Chapter 8
Inferences About More Than Two Population Central Values

STAT 441/541 Statistical Methods II

Motivating Example: Silt Content of Soils

- Eight contiguous sites
- 11 random points within each site
- All samples taken from same depth
- Soil property of interest is silt content
- Objective: Determine if there is a difference in silt content among the soils from different sites

Activity: Review the handout Silt Content of Soils

Sections Covered

- 8.1 Introduction
- 8.2 A Statistical Test About More Than Two Population Means: An Analysis of Variance (ANOVA)
- 8.3 The Model for Observations in a Completely Randomized Design
- 8.4 Checking the ANOVA Conditions

Notation Refresher

- Population or Treatment Means
  - Denoted by the Greek letter mu, μ
  - Each mean denoted by $\mu_1, \mu_2, \ldots, \mu_t$
- Population or Treatment Variances and Standard Deviations
  - Denoted by the Greek letter sigma, σ
  - Variance is σ^2 and standard deviation is σ
  - Each variance and standard deviation denoted by σ_i^2 and σ_i
- Sample sizes
  - Denoted by n
  - Each denoted by n_i
  - We will work with balanced data: $n_1 = n_2 = \cdots = n_t = n$

Notation for *t* treatments

TABLE 8.5
Summary of sample data for a completely randomized design

Population	Data				Mean
1	y_{11}	y_{12}	y_{13}	y_{14}	$\bar{y}_{1.}$
2	y_{21}	y_{22}	y_{23}	y_{24}	$\bar{y}_{2.}$
3	y_{31}	y_{32}	y_{33}	y_{34}	$\bar{y}_{3.}$
4	y_{41}	y_{42}	y_{43}	y_{44}	$\bar{y}_{4.}$
5	y_{51}	y_{52}	y_{53}	y_{54}	$\bar{y}_{5.}$

Notation Needed for the AOV of a Completely Randomized Design

y_{ij}: The jth sample observation selected from population i. For example, y_{23} denotes the third sample observation drawn from population 2.

n_i: The number of sample observations selected from population i. In our data set, n_1, the number of observations obtained from population 1, is 4. Similarly, $n_2 = n_3 = n_4 = n_5 = 4$. However, it should be noted that the sample sizes need not be the same. Thus, we might have $n_1 = 12$, $n_2 = 3$, $n_3 = 6$, $n_4 = 10$, and so forth.

n_T: The total sample size; $n_T = \sum n_i$. For the data given in Table 8.5, $n_T = n_1 + n_2 + n_3 + n_4 + n_5 = 20$.

$\bar{y}_{i.}$: The average of the n_i sample observations drawn from population i; $\bar{y}_{i.} = \sum_j y_{ij}/n_i$.

$\bar{y}_{..}$: The average of all sample observations; $\bar{y}_{..} = \sum_i \sum_j y_{ij}/n_T$.

Note: The text uses “population” but I am using “treatment”

Statistical Dataset

- Columns are variables
  - Variables denoting treatments are called factors
  - Treatments can be identified by numbers, characters, or a combination of both
- Rows are observations
  - Each observation consists of a value for the response variable and a value that identifies the treatment
- We are using Excel (.xlsx) files for our statistical datasets

- **Activity**: Complete the handout Fun with a Statistical Dataset for One-Way ANOVA

Section 8.1 Introduction

- Many practical and scientific settings want to compare the means of three or more populations, or treatments
- The testing procedure called analysis of variance will detect differences among sample means from each population, or treatment
- Two sources of variation:
  1. Within-sample variation, and
  2. Between-sample variation
- All differences between sample means are judged statistically significant (or not) by comparing them to the variation within samples

Section 8.2 A Statistical Test for more than Two Population Means: An Analysis of Variance

The analysis of variance procedures are developed under the following conditions:

1. Each of the populations, or treatments, has a normal distribution,
2. The variances of the populations, or treatments, are equal (there is a common variance), and
3. The measurements are independent random samples from their respective populations

Estimating Within-sample and Between-sample Variation

- Within-sample variance for each population:

$$s_{W_i}^2 = \frac{\sum_{j=1}^{n_i} \left(y_{ij} - \overline{y}_i\right)^2}{n_i - 1}$$

- The within-sample variances are pooled together to estimate the common variance σ^2
- Between-sample variance of sample means:

$$s_B^2 = \frac{\sum_{i=1}^{t} \left(\overline{y}_i - \overline{\overline{y}}\right)^2}{t - 1}$$

Where t=number of populations and $\overline{\overline{y}}$ is the average of the sample means

Analysis of Variance Table

Used to partition of Total (Corrected) Sum of Squares (TSS) into two components:
(1) Between Treatments (SSB), and
(2) Within Treatments (SSW)

Source	df=degrees of freedom	Sum of Squares	Mean Square	F test	p-value
Between Treatments	$t-1$	SSB	$s_B^2 = \frac{SSB}{t-1}$	$F = \frac{s_B^2}{s_W^2}$	
Within Treatments	$n_T - t$	SSW	$s_W^2 = \frac{SSW}{n_T - 1}$		
Total (Corrected)	$n_T - 1$	TSS			

Where $n_T = \sum_i^t n_i =$ total number of observations
t=number of treatments
Note that df and Sum of Squares are additive and Mean Square=SS/df

Section 8.3 Models for Observations in a Completely Randomized Design (CRD)

Some important definitions

Treatments: conditions selected by the researcher, such as different types of pesticides

Experimental Unit: the physical entity to which the treatment is randomly assigned, such as test plots

Completely Randomized Design: an experimental setting in which treatments are randomly assigned to a large group of homogeneous experimental units

Activity: Complete the handout "Fun with Experimental Unit"

Two Models for a CRD

- Means Model (based on treatment means)

$$y = \mu_i + \varepsilon$$

- Effects Model (based on treatment effects)

$$y = \mu + \tau_i + \varepsilon$$

- These two models are equivalent for completely randomized designs
- We will use the treatment effects model

Hypothesis Test for Means Model

- Hypotheses
  - H_0: $\mu_1 = \mu_2 = \cdots = \mu_t$
  - H_a: At least one μ_j is different
- Test Statistic: $F = \frac{MS(Between\ Samples)}{MS(Within\ Samples)}$ with $t - 1$ numerator df and $n_T - t$ denominator df
- Use *p*-value from output and compare to α
- Decision about the null hypothesis
- Conclusion in the context of the scenario

According to the null hypothesis, all of the treatment means are equal

If the null hypothesis is rejected, there is good evidence at least one of the treatment means is different than the others

Hypothesis Test for Effects Model

- Hypotheses
  - H_0: $\tau_1 = \tau_2 = \cdots = \tau_t = 0$
  - H_a: At least one $\tau_j \neq 0$
- Test Statistic: $F = \frac{MS(Between\ Samples)}{MS(Within\ Samples)}$ with $t - 1$ numerator df and $n_T - t$ denominator df
- Use *p*-value from output and compare to α
- Decision about the null hypothesis
- Conclusion in the context of the scenario

According to the null hypothesis, all of the treatment effects are zero

If the null hypothesis is rejected, there is good evidence at least one of the treatment effects is not equal to zero

Section 8.4 Checking on the AOV Conditions (Assumptions)

NOTE: Out text uses AOV but I will continue to use ANOVA

We will check the conditions of ANOVA by using the following assumptions:

- Errors are normally distributed
  - Normal Q-Q plot and Shapiro-Wilk Test
- There are no outliers
  - Look for unusual points in plots
- There is a common variance
  - Levene's Test for Homogeneity of Variance

Activity: Complete the handout Checking ANOVA Assumptions