AMS 315
Data Analysis
Chapter Eight Study Guide
Inferences about More than Two Population Central Values
Spring 2023

Context

The procedures in this chapter generalize the test of the equality of means of two independent populations. This generalization is often called the one-way layout. While this design has somewhat limited value in practice, the material in this chapter is fundamental for further generalizations. The key ideas that are first developed in the one-way analysis of variance are: the generalization of the t-test, the expected mean square calculation (which is described in Chapter 14 and is crucial for power calculations), and the introduction to multiple testing of hypotheses in Chapter 9.

Chapter Eight

8.1. Introduction and Abstract of Research Study

Note the case study.

8.2 A Statistical Test about More Than Two Population Means: An Analysis of Variance

This is core material that you must master and that was covered extensively in lecture.

8.3 The Model of Observations in a Completely Randomized Design

This is also fundamental material. Note that the model that specified in class was $Y_{ij} = \mu + \tau_i + \sigma_{1W} Z_{ij}$, for $i = 1, \ldots, t$ (where t is the number of treatments), $j = 1, \ldots, n_i$, and $\sum_{i=1}^{t} n_i \tau_i = 0$. The use of Z_{ij} in this model is the assumption that the dependent variable data is normally distributed. The use of the multiplier σ_{1W} is the assumption that the variances within groups are homogeneous. The distribution of the F-test is not sensitive to violations of these assumptions. The important assumption is independence of the error terms. This is guaranteed when there is a random assignment of experimental unit to treatments. Sometimes researchers apply these techniques to data not generated by a randomized experiment. In that event, checking the assumption of independence is crucial.

8.4 Checking on the AOV Conditions

The most important assumption is that of independence. This is guaranteed in a randomized experiment in which the experimental units are randomly assigned to treatment. When the data do not come from a randomize experiment, this assumption should be checked carefully. Common problems occur when time series data (for example, an exchange rate on successive days as the dependent variable) is used. Also data describing a geographical area such as a census tract have spatial autocorrelation. Data on students from the same class will be correlated because of the common instruction.

The analysis procedures for balanced analyses of variances are not sensitive to violations of the normality assumption and the homogeneity of variance assumption. The residuals in a one-way AOV are $e_{ij} = y_{ij} - \bar{y}_i$. Residual analysis is simple for this model. One can and should generate a probability plot of the residuals. Closeness of the plot to a straight suggests that the assumption of normality appears to be true. Hartley's $F_{\max} = \frac{S^2_{\max}}{S^2_{\min}}$ is sensitive to normality. The Brown-Forsythe-Levene test is more robust to violations of the assumption of normality. Many statistical packages will calculate this test, and you should use it routinely. There is another more robust test of this null hypothesis that corrects for the estimated kurtosis of the sampled random variables. Some statistical packages report this test as well or instead of Levene's test. In the event that the hypothesis of constant variance is rejected, there are two common next steps. One is to use weighted least squares (with weights reflecting the difference in variance of observations), and the other approach is to transform the data to lessen the differences in variance. These transformations are called variance stabilizing transformations and are commonly used.

8.5 An Alternative Analysis: Transformations of the Data

Lecture Material on the "Delta Method":

The random variable *Y* has expected value μ_Y and variance σ^2_Y. The function *f* has finite derivatives. The random variable $W = f(Y)$. The delta method approximates the value of *W* using the first term of the Taylor series: $W \cong f(\mu_Y) + f'(\mu_Y)(Y - \mu_Y)$. Using this approximation, $E(W) \cong f(\mu_Y)$, and $\text{var}(W) \cong [f'(\mu_Y)]^2 \text{ var}(Y)$.

Exploratory Data Analysis Tool to Identify Variance Stabilizing Transformation

The following is material presented in lecture. When the dependent variable in an analysis of variance is always positive, calculate $\bar{y}_i$ and $s_i^2 = \frac{\sum_{i=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{n_i - 1}$, where *i*

indexes the treatments in the analysis of variance. Plot $\log(s_i)$ against $\log(\bar{y}_i)$ and fit a straight line to the data. Call the slope *m*. Then analyze the transformed values $t_{ij} = y_{ij}^{1-m}$, with $m \neq 1$. When $m = 1$, use $t_{ij} = \log(y_{ij})$. When $m \cong 0$, no transformation is necessary.

There is a related set of techniques called the Box-Cox transformations that is also helpful. We will deal with this when we study multiple regression.

8.6. A Nonparametric Alternative: The Kruskal-Wallis Test

This was not covered, and you are not responsible for it.

8.7. Research Study: Effect of Timing on the Treatment of Port-Wine Stains with Lasers

This was not covered. It is a very good discussion that I recommend to you. There is no testable material here.

8.8. Summary and Key Formulas

Use this as the basis of your notes.

Example Past Examination Questions

1. A research team wishes to specify a manufacturing process so that *Y*, the area in a product affected by surface flaws is as small as possible. They have four levels of concentration of a chemical used to wash the product before the final manufacturing step and want to determine whether the concentration level causes a change in $E(Y)$. They run a balanced one-way layout with 6 observations for each concentration with level 1 set at 10%, level 2 set at 15%, level 3 set at 20%, and level 4 set at 25%. They run a balanced one-way layout with 6 observations for each treatment. They observe that $y_{1.} = 264.5$, $y_{2.} = 255.9$, $y_{3.} = 216.2$, and $y_{4.} = 263.8$, where $y_{i.}$ is the average of the observations taken on the *i*th level. They also observe that $s_1^2 = 411.9$, $s_2^2 = 522.2$, $s_3^2 = 631.8$, and $s_4^2 = 521.9$, where s_i^2 is the unbiased estimate of the variance for the observations taken on the *i*th level.

   a. Complete the analysis of variance table for these results; that is, be sure to specify the degrees of freedom, sum of squares, mean square, and F-test.
   b. What is your conclusion? Use significance levels set to 0.10, 0.05, and 0.01. Make sure that you discuss the optimal setting of the concentration level and how you could document it.
   c. In examinations, I add parts asking for the decomposition into linear, quadratic, and cubic components. See Chapter Nine Problems.

Answers:

Analysis of Variance Table

Source	Sum of Squares	Degrees of Freedom	Mean Square	
Treatments	9467.4	3	3155.8	F=6.046
Error	10439.0	20	521.95	
Total	19906.4	23		

Answer: The critical values for the F-test are 2.38 for the 0.10 level, 3.10 for the 0.05 level, and 4.94 for the 0.01 level. Reject the null hypothesis that the mean responses of the four treatments are equal. The optimum setting is to use 20% as the concentration setting. This can be confirmed with Fisher's protected t-test.

2. A balanced one-way layout uses *J* observations at each of *I* treatment settings. Specify the formulas for the treatment sum of squares, the error sum of squares, and the total sum of squares. Prove that the treatment sum of squares added to the error sum of squares equals the total sum of squares.

Answer: Since $SS_{Total} = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - Y_{\bullet\bullet})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - Y_{i\bullet} + Y_{i\bullet} - Y_{\bullet\bullet})^2$,

$$SS_{Total} = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - Y_{i\bullet})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{i\bullet} - Y_{\bullet\bullet})^2 + 2\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - Y_{i\bullet})(Y_{i\bullet} - Y_{\bullet\bullet}).$$

Note that $SSE = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - Y_{i\bullet})^2$ and that

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{i\bullet} - Y_{\bullet\bullet})^2 = \sum_{i=1}^{I}(Y_{i\bullet} - Y_{\bullet\bullet})^2\sum_{j=1}^{J}1 = \sum_{i=1}^{I}J(Y_{i\bullet} - Y_{\bullet\bullet})^2 = SS_{Treatment}.$$

Finally, $2\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - Y_{i\bullet})(Y_{i\bullet} - Y_{\bullet\bullet}) = 2\sum_{i=1}^{I}(Y_{i\bullet} - Y_{\bullet\bullet})\sum_{j=1}^{J}(Y_{ij} - Y_{i\bullet}).$

Since $\sum_{j=1}^{J}(Y_{ij} - Y_{i\bullet}) = 0,$ the cross-product term is 0. This proves the identity.

3. A balanced one-way layout has I treatments, $I \geq 2$, with J observations, $J \geq 2$, randomly assigned to each treatment. For $i = 1,\ldots,I$ and $j = 1,\ldots,J$, Let Y_{ij} denote the outcome variable, and let $Y_{i\bullet} = \dfrac{\sum_{j=1}^{J}Y_{ij}}{J}$ and $Y_{\bullet\bullet} = \dfrac{\sum_{i=1}^{I}\sum_{j=1}^{J}Y_{ij}}{IJ}$. Prove or disprove $\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - Y_{i\bullet})(Y_{i\bullet} - Y_{\bullet\bullet}) = 0$.

Answer: see answer to question 2 above.

4. A balanced one-way layout has I treatments, $I \geq 2$, with J observations, $J \geq 2$, randomly assigned to each treatment. For $i = 1,\ldots,I$ and $j = 1,\ldots,J$, the outcome variable Y_{ij} is normally distributed with expected value $\mu + \alpha_i$ (where $\sum_{i=1}^{I} \alpha_i = 0$) and variance σ_{OW}^2. The outcome variables are independent. Note that this is the usual model for the balanced one-way layout. Find the expected value of the treatment sum of squares; that is, find $E(J\sum_{i=1}^{I}(Y_{i\bullet} - Y_{\bullet\bullet})^2)$, where $Y_{i\bullet} = \frac{\sum_{j=1}^{J} Y_{ij}}{J}$ and $Y_{\bullet\bullet} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij}}{IJ}$. Prove your answer.

Answer: $E(J\sum_{i=1}^{I}(Y_{i\bullet} - Y_{\bullet\bullet})^2) = (I-1)\sigma_{OW}^2 + J\sum_{i=1}^{I}\alpha_i^2$

5. The random variable Y, $Y > 0$, has $E(Y) = \theta$ and $\text{var}(Y) = \theta^3, \theta > 0$. Find the approximate mean and variance of $W = \ln(Y)$.

Answer: $E(W) \cong \ln(\theta)$, and $\text{var}(W) \cong \theta$.

6. The random variable $Y > 0$ has $E(Y) = \theta$ and $\text{var}(Y) = \theta^{4/3}, \theta > 0$. Find the approximate mean and variance of $W = Y^{1/3}$.

Answer: $E(W) \cong \theta^{1/3}$ and $\text{var}(W) \cong 1/9$.