Contents

5

Analysis of variance

This chapter goes deeper into the Analysis of Variance (ANOVA). We consider multiple factors and we also introduce the notions of fixed and random effects. Most of these notes assume familiarity with statistics and data analysis in order to study how to make data more than how to analyze it. This chapter is a bit of an exception. We will also extend the theory to cover what might be gaps in the usual way regression courses are taught. ANOVA is a bit more complicated than just running regressions with the categories coded up as indicator variables, and so we will need to add some extra theory. Where possible, the additional theory will be anchored in things that one would remember from an earlier regression course.

Suppose that we have two categorical variables, say A and B. Each has two levels. That makes four treatment combinations. We could just study it as one categorical variable with four levels. However we benefit from working with the underlying 2×2 structure. We have two choices to make (which A) and (which B) and a third thing about interaction, which we will see includes considering whether the best A depends on B and vice versa. Note that here A and B are two different treatment options. In A/B testing A and B are two different versions of one treatment option. Experimental design is complicated enough that no single notational convention can carry us through. We have to use local notation or else we would not be able to read the literature.

5.1 Potatoes and sugar

The oldest ANOVA reference I know of is Fisher and Mackenzie (1923). They were using it to study the effects of different fertilizer on potato yields. It is (historically) interesting that even in this first paper they are thinking of non-

additivity and even consider something that looks like one term of a singular value decomposition.

To illustrate how a two factor experiment works, consider the following hypothetical yields for 3 fertilizers and 4 varieties of potatoes:

$$\begin{array}{cc} \text{Yield (kg)} & \begin{array}{cccc} V_1 & V_2 & V_3 & V_4 \end{array} \\ \begin{array}{c} F_1 \\ F_2 \\ F_3 \end{array} & \left[\begin{array}{cccc} 109.0 & 110.9 & 94.2 & 125.9 \\ 104.9 & 113.4 & 110.1 & 138.0 \\ 151.8 & 160.9 & 111.9 & 145.0 \end{array}\right] \end{array}$$

Based on these values, we can wonder which fertilizer is best, which variety is best, and the extent to which one decision depends on the other.

Taking the yield data to be a 3×4 matrix of Y_{ij} values, the overall average yield is $\bar{Y}_{\bullet\bullet} = 123$. If we think fertilizer i raised or lowered the yield it must be about yields higher or lower than 123. So we can subtract 123 from all the Y_{ij} and then take $(1/J)\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{\bullet\bullet}) = \bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}$ as the incremental effect of fertilizer i. We get:

F_1	F_2	F_3
-13.0	-6.4	19.4

By this measure, fertilizer 1 lowers the yield by 13 while fertilizer 3 raises it by 19.4. The same idea applied to varieties yields $\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet}$:

V_1	V_2	V_3	V_4
-1.1	5.4	-17.6	13.3

Variety 3 underperforms quite a bit while variety 4 comes out best.

We have just computed the ***grand mean*** $\bar{Y}_{\bullet\bullet} = 123$ and the ***main effects*** for fertilizer and variety. Note that each of the main effects average to zero, because of the way that they are constructed. We can decompose the table into grand mean, main effects and a residual term, as follows:

$$\begin{aligned} &\begin{bmatrix} 109.0 & 110.9 & 94.2 & 125.9 \\ 104.9 & 113.4 & 110.1 & 138.0 \\ 151.8 & 160.9 & 111.9 & 145.0 \end{bmatrix} = \begin{bmatrix} 123 & 123 & 123 & 123 \\ 123 & 123 & 123 & 123 \\ 123 & 123 & 123 & 123 \end{bmatrix} \\ &+ \begin{bmatrix} -13.0 & -13.0 & -13.0 & -13.0 \\ -6.4 & -6.4 & -6.4 & -6.4 \\ 19.4 & 19.4 & 19.4 & 19.4 \end{bmatrix} + \begin{bmatrix} -1.1 & 5.4 & -17.6 & 13.3 \\ -1.1 & 5.4 & -17.6 & 13.3 \\ -1.1 & 5.4 & -17.6 & 13.3 \end{bmatrix} \\ &+ \begin{bmatrix} 0.1 & -4.5 & 1.8 & 2.6 \\ -10.6 & -8.6 & 11.1 & 8.1 \\ 10.5 & 13.1 & -12.9 & -10.7 \end{bmatrix}. \end{aligned}$$

The last term captures the extent to which the yields are not additive. It is called the ***interaction***. The grand mean, main effects and interactions we

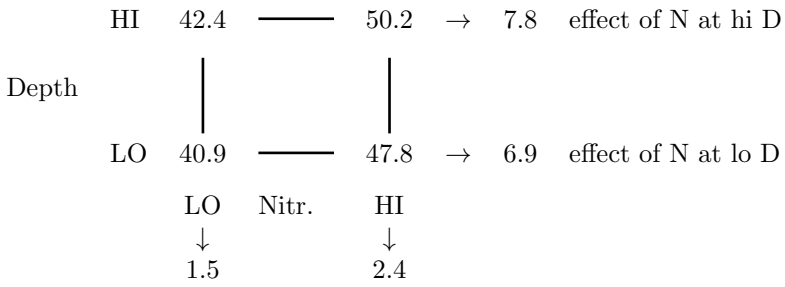

Figure 5.1: Sugar yields in a 2×2 experiment where N denotes use of nitrogen and D denotes increased depth of ploughing.

want are defined in terms of $\mathbb{E}(Y_{ij})$. The ones we get are noisy versions of those. Much of ANOVA is about coping with noise. Perhaps more is about coping with interactions.

Here is another example motivated by agriculture. My notes say that I got it from Cochran and Cox (probably a very old book of theirs) but I cannot now find it in that book. It is about the yield of sugar in 100s of pounds per acre. There's an old unit called the 'hundredweight' which is about 45.4 kilograms. They considered two treatments. Treatment N involved either no nitrogen, or application of 300 lbs of nitrogen per acre. Treatment D involved either ploughing to the usual depth of 7" or going further to 11". The results, which might or might not be hypothetical are depicted in Figure 5.1.

When both variables are at their 'low' level the yield is 40.9. We see that going to the high level of N raises the yield by either 7.8 if D is at the high level (meaning greater depth) or 6.9 if D is at the low level. The overall estimate of the treatment effect is then their average, roughly 7.4. Similarly, we see two different effects for D, one at each of the high and low levels of N, that average to about 2.

There seems to be a positive interaction of about 0.9 meaning that applying both treatments gives more than we would expect from them individually. This is a kind of synnergy.

5.2 One at a time experiments

For the sugar problem, we tried all four combinations varying both factors. We could instead have done two separate experiments, one for each factor. That is called a one at a time ***OAAT*** experiment. Sometimes people even advocate for changing just one thing at a time to learn its effects. Here we show that if you have two things to investigate it is better to investigate them in a combined experiment.

Experiment A Take n observations at $(N, D) = (0, 0)$ and n more at $(N, D) = (0, 1)$ to test test D. Then take another n at $(N, D) = (0, 0)$ and n more at

$(N, D) = (1, 0)$ to test n. Experiment A costs $4n$ and delivers $\hat{N} = \bar{Y}_{10} - \bar{Y}_{00}$ with

$$\mathrm{var}(\hat{N}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{n} = \frac{2\sigma^2}{n}.$$

Similarly $\mathrm{var}(\hat{D}) = 2\sigma^2/n$.

Experiment B Take $n/2$ observations at each of $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$. Experiment B costs only $2n$. It has $\hat{N} = [(\bar{Y}_{10} - \bar{Y}_{00}) + (\bar{Y}_{11} - \bar{Y}_{01})]/2$ with

$$\mathrm{var}(\hat{N}) = \frac{1}{4}\Big(\frac{\sigma^2}{n/2} + \frac{\sigma^2}{n/2} + \frac{\sigma^2}{n/2} + \frac{\sigma^2}{n/2}\Big) = \frac{2\sigma^2}{n} = \mathrm{var}(\hat{D}).$$

The factorial experiment B delivers the same accuracy as the OAAT one at have the cost. By that measure, it is twice as good. It is actually better than twice as good. The factorial experiment can be used to investigate whether the factors ineract.

Experiment A is not the best OAAT that we could do. It misses an opportunity to reuse the data at $(0,0)$. We could also consider experiment C below.

Experiment C Take n observations at $(0,0)$, and n more at $(0,1)$ and n more at $(1,0)$.

Experiment C costs 1.5 times as much as experiment B and has the same variance. It also makes $\mathrm{corr}(\hat{N}, \hat{D}) \neq 0$. So it's not twice as bad, only 2/3 as good. Since the observation at $(0,0)$ is used twice, we might want to give it extra samples. Of course a better idea is not to do OAAT.

OAAT could leave us with the following information

$$\begin{array}{ccc} \text{GOOD} & & \textcircled{?} \\ \uparrow & & \\ \text{FAIR} & \longrightarrow & \text{GOOD} \end{array}$$

where two changes from our starting point are both beneficial but we don't know what happens if we make both changes. We would probably try that. We might get a result like

$$\begin{array}{ccc} \text{GOOD} & \longrightarrow & \text{BAD} \\ \uparrow & & \uparrow \\ \text{FAIR} & \longrightarrow & \text{GOOD} \end{array}$$

hopefully by trying it out first before commiting to it. In a case like this we learn that making both changes is a bad idea, but we would have learned sooner with a factorial experiment.

Another possibility is that an OAAT experiment has this result

$$\begin{array}{ccc} \text{BAD} & & (?) \\ \uparrow & & \\ \text{FAIR} & \longrightarrow & \text{BAD} \end{array}$$

where both changes are adverse. We don't know what would happen if both changes were made. Based on the sketch above, many people would not even try making both changes. It is possible that the underlying truth is like this

$$\begin{array}{ccc} \text{BAD} & \longrightarrow & \text{BEST EVER} \\ \uparrow & & \uparrow \\ \text{FAIR} & \longrightarrow & \text{BAD} \end{array}$$

where making both changes would be extremely valuable. In a factorial experiment we would learn this, while n OAAT it could very well go undiscovered.

5.3 Interactions

One severe problem with OAAT is that if there are important interactions, then we don't learn about them and might be forever stuck in a suboptimal setting.

Interactions cause severe difficulties. We can think of a failure of external validity as being an interaction between treatment choices (e.g., aspirin vs tylenol) and another variable describing the past versus the future. Or that second variable could be data in our study versus data we want to generalize to. It is bad enough that 'your mileage may vary' and worse still that mileage differences may vary. That could mean that the optimal choice changes between the data we learned from and the setting where we will make future decisions.

Interactions underly lots of accidents and disasters. An accident might only have happened because of a wet road, inattentive driver, bad street lights and poor brakes. If all of those things were needed to create the accident then it is a sort of interaction. It's good that the accident is not in the grand mean or main effects, because then we could get more of them, but having it be an interaction makes it harder to prevent.

Andrew Gelman has written that interactions take 16 times as much data to estimate as main effects do: `https://statmodeling.stat.columbia.edu/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/`

It is informative to see how that 16 arises. In a 2×2 experiment the main effect estimate would have

$$\mathrm{var}(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}) = \frac{\sigma^2}{n/2} + \frac{\sigma^2}{n/2} = \frac{4\sigma^2}{n}$$

while the interaction would have variance

$$\mathrm{var}(\bar{Y}_{11} - \bar{Y}_{01} - \bar{Y}_{10} + \bar{Y}_{11})\frac{\sigma^2}{n/4} \times 4 = \frac{16\sigma^2}{n}.$$

If an interaction would have the same size as a main effect, then it would be 4 times as hard to estimate, meaning that we would need to raise n by a factor of 4 to do as well.

Now suppose that the main effect is θ_{main} and the interaction is $\theta_{\text{inter}} = \lambda\theta_{\text{main}}$. We use sample size n for the main effect and get a relative error of

$$\frac{|\theta_{\text{main}}|}{\sqrt{4\sigma^2/n}} = \frac{\sqrt{n}|\theta_{\text{main}}|}{2\sigma}.$$

If we use n' observations for the interaction our relative error would be

$$\frac{|\theta_{\text{inter}}|}{\sqrt{16\sigma^2/n'}} = \frac{\sqrt{n'}|\lambda\theta_{\text{main}}|}{4\sigma}.$$

We make these equal by solving $|\lambda|\sqrt{n'} = 2\sqrt{n}$ giving $n' = 4n/\lambda^2$. If we take $\lambda = \pm 1/2$, then we find that interactions are 16 times as hard to estimate as main effects. Relative error is important because statistical significance and power depend on the ratio of the (absolute) effect size to the standard deviation of the effect estimate. The factor of $1/2$ is a ballpark estimate. The actual size ratio between a typical interaction and typical main effect will vary with the setting.

Interactions can raise severe problems of multiple comparisons. When there are d experimental factors then there are $\binom{d}{2}$ different pairwise comparisons. If we test everything at level α then we espect up to $d\alpha$ false discoveries for main effects and up to $d(d-1)\alpha/2$ among interactions. If we think that interactions are more likely to be null (or close enough to it to be practically null) then the problem is even more severe for interactions than main effects. Or, if we use methods to control false discovery rates, then we have to be more conservative with interactions.

The problem gets worse for higher order interactions. For instance if an $A \times B$ interaction depends on the level of factor C, then we have a three factor interaction.

If in baseball a pitcher has a very good record against tall left handed batters in evening games with a light wind, then that finding is a kind of high order interaction that was probably based on slim evidence. Patterns that appear as high order interactions can sound very subtle and important but they could also be noise.

Interactions are not all bad. It is good that not everybody likes the exact same restaurants. That is an interaction between people and restaurants. More generally, personalization of services involves measuring interactions.

I believe that much of the work by George Box and his co-authors on experimental design was a response to the challenge of learning scientific facts despite the presence of many interactions. Ordinary statistical modeling is well able to handle noise. Randomization of treatments is a good way to counter missing variables and get causal estimates. Factorial designs and, later on fractional factorials, help us cope with the complexity that comes from interactions.

5.4 Multiway ANOVA

Now suppose we have 4 factors, A, B, C and D. We can write the regression model as

$$\begin{aligned}Y_{ijk\ell} &= \mu + \alpha_i + \beta_j + \gamma_k + \delta_\ell \\ &\quad + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\alpha\delta)_{i\ell} + (\beta\gamma)_{jk} + (\beta\delta)_{j\ell} + (\gamma\delta)_{k\ell} \\ &\quad + (\alpha\beta\gamma)_{ijk} + (\alpha\beta\delta)_{ik\ell} + (\alpha\gamma\delta)_{ik\ell} + (\beta\gamma\delta)_{jk\ell} \\ &\quad + \varepsilon_{ijk\ell}.\end{aligned}$$

Factor A has levels $1 \leqslant i \leqslant I$, factor B has levels $1 \leqslant j \leqslant J$, factor C has levels $1 \leqslant k \leqslant K$ and factor D has levels $1 \leqslant \ell \leqslant L$. The model above has numerous interactions. We could run out of Greek letters and so we use notation like $(\alpha\beta)$ as its own symbol to describe the parameters in an $A \times B$ interaction. Here we have an error/noise term $\varepsilon_{ijk\ell}$.

To make the model identifiable we make each main effect sum to zero and each interaction sum to zero over all values of any index in it. For instance $\sum_{j=1}^{J}(\alpha\beta\gamma)_{ijk} = 0$ for all i and k.

It is easy to work out what the estimates are when $\varepsilon_{ijk\ell} \sim \mathcal{N}(0, \sigma^2)$. We get

$$\begin{aligned}\hat\mu &= \bar Y_{\bullet\bullet\bullet\bullet} \\ \hat\alpha_i &= \frac{1}{JKL}\sum_j\sum_k\sum_\ell (Y_{ijk\ell} - \hat\mu) = \bar Y_{i\bullet\bullet\bullet} - \bar Y_{\bullet\bullet\bullet\bullet} \\ \widehat{(\alpha\beta)}_{ij} &= \frac{1}{KL}\sum_k\sum_\ell (Y_{ijk\ell} - \hat\mu - \hat\alpha_i - \hat\beta_j) \\ &= \bar Y_{ij\bullet\bullet} - \hat\mu - \hat\alpha_i - \hat\beta_j \\ &= \bar Y_{ij\bullet\bullet} - \bar Y_{\bullet\bullet\bullet\bullet} - (\bar Y_{i\bullet\bullet\bullet} - \bar Y_{\bullet\bullet\bullet\bullet}) - (\bar Y_{\bullet j\bullet\bullet} - \bar Y_{\bullet\bullet\bullet\bullet}) \\ &= \bar Y_{ij\bullet\bullet} - \bar Y_{i\bullet\bullet\bullet} - \bar Y_{\bullet j\bullet\bullet} + \bar Y_{\bullet\bullet\bullet\bullet}, \quad \text{and} \\ \widehat{(\alpha\beta\gamma)}_{ijk} &= \frac{1}{L}\sum_\ell \big(Y_{ijk\ell} - \hat\mu - \hat\alpha_i - \hat\beta_j - \hat\gamma_k - \widehat{(\alpha\beta)}_{ij} - \widehat{(\alpha\gamma)}_{ik} - \widehat{(\beta\gamma)}_{jk}\big) \\ &= \bar Y_{ijk\bullet} - \bar Y_{ij\bullet\bullet} - \bar Y_{i\bullet k\bullet} - \bar Y_{\bullet jk\bullet} + \bar Y_{i\bullet\bullet\bullet} + \bar Y_{\bullet j\bullet\bullet} + \bar Y_{\bullet\bullet k\bullet} - \bar Y_{\bullet\bullet\bullet\bullet}\end{aligned}$$

and the others are similar. In each case we subtract sub-effect estimates and average over variables not in the interaction of interest. The results are differences of certain data averages.

We also have an ANOVA identity. It is a bit ungainly:

$$\begin{aligned}\sum_{ijk\ell}(Y_{ijk\ell}-\hat\mu)^2 &= \sum_{ijk\ell}\hat\alpha_i^2+\sum_{ijk\ell}\hat\beta_j^2+\sum_{ijk\ell}\hat\gamma_k^2+\sum_{ijk\ell}\hat\delta_\ell^2\\ &+\sum_{ijk\ell}\widehat{\alpha\beta}_{ij}^2+\sum_{ijk\ell}\widehat{\alpha\gamma}_{ik}^2+\sum_{ijk\ell}\widehat{\alpha\delta}_{i\ell}^2+\sum_{ijk\ell}\widehat{\beta\gamma}_{jk}^2+\sum_{ijk\ell}\widehat{\beta\delta}_{j\ell}^2+\sum_{ijk\ell}\widehat{\gamma\delta}_{k\ell}^2\\ &+\sum_{ijk\ell}\widehat{\alpha\beta\gamma}_{ijk}^2+\sum_{ijk\ell}\widehat{\alpha\beta\delta}_{ij\ell}^2+\sum_{ijk\ell}\widehat{\alpha\gamma\delta}_{ik\ell}^2+\sum_{ijk\ell}\widehat{\beta\gamma\delta}_{jk\ell}^2\\ &+\sum_{ijk\ell}\widehat{\alpha\beta\gamma\delta}_{ijk\ell}^2+\sum_{ijk\ell}\hat\varepsilon_{ijk\ell}^2.\end{aligned}$$

For d factors there are 2^d-1 non-empty subsets of them that all explain some amount of variance that sums to the total variance among all N values. We ordinarily ignore $\sum_{ijk\ell}\hat\mu^2$ which when added to the above gives us $\sum_{ijk\ell}Y_{ijk\ell}^2$. The reason for this is that the grand mean has no impact on our choices of i or j or k or ℓ.

Later we will look at the ***functional ANOVA***. This was used by Hoeffding (1948) to study U-statistics, by Sobol' (1969) to study numerical integration and by Efron and Stein (1981) to study the jackknife. We will use a more abstract notation for it that simplifies some expressions. Suppose that $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$ where x_j are independent random inputs. Now let $Y = f(\boldsymbol{x})$. If $\mathbb{E}(Y^2) < \infty$ then $\mathrm{var}(f(\boldsymbol{x}))$ exists and there is a way to do an ANOVA on it. We proceed by analogy. The grand mean is $\mu = \mathbb{E}(f(\boldsymbol{x}))$. Then the main effect for variable j is

$$f_{\{j\}}(\boldsymbol{x}) = \mathbb{E}(f(\boldsymbol{x})-\mu\,|\,x_j) = \mathbb{E}(f(\boldsymbol{x})\,|\,x_j)-\mu$$

and the j,k interaction is

$$\begin{aligned}f_{\{j,k\}}(\boldsymbol{x}) &= \mathbb{E}(f(\boldsymbol{x})-\mu-f_{\{j\}}(\boldsymbol{x})-f_{\{k\}}(\boldsymbol{x})\,|\,x_j,x_k)\\ &= \mathbb{E}(f(\boldsymbol{x})\,|\,x_j,x_j)-\mu-f_{\{j\}}(\boldsymbol{x})-f_{\{k\}}(\boldsymbol{x}).\end{aligned}$$

We keep subtracting sub-effects and averaging over variables not in the interaction of interest. The x_j do not have to be categorical random variables like the levels of the factors we use above. They could be $\mathbb{U}[0,1]$ random variables or vectors, sound clips, images or genomes. All that matters is that they are independent and $f(\boldsymbol{x})$ is real valued with finite variance.

We will look at `https://statweb.stanford.edu/~owen/mc/A-anova.pdf` later when we get to computer experiments.

5.5 Replicates

Suppose that we were interested in a four-factor interaction. We could take $R \geqslant 2$ independent measurements at each ABCD combination. Then our model

would be

$$\begin{aligned} Y_{ijk\ell r} &= \mu + \alpha_i + \beta_j + \gamma_k + \delta_\ell \\ &\quad + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\alpha\delta)_{i\ell} + (\beta\gamma)_{jk} + (\beta\delta)_{j\ell} + (\gamma\delta)_{k\ell} \\ &\quad + (\alpha\beta\gamma)_{ijk} + (\alpha\beta\delta)_{ik\ell} + (\alpha\gamma\delta)_{ik\ell} + (\beta\gamma\delta)_{jk\ell} \\ &\quad + (\alpha\beta\gamma\delta)_{ijk\ell} + \varepsilon_{ijk\ell r}. \end{aligned} \tag{5.1}$$

where $1 \leqslant r \leqslant R$.

Suppose that we are making soup. We could go to store i, buy vegetables j, use recipe k, find person ℓ and have them taste the resulting pot of soup R times for $r = 1, \dots, R$. Or, on R separate days we could go to all I stores, buy all J vegetables at each store, try all K recipes on each set of vegetables from each store, and ask all L people to try all IJK of those soups once.

These are clearly quite different things. Not all kinds of replicate are equal. Equation (5.1) is a plausible model for the setting where the tasters taste each pot of soup R times in a row. It is an entirely unsuitable model for the setting where the whole experiment is completely redone R times. For that we would at a minimum want to include a main effect η_r for replicate r. There could even be a case for making the day of the experiment be its own fifth factor E with R levels. What is going on here is that the meaning of $r = 1$ is different in the two cases. In the first case it is just the first time one person tastes a given soup. In the second case it is one entire full replication of the experiment.

Just looking at the file of $N = IJKLR$ numbers we might not be able to tell which way the experiment was done. We will think more about replicates when they come up in specific examples.

5.6 High order ANOVA tables

If factors A and B have I and J levels, respectively then their interaction has $(I-1)(J-1)$ degrees of freedom. To see why consider this table

$(\alpha\beta)_{ij}$	j=1	j=2	j=3	j=4
i=1	✓	✓	✓	?
i=2	✓	✓	✓	?
i=3	?	?	?	??

where we've filled in the cells that are marked with a ✓. There are $(I-1)(J-1)$ of them. For any set of choices we make, the table can be completed by first making row sums zero and them making column sums zero. Had we omitted one of those $(I-1)(J-1)$ values, there would not be a unique way to complete the table.

Table 5.1 shows a portion of the ANOVA table for a four way factorial experiment where each cell has R independent values. We see $IJKL(R-1)$ degrees of freedom for error. We could get this by subtracting all of the other degrees of freedom numbers from $N - 1 = IJKLR - 1$. Or we can view it as gathering $R - 1$ degrees of freedom for error from each of $I \times J \times K \times L$ cells.

Source	DF
A or α	$I-1$
B or β	$J-1$
$\vdots$	$\vdots$
AB or $\alpha\beta$	(I-1)(J-1)
$\vdots$	$\vdots$
ABC or $\alpha\beta\gamma$	(I-1)(J-1)(K-1)
$\vdots$	$\vdots$
ABCD or $\alpha\beta\gamma\delta$	(I-1)(J-1)(K-1)(L-1)
Error	IJKL(R-1)
Total	IJKLR-1

Table 5.1: Selected rows of the ANOVA table for a four way table where each cell has R independent repeats.

It is clear from the above that these full factorial experiments are big and bulky and hence probably expensive. If $I = J = K = L = 11$ then we need $N = 14{,}641R$ observations. They will give us 10 df in each main effect, 100 per two factor interaction, 1000 df for each three factor interaction and 10,000 df in the four factor interaction which could be the least useful of them all.

Next we will look at what happens when all factors are at 2 levels. We still need $N = 2^d$ data points for d factors or $R2^d$ if we have replicated the experiment.

A common practice is to design an experiment that learns about the main effects and low order interactions partially or completely ignoring the high order interactions. To do this seems a bit hypocritical. We earlier argued that OAAT is bad because it can miss the two factor interactions and now we are getting ready to possibly ignore some other higher order interactions.

This common practice is a gamble. In statistics, we are usually against gambling and favor a cautious approach that covers all possibilities. However the cautious approach is really expensive and could be suboptimal. Experimental design has room for both bold and cautious choices. With a bold choice we do a small experiment that could be inexpensive or fast. If it goes well, then we learn more quickly. If it goes badly, then the experiment might not be informative at all and we have to do another one.

The bold strategies we will look at are motivated by a principle of ***factor sparsity***. This holds that the important quantities are mostly lower order and perhaps even many of the main effects are unimportant. Or at least relatively unimportant. If there are $2^{10} - 1$ effects and interactions they cannot all be relatively important! There is a related ***bet on sparsity*** principle (Friedman et al., 2001) motivating the use of the lasso. If things are sparse, you win. If they're not, maybe nothing would have worked.

5.7 Distributions of sums of squares

The ANOVA table involves a lot of algebraic manipulations. We get a table based on the values of $Y_{ijk\ell}$. We would rather have the table based on $\mathbb{E}(Y_{ijk\ell})$. So here we must study the distribution of the quantities in the table we get, using some model assumptions. The most common model assumptions are based on the Gaussian distribution. Sometimes that's robust due to the central limit theorem and sometimes it is not. Balanced experiments are more robust. The F-tests that we do for main effects and interactions are more robust than those we might do to compare variances of two sources of noise. Time permitting, we might cover this. For now, let's see how the mechanics of these tests work out.

These derivations are to the extent possible based on things that are easy to remember from an introductory regression class. To that we add some things about noncentral distributions that are not always included.

First, we recall that if $Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ then

$$\sum_{i=1}^{n} Z_i^2 \sim \chi^2_{(n)}.$$

Also, if $Q_j \overset{\text{ind}}{\sim} \chi^2_{(n_j)}$ then

$$F = \frac{Q_1/n_1}{Q_2/n_2} \sim F_{n_1,n_2}$$

and this is the very definition of the F distribution. The F distribution has numerator and denominator degrees of freedom and we write $F_{\text{num,den}}$ for the general case.

Now we consider noncentral distributions. These appear less often in introductory courses. The main place we need them is in power calculations, such as choosing a sample size. Choosing a sample size is maybe the simplest design problem. It does not however come up if we are just looking at pre-existing data sets.

If $Z_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, 1)$ then

$$\sum_{i=1}^{n} Z_i^2 \sim \chi'^2_{(n)}(\lambda)$$

where $\lambda = \sum_{i=1}^{n} \mu_i^2$. This is the ***noncentral*** χ^2 distribution on n degrees of freedom with noncentrality parameter λ. There are alternative parameterizations out there so we always have to check books, articles and software documentation to see which one was used.

If $Q_1 \sim \chi'^2_{(n_1)}(\lambda)$ and $Q_2 \sim \chi^2_{(n_2)}$ then

$$F' = \frac{Q_1/n_1}{Q_2/n_2} \sim F'_{n_1,n_2}(\lambda).$$

This is the ***noncentral F*** distribution. Here is how we use it. Our F statistics will have central numerators under H_0 but noncentral ones under H_A. They

will usually have central denominators under both because most of our denominators will involve sums of squares of differences of noise. If our noise model is wrong then the denominator might be noncentral. That would leave us with a ***doubly noncentral F*** distribution. That second noncentrality would make the denominator bigger and hence the F statistic smaller and might then mean lower statistical power.

Next we look at the distributions of sums of squares for simple one way layout with $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ for $i = 1, \dots, I$ and $j = 1, \dots, n$. Recall from introductory regression that when $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ then

$$(n-1)s^2 \equiv \sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \sigma^2 \chi^2_{(n-1)}.$$

Also if $Q_i \overset{\text{ind}}{\sim} \chi^2_{(n_i)}$ then

$$\sum_i Q_i \sim \chi^2_{(N)} \quad N = \sum_i n_i.$$

Now

$$\begin{aligned}
\text{SSE} &= \sum_{i=1}^{I}\sum_{j=1}^{n} (Y_{ij} - \bar{Y}_{i\bullet})^2 = \sum_{i=1}^{I}\sum_{j=1}^{n} \big((\mu + \alpha_i + \varepsilon_{ij}) - (\mu + \alpha_i + \bar{\varepsilon}_{i\bullet})\big)^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n} (\varepsilon_{ij} - \bar{\varepsilon}_{i\bullet})^2 = \sum_{i=1}^{I} Q_i \qquad \text{for } Q_i \overset{\text{iid}}{\sim} \sigma^2\chi^2_{(n-1)} \\
&\sim \sigma^2 \chi^2(I(n-1)).
\end{aligned}$$

Exercise: what is $\mathcal{L}(\text{SSE})$ in the unbalanced case with n_i observations at level i of the factor A? Notice that α_i did not affect SSE. They canceled out.

Next, under $H_0 : \alpha_i = 0$ we have

$$\begin{aligned}
\text{SSA} &= \sum_{i=1}^{I}\sum_{j=1}^{n} (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 = n\sum_{i=1}^{I} (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 \\
&= n\sum_{i=1}^{I} ((\mu + \alpha_i + \bar{\varepsilon}_{i\bullet}) - (\mu + \bar{\alpha}_\bullet + \bar{\varepsilon}_{\bullet\bullet}))^2 \\
&= n\sum_{i=1}^{I} (\bar{\varepsilon}_{i\bullet} - \bar{\varepsilon}_{\bullet\bullet})^2 \sim n\frac{\sigma^2}{n}\chi^2_{(I-1)} = \sigma^2\chi^2_{(I-1)}.
\end{aligned}$$

If H_0 does not hold then $\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet} = \alpha_i + \bar{\varepsilon}_{i\bullet} - \bar{\varepsilon}_{\bullet\bullet}$. Now

$$\bar{\varepsilon}_{i\bullet} \sim \mathcal{N}\Big(\alpha_i, \frac{\sigma^2}{n}\Big) \quad \text{and so} \quad \frac{\bar{\varepsilon}_{i\bullet}}{\sigma/\sqrt{n}} \sim \mathcal{N}\Big(\frac{\alpha_i}{\sigma/\sqrt{n}}, 1\Big).$$

Therefore

$$\sum_{i=1}^{I}\Big(\frac{\bar\varepsilon_{i\bullet}}{\sigma/\sqrt{n}}\Big)^2 \sim \chi'^2(\lambda) \quad \text{for } \lambda = \frac{n}{\sigma^2}\sum_{i=1}^{I}\alpha_i^2.$$

From this we find that

$$\sum_{i=1}^{I}\bar\varepsilon_{i\bullet}^2 \sim \frac{\sigma^2}{n}\chi'^2_{(I)}\Big(\frac{n}{\sigma^2}\sum_{i=1}^{I}\alpha_i^2\Big) \quad \text{and so} \quad \sum_{i=1}^{I}\sum_{j=1}^{n}\bar\varepsilon_{i\bullet}^2 \sim \sigma^2\chi'^2_{(I)}\Big(\frac{n}{\sigma^2}\sum_{i=1}^{I}\alpha_i^2\Big).$$

The answer we are going for is

$$\mathrm{SSE} = \sum_{i=1}^{I}\sum_{j=1}^{n}(\bar\varepsilon_{i\bullet}-\bar\varepsilon_{\bullet\bullet})^2 \sim \sigma^2\chi'^2_{(I-1)}\Big(\frac{n}{\sigma^2}\sum_{i=1}^{I}\alpha_i^2\Big).$$

Notice that the degrees of freedom drop by one for centered $\bar\varepsilon_{i\bullet}$. I have not found a nice way to get this using just things one might remember from an introductory regression class plus the noncentral distribution definitions.

Now let's look at our F statistic, under H_0:

$$\begin{aligned} F = \frac{\mathrm{MSA}}{\mathrm{MSE}} &= \frac{\frac{1}{I-1}\mathrm{SSA}}{\frac{1}{I(n-1)}\mathrm{SSE}} \\ &\sim \frac{\frac{1}{I-1}\sigma^2\chi^2_{(I-1)}}{\frac{1}{I(n-1)}\sigma^2\chi^2_{(I(n-1))}} \\ &= \frac{\frac{1}{I-1}\chi^2_{(I-1)}}{\frac{1}{I(n-1)}\chi^2_{(I(n-1))}} \\ &= F_{I-1,I(n-1)}. \end{aligned}$$

Under H_A,

$$F \sim F'_{I-1,I(n-1)}(\lambda) \qquad \lambda = \frac{n}{\sigma^2}\sum_{i=1}^{I}\alpha_i^2.$$

Under H_A, $\lambda > 0$ and the larger λ is the larger $\mathbb{E}(\mathrm{MSA})$ is. Thus H_A tends to increase F and so we should reject H_0 for unusually large values. We reject H_0 at level α when the observed value F_{obs} satisfies

$$F_{\mathrm{obs}} \geqslant F^{1-\alpha}_{I-1,I(n-1)}$$

We set a p-value of

$$p = \Pr(F_{I-1,I(n-1)} \geqslant F_{\mathrm{obs}})$$

where F_{obs} is the observed F statistic. The power of our test is

$$\Pr(F'_{I-1,I(n-1)}(\lambda) \geqslant F^{1-\alpha}_{I-1,I(n-1)}).$$

If we look at λ, then we see that it is a signal to noise ratio

$$\lambda = \frac{\sum_{i=1}^I \alpha_i^2}{\sigma^2/n} = I \times \frac{\frac{1}{I}\sum_{i=1}^I \alpha_i^2}{\sigma^2/n}.$$

Here $(1/I)\sum_{i=1}^I \alpha_i^2$ is the variance of α_i for a randomly chosen level i and σ^2/n is the variance of $\bar{\varepsilon}_{i\bullet}$. We get more power from larger n (get a bigger sample) or smaller σ^2 (e.g., buy better equipment) and from studying phenomena with larger effects.

5.8 Fixed and random effects

In an ANOVA we need to decide whether each of our factors represents a ***fixed effect*** or a ***random effect***. It is a fixed effect if we care about the exact set of levels in the experiment. That could be 3 headache pills, 4 gas additives or 5 machine operators. It is a random effect if we care about the population from which those levels were sampled. That could be 24 patients in a trial, 8 rolls of vinyl or those same 5 machine operators.

Whether we care about the levels in our experiment or a population that they represent depends on the uses we plan to make based on the data. While the health of 24 patients is very important, the reason for the trial is likely to be in order to learn how to treat some condition for people in general. Those 24 patients may then be viewed as a sample of the population to be treated. It is unlikely that they really are a random sample of the population to be treated, but hopefully they are reasonably representative. Once the 8 rolls of vinyl are used up and put into tested products we have no specific interest in them per se beyond how they represent the process that made them. For machine operators we might be interested in comparing 5 specfic employee's productivity. Or we might be interested in which of two machines will work better on average for a population of operators.

In a random effects model we write

$$Y_{ij} = \mu + a_i + \varepsilon_{ij}$$

where $a_i \sim \mathcal{N}(0, \sigma_A^2)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_E^2)$ are all independent. We might be interested in learning σ_A^2/σ_E^2 and the usual null hypothesis is $H_0 : \sigma_A^2 = 0$. The ANOVA is exactly the same as before. That is SST = SSB + SSW or SST = SSA + SSE are the two different notations we have used for the one factor ANOVA. This identity is just algebraic so it holds for any numbers we would put in. Using methods like those in Section 5.7 we find that

$$\text{SSW} \sim \sigma_E^2 \chi^2_{I(n-1)} \quad \text{and} \quad \text{SSB} \sim n(\sigma_A^2 + \sigma_E^2/n)\chi^2_{(I-1)}.$$

Now MSB/MSW $\sim F_{I-1,I(n-1)}$ under H_0 and is larger than that under H_A, though it is not noncentrally distributed any more. So we do the same F test as before. Not much changed.

Source	DF	Expected mean square
A	$I-1$	$\sigma_E^2 + n\sigma_{AB}^2 + nJ\sigma_A^2$
B	$J-1$	$\sigma_E^2 + nI\sigma_B^2$
AB	$(I-1)(J-1)$	$\sigma_E^2 + n\sigma_{AB}^2$
Err	$IJ(n-1)$	σ_E^2

Table 5.2: Expected values of the mean squares in a mixed effects model.

When there are two random effects we use this model

$$\begin{aligned} Y_{ijk} &= \mu + a_i + b_j + (ab)_{ij} + \varepsilon_{ijk} \\ a_i \sim \mathcal{N}(0,\sigma_A^2) &\qquad b_j \sim \mathcal{N}(0,\sigma_B^2) \\ (ab)_{ij} \sim \mathcal{N}(0,\sigma_{AB}^2) &\qquad \varepsilon_{ijk} \sim \mathcal{N}(0,\sigma_E^2) \end{aligned}$$

where all the Gaussian random variables are indpendent.

In the balanced case there are the same number n of observations at each ij combination. For $n \geqslant 2$ we find that

$$\begin{aligned} \text{MSA} &\sim \big(\sigma_E^2 + n\sigma_{AB}^2 + nJ\sigma_A^2\big)\frac{\chi^2_{I-1}}{I-1} \\ \text{MSB} &\sim \big(\sigma_E^2 + n\sigma_{AB}^2 + nI\sigma_B^2\big)\frac{\chi^2_{J-1}}{J-1} \\ \text{MSAB} &\sim \big(\sigma_E^2 + n\sigma_{AB}^2\big)\frac{\chi^2_{(I-1)(J-1)}}{(I-1)(J-1)}, \quad \text{and} \\ \text{MSE} &\sim \sigma_E^2\frac{\chi^2_{IJ(n-1)}}{IJ(n-1)}. \end{aligned}$$

If we use $F_A = \text{MSA}/\text{MSE}$, then we have a problem. It won't have the F distribution if $\sigma_A^2 = 0$ but $\sigma_{AB}^2 > 0$. What we must do instead is test it via $F_A = \text{MSA}/\text{MSAB}$. Similarly we test $\sigma_B^2 = 0$ using $F_B = \text{MSB}/\text{MSAB}$ and we test $\sigma_{AB}^2 = 0$ using $F_{AB} = \text{MSAB}/\text{MSE}$.

Now suppose that A is a fixed effect while B is a random effect. We are in for a bit of a surprise. This is called a ***mixed effects model***. Table 5.2 show the expected values of the mean squares for this case. What we see is that we should test the fixed effect A via the ratio MSA/MSAB and the random effect B via the ratio MSB/MSE. If the ***other effect*** is fixed use MSE in the denominator while if the other effect is random use MSAB.

To understand this result intuitively let's consider an extreme example. Suppose that the fixed effect A is about 3 headache medicines. We have 12 subjects in a random effect B. Each subject tests each medicine n times.

Let's exaggerate and suppose that $n = 10^6$. It naively looks like we have 36 million observations. But we only have 12 people in the data set. If we did let $n \to \infty$ then we would have a 3×12 table of exact means. Our sample size

would actually be 36. With a small sample n our data have to be less useful than those 36 values would have been.

Exercise: work out the limit as $n \to \infty$ of the test for A.

Bibliography

Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *Annals of Statistics*, 9(3):586–596.

Fisher, R. and Mackenzie, W. (1923). Studies in crop variation: The manurial response of different potato varieties. *Journal of Agricultural Sciences*, 13:311–320.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*. Springer, New York.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19:293–325.

Sobol', I. M. (1969). *Multidimensional Quadrature Formulas and Haar Functions*. Nauka, Moscow. (In Russian).