AMS 572 Data Analysis I
Analysis of Single Factor Experiments

Pei-Fen Kuan

Applied Math and Stats, Stony Brook University

Analysis of Variance Model

- Objective: To test hypotheses about the mean of more than 2 groups
- Definition: An *analysis of variance model* is a linear regression model in which the predictor variables are classification variables. The categories of a variable are called the *levels* of the variable.
- Categorical predictor variables are also called *qualitative factors*

Analysis of Variance Model

Data structure:

$$\begin{array}{ccc} population\,1 & population\,i & population\,K \\ \downarrow & \downarrow & \downarrow \\ sample\,1 & sample\,i & sample\,K \end{array}$$

$$n_1 \left\{ \begin{array}{l} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \end{array} \right. \qquad n_i \left\{ \begin{array}{l} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{array} \right. \qquad n_K \left\{ \begin{array}{l} Y_{K1} \\ Y_{K2} \\ \vdots \\ Y_{Kn_K} \end{array} \right.$$

Balanced design: $n_i \equiv n$

Notation

- Let Y_{ij} be the j^{th} observation in the i^{th} group
- $i = 1, \ldots, K;\ j = 1, \ldots, n_i$
- Let $N = \sum_{i=1}^{K} n_i$
- $\bar{Y}_{i\cdot} = \sum_j Y_{ij}/n_i$

ANOVA Model and Hypotheses

- Assume $Y_{ij} \sim N(\mu_i, \sigma^2)$.
  That is, equal (unknown) population variances
  $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_K^2 = \sigma^2$
- Suppose

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

  versus

  $$H_a : \text{ these } \mu_i\text{'s are not all equal}$$

Derivation of the test

- The mean square treatment is given by

$$\text{MSA} = \frac{\sum_{i=1}^{K} n_i(\bar{Y}_{i.} - \bar{Y})^2}{K-1}$$

where

$$\bar{Y} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_i} Y_{ij}}{N}$$

- A large standardized value of MSA indicates that H_0 is false.
- MSE is standardized using the pooled estimate of σ^2 which is estimated as:

$$\text{MSE} = s_p^2 = \frac{\sum_{i=1}^{K} (n_i - 1)s_i^2}{\sum_{i=1}^{K} (n_i - 1)}$$

Review: F distribution

If X_1 and X_2 are independent rvs with $X_1 \sim \chi^2_{v_1}$ and $X_2 \sim \chi^2_{v_2}$, then

$$\frac{X_1/v_1}{X_2/v_2} \sim F_{v_1,v_2}$$

Note:

- If $F \sim F_{v_1,v_2}$, then $1/F \sim F_{v_2,v_1}$
- Thus, if the F-table only gives the upper bound $F_{v_1,v_2,\alpha,U}$, i.e., $P(F \geq F_{v_1,v_2,\alpha,U}) = \alpha$, the lower bound can be obtained using the relationship above.

ANOVA: F test

- It can be shown under H_0:

$$(N-K)\text{MSE}/\sigma^2 \sim \chi^2_{N-K}$$

$$(K-1)\text{MSA}/\sigma^2 \sim \chi^2_{K-1}$$

  and MSE and MSA are independent

- Therefore, under H_0,

$$F_0 \equiv \frac{\text{MSA}}{\text{MSE}} \sim$$

ANOVA: F test

- It can be shown that $E(\text{MSE}) = \sigma^2$ whereas

$$E(\text{MSA}) = \sigma^2 + \frac{\sum_i n_i(\mu_i - \mu)^2}{K-1}$$

where $\mu = \frac{\sum_{i=1}^{K} n_i \mu_i}{N}$ is the overall mean

- Under H_0, $F_0 = 1$
- Under H_a, $F_0 > 1$
- Intuitively, we reject H_0 in favor of H_a if $F_0 \geq C$ where

$$P(\text{reject } H_0 | H_0) = P(F_0 \geq C | H_0) = \alpha$$

- The critical region:

$$C_\alpha = \{F_0 : F_0 > F_{K-1,N-K,\alpha,U}\}$$

- When $K = 2$, $H_0 : \mu_1 = \mu_2 \quad H_a : \mu_1 \neq \mu_2$

$$T_0 = \frac{\bar{y}_{1\cdot} - \bar{y}_{2\cdot}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \overset{H_0}{\sim} t_{n_1+n_2-2}$$

Cell Means Model

- The version of ANOVA model that we have looked at so far is called the *cell means model*

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

for $i = 1, 2, \ldots, K$; $j = 1, 2, \ldots, n_i$ where

$$\epsilon_{ij} \sim N(0, \sigma^2) \text{ for all } i, j$$

Factor Effects Model

- An equivalent model is the *factor effects model*

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

for $i = 1, 2, \ldots, K; \;\; j = 1, 2, \ldots, n_i$ where

$$\mu = \tfrac{1}{N} \sum_{i=1}^{K} n_i \mu_i \tag{1}$$

$$\alpha_i = \mu_i - \mu$$

and

$$\epsilon_{ij} \sim N(0, \sigma^2) \;\; \text{for all} \;\; i, j$$

- Constraint: $\sum_{i=1}^{K} n_i \alpha_i = 0$

Model Equivalence

- Equivalence of null hypotheses

$$H_0 : \mu_1 = \cdots = \mu_K \Leftrightarrow H_0 : \alpha_i = 0; \;\; i = 1, 2, \ldots, K$$

- α_i is called the i^{th} *main effect* or *factor effect*

$$\begin{aligned} Y_{ij} &= \mu_i + \epsilon_{ij} \\ &= \mu + (\mu_i - \mu) + \epsilon_{ij} \\ &= \mu + \alpha_i + \epsilon_{ij} \end{aligned}$$

ANOVA: Sum of Squares

- It can be shown that

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y})^2 = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(\bar{Y}_{i\cdot}-\bar{Y})^2 + \sum_{i=1}^{K}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\cdot})^2$$

- That is,

$$\text{SST} = \text{SSA} + \text{SSE}$$

ANOVA: F Test and ANOVA Table

ANOVA Table

Source	SS	df	MS	F
Among groups	SSA	$K-1$	$\text{MSA} = \frac{\text{SSA}}{K-1}$	MSA/MSE
Within groups	SSE	$N-K$	$\text{MSE} = \frac{\text{SSE}}{N-K}$	
Total	SST	$N-1$		

Example: A study was conducted to compare the lung function of groups of smokers and non-smokers. Test the hypothesis if the lung function differs by smoking status.

Group	n_i	Mean (L/sec)	sd (L/sec)
Non-smokers	200	3.78	0.79
Passive smokers	200	3.30	0.77
Non-inhalers	50	3.32	0.86
Light smokers	200	3.23	0.78
Mod. smokers	200	2.73	0.81
Heavy smokers	200	2.59	0.82