통계학 개론

제7장 통계적 비교

7.1 두 모집단의 비교

두 모평균의 비교는 각 모집단에서 추출된 표본이 서로 독립적으로 추출되었을 경우(독립표본)와 아닌 경우(대응표본)에 따라 검정방법이 다르다.

1) 두 독립표본의 평균 비교

두 모평균에 대한 가설검정 유형

① H_0: $\mu_1 - \mu_2 = D_0$

H_0: $\mu_1 - \mu_2 > D_0$

② H_0: $\mu_1 - \mu_2 = D_0$

H_0: $\mu_1 - \mu_2 < D_0$

③ H_0: $\mu_1 - \mu_2 = D_0$

H_0: $\mu_1 - \mu_2 \neq D_0$

모집단에서 서로 독립적으로 표본을 추출했을 때 모평균의 차 μ_1-μ_2의 추정량은 표본평균의 차 $\overline{X}_1 - \overline{X}_2$이며, 모든 가능한 표본평균의 차는 표본이 충분히 클 경우 근사적으로 평균이 μ_1-μ_2이고, 분산이 $(\sigma_1^2/n_1)+(\sigma_2^2/n_2)$인 정규분포를 따르게 된다.

$$\overline{X}_1 - \overline{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

1-1) 두 모집단의 분산 σ_1^2, σ_2^2를 모르고, 두 모분산이 같은 경우 자유도가 n1+n2-2인 t분포를 따르는 통계량 T를 사용한다.

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - D_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{ 단 공통분산 } S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

❖ (표본이 서로 독립적으로 추출되었으며, 두 모집단이 정규분포를 따르고, 두 모분산이 같은 경우) 두 모평균의 가설검정

가설의 종류	선택기준
H_0: $\mu_1 - \mu_2 = D_0$ H_0: $\mu_1 - \mu_2 > D_0$	$T > t_{n_1+n_2-2,\alpha}$ 이면 H_0 기각
H_0: $\mu_1 - \mu_2 = D_0$ H_0: $\mu_1 - \mu_2 < D_0$	$T < -t_{n_1+n_2-2,\alpha}$ 이면 H_0 기각
H_0: $\mu_1 - \mu_2 = D_0$ H_0: $\mu_1 - \mu_2 \neq D_0$	$\lvert T \rvert > t_{n_1+n_2-2,\alpha/2}$ 이면 H_0 기각

1-2) 두 모집단의 분산 σ_1^2, σ_2^2를 모르고, 두 모분산이 다른 경우 자유도가 ϕ인 t 분포를 따르는 통계량 T를 사용한다.

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - D_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_{\phi,\alpha} \quad \text{다만,} \quad \phi = \frac{\left[S_1^2/n_1 + S_2^2/n_2\right]^2}{\frac{S_1^2/n_1}{n_1-1} + \frac{S_2^2/n_2}{n_2-1}}$$

❖ (표본이 서로 독립적으로 추출되었으며, 두 모집단이 정규분포를 따르고, 두 모분산이 다른 경우) 두 모평균의 가설검정

가설의 종류	선택기준
H_0: $\mu_1 - \mu_2 = D_0$ H_0: $\mu_1 - \mu_2 > D_0$	$T > t_{\phi,\alpha}$ 이면 H_0 기각
H_0: $\mu_1 - \mu_2 = D_0$ H_0: $\mu_1 - \mu_2 < D_0$	$T < -t_{\phi,\alpha}$ 이면 H_0 기각
H_0: $\mu_1 - \mu_2 = D_0$ H_0: $\mu_1 - \mu_2 \neq D_0$	$\lvert T \rvert > t_{\phi,\alpha/2}$ 이면 H_0 기각

2) 대응표본의 평균 비교

대응비교(paired comparison): 비슷한 성질의 대응표본을 사용하여 두 모집단의 평균을 비교하는 가설검정

대응비교 일 때는 먼저 다음과 같이 관찰된 n쌍의 차(D_i)를 계산해서 평균($\overline{D}$)과 표준편차(s_D)를 구한다.

❖ 대응표본의 차, 평균, 분산

D_i의 평균 $\overline{D} = \sum D_i / n$

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

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}	$D_2 = X_{21} - X_{22}$
⋮	⋮	⋮
X_{n1}	X_{n2}	$D_n = X_{n1} - X_{n2}$

두 모집단이 모평균이 같은 정규분포일 때 $\frac{\overline{D}}{s_D/\sqrt{n}}$는 자유도가 n-1인 t 분포를 따른다.

$$\frac{\overline{D}}{s_D/\sqrt{n}} \sim t_{n-1,\alpha}$$

❖ (모집단이 정규분포이고 두 표본이 쌍으로 추출되었을 경우) 두 모평균의 가설검정

가설의 종류	선택기준
H_0: $\mu_1 - \mu_2 = D_0$ H_0: $\mu_1 - \mu_2 > D_0$	$\frac{\overline{D} - D_o}{s_D/\sqrt{n}} > t_{n-1,\alpha}$ 이면 H_0 기각
H_0: $\mu_1 - \mu_2 = D_0$ H_0: $\mu_1 - \mu_2 < D_0$	$\frac{\overline{D} - D_o}{s_D/\sqrt{n}} < -t_{n-1,\alpha}$ 이면 H_0 기각
H_0: $\mu_1 - \mu_2 = D_0$ H_0: $\mu_1 - \mu_2 \neq D_0$	$\left\lvert \frac{\overline{D} - D_o}{s_D/\sqrt{n}} \right\rvert > t_{n-1,\alpha/2}$ 이면 H_0 기각

3) 두 모분산 가설검정

두 모집단의 분산(σ_1^2, σ_2^2)을 비교하는 경우, 분산의 비(σ_1^2/σ_2^2)를 계산한다

통계량 $\frac{(S_1^2/\sigma_1^2)}{(S_2^2/\sigma_2^2)}$ 은 두 모집단이 각각 정규분포를 따를 경우 분자자유도 n_1-1, 분모자유도 n_2-1인 F분포를 따른다.

$$\frac{(S_1^2/\sigma_1^2)}{(S_2^2/\sigma_2^2)} \sim F_{n_1-1,n_2-1,\alpha}$$

❖ (두 모집단이 정규분포인 경우) 두 모분산의 가설검정

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

7.2 세 개 이상 다수 모집단의 비교

요인(factor): 실험결과에 영향을 주는 무수히 많은 원인 중에서 실험에서 직접 취급되어 관리되는 것

랜덤화(randomization): 요인의 각 수준에서 실험단위의 배정 또는 실험순서를 임의로 배정하는 것
실험계획법(design of experiments): 실험을 합리적으로 설계하는 방법

1) 분산분석의 개념
분산분석(analysis of variance):
① 세 개 이상의 평균비교에 대한 검정방법
② 특성값의 변동을 제곱합으로 나타내고, 이것을 실험과 관련된 요인의 제곱합과 오차의 제곱합으로 분해하여 오차에 비해 영향이 큰 요인이 무엇인가를 찾아내는 분석방법
③ 각 요인의 평균제곱을 구하고, 이 값이 오차의 분산에 비해 얼마나 큰가를 검토. 요인의 평균제곱값이 오차의 분산보다 매우 크다면 그 요인은 특성값의 변동을 유의하게 설명해 주는 요인임
※ 요인의 평균제곱: 요인의 제곱합을 요인의 자유도로 나눈 값

2) 일원배치법
(1) 일원배치법의 개념
일원배치법(one-way factorial design): 어떤 관심이 있는 특성값에 대하여 하나의 요인의 영향을 조사하기 위하여 쓰는 실험계획법

일원배치법에서는 반복수가 다르더라도 특별한 수정 없이 그대로 분석할 수 있다.

실험의 완전 랜덤화는 일원배치법에서 매우 중요한 특징이기 때문에 일원배치법을 완전확률화법(completely randomized design)이라고도 한다.

(2) 데이터의 구조

❖ 일원배치법 데이터의 배열

구분	인자의 수준	
	$A_1 \quad A_2 \quad \cdots \quad A_l$	
실험의 반복	$\begin{matrix} x_{11} & x_{21} & \cdots & x_{l1} \\ x_{12} & x_{22} & \cdots & x_{l2} \\ \vdots & \vdots & & \vdots \\ x_{1m} & x_{2m} & \cdots & x_{lm} \end{matrix}$	
합계	$T_1. \quad T_2. \quad \cdots \quad T_l.$	T
평균	$\overline{x}_1. \quad \overline{x}_2. \quad \cdots \quad \overline{x}_l.$	$\overline{\overline{x}}$

$$T_i. = \sum_{j=1}^{m} X_{ij} \quad \overline{x_i.} = \frac{T_i.}{m} \quad (i = 1, 2, \cdots, l)$$

$$T = \sum_{i=1}^{l} T_i. \quad \overline{\overline{x}} = \frac{T}{lm}$$

A_i 수준에서의 j번째 데이터 x_{ij}는 A_i 수준에서 특성값의 모평균 μ_i를 중심으로 오차 ε_{ij}를 가진 변량으로 일반화하면

$$x_{ij} = \mu_i + \epsilon_{ij} \quad (i = 1, 2, \cdots, l, \quad j = 1, 2, \cdots, m)$$

실험 전체의 모평균 μ는

$$\mu = \sum_{i=1}^{l} \frac{\mu_i}{l}$$

요인 A_i의 주효과(main effect) α_i는

$$\alpha = \mu_i - \mu$$

이를 정리하면

$$\begin{aligned} x_{ij} &= \mu_i + \epsilon_{ij} \\ &= \mu + (\mu_i - \mu) + \epsilon_{ij} \\ &= \mu + \alpha_i + \epsilon_{ij} \end{aligned}$$

주효과 α_i의 합은 항상 0이다.

❖ 일원배치법의 데이터 구조식

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

$$\text{단}, \quad \epsilon_{ij} \sim N(0, \sigma_E^2), \quad P(x_{ij} \cap \epsilon_{ij}) = P(x_{ij})P(\epsilon_{ij}), \quad \sum_{i=1}^{l} \alpha_i = 0$$

데이터 x_{ij}는 공통적인 평균 μ, 요인 A의 효과 α_{ij}와 A 요인으로 설명할 수 없는 오차 ε_{ij}로 구성되어 있다.

(3) 분산분석표의 작성

❖ 일원배치법의 분산분석표

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

$$S_T = S_E + S_A$$

데이터의 총변동(S_T)은 요인수준의 변화에 따른 변동(S_A)과 수준 내의 오차변동(S_E)이라는 두 개의 요인으로 분해된다.

F 통계량의 값은 평균제곱 V_A와 V_E의 비로 계산되는데, 이것을 통해 요인 A의 l 수준의 모평균이 동일하다는 가설을 검정하게 된다.

H0: $\mu_1 = \mu_2 = \cdots = \mu_l$

H1: μ_i가 모두 같지는 않다.

이 때 F 값이 커지면 요인 A에 의한 변동이 오차에 의한 변동보다 커지므로 요인 A에 의한 변동이 유의하다고 판단한다.

F 통계량은 귀무가설하에서 자유도 ϕ_A, ϕ_E인 F분포를 따른다.

F 통계량의 값이 $F \geq F(\phi_A, \phi_E : \alpha)$ 이면 유의수준 α에서 귀무가설이 기각된다.

(4) 모평균의 추정

① 각 수준의 모평균 추정

요인 A의 i수준에서의 모평균 $\mu_i = \mu + \alpha_i$

요인 A의 i수준에서 m개의 데이터 x_{ij}(j=1, 2, …, m)는 정규분포 $N(\mu_i, \sigma_E^2)$에서 얻어진 크기 m의 확률표본이다.

표본분포의 성질을 다시 보면,

기댓값 $E(\bar{X}) = \mu, Var(\bar{X}) = \frac{\sigma^2}{n}$

이를 적용하면 $\mu_i = \mu + \alpha_i$의 점추정량은 $\bar{x}_{1\cdot}$.

$\bar{x}_{1.}$의 분산은 $Var(\bar{x}_{i.}) = \frac{\sigma_E^2}{m}$

μ의 100(1-α)% 신뢰구간은 $\left[\bar{x}_{i.} - t(\phi_E;\frac{\alpha}{2})\sqrt{\frac{V_E}{n}},\ \bar{x}_{i.} + t(\phi_E;\frac{\alpha}{2})\sqrt{\frac{V_E}{n}}\right]$

여기서 $t(\phi_E)$는 자유도 $\phi_E = l(m-1)$인 t분포

② 각 수준의 모평균차의 추정과 검정

요인 A의 두 수준 i와 i'에서의 모평균 차이는 요인 A의 수준효과의 차이

$$\mu_i - \mu_{i'} = (\mu + \alpha_i) - (\mu + \alpha_{i'}) = \alpha_i - \alpha_{i'}$$

여기에서 $\mu_i - \mu_{i'} = \alpha_i - \alpha_{i'}$은 다음과 같이 두 수준의 표본평균의 차이로 추정된다.

$$\hat{\mu_i} - \hat{\mu_{i'}} = \hat{\alpha_i} - \hat{\alpha_{i'}} = \bar{x}_{i.} - \bar{x}_{i'.}$$

이 추정값의 분산은 A_i 수준의 데이터와 $A_{i'}$ 수준의 데이터가 서로 독립이므로 다음과 같다.

$$Var(\bar{x}_{i.} - \bar{x}_{i'.}) = Var(\bar{x}_{i.}) + Var(\bar{x}_{i'.}) - 2Cov(\bar{x}_{i.}, \bar{x}_{i'.})$$
$$= \frac{\sigma_E^2}{m} + \frac{\sigma_E^2}{m} = \frac{2\sigma_E^2}{m}$$

따라서, $\sigma_E^2 = V_E$를 사용하여 $\mu_i - \mu_{i'}$의 100(1-α)% 신뢰구간을 구할 수 있다

$$\left[(\bar{x}_{i.} - \bar{x}_{i'.}) - t(\phi_E;\frac{\alpha}{2})\sqrt{\frac{2V_E}{n}},\ (\bar{x}_{i.} - \bar{x}_{i'.}) + t(\phi_E;\frac{\alpha}{2})\sqrt{\frac{2V_E}{n}}\right]$$

만약 $|\bar{x}_{i.} - \bar{x}_{i'.}| \geq t(\phi_E;\frac{\alpha}{2})\sqrt{\frac{2V_E}{m}}$이면, 두 수준 A_i, $A_{i'}$의 모평균은 유의수준 α에서 유의하게 차이가 있다.

여기에서, $t(\phi_E;\frac{\alpha}{2})\sqrt{\frac{2V_E}{n}}$ 을 최소유의차(least significant difference: LSD)라고 한다.

따라서, LSD를 먼저 구해 놓고, 모든 수준 간의 표본평균의 차이 $|\bar{x}_{i.} - \bar{x}_{i'.}|$를 구하여 이 값이 LSD보다 크면 두 수준 간의 차이가 유의하고, 이것이 LSD보다 작으면 두 수준간의 차이는 유의하지 않다고 결론지을 수 있다.

3) 이원배치법

(1) 이원배치법의 개념

이원배치법: 문제가 되는 요인을 두 개 취하여 행하는 실험

(2) 데이터의 구조

$$x_{ij} = \mu + \alpha_i + \beta_i + \epsilon_{ij}$$

$\epsilon_{ij} \sim N(0, \sigma_E^2)$ 이고 서로 독립

$$i = 1,2,\cdots,l \quad j = 1,2,\cdots,m$$

❖ 반복이 없는 이원배치법의 자료배열

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

(3) 분산분석표의 작성

이원배치법에서 데이터 x_{ij}와 총평균 x의 차이는 다음과 같이 나눌 수 있다

$$(x_{ij} - \bar{\bar{x}}) = (\bar{x}_{i\cdot} - \bar{\bar{x}}) + (\bar{x}_{\cdot j} - \bar{\bar{x}}) + (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}})$$

양변을 제곱한 후에 모든 i, j에 대하여 합하면 다음과 같다.

$$\sum_{i=1}^{l}\sum_{j=1}^{m}(x_{ij} - \bar{\bar{x}})^2 = \sum_{i=1}^{l}\sum_{j=1}^{m}(\bar{x}_{i\cdot} - \bar{\bar{x}})^2 + \sum_{i=1}^{l}\sum_{j=1}^{m}(\bar{x}_{\cdot j} - \bar{\bar{x}})^2 + \sum_{i=1}^{l}\sum_{j=1}^{m}(x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}})^2$$

$$S_T = S_A + S_B + S_E$$

❖ 반복이 없는 이원배치법의 분산분석표

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_C	$\phi_E = (l-1)(m-1)$	V_E	
T	S_T	$lm-1$		

$F = V_A/V_E \geq F(\phi_A, \phi_E ;\alpha)$이면 $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_l = 0$ 이 유의수준 α에서 기각

$F = V_B/V_E \geq F(\phi_B, \phi_E ;\alpha)$이면 $H_0 : \beta_1 = \beta_2 = \cdots = \beta_l = 0$ 이 유의수준 α에서 기각

(4) 모평균의 추정

① 요인 A의 모평균의 추정

$$\hat{\mu}(\alpha_i) = \hat{\mu} + \hat{\alpha}_i = \bar{x}_{i\cdot}$$

$\mu(\alpha_i)$의 100(1-α)% 신뢰구간은

$$\left[\bar{x}_{i\cdot} - t(\phi_E;\frac{\alpha}{2})\sqrt{\frac{V_E}{m}},\ \bar{x}_{i\cdot} + t(\phi_E;\frac{\alpha}{2})\sqrt{\frac{V_E}{m}}\right]$$

② 요인 B의 모평균의 추정

$$\hat{\mu}(\beta_j) = \hat{\mu} + \hat{\beta}_j = \bar{x}_{\cdot j}$$

$\mu(\beta_j)$의 100(1-α)% 신뢰구간은

$$\left[\bar{x}_{\cdot j} - t(\phi_E;\frac{\alpha}{2})\sqrt{\frac{V_E}{l}},\ \bar{x}_{\cdot j} + t(\phi_E;\frac{\alpha}{2})\sqrt{\frac{V_E}{l}}\right]$$

③ 두 요인의 수준을 조합한 조건에서의 모평균의 추정

A 요인의 i 수준과 B 요인의 j 수준에서 모평균의 점추정량은 다음과 같다.

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

이 때 점추정량의 분산은 다음과 같다.

$$Var(\bar{x}_{i\cdot} + \bar{x}_{\cdot j} - \bar{\bar{x}}) = \frac{\sigma_E^2}{lm/(l+m-1)} = \frac{\sigma_E^2}{n_e}$$

※ 유효반복수 $n_e = \dfrac{lm}{l+m-1}$

$\mu(\alpha_i\beta_j)$의 100(1-α)% 신뢰구간은

$$\left[(\bar{x}_{i\cdot} + \bar{x}_{\cdot j} - \bar{\bar{x}}) - t(\phi_E;\frac{\alpha}{2})\sqrt{\frac{V_E}{n_e}},\ (\bar{x}_{i\cdot} + \bar{x}_{\cdot j} - \bar{\bar{x}}) + t(\phi_E;\frac{\alpha}{2})\sqrt{\frac{V_E}{n_e}}\right]$$