데이터과학과 AI를 위한 파이썬

12강. 추정과 검정

세종사이버대학교

김명배 교수

학습내용

- 범주형 변수의 가설 검정
- 연속형 변수의 가설검정
- 피어슨 상관계수

학습목표

- 범주형 변수의 가설 검정 종류와 검정 방법을 설명하고 파이썬으로 구현할 수 있다.
- 연속형 변수의 가설 검정 종류와 검정 방법을 설명하고 파이썬으로 구현할 수 있다.
- 피어슨의 상관분석의 의미를 설명하고 파이썬으로 구현할 수 있다.

[Remind] 가설검정의 절차

1. 귀무가설과 대립가설의 수립	▪ 귀무가설 : 기존의 사실, 기존에 받아들이던 가설 ▪ 대립가설 : 표본을 통해 새롭게 입증하고자 하는 가설
2. 유의수준 설정	▪ 유의수준 : 제 1종 오류(귀무가설이 참인데, 대립가설을 선택하는 오류), 보통 5% 기준으로 사용한다.
3. 통계적 분석 기법의 선택	▪ 독립변수와 종속변수의 척도(범주형 or 연속형)에 따라 확률분포를 적절하게 선택한다.
4. 검정통계량 VS 기각역 유의확률 VS 유의수준	▪ 검정통계량과 기각역 또는 유의확률과 유의수준의 대소관계를 판단한다.
5. 귀무가설 기각 여부 결정	▪ 유의확률이 유의수준(사용자 결정)보다 작거나, ▪ 검정통계량이 기각역보다 크면 귀무가설을 기각한다.
6. 최종 결론 및 의사결정	▪ H_0또는 H_1 기각 여부를 판단하여 최종 의사결정을 한다.

1. 범주형 변수의 가설검정

1) 이항 검정(binomial test)

- 이항분포를 이용한 베르누이 확률변수의 모수(μ)에 대한 검정
- 귀무가설 $H_0 : \mu = \mu_0$

  대립가설
  - 양측검정 → $H_1 : \mu \neq \mu_0$
  - 단측검정 → $H_1 : \mu > \mu_0$ 또는 $H_1 : \mu < \mu_0$

[파이썬] Scipy.stats.binom_test(x, n=None, p=0.5, alternative='two-sided')

x : 검정통계량 1이 나온 횟수

n : 총 시도 횟수

p : 귀무가설의 μ_0값

alternative : 양측검정이면 'two-sided', 단측검정이면 'one-sided'

1. 범주형 변수의 가설검정

2) 카이제곱 검정(chi-squared test)

- 카테고리분포 표본의 합 통계량을 통한 모수(μ_k)에 대한 검정
- 적합도 검정(goodness of fit test)이라고도 함

- 귀무가설 $H_0 : \mu = (\mu_1, \mu_2, \cdots \mu_k)$
  대립가설 $H_1 : \mu \neq (\mu_1, \mu_2, \cdots \mu_k)$

[파이썬] Scipy.stats.chisquare(f_obs, f_exp=None)
f_obs : 데이터 행렬
f_exp : 기댓값 행렬

1. 범주형 변수의 가설검정

3) 독립성 검정(contingency test)

- 범주형 확률변수 X가 다른 범주형 확률변수 Y와 독립인지를 검증
- 행범주와 열범주가 독립인지를 검증(범주형 변수들 간에 상관분석)
- 분할표 검정, 교차분석 이라고도 함
- 귀무가설(H_0) : 두 변수는 상관성이 없다.=서로 독립이다.
  대립가설 (H_1) : 두 변수는 연관성이 있다.=서로 종속이다.

[예시]
- 연령대별로 정당의 선호도에 차이가 있는가?
- 성별에 따라 자동차 색상 선호도의 차이가 있는가?

1. 범주형 변수의 가설검정

3) 독립성 검정(contingency test)

[기본가정]

교차표로 표현하였을 때, 기대빈도가 5보다 작은 셀의 수가 전체 셀의 25% 미만이어야 한다.

[참고] 가정을 만족하지 못하면 피셔의 정확검정(Fisher’s exact test)으로 검정하여야 함

카이제곱 통계량

$$\chi^2 = \sum \frac{(\text{관측빈도} - \text{기대빈도})^2}{\text{기대빈도}}$$

[파이썬] scipy.stats.chi2_conrtingency()

1. 범주형 변수의 가설검정

3) 독립성 검정(contingency test)

- 기대빈도란? 귀무가설 하에 각 셀의 예상되는 빈도

교차표		선호색상		
		검은색	흰색	합계
성별	남자	a	b	a+b
	여자	c	d	c+d
	합계	a+c	b+d	a+b+c+d

Ea : a의 기대빈도 = (a+b)(a+c)/(a+b+c+d)

Eb : b의 기대빈도 = (a+b)(b+d)/(a+b+c+d)

Ec : c의 기대빈도 = (c+d)(a+c)/(a+b+c+d)

Ed : d의 기대빈도 = (c+d)(b+d)/(a+b+c+d)

1. 범주형 변수의 가설검정

3) 독립성 검정(contingency test)

[예시] 두 상표(A,B)의 가전제품이 3년이내에 고장이 발생한 빈도

- **관측 빈도** : 실제 표본에서 관찰된 빈도

교차표		3년 이내 고장 여부		
		有	無	전체
상표	A	16(20%)	64(80%)	80(100%)
	B	48(40%)	72(60%)	120(100%)
	전체	60(32%)	140(68%)	200(100%)

- **기대 빈도** : 두 변수 사이에 연관성이 없다는 가정(H_0)하에 예상되는 빈도

교차표		3년 이내 고장 여부		
		有	無	전체
상표	A	25.6(32%)	554.4(68%)	80(100%)
	B	38.4(32%)	81.6(68%)	120(100%)
	전체	60(32%)	140(68%)	200(100%)

2. 연속형변수의 가설검정

1) 단일 표본 t검정(one-sample t-test)

- 정규분포의 표본에 대해 기댓값을 검정하는 방법
- 단일 집단의 연속형 변수(평균)에 대한 검정 방법

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{N}}$$

- 귀무가설(H_0) : $\mu = \mu_0$
  대립가설(H_1) : $\mu \neq \mu_0$ 또는 $\mu > \mu_0$, $\mu < \mu_0$

2. 연속형변수의 가설검정

1) 단일 표본 t검정(one-sample t-test)

[예시]

- 어떤 고등학교 A반의 수학 평균점수는 70점인가?
- 고등학생 남자의 평균신장은 175cm인가?
- 회귀계수값이 0인가?

[파이썬] scipy.stats.ttest_1samp(a, popmean)

- a : 표본 데이터 배열
- popmean : 귀무가설의 기댓값(평균값)

2. 연속형변수의 가설검정

2) 독립 표본 t검정(independent two-sample t-test)

- 독립적인 정규분포에서 나온 N_1, N_2개 데이터를 이용하여 두 정규분포의 기댓값(평균)이 같은지를 검정하는 방법
- 범주형 변수에 따른 연속형 변수의 차이 검정

- 귀무가설(H_0) : $\mu_1 = \mu_2$

  대립가설(H_1) : $\mu_1 \neq \mu_2$ 또는 $\mu_1 > \mu_2$, $\mu_1 < \mu_2$

2. 연속형변수의 가설검정

2) 독립 표본 t검정(independent two-sample t-test)

[예시]

- 어떤 학교에 A반과 B반의 수학 평균점수는 차이가 있는가?
- 성별에 따라 수능점수 평균에 차이가 있는가?
- 연령대에 따라 소득금액 평균에 차이가 있는가?

※ 변수역할에 따른 분류

A/B반 → 수학점수
성별 → 수능점수
연령대 → 소득금액

독립변수→종속변수
or 설명변수→ 목표변수
or 처리→반응변수
or predictor→ Target

2. 연속형변수의 가설검정

2) 독립 표본 t검정(independent two-sample t-test)

[기본 가정]

가) 독립성

- 두 그룹은 서로 독립이다.

나) 정규성

- 종속변수는 집단별로 각각 정규분포를 따른다.

다) 등분산성

- 두 집단간 종속변수의 분산이 같다.

2. 연속형변수의 가설검정

2) 독립 표본 t검정(independent two-sample t-test)

- 두 집단의 분산이 같은지에 따라 통계량이 다름

가) 분산이 같은 경우(등분산), $t = \frac{\bar{x}_1 - \bar{x}_2}{s \cdot \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$

나) 분산이 다른 경우(이분산), $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$

[파이썬] scipy.stats.ttest_inds(a, b, equal_var=True)

- a : 첫 번째 그룹 데이터 배열, b: 두 번째 그룹 데이터 배열
- equal_var : 등분산 여부(같으면 True, 다르면 False)

2. 연속형변수의 가설검정

3) 대응 표본 t검정(paired t-test)

- 실험 이전의 집단과 실험 이후의 집단이 동일한 집단인 경우 사용하는 검정
- 짝을 지어 처리(실험) 이외의 영향 인자들을 통제하고 비교(짝지은 t검정)

- 귀무가설(H_0) : $\mu_{pre} = \mu_{post}$ 또는 $\mu_{post} - \mu_{pre} = 0$
  대립가설(H_1) : $\mu_{pre} \neq \mu_{post}$ 또는 $\mu_{pre} < \mu_{post}$, $\mu_{pre} > \mu_{post}$
  $\mu_{pre} - \mu_{post} \neq 0$ 또는 $\mu_{post} - \mu_{pre} < 0$, $\mu_{post} - \mu_{pre} > 0$

$$t = \frac{\bar{x}_d - \mu_0}{s_d/\sqrt{N}}, \quad x_d = x_{i,pre} - x_{i,post}, \ i = 1, 2, \cdots N$$

2. 연속형변수의 가설검정

3) 대응 표본 t검정(paired t-test)

[예시]

- 동영상 교육자료를 수강하기 전과 수강한 이후에 시험점수가 차이가 있는가?
- 약물치료 전과 후의 콜레스테롤 농도의 차이가 있는가?
- A제품과 B제품 쌍의 타이어 마모도에 차이가 있는가?

→ 차이 비교에 영향을 주는 인자를 통제하기 위해 동일한 개체에서 전과 후의 값을 비교함

[파이썬] scipy.stats.ttest_rel(a, b)

- a : 1번 표본 집합 데이터
- b : 2번 표본 집합 데이터

2. 연속형변수의 가설검정

4) 등분산성 검정(equal-variance test)

- 두 정규분포를 따르는 데이터로부터 두 정규분포의 분산이 같은지 확인하는 검정
- 바틀렛(bartlett), 플리그너(fligner), 레빈(levene) 검정이 있음
- 결과가 서로 다를 수 있음
- 귀무가설(H_0) : $\sigma_1^2 = \sigma_2^2$

  대립가설(H_1) : $\sigma_1^2 \neq \sigma_2^2$

[파이썬] scipy.stats.bartlett(x1, x2), fligner(), levene()

2. 연속형변수의 가설검정

5) 정규성 검정(normality test)

- 정규분포를 따른다고 할 수 있는지에 대한 검정
- 모수적 통계분석 방법에서는 정규성을 기본 가정으로 하는 경우가 많음
- 콜모고로프-스미르노프 검정(Kolmogorov-Smirnov test)와 샤피로-윌크 검정(Shapiro-Wilk test)을 가장 많이 사용함
- 귀무가설(H_0) : 정규분포와 차이가 없다. or 정규분포를 따른다.
  대립가설(H_1) : 정규분포와 차이가 있다. or 정규분포를 따르지 않는다.

[파이썬] 콜모고로프-스미르노프 검정 : scipy.stats.ks_2samp()
샤피로-윌크 검정(Shapiro-Wilk test) : scipy.stats.shapiro()

3. 피어슨 상관계수

1) 피어슨의 상관분석(Peason's correlation analysis)

- 2개의 연속형 변수 간에 선형적인 상관성이 있는지를 검증하고 상관계수를 산출하는 분석
- 데이터 탐색(산점도)을 통해 연속형 두 변수 간에는 선형적인 관계가 있는지 판단 필요
- 두 변수 모두 정규분포를 따라야 함

$$\text{상관계수}(r) = \frac{\sum(x_1 - \bar{x})(y_1 - \bar{y})}{\sqrt{\sum x_1 - \bar{x} \sum(y_1 - \bar{y})^2}}$$

- 귀무가설 (H_0) : 두 변수 간에는 (선형적인) 관계가 없다.
  대립가설 (H_1) : 두 변수 간에는 (선형적인) 관계가 있다.

3. 피어슨 상관계수

1) 피어슨의 상관분석(Peason's correlation analysis)

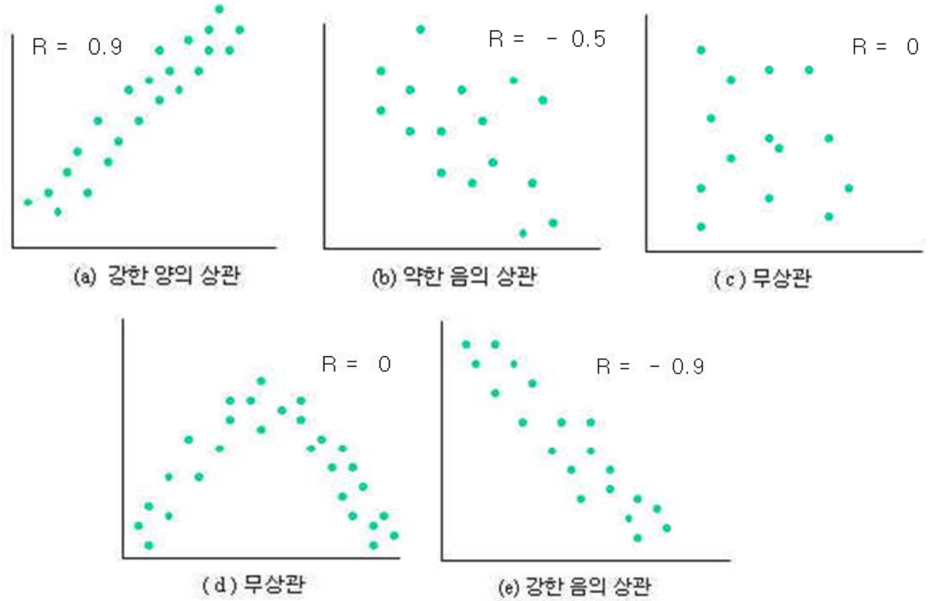
[예시]

- 혈중 중성지방 수치와 콜레스테롤 수치에 선형적인 관련성이 있는가?
- 간 기능 수치들 간의 선형적인 관련성이 있는가?
- 연령과 소득에 선형적인 관계가 있는가?

[파이썬] corr()

3. 피어슨 상관계수

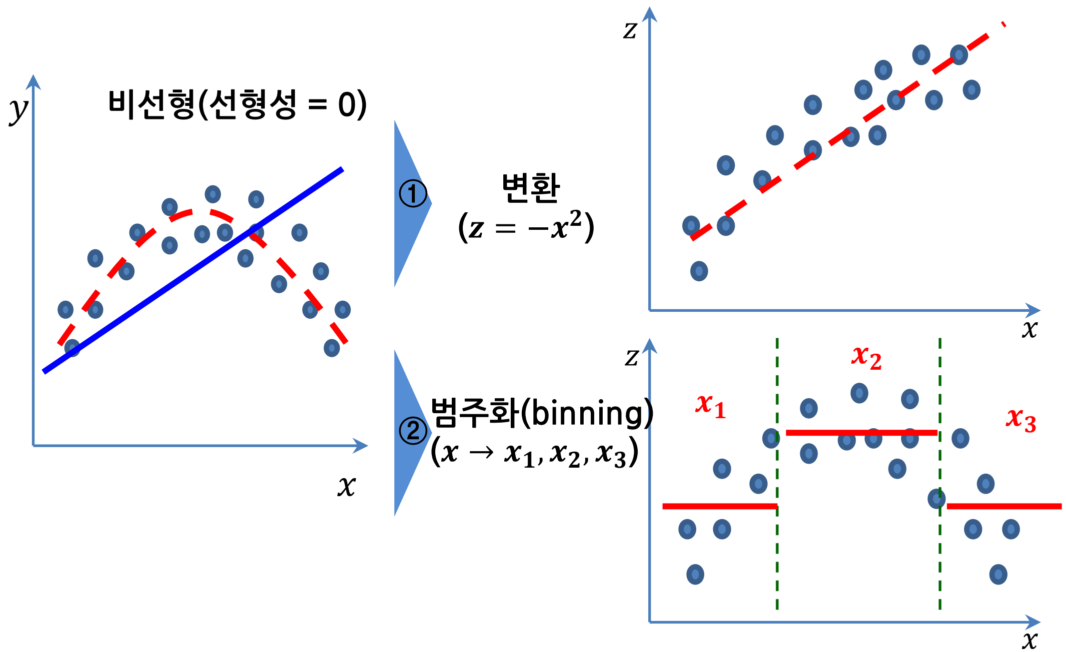
2) 상관계수의 의미



3. 피어슨 상관계수

3) 선형성 변환

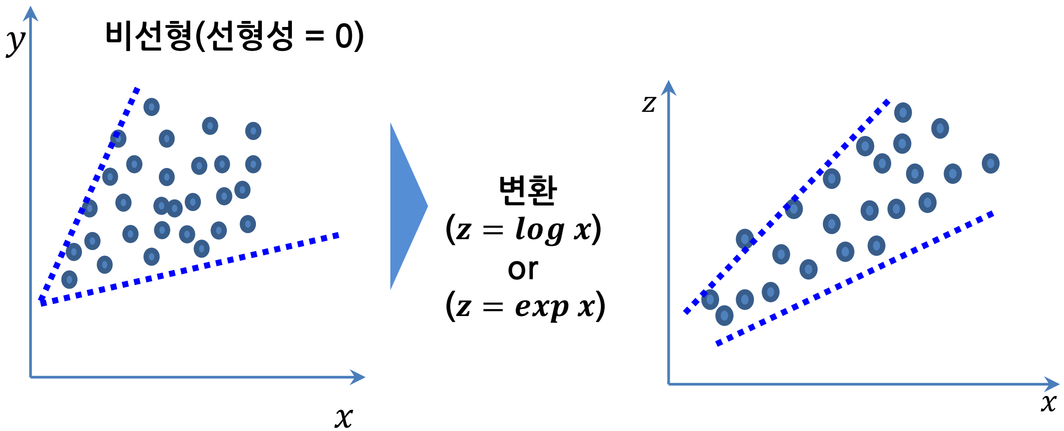
- 비선형 함수형태의 관계를 특정 변환하여 선형성을 강하게 할 수 있음



3. 피어슨 상관계수

3) 선형성 변환

- 부채꼴인 관계를 로그 또는 지수 변환을 통해 퍼짐정도를 줄일 수 있음



정리하기

1. 범주형 변수의 가설검정

가) 이항분포를 이용한 베르누이 확률변수의 모수(μ)에 대한 검정 → 이항검정

$H_0 : \mu = \mu_0,\ H_1 : \mu \neq \mu_0$

[파이썬] scipy.stats.binom_test()

나) 카테고리분포 표본의 합 통계량을 통한 모수(μ_k)에 대한 검정 → 카이제곱 검정

$H_0 : \mu = (\mu_1, \mu_2, \cdots \mu_k),\ H_1 : \mu \neq (\mu_1, \mu_2, \cdots \mu_k)$

[파이썬]scipy.stats.chisquare()

다) 행범주와 열범주가 독립인지를 검증(범주형 변수들 간에 상관분석) → 카이제곱 검정

H_0 : 두 변수는 상관성이 없다.=서로 독립이다.

H_1 : 두 변수는 연관성이 있다.=서로 종속이다.

[파이썬] scipy.stats.chi2_conrtingency()

정리하기

2. 연속형 변수의 가설검정

가) 단일 집단의 연속형 변수(평균)에 대한 검정 방법 → 단일 표본 t검정

$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$ 또는 $\mu > \mu_0, \mu < \mu_0$

[파이썬] scipy.stats.ttest_1samp()

나) 범주형 변수에 따른 연속형 변수의 차이 검정 → 독립 표본 t검정

$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$ 또는 $\mu_1 > \mu_2, \mu_1 < \mu_2$

[파이썬] scipy.stats.ttest_inds()

다) 실험 이전의 집단과 실험 이후의 집단이 동일한 집단인 경우 사용하는 검정 → 대응 표본 t검정

$H_0 : \mu_{pre} = \mu_{post}$ 또는 $\mu_{post} - \mu_{pre} = 0$

$H_1 : \mu_{pre} \neq \mu_{post}$ 또는 $\mu_{pre} < \mu_{post}, \mu_{pre} > \mu_{post}$

$\mu_{pre} - \mu_{post} \neq 0$ 또는 $\mu_{post} - \mu_{pre} < 0, \mu_{post} - \mu_{pre} > 0$

[파이썬] scipy.stats.ttest_rel()

정리하기

2. 연속형 변수의 가설검정

라) 두 정규분포를 따르는 데이터로부터 두 정규분포의 분산이 같은지 확인하는 검정 → 등분산성 검정

H_0: $\sigma_1^2 = \sigma_2^2$ H_1 : $\sigma_1^2 \neq \sigma_2^2$

[파이썬] scipy.stats.bartlett(x1, x2), fligner(), levene()

마) 정규분포를 따른다고 할 수 있는지에 대한 검정 → 정규성 검정

H_0 : 정규분포와 차이가 없다. or 정규분포를 따른다.

H_1 : 정규분포와 차이가 있다. or 정규분포를 따르지 않는다.

[파이썬] 콜모고로프-스미르노프 검정 : scipy.stats.ks_2samp()

샤피로-윌크 검정(Shapiro-Wilk test) : scipy.stats.shapiro()