Chapter 4 Notes

Brief Review

Before we discuss the ANOVA model, let's recall the classification-based testing we have done in the previous chapters.

In Chapter 3 we talked about categorical data analysis. We had observations classified (or categorized) by categorical variables. Combinations of categorical variables defined cells, and we had counts (given explicitly in the data or counted up from individual observations) for each cell.

In our discussions of descriptive analysis, we encountered a two-sample t-test which allowed us to test for significant differences between two population means. For the two-sample t-test, we had the following situation:

- a classification variable (defined in a **class** statement in SAS)
- the classifier had two possible values
- a population measure of interest (defined in the **var** statement)
- a sample of observed measurement values from each population
- an assumption that the underlying populations were normally distributed

The two-sample t-test for location is very useful in cases where it is appropriate (comparison of a population measure for two populations based on independent samples from the two populations and assuming the populations are reasonably close to normally distributed).

The two-sample t-test has limitations:

- we are limited to a single classification variable with two possible values
- we assume normality of the underlying measurements

We have seen that we can use robust rank-based methods in the case where the normality assumption is not reasonable.

What if we have a classification variable with more than 2 possible values? Or more than 1 classification variable? In such cases we can use an analysis of variance model.

Analysis of Variance (ANOVA) Model

In the Analysis of Variance model, we consider categorical predictor variables and a continuous response and we are interested in the impact of the categorical predictors on that continuous outcome. The idea of differences of population means will carry over from the two-sample t-tests, but now we may have more than one categorical variable and more than two possible values for each categorical variable. The idea of cells will carry over from categorical data analysis—rather than just having counts

we will now of continuous measurements on each of the observations, and we will be modeling those as a function of the categorical variables.

In the analysis of variance model, we have the following situation:

- a continuous response assumed to have iid normally distributed errors
- one or more nominal categorical predictor variables (basically categories for defining the subpopulations)
- no continuous predictor variables

Note that we will be assuming normality, but there are nonparametric techniques based on ranks that can be used if we choose not to assume normality.

Also note that in Chapter 4 we will be working with balanced data (equal number of observations in each cell). In Chapter 5 we will look at the more general case of unbalanced data.

Some terms

Before jumping into the model, we should define a few terms that we will want to use

One-way analysis of variance: an analysis of variance based on a single categorical predictor variable

Two-way analysis of variance: an analysis based on 2 independent categorical predictor variables

N-way analysis of variance: an analysis based on *n* independent categorical variables

Main effect: the effect of a single categorical predictor

Interaction: the combined effect of a combination of categorical predictors

First-order interaction: an interaction between two categorical predictors

N-th order interaction: an interaction of a single categorical predictor with *n* other categorical predictors

Cell: a particular combination of categories

Balanced data (or **balanced designs**): data with an equal number of observations in each cell (or experimental designs set up to have such data)

Let's take a look at what the mathematical formulation for this model would look like (see formula 4.1 in the text for a general three-way ANOVA with all interaction terms).

In the case of the t-tests for geographic differences in the UK water data, we could have specified the problem as the simplest type of ANOVA model: a one-way ANOVA where the single predictor variable could take on one of two values.

Much like in the t-test case, we want to investigate if there are significant differences between groups, but we may now have multiple groupings, and we may also need to incorporate interactions between those groupings.

An Example: Hypertension Treatment

The first example in chapter 4 is about the impact of medication, biofeedback and diet on blood pressure. So we have 3 possible categorical predictors (3 possible **drug**s, absence or present of **biofeed**back, and absence or presence of a special **diet**) for a continuous response **bp** (blood pressure).

Before we do any type of modeling of blood pressure, we should have a look at the data itself and then cross-tabulations for the response to give us some basic descriptive information about the blood pressure values under particular conditions (e.g. cells).

After that, let's take a look at a one-way ANOVA based solely on the **drug** taken.

We should note that this data set is **balanced**. With balanced data, we can use **proc anova**, which is intended for use on balanced designs and some other special types of designs described in the **Overview: ANOVA Procedure** section of the SAS documentation. With unbalanced data there are additional issues and we will want to use **proc glm** (we will see that in Chapter 5…).

With balanced data (or balanced designs), we have some nice properties including:

- the effects (main effects or interactions) can all be assumed to be orthogonal
- the cell means will be the same as the predicted means for the model
- if we decompose the model sum of squares (the variation in the data explained by the model) into the contributions from the various effects (main effect or interaction), the contributions don't change if we reorder terms with the same number of interaction components

For a one-way analysis of variance, these properties still hold and we are still OK using **proc anova**. The contributions to the model sum of squares just comes from one main effect, so the model sum of squares IS the sum of squares for main effect.

Example

Let's have a look at the one-way analysis of variance with **drug** as the only predictor variable.

- What does the overall ANOVA table tell us about the significance of the model?
- What does the R-Square value in the fit statistics tell us about the predictive power of the model?
- What does the ANOVA model ANOVA table tell us about the significance of **drug** as a predictive variable?
- How is this related to what we found from the overall table?
- What do our results tell us about the impact of **drug** on blood pressure?

Exercises

Two-way ANOVA

Let's add the **diet** variable as a main effect (so we'll have the model with both **drug** and **diet** as predictors, but no interaction between **drug** and **diet**). You may wish to look at the documentation for the **MODEL** statement to **proc anova**.

- Obtain the two-way ANOVA results.
- Answer the same questions for these tables as we answered for the one-way model, this time answering with respect to both **drug** and **diet**.

One-way Using cell Variable

- Perform a one-way analysis using the **cell** variable as the only predictor.
- What does this tell us about the individual impacts of **drug**, **diet**, and **biofeed**back on blood pressure?
- What conclusions can we make based on this analysis?

Three-way Main Effects ANOVA

- Perform a three-way analysis of variance using **drug**, **diet**, and **biofeed**back as predictors.
- What can we conclude about the impact of drug, diet, and biofeedback on blood pressure?

Three-way ANOVA with Interactions

- Perform a three-way analysis of variance with all possible interactions (the shorthand | notation will make it easier to specify the model than writing out all terms explicitly).
- Include an **ods output** statement to save the means of the three-way interaction term to a variable **outmeans**.
- How does this model compare to the one-way analysis of variance based on **cell**?
- How does it compare to the three-way analysis of variance with no interactions?
- At a .05 level, are there terms which would not be considered statistically significant on their own? What does this tell us about significant interactions?

Multiple Comparisons and Means Analysis

The analysis of variance tables and associated F statistics will allow us to determine if there is a statistically significant difference in the response (in the example we're looking at, the response is blood pressure) across groups. It's good to determine that a significant difference exists, but it would be better to determine which groups differ significantly. To answer the question of "which populations are different," we will want to compare the means of the populations and take into account the variation within populations (as estimated from our samples).

For one-way analyses of variance, we should use a **hovtest** option to check that the variance across population is roughly the same. (If we have a higher order model, this will not be possible.) If the variances are deemed unequal, we could use a Welch weighted analysis of variance to account for the non-constant variance.

When we want to check for significant differences between population means, we need to take into account that if we make multiple comparisons at a given alpha value, the overall probability of being wrong about some decision goes up.

Bonferroni's method accounts for this by dividing the alpha value for each individual comparison by the number of comparisons that will be made. This is the most conservative comparison and results in an overall error rate less than or equal alpha.

Tukey's method is intended for making all pairwise comparisons.

Scheffe's method is more complicated and factors in all possible contrasts. For our analysis, we are only looking at pairwise comparisons, so Tukey would be more appropriate than Scheffe.

There are a number of other somewhat common tests as well, but the ones mentioned above will be the one's we'll concern ourselves with for now. Note that in the SAS results for multiple comparisons, two main effect levels are considered significantly different if they share no grouping letters. So for instance, a level with grouping letters **A** and **B** is not significantly different from a level with letters **B** and **C** because they share **B**, but the level with grouping letters **A** and **B** would be significantly different from a level with only the grouping letter **C**.

Examples and Exercises

One-way example

- Test for equal variance in **drug** effect model
- Interpret significance of differences of mean for the **drug** groups

Three-way Model with Interactions

- Based on this model what terms should we keep?
- Perform pairwise means analysis on main effects
- Which test should we use?
- What do we conclude about significantly different groups?

Log Example in Text

The motivation of this example is to lessen the impact of the three-way interaction. Log-transformation of the response should lessen the significance of the higher order interactions. The authors want to make it insignificant enough in the model to throw it out.

One downside to this approach is that now we are working with the logs of blood pressures, so now all of our assumption should be reasonable for the log-transformed responses and interpretation directly in terms of blood pressure will require transformations back to the blood pressure space. The extra transforming is not such a big deal, but we might want to consider how appropriate the transformation is conceptually. If it is more conceptually meaningful to model blood pressure on a logarithmic scale, then we would definitely want to do that. If not, we may want to consider how much we're really gaining by way of the transformation.

Another possibility is to consider whether it is reasonable to include a higher order interaction without lower order interactions. If we have reason to believe that the combination of diet, drug and biofeedback should have an impact on blood pressure and that combinations of pairs of those categorical factors should not, it would be OK to leave out insignificant two-way terms and keep the significant three-way term. Generally, though, we would want to keep all lower order terms because it makes sense that if there is a higher order interaction, all the lower order interactions should be present in reality.

One more alternative is to consider how married we are to the .05 level. If a term is barely significant at that level, it might be justified to remove the term if the resulting model is more meaningful.