PART 03
데이터 분석
4장

통계 분석

17 DAY

○ 학습 목표

- 통계의 정의와 자료 획득 방법을 이해한다.
- 통계분석과 통계분석 방법을 이해한다.
- 확률 및 확률분포를 이해한다.
- 추정과 가설검정을 이해한다.

○ 눈높이 체크

✓ **통계의 정의와 통계자료의 획득 방법을 알고 계시나요?**

여러분은 아마도 일상에서 통계라는 단어를 많이 접하시고 계실 것입니다. 통계는 간단한 테이블과 그래프에서 아주 복잡한 분석 결과에 이르기까지 그 형태는 다양합니다. 그리고 통계를 만들기 위해 필요한 통계자료를 획득하는 방법으로 총조사와 샘플링 조사가 있습니다.

✓ **통계분석의 방법에 대해 알고 계신가요?**

통계자료를 획득한 후 통계분석을 하게 되는데 분석 방법에는 크게 기술통계와 통계적 추론으로 구분됩니다.

✓ **확률에 대해 이해하고 계신가요?**

확률이 복잡하고 머리 아픈 학문이지만 여러분도 일상생활에서 많이 활용하고 있습니다. 스포츠 경기에서 이길 팀을 예측할 때나 누군가를 위한 깜짝 선물을 고를 때도 나도 모르게 확률을 활용하고 있습니다.

✓ **추정과 가설검정에 대해 들어 보셨나요?**

추정은 표본으로부터 모집단이 가지는 특성(모수)을 추측하는 것입니다. 그리고 자신이 가지는 이론적 대안이 통계적으로 의미가 있는지를 확인하는 것이 가설검정이라 할 수 있겠습니다.

1절 통계분석의 이해

① 통계

- 특정집단을 대상으로 수행한 조사나 실험을 통해 나온 결과에 대한 요약된 형태의 표현이다.
  예 일기예보, 물가/실업률/GNP, 정당 지지도, 의식조사와 사회조사 분석 통계, 임상실험 등의 실험 결과 분석 통계
- 조사 또는 실험을 통해 데이터를 확보, 조사대상에 따라 총조사(Census)와 표본조사(Sampling)로 구분한다.

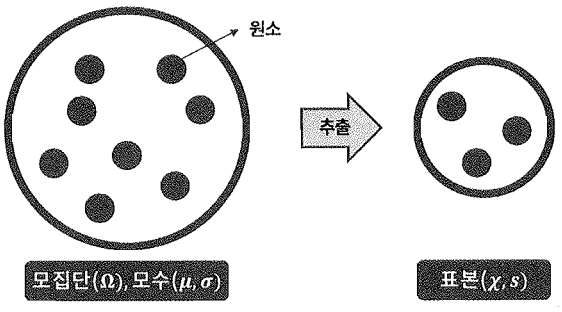
② 통계자료의 획득 방법

가. 총 조사/전수 조사(Census)

- 대상 집단 모두를 조사하는데 많은 비용과 시간이 소요되므로 특별한 경우를 제외하고는 사용되지 않는다. (ex. 인구주택 총 조사)

나. 표본조사(Sampling)

- 대부분의 설문조사가 표본조사로 진행되며 모집단에서 샘플을 추출하여 진행하는 조사이다.
- **모집단(Population)** : 조사하고자 하는 **대상 집단 전체**
- **원소(Element)** : **모집단을 구성**하는 개체
- **표본(Sample)** : 조사하기 위해 **추출한 모집단의 일부 원소**
- **모수(Parameter)** : 표본 관측에 의해 구하고자 하는 **모집단에 대한 정보**
- 모집단의 정의, 표본의 크기, 조사방법, 조사기간, 표본추출방법을 정확히 명시해야 한다.



- 표본오차 : 모집단의 일부인 표본에서 얻은 자료를 통해 모집단 전체의 특성을 추론함으로써 생기는 오차, 모집단을 대표할 수 있는 표본단위들이 조사대상으로 추출되지 못하면 발생
- 비표본오차 : 표본오차를 제외한 조사의 전체 과정에서 발생할 수 있는 모든 오차
- 표본편의 : 표본추출방법에서 기인하는 오차로 표본추출이 의도된 모집단의 일부 구성원이 다른 구성원보다 더 낮거나 더 높은 표본추출 확률을 갖는 오차

표본추출방법 4가지를 헷갈리지 않게 꼭 암기하도록 합시다. 한글뿐만 아니라 영어로 된 용어도 알도록 합시다. 한글과 영어가 혼용되어 출제될 수 있기 때문입니다. 이 표에 나오는 용어뿐 아니라 본 장에서 나오는 영어로 된 용어를 알아두도록 합니다.

다. 표본 추출 방법

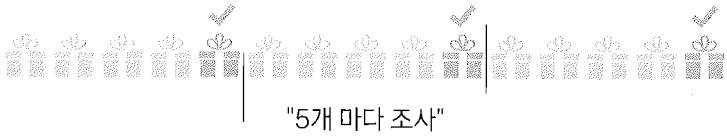
- 표본조사의 중요한 점은 모집단을 대표할 수 있는 표본 추출이므로 표본 추출 방법에 따라 분석결과의 해석은 큰 차이가 발생한다.
  (N개의 모집단에서 n개의 표본을 추출하는 경우)

1) 단순랜덤 추출법 (Simple Random Sampling)

- 각 샘플에 번호를 부여하여 임의의 n개를 추출하는 방법으로 각 샘플은 선택될 확률이 동일하다.(비복원, 복원(추출한 Element를 다시 집어넣어 추출하는 경우) 추출)

2) 계통추출법 (Systematic Sampling)

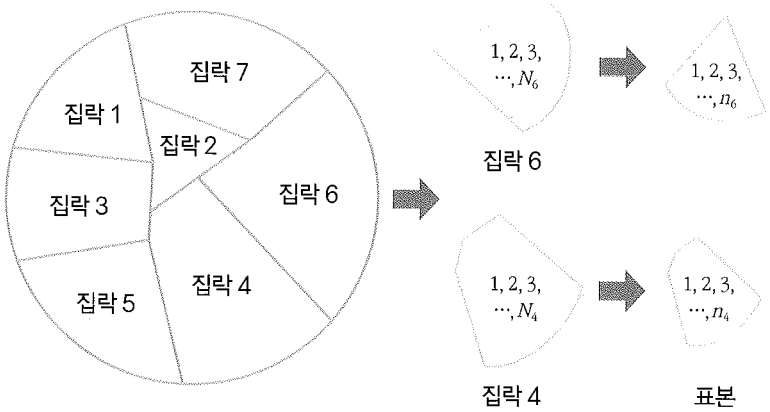
- 단순랜덤추출법의 변형된 방식으로 번호를 부여한 샘플을 나열하여 K개씩 (K=N/n) n개의 구간으로 나누고 첫 구간(1, 2, … , K)에서 하나를 임의로 선택한 후에 K개씩 띄어서 n개의 표본을 선택한다. 즉, 임의의 위치에서 매 k번째 항목을 추출하는 방법이다.



예 15개 중 3개의 샘플 추출

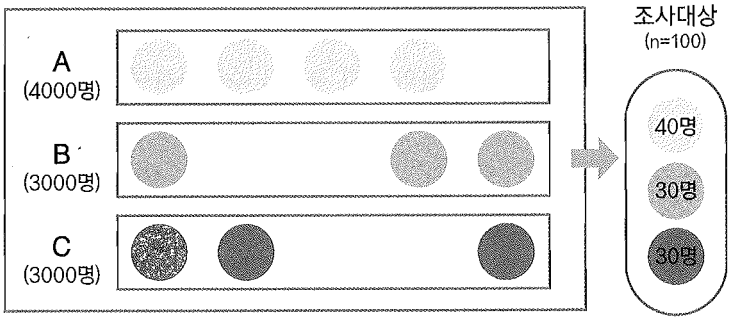
3) 집락추출법(Cluster Random Sampling)

- 군집을 구분하고 군집별로 단순랜덤 추출법을 수행한 후, 모든 자료를 활용하거나 샘플링하는 방법이다. (지역표본추출, 다단계표본추출)



4) 층화추출법(Stratified Random Sampling)

- 이질적인 원소들로 구성된 모집단에서 각 계층을 고루 대표할 수 있도록 표본을 추출하는 방법으로, 유사한 원소끼리 몇 개의 층(Stratum)으로 나누어 각 층에서 랜덤 추출하는 방법이다. (비례층화추출법, 불비례층화추출법)



※ 실험 : 특정 목적 하에서 실험 대상에게 처리를 가한 후에 그 결과를 관측해 자료를 수집하는 방법이다.

★2회에 1번꼴로 출제되는 부분입니다.
각 척도의 정의와 예시를 반드시 숙지하도록 합시다.

라. 측정(Measurement)

1) 개요

- 표본조사나 실험을 실시하는 과정에서 추출된 원소들이나 실험 단위로부터 주어진 목적에 적합하도록 관측해 자료를 얻는 것이다.

2) 측정방법

명목척도	측정 대상이 어느 **집단**에 속하는지 분류할 때 사용 (성별, 출생지 구분)	**질적척도** (범주형자료, 숫자들의 크기 차이가 계산되지 않는 척도)
순서척도	측정 대상의 **서열관계**를 관측하는 척도 (만족도, 선호도, 학년, 신용등급)	
구간척도 (등간척도)	측정 대상이 갖고 있는 **속성의 양**을 측정하는 것으로 구간이나 구간 사이의 **간격이 의미가 있는** 자료(온도, 지수)	**양적척도** (수치형자료, 숫자들의 크기 차이를 계산할 수 있는 척도)
비율척도	간격(차이)에 대한 비율이 의미를 가지는 자료, **절대적 기준인 0이 존재**하고 **사칙연산이 가능**하며 제일 많은 정보를 가지는 척도(무게, 나이, 시간, 거리)	

- 서열척도는 명목척도와 달리 매겨진 숫자의 크기를 의미있게 활용할 수 있다.(예 : 1등이 2등보다는 성적이 높다.)
- 구간척도는 절대적 크기는 측정할 수 없기 때문에 사칙연산 중 더하기와 빼기는 가능하지만 비율처럼 곱하거나 나누는 것은 불가능하다.

③ 통계분석

가. 정의

- 특정한 집단이나 **불확실한 현상**을 대상으로 자료를 수집해 대상 집단에 대한 정보를 구하고, **적절한 통계분석 방법을 이용해 의사결정을 하는** 과정이다.

나. 기술통계(Descriptive Statistic)

- 주어진 자료로부터 어떠한 판단이나 예측과 같은 주관이 섞일 수 있는 과정을 배제하여 통계집단들의 여러 특성을 수량화하여 객관적인 데이터로 나타내는 통계분석 방법론이다.
- Sample에 대한 특성인 평균, 표준편차, 중위수, 최빈값, 그래프, 왜도, 첨도 등을 구하는 것을 의미한다.

다. 통계적 추론(추측통계, Inference Statistics)

- 수집된 자료를 이용해 대상 집단(모집단)에 대한 의사결정을 하는 것으로 Sample을 통해 모집단을 추정하는 것을 의미한다.

1) 모수추정

- 표본집단으로부터 **모집단의 특성인 모수(평균, 분산 등)를 분석**하여 모집단을 추론한다.

2) 가설검정

- 대상집단에 대해 특정한 가설을 설정한 후에 그 가설이 **옳은지 그른지에 대한 채택여부를 결정**하는 방법론이다.

3) 예측

- 미래의 **불확실성을 해결해 효율적인 의사결정**을 하기 위해 활용한다. (예: 회귀분석, 시계열분석 등의 방법이 있다.)

④ 확률 및 확률분포

가. 확률

- 표본공간 S에 부분집합인 각 사상에 대해 실수값을 가지는 함수의 확률값이 0과 1사이에 있고, 전체 확률의 합이 1인 것을 의미한다. 표본공간 Ω의 부분집합인 사건 E의 확률은 표본공간의 원소의 개수에 대한 사건 E의 개수의 비율로 확률을 P(E)라고 할 때, 다음과 같이 정의한다.

$$P(E) = \frac{n(E)}{n(\Omega)}$$

1) **표본공간**(Sample Space, Ω)

- 어떤 실험을 실시할 때 나타날 수 있는 **모든 결과들의 집합**이다.

2) 사건(Event)

- 관찰자가 관심이 있는 사건으로 표본공간의 부분집합이다.

3) 원소(Element)

- 나타날 수 있는 개별의 결과들을 의미한다.

4) 확률변수(Random Variable)

- 특정값이 나타날 가능성이 확률적으로 주어지는 변수이다.
- **정의역(Domain)이 표본공간, 치역(Range)이 실수값(0<y<1)인 함수**이다.
- 0이 아닌 확률을 갖는 실수값의 형태에 따라 이산형 확률변수(Discrete Random Variable)와 연속형 확률변수(Continuous Random Variable)로 구분된다.
- 확률변수의 기대값 확률변수 X의 기대값(Expectation, Expected Value)은 다음과 같이 정의한다.

$$E(X)=\begin{cases}\sum x_i f(x_i) & : \text{이산형 변수인 경우}\\ \int xf(x)dx & : \text{연속형 변수인 경우}\end{cases}$$

일반적으로 확률변수 X의 k차 적률(k-th Moment)

$$E(X^k)=\begin{cases}\sum x_i^k f(x_i) & : \text{이산형 변수인 경우}\\ \int x^k f(x)dx & : \text{연속형 변수인 경우}\end{cases}$$

확률변수 X의 k차 중심적률(k-th Cental Moment)

$$E[(X-\mu)^k]=\begin{cases}\sum (x_i-\mu)^k f(x_i) & : \text{이산형 변수인 경우}\\ \int (x-\mu)^k f(x)dx & : \text{연속형 변수인 경우}\end{cases}$$

특히, 2차중심적률 $E[(X-\mu)^2]=\sigma^2$: 모분산(Population Variance)

기대값의 선형성을 이용하면

$$\sigma^2 = E[(X-\mu)^2]$$
$$= E[(X^2-2\mu X+\mu^2)]$$
$$= E(X^2)-2\mu E(X)+\mu^2$$
$$= E(X^2)-\mu^2$$

즉, 모분산 = 2차 적률 - 1차 적률^{2}로 해석 가능

$$E(aX+b)=\sum(ax+b)P(X=x)$$
$$=\sum(axP(X=x)+bP(X=x))$$
$$=a\sum xP(X=x)+b\sum P(X=x)$$
$$=aE(X)+b(1)=aE(X)+b$$
$$Var(aX+b)=E[(aX+b-a\mu-b)^2]$$
$$=E[a^2(X-\mu)^2]$$
$$=a^2E[(X-\mu)^2]=a^2Var(X)$$

참고

- 동전 2개를 던져서 앞/뒷면이 나오는 경우의 수(H:앞, T:뒤)

확률분포표					
표본공간(Ω)	HH(사건)	HT	TH	TT	합계
P(x)	1/4(원소)	1/4	1/4	1/4	1

- 덧셈정리(배반사건이 아닐 때) : 사건 A와 사건 B가 동시에 일어날 수 있을 때(교집합이 성립할 때), **일어날 확률** P(A 또는 B)는 P(A∪B)=P(A)+P(B)-P(A∩B)로 표현된다. **사건 B가 주어졌을 때 사건 A의 조건부 확률은** P(A|B)=P(A∩B)/P(B)로 표현된다.
- 덧셈정리(배반사건일 때) : 사건 A와 사건 B가 동시에 일어나지 않을 때, 즉 **사건 A 또는 사건 B 중 어느 한 쪽만 일어날 확률은** P(A∪B)=P(A)+P(B)로 표현된다.
- 곱셈정리 : 사건 A와 B가 서로 무관계하게 나타날 때, 즉 **독립사건(獨立事件)일 때 A와 B가 동시에 나타날 확률** P(A와 B)는 P(A∩B)=P(A)×P(B)로 표현되고, **사건 B가 주어졌을 때 사건 A의 조건부 확률은** P(A|B)=P(A)로 표현된다.

이산형 확률변수의 종류와 정의, 그리고 특징까지 알아둡시다.

나. 확률분포

1) 이산형 확률변수

- 0이 아닌 확률값을 갖는 확률 변수를 셀 수 있는 경우(확률질량함수)

$$P(X_i)>0 \quad i=1,2,\cdots,k \quad \sum_{i=1}^{k}P(X_i)=1$$

- 이산형 확률변수의 예시 : 동전 2개를 던져서 앞/뒷면이 나오는 경우의 수(H:앞, T:뒤)

표본공간(Ω)	HH(사건)	HT	TH	TT	합계
P(x)	1/4(원소)	1/4	1/4	1/4	1

가) 베르누이 확률분포(Bernoulli Distribution)

- 결과가 2개만 나오는 경우 (예시 : 동전 던지기, 시험의 합격/불합격 등)

  $P(X=x)=p^x \cdot (1-p)^{1-x}$ (x = 1 or 0),

  E(x)=p, Var(x)=p(1-p)

  예 메이저리거인 추신수 선수가 안타를 칠 확률은 베르누이 분포를 따른다. (안타를 치는 사건을 x=1이라고 할 때 안타를 칠 확률은 타율로 적용 가능)

나) 이항분포(Binomial Distribution)

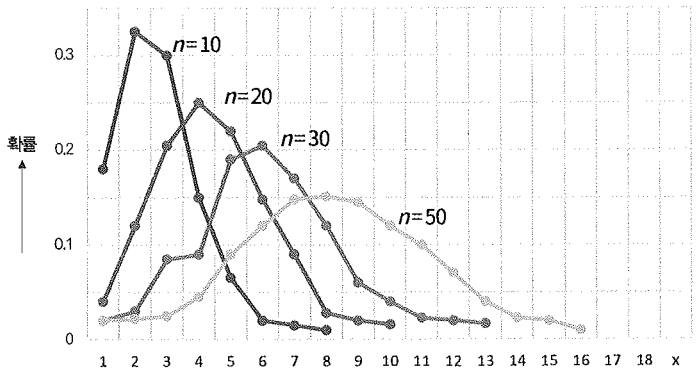
- 베르누이 시행을 n번 반복했을 때 k번 성공할 확률

  $P(X=k)={}_nC_k p^k (1-p)^{n-k}$, ${}_nC_k = \frac{n!}{k!(n-k)!}$

  x~B(n,p), E(x) = np, Var(x)= np(1-p)

  예 메이저리거인 추신수 선수가 오늘 경기에서 5번 타석에 들어와서 3번 안타를 칠 확률은 이항분포를 따른다.(n=5, k=3, 안타를 칠 확률 P(x)=타율로 적용 가능)

- 성공할 확률 p가 0이나 1에 가깝지 않고 n이 충분히 크면 이항분포는 정규분포에 가까워진다. 성공할 확률 p가 1/2에 가까우면 종모양이 된다.



1의 주사위를 10회, 20회, 30회, 50회 던졌을 때, 그 눈이 x회 나올 확률을 그래프로 나타낸 것

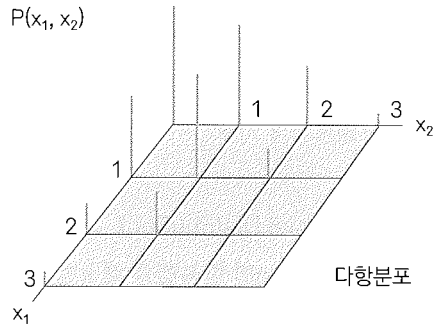
다) 기하분포(Geometric Distribution)

- 성공확률이 p인 베르누이 시행에서 첫 번째 성공이 있기까지 x번 실패할 확률

예 메이저리거인 추신수 선수가 오늘 경기에서 5번 타석에 들어와서 3번째 타석에서 안타칠 확률은 기하분포를 따른다.

라) 다항분포(Multinomial Distribution)

- 이항분포를 확장한 것으로 세가지 이상의 결과를 가지는 반복 시행에서 발생하는 확률 분포



마) 포아송분포(Poisson Distribution)

- 시간과 공간 내에서 발생하는 사건의 발생 횟수에 대한 확률분포 (예 : 책에 오타가 5page 당 10개씩 나온다고 할 때, 한 페이지에 오타가 3개 나올 확률)
- λ=정해진 시간 안에 어떤 사건이 일어날 횟수에 대한 기댓값, y=사건이 일어난 수

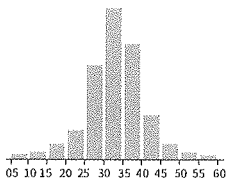
$$P(y)=\lim_{n\to\infty}\begin{bmatrix} n \\ y \end{bmatrix} p^{y}(1-p)^{n-y}=\frac{\lambda^{y}}{y!}e^{-\lambda}$$

예 메이저리거인 추신수 선수가 최근 5경기에서 10개의 홈런을 때렸다고 할 때, 오늘 경기에서 홈런을 못 칠 확률은 포아송분포를 따른다.

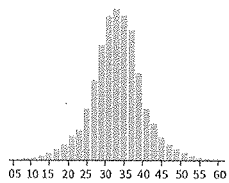
2) 연속형 확률변수

- 가능한 값이 실수의 어느 특정구간 전체에 해당하는 확률변수(확률밀도함수)

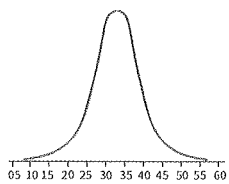
$$f(x)\geq 0 \qquad \int_{-\infty}^{\infty} f(x)dx=1$$



신생아 5,000명의 몸무게를 10개의 계급구간으로 나눈 히스토그램



신생아 5,000명의 몸무게를 30개의 계급구간으로 나눈 히스토그램

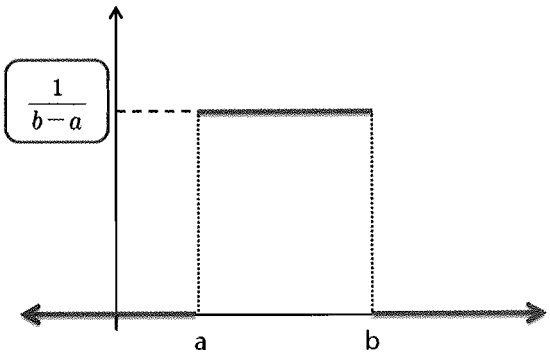


신생아 몸무게에 대한 확률밀도곡선

가) 균일분포(일양분포, Uniform Distribution)

- 모든 확률변수 X 가 균일한 확률을 가지는 확률분포 (다트의 확률분포)

$$E(X)=\frac{a+b}{2} \quad Var(X)=\frac{(b-a)^2}{12}$$



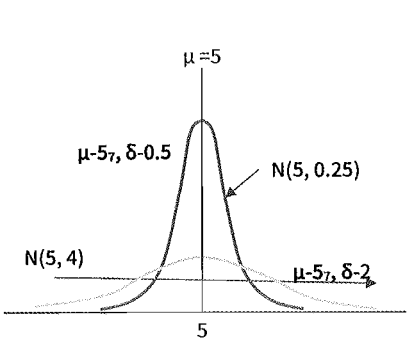
〈균일 분포〉

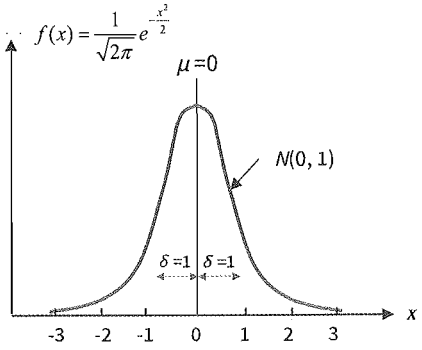
나) 정규분포(Normal Distribution)

- 평균이 μ이고, 표준편차가 σ인 x의 확률밀도함수
- 표준편차가 클 경우 퍼져보이는 그래프가 나타난다.

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty<x<\infty$$

표준정규분포의 누적분포함수





· **최소-최대 정규화 (Min-Max Normalization)**
: (X-Min)/(Max-Min), 원 데이터의 분포를 유지하면서 0~1 사이 값이 되도록 정규화함

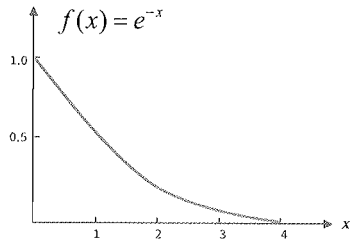
· **Z-점수 표준화 (Z-Score Standardization)**
: (X-평균)/표준편차, 원 데이터를 표준정규분포에 해당되도록 표준화함

표준정규분포는 평균이 0이고 표준편차가 1인 정규분포입니다. 정규분포를 표준정규분포로 만들기 위해선 $Z=\frac{X-\mu}{\sigma}$ 식을 이용합니다.

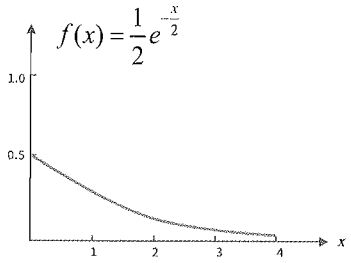
다) 지수분포(Exponential Distribution)

- 어떤 사건이 발생할 때까지 경과 시간에 대한 연속확률분포이다.

  예 전자레인지의 수명시간, 콜센터에 전화가 걸려올 때까지의 시간, 은행에 고객이 내방하는데 걸리는 시간, 정류소에서 버스가 올 때까지의 시간



(지수분포밀도함수λ=1)



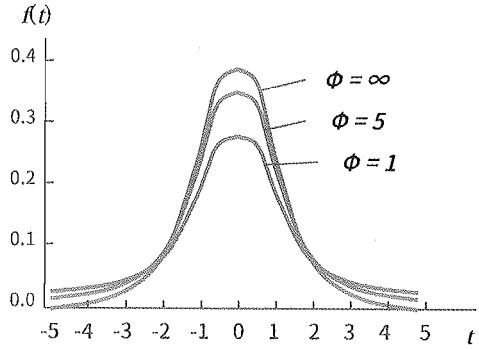
(지수분포밀도함수λ=1/2)

가설검정 시 많이 활용되는 분포가 t-분포, 카이제곱분포, F-분포입니다. 이 세 분포의 특징에 대해 정확히 기억합시다.

정규분포보다 더 퍼져있고 자유도가 커질수록 정규분포에 가까워집니다.

라) t-분포(t-Distribution)

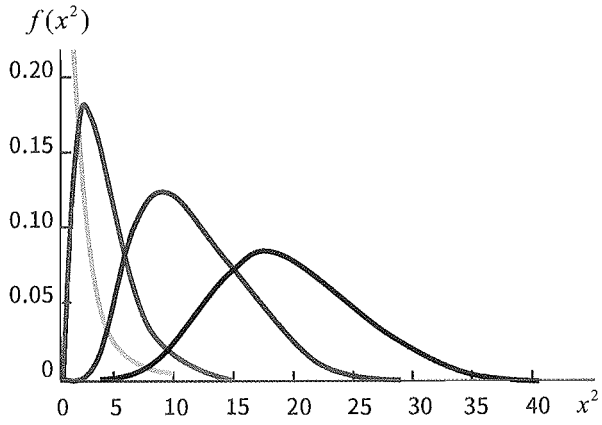
- 표준정규분포와 같이 평균이 0을 중심으로 좌우가 동일한 분포를 따른다.
- 표본의 크기가 적을 때는 표준정규분포를 위에서 눌러 놓은 것과 같은 형태를 보이지만 표본이 커져서(30개 이상) 자유도가 증가하면 표준정규분포와 거의 같은 분포가 된다.
- 데이터가 연속형일 경우 활용한다.
- **두 집단의 평균이 동일**한지 알고자 할 때 검정통계량으로 활용된다.



〈 t-분포 〉

마) x^2-분포(Chi-Square Distribution)

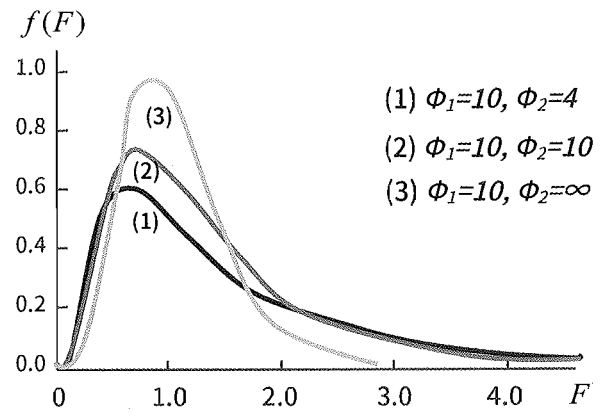
- 모평균과 모분산이 알려지지 않은 모집단의 모분산에 대한 가설 검정에 사용되는 분포이다.
- **두 집단 간의 동질성 검정에 활용**된다. (범주형 자료에 대해 얻어진 관측값과 기대값의 차이를 보는 적합성 검정에 활용)



〈 x^2- 분포 〉

바) F-분포(F-Distribution)

- **두 집단 간 분산의 동일성 검정**에 사용되는 검정 통계량의 분포이다.
- 확률변수는 항상 양의 값만을 갖고 x^2분포와 달리 자유도를 2개 가지고 있으며 자유도가 커질수록 정규분포에 가까워진다.



〈 F- 분포 〉

⑤

추정과 가설검정

가. 추정의 개요

1) 확률표본(Random Sample)

- 확률분포는 분포를 결정하는 평균, 분산 등의 모수(Parameter)를 가지고 있다.
- 특정한 확률분포로부터 독립적으로 반복해 표본을 추출하는 것이다.
- 각 관찰값들은 서로 독립적이며 동일한 분포를 갖는다.

2) 추정

- 표본으로부터 미지의 모수를 추측하는 것이다.
- 추정은 점추정(Point Estimation)과 구간추정(Interval Estimation)으로 구분된다.

가) 점추정(Point Estimation)

- **'모수가 특정한 값일 것'**이라고 추정하는 것이다.
- 표본의 평균, 중위수, 최빈값 등을 사용한다.

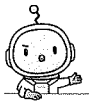
참고

점추정량의 조건, 표본평균, 분산

- 불편성(Unbiasedness) : 모든 가능한 표본에서 얻은 추정량의 기댓값은 모집단의 모수와 편의(차이)가 없다.
- 효율성(Efficiency) : 추정량의 분산이 작을수록 좋다.
- 일치성(Consistency) : 표본의 크기가 아주 커지면, 추정량이 모수와 거의 같아진다.
- 충족성(Sufficient) : 추정량은 모수에 대하여 모든 정보를 제공한다.
- 표본평균(Sample Mean) : 모집단의 평균(모평균)을 추정하기 위한 추정량. 확률표본의 평균값.

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

- 표본분산(Sample Variance) : 모집단의 분산(모분산)을 추정하기 위한 추정량



$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

나) 구간추정(Interval Estimation)

- 점추정의 정확성을 보완하기 위해 확률로 표현된 믿음의 정도 하에서 **모수가 특정한 구간에 있을 것이라고 선언**하는 것이다.
- 항상 추정량의 분포에 대한 전제가 주어져야 하고, 구해진 구간 안에 모수가 있을 가능성의 크기(신뢰수준(Confidence Interval))가 주어져야 한다.

모분산을 알 때는 분자에 σ를 넣고, 모분산을 모를 때는 분자에 s를 넣는다는 것을 기억합시다.

참고

95% 신뢰수준 하에서 모평균의 신뢰구간

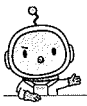
- 모분산 σ^2이 알려져 있는 경우

$$\left(\overline{X}-1.96\frac{\sigma}{\sqrt{n}},\ \overline{X}+1.96\frac{\sigma}{\sqrt{n}}\right)$$

표준정규분포 $N(0,1)$를 따르는 $Z=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ 통계량 이용

- 모분산 σ^2이 알려져 있지 않은 경우에는 모분산 대신 표본분산을 사용

$$\left(\overline{X}-2.26\frac{S}{\sqrt{n}},\ \overline{X}+2.26\frac{S}{\sqrt{n}}\right)$$



자유도가 n-1인 t-분포를 따르는 $T=\frac{\overline{X}-\mu}{S/\sqrt{n}}$ 통계량 이용

나. 가설검정

1) 정의

- 모집단에 대한 어떤 가설을 설정한 뒤에 표본관찰을 통해 그 가설의 채택여부를 결정하는 분석방법이다.
- 표본 관찰 또는 실험을 통해 귀무가설과 대립가설 중에서 하나를 선택하는 과정이다.
- 귀무가설이 옳다는 전제하에 검정통계량 값을 구한 후에 이 값이 나타날 가능성의 크기에 의해 귀무가설의 채택여부를 결정한다.

· 유의확률(p-value) : 귀무가설이 맞다고 가정할 때 얻을 수 있는 결과보다 실제값이 더 극단에 위치할 확률

· 검정력(Statistical Power) : 대립가설이 사실일 때, 대립가설을 채택하는 옳은 결정을 할 확률

가) 귀무가설(Null Hypothesis, H_0)

- **'비교하는 값과 차이가 없다, 동일하다'**를 기본개념으로 하는 가설

나) 대립가설(Alternative Hypothesis, H_1)

- **뚜렷한 증거가 있을 때 주장하는 가설**

다) 검정통계량(Test Statistic)

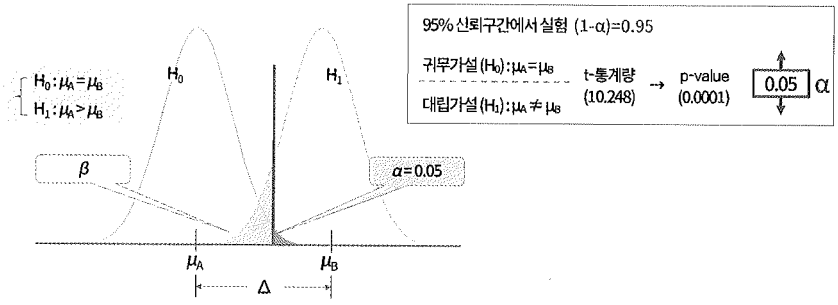
- 관찰된 표본으로부터 구하는 통계량, 검정 시 가설의 진위를 판단하는 기준

라) **유의수준**(Significance Level, α)

- 귀무가설을 기각하게 되는 확률의 크기로 '귀무가설이 옳은데도 이를 기각하는 확률의 크기'

마) 기각역(Critical Region, C)

- 귀무가설이 옳다는 전제 하에서 구한 검정통계량의 분포에서 확률이 유의수준 α인 부분(반대는 채택역(Acceptance Region))



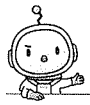
α의 크기를 0.05로 설정했다가 0.01로 줄인 경우 β값은 어떻게 될까요? 데이터마다 증가폭이 다르지만 일반적으론 증가합니다. 그래서 α값과 β값은 상충관계가 있다고 한 것입니다.

참고

제1종 오류와 제2종 오류

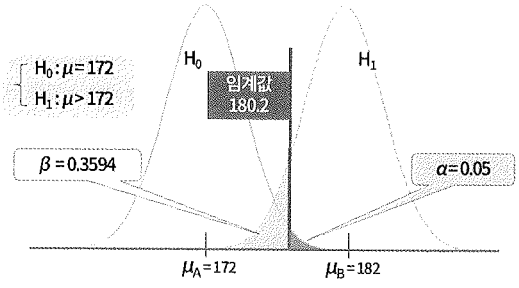
가설검정결과 / 정확한사실	H_0가 사실이라고 판정	H_0가 사실이 아니라고 판정
H_0가 사실임	옳은 결정	제1종 오류(α)
H_0가 사실이 아님	제2종 오류(β)	옳은 결정

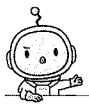
- 제1종 오류(Type I Error) : 귀무가설 H_0 가 옳은데도 귀무가설을 기각하게 되는 오류
- 제2종 오류(Type II Error) : 귀무가설 H_0 가 옳지 않은데도 귀무가설을 채택하게 되는 오류
- 두 가지 오류는 서로 상충관계가 있어서 일반적으로 가설검정에서는 제1종 오류 α의 크기를 0.1, 0.05, 0.01 등으로 고정시킨 뒤 제2종 오류 β가 최소가 되도록 기각역을 설정



예시

"데이터에듀 남자 사원의 평균 키(182cm)는 대한민국 남성 평균 키(172cm)보다 크다"에 대한 가설검정을 하고자 한다. 귀무가설 H0:μ=172, 대립가설 H1:μ>172을 세운 뒤, 모집단은 표준편차가 5인 정규분포를 따른다고 가정하였다. 데이터에듀 남자 사원의 평균 키(μ)는 182일 때, 유의수준 0.05에서의 임계값은 N(172,5)인 정규분포를 표준정규분포 Z로 변환하여 95%에 해당하는 값이므로 180.2로 계산된다. 데이터에듀 남자 사원의 평균 키인 182cm는 유의수준 0.05에서의 임계값인 180.2보다 큰 값이므로, 귀무가설 H0는 기각 된다. 따라서 "데이터에듀 남자 사원의 평균 키는 우리나라 남자 평균키보다 크다"고 할 수 있다.





⑥ 비모수 검정

통계적 검정에서 모집단의 모수에 대한 검정은 모수적 검정과 비모수적 검정으로 구분한다.

가. 모수적 방법

- 검정하고자 하는 **모집단의 분포에 대한 가정**을 하고, 그 가정하에서 검정통계량과 검정통계량의 분포를 유도해 검정을 실시하는 방법이다.

나. 비모수적 방법

비모수적 방법에 대해 해당하지 않는 것에 대한 내용의 문제가 출제됨으로 반드시 파악하고 넘어가시기 바랍니다.

- 자료가 **추출된 모집단의 분포에 대한 아무 제약을 가하지 않고 검정을 실시**하는 방법이다.
- 관측된 자료가 특정분포를 따른다고 가정할 수 없는 경우에 이용한다.
- 관측된 **자료의 수가 많지 않거나**(30개 미만) 자료가 개체 간의 **서열관계를 나타내는 경우**에 이용한다.

다. 모수적 검정과 비모수 검정의 차이점

1) 가설의 설정

가) 모수적 검정

- 가정된 분포의 모수에 대해 가설을 설정한다.

나) 비모수 검정

- **가정된 분포가 없으므로** 가설은 단지 '분포의 형태가 동일하다' 또는 '분포의 형태가 동일하지 않다'와 같이 **분포의 형태에 대해 설정**한다.

2) 검정 방법

가) 모수적 검정

- 관측된 자료를 이용해 구한 **표본평균, 표본분산** 등을 이용해 검정을 실시한다.

나) 비모수 검정

- 관측값의 절대적인 크기에 의존하지 않는 **관측값들의 순위(Rank)나 두 관측값 차이의 부호** 등을 이용해 검정한다.

라. 비모수 검정의 예

- 부호검정(Sign Test), 윌콕슨의 순위합 검정(Wilcoxon's Rank Sum Test), 윌콕슨의 부호 순위 검정(Wilcoxon's Signed Rank Test), 맨-휘트니의 U검정(Mann-Whitney U test), 런 검정(Run Test), 스피어만의 순위상관계수(Spearmans's rank correlation analysis)

2절 기초 통계분석

○ 학습 목표

- 기술통계의 정의를 이해한다.
- 통계량에 의한 자료정리와 R프로그램을 할 수 있다.
- 그래프에 의한 자료정리와 R프로그램을 할 수 있다.
- 상관관계 분석의 정의와 활용방법을 이해한다.

○ 눈높이 체크

✓ 기술통계에 대해 알고 계시나요?

데이터 분석에서 가장 먼저 수행되는 부문이 바로 기술통계입니다. 기술통계는 자료의 특성을 표, 그림 통계량 등을 사용하여 쉽게 파악할 수 있도록 정리/요약하는 통계분석 방법론입니다. 크게 기초통계량을 통한 방법과 그래프를 활용하는 방법으로 구분할 수 있습니다.

✓ 기술통계를 위한 기초통계량들은 어떤 것이 있을까요?

기술통계에 활용되는 통계량은 최솟값, 최댓값, 평균, 표준편차, 분산, 중앙값, 사분위수범위, 왜도, 첨도 등이 있습니다.

✓ 그래프를 활용한 기술통계방법에는 어떤 것이 있을까요?

그래프를 활용한 기술통계방법에는 막대그래프, 히스토그램, 줄기잎그림, 상자그림, 꺾은선그래프 등 다양한 그래프가 있습니다.

✓ 상관분석에 대해 알고 계신가요?

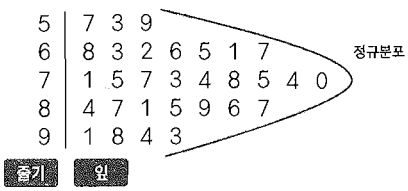
두 변수 간의 관계를 분석하기 위해서 공분산과 상관계수를 활용할 수 있습니다. 한 변수의 값이 증가할 때 상대변수의 값이 증가하면 양의 상관, 상대변수의 값이 감소하면 음의 상관이 있다고 해석하고 상관계수를 통해 상관성의 정도를 설명할 수 있습니다.

가. 기술통계의 정의

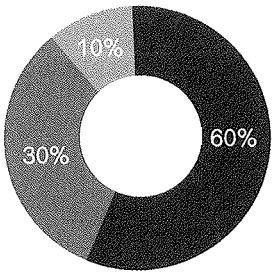
- 자료의 특성을 표, 그림, 통계량 등을 사용하여 쉽게 파악할 수 있도록 정리/요약하는 것이다.
- 자료를 요약하는 기초적 통계를 의미한다.
- 데이터 분석에 앞서 데이터의 대략적인 통계적 수치를 계산해 봄으로써 데이터에 대한 대략적인 이해와 앞으로 분석에 대한 통찰력을 얻기에 유리하다.

	N	최솟값	최댓값	평균	표준편차	분산
성별	100	1	2	1.49	0.52	0.27
혈액형	100	1	4	2.22	0.93	0.86
연령	100	19	26	22.9	2.43	5.90
학년	100	1	4	2.57	1.14	1.30
전공	100	1	5	1.89	0.86	0.74
교제기간	100	1	60	13.3	13.52	182.79
유효수(목록별)	100					

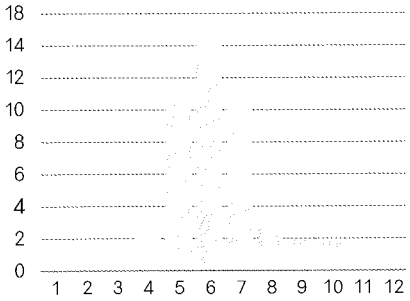
〈 기술 통계를 위한 기초 통계량(예시) 〉



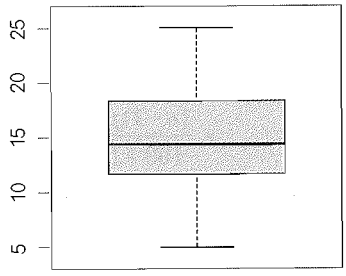
〈 줄기-잎 그림 〉



〈 도넛차트 〉



〈 히스토그램 〉



〈 상자수염그림 〉

나. 통계량에 의한 자료 정리

1) 중심위치의 측도

가) 자료(데이터) : $X_1, X_2, ..., X_n$

나) 표본평균(Sample Mean) : $\bar{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n) = \sum_{i=1}^{n} \frac{X_i}{n}$

다) 중앙값(Median) : 자료를 크기순으로 나열할 때 중앙에 위치하는 자료값이다. (중앙값의 순위는 $\frac{(n+1)}{2}$)

- n이 홀수인 경우 $\frac{(n+1)}{2}$
- n이 짝수인 경우 $\frac{n}{2}$ 번째값과 $\frac{n}{2}+1$ 번째 값의 평균

2) 산포의 측도

- 대표적인 산포도(Dispersion)는 분산, 표준편차, 범위 및 사분위수범위

가) 분산

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}(\sum_{i=1}^{n}(X_i^2 - n\bar{X}^2))$$

평균 절대 편차(Average Absolute Deviation, AAD 또는 Mean Absolute Deviation, MAD) :

평균과 개별 관측치 사이 거리(절댓값 편차)의 평균

수식 : $\sum|X_i - \bar{X}|$

나) 표준편차

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

다) 사분위수범위(Interquartile Range)

- IQR= Q3-Q1

라) 사분위수

- 제 1사분위수(Q1)=25백분위수
- 제 2사분위수(Q2)=50백분위수
- 제 3사분위수(Q3)=75백분위수

마) 백분위수(Percentile)

- $\frac{(n-1)p}{100+1}$번째 값

바) 변동계수(Coefficient of Variation) $V = \frac{S}{\bar{X}}$

사) 표본평균의 표준오차 $SE(\bar{X}) = \frac{S}{\sqrt{n}}$

예시

• 풀이

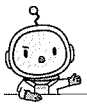
출근에 소요되는 시간(단위 : 분)

직원	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
시간	62	55	32	42	55	35	110	64	54	66	58	62	58	26	15

• 표본평균 : (62+55+31+……+15)/15 = 795/15 = 53
• 중앙값 : n이 홀수이므로 (n+1)/2인 8번째 값 55이다.

15 26 32 35 42 54 55 (55) 58 58 62 62 64 66 110

• 분산 : $(15-53)^2+(26-53)^2+\cdots\cdots+(110-53)^2/14 = 6870/14 = 487$
• 표준편차 : Sqrt(487) = 22.07
• 범위 : 최댓값-최솟값, 110-15 = 95
• 사분위수범위 : 62-38.5 = 23.5
• 사분위수 : Q1 = 38.5, Q2 = 55, Q3 = 62
• 변동계수 : 22.07/53 = 0.416
• 평균의 표준오차 : 22.07/√15



식까지 외울 필요는 없지만 왜도값이 주어졌을 때 어떻게 해석하는지 알아야 합니다. 왜도가 양수인 경우엔 왼쪽으로 밀집되어있고 오른쪽으로 긴 꼬리를 갖는 분포를 띄게 됩니다. 왜도가 음수인 경우는 오른쪽으로 밀집되어 있고 왼쪽에 긴꼬리를 갖게 됩니다. 왜도가 0일 경우 좌우대칭의 분포를 띄게 됩니다.

3) 분포의 형태에 관한 측도

가) 왜도

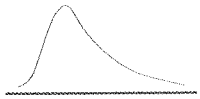
• 분포의 비대칭정도를 나타내는 측도이다.

왜도의 크기에 따라 평균값(Mean)과 중앙값(Median), 최빈값(Mode)이 변화가 있습니다. 왜도가 양수인 경우 최빈값<중앙값<평균 순으로 위치합니다. 왜도가 음수인 경우, 0인 경우도 체크하고 넘어갑시다!

$$m_3 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3}$$

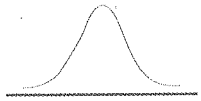
- $m_3 > 0$

  오른쪽으로 긴 꼬리를 갖는 분포



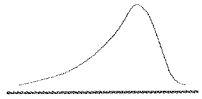
- $m_3 = 0$

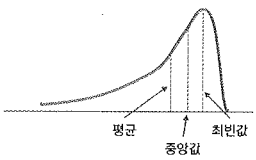
  좌우가 대칭인 분포



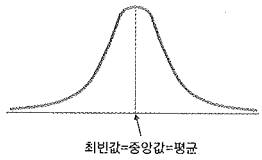
- $m_3 < 0$

  왼쪽으로 긴 꼬리를 갖는 분포

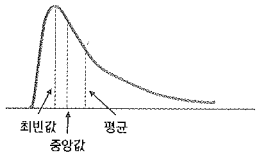




〈 왜도가 음수인 경우 〉



〈 왜도가 0인 경우 〉



〈 왜도가 양수인 경우 〉

첨도값을 보고 우리는 그 분포가 표준정규분포보다 더 뾰족한지 덜 뾰족한지 알 수 있습니다.

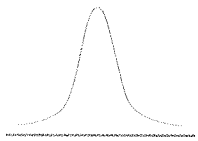
나) 첨도

• 분포의 중심에서 뾰족한 정도를 나타내는 측도이다.

$$m_4 = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3 = \frac{\mu_4}{\sigma^4} - 3$$

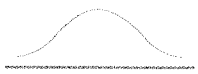
- $m_4 > 0$

  표준정규분포보다 더 뾰족함



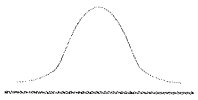
- $m_4 < 0$

  표준정규분포보다 덜 뾰족함



- $m_4 = 0$

  표준정규분포와 유사한 뾰족함



다. 그래프를 이용한 자료 정리

1) 히스토그램

• 표로 되어 있는 도수 분포를 그림으로 나타낸 것으로, 도수분포표를 그래프로 나타낸 것이다.

2) 막대그래프와 히스토그램의 비교

가) 막대그래프

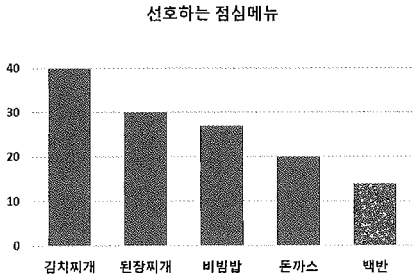
• **범주(Category)형**으로 구분된 데이터(예 : 직업, 종교, 음식 등)를 표현하며 범주의 순서를 의도에 따라 바꿀 수 있다.

나) 히스토그램

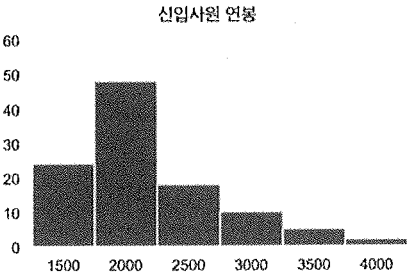
• **연속(Continuous)형**으로 표시된 데이터(예 : 몸무게, 성적, 연봉 등)를 표현하며 임의로 순서를 바꿀 수 없고 막대의 간격이 없다.

모자이크 플롯 (Mosaic Plot)

교차표(2원, 3원)를 시각화한 그래프로 사각형들이 그래프에 나열되고 사각형의 넓이는 범주에 속한 데이터 수(또는 비율)를 의미함



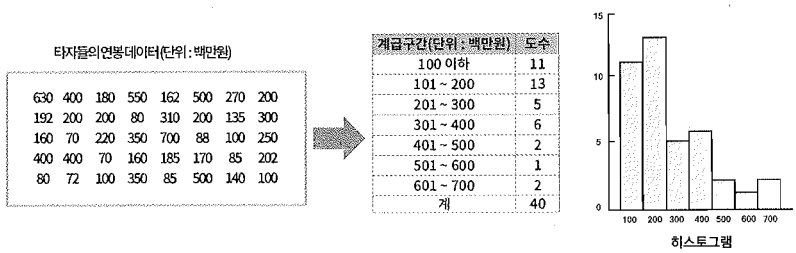
〈막대그래프〉



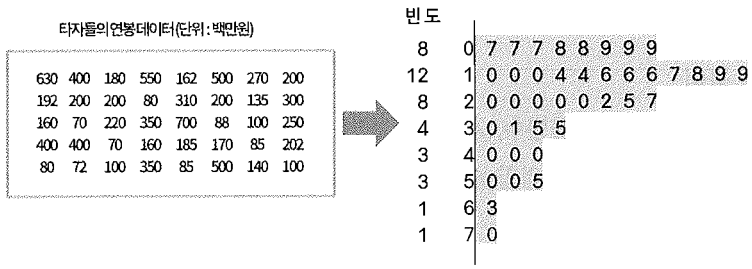
〈히스토그램〉

3) 히스토그램의 생성

- 데이터의 수를 활용해서 계급의 수와 계급간격을 계산하여 도수분포표를 만들고 히스토그램을 생성한다.
- 계급의 수는 $2^k \geq n$ 을 만족하는 최소의 정수 $\log_2 n = k$ 에서 최소의 정수이다. (k는 계급 수, n은 데이터 수)
- 계급의 간격은 (최댓값 - 최솟값)/계급수로 파악할 수 있다.
- 계급의 수와 간격이 변하면 히스토그램의 모양이 변한다.



4) 줄기-잎 그림(Stem-and Leaf Plot) : 데이터를 줄기와 잎의 모양으로 그린 그림



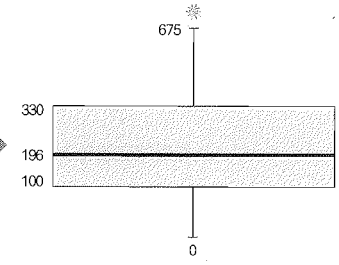
5) 상자그림(Box Plot): 다섯 숫자 요약을 통해 그림으로 표현 (최솟값, Q1, Q2, Q3, 최댓값)

- 사분위수범위(IQR) : Q3 - Q1
- 안울타리(Inner Fence) : Q1 − 1.5 × IQR 또는 Q3 + 1.5 × IQR

- 바깥울타리(Outer Fence) : Q1 − 3 × IQR 또는 Q3 + 3 × IQR
- 보통이상점(Mild Outlier) : 안쪽 울타리와 바깥 울타리 사이에 있는 자료
- 극단이상점(Extreme Outlier) : 바깥울타리 밖의 자료

타자들의연봉데이터(단위:백만원)

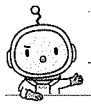
630	400	180	550	162	500	270	200
192	200	200	80	310	200	135	300
160	70	220	350	700	88	100	250
400	400	70	160	185	170	85	202
80	72	100	350	85	500	140	100



참고

R에서 활용되는 대표적 기술통계

R code	설 명
head(data명)	데이터를 기본 6줄 보여주어 데이터가 성공적으로 import되었는지 살펴볼 수 있다.
head(data명, n)	n에 숫자를 지정해주면 n번째 라인까지 살펴볼 수 있다.
summary(data명)	데이터 컬럼에 대한 전반적인 기초 통계량을 보여준다.
mean(data명$column명)	특정 컬럼의 평균을 알고 싶을 때 사용
median(data명$coulmn명)	특정 컬럼의 중앙값을 알고 싶을 때 사용
sd(data명$coulmn명)	특정 컬럼의 표준편차를 알고 싶을 때 사용
var(data명$coulmn명)	특정 컬럼의 분산을 알고 싶을 때 사용
quantile(data명$coulmn명)	특정 컬럼의 분위수를 알고 싶을 때 사용



2 인과관계의 이해

가. 용어

1) 종속변수(반응변수, y)
   - 다른 변수의 영향을 받는 변수

2) 독립변수(설명변수, x)
   - 영향을 주는 변수

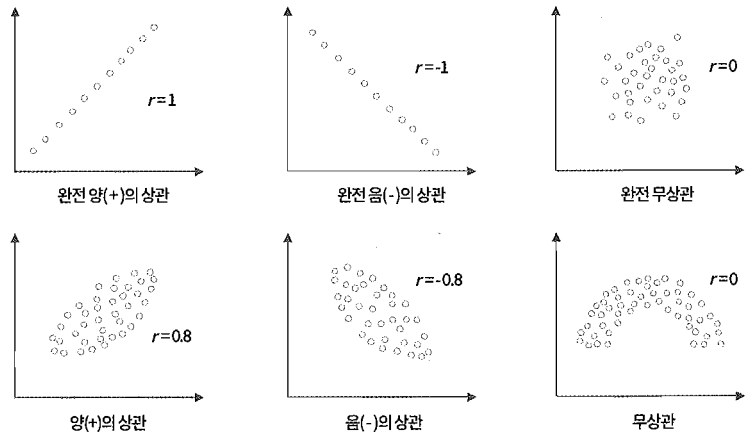
3) 산점도(Scatter Plot)
   - 좌표평면 위에 점들로 표현한 그래프

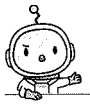
산점도에서 확인할 수 있는 사항에 대한 문제가 자주 출제되오니 꼭 기억하도록 합시다.

참고

산점도에서 확인할 사항

- 두 변수 사이의 선형관계(직선관계)가 성립하는가?
- 두 변수 사이의 함수관계(직선관계 또는 곡선관계)가 성립하는가?
- 이상값이 존재하는가?
- 몇 개의 집단으로 구분(층별)되는가?



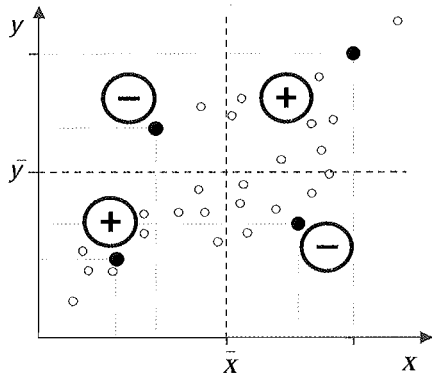


나. 공분산(Covariance)

- 두 확률변수 X , Y 의 방향의 조합(선형성)이다.

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

- 공분산의 부호만으로 두 변수 간의 방향성을 확인할 수 있다. 공분산의 부호가 +이면 두 변수는 양의 방향성, 공분산의 부호가 -이면 두 변수는 음의 방향성을 가진다.



No	X	Y
1	1	2
2	3	6
3	9	17
4	2	3
.	.	.
.	.	.
.	.	.
30	7	15

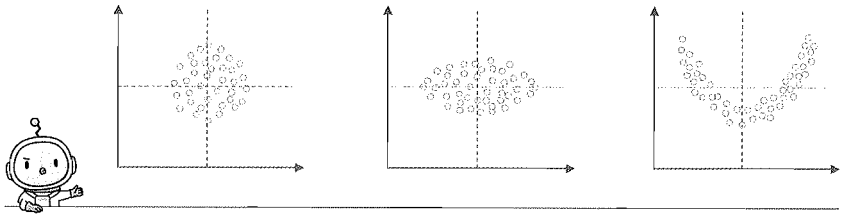
• X , Y 가 서로 독립이면, $Cov(X, Y) = 0$ 이다.

$Cov(X, Y) = \sigma_{XY} = E[XY] - E(X)E(Y)$

참고

공분산이 0인 산점도

• 두 변수 사이의 공분산이 0인 경우



③ 상관분석 (Correlation Analysis)

가. 상관분석의 정의

• 두 변수 간의 관계의 정도를 알아보기 위한 분석방법이다.

• 두 변수의 상관관계를 알아보기 위해 상관계수(Correlation Coefficient)를 이용하며, 그 공식은 아래와 같다.

$$\gamma = \frac{\mathrm{cov}(x, y)}{S_x \times S_y} = \frac{\sum_{i=1}^{n}[(x-\bar{x})(y-\bar{y})]}{(n-1)(S_x \times S_y)}$$

나. 상관관계의 특성

상관계수 범위	해 석
$0.7 < \gamma \leq 1$	강한 양(+)의 상관이 있음
$0.3 < \gamma \leq 0.7$	약한 양(+)의 상관이 있음
$0 < \gamma \leq 0.3$	거의 상관이 없음
$\gamma = 0$	상관관계(선형, 직선)가 존재하지 않음
$-0.3 \leq \gamma < 0$	거의 상관이 없음
$-0.7 \leq \gamma < -0.3$	약한 음(-)의 상관이 있음
$-1 \leq \gamma < -0.7$	강한 음(-)의 상관이 있음

피어슨과 스피어만을 구분할 때 Tip!
스피어만, 서열척도, 순서, 순위상관계수 등의 단어는 다 "ㅅ"(시옷)으로 시작합니다!

다. 상관분석의 유형

구분	피어슨	스피어만
개념	• 등간척도 이상으로 측정된 두 변수들의 상관관계 측정 방식	• 서열척도인 두 변수들의 상관관계 측정 방식
특징	• 연속형 변수, 정규성 가정 • 대부분 많이 사용	• 순서형 변수, 비모수적 방법 • 순위를 기준으로 상관관계 측정
상관계수	• 피어슨 γ(적률상관계수)	• 순위상관계수(ρ, 로우)

라. 상관분석을 위한 R 코드

구분	R code
분산	var(x, y = NULL, na.rm = FALSE)
공분산	cov(x, y = NULL, use = "everything", method = c("pearson", "kendall", "spearman"))
상관관계	cor(x, y = NULL, use = "everything", method = c("pearson", "kendall", "spearman"))
	• Hmisc 패키지의 rcorr 사용 rcorr(matrix(data명), type=c("pearson", "kendall", "spearman"))

x = 숫자형 변수
y = NULL(default) 또는 변수
na.rm = 결측값 처리

마. 상관분석의 가설 검정

- 상관계수 γ가 0이면 입력변수 x와 출력변수 y사이에는 아무런 관계가 없다. (귀무가설 : $\gamma = 0$, 대립가설 : $\gamma \neq 0$)
- t 검정통계량을 통해 얻은 p-value 값이 0.05이하인 경우, 대립가설을 채택하게 되어 우리가 데이터를 통해 구한 상관계수를 활용할 수 있게 된다.

바. 상관분석 예제

1) datasets 패키지의 "mtcars"라는 데이터셋의 마일(mpg), 총마력(hp)의 상관관계 분석을 실시한다.

R프로그램

```
data(mtcars)
a <- mtcars$mpg
b <- mtcars$hp
cor(a,b)
cov(a,b)
cor.test(a, b, method="pearson")
```

2) 결과 및 해석

```
> data(mtcars)
> a <- mtcars$mpg
> b <- mtcars$hp
> cov(a,b)
[1] -320.7321
> cor(a,b)
[1] -0.7761684
> cor.test(a, b, method="pearson")

        Pearson's product-moment correlation

data:  a and b
t = -6.7424, df = 30, p-value = 1.788e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8852686 -0.5860994
sample estimates:
       cor
-0.7761684
```

- mtcars 데이터셋의 mpg와 hp를 각각 a, b에 저장하여 mpg와 hp의 공분산, 상관계수를 구한 결과, 공분산은 -320.7321, 상관계수는 -0.7761684로 나타났다. 따라서, mpg와 hp는 공분산으로 음의 방향성을 가짐을 알 수 있고, 상관계수로 강한 음의 상관관계가 있음을 알 수 있다.
- cor.test를 이용해 mpg와 hp의 상관관계 분석을 실행한 결과, p-value가 1.788e-07로 유의수준 0.05보다 작게 나타나므로 mpg와 hp가 상관관계가 있다고 할 수 있다.

3절 회귀분석

회귀분석에선 개념에 대한 문제, R프로그램 실행 후 아웃풋을 해석하는 문제가 3~5 문제정도 출제됩니다. 내용을 반드시 숙지하도록 합시다.

○ 학습 목표

- 회귀분석의 정의와 가정을 이해한다.
- 회귀추정식의 통계적 가설 검증을 이해한다.
- R 프로그램을 통해 회귀분석을 활용하고 내용을 해석할 수 있다.
- 다중회귀분석에서 변수선택법을 이해하고 활용할 수 있다.

○ 눈높이 체크

✓ 회귀분석을 들어본 적 있으신가요?

우리 주변에서 일어나는 많은 인과현상들을 회귀분석을 통해 모형화하고 이를 활용하고 있습니다. 매출증대에 영향을 미치는 요소들, 난방비에 영향을 주는 요소들, 학습능력을 향상시키는 요소들 등 다양한 분야에서 회귀분석이 활용되고 있습니다.

✓ 단순회귀분석과 다중회귀분석을 이해하시나요?

하나의 요소가 결과에 미치는 영향을 모형화하는 방법은 단순회귀분석이고 여러 개의 요소가 결과에 미치는 영향을 모형화하는 것을 다중회귀분석이라 할 수 있습니다. 일반적으로 하나의 결과에는 여러가지 요소들이 영향을 미치므로 다중회귀분석이 많이 활용되고 있습니다.

✓ 회귀분석을 통계패키지로 구현해 본 적이 있으신가요?

회귀분석은 통계모델링에서 가장 많이 활용되고 있는 통계기법이기 때문에 대부분의 통계 패키지에서 회귀분석을 경험할 수 있습니다. SAS, SPSS 뿐만 아니라 R에서도 회귀분석은 쉽게 활용할 수 있습니다. 이번 강의를 통해 회귀분석을 이해하고 R을 통해 실습해 보도록 하겠습니다.

1 회귀분석의 개요

가. 회귀분석의 정의

- 하나나 그 이상의 독립변수들이 종속변수에 미치는 영향을 추정할 수 있는 통계기법이다.
- 변수들 사이의 인과관계를 밝히고 모형을 적합하여 관심있는 변수를 예측하거나 추론하기 위한 분석방법이다.
- 독립변수의 개수가 하나이면 단순선형회귀분석, 독립변수의 개수가 두 개 이상이면 다중선형회귀분석으로 분석할 수 있다.

나. 회귀분석의 변수

- 영향을 받는 변수(y) : 반응변수(Response Variable), 종속변수(Dependent Variable), 결과변수(Outcome Variable)
- 영향을 주는 변수(x) : 설명변수(Explanatory Variable), 독립변수(Independent Variable), 예측변수(Predictor Variable)

다. 선형회귀분석의 가정

1) 선형성

- 입력변수와 출력변수의 관계가 선형이다.(선형회귀분석에서 가장 중요한 가정)

2) 등분산성

- 오차의 분산이 입력변수와 무관하게 일정하다. 잔차플롯(산점도)을 활용하여 잔차와 입력변수간에 아무런 관련성이 없게 무작위적으로 고루 분포되어야 등분산성 가정을 만족하게 된다.

3) 독립성

- 입력변수와 오차는 관련이 없다. 자기상관(독립성)을 알아보기 위해 Durbin-Waston 통계량을 사용하며 주로 시계열 데이터에서 많이 활용한다.

4) 비상관성

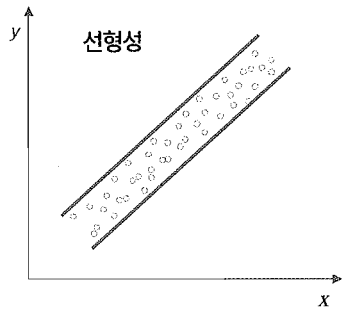
- 오차들끼리 상관이 없다.

5) 정상성(정규성)

- 오차의 분포가 정규분포를 따른다. Q-Q plot, Kolmogorov-Smirnov 검정, Shaprio-Wilk 검정 등을 활용하여 정규성을 확인한다.

라. 그래프를 활용한 선형회귀분석의 가정 검토

1) 선형성

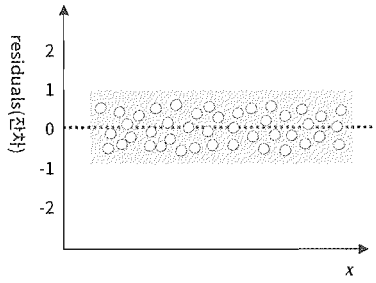


선형회귀모형에서는 왼쪽의 그래프와 같이 설명변수(x)와 반응변수(y)가 선형적 관계에 있음이 전제되어야 한다.

- Anderson-Darling Test : 콜모고로프-스미르노프 검정(K-S검정)을 수정한 적합도 검정으로 특정분포의 꼬리(Tail)에 K-S 검정보다 가중치를 더 두어 수행, 여러 분포의 적합도 검정이 가능하지만 정규성 검정에 강력하다고 알려짐
- D'Agostino-Pearson Test : 왜도와 첨도를 사용해 데이터가 정규분포를 따르는 지 검정함(표본의 크기가 20이상)
- Jarque-Bera Test : 정규분포의 기대 왜도와 첨도가 데이터에서 얻은 값과 일치하는지 검정

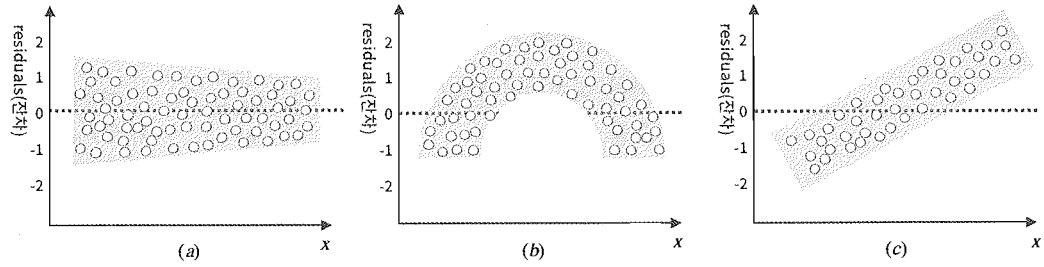
2) 등분산성

가) 등분산성을 만족하는 경우



설명변수(x)에 대한 잔차의 산점도를 그렸을 때, 왼쪽의 그림과 같이 설명변수(x) 값에 관계없이 잔차들의 변동성(분산)이 일정한 형태를 보이면 선형회귀분석의 가정 중 등분산성을 만족한다고 볼 수 있다.

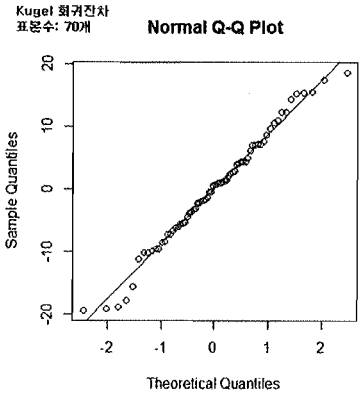
나) 등분산성을 만족하지 못하는 경우



- (a) : 설명변수(x)가 커질수록 잔차의 분산이 줄어드는 이분산의 형태
- (b) : 2차항 설명변수가 필요
- (c) : 새로운 설명변수가 필요

3) 정규성

- Q-Q Plot을 출력했을 때, 오른쪽의 그림과 같이 잔차가 대각방향의 직선의 형태를 지니고 있으면 잔차는 정규분포를 따른다고 할 수 있다.



마. 가정에 대한 검증

1) 단순선형회귀분석

- 입력변수와 출력변수간의 선형성을 점검하기 위해 산점도를 확인한다.

2) 다중선형회귀분석

- 선형회귀분석의 가정인 선형성, 등분산성, 독립성, 정상성이 모두 만족하는지 확인해야 한다.

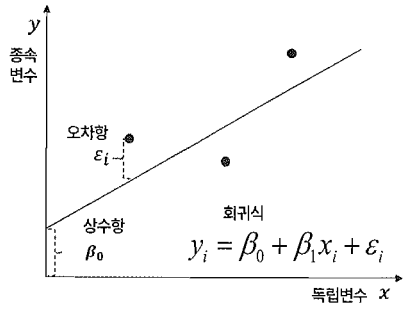
② 단순선형 회귀분석

- iid(독립이고 동일한 분포)
- independent(독립)
- identically(동일)
- distributed(분포)

- 하나의 독립변수가 종속변수에 미치는 영향을 추정할 수 있는 통계기법이다.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1,2,...,n,\ \varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$$

- y_i : i번째 종속변수 값
- x_i : i번째 독립변수 값
- β_0 : 선형 회귀식의 절편
- β_1 : 선형 회귀식의 기울기
- ε_i : 오차항, 독립적이며 $N(0,s^2)$ 의 분포를 이룬다.



가. 회귀분석에서의 검토사항

1) 회귀계수들이 유의미한가?

- 해당 계수의 t-통계량의 p-값이 0.05보다 작으면 해당 회귀계수가 통계적으로 유의하다고 볼 수 있다.

2) 모형이 얼마나 설명력을 갖는가?

- 결정계수(R^2)를 확인한다. 결정계수는 0~1값을 가지며, 높은 값을 가질수록 추정된 회귀식의 설명력이 높다.

3) 모형이 데이터를 잘 적합하고 있는가?

- 잔차를 그래프로 그리고 회귀진단을 한다.

나. 회귀계수의 추정(최소제곱법, 최소자승법)

- 측정값을 기초로 하여 적당한 제곱합을 만들고 그것을 최소로 하는 값을 구하여 측정결과를 처리하는 방법으로 잔차제곱이 가장 작은 선을 구하는 것을 의미한다.
- 추정식

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1,2,...,n,\ \varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$

$\sum_{i=1}^{n}(\varepsilon_i)^2 = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 X_i))^2$식을 각각 β_1과 β_0로 각각 편미분하여 0과 같다고 놓는다.

$\sum y_i = n\beta_0^* + \beta_1^* \sum x_i$

$\sum x_i y_i = \beta_0^* \sum x_i + \beta_1^* \sum x_i^2$

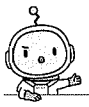
$\beta_0^* = E(y) - \beta_1^* E(x)$

$\beta_1^* = \dfrac{\sum_{i=1}^{n}(x_i - E(x))(y_i - E(y))}{\sum_{i=1}^{n}(x_i - E(x))^2}$

예시

- 10년간 에어컨 예약대수와 판매대수 (단위 : 1,000대)

예약대수(X)	19	23	26	29	30	38	39	46	49
판매대수(Y)	33	51	40	49	50	69	70	64	89



- 에어컨 판매대수에 대한 예약대수의 추정식 : 판매대수 = 6.41+1.53*예약대수

회귀분석에선 개념에 대한 문제, R프로그램 실행 후 Output을 해석하는 문제가 3~5 문제정도 출제됩니다. 내용을 반드시 숙지하도록 합시다.

다. 회귀분석의 검정

1) 회귀계수의 검정

- 회귀계수 β_1이 0이면 입력변수 x와 출력변수 y 사이에는 아무런 인과관계가 없다.
- 회귀계수 β_1이 0이면 적합된 추정식은 아무 의미가 없게 된다.

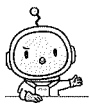
예시

- 10년간 에어컨 예약대수와 판매대수 (단위 : 1,000대)

예약대수(X)	19	23	26	29	30	38	39	46	49
판매대수(Y)	33	51	40	49	50	69	70	64	89

- 위의 데이터에 대해 단순회귀분석을 실시하여 검정을 실시한다.

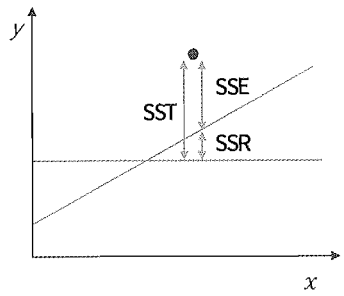
```
> x<-c(19, 23, 26, 29, 30, 38, 39, 46, 49)
> y<-c(33, 51, 40, 49, 50, 69, 70, 64, 89)
> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
      6.409        1.529
> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-12.766  -2.470  -1.764   4.470   9.412

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.4095     8.9272   0.718 0.496033
x             1.5295     0.2578   5.932 0.000581 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.542 on 7 degrees of freedom
Multiple R-squared:  0.8341,    Adjusted R-squared:  0.8104
F-statistic: 35.19 on 1 and 7 DF,  p-value: 0.0005805
```



- x의 회귀계수인 t-통계량에 대한 p-값이 0.000581로 나타나, 유의수준인 0.05보다 작으므로 회귀계수의 추정치들이 통계적으로 유의하다.

- 결정계수는 0.8341으로 높게 나타나 이 회귀식이 데이터를 적절하게 설명하고 있다고는 할 수 있다.
- 결정계수가 높아 데이터의 설명력이 높고 회귀분석결과에서 회귀식과 회귀계수들이 통계적으로 유의하므로 에어컨 판매대수를 에어컨 예약대수로 추정할 수 있다.
- 회귀분석 결과 "판매대수 = 6.4095 + 1.5295 * 예약대수"의 회귀식을 구할 수 있다.

2) 결정계수



$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

$$SSR = SST - SSE = \sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2$$

- 전체제곱합(Total Sum of Squares, SST) : $\sum_{i=1}^{n}(y_i - \bar{y})^2$
- 회귀제곱합(Regression Sum of Squares, SSR) : $\sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2$
- 오차제곱합(Error Sum of Squares, SSE) : $\sum_{i=1}^{n}(y_i - \hat{y_i})^2$
- 결정계수(R^2)는 전체제곱합에서 회귀제곱합의 비율(SSR/SST), $0 \leq R^2 \leq 1$(여기서 SST = SSR+SSE)
- 결정계수(R^2)는 전체 데이터를 회귀모형이 설명할 수 있는 설명력을 의미한다. (단순회귀분석에서 결정계수는 상관계수 r의 제곱과 같다.)

3) 회귀직선의 적합도 검토

- 결정계수(R^2)를 통해 추정된 회귀식이 얼마나 타당한지 검토한다.
  (결정계수(R^2)가 1에 가까울수록 회귀모형이 자료를 잘 설명함)
- 독립변수가 종속변수 변동의 몇 %를 설명하는지 나타내는 지표이다.
- 다변량 회귀분석에서는 독립변수의 수가 많아지면 결정계수(R^2)가 높아지므로 독립변수가 유의하든, 유의하지 않든 독립변수의 수가 많아지면 결정계수가 높아지는 단점이 있다.

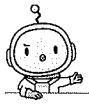
- 이러한 결정계수의 단점을 보완하기 위해 수정된 결정계수(R_a^2 : adjusted R^2)를 활용한다. 수정된 결정계수는 결정계수보다 작은 값으로 산출되는 특징이 있다.
- 수정된 결정계수 $R_a^2 = 1 - \frac{(n-1)(1-R^2)}{n-k-1} = 1 - \frac{(n-1)\left(\frac{SSE}{SST}\right)}{n-k-1} = 1-(n-1)\frac{MSE}{SST}$

(k : 독립변수 개수, n : 데이터의 개수)

참고

오차(Error)와 잔차(Residual)의 차이

- 오차 : 모집단에서 실제값이 회귀선과 비교해 볼 때 나타나는 차이(정확치와 관측치의 차이)
- 잔차 : 표본에서 나온 관측값이 회귀선과 비교해 볼 때 나타나는 차이.
  회귀모형에서 오차항은 측정할 수 없으므로 잔차를 오차항의 관찰값으로 해석하여 오차항에 대한 가정들의 성립 여부를 조사함



③ 다중선형 회귀분석

가. 다중선형회귀분석(다변량회귀분석)

1) 다중회귀식

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

2) 모형의 통계적 유의성

- 모형의 통계적 유의성은 F-통계량으로 확인한다.
- 유의수준 5% 하에서 F-통계량의 p-값이 0.05보다 작으면 추정된 회귀식은 통계적으로 유의하다고 볼 수 있다.

〈귀무가설 : $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ vs 대립가설 : $H_1 : \beta_1 \neq \beta_2 \neq \ldots \neq \beta_k \neq 0$〉

요인	제곱합	자유도	제곱평균	F-통계량
회귀	회귀제곱합(SSR)	k	MSR=SSR/k	F=MSR/MSE
오차	오차제곱합(SSE)	n-k-1	MSE=SSE/(n-k-1)	
계	전체제곱합(SST)	n-1		

- F-통계량이 크면 p-value가 0.05보다 작아지고 이렇게 되면 귀무가설을 기각한다. 즉, 모형이 유의하다고 결론지을 수 있다.

3) 회귀계수의 유의성

- 회귀계수의 유의성은 단변량회귀분석의 회귀계수 유의성 검토와 같이 t-통계량을 통해 확인한다.

- 모든 회귀계수의 유의성이 통계적으로 검증되어야 선택된 변수들의 조합으로 모형을 활용할 수 있다.

4) 모형의 설명력

- 결정계수(R^2)나 수정된 결정계수(R_a^2)를 확인한다.

5) 모형의 적합성

- 모형이 데이터를 잘 적합하고 있는지 잔차와 종속변수의 산점도로 확인한다.

6) 데이터가 전제하는 가정을 만족시키는가?

- 선형성, 독립성, 등분산성, 비상관성, 정상성

7) 다중공선성(Multicollinearity)

- 다중회귀분석에서 설명변수들 사이에 선형관계가 존재하면 회귀계수의 정확한 추정이 곤란하다.
- 다중공선성 검사 방법

  가) 분산팽창요인(VIF) : 4보다 크면 다중공선성이 존재한다고 볼 수 있고, 10보다 크면 심각한 문제가 있는 것으로 해석할 수 있다.

  나) 상태지수 : 10 이상이면 문제가 있다고 보고, 30보다 크면 심각한 문제가 있다고 해석할 수 있다.

  - 다중선형회귀분석에서 다중공선성의 문제가 발생하면, 문제가 있는 변수를 제거하거나 주성분회귀, 능형회귀 모형을 적용하여 문제를 해결한다.

④ 회귀분석의 종류

종류	모형	
단순회귀	$Y = \beta_0 + \beta_1 X + \varepsilon$	독립변수가 1개이며 종속변수와의 관계가 직선
다중회귀	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$	독립변수가 k개이며 종속변수와의 관계가 선형 (1차 함수)
로지스틱 회귀	$P(y) = \frac{1}{1+\exp[-(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k)]}$	종속변수가 범주형(2진변수)인 경우에 적용되며, 단순 로지스틱 회귀 및 다중, 다항 로지스틱 회귀로 확장할 수 있음

다항회귀	k=2이고 2차 함수인 경우 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \varepsilon$	독립변수와 종속변수와의 관계가 1차 함수 이상인 관계(단, k=1이면 2차 함수 이상)
곡선회귀	2차 곡선인 경우 $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$ 3차 곡선인 경우 $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$	독립변수가 1개이며 종속변수와의 관계가 곡선
비선형 회귀	$Y = \alpha e^{-\beta X} + \varepsilon$	회귀식의 모양이 미지의 모수들의 선형관계로 이뤄져 있지 않은 모형

⑤ 회귀분석 사례

가. R 프로그램을 통한 회귀분석

1) 분석내용 : MASS 패키지의"Cars93"라는 데이터셋의 가격(Price)를 종속변수로 선정하고 엔진 크기(Engine-Size), RPM, 무게(Weight)를 이용해서 다중회귀분석을 실시한다.

R 프로그램

```
library(MASS)
head(Cars93)
attach(Cars93)
lm(Price~EngineSize+RPM+Weight, data=Cars93)
summary(lm(Price~EngineSize+RPM+Weight, data=Cars93))
```

2) 결과 및 해석

```
> lm(Price~EngineSize+RPM+Weight, data=Cars93)
Call:
lm(formula = Price ~ EngineSize + RPM + Weight, data = Cars93)
Coefficients:
(Intercept)   EngineSize          RPM       Weight
-51.793292     4.305387     0.007096     0.007271

> summary(lm(Price~EngineSize+RPM+Weight, data=Cars93))

Call:
lm(formula = Price ~ EngineSize + RPM + Weight, data = Cars93)
Residuals:
  Min      1Q  Median      3Q     Max
-10.511  -3.806  -0.300   1.447  35.255
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -51.793292  9.106309  -5.688 1.62e-07 ***
EngineSize  4.305387  1.324961  3.249  0.00163 **
RPM        0.007096   0.001363  5.208 1.22e-06 ***
Weigh       0.007271  0.002157  3.372  0.00111 **
--Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.504 on 89 degrees of freedom
Multiple R-squared:  0.5614,    Adjusted R-squared:  0.5467
F-statistic: 37.98 on 3 and 89 DF,  p-value: 6.746e-16
```

- 여기서 F-통계량은 37.98이며 유의확률 p-value 값이 6.746e-16로 유의수준 5% 하에서 추정된 회귀 모형이 통계적으로 매우 유의함을 알 수 있다.
- 결정계수와 수정된 결정계수는 각각 0.5614, 0.5467로 조금 낮게 나타나 이 회귀식이 데이터를 적절하게 설명하고 있다고는 할 수 없다.
- 회귀계수들의 p-값들이 0.05보다 작으므로 회귀계수의 추정치들이 통계적으로 유의하다.
- 결정계수가 낮아 데이터의 설명력은 낮지만 회귀분석 결과에서 회귀식과 회귀계수들이 통계적으로 유의하여 자동차의 가격을 엔진의 크기와 RPM 그리고 무게로 추정할 수 있다.

나. R 프로그램을 통한 로지스틱 회귀분석의 사례

1) 데이터 설명

- 림프절이 전립선 암에 대해 양성인지 여부를 예측하는 데이터

변수명	설 명
양성여부(r)	전립선암에 대한 양성 여부
age	환자의 연령
stage	질병 단계 : 질병이 얼마나 진행되어 있는지 나타내는 척도
grade	종양의 등급 : 진행의 정도
xray	X-선 결과
acid	특정한 부위에 종양이 전이되었을 때 상승되는 혈청의 인산염값

2) 분석 결과

```
> library(boot)
> data(nodal)
> a<-c(2,4,6,7)
> data <- nodal[,a]
> glmModel <- glm(r~., data=data, family="binomial")
> summary(glmModel)

Call: glm(formula = r ~ ., family = "binomial", data = data)
Deviance Residuals:
Min       1Q   Median       3Q      Max
-2.1231  -0.6620  -0.3039   0.4710   2.4892
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.0518     0.8420  -3.624  0.00029 ***
stage         1.6453     0.7297   2.255  0.02414 *
xray          1.9116     0.7771   2.460  0.01390 *
acid          1.6378     0.7539   2.172  0.02983 *
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 70.252  on 52  degrees of freedom
Residual deviance: 49.180  on 49  degrees of freedom
AIC: 57.18

Number of Fisher Scoring iterations: 5
```

- 2번째 변수인 양성여부를 종속변수로 두고 5개의 변수를 독립변수로 하여 로지스틱 회귀 분석을 실시한 결과 age와 grade는 유의수준 5%하에서 유의하지 않아 이를 제외한 3개 변수 stage, xray와 acid 를 활용해서 모형을 개발한다.
- stage, xray와 acid의 추정계수는 유의수준 5% 하에서 유의하게 나타나므로 p(r=1)=1/(1+e-(-3.05+1.65stage+1.91xray+1.64acid))의 선형식이 가능하다.

⑥ 최적회귀방정식

가. 최적회귀방정식의 선택

1) 설명변수 선택

- 필요한 변수만 상황에 따라 타협을 통해 선택한다.
- y에 영향을 미칠 수 있는 모든 설명변수 x들을 y의 값을 예측하는데 참여한다.

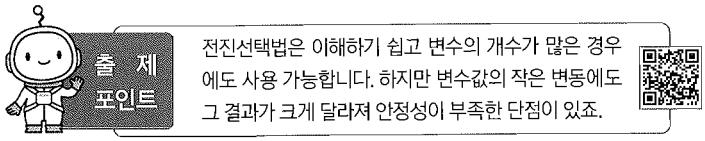
- 데이터에 설명변수 x들의 수가 많아지면 관리하는데 많은 노력이 요구되므로, 가능한 범위 내에서 적은 수의 설명변수를 포함한다.

2) 모형선택(Exploratory Analysis) : 분석 데이터에 가장 잘 맞는 모형을 찾아내는 방법이다.

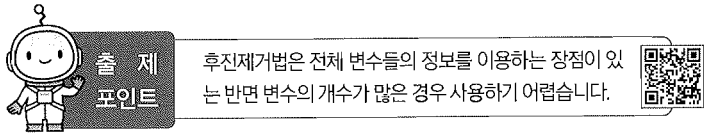
- 모든 가능한 조합의 회귀분석(All Possible Regression) : 모든 가능한 독립변수들의 조합에 대한 회귀모형을 생성한 뒤 가장 적합한 회귀모형을 선택한다.

3) 단계적 변수선택(Stepwise Variable Selection)

- 전진선택법(Forward Selection) : 절편만 있는 상수모형으로부터 시작해 중요하다고 생각되는 설명변수부터 차례로 모형에 추가한다.



- 후진제거법(Backward Elimination) : 독립변수 후보 모두를 포함한 모형에서 출발해 가장 적은 영향을 주는 변수부터 하나씩 제거하면서 더 이상 제거할 변수가 없을 때의 모형을 선택한다.



- 단계선택법(Stepwise Method) : 전진선택법에 의해 변수를 추가하면서 새롭게 추가된 변수에 기인해 기존 변수의 중요도가 약화되면 해당 변수를 제거하는 등 단계별로 추가 또는 제거되는 변수의 여부를 검토해 더 이상 없을 때 중단한다.

나. 벌점화된 선택기준

1) 개요

- 모형의 복잡도에 벌점을 주는 방법으로 AIC 방법과 BIC 방법이 주로 사용된다.

2) 방법

- AIC(Akaike Information Criterion)

$$AIC = -2\sum_{i=1}^{n} l(y_i, x_i^T\hat{\beta})/n + 2k/n,\ k$$

k는 모수의 개수, n은 자료의 수

- BIC(Bayesian Information Criterion)

$$BIC = -2\sum_{i=1}^{n} l(y_i, x_i^T\hat{\beta})/n + klog(n)/n$$

k는 모수의 개수, n은 자료의 수

가장 최적화된 모형을 선택하기 위해서는 AIC와 BIC가 최소가 되는 모형을 선택해야 함을 잊지 마세요!

3) 설명

- 모든 후보 모형들에 대해 AIC 또는 BIC를 계산하고 그 값이 최소가 되는 모형을 선택한다.
- 모형선택의 일치성(Consistency Inselection) : 자료의 수가 늘어날 때 참인 모형이 주어진 모형 선택 기준의 최소값을 갖게 되는 성질이다.
- 이론적으로 AIC에 대해서 일치성이 성립하지 않지만 BIC는 주요 분포에서 이러한 성질이 성립한다.
- AIC를 활용하는 방법이 보편화된 방법이다.
- 그밖의 벌점화 선택기준으로 RIC(Risk Inflation Criterion), CIC(Covariance Inflation Criterion), DIC(Deviation Information Criterion)가 있다.

다. 최적회귀방정식의 사례

1) 변수 선택법 예제(유의확률 기반)

- x1, x2, x3, x4를 독립변수로 가지고 y를 종속변수로 가지는 선형회귀모형을 생성한 뒤, step( ) 함수를 이용하지 않고 직접 후진제거법을 적용하는 R코드를 작성하여 변수제거를 수행해보자.

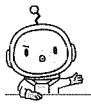
```
> # 1) 데이터 프레임 생성
> x1 <- c(7, 1, 11, 11, 7, 11, 3, 1, 2,21, 1,11, 10)
> x2 <- c(26, 29, 56, 31, 52, 55, 71,31, 54, 47, 40, 66, 68)
> x3 <- c(6, 15, 8, 8, 6, 9, 17, 22, 18, 4, 23, 9, 8)
> x4 <- c(60, 52, 20, 47, 33, 22, 6, 44, 22, 26, 34, 12, 12)
> y <- c(78.5, 74.3, 104.3, 87.6, 95.9, 109.2, 102.7, 72.5, 93.1,
115.9, 83.8, 113.3, 109.4)
> df <- data.frame(x1, x2, x3, x4, y)
> head(df)
  x1 x2 x3 x4     y
1  7 26  6 60  78.5
2  1 29 15 52  74.3
3 11 56  8 20 104.3
4 11 31  8 47  87.6
5  7 52  6 33  95.9
6 11 55  9 22 109.2
>
> # 2) 회귀모형(a) 생성
> a <- lm(y ~ x1 + x2 + x3 + x4, data=df)
> summary(a)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1750 -1.6709  0.2508  1.3783  3.9254

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.4054    70.0710   0.891   0.3991
x1            1.5511     0.7448   2.083   0.0708 .
x2            0.5102     0.7238   0.705   0.5009
x3            0.1019     0.7547   0.135   0.8959
x4           -0.1441     0.7091  -0.203   0.8441
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.446 on 8 degrees of freedom
Multiple R-squared:  0.9824,    Adjusted R-squared:  0.9736
F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```

- summary(a)에서 모형의 유의성을 판단하기 위해 F-통계량을 확인한 결과, 111.5로 나타났으며 유의확률이 4.756e-07임으로 통계적으로 유의하게 나타났다. 하지만 각각의 입력변수들의 통계적 유의성을 검토해 본 결과, t-통계량을 통한 유의확률이 0.05 보다 작은 변수가 하나도 존재하지 않아 모형을 활용할 수 없다고 판단되었다. 적절한 모형을 선정하기 위해 유의확률이 가장 높은 x3 을 제외하고 다시 회귀모형을 생성해 보았다.

참고

- 부동소수점(Floating Point) : 컴퓨터에서 실수를 표시하는 방법으로 (가수)(밑수)^(지수)와 같은 형태로 표현(가수는 유효숫자, 지수는 소수점 위치를 나타냄)

부동소수점	가수	(밑수)^(지수)	값
1e+02	1	10^(2)	100
1e+01	1	10^(1)	10
1e+00	1	10^(0)	1
1e-01	1	10^(-1)	0.1
1e-02	1	10^(-2)	0.01



- 예를 들어, 0.312e+02는 0.312×10^(2)=0.312×100=31.2를 의미하고, 4.756e-07은 4.756×10^(-7)=4.756×0.0000001=0.00004756을 의미한다

```
> # 3) 유의확률이 가장 높은 변수를 제거하고 다시 회귀모형(b)을 생성
> b <- lm(y ~ x1 + x2 + x4, data=df)
> summary(b)

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

Residuals:
    Min      1Q  Median      3Q     Max
-3.0919  -1.8016  0.2562  1.2818  3.8982

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  71.6483    14.1424   5.066 0.000675 ***
x1            1.4519     0.1170  12.410 5.78e-07 ***
x2            0.4161     0.1856   2.242 0.051687 .
x4           -0.2365     0.1733  -1.365 0.205395
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.309 on 9 degrees of freedom
Multiple R-squared:  0.9823,    Adjusted R-squared:  0.9764
F-statistic: 166.8 on 3 and 9 DF,  p-value: 3.323e-08
```

- x3 변수를 제거한 후, 모형의 유의성을 다시 검토한 결과 F-통계량에 대한 유의확률은 통계적으로 유의하게 나타났다. 모든 변수들의 t-통계량에 대한 유의확률이 0.05보다 낮아야 하지만 x1을 제외한 2개 변수의 유의확률이 0.05보다 높게 나타나 유의하지 않은 결과를 보였다. 따라서 유의확률이 가장 높은 x4 변수를 제외하고 회귀모형을 다시 생성하였다.

```
> # 4) 유의확률이 가장 높은 변수를 제거하고 다시 회귀모형(c)을 생성
> c <- lm(y ~ x1 + x2, data=df)
> summary(c)
```

```
Call:
lm(formula = y ~ x1 + x2, data = df)

Residuals:
   Min     1Q Median     3Q    Max

-2.893 -1.574 -1.302  1.363  4.048

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 52.57735    2.28617   23.00 5.46e-10 ***
x1           1.46831    0.12130   12.11 2.69e-07 ***
x2           0.66225    0.04585   14.44 5.03e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.406 on 10 degrees of freedom
Multiple R-squared:  0.9787,    Adjusted R-squared:  0.9744
F-statistic: 229.5 on 2 and 10 DF,  p-value: 4.407e-09
```

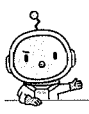
- F-통계량을 통해 유의수준 0.05 하에서 모형이 통계적으로 유의함을 확인할 수 있다.
- 다변량회귀분석에 선정된 x1, x2 변수에 대한 각각의 유의확률 값이 모두 통계적으로 유의하게 나타났다. 수정된 결정계수는 0.9744로 선정된 다변량회귀식이 전체 데이터의 97.44%를 설명하고 있는 것을 확인할 수 있었다.
- 위의 후진제거법을 통해 최종적으로 얻게 된 추정된 회귀식은 y = 52.57735 + 1.46831x1 + 0.66225x2이다.

2) 변수 선택법 예제(벌점화 전진선택법)

- 이번에는 step 함수를 사용하여 전진선택법을 적용하는 R코드를 작성하여 변수 제거를 수행해보자.

참고

- step(lm(출력변수~입력변수, 데이터세트), scope=list(lower=~1, upper=~입력변수), direction="변수선택방법")
- scope - 변수선택 과정에서 설정할 수 있는 가장 큰 모형 혹은 가장 작은 모형을 설정. scope가 없을 경우 전진선택법에서는 현재 선택한 모형을 가장 큰 모형으로, 후진제거법에서는 상수항만 있는 모형을 가장 작은 모형으로 설정한다.
- direction - 변수선택법(forward : 전진선택법, backward : 후진제거법, stepwise : 단계적선택법)
- k - 모형선택 기준에서 AIC, BIC와 같은 옵션을 사용. k=2 이면 AIC, k=log(자료의 수)이면 BIC



```
> # step함수를 이용한 전진선택법의 적용
> step(lm(y~1, data=df), scope=list(lower=~1,
upper=~x1+x2+x3+x4), direction="forward")
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

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

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

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

Coefficients:
(Intercept)           x4           x1           x2
    71.6483      -0.2365       1.4519       0.4161
```

- 벌점화 방식을 적용한 전진선택법을 실시한 결과, 가장 먼저 선택된 변수는 AIC값이 58.852으로 가장 낮은 x4였다. x4에 x1을 추가하였을 때 AIC 값이 28.742로 낮아지게 되었고, x2를 추가하였을 때 AIC 값이 24.974으로 최소화되어 더 이상 AIC를 낮출 수 없어 변수 선택을 종료하게 되었다.
- 최종적으로 선택된 추정된 회귀식은 y = 71.6483 -0.2365x4 + 1.4519 x1 + 0.4161x2 이다.

3) 변수 선택법 예제(벌점화 후진제거법)

가) 활용데이터

- 전립선암 자료(8개의 입력변수와 1개의 출력변수로 구성)
- 마지막 열에 있는 변수는 학습자료인지 예측자료인지를 나타내는 변수로 이번 분석에서는 사용하지 않는다.

변수명	설 명
lcavol	종양 부피의 로그
lweight	전립선 무게의 로그
age	환자의 연령
lbph	양성 전립선 증식량의 로그
svi	암이 정낭을 침범할 확률
lcp	capsular penetration의 로그값
gleason	Gleason 점수
pgg45	Gleason 점수가 4 또는 5인 비율
lpsa	전립선 수치의 로그

R 프로그램

```
> library(ElemStatLearn)
> Data = prostate
> data.use = Data[,-ncol(Data)]
> lm.full.Model = lm(lpsa~., data=data.use)
```

나) 후진제거법에서 AIC를 이용한 변수선택

```
> backward.aic = step(lm.full.Model, lpsa~1, direction="backward")
Start:  AIC=-60.78
lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45
```

```
     Df Sum of Sq    RSS     AIC
- gleason  1    0.0491 43.108 -62.668
- pgg45    1    0.5102 43.569 -61.636
- lcp      1    0.6814 43.740 -61.256
<none>                 43.058 -60.779
- lbph     1    1.3646 44.423 -59.753
- age      1    1.7981 44.857 -58.810
- lweight  1    4.6907 47.749 -52.749
- svi      1    4.8803 47.939 -52.364
- lcavol   1   20.1994 63.258 -25.467
Step:  AIC=-62.67
lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45

          Df Sum of Sq    RSS     AIC
- lcp      1    0.6684 43.776 -63.176
<none>                 43.108 -62.668
- pgg45    1    1.1987 44.306 -62.008
- lbph     1    1.3844 44.492 -61.602
- age      1    1.7579 44.865 -60.791
- lweight  1    4.6429 47.751 -54.746
- svi      1    4.8333 47.941 -54.360
- lcavol   1   21.3191 64.427 -25.691

Step:  AIC=-63.18
lpsa ~ lcavol + lweight + age + lbph + svi + pgg45

          Df Sum of Sq    RSS     AIC
- pgg45    1    0.6607 44.437 -63.723
<none>                 43.776 -63.176
- lbph     1    1.3329 45.109 -62.266
- age      1    1.4878 45.264 -61.934
- svi      1    4.1766 47.953 -56.336
- lweight  1    4.6553 48.431 -55.373
- lcavol   1   22.7555 66.531 -24.572

Step:  AIC=-63.72
lpsa ~ lcavol + lweight + age + lbph + svi

          Df Sum of Sq    RSS     AIC
<none>                 44.437 -63.723
- age      1    1.1588 45.595 -63.226
- lbph     1    1.5087 45.945 -62.484
- lweight  1    4.3140 48.751 -56.735
- svi      1    5.8509 50.288 -53.724
- lcavol   1   25.9427 70.379 -21.119
```

- 맨처음 AIC는 -62.67로 gleason을 제거하고 회귀분석 실시, 그 다음 차례로 lcp, pgg45 순서로 제거되어 회귀분석이 실시된다.

4절 시계열 분석

시계열 분석의 시계열 자료의 종류와 정상성에 대한 개념을 정리하고 있습니다. 그리고 시계열 모형들에 대한 이해가 중요합니다.

○ 학습 목표

- 시계열 자료를 이해한다.
- 정상 시계열과 비정상 시계열을 구분할 수 있다.
- ARIMA모형과 분해 시계열 분석을 할 수 있다.
- R 프로그램을 통해 시계열 분석과 예측을 할 수 있다.

○ 눈높이 체크

✔ 시계열 자료는 어떻게 구분할까요?

시간의 흐름에 따라 관찰된 데이터를 시계열 데이터 또는 시계열 자료라고 합니다. 이러한 시계열 자료에는 주식가격 데이터, 실업률, 기후데이터 등 우리 주위에서 많이 찾아 볼 수 있습니다.

✔ 시계열 자료의 정상성을 구분할 수 있나요?

대부분의 시계열 자료는 비정상성 데이터 입니다. 시계열 자료를 통해 미래를 예측하기 위해서는 비정상성 데이터를 정상성 데이터로 변화하여 분석모형을 설계할 수 있습니다. 그렇다면 정상성의 기준이 무엇일까요? 본문에서 자세히 확인해 보도록 하겠습니다.

✔ 시계열 분석에 대해 알고 계신가요?

시계열 분석은 시계열 자료를 통해 미래를 예측하거나 시계열 데이터의 특성을 파악하는 것을 의미합니다. 시계열 분석은 자기회귀모형과 이동평균모형으로 구분됩니다.

✔ 회귀분석을 이해하고 계신가요?

시계열 분석은 통계분석의 한 방법이지만 고급통계분석에 해당됩니다. 시계열 분석을 이해하기 위해서는 회귀분석과 상관분석에 대해 이해하고 계셔야 됩니다. 간단히 내용을 확인하고 강의에 들어가면 훨씬 쉽게 내용을 이해할 수 있을 것입니다.

① 시계열 자료

가. 개요

- 시간의 흐름에 따라 관찰된 값들을 시계열 자료라 한다.
- 시계열 데이터의 분석을 통해 미래의 값을 예측하고 경향, 주기, 계절성 등을 파악하여 활용한다.

나. 시계열 자료의 종류

1) 비정상성 시계열 자료

- 시계열 분석을 실시할 때 다루기 어려운 자료로 대부분의 시계열 자료가 이에 해당한다.

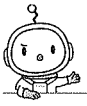
2) 정상성 시계열 자료

- 비정상 시계열을 핸들링해 다루기 쉬운 시계열 자료로 변환한 자료이다.

참고

시계열 자료의 역사

- 17세기 태양의 흑점 자료나 밀 가격지수 변동을 나타내는 함수로 sin, cos 곡선 활용
- Yule(1926) - ARMA 개념 제시, Walker(1937) - ARMA 모형 제시
- Durbin(1960), Box & Jenkins(1970) - ARMA 모형에 대한 추정
- Holt(1957) - 지수평활법(Exponential Smoothing) 제시
- Winter(1960) - 계절성(Seasonal) 지수평활법 제시



정상성은 평균이 일정할 때, 분산이 일정할 때, 공분산도 단지 시차에만 의존하고 실제 특정 시점 t, s에는 의존하지 않을 때 만족합니다.

② 정상성

가. 평균이 일정할 경우

- 모든 시점에 대해 일정한 평균을 가진다.
- 평균이 일정하지 않은 시계열은 **차분(Difference)을 통해 정상화**할 수 있다.

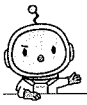
나. 분산이 일정

- 분산도 시점에 의존하지 않고 일정해야 한다.
- 분산이 일정하지 않을 경우 **변환(Transformation)을 통해 정상화**할 수 있다.

다. 공분산도 단지 시차에만 의존, 실제 특정 시점 t, s에는 의존하지 않는다.

참고

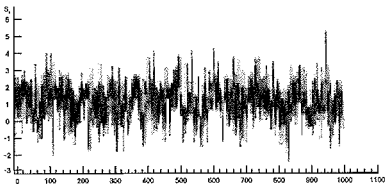
차분이란?



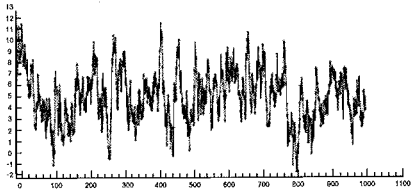
- 차분은 현시점 자료에서 전 시점 자료를 빼는 것이다.
- 일반차분(Regular Difference) : 바로 전 시점의 자료를 빼는 방법이다.
- 계절차분(Seasonal Difference) : 여러 시점 전의 자료를 빼는 방법, 주로 계절성을 갖는 자료를 정상화 하는데 사용한다.

라. 정상 시계열의 모습

$y_t = 0.5+0.5\, y_{t-1} + e_t,\ e_t \sim iid\ N(0,1)$



$y_t = 0.5+0.9\, y_{t-1} + e_t$



$E(y_t) = \mu$ [일정한 평균]

$var(y_t) = \sigma^2$ [일정한 분산]

$cov(y_t, y_{t+s}) = cov(y_t, y_{t-s}) = \gamma_s$ [공분산은 t가 아닌 s에 의존함]

1) 정상 시계열의 특징

- 정상 시계열은 어떤 시점에서 평균과 분산 그리고 특정한 시차의 길이를 갖는 자기공분산을 측정하더라도 동일한 값을 갖는다.
- 정상 시계열은 항상 그 평균값으로 회귀하려는 경향이 있으며, 그 평균값 주변에서의 변동은 대체로 일정한 폭을 갖는다.
- 정상 시계열이 아닌 경우 특정 기간의 시계열 자료로부터 얻은 정보를 다른 시기로 일반화 할 수 없다.

③ 시계열자료 분석방법

가. 분석방법

- 회귀분석(계량경제)방법, Box-Jenkins 방법, 지수평활법, 시계열 분해법 등이 있다.

- 수학적 이론모형 : 회귀분석(계량경제)방법, Box-Jenkins 방법
- 직관적 방법 : 지수평활법, 시계열 분해법으로 시간에 따른 변동이 느린 데이터 분석에 활용
- 장기 예측 : 회귀분석방법 활용
- 단기 예측 : Box-Jenkins 방법, 지수평활법, 시계열 분해법 활용

나. 자료 형태에 따른 분석방법

1) 일변량 시계열분석

- Box-Jenkins(ARMA), 지수 평활법, 시계열 분해법 등이 있다.
- 시간(t)을 설명변수로 한 회귀모형주가, 소매물가지수 등 하나의 변수에 관심을 갖는 경우의 시계열분석

2) 다중 시계열분석

- 계량경제 모형, 전이함수 모형, 개입분석, 상태공간 분석, 다변량 ARIMA 등
- 여러 개의 시간(t)에 따른 변수들을 활용하는 시계열 분석

> 계량경제(econometrics) : 시계열 데이터에 대한 회귀분석(예 : 이자율, 인플레이션이 환율에 미치는 요인)

다. 이동평균법

1) 이동평균법의 개념

- 과거로부터 현재까지의 시계열 자료를 대상으로 일정 기간별 이동평균을 계산하고, 이들의 추세를 파악하여 다음 기간을 예측하는 방법
- 시계열 자료에서 계절변동과 불규칙 변동을 제거하여 추세변동과 순환변동만 가진 시계열로 변환하는 방법으로도 사용됨

$$F_{n+1} = \frac{1}{m}(Z_n + Z_{n-1} + \ldots + Z_{n-m+1}) = \frac{1}{m}\sum_{t}^{n} Z_t,\ t = n - m + 1$$

m은 이동평균한 특정 기간, Z_n은 가장 최근 시점의 데이터

- n개의 시계열 데이터를 m기간으로 이동평균하면 n-m+1개의 이동평균 데이터가 생성된다.

2) 이동평균법의 특징

- 간단하고 쉽게 미래를 예측할 수 있으며, 자료의 수가 많고 안정된 패턴을 보이는 경우 예측의 품질(Quality)이 높음
- 특정 기간 안에 속하는 시계열에 대해서는 동일한 가중치를 부여함
- 일반적으로 시계열 자료에 뚜렷한 추세가 있거나 불규칙변동이 심하지 않은 경우에는 짧은 기간(m의 개수를 적음)의 평균을 사용, 반대로 불규칙변동이 심한 경우 긴 기간(m의 개수가 많음)의 평균을 사용함
- 이동평균법에서 가장 중요한 것은 적절한 기간을 사용하는 것, 즉, 적절한 m의 개수를 결정하는 것임

라. 지수평활법

1) 지수평활법의 개념

- 일정 기간의 평균을 이용하는 이동평균법과 달리 모든 시계열 자료를 사용하여 평균을 구하며, 시간의 흐름에 따라 최근 시계열에 더 많은 가중치를 부여하여 미래를 예측하는 방법

$$
\begin{aligned}
F_{n+1} &= \alpha Z_n + (1-\alpha)F_n \\
&= \alpha Z_n + (1-\alpha)[\alpha Z_{n-1} + (1-\alpha)F_{n-1}] \\
&= \alpha Z_n + \alpha(1-\alpha)Z_{n-1} + (1-\alpha)^2 F_{n-1} \\
&= \alpha Z_n + \alpha(1-\alpha)Z_{n-1} + (1-\alpha)^2[\alpha Z_{n-2} + (1-\alpha)F_{n-2}] \\
&= \alpha Z_n + \alpha(1-\alpha)Z_{n-1} + \alpha(1-\alpha)^2 Z_{n-2} + \alpha(1-\alpha)^3 Z_{n-3} + \ldots
\end{aligned}
$$

- 여기서 F_{n+1}은 n시점 다음의 예측값, α는 지수평활계수, Z_n은 n시점의 관측값이며, 지수평활 계수가 과거로 갈수록 지수형태로 감소하는 형태인 것을 확인할 수 있다.

2) 지수평활법의 특징

- 단기간에 발생하는 불규칙 변동을 평활하는 방법
- 자료의 수가 많고, 안정된 패턴을 보이는 경우일수록 예측 품질이 높음
- 지수평활법에서 가중치의 역할을 하는 것은 지수평활계수(α)이며, 불규칙변동이 큰 시계열의 경우 지수평활계수는 작은 값을, 불규칙변동이 작은 시계열의 경우, 큰 값의 지수평활계수를 적용함(generally, α is between 0.05 and 0.3)
- 지수평활계수는 예측오차(실제 관측치와 예측치 사이의 잔차제곱합)를 비교하여 예측오차가 가장 작은 값을 선택하는 것이 바람직함
- 지수평활계수는 과거로 갈수록 지속적으로 감소함
- 지수평활법은 불규칙변동의 영향을 제거하는 효과가 있으며, 중기 예측 이상에 주로 사용됨 (단, 단순지수 평활법의 경우, 장기추세나 계절변동이 포함된 시계열의 예측에는 적합하지 않음)

시계열의 모형은 자기회귀 모형(AR 모형), 이동평균 모형(MA 모형), 자기회귀누적이동평균 모형(ARIMA 모형)이 있습니다. 시계열 모형과 더불어 분해 시계열의 내용에 대해 정확히 숙지하셔야 합니다.

④ 시계열모형

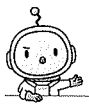
가. 자기회귀 모형(AR 모형, Autoregressive Model)

- p 시점 전의 자료가 현재 자료에 영향을 주는 모형이다.

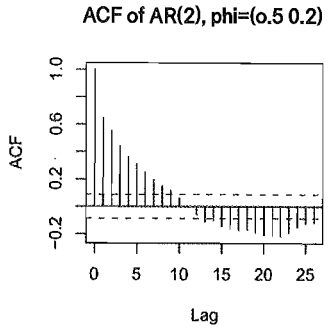
$$Z_t = \Phi_1 Z_{t-1} + \Phi_2 Z_{t-2} + \cdots + \Phi_p Z_{t-p} + \alpha_t$$

참고

- Z_t : 현재 시점의 시계열 자료
- $Z_{t-1}, Z_{t-2}, \cdots, Z_p$: 이전, 그 이전 시점 p의 시계열 자료
- Φ_p : p 시점이 현재에 어느 정도 영향을 주는지를 나타내는 모수
- α_t : 백색잡음과정(White Noise Process), 시계열 분석에서 오차항을 의미한다.
- 평균이 0, 분산이 σ^2, 자기공분산이 0인 경우를 뜻하며, 시계열간 확률적 독립인 경우 강(Strictly) 백색잡음 과정이라고 한다. 백색잡음 과정이 정규분포를 따를 경우 이를 가우시안(Gaussian)백색잡음과정이라고 한다.



- AR(1) 모형 : $Z_t = \Phi_1 Z_{t-1} + \alpha_t$, 직전 시점 데이터로만 분석
- AR(2) 모형 : $Z_t = \Phi_1 Z_{t-1} + \Phi_2 Z_{t-2} + \alpha_t$, 연속된 2시점 정도의 데이터로 분석
- 자기상관함수(ACF)는 빠르게 감소, 부분자기함수(PACF)는 어느 시점에서 절단점을 가진다. (ACF가 빠르게 감소하고, PACF가 3시점에서 절단점을 갖는 그래프가 있다면, 2시점 전의 자료까지가 현재에 영향을 미치는 AR(2) 모형이라 볼 수 있다.)
- AR(2) 모형의 자기상관함수(ACF)와 편자기상관함수(PACF)



참고

자기상관계수와 부분자기상관계수

자기상관계수

- k 기간 떨어진 값들log(k)의 상관계수

$$p_k = \frac{\lambda_k}{\lambda_0} = \frac{Cov(Y_t, Y_{t-k})}{\sqrt{Var(Y_t)Var(Y_{t-k})}}$$

where $\lambda = Cov(Y_t, Y_{t-k})$: 자기공분산(Autocavariance)

- 자기상관계수 함수 : 상관계수 k를 함수 형태로 표시한 것

AR (1) : $Z_t = \phi_1 Z_{t-1} + \alpha_t$, 만일 $-1 < \phi_1 < 1$이면 두 지점간의 거리가 멀어질수록 (k가 커질수록) ACF는 0에 수렴하게 된다.

부분(편)자기상관계수

- 서로 다른 두 시점 사이의 관계를 분석할 때 중간에 있는 값들의 영향을 제외시킨 상관관계 개념

$Corr(Y_t, Y_{t-k} | Y_{t-1}, \cdots, Y_{t-k+1})$

나. 이동평균 모형(MA 모형, Moving Average Model)

- 유한한 개수의 백색잡음의 결합이므로 언제나 정상성을 만족
- 1차 이동평균모형(MA(1) 모형)은 이동평균모형 중에서 가장 간단한 모형으로 시계열이 같은 시점의 백색잡음과 바로 전 시점의 백색잡음의 결합으로 이뤄진 모형

$Z_t = \alpha_t - \theta_1 \alpha_{t-1} - \theta_2 \alpha_{t-2} - \cdots - \theta_p \alpha_{t-p}$

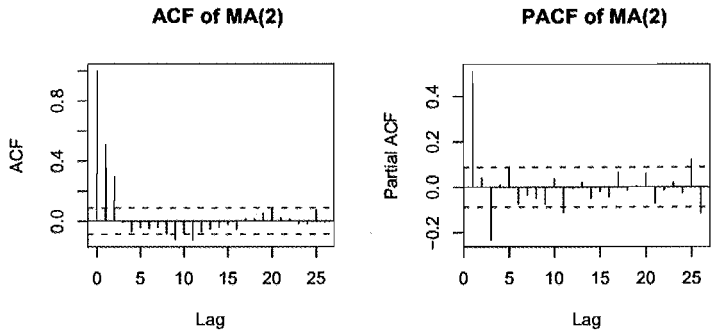
- 2차 이동평균모형(MA(2) 모형)은 바로 전 시점의 백색잡음과 시차가 2인 백색잡음의 결합으로 이뤄진 모형

$Z_t = \alpha_t - \theta_1 \alpha_{t-1}$

- AR 모형과 반대로 ACF에서 절단점을 갖고, PACF가 빠르게 감소

$Z_t = \alpha_t - \theta_1 \alpha_{t-1} - \theta_2 \alpha_{t-2}$

- MA(2) 모형의 자기상관함수(ACF)와 편자기상관함수(PACF)



자기회귀이동평균 (Autoregressive and Moving Average, ARMA) 모형

자기회귀와 이동평균 다항식으로 약한 정상성을 가진 확률적 시계열을 표현하는 데 사용되며 ARMA(p,q)로 표기함(여기서, p는 자기회귀 다항식의 차수, q는 이동평균 다항식의 차수)

다.자기회귀누적이동평균 모형 (ARIMA(p,d,q) 모형, Autoregressive Integrated Moving Average Model)

- ARIMA 모형은 비정상시계열 모형이다.
- ARIMA 모형을 차분이나 변환을 통해 AR모형이나 MA모형, 이 둘을 합친 ARMA 모형으로 정상화 할 수 있다.
- p는 AR 모형, q는 MA모형과 관련이 있는 차수이다.
- 시계열 $\{Z_t\}$의 d번 차분한 시계열이 ARMA(p,q) 모형이면, 시계열 $\{Z_t\}$는 차수가 p,d,q인 ARIMA 모형, 즉 ARIMA(p,d,q) 모형을 갖는다고 한다.
- d=0 이면 ARMA(p,q) 모형이라 부르고, 이모형은 정상성을 만족한다.(ARMA(0,0)일 경우 정상화가 불필요하다)
- p=0 이면 IMA(d,q) 모형이라고 부르고, d번 차분하면 MA(q) 모형을 따른다.
- q=0 이면 ARI(p,d) 모형이라 부르며, d번 차분한 시계열이 AR(p) 모형을 따른다.

예시

- ARIMA(0, 1, 1)의 경우에는 1 차분 후 MA(1) 활용
- ARIMA(1, 1, 0)의 경우에는 1 차분 후 AR(1) 활용
- ARIMA(1, 1, 2)의 경우에는 1 차분 후 AR(1), MA(2), ARMA(1, 2) 선택 활용
  → 이런 경우 가장 간단한 모형을 선택하거나 AIC를 적용하여 점수가 가장 낮은 모형을 선정한다.

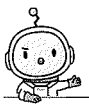
라. 분해 시계열

- 시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법을 말하며 회귀분석적인 방법을 주로 사용한다.
- 분해식의 일반적 정의

$$Z_t = f(T_t, S_t, C_t, I_t)$$

참고

- T_t : 경향(추세)요인 : 자료가 오르거나 내리는 추세, 선형, 이차식 형태, 지수적 형태 등
- S_t : 계절요인 : 요일, 월, 사계절 각 분기에 의한 변화 등 고정된 주기에 따라 자료가 변하는 경우
- C_t : 순환요인 : 경제적이나 자연적인 이유 없이 알려지지 않은 주기를 가지고 변화하는 자료
- I_t : 불규칙요인 : 위의 세 가지 요인으로 설명할 수 없는 오차에 해당하는 요인



마. R을 이용한 시계열분석

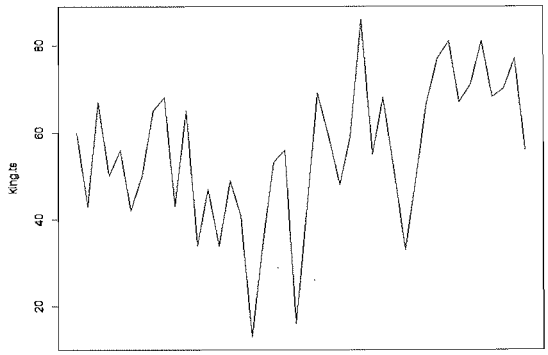
- 영국 왕들의 사망 시 나이 데이터를 이용한 시계열분석

> - 영국 왕 42명의 사망 시 나이 예제는 비계절성을 띄는 시계열 자료
> - 비계절성을 띄는 시계열 자료는 트렌드 요소, 불규칙 요소로 구성
> - 20번째 왕까지는 38세에서 55까지 수명을 유지하고, 그 이후부터는 수명이 늘어서 40번째 왕은 73세까지 생존

1) 분해 시계열

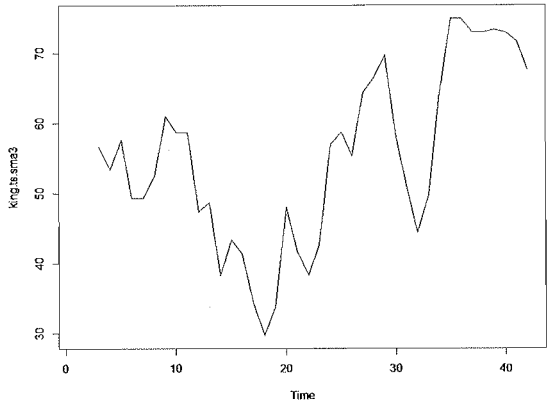
가) 자료 읽기 및 그래프 그리기

```
> library(tseries)
> library(forecast)
> library(TTR)
> king <- scan("http://robjhyndman.com/tsdldata/misc/kings.dat",skip=3)
> king.ts <- ts(king)
> plot.ts(king.ts)
```



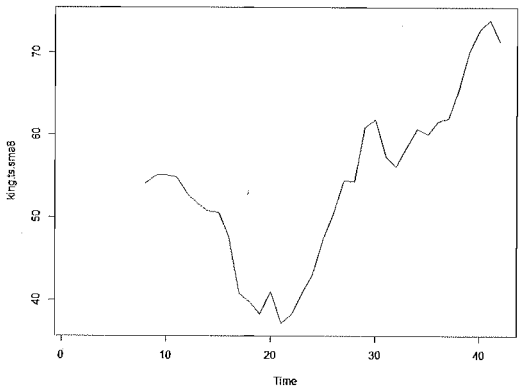
나) 3년마다 평균을 내서 그래프를 부드럽게 표현

```
> king.sma3 <- SMA(king.ts, n=3)
> plot.ts(king.sma3)
```



다) 8년마다 평균을 내서 그래프를 부드럽게 표현

```
> king.sma8 <- SMA(king.ts, n=8)
> plot.ts(king.sma8)
```

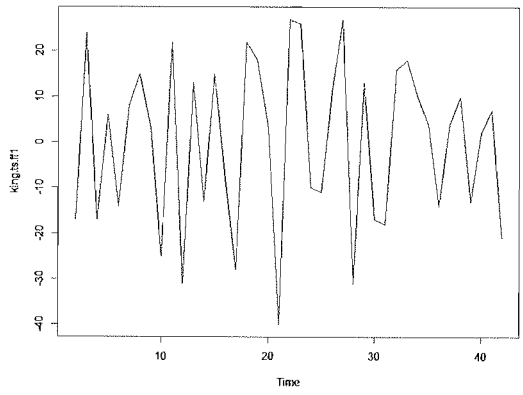


2) ARIMA 모델

가) 개요

- ARIMA 모델은 정상성 시계열에 한해 사용한다.
- 비정상 시계열 자료는 차분해 정상성으로 만족하는 조건의 시계열로 바꿔준다.
- 이전 그래프에서 평균이 시간에 따라 일정치 않은 모습을 보이므로 비정상시계열이다. 따라서 차분을 진행한다.
- 1차 차분 결과에서 평균과 분산이 시간에 따라 의존하지 않음을 확인한다.
- ARIMA(p,1,q)모델이며 차분을 1번 해야 정상성을 만족한다.

```
> king.ff1 <- diff(king.ts, differences=1)
> plot.ts(king.ff1)
```

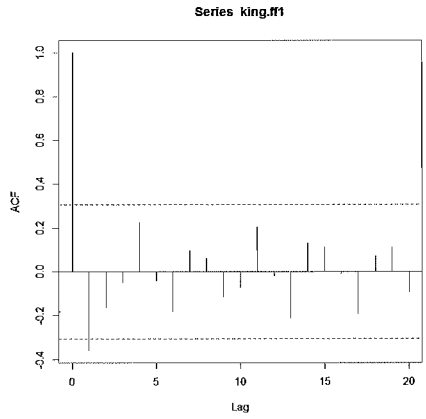


나) ACF와 PACF를 통한 적합한 ARIMA 모델 결정

① ACF

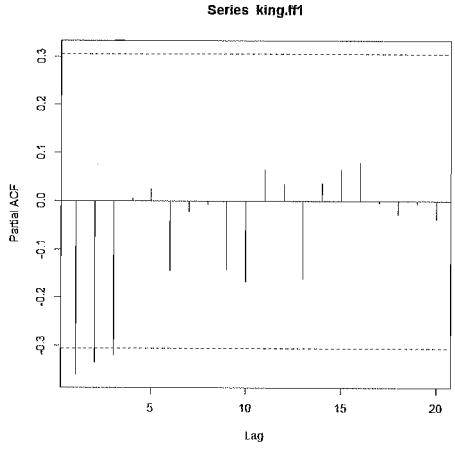
- lag는 0부터 값을 갖는데, 너무 많은 구간을 설정하면 그래프를 보고 판단하기 어렵다.
- ACF값이 lag 1인 지점 빼고는 모두 점선 구간 안에 있다.

```
> acf(king.ff1, lag.max=20)
> acf(king.ff1, lag.max=20, plot=FALSE)

Autocorrelations of series 'king.ff1', by lag

    0     1     2     3     4     5     6     7     8     9    10
1.000 -0.360 -0.162 -0.050  0.227 -0.042 -0.181  0.095  0.064 -0.116 -0.071
   11    12    13    14    15    16    17    18    19    20
0.206 -0.017 -0.212  0.130  0.114 -0.009 -0.192  0.072  0.113 -0.093
```



② PACF - PACF 값이 lag 1, 2, 3에서 점선 구간을 초과하고 음의 값을 가지며 절단점은 lag 4이다.

```
> pacf(king.ff1, lag.max=20)
> pacf(king.ff1, lag.max=20, plot=FALSE)


Partial autocorrelations of series 'king.ff1', by lag


    1     2     3     4     5     6     7     8     9    10
-0.360 -0.335 -0.321  0.005  0.025 -0.144 -0.022 -0.007 -0.143 -0.167
   11    12    13    14    15    16    17    18    19    20
 0.065  0.034 -0.161  0.036  0.066  0.081 -0.005 -0.027 -0.006 -0.037
```



다) 종합

- ARMA 후보들이 생성

  ① ARMA(3,0) 모델 : PACF 값이 lag4에서 절단점을 가짐. AR(3)모형

  ② ARMA(0,1) 모델 : ACF 값이 lag2에서 절단점을 가짐. MA(1)모형

  ③ ARMA(p,q) 모델 : 그래서 AR모형과 MA모형을 혼합

라) 적절한 ARIMA모형 찾기

- forecast 패키지에 내장된 auto.arima() 함수 이용
- 영국 왕의 사망 나이 데이터의 적절한 ARIMA모형은 ARIMA(0,1,1)이다.

```
> auto.arima(king)
Series: king
ARIMA(0,1,1)

Coefficients:
          ma1
      -0.7218
s.e.   0.1208

sigma^2 estimated as 236.2:  log likelihood=-170.06
AIC=344.13   AICc=344.44   BIC=347.56
```

마) 예측

- 42명의 영국왕 중에서 마지막 왕의 사망시 나이는 56세
- 43번째에서 52번째 왕까지 10명의 왕의 사망시 나이를 예측한 결과 67.75살로 추정된다.

- 5명 정도만 예측하고 싶다면, 옵션에 h=5를 입력한다.
- 신뢰 구간은 80%~95% 사이

```
> king.arima<- arima(king, order=c(0, 1, 1))
> king.forecasts <- forecast(king.arima)
> king.forecasts
```

Point	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
43	67.75063	48.29647	87.20479	37.99806	97.50319
44	67.75063	47.55748	87.94377	36.86788	98.63338
45	67.75063	46.84460	88.65665	35.77762	99.72363
46	67.75063	46.15524	89.34601	34.72333	100.77792
47	67.75063	45.48722	90.01404	33.70168	101.79958
48	67.75063	44.83866	90.66260	32.70979	102.79146
49	67.75063	44.20796	91.29330	31.74523	103.75603
50	67.75063	43.59372	91.90753	30.80583	104.69543
51	67.75063	42.99472	92.50653	29.88974	105.61152
52	67.75063	42.40988	93.09138	28.99529	106.50596

MEMO

5절 다차원척도법

○ 학습 목표

- 다차원 척도법(MDS)를 이해한다.
- 주성분분석(PCA)을 이해한다.
- R 프로그램을 통해 다차원 척도법과 주성분분석을 진행할 수 있다.

○ 눈높이 체크

✓ 다차원 척도법을 들어보셨나요?

다차원 척도법(Multi Dimensional Scaling, MDS)은 군집분석과 같이 개체들을 대상으로 변수들을 측정한 후, 개체들 사이의 유사성/비유사성을 측정하여 개체들을 2차원 또는 3차원 공간상에 점으로 표현하는 분석 방법입니다.

✓ 군집분석을 들어보셨나요?

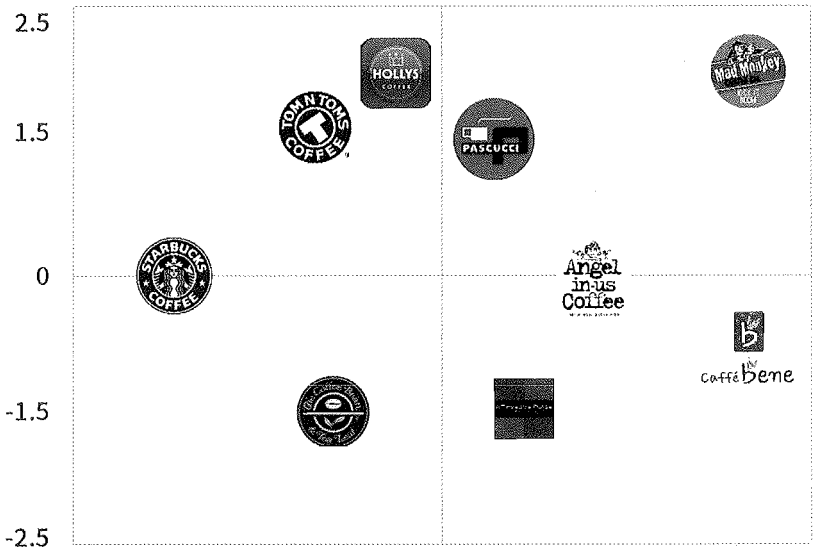
군집분석은 개체들 간의 비유사성을 이용하여 동일한 그룹들로 분류하는 것이 목적인 반면, 다차원척도법은 개체들의 비유사성을 이용하여 2차원 공간상에 점으로 표시하고 개체들 사이의 집단화를 시각적으로 표현하는 것을 목적으로 합니다.

✓ 주성분분석을 들어보셨나요?

주성분분석(Principal Component Analysis, PCA)은 상관관계가 있는 변수들의 선형결합을 통해 변수를 축약하는 기법입니다. 넓은 의미에서는 요인분석(Factor Analysis)의 한 종류로 활용되기도 합니다.

① 다차원척도법 (Multidimensional Scaling)

- 객체간 근접성(Proximity)을 시각화하는 통계기법이다.
- 군집분석과 같이 개체들을 대상으로 변수들을 측정한 후에 개체들 사이의 유사성/비유사성을 측정하여 개체들을 2차원 공간상에 점으로 표현하는 분석방법이다.
- 개체들을 2차원 또는 3차원 공간상에 점으로 표현하여 개체들 사이의 집단화를 시각적으로 표현하는 분석방법이다.



② 다차원척도법 목적

- 데이터 속에 잠재해 있는 패턴(Pattern), 구조를 찾아낸다.
- 그 구조를 소수 차원의 공간에 기하학적으로 표현한다.
- 데이터 축소(Data Reduction)의 목적으로 다차원척도법을 이용한다. 즉, 데이터에 포함되는 정보를 끄집어내기 위해서 다차원척도법을 탐색수단으로써 사용한다.
- 다차원척도법에 의해서 얻은 결과를, 데이터가 만들어진 현상이나 과정에 고유의 구조로서 의미를 부여한다.

③ 다차원척도법 방법

- 개체들의 거리 계산에는 **유클리드 거리행렬을 활용**한다.

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + \ldots + (x_{iR} - x_{jR})^2}$$

- 관측대상들의 상대적 거리의 정확도를 높이기 위해 적합 정도를 스트레스 값(Stress Value)으로 나타낸다.
- 각 개체들을 공간상에 표현하기 위한 방법은 부적합도 기준으로 Stress나 S-Stress를 사용한다.
- 최적모형의 적합은 부적합도를 최소로 하는 반복 알고리즘을 이용하며, 이 값이 일정 수준 이하가 될 때 최종적으로 적합된 모형으로 제시한다.
- 스트레스 값은 $S = \sqrt{\frac{\sum_{i=1,j=1}^{n} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i=1,j=1}^{n} (d_{ij})^2}}$

(d_{ij}=관측대상i부터 j까지 실제거리, $\hat{d}_{ij}$ = 프로그램에 의해 추정된 거리)

- Stress와 적합도 수준 M은 개체들을 공간상에 표현하기 위한 방법으로 Stress나 S-Stress를 부적합도 기준으로 사용한다.
- 최적모형의 적합은 부적합도를 최소로 하는 방법으로 일정 수준 이하로 될 때까지 반복해서 수행한다.

Stress	적합도 수준
0	완벽(Perfect)
0.05 이내	매우 좋은(Excellent)
0.05~0.10	만족(Satisfactory)
0.10~0.15	보통(Acceptable, but Doubt)
0.15 이상	나쁨(Poor)

④ 다차원척도법 종류

가. 계량적 MDS(Metric MDS)

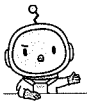
- 데이터가 **구간척도나 비율척도인 경우 활용**한다.(전통적인 다차원척도법) N개의 케이스에 대해서 p개의 특성변수가 있는 경우, 각 개체들간의 유클리드 거리행렬을 계산하고 개체들간의 비유사성 S(거리제곱 행렬의 선형함수)를 공간상에 표현한다.

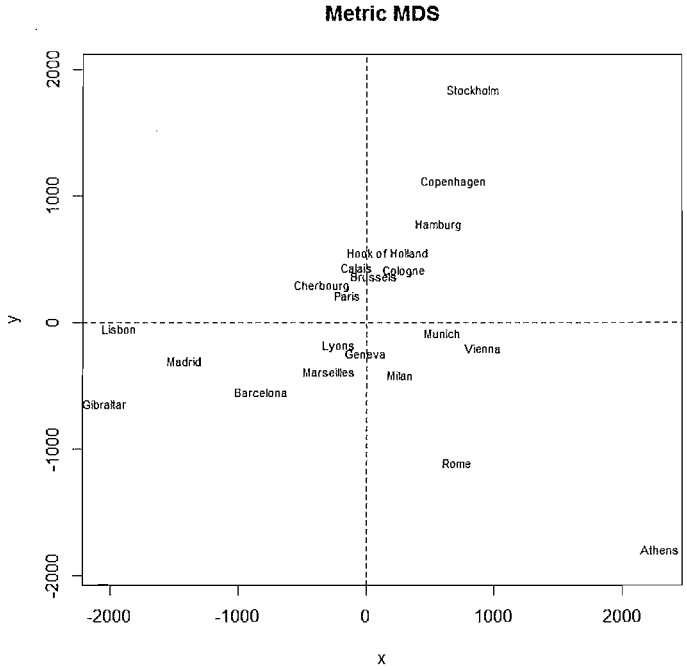
사례

cmdscale 사례

- MASS package의 eurodist 자료를 이용한다.
- 유럽의 21개 도시들 사이의 거리를 측정한다.
- cmdscale을 이용하여 2차원으로 21개 도시들을 매핑한다.
- 종축은 북쪽 도시를 상단에 표시하기 위해 부호를 바꾼다.

```
> library(MASS)
> loc <- cmdscale(eurodist)
> x <- loc[, 1]
> y <- -loc[, 2]
> plot(x, y, type="n", asp=1, main="Metric MDS")
> text(x, y, rownames(loc), cex=0.7)
> abline(v=0, h=0, lty=2, lwd=0.5)
```





나. 비계량적 MDS(nonmetric MDS)

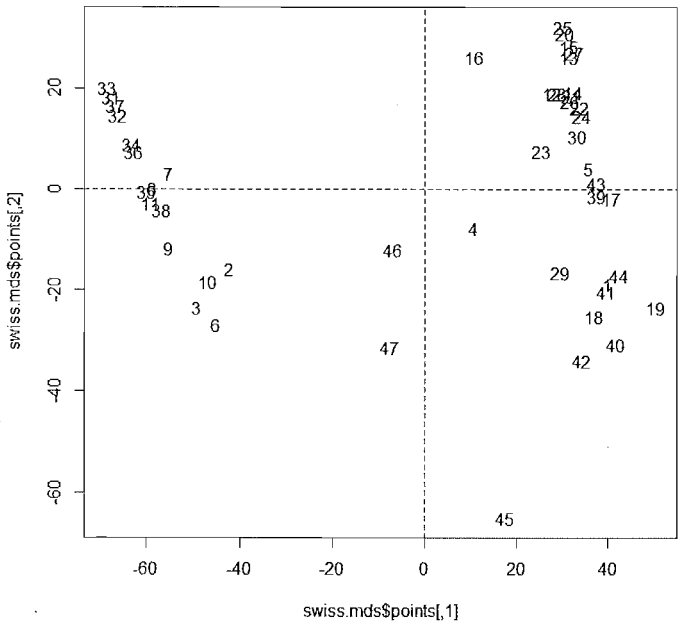
- 데이터가 **순서척도**인 경우 활용한다. 개체들 간의 거리가 순서로 주어진 경우에는 순서척도를 거리의 속성과 같도록 변환(Monotone Transformation)하여 거리를 생성한 후 적용한다.

사례

isoMDS 사례

- MASS package의 Swiss 자료를 이용하여 2차원으로 도시들을 매핑한다.
- 1888년경의 스위스연방 중 47개의 불어권 주의 토양의 비옥도 지수와 여러 사회경제적 지표를 측정한 자료이다.

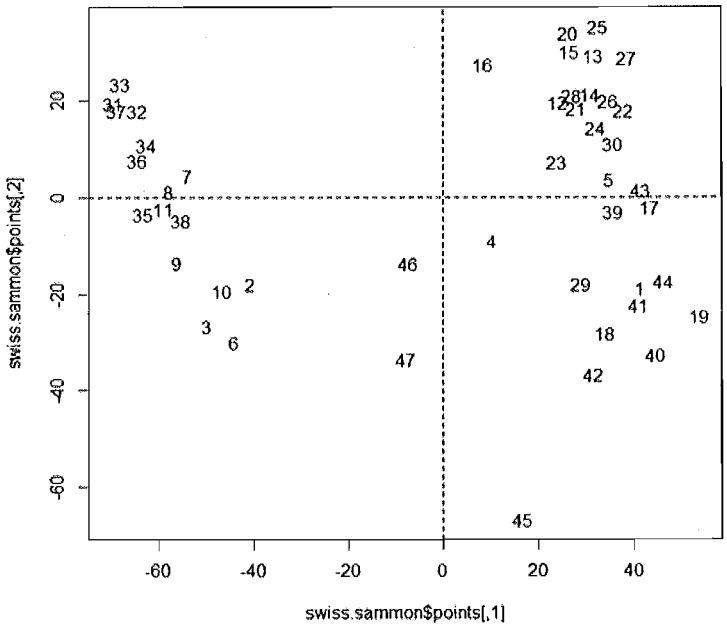
```
> library(MASS)
> data(swiss)
> swiss.x <- as.matrix(swiss[, -1])
> swiss.dist <- dist(swiss.x)
> swiss.mds <- isoMDS(swiss.dist)
> plot(swiss.mds$points, type="n")
> text(swiss.mds$points, labels=as.character(1:nrow(swiss.x)))
> abline(v=0, h=0, lty=2, lwd=0.5)
```



사례

sammon 사례

```
> swiss.x <- as.matrix(swiss[, -1])
> swiss.sammon <- sammon(dist(swiss.x))
> plot(swiss.sammon$points, type="n")
> text(swiss.sammon$points, labels=as.character(1:nrow(swiss.x)))
> abline(v=0, h=0, lty=2, lwd=0.5)
```



6절 주성분 분석

① 주성분분석 (Principal Component Analysis)

- 여러 변수들의 변량을 '주성분(Principal Component)'이라는 **서로 상관성이 높은 변수들의 선형 결합**으로 만들어 기존의 상관성이 높은 변수들을 요약, 축소하는 기법이다.
- 첫 번째 주성분으로 전체 변동을 가장 많이 설명할 수 있도록 하고, 두 번째 주성분으로는 첫 번째 주성분과는 상관성이 없어서(낮아서) 첫 번째 주성분이 설명하지 못하는 나머지 변동을 정보의 손실 없이 가장 많이 설명할 수 있도록 변수들의 선형조합을 만든다.

② 주성분분석의 목적

- 여러 변수들 간에 내재하는 상관관계, 연관성을 이용해 **소수의 주성분으로 차원을 축소**함으로써 데이터를 이해하기 쉽고 관리하기 쉽게 해준다.
- 다중공선성이 존재하는 경우, **상관성이 없는(적은) 주성분으로 변수들을 축소**하여 모형 개발에 활용된다. **회귀분석** 등의 모형 개발 시 입력변수들간의 상관 관계가 높은 **다중공선성(Multicollinearity)**이 존재할 경우 모형이 잘못 만들어져 문제가 생김
- 연관성이 높은 변수를 주성분분석을 통해 차원을 축소한 후에 **군집분석을 수행하면 군집화 결과와 연산속도를 개선**할 수 있다.
- 기계에서 나오는 다수의 센서데이터를 주성분분석으로 차원을 축소한 후에 시계열로 분포나 추세의 변화를 분석하면 기계의 고장(Fatal Failure) 징후를 사전에 파악하는데 활용하기도 한다.

③ 주성분분석 vs 요인분석

가. 요인분석(Factor Analysis)

- 등간척도(혹은 비율척도)로 측정한 두 개 이상의 변수들에 잠재되어 있는 공통인자를 찾아내는 기법이다.

나. 공통점

- 모두 데이터를 축소하는데 활용된다. 원래 데이터를 활용해서 몇 개의 새로운 변수들을 만들 수 있다.

다. 차이점

1) 생성된 변수의 수

- 요인분석은 몇 개라고 지정 없이(2 or 3, 4, 5 ….) 만들 수 있다.
- 주성분분석은 제1주성분, 제2주성분, 제3주성분 정도로 활용한다.(대개 4개 이상은 넘지 않음)

2) 생성된 변수의 이름

- **요인분석**은 분석자가 **요인의 이름을 명명**한다.
- **주성분분석**은 주로 **제1주성분, 제2주성분** 등으로 표현된다.

3) 생성된 변수들 간의 관계

- 요인분석은 새 변수들은 기본적으로 대등한 관계를 갖고 '어떤 것이 더 중요하다'라는 의미는 요인분석에서는 없다. 단, 분류/예측에 그다음 단계로 사용된다면 그때 중요성의 의미가 부여된다.
- 주성분분석은 제1주성분이 가장 중요하고, 그다음 제2주성분이 중요하게 취급된다.

4) 분석 방법의 의미

- 요인분석은 목표변수를 고려하지 않고 그냥 데이터가 주어지면 변수들을 비슷한 성격들로 묶어서 새로운 잠재변수들을 만든다.
- 주성분분석은 목표 변수를 고려하여 목표 변수를 잘 예측/분류하기 위하여 원래 변수들의 선형 결합으로 이루어진 몇 개의 주성분(변수)들을 찾아내게 된다.

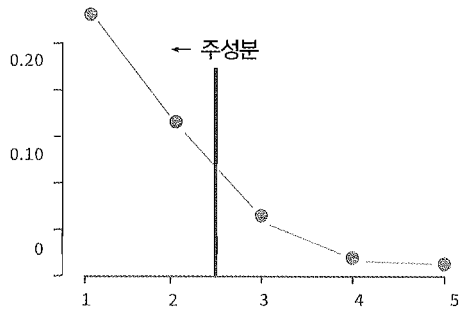
주성분 갯수의 선택방법은 누적기여율(Cumulative Proportion)이나 Scree Plot을 활용하므로 해석 방법을 정확히 숙지하시기 바랍니다.

④ 주성분의 선택법

- 주성분분석의 결과에서 누적기여율(Cumulative Proportion)이 85%이상이면 주성분의 수로 결정할 수 있다.

Importance of components :

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.6618	1.2671	0.7420	0.25311	0.13512
Proportion of Variance	0.5523	0.3211	0.1101	0.01281	0.00365
Cumulative	0.5523	0.8734	0.9835	0.99635	1.00000



Scree Plot

- Scree Plot을 활용하여 고윳값(Eigenvalue)이 수평을 유지하기 전단계로 주성분의 수를 선택한다.

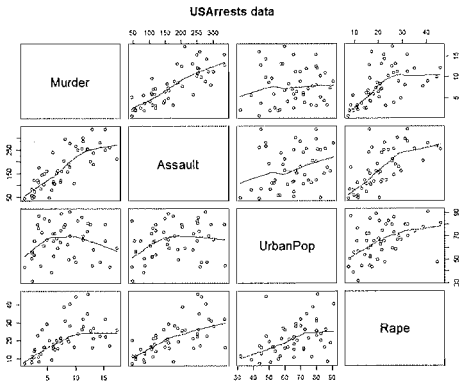
⑤ 주성분 분석 사례

가. USArrests 자료

- 1973년 미국 50개주의 100,000명의 인구 당 체포된 세 가지 강력범죄수(Assault, Murder, Rape)와 각 주마다 도시에 거주하는 인구의 비율(%)로 구성되어 있다.
- 변수들 간의 척도의 차이가 상당히 크기 때문에 상관행렬을 사용하여 분석한다.
- 특이치 분해를 사용하는 경우 자료 행렬의 각 변수의 평균과 제곱의 합이 1로 표준화되었다고 가정할 수 있다.

1) 4개의 변수들 간의 산점도

```
> library(datasets)
> data(USArrests)
> pairs(USArrests, panel = panel.smooth, main = "USArrests
data")
```



- Murder와 UrbanPop비율간의 관련성이 작아 보인다.

2) summary

- 제1주성분과 제2주성분까지의 누적 분산비율은 대략 86.8%로 2개의 주성분 변수를 활용하여 전체 데이터의 86.8%를 설명할 수 있다.
- 주성분들에 의해 설명되는 변동의 비율은 Scree Plot 을 통해 확인 가능하다.

```
> US.prin <- princomp(USArrests, cor = TRUE)
> summary(US.prin)
> screeplot(US.prin, npcs=4, type="lines")
```

Importance of components :

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	1.5748783	0.9948694	0.5971291	0.41644938
Proportion of Variance	0.6200604	0.2474413	0.0891408	0.04335752
Cumulative Proportion	0.6200604	0.8675017	0.9566425	1.00000000

3) Loading

- 네 개의 변수가 각 주성분 Comp.1-Comp.4까지 기여하는 가중치가 제시된다.
- 제1주성분에는 네 개의 변수가 평균적으로 기여한다.
- 제2주성분에서는 (Murder, Assault)와 (UrbanPop, Rape)의 계수의 부호가 서로 다르다.

```
> loadings(US.prin)

Loadings :
           Comp.1 Comp.2 Comp.3 Comp.4
Murder      0.536  0.418  0.341  0.649
Assault     0.583  0.188  0.268 -0.743
UrbanPop    0.278 -0.873  0.378  0.134
Rape        0.543 -0.167 -0.818

               Comp.1 Comp.2 Comp.3 Comp.4
SS loadings      1.00   1.00   1.00   1.00
Proportion Var   0.25   0.25   0.25   0.25
Cumulative Va    0.25   0.50   0.75   1.00
```

4) Scores

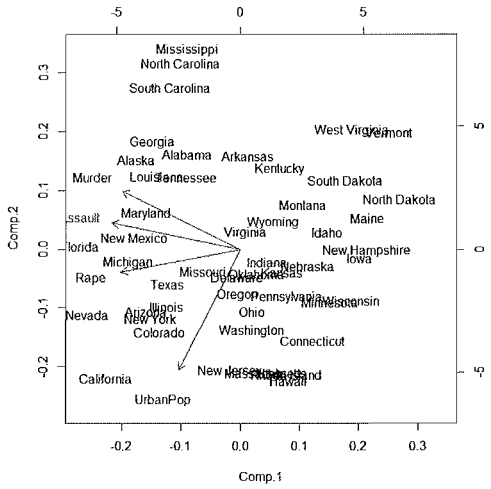
- 각 주성분 Comp.1-Comp.4의 선형식을 통해 각 지역(record)별로 얻은 결과를 계산한다.

```
> US.prin$scores

               Comp.1      Comp.2      Comp.3      Comp.4
Alabama     0.98556588  1.13339238  0.44426879  0.156267145
Alaska      1.95013775  1.07321326 -2.04000333 -0.438583440
Arizona     1.76316354 -0.74595678 -0.05478082 -0.834652924
Arkansas   -0.14142029  1.11979678 -0.11457369 -0.182810896
California  2.52398013 -1.54293399 -0.59855680 -0.341996478
Colorado    1.51456286 -0.98755509 -1.09500699  0.001464887
Connecticut -1.35864746 -1.08892789  0.64325757 -0.118469414
Delaware    0.04770931 -0.32535892  0.71863294 -0.881977637
Florida     3.01304227  0.03922851  0.57682949 -0.096284752
Georgia     1.63928304  1.27894240  0.34246008  1.076796812
Hawaii     -0.91265715 -1.57046001 -0.05078189  0.902806864
Idaho      -1.63979985  0.21097292 -0.25980134 -0.499104101
Illinois    1.37891072 -0.68184119  0.67749564 -0.122021292
Indiana    -0.50546136 -0.15156254 -0.22805484  0.424665700
Iowa       -2.25364607 -0.10405407 -0.16456432  0.017555916

중략
```

5) 제 1-2주성분에 의한 행렬도

- 조지아, 메릴랜드, 뉴 멕시코 등은 폭행과 살인의 비율이 상대적으로 높은 지역이다.
- 미시간, 텍사스 등은 강간의 비율이 높은 지역이다.
- 콜로라도, 캘리포니아, 뉴저지 등은 도시에 거주하는 인구의 비율이 높은 지역이다.
- 아이다호, 뉴 햄프셔, 아이오와 등의 도시들은 도시에 거주하는 인구의 비율이 상대적으로 낮으면서 3대 강력범죄도 낮다.

```
arrests.pca <- prcomp(USArrests, center = TRUE, scale. = TRUE)
biplot(arrests.pca, scale=0)
```



심화

통계 만점 지름길

확률변수

가. 이산형 확률변수

- 이산점(Discrete Points)에서 0이 아닌 확률값을 가지는 확률변수이다.
- **이산형 확률변수의 확률**은 $P(X = x_i) = P_i, i = 1, 2, ..., n$으로 표현한다.
- 각 이산점에 있어서 확률의 크기를 표현하는 함수를 확률질량함수(Probability Mass Function, PMF)라 한다.
- 예를 들어, 두 개의 주사위를 던지는 실험에서 확률변수 X를 'X≡두 주사위 눈금의 합'이라고 정의하면 X의 이산형 확률분포는 다음과 같다.

X	2	3	4	5	6	7	8	9	10	11	12	합
확률	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36	1

- 이산형 확률변수의 확률조건은 다음과 같다.

① $0 \le P_i \le 1, (i = 1, 2, \cdots, n)$즉, 각 x_i가 나타날 확률은 0과 1 사이의 값을 갖는다.

② $\sum_{i=1}^{n} P_i = 1$즉, 모든 가능한 경우의 확률의 합은 1이다.

나. 연속형 확률변수

- 특정 실수 구간에서 0이 아닌 확률을 갖는 확률변수이다.
- 연속형 확률변수는 특정한 실수구간 내에서 0이 아닌 확률을 가지므로 이 구간에 대한 확률은 함수의 형태로 표현한다.
- 연속형 확률변수 X의 확률함수를 f(x)라고 할 때, f(x)는 확률밀도함수(Probability Density Function, PDF)라고 부르며 다음 조건을 만족한다.

① 모든 X 값에 대하여 $f(x) \ge 0$이다. 즉, X의 모든 실수 값에 대하여 확률밀도함수는 0 이상이다.

② X의 모든 가능한 값의 확률은 적분 $\int_{-\infty}^{\infty} f(x)dx$로 구하며 이 값은 항상 1이다.

③ 구간(a, b)의 확률은 $\Pr[a < X < b] = \int_a^b f(x)dx$이다. 즉 구간 (a,b)에 대한 X의 확률은 그 구간에 있어 확률밀도함수 f(x)로 만들어지는 면적의 크기이다.

- 예를 들어, 확률변수 X가 0와 1 사이에서 균등한 분포를 가진다면 X의 확률밀도함수는 다음과 같이 표현한다.

$$f(x)=\begin{cases}1, 0\le x\le 1\\ 0, \text{otherwise}\end{cases}$$

여기에서 모든 실수값 X에 대하여 f(x)=0 또는 1이므로 f(x)≥0의 조건을 만족하며, 아래와 같이 확률밀도함수의 조건을 만족한다.

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

다. 누적 분포 함수(Cumulative Distribution Function, CDF)

- 누적 분포 함수는 특정 값 a에 대하여 확률변수 X가 X≤a인 모든 경우의 확률의 합으로 다음과 같이 표현한다.

$$F_X(a)=\Pr(X\le a)$$

- 이산형 확률변수는 $F_X(a)=\sum_{\substack{a\text{이하인}\\ \text{모든 } x_i}} P(X=x_i)$이고 연속형 확률변수는 다음과 같이 표현한다.

$$F_X(a)=\int_{-\infty}^{a} f(x)dx$$

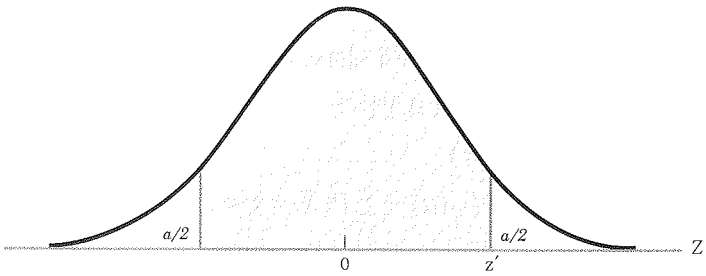
- 또한 누적분포함수는 증가 함수이고 우측 연속 함수이며, 0과 1사이의 값을 가진다.
- 확률분포와의 관계 : 확률변수의 누적분포 함수는 그 확률 분포를 유일하게 결정한다. 구간 (a,b]에 대한 X의 확률은 $\Pr(a<x\le b)=F_X(b)-F_X(a)$이다.

② 구간추정

가. 모평균과 모분산에 따른 구간추정

구분		모평균	모분산
모수		μ	σ^2
점추정량		$\bar{X}$	S^2
표본분포		$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$
(1-α)×100% 신뢰구간	σ^2을 알고있음	$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{1-(\alpha/2)}}\right)$
	σ^2을 모르고 n>30인 경우	$\bar{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$	
	σ^2을 모르고 n≤30인 경우	$\bar{X} \pm t_{\alpha/2}\frac{S}{\sqrt{n}}$	

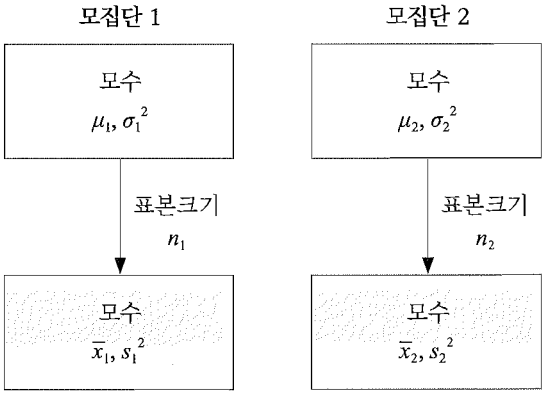
- $z_{\alpha/2}$는 $N(0,1)$에서 $\Pr[Z > z^*] = \alpha/2$를 만족하는 z^*값(임계값)으로 예를 들어, $\alpha = 0.05$(95%신뢰구간)이면, $z^* = 1.96$이다.



- $t_{\alpha/2}$는 자유도가 n-1인 t-분포에서 $\Pr[T > t^*] = \alpha/2$를 만족하는 t^*값(임계값)으로 예를 들어, $\alpha = 0.05$(95%신뢰구간)이면, $t^* = 2.093$이다(n=20).
- $\chi^2_{\alpha/2}$는 자유도가 n-1인 카이제곱분포에서 $\Pr[\chi^2 > \chi^{2*}] = \alpha/2$를 만족하는 χ^{2*}값(임계값), $\chi^2_{(1-\alpha/2)}$는 $\Pr[\chi^2 < \chi^{2*}] = \alpha/2$를 만족하는 값(임계값)을 의미한다.

나. 두 모평균 차이의 신뢰구간 추정(독립표본)

- 두 집단이 서로 독립이라는 전제조건 하에 두 모평균 차이에 대한 추정



모수	점 추정량		(1-α)×100% 신뢰구간
μ1- μ2	$\bar{X}_1 - \bar{X}_2$	σ_1^2, σ_2^2 알고있음	$(\overline{X_1} - \overline{X_2}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
		σ_1^2, σ_2^2 모르지만 $n_1 \geq 30, n_2 \geq 30$	$(\overline{X_1} - \overline{X_2}) \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$
		σ_1^2, σ_2^2 모르고 n_1, n_2 중 하나가 30미만	$(\overline{X_1} - \overline{X_2}) \pm t_{\alpha/2,(n_1+n_2-2)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

- 여기서, $z_{\alpha/2}$는 $N(0,1)$에서 $\Pr[Z > z^*] = \alpha/2$를 만족하는 z^*값이고 $t_{\alpha/2,(n_1+n_2-2)}$는 자유도가 $n_1 + n_2 - 2$인 t-분포에서 $\Pr[T > t^*] = \alpha/2$를 만족하는 t^*값
- S_p^2는 등분산($\sigma_1^2 = \sigma_2^2$)을 가정한 통합 분산으로 $S_p^2 = [(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2] / (n_1 + n_2 - 2)$

다. 두 모평균 차이의 신뢰구간 추정(대응표본)

- 투약 전후나 이벤트 성과 비교와 같이 짝을 이루는 각 쌍에 대한 표본을 대상으로 모평균의 차이 μ1- μ2에 대한 추정에는 대응 표본(Pairwise Sample)을 사용한다.

개체	관측값1(A)	관측값2(B)	차이(D=A-B)
1	A_1	B_1	$D_1=A_1-B_1$
2	A_2	B_2	$D_2=A_2-B_1$
...	...	...	...
n	A_n	B_n	$D_n=A_n-B_n$

- 대응표본의 특징은 모집단과 표본은 하나씩이지만, 각 개체들에 대해 두 개씩의 관측값이 존재하므로 모수는 두 개이고, 표본 내에 있는 각 개체별로 짝지어진 관측값 사이의 차이를 통해 두 모평균의 차이를 추정한다.

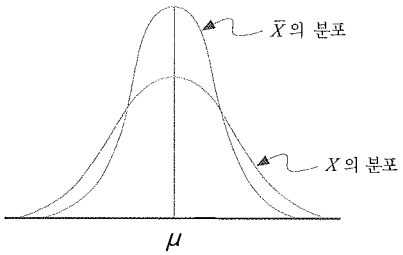
모수	점추정량	표본분포	(1-α)×100% 신뢰구간
μ_D	$\overline{D}$	$\overline{D} \sim N(\mu_D, \frac{\sigma_D^2}{n})$	$\overline{D} \pm t_{\alpha/2} \frac{S_D}{\sqrt{n}}$

- 여기서 $\mu_D = \mu_1 - \mu_2$, $\overline{D}$ 는 짝을 이룬 n개의 표본들의 차이 $(D_i = A_i - B_i)$ 들의 평균

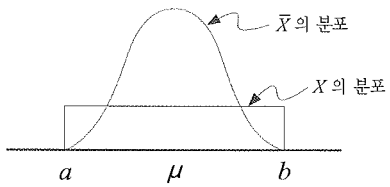
③ 중심극한정리 (Central Limit Theorem)

- 중심극한정리(Central Limit Theorem, CLT)란 평균이 μ이고 분산이 σ^2인 확률분포로부터 크기 n인 확률표본 $(X_1, X_2, ..., X_n)$을 관측할 때, 표본평균 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$는 n이 커질수록 평균이 $E(\overline{X}) = \mu$ 이고 분산 $Var(\overline{X}) = \sigma^2 / n$인 정규분포에 가까운 분포를 가진다. 즉, $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$가 된다.

- 자료가 관찰된 모집단의 분포가 실제로 정규분포가 아닌 경우에도 중심극한정리에 의하여 정규 분포를 이용한 추정량의 근사확률을 구할 수 있다.



(a) X~N(μ, σ²)인 경우

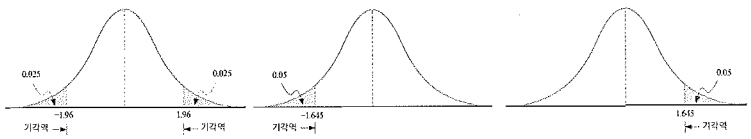


(b) X~U(a, b)인 경우

가. 대립가설 H1과 기각역 C

- 검정통계량의 분포에서 유의수준 α에 의해 기각역 C의 크기가 결정되며, 기각역의 위치는 대립가설 H1의 형태에 의해 분포의 양쪽 끝(양측검정) 또는 한 쪽 끝(단측검정)으로 나뉘어지고 오른쪽 끝에 위치하면 오른쪽 단측검정, 왼쪽 끝으로 위치하면 왼쪽 단측검정으로 분류된다.
- 귀무가설 H0가 "모수가 특정값(μ0)이다"라고 할 때 대립가설 H1

귀무가설	대립가설		
	양측검정 (모수가 특정값이 아니다)	왼쪽 단측검정 (모수가 특정값보다 작다)	오른쪽 단측검정 (모수가 특정값보다 크다)
$H_0: \mu = \mu_0$	$H_1: \mu \neq \mu_0$	$H_1: \mu < \mu_0$	$H_1: \mu > \mu_0$

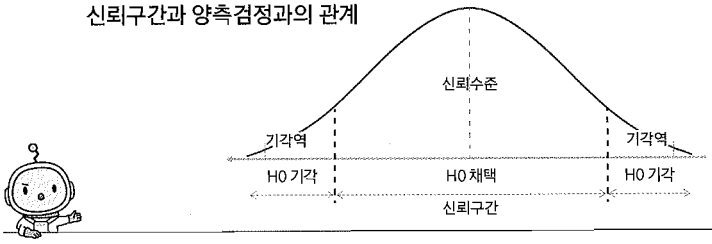


나. 가설검정 단계

- 가설검정 과정을 단계적으로 설명하면 다음과 같다.
  ① 검정하고자 하는 목적에 따라서 귀무가설(H0)과 대립가설(H1)을 설정한다.
  ② 검정통계량 T(X)을 구하고 그 분포를 구한다.
  ③ 유의수준 α를 결정하고 검정통계량 T(X)의 분포에서 대립가설의 형태에 따라 유의수준 α에 해당하는 기각역 C를 설정한다.
  ④ 귀무가설(H0)이 옳다는 전제 하에서 표본관찰에 의한 검정통계량 T(X)의 값을 구한다.
  ⑤ T(X)의 값이 기각역 C에 속하는가를 판단하여 기각역에 속하면 귀무가설(H0)을 기각하고 기각역에 속하지 않으면 귀무가설(H0)을 채택한다.

참고

신뢰구간과 양측검정과의 관계



- 유의수준이 α인 양측 검정에서의 귀무가설의 특정값(μ_0)이 $(1-\alpha)\times100\%$ 신뢰구간 내에 포함된다면, 귀무가설(H_0)을 채택한다.
- 예를 들어 95% 신뢰구간이 [49.50, 51.5]라고 가정하면, 해당 신뢰구간이 50을 포함하므로 유의수준 5%에서 귀무가설(H_0) "모평균이 50이다. (μ=50)"를 채택할 수 있다.

⑤ 상관계수

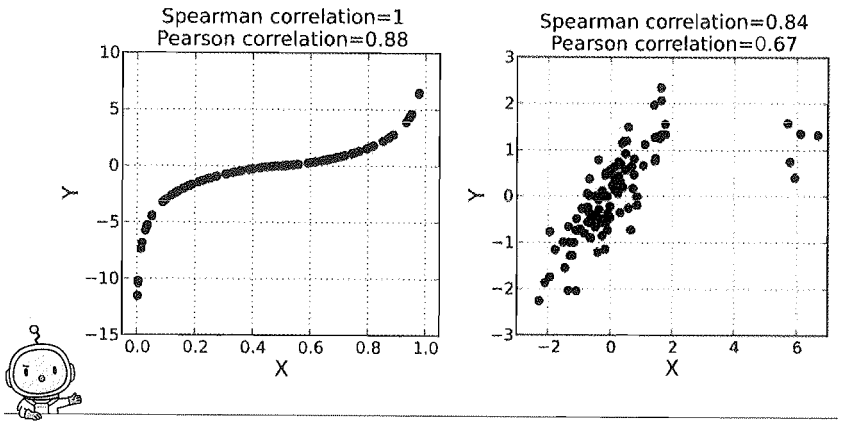
가. 피어슨-스피어만 상관계수의 관계

- 일반적으로 서열상관계수는 집단 내의 개별 관측치를 두 개의 서로 다른 관점이나 특성으로 평가한 순위값들을 이용해서 분석하는 경우에 사용
- 두 변수의 순위 사이의 의존성을 측정하는 비모수 척도로 **단조함수**(Monotonic Function)를 통해 두 변수의 관계가 얼마나 잘 설명될 수 있는 지 판단한다.
- 즉, 스피어만 상관계수는 두 변수 사이의 선형 관계를 평가하는 피어슨 상관계수와 달리, 선형 여부와 관계없이 두 변수가 단조적 관계가 있는지를 평가한다.
- 중복 데이터가 없다는 가정하에 각 변수가 다른 변수의 완벽한 단조 함수일 때 +1 또는 -1의 관계가 발생한다.

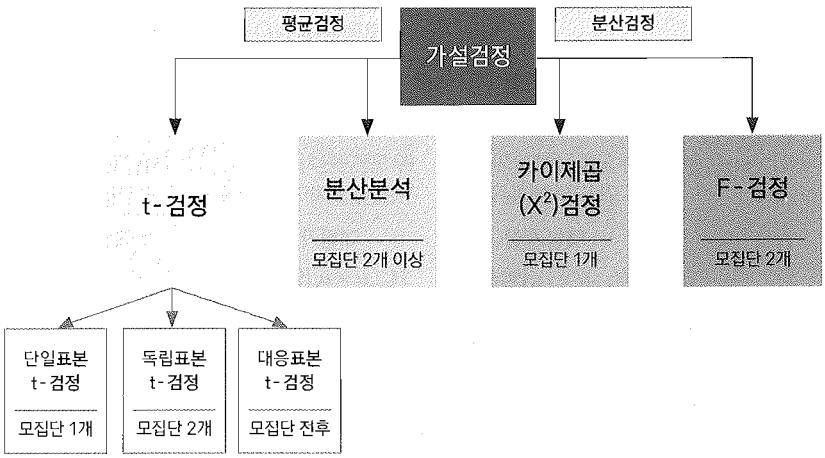
단조함수 (Monotonic Function)
순서 관계 ≤를 보존하거나 반전시키는 함수, x≤y 이면 f(x)≤f(y), x≤y 이면 f(x)≥f(y)

참고

두 변수 간의 스피어만 상관 계수 = 두 변수의 순위 값 사이의 피어슨 상관계수
두 변수 사이의 선형 관계(피어슨) vs 두 변수 사이의 단조적 관계(선형여부 아님) (스피어만)



• 모집단의 모수에 대해 추정을 한 후에는 모집단에 대해 어떤 가설(Hypothesis)을 설정한 후 그 가설의 타당성 여부를 검정한다. 한 집단, 두 집단, 독립적인 집단의 평균부터 분산 검정 등에 대한 통계 분석 방법론에 대해 학습한다.



· 단일표본t-검정 : 하나의 모집단에 대한 가설검정
· 독립표본t-검정 : 두 집단이 서로 독립적일 때 두 집단간 평균차이 검정
· 대응표본t-검정 : 동일한 모집단에 변수를 노출시키기 전과 후의 평균값 비교검정 (쌍체비교/전후비교)

▲가설검정의 종류 / t-검정 / 분산분석 / 카이제곱 검정 / F-검정

• 단일 모수의 가설검정 단계

검정단계	단일표본 t-검정	단일표본 카이제곱 검정
1단계 가설설정	모수 : 모평균(μ)	모수 : 모분산(σ^2)
	귀무가설(H_0) : $\mu = \mu_0$	귀무가설(H_0) : $\sigma^2 = \sigma_0^2$
	대립가설(H_1) : $\mu \neq \mu_0$ (양측) $\mu > \mu_0$ (우단측) $\mu < \mu_0$ (좌단측)	대립가설(H_1) : $\sigma^2 \neq \sigma_0^2$ (양측) $\sigma^2 > \sigma_0^2$ (우단측) $\sigma^2 < \sigma_0^2$ (좌단측)
2단계 유의수준 설정	유의수준(α)을 0.1, 0.05, 0.01 중에서 설정한다.	
3단계 검정통계량의 값 및 유의확률 계산	검정통계량 값 $t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t(df), df = n-1$	검정통계량 값 $\chi_0^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(df), df = n-1$
4단계 귀무가설의 기각여부 판단 및 의사결정	유의확률(p-value)<유의수준(α) : 귀무가설을 기각 유의확률(p-value)>유의수준(α) : 귀무가설을 기각하지 않는다.	

• 두 모수의 가설검정 단계

검정단계	대응표본 t-검정	독립표본 t-검정	등분산검정 (2표본)
	모수 모평균 차이(μ_D)	모수 독립인 정규 모집단의 모평균(μ_1, μ_2)	모수 독립인 정규 모집단의 모분산(σ_1^2, σ_2^2)
1단계 가설설정	귀무가설 $H_0 : \mu_D = 0$ (두 모평균은 차이가 없다. 두 모평균의 차이는 0이다)	귀무가설 $H_0 : \mu_1 = \mu_2$ (두 모평균은 차이가 없다. 두 모평균의 차이는 0이다)	귀무가설 $H_0 : \sigma_1^2 = \sigma_2^2$ (두 모분산은 같다. 두 모분산의 비는 1이다)
	대립가설 $H_1 : \mu \neq \mu_0$ (양측) $H_1 : \mu > \mu_0$ (우단측) $H_1 : \mu < \mu_0$ (좌단측)	대립가설 $H_1 : \mu \neq \mu_2$ (양측) $H_1 : \mu > \mu_2$ (우단측) $H_1 : \mu < \mu_2$ (좌단측)	대립가설 $H_1 : \sigma^2 \neq \sigma_0^2$ (양측) $H_1 : \sigma^2 > \sigma_0^2$ (우단측) $H_1 : \sigma^2 < \sigma_0^2$ (좌단측)
2단계 유의수준 설정	유의수준(α)을 0.1, 0.05, 0.01 중에서 설정한다.		
3단계 검정통계량의 값 및 유의확률 계산	검정통계량 값 $t = \frac{\overline{D} - \mu_D}{S_D / \sqrt{n}} \sim t(df)$ $df = n - 1$	검정통계량 값 (등분산 검정 선행 후) $t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(df)$ $df = n_1 + n_2 - 2$	검정통계량 값 $F = \frac{S_1^2}{S_2^2} \sim F(df_1, df_2)$ $df_1 = n_1 - 1,\ df_2 = n_2 - 1$
4단계 귀무가설의 기각여부 판단 및 의사결정	• 유의확률(p-value)<유의수준(α) : 귀무가설을 기각 • 유의확률(p-value)>유의수준(α) : 귀무가설을 기각하지 않는다.		
	(양측만) • (1-α)×100% 신뢰구간이 0을 포함하면 귀무가설을 채택한다.		(양측만) • (1-α)×100% 신뢰구간이 1을 포함하면 귀무가설을 채택한다.

⑦ 분산분석 (Analysis of Variance, ANOVA)

• 앞서 학습한 t-검정이 두 집단 간의 평균 차이를 비교하는 통계분석 방법이라면 분산분석은 두 개 이상의 다수 집단 간 평균을 비교하는 통계분석 방법이다. 분산분석은 독립변수의 개수에 따라 일원배치 분산분석, 이원배치 분산분석, 다원배치 분산분석으로 나누어진다. 분산분석의 개념을 학습 해보자.

가. 일원배치 분산분석(One-Way ANOVA)

1) 분산분석의 개념

• 분산분석은 두 개 이상의 집단에서 그룹 평균 간 차이를 그룹 내 변동에 비교하여 살펴보는 통계 분석 방법이다.

- 즉, 두 개 이상 집단들의 평균 간 차이에 대한 통계적 유의성을 검증(두 개 이상 집단들의 평균을 비교)하는 방법이다.

2) 일원배치 분산분석의 개념

- 분산분석에서 반응값에 대한 하나의 범주형 변수의 영향을 알아보기 위해 사용되는 검증 방법이다.
- 모집단의 수에는 제한이 없으며, 각 표본의 수는 같지 않아도 된다.
- F-통계량을 이용한다.
- 각 집단의 표본의 수가 동일한 경우의 예시(여기서 Y_{ij}는 i번째 집단의 j번째 관측값)

번호	집단1	집단2	…	집단k
1	Y_{11}	Y_{21}	…	Y_{k1}
2	Y_{12}	Y_{22}	…	Y_{k2}
⋮	⋮	⋮	⋱	⋮
m	Y_{1m}	Y_{2m}	…	Y_{km}

3) 일원배치 분산분석의 가정

- 각 집단의 측정치는 서로 독립적이며, 정규분포를 따른다.
- 각 집단 측정치의 분산은 같다.(등분산 가정)

4) 분산분석표

요인	제곱합(SS)	자유도(df)	평균제곱합(MS)	분산비(F)
집단 간	SSB	k-1	MSB=SSB/(k-1)	F=MSB/MSW
집단 내	SSW	N-k	MSW=SSW/(N-k)	
전체	SST	N-1		

SSB(집단 간 변동제곱합)	$m\sum_{i=k}^{k}(\overline{Y}_i-\overline{Y})^2$
SSW(집단 내 변동제곱합)	$\sum_{i=1}^{k}\sum_{j=1}^{m}(Y_{ij}-\overline{Y}_j)^2$
SST(전체 변동제곱합)	$\sum_{i=1}^{k}\sum_{j=1}^{m}(Y_{ij}-\overline{Y})^2$
SST = SSB + SSW	

- k : 집단의 수
- m : 집단별 관측치 수 (각 집단의 표본 수 동일 가정)
- N : 전체 관측치 수($N = mk$)
- $\overline{Y}_i$: 번째 집단의 표본 평균
- $\overline{Y}$: 모든 관측치 평균(총 평균)

5) 가설 검정

- 귀무가설(H_0) : k개의 집단 간 모평균에는 차이가 없다. ($\mu_1 = \mu_2 = ... = \mu_k$)
- 대립가설(H_1) : k개의 집단 간 모평균이 모두 같다고 할 수 없다.(H_0 is not true)

6) 사후 검정

- 사후 검정이란 분산분석의 결과 귀무가설이 기각되어 적어도 한 집단에서 평균의 차이가 있음이 통계적으로 증명되었을 경우, 어떤 집단들에 대해서 평균의 차이가 존재하는지를 알아보기 위해 실시하는 분석이다.
- 사후분석의 종류로는 던칸(Duncan)의 MRT(Multiple Range Test) 방법, 피셔(Fisher)의 최소유의차(LSD)방법, 튜키(Tukey)의 HSD방법, Scheffe의 방법 등이 있다.

⑧ 범주형 자료 분석

- 범주형 자료 분석은 분석에 사용되는 변수들이 범주형일 때 사용하는 분석 방법론이다.
- 설명변수와 반응변수에 따른 범주형 자료 분석 방법론은 아래와 같이 분류할 수 있다.

설명변수	반응변수	통계분석방법
범주형 자료	범주형 자료	· 분할표 분석 · 카이제곱 검정
	연속형 자료	· t-검정 · 분산분석
연속형 자료	범주형 자료	· 로지스틱 회귀분석

⑨ 분할표 (Contingency Table) 분석

가. 개요

- 여러 개의 범주형 변수를 기준으로 빈도를 표 형태로 나타낸 것을 분할표(또는 교차표)라 한다. 아래의 표는 2개의 범주형 변수(학년, 성적 등급)별 빈도를 분할표로 나타낸 것이다.

	A등급	B등급	C등급
1학년	3	10	7
2학년	4	11	5
3학년	5	10	5
4학년	7	11	2

- 범주형 변수가 1개일 때는 1원 분할표, 2개일 때는 2원 분할표, 3개 이상일 때는 다원 분할표로 표시한다.
- 분할표의 행은 설명변수, 열은 반응변수를 입력하고 범주형 자료를 분석할 때는 이 분할표를 기반으로 여러 가지 검정을 수행할 수 있다.

나. 상대위험도(Relative Risk)

- 상대위험도란 관심 집단의 위험률/비교 집단의 위험률을 의미하며, 여기서 위험률이란 특정 사건이 발생할 비율을 의미한다.
- 예를 들어 위험인자에 노출된 암환자의 확률/위험인자에 노출되지 않은 암환자의 확률이 상대 위험도이다.
- 다음과 같은 분할표에서 상대위험도를 식으로 표현하면 아래의 식과 같다.

		암 발생 여부	
		O	X
위험인자 노출 여부	O	a	b
	X	c	d

$$\text{상대위험도(RR)} = \frac{\text{위험인자에 노출된 암환자의 비율}}{\text{위험인자에 노출되지 않은 암환자의 비율}} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

가. 교차표

- 두 변수의 각 범주를 교차하여 데이터의 관측도수(빈도)를 표 형태로 나타내면 아래와 같다.

	A_1	A_2	⋯	A_N	Total
B_1	O_{11}	O_{12}	⋯	O_{1N}	$T_{1\cdot}$
B_2	O_{21}	O_{22}	⋯	O_{2N}	$T_{2\cdot}$
⋮	⋮	⋮	⋱	⋮	⋮
B_M	O_{M1}	O_{M2}	⋯	O_{MN}	$T_{M\cdot}$
Total	$T_{\cdot 1}$	$T_{\cdot 2}$	⋯	$T_{\cdot N}$	T

- 교차분석은 교차표에서 각 셀의 관찰빈도(자료로부터 얻은 빈도분포)와 기대빈도(두 변수가 독립일 때 이론적으로 기대할 수 있는 빈도분포)간의 차이를 검정한다.

나. 카이제곱검정

- 범주형 자료(명목/서열 수준)인 두 변수 간의 관계를 알아보기 위해 실시하는 분석 기법이다.
- 적합성 검정, 독립성 검정, 동질성 검정에 사용되며, 카이제곱 x^2 으로 검정통계량을 이용한다.

1) 적합성 검정

- 실험에서 얻어진 관측값들이 예상한 이론과 일치하는지 아닌지를 검정하는 방법이다.
- 관측값들이 어떠한 이론적 분포를 따르고 있는지를 알아볼 수 있다. 즉, 모집단 분포에 대한 가정이 옳게 됐는지를 관측 자료와 비교하여 검정하는 것이다.

2) 독립성 검정

- 모집단이 두 개의 변수 A, B에 의해 범주화되었을 때, 이 두 변수들 사이의 관계가 독립인지 아닌지를 검정하는 것을 의미한다.
- 모집단을 범주화하는 기준이 되는 두 변수 A, B가 서로 독립적으로 관측값에 영향을 미치는지의 여부를 검정하는 것이다.
- 검정 통계량 값을 계산할 때는 교차표를 활용한다.

3) 동질성 검정

- 모집단이 임의의 변수에 따라 R개의 속성으로 범주화되었을 때, R개의 부분 모집단에서 추출한 각 표본인 C개의 범주화된 집단의 분포가 서로 동일한지를 검정하는 것을 의미한다.
- 검정 통계량 값을 계산할 때는 교차표를 활용하며, 계산법과 검증법은 모두 독립성 검정과 같은 방법으로 진행된다.

	적합성	독립성	동질성
가설 검정	• H_0: 실제 분포와 이론적 분포 간에는 차이가 없다. (두 분포가 일치한다) • H_1: 실제 분포와 이론적 분포 간에는 차이가 있다. (두 분포가 일치하지 않는다)	• H_0: 두 변수 사이에는 연관이 없다.(독립이다) • H_1: 두 변수 사이에는 연관이 있다.(종속이다)	• H_0: $P_{1j} = P_{2j} = ... = P_{rj}$ (모든 P_{nj} 는 동일하다 $(n = 1, 2, ..., r)$) • H_1 : $not\ H_0$

	• $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ • O_i :관찰도수, E_i :기대도수, $i = 1, 2, ..., k$	• $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ • $E_{ij} = \frac{O_{i\cdot} \times O_{\cdot j}}{n}$: 기대빈도 • n : 전체관측도수, O_{ij} : 관찰빈도 • $O_{i\cdot}$: 행의 합, $O_{\cdot j}$: 열의 합	
검정 통계량	• χ^2 통계량 값이 큰 경우 : 관찰 도수와 기대도수의 차이가 크며, 적합도가 낮다. 즉, 일치한다고 볼 수 없다.	• χ^2 통계량 값이 큰 경우 : 두 변수 사이에는 연관이 있다. 즉, 두 변수는 종속 관계이다.	• χ^2 통계량 값이 큰 경우 : (P_{nj} 중 다른 값이 하나 이상 존재한다. $(n = 1, 2, ..., r)$)
	• χ^2 통계량 작은 경우 : 관찰 도수와 기대도수의 차이가 작으며, 적합도가 높다. 즉, 일치한다고 볼 수 있다.	• χ^2 통계량 작은 경우 : 두 변수 사이에는 연관이 없다. 즉, 두 변수는 독립 관계이다.	• χ^2 통계량 작은 경우 : (모든 P_{nj} 는 동일하다. $(n = 1, 2, ..., r)$)
	• df = k-1(k는 범주의 개수)	• df = (R-1)(C-1) (R:행의 수, C:열의 수)	

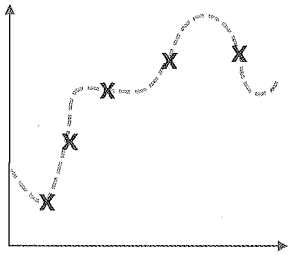
11 더빈 왓슨 (Durbin-Watson) 검정

- 회귀 분석의 주요한 가정 중 오차항이 독립성을 만족하는지를 검정하기 위해서는 더빈 왓슨(Durbin-Watson) 검정을 수행한다.
- 검정 결과 더빈 왓슨 통계량이 2에 가까울수록 오차항의 자기상관이 없음을 의미한다.
- 더빈 왓슨 통계량의 값이 0에 가까울수록 양의 상관관계가 있음을 의미하고, 4에 가까울수록 음의 상관관계가 있음을 의미한다. 따라서 더빈 왓슨 통계량이 0 혹은 4에 가까울 경우 잔차들 간의 상관관계가 있어서 회귀식이 부적합함을 의미한다.

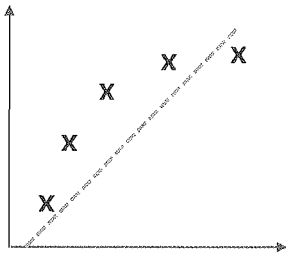
⑫

정규화 선형회귀 (Regularized Linear Regression)

- 정규화 선형회귀는 선형회귀 계수에 대한 제약 조건을 추가하여 모델이 과도하게 최적화되는 현상(과적합, Overfitting)을 막는 방법이다.
- 오른쪽의 그림에서 (a)그래프는 모델이 학습 데이터를 매우 잘 적합하고 있지만, 미래 데이터가 조금만 바뀌어도 예측값이 과도하게 변할 수 있다. 반면 (b)그래프는 정규화를 수행하여, 학습데이터에 대한 설명력을 조금은 포기하는 대신 미래 데이터의 변화에 대해 상대적으로 안정된 결과를 낼 수 있다.
- 이와 같이, 모형이 과적합되면 계수의 크기도 과도하게 증가하는 경향이 있다. 따라서 정규화 선형 회귀에서는 계수의 크기를 제한하는 방법으로 제약조건을 추가한다.
- 정규화 선형회귀에서는 제약조건의 종류에 따라 Ridge회귀, Lasso회귀, ElasticNet회귀모형이 일반적으로 사용된다.



(a) 과대적합



(b) 정규화

가. 릿지회귀

- 릿지(Ridge)회귀모형은 가중치들의 제곱합(Squared Sum of Weights)을 최소화하는 것을 제약 조건으로 추가하는 기법이다(능형 회귀모형이라고도 함).

$$\omega = \arg\min_{\omega} \left(\sum_{i=1}^{N} e_i^2 + \lambda \sum_{j=1}^{M} \omega_j^2 \right)$$

- 릿지회귀모형에서는 가중치의 모든 원소가 0에 가까워지는 것을 원하며, 이를 위해 회귀 모델에 사용하는 규제 방식을 L2 규제(Penalty)라고 한다.
- λ는 기존의 잔차 제곱합과 추가적인 제약조건의 비중을 조절하기 위한 초매개변수(Hyper Parameter)에 해당한다. λ가 커지면 가중치의 값들이 작아지며, 정규화 정도가 커진다. λ가 작아지면 정규화 정도가 작아지고, λ가 0이 되면 일반적인 선형회귀모형이 된다.

나. 라쏘회귀

- 라쏘(Least Absolute Shrinkage and Selection Operator, Lasso)회귀모형은 가중치 절대값의 합을 최소화하는 것을 제약조건으로 추가하는 기법이다.

- 릿지회귀에서는 가중치가 0에 가까워질 뿐, 실제로 0이 되지는 않는다. 하지만 라쏘회귀에서 중요하지 않은 가중치는 0이 될 수도 있다.

$$\omega = \arg\min_{\omega} \left(\sum_{i=1}^{N} e_i^2 + \lambda \sum_{j=1}^{M} | \omega_j | \right)$$

- 라쏘회귀에서 사용하는 규제 방식을 L1 규제(Penalty)라고 한다.

다. 엘라스틱넷(Elastic Net)

- 엘라스틱넷(Elastic Net)은 릿지회귀와 라쏘회귀를 결합한 모델이다.

$$\omega = \arg\min_{\omega} \left(\sum_{i=1}^{N} e_i^2 + \lambda_1 \sum_{j=1}^{M} | \omega_j | + \lambda_2 \sum_{j=1}^{M} \omega_j^2 \right)$$

- 가중치 절댓값의 합과 제곱합을 동시에 제약조건으로 가지는 모형임으로 λ_1과 λ_2라는 두 개의 초매개변수를 가진다.

16

회귀분석의 영향력 진단

- 영향력 진단이란 적합된 회귀모형의 안전성을 평가하는 통계적인 방법이다.
- 자료에서 특정 관측치가 제외됨에 따라 분석 결과의 주요 부분에 많은 변동이 있다면 안전성이 약하다고 판단한다.
- 선형회귀분석에서 회귀직선의 기울기에 영향을 크게 주는 점을 영향점이라고 한다.
- 영향력 진단의 방법에는 Leverage H, Cook's Distance, DFBETAS, DFFITS 등이 있다.

영향력 진단 방법	설 명	공 식
Leverage H (지레점, 레버리지)	레버리지는 $H = X(X^TX)^{-1}X^T$(Hat Matrix)의 i번째 대각원소로 관측치가 다른 관측치 집단으로부터 떨어진 정도를 의미하며, 2×(p+1)/n 보다 크면 영향치이거나 이상치라고 본다.	$h_{ii} = x_i^T (X^TX)^{-1} x_i$
Cook's Distance (쿡의 거리)	쿡의 거리는 Full Model에서 i번째 관측치를 포함하여 계산한 적합치와 i번째 관측치를 포함하지 않고 계산한 적합치 사이의 거리이다. 쿡의 거리가 기준값인 1보다 클 경우에 영향치로 간주한다.	$C_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{(p+1)MSE}$ $\hat{Y}_{j(i)}$: i번째 관측치를 포함하지 않고 계산한 j번째 추정치

DFBETAS (Difference in Betas)	DFBETAS의 절대값이 커지면 i번째 관측치가 영향치 혹은 이상치일 가능성이 높다. 기준값은 2나 2/√n(표본을 고려한 경우)을 사용하며, DFBETAS 값이 기준값보다 클 경우 영향치로 간주한다.	$DFBETAS_{k(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{MSE_{(i)}h_{ii}}}$ $\hat{\beta}_{k(i)}$: i번째 관측치를 포함하지 않고 계산한 k번째 추정 회귀계수
DFFITS (Difference in Fits)	i번째 관측치 제외시 종속변수 예측치의 변화정도를 측정한 값이다. DFFITS의 절대값이 기준값인 2×(p+1)/n보다 클수록 영향치일 가능성이 높다고 본다.	$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)}h_{ii}}}$ $\hat{Y}_{i(i)}$: i번째 관측치를 포함하지 않고 계산한 i번째 추정값

12 변수 선택의 기준으로 사용되는 통계량

모수 절약의 원칙 (Principle of Parsimony)

회귀모형을 구축할 때 가능한 작은 수의 독립변수를 이용해야 하는 통계학적 원칙 → 모형의 간명성

가. 수정된 결정계수(Adjusted R Square, R_a^2)

- 설명변수의 개수가 증가하면 결정계수도 함께 증가하는 속성을 가진다.
- 따라서 수정된 결정계수를 이용해 이러한 단점을 보완하고 변수를 선택할 수 있다. 수정된 결정 계수는 변수의 개수가 증가함에 따라 처음에는 감소하다가 점점 안정화되고 나중에는 약간 증가하는 경향을 가진다.

$$R_a^2 = 1 - \frac{n-1}{n-p}(1-R)^2$$

- 수정된 결정계수를 이용하여 변수를 선택할 경우, MSE값이 최소인 시점의 모형을 선택하거나 이 값의 최소와 비슷해서 더 이상 변수를 추가할 필요가 없는 시점의 모형을 선택하게 된다.

나. Mallow's Cp

- Mallow가 제안한 통계량으로 Cp값은 최소자승법(Ordinary Least Squares)을 사용하여 추정된 회귀모형의 적합성을 평가하는데 사용된다.

$$C_p = \frac{SSE_p}{MSE} + 2p - n$$

- 여기서 SSE_p는 p번째 변수를 제외함으로써 줄어드는 오차제곱합의 양
- 일반적으로 Cp값이 작고, p + 상수(변수의 개수 + 상수)에 가까운 모형을 선택한다.

Cp값	해석
Cp값이 p(변수의 개수)와 비슷한 경우	Bias(편향)가 작고 우수한 모델을 의미
Cp값이 p(변수의 개수)보다 큰 경우	Bias(편향)가 크고 추가적인 변수가 필요한 모델을 의미
Cp값이 p(변수의 개수)보다 작은 경우	Variance(분산)의 증가폭보다 Bias(편향)의 감소폭이 더 크며, 필요 없는 변수가 모델에 있다는 것을 의미

15

실험계획법 (Design Of Experiment, DOE)

가. 실험 계획법의 개념

- 시스템이나 프로세스의 결과에 영향을 미치는 인자를 도출하고, 측정 데이터를 통계적으로 분석하기 위한 실험을 설계하는 방법을 의미한다.
- 실험 방식, 데이터 수집 방법, 활용 통계 기법 등 실험의 모든 과정을 설계한다.
- 최소 실험 횟수로 최대의 정보를 얻는 것을 목적으로 한다.

나. 계획 설계의 목적

- 분산분석 및 검정과 추정의 문제 : 어떠한 요인이 특성치 변화에 유의미한 영향을 주는지, 또한 해당 요인의 영향이 어느 정도인지를 파악
- 최적 반응 조건의 결정 문제 : 어떤 인자를 사용해야 가장 원하는 결과값을 얻을 수 있는지를 파악
- 오차항 추정의 문제 : 이해하기 어렵던 오차와 그 변동에 관한 정도를 파악

다. 실험계획의 원리

- 랜덤화의 원리(Randomization) : 실험 순서를 무작위로 선택하여 실시
- 반복의 원리(Replication) : 인자의 동일 수준 내에서 최소 두 번 이상 실험을 진행
- 블록화의 원리(Blocking) : 실험 전체를 시간적·공간적으로 분할하여 블록으로 만듦
- 직교화의 원리(Orthogonality) : 요인간 직교성을 갖도록 실험을 계획
- 교락의 원리(Confounding) : 고차항의 교호효과와 블록효과를 교락시키는 방법

라. 주요 용어

- 인자(Factor) : 실제 실험의 대상, 입력변수 X
- 특성치(Characteristic Value) : 실험의 모든 결과값, 출력변수 Y
- 수준(Level) : 실험하기 위한 인자의 조건, 인자의 정도나 값
- 주효과(Main Effect) : 각 입력변수의 수준 간 차이, 인자가 독립적으로 반응에 미치는 영향
- 교호효과(Interaction Effect) : 특정한 인자 수준의 조합에서 일어나는 효과, 인자들이 혼합되어 반응에 미치는 영향
- 교락(Confounding) : 2개 이상의 효과(주효과 또는 교호효과)를 구별할 수 없도록 계획적으로 조합하는 것

- 블록(Block) : 실험 단위가 균일할 수 있도록 단위를 모은 것
- 반복(Replication) : 인자들의 동일한 수준 조합에서 다회의 실험을 진행
- 중복(Repetition) : 한 실험에서 여러 개의 대상을 측정

마. 실험 계획법의 종류

1) 요인배치법(Factorial Design)

- 모든 인자간의 수준 조합에서 실험이 이루어지는 완전랜덤화방법이다.
- 교호효과를 포함한 모든 요인효과를 추정할 수 있다.
- K^n 형 요인실험 : 인자 수가 n이고, 각 인자의 수준 수가 k인 실험계획법이다.

2) 분할법(Split-Plot Design)

- 완전랜덤화하기 힘들 경우, 몇 단계로 분할하여 각 단계별로 완전 랜덤하게 실험 순서를 결정하는 방법이다.
- 랜덤화가 가장 어려운 것을 1차 단위로, 비교적 쉬운 것을 후(後) 단위로 배치한다.

3) 교락법(Confounding Method)

- 검출할 필요가 없는 교호작용을 다른 요인과 교락하도록 배치하는 방법이다.
- 실험 횟수를 늘리지 않고 실험 전체를 몇 개의 블록으로 나누어 배치하는 방법으로, 동일 환경에서의 실험 횟수를 줄일 수 있다.
- 고차의 교호작용을 블록에 교락시키기 때문에, 주효과가 높게 추정된다.

4) 난괴법(Randomized Block Design, RBD, 랜덤화 블록 실험설계)

- 실험 단위를 몇 개의 반복으로 나누어 배치하는 방법이다.
- A가 모수인자이고, B가 변량인자일 때, A인자의 수준 수가 l이고, B인자의 수준 수가 m인 반복이 없는 이원배치 분산분석방법이다.
- 실험 오차를 줄일 수 있기 때문에 효율이 높고 비교적 분석이 간단하다.

MEMO