분 산 분 석

목 차

1. 분산분석의 목적
2. 분산분석의 예
3. 분산분석에 사용되는 용어
4. 일원배치법
5. 이원배치법

1. 분산분석의 목적

⌘ 연구대상인 특성은 몇 가지 인자에 의하여 영향을 받는다. 이 몇 가지 인자중에서 어느 인자에 따라서 가장 많이 영향을 받는지를 결정하거나 이 인자를 어떻게 조절하면 최적의 상태로 되는가의 조건을 결정하는 일이 중요하다.이에 분산분석이 사용되어진다.

2. 분산분석의 정의

⌘ 실험결과 생긴 자료의 분산을 각 인자에 기인하는 부분과 우연오차에 기인하는 부분으로 분해하고 인자에 기인하는 분산과 우연오차에 기인하는 분산차이의 유의성을 검증함으로써 각 인자의 영향을 가려내는 방법이다.

3.분산분석에 사용되는 용어

- 인자(factor) : 실험결과에 영향을 미치는 원인
- 수준(level) : 인자의 상태를 지정한 조건
- 요인 (factorial factor) : 주효과와 교호작용의 원인이 되는 요소의 전체 (주효과, 교호작용)

*확률변수 : 한 동전을 2회 던지는 시행에서 앞면이 나오는 사건은 H, 뒷면이 나오는 사건 T라 하고 표본 공간 S={TT,TH,HT,HH} 이때 앞면의 수를 X라고 하면 X에의하여 표본공간의 원소에 앞면의 수를 대응시킬 수 있다. X(S)={0, 1, 1, 2}가 된다. 표본공간 S의 각 원소를 어떤 규칙 X에 의하여 실수 x로 바꾸어 생각하는 경우 x=X(S)와 같이 X는 표본공간S를 정의구역으로 하는 함수이다.

즉 표본공간을 정의구역으로 하는 실수 함수이다.

*귀무가설 : 어떤 회사제품의 전구의 평균수명은 1500시간이라고 주장하고 있다. 지금 이 회사의 전구중에서 36개를 무작위추출하여 검사해보니 1478시간이었다.
이때 일반적으로 표본으로 뽑힌 모집단의 평균을 1500이라 가정하고 이 가정은 모집단의 평균과 회사에서 주장하는 평균 1500과 차이가 없다라는 가정을 귀무가설이라 한다.

*대립가설 : 귀무가설이 기각되는 경우 모집단의 평균은 1500이 아니라는 가정이다.

*유의적 : 표본평균값이 귀무가설를 기각할만큼 모집단평균값으로부터 떨어진 상태를 유의적이라 한다.

*위험역 : 가설이 참인데도 기각할때 참인 가설을 기각하는 과오를 범하는 확률을 어느정도 허용하느냐의 기준을 유의수준 혹은 위험영역이라 한다.

* 자유도 : 임의의 표본 $(X_1, X_2, \cdots X_n)$에서

평방합 $\sum_{i=1}^{n}\left(X_i - \bar{X}_i\right)^2$ 에대하여 자유도는 $n-1$

그 이유는 $\sum_{i=1}^{n}\left(X_i - \bar{X}_i\right)^2$ 의 값을 좌우하는 변수의 수는

겉보기로는 $X_1, X_2, \cdots X_n$의 n개로 보이지만 이들 변

수사이에는 $\sum_{i=1}^{n}\left(X_i - \bar{X}\right) = (X_1 - \bar{X}) + (X_2 - \bar{X})$

$+ \cdots + (X_n - \bar{X}) = 0$인 관계가 있으므로 이중 $(n-1)$의

개의 값은 자유로이 취할 수있으나 제 n변수 는 나머

머지값에의하여 결정이 된다. 따라서 평방합 $\sum_{i=1}^{n}\left(X_i - \bar{X}_i\right)^2$

의 값은 $(n-1)$개의 독립변수를 가지므로 자유도는 $(n-1)$

이다.

4. 일원배치법

⌘ 3개 이상의 평균의 균일성을 검증하는데 사용되는 방법

⌘ 일원배치표

인자 A	자료값	계	자료수	평균
A_1	$x_{11}\ \ x_{12} \cdots x_{1j} \cdots x_{1n1}$	$T_{1.}$	n_1	$\bar{X}_{1.} = T_{1.} / n_1$
A_2	$x_{21}\ \ x_{22} \cdots x_{2j} \cdots x_{2n2}$	$T_{2.}$	n_2	$\bar{X}_{2.} = T_{2.} / n_2$
M	M M M M	M	M	M M
A_i	$x_{i1}\ \ x_{i2} \cdots x_{ij} \cdots x_{ini}$	$T_{i.}$	n_i	$\bar{X}_{i.} = k_{i.} / n_i$
A_k	$x_{k1}\ \ x_{k2} \cdots x_{kj} \cdots x_{knk}$	$T_{k.}$	n_k	$\bar{X}_{k.} = T_{k.} / n_k$
계		$T_{..}$	n	$\bar{X}_{..} = T_{..} / n$

실측값 $x_{i,j}$로 나타난 확률변수를 $X_{i,j}$

$X_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}$

μ : 이 자료전체의평균

α_i : 수준A_i의 주효과를가리키는상수

$\varepsilon_{i,j}$: i, j와 관계없이나타나는실험오차로서정규분포$(0, \alpha^2)$

$$\bar{X}_{i.} = \frac{1}{n_i}\sum_{i=1}^{n_i} X_{ij} = \frac{1}{n_i}\sum_{i=1}^{n_i}(\mu + \alpha_i + \varepsilon_{ij})$$

$$= \mu + \alpha_i + \bar{\varepsilon}_{i.} \quad (\bar{\varepsilon}_{i.} = \frac{1}{n_i}\sum_{i=1}^{n_i}\varepsilon_{ij})$$

$$\bar{X}_{ij} = \frac{1}{n_i}\sum_{i=1}^{k}\sum_{j=1}^{n_j} X_{ij} = \mu + \bar{\alpha}_i + \bar{\varepsilon}_{i.} \quad (\bar{\alpha}_i = \frac{1}{n_i}\sum_{i=1}^{n_i}\alpha_i \ , \bar{\varepsilon}_{i.} = \frac{1}{n_i}\sum_{i=1}^{n_i}\varepsilon_{ij})$$

$$S_T(\text{totalvariation}) = \sum_{i=1}^{k}\sum_{j=1}^{n_j}(X_{ij} - \bar{X}_{..})^2 \qquad (1)$$

$S_T(\text{totalvariation})$: 각변수 X_{ij}와 자료의전체평균 $\bar{X}_{..}$ 와의 차의 제곱의 합 즉 자료전체에대한 평방합

전변동 S_T 의 자유도 $\nu_T = n-1$

(1)식을 다음과 같이 변형하면

$$S_T = \sum_i \sum_j \{(X_{ij} - \bar{X}_{i.}) + (\bar{X}_{i.} - \bar{X}_{..})\}^2$$

$$= \sum_i \sum_j (X_{ij} - \bar{X}_{i.})^2 + 2\sum_{i=1}^{k} (\bar{X}_{i.} - \bar{X}_{..}) \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})$$

$$+ \sum_{i=1}^{k} (\bar{X}_{i.} - \bar{X}_{..})^2$$

$$= \sum_i \sum_j (X_{ij} - \bar{X}_{i.})^2 + \sum_{i=1}^{k} (\bar{X}_{i.} - \bar{X}_{..})^2$$

$$S_E\ (\text{급내변동}) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2$$

$$\text{자유도}\ \nu_E = \sum_{i=1}^{k}(n_i - 1) = n - k$$

$$(\Theta \sum_{j=1}^{n}(X_{ij} - \bar{X}_{i.})^2 \text{ 의 자유도는 } n_i - 1\ )$$

$$S_A\ (\text{급간변동}) = \sum_{i=1}^{k} n_i(\bar{X}_{i.} - \bar{X}_{..})^2$$

$$\text{자유도}\ \nu_A = k - 1$$

$$S_E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\{(\mu + \alpha_i + \varepsilon_{ij}) - (\mu + \alpha_i + \bar{\varepsilon}_{i.})\}^2$$

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}\ ,\quad \bar{X}_{i.} = \mu + \alpha_i + \bar{\varepsilon}_{i.}$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\varepsilon_{ij} - \bar{\varepsilon}_{i.})^2$$

$$S_A = \sum_{i=1}^{k} n_i(\bar{X}_{i.} - \bar{X}_{..})^2 = \sum_{i=1}^{k} n_i\{(\mu + \alpha_i + \bar{\varepsilon}_{i.}) - \mu + \alpha_i + \bar{\varepsilon}_{..})\}^2$$

$$= \sum_{i=1}^{k} n_i(\alpha_i - \alpha_{i.} + \bar{\varepsilon}_{i.} - \bar{\varepsilon}_{..})^2$$

"각 수준의 주효과 사이에 차이가 없다"라는 가설

즉 귀무가설 $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_i = \alpha_k$ 를 세우면

급내불편분산 $V_E = \dfrac{S_E}{V_E} = \dfrac{S_E}{n-k}$

급간불편분산 $V_A = \dfrac{S_A}{V_A} = \dfrac{S_A}{k-1}$

이들의 분산비 $F = \dfrac{V_A}{V_E}$

이는 분자의 자유도 k-1, 분모의자유도 n-k인

F분포를 한다.

*일원배치의 분산분석

(1) 가설 H_0 : 인자 A의수준사이에 주효과의 차이가 없다.

(2) 전변동 $S_T = \sum_i \sum_j (x_{ij} - \frac{T_{..}^{\,2}}{n}) \quad \nu_T = n - 1$

급간변동 $S_A = \sum_i (\frac{T_{i.}^{\,2}}{n_i} - \frac{T_{..}^{\,2}}{n})^2 \quad \nu_a = k - 1$

급내변동(오차) $S_E = S_T - S_A \quad \nu_E = \nu_T - \nu_a = n - k$

불편분산 $V_A = \frac{S_A}{k-1} \qquad V_E = \frac{S_E}{n-k}$

분산비 $F = \frac{V_A}{V_E}$

(3) 우측검정 $F \geq F^{k-1}_{n-k}(\alpha)$ 이면 유의수준 α 로 가설을 기각
즉 인자 A에 의한 수준간 효과의차이는 유의적

(4) 분산분석표작성

요인	평방합	자유도	불편분산	분산비	판정기준
급간	S_A	$k-1$	$V_A = S_A/k-1$	$F = V_A / V_E$	$F^{k-1}_{n-k}(\alpha)$
급내	S_E	$n-k$	$V_E = S_E / n-k$		
전	S_T	$n-1$			

*두 수준간의 평균의 차의 검정
분석결과 수준간의 주효과의 유의차가 인정되면 각 수준중의 어느 두 수준의 평균차가 유의적인가를 알아보는 것이다. 이때 귀무가설을 검정할때의 유의수준의 위험역은 양측검과 우측검정으로 한다.

*평균차의 검정

모분산이 같은 두 정규분포집단 $A: N(\mu, \sigma^2)$, $B: N(\mu, \sigma^2)$

(1) A,B에대하여 모평균이 같다라는 가설을 세운다.

(2)A,B에서 서로 독립적으로 각각 크기 n_1, n_2인 표본을 추출하여 얻은 표본값 $(x_1, x_2 \dots x_{n1})$, $(y_1, y_2 \dots y_{n2})$

표본평균값 $\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_i} x_i$, $\bar{y} = \frac{1}{n_1}\sum_{i=1}^{n_2} y_i$

이때의 불편분산 $\hat{\sigma}^2 = \frac{1}{n_1 + n_2 - 2}\{\sum_{i=1}^{n_i}(x_i - \bar{x})^2 + \sum_{i=1}^{n_2}(y_i - \bar{y})^2\}$

$$t = \frac{\bar{x} - \bar{y}}{\hat{\sigma}}\sqrt{1/n_1 + 1/n_2}$$

(3)양측검정

(4)우측검정

(5)좌측검정

i째 수준의 모평균 $\mu+\alpha_{ㅑi}$와 j째 수준의 모평균을 $\mu+\alpha_j$이면

가설 : $\mu+\alpha_{ㅑi}=\mu+\alpha_j$ 즉 $\alpha_i=\alpha_j$

$$\mathrm{T}=\frac{\left|X_{i.}-X_{j.}\right|}{\sqrt{V_E(1/n_i+1/n_j)}}$$ 인 자유도 $\nu_{\mathrm{E}}=n-k$인 t분포를한다.

따라서 귀무가설을 검정할때의 유의수준 $\alpha=0.05$ 또는 0.01_)

인 위험역은

(1) 양측검정

$$\left|\mathrm{x}_{\mathrm{i.}}-x_{j.}\right|\geq t_{n-k}(\alpha)\sqrt{V_E(1/n_i+1/n_j)}$$

(2) 우측검정

$$\left|\mathrm{x}_{\mathrm{i.}}-x_{j.}\right|\geq t_{n-k}(2\alpha)\sqrt{V_E(1/n_i+1/n_j)}$$

Ex)어떤 합성반응에서 촉매의 양이 합성물의 생산량에 영향을 알기취해 촉매의 양을 2g, 4g, 6g 의 3수준으로 하여 각 수준마다 4회씩 실험한 결과 다음과 같은 자료를 얻었다. 촉매의 양은 합성물의 생성량에 영향을 미친다고 할 수 있는가?

$u_{ij} = x_{ij}-77$ (77은 가평균)과 같이 변환함으로써 자료값 x_{ij} 를 u_{ij} 로 고치면

촉매량	생성량			
$A_1(2g)$	74	69	73	67
$A_2(4g)$	79	81	75	78
$A_3(6g)$	82	85	80	79

인자 A	$u_{ij} = x_{ij} - 77$				계	(계)^2
A_1	-3	-8	-4	-10	-25	625
A_2	2	4	-2	1	5	25
A_3	5	8	3	2	18	324
계					-2	974

$CT=(-2)^2/12=0.33$

전변동 : $S_T=\{(-3)^2+(-8)^2+(-4)^2+(-10)^2+(2)^2+(4)^2+(-2)^2+(1)^2+(5)^2+(8)^2+(3)^2+(2)^2\}-CT=316-0=316$

전변동의 자유도 12-1=11

촉매량간변동 : $S_A=\{(-25^2/4)+(-5^2/4)+(18^2/4)-CT=244$

촉매량간변동의 자유도 3-1=2

촉매량내변동 : $S_E=S_T-S_A=72$

자유도 11-2=9

요인	평방합	자유도	불편분산	분산비	판정기준
급간	244	2	122	15.2	$F_9^2(0.05)=4.26$
급내	72	9	8		$F_9^2(0.01)=8.02$
전	316	11			

위의 예제 각 촉매량에 대한 평균의 차의 유의성을 개별적으로 비교하면

촉매량 2g, 4g, 6g에대한 합성물 생성량의 평균을 각각 $\bar{X}_{1.}$,

$\bar{X}_{2.}$, $\bar{X}_{3.}$ 라하면

$$\bar{X}_{1.} = \frac{-25}{4} + 77 = 70.25 \qquad \bar{X}_{2.} = \frac{5}{4} + 77 = 78.25$$

$$\bar{X}_{3.} = \frac{18}{4} + 77 = 81.50$$

오차변동의 자유도 9

$$\therefore \quad t_9(0.05) = 2.262 \quad , t_9(0.01) = 3.250$$

$$n_1 = n_2 = n_3 = n_4 = 4 \quad , \qquad \nu_E = 8$$

$$t_9(0.05)\sqrt{V_E(1/4 + 1/4)} = 4.524$$

$$t_9(0.01)\sqrt{V_E(1/4 + 1/4)} = 6.500$$

1. $\bar{x}_{1.}$ 과 $\bar{x}_{2.}$ 의차

$= |70.25 - 78.25| = 8.0 \rangle t_9(0.01)\sqrt{V_E(1/4 + 1/4)} = 6.5$

$\Rightarrow$ 유의수준 1%으로 유의적이다.

2. $\bar{x}_{1.}$ 과 $\bar{x}_{2.}$ 의차

$= |70.25 - 81.25| = 11.0 \rangle t_9(0.01)\sqrt{V_E(1/4 + 1/4)} = 6.5$

$\Rightarrow$ 유의수준 1%으로 유의적이다.

3. $\bar{x}_{2.}$ 과 $\bar{x}_{3.}$ 의차

$= |78.25 - 81.25| = 3.25 \rangle t_9(0.05)\sqrt{V_E(1/4 + 1/4)} = 4.524$

$\Rightarrow$ 유의적이 아니다.

5. 이원배치법

⌘ 두 인자 A,B가 있어서 인자 A의 수준은 k개, 인잔 B의 수준은 *l*개가 있다고 하면 A수준의 A_i와 B수준의 B_j가 결합된 조건(A_i,B_j)에서의 측정값 x_{ij}를 얻으면 이 실험을 이원배치라하고 이원배치표에서 인자의 각 수준에서의 측정값이 가로로 배열된것을 행 (행인자), 세로로 배열된 것을 열 (열인자)

	B_1	B_2	$\cdots$	B_j	$\cdots$	B_l	계	평균
A_1	x_{11}	x_{12}	$\cdots$	x_{ij}	$\cdots$	x_{1l}	$T_{1.}$	$\bar{X}_{1.}$
A_2	x_{21}	x_{22}	$\cdots$	x_{ij}	$\cdots$	x_{2l}	$T_{2.}$	$\bar{X}_{2.}$
M	M	M		M		M	M	M
A_i	x_{i1}	x_{i2}	$\cdots$	x_{ij}	$\cdots$	x_{il}	$T_{i.}$	$\bar{X}_{i.}$
A_k	x_{k1}	x_{k2}	$\cdots$	x_{kj}	$\cdots$	x_{kl}	$T_{k.}$	$\bar{X}_{k.}$
계	T_{i1}	$T_{.2}$	$\cdots$	$T_{.j}$	$\cdots$	$T_{.l}$	$T_{..}$	
평균	$\bar{X}_{.1}$	$\bar{X}_{.2}$	$\cdots$	$\bar{X}_{.j}$	$\cdots$	$\bar{X}_{.l}$		$\bar{X}_{..}$

실측값 $x_{i,j}$ 로 나타난 확률변수를 $X_{i,j}$

$X_{i,j} = \mu + \alpha_i + \beta_j + \varepsilon_{i,j}$

μ : 이 자료전체의 모평균

α_i, β_j : 수준 A_i, B_j의 주효과를 가리키는 상수

$\varepsilon_{i,j}$: i, j와 관계없이 나타나는

실험오차로서 정규분포 $(0, \alpha^2)$

$$\bar{X}_{i.} = \frac{1}{n_i}\sum_{i=1}^{n_i} X_{ij} = \frac{1}{n_i}\sum_{i=1}^{n_i}(\mu + \alpha_i + \varepsilon_{ij})$$

$$= \mu + \alpha_i + \bar{\varepsilon}_{i.} \quad (\bar{\varepsilon}_{i.} = \frac{1}{n_i}\sum_{i=1}^{n_i}\varepsilon_{ij})$$

$$\bar{X}_{ij} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_j} X_{ij} = \mu + \bar{\alpha}_i + \bar{\varepsilon}_{i.} \quad (\bar{\alpha}_i = \frac{1}{n_i}\sum_{i=1}^{n_i}\alpha_i, \bar{\varepsilon}_{i.} = \frac{1}{n_i}\sum_{i=1}^{n_i}\varepsilon_{ij})$$

일원배치법과마찬가지로

$$S_T(\text{total variation}) = \sum_{i=1}^{k}\sum_{j=1}^{l}(X_{ij} - \bar{X}_{..})^2 \qquad (1)$$

$S_T(\text{total variation})$: 각변수 X_{ij}와 자료의전체평균 $\bar{X}_{..}$ 와의 차의 제곱의 합 즉 자료전체에대한 평방합

전변동 S_T 의자유도 $\nu_T = kl - 1$

(1)식을 다음과 같이 변형하면

$$S_T = \sum_i \sum_j \{(X_{ij} - \bar{X}_{i.}) + (\bar{X}_{i.} - \bar{X}_{..}) + (\bar{X}_{.j} + \bar{X}_{..}) + (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})\}^2$$

$$= l \sum_i (\bar{X}_{i.} - \bar{X}_{..})^2 + k(\bar{X}_{.j} + \bar{X}_{..})^2 + \sum_i \sum_j (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})^2$$

$$A간변동 S_A = l\sum_i(\bar{X}_{i.} - \bar{X}_{..})^2 \qquad 자유도 \nu_A = k-1$$

$$B간변동 S_B = k(\bar{X}_{.j} + \bar{X}_{..})^2 \qquad 자유도 \nu_B = l-1$$

$$오차변동 S_E = \sum_i\sum_j(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})^2$$

$$\therefore S_T = S_A + S_B + S_E$$

$$오차변동의자유도 \nu_E = \nu_T - \nu_A - \nu_B$$

$$= (kl-1)-(k-1)-(l-1) = (k-1)(l-1)$$

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}\ ,\quad \bar{X}_{i.} = \mu + \alpha_i + \bar{\varepsilon}_{i.}$$

$$\bar{X}_{.j} = \mu + \beta_j + \bar{\varepsilon}_{.j.}\ ,\quad \bar{X}_{..} = \mu + \bar{\varepsilon}_{..}$$

$$(\text{단}\ \bar{\varepsilon}_{i.} = \frac{1}{l}\sum_{j=1}^{l}\varepsilon_{ij}\ ,\ \bar{\varepsilon}_{.j.} = \frac{1}{k}\sum_{i=1}^{k}\varepsilon_{ij}\ ,\ \bar{\varepsilon}_{..} = \frac{1}{\mathrm{kl}}\sum_{j}\sum_{i}\varepsilon_{ij}\ )$$

$$S_E = \sum_{i=1}\sum_{j=1}(\varepsilon_{ij} - \bar{\varepsilon}_{i.} - \bar{\varepsilon}_{.j} - \bar{\varepsilon}_{..})^2$$

$$S_A = l\sum_{i=1}^{k}(\alpha_i + \bar{\varepsilon}_{i.} - \bar{\varepsilon}_{..})^2$$

$$S_B = k\sum_{i=1}^{l}(\beta_i + \bar{\varepsilon}_{.ji.} - \bar{\varepsilon}_{..})^2$$

가설 $H_{0\alpha}$: $\alpha_i = 0$

가설 $H_{0\beta}$: $\beta_j = 0$

$$V_E = \frac{S_E}{V_E} = \frac{S_E}{(k-1)(l-1)}$$

가설 $H_{0\alpha}$: $\alpha_i = 0$가 참이면

$$V_A = \frac{S_A}{V_A} = \frac{S_A}{k-1} \quad , \quad \text{분산비 } F = \frac{V_A}{V_E}$$

가설 $H_{0\beta}$: $\beta_j = 0$가 참이면

$$V_A = \frac{S_B}{V_B} = \frac{S_A}{l-1} \quad , \quad \text{분산비 } F = \frac{V_B}{V_E}$$

이는 분자의 자유도 l-1, 분모의자유도 (k-1)(l-1)인 F분포를 한다.

따라서 일원배치법의 경우와 마찬가지로 인자 A ,B의 각 수준의 평균의 균일성의 검정을 분산분석표에 의해 할 수 있다.

또한 각 수준의 평균의 차를 개별적으로 검정할때의 위험역을 일원배치법의 경우와 마찬가지로 결정할 수 있다.