28 Lecture 28: April 13

Last time

- One-way ANOVA

Today

- Two-way ANOVA
- ANCOVA
- Linear contrasts of means

Additional reference

Course notes by Dr. Jason Osborne.

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 $\alpha_j = 0$ | all $\gamma_{jk} = 0$
C	$\mathbf{SS}(\beta\|\alpha,\gamma)$	$c-1$	all $\beta_k = 0$
	$\mathbf{SS}(\beta\|\alpha)$	$c-1$	all $\beta_k = 0$ | all $\gamma_{jk} = 0$
RC	$\mathbf{SS}(\gamma\|\alpha,\beta)$	(r -1)(c - 1)	all $\gamma_{jk} = 0$
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:

A three-factor example

In a balanced, complete, crossed design, $N = 36$ shrimp were randomized to $abc = 12$ treatment combinations from the factors below:

- A1: Temperature at $25°C$
- A2: Temperature at $35°C$
- B1: Density of shrimp population at 80 shrimp/40l
- B2: Density of shrimp population at 160 shrimp/40l
- C1: Salinity at 10 units
- C2: Salinity at 25 units
- C3: Salinity at 40 units

The response variable of interest is weight gain Y_{ijkl} after four weeks.

Three-way ANOVA model

$$\begin{aligned} Y_{ijkl} &= \mu + \alpha_i + \beta_j + \gamma_k \\ &\quad + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} \\ &\quad + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl} \\ i &= 1, 2 \\ j &= 1, 2 \\ k &= 1, 2, 3 \\ l &= 1, 2, 3 \\ \epsilon_{ijkl} &\overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \end{aligned}$$

Many constraints such as (over one dimension):

$$\begin{aligned} \sum_i \alpha_i &= 0 \\ \sum_i (\alpha\beta)_{ij} &= \sum_j (\alpha\beta)_{ij} = 0 \quad \text{for all } i, j \\ \sum_i (\alpha\beta\gamma)_{ijk} &= \sum_j (\alpha\beta\gamma)_{ijk} = \sum_k (\alpha\beta\gamma)_{ijk} = 0 \quad \text{for all } i, j, k \end{aligned}$$

Now, please finish the table below

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

Answer:

The three-way ANOVA model includes parameters for

- Main effects: α_i, β_j and γ_k.

- Two-way interactions between each pair of factors: $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$ and $(\beta\gamma)_{jk}$.
- Three-way interaction among all three factors: $(\alpha\beta\gamma)_{ijk}$.

Readings:

1. JF 8.3.1 on parameter estimates and hypothesis testing for three-way ANOVA model.
2. JF 8.3.2 on Higher-order classifications.

Analysis of Covariance

Analysis of covariance (ANCOVA) is a term used to describe linear models that contain both qualitative and quantitative explanatory variables. The method is, therefore, equivalent to dummy-variable regression, discussed in the previous lectures, although the ANCOVA model is parametrized differently from the dummy-regression model.

Covariate is a variable known to affect the response that

1. differs among EUs
2. reflects differences that exist independently of experimental treatment.

A nutrition example

A nutrition scientist conducted an experiment to evaluate the effects of four vitamin supplements on the weight gain of laboratory animals. The experiment was conducted in a completely randomized design with $N = 20$ animals randomized to $a = 4$ supplement groups, each with sample size $n \equiv 5$. The response variable of interest is weight gain, but calorie intake z was measured simultaneously.

Diet	$y(g)$	Diet	y	Diet	y	Diet	y
1	48	2	65	3	79	4	59
1	67	2	49	3	52	4	50
1	78	2	37	3	63	4	59
1	69	2	75	3	65	4	42
1	53	2	63	3	67	4	34
1	$\bar{y}_{1+} = 63$	2	$\bar{y}_{2+} = 57.8$	3	$\bar{y}_{3+} = 65.2$	4	$\bar{y}_{4+} = 48.8$
1	$s_1 = 12.3$	2	$s_2 = 14.9$	3	$s_3 = 9.7$	4	$s_4 = 10.9$

Question: Is there evidence of a vitamin supplement effect?

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Diet	3	797.8	265.9	1.823	0.184
Residuals	16	2334.4	145.9		

But calorie intake z was measured simultaneously:

Diet	$y(g)$	z	Diet	y	z	Diet	y	z	Diet	y	z
1	48	350	2	65	400	3	79	510	4	59	530
1	67	440	2	49	450	3	52	410	4	50	520
1	78	440	2	37	370	3	63	470	4	59	520
1	69	510	2	75	530	3	65	470	4	42	510
1	53	470	2	63	420	3	67	480	4	34	430

Question: How and why could these new data be incorporated into analysis?
Answer: ANCOVA can be used to reduce unexplained variation.

ANCOVA model,

$$y_{ij} = \mu + \alpha_i + \beta z_{ij} + \epsilon_{ij}$$

where μ is the reference level, α_i is the main effect of treatment, β is the partial regression coefficient, and $\epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. The model is equivalent as the dummy-variable regression model,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_z z_i + \epsilon_i \quad \text{for } i = 1, \dots, 20$$

Finish the table below

Source	df
Diet	
Covariate	1
Residual	
Total	

Answer:

To test for difference among treatments. The null hypothesis in terms of α_i is
$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_4 = 0$ v.s. $H_a :$ at least one $\alpha_i \neq 0$

And the null hypothesis in terms of β_i is
$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ v.s. $H_a :$ at least one $\beta_i \neq 0$

Question: which two models do we compare when testing the above null hypothesis?
Answer:

Linear contrasts of means

With ANOVA (or ANCOVA) models, we do not generally test hypotheses about individual coefficients (but we can do so if we wish). For dummy-coded regressors in one-way ANOVA, a t-test or F-test of $H_0 : \alpha_1 = 0$, for example, is equivalent to testing for the difference in means between the first group and the baseline group, $H_0 : \mu_1 = \mu_m$.

Consider the one-way ANOVA model:

$$Y_{ij} = \mu_i + \epsilon_{ij}, i = 1, 2, \ldots, t, \text{ and } j = 1, 2, \ldots, n_i$$

with $\epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

A linear function of the group means of the form

$$\theta = c_1\mu_1 + c_2\mu_2 + \cdots + c_t\mu_t$$

is called a linear combination of the treatment means. And the c_i's are the coefficients of the linear combination. If

$$c_1 + c_2 + \cdots + c_t = \sum_{j=1}^{t} c_j = 0,$$

the linear combination is called a contrast. Contrasts with more than two non-zero coefficients are called complex contrasts.

Let two contrasts θ_1 and θ_2 be given by

$$\theta_1 = c_1\mu_1 + \cdots + c_t\mu_t = \sum_{j=1}^{t} c_j\mu_j$$
$$\theta_2 = d_1\mu_1 + \cdots + d_t\mu_t = \sum_{j=1}^{t} d_j\mu_j,$$

then the two contrasts θ_1 and θ_2 are mutually orthogonal if the products of their coefficients sum to zero:

$$c_1d_1 + \cdots + c_td_t = \sum_{j=1}^{t} c_jd_j = 0$$

θ_i and θ_j are orthogonal $\implies$ $\hat{\theta}_i$ and $\hat{\theta}_j$ are statistically independent.

Types of effects

Consider the following two-way ANOVA model:

$$\begin{aligned} Y_{ijk} &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \\ & i = 1, 2 = a \text{ and } j = 1, 2 = b \text{ and } k = 1, 2, \ldots, 7 = n. \end{aligned}$$

$\epsilon_{ijk} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Parameter constraints: $\sum_i \alpha_i = \sum_j \beta_j = 0$ and $\sum_i (\alpha\beta)_{ij} = 0$ for each j and $\sum_j (\alpha\beta)_{ij} = 0$ for each i.

- Factor A: AGE has $a = 2$ levels - A_1 : younger and A_2 : older
- Factor B: GENDER has $b = 2$ levels - B_1 : female and B_2 : male

Three kinds of effects in this 2×2 design:

1. Simple effects are simple contrasts.
   - $\mu(A_1 B) = \mu_{12} - \mu_{11}$ - simple effect of gender for young folks.
   - $\mu(AB_1) = \mu_{21} - \mu_{11}$ - simple effect of age for women.
2. Interaction effects are differences of simple effects: $\mu(AB) = \mu(AB_2) - \mu(AB_1) = (\mu_{22} - \mu_{12}) - (\mu_{21} - \mu_{11})$
   - difference between simple age effects for men and women
   - difference between simple gender effects for old and young folks
   - interaction effect of AGE and GENDER.
3. Main effects are averages or sums of simple effects

$$\begin{aligned} \mu(A) &= \frac{1}{2}(\mu(AB_1) + \mu(AB_2)) \\ \mu(B) &= \frac{1}{2}(\mu(A_1 B) + \mu(A_2 B)) \end{aligned}$$