**Two-phase experiment literature review**

High-throughput biotechnologies had improved rapidly within last decade; on the other hand, the statistical methods, for analysing the data generated from these technologies, are falling further behind. One of the most important statistical theories during the process of data analysis is *experimental design*, which was initially proposed by Fisher (1935). Experimental design is a set of procedures that outlines an experiment, which allows us to confirm a presence of any valid statistical tests. Most of the experiments involves these high-throughput biotechnologies have a two-phase structure. The first phase consists of perturbing the samples with different conditions of that we would like to study or compare, while to second phase experiment measures a specific attributes of the samples from the phase 1 experiment. In the case of the biotechnologies, the cellular materials from living organisms have been experimentally perturbed, while the second phase involves making measurements (e.g. gene expression or protein abundance) on the material using laboratory-based biotechnologies. These two steps of experimentation are also known as *two-phase experiments*.

A real example can be observed with a gene expression experiment by Jarrett and Ruggiero (2008) to illustrate the two-phase experiment. In the phase 1 experiment, the samples were collected from 21 plants that were arranged in seven blocks of three plots. Seven treatments were assigned in a symmetric balanced incomplete block design. The second phase experiment contains two replicates of the samples from the first phase experiment. A loop design was used to assign these 42 samples onto the 21 microarrays to measure the gene expression levels.

The purpose of this chapter is to describe how the methods surrounding the two-phase experiments have evolved over the last few decades. This will aid in the understanding of the two-phase experiment which can also be employed as a part of important idea in the experimental design.

**The introduction of two-phase experiment by McIntyre**

Two-phase experiment was first introduced by McIntyre in 1955, where he pointed out several important aspects. The samples from the first phase experiment should be able to evaluate separately from the Phase 2 experiment, this means the analysis of variance (ANOVA) table can be constructed without considering the second phase experiment. In addition, it is essential to have replication in the first phase experiment. Since the treatment groups are normally assigned in the Phase 1 experiment and the statistical test is based on comparing between different treatment groups. Hence, the replication allows us to estimate the experimental error in the first phase experiment which is essential to perform the statistical test. On the other hand, the replicates for the second phase are only important when assessing the measurement errors of the technologies or the methods that applied. The main objective in the theory of two-phase experiment is to link the second phase design to the first phase design by considering all sources of variation from both phases of experiment.

McIntyre then discussed five different two-phase experiment examples with several different design combinations. However, his main focus was on the last example which came from a real experiment that investigated the effects of four light treatments on the synthesis of tobacco mosaic viruses in the tobacco leaves. The first phase experiment consists of eight plants and four leaf positions in two four-by-four Latin squares, then the treatment types are assigned within the plots of each Latin square. Hence, there are 32 observations from the phase one experiment. The second phase experiment consists of four four-by-four Greaco-Latin square designs. The light treatments were estimated by injecting sap from each of the first phase leaves into half-leaves of specific assay plants for the second phase experiment. This generates a total of 128. To analysis this study, McIntyre produced an analysis of variance table explaining the sources of variation introduced from overall two-phase experiment. However, the procedure of how each variance components in the table derived was not mentioned.

In the rest of this paper, McIntyre showed two important concepts. Firstly, he used the light treatments experiment on the different design structures mentioned in the first four examples. He demonstrated the different two-phase design combinations can induce different error variances, in addition, the design used in the real experiment proved to be more efficient than the rest. The second concept is about the replication in the design; McIntyre stated that by doubling the replication of the light treatment experiment, it increases the information of estimated error variance by 19%.

**Curnow’s weighted estimate**

Four years later, Curnow (1957) revisits the theory of two-phase experiment and, in particular, McIntyre’s last example. McIntyre used Latin square analysis to estimate the treatment effect which is equivalent to taking the un-weighted estimate the treatment effects. However, the light treatment experiment applied the incomplete block design; hence the treatment information is located in both “inter-block” and “intra-block” strata. If the accuracies of the two treatment estimates are different and can be estimated without excessive error, then Curnow believed that weighting the estimates according to their accuracies would be a better method. In the light treatment experiment, the treatment effects are relatively small compared to the random error from between half leaves and whole leaves of the test plants, hence it is not required to use the weighted estimate for the treatment effects. To estimate the variance components, Curnow also believed using the weighted estimates by combing the intra-block and inter-block can give a more accurate result. However, the results from Curnow’s calculation still generated similar results to McIntyre’s for this light treatment experiment. Nevertheless, Curnow’s idea, with the weighted estimates, has given a first stepping stone on the developing a better analysis procedure for the two-phase experiment.

**Further development on the two-phase experiment theory**

Since the initial introduction of two-phase experiments by McIntyre (1955) and subsequent investigation by Curnow, there was no attention on this theory until a publication by Brien in 1983. There, he proposed a set of steps for determining the experimental structures and rules for deriving the analysis of variance tables, focusing on the two-phase experiment. To describe his method briefly, first step is to identify the factors associated the experimental and observational units, then group these factors into different sets, called *tiers*. Using the Wilkinson-Rogers notation, the structure formulae can be written which shows that the relationship between the factors can be determined in each tier, i.e. crossing or nesting relationships. For the two-phase experiment, there will be at least three tiers of factors, two set of block factors and one set of treatment factors. Hence, two phase experiments also known as *multi-tiered* experiment. To derive the analysis of variance table, first step is to expand the structure formulae for each tier to obtain the linear model, then examining for any confounding of any terms between different tiers. The presence of confounding can be determined by study the contrasts that associated specific factors. For example, if a term from a lower tier is confounded to the another terms from a higher tier, then the terms from the lower tier should be within the stratum of the terms from the higher tier in the analysis of variance table. For a more detail description of these procedures see Brien (1983).

Wood et al (1988) emphasised the importance of studying the non-orthogonal block structure which forms the basis of the difficulties encountered in understanding and analysing two-phase experiments. The idea of using the efficiency factors, introduced initially by Yates (1963), was presented to compute the treatment estimators and generalized least squares estimators. In the case of two-phase experiment, there are two sets of block factors from each phase which can be non-orthogonal to each other; therefore, another efficiency factor is present between these two block structures. Wood et al (1988) showed how to incorporate these efficiency factors for the block and treatment structures, and present a set of formulae that computes these estimates of the treatment estimators and generalized least squares estimators.

In practice, the variance components, from the non-orthogonal block design of two-phase experiments cannot be estimated unless using some computationally intensive methods. Instead, Wood et al noted the idea of using restricted maximum likelihood, described by Patterson and Thompson (1971), to obtain these variance parameter estimates.

Brien and Payne (1999) described a viticultural experiment which has a two-phase structure. They then described an algorithm consists of fitting terms in the model by a series of sweeping operations for analysing the two-phase experiment. There are two types of sweeping operations originally discussed by Wilkinson (1970) and Payne and Wilkinson (1977). The pivotal sweep involves a pivot in which the effects are placed into a unit length vector. This is required for the terms in the first structure formula before sweeps for the terms in the second structure formula which they are confounded. For the two-phase experiments, another pivot sweep is required, since there are always three structure formulae present. The second type of sweep is the reanalysis sweep; this sweeps for the terms from the previous structure to which they are not orthogonal and generally performs after the pivotal sweep. These sweeping sequences allow us to compute the sum of square of the analysis of variance table in the experiment.

**Some publications based on the plant experiments**

A series of plant breeding experiments using two-phase experiment were published since then. Willis et al (2000) investigated the causes of non-indigenous invader plant grows taller and whether it is associated with the genetic level or environmental level. The first phase experiment involves the collection of seeds and the second phase includes the experimental layout of the seedling in the field. The author used the analysis of covariance (ANCOVA) to analyse the data. Due to the missing observations in the experiment, the residual maximum likelihood analysis was also conducted to investigate whether the ANCOVA results were biased by the missing data. It shows the REML results from the ANCOVA were very similar. This study found a little evidence that increased plant size is a genetically determined characteristic of invasive plants.

Smith et al (2001) investigated the genetic mapping of milling yield in wheat using the two-phase experiment. The first phase is a field experiment which the plant materials consist of double haploid line from two mapping populations in a randomised complete block design. One of two populations was divided into two subsets, and the analysis was performed on each of three populations. The grain sample from the field plots are processed in the laboratory for the subsequent second phase. The authors used resolvable incomplete block design and neighbour balanced design for two of three different sets of samples. The author discussed how the linear model and ANOVA table were built. REML was also used in this study to estimate the variance parameters. In their discussion, authors stated further research is required to provide a efficient arrangement of the sample in the field and in the laboratory with adequate replication in both phases.

Cullis et al (2003) presented a barely malting quality experiment. In their study, there were three different sets of data available from three different field trails. The first two were completed in 2000 and used un-replicated grid plot design with a single plot for each doubled haploid line and multiple plots of Arapiles and Franklin arranged in a systematic grids throughout the trial. The trial in 2001 was designed as randomised complete block design with neighbour balance. Each trial was tested separately in the laboratory for the second phase experiment using a randomised block design. These samples were further processed in order to obtain the traits of interest. This phase involves analysing samples in batches which is a nested block design. Hence, this study consists of three phases: the field trial, malting phase and the analysis of starch enzymes in batches. The author followed the methods described in Brien’s (1983) paper to classify the linear mixed model, structure formulae and ANOVA table for this three-phase experiment. REML was again used to estimate the variance components. The result showed the presence of both field spatial and laboratory variation in the experiment. In addition, the authors also confirmed that accounting this variation in both the design and analysis can substantially change the result. In the discussion, the author emphases a need for developing standard designing software that generates an efficient two-phase design which account for both field spatial and laboratory variation. This approach for an exact algorithmic efficient design is believed to be a computational issue.

Smith et al (2006) published a simulation study based on a plant breading and genetic mapping experiment to show the benefits of using the two-phase experimental design. The authors proposed a general linear mixed model that removes any restriction concerning orthogonality of block structure. A new design principle was proposed namely p/q-rep design, where p is the proportion of Phase 1 replicates to the genotypes and q is the proportion of the Phase 2 replicates to the Phase 1 replicates. This study shows that p=q=0.1 was better than the experiment without replication. The study also shows using phase 2 replication always had header realized genetic gain that those did not, which conclude the phase 2 replication is essential for the genetic mapping experiment. There was not any spatial correlation was shown in the phase 1 experiment, which indicating replication in the first is not as essential. At present, it is recommended to use “sensible” values that allow a statistical analysis to be conducted but do not lead to an excessive total number of samples.

**Brien and Bailey’s discussion on multiple randomisation of the two-phase experiment**

Randomisation is a procedure which involves the random assignment of one set of objects to another set of objects. For a two-phase experiment, it requires two stages of the randomisation procedures. The first stage of randomisation involves the treatment factors to the experimental units or the block factors from the phase 1 experiment. The second stage consists of the randomisation of the units from the first phase to the units in the second phase experiment. Randomisation is a critical technique when performing an experiment because it protects against systematic biases. The other idea behind randomisation is that it allows us to make causal inference statements with an associated probability and allows us to make probabilistic statements in connection with results due to chance. In practice, randomisation is achieved by having an initial systematic design where the treatments are orthogonally assigned to the plots of blocks. Then the permuting the blocks and the plots within blocks will gives us a randomised design. Further discussion on the generating this permutation for randomisation is still needed.

In Brien and Bailey’s (2006) paper, they first introduced the randomisation diagram which indicates the factors that are involved with the randomisation in the experiment. There are many symbols can be used to build up this randomisation diagram, these symbols are as follow. The panels indicate the list of factors, with the level number, in each tier. The order of factors listed can specify the relationship between the factors, i.e. nested or crossed, but to avoid the lines or arrows crossing between two panels, the order of the factors in each panel can also be rearranged. The arrow, from left to right, indicates the factor(s), in the left panel, that are randomised to the factor(s), in the right panel. In some cases, the arrow does not point to a specific factor, a series of different symbols are also involved. If the arrow is pointed to a solid circle with two or more lines leading from this solid circle, it signifies a combination of factors in a same tier are involved in a same randomisation. A solid square indicate a combination of factor in the different tiers involved in a same randomisation. Pseudo-factor can be applied where a combination of factors is involved in a randomisation, the initial letter of the factor and numeric subscripts are using to show the presences of pseudo-factor. If the design is non-orthogonal, an empty circle or square is used. An orthogonal design is indicated with a perpendicular sign inside an empty circle or square. A dashed oval surrounds the panel illustrates a pseudo-tier. Diamonds show which factor(s) or pseudo-factor(s) from the left panel have been used to determine the pseudo-factors in the right panel. By applying these symbols, a detailed structure of how factors are randomised can be shown; this is very useful tool especially in the case of two-phase experiment when three tiers are involved.

The authors described six different types of multiple randomisations that involves with the two-phase experiments. These are composed, coincident, independent, double, randomised-inclusive and un-randomised-inclusive multiple randomisations. For the first four types of multiple randomisations, the randomisation procedures can be performed in any order.

Composed randomisation is a selected set from the first objects randomised to another set of object, and this set of objects is randomised onto a third set of objects. Coincident randomisation is two different sets of objects are both randomised to the same un-randomised factors. On the other hand, independent randomisation is two different sets of objects are both randomised to different un-randomised factors in the third set of objects. Double randomisation is a set of objects randomised to another two sets of objects.

The last two multiple randomisation procedures are inclusive randomisation, this means the factors from both tiers of the first randomisation are explicitly involved in determining the design for the second randomisation. To show the presence of inclusive randomisation, the two tiers of the first randomisation forms a pseudo-tier which indicated by a dotted oval around the panels of these two tiers. Randomised inclusive randomisation is used when the first randomisation involves first two sets of factors, excluding the un-randomised factor, which forms a pseudo-tier. Then, the second randomisation takes into account of the first two sets of factors of first randomisation. Un-randomised inclusive randomisation is used when the first randomisation involves the second set of factors and the last set of un-randomised factors to form a pseudo-tier. Hence, the randomisation from the first set of factors will also need to consider the last set of un-randomised factors. Another different type of un-randomised inclusive randomisation is called incoherent un-randomised inclusive randomisation. This is a more complicated type of randomisation and is used when a factor that is crossed from the first randomisation become nested in the second randomisation. By treating the factors as nested instead of crossed, it has loosen the restriction and ensure the factors can be separated out easily.

If more than three tiers are involved in an experiment, then more than one type of the above multiple randomisations mentioned can be applied, but it is out of the context of the two-phase experiment. This paper has introduced a very informative concept in terms of randomisation for two-phase experiment.

**Producing the decomposition table for the two-phase experiment**

Building up from the ideas of multiple randomisations, Brien and Bailey (2009) produced another paper on the aspects of orthogonal decomposition of the data space for the two-phase experiment. They defined the decomposition as a set of orthogonal subspaces of the data space. This can be shown in the decomposition table where each rows corresponds to one of the subspaces in the decomposition, also known as sources of variation. In addition, this decomposition table will also indicate the degrees of freedom and efficiency factors for each source of variation. This is also the essence of the analysis of variance table. The decomposition procedure starts with establishing a randomisation diagram of the experiment and then determines the properties of the designs by obtaining the decomposition table.

The panel within each randomisation diagram defines a tier. To show the structure of a single tier, the authors used the Hasse diagram to illustrate the structure of the relationships between each factor in the experiment and calculate the degrees of freedom. Once the relationships between different factors within each tier are defined, the Hasse diagram can then be used to compute the orthogonal projectors for each source. These projectors are the essential scheme to carry out the decomposition.

The ideas from Nelder (1968) and James and Wilkinson (1971), for an experiment with a single randomisation, were shown in detail by combining the block and treatment structures to produce a single decomposition table. The authors then further extended this concept to the multiple randomisations, i.e. the composed and randomised inclusive randomisations.

**Moving toward a high-throughput biotechnology**

Jarrett and Ruggiero (2008) looked at the two-phase experiment for gene expression two-colour microarray. Initially, they discussed the different orders of fitting the factors to the model can affect the structure of ANOVA table and the way to analyse the treatment effects i.e. computing the F-ratio. Two different types of designs were investigated, these were multiple dye-swap and alternating loop designs. The results shows the multiple dye-swap design is a more robust method, because there are always have at least (r-1) DF for estimating the variance for treatment comparisons. REML was again shown to be a better approach to estimate the variance components. The effectiveness of the estimate was assessed using the effective degrees of freedom computed by the first two moments to those of an approximating chi-square distribution.

The most recent paper by Brien et al. (2011) discuss a systemic approach in designing the Phase 2 experiment taking into account of the Phase 1 experiment. The first principle is to formulate the skeleton ANOVA table using this factor-allocation diagram. The factor-allocation diagram is the randomisation diagram as discussed in Brien and Bailey (2006). In this paper, the author presents a list of rules for calculating the expected mean squares of the ANOVA table. The Chapter 2 of this thesis will present an R package which can automatically produce a skeleton of ANOVA table. The author also address some of the fundamentals in the designing the experiments. These are replication, to measure the random error, randomisation, to avoid systematic biases, and blocking to reduce the variation among experimental units.

Brien et al. (2011) states the replication of the Phase 1 experiment is only required where there is uncontrolled variation in the Phase 2 experiment. For the cases of MudPIT-iTRAQ experiment, the variation between MudPIT runs is known to be large; hence the replication of the Phase 1 experiment is required. The treatment effects of the Phase 2 experiment for the MudPIT-iTRAQ experiment is just tag effect which generally is not the main interest for estimation of the biologists. Thus, in Chapter 4 in finding the optimal design where the Phase 1 experiment is randomised block design, we will show that by confounding the block factor of the Phase 1 experiment with the tag effects can maximised the degrees of freedom in estimating the variance of the treatment effects.

The block factor of the Phase 1 experiment with the highest variation but without treatment information should be confounded with the block factor of the Phase 2 experiment with the highest variation. This situation can be observed in Chapter 4 in finding the optimal design where the Phase 1 experiment is randomised block design. The block factor of the Phase 1 experiment can also be assigned in such way that the block is confounded more with run.

If a given block factor of the Phase 1 experiment contain treatment information, this block factor should be confounded with the block factor of the Phase 2 experiment with the smallest variation.

The treatment should be assigned with the random effects of Phase 2 experiment which consist of the smallest variation. The treatment can be confounded with multiple random effects of Phase 2 experiment which is caused by the non-orthogonal design. However, this paper only focuses on the orthogonal design. In the chapter where searching for the optimal design, there are some cases with the non-orthogonal design will be described.

Pseudofactors can be used to group together a set of levels of a factor and to be randomised for the other factor. This will allows us to keep track of all factors in the experiments or to procedure structure-balanced designs. However, the use of pesudofactor can cause the given factor split which may result in the degradation in the estimation.

If the randomisation procedure is complicated, the factor-allocation diagram can also be complicated. Thus, the author proposed to start with a simple composed randomisation method before using a more complicated randomisation procedure. In addition, all of the Phase 1 factors should always allocate to the Phase 2 factor, and randomise them when it is possible.

Most importantly, plan both phases of the experiment before the biologists commencing any of them. However, typically the biologists would have already done the Phase 1 experiment before consulting in the strategies of designing the Phase 2 experiment.

Problems with using a single formula

Does not display confounding

With a single structure formula each contrast is associated with only a single term and so there is no indication of the multiple sources associated with some of the contrasts as a result of the confounding arising from the randomization. For example in the full decomposition table derived from three structure formula it is clear that the one degree of freedom associated with Rosemary is confounded with Meatloaves[Blocks] which in turn is confounded with Panellists#Time-orders[Sessions]. With a single formula there is only the term Rosemary associated with this one degree of freedom.

Does not correctly identify sources of variation

It is very important to understand that we do not regard the Block interactions with Rosemary and Irradiation as indicating actual sources of variation for which we are including terms in the model. Rather they are merely convenient way of obtaining the sums if squares. Summing these interactions is one way of obtaining the Residual for Meatloaves[Blocks]. However, the term in the model is Meatloaves[Blocks], not Blocks#Rosemary + Blocks#Irradiation + Blocks#Rosemary#Irradiation. So we have a term for variability between meatloaves within a block in the model and we are assuming that there is no interaction between blocks and rosemary and irradiation treatments.

Cannot obtain full analysis of nonorthogonal experiments

As explained by Brien and Payne (1999) one obtains the correct intrablock analysis with a single strucutre formula. However, to obtain the interblock analysis, for use in recombination of information, requires the use of multiple stucture formulae

pseudofactor A factor that groups levels of another factor, but does not represent a new source of differences in the experiment. It is usually used as to identify subspaces of a term that will produce a balanced analysis. They occur more frequently in multitiered than in two-tiered experiments. They were introduced by Yates in connection with balanced lattice square designs. The pseudofactors here represent groups of treatments to be applied to rows or columns in a row-column design. The pseudofactors still represent Treatment contrasts -- it is just that recognition of these Treatment groups aids the analysis in that it is balanced with respect to the pseudofactors. For a fuller discussion see Monod and Bailey (1992).