Contents

6

Two level factorials

Suppose that we have k factors to explore. If we go with 2 levels each that yields 2^k different treatment combinations which is the smallest possible product of k integers larger than 1. A variable could originally be binary, such as a choice between two polymers, or the choice between doing or not doing a step in a process. Or the variable could have more level or even take on a continuum of levels such as an oven temperature. For such a variable we could select two levels such as $1020°C$ and $1050°C$.

When we choose two levels for a continuous variable, some judgment must be exercised. If the effect of that variable on the response is monotone or even linear then we will have the greatest statistical power using widely separated values. The flip side is that we should avoid closely spaced values because then there will be low power and we might not get good relative accuracy on the effect. Then if we suspect a variable is not important we might use more widely separated values just to be able to check on that. If we know a good operating range then spanning that range makes sense for 'external validity' reasons. If we space things too far apart then our experimental analysis might give that variable too much importance compared to the others. The possibilities that we want to study might not be perfectly orthogonal. For instance if the experiment has an oven at two temperatures for two different time periods, then either the (LO,LO) combination or the (HI,HI) combination might be unsuitable. In that case one can use ***sliding levels*** such as 30 vs 50 minutes when the temperature is high and 45 vs 75 minutes when it is low. Wu and Hamada (2011) discuss these tradeoffs.

If we suspect curvature then two values will not be enough. We will look at 'response surface' designs later that let us estimate curved responses.

If we see each treatment combination$n$ times then we need to gather $n2^k$

Source	df
TRTs	$2^k - 1$
DAYS	$n - 1$
ERR	$(n-1)(2^k-1)$
TOT	$n2^k - 1$

Source	df
TRTs	$2^k - 1$
ERR	$2^k(n-1)$
TOT	$n2^k - 1$

Table 6.1: Left: Randomized blocks ANOVA table. Right: completely randomized design ANOVA table.

data points.

6.1 Replicates and more

Suppose that we get n data points from each of our 2^k treatment combinations. There are different ways that we could have done this and the way we did it can and should affect the analysis.

Suppose that we are making soup. We have two different stores, we use two different vegetables (carrots vs cabbage), cook at two different temperatures, and add two amounts of salt. There is more to our soup than just these factors but those other aspects are not varying. We now have $2^4 = 16$ precedures. The value Y is a measure of how good the soup tastes.

We are going to get $16n$ data points for some $n > 1$. One way to do this is to make all 16 kinds of soup on n different days. On each of those days we go to both stores, buy both kinds of vegetable cook at both temperature and use both salt levels. If we do it this way then we have 16 experimental treatments in n blocks, a ***randomized blocks*** design. Within each block (i.e., each day) we make the 16 soups in a random order.

We could also simply spend one day making soup 16 ways and taste each pot n times. This is really very different. It is sometimes called a ***repeated measures*** design. It is much faster, easier and cheaper than the blocked analysis but also much less informative. The order in which those tastes were made could be impactful, due to correlations between consecutive measurements. Maybe the soup is cooling off as we keep on tasting. We might just average those n values into one presumably better value.

A third way to do this is to make soup once on each of $16n$ different days. Each of the 16 combinations appears n times in a randomized order. This is a ***completely randomized design*** and it does not impose the balance that the randomized blocks version would. Table 6.1 shows ANOVA tables for the randomized blocks and completely randomized 2^k factorial experiment. If we did just average the repeated measures then we would not have any degrees of freedom for error and so Table 6.1 does not have a corresponding table. Later on we will see ways to analyze such an unreplicated experiment.

A B C D	Y
− − −−	(1)
+ − −−	a
− + −−	b
+ + −−	ab
− − +−	c
+ − +−	ac
− + +−	bc
+ + +−	abc
− − −+	d
+ − −+	ad
− + −+	bd
+ + −+	abd
− − ++	cd
+ − ++	acd
− + ++	bcd
+ + ++	abcd

Table 6.2: Standard notation for 16 runs in a 2^4 experiment.

6.2 Notation for 2^k experiments

Let's consider factors A, B, C and so forth at two levels each. Because these factors have 2 levels, they have $2-1=1$ degrees of freedom. Furthermore the interaction between a factor with I levels and one with J levels has $(I-1)(J-1)$ levels and so in this setting the two factor interactions also have 1 degree of freedom. Indeed, any interaction among any number of two level factors has just one degree of freedom.

There is something special about an effect having just one degree of freedom. We can attach a sign to it, and consider factors that increase or decrease a response. That is, given levels μ_1 and μ_2 we can compare them via $\mu_2-\mu_1$. We cannot easily and uniquely attach a sign to $(\mu_1-\mu_2, \mu_1-\mu_3, \mu_2-\mu_3) \in \mathbb{R}^3$. We could reduce them to $\mu_1-(\mu_2+\mu_3)/2$ and $\mu_3-\mu_2$ but that is just one of many potential comparisons to describe 2 degrees of freedom among three means.

To exploit this signing possibility we use a special notation for the observations in a 2^k experiment. For each factor we decide that one of its levels will be called the ***high level*** and the other will be the ***low level***. For a numerical value we would ordinarily take the higher value to be the high level. For an attribute like presence or absences of something, presence makes sense as a high level. In other settings the high level could be the new treatement, or the more expensive one, or the one that we anticipate is more likely to increase $\mathbb{E}(Y)$.

Table 6.2 shows all 16 possible combinations in a 2^4 experiment using $+$ for high levels and $-$ for low levels of factors A, B, C and D. The name of an observation is the string of letters at which it takes the high level. For instance "abd" is the observation at high levels of A and B and D with the low level

Y		A		B		AB	
(1)	b	−	−	−	+	+	−
a	ab	+	+	−	+	−	+

Table 6.3: The first square diagrams Y for a 2^2 experiment. The next ones show which observations are at high and low levels of A, B and AB.

of C. When everything is at the low level we use the symbol '(1)' instead of a blank or null string. We will be using some multiplicative formulas in which (1) makes a natural multiplicative identity.

The treatment effect for A is defined to be the average of $\mathbb{E}(Y)$ over 2^{k-1} observations at the high level of A minus the average of $\mathbb{E}(Y)$ over 2^{k-1} observations at the low level of A. It is called α_A and its estimate $\hat{\alpha}_A$ is the average of observed Y at the high level of A minus the average of observed Y at the low level of A. For instance with $k = 3$ and no replicates

$$\alpha_A = \frac{1}{4}\big(\mathbb{E}(Y_a) + \mathbb{E}(Y_{ab}) + \mathbb{E}(Y_{ac}) + \mathbb{E}(Y_{abc})\big) - \frac{1}{4}\big(\mathbb{E}(Y_{(1)}) + \mathbb{E}(Y_b) + \mathbb{E}(Y_c) + \mathbb{E}(Y_{bc})\big) \quad \text{and}$$
$$\hat{\alpha}_A = \frac{1}{4}\big(Y_a + Y_{ab} + Y_{ac} + Y_{abc}\big) - \frac{1}{4}\big(Y_{(1)} + Y_b + Y_c + Y_{bc}\big).$$

If there are replicates, we could call the obvservations $Y_{a,j}$ and $Y_{b,j}$ for $j = 1, \dots, n$ et cetera.

This is different from what we had before. With $k = 1$ we used to have $\mathbb{E}(Y_{1j}) = \mu + \alpha_1$ and $\mathbb{E}(Y_{1j}) = \mu + \alpha_2$ with $\alpha_1 + \alpha_2 = 0$. With that notation $\mathbb{E}(Y_{2j}) - \mathbb{E}(Y_{1j}) = \alpha_2 - \alpha_1 = 2\alpha_2$. Now we have $\mathbb{E}(Y_{a,j}) - \mathbb{E}(Y_{(1),j}) = \alpha_A$. That is $\alpha_a = 2\alpha_2$. In a 2^1 experiment we write

$$Y_{a,j} = \mu + \frac{1}{2}\alpha_a + \varepsilon_{a,j} \quad \text{and} \quad Y_{(1),j} = \mu - \frac{1}{2}\alpha_a + \varepsilon_{(1),j}.$$

Because interactions have one degree of freedom, we can give them a sign too. The AB interaction is the effect of A at the high level of B minus the effect of A at the low level of B. We can write it as

$$(ab - b) - (a - (1)) = ab - a - b + (1).$$

If instead we look at effect of B at the high level of A minus the effect of B at the low level of A we get

$$(ab - a) - (b - (1)) = ab - a - b + (1)$$

again That, is the AB interaction is the same as the BA interaction. Observations ab and (1) are at the high level of the AB interaction while a and b are at the low level. Table 6.3 depicts the AB interaction as a diagonal comparison in a 2×2 square of data values.

For a 2^2 experiment we may write the expected observatoins values as follows:

$$\begin{aligned}
\mathbb{E}(y_{ab}) &= \mu + \frac{1}{2}\big[\ \ \alpha_a + \alpha_b + \alpha_{ab}\big]\\
\mathbb{E}(y_{a}) &= \mu + \frac{1}{2}\big[\ \ \alpha_a - \alpha_b - \alpha_{ab}\big]\\
\mathbb{E}(y_{b}) &= \mu + \frac{1}{2}\big[-\alpha_a + \alpha_b - \alpha_{ab}\big] \quad \text{and}\\
\mathbb{E}(y_{(1)}) &= \mu + \frac{1}{2}\big[-\alpha_a - \alpha_b + \alpha_{ab}\big].
\end{aligned}$$

The estimated AB interaction parameter is $\hat\alpha_{ab}$ equal to the average of 2^{k-1} data values at the high level of the AB interaction minus the average of 2^{k-1} data values at the low level of the AB interaction. Now we find that

$$\begin{aligned}
&\mathrm{var}(\hat\alpha_a) = \mathrm{var}(\hat\alpha_b) = \mathrm{var}(\hat\alpha_{ab}) = \mathrm{var}(\hat\alpha_a) = \mathrm{var}(\hat\alpha_{abc}) = \cdots = \mathrm{var}(\hat\alpha_{abc\cdots z})\\
&\quad = \frac{\sigma^2}{N/2} + \frac{\sigma^2}{N/2} = \frac{4\sigma^2}{N} = \frac{4\sigma^2}{n2^k} = \frac{\sigma^2}{n2^{k-2}}.
\end{aligned}$$

Every main effect and every interaction of whatever order are all estimated with the same precision.

We can define interactions of all orders using this approach. The ABC interaction is the BC interaction at the high level of A minus the BC interaction at the low level of A:

$$\begin{aligned}
&(abc - ab - ac + a) - (bc - b - c + (1))\\
&= \underbrace{(abc + a + b + c)}_{\text{HI level of ABC}} - \underbrace{(ab + ac + bc + (1))}_{\text{LO level of ABC}}.
\end{aligned}$$

Just like other main effects and interactions, half the data are at the high level of ABC and half are at the low level.

Because each effect has one degree of freedom, we can use a t test for it instead of an F test. The t statistic for $H_0 : \alpha_a = 0$ is

$$\frac{\hat\alpha_a}{s/\sqrt{2^{k-2}n}},$$

where s is the standard error. The degrees of freedom are $(n-1)(2^k-1)$ if the experiment was run in n blocks of 2^k runs (and the model should then include a block effect) and the degrees of freedom are $n2^k - 1$ if it is a completely randomized allocation.

6.3 Why $n = 1$ is popular

If we take $n = 1$ then there are no replicates and there are then 0 degrees of freedom for error. Were we to compute s^2 we would get 0/0 (not a number).

To see why $n = 1$ is popular suppose that we have a choice between a 2^4 experiment to study factors ABCD with 2 replicates, and a 2^5 experiment to study factors ABCDE without replication. Given the chance to explore one more factor without a variance estimate many people would do that. Experimental design has room for both cautious approaches and bold ones.

In class we considered a famous example about an experiment to improve ball bearings. The engineers were planning $n = 4$ plant runs for two levels of one factor. A statistician persuaded them to use those same 8 plant runs to investigate 3 factors in a 2^3 layout. The additional factors proved to be much more effective than the originally contemplated one. The story is at this link `https://en.wikipedia.org/wiki/Factorial_experiment` after scrolling down a bit. What I like about this story is that it has the statistician playing the part of the bold person. As always, there is selection bias in the stories that get included in a course or text book or article.

There is a notion of ***factor sparsity*** that underlies this decision. The idea is that most of the α's are close to 0. Then the estimates that we get are mostly $\mathcal{N}(0, 4\sigma^2/N)$ plus a few real effects providing outliers. Then using QQ-plots of the estimated effects or their absolute values we can hope to spot the outliers. This approach was developed by Cuthbert Daniel (1959). Even if the α's are not exactly zero, it could happen that most of them are relatively small. They cannot all be relatively important. Suppose that the small ones can be modeled by effects that are $\mathcal{N}(\mu_*, \sigma_*^2)$. Then their estimates will look like $\mathcal{N}(\mu_*, \sigma_*^2 + 4\sigma^2/2^k)$. If the large effects have α values that are much greater than $|\mu_*| + \sqrt{\sigma_*^2 + 4\sigma^2/N}$ then they will still look like outliers.

When this factor sparsity is in play, then it would be wasteful to take $n > 1$ replicate instead of inspecting additional factors. If you replicate and leave out the most important factor, that seemingly safe choice can be very costly.

It is common to see that the important interactions in a QQ plot involve the important main effects. We might see $\hat{\alpha}_A$, $\hat{\alpha}_B$, $\hat{\alpha}_C$ and $\hat{\alpha}_{BC}$ as the outliers. In class we saw such a QQ plot in a the fractional factorial class.

Daniels' article closes with an interesting comment about why we just analyze one response at a time:

> The third set of questions plaguing the conscience of one applied statistician concerns the repeated use of univariate statistical methods instead of a single multivariate system. I know of no industrial product that has but one important property. And yet, mainly because of convenience, partly because of ignorance, and perhaps partly because of lack of full development of methods applicable to industrial research problems, I find myself using one-at-a-time methods on responses, even at the same time that I derogate a one-at-a-time approach to the factors influencing a response. To what extent does this simplification invalidate the judgments I make concerning the effects of all factors on all responses?

Factor sparsity supports a concept called ***design projection***. Suppose that in a 2^3 experiment that factor A is completely null: $\alpha_a = \alpha_{ab} = \alpha_{ac} = \alpha_{abc} = 0$.

Then our experiment is just like a 2^2 experiment in B and C with two replicates, depicted as follows:

$$\begin{bmatrix} A & B & C \\ -1 & -1 & -1 \\ -1 & -1 & +1 \\ -1 & +1 & -1 \\ -1 & +1 & +1 \\ +1 & -1 & -1 \\ +1 & -1 & +1 \\ +1 & +1 & -1 \\ +1 & +1 & +1 \end{bmatrix} \longrightarrow \begin{bmatrix} B & C \\ -1 & -1 \\ -1 & +1 \\ +1 & -1 \\ +1 & +1 \\ -1 & -1 \\ -1 & +1 \\ +1 & -1 \\ +1 & +1 \end{bmatrix} = \begin{bmatrix} B & C \\ -1 & -1 \\ -1 & -1 \\ -1 & +1 \\ -1 & +1 \\ +1 & -1 \\ +1 & -1 \\ +1 & +1 \\ +1 & +1 \end{bmatrix}.$$

In the rightmost array the runs are reordered to show the replication. The same thing happens if either B or C is null. In the multi-response setting that Daniels' was concerned with, there might be a different null factor for each of the responses being measured.

In practice it is unlikely that A is perfectly null but it might be relatively null with some other $|\alpha|$'s being orders of magnitude larger than the ones involving A. It could be tricky to decide whether A is null and would involve the issues raised in the next section on factorial experiments with no replicates.

6.4 Factorial with no replicates

If we had and $I \times J$ experiment run just $n = 1$ time, then our $N = I \times J \times 1$ data points could be used to get SSA, SSB and SSE. If the factors interact, then SSE will be inflated by that interaction. It will have a noncentral χ^2 distribution (under our Gaussian assumption). Then $F = \text{MSA}/\text{MSE}$ will be doubly noncentral. Noncentrality in SSE will make MSE larger, making F smaller, making for less signficance (larger p values). So our test would be conservative having lower power. If the resulting p-value is below 0.01 or other threshold then it likely would have been even further below it in the event of no interaction.

In the 2^k experiment we could designate a priori some high order interactions to use in forming an error estimate. We could take a mean square of the corresponding $\hat{\alpha}$'s as our estimate of $4\sigma^2/N$. This would have the same conservative property as the doubly noncentral F described above.

We could possibly just take the 10 smallest $\hat{\alpha}^2$'s but then we would have to adjust them for this selection. For instance, a Monte Carlo computation can find for us the distribution of $\sum_{j=1}^{10}(\hat{\alpha}^2)_{(i)}/(4\sigma^2/N)$.

There are lots of methods to get approximate inferences without replication. Hamada and Balakrishnan (1998) describe 24 different methods. One prominent one is due to Lenth (1989). In recent years there has been much interest in inference after model selection and perhaps progress in that area could lead to some new methods here.

y	M	A	B	C	AB	AC	BC	ABC
(1)	+	−	−	−	+	+	+	−
a	+	+	−	−	−	−	+	+
b	+	−	+	−	−	+	−	+
ab	+	+	+	−	+	−	−	−
c	+	−	−	+	+	−	−	+
ac	+	+	−	+	−	+	−	−
bc	+	−	+	+	−	−	+	−
abc	+	+	+	+	+	+	+	+

Table 6.4: Each row is an observation in a $2^k = 2^3$ experiment. Each column shows an effect of interest with M representing the grand mean. The estimate of that effect is the inner product of the corresponding column of ± 1's divided by 2^{k-1}, except that for the grand mean (column M) we divide by 2^k.

6.5 Generalized interactions

For $n = 1$, the effects that we are interested in are dot products of a vector of Y values with some coefficients ± 1 and then divided by 2^k for $\hat{\mu}$ or 2^{k-1} for the αs. Table 6.4 shows those coefficents for $k = 3$.

Using these signs we define some ***generalized interactions***. The generalized interaction of AB and BC is what we get if we mutiply the column of signs for AB times that for BC. We get $AB \times BC = AB^2C = AC$ because B^2 is a column of +1s. Similarly $AB \times ABC = C$.

We can compare our 2^k effects to a plain linear regression model

$$Y = X\beta + \varepsilon \in \mathbb{R}^N$$

with parameter estimates $\hat{\beta} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}Y$ and $\mathrm{var}(\hat{\beta}) = (X^\mathsf{T}X)^{-1}\sigma^2$ We could take X to be the matrix in Table 6.4 or a generalization of it to some other k. Now $X^\mathsf{T}X = 2^k \times I_{2^k}$ so $\hat{\beta}$ would have $\hat{\mu}$ in it and $2\hat{\alpha}$ for the other effects.

What we see here is that all of the $\hat{\beta}_j$ (and hence all of our effects) are uncorrelated because $\mathrm{var}(\hat{\beta}) = \sigma^2 I_{2^k}/2^k$. Ordinarily a regression with N observations and p parameters has a computational cost proportional to Np^2. Here $N = p = 2^k$ and the cost would be $O(N^3)$. In this instance, the orthogonality of X reduces the cost to just that of p one at a time regressions actually dot products, and the cost is $O(N^2)$.

Yates's algorithm (Yates, 1937) reduces the cost to $O(N\log(N))$. When $N = 16$ we are not much interested in the cost difference between N^2 and $N\log(N)$. If however $N = 2^{30}$ as it might in an experiment carried out solely in software, then Yates' algorithm is very useful. His algorithm is a fast Fourier transform used for hand calculations decades before the fast Fourier transform became prominent in signal processing.

We can get an algebraic expression to show which observations are at the high and low levels of an interaction. Consider the BCD interaction. We can

b	ab	(1)	a
c	ac	bc	abc
d	ad	cd	acd
bcd	abcd	bd	abd

Table 6.5: The 8 observations on the left are at the high level of BCD. The 8 observations on the right are at the low level of BCD.

write it as

$$\begin{aligned} BCD &= \frac{1}{2^{k-1}}(a+1)(b-1)(c-1)(d-1) \\ &= \frac{1}{2^{k-1}}(ab+b-1-b)(cd-c-d+1) \\ &= \cdots \\ &= \frac{1}{2^{k-1}}(b+c+d+bcd+ab+ac+ad+abcd) \\ &\quad - \frac{1}{2^{k-1}}((1)+bc+cd+bd+a+abc+acd+abd). \end{aligned}$$

The high and low levels of BCD are depicted in Table 6.5. What we see are that bcd is at the high level. Anything that is missing an odd number of those letters is at the low level. Anything missing an even number is at the high level (just like zero which is even). The presence of another factor like A does not change it.

6.6 Blocking

We might want to (or have to) run our 2^k experiment in blocks. One common choice has blocks of size 2^{k-1}. So our experiment has 2 blocks of size 2^{k-1}. Blocking could be done for greater accuracy (balancing out some source of noise) or we might be forced into it: the equipment holds exactly 8 pieces, no more.

The best plan for 2 blocks is run one block at the high level of the k-factor interaction and one at the low level. For $k = 3$ it would be

$$\boxed{a \quad b \quad c \quad abc} \quad \text{vs} \quad \boxed{1 \quad ab \quad ac \quad bc}$$

with run order randomized within the blocks. With this choice the ABC interaction is confounded with the blocks. The effects for A, B, C, AB, BC and AC are orthogonal to blocks (because they're orthogonal to ABC).

If we only have these 3 factors we might replicate the block structure as follows:

$$\boxed{\begin{array}{|c|c|}\hline abc & ab \\ a & ac \\ b & bc \\ c & 1 \\ \hline\end{array}} \quad \boxed{\begin{array}{|c|c|}\hline abc & ab \\ a & ac \\ b & bc \\ c & 1 \\ \hline\end{array}} \quad \boxed{\begin{array}{|c|c|}\hline abc & ab \\ a & ac \\ b & bc \\ c & 1 \\ \hline\end{array}}$$

Source	df
Replicates	2
Blocks = ABC	1
Blocks × replicates	2
A, B, C, AB, AC, BC	1 each
Error	12
Total	23

Table 6.6: This is the ANOVA table for a blocked 2^3 experiment with $n = 3$ replicates each having 2 blocks.

Of course if there were additional factors to investigate we might introduce those instead of replicating our blocked 2^3 experiment.

The ANOVA table for this setup is shown in Table 6.6. With 3 replicates, there are 2 degrees of freedom for replicates. One degree of freedom goes to comparing the block at the high level of ABC to the one at the low level. There are also 2 degrees of freedom for blocks by replicates and 1 degree of freedom for each of our main effects and two factor interactions. This table will let us study the main effects and two factor interactions.

To study the 3 factor interaction we can test block level averages. We form the 2×3 matrix

$$\begin{array}{c} \\ \text{ABC}{=}+ \\ \text{ABC}{=}- \end{array} \begin{array}{c} \begin{array}{ccc} r{=}1 & r{=}2 & r{=}3 \end{array} \\ \begin{bmatrix} \bar{Y}_{+1\bullet} & \bar{Y}_{+2\bullet} & \bar{Y}_{+3\bullet} \\ \bar{Y}_{-1\bullet} & \bar{Y}_{-2\bullet} & \bar{Y}_{-3\bullet} \end{bmatrix} \end{array}$$

and test the row effect. Wu and Hamada (2011) warn that there will be trouble if the blocks interact with treatments. That will also mess up other blocked designs.

With many replicates, there is also the possibility of confounding ABC with blocks in some replicates and other interactions such as BC in different replicates. The analysis of such combinations could get pretty complicated. In class I mentioned that it might be wise to simply use a probabilistic programming language (e.g., Stan or others mentioned here `https://en.wikipedia.org/wiki/Probabilistic_programming#Probabilistic_programming_languages`) to analyse such a thing instead of wrangling mean squares and ANOVA tables. That would handle the analysis. To actually choose a design one could simulate data from it many times and pipe the simulated output through the probabilistic programming language.

Bibliography

Daniel, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, 1(4):311–341.

Hamada, M. and Balakrishnan, N. (1998). Analyzing unreplicated factorial experiments: a review with some new proposals. *Statistica Sinica*, pages 1–28.

Lenth, R. V. (1989). Quick and easy analysis of unreplicated factorials. *Technometrics*, 31(4):469–473.

Wu, C. J. and Hamada, M. S. (2011). *Experiments: planning, analysis, and optimization*, volume 552. John Wiley & Sons.

Yates, F. (1937). The design and analysis of factorial experiments. Technical Report 35, Imperial Bureau of Soil Science.