Experimental Design
Note 3
Introduction to ANOVA

Keunbaik Lee

Sungkyunkwan University

What if there are more than two levels of a single factor?

- The t-test does not directly apply.
- There are lots of practical situations where there are either more than two levels of interest, or there are several factors of simultaneous interest.
- The analysis of variance (ANOVA) is the appropriate analysis "engine" for these types of experiments.

평균 μ_i를 비교하고자 하는 분포가 2개이면 t-test를 진행하면 되지만, 3개 이상이 될 경우 t-test로 비교할 수 없기 때문에 ANOVA가 만들어짐. ANOVA는 3개 이상의 독립적 분포의 mean을 비교하기 위해 고안된 것이고, mean의 분산을 통해 분석을 한다는 의미에서 Analysis of Variance라고 한다.

ANOVA - Analysis of Variance I

- Extends independent-samples t test.
- Compares the means of groups of independent observations
  - Do not be fooled by the name
  - ANOVA does not compare variances
  - The name "analysis of variance" stems from a partitioning of the total variability in the response variable into components
  - Can compare more than two groups
- The ANOVA was developed by Fisher in the early 1920s, and initially applied to agricultural experiments. Now it is used widely.

Say the sample contains K independent groups

ANOVA - Analysis of Variance II

groups $1, 2, \ldots, i, \ldots, k$

$\mu_1 \; \mu_2 \cdots \mu_i \cdots \mu_k$

- ANOVA tests the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$$

That is, “the group means are all equal”

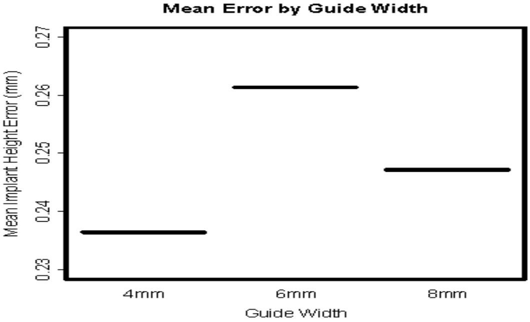
- The alternative hypothesis is

$$H_1 : \mu_i \neq \mu_j \text{ for some } i, j$$

or, “the group means are not all equal”

Example: Accuracy of Implant Placement I

- Implants were placed in a manikin using placement guides of various widths.
- 15 implants were placed using each guide.
- Error (discrepancies with a reference implant) was measured for each implant



임플란트 가이드라인의
각 평균오차

Example: Accuracy of Implant Placement II

- Does changing the guide change the mean height error?
- Is there an optimum level for guide?
- We would like to have an objective way to answer these questions
- The t-test really does not apply here - more than two factor levels
  - Pairwise comparisons will inflate type I error

    비교하고자 하는 분포가 3개 이상일 때 t-test comparison을 진행하면 type I error의 확률을 증가시킴

    Type I Error = $P(\text{reject } H_0 \mid H_0) = \alpha = 0.05$

The ANOVA Statistic

- Each time a hypothesis test is performed at significance level α, there is probability rejecting in error. $P(\text{reject } H_0 \mid H_0)$
- Performing multiple tests increases the chances of rejecting in error at least once.
- For example:
  - if you did 3 independent hypothesis tests at the $\alpha = 0.05$
  - If, in truth, H_0 were true for all three.
  - The probability that at least one test rejects H_0 is 14.3% (P(at least one rejection)=1-P(no rejections)=$1 - .95^3 = 0.143$)

$H_0 : \mu_1 = \mu_2 = \mu_3$

$H_0 : \mu_1 = \mu_2$ $H_0 : \mu_2 = \mu_3$ $H_0 : \mu_1 = \mu_3$

with $\alpha = 0.05$ each

$\Rightarrow$ $H_0 : \mu_1 = \mu_2 = \mu_3$ 이 reject 된다는 말은 적어도 하나의 비교군의 H_0 가 reject 된다는 의미

$\Rightarrow$ $P(\text{not reject } H_0 \mid H_0) = 0.95 = 1 - \alpha$

$\Rightarrow$ $P(\text{at least one rejection}) = 1 - \{P(\text{not reject } H_0 \mid H_0)\}^3$

$= 1 - (0.95)^3 = 0.143 = P(\text{Type I Error})$, 때문에 비교군이 증가할 땐 $P(\text{Type I Error})$도 같이 증가하게 된다.

ANOVA

Groups

1	2	$\cdots$	i	$\cdots$	K
y_{11}	y_{21}		y_{i1}		y_{k1}
y_{12}	y_{22}		y_{i2}		y_{k2}
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$
y_{1n}	y_{2n}		y_{in}		y_{kn}
μ_1	μ_2		μ_i		μ_K

μ = the overall mean

$$\begin{array}{cccc} (\mu_1-\mu)^2 & (\mu_2-\mu)^2 & (\mu_i-\mu)^2 & (\mu_k-\mu)^2 \\ \wr & \wr & \wr & \wr \\ (\bar{y}_{1\cdot}-\bar{y}_{\cdot\cdot})^2 + & (\bar{y}_{2\cdot}-\bar{y}_{\cdot\cdot})^2 + & (\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2 + & (\bar{y}_{k\cdot}-\bar{y}_{\cdot\cdot})^2 \end{array}$$

=> 이들 mean들 중에서 overall mean μ과 아주 다른게 존재한다면
위 숫자의 합은 0보다 확실히 클 것이다. 그러나 이 값은
observation의 measurement에 따라 크게 달라질 것이기에 Variation을 알아야 한다.
(Scale)

$$\text{Overall variation} = \sum_{i=1}^{K}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{K}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i\cdot}+\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{K}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i\cdot})^2 + \sum_{i=1}^{K}\sum_{j=1}^{n}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2 + \underbrace{2\sum_{i=1}^{K}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i\cdot})(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})}_{=0}$$

$$= \underbrace{\sum_{i=1}^{K}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i\cdot})^2}_{SS_E\,:\,\text{그룹별 차이}} + \underbrace{\sum_{i=1}^{K}\sum_{j=1}^{n}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2}_{SS_B\,:\,\text{그룹간 차이}}$$

☆ 데이터가 주어진 동시에 $\underbrace{\sum_{i=1}^{K}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{\cdot\cdot})^2}_{SST}$ 은 fixed value가 된다.

H_0가 True일 때는 SS_B가 상대적으로 작아진다. H_0가 False일 경우 상대적으로 커진다.

Given $\underbrace{SST}_{\text{fixed}} = SSE + SSB$, if $\frac{SSB}{SSE} \approx 0$, cannot reject H_0

, if $\frac{SSB}{SSE} \gg 0$, reject H_0

☆ $\frac{MSB}{MSE} \overset{H_0}{\sim} F$

Why pairwise comparisons inflates type I error?

- To combine the differences from the grand mean we
  - Square the differences
  - Multiply by the numbers of observations in the groups
  - Sum over the groups

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2 = SS_B = 15(\bar{X}_{4mm} - \bar{\bar{X}})^2 + 15(\bar{X}_{6mm} - \bar{\bar{X}})^2 + 15(\bar{X}_{8mm} - \bar{\bar{X}})^2$$

j에 대해 독립이므로

where $\bar{X}_*$ are the group means and $\bar{\bar{X}}$ is the grand mean.

$$SS_B = \text{Sum of Squares Between groups}$$

Note: This looks a bit like a variance.

How big is big?

- For the Implant Accuracy Data, $SS_B = 0.0047$
- Is that big enough to reject H_0?
- As with the t test, we compare the statistic to the variability of the individual observations.
- In ANOVA the variability is estimated by the Mean Square Error, or MSE

$$\frac{SSB/df}{SSE/df}$$

그러나 이 경우 observation의 scale에 따라 값이 상이하기 때문에 MSE로 유도한다

MSE: Mean Square Error I

The Mean Square Error is a measure of the variability after the group effects have been taken into account.

$$MSE = \frac{1}{N-a} \sum_{j=1}^{k} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

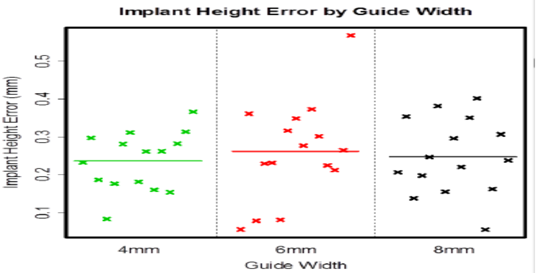
$y_{ij} = \mu_i + \varepsilon_{ij}$

$\Updownarrow$

$y_{ij} - \bar{y}_{i.} = \hat{\varepsilon}_{ij}$

$(y_{ij} - \bar{y}_{i.})^2 = \hat{\varepsilon}_{ij}^2$

where x_{ij} is the ith observation in the jth group.



MSE: Mean Square Error II

Note that the variation of the means seems quite small compared to the variance of observations within groups

Notes on MSE

- If there are only two groups, the MSE is equal to the pooled estimate of variance used in the equal-variance t test.
- ANOVA assumes that all the group variances are equal.
- Other options should be considered if group variances differ by a factor of 2 or more.

ANOVA F Test

- The ANOVA F test is based on the F statistic

$$F = \frac{SS_B/(a-1)}{MSE}$$

where a is the number of groups.

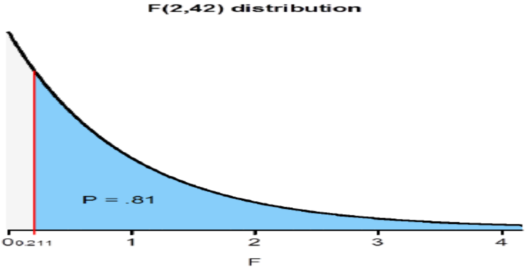
- Under H_0 the F statistic has an "F" distribution, with $a-1$ and $N-a$ degrees of freedom (N is the total number of observations). In this case N=45

F Test p-value

To get a p-value we compare our F statistic to an $F(2,42)$ distribution. In our example, The p-value is

$$F = \frac{\overset{SSB}{.0047}/\overset{3-1}{2}}{\underset{SSE}{.4664}/\underset{N-3}{42}} = .211$$

The p-value is $P(F(2,42) > .211) = 0.81$.



ANOVA Table I

Results are often displayed using an ANOVA Table

Source of Variation	Sum of Squares	df	Mean Square	F	P-value
Between Groups	.005	2	.002	.211	.811
Within Groups	.466	42	.011		
Total	.470	44			

- The name “analysis of variance” stems from a partitioning of the total variability in the response variable into components that are consistent with a model for the experiment

ANOVA Table II

- The basic single-factor ANOVA model is

$$y_{ij} \overset{= \mu_i + \epsilon_{ij}}{=} \mu + \tau_i + \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

$$= \mu + \underbrace{\mu_i - \mu}_{i\text{th treatment effect}} + \epsilon_{ij}$$

where μ =an overall mean, $\tau_i = i$th treatment effect (τ_i is constant and $\underbrace{\sum_i \tau_i = 0}_{\text{restriction}}$), ϵ_{ij} =experiemental error, $NID(0, \sigma^2)$ for $i = 1, 2, \cdots, a$ and $j = 1, 2, \cdots, n$ (Balanced design).

Models for the Data

There are two ways to write a model for the data:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \text{ is called the } \textbf{effects model}.$$

Let $\mu_i = \mu + \tau_i$. Then

$$y_{ij} = \mu_i + \epsilon_{ij} \text{ is called the } \textbf{mean model}.$$

Regression models can also be employed.

Notations for ANOVA I

- Total variability is measured by the total sum of squares:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2.$$

$$= SS_B + SSE$$

$$= \sum_{i=1}^{a} n(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2$$

Notations for ANOVA II

- The basic ANOVA partitioning is:

$$\begin{aligned}
\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{\cdot\cdot})^2 &= \sum_{i=1}^{a}\sum_{j=1}^{n}\{(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})+(y_{ij}-\bar{y}_{i\cdot})\}^2 \\
&= n\sum_{i=1}^{a}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2+\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i\cdot})^2, \\
SS_T &= SS_{Treatment}+SS_E \\
\text{or } SS_T &= SS_B+SS_E
\end{aligned}$$

- A large value of $SS_{Treatments}$ reflects large differences in treatment means.

Notations for ANOVA III

- A small value of $SS_{Treatments}$ likely indicates no differences in treatment means.
- Formal statistical hypotheses are:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a = \mu \qquad H_0 : \underset{\mu_i - \mu}{\tau_1} = \tau_2 = \cdots = \tau_a = 0$$

$$H_1 : \text{ at least one "=" does not hold.} \quad H_1 : \text{at least one is not } 0$$

- While sums of squares cannot be directly compared to test the hypothesis of equal means, mean squares can be compared.

Notations for ANOVA IV

- A mean square is a sum of squares divided by its degrees of freedom:

$$df_{Total} = df_{Treatments} + df_{Error},$$
$$an - 1 = a - 1 + a(n-1),$$
$$MS_{Treatment} = \frac{SS_{Treatment}}{a-1}, \quad MS_E = \frac{SS_E}{a(n-1)} = N - a.$$

- If the treatment means are equal, $MS_{Teatments} = 0$.

$Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu_i, \sigma^2)$

$\Rightarrow \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma}\right)^2 \sim \chi^2_n \longrightarrow \sum_{i=1}^{n} \frac{Y_i^2}{\sigma^2} \sim \chi^2_{n(\lambda)}$, where $\lambda = \frac{1}{2} \sum_{i=1}^{n} \frac{\mu_i^2}{\sigma^2}$

The Analysis of Variance is Summarized in a Table I

TABLE 3.3

The Analysis of Variance Table for the Single-Factor, Fixed Effects Model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Between treatments	$SS_{\text{Treatments}} = n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2$	$\star\ \tau_i^2$ $a-1$	$MS_{\text{Treatments}}$	$F_0 = \dfrac{MS_{\text{Treatments}}}{MS_E}$
Error (within treatments)	$SS_E = SS_T - SS_{\text{Treatments}}$	$N-a$	MS_E	
Total	$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2$	$N-1$		

- The reference distribution for F_0 is the $F_{a-1,a(n-1)}$ distribution
- Reject the null hypothesis (equal treatment means) if $F_0 > F_{\alpha,a-1,a(n-1)}$.

The Analysis of Variance is Summarized in a Table II

HW

$$\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 = SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}^2 - \frac{y_{..}^2}{N},$$

$$\sum_{i=1}^{a} n(\bar{y}_{i.} - \bar{y}_{..})^2 = SS_{Treatments} = \frac{1}{n} \sum_{i=1}^{a} y_{i.}^2 - \frac{y_{..}^2}{N},$$

$$SS_E = SS_T - SS_{Treatments}$$

Some words for coding the observations

- If every observation subtracts the same constant, then sums of squares do not change, so we can get the same conclusion.
- If we multiply each observation by the same constant, then the sums of squares change. But the F ratio is equal to the F ratio for the original data. It implies that we can still get the same conclusion.

Parameter estimation I

- Estimates for parameters:

$\hat{\mu}_i - \hat{\mu}$

$$\hat{\mu} = \bar{y}_{\cdot\cdot}, \quad \hat{\tau}_i = (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}),$$

$$\hat{\epsilon}_{ij} = y_{ij} - \bar{y}_{i\cdot} \quad \text{(residual)}$$

$- \mu_i$

So $y_{ij} = \hat{\mu} + \hat{\tau}_i + \hat{\epsilon}_{ij}$.

- So estimator of mean of the ith treatment is:

$$\hat{\mu}_i = \hat{\mu} + \hat{\tau}_i = \bar{y}_{i\cdot}$$

Parameter estimation II

- And the $100(1-\alpha)\%$ confidence interval for the ith treatment mean:

$T = \frac{Z}{\sqrt{W/V}} \Rightarrow \begin{matrix} Z \sim N(0,1) \\ W \sim \chi^2_v \end{matrix}$ 독립

$\sim t_v$

$$\bar{y}_{i\cdot} - t_{\alpha/2,N-a}\sqrt{\frac{MS_E}{n}} \le \mu_i \le \bar{y}_{i\cdot} + t_{\alpha/2,N-a}\sqrt{\frac{MS_E}{n}}$$

- C.I. for the difference of two treatment means:

$$\bar{y}_{i\cdot} - \bar{y}_{j\cdot} - t_{\alpha/2,N-a}\sqrt{\frac{2MS_E}{n}} \le \mu_i - \mu_j \le \bar{y}_{i\cdot} - \bar{y}_{j\cdot} - t_{\alpha/2,N-a}\sqrt{\frac{2MS_E}{n}}$$

$\bar{y}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} y_{ij} \sim N(\mu_i, \frac{\sigma^2}{n}) \Rightarrow \frac{\bar{y}_{i\cdot} - \mu_i}{\sigma/\sqrt{n}} \sim N(0,1)$

$\frac{SSE}{\sigma^2} \sim \chi^2_{N-a}$ ..... $T = \frac{\frac{\bar{y}_{i\cdot} - \mu_i}{\cancel{\sigma}/\sqrt{n}}}{\sqrt{\underbrace{\frac{SSE}{\cancel{\sigma^2}}/(N-a)}_{=MSE}}} \sim t_{N-a}$

$= \frac{\bar{y}_{i\cdot} - \mu_i}{\sqrt{\frac{MSE}{n}}} \Rightarrow \left|\frac{\bar{y}_{i\cdot} - \mu_i}{\sqrt{\frac{MSE}{n}}}\right| < \mu_i \Rightarrow \bar{y}_{i\cdot} - t_{\frac{\alpha}{2},N-a}\sqrt{\frac{MSE}{n}} \le \mu_i \le \bar{y}_{i\cdot} + t_{\frac{\alpha}{2},N-a}\sqrt{\frac{MSE}{n}}$

Model for unbalanced experiment I

- More general model for unbalanced experiment:

$$y_{ij} = \mu + \overset{\mu_i - \mu}{\underset{\text{treatment effect}}{\tau_i}} + \epsilon_{ij}, \text{ for } i = 1, 2, \cdots, a; j = 1, 2, \cdots, n_i,$$

where $\sum_{i=1}^{a} n_i \tau_i = 0.$, $\sum_{i=1}^{a} \tau_i = 0$ 라는 restriction이 파생된거임 ☆ n_i 들이 모두 같을 경우

- Notation: $\sum_{i=1}^{a} n_i \tau_i = \sum_{i=1}^{a} n_i \mu_i - n_i \mu = \sum_{i=1}^{a} n_i \mu_i - N\mu,$

$$y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij} \;\Rightarrow\; \bar{y}_{i\cdot} = y_{i\cdot}/n_i \text{ (treatment sample mean, or row mean}$$

$$y_{\cdot\cdot} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij} \;\Rightarrow\; \bar{y}_{\cdot\cdot} = y_{\cdot\cdot}/\underset{\sum_{i=1}^{a} n_i}{\underset{\sim}{N}} \text{ (grand sample mean)}$$

Model for unbalanced experiment II

- Decomposition of y_{ij}: $y_{ij} = \bar{y}_{\cdot\cdot} + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + \underbrace{(\bar{y}_{ij} - \bar{y}_{i\cdot})}_{\text{residual}}$
- Estimates for parameters:

$$
\begin{aligned}
\hat{\mu} &= \bar{y}_{\cdot\cdot}, \\
\hat{\tau}_i &= (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}), \\
\hat{\epsilon}_{ij} &= y_{ij} - \bar{y}_{i\cdot}, \text{ (residual)}
\end{aligned}
$$

So $y_{ij} = \hat{\mu} + \hat{\tau}_i + \hat{\epsilon}_{ij}$.

- It can be verified that

HW

$$\sum_{i=1}^{a} n_i \hat{\tau}_i = 0, \quad \sum_{j=1}^{n_i} \hat{\epsilon}_{ij} = 0, \text{ for all } i$$

Example: Tensile Strength

Investigate the tensile strength of a new synthetic fiber. The factor is the weight percent of cotton used in the blend of the materials for the fiber and it has five levels.

Percent of cotton	Tensile Strength 1	2	3	4	5	Total	Average
15	7	7	11	15	9	49	9.8
20	12	17	12	18	18	77	15.4
25	14	18	18	19	19	88	17.6
30	19	25	22	19	23	108	21.6
35	7	10	11	15	11	54	10.8

$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$

See Tensile.SAS.

$H_0: \tau_1 = \tau_2 = \cdots = \tau_5 = 0$