ANOVA의 이해

Tutor: 장재호 (jaehochang@konkuk.ac.kr)

Tutee: 오지선

1. 기초확률론

학습목표

- 확률변수의 개념을 이해한다.
- 확률변수의 독립을 이해한다.
- 확률변수의 기대값을 이해한다.
- **정규분포**를 이해한다.

1.1. 확률변수

정의: 확률변수는 **사건**을 **숫자**로 대응시키는 함수이다.

예) 동전을 던져서 본 윗면이 X면?

$$X : \{\text{앞면}, \text{뒷면}\} \to \{0, 1\}$$

$X(\{\text{앞면}\}) = 1$, $X(\{\text{뒷면}\}) = 0$

예) 축구선수가 슛을 찼을때 결과가 X면?

$$X : \{\text{노골}, \text{골}\} \to \{0, 1\}$$

$X(\{\text{노골}\}) = 0$, $X(\{\text{골}\}) = 1$

예제 1.1.1.

X = A, B, C, D, E 총 5명의 사람중 대회에 참여할 사람의 수

$$X : \{\{\emptyset\}, \{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{A, B\}, \{A, C\}, \cdots, \{A, B, C, D, E\}\} \to \{0, 1, 2, 3, 4, 5\}$$

$X(\{\emptyset\}) = 0$

$X(\{A\}) = X(\{B\}) = X(\{C\}) = X(\{D\}) = X(\{E\}) = 1$

$X(\{A, B\}) = X(\{B, C\}) = \cdots = X(\{D, E\}) = 2$

$X(\{A, B, C\}) = X(\{A, B, D\}) = \cdots = X(\{C, D, E\}) = 3$

$X(\{A, B, C, D\}) = X(\{A, B, C, E\}) = \cdots = X(\{B, C, D, E\}) = 4$

$X(\{A, B, C, D, E\}) = 5$

1.2. 확률변수의 독립

두 확률변수 X, Y는 다음을 만족하면 서로 독립이라고 한다.

$$P(X, Y) = P(X)P(Y)$$

예제 1.2.1.

동전을 두 번 던질 때 윗면을 X_1, X_2라 하자. 앞면을 1, 뒷면을 0이라 하자.

$$P(X_1 = 1, X_2 = 1) = \frac{1}{4} = \frac{1}{2} \times \frac{1}{2} = P(X_1 = 1)P(X_2 = 1)$$
$$P(X_1 = 1, X_2 = 0) = \frac{1}{4} = \frac{1}{2} \times \frac{1}{2} = P(X_1 = 1)P(X_2 = 0)$$
$$P(X_1 = 0, X_2 = 1) = \frac{1}{4} = \frac{1}{2} \times \frac{1}{2} = P(X_1 = 0)P(X_2 = 1)$$
$$P(X_1 = 0, X_2 = 0) = \frac{1}{4} = \frac{1}{2} \times \frac{1}{2} = P(X_1 = 0)P(X_2 = 0)$$

따라서 X_1, X_2는 서로 독립이다.

1.3. 확률변수의 기대값

확률변수의 기대값(Expectation)은 다음과 같이 정의된다. 먼저 이산형(discrete) 확률변수 X의 기대값은

$$E(X) := \sum_x x \cdot p_X(x)$$

여기서 $p_X(x)$는 확률변수 X의 확률질량함수(probability mass function, pmf)라 한다. 연속형(continuous) 확률변수의 Y의 기대값은 다음과 같이 정의된다.

$$E(Y) := \int_y y \cdot f_Y(y)dy$$

여기서 $f_Y(y)$는 확률변수 Y의 확률밀도함수(probability density function, pdf)라 한다.

예제 1.3.1.

확률변수 X를 주사위를 하나 던질 때 나온 눈이라 하자. 이때 확률변수 X의 기대값은

$$E(X) = \sum_x x \cdot p_X(x) = 1 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = \frac{21}{6} = \frac{7}{2} = 3.5$$

위 예제에서 보듯이, 기대값은 평균(mean)을 일반화한 개념이라고 볼 수 있다.

예제 1.3.2.

확률변수 Y가 다음 연속형 확률분포를 따른다고 하자.

$$f_Y(y) = y, \ 0 \le y \le 1$$

이때 Y의 기대값을 구하면

$$E(Y) = \int_0^1 y \cdot y dy = \frac{y^3}{3}\Big|_0^1 = \frac{1}{3}$$

확률변수의 기대값(Expectation)은 다음과 같은 형태로 일반화(generalize)할 수 있다.

$$E[g(X)] := \sum_x g(x) \cdot p_X(x)$$

$$E[g(Y)] := \int_y g(y) \cdot f_Y(y)dy$$

확률분포의 분산(Variance)는 다음과 같이 계산한다.

$$Var(X) = E[\{X - E(X)\}^2] = E(X^2) - E(X)^2$$

1.4. 정규분포와 그 성질

정규분포는 연속형 확률분포로서, 다음과 같은 확률밀도함수를 가진다.

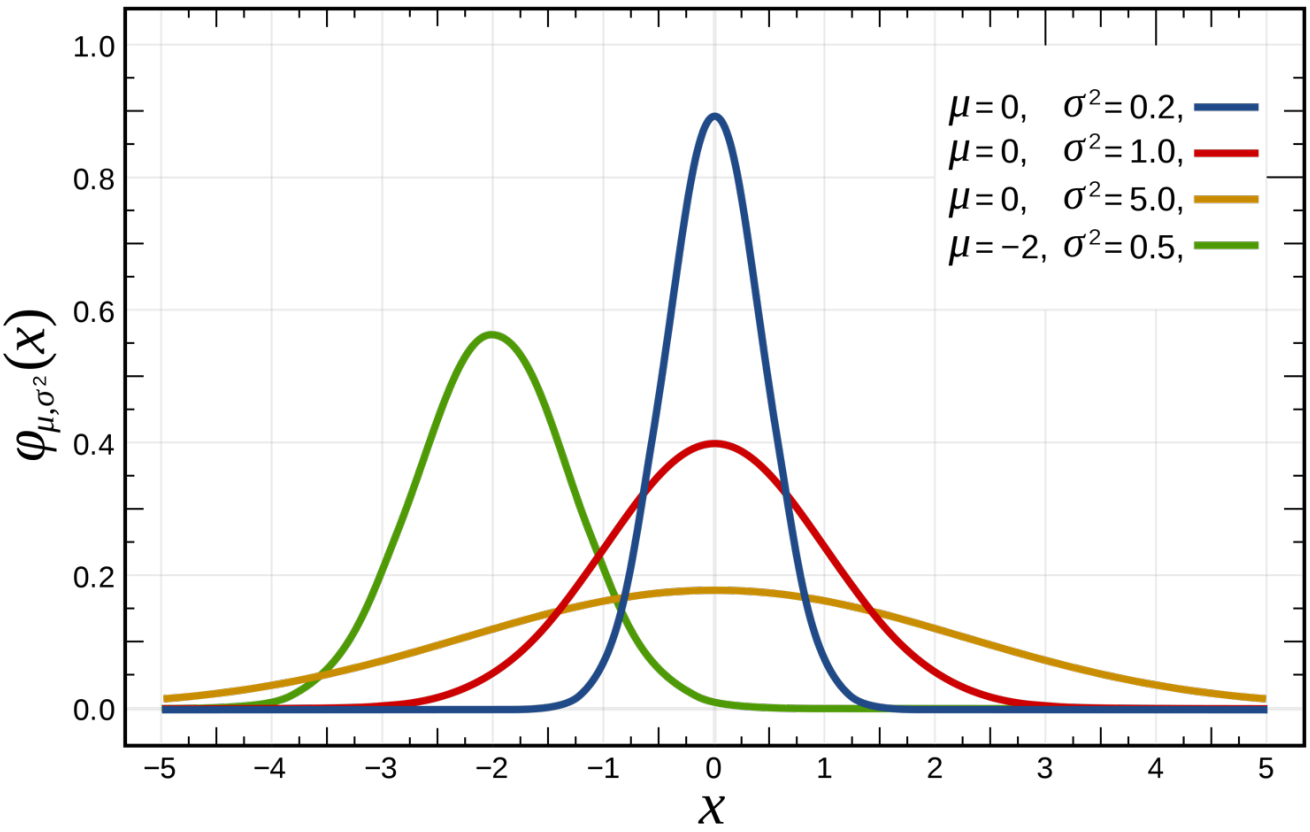
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

여기서 x는 실수의 어떤 값도 될 수 있으며, 음의 무한대나 양의 무한대도 허용한다. μ, σ^2는 각각 평균, 분산을 의미한다.

확률변수 X가 정규분포를 따를 때 다음과 같이 표현한다.

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

정규분포는 다음과 같은 형태를 가진다.



1.4.1. 정규분포의 성질

확률변수 X가 평균이 μ, 분산이 σ^2인 정규분포를 따른다 하자. 이때 다음이 성립한다.

- $X+a \sim \mathcal{N}(\mu+a, \sigma^2)$, 즉 정규분포는 평행이동이 가능하다.
- $b \cdot X \sim \mathcal{N}(\mu, b^2\sigma^2)$, 즉 정규분포는 스케일링(scaling)이 가능하다.
- 위 두 조건을 만족하는 확률분포들을 모은 집합을 **Location−scale family**라 한다.
- 확률분포의 분산은 분포의 평행이동에 영향을 받지 않는다. 즉, $a+bX \sim \mathcal{N}(\mu+a, b^2\sigma^2)$
- 정규분포는 평균에 대해서 대칭이다. 즉, $P(X<\mu) = P(X>\mu) = \frac{1}{2}$
- 확률변수 Y가 평균이 μ, 분산이 σ^2인 정규분포를 따르고 X와 서로 독립이라 하자. 이때 다음이 성립한다.

$$X+Y \sim \mathcal{N}(2\mu, 2\sigma^2)$$

예제 1.4.1.

Let $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. Define the sample mean $\bar{X}_n := \sum_{i=1}^n X_i/n$. Then

$$\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n).$$

Therefore,

$$\sqrt{n}(\bar{X}_n - \mu)/\sigma \sim \mathcal{N}(0,1).$$

2. ANalysis Of VAriance

학습목표

- 분산분석의 모형을 이해한다.
- 분산분석의 이론적 가정을 이해한다.
- 분산분석의 가설을 이해한다.
- 사후분석을 이해한다.

2.1. ANOVA 모형

일반적인 실험(experiment)은 세 가지의 요소를 가진다.

- 종속변수(Dependent Variable) = Y
- 처치(Treatment) = T
- 요인(Factor) = F

예제 2.1.1.

* 신약의 임상실험

- Y = vital signs
- T = Placebo, Treated
- F = Gender, Age bracket, etc.

* 토지시범사업의 부동산가격 인하효과 검증

- Y = 부동산 가격
- T = 비사업지역, 사업지역
- F = 행정단위, 소득수준 범주(저/중/고), 인구수 범주(소/중/대)

* ADHD case study, Pearson et al. (2003)

- Y = cognitive performance: measured by the number of correct responses to the Delay of Gratification or DOG task
- T = different dosages
- This is a repeated-measures design because each participant performed the task after each dosage.

Descriptions of Variables

Variable	Description
d0	Number of correct responses after taking a placebo
d15	Number of correct responses after taking .15 mg/kg of the drug
d30	Number of correct responses after taking .30 mg/kg of the drug
d60	Number of correct responses after taking .60 mg/kg of the drug

위와 같이 처치만 주어진 모형을 일원배치 분산분석(One-way ANOVA) 모형이라 한다. 그러나, 대부분의 실험들은 처치만 고려하진 않는다. 그 이유는 다양한 요인들이 실험결과에 영향을 미칠 수 있기 때문이다. 가령, 신약의 효능은 당연히 연령대, 사람의 컨디션에 영향을 받을 것이다. 따라서 이원배치 분산분석 혹은 다원배치 분산분석이 일반적으로 이용되고 있다.

예제 2.1.2.

Kirchhoefer(1979)

Lab 1	Lab 2	Lab 3	Lab 4	Lab 5	Lab 6	Lab 7
4.13	3.86	4.00	3.88	4.02	4.02	4.00
4.07	3.85	4.02	3.88	3.95	3.86	4.02
4.04	4.08	4.01	3.91	4.02	3.96	4.03
4.07	4.11	4.01	3.95	3.89	3.97	4.04
4.05	4.08	4.04	3.92	3.91	4.00	4.10
4.04	4.01	3.99	3.97	4.01	3.82	3.81
4.02	4.02	4.03	3.92	3.89	3.98	3.91
4.06	4.04	3.97	3.90	3.89	3.99	3.96
4.10	3.97	3.98	3.97	3.99	4.02	4.05
4.04	3.95	3.98	3.90	4.00	3.93	4.06

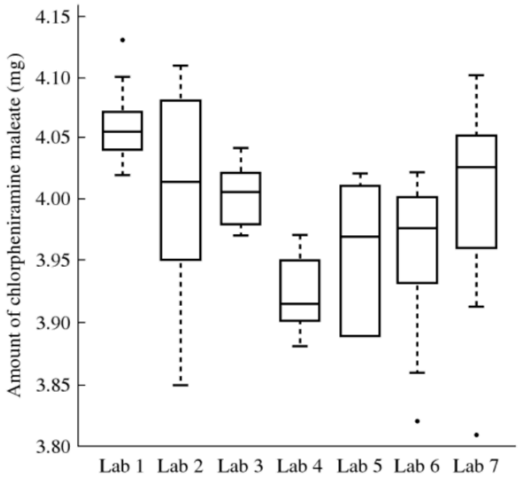


FIGURE **12.1** Boxplots of determinations of amounts of chlorpheniramine maleate in tablets by seven laboratories.

2.1.1. ANOVA 모형

이제 Two-way ANOVA(이원배치 분산분석)모형이 실제로 어떻게 생겼는지 들여다 보자.

$$Y_{ijk} = i\text{번째 처치, } j\text{번째 요인이 주어졌을 때 } k\text{번째 관측치}$$
$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

여기서 k는 관측치의 인덱스(index)를 의미한다. 각 모수(parameter)가 의미하는 바는 무엇일까?

- μ는 실험결과 전체의 평균, 즉 처치 및 요인의 영향을 받지 않는 기본적인 평균수준이 존재한다고 가정하는 것이다.
- α_i는 i번째 처치의 평균효과(mean effect)가 존재한다고 가정하는 것이다.
- β_j는 j번째 요인의 평균효과가 존재한다고 가정하는 것이다.

ANOVA는 다음 조건들이 가정된다.

- $\varepsilon_{ijk} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, 즉 오차(residual)는 서로 같은 정규분포를 따르면서 독립이다. iid는 **Independent and identically distributed**를 의미한다.
- $\sum_{i=1}^{I} \alpha_i = 0$
- $\sum_{i=j}^{J} \beta_j = 0$

자, 이제 우리가 관심있는 종속변수의 기대값을 보자. 우리가 가정하기를 잔차가 정규분포를 따른다고 했다. 그런데 앞에서 정규분포는 member of Location-scale family라고 했다. 따라서,

$$Y_{ijk} \overset{ind.}{\sim} \mathcal{N}(\mu + \alpha_i + \beta_j, \sigma^2)$$
$$E(Y_{ijk}) = \mu + \alpha_i + \beta_j := \mu_{ij}$$
$$Var(Y_{ijk}) = \sigma^2$$

각 처치, 요인 조합마다 샘플의 개수가 같을 때 balanced design, 같지 않을 경우 unbalanced design이라 한다.

마지막 두 개의 조건은 정규화(regularization) 조건이다. 이는 ANOVA의 가설검정에서 필요한 이론적 가정인데, 이 가정이 없으면 추정해야 할 모수들($\mu, \alpha_1, \cdots, \alpha_I, \beta_1, \cdots, \beta_J, \sigma^2$)이 수학적으로 해를 구하기가 불가능해진다. 통계모형은 기본적으로 최우추정(maximum-likelihood estimation)을 통해 모수들을 추정하는데, 다음은 이 모수들의 최우추정 연립방적식이다.

$$\frac{\sum_{i,j,k}(y_{ijk} - \alpha_i - \beta_j)}{I \cdot J \cdot n} = \mu$$
$$\frac{\sum_{j,k}(y_{ijk} - \mu - \beta_j)}{J \cdot n} = \alpha_i$$
$$\frac{\sum_{i,k}(y_{ijk} - \mu - \alpha_i)}{I \cdot n} = \beta_j$$

이를 $\bar{\alpha} := I^{-1} \sum_i \alpha,\ \bar{\beta} := J^{-1} \sum_j \beta_j$를 이용해 간단하게 표현하면

$$\bar{Y}_{...} - \bar{\alpha} - \bar{\beta} = \mu$$
$$\bar{Y}_{i..} - \mu - \bar{\beta} = \alpha_i$$
$$\bar{Y}_{.j.} - \mu - \bar{\alpha} = \beta_j$$

인데, 이 연립방적식 만으론 μ, α_i, β_j가 식별이 불가능(unidentifiable)하므로, $\bar{\alpha} = \bar{\beta} = 0$ 이라는 조건이 필요한 것이다. 따라서 이 조건 하에 모수들을 추정하면,

$$\hat{\mu} = \bar{Y}_{...}$$
$$\hat{\alpha}_i = \bar{Y}_{i..} - \hat{\mu} = \bar{Y}_{i..} - \bar{Y}_{...}$$
$$\hat{\beta}_j = \bar{Y}_{.j.} - \hat{\mu} = \bar{Y}_{.j.} - \bar{Y}_{...}$$

으로 ANOVA는 위처럼 처리, 요인의 효과(effect)를 추정하는 것이다.

지금까지는 처치, 요인 각각의 개별효과만 살펴보았다. 그러나 처치와 요인 간 교호작용(interaction)은 존재하지 않을까? 가령, 두 가지 식단(요인)과 두 가지 운동법(처치)에 따른 체중감량 효과를 실험했다고 하자. 이때 운동법의 효과가 식단에 따라 달라지거나, 식단의 효과가 운동법에 따라 달라지는 경우에 두 변수 간 교호작용이 존재한다고 한다. 식단의 구성에 따라 weight exercise의 효과가 더 좋을 수 있고, cardio exercise의 효과가 더 좋을 수도 있는 것이다.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

이경우 역시 마찬가지로 정규화 조건$\left(\sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0\right)$이 필요하며, 최우추정을 통한 추정은 다음과 같다.

$$\hat{(\alpha\beta)}_{ij} = \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}$$

2.2. ANOVA의 가설검정

2.2.1. F-test

이원배치 분산분석의 귀무가설 및 대립가설은 다음과 같다.

$$H_0 : \alpha_1 = \cdots = \alpha_I = 0$$
$$H_1 : \text{Not } H_0$$
$$H_0 : \beta_1 = \cdots = \beta_J = 0$$
$$H_1 : \text{Not } H_0$$
$$H_0 : (\alpha\beta)_{11} = \cdots = (\alpha\beta_{IJ}) = 0$$
$$H_1 : \text{Not } H_0$$

즉, 처치와 요인의 효과를 검정하는 것이다. 그러나 귀무가설을 자세히 보면 omnibus testing인 것을 알 수 있다. 즉, 처치 및 요인 수준 간의 쌍대비교는 검정하지 않는 것이다.

$$SSA = \sum_{i,j,k} \hat{\alpha}_i^2 = Jn \sum_i \hat{\alpha}_i \text{: Treatment A sum of squares}$$
$$SSB = \sum_{i,j,k} \hat{\beta}_j^2 = In \sum_j \hat{\beta}_j \text{: Factor B sum of squares}$$
$$SSAB = \sum_{i,j,k} \hat{(\alpha_i\beta_j)}^2 = n \sum_{ij} (\hat{\alpha}\hat{\beta})_{ij} \text{: Interaction AB sum of squares}$$
$$SSE = \sum_{i,j,k} (Y_{ijk} - \bar{Y}_{ij.})^2 \text{: Error sum of squares}$$
$$SST = \sum_{i,j,k} (Y_{ijk} - \bar{Y}_{...})^2 \text{: Total sum of squares}$$
$$SSM = SST - SSE \text{: Model sum of squares}$$
$$SSA + SSB + SSE = SST$$

이제 우린 Mean Squarred 통계량을 구할 수 있다.

$$MSA = SSA/(I-1)$$
$$MSB = SSB/(J-1)$$
$$MSAB = SSAB/\{(I-1)(J-1)\}$$
$$MSE = SSE/\{IJ(n-1)\}$$
$$MST = SST/(IJn-1)$$
$$MSM = SSM/\{IJn - 1 - IJ(n-1)\} = SSM/(IJ-1)$$

수학적으로 다음이 알려져 있다.

$$MSA/MSE \sim F(I-1, IJ(n-1))$$
$$MSB/MSE \sim F(J-1, IJ(n-1))$$
$$MSAB/MSE \sim F((I-1)(J-1), IJ(n-1))$$
$$MSM/MSE \sim F(IJ-1, IJ(n-1))$$

예제 2.1.3.

	Sunlight Exposure			
Watering Frequency	None	Low	Medium	High
Daily	4.8	5	6.4	6.3
	4.4	5.2	6.2	6.4
	3.2	5.6	4.7	5.6
	3.9	4.3	5.5	4.8
	4.4	4.8	5.8	5.8
Weekly	4.4	4.9	5.8	6
	4.2	5.3	6.2	4.9
	3.8	5.7	6.3	4.6
	3.7	5.4	6.5	5.6
	3.9	4.8	5.5	5.5

Mean table

Watering Frq.	None	Low	Medium	High
Daily	4.14	4.98	5.72	5.78
Weekly	4	5.22	6.06	5.32

$$\hat{\mu} = \bar{Y}_{...} = 5.1525$$

$$\hat{None} = \bar{Y}_{1..} - \bar{Y}_{...} = 4.07 - 5.1525 = -1.0825 \qquad \hat{Low} = \bar{Y}_{2..} - \bar{Y}_{...} = 5.1 - 5.1525 = -0.0525$$

$$\hat{Medium} = \bar{Y}_{3..} - \bar{Y}_{...} = 5.89 - 5.1525 = 0.7375 \qquad \hat{High} = \bar{Y}_{4..} - \bar{Y}_{...} = 5.5 - 5.1525 = 0.3975$$

$$\hat{Daily} = \bar{Y}_{.1.} - \bar{Y}_{...} = 5.155 - 5.1525 = 0.0025 \qquad \hat{Weekly} = \bar{Y}_{.2.} - \bar{Y}_{...} = 5.15 - 5.1525 = -0.0025$$

$$(\hat{None * Daily}) = \bar{Y}_{11.} - \bar{Y}_{1..} - \bar{Y}_{.1.} + \bar{Y}_{...} = 4.14 - 4.07 - 5.155 + 5.1525 = 0.0675 \qquad (None * \hat{Weekly}) = 4 - 4.07 - 5.15 + 5.1525 = -0.0675$$

$$\cdots \qquad (High * \hat{Weekly}) = 5.32 - 5.5 - 5.15 + 5.1525 = -0.1775$$

ANOVA

	Sum of Squares	df	Mean Square	F	p
Watering Frequency	2.50e-4	1	2.50e-4	9.21e-4	0.976
Sunlight Exposure	18.76	3	6.255	23.05	< .001
Watering Frequency ✻ Sunlight Exposure	1.01	3	0.337	1.24	0.311
Residuals	8.68	32	0.271		

결과의 해석(유의수준 = 0.05)

- Watering Frequency의 효과는 유의하지 않음
- Sunlight Exposure의 효과는 유의함
- 처치와 요인 간 교호작용은 유의하지 않음

2.2.2. Post hoc comparisons

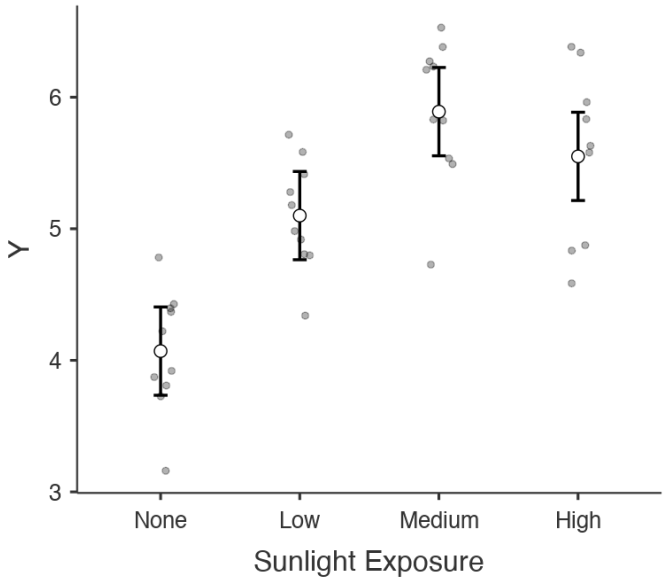
사후검정 방법으로는 LSD, Bonferroni, Sidak, Tukey, Duncan, Dunnett, Scheffe 등 여러가지 방법이 있다. 그 중에서도 Tukey, Duncan, Scheffe 의 방법이 대부분의 연구에 사용된다.

SPSS는 Tukey의 사후검정 시 unbalanced design의 경우에는 조화평균을 사용하는 Tukey-Kramer 검정을 이용하여 사후검정을 실시한다. Scheffe의 방법은 unbalanced design의 경우 사후검정을 위해 고안된 방법이다. 물론 balanced design에도 사용할 수 있다.

이 3가지 방법의 민감도에 대해 생각을 하면 Scheffe의 방법이 가장 엄격하고, Duncan의 방법이 가장 관용적이다. 그래서, Scheffe에서 차이가 있다라고 하면 Duncan에서는 차이가 있다라고 나오지만, 그 역은 성립하지 않는다.

그리고, Tukey의 방법은 Duncan과 Scheffe의 중간 정도에 위치한다고 생각하면 된다.

Estimated Marginal Means: Sunlight Exposure



Post Hoc Comparisons - Sunlight Exposure

Comparison								
Sunlight Exposure		**Sunlight Exposure**	**Mean Difference**	**SE**	**df**	**t**	**p tukey**	**p bonferroni**
None	-	Low	-1.030	0.233	32.0	-4.42	< .001	< .001
	-	Medium	-1.820	0.233	32.0	-7.81	< .001	< .001
	-	High	-1.480	0.233	32.0	-6.35	< .001	< .001
Low	-	Medium	-0.790	0.233	32.0	-3.39	0.010	0.011
	-	High	-0.450	0.233	32.0	-1.93	0.235	0.374
Medium	-	High	0.340	0.233	32.0	1.46	0.473	0.925