POINTS OF SIGNIFICANCE

Two-level factorial experiments

Simultaneous examination of multiple factors at two levels can reveal which have an effect.

Byran Smucker, Martin Krzywinski and Naomi Altman

Two-level factorial experiments, in which all combinations of multiple factor levels are used, efficiently estimate factor effects and detect interactions—desirable statistical qualities that can provide deep insight into a system. This gives them an edge over the widely used one-factor-at-a-time experimental approach, which is statistically inefficient and unable to detect interactions because it sequentially varies each factor individually while all the others are held constant.

Suppose that we would like to determine which of three candidate compounds (factors) have an effect on cell differentiation (response) and also estimate their interactions. In this case, two levels for each compound suffice: low (or zero) and high concentration, giving $2^3 = 8$ factor-level combinations (treatments). The levels for each compound should be as far apart as possible so that the effect size will be as large as possible. However, the common assumption that the response and the factor level are linearly related might not be true when the distance between factor levels is large. Thus, for accuracy, complicated designs may call for levels that are closer together. If in doubt, increase the chance of detecting factor effects by choosing levels that are too far apart, rather than too close.

Let's name our factors A, B and C, and use -1 and $+1$ for the low and high levels, respectively (Table 1). Even though there is no replication, this 2^3 full factorial design can detect factor effects, if some sensible assumptions are made.

A key quantity to estimate is the main effect, which is the average difference in response between the high and low levels of a factor. For example, we compute the main effect for A as -1.2 by taking the average of the responses when $A = +1$ (-0.063) and subtracting from it the average of the responses when $A = -1$ ($+1.1$). Equivalently, one can compute main effect estimates by taking the inner product of their column and the response column, and then dividing that value by $n/2$, where n is the number of runs. Note that this effect estimate measures the change in response mean when the factor changes by two units (from -1 to $+1$), whereas a regression parameter estimate

Table 1 | The factor-level combinations in a 2^3 full factorial experiment

Run	Intercept	Main effect			Interaction effect				y
		A	B	C	AB	AC	BC	ABC	
1	+1	-1	-1	-1	+1	+1	+1	-1	0.67
2	+1	-1	-1	+1	+1	-1	-1	+1	1.85
3	+1	-1	+1	-1	-1	+1	-1	+1	0.93
4	+1	-1	+1	+1	-1	-1	+1	-1	0.98
5	+1	+1	-1	-1	-1	-1	+1	+1	-0.38
6	+1	+1	-1	+1	-1	+1	-1	-1	-0.17
7	+1	+1	+1	-1	+1	-1	-1	-1	-0.09
8	+1	+1	+1	+1	+1	+1	+1	+1	0.39
Effect size		-1.2	0.057	0.48	0.36	-0.13	-0.21	0.35	

The y column shows a simulated response for each run using the model $y = 0.5 - 0.5A + 0.15B + 0.2C + \varepsilon$, where ε is distributed normally with mean 0 and s.d. 0.3. The "effect size" row shows the estimated factorial effects. The runs should be performed in random order.

measures the change in response when the factor changes by one unit¹. For instance, the true regression coefficient in our model for A is -0.5 , and the regression parameter estimate is $-1.2/2 = -0.6$.

The products of the main effect columns yield interaction columns (Table 1), whose effects can be calculated in the same way as the main effects. For example, though the true effect of the AB interaction is 0, its estimate is 0.36, which is the difference between the average response when the constituent factors have the same sign (that is, $AB = +1$) and the average when their sign is different ($AB = -1$). All main effects and interactions are uncorrelated with all other effect estimates, which is evident from the fact that their columns in Table 1 are all pairwise orthogonal (the inner product of any pair is zero).

Once the factorial effects have been computed, the natural question is whether they are large enough to be of statistical and scientific interest. A model can be fit using linear regression^{1,2}, but because in a 2^k full factorial experiment there are as many runs (2^k) as factorial terms (in our 2^3 example, there are 3 main effects, 3 two-factor interactions, 1 three-factor interaction and the intercept), the fitted values are just the observed data. Thus, if all factorial terms are included in the model, traditional regression-based inferences cannot be made

because there is no estimate of residual error. In a three-factor experiment, this issue can be addressed by replication, but for larger studies this might be infeasible owing to the large number of treatments.

Various methods exist to address inference in factorial experiments. Simple graphical examination (e.g., using a Pareto plot, which shows both absolute and cumulative effect sizes) can provide considerable information about important effects. A more formal method is to model only some of the factorial effects; this approach depends on the reasonable and empirically validated assumptions of effect sparsity and effect hierarchy³. Effect sparsity tells us that in factorial experiments, most of the factorial effects are likely to be unimportant. Effect hierarchy tells us that low-order terms (e.g., main effects) tend to be larger than higher-order terms (interactions). Application of these assumptions yields a reasonable analysis strategy: fit only the main effects and two-factor interactions, and use the degrees of freedom from the unmodeled higher-order interactions to estimate residual error.

We illustrate this by simulating a 2^6 full factorial design (64 runs) with the model $y = 1.5 - 0.5A + 0.15C + 0.65F + 0.2AB - 0.5AF + \varepsilon$, where ε is the same as in our 2^3 model (Table 1). Note that we have simulated only three factors (A, C and F)

Fig. 1 | The effect estimates for a 2^6 factorial design. The intercept fit is not shown. **a**, A full 2^6 factorial with fits to all terms. **b**, A full 2^6 factorial with fits to main and two-factor effects only. Bar color indicates inference of true positive (blue), true negative (gray) and false positive (orange) significant observations (tested at $P < 0.05$). **c**, Fractional 2^{6-1} factorial fitting to main and two-factor effects only. Color-coding denotes inference as in **b**, with the addition of a false negative (red). The horizontal scale for **b,c** is the same as in **a**. The factor order is the same for all panels, in descending order of effect in **a**.

and two interactions (AB and AF) to have an effect. The fit to all factorial effects provides strong visual evidence that F, A, AF and AB are important (Fig. 1a); the effect of C is uncertain, as its magnitude is similar to that of many inert effects.

If we apply the strategy of modeling only the main effects and two-factor interactions, we get $64 - (1 + 6 + 15) = 42$ degrees of freedom for error that can be used for inference (Fig. 1b). Obviously, we will not be able to detect any interactions of three or more factors. If these interactions are large, our error estimate will be inflated and our inferences will be conservative. However, on the basis of effect hierarchy, we are

willing to assume that these higher-order terms are not important. This model shows that the C regression parameter estimate of 0.11 is significant ($P = 0.01$), but also incorrectly identifies the BF estimate (0.09) as significant ($P = 0.04$). Whether considered visually or more formally via regression, the most important effects are identified, with ambiguities for a few smaller estimated effects.

For interpretations of interaction effects—how factors influence the effects of other factors—interaction plots are useful (Fig. 2). For example, the large AF interaction of -0.38 (Fig. 1a,b) tells us that the level of A has an important effect on the effect of F. Given that the regression main effect estimates of A and F are -0.56 and 0.70 , respectively, if $A = -1$, then the estimated change in the mean response for a unit change in F is $0.70 + 0.38 = 1.08$, whereas when $A = +1$, the change due to F is just $0.70 - 0.38 = 0.32$.

Full factorial designs grow large as the number of factors increases, but we can use fractional factorial designs to reduce the number of runs required by considering only a fraction of the full factorial runs (e.g., half as many in a 2^{6-1} design). These runs are chosen carefully so that under the reasonable assumptions of effect sparsity and hierarchy, the terms of interest (e.g., main effects and two-factor interactions) can be estimated.

For example, consider runs 2, 3, 5 and 8 in Table 1, which have $ABC = +1$. If we have only these four runs, we cannot distinguish the intercept from the ABC interaction (they are completely confounded) because they have the same factor levels (their inner products with the response are identical). Within these runs, A is completely confounded with BC, B with AC, and C with AB. Thus, if we found that the $A = BC$ effect was important, we would be unsure of whether this was due to a significant effect of A or of the BC interaction. However, the effect-hierarchy principle would suggest that A is probably driving the result rather than BC.

We can apply the same reasoning in a 2^6 experiment to remove half the runs. In the 32-run 2^{6-1} fractional factorial design there are 32 confounding relations (e.g., ABCDEF with the intercept, A with BCDEF, etc.), and, importantly, all of the main effects and two-factor interactions are confounded with four- and five-factor interactions. Given our assumption that these high-order effects are unlikely to be important, we have little worry that they will contaminate our estimate of the main effects and two-factor interactions.

Even if we fit the intercept, all main effects and all 15 two-factor interactions,

Fig. 2 | Interaction plot of factors A and F from the 2^6 full factorial simulation. The estimated change in mean response for one unit change in F is 1.08 when $A = -1$ (blue) and 0.32 when $A = +1$ (black).

we're still left with $32 - 22 = 10$ degrees of freedom for inference on these factorial effects (Fig. 2c), similar to the process for the full set of 64 runs (Fig. 2a,b), but with half the number of runs. With further assumptions about the model hierarchy, even smaller fractions of the full factorial experiment can provide useful information about the main effects and some interactions.

Two-level fractional factorial designs provide efficient experiments to screen a moderate number of factors when many of the factorial effects are assumed to be unimportant (sparsity) and when an effect hierarchy can be assumed. They are simple to design and analyze, while providing information that can be used to inform more detailed follow-up experiments using only the factors found to be important. More details on full and fractional factorial designs can be found in ref. 4.

Byran Smucker¹, Martin Krzywinski^{2*} and Naomi Altman³

¹Department of Statistics, Miami University, Oxford, OH, USA. ²Michael Smith Genome Sciences Centre, Vancouver, British Columbia, Canada. ³Department of Statistics, The Pennsylvania State University, University Park, PA, USA.

*e-mail: martink@bcgsc.ca

Published online: 27 February 2019
<https://doi.org/10.1038/s41592-019-0335-9>

References

- Altman, N. & Krzywinski, M. *Nat. Methods* **12**, 999–1000 (2015).
- Krzywinski, M. & Altman, N. *Nat. Methods* **12**, 1103–1104 (2015).
- Li, X., Sudarsanam, N. & Frey, D. *Complexity* **11**, 32–45 (2006).
- Mee, R. A. *A Comprehensive Guide to Factorial Two-level Experimentation* (Springer-Verlag, New York, 2009).

Competing interests

The authors declare no competing interests.