Richard Gonzalez
Psych 613
Version 3.1 (Oct 2022)

LECTURE NOTES #5: Advanced Topics in ANOVA

Reading assignment

- Read MD chs 10, 11, 12, 13, & 14; G chs 12 & 13

  When reading the four repeated measures chapters in MD, concentrate on the multivariate approach (Chapters 13 and 14). Read the "traditional univariate approach" (Chapters 11 and 12) for background knowledge so that you can communicate with other people who aren't familiar with the modern multivariate approach.

Goals for Lecture Notes #5

- Discuss issues surrounding unequal sample sizes in factorial designs
- Introduce random and nested effects
- Introduce repeated measures factors

Major Warning: The statistical content in these lecture notes is not too difficult. It is mostly about (1) how to handle the decomposition of sum of squares when we have unequal sample sizes in a factorial design, (2) how to use the correct error term when we have random effects factors, and (3) how to make the correct assumptions and use the correct error term when we have repeated measures designs. To put it simply, this is a lecture notes about using the correct sum of squares under different applications that require a deviation from what we learned in LN2 through LN4. However, both SPSS and R will make you pull your hair out. I wish the programs would make implementing the content in this set of lecture notes easier. So take my advice from the first day of class—concentrate first on understanding the statistics, then focus on implementing it in SPSS or R. This set of lecture notes will be the worst part of the entire year in terms of having to cover so much trivial detail about SPSS and R. It is the programs that make life difficult in these lecture notes, not the statistical material. After this and through the end of next semester, syntax in both SPSS and R becomes as easy as in lecture notes 1, 2 and 3, even though the statistical concepts will become a more complicated. In these lectures notes, you will spend most of your time learning how to instruct the program to use the right denominator, or sometimes just giving up, computing your own F_{obs} and doing your own table lookup for F_{critical}, or compute your own t_{obs} and table lookup t_{critical}.

1. Factorial designs having cells with unequal sample size

We have been discussing the analysis of variance in terms of a decomposition of sums of squares (e.g., with a two-way design we had SST = SSA + SSB + SSAB + SSE). Such a perfect decomposition applies in general only when every cell has the same number of subjects. When the cells contain an unequal number of subjects, then such decomposition may no longer be unique. The reason is that contrasts will no longer be orthogonal so there is going to be redundancy in some of the sums of squares. The SSE will always be the same, so all approaches under consideration have the same error term. What differs across the approaches for dealing with unequal sample sizes is how they decompose the main effect terms. The different approaches do not disagree on the interaction term, but they disagree on the main effects.

There are different ways of handling the redundancy due to unequal sample sizes across cells. I will discuss the three most popular methods. We assume that the reason there are unequal sample sizes is because of either random chance or true differences in population sizes. For example, a few subjects failed to show up to the experiment, a few subjects had to be thrown out because they did not take the task seriously and gave what appear to be random responses, etc. This kind of information should be included in the *Subjects* section of the paper. Researchers should check whether there are particular cells that are losing more subjects than others—that might be informative. If the different cell sizes arose from differential levels of attrition or differential population sizes, then the methods I present are no longer valid. Also, if the researcher intended to have unequal sample sizes, such as in a representative sampling design where some groups are over sampled, then other methods are required that take into account sampling weights. If this applies to your research problems, I recommend taking a course on sampling after you complete the 613/614 sequence.

The problem of unequal sample sizes occurs when we want to make claims about main effects, i.e., when we want to collapse cells and look at marginal means. There are different ways to collapse the main effects, and each method can give a different answer. I reiterate, the interaction term and the error term do not involve collapsing because they say something about individual cells, so the various methods will agree on the interaction term and the error term; **the methods differ on how they handle the main effects**.

The best approach, in my opinion, is what SPSS calls the *unique approach*; in some versions of SPSS it is called the regression approach. This approach tests the relevant effect having taken into account all other effects. This is the default of the **ANOVA** and **MANOVA** commands (in some versions of SPSS, however, **ANOVA** has a different default). The trick of converting a factorial into a one-way arrangement in order to test contrasts presented at the end of Lecture Notes 4 leads to the identical result as the unique approach when there are unequal n. Also, the unique approach tests contrasts in a straightforward way. In R this is the default method in the lm() command but not the default method in other commands like in the anova() command. R uses the ”helpful” term of Type III sum of squares to refer to the unique (aka regression) approach.

A second method, the *hierarchical method*, only looks at the factor of interest and, when unequal cells are present, confounds all the other effects. Here, R uses the term Type I sum of squares to refer to this method.

A third method that some people view as a compromise between the unique and the hierarchical approaches. SPSS calls this third approach *experimental* and R calls it Type II sum of squares. The experimental approach enters, in order, all main effects, then all two-way interactions, then all three-ways, etc., in sequential models. This differs from the hierarchical method, which enters every term in the model one at a time rather than in conceptual chunks (main effects, two-way interactions, etc). The unique approach enters everything at once (all in one big chunk). I'll explain what "entering terms into a model" and in different order means later. When we cover regression techniques, we will come across a situation where the experimental method makes sense, though for our purposes right now the "experimental method" probably is the least preferred of the three methods.

Let's go to an example to illustrate the basic concepts. The following example is taken from Maxwell & Delaney (1990). Consider the starting salaries (in 1000s) of a small sample of men and women who either have or don't have a college degree.

Example illustrating potential misleading effects of unequal sample sizes. The mean for the 12 females is 22.33 and the mean for the 10 males is 22.1, suggesting that females have a slightly higher score. However, if gender comparisons are made within level of college degree category, then the males have a higher score. What is the right main effect for gender? 22.33 for females vs. 22.1 for males? 21 for females v. 23.5 for males?

	College	No College
female	24	15
	26	17
	25	20
	24	16
	27	
	24	
	27	
	23	
mean	25	17
male	25	19
	29	18
	27	21
		20
		21
		22
		19
mean	27	20

If someone ignored education and simply looked at the 12 women v. the 10 men, the conclusion would be that on average women ($\hat{Y}_w$ = 22.33) earn slightly more than men ($\hat{Y}_m$ = 22.1), likely not statistically significant with such a small difference and small sample size. However, if education status is accounted in the analysis (say, by doing the 1,1,-1,-1 contrast in a 2 × 2 factorial design), then the opposite conclusion is reached. The mean for the women is 21 (average of college and no college) and the mean for the men is 23.5. This could reach significance with the right sample size.[1] So we have a situation where two ways of analyzing the data could in principle lead to opposite conclusions not just in terms of significance but also in terms of the direction of the effect.

These two ways of looking at the data answer different questions. The first asks the question: Are males paid a higher starting salary than females? The second asks the question: Within an education status are males paid a higher starting salary than females? Note that in the second

[1] This discrepancy is related to a more general problem known as “Simpson’s Paradox.”

analysis equal weight is given across the four cells. Which method of analysis is right? The answer rests completely on the research question you want to ask. If you want to compare the women in the study to the men completely ignoring college status, then the hierarchical approach may be appropriate. If you want to conditionalize comparisons by educational level, then the regression method is appropriate. SPSS and R will do both of these analyses. There are multiple names for these two methods. For example, the hierarchical approach is also referred to as a weighted mean analysis and Type I sum of squares. The unique approach is also referred to as an unweighted means analysis or Type III sum of squares.

Yet another way of distinguishing the three methods is to look at the contrast weights that are implied by each method. The contrast weights will give us some more intuition about how these methods differ because if you know the weights you can figure out the research question implied by that contrast. Recall that the cell sample sizes in this example are NFC=8, NMC=3, NFNC=4, and NMNC=7 (here “NFC” stands for “number of females who went to college”, etc). There were 12 females and 10 males in the study. Here are the contrast weights for the main effect of male v. female implied by the three methods:

method	FC	FNC	MC	MNC
unique	1	1	-1	-1
hierarchical	NFC/NF	NFNC/NF	-NMC/NM	-NMNC/NM
	8/12	4/12	-3/10	-7/10
	.667	.333	-.3	-.7
experimental	NFC*NMC/NC	NFNC*NMNC/NNC	-NFC*NMC/NC	-NFNC*NMNC/NNC
	2.18	2.54	-2.18	-2.54

As Maxwell & Delaney note (p. C-14, note 18), the experimental contrast weight is twice the harmonic mean[2] of the number of males and females at each level of the college factor. This table highlights that the clearest method is the unique method (the contrast weights are 1s and -1s), unless one wants to have contrast weights depend on group size.

This discussion illustrates another worry researchers should have: are all the relevant variables included in the analysis? It appears that in the example, education level is important and has implications for how we interpret the sex difference, but we probably need to account for type of job. Are these starting salaries for the same type of job or are men and women in this sample getting different kinds of jobs. There are probably many other factors that should be considered as well. This highlights the role blocking factors (aka covariates) can play in an analysis.

Sometimes a researcher intentionally has cells with unequal sample sizes. For example, a researcher might want a sample to represent the ethnicity composition of a population. These

[2] For a definition of the harmonic mean see MD, page C-13, note 14.

situations sometimes arise in field studies and in surveys that are intended to have a representative sample. See Kirk's advanced ANOVA textbook for a discussion of how to handle designs with this sampling feature.

More detail on the different methods is given in Appendices 2 and 3.

2. Random Effects Model

So far this semester we have been talking about the fixed effects model, which applies when you have specific treatments and want to make inferences only to those treatments. For example, in the sleep deprivation study we selected 12 hrs, 24 hrs, 36 hrs, and 48 hrs. The researcher probably wants to compare only those four conditions.

The random effects model involves the situation where the treatments are sampled from a population, and the researcher wants to make inferences about some population of possible treatments. For example, a memory researcher may want to test recall of high frequency words compared to the recall of low frequency words. The researcher probably does not care about the particular words chosen. She probably wants to make inferences about the category of high frequency and the category of low frequency words; she doesn't care about the particular words used in the study. So, the key to deciding between fixed effects and random effects models is the type of inference you want to make.

A classic piece showing the error of using a fixed effects model when one should have done a random effects model is Clark (1973), The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335-359. A more recent treatment of these issues is by Raaijmakers (2003, *Canadian Journal of Experimental Psychology, 57*, 141-151).

Random effects models are becoming more common in psychology and other social sciences. It turns out they are the foundation of repeated measures ANOVA, are equivalent to some complicated models like latent growth curve modeling and structural equation modeling. They also allow a more natural bridge between classical and Bayesian statistical methods.

But what I just said about random effects being used when you want to generalize to the population of possible levels is more of a simplifying description. I now want to get into what is really going on. The key idea in a random effects model is that you not only take into account the random noise (i.e., ϵ), but also take into account the sampling "noise" that went into the selection of the levels of a factor. The idea being that if you replicate the study you would also replicate drawing the levels of the random factor so that variability due to sampling of factor levels needs to be taken into account. For example, Clark argued that words in a psycholinguistic or memory experiment should be treated as a random effect. Note that subjects can be considered a random effect factor—more on this later.

The main differences between how a random effects factor is treated as compared to a fixed effects factor is (1) in the interpretation of the terms in the structural model and (2) sometimes a change in what is used as the error term, or what to put into the denominator of the F test.

In the fixed effects structural model all the terms in the structural model are constants, except for ϵ that has a distribution (every subject has their "own" ϵ and the ϵs are assumed normally distributed. In the random effects structural model some terms in addition to ϵ may be given a distribution.

The random effects model adds two assumptions to the usual three ANOVA assumptions (observations are independent, homogeneity of variance, and normally distributed errors):

(a) Treatment effects are normally distributed with a mean of zero and a variance that is estimated from the data.

(b) All treatment effects are independent from each other and the errors.

The structural model for the two-way factorial design with both factors being random is

$$Y = \mu + \alpha_\sigma + \beta_\sigma + (\alpha\beta)_\sigma + \epsilon \qquad (5\text{-}1)$$

I denote random effect terms with a σ subscript to highlight that they are random and have a distribution, somewhat like a neon sign saying "I have variability so I'm random." The meaning is that each term (e.g., each level of factor α) is a random draw from a random variable that has some variance. It is an abuse of notation but it communicates the key concept.

random effects for oneway, rbd and lsd

The source tables for one-way analysis of variance, randomized block designs, and latin square designs are the same for both fixed effects models and random effects models when all factors are random effects. So, the *p* values are the same, but the interpretations are slightly different with different implications for a replication. A fixed effects model would be replicated in an identical format (except for subjects who are treated as a random effect). A random effects model for the manipulation(s) would replicate using new levels that were randomly selected. Further, in a random effects model one generalizes the results to the population of interest. You can see how these two different approaches lead to different (underlying, hypothetical) sampling distributions.

random effects for two-way

In a two-way analysis of variance with both factors treated as random effects the test for the interaction is exactly the same as with the fixed effects model (i.e., F = MSAB/MSE). However, the main effects are tested differently because the "error term" (i.e., the denominator for the F observed) for the main effect becomes the mean square term for the interaction. So, for example, when both factors are random, the main effect for random effects factor A would be tested as F = MSA/MSAB rather than F = MSA/MSE. The reason is that the main effect of A is confounded with the sampling of B (and vice versa). The expected mean square terms are given by:

A: $\sigma^2_\epsilon + \mathrm{n}\sigma^2_{\alpha\beta} + \mathrm{nb}\sigma^2_\alpha$

B: $\sigma^2_\epsilon + \mathrm{n}\sigma^2_{\alpha\beta} + \mathrm{na}\sigma^2_\beta$

AB: $\sigma^2_\epsilon + \mathrm{n}\sigma^2_{\alpha\beta}$

Error: σ^2_ϵ

The calculation of the source table is identical to the fixed effects case except for the last step when the F tests are calculated. Clearly, the correct denominator for the main effects should be the interaction because that cancels the "nuisance" terms. Recall our statement of the F test requiring the "nuisance" terms in the denominator (i.e., ther terms you care about being isolated from the terms you don't care about in the ratio of the F test). Ott (1988) presents a procedure for deriving expected mean terms in general situations, it isn't straightfoward so I'm not going to cover this year.

Whenever a new error term is used be sure to use the associated degrees of freedom when comparing the observed F to the critical F from the table. So when testing a main effect in a two-factor ANOVA with both factors treated as random effects, the main effect MS is divided by the MS interaction as shown above, and the "error" degrees of freedom will be the degrees of freedom associated with the interaction term (not the degrees of freedom associated with the MSE term). This can sometimes mean a major reduction in degrees of freedom for the error term if the interaction has associated degrees of freedom lower than that of the error term.

one factor fixed and one factor random

We now consider a two-way factorial where one factor (say, factor A) is treated as fixed and the second factor is treated as a random effect (factor B). The assumptions are as you'd expect by now. The fixed effect part has the usual three assumptions, the random effects part (including the interaction) has the additional assumptions listed above.

The structural model for a two-way factorial where only one factor is fixed is

$$Y = \mu + \alpha + \beta_\sigma + (\alpha\beta)_\sigma + \epsilon \qquad (5\text{-}2)$$

The expected mean square terms for this design (one factor treated as fixed, the other factor treated as random) are identical as the ones in the previous example for two random effects factors with one exception. The fixed effect A is influenced by the random effect B. However, since A is not assumed to be random there is no “sampling variability” of treatments on A. So, the expected mean square term for factor B does not contain the interaction term.

The expected mean square terms when one factor is a fixed effect and the second factor is a random effect are

A (fixed): $\sigma_\epsilon^2 + \mathrm{n}\sigma_{\alpha\beta}^2 + \frac{\mathrm{bn}\sum\alpha^2}{\mathrm{a\text{-}1}}$

B (random): $\sigma_\epsilon^2 + \mathrm{na}\sigma_\beta^2$

AB: $\sigma_\epsilon^2 + \mathrm{n}\sigma_{\alpha\beta}^2$

Error: σ_ϵ^2

The F for the main effect for factor B will use the MSE term in the denominator, whereas the F test for the main effect for factor A will use the MSAB term in the denominator. Remember the rule that the denominator must have the same terms the numerator has except the particular term being tested.

3. Example of a random effects design

Here are data from an experiment looking at the effectiveness of three different treatments (Maxwell & Delaney, 1990). The three treatments are rational-emotive therapy (RET), client-centered therapy (CCT), and behavior modification (BMOD). The effectiveness scores come from a standardized battery of tests–the greater the score the more “effective” the treatment.

I’m going to deviate from the description given in Maxwell and Delaney. Forty-five subjects were randomly assigned to each of the three treatments (15 subjects per cell). Three therapists were enlisted to conduct the treatments (each therapist administered one of each treatment). Later, we will revise this example (same data) and treat it as though there were nine different therapists.

First, we examine the one-way analysis of variance comparing the three treatments. I’ll omit all the assumption testing (boxplots and such) and only present some of the relevant means. [To save space I list these data in two columns.]

```
data list free / thrpy thpst score.                    value labels thrpy 1 'RET' 2 'CCT' 3 'BMOD'.
```

```
 begin data.
1 1 40
1 1 42
1 1 36
1 1 35
1 1 37
1 2 40
1 2 44
1 2 46
1 2 41
1 2 39
1 3 36
1 3 40
1 3 41
1 3 38
1 3 45
2 1 42
2 1 39
2 1 38
2 1 44
2 1 42
2 2 41
2 2 45
2 2 40
2 2 48
2 2 46
2 3 41
2 3 39
2 3 37
2 3 44
2 3 44
3 1 48
3 1 44
3 1 43
3 1 48
3 1 47
3 2 41
3 2 40
3 2 48
3 2 47
3 2 44
3 3 39
3 3 44
3 3 40
3 3 44
3 3 43
 end data .
```

```
means tables=score by thrpy by thpst .
```

```
Description of Subpopulations

Summaries of     SCORE
By levels of     THRPY
                 THPST


Variable        Value  Label                       Mean    Std Dev    Cases

For Entire Population                           42.0000     3.5227       45

THRPY            1.00  RET                      40.0000     3.3166       15
  THPST          1.00                           38.0000     2.9155        5
  THPST          2.00                           42.0000     2.9155        5
  THPST          3.00                           40.0000     3.3912        5

THRPY            2.00  CCT                      42.0000     3.1396       15
  THPST          1.00                           41.0000     2.4495        5
  THPST          2.00                           44.0000     3.3912        5
  THPST          3.00                           41.0000     3.0822        5

THRPY            3.00  BMOD                     44.0000     3.0938       15
  THPST          1.00                           46.0000     2.3452        5
  THPST          2.00                           44.0000     3.5355        5
  THPST          3.00                           42.0000     2.3452        5

  Total Cases = 45
```

Here I use the good old fixed effects, one-way ANOVA that ignores therapist to serve as a fixed effects oneway ANOVA

comparison to the more complicated models that I will present below.

```
manova score by thrpy(1,3)
    /design thrpy.

Tests of Significance for SCORE using UNIQUE sums of squares
Source of Variation          SS      DF        MS         F  Sig of F

WITHIN CELLS             426.00      42     10.14
THRPY                    120.00       2     60.00      5.92      .005
```

The conclusion from the omnibus test is that there is a difference between the three treatments. As with all omnibus tests, this conclusion is not very helpful. We would need to test contrasts or pairwise tests to learn more about the specific pattern. By looking at the means we see that RET $<$ CCT $<$ BMOD, but we don't know so far whether these individual comparisons between means are statistically significant.

Before we get to random and fixed effects I want to highlight a feature of the MANOVA command in SPSS. It is possible to get a test of significance for the grand mean (the μ term in the structural model). This will come in handy when we do repeated measures designs. I thought I'd show this to you now even though it won't be of much use until we get to repeated measures designs. All that needs to be done is to include the word "constant" in the design line and out comes the test of the grand mean against zero. Note that the error term and the thrpy term are identical to the previous source table.

```
manova score by thrpy(1,3)
    /design constant thrpy.

Tests of Significance for SCORE using UNIQUE sums of squares
Source of Variation          SS      DF        MS         F  Sig of F

WITHIN CELLS             426.00      42     10.14
CONSTANT               79380.00       1  79380.00   7826.20      .000
THRPY                    120.00       2     60.00      5.92      .005
```

Explanation: The sum of squares constant is the sum of squares that comes out of doing the grand mean (1, 1, 1) contrast on the three group means for therapy. The three group means are 40, 42, and 44. The value of Î is 126. Recall that the formula for SSC is

$$\frac{\hat{\text{I}}^2}{\sum \frac{a_i}{n_i}} \qquad (5\text{-}3)$$

The cell sample size n_i is 15 per group. Plug in the numbers and you should get sum of squares for this contrast equal to 79,380, just as reported in the source table above.

Let's take into account some of the structure of the design to reduce the error variance (i.e., the within sums of squares). We might be able to get more power by including therapist as a blocking factor. Because there is more than one subject per cell we don't have a traditional randomized block design. We do have a two-way factorial design. For our purposes the main effect for therapist and the interaction between therapist will both be components of the "blocking" factor. Here are the commands and source table for this two-way analysis of variance[3].

fixed effects, two-way ANOVA

Here is the factorial ANOVA with both factors treated as fixed.

```
manova score by thrpy(1,3) thpst(1,3)
   /design thrpy thpst thrpy by thpst.
```

```
Tests of Significance for SCORE using UNIQUE sums of squares
Source of Variation          SS      DF        MS         F  Sig of F

WITHIN CELLS             316.00      36      8.78
THRPY                    120.00       2     60.00      6.84      .003
THPST                     43.33       2     21.67      2.47      .099
THRPY BY THPST            66.67       4     16.67      1.90      .132
```

Again, we're probably not interested in the differences between therapists or the interaction between therapy and therapist so those two terms will be thought of as "blocking factors." [4]

Now let's treat both factors (therapist and therapy) as random effects. The interaction term will be identical to the interaction from the fixed-effects model because you use MSW as the error in both cases. Note the SPSS syntax: I can tell SPSS which error term to use for each term in the structural model.

The syntax defines an alias for the interaction term to be "1". The syntax requests that the main effect be tested against the error term 1 (i.e., the interaction term) and that the interaction term be tested against the usual MSE (called WITHIN in SPSS jargon). The resulting table contains all the correct F's and no additional computations are required.

random effects, two-way ANOVA

```
manova score by thrpy(1,3) thpst(1,3)
  /design thrpy vs 1, thpst vs 1, thrpy by thpst = 1 vs within.
```

[3] In the following example I consider the effects of specific factors as being fixed or random. Focus on the meaning the different formulations offer in terms of the research questions being addressed (and the underlying statistical models they imply).

[4] I don't mean to imply that therapists and the interaction between therapist and therapy are not important to study. It is just that the way I framed the present study suggests that those two terms should be treated as blocking terms. There are situations where someone would want to study differences between therapists and the interaction between therapist and therapy. But, that is not the question I have posed here.

```
Tests of Significance for SCORE using UNIQUE sums of squares
Source of Variation          SS      DF        MS         F  Sig of F

WITHIN CELLS             316.00      36      8.78
THRPY BY THPST (ERRO      66.67       4     16.67      1.90      .132
R 1)

Error 1                   66.67       4     16.67
THRPY                    120.00       2     60.00      3.60      .128
THPST                     43.33       2     21.67      1.30      .367
```

Note that the therapy main effect is no longer significant. If we believe the model that both therapist and therapy should be treated as random effects, then the correct error term to test the main effect of therapy is the MSAB (mean square for interaction).

Let's examine the following case. This is the case with Therapy fixed but Therapist random. In this example I defined the error term to be MSW for the therapist/therapy interaction, and the error term for therapy to be the mean square for interaction.

mixed two-way ANOVA

```
manova score by thrpy(1,3) thpst(1,3)
   /design thrpy vs 1, thpst vs within, thrpy by thpst = 1 vs within .
```

```
Tests of Significance for SCORE using UNIQUE sums of squares
Source of Variation          SS      DF        MS         F  Sig of F

WITHIN CELLS             316.00      36      8.78
THPST                     43.33       2     21.67      2.47      .099
THRPY BY THPST (ERRO      66.67       4     16.67      1.90      .132
R 1)

Error 1                   66.67       4     16.67
THRPY                    120.00       2     60.00      3.60      .128
```

Note that now therapy is not significant, likely due to the loss of degrees of freedom.

The UNIANOVA and GLM commands in SPSS unfortunately do something different with the error terms in a mixed design. By design, SPSS tests both main effects using the interaction term, which is not the convention. First, I present the output of UNIANOVA so you can see the difference in what SPSS does automatically and what you should get by manually specifying the correct error terms, then I present an email from SPSS explaining their take on the problem. If you use the menu system, you will get the wrong answer. Here is a case where you need to use syntax to get SPSS to divide by the right error term.

```
UNIANOVA score BY thrpy thpst
/RANDOM=thpst
/DESIGN=thrpy thpst thrpy*thpst.
```

```
Tests of Between-Subjects Effects
Dependent Variable:score
|-------------------------|----------------|--|-----------|--------|----|
|Source                   |Type III Sum    |df|Mean Square|F       |Sig.|
|                         |of Squares      |  |           |        |    |
|-------------|-----------|----------------|--|-----------|--------|----|
|Intercept    |Hypothesis|79380.000       |1 |79380.000  |3663.692|.000|
|             |----------|----------------|--|-----------|--------|----|
|             |Error     |43.333          |2 |21.667a    |        |    |
|-------------|----------|----------------|--|-----------|--------|----|
|thrpy        |Hypothesis|120.000         |2 |60.000     |3.600   |.128|
|             |----------|----------------|--|-----------|--------|----|
|             |Error     |66.667          |4 |16.667b    |        |    |
|-------------|----------|----------------|--|-----------|--------|----|
|thpst        |Hypothesis|43.333          |2 |21.667     |1.300   |.367|
|             |----------|----------------|--|-----------|--------|----|
|             |Error     |66.667          |4 |16.667b    |        |    |
|-------------|----------|----------------|--|-----------|--------|----|
|thrpy * thpst|Hypothesis|66.667          |4 |16.667     |1.899   |.132|
|             |----------|----------------|--|-----------|--------|----|
|             |Error     |316.000         |36|8.778c     |        |    |
|-------------|----------|----------------|--|-----------|--------|----|
a.  MS(thpst)
b.  MS(thrpy * thpst)
c.  MS(Error)
```

SPSS email

From: nichols@spss.com (David Nichols)
Subject: Expected mean squares and error terms in GLM
Date: 1996/11/05
Message-ID: <55oa9t$1tj@netsrv2.spss.com>#1/1
organization: SPSS, Inc.
newsgroups: comp.soft-sys.stat.spss

I've had a few questions from users about expected mean squares and error terms in GLM. In particular, with a two way design with A fixed and B random, many people are expecting to see the A term tested against A*B and B tested against the within cells term. In the model used by GLM, the interaction term is automatically assumed to be random, expected mean squares are calculated using Hartley's method of synthesis, and the results are not as many people are used to seeing. In this case, both A and B are tested against A*B. Here's some information that people may find useful.

It would appear that there's something of a split among statisticians in how to handle models with random effects. Quoting from page 12 of the SYSTAT DESIGN module documentation (1987):

There are two sets of distributional assumptions used to analyze a two factor mixed model, differing in the way interactions are handled. The first, used by SAS (1985, p. 469-470), can be traced to Mood (1950). Interaction terms are assumed to be a set of i.i.d. normal random variables. The second, used by DESIGN, is due to Anderson and Bancroft (1952). They impose the constraint that the interactions sum to zero over the levels of fixed factor within each level of the random factor.

According to Miller (1986, p. 144): "The matter was more or less resolved by Cornfield and

Tukey (1956)." Cornfield and Tukey derive expected mean squares under a finite population model and obtain results in agreement with Anderson and Bancroft.

On the other side, Searle (1971) states: "The model that leads to [Mood's results] is the one customarily used for unbalanced data."

Statisticians have divided themselves along the following lines:

Mood (1950, p. 344)	Anderson and Bancroft (1952)
Hartley and Searle (1969)	Cornfield and Tukey (1956)
Hocking (1985, p. 330)	Graybill (1961, p. 398)
Milliken and Johnson (1984)	Miller (1986, p. 144)
Searle (1971, sec. 9.7)	Scheffe (1959, p. 269)
SAS	Snedecor and Cochran (1967, p. 367)
SPSS GLM*	DESIGN

The references are:

Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. Annals of Mathematical Statistics, 27, 907-949.

Graybill, F. A. (1961). An introduction to linear statistical models (Vol. 1). New York: McGraw-Hill.

Hartley, H. O., & Searle, S. R. (1969). On interaction variance components in mixed models. Biometrics, 25, 573-576.

Hocking, R. R. (1985). The analysis of linear models. Monterey, CA: Brooks/Cole.

Miller, R. G., Jr. (1986). Beyond ANOVA, basics of applied statistics. New York: Wiley.

Milliken, G. A., & Johnson, D. E. (1984). Analysis of Messy Data, Volume 1: Designed Experiments. New York: Van Nostrand Reinhold.

Mood, A. M. (1950). Introduction to the theory of statistics. New York: McGraw-Hill.

Scheffe, H. (1959). The analysis of variance. New York: Wiley.

Searle, S. R. (1971). Linear models. New York: Wiley.

Snedecor, G. W., & Cochran, W. G. (1967). Statistical methods (6th ed.). Ames, IA: Iowa State University Press.

SPSS can be added to the left hand column. We're assuming i.i.d. normally normally distributed random variables for any interaction terms containing random factors.

In my view the column on the right is the correct one (in reference to the two groups of statisticians that Nichols lists in his email). I don't know why SPSS decided to go with the arguments made by those in the left hand column. As a heuristic, if Tukey, Scheffe, Cochran, Snedecor, Anderson and Cornfield are in the same statistical camp, then that is most likely the statistical camp one wants to be in (Johnson, Mood and Searle are solid statisticians but must have had a lapse in judgment).

Summary

We examined four different source tables from four different ways of framing the same study. It will be useful to summarize the four source tables in terms of their respective structural models. Pay careful attention to which terms are random and which terms are fixed.

one-way ANOVA; therapy fixed:	$Y = \mu + \alpha + \epsilon$
two-way ANOVA; therapy & therapist fixed:	$Y = \mu + \alpha + \beta + \alpha\beta + \epsilon$
two-way ANOVA; therapy & therapist random:	$Y = \mu + \alpha_\sigma + \beta_\sigma + (\alpha\beta)_\sigma + \epsilon$
two-way ANOVA; therapy fixed & therapist random:	$Y = \mu + \alpha + \beta_\sigma + (\alpha\beta)_\sigma + \epsilon$

4. Contrasts and post hoc tests for random effects

If you want to perform contrasts or post hoc tests, you use the same formulae discussed before for ANOVA with one exception. Use the correct error term in place of MSE. That is, whenever a formula (like the Tukey, Scheffe or contrast) calls for an MSE term, be sure to use the correct error term. For example, if I have one fixed and one random factor and I want to perform a contrast over the levels of the fixed factor, then I would substitute the MS interaction term for MSE (similarly for the Tukey and Scheffe). In addition, be sure to use the same degrees of freedom corresponding to the error term you use. So, if I use MS interaction rather than MSE, I would use the degrees of freedom associated with MS interaction rather than the degrees of freedom associated with MSE. This has implications for what critical value you look up in the table for the contrast, Tukey and Scheffe tests. Unfortunately, software doesn't always cooperate when performing contrasts, Tukey and Scheffe tests, so I recommend you double check the output by recomputing by hand using the equations in LN3, but substituting the correct error term and associated degrees of freedom.

5. Nesting

Sometimes one factor is nested within another factor (as opposed to being crossed like in the factorial design). An example using the one-way ANOVA will serve to illustrate this concept. Suppose there is one factor with three levels and subjects are assigned to only one treatment. One can think about this design as a randomized-block where subjects are the blocking factor, which is treated as a random-effect.

The structural model for this design is

$$Y \quad = \quad \mu + \alpha + \pi_\sigma/\alpha \tag{5-4}$$

The last term on the right means subjects (π) are nested within the α term. This term is equivalent to what we called ϵ. The two are synonymous, but the new notation highlights the idea that subjects are nested and treated as a random effect.

Let me give you another example that is a little more useful (adapted from Maxwell & Delaney). Suppose you want to look at the effects of therapists' gender. You conduct a study with three male therapists and three female therapists. Each therapist sees four different clients and you have a battery of measures for dependent variables. Here the independent variable of interest is gender. Note that therapist is nested within the levels of gender (obviously, it can't be crossed) and that client is nested within levels of therapist. Here gender is fixed but therapist is random because presumably we don't just care about these six therapists but want to generalize to some population of therapists.

For this simple design where there are an equal number of clients per therapist and an equal number of therapist per level of gender, the analysis is simple. To test for a gender difference, compute the mean for each therapist (over the four clients). You will have six means (three for male therapists and three for female therapists). Perform a two-sample *t* test using gender as a grouping variable on the six score (the six means). This test automatically treats gender as fixed, and automatically gives you the correct error term. Note that the question is about gender of therapist, so the error should be based on the mean of therapists (not on any of the lower levels).

The structural model for the therapist example is

$$Y = \mu + \alpha + \beta_\sigma/\alpha + \pi_\sigma/\beta_\sigma \tag{5-5}$$

where α is fixed, β is random and nested within α, π is random and nested within β, which is also random. Note that π_σ/β_σ plays the role of ϵ, but the former makes it clear that subjects is a nested factor; it is error but it comes from a specific term of subjects treated as a random effect nested in another factor that, in this example, is also a random effect. Maxwell and Delaney give additional cases and a more complete list of different combinations where the several factors are treated as fixed or random.

Another example.... Suppose you study different instructions given to juries. Juries would most likely be treated as nested within instruction and treated as a random effect. But juries are themselves made up of individuals, which are nested within levels of jury because jurors are randomly selected within a jury.

You should be aware of nesting and note that when nesting is present you may need to change your level of analysis as I did in the examples above. I didn't compare individual clients or individual jurors, but therapists and juries. Again, when all sample sizes are equal (e.g., in the therapist example there were three therapists within *both* levels of gender and four clients across *all* levels of therapist), then you can use the mean trick I mentioned above. Otherwise, the complications can get a ugly as you will see.

example of nested design

A numerical example with a nested factor: Suppose I treated the ongoing therapy & therapist example as a nested design. Let's change the design a little and consider the case where we have nine therapists, such that three therapists administered one of the three treatments. Thus, the nine therapists are nested with in the three therapies, and the 45 clients are nested within the 9 therapists such that each therapist is assigned 5 clients. In this design, subjects (π) is a random, nested effect within therapist (β), therapists is a random, nested effect within therapy, and therapy (α) is fixed. The structural model is:

$$Y = \mu + \alpha + \beta_\sigma/\alpha + \pi_\sigma/\beta_\sigma \tag{5-6}$$

SPSS nesting example

The SPSS command to do this model is below. The source table is partitioned into two parts, one for the fixed effect α and one for the random effect β nested in α and each of those two term have their own error term.

```
manova score by thrpy(1,3) thpst(1,3)
 /design thrpy vs 1, thpst within thrpy = 1 vs within.
```

The resulting output is

```
 Tests of Significance for SCORE using UNIQUE sums of squares
 Source of Variation          SS      DF        MS         F  Sig of F

 WITHIN CELLS             316.00      36      8.78
 THPST WITHIN THRPY       110.00       6     18.33      2.09      .079
   (ERROR 1)

 Error 1                  110.00       6     18.33
 THRPY                    120.00       2     60.00      3.27      .109
```

The therapy fixed effect factor uses the (random) therapist within therapy MS term. It has 6 degrees of freedom. Recall that the total sample size is 45 clients and there are 9 therapists. In this experimental design, the fixed effect therapy factor is tested against an error term based on the 9 therapists rather than the 45 clients.

As I suggested above, when there are an equal number of observations in each cluster, it is possible to perform this analysis by first aggregating over subject (i.e., compute the mean for each therapist, so each therapist was assigned 5 clients and compute the mean of the 5 clients) and then perform a one-way ANOVA comparing the means of the therapists across therapy. The statistical tests for this approach will be identical to the previous approach. Note that this way to think about a nested design aggregates over the client information and "acts" as though there are only 9 observations (i.e., the mean each therapist produced over their 5 clients).

	RET	CCT	BMOD
	38	41	46
	42	44	44
	40	41	42
cell mean	40	42	44

Simply perform a one-way ANOVA (a basic LN2 ANOVA) on these 9 numbers in order to test the main effect (or any contrast for that matter) over the 3 levels of therapy. Here is the resulting source table. The degrees of freedom are correct (2 and 6) and the F value is identical to the SPSS run above. However, the sum of squares terms "look" different. The reason the sum of squares terms are different is that the present analysis is tricked into thinking each observation is one subject, but in the previous analysis there were 5 observations per therapist. If you multiply the sum of squares between and sum of squares within each by 5 (the cell sample size), you get the same sum of squares as in the previous source table. The F statistics and p-values are identical.

	SS	df	MS	F
between	24	2	12	3.272727
within	22	6	3.6667	

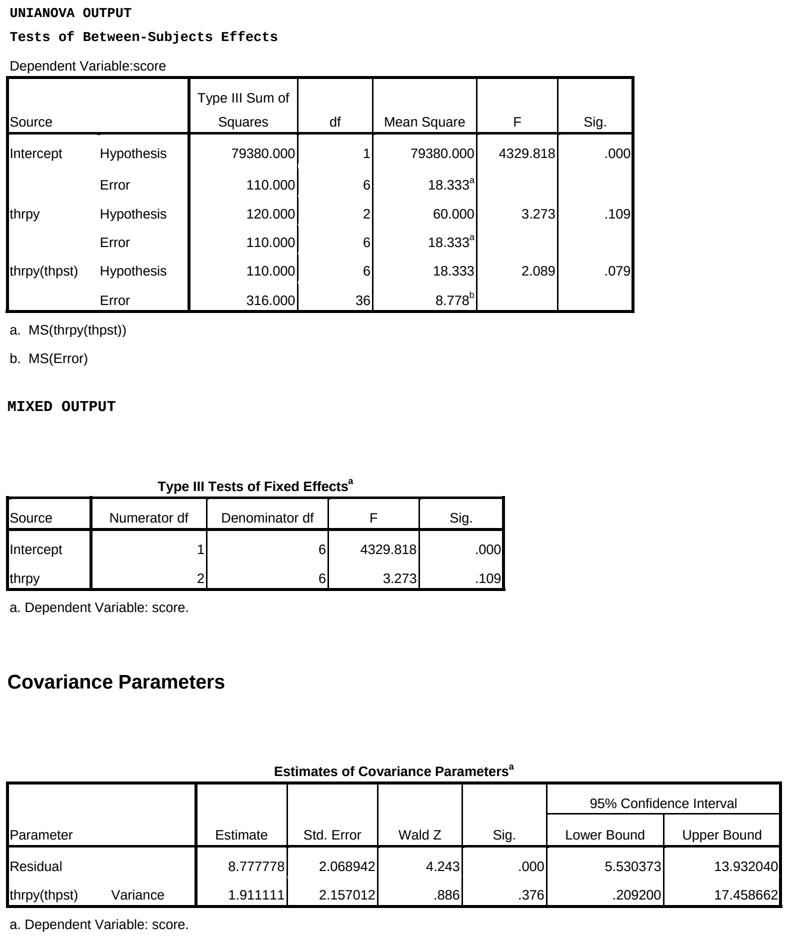
Again, this "trick" for computing nested ANOVAs works if you have equal sample sizes across groups. If you don't have equal sample sizes, then it is safer to use the model-based approach I gave above for nested factors using the SPSS syntax or analagous syntax in R as per the appendix.

There are other ways of implementing nested designs in SPSS. The UNIANOVA command and MIXED command have the nice feature of automatically figuring out the right error term. So, unlike the MANOVA command where you have to be specific about each error term, both UNIANOVA and MIXED "take out the guess work" (some people could say "take out the intelligent thinking"). Below is the syntax for both commands and some snips of the relevant output appear below.

```
UNIANOVA score by thrpy thpst
 /method=sstype(3)
 /intercept = include
 /print = descriptive
 /random = thpst
 /design = thrpy thpst(thrpy).
```

```
MIXED score by thrpy thpst
  /PRINT=SOLUTION TESTCOV
  /FIXED=thrpy
  /METHOD=REML
  /RANDOM=thpst(thrpy).
```

The UNIANOVA command defines the variable thpst as random and the design line has two terms: thryp and "thpst nested within thrpy" (note that the higher order term goes inside the parentheses–B(A) means B is nested within A). The MIXED command specifies the same model stating that the variable thrpy is fixed and the nested term "thpst(thrpy)" as random. When the nested factor is treated as random and when the design is balanced (i.e., has an equal



UNIANOVA OUTPUT

Tests of Between-Subjects Effects

Dependent Variable:score

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	79380.000	1	79380.000	4329.818	.000
	Error	110.000	6	18.333[a]		
thrpy	Hypothesis	120.000	2	60.000	3.273	.109
	Error	110.000	6	18.333[a]		
thrpy(thpst)	Hypothesis	110.000	6	18.333	2.089	.079
	Error	316.000	36	8.778[b]		

a. MS(thrpy(thpst))

b. MS(Error)

MIXED OUTPUT

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	6	4329.818	.000
thrpy	2	6	3.273	.109

a. Dependent Variable: score.

Covariance Parameters

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		8.777778	2.068942	4.243	.000	5.530373	13.932040
thrpy(thpst)	Variance	1.911111	2.157012	.886	.376	.209200	17.458662

a. Dependent Variable: score.

Figure 5-1: SPSS output for nested design using UNIANOVA and MIXED.

number of subjects at each level such as an equal number of patients for each therapist and an equal number of therapist in each therapy), the syntax for MANOVA, MIXED, UNIANOVA, and the trick of computing group means, all yield the identical test for the highest level (thrpy). Cool.

Unfortunately, things break down when there is an unequal number of observations, such as different number of patients per therapist. These different approaches do not yield the same result. The modern view is to take a special approach to such unequal cases, and this special approach is what is implemented in the MIXED command. So with unbalanced designs best to just stick with MIXED. The GLM and UNIANOVA commands use a "method of moments" approach, which is not well-behaved with unbalanced designs. Fancy programs such as HLM and MlWIN accomplish the same analyses as the SPSS MIXED command but those programs are more specialized and have many more features. Entire courses are taught on programs such as HLM. The term for this approach is "multilevel modeling", but the basic idea is that one can treat factors as fixed or random, and factors can be crossed or nested. In R, multilevel models can be done in either the nlme or lme4 packages, or within analogous Bayesian packages such as the brms package (an R wrapper for the Stan Bayesian programming language).

It turns out that repeated measures analyses can be accomplished nicely within this multilevel modeling approach and unequal sample sizes are allowed. This is particularly useful when there is missing data, such as when some subjects miss some of the repeated sessions. Under the multilevel modeling framework all data are analyzed, whereas under the traditional ANOVA repeated measures approach only subjects with complete cases can be analyzed (meaning one must throw away data from those subjects with incomplete data).

6. Repeated Measures

   It is a natural transition to move from a discussion of random/fixed effects and nested/crossed factors to the topic of repeated measures ANOVA. The basic idea of a repeated measures design is that rather than have independent groups of subjects randomly assigned to conditions, one collects more than one observation from each participant.

   Advantages

   (a) each subject serves as his/her own control

   (b) fewer subjects are needed

   (c) useful for examining change, learning, trends, etc.

(d) minimize error due to individual differences by treating subjects as a blocking factor

Disadvantages

(a) practice effects

(b) differential carry-over effects (e.g., recognition before recall)

(c) demand characteristics

(d) responses across trials need to be highly correlated for within subject designs to have high power

For additional pros and cons regarding within subject designs see Greenwald (1976), *Psychological Bulletin, 83*, 314-20.

7. Paired *t* Test: simplest repeated measures design

Each subject is measured twice on the same variable. For each subject you compute the difference between the score at time 1 and the score at time 2. Now you have, for each subject, one score—it's the difference between the two times. Let's denote the difference for subject i by $\mathrm{D_i}$. Perform the usual one sample *t* test on those difference scores. Thus,

$$t \quad = \quad \frac{\overline{\mathrm{D}}}{\mathrm{s_D}/\sqrt{\mathrm{n}}}, \tag{5-7}$$

where $\overline{\mathrm{D}}$ is the mean of the difference scores and $\mathrm{s_D}$ denotes the standard deviation of the difference scores. There are N - 1 degrees of freedom. This test is performed on difference scores. When testing more complicated repeated-measures designs we will exploit this idea of creating difference scores over the multiple time points.

Example of paired *t* test

before	after	D
12	16	4
11	9	-2
12	15	3
10	10	0
10	9	-1
9	12	3
7	9	2
6	10	4
4	9	5
4	3	-1

$$\sum \mathrm{D} = 17; \overline{\mathrm{D}} = 1.7; \sum \mathrm{D}^2 = 85$$

The mean before = 8.5 and the mean after = 10.2. Note that the difference between the two means (10.2 - 8.5) equals the mean of the difference scores 1.7.

The *t*-test is given by

$$\begin{aligned} t &= \frac{\overline{\mathrm{D}}}{\hat{\mathrm{s}}_{\mathrm{D}}/\sqrt{n}} \\ &= \frac{1.7}{2.49/\sqrt{10}} \\ &= 2.15 \end{aligned}$$

The critical *t*(9) = 2.26, so we fail to reject the null hypothesis.

The CI around the difference is given by

$$\begin{array}{ccc} \overline{\mathrm{D}} & \pm & t\,\frac{\hat{\mathrm{s}}_{\mathrm{D}}}{\sqrt{\mathrm{n}}} \\ 1.7 & \pm & (2.26)\frac{2.49}{\sqrt{10}} \\ 1.7 & \pm & 1.78 \end{array}$$

The interval is (-0.08, 3.48). It contains 0, so we fail to reject the null hypothesis.

Can you figure out how to construct a CI around the difference? The ingredients are in Equation 5-7. The answer is given in the example below (see box).

We can cast this test in terms of the hypothesis testing template we introduced in Lecture

Notes #1. As stated earlier the paired t-test is equivalent to taking difference scores of the two times and conducting a one sample t-test on the difference scores.

Hypothesis test template for the paired *t* test

Null Hypothesis

- H$_o$: population mean $D = 0$
- H$_a$: population mean $D \neq 0$ (two-sided test)

Structural Model and Test Statistic

The structural model is the same as the randomized blocking variable design treating subjects as a blocking factor that is a random effect.

The test statistic operates on the mean D and specifies its sampling distribution. Recall the general definition of a t distribution (lecture notes #1)

$$t \quad \sim \quad \frac{\text{estimate of population parameter}}{\text{estimated st. dev. of the sampling distribution}}$$

For the case of a paired t test we have

$$t_{\text{observed}} \quad = \quad \frac{\overline{D}}{s_D/\sqrt{n}}$$

Critical Test Value The critical *t* will be two tailed and based on N-1 degrees of freedom.

Statistical decision If the observed *t* computed from the raw data exceeds in absolute value terms the critical value t_{critical}, then we reject the null hypothesis. If the observed *t* value does not exceed the critical value t_{critical}, then we fail to reject the null hypothesis. In symbols, if $|t_{\text{observed}}| > t_{\text{critical}}$, then reject the null hypothesis, otherwise fail to reject.

The paired-*t* test is equivalent to doing a (1, -1) contrast on each subject to create a new variable that becomes the variable you analyze. However, contrasts over repeated measures are different than contrasts for between subjects designs. Take the paired-*t* test as a simple example. The (1, -1) contrast is defined over a subject's pair of observations (say, time 1 - time 2). The error term is defined on this difference score and not by pooling over cells (hence there is no equality of variance assumption over time).

Recall that the one-way between-subjects ANOVA decomposes SST into treatments (SSB) and error (SSW). The one-way repeated measures ANOVA performs a similar decomposition, but breaks the SSW into two independent "pieces": something that has to do with subjects and something that has to do with noise due to particular subject × treatment combinations. This is exactly the same decomposition we saw in the randomized-block design. The one-way repeated measures is equivalent to the randomized block design where *subjects* are treated as the blocking factor, which is also treated as random. The repeated-measures design allows one to eliminate the variability due to subjects because subjects is treated as a random effect blocking factor.

8. Structural Model and Expected Mean Square Terms for Repeated Measures ANOVA

The model for a one-way repeated measures ANOVA is

$$\mathrm{Y} = \mu + \alpha_{\mathrm{i}} + \pi_\sigma + \epsilon \tag{5-8}$$

where π_σ refers to the effect of subjects and is treated as a random effect. Instead of writing ϵ I could have written $(\alpha \times \pi)_\sigma$ because subjects is a random, crossed (not nested) with α because each subjects appears in every cell. For comparison, the two-sample t test (where subjects are randomly assigned to groups) subjects is also treated as a random effect but it is treated as a nested factor within the two levels of the experimental factor.

The expected mean square terms for the one-way repeated-measures ANOVA are

TIME: $\sigma^2_\epsilon + \frac{\mathrm{n}\sum\alpha^2}{\mathrm{T\text{-}1}}$
SUBJECT: $\sigma^2_\epsilon + \mathrm{T}\sigma^2_\pi$
ERROR: σ^2_ϵ

Everything generalizes in the obvious way for more complicated designs (i.e., more factors with repeated-measures and also designs that have both repeated-measures and between-subjects factors). But, detailed discussion of these issues will wait till next semester. Turns out that different ways of thinking about the error term lead to different structural models, different source tables, etc. The consensus among methodologists about running repeated measures ANOVA is to do everything in terms of contrasts and bypass the debate of how to handle error terms for the omnibus case.

Now I will build the structural model for a two-way ANOVA where one factor is between-subjects and the other factor is within-subjects. You can think of this design as having two structural models—one for the between-subjects part and one for the within-subjects part. The reason for needing two parts is that the subjects factor is nested with respect to the between factor but the subjects factor is crossed with respect to the within factor. Thus, we need to treat the two subparts of the design (the between subpart and the within subpart) differently.

Let me remind you of the structural models for the between-subject and within-subject one-way ANOVA. The structural model for the between-subjects one-way ANOVA is

$$Y = \mu + \alpha + \epsilon \tag{5-9}$$

I could have written π_σ/α instead of ϵ because subject is nested within α. The source table for this model has two lines (between groups and within groups). Recall that subject is treated as a random effects factor that is nested within treatment.

Next, recall the structural model for the one-way within-subjects ANOVA:

$$Y = \mu + \beta + \pi_\sigma + (\pi \times \beta)_\sigma \tag{5-10}$$

Here β is the fixed time factor (I'm using β instead of α to avoid confusion in the next paragraph). The source table for this model will have three lines: one for time (β), one for subject (π_σ), and one for error $(\pi \times \beta)_\sigma$. The error term is the last term: subjects crossed with β.

Finally, we combine the two structural models. A mixed ANOVA (both between and within factors) is simply a concatenation of a between ANOVA on one hand with a within ANOVA on the other. Literally sum the two structural models above (Equations 5-9 and 5-10). Be sure not to count the grand mean twice, and also include an interaction between α (a between subjects factor) and β (a within subjects factor). This results in the following structural model:

$$Y \quad = \quad \mu + \alpha + \beta + \alpha\beta + \pi_\sigma + (\pi \times \beta)_\sigma + \epsilon \tag{5-11}$$

where the ϵ comes from the between subjects model. Be sure you can account for all these terms.

Most computer programs print only five lines in the source table rather than the six that you would expect from the structural model in Equation 5-11. The five lines that are printed are usually organized in this manner:

1 α
2 ϵ

3 β
4 $\alpha\beta$
5 $(\pi \times \beta)_\sigma$

The first line is the between subjects factor and the second line is the error term that is used to test the between subjects factor. That forms the between-subjects portion of the design with its own error term.

The third and fourth lines are the main effect for time and the interaction between the two factors, respectively. The last line is the error term used for both β and $\alpha\beta$. I don't know why most programs omit from the source table the sixth possible term present in the structural model (π_σ). The calculations are correct (that is, the SS corresponding to that term is removed), but it simply is not printed in the source table by most programs.

9. Assumptions for the paired *t* test and repeated measures

A critical assumption is normality (actually, symmetry is the crucial property). If the distribution of the differences is skewed or there are outliers you may consider doing the nonparametric Wilcoxon Signed-Rank test or you could find a suitable transformation on the difference scores. One assumes that subjects are independent from each other but data from the same subject are allowed to be correlated. Equality of variances between time 1 and time 2 is *not* assumed under one version of repeated measures known as the multivariate approach.

But wait, there's more! The most critical assumption of a repeated measures ANOVA is lurking not far away. It involves the structure of the variance-covariance matrix as well as the equality of the various variance-covariance matrices for every between-subjects cell. (Sometimes we just call the variance-covariance matrix the "covariance matrix.") This critical assumption only enters the picture when you perform omnibus tests. This assumption is automatically satisfied when you do contrasts (or more generally, whenever you have a 1 degree of freedom test). This gives another justification for always doing contrasts. How do you examine the assumption on the variance-covariance matrix? Unfortunately, there are no nifty plots to look at. But there are two properties one can examine as indicators of the assumption. One property is called compound symmetry. It is a property that is sufficient for the assumption. The key idea of compound symmetry is equality of the correlation coefficients between all levels of a within-subjects factor. Indeed, compound symmetry[5] uses a pooled error term, which is the mean of the variance of each variable minus the mean covariance between all possible pairs ($\overline{\mathrm{V}} - \overline{\mathrm{C}}$).

compound symmetry

Let me give some more background and connect this to material we saw in the between-subjects case. When we have four groups, for example, the independence assumption means that we merely have to examine the equality of variance assumption because there are no correlations across groups. So in the case of four groups there are four variances that we will call V_1, V_2, V_3, and V_4. We can arrange these four variances along the diagonal in a 4 x 4 matrix where all the off diagonals are 0 because of independence (e.g., data from group 1 are not related to data in group 2, etc.).

[5] For the linear algebra buffs…(don't worry about this footnote now, we will cover this material next semester and it will make sense then). A simpler way of stating compound symmetry is to say that T - 1 of the T eigenvalues are identical. It is a strange matrix that has T - 1 identical eigenvalues, and this is indeed a strange matrix because it has the same value x in the diagonal and the same value y everywhere else. This result is related to the diagonalization of a matrix. When matrix A is real and symmetric (as variance-covariance matrices are), then the following holds:

$$Q^{-1}AQ \quad = \quad L$$

where Q is the orthonormal matrix of eigenvectors and L is the diagonal matrix of eigenvalues. A property of Q orthonormal is that the transpose of Q equals the inverse of Q. If Q includes a set of orthogonal T - 1 contrasts along with the grand mean contrast, then this equation leads to the diagonal of L having T - 1 identical values.

$$\begin{bmatrix} V_1 & 0 & 0 & 0 \\ 0 & V_2 & 0 & 0 \\ 0 & 0 & V_3 & 0 \\ 0 & 0 & 0 & V_4 \end{bmatrix}$$

In repeated measures, however, independence needs to be relaxed because data could be correlated over time. So, if we have four times instead of four groups, the variances and covariances can be arranged in a 4 x 4 matrix that involves time with the variances at each time along the diagonal and the (typically nonzero) covariances in the off diagonal (e.g., the covariance between time 1 and time 2 denoted as C_{12}, etc., for all possible pairs of times).

$$\begin{bmatrix} V_1 & C_{12} & C_{13} & C_{14} \\ C_{12} & V_2 & C_{23} & C_{24} \\ C_{13} & C_{23} & V_3 & C_{34} \\ C_{14} & C_{24} & C_{34} & V_4 \end{bmatrix} \tag{5-12}$$

This is a "free" matrix because all six covariances are free to vary and all four variances are free to vary. There are six covariances because the matrix is symmetric so the upper triangular half is equivalent to the lower triangular half. Compound symmetry imposes the constraint that the four variances are equal and the six covariances are equal, so the resulting 4 x 4 covariance matrix is

$$\begin{bmatrix} V & C & C & C \\ C & V & C & C \\ C & C & V & C \\ C & C & C & V \end{bmatrix} \tag{5-13}$$

In the R appendix subsection for repeated measures I show how different analyses lead to error terms based on either the free covariance matrix such as in 5-12 (called the *multivariate approach*) or the constrained covariance matrix such as in 5-13 (called repeated measures under the compound symmetry assumption). Basing error terms on either of these two matrices also has implications for how degrees of freedom are computed, as we will see in the subsequent examples.

sphericity

A more recent development is the less restrictive property of sphericity (which is a necessary and sufficient property with respect to the key assumption on the covariance matrix). The intuitive idea of sphericity is that the variance of the difference between any two levels of a within-subjects factor is a constant over all possible pairs of levels. That is, the following is assumed to be a constant for all pairs of variables i and j:

$$\sigma_{\mathrm{i}}^2 + \sigma_{\mathrm{j}}^2 - 2\sigma_{\mathrm{ij}} \tag{5-14}$$

Compound symmetry will automatically satisfy sphericity as you can see by the structure of the covariances in Matrix 5-13 (if all the Vs are the same and all the Cs are the same, then for all possible pairs Expression 5-14 will be a constant). But there are other patterns that could satisfy sphericity but not compound symmetry so that is why it is less restrictive.

There are some tests one can do to attach a *p* value to the degree of violation. However, as with all tests of significance involving assumptions, the logic does not make sense because with enough subjects you will always reject the assumption and the tests are very, very sensitive to violations from normality.

The reason such assumptions are made is because for an omnibus test one pools cell measures to get one estimate of the error variance. Thus, all time periods need to have the same variance and the same pairwise correlations with all other time periods to make the pooling for the use of omnibus tests interpretable. Again, if you limit your statistical tests to contrasts you automatically satisfy the sphericity assumption. So, ignore the uninformative omnibus tests and you will bypass the need to satisfy the sphericity assumption. Just stick with contrasts.

10. Remedial measures for repeated-measures: What to do when the assumptions are not met.

The first thing to realize is that if you want to test specific contrasts, then you have no problem because the assumption on the off-diagonal of the variance-covariance matrix will automatically be satisfied. This is because there are two covariance terms. Due to the fact that the variance-covariance matrix is *always* symmetric (the correlation of A and B equals the correlation of B and A), we know that the two covariances must equal each other. In general, whenever there is a test with one degree of freedom in the numerator, then the variance-covariance matrix for that test is symmetric. To see this think about a simple case such as the (1, -1) contrast. This is taking the difference of two variables. If you have factors with two levels (degrees of freedom for that factor will be 1 in the numerator) or are doing contrasts (contrasts always have one degree of freedom in the numerator), then the critical assumption of repeated measures analysis is satisfied.

Problems arise when you want to examine omnibus tests. For example, a researcher measures the number of hours a subjects spends watching television. The researcher is interested in age effects and observes the subjects at ages 10, 20, and 30 (a longitudinal design). Our solution, of course, will be to make focused comparisons (contrasts) across the time intervals, e.g., is there a linear trend over time (-1,0,1) or a quadratic trend (1, -2, 1) and forget about omnibus tests.

Suppose, however, that you are forced against your will to perform an omnibus test based on three or more time periods. That is, you must perform a test that will only tell you "the time periods differed somewhere" but without knowing where. The assumed structure of the variance-covariance matrix will probably be violated. What do you do? It turns out that transformations will not help you because of the complicated covariance matrix (there are proofs that transformations to "stabilize" a variance-covariance matrix do not exist).

- One strategy is to "play dumb," ignore the assumption on the variance-covariance matrix, and proceed with the usual repeated-measures omnibus tests. Many people I know use this strategy.
- A second strategy is to find a Welch-like adjustment for the violation of the assumption. There are two techniques that perform such an adjustment. Both of these procedures are in the spirit of Welch's *t*-test because they adjust the degrees of freedom. One version was derived by Greenhouse & Geisser (G&G), the other version was derived by Huyndt & Feldt (H&F). Typically, both yield very similar results. Years of study on these two approaches shows that the G&G approach tends to be slightly conservative (i.e., it tends to slightly overcorrect). The H&F approach has a transformation that reduces G&G's conservative bias. Therefore, most statisticians (except probably Greenhouse and Geisser) tend to favor the H&F approach.

  For one-way repeated measures ANOVA, the corrections are based on a measure developed by Box called ϵ, which indexes the discrepancy from the assumption on the covariance matrix. The measure ϵ ranges between $\frac{1}{T-1}$ and 1, where T is the number of time periods. The lower the ϵ the worse the assumption fit. If ϵ = 1, then the assumption fits perfectly.

  For completeness, here is Box's definition of ϵ:

  $$\frac{T^2(\overline{\sigma}_{ij} - \overline{\sigma})^2}{(T-1)(\sum\sum\sigma_{ij}^2 - 2T\sum\overline{\sigma}_{i.} + T^2\overline{\sigma^2}_{..})} \tag{5-15}$$

  where $\overline{\sigma^2}_{ii}$ is the average variance, $\overline{\sigma^2}_{..}$ is the average of all variances and covariances, $\overline{\sigma^2}_{i.}$ is the mean entry in row i of the covariance matrix, and T is the number of times each subjects is measured. Note that the Box ϵ is not the same as the error ϵ we have been using to denote residuals in the structural model.

  There are tests of significance for Box's ϵ but they are highly sensitive to violations of normality and one needs to be careful of rejecting very small differences in the presence of large sample size. I don't recommend using those tests.

  There are corrections in the spirit of Welch that use ϵ to adjust the degrees of freedom. The Greenhouse & Geisser and the Huyndt & Feldt corrections are two examples that adjust degrees of freedom based on ϵ. The Greenhouse and Geisser correction adjusts the degrees of freedom using Box's ϵ. Rather than use than base the F test on T-1 degrees of freedom for the numerator and (T-1)(N-1) for the denominator, the GG correction multiplies the two degrees of freedom by ϵ. The improvement proposed by HF involves a transformation of Box's ϵ to reduce bias.
- Finally, the third strategy is probably the most elegant. Derive a test that does not impute any structure to the variance-covariance matrix. This strategy is called the multivariate

analysis of variance. This is the procedure that blesses the SPSS command **MANOVA** with its name. The R appendix presents my attempt to search the R ecosystem for a proper multivariate ANOVA.

The multivariate analysis of variance will be covered in more detail next semester. Despite all the hand-waving and fancy mathematics that go along with MANOVA, the intuition is quite simple. The test finds a contrast over time periods (more generally, over dependent variables) that maximizes the F value. Of course, there are appropriate corrections for chance because one can always find a contrast that maximizes an F value. Further, the contrast itself can be interpreted in terms of the calculated weights. In other words, the multivariate analysis of variance hunts down the best set of orthogonal contrasts for a given within-subjects factor, and automatically corrects for the "fishing expedition" in a manner analogous to Scheffe.

11. More numerical examples and an illustration of a simple way to perform contrasts on designs having repeated measures

(a) One-way repeated measures

Let's consider a one-way repeated measures design with four levels (say, measurements over four different time periods). Here are some data from twelve subjects:

subject	time 1	time 2	time 3	time 4
1	92	95	96	98
2	120	121	121	123
3	112	111	111	109
4	95	96	98	99
5	114	112	110	109
6	99	100	99	98
7	124	125	127	126
8	106	107	106	107
9	100	98	95	94
10	108	110	112	115
11	112	115	116	118
12	102	102	101	101

The omnibus hypothesis that the four measurements yield the identical means is shown below. The SPSS output will be organized as follows: The test for the grand mean is first (labeled CONSTANT), then a test for sphericity, then some measures of epsilon (a measure of the departure from sphericity), then some tests for a multivariate analysis of variance approach that doesn't make the sphericity assumption, and finally the omnibus test. The title "AVERAGE" just means that the sphericity assumption is made and the error term (MSE) is estimated as the difference between the average of the variances and the average of the covariances. Also, I asked for significance tests for the HF and GG corrections (not all versions of SPSS print these by default).

We will use the **MANOVA** command. Two new subcommands are **WSDESIGN** and **WSFACTOR**.

```
data list free / id t1 t2 t3 t4.
begin data.
1 92 95 96 98
2 120 121 121 123
3 112 111 111 109
4 95 96 98 99
5 114 112 110 109
6 99 100 99 98
7 124 125 127 126
8 106 107 106  107
9 100 98 95 94
10 108 110 112 115
11 112 115 116 118
12 102 102 101 101
end data.

manova t1 t2 t3 t4
  /wsfactor time(4)
  /print signif(hf gg)
  /wsdesign time .
```

```
 Tests of Significance for T1 using UNIQUE sums of squares
 Source of Variation          SS      DF        MS         F  Sig of F

 WITHIN+RESIDUAL         4387.23      11    398.84
 CONSTANT              555775.52       1 555775.52   1393.48      .000

Tests involving 'TIME' Within-Subject Effect.


 Mauchly sphericity test, W =      .01905
 Chi-square approx. =           38.50691 with 5 D. F.
 Significance =                     .000

 Greenhouse-Geisser Epsilon =      .37577
 Huynh-Feldt Epsilon =             .38921
 Lower-bound Epsilon =             .33333

AVERAGED Tests of Significance that follow multivariate tests are equivalent to
univariate or split-plot or mixed-model approach to repeated measures.
Epsilons may be used to adjust d.f. for the AVERAGED results.

 EFFECT .. TIME
 Multivariate Tests of Significance (S = 1, M = 1/2, N = 3 1/2)

 Test Name         Value    Exact F Hypoth. DF   Error DF  Sig. of F

 Pillais          .27586    1.14287       3.00       9.00       .383
 Hotellings       .38096    1.14287       3.00       9.00       .383
 Wilks            .72414    1.14287       3.00       9.00       .383
 Roys             .27586
 Note.. F statistics are exact.
```

```
Tests involving 'TIME' Within-Subject Effect.

 AVERAGED Tests of Significance for T using UNIQUE sums of squares
 Source of Variation          SS      DF        MS         F  Sig of F

 WITHIN+RESIDUAL          123.02      33      3.73
    (Greenhouse-Geisser)           12.40
    (Huynh-Feldt)                  12.84
    (Lower bound)                  11.00
 TIME                       7.23       3      2.41       .65      .591
    (Greenhouse-Geisser)            1.13                 .65      .455
    (Huynh-Feldt)                   1.17                 .65      .459
    (Lower bound)                   1.00                 .65      .438
```

The measure ϵ ranges between $\frac{1}{T-1}$ and 1, where T is the number of time periods. The Box ϵs for these data are relatively small, suggesting some concern about violating the assumptions. If we want to test the omnibus tests, we can either look at the G-G or H-F tests, or we can dispense with the omnibus tests and just do contrasts as I show next. The rest of the tests (those printed as Pillais, Hotellings, Wilks, and Roys I'll defer to Lecture Notes 12 as they refer to multivariate tests and we have to get other concepts under our belt before those tests will make sense).

Suppose you had some planned contrasts. Then you avoid the above mess and can do the contrasts directly without needing an omnibus test and without making the sphericity assumption. Unfortunately, SPSS still prints out all that junk even though it is not necessary. The GG and HF corrections are not needed for the contrast tests. Recall that the sphericity assumption is automatically satisfied for contrasts (i.e., any testing having 1 degree of freedom in the numerator).

```
manova t1 t2 t3 t4
  /wsfactor time(4)
  /contrast(time) = special( 1 1 1 1
                    1 1 -1 -1
                    1 -1 1 -1
                    1 -1 -1 1)
  /print= parameters(estim)
  /wsdesign time .
```

```
 --- Individual univariate .9500 confidence intervals
 CONSTANT

  Parameter   Coeff.   Std. Err. t-Value     Sig. t   Lower -95%  CL- Upper

        1     215.21       5.76    37.33     .00000       202.52     227.90

Tests involving 'TIME' Within-Subject Effect.


 Mauchly sphericity test, W =      .01905
 Chi-square approx. =           38.50691 with 5 D. F.
 Significance =                      .000
 Greenhouse-Geisser Epsilon =      .37577
 Huynh-Feldt Epsilon =             .38921
 Lower-bound Epsilon =             .33333

AVERAGED Tests of Significance that follow multivariate tests are equivalent to
univariate or split-plot or mixed-model approach to repeated measures.
Epsilons may be used to adjust d.f. for the AVERAGED results.

 EFFECT .. TIME
 Multivariate Tests of Significance (S = 1, M = 1/2, N = 3 1/2)

 Test Name               Value   Exact F   Hypoth. DF  Error DF   Sig. of F

 Pillais                .27586   1.14287         3.00      9.00        .383
 Hotellings             .38096   1.14287         3.00      9.00        .383
 Wilks                  .72414   1.14287         3.00      9.00        .383
 Roys                   .27586
Note.. F statistics are exact.

Tests involving 'TIME' Within-Subject Effect.

 AVERAGED Tests of Significance for T using UNIQUE sums of squares
 Source of Variation          SS      DF        MS         F  Sig of F

 WITHIN CELLS             123.02      33      3.73
 TIME                       7.23       3      2.41       .65      .591
 Estimates for T2

[STATS FOR 95% CI OMITTED]

 TIME

  Parameter           Coeff.        Std. Err.         t-Value          Sig. t

        1      -.5416666667           .84041         -.64453          .53244

 Estimates for T3
 --- Individual univariate .9500 confidence intervals
 TIME

  Parameter           Coeff.        Std. Err.         t-Value          Sig. t

        1      -.5416666667           .44576        -1.21514          .24975

 Estimates for T4
 --- Individual univariate .9500 confidence intervals
 TIME

  Parameter           Coeff.        Std. Err.         t-Value          Sig. t

        1      -.1250000000           .16428         -.76089          .46272
```

You can also do this with GLM using the following syntax, note the MMATRIX subcommand that displays contrasts separated by semicolons

```
GLM t1 t2 t3 t4
  /wsfactor time 4
```

```
/mmatrix ”cont1” t1 1 t2 1 t3 -1 t4 -1;
        ”cont2” t1 1 t2 -1 t3 1 t4 -1;
        ”cont3” t1 1 t2 -1 t3 -1 t4 1
/wsdesign time .
```

Oh, a little detail about the MANOVA command. Apparently if a nonorthogonal set of contrasts is given in the contrast=special command, then SPSS decides that you don’t want those contrasts and instead computes something else. Whatever it does, the weird thing is that SPSS gives different results if you reorder the contrasts. Very ugly. If you want to use the MANOVA command with nonorthogonal contrasts in repeated measures designs, then use this version of the syntax instead.

```
manova t1 t2 t3 t4
 /transform = special( 1 1 1 1
   -1 0 1 0
  -1 1 0 0
  -1 0 0 1)
 /print transform
 /ANALYSIS=(T1 T2 T3 T4).
```

The key difference is that the WSFACTOR subcommand is not given, and instead the /TRANSFORM and the /ANALYSIS subcommand are used. The T1 T2 T3 T4 refers to the four contrasts listed in the special (they do not refer to Time 1 scores, Time 2 scores, etc). T1 refers to the grand mean contrast, T2 the 2nd contrast listed, etc. This MANOVA syntax gives the same result as the GLM command syntax, and like GLM, is not sensitive to the order of specifying the contrasts. Here is the GLM version of the command, which doesn’t mind nonorthogonal contrasts.

```
glm t1 t2 t3 t4
 /wsfactor time 4 special( 1 1 1 1
  -1 0 1 0
  -1 1 0 0
  -1 0 0 1).
```

Ugly SPSS stuff.

In my personal data analysis of repeated measures data I avoid all these hassles by doing a convenient shortcut that is very simple. A simple way to generate these tests without using the **MANOVA** command for the contrast values (or the **GLM** command) is to create new variables according to the contrast (e.g., the 1 1 -1 -1 contrast would be a new variable that has t1 + t2 - t3 - t4) and then do one sample *t* tests against zero for those new variables. Note that these contrasts are not the same as the contrasts used in the one-way between-subjects ANOVA, where contrast values applied to cell means consisting of different subjects. In the repeated-measures case the contrast values are used to create a new variable, then simple *t*-tests can be performed on these new variables[6].

```
compute cont1 = t1 + t2 - t3 - t4.
compute cont2 = t1 - t2 + t3 - t4.
compute cont3 = t1 - t2 - t3 + t4.

t-test /testval = 0 /variables cont1 cont2 cont3 .
```

```
Variable    Number                Standard   Standard
           of Cases     Mean    Deviation     Error
-----------------------------------------------------
CONT1
              12      -1.0833      5.823     1.681
              12        .0000       .000      .000
TEMP

(Difference) Standard   Standard   |       2-tail |    t    Degrees of  2-tail
     Mean   Deviation     Error    | Corr. Prob.  |  Value    Freedom    Prob.
-----------------------------------+-------------+---------------------------
   -1.0833      5.823     1.681    |  .      .    |   -.64      11       .532


Variable    Number                Standard   Standard
           of Cases     Mean    Deviation     Error
-----------------------------------------------------
CONT2
              12      -1.0833      3.088      .892
              12        .0000       .000      .000
TEMP

(Difference) Standard   Standard   |       2-tail |    t    Degrees of  2-tail
     Mean   Deviation     Error    | Corr. Prob.  |  Value    Freedom    Prob.
-----------------------------------+-------------+---------------------------
   -1.0833      3.088      .892    |  .      .    |  -1.22      11       .250



Variable    Number                Standard   Standard
           of Cases     Mean    Deviation     Error
-----------------------------------------------------
```

[6] There are several ways of performing a one sample *t* test in SPSS. I illustrate one method in the text, which is fairly straightforward. It involves creating a new variable using the contrast weights and performing a one sample t-test. Another equivalent method is to compute a new variable that has all the positive contrast weights, another variable that has all the negative contrast weights, and compare the two using the paired t test command in SPSS. Here is the command syntax for the first contrast in the example:

```
compute poscont1 = t1 + t2.
compute negcont1 = t3 + t4.
execute.
ttest pairs poscont1 negcont1.
```

Other ways of performing a paired t-test are through the menu system and through this syntax:

```
   ttest pairs = poscont1 with negcont1 (paired).
```

```
CONT3
               12       -.2500      1.138       .329
               12        .0000       .000       .000
TEMP

(Difference) Standard   Standard    |      2-tail |    t     Degrees of   2-tail
    Mean    Deviation     Error     | Corr. Prob. |  Value    Freedom     Prob.
-----------------------------------+-------------+---------------------------
   -.2500      1.138       .329     |  .      .   |  -.76       11        .463
```

These three contrasts are identical to those reported in the previous **MANOVA** output.

Alternate sets of contrasts over time

You are free to choose any set of orthogonal contrasts over time. Another natural set of contrasts would be the polynomial contrasts. For four times, this would test the linear, the quadratic and the cubic trends. Intuitively, this corresponds to the trend with no bend (linear), one bend (quadratic) and two bends (cubic); more generally, each term in the polynomial adds another possible bend to the model. With four times, the linear contrast is -3, -1, 1, 3. The point of this contrast is that the weights are equally spaced and equally increasing between time points. The quadratic contrast is 1, -1, -1, 1 (the two outer times have different weights than the two inner points, indicating one bend either U or inverted U shaped). The cubic is -1, 3, -3, 1; note how this contrast has two switches in sign and weight. Together, the linear, quadratic and cubic trends together can model many different possible trajectories consisting of a weighted combination of these three components. You can test the linear, quadratic and cubic contrasts over time by either computing new scores using the contrast weights and then one sample t tests, or you can use a repeated measures ANOVA program like manova in SPSS.

(b) Two-way repeated measures

The structural model for a factorial design with two repeated measures is ugly (but admittedly, very logical):

$$Y_{ijk} \quad = \quad \mu + \alpha_j + \beta_k + \pi_i + \alpha\beta_{jk} + \alpha\pi_{ki} + \beta\pi_{ki} + \epsilon_{ijk} \qquad \text{(5-16)}$$

where α and β are the two manipulated factors, and π is the factor representing subjects, which is also treated as a random effect.

Consider a simple example of a 2 $\times$ 2 repeated-measures on both factors design. This could come out of a pre/post test sequence administered in two different settings. Questions one might be interested are things such as “Collapsing over days, is there a difference between the post-tests and the pre-tests?”, “Collapsing over pre/post tests, is there a difference between the first session and the second session?”, “Looking only at the first session, is there a difference between the post-test and the pre-test?", etc. These questions translate easily into contrasts. For this example, the design is

	pre-test	post-test
session 1		
session 2		

and the contrasts for the above questions are (1 -1 1 -1), (1 1 -1 -1), and (1 -1 0 0), respectively.

Imagine the following data came from the above design (the same data set used to illustrate the one-way repeated-measures). Suppose the researcher wanted to test the main effect for session, the main effect for pre/post, and the interaction (as a specific example of possible orthogonal contrasts that can be tested).

	Session 1		Session 2	
subject	pre	post	pre	post
1	92	95	96	98
2	120	121	121	123
3	112	111	111	109
4	95	96	98	99
5	114	112	110	109
6	99	100	99	98
7	124	125	127	126
8	106	107	106	107
9	100	98	95	94
10	108	110	112	115
11	112	115	116	118
12	102	102	101	101

The contrast (1 -1 1 -1) can be tested by creating a new variable that is the sum of the two pre-tests minus the sum of the two post-tests. This can be done in SPSS with the **COMPUTE** command. The test of significance just uses that new variable and is a one-sample *t* test against the null hypothesis that the mean of that new variable (i.e., the contrast value is 0).

```
 data list free / id s1.pre s1.post s2.pre s2.post.
begin data.
1  92  95  96  98
2  120  121  121  123
3  112  111  111  109
4  95  96  98  99
```

```
5 114 112 110 109
6 99 100 99 98
7 124 125 127 126
8 106 107 106 107
9 100 98 95 94
10 108 110 112 115
11 112 115 116 118
12 102 102 101 101
end data.

compute cont1 = s1.pre + s2.pre - s1.post - s2.post.
compute cont2 = s1.pre + s1.post - s2.pre - s2.post.
compute cont3 = s1.pre - s1.post - s2.pre + s2.post .
```

```
 t-test
 /testval = 0
 /variables cont1 cont2 cont3.

Variable    Number               Standard   Standard
           of Cases     Mean    Deviation     Error
------------------------------------------------------
CONT1
               12      -1.0833      3.088      .892
               12        .0000       .000      .000
TEMP

(Difference) Standard   Standard   |      2-tail |    t    Degrees of  2-tail
     Mean   Deviation     Error    | Corr. Prob. |  Value    Freedom    Prob.
-----------------------------------+-------------+---------------------------
   -1.0833      3.088       .892    |  .      .   |  -1.22       11       .250


Variable    Number               Standard   Standard
           of Cases     Mean    Deviation     Error
------------------------------------------------------
CONT2
               12      -1.0833      5.823     1.681
               12        .0000       .000      .000
TEMP

(Difference) Standard   Standard   |      2-tail |    t    Degrees of  2-tail
     Mean   Deviation     Error    | Corr. Prob. |  Value    Freedom    Prob.
-----------------------------------+-------------+---------------------------
   -1.0833      5.823      1.681    |  .      .   |   -.64       11       .532


Variable    Number               Standard   Standard
           of Cases     Mean    Deviation     Error
------------------------------------------------------
CONT3
               12       -.2500      1.138      .329
               12        .0000       .000      .000
TEMP

(Difference) Standard   Standard   |      2-tail |    t    Degrees of  2-tail
     Mean   Deviation     Error    | Corr. Prob. |  Value    Freedom    Prob.
-----------------------------------+-------------+---------------------------
    -.2500      1.138       .329    |  .      .   |   -.76       11       .463
```

Note that these are the same results we saw in the case of the one-way repeated-measures. The reasons are the identical data set was used as well as the identical set of contrasts.

Identical results are found if one performs a "repeated-measures" analysis of variance. In SPSS this is accomplished with the **MANOVA** command. Note how I defined the factorial repeated measures ANOVA using the **MANOVA** command.

```
manova s1.pre s1.post s2.pre s2.post
 /wsfactor session(2), prepost(2)
 /wsdesign session prepost session by prepost.
```

```
 Tests of Significance for T1 using UNIQUE sums of squares
 Source of Variation          SS      DF        MS         F  Sig of F

 WITHIN CELLS            4387.23      11    398.84
 CONSTANT              555775.52       1 555775.52   1393.48      .000

Tests involving 'SESSION' Within-Subject Effect.

 Tests of Significance for T2 using UNIQUE sums of squares
 Source of Variation          SS      DF        MS         F  Sig of F

 WITHIN CELLS              93.23      11      8.48
 SESSION                    3.52       1      3.52       .42      .532

Tests involving 'PREPOST' Within-Subject Effect.

 Tests of Significance for T3 using UNIQUE sums of squares
 Source of Variation          SS      DF        MS         F  Sig of F

 WITHIN CELLS              26.23      11      2.38
 PREPOST                    3.52       1      3.52      1.48      .250

Tests involving 'SESSION BY PREPOST' Within-Subject Effect.

 Tests of Significance for T4 using UNIQUE sums of squares
 Source of Variation          SS      DF        MS         F  Sig of F

 WITHIN CELLS               3.56      11       .32
 SESSION BY PREPOST          .19       1       .19       .58      .463
```

The three *t* values presented earlier with the contrast formulation correspond to the F tests it this ANOVA (since these are 1 degree of freedom in numerator tests you square the *t*'s and they will equal the Fs).

(c) Designs with both between-subjects factors and repeated-measures factors

Now consider a "mixed-design" having one factor that is between-subjects and one factor that is within-subjects. An example comes from Ott (p 807). Suppose the researcher wanted to compare sequence 1 with sequence 2, period 1 with period 2, and test the interaction of sequence and period. One way to accomplish these tests is with the following command. Because all three tests are 1 df tests there are no “global” omnibus tests, and each F corresponds to a specific contrast.

		Period	
	Patient	1	2
Sequence 1	1	8.6	8.0
	2	7.5	7.1
	3	8.3	7.4
	4	8.4	7.3
	5	6.4	6.4
	6	6.9	6.8
	7	6.5	6.1
	8	6.0	5.7
Sequence 2	9	7.3	7.9
	10	7.5	7.6
	11	6.4	6.3
	12	6.8	7.5
	13	7.1	7.7
	14	8.2	8.6
	15	7.2	7.8
	16	6.7	6.9

```
data list free / sequence id period1 period2.
begin data.
end data.

manova period1 period2 by sequence(1,2)
 /wsfactor period(2)
 /wsdesign   period
 /design sequence .
```

```
 Source of Variation          SS      DF        MS         F  Sig of F

 WITHIN CELLS              15.86      14      1.13
 SEQUENCE                    .53       1       .53       .46      .507


Tests involving 'PERIOD' Within-Subject Effect.

 Tests of Significance for T2 using UNIQUE sums of squares
 Source of Variation          SS      DF        MS         F  Sig of F

 WITHIN CELLS                .79      14       .06
 PERIOD                      .02       1       .02       .27      .611
 SEQUENCE BY PERIOD         1.49       1      1.49     26.30      .000
```

There is an error term used for the between-subjects factor and a second error term used for any term involving the within-subjects factor (in this case, one of the main effects and the interaction). The sphericity and multivariate analysis of variance output is omitted. SPSS is "smart" enough not to print this information because with tests involving one degree of freedom in the numerator, the issue of sphericity is irrelevant.

Can this test be done with **COMPUTE** commands followed by one-sample *t* tests? I'll show you a very simple way of performing the mixed design without having to use the MANOVA command in SPSS. But one must perform between-subject ANOVAs on the difference scores and the sum scores, rather than two sample two-sample *t* tests. Let me illustrate.

```
compute persum = period2 + period1.
compute perdiff  = period2 - period1.

t-test /testval = 0 /variables perdiff .
```

```
Variable    Number                Standard   Standard
          of Cases     Mean    Deviation     Error
------------------------------------------------------
PERDIFF
               16      -.0438       .551      .138
               16       .0000       .000      .000
TEMP

(Difference) Standard   Standard   |      2-tail |    t    Degrees of  2-tail
     Mean   Deviation     Error    | Corr. Prob. |  Value    Freedom    Prob.
-----------------------------------+-------------+--------------------------
    -.0438       .551       .138    |  .      .   |   -.32      15       .755
```

```
ttest groups = sequence(1,2) / variables persum perdiff .
```

```
Variable         Number                Standard    Standard
               of Cases     Mean   Deviation      Error
------------------------------------------------------------
PERSUM
      GROUP 1       8      14.1750      1.750       .619
      GROUP 2       8      14.6875      1.212       .429

                | Pooled Variance Estimate | Separate Variance Estimate
                |                          |
    F   2-tail  |    t   Degrees of 2-tail |    t    Degrees of  2-tail
  Value  Prob.  |  Value    Freedom   Prob. |  Value    Freedom    Prob.
----------------+--------------------------+------------------------------
   2.08   .354  |   -.68     14       .507  |   -.68     12.46      .508
```

```
Variable         Number                Standard    Standard
               of Cases     Mean   Deviation      Error
------------------------------------------------------------
PERDIFF
      GROUP 1       8      -.4750       .377       .133
      GROUP 2       8       .3875       .290       .103
```

```
               | Pooled Variance Estimate | Separate Variance Estimate
               |                          |
   F   2-tail  |    t   Degrees of 2-tail |    t    Degrees of  2-tail
 Value  Prob.  |  Value   Freedom   Prob. |  Value    Freedom    Prob.
---------------+--------------------------+------------------------------
  1.69   .505  |  -5.13     14      .000  |  -5.13     13.14      .000
```

Two of the three contrasts are correct (compared to the previous output from the **MANOVA** command). What went wrong? Think about the degrees of freedom involved in a one-sample *t* test–the degrees of freedom are N - 1. Since there are 16 subjects, the **TTEST** command spits out 15 degrees of freedom for the test of persum. But, we know from the previous **MANOVA** output that the degrees of freedom should be 14. So, in this entire lecture this is the only example where there fails to be a correspondence between the contrast in the **MANOVA** output and the same contrasts tested through **COMPUTE** and **TTEST**.

What is the right way to do this analysis with difference scores? The right way to think about contrasts for mixed designs is in terms of two separate between-subjects analysis of variances. That is, one between-subjects analysis of variance is conducted on the variable persum (the sum over time) and a second between-subjects analysis is conducted on the variable perdiff (the difference over time).

```
manova persum by sequence(1,2)
  /design sequence .
```

```
Source of Variation          SS      DF        MS         F  Sig of F

WITHIN CELLS              31.72      14      2.27
SEQUENCE                   1.05       1      1.05       .46      .507
```

```
manova perdiff by sequence(1,2)
 /design constant sequence .
```

```
Source of Variation          SS      DF        MS         F  Sig of F

 WITHIN CELLS               1.58      14       .11
 CONSTANT                    .03       1       .03       .27      .611
 SEQUENCE                   2.98       1      2.98     26.30      .000
```

Note that for the first time in this course the CONSTANT term is interpretable. The grand mean for this second between-subjects analysis of variance is testing whether the difference is significantly different than zero. The test for SEQUENCE (on the perdiff variable) is identical to the interaction of SEQUENCE and PERIOD. The test for SEQUENCE in the first between-subjects analysis of variance corresponds to the main effect of SEQUENCE.

12. More complex example of a design having one between-subjects factor and one within-subjects factor, a 3 by 3 mixed design

    I'll use the MANOVA command. Suppose that you have a 3 × 3 design with one factor between and one factor within. That is, three groups with each individual measured three times. The **MANOVA** syntax is:

```
manova t1 t2 t3 by group(1,3)
 /wsfactor time(3)
 /contrast(time) = special( 1 1 1
                           1 -1 0
                           1 1 -2)
 /contrast(group) = special(1 1 1
                           1 -1 0
                           1 1 -2)
 /wsdesign time(1) time(2)
 /design group(1) group(2).
```

    The two new lines are *WSFACTOR* and *WSDESIGN*. The *WSFACTOR* call assigns a name (in this example, time) to the three dependent variables. The *WSDESIGN* subcommand gives the structural model for the within part of the design (note that I defined each contrast separately). The output will be printed in two sections. The first section will have the between part of the design (i.e., group contrast 1 and group contrast 2) tested against the MSE. The second section will have the within part, but that will also be broken up into subparts (one part for each contrast). The reason the within portion is divided into subparts is that each contrast on a within subjects factor uses its own error term (more on this later). So there will be a source table for time(1) with its own error term and also the interaction contrasts between time(1) and each of the two group contrasts. There will also be a source table for time(2), with its own source table, and it will include the interaction contrasts of time(2) with the two group contrasts. The elegance of writing *WSDESIGN* and *DESIGN* in the way shown above (i.e., by calling specific contrasts) is that NO omnibus tests are performed and we avoid the ugly mess of nasty additional assumptions. The time(1) SPSS notation may be confusing—it refers to the first contrast listed under the special command, not to the time 1 variable.

```
data list free / g t1 t2 t3.

begin data.
1 3 4 5
1 3 2 4
1 4 5 7
2 2 1 6
2 9 8 7
2 8 5 6
```

```
3 8 3 2
3 8 7 9
3 7 9 8
end data.

manova t1 t2 t3 by g(1,3)
 /wsfactor time(3)
 /contrast(time)= special(1 1 1
                          1 -1 0
                          1 1 -2)
 /contrast(g)=special(1 1 1
                      1 -1 0
                          1 1 -2)
 /wsdesign time(1) time(2)
 /design g(1) g(2).
```

The resulting three source tables are:

```
 Tests of Significance for T1 using UNIQUE sums of squares
 Source of Variation          SS      DF        MS         F  Sig of F

 WITHIN CELLS              74.00       6     12.33
 G(1)                      12.50       1     12.50      1.01      .353
 G(2)                      20.17       1     20.17      1.64      .248

Tests involving 'TIME(1)' Within-Subject Effect.

 Source of Variation          SS      DF        MS         F  Sig of F

 WITHIN CELLS              15.00       6      2.50
 TIME(1)                    3.56       1      3.56      1.42      .278
 G(1) BY TIME(1)            3.00       1      3.00      1.20      .315
 G(2) BY TIME(1)             .44       1       .44       .18      .688

Tests involving 'TIME(2)' Within-Subject Effect.

 Source of Variation          SS      DF        MS         F  Sig of F

 WITHIN CELLS              23.00       6      3.83
 TIME(2)                    2.67       1      2.67       .70      .436
 G(1) BY TIME(2)            1.00       1      1.00       .26      .628
 G(2) BY TIME(2)            5.33       1      5.33      1.39      .283
```

Notice that NO omnibus tests were printed. Yeah! That is because I did not ask for omnibus tests in the *DESIGN* and *WSDESIGN* subcommands. Instead, I asked for specific contrasts. If you want omnibus tests you are going to have to deal with the nasty repeated measures assumptions. Better to avoid omnibus tests altogether, and go directly to contrasts.

SPSS tidbit: you may find it useful to print the contrasts used by SPSS in the spirit of the /design(solution) subcommand that we used in the between-subjects case. For repeated measures factorial designs you will need to use /design(oneway); for mixed designs with both repeated and between factors, you should use /design(solution oneway).

Here is the **GLM** syntax:

```
GLM t1 t2 t3 by group
 /wsfactor time 3 special( 1 1 1
                           1 -1 0
                           1 1 -2)
 /contrast(group) = special(1 1 1
                            1 -1 0
                            1 1 -2)
 /wsdesign time
 /design group.
```

The key differences in the GLM syntax are that the special contrast is defined directly in the *WSFACTOR* line, the grouping variable doesn't need the number of levels specified, and *WSDESIGN* doesn't need to separate out each contrast because the output by default is separated by contrasts. The GLM subcommand also has nice features, such as /PLOT=PROFILE(time) to print means over time, and can be crossed with grouping variables as well. Also, the /PRINT parameters test(mmatrix) subcommand to print the tests on individual cell means as well as the contrasts actually used by SPSS.

For completeness I show the syntax for conducting the same analysis using the MIXED command in SPSS. This command will become useful when doing more complicated analyses such as hierarchical linear models (HLM). It also has the advantage over MANOVA or GLM that it allows missing data for repeated measures. Whereas MANOVA or GLM drop subjects with missing data from repeated measures analyses, the MIXED command makes use of all the available data and performs the correct test of significance.

Data need to be organized in long format for the MIXED command, that is, each row is one observation so if a subject is measured at three times that subject uses three rows. Here are the reorganized data for the previous example and the MIXED syntax. The output for the eight contrasts (two main effects for time, two main effects for group, and four interaction contrasts) is identical to that of MANOVA and GLM. I specify "unstructured" (UN) covariances, which means the multivariate version that is less restrictive, but one could specify compound symmetry (CS) to get the more traditional, assumption-laden repeated measures tests. I include columns d1 and d2 for something I may do in class at a later date; for now ignore columns d1 and d2. Note the last two columns: sub codes subject number and time codes which of the three measurements. Sometimes it is useful to sort the data and I provide sorting syntax after the MIXED command for completeness.

```
data list free /row dv gr d1 d2 sub time.
begin data
1   3  1  1  1   1 1
2   3  1  1  1   2 1
3   4  1  1  1   3 1
4   2  2  1  1   4 1
```

```
5   9  2  1  1   5 1
6   8  2  1  1   6 1
7   8  3  1  1   7 1
8   8  3  1  1   8 1
9   7  3  1  1   9 1
10  4  1 -1  1   1 2
11  2  1 -1  1   2 2
12  5  1 -1  1   3 2
13  1  2 -1  1   4 2
14  8  2 -1  1   5 2
15  5  2 -1  1   6 2
16  3  3 -1  1   7 2
17  7  3 -1  1   8 2
18  9  3 -1  1   9 2
19  5  1  0 -2   1 3
20  4  1  0 -2   2 3
21  7  1  0 -2   3 3
22  6  2  0 -2   4 3
23  7  2  0 -2   5 3
24  6  2  0 -2   6 3
25  2  3  0 -2   7 3
26  9  3  0 -2   8 3
27  8  3  0 -2   9 3
end data.

mixed dv by gr time
 /fixed =gr time gr*time |  SSTYPE(3)
 /method=REML
 /print = solution
 /repeated = time | subject(sub) covtype(un)
 /emmeans=tables(gr*time) compare(time)
 /test 'main eff 1-gr'  time .5 -.5 0 time*gr 1/6 -1/6 0 1/6 -1/6 0 1/6 -1/6 0
 /test 'main eff 2-gr'  time .5 .5 -1 time*gr 1/6 1/6 -1/3 1/6 1/6 -1/3 1/6 1/6 -1/3
 /test 'main eff 1-time' gr .5 -.5 0 time*gr 1/6 1/6 1/6 -1/6 -1/6 -1/6 0 0 0
 /test 'main eff 2-time' gr .5 .5 -1 time*gr 1/6 1/6 1/6 1/6 1/6 1/6 -1/3 -1/3 -1/3
 /test 'int1'   time*gr 1/2 -1/2 0 -1/2 1/2 0 0 0 0
 /test 'int2'   time*gr 1/2 1/2 -1 -1/2 -1/2 1 0 0 0
 /test 'int3'   time*gr 1/2 -1/2 0 1/2 -1/2 0 -1 1 0
 /test 'int4'   time*gr 1 1 -2 1 1 -2 -2 -2 4.
```

This works at producing the identical output as GLM and MANOVA. I don't completely understand how the /TEST subcommand works and how the contrast codes are specified. Basically, each /TEST subcommand is a separate contrast. For main effect contrasts, it is necessary to include information about the higher order interaction(s) that include that factor. Interaction contrasts though stand on their own and don't need lower order contrast information. The order of the interaction contrasts are T1T2T3 repeated for Group1, Group2 and Group3, yielding nine values. Good luck trying to understand the syntax for the contrasts in the /test subcommand; every time I think I have it, I realize I don't.

13. Multilevel model approach to repeated measures

A different way to represent a within-subject design is to nest time within the subjects factor, where subjects is treated as a random effect factor. Recall the usual structural model for the within-subjects design models each observation as an additive sum of the following terms

$$X_{ij} \quad = \quad \mu + \alpha_i + \pi_j + \epsilon_{ij} \tag{5-17}$$

where μ is the usual grand mean, α corresponds to the main effect of time (there is one α each each level of time), π represents the subject factor that is treated as a random effect because subjects are randomly selected from a population (there are as many πs as there are participants), and ϵ is the usual error term (there are as many ϵ terms as there are number of participants times number of observations over time). Both π and ϵ are random variables assumed to follow a normal distribution with mean 0 and variances σ_π and σ_ϵ, respectively. However, in within-subjects designs the variance σ_π represents a square matrix with as many rows and columns as there are times, with variances in the diagonal and covariances in the off-diagonal. The reason that the σ_ϵ term is not a matrix is because subjects are assumed to be independent from each other (so all the covariances between subjects are 0) and are assumed to have the same error distribution, which is the homogeneity of variance assumption.

There is another way to specify the within-subjects model. The multilevel approach takes the time factor as nested within the subject factor and writes a two-level structural model. The first level models the observed data as a function of a subject effect β_j, time and the residual as in

$$X_{ij} \quad = \quad \beta_j + \alpha_i + \epsilon_{ij} \tag{5-18}$$

There is also a second structural model for the β_j, which is written as

$$\beta_j \quad = \quad \mu + \pi_j \tag{5-19}$$

where π is a random effect parameter assumed to be sampled from a normal distribution with mean 0 and variance matrix σ_π. Note that if one substitutes the linear definition of β (Equation 5-19) into the first level Equation 5-18, one gets the identical structural model for the within-subjects design presented in Equation 5-17. So the multilevel model is not a different model as much as a different approach to handling repeated measures. The advantage of the multilevel approach comes in its generalization to other models, and a common framework for handling many different kinds of designs under one umbrella. For a review of multilevel models see Raudenbush & Bryk's *Hierarchical Linear Models* book.

One of the key benefits of the multilevel approach to repeated measures is that it handles missing data in an elegant way. Unlike ANOVA, which discards the entire subject from the dataset if there is at least one missing time point for that participant, the multilevel model makes use of all available data for each subject. Another benefit of the multilevel approach is that it can extend the standard repeated measures approach by having a random slope (and more generally random effects on higher order terms like the quadratic). The traditional approach to repeated measures only has a random intercept (the π term in the structural model).

The SPSS implementation of this approach uses the command MIXED. For example, here is a 2x2 design with repeated measures on both factors. We test the main effects and interactions through contrasts. The data are structured in a slightly different way than usual. Rather than putting all the subject's data in the same row, each time takes on its own row. So with 4 observations per subject, there will be 4 times the number of subjects rows in the data file.

```
MIXED data  BY time
  /FIXED = time
  /REPEATED = time | SUBJECT(subject) COVTYPE(UN)
  /TEST  'main effect 1'  time 1 1 -1 -1
  /TEST  'main effect 2'  time 1 -1 1 -1
  /TEST  'interaction'    time 1 -1 -1 1.
```

Data structure is important so let me reiterate. This syntax requires that data be entered in a different format. Rather than entering each subject's data in a single row, every observation is entered in a separate row and new variables are included that code which subject and which time for each observation. So if there are 12 participants and each participant provided four observations, then the data set will have 48=12*4 rows, a column of numbers 1 through 12 to indicate which subject each observation belongs, and a column of numbers 1 through 4 to indicate which time each observation belongs. This is a strange way to implement a repeated measures design for those who are well-versed in the ways of repeated measures where multiple times for the same person are entered in the same row. It is good to get in the habit of sorting the data file by subject as many programs require this sorting and yield inappropriate results if data aren't sorted.

The subcommand FIXED instructs SPSS to use time as a fixed effect, and the REPEATED subcommand sets up the structure where the time scores are nested within the subject factor. The covariance is assumed to be unstructured. This is the typical assumption that corresponds to the single degree of freedom tests that we presented earlier; there are different types of covariance structures that are possible. We refer the reader to the manual of their statistics package, which in the case of SPSS and most major programs, tend to have good documentation on the different covariance structures for the time factor that are possible within their commands for testing multilevel models. Note: instead of putting COVTYPE(UN) for unstructured covariance matrix, if you enter COVTYPE(CS) that yields the compound symmetry structure on the covariance matrix that we saw earlier.

The MIXED syntax specifies each comparison in separate TEST subcommands. This MIXED syntax produces the same output as the MANOVA and the set of one-sample *t*-tests we introduced earlier. There are several different ways of implementing this design within the MIXED command that yield the identical results.

missing data

A major benefit of the linear mixed model approach to repeated measures ANOVA (both in SPSS and as described in the Appendix for R) is that missing data is handled by using all the available data. The traditional approach to handling missing data in repeated measures ANOVA is to delete the entire subject if there that subject has any missing data. The linear mixed model approach uses all available data for each subject. If data are "missing at random" this strategy makes sense. However, if there are systematic patterns to missingness (e.g., subjects are particular times or in particular conditions) are more likely to have missing data, then more advanced methods to address missingness are needed. This is beyond the scope of the present course; UM offers an entire semester long course in Public Health/Biostatistics on missing data.

14. Simple Effects for Factorial Designs

Sometimes we want to know how factors differ at each level of another factor. For example, suppose one factor is dosage with three levels (high, med, low) and the other factor is sex. We may be interested in how dosage differs for males (essentially a one-way ANOVA on males only) and how dosage differs for females (essentially another one-way on females only). These kinds of focused tests can easily be done with contrasts, though in SPSS the syntax gets tricky if you use MANOVA.

There is another method that many people like to use. It essentially amounts to breaking up omnibus tests into smaller omnibus tests, and breaking those into even smaller omnibus tests, etc., until you are down to single degree of freedom tests. The technique is sometimes called "simple main effects" and "simple interactions". Well, by now you know exactly my reaction to this procedure. If you eventually want to test contrasts, why not go there in the first place. In any case, I thought it would be good to introduce you to this style of analysis because you may be asked to do this at some point. In some cases, this technique of finding simple main effects ends up being identical to doing contrasts.

Maxwell and Delaney have a good discussion of "simple effects" ideas pages 260-268. I'll use their example to illustrate these ideas further, as well as show some SPSS syntax. Here are the data they used (part appears in Table 7.5 on page 258, and part appears in Table 7.8 on page 263) and a traditional source table. I've also added an extra column of codes to convert the 2x3 factorial into a 6-level oneway ANOVA. Bio has two levels and drug has three levels.

```
data list free / bio drug dv group.

begin data
1 1 170 1
1 1 175 1
1 1 165 1
1 1 180 1
1 1 160 1
1 2 186 2
1 2 194 2
1 2 201 2
1 2 215 2
```

```
1 2 219 2
1 3 180 3
1 3 187 3
1 3 199 3
1 3 170 3
1 3 204 3
2 1 173 4
2 1 194 4
2 1 197 4
2 1 190 4
2 1 176 4
2 2 189 5
2 2 194 5
2 2 217 5
2 2 206 5
2 2 199 5
2 3 202 6
2 3 228 6
2 3 190 6
2 3 206 6
2 3 224 6
1 1 158 1
1 2 209 2
1 3 194 3
2 1 198 4
2 2 195 5
2 3 204 6
end data.

manova dv by bio(1,2) drug(1,3) .
```

```
Tests of Significance for DV using UNIQUE sums of squares
Source of Variation          SS      DF        MS         F  Sig of F

WITHIN CELLS            4098.00      30    136.60
BIO                     1296.00       1   1296.00      9.49      .004
DRUG                    4104.00       2   2052.00     15.02      .000
BIO BY DRUG             1152.00       2    576.00      4.22      .024

(Model)                 6552.00       5   1310.40      9.59      .000
(Total)                10650.00      35    304.29
```

Now, suppose you want to conduct the test between the two biofeedback conditions at EACH level of drug. SPSS can perform this in MANOVA as follows. The variable drug is entered as a main effect. Three main effects "bio within drug(?)" where the question mark is replaced with each level of the drug variable (WITHIN is an SPSS keyword). No interaction is entered. There is a sense in which this simple main effect confounds a main effect and an interaction.

```
manova dv by bio(1,2) drug(1,3)
 /design drug, bio within drug(1), bio within drug(2), bio within drug(3) .
```

```
Tests of Significance for DV using UNIQUE sums of squares
Source of Variation          SS      DF        MS         F  Sig of F

WITHIN+RESIDUAL         4098.00      30    136.60
DRUG                    4104.00       2   2052.00     15.02      .000
BIO WITHIN DRUG(1)      1200.00       1   1200.00      8.78      .006
```

```
 BIO WITHIN DRUG(2)         48.00       1     48.00       .35      .558
 BIO WITHIN DRUG(3)       1200.00       1   1200.00      8.78      .006

 (Model)                  6552.00       5   1310.40      9.59      .000
 (Total)                 10650.00      35    304.29
```

Compare these two source tables. Note that the drug effect is identical in both tables. However, look at the sums of squares for the three bio tests in the second table. Add them up and check that the result is the same as the sum of squares for both the main effect of bio and the interaction term (2448). This confounding of main effect and interaction is another reason why I don't like this "simple test" approach as a general rule (though I could imagine that in some cases this test could make sense–it all boils down to which contrasts you want to test as I show below). You should double check that this is identical to the test presented in Maxwell and Delaney on page 264 (e.g., the BIO WITHIN DRUG(1) yields an F = 8.78). Note that the numbers in parentheses after the word DRUG do not refer to contrast numbers as we have used them in the past (here we have not defined /contrast=special()) but instead refer to the level of the drug factor.

I think it is easier to see what is happening in this "simple effect" test if we do it through a contrast directly. I'll use ONEWAY for this and recode the six cells of the factorial into a single one-way ANOVA with six levels. Let's look at the test of bio feedback for the first drug (the BIO WITHIN DRUG(1) above). This is identical to the (1, 0, 0, -1, 0, 0) contrast because we are only comparing two cells and ignoring the rest. Here's the syntax and the output.

```
oneway dv by group(1,6)
 /contrast 1 0 0 -1 0 0 .
```

```
                                 Sum of         Mean            F      F
       Source           D.F.     Squares        Squares          Ratio  Prob.

Between Groups            5      6552.0000      1310.4000       9.5930  .0000
Within Groups            30      4098.0000       136.6000
Total                    35     10650.0000


 Contrast Coefficient Matrix

            Grp 1       Grp 3       Grp 5
                  Grp 2       Grp 4       Grp 6

Contrast  1   1.0    .0    .0  -1.0    .0    .0


                                          Pooled Variance Estimate
                     Value      S. Error       T Value    D.F.       T Prob.

Contrast  1       -20.0000        6.7478        -2.964    30.0          .006


                                        Separate Variance Estimate
                     Value       S. Error        T Value    D.F.        T Prob.

Contrast  1       -20.0000         5.6569         -3.536     9.5           .006
```

If we square the pooled t we see that it is identical to the F test we reported above ($-2.964^2 = 8.78$). Instead of all that crazy, complicated language about simple main effects you can equivalently conduct this analysis through contrasts, being very clear about which group is being compared to which. Of course, this direct connection between the two approaches occurred because the simple effects approach yielded a single degree of freedom test, which can always be put into a contrast. In general though, simple effects must be followed by simple effects, etc., until you get down to single degree of freedom tests. Another advantage of framing the problem as a one-way is that it gives you the advantage of using the Welch test in cases where the equality of variance assumption is suspect. If your research question requires that you test these two cells independent from the other four, then this is a sensible contrast.

The simple effect approach will do strange things when you have unequal sample sizes because of how it partitions the variance. The answers are much cleaner if you stick to contrasts. With more complicated factorials the "simple effects" approach becomes even uglier. For example, in a factorial with three levels, should one examine the 2-way interaction for each level of the 3rd factor? a factor for each cell in the two-way? test A at each level of B ignoring C? There are many possibilities. The only way to know for sure which is appropriate in your situation is to think about the research question you are asking, convert them into contrasts, and test the contrasts directly.

For completeness, I mention that these "simple effect" tests can also be done on within-subjects factors. For designs with two repeated-measures the syntax for /WSDESIGN is identical to the one shown above for /DESIGN. However, for designs where one factor is between and the other is fixed, SPSS introduces a slight change in syntax–the use of the keyword MWITHIN rather than WITHIN. To test the within factor W at each level of the between factor B use this syntax

```
/WSFACTOR W
/DESIGN MWITHIN B(1), MWITHIN B(2)
```

To test B at each level of W

```
/WSDESIGN MWITHIN W(1), MWITHIN W(2)
/DESIGN B
```

These tests are also identical to performing contrasts directly on the cells being compared.

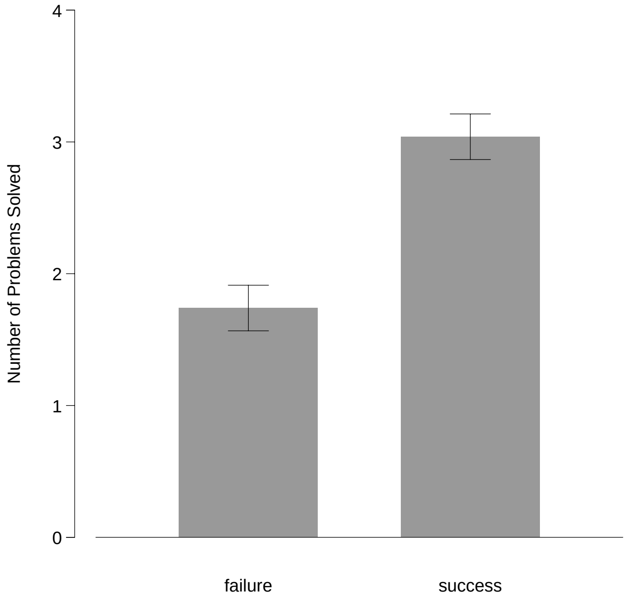
Appendix 1: Elements of a good results section

An example of a well-written results section appears in Lepper, Ross, and Lau (1986, *JPSP*, 50, 482-491).

A nice feature about this results section is that data, not inferential statistics, are emphasized. These authors state the result first and then provide a p-value as a punctuation mark. For example, "Subjects in the success conditions solved significantly more problems (M = 3.04) than did subjects in the failure conditions (M = 1.74), $F(1, 48) = 28.20$, $p < .0001$" (p. 485; note the typo in the article). Here the emphasis is on what was found—subjects in the success condition outperformed subjects in the failure condition.

Contrast this with the more common way results sections are written. Here is a typical sentence, which places emphasis on the inferential test rather than the actual result: "An ANOVA reveals a main effect of the success manipulation, $F(1, 48) = 28.20$, $p < .0001$." I've seen cases where some authors stop there. The reader is left wondering "but was the significant result in the correct direction? What were the differences between the groups? Was it a large or small effect?"

The Lepper, Ross, and Lau results section would have been even better had they shown a figure with the confidence intervals around the means. I calculated the intervals using the MSE calculated by working backwards from the means and F value that appeared in the article. I can't extract the individual standard deviations for each group (the authors should have provided that information), so I computed confidence intervals based on the pooled standard deviation (i.e., $\sqrt{\text{MSE}}$).



± one standard error (based on pooled standard deviation)

This paper also illustrates a relatively growing trend to report more information than just the means. We know that means alone can be deceiving. Two groups can have different means because of a couple of outliers in one cell or, more generally, the data violate the assumptions. Lepper, Ross, and Lau (1986) go beyond reporting the means and tell us the pattern of individual subjects. Here is an example:

> Only 3 of 26 failure-condition subjects solved as many as three problems (11 subjects solved only the one easy problem, 12 solved it and one other problem, 2 solved three problems, and 1 solved all four problems), whereas only 7 of 26 success-condition subjects solved fewer than three problems (2 subjects solved only the single easy problem and an additional 5 subjects solved only two problems, but 9 solved three problems and 10 solved all four problems). (p. 485)

This sentence could be more succinct, but you get the idea. The authors convey a sense of what happened in the study. Data analysis should go beyond just reporting means and *p*-values.

For more details on writing research papers see the APA Publication Manual, in particular the sections "Parts of a Manuscript" and "Writing Style" (pages 7-60 in the 4th Edition). I recommend that you take a look at that manual; it offers sound advice such as report all the relevant descriptive statistics so the reader can reproduce the inferential results you present. If you report means and standard deviations, the reader can reproduce any between-subjects factorial design and followup comparisons such as contrasts, Tukey, Scheffe, or Bonferroni.

While we're on the topic of writing, I bring up the issue of first sentences. Many people think that one must write science in some special, stuffy supercilious style. Here are the first sentences from a few classic papers in psychology. It may be that a necessary condition for a paper to become a classic is that the reader needs to stay awake long enough to finish reading the paper.

> My problem is that I have been persecuted by an integer. For seven years this number has followed me around, has intruded in my most private data, and has assaulted me from the pages of our most public journals. (G. Miller, *Psychological Review, 1956,* 63, 81-, famous 7 plus/minus 2 paper)

> Suppose someone shows a three-year old and a six-year old a red toy car covered by a green filter that makes the car look black, hands the car to the children to inspect, puts it behind the filter again, and asks "What color is this car? Is it red or is it black?" (Flavell, 1986, *American Psychologist, 41,* 418-)

> In the central nervous system the visual pathway from retina to striate cortex provides an opportunity to observe and compare single unit responses to several distinct levels. (Hubel & Wiesel, 1959, *J of Physiology, 148,* 574-)

And a good one from a classic economics piece

> Lassie died one night. Millions of viewers, not all of them children, grieved. At least, they shed tears. Except for the youngest, the mourners knew that Lassie didn't really exist.... Did they enjoy the episode? (Thomas Schelling, The Mind as a Consuming Organ)

While these are just examples, it is rare that a classic paper begins with the first sentence along the lines of "Smith (1982) found that .... But then Wesson (1984) failed to replicate the major result that .... So the present study attempts to solve this inconsistency." Blah. Few readers will finish that paper.

Appendix 2: Example of different methods for unequal n

Study looking at salary differences (in thousands) between men and women, and between college educated and non-college educated.

Here we'll make use of the ANOVA command in SPSS and its built-in facility for performing each of the three methods I discussed in class.

Data:

```
1 1 24
1 1 26
1 1 25
1 1 24
1 1 27
1 1 24
1 1 27
1 1 23
1 2 15
1 2 17
1 2 20
1 2 16
2 1 25
2 1 29
2 1 27
2 2 19
2 2 18
2 2 21
2 2 20
2 2 21
2 2 22
2 2 19
```

SPSS Commands

```
data list free / sex college dollars.

value labels sex 1 'female' 2 'male'.
value labels college 1 'college' 2 'nocollege'.

begin data.
end data.

anova dollars by sex(1,2) college(1,2).
```

```
                DOLLARS
           by   GENDER
                COLLEGE

                UNIQUE sums of squares
                All effects entered simultaneously


                                   Sum of                 Mean             Sig
Source of Variation               Squares      DF        Square        F    of F

   GENDER                          29.371       1        29.371   10.573  .004
   COLLEGE                        264.336       1       264.336   95.161  .000
   GENDER   COLLEGE                 1.175      1         1.175      .423  .524

Explained                         273.864      3        91.288    32.864  .000

Residual                           50.000     18         2.778

Total                             323.864     21        15.422
```

The default method for the **ANOVA** command is the unique method. This is the method that most directly answers the questions psychologists typically ask (within levels of college education is there a sex difference?). Not all versions of SPSS have the unique method as the default in **ANOVA**. To make sure that SPSS is performing the unique method, you can specify it explicitly with this subcommand:

```
  anova dollars by sex(1,2) college(1,2)
         /method = unique.
```

A different method is the hierarchical method. This is sometimes useful when one is asking about the differences between marginal means ignoring the other factors. The unique approach answers an analogous question but within each level of the other factors. Compare the main effect for sex in the unique approach (significant) with the main effect for sex in the hierarchical approach (nonsignificant) in the output below.

SPSS will do the hierarchical method if one specifies the /method = hierarchical subcommand. The source table is organized as though each term were entered one at time and all terms before it are still included. Note that for method=hierarchical the order the independent variables are listed in the first line of the **ANOVA** command makes a difference (order is irrelevant for method=unique). The reason order matters for method=hierarchical is that the hierarchical method identifies one factor as the most important factor, then the second factor, etc., whereas method=unique makes no such designation.

```
 anova dollars by sex(1,2) college(1,2)
    /method = hierarchical.
```

```
            * * *  A N A L Y S I S   O F   V A R I A N C E  * * *

                 DOLLARS
            by   GENDER
                 COLLEGE

                HIERARCHICAL sums of squares
                 Covariates entered FIRST

                                   Sum of               Mean             Sig
Source of Variation               Squares     DF        Square       F    of F

   GENDER                            .297      1          .297      .107  .747
   COLLEGE                        272.392      1       272.392    98.061  .000

   GENDER   COLLEGE                 1.175      1         1.175      .423  .524

Explained                         273.864      3        91.288    32.864  .000
```

```
Residual                          50.000     18         2.778

Total                            323.864     21        15.422
```

A third method is the experimental method. SPSS will do the experimental method if one specifies the /method = experimental subcommand. In older versions of SPSS the experimental method in the ANOVA command was known as the “sequential method” and was the default. To make matters more complicated, the MANOVA command calls the hierarchical method the “sequential” method. Keeping these terms straight is very difficult when different people use different versions of SPSS, and I have a history of confusing these terms in lecture. As long as we all write out the structural models (described in the next Appendix), then we will be okay.

```
anova dollars by sex(1,2) college(1,2)
   /method = experimental.
```

```
                     DOLLARS
                by   GENDER
                     COLLEGE

                     EXPERIMENTAL sums of squares
                     Covariates entered FIRST


                                  Sum of                Mean             Sig
Source of Variation              Squares     DF       Square        F    of F

   GENDER                         30.462      1       30.462   10.966   .004
   COLLEGE                       272.392      1      272.392   98.061   .000

   GENDER   COLLEGE                1.175      1        1.175     .423   .524

Explained                        273.864      3       91.288   32.864   .000

Residual                          50.000     18        2.778

Total                            323.864     21       15.422
```

Note how these three different methods partition the sum of squares for main effects differently (error and interaction are identical). Also, note how the sum of squares doesn’t always add up to sum of squares total, this is due to redundancy. The hierarchical method maintains the property that the two main effects, interaction and error sum of squares add up to sum of squares total, but the hierarchical method may not test the question we are asking. When we want to test contrasts that are not influenced by sample size, then we are forced to give up orthogonality in our design, live with redundancy, and have sum of squares that don’t add up.

Appendix 3: Illustrating the different methods of handling unbalanced designs: One more time

We'll use the same data looking at salaries using sex and college education. We will perform the different approaches directly with the design subcommand. The reason for showing you this is to highlight how the approaches differ by comparing the different structural models they imply. I'm going to present several runs of MANOVA, each one with a different structural model (but always the same MSE as the error term) so you can see where each of the pieces from the previous appendix come from. You might want to have Table 7.15 (Maxwell & Delaney, page 287-88) available while you work through this appendix; also read the summary section starting on page 286.

Note the use of the **/ERROR** subcommand. I'm telling the program I want the error term to be the within sums of squares. The first **/DESIGN** is the unique approach because all terms of the structural model are included simultaneously. That is, the structural model includes α, β, and $\alpha\beta$. The default is to use the within sums of squares as the error so I don't need to specify it.

The second **/DESIGN** subcommand is the beginning of the hierarchical approach. The main variable of interested is entered first. Note that the error term is the same error term used in the unique approach. This approach is testing the structural model Y = μ + α + ϵ but it is using the MSE term from the full model (all terms, like in the unique approach) rather than the MSE that falls out of the present structural model. So, in computing F, the method uses the numerator MS term from one structural model and a denominator MS term from a different structural model.

The third **/DESIGN** subcommand is step 2 of the hierarchical approach as well as step 1 of the experimental approach: all the main effects are entered simultaneously. Again, the error term is the same error term used in the unique approach. Thus, the structural model is Y = μ + α + β + ϵ. Each of the two terms (α and β are tested using the MSE term from the unique method, i.e., the structural model with all terms).

To check your understanding, you should compare the following results with the output in Appendix 2 that used the built in features of **ANOVA**.

Recall that **MANOVA** uses the unique approach to unequal sample sizes as the default. Much like the **ANOVA** command, the **MANOVA** also has a **/METHODS** subcommand where you can specify unique, hierarchical, or experimental manually. Below I will specify the models using the design line so you can see directly the structural models of each approach.

```
manova dollars by sex(1,2) college(1,2)
/error within
/design sex college sex by college
/design sex vs within
/design sex vs within college vs within .
```

UNIQUE APPROACH :

```
Tests of Significance for DOLLARS using UNIQUE sums of squares
Source of Variation          SS      DF        MS         F  Sig of F
```

```
 WITHIN CELLS              50.00      18      2.78
 GENDER                    29.37       1     29.37     10.57      .004
 COLLEGE                  264.34       1    264.34     95.16      .000
 GENDER BY COLLEGE          1.17       1      1.17       .42      .524
```

HIERARCHICAL APPROACH STEP 1 :

```
 Tests of Significance for DOLLARS using UNIQUE sums of squares
 Source of Variation          SS      DF        MS         F  Sig of F

 WITHIN CELLS              50.00      18      2.78
 GENDER                      .30       1       .30       .11      .747
```

HIERARCHICAL APPROACH STEP 2 AND EXPERIMENTAL APPROACH STEP 1 :

```
 Tests of Significance for DOLLARS using UNIQUE sums of squares
 Source of Variation          SS      DF        MS         F  Sig of F

 WITHIN CELLS              50.00      18      2.78
 GENDER                    30.46       1     30.46     10.97      .004
 COLLEGE                  272.39       1    272.39     98.06      .000
```

Compare these source tables with the ones presented in the previous Appendix. Such comparisons will help you figure out how these methods differ. The three source tables have the same error term even though they differ in the *DESIGN* subcommand because I specified that the within term should be used as the error.

Here is how to implement these three methods using the MANOVA command. I'll also ask MANOVA to print the contrasts it used so that may help you understand what is going on in these different methods. Take a look at the contrasts printed on page 5-5 so you can anticipate what this output should look like. The MANOVA command does an extra normalization of the contrast weights as I explain below so the numbers look slightly different but they are linearly related to the contrasts claimed on page 5-5.

First, the unique method. Note that the contrasts are as you would expect (e.g., the sex contrast is 1, 1, -1, -1). SPSS does a normalization so instead of 1s we see 1.08s, but the logic is the same. Recall in LN3 I showed that contrast t tests are unique up to scalar multiplication of contrast weights.

```
manova dollars by sex(1,2) college(1,2)
 /print design(solution)
 /method=unique
 /design .
```

```
 Solution Matrix for Between-Subjects Design
  1-GENDER  2-COLLEGE
 FACTOR                                    PARAMETER

    1   2            1          2          3          4

    1   1        -1.084     1.084     -1.084      1.084
    1   2        -1.084     1.084      1.084     -1.084
    2   1        -1.084    -1.084     -1.084     -1.084
    2   2        -1.084    -1.084      1.084      1.084
```

```
Tests of Significance for DOLLARS using UNIQUE sums of squares
Source of Variation          SS      DF        MS         F  Sig of F

WITHIN CELLS              50.00      18      2.78
GENDER                    29.37       1     29.37     10.57      .004
COLLEGE                  264.34       1    264.34     95.16      .000
GENDER BY COLLEGE          1.17       1      1.17       .42      .524

(Model)                  273.86       3     91.29     32.86      .000
(Total)                  323.86      21     15.42
```

Now, I'll do the hierarchical method, which in the MANOVA command is called the sequential method and order of entry is key: I need to list gender first to replicate the hierarchical method previously listed. The GENDER main effect is "hierarchical step 1" and the college effect is "hierarchical step 2".

The contrast corresponding to the main effect for Gender is based on sample size, as described in the lecture notes page 5-5.

```
manova dollars by sex(1,2) college(1,2)
 /print design(solution)
 /method=sequential
 /design.
```

```
Solution Matrix for Between-Subjects Design
 1-GENDER  2-COLLEGE
FACTOR                                    PARAMETER

   1   2             1         2         3         4

   1   1         1.706     1.557     1.221     1.084
   1   2          .853      .778    -1.221    -1.084
   2   1          .640     -.701      .962    -1.084
   2   2         1.492    -1.635     -.962     1.084

Tests of Significance for DOLLARS using SEQUENTIAL Sums of Squares
Source of Variation          SS      DF        MS         F  Sig of F

WITHIN CELLS              50.00      18      2.78
GENDER                      .30       1       .30       .11      .747
COLLEGE                  272.39       1    272.39     98.06      .000
GENDER BY COLLEGE          1.17       1      1.17       .42      .524

(Model)                  273.86       3     91.29     32.86      .000
(Total)                  323.86      21     15.42
```

The gender contrast listed here is 1.557, .7785, -.70065, -1.63485 (second column of the solution matrix listed above). This contrast is linearly related to the .6667, .333, -.3, -.7 hierarchical contrast claimed on page 5-5. That is, if you divide the contrast in this output by .4282 (which I got by dividing the weights of the two contrasts for one group, e.g., .667/.1557) you get back the hierarchical contrast on page 5-5, so the test of significance (t or F) will be identical for both of these contrast weights. Another way to connect these MANOVA contrast weights to those reported on page 5-5 is to convert them to proportions as in 1.557/(1.557+.778) = .667 (females with college) and -.701/(-.701-1.635) = -.3 (males with college).

Finally, to get the “experimental method” in MANOVA you need to rerun the same MANOVA command but reversing the order of the two independent variables (I don’t think MANOVA offers the experimental directly). You take the F for the COLLEGE variable from the source table previous to this one with GENDER entered first, and the F test for the GENDER variable from the source table below that contains COLLEGE entered first. That way, you get both F tests as though that variable were entered second (i.e., COLLEGE entered second in one analysis and GENDER entered second in the other analysis). Note how the contrasts weights for the main effects have changed.

```
manova dollars by college(1,2) sex(1,2)
/print design(solution)
/method=sequential
/design .
```

```
Solution Matrix for Between-Subjects Design
 1-COLLEGE  2-GENDER
FACTOR                                    PARAMETER

   1   2              1         2         3         4

   1   1          1.706     1.706     1.003     1.084
   1   2           .640      .640    -1.003    -1.084
   2   1           .853     -.853     1.171    -1.084
   2   2          1.492    -1.492    -1.171     1.084


Tests of Significance for DOLLARS using SEQUENTIAL Sums of Squares
Source of Variation          SS      DF        MS         F  Sig of F

WITHIN CELLS              50.00      18      2.78
COLLEGE                  242.23       1    242.23     87.20      .000
GENDER                    30.46       1     30.46     10.97      .004
COLLEGE BY GENDER          1.17       1      1.17       .42      .524

(Model)                  273.86       3     91.29     32.86      .000
(Total)                  323.86      21     15.42
```

This example shows how the order of entry of the factors can affect the main effect tests of significance for the hierarchical method—note the difference between the previous two source tables where the only difference in the syntax is the order of sex and college in the first line of the manova command.

The formulae for these contrast weights are based on sample size and are given in the lecture notes page 5-5. In the first “sequential” run (sex entered first), sex received the weights corresponding to the hierarchical approach and college received the weights corresponding to the experimental approach. In the second “sequential” run (college entered first), college received the weights corresponding to the hierarchical approach and sex received the weights corresponding to the experimental approach.

Appendix 4: R commands

Unequal sample sizes

If you want the regression approach in R, then use the lm() command. That is, run the ANOVA model as a regression model making sure to define the factors in the usual way so you have the correct degrees of freedom. For example,

```
# read in the salary data
data <- read.table("~/rich/Teach/Gradst~1/unixfiles/lectnotes/lect5/samplesize.
names(data) <- c("sex", "college", "dollars")
data[, 1] <- as.factor(data[, 1])
data[, 2] <- as.factor(data[, 2])

# manually specify contrasts on those factors this is
# important as we will see below
contrasts(data[, 1]) <- cbind(c(1, -1))
contrasts(data[, 2]) <- cbind(c(1, -1))

out.lm <- lm(dollars ~ sex * college, data = data)
summary(out.lm)


##
## Call:
## lm(formula = dollars ~ sex * college, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##     -2     -1      0      1      3
##
## Coefficients:
##                Estimate Std. Error
## (Intercept)     22.2500     0.3844
## sex1            -1.2500     0.3844
## college1         3.7500     0.3844
## sex1:college1    0.2500     0.3844
##                t value Pr(>|t|)
## (Intercept)     57.880  < 2e-16 ***
## sex1            -3.252  0.00443 **
## college1         9.755 1.31e-08 ***
## sex1:college1    0.650  0.52369
## ---
## Signif. codes:
```

```
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
##
## Residual standard error: 1.667 on 18 degrees of freedom
## Multiple R-squared:  0.8456,Adjusted R-squared:  0.8199
## F-statistic: 32.86 on 3 and 18 DF,  p-value: 1.629e-07
```

```
# square ts to show they are the same as the Fs for the
# regression method
summary(out.lm)$coef[, 3]^2
```

```
##   (Intercept)          sex1
##  3350.0844755    10.5734266
##      college1 sex1:college1
##    95.1608392     0.4229371
```

But the anova() command on the output gives the hierarchical method results. This is annoying that R produces two different approaches on the same output by running either summary() or anova().

```
anova(out.lm)
```

```
## Analysis of Variance Table
##
## Response: dollars
##              Df  Sum Sq Mean Sq F value
## sex           1   0.297   0.297  0.1069
## college       1 272.392 272.392 98.0611
## sex:college   1   1.175   1.175  0.4229
## Residuals    18  50.000   2.778
##                 Pr(>F)
## sex             0.7475
## college      1.038e-08 ***
## sex:college     0.5237
## Residuals
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
```

There are two ways to get the regression method source table from the lm output instead of the hierarchical method that the anova() command provides. One approach is with the drop1 command

and the other is with the Anova() command in the car package. R is case sensitive so anova is not the same as Anova.

```
# one approach
drop1(out.lm, . ~ ., test = "F")


## Single term deletions
##
## Model:
## dollars ~ sex * college
##             Df Sum of Sq     RSS    AIC
## <none>                    50.000 26.062
## sex          1    29.371  79.371 34.228
## college      1   264.336 314.336 64.507
## sex:college  1     1.175  51.175 24.573
##             F value    Pr(>F)
## <none>
## sex         10.5734  0.004429 **
## college     95.1608 1.306e-08 ***
## sex:college  0.4229  0.523690
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1


# another approach
library(car)
Anova(out.lm, type = "III")


## Anova Table (Type III tests)
##
## Response: dollars
##             Sum Sq Df   F value
## (Intercept) 9305.8  1 3350.0845
## sex           29.4  1   10.5734
## college      264.3  1   95.1608
## sex:college    1.2  1    0.4229
## Residuals     50.0 18
##                Pr(>F)
## (Intercept) < 2.2e-16 ***
## sex          0.004429 **
## college     1.306e-08 ***
## sex:college  0.523690
## Residuals
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
```

And if you want the experimental method, use `Type="II"` . Compare these to the output in SPSS so you confirm the exact source tables and how different they can be depending on which command and which option you specify in R.

```
Anova(out.lm, type = "II")


## Anova Table (Type II tests)
##
## Response: dollars
##                Sum Sq Df F value
## sex            30.462  1 10.9662
## college       272.392  1 98.0611
## sex:college     1.175  1  0.4229
## Residuals      50.000 18
##                  Pr(>F)
## sex            0.003881 **
## college       1.038e-08 ***
## sex:college    0.523690
## Residuals
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
```

If you want to use the aov() command be careful because a summary() on the aov output does not produce the regression method but rather the hierarchical method as we saw with the lm() command. You need to use the drop1 command or the Anova command with Type=”III” to print the source table corresponding to the regression method as in

```
out.aov <- aov(dollars ~ sex * college, data)
drop1(out.aov, . ~ ., test = "F")


## Single term deletions
##
## Model:
## dollars ~ sex * college
##               Df Sum of Sq     RSS    AIC
## <none>                      50.000 26.062
## sex            1    29.371  79.371 34.228
## college        1   264.336 314.336 64.507
## sex:college    1     1.175  51.175 24.573
##               F value    Pr(>F)
## <none>
## sex           10.5734  0.004429 **
## college       95.1608 1.306e-08 ***
## sex:college    0.4229  0.523690
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1


# Anova output omitted Anova(out.aov, type='III')
```

What happens if you don't manually define the contrasts. For example, if you copy someone's syntax from a blog post and didn't realize there were two earlier contrast commands (or the blog poster didn't realize the importance of those two contrast definitions). I'll show by re-reading the data from scratch and relying on the default contrasts R provides, which are dummy codes.

```
data <- read.table("samplesize.dat")
names(data) <- c("sex", "college", "dollars")
data[, 1] <- as.factor(data[, 1])
data[, 2] <- as.factor(data[, 2])

# let's see what contrasts R puts in by default these are 0
# and 1, called dummy codes we'll see in LN8 these are
# completely incorrect in the case of a factorial model
# with interactions
contrasts(data$sex)


##   2
## 1 0
## 2 1


contrasts(data$college)


##   2
## 1 0
## 2 1
```

```
out.lm <- lm(dollars ~ sex * college, data = data)
summary(out.lm)


##
## Call:
## lm(formula = dollars ~ sex * college, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##     -2     -1      0      1      3
##
## Coefficients:
##                Estimate Std. Error
## (Intercept)     25.0000     0.5893
## sex2             2.0000     1.1283
## college2        -8.0000     1.0206
## sex2:college2    1.0000     1.5377
##                t value Pr(>|t|)
## (Intercept)     42.426  < 2e-16 ***
## sex2             1.773   0.0932 .
## college2        -7.838 3.27e-07 ***
## sex2:college2    0.650   0.5237
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
##
## Residual standard error: 1.667 on 18 degrees of freedom
## Multiple R-squared:  0.8456,Adjusted R-squared:  0.8199
## F-statistic: 32.86 on 3 and 18 DF,  p-value: 1.629e-07


# square ts to get Fs none of these Fs are the same as the
# ones we saw before
summary(out.lm)$coef[, 3]^2


##   (Intercept)          sex2
##  1800.0000000     3.1418182
##      college2 sex2:college2
##    61.4400000     0.4229371
```

The lm() output is equal to one of the group means rather than the grand mean as we have been seeing all along.

Bottom line: be careful with R's default contrast definitions because it uses dummy codes and those do not work correctly with unbalanced designs. If you don't define your own contrasts based on specific hypotheses, you should at least define what are called effect codes using the contr.sum() command, which takes the number of levels as an argument, to avoid dummy codes. Using effect codes as contrasts will then give the same results as the regression approach.

```
contrasts(data$sex) <- contr.sum(2)
contrasts(data$college) <- contr.sum(2)
out.aov <- aov(dollars ~ sex * college, data)
drop1(out.aov, . ~ ., test = "F")


## Single term deletions
##
## Model:
## dollars ~ sex * college
##              Df Sum of Sq     RSS    AIC
## <none>                     50.000 26.062
## sex           1    29.371  79.371 34.228
## college       1   264.336 314.336 64.507
## sex:college   1     1.175  51.175 24.573
##              F value    Pr(>F)
## <none>
## sex          10.5734  0.004429 **
## college      95.1608 1.306e-08 ***
## sex:college   0.4229  0.523690
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1


# or could equivalently use the Anova command with
# type='III'
```

The reason this works is because the contr.sum command gives effect coding, which is a particular set of nonorthogonal contrasts where one group is a defined as the reference group and each of the other groups is then compared to the reference group. An example for a factor with 4 levels is the following with one level becoming the reference group (always assigned -1) and each of the three contrasts compare a different group (assigned 1) to the reference group[7]. We will discuss effects coding in more detail in LN7 and LN8.

[7] If you want the intercept in the summary(lm()) command, with properly defined contrasts on the factors, to correspond to the mean of the data (aka the weighted mean), then use the trick in Lecture Notes 4 to recode the factorial into a one-way ANOVA with as many groups as there are cells, apply effects contrasts on those groups but replace the -1 weights with the ratio of the sample size of the group that has the positive 1 over the size of the group that has negative 1. For example, if the four groups have sizes 10, 20, 30 and 5 then the first contrast in the effect coding would be 1,0,0,-10/5;

```
contr.sum(4)
```

```
##   [,1] [,2] [,3]
## 1    1    0    0
## 2    0    1    0
## 3    0    0    1
## 4   -1   -1   -1
```

In Lecture Notes #8 I will provide more explain one how to work with dummy codes, and why effect coding and contrasts are better than dummy codes (it has to do with way that the interaction on dummy codes messes up the main effects and messes up with unbalanced designs as we see here).

Random and nested effects

I suggest you compute ANOVAs in the usual way (e.g., aov) and manually compute the correct F tests where you use the appropriate error term.

You can use the lme4 or nlme libraries, but they introduce more complication. For example, the author of the lme4 library believes that users should know their appropriate degrees of freedom so models you run in lme4 do not print degrees of freedom and p-values. Someone else wrote an additional package that adds p-value functionality to the lme4 package, called the LMERConvenienceFunctions library (there is also a similar package called lmerTest and the afex package), which provide a wrapper to lmer to give p-values and degrees of freedom. Sometimes the additional packages get it right, other times they don't as we will see later in this appendix.

As you are learning this material in R, it is best to start by computing the correct tests by hand; that is, the proper F tests with the correct error term, performing contrasts by hand with the correct error term, etc., and as you gain familiarity with the random effect packages like lme4 or nlme, then add the bells and whistles always double checking that you are recreating in syntax what you do by hand. Using lme4 or nlme or mixed in the afex package requires putting the data in long format and handling contrasts slightly differently, issues we'll cover briefly here but in more detail in Lecture Notes 7 and 8.

For random and nested effects, there are the same issues we saw before when you rely on R's default dummy code contrasts. Either manually define contrasts on each factor or use

the second contrast would be 0,1,0,-20/5; the third contrast would be 0,0,1,-30/5. For the standard effects coding where the intercept is equal to the mean of the cell means, the effect coding ignores sample size and becomes like above with the first contrast being 1,0,0,-1; second contrast 0,1,0,-1; and third contrast 0,0,1,-1. This gets closer to what the hierarchical approach is doing but this doesn't give the tests of the main effects (just tests of the treatment effect αs relative to the grand mean). Similarly, one could define these weighted contrasts on the main effects as in 1,-12/10 for sex and 1,-1 for college to correspond to the 12 vs 10 ratio in sex and 10 vs 10 ratio in college. The anova() command on the aov() or lm() output reproduces the hierarchical method, but the direct output from the lm() command is a little difficult to interpret, and we'll cover that in LN8.

the `contrasts=list()` argument as discussed earlier within the command call such as for aov(), lm(), etc.

Here I show how to do this the old fashion way in R using the aov() command. This requires balanced data (meaning equal number of subjects across cells); for more general settings one should use packages like lme4 or nlme, we just need to be careful about making sure we are using the right error term.

```
# read data, define factors
random.data <- read.csv("random-data.csv", header = T)
random.data[, "thrpy"] <- factor(random.data[, "thrpy"])
random.data[, "thpst"] <- factor(random.data[, "thpst"])
# print first few rows for structure of data
head(random.data)


##   thrpy thpst score id
## 1     1     1    40  1
## 2     1     1    42  2
## 3     1     1    36  3
## 4     1     1    35  4
## 5     1     1    37  5
## 6     1     2    40  6


# rest of output parallels the main body of LN (roughly
# pages 10-13) in SPSS one way anova with thrpy as fixed
# effect factor
summary(aov(score ~ thrpy, data = random.data))


##             Df Sum Sq Mean Sq F value
## thrpy        2    120   60.00   5.915
## Residuals   42    426   10.14
##              Pr(>F)
## thrpy       0.00545 **
## Residuals
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1


# two way anova with both factors as fixed
summary(aov(score ~ thrpy * thpst, data = random.data))


##             Df Sum Sq Mean Sq F value
## thrpy         2 120.00   60.00   6.835
## thpst         2  43.33   21.67   2.468
## thrpy:thpst  4  66.67   16.67   1.899
## Residuals   36 316.00    8.78
##              Pr(>F)
## thrpy       0.00305 **
## thpst       0.09895 .
## thrpy:thpst 0.13188
## Residuals
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
```

The last source table gives you everything you need for computations for random effects. If both A and B are random, leave the interaction as is but recompute the two main effect Fs using MS for interaction. For example, for random factor thrpy, the new F observed would be 60/16.67=3.6 (rather than 6.835 in the output), the critical F is 6.94 (as seen from the command qf(.95,2,4) since there are 2 df in the numerator MS and 4 df in the denominator df). The recomputed observed F is not statistically significant when recomputed and compared to the new F critical (note that the original computation as thrpy fixed yielded a statistically significant omnibus test).

For contrasts and post-hoc tests like Tukey and Scheffe, you can compute contrasts in the usual way to get R to compute Ihat, but the standard error of Ihat would have to be recomputed with the appropriate error term depending on whether the factor on which the contrast is computed is fixed or random. Basically, whenever a formula calls for MSE you may need to substitute MSInteraction (and its degrees of freedom). This applies to Tukey and Scheffe as well.

For nested effects you can still use the source table of the full factorial because it has all the ingredients you need but you will have to recompute Fs much like the case of random effects. Note that the model that has one factor nested in another merely collapses the sums of squares of the nested factor and the interaction. So if thrpy is a fixed effect and thrpst are nested in thrpy, the SS for thrpy is the same as for the two way anova, the nested thrpst is SSthpst + SSint (43.33+66.66=110), and the SSerror is the same as the two way anova. You would then need to recompute the Fs using the appropriate error term for the test you are using.

For completeness, here is R code to avoid hand computation. First, the case for two random effects. We need to run two models, one to get the two main effects with the correct interaction term as the error, and a second model that uses the usual MSW as the error term and from this model we only examine the interaction term (ignore the two main effects from this second model). The degrees of freedom, F and p values are identical to what we saw with SPSS. The SS terms are slightly different because of how lmer normalizes the data in the context of the models, but the ratio of mean square terms (the Fs) are appropriately preserved, so the results of the statistical decisions are equivalent to SPSS and hand computation, even though the SS terms are slightly different. The summary() command prints the results of the contrasts; you use the contrasts in the appropriate model (i.e., for the main effect, you use the first model because you need the interaction term as the error; for the interaction contrasts, you use the second model because you need the error term based on MSW).

```
library(lmerTest)
library(pbkrtest)
contrasts(random.data[, "thrpy"]) <- cbind(c(1, -1, 0), c(1,
    1, -2))
contrasts(random.data[, "thpst"]) <- cbind(c(1, -1, 0), c(1,
    1, -2))

# two random effects; both main effects are correctly
# divided by the interaction term; use summary to get tests
# of contrasts for random effect terms having correct error
# term
out.lmer <- lmer(score ~ thpst + thrpy + (1 | thrpy:thpst), data = random.data)
anova(out.lmer, ddf = "Satterthwaite", type = "III")


## Type III Analysis of Variance Table with Satterthwaite's method
##       Sum Sq Mean Sq NumDF DenDF
## thpst 22.822  11.411     2     4
## thrpy 63.200  31.600     2     4
##       F value Pr(>F)
## thpst     1.3 0.3673
## thrpy     3.6 0.1276


summary(out.lmer)


## Linear mixed model fit by REML.
##   t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula:
## score ~ thpst + thrpy + (1 | thrpy:thpst)
##    Data: random.data
##
## REML criterion at convergence: 222.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5871 -0.8528  0.1598  0.7936  1.8654
##
## Random effects:
##  Groups      Name        Variance
##  thrpy:thpst (Intercept) 1.578
##  Residual                8.778
##  Std.Dev.
##  1.256
##  2.963
## Number of obs: 45, groups:
## thrpy:thpst, 9
##
## Fixed effects:
##             Estimate Std. Error      df
## (Intercept)  42.0000     0.6086  4.0000
## thpst1       -0.8333     0.7454  4.0000
## thpst2        0.5000     0.4303  4.0000
## thrpy1       -1.0000     0.7454  4.0000
## thrpy2       -1.0000     0.4303  4.0000
##             t value Pr(>|t|)
## (Intercept)  69.013 2.64e-07 ***
## thpst1       -1.118   0.3262
## thpst2        1.162   0.3099
## thrpy1       -1.342   0.2508
## thrpy2       -2.324   0.0808 .
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##        (Intr) thpst1 thpst2 thrpy1
## thpst1 0.000
## thpst2 0.000  0.000
## thrpy1 0.000  0.000  0.000
## thrpy2 0.000  0.000  0.000  0.000
```

The Satterthwaite is a Welch-like correction to degrees of freedom that is typically done in random effects modeling as a matter of course regardless of whether the equal variance assumption holds or not.

The reason I did a random intercept within each cell (the interaction) is so that the degrees of freedom would be correct. By specifying a random intercept for each cell in the interaction, lmer interprets that term as the error term and will use the interaction term when testing the two main effects on the random factors. If I had done what may seem like a more natural specification of calling each of the two factors random effects by assigning them each a random intercept, the terms are incorrect as I show below.

Next, test the interaction term (ignore main effects here because these were already tested in the lmer syntax above with the correct error term—the main effects printed in this next output use the incorrect error term so you ignore them).

```
# second model; only examine the interaction term here;
# this uses the correct MSW as the error term
summary(aov(score ~ thrpy * thpst, data = random.data))


##              Df Sum Sq Mean Sq F value
## thrpy         2 120.00   60.00   6.835
## thpst         2  43.33   21.67   2.468
## thrpy:thpst   4  66.67   16.67   1.899
## Residuals    36 316.00    8.78
##               Pr(>F)
## thrpy        0.00305 **
## thpst        0.09895 .
## thrpy:thpst  0.13188
## Residuals
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
```

Here is an example of what may seem more natural syntax by making each factor "random", but it yields incorrect results for this model. While the interaction is correctly tested with the MSE term having 36 degrees of freedom, the two main effects are incorrectly tested with MSE rather than the MSInteraction. I've seen this form of syntax suggested on the web in several places, so as usual be careful when using what you find on the web.

```
# this output contains warnings that for purposes of this
# demo can be ignored, but the warnings are a clue
# something is awry with this 'natural' syntax
out.lmer <- lmer(score ~ thrpy * thpst + (1 | thpst) + (1 | thrpy),
    data = random.data)


## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv,
:  unable to evaluate scaled gradient

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv,
:   Hessian is numerically singular:  parameters are not uniquely determined

## Warning in as_lmerModLT(model, devfun):  Model may not have converged
with 2 eigenvalues close to zero:  1.1e-10 -1.1e-10


anova(out.lmer, ddf = "Satterthwaite", type = "III")
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##              Sum Sq Mean Sq NumDF DenDF
## thrpy        21.038 10.5192     2    36
## thpst        16.058  8.0288     2    36
## thrpy:thpst  66.667 16.6667     4    36
##              F value Pr(>F)
## thrpy         1.1984 0.3134
## thpst         0.9147 0.4098
## thrpy:thpst   1.8987 0.1319
```

This output uses 36 degrees of freedom for the denominator for all three tests, when it should be 4 for the two random main effects. The interaction test is right (up to round off error) but the main effect tests treating both factors as random effects is not correct. So even though this looks like in could be the right syntax, it doesn't produce the correct tests.

In a mixed effects case where thrpy is fixed and thpst is random, run two separate models. The first model yields the fixed effect term that is tested with the correct interaction term; the second model has the random main effect and the interaction that use the MSW as the error term (ignore the fixed effect term in this model because it is incorrectly tested, use test from the first model).

```
# note the interaction defined as random effect; this will
# then prompt the interaction term as the error term for
# therapy; ignore the other tests in this output because
# they have the incorrect error term
out.lmer <- lmer(score ~ thrpy + thpst + (1 | thrpy:thpst), data = random.data)
anova(out.lmer, ddf = "Satterthwaite", type = "III")


## Type III Analysis of Variance Table with Satterthwaite's method
##        Sum Sq Mean Sq NumDF DenDF
## thrpy 63.200  31.600     2     4
## thpst 22.822  11.411     2     4
##        F value Pr(>F)
## thrpy      3.6 0.1276
## thpst      1.3 0.3673


# tests of individual contrasts on therapy
summary(out.lmer)


## Linear mixed model fit by REML.
##   t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula:
```

```
## score ˜ thrpy + thpst + (1 | thrpy:thpst)
##    Data: random.data
##
## REML criterion at convergence: 222.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5871 -0.8528  0.1598  0.7936  1.8654
##
## Random effects:
##  Groups      Name        Variance
##  thrpy:thpst (Intercept) 1.578
##  Residual                8.778
##  Std.Dev.
##  1.256
##  2.963
## Number of obs: 45, groups:
## thrpy:thpst, 9
##
## Fixed effects:
##             Estimate Std. Error      df
## (Intercept)  42.0000     0.6086  4.0000
## thrpy1       -1.0000     0.7454  4.0000
## thrpy2       -1.0000     0.4303  4.0000
## thpst1       -0.8333     0.7454  4.0000
## thpst2        0.5000     0.4303  4.0000
##             t value Pr(>|t|)
## (Intercept)  69.013 2.64e-07 ***
## thrpy1       -1.342   0.2508
## thrpy2       -2.324   0.0808 .
## thpst1       -1.118   0.3262
## thpst2        1.162   0.3099
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##        (Intr) thrpy1 thrpy2 thpst1
## thrpy1 0.000
## thrpy2 0.000  0.000
## thpst1 0.000  0.000  0.000
## thpst2 0.000  0.000  0.000  0.000
```

```
# second model; only use the interaction term and the
# random main effect; these have MSW as the error; ignore
# main effect of the fixed effect factor because it was
# tested by the previous model
summary(aov(score ~ thrpy * thpst, data = random.data))


##              Df Sum Sq Mean Sq F value
## thrpy         2 120.00   60.00   6.835
## thpst         2  43.33   21.67   2.468
## thrpy:thpst   4  66.67   16.67   1.899
## Residuals    36 316.00    8.78
##               Pr(>F)
## thrpy        0.00305 **
## thpst        0.09895 .
## thrpy:thpst  0.13188
## Residuals
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
```

For nested effects, you can also use the aov command to reproduce the appropriate source table, then you can compute F observed by hand and F critical from the details of the source table. In R the A/B term inside the Error() specification is interpreted as B nested within A.

```
out.nested.aov <- aov(score ~ thrpy + Error(thrpy/thpst), data = random.data)
summary(out.nested.aov)


##
## Error: thrpy
##       Df Sum Sq Mean Sq
## thrpy  2    120      60
##
## Error: thrpy:thpst
##           Df Sum Sq Mean Sq F value
## Residuals  6    110   18.33
##           Pr(>F)
## Residuals
##
## Error: Within
##           Df Sum Sq Mean Sq F value
## Residuals 36    316   8.778
##           Pr(>F)
## Residuals
```

For a more direct approach to nested designs, I started tinkering with the lmer command in the lmerTest package, which is an augmented lmer command from the lme4 package with extras such as degrees of freedom and p-values included in the output. This gives the right df, F observed and p-values (SS are rescaled), but this package needs to be studied more carefully to make sure the logic is right and not just the right output for this particular design by accident. This package also go the random effects analyses presented above correct, so overall looks promising.

```
library(lmerTest)
library(pbkrtest)
out.nested <- lmer(score ~ thrpy + (1 | thrpy:thpst), data = random.data)
summary(out.nested)


## Linear mixed model fit by REML.
##   t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula:
## score ~ thrpy + (1 | thrpy:thpst)
##    Data: random.data
##
## REML criterion at convergence: 226.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5117 -0.8367  0.0000  0.8510  1.6876
##
## Random effects:
##  Groups      Name        Variance
##  thrpy:thpst (Intercept) 1.911
##  Residual                8.778
##  Std.Dev.
##  1.382
##  2.963
## Number of obs: 45, groups:
## thrpy:thpst, 9
##
## Fixed effects:
##             Estimate Std. Error      df
## (Intercept)  42.0000     0.6383  6.0000
## thrpy1       -1.0000     0.7817  6.0000
## thrpy2       -1.0000     0.4513  6.0000
##             t value Pr(>|t|)
## (Intercept)  65.801 8.29e-10 ***
## thrpy1       -1.279   0.2481
## thrpy2       -2.216   0.0686 .
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##        (Intr) thrpy1
## thrpy1 0.000
## thrpy2 0.000  0.000
```

```
anova(out.nested)
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##       Sum Sq Mean Sq NumDF DenDF
## thrpy 57.455  28.727     2     6
##       F value Pr(>F)
## thrpy  3.2727 0.1094
```

```
summary(aov(score ~ thrpy + Error(thrpy/thpst), data = random.data))
```

```
##
## Error: thrpy
##       Df Sum Sq Mean Sq
## thrpy  2    120      60
##
## Error: thrpy:thpst
##           Df Sum Sq Mean Sq F value
## Residuals  6    110   18.33
##           Pr(>F)
## Residuals
##
## Error: Within
##           Df Sum Sq Mean Sq F value
## Residuals 36    316   8.778
##           Pr(>F)
## Residuals
```

Repeated measures

Repeated measures ANOVA in R is not well developed at this point. Most people end up using a random effect approach to repeated measures, such as the implementation in the lme4 library, because simple repeated measures models have a subjects factor that is treated as a random effect, as I showed earlier in these lecture notes. This requires putting the repeated measures data in long format as opposed to wide format. Wide format has one row per subject and each time point as a separate column; long format has all time data stacked so if there are three time points and 5 subjects, the long format has 15 (3 times 5) rows. When using long format, one also needs to add a variable that codes for time (e.g., if three times, then a grouping code of 1s, 2s and 3s) and codes for subjects (i.e., subject number). One can specify contrasts on the time factor, such as 1, 0, -1 and 1, -2, 1 in the case of three times. Long format is the preferred format for random effect models because it generalizes to more advanced techniques such as multilevel modeling and latent growth curve models. However, not all implementations in R produce the correct error term and correct degrees of freedom.

If all the factors in a study are repeated measures factors, I typically use the short cut presented in the body of the lecture notes that creates new variables by weighting the repeated measures with the contrast weights and running straightforward t tests. But if I have a mixed design with both repeated and between subjects designs, things get a little tricky (e.g., keeping track of degrees of freedom, sometimes requiring manual changes to the degrees of freedom used in the output as I showed earlier in these lecture notes), so I always double check my work with another program for consistency.

The car package's Anova() command can run repeated measures ANOVA with the Greenhouse-Geisser correction. See the help file for the Anova command in R; their example for the multivariate linear model for repeated-measures data shows a 3 by 5 ANOVA with repeated measures on both factors. You specify the design and contrasts on the repeated measures factor through the idesign and icontrast arguments to the Anova() command. Unlike the random effects approach, the car package requires data in wide format and uses a construction of putting all the repeated measures variables on the dependent variable side by creating a dv matrix. So, if you want to verify analyses in R across packages you sometimes need to switch back and forth between long and wide formats of the data frame. The dplyr package that is part of the tidyverse ecosystem, has some nice functions that help with these conversions. The dplyr package is actively developed. The current functions for moving between wide and long format data.frames are pivot_wider() and pivot_longer(), though earlier versions of dplyr had completely different approaches and you'll find some of those outdated approaches still lingering on the web.

Baron and Li maintain some R notes for running various statistics (`http://www.psych.upenn.edu/~baron/rpsych/rpsych.html`). You may find these notes helpful for various things like computing sphericity, Greenhouse-Geisser, using the Error() feature within the aov() command to specify the appropriate error term (but that is only for balanced designs), etc. They do a nice job of explaining so I'll just direct you their notes. It is convenient that they use some of the same standard data sets I've included in my lecture notes.

You can try a few newer libraries that add functionality such as the ezANOVA() command in the ez library. Be careful because many commands in the ez library assume balanced data (e.g., no missing data), presumably so they can avoid the extra computational hassles of unequal samples including providing options for the various ways of handling the decomposition of sums of squares (regression, hierarchical and sequential).

I suggest you run through some of the examples in the lecture notes with some of these newer functions to ensure that you are using them correctly and getting the same output as the lecture notes to confirm that you using the R syntax correctly.

This is probably the only occasion I'll say this: for repeated measure ANOVA, it is usually easier and safer to use SPSS and the manova command. There isn't much yet available in R on the fancy stuff like Box's ϵ, Greenhouse-Geisser corrections, etc. If you are interested in developing your own package, one on repeated measures ANOVA will be a useful one to contribute. But, it will be a lot of work to write this for general use. Some insight into how to accomplish this task will be given in LN12, where I introduce an approach using linear algebra that simplifies much of the hassles in these lecture notes (the logic generalizes the material in Maxwell and Delaney's two textbook chapters on the multivariate approach to repeated measures ANOVA, which doesn't make the sphericity assumption) and provides a general framework for handling repeated measures analyses.

One way repeated measures

One way repeated measures

For those who insist on using R for repeated measures. I offer a few examples. I'll illustrate how one gets different results with different R commands, which is not a good thing. The major lesson here is know your R commands and test them out on examples where you know the right answer to make sure you are reproducing those answers in R.

Let's read in the data I used for the one way repeated measure ANOVA presented earlier in the lecture notes (page 5-32). First, I'll show how to do this by creating three contrast scores and then following up with one sample t tests. This replicates the one way repeated measures example in the lecture notes using SPSS. Note that degrees of freedom are 11 (11 is the magic number in this example), which we will compare with some R output later where it is a different value (e.g., sometimes 33, sometimes 44).

This example has the data in wide format: each row is a different subject and the columns correspond to the four time points (plus an additional column for subject number).

```
data <- cbind(1:12, c(92, 120, 112, 95, 114, 99, 124, 106, 100,
    108, 112, 102), c(95, 121, 111, 96, 112, 100, 125, 107, 98,
    110, 115, 102), c(96, 121, 111, 98, 110, 99, 127, 106, 95,
    112, 116, 101), c(98, 123, 109, 99, 109, 98, 126, 107, 94,
    115, 118, 101))
colnames(data) <- c("subjectID", paste("time", 1:4, sep = ""))
```

```
data

##       subjectID time1 time2 time3 time4
##  [1,]         1    92    95    96    98
##  [2,]         2   120   121   121   123
##  [3,]         3   112   111   111   109
##  [4,]         4    95    96    98    99
##  [5,]         5   114   112   110   109
##  [6,]         6    99   100    99    98
##  [7,]         7   124   125   127   126
##  [8,]         8   106   107   106   107
##  [9,]         9   100    98    95    94
## [10,]        10   108   110   112   115
## [11,]        11   112   115   116   118
## [12,]        12   102   102   101   101

contrast1 <- data[, 2] + data[, 3] - data[, 4] - data[, 5]
contrast2 <- data[, 2] - data[, 3] + data[, 4] - data[, 5]
contrast3 <- data[, 2] - data[, 3] - data[, 4] + data[, 5]

t.test(contrast1)

##
##  One Sample t-test
##
## data:  contrast1
## t = -0.64453, df = 11, p-value =
## 0.5324
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -4.782774  2.616107
## sample estimates:
## mean of x
## -1.083333

t.test(contrast2)

##
##  One Sample t-test
##
## data:  contrast2
## t = -1.2151, df = 11, p-value =
## 0.2498
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -3.0455745  0.8789079
## sample estimates:
## mean of x
## -1.083333
```

```
t.test(contrast3)
```

```
##
##  One Sample t-test
##
## data:  contrast3
## t = -0.76089, df = 11, p-value =
## 0.4627
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.9731653  0.4731653
## sample estimates:
## mean of x
##     -0.25
```

Here is the usual repeated measures omnibus test using the afex package, which requires data in long format. The afex package produces the correct omnibus F, degrees of freedom, Greenhouse-Geisser, Box ϵ, etc. For completelness with commands I'll show later, I also create three within subject contrasts m1, m2, and int as separate columns to include in the data table.

```
dv <- c(data[, 2], data[, 3], data[, 4], data[, 5])
sub <- factor(rep(1:12, 4))
# let's select 3 orthogonal contrastas
m1 <- (rep(c(1, 1, -1, -1), each = 12))
m2 <- (rep(c(1, -1, 1, -1), each = 12))
int <- (rep(c(1, -1, -1, 1), each = 12))
time <- factor(rep(1:4, each = 12))
data.oneway <- cbind(sub, dv, m1, m2, int, time)
data.oneway
```

```
##       sub  dv m1 m2 int time
##  [1,]   1  92  1  1   1    1
##  [2,]   2 120  1  1   1    1
##  [3,]   3 112  1  1   1    1
##  [4,]   4  95  1  1   1    1
##  [5,]   5 114  1  1   1    1
##  [6,]   6  99  1  1   1    1
##  [7,]   7 124  1  1   1    1
##  [8,]   8 106  1  1   1    1
##  [9,]   9 100  1  1   1    1
## [10,]  10 108  1  1   1    1
## [11,]  11 112  1  1   1    1
## [12,]  12 102  1  1   1    1
## [13,]   1  95  1 -1  -1    2
## [14,]   2 121  1 -1  -1    2
## [15,]   3 111  1 -1  -1    2
## [16,]   4  96  1 -1  -1    2
## [17,]   5 112  1 -1  -1    2
## [18,]   6 100  1 -1  -1    2
## [19,]   7 125  1 -1  -1    2
## [20,]   8 107  1 -1  -1    2
## [21,]   9  98  1 -1  -1    2
## [22,]  10 110  1 -1  -1    2
## [23,]  11 115  1 -1  -1    2
## [24,]  12 102  1 -1  -1    2
## [25,]   1  96 -1  1  -1    3
## [26,]   2 121 -1  1  -1    3
## [27,]   3 111 -1  1  -1    3
## [28,]   4  98 -1  1  -1    3
## [29,]   5 110 -1  1  -1    3
## [30,]   6  99 -1  1  -1    3
## [31,]   7 127 -1  1  -1    3
## [32,]   8 106 -1  1  -1    3
## [33,]   9  95 -1  1  -1    3
## [34,]  10 112 -1  1  -1    3
## [35,]  11 116 -1  1  -1    3
## [36,]  12 101 -1  1  -1    3
## [37,]   1  98 -1 -1   1    4
## [38,]   2 123 -1 -1   1    4
## [39,]   3 109 -1 -1   1    4
## [40,]   4  99 -1 -1   1    4
## [41,]   5 109 -1 -1   1    4
## [42,]   6  98 -1 -1   1    4
## [43,]   7 126 -1 -1   1    4
## [44,]   8 107 -1 -1   1    4
## [45,]   9  94 -1 -1   1    4
## [46,]  10 115 -1 -1   1    4
## [47,]  11 118 -1 -1   1    4
## [48,]  12 101 -1 -1   1    4
```

```
library(afex)
library(car)
out.aovez <- aov_ez("sub", "dv", data = data.oneway, within = c("time"),
    type = 3)
summary(out.aovez)
```

```
##
## Univariate Type III Repeated-Measures ANOVA Assuming Sphericity
##
##             Sum Sq num Df Error SS
## (Intercept) 555776      1   4387.2
## time             7      3    123.0
##             den Df   F value    Pr(>F)
## (Intercept)     11 1393.4833 6.154e-13
## time            33    0.6464    0.5908
##
## (Intercept) ***
## time
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
##
##
## Mauchly Tests for Sphericity
##
##      Test statistic    p-value
## time        0.01905 3.6732e-07
##
##
## Greenhouse-Geisser and Huynh-Feldt Corrections
##  for Departure from Sphericity
##
##       GG eps Pr(>F[GG])
## time 0.37577     0.4547
##
##         HF eps Pr(>F[HF])
## time 0.3892073  0.4594742
```

The omnibus test with compound symmetry is provided and you get some corrections for sphericity as well. The package afex doesn't appear to do multivariate tests (though see below for contrasts on afex output through the extra emmeans package); or another attack on the problem is to use the lm command, which can handle a dv that has multiple columns and then use Anova() from the car package to get the multivariate omnibus tests. This replicates the multivariate tests in SPSS. But this is a hassle if you want both multivariate and sphericity, you need to run two commands and afex wants data in long format, whereas lm wants data in wide format. The `aov_ez` program stores data in both long and wide format to facilitate moving back and forth between the two formats (e.g., look at names(out.aovez), the object data in the list contains the raw data in both long and wide, which you can see with `out.aovez[["data"]][["wide"]]` and `out.aovez[["data"]][["long"]]`). Storing data in both long and wide format is not very efficient, and could create issues if you have large datasets such as 10 million observations—a case where a single data frame is large, now that memory need is essentially doubled by storing the data in both formats.

Here is the output from the lm() followed by the Anova() command combination I mentioned in the previous paragraph.

```
# this commmand uses wide format data there is no grouping
# variable so the predictor is just an intercept
out.lm <- lm(data[, 2:5] ~ 1)
idata <- data.frame(time = factor(c(1:4)))
mv.anova <- Anova(out.lm, idata = idata, idesign = ~time, type = 3)
summary(mv.anova)


##
## Type III Repeated Measures MANOVA Tests:
##
## ------------------------------------------
##
## Term: (Intercept)
##
##  Response transformation matrix:
##       (Intercept)
## time1           1
## time2           1
## time3           1
## time4           1
##
## Sum of squares and products for the hypothesis:
##             (Intercept)
## (Intercept)     2223102
##
## Multivariate Tests: (Intercept)
##                   Df test stat approx F
## Pillai             1   0.99217 1393.483
## Wilks              1   0.00783 1393.483
## Hotelling-Lawley  1 126.68030 1393.483
## Roy               1 126.68030 1393.483
##                  num Df den Df
## Pillai                1     11
## Wilks                 1     11
## Hotelling-Lawley      1     11
## Roy                   1     11
##                      Pr(>F)
## Pillai           6.1539e-13 ***
## Wilks            6.1539e-13 ***
## Hotelling-Lawley 6.1539e-13 ***
## Roy              6.1539e-13 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
##
## ------------------------------------------
##
## Term: time
##
##  Response transformation matrix:
##       time1 time2 time3
## time1     1     0     0
## time2     0     1     0
## time3     0     0     1
## time4    -1    -1    -1
##
## Sum of squares and products for the hypothesis:
##           time1    time2    time3
## time1 14.083333 5.416667 5.416667
## time2  5.416667 2.083333 2.083333
## time3  5.416667 2.083333 2.083333
##
## Multivariate Tests: time
##                  Df test stat approx F
## Pillai            1 0.2758637 1.142866
## Wilks             1 0.7241363 1.142866
## Hotelling-Lawley  1 0.3809554 1.142866
## Roy               1 0.3809554 1.142866
##                  num Df den Df  Pr(>F)
## Pillai                3      9 0.38322
## Wilks                 3      9 0.38322
## Hotelling-Lawley      3      9 0.38322
## Roy                   3      9 0.38322
##
## Univariate Type III Repeated-Measures ANOVA Assuming Sphericity
##
##             Sum Sq num Df Error SS
## (Intercept) 555776      1   4387.2
## time             7      3    123.0
##             den Df   F value    Pr(>F)
## (Intercept)     11 1393.4833 6.154e-13
## time            33    0.6464    0.5908
##
## (Intercept) ***
## time
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
##
##
## Mauchly Tests for Sphericity
##
##      Test statistic    p-value
## time        0.01905 3.6732e-07
##
##
## Greenhouse-Geisser and Huynh-Feldt Corrections
##  for Departure from Sphericity
##
##       GG eps Pr(>F[GG])
## time 0.37577     0.4547
##
##         HF eps Pr(>F[HF])
## time 0.3892073  0.4594742
```

To repeat, checking assumptions such as compound symmetry or sphericity come into play when one wants to test the omnibus test that there is a main effect of time. Things become a little more straightforward when you focus on contrasts.

contrasts

Let's switch from omnibus tests to contrasts. We know we get the correct answer with the one sample t test approach using contrast weights to create a new score for each subject. Let's see if we can get R to cooperate with some of its built in tools.

Here is an attempt using emmeans with the `aov_ez` output we ran above. This reproduces the manual t-tests and the SPSS results but I had to specify `model="multivariate"` . Apparently, `model="multivariate"` is now the default. Later, I'll show how to do the univariate version that assumes compound symmetry (with standard errors constrained to be identical for every single contrast rather than each contrast having its own standard error, as in the multivariate approach). The emmeans package expects you to provide a list of contrasts to test.

The afex and emmeans combination provides a variety of bells and whistles, including the ability to perform interaction contrasts automatically, apply various corrections for multiple tests, and other nice features. But you need to be careful about the various options and that you are using them correctly such as including the `model="multivariate"` argument to the emmeans call.

```
library(emmeans)
out.emmeans <- emmeans(out.aovez, specs=˜time,
                       model="multivariate")
cont <- list(m1=c(1,1,-1,-1),m2=c(1,-1,1,-1),
             int=c(1,-1,-1,1))
contrast(out.emmeans, cont)


##  contrast estimate    SE df t.ratio
##  m1          -1.08 1.681 11  -0.645
##  m2          -1.08 0.892 11  -1.215
##  int         -0.25 0.329 11  -0.761
##  p.value
##   0.5324
##   0.2498
##   0.4627
```

An aside: understanding the error term in repeated measures anova. The emmeans command having specified the multivariate option uses a full covariance matrix without forcing variances to be equal and covariances to be equal. Here is the covariance matrix of the four times that emmeans is using; he labels of “Intercept” seem strange here but makes sense in terms of the modeling strategy the afex/emmeans combination uses to estimate variances and covariances for each time period.

```
print(round(slot(out.emmeans, "V"), 4), width = 80)


##                 X1:(Intercept) X2:(Intercept) X3:(Intercept) X4:(Intercept)
## X1:(Intercept)          8.0758         7.8939         8.0682         7.8939
## X2:(Intercept)          7.8939         7.9444         8.2854         8.2828
## X3:(Intercept)          8.0682         8.2854         8.8535         8.9646
## X4:(Intercept)          7.8939         8.2828         8.9646         9.2948
```

These are identical to the observed covariance matrix

```
round(var(data[, 2:5])/12, 4)


##        time1  time2  time3  time4
## time1 8.0758 7.8939 8.0682 7.8939
## time2 7.8939 7.9444 8.2854 8.2828
## time3 8.0682 8.2854 8.8535 8.9646
## time4 7.8939 8.2828 8.9646 9.2948
```

The standard errors reported by emmeans follow from this covariance matrix. The explanation of the formala will come next semester once we know some linear algebra, but for now we pre- and post-multiply the covariance matrix by the contrast matrix and take the square root of the diagonal. You'll see these three numbers correspond to the standard errors reported in the contrast(out.emmeans, cont) command above.

```
covarmat <- var(data[, 2:5])/12
contrastmat <- cbind(c(1, 1, -1, -1), c(1, -1, 1, -1), c(1, -1,
    -1, 1))
sqrt(diag(t(contrastmat) %*% covarmat %*% contrastmat))


## [1] 1.6808112 0.8915286 0.3285644
```

To compare this output with the univariate approach that, for example, implements the compound symmetry symmetry, let's rerun afex/emmeans using the univariate option and look at the output and the covariance matrix this command uses. Under this approach the standard error for the contrasts is 1.11 for every contrast (and also the df is now 33 when before it was 11).

```
#uses univariate rather than multivariate to compare
out.aovex.u <- aov_ez("sub","dv", data=data.oneway,
                      within="time",type=3, include_aov=T)
out.emmeans.u <- emmeans(out.aovex.u, specs=~time,
                         model="univariate")
contrast(out.emmeans.u,cont)


##  contrast estimate   SE df t.ratio
##  m1          -1.08 1.11 33  -0.972
##  m2          -1.08 1.11 33  -0.972
##  int         -0.25 1.11 33  -0.224
##  p.value
##   0.3382
##   0.3382
##   0.8239
```

How did this standard error of 1.11 come about? We assume compound symmetry and compute the mean variance and the mean covariance, then create a covariance matrix using those terms so the mean variance is everywhere in the diagonal and the mean covariance is everywhere in the off-diagonal.

```
meanV <- mean(diag(var(data[, 2:5])/12))
meanC <- mean(var(data[, 2:5])[lower.tri(var(data[, 2:5]))]/12)
compsymmat <- matrix(meanC, 4, 4)
diag(compsymmat) <- meanV
round(compsymmat, 5)


##         [,1]    [,2]    [,3]    [,4]
## [1,] 8.54214 8.23148 8.23148 8.23148
## [2,] 8.23148 8.54214 8.23148 8.23148
## [3,] 8.23148 8.23148 8.54214 8.23148
## [4,] 8.23148 8.23148 8.23148 8.54214
```

Finally, pre- and post-multiply this covariance matrix by the contrast matrix as we did before

```
sqrt(diag(t(contrastmat) %*% compsymmat %*% contrastmat))


## [1] 1.114735 1.114735 1.114735
```

We see that a subtle change in syntax led to a different standard error. Also, the original afex/emmeans multivariate approach used df=11 for each contrast, but in this univariate afex/emmeans version we not only see a different standard error but also that the df are 33.

Another common approach to testing contrasts in repeated measures designs is as a linear mixed model. It is true that repeated measures can be converted to a linear mixed model, as I have said several times. However, it depends on the details whether you get the correct results (i.e., correct degrees of freedom and error term). I'll just focus on the three contrast tests rather than the omnibus test for this part. We need to list the contrasts as predictors. This implementation yields the compound symmetry version of repeated measures that mimics the incorrect p values that came out of the contrasts in the afex package above. I also repeat the example using another function in the nlme package that you'll find several mentions when searching the web for repeated measures ANOVA in R.

```
library(nlme)
#playing with form to allow different variances
#over time; commented out here
#out.lme <- lme(dv ~ m1+m2+int, random = ~1|sub,
#             data=data.frame(data.oneway), weight#=varIdent(form=~1|time),
#             method="REML")
```

```
out.lme <- lme(dv ~ m1+m2+int, random = ~1|sub,
            data=data.frame(data.oneway),
            method="REML")
summary(out.lme)
```

```
## Linear mixed-effects model fit by REML
##   Data: data.frame(data.oneway)
##        AIC      BIC    logLik
##   261.6485 272.3536 -124.8242
##
## Random effects:
##  Formula: ~1 | sub
##         (Intercept) Residual
## StdDev:    9.938701 1.930778
##
## Fixed effects:  dv ~ m1 + m2 + int
##                 Value Std.Error DF
## (Intercept) 107.60417 2.8825590 33
## m1           -0.27083 0.2786838 33
## m2           -0.27083 0.2786838 33
## int          -0.06250 0.2786838 33
##               t-value p-value
## (Intercept) 37.32939  0.0000
## m1          -0.97183  0.3382
## m2          -0.97183  0.3382
## int         -0.22427  0.8239
##  Correlation:
##     (Intr) m1 m2
## m1  0
## m2  0      0
## int 0      0  0
##
## Standardized Within-Group Residuals:
##        Min          Q1         Med
## -1.7250141 -0.5529223  0.0270251
##         Q3        Max
##  0.4366413  1.9436278
##
## Number of Observations: 48
## Number of Groups: 12
```

```
out.gls <- gls(dv ~ m1+m2+int,
            data=data.frame(data.oneway),
            method="REML")
summary(out.gls)

## Generalized least squares fit by REML
##   Model: dv ~ m1 + m2 + int
##   Data: data.frame(data.oneway)
##        AIC      BIC    logLik
##   354.0678 362.9887 -172.0339
##
## Coefficients:
##                 Value Std.Error
## (Intercept) 107.60417  1.461347
## m1           -0.27083  1.461347
## m2           -0.27083  1.461347
## int          -0.06250  1.461347
##              t-value p-value
## (Intercept) 73.63355  0.0000
## m1          -0.18533  0.8538
## m2          -0.18533  0.8538
## int         -0.04277  0.9661
##
##  Correlation:
##     (Intr) m1 m2
## m1   0
## m2   0      0
## int 0      0  0
##
## Standardized residuals:
##         Min          Q1          Med
## -1.48155333 -0.86629716  0.01234628
##          Q3         Max
##  0.68521842  1.90955763
##
## Residual standard error: 10.12451
## Degrees of freedom: 48 total; 44 residual
```

The lme and gls commands give different results to each other, and both results are different from the `aov_ez` multivariate command I ran above (though lme mimics the univariate version in terms of t test and p-values we saw for afex/emmeans using univariate). Strange. If you do a web search you will see repeated measures tutorials that present these but hardly any discussion on the vast differences between them and how they can lead to incorrect results.

Let's move toward something that works (almost). Recall that the key to contrasts in repeated measures is that we should use the multivariate approach to avoid the compound symmetry/sphericity assumptions. In order to do that we need to extend the linear mixed model to have a free covariance matrix rather than a covariance matrix that makes the restrictive assumptions. The end result is a better error term from the multivariate approach. SPSS does this automatically in MANOVA and the method of creating scores for each subject followed by one sample t tests also does this automatically. Here is how to do allow a free covariance matrix in lme and gls without imposing compound symmetry. I don't think this multivariate approach can be done with the lmer function in the lme4 package because lme4 doesn't yet implement ways of specifying covariance matrices.

```
#had to increase iteration number for convergence
#note how both correlations and variances are set
#to vary; lme wants correlation to correspond to
#the random specification but the variance needs
#to be defined by time as we want different
#time variances not different subject variances
out.lme <- lme(dv ~ m1+m2+int, random = ~1|sub,
               correlation = corSymm(form = ~1|sub),
            data=data.frame(data.oneway),
            weight=varIdent(form=~1|time),
            method="REML",
            control=lmeControl(maxIter=100,msMaxIter=200))
summary(out.lme)

## Linear mixed-effects model fit by REML
##   Data: data.frame(data.oneway)
##        AIC      BIC    logLik
##   234.3039 261.0667 -102.1519
##
## Random effects:
##  Formula: ~1 | sub
##         (Intercept) Residual
## StdDev:    9.210681 3.474544
##
## Correlation Structure: General
##  Formula: ~1 | sub
##  Parameter estimate(s):
##  Correlation:
##   1     2     3
## 2 0.879
## 3 0.745 0.973
## 4 0.551 0.870 0.951
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | time
##  Parameter estimates:
##         1         2         3         4
## 1.0000000 0.9324562 1.3315815 1.4871948
## Fixed effects:  dv ~ m1 + m2 + int
##                  Value Std.Error DF
## (Intercept) 107.60417 2.8825590 33
## m1            -0.27083 0.4202027 33
## m2            -0.27083 0.2228822 33
## int           -0.06250 0.0821411 33
##               t-value p-value
## (Intercept) 37.32939  0.0000
## m1          -0.64453  0.5237
## m2          -1.21514  0.2329
## int         -0.76089  0.4521
##  Correlation:
##     (Intr) m1     m2
## m1  -0.220
## m2  -0.072  0.905
## int -0.055  0.216  0.045
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med
## -1.5744230 -0.8529488 -0.2596430
##          Q3         Max
##  0.9401155  1.6252264
##
## Number of Observations: 48
## Number of Groups: 12
```

```
#correct specification of the covariance matrix
out.gls <- gls(dv ~ m1+m2+int,
            correlation = corSymm(form = ~time|sub),
            data=data.frame(data.oneway),
            weight=varIdent(form=~1|time),
            method="REML")
summary(out.gls)
```

```
## Generalized least squares fit by REML
##   Model: dv ~ m1 + m2 + int
##   Data: data.frame(data.oneway)
##         AIC      BIC    logLik
##   232.3039 257.2825 -102.1519
##
## Correlation Structure: General
##  Formula: ~time | sub
##  Parameter estimate(s):
```

```
##  Correlation:
##   1     2     3
## 2 0.986
## 3 0.954 0.988
## 4 0.911 0.964 0.988
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ˜1 | time
##  Parameter estimates:
##         1         2         3         4
## 1.0000000 0.9918367 1.0470485 1.0728254
##
## Coefficients:
##                 Value Std.Error
## (Intercept) 107.60417 2.8825589
## m1           -0.27083 0.4202029
## m2           -0.27083 0.2228822
## int          -0.06250 0.0821411
##              t-value p-value
## (Intercept) 37.32939  0.0000
## m1          -0.64453  0.5226
## m2          -1.21514  0.2308
## int         -0.76089  0.4508
##
##  Correlation:
##     (Intr) m1     m2
## m1  -0.220
## m2  -0.072  0.905
## int -0.055  0.216  0.045
##
## Standardized residuals:
##         Min          Q1         Med
## -1.52373366 -0.84563270  0.00925861
##          Q3         Max
##  0.66895594  1.87567567
##
## Residual standard error: 9.84424
## Degrees of freedom: 48 total; 44 residual
```

One thing to notice about the lme and gls output is that the t tests for the fixed effects contrasts are correct, but the pvalues are off. In other words, it uses the correct error term we saw for the afex/emmeans multivariate approach but this approach doesn't properly adjust the degrees of freedom. This is, the incorrect DFs are used; both tests use df = 33, which is (n-1)(t-1). If we use the right DFs (which should be 11=n-1) we get the correct p-values that match the afex/emmeans multivariate approach and the one sample t test method. To illustrate I extract the t values and use df = n - 1 to recompute the pvalues. This corresponds to the correct p-values because the correct error term was also used. In general, if you see the wrong df, it doesn't mean that you can correct the issue by recomputing the df because it may be that the wrong standard error was also used.

```
# for lme
tvals <- summary(out.lme)[[20]][2:4, 4]
tvals

##         m1         m2        int
## -0.6445301 -1.2151414 -0.7608860

(1 - pt(abs(tvals), 11)) * 2

##        m1        m2       int
## 0.5324423 0.2497504 0.4627244

# for gls
tvals <- summary(out.gls)[[18]][2:4, 3]
tvals

##         m1         m2        int
## 0.90452939 1.00000000 0.04525917

(1 - pt(abs(tvals), 11)) * 2

##        m1        m2       int
## 0.3850972 0.3388007 0.9647119
```

To convince you that this is simply a degrees of freedom issue, let me take the same t values and reproduce the lme pvalues with 33 degrees of freedom and the gls degrees of freedom with 44 degrees of freedom.

```
# reproduce lme pvalues with 33 df
(1 - pt(abs(tvals), 33)) * 2

##        m1        m2       int
## 0.3722712 0.3245871 0.9641735
```

```
# reproduce lme pvalues with 44 df
(1 - pt(abs(tvals), 44)) * 2


##        m1        m2       int
## 0.3706409 0.3227785 0.9641055
```

Here is how to do contrasts in lme but using a time factor rather than building the contrasts directly into the predictors. This is just a different syntax approach to accomplish the same thing; it reproduces the lme output above also with 33 df so we would need to recompute the p value using 11 rather than 33 degrees of freedom. This form allows for an unconstrained covariance matrix consistent with the covariance matrix used in the multivariate approach, though degrees of freedom are still incorrectly 33 rather than 11.

```
data.oneway <- data.frame(data.oneway)
data.oneway$time <- factor(data.oneway$time)
#specify contrasts to be consistent with previous analyses
contrasts(data.oneway$time) <- cbind(m1=c(1,1,-1,-1),
                    m2=c(1,-1,1,-1),int=c(1,-1,-1,1))
out.lme.time <- lme(dv ˜ time, random = ˜1|sub,
            correlation = corSymm(form = ˜1|sub),
            data=data.oneway, weight=varIdent(form=˜1|time),
            method="REML",
            control=lmeControl(maxIter=100,msMaxIter=200))
summary(out.lme.time)


## Linear mixed-effects model fit by REML
##   Data: data.oneway
##        AIC      BIC    logLik
##   234.3039 261.0667 -102.1519
##
## Random effects:
##  Formula: ˜1 | sub
##         (Intercept) Residual
## StdDev:    9.210681 3.474544
##
## Correlation Structure: General
##  Formula: ˜1 | sub
##  Parameter estimate(s):
##  Correlation:
##   1     2     3
## 2 0.879
## 3 0.745 0.973
## 4 0.551 0.870 0.951
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | time
##  Parameter estimates:
##         1         2         3         4
## 1.0000000 0.9324562 1.3315815 1.4871948
## Fixed effects:  dv ~ time
##                 Value Std.Error DF
## (Intercept) 107.60417 2.8825590 33
## timem1       -0.27083 0.4202027 33
## timem2       -0.27083 0.2228822 33
## timeint      -0.06250 0.0821411 33
##              t-value p-value
## (Intercept) 37.32939  0.0000
## timem1      -0.64453  0.5237
## timem2      -1.21514  0.2329
## timeint     -0.76089  0.4521
##  Correlation:
##         (Intr) timem1 timem2
## timem1  -0.220
## timem2  -0.072  0.905
## timeint -0.055  0.216  0.045
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med
## -1.5744230 -0.8529488 -0.2596430
##         Q3        Max
##  0.9401155  1.6252264
##
## Number of Observations: 48
## Number of Groups: 12


#unfortunately the "multivariate" option
#in emmeans doesn't
#have the same effect it has as when it is used
#with afex that we saw earlier of giving df=11;
#ts are correct but df and pvalues are off
#one can recompute the p values as shown above
out.emmeans <- emmeans(out.lme.time, ~time, model="multivariate")


## Warning:  contrasts dropped from factor time


cont <- list(m1=c(1,1,-1,-1),m2=c(1,-1,1,-1),int=c(1,-1,-1,1))
```

```
contrast(out.emmeans,cont)
```

```
##  contrast estimate    SE df t.ratio
##  m1          -1.08 1.681 33  -0.645
##  m2          -1.08 0.892 33  -1.215
##  int         -0.25 0.329 33  -0.761
##  p.value
##   0.5237
##   0.2329
##   0.4521
##
## Degrees-of-freedom method: containment
```

For completeness here is the version of the lme command that assumes compound symmetry.

```
#this form reproduces the compound symmetry approach
#NOTE HOW THE RANDOM PART HAS SUB/TIME
#based on some websites
out.lme.time <- lme(dv ~ time, random = ~1|sub/time,
            data=data.oneway,
            method="REML",
            control=lmeControl(maxIter=100,msMaxIter=200))
summary(out.lme.time)
```

```
## Linear mixed-effects model fit by REML
##   Data: data.oneway
##        AIC      BIC    logLik
##   263.6485 276.1378 -124.8242
##
## Random effects:
##  Formula: ~1 | sub
##         (Intercept)
## StdDev:    9.938701
##
##  Formula: ~1 | time %in% sub
##         (Intercept) Residual
## StdDev:    1.547317 1.154866
##
## Fixed effects:  dv ~ time
##                 Value Std.Error DF
## (Intercept) 107.60417 2.8825590 33
## timem1       -0.27083 0.2786838 33
## timem2       -0.27083 0.2786838 33
## timeint      -0.06250 0.2786838 33
##                t-value p-value
## (Intercept) 37.32939  0.0000
## timem1      -0.97183  0.3382
## timem2      -0.97183  0.3382
## timeint     -0.22427  0.8239
##  Correlation:
##         (Intr) timem1 timem2
## timem1  0
## timem2  0      0
## timeint 0      0      0
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med
## -1.03179126 -0.33072216  0.01616466
##          Q3         Max
##  0.26117044  1.16255178
##
## Number of Observations: 48
## Number of Groups:
##           sub time %in% sub
##            12            48

#THIS VERSION HAS RANDOM SUB
#output (t, df, and p) is the same but residual
#variance is different than sub/time version above
#need to figure this out
out.lme.time <- lme(dv ~ time, random = ~1|sub,
            data=data.oneway,
            method="REML",
            control=lmeControl(maxIter=100,msMaxIter=200))
summary(out.lme.time)

## Linear mixed-effects model fit by REML
##   Data: data.oneway
##        AIC      BIC    logLik
##   261.6485 272.3536 -124.8242
##
## Random effects:
##  Formula: ~1 | sub
##         (Intercept) Residual
## StdDev:    9.938701 1.930778
##
## Fixed effects:  dv ~ time
##                 Value Std.Error DF
## (Intercept) 107.60417 2.8825590 33
## timem1       -0.27083 0.2786838 33
## timem2       -0.27083 0.2786838 33
## timeint      -0.06250 0.2786838 33
##              t-value p-value
## (Intercept) 37.32939  0.0000
## timem1      -0.97183  0.3382
## timem2      -0.97183  0.3382
## timeint     -0.22427  0.8239
##  Correlation:
##         (Intr) timem1 timem2
## timem1  0
## timem2  0      0
## timeint 0      0      0
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med
## -1.7250141 -0.5529223  0.0270251
##         Q3        Max
##  0.4366413  1.9436278
##
## Number of Observations: 48
## Number of Groups: 12
```

The residual standard deviation in the latter output is 1.93 and the error in the former is partitioned into two chunks (“intercept” and “time % in % sub”). But if you square each to put them in variance form, sum and take sqrt to get back standard error, the latter equals 1.93 (= $\sqrt{(1.547317^2 + 1.154866^2)}$). Right now I favor the sub form of the error term (the latter approach) over the sub/time form (the former approach).

I wish using lme would be easier so one doesn’t have to adjust pvalues manually. The primary advantage of the linear mixed model approach is that it allows for missing data. The traditional repeated measures requires subjects to have data for all times or they are excluded from the analysis. The linear mixed model keeps all available data. It also handles unbalanced designs pretty well and can take advantage of all the generalizations we will add to the mix when we cover regression.

I’ll next use the lmer function in the lme4 package to illustrate a different way to implement the linear mixed model. The data need to be in long format. I’ll make the contrasts into factors this time to illustrate another complication of using R I will discuss right after examining the output.

```
dv <- c(data[, 2], data[, 3], data[, 4], data[, 5])
sub <- factor(rep(1:12, 4))

m1 <- factor(rep(c(1, 1, -1, -1), each = 12))
```

```
m2 <- factor(rep(c(1, -1, 1, -1), each = 12))
int <- factor(rep(c(1, -1, -1, 1), each = 12))
# print to see the data matrix
cbind(sub, dv, m1, m2, int)
```

```
##       sub  dv m1 m2 int
##  [1,]   1  92  2  2   2
##  [2,]   2 120  2  2   2
##  [3,]   3 112  2  2   2
##  [4,]   4  95  2  2   2
##  [5,]   5 114  2  2   2
##  [6,]   6  99  2  2   2
##  [7,]   7 124  2  2   2
##  [8,]   8 106  2  2   2
##  [9,]   9 100  2  2   2
## [10,]  10 108  2  2   2
## [11,]  11 112  2  2   2
## [12,]  12 102  2  2   2
## [13,]   1  95  2  1   1
## [14,]   2 121  2  1   1
## [15,]   3 111  2  1   1
## [16,]   4  96  2  1   1
## [17,]   5 112  2  1   1
## [18,]   6 100  2  1   1
## [19,]   7 125  2  1   1
## [20,]   8 107  2  1   1
## [21,]   9  98  2  1   1
## [22,]  10 110  2  1   1
## [23,]  11 115  2  1   1
## [24,]  12 102  2  1   1
## [25,]   1  96  1  2   1
## [26,]   2 121  1  2   1
## [27,]   3 111  1  2   1
## [28,]   4  98  1  2   1
## [29,]   5 110  1  2   1
## [30,]   6  99  1  2   1
## [31,]   7 127  1  2   1
## [32,]   8 106  1  2   1
## [33,]   9  95  1  2   1
## [34,]  10 112  1  2   1
## [35,]  11 116  1  2   1
## [36,]  12 101  1  2   1
## [37,]   1  98  1  1   2
## [38,]   2 123  1  1   2
## [39,]   3 109   1  1   2
## [40,]   4  99   1  1   2
## [41,]   5 109   1  1   2
## [42,]   6  98   1  1   2
## [43,]   7 126   1  1   2
## [44,]   8 107   1  1   2
## [45,]   9  94   1  1   2
## [46,]  10 115   1  1   2
## [47,]  11 118   1  1   2
## [48,]  12 101   1  1   2

library(lme4)

# this will be compound symmetric assumed; lmer doesn't
# have a way to estimate free covariance matrices (i.e.,
# the correlation and weight arguments in lme)
out.lmer <- lmer(dv ˜ m1 + m2 + int + (1 | sub), REML = T)
summary(out.lmer)

## Linear mixed model fit by REML [lmerMod]
## Formula: dv ˜ m1 + m2 + int + (1 | sub)
##
## REML criterion at convergence: 245.5
##
## Scaled residuals:
##      Min       1Q   Median       3Q
## -1.72501 -0.55292  0.02703  0.43664
##      Max
##  1.94363
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  sub      (Intercept) 98.778   9.939
##  Residual              3.728   1.931
## Number of obs: 48, groups:  sub, 12
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 108.2083     2.9227  37.023
## m11          -0.5417     0.5574  -0.972
## m21          -0.5417     0.5574  -0.972
## int1         -0.1250     0.5574  -0.224
##
## Correlation of Fixed Effects:
```

```
##       (Intr) m11    m21
## m11  -0.095
## m21  -0.095  0.000
## int1 -0.095  0.000  0.000
```

There are no pvalues or degrees of freedom listed in this output. That is because the developer of the lme4 package recognizes that it is not easy to write general code to do this correctly across different types of packages. Other people have written additional packages to add onto lme4 (like pbkrtest, lmertest, lmerconveniencefunctions), but these should be used with caution and test them on data where you know the correct answer.

Here is the version with p-values that emerges from using the lmer function (which is a tweaked version of the original lmer in the lme4 package) in the lmerTest package. The degrees of freedom are 33 instead of 11. R just can't seem to get this right.

```
library(lmerTest)
out.lmer <- lmer(dv ˜ m1 + m2 + int + (1 | sub), REML = T)
summary(out.lmer)


## Linear mixed model fit by REML.
##   t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: dv ˜ m1 + m2 + int + (1 | sub)
##
## REML criterion at convergence: 245.5
##
## Scaled residuals:
##      Min       1Q   Median       3Q
## -1.72501 -0.55292  0.02703  0.43664
##      Max
##  1.94363
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  sub      (Intercept) 98.778   9.939
##  Residual              3.728   1.931
## Number of obs: 48, groups:  sub, 12
##
## Fixed effects:
##             Estimate Std. Error
## (Intercept) 108.2083     2.9227
## m11          -0.5417     0.5574
## m21          -0.5417     0.5574
## int1         -0.1250     0.5574
##                    df t value Pr(>|t|)
## (Intercept)  11.6225  37.023    2e-13
## m11          33.0000  -0.972    0.338
## m21          33.0000  -0.972    0.338
## int1         33.0000  -0.224    0.824
##
## (Intercept) ***
## m11
## m21
## int1
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr) m11    m21
## m11  -0.095
## m21  -0.095  0.000
## int1 -0.095  0.000  0.000
```

```
# one could do this in terms of a time factor and then
# specify contrasts to get equivalent results; df are 33
# here because we have the compound symmetry form
out.lmer <- lmer(dv ˜ time + (1 | sub), REML = T)
library(emmeans)
out.emmeans <- emmeans(out.lmer, ˜time, lmer.df = "satterthwaite")
cont <- list(m1 = c(1, 1, -1, -1), m2 = c(1, -1, 1, -1), int = c(1,
    -1, -1, 1))
contrast(out.emmeans, cont)
```

```
##  contrast estimate   SE df t.ratio
##  m1           -1.08 1.11 33  -0.972
##  m2           -1.08 1.11 33  -0.972
##  int          -0.25 1.11 33  -0.224
##  p.value
##   0.3382
##   0.3382
##   0.8239
##
## Degrees-of-freedom method: satterthwaite
```

An Aside:

There is an R detail here to point out about the use of factors. Even though three contrasts were defined as factors with values of 1 and -1, when they are printed as numerical values (like after a cbind command) the original numbers are converted to integers starting with 1 as in 1s and 2s in the case of two groups. When factors are used in formulas, their contrast values are used rather than the numerical codes used for the levels. So a weird thing can happen when using R. A factor can have one set of values, a different set of values when it is converted into numerical values, and a third set of values when used in a regression command like lmer. So, yet another warning for R users: be careful when interpreting output from a statistics command like a regression. For example,

```
# define test as 1s and -1s
test <- c(1, 1, 1, -1, -1, -1)
test

## [1]  1  1  1 -1 -1 -1

# convert to factor, remains as 1s and -1s
test.factor <- factor(test)
test.factor

## [1] 1  1  1  -1 -1 -1
## Levels: -1 1

# put into a matrix, test remains 1s and -1s but
# test.factor is converted to 1s and 2s
cbind(test, test.factor)

##      test test.factor
## [1,]    1           2
## [2,]    1           2
## [3,]    1           2
## [4,]   -1           1
## [5,]   -1           1
## [6,]   -1           1

# regressions like lm, lmer, glm, etc., use the contrast
# codes of factor, which default to dummy codes 1s and 0s;
# a different set of contrasts than maybe one would expect
# with 1 and -1
contrasts(test.factor)

##    1
## -1 0
## 1  1
```

General point: be careful. You can see that one can have different results for a simple one-way repeated measures ANOVA just by making tiny changes in the commands. Sometimes a syntax depends on specific details such as how factors and contrasts are defined; sometimes a syntax looks like it should be right but it doesn't necessarily do the right thing.

Two-way repeated measure ANOVA

Be sure to work through the previous subsection in this R Appendix for the case of a one-way repeated measures ANOVA. I won't re-explain issues of multivariate and univariate approaches (i.e., having a free or constrained covariance matrix).

Two-way repeated measures

Here is an example with two within subject factors. For simplicity, I took the one way ANOVA data with 4 times and arranged it as a 2x2 repeated measures anova. There are two factors (m1 and m2) each with two levels. We'll examine the two main effects and the interaction.

```
library(afex)
summary(aov_ez("sub", "dv", data = data.oneway, within = c("m1",
    "m2")))


##
## Univariate Type III Repeated-Measures ANOVA Assuming Sphericity
##
##               Sum Sq num Df Error SS
## (Intercept) 555776      1   4387.2
## m1               4      1     93.2
## m2               4      1     26.2
## m1:m2            0      1      3.6
##             den Df   F value    Pr(>F)
## (Intercept)     11 1393.4833 6.154e-13
## m1              11    0.4154    0.5324
## m2              11    1.4766    0.2498
## m1:m2           11    0.5789    0.4627
##
## (Intercept) ***
## m1
## m2
## m1:m2
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
```

This yields the correct 3 p-values per the t tests I did earlier, the one where I manually computed the contrast values for each subject and performed one sample t tests; it happens that those three contrasts on the one way ANOVA with 4 time points correspond to the same contrasts as the main effects and interaction in the 2x2 repeated measures ANOVA. But this result has 0 sum of squares for the interaction, which can't be right if the F is .57 likely because of roundoff error. We see from the SPSS output that the sum of squares for the interaction is .19. The p-values are correct so the end computation is right. We can force R to print more significant digits through this command (the fourth element in the summary list contains the source table, and we wrap it around the print command to increase the number of digits printed).

```
print(summary(aov_ez("sub", "dv", data = data.oneway, within = c("m1",
    "m2")))[[4]], digits = 10)


##                   Sum Sq num Df
## (Intercept) 555775.5208      1
## m1               3.5208      1
## m2               3.5208      1
## m1:m2            0.1875      1
##                 Error SS den Df
## (Intercept) 4387.229167     11
## m1            93.229167     11
## m2            26.229167     11
## m1:m2          3.562500     11
##                F value     Pr(>F)
## (Intercept) 1393.48333 6.1539e-13 ***
## m1             0.41542    0.53244
## m2             1.47657    0.24975
## m1:m2          0.57895    0.46272
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
```

Note that for a two way ANOVA where each factor has two levels, the `aov_ez` command produced the correct pvalues, but recall that for the same data represented as a one-way ANOVA with four levels the same program imposed the compound symmetry assumption and yielded incorrect t tests and p values. Another weird aspect about repeated measures in R. I'll leave it as homework to go through the analogous series of analyses I did above for the one-way repeated measures ANOVA for this two-factor repeated measures problem and compare the results across different programs (e.g., lme, lmer, gls, contrasts with emmeans, etc). There won't be much difference between the results of these analyses because for a 2 x 2 things become very trivial. It would be good to compare the different approaches on a more complex design such as a 3 x 3 repeated measures design. It would also be useful to compare to SPSS MANOVA results, which we know are correct.

One Between and One Repeated Measures factor

Check out the earlier subsection on one-way repeated measures ANOVA (omnibus tests and tests of contrasts) as I won't repeat some of those issues here.

one between and one repeated measures

Finally, an example with one repeated measures and one between subjects. You could just do the contrasts on the sum and difference like I show in the lecture notes (being mindful of the degrees of freedom), but if you want to do this with R packages here is how you would go about it. There are fancier ways of moving data between wide format and long format as described earlier, here I just do it the bonehead way so you can see clearly what is going on. I illustrate with lmer in the lme4 package, replicate with mixed in the afex package, replicate again with lme from the nlme package, and finally replicate with the Anova function from the car package (interesting way of converting a regular linear model from lm into a mixed design). These are just for illustration; there are other packages that do this too and you may find them easier. I suggest you learn all the options in the program you choose (lmer, etc) and try out several data sets from textbooks you trust until you can recreate the example output. Next semester I will share a very concise function that allows us to test any contrast in any repeated measures, between subjects or mixed designs. It uses some concepts that will be covered in Lecture Notes 10 and 11. It is what I typically use in my own data; or you can simply use SPSS manova and get the correct results. I do not recommend you just run one program in R because that is what you may know. It is likely that you'll have an incorrect answer unless you know exactly what you are doing.

So, what follows is five different ways of running the same analyses. Each of the out files are numbered 1 to 5 (out1, out2, etc).

```
data <- cbind(1:16, rep(1:2,each=8),
c(8.6,7.5,8.3,8.4,6.4,6.9,6.5,6,7.3,7.5,6.4,6.8,7.1,8.2,7.2,6.7),
c(8,7.1,7.4,7.3,6.4,6.8,6.1,5.7,7.9,7.6,6.3,7.5,7.7,8.6,7.8,6.9))
#data
longdata <- rbind(data[,-4],data[,-3])
longdata <- cbind(longdata,rep(1:2,each=16))
colnames(longdata) <- c("sub", "sequence", "dv", "period")
longdata <- data.frame(longdata)
longdata$sequence <- factor(longdata$sequence)
longdata$period <- factor(longdata$period)
longdata$sub <- factor(longdata$sub)
#some functions like gee require data to be ordered by subject
#longdata <- longdata[order(longdata$sub),]
longdata


##    sub sequence  dv period
## 1     1        1 8.6      1
## 2     2        1 7.5      1
## 3     3        1 8.3      1
## 4     4        1 8.4      1
## 5    5        1 6.4      1
## 6    6        1 6.9      1
## 7    7        1 6.5      1
## 8    8        1 6.0      1
## 9    9        2 7.3      1
## 10  10        2 7.5      1
## 11  11        2 6.4      1
## 12  12        2 6.8      1
## 13  13        2 7.1      1
## 14  14        2 8.2      1
## 15  15        2 7.2      1
## 16  16        2 6.7      1
## 17   1        1 8.0      2
## 18   2        1 7.1      2
## 19   3        1 7.4      2
## 20   4        1 7.3      2
## 21   5        1 6.4      2
## 22   6        1 6.8      2
## 23   7        1 6.1      2
## 24   8        1 5.7      2
## 25   9        2 7.9      2
## 26  10        2 7.6      2
## 27  11        2 6.3      2
## 28  12        2 7.5      2
## 29  13        2 7.7      2
## 30  14        2 8.6      2
## 31  15        2 7.8      2
## 32  16        2 6.9      2
```

```
library(car)
#first with lmer in the lme4 package;
#replicates MANOVA output in SPSS, Anova is in the car package
library(lme4)
out1 <- lmer(dv˜ period*sequence + (1|sub),REML=T,data=longdata)
temp <- Anova(out1, type="II", test.statistic="F")
temp
```

```
## Analysis of Deviance Table (Type II Wald F tests with Kenward-Roger df)
##
## Response: dv
##                       F Df Df.res
## period           0.2707  1     14
## sequence         0.4637  1     14
## period:sequence 26.3039  1     14
##                     Pr(>F)
## period          0.6109803
## sequence        0.5070293
## period:sequence 0.0001534 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1

#demonstrate that type="III" yields an incorrect main effect
Anova(out1, type="III", test.statistic="F")

## Analysis of Deviance Table (Type III Wald F tests with Kenward-Roger df)
##
## Response: dv
##                        F Df Df.res
## (Intercept)     721.691  1 15.394
## period           15.956  1 14.000
## sequence          0.206  1 15.394
## period:sequence  26.304  1 14.000
##                     Pr(>F)
## (Intercept)     2.359e-14 ***
## period          0.0013304 **
## sequence        0.6562852
## period:sequence 0.0001534 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1

#example of doing contrasts on lmer;
#note df=14 for all three tests as expected;
#the emmeans package is impressing me, recall
#how it gave the correct results in the one way
#repeated measures anova with 4 levels
library(emmeans)
out.emmeans <- emmeans(out1, ~ period*sequence,
                       model="multivariate")
#print the means so we know the order to
#specify contrasts (though ignore df in this part)
out.emmeans

##  period sequence emmean    SE   df
##  1      1          7.33 0.273 15.4
##  2      1          6.85 0.273 15.4
##  1      2          7.15 0.273 15.4
##  2      2          7.54 0.273 15.4
##  lower.CL upper.CL
##      6.75     7.90
##      6.27     7.43
##      6.57     7.73
##      6.96     8.12
##
## Degrees-of-freedom method: kenward-roger
## Confidence level used: 0.95
```

```
#test the contrasts that you specify
contrast(out.emmeans, list(me.period =c(1,-1,1,-1),
                           me.sequence=c(1,1,-1,-1),
  interaction=c(1,-1,-1,1)))
```

```
##  contrast    estimate    SE df t.ratio
##  me.period     0.0875 0.168 14   0.520
##  me.sequence  -0.5125 0.753 14  -0.681
##  interaction   0.8625 0.168 14   5.129
##  p.value
##   0.6110
##   0.5070
##   0.0002
##
## Degrees-of-freedom method: kenward-roger
```

This reproduced the correct tests for the contrasts, even using different error terms for the between and the within parts.

Now switch to another approach using afex.

```
#second with the package afex, replicates MANOVA in SPSS
library(afex)
out2 <- mixed(dv˜ period*sequence + (1|sub),longdata,
              method="KR")
out2
```

```
## Mixed Model Anova Table (Type 3 tests, KR-method)
##
```

```
## Model: dv ~ period * sequence + (1 | sub)
## Data: longdata
##             Effect    df         F
## 1           period 1, 14      0.27
## 2         sequence 1, 14      0.46
## 3 period:sequence 1, 14 26.30 ***
##   p.value
## 1    .611
## 2    .507
## 3   <.001
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '+' 0.1 ' ' 1

#summary(out2)

#third with the package nlme, replicates MANOVA output in SPSS
library(nlme)
out3 <- lme(fixed=dv ~ sequence*period, random=~1|sub, data=longdata)
print(anova(out3))

##                 numDF denDF   F-value
## (Intercept)         1    14 1470.5256
## sequence            1    14    0.4637
## period              1    14    0.2707
## sequence:period     1    14   26.3038
##                 p-value
## (Intercept)      <.0001
## sequence         0.5070
## period           0.6110
## sequence:period  0.0002

#fourth approach with car package and lm
data <- data.frame(data)
colnames(data) <- c("sub","sequence","P1","P2")
period.factor <- factor(c("period1","period2"))
idata <- data.frame(period.factor=period.factor)
temp <- lm(cbind(P1,P2) ~sequence, data=data)
out4 <- Anova(temp, idata=idata, idesign=~period.factor, type="II")
summary(out4)

##
```

```
## Type II Repeated Measures MANOVA Tests:
##
## ------------------------------------------
##
## Term: (Intercept)
##
##  Response transformation matrix:
##    (Intercept)
## P1           1
## P2           1
##
## Sum of squares and products for the hypothesis:
##             (Intercept)
## (Intercept)    3332.176
##
## Multivariate Tests: (Intercept)
##                  Df test stat approx F
## Pillai            1   0.99057 1470.522
## Wilks             1   0.00943 1470.522
## Hotelling-Lawley  1 105.03726 1470.522
## Roy               1 105.03726 1470.522
##                  num Df den Df
## Pillai                1     14
## Wilks                 1     14
## Hotelling-Lawley      1     14
## Roy                   1     14
##                      Pr(>F)
## Pillai           1.3955e-15 ***
## Wilks            1.3955e-15 ***
## Hotelling-Lawley 1.3955e-15 ***
## Roy              1.3955e-15 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
##
## ------------------------------------------
##
## Term: sequence
##
##  Response transformation matrix:
##    (Intercept)
## P1           1
## P2           1
```

```
##
## Sum of squares and products for the hypothesis:
##             (Intercept)
## (Intercept)    1.050625
##
## Multivariate Tests: sequence
##                  Df test stat  approx F
## Pillai            1 0.0320563 0.4636511
## Wilks             1 0.9679437 0.4636511
## Hotelling-Lawley  1 0.0331179 0.4636511
## Roy               1 0.0331179 0.4636511
##                  num Df den Df  Pr(>F)
## Pillai                1     14 0.50703
## Wilks                 1     14 0.50703
## Hotelling-Lawley      1     14 0.50703
## Roy                   1     14 0.50703
##
## ------------------------------------------
##
## Term: period.factor
##
##  Response transformation matrix:
##    period.factor1
## P1              1
## P2             -1
##
## Sum of squares and products for the hypothesis:
##                period.factor1
## period.factor1       0.030625
##
## Multivariate Tests: period.factor
##                  Df test stat  approx F
## Pillai            1 0.0189702 0.2707182
## Wilks             1 0.9810298 0.2707182
## Hotelling-Lawley  1 0.0193370 0.2707182
## Roy               1 0.0193370 0.2707182
##                  num Df den Df  Pr(>F)
## Pillai                1     14 0.61098
## Wilks                 1     14 0.61098
## Hotelling-Lawley      1     14 0.61098
## Roy                   1     14 0.61098
##
## ------------------------------------------
##
```

```
## Term: sequence:period.factor
##
##  Response transformation matrix:
##    period.factor1
## P1              1
## P2             -1
##
## Sum of squares and products for the hypothesis:
##                 period.factor1
## period.factor1       2.975625
##
## Multivariate Tests: sequence:period.factor
##                  Df test stat approx F
## Pillai            1 0.6526388 26.30387
## Wilks             1 0.3473612 26.30387
## Hotelling-Lawley  1 1.8788477 26.30387
## Roy               1 1.8788477 26.30387
##                  num Df den Df
## Pillai                1     14
## Wilks                 1     14
## Hotelling-Lawley      1     14
## Roy                   1     14
##                      Pr(>F)
## Pillai           0.00015337 ***
## Wilks            0.00015337 ***
## Hotelling-Lawley 0.00015337 ***
## Roy              0.00015337 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
##
## Univariate Type II Repeated-Measures ANOVA Assuming Sphericity
##
##                          Sum Sq num Df
## (Intercept)            1666.09      1
## sequence                  0.53      1
## period.factor             0.02      1
## sequence:period.factor    1.49      1
##                        Error SS den Df
## (Intercept)             15.8619     14
## sequence                15.8619     14
## period.factor            0.7919     14
## sequence:period.factor   0.7919     14
##                           F value
## (Intercept)            1470.5216
## sequence                  0.4637
## period.factor             0.2707
## sequence:period.factor   26.3039
##                            Pr(>F)
## (Intercept)            1.395e-15 ***
## sequence               0.5070293
## period.factor          0.6109803
## sequence:period.factor 0.0001534 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
```

```
#fifth approach using the lmerTest package which redefines the lmer function
#no need to use the Anova() function from the car package; lmerTest
#redefines the anova function that works on the lmer object
library(lmerTest)
out5 <- lmer(dv˜ period*sequence + (1|sub),REML=T,data=longdata)
temp <- anova(out5)
temp
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##                  Sum Sq Mean Sq NumDF
## period          0.01531 0.01531     1
## sequence        0.02623 0.02623     1
## period:sequence 1.48781 1.48781     1
##                 DenDF F value    Pr(>F)
## period             14  0.2707 0.6109803
## sequence           14  0.4637 0.5070293
## period:sequence    14 26.3039 0.0001534
##
## period
## sequence
## period:sequence ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
```

```
anova(out5,  type=2)
```

```
## Type II Analysis of Variance Table with Satterthwaite's method
##                    Sum Sq Mean Sq NumDF
## period            0.01531 0.01531     1
## sequence          0.02623 0.02623     1
## period:sequence 1.48781 1.48781     1
##                  DenDF F value    Pr(>F)
## period              14  0.2707 0.6109803
## sequence            14  0.4637 0.5070293
## period:sequence    14 26.3039 0.0001534
##
## period
## sequence
## period:sequence ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
```

```
#but Anova with type="III" yields an incorrect main effect
#even though anova redefined by lmerTest uses Type="III"
#and gives the correct result
#we saw issue in the first analysis in this series using
#lmer in the lme4 package
Anova(out5,type="III",test.statistic ="F")
```

```
## Analysis of Deviance Table (Type III Wald F tests with Kenward-Roger df)
##
## Response: dv
##                         F Df Df.res
## (Intercept)     721.691  1 15.394
## period           15.956  1 14.000
## sequence          0.206  1 15.394
## period:sequence  26.304  1 14.000
##                       Pr(>F)
## (Intercept)     2.359e-14 ***
## period          0.0013304 **
## sequence        0.6562852
## period:sequence 0.0001534 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
```

The first run with lmer (out1) shows degrees of freedom based on Kenward-Roger. This is an alternative approach and extends the Satterthwaite extension of the Welch correction for degrees of freedom. Some R commands let you choose between Kenward-Roger and Satterthwaite. In the first run for out1 I used lmer but I had to run Anova() with type="II" rather than type="III" because otherwise one of the main effects are off. I don't understand why this happens when in the fifth approach (out5) I also ran lmer() as implemented in the lmerTest package, which uses type="III" by default, but in this case type="III" gave the correct results. This demonstrates why it is so important to run a program through examples with known results so you can evaluate the output. I wouldn't have been able to diagnose this type="III" issue had I ran an actual data set in my research where I didn't already know the right answer.

This is a relatively easy example because both factors have two levels. You'll need to double check that this all works in your setting where the repeated measures factor may have more than two levels and the R command you are using may invoke additional assumptions like compound symmetry. I suspect things will get even messier in the case where factors have more than two levels.

What a mess. So many different ways to do this in R, and one needs to be careful because each different approach requires one to be mindful of a particular set of details: the assumptions on the covariance matrix, the resulting error term and the corresponding degrees of freedom.