5. ANOVA (Analysis of Variance)

[ANOVA 1] One-way ANOVA $\Rightarrow$ Fixed effect regression model

(1) $Y_{ij} = \mu_i + \epsilon_{ij}$ $\quad i=1\cdots k, \quad j=1,\cdots,n_i$

(2) $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ $\quad \sum_{i=1}^{k} \alpha_i = 0$

[ANOVA 2] (2.1) One-way ANOVA global F-test

Under $H_0: \mu_1 = \mu_2 = \cdots = \mu_k \Leftrightarrow H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_k$

$$F_{stat} = \frac{(RSS_H - RSS_{Full})/(k-1)}{RSS_{Full}/(n-k)} \sim F_{k-1,\, n-k}$$

$$RSS_{Full} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 \qquad \hat{\mu}_i = \bar{Y}_{i.}$$

$$RSS_H = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 \qquad \hat{\mu}_H = \bar{Y}_{..}$$

(2.2) ANOVA table

Source	Sum of Squares	Degrees of freedom
Between groups	$SS_{Between} = \sum_{i=1}^{k} n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2$	$k-1$
Within groups	$SS_{Within} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$	$n-k$
Total	$SS_{Total} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2$	$n-1$

△ When the global F-test is reject $\Rightarrow$ look for contrasts

[ANOVA 3] One contrast

① Hypothesis test : under $H_0 : \mu_i = \mu_j$

$$\frac{\hat{\mu}_i - \hat{\mu}_j}{S.E.(\hat{\mu}_i - \hat{\mu}_j)} \sim t_{df} \quad \text{with} \quad S.E.(\hat{\mu}_i - \hat{\mu}_j) = \hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

$$df = n - k$$

② Confidence interval for $\mu_i - \mu_j$

$$\hat{\mu}_i - \hat{\mu}_j \pm t_{df}^{(\alpha/2)} \times S.E.(\hat{\mu}_i - \hat{\mu}_j)$$

△ K groups $\Rightarrow$ $\binom{k}{2}$ pairwise comparisons $\Rightarrow$ multiple contrasts

[ANOVA 4] Multiple contrasts (Simultaneous inference)

Given multiple hypotheses $H_{01}, \cdots, H_{0m}$ for parameters $\theta_1 \cdots \theta_m$

(4.1) **Familywise Error Rate (FWER)** is probability of rejecting at least one of $H_{01}, \cdots, H_{0m}$ when they are all true

(4.2) **Simultaneous confidence intervals** at $100(1-\alpha)\%$ level are intervals (L_i, U_i) $i = 1 \cdots m$ with

$$P[L_i \le \theta_i \le U_i \text{ for all } i = 1, \cdots, m] > 1 - \alpha$$

(4.3) **Bonferroni's correction** Let P_i be p-value of H_{0i}.

① Hypothesis test :

VI : reject H_{0i} if $P_i < \frac{\alpha}{m}$

V2: define $p_{i,adj} = \min\{m \times p_i, 1\}$, reject H_{0i} if $p_{i,adj} < \alpha$

② Simultaneous confidence intervals: change quantile to at level $\frac{\alpha}{2m}$

E.g. $\hat{\mu}_i - \hat{\mu}_j \pm t_{df}^{(\frac{\alpha}{2m})} \times S.E.(\hat{\mu}_i - \hat{\mu}_j)$ with $m = \binom{K}{2}$, $df = n-K$

Differences and connections between ① one-way ANOVA F-test ② one contrast t-test ③ multiple contrast tests

- Bonferroni's correction is a general method.
- But it tends to be too conservative if $k > 10$.
- In specific problem, we may be able to improve.

- Suppose now we are interested in ALL pairwise comparisons $\mu_i = \mu_j$ for $i \neq j$

(4.4) Tukey - Kramer procedure for pairwise comparison

- Studentized range distribution
  - Let $X_1, \cdots, X_m$ be iid $N(\mu, \sigma^2)$
  - Let $R = \max_{1 \le i \le m} X_i - \min_{1 \le j \le m} X_j$ be the range
  - $\frac{R}{\hat{\sigma}}$ follows the studentized range distribution $q_{m,\nu}$

    where m is # of normal r.v.

    ν is # of degrees of freedom used in estimating σ.

We apply the property to pairwise difference

△ Consider a balanced design $n_1 = \cdots = n_K = n_B$

k groups have a common mean μ_0

$y_{ij} \sim N(\mu_0, \sigma^2)$ for $i = 1 \cdots K$, $j = 1 \cdots, n_B$

$$|t_{ij}| = \frac{|\hat{\mu}_i - \hat{\mu}_j|}{\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} = \frac{|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}|}{\hat{\sigma}\sqrt{\frac{1}{n_B} + \frac{1}{n_B}}}$$

$$= \frac{1}{\sqrt{2}} \frac{1}{\hat{\sigma}} \left| \sqrt{n_B}\, \bar{y}_{i\cdot} - \sqrt{n_B}\, \bar{y}_{j\cdot} \right|$$

$$= \frac{1}{\sqrt{2}} \times \frac{1}{\hat{\sigma}} \left| \sqrt{n_B}(\bar{y}_{i\cdot} - \mu_0) - \sqrt{n_B}(\bar{y}_{j\cdot} - \mu_0) \right|$$

$$= \frac{1}{\sqrt{2}} \times \frac{1}{\hat{\sigma}} |X_i - X_j| \qquad *$$

where we denote $X_i = \sqrt{n_B}(\bar{y}_{i\cdot} - \mu_0)$

- As $X_i \overset{iid}{\sim} N(0, \sigma^2)$, for $i = 1 \cdots K$

$q_{K, n-K}$ is the distribution of

$$\frac{\max_{1 \le i \le K} X_i - \min_{1 \le i \le K} X_i}{\hat{\sigma}}$$

$$= \max_{1 \le i \ne j \le K} (X_i - X_j) \times \frac{1}{\hat{\sigma}}$$ By equivalence of definitions

$$= \max_{1 \le i < j \le K} |X_i - X_j| \times \frac{1}{\hat{\sigma}}$$ ⇒ By $|X_1 - X_2| = \max\{X_1 - X_2, X_2 - X_1\}$

$$= \sqrt{2} \times \max_{1 \le i < j \le K} |t_{ij}| \quad \text{(By } \circledast \text{ above)}$$

In summary $\sqrt{2} \times \max_{1 \le i < j \le K} |t_{ij}| \sim q_{K,\, n-K}$

△ Let $q_{K,\, n-K}^{(\alpha)}$ denote the upper α-level quantile of the studentized range distribution

△ [Hypothesis test with FWER control]

Reject $H_{0,ij}: \mu_i = \mu_j$ if

$$|t_{ij}| = \frac{|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}|}{\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} > \frac{1}{\sqrt{2}} q_{K,\, n-K}^{(\alpha)}$$

- It controls FWER exactly at α for balanced design $n_1 = \cdots = n_K$

  approximately at α for unbalanced design

Proof of FWER $= \alpha$ under balanced design

FWER

$$= P\left(\text{at least one of } |t_{ij}| > \frac{1}{\sqrt{2}} q_{k,n-k}^{(\alpha)} \;\middle|\; \mu_i = \mu_j \text{ for all } i \neq j\right)$$

$$= P\left[\max_{1 \le i < j \le k} |t_{ij}| > \frac{1}{\sqrt{2}} q_{k,n-k}^{(\alpha)} \;\middle|\; \mu_i = \mu_j \text{ for all } i \neq j\right]$$

$$= \alpha$$

(Because we have shown $\sqrt{2} \times \max_{1 \le i < j \le k} |t_{ij}| \sim q_{k,n-k}$)

△ [Simultaneous confidence intervals]

Tukey's HSD intervals

(Honest Significance Difference)

$$\bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm \frac{1}{\sqrt{2}} q_{k,n-k}^{(\alpha)} \times \hat{\sigma} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

(4.5) False Discovery Rate (FDR)

- Perform m simultaneous hypothsis tests
  Classify the null hypotheses based on truth and results

	H_0 rejected	H_0 not rejected	Total
H_0 true	FD (V)	TN	m_0
H_0 false	TD	FN	m_1
Total	D (R) Rejected	N	m

T / F = true / false   P/N = Discovery / Nondiscovery
(Rejected)

- All quantities except m, D and N are unobserved

False Discovery Proportion

$$FDP = \frac{\#\ \text{False Discoveries}}{\#\ \text{Discoveries}} = \frac{V}{\max\{R, 1\}}$$

where $\max\{R, 1\} = \begin{cases} R & \text{if } R \geq 1 \\ 0 & \text{if } R = 0 \end{cases}$

False Discovery Rate

$$FDR = E(FDP) = E\left(\frac{V}{\max\{R, 1\}}\right)$$

- **Benjamini-Hochberg (BH)** procedure (JRSS-B 1995)

① Rank the p-values from smallest to largest with

$P_{(0)} = 0 \quad P_{(1)} = \min_i P_i, \quad P_{(2)}, \quad P_{(3)} \cdots P_{(m)}$

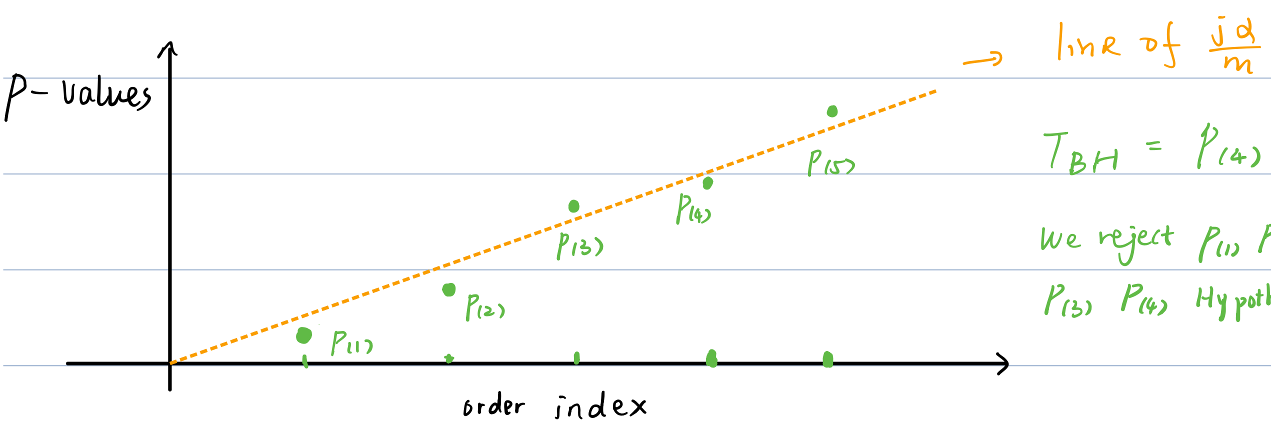
② Let the thresholds for $j = 0, \cdots, m$ be

$0, \quad \frac{\alpha}{m}, \quad \frac{2\alpha}{m}, \quad \frac{3\alpha}{m}, \quad \cdots, \quad \alpha$

③ BH threshold is defined for pre-specified $\alpha \in (0,1)$

$$T_{BH} = \max_{1 \le j \le m} \left\{ P_{(j)} : P_{(j)} \le \frac{j\alpha}{m} \right\}$$

Reject null hypothesis H_{0l} if $P_{(l)} \le T_{BH}$



$T_{BH} = P_{(4)}$

We reject $P_{(1)}$ $P_{(2)}$ $P_{(3)}$ $P_{(4)}$ Hypotheses

Equivalent procedure 1:

$\hat{j} = 4$ in above example

Let $\hat{j} = \max_{1 \le j \le m} \left\{ j : P_{(j)} \le \frac{j\alpha}{m} \right\}$

Reject null hypotheses corresponding to $P_{(1)}, P_{(2)}, \cdots, P_{(\hat{j})}$

Proof: Since $T_{BH} = P_{(\hat{j})}$ by definition,

$P_{(l)} \le T_{BH} \Leftrightarrow P_{(l)} \le P_{(\hat{j})}$

$\Leftrightarrow$ By the order of $P_{(l)}$

l takes $\{1, 2, \cdots, \hat{j}\}$

Equivalent procedure 2:

Define BH adjusted p-values

$$P_{(j), adj} = \min \left\{ \min_{l \ge j} \left\{ \frac{m P_{(l)}}{l} \right\}, 1 \right\}$$

Reject H_{0j} if $P_{(j), adj} \le \alpha$

Proof: $P_{(j),adj} \le \alpha \iff \min_{l \ge j} \frac{m P_{(l)}}{l} \le \alpha$

Equivalent to saying $P_{(j)}$ is rejected if one of index $l \ge j$ gives $P_{(l)} \le \frac{l}{m} \alpha$

Looking at graphical example above

$P_{(3)}$ is rejected because $P_{(4)} \le \frac{4}{m} \alpha$ (take $l=4$)

$P_{(4)}$ is rejected because $P_{(4)} \le \frac{4}{m} \alpha$ (take $l=4$)

Theoretical guarantee

BH (1995) proved that for independent tests the BH procedure guarantees that

$$FDR \le \frac{m_0}{m} \alpha \le \alpha$$

Intuition (only for students of interest, not required)

Suppose H_{0i} is true with probability $p = \frac{m_0}{m}$ independently

Reject at level t_{BH}

$$FDR = E\left(\frac{FD}{D}\right) \approx \frac{E(FD)/m}{E(D)/m}$$

$$= \frac{E\left\{\frac{1}{m}\sum_{i=1}^{m} I(P_i \le t_{BH}) \times I(H_i \text{ holds})\right\}}{E\left\{\frac{1}{m}\sum_{i=1}^{m} I(P_i \le t_{BH})\right\}}$$

$$\approx \frac{P(P_i \le t_{BH} \mid H_i \text{ true})\, P(H_i \text{ is true})}{\hat{G}(t_{BH})}$$

(Let $\hat{G}(t) = \frac{1}{m}\sum_{i=1}^{m} I(P_i \le t)$ denote empirical CDF of P_i's)

$$\approx \frac{t_{BH} \times \frac{m_0}{m}}{\hat{G}(t_{BH})}$$

[Use $P(P_i \le t_{BH} \mid H_i \text{ true}) = t_{BH}$

$P(H_i \text{ is true}) = \frac{m_0}{m}$

We want t_{BH} to be as large as possible while $\star \leq \frac{m_0}{m}\alpha$

$$\hat{t}_{BH} = \sup\left\{ t : \frac{t \times \frac{m_0}{m}}{\hat{G}(t)} \leq \frac{m_0}{m}\alpha \right\}$$

$$\approx \max_{1 \leq j \leq m} \left\{ t_{(j)} : t_{(j)} \leq \hat{G}(t_{(j)}) \times \alpha \right\}$$

(Since $t_{(j)}$ j-th smallest p-value, $\hat{G}(t_{(j)}) = j/m$)

$$\approx \max_{1 \leq j \leq m} \left\{ t_{(j)} : t_{(j)} \leq \frac{j}{m} \times \alpha \right\}$$