Simulation Studies for Statistical Procedures: Why Can't We Practice What We Preach?

Hugh Chipman

Acadia University

April 15, 2015

Joint work with Pritam Ranjan, Fatimah Al-Ahmad. Thanks also to R. Jock MacKay.

Motivating example

78 K. KRISHNAMOORTHY, AVISHEK MALLICK, AND THOMAS MATHEW

Table 1. Type I error rates of the tests for a lognormal mean under Type I censoring for the choice $\mu = 0$; L = left-tailed test, R = right-tailed test, T = two-tailed test (p_0 is the proportion of left-censored observations)

		$\sigma = 1$			$\sigma = 2$			$\sigma = 3$		
p_0	Method	L	R	T	L	R	T	L	R	T
						$n = 20$				
0.2	GV	0.054	0.044	0.050	0.060	0.050	0.052	0.049	0.049	0.053
	AN	0.021	0.105	0.082	0.004	0.124	0.091	0.002	0.152	0.115
	SLRT	0.046	0.069	0.059	0.043	0.040	0.040	0.035	0.077	0.062
	MSLRT	0.053	0.042	0.047	0.057	0.039	0.049	0.062	0.038	0.052
0.3	GV	0.053	0.044	0.058	0.059	0.040	0.052	0.055	0.049	0.050
	AN	0.021	0.110	0.083	0.004	0.137	0.102	0.001	0.151	0.117
	SLRT	0.044	0.058	0.051	0.034	0.064	0.051	0.040	0.063	0.051
	MSLRT	0.053	0.036	0.043	0.060	0.037	0.046	0.064	0.038	0.052
0.5	GV	0.059	0.036	0.044	0.080	0.033	0.053	0.055	0.054	0.055
	AN	0.022	0.108	0.085	0.002	0.157	0.122	0.000	0.176	0.139
	SLRT	0.048	0.059	0.056	0.041	0.066	0.058	0.039	0.070	0.056
	MSLRT	0.049	0.029	0.036	0.060	0.028	0.051	0.078	0.027	0.053
0.7	GV	0.062	0.016	0.043	0.070	0.023	0.053	0.051	0.044	0.046
	AN	0.039	0.003	0.017	0.001	0.197	0.164	0.000	0.213	0.184
	SLRT	0.052	0.040	0.043	0.043	0.070	0.056	0.033	0.073	0.054
	MSLRT	0.044	0.030	0.037	0.079	0.014	0.046	0.102	0.015	0.061
						$n = 30$				
0.2	GV	0.054	0.047	0.054	0.055	0.046	0.049	0.060	0.044	0.052
	AN	0.028	0.096	0.071	0.009	0.119	0.084	0.006	0.130	0.091
	SLRT	0.046	0.060	0.055	0.043	0.065	0.056	0.037	0.066	0.052
	MSLRT	0.051	0.044	0.047	0.056	0.041	0.049	0.055	0.042	0.049
0.3	GV	0.056	0.046	0.054	0.060	0.045	0.050	0.054	0.040	0.048
	AN	0.025	0.092	0.069	0.009	0.126	0.092	0.003	0.138	0.104
	SLRT	0.050	0.061	0.054	0.041	0.064	0.052	0.039	0.067	0.053
	MSLRT	0.055	0.036	0.044	0.063	0.041	0.052	0.065	0.039	0.050
0.5	GV	0.060	0.039	0.052	0.062	0.033	0.050	0.061	0.033	0.043
	AN	0.024	0.098	0.076	0.004	0.141	0.107	0.002	0.149	0.116
	SLRT	0.052	0.039	0.051	0.047	0.066	0.057	0.036	0.068	0.054
	MSLRT	0.048	0.033	0.039	0.064	0.034	0.047	0.073	0.033	0.052
0.7	GV	0.066	0.024	0.048	0.073	0.040	0.053	0.075	0.022	0.047
	AN	0.039	0.042	0.023	0.002	0.168	0.138	0.000	0.190	0.156
	SLRT	0.045	0.040	0.042	0.040	0.062	0.053	0.036	0.069	0.052
	MSLRT	0.046	0.027	0.032	0.073	0.018	0.046	0.088	0.018	0.055
						$n = 50$				
0.2	GV	0.053	0.053	0.048	0.051	0.050	0.049	0.055	0.041	0.044
	AN	0.028	0.080	0.060	0.017	0.100	0.067	0.012	0.108	0.073
	SLRT	0.043	0.059	0.055	0.043	0.060	0.054	0.043	0.063	0.056
	MSLRT	0.051	0.047	0.049	0.058	0.044	0.050	0.057	0.048	0.053
0.3	GV	0.067	0.044	0.049	0.066	0.047	0.063	0.054	0.053	0.057
	AN	0.026	0.083	0.060	0.014	0.101	0.071	0.008	0.115	0.081
	SLRT	0.049	0.059	0.058	0.041	0.057	0.051	0.044	0.059	0.054
	MSLRT	0.050	0.047	0.048	0.056	0.041	0.048	0.057	0.039	0.049
0.5	GV	0.061	0.041	0.052	0.066	0.032	0.051	0.062	0.044	0.049
	AN	0.027	0.086	0.065	0.008	0.125	0.095	0.004	0.124	0.095
	SLRT	0.049	0.058	0.051	0.044	0.064	0.055	0.042	0.067	0.056
	MSLRT	0.050	0.039	0.043	0.062	0.036	0.052	0.062	0.037	0.051
0.7	GV	0.079	0.015	0.048	0.080	0.040	0.054	0.073	0.022	0.053
	AN	0.035	0.056	0.040	0.004	0.143	0.110	0.001	0.151	0.119
	SLRT	0.048	0.045	0.047	0.042	0.058	0.049	0.037	0.066	0.052
	MSLRT	0.047	0.041	0.042	0.065	0.024	0.044	0.071	0.025	0.049

TECHNOMETRICS, FEBRUARY 2011, VOL. 53, NO. 1

Krishnamoorthy, Mallick and Mathew (2011, Technometrics)

Inference for the Lognormal Mean and Quantiles Based on Samples with Left and Right Type I Censoring

Simulation study: size of a nominal $\alpha = 0.05$ test for a lognormal mean.

Table: Type I error, for 432 different combinations of 5 experimental factors.

Results summarized by text, e.g. "The test based on the asymptotic normality of the MLE seems to be the worst among all tests."

Motivating example

Table 1. Type I error rates of the tests for a lognormal mean under Type I censoring for the choice $\mu = 0$; L = left-tailed test, R = right-tailed test, T = two-tailed test (p_0 is the proportion of left-censored observations)

		$\sigma = 1$			$\sigma = 2$			$\sigma = 3$		
p_0	Method	L	R	T	L	R	T	L	R	T
						$n = 20$				
0.2	GV	0.054	0.044	0.050	0.060	0.050	0.052	0.049	0.049	0.053
	AN	0.021	0.105	0.082	0.004	0.124	0.091	0.002	0.152	0.115
	SLRT	0.046	0.069	0.059	0.043	0.040	0.040	0.035	0.077	0.062
	MSLRT	0.053	0.042	0.047	0.057	0.039	0.049	0.062	0.038	0.052
0.3	GV	0.053	0.044	0.058	0.059	0.040	0.052	0.055	0.049	0.050
	AN	0.021	0.110	0.083	0.004	0.137	0.102	0.001	0.151	0.117
	SLRT	0.044	0.058	0.051	0.034	0.064	0.051	0.040	0.063	0.051
	MSLRT	0.053	0.036	0.043	0.060	0.037	0.046	0.064	0.038	0.052
0.5	GV	0.059	0.036	0.044	0.080	0.033	0.053	0.055	0.054	0.055
	AN	0.022	0.108	0.085	0.002	0.157	0.122	0.000	0.176	0.139
	SLRT	0.048	0.059	0.056	0.041	0.066	0.058	0.039	0.070	0.056
	MSLRT	0.049	0.029	0.036	0.060	0.028	0.051	0.078	0.027	0.053
0.7	GV	0.062	0.016	0.043	0.070	0.023	0.053	0.051	0.044	0.046
	AN	0.039	0.003	0.017	0.001	0.197	0.164	0.000	0.213	0.184
	SLRT	0.052	0.040	0.043	0.043	0.070	0.056	0.033	0.073	0.054
	MSLRT	0.044	0.030	0.037	0.079	0.014	0.046	0.102	0.015	0.061
						$n = 30$				
0.2	GV	0.054	0.047	0.054	0.055	0.046	0.049	0.060	0.044	0.052
	AN	0.028	0.096	0.071	0.009	0.119	0.084	0.006	0.130	0.091
	SLRT	0.046	0.060	0.055	0.043	0.065	0.056	0.037	0.066	0.052
	MSLRT	0.051	0.044	0.047	0.056	0.041	0.049	0.055	0.042	0.049
0.3	GV	0.056	0.046	0.054	0.060	0.045	0.050	0.054	0.040	0.048
	AN	0.025	0.092	0.069	0.009	0.126	0.092	0.003	0.138	0.104
	SLRT	0.050	0.061	0.054	0.041	0.064	0.052	0.039	0.067	0.053
	MSLRT	0.055	0.036	0.044	0.063	0.041	0.052	0.065	0.039	0.050
0.5	GV	0.060	0.039	0.052	0.062	0.033	0.050	0.061	0.033	0.043
	AN	0.024	0.098	0.076	0.004	0.141	0.107	0.002	0.149	0.116
	SLRT	0.052	0.039	0.051	0.047	0.066	0.057	0.036	0.068	0.054
	MSLRT	0.048	0.033	0.039	0.064	0.034	0.047	0.073	0.033	0.052
0.7	GV	0.066	0.024	0.048	0.073	0.040	0.053	0.075	0.022	0.047
	AN	0.039	0.042	0.023	0.002	0.168	0.138	0.000	0.190	0.156
	SLRT	0.045	0.040	0.042	0.040	0.062	0.053	0.036	0.069	0.052
	MSLRT	0.046	0.027	0.032	0.073	0.018	0.046	0.088	0.018	0.055
						$n = 50$				
0.2	GV	0.053	0.053	0.048	0.051	0.050	0.049	0.055	0.041	0.044
	AN	0.028	0.080	0.060	0.017	0.100	0.067	0.012	0.108	0.073
	SLRT	0.043	0.059	0.055	0.043	0.060	0.054	0.043	0.063	0.056
	MSLRT	0.051	0.047	0.049	0.058	0.044	0.050	0.057	0.048	0.053
0.3	GV	0.067	0.044	0.049	0.066	0.047	0.063	0.054	0.053	0.057
	AN	0.026	0.083	0.060	0.014	0.101	0.071	0.008	0.115	0.081
	SLRT	0.049	0.059	0.058	0.041	0.057	0.051	0.044	0.059	0.054
	MSLRT	0.050	0.047	0.048	0.056	0.041	0.048	0.057	0.039	0.049
0.5	GV	0.061	0.041	0.052	0.066	0.032	0.051	0.062	0.044	0.049
	AN	0.027	0.086	0.065	0.008	0.125	0.095	0.004	0.124	0.095
	SLRT	0.049	0.058	0.051	0.044	0.064	0.055	0.042	0.067	0.056
	MSLRT	0.050	0.039	0.043	0.062	0.036	0.052	0.062	0.037	0.051
0.7	GV	0.079	0.015	0.048	0.080	0.040	0.054	0.073	0.022	0.053
	AN	0.035	0.056	0.040	0.004	0.143	0.110	0.001	0.151	0.119
	SLRT	0.048	0.045	0.047	0.042	0.058	0.049	0.037	0.066	0.052
	MSLRT	0.047	0.041	0.042	0.065	0.024	0.044	0.071	0.025	0.049

Study looked at several comparisons:

- Tail of test (L/R/two-sided)
  3 levels
- Sample size
  3 levels
- Population variance σ^2
  3 levels
- Censoring level ("p0")
  4 levels
- 4 different hypothesis tests
  4 levels

Full factorial design with
$3 \times 3 \times 3 \times 4 \times 4 = 432$ runs

This simulation study is a designed experiment

Viewing the study as a designed experiment leads me to ask some questions:

1. **Design:** Are so many runs necessary? Could we reduce the number of runs and/or use a fractional factorial?
2. **Analysis:** Why not use a statistical analysis to report results instead of presenting a massive table?

For this example, let's try to answer these questions, with "design of experiments 101" tools:

1. Analysis of full factorial experiment.
2. Design of smaller study.
3. Re-analysis of smaller study.
4. Repeat #2 and #3 with fractional factorial.

Analysis of full factorial experiment

Include:

- main effects ($2 + 2 + 2 + 3 + 3 = 12$ df)
- two-factor interactions (57 df)
- three-factor interactions (134 df)

... leaving 228 df for residuals

(main effects: $R^2 = 22.3\%$, 2fi: $R^2 = 82.5\%$, 3fi: $R^2 = 95.4\%$)

And we might as well remove insignificant terms from the model.

Analysis of full factorial experiment

ANOVA table, ordered by Mean SS terms:

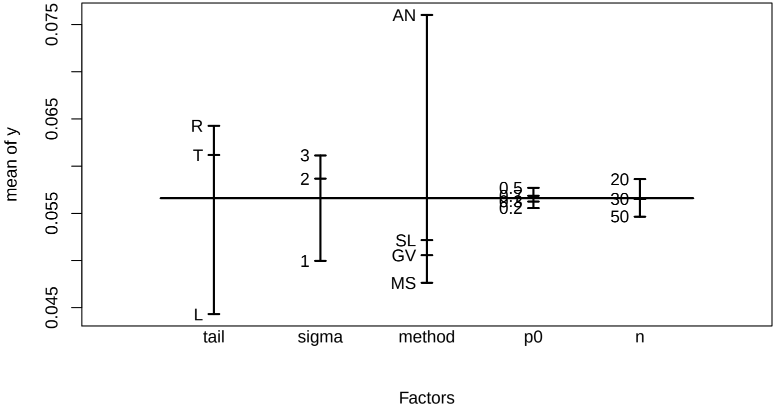
```
summary(aov(y ~ (tail + sigma + method + p0 + n)^3,data=mydata))
                    Df   Sum Sq  Mean Sq  F value     Pr(>F)
tail:method          6 0.225805 0.037634 470.3605 < 2.2e-16 ***
method               3 0.055513 0.018504 231.2728 < 2.2e-16 ***
tail                 2 0.033227 0.016613 207.6383 < 2.2e-16 ***
sigma                2 0.009920 0.004960  61.9892 < 2.2e-16 ***
tail:sigma:method   12 0.029806 0.002484  31.0431 < 2.2e-16 ***
sigma:method         6 0.013771 0.002295  28.6862 < 2.2e-16 ***
tail:sigma           4 0.009015 0.002254  28.1669 < 2.2e-16 ***
sigma:p0             6 0.008071 0.001345  16.8126 < 2.2e-16 ***
method:n             6 0.005038 0.000840  10.4933 1.619e-10 ***
sigma:method:p0     18 0.010329 0.000574   7.1721 4.240e-15 ***
n                    2 0.001137 0.000569   7.1077 0.0009726 ***
...
Residuals          284 0.022723 0.000080
```

Analysis of full factorial experiment

Plot of main effects:

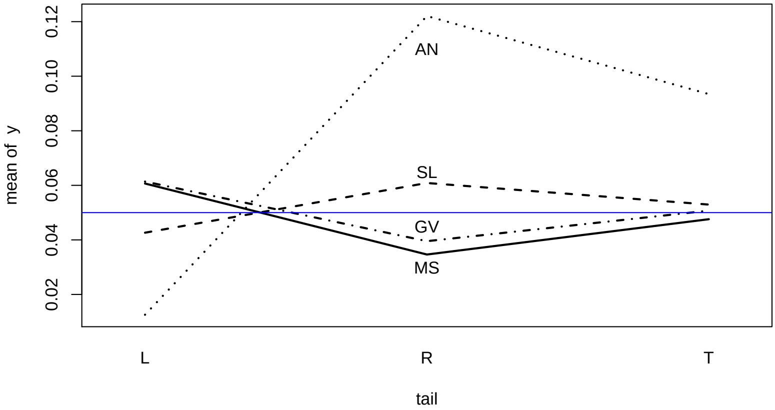
Large effects: `tail`, `sigma`, `method`

Small effects: censoring (`p0`) and `n`.



Analysis of full factorial experiment

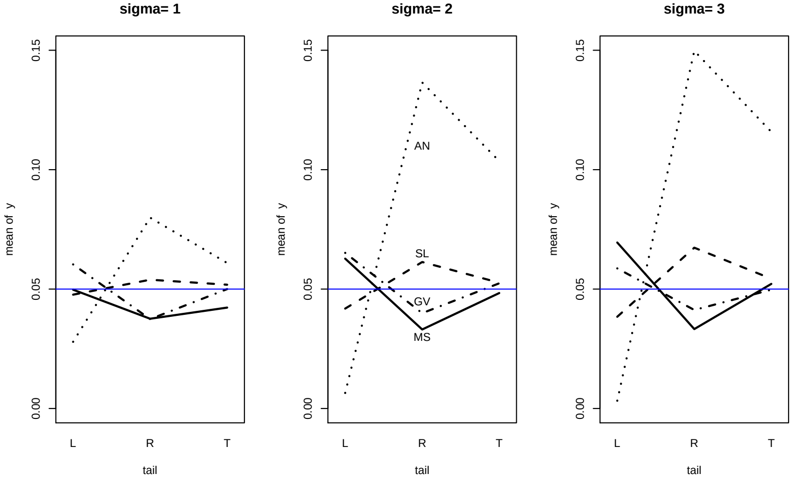
Interaction plot for `tail:method`



Analysis of full factorial experiment

Interaction plot for `tail:method:sigma`

Asymptotic Normal ("AN") test has higly variable α-level.



Design of smaller study

Full factorial with fewer levels?

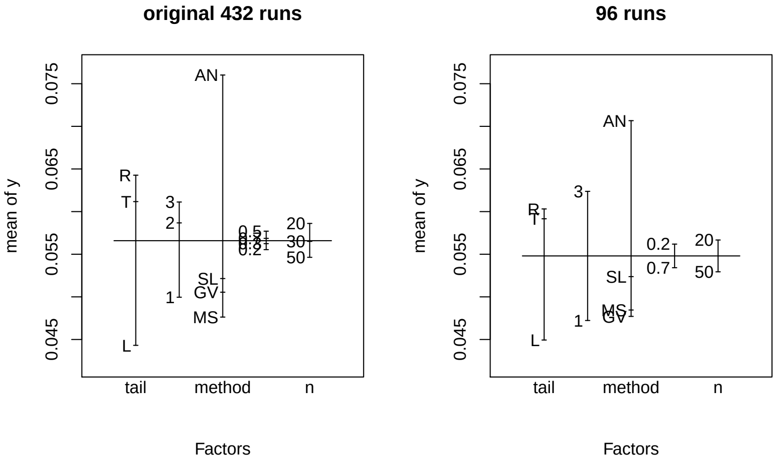
We had (and can reduce to):

- Tail of test (L/R/two-sided): 3 levels
- 4 different hypothesis tests: 4 levels
- Sample size: 3 levels [reduce to 2 levels]
- Population variance σ^2: 3 levels [reduce to 2 levels]
- Censoring level ("p0"): 4 levels [reduce to 2 levels]

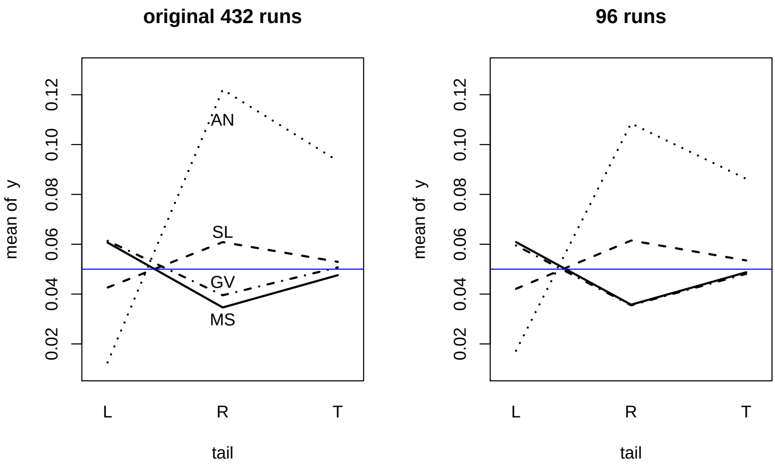
So we go from $3 \times 3 \times 3 \times 4 \times 4 = 432$ runs to $3 \times 2 \times 2 \times 2 \times 4 = 96$ runs.

Note that because we have results for the full factorial, we can "run" the simplified design.

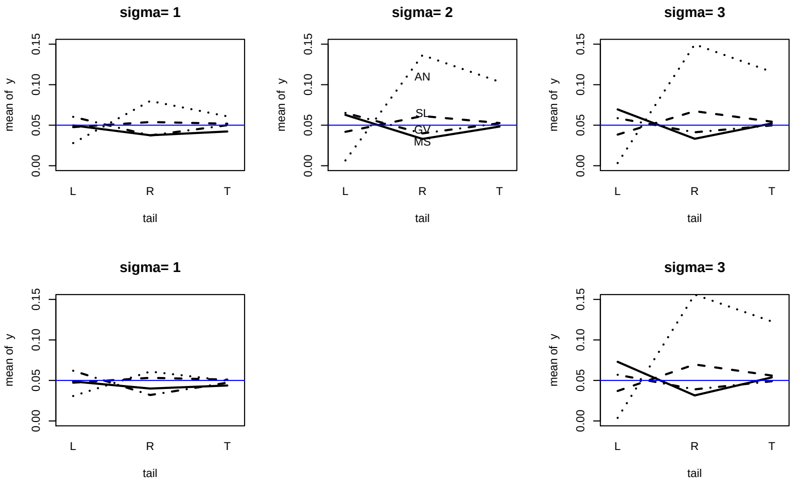
Analysis of small full factorial experiment



Analysis of small full factorial experiment



Analysis of small full factorial experiment



Analysis of small full factorial experiment

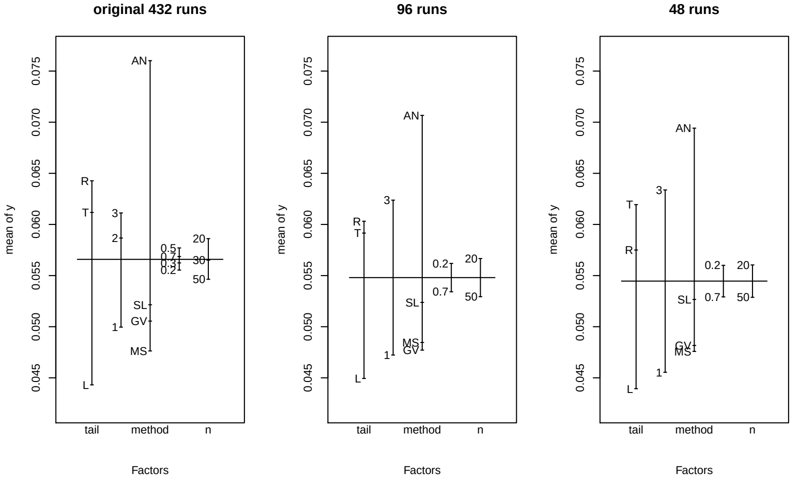
Conclusions:

- Similar significant terms.
- Analysis has less power, but most terms still significant.
- `tail`, `sigma`, `method` and interactions still most important.
- Dropping levels of numeric factors (`n`, `sigma`, `p0`) didn't hurt.

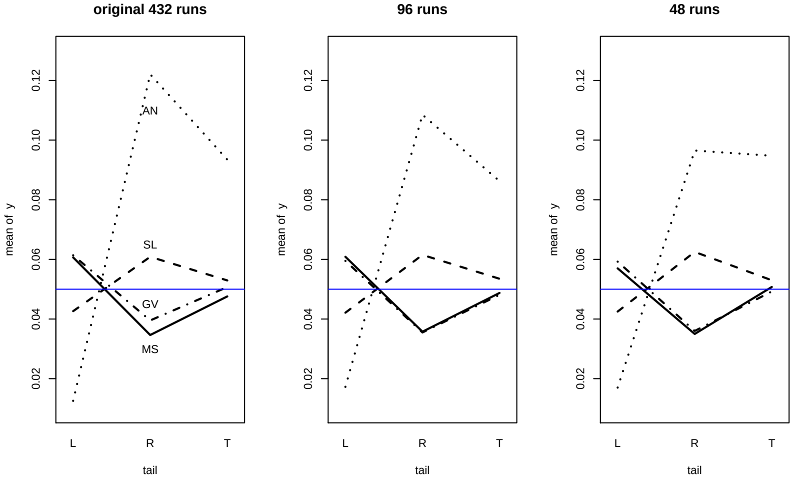
Can we go further? Fractional factorials?

Using JMP, we get a balanced 48-run D-optimal design (half-fraction of 96 runs).

Analysis of small half fraction experiment



Analysis of small half fraction experiment



Analysis of small half fraction experiment

Conclusions:

- Fewer terms are statistically significant (now using 32 df for effects out of 48 runs)
- Similar conclusions on effect sizes
- 3fi's no longer feasible.

General remarks, I

- Statistical analysis of results a clear win.
- Reduce experimental effort through fewer levels.
- Fractional factorials possible, but scope limited (this and other studies limited to 5 or fewer experimental variables).
- Open source tools for mixed-level factorial designs aren't readily available.
- Even if we're interested in exploring nonlinearities and higher-order interactions, smaller designs are a good place to start.
  - Screen out irrelevant factors, then study important factors in greater detail.

General remarks, II

- Isn't this a computer experiment?
  - Simulation of pseudo-random samples makes for a non-deterministic process.
  - Do we care about exploring numeric variables on a continuous scale?
- More complex experimental designs...
  - What if we apply each method (here, 4 hypothesis tests) to the same simulated data?
    $\Rightarrow$ Split-plot experiment.
- Sensible choice of factor levels is important.
  - Is $n = 50$ sufficiently large for large sample asymptotics?
  - More generally, a significant factor can seem insignificant if we choose levels badly.
- Replications?
- Transformations?
- Model checking?