1강 통계학의 기본 개념 및 엑셀 기초

정보통계학과 이기재 교수

학습목표

- 데이터 수집방법과 데이터 종류에 대해서 설명
- 엑셀 기본 용어와 리본메뉴의 기능의 활용
- 엑셀 연산자와 함수를 사용한 계산
- 상대참조와 절대참조의 차이점을 설명하고 바르게 활용

학습목차

1 통계학의 기본개념

1 통계학의 기본 개념

1 통계학의 적용 과정

1. 문제 설정
2. 조사, 관측 실험을 통한 데이터의 수집
3. 수집된 데이터의 정리 · 요약을 통한 새로운 정보추출
4. 통계적 추론과정을 통해서 문제 해결

1 통계학의 기본 개념

2 통계조사와 실험

☑ 통계 조사

- 복잡한 사회 또는 집단의 어떤 현상을 수량화 함으로써 객관적이고, 구체적인 특징을 파악하기 위한 일련의 과정

☑ 실험

- 연구자가 실험환경을 통제하고 조작을 가함으로써, 특정 처리의 효과를 파악하는 과정

1 통계학의 기본 개념

통계조사의 예

조사 목적 2007년 12월에 실시된 제17대 대통령 선거 예측조사

조사 개요 전국의 남, 여 유권자를 대상으로 2,000명을 표본 추출

조사 결과

	이명박	정동영	이회창	문국현	기타
예상득표율	49.1%	24.4%	15.9%	7.0%	3.6%
실제득표율	48.7%	26.1%	15.1%	5.8%	4.3%

1 통계학의 기본 개념

통계조사의 특징

- 모집단의 특성을 수치로 보여주는데 목적이 있음
- 연구자는 모집단(조사대상)으로부터 조사를 위해서 누구를 뽑을 것인가(표본추출)를 정하게 됨
- 표본으로 추출된 사람들에 대해서 어떤 통제나 조정을 가하지 않고 조사하게 됨

1 통계학의 기본 개념

실험의 예

실험 목적

아스피린이 심장마비를 예방하는 데에 효과가 있는가?

실험 방법

① 자원한 22,000명의 내과 환자들을 랜덤하게 11,000명씩 두 그룹으로 구분

실험집단	이틀에 한 알씩 아스피린 복용
대조집단	이틀에 한 알씩 가짜 약 복용

② 몇 년 후, 이들 두 집단을 관찰하여 결과 분석

1 통계학의 기본 개념

실험의 기본개념

실험

- 합리적이고 공정한 방법으로 특정처리의 효과 유무를 알고자 하는 것임

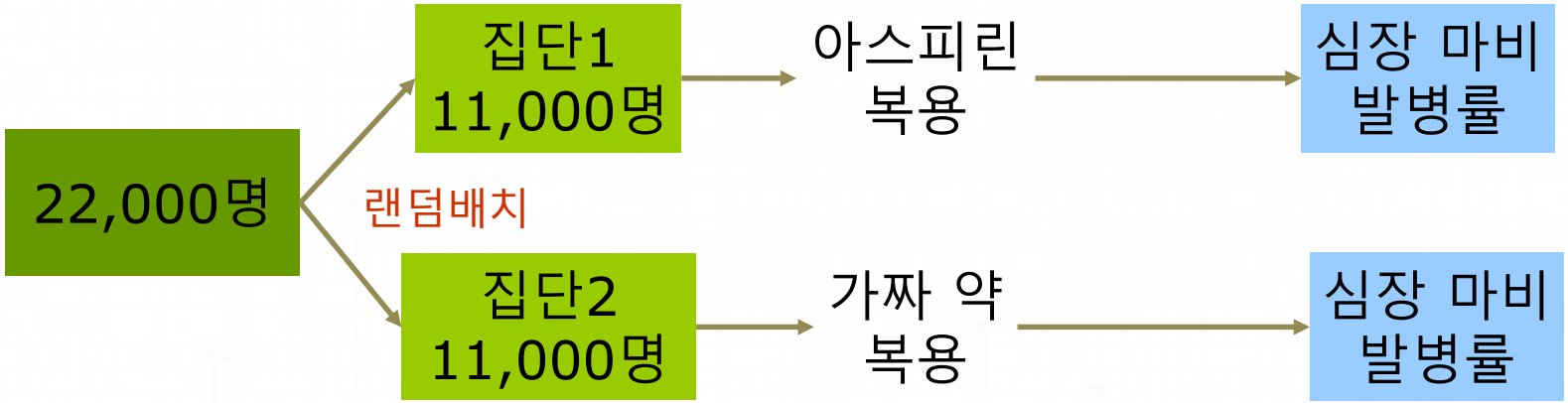
공정한 실험 과정

- 무작위로 통제(랜덤화)
  - 실험 대상자들을 랜덤한 방법으로 실험집단과 통제집단으로 구분하여 실험
- 이중눈가림(double blindness)
  - 피험자와 연구자 모두 누가 처리집단 또는 통제집단에 속해 있는지 알지 못해야 함

➡ **무작위로 통제된 이중 눈가림 실험**

1 통계학의 기본 개념

실험의 예 : 개선된 실험 방법



- 무작위로 통제된 이중 눈가림 실험
- 집단 1의 심장마비 발병률이 집단 2보다 현저히 작을 때 그 차이를 아스피린 복용의 효과로 해석

1 통계학의 기본 개념

3 질적 자료와 양적 자료

☑ 질적 자료 (Qualitative data)

- 명목척도나 순서척도에 의해서 측정된 자료

☑ 양적 자료 (Quantitative data)

- 구간척도나 비율척도로 측정된 자료
- 연속형 자료(continuous data)

1 통계학의 기본 개념

4 일변량 자료와 다변량 자료

☑ 일변량 자료

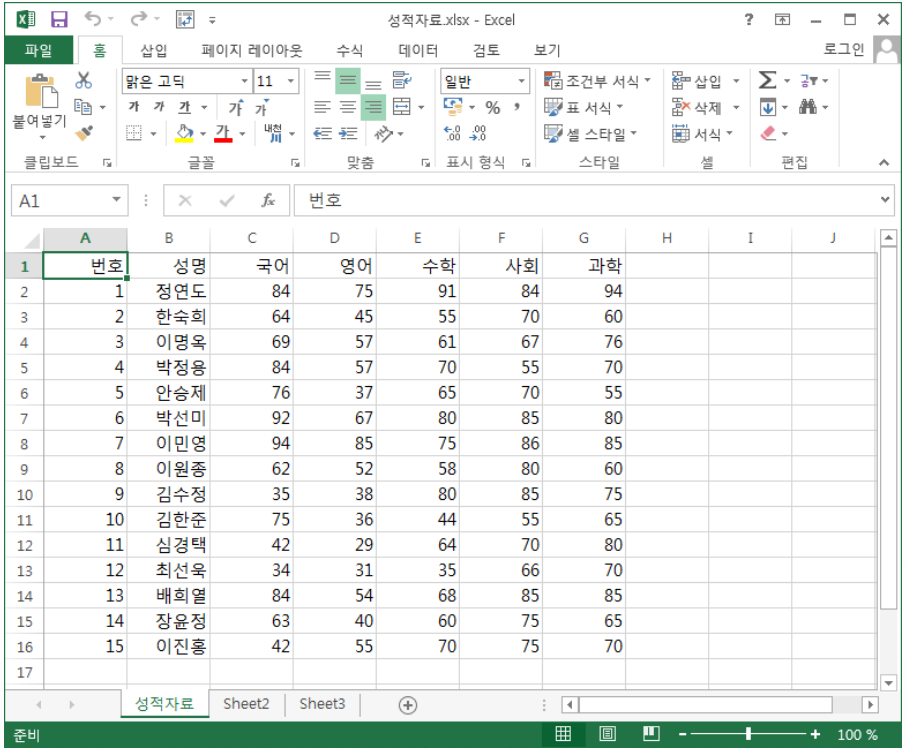
- 각 조사단위에서 한 변수만을 측정하여 얻은 데이터

☑ 다변량 자료

- 각 조사단위에서 두 개 이상의 변수를 측정하여 얻은 자료

1 통계학의 기본 개념

5 데이터 관련 기본용어



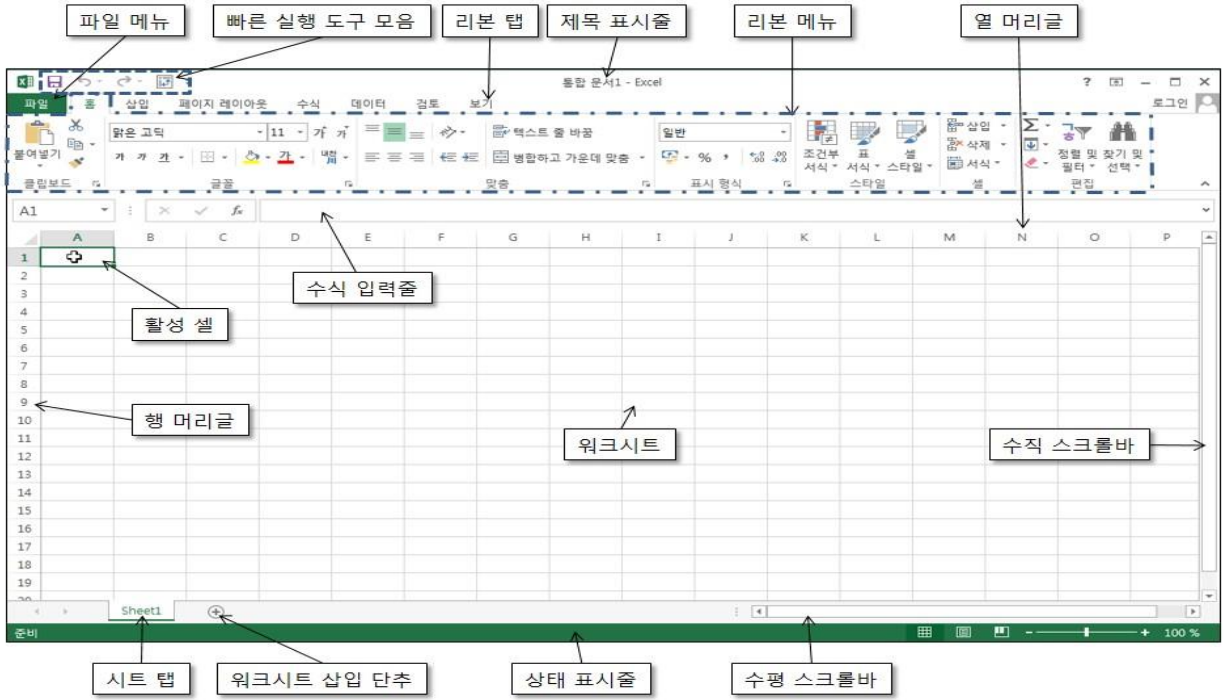
번호	성명	국어	영어	수학	사회	과학
1	정연도	84	75	91	84	94
2	한숙희	64	45	55	70	60
3	이명옥	69	57	61	67	76
4	박정용	84	57	70	55	70
5	안승제	76	37	65	70	55
6	박선미	92	67	80	85	80
7	이민영	94	85	75	86	85
8	이원종	62	52	58	80	60
9	김수정	35	38	80	85	75
10	김한준	75	36	44	55	65
11	심경택	42	29	64	70	80
12	최선욱	34	31	35	66	70
13	배희열	84	54	68	85	85
14	장윤정	63	40	60	75	65
15	이진홍	42	55	70	75	70

- **변수**
  : 각 조사 단위로부터 측정된 개별적인 속성
- **케이스**
  : 한 조사단위에 대한 정보의 집합체
- **변수명 :** 변수의 이름

2 엑셀 사용법 기초

2 엑셀 사용법 기초

1 엑셀 기본용어(1)



2 엑셀 사용법 기초

1 엑셀 기본용어(2)

☑ 워크시트(worksheet)

- 셀로 구성된 작업 공간으로 통합문서의 일부분
- 데이터를 입력하고 작업하는 페이지라고 할 수 있음

☑ 통합문서(workbook)

- 데이터 작업을 하고 저장하는 하나의 파일
- 여러 개의 워크시트로 구성됨
- 기본적으로 통합문서는 3개의 워크시트 포함

2 엑셀 사용법 기초

1 엑셀 기본용어(3)

☑ 셀(cell)

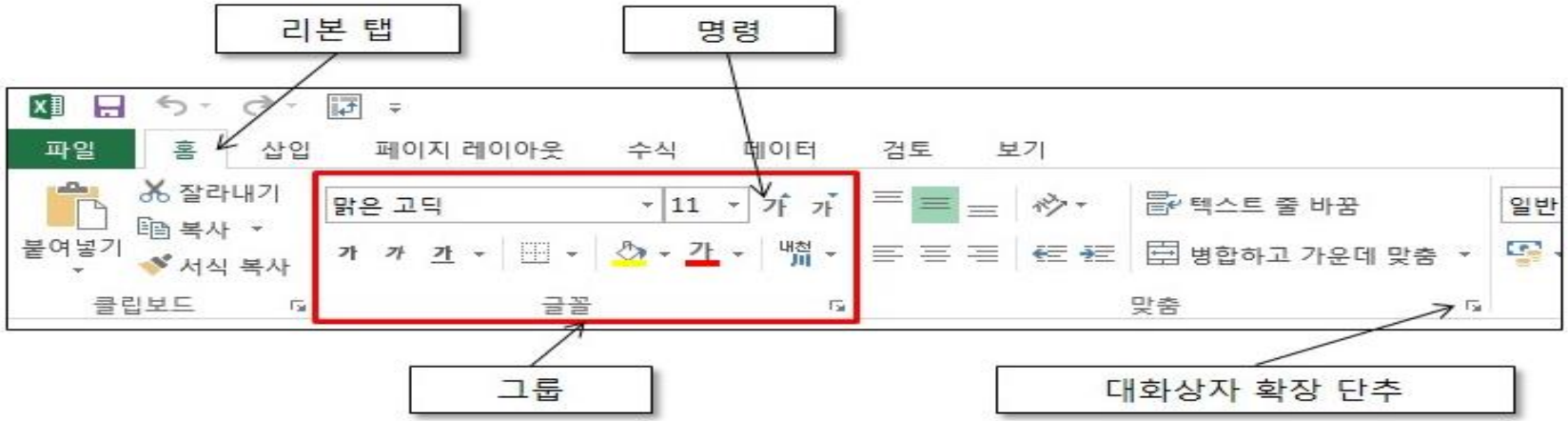
- 워크시트에서 자료를 입력하는 가장 작은 단위
- 열 범위 : A, B, ..., AA, BB, ..., XFD까지 16,384개
- 행 범위 : 1부터 1,048,576까지
- 셀 참조 : 각 셀은 열과 행의 조합으로 주소를 가짐
  (예 : C열의 5행의 셀은 C5가 그 주소가 됨)

3 엑셀의 리본 메뉴

3 엑셀 리본 메뉴

1 엑셀 리본메뉴(1)

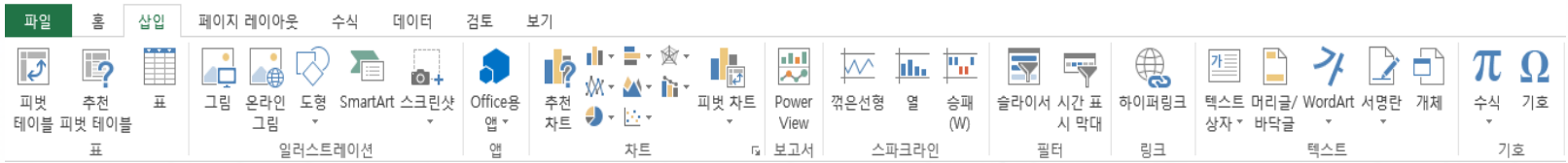
- 과거 엑셀에서 사용되던 메뉴와 도구모음 역할을 대신함
- 리본메뉴의 구성



3 엑셀 리본 메뉴

1 엑셀 리본메뉴(2)

- **[삽입] 탭의 리본메뉴**
  워크시트에 삽입할 수 있는 피벗테이블, 표, 그림이나 차트 등을 넣을 수 있는 명령들이 나타남



4 엑셀 함수 사용법

4 엑셀 함수 사용법

엑셀의 연산자(1)

1. 산술 연산자
2. 비교 연산자
3. 문자 연산자
4. 참조 연산자

4 엑셀 함수 사용법

엑셀의 연산자(2)

산술 연산자	+ (더하기), - (빼기), /(나누기), *(곱셈), ^(지수)
비교 연산자	=(같다), >(크다), <(작다) >=(크거나 같다), <=(작거나 같다), <>(같지 않다)
문자 연산자	&(두 값을 연결하여 하나의 문자 값 산출)

4 엑셀 함수 사용법

엑셀의 연산자(3)

참조 연산자

- : (범위)
  - 예 = SUM(A1:A10)
- , (합집합)
  - 예 = SUM(A1:A10, B1:B5)
- 공백 (교집합)
  - 예 = SUM(A1:A10 A8:A15)

4 엑셀 함수 사용법

엑셀 함수 기능과 형식

기능

값, 셀 참조, 함수 등을 사용하여 새로운 값의 생성

형식

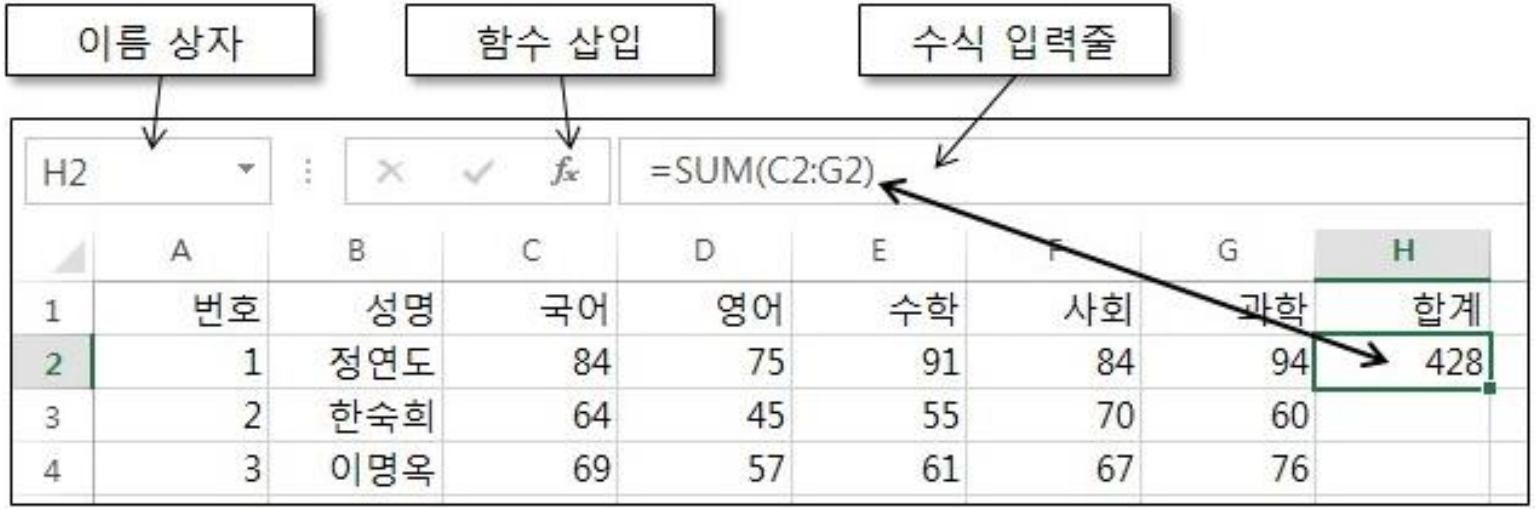
= 함수이름 (인수, 인수,…, 인수)

예제

= SUM(A1:A3)

4 엑셀 함수 사용법

함수식의 입력



4 엑셀 함수 사용법

함수 사용의 규칙

- 수식은 등호(=)로 시작
- 인수를 묶는 양쪽의 괄호는 반드시 필요
  - 예 = RAND( )
- 함수의 인수로는 숫자, 셀 범위, 논리 값, 문자 값, 다른 함수 사용 가능
- 인수가 여러 개 사용되는 함수는 콤마(,)를 사용하여 분리

4 엑셀 함수 사용법

함수 사용에 대한 예

- =SUM(A1:A3)
- =SUM(2, 3, 4, 5, 5)
- =SUM(5+2, AVERAGE(5, 7), 5)
- =RAND()

4 엑셀 함수 사용법

셀 참조 방법

☑ **상대 참조 : 행 이름이나 열 이름만을 사용하는 형식**

예) A1, D3, =AVERAGE(C2:G2)

☑ **절대 참조 : 셀 참조는 셀 참조에 "$" 표시 사용**

예) $A$1, $D$3, =AVERAGE($C$2:$G$2)

☑ **혼합 참조 : 행 이름이나 열 이름의 한쪽에만 "$" 표시**

예) $A1, D$3

4 엑셀 함수 사용법

상대 참조의 예

	A	B	C	D	E	F	G	H	I	J	K	L
1	번호	성명	국어	영어	수학	사회	과학	평균				
2	1	정연도	84	75	91	84	94	85.6		=AVERAGE(C2:G2)		
3	2	한숙희	64	45	55	70	60	58.8		=AVERAGE(C3:G3)		
4	3	이명옥	69	57	61	67	76	66				
5	4	박정용	84	57	70	55	70	67.2				
6	5	안승제	76	37	65	70	55	60.6				
7	6	박선미	92	67	80	85	80	80.8				
8	7	이민영	94	85	75	86	85	85				
9	8	이원종	62	52	58	80	60	62.4				
10	9	김수정	35	38	80	85	75	62.6		=AVERAGE(C9:G9)		
11	10	김한준	75	36	44	55	65	55				

- [채우기]기능을 이용하면 새 위치에 맞게 참조 영역이 자동 조정됨
- 복사 또는 이동된 수식에서 참조되는 영역은 원래의 수식에 들어 있는 셀의 영역과는 다른 셀을 참조함

4 엑셀 함수 사용법

절대 참조의 예

	A	B	C	D	E	F	G	H	I	J	K	L
1	번호	성명	국어	영어	수학	사회	과학	평균				
2	1	정연도	84	75	91	84	94	85.6		=AVERAGE($C$2:$G$2)		
3	2	한숙희	64	45	55	70	60	85.6		=AVERAGE($C$2:$G$2)		
4	3	이명옥	69	57	61	67	76	85.6				
5	4	박정용	84	57	70	55	70	85.6				
6	5	안승제	76	37	65	70	55	85.6				
7	6	박선미	92	67	80	85	80	85.6				
8	7	이민영	94	85	75	86	85	85.6				
9	8	이원종	62	52	58	80	60	85.6				
10	9	김수정	35	38	80	85	75	85.6		=AVERAGE($C$2:$G$2)		
11	10	김한준	75	36	44	55	65	85.6				

- [채우기]기능을 이용하거나 복사하면 원래 수식에 나타난 것과 똑같은 참조 영역이 복사됨

4 엑셀 함수 사용법

수식 사용에서 나타 날 수 있는 오류 값들

오류 값	내용
#DIV/0?	수식에서 값을 0으로 나누려고 했을 때 발생함
#NAME	엑셀에서 인식할 수 없는 이름을 사용하는 경우에 발생
#VALUE	잘못된 인수나 피 연산자를 사용했을 때 발생
#N/A	원하는 연산을 수행하기 위한 데이터가 없는 경우
#REF!	다른 수식에서 참조하는 셀을 삭제하였거나, 다른 수식에서 참조하는 셀 위에 다른 셀을 붙여 넣기를 한 경우
#NUM!	숫자 인수를 필요로 하는 함수에서 사용할 수 없는 인수를 지정한 경우
#NULL!	부적당한 범위 연산자나 셀 참조를 사용한 경우

- #### 로 표기 : 수식 결과값이 너무 길어서 셀 안에 모두 표시할 수 없을 때 나타남

연습문제 1.

1. 다음은 엑셀의 통합문서와 워크시트의 관계를 설명한 것이다. 다음 중 올바르지 **않은** 것은?

① 워크시트는 행과 열의 구조를 갖는 많은 셀로 구성된 데이터 입력과 분석의 작업 공간이다.

② 통합문서는 여러 개의 워크시트를 가질 수 있다

③ 워크시트는 16,384개의 열과 1,048,576개의 행으로 구성된다.

④ 엑셀에서 파일로 저장하고, 불러오는 기본 단위는 워크시트이다.

연습문제 2.

2. 통계 분석에서 자료(데이터)를 양적 자료와 질적 자료로 구분할 수 있다. 양적 자료와 질적 자료로 구분하는 기준은 무엇인가?

① 자료 품질의 높고 낮음

② 자료를 얻는 데 사용된 측정의 척도

③ 자료에 있는 변수의 개수

④ 자료가 실험에 의해서 얻어졌는지 여부

연습문제 3.

3. 다음 자료의 변수의 수와 케이스의 수는 각각 몇 개인가?

행 성	태양과의 거리 (백만 km)	적도 반경 (km)	극 반경 (km)	탈출 속도 (km/sec)	비행 속도 (km/h)	위성 개수
수성	58	2439	2439	4.3	172410	0
금성	108	6050	6050	10.3	126110	0
지구	150	6378	6356	11.2	107250	1
화성	229	3397	3377	5	86870	2
목성	779	71398	66850	59.5	47020	12
토성	1427	60000	53880	35.6	34710	10
천왕성	2871	26145	25500	21.2	24520	5
해왕성	4497	24300	23654	23.6	19550	2

연습문제 4.

4. 셀 C6에 입력되어 있는 수식을 드래그 & 드롭으로 C11까지 채워서 1월부터 6월까지의 미 달러($)기준의 수출액을 원화(₩) 기준으로 바꾸고자 한다. 셀 C6에 입력할 수식은?

	A	B	C
1			
2	환율	940	
3			
4	기간	수출액	
5		미달러($)	원화(₩)
6	1월	12,000	11,280,000
7	2월	11,000	
8	3월	15,000	
9	4월	14,500	
10	5월	16,000	
11	6월	15,000	
12			

2강 데이터의 그래프 표현과 수치요약 (1)

정보통계학과 이기재 교수

학습목표

- 다양한 함수 사용
- 데이터 특성과 분석 목적에 알맞은 그래프의 작성
- 엑셀을 이용하여 작성한 차트의 편집·수정
- 히스토그램을 이용한 분포의 특징 파악

학습목차

1 엑셀 함수

1 엑셀 함수

수학함수 예(1)

SUM 함수

- 목 적 : 합 계산
- 구 문 : =SUM(number1, number2, …)
- 예 : =SUM(A2:C2) , =SUM(2,3,7)

1 엑셀 함수

수학함수 예(2)

ROUND 함수

- 목 적 : 숫자를 지정한 자릿수로 반올림
- 구 문 : = ROUND(number, num_digits)

  Number : 반올림 할 수

  Num_digits : 반올림 할 자릿 수
- 예 : = ROUND(1.475, 2), = ROUND(A1, 1)

1 엑셀 함수

통계함수 예

AVERAGE 함수

- 목 적 : 인수의 산술 평균을 구함
- 구 문 : = AVERAGRE(number1, number2, …)

STDEV 함수

- 목 적 : 표본자료의 표준편차를 구함
- 구 문 : = STDEV(number1, number2, …)

1 엑셀 함수

IF 함수

IF 함수

- 목 적 : 주어진 조건에 따라서 다른 값을 표시
- 구 문 : =IF(주어진 조건, value_if_true, value_if_false)

1 엑셀 함수

IF 함수 사용 예제

- 성적이 70점 이상이면 “합격”, 70점 미만이면 “불합격”

  = IF(A2>=70, "합격”, “불합격”)

- = IF(A2>=90, "A", IF(A2>=80, "B", IF(A2>=70, "C", IF(A2>=60, "D“ , "F"))))

2 엑셀 그래프의 종류 및 특성

2 엑셀 그래프의 종류 및 특성

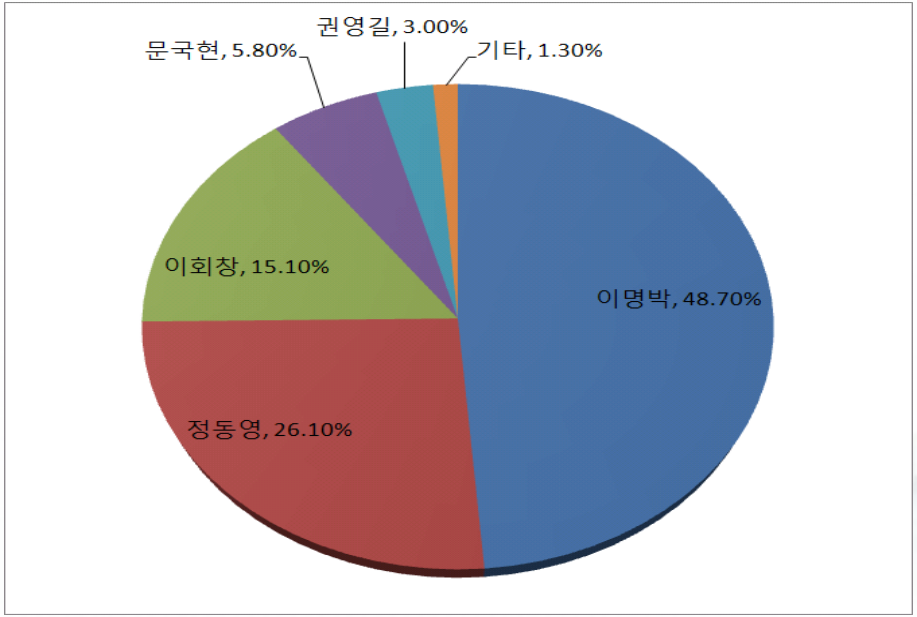
데이터 분석에서 그래프의 사용

- 데이터가 갖고 있는 정보를 요약하여 시각적으로 보여줌
- 데이터의 특징을 이해하기 쉽게 전달해 줌
- 분석 목적과 데이터 종류에 따라서 사용할 수 있는 그래프가 달라짐

2 엑셀 그래프의 종류 및 특성

엑셀 그래프의 종류(1)

구성비 ▸ 원형 차트

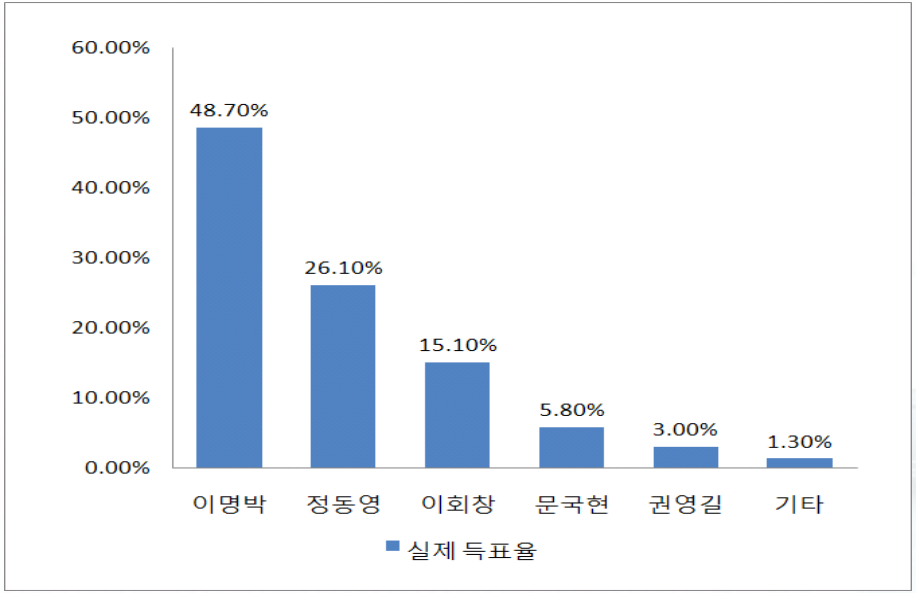


(구성비는 백분율로 주어지고 합하여 100%가 되는 경우를 말함)

2 엑셀 그래프의 종류 및 특성

엑셀 그래프의 종류(2)

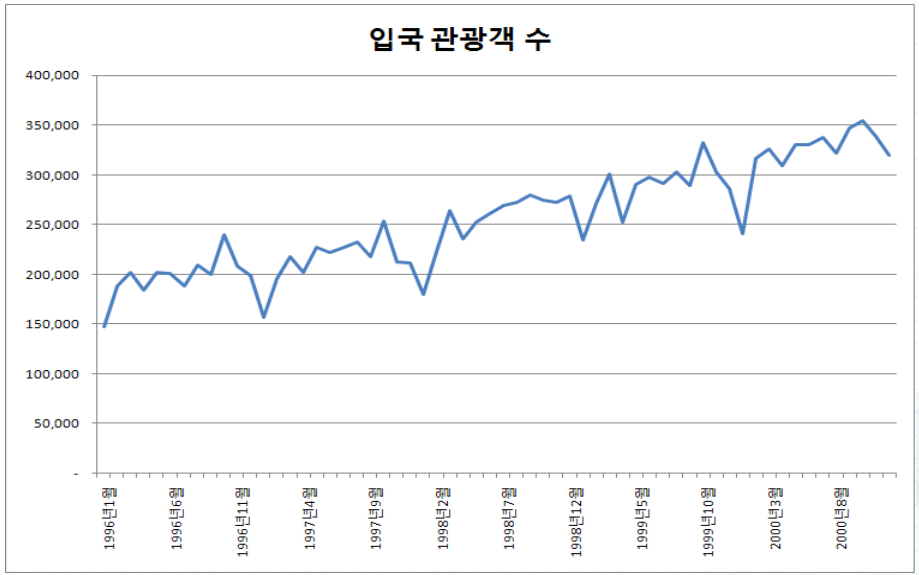
항목별 비교 ▶ 세로 막대형, 가로 막대형



2 엑셀 그래프의 종류 및 특성

엑셀 그래프의 종류(3)

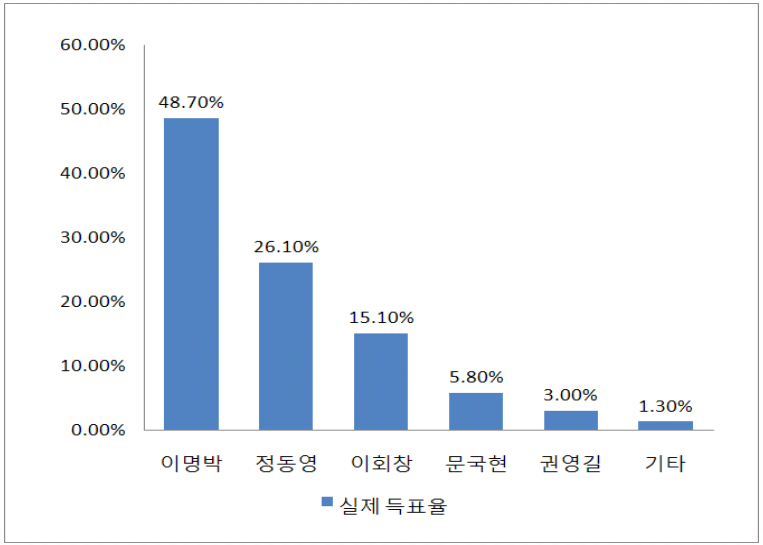
시간적 추이 ▶ 꺾은선형, 세로 막대형

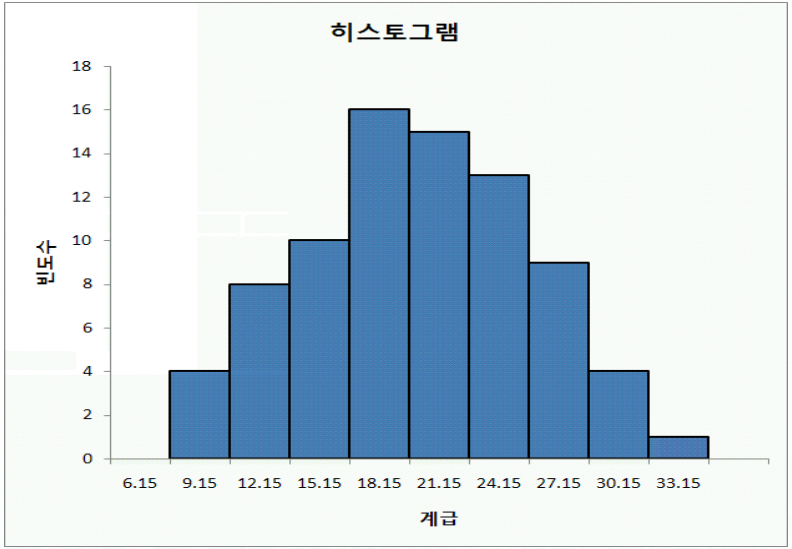


2 엑셀 그래프의 종류 및 특성

엑셀 그래프의 종류(4)

☑ **도수 분포** ▶ 세로 막대형, 꺾은선형

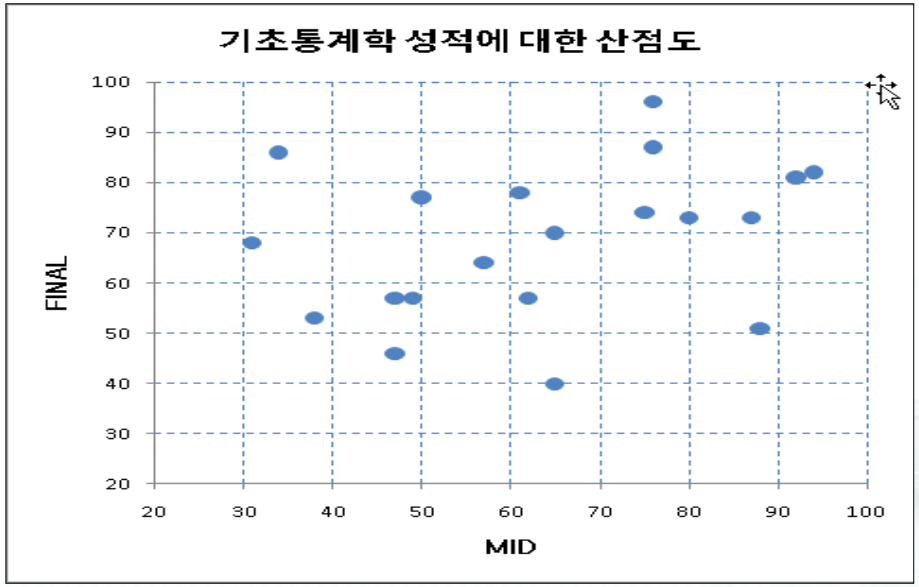




2 엑셀 그래프의 종류 및 특성

엑셀 그래프의 종류(5)

연관성 검토 ▶ 분산형



3 차트의 작성

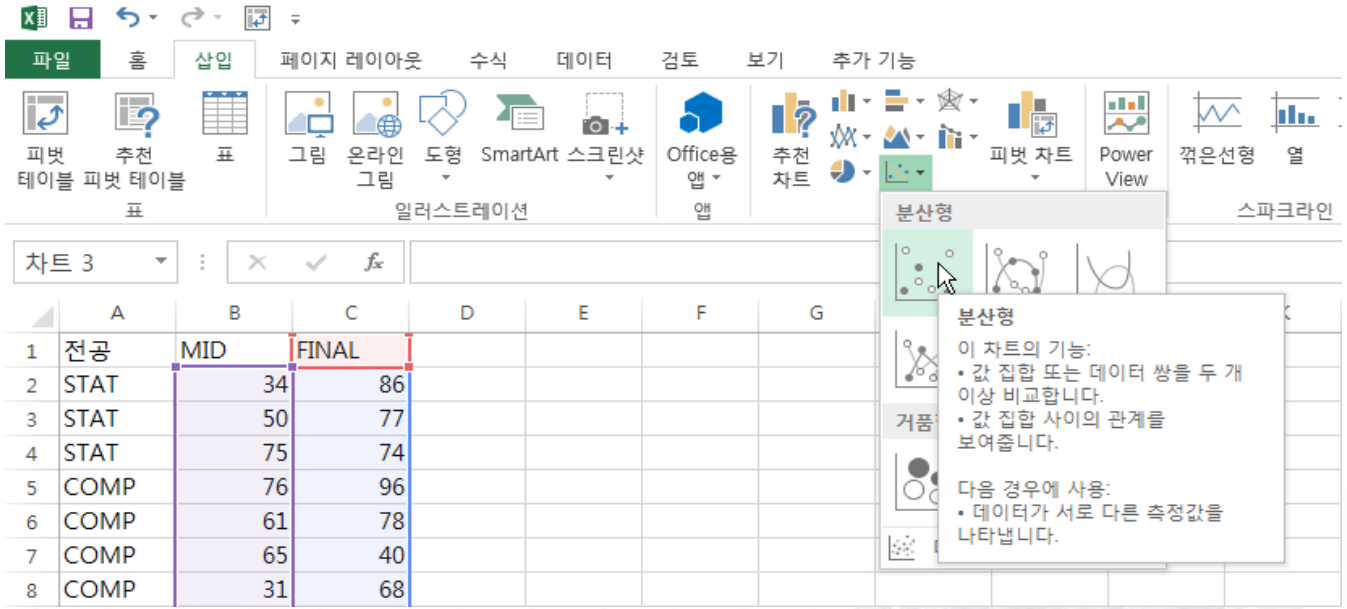
3 차트의 작성

차트 작성 과정

1. 차트로 나타내고자 하는 데이터 선택
2. [삽입] 리본탭의 [차트] 그룹에서 알맞은 차트 유형 선택
3. 차트를 구성하는 각 구성요소의 옵션 조정
   - 데이터에 적합한 보기 좋은 차트 작성

3 차트의 작성

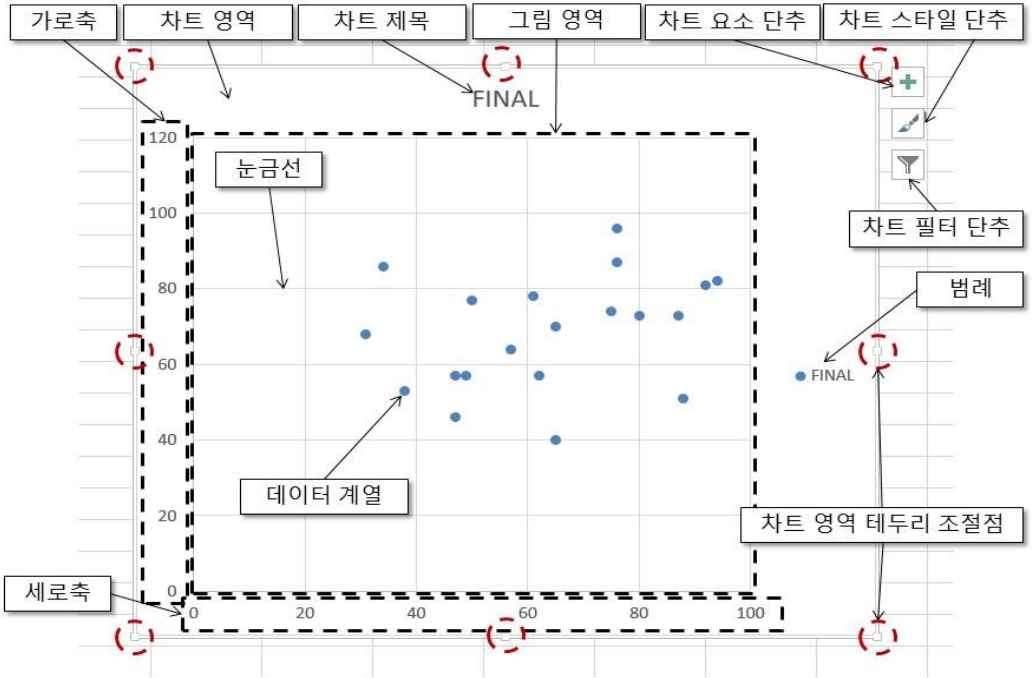
분산형 차트 그리기의 예



	A	B	C	D	E	F	G
1	전공	MID	FINAL				
2	STAT	34	86				
3	STAT	50	77				
4	STAT	75	74				
5	COMP	76	96				
6	COMP	61	78				
7	COMP	65	40				
8	COMP	31	68				

3 차트의 작성

차트의 구성요소 1



3 차트의 작성

차트의 구성요소 2

차트 영역	모든 차트 구성요소들의 배경이 되는 영역
그림 영역	차트 영역 내부에 있고, 실제 차트를 포함하고 있는 프레임
차트 제목	차트의 제목 표시
세로축/ 가로축	Y축/X축의 단위 및 제목 표시

3 차트의 작성

차트의 구성요소 3

범례 (legend)

여러 계열의 데이터를 이용하는 경우에 각 계열의 구분을 위해 사용

눈금선

데이터 값을 정확히 알 수 있게 하는 기능,
주 눈금선과 보조 눈금선이 있음

데이터 계열

각각의 데이터, 하나하나를 가리킴

3 차트의 작성

차트의 편집

- 차트가 선택된 상태에서 나타나는 [차트 도구] 리본탭 이용
- 차트의 해당 구성 요소를 마우스 오른쪽 단추를 클릭하여 나타나는 [미니 도구 모음] 또는 [단축 메뉴]를 사용할 수도 있음
- 차트의 각 구성요소를 독립적으로 편집할 수 있음

3 차트의 작성

히스토그램 그리기

- 연속형 데이터의 도수분포표를 그래프로 표현
- 수평축에 계급구간을 표시하고 각 계급의 상대도수에 비례하는 넓이의 직사각형을 그린 것
- 엑셀에서 기본적으로 제공되는 차트에는 히스토그램이 없음
- [데이터] 리본탭의 [분석] 그룹의 [데이터 분석] 이용

연습문제 1.

1. 논리함수를 사용하여 다음 식에 의해 출력되는 것은?

(1) =NOT(FALSE)

(2) =OR(TRUE, FALSE)

(3) =OR(2+ 2=3, 3-2=2)

(4) =NOT(AND(2+ 2=3, 3-2=2))

연습문제 2.

2. 다음 워크시트에서 ‘통계학개론', ‘조사방법론'이 각각 60점 이상이고, ‘평균’이 70점 이상일 때에만 “합격”을 표시하고 나머지의 경우에는 “불합격”을 표시하고자 한다. 이를 위해 E2 셀에 입력해야 할 수식은?

	A	B	C	D	E
1	이름	통계학개론	조사방법론	평균	판정
2	홍길동	67	80	73.5	
3	임꺽정	56	78	67.0	

연습문제 3.

3. 다음 설명 중에서 옳은 것을 모두 고른 것은?

Ⅰ.여론조사의 각 정당별 지지율은 항목별 비교로 볼 수 있어 세로 막대형으로 그릴 수 있다.

Ⅱ.두 변수의 연관성을 검토하기 위해서는 분산형 차트를 이용한다.

Ⅲ.연도별 입국 관광객 수와 같이 시간적 추이변화를 보기 위해서는 방사형 차트를 이용한다.

① Ⅰ, Ⅱ
② Ⅰ, Ⅲ
③ Ⅱ, Ⅲ
④ Ⅰ, Ⅱ, Ⅲ

3강

데이터의 그래프 표현과 수치 요약

정보통계학과 이기재 교수

학습목차

학습목표

- 중심위치 측도의 개념 이해
- 산포 측도의 개념 이해
- 상대적 위치 측도의 개념 이해
- 엑셀과 KESS를 이용해 구한 요약통계량으로 분포의 특징 설명

1 데이터의 수치 요약

1 데이터의 수치 요약

데이터의 수치요약

- 자료를 객관적으로 파악할 수 있다는 장점이 있음
- 도수분포표나 그래프 등을 이용해서 분포 상태를 확인하고,

  수치적 측도를 이용해서 자료를 요약 · 정리하는 것이 바람직함
- 중심위치의 측도와 산포의 측도가 대표적인 수치임

1 데이터의 수치 요약

중심위치 측도의 종류

- ☑ 평균(mean, average)
- ☑ 중앙값(median)
- ☑ 최빈값(mode)

1 데이터의 수치 요약

중심위치 측도의 종류(1)

☑ 평균(mean, average)

- 양적 자료에 사용
- 평균 $= \overline{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$
- 엑셀 AVERAGE 함수 이용
- 아주 크거나 작은 극단값의 영향을 많이 받음

1 데이터의 수치 요약

중심위치 측도의 종류(1)

☑ α% 절사 평균(trimmed mean)

- 데이터 중심에서 멀리 벗어난 값들을 양쪽에서 각각 α%씩 제외하고 평균 계산
- 극단값의 영향을 덜 받음
- TRIMMEAN 함수

  참고 : "10% 절사 평균 계산" ⇔ =TRIMMEAN(array, 0.2)

1 데이터의 수치 요약

중심위치 측도의 종류(2)

☑ 중앙값(median)

- 데이터를 크기 순으로 나열할 때 가운데 위치한 값
- 중앙값 = $\begin{cases} (n+1)/2\text{번째 위치한 자료 값}, & n\text{이 홀수} \\ n/2\text{번째와 } n/2+1\text{번째 자료 값의 평균}, & n\text{이 짝수} \end{cases}$
- 엑셀 MEDIAN 함수 이용
- 아주 크거나 작은 극단값에 영향을 거의 받지 않음

1 데이터의 수치 요약

중심위치 측도의 종류(3)

☑ 최빈값(mode)

- 주로 질적 자료에서 사용함
- 데이터 중에서 빈도가 가장 높은 값
- 엑셀 MODE 함수 이용

1 데이터의 수치 요약

중심위치 측도 계산을 위한 엑셀 함수들

함 수	설 명
AVERAGE (number1, [number2],…)	자료의 표본평균 계산
TRIMMEAN(array, percent)	절사평균 계산 α = percent/2
MEDIAN (number1, [number2],…)	자료(number1, number2,…)의 중앙값 계산
MODE (number1, [number2],…)	자료(number1, number2,…)의 최빈값 계산

1 데이터의 수치 요약

산포 측도의 종류

- ☑ 분산(variance)과 표준편차(standard deviation)
- ☑ 범위(range)
- ☑ 사분위수범위(interquartile range : IQR)
- ☑ 변동계수(coefficient of variation : CV)

1 데이터의 수치 요약

산포 측도의 종류(1)

분산(variance)과 표준편차(standard deviation)

- 분산(variance)

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

  - VAR 함수 이용

- 표준편차(standard deviation)

$$S = \sqrt{S^2}$$

  - STDEV 함수 이용

1 데이터의 수치 요약

산포 측도의 종류(2)

☑ 범위(range)

- 데이터의 최대값과 최소값 차이
- 범위 = MAX - MIN
- 이상점이 있는 경우에는 부적절함

1 데이터의 수치 요약

산포 측도의 종류(3)

사분위수범위(interquartile range : IQR)

- 위사분위수와 아래사분위수의 차이
  - 사분위수범위 = Q3 - Q1
- 이상점의 영향을 적게 받음
- QUARTILE(array, quart) 이용
  quart = 0(최소값), 1, 2, 3, 4(최대값)

1 데이터의 수치 요약

산포 측도의 종류(4)

☑ 변동계수(coefficient of variation : CV)

- $cv = \frac{s}{\bar{x}} \times 100(\%)$
- 평균이 크게 다른 집단의 산포 정도 비교에 사용함
- STDEV 함수와 AVERAGE 함수 이용

1 데이터의 수치 요약

상대적 위치의 측도

백분위수

- 어떤 숫자들의 집합에서 제 c 백분위수 (c–th percentile)는 수들의 c%가 그 값보다 작고 나머지는 그 값보다 큰 값을 말함

데이터의 z-값(표준화 점수)

- $$z = \frac{x - \bar{x}}{s}$$

1 데이터의 수치 요약

상대적 위치의 측도 계산을 위한 엑셀 함수들

함 수	설 명
PERCENTILE (array, *k*)	자료(array)의 제 *k* 백분위수 계산 단, 0 < *k* < 1
PERCENTRANK (array, *x*, significance)	자료(array)에서 *x*의 백분율 상대 위치를 계산 significance는 계산할 백분율 값의 유효 자리 수 개수를 나타냄
STANDARDIZE (*x*, mean,	평균(mean)과 표준편차(standard_dev)를 이용하여 *x*의 *z* -값 계산

1 데이터의 수치 요약

기술통계량 구하기

1. [데이터] 리본 탭의 [분석] 그룹에서 [데이터 분석] 선택
2. [통계 데이터 분석] 대화상자에서 [기술 통계법] 선택
3. 입력범위에 데이터 범위 지정
4. 첫째 행을 이름표 사용(L) 선택
5. 요약통계량(S) 선택

1 데이터의 수치 요약

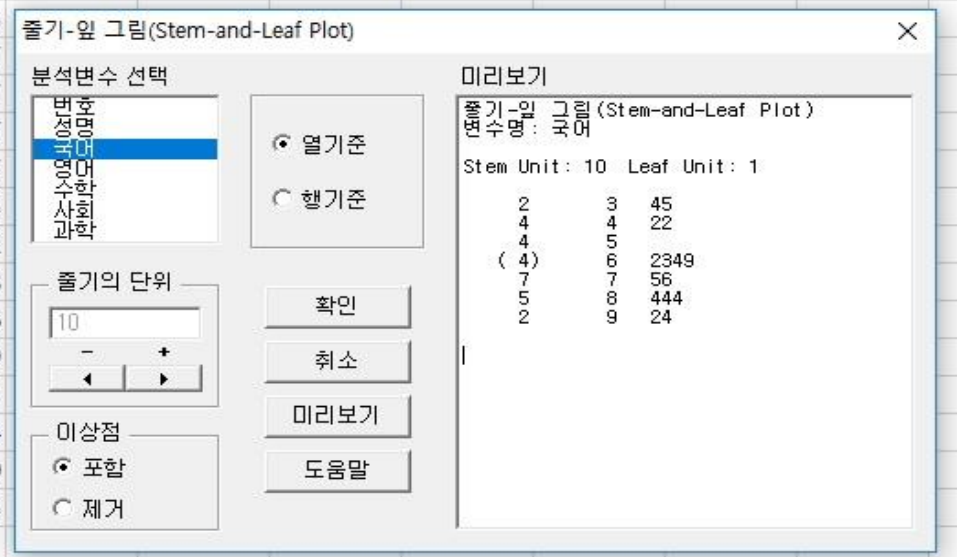
줄기-잎 그림 (stem-leaf plot)

- 연속형 데이터의 분포형태를 살펴보기 위해 사용
- 계급의 기둥 높이가 분포 정도를 나타내지만, 높이를 나타낼 때 데이터 관측값의 끝자리 숫자 이용

1 데이터의 수치 요약

줄기-잎 그림 (stem-leaf plot)

	A	B	C	D	E	F	G
1	번호	성명	국어	영어	수학	사회	과학
2	1	정연도	84	75	91	84	94
3	2	한숙희	64	45			
4	3	이명옥	69	57			
5	4	박정용	84	57			
6	5	안승제	76	37			
7	6	박선미	92	67			
8	7	이민영	94	85			
9	8	이원종	62	52			
10	9	김수정	35	38			
11	10	김한준	75	36			
12	11	심경택	42	29			
13	12	최선욱	34	31			
14	13	배희열	84	54			
15	14	장윤정	63	40			
16	15	이진홍	42	55			
17							
18							



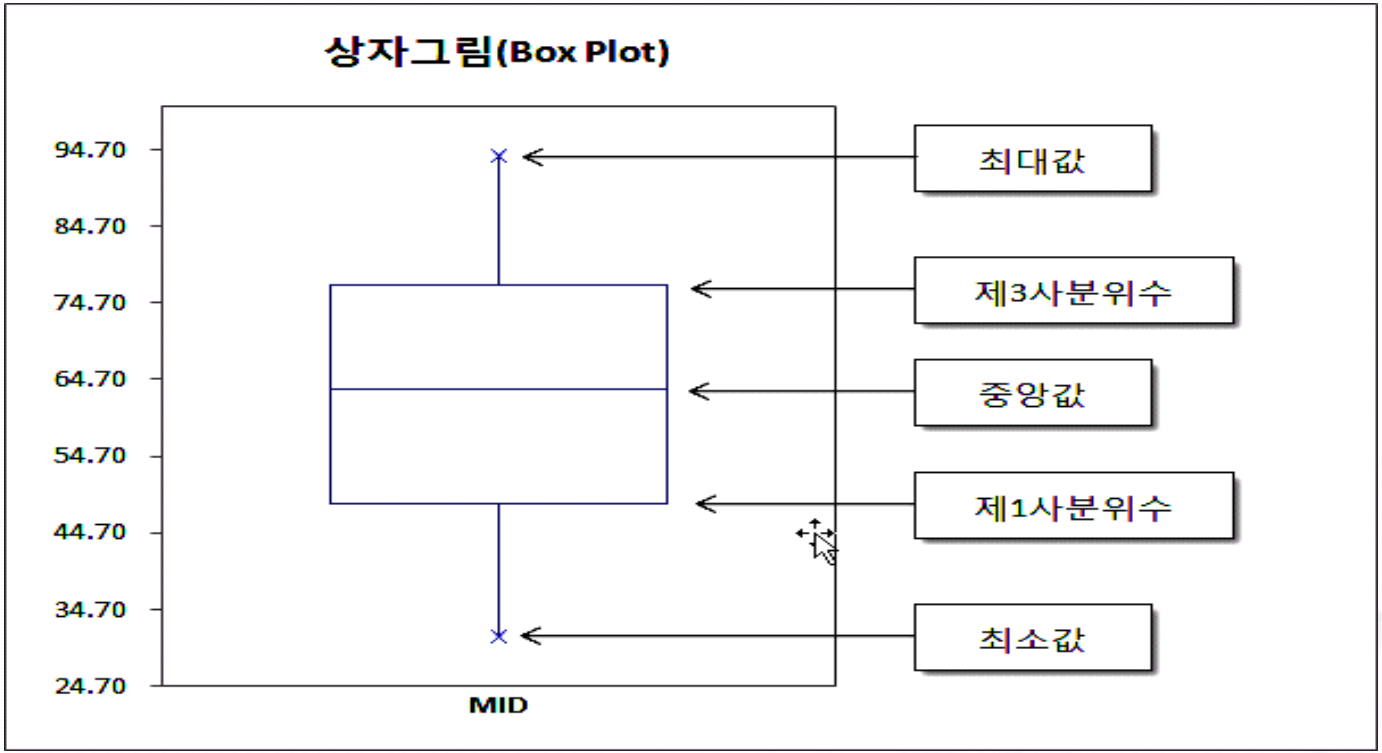
1 데이터의 수치 요약

상자그림 (box plot)

- 다섯 숫자 요약을 이용해서 데이터의 분포 형태를 표시함
- 이상치를 찾는데 유용함
- 여러 데이터 계열의 분포 형태를 비교할 때 유용함

1 데이터의 수치 요약

상자그림 (box plot)



연습문제 1.

1. 다음 설명 중에서 올바른 것을 모두 고른 것은?

Ⅰ.절사평균은 평균에 비해서 특이치(outlier)의 영향을 적게 받는다.
Ⅱ.표준편차는 사분위수 범위에 비해서 특이치(outlier)의 영향을 적게 받는다.
Ⅲ.범위는 사분위수 범위에 비해서 특이치(outlier)의 영향을 적게 받는다.

① Ⅰ
② Ⅰ, Ⅱ
③ Ⅰ, Ⅲ
④ Ⅱ, Ⅲ

연습문제 2.

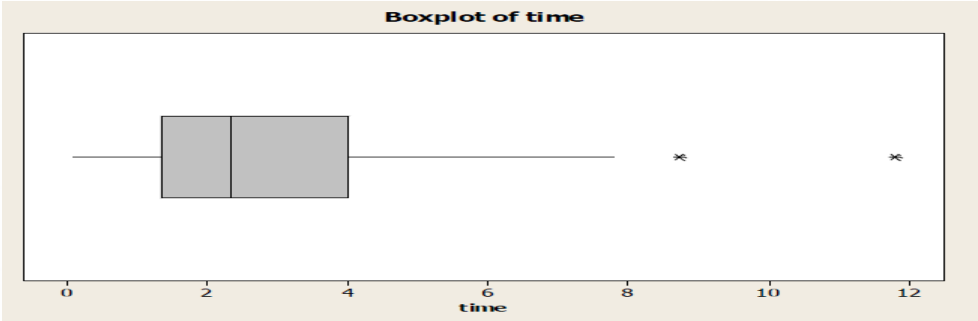
2. 어느 대학교에서 전체 학생을 대상으로 0점부터 100점까지 표시되는 직업 적성 시험을 치렀다. 전체 학생에 대한 평균과 표준편차는 각각 53과 16이다. A 학생의 표준화 점수(z-score)는 1.19이고, B 학생의 표준화 점수(z-score)는 2.06이었다면 B는 A보다 몇 점을 더 많이 얻은 것인가?

① 0.87　　② 14

③ 19　　④ 33

연습문제 3.

3. 어느 연구자는 멈춤 신호가 있는 교차로에서 300명의 운전자를 대상으로 몇 분을 기다리게 되는가를 조사하여 이를 상자그림으로 표시하였다. 다음 설명 중에서 옳지 않은 것은?



① 기다리는 시간의 중위수(중앙값)은 약 2.3이다.

② 전체 운전자 중 25%는 멈춤 신호에서 4분 이상 기다리게 된다.

③ 기다리는 시간의 평균은 중위수보다 작은 값을 갖게 된다.

④ 주어진 상자그림에서 이상치로 두 개의 관측치가 있다.

4강 엑셀 데이터베이스 기능과 해 찾기

정보통계학과 이기재 교수

학습목차

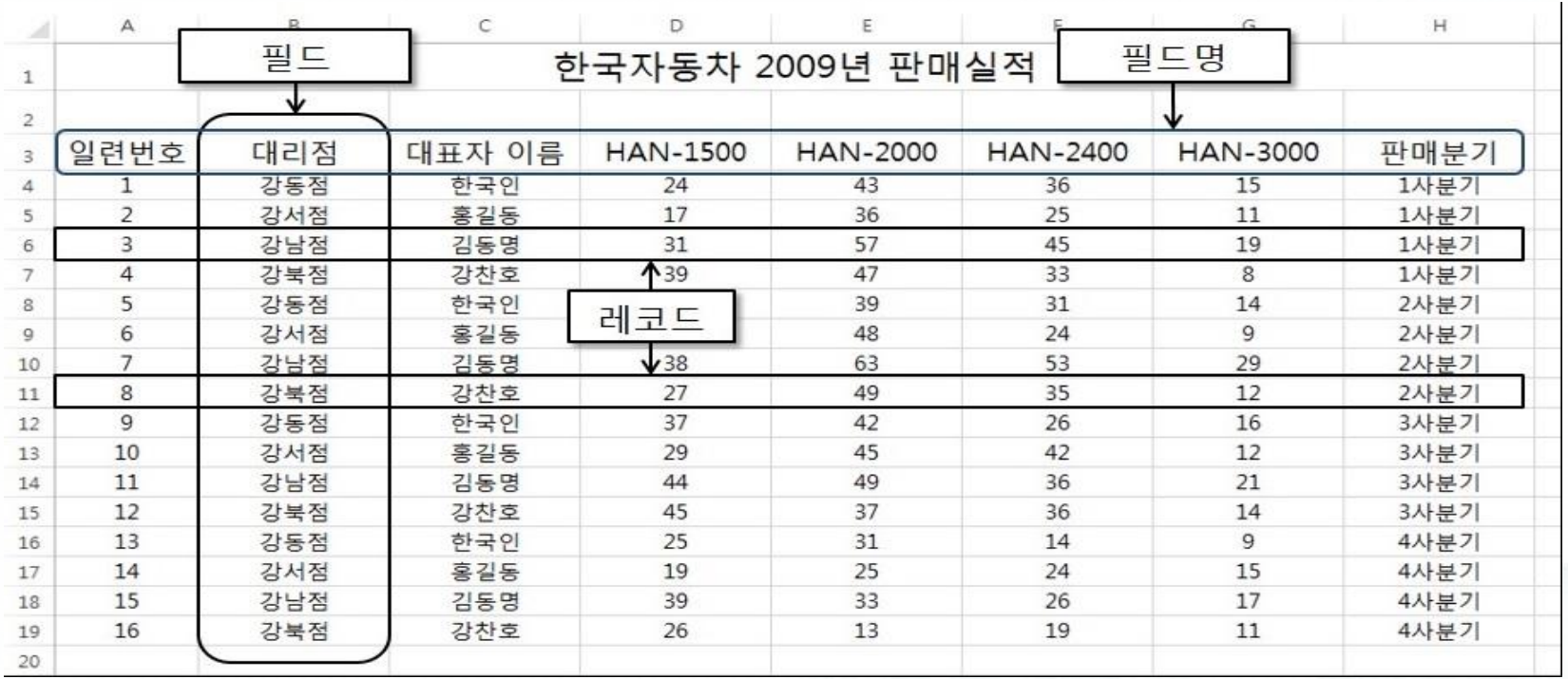
학습목표

- 데이터 목록의 개념을 이해하고 작성
- 데이터 목록의 정렬 및 필터 기능의 활용
- 부분합 기능을 이용한 데이터 목록의 요약
- 목표값 찾기와 해 찾기 기능의 활용

1 데이터 목록 작성 및 관리

1 데이터 목록 작성 및 관리

데이터 목록의 주요 용어



일련번호	대리점	대표자 이름	HAN-1500	HAN-2000	HAN-2400	HAN-3000	판매분기
1	강동점	한국인	24	43	36	15	1사분기
2	강서점	홍길동	17	36	25	11	1사분기
3	강남점	김동명	31	57	45	19	1사분기
4	강북점	강찬호	39	47	33	8	1사분기
5	강동점	한국인	39	31	14	2사분기	
6	강서점	홍길동	48	24	9	2사분기	
7	강남점	김동명	38	63	53	29	2사분기
8	강북점	강찬호	27	49	35	12	2사분기
9	강동점	한국인	37	42	26	16	3사분기
10	강서점	홍길동	29	45	42	12	3사분기
11	강남점	김동명	44	49	36	21	3사분기
12	강북점	강찬호	45	37	36	14	3사분기
13	강동점	한국인	25	31	14	9	4사분기
14	강서점	홍길동	19	25	24	15	4사분기
15	강남점	김동명	39	33	26	17	4사분기
16	강북점	강찬호	26	13	19	11	4사분기

1 데이터 목록 작성 및 관리

데이터 목록 이용의 장점

- 데이터 구성을 쉽게 이해할 수 있고, 작업이 효율적임
- 정렬, 필터 기능을 사용한 빠른 검색 기능
- 데이터의 추가, 삭제가 편리함

1 데이터 목록 작성 및 관리

데이터 목록의 작성 방법

- 워크시트에는 한 종류의 데이터만 입력
- 같은 열에는 같은 종류의 데이터 입력
- 데이터 목록의 첫 행에 열 이름을 입력
- 데이터에 지정한 서식과는 다른 글꼴 등 사용
- 문자 입력에서 셀 시작 부분에 필요 없는 공백을 없앰

  (공백은 정렬과 찾기에 영향을 줌)

2 데이터 목록의 정렬 및 필터

2 데이터 목록의 정렬과 필터

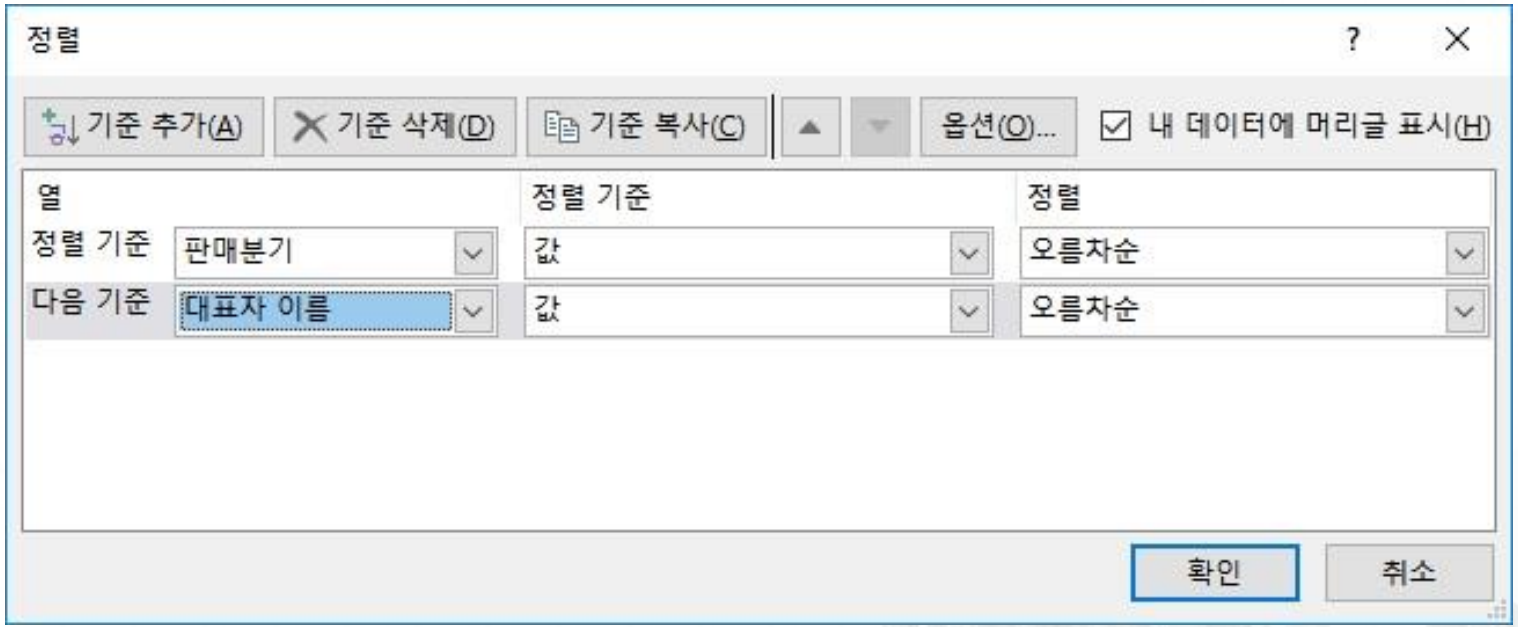
데이터 정렬 방법

- 데이터 목록 내의 하나의 셀 선택
- [데이터] 리본 메뉴 → [정렬 및 필터] 리본 탭 → [정렬] 선택
- 정렬 대화 상자에서 필요한 사항 선택
  - 정렬방법 : 오름차순 , 내림차순
- [기준추가] 단추를 누르면,

  원하는 추가적인 기준 필드를 정할 수 있음

2 데이터 목록의 정렬과 필터

데이터 정렬 방법



2 데이터 목록의 정렬과 필터

필터(filter)의 사용

기 능

데이터 목록에서 어떤 조건을 만족하는
데이터만을 보여 주는 기능

이용 방법

- **자동 필터**

  [데이터] 리본 메뉴 → [정렬 및 필터] 리본 탭 → [필터] 선택

- **고급 필터**

  [데이터] 리본 메뉴 → [정렬 및 필터] 리본 탭 → [고급] 선택

2 데이터 목록의 정렬과 필터

필터(filter)의 사용

☑ 고급필터

- 복합조건이나 계산 조건을 설정하여 목록을 쉽게 골라 낼 수 있음
- 고급 필터에서 '그리고'나 '또는'의 조건식은
  같은 행 또는 다른 행에 입력하느냐에 따라 결정됨

2 데이터 목록의 정렬과 필터

고급 필터 사용의 예

한국자동차 2009년 판매실적

일련번호	대리점	대표자 이름	HAN-1500	HAN-2000	HAN-2400	HAN-3000	판매분기
	강남점			>=40		>=20	
일련번호	대리점	대표자 이름	HAN-1500	HAN-2000	HAN-2400	HAN-3000	판매분기
1	강동점	한국인	24	43			1사분기
2	강서점	홍길동	17	36			1사분기
3	강남점	김동명	31	57			1사분기
4	강북점	강찬호	39	47			1사분기
5	강동점	한국인	19	39			2사분기
6	강서점	홍길동	24	48			2사분기
7	강남점	김동명	38	63			2사분기
8	강북점	강찬호	27	49			2사분기
9	강동점	한국인	37	42			3사분기
10	강서점	홍길동	29	45			3사분기
11	강남점	김동명	44	49			3사분기
12	강북점	강찬호	45	37	36	14	3사분기
13	강동점	한국인	25	31	14	9	4사분기
14	강서점	홍길동	19	25	24	15	4사분기
15	강남점	김동명	39	33	26	17	4사분기
16	강북점	강찬호	26	13	19	11	4사분기

고급 필터

결과
- (●) 현재 위치에 필터(F)
- ( ) 다른 장소에 복사(O)

목록 범위(L): $A$5:$H$21
조건 범위(C): $A$2:$H$3
복사 위치(T):
☐ 동일한 레코드는 하나만(R)

확인 취소

한국자동차 2009년 판매실적

일련번호	대리점	대표자 이름	HAN-1500	HAN-2000	HAN-2400	HAN-3000	판매분기
	강남점			>=40		>=20	
일련번호	대리점	대표자 이름	HAN-1500	HAN-2000	HAN-2400	HAN-3000	판매분기
7	강남점	김동명	38	63	53	29	2사분기
11	강남점	김동명	44	49	36	21	3사분기

3 부분합 계산 및 데이터 베이스 함수

3 부분합 계산 및 데이터 베이스 함수

부분합 기능의 사용

기 능

- 그룹화 할 항목에 따라 부분적인 합계 계산
- 데이터 목록의 요약, 정리

이용 방법

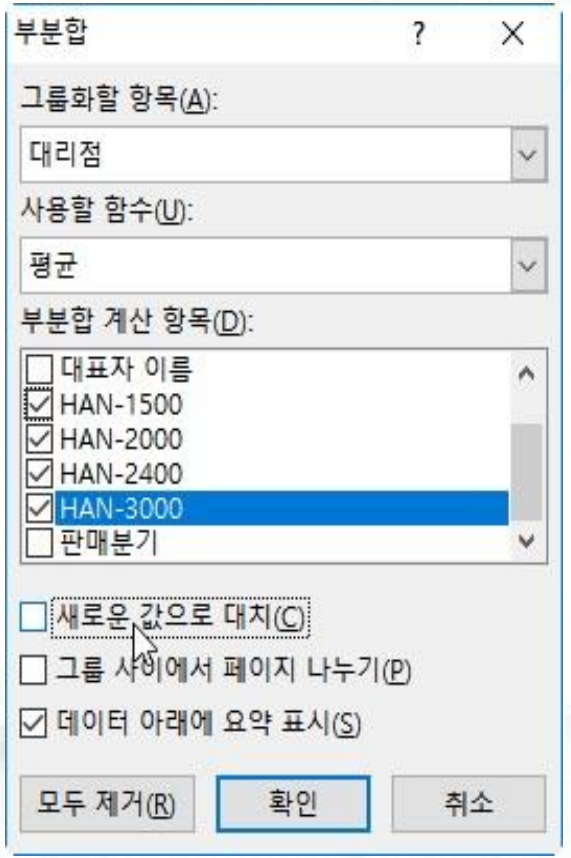
[데이터] 리본 메뉴 → [윤곽선] 리본 탭 → [부분합] 선택

3 부분합 계산 및 데이터 베이스 함수

부분합 기능 이용절차

- 데이터 목록의 정렬
  - 부분합을 구하려는 항목끼리 그룹화
- [데이터] 리본 메뉴

  → [윤곽선] 리본 탭 → [부분합] 선택
- 부분합 대화 상자에서 그룹화할 항목(A)과 사용할 함수(U) 선택



3 부분합 계산 및 데이터 베이스 함수

데이터베이스 함수의 사용

- 사용자가 원하는 조건에 따라서 데이터 목록을 요약하고, 검색할 수 있음
- 데이터베이스 관련 함수 형식
  - = 데이터베이스 함수명 (데이터베이스 범위, 필드명, 비교 조건 범위)
    - 예 : DAVERAGE (데이터베이스 범위, 필드명, 비교 조건 범위)

3 부분합 계산 및 데이터 베이스 함수

데이터베이스 함수의 사용

함 수	설 명
DSUM(database, field, criteria)	필드의 합을 구한다.
DAVERAGE(database, field, criteria)	필드의 평균을 구한다.
DCOUNT(database, field, criteria)	필드에서 숫자를 포함한 셀의 수를 구한다.
DGET(database, field, criteria)	찾을 조건에 맞는 레코드를 추출한다.
DMAX(database, field, criteria)	필드 값 중에서 최대값을 구한다.
DMIN(database, field, criteria)	필드 값 중에서 최소값을 구한다.
DPRODUCT(database, field, criteria)	필드 값들의 곱을 구한다.
DSTDEV(database, field, criteria)	필드 값들의 표본표준편차를 구한다.
DVAR(database, field, criteria)	필드 값들의 표본분산을 구한다.

4 엑셀의 해 찾기 기능

4 엑셀의 해 찾기 기능

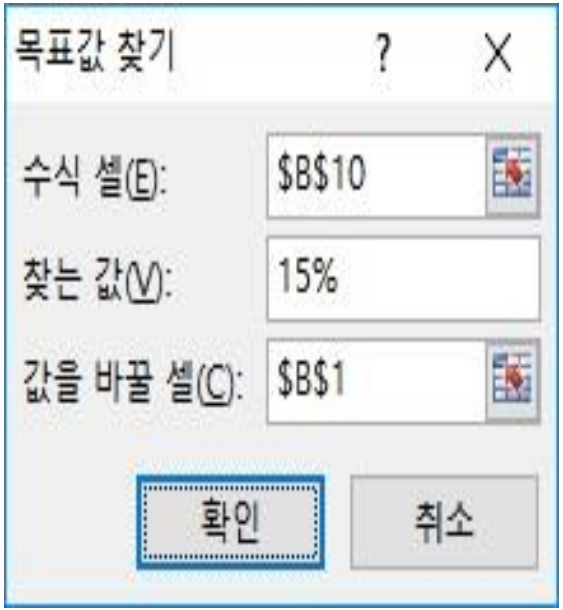
목표값 찾기

- 수식의 결과값이 목표하는 값과 같아지도록 입력값을 조정하여 찾는 기능
- [데이터] 리본 메뉴 → [데이터 도구] 리본 탭

  → [가상분석] 선택 → [목표값] 선택

4 엑셀의 해 찾기 기능

목표값 찾기 대화상자



- **수식 셀**
  셀 주소나 이름 중 하나를 입력
- **찾는 값**
  수식의 결과로 얻을 값 즉, 목표값 입력
- **바꿀 셀**
  목표값을 찾을 때까지 조정하려는 변수가 있는 셀 입력

4 엑셀의 해 찾기 기능

목표값 찾기 예 : 컴퓨터용 소모품 대리점

- 디스켓 가격 : 300원(공장도), 400원(소비자)
- 한 달 평균 10,000 장의 디스켓 판매
- 총수입 = 판매가×판매개수 = 4,000,000원

  총비용 = 원가×판매개수+고정비용 = 3,250,000원
- 이익 = 총수입-총비용 = 750,000원,

  이익률 = 이익/총수입 = 19%

➡ **만약 디스켓 판매의 이익률을 15%로 낮추려면 소매점 판매가는?**

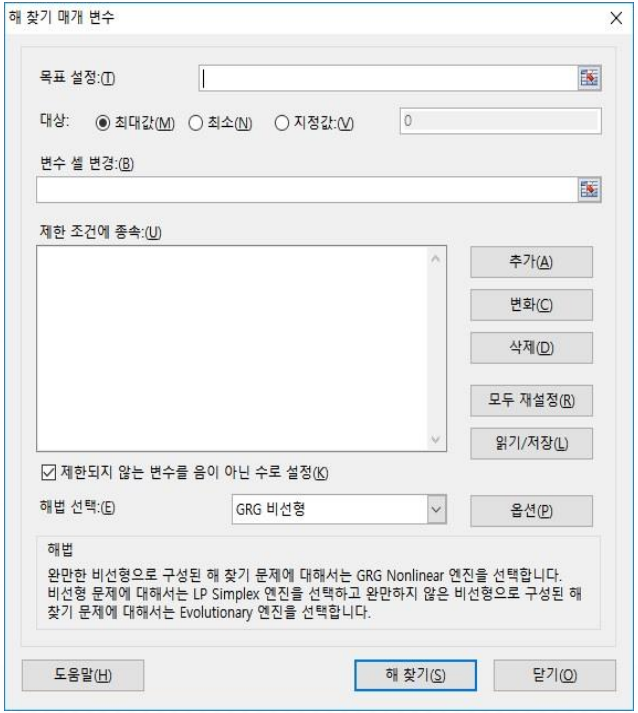
4 엑셀의 해 찾기 기능

해 찾기 기능

- 입력변수가 2개 이상인 경우 최적값을 만드는 입력변수 값들을 찾는 기능
- [해 찾기] 기능은 Microsoft Office Excel에서 추가적으로 설치되어야 함

4 엑셀의 해 찾기 기능

해 찾기 대화상자



목표 셀

최대화, 최소화, 또는 지정값으로 만들려는 셀

값을 바꿀 셀

목표 셀을 최적화하기 위해서 변경할 셀

제한 조건

최적화를 위한 입력 셀들의 값이 만족해야 할 제한 조건을 입력함

4 엑셀의 해 찾기 기능

해 찾기 예

- 가구 제조 공장의 생산 현황

	식탁용 의자	식탁	가용자원
목재 (board feet)	5	20	400
노동시간 (시간)	10	15	450
이익 (천원/개)	45	80	

- 의자와 식탁을 각각 얼마씩 생산하면 이익이 최대가 되는가?

4 엑셀의 해 찾기 기능

해 찾기 예제에 대한 수식 표현 1

가구 제조 공장의 생산 현황

	식탁용 의자	식탁	가용자원
목재 (board feet)	5	20	400
노동시간 (시간)	10	15	450
이익 (천원/개)	45	80	
생산량	X_1	X_2	

4 엑셀의 해 찾기 기능

해 찾기 예제에 대한 수식 표현 2

- 제약식

$$5 \times X_1 + 20 \times X_2 \leq 400$$

$$10 \times X_1 + 15 \times X_2 \leq 450$$

$$X_1 , X_2 \geq 0$$

- 총이익 $45 \times X_1 + 80 \times X_2$ 가 최대가 되도록 X_1 , X_2 를 정하는 문제

연습문제 1.

1. 다음의 고급 필터 검색조건을 통해서 추출되는 레코드는 몇 건인가?

<데이터 목록>

	A	B	C	D	E
1	대표자 이름	HAN-1500	HAN-2000	HAN-2400	HAN-3000
2	김동명	31	57	45	19
3	한국인	24	43	36	15
4	강찬호	39	47	33	8
5	홍길동	17	36	25	11

<검색 조건>

HAN-1500	HAN-2000	HAN-3000
>=30	>=50	
		>=15

① 없음　　② 1건

③ 2건　　④ 3건

연습문제 2.

2. 어느 부분합을 구하기에 앞서 해야 할 작업을 설명한 것이다. 올바른 것은?

① 부분합을 구하는 그룹화 항목에 대하여 필터링 되어야 한다.

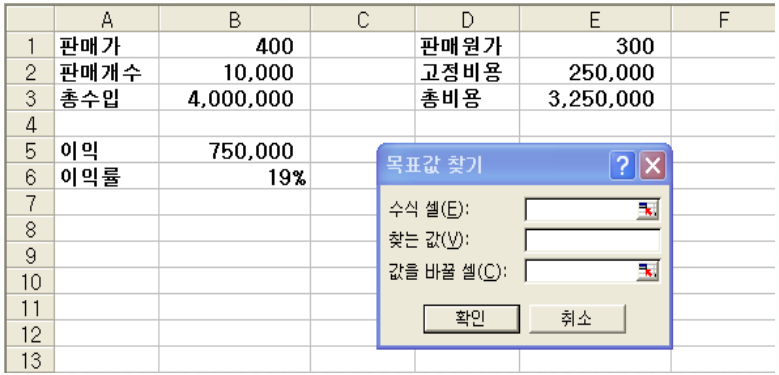
② 그룹화 항목에 대하여 오름차순이나 내림차순으로 정렬해야 한다.

③ 부분합을 구하는 그룹화 항목이 첫 번째 필드에 있어야 한다.

④ 부분합을 구하는 그룹화 항목이 반드시 숫자이어야 한다.

연습문제 3.

3. 목표값 찾기를 이용해서 이익률을 25%로 높이고자 한다. B6 셀에 =B5/B3의 수식이 입력되어 있다. 판매가를 얼마로 해야 되는지 찾고자 한다. 대화상자의 '수식 셀', '찾는 값', '값을 바꿀 셀'의 내용을 순서대로 올바르게 나열한 것은 어느 것인가?



① $B$6, 15%, $B$1

② $B$6, $B$2, $B$1

③ $B$6, $B$1, $B$2

④ $B$6, 15%, $E$1

5강

확률 분포 (1)

정보통계학과 이기재 교수

학습목차

4.1 확률변수와 분포함수

- 확률의 개념에 대한 이해
- 확률변수와 분포함수에 대한 이해

1 확률 변수와 분포 함수

1 확률변수와 분포함수

확률(Probability)의 개념

- 공정한 주사위를 던질 때 6의 눈이 나올 확률은 얼마일까?
- 윷놀이를 할 때 윷이 나올 확률은 얼마일까?
- 주사위를 던질 때 6의 눈이 나올 확률이 1/6이라는 의미는?
- 처음 몇 번 던지면 6의 눈이 나오는 비율이 1/6로 나타나지 않지만 던지는 횟수를 늘리면 결국 6의 눈이 나오는 비율은 1/6로 가까이 간다.

확률이란?

동일한 상태에서 동일한 시행을 무한 번 반복한다고 할 때 궁극적으로 전체 시행 중에서 특정 사건이 발생할 비율을 나타내는 개념이다.

1 확률변수와 분포함수

확률적 실험

- 실험의 결과로서 일어날 수 있는 전체 결과를 실험 전에 알고 있고, 결과가 여러 가지로 나타날 수 있으며 반복할 수 있는 실험이나 현상

- 확률적(통계적) 실험
  "실험의 결과가 구체적으로 어떤 것인가는 알 수 없지만 전체 가능한 모든 결과들은 알고 있고 반복이 가능한 경우를 확률적(통계적)실험이라고 함"

1 확률변수와 분포함수

기본 용어

☑ 확률 (Probability)

- 아직 실현되지 않은 현상에 대하여
  그 현상의 실현가능성을 0과 1 사이의 숫자로 표현한 것

☑ 원소(element)

- 어떤 실험의 시행 결과로 나타날 수 있는 가능한 경우들

1 확률변수와 분포함수

확률적 실험의 예

동전을 한번 던지는 실험

- 원소 : H (=앞면) 또는 T (=뒷면)
- 표본공간 : $S = \{H, T\}$

주사위를 한 번 던지는 실험

- 원소 : 1, 2, 3, 4, 5, 6
- 표본공간 $S = \{1, 2, 3, 4, 5, 6\}$
- 사상 : 실험의 결과 짝수 눈이 나오는 경우

1 확률변수와 분포함수

확률의 고전적 정의

- 표본공간 내 각 원소의 발생 가능성이 같다고 하자. 사건 A가 발생할 확률은 다음과 같이 정의한다.

- $$P(A) = \frac{\text{사건 A에 속한 원소수}}{\text{표본공간의 전체 원소수}}$$ (이산형)

- $$P(A) = \frac{\text{사건 A에 속한 원소에 대한 측도}}{\text{표본공간의 전체 원소에 대한 측도}}$$ (연속형)

여기서, 측도란 길이, 면적, 부피 등을 뜻함

1 확률변수와 분포함수

확률 P(A)의 상대도수적 정의

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

단, : n_A사상 A가 일어나는 빈도수, n : 실험회수

- 상대빈도의 극한의 정의됨

1 확률변수와 분포함수

확률변수(Random variable)

- 확률적 실험의 결과에 수치를 대응시켜 주는 것

이산형 확률변수(discrete random variable)

- 취할 수 있는 값이 유한하거나 셀 수 있는 값을 취하는 확률변수

- 예 : 책의 페이지 당 오자의 수,
  제품 한 상자에서 나온 불량품의 개수,
  주사위를 5번 던질 때 짝수가 나오는 횟수

1 확률변수와 분포함수

확률변수(Random variable)

☑ 연속형 확률변수(continuous random variable)

- 어떤 구간의 값을 취할 수 있는 확률 변수
- 예 : 키, 몸무게, 강도, 성적 등

1 확률변수와 분포함수

확률변수 예: 동전을 4번 던지는 실험

- P{X=0} = P{(TTTT)} = 1/16
- P{X=1} = P{(HTTT), (THTT), (TTHT), (TTTH)} = 4/16
- P{X=2} = P{(HHTT), (HTHT), (HTTH), (THTH), (TTHH)} = 6/16
- P{X=3} = P{(HHHT), (HHTH), (HTHH), (THHH)} = 4/16
- P{X=4} = P{(HHHH)} = 1/16

확률분포표

x	0	1	2	3	4
P(X = x)	1/16	4/16	6/16	4/16	1/16

1 확률변수와 분포함수

확률변수의 예

- 200가구를 대상으로 지난 일 년 동안 각 가구에서 병원 방문 횟수 조사
- 확률변수를 X = '병원 방문 횟수'로 정의

병원방문 횟수	0	1	2	3	4	계
가구 수	74	80	30	10	6	200

x	0	1	2	3	4	계
P(X=x)	0.37	0.40	0.15	0.05	0.03	1

<참고> P(X=x)의 의미

확률변수 X= '병원 방문 횟수'가 x 일 확률을 의미함

1 확률변수와 분포함수

확률질량함수 및 확률밀도함수의 성질

확률질량함수 (이산형)

$$0 \le p(x) \le 1$$

$$P(a < X \le b) = \sum_{a<x\le b} p(x)$$

$$\sum_{\text{모든 } x} p(x) = 1$$

1 확률변수와 분포함수

확률질량함수 및 확률밀도함수의 성질

확률밀도함수 (연속형)

$$f(x) \geq 0$$

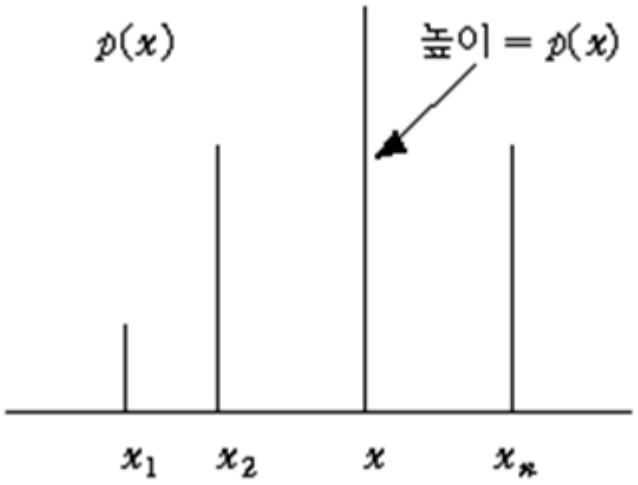
$$P(a < X \leq b) = \int_a^b f(x)dx$$

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

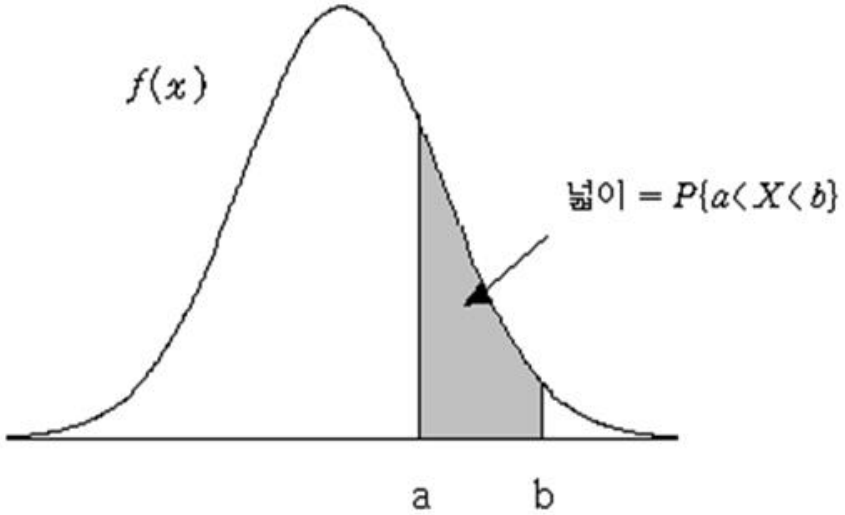
1 확률변수와 분포함수

확률함수의 분포

(a) 이산형



(b) 연속형



연습문제 1.

1. **어느 도시에서 한 가구에 살고 있는 성인 수 분포를 파악하고자 1,000가구를 단순임의추출하여 조사한 결과로 다음과 같은 확률분포표를 만들었다. 한 가구를 조사할 때 추출된 표본가구의 성인이 4명 이상일 확률은 얼마인가?**

성인 수(x)	1	2	3	4인 이상
$P(X=x)$	0.25	0.50	0.15	?

① 0.10 ② 0.15 ③ 0.20 ④ 0.25

연습문제 2

2. 다음 설명 중에서 올바른 것은 것은?

① 확률변수 값은 항상 양수이어야 한다.

② 확률변수의 기댓값은 항상 0보다 크거나 같다.

③ 확률변수의 분산은 항상 0보다 크거나 같다.

④ 확률변수의 기댓값은 항상 0보다 크거나 같다.

연습문제 3.

3. 다음 중 이산형 확률변수 X 에 대해서 항상 성립하는 것은?

① $E(X) = np$

② 확률변수 X 의 분포모양은 좌우대칭이다.

③ 확률변수 X 의 값을 모두 더하면 1이다.

④ $\sum_{\text{모든 } x_i} P(X = x_i) = 1$

6강

확률 분포 (2)

정보통계학과 이기재 교수

학습목차

4.2 이산 확률 분포

- 이항분포의 이해와 활용
- 포아송분포의 이해와 활용

1 이산확률분포

1 이산확률분포

베르누이 시행

- 각 시행의 결과는 성공(S) 또는 실패(F) 중의 하나로 나타남
- 매 시행의 성공확률을 $p=P(S)$ 로 나타내면 실패확률은 $q=1-p$ 임
- 각 시행은 독립적임

"시행의 성공여부가 다음 시행의 성공여부에 영향을 미치지 않음"

1 이산확률분포

이항분포

- 동일한 성공확률 p 를 가진 베르누이 시행을 n회 반복하여 시행할 때 성공 횟수를 확률변수 X라고 할 때, 확률변수 X는 이항분포(binomial distribution)를 따름
- $X \sim B(n, p)$ 로 표현
- $X_1, X_2, \ldots, X_n$을 성공확률 p , 서로 독립인 베르누이 확률변수라고 할 때, $X_1 + X_2 + \ldots + X_n$ 로 정의되는 확률변수 X는 이항확률변수임

1 이산확률분포

이항분포의 예

- 앞면의 확률이 p 인 동전을 n 회 던졌을 때 나타나는 앞면의 횟수
- 공정한 주사위를 n 회 던졌을 때 1의 눈이 나타나는 횟수
- 불량률 p 인 제품 더미 중에서 n 개를 추출하였을 때

  그 중에 포함되는 불량품의 수

1 이산확률분포

이항분포 정의

기 호

$X \sim B(n, p)$ 시행의 회수가 n 이고, 성공확률이 p

이항분포 분포함수

$$p(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x},$$

$$x = 0, 1, 2, \cdots, n$$

평균과 분산

$$E(X) = np, \quad Var(X) = np(1-p)$$

1 이산확률분포

이항분포 적용 예1

- 어떤 질병에 대한 치유율이 70%로 알려진 의약품을 이용하여 환자 10명을 치료할 때 적어도 7명의 환자가 치유될 확률은?

<풀이> 각 환자의 치유 확률은 0.7이며 실험의 결과로 나타나는 것은 성공 또는 실패로 나타남. 각 시행은 베르누이 시행임.
10명의 환자 중 치유 환자 수는 이항분포를 따름.

➡ "X=10명의 환자 중 치유된 환자 수"로 정의
➡ $X \sim B(10, 0.7)$
➡ $P(X \geq 7) = 1-P(X < 7) = 1-P(X \leq 6)$

1 이산확률분포

BINOMDIST(number_s, trials, probability_s, cumulat

number_s	성공한 횟수
trials	독립시행의 전체 횟수
probability_s	각 시행에서 성공 확률
cumulative	함수 형태를 결정하는 논리값 - 1 또는 TRUE: 누적확률값 계산 - 0 또는 FALSE: 확률값 계산

1 이산확률분포

BINOMDIST(number_s, trials, probability_s, cumulat

- 예 : $\mathrm{BINOMDIST}(x, n, p, 1)$

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

$$\mathrm{BINOMDIST}(x, n, p, 0)$$

$$= P(X = x) = \binom{n}{k} p^x (1-p)^{n-x}$$

1 이산확률분포

이항분포 적용 예2

- 어느 생산공정의 불량률이 5%라고 한다. 이 공정에서 임의로 10개를 추출하였을 때 이 중 불량품이 3개 이상 포함될 확률은?

<풀이> 10개 제품 중 포함될 불량품 수를 X로 정의하면 $B(10, 0.05)$를 따름

$$P(X \geq 3) = 1 - P(X \leq 2)$$

$$= 1 - \binom{10}{0}.05^1.95^{10} - \binom{10}{1}.05^1.95^9 - \binom{10}{2}.05^2.95^8$$

$$= 1 - 0.5987 - 0.3151 - 0.0746 = 0.0116$$

엑셀함수 '=1-BINOMDIST(2, 10, 0.05, 1)' 이용

1 이산확률분포

포아송분포가 적용될 수 있는 사례

- 어떤 사무실에 한 시간 동안에 전화가 걸려오는 횟수
- 고속도로 상에서 하루 동안에 발생하는 교통사고의 수
- 신문 1면 중의 오자의 개수

"어떤 특정한 사건이 일어날 확률이 아주 작은 경우에 적용되는 분포"

1 이산확률분포

포아송분포 적용 예1

- 어떤 보험회사에서 하루에 접수되는 보험금 청구건수는 평균 2건이다.

  (1) 어느 날 보험금 청구가 한 건도 없을 확률은?

  (2) 3건 이상의 청구가 있을 확률은?

<풀이> 확률변수 "X = 하루에 접수되는 보험금 청구건수"로 정의

확률변수 X는 대한 확률 계산은 포아송분포를 이용함

(1) $P(X = 0) = ?$

(2) $P(X \geq 3) = ?$

1 이산확률분포

포아송분포 X~Poisson(m)

- 단위당 희귀현상의 평균 발생횟수를 m 일 경우
- 특정 단위 동안에 발생한 희귀현상의 횟수를 확률변수 X 로 정의함
- 확률변수 X 는 0, 1, 2, …의 값을 취함
  - X 는 모수 m 인 포아송분포(Poissondistribution)를 따름
- 분포함수

$$p(x) = P(X = x) = \frac{e^{-m} m^{x}}{x!}, \quad m > 0, \quad x = 0, 1, 2, \cdots$$

1 이산확률분포

POISSON(x, mean, cumulative)

x	발생 횟수
mean	단위시간 동안의 평균 발생 횟수
cumulative	함수 형태를 결정하는 논리값 - 1 또는 TRUE: 누적확률값 계산 - 0 또는 FALSE: 확률값 계산

1 이산확률분포

포아송분포 적용 예1

- 어떤 보험회사에서 하루에 접수되는 보험금 청구건수는 평균 2건이다.

  (1) 어느 날 보험금 청구가 한 건도 없을 확률은?

  (2) 3건 이상의 청구가 있을 확률은?

<풀이> 엑셀함수 POISSON을 이용하면 각각의 확률을 구할 수 있음

(1) $P(X = 0)$ ➔ "= POISSON(0, 2, 0)" 이용

(2) $P(X \geq 3) = 1 - P(X < 3) = 1 - P(X \leq 2)$

➔ "=1 - POISSON(2, 2, 1)" 이용

연습문제 1.

1. 공정한 동전을 세 번 던질 때 나타나는 앞면의 수를 확률변수 X라고 정의하자. 앞면이 1회 나올 확률을 구하고자 한다. 다음 설명 중 옳지 않은 것은?

① 확률변수 X 는 이항분포 $B(3,\ 0.5)$ 를 따른다.

② 동전을 세 번 던질 때 앞면이 1회 나올 확률은 $P(X=1)$ 로 표현된다.

③ 확률변수 X 는 0, 1, 2, 3의 값을 갖는다.

④ 공정한 동전을 세 번 던질 때 앞면이 1회만 나올 확률은 ¼이다.

연습문제 2.

2. 4개 보기 중 하나를 선택하는 선다형 문제가 20문항이 있고, 각 문항의 배점은 5점이다. 60점 이상이면 합격이다. 시험에서 랜덤하게 답을 써넣을 때 합격할 확률을 구하고자 한다. 엑셀 함수식은?

㉠ =BINOMDIST(12, 20, 0.25, 1)

㉡ =1-BINOMDIST(11, 20, 0.25, 1)

㉢ =BINOMDIST(12, 20, 0.25, 0)

④ =1-BINOMDIST(12, 20, 0.25, 1)

연습문제 3.

3. 어느 지역에서 1주일 동안 발생하는 교통사고 건수는 평균이 1.3인 포아송분포를 따른다. 특정 2주일 동안에 교통사고 건수가 3건 이상 발생할 확률을 구하고자 한다. 함수식으로 알맞은 것은?

① =Poisson(3, 1.3, 0)　　② =Poisson(3, 2.6, 0)

③ =1-Poisson(2, 1.3, 1)　　④ =1-Poisson(2, 2.6, 1)

7강

확률 분포 (3)

정보통계학과 이기재 교수

학습목표

- 연속형 확률분포의 특징에 대해서 설명
- 균등분포의 개념의 이해와 활용
- 정규분포를 이해하고 실제 문제에 적용
- 엑셀함수를 활용한 확률 계산

1 연속확률분포

1 연속확률분포

확률변수(Random variable)

- 확률적 실험에서 실험의 결과에 수치를 대응시켜 주는 것

☑ 이산형 확률변수(discrete random variable)

- 예 : 불량품의 개수, 동전을 5번 던질 때 나온 앞면의 수 등

☑ 연속형 확률변수(continuous random variable)

- 어떤 구간의 값을 취할 수 있는 확률변수
- 예 : 키, 몸무게, 강도, 성적 등

1 연속확률분포

연속형 확률변수의 예

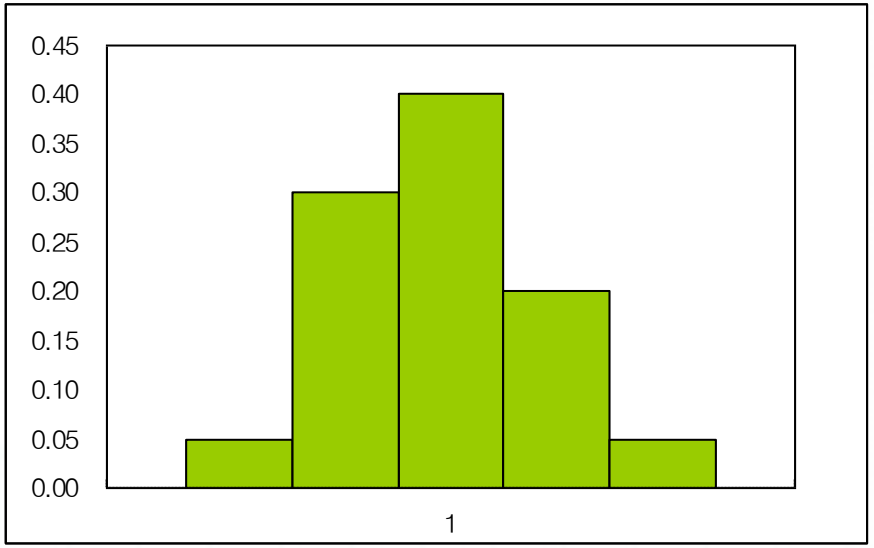
- X = 회사까지의 출근 소요 시간
- 집에서 회사까지 출근에 걸리는 시간을 100일 동안 수집

$(a \le X < b)$	도수	상대도수
$10 \le X < 20$ 분	5일	5/100
$20 \le X < 30$ 분	30일	30/100
$30 \le X < 40$ 분	40일	40/100
$40 \le X < 50$ 분	20일	20/100
$50 \le X < 60$ 분	5일	5/100

1 연속확률분포

연속형 확률변수의 예

출근 시간에 대한 히스토그램 1



- 출근에 걸리는 시간이 30분에서 50분 사이가 될 확률은?

$P(30 \le X < 50)$

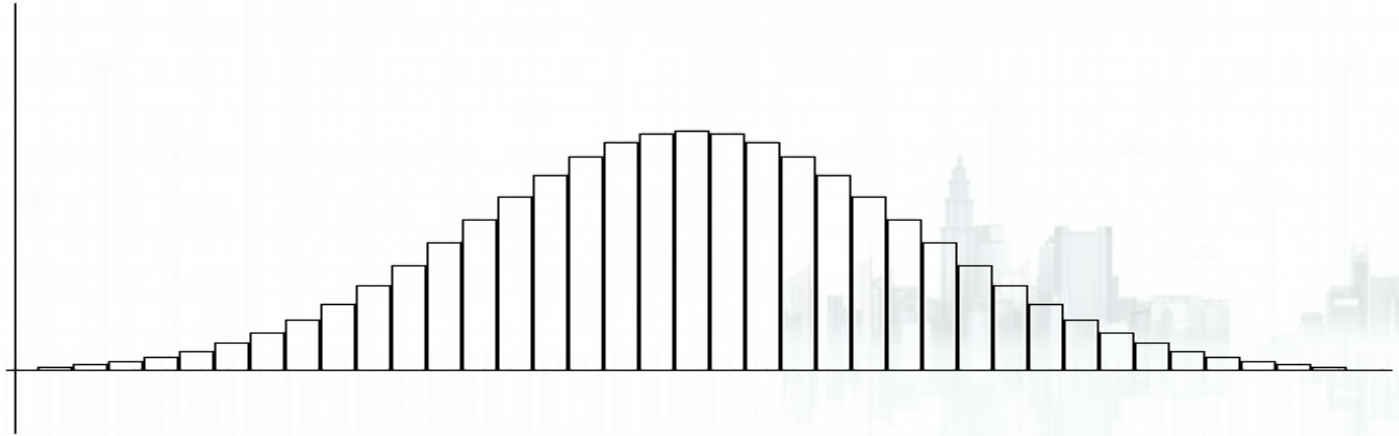
$= 40/100 + 20/100 = 0.6$

1 연속확률분포

연속형 확률변수의 예

출근 시간에 대한 히스토그램 2

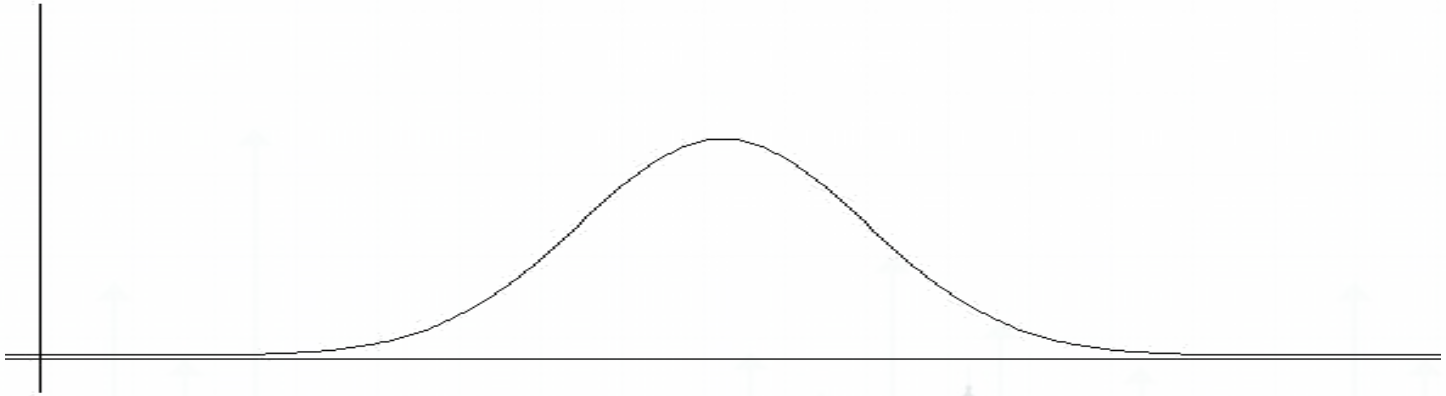
- X = 회사까지의 출근 소요 시간
- 1년 동안 집에서 회사까지 출근에 걸리는 시간 측정



1 연속확률분포

연속형 확률변수의 예

회사까지의 출근 소요 시간에 대한 확률분포함수



- 조사일수를 늘려서, 히스토그램의 구간 너비를 0에 가깝게 하면

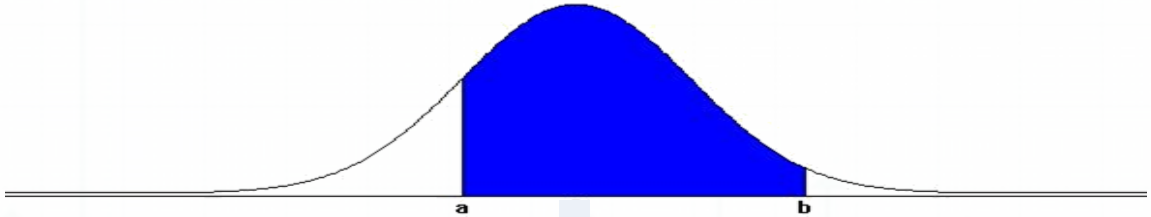
  앞선 히스토그램은 연속함수가 될 것임 ➔ **연속형 확률변수의 확률분포함수**

1 연속확률분포

확률밀도함수 f(x)의 성질

1. $f(x) \geq 0$

2. 확률변수 X에 대한 $P(a \leq X < b)$ 계산



➡ 수학적 표현 : $P(a \leq X < b) = \int_a^b f(x)dx$

3. $\int_{-\infty}^{+\infty} f(x) = 1$

1 연속확률분포

확률질량함수의 성질 : 이산형 확률 변수

1 $0 \le p(x) \le 1$

2 $\mathrm{P}(a < \mathrm{X} \le b) = \sum_{a<x\le b} p(x)$

3 $\sum_{\text{모든 } x} p(x) = 1$

1 연속확률분포

균등분포

- X ~ U(*a*, *b*)
- 분포함수 $$f(x)=\begin{cases}\dfrac{1}{b-a} & a\le x\le b\\ 0 & \text{기타}\end{cases}$$
- 구간 (a, b) 사이의 임의의 값들이 같은 가능성으로 발생함
- 엑셀에서 균등분포 난수(random number) 발생
  - 구간 (0, 1) 사이의 난수 발생 : '=RAND( )' 이용
  - 구간 (a, b) 사이의 난수 발생 : '=(b-a) * RAND( )+a' 이용

1 연속확률분포

균등분포 난수 발생 실습

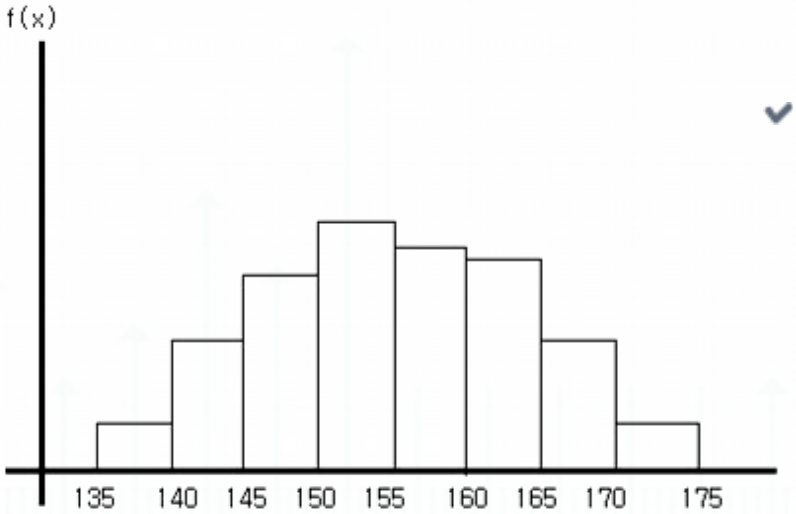
- 엑셀 함수 =RAND( )를 이용해서 1000개 난수 발생
- 생성 난수를 히스토그램으로 나타냄
- 생성된 난수가 0과 1사이의 균등분포를 따른다고 볼 수 있는가?

1 연속확률분포

연속형 분포 사례

학생의 키(1)

- 중학교 1학년 남학생 200명의 키 조사



- 주위에서 흔히 볼 수 있는 자료의 형태

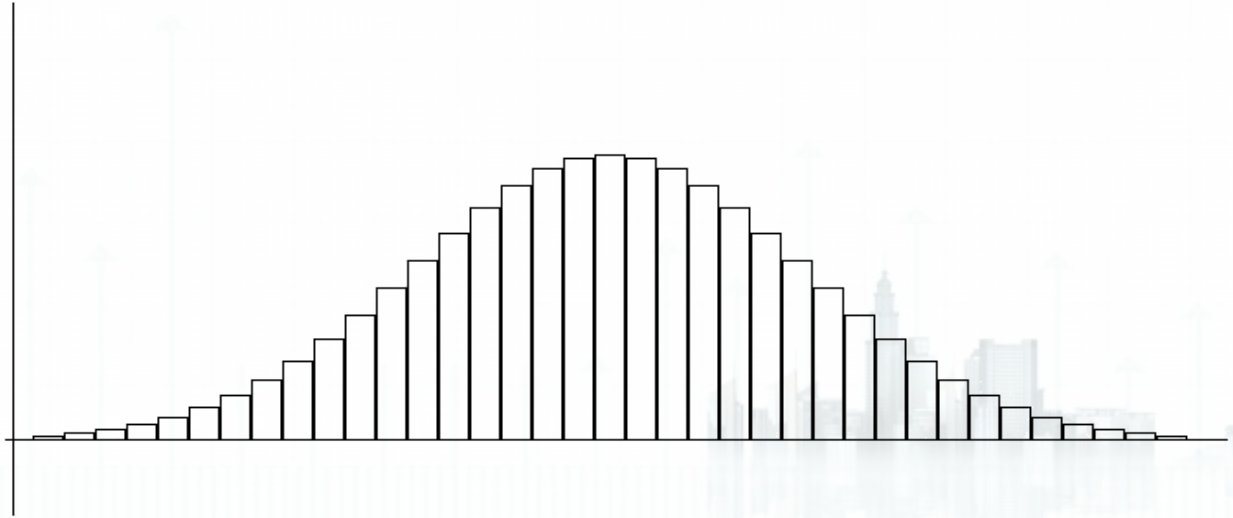
"평균 근처에 많이 모여 있고,
평균을 중심으로 좌우로 대칭인 형태"

1 연속확률분포

연속형 분포 사례

학생의 키(2)

- 중학교 1학년 남학생 3000명의 키 조사

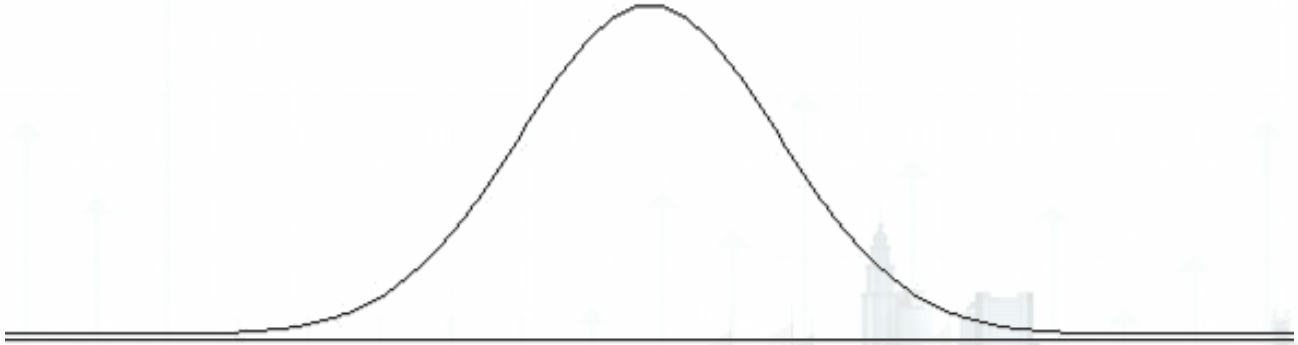


1 연속확률분포

연속형 분포 사례

학생의 키(3)

- 중학교 1학년 남학생 키에 대한 개념상의 분포



1 연속확률분포

정규분포(Normal Distribution)

- 평균을 중심으로 좌우 대칭이고, 종모양을 갖는 확률분포
- 수학자 드 므와브르(A. de Moivre, 1667 - 1754)
  : 정규분포 함수식 발견
- 가우스(C. F. Gauss, 1777 - 1855)
  : 물리학과 천문학 분야에 응용

"정규분포함수 또는 가우스 분포함수라고 함"

1 연속확률분포

정규분포의 특징

- "확률변수 X는 평균 μ, 표준편차 σ인 정규분포를 따른다."

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

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

  " μ(평균)와 σ(표준편차)의 값에 의해서 정규분포 모양이 결정됨"

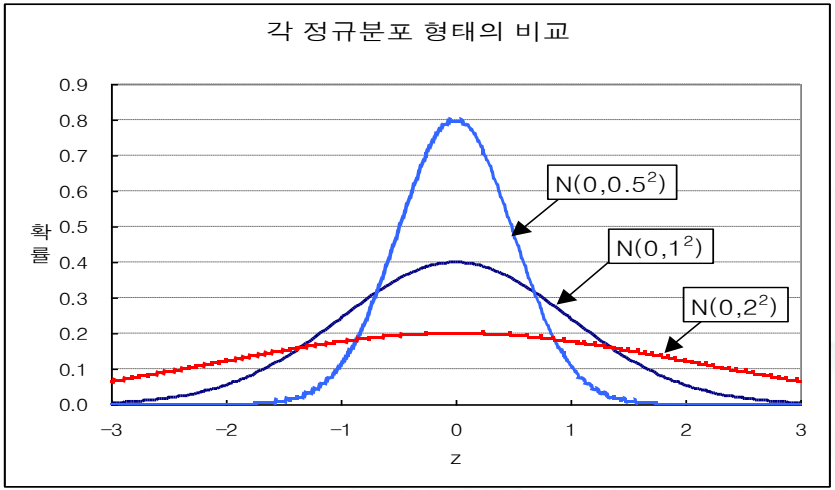
- 종모양이고, 평균 μ 에 관해 좌우 대칭

1 연속확률분포

정규분포의 모양

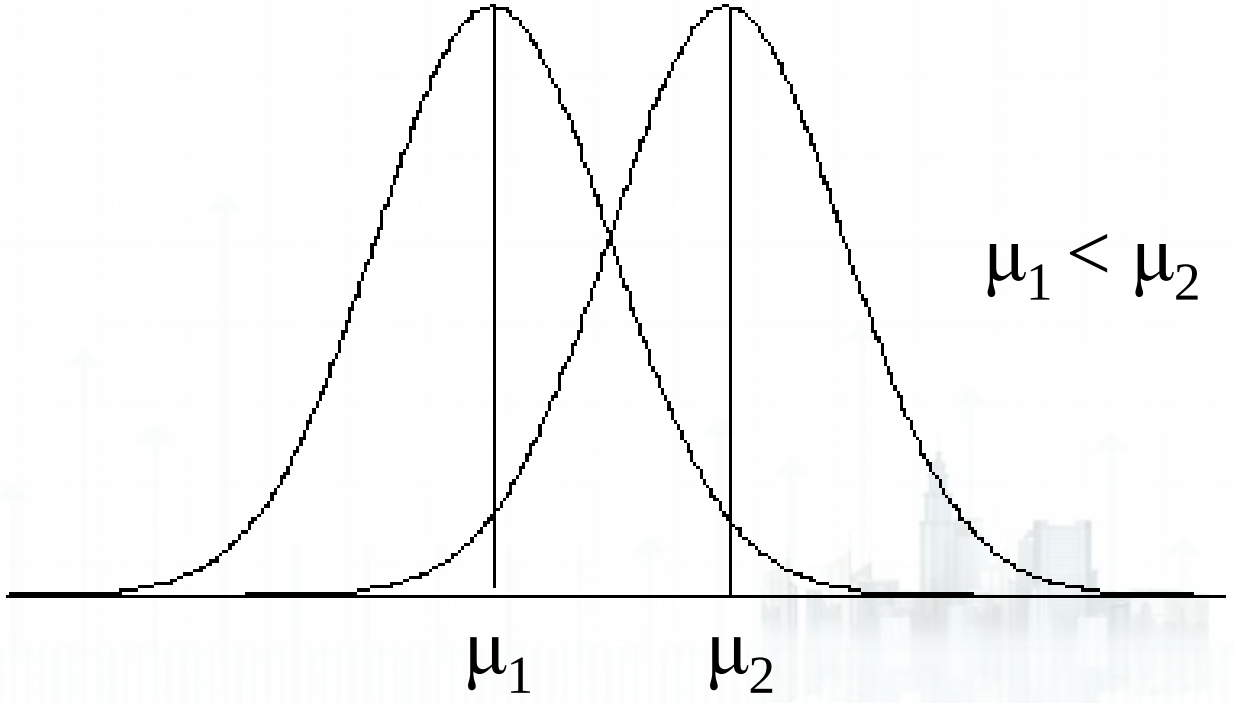
μ (평균)	분포의 중심위치
σ (표준편차)	평균을 중심으로 얼마나 넓게 퍼져 있는가

- 정규분포에서
  평균과 표준편차의 관계



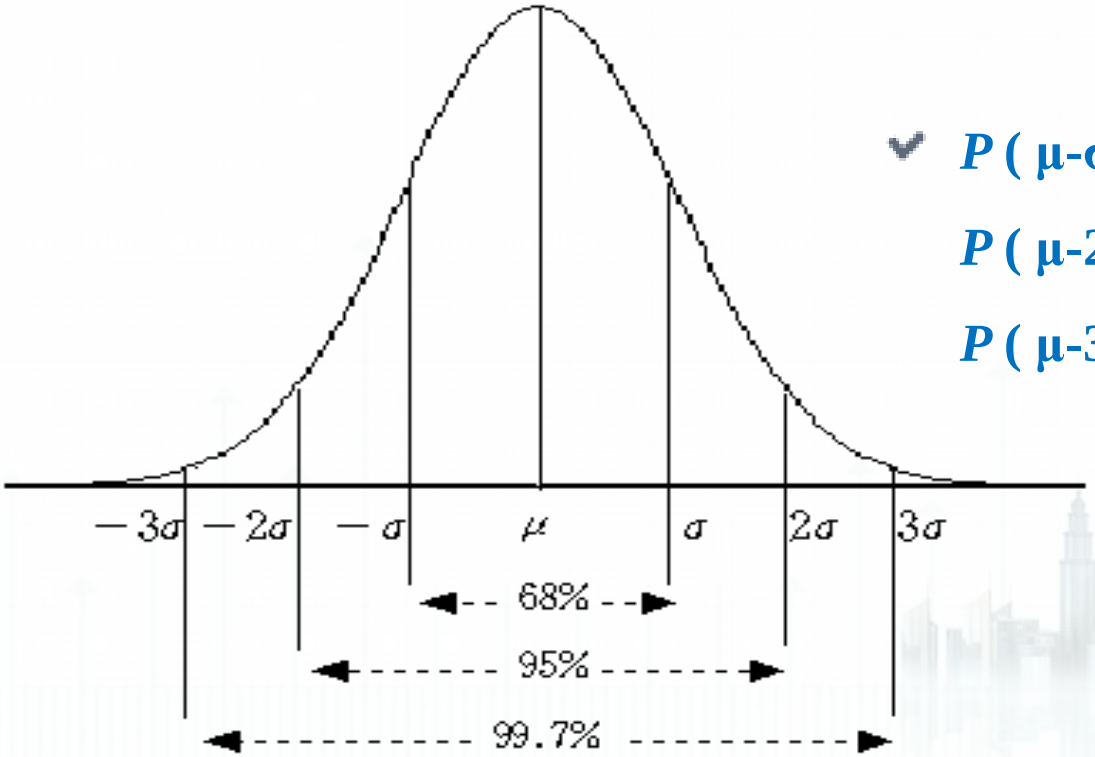
1 연속확률분포

정규분포의 모양 : 평균이 다른 경우



1 연속확률분포

정규분포의 성질 : 68 – 95- 99.7 법칙



- $P(\mu-\sigma < X < \mu+\sigma) = 0.683$
- $P(\mu-2\sigma < X < \mu+2\sigma) = 0.954$
- $P(\mu-3\sigma < X < \mu+3\sigma) = 0.997$

1 연속확률분포

정규분포의 표준화

- 확률변수 X가 정규분포 $N(\mu, \sigma^2)$를 따르는 경우

  $Z = \frac{X-\mu}{\sigma}$ 는 표준정규분포 $N(0, 1)$를 따름

- 표준화를 통한 확률 계산

$$P(X < x) = (\frac{X-\mu}{\sigma} < \frac{x-\mu}{\sigma}) = P(Z < \frac{x-\mu}{\sigma})$$

$$P(a < X < b) = P(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma})$$ 표준정규분포표 이용

1 연속확률분포

표준정규분포에서 확률계산

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

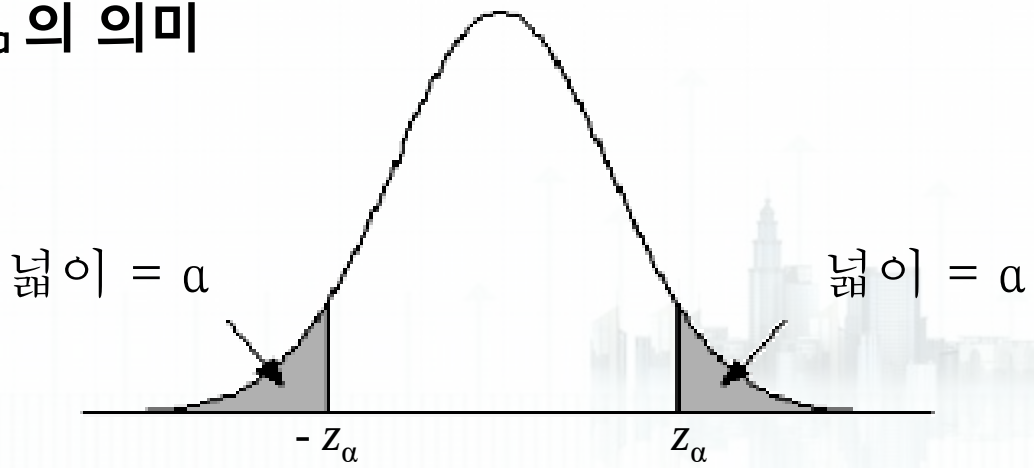
2. $P(Z \le -a) = P(Z \ge a)$

3. $P(Z \ge a) = 1 - P(Z < a)$

1 연속확률분포

표준정규분포에서 z_α의 정의

- $P(Z \geq z_\alpha) = \alpha$ 을 만족하는 값
  (즉, 위쪽 꼬리부분의 확률이 α가 되게 하는 값)
- 기호 z_α의 의미



1 연속확률분포

NORMDIST(x, mean, standard_dev, cumulative)

x	확률분포를 구하려는 값
mean	평균
standard_dev	표준편차
cumulative	함수 형태를 결정하는 논리값 1 또는 TRUE이면 누적확률값 계산, 0 또는 FALSE이면 확률값 계산

- "정규분포에서 $P(X \leq x)$의 계산 "

1 연속확률분포

정규분포의 사례1

- 집에서 회사까지 통근 시간 X(분)은 정규분포 $N(40, 5^2)$를 따름.

  통근 시간이 50분 이상 걸릴 확률은?

<풀이> X : 집에서 회사까지 통근 시간

$X \sim N(40, 5^2)$, $P(X \geq 50) = ?$

➡ $P(X \geq 50) = 1 - P(X < 50)$

➡ = 1 - NORMDIST(50, 40, 5, 1) 이용

1 연속확률분포

정규분포의 사례2

- 60만 명이 수학능력시험을 본 결과, $\mu=220$, $\sigma^2=30^2$인 정규분포를 따름.

  ① 250점을 받은 학생은 대략 몇 등쯤 될까?

<풀이> X = 수험생이 받은 점수

- $X \sim N(220, 30^2)$

➡ $P(X \geq 250) = 1 - P(X < 250)$

➡ 엑셀 함수 '= 1 - NORMDIST(250, 220, 30, 1)' 이용

1 연속확률분포

정규분포의 사례2

- 60만 명이 수학능력시험을 본 결과, $\mu=220$, $\sigma^2=30^2$인 정규분포를 따름.

② 상위 10%에 해당되는 학생은 대략 몇 점 정도인가?

<풀이> 상위 10%에 해당되는 학생은

아래에서부터 90%에 해당하는 학생

➡ 엑셀 함수'= NORMINV(0.9, 220, 30)'이용

1 연속확률분포

NORMINV(probability, mean, standard_dev)

probability	확률 값
mean	평균
standard_dev	표준편차

- "정규분포에서 왼쪽 부분의 확률이 주어진 경우에 이에 해당하는 *x*값 계산"

연습문제 1.

1. 다음은 정규분포에 대한 설명이다. 올바른 것을 모두 고른 것은?

Ⅰ. 정규분포의 모양은 평균과 표준편차에 의해서 결정된다.
Ⅱ. 정규분포는 표준편차에 대해서 좌우대칭이다.
Ⅲ. 정규분포를 따르는 확률변수의 평균과 중앙값은 같다.

① Ⅰ
② Ⅰ, Ⅱ
③ Ⅰ, Ⅲ
④ Ⅱ, Ⅲ

연습문제 2.

2. 어떤 사람이 자기 집에서 직장까지 차를 몰고 가는데 걸리는 시간 X(분)는 평균이 30분, 표준편차가 5분인 정규분포를 따른다고 한다. 이 사람이 집에서 아침 8시 20분에 출발하였을 때 아침 9시까지 직장에 도착할 확률을 구하고자 한다. 알맞은 엑셀 함수식은?

⊖ =NORMDIST(40, 30, 5, 0) ⊜ =1 - NORMDIST(40, 30, 5, 0)

⊛ =NORMDIST(40, 30, 5, 1) ④ =1 - NORMDIST(40, 30, 5, 1)

연습문제 3.

3. 어떤 자격시험의 성적분포가 근사적으로 평균이 70, 표준편차가 8인 정규분포를 따른다고 한다. 내년에도 비슷한 수준의 자격시험을 실시할 예정이며, 과거 성적분포에 따른 상위 5%에 해당하는 점수를 얻으면 포상금을 주려고 한다. 내년 시험에서 포상금을 받기 위해서는 몇 점을 받아야 하는지를 구하고자 한다. 알맞은 엑셀 함수식은?

① =NORMINV(0.95, 70, 8)

② =NORMINV(0.05, 70, 8)

③ =1-NORMDIST(0.95, 70, 8, 1)

④ =NORMDIST(0.05, 70, 8, 1)