Theoretical introduction to `ExpertChoice`

Jed Stephens

March 31, 2020

1 Purpose

The purpose of this vignette is to provide a theoretical explanation of how to design efficiently. Understanding this in conjunction with the `ExpertChoice` package will allow you to design experiments and discrete choice questionnaires in one paradigm of discrete choice experiments.

Hensher et al. (2015, 287) conclude by explaining that Burgess and Street (2005) and Street and Burgess (2007) launched a literature of optimal stated choice experiments which are based on the multinomial logistic regression and optimal linear experiments. Since then the stated choice/discrete choice literature has expanded in many directions and Hensher et al. (2015) explore this in their comprehensive introduction to Applied Choice Analysis. One large contention that this literature is currently grappling with is the difference between D-optimal and D-efficient (Rose and Bliemer, 2009; Walker et al., 2018). This package, `ExpertChoice`, provides an easy, detailed implementation of the Burgess and Street (2005) literature explaining how to create a DCE in this paradigm. As a paradigm it is gaining renewed interest in particular because it creates experiments that are very robust and are especially suited to situations where there is no prior knowledge.

Why is the term 'experiment' used to describe this process and its literature? Oxford dictionary (2019) gives the following definition for experiment: "A scientific procedure undertaken to make a discovery, test a hypothesis, or demonstrate a known fact." The idea that this definition expresses is that there are variables or attributes which are altered, systematically, and it is the effects of these alterations which is of interest to discover.

2 Designing an experiment

The objective of designing a good experiment is a simple one: ensure that the variables of interest are transparently testable to a chosen satisfactory level. Designing an experiment starts before becoming wrapped up in theoretical considerations as to how best to achieve this. The important first step is defining the relevant attributes and their levels. Doing this requires some prior insight. Problem structuring methods (such as those described in Belton and Stewart (2002)) can be incorporated with the first two design stages described by Hensher et al. (2015, 194-201). In the linked Practical Introduction to `ExpertChoice` this process is described as Step 0.

The next section will introduce two examples of experiments where this groundwork has already been done. The first, a restaurant experiment, is relatively of a small size, but has interesting considerations. The second, a silver object experiment, is larger and was the motivation for this package being written. Both examples are fully worked in the linked Practical Introduction to `ExpertChoice`.

2.1 Describing your experiment

An experiment starts by describing the variables which change. It is Step 0. The variables, also referred to as attributes, can be ordered (ordinal) or unordered (categorical). This distinction is important if the intention is to convert the experiment into a discrete choice experiment, but not so much otherwise. This section will describe two experiments: one with all variables unordered (categorical) and another with all variables ordered. It is possible to also have an experiment with a mixture of the two.

2.1.1 Unordered Variable Experiment: Restaurant Experiment

Imagine you own a restaurant that only serves a set menu. As the menu is set your patrons never choose what they are getting for each part of the menu or infact what they are getting on the day. You want to experiment with different set menus to see not only what meals patrons enjoy, but also which starter, main and dessert combinations work well together. In your repertoire you have the following recipes[1]:

starter = {Tomato Soup, Duck Rillettes, Seafood Chowder}
main = {Roast Pheasant, Pan Fried Hake, Pork Belly, Mushroom Risotto, Sirloin Steak, Vegetable Bake}
dessert = {Sticky Toffee Pudding, Chocolate & Hazelnut Brownie, Cheesecake}

Table 1 formally describes this experiment. The different levels of the variables (z_l) are labelled starting from 1 upwards. Starting at 1 is the convention for unordered variables. This design would be called a $3^2 6^1$ design because there are two variables each with three levels and one variable with six levels. As a result there would be 54[2] different variations of set menus that could be offered.

z	**attribute/variable name** (z)	z_l	**level name** (z_l)
1	Starter	1	Tomato Soup
1	Starter	2	Duck Rillettes
1	Starter	3	Seafood Chowder
2	Main	1	Roast Pheasant
2	Main	2	Pan Fried Hake
2	Main	3	Pork Belly
2	Main	4	Mushroom Risotto
2	Main	5	Sirloin Steak
2	Main	6	Vegetable Bake
3	Dessert	1	Sticky Toffee Pudding
3	Dessert	2	Chocolate & Hazelnut Brownie
3	Dessert	3	Cheesecake

Table 1: Design for research on different set menus. (A $3^2 6^1$ design.)

2.1.2 Ordered Variable Experiment

The silver experiment was an attempt to determine utility functions for the different variables from experts in antique silver. Notice here that unlike in Table 1 the levels of the variables (z_l) are labelled

[1](This menu is adapted from the restaurant 101 Talbot http://www.101talbot.ie/menus/)
[2]This number should not be too suprisingly it comes from the design notation that is $54 = 3^2 \times 6^1$

starting from 0 upwards. This is to make explicit the fact that the $z_l = 0$ level will be used as the base level. In unordered variables the base level can be chosen arbitrarily. This experiment would be described as 5^5 and hence there are 3125 possible combinations.

z	**attribute name** (z)	z_l	**level name** (z_l)	**Description**
1	Makers Renown	0	bottom 50% of makers	common
1	Makers Renown	1	50% to 65% of makers	known to specialists
1	Makers Renown	2	65% to 80%	recognised
1	Makers Renown	3	80% to 90%	famous
1	Makers Renown	4	top 10%	celebrated
2	Technical Perfection	0	below 50% of craftmanship	below average
2	Technical Perfection	1	50% to 65%	good
2	Technical Perfection	2	65% to 80%	meritorious
2	Technical Perfection	3	80% to 90%	distinguished
2	Technical Perfection	4	top 10%	exquisite
3	Category Rarity	0	bottom 20%	common
3	Category Rarity	1	20% to 40%	uncommon
3	Category Rarity	2	40% to 60%	rare
3	Category Rarity	3	60% to 80%	very rare
3	Category Rarity	4	top 20%	exceptional
4	Size (of object)	0	under 125g	petite
4	Size (of object)	1	between 126g and 275g	small
4	Size (of object)	2	between 276g and 600g	medium
4	Size (of object)	3	between 601g and 1200g	large
4	Size (of object)	4	exceeds 1200g	extra large
5	Age (of object)	0	1951-present	
5	Age (of object)	1	1900-1950	
5	Age (of object)	2	1851-1899	
5	Age (of object)	3	1801-1850	
5	Age (of object)	4	before 1800	

Table 2: Design for research on antique silver objects to be answered by experts. (A 5^5 design.)

3 The Factorial Designs

3.1 The Full Factorial Design

The full factorial design is constructed from z attributes each with l levels denoted as z_l. It contains all possible combinations of the levels of the attributes and each row is unique. The full factorial for any design has the unique number of combinations (hence the number of rows of the full factorial) given by the design description. Recall for the restaurant and silver experiments that was $3^2 6^1$ and 5^5 respectively. Table 5, at the end of this document gives the full factorial design for the restaurant experiment.

3.2 The Fractional Factorial Design

It is clear that the full factorial design can quickly become overwhelming if the intention was to implement an experiment where each scenario (i.e. row) as conducted. It would also be onerous, costly and more so depending on what is desired unnecessary. The aim of all experimental design is to get to the results faster and as effectively as possible. Therefore for all but toy examples the researcher must concern themselves with how best to select from the full factorial design to form the fractional factorial design. The fractional factorial design is always (by definition) be contained in the full factorial design.

In general there are three methods to select from the full factorial: column based methods (typically methods originating with Federov), row based methods (mixed integer programming) and the construction of orthogonal arrays (Grömping, 2018; Kuhfeld, 2010). There are merits to each method, but before an extensive discussion can be had it is necessary to explain further some of the efficacy measures for factorial designs. These efficacy measures are meant to guide the selection process.

3.2.1 Efficacy Measures for Factorial Designs

A design's main effects are the effects of the each of the attributes measured at each of the levels. Two-attribute interactions are the interaction effects between z^a and z^b where $a \neq b$. In order to have a two-attribute interaction effect there must be at least two factors. Similarly for three factors. The number of estimable n-attribute interactions are related to a efficacy measure described by Xu and Wu (2001) and Grömping and Xu (2014) as generalised word lengths. That is the nth world length is the n-attribute interactions that the full factorial would support. For the menu example there are 3 attributes hence 3-attribute interactions as such the there are generalised word lengths 0,1,2 and 3. For the silver object design there are 5 attributes hence 5-attribute interactions as such the there are generalised word lengths 0,1,2,3,4,5 and 6. Only the full factorial design is capable of supporting the nth world length. But in many instances one's interest is only in the main effects and/or possibly two level interactions. This comes with the major advantage that a fraction of the full factorial may now be used.

Generalised word lengths are a powerful method of assessing design efficacy which has a strong relationship to two more familiar concepts from the DoE literature: resolution and strength. Strength[3] s is equal to the resolution[4] (r) less 1. The first generalised world length, always the zero word length is always 1 i.e. generalised word length (0) = (1). Thereafter the number of zero length words is the strength of the design. For example in the menu design there are 3 word lengths (as explained previously). For the silver experiment if generalised world length (1) = 0, generalised word length (2) = 0 and generalised word length (3) = 0, but generalised word length (4) = 25, i.e. 3 zero length words, then the strength of the design is 3, hence resolution four.

This is significant because, following Kuhfeld (2010), stated generally if resolution (r) is odd then the effects of order $e = (r-1)/2$ or less are estimable free of each other. However at least some of the effects of order e are confounded with interactions of order $e+1$. If r is even then effects of order $e = (r-2)/2$ are estimable free of each other and are also free of interactions of order $e+1$. Table 3 gives some commonly chosen designs and an interpreted description.

[3] Strength is traditionally denoted with a number: 1,2,3,4

[4] Resolution is traditionally denoted in roman numerals or in words

resolution	strength	number of zero length words	description
III	2	2	all main effects are estimable free of each other, but some are confounded with two-attribute interactions
IV	3	3	all main effects are estimable free of each other and free of all two-factor interactions, but some two-attribute interactions are confounded with other two-attribute interactions
V	4	4	all main effects and two-factor interactions are estimable free of each other

Table 3: Commonly chosen designs and their efficacy

The full factorial design has the maximum achievable resolution, strength and number of zero length words. Hence, although it is only possible in toy examples, it is best possible design. It also has two other desirable properties: orthogonality and level balance.

The efficacy of a design for a particular specification can be calculated for a specific stipulation. Let $\mathbf{X}$ be the design matrix of the proposed design (typically this is the fractional factorial design) with an intercept and its attributes expanded using standardised orthogonal contrast coding[5]. The information matrix (familiar from theory of the linear model) is $\mathbf{X^TX}$. The number of rows in the proposed design is denoted N_D. The number of rows (or columns) in the symmetric information matrix ($\mathbf{X^TX}$) is denoted as p. The A-efficiency is defined as

$$\frac{100}{N_D} \times \frac{1}{\text{trace}((\mathbf{X^TX})^{-1})/p} \tag{1}$$

and the D-efficiency[6] as

$$\frac{100}{N_D} \times \frac{1}{\det((\mathbf{X^TX})^{-1})^{(1/p)}}. \tag{2}$$

It cannot be overemphasised that A-efficiency and D-efficiency of a design is specific to the particular model matrix expansion of the proposed design. For example assume that a suitable fractional factorial design for the silver objects experiment is given by the matrix $\mathbf{B}$. The design expansion of the matrix $\mathbf{B}$ would be different when estimating only the main effects (viz. Makers Renown, Technical Perfection, Category Rarity, Size and Age) as it would be when an expansion that included some (or all) interactions (viz. Makers Renown, Technical Perfection, Makers Renown $\times$ Technical Perfection, Category Rarity, Size). Let us assume that matrix $\mathbf{B}$ is of resolution IV then, by definition, the A- and D- efficiency of the main effects design will be 100% i.e. fully efficient. Yet, the second proposed expansion of $\mathbf{B}$ (the one including interactions) may or may not be 100% efficient. It will depends on whether those particular interactions ailise each-other. In general to inspect which effects ailise which other expand the $\mathbf{X}$ up to the value of e (see definition earlier) and investigate the information matrix ($\mathbf{X^TX}$).

To make this discussion about model expansions more concrete, Step 5 of the menu experiment, demonstrates how to programme these tests. The following table, Table 4, summarises the results for the different formulations. The notation '+' indicates the variable is added linearly, while the '$\times$' indicates that the variables and its interactions are added. A 36 run orthogonal array with generalised

[5] Any coding can be used in analysis of the completed experiment. The standardised orthogonal contrast coding has attractive properties when designing an experiment as its efficiency measures are normalised to 100% in the case of the optimal design.

[6] D-efficiency is, in general, a relationship between $[\det(C)/\det(C_{\text{optimal}})]$ where C is the information matrix in the case of the linear model viz. $\mathbf{X^TX}$. For linear models the $\det(C_{\text{optimal}})$ is well known.

word lengths $(0) = 1, (0) = 0, (1) = 0, (2) = 0, (3) = 0.5$ was used. This design hence has strength of 2 (the number of zero lengths words is $(3) - (1) = 2$).

Design Expansion	**minimum strength for full efficiency**	**A-efficiency (%)**	**D-efficiency (%)**
starter + main + dessert	2	100	100
starter × main + dessert	3	93.75	97.164
starter + main × dessert	3	93.75	97.164
starter × dessert + main	3	91.304	95.975
starter × main× dessert[7]	4	NA	NA

Table 4:

The information matrix $(\mathbf{X^TX})$ is also telling about the balance and orthogonality of the design. A design is orthogonal when the sub-matrix of $(\mathbf{X^TX})^{-1}$ (excluding the row and column for the intercept) is diagonal. (There may be off-diagonal non-zeros for the intercept.) A design is balanced when all off diagonal elements in the intercept row and column are zero. When a design is both simultaneously balanced and orthogonal, the $(\mathbf{X^TX})^{-1}$ matrix is diagonal and $(\mathbf{X^TX})^{-1}$ is equal to $\frac{1}{N_D}\mathbf{I_{(p\times p)}}$ (Kuhfeld, 2010, 63). Such a design is a 100% efficient design. That is, practically speaking, the design does not in any way influence the results – it has no systematic bias. All designs less than 100% efficient may have balance or orthogonality or neither.

In the conjoint literature there existed a historical preference for designs that are orthogonal, despite the fact that some of these designs may have been very imbalanced and hence rather inefficient. This literature has now migrated to choosing designed based on D-efficiency (Kuhfeld, 2010). (Which may result in designs which are neither orthogonal nor balanced, but are relatively orthogonal and balanced.) In the discrete choice literature there exists a strong preference for a balanced design (Hensher et al., 2015).

4 Experiments without blocks

A major technique in the design of both conjoint and discrete choice experiments is blocking. Blocking is a design of experiment (DoE) term used to describe a situation where different respondents answer different portions of the chosen design. (In more technical terms blocking is the division of the chosen fractional factorial design.) Blocking is a systematic technique of division – typically blocks are mutually exclusive of one-another (so called "no-overlap designs"), but increasingly often with purposeful overlap (so called "minimal overlap designs").

Why block? Sometimes a fractional factorial design may be too large that it can be reasonably answered by one respondent. Blocking breaks the experiment into smaller "bites" for respondents.

The appropriateness of blocking has come under strong theoretical scrutiny (Hensher et al., 2015; Rose and Bliemer, 2009). Their arguments can be summarised as this, typically in large respondent surveys when blocking is used the result can be imbalance of administration of the blocks. Let us assume that a design is separated into four blocks (A, B, C, D). The study has 50 participants. Firstly, four does not divide 50 equally so the researcher must make the choice of which of the blocks to give to the 49th and 50th respondent. Secondly many things could foul a response: the respondent may wish to withdraw from the study, they may be missing questions or have answered illegibly, etc. Let us assume that there are 47 usable responses and that these consist of 12 A blocks, 7 B blocks, 11 C blocks and 17 D blocks which collectively sum to 47. The problem is now self-evident: analysis happens based on the original chosen design. The blocked reconstruction of the original chosen design is fatally flawed it will introduce imbalances (where they never existed) and will struggle to estimate with the same efficacy. This is strong motivation to avoid blocked designs.

4.1 Selecting from the Full Factorial Design

Constructing the fractional factorial design is very much an iterative process of using a selection method evaluating the design and then typically reiterating. Step 3 in the Practical Introduction to `ExpertChoice` demonstrate how to do so using a column based, row based and orthogonal array approach. Hensher et al. (2015) provide a good introduction to these different methods.

5 Moving from Factorial Design to Discrete Choice Design

A discrete choice experiment consists of several choice sets (denoted as the number of runs) with each choice set containing two or more options (denoted as alternatives). The most apparent difference, the fact that in discrete choice there must be choice within choice sets soon takes forefront concern in converting from a fractional design to a discrete choice design.

For now the focus is to review the different techniques currently available to convert from a factorial design to a discrete choice design. The following section is quashed in a warning about efficiency measures for discrete choice experiments. Methods for laying out DCE are the same regardless of the paradigm for evaluating their end results.

- Modulo Methods (these are proposed by Street and Burgess (2007) and have many advantages)
- L^{MA}[8]
- Rotation Method[9]
- mix-and-match Method [10]

5.1 Optimality Measures for MNL Discrete Choice Designs

The efficacy measures for discrete choice designs inherit much of the literature from the fractional designs (used in conjoint analysis). This can be incorporated into a measure of efficacy for DCE that are used to estimate the main effects or the main effects plus two-factor interactions. Optimal designs will, when using D-optimally criterion have the maximum determinant of the Fisher information matrix. For DCE the information matrix is defined to be: $C = B\Lambda B'$. The B matrix is the matrix of contracts for the effects that are to be estimated. Although theoretically possible, currently the `ExpertChoice` package only supports the construction of the B matrix for main effects. Hence the D-optimallity calculated

[8] "The L^{MA} method directly creates a choice experiment design from an orthogonal main-effect array (Johnson et al. 2007). In this method, an orthogonal main-effect array with M times A columns of L level factors is used to create each choice set that contains M alternatives of A attributes with L levels. Each row of the array corresponds to the alternatives of a choice set." (Aizaki, 2012)

[9] "The rotation method uses an orthogonal main-effect array as the first alternative in each choice set; this method creates one or more additional alternative(s) by adding a constant to each attribute level of the first alternative; the kth(¿= 2) alternative in the jth (= 1, 2, ..., J) choice set is created by adding one to each of the m attributes in the k - 1 th alternative in the jth choice set. If the level of the attribute in the k - 1 th alternative is maximum, then the level of the attribute in the k th alternative is assigned the minimum value." (Aizaki, 2012)

[10] "The mix-and-match method modifies the rotation method by introducing the randomizing process. After placing a set of N alternatives created from the orthogonal main-effect array into an urn, one or more additional set(s) of N alternatives are created using the rotation method and placed into different urn(s). A choice set is generated by selecting one alternative from each urn at random. This selection process is repeated, without replacement, until all the alternatives are assigned to N choice sets. These N choice sets correspond to a choice experiment design." Aizaki (2012)

by `ExpertChoice` will, at this stage, always be for main effects only. This may sound limiting, but it will become apparent that achieving a D-optimal DCE for main effects is sufficiently challenging not to warrant further complication. The Λ matrix is the matrix of second derivatives of the likelihood function which "under the null hypothesis of no differences between the effects of the levels of each attribute turns out that Λ contains the proportions of choice sets in which pairs of profiles appear together" (Street and Burgess, 2007, 462). The entries in Λ can be evaluated by counting the occurrences of pairs of profiles and diving by m^2N where N is the number of choice sets (Street and Burgess, 2007, 462). The D-efficency of the design of any design can be given by $[\det(C)/\det(C_{\text{optimal}})]$.

There exists a theoretical $\det(C_{\text{optimal}})$ for main effects only, first determined by Burgess and Street (2005), for the multinomal logit model. The mathematics of this are given most plainly in Street and Burgess (2007). The `ExpertChoice` package has this functionally built into it. See the Practical Introduction, step 9.

Since then "Bliemer and Rose (2014) were able to show that Street and Burgess designs are simply a special case of the more general methods used by other researchers" (Hensher et al., 2015, 310). In particular this more general method requires a larger introduction than is possible here. For the most neutral introduction see Walker et al. (2018). Much of the dififcult boils down to: "D-optimal designs attempt to maximize attribute level differences whereas D-efficient designs attempt to minimize the elements that are likely to be contained within the AVC matrices of models estimated from data collected using the design." (Rose and Bliemer, 2009)

starter	main	dessert
1	1	1
2	1	1
3	1	1
1	2	1
2	2	1
3	2	1
1	3	1
2	3	1
3	3	1
1	4	1
2	4	1
3	4	1
1	5	1
2	5	1
3	5	1
1	6	1
2	6	1
3	6	1
1	1	2
2	1	2
3	1	2
1	2	2
2	2	2
3	2	2
1	3	2
2	3	2
3	3	2
1	4	2
2	4	2
3	4	2
1	5	2
2	5	2
3	5	2
1	6	2
2	6	2
3	6	2
1	1	3
2	1	3
3	1	3
1	2	3
2	2	3
3	2	3
1	3	3
2	3	3
3	3	3
1	4	3
2	4	3
3	4	3
1	5	3
2	5	3
3	5	3
1	6	3
2	6	3
3	6	3

Table 5: The full factorial design for the restaurant experiment. (A $3^2 6^1$ design.)

References

Hideo Aizaki. Basic functions for supporting an implementation of choice experiments in R. *Journal of Statistical Software, Code Snippets*, 50(2):1–24, 2012. URL `http://www.jstatsoft.org/v50/c02/`.

Valerie Belton and Theodor Stewart. *Multiple criteria decision analysis: an integrated approach*. Springer Science & Business Media, 2002.

Leonie Burgess and Deborah J Street. Optimal designs for choice experiments with asymmetric attributes. *Journal of Statistical Planning and Inference*, 134(1):288–301, 2005.

Ulrike Grömping. R package doe.base for factorial experiments. *Journal of Statistical Software, Articles*, 85(5):1–41, 2018. ISSN 1548-7660. doi: 10.18637/jss.v085.i05. URL `https://www.jstatsoft.org/v085/i05`.

Ulrike Grömping and Hongquan Xu. Generalized resolution for orthogonal arrays. *The Annals of Statistics*, 42(3):918–939, 2014.

David A. Hensher, John M. Rose, and William H. Greene. *Applied Choice Analysis*. Cambridge University Press, 2 edition, 2015. doi: 10.1017/CBO9781316136232.

Warren F. Kuhfeld. Marketing Research Methods in SAS Experimental Design, Choice, Conjoint, and Graphical Techniques, 2010.

John M. Rose and Michiel C. J. Bliemer. Constructing efficient stated choice experimental designs. *Transport Reviews*, 29(5):587–617, 2009. doi: 10.1080/01441640902827623. URL `https://doi.org/10.1080/01441640902827623`.

D.J. Street and L. Burgess. *The Construction of Optimal Stated Choice Experiments: Theory and Methods*. Wiley Series in Probability and Statistics. Wiley, 2007. ISBN 9780470053324.

Joan L Walker, Yanqiao Wang, Mikkel Thorhauge, and Moshe Ben-Akiva. D-efficient or deficient? a robustness analysis of stated choice experimental designs. *Theory and Decision*, 84(2):215–238, 2018.

Hongquan Xu and CF Jeff Wu. Generalized minimum aberration for asymmetrical fractional factorial designs. *The Annals of Statistics*, 29(4):1066–1077, 2001.