ANOVA and experimental design

Patrick Breheny

March 1

ANOVA

- The *analysis of variance*, or *ANOVA*, model is essentially a special case of the linear regression model in which all of the explanatory variables are categorical
- Thus, we have seen, fit, and conducted inference for ANOVA models already in this course, without calling them "ANOVA models"
- A note on terminology: the ANOVA is so named because it was the first sort of linear model for which the decomposition of variance was derived and used to carry out an F test (in modern statistics, the name is perhaps misleading, as it doesn't analyze variance any more than any other linear model does)

ANOVA terminology

- Nevertheless, certain concepts are particularly important to ANOVA models, and these models tend to have a certain terminology and parameterization
- Models with one categorical factor are called *one-way analysis of variance models*, models with two factors called *two-way analysis of variance models*, and so on
- So, for example, our model for the alcohol data set in which both sex and alcoholism were explanatory variances was a two-way analysis of variance model

ANOVA parameterization

- The parameterization used in linear models (in which one category is chosen as a reference and then indicator variables introduced as needed) is somewhat unnatural
- An alternative parameterization in which the interpretation of the parameters is simpler – but the inferential derivations much more complicated – is as follows:

$$\mathrm{E}(y_{ij}) = \mu + \alpha_j$$

In words, the expected value of the ith observation in group j is equal to the overall mean plus the effect of group j (where effect is taken to mean the departure of the group mean from the overall mean)

ANOVA parameterization (cont'd)

- Note that this model is deliberately overparameterized: it has four parameters to represent three groups, and $\mathbf{X}$ will not be full rank
- It turns out, however, that inference can still be made about certain linear combinations – for example, the mean in group j, or the difference between groups j and k – and that these results are equivalent to what you obtain if reparameterize the model in a full-rank fashion
- Other linear combinations, however, like μ itself, cannot be estimated

Constraints

- An alternative way of thinking about ANOVA models is with *constraints*
- For example, the one-way ANOVA model is often constrained so that $\sum_j \alpha_j = 0$
- With this constraint, everything is once again full-rank and can be estimated
- For example, letting $\boldsymbol{\beta}$ denote the full-rank parameterization $(\mu_1, \mu_2, \mu_3)^T$,

$$\mu = \frac{\mu_1 + \mu_2 + \mu_3}{3}$$

$$\alpha_1 = (\frac{2}{3}, -\frac{1}{3}, -\frac{1}{3})\boldsymbol{\beta}$$

Experimental design

- Another important topic that tends to be tied to ANOVA models is the issue of experimental design
- In controlled experiments, the most important statistical consideration is often the design and efficiency of the experiment
- For example, the $\sum_j \alpha_j = 0$ constraint is most sensible if the number of observations in each group are the same (*i.e.*, the design is *balanced*)
- If the design is not balanced, we have the question of whether it still makes sense for each group to receive the same weight in the constraint that they sum to zero

Experimental design and balance

- This in turn leads to confusion regarding exactly how to parameterize the model and carry out certain tests
- Although this is not crucially important in the one-way ANOVA model, unbalanced designs can lead to large problems in more complicated settings
- But what needs to be balanced, and what doesn't?

Our hypothetical experiment

- Suppose several ways of teaching an introductory statistics class had been proposed, and we were interested in conducting an experiment to determine which was the best
- Specifically, we plan on giving one of several lesson plans to an instructor and having them teach a course out of that lesson plan; the outcome will be how well their students score on a common test given to all the students

Which sample size?

- We know that in order to have higher power and lower standard error, we should increase the sample size
- But which sample size? The number of students per teacher, or the number of teachers?

Which sample size? (cont'd)

- After a bit of careful thinking, we can see that no matter how many students are assigned to each teacher, we will never be able to conclude that there is a difference between the teaching methods unless we have a sufficient number of teachers
- The sample size, in terms of the number of independent observations of each teaching method, is the number of teachers, not the number of students
- In statistical parlance, students and teachers are sometimes said to be *nested* (*e.g.*, "students are nested inside teachers")

Multiple sources of variability

- The random error ϵ_i in this case is composed of two sources, the teacher-to-teacher variability and the student-to-student variability
- Although larger numbers of students will cause the student-to-student variability for each teacher to drop to zero, it will have no effect on the teacher-to-teacher variability
- This idea comes up often in biological studies, where assays can be noisy and are often repeated more than once on the same patient, but it is important to distinguish between *biological replication* (subjects) and *technical replication* (measurements per subject)

A paired design

- Multiple sources of variation may add complication, but can also be exploited to yield more efficient designs
- For example, we know that the varying levels of skill from teacher to teacher is going to be large source of variability in our study
- So consider a design in which each teacher teaches from lesson plan A in the fall semester and then lesson plan B in the spring semester
- In this design, each teacher contributes information about the difference between the lesson plans, and teacher-to-teacher variability is completely eliminated

Blocking

- This is an illustration of a *paired* or *matched* design, which is a special case of the very important concept of *blocking* (the special case of two groups)
- In our example, each teacher would represent a *block*; by making comparisons within each block, we increase our signal-to-noise ratio and lower our standard error by eliminating variability between blocks
- Blocking is often crucially important in well-designed experiments, although it can lead to complicated designs that require careful thought

Randomized block designs

- For example, our earlier experiment was potentially flawed – if approach A in always taught in the fall semester and approach B in the spring semester, then perceived differences between A and B could really be caused by differences in students who take the course in fall and spring semesters
- One solution is to randomly decide, for each teacher, which approach will be taught first, thereby eliminating the confounding
- This is called a *randomized block design*

Manually balancing blocks

- Rather than relying on randomization, we could also manually maintain a balance between courses and semesters by, for example, assigning every other instructor to teach with method A first
- Maintaining this balance is fairly trivial with two lesson plans, but suppose we had four (A, B, C, and D)
- Now there are $4! = 24$ possible orders; requiring each of these orders to appear an equal number of times means that our sample size must be divisible by 24 – an unattractive limitation
- This is an unattractive limitation, and becomes very unattractive as we start to consider larger and larger numbers of treatments (lesson plans)

Latin squares

- However, this is not the smallest number of orders needed to preserve balance
- Consider instead the design

	Fall	Spring	Fall	Spring
Teacher	2011	2012	2012	2013
1	A	B	C	D
2	B	C	D	A
3	C	D	A	B
4	D	A	B	C

- This is an example of a *Latin square*, a technique for maintaining balance across more than one blocking component (here teacher and semester)

Incomplete blocks

- As a final example, suppose that we still have four methods we wish to compare, that we only have one semester in which to gather data, but that each instructor can teach two sections of the course
- Thus, we don't need to worry about confounding due to semesters anymore, but now we have the problem that we can't balance across each block (four methods, but each teacher can only teach two classes)
- In other words, *complete blocks* are infeasible, and we must consider an *incomplete block design*

Unbalanced incomplete block design

However, not all incomplete block designs are good ones; consider the following:

Teacher		
1	A	B
2	C	D
3	A	B
4	C	D
	...	

Balanced incomplete block designs

- This design suffers from the fatal flaw that method C is always compared against method D; this makes it impossible to compare methods A and B
- Now, we must ensure that each treatment appears in a block with each other treatment the same number of times
- Such designs are called *balanced incomplete block designs*

Balanced incomplete block designs (cont'd)

A balanced incomplete block design:

Teacher		
1	A	B
2	A	C
3	A	D
4	B	C
5	B	D
6	C	D
	...	

More than one factor

- These concepts extend into multi-way studies as well
- Suppose, for example, that we were interested in varying the textbook (among three choices), the computer software used (R, SAS, or none), and the order of topics (2 choices)
- Considering all combinations (the so-called *factorial* design), we have $3 \times 3 \times 2 = 18$ treatments to investigate now
- The details of the design get more complicated (incomplete blocks are a much bigger issue now, for example), but the ideas are the same

Final remarks

- As a final remark, it is worth noting that blocking can result in significant gains in efficiency, but that it is not always possible to block
- Furthermore, one can only block on known factors, and there are limits to how many factors one can block on
- Thus, the general maxim of experimental design is: "Block what you can and randomize what you cannot"