

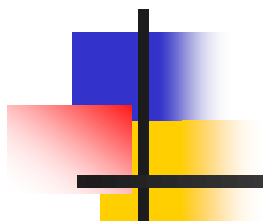
[illegible]

호서대학교 빅데이터경영공학부 연구필

# 탐색적 자료분석 (Exploratory Data Analysis)



1. 통계와 통계학
  - 1.1 통계학이란
  - 1.2 자료의 형태와 변수의 종류
2. 양적자료에 대한 EDA
  - 2.1 기술통계량
  - 2.2 R을 활용한 기초통계량 계산
  - 2.3 양적자료의 그래프 표현
3. 질적자료에 대한 EDA
  - 3.1 질적자료의 요약
  - 2.2 질적자료의 그래프 표현



---

# 1. 통계와 통계학

---



## 1.1 통계학이란

### ■ 통계(statistic)

- 사회적 현상 또는 자연현상을 규명하기 위해 수집된 각종 데이터를 요약하거나 적절한 방법을 통하여 일차적으로 가공되어 나오는 **정보**를 말함
- 불확실성을 규명하기 위해 얻어지는 **데이터**로부터 생성된 정보
- 일상생활에서 통계적인 수치로 표현되는 여러 가지 정보를 자주 접하게 되며 주로 **숫자, 그래프, 도표 또는 그림** 등 적절한 표현 방법으로 나타냄
- 동일한 현상도 데이터에 따라 다양한 통계를 얻을 수 있으며 이를 종합하거나 따로 따로 해석함으로써 **현상을 이해하고 예측하는데** 사용
- 국가는 국가정책을 세우고, 기업에서 경영계획, 생산계획 등을 세우고, 개인은 사회생활이나 경제생활을 하는 데 지표로써 활용함
- **왜곡이 된다면 엄청난 재앙**이 따를 수 있음
- 예) 물가지수, 경기종합지수, 산업생산지수, 경제성장률, 실업률, 인구증가율, 종합주가 지수, 국민총생산고(GNP), 가용외환보유액, 부채비율 등 (<http://www.nso.go.kr>)



## 1.1 통계학이란

### ■ 통계학(statistics)

- 자연과 사회적 집단 혹은 인간사회 등에서 나타나는 현상 등에서 보이는 **불확실성 (uncertainty)**을 규명하기 위해 다양한 데이터를 기초로 **수학 또는 확률적 수단**을 통해 학문적으로 분석하기 위한 **설계 · 조사 · 분석 · 처리 · 추론**에 대한 방법 또는 **의사결정에 도움을 주는 방법을 연구하는 학문**
- 적절한 데이터 또는 데이터의 가공을 통해 지식을 창출하는 방법을 다루는 학문
- 과학적인 이론에 근거하여 관심사에 대한 정확한 대상이 선정되어야 하며, 연구 목적에 필요한 자료와 정보가 경제성과 정밀도를 고려하여 최적의 방법으로 수집되고, 수집된 자료는 과학적인 이론에 의하여 정리 · 분석하여 최적의 의사결정을 제공하는 방법을 다루는 학문임
- 통계학적 지식은 의사결정시 과학적이고 합당한 이유와 근거를 제시하여 줌
- 통계적 도구들은 우리들의 일상생활에 영향을 끼치는 것에 대한 의사결정을 할 때 합리적이고 과학적인 도움을 줄 수 있음

# 1.1 통계학이란

## ■ 통계학이란?

- 통계학이란 불확실성을 연구하는 학문임

### 통계학의 어원

- 고대 로마시대에 국가(State)의 상태(State)를 살피는데 관심을 가졌는데 이것을 'Statistics'라 불렀음



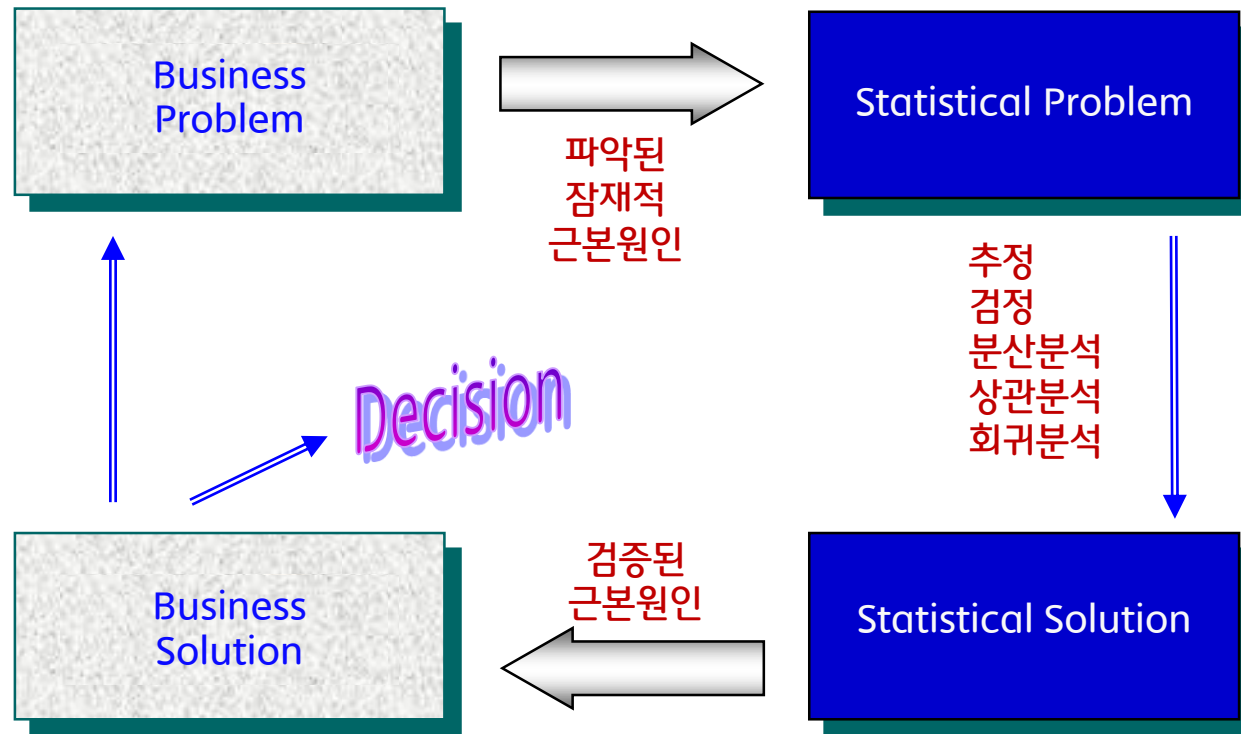
### 통계학이란?

- 연구 목적에 필요한 자료 및 정보를 최적의 방법으로 수집하고, 수집한 자료를 과학적이고 논리적인 이론에 의해 정리, 분석하는 학문
- 불확실한 상황에서 현명한 의사결정을 하기 위한 이론과 방법의 체계
- 일부로서 전체를 파악하기 위한 이론과 방법의 체계

# 1.1 통계학이란

## ■ 통계적 문제 해결

- 현실에서 발생하는 의사결정 문제를 통계적 방법을 통해 실제 문제에 대한 근본원인을 파악 및 해결할 수 있음





## 1.1 통계학이란

### ■ 통계학의 필요 분야

- 인문 · 사회 · 자연과학, 경험과학 등 모든 분야에 이용되고 있음
  - 경험과학: 자연과 사회에서 일어나는 현상을 체계적으로 설명하기 위하여 새로운 이론을 수립하거나, 기존의 이론을 지지하거나 거부하기 위하여 자료를 수집하여 분석하는 과학
  - 주관적인 판단에 의한 이론이나 의견의 타당성을 주장하기 보다는 객관적 자료에 입각하여 이론이나 의견을 주장하는 것이 보다 합리적이고 바람직함

=> 자료분석에 의한 통계가 필수적인 수단

예) 멀티미디어를 사용한 교수법이 전통적 교수법보다 학업성취도 향상에 효과적인지의 여부를 검증

- 경제통계, 경영통계, 사회통계, 교육통계, 의학통계, 보건통계, 행정통계, 공업통계, 농업통계, 인구통계, 스포츠통계 등 모든 분야에 필요함
- 최근에는 많은 기업에서 자사가 보유하고 있는 데이터베이스(Database)를 이용하여 고객관계관리(CRM: customer relationship management)를 활용하는 사례(통신사, 카드사, 보험사, 증권사, 백화점, 인터넷 쇼핑몰, 각종 전자상거래 사이트 등)가 큰 이슈로 부각





## 1.1 통계학이란

### ■ 통계 적용 사례

- **소아마비용 쇼크 백신** : 1954년 백신실험은 엄격한 통제하에 약 40만 명의 어린이들에게 실시. 그 결과에 대한 훌륭한 통계적 분석은 백신의 효능을 확실하게 믿게 함
- **챌린저호 폭발** : 1986년 우주왕복선 챌린저호가 폭발하여, 7명의 우주비행사들 사망. 이것은 우주선이 낮은 온도에서도 그 기능을 제대로 수행하는지에 대한 자료분석도 없이 우주선 발사를 결정한 결과임.
- **종이 헬리콥터의 설계** : 미국의 저명한 통계학자인 박스교수가 제시한 종이 헬리콥터의 설계문제에 있어 설계 파라미터인 날개길이, 몸체길이, 몸체 폭을 얼마로 설계하여야 종이 헬리콥터가 가장 오래 날 것인가를 알아보는 문제, 날개길이와 몸체길이가 체공시간에 가장 큰 영향을 준다는 것을 통계분석의 결과 발견하게 됨. 체공시간에 맞는 헬리콥터를 설계가능
- **품질향상** : 많은 기업들은 초우량 기업의 달성 기준을 산업표준으로서 6-시그마(6-sigma) 수준을 설정하고 있음. 이는 품질에 있어 99.9999998%가 결함이 없는 프로세스이다. 통신산업의 예를 들면 10년 동안 2.6분 서비스가 중단되는 것과 동일할 정도로 완전무결한 수준을 유지 가능



## 1.1 통계학이란

### ■ 통계 적용 사례

- **고객만족도 조사** : 고객만족도 조사를 통해 현재 소유하고 있는 TV에 대한 만족도, 광고인지도, TV 선택 시 중요 요인, 라이프 스타일, 인구학적 특성 등을 조사하여 제품개발, 브랜드 로열티 분석, 향후 포트폴리오 전략수립, GAP분석, 포지셔닝(positioning) 분석 등에 활용한다
- **사회여론조사** : 대통령 선거나 국회의원 선거 등과 같은 중요 정치상황에 대해 선거를 하기 전에 유권자들의 여론 방향을 조사하고, 또한 실제선거에 있어서의 당선자를 예측함.
- **인간 게놈 프로젝트** : 대용량의 데이터 속에서 숨겨진 패턴과 지식을 찾아내는 데이터마이닝(Data Mining)과 빅데이터(Big Data) 통계기술이 인간 게놈의 비밀을 푸는데 결정적으로 기여하였다. 과학자들은 30억 개의 인간 염기 서열 중 3% 정도만이 유용한 것으로 보고 있는데, 30억 개 중 어떤 것이 3%에 해당되는지, 또 그 속에 담겨져 있는 유전자 정보를 찾아내 질병 발생 원인을 밝혀내는데 데이터마이닝 기술이 절대적 기여를 함



## 1.1 통계학이란

---

### ■ 기업에서 통계를 통한 의사결정 사례

- 보험회사들은 적절한 통계적 방법론을 이용하여 **자동차 혹은 생명보험료의 적정 수가를** 결정한다.
- 수자원공사는 주기적으로 **한강의 오염도를 측정하여** 정화작업의 필요성 여부에 이용한다.
- 의료업계에서 근무하는 연구자들은 새롭게 개발된 여러 가지 약물들에 대한 **임상실험을 통하여 특정 질병의 생존율에** 관하여 연구한다.
- 통신회사는 **이용금액이 높고 이탈가능성이 높은 고객에게** 기기변경에 대한 보조금을 차별 지급하며, **가입기간, 이용금액, 연체횟수에 따라** 고객을 점수화하여 멤버십등급을 구별한다.



## 1.1 통계학이란

---

### ■ 마케팅 분야에서의 통계학의 활용

- **모델링(modeling):** 이탈모델, 예측모델, 반응모델, Cross-sell 모델 등에 의한 타겟팅 제공하여 비용 감소 및 수익 증대
- **세분화(segmentation):** 다양한 개별 고객을 활용 목적에 따라 유사한 몇몇의 집단으로 그룹화하여 회원의 속성을 규명하는 마케팅 체계 구축 지원
- **분석(analysis):** 고객, 상품, 가격, 영업, 경쟁사, 시장변화, 채널 등 다양한 마케팅 이슈에 대한 분석을 통한 마케팅 Insight 발견
- **성과 측정 :**
  - 마케팅 성과측정 방법 및 기준을 수립하고, 마케팅 효율 및 성과 측정
  - 고객, 상품, Segment에 대한 수익, 비용을 통해 정확한 수익성 평가



## 1.1 통계학이란

---

### ■ 신용(Risk)관리 분야에서의 통계학의 활용

- **신용 평가/심사:** 고객 심사(underwriting), 연체 예측과 연체 상태 파악, 통계적/계량적 대손 추정
- **신용평점 산출(scoring):** 고객 정보와 대외정보, 연체정보를 고려한 모델링을 통한 신용평점 개발 및 신용평점 시스템 관리
- **신용등급 세분화(segmentation):** 대외정보, 연체정보, 신용평점 등 리스크에 기반한 고객 세분화와 고객 리스크 관리
- **사기 방지(Fraud Detect System):** 제 3자 사기, 명의도용, 사고매출, 보험사기 등을 방지하기 위한 모델링 및 사기 방지 시스템 개발



## 1.1 통계학이란

---

### ■ 통계 소프트웨어

SAS, SPSS, Minitab, R, Excel, Statistica

- **SAS** : 대용량 데이터 사용이 가능하며 프로그램 중심으로 통계전문가들이 선호하며 기업에서 주로 많이 활용됨
- **SPSS** : 쉬운 메뉴방식으로 여론조사, 사회과학 분야에서 많이 사용됨
- **Minitab** : 프로그램의 부피가 작아 사용하기 편함. 공업통계, 품질관리와 6-시그마 분야에 특화되어 있음
- **R** : 프리 소프트웨어이며 프로그램 방식으로 대학에서 많이 사용되며 최근 빅데이터 분야에서 선호됨
- **Excel** : 쉽고, 기초 통계 자료분석에 적합함

# 1.1 통계학이란

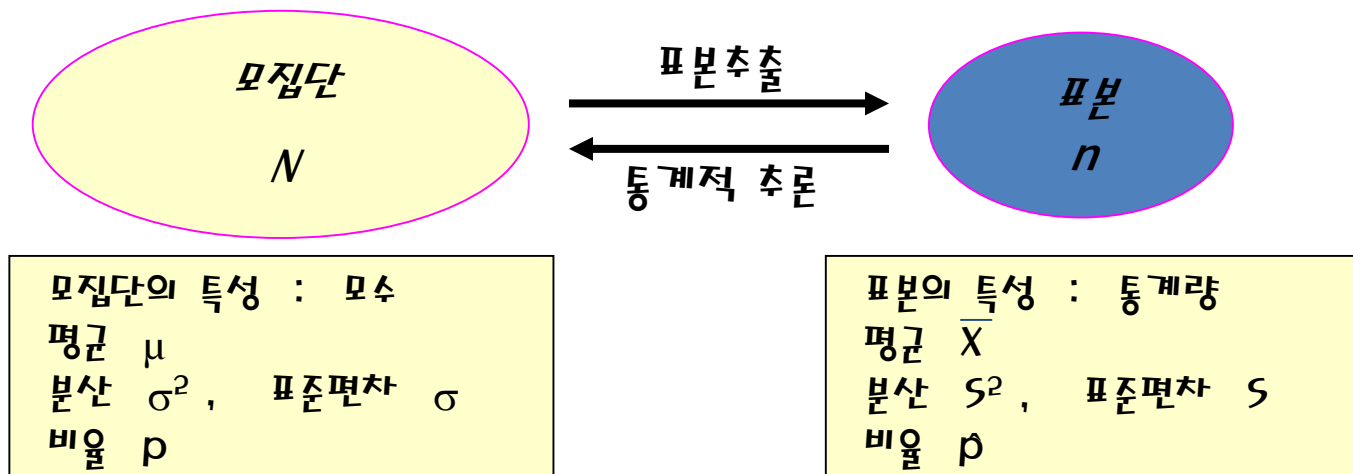
## ■ 통계학의 두 방향

- **기술통계학 (Descriptive Statistics)**

- 방대한 자료를 그래프나 표 또는 몇 개의 숫자로 요약하여, 주어진 자료의 전반적인 내용을 쉽고 빠르게 파악할 수 있는 기법을 다루는 분야

- **추측통계학 (Inferential Statistics)**

- 관심의 대상이 되는 전체집단(모집단)으로부터 모집단의 일부(표본)를 추출하고, 표본으로부터 관측된 내용 (통계량)을 근거로 하여 모집단의 특성(모수)을 추측하고 검정하는 방법을 다루는 통계학
- 통계적 모형을 설정하고, 설정된 모형이 합리적인지 여부를 평가하며, 주어진 자료로부터 얻어지는 정보를 근거로 미지의 특성에 대한 결론을 내리고 미래를 예측하는 분야





# 1.1 통계학이란

## ■ 모집단과 표본

- **모집단(population):** 연구대상이 되는 모든 개체의 관측값이나 측정값의 집합 – 반드시 실존하는 개체의 집합이 아님
- **표본(sample):** 통계적 처리를 위하여 추출한 모집단의 일부 또는 부분집합
- **추출단위(sampling unit):** 모집단의 가능한 관측값이나 측정값이 얻어지는 개체
- **모수 (Parameter)**
  - 모집단의 특성치, 고정된 미지의 상수
- **통계량 (Statistic)**
  - 표본의 특성치
  - 통계량의 값은 데이터로부터 얻어지고 표본마다 다를 수 있다.
  - 표본평균, 표본비율, 표본분산 등

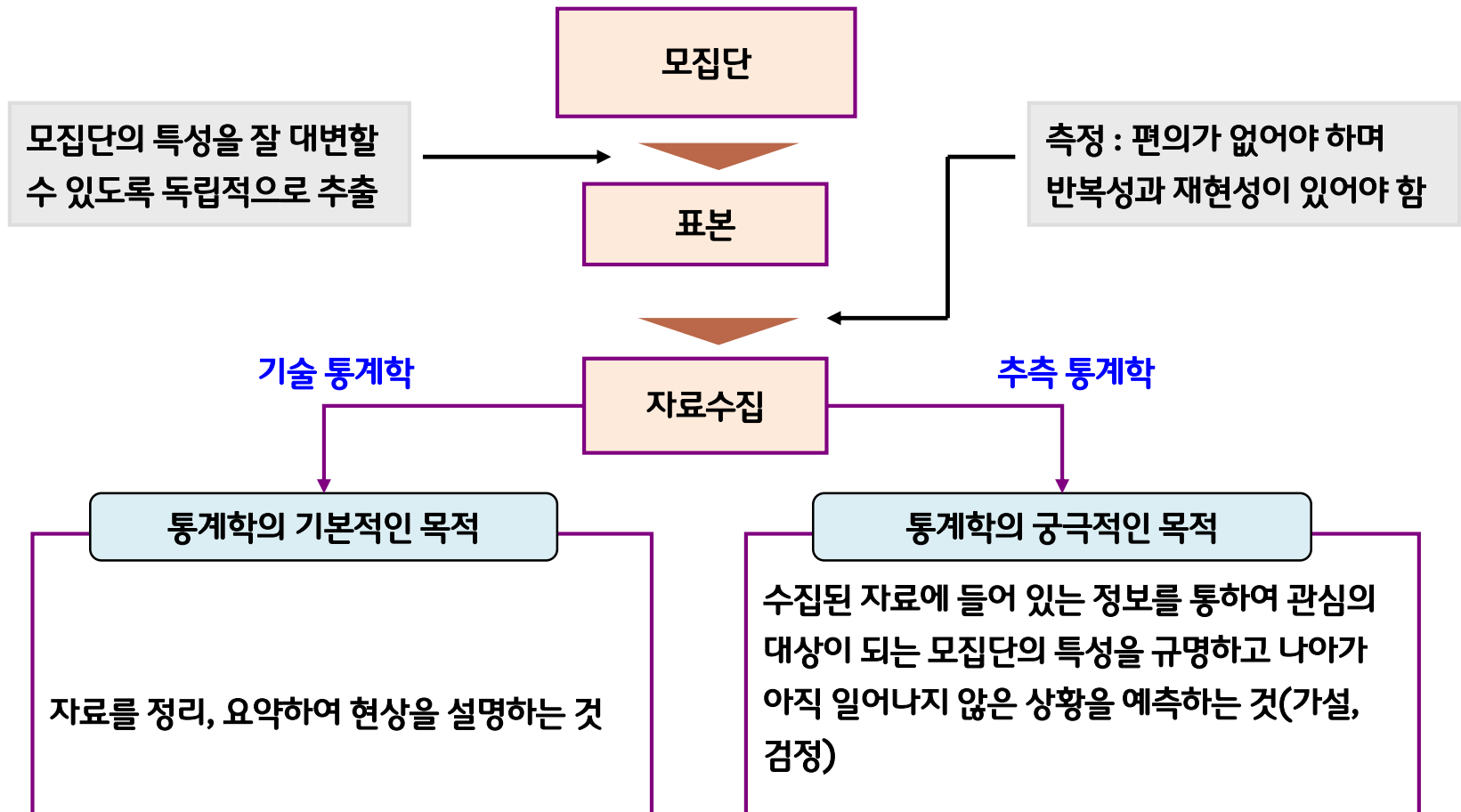
예) 한국의 성인전체(모집단)의 흡연자비율(모수)를 알기 위해서 성인남자 1000명(표본)을 추출하여 이 중 흡연자비율(통계량)이 얼마나 되는지를 조사하였다.



# 1.1 통계학이란

## ■ 모집단과 표본

### ■ 모집단과 표본



## 1.2 자료의 형태와 변수의 종류

### ■ 자료 (데이터)

- 데이터: 어떤 정황(Context)에 대한 정보를 포함한 숫자나 형태
- 정황은 우리가 어떠한 상황에 대한 배경지식을 이용하여 의사결정 또는 판단을 할 수 있도록 도와주는 역할을 함





## 1.2 자료의 형태와 변수의 종류

---

### ■ 자료의 설명

- 자료집합(데이터세트, data set, table): 자료를 모아 놓은 표
- 변수(variable): 개체의 속성(attribute)을 나타내며 열(column)로 표시함
- 변수값: 해당된 변수에 속한 값들
- 자료값(data value): 각 개인의 임의 변수에 해당되는 측정값
- 관측값(observation): 관측대상이 되는 개인에 관한 자료값들의 모임
- 다변량자료(multivariate): 각 개체 별로 알고자 하는 속성이 한 개 이상의 변수로 구성되어 있는 경우의 자료
- 일변량자료(univariate data): 한 변수만의 속성을 다루는 경우의 자료

## 1.2 자료의 형태와 변수의 종류

### ■ 자료의 설명

통계학 과목 수강학생에 대한 자료

변수 필드 속성

번호	나이	성별	학년	키(cm)	몸무게(kg)
1	28	0	3	183	82
2	18	1	1	168	52
3	46	1	2	165	52
4	18	1	1	158	55
5	55	0	5	180	93
:	:	:	:	:	:
:	:	:	:	:	:
33	19	1	2	170	54
34	19	0	2	168	79
35	19	0	2	183	70
36	20	1	2	145	37

관측값 개체 레코드

자료값

성별의 0은 남자, 1은 여자  
학년의 5는 대학원생을 의미함



## 1.2 자료의 형태와 변수의 종류

### ■ 자료의 형태

#### ■ 질적 (qualitative) 자료

- 범주형자료라고도 하며, 명목형 자료와 순서형 자료로 구분됨.

##### • 명목형(nominal) 자료

- 자료 값의 크기나 순서가 없고 단지 자료 값 자체의 이름만 부여할 수 있는 자료  
예) 성별, 지역, 직업, 혈액형, 종교, 운동선수의 번호, 인종 등이 있음

##### • 순서형(ordinal) 자료

- 기준에 의해 자료 값 들의 순서에 의미를 부여할 수 있는 자료, 비율이나 차이는 없음  
예) 달리기 1위, 2위, 3위, 연비에 대한 등급, 성적에서의 수, 우, 미, 양, 가

■ 질적자료는 계수형 자료와 함께 이산적인 값을 갖기 때문에 이들을 한데 묶어 **이산형 (discrete) 자료**라고도 함



## 1.2 자료의 형태와 변수의 종류

### ■ 자료의 형태

#### ■ 양적 (quantitative) 자료

- 측정자료(measurement data) 또는 연속자료(continuous data)라고 함
- 계수형자료와 연속형자료의 구분
  - 계수형(count) 자료: 값이 셀 수 있는 정수 형태인 자료  
예) 형제의 수, 자동차 보유대수, 입사 지원자수, 보험 해약건수
  - 연속형(continuous) 자료: 자료의 측정이 셀 수 없는 소수점을 포함하는 자료  
예) 키, 무게, 길이 등
- 비율형자료와 등간형자료의 구분
  - 비율형(ratio) 자료: 값들 사이의 차이 또는 비율에도 의미를 부여 있는 자료  
예) 무게 (0은 절대 영점임)
  - 등간형(interval) 자료: 값들 사이의 차이에는 의미를 부여할 수 있지만 비율에는 의미를 부여할 수 없는 자료.  
예) 온도(0은 상대 영점임)

## 1.2 자료의 형태와 변수의 종류

### ■ 자료의 형태

- 범주형 변수 (categorical variable)
  - 자료의 형태가 명목형, 순서형, 계수형 자료를 다루는 변수
    - **명목형 자료** : 성별(남, 여), 지역(서울, 부산, 광주 ... ), 혈액형
    - **순서형 자료** : 성적(A, B, C, D, E), 순위(1등, 2등, 3등)
    - **계수형(count) 자료** : 불량품 수, 결석 인원 수, 방문 수, 보험 해약건수
- 연속형 변수 (continuous variable)
  - 비율형자료, 등간형 자료를 다루는 연속적인 값을 갖는 변수
    - **구간척도** : 온도
    - **비율척도** : 키, 몸무게, 길이, 소요시간
- 질적 변수(quantitative variable)와 양적 변수(qualitative variable)
  - 질적 변수: 명목형 변수, 순서형 변수 => 사칙연산이 가능하지 않음
  - 양적 변수: 계수형 변수, 연속형 변수 => 사칙연산이 가능

※ **주의**: 자료의 종류에 따라 요약과 통계적 분석의 방법이 다르다.



## 1.2 자료의 형태와 변수의 종류

---

### ■ 변수의 유형

- 종속변수(dependent variable) 또는 반응변수(response variable)
  - 가정된 원인에 의해 영향을 받아 변화되는 변수
- 독립변수(independent variable) 또는 설명변수(exploratory variable)
  - 다른 변수를 변화시키는 원인이 되는 변수
- 반응변수와 설명변수가 어떠한 자료의 형태를 따르느냐에 따라 사용 가능한 통계분석 방법들이 다르게 선택됨

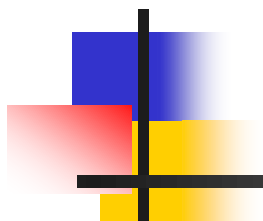


## 1.2 자료의 형태와 변수의 종류

### ■ 자료의 형태에 따른 분석 기법 분류

- 자료의 종류에 따라 요약과 통계적 분석의 방법이 매우 다르게 적용된다.

분석의 목적	반응변수의 유형	설명변수의 유형	분석 기법
연관관계 분석 (반응변수 없음)		연속형 (구간형, 비율형)	단순상관분석 (변수가 2개) 주성분분석, 인자분석, 정준상관분석 (변수가 여러 개)
		이산형 (명목형, 순서형)	Fisher's 정확검정, 카이제곱 검정, 대응분석 (변수가 2개) 감마, 람다 등 연관성의 측도 다중대응분석, 로그선형모형 (변수가 여러 개)
인과관계 분석 (반응변수 있음)	연속형 (구간형, 비율형)	연속형 (이산형 포함)	회귀분석 (선형회귀, 비선형회귀, Regression Tree 등) 시계열자료분석
		이산형 (연속형 포함)	t-검정 (이산형 설명변수 1개, 수준 2) 일원 분산분석 (이산형 설명변수 1개, 수준 여러 개) 다원 분산분석 (설명변수 여러 개) 공분산분석 (이산형과 연속형 설명변수의 혼합)
	이산형 (명목형, 순서형)	연속형 (이산형 포함)	판별분석 (선형판별, 2차판별, Classification Tree 등) 이항, 순서, 다항 로지스틱(프로빗)
		이산형 (연속형 포함)	로그선형모형
기타			경로분석 (반응변수 여러 개, 반응변수들 간의 인과관계 존재) 구조방정식모형 (경로분석과 인자분석의 혼합 형태) 군집분석 (개체들 간의 군집화), 다차원척도법(MDS) 비모수적 분석 (정규성, 등분산성 등의 가정이 적절하지 않은 경우)



---

## 2. 양적 자료에 대한 EDA

---



## 2.1 기술통계량 (Descriptive Statistics)

---

### ■ 대표값

- 위치를 나타내는 통계량
  - 자료들을 수치로 보았을 때 어느 위치에 있는지를 나타냄(위치측도)
  - 자료들이 대략 어떠한 값을 갖는 지를 알아보기 위하여, 어느 위치를 중심으로 자료들이 모여 있는 지를 나타내는 척도
  - 평균(mean), 중앙값(median),
  - 최빈값(mode),
  - 가중평균(weighted mean), 절사평균(trimmed mean),
  - 기하평균(geometric mean), 조화평균(harmonic mean),
  - 백분위수(percentile) 등이 있음
  - 자료에 특이하게 작거나 큰 값이 들어있게 되면 평균값은 대표값으로 부적절한 경우가 생김
  - 최대값, 최소값

## 2.1 기술통계량 (Descriptive Statistics)

### ■ 대표값

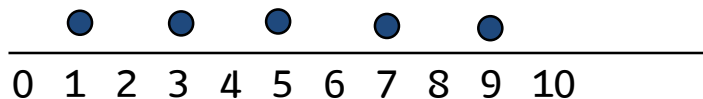
- 자료의 평균값과 중앙값

(1) 자료의 평균  $\bar{x} = \sum_{i=1}^n x_i / n$

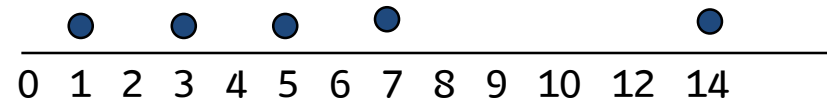
(2) 자료의 중앙값은 크기에 따라 늘어 놓을 때 가운데에 놓이는 값을 말함

중앙값의 위치는  $(n+1)/2$ 에 의해 결정되며 자료의 개수가 짝수일 때에는 2개의 중앙의 값의 평균이 중앙값이 된다.

(3) 평균은 이상점에 크게 영향을 받지만 중앙값은 별로 영향을 받지 않아 분포상태가 극도로 비대칭인 경우에는 중앙값이 평균보다 더 큰 의미를 가진다.



평균 = 5, 중앙값=5



평균 = 6, 중앙값=5

## 2.1 기술통계량 (Descriptive Statistics)

### ■ 대표값

- 가중평균(weighted mean)

$$\bar{x}_w = \frac{\sum_{i=1}^k w_i x_i}{\sum_{i=1}^k w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_k x_k}{w_1 + w_2 + \cdots + w_k}$$

- 최빈값(mode): 자료 중 그 빈도수가 최대인 값
- 범위의 중앙값(midrange): 최대값과 최소값의 평균
- 기하평균: 자료가 양수인 경우 증감율의 평균을 구할 때 사용

$$(x_1 x_2 \cdots x_n)^{1/n}$$

- 조화평균

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



## 2.1 기술통계량 (Descriptive Statistics)

---

### ■ 산포도 (Measure of Dispersion)

- 통계자료가 얼마나 서로 다른 값을 가지는가를 나타냄
- 자료들이 얼마나 변동하거나 퍼져있는 지를 표시함
- 변동성 척도(measure of variability), 퍼짐척도(measure of spread)
- 분산(Variance), 표준편차(Standard deviation),
- 변동계수(Coefficient of variation),
- 범위(Range),
- 4분위 범위(Interquartile range : I.Q.R.) 등이 있음

## 2.1 기술통계량 (Descriptive Statistics)

### ■ 산포도 (Measure of Dispersion)

- 자료의 분산과 표준편차

#### (1) 자료의 분산

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

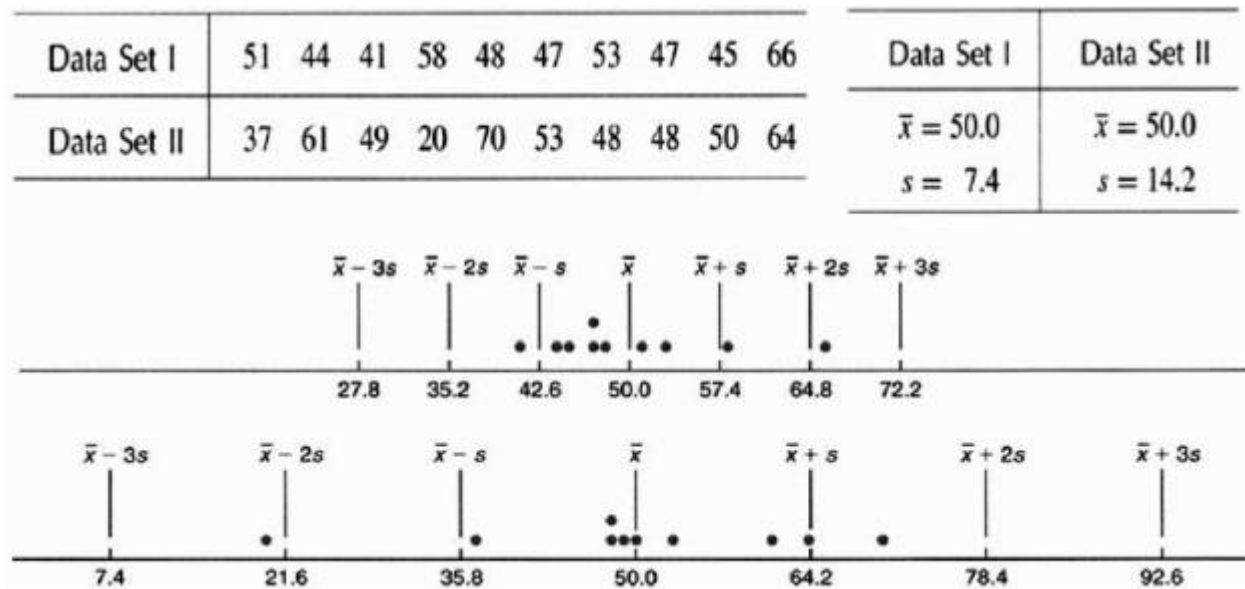
#### (2) 자료의 표준편차

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

## 2.1 기술통계량 (Descriptive Statistics)

### ■ 산포도 (Measure of Dispersion)

- 자료의 분산과 표준편차



<그림> 평균은 같고 분산이 다른 두 표본



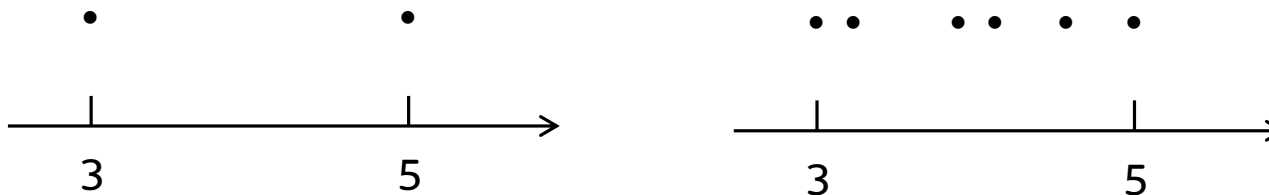
## 2.1 기술통계량 (Descriptive Statistics)

### ■ 산포도 (Measure of Dispersion)

- 자료의 범위(range):

범위 = 최대값 - 최소값

- 쉽고 빠르게 구할 수 있음
  - 특이하게 크거나 작은 값이 있을 경우 자료의 범위가 왜곡됨
  - 자료의 개수와 상관없이 같게 나올 수 있음
- ⇒ 자료의 변동성을 대표하지 못하는 경우가 많음



<그림> 같은 범위를 갖는 두 표본

## 2.1 기술통계량 (Descriptive Statistics)

### ■ 대표값

- 사분위수와 백분위수
  - 제 1 사분위  $Q_1$  = 제 25 백분위수
  - 제 2 사분위  $Q_2$  = 제 50 백분위수 (=중위수)
  - 제 3 사분위  $Q_3$  = 제 75 백분위수
  - 제  $P$  백분위수라 함은 자료를 크기 순서로 늘어 놓았을 때 적어도  $P$  %의 관측값이 그 값보다 작거나 같고, 또한 적어도  $(100 - P)$  %의 관측값이 그 값보다 크거나 같게 되는 값을 말한다.
- 다섯숫자 요약 (5-number summary )
  - minimum,  $Q_1$ ,  $Q_2$ ,  $Q_3$ , maximum
- 변동계수 (coefficient of variation)
  - 자료의 측정단위에 의존하지 않는 상대적인 산포의 측도
  - 서로 측정단위가 다른 여러 개의 자료의 산포를 비교할 때 사용될 수 있음

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

## 2.1 기술통계량 (Descriptive Statistics)

### ■ 산포도 (Measure of Dispersion)

- 자료의 사분위범위(Interquartile range : I.Q.R.)

- 사분위범위(IQR) =  $Q_3 - Q_1$
- 사분위편차(quartile deviation) =  $(Q_3 - Q_1) / 2$
- 양쪽 극단 값에서 자료의 25%씩 안쪽으로 들어와 있는 값의 거리  
⇒ 특이값의 영향을 거의 받지 않음

예) ( 순서대로 나열 ): 11 12 13 16 16 17 17 18 21

$$\text{사분위범위} = Q_3 - Q_1 = 17 - 13 = 4$$

- 절대평균편차 (mean absolute deviation, MAD)

$$\text{MAD} = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

- 범위가 가지는 단점을 보완한 척도
- 산술평균으로부터의 평균적인 거리
- 관찰값들의 측정단위와 같은 측정단위를 가짐

## 2.1 기술통계량 (Descriptive Statistics)

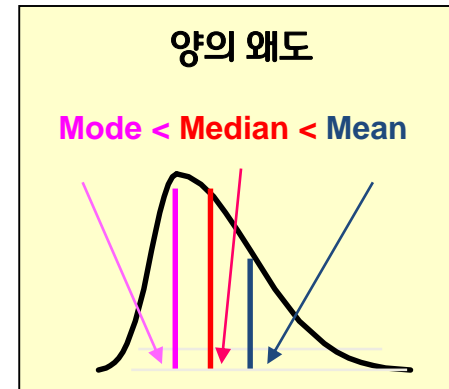
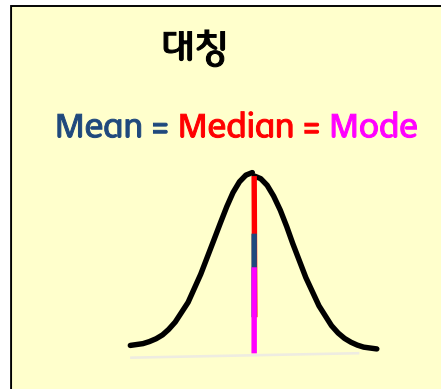
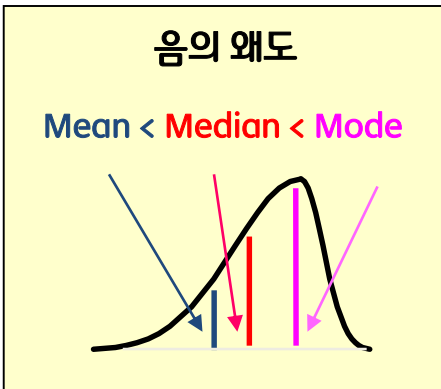
### ■ 왜도와 첨도

#### ▪ 표본 왜도 (Skewness)

- 자료의 분포에 대한 비대칭의 정도를 나타냄

$$\hat{\mu}_3 = \frac{m_3}{(m_2)^{3/2}} = \frac{\sum_{i=1}^n (x_i - \bar{x})^3 / n}{\left( \sum_{i=1}^n (x_i - \bar{x})^2 / n \right)^{3/2}}$$

- 왜도가 0이면 자료분포의 형태가 좌우 대칭이라는 것을 의미
- 왜도가 0보다 크면 자료가 왼쪽으로 치우쳐진(오른쪽으로 긴 꼬리) 분포
- 왜도가 0보다 작으면 자료가 오른쪽으로 치우쳐진(왼쪽으로 긴 꼬리) 분포



## 2.1 기술통계량 (Descriptive Statistics)

### ■ 왜도와 첨도

#### ▪ 첨도 (Kurtosis)

- 자료의 분포의 뾰족한 정도를 나타냄

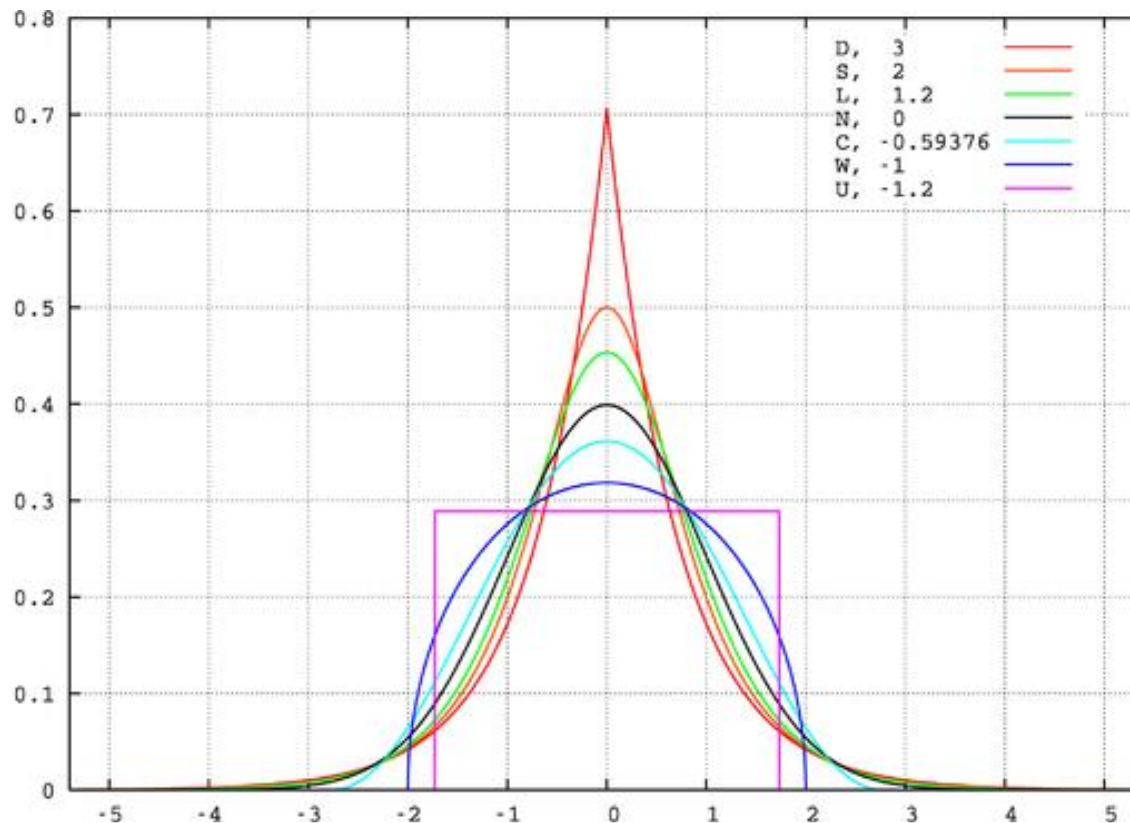
$$\hat{\mu}_4 = \frac{m_4}{m_2^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^4 / n}{\left( \sum_{i=1}^n (x_i - \bar{x})^2 / n \right)^2} - 3$$

- 대칭이면 0, 표준정규분포의 첨도는 0
- 첨도가 0보다 크면 자료분포의 형태가 표준정규분포보다 더 뾰족하다는 것을 의미
- 첨도가 0보다 작으면 자료분포의 형태가 표준정규분포보다 더 납작하다는 것을 의미

## 2.1 기술통계량 (Descriptive Statistics)

### ■ 왜도와 첨도

#### ▪ 첨도 (Kurtosis)



## 2.2 R을 활용한 기초통계량 계산

### ■ 기초 통계량 계산 함수

#### ▪ 기초 통계량 계산 함수

함수	내용	함수	내용		
ave, mean	평균	median	중앙값	summary	다섯수치요약, 평균
var	분산	sd	표준편차	fivenum	다섯수치요약
sum	합계	range, IQR	범위, 사분위범위	scale	표준화
weighted.mean	가중평균	quantile	사분위수	skewness	왜도
min, max	최소값, 최대값	rank	순위	kurtosis	첨도

```
> stat = c(87,85,86,96,78,83,89,95,92,68)
> sum(stat)
[1] 859
> mean(stat)    # 평균
[1] 85.9
> max(stat)     # 최대값
[1] 96
> min(stat)     # 최소값
[1] 68
> range(stat)   # 범위
[1] 68 96
```

```
> var(stat)     # 분산
[1] 69.43333
> sd(stat)      # 표준편차
[1] 8.332667
> median(stat)  # 중앙값
[1] 86.5
> rank(stat)    # 순위
[1] 6 4 5 10 2 3 7 9 8 1
> summary(stat) # 다섯수치요약
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 68.00  83.50   86.50   85.90  91.25   96.00
```

## 2.2 R을 활용한 기초통계량 계산

### ■ 기초 통계량 계산 함수

#### ▪ 기초 통계량 계산 함수

함수	의미
<code>table(...)</code>	<ul style="list-style-type: none"><li>• 분할표를 작성 한다.</li><li>• 반환 값은 <code>table</code> 클래스의 인스턴스로, 인자의 ...에 지정한 factor들의 모든 조합에 대해 빈도수를 구한 결과를 저장함</li></ul>
<code>which.max(x)</code>	<ul style="list-style-type: none"><li>• 최대값이 저장된 위치의 색인을 반환한다.</li></ul>

```
> fivenum(stat)
[1] 68.0 83.0 86.5 92.0 96.0
> y <- rnorm(500, 0, 5)
> mean(y)
[1] -0.3851284
> sd(y)
[1] 5.244368
> install.packages("moment")
> library(moments)
> skewness(y)
[1] -0.009739736
> kurtosis(y)
[1] 2.682978
```

```
> x =factor(c("a","b","c","c","c","d","d"))
> x
[1] a b c c c d d
Levels: a b c d
> table(x)
x
a b c d
1 1 3 2
> which.max(table(x))    # 최빈값
c
3
> names(table(x))[3]
[1] "c"
```



## 2.2 R을 활용한 기초통계량 계산

### ■ apply 계열 함수

- R에는 벡터, 행렬 또는 데이터 프레임에 임의의 함수를 적용한 결과를 얻기 위한 apply 계열 함수가 있다.
- 이 함수들은 데이터 전체에 함수를 한번에 적용하는 벡터 연산을 수행함으로 속도가 빠르다.

#### < apply 계열 함수 >

함수	의미	비고
apply( ... )	• 배열 또는 행렬에 주어진 함수를 적용한 뒤 그 결과를 벡터, 배열 또는 리스트로 반환한다.	• 배열 또는 행렬에 적용
lapply(...)	• 벡터, 리스트 또는 표현식에 함수를 적용하여 그 결과를 리스트로 반환한다.	• 결과가 리스트
sapply(...)	• lapply와 유사하지만 결과를 벡터, 행렬 또는 배열로 반환함	• 결과를 벡터, 행렬 또는 배열
tapply(...)	• 벡터에 있는 데이터를 특정 기준에 따라 그룹으로 묶은 뒤 각 그룹마다 주어진 함수를 적용하고 그 결과를 반환한다.	• 데이터를 그룹으로 묶은 뒤 함수를 적용
mapply(...)	• sapply의 확장된 버전으로, 여러 개의 벡터 또는 리스트를 인자로 받아 함수에 각 데이터의 첫째 요소들을 적용한 결과, 둘째 요소들을 적용한 결과, 셋째 요소들을 적용한 결과 등을 반환한다.	• 여러 데이터를 함수의 인자로 적용

## 2.2 R을 활용한 기초통계량 계산

### ■ apply 함수

- apply 함수는 행렬의 행 또는 열 방향으로 특정 함수를 적용하는 데 사용된다.
- rowSums, colSums, rowMeans, colMeans을 통해서도 합 또는 평균을 구할 수 있다.

#### <apply, rowSums, colSums, rowMeans, colMeans 함수>

함수	의미
apply(x, MARGIN, FUN)	<ul style="list-style-type: none"><li>• 배열 또는 행렬에 FUN을 MARGIN 방향으로 적용하여 그 결과를 벡터, 배열 또는 리스트로 반환한다.</li><li>• margin=1 이면 행, margin=2 이면 열</li><li>• 반환 값은 FUN이 길이 1인 벡터들이면 벡터, 1보다 큰 벡터들이면 행렬, 서로 다른 길이의 벡터들이면 리스트이다.</li></ul>
rowSums(x, na.rm=FALSE)	<ul style="list-style-type: none"><li>• 숫자 배열 또는 데이터 프레임에서 행의 합을 구한다.</li></ul>
colSums(x, na.rm=FALSE)	<ul style="list-style-type: none"><li>• 숫자 배열 또는 데이터 프레임에서 열의 합을 구한다.</li></ul>
rowMeans(x, na.rm=FALSE)	<ul style="list-style-type: none"><li>• 숫자 배열 또는 데이터 프레임에서 행의 평균을 구한다.</li></ul>
colMeans(x, na.rm=FALSE)	<ul style="list-style-type: none"><li>• 숫자 배열 또는 데이터 프레임에서 열의 평균을 구한다.</li></ul>

## 2.2 R을 활용한 기초통계량 계산

### ■ apply 함수

```
> (d = matrix(1:9, ncol=3))
      [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
> apply(d,1, sum)           # margin=1 이면 행
[1] 12 15 18
> apply(d,2, sum)           # margin=2 이면 열
[1]  6 15 24
> head(iris,3)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1          3.5          1.4          0.2  setosa
2          4.9          3.0          1.4          0.2  setosa
3          4.7          3.2          1.3          0.2  setosa
> apply(iris[,1:4],2,sum)
Sepal.Length Sepal.Width Petal.Length Petal.Width
      876.5       458.6       563.7       179.9
> colSums(iris[,1:4])
Sepal.Length Sepal.Width Petal.Length Petal.Width
      876.5       458.6       563.7       179.9
```

## 2.2 R을 활용한 기초통계량 계산

### ■ lapply 함수

- lapply 함수는 리스트로 반환하는 특징이 있는 apply 계열 함수이다.

<lapply, unlist 함수>

함수	의미
<code>lapply(x, FUN, ...)</code>	<ul style="list-style-type: none"><li>벡터, 리스트, 표현식, 데이터 프레임 등에 함수를 적용하고 그 결과를 <b>리스트로 반환</b>한다. 반환 값은 x와 같은 길이의 리스트이다.</li></ul>
<code>unlist(x, recursive=FALSE, use.names=TRUE)</code>	<ul style="list-style-type: none"><li>리스트 구조를 벡터로 변환한다.</li><li>반환 값은 벡터다.</li></ul>

```
> (result=lapply(1:3, function(x) {x*2}))
[[1]]
[1] 2

[[2]]
[1] 4

[[3]]
[1] 6
> str(result)
List of 3
 $ : num 2
 $ : num 4
 $ : num 6
```

```
> (t=unlist(result))
[1] 2 4 6

> x=list(a=1:3, b=4:6)

> (y=lapply(x, mean))
$a
[1] 2
$b
[1] 5
> str(y)
List of 2
 $ a: num 2
 $ b: num 5
```

## 2.2 R을 활용한 기초통계량 계산

### ■ lapply 함수

```
> y=lapply(iris[, 1:4], mean); y  # 리스트
$Sepal.Length
[1] 5.843333

$Sepal.Width
[1] 3.057333

$Petal.Length
[1] 3.758

$Petal.Width
[1] 1.199333

> t=data.frame(y); t          # 데이터프레임
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1    5.843333    3.057333     3.758     1.199333
> u=unlist(y); u             # 벡터
Sepal.Length Sepal.Width Petal.Length Petal.Width
    5.843333    3.057333    3.758000    1.199333

> colMeans(iris[, 1:4]) # 벡터
Sepal.Length Sepal.Width Petal.Length Petal.Width
    5.843333    3.057333    3.758000    1.199333
```

## 2.2 R을 활용한 기초통계량 계산

### ■ apply 함수

- apply 함수는 lapply 함수와 유사하지만 리스트 대신 행렬, 벡터 등의 데이터 타입으로 결과를 반환함

함수	의미
apply(x, FUN, ...)	<ul style="list-style-type: none"><li>• 벡터, 리스트, 표현식, 데이터 프레임 등에 함수를 적용하고 그 결과를 벡터 또는 행렬로 반환한다.</li><li>• 반환 값은 FUN이 길이 1인 벡터들이면 벡터, 1보다 큰 벡터들이면 행렬이다.</li></ul>

```
> x=apply(iris[, 1:4], mean); x
Sepal.Length Sepal.Width Petal.Length Petal.Width
      5.843333      3.057333      3.758000      1.199333
> class(x)
[1] "numeric"
> as.data.frame(x)
              x
Sepal.Length 5.843333
Sepal.Width  3.057333
Petal.Length 3.758000
Petal.Width  1.199333
> as.data.frame(t(x) )
      Sepal.Length Sepal.Width Petal.Length Petal.Width
1      5.843333      3.057333      3.758      1.199333
```

```
> apply(iris, class)
Sepal.Length Sepal.Width Petal.Length Petal.Width
Species
      "numeric"      "numeric"      "numeric"      "numeric"
"factor"

> y=apply(iris[, 1:4], function(x) {x>3}); y
> class(y)
[1] "matrix"

> head(y)
      Sepal.Length Sepal.Width Petal.Length Petal.Width
[1,]           TRUE           TRUE          FALSE          FALSE
[2,]           TRUE          FALSE          FALSE          FALSE
```

## 2.2 R을 활용한 기초통계량 계산

### ■ tapply 함수

- tapply 함수는 **그룹별로 함수를 적용**하기 위한 apply 계열 함수로 배열을 반환한다.

함수	의미
tapply(x, INDEX, FUN, ...)	<ul style="list-style-type: none"><li>• 벡터 등에 저장된 데이터를 주어진 기준에 따라 <b>그룹으로 묶은 뒤</b> 각 그룹에 함수를 적용하고 그 결과를 반환한다.</li><li>• 반환 값은 배열이다.</li></ul>

```
> tapply(1:10, rep(1,10), sum)
1
55
> tapply(1:10, 1:10 % 2, sum)
0 1
30 25
> tapply(iris$Sepal.Length, iris$Species,
mean)
      setosa versicolor  virginica
      5.006      5.936      6.588
```

```
> m <- matrix(1:8, ncol=2,
  dimnames=list(c("spring", "summer", "fall",
    "winter"), c("male", "female")))
> m
      male female
spring     1      5
summer     2      6
fall       3      7
winter     4      8
> tapply(m, list(c(1,1,2,2,1,1,2,2),
  c(1,1,1,1,2,2,2,2)), sum)
      1 2
1 3 11
2 7 15
```

## 2.2 R을 활용한 기초통계량 계산

### ■ mapply 함수

- mapply 함수는 sapply 함수와 유사하지만 **다수의 인자를 함수에 넘긴다는** 점에서 차이가 있다.

함수	의미
mapply(FUN, ...)	<ul style="list-style-type: none"><li>• 함수에 리스트 또는 벡터로 주어진 인자를 적용한 결과를 반환한다.</li><li>• ...에 주어진 여러 데이터가 있을 때 FUN에 이들 데이터 각각의 첫째 요소를 인자로 전달하여 실행한 결과, 각각의 둘째 요소를 인자로 전달하여 실행한 결과 등을 반환한다.</li></ul>

```
> rnorm(10, 0, 1)
[1] -0.277173007  0.256457498  1.424091000  0.715459531 -0.005594503  0.896659832
[7] -0.175640622 -0.194556803  1.279371985  0.428433761
```

```
> mapply(rnorm,
         c(1, 2, 3),      # n
         c(0, 10, 100),   # mean
         c(1, 1, 1))      # sd
```

```
[[1]]
[1] -1.323438
```

```
[[2]]
[1] 9.447432 9.494280
```

```
[[3]]
[1] 99.80582 100.97717 100.82098
```

```
> mapply(mean, iris[, 1:4])
Sepal.Length Sepal.Width Petal.Length Petal.Width
5.843333      3.057333      3.758000      1.199333
```



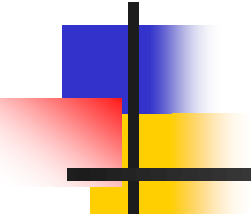
## 2.2 R을 활용한 기초통계량 계산

### ■ aggregate 함수

- aggregate 함수는 그룹별 연산을 위한 함수임. tapply와 같은 동일한 결과를 출력함.

함수		Arguments
aggregate(x, by, FUN, ...)	x	• R 객체
	by	• 데이터프레임안의 변수로서 그룹을 하고자 하는 변수로 factor이어야 함
	FUN	• 적용하고자 하는 함수
aggregate(formula , data, FUN, ...)	formula	• $y \sim x$ or $\text{cbind}(y1, y2) \sim x1 + x2$ 와 같은 <a href="#">formula</a> . x 또는 x1, x2는 factor이어야 함
	data	• 적용할 데이터프레임 또는 리스트
	FUN	• 적용하고자 하는 함수

```
> aggregate(Sepal.Width ~ Species, iris, mean)
  Species Sepal.Width
1  setosa      3.428
2 versicolor  2.770
3 virginica   2.974
> tapply(iris$Sepal.Width, iris$Species, mean)
  setosa versicolor virginica 
  3.428      2.770      2.974
```



## 2.3 자료의 그래프 표현

---

### ■ 자료의 그래프 표현

- 수집된 자료를 효과적으로 정리, 요약하기 위해서 적절한 도표나 그래프를 사용하여 시각화 할 필요가 있음
- 자료의 시각화를 통해 자료의 대략적인 분포 형태 및 특성 등을 파악할 수 있음
  - 자료의 시각화를 통해 대칭 혹은 비대칭의 정도, 대부분의 자료로부터 동떨어진 이상점(outliers)의 유무, 그리고 상대적으로 많은 자료가 분포되어 있는 봉우리의 위치 등
- 그래프 표현은 수치를 이용하는 것을 보완하여 자료의 내포된 정보를 보다 쉽고 빠르게 파악할 수 있게 함
- 탐색적 자료 분석의 단계에서는 필수적인 요소임

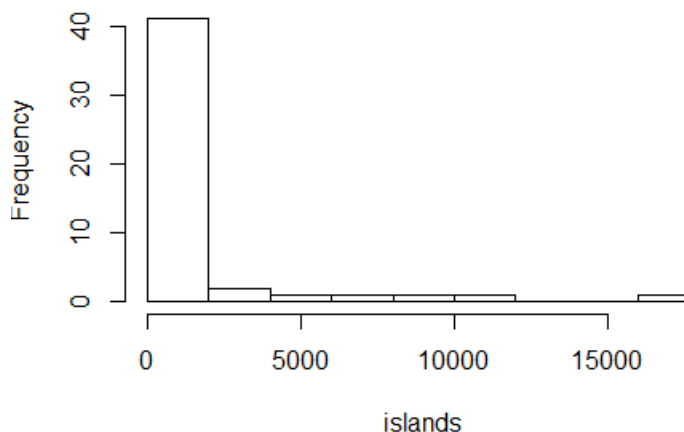
## 2.3 자료의 그래프 표현

### ■ hist 함수

함수	의미
<pre>hist(x, breaks = "Sturges",      freq = NULL, probability = !freq,      include.lowest = TRUE, right = TRUE,      density = NULL, angle = 45, col = NULL, border = NULL,      main = paste("Histogram of" , xname),      xlim = range(breaks), ylim = NULL,      xlab = xname, ylab,      axes = TRUE, plot = TRUE, labels = FALSE,      nclass = NULL, warn.unused = TRUE, ...)</pre>	<ul style="list-style-type: none"><li>• x: 히스토그램을 그려야할 벡터</li><li>• breaks: breakpoint를 주어야 할 벡터, 함수, 수, 문자 또는 셀의 수를 계산 할 함수</li><li>• freq: 빈도이면 TRUE, 확률이면 FALSE</li><li>• include.lowest: 첫번째를 포함하면 TRUE</li><li>• density: 밀도함수 선을 그림. Default는 NULL</li></ul>

Histogram of islands

```
> hist(islands)
```

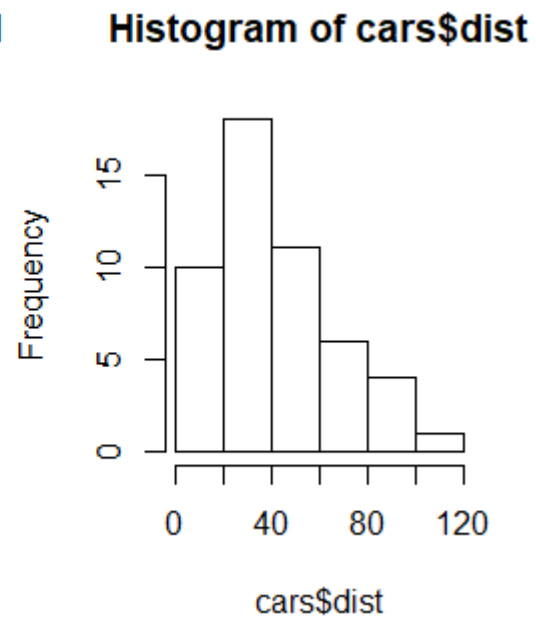
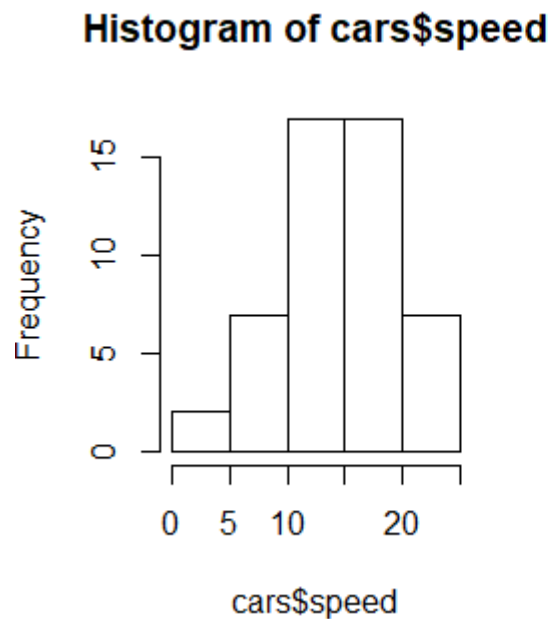


## 2.3 자료의 그래프 표현

### ■ hist 함수

- cars example : speed, dist

```
> par(mfrow = c(1, 2))  
> hist(cars$speed)  
> hist(cars$dist)
```

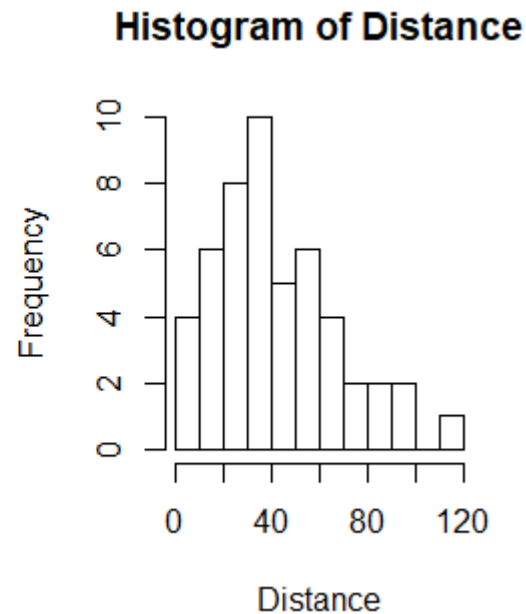
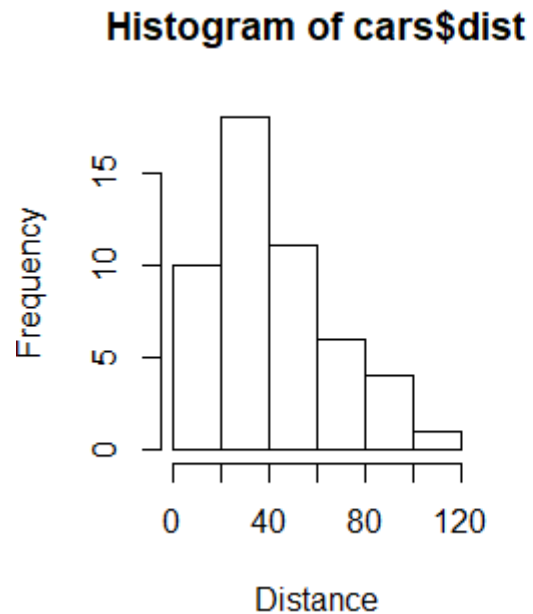


## 2.3 자료의 그래프 표현

### ■ hist 함수

- histogram 장식

```
> hist(cars$dist, xlab = "Distance")  
> hist(cars$dist, 15, xlab = "Distance", main = "Histogram of Distance")
```



## 2.3 자료의 그래프 표현

### ■ hist 함수

- histogram에 추정 밀도 추가하기

```
> par(mfrow = c(1, 1))  
> hist(cars$dist, 15, xlab = "Distance", main = "Histogram of Distance", prob = T)  
> lines(density(cars$dist))
```

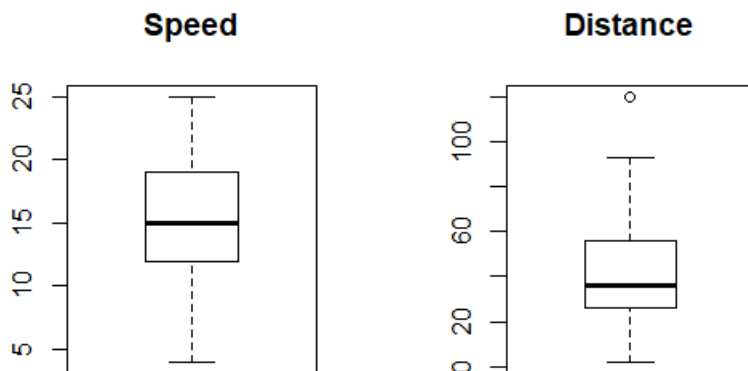


## 2.3 자료의 그래프 표현

### ■ boxplot 함수

함수	의미
<code>boxplot(formula, data = NULL, ..., subset, na.action = NULL, drop = FALSE, sep = ".", lex.order = FALSE)</code>	<ul style="list-style-type: none"> <li>• <code>y~grp</code> 와 같은 식</li> <li>• <code>data</code>: 데이터프레임</li> </ul>
<code>boxplot(x, ..., range = 1.5, width = NULL, varwidth = FALSE, notch = FALSE, outline = TRUE, names, plot = TRUE, border = par("fg"), col = NULL, log = "", pars = list(boxwex = 0.8, staplewex = 0.5, outwex = 0.5), horizontal = FALSE, add = FALSE, at = NULL)</code>	<ul style="list-style-type: none"> <li>• <code>x</code>: 그림을 그려야할 벡터</li> </ul>

```
> par(mfrow = c(1, 2))
> boxplot(cars$speed, main = "Speed")
> boxplot(cars$dist, main = "Distance")
```



```
> boxplot.stats(cars$dist)
$stats
[1]  2 26 36 56 93

$n
[1] 50

$conf
[1] 29.29663 42.70337

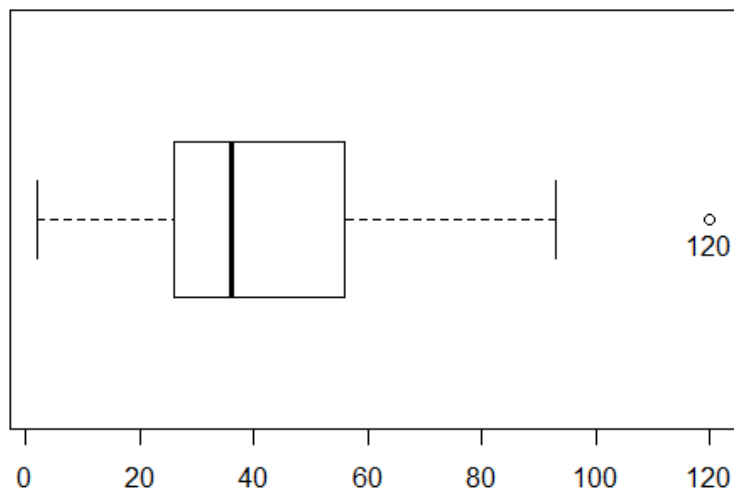
$out
[1] 120
```

## 2.3 자료의 그래프 표현

### ■ boxplot 함수

- outlier에 label 붙이기
- pos는 문자열은 하단에 표시하라는 옵션

```
> boxstats <- boxplot(cars$dist, horizontal = TRUE)  
> boxstats  
> text(boxstats$out, rep(1, NROW(boxstats$out)), label = boxstats$out,  
      pos = rep(1, NROW(boxstats$out)))
```



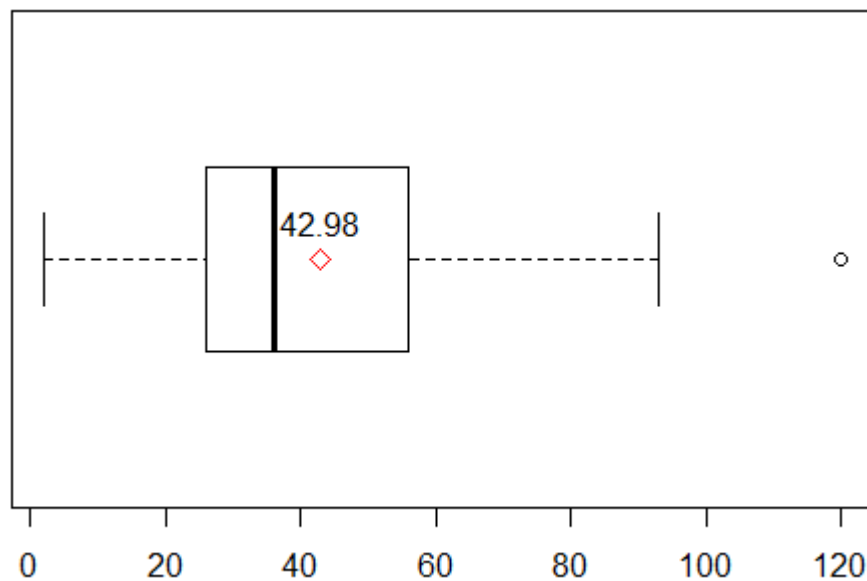


## 2.3 자료의 그래프 표현

### ■ boxplot 함수

- boxplot에 평균 추가

```
> boxstats <- boxplot(cars$dist, horizontal = TRUE)  
> mean.dist <- mean(cars$dist)  
> text(mean.dist, 1 + 0.08, labels = mean.dist)
```

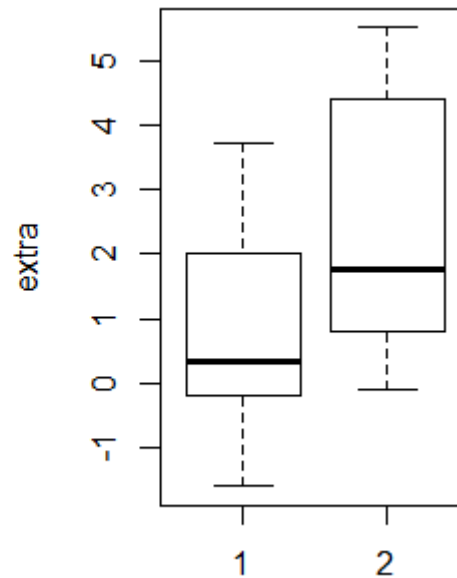
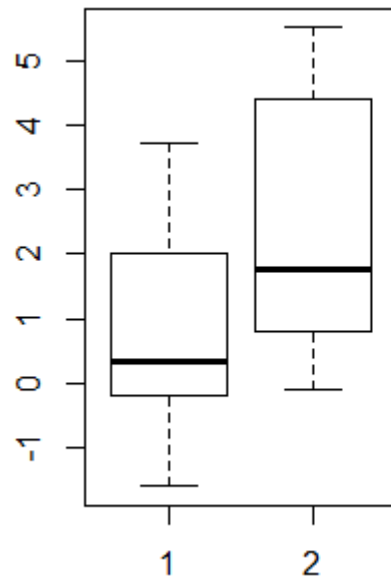


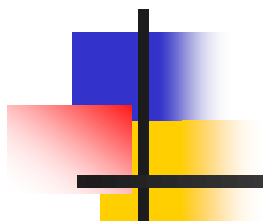
## 2.3 자료의 그래프 표현

### ■ boxplot 함수

- boxplot을 이용한 그룹간 분포 비교

```
> par(mfrow = c(1, 2))  
> boxplot(extra ~ group, data = sleep)  
> plot(extra ~ group, data = sleep)
```





---

### 3. 질적 자료에 대한 EDA

---

### 3.1 질적자료의 요약

#### ■ 빈도표

Drink 데이터

ID	Age	Drink	ID	Age	Drink
1	20대	C	55	10대	C
2	10대	D	56	10대	D
3	10대	D	57	30대	D
4	10대	A	58	10대	D
5	10대	B	59	20대	C
6	10대	D	60	20대	B
7	30대	B	61	10대	B
...	...	...	...	...	...
54	10대	D	108	10대	B

```
Drink = read.csv("D:/Drink.csv",header=TRUE)
table(Drink$Age) # 1차원 빈도표
table(Drink$Age,Drink$Drink) # 2차원 분할표(교차표)
```

## 3.1 질적자료의 요약

### ■ 빈도표

```
Drink = read.csv("D:/Drink.csv",header=TRUE)
table(Drink$Age) # 1차원 빈도표
table(Drink$Age,Drink$Drink) # 2차원 분할표(교차표)
```

```
> table(Drink$Age) # 1차원 빈도표
```

10대	20대	30대
38	38	32

```
> table(Drink$Age,Drink$Drink) # 2차원 분할표(교차표)
```

	A	B	C	D
10대	10	14	2	12
20대	13	7	10	8
30대	12	4	6	10

## 3.1 질적자료의 요약

### ■ 주변합계의 추가

```
addmargins(table(Drink$Age))  
addmargins(table(Drink$Age,Drink$Drink),margin=2)
```

```
> addmargins(table(Drink$Age))
```

10대	20대	30대	Sum
38	38	32	108

```
> addmargins(table(Drink$Age,Drink$Drink),margin=2)
```

	A	B	C	D	Sum
10대	10	14	2	12	38
20대	13	7	10	8	38
30대	12	4	6	10	32

## 3.1 질적자료의 요약

### ■ 비율(백분율)의 출력

```
Age.table = table(Drink$Age)
prop.table(Age.table)*100
round(prop.table(Age.table)*100,1)
```

```
> Age.table = table(Drink$Age)
> prop.table(Age.table)*100
```

10대	20대	30대
35.18519	35.18519	29.62963

```
> round(prop.table(Age.table)*100,1)
```

10대	20대	30대
35.2	35.2	29.6

```
Drink.table = table(Drink$Age,Drink$Drink)
round(addmargins(prop.table(Drink.table,margin=1)*100,margin=2),1)
```

```
> Drink.table = table(Drink$Age,Drink$Drink)
> round(addmargins(prop.table(Drink.table,margin=1)*100,margin=2),1)
```

	A	B	C	D	Sum
10대	26.3	36.8	5.3	31.6	100.0
20대	34.2	18.4	26.3	21.1	100.0
30대	37.5	12.5	18.8	31.2	100.0

### 3.1 질적자료의 요약

#### ■ Xtabs 함수를 이용한 분할표 출력

```
xtabs(~Age,data=Drink)  
xtabs(~Age+Drink,data=Drink)
```

```
> xtabs(~Age,data=Drink)
```

Age

10대	20대	30대
38	38	32

```
> xtabs(~Age+Drink,data=Drink)
```

Drink

Age	A	B	C	D
10대	10	14	2	12
20대	13	7	10	8
30대	12	4	6	10



### 3.1 질적자료의 요약

#### ■ 요약데이터에 대한 빈도표

Count 데이터

Obs.	Age	Drink	Freq	Obs.	Age	Drink	Freq
1	10대	A	10	7	20대	C	10
2	10대	B	14	8	20대	D	8
3	10대	C	2	9	30대	A	12
4	10대	D	12	10	30대	B	4
5	20대	A	13	11	30대	C	6
6	20대	B	7	12	30대	D	10

```
Count = read.csv("data/Count.csv",header=TRUE)
Drink.table = xtabs(Freq~Age+Drink,data=Count)
Drink.table
```

```
> Drink.table
```

	Drink			
Age	A	B	C	D
10대	10	14	2	12
20대	13	7	10	8
30대	12	4	6	10

### 3.1 질적자료의 요약

#### ■ 요약데이터에 대한 빈도표

```
round(addmargins(prop.table(Drink.table,margin=1)*100,margin=2),2)
```

```
> round(addmargins(prop.table(Drink.table,margin=1)*100,margin=2),2)
```

	Drink				
Age	A	B	C	D	Sum
10대	26.32	36.84	5.26	31.58	100.00
20대	34.21	18.42	26.32	21.05	100.00
30대	37.50	12.50	18.75	31.25	100.00

## 3.2 그래프를 이용한 질적자료의 표현

### ■ 막대도표

```
Chart.table = table(Drink$Drink)  
barplot(Chart.table,col="green",main="선호하는 음료수",cex.axis=1)
```

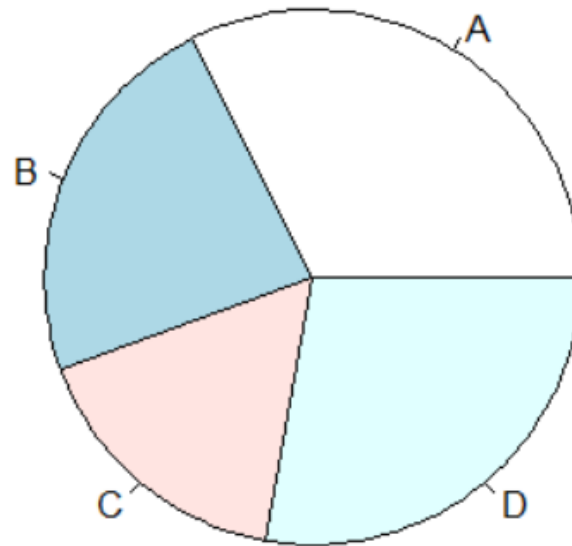


## 3.2 그래프를 이용한 질적자료의 표현

### ■ 원도표

```
pie(Chart.table,main="선호하는 음료수",cex=1)
```

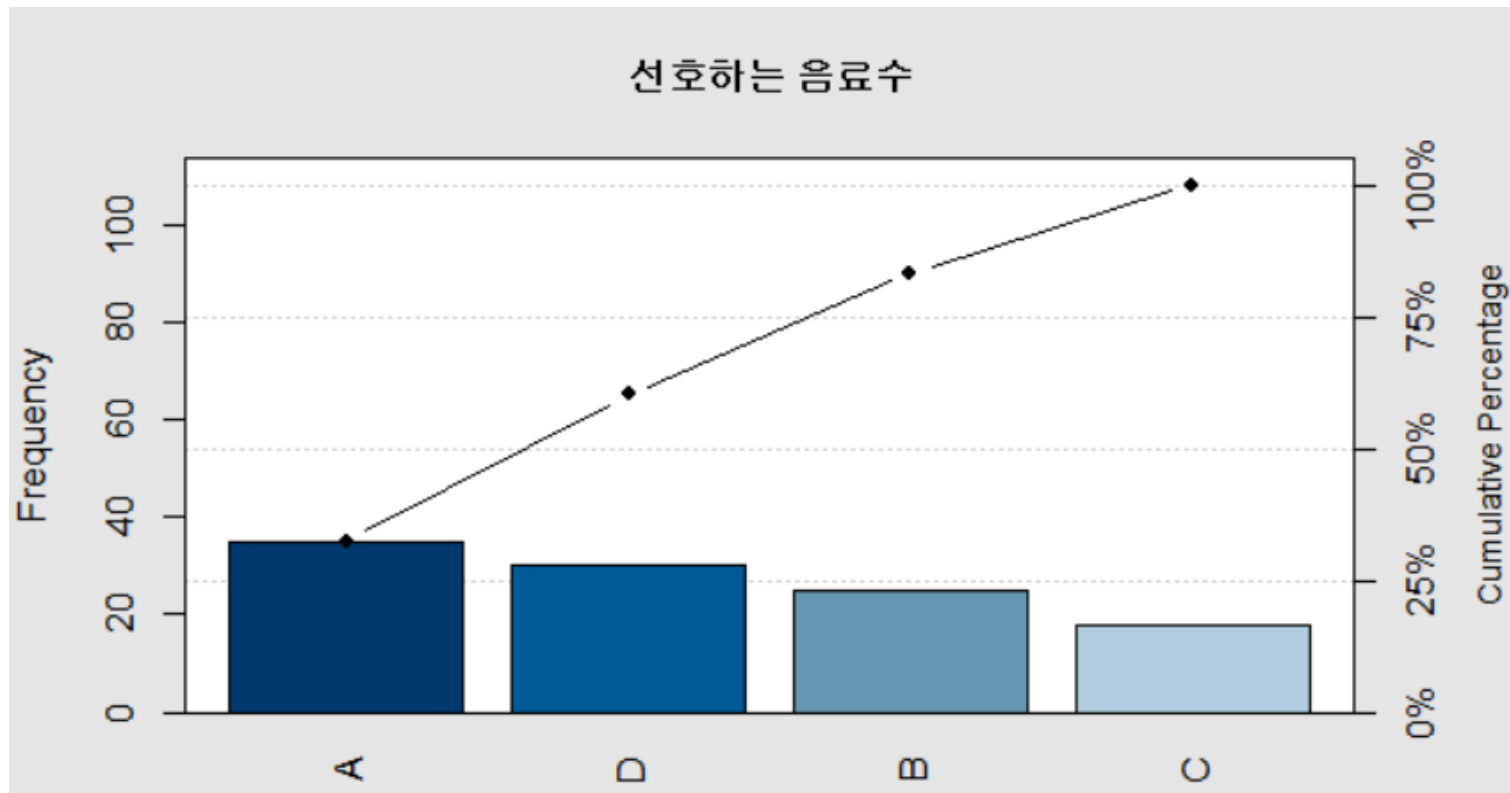
선호하는 음료수



## 3.2 그래프를 이용한 질적자료의 표현

### ■ 파레토 도표

```
install.packages("qcc")  
library(qcc)  
pareto.chart(Chart.table,main="선호하는 음료수")
```



## 3.2 그래프를 이용한 질적자료의 표현

### ■ 모자이크 도표

```
Drink.table = table(Drink$Age,Drink$Drink)
mosaicplot(Drink.table,color=TRUE,cex.axis=1)
mosaicplot(Age~Drink,data=Drink,color=TRUE,cex.axis=1)
```



## ■ 연결선 그래프

Temperature 데이터: 우리나라 1984년도 8월의 일자별 기온

```
temp = read.csv("data/temperature.csv",header=TRUE)
head(temp)
plot(temp$Date,temp$Temperature,type="b",lty="solid",lwd=1,pch=1,cex=2,
      col="blue",xlab="Date",ylab="Temperature",xlim=c(1,30),ylim=c(20,40))
plot(Temperature~Date,data=temp,type="b",lty="solid",lwd=1,pch=1,cex=2,
      col="blue",xlab="Date",ylab="Temperature",xlim=c(1,30),ylim=c(20,40))
```

```
> head(temp)
  Date Temperature
1    1          24
2    2          27
3    3          28
4    4          32
5    5          30
6    6          35
```

