ANALYZING UNREPLICATED FACTORIAL EXPERIMENTS:
A REVIEW WITH SIMULATION COMPARISON

Presented by Quoc Tran
Based on

Outline

• Motivation

• Notation and List of Methods on Comparison
  – Important Notation
  – Methods on Comparison
    * Existing Methods
    * Modifications and New Proposals

• Comparison
  – Grouping The Methods
  – A Simulation Study

• Conclusion and Discussion

• Summary and Comments
Motivation

• **Starting point**: Since the 1980s, the objective analysis of unreplicated two-level factorial and fractional factorial designs has attracted much attention. 24 methods are listed.

• **Why**: Unreplicated experiment is economical. But, there are no degrees of freedom left to estimate the error variance. Consequently, standard t tests cannot be used to identify the active effects.


• **What**: This paper reviews various methods for analyzing unreplicated experiments.

• **How**: By simulation study on an equitable basis.
Important Notation

- $n$: number of runs.
- $k$: number of contrasts
- $\sigma$: error standard deviation
- $\tau$: contrast standard error
- $\kappa_i$: $i$th effect
- $c_i$: $i$th contrast - estimate of $i$th effect
- $p_i$: probability of declaring $i$ effects active under all inert effects
- EER: experimentwise error rate
- IER: individual error rate
- #AE: number of active effects
Existing Methods

- BEN89: Benski (1989)
- BI92: Bissell (1992)
- BM86: Box and Meyer (1986)
- BP91: Berk and Picard (1991)
- DAN59: Daniel (1959)
- DONG93: Dong (1993)
- HB69: Holms and Berrettoni (1969)
- JP92: Juan and Pena (1992)
- JTUK87: Johnson and Tukey (1987)
- LEN89: Lenth (1989)
- SKW93: Schneider et al. (1993)
- STUK82: Seheult and Tukey (1982)
• VS96: Venter and Steel (1996)
• ZAHN75: Zahn (1975a, version S)
• ZAHN75(m): uses m smallest contrasts to estimate $\tau$
Modifications and New Proposals

- CORR: correlation coefficient probability plot
- DISP: dispersion test
- HLOH92: half-normal version of Loh (1992)
- HSW: half-normal Shapiro-Wilk
- MDONG: uses iterative DONG93 estimator
- MLEN: iterative LEN89
- MLZ92: modified Le and Zamar
- MSKW: modified SKW93
Comparison - Grouping The Methods

• Directed vs. Composite:
  – Directed methods test the individual effects directly.
    * DAN59, DONG93, JP92, LEN89, SKW93 and ZAHN75 (also the modified versions MDONG, MLEN, MSKW) standardize the contrasts by various estimates of $\sigma$.
    * BM86 and JTUK87 use individual posterior probabilities and ratio-to-scale statistics, respectively.
    * HB69, BP91 and VS96 use mean squares
  – Composite methods test all the effects as a group. These include BI89, BI92, CORR, DISP, HSW, MLZ92, as well as the first parts of BEN89 and HLOH92.

• Sequential BI89, BI92, DAN59, HB69, JP92, JTUK87, MLZ92, STUK82, and ZAHN75 as proposed are sequential meaning that some computation is done at each stage with the remaining contrasts. Others are not.
Comparison - A Simulation Study (10,000 simulations)

Table 1. Off-the-shelf performance of existing methods $p_i = \text{observed proportion of simulations detecting } i \text{ effects under all inert effects for } 16 \text{ run design}$ (* indicates $\geq 8$ declared effects)

<table>
<thead>
<tr>
<th>method</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>IER</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEN89</td>
<td>.975</td>
<td>.020</td>
<td>.003</td>
<td>.002</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.0022</td>
<td>.025</td>
<td></td>
</tr>
<tr>
<td>BI89</td>
<td>.948</td>
<td>.023</td>
<td>.007</td>
<td>.004</td>
<td>.002</td>
<td>.003</td>
<td>.002</td>
<td>.002</td>
<td>.009*</td>
<td>.0139</td>
<td>.052</td>
</tr>
<tr>
<td>BI92</td>
<td>.834</td>
<td>.118</td>
<td>.032</td>
<td>.011</td>
<td>.004</td>
<td>.001</td>
<td>.000</td>
<td></td>
<td></td>
<td>.0157</td>
<td>.166</td>
</tr>
<tr>
<td>BM86</td>
<td>.748</td>
<td>.176</td>
<td>.044</td>
<td>.016</td>
<td>.007</td>
<td>.004</td>
<td>.003</td>
<td>.002</td>
<td>.000</td>
<td>.0262</td>
<td>.252</td>
</tr>
<tr>
<td>BP91</td>
<td>.555</td>
<td>.259</td>
<td>.119</td>
<td>.050</td>
<td>.017</td>
<td>.004</td>
<td>.000</td>
<td></td>
<td></td>
<td>.0492</td>
<td>.445</td>
</tr>
<tr>
<td>DAN59</td>
<td>.598</td>
<td>.193</td>
<td>.093</td>
<td>.050</td>
<td>.065</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.0527</td>
<td>.402</td>
</tr>
<tr>
<td>DONG93</td>
<td>.569</td>
<td>.302</td>
<td>.085</td>
<td>.029</td>
<td>.011</td>
<td>.004</td>
<td>.001</td>
<td>.000</td>
<td></td>
<td>.0418</td>
<td>.431</td>
</tr>
<tr>
<td>HB69</td>
<td>.629</td>
<td>.157</td>
<td>.067</td>
<td>.045</td>
<td>.033</td>
<td>.029</td>
<td>.042</td>
<td></td>
<td></td>
<td>.0634</td>
<td>.371</td>
</tr>
<tr>
<td>HLOH92</td>
<td>.951</td>
<td>.017</td>
<td>.018</td>
<td>.010</td>
<td>.004</td>
<td>.001</td>
<td>.000</td>
<td></td>
<td></td>
<td>.0070</td>
<td>.049</td>
</tr>
<tr>
<td>LEN89</td>
<td>.755</td>
<td>.144</td>
<td>.054</td>
<td>.024</td>
<td>.013</td>
<td>.007</td>
<td>.003</td>
<td>.001</td>
<td></td>
<td>.0290</td>
<td>.245</td>
</tr>
<tr>
<td>JP92</td>
<td>.799</td>
<td>.104</td>
<td>.039</td>
<td>.021</td>
<td>.014</td>
<td>.010</td>
<td>.006</td>
<td>.004</td>
<td>.003*</td>
<td>.0294</td>
<td>.201</td>
</tr>
<tr>
<td>JTUK87</td>
<td>.950</td>
<td>.034</td>
<td>.010</td>
<td>.003</td>
<td>.002</td>
<td>.001</td>
<td>.001</td>
<td>.000</td>
<td>.000</td>
<td>.0054</td>
<td>.050</td>
</tr>
<tr>
<td>MLZ92</td>
<td>.953</td>
<td>.023</td>
<td>.007</td>
<td>.007</td>
<td>.003</td>
<td>.003</td>
<td>.002</td>
<td>.001</td>
<td>.001</td>
<td>.0074</td>
<td>.047</td>
</tr>
<tr>
<td>SKW93</td>
<td>.590</td>
<td>.254</td>
<td>.105</td>
<td>.038</td>
<td>.011</td>
<td>.002</td>
<td>.000</td>
<td></td>
<td></td>
<td>.0421</td>
<td>.410</td>
</tr>
<tr>
<td>STUK82</td>
<td>.742</td>
<td>.129</td>
<td>.054</td>
<td>.026</td>
<td>.017</td>
<td>.012</td>
<td>.008</td>
<td>.005</td>
<td>.005*</td>
<td>.0387</td>
<td>.258</td>
</tr>
<tr>
<td>VS92</td>
<td>.900</td>
<td>.083</td>
<td>.015</td>
<td>.001</td>
<td>.000</td>
<td>.000</td>
<td></td>
<td></td>
<td></td>
<td>.0079</td>
<td>.100</td>
</tr>
<tr>
<td>ZAHN75</td>
<td>.618</td>
<td>.190</td>
<td>.089</td>
<td>.048</td>
<td>.055</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.0487</td>
<td>.382</td>
</tr>
</tbody>
</table>
Comparison - Equitable Basis

- Let $p_i$ denote the proportion of simulations for $i$ inert effects declared active.

- Then EER is proportion of the simulations when one or more effects is declared active, $1 - p_0$.

- IER is the average proportion of inactive effects declared active, $(i/(n - 1))p_i$.

- Note that the EER and IER given in Table 1 vary across the different methods. These methods try to control either IER or EER or both.

- The author try to tune the parameters of all these methods so that they have the same IER for all inert effects (=0.044) without destroying the essence of the methods. The paper also provides details for the particular versions of the methods used in the study.
Comparison - Equitable Basis

Figure 1. EER, IER for $N(0, 1)$ and std. $t(9)$ errors (1st, 2nd of pair) and $n = 16$ no active effects ($e$ = EER, $i$ = IER)
Comparison - Power Investigation (10,000 simulations) - one active effect

Figure 2. Power for $N(0, 1)$ and std. $t(9)$ errors (1st, 2nd of pair) and $n = 16$
one active effect = 0, 5(.5)3, 4 σ (labels 0-7)
Comparison - Power Investigation (10,000 simulations) - two active effect

Figure 3. Power for $N(0,1)$ and std. $t(9)$ errors (1st, 2nd of pair) and $n = 16$ two active effects $= 0, .5(.5)3, 4 \sigma$ (labels 0-7)
Comparison - Power Investigation (10,000 simulations) - four active effect

Figure 4. Power for $N(0,1)$ and std. $t(9)$ errors (1st, 2nd of pair) and $n = 16$ four active effects = 0, .5(.5)3, 4 $\sigma$ (labels 0-7)
Comparison - Power Investigation (10,000 simulations) - six active effect

Figure 5. Power for \( N(0,1) \) and std. \( t(9) \) errors (1st, 2nd of pair) and \( n = 16 \)
six active effects = 0, .5(.5)3, 4 \( \sigma \) (labels 0-7)
Conclusion

• There is little difference between the methods for small size effects, say $0.5\sigma$, which exhibit little power. Also, there is not much of a difference for large size effects, say $4\sigma$. There are marked differences between the methods for intermediate size effects ($\sigma$ to $3\sigma$) which become more pronounced for larger size effects in this range as the $\#AE$ increases; there is less of a difference for smaller size effects as the $\#AE$ increases.

• The power decreases as the number of active effects increases.

• Except for BI92 (see the six active effect case), the power increases as the size of the active effects increases. Note that the effects need to be rather large relative to the process standard deviation $\sigma$. For example, the power is around 0.7 for a single $1.5\sigma$ effect.

• The directed methods which focus on the current largest unsigned contrast perform better than the composite methods, BI89, BI92, MLZAM, CORR, DISP and HSW. MLZ92 is an exception, performs surprisingly well, especially for large $\#AE$. CORR is clearly the worst of all the composite methods.
Conclusion (cont.)

• BI92, a composite procedure appears promising say for up to four active effects but then its performance seriously degrades for six active effects. This is clearly an undesirable property.

• Many of the directed methods only differ in the estimator used for $\tau$. Various proposals were motivated by better MSE properties of the estimators. Yet, the gains in estimator performance appear to have little impact on the test performance. Rather, the $#AE$ seems to affect two groups of the methods differently. DAN59, DONG94, MDONG and SKW93 do better than JP92, LEN89, MLEN, MSKW and ZAHN75 for small $#AE$ whereas the latter group perform better for large $#AE$. Overall, DAN59 and SKW93 perform the best in the first group. ZAHN75 performs the best in the second group although there is not much of a difference between these methods.

• Among the other directed methods (BM86, JTUK87, HB69, BP91, VS96), BP91 performs the best. BM86 is quite competitive for small $#AE$; BM86 performs poorly for six active effects, it was designed for three active effects.
Conclusion (cont.)

- The modified procedures MLEN and MSKW provide little if any improvement over LEN89 and SKW92, except for large \( #AE \).

- MDONG has almost the same power as DONG93 for small \( #AE \), and actually performs worse as \( #AE \) increases. Thus, there is no real benefit offered by the iteration in estimating \( \tau \).

- STUK82 is quite competitive with the best of the other methods for \( #AE \) up to four and is slightly worse for six active effects. Recall that it is at a slight disadvantage since its IER is 0.038 as compared with 0.044 for the other methods.

- Among the two hybrid methods, HLOH92 performs better than BEN89 although BEN89s IER and EER is smaller. BEN89s performance degrades dramatically for large \( #AE \). HLOH92 also outperforms BP91 which was adjusted to have the same IER of 0.007. Recall that BP91 is one of the best methods.
Discussion on evaluating new procedures

- A simulation study is needed to evaluate the performance of the new method with existing ones. Calibration of the new procedure needs to be done in order to provide a fair comparison. At a minimum, it should report the $p_i$ values under all inert effects.

- In addition to examining the power of the proposed procedure, its IER behavior should also be studied.

- One should check to see if the method is exploiting the following properties of the contrasts: (1) The contrasts have equal variances and are normally distributed. (2) Contrasts for inert effects have zero means while those for estimating active effects have non-zero means.

- The method should not depend on the arbitrariness of the factor level labels. For example, methods that work with unsigned or squared contrasts avoids this problem.
Discussion on issues and possibility of connections with other areas

- What are desirable EER, IER and $p_i$ when all effects are inert? For example, using an IER of 0.044 for $n = 16$, some methods had an EER of 0.40, which some might consider large.

- Other measures of overall performance for both a procedures ability to detect active effects as well as its tendency to identify inactive effects as active?


- Are there non-sequential procedures which have better performance or are sequential directed tests preferable?

- Can gains be made by combining methods, i.e., hybrid methods? The simulation study showed that the half normal version of Loh (1992) is promising.
Discussion on issues and possibility of connections with other areas (cont.)

- Can information on how many active effects there are likely to be present in the experiment be exploited?

- How robust are the methods to other nonnormal distributions? In such a situation, would a nonparametric method be preferable? For example, Loughin and Noble (1997) propose a permutation test procedure.
Specific Recommendations

- For up to six active effects, overall DAN59, SKW93, ZAHN75, BP91 and HLOH92 performed well, although others are competitive if one has a good idea about how many active effects there are. For example, for eight active effects, the versions of DAN59 and ZAHN75 used here would not be expected to do well since they assumed there would be no more than six active effects.

- The power for BI92 seriously degrades when there are many active effects so that this method is not recommended. For the same reasons, BEN89 is also not recommended.

- Especially for \( n = 8 \), the substantial variability exhibited in the probability plots when all effects are inert makes it difficult to both identify the active effects and to not choose the inert effects. Objective methods are therefore preferable. Nevertheless, Balakrishnan and Hamada (1994) showed that for such a small run size, the active effects need to be large relative to the process standard deviation for any hope of detecting them. Consequently, a larger run size \( (n = 16) \) is recommended.
Summary and Comments

- A very extensive review concerning a wide variety of methods used to identify active effects in unreplicated factorial experiments.

- The complexity of this problem stems for the multiple types of comparisons and hypotheses that can be made.

- Hamada and Balakrishnan have chosen to compare the different methods by measuring only their statistical power while fixing the number of experimental runs at 16 and tuning all the methods so that their IER is controlled at 0.044.

- The author have proposed and used a common notation for describing the various methods and their properties. This sets a high standard for future work.
Summary and Comments

• Active effects as fixed. A more realistic approach is to consider active effects as random.

• Need a comparison of these methods arises from how these designs are used in practice.

• Nested design.

• Active effects are of the same magnitude, the number of runs is fixed.

• Due to practical constraints, a Monte Carlo study is often very limited in scope, and this one is no exception.
References


Technometrics **31**, 469-473.

