제4장. 통계분석

1. 통계 분석의 이해

1. **통계**

- _ _ _ _ : 특정집단을 대상으로 수행한 _ _ _ _ 나 _ _ _ _ 통해 나온 결과에 대한 요약된 형태의 표현
- 획득방법 : _ _ _ _ 와 _ _ _ _ _ _
- _ _ _ _ 기법 : _ _ _ _ _ _ _ _, _ _ _ _ _ _ _ _ _ _, _ _ _ _ _ _ _ _ _
- 자료의 형태 : _ _ _ _ _ _, _ _ _ _ _ _, _ _ _ _ _ _, _ _ _ _ _ _ _

1. **통계분석**

- _ _ _ _ _ _ _ _ : 추설, 가설검정, 예측
- _ _ _ _ _ _ _ _ : 평균, 표준편차, 중위수, 최빈값,, 그래프의 표현

2. **확률 및 확률 분포**

- _ _ _ _ _ _ : 특정값이 나타날 확률의 변수
- 이산형 확률분포(_ _ _ _ _ _ _ _) : _ _ _ _ _ _ 분포, _ _ _ _ 분포, _ _ _ _ 분포, _ _ _ _ _ _ 분포, _ _ _ _ _ _ 분포
- 역속형 확률분포(_ _ _ _ _ _ _ _) : _ _ _ _ _ _ 분포, _ _ _ _ _ _ 분포, _ _ _ _ _ _ 분포, _ _ 분포, _ _ 분포, _ _ 분포

3. **추정 및 가설 검정**

- _ _ _ _ : 표본으로 미지의 모수를 추측하는 것
- _ _ _ _ _ _ : 모수가 특정 값으로 추정. _ _ _ _ , _ _ _ _ _ _ , _ _ _ _ _ _ 등을 추정.
  - _ _ _ _ _ _ (_ _ _ _ _ _ _ _) 조건 : _ _ _ _ , _ _ _ _ _ _ , _ _ _ _ , _ _ _ _ _ _ _ _
- _ _ _ _ _ _ (_ _ _ _ _ _ _ _)
  - 모수가 특정 구간에 있을 것으로 추정
  - 모분산을 알경우 _ _ _ _ _ _ _ _ 활용, 모를경우 _ _ _ _ _ _ _ _ 활용
- 가설검정
  - _ _ _ _ _ _ _ _ (_ _ _ _ _ _ _ _), _ _ _ _ _ _ _ _ (_ _ _ _ _ _ _ _)
  - _ _ _ _ (_ _ _ _) : _ _ _ _ 가설이 옳은데 귀묵가설을 기각하는 오류
  - _ _ _ _ (_ _ _ _) : _ _ _ _ 가설이 옳지 않은데 _ _ _ _ 을 채택하는 오류
  - _ _ _ _ _ _ 크기를 0.1, 0.05, 0.01로 고정시키고 _ _ _ _ _ _ 가 최소가 되고록 _ _ _ _ 을 설정

4. **비모수 검정**

- 모집단의 분포에 아무 _ _ _ _ 을 가하지 않고 실행하는 검정
- "동일하다/동일하지 않다" 식으로 가설 설정
- 순위나 두 관측값 차이의 부호를 이용해 검정

- _ _ _ _ _ _ _ _(_ _ _ _ _ _ _ _ _), _ _ _ _의 _ _ _ _ _ _ _ _(_ _ _ _ _ _ _ _), _ _ _ _의 _ _ _ _ _ _ _ _(_ _ _ _ _ _ _ _), _ _ _ _의 _ _ _ _ _ _ _ _(_ _ _ _ _ _ _ _), _ _ _ _ _, _ _ _ _ _ _의 _ _ _ _ _ _ _ _(_ _ _ _ _ _ _ _)

2. 기술 통계 분석

1. **기술통계**

- 자료의 특성을 _ _, _ _ _ _, _ _ _ _ 등을 사용하여 쉽게 파악할 수 있도록 정리/요약

2. **통계량에 의한 자료 분석**

- _ _ _ _ _ _ : 평균, 중앙값, 최빈값
- _ _ _ _의 척도 : _ _, _ _ _ _, _ _ _ _, _ _ _ _ _ _ _ _, 변동계수, 표준오차
- 분포의 형태 : _ _ _ _, _ _ _ _

3. **그래프를 통한 자료 분석**

- _ _ _ _ 형 자료 : 막대그래프, 파이차트 등
- _ _ _ _ 형 자료 : 히스토그램, 줄기-잎 그림, 상자그림
- _ _ _ _ 자료 : 꺽은선 그래프

4. **연관성 분석**

- _ _ _ _ _ _ _ _, _ _ _ _ _ _ _ _
- _ _ _ _ _ _ _ _(_ _ _ _ _ _ _ _ _)로 확인할 수 있는 것
  - 두 변수사의 선형관계, 함수관계 성립
  - 이상값의 존재 여부와 몇 개의 집단으로 구분
- _ _ _ _ _ _ _ _(_ _ _ _ _ _ _ _ _)
  - 두 확률변수 간의 방향성을 확인

5. **상관분석**

- 두 변수간의 _ _ _ _ _ _ 정도를 _ _ _ _ _ _ _ _ 를 통해 확인할 수 있음
- -1에서 1사이의 값으로 표현, 0이면 _ _ _ _ _ _가 없음
- _ _ _ _ _ _ _ _ _ _ _ _ _ : _ _ _ _ 척도 이상으로 측정된 계수
- _ _ _ _ _ _ _ _ _ _ _ _ _ : _ _ _ _ 및 _ _ _ _ 척도로 측정된 계수
- 프로그램
  - _ _ _ _ _ _ : _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
  - _ _ _ _ _ _ : _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

3. 회귀분석

1. **회귀분석**

- _ _ _ _ _ _ : 하나 이상의 _ _ _ _ 변수들이 _ _ _ _ 변수에 미치는 영향을 추정할 수 있는 통계기법
- _ _ _ _ 선형회귀 / _ _ _ _ 선형회귀

2. **회귀분석 특징**

- 회귀식(모형)에 대한 검증 : _ _ _ _ 검정
- 회귀계수 검정 : _ _ _ _ 검정

- 모형의 설명력 : _ _ _ _ 계수 = _ _ _ _ _ _ / _ _ _ _ _ _ (_ _ _ _ / _ _ _ _) 단 0과 1사이값
- 선형회귀분석의 가정
  - _ _ _ _ _ _ : 입력변수와 출력변수와 관계가 _ _ _ _ : _ _ _ _ 와 출력변수의 _ _ _ _ 로 확인
  - _ _ _ _ _ _ : 잔차와 _ _ _ _ 값이 관련되어 있지 않음 : _ _ _ _ 와 _ _ _ _ 의 _ _ _ _ 로 확인
  - _ _ _ _ _ _ : _ _ _ _ _ _ 의 모든 값에 대한 오차들의 _ _ _ _ 이 일정
  - _ _ _ _ _ _ : 관측치들의 _ _ _ _ 들이 상관이 없어야 함
  - _ _ _ _ _ _ : 잔차항이 _ _ _ _ 분포를 이뤄야 함

3. **다중선형회귀분석**

- _ _ _ _ _ _ : 변수들 사이에 _ _ _ _ 관계가 존재하면 _ _ _ _ _ _ 의 정확한 추정이 어려움
- 검사방법
  A. _ _ _ _ _ _ _ _ (_ _ _ _ _ _) : 10보다 크면 심각한 문제
  B. _ _ _ _ _ _ _ : 10이상이면 문제, 30 보다 크면 심각
  C. _ _ _ _ _ _ 가 강한 변수 제거

4. **변수선택법**

- 모든 가능한 조합 : 모든 가능한 변수의 조합에 대한 회귀모형을 분석
- _ _ _ _ _ _ _ _ (_ _ _ _ _ _ _ _) : 중요한 변수를 차례로 추가하는 방법
  - 이해 쉬운, 많은 변수 활용, 작은 변동에 결과가 달라져 _ _ _ _ _ _ 이 부족.
- _ _ _ _ _ _ _ _ (_ _ _ _ _ _ _ _) : 독립변수 후보를 모두 포함한 후 영향력이 적은 변수를 제거
  - 전체 변수 정보 이용 가능, 변수가 많을 경우 활용 어려움, _ _ _ _ _ _ 부족
- _ _ _ _ _ _ _ _ (_ _ _ _ _ _ _ _) : 새롭게 추가한 변수의 _ _ _ _ _ _ 가 약화되면 제거

4. 시계열 분석

1. **시계열 자료**

- _ _ _ _ _ _ _ _ : 시간의 흐음에 따라 관찰된 값들

2. **정상성**

- 3가지를 모두 만족 : _ _ _ _ 이 일정 / _ _ _ _ 이 일정 / _ _ _ _ 특정시점에서 t,s에 의존하지 않고 일정

3. **정상시계열의 특징**

- 어떤 시점에서 _ _ _ _ 과 _ _ _ _ 그리고 특정한 _ _ _ _ 의 길이를 갖는 _ _ _ _ _ _ _ _ 을 측정하면 동일한 값
- 항상 _ _ _ _ _ _ 으로 회귀하려는 경향, _ _ _ _ 은 평균값 주변에서 일정한 _ _ 유지
- _ _ _ _ 시계열은 특정 기간의 시계열 자료애서 얻은 정보를 다른 시기로 _ _ _ _ _ _ 할 수 없음

4. **시계열 모형**

- _ _ _ _ _ _ _ _ (_ _ _ _ _ _ _ _) : ACF는 빠르게 감소, PACF는 절단점이 존재 : _ _ _ _ (절단점-1)
- _ _ _ _ _ _ _ _ (_ _ _ _ _ _ _ _) : ACF는 절단점이 존재, PACF는 빠르게 감소
- _ _ _ _ _ _ _ _ _ _ _ _ _ (_ _ _ _ _ _ _ _ _ _ _ _)
  - d=0 이면 _ _ _ _ _ _ 모형이라고 부르고, 정상성을 만족
  - p=0 이면 _ _ _ _ _ _ 모형이라고 부르고, d번 _ _ _ _ 하면 _ _ _ _ 모형을 따름
- _ _ _ _ _ _ _ _ _ _
  - _ _ _ _ _ _ _ _ (_ _ _ _ _ _ _ _) : 형태가 오르거나 또는 내리는 추세를 따르는 경우
  - _ _ _ _ _ _ _ _ (_ _ _ _ _ _ _ _) : 요일, 월, 사분기 등 고정된 주기에 따라 변화하는 경우
  - _ _ _ _ _ _ _ _ (_ _ _ _ _ _ _ _) : 알려지지 않은 주기로 변화하는 경우

- _ _ _ _ _ _ _ _ _ (_ _ _ _ _ _ _ _ _) : 이 세가지 요인으로 설명할 수 없는 회귀분석에서 _ _ _ _ 에 해당하는 요인

5. 다차원척도법(MDS)

1. **다차원 척도법** 군집분석과 같이 개체들 사이의 _ _ _ _ / _ _ _ _ _ _ 을 측정하여 2차원, 3차원 공간상에 점으로 표현하는 방법

- 목적 : 군집분석은 개체들간의 동일한 그룹으로 분류 / 다차원척도법은 점으로 표시하여 개체들 사이의 _ _ _ _ 를 시각적으로 표현
- 종류
  - _ _ _ _ _ _ _ _ _ (= _ _ _ _ _ _ _ _ _ _ _ _ _ _ _) : 구간이나 비율척도일 경우, 각 개체들간의 _ _ _ _ _ _ 거리를 계산
  - _ _ _ _ _ _ _ _ _ : 순서척도일 경우 활용. 순서척도일 경우 거리 속성과 같도록 _ _ _ _ _ _ 하여 거리르 생성
- _ _ _ _ _ _ 와 _ _ _ _ _ _ 수준 M
  - 개체들을 공간상에 표현하기 위한 방법
  - _ _ _ _ _ _ 나 _ _ _ _ _ _ _ _ 를 부적합도 기준으로 사용
  - 부적합도를 최소로 하는 방법을 반복실행

6. 주성분분석(PCA)

1. **주성분 분석**

- _ _ _ _ 관계가 있는 변수들을 결합하여 _ _ _ _ _ _ 관계가 없는 변수로 분산을 극대화하는 방법
- 변수를 _ _ _ _ 하는데 사용
- _ _ _ _ 분석 : 잠재된 변수를 추출하기 위한 작업
- _ _ _ _ 분석 : 그 중 가장 많이 사용되는 방법
- 공통점 : 데이터를 _ _ _ _ 하는데 사용
- 차이점
  - 생성된 변수의 _ _ 와 _ _ _ _ : 요인분석은 지정 가능, 주성분 분석은 보통 2개
  - 변수와의 _ _ _ _ : 요인분석은 _ _ _ _ 한 관계, 주성분 분석은 중요도에 따라 차이
  - _ _ _ _ 변수와의 관계 : 요인분석은 _ _ _ _ 변수 고려안함, 주성분은 _ _ _ _ 변수를 고려하여 변수 생성
-

2. **주성분 분석의 활용**

- 여러 분석의 _ _ _ _ _ _, _ _ _ _ _ _ 을 이용해 주성분차원으로 변수를 축소
- 회귀나 의사결정나무 등에서 _ _ _ _ _ _ _ _ 이 존재할 경우, _ _ _ _ 가 높은 변수를 축소

3. **R결과 해석**

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6    PC7
Standard deviation     2.1119 1.0928 0.72181 0.67614 0.49524 0.27010 0.2214
Proportion of Variance 0.6372 0.1706 0.07443 0.06531 0.03504 0.01042 0.0070
Cumulative Proportion  0.6372 0.8078 0.88223 0.94754 0.98258 0.99300 1.0000
```

- 제 1주성분, 제 2주성분 누적 기여율은 _ _ _ _ _ _ %
- 1 주성분의 기여울(해석률)은 _ _ _ _ %, 2 주성분은 _ _ _ _ %