ST102 Week 19

Analysis of Variance (ANOVA)

One-way ANOVA

Given k independent random samples, i.e. for $j \in [k]$, i.i.d. random sample $\{X_{ij}\}_{i=1}^{n_j}$ from $N(\mu_j, \sigma^2)$. Test:

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

$$H_1: \text{not all } \mu_j\text{'s are the same}$$

Some definitions:

1) The j-th sample mean: $\bar{X}_{\cdot j} := \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}$

2) Overall sample mean: $\bar{X} := \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{ij} = \frac{1}{n} \sum_{j=1}^{k} n_j \bar{X}_{\cdot j}$

where $n := \sum_{j=1}^{k} n_j$.

3) Total variation: $\sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2$ with $df = n-1$

4) Between-groups variation: $B := \sum_{j=1}^{k} n_j (\bar{X}_{\cdot j} - \bar{X})^2$, $df = k-1$

5) Within-groups variation (residual sum of squares):

$$W := \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{\cdot j})^2, \quad df = n-k$$

6) Total variation / total sum of squares: $\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\overline{X})^2$

7) The ANOVA decomposition: Total variation $= B + W$.

8) Equivalent forms: Total variation $= \sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ij}^2 - n\overline{X}^2$

$$B = \sum_{j=1}^{k} n_j \overline{X}_{\cdot j}^2 - n\overline{X}^2$$

$$W = \sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ij}^2 - \sum_{j=1}^{k} n_j \overline{X}_{\cdot j}^2$$

$$= \sum_{j=1}^{k}(n_j - 1)S_j^2$$

Under H_0, it can be shown that:

$$F := \frac{\sum_{j=1}^{k} n_j(\overline{X}_{\cdot j} - \overline{X})^2/(k-1)}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_{\cdot j})^2/(n-k)} = \frac{B/(k-1)}{W/(n-k)} \sim F_{k-1,n-k}$$

Reject H_0 if $f > F_{k-1,n-k,\alpha}$ ($F_{\alpha;k-1,n-k}$).

One-way ANOVA table:

Source	DF	SS	MS	F	p-value
Factor	$k-1$	B	$B/(k-1)$	$\frac{B/(k-1)}{W(n-k)}$	p
Error	$n-k$	W	$W/(n-k)$		
Total	$n-1$	$B+W$			

Without further assumptions, we also have:

1) estimator of σ^2: $\hat{\sigma} := S := \sqrt{\frac{W}{n-k}}$

2) $\alpha \cdot 100\%$ - confidence interval for μ_j:

$$\bar{X}_{\cdot j} \pm t_{\frac{\alpha}{2}, n-k} \cdot \frac{S}{\sqrt{n_j}} \qquad \text{for } j \in [k]$$

<u>A system view of one-way ANOVA</u>

$$X_{ij} = \mu + \beta_j + \varepsilon_{ij} \qquad \text{for } i \in [n_j] \text{ and } j \in [k],$$

with $\varepsilon_{ij} \sim N(0, \sigma^2)$, all independent.
$\sum_j \beta_j = 0$ is required for identifiability.

μ is the average effect and β_j the j-th level treatment effect.

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

<u>Two-way ANOVA</u>

$$X_{ij} = \mu + \gamma_i + \beta_j + \varepsilon_{ij} \qquad \text{for } i \in [r] \text{ and } j \in [c].$$

μ: average treatment effect
β_j: treatment (column) levels
γ_i: different block (row) levels
ε_{ij}: follow $N(0, \sigma^2)$, all independent

Conditions for identifiability : $\sum_i \gamma_i = \sum_j \beta_j = 0$.

Two directions

$H_0: \beta_1 = \beta_2 = \cdots = \beta_c = 0$ $\qquad\qquad$ $H_0: \gamma_1 = \gamma_2 = \cdots = \gamma_r = 0$

Some definitions :

1) i-th block sample mean : $\bar{X}_{i\cdot} := \frac{1}{c} \sum_{j=1}^{c} X_{ij}$

2) j-th treatment sample mean : $\bar{X}_{\cdot j} := \frac{1}{r} \sum_{i=1}^{r} X_{ij}$

3) overall sample mean : $\bar{X} := \frac{1}{n} \sum_{j=1}^{c} \sum_{i=1}^{r} X_{ij}$

4) total variation : Total SS $:= \sum_{i=1}^{r} \sum_{j=1}^{c} (X_{ij} - \bar{X})^2$, $df = rc-1$

5) between-blocks variation : $B_{row} := c \sum_{i=1}^{r} (\bar{X}_{i\cdot} - \bar{X})^2$, $df = r-1$

6) between-treatments variation : $B_{col} := r \sum_{j=1}^{c} (\bar{X}_{\cdot j} - \bar{X})^2$, $df = c-1$

7) residual sum of squares : Residual SS $:= \sum_{i=1}^{r} \sum_{j=1}^{c} (X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})^2$

$df = (r-1)(c-1)$

8) two-way ANOVA decomposition : Total SS = Brow + Bcol + Residual SS

Equivalent forms: $\text{Total } SS = \sum_{i=1}^{r}\sum_{j=1}^{c} X_{ij}^2 - rc\overline{X}^2$

$$B_{row} = c\sum_{i=1}^{r} \overline{X}_{i\cdot}^2 - rc\overline{X}^2$$

$$B_{col} = r\sum_{j=1}^{c} \overline{X}_{\cdot j}^2 - rc\overline{X}^2$$

$$\text{Residual } SS = \text{Total } SS - B_{row} - B_{col}$$

Case I: $H_0: \gamma_1 = \gamma_2 = \cdots = \gamma_r = 0$

$$F := \frac{B_{row}/(r-1)}{\text{Residual } SS/[(r-1)(c-1)]} = \frac{(c-1)B_{row}}{\text{Residual } SS}$$

Under H_0, $F \sim F_{r-1,\,(r-1)(c-1)}$

reject H_0 if $f > F_{\alpha;\, r-1,\,(r-1)(c-1)}$

Case II: $H_0: \beta_1 = \beta_2 = \cdots = \beta_c = 0$

$$F := \frac{B_{col}/(c-1)}{\text{Residual } SS/[(r-1)(c-1)]} = \frac{(r-1)B_{col}}{\text{Residual } SS}$$

Under H_0, $F \sim F_{c-1,\,(r-1)(c-1)}$.

Reject H_0 if $f > F_{\alpha;\, c-1,\,(r-1)(c-1)}$

Two-way ANOVA table

Source	DF	SS	MS	F	p-value
Row factor	$r-1$	B_{row}	$B_{row}/(r-1)$	$(c-1)B_{row}/RSS$	p_1
Column factor	$c-1$	B_{col}	$B_{col}/(c-1)$	$(r-1)B_{col}/RSS$	p_2
Residual	$(r-1)(c-1)$	Residual SS	$\frac{\text{Residual } SS}{(r-1)(c-1)}$		
Total	$rc-1$	Total SS			

Remark on residuals:

By model we have $X_{ij} = \mu + \gamma_i + \beta_j + \varepsilon_{ij}$,

While from data we have the decomposition:

$$X_{ij} = \bar{X} + (\bar{X}_{i\cdot} - \bar{X}) + (\bar{X}_{\cdot j} - \bar{X}) + (X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})$$

$\Rightarrow \hat{\mu} = \bar{X}$

$\hat{\gamma}_i = \bar{X}_{i\cdot} - \bar{X}$

$\hat{\beta}_j = \bar{X}_{\cdot j} - \bar{X}$

$\hat{\varepsilon}_{ij} = X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}$