# Chapter 8

# Analysis of Variance II

CONTENTS



## 8.1 OVERVIEW

Chapter 7 covered the basics of ANOVA. Recall that ANOVA is structured as a two-step process. First, F-tests are used to determine whether there is evidence of differences in **treatment** means at the population level. Then, if differences are found, pairwise comparison procedures are used to identify the nature of those differences. In particular, it is important to incorporate confidence intervals to assess the magnitude of differences and make clinical application of results. In this chapter we will now consider some more advanced topics in ANOVA and experimental design, including:

- Repeated Measures ANOVA (a.k.a. ANOVA with blocks)
- Two-factor interaction model; analyzing interactions
- Unbalanced design; potential for confounded factors
- Multi-factor models
- Individual prediction
- Random effects

**What is a "treatment"?**

The statistical (not medical) definition of "treatment" becomes much more important in multi-factor ANOVA, as the overall F-test examines differences in "treatment" averages. A **treatment** is a *particular combination* of levels for **all** factors involved in a study. For example, suppose that a study considers three factors including gender, exercise level, and intervention. In such a study, one treatment might be "male, active, receiving placebo". Another might be "female, very active, receiving medication A". It is important to understand that a difference in treatments means for such a study could be associated to gender, exercise level, intervention or any combination of the three. An overall F-test for the model will not distinguish where differences lie; it can only identify that some exist. Term-by-term F-tests also exist and these do help us to distinguish which factors are involved and should be considered for analysis of pairwise comparisons to more fully identify the nature of differences.

## 8.2 REPEATED MEASURES ANOVA

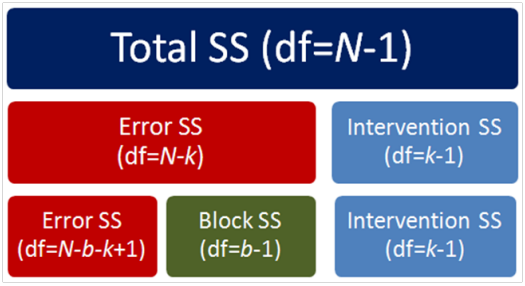
**Repeated measures ANOVA** (RM-ANOVA) employs a design that is similar to the matched pairs T-test. The difference is that experimental units will each be measured multiple times, generally once for each intervention. This type of analysis is also sometimes called "ANOVA with blocks", referring to the fact that (just like in a matched pairs study) variability due to the participant is "blocked out" of the analysis. As was the case with matched pairs, this reduction in variability allows for greater statistical power, yielding more precise estimation related to the interventions under study.

While the classic example involves people who are being measured multiple times on different interventions, the experimental unit certainly can be something other than people. For example, perhaps one wishes to study the effectiveness of a four laboratory procedures, all of which involve the use of a centrifuge. Further suppose that each of 30 available centrifuges holds four vials. By running each centrifuge with one vial corresponding to each lab procedure, the individual centrifuges may be used as a blocking variable, thus removing variability associated to the centrifuge itself and also perhaps other conditions under which each centrifuge is used.

## 8.2.1 MODIFYING THE ANOVA TABLE AND F-TESTS

Ultimately any repeated measures variable is simply considered an additional **factor** in the RM-ANOVA. In a repeated measures design, each of $k$ interventions should occur exactly one time in each of $b$ blocks; meaning that there will be $N = k*b$ measurements. Table 8.1 illustrates the break-down of sums of squares and degrees of freedom from a one-factor ANOVA (row 2) into a two-factor design with the incorporation of the blocking factor (row 3). Notice that it is the Error SS (red) that will be reduced by the blocking factor (green) – meaning that our statistical power should be substantially improved whenever the blocking factor is indeed related to the response variable of interest. Note also that if the blocking factor is unimportant the Error SS would still be reduced, but in a manner that is proportional to the degrees of freedom ($b - 1$) so that the Mean Square Error (which represents error variability) would not be greatly affected. Refer back to Section 7.5 for a review of the basic concepts for SS and MS.

Table 8.1 Breakdown of SS




**Algebraic Note**

In this commentary and in Table 8.2, the letters $b$, $k$, and $N$ are algebraic symbols representing actual counts for a specific experiment. Likewise, $MS_{INT}$, SSE, etc. would represent actual numbers (that would be computed using technology). It is not necessary to memorize these things, but rather it is appropriate to come to an understanding of how this table works. In particular, advance consideration of sample size and its impact on error degrees of freedom is one of the main goals of a power analysis.

The RM-ANOVA table (Table 8.2) is likewise broken into three lines instead of two. When using a repeated measures factor, one does so because of a firm belief that factor will be important in explaining variability in the response. Thus the F-test associated with the blocks is generally expected to show a significant p-value. A significant test result for the blocking factor simply confirms that blocking was useful. There is little interest in actually comparing the blocks, as comparing specific individuals from the sample does not extend to the population. For this reason, the blocking F-test is seldom discussed as part of the analysis. Note also that if this F-test turns out to show lack of evidence of the effectiveness of blocking – that does not mean that the blocks should be removed from the model. In fact, their removal could lead to a small amount of bias due to post-hoc change in the analysis plan (the result of removal would be that MSE would change in a small, but random, way).

Table 8.2 ANOVA table for one-way repeated measures design.

| Source | DF | SS | MS** | F-ratio andp-value |
|---|---|---|---|---|
| Blocks | $b-1$ | $SS_{BLK}$ | $MS_{BLK}$ | $F = MS_{BLK} / MSE$ |
| **Interventions** | $k-1$ | $SS_{INT}$ | $MS_{INT}$ | $F = MS_{INT} / MSE$ |
| Error | $N-k-b+1$ | SSE | MSE | |
| Total | $N-1$ | Total SS | | |

** MS are calculated as SS divided by DF.

The F-test for the interventions is the test of interest. As in one-factor ANOVA, that F-test assesses whether or not differences exist among the intervention means at the population level. Possible conclusions from this F-test are:

- Lack evidence of any differences in the population means across all interventions.
- Find evidence that the population means for at least two of the interventions are different.

If differences are found, the next appropriate step will again be pairwise comparisons; the structure is nearly identical to one-factor ANOVA. The only difference will be that variability is lessened due to the blocking variable and therefore pairwise comparison confidence intervals should be more precise. Any pairwise comparison procedure discussed in Chapter 7 could be similarly applied to an RM-ANOVA model.

**Basic RM-ANOVA Procedure**:

1. Check assumptions of the model.
2. Produce ANOVA table and check for differences in the population means for the factor of interest (Blocking factor should be significant by design; its p-value is not of interest)
3. If differences are found, use an appropriate multiple comparison procedure to compare means for the factor of interest.

## 8.2.2 ASSUMPTIONS IN RM-ANOVA

The response variable for any ANOVA needs to be measured at the interval-ratio level. Repeated measures ANOVA makes the same assumptions as a one-factor ANOVA as relates to the error component of the model – namely that errors are independent, normally distributed, and homogeneity of variance applies. Additionally, RM-ANOVA typically makes one additional assumption that there is no **interaction** between the blocking factor and the interventions**.** Quite often, this assumption is referred to as **additivity**. Fundamentally, this means that the interventions have the same impact regardless of which block is chosen. Of course in most medical studies, the blocking factor will be the patients. The additivity assumption would be violated, for example, if some patients are better off with one intervention while other patients respond more favorably to another. Consider if this were the case, when we look at averages as

we do in ANOVA, there may be a third intervention that will appear to be the best (when in fact the third can always be bettered by one of the first two, the choice of which depends on the patient). If interaction like this is expected, one should not set up a repeated measures model. Instead a search for other characteristics (e.g. gender, physical attributes, etc.) that might help to assign the best intervention would be more appropriate.

The additivity assumption is not particularly easy to assess in a formal manner. We must determine whether or not **we can expect the interventions to be best/worst in similar patterns across all participants**. If the answer is no, the expectation is that there will be interaction and a different experimental design would be needed. The additivity assumption can be checked to some extent with an **interaction plot**. In the case of RM-ANOVA, an interaction plot displays the data (we will later see for multi-factor ANOVA, interaction plots will display sample means). In general, an interaction plot for the purpose of RM-ANOVA assumption checking is a plot of the data having the following attributes:

- Response variable plotted on the vertical axis.
- Repeated Measures factor plotted on the horizontal axis (the order in which participants are listed generally wouldn't matter).
- Interventions represented by different colors and lines.

If there is no interaction, "trends" as represented by the colored line seen in the plot will generally be reasonably the same (e.g. if the "red" intervention is best for one subject, by around three units over the "orange" intervention, it would be best by around the same amount for all subjects). Further discussion of interaction plots will be found as part of the example in Section 8.3 as well as in a discussion of general two-factor ANOVA beginning in Section 8.4.

### 8.2.3 RM-ANOVA AS BALANCED DESIGN – DF IMPLICATIONS

The astute reader may have noticed that a repeated measures design will typically have a **balanced design** with one **replication**, meaning that every **treatment** (subject-intervention combination, in this case) appears exactly once in the study. The total sample size is therefore the product of the number of blocks and the number of interventions: $N = b*k$. Incidentally, this is the primary reason that RM-ANOVA requires the additivity assumption. As we will later discuss, an interaction term would require allocation of $(b\text{-}1)*(k\text{-}1)$ degrees of freedom to estimate. In a model where the number of observations is $N = b*k$, that allocation would mean that zero degrees of freedom would remain to estimate error (this would be catastrophic as it would ensure that no model estimation would be possible).

**Algebraic Note**

The RM-ANOVA in Table 8.2 has $N - k - b + 1$ degrees of freedom for error. Because $N = b*k$, it is not difficult to show algebraic equivalence to the $(b\text{-}1)*(k\text{-}1)$ degrees of freedom necessary to estimate interaction.

## 8.3 A REPEATED MEASURES ANOVA EXAMPLE

In section 7.8, we discussed a fictitious study that examined the effect of four different diets on the total length of nightly REM sleep cycles. In this study, we concluded that diets 2 and 3 were superior to diets 1 and 4. Diet 4, due to the large amount of difference when compared to 2 and 3, would likely not be advised in any circumstance. But suppose that diet 1 is cheaper and far more convenient when compared to diets 2 and 3. The analysis in section 7.8 showed that diets 2 and 3 resulted in greater average REM sleep; the confidence intervals indicate this average is better by roughly 20 to 60 minutes. It also suggests that diets 2 and 3 might have as much as 24 minutes difference in their true population averages. What can we do if we want to pin these numbers down more precisely? There are two possibilities in constructing a follow-up study:

- **Increase sample size**: We had 200 participants in this study already (50 in each group). Increasing sample size would yield tighter confidence intervals; but we would need to increase sample size by large amounts (probably well into the thousands) in order to make serious headway. Such a large study may not be feasible.

- **Use repeated measures**: Instead of 200 participants, we may get by with only 50 if we lengthen the study. The original study was one month in length; but each participant was measured on only one of the four diets. For the follow-up, suppose we design a seven month study in which participants will engage, at different times, in all four diets. The particulars: each participant will be randomly assigned an order in which they will participate in the four diets, and they do so during months 1, 3, 5, and 7. Their REM sleep will be measured at the end of **each** of those months. Note that we should allow them to eat whatever they normally would during months 2, 4, and 6 – these are called **washout periods** and help to ensure that the previous treatment doesn't affect results for the current treatment. The best part of this scheme is that we likely get greater precision while at the same time using only 25% of the original participants.

The lower half of Table 8.3 illustrates what might happen if we follow-up with the repeated measures study. The upper half shows the original results from the one-factor ANOVA in Section 7.8 for comparison; making this comparison should help you to understand the clear benefits that repeated measures can add to the design. The items of note:

- $R^2$ jumps by a large amount for the RM-ANOVA study. This is an indication that repeated measures in this case is very successful.
- Mean Square Error (MSE) drops from 1737 in the first model to 137 in the RM-ANOVA model. This number controls the standard errors for confidence intervals, which is why confidence intervals in the RM case are only about 12 minutes wide

Most importantly, we now have much better information comparing diets 1, 2, and 3. We are still unable to tell a difference between diet 2 and diet 3. Diet 2 results in an average of at least 24 minutes additional sleep as compared to Diet 1. Diet 3 results in an average of at least 41 minutes additional sleep as compared to Diet 1. And Diets 2 and 3 can now be differentiated – Diet 3 is better by an average between 11 and 23 minutes. Diet 4 is worse than Diet 3 by at least an hour

**Table 8.3. One-way ANOVA vs. Repeated Measures ANOVA comparison**

**Standard One-Factor ANOVA**

Analysis of Variance  R-squared = 36.14%

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Diet | 3 | 192636 | 64212 | 36.97 | 0.000 |
| Error | 196 | 340424 | 1737 | | |
| Total | 199 | 533060 | | | |

| Difference | 95% CI |
|---|---|
| Diet 2 - Diet 1 | ( 17.11, 60.25) |
| Diet 3 - Diet 1 | ( 20.03, 63.17) |
| Diet 4 - Diet 1 | (-55.63, -12.49) |
| Diet 3 - Diet 2 | (-18.65, 24.49) |
| Diet 4 - Diet 2 | (-94.31, -51.17) |
| Diet 4 - Diet 3 | (-97.23, -54.09) |

**Repeated Measures ANOVA**

Analysis of Variance  R-squared = 94.82%

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Patient | 49 | 193047 | 3939.7 | 28.76 | 0.000 |
| Diet | 3 | 175530 | 58509.9 | 427.19 | 0.000 |
| Error | 147 | 20134 | 137.0 | | |
| Total | 199 | 388711 | | | |

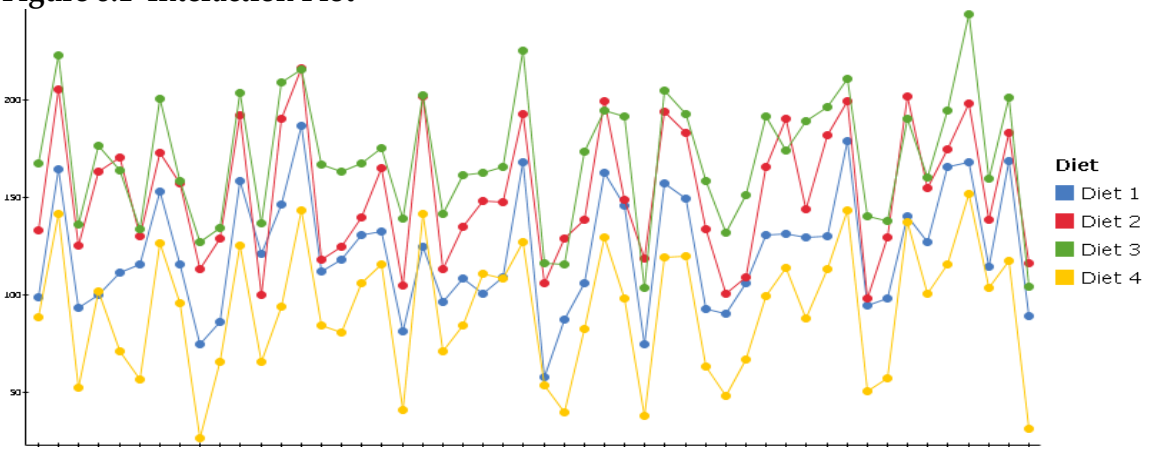
| Difference | 95% CI |
|---|---|
| Diet 2 - Diet 1 | ( 24.30, 36.48) |
| Diet 3 - Diet 1 | ( 41.53, 53.72) |
| Diet 4 - Diet 1 | (-36.03, -23.85) |
| Diet 3 - Diet 2 | ( 11.14, 23.32) |
| Diet 4 - Diet 2 | (-66.43, -54.24) |
| Diet 4 - Diet 3 | (-83.66, -71.48) |

***Output shown is from Minitab®**

(on average), and at this point may clearly be removed from consideration. In comparing Diets 1, 2, and 3, we now have a clear ranking and a reasonably good idea of the magnitude of differences between the three diets. It should now be fairly simple to combine this information with costs and make recommendations to patients.

Before leaving this example, let's look briefly at the additional validity condition required for RM-ANOVA. One way to check additivity is to look at an interaction plot (see section 8.2.2). This plot for this example is shown in Figure 8.1. The 50 patients are listed along the horizontal axis here (note that the order of this list is arbitrary). As you view the four colored lines representing the diets, you'll note that for the most part they traverse similar patterns as you go from one patient to the next. This graph shows no hint of any major interaction, and we should feel quite comfortable with RM-ANOVA in this situation. What would interaction look like? If major interaction were present, the order of the colors would most likely differ a lot by patient.

**Figure 8.1 Interaction Plot**



## 8.4 TWO-FACTOR ANOVA: INTERACTION & MAIN EFFECTS

As seen in Section 8.2, RM-ANOVA is used when there are two factors in the model but only one factor is of primary interest (while the other is present due to repeated measurements on people and an expectation of variability in response across different people). Technically, the simplest RM-ANOVA is in fact a two-factor ANOVA model that incorporates an "additivity" assumption. At this point we now transition into ANOVA where two factors are of equal interest. How do things change? The following issues must be considered:

1. There will be more than one relevant F-test. Pairwise comparisons likewise become more numerous (and only some of them will be of interest).

2. **Interaction** between factors is possible and can be modeled, provided that the design has **replication**. When present, its interpretation takes precedence over that of individual factor **main effects**.

3. **Cell Sizes** become more important (differing cell sizes now create the possibility for **confounding**). For this reason, **balanced design** is ideal. A balanced design is one that has equal cell sizes – i.e. the same number of observations for each **treatment** (combination of factor levels). This chapter will focus on designs that are balanced.

The idea of interaction is of key importance. You have already seen the idea of an **interaction plot** in assessing the additivity assumption for RM-ANOVA. It is now important to consider what things look like when we do have interaction. Before considering examples, two formal definitions are needed:

> **Main Effect**: Differences in population means that are associated to a single factor.
>
> **Interaction**: Differences in population means that are associated to multiple factors in such a way that complete separation of these factors in discussion of these differences *cannot* be achieved.

When analyzing two-factor ANOVA models, it is possible observe any one of the following:

- No evidence of effects associated to either factor.
- Only main effects (the factors may be discussed separately).
- Interaction that does not allow for any separation of main effects (the factors must be discussed simultaneously).
- Interaction that still allows for some separation of main effects (general aspects of the factors may be discussed separately, but specific aspects still require simultaneous consideration). An example of this would be a scenario in which an overall "best" treatment is identified, but the amount by which it is "best" depends on one or more other factors.

These situations will be easiest to identify in the context of examples. Consider a study that has as its main goal to determine if a certain medication is effective for weight loss. The medication is to be used in conjunction with a standard exercise program. Biochemists believe that the effectiveness of the medication may be different for men and women (in other words, there may be an interaction effect between gender and the medication).

Forty participants (20 men and 20 women) are available for study, and are randomized into two groups. Group A, containing 10 men and 10 women, will receive the medication in addition to the exercise program. Group B, also containing 10 men and 10 women, will receive a placebo in addition to the exercise program. The response variable for this study will be the amount of weight lost over the course of one month on this program (note that negative values are possible here and would imply a weight gain).

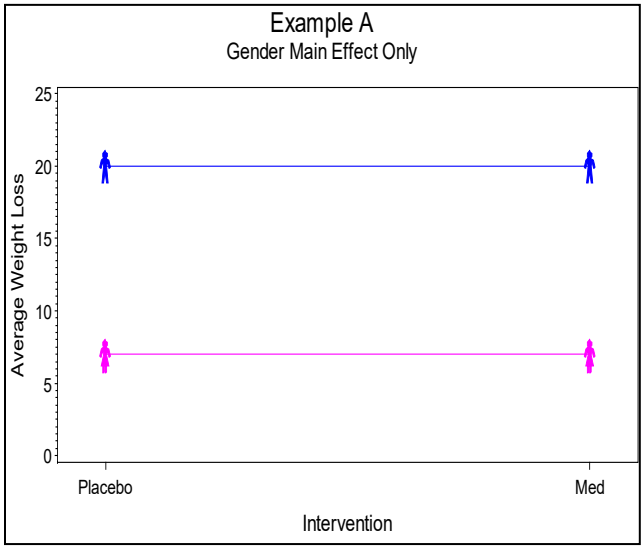
By definition, an **interaction plot** for two-factor ANOVA shows the **sample average** for the different levels of Factor A at each different level of Factor B (i.e. both factors are relevant to the plot). The factors are arbitrarily assigned here and can be interchanged as desired. What this means is:

- The vertical axis will always represent the average response (average weight lost in the case of our example).
- The horizontal axis will be labeled with the levels of one factor (for our example we will label it with the medication).
- Different symbols/colors/lines will be used to represent the levels of the other factor (we'll use blue for men and pink for women to differentiate the genders).

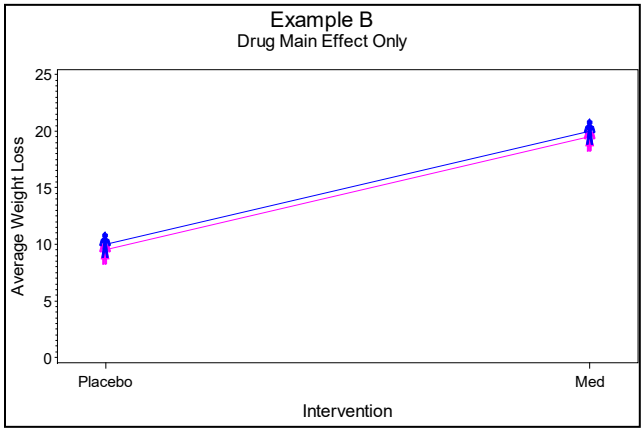
It is common practice (but not a necessity) to place the factor with the most levels on the horizontal axis so that there will be fewer colored lines. This may lead to an interaction plot that is easier to understand and interpret.

> **Important Point**: Keep in mind interaction plots show SAMPLE averages. They do not illustrate the inherent variability involved in sampling, and therefore they cannot be interpreted for population inference without further analysis using F-tests and pairwise comparisons.

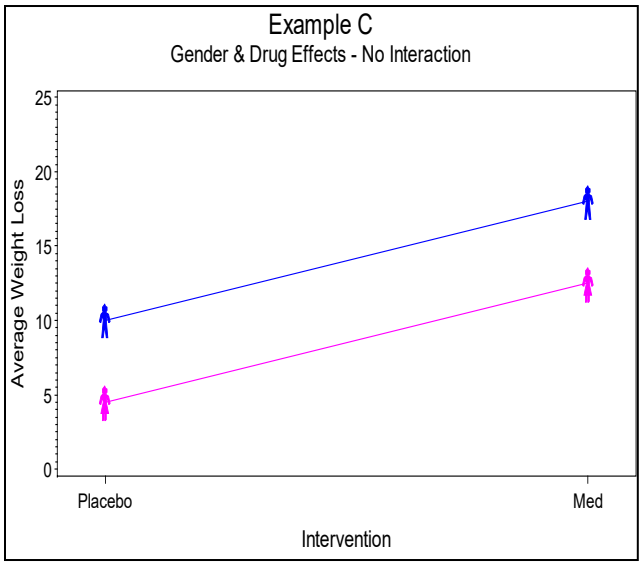
While recognizing that statistical assessment is needed for inference, perhaps the best way to understand the nuances of interaction is by examining several different interaction plots in an attempt to understand the ideas of **main effects**, **interactions**, and the scenarios described above. Let's do this with the weight-loss example; in doing so, we shall assume that appropriate statistical analyses have been completed and that all sample mean differences greater than 1 pound have been identified as statistically significant differences. The six graphs on the following pages illustrate the different types of relationships that could potentially exist between gender, medication, and average weight loss.



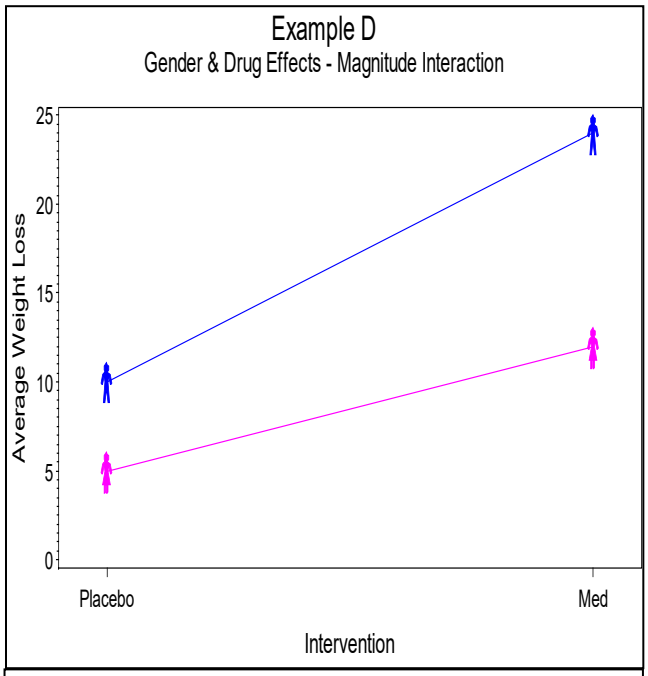

Example A: In this example, we have only one main effect. Gender differences may be described as follows: **Men lose more weight on average as compared to women when using the exercise program involved in the study.** Unfortunately, in this case the medication did not show any evidence of being useful (note that both men and women lost about the same average amount of weight whether or not they used the medication). There also is NO evidence of interaction here; gender has no bearing on a discussion of the effectiveness of medication.

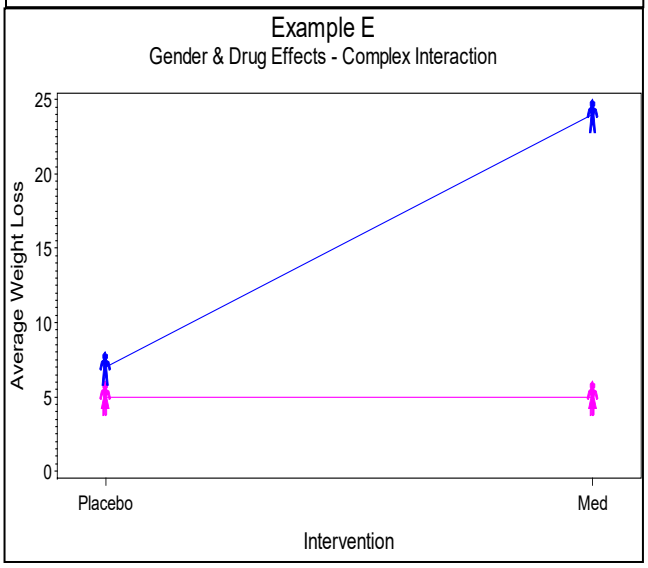



Example B: This interaction plot again displays a single main effect – this time an effect of the medication. There is no evidence of any difference in average weight loss in comparing men and women. **Irrespective of gender, there is evidence that the medication increases average weight loss**. As we have no evidence that gender has any bearing on this conclusion, there is no interaction to consider.




Example C: In this interaction plot, there is still no evidence of interaction (one should have noticed by now that when there is no interaction, the lines in the plot are **reasonably close** to *parallel*, although again the actual conclusion should always rely on the associated F-test). There are main effects here: **men lose on average more weight than women (whether the medication is used or not); and the medication seems to be effective in increasing average weight loss (irrespective of gender)**. The idea that there is no interaction here is what allows separation of these two conclusions.



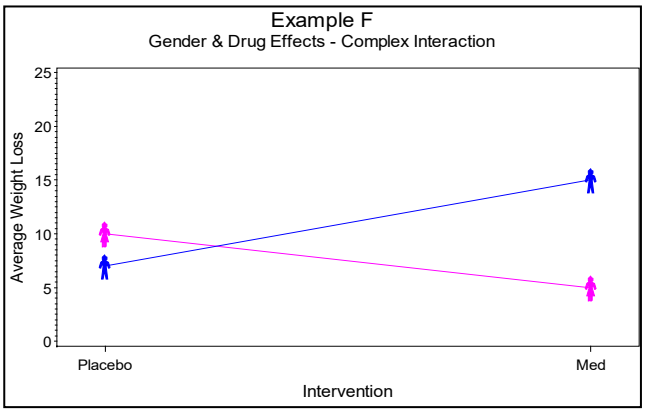




Example D: The remaining three examples involve interaction (note the non-parallel lines in the plots). From this plot, we may identify three things: (1) Men lose on averages more weight than women, regardless of whether the medication is used; (2) the medication increases the average weight loss irrespective of gender; (3) **the medication increases average weight loss by a greater amount for the men**. This last statement is the statement of interaction. Now gender and medication cannot be fully separated (while the medication is effective for both males and females, it is *more effective* for the men.) This **mild interaction** impacts only the magnitude of differences.

Example E: In this example there exists a **complex interaction** and it is no longer possible to separate main effects at all. That is, no conclusions can be made that relate only to gender; similarly no conclusions can be made relating only to the medication. What conclusions can be made here? The medication is effective in increasing average weight loss for the men, but not for the women. Notice how this conclusion carefully incorporated both gender and the intervention. Also note that if one looked only at the medication, completely ignoring gender, one would see an increase in the average. That is to say that the main effect for the intervention *will test significant in this model*. But one can see from the interaction plot that no blanket statements about the effectiveness of the medication would be correct (it is only effective for men, not women). This brings up an important point: when interaction is both statistically significant and clinically important, individual assessments of the main effects are quite often inappropriate.

Returning momentarily to Example D, in that case we were able to make a statement with regard to overall effectiveness of the medication – but only in so much as the both genders saw improvement. The magnitude of the improvement depended on gender and therefore even in that case, discussion of the size of any effect for the medication cannot occur unless gender is also involved in the discussion.

**Key Point**: When interaction is both statistically significant and clinically important, individual assessments of the main effects are quite often inappropriate.




Example F: Again there is interaction in this case and thus any discussion must relate to both gender and the intervention. What is seen here is that the medication appears to be somewhat helpful for men – improving their average weight loss. Women on the other hand should not use this medication if they wish to lose weight, as in their case it appears to inhibit weight loss (women lose more weight through exercise without the medication.

**Additional Point**: Much importance is given to the **effect size** or the magnitude of a change based on an intervention. It should be noted that this will be a necessary part of any discussion of the clinical importance of any result.

## 8.5 TWO-FACTOR ANOVA: MODEL, SIGNIFICANCE, AND COMPARISONS

As mentioned many times in the previous section, evaluation of interaction plots is a descriptive analysis; to draw inference about the population one must return to hypothesis tests and confidence intervals. For a two-factor ANOVA model with interaction, the following changes are necessary to inferential analysis:

1. The "model" portion of the ANOVA table will now have three lines – one for each factor's main effect and one for interaction. There will likewise be three F-tests given in the ANOVA table.

2. The F-tests must be considered in a specific order, ***beginning with the test for interaction***. In the event that a significant interaction is found, the tests for main effects will not be meaningful and should no longer be considered.

3. Interaction plots can be useful to aid in identification of differences, but all differences should be confirmed using Tukey adjusted pairwise comparisons (occasionally a procedure other than Tukey may be more appropriate).

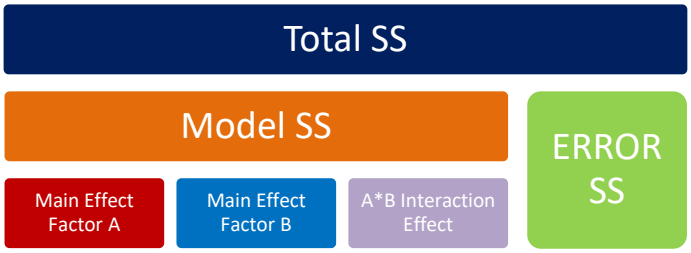
The strategy below lays out the flow of analysis in a typical two-factor ANOVA. One should generally follow the outlined steps, in order:

**Strategy for analyzing Two-Way ANOVA Design with Interaction**

1. Check assumptions for the model (normality, homogeneity, independence).
2. Check for an interaction effect. Use the F-test from the ANOVA table for the interaction line. If the F-test is significant, produce an "interaction plot" to determine whether interaction seems to be complex. If so, interpret the plot **in conjunction with Tukey intervals** at the interaction level and proceed to Step 4. If no interaction is found, continue to 3A; for mild interaction, go to 3B.
3. Consider main effects for each factor (individually).
   a. If interaction was insignificant: F-tests for main effects are valid; if significant use a pairwise comparison procedure (such as Tukey) to assess differences **individually for each factor**.
   b. If interaction was significant: F-tests for main effects are not valid. If there was an interaction, then both factors are important (even though main effects F-tests may have insignificant p-values). If there is still a "visible" main effect, then the main effect may in some sense be "much stronger" than the interaction. The interaction plot and Tukey intervals should provide some idea of what if anything one can say about main effects. Pairwise comparisons however should occur only at the interaction level unless the main effects are so much larger as compared to interactions that the interactions may be deemed clinically unimportant.
4. Summarize your conclusions in the context of the problem. Include ANOVA output, interaction plots (quite useful in helping to see likely differences), and pairwise comparisons (needed to confirm differences statistically).

## 8.5.1 A FEW DETAILS OF THE ANOVA TABLE

**Figure 8.2 Sums of Squares Breakdown**




A generic ANOVA table for a balanced two-factor ANOVA with interaction is shown in Figure 8.3. Keep in mind that one would not calculate sums of squares, mean squares, or F-statistics by hand. Total SS are still broken into "Model" and "Error" but now the Model SS may be considered as the sum of three components related to the two factors and their interaction (see also the concept map in Figure 8.2). With regard to the algebra of the ANOVA table, suppose that Factor A has $a$ levels and Factor B has $b$ levels. In this case there will be $a*b$ treatments (combinations of the levels of factors A and B). The term "balanced" implies the same number of observations ($n$) will be taken for each treatment, so that in total there will be $N = n*a*b$ observations available.

**Figure 8.3 ANOVA Table for Two-Factor ANOVA**

| Source | DF | Sums of Squares | Mean Squares | F-Statistic |
|---|---|---|---|---|
| Factor A | $a-1$ | SSA | MSA | F = MSA / MSE |
| Factor B | $b-1$ | SSB | MSB | F = MSB / MSE |
| A*B Interaction | $(a-1)*(b-1)$ | SSAB | MSAB | F = MSAB / MSE |
| Error | DFE | SSE | MSE | |
| Total | $N-1$ | SSTOT | | |

As in previous ANOVA models, mean squares are obtained by dividing the corresponding sums of squares by associated degrees of freedom. F-statistics are obtained by looking at ratios of the mean squares; these result in p-values that inform as to the significance of interaction and main effects.

Sample size yields statistical power which is necessary to find differences between treatment means when they exist. As such, perhaps the most important item in the ANOVA table is **degrees of freedom**. In particular, the **degrees of freedom for error** (labeled DFE in the table above) are typically responsible for statistical power. One might think of DFE as the **effective sample size** for an analysis. In fact, one ought to calculate DFE prior to the analysis in order to insure that the effective sample size will not be too small. That calculation is not terribly difficult – the steps are as follows:

1. Calculate the total DF by subtracting one from the overall sample size.
2. Calculate the DF needed for each factor by subtracting 1 from its number of levels.
3. Calculate the interaction DF by multiplying together the DF for each factor involved in the interaction.
4. Calculate DF for error by subtracting all DF calculated in (2) or (3) from the total.

If the error DF will be too small, the solution is to add **replication** (i.e. increase the number of observations, $n$, that are allocated to each treatment). As models grow more complex in nature, determination of appropriate sample size can become substantially more complicated, so this is typically something best accomplished with the aid of a practicing statistician.

## 8.5.2 DESIGN ASPECTS: REPLICATION AND COMPLETENESS

A **complete factorial design** is one in which every combination of factor levels is implemented. That is to say that each **cell** – each combination of factor levels – has available data. A **balanced design** is a complete design in which each factor-level combination is used the same number of times. Note that balanced does imply complete. In an ideal world, designs will be both complete and balanced (and if not, very close to it). When either is lacking, **confounding** can occur. With regard to ANOVA, confounding means that for factors A and B which are confounded, it is not possible to fully and separately assign their effects (this is different from interaction because if effects are confounded one simply cannot know which factor is responsible for an observed difference!).

As an example of confounding, suppose that it is of interest to consider the effects of gender and age on height. A sample of 30 people is collected, but no attempts at stratification are made. The resulting sample is illustrated in the **design chart** below (note that a design chart illustrates all of the possible treatments and each "x" represents a single observation for that treatment).

| | Children | Young Adults | Elderly |
|---|---|---|---|
| Male | xxxxxxxxxx | xxxx | x |
| Female | x | xxxxxxxxxx | xxxx |

This sample is problematic, as it is likely to suggest that females are taller on average as compared to males. The reality is that adults are taller than children – and in this sample we have substantially more male children than female children (and conversely more female adults than male). In fact, all but one of the children in this sample are male. The main effects of age and gender will be confounded. We cannot accurately compare male and female (because we'll wind up seeing difference due to age). Nor are we able to accurately compare, for example, heights of children and elderly – the magnitude of differences we find are going to be affected by the fact that most of the children in this sample are male while most of the elderly are female. With substantial confounding present, conclusions from an ANOVA model will not be useful.

Structuring a design in this way is to be avoided. This sample should be **stratified** in such a way that confounding will be avoided – as is the case in the design chart below. **When designs are balanced, comparisons will be fair and confounding is not possible.** Having equal numbers of children, young adults, and elderly across gender will allow comparisons of the two genders to remain fair. Likewise a comparison of children to elderly would be fair since males and females are equally represented. With balanced design, we are able to differentiate effectively between the factors and conclusions based on ANOVA will be sound.

| | Children | Young Adults | Elderly |
|---|---|---|---|
| Male | xxxxx | xxxxx | xxxxx |
| Female | xxxxx | xxxxx | xxxxx |

### 8.5.3 POST-HOC COMPARISON

Main effects comparisons, when they are valid (i.e. when interaction lacks evidence of importance), are constructed a similar manner to a one-factor ANOVA. Adjustment for multiple comparisons (see Section 7.7) applies and confidence intervals can be obtained and interpreted for each main effect separately when no interaction is present.

If interaction is present, it is necessary to obtain confidence intervals at the interaction level of the model. What does this mean? Suppose that we do have a balanced 3x2 factorial design in which the factors are age and gender. For convenience, we number the cells of the design table as seen below.

| | Children | Young Adults | Elderly |
|---|---|---|---|
| Male | (1-1) | (1-2) | (1-3) |
| Female | (2-1) | (2-2) | (2-3) |

A main effects comparison for gender would compare the **combined average** results from (1-1), (1-2), and (1-3) with the **combined average** results of (2-1), (2-2), and (2-3). Main effects comparisons for age would be very similar, but combining the other direction in this table. Interaction level comparisons, on the other hand, compare individual cells within the table without any combining of cells. Thus at the interaction level the following 15 different comparisons might be made in this example:

| | | | | |
|---|---|---|---|---|
| **(1-1) vs (1-2)** | **(1-1) vs (1-3)** | **(1-1) vs (2-1)** | (1-1) vs (2-2) | (1-1) vs (2-3) |
| | **(1-2) vs (1-3)** | (1-2) vs (2-1) | **(1-2) vs (2-2)** | (1-2) vs (2-3) |
| | | (1-3) vs (2-1) | (1-3) vs (2-2) | **(1-3) vs (2-3)** |
| | | | **(2-1) vs (2-2)** | **(2-1) vs (2-3)** |
| | | | | **(2-2) vs (2-3)** |

But these are not all important to us. For example, it makes little sense to compare heights of male children to those of female adults. There are several that we might care about, and these have been illustrated using different colors in the table above:

- **Green Comparisons: Compare the three different ages for men**
- **Blue Comparisons: Compare the three different ages for women**
- **Orange Comparisons: Compare men and women at each different age.**

Note that key to these comparisons being interesting is that they share something in common: **men**, **women**, or **age**. The remaining comparisons have nothing in common and would typically not be of interest.

> **Key Point**: Computer programs do not recognize which comparisons may or may not be important. Typically, they simply generate all possible comparisons. It is up to the researcher to determine which are useful and which are not.

## 8.5.4 PREDICTION FOR INDIVIDUALS

In the field of healthcare, individual prediction can often be important. However very few statistical software programs produce any sort of prediction by default. And such a focus is rare within medical literature as well. For **fixed effects** models (see Section 8.7 for a definition of this, and also note that repeated measures automatically involve a **random effect**), prediction intervals can be developed by applying the empirical rule (see Section 2.6). The sample mean for the group will be taken as the center of the interval, and the square root of MSE is the estimate that should be used for standard deviation. As a simple example, suppose that MSE=25 and the sample-mean height for young-adult men was 70 inches. Based on MSE, the standard deviation estimate would be 5 and therefore the empirical rule estimates that most young-adult males would have heights between 60 and 80 inches. While this technique does require some hand-calculation, it can also be quite useful when evaluating results provided as part of published research.

## 8.6 TWO-FACTOR ANOVA EXAMPLE

Consider a study published in the *Journal of Clinical Nursing* in 2013. Authors designed an experiment to examine how a nurse-led intensive educational program impacted control of hyperphosphataemia in a certain group of patients. The details of the study are available in the published results of Shi, et. al.[67], but suffice it to say that in the opinion of at least one statistician, this particular study was fairly well constructed. Here we will focus on one particular model that the authors of this study used to evaluate the effects of their intervention on serum phosphorus levels. In statistical terms, here are the details of that two-factor repeated measures ANOVA model:

- Response variable: Phosphorous level in mmol/L
- Repeated measures: Participants' phosphorous levels were measured initially to obtain a baseline and then measured again at 3 months and 6 months of involvement in the educational intervention. Three measurements were taken for each participant.
- Factor 1: **Intervention** (80 participants were randomized 40 each into either intervention group or control group)
- Factor 2: **Time** (participants were measured at baseline, or prior to intervention; they were then measured again after three months and six months of the intervention, respectively)

Before proceeding to their analysis, it is important to recognize two additional design elements here that go beyond a simple two-factor ANOVA. The first is repeated measures, and it is sufficient to understand that this is being incorporated reasonably and for the simple purpose of reducing error variability by accounting for difference among the participants. The second design element is more advanced. In addition to repeated measures, participant is a **nested factor** in this model. Nested factors are factors which the levels (in this case the participants) occur in combination with only <u>one</u> level of another factor. Here, participant is nested within intervention. Each participant is in only one specific intervention group (either treatment or control). In contrast, each participant in each intervention group is measured at each of the three different time-points. Hence the intervention and time factors are called **crossed factors** since each combination of levels does occur in the study. Why does this matter? Interaction effects can be considered for crossed factors, but they cannot be studied (and in fact would be assumed not to exist) for two factors that are nested.

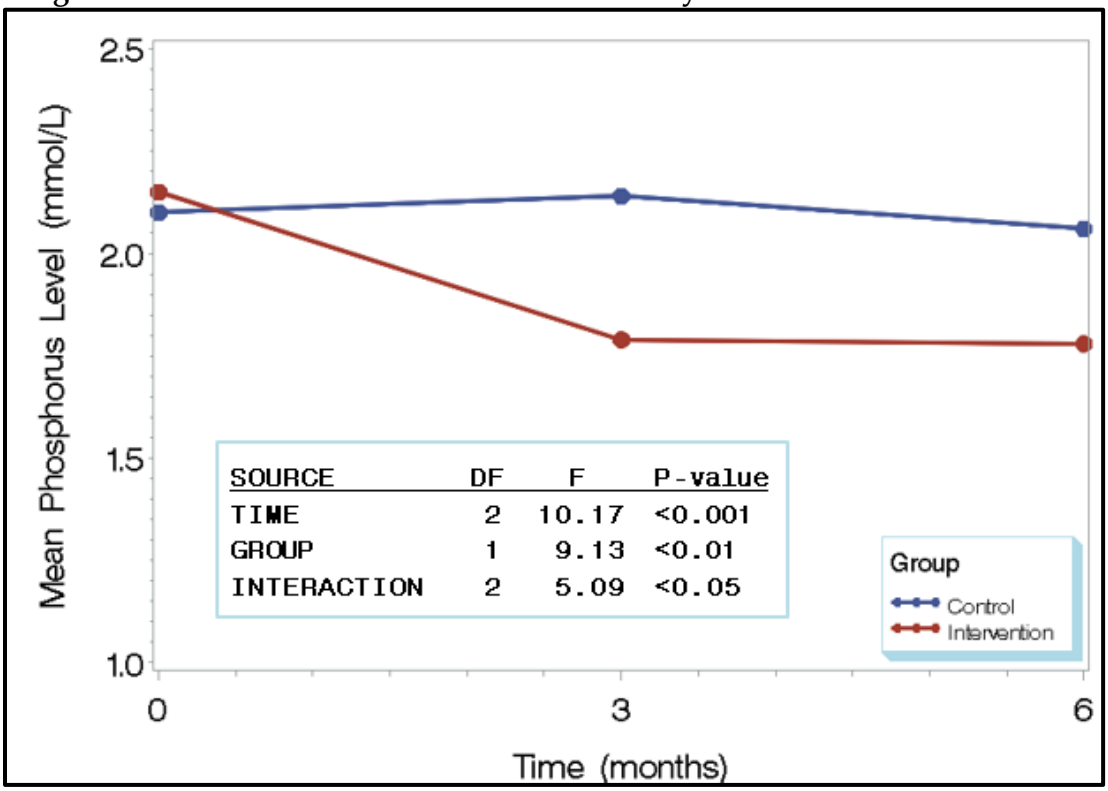
Given the complex nature of this design, the statistical analysis should almost certainly be done in consultation with a statistician. In terms of understanding results, however, the repeated measures factor (participant) doesn't play a large role; for interpretative purposes we can view this as a two-factor ANOVA (factors of interest being intervention and time). In the first two line of their second table, the authors of this study have reported group (sample) means, F-statistics, and p-values from this model. They've also reported some post-hoc t-tests (multiple comparisons) in the third line, presumably as an attempt to compare months 3 and 6 to baseline. Based on available information in their manuscript, we can recreate a portion of the Shi study's

[67] Shi, Y.X., Fan, X.Y, et. al. (2013). Effectiveness of a nurse-led intensive educational programme on chronic kidney failure patients with hyperphosphataemia: randomized controlled trial. *Journal of Clinical Nursing*, 22, 1189-1197.

ANOVA table as well as an interaction plot for this model. These are shown in Figure 8.4. Given this information, we will now follow analysis through the steps of the ANOVA procedure.

**Figure 8.4 Interaction Plot related to the Shi study[1]**




**Step 1: Check assumptions.** As we don't have the authors' actual data here, we will assume the authors collected a reasonable sample and that they checked to ensure that the three main assumptions (normality, constant variability, and independence) of ANOVA are satisfied.

**Step 2: Check for interaction effect**. The p-value for interaction in this model is significant. Since we have evidence of interaction, we would not consider main effects (except in the rare circumstance where we might deem the interaction to be statistically significant but clinically unimportant).

**Step 3: Interpret interaction effect**. Ideally we would have some confidence intervals here, but the authors of this study do not provide these. They do provide two statistically significant post-hoc t-tests, presumably showing that the 3-month and 6-month time-points differ when compared to baseline *for patients who receive the intervention.* Notice how interaction is built in here. The same statement is not true *for patients who do not receive the intervention*; for those patients the authors didn't include post-hoc t-test results, but we can see from the interaction plot that pairwise comparisons within the control group would not likely show evidence of differing means.

**Caution**: **We should <u>not</u> attempt to analyze main effects in this example.** Though they result in low p-values, the F-tests for group and time separately are not relevant here. It isn't true, for example, that the intervention group always has lower average phosphorus levels compared to control. That's likely a valid statistical conclusion at 3 months and 6 months, but not at baseline. Notice how to talk about the intervention here one must also reference time (and vice versa). This duality is the essence of an important interaction. Taking the time to understand the manner in which interaction requires looking at both factors to explain the impact of either is perhaps the most important thing one can do when trying to understand the idea of interaction.

## 8.7 IMPORTANT ADVANCED TOPICS

ANOVA designs need not stop with two factors. Models having three or more factors may be constructed and complex interactions between any and all factors may be examined. Methods from one and two-factor ANOVA do "scale up" in some of the ways one might expect:

- The "model" portions of ANOVA tables have lines for each main effect or interaction term that is to be examined.
- Pairwise comparisons may be considered for any non-interacting main effects as well as appropriate interactions. Analysis begins with the most complex interactions.

It is generally true, however, that the more factors one adds to the model, the more complex interpretations will become. As more factors are used, balance in design may become an issue (after all, it must remain feasible to collect the data). As three or more factors become involved in an interaction effect, interpretations lose substantial clarity. For these more complex designs, one is usually best to consult with a practicing statistician.

Another issue that may substantially complicate an ANOVA is the presence of a **random effect**. A random effect is a factor for which not all possible levels of the factor will be represented in the analysis (if all levels of interest are studied, we call them **fixed effects**). A common example of this occurs in repeated measures analysis – where "subject" is a random factor. In that particular case, due to the additive nature of the repeated measures model, the fact that subject is a random factor has no real impact on results. But in multi-factor ANOVA with interaction models, the presence of random factors typically will rewrite the rules for many of the F-tests – making consultation with a statistician very important in such a case.

Still another issue arises when there exist quantitative variables (in addition to your factor of interest) that may affect the response variable. If such variables are available, we can account for them using an **Analysis of Covariance** (ANCOVA) model. Provided that the quantitative **covariate(s)** are not variables of particular interest to the researcher, they are often treated in the same manner as blocks. That is to say that they are included in the model to reduce variability (usually under the assumption that these variables have no interaction with the factors of interest) and otherwise do not play a large role in conclusions of the analysis. In healthcare literature, authors often reference as part of their methods that they are "adjusting" for certain quantitative variables that they do not otherwise discuss. This usually implies they are using ANCOVA.

## 8.8 CHAPTER SUMMARY

While it may be difficult to believe – this chapter provides only a small glimpse into the potential complexities of ANOVA models. This text is not intended to cover substantially advanced design issues, nor would it be recommended that complex designs be handled without the aid of a statistician. Two advanced statistics texts that would cover such designs are suggested in the footnotes.[68,69]

Ultimately, we examined two models that behave very nicely: repeated measures ANOVA and two-factor ANOVA using balanced design (and assuming fixed effects). The reality is that in practice, analyses are seldom this simple. Two factors often are not enough to appropriately model a situation; random factors are often involved; and missing data are common – meaning that balance can be lost even in studies which are designed to attain it. The strategies when these things happen are two-fold:

1. Understand the issues involved (e.g. confounding, changes to the F-tests required for random factors, etc.)
2. Consult with a statistician to ensure that analyses are done correctly and that you understand how to properly interpret the results.

While the second of these may seem apparent, all too often results are published that make it clear that this strategy was not employed.

## 8.9 EXERCISES

1. A study involves 120 participants randomized 60 each into an intervention group and a control group. The study also involves each participant being measured twice (a pre-test and post-test scenario).
    a. Explain how this could be handled using repeated measures ANOVA.
    b. Explain how this could likewise be handled using an **independent** two-sample T-test.

    Note: Neither method is actually better than the other (they are equivalent).
2. A researcher proposes to conduct a study of three medications proposed for use in preventing seizures. He proposes to place patients on medication 1 for the first month, medication 2 for the second month, and medication 3 for the third month. Each month he will record the number of seizures for each patient. This is poor design for two reasons. Explain what this researcher did wrong and how this particular study ought to be done differently.

[68] Montgomery, D.C. (2013). *Design and Analysis of Experiments*. New York: John Wiley & Sons.
[69] Kutner, M.H., Nachtsheim, C.J., Neter, J., and Li, W. (2005). Boston: McGraw-Hill.

3. In a repeated measures ANOVA, authors of a manuscript report $R^2$ for the ANOVA of 99.8% and suggest that this is clear evidence that the treatment they are studying has been effective. Explain why they are wrong to jump to this conclusion.

4. A study is to examine blood pressure by gender and medication (there are two medications being tested and some participants will be given a placebo). Suppose there is found to be an interaction between medication and gender. Explain what this means. Could there also be an important main effect of gender or medication at the same time? Explain.

5. Consider Table 2 in the Shi manuscript.[70]

    a. Perform an analysis (similar to the example given in Section 8.6) for the response variable representing the product of calcium and phosphorus concentrations.

    b. Perform an analysis of the calcium level response variable.

    c. Perform an analysis of the albumin level response variable.

    d. Summarize all four analyses (including the one discussed in Section 8.6). What have the authors truly learned about their intervention?

    e. Discuss the authors analysis of the "Knowledge Score" variable in Table 2. Consideration of those results should be very different from the other variables in this table; explain why and tell what you believe these results mean.

[70] Shi, Y.X., Fan, X.Y, et. al. (2013). Effectiveness of a nurse-led intensive educational programme on chronic kidney failure patients with hyperphosphataemia: randomized controlled trial. *Journal of Clinical Nursing*, 22, 1189-1197.