12강

통계적 비교 I

통계·데이터과학과 장영재 교수

한국방송통신대학교

목차

01

두 모집단의 비교

두 모집단의 비교 사례

- 제품 A를 사용한 집단과 B를 사용한 집단 간 선호도 차이는 있을까
- 두 생산 라인에서 생산되는 제품 간 수율 차이는 있을까
- 어느 직장의 직무연수가 연수 이전에 비해 직원들의 직무능력을 향상시켰는가

➡ 각 모집단의 특성을 나타내는 값, 평균을 고려한다면
두 모집단의 비교는 모평균의 비교 문제로 귀결

독립표본 시 두 모집단의 비교

- 두 모집단의 모평균 μ_1, μ_2 두 모집단의 차이의 비교 기준 값 δ_0일 때, 세 가지 가설

① $H_0 : \mu_1 - \mu_2 = \delta_0$

$H_1 : \mu_1 - \mu_2 > \delta_0$

② $H_0 : \mu_1 - \mu_2 = \delta_0$

$H_1 : \mu_1 - \mu_2 < \delta_0$

③ $H_0 : \mu_1 - \mu_2 = \delta_0$

$H_1 : \mu_1 - \mu_2 \neq \delta_0$

독립표본 시 두 모집단의 비교

- <u>표본 수가 충분히 큰 경우(통상 30보다 큰 경우)</u>에는 모집단의 분포와 관계 없이 다음과 같은 검정통계량을 산출하고 표준정규분포를 이용하여 검정

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

독립표본 시 두 모집단의 비교

- 표본 수가 작은 경우: 정규분포를 따르고 두 집단의 모분산이 서로 같다면, 다음 검정통계량을 산출하고 t분포를 이용하여 검정

가설의 종류	선택기준
① $H_0: \mu_1 - \mu_2 = \delta_0$ $H_1: \mu_1 - \mu_2 > \delta_0$	$\dfrac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\dfrac{S_p^2}{n_1} + \dfrac{S_p^2}{n_2}}} > t_{n_1+n_2-2,\,\alpha}$ 이면 H_0 기각
② $H_0: \mu_1 - \mu_2 = \delta_0$ $H_1: \mu_1 - \mu_2 < \delta_0$	$\dfrac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\dfrac{S_p^2}{n_1} + \dfrac{S_p^2}{n_2}}} < -t_{n_1+n_2-2,\,\alpha}$ 이면 H_0 기각
③ $H_0: \mu_1 - \mu_2 = \delta_0$ $H_1: \mu_1 - \mu_2 \neq \delta_0$	$\left\lvert \dfrac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\dfrac{S_p^2}{n_1} + \dfrac{S_p^2}{n_2}}} \right\rvert > t_{n_1+n_2-2,\,\alpha/2}$ 이면 H_0 기각

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

독립표본 시 두 모집단의 비교

- <u>표본 수가 작은 경우</u>: 정규분포를 따르고 두 집단의 모분산이 서로 다를 때에는 t 분포의 자유도를 ϕ로 수정[새터스웨이트(Satterthwaite) 근사]
- 검정통계량과 자유도를 계산하고 앞의 표를 이용하여 검정

$$T = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$\phi = \frac{\left[\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right]^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

대응표본 시 두 모집단의 비교

- 짝지어진 n쌍(pair)의 표본의 차를 계산하여 단일표본의 검정문제로 단순화

모집단 1의 표본(X_{i1})	모집단 2의 표본(X_{i2})	$D_i = X_{i1} - X_{i2}$
X_{11}	X_{12}	$D_1 = X_{11} - X_{12}$
X_{21}	X_{22}	$D_2 = X_{21} - X_{22}$
$\vdots$	$\vdots$	$\vdots$
X_{n1}	X_{n2}	$D_n = X_{n1} - X_{n2}$

D_i의 평균 $\overline{D} = \sum D_i / n$

D_i의 분산 $s_D^2 = \sum (D_i - \overline{D})^2 / (n-1)$

대응표본 시 두 모집단의 비교

가설의 종류	선택기준
① $H_0: \mu_1 - \mu_2 = D_0$ $H_1: \mu_1 - \mu_2 > D_0$	$\dfrac{\bar{D} - D_0}{\dfrac{S_D}{\sqrt{n}}} > t_{n-1,\,\alpha}$이면 H_0 기각
② $H_0: \mu_1 - \mu_2 = D_0$ $H_1: \mu_1 - \mu_2 < D_0$	$\dfrac{\bar{D} - D_0}{\dfrac{S_D}{\sqrt{n}}} < -t_{n-1,\,\alpha}$이면 H_0 기각
③ $H_0: \mu_1 - \mu_2 = D_0$ $H_1: \mu_1 - \mu_2 \neq D_0$	$\left\lvert \dfrac{\bar{D} - D_0}{\dfrac{S_D}{\sqrt{n}}} \right\rvert > t_{n-1,\,\alpha/2}$이면 H_0 기각

두 모분산의 비교

- 두 모집단의 모분산 σ_1^2과 σ_2^2, 각 모집단에서 추출한 크기 n_1, n_2개의 독립표본의 표본분산 각각 S_1^2과 S_2^2 라 할 때,
- 검정통계량 $F = \left(\dfrac{S_1^2}{\sigma_1^2}\right) \Big/ \left(\dfrac{S_2^2}{\sigma_2^2}\right)$ 는 두 모분산이 같다는 귀무가설 하에서 자유도$(n_1 - 1, n_2 - 1)$인 F 분포를 따르므로 아래와 같이 검정

가설의 종류	선택기준
$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 \neq \sigma_2^2$	$\dfrac{S_1^2}{S_2^2} > F_{n_1-1,\, n_2-1,\, \alpha/2}$ 또는 $\dfrac{S_1^2}{S_2^2} < F_{n_1-1,\, n_2-1,\, 1-\alpha/2}$ 이면 H_0 기각

02

다수 모집단의 비교 1

분산분석의 원리

- 3개 이상 모집단의 비교 : 두 모집단의 비교 중 독립표본의 표본평균을 이용한 모평균 비교의 확장 → 분산분석(Analysis of Variance)
- 분산분석이란 반응값의 변동을 제곱합(sum of square)으로 나타내고, 이것을 실험과 관련된 요인의 제곱합과 오차의 제곱합으로 분해하여 오차에 비해 영향이 큰 요인이 무엇인가를 찾아내는 분석방법

$$S_T = S_A + S_E$$

- 각 모집단(요인의 수준)의 분포가 정규분포이고, 개별 관측값은 서로 독립이며, 각 모집단의 분산이 동일하다는 가정이 필요

일원배치법 데이터의 배열

- 일원배치법은 여타 조건이 동일할 때, 어느 하나의 요인이 반응값에 영향을 주는지 파악할 수 있는 실험계획법 (랜덤화)

구분	요인의 수준				
	A_1	A_2	$\cdots$	A_l	
실험의 반복	x_{11} x_{12} $\vdots$ x_{1m}	x_{21} x_{22} $\vdots$ x_{2m}	$\cdots$ $\cdots$ $\vdots$ $\cdots$	x_{l1} x_{l2} x_{lm}	
합계	$T_{1\cdot}$	$T_{2\cdot}$	$\cdots$	$T_{l\cdot}$	T
평균	$\bar{x}_{1\cdot}$	$\bar{x}_{2\cdot}$	$\cdots$	$\bar{x}_{l\cdot}$	$\bar{\bar{x}}$

$$T_{i\cdot} = \sum_{j=1}^{m} x_{ij} \quad \overline{x_{i\cdot}} = \frac{T_{i\cdot}}{m} \quad (i = 1,\ 2,\ \cdots,\ l)$$

$$T = \sum_{i=1}^{l} T_{i\cdot} \quad \bar{\bar{x}} = \frac{T}{lm}$$

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

ε_{ij}는 $N(0,\ \sigma_E^2)$에 따르고 서로 독립

$$\left(\text{단, } \sum \alpha_i = 0 \right)$$

일원배치법에서의 검정

일원배치법의 분산분석표

요인	제곱합	자유도	평균제곱	F
A	$S_A=\sum_{i=1}^{l}\sum_{j=1}^{m}(\bar{x}_{i.}-\bar{\bar{x}})^2$	$\phi_A=l-1$	$V_A=\dfrac{S_A}{\phi_A}$	$F=\dfrac{V_A}{V_E}$
E	$S_E=S_T-S_A$	$\phi_E=l(m-1)$	$V_E=\dfrac{S_E}{\phi_E}$	
T	$S_T=\sum_{i=1}^{l}\sum_{j=1}^{m}(x_{ij}-\bar{\bar{x}})^2$	$\phi_T=lm-1$		

- 가설

$$\begin{cases} H_0: \mu_1=\mu_2=\cdots=\mu_l \\ H_1: \mu_i\text{가 모두 같지는 않다.} \end{cases} \text{ or } \begin{cases} H_0: \alpha_1=\alpha_2=\cdots=\alpha_l=0 \\ H_1: \alpha_i\text{가 모두 0은 아니다.} \end{cases}$$

- 검정

F가 $F(\phi_A, \phi_E\,;\,\alpha)$ 보다 크면 유의수준 α에서 귀무가설 기각

일원배치법에서의 추정

- 각 수준의 모평균의 추정 : $\boldsymbol{\mu_i}$ 의 100($\mathbf{1} - \boldsymbol{\alpha}$)% 신뢰구간

$$\widehat{\mu_i} = \bar{x}_{i\cdot} = \sum_{j=1}^{m} x_{ij} / m$$

$$Var(\bar{x}_{i\cdot}) = Var\left(\sum_{i=1}^{m} x_{ij}/m\right) = m\,Var(x_{i1})/m^2 = \frac{\sigma_E^2}{m}$$

$$\bar{x}_{i\cdot} \pm t\left(\phi_E ; \frac{\alpha}{2}\right)\sqrt{\frac{V_E}{m}}$$

- 각 수준의 모평균 차의 추정 : $\boldsymbol{\mu_i} - \boldsymbol{\mu_{i'}}$의 100($\mathbf{1} - \boldsymbol{\alpha}$)% 신뢰구간

$$(\bar{x}_{i\cdot} - \bar{x}_{i'\cdot}) \pm t\left(\phi_E ; \frac{\alpha}{2}\right)\sqrt{\frac{2V_E}{m}}$$

* $\boldsymbol{t}$ 분포 자유도 $\emptyset_E = \boldsymbol{l(m-1)}$

일원배치법에서의 다중비교

- 다중비교란 분산분석에서 F검정을 통해 귀무가설이 기각되었음을 확인한 이후, 어느 수준에서 평균이 차이 나는지 비교하는 방법
- 아래는 두 수준 $A_i, A_{i'}$의 모평균이 유의수준 α에서 유의한 차이가 있음을 의미

$$|\bar{x}_{i\cdot} - \bar{x}_{i'\cdot}| \geq t\left(\phi_E; \frac{\alpha}{2}\right)\sqrt{\frac{2V_E}{m}}$$

- 위에서 기준이 되는 우변의 값을 LSD(Least Significant Difference; 최소유의차)라고 함

  → LSD를 구하고 각 두 수준 조합 간의 표본평균 차이를 구하여 비교(피셔의 LSD 방법)

03

다수 모집단의 비교 2

이원배치법 데이터의 배열

- 이원배치법은 관심 대상인 요인이 2개 존재하여 이 두 요인을 동시에 고려하여 행하는 실험계획법 ($S_T = S_A + S_B + S_E$)

요인 B \ 요인 A	A_1	A_2	$\cdots$	A_l	합	평균
B_1	x_{11}	x_{21}	$\cdots$	x_{l1}	$T_{\cdot 1}$	$\bar{x}_{\cdot 1}$
B_2	x_{12}	x_{22}	$\cdots$	x_{l2}	$T_{\cdot 2}$	$\bar{x}_{\cdot 2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
B_m	x_{1m}	x_{2m}	$\cdots$	x_{lm}	$T_{\cdot m}$	$\bar{x}_{\cdot m}$
합	$T_{1\cdot}$	$T_{2\cdot}$	$\cdots$	$T_{l\cdot}$	T	
평균	$\bar{x}_{1\cdot}$	$\bar{x}_{2\cdot}$	$\cdots$	$\bar{x}_{l\cdot}$		$\bar{\bar{x}}$

$$x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

$\varepsilon_{ij} \sim N(0, \sigma_E^2)$이고 서로 독립 $\sum \alpha_i = 0, \ \sum \beta_j = 0$

$$i = 1, 2, \cdots, l$$

$$j = 1, 2, \cdots, m$$

이원배치법에서의 검정

이원배치법의 분산분석표(반복이 없는 경우)

요인	S	ϕ	V	F
A	S_A	$\phi_A = l-1$	V_A	V_A/V_E
B	S_B	$\phi_B = m-1$	V_B	V_B/V_E
E	S_E	$\phi_E = (l-1)(m-1)$	V_E	
T	S_T	$lm-1$		

- 가설
  A요인 : $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_l = 0$
  B요인 : $H_0 : \beta_1 = \beta_2 = \cdots = \beta_m = 0$

- 검정
  $F=V_A/V_E$가 $F(\phi_A, \phi_E ; \alpha)$ 보다 크면 유의수준 α에서 A 귀무가설 기각
  $F=V_B/V_E$가 $F(\phi_B, \phi_E ; \alpha)$ 보다 크면 유의수준 α에서 B 귀무가설 기각

한국방송통신대학교

이원배치법에서의 추정

- 각 수준 모평균 $\mu(\alpha_i)$ 및 $\mu(\beta_j)$ 의 점추정 및 100$(1-\alpha)$% 신뢰구간

$$\hat{\mu}(\alpha_i) = \hat{\mu} + \hat{\alpha}_i = \bar{x}_{i\cdot}$$ 이고 $$\bar{x}_{i\cdot} \pm t\left(\phi_E ; \frac{\alpha}{2}\right)\sqrt{\frac{V_E}{m}}$$

$$\hat{\mu}(\beta_j) = \hat{\mu} + \hat{\beta}_j = \bar{x}_{\cdot j}$$ 이고 $$\bar{x}_{\cdot j} \pm t\left(\phi_E ; \frac{\alpha}{2}\right)\sqrt{\frac{V_E}{l}}$$

- 요인 A의 i 수준과 요인 B의 j 수준에서의 100$(1-\alpha)$% 신뢰구간

$$\begin{aligned}\hat{\mu}(\alpha_i\beta_j) &= \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j \\ &= \hat{\mu} + \hat{\alpha}_i + \hat{\mu} + \hat{\beta}_j - \hat{\mu} \\ &= \bar{x}_{i\cdot} + \bar{x}_{\cdot j} - \bar{\bar{x}}\end{aligned}$$ 이고

$$\begin{aligned}&Var\left(\frac{1}{m}\sum_{j=1}^{m} x_{ij} + \frac{1}{l}\sum_{i=1}^{l} x_{ij} - \frac{1}{lm}\sum_{i=1}^{l}\sum_{j=1}^{m} x_{ij}\right) \\ &= \frac{\sigma_E^2}{m} + \frac{\sigma_E^2}{l} + \frac{\sigma_E^2}{lm} + 2Cov\left(\frac{1}{m}\sum_{j=1}^{m} x_{ij}, \frac{1}{l}\sum_{i=1}^{l} x_{ij}\right) \\ &- 2Cov\left(\frac{1}{l}\sum_{i=1}^{l} x_{ij}, \frac{1}{lm}\sum_{i=1}^{l}\sum_{j=1}^{m} x_{ij}\right) \\ &- 2Cov\left(\frac{1}{m}\sum_{j=1}^{m} x_{ij}, \frac{1}{lm}\sum_{i=1}^{l}\sum_{j=1}^{m} x_{ij}\right)\end{aligned}$$

$$Var(\bar{x}_{i\cdot} + \bar{x}_{\cdot j} - \bar{\bar{x}}) = \sigma_E^2 \Big/ \frac{lm}{l+m-1} = \frac{\sigma_E^2}{n_e}$$ 이므로 $$(\bar{x}_{i\cdot} + \bar{x}_{\cdot j} - \bar{\bar{x}}) \pm t\left(\phi_E ; \frac{\alpha}{2}\right)\sqrt{\frac{V_E}{n_e}}$$

04

R을 이용한 실습

두 모집단의 비교 - 대응표본

- t.test는 두 모평균을 비교하는 t검정을 실시하는 함수
- 'mu=0, alternative="less"는 대립가설이 '두 모평균의 차가 0보다 작다'를, 'paired=T'는 대응표본을 이용한 검정을 의미

```
pre <- c(72,80,83,63,66,76,82)

post <- c(78,82,82,68,70,75,88)

exam1 <- data.frame(pre, post)

t.test(exam1$pre, exam1$post, mu=0, alternative="less", paired=T)
```

```
        Paired t-test

data:  exam1$pre and exam1$post
t = -2.5981, df = 6, p-value = 0.02038
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.7562087
sample estimates:
mean of the differences
                     -3
```

다수 모집단의 비교 - 분산분석법

- 일원배치법
- factor 함수를 이용하여 요인 A의 각 수준을 지정
- aov함수를 이용하여 분산분석을 실시

```
x <- c(84,83,82,85,89,86,93,94,96,89,89,87)
A <- c(rep(1,3), rep(2,3),rep(3,3),rep(4,3))
A <- factor(A)
aovdat1 <- data.frame(x, A)
aovmodel1 <- aov(x ~ A, data=aovdat1)
summary(aovmodel1)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)    
A            3  200.9   66.97   29.77 0.000109 ***
Residuals    8   18.0    2.25                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

한국방송통신대학교

다수 모집단의 비교 - 분산분석법

- 이원배치법
- 각 요인(factor)을 나타내는 변수의 합으로 반응값을 설명

```
y <- c(97.8,97.5,96.9,98.5,98.8,97.1,99.2,98.4,98.1,98.2,97.5,96.8)
surface <- c(rep(1,3), rep(2,3),rep(3,3),rep(4,3))
manu <- rep(c(1,2,3),4)
surface <- factor(surface)
manu <- factor(manu)
aovdat2 <- data.frame(surface, manu)
aovmodel2 <- aov(y ~ surface + manu, data=aovdat2)
summary(aovmodel2)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
surface      3 2.7267  0.9089   8.039 0.0159 *
manu         2 3.0150  1.5075  13.334 0.0062 **
Residuals    6 0.6783  0.1131
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

정리하기

- 서로 독립적으로 추출된 표본의 수가 충분히 큰 경우(통상 30보다 큰 경우)에는 두 모평균이 같다는 가설은 모집단의 분포와 관계 없이 표준정규분포를 이용하여 검정한다.
- 서로 독립적으로 추출된 표본 수가 작을 경우,
  두 모평균이 같다는 가설은 두 모집단이 정규분포를 따르고
  두 모분산이 같다는 가정 하에서 t 분포를 이용하여 검정한다.
- 모집단이 정규분포이고 두 표본이 쌍(종속적)으로 추출되었을 경우,
  두 모평균의 가설검정은 짝지어진 n쌍(pair)의 표본의 차를 계산하여
  단일표본의 검정문제로 단순화하여 검정한다.

정리하기

- 두 모집단이 정규분포인 경우, 두 모분산이 같다는 가설은 표본분산비를 계산하고 F 분포를 이용하여 검정한다.
- 분산분석이란 실험계획법에 의하여 얻어진 특성값의 분포를 총제곱합으로 나타내고, 이 총제곱합을 요인마다 제곱합으로 분해하여 오차에 비해 특히 큰 영향을 주는 요인이 무엇인가를 검토하는 분석방법이다.

다음 시간 안내

13강

통계적 비교 II

한국방송통신대학교