The University of Georgia

The Linear Model and Its Assumptions
Chapter 7

ERSH 8310

Lecture 6

September 23, 2009

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

But First…

- Tomorrow is National Punctuation Day…
- Laine has highlighted a very cool punctuation mark: the interrobang.

‽

- The combination of the question mark and the exclamation point.
  - “They did what‽”
- Find it with ALT+8253 on your keyboard…

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

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 j
  - μ_T is the grand mean of the treatment populations
  - $\alpha_j = \mu_j - \mu_T$ is the treatment effect for α_j (must sum to zero)
  - $E_{ij} = Y_{ij} - \mu_j$ is the experimental error

Putting Numbers on the Model

- Recall our vigilance task example
  - 4 conditions of sleep deprivation (4, 12, 20, 28 hours)
  - 4 subjects per condition
  - DV is number of errors made in 30 minutes

- Grand mean number of errors = 45.875
  - 4 hour mean = 26.5
  - 12 hour mean = 37.75
  - 20 hour mean = 57.5
  - 28 hour mean = 61.75

The Estimated Linear Model

- Here, we have several quantities:

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

- μ_T (grand mean) = 45.875

- α_1 (effect for group 1: 4 hour condition) =
  - 26.5 – 45.875 = -19.375
- α_2 (effect for group 2: 12 hour condition) =
  - 37.75 – 45.875 = -8.125
- α_3 (effect for group 3: 20 hour condition) =
  - 57.5 – 45.875 = 11.625
- α_4 (effect for group 4: 28 hour condition) =
  - 61.75 – 45.875 = 15.875

Hours	Errors	μ_T	α_1	α_2	α_3	α_4	$\hat{Y}_{ij}$	E_{ij}
4	37	45.875	-19.375				26.50	10.50
4	22	45.875	-19.375				26.50	-4.50
4	22	45.875	-19.375				26.50	-4.50
4	25	45.875	-19.375				26.50	-1.50
12	36	45.875		-8.125			37.75	-1.75
12	45	45.875		-8.125			37.75	7.25
12	47	45.875		-8.125			37.75	9.25
12	23	45.875		-8.125			37.75	-14.75
20	43	45.875			11.625		57.50	-14.50
20	75	45.875			11.625		57.50	17.50
20	66	45.875			11.625		57.50	8.50
20	46	45.875			11.625		57.50	-11.50
28	76	45.875				15.875	61.75	14.25
28	66	45.875				15.875	61.75	4.25
28	43	45.875				15.875	61.75	-18.75
28	62	45.875				15.875	61.75	0.25

The Linear Model

The null hypothesis,

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_a$$

is equivalent to

$$H_0 : \alpha_1 = 0, \alpha_2 = 0, \ldots, = \alpha_a = 0$$

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,

$$E(MS_{S/A}) = \sigma^2_{error}$$

- The expected value of the treatment mean square is:

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

Expected Mean Squares and the F Ratio

- $E(MS_A)$ reflects the fixed-effects
  - The levels of the treatment variable have been selected arbitrarily

- 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
  - Under the null, all α_j are equal to zero
  - 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 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 preexising 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 a new factor can be introduced
  - A blocking factor

- Other times the analysis of covariance is used

Loss of Subjects

- When subjects are lost:
  - Now an unbalanced design
    - Not equal n in all conditions
  - Calculations needed to analyze the design are more complex
    - Not an issue in modern statistical computing
  - Potentially damages the equivalence of groups that were originally created randomly
    - Biggest problem, particularly if loss is not random

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
  - Ignorable

- Loss that does disturb the random formation of the groups is missing not at random
  - 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 a, the nominal significance level
- The test is then robust with respect to the violation of such assumptions
- If the observed proportion exceeds the nominal a level, the test is liberal (i.e., positively biased)
- If the observed proportion is less than the nominal a level, the test is conservative (i.e., negatively biased)

Independence of the Scores

- The scores are assumed independent within treatment groups and independent between treatment groups

- Independence means that each observation is in no way related to any other observations in the experiment

- Violations of independence can be quite serious
  - Other statistical techniques used
    - Hierarchical Linear Models (A.K.A. mixed models or multilevel models)

Identical Within-Group Error Distribution

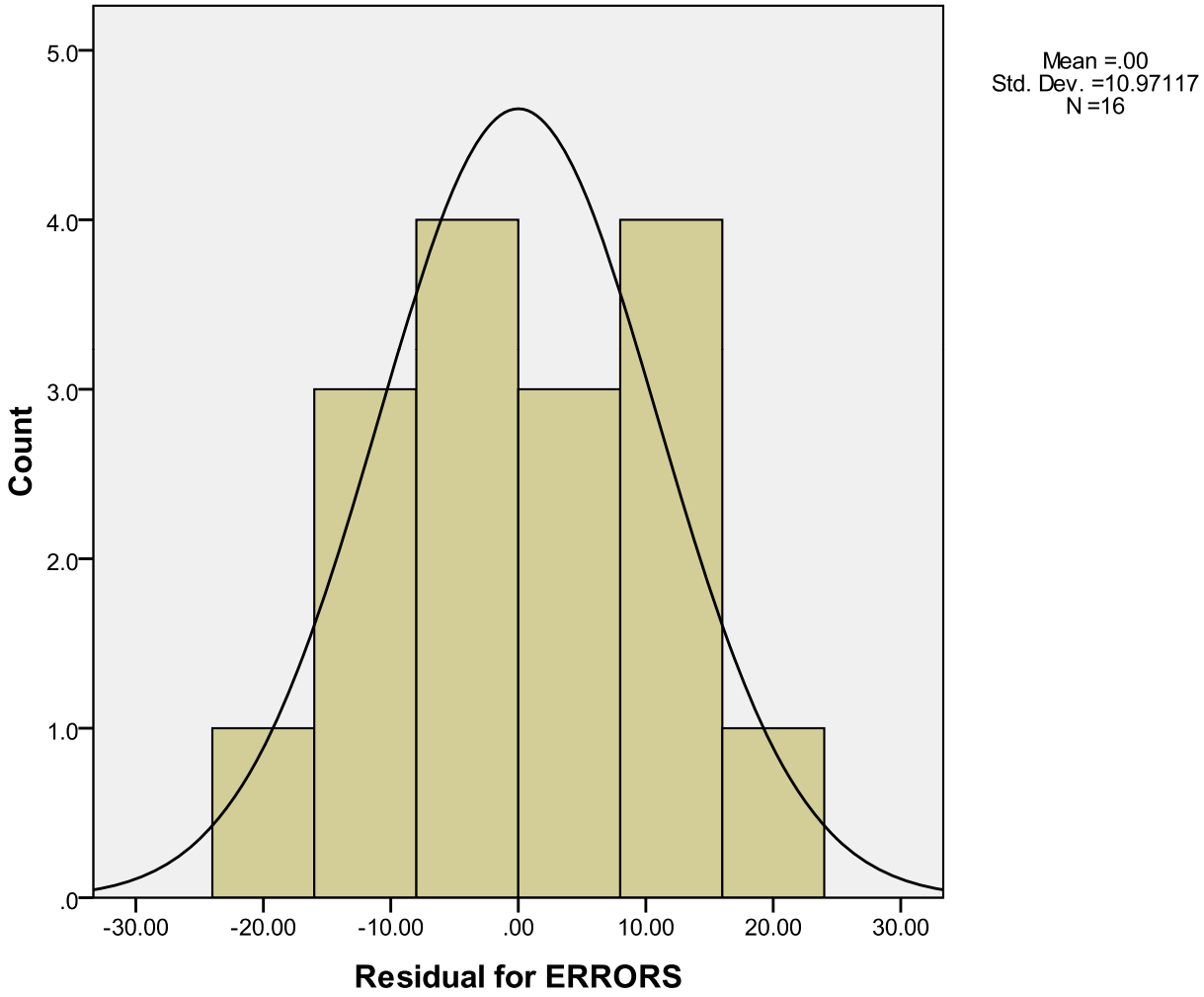
- The most common violation of the within-group identical distribution assumption occurs when the population contains subgroups of subjects with substantially different statistical properties

- 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 residuals, E_{ij}, 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)

Vigilance Errors



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

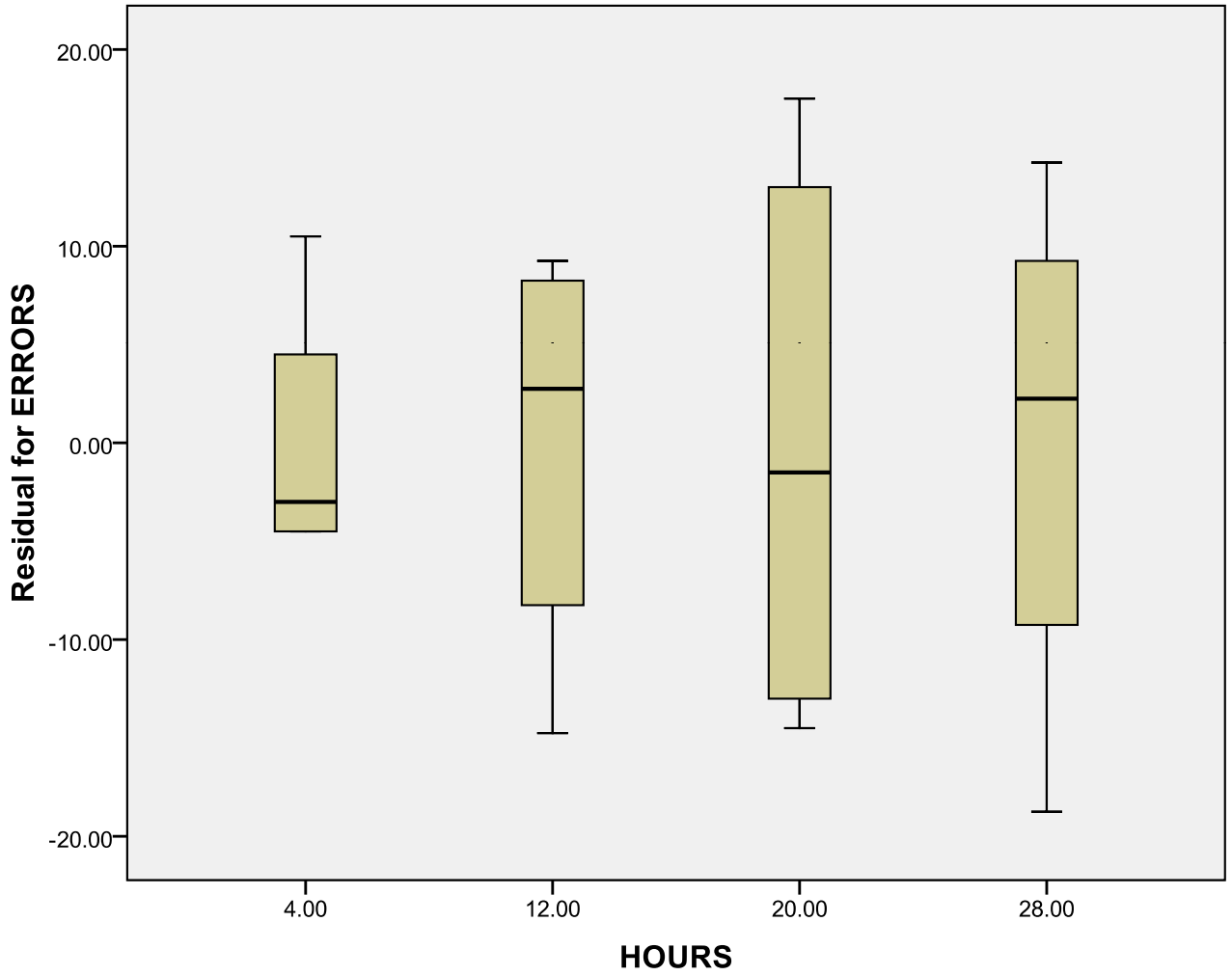
- A rule of thumb is that the largest group variance should be no larger than 9 times the smallest group variance:

$$F_{Max} = \frac{s^2_{largest}}{s^2_{smallest}} \leq 9$$

Vigilance Task (Again)

- Largest Variance:
  - 240.333 (20 hour condition)
- Smallest Variance:
  - 51.000 (4 hour condition)
- Ratio:
  - 4.71

Boxplot of Errors by Group



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
- No good tests exist for testing variance heterogeneity

Testing the Differences Among Variances

- The null hypothesis is

$$H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_a^2$$

- The alternative hypothesis is

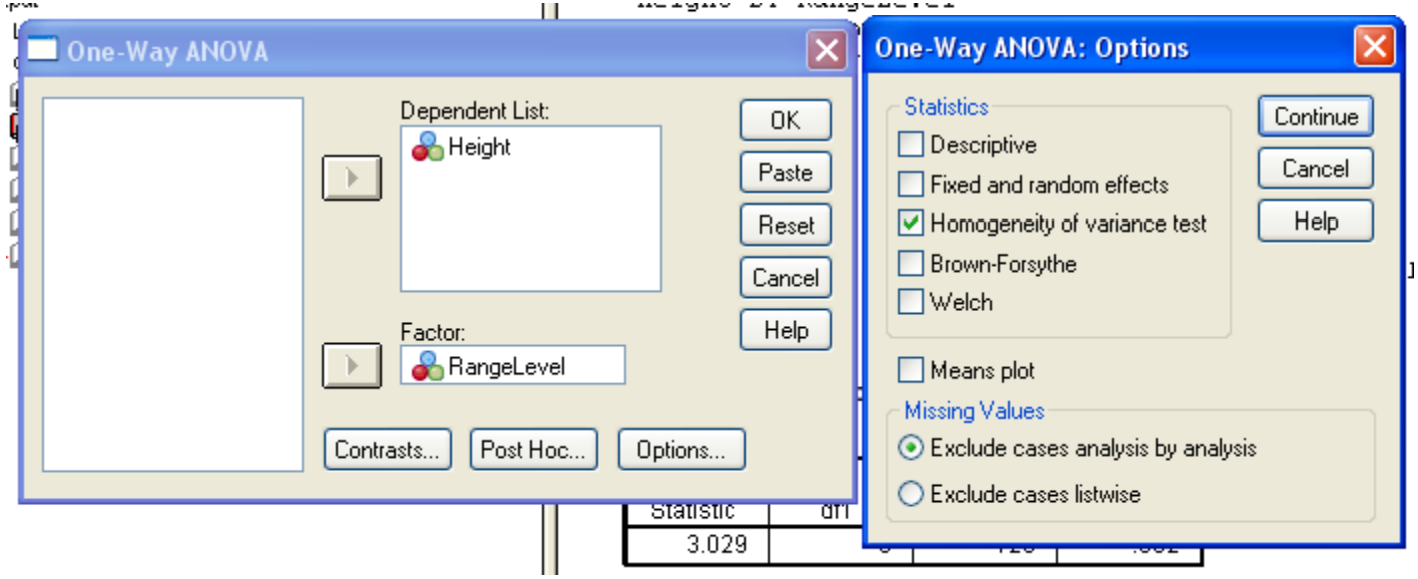
$$H_1: \text{Not all } \sigma_j^2 \text{ equal}$$

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

ERRORS

Levene Statistic	df1	df2	Sig.
1.245	3	12	.337

This tests the null hypothesis of equal variances

Brown and Forsythe Test

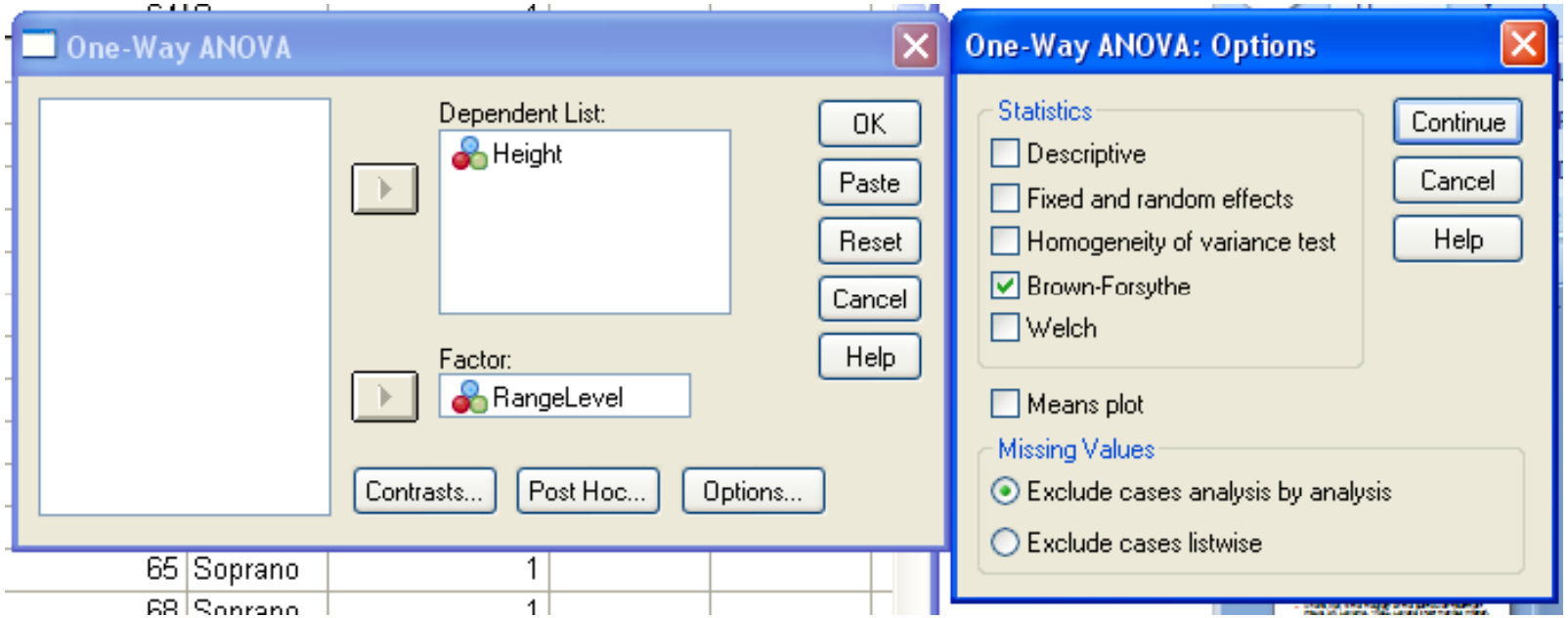
- Brown and Forsythe (1974a) proposed a test based on transformed Y_{ij} to Z_{ij} using medians of the treatment groups:

$$Z_{ij} = \left| Y_{IJ} - Md_j \right|$$

- Here, Md_j is the median of treatment group j

- Once Z_{ij} are calculated, an ordinary ANOVA is run with Z as the DV
  - If the F is significant, there exists heterogeneity of variances

SPSS: Brown and Forsythe Test



SPSS Output

Robust Tests of Equality of Means

ERRORS

	Statistic[a]	df1	df2	Sig.
Brown-Forsythe	7.343	3	9.780	.007

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, a = .025 can be used instead of the conventional .05 level

- Transform the data
  - We can apply such transformations as:

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

$$Y'_{ij} = \log\left(Y_{ij} + 1\right)$$

$$Y'_{ij} = 2\sin^{-1}\left(\sqrt{Y_{ij}}\right)$$

Testing the Means When the Variances Differ

- Alternatives to ANOVA
  - 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
  - i.e., Contrasts

Contrasts with Heterogeneous Variance

- The t statistic is:

$$t_{\psi} = \frac{\hat{\psi}}{s_{\hat{\psi}}}$$

- Where:

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

Contrasts with Heterogeneous Variance

- And…

$$s_{\hat{\psi}} = \sqrt{\sum_{j=1}^{a} c_j^2 s_{M_j}^2}$$

- With the degrees of freedom (also known as the Satterthwaite df, 1941, 1946):

$$df = \frac{s_{\hat{\psi}}^4}{\sum_{j=1}^{a} \frac{c_j^4 s_{M_j}^4}{n_j - 1}}$$

Recalling the Contrast Output

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
ERRORS	Assume equal variances	1	-35.2500	8.67347	-4.064	12	.002
	Does not assume equal variances	1	-35.2500	7.77684	-4.533	4.496	.008

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

Wrapping Up

- Formulating the ANOVA model as a linear model allows for us to understand how assumptions are placed on our data
  - Also will generalize to other statistics courses you may take

- 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

Up Next…

- In Lab:
  - How to do assumption checking in SPSS
- Homework:
  - Posted tomorrow morning
  - Due next week at the start of class
- Next week:
  - Chapter 8: Effect size, power, and sample size