Chapter 12
Hypothesis Test II

경영대학 재무금융학과
윤선중

Objectives

- 한 모집단의 특성를 파악하는 문제
  - 구간데이터
    - 추정 모수: 모평균, 모분산
    - Student t분포: 모표준편차가 알려져 있지 않을 때 모평균 검정 및 추정
    - 카이제곱분포: 모분산에 관한 추론에 사용
  - 범주데이터
    - 추정 모수: 모비율 p
    - 표본비율은 근사적으로 정규분포
    - 모비율에 대한 검정 통계량과 구간추정량 만듬

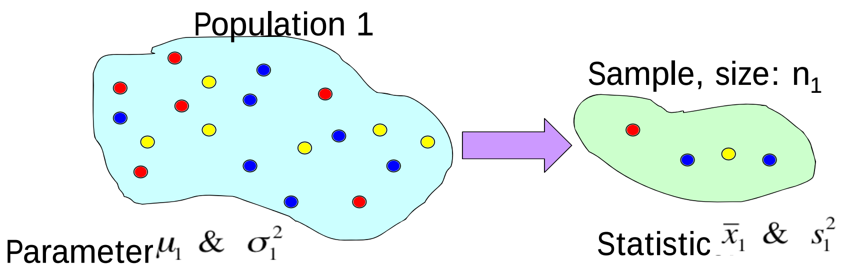
Objectives (2)

■ 두 집단을 비교하는 통계기법

- 구간데이터
  - 중심위치의 척도비교
    - 독립표본: 동분산 공식 or 이분산 공식
    - 짝진표본: 한가지 공식 적용
    - F 통계량
  - 모분산의 척도 비교
- 범주데이터
  - 두 모비율의 차이
    - 두가지 검정 통계량과 한가지 구간추정량 제공
    - 관측데이터와 실험데이터

Introduction

■ 두 모집단의 평균 비교



- 모집단 1과 유사하게 모집단 2에 대한 표본크기, 통계량, 모수들이 존재

■ 경우의 수

- (1) 두 모집단의 분산을 아는 경우 vs. 분산을 모르는 경우
- (2) 정규분포를 따르는 경우 vs. 그렇지 않음 경우
- (3) 표본의 크기가 큰 경우 vs. 그렇지 않은 경우

Properties of Difference in Two Samples

- $\bar{x}_1 - \bar{x}_2$ 가 정규분포를 따르면, 정규분포를 따르고, 모집단이 정규분포를 따르지 않지만 표본크기가 크면 근사적으로 정규분포!
- 기대값 $E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$
- 분산 $$V(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$
- 표준편차 $$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Properties of Difference in Two Samples

- 표준정규확률변수와 구간추정량

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- 모분산이 알려져 있지 않지만 그 크기가 같을 때,

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \ \nu = n_1 + n_2 - 2$$

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \ \nu = n_1 + n_2 - 2$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Independent Sample

- **Example**
  - A: 아침에 시리얼을 먹는 사람
  - B: 아침에 시리얼을 먹지 않는 사람
  - A보다 B가 점심식사 때, 더 많은 칼로리를 섭취한다는 주장을 5% 유의수준에서 검정하시오. 단, A와 B의 분산은 각각 2900, 3500인 것으로 알려져 있으며, A와 B의 점심 칼로리 섭취량은 정규분포를 따른다고 가정
- **Step 1:** 가설 설정
  - $$H_0 : \mu_A > \mu_B \Rightarrow \mu_A - \mu_B > 0$$ $$H_1 : \mu_A \le \mu_B \Rightarrow \mu_A - \mu_B \le 0$$
- **Step 2:** 검정통계량 산출
  - 각각 100명 씩의 sample을 조사한 결과, A는 평균 600칼로리, B는 평균 610 칼로리를 섭취하는 것으로 조사되었다고 함
  - 귀무가설을 기각? 귀무가설을 기각 못함?

Independent Sample

- **Step 3:** 가설검정의 결론 (의사결정)
  - 귀무가설이 옳다면,

$$\bar{X}_A - \bar{X}_B \sim N\left(\mu_A - \mu_B, \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}\right) \sim N\left(0, \frac{2900}{100} + \frac{3500}{100}\right) \sim N(0, 8^2)$$

  - 두 표본평균의 차이가 -10이라는 관찰에 근거하여 귀무가설을 기각하였을 때, 잘못된 오류를 범할 확률은,

$$P(\bar{X}_A - \bar{X}_B \le -10) = P\left(\frac{\bar{X}_A - \bar{X}_B - 0}{8} \le \frac{-10 - 0}{8}\right) = P(Z \le -1.25) = 0.056$$

  - 0.1056는 허용된 오류확률 5%보다 크므로, 귀무가설을 ________

Independent Sample

- 만약 표본평균의 파이가 얼마 이하이면, 5% 유의수준에서 귀무가설을 기각할 수 있는가?

$$P(\bar{X}_A - \bar{X}_B \le X_H) = P\left(\frac{\bar{X}_A - \bar{X}_B - 0}{8} \le \frac{X_H - 0}{8}\right) = P(Z \le -1.645) = 0.05$$

$$\therefore X_H = -1.645 \times 8 + 0 = -13.16$$

- 앞 예제에서, 표본평균의 차이가 -10이므로 임계치를 벗어나지 못함
- => 따라서 5% 유의수준에서 귀무가설을 ________
- =>위 예에서 기각역은 {표본평균의 차이 < -13.16}

Independent Sample (Review)

- **Case 1**
  - 두 모집단의 모분산이 모두 알려져 있는 경우

$$\bar{X}_A - \bar{X}_B \sim N\left(\mu_A - \mu_B, \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}\right)$$

  - (1) 정규분포를 따르면, 표본평균의 차이도 정규분포
  - (2) 정규분포를 따르는지 모르면, 표본의 크기가 큰 경우에만 표본평균의 차이도 정규분포

- **Case 2**
  - 두 모집단의 분산이 알려져 있지 않지만, 동일한 경우 (등분산)

$$\bar{X}_A - \bar{X}_B \sim N\left(\mu_A - \mu_B, s_p^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)\right), \quad \text{단,} \quad s_p^2 = \frac{(n_A - 1)s_1^2 + (n_B - 1)s_2^2}{n_A + n_B - 2}$$

  - 두 모집단이 정규분포를 따르더라도, 표본평균의 차이는 student-t분포 (자유도 n1+n2-2)를 따름

Independent Sample

- **Case 3**
  - 두 모집단의 분산이 알려져 있지 않으며, 동일하지 않은 경우 (이분산)

$$\bar{X}_A - \bar{X}_B \sim N\left(\mu_A - \mu_B, \frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)$$

  - 두 모집단이 정규분포를 따르더라도: 표본평균의 차이는 student-t 분포를 따르며, 자유도는 아래와 같음 (가장 가까운 정수)

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}, \quad \nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Testing the Population Variances

- 동분산과 이분산 검정중 어떤 것을 이용할 것인가?
  - 분산의 동일성에 대한 검정 필요!!!

H_0: $\sigma_1^2 / \sigma_2^2 = 1$

H_1: $\sigma_1^2 / \sigma_2^2 \neq 1$

  - s_1^2 / s_2^2 는 F-분포를 따르며 자유도는 ν1 = n1– 1 and ν2 = n2 –1.
  - 동분산에 대한 기각 영역 (rejection region)

$$F > F_{\alpha/2,\nu_1,\nu_2} \qquad F < F_{1-\alpha/2,\nu_1,\nu_2}$$

Exercise I

- 개요 (Xm 12-01: 예제 12.1) 직접 구매한 뮤추얼펀드와 브로커를 통해 구매한 뮤추얼 펀드

$H_1: \mu_1 - \mu_2 > 0 \qquad H_0: \mu_1 - \mu_2 = 0$

- **F-test of σ_1^2 / σ_2^2**

$s_1^2 = 37.49$ and $s_2^2 = 43.34$

Test statistic: $F = 37.49/43.34 = 0.86$

$$F > F_{\alpha/2,\nu_1,\nu_2} = F_{.025,49,49} \approx F_{.025,50,50} = 1.60$$

$$F < F_{1-\alpha/2,\nu_1,\nu_2} = F_{.975,49,49} = 1/F_{.025,49,49} \approx 1/F_{.025,50,50} = 1/1.60 = .63$$

따라서 동분산검정 가능!!!

Exercise II

- 개요 (Xm 12-02: 예제 12.2) 가족경영의 새로운 CEO 효과

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 \neq 0$

- $s_1^2 = 3.79$ and $s_2^2 = 8.03$
- $F = 3.79/8.03 = 0.47$
- 기각영역 $F > F_{\alpha/2,\nu_1,\nu_2} = F_{.025,41,97} \approx F_{.025,40,100} = 1.54$
- 이분산 검정 사용!!!

Exercise 3

- 개요 (Xr 12-15, 연습문제 12-15)
  - 성별에 따른 운전거리의 차이를 조사 (남성=1, 여성=2)
  - 남성 운전자와 여성운전자의 연간 운전거리가 다르다는 주장을 5% 유의 수준에서 검정하시오
  - 남성운전자와 여성운전자의 평균 운전거리 차이에 대한 95% 신뢰구간을 추정하시오

Matched Pairs Sample

- 짝진 실험
  - 두 표본의 관측치들이 서로 상대표본의 한 관측치와 짝이 되게끔 설계
- 예제
  - Xm 12-4; 예제 12-4
  - MBA 금융전공자와 마케팅 전공자 간의 급여 비교
  - 금융 전공자들이 더 많은 연봉을 받는다는 주장을 5% 유의수준에서 검정하시오
    - 단, 우선 독립표본으로 가정하여 검정
- Step 1: 가설설정

$$H_0 : \mu_F = \mu_M \Rightarrow \mu_F - \mu_M = 0$$

$$H_1 : \mu_F > \mu_M \Rightarrow \mu_F - \mu_M > 0$$

- Step 2: 검정 통계량 산출
  - 엑셀 파일 Xm 12-4에서 조사

Matched Pairs Sample

- **Step 3: 가설검정의 결론 (의사결정)**
  - 귀무가설이 옳다면,

$$\bar{X}_F - \bar{X}_M \sim \left(\mu_F - \mu_M, s_p^2 \left(\frac{1}{n_F} + \frac{1}{n_M} \right) \right) \sim \left(0, 4991^2 \right)$$

  - 두 표본평균의 차이가 5,201이라는 관찰에 근거하여 귀무가설을 기각하였을 때, 잘못된 오류를 범할 확률은

$$P(\bar{X}_F - \bar{X}_M > 5201) = P\left(\frac{\bar{X}_F - \bar{X}_M - 0}{4991} \leq \frac{5201 - 0}{4991} \right) = P(t > 1.04) = 0.1503$$

  - 0.1513은 허용된 오류확률 5%보다 크므로, 귀무가설을 ________
  - Excel을 통한 실습: 도구-데이터분석-t검정 이분산 가정 두집단

Matched Pairs Sample

- 결과

	A	B	C
1	t-Test: Two-Sample Assuming Equal Variances		
2			
3		*Finance*	*Marketing*
4	Mean	65,624	60,423
5	Variance	360,433,294	262,228,559
6	Observations	25	25
7	Pooled Variance	311,330,926	
8	Hypothesized Mean Difference	0	
9	df	48	
10	t Stat	1.04	
11	P(T<=t) one-tail	0.1513	
12	t Critical one-tail	1.6772	
13	P(T<=t) two-tail	0.3026	
14	t Critical two-tail	2.0106	

Matched Pairs Sample

- 예제 Xm 12-5;MBA 재무와 마케팅 전공자 비교 PART 2
  - 단, 각 GPA에 따라 짝을 정하여 상호 비교 (즉, 짝진표본에서 검정)
  - 금융 전공자들이 더 많은 연봉을 받는다는 주장을 5% 유의수준에서 검정하시오
- Step 1: 가설 설정
  - $H_0 : \mu_D = 0$ (단, $\boxed{\mu_1 - \mu_2 = \mu_D}$
  - $H_1 : \mu_D > 0$
- Step 2: 검정통계량 산출
  - 엑셀 파일 Xm 12-5에서 조사

Matched Pairs Sample

- Step 3: 가설검정의 결론 (의사결정)
  - 귀무가설이 옳다면,

$$t = \frac{\bar{x}_D - \mu_D}{s_D / \sqrt{n_D}} \qquad \bar{X}_D = \left(\mu_D, \frac{s_D^2}{n_D} \right) \sim (0, 1329^2)$$

  - 두 표본평균의 차이가 5,065이라는 관찰에 근거하여 귀무가설을 기각하였을 때, 잘못된 오류를 범할 확률은

$$P(\bar{X}_D > 5065) = P\left(\frac{\bar{X}_D - 0}{1329} \leq \frac{5065 - 0}{1329} \right) = P(t > 3.81) = 0.0004$$

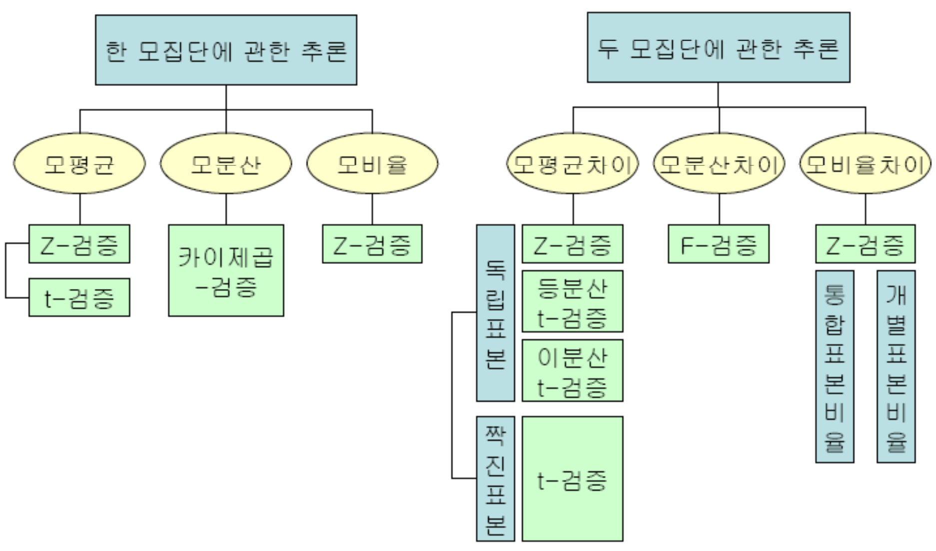
  - 0.0004은 허용된 오류확률 5%보다 작으므로, 귀무가설을 ________

Matched Pairs Sample

- 독립표본 vs. 짝진표본
  - 짝진표본을 설계할 때 짝을 짓는 기준이 의미있는 변수인 경우
    - 독립표본보다 짝진표본에서 표준오차가 줄어드는 경향이 존재함
    - 짝진표본의 검정력이 더 우수할 수 있음
  - 짝진표본에서는 자유도가 감소
    - 검정력이 항상 독립표본보다 우수하다고 장담할 수 없음
  - 따라서,
    - (1) 표본집단 1과 표본집단 2의 각 관측치들 사이에 특정한 관계가 존재한다라고 하면, _______표본에서의추론을 적용하는 것이 바람직
    - (2) 그렇지 않은 경우에는 _______ 표본에서의 추론을 적용하는 것이 바람직

Summary

- 요약



- 유의
  - 모 분산을 모르는 경우의 t-검증
    - 엄밀하게는 모집단이 정규분포여야 하지만, 모집단이 극단적인 비정규분포인 경우를 제외하고는 대부분 활용됨
    - 표본의 크기가 아주 큰 경우에는 Z검증으로 대체될 수 있음