Block designs with missing observations: Incorporating MLEs for the missing data in ANOVA models

Laura Lartey

Department of Mathematical Sciences
Montana State University

April 30, 2019

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

Laura Lartey

This writing project has been read by the writing project advisor and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

Date	Prof. John Borkowski Writing Project Advisor
Date	Mark C. Greenwood Writing Projects Coordinator

Contents

Abstract

The problem of missing observations is a common one in experimental designs especially in block designs. In this study, I considered the randomized complete block design with 3 and 4 treatments and 4,6,8 and 10 blocks as well as the latin square design with dimensions 3, 4, 5, 6 and 7. The goal was to study the effect of F-tests on power and Type I error when one experimental observation was missing. Two methods of data analysis were considered: the exact analysis and the analysis using a Maximum Likelihood Estimator (MLE) for the missing observation. Simulation studies were performed to compare these methods to each other and to the ANOVA with complete data. For both block designs, the power of the analysis done with the MLE was found to be higher than that of the exact analysis. The Type I error of the analysis with the MLE was, however, higher than that of the Exact analysis when the null hypothesis was true in both block designs.

Introduction

Missing observations are common occurrences in statistical analyses especially in scientific experiments. Causes of missing observations include coding and data entry errors. Limited time and resources, as well as insufficient logistics, often prevent researchers from adjusting data collection methods to get a complete data set. Most researchers just drop these observations and carry out their analyses with the incomplete data sets. This is referred to as an exact analysis

Though commonly accepted, there appear to be many problems associated with using incomplete data sets in scientific experiments. Kang (2013) stated that one primary issue associated with missing observations is the reduction in the statistical power of a test. That is, we may have biased parameter estimates, reduced generalizability of the data and a complicated statistical analysis.

A significant amount of work has been done in this area and some suggested solutions include analysis with incomplete data sets, analyses that replace the missing observations with the mean of the complete cases, multiple imputation methods and the use of maximum likelihood estimates for missing data (2013).

Statement of Problem

In this study, the focus is on two block designs with missing observations. Specifically, the latin square designs and randomized complete block designs will be studied. Estimates of the missing observations that minimize the ANOVA sums of squares errors of the three designs stated above will be derived.

Comparison of two analysis methods

Through simulation, the power and type 1 error between ANOVA models incorporating an estimate of the MLEs for the missing data and the exact analysis using unbalanced ANOVA with the missing observations are compared. It is my expectation that the ANOVA models incorporating MLEs for the missing data would have higher power.

Randomized Complete Block Designs

Definition and Notation

Randomized Complete Block Designs (RCBD) is one of the common block design methods used in statistics (Montgomery 2017). This design is called a complete block design since each block contains all the treatment levels considered for the treatment factor. Blocking is done to reduce random error. Typical blocking factors are usually the observational units in the study but some other common ones include time, people, batches of material and machinery. Treatments in a block are randomly ordered such that the treatment order differs for each block. Though we are interested in the treatments, we can also treat blocks as fixed or random factors. In either case, the F-test for treatment effects is same. That is, the F-statistic is MS_{Trt}/MS_E. In this study, a will denote the number of treatments and b, the number of blocks. (Montgomery 2017). Table 1 is an example of an RCBD with treatment levels A, B, C, D and E.

Blocks	Treatments				
1	A	B	C	D	E
2	B	C	D	E	A
3	C	D	E	A	B
4	D	E	A	B	C

Table 1: RCBD with 4 blocks and 5 treatments

The common statistical model is the additive effects model:

$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij}$$

where

y_{ij} is the response for an observation in block j and under treatment i

μ is the overall mean

τ_i is the effect of treatment i

β_j is the effect of block j

$\epsilon_{ij} \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$ is the identical and independent random error

The partition of the total sums of squares for the RCBD is

$$\Sigma_{i=1}^{a}\Sigma_{j=1}^{b}(y_{ij} - \bar{y}_{..})^2 = b\Sigma_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 + a\Sigma_{j=1}^{b}(\bar{y}_{.j} - \bar{y}_{..})^2 + \Sigma_{i=1}^{a}\Sigma_{j=1}^{b}(\bar{y}_{ij} - \bar{y}_{.j} - \bar{y_{i}}. + \bar{y}_{..})^2$$

$$SS_T = SS_{Treatment} + SS_{Blocks} + SS_{Error}$$

where

$\bar{y}_{i.}$ is the mean of the observations that are taken under treatment i

$\bar{y}_{.j}$ is the mean of the observations in block j

$\bar{y}_{..}$ is the mean of all observations

Table 2 is the ANOVA table for a Randomized Complete Block Design.

Source	Df	SS	MS	F
Blocks	b-1	SS_{Blocks}	MS_{Blocks}	
Treatment	a-1	$SS_{Treatment}$	$MS_{Treatment}$	$F = \frac{MS_{Treatment}}{MSE}$
Error	$(a$-1$)(b$-1$)$	SS_{Error}	MS_{Error}	
Total	ab-1	SS_T		

Table 2: ANOVA table for RCBD

Derivation of missing value estimate

Even though treatments are supposed to be orthogonal to blocks, the absence of an observation can lead to nonorthogonality. In such cases, there are two approaches of carrying out an analysis: the exact analysis and the approximate analysis. The exact analysis is carried out using the incomplete data or the data with the missing observation. This is a sequential sums of squares ANOVA with $SS_{Treatment(adj)}$ being the sum of squares for treatments adjusted for blocks being in the model. With the approximate analysis, we would replace the missing observation with an estimate and reduce the error degrees of freedom by 1.

Source	Df	SS	MS	F
Blocks	b-1	SS_{Blocks}	MS_{Blocks}	
Treatment	a-1	$SS_{Treatment(adj)}$	$MS_{Treatment}$	$F = \frac{MS_{Treatment}}{MSE}$
Error	$(a$-1$)(b$-1$)$-1	SS_{Error}	MS_{Error}	
Total	ab-2	SS_T		

Table 3: ANOVA table for RCBD with a missing observation

In this study, the Maximum Likelihood Estimator is used to estimate the missing value. The MLE is the estimate that the minimizes the sums of square error. This requires finding the derivative of the sums of square error, and equating this to zero to solve for the missing observation, X. Table 4 shows an RCBD with a missing observation X for treatment E in block 2

Blocks	Treatments				
1	A	B	C	D	E
2	B	C	D	X	A
3	C	D	E	A	B
4	D	E	A	B	C

Table 4: RCBD with a missing observation

Theorem: Suppose X represents the missing y_{ij} value, then the MLE of x is $x = \frac{y'_{i.}a + y'_{.j}b - y'_{..}}{(a-1)(b-1)}$

Proof:

By definition, $SSE = \Sigma_{i=1}^{a}\Sigma_{j=1}^{b}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$

With missing observation x to be estimated:

$$\begin{aligned} SSE &= \Sigma_{i=1}^{a}\Sigma_{j=1}^{b}y_{ij}^2 - \frac{1}{b}\Sigma_{i=1}^{a}(\Sigma_{j=1}^{b}y_{ij})^2 - \frac{1}{a}\Sigma_{j=1}^{b}(\Sigma_{i=1}^{a}y_{ij})^2 + \frac{1}{ab}(\Sigma_{i=1}^{a}\Sigma_{j=1}^{b}y_{ij})^2 \\ &= x^2 - \frac{1}{b}(y'_{i.} + x)^2 - \frac{1}{a}(y'_{.j} + x)^2 + \frac{1}{ab}(y'_{..} + x)^2 + F(y^*) \end{aligned}$$

where

y^* is the set of y_{ij} values with x removed

$y'_{..} = y_{..}$ - x

$y'_{i\cdot}$ is the sum of responses with x removed from treatment group i

$y'_{\cdot j}$ is the sum of responses with x removed from block j

$F(y^*)$ is a function of y values that do not contain x

Differentiating the SSE with respect to x,

$$\frac{dSSE}{dx} = 2x - \frac{2}{b}(y'_{i\cdot} + x) - \frac{2}{a}(y'_{\cdot j} + x) + \frac{2}{ab}(y'_{\cdot\cdot} + x)$$

To minimize the SSE, we set the derivative to 0

$$\begin{aligned}
0 &= 2x - \frac{2}{b}(y'_{i\cdot} + x) - \frac{2}{a}(y'_{\cdot j} + x) + \frac{2}{ab}(y'_{\cdot\cdot} + x) \\
\rightarrow 0 &= x - \frac{y'_{i\cdot}}{b} - \frac{x}{b} - \frac{y'_{\cdot j}}{a} - \frac{x}{a} + \frac{y'_{\cdot\cdot}}{ab} + \frac{x}{ab} \\
\rightarrow 0 &= abx - y'_{i\cdot}a - xa - by'_{\cdot j} - bx + y'_{\cdot\cdot} + x \\
&= (1 + ab - a - b)x - y'_{i\cdot}a - y'_{\cdot j}b + y'_{\cdot\cdot}
\end{aligned}$$

Finally, solving for x:

$$x = \frac{y'_{i\cdot}a + y'_{\cdot j}b - y'_{\cdot\cdot}}{(a-1)(b-1)}$$

Latin Square Design

Definition and Notation

The Latin Square Design (LSD) has 3 factors: one factor is of interest and the other two considered to be nuisance factors. The LSD blocks on the nuisance factors with the aim of separating their variability from the variability due to random error (Montgomery 2017). All factors have the same number of levels, p and form a *pxp* array with one of the blocking factors on the rows and the other, on the columns. These rows and columns are considered to be restrictions on randomization (Montgomery 2017). The *pxp* array is orthogonal and transposing the square or rearranging the rows and columns still gives an orthogonal array. The variable of interest is the treatment and it is randomized such that it occurs once in every row and column. Treatment levels are typically labelled with roman letters (Sirikasemsuk and Leerojanaprapa 2017). Latin squares can be in two forms: reduced and non-reduced. The reduced design has the treatments labelled in alphabetical order down the first row and the first column and then randomized in the other cells. The non-reduced design however has the treatments randomized in every row and column (Fisher and Yates 1934).

The statistical model is an additive model and is written as

$$y_{ijk} = \mu + \alpha_i + \tau_j + \beta_k + \epsilon_{ijk}$$

for

$$\begin{cases} i = 1, 2, 3, ..., p \\ j = 1, 2, 3, ..., p \\ k = 1, 2, 3, ..., p \end{cases} \tag{1}$$

where

y_{ijk} is the response for the observation in the ith row, kth column and and jth treatment

μ is the overall mean

α_i is the effect of the ith row

τ_j is the effect of treatment j

β_k is the effect of the kth column

$\epsilon_{ij} \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$ is the identical and independent random error

The sums of squares for the LSD is partitioned as

$$SSE = SS_T - SS_{Rows} - SS_{Columns} - SS_{Treatment}$$

$$\begin{aligned} SSE &= \Sigma_i \Sigma_j \Sigma_k y_{ijk}^2 - \frac{y_{...}^2}{N} - \frac{1}{p}\Sigma_{j=1}^p y_{.j.}^2 + \frac{y_{...}^2}{N} - \frac{1}{p}\Sigma_{i=1}^p y_{i..}^2 + \frac{y_{...}^2}{N} - \frac{1}{p}\Sigma_{k=1}^p y_{..k.}^2 + \frac{y_{...}^2}{N} \\ &= \Sigma_i \Sigma_j \Sigma_k y_{ijk}^2 - \frac{1}{p}\Sigma_{j=1}^p y_{.j.}^2 - \frac{1}{p}\Sigma_{i=1}^p y_{i..}^2 - \frac{1}{p}\Sigma_{k=1}^p y_{..k.}^2 + 2\frac{y_{...}^2}{N} \end{aligned}$$

where

$y_{i..}$ is the sum of all responses for row i

$y_{.j.}$ is the sum of all responses for treatment j

$y_{..k}$ is the sum of all responses for column k

$y_{...}$ is the sum of all responses

$N = p^2$ is the total number of observations

Table 5 is the ANOVA table for a Latin Square Design

Source	Df	SS	MS	F
Treatment	p-1	$SS_{Treatment}$	$MS_{Treatment}$	$F = \frac{MS_{Treatment}}{MSE}$
Rows	p-1	SS_{Rows}	MS_{Rows}	
Columns	p-1	$SS_{Columns}$	$MS_{Columns}$	
Error	(p-2)(p-1)	SS_{Error}	MS_{Error}	
Total	p^2-1	SS_T		

Table 5: ANOVA table for LSD

Derivation of missing value estimate

	Treatments			
Blocks	1	2	3	4
1	A	B	C	D
2	B	X	D	A
3	C	D	A	B
4	D	A	B	C

Table 6: LSD with one missing observation, X, for treatment C in block 2.

Missing observations in LSD can lead to nonorthoganility in the array. This is because there would be an unequal number of replicates for treatments in the rows and columns. Approaches to tackling the issue of missing observations are very similar to those used in the RCBD. The exact analysis is carried out using the incomplete data or the data with the missing observation while for the approximate analysis, the missing observation is replaced with the Maximum Likelihood Estimator (MLE). Both methods reduce the error degrees of freedom by 1. (Yates 1933). Table 7 shows the ANOVA table for an LSD with one missing observation.

Source	Df	SS	MS	F
Rows	p-1	SS_{Rows}	MS_{Rows}	
Columns	p-1	$SS_{Columns}$	$MS_{Columns}$	
Treatment(adj)	p-1	$SS_{Treatment}$	$MS_{Treatment}$	$F = \frac{MS_{Treatment}}{MSE}$
Error	(p-2)(p-1)-1	SS_{Error}	MS_{Error}	
Total	p^2-2	SS_T		

Table 7: ANOVA table for LSD with one missing observation

Theorem: Suppose X represents the missing y_{ij} value, then the MLE of X is $x = \frac{p(y'_{.j.}+y'_{i..}+y'_{..k})-2y'_{...}}{p^2-3p+2}$

Proof:

By definition,

$$\begin{aligned} SSE &= \Sigma_i\Sigma_j\Sigma_k y_{ijk}^2 - \frac{y_{...}^2}{N} - \frac{1}{p}\Sigma_{j=1}^p y_{.j.}^2 + \frac{y_{...}^2}{N} - \frac{1}{p}\Sigma_{i=1}^p y_{i..}^2 + \frac{y_{...}^2}{N} - \frac{1}{p}\Sigma_{k=1}^p y_{..k.}^2 + \frac{y_{...}^2}{N} \\ &= \Sigma_i\Sigma_j\Sigma_k y_{ijk}^2 - \frac{1}{p}\Sigma_{j=1}^p y_{.j.}^2 - \frac{1}{p}\Sigma_{i=1}^p y_{i..}^2 - \frac{1}{p}\Sigma_{k=1}^p y_{..k.}^2 + 2\frac{y_{...}^2}{N}. \end{aligned}$$

With missing observation x to be estimated, the SSE is

$$SSE = \Sigma_i \Sigma_j \Sigma_k y_{ijk}^2 - \frac{1}{p}\Sigma_{j=1}^{p} y_{.j.}^2 - \frac{1}{p}\Sigma_{i=1}^{p} y_{i..}^2 - \frac{1}{p}\Sigma_{k=1}^{p} y_{..k.}^2 + \frac{2y_{...}^2}{p^2}$$

Let y' represent the sum of observations without the missing observation x. Then

$$SSE = x^2 - \frac{1}{p}(y'_{.j.} + x)^2 - \frac{1}{p}(y'_{i..} + x)^2 - \frac{1}{p}(y'_{..k.} + x)^2 + \frac{2(y'_{...} + x)}{p^2} + F(y*)$$

where

$N = p^2$ is the total number of observations

$F(y*)$ is a function of y_{ijk} values that do not contain x

$y_{i..'}$ is the sum of all responses with x removed from row i

$y_{.j.'}$ is the sum of all responses with x removed from treatment j

$y_{..k'}$ is the sum of all responses with x removed from column k

$y_{...'}$ is the sum of all responses of all observations with x removed

Differentiating the SSE with respect to x,

$$\frac{dSSE}{dx} = 2x - \frac{2}{p}(y'_{.j.} + x) - \frac{2}{p}(y'_{i..} + x) - \frac{2}{p}(y'_{..k} + x) + \frac{4(y'_{...} + x)^2}{p^2}$$

To minimize the SSE, we set the derivative to 0

$$\begin{aligned} 0 &= x - \frac{(y'_{.j.} + x)}{p} - \frac{(y'_{i..} + x)}{p} - \frac{(y'_{..k} + x)}{p} + \frac{2(y'_{...} + x)}{p^2} \\ \rightarrow 0 &= p^2 x - p(y'_{.j.} + x) - p(y'_{i..} + x) - p(y'_{..k} + x) + 2(y'_{...} + x) \\ &= x(p^2 - 3p + 2) - p(y'_{.j.} + y'_{i..} + y'_{..k}) + 2y'_{...} \end{aligned}$$

Finally, solving for x:

$$x = \frac{p(y'_{.j.} + y'_{i..} + y'_{..k}) - 2y'_{...}}{p^2 - 3p + 2}$$

Simulation

An additive effects model for each of the designs is considered. A random sample of response values were generated for each block and treatment combination. An ANOVA was carried out with the complete data set and on the data set after one random observation was deleted. Both the exact and approximate analysis were performed and this process was repeated 50,000 times. The estimated proportion of times that the null hypothesis is rejected was reported.

Results

Randomized Complete Block Designs

When the null hypothesis was true, both the exact and complete analysis attained the 0.01, 0.05 and 0.1 significance levels for all block and treatment combinations. The Type I error for the approximate analysis also attained the 0.01 significance level in all block-treatment combinations but was slightly higher than that of the exact and complete analyses. The largest difference between the Type I error of the approximate analysis and that of the exact and complete analyses was about 0.025, and this occurred when there were 4 blocks and 3 treatments. With an increasing number of blocks or treatments, the Type I error of the MLE decreased gradually.

In the presence of treatment effects, the approximate analysis and the analysis done on the complete data set attained very similar power values whiles the exact analysis attained comparatively lower power levels. With an increasing number of treatments or blocks, the power from all 3 analyses increased. The difference between the power for the exact analysis and power of either the MLE analysis or the ANOVA with the complete data set also decrease with increasing number of blocks and treatments.

Latin Square Design

When the null hypothesis was true, both the exact analysis and the ANOVA on the complete dataset attained the nominal α levels. The Type I error for the MLE analysis was higher than that of the analyses on the complete dataset and exact analysis. The difference, however, decreased with increasing dimensions for the latin square design.

When there were 3 treatments, the analysis done with the complete dataset had a higher power than the exact and approximate analyses. The power from the MLE analysis was in turn greater than that of the exact analysis. The power the analysis with the MLE was greater than that of the complete analysis when the dimensions were 4,5,6 or 7 and when the treatment effects were not very large. With increasing dimensions, the power from all 3 analyses increased and they all appeared to attain very similar power values when there were 7 treatments.

Conclusion

In a RCBD or LSD, when the null hypothesis is true, the MLE has a higher Type I error than the exact and complete analyses. However, when there are treatment effects present, the power of the analysis done on a complete dataset and the analyses using an MLE are very similar and are both greater than the power of the exact analysis.

With a high number of blocks and treatment levels, the power of all three analyses attain very similar power values, and the power of the test increases at a decreasing rate.

Appendix

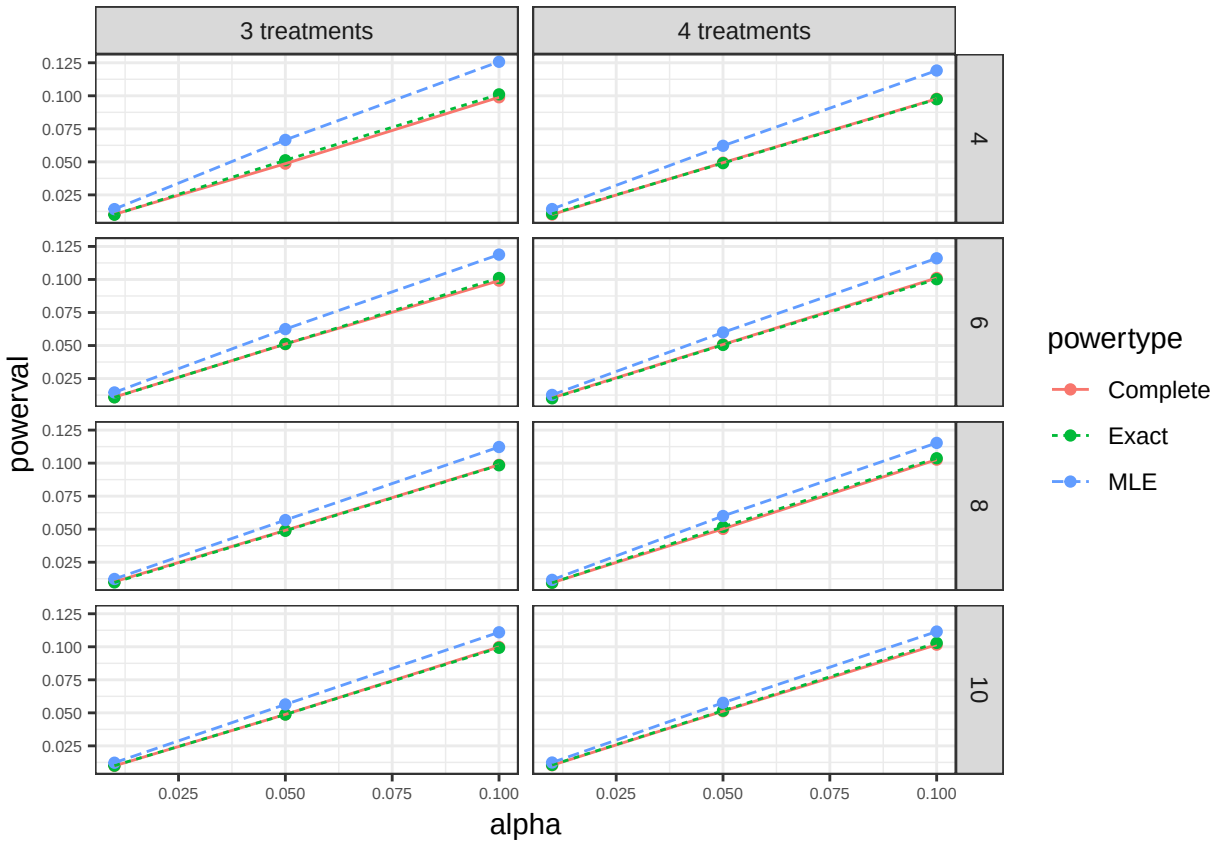

Figure 1: Plot for the Type I error for RCBD when the null hypothesis is true

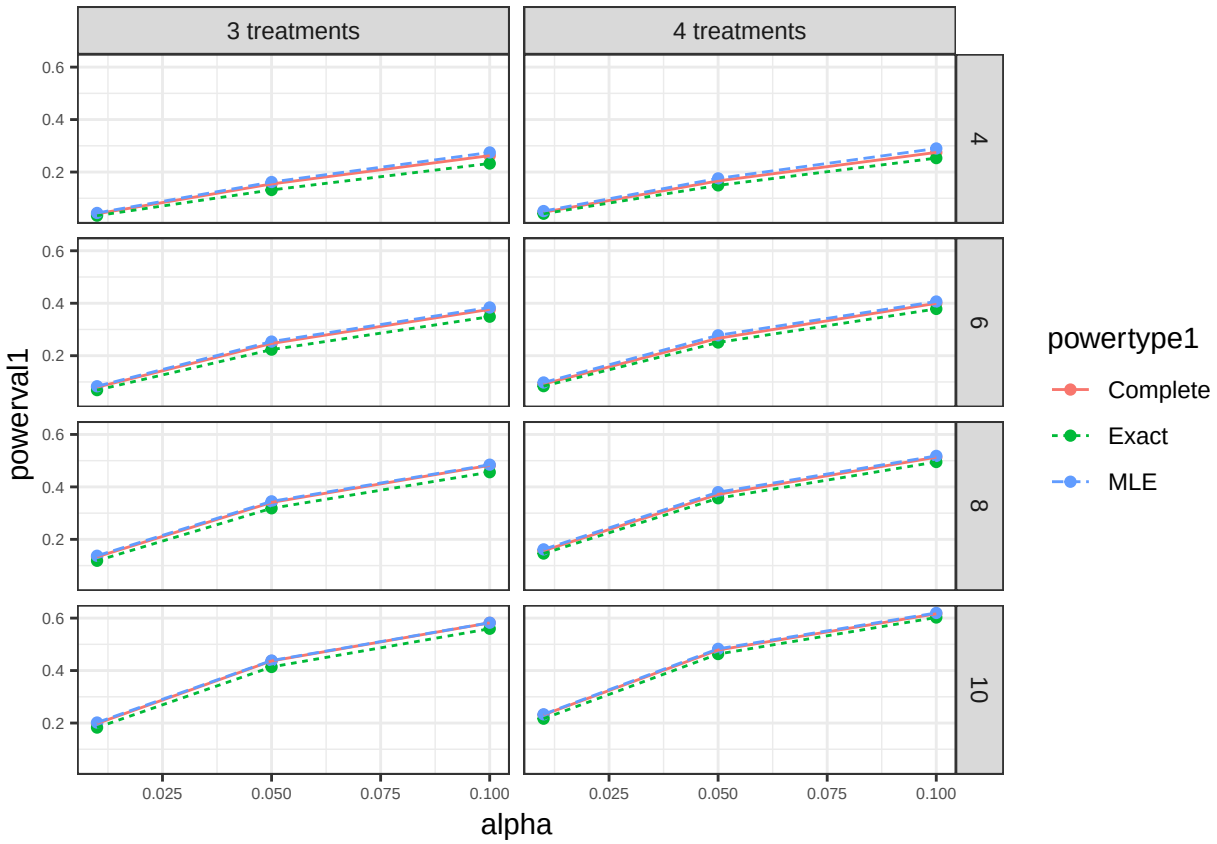

Figure 2: Plot for the Power for RCBD when with treatment effects (τ_1, τ_2, τ_3)=(-0.5,0,0.5) and block effects $(\beta_1, \beta_2, \beta_3, \beta_4)$=(-0.5,0.5,-0.25,0.25)

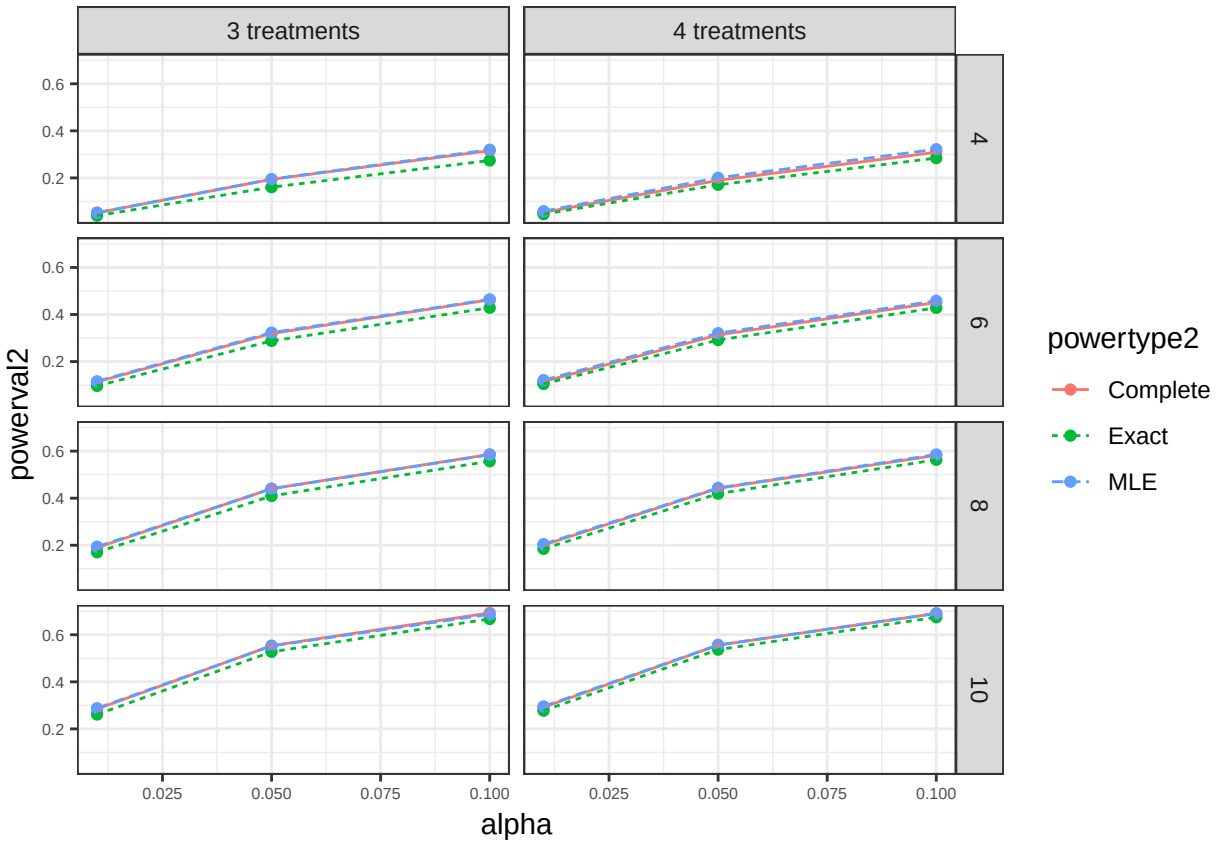

Figure 3: Plot for the Power for RCBD when with treatment effects (τ_1, τ_2, τ_3)=(1,0,0) and block effects $(\beta_1, \beta_2, \beta_3, \beta_4)$=(1,0,0,0)

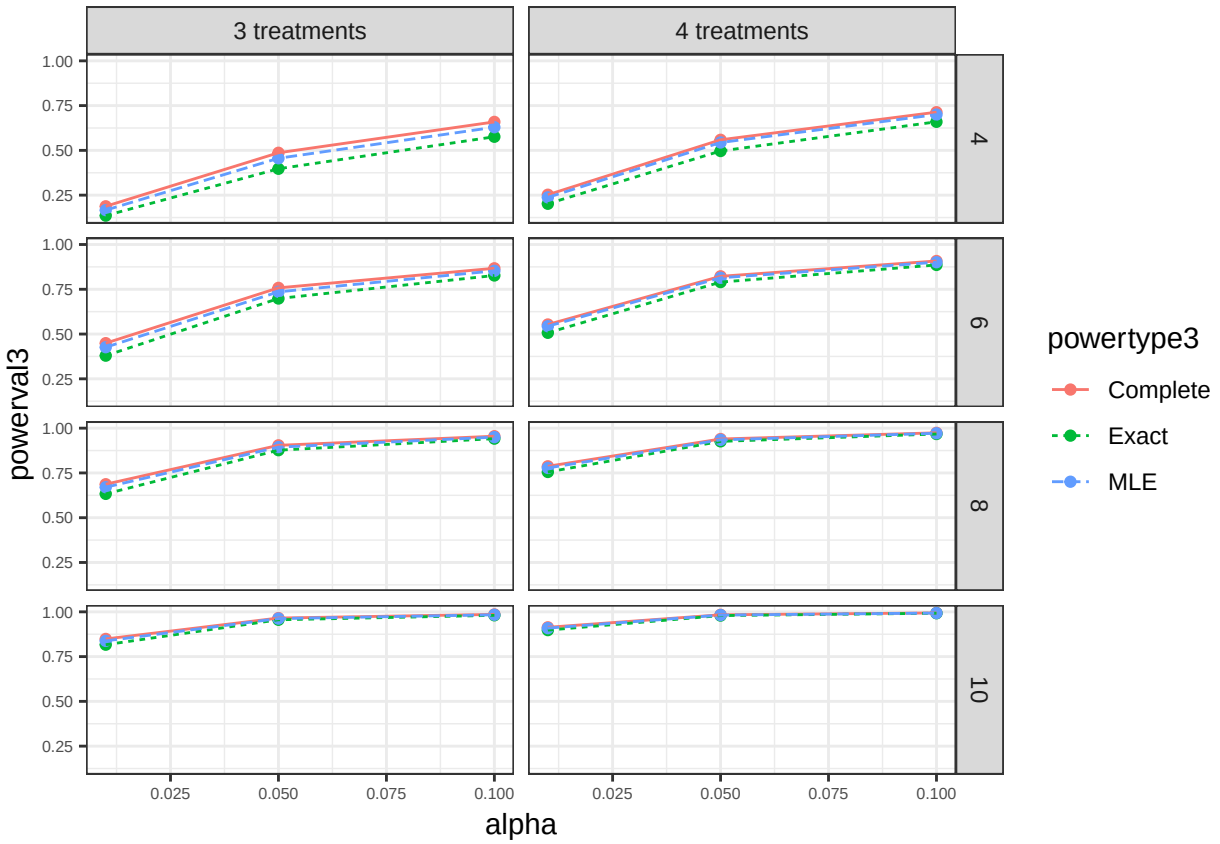

Figure 4: Plot for the Power for RCBD when with treatment effects (τ_1, τ_2, τ_3)=(-1,0,1) and block effects $(\beta_1, \beta_2, \beta_3, \beta_4)$=(-1,1,-0,5,0.5)

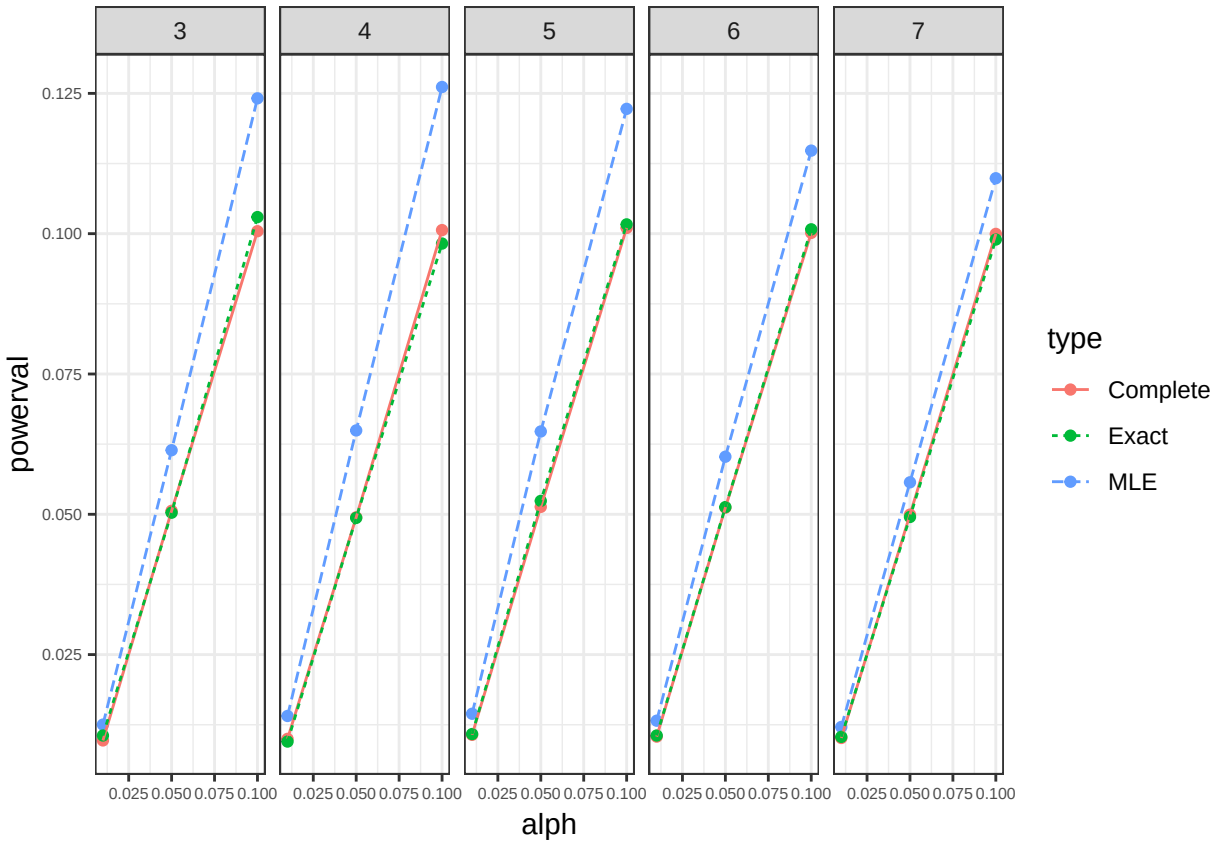

Figure 5: Plot for the Power for the LSD when the null is true

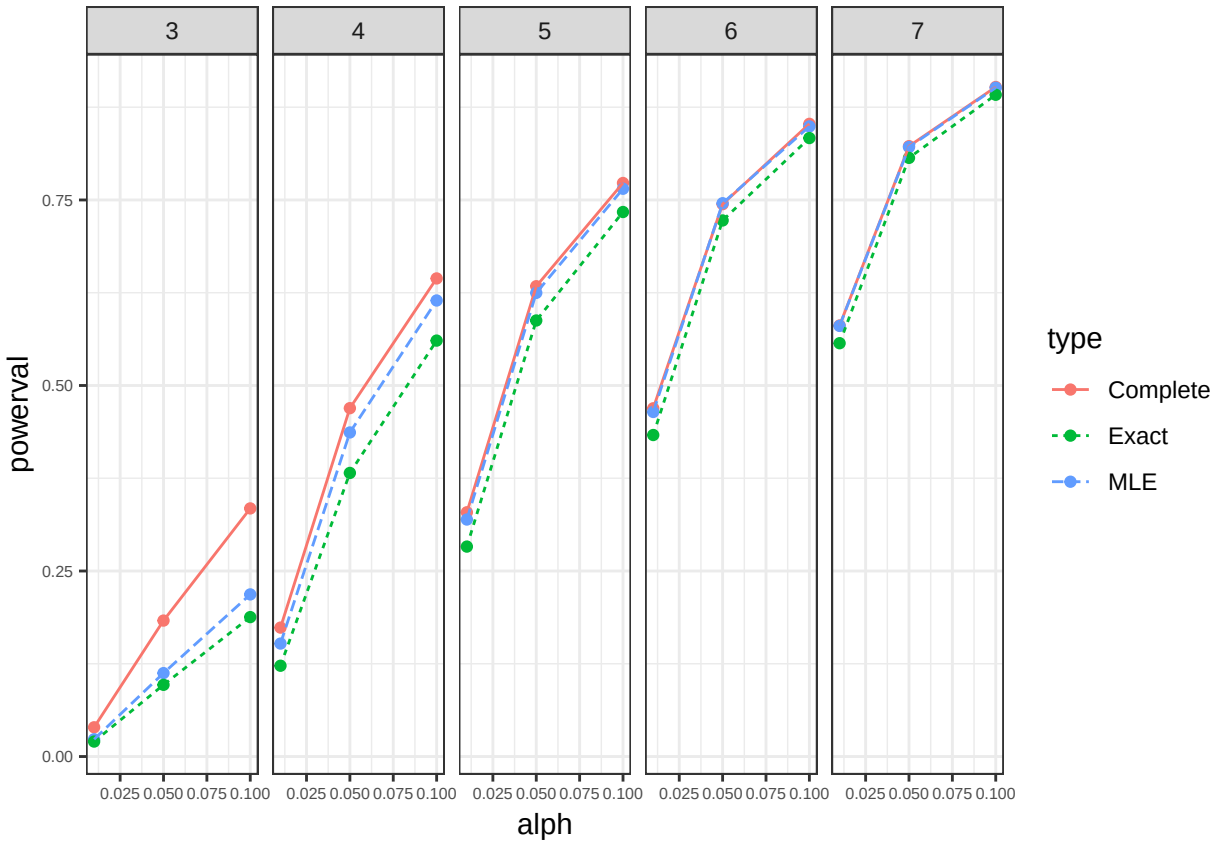

Figure 6: Plot for the Power for LSD when with treatment effects (τ_1, τ_2, τ_3)=(-1,0,1), $(\tau_1, \tau_2, \tau_3, \tau_4)$=(-1,1,-0.5,0.5),$(\tau_1, \tau_2, \tau_3, \tau_4, \tau_5)$=(-1,1,0,-0.5,0.5),$(\tau_1, \tau_2, \tau_3, \tau_4, \tau_5, \tau_6)$=(-1,0,1,0,-0.5,0.5),$(\tau_1, \tau_2, \tau_3, \tau_4, \tau_5, \tau_6, \tau_7)$=(-1,0,0,1,0,-0.5,0.5)

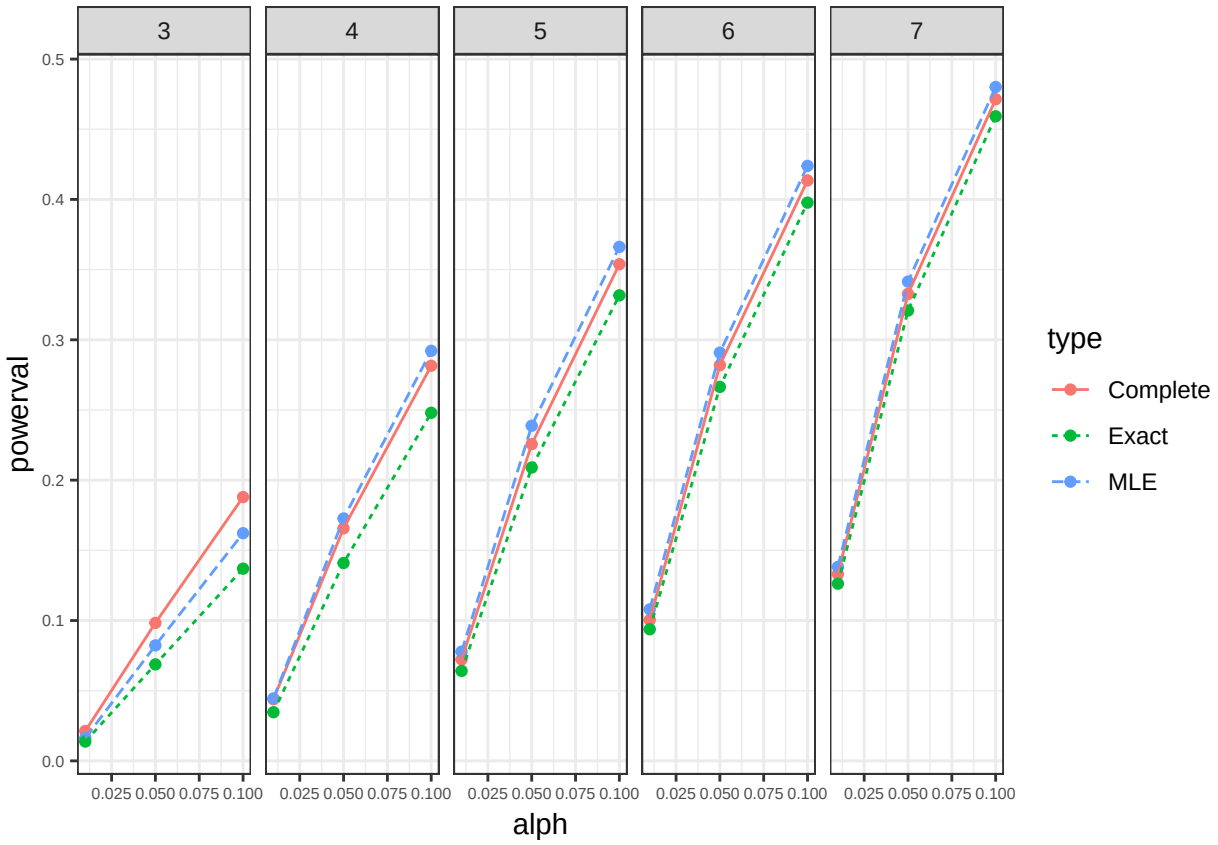

Figure 7: Plot for the Power for LSD when with treatment effects: (τ_1, τ_2, τ_3)=(1,0,0), $(\tau_1, \tau_2, \tau_3, \tau_4)$=(1,0,0,0),$(\tau_1, \tau_2, \tau_3, \tau_4, \tau_5)$=(1,0,0,0,0),$(\tau_1, \tau_2, \tau_3, \tau_4, \tau_5, \tau_6)$=(1,0,0,0,0,0),$(\tau_1, \tau_2, \tau_3, \tau_4, \tau_5, \tau_6, \tau_7)$=(1,0,0,0,0,0,0)

References

Cochran, W. G, and G. M Cox. 1957. "Experimental Designs." John Wiley & Sons.

Davies, O. L. 1954. *The Design and Analysis of Industrial Experiments.* Oliver & Boyd, London.

Fisher, R. A., and F. Yates. 1934. "The 6× 6 Latin Squares." In *Mathematical Proceedings of the Cambridge Philosophical Society*, 30:492–507. 4. Cambridge University Press.

Kang, H. 2013. "The Prevention and Handling of the Missing Data." *Korean Journal of Anesthesiology* 64 (5): 402–6.

Kempthorne, O. n.d. "The Design and Analysis of Experiments, 1952." New York, John Wiley & Sons, Inc.

Montgomery, D. C. 2017. *Design and Analysis of Experiments.* John Wiley & Sons.

Sirikasemsuk, K., and K. Leerojanaprapa. 2017. "One Missing Value Problem in Latin Square Design of Any Order: Exact Analysis of Variance." *Cogent Engineering* 4 (1). Taylor & Francis: 1411222.

Yates, F. 1933. "The Analysis of Replicated Experiments When the Field Results Are Incomplete." *Empire Journal of Experimental Agriculture* 1 (2): 129–42.