STAT 500

Inference about Means for Several Populations

Scenario

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

Notation

- Population or Treatment Group Parameters
  - Number: $\boldsymbol{i = 1, 2, \ldots, r}$
  - Population or Treatment means $= \boldsymbol{\mu_1, \mu_2, \ldots, \mu_r}$
  - Population or Treatment Variances $= \boldsymbol{\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2}$
  - Population or Treatment Std. Dev. $= \boldsymbol{\sigma_1, \sigma_2, \ldots, \sigma_r}$

Notation

- Data and Summary Statistics
  - n_i = sample size for ith sample or treatment group
  - Y_{ij} = jth 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

- Summary Statistics
  - Total number of observations: $N = \sum_{i=1}^{r} n_i$
  - Overall mean: $\bar{Y} = \bar{Y}_{..} = \frac{1}{N} \sum_{i=1}^{r} \sum_{j=1}^{n_i} Y_{ij}$
  - 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 about Means for Several Populations

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

Research Question

- 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?

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}
- Gives an equation for each of the $N = \Sigma_{i=1}^{r} n_i$ observations

Cell Means Model in Matrix Notation

We can write this system of N equations in matrix notation

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

Cell Means Model in Matrix Notation

Let

$$\mathbf{Y} = \begin{bmatrix} \boldsymbol{Y}_{11} \\ \boldsymbol{Y}_{12} \\ \vdots \\ \boldsymbol{Y}_{1n_1} \\ \boldsymbol{Y}_{21} \\ \vdots \\ \boldsymbol{Y}_{2n_2} \\ \boldsymbol{Y}_{31} \\ \vdots \\ \boldsymbol{Y}_{rn_r} \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_3 \\ \vdots \\ \boldsymbol{\mu}_r \end{bmatrix}, \quad \text{and } \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}_{11} \\ \boldsymbol{\epsilon}_{12} \\ \vdots \\ \boldsymbol{\epsilon}_{1n_1} \\ \boldsymbol{\epsilon}_{21} \\ \vdots \\ \boldsymbol{\epsilon}_{2n_2} \\ \boldsymbol{\epsilon}_{31} \\ \vdots \\ \boldsymbol{\epsilon}_{rn_r} \end{bmatrix}$$

Linear Model

Cell Means Model is an example of a **linear model** in matrix form:

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

- The vector $\mathbf{Y}$ is length $\boldsymbol{N}$ and is the vector of observations.
- The matrix $\boldsymbol{X}$ is size $\boldsymbol{N}$ x $\boldsymbol{r}$ 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 $\boldsymbol{r}$ and is the vector of parameter values.
- The vector $\boldsymbol{\epsilon}$ is length $\boldsymbol{N}$ and is the vector of random error terms.

Expected Values: Linear Model

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

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

Cell Means Model

Expected Value for the cell means model:

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

Least Squares Estimation

- Using our data, we will estimate the parameters in the $\boldsymbol{\beta}$ 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,

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

Least Squares Estimation

- If the design matrix $\boldsymbol{X}$ is of full column rank, then
  - value of the parameter vector $\boldsymbol{\beta}$ that minimizes the squared errors is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\mathbf{Y}$$

  - value $\hat{\boldsymbol{\beta}}$ is unique since $(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}$ is unique.

Cell Means Model

- Design matrix $\boldsymbol{X}$ has full column rank.
- Unique least squares estimator for the parameter vector $\boldsymbol{\beta}$ is:

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

- This is the least squares estimator for the population parameters (population means)

Predicted Values: Linear Model

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

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

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