VE401 Recitation Class Note

Final Part3

Chen Siyi
siyi.chen_chicy@sjtu.edu.cn

1 Basic Model of ANOVA

Population	Responses					Total	Mean
1	Y_{11}	Y_{12}	Y_{13}	$\dots$	Y_{1n_1}	$T_{1\cdot}$	$\overline{Y}_{1\cdot}$
2	Y_{21}	Y_{22}	Y_{23}	$\dots$	Y_{2n_2}	$T_{2\cdot}$	$\overline{Y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
k	Y_{k1}	Y_{k2}	Y_{k3}	$\dots$	Y_{kn_k}	$T_{k\cdot}$	$\overline{Y}_{k\cdot}$
Overall						$T_{\cdot\cdot}$	$\overline{Y}_{\cdot\cdot}$

1.1 Settings

Goal:
Testing H_0: $\mu_1 = \cdot\ \cdot\ \cdot = \mu_k$

Notation:

1. k populations, Y | i = Y_i N(μ_i,σ^2), i = 1,...,k.
2. Sample sizes n_i. Totally N = $n_1 + \cdot\ \cdot\ \cdot + n_k$ observations.
3. Y_{ij} is the jth response (measurement) for the i th population.
4. $T_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}$ $\quad \overline{Y}_{i\cdot} = \frac{T_{i\cdot}}{n_i}$
5. $T_{i\cdot\cdot} = \sum_{i=1}^{k} T_{i\cdot}$ $\quad \overline{Y_{i\cdot\cdot}} = \frac{T_{\cdot\cdot}}{N}$

Assumption:
Y_{ij} are independent, normally distributed random variables with mean μ_i and variance σ^2 (independent of i and j).

1.2 Model

$$Y_{ij} = \mu_{ij} + E_{ij} = \mu + \alpha_i + E_{ij}$$

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$$

Unbiased Estimators:

$$\widehat{\mu} = \overline{Y_{\cdot\cdot}}, \quad \widehat{\mu}_i = \overline{Y_{i\cdot}}, \quad \widehat{\alpha}_i = \widehat{\mu}_i - \widehat{\mu}$$

Then

$$\begin{aligned} Y_{ij} &= \overline{Y_{\cdot\cdot}} + \left(\overline{Y}_{i.} - \overline{Y_{\cdot\cdot}}\right) + \left(Y_{ij} - \overline{Y}_{i.}\right) \\ &= \widehat{\mu} + \widehat{\alpha}_i + e_{ij} \end{aligned}$$

1.3 Sum of Squares(SS) and Mean Square(MS)

Error Sum of Squares:

$$\mathrm{SS_E} := \sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i}\cdot\right)^2$$

Total Sum of Squares:

$$\begin{aligned} \mathrm{SS_T} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}\cdot\cdot\right)^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\overline{Y}_{i.} - \overline{Y}\cdot\cdot\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i.}\right)^2 \\ &= SS_{Tr} + SS_E \end{aligned}$$

Treatment Sum of Squares:

$$SS_{Tr} = \sum_{j=1}^{n_i}\left(\overline{Y}_{i.} - \overline{Y}\cdot\cdot\right)^2$$

Analysis on SS_E:

Since

$$S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i}\cdot\right)^2, \quad \mathrm{SS_E} := \sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i}\cdot\right)^2$$

hence

$$\mathrm{SS_E} = \sum_{i=1}^{k}\left(n_i - 1\right) S_i^2$$

Since

$$\chi^2_{n_i-1} = \frac{(n_i-1)\,S_i^2}{\sigma^2}$$

hence

$$\chi^2_{N-k} = \frac{SS_E}{\sigma^2}$$

Error Mean Square:

$$\mathrm{MS_E} = \frac{\mathrm{SS_E}}{N-k}$$

$E[\mathrm{MS_E}] = \sigma^2$, is an unbiased estimator for the variance.(Why?)

Treatment Mean Square:

$$\mathrm{MS_{Tr}} := \frac{\mathrm{SS_{Tr}}}{k-1}$$

1.4 Statistic

Theorem1:

$$\frac{(N-k)\mathrm{MS_E}}{\sigma^2} = \frac{\mathrm{SS_E}}{\sigma^2}$$

follows a chi-squared distribution with N − k degrees of freedom. (Why?)

Theorem2:

If $\alpha_1 = \cdot\ \cdot\ \cdot = \alpha_k = 0$,then

$$\frac{\mathrm{SS_{Tr}}}{\sigma^2} = \frac{(k-1)\mathrm{MS_E}}{\sigma^2}$$

follows a chi-squared distribution with k − 1 degrees of freedom.

Theorem3:

SS_E and SS_{Tr} are independent.

Statistic:

If $\alpha_1 = \cdot\ \cdot\ \cdot = \alpha_k = 0$,then

$$F_{k-1,N-k} = \frac{MS_{Tr}}{MS_E}$$

follows an F -distribution with k − 1 and N − k degrees of freedom.

1.5 ANOVA F-Test

Suppose Y_1,...,Y_k are normally distributed random variables with common variance σ^2. Suppose that samples of size n_1, ... , n_k are taken from each population and N $= n_1 + \cdot\ \cdot\ \cdot + n_k$. Then we reject

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

in favor of

$$H_1 : \mu_i \neq \mu_j \text{ for at least one pair } (i, j), 1 \leq i < j \leq k$$

at significance level α if the test statistic

$$F_{k-1,N-k} = \frac{MS_{Tr}}{MS_E}$$

satisfies $F_{k-1,N-k} > f_{\alpha,k-1,N-k}$. (Or when the P-value is small enough.)

1.6 Afterwards Comments

1. Remember to give a conclusion.
2. When the sample size is reasonably large, if Y_{ij} are not strictly normally distributed, the result of the F-test is still reliable
3. But it is essential for the F-test that all populations have equal variance.
4. To determine specifically which seams have a higher sulphur content requires pairwise comparisons of means.

2 Test for Equality of Variances

Goal:

We wish to test

$$H_0 : \sigma_1^2 = \cdots = \sigma_k^2$$

Define:

$$Q := (N - k) \ln MS_E - \sum_{i=1}^{k} (n_i - 1) \ln S_i^2$$

$$h := 1 + \frac{1}{3(k-1)} \left(\sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{N - k} \right)$$

Bartlett statistic:

$$B := \frac{Q}{h}$$

follows approximately a chi-squared distribution with k − 1 degrees of freedom.

Bartlett's Test:

Suppose Y_1, ... , Y_k are normally distributed random variables. Suppose that samples of size n_1, ... , n_k are taken from each population and N = $n_1 + \cdot \cdot \cdot + n_k$. Then we reject

$$H_0 : \sigma_1^2 = \cdots = \sigma_k^2$$

in favor of

$$H_1 : \sigma_i^2 \neq \sigma_j^2 \text{ for at least one pair } (i, j), 1 \leq i < j \leq k$$

at significance level α if the test statistic

$$B = \frac{Q}{h}$$

satisfies

$$B > \chi^2_{\alpha, k-1}$$

3 Pairwise Comparisons (Post hoc Tests)

Assuming we have established the equality of variances and obtained an ANOVA table showing a significant difference in the k treatment means, we would like to know which particular treatment means differ and which are statistically similar.

The canonical strategy is to then perform pairwise tests.

$$H_0 : \mu_i = \mu_j, \quad H_1 : \mu_i \neq \mu_j$$

3.1 Fisher's Least Significant Difference(LSD) Test

Statistic

Compared with the statistic in the pooled T-Test:

$$T_{n_i+n_j-2} = \frac{\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot}}{S_p\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}$$

With the estimator $\widehat{\sigma^2} = \mathrm{MS_E}$, we have:

$$T_{N-k} = \frac{\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot}}{\sqrt{MS_E\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

Fisher's LSD Test:

Suppose that $Y_{i.}$, i = 1, ... , k , are the populations means in an ANOVA that yielded a significant F -test. Then for each i , j = 1, ... , k we reject

$$H_0 : \mu_i = \mu_j \quad \text{in favor of} \quad H_1 : \mu_i \neq \mu_j$$

at significance level α if

$$\left|\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot}\right| > \tau_{ij} := t_{\alpha/2,N-k} \cdot \sqrt{\mathrm{MS}_E\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Here τ_{ij} is called the least significant difference (LSD).

3.2 Bonferroni's Test

When perform Fisher LSD tests for all possible combinations of means.

Error Rates:

1. α is the comparisonwise error rate,
2. α' is the overall or experimentwise error rate.

If the test are independent:

$$\alpha' = 1 - (1-\alpha)^m$$

No matter whether the tests are independent:

$$\alpha' \leq m\alpha$$

m is the total number of tests, $\frac{k(k-1)}{2}$.

Bonferroni's Test:

A Fisher LSD Test of k means performed at a comparisonwise error rate of

$$\alpha = \frac{\alpha'}{k(k-1)/2}$$

is said to be a Bonferroni test with controlled experimentwise error rate α'.

The least significant differences τ_{ij} are sometimes called Bonferroni critical points.

3.3 Tukey's Honestly Significant Difference (HSD) Test

Studentized Sample Range:

$$R_n = \frac{\max\{X_1, \dots, X_n\} - \min\{X_1, \dots, X_n\}}{\hat{\sigma}}$$

where $\widehat{\sigma^2}$ is an estimator for σ^2. For example, one could take the usual sample standard deviation $\hat{\sigma} = S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$.

Least Significant Studentized Ranges $r_{n,\alpha,\gamma}$:

$$P\left[R_n > r_{n,\alpha,\gamma}\right] = \alpha$$

R_n in ANOVA and Tests:

Suppose $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ is true and $n_1 = n_2 = \cdots = n_k$, then $\overline{Y_i}$ has variance $\frac{\sigma^2}{n}$. And we estimate:

$$\frac{\widehat{\sigma^2}}{n} = \frac{MS_E}{n}$$

$$R_n = \frac{\max\left\{\overline{Y}_{1\cdot}, \dots, \overline{Y}_{k\cdot}\right\} - \min\left\{\overline{Y}_{1\cdot}, \dots, \overline{Y}_{k\cdot}\right\}}{\sqrt{MS_E/n}}$$

With probability $1-\alpha$, all means will satisfy

$$\left|\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot}\right| \leq r_{k,\alpha,N-k}\sqrt{MS_E/n}$$

Further we extend this method to unequal sample sizes by set:

$$\frac{1}{n^*} = \frac{1}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right) \quad \Leftrightarrow \quad n^* = \frac{2n_i n_j}{n_i + n_j}$$

Based on this, we reject $H_0 : \mu_i = \mu_j$ for any i , j if

$$\left|\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot}\right| > r_{k,\alpha,N-k}\sqrt{MS_E/n^*}$$

Summary-Tukey's HSD Test:

Suppose that $Y_{i\cdot}$, i = 1, ... , k , are the populations means in an ANOVA that yielded a significant F -test. Then for each i , j = 1, ... , k we reject

$$H_0 : \mu_i = \mu_j \quad \text{in favor of} \quad H_1 : \mu_i \neq \mu_j$$

at an experimentwise error rate α' if

$$\left|\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot}\right| > r_{k,\alpha',N-k}\sqrt{\frac{MS_E}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

3.4 Duncan's New Multiple Range Test

Duncan's test modifies Tukey's test as follows: instead of conducting pairwise comparisons, we compare groups of means at once. For this reason it is said to be a multiple range test.

Also Duncan's test use a modification of the least significant range $r'_{p,\alpha'}N - k$. p is the number of means in the chosen group, k is the number of means in total.

We try to link as much $\overline{Y_i}$ and $\overline{Y_j}$ as we can. Linking means small difference. We link from large groups to small groups so that we can perform less comparisons.

Requirements:

1. All treatment sample sizes are equal
2. Order sample means from smallest to largest

Steps/Example:

32.8. Example. Suppose we have the following data:

Population i	Response Y_{ij}					
1	0.30	0.64	0.68	0.53	0.78	0.84
2	0.69	1.08	1.30	1.11	0.51	0.96
3	0.96	0.93	1.02	1.19	0.89	1.33
4	1.21	0.96	1.02	1.17	1.28	1.36
5	1.70	1.16	0.95	1.18	1.56	1.07
6	1.52	1.34	1.58	1.41	1.39	1.46

We find

$$\overline{Y}_{1\cdot} = 0.63, \qquad \overline{Y}_{2\cdot} = 0.94, \qquad \overline{Y}_{3\cdot} = 1.05,$$
$$\overline{Y}_{4\cdot} = 1.17, \qquad \overline{Y}_{5\cdot} = 1.27, \qquad \overline{Y}_{6\cdot} = 1.45.$$

and

$$\mathrm{MS}_\mathrm{E} = 0.0450.$$

We have $k = 6$ populations with $n = 6$ measurements each, so $N = 36$ and $N - k = 30$. The least significant ranges for $\alpha = 0.01$ and 30 degrees of freedom are given by

$$r'_{2,0.01,30} = 3.889, \qquad r'_{3,0.01,30} = 4.056, \qquad r'_{4,0.01,30} = 4.168,$$
$$r'_{5,0.01,30} = 4.250, \qquad r'_{6,0.01,30} = 4.314.$$

The critical differences are given by $d_p = r'_{p,0.01,30}\sqrt{\mathrm{MS}_\mathrm{E}/n}$, i.e.,

$$d_2 = 0.337, \quad d_3 = 0.351, \quad d_4 = 0.361, \quad d_5 = 0.368, \quad d_6 = 0.373.$$

$$\overline{Y}_{6\cdot} - \overline{Y}_{1\cdot} = 1.45 - 0.63 = 0.82 > d_6 = 0.373$$

so we conclude that there is a significant difference among the means.

Next, we consider two groups of five means,

$$\overline{Y}_{5.} - \overline{Y}_{1.} = 1.27 - 0.63 = 0.64 > d_5 = 0.368,$$
$$\overline{Y}_{6.} - \overline{Y}_{2.} = 1.45 - 0.94 = 0.51 > d_5 = 0.368,$$

so there significant differences in these groups also. We next test groups of four means,

$$\overline{Y}_{4.} - \overline{Y}_{1.} = 1.17 - 0.63 = 0.64 > d_4 = 0.361,$$
$$\overline{Y}_{6.} - \overline{Y}_{3.} = 1.45 - 1.05 = 0.40 > d_4 = 0.361,$$
$$\overline{Y}_{5.} - \overline{Y}_{2.} = 1.27 - 0.94 = 0.33 \not> d_4 = 0.361.$$

We stop testing those groups were there are no significant differences among the means. This leaves

$$\overline{Y}_{6.} - \overline{Y}_{4.} = 1.45 - 1.17 = 0.28 \not> d_3 = 0.337,$$
$$\overline{Y}_{3.} - \overline{Y}_{1.} = 1.05 - 0.63 = 0.42 > d_3 = 0.337,$$
$$\overline{Y}_{2.} - \overline{Y}_{1.} = 1.05 - 0.94 = 0.11 \not> d_2 = 0.337.$$

We summarize by underlining those groups of means that are not significantly different:

$$\overline{Y}_{1.} \quad \overline{Y}_{2.} \quad \overline{Y}_{3.} \quad \overline{Y}_{4.} \quad \overline{Y}_{5.} \quad \overline{Y}_{6.}$$

We conclude: there is evidence that μ_1 is significantly smaller that μ_3, μ_4, μ_5 and μ_6. Furthermore, μ_2 and μ_3 are significantly smaller than μ_6. There is no significant difference between any other means.

4 Tips for Final

1. Prepare certain mma codes, but don't rely on it. Horst says all the problems can be solved without mma.
2. If you use mma or other tools, zip your codes, and write certain explanations.
3. Understand essential concepts, be flexible.


逢考必过
EVERY EXAM MUST PASS.
good luck