STAT 517

Analysis of Variance (ANOVA)

Prof. Michael Levine

April 14, 2016

Two-way ANOVA: an additive model

- There are two factors A and B with levels a and b, respectively
- The response is $X_{ij} \sim N(\mu_{ij}, \sigma^2)$ where $i = 1, \dots, a$ and $j = 1, \dots, b$
- The total sample size is $n = ab$
- The **additive model** is

$$\mu_{ij} = \bar{\mu} + (\bar{\mu}_{i.} - \bar{\mu}) + (\bar{\mu}_{.j} - \bar{\mu})$$

Two-way ANOVA model

- In the above, denote $\alpha_i = \bar{\mu}_{i.} - \bar{\mu}$, $\beta_j = \bar{\mu}_{.j} - \bar{\mu}$, $\mu = \bar{\mu}$
- The resulting form of the model is

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

  where $\sum_{i=1}^{a} \alpha_i = 0$ and $\sum_{j=1}^{b} \beta_j = 0$
- This is a **two-way ANOVA model** with constraints
- For each j, plots of μ_{ij} vs j are called **mean profile plots**
- They are always parallel for all additive models

Hypotheses of interest

- The first one is

$$H_{0A} : \alpha_1 = \cdots = \alpha_a = 0 \text{ vs. } H_{1A} : \alpha_i \neq 0 \text{ for some } i$$

- The second one is

$$H_{0B} : \beta_1 = \cdots = \beta_b = 0 \text{ vs. } H_{1B} : \beta_j \neq 0 \text{ for some } j$$

- If H_{0A} is true, the mean of the (i, j)th cell does not depend on the level of A
- These are **main effect** hypotheses

One-way equality of means test: a reminder

- Recall the equality of means test...the important quadratic forms were Q, Q_3 and Q_4
- We had $Q = Q_3 + Q_4$ or

$$(ab-1)S^2 = \sum_{j=1}^{b}\sum_{i=1}^{a}(\bar{X}_{.j} - \bar{X}_{..})^2 + \sum_{j=1}^{b}\sum_{i=1}^{a}(X_{ij} - \bar{X}_{.j})^2$$

- Here Q_4 is the sum of squares *among the column means* Q_3 is *within the column means*
- Q_3/ab is the $\hat{\sigma}^2_{\Omega}$ - MLE of σ^2 under Ω
- $\hat{\sigma}^2_{\omega} = (ab-1)S^2/ab$ is the MLE of σ^2 under ω
- Thus, the likelihood ratio is the monotone function of statistic

$$F = \frac{Q_4/(b-1)}{Q_3/[b(a-1)]}$$

A two-way equality of means test for columns

- A similar expansion is $Q = Q_2 + Q_4 + Q_5$ or

$$(ab-1)S^2 = \sum_{j=1}^{b}\sum_{i=1}^{a}(\bar{X}_{i.} - \bar{X}_{..})^2 + \sum_{j=1}^{b}\sum_{i=1}^{a}(\bar{X}_{.j} - \bar{X}_{..})^2$$

$$+ \sum_{j=1}^{b}\sum_{j=1}^{a}(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})^2$$

- Q_2 is the sum of squares among *rows*, Q_4 among *columns*, and Q_5 is the *remainder*
- $\hat{\sigma}^2_{\Omega} = Q_5/ab$ is the MLE of σ^2 under Ω

A two-way equality of means test for columns

- $$\hat{\sigma}^2_\omega = \frac{(Q_4 + Q_5)}{ab} = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(X_{ij} - \bar{X}_{i.})^2}{ab}$$

  is the MLE of σ^2 under ω
- A good monotone function of the likelihood ratio $\Lambda = (\sigma^2_\Omega/\hat{\sigma}^2_\omega)^{ab/2}$ is the F-statistic

  $$F = \frac{Q_4/(b-1)}{Q_5/[(a-1)(b-1)]} \sim F_{b-1,(a-1)(b-1)}$$

  under H_0

Power of the two-way ANOVA column mean test

- Under the alternative H_{1B}, due to Hogg-Craig's Theorem, Q_4/σ^2 and Q_5/σ^2 are independent
- Check that $E(X_{ij}) = \mu + \alpha_i + \beta_j$, $E(\bar{X}_{.j}) = \mu + \beta_j$, $E(\bar{X}_{..}) = \mu$
- Substitute means for real random variables in the quadratic form Q_4 to find the non-centrality parameter

$$\frac{a}{\sigma^2}\sum_{j=1}^{b}(\mu + \beta_j - \mu)^2 = \frac{a}{\sigma^2}\sum_{j=1}^{b}\beta_j^2$$

Power of the two-way ANOVA column mean test

- Do the same for Q_5 to obtain

$$\frac{1}{\sigma^2}\sum_{j=1}^{b}\sum_{i=1}^{a}(\mu+\alpha_i+\beta_j-\mu-\alpha_i-\mu-\beta_j+\mu)^2=0$$

- Thus, under H_{1B}, we have

$$F \sim F_{b-1,(a-1)(b-1)}\left(\frac{a}{\sigma^2}\sum_{j=1}^{b}\beta_j^2\right)$$

Two-way ANOVA row mean test

- A very similar argument leads to the row mean test
- The test statistic is

$$F = \frac{Q_2/(a-1)}{Q_5/[(a-1)(b-1)]} \sim F_{a-1,(a-1)(b-1)}$$

under H_{0A}

Two-way classification with multiple replicates

- The ANOVA problem just considered is called **two-way classification with one observation per cell**
- Sometimes, more than 1 observation is available for each of ab cells
- For simplicity, assume that each cell has the same number c replicates
- Notation: X_{ijk}, $i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, c$; the sample size is $n = abc$
- All of $X_{ijk} \sim N(\mu_{ij}, \sigma^2$

Interaction parameters

- Now, we can make a mean fully cell-specific
- Define

$$\gamma_{ij} = \mu_{ij} - \{\mu + (\bar{\mu}_{i.} - \mu) + (\bar{\mu}_{.j} - \mu)\} = \mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \mu$$

- γ_{ij} is an **interaction parameter**
- The cell means model is then

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

- The constraints are $\sum_{i=1}^{a} \alpha_i = 0$, $\sum_{j=1}^{b} \beta_j = 0$, $\sum_{i=1}^{a} \gamma_{ij} = 0$, $\sum_{j=1}^{b} \gamma_{ij} = 0$

Remarks

- Note that it is only due to the presence of replicates that we can estimate interaction parameters
- The additive constraints are not the only possible...another option is to assume that e.g. $\alpha_a = \beta_b = \gamma_{a1} = \gamma_{1b} = 0$

Hypothesis of interest

- $H_{0AB} : \gamma_{ij} = 0$ for all i, j vs. $H_{1AB} : \gamma_{ij} = 0$ for at least one pair (i, j)
- The general form of the test statistic is

$$F = \frac{\left[\sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{X}_{ij.} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})^2\right] / [(a-1)(b-1)]}{\left[\sum \sum \sum (X_{ijk} - \bar{X}_{ij.})^2\right] / [ab(c-1)]}$$

- Under H_{0AB}, we have $F \sim F_{(a-1)(b-1), ab(c-1)}$