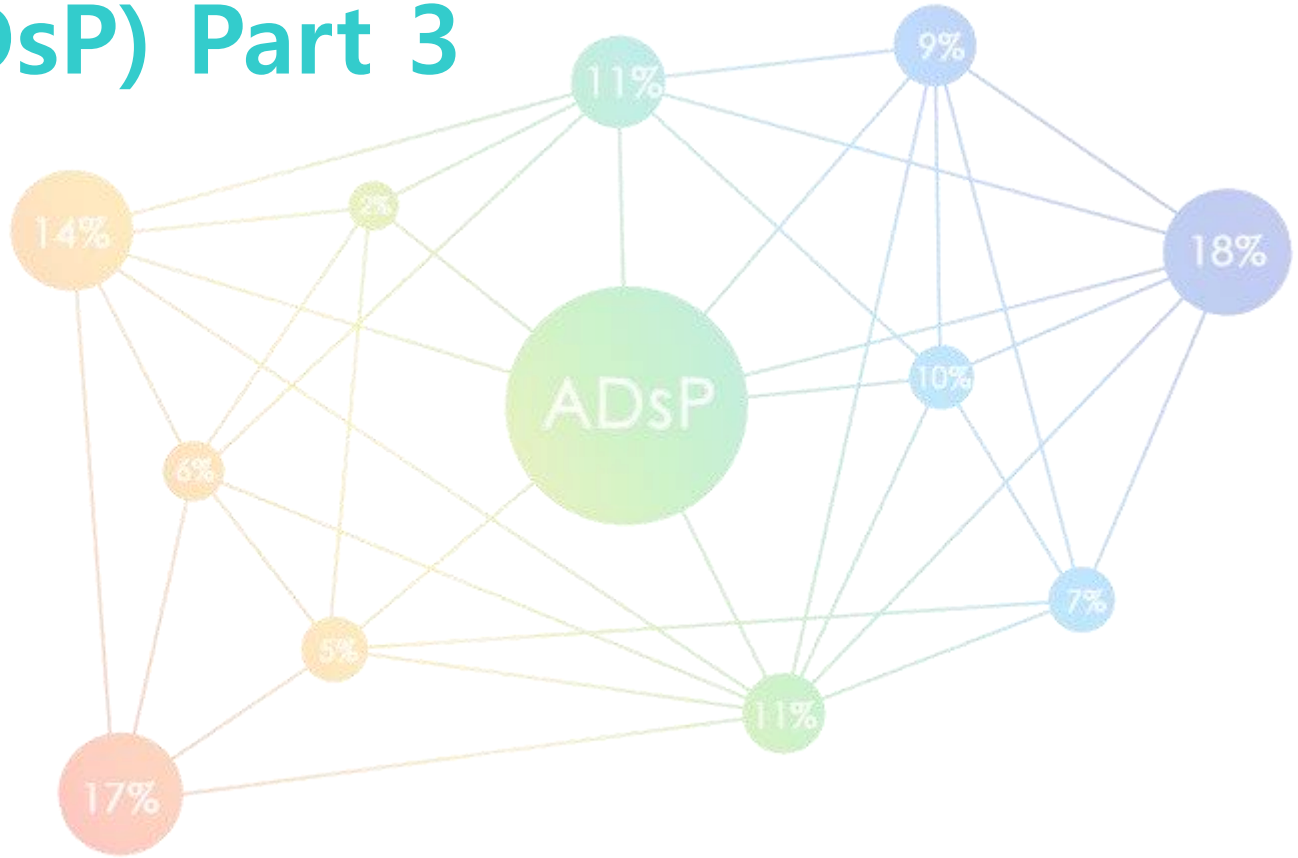
데이터분석준전문가(ADsP) Part 3

데이터분석



02 제2장 통계분석

제1절 통계학 개론

1. 통계분석 개요
2. 확률 및 확률분포
3. 표본의 분포
4. 점추정과 구간추정
5. 가설검정
6. 비모수 검정
7. 정규성검정과 데이터 타입에 따른 분석

1.통계분석 개요

통계학의 정의

통계학(Statistics)은 수량적인 비교를 기초로 많은 사실을 관찰하고
분석하는 방법을 연구하는 학문

→(목적) 신뢰성을 확보, 의사결정에 근거,문제해결을 위한 원인 파악 가능

기술통계(Descriptive Statistics) :

표본에 대한 분석 결과의 각종 수치들을 활용하여 집단의 특성을 설명

ex) 자료의 요약, EDA의 시각화

추론통계(Inference statistics) :

표본을 활용하여 모집단의 특성을 나타내는 것

ex)추정

1.통계분석 개요

(1)모집단과 표본

모집단

모집단(Population) : 통계분석 방법을 적용할 관심 대상의 전체 집합

표본

표본(Sample) : 과학적인 절차를 적용하여 모집단을 대표할 수 있는

일부를 추출하여 직접적인 조사 대상이 된 모집단의 일부

1.통계분석 개요

(2)모수 vs 통계량

모수(전수조사)

모수(Parameter) : 모집단을 분석하여 얻어지는 결과 수치

ex) 모평균, 모분산, 모표준편차, 모비율

통계량(표본조사)

통계량(Statistic) : 표본을 분석하여 얻어지는 결과 수치(통계치)

ex) 표본평균, 표본분산, 표본표준편차, 표본비율

1.통계분석 개요

모집단

모수

- 모평균(μ)
- 모분산(σ^2)
- 모표준편차(σ)
- 모비율(p)

표본

통계량

- 표준평균($\bar{x}$)
- 표본분산(s^2)
- 표본표준편차(s)
- 표본비율($\hat{p}$)

1.통계분석 개요

(3)표준추출 방법

모집단을 대상으로 조사하는 것

확률적 표본추출 방법(probability sampling method) :

표본추출의 방법은 동일한 확률 하에서 표본을 구성

- **단순 무작위(랜덤) 추출법**

모집단에서 일정한 규칙에 따라 표본을 기계적으로 추출하는 방법

ex) 난수 추출

1.통계분석 개요

계통(체계적)추출법

모집단에 번호를 부여하고 일정한 n개의 간격으로 표본을 추출하는 방법

군집(집락) 추출

모집단의 구성이 내부 이질적이면서 외부 동질적으로 구성되어 있다면,
모집단 전체를 조사하지 않고 몇 개의 군집을 표본으로
선택해서 조사하는 방법

층화 추출

모집단을 여러 개의 이질적 집단으로 구분한 후, 각 계층(집단)을
대표할 수 있는 표본을 추출하는 방법

1.통계분석 개요

비확률적 표본추출 방법(non-probability sampling method)

확률과는 상관없이 조사자가 자신의 의지로 표본 추출

편의 표본추출

조사자의 편의에 따라 시간이나 장소에 구애 받지 않고 임의적으로
표본을 추출하는 방법

할당 표본추출

조사자가 결정한 표본의 개수에 따라 임의적으로
표본을 추출하는 방법

1.통계분석 개요

(4)자료의 종류

○ 통계학의 자료

질적변수

- 명목형 자료 : 분류를 목적으로 사용하는 자료
- 순위형 자료 : 순서로 분류할 때 사용하는 자료

양적변수

- 이산형 자료 : 셀 수 있는 값을 나타낼 때 사용하는 자료
- 연속형 자료 : 측정 대상의 크기 변화가 연속적일 때 사용하는 자료 '절대영점'

1.통계분석 개요

변수 유형	자료 유형	인구주택총조사 결과	예
질적변수	명목형	성별, 배우자와의 관계	거주지역, 혈액형 등
	순서형	학력	학점(A,B,C)
양적변수	이산형	출생아 수	형제 수, 수강과목 수, 온도
	연속형	연령	키, 몸무게 등

2.확률 및 확률분포

용어

표본공간(sample space) : Ω

- 확률실험으로부터 출현 가능한 모든 결과들의 모임을 표본공간이라 합니다.
- 예제 : 동전던지기
- Ω = {앞면, 뒷면} = {H, T}

사건(Event) : 표본공간의 각 원소(즉, 출현 가능한 개별 결과)들의
부분집합을 사건이라 합니다.

- 근원사건: 어떤 사건이 표본공간상의 하나의 원소로 구성된 사건

ex) 동전던지기 근원사건 {H}와 {T}

2.확률 및 확률분포

(1)확률

아무리 정교하게 분석된 통계자료 일지라도 100% 맞을 수는 없기 때문에,

→ 그 결과를 확률(Probability)과 함께 표현

일정 조건 하에서 동일한 실험을 지속적으로 N 회 반복했을 때,

사건 A가 n번 발생할 확률은

$$P(A) = \frac{n(A)}{N}$$

2.확률 및 확률분포

확률이 가지는 조건

- 확률은 0~1의 값을 가진다
- 모든 사건에 대한 확률의 합은 1이다.

$$\sum_{i=1}^{n} P(E_i) = 1 \quad (E : \text{사건(Event)},\ i : \text{시행 횟수},\ P : \text{확률})$$

2.확률 및 확률분포

(2)확률법칙

덧셈법칙

- 임의의 사건 A와 사건 B의 합사건에 대한 확률

$P(A \cup B)=P(A)+P(B)-P(A \cap B)$

만일 두 사건 A와 B가 서로 배반이라면 $(A \cap B=\phi)$

$P(A \cup B)=P(A)+P(B)$

2.확률 및 확률분포

곱셈법칙

- **조건부 확률**

  두 사건 A와 B에 대해

  P(A | B) : 사건 B가 발생했을 때 사건 A가 발생할 확률

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \qquad P(B) > 0$$

  P(B | A) : 사건 A가 발생했을 때 사건 B가 발생할 확률

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \qquad P(A) > 0$$

2.확률 및 확률분포

- **조건부 확률 예)**

  주사위를 던지는 실험에서 주사위의 눈이 짝수인 사건을 A,

  주사위의 눈이 4이상인 사건을 B라 할 때 P(A|B)를 구해봅시다.

  $P(A|B) = \frac{P(A \cap B)}{P(B)}$ 이므로 P(B)와 $P(A \cap B)$를 구합니다.

  P(B) : 주사위의 눈이 4이상이 나올 확률 = 1 / 2

  $P(A \cap B)$: 주사위의 눈이 짝수이고 4이상인 경우는

  {4, 6} 이므로 확률은 1/3 입니다.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/3}{1/2} = \frac{2}{3}$$

2.확률 및 확률분포

- **곱셈법칙**

  두 사건 A와 B에 대해 조건부 확률을 이용하여

$$P(A \cap B) = \begin{cases} P(A)P(A|B), & P(A)>0 \\ P(B)P(B|A), & P(B)>0 \end{cases}$$

  만일 두 사건 A와 B가 독립이라면

$$P(A \cap B) = P(A)P(B)$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

2.확률 및 확률분포

(3)확률변수와 확률함수

○ 확률변수(random variable)

실험의 결과(사건)에 수치로 대응시킨 값

이산 확률변수(discrete random variable) :

수집된 데이터의 확률변수 중 셀 수 있는 특정한 값들로 구성되거나

일정한 범위로 나타나는 확률변수

연속 확률변수(discrete random variable) :

연속형 이거나 무한한 경우와 같이 셀 수 없는 확률변수

2.확률 및 확률분포

확률함수(probability function)

반복적으로 어떤 실험을 할 때, 각각의 실험 결과가 어떨지는
그 순간에는 알 수 없지만, 실험을 다 마친 후에는 어떤 결과가
몇 번씩 발생했는지를 총체적으로 살펴볼 수 있는데,
이 결과의 수에 확률이 부여된 것

2.확률 및 확률분포

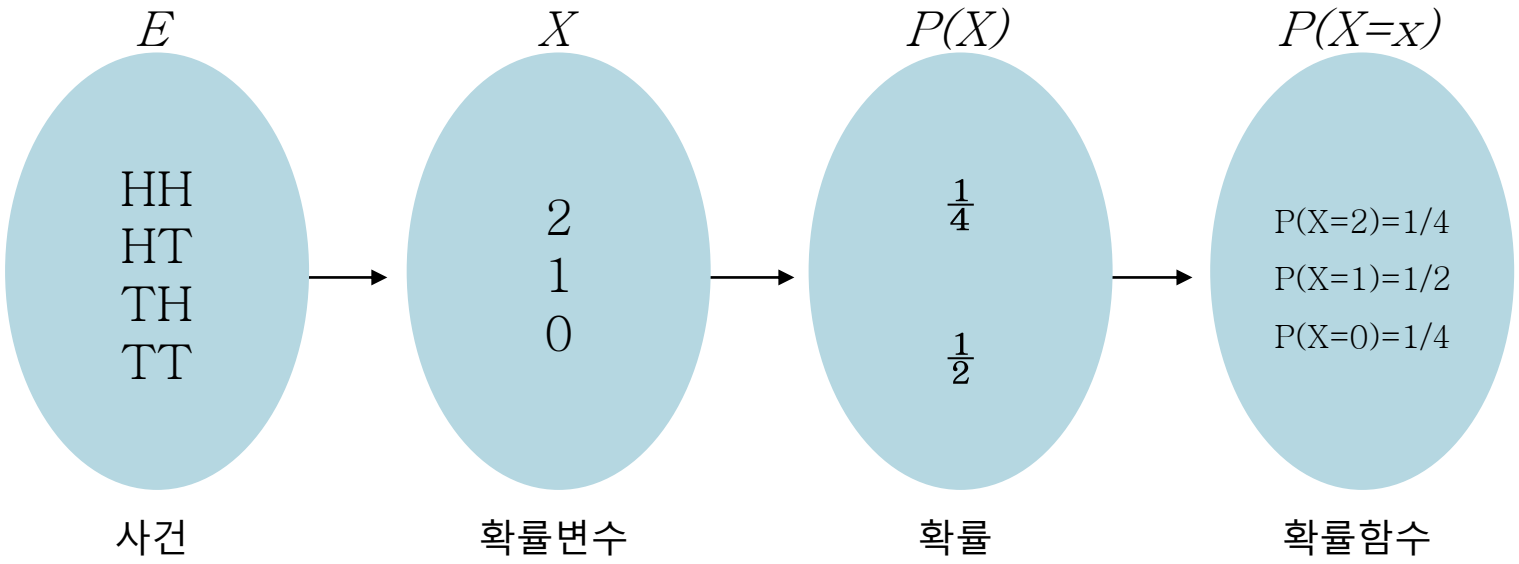


그림) 동전 던지기의 사건, 확률변수, 확률, 확률함수의 관계

2.확률 및 확률분포

(4)확률변수의 기대값

기대값 (expected value)

어떤 사건에 대해 그 사건이 벌어질 확률을 곱해서 전체 사건에 대해 합한 값

기대값 $E(X) = \sum x P(x)$

ex) 주사위를 던졌을 때의 기대값

$$1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6}$$

$$= \frac{1+2+3+4+5+6}{6} = 3.5$$

2.확률 및 확률분포

(5)확률변수의 분산

확률변수의 분산은 기대값의 특성을 나타내는 값으로,
확률변수들이 기대값으로부터 벗어나는 정도를 나타낸다.

→ (평균으로부터 산포되어 있는 정도를 분산이라 한 것과 같이)

확률에서 분산은 기대값과 어느 정도 차이가 있는지를 나타낸다.

$$Var(X) = E(X-\mu)^2 = \sum (X-\mu)^2 P(X)$$

2.확률 및 확률분포

(6)이산형 확률분포 vs 연속형 확률분포

확률분포

확률변수가 취할 수 있는 값과 각 값이 나타날 확률을 대응시킨 관계

이산 확률분포: 베르누이 분포,이항분포,초기화분포, 포아송분포

ex)정수로 구분할 수 있는 경우 이산형 ex)안경 쓴 학생수

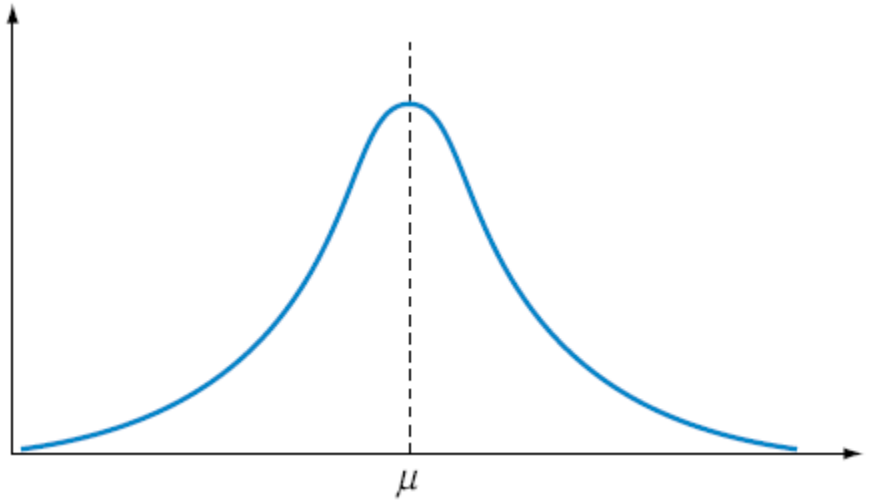
연속 확률분포: 균등분포,정규분포,t분포, 카이제곱분포, F분포

Ex)정수로 셀 수 없기 때문에 일정구간 설정ex)키170~171 학생수

3 표본의 분포

표본을 추출한 후, 표본의 특성을 파악하기 위해 표본분포의 확인이 필요

정규분포: 중심(평균)을 기준으로 좌우가 대칭되는 분포



3.표본의 분포

■ z분포

표본의 개수가 충분할 때 표준화 과정을 거친 정규분포를 표준정규분포(standard normal distribution), 혹은 z분포라고 한다.

→ 표준정규분포는 '평균=0, 분산=1'인 정규분포를 따른다.

$$z = \frac{X - \mu}{\sigma / \sqrt{n}}$$

$(X : 측정치,\ \mu : 평균,\ \sigma/\sqrt{n} : 표준오차)$

중심극한정리(Central Limit Theorem : CLT)

표본의 개수(n)가 충분하다면 모수를 모르는

상황에서도 표본 통계량으로 정규분포를 구성하여 모수를 추정할 수 있다는 것이다.

3.표본의 분포

■ t분포

표본이 충분하지 못한 경우, 즉 표본의 개수가 30개를 넘지 못하는 경우에는 t분포를 사용

$$t_{n-1} = \frac{X-\mu}{s/\sqrt{n}}$$

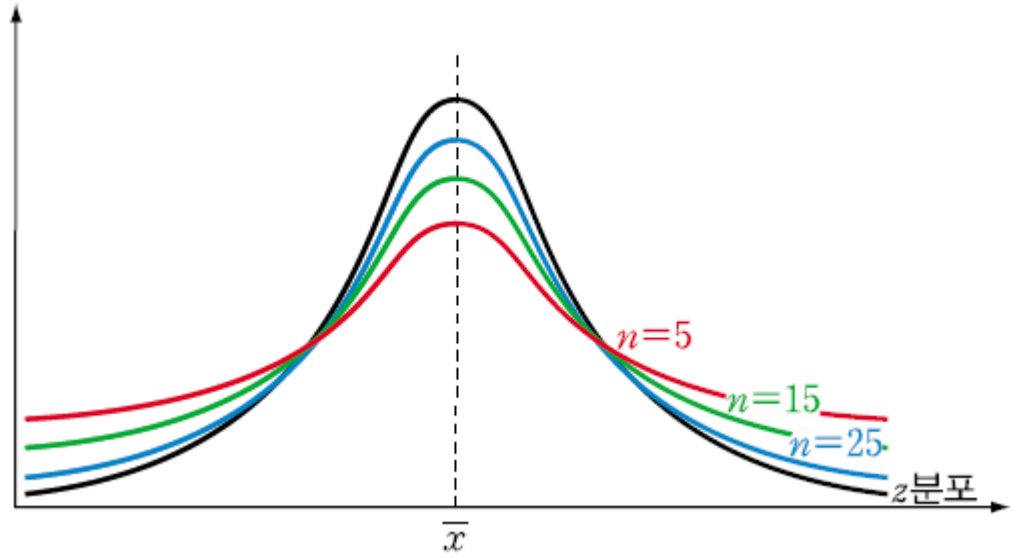
(X : 측정치, μ : 평균, $s/\sqrt{n}$: 표준오차, $n-1$: 자유도)

3.표본의 분포

■ Z분포와 t분포의 관계

$z = \frac{X-\mu}{\sigma/\sqrt{n}}$ 와 $t_{n-1} = \frac{X-\mu}{s/\sqrt{n}}$ 는 n과 $n-1$을 제외하고 식이 동일

$n \to \infty$ 면 두 분포는 동일한 분포

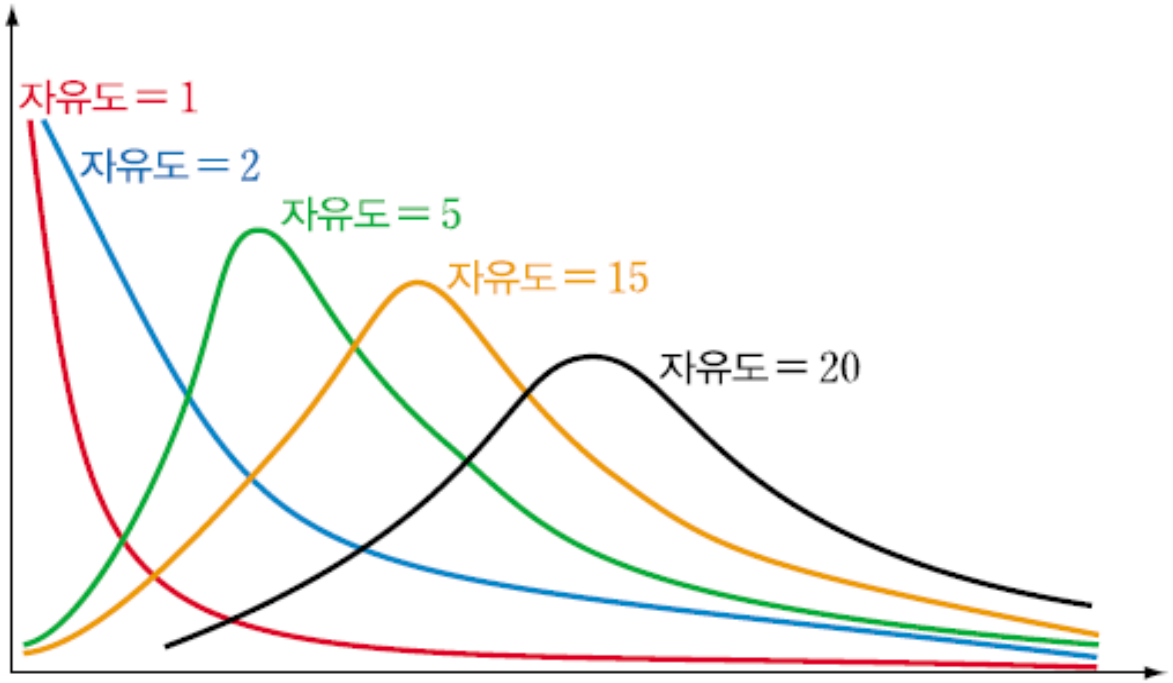


3.표본의 분포

■ χ^2분포

χ^2분포는 정규분포로부터 도출되고, z분포의 제곱에 대한 분포

∴ 항상 0보다 큰 값



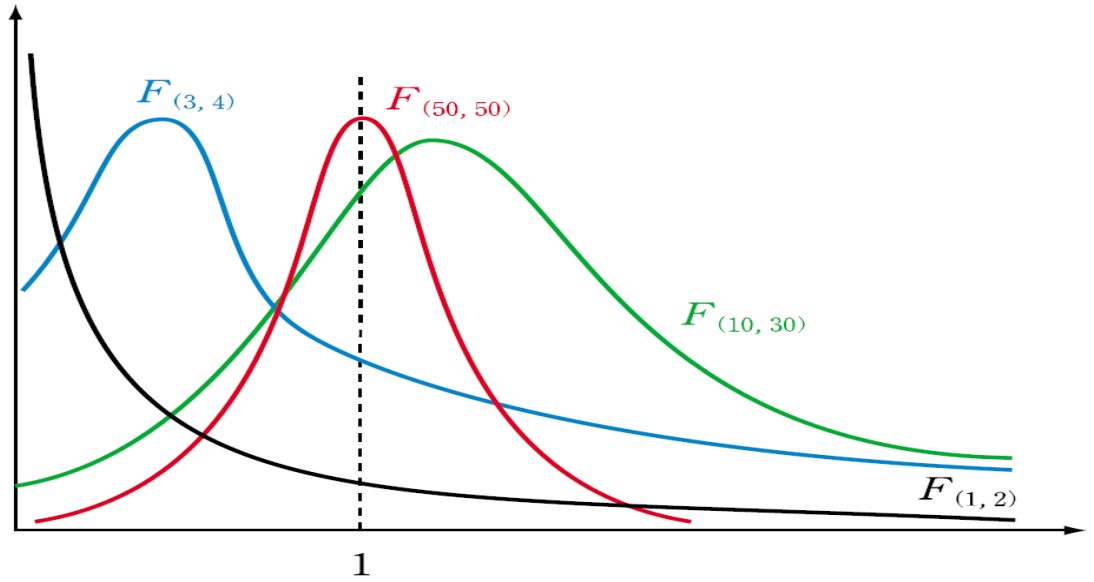
3.표본의 분포

■ F분포

F분포는 두 개의 분산에 관한 추론 → F(v_1, v_2)

$\therefore$ v_1, v_2는 각각의 X^2에 대한 분산

$F = \frac{v_1}{v_2}$ 는 각 비율을 나타냄



4.추정과 구간추정

(1)점추정이란

점추정(point estimation)

점추정은 모수를 특정한 수치로 표현하는 것

ex)통학 시간에 대해 점추정은 30분, 40분과 같이 특정한 수치로 표현

구간추정(interval estimation)

구간추정은 모수를 최소값과 최대값의 범위로 추정하는 것

ex) 통학 시간에 대해 구간추정은 30분~40분과 같이 범위로 표현

4.추정과 구간추정

추정이란

- **추정치(estimate)**

  모수를 추정하기 위해 선택된 표본을 대상으로 구체적으로 도출된 통계량

- **추정량(estimator)**

  표본에서 관찰된 값으로 추정치를 계산하기 위한 도출 함수

4.추정과 구간추정

점 추정과 바람직한 점 추정량의 조건

○ 바람직한 점 추정량 조건

01 불편성 : 추정량이 모수와 같아야 한다.

추정량 $\hat{\theta}$로 모수 θ를 추정하여 $E(\hat{\theta}) = \theta$가 되면 가장 바람직한 추정

이때의 추정량을 불편 추정량(unbiased estimator)

4.추정과 구간추정

02 일치성 : 표본의 크기가 모집단 규모에 근접해야 한다.

일치성(consistency)은 표본이 모집단의 규모에 근접할수록 오차가 작아진다는 의미

03 유효성 : 추정량의 분산이 최소값이어야 한다.

유효성(efficiency)은 모수에 대한 추정량의 분산(분포)이 작을수록 추정량이 바람직하다는 의미

04 충분성 : 표본이 모집단의 대표성을 가져야 한다.

표본은 모집단에 대해 대표성을 가져야 통계적인 의미가 있으므로,

4.추정과 구간추정

(2)구간추정

○ 구간추정(interval estimation)

구간추정은 모수를 최소값과 최대값의 범위로 추정하는 것

○ 구간추정을 사용하는 이유

조사자의 입장에서 오차를 줄이기 위하여 명확한 수치를 제시하는 점 추정 대신 신뢰도를 제시하면서 상한값과 하한값으로 모수를 추정하는 구간추정을 사용

4.추정과 구간추정

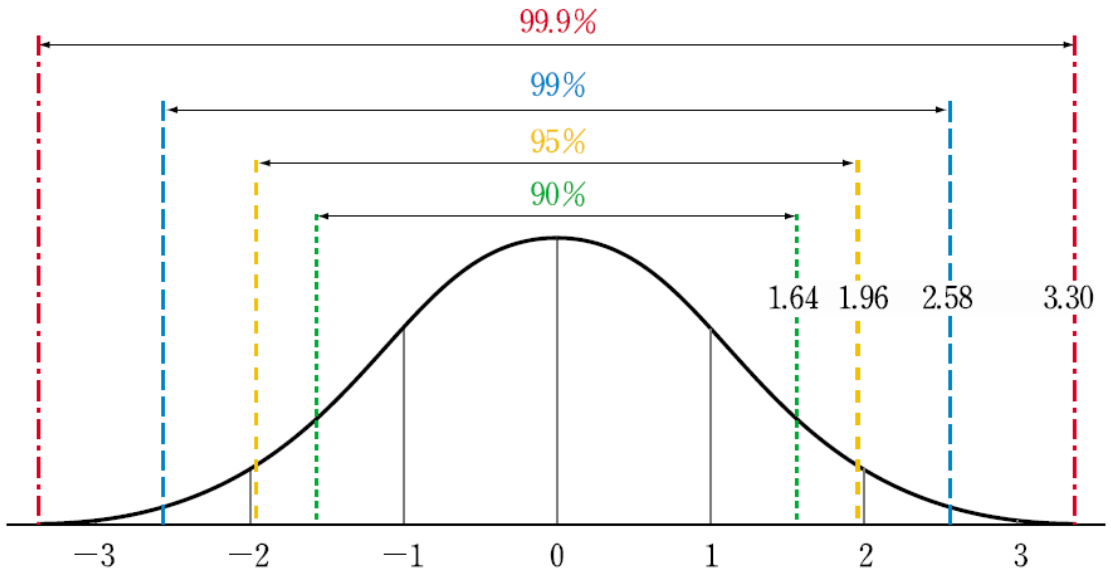
신뢰구간

상한값과 하한값의 구간으로 표시되며, 신뢰수준을 기준으로 추정된 점으로부터 음(-)의 방향과 양(+)의 방향으로 하한과 상한을 표시

모평균(μ)을 추정할 때 표본평균을 $\overline{X}$
표준오차를 SE
표본평균 $\overline{X}$

모집단 평균에 대한 신뢰구간

$$\overline{X} - z \cdot SE \leq \mu \leq \overline{X} + z \cdot SE$$



4.추정과 구간추정

Tip

표준오차

표준오차는 평균들 간의 분포를 나타내므로 표준오차가 줄어들수록 평균을 나타내는 점들이 집중적으로 모여 있다. 따라서 모수의 추정이 정확하게 이루어졌음을 판단할 수 있다. 평균의 표준오차는 다름과 같이 구한다.

모분산을 알 경우 : $S.E = \frac{\sigma}{\sqrt{n}}$

모분산을 모를 경우 : $S.E = \frac{s}{\sqrt{n}}$

4.추정과 구간추정

신뢰도 90%에서의 구간추정

$\overline{X} = 500,\ SE = 100,\ z = 1.64$ 이므로 다음과 같이 구할 수 있다.

$$500 - 1.64 \cdot 100 \leq \mu \leq 500 + 1.64 \cdot 100$$

$$336 \leq \mu \leq 664$$

신뢰도 95%에서의 구간추정

$\overline{X} = 500,\ SE = 100,\ z = 1.96$ 이므로 다음과 같이 구할 수 있다.

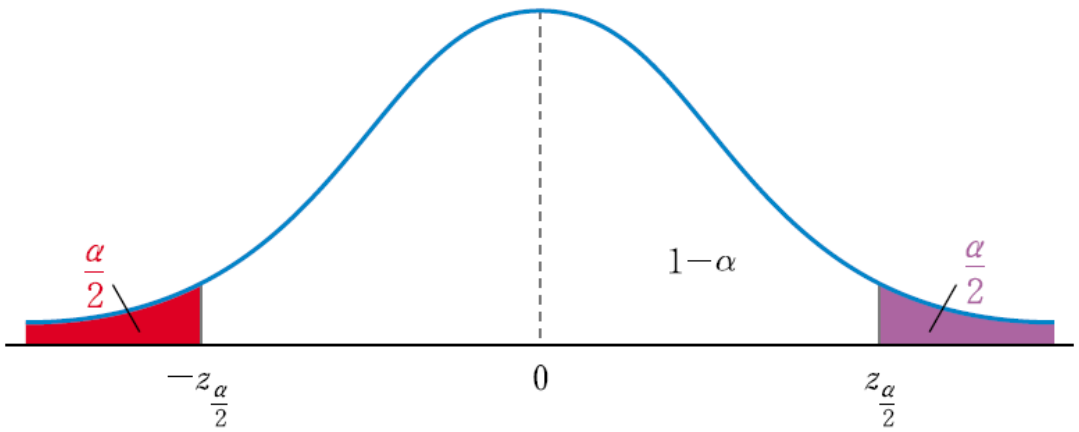
$$500 - 1.96 \cdot 100 \leq \mu \leq 500 + 1.96 \cdot 100$$

$$304 \leq \mu \leq 696$$

4.추정과 구간추정

모평균의 신뢰구간 (모집단의 표준편차를 아는 경우)

$$100(1-\alpha)\% = P\left(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}\right)$$



모집단의 표준편차를 알고 있으므로 평균이 μ, 표준오차가 $\frac{\sigma}{\sqrt{n}}$ 인 정규분포를 이룬 다고 가정하면, 신뢰구간은

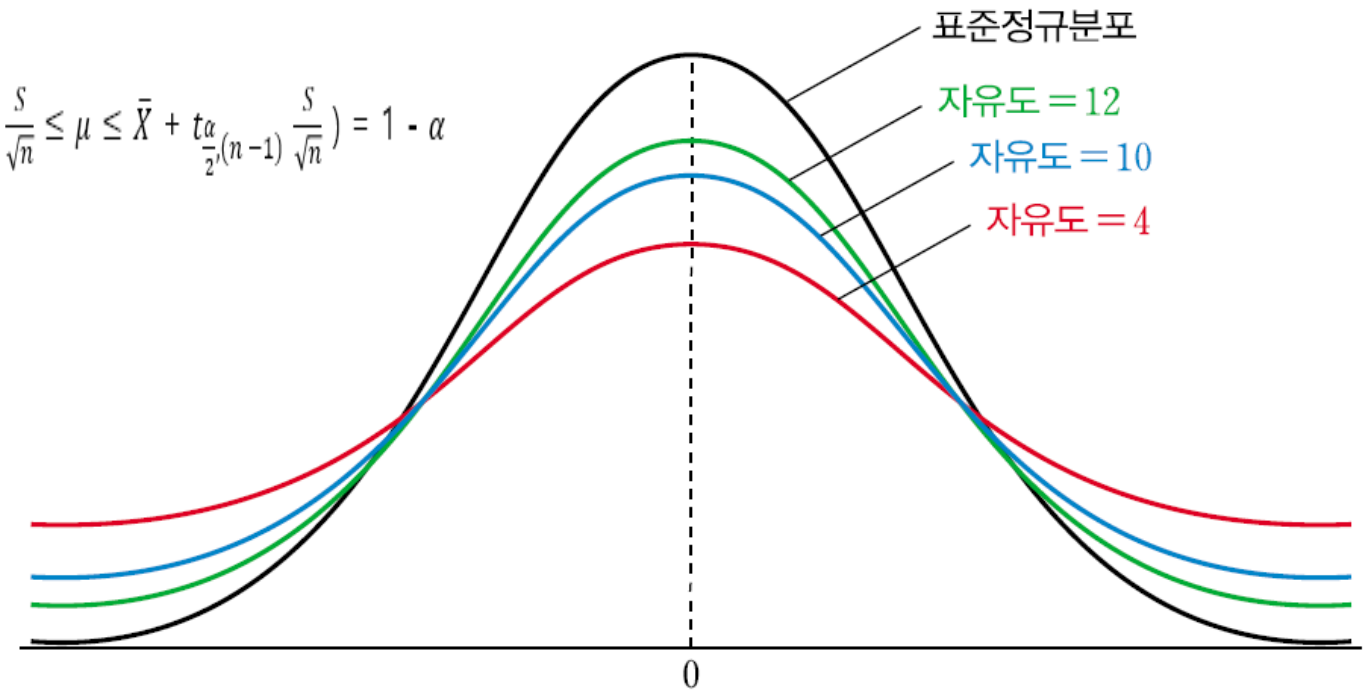
$$\overline{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

4.추정과 구간추정

모평균의 신뢰구간 (모집단의 표준편차를 모르는 경우)

표본의 표준편차를 이용한 신뢰구간은 모표준편차를 이용한 신뢰구간보다 틀릴 가능성이 더 크므로, 신뢰구간의 범위가 더 커질 수밖에 없으므로 t분포를 이용 t분포는 모수를 알지 못한 상황에서 정규분포를 이루는 모집단에서 추출한 표본의 크기가 작을 때의 추정과 검정에 사용 특히 t분포는 자유도에 따라 서로 다른 분포를 가짐.

$$P\left(\bar{X} - t_{\frac{\alpha}{2},(n-1)}\frac{S}{\sqrt{n}} \le \mu \le \bar{X} + t_{\frac{\alpha}{2},(n-1)}\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$



5.가설검정

가설(hypothesis)

주어진 사실 혹은 조사하고자 하는 사실이 어떠하다는 주장이나 추측

→ 모수를 추정할 때, 모수가 어떠하다(혹은 어떠할 것이다)는 조사자의 주장이나 추측을 일컬음

귀무가설(null hypothesis)

귀무가설은 일반적으로 믿어왔던 사실을 가설로 설정한 것으로,

영가설(零假設)이라고도 하며, 영(零,0) 이라는 의미로 H_0 로 표기

→ 귀무가설에 대한 조사는 의미가 없다고 볼 수 있다.

∵ 연구를 하더라도 일반적으로 모두 인정하고 받아들이는 사실이므로

어떤 의미를 찾아내기 어렵기 때문

ex) 스포츠 이온음료의 용량은 제품에 표기된 300ml가 맞는지에 대한 조사

H_0 : 스포츠 이온음료의 용량은 제품에 표기된 300ml가 맞다.

5.가설검정

대립 가설(antihypothesis)

대립가설은 공공연하게 사실로 받아들여진 현상에 대립되는 가설로, 일반적으로 연구를 통한 대립가설의 조사는 의미가 있다고 받아들인다.

대립가설은 연구가설이라고도 하며, 영(零,0)에 반대가 된다는 의미로 H_1로 표기

ex) 스포츠 이온음료의 용량이 제품에 표기된 300ml가 맞는지에 대한 조사

→ 귀무가설과 대립하여, 스포츠 이온음료의 용량이 300ml라고 표기된 것이 사실이 아니라고 설명하면 됨.

H_0 : 스포츠 이온음료의 용량은 제품에 표기된 300ml가 아니다.

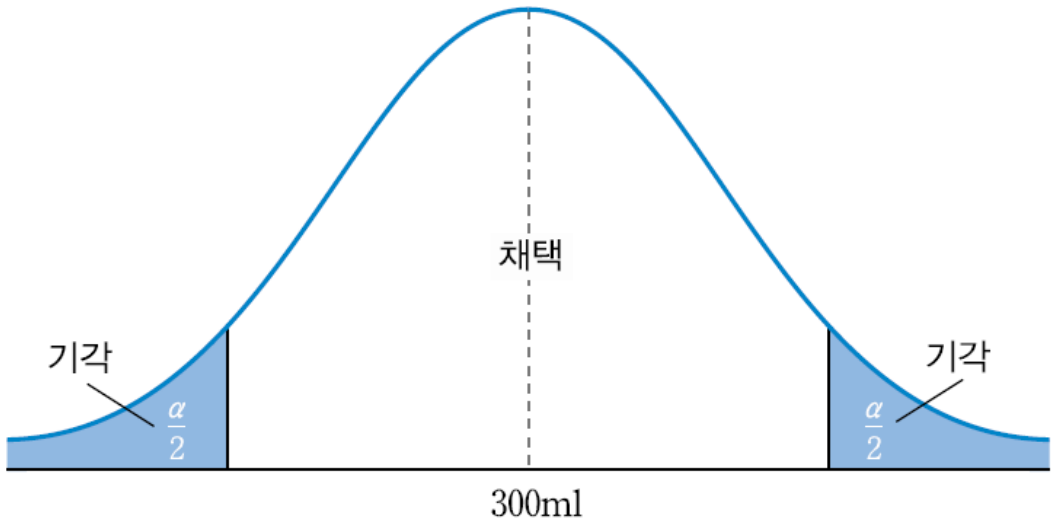
5.가설검정

양측검정(two-sided test)

조사하고자 하는 대립가설, 즉 '사실이 아니다'라는 것을 검정하여
귀무가설을 기각하고 대립가설을 채택하고자 하는 것

$$H_0 : \mu = 300\,\mathrm{ml}$$

$$H_1 : \mu \neq 300\,\mathrm{ml}$$



5.가설검정

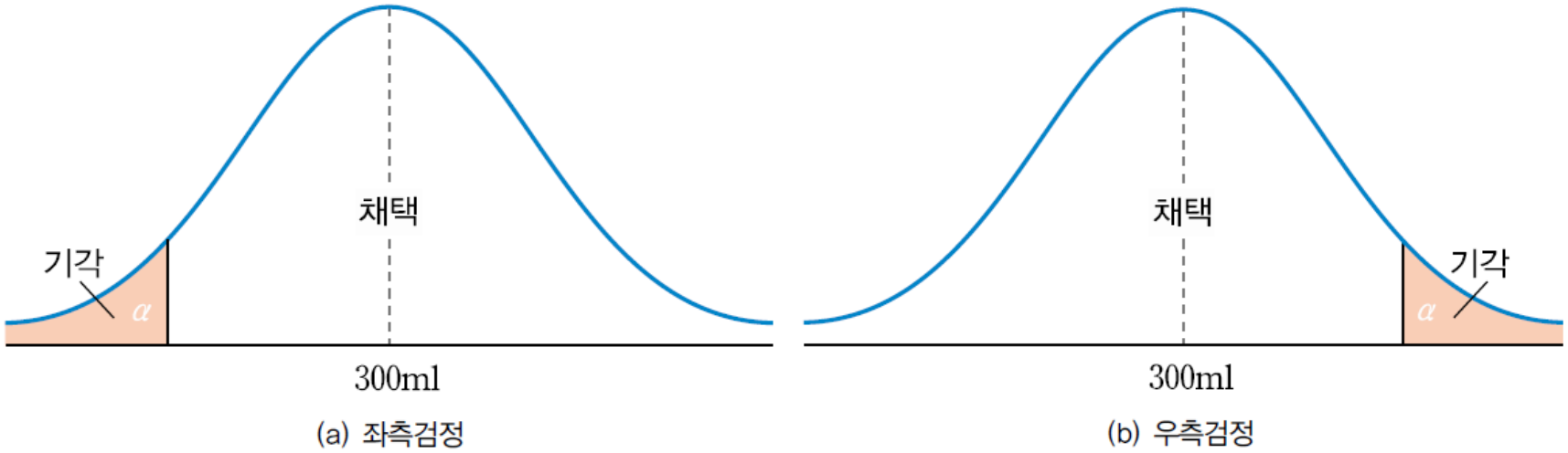
단측검정(one-sided test)

조사의 목적에 따라 대립가설을 스포츠 이온음료의 용량이 300ml보다 적다고 수립하거나, 혹은 300ml보다 많다고 수립하여 한 쪽만 살펴보는 것

$H_0: \mu = 300\,\mathrm{ml}$
$H_1: \mu < 300\,\mathrm{ml}$

혹은

$H_0: \mu = 300\,\mathrm{ml}$
$H_1: \mu > 300\,\mathrm{ml}$



5.가설검정

유의수준과 통계적 오류

- **통계적 판단**
  → 모수를 추정한다는 의미

- **추정은 틀릴 가능성을 내포**
  → 모수를 추정할 때는 항상 오류 가능성(확률)을 제시

  - 통계학에서는 모수의 추정이 맞을 확률을 1-α로 표시
    α는 유의수준(significance level)으로,
    확률(prpbability)로 표시되므로 약자를
    사용하여 p값($p-value$)으로 표시

5.가설검정

오류

모수를 추정한 결과가 실제와는 다른 결론에 도달하는 것

1종 오류(type Ⅰ error)와 2종 오류(typeⅡ error)

→ 1종 오류: 귀무가설을 채택해야 함에도 귀무가설을 기각하는 경우

→ 2종 오류: 귀무사설을 기각해야 함에도 귀무가설을 채택하는 경우

5.가설검정

○ 가설검정의 검정력(power of hypothesis testing)

귀무가설을 채택해야 하지만 귀무가설을 기각하는 경우의 확률은

→ 유의수준 α로 표시

귀무가설을 기각해야 함에도 귀무가설을 채택하는 경우의 확률은

→ β로 표시

∴ 기각해야 할 귀무가설을 기각하는 확률은 $1-\beta$가 되는데,

이를 2종 오류가 발생하지 않을 확률

이를 가설검정의 검정력(power of hypothesis testing)이라 함.

5.가설검정

가설검정의 절차

01 가설수립

귀무가설 H_0와 대립가설 H_1을 수립한다.

02 유의수준 결정가설수립

수립된 귀무가설과 대립가설 중 어떤 가설을 채택할 것인지
판단하는 유의수준 a를 결정한다.

03 기각역 설정 유의수준 결정가설수립

조사의 성격에 따라 양측검정을 할 것인지
단측검정을 할 것인지를 정해 기각역을 설정한다.

5.가설검정

04 통계량의 계산

수집된 표본을 대상으로 조사에 필요한 통계량을 계산한 후 기각역과 비교한다.

05 의사결정

가각역과 비교한 후에 귀무가설과 대립가설 중 어떤 가설을

채택할 것인지를 결정한다.

5.가설검정

모분산을 모르는 경우의 가설검정

○ 표본이 큰 경우의 가설검정

모수를 모르더라도 표본이 아주 큰 경우, 즉 $n \longrightarrow \infty$ 이면 모집단의 정규성과 상관없이

중심극한정리에 의해 s^2 은 σ^2 으로 수렴 그러므로 표본이 큰 경우의

검정통계량은 모분산이 주어진 것과 동일하게 $z = \frac{x-\mu}{\sigma/\sqrt{n}}$ 로 계산

○ 표본이 작은 경우의 가설검정

표본이 충분하지 못한 경우는

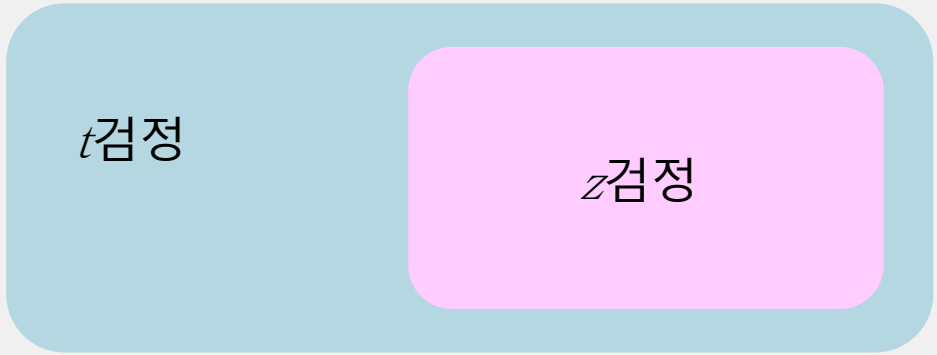
표본통계량 $t_{n-1} = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ 는 자유도 $(n-1)$ 에서 t분포를 따르게 된다.

5.가설검정

Tip

t검정과 z검정의 관계

통계학을 학습할 때 검정 방법이 다양하여 혼란스러울 수 있다.
t검정과 z검정의 관계만 보더라도 검정통계량을 구하는 공식은 비슷한데 도축 결과가 다르다. 이 두 가지 검정 방법은 중심극한정리에 의해 구분되는 것으로 판단하면 이해하기 쉽다.
즉 표본이 30보다 적으면 t분포를 이용하고, 그보다 많으면 z분포를 이용한다. 다만 t분포표를 보면 알 수 있듯이 t분포는 30보다 적은 표본의 분포를 나타내는 동시에 30보다 많은 표본의 분포도 포함한다. 다시 말해, z분포는 t분포에 포함된다고 할 수 있다.

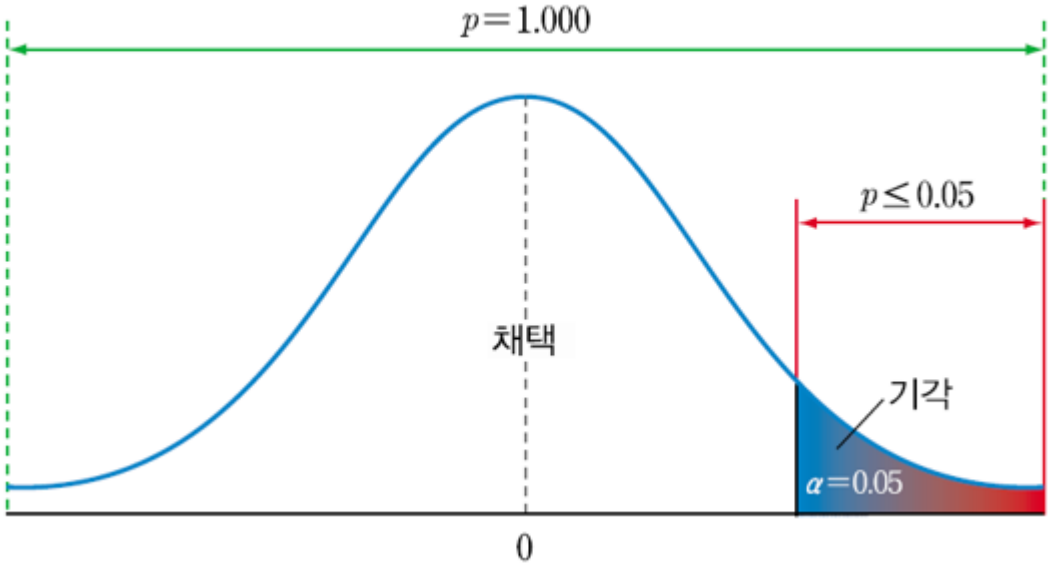


5.가설검정

p값을 이용한 가설검정

p값은 귀무가설을 기각하기 위한 최대한의 한계점을 나타내는데,

유의수준 α를 기준으로 보면 α로부터 멀리 떨어져 있는 확률을 나타낸다.



p값이 0.05.보다 작거나 같다면 귀무가설을 기각

6.비모수 검정

비모수 검정

- 비모수적 검정
  - ① 모집단이 데이터가 정규분포, 등분산 가정 따르고 있지 않을 때
  - ② 이상치를 포함한 분석을 실시하고 싶을 때
  - ③ 얻어진 데이터가 순서척도 일때

6.비모수 방법

모수적 방법	비모수 방법
두 개의 모평균에 대한 t검정	Mann-Whitney 검정
	중위수 검정
대응이 있는 t 검정	Wilcoxon 부호순위감정
	부호검정
일원배치 분산분석	Kurskal-Wallis 검정
	중위수 검정
Pearson 상관계수	Spearman의 순위상관계수
	Kendall의 순위상관계수

6. 비모수 검정

ex)12개 치즈 있고, 12개 중 6개는 상표 A, B있음. A와B에세 무게의 평균치 차가 있는지

상표	A	A	A	A	A	A	B	B	B	B	B	B
무게	5.2	6.3	7.2	8.4	9.5	9.8	5.1	6.6	7.3	8.7	9.1	9.3

A 평균치 : 7.733 B평균치 : 7.733

순위	1	2	3	4	5	6	7	8	9	10	11	12
상표	B	A	A	B	A	B	A	B	B	A	B	A
무게	5.1	5.2	6.3	6.6	7.2	7.3	8.4	8.7	9.1	9.5	9.6	9.8

A 순위의 합계(39) B 순위의 합계(39)

순위	1	2	3	4	5	6	7	8	9	10	11	12
상표	B	A	A	B	A	B	A	B	B	A	B	A
무게	5.1	5.2	6.3	6.6	7.2	7.3	8.4	8.7	9.1	9.5	9.6	19.8

순위12에서 이상치 19.8 변경하면 평균치(A)=9.45,평균치(B)=7.733
비모수검정을 하면 이상치에 영향을 받지 않고 유의미한 결과를 얻을 수 있음

7.정규성 검정

데이터의 정규성 검정 3가지

01 Q-Q plot

그래프를 그려서 정규성 가정이 만족되는지 시각적으로 확인하는 방법이다.

Q-Q plot은 아래와 같이 대각선 참조선을 따라서 값들이 분포하게

되면 정규성을 만족한다고 할 수 있다.

만약 한 쪽으로 치우치는 모습이라면

정규성 가정에 위배되었다고 볼 수 있다.

7.정규성 검정

02 Shapiro-Wilk test(샤피로-윌크 검정)

오차항이 정규분포를 따르는지 알아보는 검정으로, 회귀분석에서
모든 독립변수에 대해서 종속변수가 정규분포를
따르는지 알아보는 방법이다.
귀무가설은 'H0:정규분포를 따른다'는 것으로
p-value가 0.05보다 크면 정규성을 가정하게 된다.

7.정규성 검정

03 Kolmogorov-Smirnov test(콜모고로프-스미노프 검정)

자료의 평균/표준편차와 히스토그램을 표준정규분포와 비교하여 적합도를 검정한다.

Shapiro-Wilk test와 마찬가지로 p-value가

0.05보다 크면 정규성을 가정하게 된다.

```
> shapiro.test(DF$x)
   Shapiro-Wilk normality test

data:  DF$Happiness
W = 0.96442, p-value < 0.0000000000000022
```

정규성 검정 결과 유의확률이 유의수준보다 작으므로 x변수의 정규분포를 가정할 수 없다.
정규분포를 가정하고 있는 분석방법을 사용할 경우 주의해야 한다.

→

7.데이터 타입에 따른 분석

분석기법을 적용하기 위해서는 각 변수를 독립변수(X), 종속변수(Y)로 구분하는 것이 유용

독립변수 X	종속변수 Y	적용
수치형	수치형	아버지키(X) 로 아들키(Y) 예측
범주형	수치형	수면제의 종류(X)로 수면시간 증가(Y) 예측
수치형	범주형	온도(X)에 거북이 암수(Y) 예측
범주형	범주형	가족규모그룹(X)에 따라 세탁기 크기(Y) 다른가?

7. 데이터 타입에 따른 분석

데이터 타입에 따른 분석기법

데이터형 타입		대상	분석기법 (R 함수)
독립변수	종속변수		
1개 범주형	수치형	평균비교	수면제의 종류(X)로 수면시간 증가(Y) 예측
2개 범주형	수치형	평균비교	①두집단 평균비교 위해서는 등분산검정을한다 (var.test) p값이>유의수준,등분산동일가정 ② t.test(종속변수~독립변수,var.equal=T) var.equal=TRUE, 등분산 동일가정 옵션 * 독립변수 예 남/녀

7.데이터 타입에 따른 분석

3개 범주형	수치형	평균비교	anova(lm(종속변수~독립변수))분산분석 *주의 독립변수가 factor가 아니면 회귀분석과 같은 분석이 되므로 꼭 factor 확인해야함 * 분산의 동질성 검정 F분포 이용
범주형	범주형		①적합도검정: (1개의 factor) 관측값들이 어떤 이론적 분포를 따르고 있는지 ②독립성 검정:(2개의 factor) 서로 연관이 있는지 ③동질성검정:(2개 factor 교차표) 관측값들이 정해진 범주내에서 서로 비슷한지 chisq.test() * 모분산 가설검정 카이제곱분포 이용

7. 데이터 타입에 따른 분석

수치형	수치형		상관계수:cor() 단순회귀:lm() 산점도:plot()
수치 or 범주	범주형		로지스틱 회귀분석:glm(family='binom')

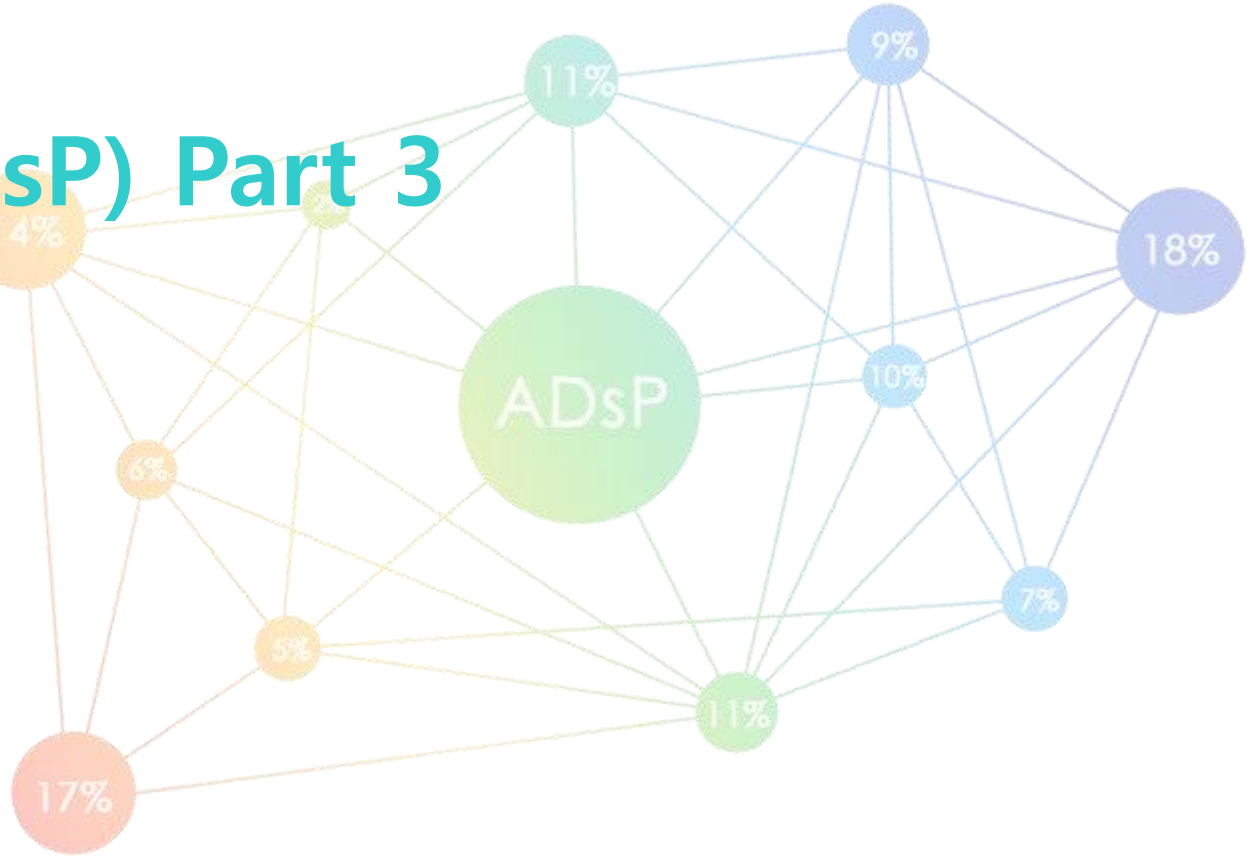
*등분산검정-분산 분석과 회귀 분석 등 대부분의 통계 절차에서는 표본들이 서로 다른 평균을 갖는 모집단에서 추출되었더라도 분산이 동일(등분산), 정규성 가정함 등분산 가정 검정 방법으로 레빈의 검정(Levene's test)과 바틀렛 검정(Bartlett's test)있음

데이터분석준전문가(ADsP) Part 3

데이터분석

02

제2장 통계분석
제2절 기초통계분석

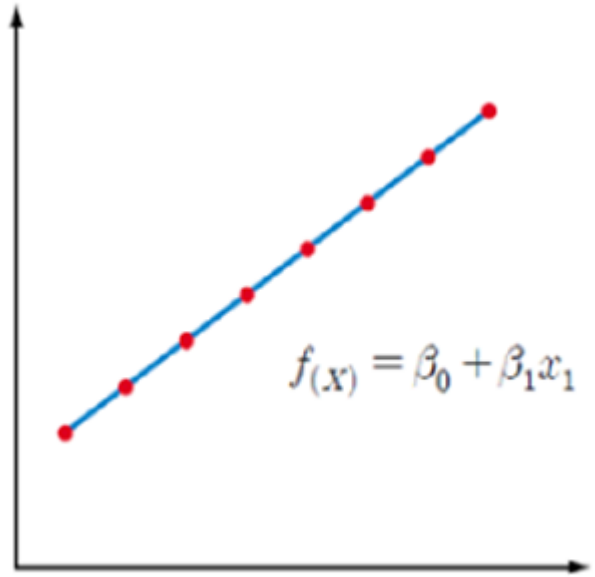


1. 회귀분석

2. 회귀분석 변수 선택방법

1.회귀분석

(1)단순회귀분석

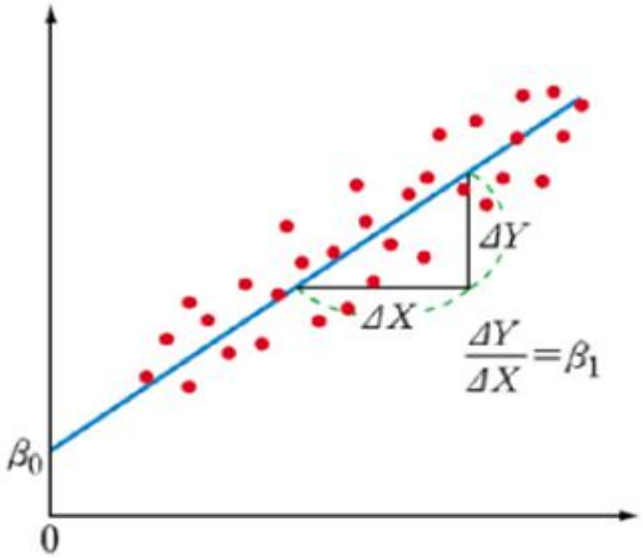


→ x의 변화에 따른 y가 받는 영향 받은 회귀식

독립 변수에 따라 종속변수 추정,예측, 설명 하는 것이 회귀분석이다.

1.회귀분석

(1)단순회귀분석



β_0는 x_i=0일 때 y의 값, 즉 β_0는 상수값이며,y축의 절편

→β_1은 x가 증가할 때 y의 변화량, 회귀선의 기울기를 의미

기울기(회귀계수)는 독립변수의 값이 1단위 증가할 때
종속변수의 증가(감소)을 의미

1.회귀분석

(2)잔차를 포함한 단순회귀 모형

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1, 2, \cdots, n$$

단, Y_i : i 번째 종속변수의 값

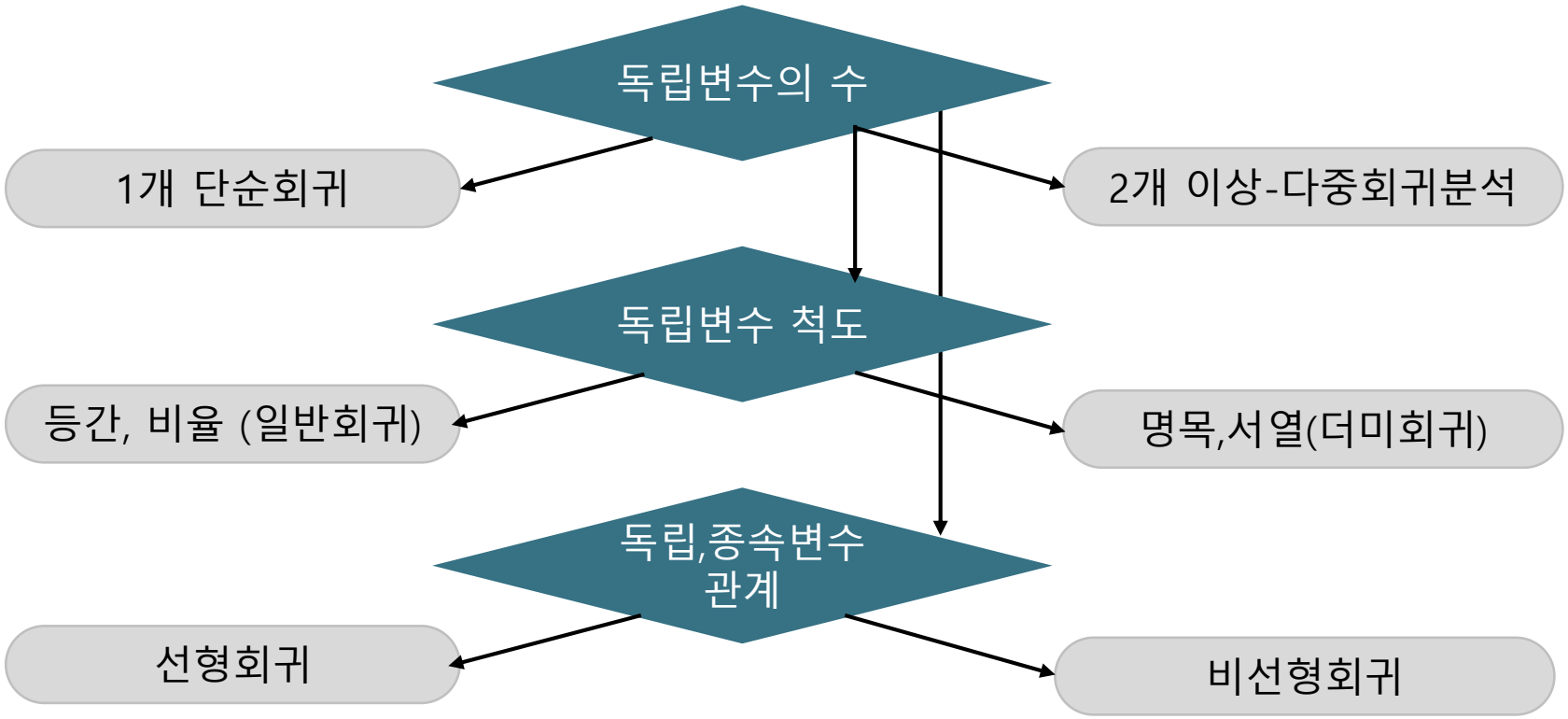
X_i : i번째 독립변수의 값

β_0 : 선형회귀식의 절편

β_1 : 선형회귀식의 기울기

ε_i : 오차항으로 ε_i는 독립적으며 $N(0,\ \delta^2)$의 분포를 이룬다.

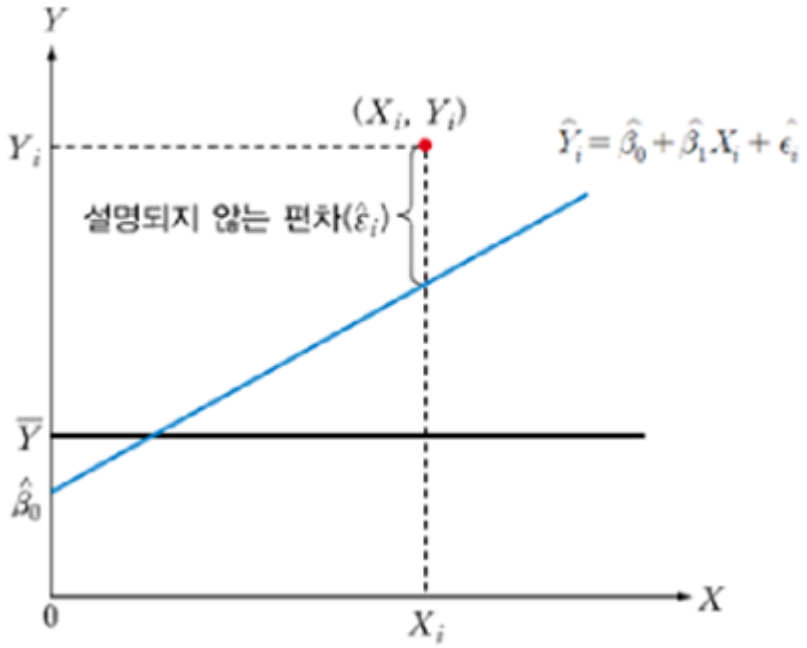
1.회귀분석


독립변수의 수
1개 단순회귀
2개 이상-다중회귀분석
독립변수 척도
등간, 비율 (일반회귀)
명목,서열(더미회귀)
독립,종속변수 관계
선형회귀
비선형회귀

1.회귀분석

(3)최소자승법 또는 최소제곱법

잔차의 제곱합을 최소->최소제곱법(OLS, Ordinary Least Square)



$Y = \beta_0 + \beta_1 X$ 모회귀선

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \epsilon$ 표본 회귀선

오차항 가정: 오차의 기대값=0

1.회귀분석

(4)회귀모형에 대한 가정

① 선형성 [독립변수의 변화에 따라 종속변수도 변화하는 선형(linear)인 모형이다.]

② 독립성 (잔차와 독립변수의 값이 관련되어 있지 않다.)->더빗-왓슨 통계량 이용

③ 등분산성 (오차항들의 분포는 동일한 분산을 갖는다.)

④ 비상관성 (잔차들끼리 상관이 없어야 한다.)

⑤ 정상성 (잔차항이 정규분포를 이뤄야 한다.)

1.회귀분석

(5)회귀분석 함수

lm(formula,data,...)

formula:y~x, y=종속변수 x=독립변수 독립변수가 1개 이상 + 설정

x,y가 데이터프레임의 변수 이름일 때 데이터프레임의 이름 설정함

ex) x$() , y$()

m<-lm( )

summary(m) 객체 저장하면 회귀분석 결과를 확인

1.회귀분석

(6)회귀분석 해석하기

① 모형이 통계적으로 유의미한가? → F분포값과 유의확률(p-value)로 확인한다.

귀무가설: 모형이 유의하지 않다.

② 회귀계수들이 유의미한가? → 회귀계수의 t값과 유의확률(p-value)로 확인한다.

귀무가설 : $\beta = 0$

③ 모형이 얼마나 설명력을 갖는가? → 결정 계수를 확인한다.

④ 모형이 데이터를 잘 적합하고 있는가? → 잔차 통계량을 확인하고 회귀진단을 한다.

1. 회귀분석


```
m<-lm(y~u+v+w,dfrm1)
summary(m)

Call:
lm(formula = y ~ u + v + w, data = dfrm1) # 회귀모형 출력

Residuals: # 잔차의 분포 정보
      Min        1Q    Median        3Q       Max
-0.188562 -0.058632 -0.002013  0.080024  0.143757
```

1.회귀분석

Coefficients:

회귀계수라고(coefficients)-> 비표준화 회귀계수

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.041653	0.264808	11.486	0.0000261520040471	***
u	0.123173	0.012841	9.592	0.0000733951159523	***
v	1.989017	0.016586	119.923	0.0000000000226681	***
w	-2.997816	0.005421	-552.981	0.0000000000000236	***

종속변수 y 대해 u,v비표준화(estimate)계수 양의 관계,w는 음의관계

1.회귀분석

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

회귀계수 유의성 검정->t통계량 때 p값이 모두 <0.05 귀무가설 기각

회귀계수 모두 유의미함

Residual standard error: 0.1303 on 6 degrees of freedom

Multiple R-squared: 0.799 Adjusted R-squared: 0.855

F-statistic: 1.038e+05 on 3 and 6 DF, p-value: 0.0000000000001564

모형의 통계적 유의성 검정 -> F통계량 값,p값(0.0000000000001564)<0.05

귀무가설: 회귀모형은 의미가 없다-> 기각

#결정계수(R^2)=종속변수의 분산 중에서 독립변수의 의해 설명된 분산 비율=85.5%

R^2값은 독립변수의 수가 많아질수록 커지는 특정-> 수정된 R^2을 사용함

1.회귀분석

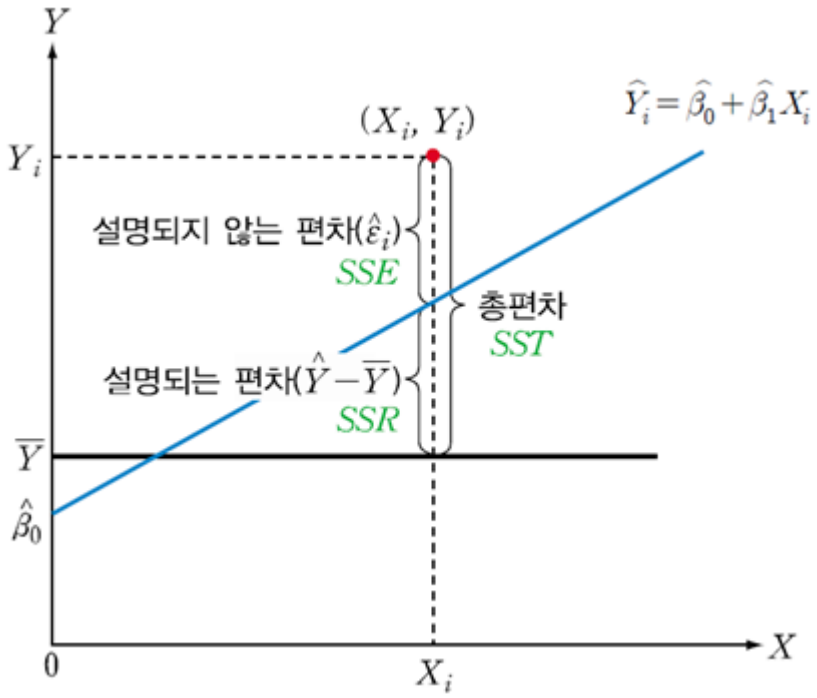
즉 모형의 설명력이 85.5% 참고로 결정 계수 0~1 범위

최종 회귀식 추정

y=3.041653+u*0.123173+v*1.989017-w*2.7997816

1.상관분석

(7)결정계수


Y
$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$
(X_i, Y_i)
Y_i
설명되지 않는 편차$(\hat{\varepsilon}_i)$
SSE
총편차
SST
설명되는 편차$(\hat{Y} - \overline{Y})$
SSR
$\overline{Y}$
$\hat{\beta}_0$
0
X_i
X

1.상관분석

(7)결정계수

① 회귀의 분산분석에서 총제곱합(총변동, SST) = 회귀제곱합(설명된 변동, SSR) + 오차제곱합(설명 안 된 변동, SSE)이며, SSR은 추정회귀방정식에 의해 설명되는 부분이다.

② R^2(결정계수) = 회귀제곱합(SSR)/총제곱합(SST)

③ 결정계수는 독립변수(X)가 종속변수(Y)를 얼마나 설명

R^2값이 클수록 회귀방정식의 설명력은 높아진다.

1.회귀분석

(8)더빗-왓슨 검정

잔차의 독립성 검정(Durbin-Watson)

0~4 사이에 나오며 2에 가까울수록 자기상관이 없이 독립이며, 독립인 경우 회귀분석을 사용

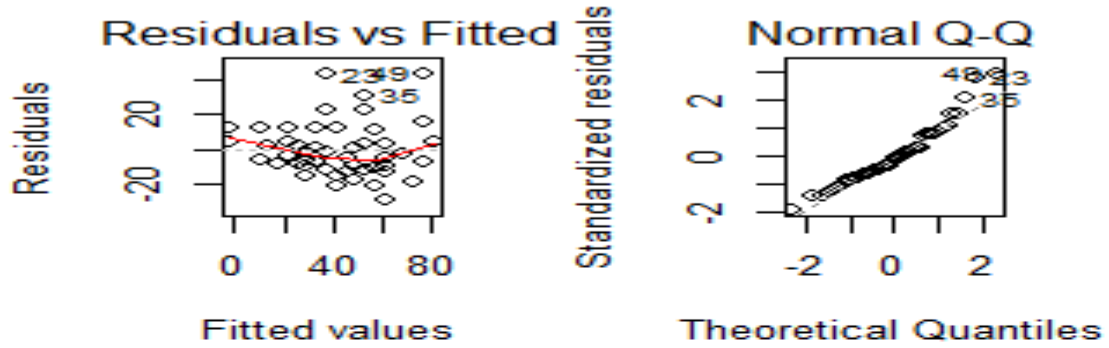
```
m1<-residuals(m)
durbinWatsonTest(m1)
```

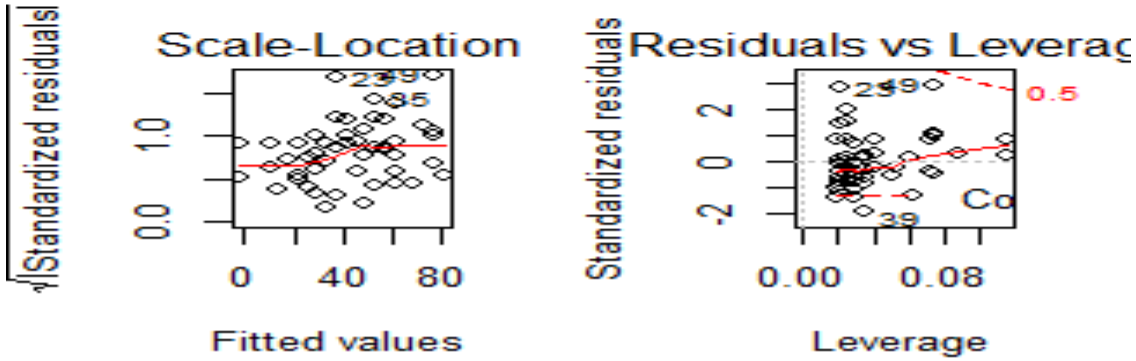
```
[1] 2.234025
```

1.회귀분석

(9)잔차 분석

plot(m)





1.회귀분석

① Normal Q-Q: 정규성 가정을 만족한다면 이 그래프의 점들은 45도 각도의 직선 위에 있어야 한다.

② Residuals vs Fitted:선형관계라면 예측값(predicted values)과 잔차(Residuals)가 일정 관계를 가지면 안된다.
즉 점이 랜덤하게 산포되어 있어야 함

③ Scale-Location: 분산이 일정하다면 값이 무작위로 찍혀야 한다.

④ Redisuals vs Leverage: 개개 관측치의 이상치, 통계 모형 계수에 영향을 줄 수 있는 관측치를 확인하기 위해 Cook's distance가 빨간 점선으로 표시

1.회귀분석

(10)표준화 계수 의미

lm 함수에서 얻은 독립변수의 회귀계수는 독립변수가 한 단위 증가할 때 종속변수가 변하는 양을 의미. 만약 두 독립변수의 단위가 다르게 측정되었다면 상대적 비교가 불가능 종속변수에 대한 독립변수의 영향을 의미하는 회귀계수를 비교하기 위해 표준화된 회귀계수 비교

```
> lm.beta(m)
          u           v           w
 0.0174509   0.2272843 –1.0443563
```

w>v>u 순서로 y에 영향력이 있음

1.회귀분석

(11)다중공선성

다중 공선성이란 모형의 일부 독립변수가 다른 독립변수와 상관되어 있을 때 발생하는 조건이다. 다중 공선성이 존재할 때 회귀계수의 추정치의 안전성과 신뢰성 문제가 발생한다.

```
> vif(m)
       u        v        w
1.030794 1.118732 1.110882
```

→ vif 값이 10미만이면 다중 공선성 문제가 없다고 간주함

2.변수선택방법

변수선택은 step 함수를 이용하여 분석

AIC : 모형의 적합도와 모형의 복잡성 사이의 균형을 다루는 통계량

모형의 적합도가 높으면 AIC의 값은 낮아진다.

→ AIC값이 낮을수록 좋은 모형

step함수에서 변수 선택의 방법은 전진(forward)선택방법,

후진(backward)제거방법,단계별 선택(both) 있다.

① 단계별 선택(Stepwise Selection)

전진선택방법과 후진제거방법을 함께 사용

② 후진 제거법(Backward Elimination): 모든 변수가 포함된 모델에서

기준 통계치에 가장 도움이 되지 않는 변수를 하나씩 제거하는 방법

→ AIC값이 제거되기 이전의 AIC값보다 낮아지면 해당변수 제거

2.변수선택방법

③ 전진 선택법(Forward Selection): 절편만 있는 모델에서 기준 통계치를 가장 많이 개선시키는 변수를 차례로 추가하는 방법

→ AIC값 비교 낮아진다면 추가

```
> step(lm(y~1,df),scope=list(lower=~1,upper=~x1+x2+x3+x4),direction="forward")
# step(lm(종속변수~설명변수,데이터셋), scope=list(lower=~1,upper=~설명변수,
direction="변수선택 방법")
Start:  AIC=71.44
y ~ 1

       Df Sum of Sq     RSS    AIC
+ x4    1   1831.90  883.87 58.852
+ x2    1   1809.43  906.34 59.178
+ x1    1   1450.08 1265.69 63.519
+ x3    1    776.36 1939.40 69.067
<none>               2715.76 71.444
```

2.변수선택방법

Step: AIC=58.85

y ~ x4

	Df	Sum of Sq	RSS	AIC
+ x1	1	809.10	74.76	28.742
+ x3	1	708.13	175.74	39.853
<none>			883.87	58.852
+ x2	1	14.99	868.88	60.629

2.변수선택방법

Step: AIC=28.74
y ~ x4 + x1

	Df	Sum of Sq	RSS	AIC
+ x2	1	26.789	47.973	24.974
+ x3	1	23.926	50.836	25.728
<none>			74.762	28.742

Step: AIC=24.97
y ~ x4 + x1 + x2

	Df	Sum of Sq	RSS	AIC
<none>			47.973	24.974
+ x3	1	0.10909	47.864	26.944

2.변수선택방법

Call:

lm(formula = y ~ x4 + x1 + x2, data = df)

Coefficients:

(Intercept)	x4	x1	x2
71.6483	-0.2365	1.4519	0.4161

Quiz

Q01 회귀분석과 결정계수 설명이 부적절한 것은?

① 결정계수는 총변동과 오차에 대한 변동 비율이다.

② 결정계수가 커질수록 회귀방정식의 설명력이 높아진다.

③ 결정계수는 0~1 사이의 범위를 갖는다.

④ 회귀계수의 유의성 검증은 t값과 p값을 통해 확인한다.

Quiz

Q01 회귀분석과 결정계수 설명이 부적절한 것은?

정답

① 결정계수는 총변동과 오차에 대한 변동 비율이다.

② 결정계수가 커질수록 회귀방정식의 설명력이 높아진다.

③ 결정계수는 0~1 사이의 범위를 갖는다.

④ 회귀계수의 유의성 검증은 t값과 p값을 통해 확인한다.

Quiz

Q02 다음은 회귀분석의 분산분석표이다. 결정계수의 값은?

```
> anova(df1) Analysis of Variance Table
  Response: Happiness
              Df    Sum Sq Mean Sq F value     Pr(>F)
  BM          1    280.00  287.95  702.2 < 2.2e-16 ***
Residuals 1923  720.00    0.41
```

Quiz

Q02 다음은 회귀분석의 분산분석표이다. 결정계수의 값은?

```
> anova(df1) Analysis of Variance Table
  Response: Happiness
            Df    Sum Sq Mean Sq F value    Pr(>F)
  BM         1    280.00  287.95  702.2 < 2.2e-16 ***
Residuals 1923  720.00     0.41
```

정답

280/100=28%