

# An illustrated guide to medical statistics using SPSS

그림으로 이해하는 닥터 배의 술술 보건의학통계

01

# Statistics, 통계, 統計

- 어원: Status (국가적 상태), 정치가(Stateman)들이 국가와 관련된 상황들을 수학적으로 기술하는 정치학적 수단으로서 역할
- 統 거느릴 통 計 셀 계: 한데 몰아서 셈함

# Part 1 보건의학통계 시작하기

## 01 이 여덟 가지만은 꼭 알고 통계를 시작하자

가. 자료의 분류

나. 가설을 검정하는 방법

다. 5% 유의수준과 다중비교

라. 자유도

마. 확률분포와 검정통계량

바. 중심극한정리

사. 모수적 방법과 비모수적 방법

아. 자료의 탐색과 정규성 검정

# 참고자료

- 그림으로 이해하는 닥터 배의 술술 보건의학통계 배정민 2012
- 한눈에 쓱쓱 의학통계 배우기 김지형 2014
- 의학통계 ABC 김수녕 2014

# 자료 처리 Data Management

- 자료 수집

1. 연구목적에 맞는 **대상자(표본)**를 선정
2. 연구하려는 **변수(variable)**들을 정함
3. 데이터베이스 표 작성
4. 설문지, 의무기록, 실험자료에서 얻은 자료입력
5. 입력된 자료의 오류를 점검하고 표를 완성함.

- 독립변수 Independent variable // 원인
  - 설명변수 Explanatory variable
  - 예측변수 Predictor variable
  - 위험인자 Risk factor
  - 공변량 Covariate – 회귀분석에서 연속형 자료
  - 요인 Factor – 범주형 자료
- 종속변수 dependent variable // 결과
  - 반응변수 Response variable
  - 결과변수 Outcome variable
  - 표적변수 Target variable

## 변수의 종류 Variable Type

- (1) 문자형 (character, string, text)
  - 범주형(categorical)로 분류되는 명목형(nominal) 데이터로 구성
  - 비연속적인 질적(qualitative) 변수
- (2) 숫자형 (numeric)
  - 연산처리가 필요한 연속적인 양적(quantitative) 변수

# 변수의 종류

## 자료의 특성에 의한 구분

이산형 변수(discrete variable)

= 범주형 변수(categorical variable)

명목변수(nominal variable)

순위변수(ordinal variable)

연속형 변수(continuous variable)

간격변수(interval variable)

비율변수(ratio variable)

- 남/여, 수술군/비수술군
- 성적(수우미양가), 만족도(만족,보통,불만족)
- Grade 1, 2, 3, 4 / Stage 1, 2, 3, 4
- 온도
- 키, 몸무게, 나이



## 범주형 자료 Categorical data

- (1) 명목척도 (nominal scale)
  - 서열(순서)에 관계없이 항목의 이름으로만 의미있음
- (2) 순서/순위 자료(ordinal scale)
  - 명목자료를 크기의 순서대로 나열한 자료
  - 숫자형, 문자형 변수 가능

## 가. 자료의 분류 (17쪽)

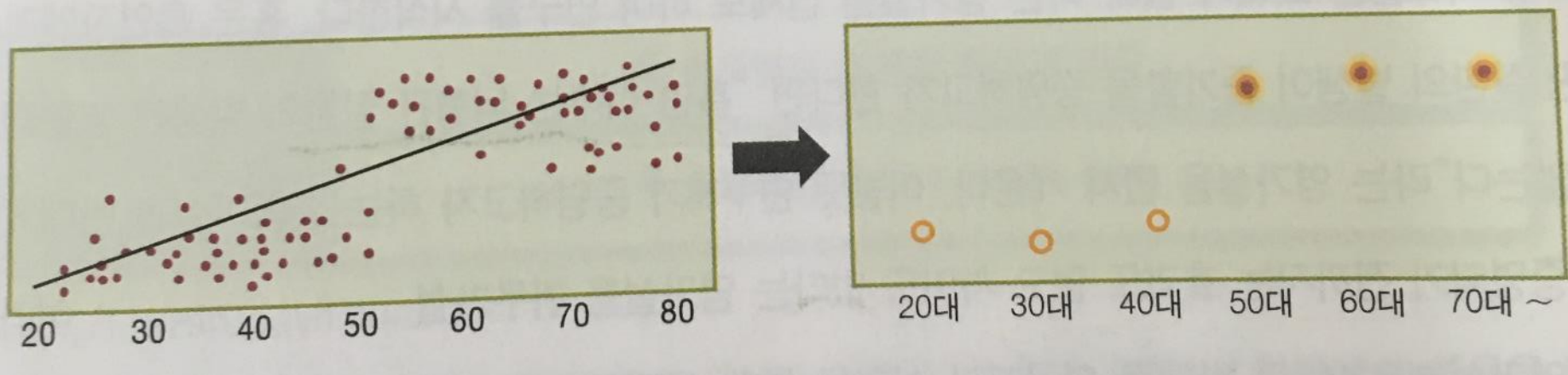
# 연속형 자료 (Numerical data) (숫자형)

- (1) 간격 척도 (interval scale) (구간)(등간자료)
  - 구간 사이의 간격이 동일
  - 0 – 임의로 정한 수치
  - 온도 / 설문조사에서 우울증 정도(0-40점)
  - 더하기 빼기는 의미가 있으나, 비율을 나타내는 곱하기, 나누기 계산은 의미가 없다.
- (2) 비율/비 척도 (ratio scale)
  - 간격 척도의 특징을 가지며, 측정하고자 하는 속성이 없는 절대 영점이 존재
  - 키, 체중, 혈압, 콜레스테롤 수치

## 가. 자료의 분류 (17쪽)

# 자료의 성격은 변할 수 있다.

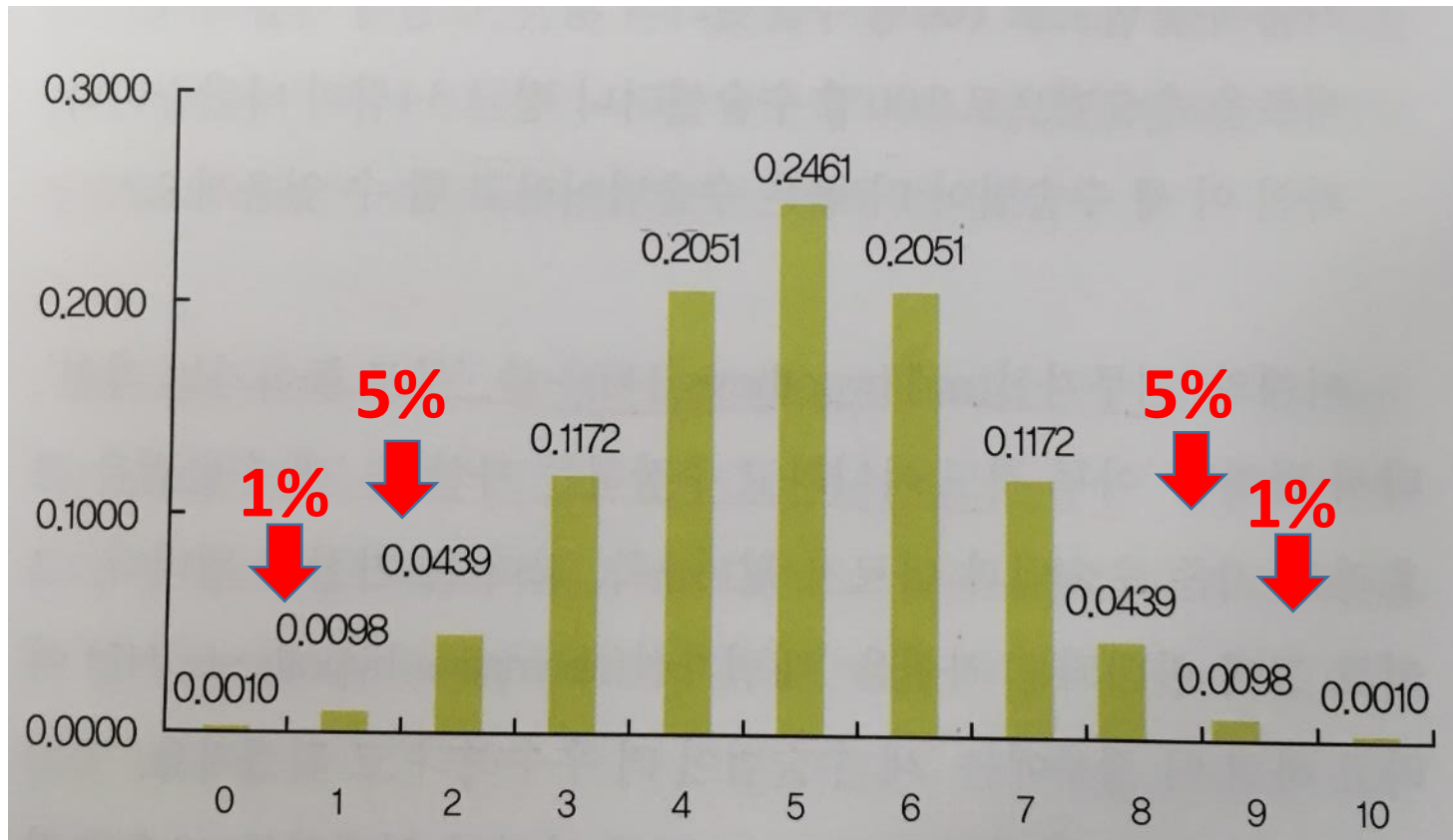
- 나이 = 연속형 자료(비 척도)
  - ➔ 10대, 20대, 30대 등 순위척도
  - ➔ 미성년/성년 명목척도
- 자료 변환 Data Transformation



## 다. 5% 유의수준과 다중비교

- 전체 인구집단에서 표본을 추출하여 수행하는 임상연구의 특성상 표본에서의 결과가 그대로 전체 인구집단으로 확대 적용될 수 있는 것은 아니며, 모든 통계검정에는 오류가 있을 수 있다.
- 제 1종 오류,  $\alpha$ : 실제로는 효과가 없는데 효과가 있다고 나오는 오류 (확률 5% 미만으로 지킬 것)
- 제 2종 오류,  $\beta$ : 실제로 효과가 있는데도 실험결과 효과를 증명하지 못하는 경우

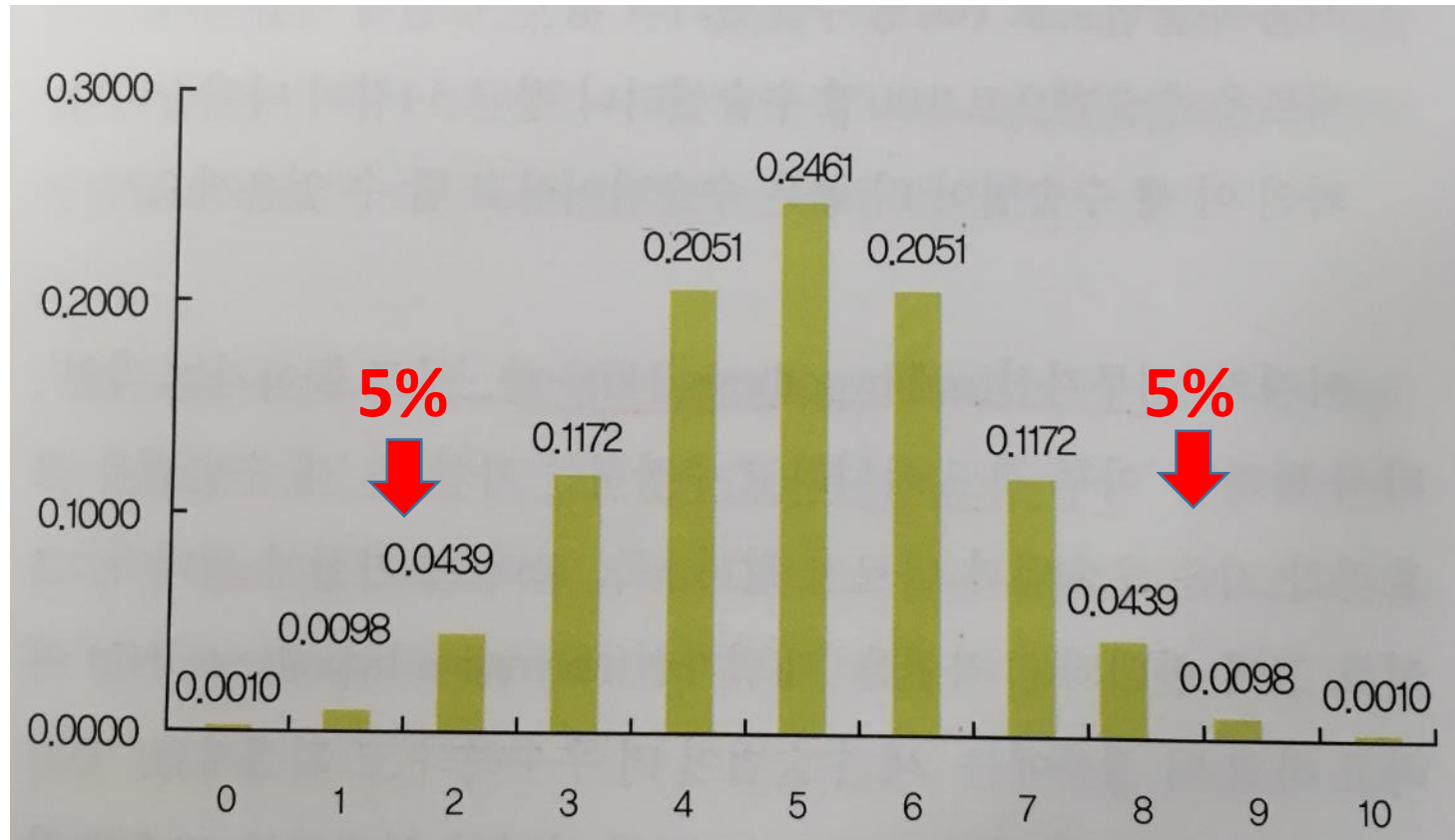
# 동전을 10번 던져 앞면이 나올 확률



유의수준: Level of Significance – 유의하다고 판단하는 기준

신뢰구간: Confidence interval: CI - 5% 유의수준 – 95% 평범한 경우의 기준

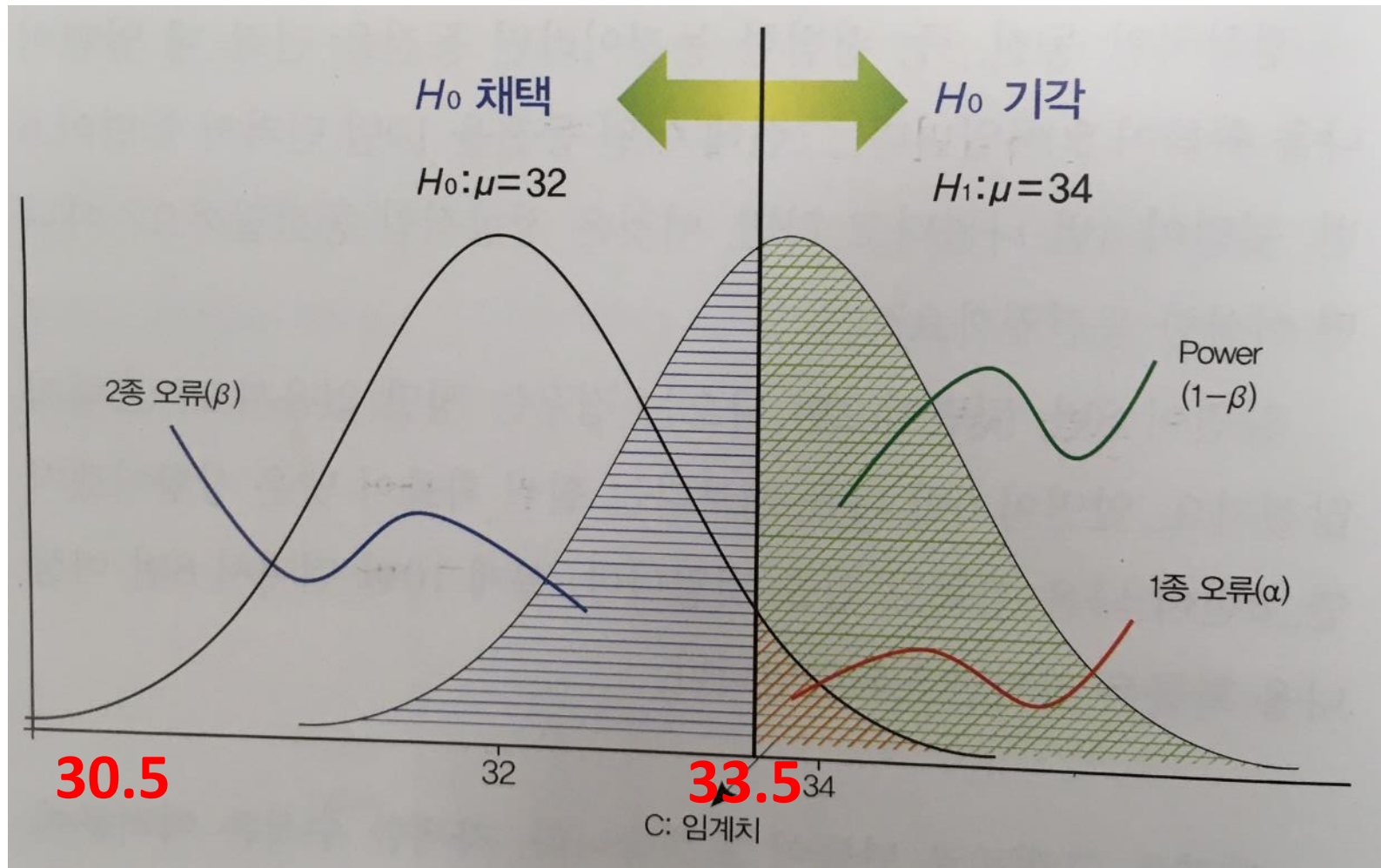
# 동전을 10번 던져 앞면이 나올 확률



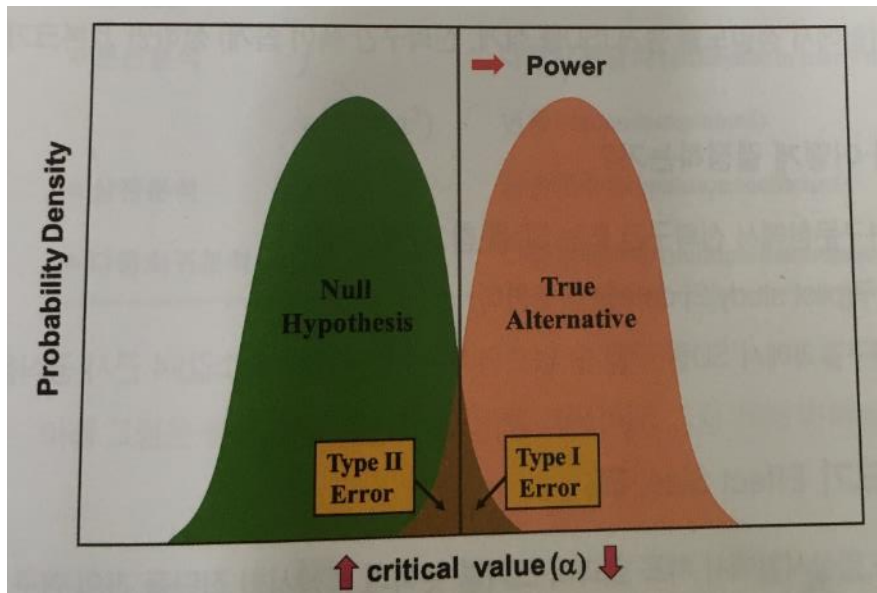
95% 선을 벗어나더라도 사실은 정상적인 동전일 확률은 5% 있다.  
5%의 오류를 감수함 - 1종 오류 ( $\alpha - error$ )

기존수술법  
평균 32

새로운수술법  
평균 34



# 검정력 Power



- $1 - \beta = 0.8$  (80%)
- 임상적으로 유의한 효과차이가 있을 때에 통계적으로 유의하다고 판정할 확률이 80%

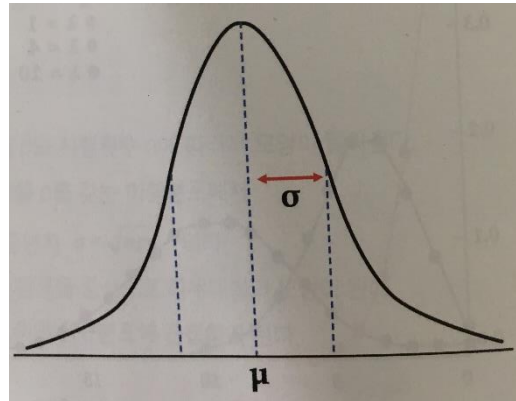
- 표본크기가 커질수록 I 형 오류와 II 형 오류가 작아지며 검정력이 높아진다.



## 나. 가설을 검정하는 방법 (20쪽)

### 귀무가설 null hypothesis, $H_0$

- 보건의학 연구 – “다른 상황을 생각하게 하는 현저한 근거가 없는 한 현상적인 모든 차이는 “무 0” 이다” 라는 과학철학적 명제로 부터 시작
- 새로운 치료법 =>
- 흡연자에서 뇌혈관 질환의 관계 =>



## 나. 가설을 검정하는 방법 (20쪽)

### 대립가설, alternative hypothesis, H1

- 연구가가 실험을 통해 규명하고자 하는 가설
  - 통계적으로 분석하여  $p < 0.05$  라는 결과를 얻은 경우
  - 가능성은 5% 미만으로 매우 희박
  - 귀무가설이 거짓일 가능성이 매우 높음
  - 귀무가설을 기각하고 대립가설을 채택
- 
- 새로운 치료법 =>
  - 흡연자에서 뇌혈관 질환의 관계 =>

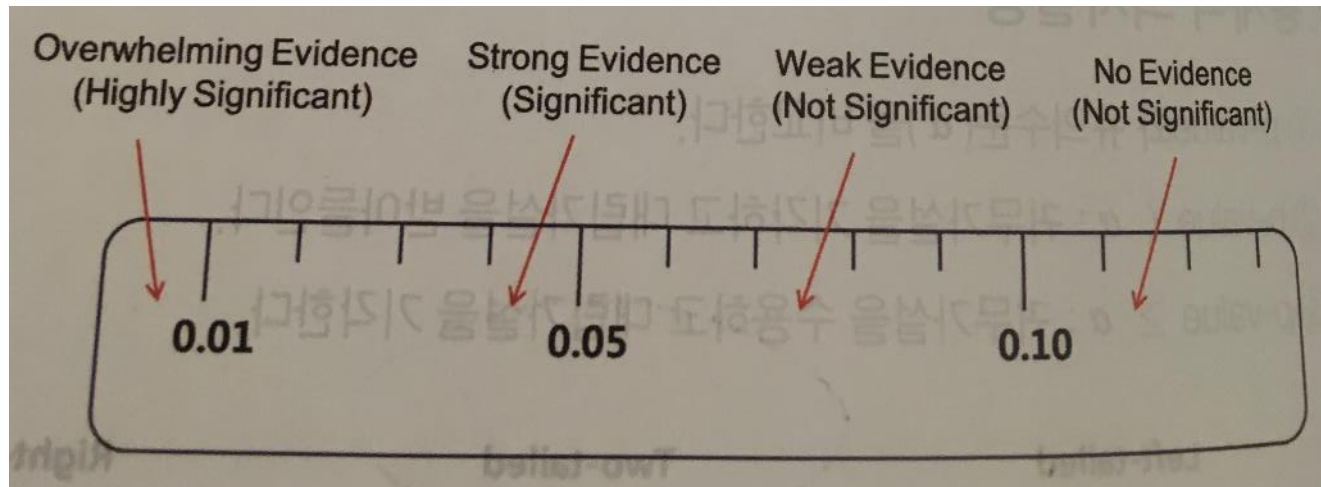
## 나. 가설을 검정하는 방법 (20쪽)

- 우위성 검정의 예
  - 기존 검사보다 효과가 더 있을 것이다.
  - 귀무가설:
  - 대립가설:
- 동등성 검정의 예
  - 기존 검사 정도로 유사하게 효과가 있을 것이다.
  - 귀무가설:
  - 대립가설:

## 다. 5% 유의수준과 다중비교

- 5% 유의수준: 1종 오류를 범할 가능성을 5% 미만으로 제한한다는 의미
- 검정력 power of test,  $1 - \beta$ : 실제 효과가 있는 것을 통계적으로 효과가 있다고 보여 줄 수 있는 힘

유의수준	통계적유의성	검정결과 해석
1%	$p < 0.01$	매우 의의있다 (highly significant)
5%	$p < 0.05$	의의있다 (significant)
10%	$p < 0.10$	의의가 별로 없다 (less significant)



### 왜 유의수준 5%를 사용할까?

통계학의 대가인 Fisher 가 1930년대에 유의수준 5%(0.05)를 사용.  
아직까지 그대로 사용해오고 있음.

다. 5% 유의수준과 다중비교

## 다중 비교의 문제

- 다중 비교의 문제

- 세가지 치료법 비교
- 5% 유의수준이 3번 사용
- 각각이 독립인 경우  $1-(1-0.05)^3 = 0.143(14.3\%)$ 의 유의수준을 적용
- 전체 유의수준이 지정된 5%를 넘지 않도록 주의해야 함

1. 세가지 치료법이 다르다 – 적어도 하나에 차이가 있다.

➔ 어떤 군 사이에 차이가 있는 지 밝혀야 한다 (사후 분석 post hoc)

2. Bonferroni, Tukey, Duncan, Scheffe method

## STEP 1

### 전반적인 검정



$H_0$

A군 ≡ B군 ≡ C군



$H_1$

A군 ≠ B군 OR B군 ≠ C군 OR C군 ≠ A군

## STEP 2 사후분석 유의수준을 보정하여 추가로 세 번의 검정을 시행

두 군의 크기 비교

$H_0$

A군 ≡ B군

$H_1$

A군 ≠ B군

두 군의 크기 비교

$H_0$

B군 ≡ C군

$H_1$

B군 ≠ C군

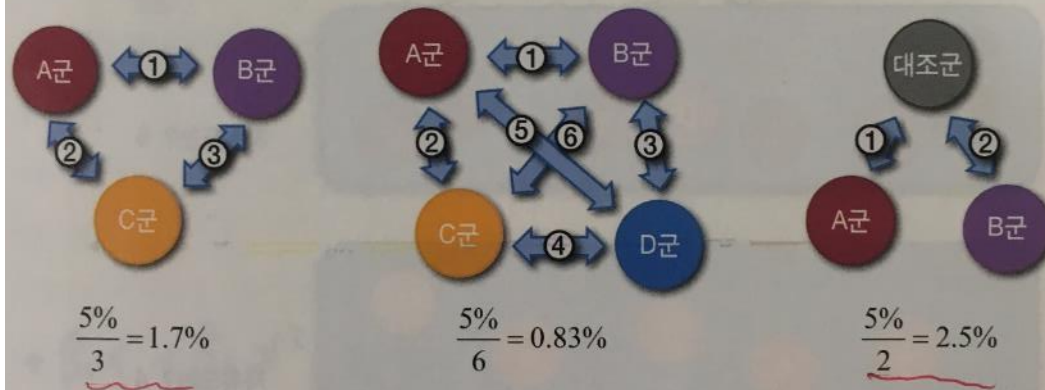
두 군의 크기 비교

$H_0$

C군 ≡ A군

$H_1$

C군 ≠ A군



## 라. 자유도

- 실질적으로 독립인 값들의 개수를 의미
- 평균이  $\mu$  인 10개의 자료
  - 1개의 값은 나머지 9개의 값에 의해 정의되므로 자유도는 9 ( $n-1$ )
- 교차표
  - $2 \times 2$  : 1개의 값만 정해지면 나머지 값은 정해짐 (자유도1)
  - $n \times m$  ( $n-1$ )  $\times$  ( $m-1$ )의 자유도를 가짐

	당뇨	정상	전체
고혈압	$a$	$b$	20
정상	$c$	$d$	80
전체	25	75	100



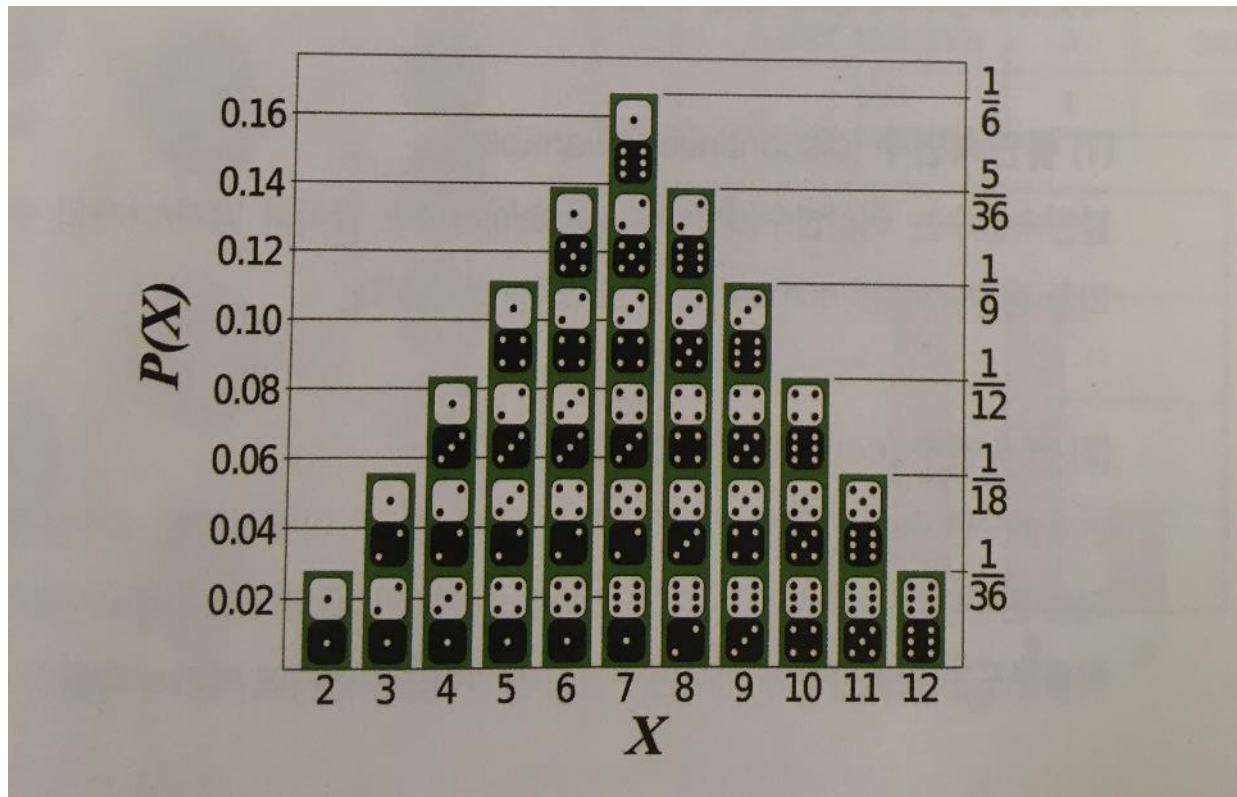
마. 분포와 검정통계량

## 확률분포, probability distribution

- 어떤 집단에서 개개의 관측값을 측정하거나 혹은 동일한 현상에 대해 반복적으로 관측된 결과를 집합적으로 살펴보면 특정한 분포를 이루는 경향이 있다.
- 정규분포,  $t$  분포,  $F$  분포, 카이제곱분포

# 확률분포

- 모든 가능한 결과와 각각의 결과가 일어날 빈도를 기술하여 도수분포로 나타낸 것



# 확률분포의 유형

## 1. 불연속확률분포 – 정수로만 표시되는 변수

- 이항분포 (binomial distribution)
- 포아송분포 (Poisson distribution)

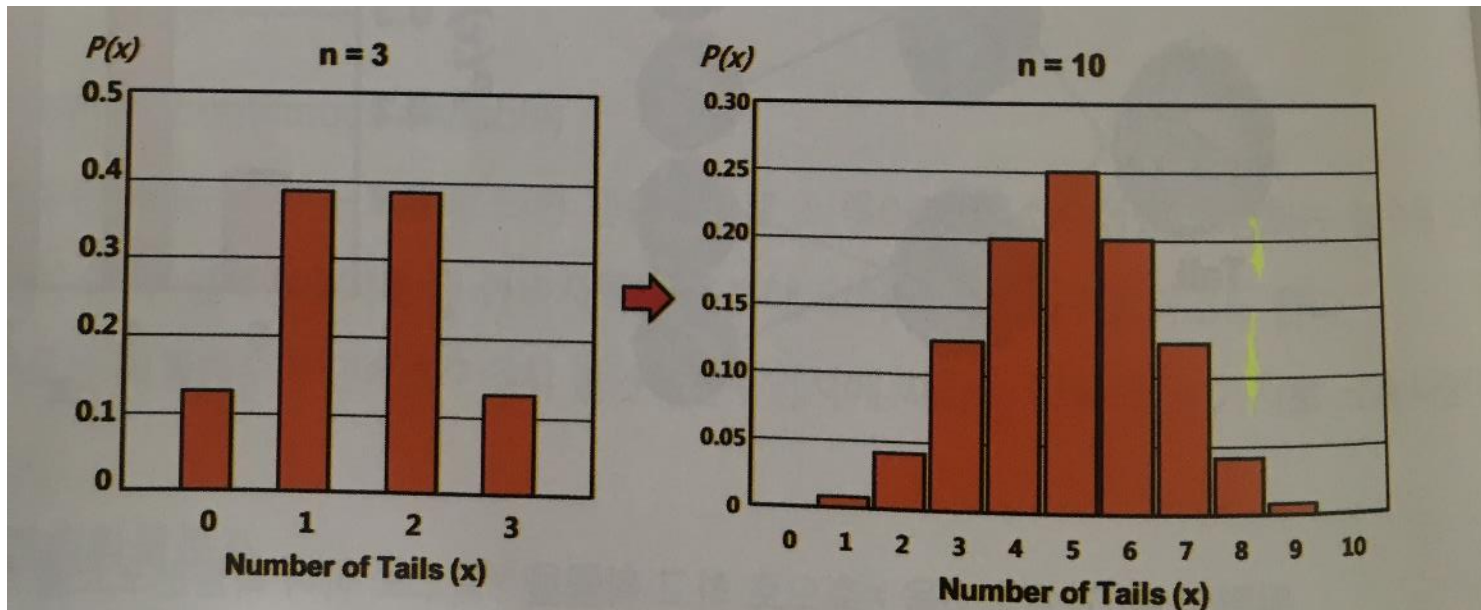
## 2. 연속확률분포 – 측정에 의한 관측값, 실수로 표시

- 정규분포 (Normal distribution)

# 1. 불연속확률분포

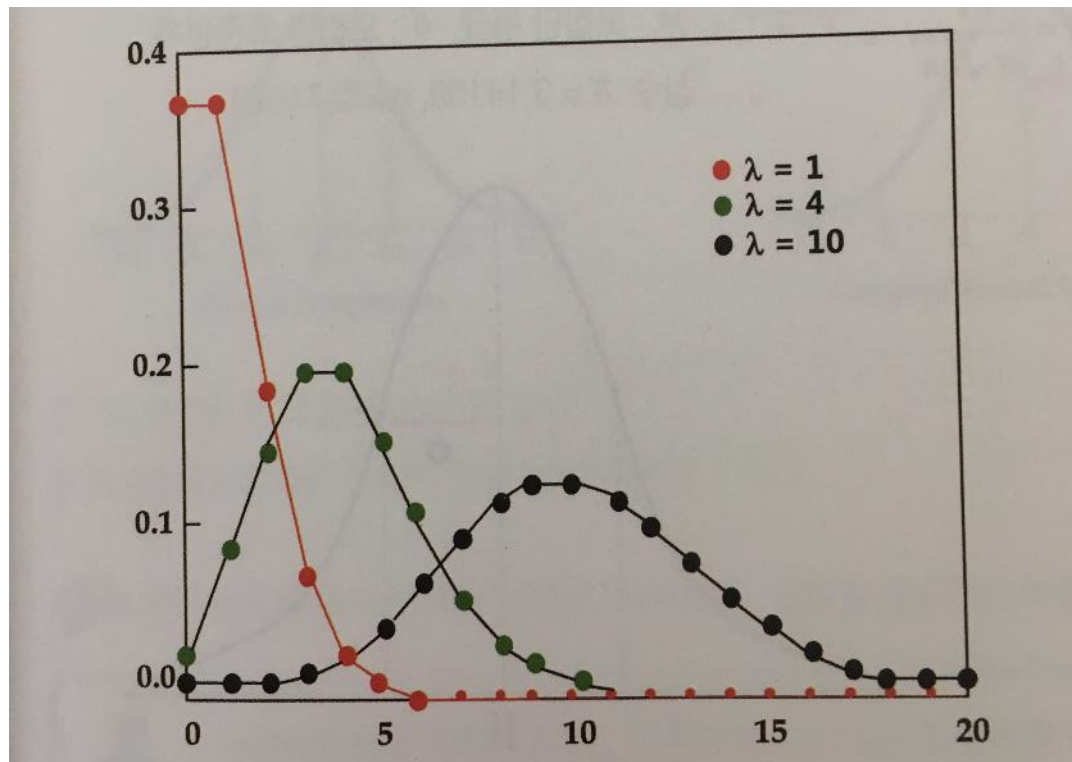
## 1) 이항분포 binomial distribution (동전 던지기)

- 베르누이 시행(Bernoulli trial) – 실험결과가 “성공-실패” 두가지만 가능한 경우
- 성공률  $p$ 와 시행횟수  $n$ 에 따라서 모양이 달라진다.
- $p=0.5$  인 경우는 평균을 중심으로 좌우대칭의 모양이 된다.



## 2) 포아송분포 Poisson distribution

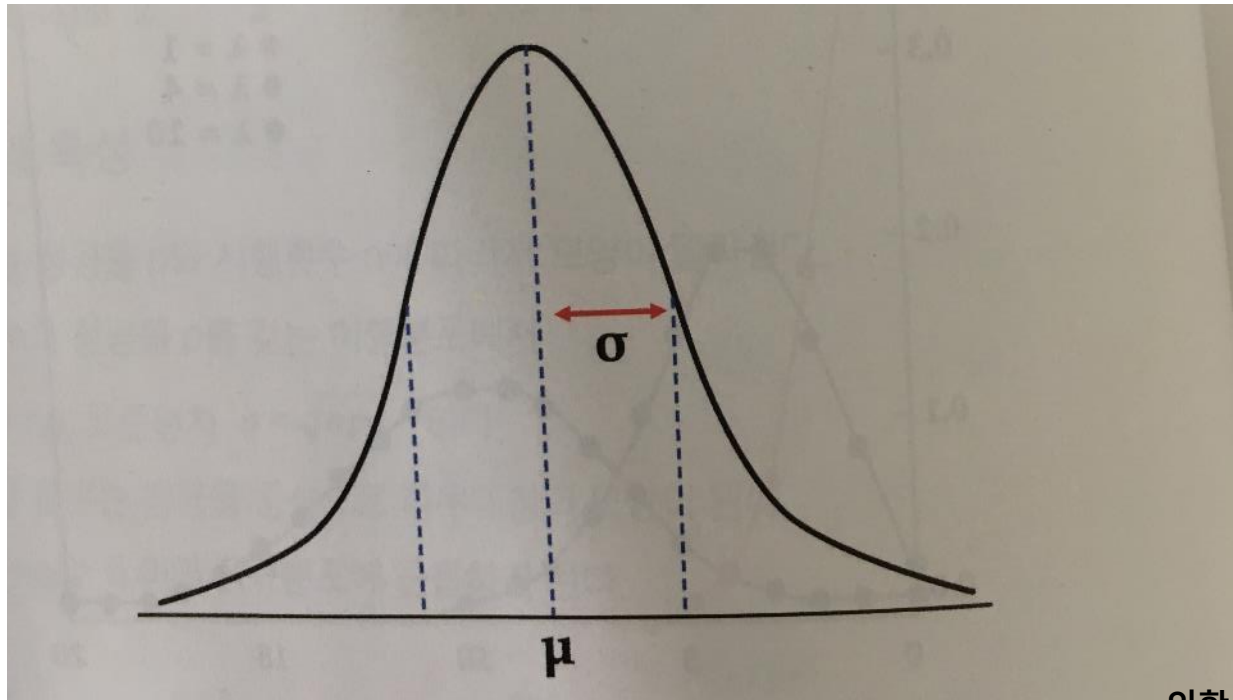
- 시간 또는 공간 단위 당 사건의 발생 횟수 ( $\lambda$ )가 나타내는 분포
- 매우 드물게 발생하는 사건에 대한 확률분포
- 이항분포에서  $n$ 은 매우 큰 반면에  $p$ 는 매우 작은 경우에 나타나는 분포의 모양을 갖는다.



## 2. 연속확률분포

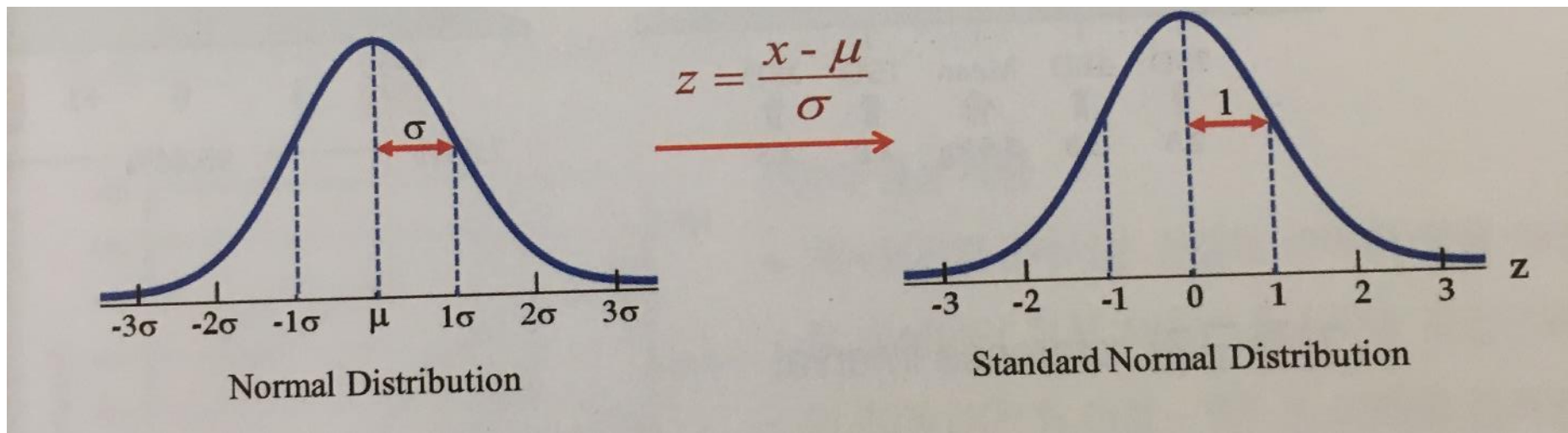
### 1) 정규분포 Normal distribution

- 모집단을 근거로 함. 평균  $\mu$  분산  $\sigma^2$
- 정규분포의 곡선 모양은 평균 (또는 표준편차  $\sigma$ )에 따라 변한다.
- 평균을 중심으로 좌우대칭인 종 모양의 분포곡선을 나타낸다.



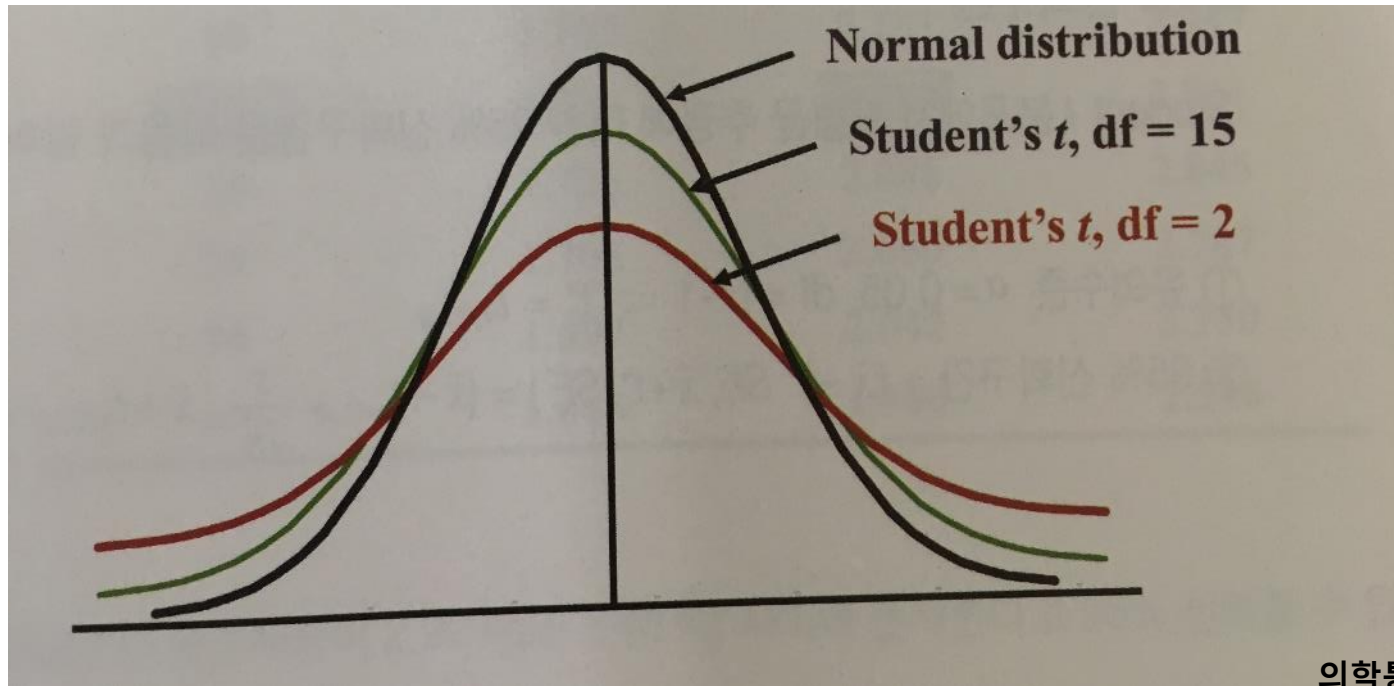
## 2) 표준정규분포 standard Normal distribution

- 관측값( $x$ )를 변형하여 모집단의 평균값과 표준편차에 대해 상대적인 값(표준점수,  $z$ )으로 만든 후  $z$  값에 대한 분포를 구함
- $z$ 의 평균은 0, 표준편차는 1
- 전체 면적(확률)은 1
- $z$ -분포( $z$ -distribution)라고도 한다.



### 3) Student t 분포 Student's t test

- 표준정규분포는 오직 하나만 존재하지만 t 분포는 자유도에 따라서 산포도가 다른 여러 개 분포가 존재함.
- T-분포에서 자유도는 표본크기-1 ( $n-1$ )이다.
- 표본크기 30 미만인 소표본들의 표본평균분포는 Student t 분포를 따른다.





- 정규분포 Normal distribution
  - 하나의 꼭지를 가진 완벽한 좌우대칭
  - 평균( $\mu$ )와 표준편차( $\sigma$ )에 의해 결정
  - 평균과 표준편차만 같다면 동일한 정규분포
- 표준정규분포 Standard Normal distribution
  - 관측치( $x$ )를  $z=(x-\mu / \sigma)$ 로 치환하면 평균이 0이고 표준편차가 1인 정규분포
  - z- 분포
  - 어떤 집단이 정규분포를 따른다고 가정하고 그 평균을 비교하는 것: z- 검정

# T 분포

- 표본수( $n$ )가 충분히 클 때 표본평균들의 분포는 정규분포를 보인다고 알려져 있지만, 실제로 표본수가 정규분포를 이룰 만큼 충분치 않을 때에는  $t$ -분포를 더 근사적으로 따른다
- 좌우대칭의 종모양, 좀더 긴꼬리를 가진 형태로 넓게 퍼져있다.
- 자유도의 따라 달라지며, 자유도가 증가함에 따라 표준정규분포의 형태에 근접
- 모집단의 모주를 모르는 상태에서 모집단의 표준편차 대신에 표본의 표준편차를 사용하고, 이때 표본 평균의 분포가  $t$ -분포를 이룬다는 가정하에 분석
- 이  $t$  분포를 이용하여 독립인 두 군의 평균을 비교하는 방법 : 독립표본  $t$  검정, Student's  $t$ -test

- 카이제곱 분포

- 표준정규분포를 따르는 독립적인 확률변수를 제공하여 합해서 얻어지는 분포
- 자유도에 따라 달라지는데 자유도가 커짐에 따라 오른쪽으로 길게 늘어진 모양이 점차 좌우대칭의 모양에 가까워 짐

- F-분포

- 비율에 의해 정의되는 분포로 왼쪽으로 치우친 형태
- 2개의 모집단 각각의 분산 추정치의 비(변동의 평균의 비)로 정의됨
- 분자와 분모 각각 자유도에 따라 형태가 결정
- Analysis of variance ANOVA 분산분석에 사용

- 분포와 검정통계량

- 가설검정 시에는 관찰된 표본으로 부터 가설검정을 위한 검정통계량 test statistic을 계산
- 귀무가설이 옳다는 전제하에, 가정된 분포에서 관찰된 표본이 발견될 수 있는 확률 probability value, p value은 아래 면적을 통해 알 수 있다.

## 바. 중심극한정리

# 중심극한정리 Central limit theorem

- 정규분포인 모집단으로부터 무작위로 표본을 추출할 때 표본의 크기가 충분히 크다면 표본 변수들의 그 합 또는 평균의 히스토그램은 정규분포 곡선에 수렴한다.

# 사. 모수적 방법과 비모수적 방법

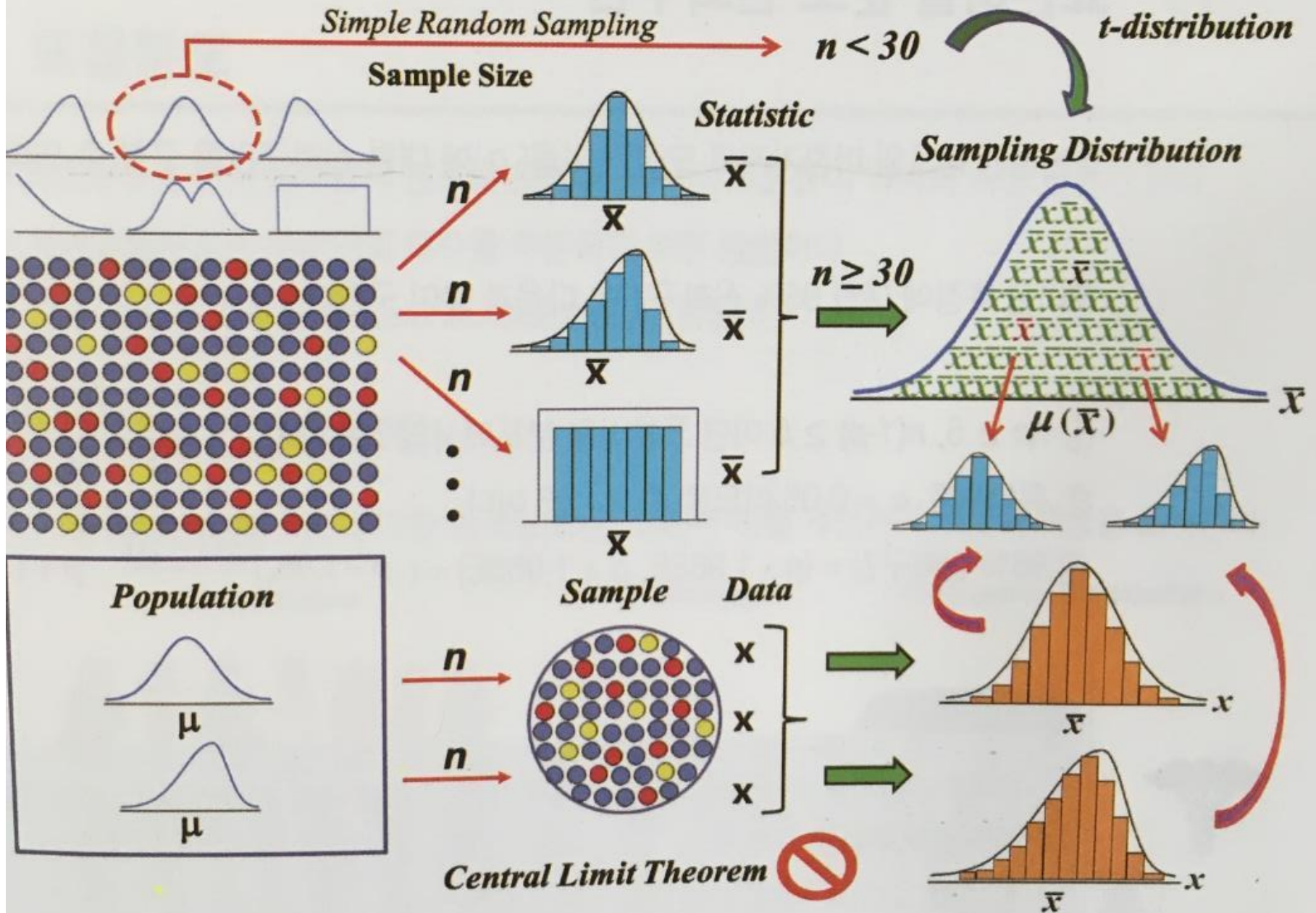
## 모수적 방법

- 통계량 Statistic
  - 추출된 표본의 평균, 표준편차, 분산
- 모수 parameter
  - 모집단의 모평균, 모표준편차, 모분산
- **중심극한정리 - 표본수가 크다면 정규분포를 따른다.**
  - 표본의 크기가 충분히 크다면 모집단의 정규성 유무와 관계 없이 모평균에 대한 가설검정을 시행할 수 있다.
  - 정규성을 갖는다는 모수적 특성을 이용하는 통계적 방법: 모수적 방법  
Parameteric method
  - **군당 30명이상으로 구성된 표본의 경우 표본평균이 정규분포를 따른다고 가정할 수 있다.**

# 표본수와 중심극한정리

- 중심극한정리는 모집단으로부터 추출가능한 모든 표본을 대상으로 한 이론적분포이다
  - ① 중심극한정리는 추정값 특히 표본평균 분포에 적용되는 추론통계의 이론적 기초이다.
  - ② 표본크기  $n \geq 30$  이면 모집단의 분포에 무관하게 z-분포(표준정규분포)를 따른다.
  - ③ 표본크기  $n < 30$  이면 모집단이 정규분포이면 t-분포를 따른다.
- 표본수 = 1(표본추출 1회)인 실제표본에서는 중심극한정리가 적용되지 않는다.
  - ① 실제 표본은 표본크기  $n \geq 30$  의 법칙이 적용되지 않는다.
  - ② 실제 표본에서는 표본크기가 클수록 모집단의 분포에 따르게 된다.

# Central Limit Theorem





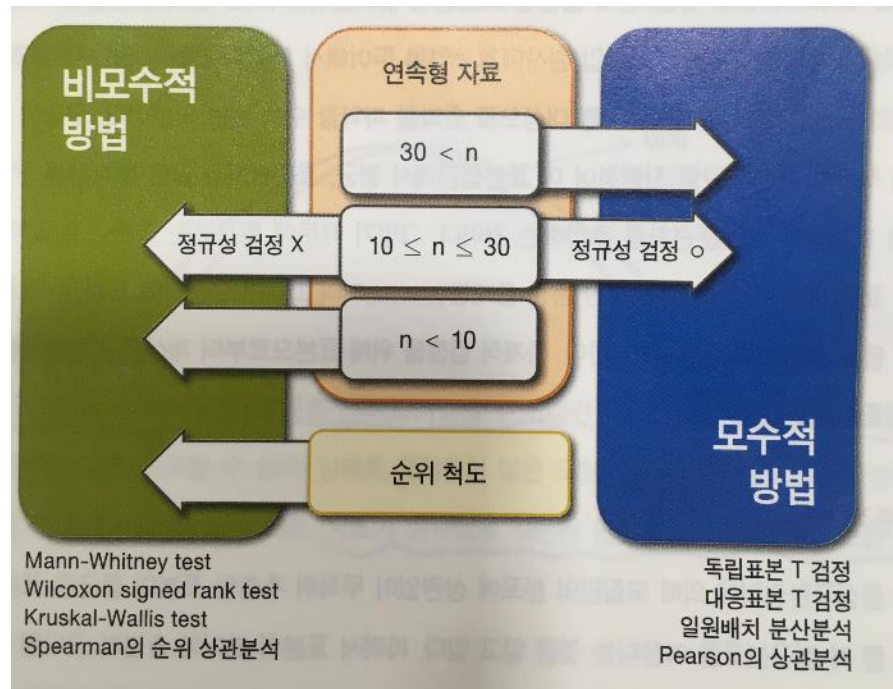
# 사. 모수적 방법과 비모수적 방법

## 모수적 방법

- 통계량 Statistic
  - 추출된 표본의 평균, 표준편차, 분산
- 모수 parameter
  - 모집단의 모평균, 모표준편차, 모분산
- 중심극한정리 - 표본수가 크다면 정규분포를 따른다.
  - 표본의 크기가 충분히 크다면 모집단의 정규성 유무와 관계 없이 모평균에 대한 가설검정을 시행할 수 있다.
  - 정규성을 갖는다는 모수적 특성을 이용하는 통계적 방법: 모수적 방법  
Parameteric method
  - 군당 30 명 이상으로 구성된 표본의 경우 표본평균이 정규분포를 따른다고 가정할 수 있다.

# 비모수적 방법 nonparametric method

- 정규성 검정에서 정규분포를 따르지 않는다고 증명되거나 군당 10명 미만의 소규모 실험에서는 정규분포임을 가정할 수 없는 경우
- 순위합검정: 자료를 크기 순으로 배열하여 순위를 매긴 다음 순위의 합을 통해 차이를 비교



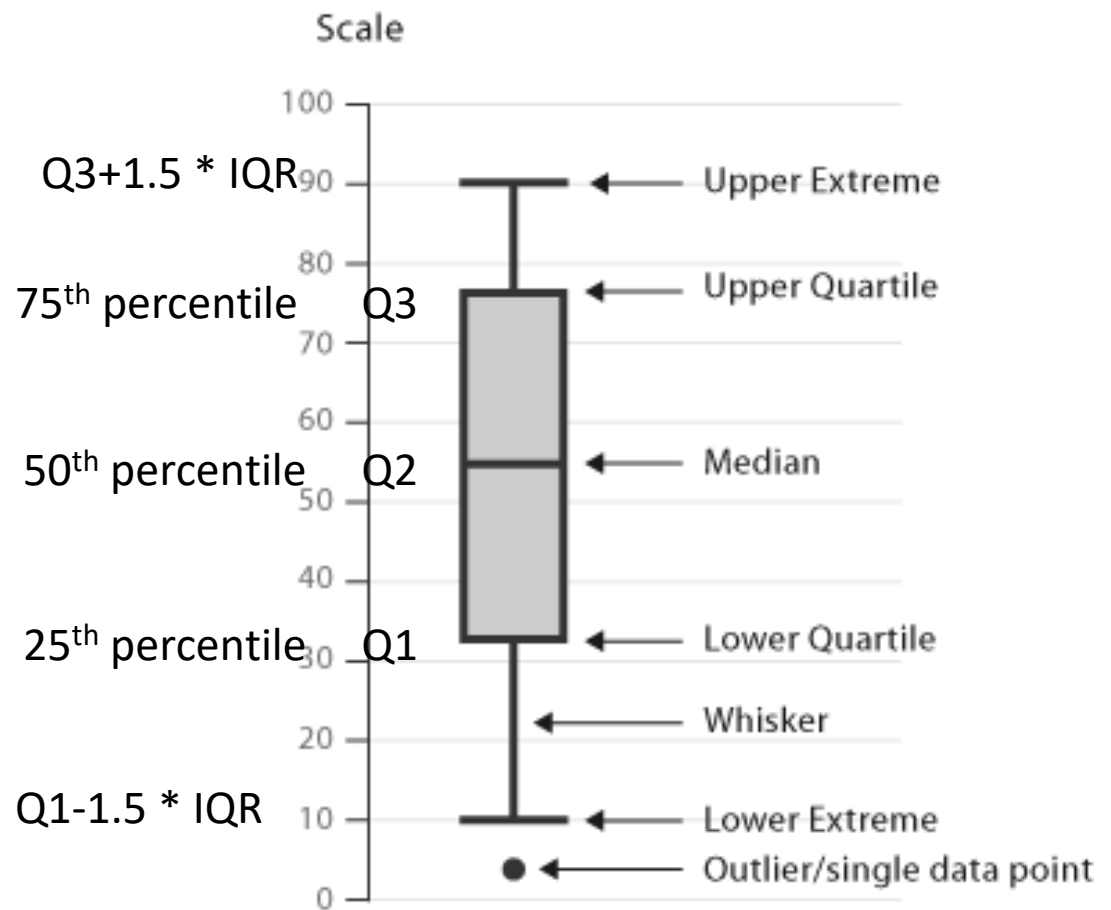
## 아. 자료의 탐색 및 정규성 검정

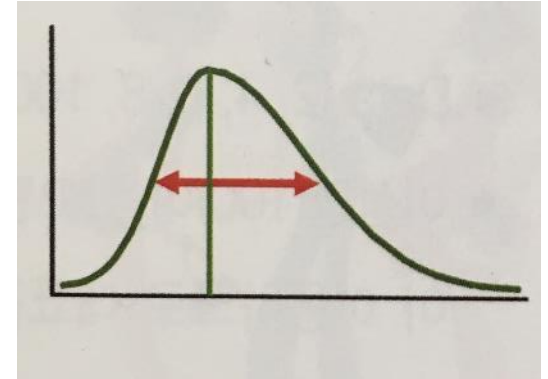
### 탐색적 자료분석 exploratory data analysis

- 자료의 특징과 내재하는 구조적 관계를 알아내기 위해 시행하는 모든 방법
- 자료의 분포가 실제로 정규분포에 근사하는지
- 두 변수가 실제로 선형의 관계를 보이는지 파악

- 대표값 representative value
  - 자료를 대표할 수 있는 하나의 값을 의미
  - 평균 mean
  - 중앙값(중위수) median
  - 최빈값 mode
- 산포도 degree of scattering
  - 자료의 흩어진 정도를 의미
  - 분산 variance: 관측치와 평균의 차이의 제곱의 평균
  - 표준편차 standard deviation: 분산의 제곱근을 구하여 원래 측정값들과 차수를 맞추는 것

- 범위 range
- 특이점 outlier
- 사분위수 quartiles
- 사분위수 범위 interquartile range
- 다섯수치요약 five number summary
  - 최소값, 최대값, 중앙값, 1사분위수와 3사분위수



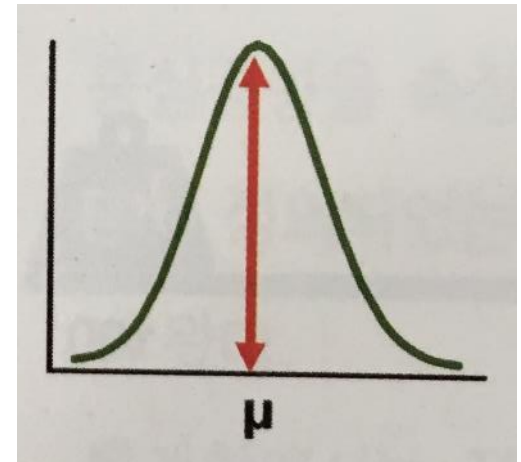


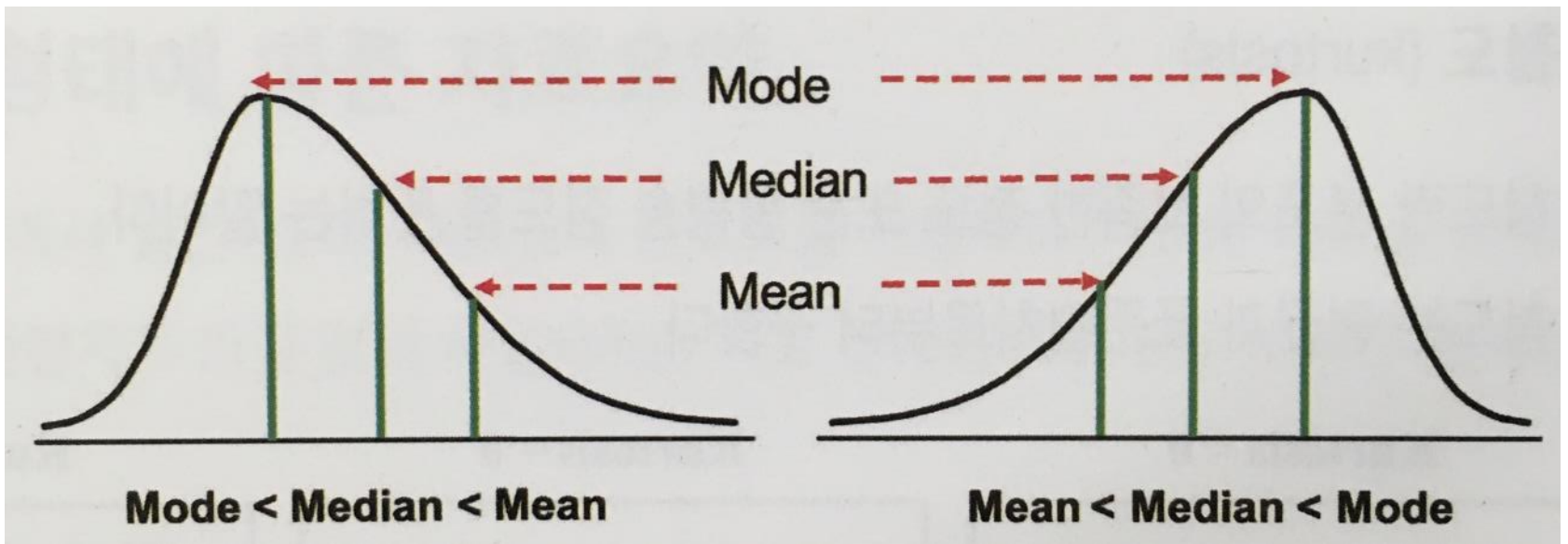
- 왜도 Skewness

- 자료의 치우친 정도
- Negative skewness ( $S < 0$ ): 왼쪽으로 긴 꼬리 (시험성적)
- Positive skewness ( $S > 0$ ): 오른쪽으로 긴 꼬리 (소득분포)
- 정규분포: 0 (-2 ~ +2 : 대칭성 만족)

- 첨도 Kurtosis

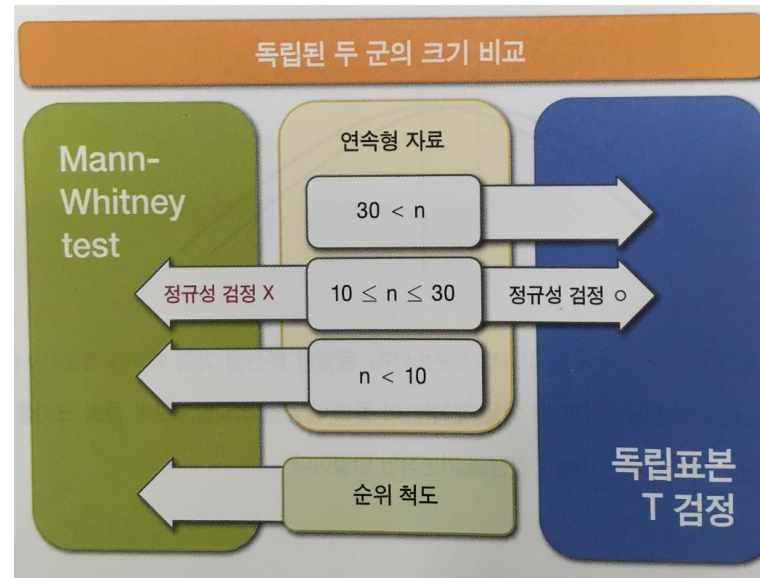
- Negative kurtosis ( $K < 0$ ) : 납작한 분포
- Positive kurtosis ( $K > 0$ ): 뾰족한 분포





# 정규성 검정

- 표본수가 30을 넘는 경우 중심극한정리에 의해 모집단의 분포에 상관없이 표본평균이 정규성을 갖는다고 가정할 수 있다
- 표본수가 10~30인 경우 정규성 검정을 통해 정규성을 검정
- 표본수가 10미만: 비모수적인 통계방법사용
- Kolmogorov-Smirnov test
- Shapiro-Wilk test
- $P > 0.05$  귀무가설 채택, 정규분포를 따른다





## 표본크기가 30 이상이면 정규분포 검정이 필요하지 않는가?

- 중심극한정리에 의하면 표본크기  $\geq 30$  인 표본들은 모집단으로부터 가능한 모든 표본을 추출하면 **표본평균 분포**는 모집단 분포의 모양에 무관하게 정규분포를 이룬다.
- 실제연구에서 사용하는 자료는 모집단으로부터 1회 표본추출한 단일표본이고 표본 자료에 대한 빈도분포는 표본평균 분포와는 완전히 다르다.
- 표본크기가 모집단의 크기와 같다고 가정하면 표본자료의 분포는 모집단 분포가 된다. 따라서 단일표본에서 모집단이 비정규분포이면 표본도 비정규분포를 이룬다.

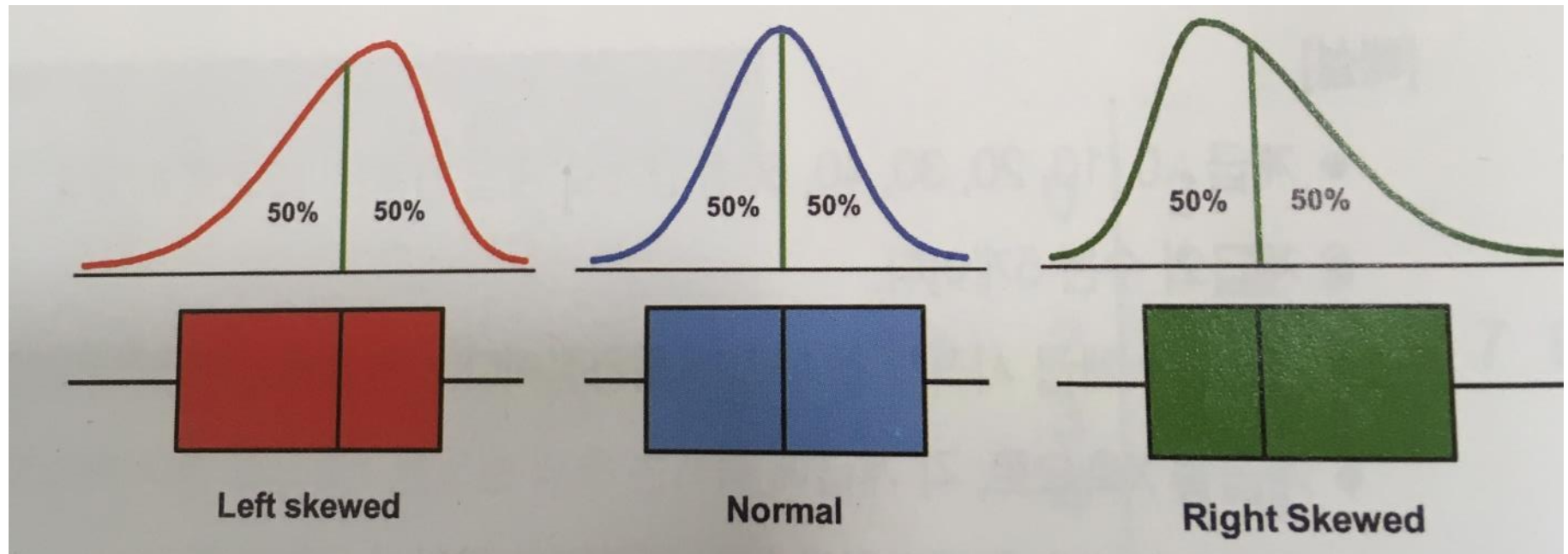
# 정규분포 검정 Normality test

## 1. 그래프를 이용하는 방법 Graphic method

- 1) Histogram
- 2) Box plot
- 3) Normal Q-Q plot

## 2. 통계학적 방법 Statistical method

- 1) 95% 범위 (평균 $\pm$ 2 표준편차)
- 2) 왜도와 첨도
- 3) 이상값 검사
- 4) 통계학적 방법
  - ① Kolmogorov-Smirnov test
  - ② Lilliefors test
  - ③ Shapiro-Wilk test



# 정규분포 검정 Normality test

## 1. 그래프를 이용하는 방법 Graphic method

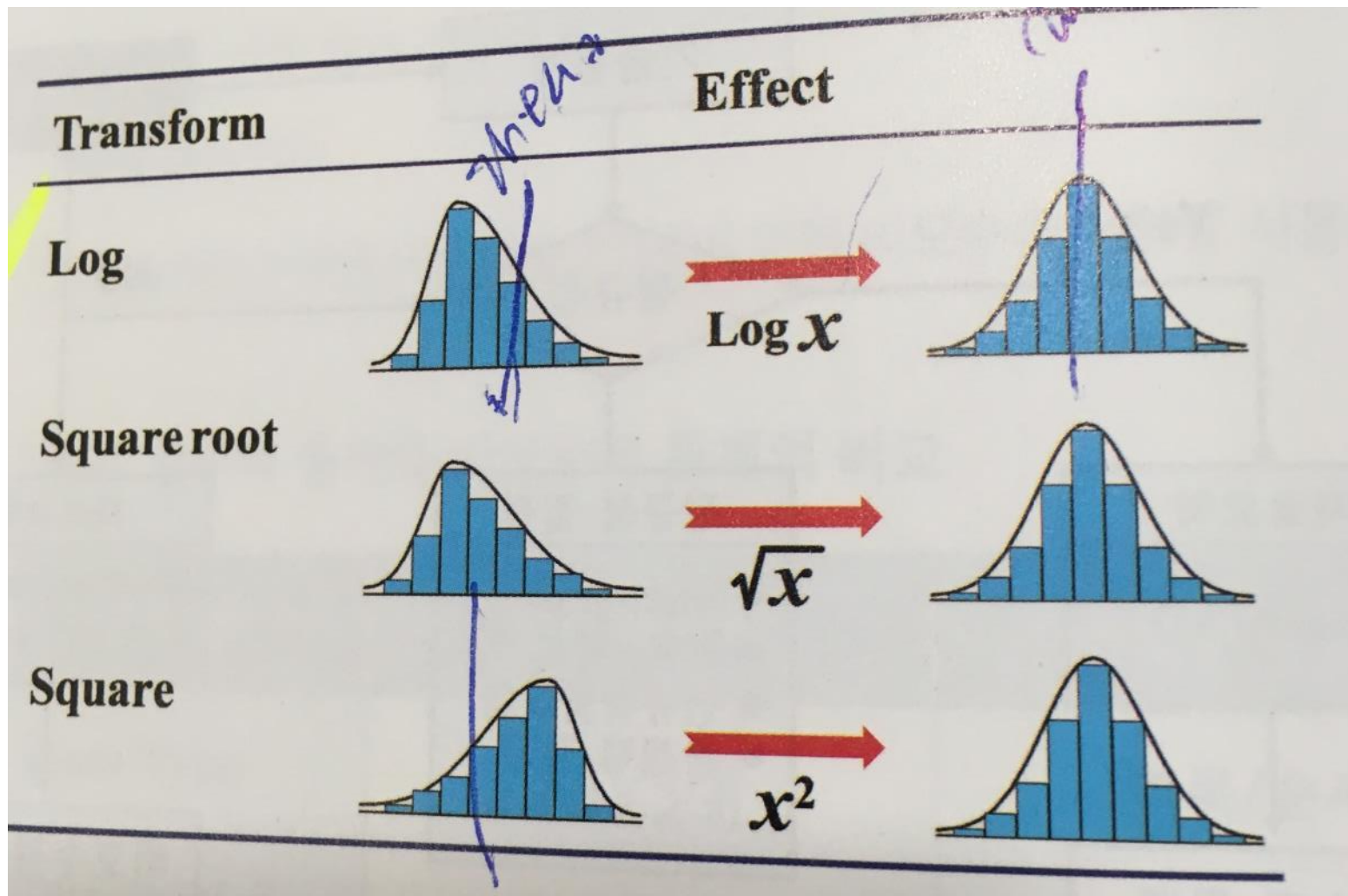
- 1) Histogram
- 2) Box plot
- 3) Normal Q-Q plot

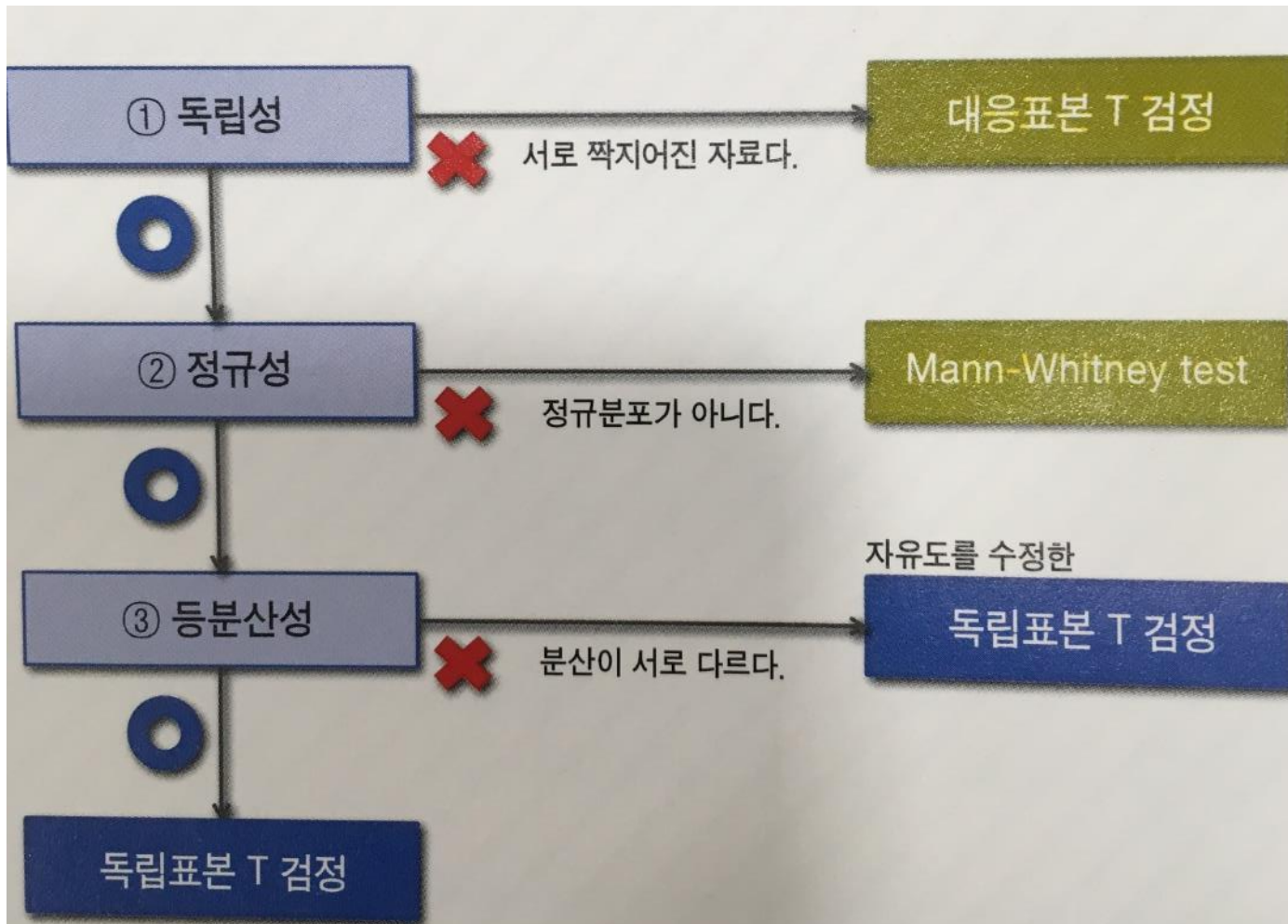
**$40.5 \pm 12.4 \text{ kg}$**

## 2. 통계학적 방법 Statistical method

- 1) 95% 범위 (**평균  $\pm 2$  표준편차**)
- 2) 왜도와 첨도 (**절대값이 3보다 큰 경우**)
- 3) 이상값 검사
- 4) **통계학적 방법**
  - ① Kolmogorov-Smirnov test
  - ② Lilliefors test
  - ③ Shapiro-Wilk test

**$40.5 \pm 39.2 \text{ kg}$**





## 02 두 군의 크기 비교

# 독립표본 t 검정의 가정

- 결과변수가 연속형인 독립된 두 군의 크기를 비교할 경우 사용
- 독립표본 t 검정의 가정
  - 독립성
  - 정규성
  - 등분산성
- 독립성,  $X$ : 짝지어진 자료  $\rightarrow$  대응표본 t 검정
- 정규성,  $X$ : 정규분포  $X \rightarrow$  Mann-Whitney test
- 등분산성,  $X$ : 분산이 다르다  $\rightarrow$  자유도 수정 후
- 3가지 모두 만족  $\rightarrow$  독립표본 t 검정



- 정규성 검정
  - Kolmogorov-Smirnov test
  - Shapiro-Wilk test
- Levene의 등분산 검정

# Mann-Whitney test

- 정규분포를 따른다고 가정할 수 없는 두 군은 평균을 통해 크기의 차이를 비교할 수 없다.
- 모집단의 분포를 가정하지 않는 비모수적인 방법이 사용
- Wilcoxon rank sum test
- 독립표본  $t$  검정에 비해 검정력이 낮다.
- 순위만 비교한 것이기 때문에 두군의 평균의 차이를 언급할 수 없다.