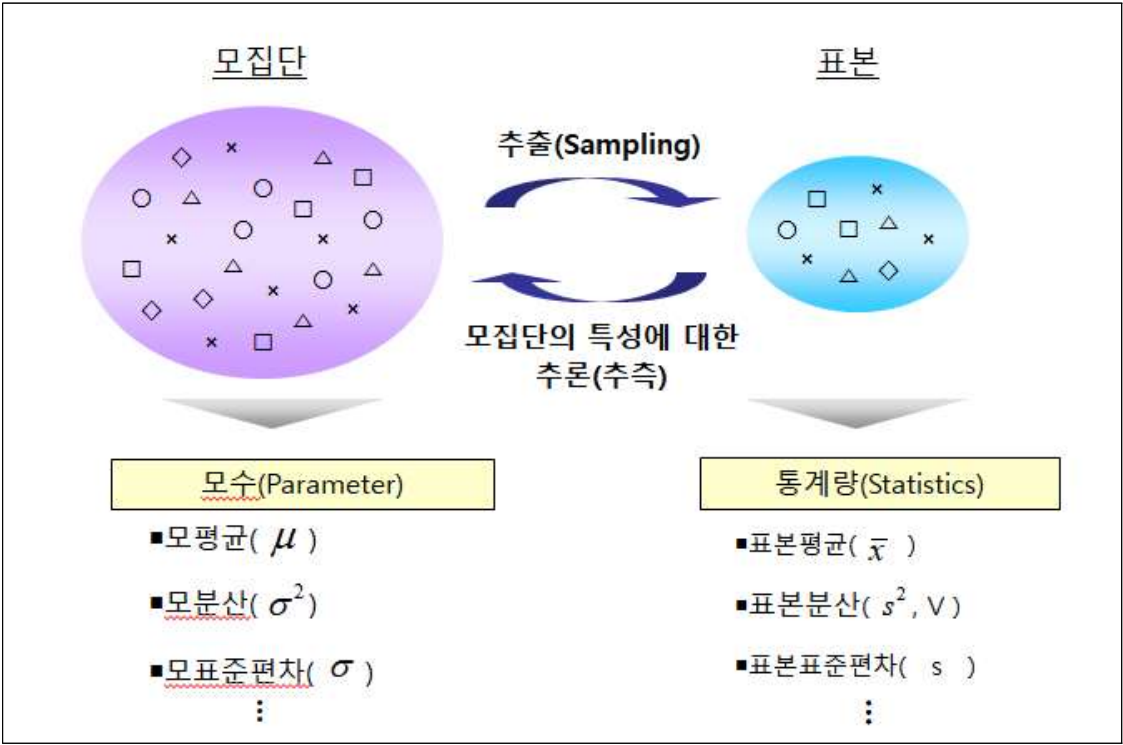


기초통계학 특강자료

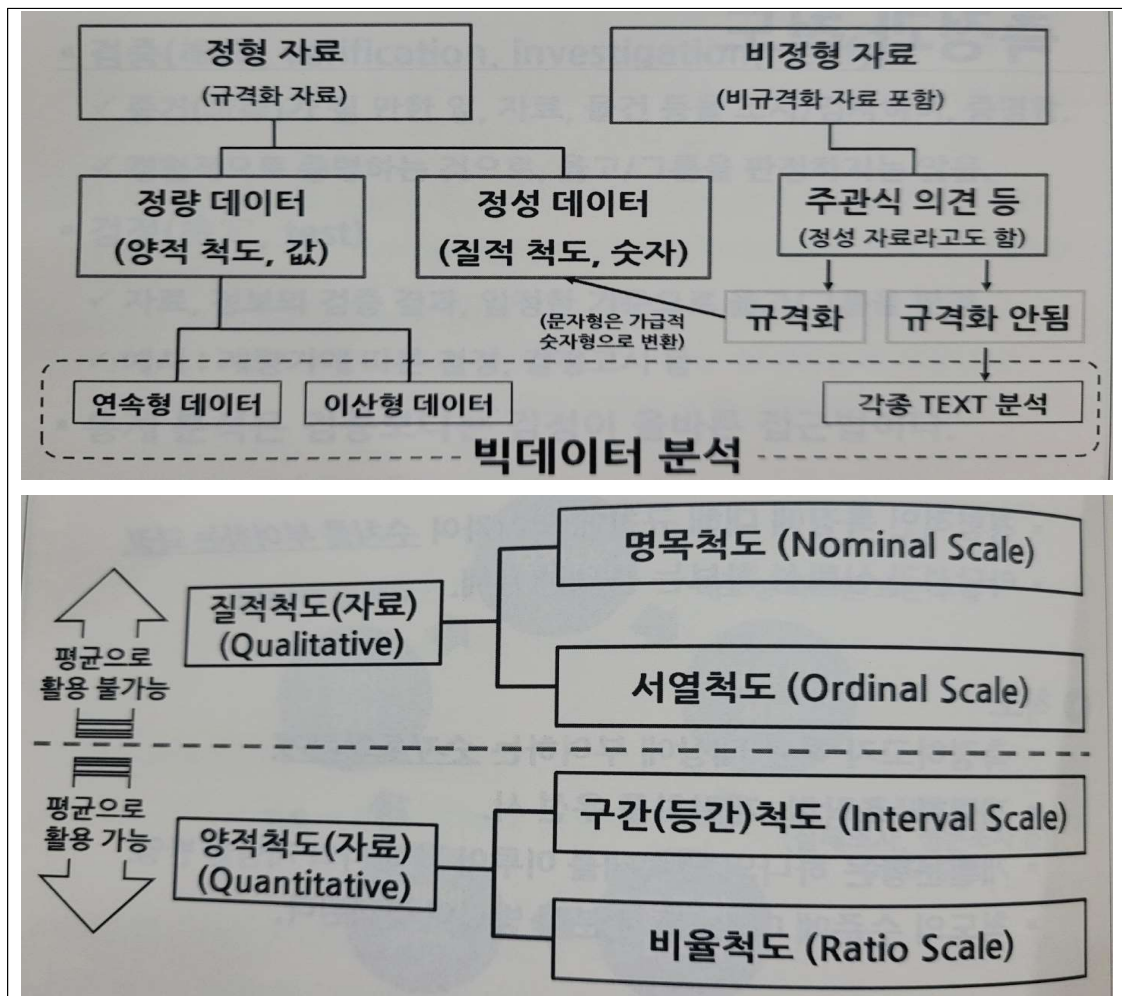
1. 기초통계의 목적(부제:통계를 왜 하는가?)
알고 싶은 대상(모집단)의 특성을 수치(모수)로 알고 싶은 것
2. 통계량(모집단-표본집단) 비교



3. 변수

구분	독립변수	종속(결과)변수에 영향을 미치는 원인 변수
	종속변수	독립(원인)변수의 영향으로 변화하는 결과 변수

종류	범주형 데이터 (qualitative) 질적, 정성적	범주 (categorical) 로 표현	명목형 데이터	범주 간에 순서의 의미가 없는 경우	성별, 지역 등
			순서형 데이터	범주 간에 순서의 의미가 있는 경우	상/중/하, 학점 등
	수치형 데이터 (quantitative) 양적, 정량적	수치 (numeric) 로 표현	이산형 데이터	관측 가능한 값이 정수로 셀 수 있는 경우	인구수, 불량수 등
			연속형 데이터	관측 가능한 값이 연속적인 경우	길이, 무게 등



4. 자료의 특성 분석

1) 중심경향치(central tendency) : 자료의 분포가 중심을 향해 밀집하는 경향

구분	설명
평균(mean) 산술평균, 기하평균, 조화평균, 평방평균 등	모든 관측값의 합을 관측값의 개수로 나눈 값 극단값(extreme value)에 의해 영향을 많이 받음
중앙값(median) 중앙치, 중위수	자료를 크기순으로 나열했을 때 중앙에 위치하는 값
최빈값(mode) 최빈수, 최빈치	빈도가 가장 많은 관측값 자료에서 반드시 하나만 존재하는 것은 아님

※ 극단치/이상점(outlier) : 일반적인 기대를 벗어나서 나타나는 관측값

2) 산포도(degree of sccttering) : 자료의 흩어진 정도

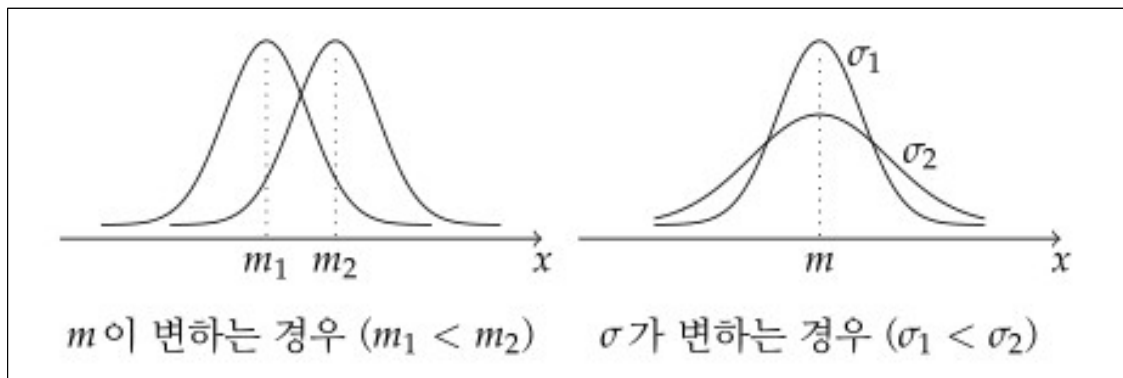
구분	설명
범위(range)	최대값 - 최소값
사분위수 범위 (InterQuartile Range)	자료를 오름차순으로 정렬했을 때, 상위 25%와 하위 25%를 제외하고 범위를 구한 값. 극단값의 영향을 크게 받지 않음 $IQR = Q_3 - Q_1$
사분위수 편차 (interquartile range)	오름차순으로 정렬했을 때, 3사분위수와 1사분위수의 평균 $Q = \frac{Q_3 - Q_1}{2}$
분산(variance)	관측값들이 평균으로부터 얼마나 떨어져 있는 정도(거리)의 제곱을 배제하기 위해 제곱 사용 $s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$
표준편차 (standard deviation)	분산의 양의 제곱근 관측치의 단위와 동일하므로 두 집단을 상대적으로 비교가능
표준오차 (standard error)	표본추출분포의 표준편차(표준편차를 자료의 수의 제곱근으로 나눈 값) 표본이 모집단에서 떨어져 있는 정도를 나타냄 → 표준오차가 작을수록 표본의 대표성이 높다고 할 수 있음 $S.E = \frac{s}{\sqrt{n}}$

※ 편차(deviation) : 실제 데이터의 값(관측값)에서 표본평균을 차감한 것. 편차의 합은 0

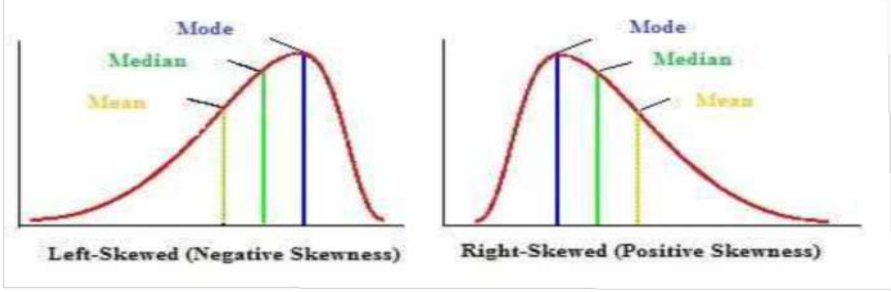
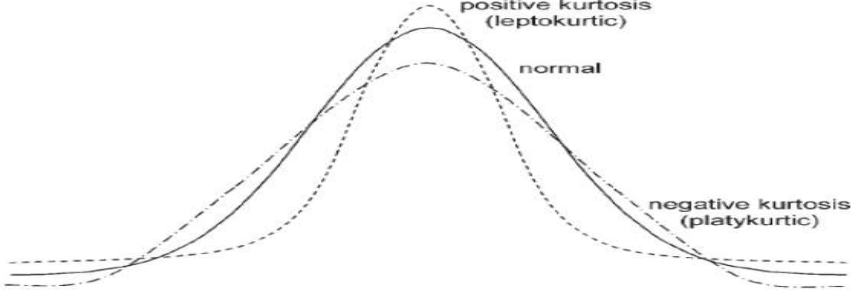
※ 오차(error) : 측정대상의 참값과 측정도구를 통한 측정값 사이의 불일치 정도

모집단으로부터 추정된 회귀식으로부터 얻은 예측값과 실제 관측값의 차이

※ 잔차(residual) : 표본으로 추정된 회귀식의 값(예측값)과 실제 데이터의 값(관측값)의 차이



3) 분포도 : 자료가 분포되어 있는 전체적인 모양 설명

구분	설명
왜도 (skewness)	<p>자료의 분포가 기울어진 방향과 정도. 분포의 비대칭을 의미</p>  <p>왜도<0, 오른쪽으로 치우친 분포 왜도=0, 비대칭성 정도가 정규분포와 유사한 분포 왜도>0, 왼쪽으로 치우친 분포</p>
첨도 (kurtosis)	<p>자료의 분포가 얼마나 중심에 집중되어 있는가를 나타냄. 분포의 뾰족한 정도</p>  <p>첨도<0, 상대적으로 평평한 분포 첨도=0, 뾰족한 정도가 정규분포와 유사한 분포 첨도>0, 상대적으로 뾰족한 분포 ※ 정규분포와 비교를 위해 첨도값에서 3을 뺀 값을 첨도로 사용하기도 함</p>

5. 가설

: 아직 검증되지 않은 추측적 예비 이론.

흔히 독립변수와 종속변수의 관계로 표현됨

종류	설명	예시
귀무가설 (H_0) null hypothesis 영가설	기각을 목적으로 수립되는 가설	성별에 따라 키 차이가 있다
대립가설 ($H_1 = H_a$) alternative hypothesis 연구가설, 채택가설	연구자가 제시한 (밝히고 싶은) 내용	성별에 따라 키 차이가 없다

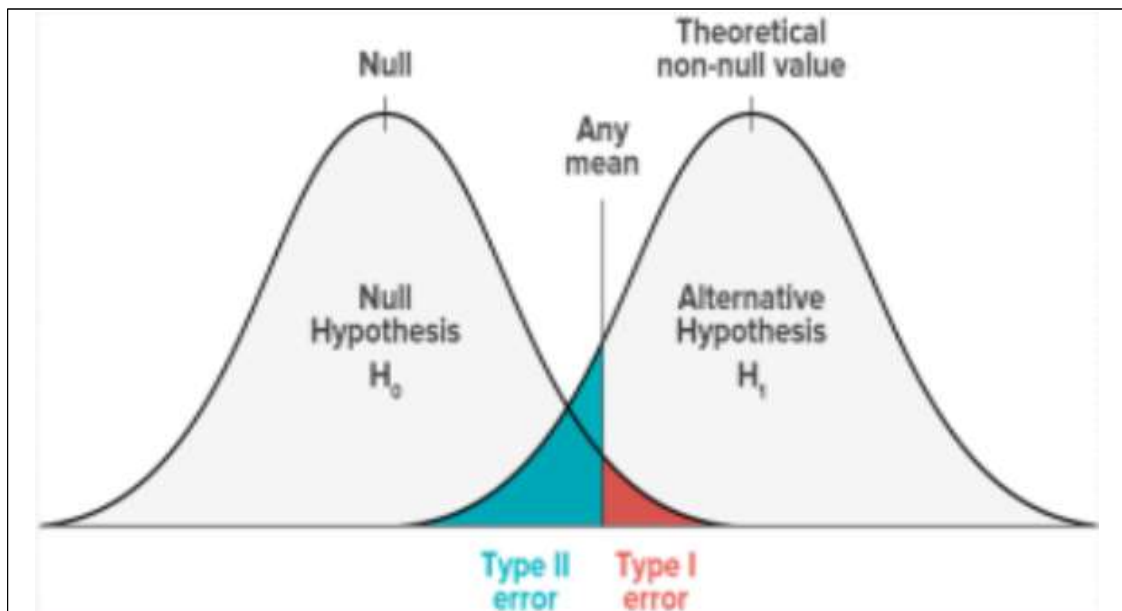
● 가설 검증 방식

- 1) 검정통계량의 값이 기각역과 채택역 중 어디에 위치하는지를 보는 방식
- 2) p-value를 계산하여 유의수준과 비교하는 방식

● 추계통계기법에서 사용되는 검정방법

평균검정(일표본/독립표본 평균분석)	Z-검정, t-검정
평균차이검정(대응표본 평균분석)	Z-검정, t-검정
분산분석	F-검정
상관분석	t-검정
회귀분석	F-검정, t-검정
χ^2 독립성검정(교차분석), χ^2 적합성검정	χ^2 -검정
판별분석	F-검정, χ^2 -검정

6. 유의수준



Ch.5 8페이지

- 제1종 오류: Type I error
 - 귀무가설이 참이지만, 검정 결과에 따라 귀무가설을 기각하는 오류(α)
- 제2종 오류: Type II error
 - 귀무가설이 거짓이지만, 검정 결과에 따라 귀무가설을 채택하는 오류(β)

가설검정에 따른 판단

		H_0 를 채택	H_0 를 기각
귀무가설의 실제 상황	H_0 가 참	신뢰수준 ($1 - \alpha$)	1종 오류 ($\alpha = \text{유의수준}$)
	H_0 가 거짓	2종 오류 (β)	검정력 ($1 - \beta$)

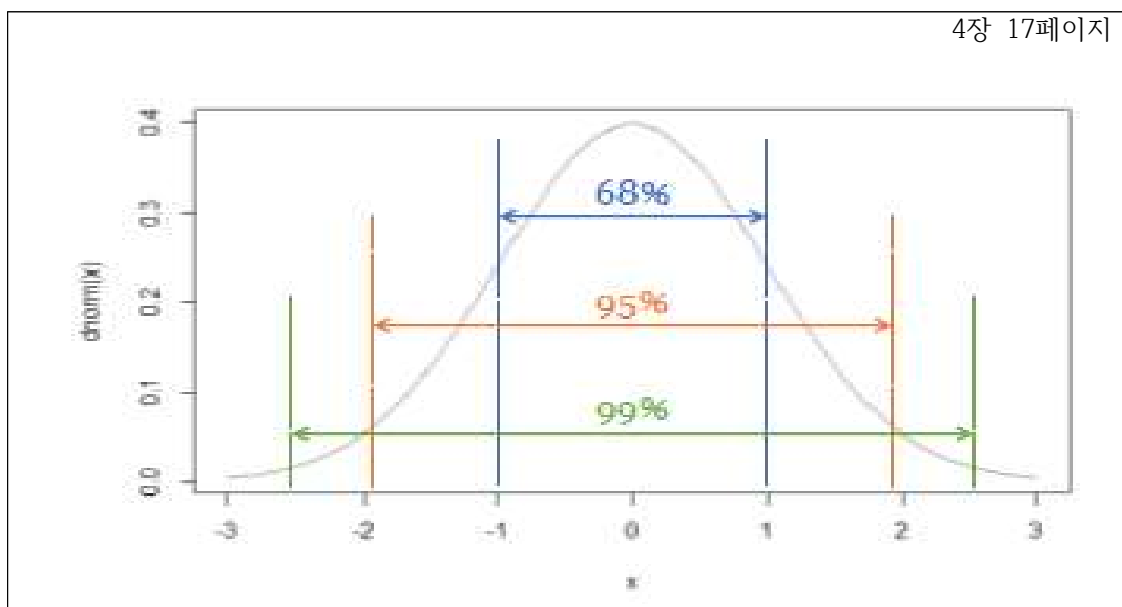
- 유의수준(α) : 1종 오류를 범할 통계적 확률
 - 유의확률($p-value$) : 귀무가설이 참일 때, 해당 값이 나올 확률
(표본에서 관측한 통계량보다 더 극단적인 값이 발생할 확률)
- $\Rightarrow \alpha$ 은 귀무가설의 기각 여부를 결정하는 데 사용하는 기준이 되는 확률
 $p-value$ 은 표본으로부터 계산된 확률.
- $\Rightarrow p-value < \alpha \rightarrow$ 귀무가설을 기각할 수 있는 증거가 충분함
 \rightarrow 대립가설을 채택할 수 있음
- ※ ‘통계적으로 유의하다’ 뜻 : 이러한 관측결과가 나타날 확률이 매우 낮다

7. 정규분포(normal distribution) : $N(m, \sigma^2)$

- 정규분포의 모양은 μ 와 σ 에 의해서 결정됨.
- 좌우 대칭인 종 모양 곡선(=가우스분포)
- 전체 면적(확률밀도함수, PDF, Probability density function)은 1이다

8. 표준정규분포(standard normal distribution)

- 정규분포의 개별값을 표준화한 값의 분포
- 평균이 0, 표준편차가 1인 정규분포 : $X \sim N(m, \sigma^2) \rightarrow Z = \frac{X-m}{\sigma}, Z \sim N(0, 1^2)$
- 표준화(standardization)를 하는 이유
: 평균과 표준편차가 다른 정규분포를 따르는 두 변수의 값을 비교



- 신뢰수준(confidence level)
: 모수가 추정한 구간 안에 있을 것이라 믿을 수 있는 정도(95%, 99%)
- 신뢰구간(confidence interval)
: 신뢰도에 따라 모수가 포함될 것이라 믿을 수 있는 구간

9. 자유도 (degrees of freedom)

자유로운 정도, n개에서 고정되는 한 값을 제외

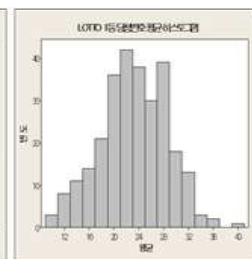
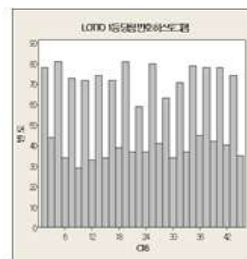
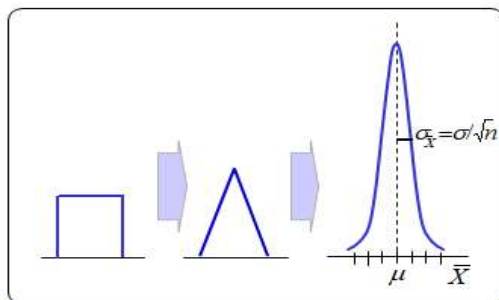
1, 9, 2, 8, 5	3, 4, 5, 6, 7	5, 5, 5, 5, 5
---------------	---------------	---------------

10. 중심 극한 정리

: 표본의 크기가 충분히 클 때($n \geq 30$),

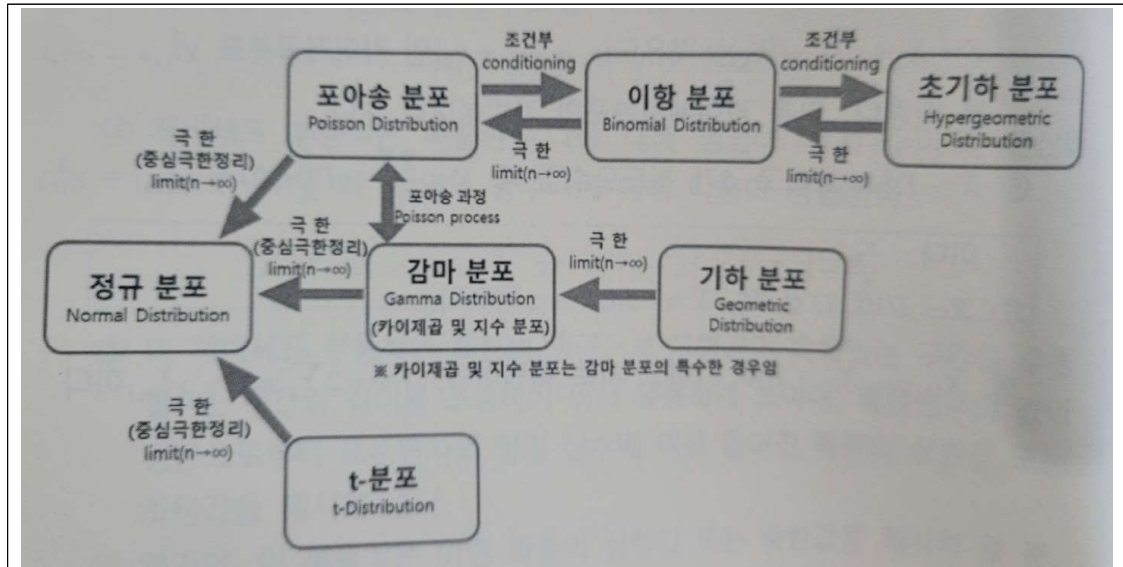
표본분포는 모집단의 분포와 상관없이 정규분포를 따른다.

- ☐ 모집단이 정규분포를 따르지 않는다고 해도 표본의 평균은 정규분포를 따른다.
- ☐ 표본 평균의 중심치 : 모평균 중심치와 동일
- ☐ 표본 평균의 표준편차 : $\sqrt{\sigma^2 / n} = \sigma / \sqrt{n}$
- ☐ 표본평균의 정규화 변환: $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$



11. 분포

1) 확률변수의 분포



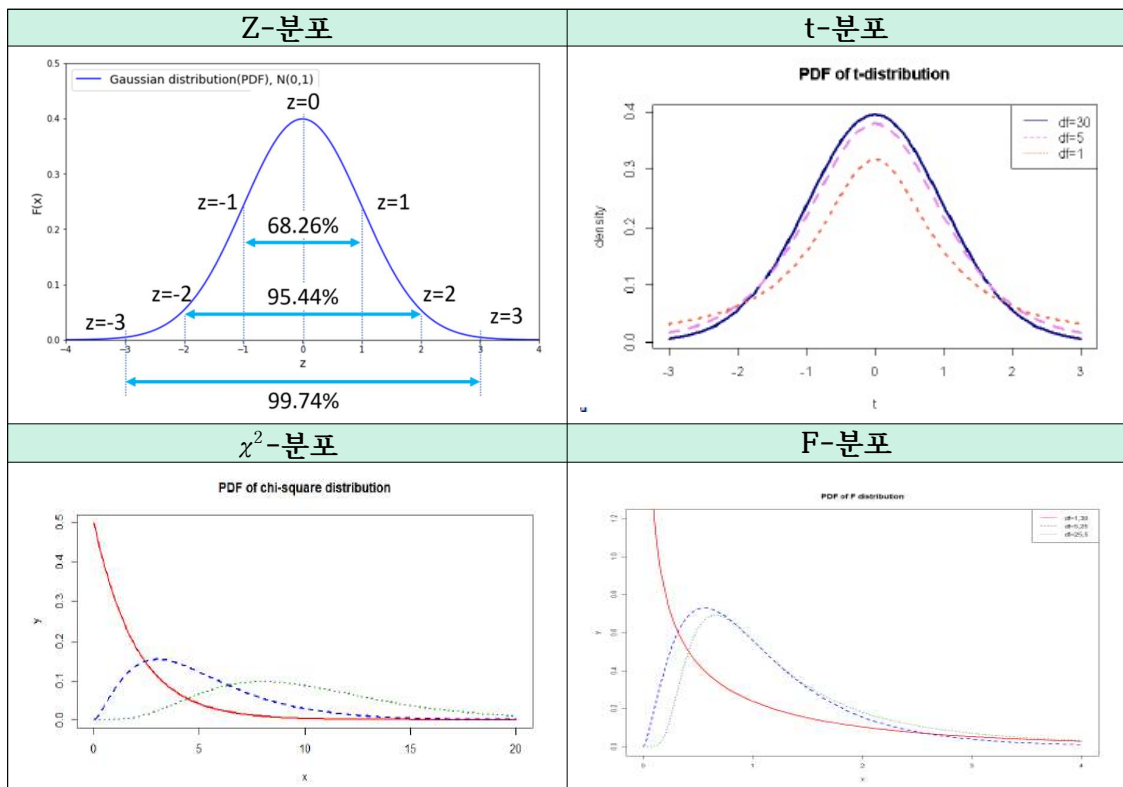
확률분포	평균	분산	비고
베르누이 시행	p	pq	성공, 실패만 존재
이항분포	np	npq	$np \geq 5, nq \geq 5$ 정규분포 근사 $p \leq 0.1, n \geq 50$ 포아송 근사
포아송분포	λ	λ	$\lambda \geq 5$ 정규분포 근사
기하분포	$\frac{1}{p}$	$\frac{q}{p^2}$	처음 성공할 때까지의 [시행]횟수
초기하분포	$\frac{nk}{N}$	$\frac{nk}{N} \frac{N-k}{N} \frac{N-n}{N-1}$	비복원 추출
균일분포	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	
정규분포	μ	σ^2	왜도 = 0, 첨도 = 3
표준정규분포	0	1	

2) 통계적 검정분포

확률분포	설명	
정규분포		
Z-분포	표준정규분포	
t-분포	정규분포를 따르는 모집단으로부터 추출한 표본의 확률분포 중 모양의 형태를 가지면서 표본크기에 따라 종 모양이 달라짐 표본의 크기가 충분히 커지면 t-분포와 정규분포가 거의 유사함	
F-분포	(집단의 개수-1), (표본개수-집단의 개수)	두 개의 자유도에 의해 분포의 모양이 결정 대체로 오른쪽으로 긴 꼬리를 가짐
χ^2 -분포	(행 개수-1), (열 개수-1)	

분산분석표(ANOVA)				
변동요인	SS	df	MS	F
요인	SSB	a-1	MSB=SSB/(a-1)	MSB/MSW
오차	SSW	N-a	MSW=SSW/(N-a)	
합계	SST	N-1		

● 분포별 모양



Q. 왜 t분포의 $df=29$ 와 정규분포가 일치하는가?

12. 변수 종류(수치형/범주형)별 통계분석방법(정규성 가정)

분석방법	적용분야	변수척도
빈도분석	가장 기초적이고 간단한 분석방법	모든 척도
교차분석 (카이제곱)	변수 간의 교차표 작성	명목척도, 서열척도
요인분석	<ul style="list-style-type: none"> 타당성 검정 설명력 부족한 변수 제거 	등간척도,비율척도
신뢰도분석	추출된 요인들의 동질적인 변수 구성	등간척도,비율척도
상관관계분석	측정변수들 간의 관계 정도를 제시	피어슨 - 등간척도, 비율척도 스피어만 - 서열척도
회귀분석	인과관계 분석	독립변수, 종속변수 : 등간척도/비율척도
t-검정	집단 간 평균 차이 검정	독립변수 : 명목척도 종속변수 : 등간척도 또는 비율척도
분산분석 (ANOVA)	3집단 이상의 평균 검정	독립변수 : 명목척도 종속변수 : 등간척도 또는 비율척도

독립변수(영향을 주는 변수)	종속변수(영향을 받는 변수)	분석 방법
범주형 변수	범주형 변수	카이제곱검정(교차표 분석)
	연속형 변수	t 검정 분산분석(집단이 세 개 이상)
연속형 변수	범주형 변수	로지스틱 회귀분석 다항 로지스틱 회귀분석
	연속형 변수	단순/다중회귀분석 구조방정식
연속형+범주형 변수	범주형 변수	로지스틱 회귀분석 의사결정나무
	연속형 변수	공분산분석(ANCOVA)

[출처]

이종환, 『조사방법론 및 SPSS 통계분석』, 공동체(2009)

이학식, 임지훈, 『SPSS 26 매뉴얼』, 집현재(2021)

사경환,

<https://blog.naver.com/restartq/222297096228>