MATH 2411 T1B Tutorial 10

HOU Zhen

zhouah@connect.ust.hk

April 28, 2026

Analysis of Variance

ANOVA model

Our assumption setting is:

- Group 1: $Y_{1,1}, Y_{1,2}, \cdots, Y_{1,n_1}$ iid samples from $N(\mu_1, \sigma^2)$.
- Group 2: $Y_{2,1}, Y_{2,2}, \cdots, Y_{2,n_2}$ iid samples from $N(\mu_2, \sigma^2)$.
- $\cdots$
- Group k: $Y_{k,1}, Y_{k,2}, \cdots, Y_{k,n_k}$ iid samples from $N(\mu_k, \sigma^2)$.
- k is the number of levels of the factor.
- Samples from different groups are independent and σ^2 is same but unknown.
- $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ vs H_1 : some means are different.

Types of variance

- Total variance, SST (sum of square total):

$$\mathrm{SST} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{\cdot\cdot})^2.$$

- Between (group) variance, $\mathrm{SS}_{\mathrm{Treat}}$ (sum of square treatment):

$$\mathrm{SS}_{\mathrm{Treat}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2. = \sum_{i=1}^{k} n_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2.$$

- Within (group) variance, SSE (sum of square error):

$$\mathrm{SSE} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2.$$

- $\mathrm{SST} = \mathrm{SS}_{\mathrm{Treat}} + \mathrm{SSE}.$

ANOVA table

- $\mathrm{MS}_{\mathrm{Treat}} = \dfrac{\mathrm{SS}_{\mathrm{Treat}}}{k-1}$, $\mathrm{MSE} = \dfrac{\mathrm{SSE}}{n-k}$, $\mathrm{MST} = \dfrac{\mathrm{SST}}{n-1}$.
- Under H_0, $\mathbb{E}(\mathrm{MS}_{\mathrm{Treat}}) = \mathbb{E}(\mathrm{MSE}) = \mathbb{E}(\mathrm{MST}) = \sigma^2$.
- $F = \dfrac{\mathrm{SS}_{\mathrm{Treat}}/(k-1)}{\mathrm{SSE}/(n-k)} = \dfrac{\mathrm{MS}_{\mathrm{Treat}}}{\mathrm{MSE}}$.
- Under H_0, $F \sim F(k-1, n-k)$.
- Rejection region: $F > F_{k-1,n-k,\alpha}$.

R code: `qf(1 - alpha, df1 = k - 1, df2 = n - k)`

Source	d.f.	SS	MS	F-value
Between	$k-1$	$\mathrm{SS}_{\mathrm{Treat}}$	$\mathrm{MS}_{\mathrm{Treat}} = \mathrm{SS}_{\mathrm{Treat}}/(k-1)$	$F = \mathrm{MS}_{\mathrm{Treat}}/\mathrm{MSE}$
Within	$n-k$	SSE	$\mathrm{MSE} = \mathrm{SSE}/(n-k)$	
Total	$n-1$	SST	$\mathrm{MST} = \mathrm{SST}/(n-1)$	

Exercise

To study if exam performance is affected by the background sound, 12 student volunteers from MATH2411 class are randomly assigned to 3 exam rooms to complete the same standardized test in statistics, each exam room has 4 students.

Rock music is played in Room X, light music is played in Room Y, while there is no special background sound in Room Z. The test scores of the 12 students are shown below:

Room X	Room Y	Room Z
50	75	65
55	65	50
45	60	65
40	60	70

Solution

- $k = 3,\ n_1 = 4,\ n_2 = 4,\ n_3 = 4,\ n = n_1 + n_2 + n_3 = 12.$
- Group means:

$$\bar{x} = \frac{50+55+45+40}{4} = 47.5, \quad \bar{y} = \frac{75+65+60+60}{4} = 65,$$
$$\bar{z} = \frac{65+50+65+70}{4} = 62.5, \quad M = \frac{n_1\bar{x}+n_2\bar{y}+n_3\bar{z}}{n} = 58\frac{1}{3}.$$

- SS:

$$\mathrm{SS}_{\mathrm{Treat}} = n_1(\bar{x}-M)^2 + n_2(\bar{y}-M)^2 + n_3(\bar{z}-M)^2 = 716\frac{2}{3},$$
$$\mathrm{SSE} = \sum_{i=1}^{n_1}(x_i-\bar{x})^2 + \sum_{i=1}^{n_2}(y_i-\bar{y})^2 + \sum_{i=1}^{n_3}(z_i-\bar{z})^2 = 500.$$

- Degrees of freedom: $df_1 = k - 1 = 2$, $df_2 = n - k = 9$.
- Mean squares:

$$\mathrm{MS}_{\mathrm{Treat}} = \frac{716\frac{2}{3}}{2} = 358\frac{1}{3}, \quad \mathrm{MSE} = \frac{500}{9} = 55\frac{5}{9}.$$

- Test statistic:

$$F = \frac{\mathrm{MS}_{\mathrm{Treat}}}{\mathrm{MSE}} = \frac{358\frac{1}{3}}{55\frac{5}{9}} = 6.45.$$

- F-quantile: $F_{2,9,0.05} = 4.256$.
- As $F = 6.45 > 4.256$, we reject H_0 at significance level 0.05.

Welch's ANOVA

- If equal-variance assumption fails, we use Welch's ANOVA.
- Weighted mean setup:

$$w_i = \frac{n_i}{s_i^2}, \qquad \bar{y}_W = \frac{\sum_{i=1}^k w_i \bar{y}_{i\cdot}}{\sum_{i=1}^k w_i}, \qquad \mathrm{SS}_{\mathrm{Treat},W} = \sum_{i=1}^k w_i(\bar{y}_{i\cdot} - \bar{y}_W)^2.$$

- Test statistic:

$$F_W = \frac{\mathrm{MS}_{\mathrm{Treat},W}}{1 + \dfrac{2\Lambda(k-1)}{3}}, \qquad \Lambda = \frac{3\sum_{i=1}^k \dfrac{\left(1 - \frac{w_i}{\sum_{j=1}^k w_j}\right)^2}{n_i - 1}}{k-1}.$$

- Approximation and rejection rule:

$$F_W \approx F\left(k-1, \frac{1}{\Lambda}\right), \quad \text{reject } H_0 \text{ if } F_W > F_{k-1,\Lambda^{-1},\alpha}.$$

Relation to t-test

- **Theorem:** ANOVA and t-test are equivalent for $k = 2$. They have the same p-value, and
$$F = t^2.$$
- **Theorem:** Welch's ANOVA and Welch's t-test are equivalent when $k = 2$. They have the same p-value, and
$$F_W = t^2.$$

Exercise

Show that ANOVA and t-test are equivalent for $k = 2$.

Tukey's HSD

- For each pair (i, j), construct a $(1-\alpha)$ CI for $\mu_i - \mu_j$:

$$\left[\bar{x}_i - \bar{x}_j - q_{k,n-k,\alpha}\, s\sqrt{\frac{1}{n_i}+\frac{1}{n_j}},\ \bar{x}_i - \bar{x}_j + q_{k,n-k,\alpha}\, s\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}\right],$$

  where $s = \sqrt{\mathrm{MSE}}$ estimates σ, and $q_{k,n-k,\alpha}$ is from a special distribution family(range distribution).
- Decision rule: if $0 \notin$ the CI, conclude $\mu_i - \mu_j \neq 0$; the direction depends which side of the CI 0 lies.

Exercise: Type I Error Inflation

You want to compare **five different diets** to see whether they lead to different amounts of weight loss.

A friend suggests: "Instead of using ANOVA, just run separate t-tests for every pair of diets."

Suppose each individual t-test is conducted at significance level

$$\alpha = 0.05.$$

There are 10 pairwise comparisons in total. **For simplicity, assume the 10 tests are independent.**

Questions

1. For a single t-test, what is the probability of making a Type I error?
2. What is the probability of making **no Type I error** across all 10 tests?
3. What is the probability of making **at least one Type I error** across all 10 tests?
4. Interpret the result in the context of comparing five diets.

Solution

1. For one test, $P(\text{Type I error}) = \alpha = 0.05$.
2. For one test, $P(\text{no Type I error}) = 1 - \alpha = 0.95$. Hence across 10 independent tests,

$$P(\text{no Type I error in all 10}) = (1-\alpha)^{10} = 0.95^{10} \approx 0.5987.$$

3. $$P(\text{at least one Type I error}) = 1 - P(\text{none}) = 1 - 0.95^{10} \approx 0.4013.$$

4. Interpretation: even if all null hypotheses are true, doing 10 separate tests at level 0.05 gives about a **40%** chance of at least one false positive, so multiple testing inflates Type I error.

Linear Regression

Linear Regression Model

Suppose we have n samples $(x_1, y_1), \ldots, (x_n, y_n)$ and assume

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \text{ i.i.d.} \sim N(0, \sigma^2).$$

- β_1 is the slope.
- β_0 is the intercept.

Least Squares Estimation

The least squares estimators are

$$\hat{\beta}_0, \hat{\beta}_1 = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Their closed forms are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Useful Sums

Define

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2,$$

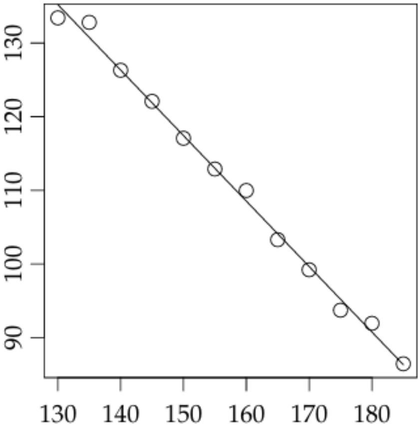
$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y},$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2.$$

Hence

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Exercise



A linear regression plot is given. Which parameter set is most plausible?

(A) $\beta_0 = 0,\ \beta_1 = -0.9,\ \sigma = 36$

(B) $\beta_0 = 0,\ \beta_1 = 0.9,\ \sigma = 3.6$

(C) $\beta_0 = 252,\ \beta_1 = -0.9,\ \sigma = 3.6$

(D) $\beta_0 = -252,\ \beta_1 = -0.9,\ \sigma = 36$

(E) $\beta_0 = 252,\ \beta_1 = -0.9,\ \sigma = 36$

Exercise

The i-th residual is $e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$.
Prove that:

- $\sum_{i=1}^{n} e_i = 0$.
- The regression line goes through $(\bar{x}, \bar{y})$.
- $\sum_{i=1}^{n} x_i e_i = 0$.
- $\sum_{i=1}^{n} \hat{y}_i e_i = 0$.