

THE LINEAR MODEL AND ITS ASSUMPTIONS

Today's Class

- Framing ANOVA as a linear model for data.
 - ▣ Model assumptions.
- Violations of Assumptions
 - ▣ Violations of Distributional Assumptions
- Dealing with Heterogeneity of Variance



The Linear Model and Its Assumptions

The Linear Model and Its Assumptions

- The statistical model of the F test provides the machinery to derive the properties of the statistical tests.
- The model for the analysis of variance is an idealization.
- Real data always deviate from it to some degree.
- A researcher needs to understand the most likely violations, their effects on the analysis, and ways to avoid them.

The Statistical Model

- A random variable is a mathematical device used to represent a numerical quantity whose value is uncertain and that may differ each time we observe it.
- The value of a random variable (e.g., F or Z) is determined by its probability distribution (i.e., a density function).
 - ▣ Note that the text used Y_{ij} to designate the potential value of the dependent variable and y_{ij} to designate the actual value of the dependent variable that is observed.
 - ▣ Without loss of generality, we may use Y_{ij} .

The Linear Model

- The linear model of the analysis of variance is a mathematical statement expressing the score of any subject in any treatment condition as the linear sum of the parameters of the population.
- The model for the completely randomized single-factor design states:

$$Y_{ij} = \mu_T + \alpha_j + E_{ij},$$

- Where:
 - Y_{ij} is the i^{th} observation under treatment α_j
 - μ_T is the grand mean of the treatment populations
 - $\alpha_j = \mu_j - \mu_T$ is the treatment effect for α_j
 - $E_{ij} = Y_{ij} - \mu_j$ is the experimental error.

The Linear Model

The null hypothesis,

$$H_0: \mu_1 = \mu_2 = \mu_3 = \text{etc.},$$

is equivalent to

$$H_0: \alpha_1 = 0, \quad \alpha_2 = 0, \quad \alpha_3 = 0, \quad \text{etc.}$$

and

$$H_0: \sum_{j=1}^a \alpha_j^2 = 0.$$

The Experimental Error

- The properties of the random variable E_{ij} are determined by a series of assumptions.
 1. Independence: The value of E_{ij} is independent of its value for all other subjects.
 2. Identical distribution within group: The distribution of E_{ij} is the same for every subject in a treatment group.
 3. Identical distribution between groups: The distribution of E_{ij} is the same for all treatment groups.

The Experimental Error

- The properties of the random variable E_{ij} are determined by a series of assumptions.
 4. Homogeneity of variance: The variance of the random variable E_{ij} is the same for all groups.
 5. Normal distribution: The random variable E_{ij} has a normal distribution centered around a mean of zero.

Expected Value

- The expected value of a statistic is the mean of the sampling distribution of that statistic obtained from repeated random sampling from the population.
- It is denoted by the letter E.
- For a given variable Y , we typically say $E(Y) = \mu_Y$

Expected Mean Squares and the F Ratio

- The within-groups mean square $MS_{S/A}$ provides an unbiased estimate of error variance,

$$E(MS_{S/A}) = \sigma_{\text{error}}^2.$$

- The expected value of the treatment square is:

$$E(MS_A) = n \frac{\sum_{j=1}^a \alpha_j^2}{a - 1} + \sigma_{\text{error}}^2.$$

Expected Mean Squares and the F Ratio

- Note that $E(MS_A)$ in the above equations reflects the fixed-effects model where the levels of the treatment variable have been selected arbitrarily (cf. random-effects model).
- Under the null hypothesis the ratio $F = MS_A / MS_{S/A}$ is distributed as $F(df_A, df_{S/A})$ provided that the assumptions are satisfied.
 - ▣ This means the Expected Value of the ratio is 1.0.



Violations of Assumptions

Violations of the Assumptions

- The F test is robust to some violations of the assumptions.
- There are two categories of violations:
 1. Some violations, particularly those affecting the randomness of the sampling, compromise the entire set of inferences drawn from the study (see Section 7.2).
 2. Others, such as concerns about the distributional shape, affect mainly the accuracy of the statistical tests themselves-their Type I error probability and their power (see Sections 7.3 and 7.4).

Sampling Bias and the Loss of Subjects

- The ideal is to randomly draw the sample from the population, so that each member of the population is equally likely to appear.
- The random assignment may eliminate any bias in recruiting the subjects.
- The potential for bias exists in nonexperimental designs that compare preexisting groups (e.g., men versus women).

Sampling Bias and the Loss of Subjects

- Researchers sometimes try to find evidence that no differential sampling bias has occurred (e.g., showing no differences in age, education, and so forth).
- Sometimes the new factor can be introduced as a blocking factor, and other times the analysis of covariance is used.

Loss of Subjects

- The researcher now has an unbalanced design.
- The calculations needed to analyze the design are more complex.
- The loss of subjects potentially damages the equivalence of groups that were originally created randomly.

Examples of Subject Loss

- In animal studies, subjects are frequently lost through death and sickness.
- In a human study in which testing continues over several days, subjects are lost when they fail to complete the experimental sequence.

Ignorable and Nonignorable Loss of Subjects

- Faced with subject loss, a researcher must decide whether it is random, that is, whether it is ignorable or nonignorable.
- We will call a loss that does not disturb the random formation of the groups is missing at random or ignorable and one that does, missing not at random or nonignorable.



Violations of Distributional Assumptions

Violations of Distributional Assumptions

- If the violation is not critical, the proportion of rejection of the null hypothesis is essentially the same as α , the nominal significance level.
- The test is then robust with respect to the violation of such assumptions.
- If the observed proportion exceeds the nominal α level, the test is liberal (i.e., positively biased).
- If the observed proportion is less than the nominal α level, the test is conservative (i.e., negatively biased).

Independence of the Scores

- The scores are independent within treatment groups as well as independent between treatment groups. Independence means that each observation is in no way related to any other observations in the experiment.
- Violations of independence, if substantial, can be quite serious.

Identical Within-Group Error Distribution

- The most common violation of the within-group identical distribution assumption occurs when the population contains subgroups of subjects with substantially different statistical properties.
- We may introduce additional factors into the design to correct this problem.

Normally Distributed Error

- The shape of the normal distribution is characterized by three properties; unimodality, symmetry (cf. skewness), and moderate spread (cf. kurtosis).
- The simplest way to check for these characteristics is to construct a histogram of your scores and look at its shape.
- The residual, $E_{ij} = Y_{ij} - [\bar{Y}]_i$, can also be plotted for all groups at once.
- The F test is not particularly affected when samples become as large as a dozen (i.e., $n = 12$) (Clinch & Keselman, 1982; Sawilowsky & Blair, 1992; Tan, 1982).
- We may use nonparametric tests instead of the analysis of variance when data are not normal (e.g., Kruskal-Wallis test).

Between-Group Differences in Distribution-Homogeneity of Variance

- Box (1954b) suggested that the F test was relatively insensitive to the presence of variance heterogeneity, except when unequal sample sizes were involved.
- More recent work summarized by Wilcox (1987a), however, questions this earlier conclusion even with equal samples.

$$(especially, F_{\max} = s_{\text{largest}}^2 / s_{\text{smallest}}^2 \geq 9)$$



Dealing with Heterogeneity of Variance

Dealing with Heterogeneity of Variance

- Most researchers do not assess the validity of the homogeneity assumption (violations have little consequences for the F test; and no good tests exist for testing variance heterogeneity).

Testing the Differences Among Variances

- The null hypothesis is

$$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \text{etc.}$$

- The alternative hypothesis is

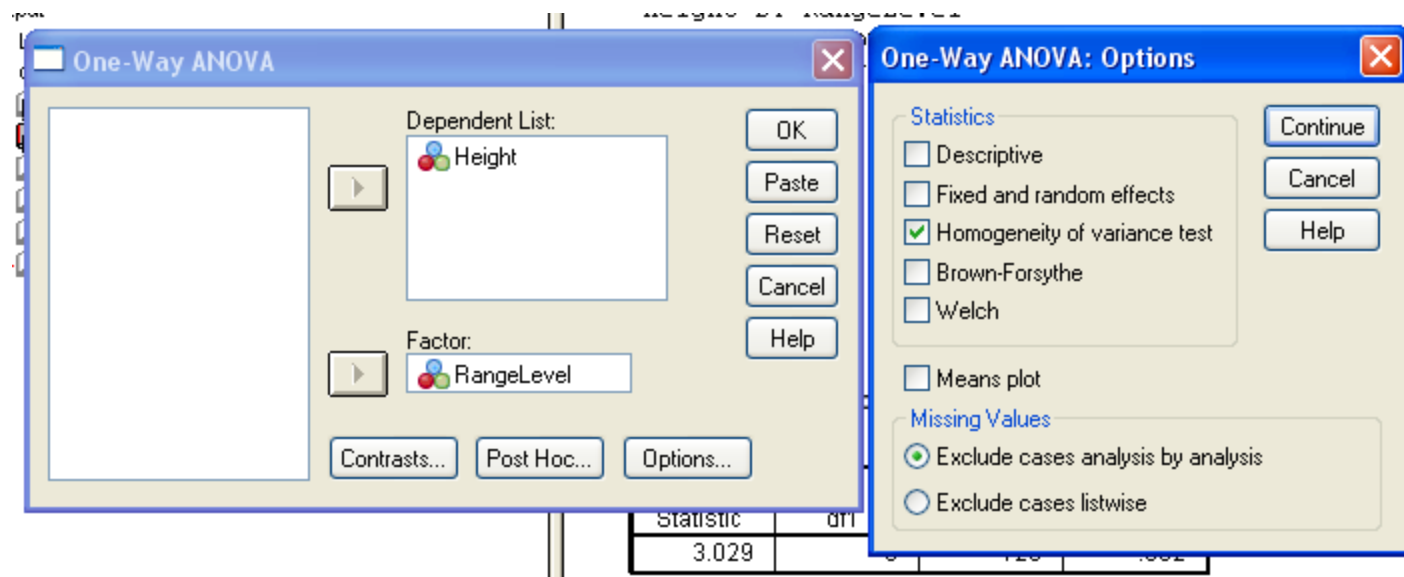
$$H_1: \text{not all } \sigma_j^2 \text{'s are equal.}$$

Testing the Differences Among Variances

- Many tests are available for testing H_0 (e.g., Hartley test using the F_{\max} statistic, Cochran test, and Bartlett test).
- ▣ See Conover, Johnson, and Johnson (1981) for the evaluation of the 56 different tests.
- ▣ The F_{\max} test may not be satisfactory if scores are not normally distributed.

SPSS: Levine's Test

- You can test for the homogeneity of variance in SPSS:



SPSS Output:

Test of Homogeneity of Variances

Height

Levene Statistic	df1	df2	Sig.
3.029	3	126	.032

This tests the null hypothesis from slide 28.

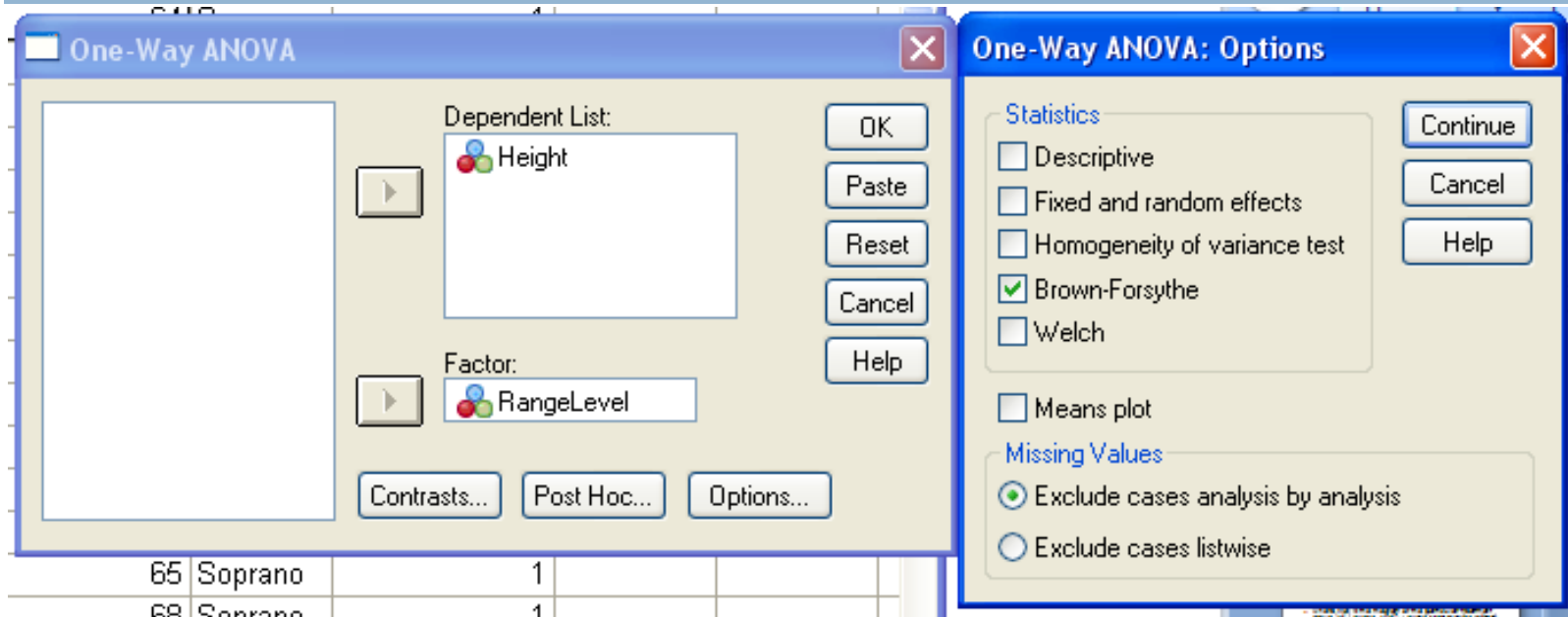
Brown and Forsythe Test

- Brown and Forsythe (1974a) proposed a test based on transformed Y_{ij} to Z_{ij} using medians of the treatment groups:


$$Z_{ij} = | Y_{ij} - Md_j |,$$

- where Md_j is the median of the particular treatment group (cf. Levene, 1960; Levene's test may be inferior).
- Once Z_{ij} are calculated, conduct an ordinary analysis of variance. If the F is significant, there exists heterogeneity.

SPSS: Brown and Forsythe Test



SPSS Output



Height	Statistic ^a	df1	df2	Sig.
Brown-Forsythe	50.648	3	78.346	.000

a. Asymptotically F distributed.

This is treated like the Omnibus F test...it is just more robust to heteroscedasticity.

Testing the Means When the Variances Differ

- **Use a More Stringent Significance Level.**

- A more stringent criterion, say, $\alpha = .025$ can be used instead of the conventional .05 level.

- **Transform the Data.** We can apply such transformations as:

$$Y_{ij}' = \sqrt{Y_{ij} + 0.5}$$

$$Y_{ij}' = \log(Y_{ij} + 1)$$

$$Y_{ij}' = 2 \arcsin\left(\sqrt{\frac{Y_{ij}}{Y_{ij} + 1}}\right)$$

Testing the Means When the Variances Differ

□ **Alternatives to the Analysis of Variance.**

- Some of the more commonly referenced tests are by Welch (1938, 1951), Brown and Forsythe (1974b), and two versions by James (1951) (see Coombs, Algina, & Oltman, 1966, for a summary of these methods and Johansen, 1980, for a formulation that links them).
- The best choice appears to be the second version of James's method, usually known as James's second-order method.

□ **Emphasize Single-df Tests.**

- It is easy to accommodate unequal variances into the single-df tests.

Contrasts with Heterogeneous Variance

- The t statistic is:

$$t_{\psi} = \frac{\hat{\psi}}{s[\hat{\psi}]}$$

- Where:

$$\hat{\psi} = \sum_{j=1}^a c_j \bar{Y}_j$$

Contrasts with Heterogeneous Variance

□ And...

$$s_{[\gamma(\psi)]} = \sqrt{\sum_{j=1}^a c_j^2 s_{Mj}^2}$$

□ With the degrees of freedom (Satterthwaite, 1941, 1946):

$$df = \frac{s_{[\gamma(\psi)]}^4}{\sum_{j=1}^a \frac{c_j^4 s_{Mj}^4}{n_j - 1}}$$

Recalling the Contrast Output

Contrast Tests							
Contrast			Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
Height	Assume equal variances	1	-10.73	.914	-11.748	126	.000
	Does not assume equal	1	-10.73	.990	-10.837	58.818	.000

The “Does not assume equal” box adjusts the contrast DF according to unequal variances.

Final Thought

- Phrasing the ANOVA model as a linear model allows for us to understand how assumptions are placed on our data.
- To the extent the assumptions are valid so to will our hypothesis tests.
 - If assumptions are violated, decisions based on hypothesis test may be incorrect.



Next Class

- Chapter 8: Effect size, power, and sample size.