26 Lecture 26: March 31

Last time

- One-way ANOVA

Today

- Announcement: alternative grading path didn't pass (5:5 from last poll + a fail on the first poll)
- Analysis of Variance (JF chapter 8)
  - two-way ANOVA

Additional reference

Course notes by Dr. Jason Osborne.

Two-Way ANOVA

The inclusion of a second factor permits us to model and test partial relationships, as well as to introduce interactions. Let's take a look at the patterns of relationship that can occur when a quantitative response variable is classified by two factors.

Patterns of Means in the two-way classification

Consider the following table:

	C_1	C_2	$\dots$	C_c	
R_1	μ_{11}	μ_{12}	$\dots$	μ_{1c}	$\mu_{1\cdot}$
R_2	μ_{21}	μ_{22}	$\dots$	μ_{2c}	$\mu_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
R_r	μ_{r1}	μ_{r2}	$\dots$	μ_{rc}	$\mu_{r\cdot}$
	$\mu_{\cdot 1}$	$\mu_{\cdot 2}$	$\dots$	$\mu_{\cdot c}$	$\mu_{\cdot\cdot}$

The factors, R and C (for "rows" and "columns" of the table of means), have r and c categories, respectively. The factor categories are denoted R_j and C_k. Within each cell of the design - that is, for each combination of categories $\{R_j, C_k\}$ of the two factors - there is a population cell mean μ_{jk} for the response variable. Extending the dot notation, we have

$$\mu_{j\cdot} \equiv \frac{\sum_{k=1}^{c} \mu_{jk}}{c}$$

is the marginal mean of the response variable in row j.

$$\mu_{\cdot k} \equiv \frac{\sum_{j=1}^{r} \mu_{jk}}{r}$$

is the marginal mean in column k. And

$$\mu_{\cdot\cdot} \equiv \frac{\sum_j \sum_k \mu_{jk}}{r \times c}$$

is the grand mean.

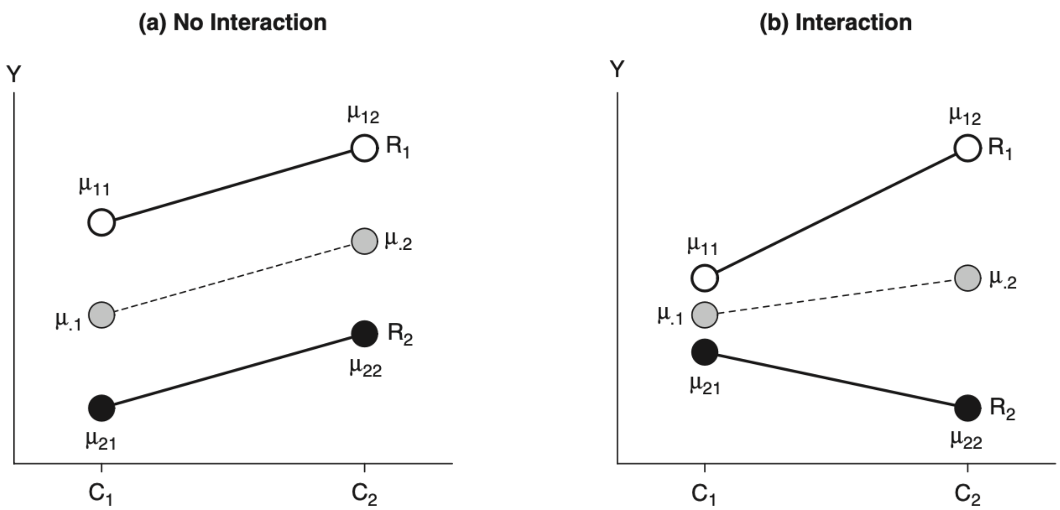

Figure 26.1: Interaction in the two-way classification. In (a), the parallel profiles of means (given by the white and black circles connected by solid lines) indicate that R and C do not interact in affecting Y. The R-effect – that is, the difference between the two profiles – is the same at both C_1 and C_2. Likewise, the C-effect – that is , the rise in the line from C_1 to C_2 – is the same for both profiles. In (b), the R-effect differs at the two categories of C, and the C-effect differs at the two categories of R: R and C interact in affecting Y. In both graphs, the column marginal means $\mu_{\cdot 1}$ and $\mu_{\cdot 2}$ are shown as averages of the cell means in each column (represented by the gray circles connected by broken lines). JF Figure 8.2.

Two-way ANOVA model

The two-way ANOVA model, suitably defined, provides a convenient means for testing the hypotheses concerning interactions and main effects. The model is

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \epsilon_{ijk}$$

where Y_{ijk} is the ith observation in row j, column k of the RC table; μ is the general mean of Y; α_j and β_k are the main-effect parameters; γ_{jk} are interaction effect parameters; and

ϵ_{ijk} are errors satisfying the usual linear-model assumptions (i.e. $\epsilon_{ijk} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$). By taking expectations, we have

$$\mu_{jk} \equiv E(Y_{ijk}) = \mu + \alpha_j + \beta_k + \gamma_{jk}$$

We have $r \times c$ population cell means with $1 + r + c + r \times c$ model parameters. Similar to one-way ANOVA model, we add in additional constraints to make the model identifiable.

$$\begin{aligned}
&\sum_{j=1}^{r} \alpha_j = 0 \\
&\sum_{k=1}^{c} \beta_k = 0 \\
&\sum_{j=1}^{r} \gamma_{jk} = 0 \quad \text{for all } k = 1, \ldots, c \\
&\sum_{k=1}^{c} \gamma_{jk} = 0 \quad \text{for all } j = 1, \ldots, r
\end{aligned}$$

The constraints produce the following solution for model parameters in terms of population cell and marginal means (and we add a hat for their estimates using the sample means):

$$\begin{aligned}
\mu &= \mu_{\cdot\cdot} \\
\alpha_j &= \mu_{j\cdot} - \mu_{\cdot\cdot} \\
\beta_k &= \mu_{\cdot k} - \mu_{\cdot\cdot} \\
\gamma_{jk} &= \mu_{jk} - \mu - \alpha_j - \beta_k \\
&= \mu_{jk} - \mu_{j\cdot} - \mu_{\cdot k} + \mu_{\cdot\cdot}
\end{aligned}$$

Hypotheses with two-way ANOVA

Some interesting hypotheses:

1. Are the cell means all equal? (Equivalent to one-factor ANOVA's "overall F-test")
   $H_0 : \mu_{11} = \mu_{12} = \cdots = \mu_{rc}$ vs. H_a : At least two μ_{ij} differ
2. Are the marginal means for row main effect equal?
   $H_0 : \mu_{1\cdot} = \mu_{2\cdot} = \cdots = \mu_{r\cdot}$ vs H_a : At least two $\mu_{j\cdot}$ differ
   which is equivalent as testing for no row main effects H_0 : all $\alpha_j = 0$ (why?)
3. Are the marginal means for column main effect equal?
   $H_0 : \mu_{\cdot 1} = \mu_{\cdot 2} = \cdots = \mu_{\cdot c}$ vs H_a : At least two $\mu_{\cdot k}$ differ
4. Do the factors interact? In other words, does effect of one factor depend on the other factor? $H_0 : \mu_{ij} = \mu_{\cdot\cdot} + (\mu_{i\cdot} - \mu_{\cdot\cdot}) + (\mu_{\cdot j} - \mu_{\cdot\cdot})$ vs H_a : At least one $\mu_{ij} \neq \mu_{\cdot\cdot} + (\mu_{i\cdot} - \mu_{\cdot\cdot}) + (\mu_{\cdot j} - \mu_{\cdot\cdot})$
   The null hypothesis is also equivalent as H_0 : all $\gamma_{jk} = 0$.

Testing hypotheses in two-way ANOVA

We follow the notations of JF for incremental sums of squares in ANOVA:

$$\begin{aligned}
\mathbf{SS}(\gamma|\alpha,\beta) &= \mathbf{SS}(\alpha,\beta,\gamma) - \mathbf{SS}(\alpha,\beta) \\
\mathbf{SS}(\alpha|\beta,\gamma) &= \mathbf{SS}(\alpha,\beta,\gamma) - \mathbf{SS}(\beta,\gamma) \\
\mathbf{SS}(\beta|\alpha,\gamma) &= \mathbf{SS}(\alpha,\beta,\gamma) - \mathbf{SS}(\alpha,\gamma) \\
\mathbf{SS}(\alpha|\beta) &= \mathbf{SS}(\alpha,\beta) - \mathbf{SS}(\beta) \\
\mathbf{SS}(\beta|\alpha) &= \mathbf{SS}(\alpha,\beta) - \mathbf{SS}(\alpha)
\end{aligned}$$

where $\mathbf{SS}(\alpha,\beta,\gamma)$ denotes the regression sum of squares for the full model which includes both sets of main effects and the interaction. $\mathbf{SS}(\alpha,\beta)$ denotes the regression sum of squares for the no-interaction model and $\mathbf{SS}(\alpha,\gamma)$ denotes the regression for the model that omits the column main-effect regressors. Note that the last model violates the principle of marginality because it includes the interaction regressors but omits the column main effects. However, it is useful for constructing the incremental sum of squares for testing the column main effects.

Additional readings: Notes on 3 types of Sum of Squares by Dr. Nancy Reid.

We now have the two-way ANOVA table

Table 1: Two-way ANOVA table

Source	Sum of Squares	df	H_0
R	$\mathbf{SS}(\alpha\|\beta,\gamma)$	$r-1$	all $\alpha_j = 0$
	$\mathbf{SS}(\alpha\|\beta)$	$r-1$	all $\alpha_j = 0 \mid$ all $\gamma_{jk} = 0$
C	$\mathbf{SS}(\beta\|\alpha,\gamma)$	$c-1$	all $\beta_k = 0$
	$\mathbf{SS}(\beta\|\alpha)$	$c-1$	all $\beta_k = 0 \mid$ all $\gamma_{jk} = 0$
RC	$\mathbf{SS}(\gamma\|\alpha,\beta)$	(r -1)(c - 1)	all $\beta_k = 0$
Residuals	$\mathbf{TSS} - \mathbf{SS}(\alpha,\beta,\gamma)$	n - rc	
Total	$\mathbf{TSS}$	n -1	

where the residual sum of squares

$$RSS = \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{jk})^2$$

When test for the hypothesis, use the corresponding SS and df together with the residual SS and df to construct the F-statistic.

$$F = \frac{SS/df}{RSS/df_{residual}}$$

There are two reasonable procedures for testing main-effect hypotheses in two-way ANOVA:

1. Tests based on $\mathbf{SS}(\alpha|\beta,\gamma)$ and $\mathbf{SS}(\beta|\alpha,\gamma)$ ("type III" tests) employ models that violate the principle of marginality, but the tests are valid whether or not interactions are present.

2. Tests based on $\mathbf{SS}(\alpha|\beta)$ and $\mathbf{SS}(\beta|\alpha)$ ("type II" tests) conform to the principle of marginality but are valid only if interactions are absent, in which case they are maximally powerful.

Some more jargon:

- Experimental unit (EU): entity to which experimental treatment is assigned.
  For example, Assign fertilizer treatment to fields. Fields = EU.

- Measurement unit (MU): entity that is measured.
  For example, Measure yields at several subplots within each field. MU: subplot

- Treatment structure: describes how different experimental factors are combined to generate treatments.
  For example, Fertilizers: A, B, C; Irrigation: High, Low.

- Randomization structure: how treatments are assigned to EUs.

- Simplest treatment structure: single experimental factor with multiple levels. Ex. Fertilizers A vs B vs C.

- Simplest randomization structure: Completely randomized design – Experimental treatments assigned to EUs entirely at random.

Example: Honeybee data

Entomologist records energy expended (y) by $N = 27$ honeybees at $a = 3$ temperature (A) levels $(20, 30, 40°C)$ consuming liquids with $b = 3$ levels of sucrose concentration (B) $(20\%, 40\%, 60\%)$ in a balanced, completely randomized crossed 3×3 design.

Temp	Suc	Sample		
20	20	3.1	3.7	4.7
20	40	5.5	6.7	7.3
20	60	7.9	9.2	9.3
30	20	6	6.9	7.5
30	40	11.5	12.9	13.4
30	60	17.5	15.8	14.7
40	20	7.7	8.3	9.5
40	40	15.7	14.3	15.9
40	60	19.1	18.0	19.9

1. What is the experimental unit?

2. What is the treatment structure?

3. Finish the table below

Source	df
A	
B	
$A \times B$	
Residual	
Total	

Answer:

4. Consider the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

where $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, b$ and $k = 1, 2, \ldots, n$ for a balanced design.
Deviation:

- total: $y_{ijk} - \bar{y}_{+++}$
- due to level i of factor A: $\bar{y}_{i++} - \bar{y}_{+++}$

- due to level j of factor B: $\bar{y}_{+j+} - \bar{y}_{+++}$
- due to levels i of factor A and j of factor B after subtracting main effects:

$$\bar{y}_{ij+} - \bar{y}_{+++} - (\bar{y}_{i++} - \bar{y}_{+++}) - (\bar{y}_{+j+} - \bar{y}_{+++}) = \bar{y}_{ij+} - \bar{y}_{i++} - \bar{y}_{+j+} + \bar{y}_{+++}$$

Use the following equations to calculate the Sum of Squares and fill out the ANOVA table.

$$\begin{aligned}
SS[Tot] &= \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{+++})^2 \\
SS[A] &= \sum_i \sum_j \sum_k (\bar{y}_{i++} - \bar{y}_{+++})^2 \\
SS[B] &= \sum_i \sum_j \sum_k (\bar{y}_{+j+} - \bar{y}_{+++})^2 \\
SS[AB] &= \sum_i \sum_j \sum_k (\bar{y}_{ij+} - \bar{y}_{i++} - \bar{y}_{+j+} + \bar{y}_{+++})^2 \\
SS[E] &= \sum_i \sum_j \sum_k (\bar{y}_{ijk} - \bar{y}_{ij+})^2
\end{aligned}$$

where

$$\begin{aligned}
\bar{y}_{ij+} &= \frac{1}{n} \sum_k y_{ijk} \\
\bar{y}_{i++} &= \frac{1}{b} \sum_j \bar{y}_{ij+} = \frac{1}{bn} \sum_j \sum_k y_{ijk} \\
\bar{y}_{+j+} &= \frac{1}{a} \sum_i \bar{y}_{ij+} = \frac{1}{an} \sum_i \sum_k y_{ijk} \\
\bar{y}_{+++} &= \frac{1}{a} \sum_i \bar{y}_{i++} = \frac{1}{b} \sum_j \bar{y}_{+j+} \\
&= \frac{1}{abn} \sum_i \sum_j \sum_k y_{ijk}
\end{aligned}$$

Source	df	Sum of Squares	Mean Square	F
A				
B				
$A \times B$				
Residual				
Total				

Answer: