빅데이터의이해 HW14

– 통계분석 기법 요약 정리하기 –

201614010 영어영문학과(휴먼ICT) 박도준

1. 교차분석

1) 사용 목적

- 명목척도나 순서척도 변수인 두 개의 변수들 간의 관련성을 알아보기 위함

2) 사용 척도

- 두 변수 모두 명목, 또는 순서척도
- 비율척도의 경우 코딩변경을 통해 범주형 변수로 변환시켜주어야 함

3) 분석 순서

① 가설 설정

- 귀무가설 H_0: 두 변수는 서로 독립적이다.
- 대립가설 H_1: 두 변수는 연관성이 있다.

② 교차표(이차원 도수분포표)를 작성

③ 관측빈도(각 행와 열이 교차하는 셀에 해당하는 케이스 수)와 기대빈도(관측 전에 고르게 나올 경우 예상되는 빈도)를 계산

④ 검정통계량(X^2) 계산

⑤ 임계값에 의한 의사결정

- 설정한 신뢰구간에 따른 유의수준(e.g. 0.05, 0.01, 0.1)과 유의확률(p값)을 비교
- 유의수준보다 p값이 작을 경우 통계적으로 유의하기에 귀무가설을 기각하고, 대립가설을 채택한다.
- 유의수준보다 p값이 클 경우 통계적으로 유의하지 못하므로 귀무가설을 채택한다.

4) SPSS 절차

- 분석 → 기술통계량 → 교차분석

2. 상관분석

1) 사용 목적

- 서로 관련된다고 예측되는 두 개의 구간, 비율척도 변수들에 대해 얼마나 선형적 연관성이 있는 지 알아보기 위함

2) 사용 척도

- 두 변수 모두 구간, 또는 비율척도

3) 상관계수 종류

- Pearson 상관계수: 대표본(케이스 수 30이상)이거나, 정규분포인 경우
- Spearman 상관계수: 소표본(케이스 수 30미만)이고, 정규분포가 아닌 경우
- Kendall의 tau 상관계수: 소표본(케이스 수 30미만)이고, 정규분포가 아닌 경우
- 편상관계수: 두 변수 간의 관련성에 영향을 미치는 다른 변수를 통제하고, 순수한 두 변수 간의 상관계수를 계산

4) 분석 순서

① 가설 설정

- 귀무가설 H_0: $p = 0$ (모집단에서는 상관관계가 없다.)
- 대립가설 (양쪽 검정) H_1: $p \neq 0$ (모집단에서는 상관관계가 있다.)
  (한쪽 검정) H_1: $p > 0$ (모집단에서는 양의 상관관계가 있다.)
  혹은 H_1: $p < 0$ (모집단에서는 음의 상관관계가 있다.)

② 산점도 출력

③ 검정통계량(T) 계산

- 비모수적 검정 방법에서는 계산하지 않음

④ 임계값에 의한 의사결정

- 설정한 신뢰구간에 따른 유의수준(e.g. 0.05, 0.01, 0.1)과 유의확률(p값)을 비교
- 유의수준보다 p값이 작을 경우 통계적으로 유의하기에 귀무가설을 기각하고, 대립가설을 채택한다.
- 유의수준보다 p값이 클 경우 통계적으로 유의하지 못하므로 귀무가설을 채택한다.

5) SPSS 절차

- 분석 → 상관분석 → 이변량 분석

3. T검정

1) 사용 목적

- 두 집단의 평균을 비교하기 위함
- 대표본의 경우 바로 T검정 적용이 가능하고, 소표본의 경우 정규모집단 이어야 가능

2) 사용 척도

- 집단 변수는 명목, 순서척도
- 검정 변수는 구간, 비율척도

3) T검정 종류

- 독립표본 T 검정: 독립적인 두 집단의 평균을 비교하기 위해 사용
- 대응표본 T 검정: 한 집단의 두 변수의 값을 비교하기 위해 사용

4) 분석 순서

① 가설 설정

- 귀무가설 H_0: $\mu_1 = \mu_2$ (두 변수의 평균은 같다)
- 대립가설 H_1: (유형 1) $\mu_1 > \mu_2$ (변수 1의 평균이 변수 2의 평균보다 크다)
  (유형 2) $\mu_1 < \mu_2$ (변수 2의 평균이 변수 1의 평균보다 크다)
  (유형 3) $\mu_1 \neq \mu_2$ (변수 1의 평균이 변수 2의 평균과 다르다)

② 분석결과 출력

③ 검정통계량(T) 계산

④ 임계값에 의한 의사결정

- 설정한 신뢰구간에 따른 유의수준(e.g. 0.05, 0.01, 0.1)과 유의확률(p값)을 비교
- 유의수준보다 p값이 작을 경우 통계적으로 유의하기에 귀무가설을 기각하고, 대립가설을 채택한다.
- 유의수준보다 p값이 클 경우 통계적으로 유의하지 못하므로 귀무가설을 채택한다.

5) SPSS 절차

- 분석 → 평균비교 → 독립표본 T 검정 / 대응표본 T 검정

4. 회귀분석

1) 사용 목적

- 한 개 이상의 독립변수를 가지고 종속변수에 대한 선형회귀모형을 만들기 위함
- 대표본의 경우 바로 T검정 적용이 가능하고, 소표본의 경우 정규모집단 이어야 가능

2) 사용 척도

- 독립변수는 구간, 비율척도, 혹은 더미(dummy)변수
- 종속변수는 구간, 비율척도

3) 회귀분석 종류

- 단순선형회귀모형: 종속변수를 설명하는 독립변수가 하나일 경우 사용
- 다중회귀모형: 종속변수를 설명하는 독립변수가 두개 이상일 경우 사용

4) 분석 순서

① 가설 설정

- 귀무가설 H_0: $\beta_1 = \beta_2 = \dots = \beta_l = 0$ (모든 독립변수들이 종속변수에 영향을 주지 않는다)
- 대립가설 H_1: 적어도 하나의 $\beta_i \neq 0$, $i = 1, \dots, k$
  (적어도 하나의 독립변수가 종속변수에 영향을 준다)

② 변수의 선택

- 전진 선택법: 영향력이 높다고 여겨지는 독립변수부터 순서대로 선택함
- 후진 제거법: 영향력이 낮다고 여겨지는 독립변수부터 순서대로 제거함
- 단계별 회귀방법: 전진 선택법에 후진 제거법을 가미한 방법
- 모두 입력: 분석자가 선택하여 변수를 제거

③ 검정통계량(T) 계산

④ 다중공선성 분석: 독립변수들 간의 상관관계 여부 검정

⑤ 더빈-왓슨 검정: 자기상관성 검정

④ 임계값에 의한 의사결정

- 설정한 신뢰구간에 따른 유의수준(e.g. 0.05, 0.01, 0.1)과 유의확률(p값)을 비교
- 유의수준보다 p값이 작을 경우 통계적으로 유의하기에 귀무가설을 기각하고, 대립가설을 채택한다.
- 유의수준보다 p값이 클 경우 통계적으로 유의하지 못하므로 귀무가설을 채택한다.

5) SPSS 절차

- 분석 → 회귀분석 → 선형

5. 분산분석 (일원배치 분산분석)

1) 사용 목적

- 세 집단의 평균을 비교하기 위함

2) 사용 척도

- 요인변수는 명목, 순서척도
- 종속변수는 구간, 비율척도

3) 분석 순서

① 가설 설정

- 귀무가설 H_0: $\mu_1 = \mu_2 = \dots = \mu_k$
- 대립가설 H_1: 적어도 하나의 μ_i는 다르다.

② 분산분석표 출력

③ 다중비교

- 집단 간 어떤 차이가 있는지를 검정

④ 임계값에 의한 의사결정

- 설정한 신뢰구간에 따른 유의수준(e.g. 0.05, 0.01, 0.1)과 유의확률(F통계량)을 비교
- 유의수준보다 유의확률이 작을 경우 통계적으로 유의하기에 귀무가설을 기각하고, 대립가설을 채택한다.
- 유의수준보다 유의확률이 클 경우 통계적으로 유의하지 못하므로 귀무가설을 채택한다.

4) SPSS 절차

- 분석 → 평균비교 → 일원배치 분산분석