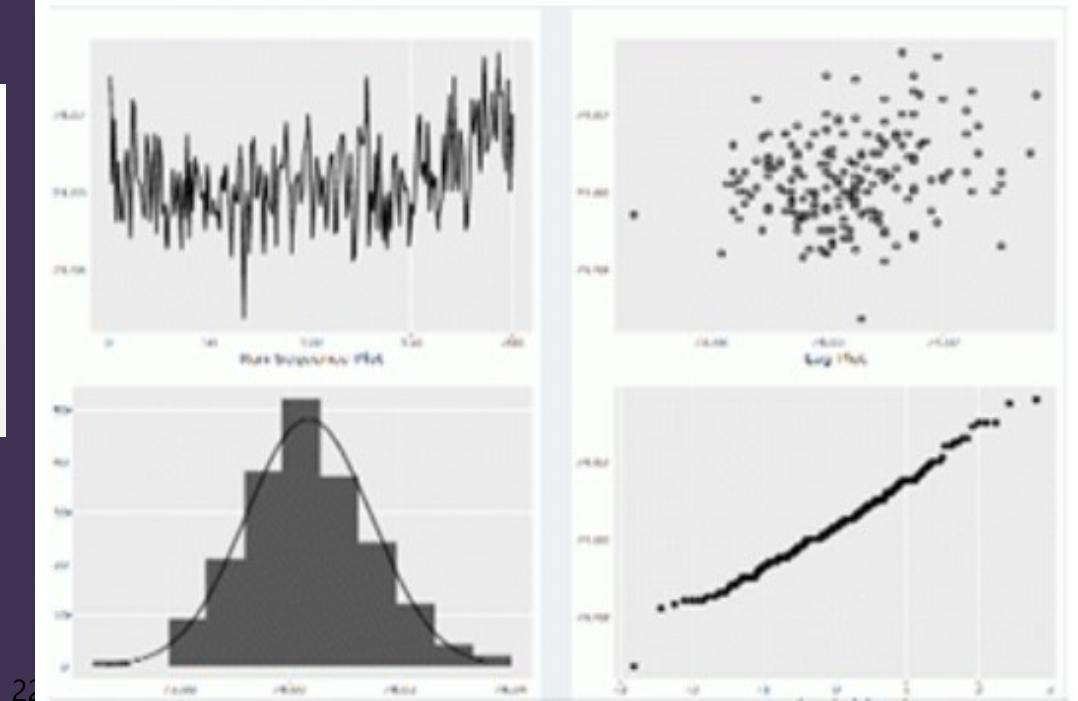


통계의 이해

학습목표

- ▶ 데이터의 특성인 기술통계량 이해
- ▶ 데이터의 특성을 고려한 확률분포 이해
- ▶ 통계적 가설검정의 이해



통계(統計)



- : 한 곳에 몰아서 어림잡아 계산함...
현상을 보기 쉽게 일정체계에 의해 숫자로 나타냄
- : 수량적인 비교를 기초로 많은 사실을 다양한 방법으로
관찰, 처리한다

통계학(統計學)



인문. 사회 및 인간 생활의 온갖 현상을 연구하기 위하여,
불확실성이 내포된 데이터(data)의 선택, 관찰, 분석 및
추정을 통하여 의사결정에 필요한 정보의 획득과 처리
방법을 연구하는 학문

■ 통계의 기원

- : 유럽에서 라틴어를 사용하는 국가에서 기원했으며, 국가를 기술하는 것과 밀접한 관계가 있음.
 - 그리스와 로마시대에는 국가(State)의 상태(state)를 살피는 것에 주로 관심을 가졌는데 이것을 'statistics'라 부름.
 - 통계는 주로 정치적인 필요에서 인구와 종교, 산업에 대한 많은 정보를 수집하는 형태로 시작됨.

통계(Statistics)의 구분

기술(Descriptive, 記述) 통계

- 수집된 데이터로부터 평균, 분산 등의 요약 통계량이나 그래프를 이용하여 체계적으로 정리/요약하여 전반적 특성을 파악하는 통계 기술

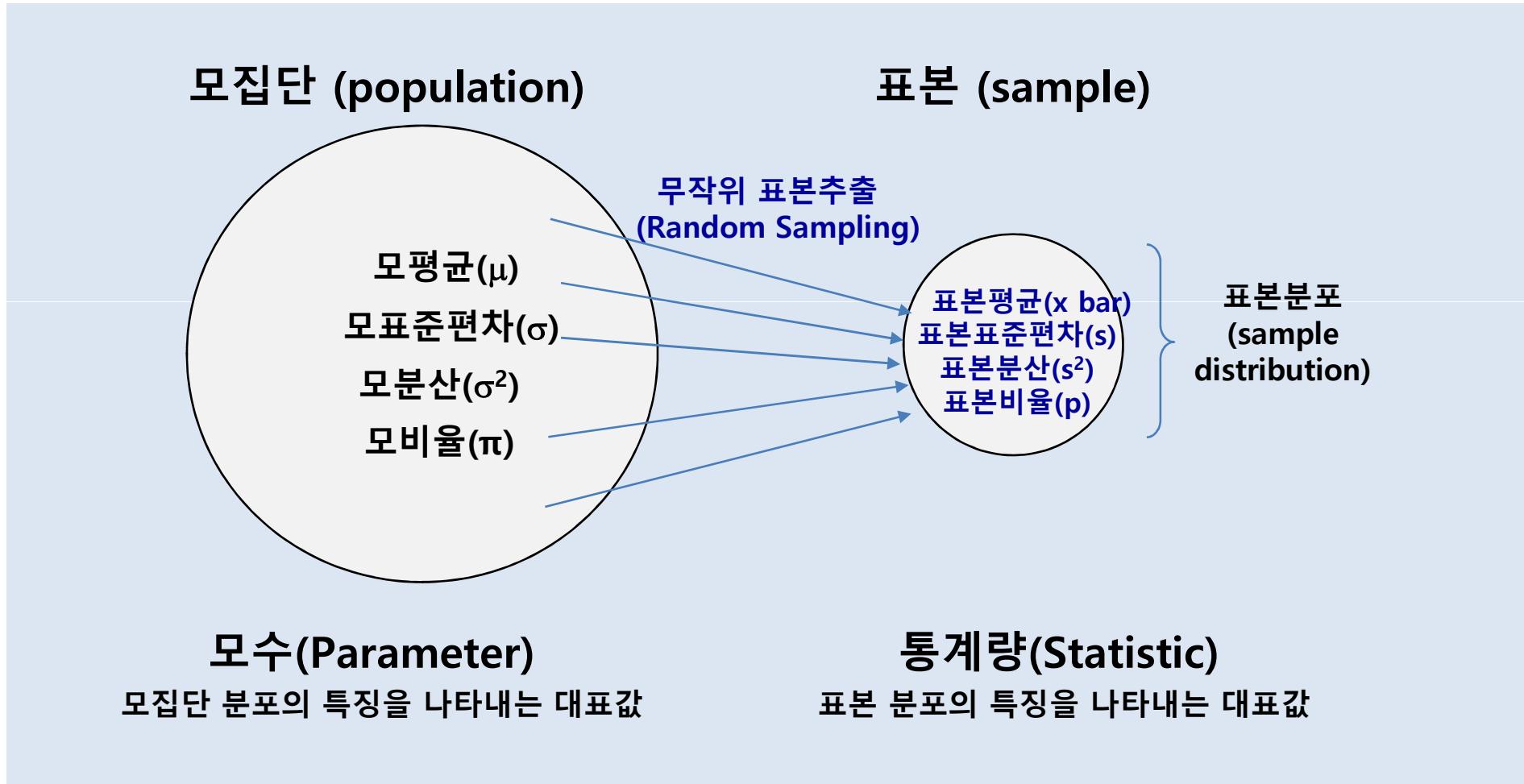
추론(Inferential, 推論) 통계

- 표본에 포함된 정보로부터 모집단의 특성 파악 및 타당성을 검토하여 모수를 추론하거나 미래를 예측하는 통계기술



▪ 모집단과 표본

[모집단, 모수, 표본, 통계량, 표본분포의 관계]



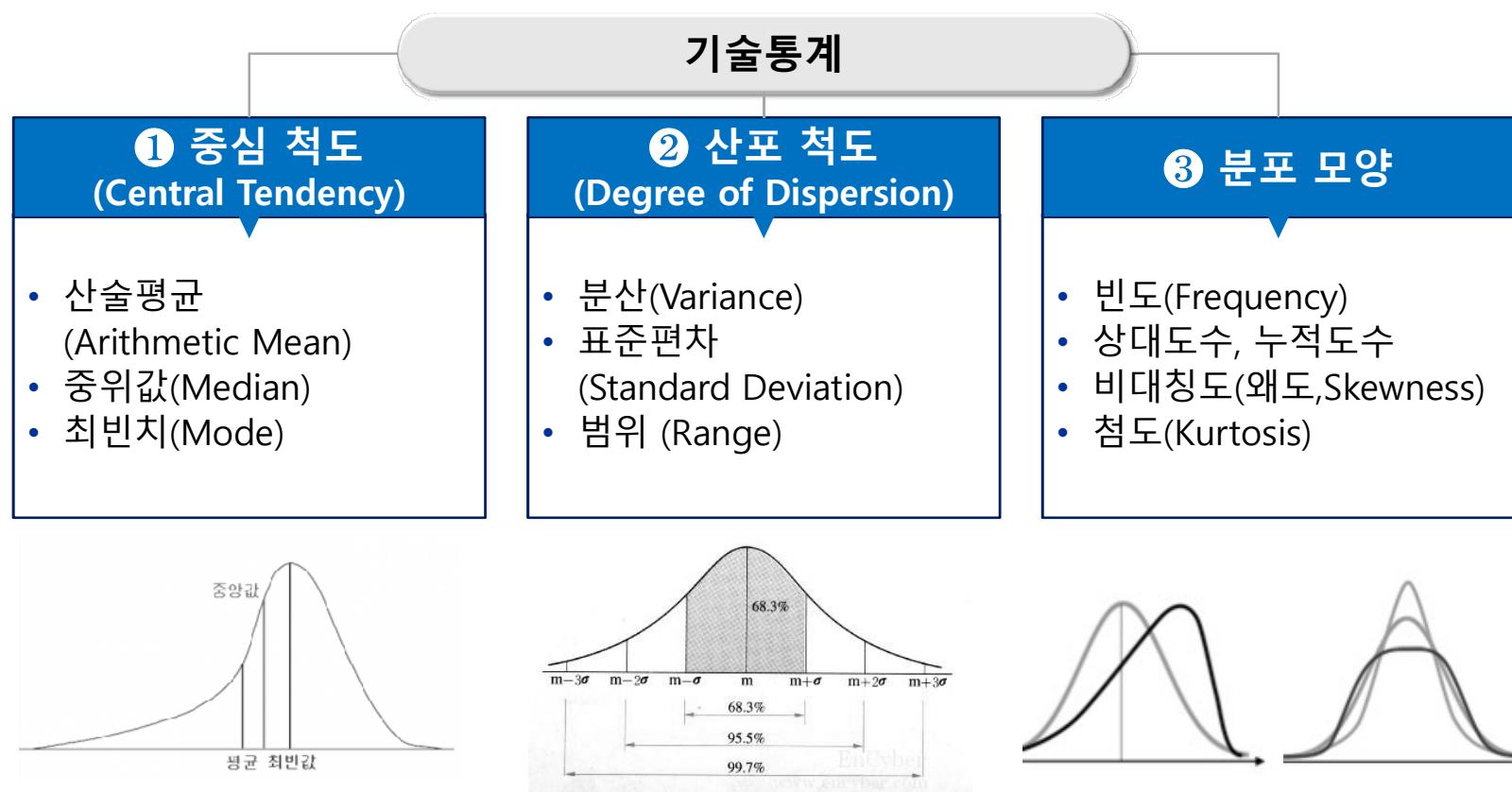
기호	설명	기호	설명
μ	모집단의 평균(모평균)	α	유의수준 [1- α :신뢰수준]
\bar{X}	표본집단의 평균(표본평균)	β	회귀계수
σ^2	모집단의 분산(모분산)	ϵ	오차
σ	모집단의 표준편차(모표준편차)	θ	모수
s^2	표본의 분산(표본분산)	p	모집단 상관계수(모집단, 표본)
s	표본의 표준편차(표본표준편차)	r	표본 상관계수

※ 모집단(population) vs. 표본(sample)

■ 정의

: 데이터의 속성을 특정한 통계량을 사용해 정리, 요약, 설명하는 방법

- 기술통계는 매우 간단한 통계량으로 엄청난 양의 데이터가 갖는 속성을 합리적인 방법으로 간명하게 요약해 줌으로써 독자가 데이터의 속성을 쉽게 이해할 수 있도록 도와준다



▪ 통계 관련 패키지 불러오기

```
import numpy as np
from scipy import stats
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from statsmodels.stats.proportion import proportions_ztest
```

▪ 연습

예제

자동차 연비 Data Set(mycars.csv)에서 자동차모델(model)별로 연비(mpg)에 대한
기술통계량을 구하시오.

- 항목명 : manufacturer, model, displacement(배기량), year, cylinder, automatic
(자동/수동), driving(전륜/후륜 등), mpg(연비), highway_mileage, fuel,
class(compact/suv 등)

	manufacturer	model	displacement	year	cylinder	automatic	driving	mpg	highway_mileage	fuel	class
0	audi	a4	1.8	1999	4	auto	f	18		p	compact
1	audi	a4	1.8	1999	4	manual	f	21		p	compact
2	audi	a4	2.0	2008	4	manual	f	20		p	compact
3	audi	a4	2.0	2008	4	auto	f	21		p	compact
4	audi	a4	2.8	1999	6	auto	f	16		p	compact

▪ 연습

데이터 불러오기

```
ds_mycars = pd.read_csv("C:/삼성멀티캠퍼스/data/mycars.csv")
```

df = ds_mycars[['model','mpg']]

모델별 기술통계량 표시

```
df.groupby('model').describe()
```

model	mpg								
	count	mean	std	min	25%	50%	75%	max	
AMC Javelin	1.0	15.2	NaN	15.2	15.2	15.2	15.2	15.2	
Cadillac Fleetwood	1.0	10.4	NaN	10.4	10.4	10.4	10.4	10.4	
Camaro Z28	1.0	13.3	NaN	13.3	13.3	13.3	13.3	13.3	
Chrysler Imperial	1.0	14.7	NaN	14.7	14.7	14.7	14.7	14.7	
Datsun 710	1.0	22.8	NaN	22.8	22.8	22.8	22.8	22.8	
Dodge Challenger	1.0	15.5	NaN	15.5	15.5	15.5	15.5	15.5	
Duster 360	1.0	14.3	NaN	14.3	14.3	14.3	14.3	14.3	
Ferrari Dino	1.0	19.7	NaN	19.7	19.7	19.7	19.7	19.7	
Fiat 128	1.0	32.4	NaN	32.4	32.4	32.4	32.4	32.4	
Fiat X1-9	1.0	27.3	NaN	27.3	27.3	27.3	27.3	27.3	
Ford Pantera L	1.0	15.8	NaN	15.8	15.8	15.8	15.8	15.8	
Honda Civic	1.0	30.4	NaN	30.4	30.4	30.4	30.4	30.4	
Hornet 4 Drive	1.0	21.4	NaN	21.4	21.4	21.4	21.4	21.4	
Hornet Sportabout	1.0	18.7	NaN	18.7	18.7	18.7	18.7	18.7	
Lincoln Continental	1.0	10.4	NaN	10.4	10.4	10.4	10.4	10.4	

Lotus Europa	1.0	30.4	NaN	30.4	30.4	30.4	30.4	30.4	30.4
Maserati Bora	1.0	15.0	NaN	15.0	15.0	15.0	15.0	15.0	15.0
Mazda RX4	1.0	21.0	NaN	21.0	21.0	21.0	21.0	21.0	21.0
Mazda RX4 Wag	1.0	21.0	NaN	21.0	21.0	21.0	21.0	21.0	21.0
Merc 230	1.0	22.8	NaN	22.8	22.8	22.8	22.8	22.8	22.8
Merc 240D	1.0	24.4	NaN	24.4	24.4	24.4	24.4	24.4	24.4
Merc 280	1.0	19.2	NaN	19.2	19.2	19.2	19.2	19.2	19.2
Merc 280C	1.0	17.8	NaN	17.8	17.8	17.8	17.8	17.8	17.8
Merc 450 SE	1.0	16.4	NaN	16.4	16.4	16.4	16.4	16.4	16.4
Merc 450 SL	1.0	17.3	NaN	17.3	17.3	17.3	17.3	17.3	17.3
Merc 450 SLC	1.0	15.2	NaN	15.2	15.2	15.2	15.2	15.2	15.2
Pontiac Firebird	1.0	19.2	NaN	19.2	19.2	19.2	19.2	19.2	19.2
Porsche 914-2	1.0	26.0	NaN	26.0	26.0	26.0	26.0	26.0	26.0
Toyota Corolla	1.0	33.9	NaN	33.9	33.9	33.9	33.9	33.9	33.9
Toyota Corona	1.0	21.5	NaN	21.5	21.5	21.5	21.5	21.5	21.5
Valiant	1.0	18.1	NaN	18.1	18.1	18.1	18.1	18.1	18.1
Volvo 142E	1.0	21.4	NaN	21.4	21.4	21.4	21.4	21.4	21.4

▪ 확률 정의

: 여러 가능한 결과 중 하나 또는 일부가 일어날 가능성 (0~1 사이의 값으로 정의)

동전을 던졌을 때
앞면이 나올 가능성



동전을 던졌을 때
앞면이 나올 확률 0.5

▪ 확률 용어

실험 or 시행

여러 가능한 결과 중 하나가 일어나도록 하는 행위

표본 공간(S)

실험에서 나타날 수 있는 모든 결과들을 모아둔 집합

사건

표본공간의 일부분, 사건 A가 일어날 확률 : $P(A)$

예) 동전을 던지는 실험 - 앞면 : H, 뒷면 : T

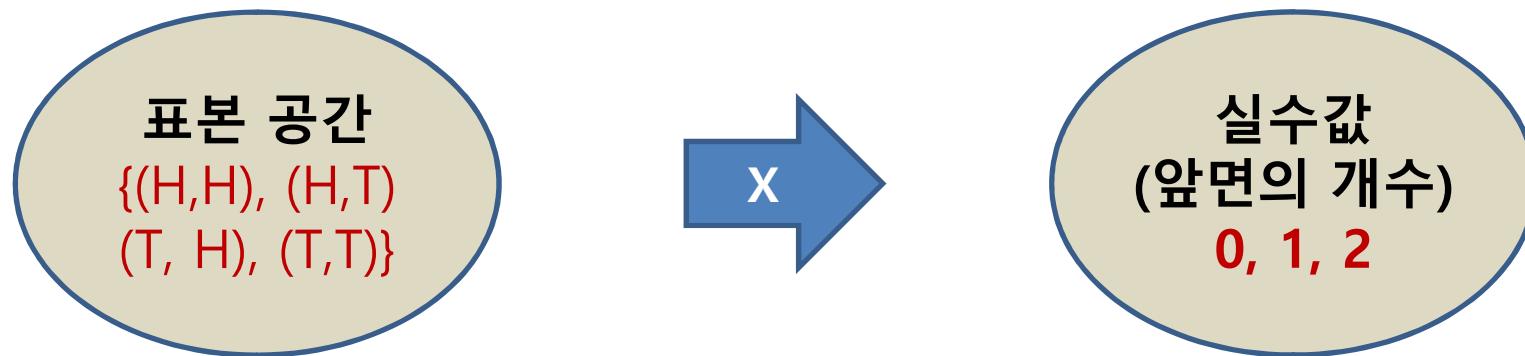
표본 공간 = {H, T}

앞면이 나오는 사건 A = {H} 이므로, $P(A) = \frac{1}{2} = 0.5$

- 확률변수(Random variable)

각각의 사건에 실수 값을 대응시킨 함수 : X, Y 처럼 대문자로 표시

[동전 2개를 던질 경우]



동전을 2개 던졌을 때 앞면의 수 : 확률 변수 X

$$\begin{aligned}P(X=0) &= \frac{1}{4} \\P(X=1) &= \frac{1}{2} \\P(X=2) &= \frac{1}{4}\end{aligned}$$

▪ 확률변수(Random variable)

이산, 연속 확률 변수

이산형 확률 변수

확률변수의 값의 개수를 셀 수
있는 경우



확률질량 함수

$$y = f(x), \quad 0 \leq y \leq 1$$

$$\sum_{all xi} f(xi) = 1$$

연속형 확률 변수

확률변수의 값이 연속적인
구간에 속하는 경우



확률밀도 함수

$$y = f(x), \quad 0 \leq y \int_{-\infty}^{\infty} f(x)dx = 1$$

이산형 확률 분포

동전을 던질 때 앞면(H)과 뒷면(T)이 나타날 확률이 동일하다는 가정하에,
동전을 3개 동시에 던져서 앞면이 나오는 개수를 확률변수 X 로 정의

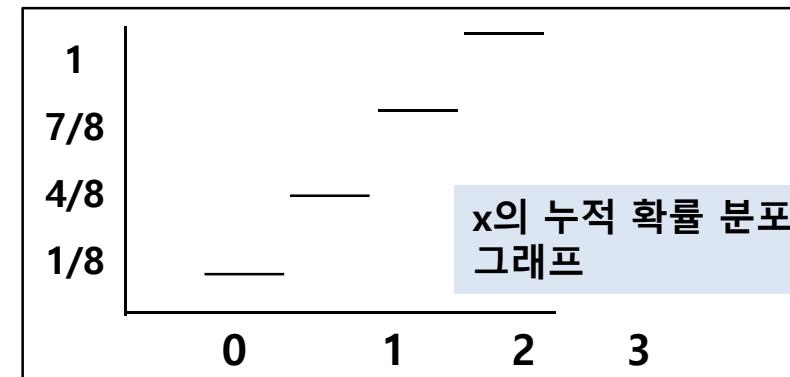
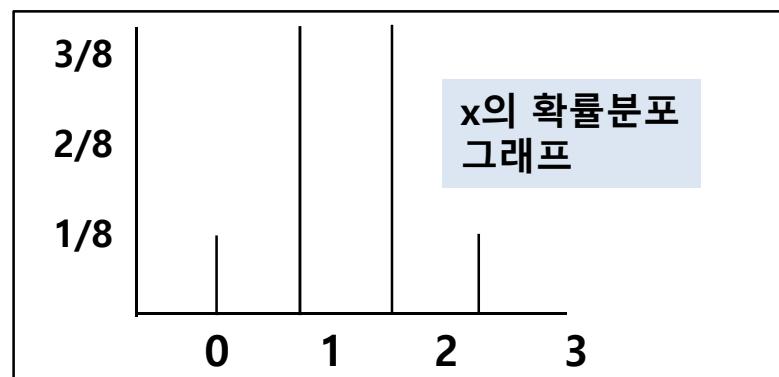
표본공간은 $S = \{(H,H,H), (H,H,T), (H,T,H), (H,T,T), (T,H,H), (T,H,T), (T,T,H), (T,T,T)\}$

$x=0$ 인 경우는 (T,T,T) 뿐이므로 확률은 $1/8$

$x=1$ 인 경우는 (H,T,T), (T,H,T), (T,T,H)로 3개이므로 확률은 $3/8$

이산형 확률 분포표

x 의 값	$P(X=x)$	$f(x)=P(X\leq x)$
0	$1/8$	$1/8$
1	$3/8$	$4/8$
2	$3/8$	$7/8$
3	$1/8$	1



연속형 확률 분포

전구의 수명이 다할 때 까지 관찰하는 실험에서 확률변수 X 를 전구의 수명으로 정의

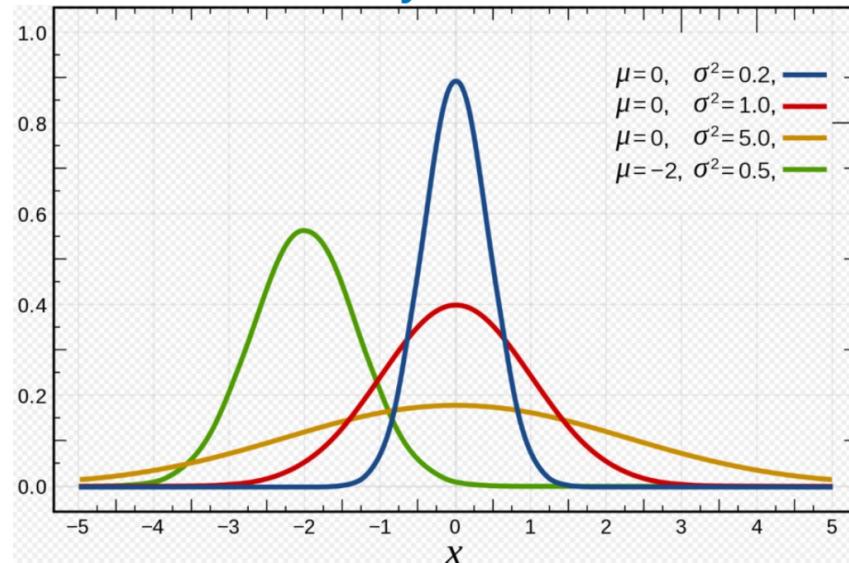
$$x : 0 \sim \infty, f(x) \geq 0$$

수명이 1.5시간에서 5.5시간 사이의 확률은

$$P(1.5 \leq y \leq 5.5) = \int_{1.5}^{5.5} f(x)dx$$

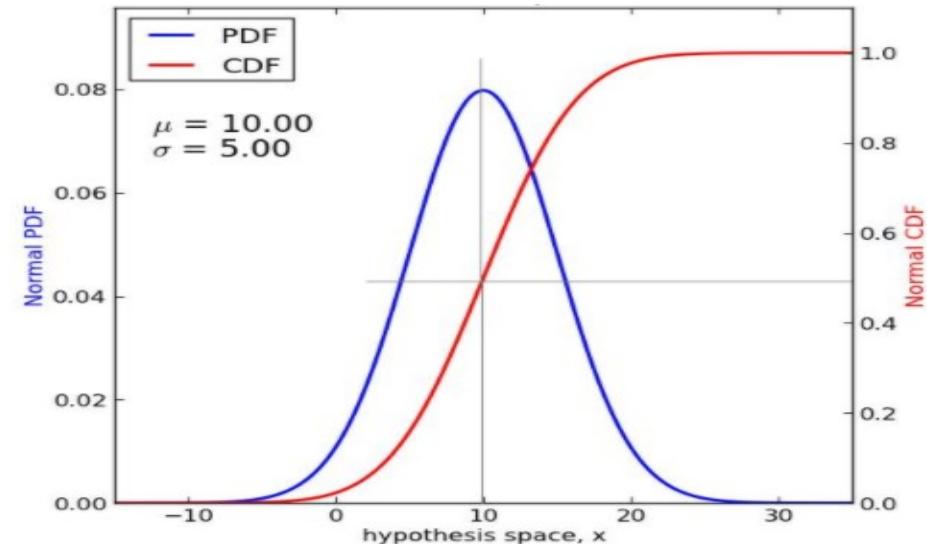
연속확률분포

Probability Distribution



누적확률분포

Cumulative Distribution



▪ 확률분포 종류



분포의 종류

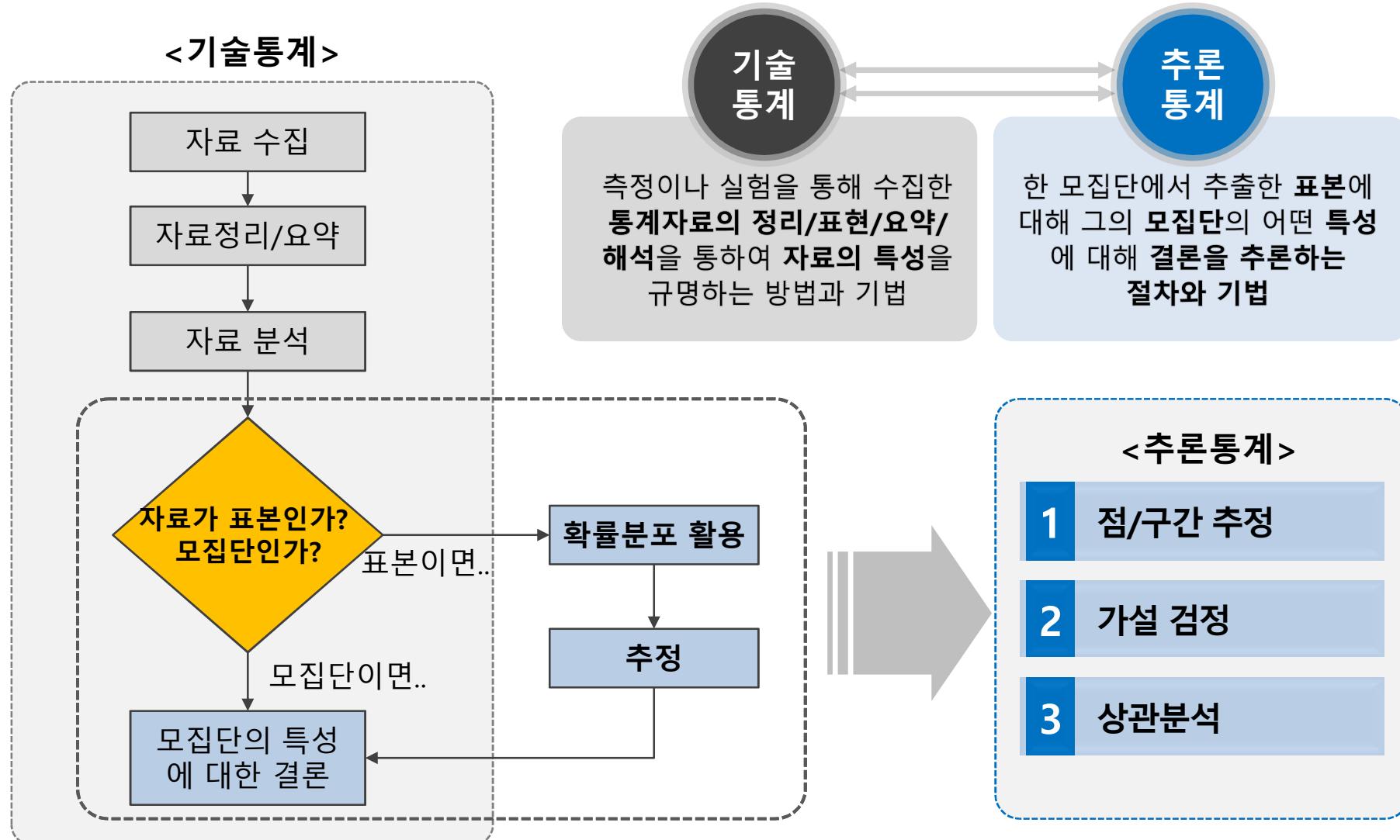
연속 확률분포

- 정규분포
- 표준정규분포
- t 분포(student t)
- χ^2 분포
- F 분포
- 와이블 분포

이산 확률분포

- 베르누이 분포
- 이항 분포
- 포아송 분포
- 초기하 분포

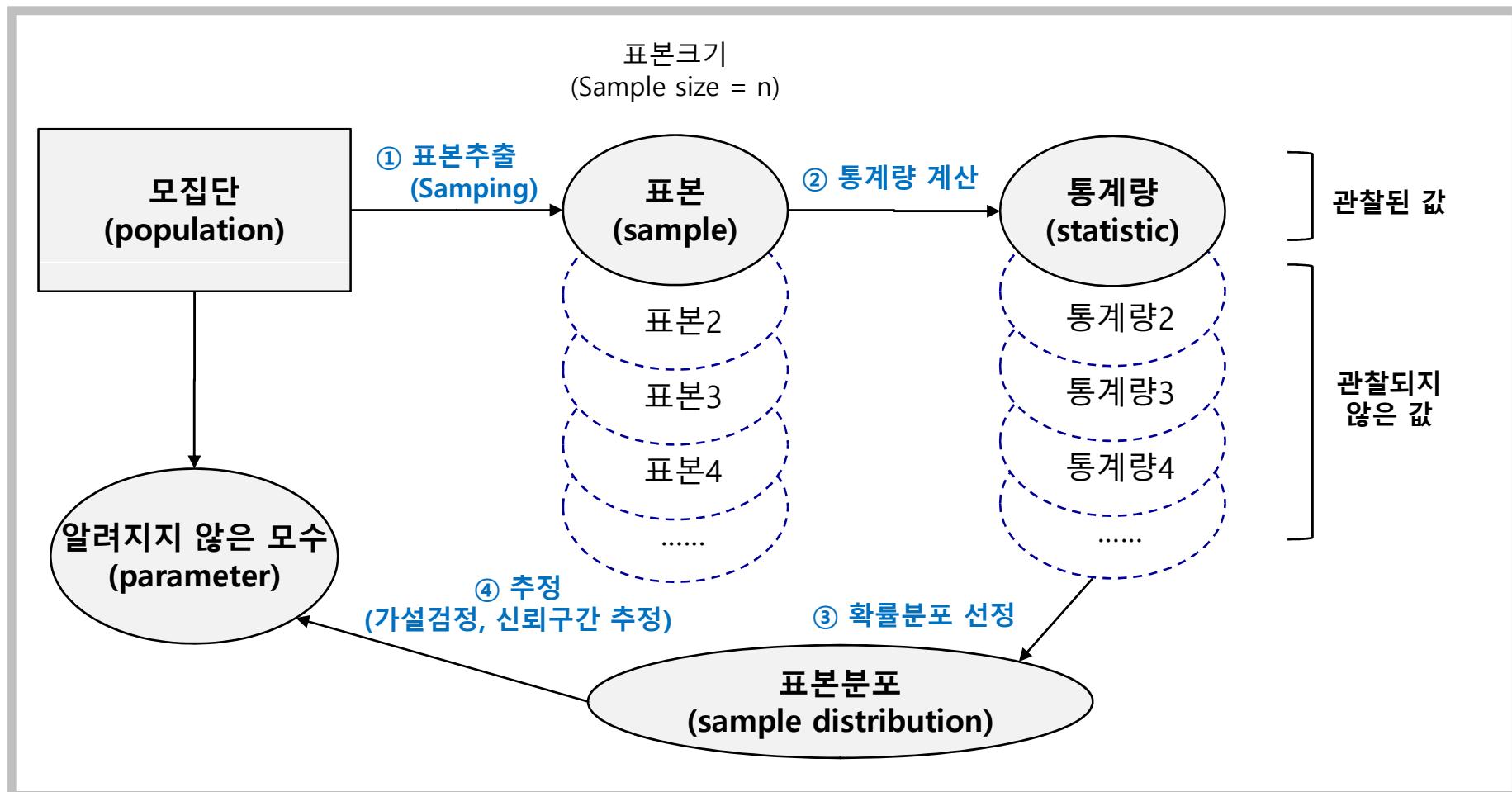
▪ 추론통계 과정



▪ 추론통계와 검정

: 표본데이터에 기초하여 모집단의 특성에 관한 결론을 얻거나 또는 추론을 하기 위해 사용되는 통계방법

「 모집단, 모수, 표본, 통계량, 표본분포의 관계 」



▪ 정의

: 표본에 포함된 정보로부터 모집단의 특성을 파악하고 타당성을 검토하여 모수를 추론하거나 미래를 예측하는 것

▪ 활용용도

▶ 통계적 추정(statistical estimation)

: 표본의 성격을 나타내는 통계량을 기초로 하여 모수를 추정하는 통계적 분석 방법

▶ 가설검정(hypothesis test)

: 모수에 대하여 특정한 가설을 세워놓고, 표본을 선택하여 통계량을 계산한 다음 이를 기초로 하여 모수에 대한 가설의 진위를 판단하는 방법

▪ 이론

모수(Parameter)

- 모집단(population)의 기술적 척도

통계량(Statistic)

- 표본(sample)의 기술적 척도

표본≠모집단 → 표본으로 구해지는 결론과 추정치들은 항상 옳은 것이 아님

그래서, 통계적 추론에 신뢰의 척도, 즉 신뢰수준(confidence level)과 유의수준(significance level) 필요

▪ 정의

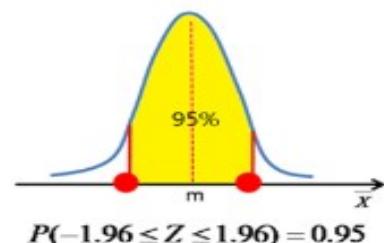
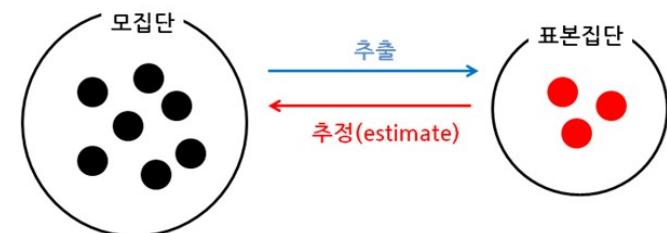
: 모집단에서 추출한 표본에서 얻은 정보를 이용하여 모집단의 평균, 표준편차 등을 추측하는 것

▪ 종류

- ▶ 점 추정 : 하나의 값으로 추정
- ▶ 구간추정 : 값을 포함하는 구간을 추정

▪ 이론

- ▶ 신뢰도 : 추정하고자 하는 모평균이 신뢰구간에 포함될 확률
- ▶ 모평균 신뢰도 (95%, 99%)



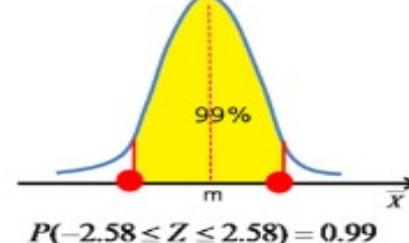
$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

$$-1.96 \leq \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} \leq 1.96$$

$$\rightarrow \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \quad (\text{신뢰도 } 95\% \text{ 신뢰구간})$$

의미: \bar{X} 를 이용해서 구한 범위내에 평균이 있을 확률이 95%라는 뜻

95%



$$P(-2.58 \leq Z \leq 2.58) = 0.99$$

$$-2.58 \leq \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} \leq 2.58$$

$$\rightarrow \bar{X} - 2.58 \frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X} + 2.58 \frac{\sigma}{\sqrt{n}} \quad (\text{신뢰도 } 99\% \text{ 신뢰구간})$$

의미: \bar{X} 를 이용해서 구한 범위내에 평균이 있을 확률이 99%라는 뜻

99%

▪ 정의

: 표본을 통해 얻은 정보를 이용하여 모집단의 특성에 대한 가설의 진위를 판단하는 과정

▪ 절차

1) 가설 수립

- ▶ 가설 검정의 목적 확인
- ▶ 새롭게 주장하고자 하는 판단 결과를 대립가설(H_1), 그 반대를 귀무가설(H_0)로 설정
- ▶ 유의수준(α)을 결정함 (보통 5% 혹은 1%)

2) 가설 검정의 수행

- ▶ 적절한 검정통계량을 결정함(t , F , χ^2)
- ▶ 데이터로부터 검정통계량을 계산함
- ▶ 데이터로부터 p-value를 계산함

3) 검정결과의 판단

- ▶ 검정통계량 > 임계값이면 H_0 기각,
검정통계량 < 임계값이면 H_0 기각할 수 없음
- ▶ $p\text{-value} < \text{유의수준}(\alpha)$ 이면 H_0 를 기각,
 $p\text{-value} > \text{유의수준}(\alpha)$ 이면 H_0 를 기각할 수 없음

▪ 가설 (Hypothesis)

H_0 : 귀무가설(Null Hypothesis)

- : 기존의 사실에 대한 가설 → 개선 전후 비교 시 개선 전 사실
- : 검정통계량은 귀무가설의 분포에서 나옴
- : 검정의 대상으로 삼는 가설, 차이가 없음. 영향을 주지 않는다는 입장

H_1 : 대립가설(Alternative Hypothesis)

- : 새롭게 확인하고자 하는 사실에 대한 가설 → 개선 전후 비교 시 개선 후 사실
- : 검정통계량이 귀무가설에서 나왔다라고 보기 어려울 경우(p-value가 작을 경우) 대립가설을 선택함
- : 귀무가설을 부정하는 가설, 차이가 있다, 영향을 준다는 가설

▪ 가설 수립의 예

귀무가설	대립가설
H_0 : 직원의 평균몸무게는 75kg이다	H_1 : 직원의 평균몸무게는 75kg이 아니다
H_0 : 고객만족도는 90%이다	H_1 : 고객만족도는 90%이 아니다
H_0 : 남자와 여자의 평균키는 같다	H_1 : 남자와 여자의 평균키는 다르다

▪ 유의수준(α)

- ▶ 귀무가설을 기각한다는 결정을 내릴 때, 귀무가설을 기각하는 결정이 잘못될 수 있을 최대가능성(확률)을 유의수준이라고 함
- ▶ 유의수준을 1% or 5% or 10%로 설정하는 것은 가설검정을 실시하는 분석자가 새로운 가설을 받아들이는데 있어서 발생될 수 있는 판단 오류에 대한 위험을 얼마나 Tight한 기준으로 가져갈 것인가에 따라 결정되며, 이는 분석상황과 분석자의 주관적인 가치에 따라 달라질 수도 있음

▪ 임계값

- ▶ 주어진 유의수준에서 귀무가설의 채택과 기각에 관련된 의사결정을 할 때, 그 기준이 되는 통계량을 의미함
- ▶ 임계값을 중심으로 귀무가설의 기각영역과 채택영역이 결정됨

▪ 가설검정의 오류(Testing Error)

제 1종 오류(α - Risk)

- 생산자 위험
- 귀무가설을 채택했어야 함에도 불구하고 이를 기각하는 위험, α 는 전형적으로 5%로 설정함
- 1종 오류는 초반에 사용자가 결정(신뢰수준은 $1 - \alpha$)

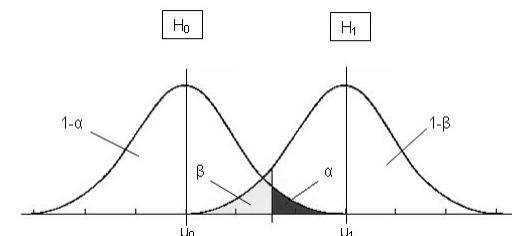
제 2종 오류(β - Risk)

- 소비자 위험
- 귀무가설을 기각했어야 함에도 불구하고 이를 채택하는 위험, β 는 전형적으로 10%를 설정함
- 다른 모든 값이 동일할 때, α 값이 작아지면, β 값은 증가
- 귀무가설을 기각하는데 많은 증거를 요구하게 되면 제 2종 오류가 일어날 확률이 높아짐
- $1 - \beta$ 는 귀무가설이 거짓일 때 이를 기각할 확률 검정력 (Power of the test).

검정결과	실제현상	H_0 참	H_1 참
	귀무가설 H_0 채택	옳은 결정	제 2종 오류
대립가설 H_1 채택	제 1종 오류	옳은 결정	

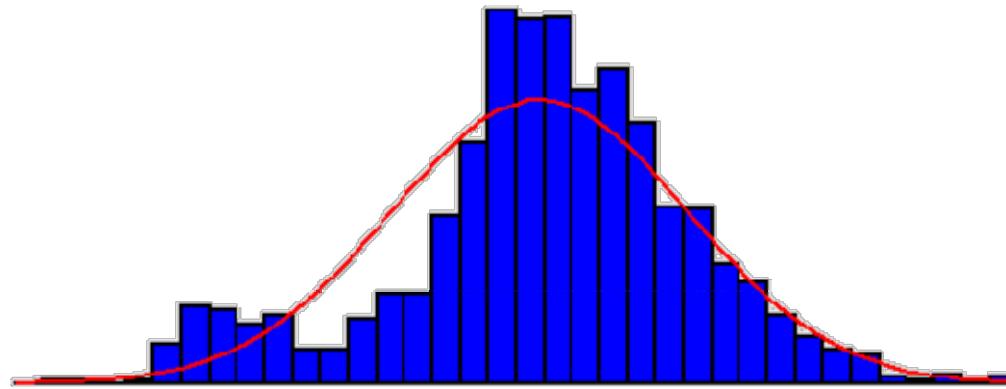
“실제로 죄가 없는데도 불구하고 유죄판결을 받음”

“실제로 범죄를 저질렀는데 무죄판결을 받음”



■ 정규성 검정(Normality Test)

- : 관측값들이 정규분포를 따르는 모집단에서 취해졌는지를 검정하는 것
 - * 정규분포란 평균을 중심으로 좌우 대칭의 종 모양의 곡선을 갖는 분포임



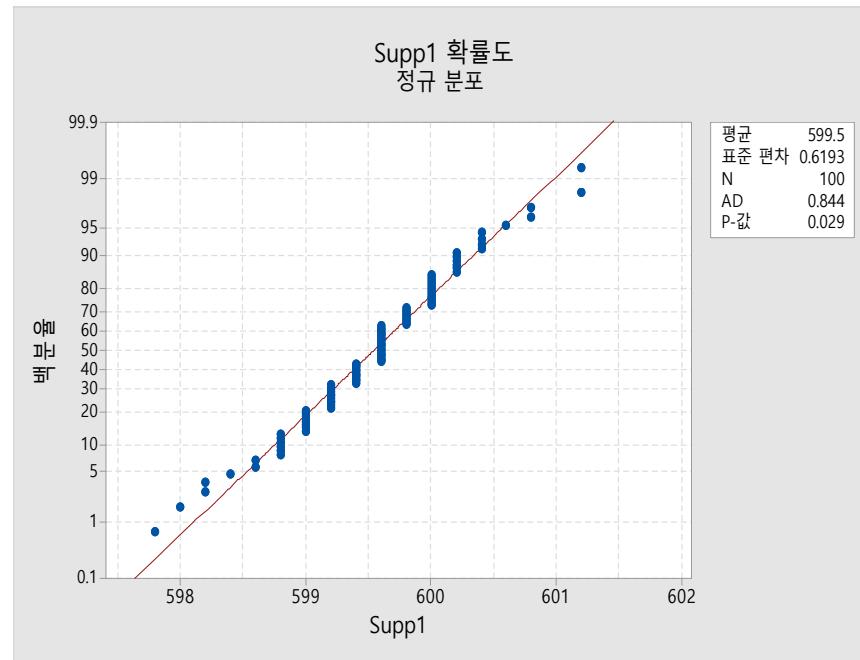
정규분포가 아닌 경우 착안사항

- ▶ 정규성을 갖지 않는 분포가 반드시 쓸모 없는 것 만은 아니다. 우리는 때로는 비정규성로부터 이상요인을 파악할 수 있다. 반드시 원인 파악이 필요하다
- ▶ 비정규성을 갖는 데이터는 우리에게 매우 귀중한 정보를 제공하고 있으며, 개선의 방향을 제공해 준다
- ▶ 비정규성의 유형과 원인 파악을 할 수 있어야 한다

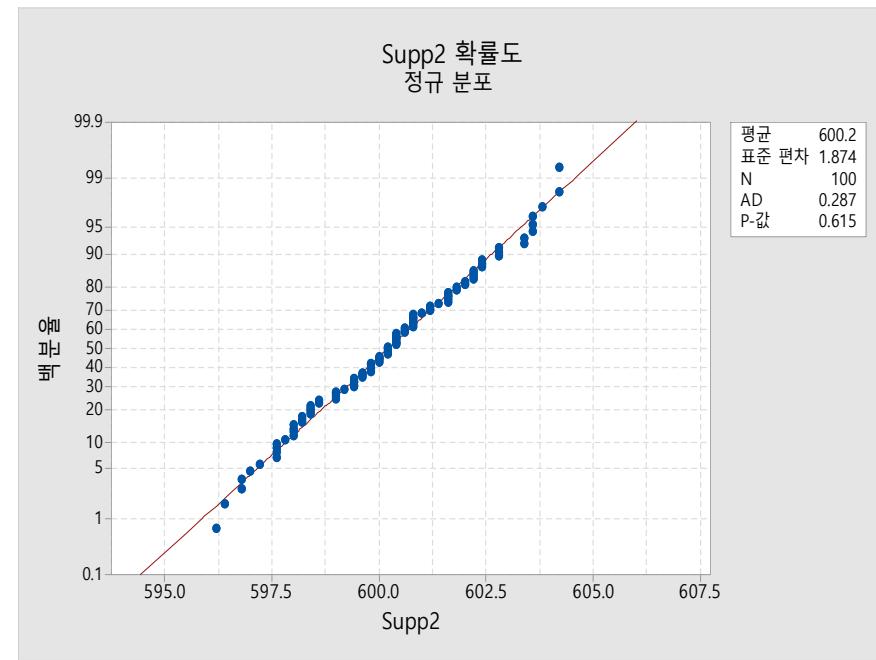
▪ 정규성 검정(Normality Test)

Anderson-Darling Normality :

- A-Squared : Anderson-Darling 검정통계량
- P-value : 정규성 검증결과(95%신뢰수준 사용시) 0.05보다 크면 정규, 0.05보다 작으면 비정규로 판단



P-value : 0.029 => 정규성이 없음



P-value : 0.615 => 정규성이 있음

■ t - test 종류

: 집단의 평균값에 대한 가설을 검정하는 도구이며 그 종류와 내용은 다음과 같음

- 한 집단의 평균이 특정 값과 같은지 비교

1-sample t - test

고객서비스 센터의
고객만족도(Yield) 평균은
76.7이 맞을까?

고등학교 3학년 남학생의
키는 175cm 인가?

- 두 집단 간 평균이 같은지 비교

2-sample t - test

재료 A와 재료 B로 만든
각 제품의 강도는 같은가?

서울지역과 경기지역의
고등학교 3학년 남학생의
키는 같은가?

- 쌍을 이룬 두 집단간 평균이 같은지 비교

Paired t - test

동일신발 바닥에 고무자재
A와 B를 사용한 경우 바닥의
마모 정도의 차이가 있는가?

교육 전과 후의
시험 점수에 차이가 있는가?

▪ 검정절차

절차1. 가설 수립 유의수준 설정

- . 귀무가설 (H_0) : "두 집단의 평균은 같다"
- . 대립가설 (H_1) : "두 집단의 평균은 같지 않다"

절차2. 가설검정 수행

- . 유의수준(α) 설정 : 0.10 or 0.05 or 0.01,
- . 정규성 검정
- . 등분산성 검정
- . 검정통계량 계산 : 등분산성 고려

절차3. 검정 결과 판단

- . 검정통계량과 임계치 비교, p-value 확인

▪ 1 sample t-test 연습

예제

고객서비스센터의 고객만족도 평균은 76.7이다. 개선활동을 완료한 후 다음과 같은 10개의 고객만족도 데이터를 얻었다. 개선활동이 만족도를 변화시켰는가? ($\alpha = 0.05$)
데이터 : data_norm.csv

```
# 데이터 불러오기  
df = pd.read_csv('C:/삼성멀티캠퍼스/data/data_norm.csv')  
# 정규성 검정  
x=stats.shapiro(df)  
print('The Shapiro-Wilkis Statistic is %.3f and the p-value is %.3f' %x)
```

```
The Shapiro-Wilkis Statistic is 0.933 and the p-value is 0.480
```

```
# 1-Sample t-test  
t_result = stats.ttest_1samp(df, 76.7)  
t, p = t_result.statistic.round(3), t_result.pvalue.round(3)  
print("1-Sample t-test")  
print("t검정통계량 : {} ".format(t))  
print("p-value : {} ".format(p))
```

```
1-Sample t-test  
T검정통계량 : 0.864  
P-value : 0.4098
```

▪ 2 sample t-test 연습

예제

고객만족도 확인을 위하여 A고객과 B고객에 대해 Survey하여 다음의 데이터를 얻었다.
A고객과 B고객의 모평균에 차이가 있다고 할 수 있는가? (유의수준 $\alpha = 0.05$)

- 데이터(10점 척도, $n = 20$)
 - A고객 : data_sampt1.csv
 - B고객 : data_sampt2.csv

데이터 불러오기

```
df1 = pd.read_csv('C:/삼성멀티캠퍼스/data/data_sampt1.csv')
df2 = pd.read_csv('C:/삼성멀티캠퍼스/data/data_sampt2.csv')
```

정규성 검정

```
x=stats.shapiro(df1)
print('Shapiro-Wilkis검정통계량은 %.3f, p-value는 %.3f' %x)
```

```
Shapiro-Wilkis검정통계량은 0.909, p-value는 0.062
```

```
y=stats.shapiro(df2)
```

```
print('Shapiro-Wilkis검정통계량은 %.3f, p-value는 %.3f' %y)
```

```
Shapiro-Wilkis검정통계량은 0.898, p-value는 0.038
```

등분산성 검정

```
stat, p = stats.bartlett(df1['sample'],df2['sample'])
print('등분산검정결과 p-value는 %.3f' % p)
```

```
등분산검정결과 p-value는 0.642
```

■ 2 sample t-test 연습

```
# 2-Sample t-test  
t_result = stats.ttest_ind(df1['sample'], df2['sample'], equal_var = True)  
t, p = t_result.statistic.round(3), t_result.pvalue.round(3)  
  
print( "2-Sample t-test ")  
print( "t검정통계량: {} ".format(t))  
print( "p-value: {} ".format(p))
```

```
2-Sample t-test  
t검정통계량: -3.122  
p-value: 0.003
```

▪ Paired t-test 연습

예제

회사원 10명에 대해 영어점수가 학원가기 전의 점수와 학원수강 후에 영어점수의 차이가 있는지 확인하고 싶다. 아래의 데이터를 가지고 차이가 있는지 검정하라. ($\alpha : 5\%$)

	man 1	man 2	man 3	man 4	man 5	man 6	man 7	man 8	man 9	man 10
test before	720	589	780	648	720	589	780	648	780	648
test after	810	670	790	712	810	670	790	712	790	712

- 데이터 : data_pair.csv

```
# 데이터 불러오기
```

```
df = pd.read_csv('C:/삼성멀티캠퍼스/data/data_pair.csv')
```

```
# paired t-test
```

```
t_result = stats.ttest_rel(df['before'], df['after'])
```

```
t, p = t_result.statistic.round(3), t_result.pvalue.round(3)
```

```
print("Paired t-test")
```

```
print("t검정통계량 : {} ".format(t))
```

```
print("p-value : {} ".format(p))
```

```
Paired t-test  
t검정통계량 : -5.324  
p-value : 0.0
```

■ 비율 검정

: 한 집단 또는 두 집단의 비율이 같은지를 검정하는 도구이며 그 내용은 다음과 같음

1 Proportion test

- 한 집단의 비율이 특정 비율과 같은지 비교

공정의 불량률이 10%인가?

2 Proportion test

- 두 집단간 비율 비교

동일한 제품을 생산하는 두 공장에서 불량률을 측정한 데이터가 있다. 두 공정의 불량률이 같다고 할 수 있는가?

■ 분산 검정

: 두 개 이상의 집단간 모 분산이 동일한지를 검정하는 도구이며 그 종류와 내용은 다음과 같음

2 Variances (F-test)

- 두 집단간의 분산 비교

과제 前後 배송 시간의 산포가 같은가?

Equal Variances

- 3 집단간의 분산 비교

제품(1 쿠키, 2 과자, 3 빵)별 오븐 가동 시간의 산포가 같은가?

▪ 연습

예제

전자제품 A를 사용하는 국내 사용자의 10% 정도가 만족을 표시했다. 한 해 동안 전자제품 A의 품질에 대해 노력을 하여 전체 사용자 중 100명을 표본으로 하여 사용 품질을 확인한 후 15명이 만족을 표현했다. 과연 품질개선을 한 결과로 기존보다 만족도의 차이가 있는 것인가? (유의수준(α)=0.05)

```
count = 15
nobs = 100
value = .1
# proportion test 실행
stat, pval = proportions_ztest(count, nobs, value)
```

```
print( "1 Proportion test ")
print( 'p검정통계량 : {0:0.3f}'.format(stat))
print( 'p-value : {0:0.3f}'.format(pval))
```

```
1 Proportion test
p검정통계량 : 1.400
p-value : 0.161
```

결론 : 결과적으로 전자제품 A에 대해 품질개선 결과 기존 대비 만족도의 차이가 있다고 할 수 없다.

▪ 연습

예제

동일한 제품을 생산하는 두 공장에서 불량률을 측정한 결과 아래와 같다.

두 공정의 불량률이 같다고 할 수 있는가? (유의수준(α)=0.05)

- 공장 1 : $N_1 = 1000$, $X_1 = 4$
- 공장 2 : $N_2 = 1200$, $X_2 = 1$

```
count = np.array([4, 1])
nobs = np.array([1000, 1200])

# proportion test 실행
stat, pval = proportions_ztest(count, nobs)

print( "2 Proportion test ")
print( 'p검정통계량 : {0:0.3f} '.format(stat))
print( 'p-value : {0:0.3f}'.format(pval))
```

```
2 Proportion test
p검정통계량 : 1.553
p-value : 0.120
```

결론 : P-value > 0.05 이므로 귀무가설을 기각할 수 없음. 따라서, 두 공장의 불량률은 차이가 있다고 할 수 없다.

▪ 연습

예제

고객만족도 확인을 위하여 A고객과 B고객에 대해 Survey하여 다음의 데이터를 얻었다.

A고객과 B고객의 모분산에 차이가 있다고 할 수 있는가? (유의수준 $\alpha = 0.05$)

- 데이터(10점 척도, $n = 20$)
 - data_var.csv

```
df = pd.read_csv('C:/삼성멀티캠퍼스/data/data_var.csv')
df
```

```
# 등분산성 검정
```

```
stats.bartlett(df['A'],df['B'])
```

```
BartlettResult(statistic=0.2158392670118528, pvalue=0.6422286416307027)
```

결론 : 유의수준 5%에서 검정결과 P값이 0.642 이므로 고객간 고객만족도의 분산은 차이가 없다.

■ 카이제곱 검정

- : 관찰된 빈도가 기대되는 빈도와 의미있게 다른지의 여부를 검증하는 검증방법
- 자료가 빈도로 주어졌을 때, 특히 범주형 자료의 분석에 이용된다

동일성 검정

- 특성 별 두 가지 이상으로 분류된 범주간에
상호 동일한 비율로 나타나는가를
검정하고자 할 때 사용함

A공장에서 작업하는 3개의 전자제품에 대하여
4개 조별 작업분에 대하여 각 전자제품별
조별로 부하율의 차이가 있는가?

독립성 검정

- 특성 별 두 가지 이상으로 분류된 범주간에
상호 관련성이 있는지를 검정하고자 할 때
사용함

현대/GM/기아 자동차 소유주와 강남/강북
거주자 사이의 관계가 있는가?
(브랜드와 거주지 사이에 관계가 있는가?)

적합도 검정

- 어떤 특성치** 또는 사건이 **기대치**(또는
이론치)에 따라 **발생했는지 여부를 검정**
하고자 할 때 사용함

완두콩 수확을 해 보니 외관이 4가지 종류이고
수확량이 170, 60, 80, 30이라면 멘델의
유전 법칙(9:3:3:1)과 일치한다고 할 수 있나?

▪ 연습

예제

공장별로 전자제품을 생산하고 있는데, 전자제품의 종류가 3가지가 있다.
공장별로 전자제품을 생산하는 차이가 있는가?

구분	1공장	2공장	3공장	4공장
냉장고	270	260	236	234
세탁기	228	285	225	262
건조기	277	284	231	208

- 데이터 : data_chi.csv

데이터 불러오기

```
df = pd.read_csv('C:/삼성멀티캠퍼스/data/data_chi.csv', encoding = 'euc-kr')
```

chi-square test 실행

```
chi, pval, dof, expected = stats.chi2_contingency(df)
```

```
print("chi-square test")
```

```
print(' chisq: {0:0.3f}'.format(chi))
```

```
print(' p: {0:0.3f}'.format(pval))
```

```
print(' degree pf freedom: {}'.format(dof))
```

```
print(' expected value: \n{}' .format(expected.round(3)))
```

chi-square test

chisq: 13.366

p: 0.038

degree pf freedom: 6

expected value:

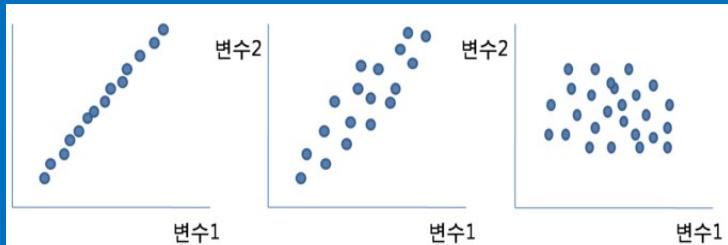
```
[ [258.333 258.333 258.333]
  [276.333 276.333 276.333]
  [230.667 230.667 230.667]
  [234.667 234.667 234.667] ]
```

결론 : 공장별로 제품을 생산하는 부하 차이가 있다고 할 수 있다.

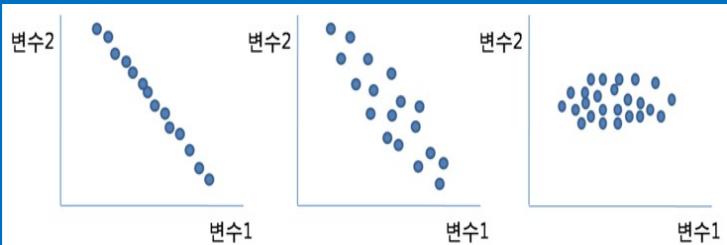
■ 상관분석

- 두 수량형 변수간에 선형적 관계의 강도와 방향을 분석하는 통계방법
- 한 변수가 증가할 때 다른 한 변수가 선형적인 증가 또는 감소하는지를 나타낸 것. 산점도 활용

양의
선형
관계



음의
선형
관계



■ 선형관계의 척도

공분산(covariance)

- 둘 이상의 변량이 연관성을 가지며 분포하는 모양을 전체적으로 나타낸 분산
- 두 변수가 동일한 방향으로 움직일 때 (두 변수 모두 증가하거나 또는 감소할 때), 공분산은 크고 양의 값을 가짐
- 두 변수가 반대방향으로 움직일 때, 공분산은 크고 음의 값을 가짐

$$\sigma_{xy} = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{N} \quad s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

상관계수(coefficient of correlation)

- 두 변수 간의 선형적인 관계 정도와 방향을 수치로 표시한 표준화된 지수
- 공분산은 척도의 단위에 따라 달라짐
→ 상관계수 활용

▪ 연습

예제

부품수리시간과 부품 수간에 관계를 분석하기 위한 상관분석을 실시하시오.

구분	1	2	3	4	5	6	7	8	9	10
MINUTES	1	2	3	4	4	5	6	6	7	8
UNITS	23	29	49	64	74	87	96	97	109	119

- 데이터 : data_cor.csv

```
# 데이터 불러오기  
df = pd.read_csv('C:/삼성멀티캠퍼스/data/data_cor.csv')  
# 상관분석 실행  
corr, pval = stats.pearsonr(df['minutes'], df['units'])  
  
print( "Correlation Analysis ")  
print( ' corr: {0:0.3f}' .format(corr))  
print( ' p: {0:0.3f}' .format(pval))
```

Correlation Analysis
corr: 0.989
p: 0.000

결론 : 부품수리시간과 부품수간에 매우 강한 상관성이 있다고 할 수 있다