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:

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

$$
n_1 \begin{cases} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \end{cases}
\qquad
n_i \begin{cases} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{cases}
\qquad
n_K \begin{cases} Y_{K1} \\ Y_{K2} \\ \vdots \\ Y_{Kn_K} \end{cases}
$$

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 $N = \sum_{i=1}^{K} n_i$
- $\bar{Y}_{i\cdot} = \sum_j Y_{ij}/n_i$

sample mean from i group.

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

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

versus

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

Derivation of the test

sum of squares treatment

- The mean square treatment is given by

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

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:

sum of square error

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

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_{ν_2,ν_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.

F_{ν_2,ν_1}

$$\alpha = P(F \geq F_{\nu_1,\nu_2,\alpha,U}) = P\left(\frac{1}{F} \leq \frac{1}{F_{\nu_1,\nu_2,\alpha,U}}\right)$$

$$\Rightarrow \frac{1}{F_{\nu_1,\nu_2,\alpha,U}} = F_{\nu_2,\nu_1,\alpha,L}$$

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,

$$\frac{(K-1)\text{MSA}/\sigma^2/K-1}{(N-K)\text{MSE}/\sigma^2/N-K} \Rightarrow F_0 \equiv \frac{\text{MSA}}{\text{MSE}} \sim F_{K-1,N-K}$$

ANOVA: F test

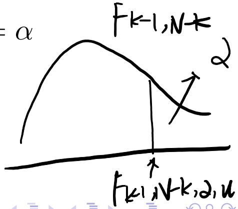
- It can be shown that $E(\text{MSE}) = \sigma^2$ whereas

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

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

- Under H_0, $F_0 = 1$ since $\mu_1 = \mu_2 = \cdots = \mu_K$
- 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 $K = 2$, $H_0: \mu_1 = \mu_2 \quad H_a: \mu_1 \neq \mu_2$

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

Note: If $T \sim t_k$, $T^2 \sim F_{1,k}$

$$T = \frac{Z}{\sqrt{W/k}} \Rightarrow T^2 = \frac{Z^2}{W/k}$$

$Z \sim N(0,1) \quad Z \perp W \qquad Z^2 = \chi^2_1$

$W \sim \chi^2_k$

ANOVA: Sum of Squares

- It can be shown that

$$\underbrace{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y})^2}_{\text{SST}} = \overbrace{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(\bar{Y}_{i\cdot}-\bar{Y})^2}^{\text{SSA}} + \overbrace{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\cdot})^2}^{\text{SSE}}$$

- That is, SST = SSA + SSE

ANOVA: F Test and ANOVA Table

p-value

ANOVA Table

Source of variation	df	MS	F
Among groups	$K-1$	$\text{MSA} = \frac{\sum_{i=1}^{K} n_i(\bar{Y}_{i\cdot} - \bar{Y})^2}{K-1}$	MSA/MSE
Within groups	$N-K$	$\text{MSE} = \frac{\sum_{i=1}^{K}(n_i-1)s_i^2}{N-K}$	
Total	$N-1$		

$$MSE = \sum_{i=1}^{K} \frac{(n_i-1)s_i^2}{N-K} = \frac{199(0.79)^2 + \cdots + 199(0.82)^2}{1050-6} = 0.636$$

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 $\bar{Y}_1$	0.79 s_1
Passive smokers	200	3.30 $\bar{Y}_2$	0.77 s_2
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

$$F_0 = \frac{MSA}{MSE} = 57.98 > F_{5,1044,0.05,u} = 2.22$$

$\therefore$ Reject H_0 & conclude that