Keppel, G. & Wickens, T. D. *Design and Analysis*
Chapter 7: The Linear Model and Its Assumptions

7.1 The Statistical Model

The Linear 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 linear model that underlies ANOVA expresses each person's score as the linear sum of parameters

$$Y_{ij} = \mu_T + \alpha_j + \varepsilon_{ij} \qquad (7.3)$$

where Y_{ij} is the score for the i^{th} person in treatment a_j,

μ_T is the grand mean (over all treatment means)

α_j is the treatment effect (μ_j - μ_T)

ε_{ij} is the experimental error (Y_{ij} - μ_j)

• Thus, when there is no treatment effect ($\alpha_j = 0$), the group means would be the same and your best prediction of Y_{ij} would be μ_T, though (as always) there will be error variability (ε_{ij}). In other words, H_0 would be true.

The Experimental Error

• Two assumptions underlying ANOVA are likely to be met as a result of random sampling: Independence and Identical Distribution. Basically, these assumptions mean that participants' performance is unaffected by the performance of other participants and that each participant's score is unconstrained (free to vary), so that you have no better prediction of any participant's score.

• Another assumption is that the error variability is the same for all groups. That is, their means may differ, but their variances do not. In essence, this is the homogeneity of variance assumption. Violation of this assumption can be fairly serious, especially if sample sizes are unequal (so keep *n* the same for each group). As a rough estimate of the extent to which the situation is serious, compute a ratio of the largest sample variance to the smallest sample variance, called F_{Max}. If $F_{Max} \geq 9$, then you might have a problem. When the assumption is violated, you have an increased chance of making a Type I error. Thus, your critical value is cutting off more of the null distribution than the α you think you're using.

• We typically assume that error is normally distributed around a mean of zero. If this assumption is violated, the *F* test is not much affected as long as the population is symmetrical and *n* is fairly high ($n \geq 12$). One strategy might be to transform the data to make it more normal. If you're concerned about lack of symmetry, lower α to .025 or .01.

Expected Mean Squares and the F Ratio

• "...the **expected value** of a statistic is the mean of the sampling distribution of that statistic obtained from repeated sampling from the population. (K3rd)"

• $E(MS_{S/A}) = \sigma^2_{error}$ means that the expected MS_{Error} is the population variance (7.4)

• $E(MS_A) = \sigma^2_{error} + n\frac{\sum(\alpha_i)^2}{a-1}$ or $E(MS_A) = \sigma^2_{error} + n(\theta^2_A)$ (7.5)
means that the expected mean square treatment is a combination of population variance and treatment effect in a fixed effects model (which assumes that the levels of the treatment were selected by the experimenter in a purposeful fashion, rather than randomly)
• The ANOVA is based on the fact that the F ratio is a ratio of these two expected mean squares. Thus,

$$F = \frac{\sigma^2_{error} + n(\theta^2_A)}{\sigma^2_{error}} \quad \text{which means that} \quad F = \frac{Error + Treatment}{Error}$$

• Under the null hypothesis, these F values will be distributed as $F(df_A, df_{S/A})$.

Violations of the Assumptions
• It turns out that the F is fairly robust, which means that it is not affected by some violations of the distributional assumptions (normal distribution and homogeneity of variance). Some researchers presume that the F is so robust that they need not concern themselves with implications of violations of distributional assumptions. Other researchers are really concerned about the impact of violation of distributional assumptions on the F. (I imagine that most researchers would agree that it's important to meet the random sampling assumptions.)

7.2 Sampling Bias and the Loss of Subjects

Exclusion of Subjects from the Experiment (Sample Bias)
• If we are interested in extrapolating to a population from the sample(s), then we need to be certain that our sampling procedure is not biased in any way. That is, are we systematically excluding a particular type of participant from our samples that will lead us to a biased estimate of population parameters?
• Often, however, we're simply interested in determining the impact of our treatments (cf. Mook). Under such circumstances, with random assignment to condition, it's unlikely that differences that emerge among the treatments are due to anything but the treatment differences.
• Such logic is of no use, however, when we're dealing with non-manipulated characteristics of participants (gender, age, intelligence, etc.). With no manipulation, the possibility of a biased sample means that we must be very much concerned about our sampling procedure.

Loss of Subjects (Mortality)
• "When observations are lost or unavailable for some reason, we should ask what effects the exclusion has on the results." However, it's probably not a big problem if the "lost" participants are roughly equivalent in number across our conditions. Furthermore, if the reasons for the loss are identifiable as random (e.g., equipment malfunction), that's not a problem either.
• Needless to say, regardless of the loss of participants, we would be able to compute an ANOVA. The problem is that we would not be able to interpret the outcome of the ANOVA. "...when the loss is related to the phenomena under study, randomness is destroyed and

systematic bias may be added to the differences among the means. This bias cannot be disentangled from the treatment effects."

7.3 Violations of Distributional Assumptions

• K&W discuss violations of distributional assumptions in separate sections, but note that they are all interrelated.
• Using a Monte Carlo approach, the logic is to construct a set of scores such that a particular distributional assumption is violated, but for which H_0 is true. A large number of sets of samples are then run through ANOVAs to see if the proportion of "reject" decisions is what it should be (α). If so, then the test is robust relative to the violation of that assumption (in that way). A positive bias arises when the proportion of incorrect reject decisions is greater than α (liberal), and a negative bias arises when the proportion is less than α (conservative).

Independence of the Scores
• Violation of the independence assumption is serious. In essence, this bias emerges when scores are not independent of one another. As K&W note, this bias can emerge in many ways. One example they provide is that when collecting data in groups, the participants may influence one another, which means that their scores are not truly independent. In a carefully planned experiment, you should be able to eliminate the opportunity for scores to be dependent.

Identical Within-Group Error Distribution
• "The most common violation of the within-group identical-distribution assumption in experimental studies occurs when the population contains subgroups of subjects with substantially different statistical properties." For example, suppose that you include males and females in your study (distributed throughout your groups) that come from populations with different means and variances. The major negative impact of such a situation, I think, is the inflation of the error term. (You should be able to articulate why.) Introducing gender as a factor, however, would deal with this problem. So, if you can anticipate sources of violation of this assumption, you may be able to address the violations via addition of factors to your study.

Normally Distributed Error
• The normal distribution (determined by a mathematical formula) is unimodal, symmetrical (skewness measure), and has moderate spread (kurtosis measure). To assess deviations from normality, it's a good idea to have the computer construct a graph of your data.
• However, remember that even when the population from which scores are sampled may be non-normal, the sampling distribution of the mean will approach normality as sample size increases (central limit theorem). As a result, once sample sizes are sufficiently large (≥ 12) we need not worry much about violations of the normality assumption.
• You should still investigate your data for the presence of outliers (i.e., more than 3 standard deviations from the mean), because they may have an influence on your interpretation of results.
• If you are concerned about departures from normality, you should probably shift to nonparametric tests such as Kruskal-Wallis.

Between-Group Differences in Distribution—Homogeneity of Variance
• Differences in the variances of the populations from which our groups were sampled may well pose serious problems. K&W discuss ways in which this heterogeneity of variance may emerge.
• K&W illustrate the impact of heterogeneity of variance in Table 7.1 Below is a portion of that table to illustrate the point with sample size (*n*) of 10. These results show the actual proportion of Type I errors based on 100,000 simulated experiments.

# of Groups (*a*)	Pattern of Variances	Nominal α = .05
3	1,1,1	.050
3	1,2,3	.054
3	1,4,9	.063
3	1,9,25	.069
6	1,1,4,4,9,9	.073
9	1,1,1,4,4,4,9,9,9	.078

• Note, first of all, that when the variances are homogeneous (or nearly so), the actual proportion of Type I errors is quite close to α. However, as the pattern of variances becomes more heterogeneous, the proportion of Type I errors increases. Not shown here, but seen in Table 7.1 is the fact that sample size doesn't seem to have that big an influence—as long as sample size is equal among groups. However, with samples of different size, the ill effects of heterogeneity of variance are exacerbated.

7.4 Dealing with Heterogeneity of Variance

Testing the Differences Among Variances
• We need to test the assertion that the homogeneity of variance assumption is reasonable. In essence, we need to test the following set of statistical hypotheses:

$$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = etc.$$
$$H_1: not H_0$$

• Several tests of these hypotheses are problematic, such as the Hartley's *F*-Max approach, which you might have learned in Stats I. To test for heterogeneity of variance with the Hartley's *F*-Max approach, you would first create a ratio of the largest to the smallest variance in your groups. Then you would compare that ratio to the table of critical *F*-Max values. The table (from Gravetter & Wallnau) is seen below:

CRITICAL VALUES FOR THE *F*-MAX STATISTIC*

*The critical values for $\alpha = .05$ are in lightface type, and for $\alpha = .01$, they are in boldface type.

	k = NUMBER OF SAMPLES										
$n-1$	2	3	4	5	6	7	8	9	10	11	12
4	9.60	15.5	20.6	25.2	29.5	33.6	37.5	41.4	44.6	48.0	51.4
	23.2	**37.**	**49.**	**59.**	**69.**	**79.**	**89.**	**97.**	**106.**	**113.**	**120.**
5	7.15	10.8	13.7	16.3	18.7	20.8	22.9	24.7	26.5	28.2	29.9
	14.9	**22.**	**28.**	**33.**	**38.**	**42.**	**46.**	**50.**	**54.**	**57.**	**60.**
6	5.82	8.38	10.4	12.1	13.7	15.0	16.3	17.5	18.6	19.7	20.7
	11.1	**15.5**	**19.1**	**22.**	**25.**	**27.**	**30.**	**32.**	**34.**	**36.**	**37.**
7	4.99	6.94	8.44	9.70	10.8	11.8	12.7	13.5	14.3	15.1	15.8
	8.89	**12.1**	**14.5**	**16.5**	**18.4**	**20.**	**22.**	**23.**	**24.**	**26.**	**27.**
8	4.43	6.00	7.18	8.12	9.03	9.78	10.5	11.1	11.7	12.2	12.7
	7.50	**9.9**	**11.7**	**13.2**	**14.5**	**15.8**	**16.9**	**17.9**	**18.9**	**19.8**	**21.**
9	4.03	5.34	6.31	7.11	7.80	8.41	8.95	9.45	9.91	10.3	10.7
	6.54	**8.5**	**9.9**	**11.1**	**12.1**	**13.1**	**13.9**	**14.7**	**15.3**	**16.0**	**16.6**
10	3.72	4.85	5.67	6.34	6.92	7.42	7.87	8.28	8.66	9.01	9.34
	5.85	**7.4**	**8.6**	**9.6**	**10.4**	**11.1**	**11.8**	**12.4**	**12.9**	**13.4**	**13.9**
12	3.28	4.16	4.79	5.30	5.72	6.09	6.42	6.72	7.00	7.25	7.48
	4.91	**6.1**	**6.9**	**7.6**	**8.2**	**8.7**	**9.1**	**9.5**	**9.9**	**10.2**	**10.6**
15	2.86	3.54	4.01	4.37	4.68	4.95	5.19	5.40	5.59	5.77	5.93
	4.07	**4.9**	**5.5**	**6.0**	**6.4**	**6.7**	**7.1**	**7.3**	**7.5**	**7.8**	**8.0**
20	2.46	2.95	3.29	3.54	3.76	3.94	4.10	4.24	4.37	4.49	4.59
	3.32	**3.8**	**4.3**	**4.6**	**4.9**	**5.1**	**5.3**	**5.5**	**5.6**	**5.8**	**5.9**
30	2.07	2.40	2.61	2.78	2.91	3.02	3.12	3.21	3.29	3.36	3.39
	2.63	**3.0**	**3.3**	**3.5**	**3.6**	**3.7**	**3.8**	**3.9**	**4.0**	**4.1**	**4.2**
60	1.67	1.85	1.96	2.04	2.11	2.17	2.22	2.26	2.30	2.33	2.36
	1.96	**2.2**	**2.3**	**2.4**	**2.4**	**2.5**	**2.5**	**2.6**	**2.6**	**2.7**	**2.7**

Table 31 of E. Pearson and H. O. Hartley, *Biometrika Tables for Statisticians*, 2nd ed. New York: Cambridge University Press, 1958. Adapted and reprinted with permission of the Biometrika trustees.

• For the K&W51 data set, the largest variance is 240.33 and the smallest variance is 51.00. Thus, the F_{Max} ratio is 4.7. Using the table is a bit difficult because of the small n (4), chosen by K&W for ease of computation, I'm sure. (The F_{Max} table presumes that you would have no n smaller than 5.) Nonetheless, you can intuit that the critical F_{Max} value would be larger than 20.6, so you would conclude that you have no reason to be concerned about heterogeneity of variance.

• In light of problems with Hartley's F_{Max}, K&W suggest the Brown-Forsythe procedure. At the heart of this procedure is the z_{trans} score (which isn't really a z score as you learned them in Stats I). In essence, you first compute a new score for every data point in your study using

$$z_{ij} = \left| Y_{ij} - Md_j \right| \qquad (7.7)$$

where Y_{ij} is an individual's score and Md_j is the median of the group from which that score was drawn. Then, you compute an ANOVA on the z_{trans} scores just as you would on raw scores. If the F ratio is significant, you have cause for concern that you've violated the homogeneity of variance assumption. If you're particularly concerned about detecting violations of the assumption, use $\alpha > .05$. Keep in mind, however, that unlike the usual F, when testing for heterogeneity of variance, we typically want to retain H_0.

• K&W show the computation of Brown-Forsythe for one data set (Table 7.2). Below, I'll show the analyses of the K&W51 data set.

	4 Hour (Md = 23.5)	12 Hour (Md = 40.5)	20 Hour (Md = 56)	28 Hours (Md = 64)
	37 – 23.5 = |13.5|	36 – 40.5 = |-4.5|	43 – 56 = |-13|	76 – 64 = |12|
	22 – 23.5 = |-1.5|	45 – 40.5 = |4.5|	75 – 56 = |19|	66 – 64 = |2|
	22 – 23.5 = |-1.5|	47 – 40.5 = |6.5|	66 – 56 = |10|	43 – 64 = |-21|
	25 – 23.5 = |1.5|	23 – 40.5 = |-17.5|	46 – 56 = |-10|	62 – 64 = |-2|
Sum (Z)	18	33	52	37
ΣZ^2	189	389	730	593

$[Z] = 1901$ $[A] = 1371.5$ $[T] = 1225$

Source	***SS***	***df***	***MS***	***F***
ztrans	146.5	3	48.8	1.11
error	529.5	12	44.1	
total	676	15		

With $\alpha = .05$ and $F_{Crit}(3,12) = 3.49$, we would retain H_0, indicating that the homogeneity of variance assumption is reasonable. (Remember the H_0 that we're testing in the Brown-Forsythe test!) Even if we set α to .25 (F_{Crit} would become 1.56), we would retain H_0.

Testing the Means When the Variances Differ

• So, what should you do if you conduct a Brown-Forsythe analysis and determine that your data violate the homogeneity of variance assumption? As K&W suggest, the best strategy appears to be to adopt $\alpha = .025$ (or smaller) when you're concerned that you've violated the homogeneity of variance assumption. Doing so, of course, will reduce power, so you'll need a larger *n*. But you won't know that your data are heterogeneous until you've collected them! If you think that your data may exhibit heterogeneity, then the best strategy is to start out with a larger sample size than usual strategy might suggest.

• You could try to transform the data to eliminate the heterogeneity. For example, with reaction times, there is often a relationship between the size of scores and variability (groups with higher mean RTs have higher variances). The square root transformation K&W suggest is:

$$Y'_{ij} = \sqrt{Y_{ij} + \frac{1}{2}}$$

• K&W show an example of how this transformation might be applied to a data set (Table 7.3). Once data are analyzed, they suggest writing in terms of the original units (not the transformed units). My own advice would be to use such transformations sparingly and when there is a tradition in a literature for conducting analyses on such transformed data.

• You will also encounter transformations of data such as the logarithmic transformation and the arcsine transformation. Once again, I'd advise you to use such transformations sparingly.

• K&W also mention alternatives to the ANOVA that are less sensitive to heterogeneity of variance. They also consider the possibility that you may conduct comparisons that don't make use of the pooled error term.

7.5 Contrasts with Heterogeneous Variance

• "The standard tests of a single-*df* contrast ψ presented in Chapter 4 are severely compromised by heterogeneity of variance."
• We've been computing $F_{Comparison}$ using the $MS_{S/A}$ from the overall ANOVA. Doing so makes sense when we have little concern about heterogeneity of variance. However, when we've determined that our variances may not be equal (e.g., as a result of the Brown-Forsythe test), pooling over all the variances may not make as much sense. Imagine a situation in which an experiment with n = 20 and 5 conditions has group variances as seen below:

s_1^2	s_2^2	s_3^2	s_4^2	s_5^2
1	1	3	6	9

These group variances would lead to a pooled variance estimate ($MS_{S/A}$) of 4. However, note that when you're comparing groups 1 and 2, using the pooled error term (4) is relatively large, given the two group variances. If you used only the two groups involved in the comparison to estimate the population variance from which they were drawn, you'd get an estimate of 1. If you're comparing groups 4 and 5, the pooled error term may be smaller than it should be. To address the issue of heterogeneity of variance when computing comparisons, you may choose to use only the variances involved in the comparison when computing your test.
• The t statistic for a comparison would be:

$$t = \hat{\psi} \big/ s_{\hat{\psi}} \qquad (7.11)$$

However, when concerned about heterogeneity of variance, instead of using a pooled estimate of the standard error, you could use just the group variances involved in the comparison:

$$s_{\hat{\psi}} = \sqrt{\sum c_j^2 s_{M_j}^2} = \sqrt{\sum c_j^2 \left(s_j^2 \big/ n_j \right)} \qquad (7.12)$$

• For example, if you were using the example above and wanted to compare groups 1 and 4, your contrast set would be {1,0,0,-1,0}. Using just the two group variances would yield an estimated standard error of .59.
• I'm sure that it is obvious to you, but you can't simply choose to use the overall pooled variance estimate when it produces a smaller error term and then use the separate variance estimate when *that* produces the smaller error term!
• There's a further complication! You'll need to compute the df_{Error} using a beast of a formula:

$$df_{Error} = \frac{s_{\hat{\psi}}^4}{\sum \frac{c_j^4 s_{M_j}^4}{n_j - 1}} \tag{7.13}$$

The formula will typically produce a non-integer *df*, so you'll need to round down to the next smallest integer in order to find a value in the table.

• Let's pretend that K&W51 exhibits heterogeneity of variance (which it does not). Remember, the four groups have the following statistics associated with them:

	4 Hour	12 Hour	20 Hour	28 Hour
Sum (A)	106	151	230	247
Mean	26.5	37.75	57.5	61.75
Variance	51	119.6	240.3	190.9

[Note that F_{Max} = 240.3/51 = 4.7, which is greater than 3. Note, also, that the average variance is 150.46 = $MS_{S/A}$.]

• In comparing groups 1 and 3 (4Hr vs. 20Hr) in a simple pairwise comparison, we obtained MS_{Comp} = 1922 and F_{Comp} = 12.77 using the pooled variance estimate ($MS_{S/A}$), which would also be t_{Comp} = 3.57. If we were concerned about heterogeneity of variance, we might use an estimate of the error term from the two groups involved in the analysis.

$$t = \frac{31}{\sqrt{51/4 + 240.3/4}} = \frac{31}{8.53} = 3.63$$

Prove to yourself that this result is the same as computing F_{Comp} using the mean variance from the two groups.

Variance 4 Hr	Variance 20 Hr	Mean Variance	F_{Comp}	t_{Comp}

• Suppose that we wanted to compare group 1 with group 2 (4Hr vs. 12 Hr). Using the pooled variance approach we would get SS_{Comp} = 253.13, MS_{Comp} = 253.13, F_{Comp} = 253.15/150.46 = 1.68. Using the separate variance approach, we would divide 253.13 by 85.3 (instead of 150.46), giving F_{Comp} = 2.97, which is almost twice as large as we got with the pooled approach.

• Unfortunately, for complex comparisons the calculation of the error term isn't a simple average, so by "hand" you need to use formulas 7.11 and 7.12.

Using SPSS to Compute the Brown-Forsythe Test

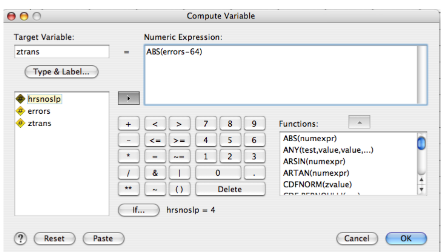
• First, to get the medians for each condition, use Analyze->Descriptive Statistics->Explore... Enter your IV (Factor list) and your scores on the DV (Dependent list). For K&W51, you'd get an output that looks (in part) like this:

Descriptives

	HRSNOSLP			Statistic	Std. Error
ERRORS	1	Mean		26.50	3.571
		95% Confidence Interval for Mean	Lower Bound	15.14	
			Upper Bound	37.86	
		5% Trimmed Mean		26.17	
		Median		23.50	
		Variance		51.000	
		Std. Deviation		7.141	
		Minimum		22	
		Maximum		37	
		Range		15	

You can see that the median for the first group in K&W51 is 23.5. The other medians are also displayed in the output.

• Next, you could use the Transform->Compute procedure to create a new variable *ztrans*. You need to use the *if*... button to compute *ztrans* separately for each of the four groups, with the different median for each group, as seen below on the left for the 4th group:



hrsnoslp	errors	ztrans
1	37	13.50
1	22	1.50
1	22	1.50
1	25	1.50
2	36	4.50
2	45	4.50
2	47	6.50
2	23	17.50
3	43	13.00
3	75	19.00
3	66	10.00
3	46	10.00
4	76	12.00
4	66	2.00
4	43	21.00
4	62	2.00

As a result of your computations, you'll now have a new column (*ztrans*), as seen above on the right. When you compute an ANOVA on the *ztrans* score (as your DV), you'll get a source table like the one below, where it's clear that there is no evidence of heterogeneity of variance.

ANOVA

ZTRANS

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	146.500	3	48.833	1.107	.384
Within Groups	529.500	12	44.125		
Total	676.000	15			

• Unfortunately, SPSS is not very sporting, and it builds in a correction for heterogeneity using the Brown-Forsythe procedure. To do so involves a simple click of the Browne-Forsythe button under Analyze->Compare Means->One-Way ANOVA, and you'll see this 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.

Once again, the computer provides results in which you invest little "sweat equity." What it's telling you is that your F is significant with a corrected p-value of .007. The original (uncorrected) F had a p-value of .005, which means that there was a tiny correction due to departures from homogeneity.

• Instead of the Brown-Forsythe approach, SPSS offers the Levene test for homogeneity of variance. That is, instead of correcting the p-value of the F to take deviations from homogeneity into consideration, the Levene test functions in a fashion like that proposed by K&W for the Brown-Forsythe test.

Test of Homogeneity of Variances

errors

Levene Statistic	df1	df2	Sig.
1.245	3	12	.337

In this case, the Levene test indicates that there is little concern about heterogeneity of variance. That, of course, is the same conclusion you'd reach had you computed the Brown-Forsythe test. Having the results of this test, you could then determine what approach you want to take to evaluating the F from your ANOVA.

• For comparisons, as we'd noted earlier, SPSS provides the *Contrasts* button. When SPSS computes contrasts, it does so both assuming homogeneity of variance (pooled error term) and not assuming homogeneity of variance (separate error terms, with modified *df* for the error term). Note the results for the simple and complex comparisons below:

Contrast Coefficients

Contrast		hrsdep			
		1	2	3	4
dimension0	1	1	0	0	-1
	2	1	1	0	-2

Contrast Tests

	Contrast			Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
errors	Assume equal variances	dimension2	1	-35.25	8.673	-4.064	12	.002
			2	-59.25	15.023	-3.944	12	.002
	Does not assume equal variances	dimension2	1	-35.25	7.777	-4.533	4.496	.008
			2	-59.25	15.283	-3.877	4.363	.015