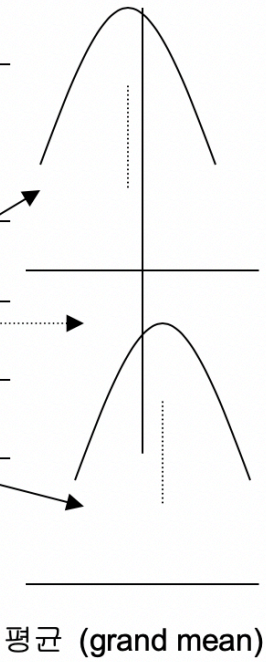
개념

상황

서로 독립인 모집단이 3개 이상인 경우 평균을 비교하거나(일원 분산 분석) 범주형 설명 변수(이를 요인, 처리 효과라 함)가 2개 이상인 인과 관계(이원, 다원 분산 분석) 분석을 분산 분석이라 한다. 분산 분석의 개념은 Fisher에 의해 정립되었으며, 분석 데이터는 주로 실험 계획에 의해 수집된다.

요인의 수준이 k인 일원 배치 실험과 k개의 모집단 중앙 위치에 대한 비교를 위한 자료 수집을 비교하면 다음과 같다. $k \geq 3$ 독립 모집단 평균 차이 검정에 이용된다. $k = 2$인 경우는 두 독립 모집단 t-검정과 동일하다.

	일원 배치 설계				모집단 중앙 위치 비교
요인	1	2	…	반복 수	
1	y_{11}	y_{12}	$\cdots y_{1n1}$	n_1	y_{11} y_{12} … y_{1n1}
2	y_{21}	y_{22}	$\cdots y_{2n2}$	n_2	y_{21} y_{22} … y_{2n2}
…	…	…		…	…
k	y_{k1}	y_{k2}	$\cdots y_{knk}$	n_k	y_{k1} y_{k2} … y_{1nk}



실험

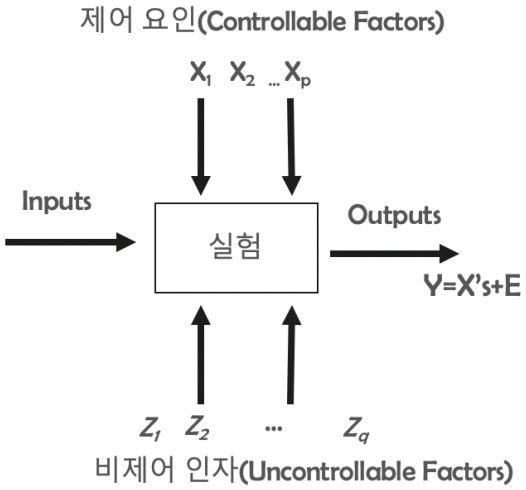
관심 대상에 대한 정보를 얻기 위한 계획된 실험 혹은 관측을 한다.

절대실험 absolute experiment

3G 서비스에 대한 고객 만족도 : 현상을 관찰(관측, 조사)하여 관심 대상에 어떤 현상이 나타나는지 분석

비교실험 comparative experiment

기존 마케팅 전략과 새로운 마케팅 전략 비교 : ¨관심 현상에 영향을 요인(factor)을 조절하여 반응(response) 변화 분석, 분산분석은 비교실험에 해당된다.



용어

물고기 사료 (기존 사료A, 새로 개발된 사료=B)의 몸무게 증가 효과를 보기 위하여 어항 4개(성장 환경 동일), 물고기 12마리(몸무게 유사, 성장 속도 유사)를 랜덤하게 선택하여, 각 어항에 3마리 물고기 배정 후, 2개 어항에 각각 A, B 사료를 주고 2주 후에 몸무게를 측정하였다.

- 실험단위: 처리(실험조건, 요인 수준)가 가해지는 최소단위 - 어항
- 관찰단위: 개체의 최소단위 ; 물고기
- 반응(response): 관심 대상의 측도, output, Y - 2주후 물고기 몸무게
- 요인(factor): 제어 가능하며 반응에 영향을 주는 인자 X - 사료
- 수준(level): 실험에 사용되는 요인의 값 - A, B

Control vs. Experimental group

- 통제집단 : 기존 사료(A)를 준 물고기 / 비교집단 : 새로 개발된 제품(B)을 실험하는 집단
- 사전 vs. 사후 검정 : 실험 전 후 물고기 몸무게에 관심이 있다면 짝진 표본 검정, 만약 사료 효과 보는 것이 주 실험설계 목적이라면 사전 몸무게는 공변량(covariate 비료 이외 사전 효과 제거) 역할을 한다.

Placebo 효과

위염 치료 인공적 가스 주입, 효과가 있는 것처럼 나타나 내과에서 한동안 사용, 즉, 치료를 받았기 때문에 통증이 완화되었을 것이라고 믿는 효과 (예) 노인 병원을 자주 찾는 이유와 동일함 - 병원 의사로부터 치료에 의한 효과보다 의사와 상담했다는 이유로 병이 낫는다.

(해결책) Blind 실험 : 위약과 치료약을 환자가 구별하지 못하게 하는 실험, 그러나 실험자가 안다면? 치료약 복용에 주 관심 - 이를 해결하는 방법 double blind (실험자도 어느 것이 위약인지 치료약인지 모르게 함)

실험설계 원리

Randomization (랜덤화)

- 실험단위의 배정과 실험순서 랜덤하게 결정 => 실험의 객관성 보장 : CRD
- 관심 요인 외에 기타 원인들의 영향이 실험결과에 미치지 않게 함
- 시간에 따라 변하는 인자의 효과나 경향을 실험을 시간대에 대해 균일하게 배치함으로써 약화시킬 수 있다.

Blocking(블록화)

- 랜덤화 불가능, 실험의 정도를 높인다.
- 동일한 성질을 가진 단위들의 집합 (Block)
- 실험전체를 시간적 혹은 공간적으로 분할하여 Block 을 만들어 주면 각 Block 내에서는 실험환경이 균일하게 되어 좀더 좋은 결과를 얻을 수 있다.

- Block 은 실험 계획 시 또 다른 독립변수(요인)로 취급해야 한다.
- 실험이 이틀에 걸쳐서 수행되었다고 하면 실험 일을 Block 이라 한다.

Replication(반복)

- 동일 처리를 2개 이상의 실험단위에 가함 => 실험오차 계산
- 실험조건을 처음부터 다시 setup 하여 실험하는 것.
- 실험의 재현성을 알아보기 위한 방법 : 실험결과의 신뢰성을 높일 수 있다.
- 오차 변동을 계산할 수 있다.

Repetition : 같은 조건에서 여러 번 반복 실험, 관측치는 평균 하나만 사용하게 된다. 이는 실험의 정도(precision)을 높이기 위함이다. 즉 집단 내 변동의 크기를 줄이는 역할을 한다.

실험설계 방법

CRD

요인의 수준이 개이고 각 수준의 반복 수를 라고 한다면 다음과 같은 실험 설계 표를 만들고 각 셀에 1-n까지 번호를 차례로 매긴다. 난수표에 의해 실험 순서를 정하고 실험을 실험 단위(unit)에 실시하면 된다. 그러므로 일원 분산의 실험 설계는 각 셀에 번호를 부여한 후 난수를 이용하여 실험 순서를 결정하므로 이를 완전 임의 실험 설계(CRD: Completely Randomized Design)라 부르기도 한다.

① A	② A	③ C
④ B	⑤ A	⑥ B
⑦ A	⑧ C	⑨ C
⑩ B	⑪ C	⑫ B

비료 종류(A, B, C)에 따른 벼 수확량의 차이를 알아보기 위하여 실험을 한다고 하자. 반복 수는 각 비료에 대해 4번을 한다면 실험을 위해 총 12곳의 경작지가 필요하다. 12 경작지가 땅의 비옥도가 동일하다면 CRD 방법을 이용하여 실험하면 된다. 난수표를 이용하여 임의로 비료를 배정할 수도 있지만 총 실험 수가 많지 않으므로 12 장의 종이에 각 번호를 적고 던진 후 하나씩 선택하여 실험을 배정하면 된다. 만약 1, 5, 2, 7, 12, 10 …, 3 순으로 선택하였다면 다음과 같다.

RBD

만약 경작지 아래 개천이 흐른다면 물로부터 거리에 따라 땅의 비옥도의 차이는 있을 것이므로 CRD 방법은 적합하지 않다. 물의 위치에 따라 경작지를 나누고(블록화:block) 각 블록 내에서 각 비료를 하나씩 임의(randomized)로 배정하는 실험 계획을 실시하면 된다. 이를 Randomized Block Design이라 한다. 행이 (물에서 먼 정도) 블록이 되는 RBD 설계의 예이다.

A	B	C
A	C	B
C	A	B
B	C	A

일원분산분석 모형

모형 및 가정

$y_{ij} = \mu + A_i + e_{ij}$, 가정 $e_{ij} \sim N(0,\sigma^2)$

- i : 요인A의 수준 첨자 $i = 1,2,...,k$, k : 요인의 수준 수, 모집단 개수
- j : i-번째 요인의 반복 첨자
- y_{ij} : 요인 A의 수준 i 의 j-번째 관측치, 수준 i의 반복 수 $j = 1,2,...,n_{i+}$
- 총 데이터 크기 : $\sum\sum n_{ij} = n$

- μ : 모집단 총평균
- A : 요인 A의 주효과를 나타내는 알파벳, 실험 처리효과라고도 한다.
- $\mu + A_i = \mu_i$: 수준 i 모집단 평균
- 오차항은 독립이며 정규분포를 따르고 분산은 동일하다.

모수와 추정 (MVUE)

모수 : $\mu, \mu_1, \mu_2, \ldots, \mu_k$ ($k+1$개)

총평균 점추정치 : $\hat{\mu} = \dfrac{\sum_i \sum_j y_{ij}}{n} = \bar{\bar{y}}$

요인 수준 i–(모집단)- 평균 점추정치: $\hat{\mu}_i = \dfrac{\sum_j y_{ij}}{n_i} = \bar{y}_i$

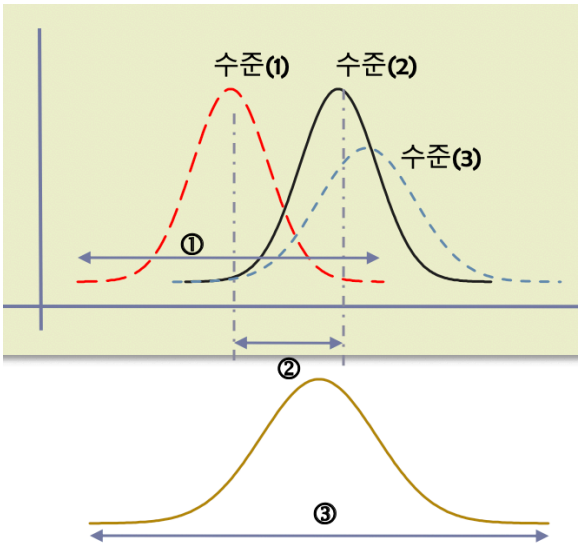
OLS(최소자승합) 추정치 $min_{\mu,A_i} \sum\sum (y_{ij} - \hat{y}_{ij})^2 \Rightarrow \underline{\hat{\beta}} = (X'X)^{-1}X'\underline{y}$

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ A_1 \\ A_2 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \end{bmatrix} \Leftrightarrow \underline{y} = X\underline{\beta} + \underline{e}$$, 요인 2개, 반복 2회

모수($\underline{\beta}$)는 3개이나 $rank(X) = 2$ (독립인 방정식 개수)이므로 모수 3개를 모두 추정할 수 없다. 하여, 모수의 수를 줄이는 제약 조건 하에서 모수를 추정한다. $\sum \mu_i = 0$ 가정 하에 모수를 추정한다.

변동분해

총변동(③)=집단간 변동(②)+집단내 변동(①)



③ : 반응변수의 총변동(SST, Total Sum of Squares)

$$SST = \sum_i \sum_j (y_{ij} - \bar{\bar{y}})^2$$

② : 요인(집단간, 설명) 변동 (SS**Between**)

$$SSB = \sum_i \sum_j (\bar{y}_i - \bar{\bar{y}})^2$$

① : 오차(집단내) 변동 (**SS**Error, **SSWithin**)

$$SSE = \sum_i \sum_j (y_{ij} - \bar{\bar{y}})^2$$

평균변동 Mean Sum of Squares

변동(Sum of Squares) 값을 자유도로 나눈 값, 변동의 평균적 개념

자유도 : 데이터가 가진 정보이다. 데이터가 n개인 경우 데이터가 가진 정보는 n개이다. 만약 이 데이터에서 평균을 계산하여 주어졌다면 다른 하나의 값을 없애도 알 수 있으므로 데이터가 가진 정보는 $(n-1)$이다.

총변동(SST)의 경우 총 데이터 크기가 n_{++}이고 $\bar{\bar{y}}$를 하나 추정했으므로 $(n-1)$이 자유도이다.

요인변동(SSB)의 경우 요인 수준 개수(집단 개수 k) 평균에서 $\bar{\bar{y}}$을 하나 추정했으므로 $(k-1)$이 자유도이다.

그리고 Cochran 정리(변동 자유도의 합의 동일하다)에 의해 오차변동($SSE,\ SSW$) 자유도는 $(n-k)$이다.

평균요인변동(Mean Squared Between)

$$MSB = \frac{SSB}{k-1}$$

평균오차변동(Mean Squared Error)

$$MSE = \frac{SSE}{n-k}$$

변동의 분포

오차의 가정 $e_{ij} \sim N(0,\sigma^2)$으로부터 $y_{ij} \sim N(\mu + A_i, \sigma^2)$이므로 다음이 증명된다.

$$\frac{SSB}{\sigma^2} \sim \chi^2(k-1),\ \frac{SSE}{\sigma^2} \sim \chi^2(n-k)$$

오차 분산의 추정치 : $\hat{\sigma^2} = MSE$

변동비의 분포

$\frac{SSB}{\sigma^2} \sim \chi^2(k-1), \frac{SSE}{\sigma^2} \sim \chi^2(n-k)$이므로 서로 독립인 카이제곱 분포의 비는 F-분포를 따르므로

$$TS = \frac{SSB/(k-1)}{SSE/(n-k)} \sim F(k-1, n-k)$$가 성립한다.

평균변동 기대값

$$E(MSE) = \sigma^2$$

$$E(MSB) = \sigma^2 + \frac{\sum n_i(\mu_i - \mu)^2}{k-1}$$

만약 모든 집단의 평균이 동일하다면 $(\mu_i - \mu) = 0$이므로 $E(MSE) = E(MSB)$가 된다.

그러므로 $TS = \frac{SSB/(k-1)}{SSE/(n-k)} = 1$이 된다. 만약 집단 간 평균의 차이가 커지면 TS값은 커지게 된다.

분산분석표

귀무가설 : 요인 수준별 평균은 동일하다(집단 평균은 동일하다). $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$

대립가설 : 적어도 하나의 집단 평균은 다르다. <=> 모든 집단 평균이 동일한 것은 아니다.

독립인 두 모집단

검정통계량 : $TS = \frac{SSB/(k-1)}{SSE/(n-k)} \sim F(k-1, n-k)$

요인	자승합 Sum of Squares	자유도 df	평균자승합 Mean Squares	F-통계량
집단간 변동 처리변동 Between	SSB $\sum_i \sum_j (\bar{y}_i - \bar{\bar{y}})^2$	$k-1$	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSE}$
오차변동 Error	SSE, SSW\ $\sum_i \sum_j (y_{ij} - \bar{\bar{y}})^2$	$n-k$	$MSE = \frac{SSE}{n-k}$	$\sim F(k-1, n-k)$
수정총합 Corrected Total	SST $\sum_i \sum_j (y_{ij} - \bar{\bar{y}})^2$	$n-1$		

사후 검정 (Post-hoc test) 혹은 다중 비교 (multiple comparison)

개념

분산분석의 F-검정은 단지 귀무가설 즉 전체적인 차이를 검정하는 것이다. 그러므로 수준별 차이(pairwise: 예:)가 있는지 혹은 수준의 선형 결합 대비(contrast: 예:)의 차이가 있는지 검정할 필요가 있는데 이를 사후 검정 혹은 다중 비교(대비 포함)라 한다. 사후 검정이므로 비록 F-검정 결과와 관계없이 (귀무가설을 채택하더라도) 시행하게 된다.

다중 비교에서는 여러 개의 가설을 동시에 검정하므로 유의수준을 조정해야 한다.

이를 조정된 실험 유의수준 (controlled experimental error rate)이라 하고 $1-(1-\alpha)^c$이다. 여기서 c는 가설 수를 의미한다. $c=1$인 경우 유의수준은 $\alpha=0.05$이지만, 요인의 수준 개수가 $k=4$개인 경우 쌍체 비교 시 검정되어야 하는 귀무가설 개수는 ${}_4C_2=6$개이므로 조정된 유의수준은 $1-(1-0.05)^6=0.264$로 매우 높다. 즉, 6개 귀무가설을 동시에 검증하는 경우 유의수준은 5%가 아니라 26.4%가 된다.

Fisher's Least Significant Difference : $t(\alpha/2, n_i+n_j-2)MSE\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}$

요인 수준 i와 j의 평균 차이가 위의 값 이상이어야 유의하다고 판단한다. pairwise (두 수준별 평균 비교) 검정에 사용하나 이는 다중 비교에 해당되지는 않는다. 독립인 두 모집단 평균차이 검정과 동일하나 통합분산 대신 MSE(집단 통합분산 개념)을 사용한다.

Tukey HSD(honestly significant difference) procedure $HSD=\frac{q_{\alpha;k,n-k}}{\sqrt{2}}\sqrt{MSE(1/n_{i.}+1/n_{j.})}$

$q_{\alpha,k,N-k}$는 두 모집단 평균 차이검정의 $t-$검정통계량과 동일한 개념으로 $q_s=\frac{\bar{y}_{max}-\bar{y}_{min}}{SE}$의 표준화 범위 분포를 가정한다.

보수적인(귀무가설 기각하지 않음) 방법이다. 자연 과학에서 가장 많이 이용한다.

Scheffe's S method

대비(contrast)까지 고려하여 유의수준을 고려한 다중 비교 방법으로 (Tukey >Scheffe > Duncan 순으로 보수적) 사회 과학 분야에서 주로 사용

Duncan Multiple range test

Tukey 방법과 매우 유사하나 수준별 표본 평균을 크기 순으로 나열하여 차이가 가장 큰 것을 비교해 가면서 유의수준을 $1-(1-\alpha)^r$으로 조정해 가면서 검정한다. r은 검정 단계 순서이다.

귀무가설을 기각할 확률이 매우 높아 자주 사용하지 않는다.

Dunnett's procedure

처리 효과의 수준 하나가 control (실험 집단)인 경우 (예: placebo 집단, 교육을 하지 않는 집단, 이전 약 투여 집단) 이 집단과 다른 집단들을 pairwise 비교한다.

대비 (contrast) : 집단 간 차이 분석

$$Q = \sum_{i}^{k} c_i \mu_{i'} \sum c_i = 1$$

만약 두 집단 (i, j) 평균 비교인 경우에는 $c_i = 1, c_j = -1$이고 나머지는 $c_i = 0$이다. $Q = c_i - c_j$

만약 집단이 3개이고 처음 2개집단과 3번째 집단의 평균 차이를 비교한다면, $c_1 = 1/2, c_2 = 1/2, c_3 = -1$ 으로 설정한다.

검정통계량 : $TS = \dfrac{n(\sum c_i \bar{y}_i)^2 / \sum c_i^2}{MSE} \sim F(m-1, n-k), m = c_i$가 0이 아닌 개수

사례 실습

분꽃 3종(Versicolor, Setosa, Virginica)에 대한 sepal(꽃받침 조각) 길이, 넓이, petal(꽃잎) 길이, 넓이 데이터이다.

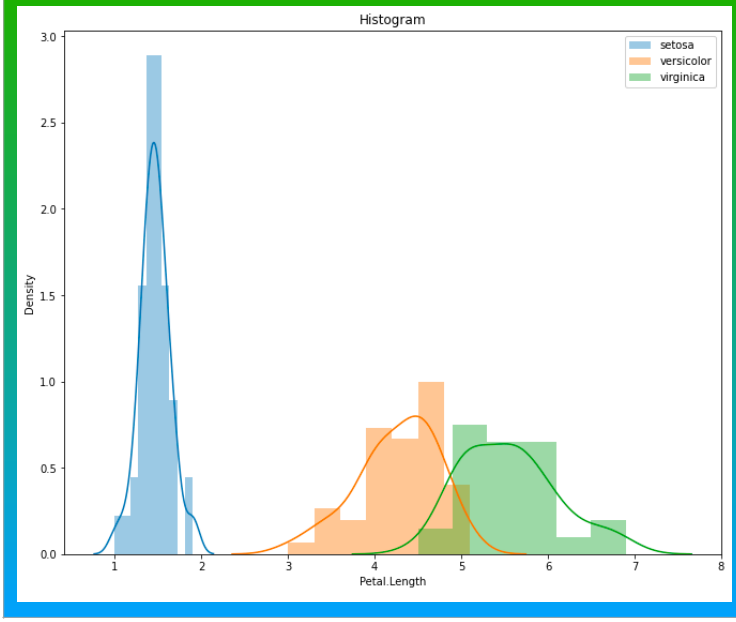
[분석 사례] 분꽃 3종에 따른 꽃잎(petal) 길이의 차이가 있는지 알아보자.

```
import pandas as pd
iris=pd.read_csv('https://vincentarelbundock.github.io/Rdatasets/csv/datasets/iris.csv')
iris.info()
```

```
RangeIndex: 150 entries, 0 to 149
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Unnamed: 0     150 non-null    int64
 1   Sepal.Length   150 non-null    float64
 2   Sepal.Width    150 non-null    float64
 3   Petal.Length   150 non-null    float64
 4   Petal.Width    150 non-null    float64
 5   Species        150 non-null    object
```

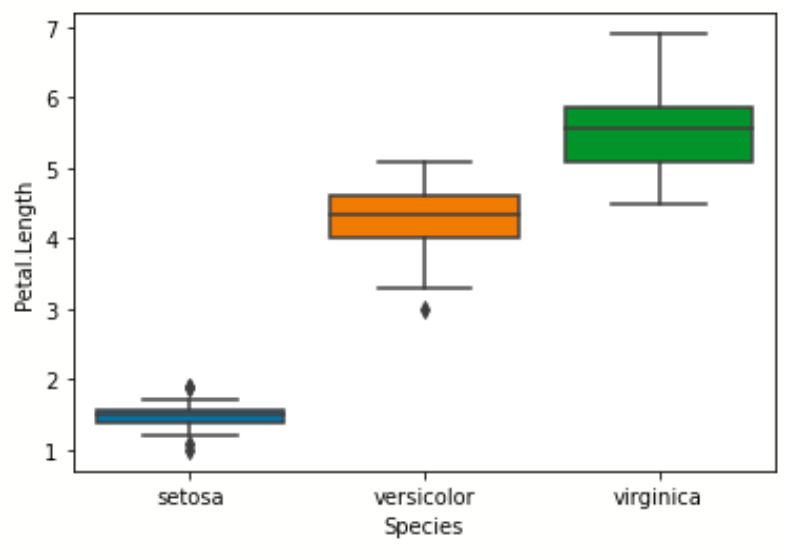
히스토그램 그리기

```
import matplotlib.pyplot as plt
import seaborn as sns
f, ax = plt.subplots(figsize=(11,9))
sns.distplot(iris.loc[iris['Species']=='setosa','Petal.Length'], ax=ax, label='setosa')
sns.distplot(iris.loc[iris['Species']=='versicolor','Petal.Length'], ax=ax, label='versicolor')
sns.distplot(iris.loc[iris['Species']=='virginica','Petal.Length'], ax=ax, label='virginica')
plt.title('Histogram'); plt.legend()
```



나무상자 그리기 : 이상치 진단

```
import seaborn as sns
ax=sns.boxplot(x='Species',y='Petal.Length',data=iris)
```



setosa, versicolor 품종에는 이상치가 존재한다. 분산분석 경우에는 이상치를 제외하는 것이 적절함

요인별 기초통계량

품종별 꽃잎 길이는 VI종이 가장 길고(평균 0.552) SE종이(평균 0.174) 가장 짧다.

```
import pandas as pd
iris=pd.read_csv('https://vincentarelbundock.github.io/Rdatasets/csv/datasets/iris.csv')
iris.info()
```

Species	count	mean	std	min	25%	50%	75%	max
setosa	50.0	1.462	0.173664	1.0	1.4	1.50	1.575	1.9
versicolor	50.0	4.260	0.469911	3.0	4.0	4.35	4.600	5.1
virginica	50.0	5.552	0.551895	4.5	5.1	5.55	5.875	6.9

분산분석 anova

```
import statsmodels.api as sm
from statsmodels.formula.api import ols
y=iris['Petal.Length']
results=ols('y~C(Species)',data=iris).fit()
sm.stats.anova_lm(results,typ=2)
```

	sum_sq	df	F	PR(>F)
C(Species)	437.1028	2.0	1180.161182	2.856777e-91
Residual	27.2226	147.0	NaN	NaN

[요인이 A, B 두 개인 경우]

Type 1 : (sequential SS) 모형에 삽입된 요인 순서대로 자승합 출력 ($SS(A)$-$SS(B|A)$-$SS(AB|A,B)$)

Type 2 : $SS(A|B)$, $SS(B|A)$ 주효과만 관심

Type 3 : (partial SS) $SS(A|B,AB)$, $SS(B|A,AB)$, $SS(AB|A,B)$

모형 $y_{ij} = \mu + A_i + e_{ij}$

```
results.summary()
```

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.4620	0.061	24.023	0.000	1.342	1.582
C(Species)[T.versicolor]	2.7980	0.086	32.510	0.000	2.628	2.968
C(Species)[T.virginica]	4.0900	0.086	47.521	0.000	3.920	4.260

그러므로 setosa 꽃잎 길이 평균 1.462(기초 통계량 평균과 동일), versicolor 품종 평균 길이는 1.462+2.798=4.26, 그리고 virginica 꽃잎 길이 평균은 1.462+4.09=5.552이다.다중비교 : Tukey HSD

```
from statsmodels.stats.multicomp import pairwise_tukeyhsd
from statsmodels.stats.multicomp import MultiComparison
mc = MultiComparison(iris['Petal.Length'],iris['Species'])
mc_results = mc.tukeyhsd()
print(mc_results)
```

```
   Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
  group1     group2   meandiff p-adj lower  upper  reject
-----------------------------------------------------
    setosa versicolor     2.798 0.001 2.5942 3.0018   True
    setosa  virginica      4.09 0.001 3.8862 4.2938   True
versicolor  virginica     1.292 0.001 1.0882 1.4958   True
-----------------------------------------------------
```

SE종과 VE종 간 꽃잎 길이 차이(meandiff +이므로 VE종의 꽃잎 길이 큼)는 유의함 (True) – 첫째 행
둘째, 셋째 모두 유의한 차이가 있으므로 VI종(5.55) > VE종(4.26) > SE종(1.46) 순으로 꽃잎의 길이가 유의한 차이를 보이고 있다.

이원 분산분석 two-way anova

모형 및 가정

$y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + e_{ijk}$, 가정 $e_{ij} \sim N(0,\sigma^2)$

- i : 요인 A의 수준 첨자, j : 요인 B의 수준 첨자, k : 반복 첨자
- 요인 A 수준 개수 a, 요인 B 수준 개수 b
- y_{ijk} : 요인 A i 수준, 요인 B의 수준 j의 k-번째 관측치

- μ : 모집단 총평균
- 주효과 main effect : A, B
- 상호효과 interaction effect : AB

주효과, 상호효과 추정

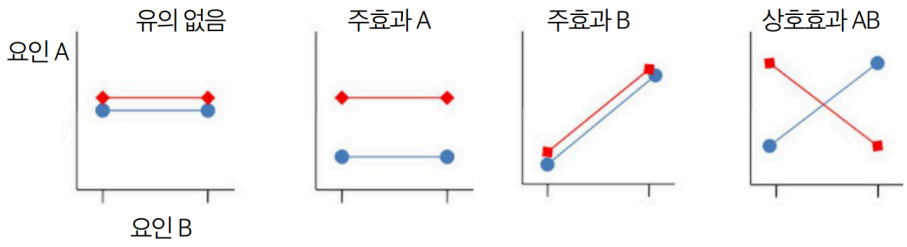
주효과

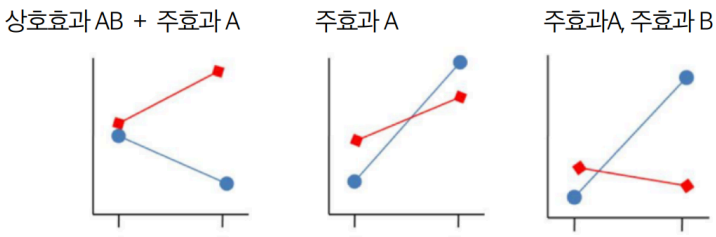
- 요인 A의 i 집단(수준) 모평균 $\mu + A_i$ - 모수 개수 a 개 : 점추정치 : $(\hat{\mu + A_i}) = \sum_j \sum_k \frac{y_{ijk}}{n_{i..}} = \bar{y}_{i..}$
- 요인 B의 j 집단 모평균 $\mu + B_j$ - 모수 개수 b 개 :점추정치 : $(\hat{\mu + B_j}) = \sum_j \sum_k \frac{y_{ijk}}{n_{.j.}} = \bar{y}_{.j.}$

상호효과

- 요인 A의 i 집단(수준), 요인 B의 j 집단 모평균 $\mu + (AB)_{ij}$ - 모수 개수 ab 개

점추정치 : $(\mu + (\hat{AB})_{ij}) = \sum_k \frac{y_{ijk}}{n_{ij.}} = \bar{y}_{ij.}$





총변동분해

총변동 $SST = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{i..})^2$

요인 A 변동 $SSA = \sum_i \sum_j \sum_k (\bar{y}_{i..} - \bar{y}_{...})^2$

요인 B 변동 $SSB = \sum_i \sum_j \sum_k (\bar{y}_{.j.} - \bar{y}_{...})^2$

요인 AB 변동 $SSAB = \sum_i \sum_j \sum_k (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$

잔차 변동 $SSE = \sum_i \sum_j \sum_k (\bar{y}_{ijk} + \bar{y}_{ij.})^2$

요인	자승합 Sum of Squares	자유도 df	평균자승합 Mean Squares	F-통계량
요인 A 주효과	SSA	$a-1$	MSA	$\frac{MSA}{MSE}$
요인 B 주효과	SSB	$b-1$	MSB	$\frac{MSB}{MSE}$
요인 AB 상호효과	$SSAB$	$(a-1)(b-1)$	$MSAB$	$\frac{MSAB}{MSE}$
오차변동	SSE	차이	MSE	
수정총합 Corrected Total	SST	$n-1$		

사례 실습

스텝업 운동 후 힘든 정도를 파악하기 위하여 운동 후 맥박과 일반 상태에서 맥박을 측정하였다.

- 계단 높이 수준=2 : Height: 0 if step at the low (5.75") height, 1 if at the high (11.5") height
- 운동 빈도 수준=3 : Frequency: the rate of stepping. 0 if slow (14 steps/min), 1 if medium (21 steps/min), 2 if high (28 steps/min)
- 쉬는 상태의 맥박 : Rest_HR: the resting heart rate of the subject before a trial, in beats per minute
- 운동 후 맥박 HR: the final heart rate of the subject after a trial, in beats per minute

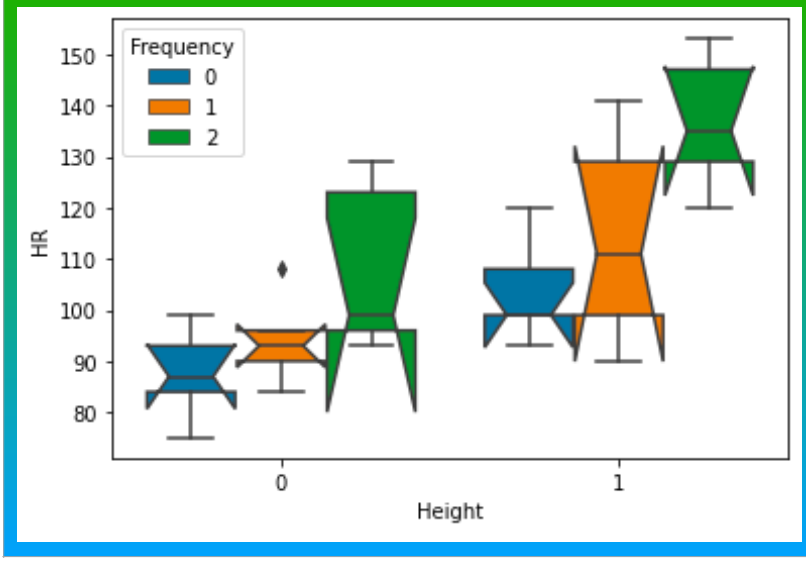
[운동 후 맥박에 계단 높이, 운동빈도] 요인이 영향을 주는가?

```
import pandas as pd
df=pd.read_csv('http://203.247.53.31/Stat_Notes/example_data/Heart.csv')
df.info()
```

```
RangeIndex: 30 entries, 0 to 29
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Height     30 non-null     int64
 1   Frequency  30 non-null     int64
 2   RestHR     30 non-null     int64
 3   HR         30 non-null     int64
```

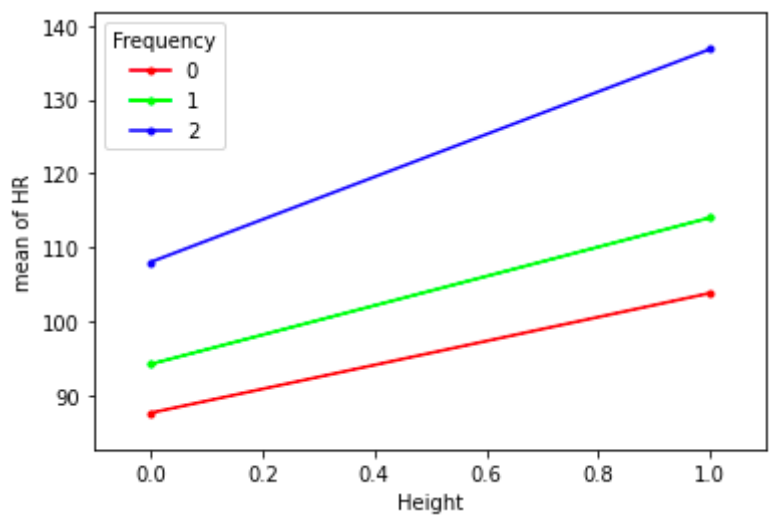
Box Plot 그리기

```
import seaborn as sns
sns.boxplot(x="Height",y="HR",hue='Frequency',notch=True,data=df)
plt.show()
```



평균 플롯 그리기

```
from statsmodels.graphics.factorplots import interaction_plot
fig = interaction_plot(df['Height'],df['Frequency'],df['HR'])
```



평균, 표준편차

```
df.groupby(['Frequency','Height'])['HR'].describe()
```

		count	mean	std	min	25%	50%	75%	max
Frequency	**Height**								
0	**0**	5.0	87.6	9.099451	75.0	84.0	87.0	93.0	99.0
	1	5.0	103.8	10.521407	93.0	99.0	99.0	108.0	120.0
1	**0**	5.0	94.2	8.899438	84.0	90.0	93.0	96.0	108.0
	1	5.0	114.0	21.000000	90.0	99.0	111.0	129.0	141.0
2	**0**	5.0	108.0	16.703293	93.0	96.0	99.0	123.0	129.0
	1	5.0	136.8	13.349157	120.0	129.0	135.0	147.0	153.0

분산분석표

- 계단 높이 : 심장박동에 유의한 영향을 미치며 계단 높이가 높을수록 박동이 높아진다. 유의확률 <0.001
- 운동 빈도 : 운동빈도 많을수록 박동 높아진다. 유의하다. 유의확률<0.001
- (계단 높이)*(운동 빈도) 상호효과는 유의하지 않음

```
import statsmodels.api as sm
from statsmodels.formula.api import ols
y=df['HR']
results=ols('y~C(Frequency)+C(Height)+C(Frequency):C(Height)',data=df).fit()
sm.stats.anova_lm(results,typ=2)
```

	sum_sq	df	F	PR(>F)
C(Frequency)	3727.8	2.0	9.551115	0.000888
C(Height)	3499.2	1.0	17.930822	0.000291
C(Frequency):C(Height)	210.6	2.0	0.539585	0.589898
Residual	4683.6	24.0	NaN	NaN

```
results.summary()
```

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	87.6000	6.247	14.022	0.000	74.706	100.494
C(Frequency)[T.1]	6.6000	8.835	0.747	0.462	-11.635	24.835
C(Frequency)[T.2]	20.4000	8.835	2.309	0.030	2.165	38.635
C(Height)[T.1]	16.2000	8.835	1.834	0.079	-2.035	34.435
C(Frequency)[T.1]:C(Height)[T.1]	3.6000	12.495	0.288	0.776	-22.188	29.388
C(Frequency)[T.2]:C(Height)[T.1]	12.6000	12.495	1.008	0.323	-13.188	38.388

공분산분석 ANOCOVA ANalysis Of COVarince Analysis

공변량 covariate

분산분석 모형에서 종속변수 값에 대한 요인들의 유의성 검정을 제대로 하기 위해 고려되는 변량(측정형 변수)

일반적으로 종속변수의 실험 전 값이다. (예) 교육효과에서의 사전점수

공변량은 관심의 대상이 아니라 요인의 유의성 검정을 정확하기 위하여 고려함 : 항상 유의하다.

2원 공분산분석

- (계단 높이), (운동 빈도) - 요인
- 종속변수 : 운동 후 심장박동
- 공변량 : 운동 전 심장박동

```
import pandas as pd
df=pd.read_csv('http://203.247.53.31/Stat_Notes/example_data/Heart.csv')
import statsmodels.api as sm
from statsmodels.formula.api import ols
y=df['HR']
results=ols('y~C(Frequency)+C(Height)+C(Frequency):C(Height)
+RestHR',data=df).fit()
sm.stats.anova_lm(results,typ=2)
```

운동빈도, 계단높이 주효과만 유의하고 상호효과는 유의하지 않다.

	sum_sq	df	F	PR(>F)
C(Frequency)	4019.033862	2.0	17.126951	0.000028
C(Height)	1799.423568	1.0	15.336342	0.000693
C(Frequency):C(Height)	213.719291	2.0	0.910756	0.416242
RestHR	1984.994148	1.0	16.917945	0.000425
Residual	2698.605852	23.0	NaN	NaN