

01

# 통계의 개념 및 통계량



# 1. 통계의 개념 및 통계량



## 가. 통계의 기초개념

### 통계학이란

- 자연 및 사회현상에서 나타나는 다양한 상황이나 측정값들을 요약하여 표현하는 것
- 자연이나 사회현상 가운데서 발생하는 다양한 데이터나 정보를 활용해 의미하는 바를 요약(summary), 분포(distribution) 이해, 추세(trend)나 일정한 패턴(pattern), 방향성(direction), 스타일(style)이나 군집유형(cluster type) 등으로 요약하여 의사결정에 활용하는 학문
- 기술통계(descriptive statistics)와 추론통계(inferential statistics)가 있음.

# 1. 통계의 개념 및 통계량



## 가. 통계의 기초개념

### ✓ 기술통계(descriptive statistics)

- 단지 관찰 혹은 측정된 데이터의 특성을 기술하는 것



경영학과 학생 100명 대상으로 안드로이드폰과 아이폰 보유 여부 현황조사

- 특정집단의 데이터를 요약하고 정리하기 위하여 사용

# 1. 통계의 개념 및 통계량



## 가. 통계의 기초개념

### ④ 추론통계(inferential statistics)

- 데이터의 특성을 기초로 하여 모집단의 특성을 일반화하거나 예측하는데 활용되는 통계
- 기본적인 데이터를 근거로 해서 **모집단(population)**의 특성을 **예측**하거나 미루어 짐작하는 것

#### » 모집단(population)

조사대상이 되는 모든 개체(사람 혹은 사물)들의 전체 집합

예

대통령선거나 국회의원 선거 시 전국 1000~2000명을 표본 추출해서 조사한 지지도

# 1. 통계의 개념 및 통계량



## 가. 통계의 기초개념

### ● 통계와 빅데이터의 관계

#### » 장점

통계적인 분석 방법을 활용하여 빅데이터를 분석 및 활용한다면 이미 구축된 DB에 저장되어 있는 데이터의 특성뿐만 아니라 새로 DB에 저장되는 데이터도 빠르게 분석할 수 있음

#### » 기업경영이나 전략수립, 시장 개척, 신제품 개발, 광고컨셉트 개발 등에 활용하여 합리적 의사결정을 할 수 있도록 지원하는 역할

# 1. 통계의 개념 및 통계량



## 나. 모수와 통계량

### 모수(parameter)

모집단의 특성을 나타내는  
수치자료

- » 조사대상 집단 모두를 조사할 때 측정이 되는 수치
- » 현실에서는 비용이나 시간, 조사지역의 광범위성 등의 한계로 거의 불가능

### 통계량(statistic)

모수를 대체하기 위해  
표본조사를 실시하여  
산출되는 수치

- » 표본집단을 특정하여 얻은 값
- » 평균(mean), 중앙값(median), 최빈값(mode), 분산(variance), 표준편차(standard deviation) 등

# 1. 통계의 개념 및 통계량



## 다. 통계분석 패키지의 종류

### ✓ 통계 패키지별 특화 유형

특화유형	통계 패키지
일반 기술 및 추론통계	SAS, SPSS, EXCEL, Matlab,
구조방정식(SEM)	LISREL, AMOS
6시그마 통계	Minitab
메타분석	MIX
빅데이터	R, R-studio, Tableau

02

# 변수의 측정과 척도





## 2. 변수의 측정과 척도



### ● 통계분석

연구의 목적에 따라 수집된 데이터들을 분석하여 정보나 결론을 얻는 일련의 과정

- ≫ **데이터 분석** : 분석자가 관심을 가지고 있는 각 개체들의 특성
- ≫ **측정** : 일반적으로 일정한 규칙에 따라 대상에 숫자를 할당하는 과정
- ≫ **척도** : 측정대상이 갖는 특성을 측정하는 잣대

## 2. 변수의 측정과 척도



### 가. 변수

- 가변적인 요인이면서 동시에 여러 가지 값으로 변할 수 있는 수를 의미
- 유형

#### ≫ 독립변수(independent variable)

원인이 되는 변수 혹은 영향을 미치는 변수

#### ≫ 종속변수(dependent variable)

결과가 되는 변수 혹은 영향을 받는 변수

$$\underline{Y} = \underline{a} + \underline{bX}$$

≫ Y: 종속변수

≫ X: 독립변수

≫ a: 상수(오차항)로 종속변수 Y의 절편

≫ b: 종속변수 Y에 대한 독립변수 X의 기울기

## 2. 변수의 측정과 척도



### 가. 변수

#### ● 양적변수 (quantitative variable)

##### 이산형 변수 (discrete variable)

특정 구간에서 서로 떨어져  
있는 자료로 정수의 값을 갖는  
셀 수 있는 자료

예

종업원의 수나 특정  
제품의 불량품 개수 등

##### 연속형 변수 (continuous variable)

특정 구간에서 어떠한  
값이라도 가질 수 있는 자료

예

물 탱크에 저장된 물의  
양, 화물차의 적재량,  
매출액, 영업이익, 원가,  
몸무게, 키, 온도 등

» 무한한 소수점으로 표현할  
수 있는 자료

## 2. 변수의 측정과 척도



### 가. 변수

- 그 외 변수의 종류

통제변수, 매개변수 등 다양한 관점에서 구분

## 2. 변수의 측정과 척도



### 나. 측정과 척도

- 측정(measurement)

특정하게 명시된 규정에 의해 수치나 혹은 다른 기호들을 통해 조사한 대상물의 특성을 기록하는 것

» 측정변수에 특정한 값을 부여하는 것

- 척도(scale)

측정대상이 가지는 고유한 특성을 기록하는 것은 측정 대상물의 유형에 따라 적절하게 기록해야 하며, 이때 적절하게 기록하는 기준을 의미

» 측정과정의 연장선상에 있는 개념으로 측정된 대상이 갖는 일직선상에서의 위치를 지정해주는 것

## 2. 변수의 측정과 척도



### 나. 측정과 척도

척도	기본특성	일상적인 활용사례	허용되는 통계량	
			기술통계	추론통계
명목	대상을 확인, 분류	주민등록번호, 운동선수 유니폼 번호 등	퍼센트 최빈값	카이스퀘어 이변량검증
서열	대상의 상대적 순서 위치	품질순위, 결승선 통과순위, 팀간의 순위 등	퍼센트 중앙값	순위서열상관 ANOVA
등간	비교대상들간 차이, 크기 등	온도계의 온도 등	범위, 평균 분산과 표준편차	단순상관, t검증, ANOVA, 회귀분석, 요인분석
비율	절대영점이 존재하고 척도값 비율을 계산하여 이용	길이 무게 등	기하학적 평균, 조화평균	분산의 계수

03

# 평균, 분산 및 표준편차



### 3. 평균, 분산 및 표준편차



#### 가. 평균

- 모집단이 지니고 있는 양적 구조의 특성치인 대표치를 나타내는 수치로 측정된 데이터(값)의 중앙으로의 집중화 경향을 파악하는 통계량
- 측정된 대부분의 값들은 평균을 중심으로 주변에 흩어져 분포
- 각 측정값들을 모두 합하여 측정치의 개수(n)로 나누면 얻을 수 있음

#### ☑ 평균의 계산식

$$\begin{aligned}\bar{x} &= \frac{1}{n}(x_1 + x_2 + \dots + x_n) \\ &= \frac{1}{n} \sum_{i=1}^n x_i\end{aligned}$$

$x_i$  :  $i$ 번째 관찰치



### 3. 평균, 분산 및 표준편차



#### 가. 평균

##### ✓ 평균의 성질

평균의 크기는 변수의 크기와 빈도 수에 의존한다.

평균 개개의 변수 값은 모르더라도 총계와 빈도 수만으로 평균을 계산할 수 있다.

반대로 평균과 빈도 수만 알면 총계를 알 수 있다.

평균은 변수들 중에서 극히 큰 값 혹은 작은 값에 의해 크게 영향을 받는다.

### 3. 평균, 분산 및 표준편차



#### 나. 분산

- 분산은 데이터를 분석하고 해석하는데 있어 가장 빈번하게 사용되는 통계량으로 데이터가 평균을 중심으로 어느 정도 흩어져 있는가를 측정하는 값
- n개의 측정된 값과 평균의 차이를 제곱해서 합한 값으로 구함
  - ≫ 분산의 크기가 클수록 자료의 흩어진 정도가 크다고 할 수 있으며, 측정에 따른 편차가 크다는 것을 의미
  - ≫ 분산이 작을수록 측정데이터의 측정신뢰성은 높다고 할 수 있음

#### ✓ 표본집단의 분산 계산식

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

$x_i$  : 관찰치  
 $\bar{x}$  : 평균치  
 $n$  : 표본수

### 3. 평균, 분산 및 표준편차



#### 다. 표준편차

- 데이터의 흩어진 정도를 측정하는 것으로 분산에 제곱근을 취한 값
- 측정된 값과 평균의 차이를 합해서 제곱한 값으로 산출되는데 여기에 제곱근을 취해서 원래대로 맞춰주는 값
  - ≫ 자료의 분산에 제곱근을 취해서 얻는 이유는 **분산이 편차의 제곱을 취하게 때문에 원래의 데이터 특성과 단위가 달라지기 때문**
  - ≫ 원래 데이터의 단위에 맞도록 수정해 주기 위해서 제곱근을 취하는 것

#### ✓ 표준편차 계산식

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

$x_i$  : 관찰치  
 $\bar{x}$  : 평균치  
 $n$  : 표본수

04

# 통계적 가설 검정



## 4. 통계적 가설검증



### 가. 가설검증

모집단의 모수에 대한 가설을 설정하고  
표본으로부터 조사한 결과에 따라  
그 가설을 선택할 것인지를  
통계적으로 결정하는 분석방법

## 4. 통계적 가설검증



### 가. 가설검증

#### 귀무가설

- 영가설 (Null Hypothesis)
- $H_0$

#### 대립가설

- 연구가설 (Alternative Hypothesis)
- $H_1$

» 귀무가설을 기각함으로써 대립가설을 채택하기 위한 의도로 실시함

## 4. 통계적 가설검증



### 나. 가설검증 절차

- 01 검증하고자 하는 연구목적 확인
- 02 귀무가설과 연구가설 설정
- 03 적합한 통계적 기법과 부합되는 검증통계량 선택
- 04 유의수준 (p-value) 알파( $\alpha$ )값 결정
- 05 표본의 크기를 결정하고 데이터 수집한 후 검증통계에 활용할 임계값 계산

## 4. 통계적 가설검증



### 나. 가설검증 절차

06

귀무가설 : 검증통계량의 표본분포를 이용하여  
검증통계와 연관된 확률 결정

07

유의수준과 검증통계에서 산출된 확률을 비교하여  
기각역에 위치하는지, 채택역에 위치하는지 결정

08

귀무가설을 기각할 것인지 채택할 것인지의  
통계적 의사결정 실시



## 4. 통계적 가설검증



### 다. 가설검증통계의 유형

#### ✓ 검증통계기법

- 평균검증
- 비율검증
- 평균의 차이 검증
- 비율의 차이 검증
- 분산분석
- 상관관계분석
- 회귀분석
- $\chi^2$ (카이스퀘어)독립성 검증
- 적합성 검증
- 판별분석 등

## 4. 통계적 가설검증



### 라. t-검증

두 집단 간의 평균 차이 여부를 검증하는 방법

- t - 검증의 구분  
단일표본 t-검증, 독립표본 t - 검증, 대응표본 t-검증
- t-검증 활용사례
  - ≫ 두 회사 가전제품 간의 선호도 차이 검증
  - ≫ 두 회사 다이어트 제품의 효과 차이 검증
  - ≫ 일본과 한국의 초등학생 IQ차이 검증 등
- 비율척도 및 등간척도 데이터 검증

## 4. 통계적 가설검증



### 마. F-검증

3개 이상의 집단들에 대한 평균을 비교하여 한 개 이상 집단 간에 차이가 있는지를 검증하는 방법

- 집단을 구분하는 인자의 수에 따른 **분산분석의 유형**  
**일원배치** 분산분석, **이원배치** 분산분석, **다원배치** 분산분석
- F-검증 활용사례
  - ≫ 20대, 30대, 40대, 50대 연령별 생활만족도 차이 검증
  - ≫ 대도시, 중도시, 소도시 간의 1인당 노인복지 만족도 차이 검증
  - ≫ 서울, 대구, 인천, 부산 지역의 주민 평균소득 차이 검증 등
- 비율척도 및 등간척도 데이터 검증  
(ANOVA분석이라고도 함)

05

$\chi^2$ 검정



# 5. $\chi^2$ 검정



## 가. $\chi^2$ 검정

범주형 변수 간의 독립성이나 적합성을 검증하는 방법

- $\chi^2$  검정의 구분  
2×2 분할표 검정, 2×m 분할표 검정, n×m 분할표 검정
- $\chi^2$  검증 활용사례
  - ≫ 성별 스마트폰 인지도 차이 검증
  - ≫ 학년별 축제 참석여부 차이 검증
  - ≫ 학년별 취미경향 차이 검증
  - ≫ 지역별 선호 정당 차이 검증
  - ≫ 가족의 규모와 세탁기 크기 독립성 검증
  - ≫ 소비자들의 자동차 색상에 대한 선호도 적합성 검증
- 명목척도 및 서열척도 데이터 검증

06

# 상관관계 분석

BIG

DATA

## 6. 상관관계 (correlation analysis) 분석



### 가. 상관관계 분석의 개념

특정한 변수 X와 또 다른 변수 Y 사이에 존재하는 상호관련성을 분석하는 기법

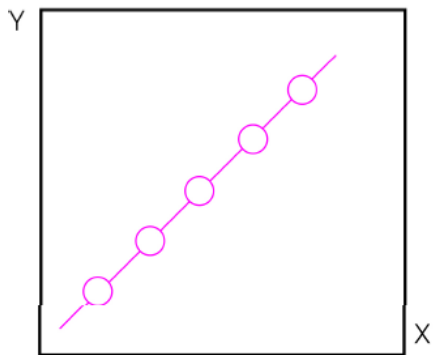
- 측정척도에 따라 다르게 분석하는 기법
- 칼 피어슨의 **피어슨 상관계수**, 스피어만의 서열상관, 켄달의 타우계수 등이 있음
- 연관성을 파악하는 통계량
  - ≫ 상관계수 (correlation coefficient)인  $r$
  - ≫ 결정계수 (coefficient of determinant)인  $r^2$

## 6. 상관관계(correlation analysis) 분석

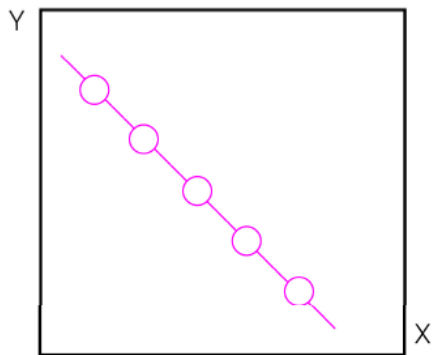


### ☑ 상관계수 평가기준

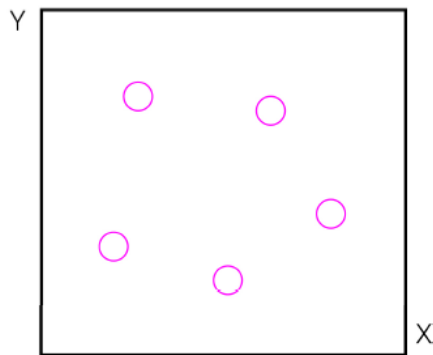
- $r = 1.00$  (완전한 상관관계)
- $r = 0.90$  (매우 높은 상관관계)
- $r = 0.70 \sim 0.80$  (높은 상관관계)
- $r = 0.50 \sim 0.60$  (보통 상관관계)
- $r = 0.30 \sim 0.40$  (약한 상관관계)
- $r = 0.20$  이하 (상관관계 없음)



〈양의 상관관계〉



〈음의 상관관계〉



〈상관관계 없음〉

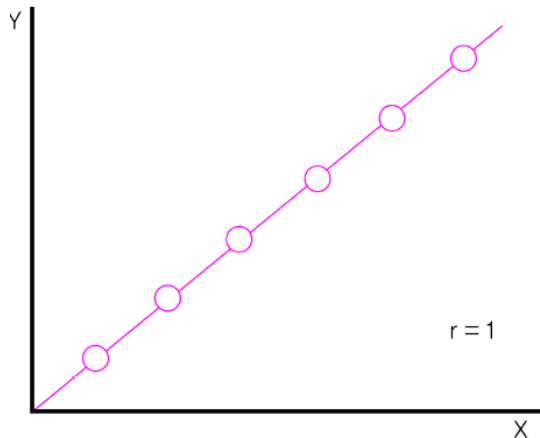


## 6. 상관관계 (correlation analysis) 분석

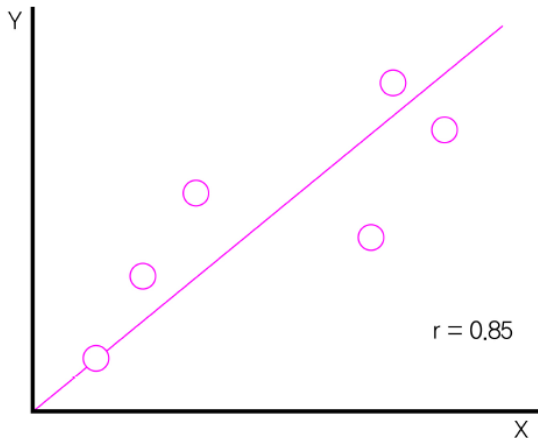


### ☑ 상관계수 평가기준

- $r = 1.00$  (완전한 상관관계)
- $r = 0.90$  (매우 높은 상관관계)
- $r = 0.70 \sim 0.80$  (높은 상관관계)
- $r = 0.50 \sim 0.60$  (보통 상관관계)
- $r = 0.30 \sim 0.40$  (약한 상관관계)
- $r = 0.20$  이하 (상관관계 없음)



가. 완전상관관계



나. 높은상관관계

## 6. 상관관계 (correlation analysis) 분석



### ☑ 상관관계 계산 식

$$r = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2}}$$

$x$  :  $x$  변수

$y$  :  $y$  변수

$s_x^2$  :  $x$  변수의 분산

$s_y^2$  :  $y$  변수의 분산

$s_x s_y$  :  $x$  변수와  $y$  변수의 공분산

$$= \frac{s_x s_y}{\sqrt{s_x^2 s_y^2}}$$

07

# 요인분석



# 7. 요인분석(factor analysis) 분석



## 가. 요인분석의 개념

자료의 감축(reduction)과 요약(summarization)을 위한 분석기법으로 실제로 존재하는 어떤 사회현상에 관한 다양한 변수들을 측정하여 분석할 때 직접 측정할 수 없는 일련의 개념 혹은 요인들을 확인하기 위한 분석방법



다양한 변수들을 몇 개의 개념이나 요인으로 결합시켜서 측정변수들의 내용을 단순화(simplify)시켜 통찰력을 높이하고자 할 때 사용

# 7. 요인분석(factor analysis) 분석



## 가. 요인분석의 개념

### ☑ 요인분석을 실시하는 목적

- 데이터 축소
- 자료 요약
- 불필요한 자료 제거
- 요인의 구조 파악
- 측정도구의 타당성 평가
- 다중공선성 문제 해결

# 7. 요인분석(factor analysis) 분석



## 나. 요인분석의 절차

- 01 요인분석 문제를 정의하고 요인분석을 하기 위한 변수의 확인
- 02 변수들의 상관관계행렬을 구성하고 요인분석방법의 선택
- 03 추출하고자 하는 요인의 수와 회전방법의 결정
- 04 요인의 회전 (배리맥스, 쿼티맥스, 이쿼맥스, 직접 오블리민 등)
- 05 회전된 요인의 설명
- 06 목적에 근거하여 요인점수 계산
- 07 요인분석 모델의 적합성 결정

08

# 회귀분석



## 8. 회귀분석(regression analysis)



### 가. 단순회귀분석

- 사회현상이나 자연현상에서 존재하는 원인과 결과간의 인과성(causality)을 규명하는 분석방법
- 단순회귀분석  
독립변수와 종속변수가 각 1개씩인 분석기법
- 다중회귀분석  
독립변수가 2개 이상인 분석기법



## 8. 회귀분석(regression analysis)



### 가. 단순회귀분석

#### ✓ 회귀방정식

$$Y_i = \alpha X_i + c$$

$Y_i$ : 종속변수,

$X_i$ : 독립변수,

$\alpha$ : 종속변수  $Y$ 에 대한  
독립변수  $X$ 의 기울기(회귀계수)

$c$ : 독립변수  $X$ 에 의해 설명이 되지 않는 변량,  
 $Y$ 의 절편(상수)

## 8. 회귀분석(regression analysis)



### 나. 다중회귀분석

#### ✓ 회귀방정식

$$Y_i = \alpha X_1 + bX_2 + cX_3 + \dots + zX_i + C$$

$Y_i$ : 종속변수 //  $X_i$ : 독립변수

$\alpha \dots z$ : 종속변수  $Y$ 에 대한 독립변수  $X_i$ 의 기울기(회귀계수)

$C$ : 독립변수  $X$ 에 의해 설명이 되지 않는 변량,  $Y$ 의 절편(상수)