Module3

확률분포와 통계적 검정

◆ 학습목표

다양한 확률분포에 대해 학습하고, 추정과 가설검정 방법에 대해 학습한다.

I. 표본추출과 표본오차

II. 확률분포에 대한 이해

III. 추정과 가설검정

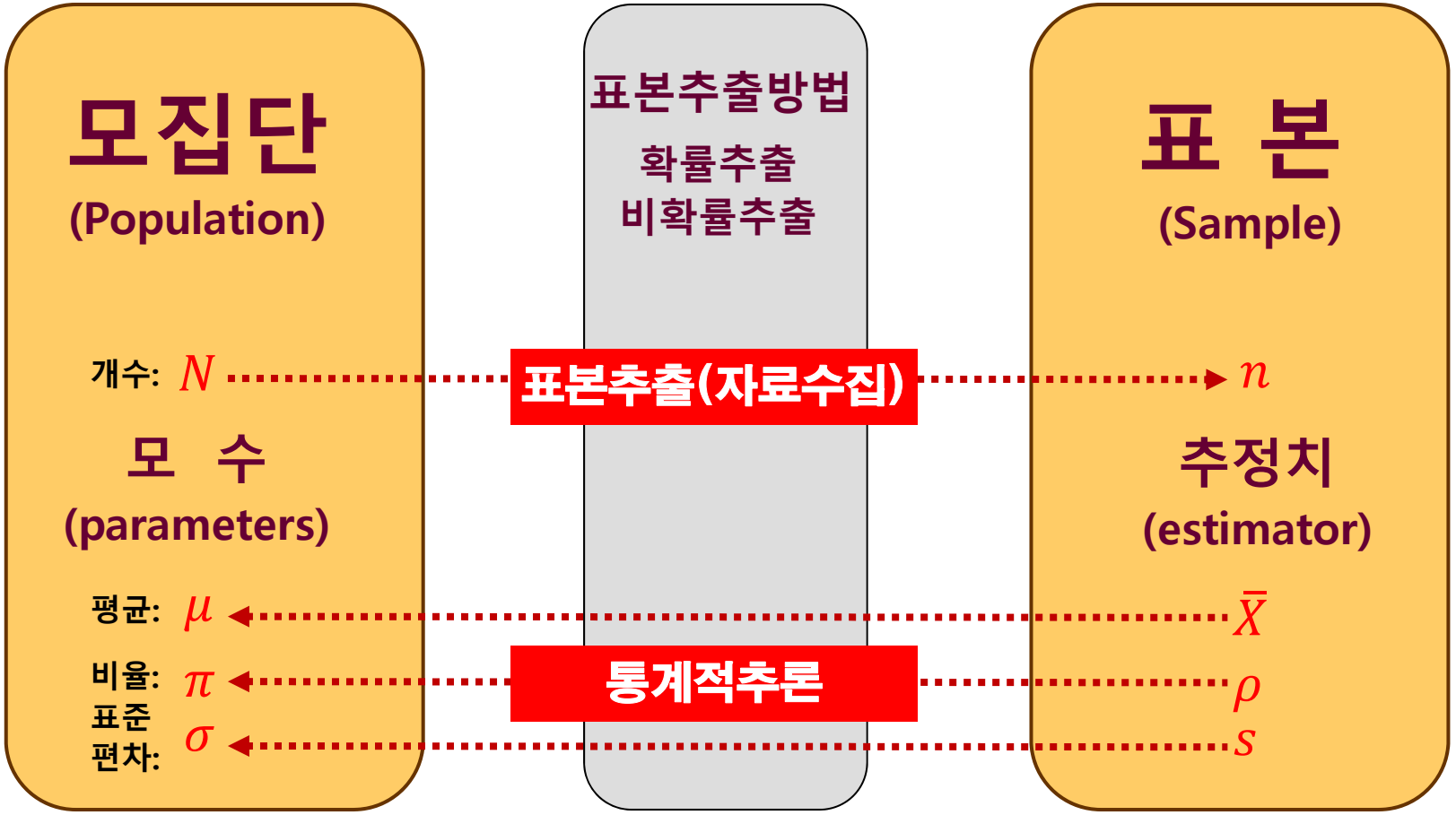
IV. 확인적(CDA) 분석방법

I. 표본추출과 표본오차

모집단과 표본

❖ 통계학(statistics)이란

– 표본의 자료를 이용하여 모집단의 특성에 대해 확률을 이용해 추론



모집단과 전수조사

❖ 모집단(population)

- 관심 있는 연구대상 전체의 집합
- 무한모집단: 모집단의 크기가 무한한 경우 (전세계 인구, 자판기 커피)
- 유한모집단: 모집단의 크기가 유한한 경우 (강서대학교 재학생)

❖ 전수조사

- 관심 있는 모집단 전체를 조사하는 경우로서 주로 모집단의 규모가 작을 경우에 실시

❖ 전수조사의 어려움

- 조사불가능: 모집단 전체를 대상으로 조사하기는 불가능
- 시간과 비용: 모집단을 다 조사하는 데는 많은 시간과 비용 소요

❖ 해결책: 전수조사 → 표본조사

표본과 표본조사

❖ 표본(sample)

- 실제로 조사 및 측정되는 모집단의 일부

❖ 표본조사

- 모집단에서 추출된 일부분인 표본을 가지고 하는 조사

❖ 수집방법

- 실험(experiment)
- 조사(survey)
- 출판 자료 (Published data)

모수와 통계량

❖ 모수(parameter):

- 모집단에 대한 수치 특성값
- 모집단의 특성을 나타내는 양적인 측도로서 주어진 모집단을 따르는 고유의 상수 (상수=모집단은 진실된 하나의 값임)
- 모평균(μ), 모표준편차(σ)
- 예) 우리나라 고등학교 사교육비 평균

❖ 통계량(statistic):

- 표본에서 얻은 수치 특성값
- 표본의 특성을 나타내는 양적인 측도로서 모집단의 분포를 따르는 확률변수(확률변수=표본에 따라 값이 변함)
- 표본평균($\bar{X}$), 표본표준편차(s)
- 예) 우리나라 고등학교 1학년 중에서 1000명만 뽑아 조사하여 얻은 평균 사교육비

표본오차와 통계적 추론

❖ 표본오차(sampling error)

- 모집단에서 표본을 추출해서 조사하기 때문에 모수와 표본 통계량 사이에 생기는 오차
- 표본의 크기를 크게 함으로써 표본오차를 감소 → 통계학에서 표본의 크기를 크게 하라는 이유
- 표본오차는 아무리 표본을 크게 해도 전수조사를 하지 않는 이상 존재
- 표본오차의 허용범위를 확률로 구하는 것이 통계의 목적

❖ 통계적 추론

- 우리가 실제로 알고 싶은 것은 표본의 값(통계량: statistic)이 아니고 모집단의 값(모수: Parameter)
- 통계학의 목적: 추론(Inference) → 표본에서 구한 값을 이용해 우리가 구하고자 하는 모집단의 값도 이럴 것이다라고 추론

표본조사의 중요성

❖ 1936년 미국 대통령 선거

- Landon(공화당) : Roosevelt(민주당)

❖ Literary Digest

- 조사대상자 : 구독자, 전화기, 자동차 보유자 1,000만 명에게 엽서
- 응답자: 237만 명
- Landon(57%) : Roosevelt(43%)

❖ Gallup

- 조사대상자 : 50,000명
- 응답자: 1,500명 (할당추출)
- Landon(44%) : Roosevelt(56%)

❖ 최종결과

- Landon(37%) : Roosevelt(63%)

출처: http://www.hani.co.kr/arti/PRINT/280849.html

표본조사 방법

❖ 확률추출(probability sampling)

- 모집단에 속하는 모든 추출단위에 대해 사전에 일정한 추출확률이 주어지는 표본추출법
- 표본추출틀 존재
- 단순확률추출(simple random sampling)
- 계통추출법(systematic sampling)
- 층화확률추출(stratified random sampling)
- 집락추출법(cluster sampling)

❖ 비확률추출

- 추출단위가 표본에 추출될 확률을 객관적으로 나타낼 수 없는 표본추출법
- 편의표출(convenience sampling)
- 할당추출(Quota sampling)
- 포커스 그룹(Focus Groups)

표본조사 방법

❖ 기본용어

- 기본단위(elementary unity) : 조사의 대상이 되는 가장 최소의 요소
- 예)여론조사 : 개인, 가계조사 : 가구, 농작물조사: 일정 면적의 경지

- 추출틀(sampling frame) : 모집단에 속하는 모든 추출단위의 목록
- 예)개인, 가구, 사업체 등의 명부, 문서철, 지도 등

- 목표모집단(target population): 관심을 갖고 특성을 알아보고자 하는 집단에 속하는 모든 기본단위들의 집합
- 예) LG전자 회원

- 조사모집단(target population): 표본 추출틀을 통해 추출될 수 있는 기본단위들의 집합(실제 조사 가능한 집단)
- 예) 전화조사 : LG전자 회원 중 전화번호가 있는 사람

확률추출

❖ 단순확률추출(simple random sampling)

- Random의 의미

  '아무렇게나' → 추출하는 사람의 주관을 일체 배제시키는 것

  모집단을 구성하는 요소 하나하나가 뽑힐 확률이 동일한 상황에서 뽑는 방법

- 조건

  모집단 전체에 대한 추출틀(sample frame)이 있어야 함

  난수표를 이용한 샘플링

- 사례)

  LG전자에 등록된 고객 1,000만 명의 명단을 이용해서 10,000명을 난수로 추출
  (1만명/1,000명)

  강서대학 재학생 1,200명 중에서 100명을 난수를 이용해 추출(100/1,200)

  입장전에 행운권을 나누어주고 무작위로 행운권을 추첨(1/n)

확률추출

❖ 계통추출법(systematic sampling)

- 모집단의 추출틀에서 k번째 간격마다 하나씩 표본으로 추출
- 처음 추출은 난수표로 한 개의 표본을 추출하고, 그 난수표에 일정숫자를 더해서 표본추출: 7, 17, 27, 37,...
- 표본추출의 단위가 클 때 이용
- 사례1) 형광등 불량품을 조사하기 위해 5,000개의 slot에서 20개를 샘플로 추출하려고 할 때

  난수표를 이용해 숫자 하나를 선택: 7

  20개 샘플을 뽑으려면: 5,000/20 = 250 (확률: 1/250)

  7, 257, 507,757,...,4757 (20개)

- 모집단이 무한모집단일 경우
- 사례2) K 레스토랑 손님 만족도를 조사하기 위해 10, 20, ...번째 손님을 추출
- 사례3) 대통령 선거 시 사전조사를 위해 선거구 출구에서 매 20번째 나오는 사람을 대상으로 조사

확률추출

❖ 층화확률추출(stratified random sampling)

- 모집단을 먼저 서로 겹치지 않는 여러 개의 층으로 분할한 후, 각 층별로 단순임의추출법을 적용시켜 표본을 얻는 방법
- 층화: 표본을 추출하는 과정이 아니고 추출을 위해 모집단을 몇 개의 부분군으로 나누는 작업

  예) 직역적 특성, 성별, 연령대 등
- 사례1)

  서울시장 후보에 대한 선호도를 조사하기 위해 1,000명 조사할 때,

  강서구 인구비율이 10% → 강서구에서 100명 표본추출
- 사례2)

  K 대학교 학생들의 주당 평균 학습시간을 계산하기 위해서 단과대학별, 학년별, 성별 등을 고려하여 학생을 추출

  경영학과 30% → 1학년 30% → 남자 40%

  K대학 학생 1,000명 중에서 100명을 샘플로 뽑을 경우

  경영학과 1학년 남자: 100명*0.3=30명*0.3=9명*0.4=4명 추출

확률추출

❖ 집락추출법(cluster sampling)

- 서로 인접한 기본단위들로 구성된 군집을 만들고, 추출된 군집 내의 일부 또는 전체를 조사
- 조사단위가 산재되어 있어서 조사비용이 많이 들 때
- 사례) 서울에 있는 대학생 월 평균 용돈 추정

  모집단: 서울에 있는 56개 대학에 다니는 대학생 전체

  서울에 있는 대학교 리스트를 이용해 무작위로 몇 개 대학 추출(PSU: primary sampling unit)

  계열 → 학과 → 학년 → 대학생 추출
- 표본 추출틀을 모를 경우

비확률추출법

❖ 편의표출(convenience sampling)

- 추출틀 없이 자발적인 참여, 인터넷 조사
- 연구자가 이용 가능한 대상을 임의대로 선택

❖ 할당법(Quota sampling)

- 확률적 근거없이 임의로 연구자가 표본을 분류하여 추출
- 예) 성별에 따라 다를 것 같은데, 나이에 따라 다를 것 같은데(연구자 추측)

❖ 포커스 그룹(Focus Groups)

- 깊이 있는 연구를 위해 대표하는 사람을 추출 (panel)
- K전자는 새로운 휴대폰을 개발하기 위해 소비자 10명을 추출하여 새로운 상품에 대한 아이디어 회의를 진행

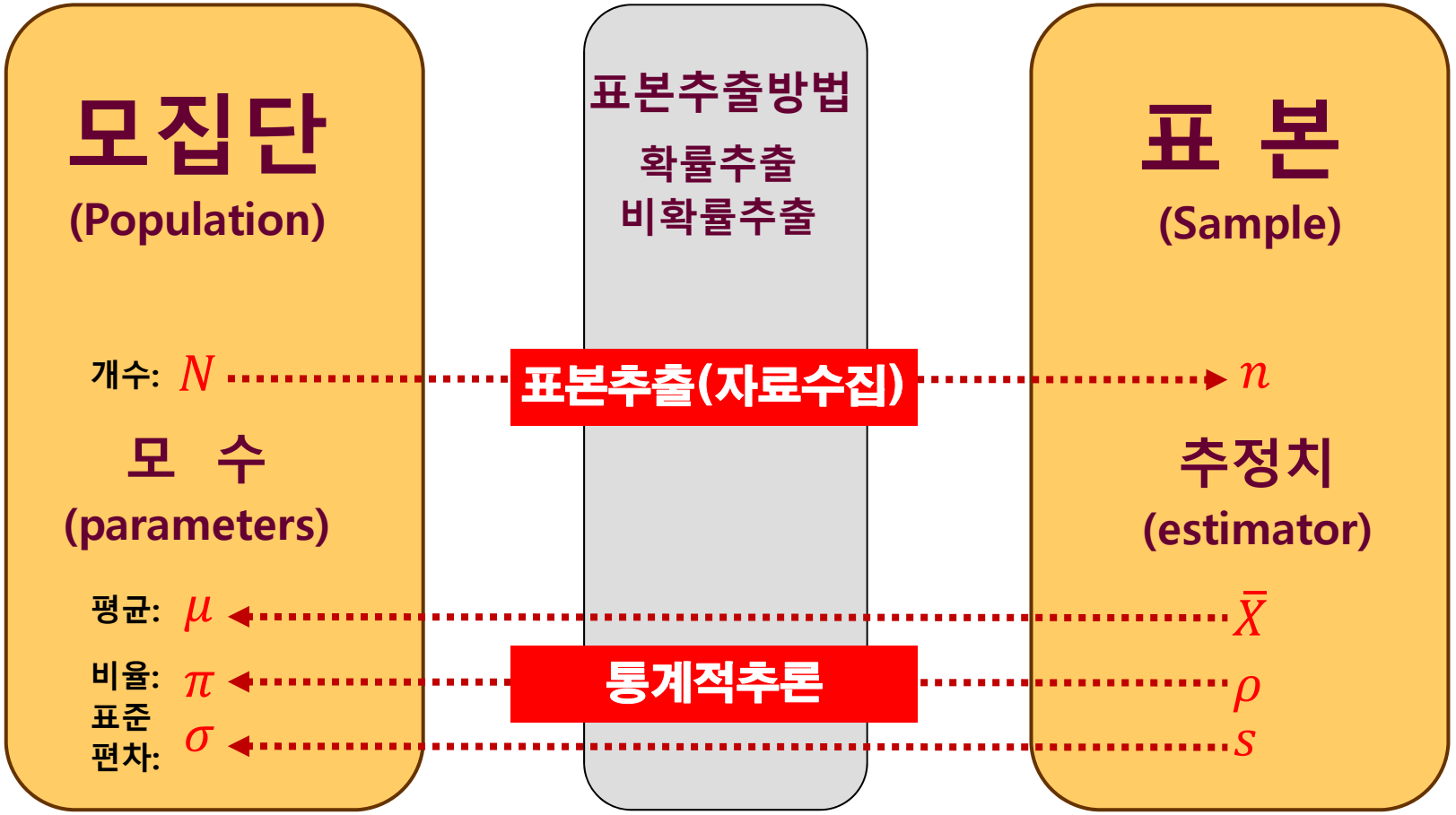
II. 확률분포에 대한 이해

통계는 항상 확률과 같이 배우는데,
왜 통계에서 확률이 중요한가요?

통계학과 확률

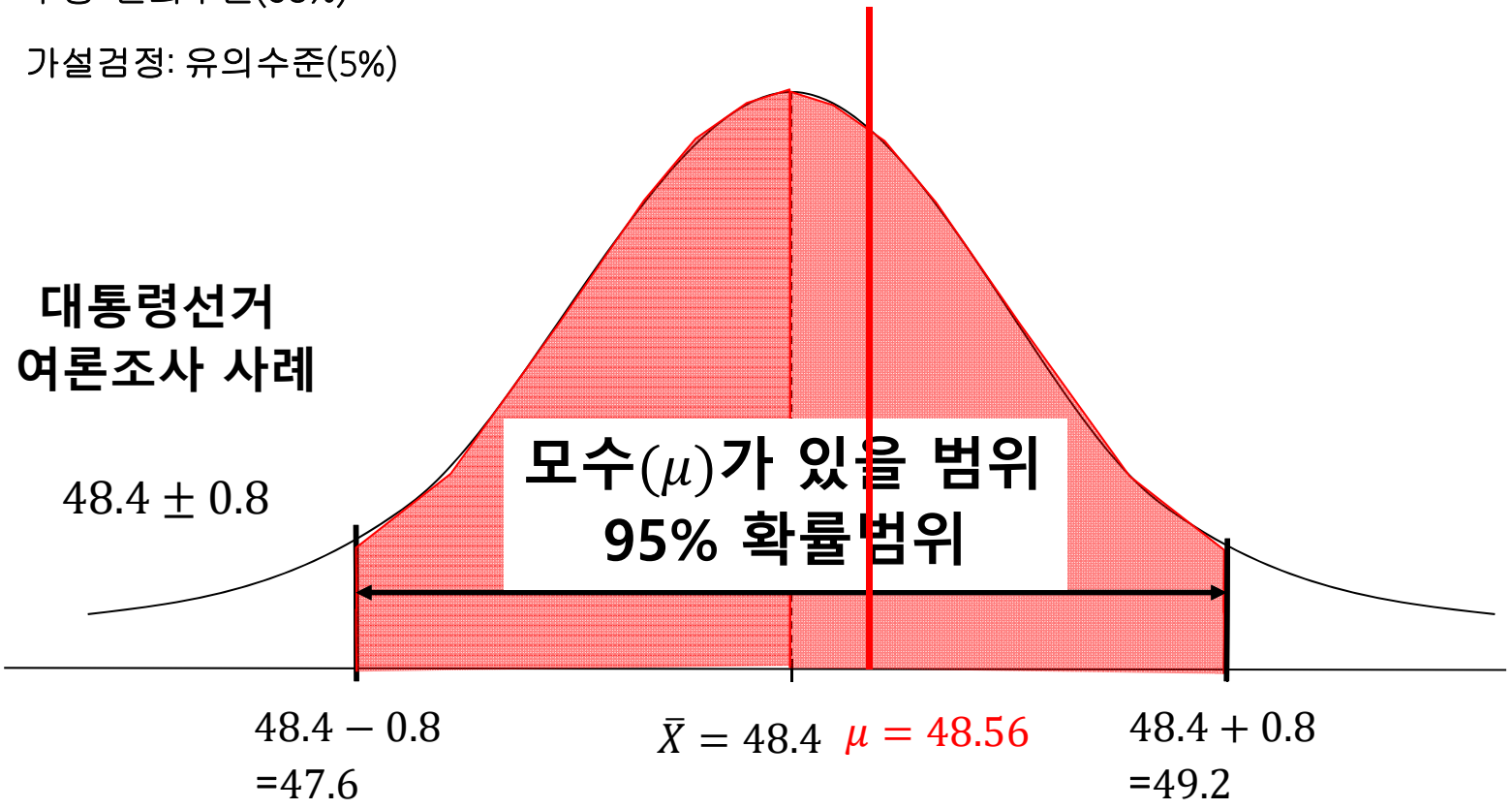
❖ 통계학(statistics)이란

– 표본의 자료를 이용하여 모집단의 특성에 대해 확률을 이용해 추론



확률과 오차범위

- 통계적 방법론의 기초이론으로써 중요한 역할을 함
- 확률을 이용하여 모수를 추측
- 모수가 있을 범위
  - 추정: 신뢰수준(95%)
  - 가설검정: 유의수준(5%)



확률의 개념

❖ 확률의 개념

- 확률 = 가능성 = %
- 비슷한 현상이 반복해서 일어날 경우에 어떤 사건이 발생할 가능성을 0과 1사이의 숫자로 표현한 것

❖ 확률의 계산

$$P(A) = \frac{n_A}{n_S} = \frac{A\text{의 개수}}{S\text{의 개수}} = \frac{\text{관심사건}}{\text{전체(표본공간)}} = \%$$

- S : 표본공간 (sample space) = 전체 개수
- A : 사건(event) 또는 사상 = 관심있는 부분

확률변수

❖ 확률변수(random variable)

- 확률: random or probability
- 표본공간에 있는 각 원소에 값을 대응시켜 주는 규칙 또는 함수
- 표본공간에 있는 값을 숫자로 변경
- 확률변수 = X, 확률변수의 값 = x
- 예) X =동전 2개를 던질 때 동전 앞면의 수
  표본공간: $S = \{HH, HT, TH, TT\}$ → 확률변수: $X = \{2,1,1,0\}$

❖ 종류:

- 이산확률변수(discrete random variable)

  특정한 수치만을 가지는 확률변수(정수)

  불량품의 수, 고속도로에서 사고건수, 방문자수 등

- 연속확률변수(continuous random variable)

  어떤 범위에서 연속적인 값을 치할 수 있는 확률변수(실수)

  전구의 수명, 몸무게, 체온, 통근시간 등

확률분포

❖ 확률분포(Probability Distributions)

- 확률변수X의 각 값(x)에 대응하는 확률(0~1)을 표시
- 이산확률분포(Discrete Probability Distribution)

  일양분포, 이항분포, 포아송분포, 초기하분포, 기하분포
- 연속확률분포(Continuous Probability Distribution)

  평균분포(정규분포, t-분포), 분산분포(x^2분포, f분포)
- 규칙

$$0 \leq P(x_i) \leq 1 \qquad \sum_{i}^{n} P(x_i) = 1$$

- 확률분포 = 표본의 분포(X) → 모집단의 형태(확률구조)를 나타냄

❖ 확률분포표

- 확률변수 X에 대한 확률을 표형태로 만든 것

이산확률분포

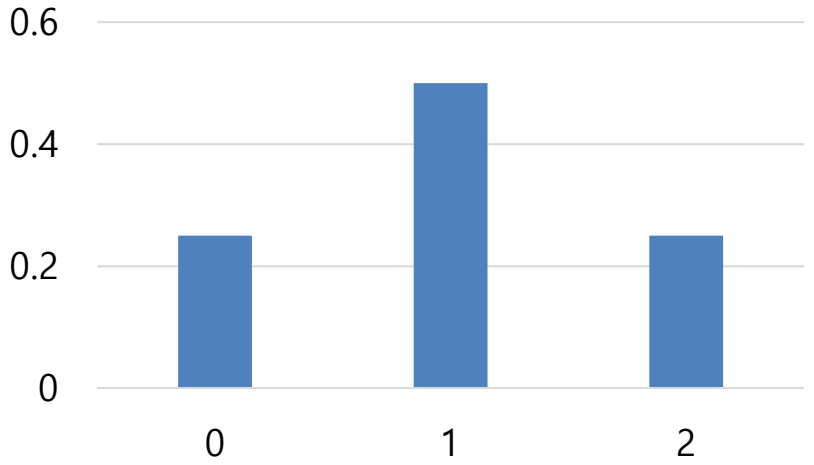
❖ 이산확률변수

- 표본공간 : {(T,T), (T,H), (H,T),(H,H)}
- 확률변수 X=동전 2개를 던질 때 동전 앞면의 수

❖ 이산확률분포

사건	X	P(X)
{T, T}	0	1/4
{H T}, {T,H)	1	2/4
{H H}	2	1/4
합계		4/4

x_i $P(x_i)$



이산확률분포 계산

❖ K대학 이교수는 통계학 수업을 들은 100명을 대상으로 결석일 수를 조사하였다.

0	1	1	1	2	2	0	2	1	2
0	3	1	1	1	0	1	0	1	0
0	4	3	1	0	0	1	1	2	1
0	1	0	3	0	0	2	0	1	2
0	0	1	1	1	1	3	2	0	1
1	2	1	2	4	3	1	0	2	2
0	2	1	1	1	1	1	1	1	0
0	2	0	0	0	0	1	0	1	0
4	0	1	0	0	0	0	1	1	1
1	3	2	1	0	1	0	0	1	3

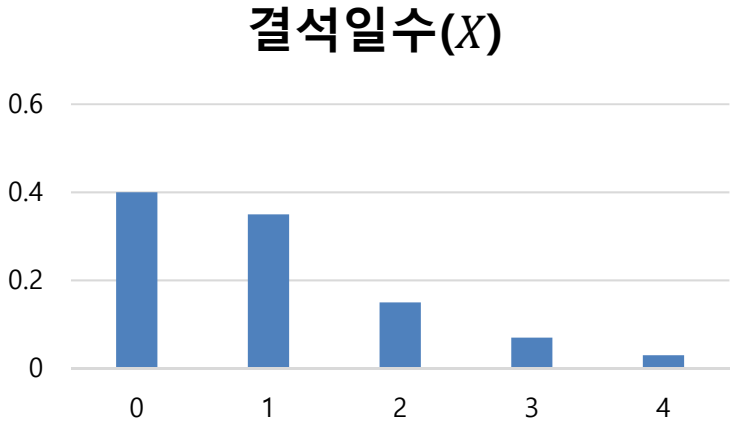
$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{100}(0 + 1 + \cdots + 3) = 0.98$$

이산확률분포 계산

❖ 이산확률분포

- 확률변수 : X = 결석일수
- 확률분포

결석일수(X)	학생수	P(X)
0	40	0.40
1	35	0.35
2	15	0.15
3	7	0.07
4	3	0.03
	100	1



- 확률분포는 도수분포(=Frequency Table)의 %
- 확률은 도수분포표의 상대도수

이산확률분포 계산

❖ 이산확률분포 계산

- 통계학 수업을 듣는 학생이 1회 결석할 확률은?

$$P(x = 1) = 0.35$$

- 통계학 수업을 듣는 학생이 3회 이상 결석할 확률은?

$$P(x \geq 3) = P(x = 3) + P(x = 3) = 0.07 + 0.03 = 0.10$$

- 통계학 수업을 듣는 학생이 1회 이하 결석할 확률은?

$$P(x \leq 1) = P(x = 0) + P(x = 1) = 0.40 + 0.35 = 0.75$$

결석일수 (X)	$P(X)$
0	0.40
1	0.35
2	0.15
3	0.07
4	0.03
	1

기대값

❖ 기대값(Expected Value)

- 확률변수 X의 평균 또는 중심위치
- 가중평균의 개념
- 기대값 = 평균

$$E(X) = \sum_{i=1}^{n} x_i P(x_i)$$
$$= (0 \times 0.40) + \cdots + (4 \times 0.03)$$
$$= 0.98$$

||

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{100}(0 + 1 + \cdots + 3) = 0.98$$

결석일수(X)	학생수	$P(X)$	$xP(X)$
0	40	0.40	0
1	35	0.35	0.35
2	15	0.15	0.30
3	7	0.07	0.21
4	3	0.03	0.12
	100	1	0.98

기대값

❖ 28살인 이길동씨는 K 화재보험에 자동차 보험을 들고 있다. 대물보상 2,000만원에 대해 매월 1만원을 내고 있을 때 누가 이득인가?

- 대물대상 보험가액: 20,000,000
- 20대 사고확률: 0.002(년)

$$E(X) = 0 \times 0.998 + 20{,}000{,}000 \times 0.002$$
$$= 40{,}000$$
$$\text{회사수익(년)} = 120{,}000 - 40{,}000 = 80{,}000$$

보험액(X)	$P(X)$	$xP(X)$
0	0.998	0
20,000,000	0.002	40,000
	1	40,000

❖ 만약 K 화재보험사의 경우 인건비, 사고처리비, 수익으로 30%를 책정하였다. 얼마를 보험료로 책정해야 하는가?

$$\text{보험료} = 40{,}000 + (40{,}000 \times 0.30) = 52{,}000$$

기대값

❖ 로또의 기대값

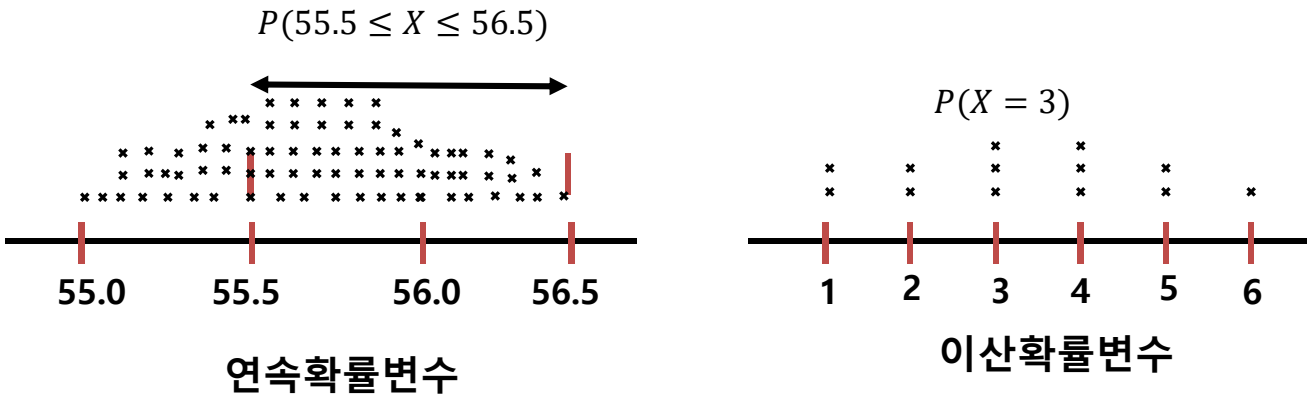
등위	당첨내용	당첨확률	당첨금 배분 비율	당첨금액	기대값(xP(X))
1등	6개 번호 일치	1 / 8,145,060	총 당첨금 중 4등과 5등 금액을 제외한 금액의 75%	2,075,050,516	255
2등	5개 번호 + 보너스 번호일치	1 / 1,357,510	총 당첨금 중 4등과 5등 금액을 제외한 금액의 12.5%	59,510,656	44
3등	5개 번호 일치	1 / 35,724	총 당첨금 중 4등과 5등 금액을 제외한 금액의 12.5%	1,550,680	43
4등	4개 번호 일치	1 / 733	50,000원	50,000	68
5등	3개 번호 일치	1 / 45	5,000원	5,000	111
미당첨		1-당첨확률 (97.64%)	0원	0	0
					521

$$\text{이익} = 521 - 1{,}000 = -479\text{원}$$

연속확률분포

❖ 연속확률변수(continuous random variable)

- 어떤 범위에서 연속적인 값을 취할 수 있는 확률변수(실수)
- 전구의 수명, 몸무게, 체온, 통근시간 등



❖ 연속확률분포(Continuous Probability Distribution)

- 연속확률변수의 값에 대응하는 확률을 표시
- 정규분포, 표준정규분포, 지수분포

연속확률분포

❖ G대학 경영통계 수강생의 몸무게 분석

- 몸무게가 55-60일 확률은?
- 몸무게가 55-65일 확률은?
- 몸무게 평균을 중심으로 95%확률로 예측할 수 있는 몸무게의 범위는?

40	56	57	56	62	60	57	65	61	50
50	72	55	48	59	59	49	67	47	55
56	46	67	50	71	53	54	57	48	56
51	63	52	68	58	47	53	62	54	62
55	56	55	53	41	62	58	53	52	46
61	52	62	58	58	56	58	58	56	49
69	48	67	62	48	59	59	60	59	60
44	57	49	57	69	50	60	51	59	63
66	52	49	68	53	61	61	54	68	69
60	45	52	54	57	64	61	59	57	66

(단위: kg)

연속확률분포

❖ K대학 통계학 수업 수강생 100명 몸무게를 조사

- 평균, 표준편차

$\mu = 56.78$

$\sigma = 6.80$

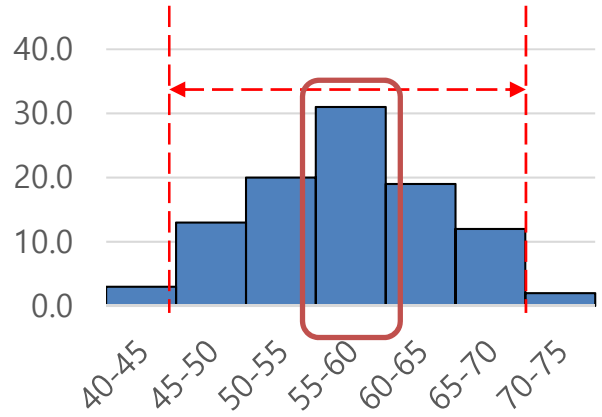
- 몸무게가 55-60일 확률은?

$P(55 \leq X \leq 60) = 0.31$

- 몸무게 평균을 중심으로 95%확률로 예측할 수 있는 몸무게의 범위는?

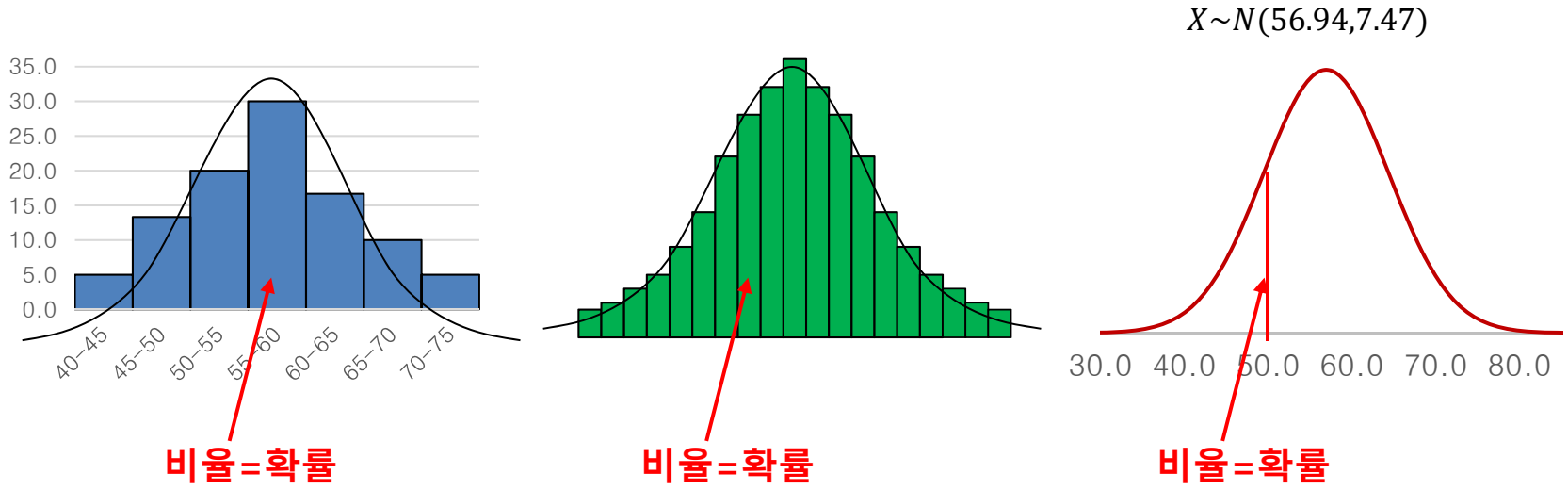
$0.95 = P(45 \leq X \leq 70)$

X	빈도수	%	확률(p)
40-45	3	3	3
45-50	13	13	13
50-55	20	20	20
55-60	31	31	31
60-65	19	19	19
65-70	12	12	12
70-75	2	2	2
합계	100	1	1



연속확률분포

❖ 관측횟수(표본)를 무한히 늘리면 정규분포



Uniform Distribution

Uniform Distribution

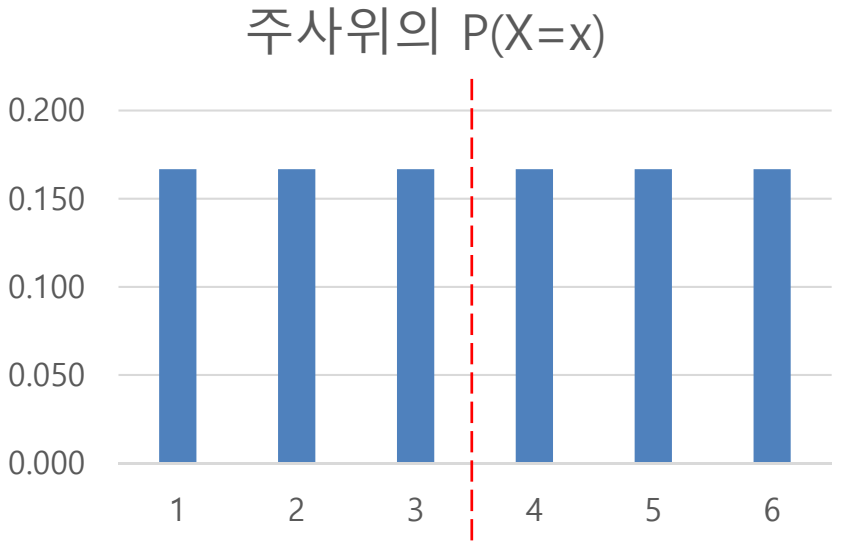
❖ 일양분포 (Uniform Distribution)

– 확률변수의 값이 특정한 두 수 사이에서 일정한 확률값을 가질때

$$\mathrm{P}(X = x) = \frac{1}{b - a + 1}$$

$$\mu = E(X) = \frac{a + b}{2}$$

$$\sigma = \sqrt{\frac{[(b - a) + 1]^2 - 1}{12}}$$



– 주사위의 확률값

$$\mathrm{P}(X = x) = \frac{1}{b - a + 1} = \frac{1}{6 - 1 + 1} = \frac{1}{6}$$

$$\mu = E(X) = \frac{a + b}{2} = \frac{1 + 6}{2} = 3.5$$

$$\sigma = \sqrt{\frac{[(b - a) + 1]^2 - 1}{12}} = \sqrt{\frac{[(6 - 1) + 1]^2 - 1}{12}} = \sqrt{\frac{36 - 1}{12}} = 1.708$$

Uniform Distribution

❖ 로또의 숫자는 랜덤한가?

– 로또 당첨번호 합계 (1097회까지)

x	당첨횟수	$P(x_i)$	$x_iP(x_i)$	$(x_i-\mu)^2$	$(x_i-\mu)^2P(x_i)$
1	152	0.02	0.02	486.91	11.24
2	142	0.02	0.04	443.78	9.57
3	147	0.02	0.07	402.65	8.99
4	147	0.02	0.09	363.52	8.12
...					
43	153	0.02	1.00	397.36	9.24
44	146	0.02	0.98	438.23	9.72
45	159	0.02	1.09	481.10	11.62
		1.00	23.07		169.2

$$\mu = 22.945 \qquad \sigma = \sqrt{169.2} = 13.009$$

출처: https://www.lotto.co.kr/article/list/AC01

Uniform Distribution

❖ 로또의 숫자는 랜덤한가?

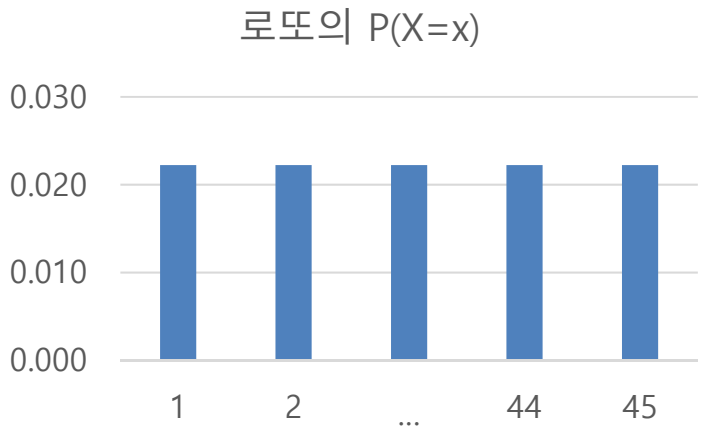
– 만약 모든 숫자가 랜덤하게 나온다면 아래와 같이 계산

$$P(X = x) = \frac{1}{45 - 1 + 1} = \frac{1}{45}$$

$$\mu = E(X) = \frac{1 + 45}{2} = 23.0 \qquad \approx \mu = 22.946$$

$$\sigma = \sqrt{\frac{(45)^2 - 1}{12}}$$

$$= \sqrt{\frac{2025 - 1}{12}} = 12.99 \qquad \approx \sigma = 13.009$$



– 로또 번호는 랜덤한가? 아니면 특정번호가 선택되는가?

Binominal Distribution

Binominal Distribution

❖ 베르누이 시행 (Bernoulli Experiments)

- 결과가 두 가지 이며, 각각의 결과가 서로 독립적인 시행
- 각 시행의 결과가 두 가지의 결과만이 가능한 시행
- 모든 실험에서 결과의 확률은 항상 동일
- 예)제품을 검사하며 불량품과 양호품으로 구분하는 경우
- 유권자에게 정부정책에 대한 찬성과 반대를 묻는 경우

$$X = \begin{cases} 1, \text{성공} \\ 0, \text{실패} \end{cases}$$

$$P(X = 1) = \pi$$

$$P(X = 0) = 1 - \pi$$

Binominal Distribution

❖ 베르누이 시행

- 기대값과 분산

$$E(X) = \sum_{i=1}^{2} x_i P(x_i) = (0)(1-\pi) + (1)(\pi) = \pi$$

$$V(X) = \sum_{i=1}^{2} \big(x_i - E(X)\big)^2 P(x_i) = \pi(1-\pi)$$

- K전자에서 생산하는 제품의 불량률이 0.01이라고 했을 때, 제품의 불량률의 기대값과 분산은?

$$E(X) = (0)(0.99) + (1)(0.01) = 0.01$$

$$V(X) = 0.01(1-0.01) = 0.0099$$

Binominal Distribution

❖ 이항분포 (Binominal Distribution)

- 성공률이 π인 베르누이 시행을 n번 시행했을 때의 확률분포
- 예)100개의 제품을 불량품과 양호품으로 구분하는 경우
- 1,000명의 유권자에게 정부정책에 대한 찬성과 반대를 묻는 경우

$X =$ 성공확률변수

$\pi =$ 1회시행시 성공확률

$n =$ 시행회수

$$E(X) = (\pi + \pi + \cdots + \pi) = n\pi$$

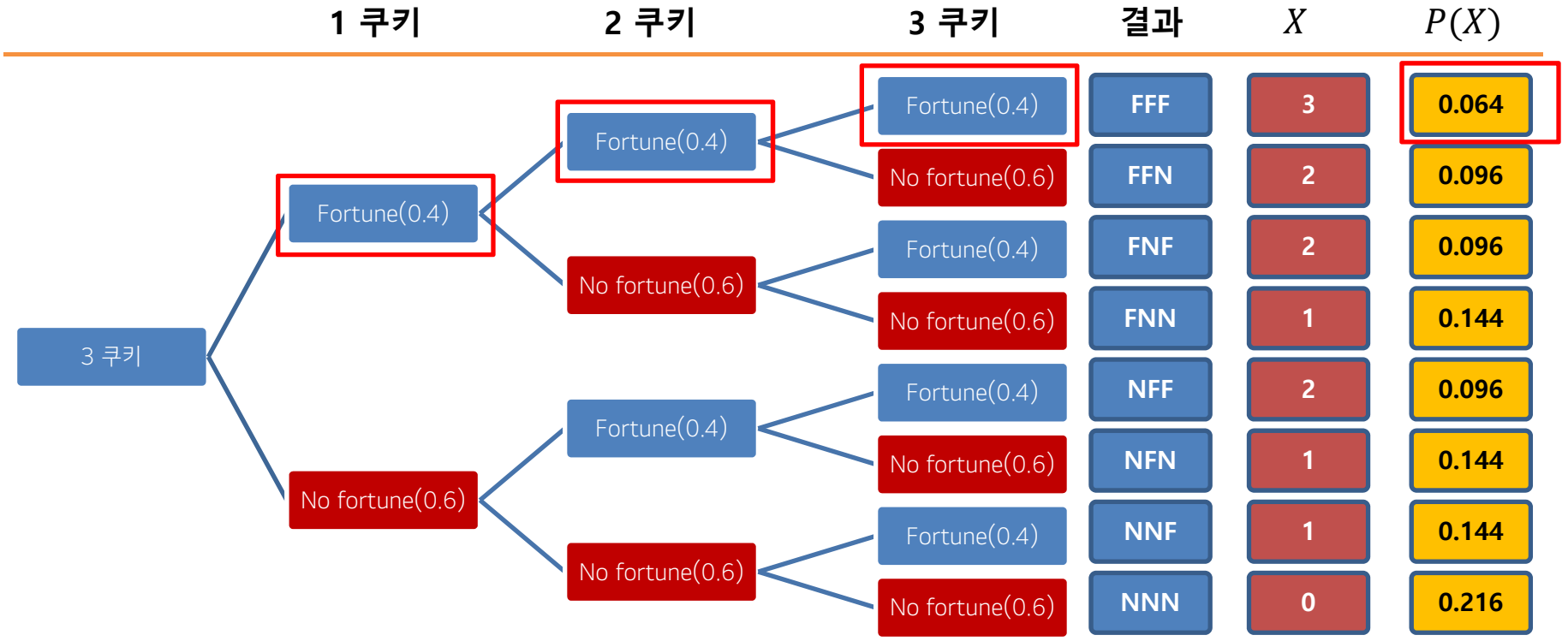
$$V(X) = \pi(1-\pi) + \pi(1-\pi) + \cdots \pi(1-\pi) = n\pi(1-\pi)$$

$$P(X) = \frac{n!}{x!\,(n-x)!}\pi^x(1-\pi)^{n-x}$$

Binominal Distribution

❖ K식당에서는 손님에게 fortune 쿠키를 제공한다. 성공율은 40%이다. 3명의 손님이 앉은 테이블에서 3명이 모두 못 받을 확률은?

확률트리



Binominal Distribution

❖ 이산확률분포로 정리하면

- 평균 몇 명이 받을 수 있나?

$$E(X) = \mu = \sum_{i=1}^{n} x_i P(x_i) = 1.2$$

X	P(X)	x*P(X)
0	0.216	0.000
1	0.432	0.432
2	0.288	0.576
3	0.064	0.192
		1.2

- 한 명도 못 받을 확률

$$P(X = 0) = 0.261$$

- 세 명 다 받을 확률

$$P(X = 3) = 0.064$$

- 한 명 이상 받을 확률

$$P(X \geq 1) = P(X = 1) + P(X = 2) + P(X = 3) = 0.784$$

Binominal Distribution

❖ 이항분포로 정리하면 (베르누이 시행이므로)

- 평균 당첨횟수

$X = forturn$ 받은 횟수
$\pi = 0.4$
$n = 3$

$$E(X) = (0.4 + 0.4 + 0.4) = 3(0.4) = 1.2$$

- 분산

$$V(X) = n\pi(1-\pi) = 3(0.4)(0.6) = 0.72$$

- 한 명도 못 받을 확률

$$P(0) = \frac{n!}{x!\,(n-x)!}\pi^x(1-\pi)^{n-x} = \frac{3!}{0!\,(3-0)!}0.4^0(1-0.4)^{3-0} = \frac{6}{1\times 6}1(0.6)^3$$

- 세 명 다 받을 확률?

어려워요.TT
→
통계학자들이 정리해준 표를 이용해 봅시다.^^

Binominal Distribution

❖ 성공률 40%일때, 3명의 손님이 앉은 테이블에서 3명이 모두 못 받을 확률은?

1. 이항분포표

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

		π									
n	x	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

Binominal Distribution

❖ 10명 중 8명 이상이 fortune 쿠키를 받을 확률은?

- 이산확률분포를 직접 만드는 방법
- 이항분포식을 이용하는 방법
- 이항분포표(통계학자들이 정리)를 이용하는 방법
- 엑셀 등 SW 함수 이용방법

		π									
n	x	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
10	0	0.5987	0.3487	0.1969	0.1074	0.0563	0.0282	0.0135	0.0060	0.0025	0.0010
	1	0.3151	0.3874	0.3474	0.2684	0.1877	0.1211	0.0725	0.0403	0.0207	0.0098
	2	0.0746	0.1937	0.2759	0.3020	0.2816	0.2335	0.1757	0.1209	0.0763	0.0439
	3	0.0105	0.0574	0.1298	0.2013	0.2503	0.2668	0.2522	0.2150	0.1665	0.1172
	4	0.0010	0.0112	0.0401	0.0881	0.1460	0.2001	0.2377	0.2508	0.2384	0.2051
	5	0.0001	0.0015	0.0085	0.0264	0.0584	0.1029	0.1536	0.2007	0.2340	0.2461
	6	0.0000	0.0001	0.0012	0.0055	0.0162	0.0368	0.0689	0.1115	0.1596	0.2051
	7	0.0000	0.0000	0.0001	0.0008	0.0031	0.0090	0.0212	0.0425	0.0746	0.1172
	8	0.0000						0.0043	0.0106	0.0229	0.0439
	9	0.0000						0.0005	0.0016	0.0042	0.0098
	10	0.0000						0.0000	0.0001	0.0003	0.0010

어려워요.TT
→
통계 package를 이용해 봐요.^^

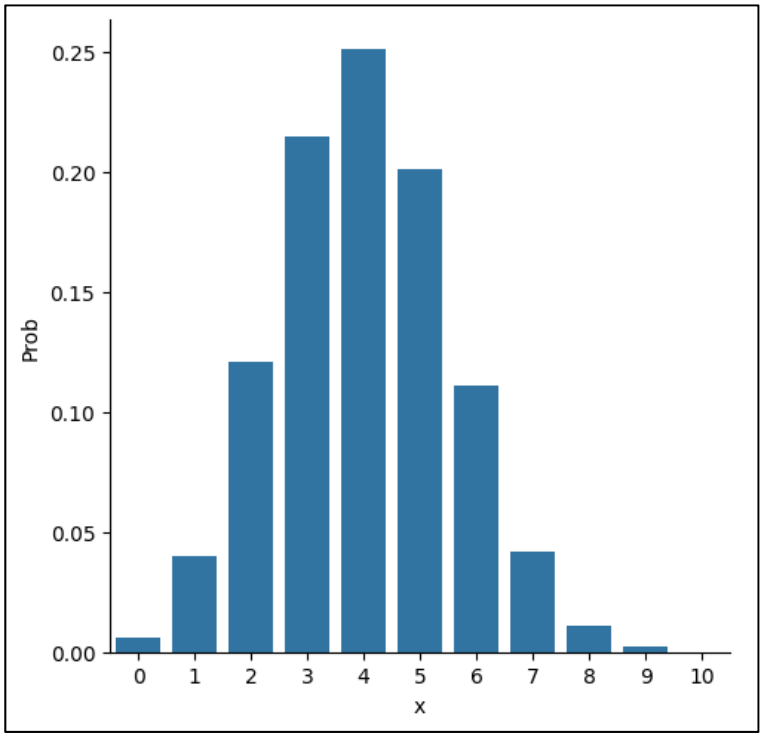
Binominal Distribution

❖ 이항분포 그래프(r, python)

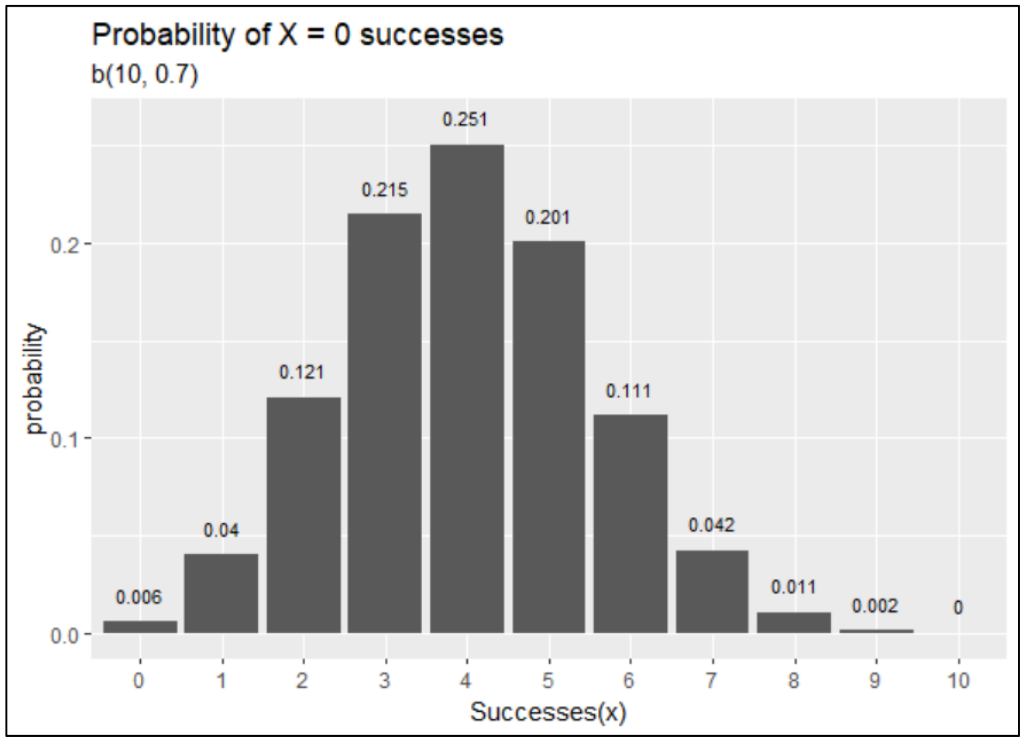
– 성공률에 따라 다른 분포

$$\pi = 0.7$$
$$n = 10$$

〔Python〕



〔R〕

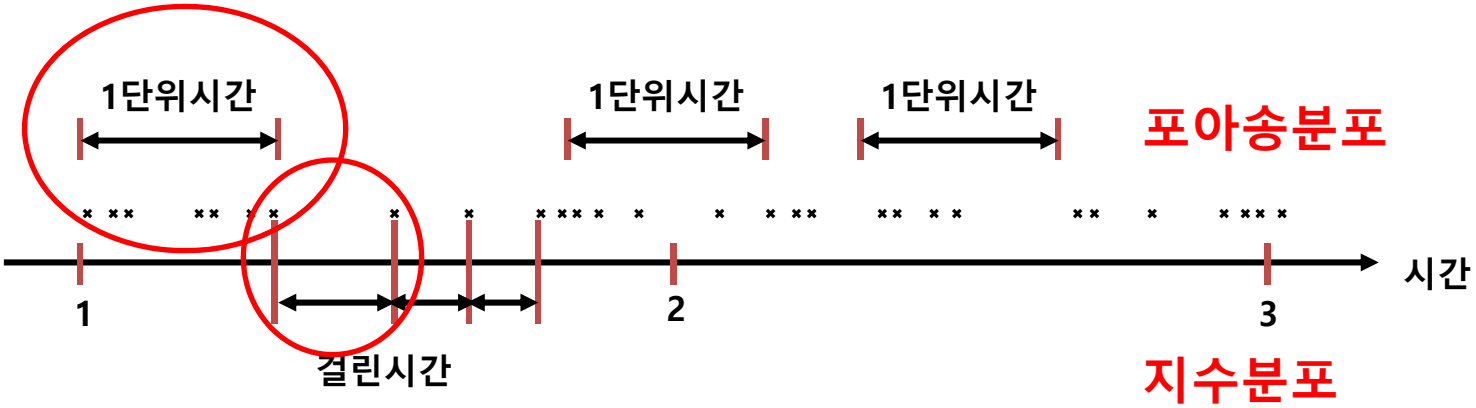


Poisson Distribution

Poisson Distribution

❖ 포아송분포(Poisson distribution)

- 램덤하게 선택한 일정한 단위 시간(시, 분, 초)이나 공간 (1평, 30cm^2 등) 내에 발생하는 사건의 개수를 설명
- 보통 단위시간당 도착(arrivals)에 대한 모델에 많이 사용되므로 시간이 주로 사용



- 반대로 도착에 따른 시간을 측정하기 위해서는 연속분포인 지수분포(Exponential Distribution)을 사용함
- 경영학에서는 대기시간 모형에서 많이 사용

Poisson Distribution

❖ 수식

– 단위당 평균 발생건이 λ인 현상에 대한 확률분포

$X =$ 단위시간당 발생건수

$\lambda =$ 단위시간당 평균발생 건수

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, e = 2.71828$$

$$E(X) = \mu = \lambda$$

$$V(X) = \sigma^2 = \lambda$$

Poisson Distribution

❖ G서비스 센터는 10분에 평균 1회의 전화가 온다. 10분 동안에 2회의 전화를 받을 확률은? 2회 이상 전화 받을 확률은

- 이산확률분포 이용

x	P(X=x)	P(X<=x)
0	0.368	0.368
1	0.368	0.736
2	0.184	0.920
3	0.061	0.981
4	0.015	0.996
5	0.003	0.999
6	0.001	1.000
7	0.000	1.000

- 포아송분포 수식 이용

$$P(X = 2) = \frac{1.0^2 e^{-1.0}}{2!} = 0.184$$

- 2회 이상 전화 받을 확률

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

Poisson Distribution

❖ 포아송 분포표

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

	λ									
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	0.9048	0.8187	0.7408	0.9703	0.6065	0.5488	0.4966	0.4493	0.4066	0.3679
1	0.0905	0.1637	0.2222	0.2681	0.3033	0.3293	0.3476	0.3595	0.3659	0.3679
2	0.0045	0.0164	0.0333	0.0536	0.0758	0.0988	0.1217	0.1438	0.1647	0.1839
3	0.0002	0.0011	0.0033	0.0072	0.0126	0.0198	0.0284	0.0383	0.0494	0.0613
4	0.0000	0.0001	0.0003	0.0007	0.0016	0.0030	0.0050	0.0077	0.0111	0.0153
5	0.0000	0.0000	0.0000	0.0001	0.0002	0.0004	0.0007	0.0012	0.0020	0.0031
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003	0.0005
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001

	λ									
x	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	0.3329	0.3012	0.2725	0.2466	0.2231	0.2019	0.1827	0.1653	0.1496	0.1353
1	0.3662	0.3614	0.3543	0.3452	0.3347	0.3230	0.3106	0.2975	0.2842	0.2707
2	0.2014	0.2169	0.2303	0.2417	0.2510	0.2584	0.2640	0.2678	0.2700	0.2707
3	0.0738	0.0867	0.0998	0.1168	0.1255	0.1378	0.1496	0.1607	0.1710	0.1804
4	0.0203	0.0260	0.0324	0.0395	0.0471	0.0551	0.0636	0.0723	0.0812	0.0902
5	0.0045	0.0062	0.0084	0.0111	0.0141	0.0176	0.0216	0.0260	0.0309	0.0361
6	0.0008	0.0012	0.0018	0.0026	0.0035	0.0047	0.0061	0.0078	0.0098	0.0120

Poisson Distribution

❖ G서비스 센터는 10분에 평균 1회의 전화가 온다. 10분 동안에 2회의 전화를 받을 확률은? 2회 이상 전화 받을 확률은

$$P(X = 2) = 0.3679$$

$$P(X \geq 2) = 1 - P(X \leq 1)$$
$$= 1 - (0.3679 + 0.3679)$$
$$= 0.264$$

$$P(1 \leq X \leq 2) = 0.552$$

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

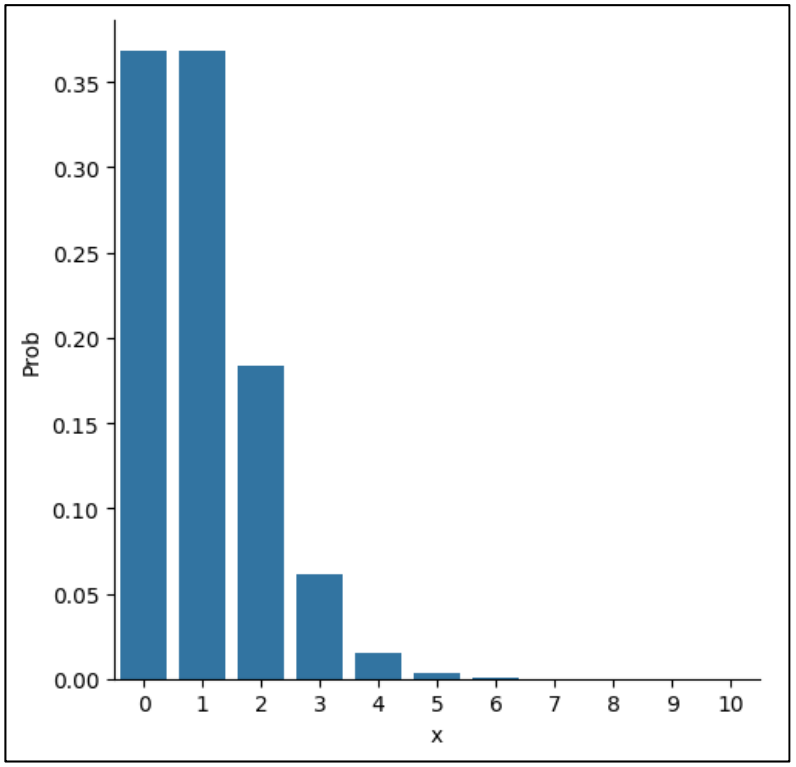
	λ									
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	0.9048	0.8187	0.7408	0.9703	0.6065	0.5488	0.4966	0.4493	0.4066	0.3679
1	0.0905	0.1637	0.2222	0.2681	0.3033	0.3293	0.3476	0.3595	0.3659	0.3679
2	0.0045	0.0164	0.0333	0.0536	0.0758	0.0988	0.1217	0.1438	0.1647	0.1839
3	0.0002	0.0011	0.0033	0.0072	0.0126	0.0198	0.0284	0.0383	0.0494	0.0613
4	0.0000	0.0001	0.0003	0.0007	0.0016	0.0030	0.0050	0.0077	0.0111	0.0153
5	0.0000	0.0000	0.0000	0.0001	0.0002	0.0004	0.0007	0.0012	0.0020	0.0031
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003	0.0005
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001

	λ									
x	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	0.3329	0.3012	0.2725	0.2466	0.2231	0.2019	0.1827	0.1653	0.1496	0.1353
1	0.3662	0.3614	0.3543	0.3452	0.3347	0.3230	0.3106	0.2975	0.2842	0.2707
2	0.2014	0.2169	0.2303	0.2417	0.2510	0.2584	0.2640	0.2678	0.2700	0.2707
3	0.0738	0.0867	0.0998	0.1168	0.1255	0.1378	0.1496	0.1607	0.1710	0.1804

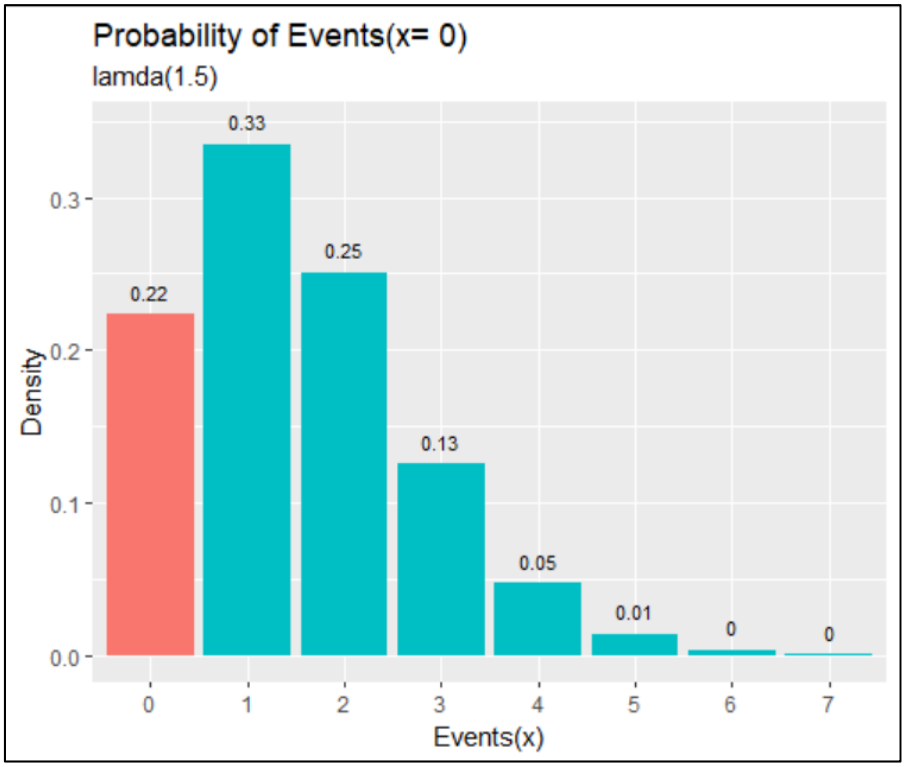
Poisson Distribution

❖ 포아송분포 그래프(R)

〔Python〕



〔R〕



연속확률분포

❖ 종류

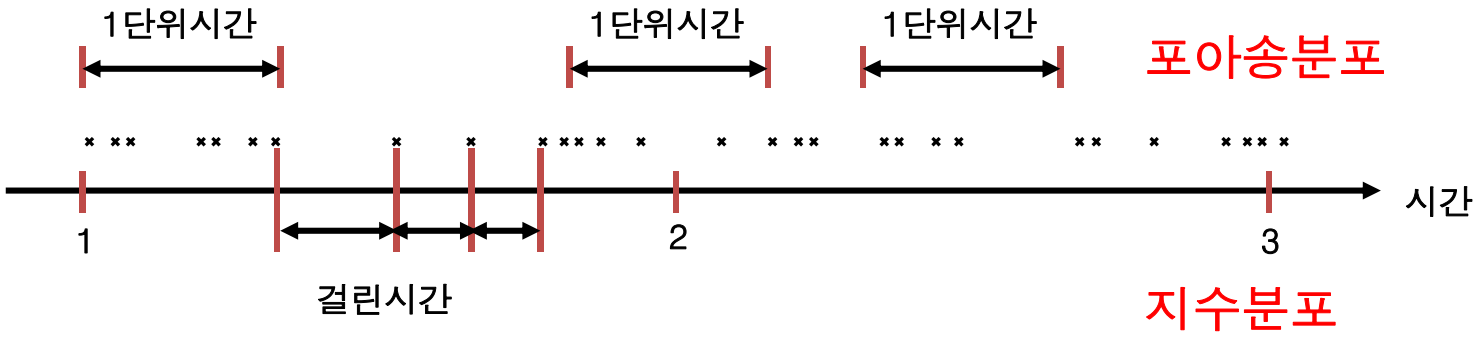
- 지수분포(Exponential Distribution)
- 정규분포(Normal Distribution)
- 표쥰정규분포(Standard Normal Distribution)
- T분포(t-Distribution)
- F분포(F-Distribution)
- χ^2분포(Chi square-Distribution)

Exponential Distribution

Exponential Distribution

❖ 지수분포 (Exponential Distribution)

- 연속확률분포
- 포아송 분포가 단위시간당 사건의 개수라면 지수분포는 두 사건 사이의 시간에 대한 확률



- 수식

$$\mu = \frac{1}{\lambda} \qquad \sigma = \frac{1}{\lambda}$$

$$P(X \le x) = 1 - e^{-\lambda x} \qquad P(X > x) = e^{-\lambda x}$$

Exponential Distribution

❖ G서비스 센터는 10분에 평균 2회의 전화가 온다. 대기시간이 2분 이내일 확률은?

- 평균 대기시간

$$\lambda = 2$$

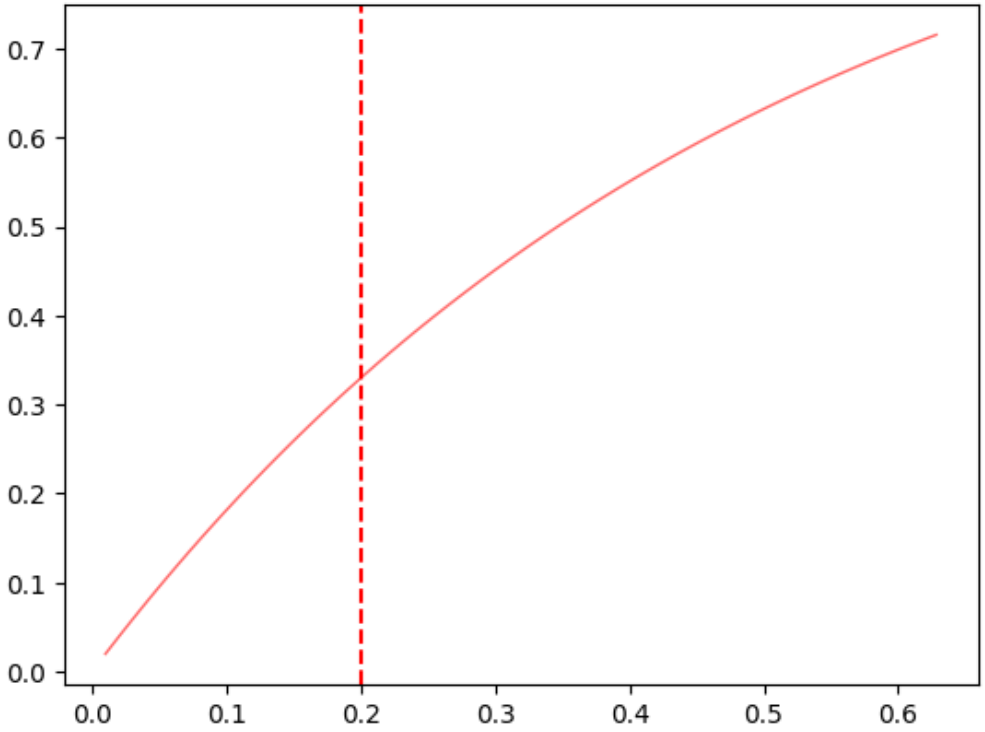
$$\mu = \frac{1}{\lambda} = \frac{1}{2} = 0.5$$

$$\mu = 10 \times 0.5 = 5min$$

$$\sigma = \frac{1}{\lambda} = 0.5$$

- 대기시간

$$x = \frac{2min}{10min} = 0.2$$



$$P(X \leq 0.2) = 1 - e^{-(2)(0.2)} = 0.330$$

Normal Distribution

Normal Distribution

❖ 정규분포(Normal Distribution)

- 평균을 중심으로 좌우대칭이고 종모양을 갖는 확률분포
- 개발자 : 프랑스 수학자 드 므와브르 (Abraham de Moivre; 1667 – 1754)
- 확산자 : 독일 수학자 가우스 (Carl Friedrich Gauss; 1777 – 1855) : 물리학 및 천문학에서 사용
- 정규분포 이름 사용: 피어슨
- 통계이론에 있어서 매우 중요한 확률분포
- 통계분석 시 모집단의 분포를 정규분포라고 대부분 가정하고 통계분석을 하기 때문 (중심극한 정리)

Normal Distribution

❖ 연속확률변수의 계산방법

- 평균

연속확률변수

$$E(X) = \mu = \int_{-\infty}^{+\infty} x f(x) dx$$

이산확률변수

$$E(X) = \mu = \sum_{i=1}^{n} x_i P(x_i)$$

- 분산

$$V(X) = \sigma^2 = \int_{-\infty}^{+\infty} (x_i - \mu)^2 f(x) dx$$

$$V(X) = \sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 P(x_i)$$

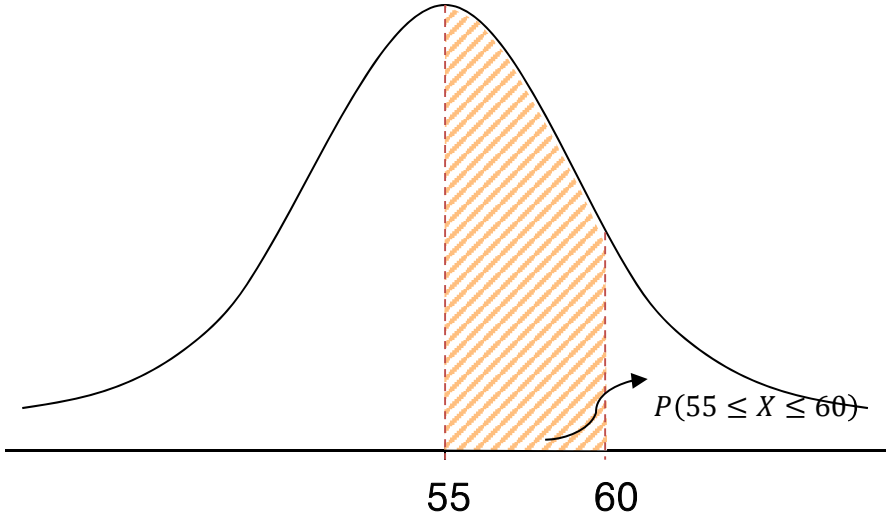
- 정규분포확률 (평균과 표준편차)

$$P(a \le X \le b) = \int_a^b f(x) dx = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \qquad X \sim N(\mu, \sigma)$$

Normal Distribution

❖ 정규분포의 계산방법

- 몸무게가 55-60일 확률은?
- 실제로 적분으로 계산하기 어려움
- 표준정규분포표를 이용
- 통계소프트웨어 이용



$$P(55 \le X \le 60) = \int_{55}^{60} f(x)dx$$

$$= \int_{55}^{60} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$=?$$

Normal Distribution

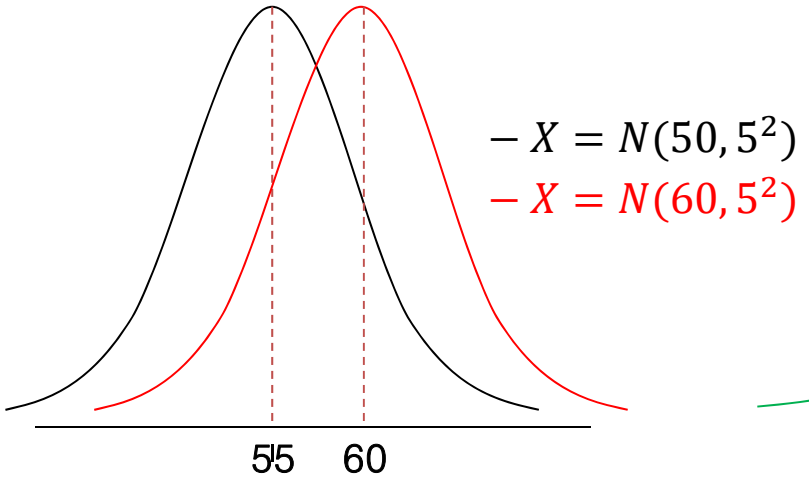
❖ 정규분포

- 평균과 분산으로 표시

$$X = N(\mu, \sigma^2)$$

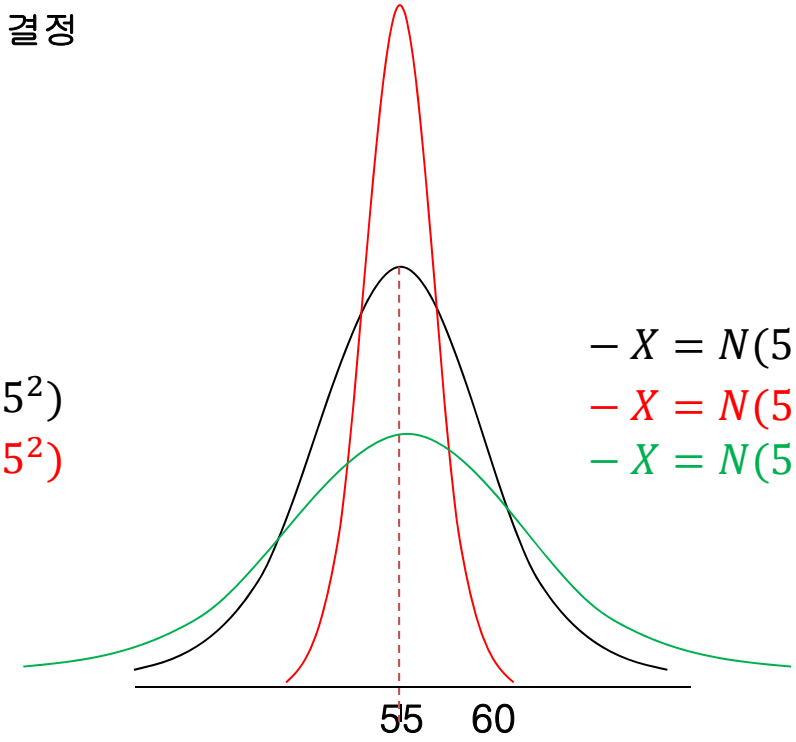
- 분산(표준편차)에 의해 분포의 넓이가 결정

  * 표준편차에 의해 오차범위가 결정됨



$-\ X = N(50, 5^2)$

$-\ X = N(60, 5^2)$



$-\ X = N(50, 5^2)$

$-\ X = N(50, 1^2)$

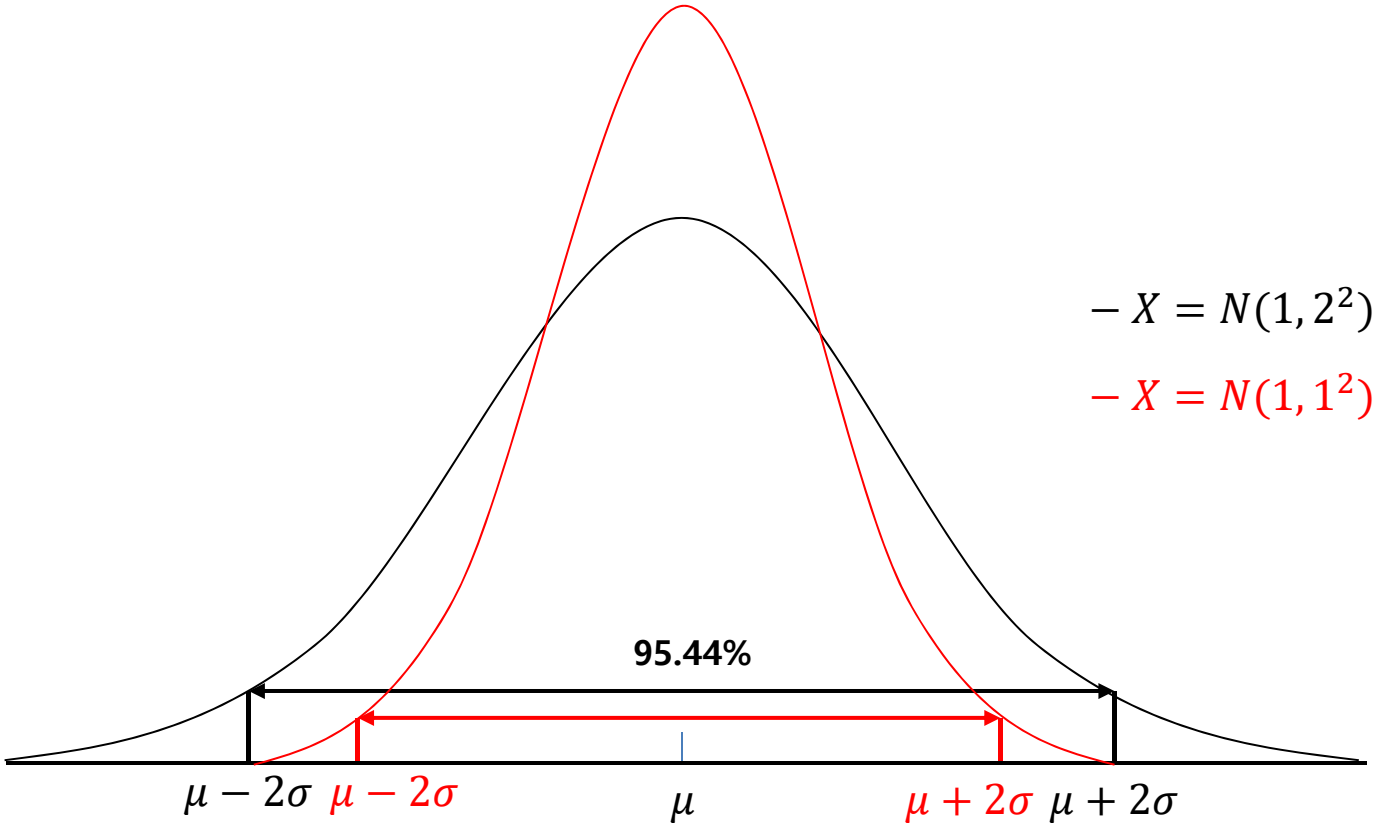
$-\ X = N(50, 10^2)$

Normal Distribution

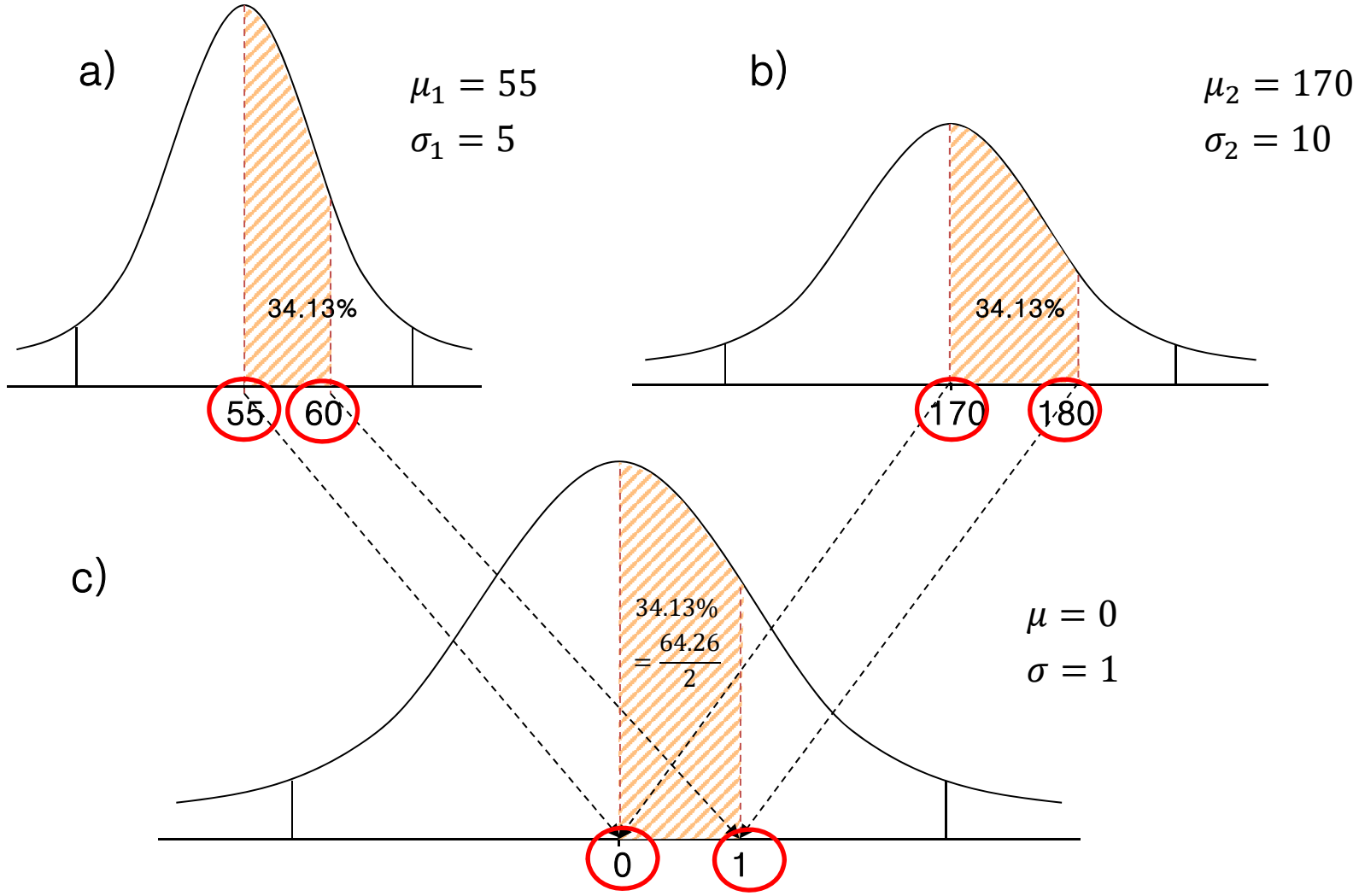
❖ Empirical Rule (경험적 법칙 - μ)

- $k = 2,\ 95.44\%$ 이상의 데이터가 $\mu \pm 2\sigma$ 사이에 있음

❖ 표준편차(σ)의 크기에 의해 오차범위의 크기가 결정됨



Standard Normal Distribution



Standard Normal Distribution

❖ 표준확률변수(standardized random variable)

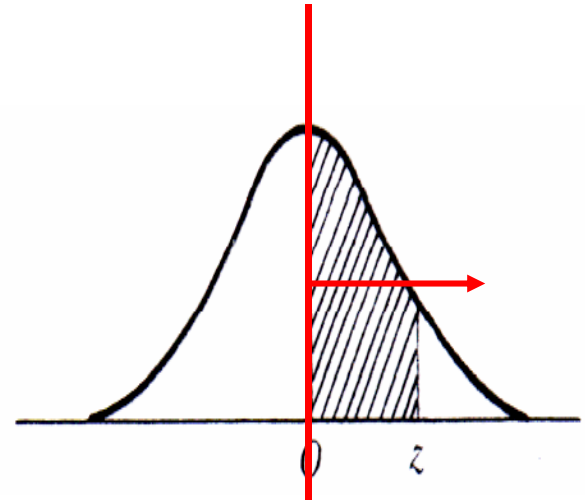
- 임의의 확률변수를 표준화 시킴
- 측정단위 등과 관계없이 자료를 표준화 시킨 값
- 평균으로부터 떨어진 거리를 σ로 계산
- Empirical Rule (경험적 법칙)

$$z = \frac{x - \mu}{\sigma}, \qquad X \sim N(\mu, \sigma^2)$$

❖ 표준정규분포(Standard Normal Distribution)

- 표준확률변수의 확률분포

Standard Normal Distribution



z = 평균으로 부터 떨어진 거리

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.012	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.43				.4406	.4418	.4429	.4441
1.6	.4452	.4463	.44				.4515	.4525	.4535	.4545
1.7	.4554	.4564	.45				.4608	.4616	.4625	.4633
1.8	.4641	.4649	.46				.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0										

정규분포

$$z = 1.96$$

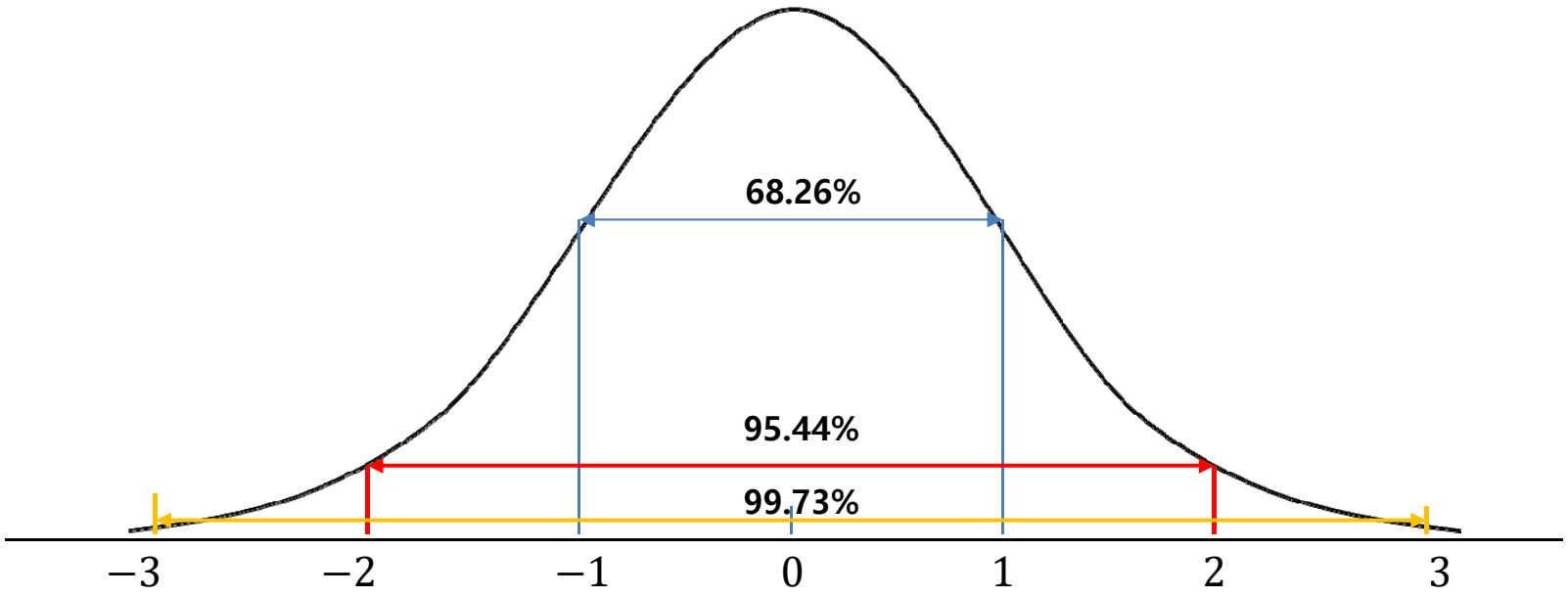
$$P(z) = 0.475(\frac{0.95}{2})$$

z = 1.96 의 의미 → 평균으로 1.96 발자국(거리)안에 전체 데이터의 47.5%가 있음

Standard Normal Distribution

❖ Empirical Rule (경험적 법칙 - z or t)

- $P(-1 \leq z \leq 1) = 2 \times 0.3413 = 0.6826$
- $P(-2 \leq z \leq 2) = 2 \times 0.4772 = 0.9544$
- $P(-3 \leq z \leq 3) = 2 \times 0.49865 = 0.9973$



Standard Normal Distribution

❖ G대학 통계학 수업 수강생 100명이 몸무게를 조사

- 몸무게 평균을 중심으로 95%확률로
  예측할 수 있는 몸무게의 범위는?
- 확률분포표로 계산

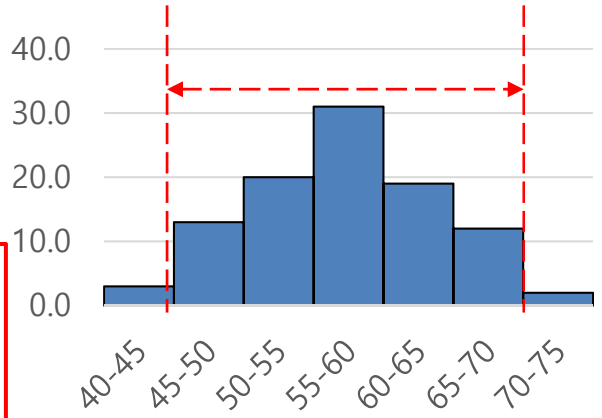
$$0.95 = P(45 \leq X \leq 70)$$

- 평균, 표준편차를 알면 표준정규분포로 계산

$$\mu = 56.78 \qquad \sigma = 6.80$$

$$z = \frac{x - \mu}{\sigma} \rightarrow x = \mu \pm z\sigma$$

$$x = 56.78 \pm 1.96(6.80) = 56.94 \pm 13.33 = [43.32, 70.24]$$

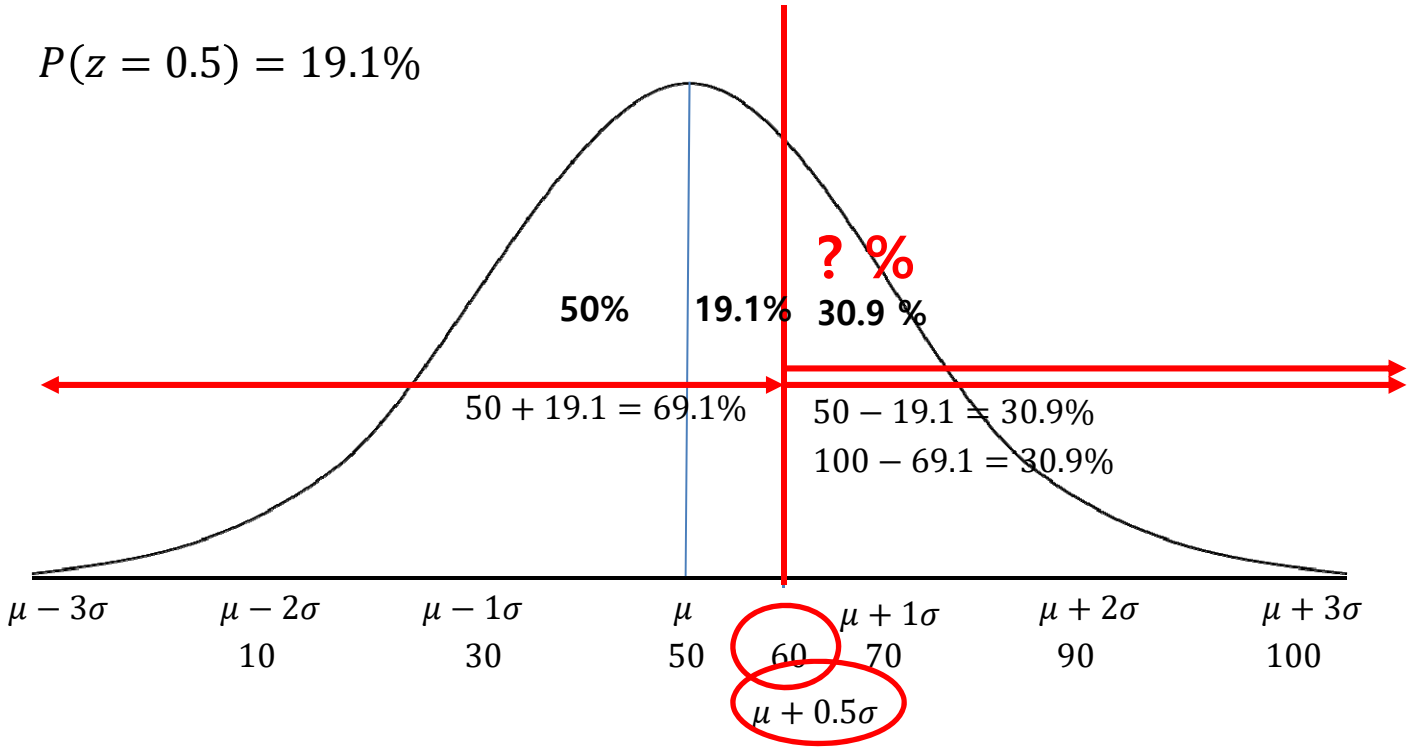


Standard Normal Distribution

❖ G대학 경영통계 수업 듣는 학생의 시험점수가 평균 50, 표준편차가 20점이라고 한다.

- 60점을 받았다면 상위 몇 %인가?
- 평균으로 부터 몇 발자국 떨어져 있는가?

$$z = \frac{x - \mu}{\sigma} = \frac{60 - 50}{20} = 0.5$$



Standard Normal Distribution

- 상위 35%에 들기 위한 점수는?
- 평균으로부터 몇 발자국 떨어져 있어야 하는가?

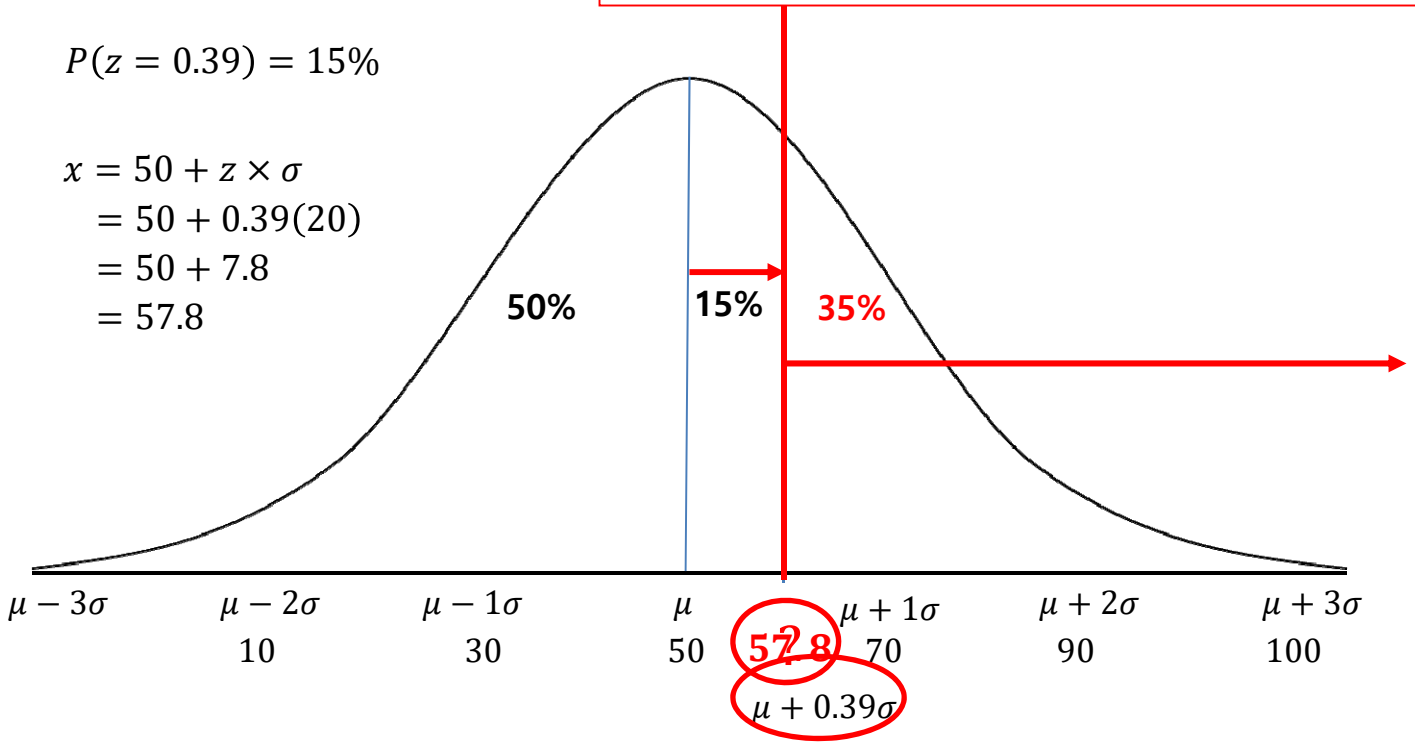
$50 - 35 = 15\%$

$P(z = ?) = 15\%$

$P(z = 0.39) = 15\%$

$$x = 50 + z \times \sigma$$
$$= 50 + 0.39(20)$$
$$= 50 + 7.8$$
$$= 57.8$$

z	.00	.01	.02	.03	.04	.05	.06	.07	08	.09
0.0	.0000	.0040	.0080	.012	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	0.478	0.517	0.557	0.596	0.636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879



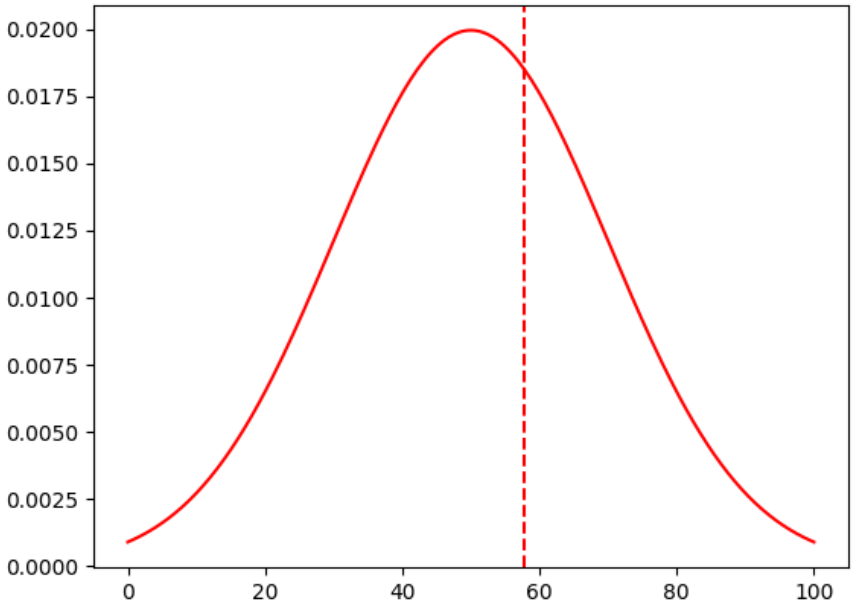
Normal Distribution

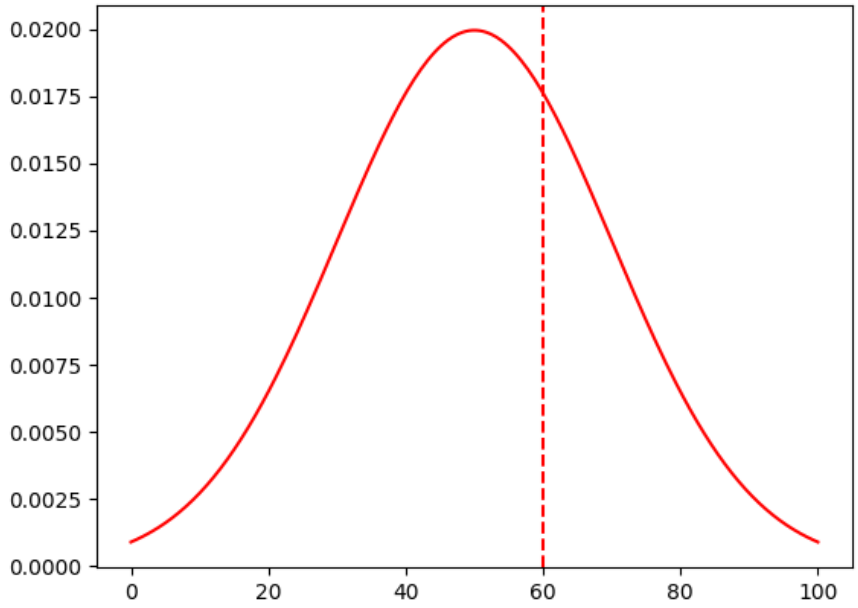
❖ K대학 경영통계 수업 듣는 학생의 시험점수가 평균 50, 표준편차가 20점이라고 한다. (R)

- 점수가 60점일 때 상위 %는
- 상위 35%에 들기 위한 점수는?

$$P(x = 60) = 1 - 69.1 = 30.9\%$$

$$65\% = P(x \geq 57.7)$$





실습

3_1.확률분포

❖ K식당에서는 손님에게 fortune 쿠키를 제공한다. 성공율은 70%이다. 10명의 손님이 앉은 테이블에서

- 1명도 못 받을 확률

$$P(X = 0) = 0.000$$

- 10명 다 받을 확률

$$P(X = 10) = 0.028$$

- 7명 이하로 받을 확률

$$P(X \leq 7) = 0.617$$

- 7명 이상으로 받을 확률

$$P(X \geq 7) = 0.650$$

- 7명 이상으로 받을 확률

$$P(2 \leq X \leq 8) = 0.851$$

2.Binominal Distribution

3_1.확률분포

1.기본 package 설치

[1]
```
## 1.기본
import numpy as np  # numpy 패키지 가져오기
import matplotlib.pyplot as plt # 시각화 패키지 가져오기
import seaborn as sns # 시각화

## 2.데이터 가져오기
import pandas as pd # csv -> dataframe으로 전환

## 3.확률분포
from scipy import stats
from scipy.stats import binom # 이항분포
from scipy.stats import poisson # 포아송분포
from scipy.stats import expon # 지수분포
from scipy.stats import norm # 정규분포
```

2.이항분포

- https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.binom.html

[2]
```
# 이항분포 확률값구하기
# 성공률 40%일때, 3명의 손님이 앉은 테이블에서 3명이 모두 못 받을 확률은?
n = 3   # 시행횟수
p = 0.4  # 성공확률
x = 0  # 성공횟수
binom.pmf(k = x, n = n, p = p).round(3)
```

```
0.216
```

✓ 0초 오후 7:31에 완료됨

2.Binominal Distribution

2.이항분포

- https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.binom.html

[2]
```
# 이항분포 확률값구하기
# 성공률 40%일때, 3명의 손님이 앉은 테이블에서 3명이 모두 못 받을 확률은?
n = 3   # 시행횟수
p = 0.4  # 성공확률
x = 0  # 성공횟수
binom.pmf(k = x, n = n, p = p).round(3)
```

0.216

[3]
```
# 이항확률분포 확률분포표
# 성공률 70%일때
n = 10   # 시행횟수
p = 0.7  # 성공확률
x = [0,1,2,3,4,5,6,7,8,9,10] # 성공횟수
results = binom.pmf(x, n = n, p = p).round(3)
results1 = binom.cdf(x, n = n, p = p).round(3)

binom_df = pd.DataFrame({'x': x, 'pmf': results, 'cdf': results1})
binom_df
```

	x	pmf	cdf
0	0	0.000	0.000
1	1	0.000	0.000
2	2	0.001	0.002
3	3	0.009	0.011
4	4	0.037	0.047
5	5	0.103	0.150
6	6	0.200	0.350

✓ 0초 오후 7:31에 완료됨

2.Binominal Distribution

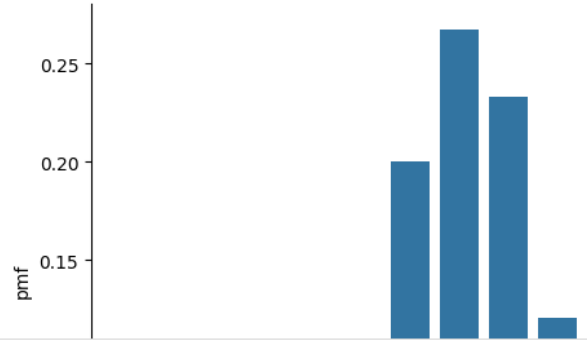
```
[4] # 이항분포 누적 확률값 구하기
    # 성공률 70%일때, 7명이하 받을 확률
    n = 10   # 시행횟수
    p = 0.7  # 성공확률
    x = 7    # 성공횟수
    binom.cdf(k = x, n = n, p = p).round(3)
```

```
0.617
```

```
[5] # 이항분포 누적 확률값 구하기
    # 성공률 70%일때, 7명이상 받을 확률
    n = 10   # 시행횟수
    p = 0.7  # 성공확률
    x = 6    # 성공횟수
    1- binom.cdf(k = x, n = n, p = p).round(3)
```

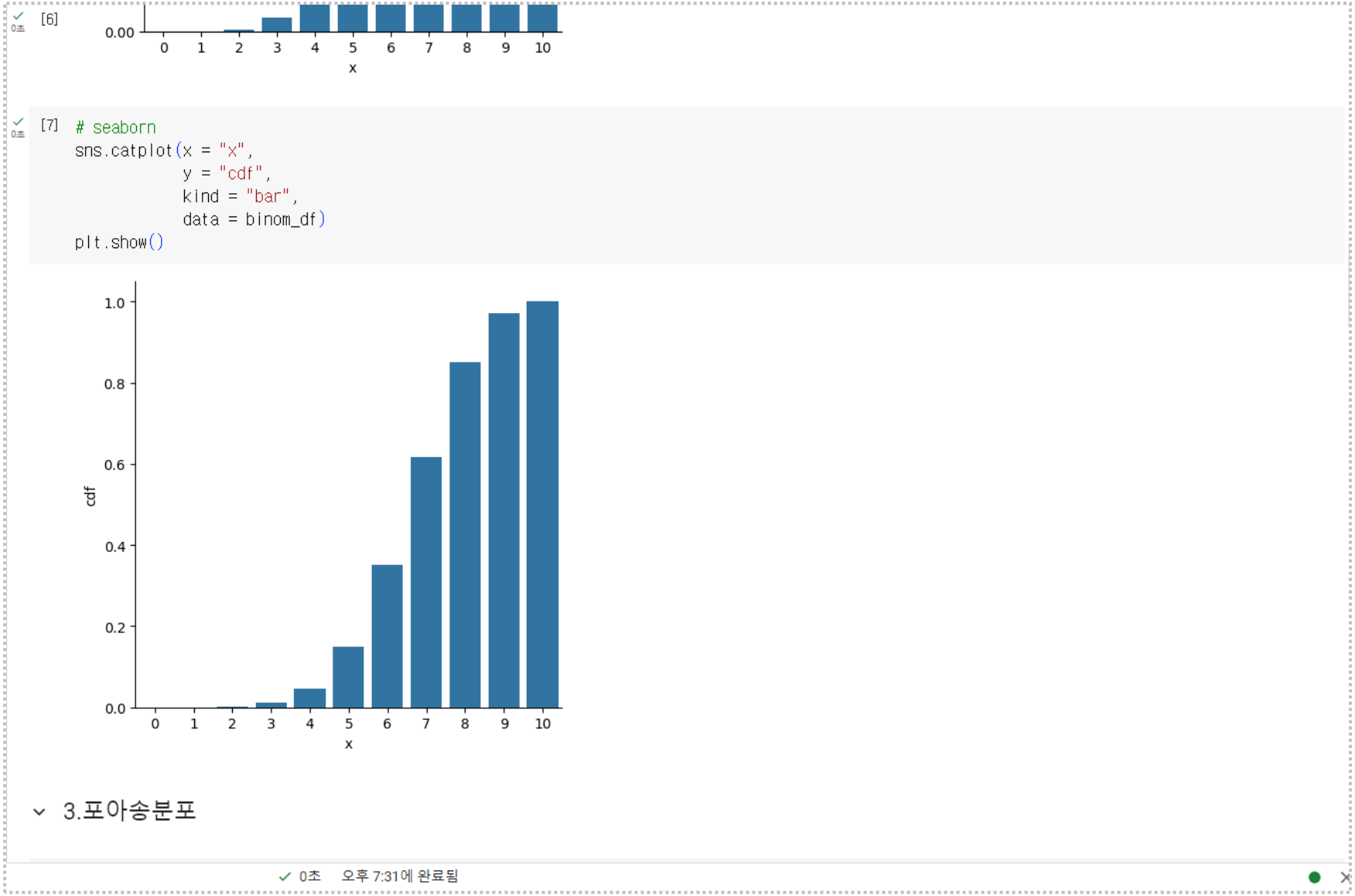
```
0.65
```

```
[6] # seaborn
    sns.catplot(x = "x",
                y = "pmf",
                kind = "bar",
                data = binom_df)
    plt.show()
```



✓ 0초 오후 7:31에 완료됨

2.Binominal Distribution



3.Poisson Distribution

3.포아송분포

[8]
```
# 포아송분포 확률값구하기
# G서비스 센터는 10분에 평균 1회의 전화가 온다. 10분 동안에 2회의 전화를 받을 확률은?
lamb = 1   # 평균발생건수
x = 2      # 발생건수
poisson.pmf(k = x, mu = lamb).round(3)
```

0.184

[9]
```
# 포아송분포 확률분포표
# G서비스 센터는 10분에 평균 1회의 전화가 온다.
lamb = 1   # 평균발생건수
x = [0,1,2,3,4,5,6,7,8,9,10] # 발생건수
results = poisson.pmf(x, mu = lamb).round(3)
results1 = poisson.cdf(x, mu = lamb).round(3)

poisson_df = pd.DataFrame({'x': x, 'pmf': results, 'cdf': results1})
poisson_df
```

	x	pmf	cdf
0	0	0.368	0.368
1	1	0.368	0.736
2	2	0.184	0.920
3	3	0.061	0.981
4	4	0.015	0.996
5	5	0.003	0.999
6	6	0.001	1.000
7	7	0.000	1.000
8	8	0.000	1.000
9	9	0.000	1.000

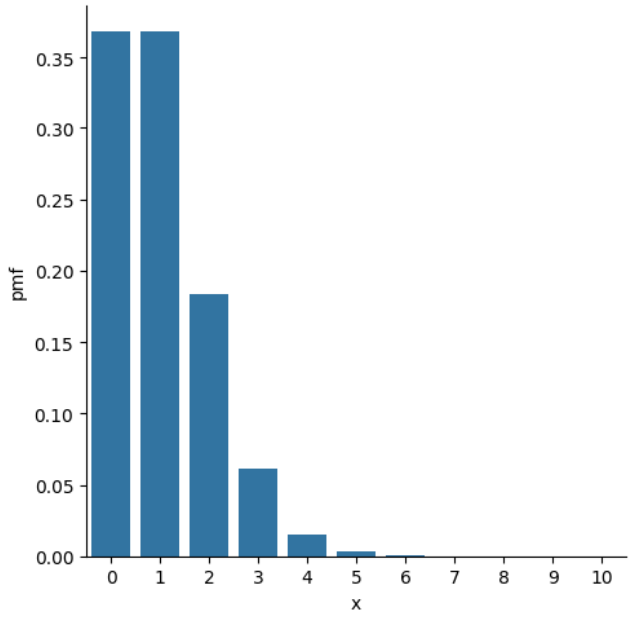
✓ 0초 오후 7:31에 완료됨

3.Poisson Distribution

```
[10] # 포아송분포 누적 확률값구하기
     # 2회 이상 전화 받을 확률은
     lamb = 1   # 평균발생건수
     x = 1      # 발생건수
     1- poisson.cdf(k = x, mu = lamb).round(3)
```

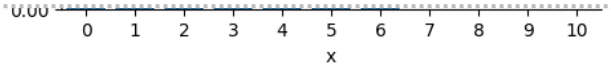
```
0.264
```

```
[11] # seaborn
     sns.catplot(x = "x",
                 y = "pmf",
                 kind = "bar",
                 data = poisson_df)
     plt.show()
```

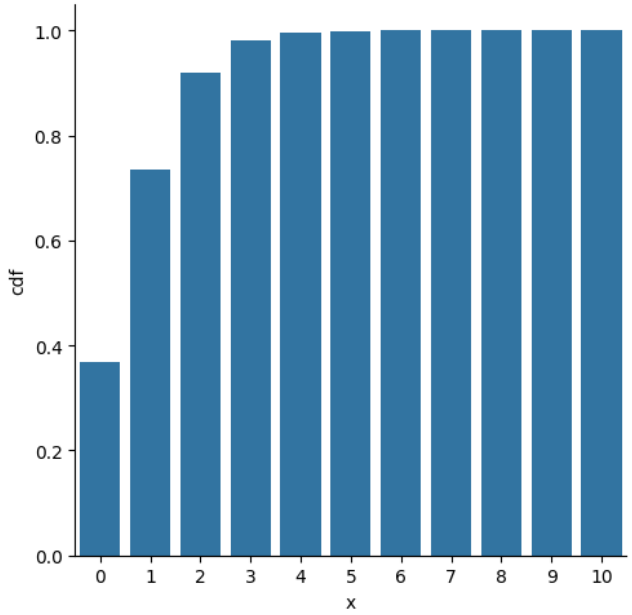


✓ 0초 오후 7:31에 완료됨

3.Poisson Distribution

[11]


[12]
```
# seaborn
sns.catplot(x = "x",
            y = "cdf",
            kind = "bar",
            data = poisson_df)
plt.show()
```



4.지수분포

- https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.expon.html

✓ 0초 오후 7:31에 완료됨

04.Exponential Distribution

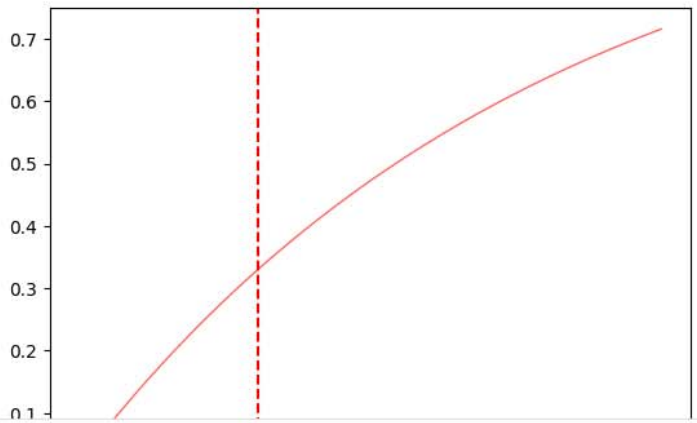
4.지수분포

- https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.expon.html

```
[12] # 지수분포 확률값구하기
     # G서비스 센터는 10분에 평균 1회의 전화가 온다. 대기시간이 2분 이내일 확률은?
     lamb = 2          # 평균발생건수
     scale = 1/lamb    # 평균대기시간
     time = 10         # 기준시간
     queue = 2         # 대기시간
     x = queue/time    # 대기시간비율
     expon.cdf(scale = scale, x = x).round(3)
```

```
0.33
```

```
[13] fig, ax = plt.subplots(1, 1)
     x = np.linspace(expon.cdf(0.01), expon.cdf(0.99), 100)
     ax.plot(x, expon.cdf(x, scale = 1/lamb), 'r-', lw=1, alpha=0.6)
     plt.axvline(x = 0.2, color='r', linestyle='--')
     plt.show()
```



✓ 0초 오후 6:36에 완료됨

05.Normal Distribution

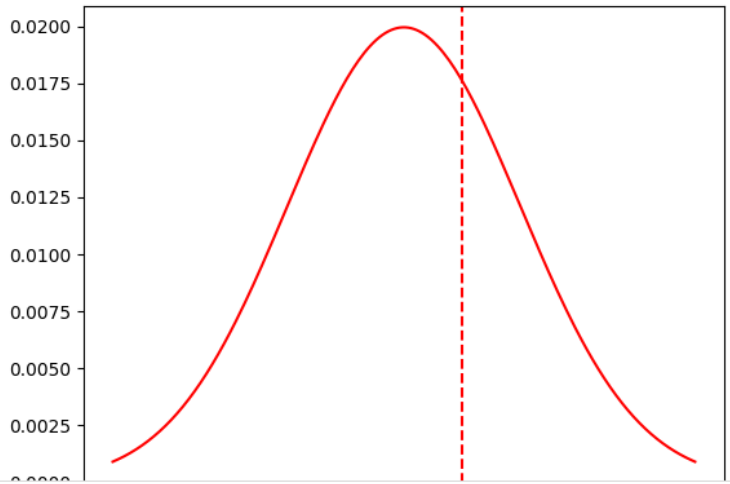
5.정규분포

[15]
```
# 정규분포 확률값구하기
# G대학 경영통계 수업 듣는 학생의 시험점수가 평균 50, 표준편차가 20점이라고 한다.
# 60점을 받았다면 상위 몇 %인가?

mu = 50 # 평균
std = 20 # 표준편차
x = 60
1-norm.cdf(x, loc = mu, scale = std).round(3)
```

```
0.30900000000000005
```

[16]
```
x_data = np.linspace(0, 100, 200)
mu = 50 # 평균
std = 20 # 표준편차
plt.plot(x_data, norm.pdf(x_data, loc = mu, scale = std), 'r-')
plt.axvline(x = 60, color='r', linestyle='--')
plt.show()
```



✓ 0초 오후 7:39에 완료됨

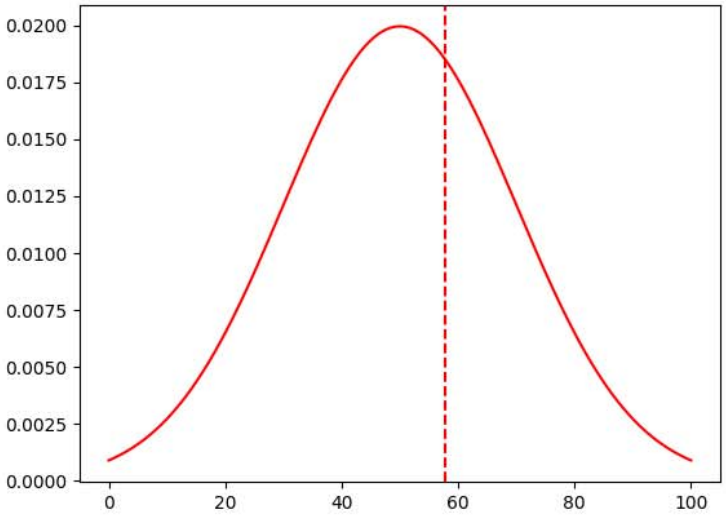
05.Normal Distribution

```
[17] # 정규분포 확률값구하기
     # 상위 35%에 들기 위한 점수는?

     mu = 50 # 평균
     std = 20 # 표준편차
     q = 0.65
     norm.ppf(q, loc = mu, scale = std).round(3)
```

```
57.706
```

```
[18] x_data = np.linspace(0, 100, 200)
     mu = 50 # 평균
     std = 20 # 표준편차
     plt.plot(x_data, norm.pdf(x_data, loc = mu, scale = std), 'r-')
     plt.axvline(x = 57.706, color='r', linestyle='--')
     plt.show()
```



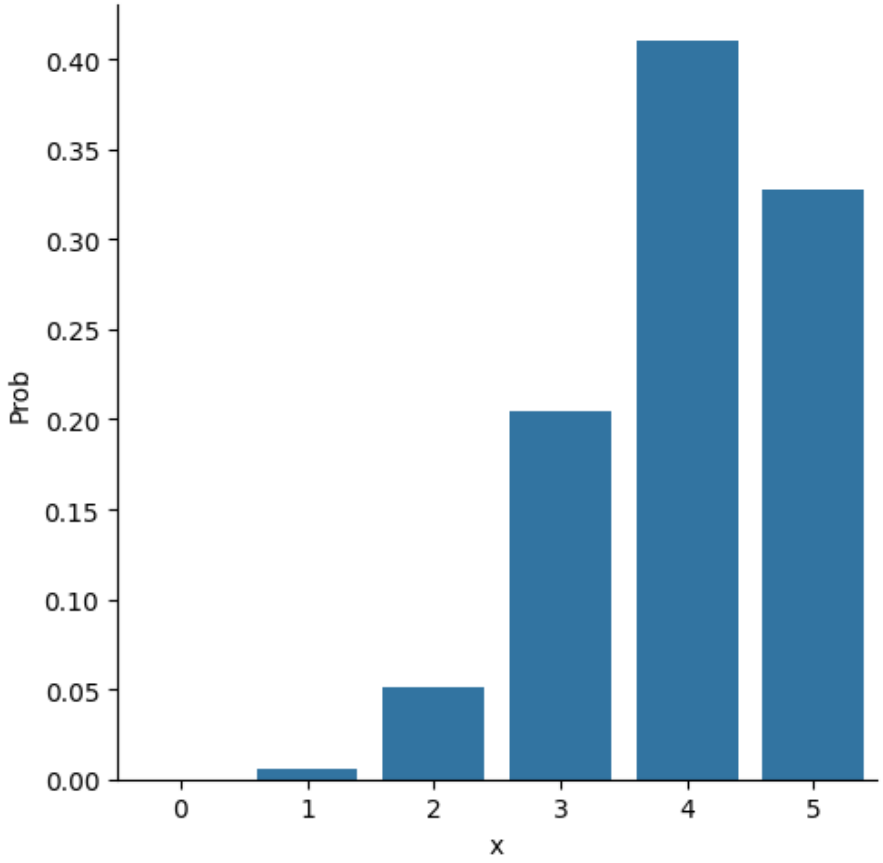
✓ 0초 오후 7:39에 완료됨

연습문제

연습문제1

- G식당에서는 손님에게 fortune 쿠키를 제공한다. 성공율은 80%이다. 5명의 손님이 앉은 테이블에서 3명이상 받을 확률은?
- 정답: 0.942

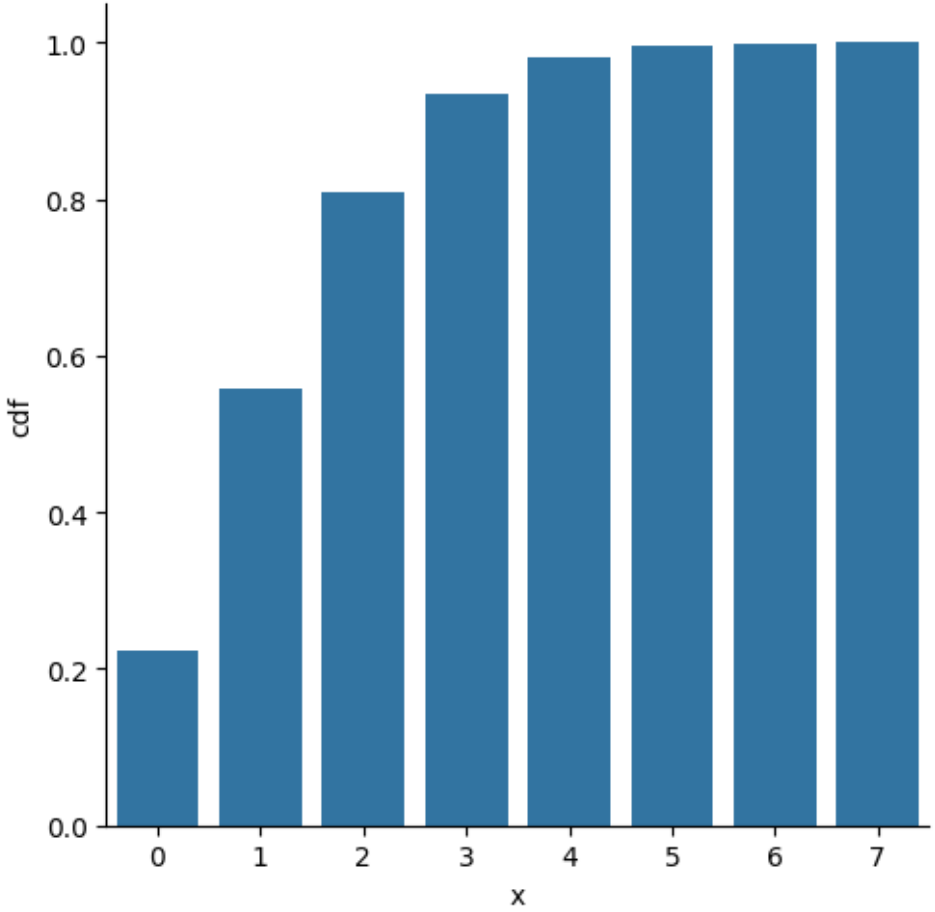
	x	pmf	cdf
0	0	0.000	0.000
1	1	0.006	0.007
2	2	0.051	0.058
3	3	0.205	0.263
4	4	0.410	0.672
5	5	0.328	1.000



연습문제2

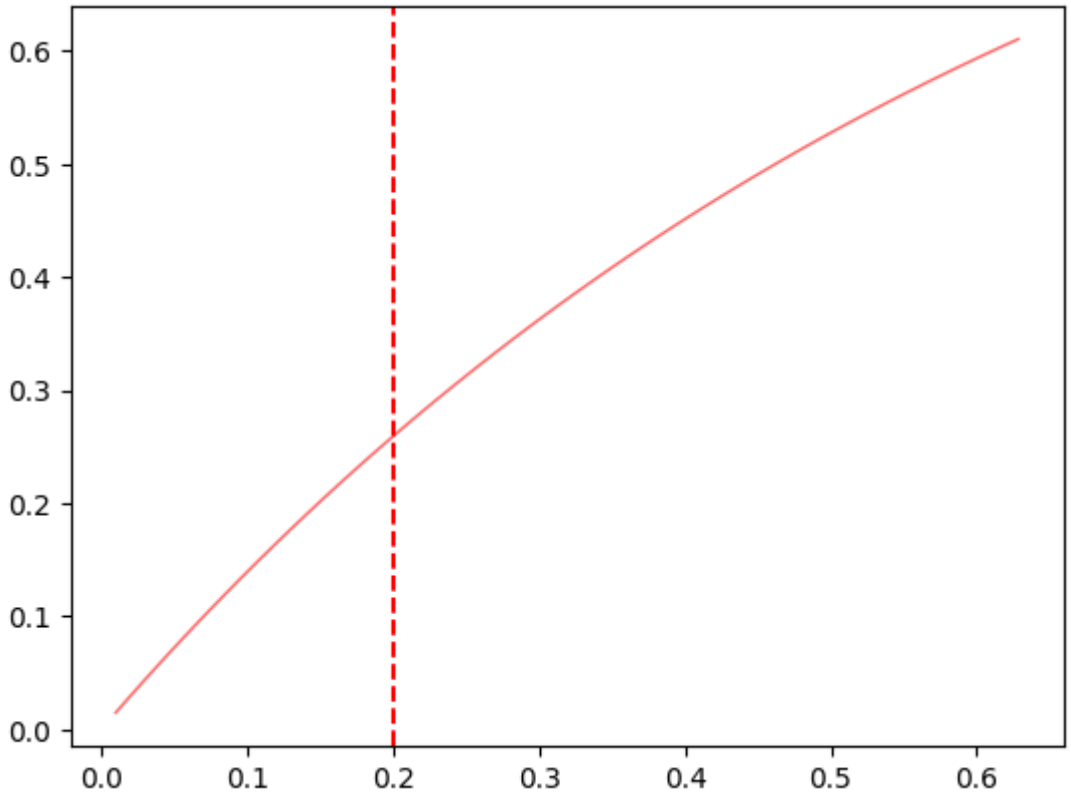
- G서비스 센터는 5분에 평균 1.5회의 전화가 온다. 5분 동안에 1회 이하로 전화를 받을 확률은?
- 정답: 0.558

	x	pmf	cdf
0	0	0.223	0.223
1	1	0.335	0.558
2	2	0.251	0.809
3	3	0.126	0.934
4	4	0.047	0.981
5	5	0.014	0.996
6	6	0.004	0.999
7	7	0.001	1.000



연습문제3

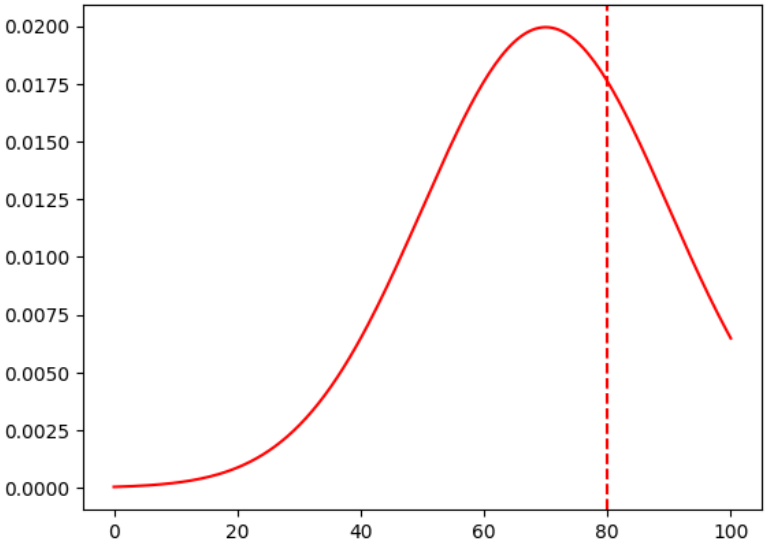
- G서비스 센터는 5분에 평균 1.5회의 전화가 온다. 대기시간이 1분 이내일 확률은?
- 정답: 0.259

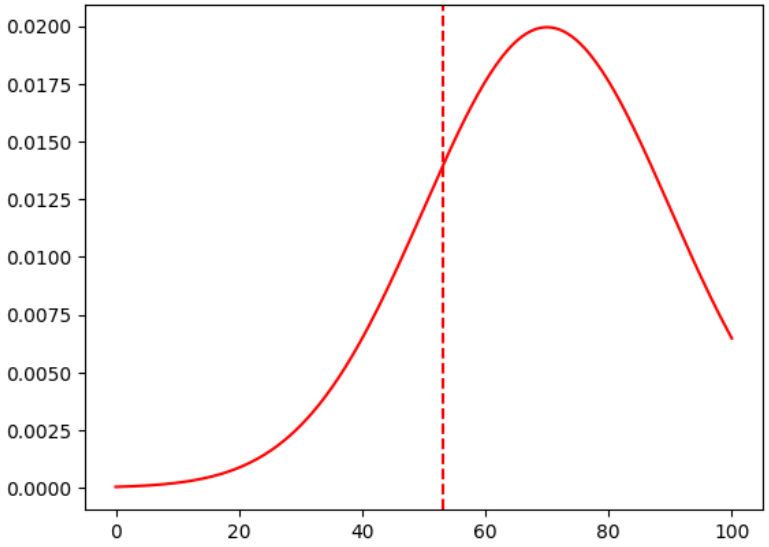


연습문제5

❖ G대학 경영통계 수업 듣는 학생의 시험점수가 평균 70, 표준편차가 20점이라고 한다.

- 80점이면 상위 몇%인가? 정답:30.1%
- (80%까지 B가 주어진다.) B이상을 받기 위한 점수는? 58.168점





III. 추정과 가설검정

표본분포

Sampling Distribution

❖ 표본분포 (sampling distribution)

- 모집단에서 추출한 표본크기 n개인 추정치(estimator)의 확률분포
- 모집단에서 일정한 크기(n)로 표본을 모두(k개) 뽑아서 각 표본의 평균을 계산하였을 때, 그 표본의 평균 $\bar{X}_1, \bar{X}_2, \bar{X}_3, \ldots, \bar{X}_k$ 의 확률분포

❖ (μ, σ^2)인 분포에서 n개의 확률표본을 추출했을 때 표본평균($\bar{X}$)의 분포의 특징

- 평균 : 모집단의 평균과 표본들의 평균은 같음
- 표준오차 (standard error of the mean, SE): 통계량의 표준편차

$$E(\bar{X}) = \mu \qquad SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

- 표본오차(sampling error) vs 표준오차 (standard error) vs 표준편차(standard deviation)

$$error = \bar{x} - \mu \qquad SD = \sigma \qquad SE = \frac{\sigma}{\sqrt{n}}$$

Sampling Distribution

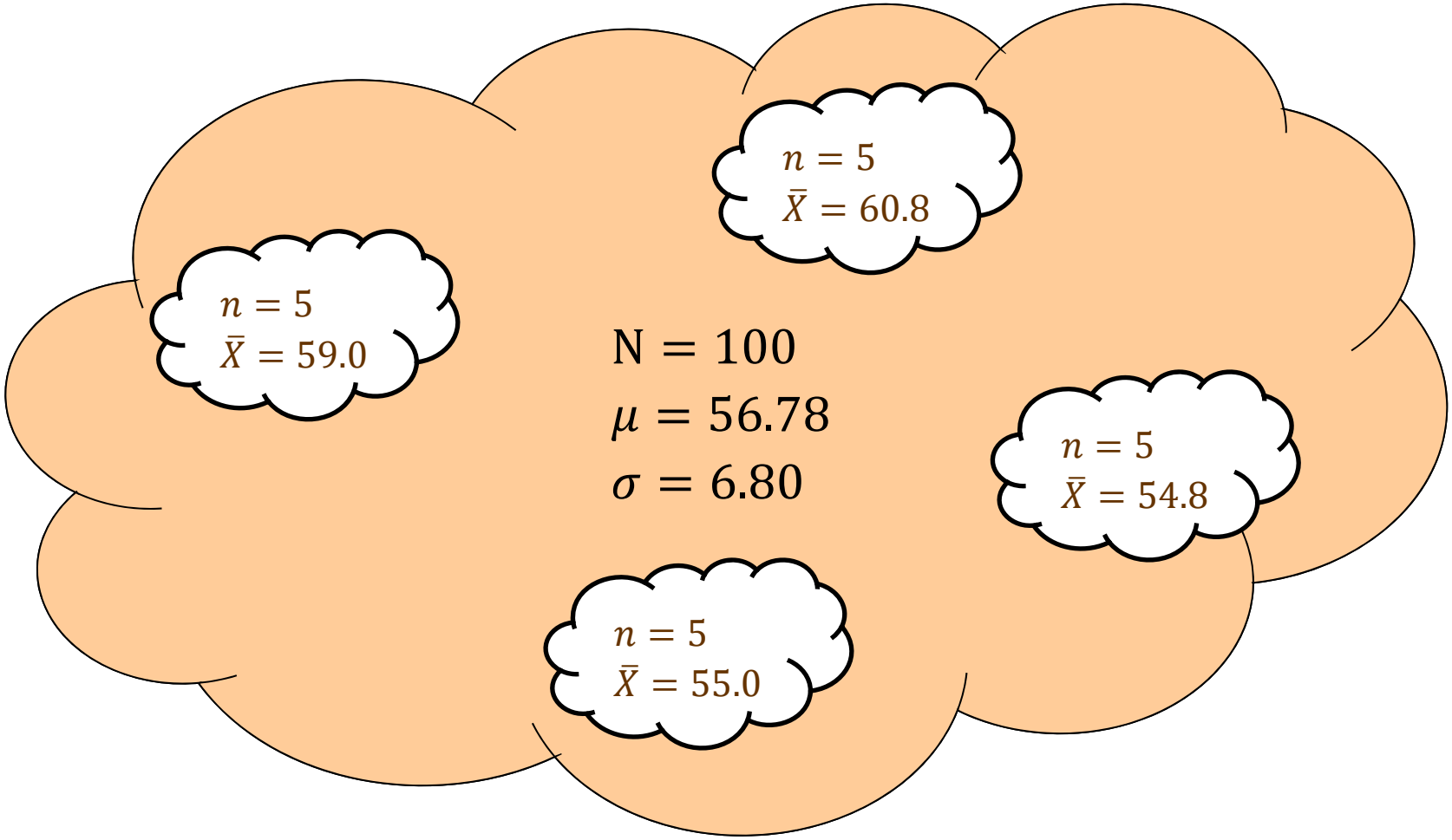
❖ G대학 경영통계 수강생의 몸무게 분석

(단위: kg)

40	56	57	56	62	60	57	65	61	50
50	72	55	48	59	59	49	67	47	55
56	46	67	50	71	53	54	57	48	56
51	63	52	68	58	47	53	62	54	62
55	56	55	53	41	62	58	53	52	46
61	52	62	58	58	56	58	58	56	49
69	48	67	62	48	59	59	60	59	60
44	57	49	57	69	50	60	51	59	63
66	52	49	68	53	61	61	54	68	69
60	45	52	54	57	64	61	59	57	66

Sampling Distribution

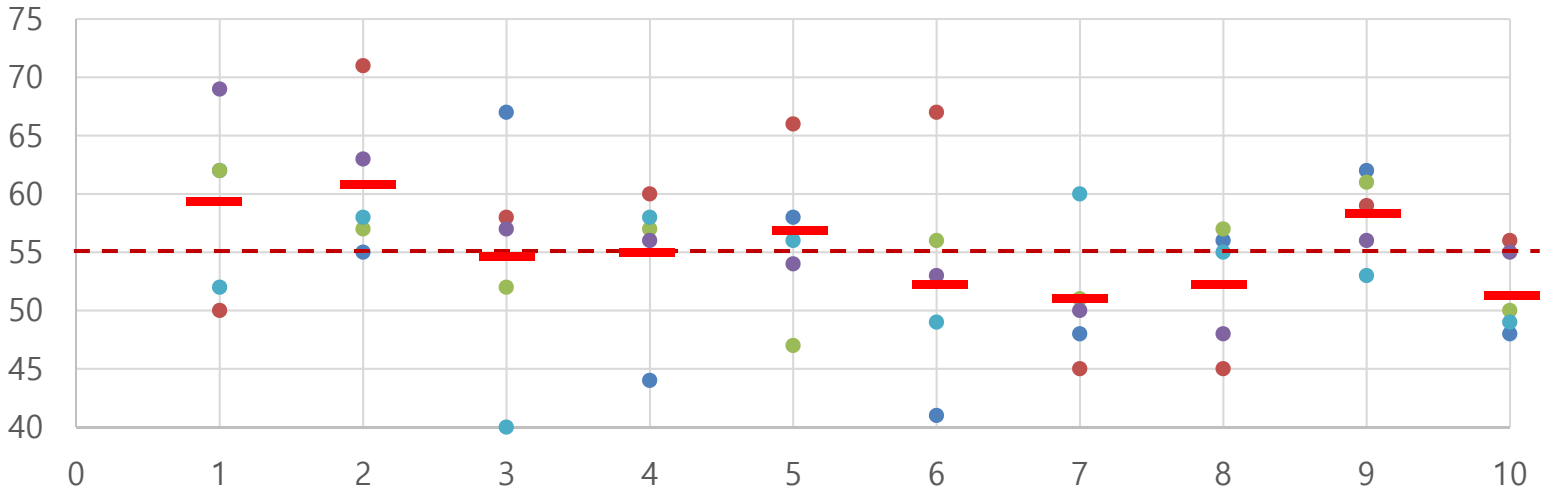
❖ 모집단(100명)에서 표본을 추출한다면



Sampling Distribution

❖ n=5일 때, 평균($\bar{\bar{X}}$)

	sample1	sample2	sample3	sample4	sample5	sample6	sample7	sample8	sample9	sample10	
1	62	55	67	44	58	41	48	56	62	48	
2	50	71	58	60	66	67	45	45	59	56	
3	62	57	52	57	47	56	51	57	61	50	
4	69	63	57	56	54	53	50	48	56	55	
5	52	58	40	58	56	49	60	55	53	49	
$\bar{x}$	59.0	60.8	54.8	55.0	56.2	53.2	50.8	52.2	58.2	51.6	$\bar{\bar{x}}$=55.18
s	7.9	6.4	9.9	6.3	6.9	9.5	5.6	5.4	3.7	3.6	$s_{\bar{x}}$ =3.36



$$\bar{\bar{X}} = 55.18 \approx \mu = 56.78$$

Sampling Distribution

❖ n=5일 때, SE($\sigma_{\bar{x}}$)

	sample1	sample2	sample3	sample4	sample5	sample6	sample7	sample8	sample9	sample10	
1	62	55	67	44	58	41	48	56	62	48	
2	50	71	58	60	66	67	45	45	59	56	
3	62	57	52	57	47	56	51	57	61	50	
4	69	63	57	56	54	53	50	48	56	55	
5	52	58	40	58	56	49	60	55	53	49	
$\bar{x}$	59.0	60.8	54.8	55.0	56.2	53.2	50.8	52.2	58.2	51.6	$\bar{\bar{x}}$=55.18
s	7.9	6.4	9.9	6.3	6.9	9.5	5.6	5.4	3.7	3.6	$s_{\bar{x}}$ =3.36

$$E(\bar{X}) = \mu \qquad 55.18 \approx 56.78$$

$$\sigma_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^{n}(\bar{x}_1 - \mu)^2}{n-1}}$$

$$= \sqrt{\frac{(59.0 - 55.18)^2 + \cdots + (51.6 - 55.18)^2}{9}}$$

$$= \sqrt{11.26}$$

$$= 3.36$$

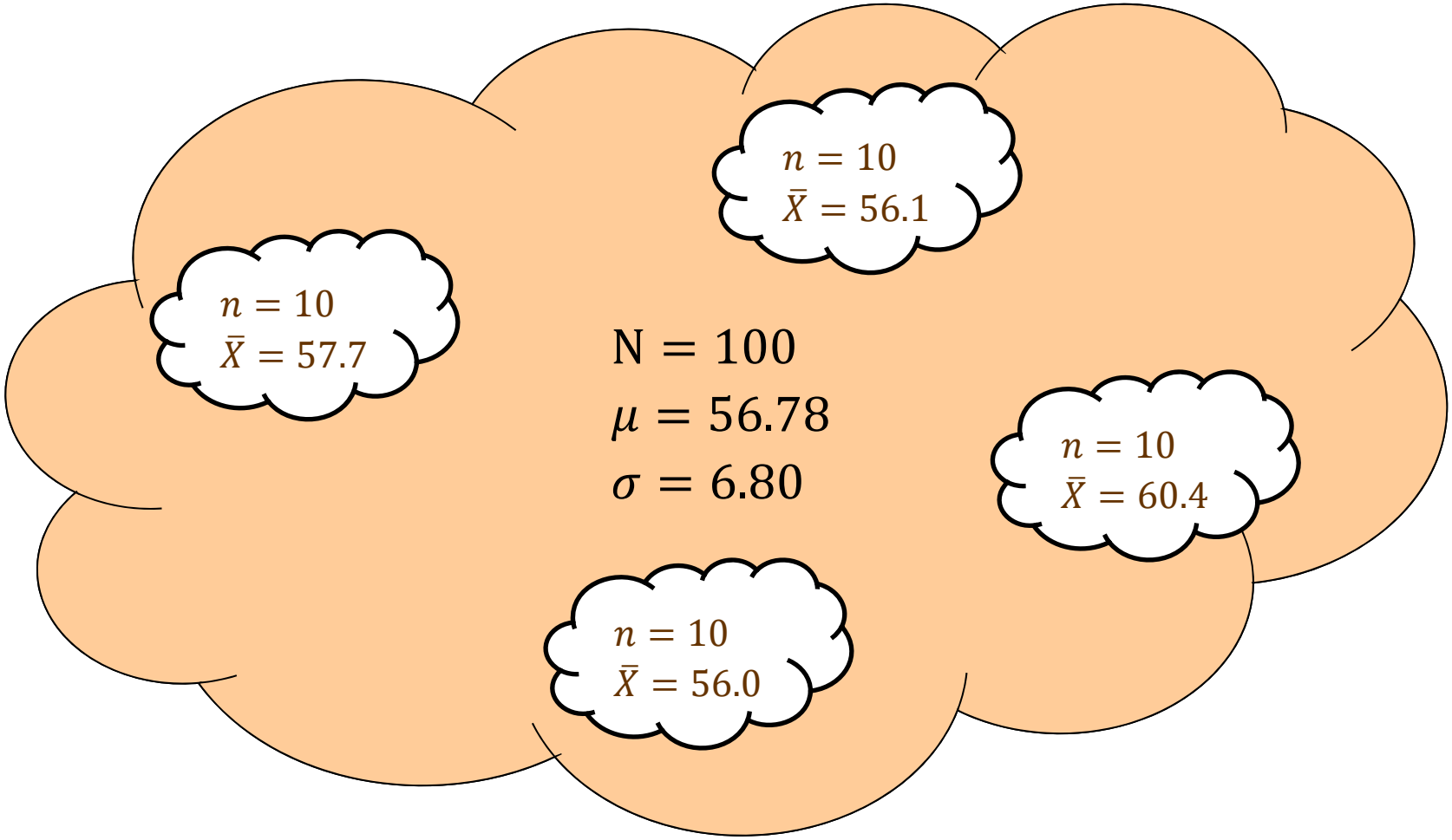
$$\approx \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$= \frac{6.80}{\sqrt{5}}$$

$$= 3.04$$

Sampling Distribution

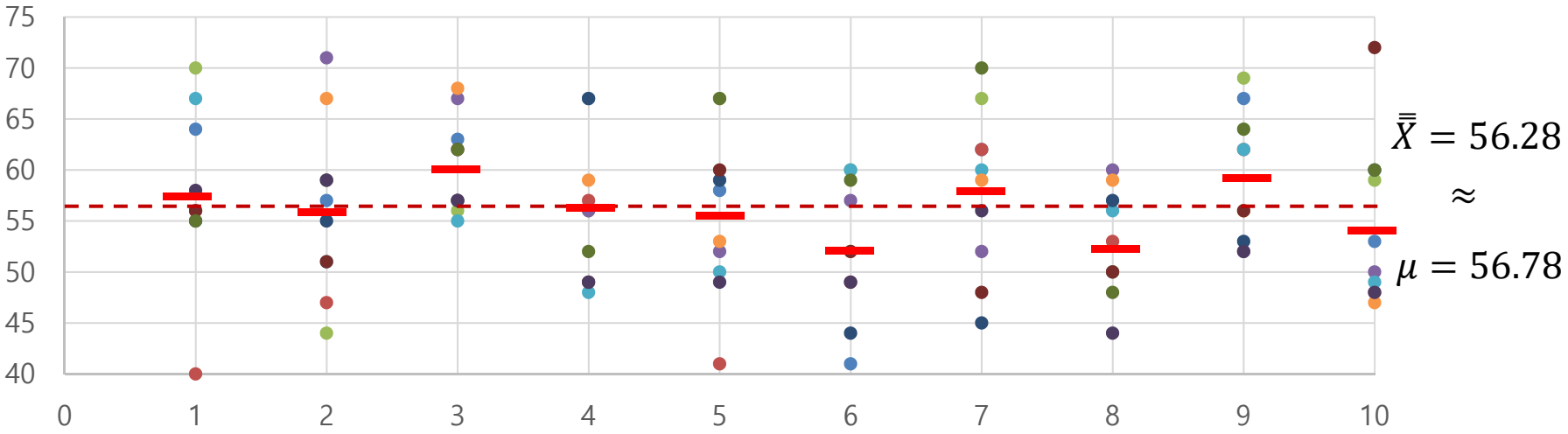
❖ 모집단(100명)에서 표본을 추출한다면



Sampling Distribution

❖ n=10일 때

	sample1	sample2	sample3	sample4	sample5	sample6	sample7	sample8	sample9	sample10	
1	64	57	63	67	58	41	62	50	67	53	
2	40	47	57	57	41	49	62	53	62	48	
3	70	44	56	56	67	60	67	50	69	59	
4	56	71	67	56	52	57	52	60	52	50	
5	67	51	55	48	50	60	60	56	62	49	
6	56	67	68	59	53	52	59	59	56	47	
7	55	55	57	67	59	44	45	57	53	60	
8	56	51	62	49	60	52	48	50	56	72	
9	55	59	62	52	67	59	70	48	64	60	
10	58	59	57	49	49	49	56	44	52	48	
$\bar{x}$	57.7	56.1	60.4	56.0	55.6	52.3	58.1	52.7	59.3	54.6	56.28
s	8.3	8.4	4.7	6.9	8.2	6.7	8.0	5.2	6.3	8.0	2.66



$\bar{\bar{X}} = 56.28$

$\approx$

$\mu = 56.78$

Sampling Distribution

❖ n=10일 때, SE($\sigma_{\bar{x}}$)

	sample1	sample2	sample3	sample4	sample5	sample6	sample7	sample8	sample9	sample10	
1	64	57	63	67	58	41	62	50	67	53	
2	40	47	57	57	41	49	62	53	62	48	
3	70	44	56	56	67	60	67	50	69	59	
4	56	71	67	56	52	57	52	60	52	50	
5	67	51	55	48	50	60	60	56	62	49	
6	56	67	68	59	53	52	59	59	56	47	
7	55	55	57	67	59	44	45	57	53	60	
8	56	51	62	49	60	52	48	50	56	72	
9	55	59	62	52	67	59	70	48	64	60	
10	58	59	57	49	49	49	56	44	52	48	
$\bar{x}$	57.7	56.1	60.4	56.0	55.6	52.3	58.1	52.7	59.3	54.6	56.28
s	8.3	8.4	4.7	6.9	8.2	6.7	8.0	5.2	6.3	8.0	2.66

$$E(\bar{X}) = \mu \qquad 56.28 \approx 56.78$$

$$\sigma_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^{n}(\bar{x}_1 - \mu)^2}{n-1}}$$

$$= \sqrt{\frac{(57.7 - 56.28)^2 + \cdots + (54.6 - 56.28)^2}{9}}$$

$$= \sqrt{7.05}$$

$$= 2.66$$

$$\approx \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$= \frac{6.80}{\sqrt{10}}$$

$$= 2.15$$

Sampling Distribution

❖ 표본의 구간

– 만약 모집단 평균이 μ이고 표준편차가 σ 라고 알고 있는 경우, 모집단에서 표본들이 추출될 95% 구간

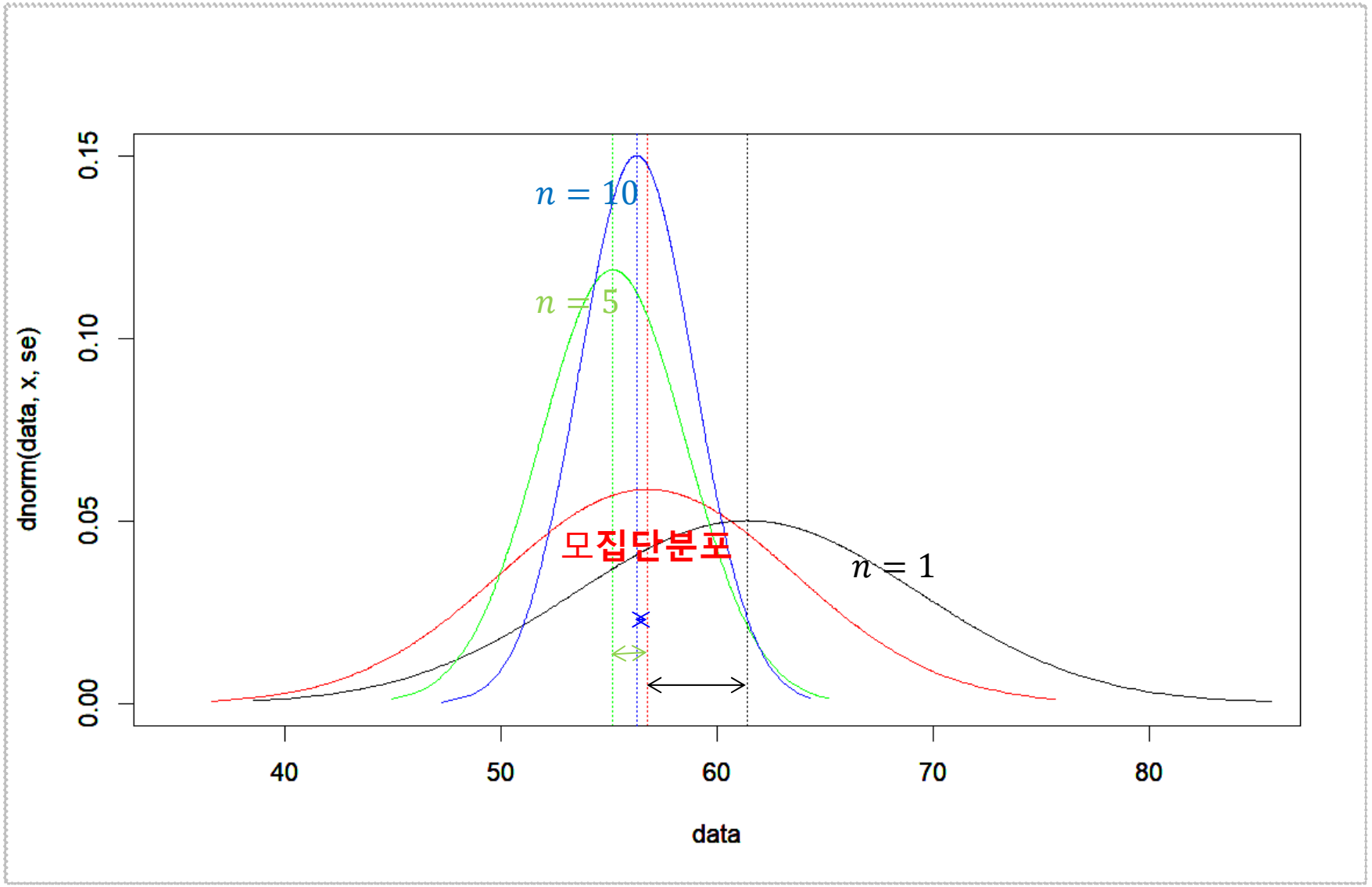
		Sample1	sample2	sample3	sample4	sample5	sample6	sample7	sample8	sample9	sample10	$\bar{\bar{X}}$	$\sigma_{\bar{x}}$
$n=1$	$\bar{X}$	59.0	59.0	53.0	50.0	72.0	68.0	53.0	70.0	69.0	61.0	61.40	7.96
$n=5$	$\bar{X}$	59.0	60.8	54.8	55.0	56.2	53.2	50.8	52.2	58.2	51.6	55.18	3.36
$n=10$	$\bar{X}$	57.7	56.1	60.4	56.0	55.6	52.3	58.1	52.7	59.3	54.6	56.28	2.66

$$n=1 \qquad 56.43 \pm 1.96\frac{7.41}{\sqrt{1}} = 56.43 \pm 14.52 \qquad [41.91, 70.95]$$

$$n=5 \qquad 56.43 \pm 1.96\frac{7.41}{\sqrt{5}} = 56.43 \pm 6.50 \qquad [49.93, 62.93]$$

$$n=10 \qquad 56.43 \pm 1.96\frac{7.41}{\sqrt{10}} = 56.43 \pm 4.59 \qquad [51.84, 61.02]$$

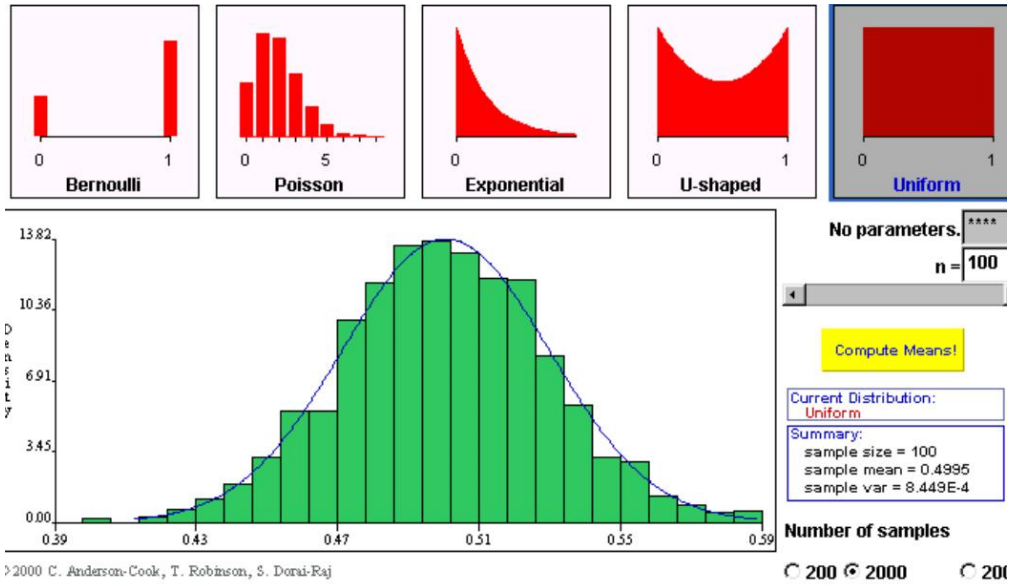
Sampling Distribution

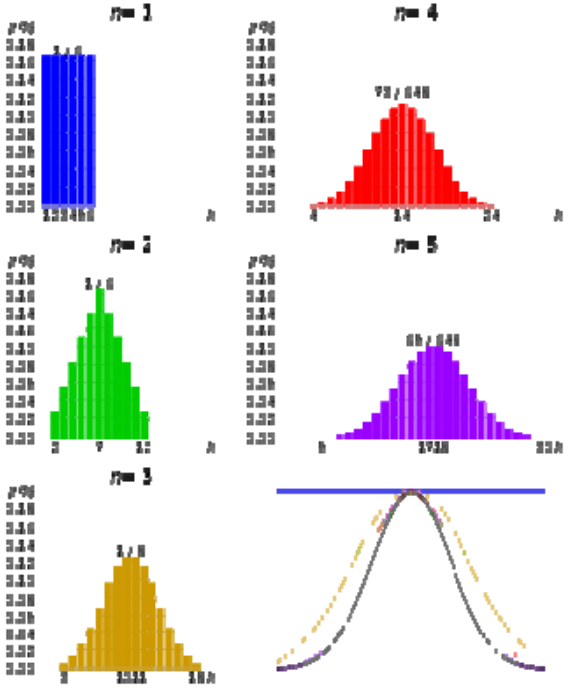


중심극한 정리

❖ 중심극한 정리(Central limit theorem)

- n 이 커질수록 모집단의 형태와 상관없이 표본분포($\bar{X}$)는 정규분포에 근사
- 평균이 μ이고 표준편차가 σ인 모집단으로 부터 추출된 표본의 개수가 n개라면, 표본의 표본분포는 평균이 $\boldsymbol{\mu}$이고 분산이 $\frac{\sigma^2}{n}$인 정규분포에 근사



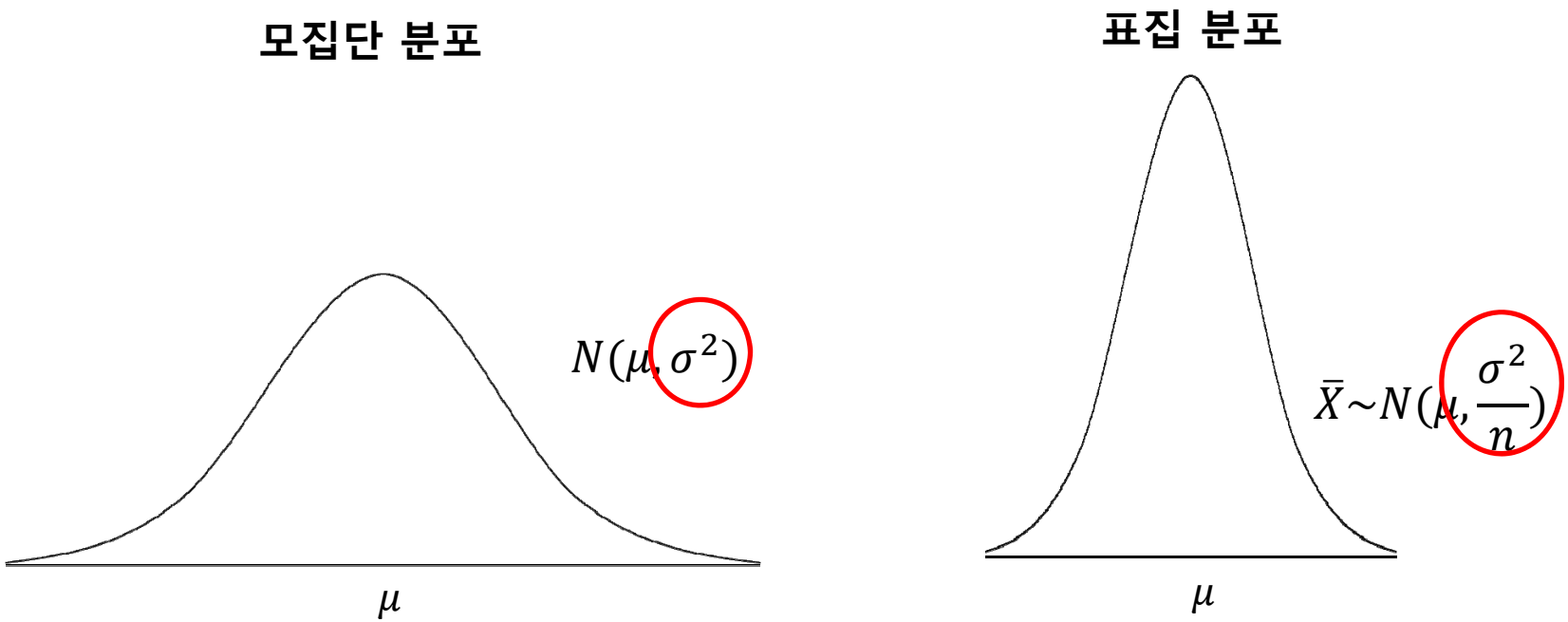


출처: https://serc.carleton.edu/sp/compadre/teachingwdata/Statcentral.html
출처: https://en.wikipedia.org/wiki/Central_limit_theorem

중심극한 정리

❖ 중심극한 정리(Central limit theorem)

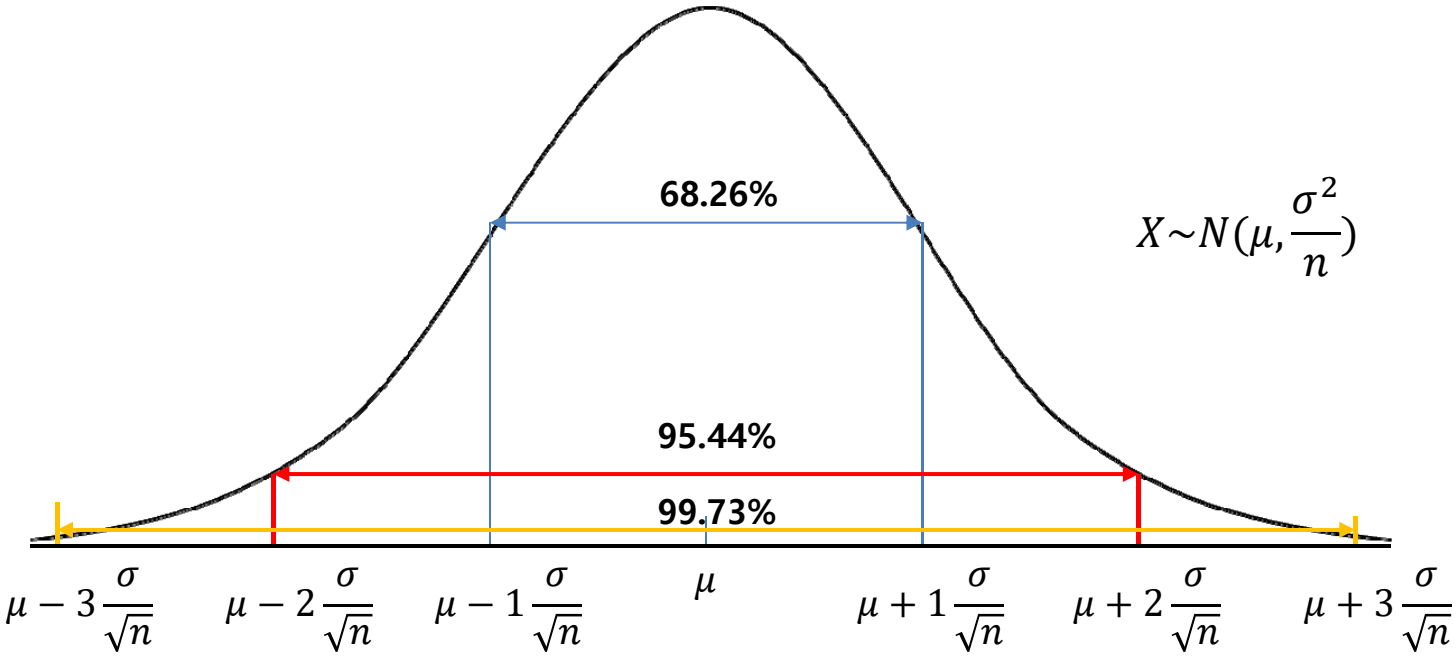
- n 이 커질수록 모집단의 형태와 상관없이 표본분포($\bar{X}$)는 정규분포에 근사
- 평균이 μ이고 표준편차가 σ인 모집단으로 부터 추출된 표본의 개수가 n개라면, 표본의 표본분포는 평균이 $\boldsymbol{\mu}$이고 분산이 $\frac{\sigma^2}{n}$인 정규분포에 근사



중심극한 정리

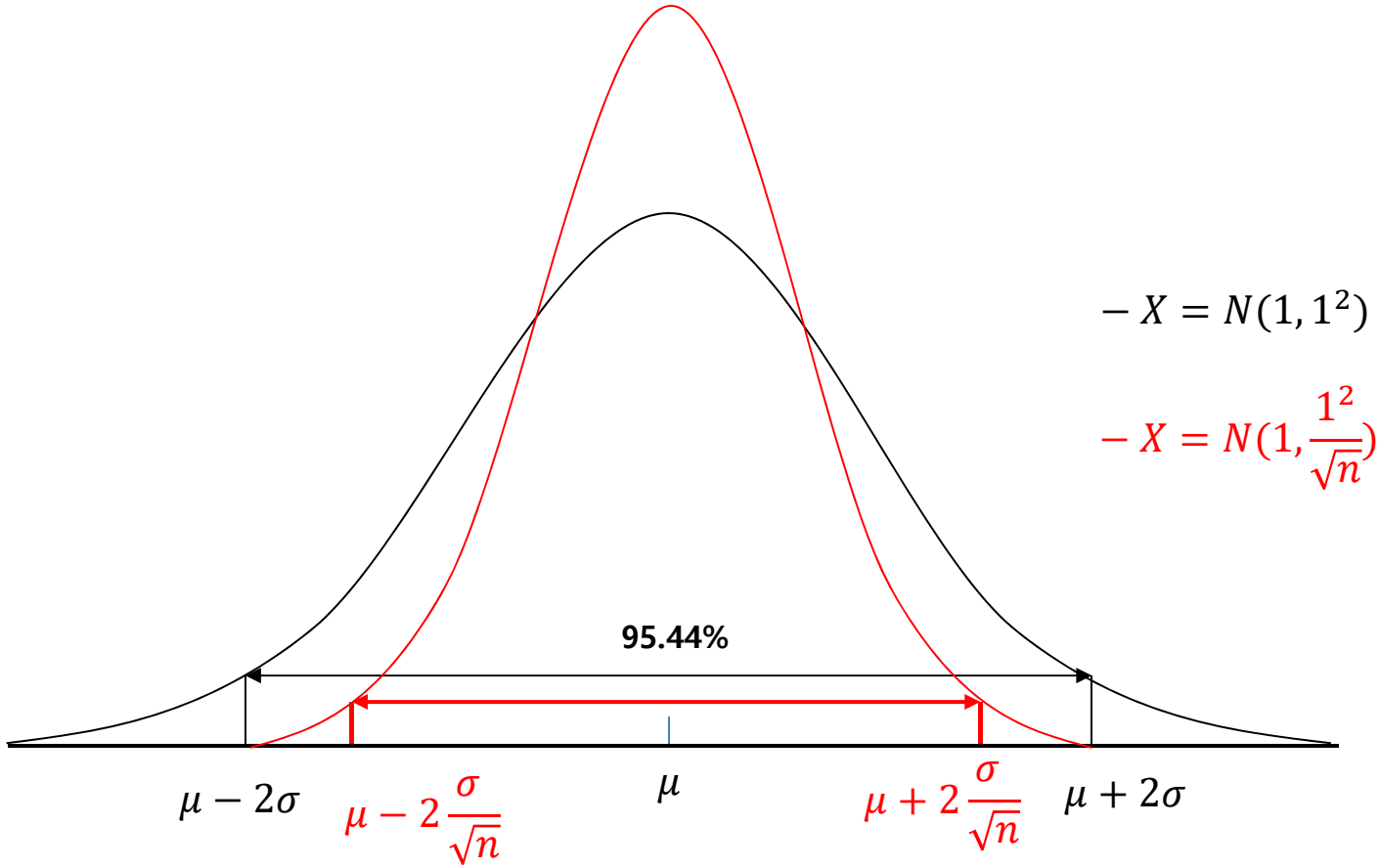
❖ Empirical Rule (경험적 법칙)

- $k = 1,\ 68.26\%$ 이상의 데이터가 $\mu \pm 1\frac{\sigma}{\sqrt{n}}$ 사이에 있음
- $k = 2,\ 95.44\%$ 이상의 데이터가 $\mu \pm 2\frac{\sigma}{\sqrt{n}}$사이에 있음
- $k = 3,\ \ 99.73\%$ 이상의 데이터가 $\mu \pm 3\frac{\sigma}{\sqrt{n}}$사이에 있음



중심극한 정리

- Empirical Rule에 적용
  - $k = 2,\ 95.44\%$ 이상의 데이터가 $\mu \pm 2\frac{\sigma}{\sqrt{n}}$사이에 있음
- 표본분포는 표준오차(SE)의 크기에 의해 오차범위의 크기가 결정됨



통계적 추론

통계적 추론의 종류

❖ 모수통계(parametric)

- 모집단의 분포가 정규분포임을 가정
- 중심극한 정리 : n이 30개 이상이면 정규분포 가능
- 정규분포 분석방법: t-test, ANOVA, 회귀분석 등

❖ 비모수통계(nonparametric)

- 모집단의 분포를 가정하지 않음
- 주로 n이 30개 미만일 경우, 이상치가 많을 경우, 척도가 순위척도일 경우
- 분석방법: Wilcoxon test, Mann-Whitney U test 등

원자료

분류	실험군	대조군
자료	13 15 10	6 12 8
합계	**38**	**26**

→

순위자료변환

분류	실험군	대조군
자료	5 6 3	1 4 2
합계	**14**	**7**

통계적 추론의 종류

❖ t-test와 대응되는 비모수 방법

독립		종속	모수	비모수
유형	집단	유형		
C	1	M	One-Sample T test	Wilcoxon test
C	2	M	Independent –Samples T-test	Mann-Whitney U test
			Paired Samples t-test	Wilcoxon test

❖ ANOVA와 대응되는 비모수 방법

독립		종속	모수	비모수
유형	수	유형		
C	1	M	One-Way ANOVA	Kruskal Wallis Test
	1	M	Repeated Measured ANOVA	Friedman Test
	2	M	Two-Way ANOVA	

통계적 추론의 종류

❖ 추정(Estimation)

- 표본의 평균과 표준오차(SE)를 구해서 모수의 범위를 구하는 것
- 신뢰구간: 일정한 확률범위 내에서 모수의 값이 포함될 가능성이 있는 범위
- 90%, 95%, 99% 의 확률 값 중에서 95%
- 종류: 점추정, 구간추정
- 사례: $\mu = 295.4 \pm 7.26, [288.14, 302.66]$

❖ 가설검정

- 모수는 얼마이다라고 정하고 그것이 맞는지 틀리는지를 검증하는 방법
- 유의수준 (α) : 모수와 통계량의 차이가 커서 확률적으로 가설을 기각할 수 있는 값
- 1%, 5%, 10%의 값 중에서 5%
- 종류: 귀무가설 (Null Hypothesis), 연구가설 (Alternative Hypothesis)
- 사례: $t_{cal} = -12.25 < t_{critical} = -1.984, p - value = 0.000 < \alpha = 0.05$

추정

점추정

❖ 점추정(Point estimation)

- 모수에 가장 가까울 것으로 생각되는 하나의 값으로 모수를 추정
- 추정량(확률변수, 측정전), 추정값(표본으로 부터 구한 측정값)
- 예) 모집단의 평균은 55kg이다

❖ 점추정 방법

- 적률법(method of moments): 기대값 이용
- 최대가능도(우도)추정법(maximum likelihood estimation): 조건부 확률
- 최소제곱법(least squares estimation): 회귀분석에서 사용

모수		점추정	
		추정량	추정값
평균	μ →	$\bar{X}$ →	$\bar{x}$
비율	θ	P	p
분산	σ^2	S^2	s^2
표준편차	σ	S	s

출처: 통계학(기본개념과 원리), 여인권, 자유아카데미

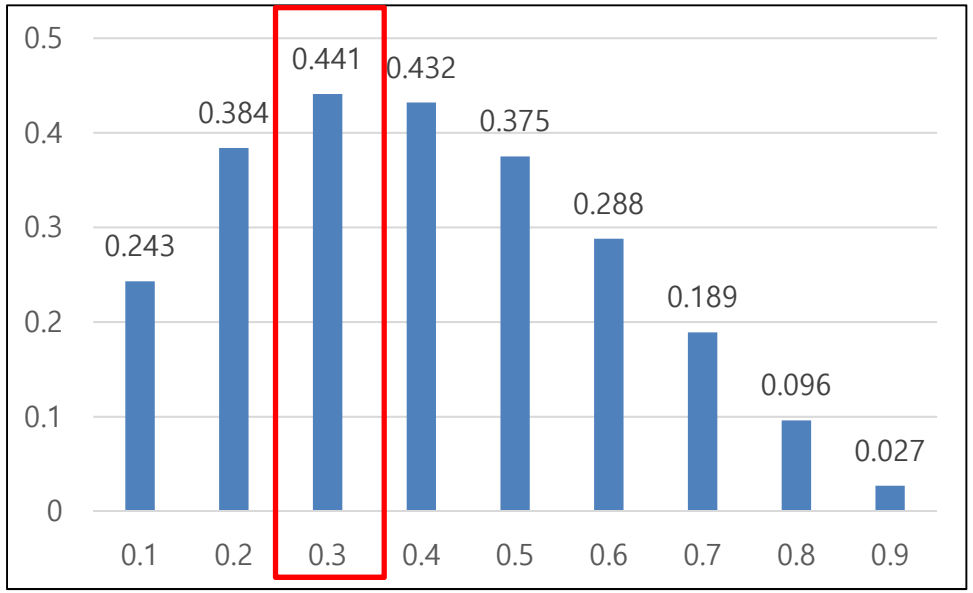
점추정

❖ 최대가능도(우도)추정법(maximum likelihood estimation)

- 결과가 주어졌을 때 속성값이 나올 확률 (사후 확률과 연계)
- fortune 쿠키 성공율이 90%일 때, 3명의 손님이 앉은 테이블에서 1명이 못 받을 확률은?

  → 3명의 손님 중 1명이 쿠키를 못 받았을 때 쿠키의 성공율은?

n	3			
x	0	1	2	3
0.1	0.729	0.243	0.027	0.001
0.2	0.512	0.384	0.096	0.008
0.3	0.343	0.441	0.189	0.027
0.4	0.216	0.432	0.288	0.064
0.5	0.125	0.375	0.375	0.125
0.6	0.064	0.288	0.432	0.216
0.7	0.027	0.189	0.441	0.343
0.8	0.008	0.096	0.384	0.512
0.9	0.001	0.027	0.243	0.729

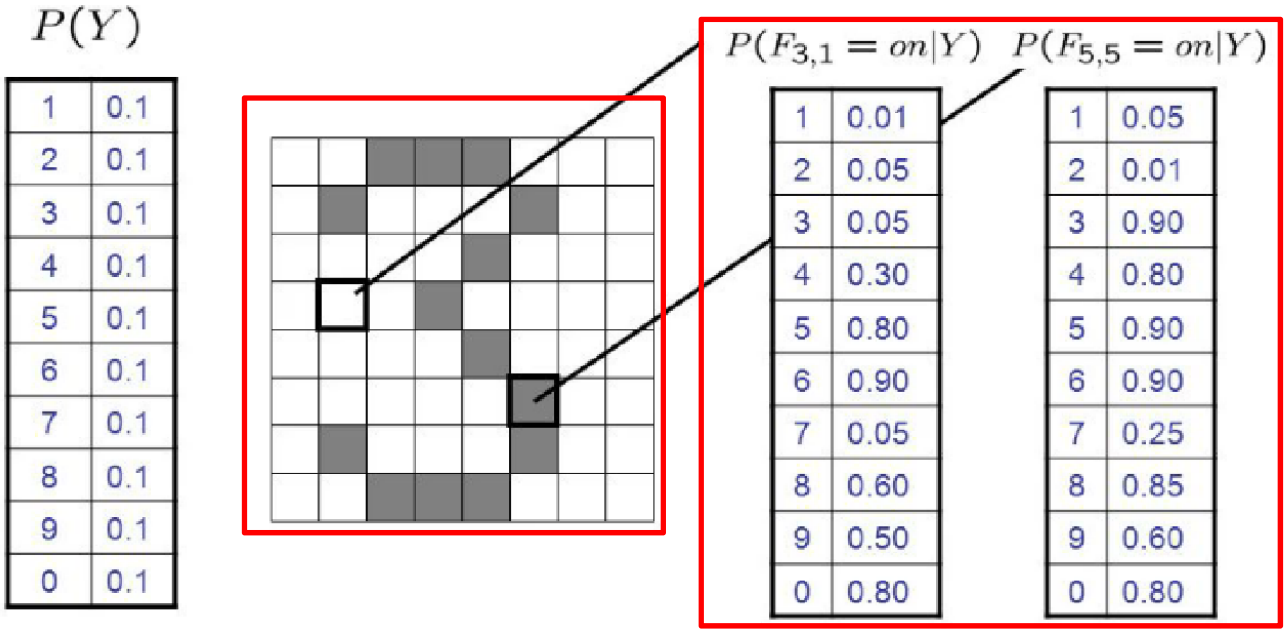


점추정

❖ 나이브베이즈 확률(데이터마이닝)

$$P(Y|F_{0,0},\cdots F_{7,7}) \propto P(Y)\prod_{i,j} P(F_{i,j}|Y) = P(Y)[P(F_{0,0}=on|Y)\dots P(F_{7,7}=on|Y)]$$

$$P(C_3)[\dots, P(F_{3,1}=on|C_3)\dots P(F_{5,5}=on|C_3)\dots] = 0.1\times[\dots, 0.05\times\cdots\times 0.9\times\cdots]$$



출처: CS 188: Artificial Intelligence Fall 2009, Dan Klein – UC Berkeley, https://slideplayer.com/slide/5120180/

구간추정

❖ 구간추정(Interval estimation)

- 모수의 값이 속할 것으로 기대되는 일정한 범위(신뢰구간)를 이용하여 모수를 추정
- 표본오차가 존재→ 추정 시에는 점추정 보다는 구간추정을 이용

❖ 신뢰구간 (Confidence Interval)

- 일반적으로 신뢰수준은 $(1-\alpha)$ 로 측정
- 신뢰수준의 확률: 90%, 95%, 99% 중에서 95%
- 신뢰구간 95%의 의미

  동일한 모집단에 대해서 동일한 방법으로 표본을 다시 뽑아서 신뢰구간을 구하게 되면 100번 중 95번은 모수를 포함 (표본분포의 의미)

❖ 방법

- (가정)모집단의 표준편차 σ를 알 경우 : 표준정규분포 $\mu = \bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
- (실제)모집단의 표준편차 σ를 모를 경우 : Student t 분포 $\mu = \bar{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$

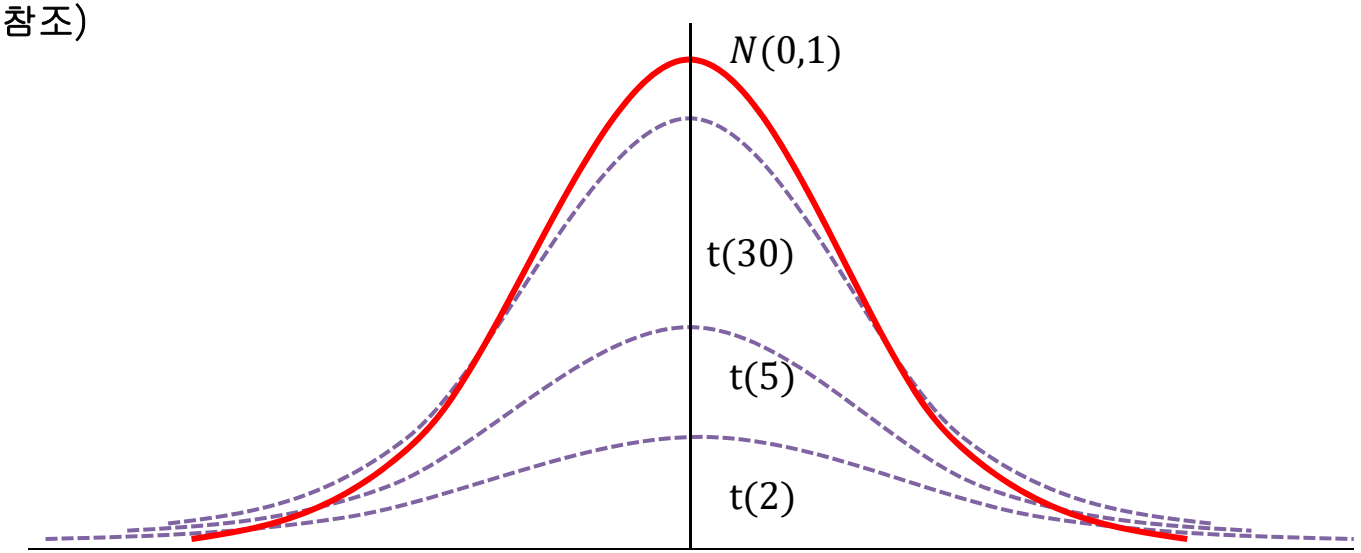
Student t 분포

❖ t분포(Student t Distribution)

- 정규분포나 표준정규분포: 모집단의 분산을 알고 있을 경우에 사용
- 실제로 우리는 모집단의 분산을 모르고 표본을 통해서 가설을 검정하므로 $\boldsymbol{t}$분포 이용
- 자유도(degree of freedom: $n-1$)에 의해서 모양이 결정
- 표본의 크기가 증가($\boldsymbol{n \geq 30}$)하면 분포는 정규분포에 접근
- 특징

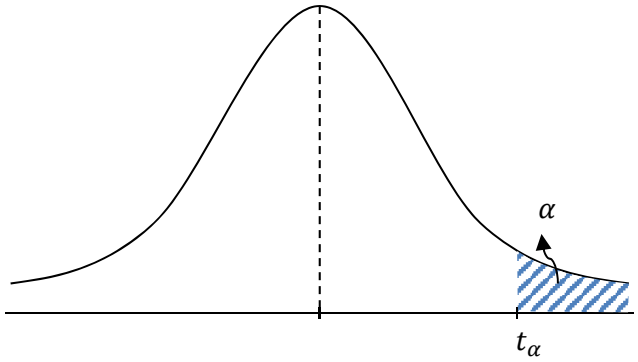
  정규분포처럼 종모양의 대칭분포

  (뒷장 그림참조)

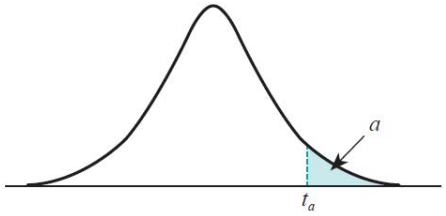


Student t 분포

❖ t분포표



4. t-분포표

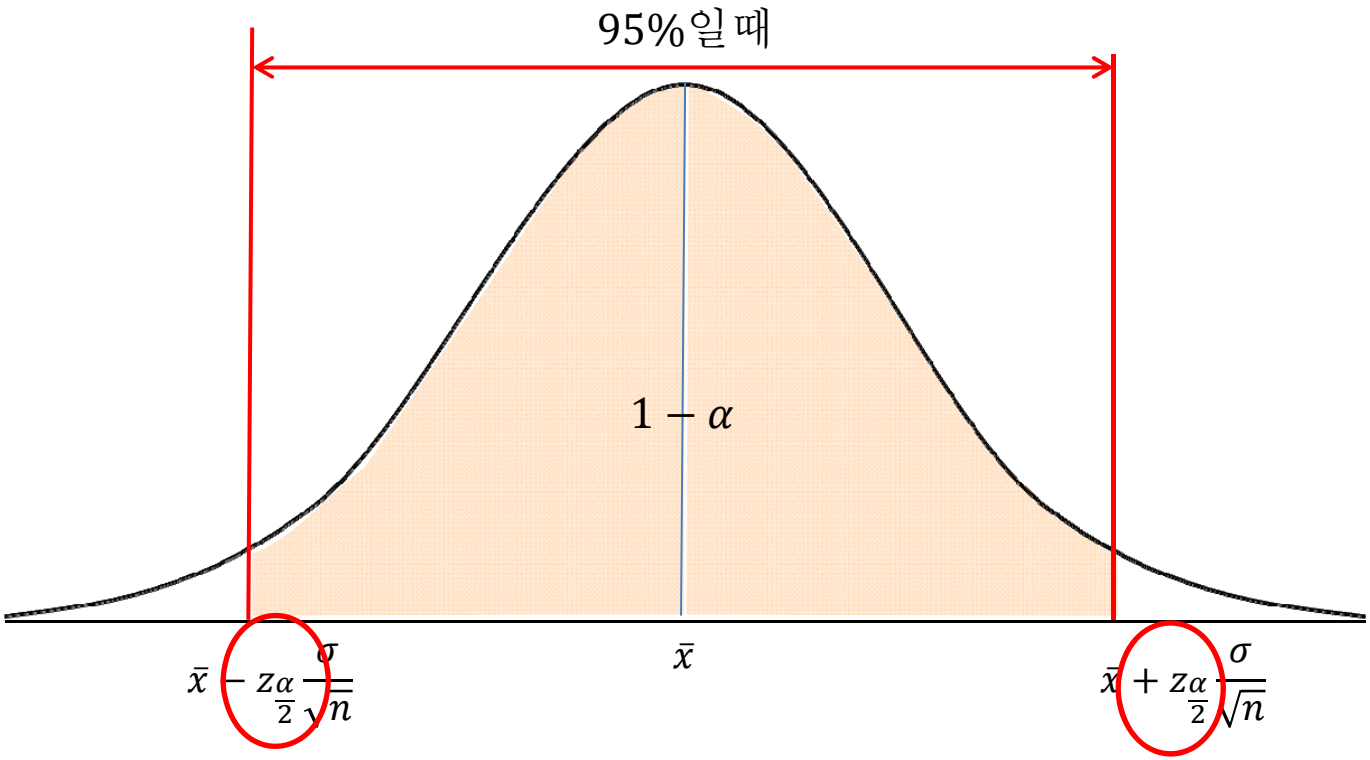


d.f.	$t_{.250}$	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.745	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.499
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.876	3.169
30	0.683	1.310	1.697	2.042	2.457	2.750
40	0.681	1.303	1.684	2.021	2.423	2.704
60	0.697	1.296	1.671	2.000	2.390	2.660
120	0.677	1.289	1.658	1.980	2.358	2.617
∞	0.674	1.282	1.645	1.960	2.326	2.576

구간추정

❖ 모집단의 표준편차 σ를 알 경우 : 표준정규분포

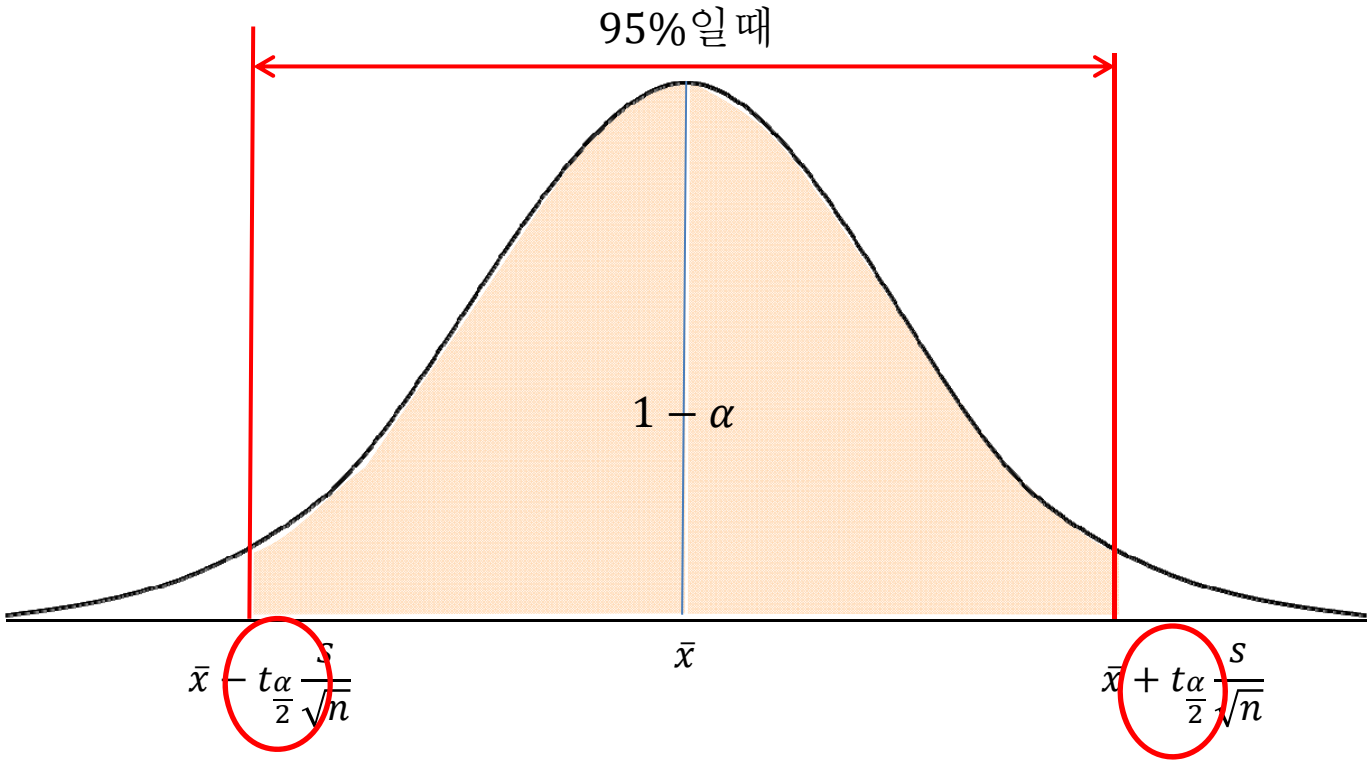
$$P\left(\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$



구간추정

❖ 모집단의 표준편차 σ를 모를 경우 : (Student) t분포

$$P\left(\bar{x} - t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$



구간추정(표준정규분포)

❖ 표준정규분포 사례)

- H자동차에서는 새로운 하이브리드 차량을 만들었다.
- 하이브리드 자동차의 평균연비는 정규분포를 따른다고 알려져 있고,
- 표준편차(σ)는 $1.0km/l$ 로 알려져 있다.
- 랜덤하게 표본 10개의 자동차를 추출해서 연비를 조사했더니 $16km/l$가 나왔다.
- 새로운 하이브리드 차량 연비의 95% 신뢰구간은?

$$P\left(\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$\mu = \bar{x} \pm z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

$$= 16 \pm 1.96\frac{1.0}{\sqrt{10}}$$

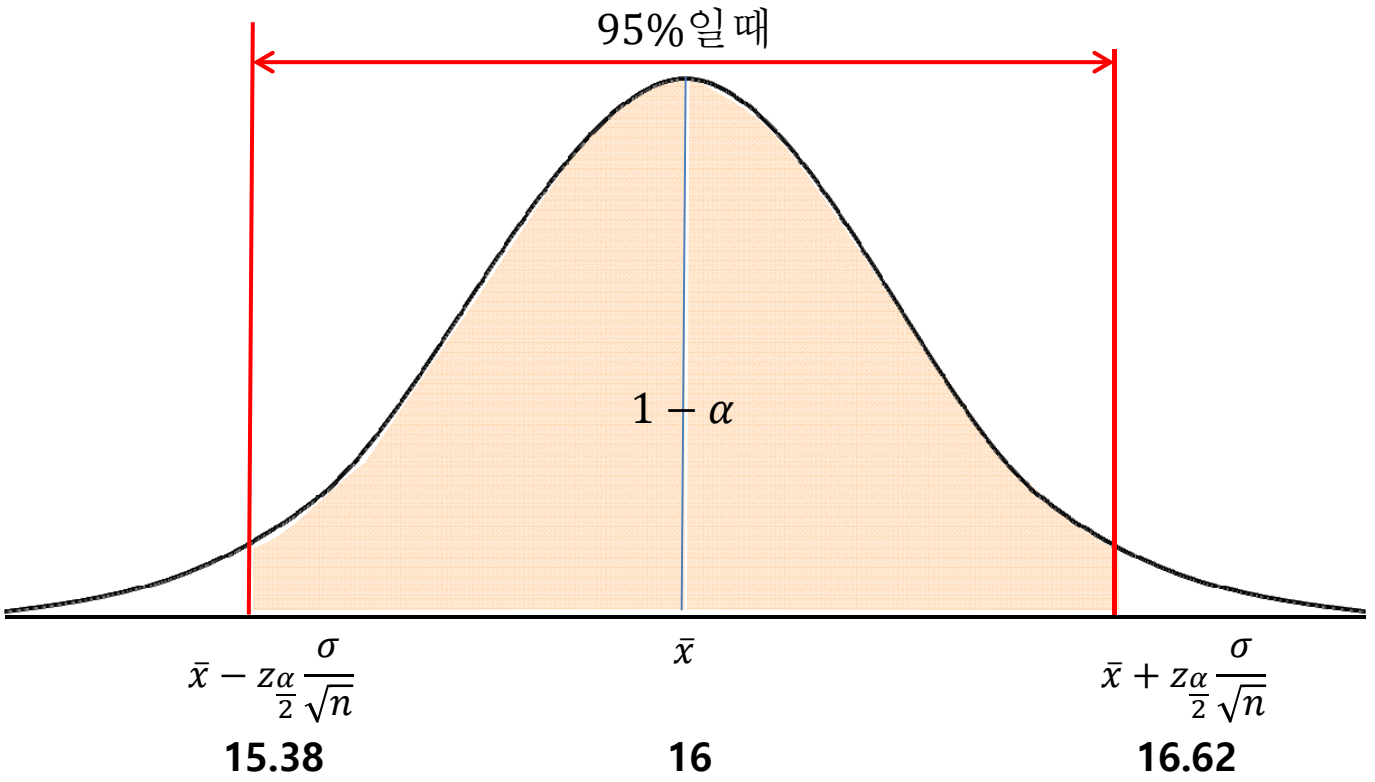
$$= 16 \pm 0.62$$

$$[15.38, 16.62]$$

구간추정(표준정규분포)

❖ 모집단의 표준편차 σ를 알 경우 : 표준정규분포

$$\left[\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] = [15.38, 16.62]$$



구간추정(t분포)

❖ t분포 사례)

- 새로운 하이브리드 자동차의 평균연비를 측정하기 위해 표본 100개를 추출하여 연비를 측정하였다.
- 표본(n) : 100
- 표본평균($\bar{X}$): $16km/l$
- 표본표준편차(s): 1.5, 표준오차 ($\frac{s}{\sqrt{n}}$): 0.15

$$P\left(\bar{x} - t_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$\mu = \bar{x} \pm t_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

$$= 16 \pm 1.984\frac{1.5}{\sqrt{100}}$$

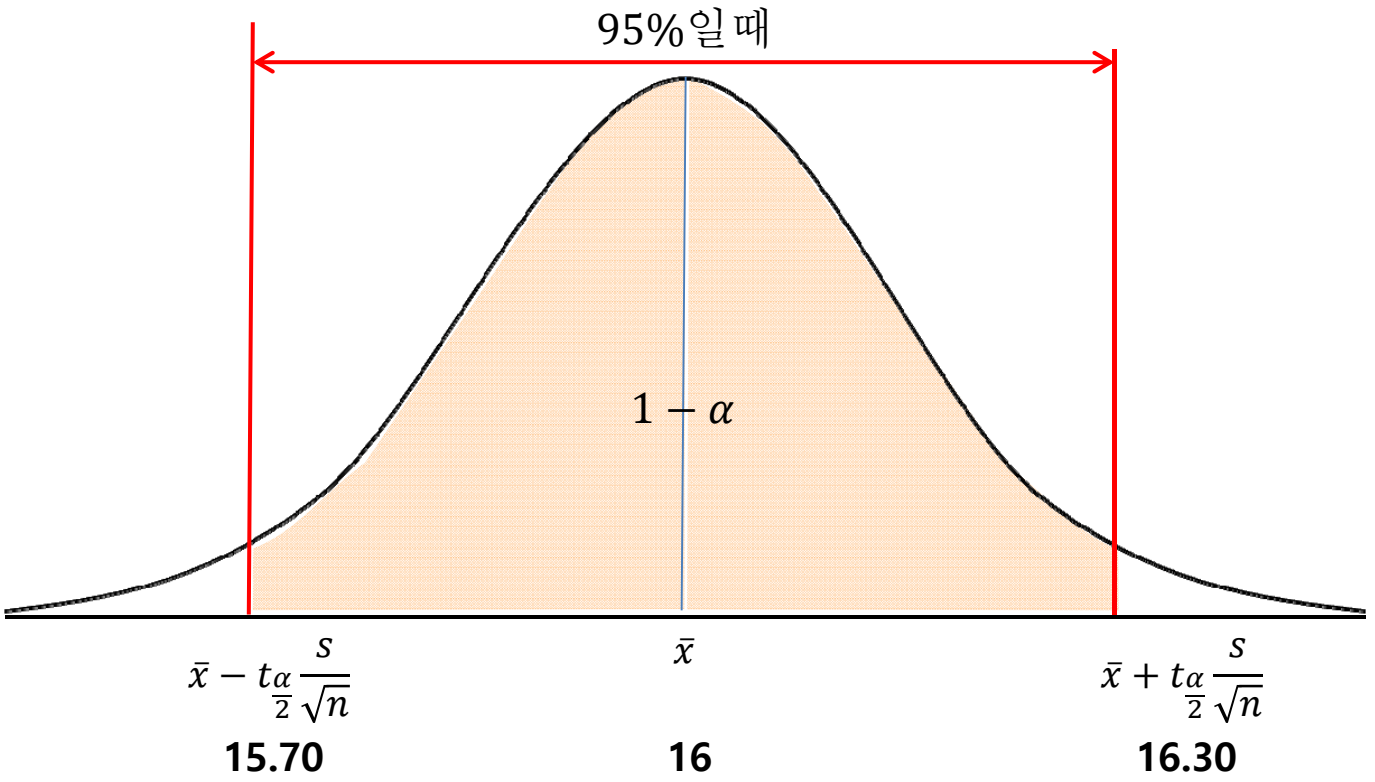
$$= 16 \pm 0.30$$

$$[15.70, 16.30]$$

구간추정(표준정규분포)

❖ 모집단의 표준편차 σ를 모를 경우 : (Student) t분포

$$\left[\bar{X} - t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right] = [15.70, 16.30]$$



모비율 추정

❖ 비율의 추정

- 비율의 평균

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

- 비율의 표준오차(중심극한정리)

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} \qquad \sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

- 비율의 정규분포 조건 (양쪽 표본 모두 5명 이상)

$$n\pi \geq 5 \; and \; n(1-\pi) \geq 5$$

* 만약, 조건이 충족되지 않으면 이항분포 또는 포아송 분포로

- 비율의 신뢰구간(정규분포일 경우)

$$\pi = p \pm z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

모비율 추정

❖ G텔레콤은 고객의 이탈율을 조사하기 위해 500명의 고객을 샘플로 이탈가능성을 조사하였다. 500명 중 50명이 앞으로 이탈할 것으로 나타났다. 이탈율의 신뢰구간은?

$$p = \frac{x}{n} = \frac{50}{500} = 0.100 \qquad np = 50 \times 0.1 = 5$$

$$\pi = p \pm z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

$$= 0.100 \pm 1.96 \sqrt{\frac{0.100(1-0.100)}{500}}$$

$$= 0.100 \pm 0.026$$

$$[0.074, 0.126]$$

가설검정

가설

❖ 가설검정(Hypothesis Test)

- 모집단 모수의 값을 설정하고 (가설설정), 표본 통계치를 통해 확률적으로 진위를 판정하는 과정

❖ 가설(Hypothesis)

- H_0: 귀무가설 (Null Hypothesis)

  기존에 알려져 있는 사실 (status quo), 통계적 검정 대상

- H_1: 대립가설 또는 연구가설 (Alternative Hypothesis)

  새로운 사실, 현재 믿음에 변화가 있는 사실, 뚜렷한 증거로 입증하려고 하는 주장

- 사례) H자동차에서 만든 새로운 하이브리드 차량의 연비는 $16.5km/l$로 알려져 있다. 과연 진짜로 $16.5km/l$ 인가를 검증
- 남자의 키와 여자의 키는 차이가 있는가?
- 주택면적과 가격은 관계가 있는가?

통계적 가설검정

❖ 통계적 가설검정 (Statistical Hypothesis Test)

- 통계적 가설 : 통계 검정을 위해서 사용하는 가설
- 일반적으로 검증할 모집단의 모수 (μ_0, π_0, σ_0)를 알고 있음
- H_0 은 현재 알려진 사실이기 때문에 명확히 가설로 진술
- 통계적 가설검정 : 모집단의 모수 (μ_0, π_0, σ_0)를 추측
- 통계적 가설검정: 귀무가설을 받아들일 것인지, 아니면 기각(reject)할 것인지를 검증
- 통계분석 귀무가설 설정(무죄주의 원칙)

구분	방법	귀무가설
차이검정	$t-test$, ANOVA	$H_0 : \mu_1 = \mu_2$
관계검정	regression	$H_0 : \beta = 0$

- 가설검정은 H_0가 진실인지를 검증함

가설검정의 오류

❖ 가설검정의 오류

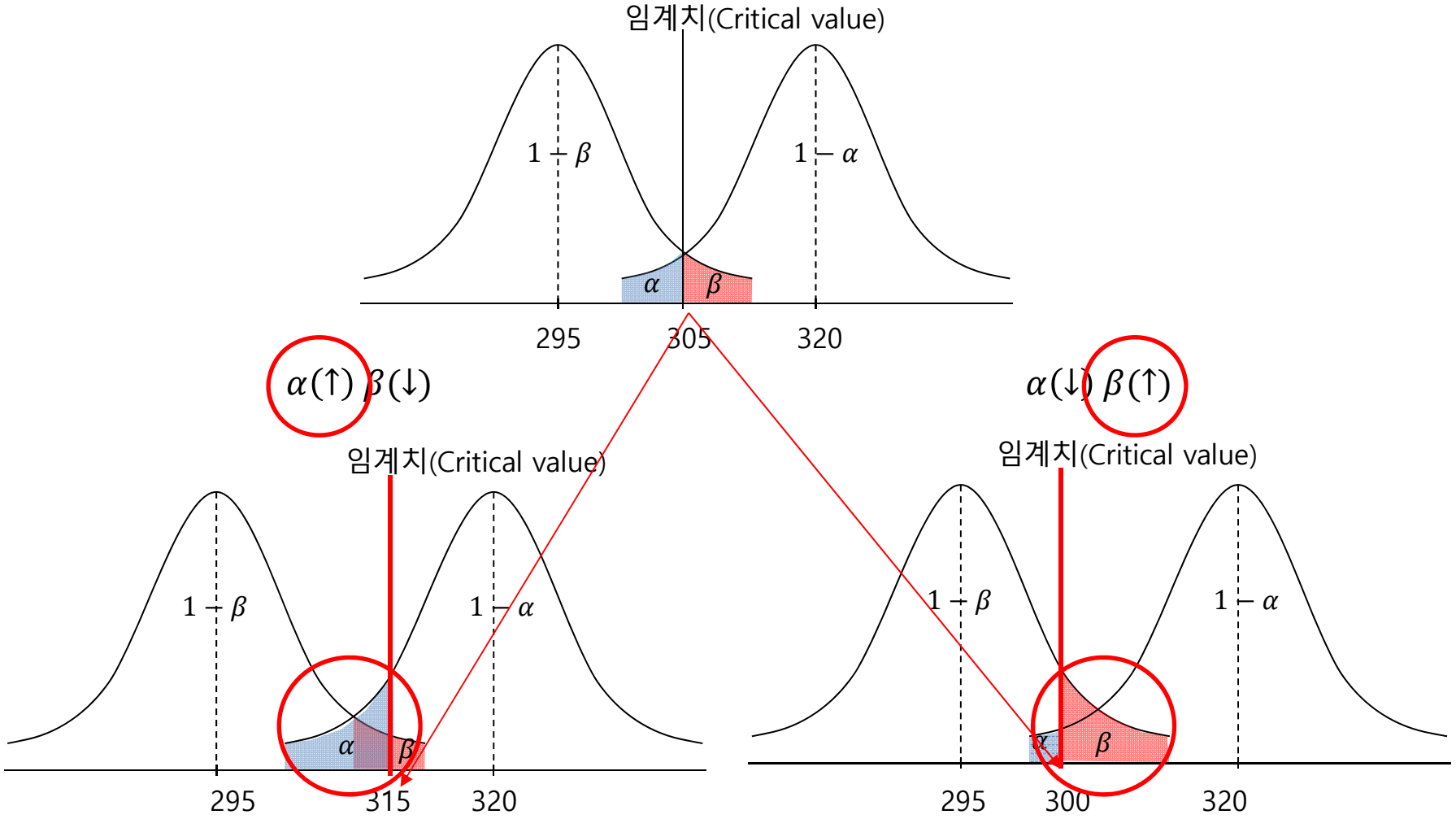
- 검사가 피고의 유죄를 증명하지 못했다고 진짜 피고가 무죄인가?
- Reject H_0 와 fail to reject H_0
- 100% 완벽한 증거(사실)을 수집할 수 없음
- 오류(error)가 존재
- 잘못된 의사결정을 내릴 수 있는 위험(Risk)를 결정해야 함

		실제	
		H_0진실	H_1진실
가설검정	H_0채택	옳은 결정(1-α)	제2종 오류(β)
	H_1채택	제1종 오류 (α)	옳은 결정(1-β)

- α, β 를 동시에 최소화 : 하나의 오류를 작게 하면 다른 오류가 커짐 (trade-off)

가설검정의 오류

❖ α와 β와의 관계



가설검정

❖ 유의수준(α)

- α를 이용한 검증 : H_0가 진실임을 증명
- H_0가 진실인데 H_0를 기각할 수 있는 오류를 범할 확률의 최대 허용치
- 0.01, 0.05, 0.10 중에서 0.05(5%)

❖ 검정력(Power)

- β를 이용한 검증 : H_1이 진실임을 증명
- H_1이 진실일 때 H_1를 채택할 수 있는 확률 : $(1-\beta)$
- 80%~95%

❖ 가설검정

- α는 현재사실을 근거로 검증하기 때문에 통계적 검정에서는 β를 통한 검증 보다는 $\boldsymbol{\alpha}$를 이용한 가설검정을 실시

가설검정

❖ 통계적 가설검정의 절차

- 가설 설정

  귀무가설(H_0), 연구가설(H_1)

  우측검정(right-sided test), 좌측검정(left-sided test), 양측검정(two-sided test)

- 유의수준(α) 및 임계치(critical value) 결정

  주로 0.05(5%)

  임계치 : 귀무가설을 기각하거나 채택할 수 있는 기준(z 또는 t)

- 검정통계량 (test statistics), 유의확률(p-value) 계산

- 임계치와 통계량 검정 및 결론도출

가설검정

❖ 임계치($\alpha = 0.05$일 때)

- 귀무가설(H_0)을 기각하거나 채택할 수 있는 기준

α	$z_{critical}$		
	양측검정	우측검정	좌측검정
0.10	± 1.645	1.282	-1.282
0.05	± 1.960	1.645	-1.645
0.01	± 2.576	2.326	-2.326

- 예) $z_{critical} = 1.96$ 일 때, $x_{critical} = \mu_0 \pm 1.96\frac{\sigma}{\sqrt{n}}$

가설검정

❖ 검정통계량 (test statistics)

- 귀무가설이 맞는지를 결정하기 위한 통계량

모표준편차 $\boldsymbol{\sigma}$알 경우($\mathbf{z}$)	모표준편차 모를 경우($\mathbf{t}$)
$z_{cal} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$	$t_{cal} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$

❖ 유의확률(p-value) 계산

- 귀무가설이 맞는다는 가정하에서 표본 통계량의 값이 나타날 확률
- 통계 패키지에서는 유의확률(p-value)을 계산해 줌

$$p - value = P(|z| \geq 1.96), \text{양측검정일때}$$

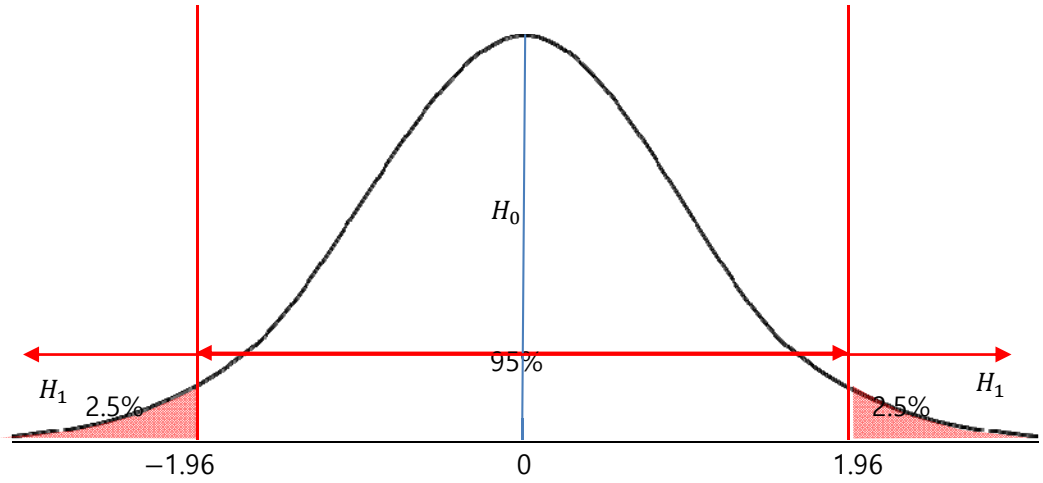
$$p - value = P(z \geq 1.96) + P(z \leq -1.96)$$

가설검정의 종류

❖ 가설검정 종류

귀무가설(H_0)	대립가설(H_1)	
$H_0: \mu = 16.5$	$H_1: \mu \neq 16.5$	양측검정(two-sided test)
	$H_1: \mu > 16.5$	우측검정(right-sided test)
	$H_1: \mu < 16.5$	좌측검정(left-sided test)

양측검정(two-sided test)



가설검정의 종류

❖ 가설검정 종류

귀무가설(H_0) 대립가설(H_1)

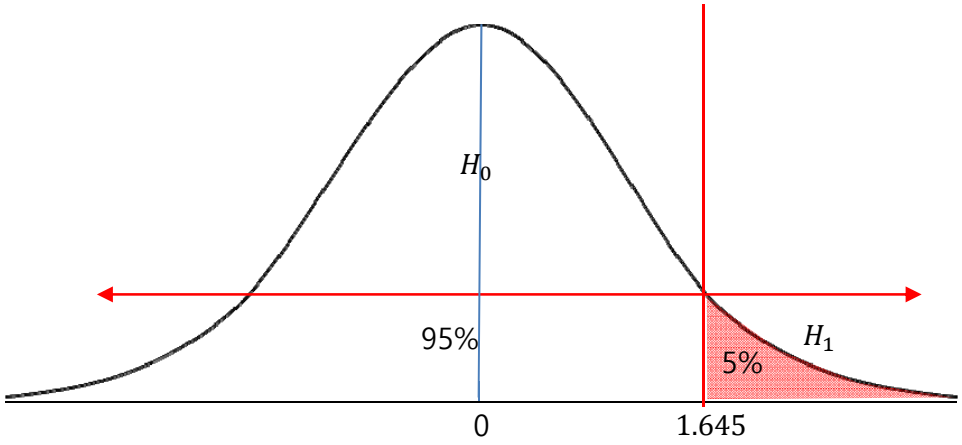
$H_0: \mu = 16.5$

$H_1: \mu \neq 16.5$ 양측검정(two-sided test)

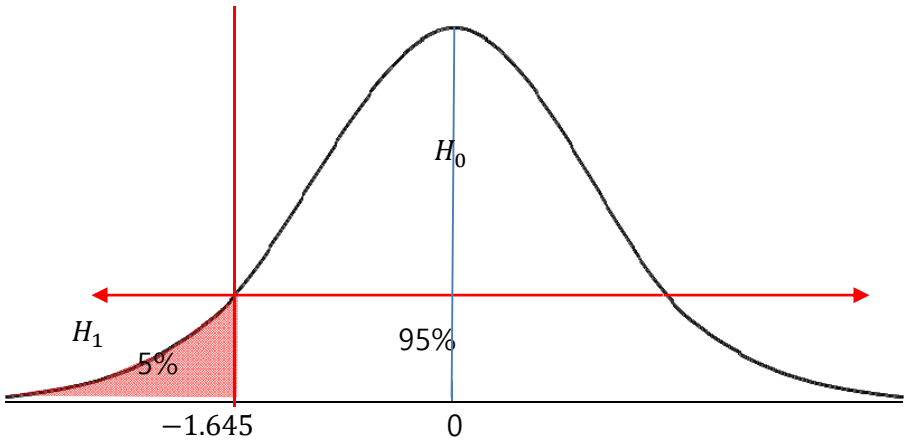
$H_1: \mu > 16.5$ 우측검정(right-sided test)

$H_1: \mu < 16.5$ 좌측검정(left-sided test)

H_1:우측검정(right-sided test)



H_1:좌측검정(left-sided test)



모평균 가설검정

❖ 모집단의 표준편차 σ를 알 경우 : 표준정규분포

- H자동차에서는 새로 개발한 하이브리드 차량의 평균연비는 $16.5km/l$ 이고, 표준편차는 $1.0km/l$로 알려져 있다.
- 표본 10개의 자동차를 추출해서 연비를 조사했더니 $16km/l$가 나왔다.
- 새로 개발한 하이브리드 차량의 평균연비는 5% 유의수준 내에서 $16.5km/l$ 라고 말할 수 있는가 ?

- 가설

귀무가설(H_0) 대립가설(H_1)

$H_0: \mu = 16.5$

$$\begin{cases} H_1: \mu \neq 16.5 & \text{양측검정(two-sided test)} \\ H_1: \mu > 16.5 & \text{우측검정(right-sided test)} \\ H_1: \mu < 16.5 & \text{좌측검정(left-sided test)} \end{cases}$$

- 임계치

$\alpha = 0.05$일 때, $z = -1.96$

모평균 가설검정

❖ 임계치

$$x_{critical} = \mu_0 - 1.96\frac{\sigma}{\sqrt{n}} = 16.5 - 1.96\frac{1.0}{\sqrt{10}} = 16.5 - 0.62 = 15.88$$

❖ 검정통계량 (test statistics)

$$z_{cal} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{16 - 16.5}{\frac{1.0}{\sqrt{10}}} = \frac{-0.5}{0.31} = -1.581$$

$$* \; z_{cal} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

❖ 유의확률(p-value) 계산

$$P(z \leq -1.581\ ) = 0.057$$

$$p - value = P(|z| \geq 1.581) = 0.114$$

$$* \; p - value = P(z \geq 1.96) + P(z \leq -1.96)$$

모평균 가설검정

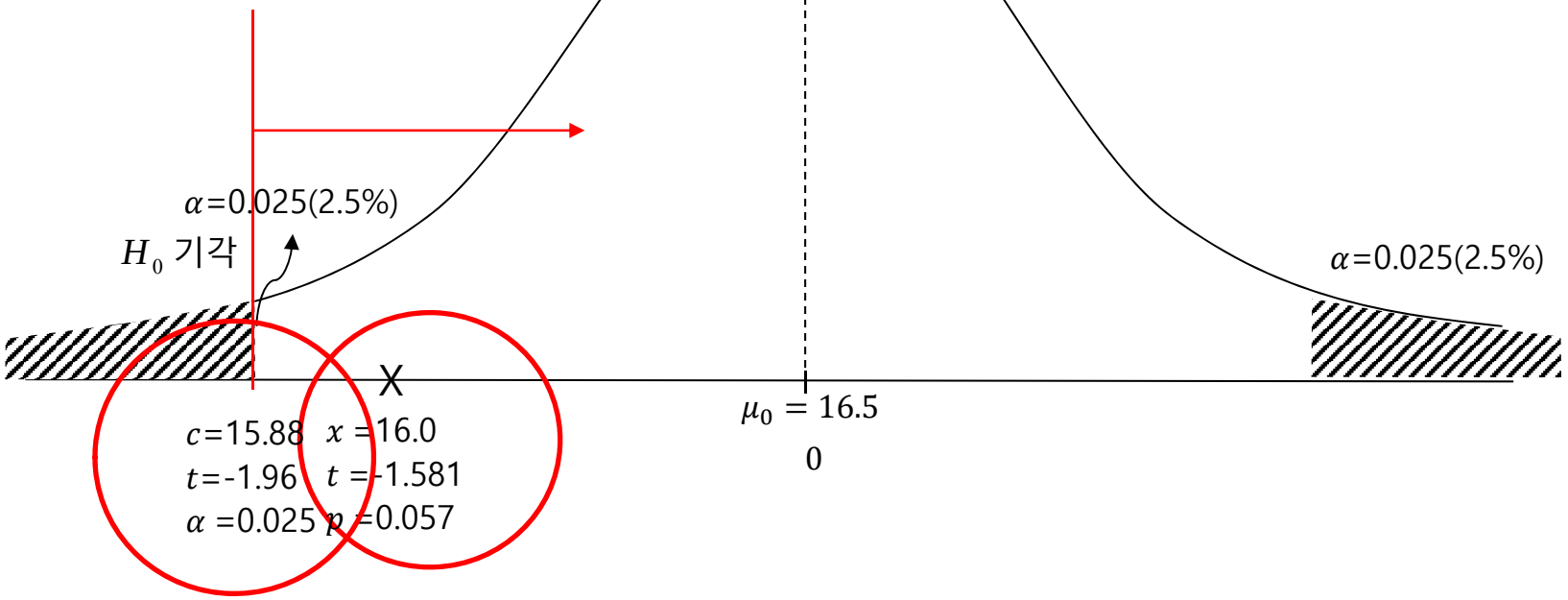
❖ 검정결과

검정통계량	임계치
$z_{cal} = -1.581 > z_{critical} = -1.96$	
$\bar{x} = 16.0 > x_{critical} = 15.88$	
$p - value = 0.114 > \alpha = 0.05$	

1.00
0.999
...
0.051
0.05
0.04
0.03
0.02
0.01
0.00

귀무가설 채택 $H_0: \mu = 16.5$

연구가설 채택 $H_1: \mu \neq 16.5$



모평균 가설검정

❖ 모집단의 표준편차 σ를 모를 경우 : t분포(Student t)

- 새로운 하이브리드 자동차의 평균연비를 측정하기 위해 표본 100개를 추출하여 연비를 측정하였다.
- 표본(n) : 100
- 표본평균$(\bar{X})$: $16km/l$
- 표본표준편차(s): 1.5, 표준오차 $(\frac{s}{\sqrt{n}})$: 0.15
- 가설

귀무가설(H_0)	대립가설(H_1)	
$H_0: \mu = 16.5$	$H_1: \mu \neq 16.5$	양측검정(two-sided test)
	$H_1: \mu > 16.5$	우측검정(right-sided test)
	$H_1: \mu < 16.5$	좌측검정(left-sided test)

- 임계치

$n = 100, \alpha = 0.05$일 때, $t = \pm 1.984$

모평균 가설검정

❖ 임계치

$$x_{critical} = \mu_0 - 1.984\frac{s}{\sqrt{n}} = 16.5 - 1.984\frac{1.5}{\sqrt{100}} = 16.5 - 0.30 = 16.20$$

❖ 검정통계량 (test statistics)

$$t_{cal} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{16 - 16.5}{\frac{1.5}{\sqrt{100}}} = \frac{-0.5}{0.15} = -3.333$$

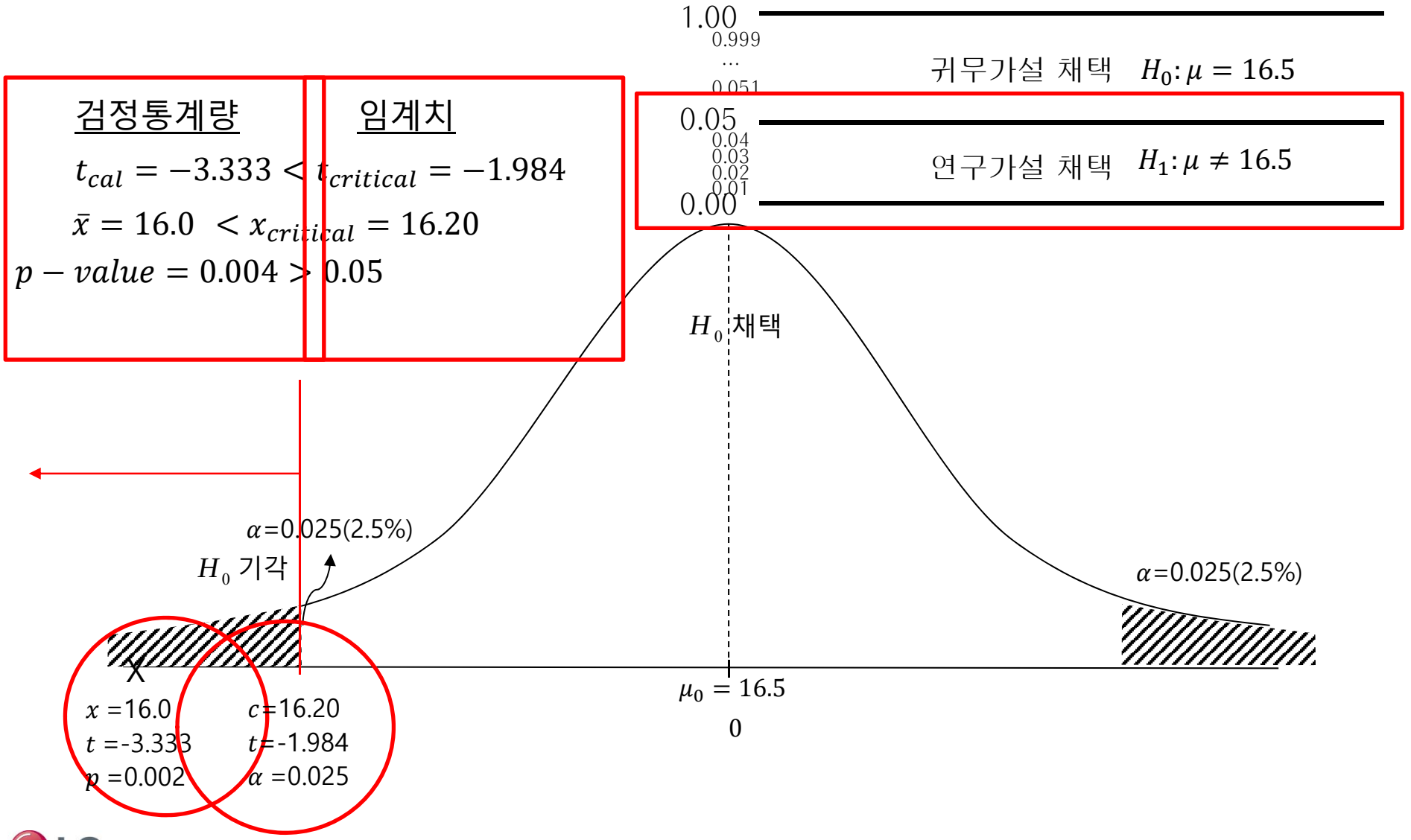
$$* \; t_{cal} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

❖ 유의확률(p-value) 계산

$$p - value = P(|t| \geq 3.333) = 0.004$$

모평균 가설검정

❖ 검정결과



검정력

Power

❖ 검정력(Power)

- H_1이 참 일때 H_1을 채택할 수 있는 확률 : $(1-\beta)$
- 실험결과 H_0이 기각되지 않았을 때 표본수의 문제인지 검정할 때 사용하며, 검정력을 이용해 새로운 적정 표본수 계산
- 의료분야에서는 질병을 정확하게 진단할 수 있는 검사의 검정력을 민감도(sensitivity)라고 함
- 새로운 연구 결과에 대한 추가 검정, 새로운 약의 효과 등과 같이 중요한 사안에 대한 검정

$$\bar{x}_{critical} = \mu_0 \pm z_{critical}\frac{\sigma}{\sqrt{n}}$$

$$z_p = \frac{\bar{x}_{critical} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$\beta = P(\bar{X} > \bar{x}_{critical} | \mu = \mu_1)$$

$$Power = P(\bar{X} < \bar{x}_{critical} | \mu = \mu_1)$$
$$= 1 - \beta$$

❖ 하이브리드 차량의 평균연비 사례

- 평균 (μ_0): $16.5km/l$
- 표준편차 (σ): $1.0km/l$
- 표본수 (n): 10
- 표본평균 ($\bar{x}$): $16km/l$
- 가설

귀무가설(H_0) 대립가설(H_1)

$$H_0: \mu = 16.5$$

$$\begin{cases} H_1: \mu \neq 16.5 & \text{양측검정(two-sided test)} \\ H_1: \mu > 16.5 & \text{우측검정(right-sided test)} \\ H_1: \mu < 16.5 & \text{좌측검정(left-sided test)} \end{cases}$$

- 임계치

$\alpha = 0.05$일 때, $z = -1.96$

Power

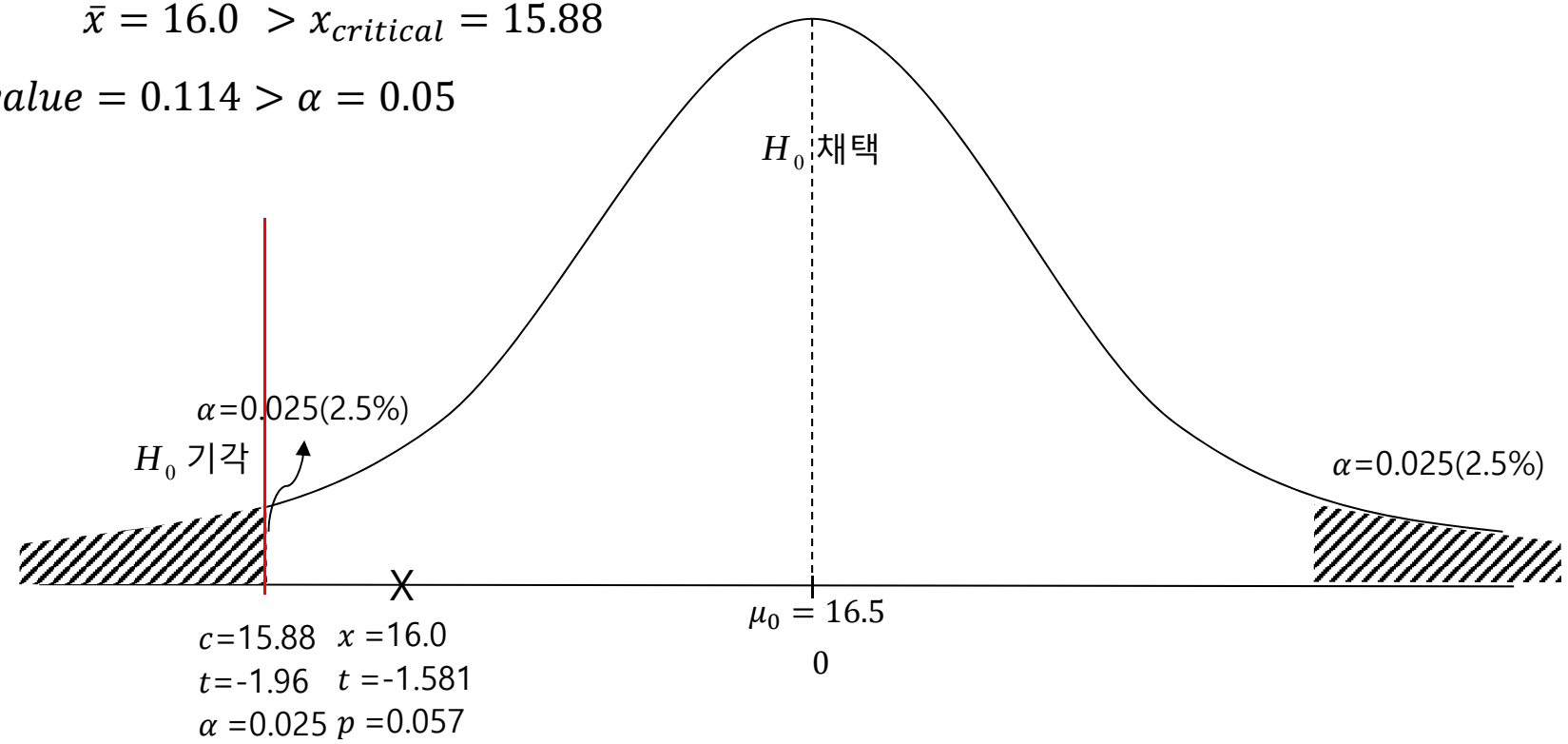
❖ 검정결과

검정통계량 임계치

$z_{cal} = -1.581 > z_{critical} = -1.96$

$\bar{x} = 16.0 > x_{critical} = 15.88$

$p - value = 0.114 > \alpha = 0.05$



Power

❖ 검정력$(1-\beta)$, $\bar{x}=\mu_1$이면

❖ 임계치

$$x_{critical}=\mu_0-1.96\frac{\sigma}{\sqrt{n}}=16.5-1.96\frac{1.0}{\sqrt{10}}=16.5-0.62=15.88$$

❖ 검정통계량 (test statistics)

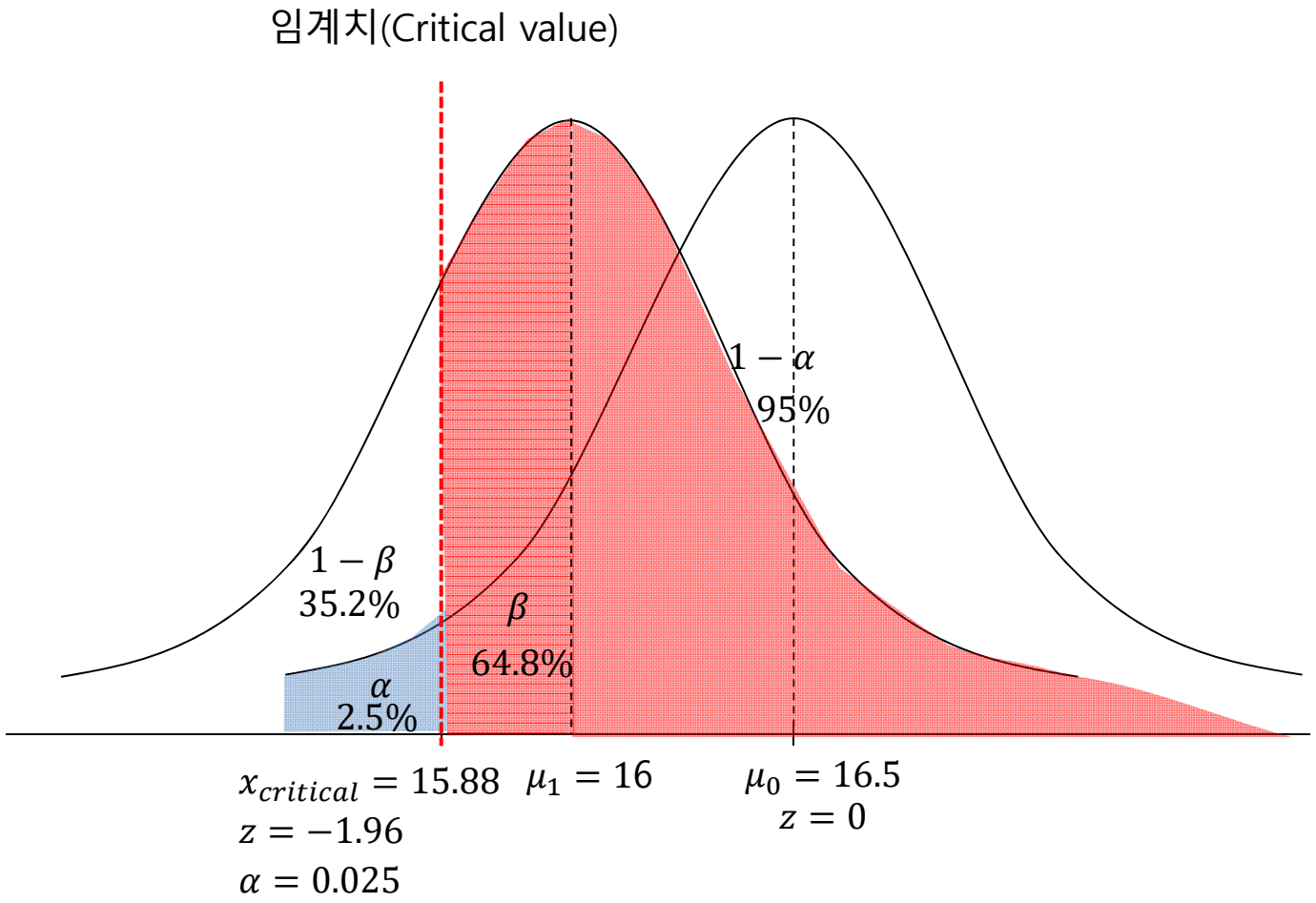
$$z_\beta=\frac{\bar{x}_{critical}-\mu_1}{\frac{\sigma}{\sqrt{n}}}=\frac{15.88-16}{\frac{1.0}{\sqrt{10}}}=\frac{-0.12}{0.316}=-0.380$$

❖ Power 계산

$$\begin{aligned}\beta&=P(\bar{X}>\bar{x}_{critical}|\mu=16)\\&=P(z>-0.380)\\&=0.5+0.148=0.648\end{aligned}$$

$$\begin{aligned}Power&=P(\bar{X}<\bar{x}_{critical}|\mu=16)\\&=1-\beta\\&=1-0.648\\&=0.352\end{aligned}$$

Power


임계치(Critical value)
$1-\alpha$
95%
$1-\beta$
35.2%
β
64.8%
α
2.5%
$x_{critical} = 15.88$ $\mu_1 = 16$
$z = -1.96$
$\alpha = 0.025$
$\mu_0 = 16.5$
$z = 0$

Power

❖ 만약 H_1과 H_0가 멀 다면,

- 평균 (μ_0): $16.5km/l$
- 표준편차 (σ): $1.0km/l$
- 표본수 (n): 10
- 표본평균 ($\bar{x}$): $15km/l(16 \rightarrow 15)$
- 임계치

$$x_{critical} = \mu_0 - 1.96\frac{\sigma}{\sqrt{n}} = 16.5 - 1.96\frac{1.0}{\sqrt{10}} = 16.5 - 0.62 = 15.88$$

- 검정통계량 (test statistics)

$$z_\alpha = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{15 - 16.5}{\frac{1.1}{\sqrt{10}}} = \frac{-1.5}{0.316} = -4.743$$

- 유의확률(p-value) 계산

$$p - value = 0.000$$

Power

❖ 검정력$(1-\beta)$, $\bar{x}=\mu_1$이면

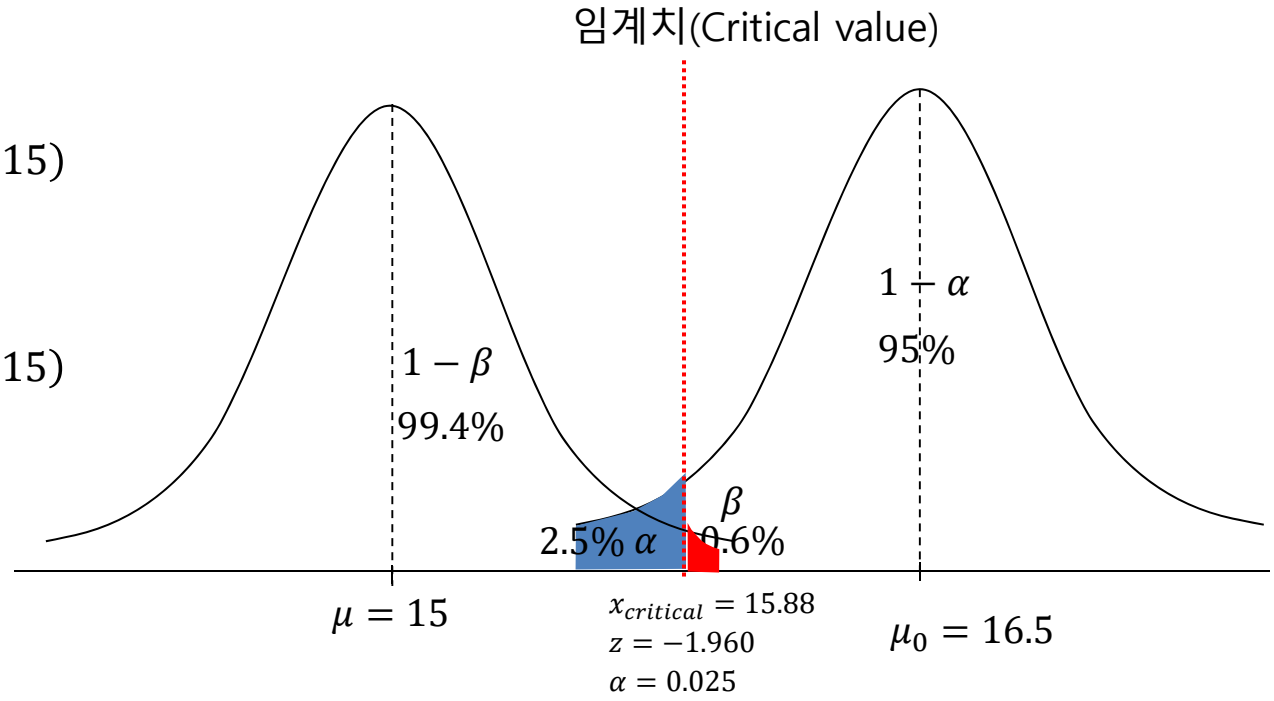
❖ 검정통계량 (test statistics)

$$z_\beta = \frac{\bar{x}_{critical}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{15.88-15}{\frac{1.0}{\sqrt{10}}} = \frac{0.88}{0.316} = 2.783$$

❖ Power 계산

$$\begin{aligned}\beta &= P(\bar{X} > \bar{x}_{critical} | \mu = 15)\\ &= P(z > 2.783)\\ &= 0.006\end{aligned}$$

$$\begin{aligned}Power &= P(\bar{X} < \bar{x}_{critical} | \mu = 15)\\ &= 1-\beta\\ &= 1-0.006\\ &= 0.994\end{aligned}$$



* 만약, 표본수가 작아서(10개) 이런 결과가 나온 것은 아닌지?

Power

❖ 표본수에 따른 영향

- 평균 (μ_0): $16.5km/l$
- 표준편차 (σ): $1.0km/l$
- 표본수 (n): 100
- 표본평균 ($\bar{x}$): $16km/l$
- 임계치

$$x_{critical} = \mu_0 - 1.96\frac{\sigma}{\sqrt{n}} = 16.5 - 1.96\frac{1.0}{\sqrt{100}} = 16.5 - 0.20 = 16.30(\leftarrow 15.88)$$

- 검정통계량 (test statistics)

$$z_{cal} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{16 - 16.5}{\frac{1.0}{\sqrt{100}}} = \frac{-0.5}{0.10} = -5.000$$

- 유의확률(p-value) 계산

$$p - value = 0.000$$

Power

❖ 검정력$(1-\beta)$, $\bar{x}=\mu_1$이면

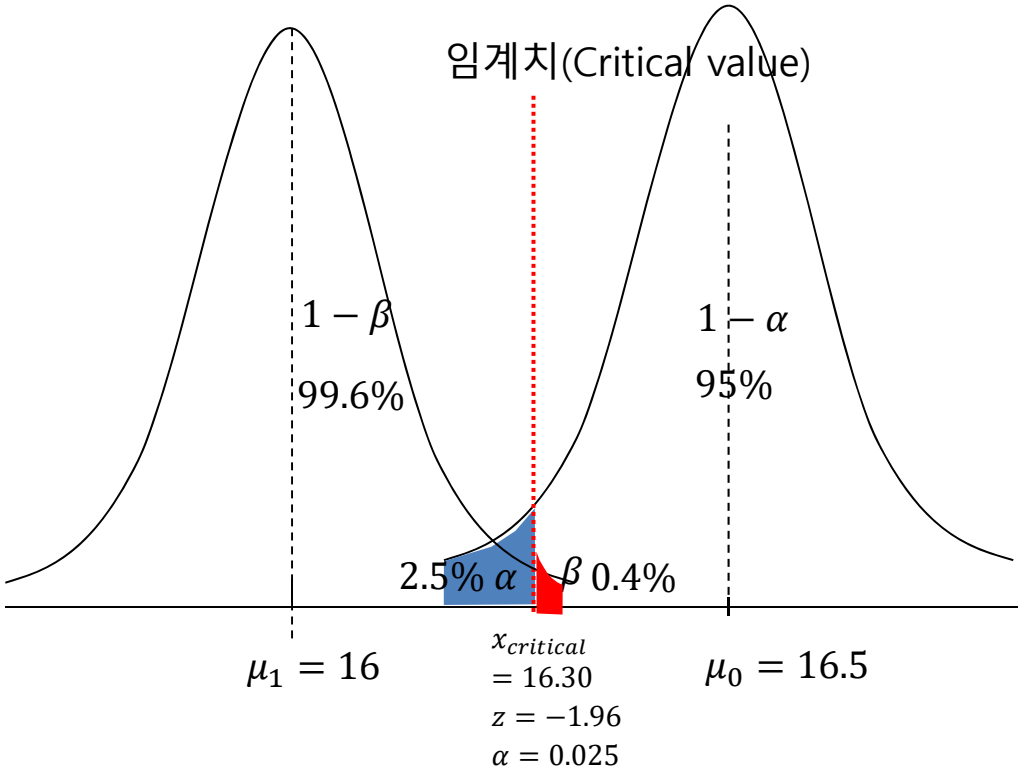
❖ 검정통계량 (test statistics)

$$z_p = \frac{\bar{x}_{critical} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{16.30 - 16}{\frac{1.0}{\sqrt{100}}} = \frac{0.30}{0.10} = 3.000$$

❖ Power 계산

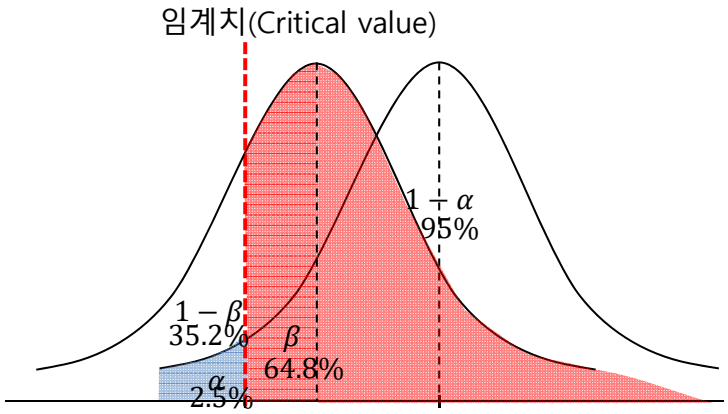
$$\begin{aligned}\beta &= P(\bar{X} > \bar{x}_{critical} | \mu = \mu_0) \\ &= P(z > 3.000) \\ &= 0.004\end{aligned}$$

$$\begin{aligned}Power &= P(\bar{X} < \bar{x}_{critical} | \mu = \mu_0) \\ &= 1 - \beta \\ &= 1 - 0.004 \\ &= 0.996\end{aligned}$$

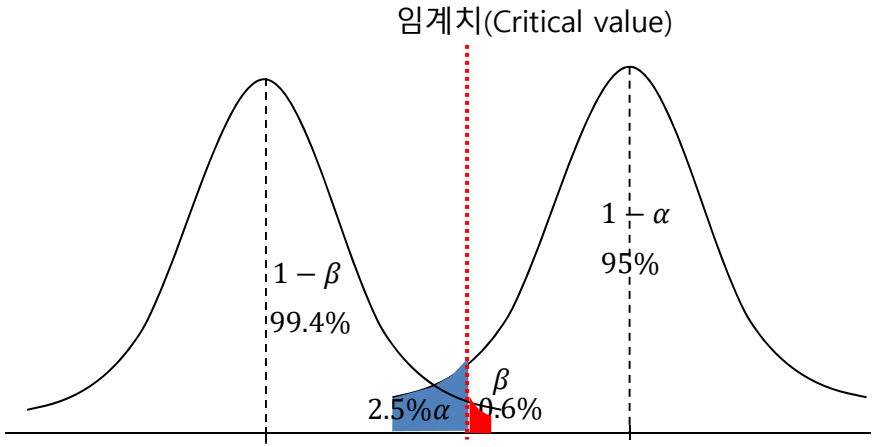


Power

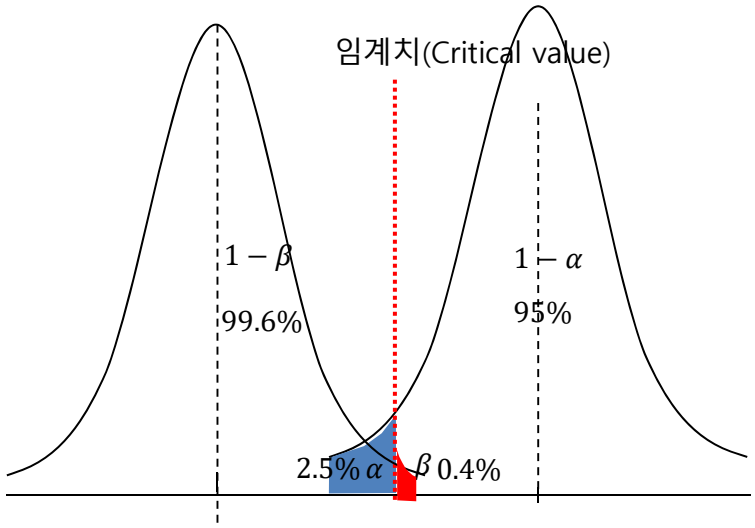
❖ 표본평균과 표본에 따른 변화



표본평균 $(\bar{x})$: $15\text{km/l}(16 \rightarrow 15)$



표본수 (n): 10 → 100



Power

❖ 검정력(power)에 영향을 미치는 요인

$$\bar{x}_{critical} = \mu_0 \pm z_{critical}\frac{\sigma}{\sqrt{n}}$$

- 표본수(n)에 민감 : 사전연구를 통해 연구결과의 신뢰성을 높일 수 있도록 (검정력을 높일 수 있는) 표본크기 계산
- 유의수준($\alpha, z_{critical}$) : 유의수준을 낮게 하면 H_0를 채택할 확률은 높아지지만 반대로 H_1을 기각할 확률도 높아짐. 실제로 의미 있는 결과를 이용하지 못할 수도 있음
- 따라서 의학연구나 실험연구의 경우에 사전연구결과와 검정력을 이용해서 의미 있는 표본수를 추출해야 함.

Power

❖ 가설검정을 위한 표본크기

- 의학연구나 실험연구의 경우에 유의수준($\alpha = 0.05$)과 검정력(80%)을 이용해서 표본수 추출
- 유효효과: $\delta = \mu_1 - \mu_0$
- 표본크기

$$n = \frac{\sigma^2(z_\alpha + z_\beta)^2}{d^2}$$

- 사례) K병원에서는 새로운 진통제를 개발하였다. 사전 연구를 통해 새로운 진통제의 효과크기가 5시간 이었으며, 표준편차는 20이었다. 5% 유의수준과 80% 검정력으로 평가하려면 몇 개의 표본을 이용해야 하는가?

$$n = \frac{\sigma^2(z_{\alpha/2} + z_\beta)^2}{d^2} = \frac{20^2(1.96 + 0.842)^2}{5^2} = 125.62 \cong 126$$

연습문제

연습문제1

- G 대학교에 근무하는 교직원의 연령은 평균 55세, 표준편차가 15세로 알려져 있다. 25명을 추출해서 연령을 조사했더니 50세로 나타났다.
- 교직원 연령의 신뢰구간은?

연습문제2

- G대학교에 근무하는 교직원의 연령은 평균 55세, 표준편차가 15세로 알려져 있다. 25명을 추출해서 연령을 조사했더니 50세로 나타났다.
- G 대학 전체 교직원의 평균연령을 55세라고 말할 수 있는가? 유의수준 5%에서 가설검증하시오.

연습문제3

- K영화관에서 판매하는 팝콘의 무게는 500g이라고 되어 있다.
- 이를 조사하기 위해 100개의 샘플을 조사했더니 평균 495g으로 나타났으며, 표준편차는 30g으로 나타났다.
- 샘플 팝콥 무게의 신뢰구간은?

  * $t_{(99,0.025)=}1.984$

연습문제4

- K영화관에서 판매하는 팝콘의 무게는 500g이라고 되어 있다.
- 이를 조사하기 위해 100개의 샘플을 조사했더니 평균 495g으로 나타났으며, 표준편차는 30g으로 나타났다.
- 팝콘 무게는 500g인가? 유의수준 5%에서 가설검증하시오.

  * $t_{(99,0.025)=}1.984$

Ⅳ. 확인적(CDA) 분석방법

통계분석방법

❖ 통계학(statistics)이란

- 관심 대상인 모집단의 특성을 파악하기 위해
- 모집단으로부터 관련된 일부 자료(표본)를 수집하고 (sampling)
- 수집된 표본의 자료를 요약 하여 표본의 특성을 파악하고(EDA)
- 표본의 자료를 이용하여 모집단의 특성에 대해 확률을 이용해 추론(CDA)

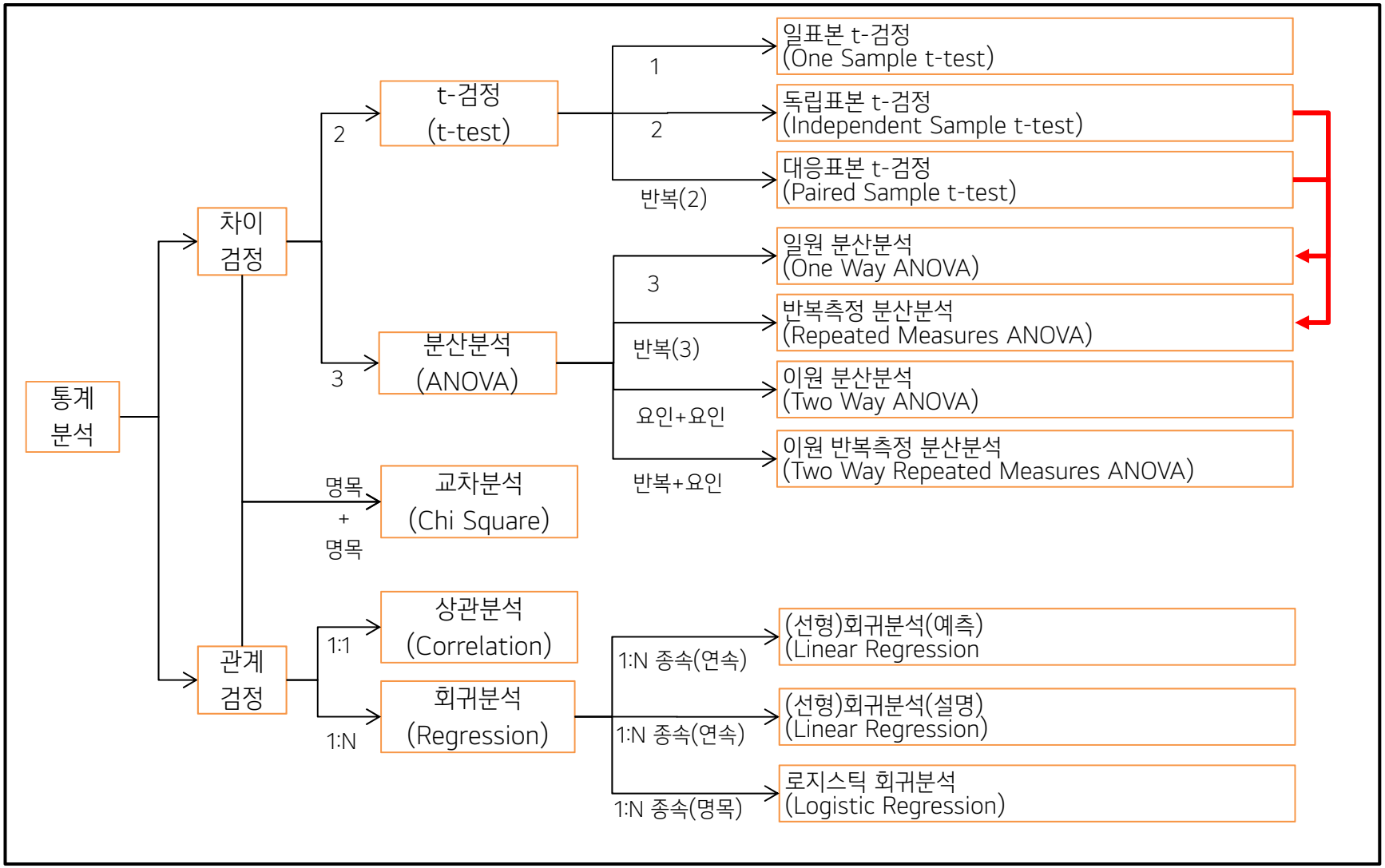
❖ 품질관리

- 1. 문제정의: DDA(Descriptive Data Analysis)
- 2. 현상확인: EDA(Exploratory Data Analysis)
- 3. 원인파악: CDA (Confirmatory Data Analysis)
- 4. 대안마련: PDA(Predictive Data Analysis)
- 5. 실행

통계분석방법

- 1. 기술적 분석 (DDA ; Descriptive Data Analysis): 문제발생
  - 표본추출을 통해 현재 상황을 기술
  - 평균, 표준편차, 빈도 등
- 2. 탐색적 분석(EDA ; Exploratory Data Analysis)
  - X-Y관계의 가설 도출
  - 그래프 분석
- 3. 확증적 분석(CDA ; Confirmatory Data Analysis)
  - 도출된 가설 검정
  - 통계분석방법
- 4. 예측적 분석 (PDA ; Predictive Data Analysis)
  - 모델링 및 최적화
  - 가설 검증된 모델의 영향력 검사
  - 최적화, 시뮬레이션 등
- 5. 실행

통계분석방법


통계 분석
차이 검정
관계 검정
2
3
t-검정 (t-test)
분산분석 (ANOVA)
명목 + 명목
교차분석 (Chi Square)
1:1
1:N
상관분석 (Correlation)
회귀분석 (Regression)
1
2
반복(2)
3
반복(3)
요인+요인
반복+요인
1:N 종속(연속)
1:N 종속(연속)
1:N 종속(명목)
일표본 t-검정 (One Sample t-test)
독립표본 t-검정 (Independent Sample t-test)
대응표본 t-검정 (Paired Sample t-test)
일원 분산분석 (One Way ANOVA)
반복측정 분산분석 (Repeated Measures ANOVA)
이원 분산분석 (Two Way ANOVA)
이원 반복측정 분산분석 (Two Way Repeated Measures ANOVA)
(선형)회귀분석(예측) (Linear Regression
(선형)회귀분석(설명) (Linear Regression)
로지스틱 회귀분석 (Logistic Regression)

One Sample t-test

❖ 문제의 정의

- B아이스크림회사에서 판매하는 아이스크림 중 파인트의 무게는 320g이다.
- 그러나 G대학 앞에 있는 점포에서 파는 아이스크림의 무게가 320g이 아니라는 소비자들의 불만이 있었다.
- 이에 따라 소비자단체에서는 B아이스크림회사에서 만든 아이스크림이 320g인지를 검사하고자 한다.

One Sample t-test(일표본 t-검정)

- 가설검정

$H_0: \mu = 320$

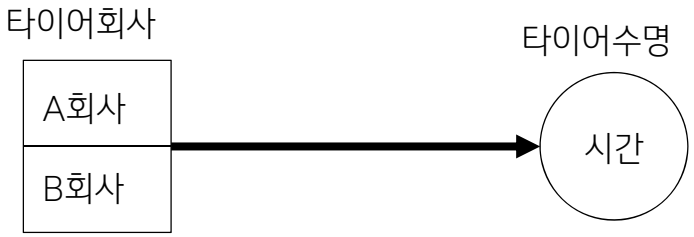
$H_1: \mu \neq 320$

무게1
304.9
305.2
304.0
304.7
305.2
307.2
309.9
313.3
314.8
314.9
315.4
315.8
317.9
315.1
315.8
316.8
317.6

Independent Sample t-test

❖ 문제의 정의

- 이교수는 이번에 자동차 타이어를 교체하려고 하는데 수명이 긴 타이어로 교체하려고 한다.
- 시중에는 A회사의 타이어와 B회사의 타이어가 있는데, 이 교수는 이 중에서 어느 타이어를 골라야 하는가?



Independent Sample t-test(독립표본 t-검정)

- 가설검정

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

회사	수명1
A타이어	50
A타이어	52
B타이어	51
B타이어	52
A타이어	52
B타이어	56
B타이어	50
B타이어	56
A타이어	46
A타이어	44
A타이어	56
A타이어	49
B타이어	51
B타이어	52
A타이어	52
B타이어	56
B타이어	50
B타이어	56
B타이어	52
B타이어	44

Paired Sample t-test

❖ 문제의 정의

- K제약회사는 새롭게 개발한 다이어트약이 효과가 있는지를 보고하기 위하여 약의 효능을 검증하였다.
- 약을 먹기 전의 체중과 약을 먹은 후 3개월 후의 체중을 조사하였다.
- 과연 새로운 약은 다이어트에 효과가 있는가?

Paired Sample t-test(대응표본 t-검정)

- 가설검정

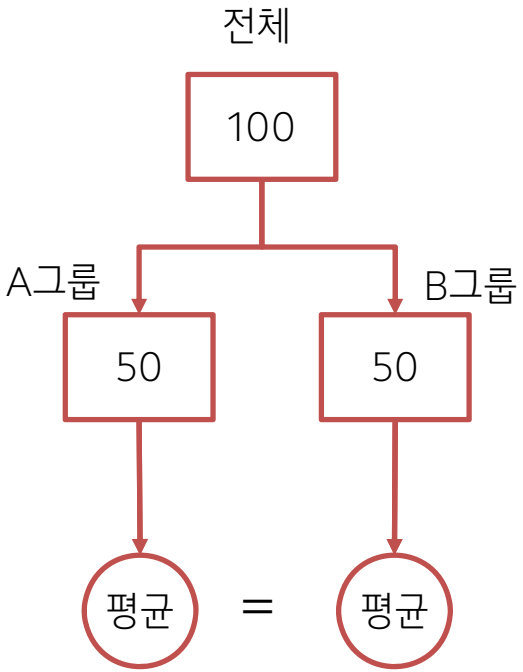
$$H_0: \mu_d = 0$$
$$H_1: \mu_d \neq 0$$

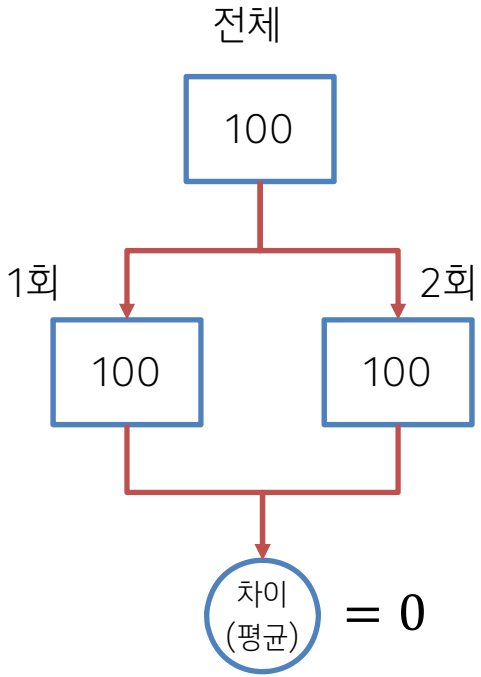
사전	사후1
83.69	77.01
71.80	69.03
78.45	71.03
75.11	71.04
78.19	71.06
71.70	74.08
62.43	62.08
76.16	73.10
82.23	75.10
74.41	74.10
76.36	73.10
76.17	73.11
75.68	76.11
71.01	69.12
75.85	77.12
54.43	50.12
76.00	71.13
68.65	67.13
75.73	71.14
75.11	71.15

Paired Sample t-test

❖ 독립표본과 대응표본의 차이점

- 독립표본 : 대상에서 1번만 측정
- 대응표본 : 동일 대상에서 반복해서 측정

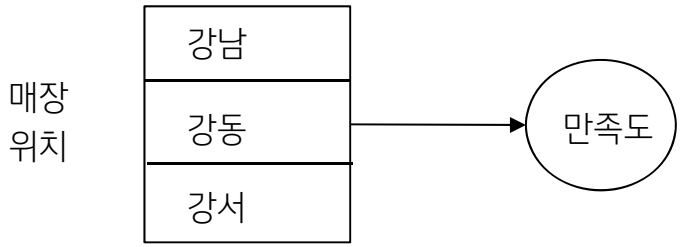




One-Way ANOVA

❖ 문제의 정의

- G거피회사는 강남(1), 강동(2), 강서(3)에 매장을 보유하고 있다. 매장별로 고객만족도가 차이가 있는지를 조사하기 위해 매장별 고객만족도를 조사하였다.
- 과연 3곳 매장의 고객만족도는 차이가 있는지? 있다면 어느 레스토랑의 서비스 만족도가 가장 안 좋은가 확인해보자



One Way ANOVA(일원 분산분석)

group	만족도1
강동	85
강동	82
강서	90
강동	88
강남	93
강동	83
강서	89
강서	88
강남	94
강동	93
강서	97
강서	94
강남	95
강동	84
강서	82
강동	87
강남	86
강동	85
강서	91
강동	92

- 가설검정

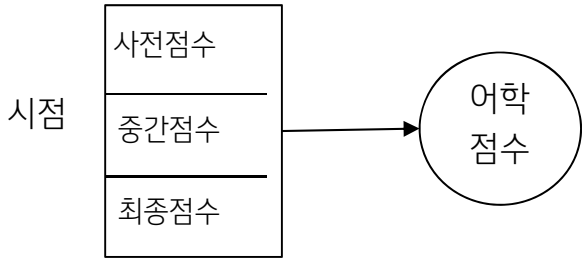
$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$H_1: not\ H_0, \mu_1 \neq \mu_2\ or\ \mu_1 \neq \mu_3\ or\ \mu_2 \neq \mu_3$$

Repeated Measures ANOVA

❖ 문제의 정의

- G대학에서 운영하는 6개월 어학 프로그램이 있다.
- 어학 프로그램의 효과를 측정하기 위해 학습 프로그램에 참여한 학생을 대상으로 프로그램 참여전, 중간시험, 최종시험 영어실력을 테스트하였다.
- 과련 어학프로그램은 효과가 있는지? 있다면 언제부터 효과가 나타났을지를 검증해 보자



Repeated Measures ANOVA(반복측정 분산분석)

사전점수	중간점수	최종점수1
63	63	63
60	60	60
61	61	61
57	57	57
58	58	58
58	58	58
53	53	53
58	58	58
58	58	58
57	57	57
60	59	60
57	64	62
65	66	65
65	66	64
63	65	63
59	58	60
55	59	60
61	64	62
61	61	59
57	57	59

- 가설검정

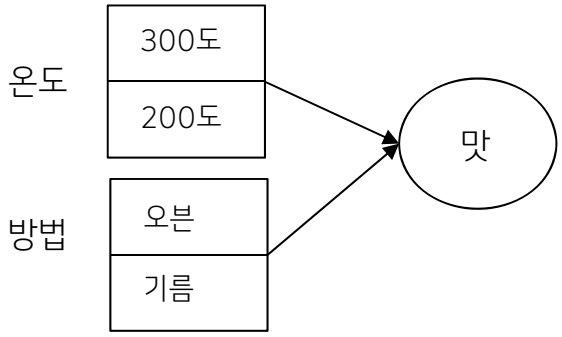
$H_0: \delta_1 = \delta_2 = \delta_3$

$H_1: not\ H_0$

Two-Way ANOVA

❖ 문제의 정의

- 치킨의 맛을 결정하는 두 가지 요인은 튀길 때의 기름 온도와 튀기는 방법이다.
- 튀길 때의 온도를 200도(1)와 300도(2)로 하고, 튀기는 방법을 오븐(1)과 기름(2)으로 하여 치킨을 튀긴 후에 사람들에게 맛을 평가하도록 하였다.
- 과연 온도와 방법이 맛을 결정하는데 중요한 요인인가? 이 두 가지 요인들 간의 상호작용효과는 없었는가?



Two Way ANOVA(이원 분산분석)

method	temp	맛점수1
오븐	200도	84
오븐	200도	87
오븐	200도	85
오븐	200도	89
오븐	200도	85
오븐	200도	87
오븐	200도	89
오븐	200도	88
오븐	200도	89
오븐	200도	87
오븐	200도	86
오븐	200도	88
오븐	200도	90
오븐	200도	91
오븐	300도	95
오븐	300도	93
오븐	300도	94
오븐	300도	98
오븐	300도	97
오븐	300도	94

Two-Way ANOVA

❖ 가설1

- 귀무가설(H_0): 튀기는 온도와 방법은 상호작용이 없다.

$$H_0: \alpha\beta = 0$$

- 연구가설(H_1): 튀기는 온도와 방법은 상호작용이 있다.

$$H_1: \alpha\beta \neq 0$$

❖ 가설2

- 귀무가설(H_0): 튀기는 온도에 따라 맛은 차이가 없다.

$$H_0: \mu_{A1} = \mu_{A2}$$

- 연구가설(H_1): 온도에 따라 맛은 차이가 있다.

$$H_1: \mu_{A1} \neq \mu_{A2}$$

❖ 가설3

- 귀무가설(H_0): 튀기는 방법에 따라 맛은 차이가 없다.

$$H_0: \mu_{B1} = \mu_{B2}$$

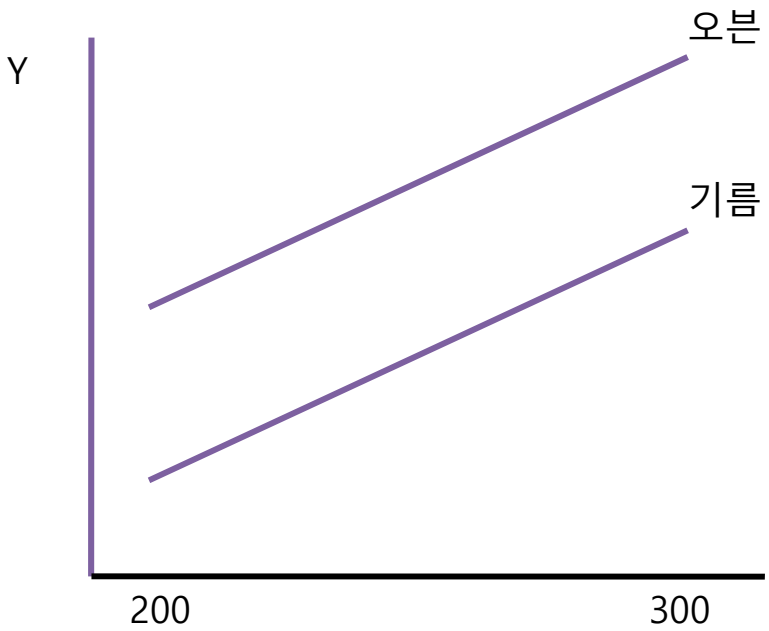
- 연구가설(H_1): 튀기는 방법에 따라 맛은 차이가 있다.

$$H_1: \mu_{B1} \neq \mu_{B2}$$

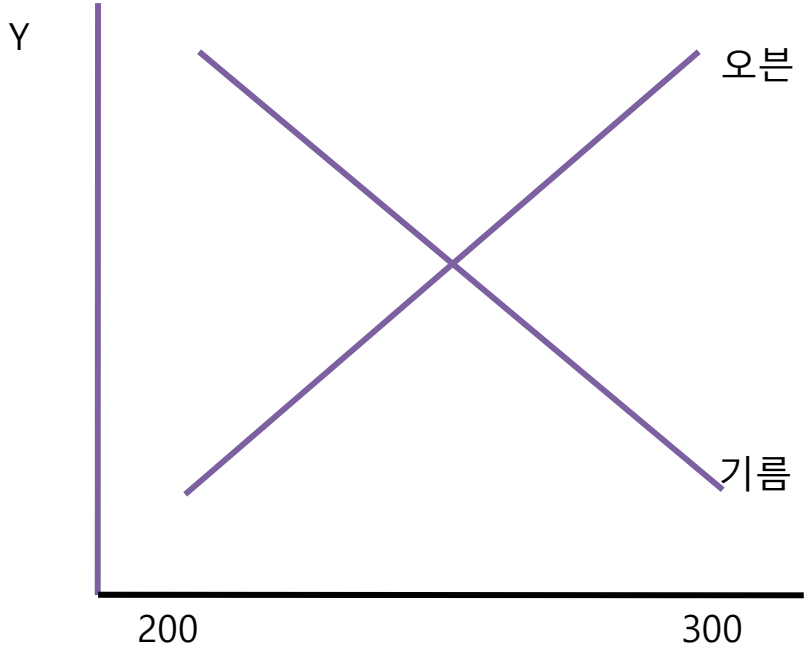
Two-Way ANOVA

❖ 평균반응 프로파일(average response profile)

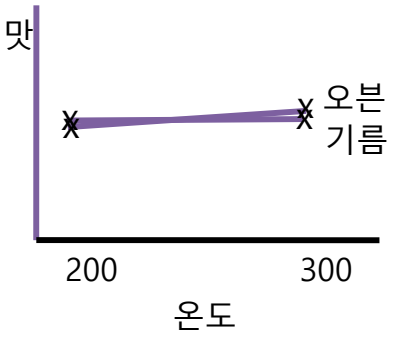
– 상호작용 없는 경우



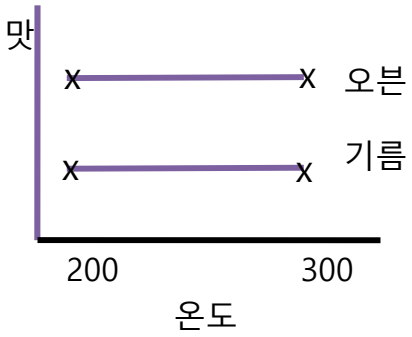
상호작용 있는 경우



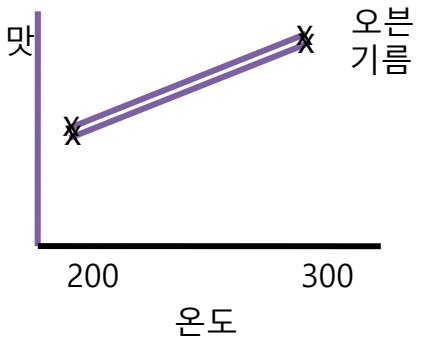
Two-Way ANOVA



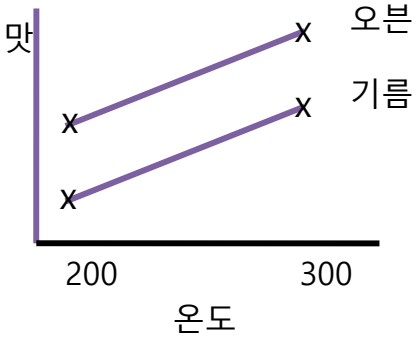
- 방법: X
- 온도: X
- 상호작용: X



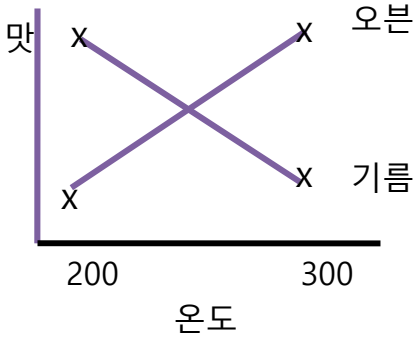
- 방법: O
- 온도: X
- 상호작용: X



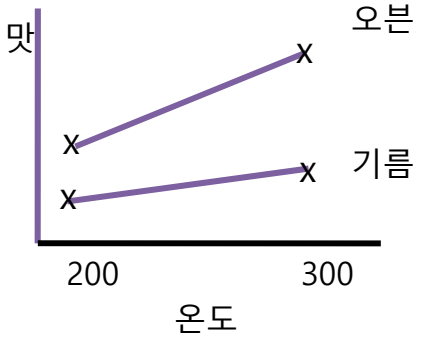
- 방법: X
- 온도: O
- 상호작용: X



- 방법: O
- 온도: O
- 상호작용: X



- 방법: X
- 온도: X
- 상호작용: O

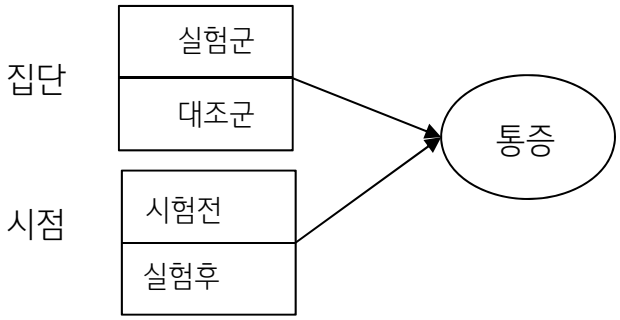


- 방법: O
- 온도: X
- 상호작용: O

Two-Way Repeated Measures ANOVA

❖ 문제의 정의

- G병원에서는 이번에 새롭게 아로마테라피 치료를 개발하였다.
- 이 치료가 통증에 효과가 있는지를 검증하기 위해
- 새롭게 개발한 치료제로 향기요법을 처치 받는 실험군(2)과 일반 향기치료제로 가짜 향기요법을 처치 받는 대조군(1)을 나누고, 치료전과 후에 통증이 차이가 있는지를 검증하였다.
- 아로마테라피 치료제는 효과가 있었는가?



Two Way Repeated Measures(Mixed) ANOVA
(이원 반복측정 분산분석)

group	사전통증	사후통증1
대조군	12.32	23.09
대조군	57.16	40.45
대조군	12.32	23.09
대조군	16.62	44.97
대조군	18.63	40.78
대조군	20.12	35.12
대조군	20.12	35.12
대조군	22.78	42.07
대조군	22.78	42.07
대조군	24.36	31.08
대조군	28.38	51.63
대조군	28.38	51.63
대조군	30.10	36.92
대조군	31.29	39.58
대조군	31.29	39.58
대조군	33.54	50.24
대조군	33.54	50.24
대조군	35.27	49.26
대조군	35.27	49.26
대조군	36.87	22.79

Two-Way Repeated Measures ANOVA

❖ 가설1

- 귀무가설(H_0): 실험전에 그룹간 통증은 차이가 없다.

$$H_0: \mu_{befor\_treatment} = \mu_{before\_control}$$

- 연구가설(H_1): 실험전에 그룹간 통증은 차이가 있다.

$$H_0: \mu_{befor\_treatment} \neq \mu_{before\_control}$$

❖ 가설2

- 귀무가설(H_0): 그룹과 시점간에는 상호작용이 없다.

$$H_0: \alpha\beta = 0$$

- 연구가설(H_1): 그룹과 시점간에는 상호작용이 있다.

$$H_1: \alpha\beta \neq 0$$

❖ 가설3

- 귀무가설(H_0): 실험후에 그룹간 통증은 차이가 없다.

$$H_0: \mu_{after\_treatment} = \mu_{after\_control}$$

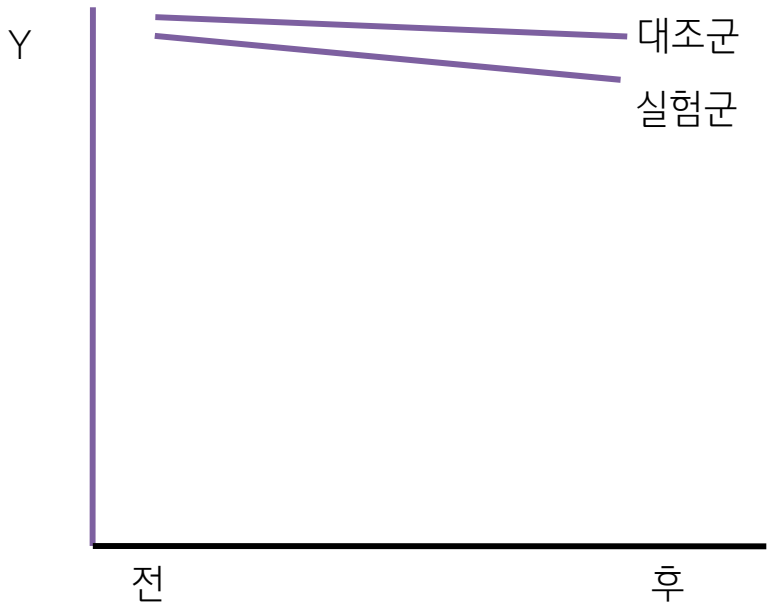
- 연구가설(H_1): 실험전에 그룹간 통증은 차이가 있다.

$$H_0: \mu_{after\_treatment} \neq \mu_{after\_control}$$

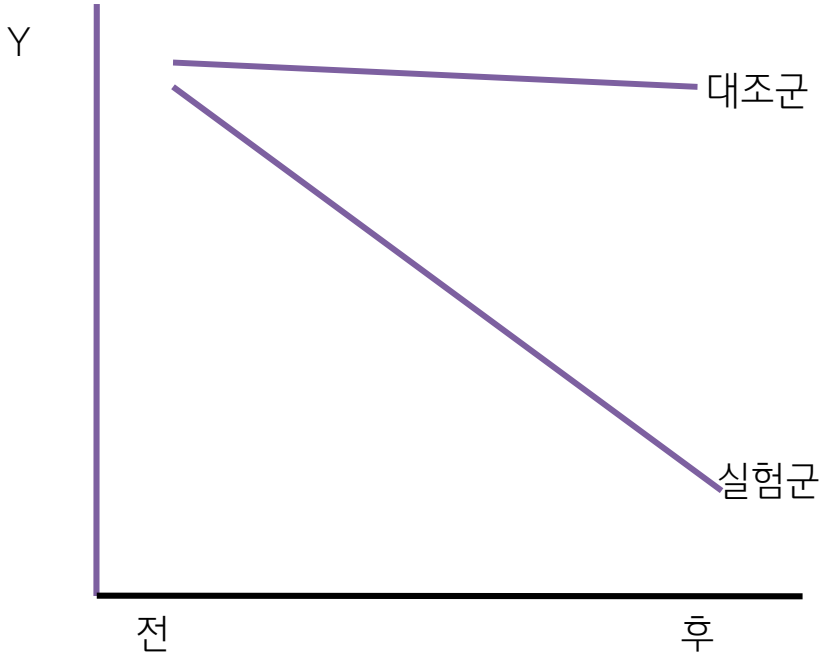
Two-Way Repeated Measures ANOVA

❖ 평균반응 프로파일(average response profile)

– 상호작용 없는 경우



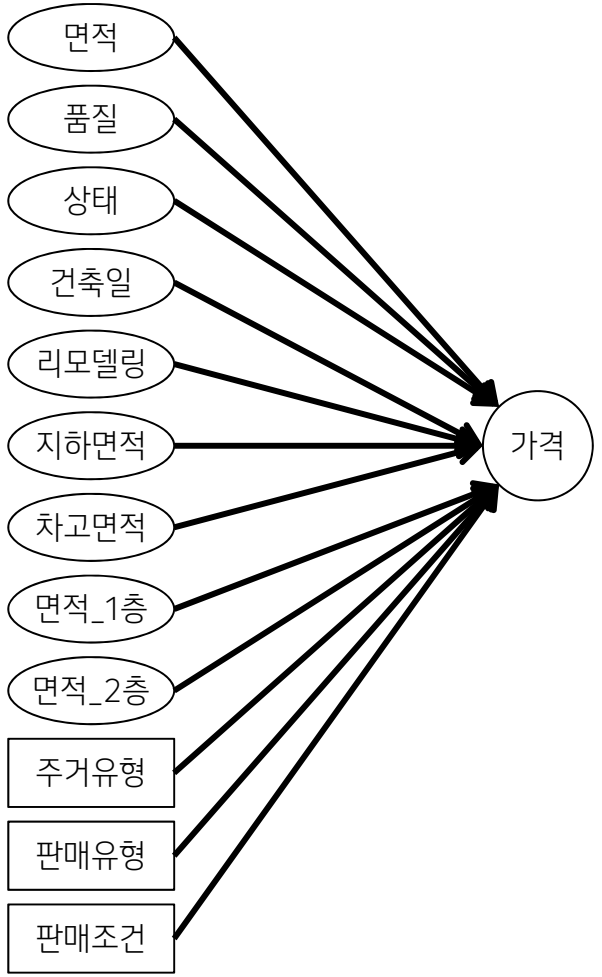
상호작용 있는 경우



Regression

❖ 문제의 정의

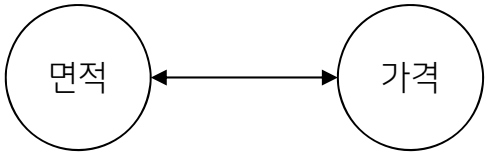
- 가격 – 부동산의 판매 가격(달러)
- 면적: 부지 크기(제곱피트)
- 품질: 전반적인 재료 및 마감 품질(10점)
- 상태: 전반적인 상태 등급 (10점)
- 건축일: 원래 건설 날짜
- 리모델링: 리모델링 날짜
- 지하면적: 지하 면적의 총 평방피트
- 차고면적: 평방 피트 단위의 차고 크기
- 면적_1층: 1층 평방 피트
- 면적_2층: 2층 평방 피트
- 주거유형BldgType: 주거 유형
- 판매유형SaleType: 판매 유형
- 판매조건SaleCondition: 판매 조건



Correlation

❖ 문제의 정의

- 부동산의 판매 가격과 면적 간의 관계를 살펴보고자 한다. 가격과 면적은 관계가 있는가?
- 또한 다른 변수들(면적 ~ 면적_2층까지의 변수)와는 관계가 있는가?



Correlation(상관분석)

- 가설검정

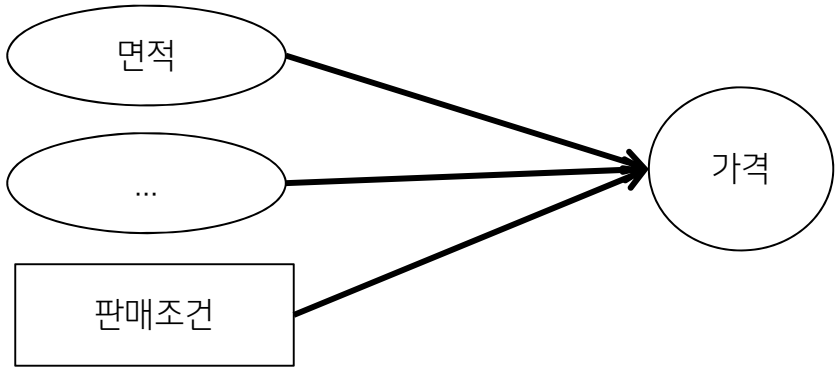
$$H_0: \rho = 0$$
$$H_0: \rho \neq 0$$

가격	연면적
150750	7388
131500	4435
160000	8800
187500	13031
153900	7892
129900	4224
165500	9600
173000	10852
167000	9937
184000	12416
165000	9500
149000	7200
180500	11851
177000	11512
155000	7931
167000	10004
156500	8400
163000	9120
170000	10192
160000	8658

Linear Regression(예측)

❖ 문제의 정의

- 부동산의 판매 가격을 예측하고자 한다. 부동산 가격과 관계가 있는 변수들은 무엇인가?
- 부동산 판매 가격을 예측할 수 있는 예측모델을 만들어 보자.



Regression(회귀분석)

- 가설검정

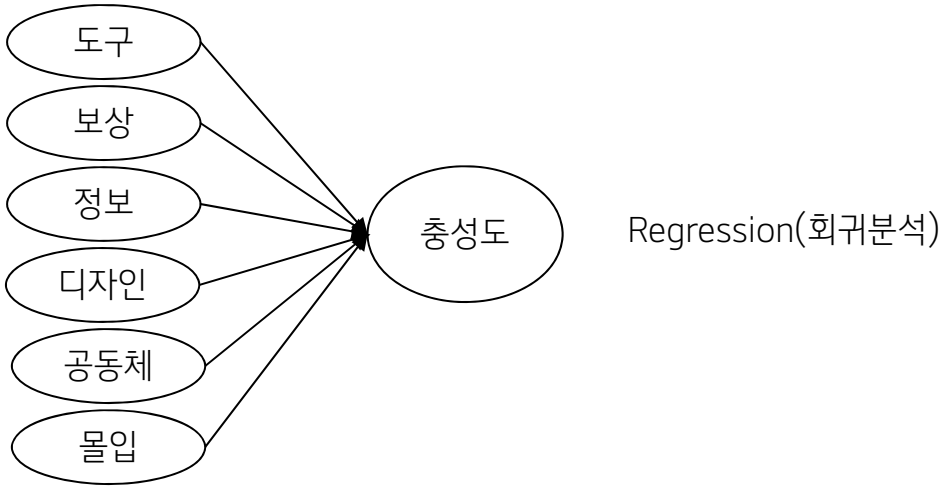
$$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_n = 0$$

$$H_1: not\ H_0$$

Linear Regression(설명)

❖ 문제의 정의

- 온라인게임의 충성도에 영향을 주는 요인이 무엇인지를 연구하고자 한다.
- 영향을 주는 변수로는 도구, 보상, 정보, 디자인, 공동체, 몰입이 있다.
- 온라인게임 몰입에 영향을 주는 변수는 무엇이고, 어떤 변수가 온라인게임 몰입에 가장 큰 영향을 주는 변수인가?



- 가설검정

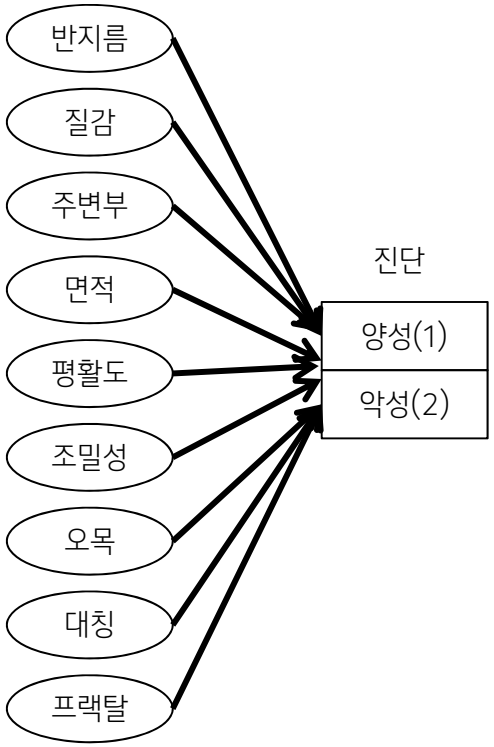
$$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_n = 0$$

$$H_1: not\ H_0$$

Logistic Regression

❖ 문제의 정의

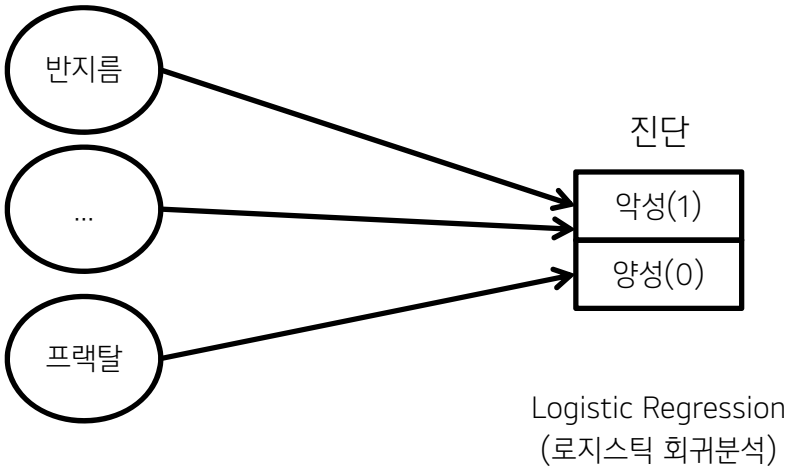
- 유방암 진단 (1 = 양성, 2 = 악성)
- 유방 덩어리의 FNA(Fine Needle Aspirate)의 디지털화된 이미지
- 반지름(주변의 중심에서 점까지의 거리 평균)
- 질감(그레이 스케일 값의 표준 편차)
- 주변부
- 면적
- 평활도(반지름 길이의 국지적 변동)
- 조밀성(주변^2 /면적 - 1.0)
- 오목(윤곽의 오목 부분의 심각도)
- 대칭
- 프랙탈 차원: 공간에 패턴을 얼마나 조밀하게 채우는지 나타내는 비율
- 분류 분포: 양성 357명, 악성 212명



Logistic Regression(설명)

❖ 문제의 정의

- 유방암 진단(악성)에 영향을 주는 요인이 무엇인지 파악하고자 한다.
- 유방암 악성에 영향을 주는 요인은 무엇인가?



- 가설검정

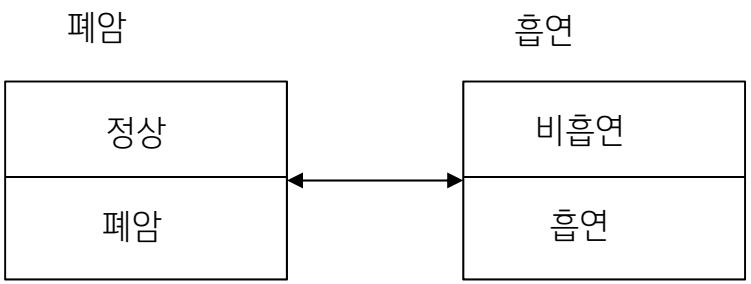
$$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_n = 0$$

$$H_1: not\ H_0$$

Chi-Square test

❖ 문제의 정의

- G병원에서는 흡연을 많이 하는 사람일수록 폐암에 걸릴 확률일 높다는 것을 발표하였다. 관련 자료는 국립보건소를 통해 2차 자료를 구했다고 하자.
- 과연 흡연이 폐암과 연관이 있는지를 검증해 보자.



Chi Square(교차분석)

폐암	흡연	관측치
정상	비흡연	170867
정상	흡연	27784
폐암	비흡연	723
폐암	흡연	504

- 가설검정

$$H_0: \theta_{11} = \theta_{21}, \theta_{12} = \theta_{22}$$
$$H_1: \theta_{11} \neq \theta_{21}, \theta_{12} \neq \theta_{22}$$