누적상대빈도	누적백분율
10~14	9.5~14.5	5	0.16	0.16	16
15~19	14.5~19.5	7	0.23	0.39	39
20~24	19.5~24.5	10	0.32	0.71	71
25~29	24.5~29.5	6	0.19	0.90	90
30~34	29.5~34.5	3	0.10	1.00	100(%)
합계		31	1.00		

Chapter 3. 분포의 특성

Daniel Rex Park
(박영식)

Chapter 3 Intro

- 자료의 분포를 나타내는 특성
- 집중화경향(Central tendency)
- 분산도(degree of dispersion)
- 비대칭도(skewness)
- 집중화를 대표하는 값, 산술평균
- 분산도를 나타내는 값, 표준편차/분산

Daniel Rex Park
(박영식)

Chapter 3 분포의 특성

- 어느 등급에는 많은 빈도수가 있으며, 다른 곳에는 빈도수가 적다든지 하는 것
- 이러한 것을 나타내는 것이 집중화 경향(central tendency)
- 집중화 경향을 나타내는 수치는 분포의 대표값

집중화 경향

집중화 경향은 관찰된 자료들이 어디에 집중되어 있는가를 나타내주는 것으로서, 이 중 대표적인 것으로는 산술평균, 중앙값, 최빈값등이 존재

- 집중화 경향을 나타내는 분포의 대표값은 산술평균, 중앙값, 최빈값 외에도 기하평균, 조화평균 등이 존재하나, 기하평균과 조화평균 등은 자주 사용되지 않으므로 여기 기초과정에서는 3가지의 대표값만 설명

Chapter 3 분포의 특성

◎ 최빈값

- 최빈값(mode)은 자료의 분포에서 빈도수가 어느 곳에 가장 많이 모여 있는가를 나타냄
- 최빈값은 양적 자료, 질적 자료에서 모두 쓰임

최빈값

최빈값은 빈도수가 가장 많이 발생한 관찰값

질적 자료에서의 최빈값

- 어느 판매점에서 가방 판매량을 크기에 따라 조사하여 본 결과 아래와 같았음
- 이에 따르면 크기가 '중'인 티셔츠가 60벌로 가장 많으므로 최빈값(가장 빈번한 값)은 '중'

크 기	수 량
소	10
중	60
대	10
특대	5

◎ 최빈값

양적 자료에서의 최빈값

이산적 자료의 경우

- 도시와 농촌에서 각각 **100**세대를 표본으로 뽑아서, **자녀의 수**가 **몇 명**인가를 알아보았더니 아래와 같음
  도시에서는 자녀의 수가 **2**명인 세대가 가장 많으며, 농촌에서는 자녀의 수가 **3**명인 세대가 가장 많음
  그러므로 도시에서 자녀 수의 최빈값은 **2**명이고, 농촌에서 자녀수의 최빈값은 **3**명

도시의 자녀수		**농촌의 자료수**	
자녀수	세대수	자녀수	세대수
0	5	0	3
1	21	1	5
2	40	2	10
3	17	3	37
4	12	4	25
5	5	5	20
합계	100		100

◎ 최빈값

양적 자료에서의 최빈값

연속적 자료의 경우

- **연속적 자료**의 경우에서 **관찰값**이 모두 다르다면, 이때는 **최빈값**이란 있을 수 없음
- **연속적 자료**의 경우 자료를 등급으로 묶어서 각 등급에 해당되는 빈도수를 통해 **최빈등급(modal class)**를 **추출**
- **같은 등급**안에서는 **등급의 중간**에 **빈도수가 집중**되어 있을 것이라는 가정하에 그 **등급의 중간점**이 **최빈값**이 됨
- 유의점 : **등급의 중간점**은 **등급**을 어떻게 나누느냐에 따라 달라지므로 **최빈값**도 달라질 수 있음
  등급의 **중간점**은 보기에 **편리한 숫자로 정하는 것**이 좋음
- 어떤 회사에서 생산되는 타이어의 수명이 어느 정도인가를 알아보기 위하여 타이어 100개의 수명을 조사했더니 아래와 같았다. 이 표에서는 타이어 수명의 등급구간을 2,000~2,199시간 등으로 구분하고 그 구간의 중간점을 2,100시간 등으로 표시하였음 :: 최빈값 2300

수명시간	중간점	수 량
1,800~1,999	1,900	8
2,000~2,199	2,100	25
2,200~2,399	2,300	32
2,400~2,599	2,500	18
2,600~2,799	2,700	12
2,800~2,999	2,900	5
합 계		100

Chapter 3 분포의 특성

◎ 중앙값

- 중앙값(median)은 숫자로 표시되는 양적 자료에만 사용되는 것으로서 그 의미 아래와 같음

중앙값

중앙값은 수치로 된 자료를 크기순서대로 나열할 때, 가장 가운데에 위치하는 관찰값
표기법은 Md이며, Md의 공식은 아래를 참고

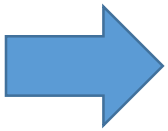
$$Md = \frac{(n+1)}{2} \text{ 번째 관찰값}$$

- **n**개의 관찰값이 있을 때, **(n+1)/2**번째로 큰 값이 바로 중앙값
- 그런데 **(n+1)/2**번째로 큰 수치는 **(n+1)/2**번째로 작은 수치와 동일
- 즉, 중앙값은 자료를 크기순서로 정리하였을 때 한가운데 있는 수치이기 때문에, 큰 쪽부터 계산하나 작은 쪽부터 계산하나 동일

이산적 자료

- 9명의 학생이 시험을 본 결과 점수가 다음과 같았다.

7,3,6,8,2,7,9,5,4



- 중앙값을 구하기 위한 정렬

2,3,4,5,**6**,7,7,8,9

(n+1)/2 → (9+1)/2 = 5번째 수 → 6

Chapter 3 분포의 특성

◎ 중앙값

- 만일, 한 사람이 더 포함되어 관찰값의 수가 n=10인 경우라면 ???

2,3,4,5,**6,7**,7,8,9,9

연속적 자료

- 연속적 자료에서 중앙값을 구하는 방법은 이산적 자료와 동일
- 그러나 이산적 자료와 연속적 자료가 도수 분포표로 표시되어 있을 때에는 그 계산의 복잡도 증가
- 아래의 도수 분포표에서 중앙값을 계산해보자.

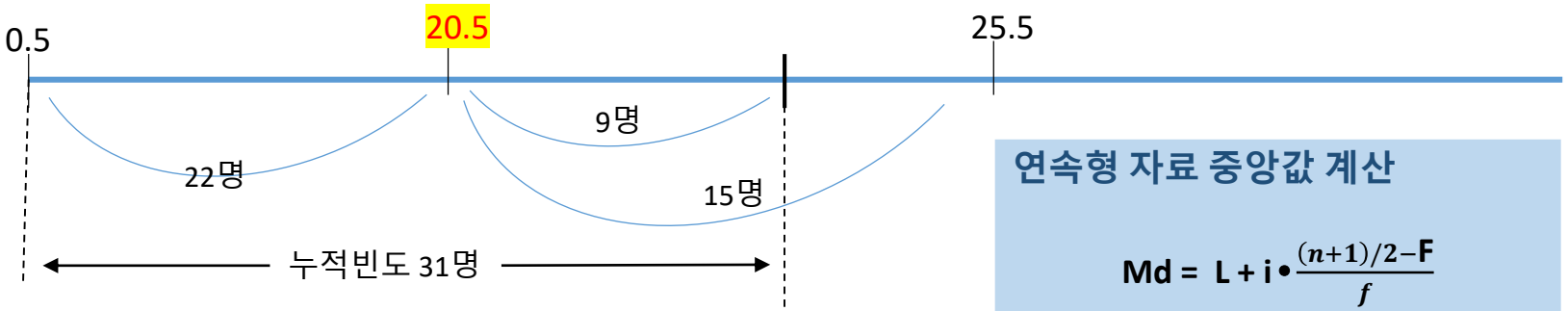
측정치(X)	빈도(f)	누적빈도
1 ~ 5	2	2
6 ~ 10	4	6
11 ~ 15	9	15
16 ~ 20	7	22
21 ~ 25	15	37
26 ~ 30	6	43
31 ~ 35	8	51
36 ~ 40	5	56
41 ~ 45	5	61

Chapter 3 분포의 특성

◎ 중앙값

연속적 자료

- **총관찰대상수가 n=61**이므로, **중앙값은(n+1)/2 = (61+1) /2 = 31, 즉 31번째로 큰 수치**가 **중앙값**
- 앞 페이지에서의 도수분포표에서 31번쨰로 큰 수는 21~25인 등급에 포함되어 있음
- 이 등급의 정확한계는 20.5~25.5고, 등급의 구간은 5, 이 등급 이하에는 **22명의 누적빈도**,
  이 등급 구간에는 **15명**이 있으므로 **15명 중에서 9번째에 있는 수치**를 구하면 **중앙값**



15명 : 9명 = 5 : x

$x = \frac{9}{15} \times 5 = 3$, 20.5 + 3 = 23.5

연속형 자료 중앙값 계산

$$Md = L + i \cdot \frac{(n+1)/2 - F}{f}$$

L : 중앙값이 있는 구간의 정확하한계
F : L까지의 누적빈도
f : 중앙값이 있는 구간의 빈도
n : 총관찰수
i:구간의 크기

Chapter 3 분포의 특성

◎ 산술평균

- 산술평균(arithmetic mean)은 양적 작료에만 사용되는 것
- 집중화경향을 나타내는 척도 중에서 가장 많이 사용되는 값
- 산술평균은 간단히 평균이라고도 하며, 기술통계뿐만 아니라 추리통계에서도 매우 중요한 역할

산술평균의 계산

- N개로 구성된 모집단의 관찰값을 $X_1, X_2, \dots, X_n$이라 할 때, 모집단의 평균 μ(뮤)는 다음과 같이 계산

모집단의 평균

$$\mu = \frac{X_1 + X_2 + \dots + X_n}{N} = \frac{\sum X_i}{N}$$

예제1) N사의 평균 재직 기간이 아래와 같다고 한다면, 이들의 평균 재직기간은?

재직기간(년) : 12, 25, 3, 8, 6, 16, 5, 3

$$\mu = \frac{X_1 + X_2 + \dots + X_n}{N} = \frac{\sum X_i}{N} = \frac{12+25+3+8+6+16+5+3}{8} = 9.8(\text{년})$$

◎ 산술평균

- 앞선 예제는 관심의 대상이 되는 집단에 포함된 모든 구성원들이 고려되었으므로 9.8년은 모집단의 산술평균이 됨.
- 모집단 평균은 μ vs 표본평균은 $\bar{X}$(엑스바) 계산방법은 **기술 통계학에서는 동일**
- **'기술(Descriptive) 통계학'**에서는 **모집단**과 **표본**을 구분할 필요가 없기에, 여기서는 산술평균을 **$\bar{X}$로 표시**

- N개로 구성된 모집단에서 n개 크기의 표본을 뽑았을 때 표본 관찰값을 $X_1, X_2, \ldots, X_n$이라 한다면, 표본평균 $\bar{X}$ 는 다음과 같이 계산된다.

표본의 평균

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum X_i}{n}$$

예제2) 인적성검사의 점수가 평균 600점 이상이어야 입사할 수 있다는 S사는 신입사원이 입사 후 얼마나 성적을 유지하고 있는지 궁금하여 입학한지 6개월이 지난 신입사원들 중 무작위로 10명을 뽑아 인적성검사 시험을 보게 하였다. 이들의 평균 점수는?

점수: 560, 625, 582, 505, 480, 650, 532, 510, 615, 632

$$\bar{X} = \frac{\sum X_i}{n} \frac{560+625+582+505+480+650+532+510+615+632}{10} = 569.1(\text{점})$$

◎ 가중산술평균

- **A = (3,4,5,8)**의 값을 가진 집단
- **B = (5,6,7,8,10,12)**의 값을 가진 집단
  각각의 평균을 구한다고 한다면, $\overline{X_A}$ = 5, $\overline{X_B}$ = 8
- **두 집단의 평균은 (5 + 8) / 2 = 6.5로 계산하면 틀린 계산**
- $\overline{X_A}$인 집단의 가중치는 4/10이고, $\overline{X_B}$인 집단의 가중치는 6/10이므로 전체 평균은 6.8이 된다.

$$\bar{X} = \frac{4 * 5 + 6 * 8}{10} = 6.8$$

가중산술평균

$$\overline{X} = \frac{n_1\overline{X_1} + n_2\overline{X_2} + \ldots + n_k\overline{X_k}}{n_1 + n_2 + \ldots + n_k} = \frac{\sum n_i\overline{X_i}}{\sum n_i}$$

◎ 가중산술평균

예제3) 서울시티투어 버스는 청계·고궁코스, 도심순환코스, 야간코스 세 가지가 있다. 서울시가 시티투어버스를 이용한 외국인을 상대로 시티투어버스의 평가점수를 알아보았더니 청계·고궁코스의 평균점수는 90점이고, 도심 순환코스의 평균 점수는 85점이며, 야간코스의 평균점수는 80점이었다. 세 가지 코스 전체에 대한 평균점수를 계산하면 얼마인가? 청계·고궁코스,이용객수는 60명, 도심순환코스 이용객 수는 40명, 야간코스 이용객 수는 50명이라 가정한다.

$\boldsymbol{n}_1$=60, $\overline{X_1}$ = 90
$\boldsymbol{n}_2$=40, $\overline{X_2}$ = 85
$\boldsymbol{n}_3$=50, $\overline{X_3}$ = 80

$$\bar{X} = \frac{\boldsymbol{n}_1\overline{X_1}+\boldsymbol{n}_2\overline{X_2}+\boldsymbol{n}_3\overline{X_3}}{\boldsymbol{n}_1+\boldsymbol{n}_2+\boldsymbol{n}_3} = \frac{60*90 + 40*85 +50*80}{60 +40+50} = \frac{12{,}800}{150} = 85.33(\text{점})$$

Chapter 3 분포의 특성

◎집중화 경향 대표값들의 위치와 특징

대표값의 종류	사례
중앙값(Median)	• 극단적인 관찰값의 영향을 거의 받지 않음 15,15,17,18,21,22,23,60 (모집인원 사례)
산술평균(Mean)	• 중앙값, 최빈값 : 수학적 연산 불가 • 산술평균: 수학적 연산이 가능하여 통계에 널리 사용됨
최빈값(Mode)	• 양적자료와 질적자료에 활용 가능 • 분포가 정규분포가 아닌 경우에 신뢰할 만한 대푯값이 아님 Ex) 구두 또는 책상 설계시 :: 최빈값의 키, 발의 크기, 몸무게 등을 사용

◎ 분산도

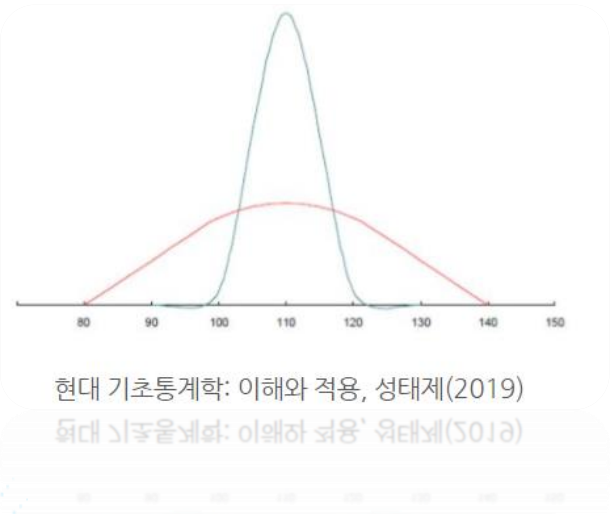
- 한 분포의 특성을 알아볼 때 집중화경향을 대표값이라고 하지만, 집중화경향만으로는 자료의 분포에 대한 충분한 정보를 얻을 수 없는 경우가 많다. 예를 들어 빅데이터 수업 쪽지 시험 결과 평균이 A클래스는 70점, B클래스는 65점이었다고 하자. 어느 학급이 더욱 바람직하다 할 수 있는가?

- 단순한 평균 값으로만 본다면, A클래스이나 점수의 분포를 살펴보면 단순한 문제가 아님을 알 수 있다.

Chapter 3 분포의 특성

◎ 분산도

- 어느 학급이 바람직스러운가에 대하여는 사람마다 다른 의견을 가질 수 있으나, 단지 평균이 높다는 것만으로 그 클래스의 점수분포가 더 좋다고 평가할 수는 없음
- 관찰값들이 흩어져 있는 정도를 살펴보는 것도 중요
- 분산도(degree of dispersion) : 관찰값의 흩어져 있는 정도



현대 기초통계학: 이해와 적용, 성태제(2019)

분산도

분산도란 관찰된 자료가 흩어져 있는 정도를 말하며, 분산도를 나타내는 방법으로는 범위, 평균편차, 표준편차 그리고 분산 등이 있다.

Chapter 3 분포의 특성

◎ 범위, 평균편차

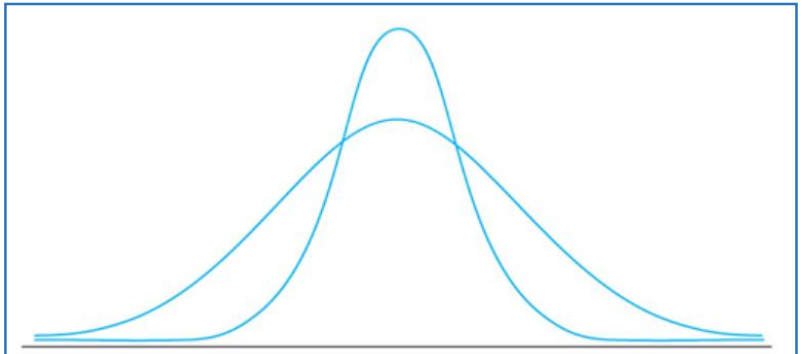
범 위(range)

범 위(range)

범위란 관찰값들 중에서 가장 큰 수치와 가장 작은 수치의 차이다.

* 한 강좌의 최고 매출이 36만원이고 최하 매출이 12만원이라고 한다면 매출의 범위는 36-12 = 24가 된다.

- **범위의 단점:** 극단적인 수치들 사이의 차이만 나타낼 뿐 그 극단적인 수치들 사이에서의 분포양상은 전혀 설명을 못하고 있음



범위는 같지만 분포의 성격이 다른 경우

Chapter 3 분포의 특성

◎ 범위, 평균편차

평균편차 (average deviation)

- **평균편차의 장점 :** 범위보다 과학적으로 분산도를 측정하는 방법

평균편차(average deviation)

평균편차는 관찰값과 산술평균과의 차이들의 평균으로써, 평균편차를 AD라고 하며, 이를 구하기 위한 식은 다음과 같다.

$$\mathbf{AD} = \frac{\sum|X_i - \bar{X}|}{n}$$

- **평균편차의 단점 :** 관찰값과 산술평균이 차이를 절대치를 사용해서 계산한다는 점이다. 절대값부호가 없는 평균의 특성 때문에 언제나 0이 되기 때문이다.

Chapter 3 분포의 특성

◎ 분산, 표준편차

분산과 표준편차의 개념 (A concept of Variance and Stadard deviation)

- 분산과 표준편차는 분포의 분산도를 나타내는 개념 중에서 가장 많이 쓰인다.
- **모집단의 분산**과 **표준편차**는 σ^2, σ라고 **표시**하고
- **표본을 대상**으로 한 **분산**과 **표준편차**는 s^2, s 로 **표시**
- 산술평균과 마찬가지로 기술통계학에서는 그 구분을 엄격히 할 필요가 없기 때문에 분산은 s^2, 표준편차는 s로 표시하여 설명하겠다.

분산과 표준편차의 계산 (Calculation of Variance and Stadard deviation) - 1

- 한 집단의 분산도를 구하기 위해서는 특정한 기준치를 설정하여 우선 계산 (일반적으로는 대표값=평균이 된다)
- 그런데 각 차이의 합인 $\sum(X_i - \bar{X})$는 0이 되므로, 단순히 관찰값과 평균과의 차이(이하 편차)를 더한 것만으로는 분산도를 구할 수 없다.

Chapter 3 분포의 특성

◎ 분산, 표준편차

분산과 표준편차의 계산 (Calculation of Variance and Stadard deviation) - 2

- 분산이란 각각의 관찰값에 대한 평균과의 편차를 제곱하여 그 평균을 구한 것!!!
- 만약 모든 관찰값들이 다 같다면, $X_1 = X_1 = X_1 \ldots = X_N = \mu$ 가 성립한다면, 분산은 σ^2은 **0**
- 분산의 단점 :: 편차를 제곱하였기 때문에 분산의 단위로 관찰 값의 단위를 그대로 사용할 수 없는 문제
  Ex) 학생들의 키를 cm단위로 재었다면, 분산의 경우 제곱단위여서 분산의 단위는 cm^2 가 됨. 이는 면적의 단위이지 키를 재는 단위가 아님
- 이와 같은 **분산의 한계를 극복**하기 위하여 분산을 계산한 뒤 그것의 제곱근인 **표준편차**를 자주 **사용**!!!

모집단의 분산	모집단의 표준편차
$\sigma^2 = \frac{\sum(X_i - \mu)^2}{N}$	$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(X_i - \mu)^2}{N}}$

Chapter 3 분포의 특성

◎ 분산, 표준편차

분산과 표준편차의 개념 (aconcept of Variance and Stadard deviation) - 3

- 분산과 표준편차는 분포의 분산도를 나타내는 개념 중에서 가장 많이 쓰인다. **모집단의 분산**과 **표준편차**는 σ^2, σ라고 표시하고 표본을 대상으로 한 **분산**과 **표준편차**는 s^2, s 로 나타낸다. 평균과 마찬가지로 그 구분을 기술통계학에서는 엄격히 할 필요가 없기 때문에 분산은 s^2, 표준편차는 s로 표시하여 설명

표본의 분산	표본의 표준편차
$S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$	$S = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$

- n대신 (n-1)로 나누어줌으로써 모집단의 σ를 추정하는데 적절한 표준편차를 구하기 위함
  :: 모집단에서 표본을 활용하는 **자유도**의 개념

Chapter 3 분포의 특성

◎ 분산, 표준편차

기술통계에서 분산과 표준편차

- 평균값$\bar{X}$를 게산할 때는 그 대상이 모집단이건 표본이건 상관없이 기호만 달리 사용할 뿐 계산 방식은 같음 그러나 분산과 표준편차를 계산할 때는 위에서 본 것과 같이 그 대상이 모집단인지 또는 표본인지에 따라 계산 방식이 약간 다름

- 그것은 표본에서 통계값을 구하는 것은 모수를 추정하기 위해서인데, 평균의 경우에는 표본의 통계량 $\bar{X}$가 모수 **μ에 대한 추정치의 역할을 할 수 있지만, 표준편차의 경우에는 계산공식에서 분모의 n대신 (n-1)을 사용해야만 모수의 추정값 역할을 할 수 있기 때문**

- 기술통계에서는 모집단과 표본을 구분하지 않고 연구자료에 대한 분포만을 알아보는 것이 목적이므로 분산과 표준편차를 구하는 식에서 분모에 **n을 그대로 사용**

기술통계에서의 분산과 표준편차

$$S^2 = \frac{\sum(X_i - \bar{X})^2}{n}$$

$$S = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n}}$$

Chapter 3 분포의 특성

◎ 분산, 표준편차

예제 5

- 헬스장 100일 챌린지에서 6명의 회원들이 감량한 자료는 아래와 같다. 이들의 분산과 표준편차를 구하시오.

5, 7, 12, 10, 4, 7(단위 : kg)

X_i		$X_i - \bar{X}$	$(X_i - \bar{X})^2$
5		-2.5	6.25
7		-0.5	0.25
12		4.5	20.25
10		2.5	6.25
4		-3.5	12.25
7		-0.5	0.25
합계	45		45.5

$$\bar{X} = \frac{45}{6} = 7.5 \text{ (kg)}$$

$$S^2 = \frac{\sum(X_i - \bar{X})^2}{n} = \frac{45.5}{6} = 7.58333$$

$$S = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n}} = 2.754\text{(kg)}$$

◎ 분산, 표준편차

예제 5-1

- 예제 5-1의 자료를 가지고 새로운 식을 적용하여 표준편차를 계산해 보자.

표준편차를 간단히 계산하는 방법

$$S = \sqrt{\frac{\sum X_i^2}{n} - \left(\frac{\sum X_i}{n}\right)^2}$$

$n = 6, \quad \sum X_i = 45$

$$\sum X_i^2 = (5)^2 + (7)^2 + (12)^2 + (10)^2 + (4)^2 + (7)^2 = 383$$

$$\therefore S = \sqrt{\frac{383}{6} - \left(\frac{45}{6}\right)^2} = \sqrt{7.583} = 2.754(\text{kg})$$

Chapter 3 분포의 특성

◎ 비대칭도

- 분포의 성격을 더 잘 알기 위해 집중화 경향이나 분산도 외 분포의 모양이 대칭분포에서 얼마나 벗어났는가를 알아보는 방법도 존재
- 비대칭도(skewness) or 왜도 : 관찰값들이 어느 쪽으로 치우쳐 있는가

◎ 피어슨의 비대칭도

- 피어슨의 비대칭도계수 (Pearson`s coefficient of skewness) : 비대칭도를 측정하는 방법

피어슨의 비대칭도

$$sk = \frac{3(\bar{X} - Md)}{S}$$

sk : 피어슨의 비대칭도

S : 표준편차

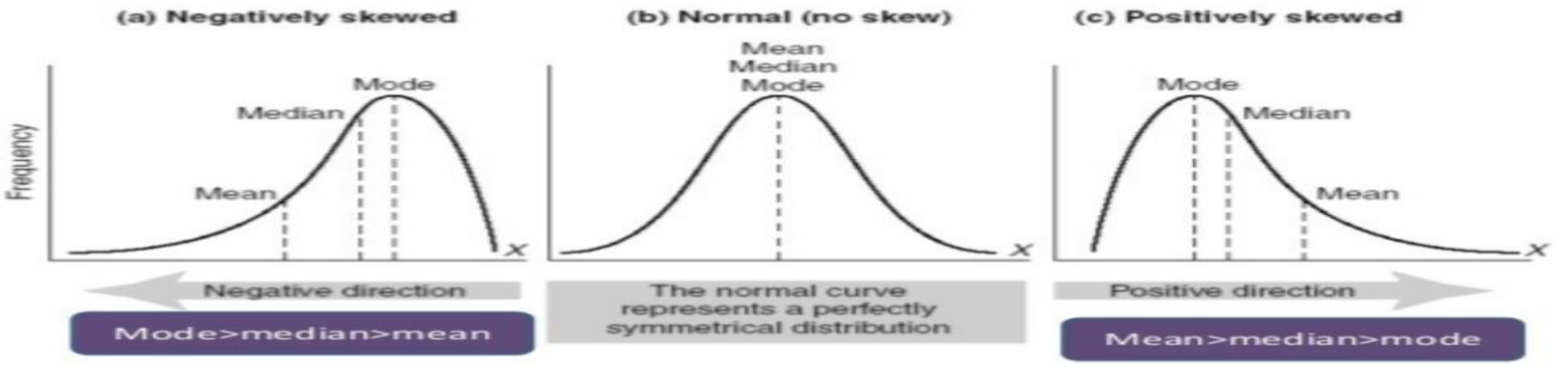
Md : 중앙값

- 피어슨의 비대칭도를 계산한 결과가 양(+)의 값인지, 음(-)의 값인지에 따라 방향과 정도를 알 수 있음
- 왼쪽꼬리분포에서는 음(-)의 값이
  오른쪽꼬리분포에서는 양(+)의 값이

Chapter 3 분포의 특성

◎집중화 경향 대표값들의 위치와 특징

- 분포의 양상에 따라서 평균, 중앙값, 최빈값이 차지하는 위치는 서로 다르다. 분포의 모양이 (b)와 같이 좌우대칭일 경우에는 **3개(평균, 중앙값, 최빈값)**의 값이 **일치**하게 되나, 완전하게 대칭이 이루어지지 않은 분포에서는 평균, 중앙값, 최빈값이 각각 **서로 다른 경우**가 대부분이다. 대칭이 아닐 때에 **중앙값**은 항상 **평균과 최빈값** 사이에 있게 된다.



(a)왼쪽꼬리분포
(skewed to the left)

(b)대칭분포

(c)오른쪽꼬리분포
(skewed to the right)

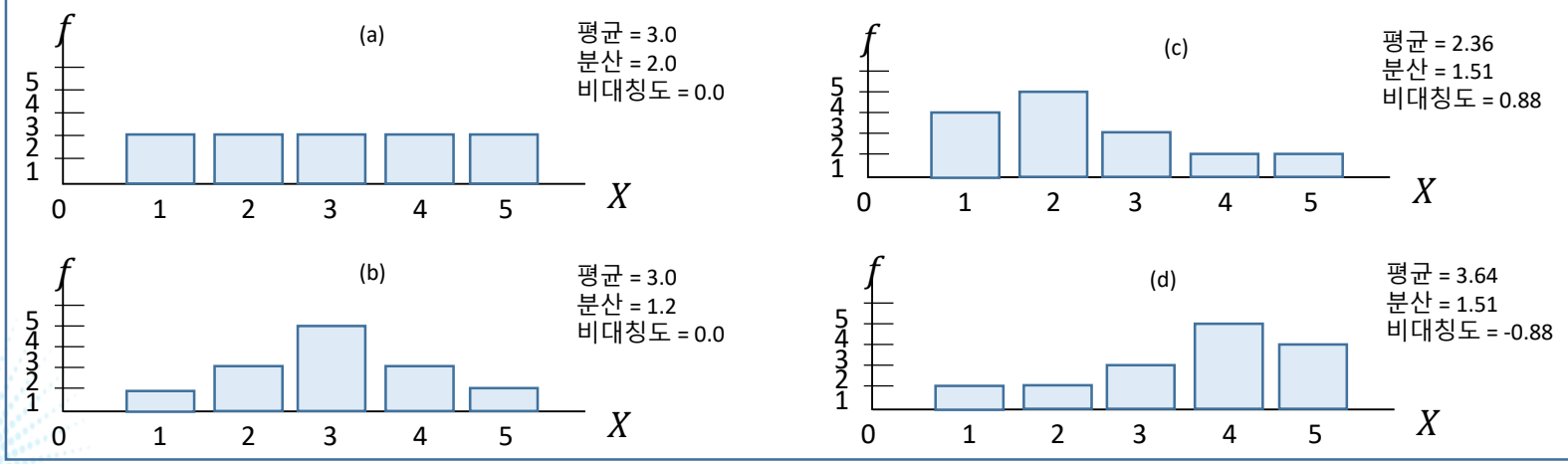
Chapter 3 분포의 특성

◎ 피어슨의 비대칭도

예제 6 우리나라의 국민 소득을 조사한 도수분포표에서 가구당 소득의 평균이 350만원, 중앙값이 335만 원이고, 가구당 소득의 표준편차가 12라면, 피어슨의 비대칭도는??

$$sk = \frac{3(\bar{X} - Md)}{S} = \frac{3(350-335)}{12} = 3.75$$ ::피어슨 비대칭도가 양(+)의 값을 갖고 있으므로, 우리나라 가구당 소득의 분포는 오른쪽으로 긴 꼬리를 갖고 있다는 것을 알 수 있음

분포가 다를 때 평균, 분산 및 비대칭도의 값



Exercise Chapter 3

(1) 산술평균, 중앙값, 최빈값을 구하라

(A) 3,5,2,6,5,9,5,2,8,6

(B) 51.6, 48.7, 50.3, 49.5, 48.9

(2) 각 15,20,25,30으로 구성된 4개의 학급이 있다. 각 학급의 평균 성적이 87,85,95,75 점이라고 한다면 전체 학생들의 평균 점수는?

(3) 용어 되짚어보기!!!

(A) 집중화경향

(B) 표준편차와 분산

(C) 비대칭도

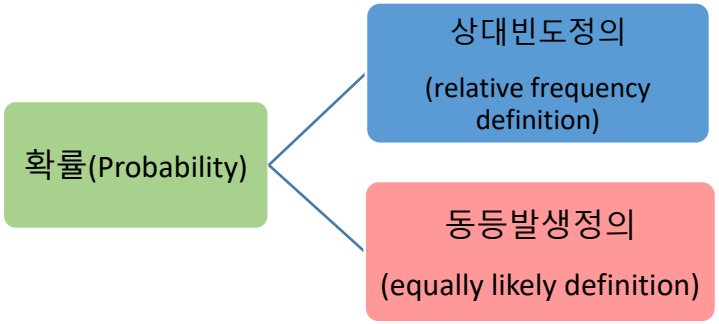
Chapter 4. 확률이론

Daniel Rex Park
(박영식)

Chapter 4 확률이론

◎ 상대빈도정의

- **확률(probability) :** 어떠한 상황이 발생할 가능성, 어떠한 사건(event)이 발생할 가능성



- 상대빈도정의 : **어떤 사건이 나타날 확률**은 **실험을 무한에 가깝게 계속적으로 시행**했을 때, 전체 시행횟수에서 그 사건이 나타나는 빈도수를 상대적으로 나타낸 것

상대빈도정의

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

$P(A)$: A사건이 발생할 확률

N : 총시행횟수

n : A사건이 발생한 횟수

◎ 동등발생정의

- **상대빈도정의의 한계** : 현실에서 실험을 무한에 가깝게 반복적으로 시행한다는 것에 대한 한계가 존재
동등발생정의라는 새로운 확률정의 방법의 대두

ex) 동전을 던질 때 **앞면이 나올 확률이 ½, 뒷면이 나올 확률 ½**, 라는 것은 무한히 실험을 시행하지 않더라도 추론이 가능

이유: **앞면**과 **뒷면**이 존재하고 두 가지 경우가 **발생할 확률이 동일**하다는 것에 대한 **논리적 추론**인 것

동등발생정의

$$P(A) = \frac{\text{사건 } A\text{에 속하는 경우의 수}}{\text{발생할 가능성이 동일한 전체 경우의 수}}$$

ex2) 우리 빅데이터 수업에 학생이 30명으로 구성되어 있는데, 남학생은 10명이고 여학생은 20명이다. 한 사람을 무작위(random)하게 뽑을 때 여학생일 확률은?

$$P(\text{여학생}) = \frac{\text{사건 } A\text{여학생 에 속하는 경우의 수}}{\text{전체 빅데이터 반의 학생의 수}} = \frac{20}{30} = \frac{2}{3}$$

Chapter 4 확률이론

◎ 상대빈도정의 vs 동등발생정의

Conclusion

- **동등발생정의**에 **입각**한 **확률**은 전체에서 어떤 특정사건이 차지하는 경우의 **구성비율(proportion)**과 같은 **의미**
- **상대빈도정의와 동등발생정의**는 서로 상호보완관계
  ex) 상대빈도정의의 사용이 어려울 경우에는 동등발생정의에 입각하여 확률 정의가 가능
  동전이 **불균형한 상태**의 경우와 같이 동등발생정의의 이용이 어려울 때에는 **상대빈도정의**에 의한 **확률 정의**

:: 어떠한 방법으로 확률을 정의하든 결과는 동일

제 2절 집합이론과 확률

집합이론

- 확률이론을 쉽게 설명하기 위한 집합이론의 용어와 부호를 사용하는 방법을 살펴보는 것이 좋음
- 집합(set) : 원소(element)의 모임 ex) A = {남자, 여자} , B = {아버지, 어머니, 나, 동생} , C = {1,2,3,4,5,6}
- 전체집합(universial set) : 특정 문제에서 가능한 모든 원소의 집합
- 부분집합(subset) : 전체 집합에서의 일부집합

Chapter 4 확률이론

◎ 집합이론

여집합

- 전체집합과 관심이 되는 부분집합 A가 정의되면 **전체집합A** 중 부분집합을 **포함하지 않는 부분**이 생기는데, 이를 여집합이라고 하며 A^c로 표시

합집합

- 만약 **집합 A**와 **집합 B**가 있다고 할 때 **합집합**은 **집합 A**나 **집합 B** 중 **하나**에 속하거나 **전체 다 속할 때**를 **의미**
- AND, OR 중 **OR의 성격**을 지닌다.
- 집합 A와 B의 합집합은 A ∪ B로 표시 ex) A={1,2,3}, B={4,5,6}, A ∪ B = {1,2,3,4,5,6}

교집합

- 두 개의 **집합 A**와 **집합 B**의 **교집합**은 **집합 A**나 **집합 B**에 공통적으로 속해 있는 원소를 뜻함
- AND, OR 중 **AND의 성격**을 지닌다.
- 집합 A와 B의 교집합은 A ∩ B로 표시 ex) A={1,2,3}, B={3,4,5,6}, A ∩ B = {3}

◎ 집합이론

합집합의 계산

- A U B는 A={1,2,3} + B={3,4,5,6}이라면 3이 두 번 세어진다(Counting). 두 번 세어진 3은 집합 A와 B의 교집합이므로 A U B = {1,2,3} + {3,4,5,6} – {3} = {1,2,3,4,5,6}으로 표현이 가능

합집합 계산

$$AUB = A + B - A \cap B$$

배타적 집합

- 배타적 집합(Mutually Exclusive) : $A \cap B = \emptyset$ 일 때 A와 B는 서로 배타적인 집합
  cf) 배타적 집합의 합집합 : AUB = A + B

- 예를 들어 집합 A={1,2,3}이고, 집합 B = {4,5,6,7}일 때에는 → {1,2,3} + {4,5,6,7} = {1,2,3,4,5,6,7}

Chapter 4 확률이론

◎ 집합이론

예제 4 -1 합집합 문제

100명의 학생이 시험을 보았다. 통계학 시험에서 60점 이상 받은 학생은 40명이었으며, 데이터 기획 시험에서 60점 이상 받은 학생이 50명이었다. 통계학 시험과 데이터 기획 시험에서 모두 60점 이상 받은 학생이 **15명**이라면, 적어도 한 과목에서 60점 이상 받은 학생은 몇 명 일까?

A. 60점 이상의 **통계학 집단**은 **40**명, **데이터 기획 집단**은 **50**명,
통계학 ∩ 데이터 기획 = **15**명

적어도 한 과목에서 60점 이상을 받는다는 의미는 **통계학 시험**에서 **60**점 이상을 취득하거나,
데이터 기획 시험에서 **60점 이상**을 **취득**했다는 의미!!!
따라서 **AND**가 아닌 **OR문제**임을 알 수 있음

OR :: 합집합 문제이므로 – 통계학 U 데이터 기획 - 통계학 ∩ 데이터 기획 = 40 + 50 - 15 =75명

Chapter 4 확률이론

◎ 집합이론과 확률이론

- 확률과 관련된 법칙은 매우 많음, 그러므로 기본적으로 잘 쓰이는 확률 법칙만 설명
- 확률의 **덧셈법칙, 곱셈법칙, 조건부확률, 동시확률, 독립 사건** 등에 대한 **개념 설명**

확률의 덧셈법칙

확률의 덧셈법칙

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

확률의 덧셈법칙 :: 배타적 사건이라면?

$$P(A \cup B) = P(A) + P(B)$$

예제 4 -2 **배타적 사건의 예**

- 빅데이터 반의 ADsP 자격증 모의고사 점수를 알아보았더니 60명 중에서 10명이 90점 이상, 20명이 80점 이상 90점 미만, 30명이 80점 미만이었다. 한 학생을 Random(무작위)로 추출했을때 90점 이상이거나 80점 미만일 확률을 구하시오.

A. P(90점이상) = $\frac{10}{60}$,

P(80점미만) = $\frac{30}{60}$ 이므로,

P(90점 이상 U 80점 미만) = P(90점이상) + P(80점미만)

$\frac{10}{60} + \frac{30}{60} = \frac{40}{60}$ = **0.67**

Chapter 4 확률이론

◎ 집합이론과 확률이론

예제 4 -3 배타적 사건의 예

• 주사위의 점이 4 또는 5가 나올 확률은 P(4 U 5)로 표시가 가능하다. 두 사건은 동시에 점이 호출될 수 없으므로 교집합이 존재하지 않는다. 즉, 4가 나오면 5가 주사위 점으로 나오지 않는다는 것이며, 5가 나온다는 것은 4가 나오지 않음을 의미한다.

P(4) = 1/6, P(5) = 1/6

P(4U5) = **1/6 + 1/6** = 2/6 = 1/3

◎ 조건부확률

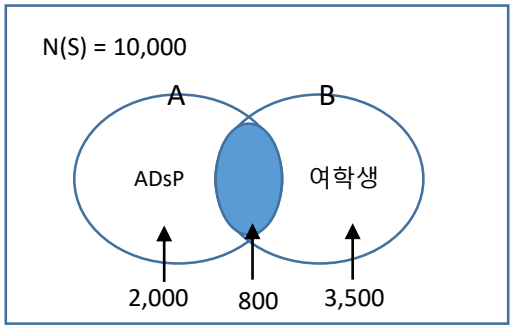
- 앞서 살펴본 확률의 경우 **한 번의 실험**에서 **발생**한 **상황**이었으나, **실제로는 두 단계** 혹은 **그 이상의 실험단계**가 **발생**
- 비복원추출의 흰 공 2개와 빨간 공 3개가 존재한다고 하자.
- 흰 공을 처음 뽑았을 때 나오는 확률 2/5, 그러나 비복원추출로써 다음에 흰 공을 뽑을 확률은 1/4이다
- 앞서 발생한 사건으로 인하여 두번째 실험의 표본공간이 변화하게 되는데, 이를 **조건부확률(conditional probability)**라 함

Chapter 4 확률이론

◎ 조건부확률

- 앞서 살펴본 확률의 경우 **한 번의 실험**에서 **발생**한 **상황**이었으나, **실제로는 두 단계** 혹은 **그 이상의 실험단계**가 **발생**함
- **조건부 확률 (Conditional Probability) :** **어떤 사건이 일어난** 또는 **일어날 조건**에서, 즉 **변화된 표본공간**에서 **어떤 사건**이 **일어날 확률**

예제 4 -4 **빅데이터 반**의 **전체 학생 10,000명** 중에서 **여학생은 3,500명**이다. **2,000명이 ADsP 시험에** 응시했으며, 이 중 **여학생**은 **800**명이다. **전체학생**을 **S**, **여학생**을 **B**, 그리고 **ADsP 응시자**를 **A**라고 하자. 그렇다면 이 때 **A∩B**는 여학생이며 **ADsP 응시학생**을 나타낸다. **ADsP**를 응시한 학생들 중 **여학생**일 **확률**은?



사건 B가 발생했다는 조건하에서 사건 A가 발생할 확률은 기호를 이용하여 나타내면 **P(A|B)**로 표시하며, **"사건 B**가 발생할 조건하에서 **사건 A**가 **발생할 확률"** 이라 함

A. 뽑힌 학생이 **ADsP 응시자**이라는 조건하에서 그 학생이 **여학생**일 가능성은 **P(여학생 | ADsP)**로 **표시**할 수 있으며 아래와 같이 계산이 가능

A- (1) :: P(여학생 | $ADsP$) $= \frac{\text{P(여학생} \cap ADsP)}{ADsP} = \frac{800}{2{,}000} = 0.40$

A- (2) :: P(여학생 | $ADsP$) $= \frac{\text{P(여학생} \cap ADsP)}{ADsP} = \frac{\text{P(여학생} \cap ADsP)/S}{ADsP/S} = \frac{800/10000}{2{,}000/10000}$

Chapter 4 확률이론

◎ 확률의 곱셈 법칙

- **확률의 덧셈법칙**은 집합이론에서 **합집합의 개념**에 대응되는 **확률**
- **확률의 곱셈법칙**은 집합이론에서 **교집합의 개념**에 대응되는 **확률**
- **교집합**의 **확률**은 **확률**의 **곱셈법칙**을 통해 쉽게 **이해**가 **가능**

확률의 곱셈법칙

$$P(A \cap B) = P(B) * P(A \mid B) = P(A) * P(B \mid A)$$

예제 4 -5 동시 확률의 예

- **사건 A**와 **B**가 **동시**에 일어날 **확률**은 **사건 A가 일어날 확률**과 사건A가 일어난 다음 **사건B가 일어날 확률의 곱**

예1) 파란공 2개, 빨간공 3개가 한 상자에 있고 비복원추출로 공 두 개를 차례로 뽑았을 때, 파란공 다음에 빨간공이 확률을 구하면!

P(파란 공∩빨간 공) =

P(파란 공) $*$ P(빨간 공 | 파란공)

$= \frac{2}{5} * \frac{3}{4} = \frac{6}{20}$

예2) 빅데이터 수강생은 10,000명이다. 그 중 빅데이터분석기사에 응시한 학생이 2,000명이며, 남학생은 6,500명이다. 남학생 중에서 빅데이터분석기사에 응시한 학생은 **15/65**이다. 그러면 어느 학생을 무작위로 선택했을 때, 그 학생이 빅데이터분석기사에 응시한 남학생인 확률은??

P(빅분기∩남학생) $= P$(남학생) $* P$(빅분기 | 남학생)

$= \frac{6{,}500}{10{,}000} * \frac{15}{65} = \frac{1500}{10{,}000} = \frac{1500}{10{,}000} = 0.15$

Chapter 4 확률이론

◎ 독립사건과 종속사건

독립사건의 정의

P(A | B) = P(A)
P(B | A) = P(B)

종속사건의 곱셈법칙

$$P(A \cap B) = P(B) * P(A \mid B) = P(A) * P(B \mid A)$$
$$P(A \cap B) = P(A) * P(B)$$

- **독립사건 (independent event) : 처음의 사건**이 다음에 **일어날 사건**에 **아무런 영향을 주지 않을 때** 두 사건은 독립사건
- **종속사건 (dependent event):** 조건부확률처럼 **한 사건의 발생이 다음 발생할 사건에 영향**을 주는 경우

독립사건의 예:
동전을 던질 때 **앞면이 나온 경우**와 **뒷면이 나올 확률**이 처음에 **앞면**이 나왔다고 해서 **영향이 미치지 않음**

종속사건의 예:
남녀 각각 **50명**씩 **100명**으로 구성된 모임에서 한 사람씩 차례로 성별을 조사하는 실험의 경우, **처음에 조사한 결과**가 **남자**였다면 **다음번**에 **또 다시 남자**가 뽑힐 확률은 **첫번째의 경우**와 달라지게 된다.

Chapter 4 확률이론

◎ 베이즈정리

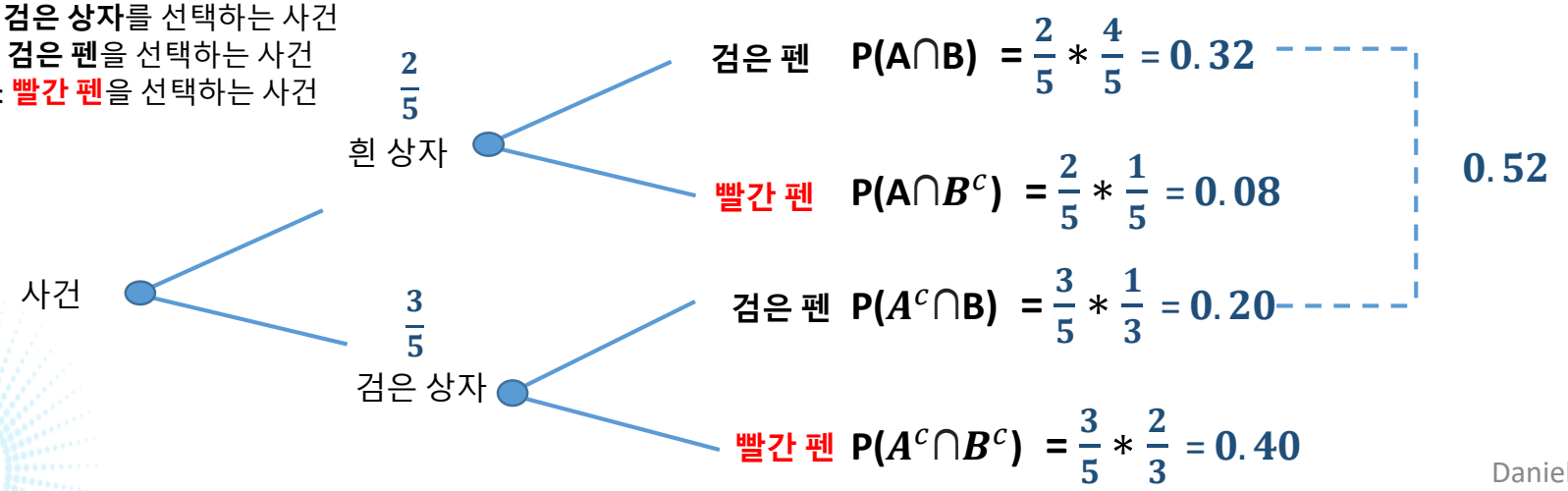
• **베이즈 정리 (Bayes' theorem) : 실험의 사건이 다음에 일어날 사건에 아무런 영향을 주지 않을 때 두 사건을 독립사건이라고 정의**

베이즈 정리의 예:
상자가 **5개** 있다고 가정할 때, **2개**는 **흰 상자**, **3개**는 **검은 상자**이다. 이 때 **흰 상자에는 빨간 펜 1개**, **검은 펜 4개**, **검은 상자에는 빨간 펜 2개**, **검은 펜 1개씩** 들어 있다. 어느 사람이 무작위로 1개의 펜을 뽑았을 때 **검은 펜**이 나왔다면, 이 사람이 **흰 상자를 택했을 확률을 구하시오.**

풀이 ::

해당 문제를 풀기 위해 **의사결정 나무(Decision tree)**로 도식화 해서 문제를 풀어보도록 하자.

A : 흰 상자를 선택하는 사건
$\bar{A}$**: 검은 상자**를 선택하는 사건
B : 검은 펜을 선택하는 사건
$\bar{B}$: **빨간 펜**을 선택하는 사건



◎ 베이즈정리

• **베이즈 정리 (Bayes' theorem)의 예 - 계속 -**

P(A∩B)와 **P(A^c∩B)**이 **검은 펜**이 나올 **확률**이므로 두 확률을 더하면 **검은 펜이 나올 확률**이 된다.
검은 펜이 흰 상자에서 나올 확률은 어떻게 구할까?
검은 펜이 흰 상자에서 나올 확률은 **P(A|B)**로 표시가 가능하다.

이를 식으로 표현하면,

$P(A|B) = \frac{P(A \cap B)}{P(A \cap B) + P(A^c \cap B)}$ 로 나타낼 수 있다.

이를 확장시켜서 정리하면 아래와 같이 정리가 가능하다.

베이즈 정리 (Bayes' theorem)

$$P(A_k|B) = \frac{P(A_k \cap B)}{\sum P(A_i \cap B)} \Rightarrow P(A_k|B) = \frac{P(A_k) * P(B|A_k)}{\sum P(A_i) * P(B|A_i)}$$

그런데 곱셈법칙으로 $P(A_i \cap B) = P(A_i) * P(B|A_i)$이므로

◎ 베이즈정리

- **베이즈 정리 (Bayes' theorem)의 예 - 계속 -**

베이즈 정리로 해당 문제를 풀게 된다면,

$$P(A \mid B) = \frac{P(A \cap B)}{P(A \cap B) + P(A^c \cap B)} = \frac{0.32}{0.32 + 0.20} = 0.615$$

또는

$$P(A \mid B) = \frac{P(A) * P(B|A)}{P(A) * P(B|A) + P(A^c) * (B|A^c)} = \frac{0.4 * 0.8}{0.4 * 0.8 + 0.6 * 0.33} = 0.615$$

예제 4 -3 비가 많이 올 확률 $P(A)$ = 0.4, 비가 중간 정도 올 확률 $P(B)$ = 0.3, 비가 아주 적게 올 확률은 $P(C)$ = 0.3 이고 비가 많이 올 때 풍년이 될 $P(K|A)$ = 0.6, 흉년이 될 확률 $P(K^c|A)$ = 0.4, 비가 중간 정도 올 때 풍년이 될 확률 $P(K|B)$ = 0.5, 흉년이 될 확률 $P(K^c|B)$ = 0.5, 비가 아주 적게 왔을 때 풍년이 될 확률 $P(K|C)$ = 0.2, 흉년이 될 확률 $P(K^c|C)$ = 0.8이라면,

* 풍년이 됐는데 비가 아주 적게 왔을 가능성은??

◎ 베이즈정리

예제 4 -3

풀이 : : 풍년이 됐는데, 비가 아주 적게 왔을 확률은 $P(C|K)$ 로 나타낼 수 있음

$$P(C|K) = \frac{P(C \cap K)}{P(A \cap K) + P(B \cap K) + P(C \cap K)}$$

$$= \frac{P(c) * P(K|c)}{P(A) * P(K|A) + P(B) * (K|B) + P(c) * P(K|c)}$$

$$= \frac{0.3 * 0.2}{0.4 * 0.6 + 0.3 * 0.5 + 0.3 * 0.2} = \frac{0.06}{0.45} = 0.133$$

Exercise Chapter 4

(1) 다음의 확률을 계산해보자.

① 1개의 주사위를 2회 던졌을 때 1이 계속해서 나올 확률

② 2개의 주사위를 동시에 던졌을 때 나온 수가 같을 확률

③ 1개의 주사위를 3회 던졌을 때 1이 한 번만 나올 확률

④ 3개의 주사위를 동시에 던졌을 때 1,2,3의 수가 나올 확률

Exercise Chapter 4

풀이

① 첫회에 1이 나올 확률은 1/6이고, 다음 회에 1이 나올 확률도 1/6이므로 곱셍정리에 의해 $\frac{1}{6} * \frac{1}{6} = \frac{1}{36}$ 이다.

② 두 개의 주사위 점이 모두 같은 수가 나오는 경우는 (1,1), (2,2), (3,3), (4,4), (5,5), (6,6) 총 6번이다. 이들은 배타적 사건이며 확률을 구하면, $6 * \frac{1}{36} = \frac{1}{6}$ 이다.

③ 처음에 1이 나오고 다음 2회 시행에는 1이 나오지 않을 확률은 $\frac{1}{6} * \frac{5}{6} * \frac{5}{6} = \frac{25}{216}$ 이다.
두 번째 1이 나오는 경우는 $\frac{5}{6} * \frac{1}{6} * \frac{5}{6} = \frac{25}{216}$, 세 번째 1이 나오는 경우는 $\frac{5}{6} * \frac{5}{6} * \frac{1}{6} = \frac{25}{216}$
즉, 확률 P는 $\frac{75}{216}$가 된다.

④ 3개의 주사위를 동시에 던졌을 때 1,2,3의 수가 나올 확률은 3! = 6이며, 세 개의 주사위를 던졌을 때 생길수 있는 모든 경우는 6 * 6 *6 = 216이다. 따라서 확률 P는 $\frac{6}{216}$

Exercise Chapter 4

(2) 어느 지역에 살고 있는 남녀성인의 집합 표본공간 S가 있다. 이들 중 한 사람을 상업회장으로 선출할 때 아래 표를 이용하여 P(M|E)를 구하시오.

M: 남자가 선택되는 경우
E: 상업에 종사하는 사람일 경우

구 분	상업 종사자(E)	상업외 종사자($\overline{E}$)
남(M)	360	140
여(W)	100	220

풀이 $P(M|E) = \frac{360}{460} =$ 약 0.783

Exercise Chapter 4

(3)빅데이터 반 학생이 ADsP에서 합격할 확률은 2/3이고, ADsP와 빅데이터 분석기사 모두 합격할 확률은 1/2 이다. 만일 그가 ADsP에 합격했음을 알고 있다면 빅데이터분석기사에서 합격할 확률은 얼마인가요?
(단, 두 자격증의 합격가능성은 서로 독립적이다.)

풀이

ADsP A라고 하고,
빅데이터분석기사를 B라고 하자.

$$\boldsymbol{P}(\mathrm{B} \mid \mathrm{A}) = \frac{\boldsymbol{P(A \cap B)}}{\boldsymbol{P(A)}} = \frac{1/2}{2/3} = \frac{3}{4} = 0.75$$

Exercise Chapter 4

(4) 우리나라에 비가 많이 올 때 풍년이 될 가능성은 0.8, 흉년이 될 가능성은 0.2고, 비가 적게 올 때 풍년이 될 가능성은 0.4, 흉년이 될 가능성은 0.6이다. 그렇다면 비가 많이 왔을 때 풍년과 흉년이 될 확률과 비가 적게 왔을 때 풍년과 흉년이 될 확률을 조건부확률로 표시하시오.

풀이

P(풍년 | 비가 많음) = 0.8

P(흉년 | 비가 많음) = 0.2

P(풍년 | 비가 적음) = 0.4

P(흉년 | 비가 적음) = 0.6

Exercise Chapter 4

(5) 시험을 치른 전체학생 중에서 국어합격자는 50%, 영어합격자는 60%이며, 두 과목 모두 합격한 학생은 15%라고 한다. 이때 임의로 한 학생을 뽑았을 경우, 이 학생이 국어에 합격한 학생이라면 영어에도 합격했을 확률은 몇 %나 될까?

풀이 $P(\text{영어} \mid \text{국어}) = \frac{P(\text{영어} \cap \text{국어})}{P(\text{국어})} = \frac{15\ \%}{50\ \%} = 0.3 = 30(\%)$

Chapter 5. 확률분포의 개념

Daniel Rex Park
(박영식)

Chapter 5 Intro

- 확률변수란 무엇인가?
- 확률분포란 무엇인가?
- 확률분포에 필요한 기대값과 분산
- 결합확률분포란?

Daniel Rex Park
(박영식)

Chapter 5 확률분포의 개념

◎ 확률변수의 개념

- **확률변수 : 일정한 확률을 가지고 발생하는 사건에 수치를 부여한 것**
- **보통 x로 표시**

예제 5 -1 주사위를 한 번 던질 때 나오는 숫자를 확률변수로 하여, 가능한 모든 결과, 즉 표본공간과 확률변수의 확률을 계산한다고 가정하자, 주사위를 던졌을 때의 가능한 결과는 6가지며, 이 때의 표본공간 S는 다음과 같다.

$$S = \{1,2,3,4,5,6\}$$

- 주사위에 나오는 숫자를 확률변수로 표시하여 나타내면 아래와 같이 나타낼 수 있다.

$$\boldsymbol{P}(X=1) = \frac{1}{6}, \quad \boldsymbol{P}(X=2) = \frac{1}{6}$$
$$\boldsymbol{P}(X=3) = \frac{1}{6}, \quad \boldsymbol{P}(X=4) = \frac{1}{6}$$
$$\boldsymbol{P}(X=5) = \frac{1}{6}, \quad \boldsymbol{P}(X=6) = \frac{1}{6}$$

Chapter 5 확률분포의 개념

◎ 확률분포

• 확률변수와 관련하여 동전의 사례를 들어보자. 동전을 두 번 던질때 나타날 수 있는 가능한 사건들과 각 사건이 나타나는 확률에 대해서는 아래와 같이 표현될 수 있다. 표본공간(S)에서 H를 앞면이라고 하고 T를 뒷면이라 하겠다.

표본공간		앞면의 수	각 사건의 확률
첫번째	두번째		
H	H	2	1/4
H	T	1	1/4
T	H	1	1/4
T	T	0	1/4

• **S = { HH, HT, TH, TT }**로 표시가 가능하다. 만약 확률변수를 앞면으로 하고자 한다면, 표본공간 S={2, 1, 0} 이 된다. 그러나 동전을 던질 때 꼭 앞면의 수를 확률변수로 정하지 않더라도, 뒷면을 기준으로 할 수도 있다.

• **어떤 확률 변수가 취할 수 있는 모든 값들**과, 이에 대응하는 각각의 확률을 정의해 놓은 것을 우리는 **확률분포(probability distribution)**라고 한다.

Chapter 5 확률분포의 개념

◎ 확률분포

확률변수(X_i) = (앞면의 수)	P(X_i) = (각 사건의 확률)
0	1/4
1	1/2
2	1/4

확률분포

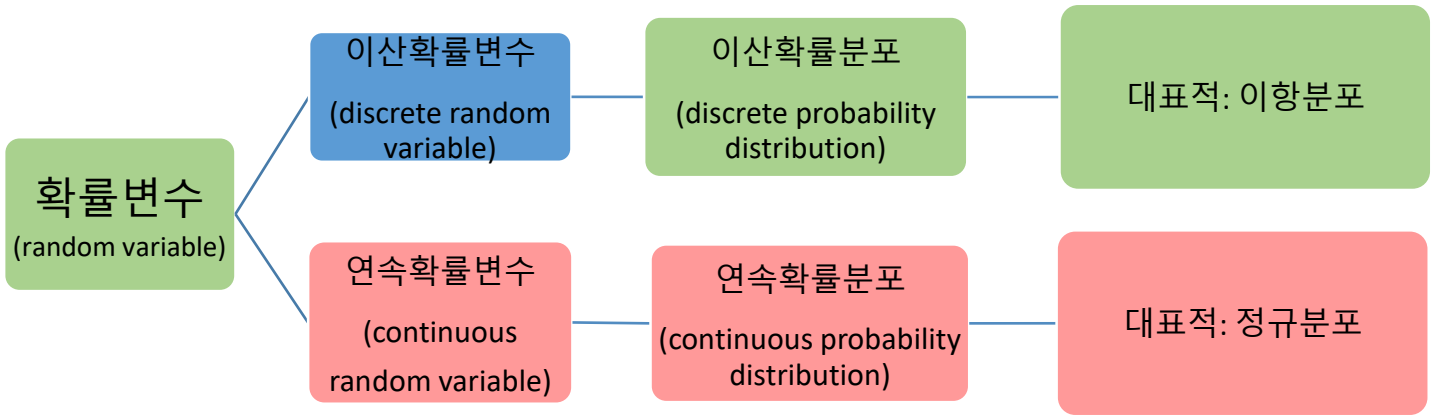
확률분포 : 어떤 **확률변수**가 취할 수 있는 **모든 값**들과 **이 값**들이 나타날 **확률**을 **표시**한 것

- 확률분포는 기계적인 결과만을 위한 것이 아니며, 그 외에도우리가 처해 있는 불확실한 상황에서 일어날 가능성이 있는 어떠한 사건이라도 주관적이건 객관적이건 간에 확률을 부여함으로써 확률분포로 표현할 수 있다.

예상 연봉 상승률(X_i)	P(X_i)
5 %	0.3
3 %	0.5
0 %	0.2

Chapter 5 확률분포의 개념

◎ 이산확률변수와 연속확률변수



◎ 확률함수

- 확률변수 : 확률시행으로 나타날 수 있는 여러 사건들에 일정한 수치를 부여한 것
- 확률분포 : 확률변수가 취하는 값에 대하여 합이 1인 확률이 어떻게 분포되어 있는지를 나타낸 것

- 확률함수(probability function) 또는
  확률밀도함수(probability density function) : 확률변수가 취할 수 있는 수치에 대해 그 확률 값이 얼마인지를 알려주는 함수

Chapter 5 확률분포의 개념

◎ 확률함수

- **확률함수는 확률변수가 취할 수 있는 모든 값에 대해 그 값을 가질 확률이 얼마인지를 알려주는 함수**

- 이산확률함수의 표기법 $\boldsymbol{P}(\mathrm{X}_i)$,
  이산확률함수 $\boldsymbol{P}(\cdot)$에서의 X_i의 확률값 : 확률변수 X가 X_i의 값을 가질 확률을 의미

Ex) 동전실험의 경우 :: 앞면(H)가 나올 확률 $\boldsymbol{P}(H)$가 $1/2$,
뒷면(T)이 나올 확률 $\boldsymbol{P}(T)$역시 $1/2$

$$\boldsymbol{P}(\mathrm{X}_i) = \frac{1}{2}\ ,\ \mathrm{X}_i\ =\ \{H, T\}$$

※ 이산확률변수의 확률분포를 나타내는 확률함수의 조건 2가지::

① 이산확률분포에서 특정한 값 X_i가 발생할 확률은 $0 \leq \boldsymbol{P}(\mathrm{X}_i) \leq 1$임

② $\sum \boldsymbol{P}(\mathrm{X}_i)\ =\ 1$

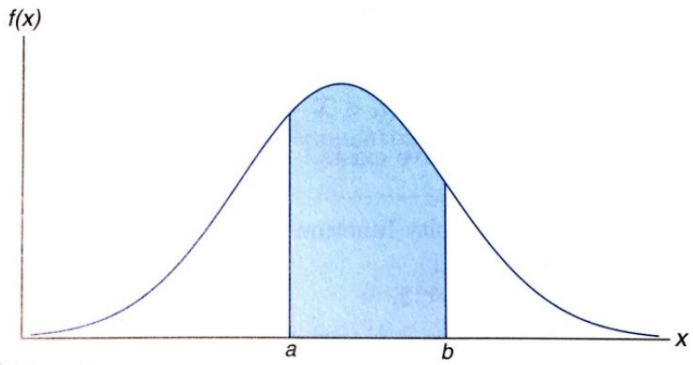
Chapter 5 확률분포의 개념

◎ 확률밀도함수

- **연속확률변수가 일정한 범위** 내에서 취할 수 있는 값은 **무한히 많음**
- 이러한 논리로 $\boldsymbol{P}(X_i) = 0$ :: 어떠한 값에만 국한된 확률을 말할 수 없으므로...
- 그러나 구간에 대한 확률은 계산이 가능하다

Ex) 1) 박쌤의 키가 165이상 170미만에 있을 확률,
2) 오늘 저녁의 기온이 0~5 ℃일 확률이 **45%**라고 한다면,

$\boldsymbol{P}\ (0 \leq (X) \leq 5) =$ **0.45**로 표현이 가능

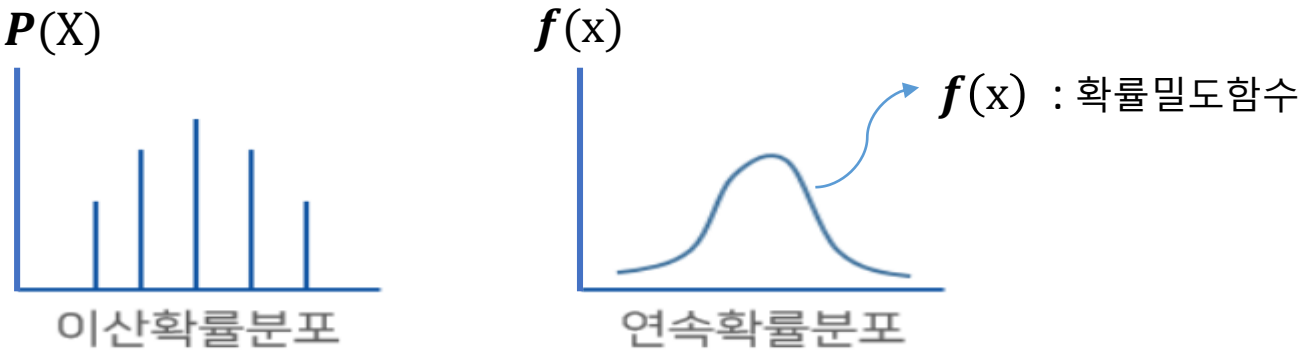


$$P(a < X < b) = \int_a^b f(x)\ dx.$$

- 옆의 **연속확률변수**의 **높이**와 확률과는 직접적인 관계가 없음
  :: 이유는 정해진 a와 b사이에는 무수한 변수의 값들이 존재하므로
  색칠된 범위의 넓이만 구할 수 있기 때문임
- 0도가 a이고 5도가 b라면, 그 사이의 확률은 계산이 가능해짐
- **if** 곡선 아래의 범위가 1이라면, 파란색 범위의 넓이는 **0.45**임
- **확률밀도함수(probability density function)**은 $f(X)$ 라고 표현되는데
  이는 단순히 그래프 모양을 나타내는 식에 불과함
  Cf) 이산확률변수에서의 $\mathrm{P}(X)$는 확률을 나타내고 있으나,
  확률밀도함수에서의 $f(X)$는 모양만 나타냄

Chapter 5 확률분포의 개념

◎ 확률밀도함수



연속확률분포의 특징

① 연속확률분포에서 어느 한 특정값 X_i가 발생할 확률 $P(X_i) = 0$

② 연속확률분포에서의 확률은 일정구간 사이의 값을 취할 확률로 계산됨
즉, $P(a \leq (X) \leq b)$**는 구간 [a,b] 사이의 확률밀도함수** $f(x)$와 X축 사이의 면적

③ 확률밀도함수는 언제나 음의 값이 아닌 양의 값(비음의 값)을 갖는다. $0 \leq f(x)$

④ 확률밀도함수 아래에 있는 전체 면적은 언제나 1

Chapter 5 확률분포의 개념

◎ 확률분포의 기대값과 분산

기대값의 계산

- 확률분포의 성격은 **'집중화 경향'**과 **'분산도'**로 **'분석'**
- 기대값(Expected Value) : 확률분포의 집중화경향 == 평균값(average, weighted average)과 같은 개념
- 기댓값은 다만 **확률변수**가 취할 수 있는 **모든 값들의 평균**의 의미
- 분산, 표준편차 : 확률분포의 분산도

기댓값

- 기댓값은 $E(x) = \sum X_i * P(X_i)$ 로 계산됨
- 기댓값은 간단히 말해 평균값(average, weighted average)과 같은 개념
- 기댓값은 **미래 발생확률이** **가장 높은 것을 의미하는 것이 아님**에 유의!!!

Ex) 어느 주식의 주가가 1,000원이 될 확률은 50%이고, 500원이 될 확률도 50%라면
기댓값은 750원이 된다.

:: 이 의미는 실제 주가가 750원이 되지 않더라도!
상황이 계속 진행되다 보면 결국 해당 주가는 750원이 될 것이라는 의미이다.

Chapter 5 확률분포의 개념

◎ 확률분포의 기댓값과 분산

예제 5 -2

- P기업이 A라는 프로젝트를 진행할 대 예상되는 수익과 확률은 다음과 같다.
  이 때의 기대수익은 얼마인가?

(단위: 억원)

X_i	$P(X_i)$	$X_i * P(X_i)$
- 600	0.5	-300
600	0.2	120
1,000	0.3	300
합계	0	120

- 풀이 :: 코리아 기업의 기대수익률은 -600 * 0.5 + 600 * 0.2 + 1,000 * 0.3 = **120억**

Chapter 5 확률분포의 개념

◎ 확률분포의 기댓값과 분산

- **기댓값의 특성**

- 확률변수 X의 기댓값 $E(x)$를 알고 있으면, 확률변수 X를 1차식으로 변환한 다른 확률변수의 기댓값도 이를 이용하여 쉽게 구할 수 있음

기댓값의 특성

① 확률변수 X에 일정한 상수 a를 곱한 확률변수의 기댓값은 확률변수 X의 기댓값에 a를 곱한 것임

$$E(ax) = a * E(x)$$

② 확률변수 X에 일정한 상수 b만큼을 가감한 확률변수의 기댓값은 확률변수 X의 기댓값에 b를 가감한 것과 같음

$$E(x \pm b) = E(x) \pm b$$

③ 위의 두 가지를 결합하면 아래가 성립 가능

$$E(ax \pm b) = a*E(x) \pm b$$

Chapter 5 확률분포의 개념

◎ 확률분포의 기댓값과 분산

예제 5 -3 동전을 던져 앞면이 나오면 400원을 받고, 뒷면이 나오면 800원을 받는 게임이 있다고 하자. 이 게임의 기댓값은 600원이 된다. 상금을 2배로 올린 후에 일률적으로 200원씩 더 올린다면 그때의 기댓값은???

Before

X_i	$P(X_i)$	$X_i * P(X_i)$
400	0.5	200
800	0.5	400
합계		600

After

$2X_i+200$	$P(X_i)$	$(2X_i+200) * P(X_i)$
1,000	0.5	500
1,800	0.5	900
합계		1,400

- 풀이 :: $E(x) = 600$(원), $E(2x+200) = 600$(원),
  $E(2x) + 200 = 2 * E(x) + 200 = \mathbf{2 * 600 + 200 = 1{,}400}$(원)

◎ 분산(variance)과 표준편차(standard deviation)

분산의 계산

$$\begin{aligned} \text{Var}(X) &= \sum[X_i - E(x)]^2 * \boldsymbol{P(X_i)} \\ &= E\,[\,[X - E(x)]^2] \\ &= E(X^2) - [E(x)]^2 \end{aligned}$$

- 위의 식은 분산이라는 것이 기댓값 $E(X)$를 중심으로 확률변수들이 얼마나 흩어져 있는 가를 나타내는 것
- 분산의 표시 : $\text{Var}(X)$ 또는 $\boldsymbol{\sigma^2}_x$ 로 표시
- 표준편차의 표시 : **분산의 제곱근**, $\boldsymbol{\sigma}_x$ 로 표시

표준편차의 계산

$$\boldsymbol{\sigma}_x = \sqrt{\sum[X_i - E(x)]^2 * \boldsymbol{P(X_i)}}$$

Chapter 5 확률분포의 개념

◎ 분산(variance)과 표준편차(standard deviation)

- 동전을 두 번 던지는 사례를 통해 분산과 표준편차를 구하는 방법

X_i	$P(X_i)$	$X_i * P(X_i)$	$X_i - E(X)$	$[X_i - E(x)]^2$	$[X_i - E(x)]^2 * P(X_i)$
0	1/4	0	-1	1	1 / 4
1	1/2	1/2	0	0	0
2	1/4	2/4	+1	1	1 / 4
합계		1			1 / 2

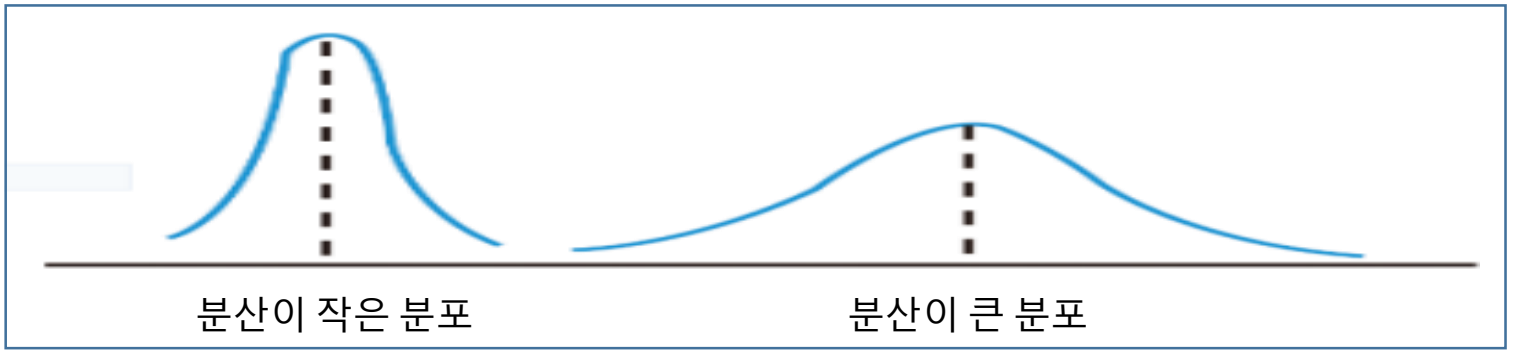
$$E(X) = \sum X_i * P(X_i) = 1$$

$$\text{Var}(X) = \sum [X_i - E(x)]^2 * P(X_i) = \frac{1}{2}$$

$$\sigma_x = \sqrt{\frac{1}{2}} = 0.71$$

Chapter 5 확률분포의 개념

◎ 분산(variance)과 표준편차(standard deviation)



분산과 표준편차의 특성

① 확률변수 X에 일정한 상수 b를 더한 확률변수의 분산은 본래의 확률변수의 분산과 같다. 확률변수에 상수를 더하는 것은 분포의 분산도에는 아무런 영향을 미치지 못함

$$\mathrm{Var}(x+b) = \mathrm{Var}(x)$$
$$\sigma(x+b) = \sigma(x)$$

② 확률변수 X에 일정한 상수 a를 곱한 확률변수의 분산은 본래의 확률변수의 분산에 a^2을 곱한 것과 같음

$$\mathrm{Var}(ax) = a^2\,\mathrm{Var}(x)$$
$$\sigma(ax) = a\,\sigma(x)$$

③ 위의 두 가지를 결합하면 아래가 성립 가능

$$\mathrm{Var}(ax+b) = a^2\,\mathrm{Var}(x)$$
$$\sigma(ax+b) = a\,\sigma(x)$$

Chapter 5 확률분포의 개념

◎ 분산(variance)과 표준편차(standard deviation)

예제 5 -5 어느 주식의 기대수익(X)은 $E(x) = 500$만 원이었으며, $\mathrm{Var}(x) = 100$이었다. 이 투자대상에 비해 수익이 절반이 되는 주식이 있다면, 기대수익과 표준편차는?

풀이 :: $E(0.5x) = 0.5 * 500$만 원 $= 250$만 원

$\mathrm{Var}(0.5x) = 0.5^2 * 100$만 원 $= 25$만 원

$\sigma(0.5x) = 0.5 * \sqrt{100} = 5$만 원

◎ 결합확률분포와 주변확률분포

Y_j(경제성장률) / X_i (실업률)	5%	10%	20%	X_i의 주변확률
3%	0.0	0.2	0.3	0.5
5%	0.1	0.1	0.2	0.4
10%	0.1	0.0	0.0	0.1
Y_j 의 주변확률	0.2	0.3	0.5	1.0

$E(x) = 3 * 0.5 + 5 * 0.4 + 10 * 0.1 = 4.5\%$

$E(y)$????

Chapter 5 확률분포의 개념

◎ 결합확률분포와 주변확률분포

결합확률분포

두 개 이상의 확률변수가 관련된 확률분포를 결합확률분포라고 함

주변확률분포

주변확률분포는 X와 Y의 결합분포에서 X 또는 Y의 어느 하나만의 확률분포를 말하며, 결합확률분포의 주변(margin)에 표시되기 때문에 이를 주변확률분포라고 함.

예를 들어, $P(X=5\%)$를 알고 싶다면 $P(X=5\%, Y=5\%)$, $P(X=5\%, Y=10\%)$, $P(X=5\%, Y=20\%)$를 모두 더하면 된다. 그 결과는 **0.4**이다. 이런 식으로 각 구한 값들은 X_i**의 주변확률**이라 할 수 있다.

Chapter 5 확률분포의 개념

◎ 공분산

- 결합확률분포와 관련된 가장 많이 사용되는 개념 중 하나
- 두 확률 분포가 결합될 때 그 결합확률분포의 분산을 측정하는 것!
- 표시 : Cov(X, Y)

확률변수 X와 Y의 공분산

$$\begin{aligned} \text{Cov}(X, Y) &= E\,[[X - E(x)][Y - E(Y)]] \\ &= E(XY) - E(X)E(Y) - E(Y)E(X) + E(X)E(Y) \\ &= E(XY) - E(X) * E(Y) \end{aligned}$$

$\text{Cov}(X, Y) = -6.5\ \%^2$

X_i (실업률) ①	Y_j(경제성장률)②	$P(X_i, Y_j)$ ③	$X_i Y_j$ ④	③ * ④	① * ③	② * ③
3%	10	0.2	30	6	0.6	2.0
3%	20	0.3	60	18	0.9	6.0
5%	5	0.1	25	2.5	0.5	0.5
5%	10	0.1	50	5	0.5	1.0
5%	20	0.2	100	20	1.0	4.0
10%	5	0.1	50	5	1.0	0.5
합 계				$E(XY) = 56.5$	$E(X) = 4.5$	$E(Y) = 14$

Exercise Chapter 5

(1) 다음 용어들을 간단히 정의해보자.
①확률변수 ②확률분포 ③기댓값
④표본공간 ⑤주변확률분포 ⑥공분산

(2) 동전을 세 번 던질 때 앞면이 나올 확률변수의 확률분포는 다음의 표와 같다. 이 확률변수의 기댓값과 분산, 표준편차를 구하라.

X_i	$P(X_i)$
0	1/8
1	3/8
2	3/8
3	1/8

Exercise Chapter 5

(3) T그룹에서 포장용 원두커피를 생산하고 있다. 그런데 그 원두커피의 용량을 조사하니 다음과 같았을 때, 포장용 원두커피(X)의 기댓값, 분산, 표준편차를 구하시오.

용량(X_i)	$P(X_i)$
340	0.2
360	0.6
380	0.2

Exercise Chapter 5

(3) T그룹에서 포장용 원두커피를 생산하고 있다. 그런데 그 원두커피의 용량을 조사하니 다음과 같았을 때, 포장용 원두커피(X)의 기댓값, 분산, 표준편차를 구하시오.

풀이 ::

X_i	$P(X_i)$	$X_i * P(X_i)$	$X_i - E(X)$	$[X_i - E(x)]^2$	$[X_i - E(x)]^2 * P(X_i)$
340	0.2	68	-20	400	80
360	0.6	216	0	0	0
380	0.2	76	+20	400	80
		$E(X) = 360$			$\sigma^2 = 160$
					$\sigma = \sqrt{160} = 12.65$

Exercise Chapter 5

(4) 어느 변수의 확률분포에서 기댓값 $E(x) = 50$, 표준편차 $\sigma_x = 30$을 구하였다. 이 확률변수를 2배하여 새로운 확률변수를 만들 때, 기댓값과 표준편차는 어떻게 될까?

풀이 :: $E(2x) = 2E(x) = 2 * 50 = 100$

$\sigma^2(2x) = 2^2\sigma^2(x) = 4 * 900 = 3{,}600$

$\sigma(2x) = 2\sigma(x) = 2 * 30 = 60$

Chapter 6. 이산확률분포

Daniel Rex Park
(박영식)

Chapter 6 Intro

- 이산확률분포란 무엇인가?
- 이항분포란 무엇인가?
- 다항분포란 무엇인가?
- 초기하분포란 무엇인가?
- 연속확률분포와 이산확률분포는 어떻게 다른가?

Daniel Rex Park
(박영식)

Chapter 6 이산확률분포

◎ 베르누이시행

베르누이시행의 조건

①각 시행의 결과는 상호배타적인 두 사건으로 구분가능함. 즉, 한 사건은 "성공"(S), 다른 사건은 "실패" (F)로 나타낸다.

②각 시행에서 성공의 결과가 나타날 확률은 p= $P(S)$로 나타내며, 실패가 나타날 확률은 q= $P(F)$ = 1-p로 나타낸다. 그러므로 각 시행에서 성공이 나타날 확률과 실패가 나타날 확률의 합은 **p + q = 1**이 된다.

③각 시행은 서로 독립적. 한 시행의 결과는 다음 시행의 결과에 아무런 영향을 주지 못함

Ex) 동전의 실험 :: 앞면 아니면 뒷면
주사위실험 :: 원하는 숫자가 나오면 "Success", 그 외 숫자는 "Fail"
코트색실험 :: 원하는 색깔이 나오면 "성공", 그 외 색깔은 "Fail"

※ 동전 던지기의 경우 앞면(H)을 목표로 한다면, 성공확률은 1/2 이며, 실패확률은 1-(1/2) = 1/2이다.
또한 지금 시행에서 나온 결과가 다음 시행에는 영향을 주지 않음

Chapter 6 이산확률분포

◎ 이항확률변수와 이항확률분포

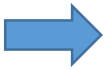
- 한 번의 베르누이시행을 통해 성공확률이나 실패확률을 알고 싶어하기 보다, 여러 번의 베르누이시행시 특정 횟수의 성공이 나타날 확률에 일반적으로 관심이 많음.

- 예를 들어 동전을 10번 던졌을 때, 2 번 앞면이 나올 확률, 주사위 2개를 동시에 던지는 실험을 3번 시행할 때 합이 3이 될 확률 등

- **이항확률변수(binomial random variable) : 어떤 시행에서 성공의 확률 또는 실패의 확률**
- **이항확률분포(binomial probability distribution) : 이항확률변수의 확률분포**

예제 6 -1 • 동전던지기를 3 번 시행할 때의 가능한 결과와 이항확률분포를 살펴보면 아래와 같다.

가능한 결과	첫째 시행	둘째 시행	셋째 시행	앞면의 수 (X_i)
1	H	H	H	3
2	H	H	T	2
3	H	T	H	2
4	H	T	T	1
5	T	H	H	2
6	T	H	T	1
7	T	T	H	1
8	T	T	T	0



앞면의 수(X_i)	$P(X_i)$
0	1/8
1	3/8
2	3/8
3	1/8

Chapter 6 이산확률분포

◎ 이항분포의 확률계산

- 시행횟수 n이 커지면 가능한 모든 경우를 감안하여 확률을 계산하고 분포의 모양을 알아내는 일은 대단히 복잡
- 이항실험의 결과인 이항분포는 선험적 분포
- **시행횟수 n과 성공 확률 p값만 알고 있다면 그 분포의 모양과 확률변수의 확률을 알 수 있음**
  1) 이항확률함수의 이용
  2) 이항분포표의 이용

이항확률함수의 이용

이항확률함수

$$P(X=x) = {}_nC_x p^x (1-p)^{n-x}$$

x: 성공횟수
n: 시행횟수
p: 성공확률
1- $p = q$: 실패확률

Chapter 6 이산확률분포

◎ 이항분포의 확률계산

예제 6 -2 동전을 세 번 던졌을 때 앞면이 나오는 횟수 각각의 확률을 알고자 한다.
이 때 $n = 3, p = 1/2, q = 1\text{-}p = 1/2$이다.

앞면이 나올 수 있는 확률은 다음과 같다.

$$P(X = 0) = {}_3C_0 * \left(\frac{1}{2}\right)^0 * \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

$$P(X = 1) = {}_3C_1 * \left(\frac{1}{2}\right)^1 * \left(\frac{1}{2}\right)^2 = \frac{3}{8}$$

$$P(X = 2) = {}_3C_2 * \left(\frac{1}{2}\right)^2 * \left(\frac{1}{2}\right)^1 = \frac{3}{8}$$

$$P(X = 3) = {}_3C_3 * \left(\frac{1}{2}\right)^3 * \left(\frac{1}{2}\right)^0 = \frac{1}{8}$$

Chapter 6 이산확률분포

◎ 이항분포의 확률계산

예제 6 -3 주사위를 10번 던졌을 때, 3이 두 번 나올 확률은 얼마인가를 알려고 한다면, 다음과 같이 계산이 가능하다. 주사위를 한 번 던질때 3이 나올 확률은 1/6이며, 3이 아닌 수가 나올 확률은 **5/6**이다. 그러므로 p=1/6, q=5/6, n=10이며, 3이 나오는 횟수를 X라 하면 3이 두 번 나올 확률, 즉 $\boldsymbol{P}(\mathrm{X}=2)$는 29%가 된다.

$$\boldsymbol{P}(\mathrm{X}=2) = {}_{10}c_2\ \boldsymbol{p}^2 q^{10-2}$$

$$= \frac{10!}{8!\,2!} * \left(\frac{1}{6}\right)^2 * \left(\frac{5}{6}\right)^8 = 0.29$$

◎ 이항분포표의 이용

이항확률함수를 이용하더라도 n이 커지고 p의 값에 소수점 이하의 숫자가 많아지면 계산이 매우 복잡해짐

이 때엔 **이미 계산되어 있는 표**를 이용하는 방법을 고려할 수 있는데, **'이항분포표'**라고 함
이항분포는 **n과 p에** 따라 그 모양이 달라지므로, 표를 이용할 때에는 반드시 **n과 p**를 알고 해당하는 확률값을 계산해야 함

Chapter 6 이산확률분포

◎ 이항분포표 (일부)

n	x	p = 0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
1	0	0.9500	0.9000	0.8500	0.8000	0.7500	0.7000	0.6500	0.6000	0.5500	0.5000
	1	0.0500	0.1000	0.1500	0.2000	0.2500	0.3000	0.3500	0.4000	0.4500	0.5000
2	0	0.9025	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500
	1	0.0950	0.1800	0.2550	0.3200	0.3750	0.4200	0.4550	0.4800	0.4950	0.5000
	2	0.0025	0.0100	0.0225	0.0400	0.0625	0.0900	0.1225	0.1600	0.2025	0.2500
3	0	0.8574	0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250
	1	0.1354	0.2430	0.3251	0.3840	0.4219	0.4410	0.4436	0.4320	0.4084	0.3750
	2	0.0071	0.0270	0.0574	0.0960	0.1406	0.1890	0.2389	0.2880	0.3341	0.3750
	3	0.0001	0.0010	0.0034	0.0080	0.0156	0.0270	0.0429	0.0640	0.0911	0.1250
4	0	0.8145	0.6561	0.5220	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625
	1	0.1715	0.2916	0.3685	0.4096	0.4219	0.4116	0.3845	0.3456	0.2995	0.2500
	2	0.0135	0.0486	0.0975	0.1536	0.2109	0.2646	0.3105	0.3456	0.3675	0.3750
	3	0.0005	0.0036	0.0115	0.0256	0.0469	0.0756	0.1115	0.1536	0.2005	0.2500
	4	0.0000	0.0001	0.0005	0.0016	0.0039	0.0081	0.0150	0.0256	0.0410	0.0625
5	0	0.7738	0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0313
	1	0.2036	0.3281	0.3915	0.4096	0.3955	0.3602	0.3124	0.2592	0.2059	0.1563
	2	0.0214	0.0729	0.1382	0.2048	0.2637	0.3087	0.3364	0.3456	0.3369	0.3125
	3	0.0011	0.0081	0.0244	0.0512	0.0879	0.1323	0.1811	0.2304	0.2757	0.3125
	4	0.0000	0.0005	0.0022	0.0064	0.0146	0.0284	0.0488	0.0768	0.1128	0.1563
	5	0.0000	0.0000	0.0001	0.0003	0.0010	0.0024	0.0053	0.0102	0.0185	0.0313
6	0	0.7351	0.5314	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156
	1	0.2321	0.3543	0.3993	0.3932	0.3560	0.3025	0.2437	0.1866	0.1359	0.0938
	2	0.0305	0.0984	0.1762	0.2458	0.2966	0.3241	0.3280	0.3110	0.2780	0.2344
	3	0.0021	0.0146	0.0415	0.0819	0.1318	0.1852	0.2355	0.2765	0.3032	0.3125
	4	0.0001	0.0012	0.0055	0.0154	0.0330	0.0595	0.0951	0.1382	0.1861	0.2344
	5	0.0000	0.0001	0.0004	0.0015	0.0044	0.0102	0.0205	0.0369	0.0609	0.0938
	6	0.0000	0.0000	0.0000	0.0001	0.0002	0.0007	0.0018	0.0041	0.0083	0.0156
7	0	0.6983	0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078
	1	0.2573	0.3720	0.3960	0.3670	0.3115	0.2471	0.1848	0.1306	0.0872	0.0547
	2	0.0406	0.1240	0.2097	0.2753	0.3115	0.3177	0.2985	0.2613	0.2140	0.1641
	3	0.0036	0.0230	0.0617	0.1147	0.1730	0.2269	0.2679	0.2903	0.2918	0.2734
	4	0.0002	0.0026	0.0109	0.0287	0.0577	0.0972	0.1442	0.1935	0.2388	0.2734
	5	0.0000	0.0002	0.0012	0.0043	0.0115	0.0250	0.0466	0.0774	0.1172	0.1641
	6	0.0000	0.0000	0.0001	0.0004	0.0013	0.0036	0.0084	0.0172	0.0320	0.0547
	7	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0006	0.0016	0.0037	0.0078
8	0	0.6634	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039
	1	0.2793	0.3826	0.3847	0.3355	0.2670	0.1977	0.1373	0.0896	0.0548	0.0313
	2	0.0515	0.1488	0.2376	0.2936	0.3115	0.2965	0.2587	0.2090	0.1569	0.1094
	3	0.0054	0.0331	0.0839	0.1468	0.2076	0.2541	0.2786	0.2787	0.2568	0.2188
	4	0.0004	0.0046	0.0185	0.0459	0.0865	0.1361	0.1875	0.2322	0.2627	0.2734
	5	0.0000	0.0004	0.0026	0.0092	0.0231	0.0467	0.0808	0.1239	0.1719	0.2188
	6	0.0000	0.0000	0.0002	0.0011	0.0038	0.0100	0.0217	0.0413	0.0703	0.1094
	7	0.0000	0.0000	0.0000	0.0001	0.0004	0.0012	0.0033	0.0079	0.0164	0.0313
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0007	0.0017	0.0039

Chapter 6 이산확률분포

◎ 이항분포표 (일부)

- 앞의 표는 이항분포표의 일부다 이항분포표에서 n은 시행횟수, x는 성공횟수를 의미한다.
- 예를 들어 동전을 3번 던져서 앞면이 2번 나올 확률은 $n = 3, p = 0.5, x = 2$이므로 이항분포표에서 0.3750이라는 것을 확인가능하다.

예제 6 -4 다섯 개의 4지선다 문항이 있다. 순전히 추측으로 세 개의 문제를 맞힐 확률을 얼마나 되는지 구하시오.

n	x	p = 0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
1	0	0.9500	0.9000	0.8500	0.8000	0.7500	0.7000	0.6500	0.6000	0.5500	0.5000
	1	0.0500	0.1000	0.1500	0.2000	0.2500	0.3000	0.3500	0.4000	0.4500	0.5000
2	0	0.9025	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500
	1	0.0950	0.1800	0.2550	0.3200	0.3750	0.4200	0.4550	0.4800	0.4950	0.5000
	2	0.0025	0.0100	0.0225	0.0400	0.0625	0.0900	0.1225	0.1600	0.2025	0.2500
3	0	0.8574	0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250
	1	0.1354	0.2430	0.3251	0.3840	0.4219	0.4410	0.4436	0.4320	0.4084	0.3750
	2	0.0071	0.0270	0.0574	0.0960	0.1406	0.1890	0.2389	0.2880	0.3341	0.3750
	3	0.0001	0.0010	0.0034	0.0080	0.0156	0.0270	0.0429	0.0640	0.0911	0.1250
4	0	0.8145	0.6561	0.5220	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625
	1	0.1715	0.2916	0.3685	0.4096	0.4219	0.4116	0.3845	0.3456	0.2995	0.2500
	2	0.0135	0.0486	0.0975	0.1536	0.2109	0.2646	0.3105	0.3456	0.3675	0.3750
	3	0.0005	0.0036	0.0115	0.0256	0.0469	0.0756	0.1115	0.1536	0.2005	0.2500
	4	0.0000	0.0001	0.0005	0.0016	0.0039	0.0081	0.0150	0.0256	0.0410	0.0625
5	0	0.7738	0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0313
	1	0.2036	0.3281	0.3915	0.4096	0.3955	0.3602	0.3124	0.2592	0.2059	0.1563
	2	0.0214	0.0729	0.1382	0.2048	0.2637	0.3087	0.3364	0.3456	0.3369	0.3125
	3	0.0011	0.0081	0.0244	0.0512	0.0879	0.1323	0.1811	0.2304	0.2757	0.3125
	4	0.0000	0.0005	0.0022	0.0064	0.0146	0.0284	0.0488	0.0768	0.1128	0.1563
	5	0.0000	0.0000	0.0001	0.0003	0.0010	0.0024	0.0053	0.0102	0.0185	0.0313
6	0	0.7351	0.5314	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156
	1	0.2321	0.3543	0.3993	0.3932	0.3560	0.3025	0.2437	0.1866	0.1359	0.0938
	2	0.0305	0.0984	0.1762	0.2458	0.2966	0.3241	0.3280	0.3110	0.2780	0.2344
	3	0.0021	0.0146	0.0415	0.0819	0.1318	0.1852	0.2355	0.2765	0.3032	0.3125
	4	0.0001	0.0012	0.0055	0.0154	0.0330	0.0595	0.0951	0.1382	0.1861	0.2344
	5	0.0000	0.0001	0.0004	0.0015	0.0044	0.0102	0.0205	0.0369	0.0609	0.0938
	6	0.0000	0.0000	0.0000	0.0001	0.0002	0.0007	0.0018	0.0041	0.0083	0.0156
7	0	0.6983	0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078
	1	0.2573	0.3720	0.3960	0.3670	0.3115	0.2471	0.1848	0.1306	0.0872	0.0547
	2	0.0406	0.1240	0.2097	0.2753	0.3115	0.3177	0.2985	0.2613	0.2140	0.1641
	3	0.0036	0.0230	0.0617	0.1147	0.1730	0.2269	0.2679	0.2903	0.2918	0.2734
	4	0.0002	0.0026	0.0109	0.0287	0.0577	0.0972	0.1442	0.1935	0.2388	0.2734
	5	0.0000	0.0002	0.0012	0.0043	0.0115	0.0250	0.0466	0.0774	0.1172	0.1641
	6	0.0000	0.0000	0.0001	0.0004	0.0013	0.0036	0.0084	0.0172	0.0320	0.0547
	7	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0006	0.0016	0.0037	0.0078
8	0	0.6634	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039
	1	0.2793	0.3826	0.3847	0.3355	0.2670	0.1977	0.1373	0.0896	0.0548	0.0313
	2	0.0515	0.1488	0.2376	0.2936	0.3115	0.2965	0.2587	0.2090	0.1569	0.1094
	3	0.0054	0.0331	0.0839	0.1468	0.2076	0.2541	0.2786	0.2787	0.2568	0.2188
	4	0.0004	0.0046	0.0185	0.0459	0.0865	0.1361	0.1875	0.2322	0.2627	0.2734
	5	0.0000	0.0004	0.0026	0.0092	0.0231	0.0467	0.0808	0.1239	0.1719	0.2188
	6	0.0000	0.0000	0.0002	0.0011	0.0038	0.0100	0.0217	0.0413	0.0703	0.1094
	7	0.0000	0.0000	0.0000	0.0001	0.0004	0.0012	0.0033	0.0079	0.0164	0.0313
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0007	0.0017	0.0039

$n = 5, p = 0.25, x = 3$

$\boldsymbol{P}(\mathrm{X} = 3) = 0.0879$

Chapter 6 이산확률분포

◎ 이항분포표 (일부)

예제 6 -5 학원에 올 때 비가 내릴 때 비를 맞을 확률이 40%라고 한다. 5명의 학생이 학원에 왔다면, 이들 중 2명이 비를 맞았을 확률은?

$$n = 5, p = 0.40, x = 2$$
$$\boldsymbol{P}(\mathrm{X} = 2) = 0.3456$$

n	x	p = 0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
1	0	0.9500	0.9000	0.8500	0.8000	0.7500	0.7000	0.6500	0.6000	0.5500	0.5000
	1	0.0500	0.1000	0.1500	0.2000	0.2500	0.3000	0.3500	0.4000	0.4500	0.5000
2	0	0.9025	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500
	1	0.0950	0.1800	0.2550	0.3200	0.3750	0.4200	0.4550	0.4800	0.4950	0.5000
	2	0.0025	0.0100	0.0225	0.0400	0.0625	0.0900	0.1225	0.1600	0.2025	0.2500
3	0	0.8574	0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250
	1	0.1354	0.2430	0.3251	0.3840	0.4219	0.4410	0.4436	0.4320	0.4084	0.3750
	2	0.0071	0.0270	0.0574	0.0960	0.1406	0.1890	0.2389	0.2880	0.3341	0.3750
	3	0.0001	0.0010	0.0034	0.0080	0.0156	0.0270	0.0429	0.0640	0.0911	0.1250
4	0	0.8145	0.6561	0.5220	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625
	1	0.1715	0.2916	0.3685	0.4096	0.4219	0.4116	0.3845	0.3456	0.2995	0.2500
	2	0.0135	0.0486	0.0975	0.1536	0.2109	0.2646	0.3105	0.3456	0.3675	0.3750
	3	0.0005	0.0036	0.0115	0.0256	0.0469	0.0756	0.1115	0.1536	0.2005	0.2500
	4	0.0000	0.0001	0.0005	0.0016	0.0039	0.0081	0.0150	0.0256	0.0410	0.0625
5	0	0.7738	0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0313
	1	0.2036	0.3281	0.3915	0.4096	0.3955	0.3602	0.3124	0.2592	0.2059	0.1563
	2	0.0214	0.0729	0.1382	0.2048	0.2637	0.3087	0.3364	0.3456	0.3369	0.3125
	3	0.0011	0.0081	0.0244	0.0512	0.0879	0.1323	0.1811	0.2304	0.2757	0.3125
	4	0.0000	0.0005	0.0022	0.0064	0.0146	0.0284	0.0488	0.0768	0.1128	0.1563
	5	0.0000	0.0000	0.0001	0.0003	0.0010	0.0024	0.0053	0.0102	0.0185	0.0313
6	0	0.7351	0.5314	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156
	1	0.2321	0.3543	0.3993	0.3932	0.3560	0.3025	0.2437	0.1866	0.1359	0.0938
	2	0.0305	0.0984	0.1762	0.2458	0.2966	0.3241	0.3280	0.3110	0.2780	0.2344
	3	0.0021	0.0146	0.0415	0.0819	0.1318	0.1852	0.2355	0.2765	0.3032	0.3125
	4	0.0001	0.0012	0.0055	0.0154	0.0330	0.0595	0.0951	0.1382	0.1861	0.2344
	5	0.0000	0.0001	0.0004	0.0015	0.0044	0.0102	0.0205	0.0369	0.0609	0.0938
	6	0.0000	0.0000	0.0000	0.0001	0.0002	0.0007	0.0018	0.0041	0.0083	0.0156
7	0	0.6983	0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078
	1	0.2573	0.3720	0.3960	0.3670	0.3115	0.2471	0.1848	0.1306	0.0872	0.0547
	2	0.0406	0.1240	0.2097	0.2753	0.3115	0.3177	0.2985	0.2613	0.2140	0.1641
	3	0.0036	0.0230	0.0617	0.1147	0.1730	0.2269	0.2679	0.2903	0.2918	0.2734
	4	0.0002	0.0026	0.0109	0.0287	0.0577	0.0972	0.1442	0.1935	0.2388	0.2734
	5	0.0000	0.0002	0.0012	0.0043	0.0115	0.0250	0.0466	0.0774	0.1172	0.1641
	6	0.0000	0.0000	0.0001	0.0004	0.0013	0.0036	0.0084	0.0172	0.0320	0.0547
	7	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0006	0.0016	0.0037	0.0078
8	0	0.6634	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039
	1	0.2793	0.3826	0.3847	0.3355	0.2670	0.1977	0.1373	0.0896	0.0548	0.0313
	2	0.0515	0.1488	0.2376	0.2936	0.3115	0.2965	0.2587	0.2090	0.1569	0.1094
	3	0.0054	0.0331	0.0839	0.1468	0.2076	0.2541	0.2786	0.2787	0.2568	0.2188
	4	0.0004	0.0046	0.0185	0.0459	0.0865	0.1361	0.1875	0.2322	0.2627	0.2734
	5	0.0000	0.0004	0.0026	0.0092	0.0231	0.0467	0.0808	0.1239	0.1719	0.2188
	6	0.0000	0.0000	0.0002	0.0011	0.0038	0.0100	0.0217	0.0413	0.0703	0.1094
	7	0.0000	0.0000	0.0000	0.0001	0.0004	0.0012	0.0033	0.0079	0.0164	0.0313
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0007	0.0017	0.0039

Chapter 6 이산확률분포

◎ 이항분포표 (일부)

예제 6 -6 동전을 네 번 던졌을 때 앞면이 세 번 이하로 나올 확률은?

$n = 4, p = 0.5, x \leq 3$

풀이 1 :: $\boldsymbol{P}(\mathrm{X} \leq 3) = \boldsymbol{P}(\mathrm{X} = 0) + \boldsymbol{P}(\mathrm{X} = 1) + \boldsymbol{P}(\mathrm{X} = 2) + \boldsymbol{P}(\mathrm{X} = 3)$

$= 0.0625 + 0.2500 + 0.3750 + 0.2500 = 0.9375$

풀이 2 :: $\boldsymbol{P}(\mathrm{X} \leq 3) = 1 - 0.0625 = \mathbf{0.9375}$

n	x	p = 0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
1	0	0.9500	0.9000	0.8500	0.8000	0.7500	0.7000	0.6500	0.6000	0.5500	0.5000
	1	0.0500	0.1000	0.1500	0.2000	0.2500	0.3000	0.3500	0.4000	0.4500	0.5000
2	0	0.9025	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500
	1	0.0950	0.1800	0.2550	0.3200	0.3750	0.4200	0.4550	0.4800	0.4950	0.5000
	2	0.0025	0.0100	0.0225	0.0400	0.0625	0.0900	0.1225	0.1600	0.2025	0.2500
3	0	0.8574	0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250
	1	0.1354	0.2430	0.3251	0.3840	0.4219	0.4410	0.4436	0.4320	0.4084	0.3750
	2	0.0071	0.0270	0.0574	0.0960	0.1406	0.1890	0.2389	0.2880	0.3341	0.3750
	3	0.0001	0.0010	0.0034	0.0080	0.0156	0.0270	0.0429	0.0640	0.0911	0.1250
4	0	0.8145	0.6561	0.5220	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625
	1	0.1715	0.2916	0.3685	0.4096	0.4219	0.4116	0.3845	0.3456	0.2995	0.2500
	2	0.0135	0.0486	0.0975	0.1536	0.2109	0.2646	0.3105	0.3456	0.3675	0.3750
	3	0.0005	0.0036	0.0115	0.0256	0.0469	0.0756	0.1115	0.1536	0.2005	0.2500
	4	0.0000	0.0001	0.0005	0.0016	0.0039	0.0081	0.0150	0.0256	0.0410	0.0625
5	0	0.7738	0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0313
	1	0.2036	0.3281	0.3915	0.4096	0.3955	0.3602	0.3124	0.2592	0.2059	0.1563
	2	0.0214	0.0729	0.1382	0.2048	0.2637	0.3087	0.3364	0.3456	0.3369	0.3125
	3	0.0011	0.0081	0.0244	0.0512	0.0879	0.1323	0.1811	0.2304	0.2757	0.3125
	4	0.0000	0.0005	0.0022	0.0064	0.0146	0.0284	0.0488	0.0768	0.1128	0.1563
	5	0.0000	0.0000	0.0001	0.0003	0.0010	0.0024	0.0053	0.0102	0.0185	0.0313
6	0	0.7351	0.5314	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156
	1	0.2321	0.3543	0.3993	0.3932	0.3560	0.3025	0.2437	0.1866	0.1359	0.0938
	2	0.0305	0.0984	0.1762	0.2458	0.2966	0.3241	0.3280	0.3110	0.2780	0.2344
	3	0.0021	0.0146	0.0415	0.0819	0.1318	0.1852	0.2355	0.2765	0.3032	0.3125
	4	0.0001	0.0012	0.0055	0.0154	0.0330	0.0595	0.0951	0.1382	0.1861	0.2344
	5	0.0000	0.0001	0.0004	0.0015	0.0044	0.0102	0.0205	0.0369	0.0609	0.0938
	6	0.0000	0.0000	0.0000	0.0001	0.0002	0.0007	0.0018	0.0041	0.0083	0.0156
7	0	0.6983	0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078
	1	0.2573	0.3720	0.3960	0.3670	0.3115	0.2471	0.1848	0.1306	0.0872	0.0547
	2	0.0406	0.1240	0.2097	0.2753	0.3115	0.3177	0.2985	0.2613	0.2140	0.1641
	3	0.0036	0.0230	0.0617	0.1147	0.1730	0.2269	0.2679	0.2903	0.2918	0.2734
	4	0.0002	0.0026	0.0109	0.0287	0.0577	0.0972	0.1442	0.1935	0.2388	0.2734
	5	0.0000	0.0002	0.0012	0.0043	0.0115	0.0250	0.0466	0.0774	0.1172	0.1641
	6	0.0000	0.0000	0.0001	0.0004	0.0013	0.0036	0.0084	0.0172	0.0320	0.0547
	7	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0006	0.0016	0.0037	0.0078
8	0	0.6634	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039
	1	0.2793	0.3826	0.3847	0.3355	0.2670	0.1977	0.1373	0.0896	0.0548	0.0313
	2	0.0515	0.1488	0.2376	0.2936	0.3115	0.2965	0.2587	0.2090	0.1569	0.1094
	3	0.0054	0.0331	0.0839	0.1468	0.2076	0.2541	0.2786	0.2787	0.2568	0.2188
	4	0.0004	0.0046	0.0185	0.0459	0.0865	0.1361	0.1875	0.2322	0.2627	0.2734
	5	0.0000	0.0004	0.0026	0.0092	0.0231	0.0467	0.0808	0.1239	0.1719	0.2188
	6	0.0000	0.0000	0.0002	0.0011	0.0038	0.0100	0.0217	0.0413	0.0703	0.1094
	7	0.0000	0.0000	0.0000	0.0001	0.0004	0.0012	0.0033	0.0079	0.0164	0.0313
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0007	0.0017	0.0039

Chapter 6 이산확률분포

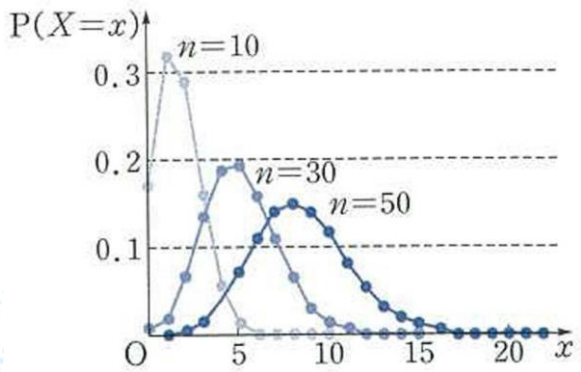
◎ 이항분포표 (변형)

- 성공확률 p가 0.5 이상의 경우는 식을 변형

예제 6 -7 예를 들어 어느 상품이 불량품일 가능성이 0.9라면 5개 골라서 이 중에 불량품이 3개일 확률이 얼마인지를 알아보는 질문에서 p= 0.9이므로 식은 아래와 같이 표현이 가능하다.

$$\boldsymbol{P}(\mathrm{X}=3) = {}_5C_3\ (0.9)^3(0.1)^2 = \boldsymbol{P}(\mathrm{X}=2) = {}_5C_2\ (0.1)^2(0.9)^3$$

${}_5C_3$ = ${}_5C_2$ 이므로 두 식은 결과적으로 동일한 식이 되며, 그 결과도 **0.0729**로 일치한다.



이항분포의 모양

① p=0.5일 때에는, 이항실험횟수 n이 작더라도 확률분포는 언제나 대칭을 이룬다

② p=0.5가 아니라고 할 지라도 이항실험횟수 n이 커짐에 따라 확률분포는 대칭에 가까워진다

Chapter 6 이산확률분포

◎ 이항분포의 기대값과 분산

- 앞선 예제에서 이항분포의 확률변수와 확률값을 대입하여 이항분포의 기대값과 분산을 구할 수 있음
- 동전을 두 번 던질 때의 확률분포는 아래와 같이 나타낼 수 있음

X_i	$P(X_i)$	$X_i * P(X_i)$	$X_i - \mu$	$(X_i - \mu)^2$	$P(X_i) * (X_i - \mu)^2$
0	1/4	0	**0-1 = -1**	**1**	1/4
1	2/4	2/4	**1-1 = 0**	**0**	0
2	1/4	2/4	**2-1 = 1**	**1**	1/4

$$\mu = E(x) = \sum X_i * P(X_i) = 0 * \frac{1}{4} + 1 * \frac{2}{4} + 2 * \frac{1}{4} = 1$$

$$\sigma^2 = \sum P(X_i) * (X_i - \mu)^2 = \frac{1}{4} + 0 + \frac{1}{4} = \frac{1}{2}$$

Chapter 6 이산확률분포

◎ 이항분포의 기댓값과 분산

- 하지만 앞에서와 같이 확률분포를 이용하여 기댓값과 분산을 일일이 계산하는 것은 번거로움
- 이에 따라 공식을 활용

이항분포의 기댓값과 분산, 표준편차

기 댓 값 $\mu = E(x) = np$

분 산 $\sigma^2 = \mathrm{Var}(x) = np(1-\mathrm{p}) = npq$

표준편차 $\sigma = \sqrt{np(1-\mathrm{p})} = \sqrt{npq}$

예제 6 -8

한 번의 시행에서 A사건이 일어날 확률은 1/5이다.
100번의 독립시행에서 A사건이 나타날
횟수의 기댓값과 분산을 구하시오.

풀이 :: $\mu = E(x) = np = 100 * \frac{1}{5} = 20$(회)

$\sigma^2 = npq = 100 * \frac{1}{5} * \frac{4}{5} = 16$

예제 6 -9

A공장의 제품은 70%만이 정상품이고, 30%는 불량품이라 한다. 품질관리자가 임의 추출방법을 활용해 임의로 100개를 선택했을 때 몇 개가 정상 제품일지 기댓값과 분산을 구하시오.

풀이 :: $\mu = np = 100 * 0.7 = 70$ (개)

$\sigma^2 = npq = 100 * 0.7 * 0.3 = 21$

Chapter 6 이산확률분포

◎ 다항분포

- 실제 우리 사례의 경우 이항분포처럼 Binary 적 문제보다,다항분포적 문제 Categorical issue가 더 많다
  만약, 사람이 배를 탈 때 어느 선실등급에 들어갈 확률, (1등급에 들어갈 확률 0.2, 2등급에 들어갈 확률 0.3, 3등급에 들어갈 확률은 0.5라고 할 수 있다면 이는 다항분포적 문제라 할 수 있음)

- **다항분포**는 **확률실험의 결과**로 k**개의 가능한 경우**가 발생할 때 나타나는 분포

다항확률함수

$$P(\boldsymbol{X} = X_1, X_2, \ldots, X_k) = \frac{\boldsymbol{N}!}{\boldsymbol{n_1}!\boldsymbol{n_2}! \ldots \boldsymbol{n_k}!} p_1^{\boldsymbol{n_1}} p_2^{\boldsymbol{n_2}} \ldots p_k^{\boldsymbol{n_k}}$$

- 위의 식에서 $\boldsymbol{k} = 2$인 경우에는 $p_2 = 1 - p_1$ 이고 $\boldsymbol{n_2} = N - \boldsymbol{n_1}$ 이 되며, 이항분포 확률식과 같게 됨

Chapter 6 이산확률분포

◎ 다항분포

예제 6 -10 한 상자에 색깔이 서로 다른 구슬이 있는데, 구슬들은 아래와 같이 분포되어 있음

구슬색깔	확률
빨강	0.40
파랑	0.30
노랑	0.20
주황	0.10

이 상자에서 복원추출 방법으로 10개의 구슬을 추출할 때, 빨강 구슬이 3개, 파랑 구슬 4개, 노랑 구슬 3개, 주황 구슬은 하나도 뽑히지 않을 확률은 얼마일까? (단, 0! = 1이고, $(0.10)^0$ =1로 계산된다)

$$P(\text{빨강 } 3\text{개, 파랑 } 4\text{개, 노랑 } 3\text{개, 주황 } 0\text{개}) = \frac{10!}{3!4!3!0!} * (0.40)^3 * (0.30)^4 * (0.20)^3 * (0.10)^0 = 0.017$$

위와 같이 구슬들이 뽑힐 확률은 0.017이다

Chapter 6 이산확률분포

◎ 다항분포

예제 6 -11 우리나라 주요 일간지 4개 신문사의 구독률이 아래와 같다면, 10명의 구독자들을 임의로 추출하여 그들의 구독 신문을 알아볼때, A와 B신문을 각각 3명이 구독하고 C와 D 신문을 각각 2명이 구독할 확률은 ??

신문사	구독률
A신문사	0.30
B신문사	0.25
C신문사	0.20
D신문사	0.15
기 타	0.10

$$P(A\text{신문 } 3\text{명}, B\text{신문 } 3\text{명}, C\text{신문 } 2\text{명}, D\text{신문 } 2\text{명}, \text{기타 } 0\text{명},)$$

$$= \frac{10!}{3!3!2!2!0!} * (0.30)^3 * (0.25)^3 * (0.20)^2 * (0.15)^2 * (0.10)^0 = 0.019$$

Chapter 6 이산확률분포

◎ 이항분포와 초기하분포

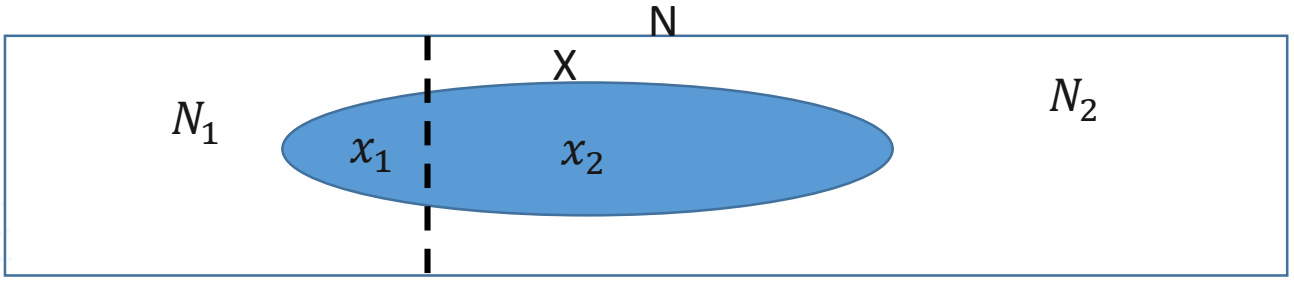
- 이항분포 : 가장 중요한 가정 중의 하나는 매 시행마다 어떤 사건이 일어날 가능성, 즉 성공의 확률은 일정
- **비복원추출**인 경우에는 **앞서의 시행결과**에 따른 **다음의 시행결과**가 달라지므로 두 시행은 서로 **종속적**

※ 베르누이시행의 조건 중 성공확률이 일정하다는 조건이 만족하지 않음 → 초기하분포(hypergeometric distribution)

:: 매 시행이 독립적 → 이항분포,
매 시행이 종속적 → 초기하분포

◎ 초기하분포의 확률변수

- 일반적인 초기하분포의 확률계산방법을 알아보기 위한 N명 중 X명을 뽑는 경우를 살펴보자.
  이 N명을 성별로 구분할 경우, N_1 명이 남자와 N_2 명의 여자로 구분된다.
  X 또한 N_1에 속하는 x_1과 N_2에 속하는 x_2로 구분된다.



Chapter 6 이산확률분포

◎ 초기하분포의 확률변수

초기하분포의 확률함수

$$P(N_1 \text{ 중 } x_1,\ N_2 \text{ 중 } x_2) = \frac{{}_{N_1}C_{x_1} * {}_{N_2}C_{x_2}}{{}_{(N_1+N_2)}C_{(x_1+x_2)}}$$

예제 6 -12 어느 모임에 다섯 사람이 참석하였는데, 이 중 여성이 2명이었다. 임의 추출로 두 사람을 뽑는다고 할 때, 여자가 한 명 뽑힐 확률은 얼마인가?

풀이 :: X를 여자의 수, N = 5, $N_1 = 2$, $N_2 = 3$인 초기하분포를 따르게 된다.

$$P(X=1) = \frac{{}_2C_1 * {}_3C_1}{{}_5C_2} = 0.6$$

예제 6 -13 ys 컴퍼니에서 생산하는 제품 20개 중에 5개의 불량품이 있다고 가정하자. 이 중 4개를 선택했을 때 2개가 불량품일 확률을 구하시오.

풀이 :: $P(X=2) = \frac{{}_5C_2 * {}_{15}C_2}{{}_{20}C_4} \approx 0.217$

Exercise Chapter 6

(1) 다음 용어들을 간단히 정의해보자.
①이항분포 ②다항분포 ③초기하분포

(2) 주사위를 다섯 번 던졌을 때 점의 수 3이 다음과 같은 횟수만큼 나올 확률은?
①한 번도 나오지 않을 확률
②한 번 나올 확률
③세 번 나올 확률

(3) 입학시험에 응시한 학생들의 합격가능성이 40%로 일정하다면, 임의추출로 5명의 학생을 고른다면, 이들 중 적어도 한명이 합격할 확률은?? (복원추출을 가정)

(4) 어느 대학에서 빅데이터 교수를 채용하려고 한다. 원서를 제출한 사람이 남자가 5명, 여자가 3명이었다면, 이들 중 무작위로 3명을 뽑을 때, 2명이 여자일 확률은?

Exercise Chapter 6

(2)번 문제 풀이

: : ①한 번도 나오지 않을 확률

$$P(X=0) = {}_5C_0 * \left(\frac{1}{6}\right)^0 * \left(\frac{5}{6}\right)^5 = 1 * 1 * \left(\frac{5}{6}\right)^5 = \frac{3,125}{7,776}$$

②한 번 나올 확률

$$P(X=1) = {}_5C_1 * \left(\frac{1}{6}\right)^1 * \left(\frac{5}{6}\right)^4 = 1 * \left(\frac{5}{6}\right)^1 * \left(\frac{5}{6}\right)^4 = \frac{3,125}{7,776}$$

③세 번 나올 확률

$$P(X=3) = {}_5C_3 * \left(\frac{1}{6}\right)^3 * \left(\frac{5}{6}\right)^2 = 10 * \left(\frac{1}{216}\right) * \left(\frac{25}{36}\right) = \frac{125}{3,888}$$

Exercise Chapter 6

(3)번 문제 풀이

:: $p = 0.4$, $q = 0.6$, $n = 5$

전체 중에서 한 명도 붙지 못할 확률을 빼면 됨

$$P(X \geq 1) = 1 - P(X = 0) = 1 - {}_5C_0 * (0.4)^0 * (0.6)^5 = \mathbf{0.9222}$$

(4)번 문제 풀이

::

$$\left(\frac{{}_5C_1 * {}_3C_2}{{}_8C_3} \right) = \left(\frac{5 * 3}{56} \right) = \left(\frac{15}{56} \right)$$

Chapter 7. 연속확률분포

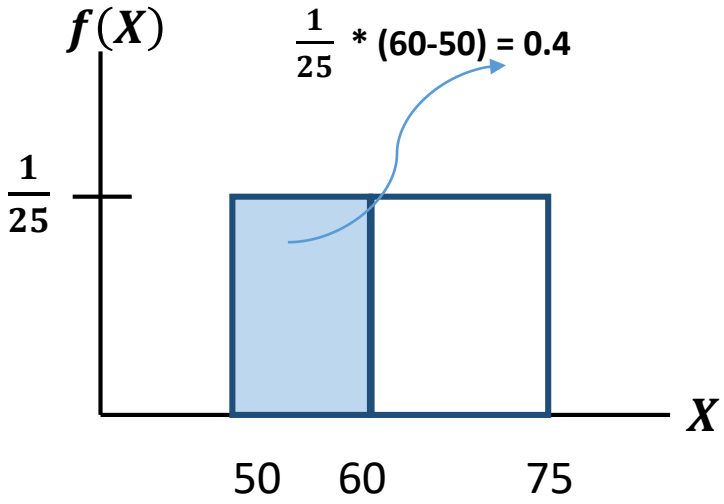
Daniel Rex Park
(박영식)

균일분포의 확률밀도함수

$$f(X) = \frac{1}{b-a}, \quad a \le X \le b$$

균일분포(uniform distribution) : **확률변수가 취하는 모든 구간**에서 각 **사건의 발생확률**이 **일정**한 것

균일분포는 **이산확률분포**와 **연속확률분포** 모두에서 통용되는 말
그러나 이 번 chapter에서는 **연속확률분포에서의 균일분포만** 설명



Chapter 7 연속확률분포

정규분포

- **정규분포**는 가우스분포(Gaussian distribution)이라고도 하며 연속확률분포 중 가장 널리 이용되는 중요한 분포
- 표본을 통한 통계적 추정 및 가설검정이론의 기본이 됨
- 사회과학현상에서 접하는 **여러 자료들의 분포**도 **정규분포**를 띠게 됨 (끄뜨레 :: 1796~1874)

정규분포의 확률밀도함수

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} * e^{-(x-\mu)^2 / 2\sigma^2}, \quad -\infty \leq X \leq +\infty$$

π 3.1416(원주율 : 상수)

e 2.7183(자연대수 : 상수)

μ 분포의 평균

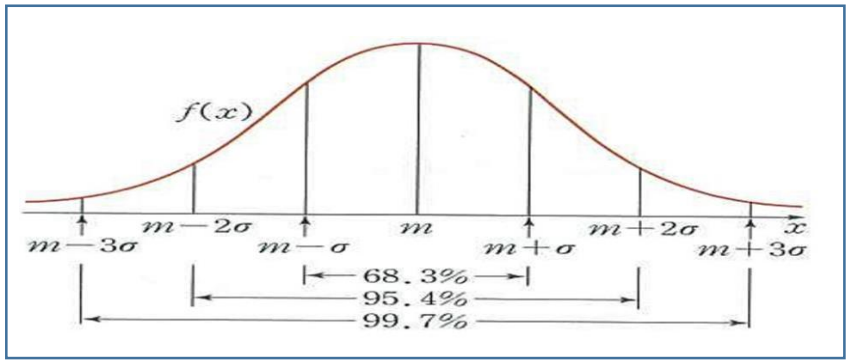
σ 분포의 표준편차

※ **정규분포**의 **모양**과 **위치**는 **분포의 표준편차**와 **평균** 두 요인으로 결정

정규분포공식에서 $\bar{X}$ 와 s 대신 μ와 σ 를 사용한 이유는 모집단의 기초를 통한 이론적 모형을 설명하기 위함

Chapter 7 연속확률분포

정규분포



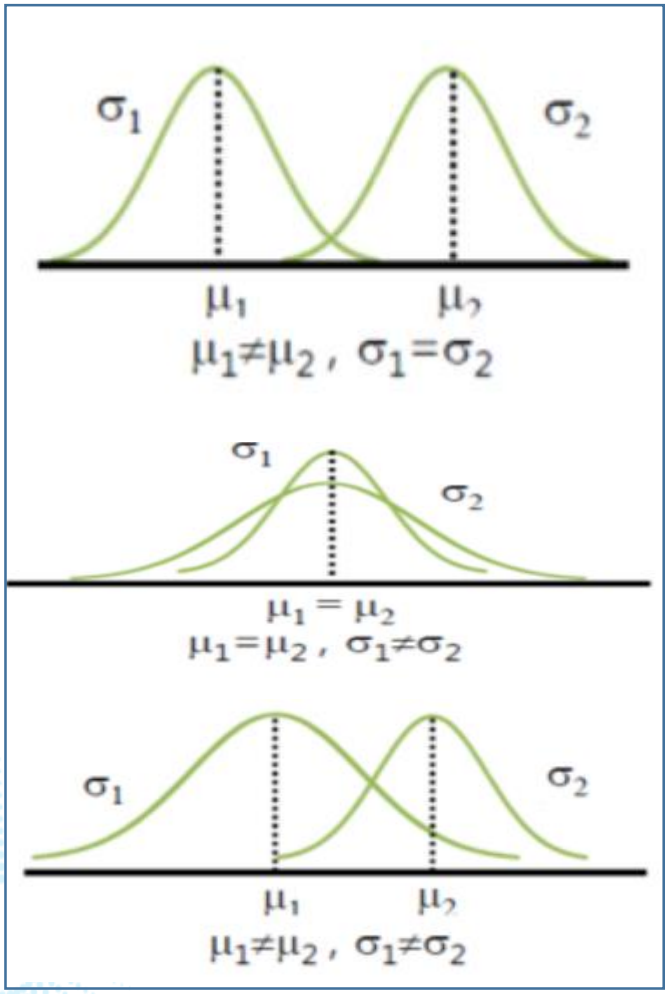
정규분포의 확률밀도함수

① 정규분포의 모양과 위치는 분포의 평균과 표준편차로 결정된다.
② 정규분포의 확률밀도함수는 평균(μ)을 중심으로 종모양(bell shape)이다.
③ 정규곡선은 X축에 맞닿지 않으므로 확률변수 X가 취할수 있는 값의 범위는 $-\infty \leq X \leq +\infty$ 이다
④ 분포의 평균(μ)과 표준편차(σ)가 어떤 값을 갖더라도, 정규곡선과 X축 사이의 전체 면적은 1이다

※ **정규분포**에서 관찰값의 99.7%가 $\pm 3\sigma$ 안에 속해 있음

Chapter 7 연속확률분포

◎ μ와 σ에 따른 정규분포 모양



1) 첫 번째 정규분포는 두 분포의 표준편차는 같고 평균이 다른 경우

2) 두 번째 정규분포는 표준편차는 다르고 평균이 같은 경우

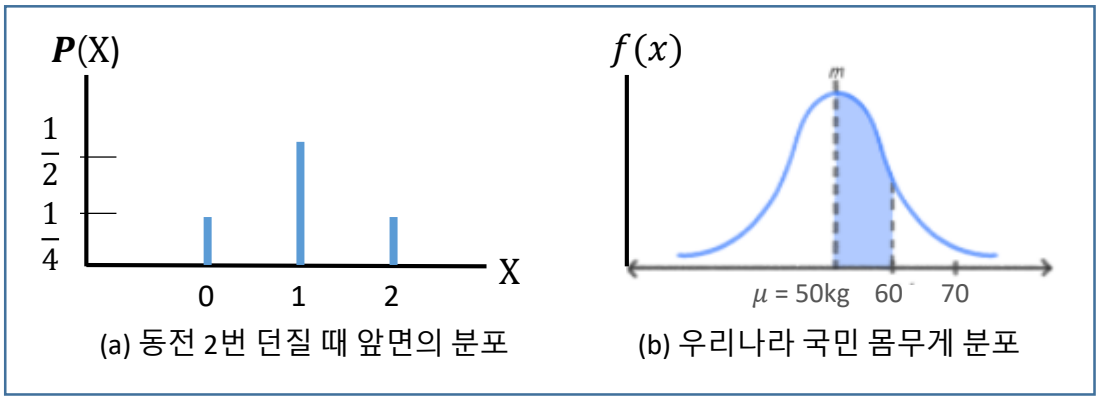
3) 세 번째 정규분포는 표준편차와 평균이 모두 다른 경우

Chapter 7 연속확률분포

◎ 정규분포에서의 확률

이산확률분포에서는 확률변수에 대응되는 확률이 존재하지만,
연속확률분포에서는 확률을 측정하는데 큰 차이가 존재

Ex) 우리나라 모든 국민의 **몸무게 분포**를 **조사**했을 때, 하나의 **고정된 값**으로 **몸무게(kg)을 측정하기란 불가능**함 하지만 **확률변수의 일정 구간**에 대한 **확률을 면적으로 구하는 것**은 가능



(a) 동전 2번 던질 때 앞면의 분포

(b) 우리나라 국민 몸무게 분포

※ **파란부분의 면적이 필요할 때마다 확률밀도함수를 계산하는 번거로움 ~> 표준정규분포의 필요성**

Chapter 7 연속확률분포

◎ 표준정규분포

정규분포는 평균과 표준편차에 따라 모양과 위치가 각기 다르므로 두 개 이상의 분포의 성격을 비교하거나 특정 정규분포에서 확률을 계산하기 위해서는, 먼저 모든 정규분포의 평균과 표준편차를 표준화하여 표준적인 정규분포를 만들어야 한다.

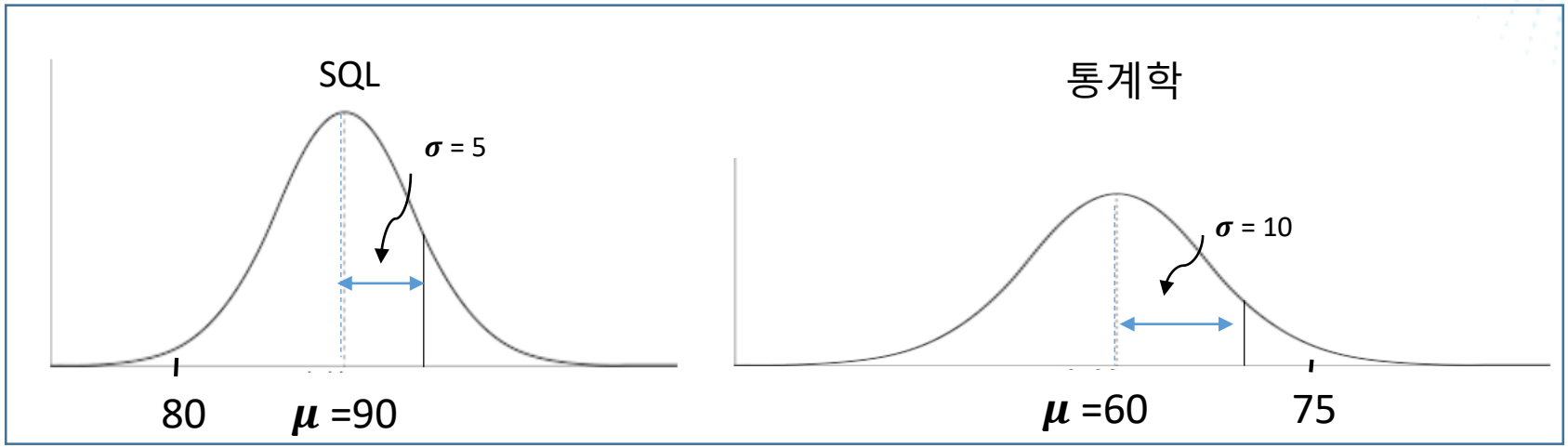
예제 7-1

어느 빅데이터반 학생이 시험을 치렀다. SQL시험은 80점이고, 통계학 시험은 75점이었다. 이 학생은 과연 어느 과목을 더 잘했다고 할 수 있는가?

A. 점수만 보면 **SQL시험**이라고 할 수 있겠지만, 해당 과목의 평균과 표준편차를 살펴본다면, **SQL 전체 학급의 평균**은 **90점**, **표준편차**는 **5점** 그리고 **통계학 시험**에서의 전체 학급의 **평균은 60점**, **표준편차가 10점**이라고 한다면, 학생들은 정규분포상의 위치가 다음 페이지와 같이 나타날 것이다.

Chapter 7 연속확률분포

◎ 표준정규분포



여러 확률분포의 경우 평균과 표준편차가 다를 수 있으므로 이를 표준정규분포로 변환하는 작업을 표준화(Standardization)라고 한다.

표준정규분포

- 모든 정규분포의 평균 $\mu = 0$이 되고, 표준편차 $\sigma = 1$이 됨
- 어떤 확률변수 X의 관찰값이 그 분포의 평균으로부터 표준편차의 몇 배 정도나 떨어져 있는가를 확률 변수 Z로 나타내기에 표준정규분포를 Z-분포라고도 함

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

◎ 표준정규분포

앞의 사례를 Z-scoring을 한다면 아래와 같이 변환이 가능

$$\text{SQL}\quad z = \frac{X-\mu}{\sigma} = \frac{80-90}{5} = -2$$

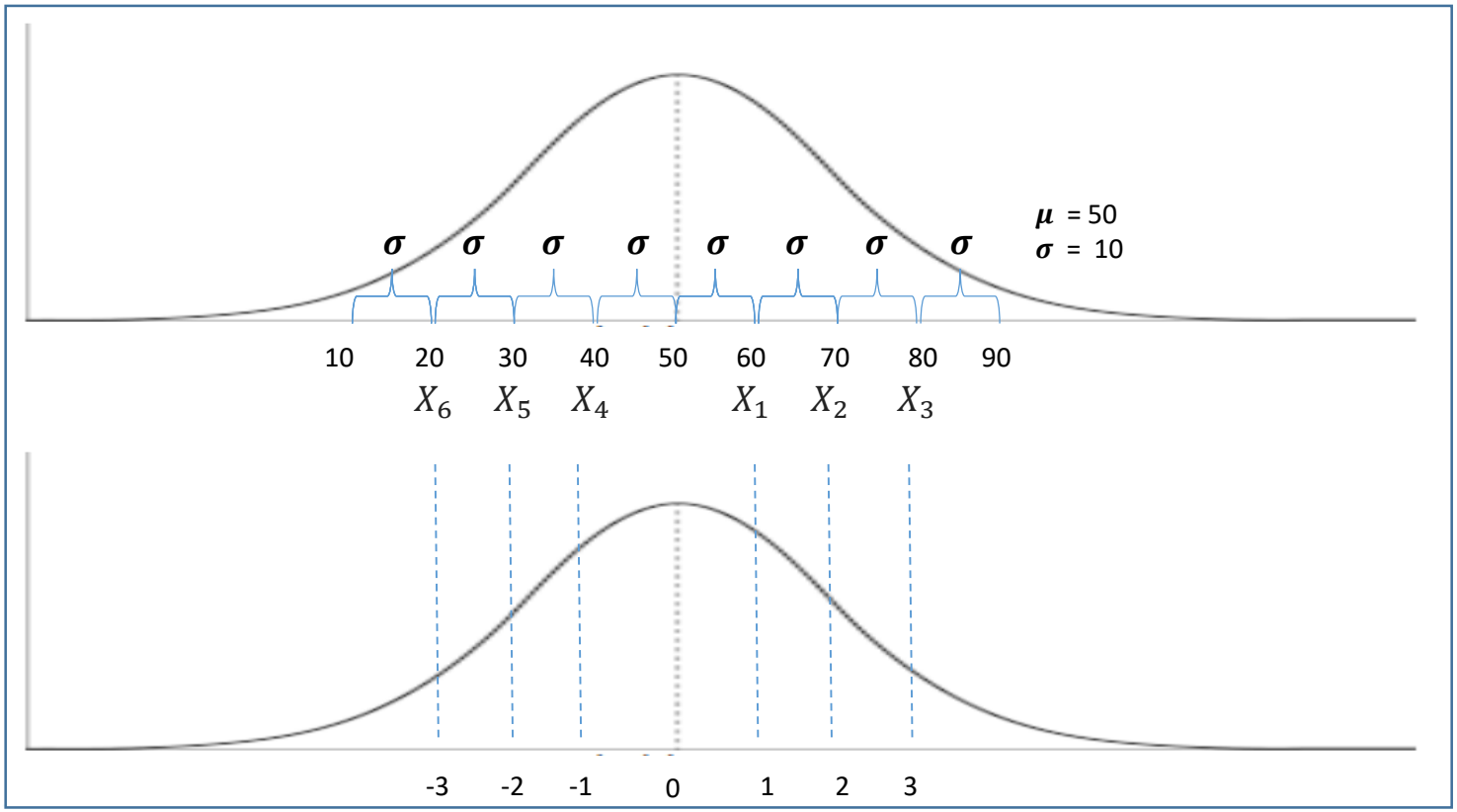
$$\text{statistics}\quad z = \frac{X-\mu}{\sigma} = \frac{75-60}{10} = 1.5$$

- z가 양수 (+)일 때는 과목평균을 중심으로 개인의 과목점수(X)가 오른쪽에 있고, z가 음수 (-)일 때는 과목점수(X)가 왼쪽에 있음을 의미

∴ SQL이 처음엔 더 높은 점수인 듯 하나, 분포상에서는 통계학이 평균보다 더 높은 점수 이므로 이 학생은 다른 학생들에 비해 **통계학**을 **SQL**보다 **더 잘한다**고 결론지을 수 있다.

Chapter 7 연속확률분포

◎ 표준정규분포



※ z값은 추리통계학에서 많이 사용되기에 표준정규분포의 이용법을 익혀두어야 함

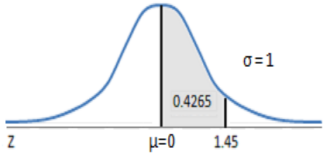
Chapter 7 연속확률분포

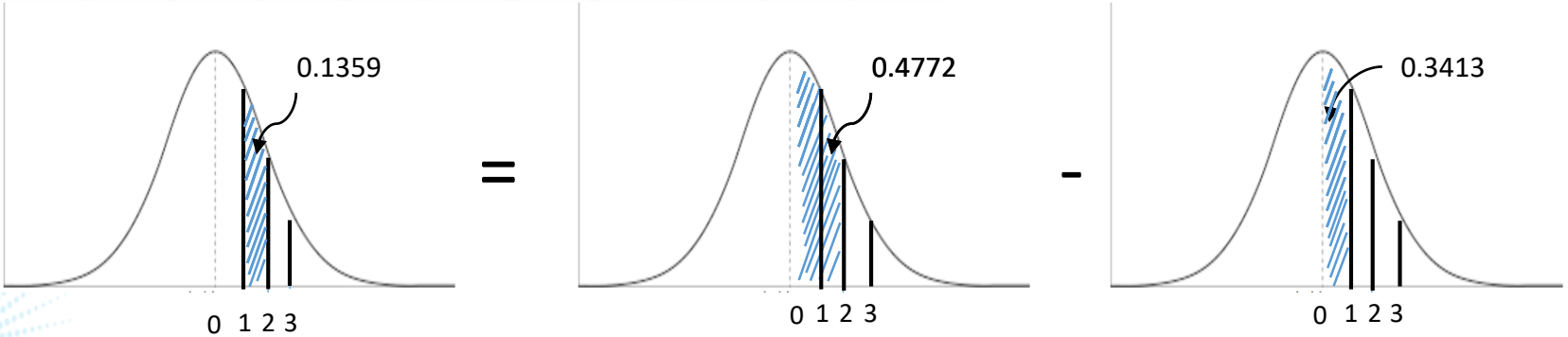
◎ 표준정규분포

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974

Areas Under the One-Tailed Standard Normal Curve

This table provides the area between the mean and some Z score.
For example, when Z score = 1.45 the area = 0.4265.



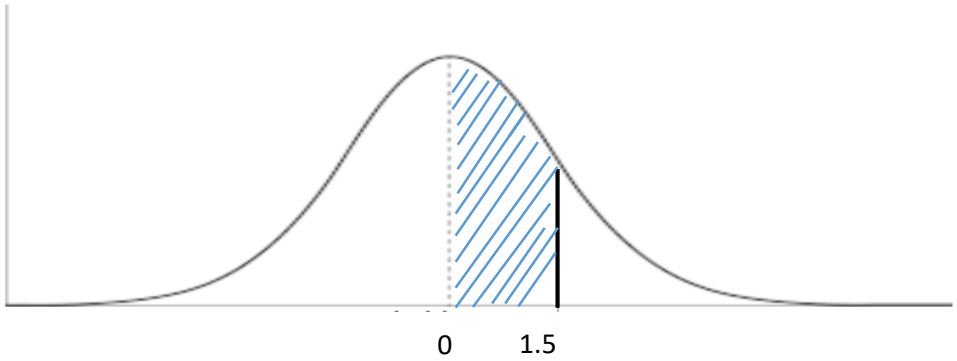


Chapter 7 연속확률분포

◎ 표준정규분포

① Z = 0부터 Z = 1.5 사이에 있을 확률

$P(0 \leq Z \leq 1.5) = 0.4332$



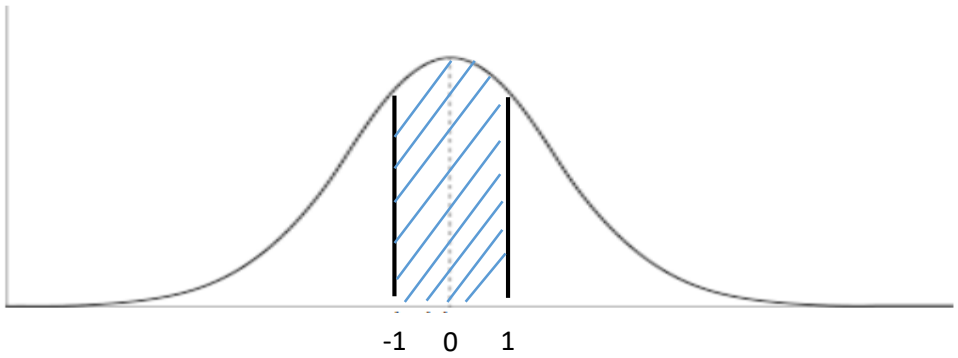
② Z = -1에서 Z = 1 사이에 있을 확률

$P(-1 \leq Z \leq 1)$

$= P(-1 \leq Z \leq 0) + P(0 \leq Z \leq 1)$

$= 2 * P(0 \leq Z \leq 1)$

$= 2 * 0.3413 = 0.6826$

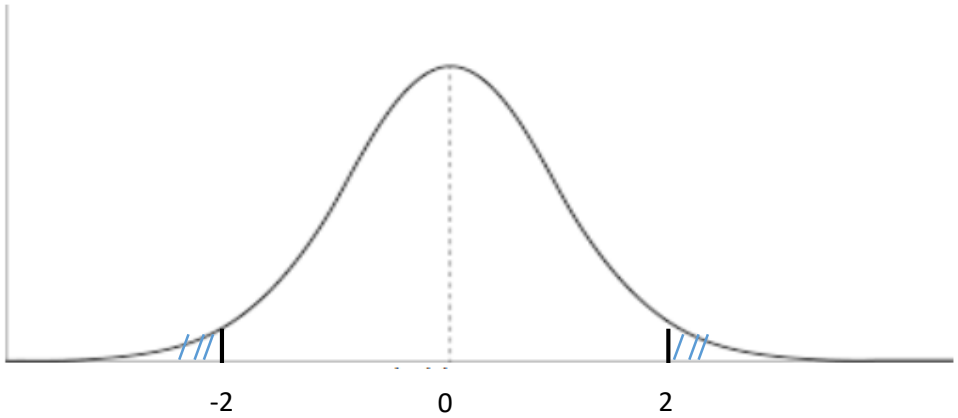


Chapter 7 연속확률분포

◎ 표준정규분포

③ Z = -2보다 작거나 Z = 2보다 큰 사이에 있을 확률

$P(Z \leq -2) + P(Z \geq +2)$
$= 1 - P(-2 \leq Z \leq +2)$
$= 1 - 2 * P(0 \leq Z \leq +2)$
$= 1 - 2 * 0.4772 = 0.0456$

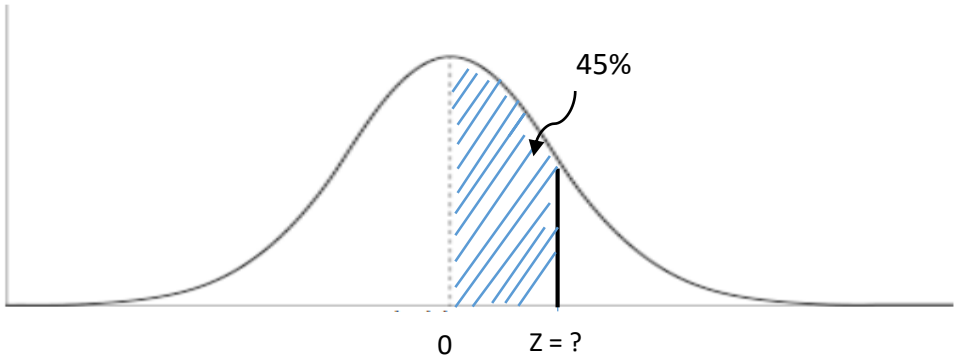


- 면적을 알고 있을 때 정규분포표를 활용하여 이에 대응하는 Z값을 찾는 문제

① 평균(Z = 0)에서 오른쪽으로 45%에 해당하는 Z값은?

A. 표에서 0.45에 가장 가까운 값을 찾으면 된다

Z = 1.64



Chapter 7 연속확률분포

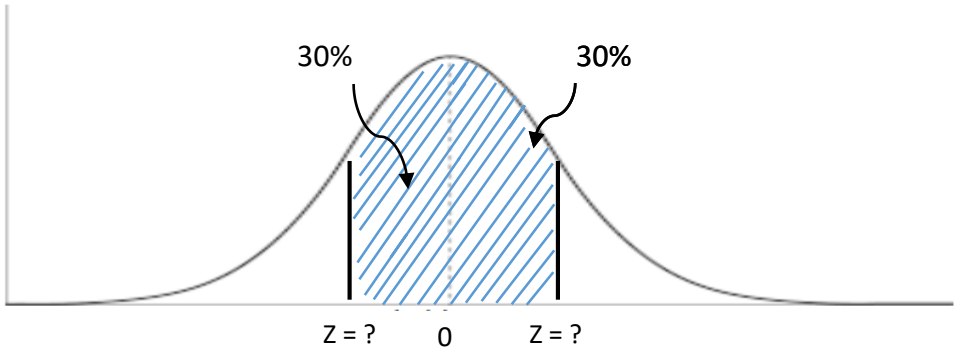
◎ 표준정규분포

- 면적을 알고 있을 때 정규분포표를 활용하여 이에 대응하는 z값을 찾는 문제

② 평균에서 각각 30%에 해당하는 z의 양쪽값은?

A. 표에서 0.30에 가장 가까운 값을 찾으면 된다

Z = ±0.84



③ 평균을 중심으로 양쪽을 합하여 90%, 95%, 99%가 되는 z값은 각각 얼마인가?

A. Z= 0에서부터 45%, 47.5%, 49.5%에 해당하는 z값을 구하면 됨

90% : Z = ± 1.64

95% : Z = ± 1.96

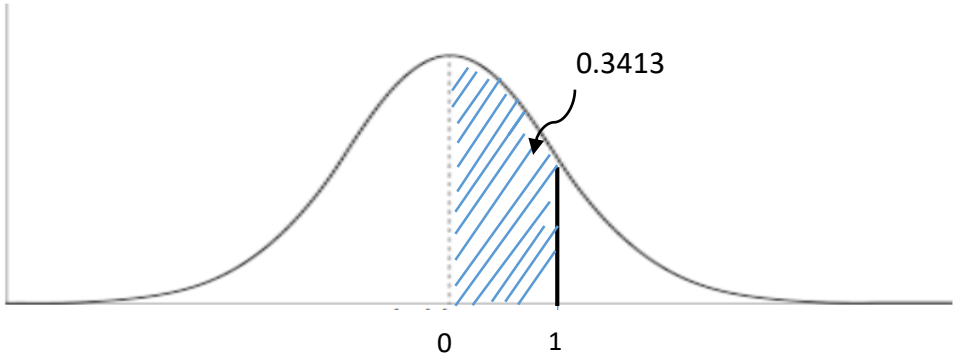
99% : Z = ± 2.57

Chapter 7 연속확률분포

◎ 정규분포의 확률계산 예

예제 7-2 Y초등학교 전교생의 IQ를 측정해 본 결과 평균 μ = 100, 표준편차 σ = 10이었다. 이 초등학교 학생들의 IQ분포가 정규분포를 이룬다고 가정할 때, IQ가 100~ 110사이인 학생의 비율은 ??

풀이 :: $P(100 \leq X \leq 110) = P(0 \leq Z \leq +1)$
$= 0.3413$

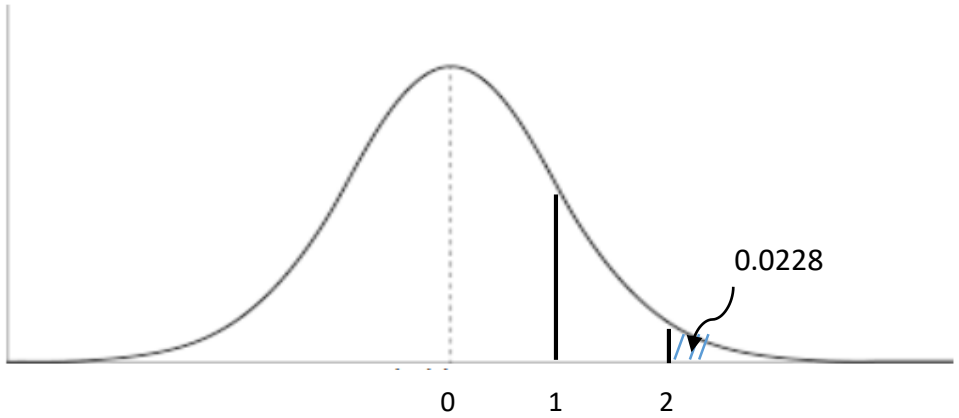


Chapter 7 연속확률분포

◎ 정규분포의 확률계산 예

예제 7-3 7-2의 예제 2에서 IQ가 120 이상인 학생의 비율은 어떻게 될까?
Z-Scoring을 활용하면 풀이는 아래와 같다

풀이 :: $P(X \geq 120) = P(Z \geq 2)$
$= 0.5 - 0.4772 = 0.0228$



Chapter 7 연속확률분포

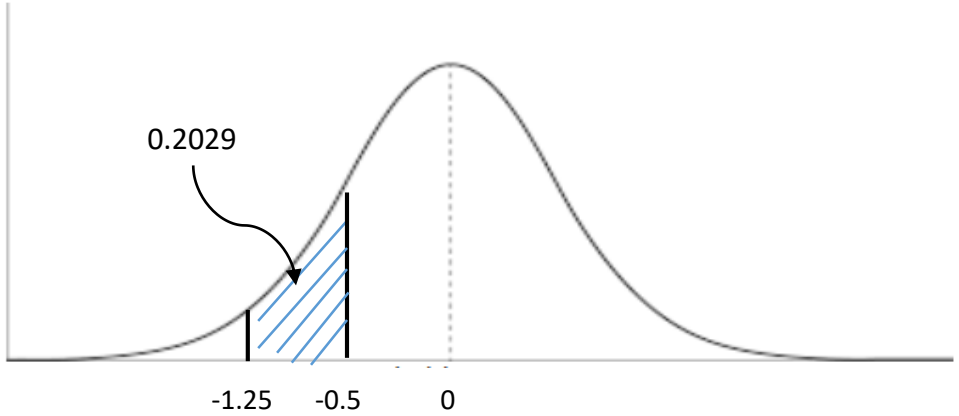
◎ 정규분포의 확률계산 예

예제 7-4 J중학교 1학년 평균 키가 145cm이고, 표준편차는 20cm이며, 정규분포를 이룬다고 하면 전체학생 중에서 키가 120 ~ 135 cm인 학생의 비율은???

풀이 :: $P(120 \le X \le 135) = P(-1.25 \le Z \le -0.5)$

$= P(-1.25 \le Z \le 0) - P(-0.5 \le Z \le 0)$

$= 0.3944 - 0.1915 = 0.2029$

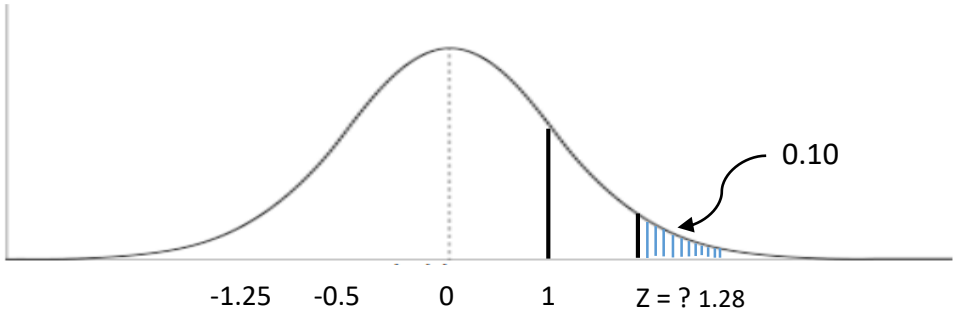


Chapter 7 연속확률분포

◎ 정규분포의 확률계산 예

예제 7-5 Y중학교의 중간고사 시험점수가 정규분포를 이루고 있고, 평균이 60점이며 표준편차가 15점이라 한다. 이 학교에서 상위 10%에 해당하는 사람에게 장학금을 주려 하는데, 몇 점 이상의 학생에게 지급하여야 하는가?

풀이 :: 상위 10%에 해당하는 z값을 알기 위해서는 표준정규분포표(Z분포표)에서 면적이 40%에 해당하는 점(Critical value)을 찾아야 한다. Z=1.28일 때 면적이 0.3997로 0.4에 가장 가까우므로 1.28을 택하는 것이 좋다



풀이 :: $Z = \frac{X-\mu}{\sigma} \rightarrow 1.28 = \frac{X-60}{15} \rightarrow$

$1.28 \times 15 = X - 60 \rightarrow X = 79.2$

따라서 장학금을 받으려면 최소 79.2점 이상을 받아야 함

Chapter 7 부록 Z정규분포표

Areas Under the One-Tailed Standard Normal Curve

This table provides the area between the mean and some Z score.
For example, when Z score = 1.45 the area = 0.4265.

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000

Daniel Rex Park
(박영식)

Chapter 8. 표본 및 표집분포

Daniel Rex Park
(박영식)

Chapter 8 표본 및 표집분포

◎ 표본조사의 필요성

장점

1) 경제성 : 시간과 비용의 절약
2) 시간의 제약 : Time Limit 과 Due Date의 존재
3) 무한 모집단 : 모집단이 무한히 큰 경우
4) 조사가 불가능한 모집단 : h학교 졸업생 전체 조사 → 사망한 사람 또는 주소가 변경되어 추적이 불가한 사람
5) 정확성 : 적은 수의 표본을 관찰할 때의 좀 더 조심스럽고 세심한 접근
6) 그 밖의 이유 : 때때로 대상을 조사하는 행위 자체가 분석 대상의 성격과 형질을 변형을 야기

단점

1) 모집단과 표본 사이에서의 오차 발생

Chapter 8 표본 및 표집분포

◎ 표본추출방법

- 표본조사에 의한 모집단의 이해는 필연적인 오차를 수반
- 표본조사 – 1) 편의(bias)에 의한 오차
  2) 우연성(chance)에 의한 오차
- 오차감소방법 – 1) 표본추출방법을 과학적으로 계획
  2) 표본의 크기를 증가시킴으로써 감소

◎ 확률표본추출

- 확률표본추출(probability sampling)이란 모집단에 속해 있는 각 구성원이 표본으로 선택될 가능성이 일정하게 되도록 하는 표본추출방법

단순무작위추출(Simple random sampling) 난수표 활용 및 기타방법 동원

층화추출(stratified sampling) 모집단의 성격에 따라 여러 집단 또는 여러 층으로 분류한 후 추출

군집추출(cluster sampling) 직접 개별적인 구성원이 아닌 자연적 또는 인위적인 집단을 추출

체계적 추출(systematic sampling) 모집단배열이 무작위일 때 체계적 수단을 동원하여 추출

Chapter 8 표본 및 표집분포

◎ 비확률표본추출

- 비확률표본추출(nonprobability sampling)은 확률표본추출(무작위추출)이 불가능하거나 비경제적일 경우, **연구자가** 모집단과 비슷하다고 생각되는 **표본을 임의로 추출**해 내는 방법

편의추출(convenience sampling) 연구자가 가장 손쉽게 구할 수 있는 구성원을 선택하여 표본으로 삼는 표본추출

판단추출(judgement sampling) 전문성이 있는 연구자가 임의로 표본추출을 하는 방법

할당추출(quota sampling) 모집단의 특성을 대표할 수 있게 몇 개의 하위집단을 구성한 후 각 집단별로 표본의 수를 할당하여 임의로 표본을 추출하는 방법

눈덩이추출(snowball sampling) 이미 참가하고 있는 사람들에게 그들이 알고 있는 사람들로부터 다른 설문조사 참가자들을 모집해 달라고 요청하는 것

Chapter 8 표본 및 표집분포

◎ 표본추출오차와 비표본추출오차

예제 8-1

어느 회사의 신입사원 IQ를 측정하기 위해 임의로 3명을 선정하여 조사하였다. 그 결과 평균 124라는 결론을 얻어 다른 회사에 비해 매우 높은 수준이라 판단하였다. 그러나 사실 신입사원 1,025명 전체의 IQ 평균은 96이었다.

표본추출오차

예제 8-2

1936년 리터러리 다이제스트(Literary Digest) 잡지사는 미국 대통령 후보인 민주당의 프랭클린 루스벨트와 공화당의 알프레드 랜든의 선거결과를 예측하였다. 다이제스트사는 수백만이나 되는 대규모 표본을 뽑아 유권자의 반응을 조사한 결과를 토대로 랜든의 압도적인 승리를 예상 및 발표하였다. 그러나 선거결과 미국 역사상 손꼽힐 만큼의 큰 차이로 루스벨트가 당선되었다.

표본추출오차

예제 8-3

어느 회사에서 부서별로 건강검진을 실시하였다. 검진을 끝마친 후 체중 데이터에서 이상한 점을 발견하였다. 총 6개의 평균 체중이 68kg, 72kg, 62kg, 70kg, 89kg, 95kg으로 마지막 두 개 부서의 평균이 다른 부서들에 비하여 월등히 높았던 것이다. 사실을 조사해본 결과, 검진시간이 부족하여 마지막 두 개 부서는 장소를 옮겨 평소에 쓰이지 않던 저울로 체중을 측정하였다고 한다.

비표본추출오차

Chapter 8 표본 및 표집분포

◎ 표본추출오차와 비표본추출오차

표본추출오차

표본추출오차란 모집단을 대표할 수 있는 전형적인 구성요소를 표본으로 선택하지 못했기 때문에 발생하는 오류.

표본추출상의 오류는 두 가지 요인(Factor)에 의해 발생하게 되는데,

1) 표본의 크기에 따른 우연적 오류
2) 모집단을 대표할 수 없는 비전형적인 구성요소를 표본으로 뽑았기에 일어나는 오류

비표본추출오차

표본의 특성 값을 측정하는 방법이 부정확하기 때문에 발생하는 오류 (measurement error)

Ex)한 사람에게 똑같은 질문을 서로 다른 두 사람이 할 경우에 다른 답변이 나오는 경우

(1) 단순히 표본의 수를 늘려도
(2) 모집단 전체를 연구대상으로 하여도
(3) 표본추출계획을 면밀히 수립하여도 감소하지 않음

Chapter 8 표본 및 표집분포

◎ 통계량과 표집분포

모수

- 연구자의 관심 연구대상에 따라 많을수도, 적을수도 있음
- 우리나라 대학생이 모두 연구 대상이라면 수백만명의 학생들이 모집단이 됨
- 우리 가족이 연구대상이 된다면 매우 적은 수가 모집단을 구성하게 됨
- 유한모집단(finite population) : 모집단이 크건 작건 관계없이 모집단의 구성원수가 유한한 모집단
- 무한모집단(infinite population) : 모집단을 구성하고 있는 요소가 무한한 것, 무한대의 실험이 행해질 때 가능한 모든 결과를 포함시켜 모집단으로 보는 경우

유한모집단의 모수 계산

평균 $\mu = \frac{\sum X_i}{N}$

분산 $\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$

표준편차 $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$

평균 $\mu = E(x) = \sum P_i * X_i$

분산 $\sigma^2 = \sigma^2(x) = \sum [X_i - E(x)]^2 * (X_i)$

표준편차 $\sigma = \sqrt{\sigma^2} = \sqrt{\sum [X_i - E(x)]^2 * (X_i)}$

Chapter 8 표본 및 표집분포

◎ 통계량과 표집분포

예제 8-4 어느 기업에 입사한 5명의 신입직원들 5명의 대학졸업성적이 3.74, 3.89, 4.00, 3.68, 3.69라고 할 때 이들의 평균, 분산, 표준편차를 구하라.

$$\mu = \frac{3.74 + 3.89 + 4.00 + 3.68 + 3.69}{5} = 3.80$$

$$\sigma^2 = \frac{(3.74-3.80)^2 + (3.89-3.80)^2 + (4.00-3.80)^2 + (3.68-3.80)^2 + (3.69-3.80)^2}{5}$$

$$= 0.016$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{0.016} = 0.126$$

예제 8-5 여행가이드가 공항에서 2명의 여행객을 기다리고 있다. 지금까지의 경험으로 볼 때, 2명 모두 정해진 시간 안에 공항에 도착할 확률은 75%이며, 1명이 지각할 확률은 15%, 두 명 모두 지각할 확률은 10% 이다. 2명의 여행객 중 지각할 사람 수에 대한 평균, 분산, 표준편차는?

$$\mu = E(x) = 0*(0.75) + 1*(0.15) + 2*(0.10) = 0.35$$

$$\sigma^2 = (0-0.35)^2 * 0.75 + (1-0.35)^2 * 0.15 + (2-0.35)^2 * (0.10) = 0.4275$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{0.4275} = 0.6538$$

Chapter 8 표본 및 표집분포

◎ 통계량과 표집분포

- 통계량 : 표본을 구해서 **표본의 분포특성**을 **계수화**하는 방법
- 대표값 : 모수와 마찬가지로 평균, 분산, 표준편차
- 대표값 표시방식 : 평균 $\bar{X}$, 분산 S^2 , 표준편차 S

통계량 계산

평균 $\bar{X} = \frac{\sum X_i}{n}$

분산 $S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$

표준편차 $S = \sqrt{S^2} = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$

Chapter 8 표본 및 표집분포

◎ 통계량과 표집분포

예제 8-6 과천에 있는 서울대공원 입장객들의 평균 연령을 알아보기 위하여 4명을 표본으로 뽑은 결과, 그들 연령이 5 ,17 ,12 ,10세였다. 이 표본의 평균, 분산, 표준편차는 얼마인가?

$$\bar{X} = \frac{5+17+12+10}{4} = 11(\text{세})$$

$$S^2 = \frac{(5-11)^2 + (17-11)^2 + (12-11)^2 + (10-11)^2}{4-1} = 24.67$$

$$S = \sqrt{S^2} = \sqrt{24.67} = 4.97(\text{세})$$

Chapter 8 표본 및 표집분포

◎ 통계량과 표집분포

• 모집단의 특성을 추정하기 위해 표본을 대신 분석한다는 것은 표본이 포함하고 있는 오차를 추정해 낼 수 있음을 의미

• 만일 표본의 오차를 추정할 수 없다면 표본은 소용없음

• 한 모집단에서 n의 크기를 가진 선택가능한 표본을 모두 뽑는다고 가정한 다음 그 표본들의 통계량의 분포를 분석해 봄으로써 우리가 뽑은 표본의 분포에서의 위치와 모집단의 분포 및 특성을 추론할 수 있음

• 똑같은 크기를 가진 표본을 여러 번 추출했을 때 각 **표본의 특성치인 통계량들 역시 분포를 갖게 됨 이 분포**를 **표집분포(sampling distribution)**이라고 함

표집분포

표집분포란 모집단에서 일정한 크기로 뽑을 수 있는 표본을 모두 뽑았을 때, 그 모든 표본의 특성치, 즉 **통계량의 확률분포**

Chapter 8 표본 및 표집분포

◎ 평균의 표집분포

- 평균의 표집분포(sampling distribution of means)란 선택가능한 모든 표본들로 부터 계산된 평균($\bar{X}$)들의 확률분포를 의미

평균의 표집분포

평균의 표집분포란 특정 모집단에서 동일한 크기로 가능한 **모든 표본을 뽑아서 각각의 표본들의 평균**을 계산했을 때, 그 **평균들의 확률분포**

예제 8-7

어느 상자에 30, 60, 90이 쓰여진 카드 3장이 있다. 이 세 장의 카드 중 **복원추출**로 두 장의 카드를 표본으로 뽑을 때 그 표본들의 평균의 분포양상을 알아보려고 한다.

X_i	$P(X_i)$	
30	1/3	$\mu = 60$
60	1/3	$\sigma^2 = 600$
90	1/3	

Chapter 8 표본 및 표집분포

◎ 평균의 표집분포

예제 8-7

어느 상자에 30, 60, 90이 쓰여진 카드 3장이 있다. 이 세 장의 카드 중 **복원추출**로 두 장의 카드를 표본으로 뽑을 때 그 표본들의 평균의 분포양상을 알아보려고 한다 - (계속)

n = 2 일때의 표본

가능한 표본	**표본의 평균($\bar{X}_i$)**
30, 30	$\bar{X}_1$ = 30
30, 60	$\bar{X}_2$ = 45
30, 90	$\bar{X}_3$ = 60
60, 30	$\bar{X}_4$ = 45
60, 60	$\bar{X}_5$ = 60
60, 90	$\bar{X}_6$ = 75
90, 30	$\bar{X}_7$ = 60
90, 60	$\bar{X}_8$ = 75
90, 90	$\bar{X}_9$ = 90

Chapter 8 표본 및 표집분포

◎ 평균의 표집분포

예제 8-7

어느 상자에 30, 60, 90이 쓰여진 카드 3장이 있다. 이 세 장의 카드 중 **복원추출**로 두 장의 카드를 표본으로 뽑을 때 그 표본들의 평균의 분포양상을 알아보려고 한다 - (계속)

평균의 표집분포

$\bar{X}_i$	$P(\bar{X}_i)$
30	1/9
45	2/9
60	3/9
75	2/9
90	1/9

◎ 평균의 표집분포의 평균

평균의 표집분포의 평균

$$\mu_{\bar{X}} = \sum \bar{X}_i * P(\bar{X}_i)$$

$$\mu_{\bar{X}} = \sum \bar{X}_i * P(\bar{X}_i) = 30 * \frac{1}{9} + 45 * \frac{2}{9} + 60 * \frac{3}{9} + 75 * \frac{2}{9} + 90 * \frac{1}{9} = 60$$

Chapter 8 표본 및 표집분포

◎ 평균의 표집분포의 평균

평균의 표집분포의 평균과 모집단의 평균

$$\mu_{\bar{X}} = \mu$$

표본평균의 평균 $\mu_{\bar{X}}$ 는 모집단의 평균인 μ 와 언제나 일치한다.

예제 8-8

어느 회계법인의 전체회원이 500명일 때, 이들의 평균 연령이 32.3세라고 한다. 이들 중에서 $n = 10$인 표본을 모두 뽑았을 때, 이 표집분포의 평균은 얼마인가?

$$\mu_{\bar{X}} = \mu = 32.3(\text{세})$$

◎ 평균의 표집분포의 분산

평균의 표집분포의 분산

$$\sigma^2_{\bar{X}} = E(\bar{X}_i - \mu_{\bar{X}})^2 = \sum(\bar{X}_i - \mu_{\bar{X}})^2 * P(\bar{X})$$

Chapter 8 표본 및 표집분포

◎ 평균의 표집분포의 분산

평균의 표집분포의 분산

$$\sigma^2_{\bar{X}} = E(\bar{X}_i - \mu_{\bar{X}})^2 = \sum(\bar{X}_i - \mu_{\bar{X}})^2 * P(\bar{X})$$

모분산 $\sigma^2 = (30-60)^2 * \frac{1}{3} + (60-60)^2 * \frac{1}{3} + (90-60)^2 * \frac{1}{3} = 600$

표본평균의 분산 $\sigma^2_{\bar{X}} = (30-60)^2 * \frac{1}{9} + (45-60)^2 * \frac{2}{9} + (60-60)^2 * \frac{3}{9} + (75-60)^2 * \frac{2}{9} + (90-60)^2 * \frac{1}{9} = 300$

표본의 수가 많다면
(ex : 500명의 회원들 중 복원추출로 두 명 뽑을때 선택가능 표본 수→ 500 * 500 = 250,000개
세 명 뽑을때 선택가능 표본 수→ 500 * 500 * 500 = 125,000,000개

◎ 평균의 표집분포의 분산

평균의 표집분포의 분산

분 산 $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$

표준편차 $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

예제 8-9

어느 음료수 공장에서 생산되는 콜라의 평균 용량은 360ml이고 이 모집단의 분산은 50이라고 하자. $n = 10$인 표본을 모두 뽑았을 때, 그 표집분포(= 표본평균)의 기댓값과 분산은 얼마인가?

$$\mu_{\bar{X}} = \mu = 360\text{ml}$$

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} = \frac{50}{10} = 5$$

Chapter 8 표본 및 표집분포

◎ 모집단의 분포와 평균의 표집분포

모집단이 정규분포일 때

평균의 표집분포의 모양은 표본이 뽑힌 모집단이 정규분포인가 아닌가와, 표본의 크기 n 에 따라 달라진다. 먼저 모집단이 정규분포일 때를 살펴보자.

모집단이 정규분포일 때 평균의 표집분포

모집단이 정규분포일 때 평균의 표집분포는 표본의 크기 n 과 상관없이 언제나 정규분포를 이루며, 표집분포의 평균 $\mu_{\bar{X}}$ 표준편차 $\sigma_{\bar{X}}$ 는 아래와 같다

$$\mu_{\bar{X}} = \mu$$
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Chapter 8 표본 및 표집분포

◎ 모집단의 분포와 평균의 표집분포

예제 8-10

전교생이 3,000명인 A고등학교에서 학생들의 IQ를 조사하였더니, 평균 μ 이 102이고, 표준편차 σ 는 16이었다. 이 학생들의 IQ분포가 정규분포라고 가정한다면,

(1) IQ가 94이하인 학생들은 몇 %S나 존재하는가?
(2) 만약 4명씩 임의로 표본을 뽑는다면 표본평균 $\bar{X}$의 분포는 어떠한 모양일까?

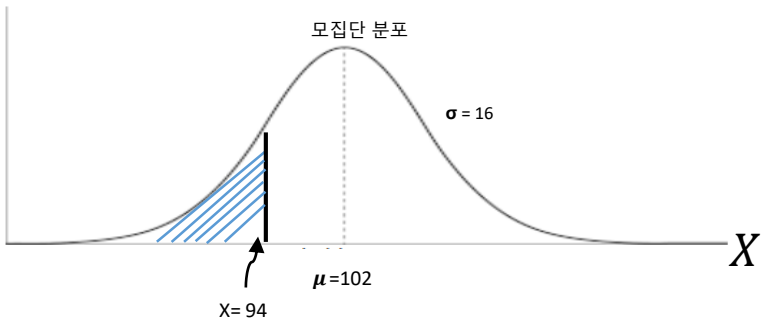
풀이 :: 여기서 (1)은 모집단, (2)는 평균의 표집분포에 관한 이야기이다.

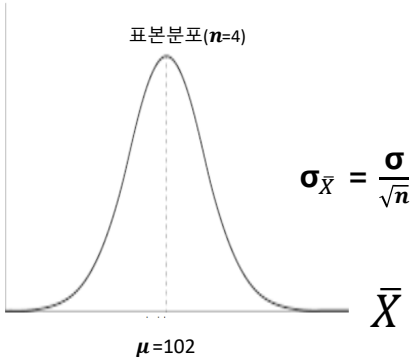
(1) IQ가 94이하인 학생들의 비율을 알려면 모집단 분포 그림에서 파란 면적을 구해야만 한다

$P(X \leq 94) = P(z \leq \frac{X-\mu}{\sigma} = \frac{94-102}{16}) = P(Z \leq -0.5)$

→ 0.5 – 0.1915 = 0.3085

(2) 풀이 : :





$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{16}{\sqrt{4}} = 8$

Chapter 8 표본 및 표집분포

◎ 모집단의 분포와 평균의 표집분포

$\bar{X}$에 해당되는 z 값

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

예제 8-11

어느 영화사에서의 평균 영화재생시간은 5,600시간, 표준편차는 4,000시간이며 분포는 정규분포를 이룬다고 한다. 이 영화사에서 전국에서 100개의 영화관을 임의로 추출하여 평균 음악재생시간을 체크했을 때, 6,400시간 이상일 확률은 얼마인가?

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{4{,}000}{100} = 400$$

$\bar{X}$ = 6,400에 해당하는 z값은

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{6{,}400 - 5{,}600}{400} = 2$$

$P(\bar{X} \geq 6{,}400) = P(Z \geq 2.0)$ → 0.5-0.4772 = 0.0228

※ 표본의 평균 $\bar{X}$가 6,400시간 이상일 확률은 2.28%이다.

Chapter 8 표본 및 표집분포

◎ 모집단의 분포와 평균의 표집분포

모집단이 정규분포가 아닐 때

표집분포는 표본의 크기 n 을 크게 할수록 정규분포에 접근하게 됨. 이를 **중심극한정리(central limit theorem)**라 하는데, 이때에도 역시 **표집분포의 평균은 모집단의 평균**과 **일치**하고 **표집분포의 표준편차는 모집단의 표준편차를 표본크기의 제곱근으로 나눈 것과 같다.**

중심극한정리

중심극한정리란 모집단의 분포모양과는 상관없이 평균 μ, 분산 σ^2인 모집단에서 크기가 n 인 선택가능한 모든 표본을 뽑을 때, 평균의 표집분포는 n을 증가시킬수록 정규분포에 접근하게 된다는 것!

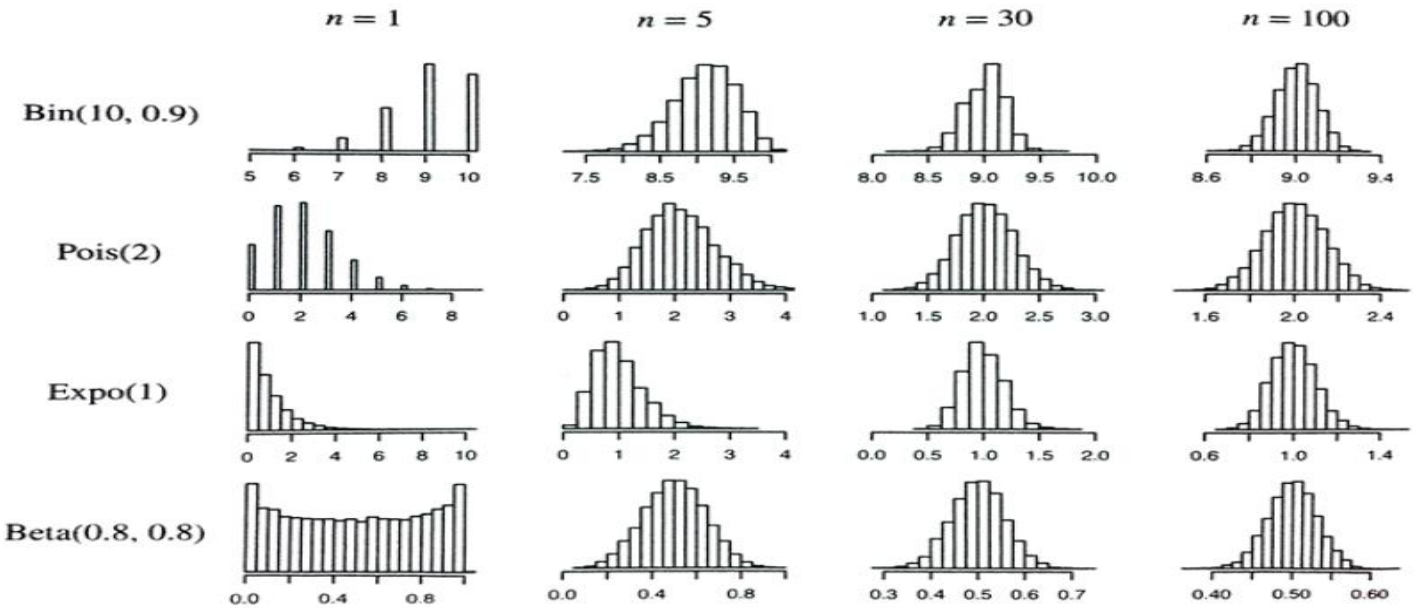
$$\mu_{\bar{X}} = \mu$$
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Chapter 8 표본 및 표집분포

◎ 모집단의 분포와 평균의 표집분포

모집단이 정규분포가 아닐 때

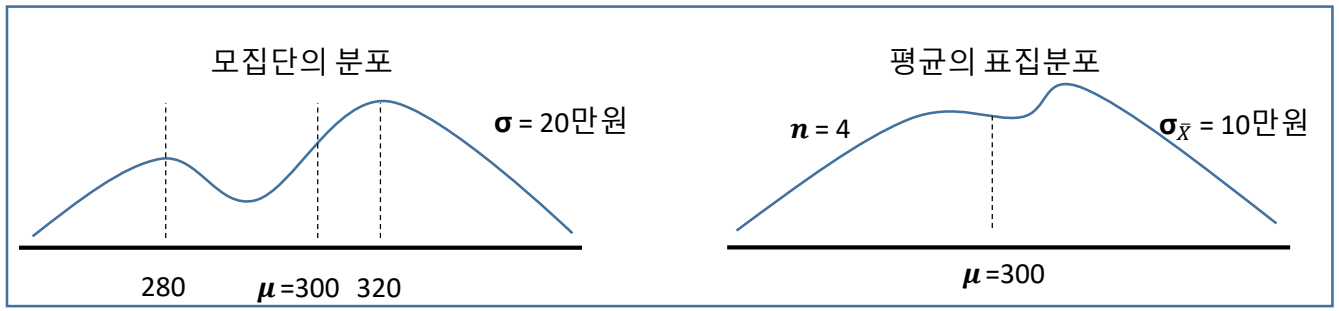
표집분포는 표본의 크기 n 을 크게 할수록 정규분포에 접근하게 됨. 이를 **중심극한정리(central limit theorem)**라 하는데, 이때에도 역시 **표집분포의 평균은 모집단의 평균**과 **일치**하고 **표집분포의 표준편차는 모집단의 표준편차를 표본크기의 제곱근으로 나눈 것과 같다.**



Chapter 8 표본 및 표집분포

예제 8-12

우리나라의 가구당 소득분포가 대체로 월수입 280만원과 320만원에 많이 몰려 있고, 평균은 300만원, 표준편차는 20만원이라고 하면, 4가구를 임의로 뽑았을 경우에 평균의 표집분포는 어떻게 될까?



A. 모집단이 정규분포가 아니고 표본의 크기 $n = 4$ 밖에 되지 않으므로 $\bar{X}$ –분포에도 모집단의 양봉형 흔적이 남게 되며,

$\mu_{\bar{X}} = \mu = 3{,}000{,}000$ (원)

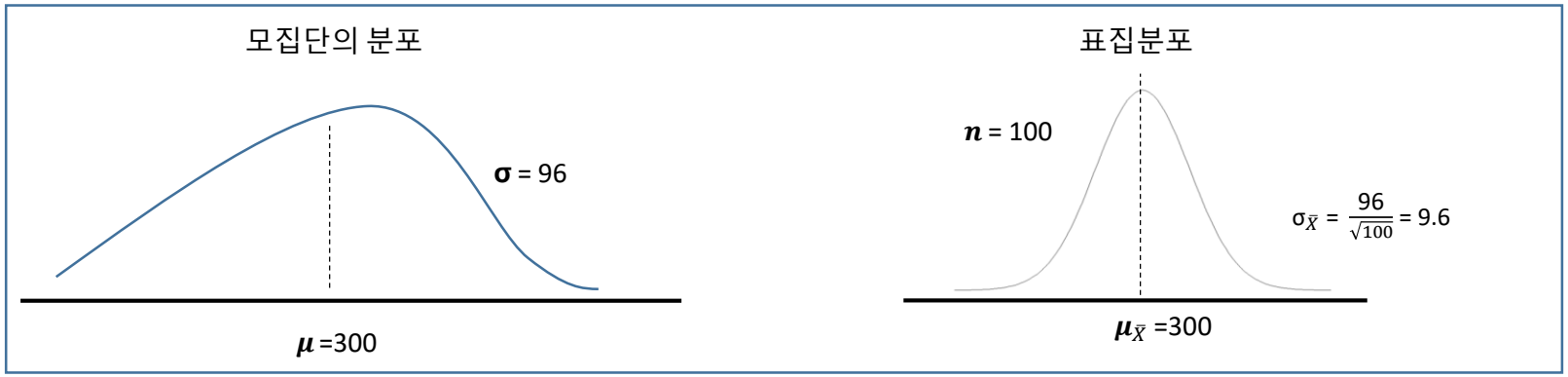
$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{200{,}000}{2} = 100{,}000$(원)

n이 커질수록 정규분포에 가까워진다

Chapter 8 표본 및 표집분포

예제 8-13

수능시험 점수의 분포가 평균 300점이며, 표준편차는 96이고, 분포의 모양이 아래와 같다고 하자. 이 중 임의로 100명씩을 계속 뽑을 때 평균의 표집분포 모양은 어떻게 될 것인가?



A. 표본의 크기가 100이므로 모집단 분포의 모양에 관계없이 $\bar{X}$-분포는 정규분포를 이룬다고 볼 수 있으며, 이 때의 표집분포의 평균과 분산은 아래와 같다

$$\mu_{\bar{X}} = \mu = 300$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{96}{10} = 9.6$$

Chapter 8 표본 및 표집분포

◎ 모집단의 분포와 평균의 표집분포

모집단의 크기가 작을 때

- 표본의 크기 n 이 30이상이 되면 모집단의 분포가 정규분포든지, 비정규분포든지 상관없이 **평균의 표집분포는 정규분포에 근사**하게 되며, 평균 $\mu_{\bar{X}}$는 모평균μ와 같고 표준편차 $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ 임을 알았다
- 다만 위와 같은 결론은 **모집단이 매우 크거나 무한(infinite)하다는 가정**하에서 **성립**되는 것이며, 모집단의 크기 N이 작은 유한모집단에서는 표준편차에 대한 조정이 필요
- 모집단의 크기가 작더라도 복원추출을 한다면 무한모집단과 같은 효과를 얻을 수 있기에 조정이 필요없음
- 하지만, 비복원추출인 경우 조정을 해야함
- $\sqrt{(N-n)/(N-1)}$ 은 유한 모집단일 때 $\bar{X}$-분포의 표준편차 계산을 위한 조정계수(correction factor)

모집단이 작을 때 표본평균($\bar{X}$)의 표준편차

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}}$$

N : 모집단의 크기
n : 표본의 크기

Daniel Rex Park
(박영식)

◎ 모집단의 분포와 평균의 표집분포

예제 8-14

S회사의 종업원은 500명인데 이들의 임금이 평균 2,350,000원, 표준편차 200,000원인 정규분포를 이룬다면, 이들 중에서 비복원추출로 100명을 표본으로 뽑을 때, $\bar{X}$-분포의 평균($\mu_{\bar{X}}$)와 표준편차($\sigma_{\bar{X}}$)는?

- 표본의 크기 n 이 100이며, 모집단의 크기가 500명이므로 **표본의 크기가 5%를 넘는다** :: 따라서 $\bar{X}$-분포의 표준편차는 조정계수를 사용하여 조정하여야 함

$$\mu_{\bar{X}} = \mu = 2{,}350{,}000(\text{원})$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}} = \frac{200{,}000}{\sqrt{100}} * \sqrt{\frac{500-100}{500-1}} = 20{,}000 * 0.8953 = 17{,}906\ (\text{원})$$

- $\bar{X}$-분포의 표준편차($\sigma_{\bar{X}}$)를 계산한 후, 그 분포에서 $\bar{X}$에 해당되는 Z값을 아래와 같이 구함

유한모집단 일 때 표본평균($\bar{X}$)에 해당되는 z값

$$Z = \frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X}-\mu_{\bar{X}}}{\frac{\sigma}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}}}$$

Chapter 8 표본 및 표집분포

◎ 분산의 표집분포

모집단이 σ^2의 분산을 가질 때, 이 모집단으로부터 크기가 동일하게 선택가능한 모든 표본을 뽑아서 각각의 분산을 계산했을 때, 표본분산 S^2 들은 일정한 분포를 이루게 되는데, 이것이 **분산의 표집분포**이다

A = {2, 4, 6, 8} , n = 2

:: 모집단의 평균과 분산

μ = (2+4+6+8) / 4 = 5

$$\sigma^2 = \sum(X_i - \mu)^2 * P(X_i) = (2-5)^2 * \frac{1}{4} + (4-5)^2 * \frac{1}{4} + (6-5)^2 * \frac{1}{4} + (8-5)^2 * \frac{1}{4} = 5$$

표본	뽑힐 확률	분산(S^2)	표본	뽑힐 확률	분산(S^2)
2,2	1/16	0	6,2	1/16	8
2,4	1/16	2	6,4	1/16	2
2,6	1/16	8	6,6	1/16	0
2,8	1/16	18	6,8	1/16	2
4,2	1/16	2	8,2	1/16	18
4,4	1/16	0	8,4	1/16	8
4,6	1/16	2	8,6	1/16	2
4,8	1/16	8	8,8	1/16	0

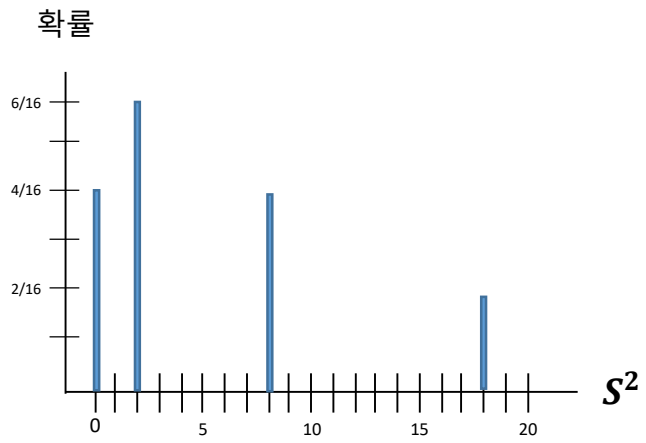
#) 각 표본에서의 분산은 $S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$ 으로 계산하였음 예를 들어, 2와 4로 구성된 표본의 분산은 $\bar{X}$가 3이므로, S^2은 $[(2-3)^2 + (4-3)^2]$ / (2 − 1) = 2가 됨

Chapter 8 표본 및 표집분포

◎ 분산의 표집분포

모든 표본 16개로부터 구한 분산의 확률분포는 아래의 표와 같다. 바로 이것이 **분산의 표집분포**이다. 그리고 이를 그래프로 나타내면 아래의 그래프와 같다.

분산(S^2)	뽑힐 확률
0	4/16
2	6/16
8	4/16
18	2/16



$E(s^2) = 0 * \frac{4}{16} + 2 * \frac{6}{16} + 8 * \frac{4}{16} + 18 * \frac{2}{16} = 5$ :: 모집단의 분산 // 표본 분산의 기댓값 $E(S^2)$ 과 같음 – 이 특성은 **표본의 크기 n** 와 상관없이 언제나 성립

분산의 표집분포의 평균과 모집단 분산

$$E(S^2) = \sigma^2$$

Chapter 8 표본 및 표집분포

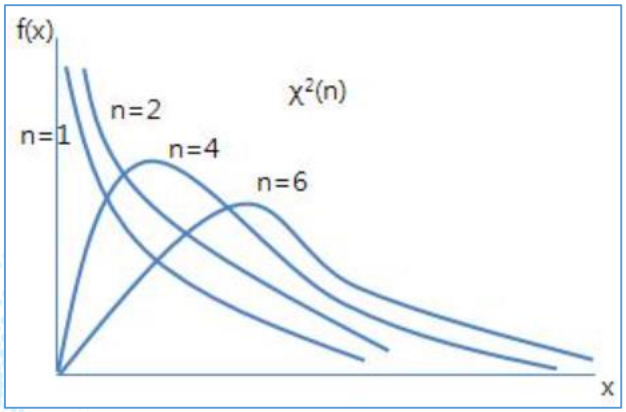
◎ X^2-분포

- S^2의 표집분포의 평균이 모집단의 분산인 σ^2과 같음
- 분산의 표집분포 모양은 n 의 크기에 따라 X^2- 분포(chi-square distribution)를 이룸

X^2-분포

분산이 σ^2 을 갖는 정규분포를 이루는 모집단으로부터 표본의 크기가 n인 선택가능한 모든 표본을 뽑을 때, 각 표본의 분산을 S^2 이라고 하면, X^2-분포는 다음과 같다.

$$X^2_{n-1} = \frac{(n-1)*s^2}{\sigma^2}$$



이와 같은 분포를 X^2- 분포(chi-square distribution)라고 한다. X^2- 분포는 **비대칭의 모양으로 오른쪽으로 긴 꼬리**를 가지며, 항상 양수값만을 갖는 특징을 지닌다

자유도가 커질수록X^2- 분포는 정규분포에 가깝게 된다. X^2- 분포의 모양은 자유도에 따라 달라지며 평균은 $(n-1)$이다

Chapter 8 표본 및 표집분포

◎ 비율의 표집분포

- 어느 대학의 **신입생 1,000명 중에 여학생이 400명**이라고 할 때, 여학생의 비율을 모르는 사람이 여학생의 비율을 알고 싶어 **100명을 뽑은 결과 여학생이 30명**이었다고 하자
- 즉, **표본으로부터 30%가 여학생**이라는 결과가 도출됨
- 각 표본으로부터 비율을 구했다면, **모집단의 비율에 가까운 수치가 많을 것이며**, 모집단의 비율과 차이가 큰 표본은 적을 것
- 동일한 모집단에서 선택가능한 모든 표본을 뽑아 구한 비율들의 분포를 **비율의 표집분포(sampling distribution of proportion)**이라고 함

$$\pi = \frac{X}{N} \qquad p = \frac{X}{n}$$

N : 모집단의 수
X : 총성공횟수

n : 표본의 크기
X : 표본에서의 성공횟수

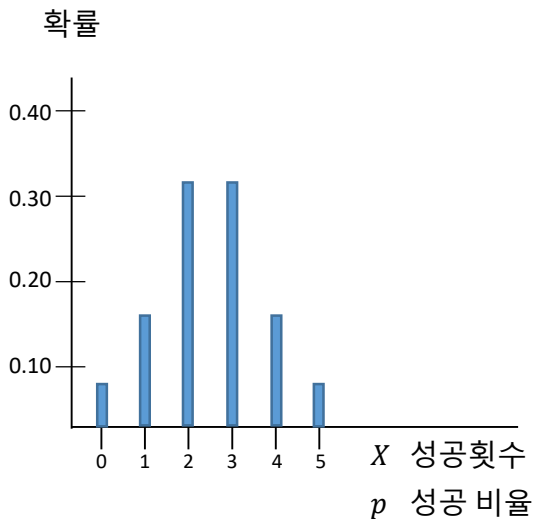
Chapter 8 표본 및 표집분포

◎ 비율의 표집분포

- 동전을 던져서 앞면이 나오는 것을 성공이라 한다면, 모집단의 성공비율 $\pi = 0.5$
- 동전을 다섯 번 던졌을 때의 비율의 표집분포는? :: 5번의 시행은 아래와 같이 표현됨

X(성공횟수)	*p(성공비율)*	뽑힐 확률
0	0.0	0.031
1	0.2	0.156
2	0.4	0.313
3	0.6	0.313
4	0.8	0.156
5	1.0	0.031

$$P(X = x) = {}_nC_x p^x (1-p)^{n-x}$$



Chapter 8 표본 및 표집분포

◎ 비율의 표집분포

- 모집단의 성공비율 π 일 때, 표본의 성공비율 p-분포의 변화를 알아보면 아래와 같은 식이 성립

성공횟수의 기대값과 분산

$$E(X) = n\pi$$
$$Var(X) = n\pi(1-\pi)$$

- 위의 식을 비율로 표현하여 비율의 표집분포의 평균(기댓값)과 분산을 구하면 아래와 같음

비율의 표집분포의 기댓값과 분산

$$E(p) = \pi$$
$$\sigma^2_p = \pi(1-\pi)/n$$

Chapter 9. 통계적 추정 /가설검정

Daniel Rex Park
(박영식)

Chapter 9 통계적 추정

◎ Intro

- 모집단의 성격을 모르는 상황에서 모집단의 성격 규명이란?
- 추리통계학이란?
- 추리통계학- 1) 통계적 추정
- 추리통계학- 2) 가설 검정

Daniel Rex Park
(박영식)

Chapter 9 통계적 추정

◎ 통계적 추정의 기본 개념

- 전국 고등학교 3학년 학생들의 수능 평균점수를 알아본다고 가정
- 전국 각 고등학교 3학년 졸업생들 중 적정한 수의 학생을 표본으로 뽑아 모집단의 모수, 수능 점수를 추정해야 함
- 만약 평균점수가 300점이었다면 모수의 평균점수는?

A. 300점일 것이다.
B. 270~330점 사이일 것이다.
C. 170~330점 사이일 것이다.

모집단의 특성을 추정하는 방법

모집단의 특성을 하나의 값으로 추정하는 점추정(point estimation)방법과 B나 C처럼 적절한 구간을 가지고 모수를 추정하는 구간추정(interval estimation)방법이 있다.

Chapter 9 통계적 추정

◎ 점추정과 추정량

- 점추정을 위해서는 추정값(estimate)과 추정량(estimator)를 정의해보자

예를 들어 모집단의 평균 μ 을 알기 위해 표본의 평균 통계량 $\bar{X}$를 이용하게 되는데 이때 $\bar{X}$ 는 μ의 추정량이 되며, 표본평균의 구체적인 수치, 예를 들어 $\bar{X}$ = 300은 모집단의 평균을 추정하는 추정값이 됨

◎ 추정량의 결정기준

(1) 불편성 (unbiasedness)
(2) 효율성 (efficiency)
(3) 일치성 (consistency)
(4) 충분성 (sufficiency)

- 전체를 다 충족하는 추정량이 선택된다면, 이상적이나 모든 조건을 충족시키지 못한다면, 첫째 조건인 불편성에 가장 큰 비중을 두어 적정 추정량을 선택해야 함

Chapter 9 통계적 추정

◎ 추정량의 결정기준

(1) 불편성 (unbiasedness)

- 추정량의 기대값이 추정할 모수의 실제값과 일치하거나 그 값에 가까울 수록 바람직한 추정량
- 기대값과 실제값과 차이가 나면 그 추정량은 편의(bias)가 있다고 함
- 이상적 추정량은 0의 편의를 가짐
- 이때의 추정량을 불편추정량(unbiased estimator)라고 함

(2) 효율성 (efficiency)

- 한 표본에서 계산된 추정량은 되도록 모수에 접근하여야 함
- 즉, 분산이 작을수록 모수를 정확하게 추정할 수 있음

효율성

추정량 중에서 최소의 분산을 가진 추정량이 가장 효율적

Chapter 9 통계적 추정

◎ 추정량의 결정기준

(3) 일치성 (consistency)

- 표본의 크기 $\boldsymbol{n}$ 이 무한히 증가하면 그 표본에서 얻은 추정량이 모수에 근접하게 되는 것을 의미
- 표본의 크기가 커지면 커질수록 추정량과 모수 간의 차이는 차차 작아지게 됨
- 이러한 특성을 지는 추정량을 일치성이 있다고 표현

(4) 충분성 (sufficiency)

- 동일한 표본으로부터 얻은 추정량이 다른 추정량보다 모수에 관해 가장 많은 정보를 제공
- 다시 말해 그 어떤 추정량도 선택된 추정량보다 더많은 정보를 제공할 수 없을때, 이 추정량을 충분성이 있는 추정량이라 함

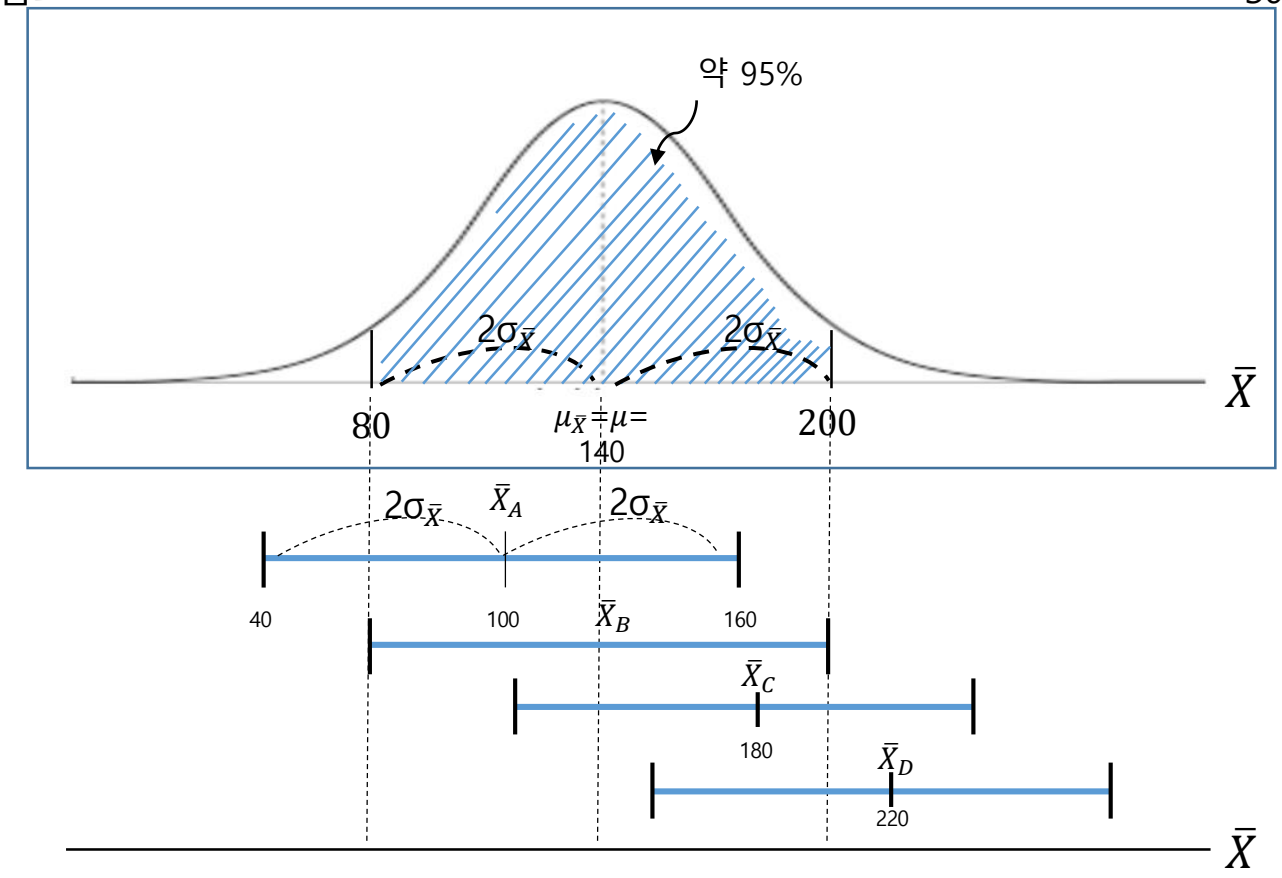
Ex) 표본평균은 모수의 충분추정량 vs 표본의 중앙값은 모수에 대한 충분추정량이 되지 못함

Chapter 9 통계적 추정

◎ 구간추정

예제 9-1 한 회사의 승진시험을 실시한 결과 μ = 140점, 표준편차 σ = 60점이며, 이 점수는 정규분포를 가정

임의로 4명씩 표본으로 뽑는다면, 이들의 표본분포는 평균 30인 정규분포



Chapter 9 통계적 추정

◎ 구간추정

• 예제 9-1에서 표현된 임의로 표본크기가 4인 A, B, C, D의 네 개의 표본을 선택할 때, 그
평균이 $\bar{X}_A$ = 100 , $\bar{X}_B$ = 140, $\bar{X}_C$ = 180, $\bar{X}_D$ = 220이었다.

• 각각의 표본평균에 $2\sigma_{\bar{X}}$만큼씩 양쪽으로 더하거나 빼거나 하는 경우에 $2\sigma_{\bar{X}}$는 2 * 30 = 60점이 된다.

• $\bar{X}_A$ = 100에서 양쪽으로 $\pm 2\sigma_{\bar{X}}$점을 하면, 40 ~ 160점이 된다.
이 구간은 $\mu_{\bar{X}}$ = 140인 모수를 포함한다.

• $\bar{X}_B$ = 140에서 양쪽으로 $\pm 2\sigma_{\bar{X}}$점을 하면, 80 ~ 200점이 된다.
이 구간도 $\mu_{\bar{X}}$ = 140인 모수를 포함한다.

• $\bar{X}_C$ = 180에서 양쪽으로 $\pm 2\sigma_{\bar{X}}$점을 하면, 120 ~ 240점이 된다.
이 구간도 $\mu_{\bar{X}}$ = 140인 모수를 포함한다.

• $\bar{X}_D$ = 220에서 양쪽으로 $\pm 2\sigma_{\bar{X}}$점을 하면, 160 ~ 280점이 된다.
이 구간은 $\mu_{\bar{X}}$ = 140인 모수를 포함하지 않는다.

Chapter 9 통계적 추정

◎ 구간추정

정규모집단으로부터의 표본평균 $\bar{X}$가 $[\mu - 2\sigma_{\bar{X}},\ \mu + 2\sigma_{\bar{X}}]$의 구간에 포함될 확률이 약 0.95인 사실은 이미 증명
되어 있다.

$$P(\mu - 2\sigma_{\bar{X}} \leq \bar{X} \leq \mu + 2\sigma_{\bar{X}}) \approx 0.95$$

이를 예제 9-1에 적용하면 표본의 평균 점수가 $\mu \pm 2\sigma_{\bar{X}}$인 80점과 200점 사이에 있을 확률은 약 0.95가 된다.

표본의 평균 $\bar{X}$를 중심으로 $[\bar{X} - 2\sigma_{\bar{X}},\ \bar{X} + 2\sigma_{\bar{X}}]$의 구간이 μ를 포함할 가능성을 고려하면 위의 식을 변형하여 아래와 같이 표현이 가능하다.

$$P(\bar{X} - 2\sigma_{\bar{X}} \leq \mu \leq \bar{X} + 2\sigma_{\bar{X}}) \approx 0.95$$

위의 식을 표현하는 방법

"위의 식을 읽을 때 ~~μ가 $[\bar{X} - 2\sigma_{\bar{X}},\ \bar{X} + 2\sigma_{\bar{X}}]$의 구간에 포함될 확률이 약 0.95~~"라는 표현인 것 같지만 이는 틀렸다.

:: "**모수μ에 대한 구간추정량인 $[\bar{X} - 2\sigma_{\bar{X}},\ \bar{X} + 2\sigma_{\bar{X}}]$가 모수의 평균 μ를 포함할 확률은 약 0.95**"

Chapter 9 통계적 추정

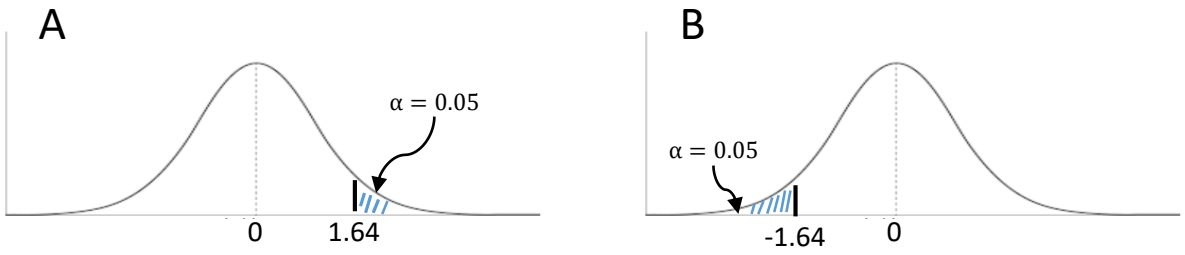
◎ 모집단 평균의 구간추정

σ를 알고 있는 경우 - Z_α값의 계산

앞에서 어떤 표본평균 $\bar{X}$에 $\pm 2\sigma_{\bar{X}}$를 하면, 이 구간 추정량이 모집단의 평균 μ를 포함하고 있을 가능성은 약 95%라는 것을 알았다. 이제 보다 일반적인 신뢰구간의 추정방법을 설명하여 보자.

- 아래의 그림과 같이 Z가 Z_α보다 클 가능성을 α로 표시하여 보자.
- 예를 들어 α = 0.05일 때, $P(Z \geq Z_{0.05})$ = 0.05,
- 즉 파란 면적이 **0.05**인 $Z_{0.05}$ 값을 ***Z* 표준정규분포표에서 찾아보면 1.64**

- 표준정규분포는 좌우대칭이므로 $P(Z \leq Z_{0.05})$도 = 0.05 이에 해당되는 왼쪽의 $-Z_{0.05}$값은 **-1.64**

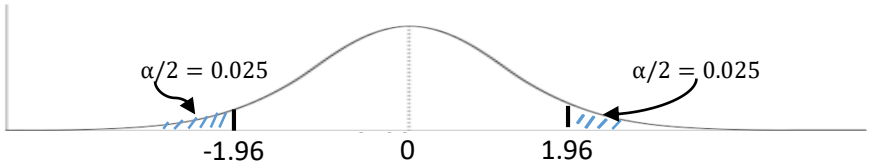


Chapter 9 통계적 추정

◎ 모집단 평균의 구간추정

Σ를 **알고 있는 경우** - Z_α값의 계산 (계속)

- Z_α 또는 $-Z_\alpha$ 보다는 양쪽으로 α/2씩의 확률을 갖는 $\pm Z_{\alpha/2}$ 로 표시하는 경우가 더 많음
- 왜냐하면 한 모집단에서 선택가능한 모든 표본들의 평균은 모집단 평균을 중심으로 **좌우대칭**이므로 신뢰구간을 설정할 때도 한 쪽으로 치우친 구간보다는 **α의 면적**을 **둘로 나누어 양쪽으로 똑같은 넓이**를 구간으로 설정
- 만약 α = 0.05라고 한다면, 양쪽 끝이 각각 α의 1/2, 즉 0.025씩의 면적을 갖게 됨
- $Z_{0.025} = 1.96$, $-Z_{0.025} = -1.96$ 이를 표현하면 아래의 그림으로 표현이 가능



Z 값에 대한 신뢰구간

$$P(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}) = 1 - \alpha$$

Daniel Rex Park
(박영식)

Chapter 9 통계적 추정

◎ 모집단 평균의 구간추정

Σ를 알고 있는 경우 - Z_α값의 계산 (계속)

- $P(-Z_{\alpha/2} \le Z \le Z_{\alpha/2})$ = **1 – α 를 적용하면 →** $P(-1.96 \le Z \le 1.96)$ = **0.95**
- 엄밀한 표현하면, 95%의 구간추정량은 $\bar{X} \pm 2\sigma_{\bar{X}}$ 가 아닌 $\bar{X} \pm 1.96\sigma_{\bar{X}}$ 로 표현

1 – α는 **신뢰도(confidence level)** 또는 **신뢰수준**이라 한다.

신뢰도는 이 같이 구간으로 추정된 추정값이 실제 모집단의 모수를 포함하고 있을 가능성
이 때 모수가 포함될 것으로 추정된 구간 :: **신뢰구간(confidence interval)**이라 함

신뢰도에 따른 $Z_{\alpha/2}$값

신뢰도 (1 – α)	$Z = 0$ 에서 $Z_{\alpha/2}$까지 면적	$Z_{\alpha/2}$
0.90	0.450	1.64
0.95	0.475	1.96
0.99	0.495	2.57

Chapter 9 통계적 추정

◎ 모집단 평균의 구간추정

신뢰구간의 예

예제 9-2

우리나라 학생들의 월평균 용돈을 조사하였다. 100명을 임의추출한 결과, 그들의 평균이 282,000인 것을 확인하였고, 모집단의 표준편차가 10만원이고, 모집단이 정규분포를 이룬다 가정할 때 우리나라 학생들 평균 용돈의 90% 신뢰구간, 95% 신뢰구간, 99%신뢰구간을 구하시오.

n= 100, σ = 100,000원이므로 $\boldsymbol{\sigma}_{\bar{X}} = \frac{\boldsymbol{\sigma}}{\sqrt{n}} = \frac{100{,}000}{\sqrt{100}}$ = 10,000

90% 신뢰구간

$\bar{X}$- $Z_{0.05}$ $\boldsymbol{\sigma}_{\bar{X}} \le \mu \le \bar{X}$+ $Z_{0.05}$ $\boldsymbol{\sigma}_{\bar{X}}$
↔ $\bar{X}$- 1.64$\boldsymbol{\sigma}_{\bar{X}} \le \mu \le \bar{X}$+ 1.64$\boldsymbol{\sigma}_{\bar{X}}$
↔ 282,000 – 1.64 * 10,000 ≤ μ ≤ 282,000 + 1.64 * 10,000
↔ 265,600 ≤ μ ≤ 298,400

95% 신뢰구간

$\bar{X}$- $Z_{0.025}$ $\boldsymbol{\sigma}_{\bar{X}} \le \mu \le \bar{X}$+ $Z_{0.025}$ $\boldsymbol{\sigma}_{\bar{X}}$
↔ $\bar{X}$- 1.96$\boldsymbol{\sigma}_{\bar{X}} \le \mu \le \bar{X}$+ 1.96$\boldsymbol{\sigma}_{\bar{X}}$
↔ 282,000 – 1.96 * 10,000 ≤ μ ≤ 282,000 + 1.96 * 10,000
↔ 262,400 ≤ μ ≤ 301,600

99% 신뢰구간

$\bar{X}$- $Z_{0.005}$ $\boldsymbol{\sigma}_{\bar{X}} \le \mu \le \bar{X}$+ $Z_{0.005}$ $\boldsymbol{\sigma}_{\bar{X}}$
↔ $\bar{X}$- 2.57$\boldsymbol{\sigma}_{\bar{X}} \le \mu \le \bar{X}$+ 2.57$\boldsymbol{\sigma}_{\bar{X}}$
↔ 282,000 – 2.57 * 10,000 ≤ μ ≤ 282,000 + 2.57 * 10,000
↔ 256,300 ≤ μ ≤ 307,700

요약:: 신뢰도가 높을수록 신뢰구간이 넓어짐

- 범위가 넓을수록 그 속에 모집단의 평균이 포함될 가능성이 더 높아짐
- 범위가 넓을수록 신뢰구간이 갖는 정보의 가치 하락

Chapter 9 통계적 추정

◎ 모집단 평균의 구간추정

신뢰구간식의 2가지 가정

- 모집단이 정규분포를 따른다 → 모집단이 정규분포가 아니라면? A. 표본의 크기 n 을 충분히 키워서 **중심극한정리** 활용
- 모집단의 표준편차 **σ**를 알고 있다. → 모른다면 아래의 방법 참조
- 위의 두 가정을 성립한다면, **표본의 크기와 관계없이**
  $\boldsymbol{P}(-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2})$ **= 1 – α 성립**

모집단이 정규분포가 아니라면?

- 표본에서 구한 불편추정량 $\boldsymbol{S}$를 모집단의 표준편차 대신으로 활용

$$S = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$$

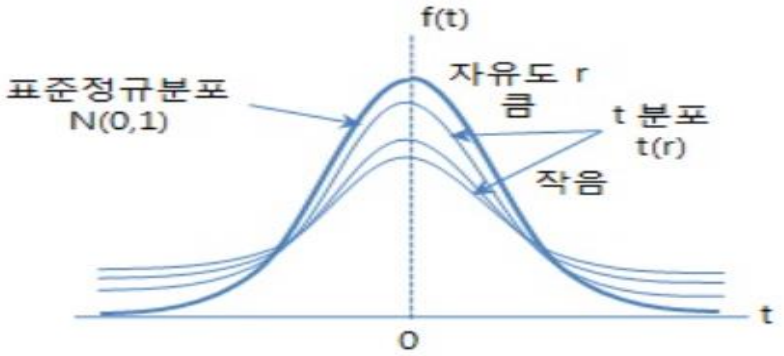
- 표본의 크기 n이 작고 또한 **σ**를 모르고 $\boldsymbol{S}$ 만 알 때에는 Z분포$(\bar{X} - \boldsymbol{\mu}_{\bar{X}}) / \boldsymbol{\sigma}/\sqrt{\boldsymbol{n}}$를 통해 신뢰구간을 계산할 수 없음
- 이런 상황에서는 $\boldsymbol{\sigma}_{\bar{X}} = \boldsymbol{\sigma}/\sqrt{\boldsymbol{n}}$ 대신 $\boldsymbol{S}_{\bar{X}} = \boldsymbol{S}/\sqrt{\boldsymbol{n}}$을 사용해야 하는데, 표본통계량 $((\bar{X} - \boldsymbol{\mu}_{\bar{X}}) / \boldsymbol{S}/\sqrt{\boldsymbol{n}})$은 표준정규분포를 따르지 않고 **자유도 $(\boldsymbol{n-1})$의 $\boldsymbol{t}$-분포**를 이루기 때문에 $\boldsymbol{t}$**-분포**를 이용하여 신뢰구간을 구해야 함

Chapter 9 통계적 추정

◎ 모집단 평균의 구간추정

t-통계량 공식

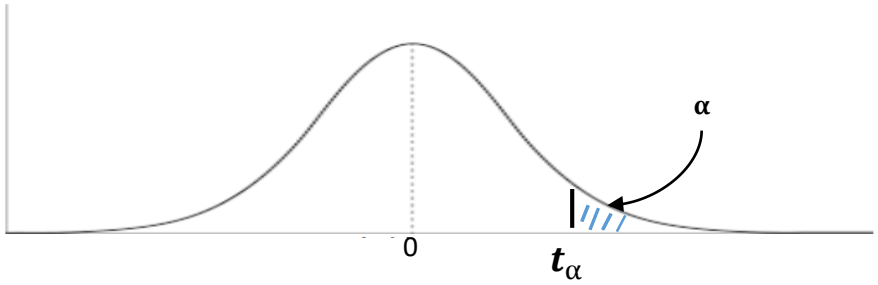
$$t = \frac{\bar{X} - \mu_{\bar{X}}}{s_{\bar{X}}}$$



- 표준정규분포와 유사하게, 0을 중심으로 좌우대칭
- 표준정규분포보다 평평하고 기다란 꼬리를 갖음(양쪽 꼬리가 두터움)
  즉, 표준정규분포보다 분산이 크므로 보다 평평한 모양
- 자유도에 따라 다른 모양을 나타냄 (χ^2 분포 도 이와 유사함)
  자유도(=표본의 수 n - 1)가 증가할수록, **표준정규분포**에 가까워짐
  대개, 자유다가 30 이 넘으면 표준정규분포와 비슷하게 됨

Chapter 9 통계적 추정

◎ 모집단 평균의 구간추정



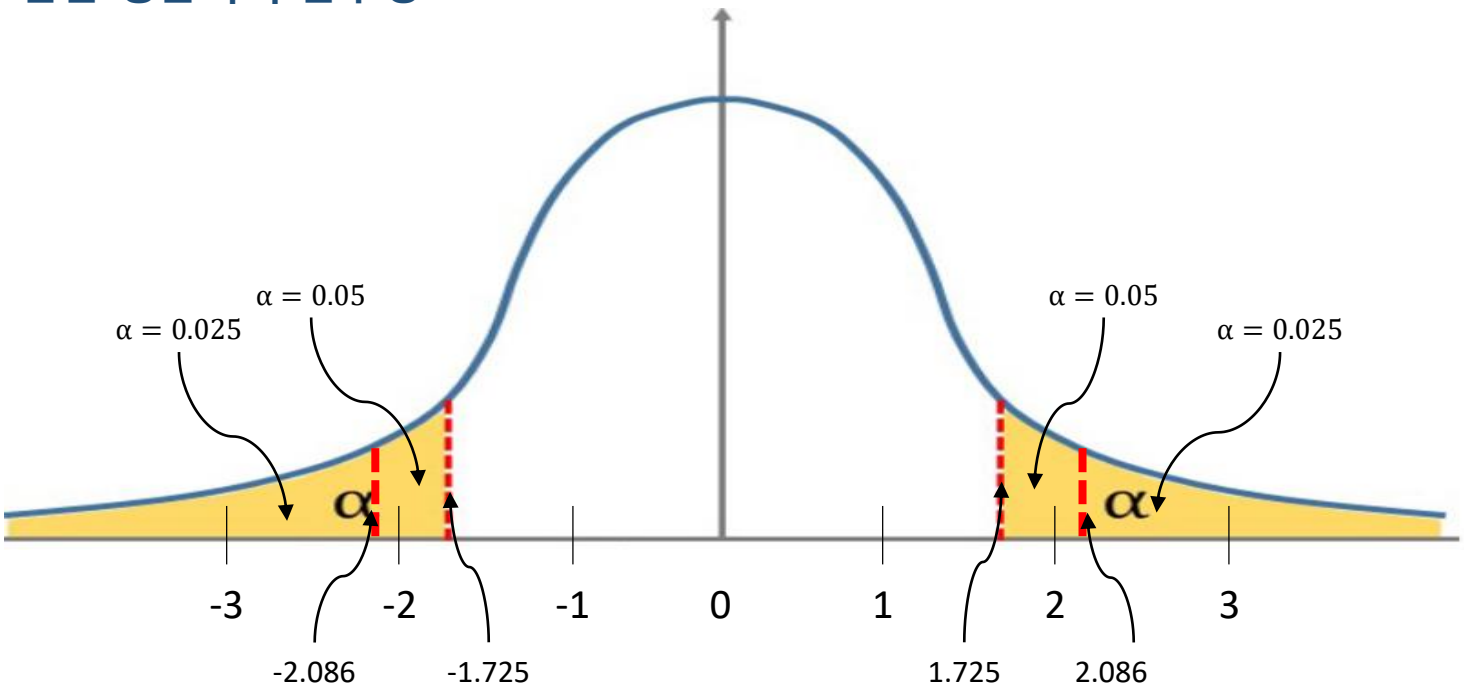
$P(t \geq t_\alpha) = \alpha$는 위와 같이 표현할 수 있다. 어떤 t값 이상되는 면적을 α라 하고, 그 α에 맞는 t값, 즉 t_α를 나타낸 것이t-분포표이다.

t-분포표를 보면 df = 20인 상태에서 α=0.05와 α=0.025에 해당되는 t값이 각각 1.725와 2.086인 것을 확인할 수 있다. 이는 df 가 20일 때, t가 1.725보다 클 확률이 0.05이며, t가 2.086 보다 클 확률은 0.025임을 의미한다.

$$P(t \geq 1.725) = P(t \leq -1.725) = 0.05$$
$$P(t \geq 2.086) = P(t \leq -2.086) = 0.025$$

Chapter 9 통계적 추정

◎ 모집단 평균의 구간추정



- **α**=0.05, **α**=0.025 일 때의 임계치(Critical Value) ::
  1.725, 2.086 (***t***분포)
  1.64, 1.96(Z분포)

- ***t***분포가 **추정구간을 넓게 하는 효과**를 가져 **모집단 표준편차를 모르는데 보상**으로 작용

Chapter 9 통계적 추정

◎ t분포를 이용한 신뢰구간 추정

$$P(-t_{\alpha/2} \leq \frac{\bar{X}-\mu_{\bar{X}}}{S_{\bar{X}}} \leq t_{\alpha/2}) = 1 - \alpha$$

- $\pm t_{\alpha/2}$는 자유도가 $n-1$ 인 t분포에서 분포 양 끝에 $\alpha/2$씩의 확률을 갖는 t값을 의미
- 평균 μ에 대하여 정리하면 t-분포에서 모집단 평균 μ의 신뢰구간의 일반식은 아래와 같이 정의 할 수 있음

t-분포에서의 신뢰구간

$$P(\bar{X} - t_{\alpha/2} * S_{\bar{X}} \leq \mu \leq \bar{X} + t_{\alpha/2} * S_{\bar{X}}) = 1 - \alpha$$

Chapter 9 통계적 추정

◎ t분포를 이용한 신뢰구간 추정

예제 9-3

어느 A학교 학생 2,000명의 평균 통학거리를 알아보려 한다. 모든 학생을 대상으로 조사하는 것은 어려워 16명을 임의추출 하여 조사한 결과 그들 평균 통학거리는 1,640m이며, 표본에서 계산된 표준편차는 2,000m였다. 통학거리의 분포가 정규분포라면, 이 학교 학생들의 평균 통학거리는 얼마인지 구하시오.

- 이 예에서 $\bar{X}$ = 1,640, $\boldsymbol{S} = \mathbf{2,000}$ 이며 $\boldsymbol{n} = \mathbf{16}$ 이므로, $\boldsymbol{df} = \mathbf{15}$ 가 된다. $\boldsymbol{df} = \mathbf{15}$ 일 때 90% 신뢰도를 만족시키는 t값은 [표]에서 보면 $t = \pm 1.753$이므로 아래와 같이 표현이 가능하다.

$$\bar{X} - t_{\alpha/2} * \boldsymbol{S}_{\bar{X}} \leq \boldsymbol{\mu} \leq \bar{X} + t_{\alpha/2} * \boldsymbol{S}_{\bar{X}}$$

$$\leftrightarrow 1{,}640 - 1.753 * \frac{2{,}000}{\sqrt{16}} \leq \mu \leq 1{,}640 + 1.753 * \frac{2{,}000}{\sqrt{16}}$$

$$\leftrightarrow 1{,}640 - 876 \leq \mu \leq 1{,}640 + 876$$

$$\leftrightarrow 764 \leq \mu \leq 2516$$

따라서, $\boldsymbol{\mu}$ 에 대한 90% 신뢰구간은 [764, 2,516]

Chapter 9 통계적 추정

◎ 표본이 큰 경우의 신뢰구간

예제 9-4

자전거를 생산하는 트윈산업에서 하루 평균 자전거생산량을 조사하려 한다. 과거 121일간의 하루 생산량의 평균과 표준편차를 계산하였더니 $\bar{X}$ = 500대, $\boldsymbol{S}$ = **110**대였다. 평균 생산량 $\boldsymbol{\mu}$에 대한 90% 신뢰구간을 구하시오.

1) t분포로 신뢰구간을 결정하는 경우,
$\alpha/2 = 0.05$, df = 120에서 t값은 1.658이므로
신뢰구간 식은 아래와 같이 쓰인다

$$\bar{X} - t_{0.05} * \boldsymbol{S}_{\bar{X}} \leq \boldsymbol{\mu} \leq \bar{X} + t_{0.05} * \boldsymbol{S}_{\bar{X}}$$

$$\leftrightarrow 500 - 1.658 * \frac{110}{\sqrt{121}} \leq \mu \leq 500 + 1.658 * \frac{110}{\sqrt{121}}$$

$$\leftrightarrow 500 - 16.58 \leq \mu \leq 500 + 16.58$$

$$\leftrightarrow 483.42 \leq \mu \leq 516.58$$

2) Z분포로 신뢰구간을 결정하는 경우,
$Z_{0.05}$ = 1.645이므로 신뢰구간은 다음과 같다

$$\bar{X} - Z_{0.05} * \boldsymbol{S}_{\bar{X}} \leq \boldsymbol{\mu} \leq \bar{X} + Z_{0.05} * \boldsymbol{S}_{\bar{X}}$$

$$\leftrightarrow 500 - 1.645 * \frac{110}{\sqrt{121}} \leq \mu \leq 500 + 1.645 * \frac{110}{\sqrt{121}}$$

$$\leftrightarrow 500 - 16.45 \leq \mu \leq 500 + 16.45$$

$$\leftrightarrow 483.55 \leq \mu \leq 516.45$$

- 결론 t값이나 Z값으로 구한 90% 신뢰구간은 매우 근사하다
- 즉, 모집단의 표준편차 "σ" 를 모르는 경우라도 표본의 크기가 큰 경우에는 정규분포를 이용하여 신뢰구간을 추정할 수 있다

Chapter 9 통계적 추정

◎ 표본크기의 결정

모집단 평균을 추정할 때 표본크기의 결정

모집단의 평균 $\boldsymbol{\mu}$를 추정할 때 신뢰구간의 양 끝점을 나타내는 다음 식을 살펴보면 표본크기의 역할을 뚜렷이 알 수 있음

$$\text{신뢰구간의 양 끝점} : \bar{X} \pm Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$$

또는

$$\text{신뢰구간의 양 끝점} : \bar{X} \pm t_{\alpha/2} * \frac{s}{\sqrt{n}}$$

σ 나 s는 연구자가 임의로 조정하기가 힘들거나 불가능할 때도 있으며 신뢰도 α 또한 어느 수준 이상 확보해야 하므로, 신뢰구간의 폭은 n에 달려 있다고 할 수 있다

표본크기 n에 따라서 신뢰구간의 폭이 어떻게 달라지는지를 보자

◎ 표본크기의 결정

모집단 평균을 추정할 때 표본크기의 결정

예제 9-5

모집단이 정규분포를 이루며, 그 모집단의 표준편차가 400이라고 가정하자.
표본의 크기가 4인경우 $\bar{X}$**가 600**이었을 때의 **90%신뢰구간**,
표본의 크기가 16인경우 $\bar{X}$**가 600**이었을 때의 **90%신뢰구간**,
표본의 크기가 100인경우 $\bar{X}$**가 600**이었을 때의 **90%신뢰구간**,
을 각각 비교하시오

풀이:: 모집단의 표준편차를 알고 있으므로 $\bar{X} \pm Z_{\alpha/2} * \sigma_{\bar{X}}$ 식을 활용하면 90%의 신뢰구간은,

n= 4일 때 $600 \pm 1.64 * \frac{400}{\sqrt{4}} = 600 \pm 328$

n= 16일 때 $600 \pm 1.64 * \frac{400}{\sqrt{16}} = 600 \pm 164$

n= 100일 때 $600 \pm 1.64 * \frac{400}{\sqrt{100}} = 600 \pm 65.6$

Chapter 9 통계적 추정

◎ 표본크기의 결정

* $Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} = e$ :: 신뢰구간의 한쪽 폭을 오차 e라고 함

표본크기의 결정

$$\sigma^2\text{을 알 때} \quad n = \frac{Z^2_{\alpha/2} \times \sigma^2}{e^2}$$

$$\sigma^2\text{을 모를 때} \quad n = \frac{t^2_{\alpha/2} \times \sigma^2}{e^2}$$

- 표본크기와 관련있는 요인 3가지

1) 신 뢰 구 간: 오차의 크기와 표준편차가 정해졌을 때, 신뢰구간을 크게 할수록 표본의 크기를 크게 해야 한다.
2) 표 준 편 차: 오차의 크기와 신뢰구간이 정해졌을 때, 표준편차 또는 분산이 클수록 표본의 크기도 커야 한다.
3) 오차의 크기: 신뢰구간과 표준편차가 정해졌을 때, 오차를 작게 하기를 원하면 표본의 크기는 커야 한다.

Chapter 9 통계적 추정

◎ 모집단 평균을 추정할 때 표본크기의 결정

예제 9-6

어느 참치회사에서는 16g 용량의 참치통조림을 생산하고 있다. 실제 무게가 그러한지를 조사하기 위해 표본을 추출하려 한다. 모집단 평균 무게에 대한 추정값의 모집단 평균 무게에 대한 추정 오차가 0.2g 이상이 되지 않을 것을 원하며, 결과는 99% 신뢰도를 갖기를 원한다. 지금까지의 데이터로 σ = 1.34라는 것을 확인하였을 때, 표본의 크기를 얼마로 해야 해당 요구가 충족되겠는가?

99% 신뢰도를 충족시키는 Z값은 2.57이며, σ = 1.34, e =0.2 이므로

$$n = \left[\frac{Z^2 \times \sigma^2}{e^2} \right] = \left[\frac{2.57 \times 1.34}{0.2} \right]^2 = 296.5$$

따라서 표본크기로 **297개 이상**을 선택함으로써 가능

Chapter 9 통계적 추정

◎ 모집단 비율을 추정할 때 표본크기의 결정

- 모집단 평균을 추정하기 위해 필요한 표본의 크기를 정할 때와 마찬가지로 모집단 비율의 신뢰구간을 추정할 때에도 **일정한 신뢰도에서 허용가능한 오차를 가져올 최소한의 표본크기를** 계산할 수 있음

- 비율의 표집분포는 이항분포를 이루지만 $\boldsymbol{n}$이 비교적 클 때에는 정규분포에 접근하므로, 정규분포의 분석방법을 사용하면 필요한 표본의 크기가 쉽게 결정됨

- 허용오차 $\boldsymbol{e}$ 라고 표시하면

$$\boldsymbol{e} = |p - \pi|$$

그런데 $\boldsymbol{Z} = \frac{p-\pi}{\sigma_p}$ 이므로 $p - \pi = Z * \sigma_p$ 가 되며 , 이를 위의 식에 대입한다면?

$$\boldsymbol{e} = |Z * \sigma_p|$$

로 표시할 수 있으며, 이 때 $\sigma_p = \sqrt{\pi\,(1-\pi)\,/n}$

- π가 미지수라는 것을 의미하므로 $\pi\,(1-\pi)\,/n$ 의 값을 계산할 수 없음

- $\boldsymbol{\sigma_p}$의 추정 값으로 표본비율의 표준편차 $S_p = \sqrt{p\,(1-p)\,/n}$을 이용할 수도 없음 :: 그러므로 이 경우 σ_p가 최대인 값, 즉, π = 0.5일 때의 $\boldsymbol{\sigma_p}$의 값으로 n을 정하는 것이 안전

$$\boldsymbol{e} = \left| Z * \sqrt{\frac{\pi\,(1-\pi)}{n}} \right| \rightarrow \boldsymbol{e} = \left| Z * \sqrt{\frac{0.5 * 0.5}{n}} \right| \rightarrow n = \frac{\boldsymbol{Z^2}}{\boldsymbol{4e^2}}$$

End of This Chapter!

주요 이력

現) ㈜W사 Recommendation System, Time Series etc) ing
前) ㈜RTMC The Head of Strategic Planning Department
前) ㈜biz사 Web Analysis & Data Voucher Business
前) H금속 FX, Amortization & Depreciation Planning
前) B건설 Mgmt Performance Report & Footnote
前) K문고 CRM VIP Clustering Strategy
前) L백화점 CRM Alert Strategy

학력

BSL(Business School of Lausanne) Big Data MBA
ASSIST Big Data statistic MBA

現) 대학교 및 관계기관 Big-Data 다수 강의
現) 코리아IT아카데미 Big-Data R
現) 코리아IT아카데미 Big-Data Python
現) 코리아IT아카데미 Big-Data Principles of Statistics