STAT 5000

Statistical Methods I

Week 5

Fall 2024

Dr. Danica Ommen

Unit 2

Introduction to ANOVA

Scenario

- Observational Studies
  - More than 2 Populations
- Experiments
  - One factor with more than 2 levels
- Compare observations of variable (quantitative) for multiple treatment groups or populations.

Notation: Population or Treatment Group Parameters

- Number of groups: $i = 1, 2, \ldots, r$
- Group means $= \mu_1, \mu_2, \ldots, \mu_r$
- Group Variances $= \sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2$
- Group Std. Dev. $= \sigma_1, \sigma_2, \ldots, \sigma_r$

Notation: Data and Summary Statistics

- n_i = sample size for ith sample or treatment group
- $Y_{ij} = j$th observation in the ith sample or treatment group, where $j = 1, 2, \ldots, n_i$
- Mean for ith sample or treatment group

$$\bar{Y}_i = \bar{Y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

- Variance for ith sample or treatment group

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2$$

Notation: More Summary Statistics

- Total number of observations: $N = \sum_{i=1}^{r} n_i$
- Overall mean: $\bar{Y} = \bar{Y}_{\cdot\cdot} = \frac{1}{N} \sum_{i=1}^{r} \sum_{j=1}^{n_i} Y_{ij}$
- Pooled variance estimate:

$$S_p^2 = \frac{\sum_{i=1}^{r} (n_i - 1) S_i^2}{N - r} \quad \text{with} \quad \text{df} = \sum_{i=1}^{r} (n_i - 1) = N - r$$

Inference Strategies about Means for Several Populations:

- Basic linear model
- Analysis of Variance (ANOVA)
  - F-tests
  - Contrasts
- Model diagnostics
- Nonparametric tests

Research Questions

- Do the populations or treatment groups have the same mean values for the variable?
- Two sources of variation
  - Variability among observations within each treatment group (or within each population)
  - Variability among mean responses for treatments (or between populations)
- Question:
  - Are differences among group means large relative to variation within groups?
  - Do all populations have the same mean?

Linear Model Set-up

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

- Each observation Y_{ij} can be described by two components:
  - Fixed mean value μ_i
  - Random error term ϵ_{ij}
- Gives an equation for each of the $N = \sum_{i=1}^{r} n_i$ observations

Matrix Notation

We can write this system of N equations in matrix notation

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ Y_{31} \\ \vdots \\ Y_{rn_r} \end{pmatrix} = \begin{pmatrix} \mu_1 + \epsilon_{11} \\ \mu_1 + \epsilon_{12} \\ \vdots \\ \mu_1 + \epsilon_{1n_1} \\ \mu_2 + \epsilon_{21} \\ \vdots \\ \mu_2 + \epsilon_{2n_2} \\ \mu_3 + \epsilon_{31} \\ \vdots \\ \mu_r + \epsilon_{rn_r} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \vdots \\ \mu_r \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \\ \epsilon_{31} \\ \vdots \\ \epsilon_{rn_r} \end{pmatrix}$$

Matrix Notation

Let

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ Y_{31} \\ \vdots \\ Y_{rn_r} \end{bmatrix}, X = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \beta = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \vdots \\ \mu_r \end{bmatrix}, \text{ and } \epsilon = \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \\ \epsilon_{31} \\ \vdots \\ \epsilon_{rn_r} \end{bmatrix}$$

Linear Model: Matrix Form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- The vector $\mathbf{Y}$ is length N and is the vector of observations.
- The matrix $\mathbf{X}$ is size N x r and is called the design matrix. It relates the observations to the parameters according to the model. It is fixed (non-random).
- The vector $\boldsymbol{\beta}$ is length r and is the vector of parameter values.
- The vector $\boldsymbol{\epsilon}$ is length N and is the vector of random error terms.

Expected Value

- Assuming $E(\epsilon) = \mathbf{0}$, we have

$$\begin{aligned} E(\mathbf{Y}) &= E(\mathbf{X}\beta + \epsilon) \\ &= \mathbf{X}\beta + E(\epsilon) \\ &= \mathbf{X}\beta + \mathbf{0} \\ &= \mathbf{X}\beta \end{aligned}$$

Expected Value

$$E(\mathbf{Y}) = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \vdots \\ \mu_r \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_1 \\ \vdots \\ \mu_1 \\ \mu_2 \\ \vdots \\ \mu_2 \\ \vdots \\ \mu_r \end{pmatrix}$$

Using our data, we will estimate the parameters in the β vector using the method of least squares.

Least Squares Estimation

Find the estimates of the population parameters that minimize the sum of squared deviations between the observed outcomes and the estimates of the expected outcomes.

For cell means model, find $\hat{\mu}_1, \hat{\mu}_2, \cdots \hat{\mu}_r$ that minimize

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \hat{\mu}_i)^2$$

or, equivalently,

$$(\mathbf{Y} - \mathbf{X}\hat{\beta})^T(\mathbf{Y} - \mathbf{X}\hat{\beta})$$

Least Squares Estimation

- If the design matrix X is of full column rank, then
  - value of the parameter vector β that minimizes the squared errors is
$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$
  - value $\hat{\beta}$ is unique since $(\mathbf{X}^T\mathbf{X})^{-1}$ is unique.
- Unique least squares estimator for the parameter vector β is:

$$\hat{\beta} = \begin{bmatrix} n_1 & 0 & 0 & \cdots & 0 \\ 0 & n_2 & 0 & \cdots & 0 \\ 0 & 0 & n_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \cdots & n_r \end{bmatrix}^{-1} \begin{pmatrix} \sum_{j=1}^{n_1} Y_{1j} \\ \sum_{j=1}^{n_2} Y_{2j} \\ \vdots \\ \sum_{j=1}^{n_r} Y_{rj} \end{pmatrix} = \begin{pmatrix} \bar{Y}_{1\cdot} \\ \bar{Y}_{2\cdot} \\ \vdots \\ \bar{Y}_{r\cdot} \end{pmatrix}$$

Predicted Values Using the least squares estimator $\hat{\beta}$, the predicted value for $\mathbf{Y}$ is:

$$\begin{aligned} \hat{\mathbf{Y}} &= X\hat{\beta} \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= P_X\mathbf{Y} \end{aligned}$$

where $P_X = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the orthogonal projection operator onto the column space of matrix $\mathbf{X}$.

Unit 2

ANOVA Tables

For the Cell Means Model,

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

the variation in Y_{ij} comes from two sources:

- **Population Means**: $\mu_1, \mu_2, \ldots, \mu_r$
- **Random errors**: $\epsilon_{ij} = Y_{ij} - \mu_i$

- Random errors and population means are unknown.
- Study sample means and residuals.
  - Sample Means: $\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot}, \ldots, \bar{Y}_{r\cdot}$
  - Residuals: $e_{ij} = Y_{ij} - \bar{Y}_{i\cdot}$ for all i, j

Research Question: Do the populations or treatment groups have the same mean values for the variable?

- Yes:
  - Sample Means will vary, but should be similar in value.
  - Most variation in Y_{ij} will be in residual e_{ij}.
- No:
  - Sample Means will vary, differences will reflect differences in population means.
  - Will still have variation in Y_{ij} from residual e_{ij}.

Analysis of Variance (ANOVA):

- Calculate three variations based on observations Y_{ij}
  - Variation due to group means
  - Variation due to residuals
  - Total Variation
- These are called the *sums of squares* (SS)

Variation due to Group Means

$$SS_{\text{among groups}} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{r} n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$$

- Also called SS_{model}
- If the population means are the same (different), this value should be small (large).

Variation due to Residuals

$$
\begin{aligned}
SS_{\text{within groups}} &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2 \\
&= \sum_{i=1}^{r}(n_i - 1)S_i^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{n_i} e_{ij}^2
\end{aligned}
$$

- also called SS_{error} or $SS_{\text{residuals}}$

Total Variation

$$SS_{\text{total}} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2 = SS_{\text{model}} + SS_{\text{error}}$$

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot} + \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2 + \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 + 2\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})$$

$$= \sum_{i=1}^{r} n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 + \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2$$

ANOVA Table

source of variation	degrees of freedom	sums of squares	mean square	F
Model	$r-1$	SS_{model}	$MS_{\text{model}} = \frac{SS_{\text{model}}}{(r-1)}$	$\frac{MS_{\text{model}}}{MS_{\text{error}}}$
Error	$N-r$	SS_{error}	$MS_{\text{error}} = \frac{SS_{\text{error}}}{(N-r)}$	
Total	$N-1$	SS_{total}		

Note: $MS_{\text{error}} = S_p^2$

Model Assumptions

- Assumptions on random error terms
  - ϵ_{ij} are i.i.d. from a normal distribution with mean 0 and variance σ^2.
  - $\boldsymbol{\epsilon}$ is multivariate normal with mean $\mathbf{0}$ and variance $\sigma^2\mathbf{I}$.
- This implies that
  - Y_{ij} are i.i.d. from a normal distribution with mean μ_i and variance σ^2.
  - $\mathbf{Y}$ is multivariate normal with mean $\mathbf{X}\boldsymbol{\beta}$ and variance $\sigma^2\mathbf{I}$.
- In addition, we assume groups are independent of each other.

Results: Mean Squares

- $E(MS_{\text{error}}) = E(S_p^2) = \sigma^2$
- $E(MS_{\text{model}}) = \sigma^2 + \frac{1}{r-1}\sum_{i=1}^{r} n_i(\mu_i - \bar{\mu})^2$ where $\bar{\mu} = \frac{1}{N}\sum_i n_i\mu_i$
- MS_{error} and MS_{model} are independent so

$$\frac{E\left(MS_{\text{model}}\right)}{E\left(MS_{\text{error}}\right)} = \frac{\sigma^2 + \frac{1}{r-1}\sum_{i=1}^{r} n_i(\mu_i - \bar{\mu})^2}{\sigma^2}$$

Hypothesis Test

- $H_o : \mu_1 = \mu_2 = \cdots = \mu_r$
- H_a : at least one μ_i is different for $i = 1, \ldots, r$
- Test Statistic:

$$F = \frac{MS_{\text{model}}}{MS_{\text{error}}}$$

- Large values of F provide evidence against the null hypothesis.

Recall:

- Let W_1 has a χ^2 distribution with ν_1 degrees of freedom.
- Let W_2 has a χ^2 distribution with ν_2 degrees of freedom.
- Assume W_1 and W_2 are independent.

$$F = \frac{W_1/\nu_1}{W_2/\nu_2}$$

has a central F distribution with ν_1 numerator and ν_2 denominator degrees of freedom.

Distribution of Test Statistic

Under model assumptions and $H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$ true:

- $(N - r)MS_{\text{error}}/\sigma^2 \sim \chi^2_{N-r}$
- $(r - 1)MS_{\text{model}}/\sigma^2 \sim \chi^2_{r-1}$
- MS_{error} and MS_{model} are independent

$$F = \frac{((r-1)MS_{\text{model}}/\sigma^2)/(r-1)}{((N-r)MS_{\text{error}}/\sigma^2)/(N-r)} = \frac{MS_{\text{model}}}{MS_{\text{error}}}$$

has a central F-distribution with $r - 1$ numerator and $N - r$ denominator degrees of freedom.

ANOVA F-test

- $H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$
- H_a : at least one μ_i is different for $i = 1, \ldots, r$
- Test Statistic:

$$F = \frac{MS_{\text{model}}}{MS_{\text{error}}}$$

- P-value:

$$P(F_{r-1,N-r} > F)$$

Donut Example

- An experiment was conducted to determine the best type of oil for frying donuts.
  - For anything fried, the goal is to use the oil to heat the food, but not have the food absorb the oil.
- Treatments: Four types of oil for cooking donuts
- Experimental units: batches of donuts
- Randomization: Assignment of batches to cooking oils
- Measured response: grams of oil absorbed when one batch is cooked minus 150 grams (to account for spilling?)

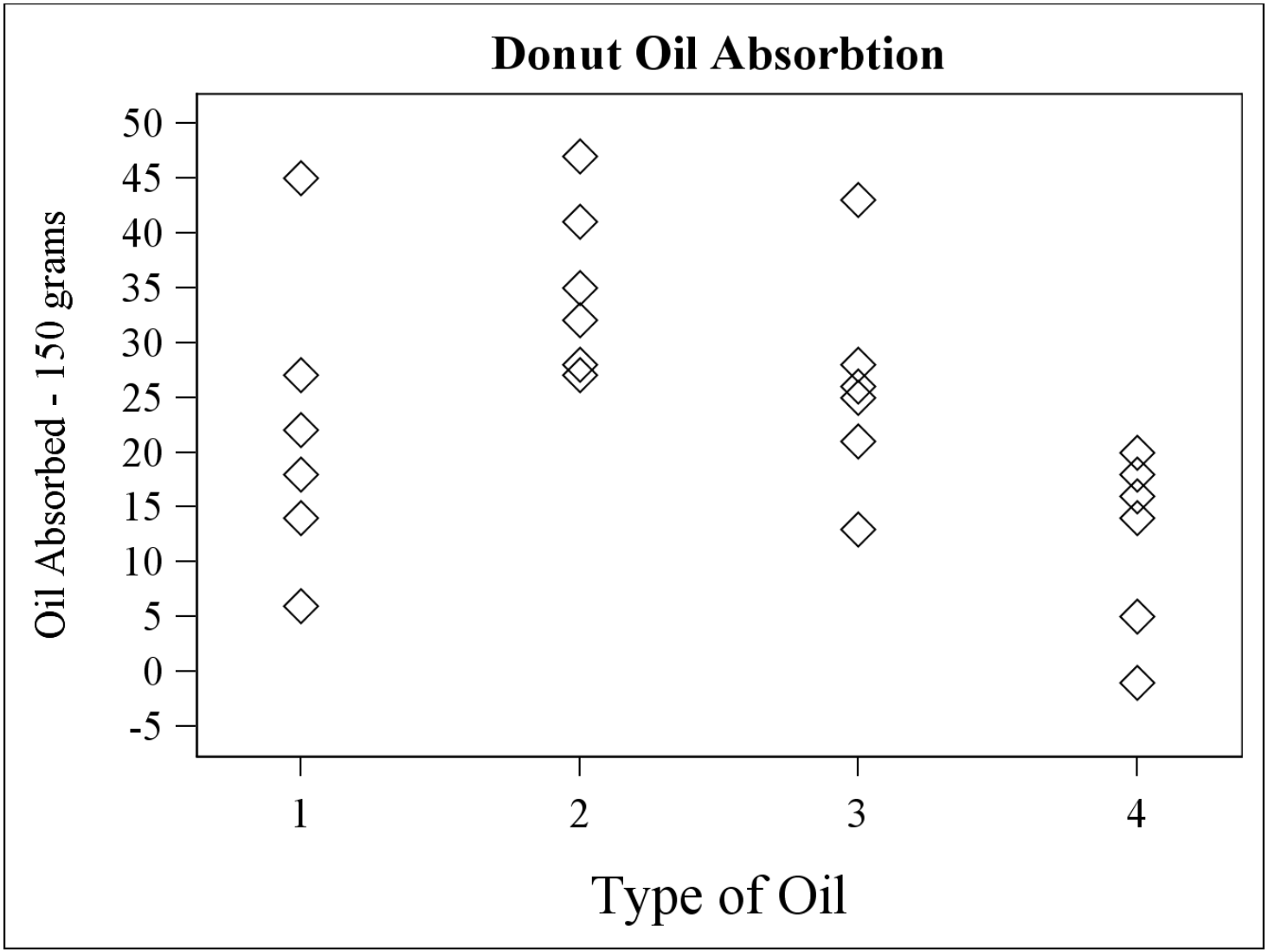
Donut Example: Data

The amount of oil absorbed (-150gram) in cooking 24 batches of donuts is given below.

Grams of oil absorbed (-150g) for different types of cooking oil

Type 1	Type 2	Type 3	Type 4
14	28	25	5
22	41	43	16
18	47	28	-1
27	32	21	14
6	35	13	20
45	27	26	18

Donut Example



Donut Example: Summary Statistics

Oil	n_i	$\bar{Y}_{i\cdot}$	S_i^2
Type 1	6	22	178.0
Type 2	6	35	60.4
Type 3	6	26	97.6
Type 4	6	12	67.6

Donut Example: Compute Total Sums of Squares

The table below provides $Y_{ij} - \bar{Y}_{..}$ where $\bar{Y}_{..} = 23.75$.

Type 1	Type 2	Type 3	Type 4
-9.75	4.25	1.25	-18.75
-1.75	17.25	19.25	-7.75
-5.75	23.25	4.25	-24.75
3.25	8.25	-2.75	-9.75
-17.75	11.25	-10.75	-3.75
21.25	3.25	2.25	-5.75

To find SS_{Total}, we square all values in the table and sum them to get 3654.5.

> *There are 23 degrees of freedom for this SS since we are calculating it by subtracting 24 observation values from the overall mean value.*

Donut Example: Compute Model Sums of Squares

The table below provides $\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}$

Type 1	Type 2	Type 3	Type 4
-1.75	11.25	2.25	-11.75
-1.75	11.25	2.25	-11.75
-1.75	11.25	2.25	-11.75
-1.75	11.25	2.25	-11.75
-1.75	11.25	2.25	-11.75
-1.75	11.25	2.25	-11.75

To find SS_{Model}, we square all values in the table and sum them to get 1636.5.

> *There are 3 degrees of freedom for this SS since we are calculating it by subtracting the 4 group means from the overall mean.*

Donut Example: Compute Error Sums of Squares

The table below provides $Y_{ij} - \bar{Y}_{i\cdot}$

Type 1	Type 2	Type 3	Type 4
-8	-7	-1	-7
0	6	17	4
-4	12	2	-13
5	-3	-5	2
-16	0	-13	8
23	-8	0	6

To find SS_{Error}, we square all values in the table and sum them to get 2018.

> *There are* $4 * (6 - 1) = 20$ *degrees of freedom for this SS since since it is calculated by subtracting the 6 observations in each of the 4 groups from the corresponding group mean.*

Donut Example: ANOVA Table

source of variation	degrees of freedom	sums of squares	mean square	F
Model	3	1636.5	545.5	5.41
Error	20	2018.0	100.9	
Total	23	3654.5		

$$F = 5.41 > F_{(3,20),\ .99} = 4.94$$

From computer output: p-value=0.0069

> *The average amount of absorbed oil is not the same for all four types of oil.*

Unit 2

ANOVA: Effects Model

Recall: Cell Means Model

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

- Each observation Y_{ij} can be described by two components:
  - Fixed mean value: μ_i
  - Random error term: ϵ_{ij}
- Test
  - $H_o : \mu_1 = \mu_2 = \cdots = \mu_r$
  - H_a : at least one μ_i different
  - Caution: No way to tell which one(s) is different if H_o rejected

Linear Effects Model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

- Each observation Y_{ij} can be described by two components:
  - Fixed mean value: $\mu_i = \mu + \alpha_i$
    - Overall mean value: μ
    - Treatment effects compared with overall mean: α_i
    - Goal: find which α's are different from 0
  - Random error term: ϵ_{ij}

Linear Effects Model: Matrix Notation

We can write this system of N equations in matrix notation:

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ Y_{31} \\ \vdots \\ Y_{rn_r} \end{pmatrix} = \begin{pmatrix} \mu + \alpha_1 + \epsilon_{11} \\ \mu + \alpha_1 + \epsilon_{12} \\ \vdots \\ \mu + \alpha_1 + \epsilon_{1n_1} \\ \mu + \alpha_2 + \epsilon_{21} \\ \vdots \\ \mu + \alpha_2 + \epsilon_{2n_2} \\ \mu + \alpha_3 + \epsilon_{31} \\ \vdots \\ \mu + \alpha_r + \epsilon_{rn_r} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ 1 & 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ 1 & 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ 1 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_r \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \\ \epsilon_{31} \\ \vdots \\ \epsilon_{rn_r} \end{pmatrix}$$

Linear Effects Model: Problems

- Model has too many parameters:
  estimates r means with $r + 1$ parameters
- Design matrix $\mathbf{X}$ is not full column rank.
- Usual inverse for $(\mathbf{X}^T\mathbf{X})$ does not exist.
- There are an infinite number of least squares estimators.
- Solution: constrain the parameters in the model
  - Set $\alpha_r = 0$ (baseline)
  - Set $\sum_{i=1}^{r} \alpha_i = 0$ (sum to zero)

Linear Effects Model: Revised #1

Using the constraint $\alpha_r = 0$, we have

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{r-1,1} \\ \vdots \\ Y_{rn_r} \end{pmatrix} = \begin{pmatrix} \mu + \alpha_1 + \epsilon_{11} \\ \mu + \alpha_1 + \epsilon_{12} \\ \vdots \\ \mu + \alpha_1 + \epsilon_{1n_1} \\ \mu + \alpha_2 + \epsilon_{21} \\ \vdots \\ \mu + \alpha_2 + \epsilon_{2n_2} \\ \vdots \\ \mu + \alpha_{r-1} + \epsilon_{r-1,1} \\ \vdots \\ \mu + \epsilon_{rn_r} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_{r-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \\ \vdots \\ \epsilon_{r-1,1} \\ \vdots \\ \epsilon_{rn_r} \end{pmatrix}$$

Linear Effects Model: Population Parameters #1

The population or treatment means in model with $\alpha_r = 0$ are:

$$\begin{aligned}
\mu_1 &= \mu + \alpha_1 \\
\mu_2 &= \mu + \alpha_2 \\
\vdots &= \vdots \\
\mu_{r-1} &= \mu + \alpha_{r-1} \\
\mu_r &= \mu
\end{aligned}$$

Least Squares Estimator of β

When $\alpha_r = 0$:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \bar{Y}_{r\cdot} \\ \bar{Y}_{1\cdot} - \bar{Y}_{r\cdot} \\ \bar{Y}_{2\cdot} - \bar{Y}_{r\cdot} \\ \vdots \\ \bar{Y}_{(r-1)\cdot} - \bar{Y}_{r\cdot} = \end{pmatrix} = \begin{pmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \\ \vdots \\ \hat{\alpha}_{r-1} \end{pmatrix}$$

Using the constraint $\sum_{i=1}^{r} \alpha_i = 0$, we have $\alpha_r = -\sum_{i=1}^{r-1} \alpha_i$.

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{r-1,1} \\ \vdots \\ Y_{rn_r} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & -1 & \cdots & -1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_{r-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \\ \vdots \\ \epsilon_{r-1,1} \\ \vdots \\ \epsilon_{rn_r} \end{pmatrix}$$

Linear Effects Model: Population Parameters #2

The population or treatment means in model with $\sum_{i=1}^{r} \alpha_i = 0$ are:

$$
\begin{aligned}
\mu_1 &= \mu + \alpha_1 \\
\mu_2 &= \mu + \alpha_2 \\
\vdots &= \vdots \\
\mu_{r-1} &= \mu + \alpha_{r-1} \\
\mu_r &= \mu + \alpha_r = \mu - \sum_{i=1}^{r-1} \alpha_i
\end{aligned}
$$

Least Squares Estimator of β

When $\sum_{i=1}^{r} \alpha_i = 0$:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \frac{1}{r}\sum_{i=1}^{r} \bar{Y}_{i\cdot} \\ \bar{Y}_{1\cdot} - \frac{1}{r}\sum_{i=1}^{r} \bar{Y}_{i\cdot} \\ \bar{Y}_{2\cdot} - \frac{1}{r}\sum_{i=1}^{r} \bar{Y}_{i\cdot} \\ \vdots \\ \bar{Y}_{(r-1)\cdot} - \frac{1}{r}\sum_{i=1}^{r} \bar{Y}_{i\cdot} \end{pmatrix} = \begin{pmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \\ \vdots \\ \hat{\alpha}_{r-1} \end{pmatrix}$$

Cautions:

- The above two types of constraints for linear effects models are not the only ways to model the means.
- The choice of constraint will affect your least squares estimator $\hat{\beta}$.
- You must determine which constraint was applied before interpreting parameter estimates.
- The interpretation of the parameters (elements of β) depends on the parametrization.

What about the ANOVA Table?

If your question is $H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$ or not, then

- The analysis is the the same no matter which set of parameters you use.
- **Estimable function**: a quantity that does not depend on the arbitrary choice of constraint.
- The population means are estimable, i.e., P_X is invariant to the choice of constraints.
- Use of effects model with different constraints does not affect the ANOVA Table.

ANOVA: Fixed or Random Effects?

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

Fixed effects

- The r treatments (or groups) examined in the study are the only ones under consideration
- Research questions are about treatment means or difference in means
  - e.g., two drugs, four pesticides

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

Random effects

- The r treatments (or groups) are a random sample from some larger population of treatments (or groups) that could have been included in the study
- Research questions are about variability in sets of treatments (or groups) the could be selected for different studies
- Additional assumptions that

$$\alpha_i \sim N(0, \sigma_\alpha^2)$$

and any α_i is independent of any ϵ_{ij}

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

- Assumptions:
  - $\alpha_i \sim N(0, \sigma_\alpha^2)$ and $\epsilon_{ij} \sim N(0, \sigma_e^2)$
  - any α_i is independent of any ϵ_{ij}
- Parameter of interest: μ, σ_α^2, and σ_e^2
- Intraclass Correlation:

$$\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_e^2}$$

is the correlation between any pair of observations in the same group

Example:

Examine variability in student performance in AP (high school) statistics classes.

- Random sample of 8 AP (high school) statistics classes
- Random sample of 10 students from each class (students are *nested* in classes)
- Give ISU Stat 101 final exam to each student and record the scores
- Not interested in just the 8 classes selected for the study

ANOVA: Random Effects

- Model: $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$
- μ measures overall performance on a college exam
- σ^2_α measures variability among AP classes
- σ^2_e measures variability among students within classes
- Intraclass correlation measures correlation between any pair of students in same class
- Focus on fixed effects in STAT 5000

Questions?

Contact me:

Email: dmommen@iastate.edu

Student Office Hours: Thursdays @ 10-11 am