Common Errors in Marketing Experiments and How to Avoid Them

Tanya Kolosova
Associates In Analytics Inc.

Samuel Berestizhevsky
Innovator and Actionable Analytics Expert

Classifications, Key Words:

- Design of experiments
- Marketing experiments
- Split-unit design
- Split-plot design
- Bias correction
- Block variables
- SAS

Abstract

A methodology for designing experiments developed by Sir Ronald Fisher is more than 80 years old, but many marketers still rely on simple A/B tests to compare the performance of marketing campaigns and to find conditions to achieve the best results. Because marketing efficiency depends on a combination of factors and not on factors acting independently, A/B tests are not only inefficient but are actually not suitable for conducting marketing experiments.

In this article, we describe the very useful and efficient split-unit (or split-plot) design of marketing experiments. Split-unit design is often used in marketing experiments but is not recognised; often missed or inappropriately analysed. This, in turn, produces misleading results that may be very costly in marketing. We use a real-life example to demonstrate some of the ideas involved and ways to correctly analyse split-unit design.

1. Introduction

A very common but inefficient approach to studying the effects of multiple factors is to carry out successive experiments in which the levels of each factor are changed one at a time (A/B testing). Sir Ronald Fisher showed that a better approach is to vary the factors simultaneously and to study response at each possible factor-level combination. A methodology of design of experiments (DoE) was developed by Fisher in his ground-breaking book “The Design of Experiments” in 1935. For his contribution in statistics, Fisher has been described as "a genius who almost single-handedly created the foundations for modern statistical science" (Hald, 1998) and "the single most important figure in 20th-century statistics" (Efron, 1998). Since then, DoE methodology has been broadly adopted in agricultural engineering, physical and social sciences, advertising and marketing.

Surprisingly, many marketers still rely on simple A/B tests to compare the performance of marketing campaigns and to find conditions to achieve the best results. There are multiple reasons to replace A/B tests by design of experiments:

a) In DoE, the approach is completely different from A/B testing, as all parameters (factors) are changed together, simultaneously, and not one parameter at a time. Thus, in DoE the required number of experiments is limited and significantly smaller than with A/B testing.

b) DoE provides a way to account for different sources of errors and compares averages to other averages rather than individual values to other individual values (as A/B testing). This achieves much greater accuracy in the estimation of effective factors for a given number of experiments, and thus the influential factors and their combinations are much more likely to emerge from the noise of the experimental errors.

c) But what is more critical, DoE allows for estimating of the impact of factor interactions which is not available in A/B testing. In fact, because marketing efficiency depends on a combination of factors and not on factors acting independently, A/B tests are not really suitable for conducting marketing experiments.

DoE methodology creates a framework for planning, analysing and executing marketing experiments. There are 3 main principles of DoE: randomisation, replication, and blocking. Randomisation is a deliberate process to eliminate potential biases from the conclusions through random assignment of "treatments". Replication is, in some sense, the heart of all of statistics. Replication is the basic issue behind every method. We always want to estimate or control the uncertainty in our results. We achieve this estimate through replication; and blocking is a technique to include other factors in our experiment which contribute to undesirable variation. We want the unknown error variance at the end of the experiment to be as small as possible. Our goal is usually to find out something about treatment factors (or factors of primary interest), but in addition to this, we want to include any blocking factors that will explain variation.

One of the most efficient and frequently used designs is split-unit (also referred as split-plot) design: when one experimental unit is split into subunits, to which subsequent treatments are applied. Marketing usually involves a number of sequential steps, which makes split-unit design not only feasible and desirable but actually necessary.

The challenge is that split-unit experiments are often used but can be difficult to recognise. As a result, split-unit experiments are often inappropriately analysed. A spreadsheet of data can look like a variety of multifactor experiments, and it is very tempting to consider the experiment as completely randomised design (CRD) and then to apply straightforward analysis. In split-unit designs of experiments, it can take some research work to find out what factors (if any) are blocking factors and which are treatment factors, and (most importantly) what were the experimental units (EU) to which treatment factors were applied.

As with any statistical method, to receive correct results the method should be correctly applied. In the case of complex split-unit design, miss-interpretation of EU and incorrect error structure lead to inappropriate analysis, which produces misleading results that may be very costly in marketing.

Description of the Marketing Experiment

As an example, let's consider the real-life case in which office supply retailing Company A needs to test the impact of marketing emails to find the optimal combination of factors-levels and achieve maximum sales as a response to marketing emails. To quantify the success of the marketing experiment, Company A uses total sales generated by the customers who participated in the marketing campaign.

There are multiple factors which affect the success of email marketing. For example, factors that describe the marketing message, format of the message, type of customers that receive the messages, etc.

In our real-life marketing experiment, the following 4 factors were included:

Factor Name	Levels	Factor Description
customer	C1 C2 C3	The Company A differentiates their customers into 3 types according to customers purchasing behaviour.
minimal_order	$50 $100	To become eligible for the discount, a customer has to make an order for a specific dollar value (at least).
discount	5% 10% 15%	If eligible according to the order dollar value, the customer receives a discount on the whole order.
subject_line	SL1 SL2	2 versions of email subject lines are developed by the marketers for the marketing experiment

First, lists of the 3 different types of customers were created. These lists were created by a random selection from the repository of the company's customers without replacement, which ensured that each selected customer appeared only once. Customers were selected according to customer types, producing 600,000 email recipients in each list. 4 replications of each type of customer were obtained, 12 Lists with 7.2million recipients in total.

Then, each List was randomly divided into 6 Batches of 100,000 recipients, and these Batches were randomly assigned combinations of the minimal order value and discount: ($50, 5%), ($50, 10%), ($50, 15%), ($100, 5%), ($100, 10%), ($100, 15%).

Next, each Batch was randomly divided into 2 Groups of 50,000 recipients each. Each Group was randomly assigned SL1 or SL2 version of email subject line.

The table below (the experimental table) presents the full factorial experiment 2^2 3^2 (36 treatment combinations) where each experiment cell contains 50,000 email recipients. The 4 replications of this experiment were conducted with an interval of 3 days.

Exp. run	customer	minimal_ order	discount	subject_ line
1	C1	$50	5%	SL1
2	C1	$50	5%	SL2
3	C1	$50	10%	SL1
4	C1	$50	10%	SL2
5	C1	$50	15%	SL1
6	C1	$50	15%	SL2
7	C1	$100	5%	SL1
8	C1	$100	5%	SL2
9	C1	$100	10%	SL1
10	C1	$100	10%	SL2
11	C1	$100	15%	SL1
12	C1	$100	15%	SL2
13	C2	$50	5%	SL1
14	C2	$50	5%	SL2
15	C2	$50	10%	SL1
16	C2	$50	10%	SL2
17	C2	$50	15%	SL1
18	C2	$50	15%	SL2

Exp. run	customer	minimal_ order	discount	subject_ line
19	C2	$100	5%	SL1
20	C2	$100	5%	SL2
21	C2	$100	10%	SL1
22	C2	$100	10%	SL2
23	C2	$100	15%	SL1
24	C2	$100	15%	SL2
25	C3	$50	5%	SL1
26	C3	$50	5%	SL2
27	C3	$50	10%	SL1
28	C3	$50	10%	SL2
29	C3	$50	15%	SL1
30	C3	$50	15%	SL2
31	C3	$100	5%	SL1
32	C3	$100	5%	SL2
33	C3	$100	10%	SL1
34	C3	$100	10%	SL2
35	C3	$100	15%	SL1
36	C3	$100	15%	SL2

How Analysis Was Performed

This experiment was considered by Company A as a Completely Randomised Design (CRD) and analysed as such. The randomisation structure of the CRD implies that there is only one error term (the within error) and all factors effects are tested against it. The analysis was performed using a user-written computer program that utilises SAS® Software PROC MIXED (see SAS code with explanations in Appendix 1). The results are presented in the table below:

Effect	Numerator DF	Denominator DF	F Stat	P-value
customer	2	108	208.37	<.0001
minimal_order	1	108	0.57	0.4525
customer*minimal_order	2	108	2.08	0.1304
discount	2	108	10.65	<.0001
customer*discount	4	108	5.22	0.0007
minimal_order*discount	2	108	0.00	0.9956
customer*minimal_order*discount	4	108	1.29	0.2784
subject_line	1	108	1.61	0.2072
customer*subject_line	2	108	2.70	0.0717
minimal_order*subject_line	1	108	9.89	0.0021
customer*minimal_order*subject_line	2	108	0.69	0.5059
discount*subject_line	2	108	3.42	0.0364
customer*discount*subject_line	4	108	3.11	0.0183
minimal_order*discount*subject_line	2	108	2.53	0.0843
customer*minimal_order*discount*subject_line	4	108	2.17	0.0767

This table contains hypothesis tests for the significance of each of the fixed effects listed in the column “Effect”. The following factors and their interactions were identified as significant (on 95% confidence level): customer, discount, customer*discount, minimal_order*subject_line, discount*subject_line, and customer*discount*subject_line.
Using significant factors, we built a regression model and found conditions (factors and their levels) that maximised response (sales). See SAS PROC MIXED code in Appendix 2. For each customer type (C1, C2, and C3), the conditions (factor-level combinations) that would generate maximum sales are presented in the table below:

customer	minimal_ order	discount	subject_ line	predicted sales
C1	$50	15%	SL1	$130,681
C2	$50	10%	SL1	$168,058
C3	$100	15%	SL2	$179,607

These results mean that if the email marketing campaign with the factors and levels presented in the above table is deployed for 50,000 customers of each type, then the Company A should expect, on average, the sales amount presented in “Predicted Sales” column.

How Analysis Should Be Performed

We suggest a closer look at how the experiment was executed to understand if the analysis of the experiment was performed correctly.
First, customers were randomly selected by customer types, producing 12 Lists: 4 replications of each of 3 types of customers. This created a completely randomised design. Each List was an experimental unit (EU) for different types of customers (3 levels) – the entity to which types of customers are randomly assigned (see **Figure 1**).

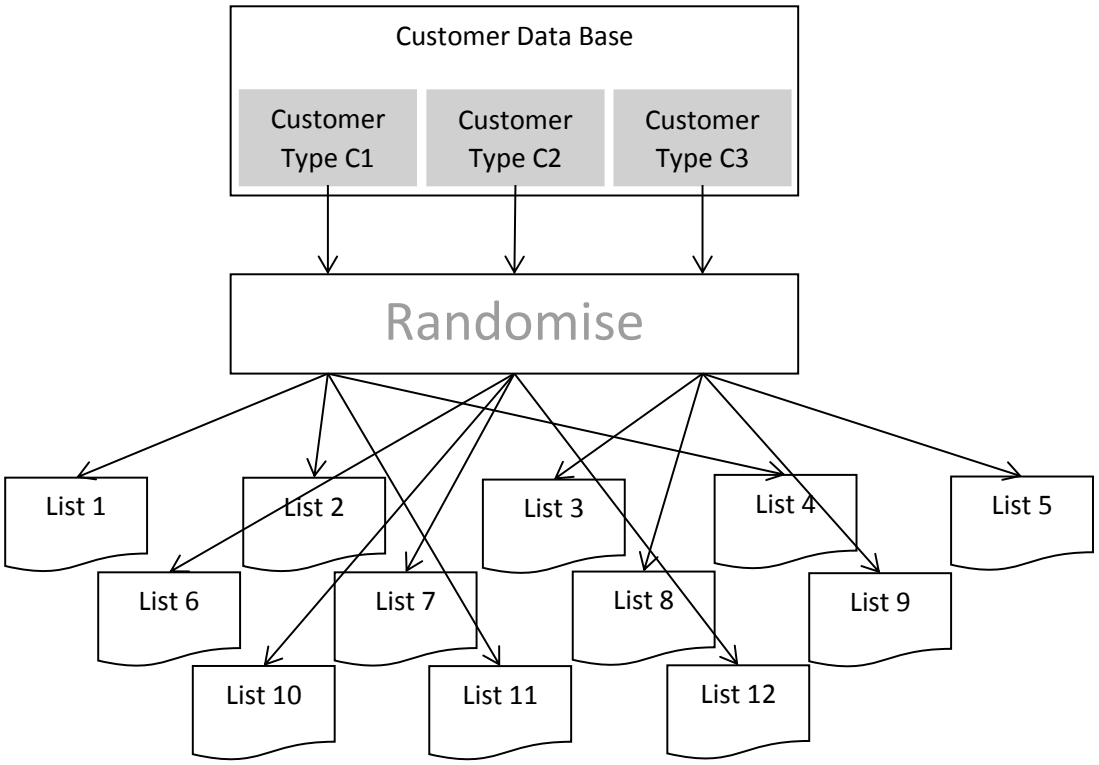

Figure 1. Lists Randomisation

Then, each List was randomly divided into 6 Batches. The act of grouping the experimental units together into homogenous groups is called blocking. Thus, the List was a block of 6 Batches, and the Batch was an experimental unit for combinations of the minimal_order and discount. In other words, the Batch design is a randomised complete block design, where the List is the blocking factor (see **Figure 2**).

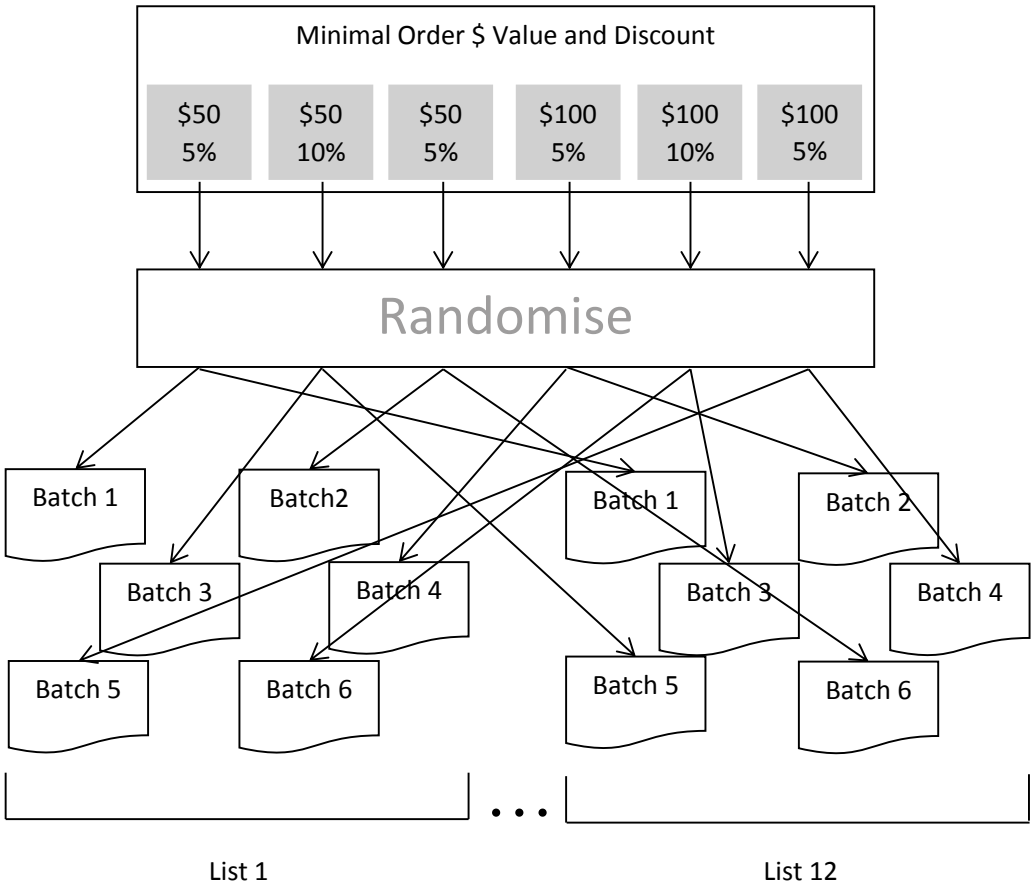

Figure 2. Batch Randomisation

And when each Batch was randomly divided into 2 Groups for 2 versions of email subject lines, Batch*List was a block for levels of email subject lines (see **Figure 3**).

Thus, the appropriate model includes:

- Factorial effects for levels of customer * minimal_order * discount * subject_line,
- and 3 experimental units: List, Batch, Group.

Using split-unit error structure, we analysed the results of the same experiment. SAS PROC MIXED code is presented in Appendix 3. Results of the analysis are presented below:

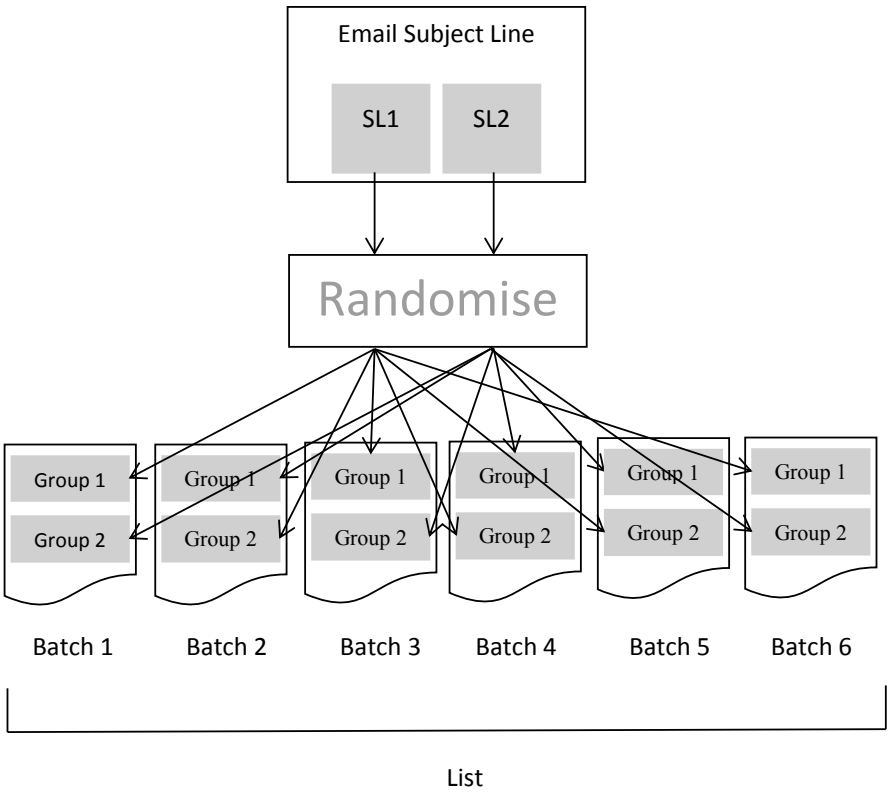

Figure 3. Groups Randomisation

Effect	Numerator DF	Denominator DF	F Stat	P-value
customer	2	108	208.37	<.0001
minimal_order	1	108	0.57	0.4525
customer*minimal_order	2	108	2.08	0.1304
discount	2	108	10.65	<.0001
customer*discount	4	108	5.22	0.0007
minimal_order*discount	2	108	0.00	0.9956
customer*minimal_order*discount	4	108	1.29	0.2784
subject_line	1	108	1.61	0.2072
customer*subject_line	2	108	2.70	0.0717
minimal_order*subject_line	1	108	9.89	0.0021
customer*minimal_order*subject_line	2	108	0.69	0.5059
discount*subject_line	2	108	3.42	0.0364
customer*discount*subject_line	4	108	3.11	0.0183
minimal_order*discount*subject_line	2	108	2.53	0.0843
customer*minimal_order*discount*subject_line	4	108	2.17	0.0767

The following significant factors and interactions were identified:

customer, discount, subject_line, customer*discount, customer*subject_line, customer*discount*subject_line, minimal_order*discount*subject_line, and customer*minimal_order*discount*subject_line.

Now, we built a new regression model and estimated conditions (factor-level combinations) that maximised response (sales). See SAS PROC MIXED code in Appendix 4. For each customer type (C1, C2, C3), the conditions generating maximum sales are presented in the table below:

customer	minimal_ order	discount	subject_ line	predicted sales
C1	$50	10%	SL1	$140,174
C2	$100	10%	SL1	$156,830
C3	$100	10%	SL2	$191,097

In other words, if an email marketing campaign with factors and levels presented in the above table are deployed for 50,000 customers of each type, then Company A should expect, on average, the sales amount presented in "Predicted Sales" column.

Impact on the Business

The split-unit error structure allowed us to discover different interactions that existed in the experimental data. This is because the CRD analysis pools the three error terms – List, Batch, and Group – together, and the resulting error is not appropriate for any of the comparisons. In fact, the split-unit design is more complex, and it has more relationships among factors than CRD could discover.

CRD analysis found that the interactions minimal_order*subject_line and discount*subject_line were significant, while in reality, they were not. On the other hand, split-unit found that subject_line factor and interactions customer*subject_line, minimal_order*discount*subject_line and customer*minimal_order*discount*subject_line were significant, while CRD did not recognise it.

As a result, CRD analysis identified incorrectly the conditions (factor-level combinations) generating a maximum response (sales).

According to the CRD analysis, the best conditions for customer type C1 are 15% discount with minimum purchase of $50 while the email is sent with subject line SL1. These conditions should bring $130,681 in sales on average per 50,000 recipients. However, per our analysis, the model based on CRD analysis is incorrect. If we plug these conditions into the model that was built based on the split-unit analysis, the result will be $127,094, which is 2.7% less. If the campaign is sent to 1,000,000 recipients, it would translate to about $71,000 lower sales than expected.

For the same type of customers, the split-unit analysis identified conditions of 10% discount with minimum purchase of $50 while the email is sent with subject line SL1. Under these conditions, the expected sales from 50,000 of email recipients are $140,174. In comparison with the $127,094 that would be received under conditions identified by CRD analysis, the correct conditions would generate 10.3% more sales. And if the marketing emails with the conditions identified by split-unit analysis is sent to 1,000,000 recipients it would translate to $261,600 higher sales.

When we perform a similar examination for customer type C2, the results are the following:

- CRD analysis suggests that the best conditions (10%, $50, SL1) will generate $168,058.
- If we plug in these conditions into the predictive model based on the split-unit design, the

expected sales are $155,070, which is 7.73% less. Applied to a campaign for 1,000,000 recipients this will produce $259,760 less than expected.

- The split-unit analysis suggests that the best conditions (10%, $100, SL1) will generate $156,830. For 1,000,000 recipients this will produce $35,200 more than based on the conditions identified by CRD analysis.

For customer type C3, the results are the following:

- CRD analysis suggests that the best conditions (15%, $100, SL2) will generate $179,607.

- Plugged in into the split-unit model, these conditions will lead to $168,914 expected sales, 5.95% less. Applied to a campaign for 1,000,000 recipients this will produce $213,860 less than expected.

- The split-unit analysis suggests that the best conditions (10%, $100, SL2) will generate $191,097. For 1,000,000 recipients this will produce $443,660 more than expected from CRD conditions.

Summary

Design of Experiment applied to marketing helps identify factors and their interactions that maximise a marketing campaign's performance (sales or customer purchases).

Failure to identify the appropriate design structure leads to an incorrect analysis of the experiment, and as a result, produces misleading inferences.

Using the real-life example, we demonstrated how to analyse a marketing experiment and identify correct error structure. We showed how to incorporate the split-unit error structure, perform appropriate analyses and build correct predictive models. The comparison of results obtained from CRD vs. split-unit design demonstrated immediate impact on business performance.

References

1. Box, G.E.P., Hunter, W.G., Hunter, J.S. (1978). Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. New York: Wiley.

2. Efron, B. (1998). R. A. Fisher in the 21st century. Statistical Science, 13: 95–122.

3. Hald, A. (1988). A History of Mathematical Statistics. New York: Wiley.

4. Kenward, M., Roger, J. (1997). Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. Biometrics 53, 983-997.

5. Kolosova, T., Berestizhevsky, S. (1998). Programming Techniques for Object-Based Statistical Analysis with SAS Software. Cary, NC: SAS Institute Inc.

6. Littell, R.C.L., Milliken, G.A., Stroup, W.W., Wolfinger, R.D. (1996). SAS System for Mixed Models. Cary, NC: SAS Institute Inc.

7. Wu, C.F.J, Hamada, M. (2000). Experiments: Planning, Analysis, and Parameter Design Optimisation. New York: Wiley.

Tanya Kolosova is an expert in the area of actionable analytics and analytical software development having served as the Senior Vice President of Research and Analytics at IPG Inc, Principal Researcher at Yahoo!, Vice President of Analytics at Nielsen and Chief Analytics Officer at YieldWise Inc. Tanya developed her expertise with extensive depth and breadth of experience in bringing mathematical disciplines to bear on marketing and other business problems. She has extensive knowledge of audience intelligence, design and analysis of marketing experiments, market-mix modelling, and multi-channel commerce and has worked in a variety of industries like online and offline retail, telecom, finance, and more. Tanya also co-authored two books on statistical analysis and metadata-based applications development with SAS, which are used in universities globally and she was featured in Forbes Magazine (2006) for her work for GAP. In 2017 Tanya co-founded InProfix Inc, a stealth mode startup that develops AI solutions for the insurance industry. She is currently a Principal at Associates In Analytics Inc.

Samuel Berestizhevsky is an innovator and actionable analytics expert having served as the Chief Technology Officer at YieldWise Inc. Samuel has extensive knowledge of software development methods and technologies; artificial intelligence methods and algorithms; as well as statistically designed experiments. He has developed and deployed analytical software solutions for a variety of industries like online and offline retail, telecom, finance, and more. Samuel co-authored two books on statistical analysis and metadata-based applications development with SAS which are used in universities globally and was featured in Forbes Magazine (2006) for his work for GAP. In 2017 Samuel and Tanya Kolosova co-founded InProfix Inc, a stealth mode startup that develops AI solutions for the insurance industry.

Appendix 1

The following statements of PROC MIXED fit the completely randomised design model.

```
proc mixed data=experiment cl;
class replication customer minimal_order
discount subject_line;
model sales=customer|minimal_
order|discount|subject_line;
run;
```

The dataset `experiment` contains the experimental table described in the article. The variables `replication`, `customer`, `minimal_order`, `discount`, and `subject_line` are listed as classification variables in the CLASS statement.

`Customer|minimal_order|discount|subject_line` listed on the right side of the MODEL statement mean that the model is built of all possible combinations of these factors. The dependent variable sales is listed on the left side of the MODEL statement.

Appendix 2

The following statements fit the completely randomised design model and estimate prediction according to this model.

```
proc mixed data=experiment cl;
class replication customer min_order discount
subject_line;
model sales=customer discount customer*discount
customer*subject_line
      minimal_order*subject_line
customer*discount*subject_line
      /solution singular=1e-11 ddfm=kr
outpm=prediction;
run;
```

Variables and their combinations listed on the right side of the MODEL statement contain all effects that were identified as significant at the previous step. `ddfm=kr` means that the degrees-of-freedom method of Kenward and Roger (1997) is in effect. `outpm=prediction` requests to create the dataset with predicted sales values.

Appendix 3

The following statements fit the split-plot model assuming random block effects.

```
proc mixed data=experiment cl;
class replication customer minimal_order
discount subject_line;
model sales=customer|minimal_
order|discount|subject_line;
random replication(customer) minimal_
order*discount*replication(customer);
run;
```

Variables and their combinations listed in the RANDOM statement define random block effects.

Appendix 4

The following statements fit the split-plot model with random block effects and estimate prediction according to this model.

```
proc mixed data=experiment cl;
class replication customer minimal_order
discount subject_line;
model sales=customer discount subject_line
customer*discount customer*subject_line
      customer*discount*subject_line minimal_
order*discount*subject_line
      customer*minimal_order*discount*subject_
line
   /solution singular=1e-11 ddfm=kr
outpm=prediction_split;
random replication(customer) minimal_
order*discount*replication(customer);
run;
```

`outpm=prediction_split` requests to create the dataset with predicted sales values according to split-plot model.