Published in final edited form as:

Stat Med. 2017 October 30; 36(24): 3910–3918. doi:10.1002/sim.7396.

LINEAR COMBINATIONS COME ALIVE IN CROSSOVER DESIGNS

Jonathan J Shuster¹

¹Department of Health Outcomes and Policy, College of Medicine, University of Florida, Gainesville, FL 32610, shusterj@ufl.edu

Abstract

Before learning anything about statistical inference in beginning service courses in biostatistics, students learn how to calculate the mean and variance of linear combinations of random variables. Practical pre-calculus examples of the importance of these exercises can be helpful for instructors, the target audience of this paper. We shall present applications to the “one-sample” and “two-sample” methods for randomized short term two-treatment crossover studies, where patients experience both treatments in random order with a “washout” between the active treatment periods. First, we show that the two-sample method is preferred as it eliminates “conditional bias” when sample sizes by order differ and produces a smaller variance. We also demonstrate that it is usually advisable to use the differences in post-tests (ignoring baseline and post washout values) rather than the differences between the changes in treatment from the start of the period to the end of the period (“delta of delta”). Although the intent is not to provide a definitive discussion of crossover designs, we provide a section and references to excellent alternative methods, where instructors can provide motivation to students to explore the topic in greater detail in future readings or courses.

Keywords

Crossover Trial; Delta of Delta; Linear Combination; Relative Efficiency; Variance

1. INTRODUCTION

The prime purpose of this paper is to provide instructors of biostatistics service courses with a practical example of finding and using the mean and variance of a linear combination of random variables. The randomized crossover study is utilized, as it has a rich set of problems students can grasp, without the need for complicated mathematics. Assume that by a 50-50 random assignment, patients will be assigned one treatment (either A or B) for a period of 2-weeks, “washed out” for 2-weeks, and then given the alternate treatment (B or A) for 2-weeks. We shall compare the means and variances of two potential outcome (dependent) variables, (i) the difference between the changes from the start of the period to the end of the period (delta of delta) vs. (ii) the difference between the end of period values (delta). The issues we seek to resolve by linear combinations are the following. (a) How do we use the information, whether (i) or (ii)? and (b) Which of the strategies (i) or (ii) is superior?

The students learn that if $Y = \sum A_j X_j$ is a linear combination of random variables $X_1, X_2, \dots, X_N$ then

$$E(Y) = \sum_j A_j E(X_j) \quad (1)$$

$$\text{Var}(Y) = \sum_j A_j^2 \text{Var}(X_j) + 2 \sum_{j < k} A_j A_k \text{Cov}(X_j, X_k). \quad (2)$$

From (1) and (2) students learn that if $X_1, \dots, X_N$ are independent identically distributed random variables with mean μ and variance σ^2 , then by setting $A_j = 1/N$ in (1) and (2), $E(\bar{X}) = \mu$ and $\text{Var}(\bar{X}) = \sigma^2/N$ where $\bar{X}$ is the arithmetic mean of the N observations.

They also learn that a small variance is a desirable objective.

Section 2 is devoted to the question as to demonstrating that the “two-sample” method is preferred to the “one-sample” method both because of (a) smaller variance and (b) when the actual sample sizes differ by order, elimination of “conditional bias” is accomplished by the two-sample, but not the one-sample method. Section 3 provides a detailed look at how to define and estimate the “effect size” in both the difference of results after treatment (ignoring the start of period, delta) and for the delta of delta method, the difference in changes from beginning to the end of the periods. Section 4, under two different sets of assumptions concerning repeated measures correlation, will demonstrate that if a strong washout can be achieved, ignoring start of period data will generally produce a smaller variance, while the two methods will estimate essentially the same treatment effect. Section 5 presents a numerical example. For future interest of instructors and some more advanced students, Section 6 is devoted to a discussion and to important references on the subject. In fact, the topic of crossover studies is rich enough to include as a full semester course. Two excellent texts on the subject are Senn [1] and Jones and Kenward [2]. We restrict the content to two-treatment crossover studies. A discussion of missing data is beyond the scope of this tutorial. Section 7 presents two exercises involving completely randomized designs: One is the likely advantage of using the difference from baseline rather than just the end of study value as an endpoint, and second, under certain assumptions, the crossover study will produce a variance of about half as much as a corresponding completely randomized study. In other words, we expect full mileage from the second period.

2. ONE-SAMPLE VS. 2-SAMPLE APPROACH

In this section, we shall first presume we have a randomized crossover study with an outcome for each subject for each period, denoted by X_{ij} where i =period (1 or 2) and j =treatment order ($j=1$ for AB and $j=2$ for BA). Table I provides the population means. A subject will experience only one treatment order.

We define important parameters of interest as follows for issues surrounding the mean of treatment B–treatment A:

Effect size for Period 1: 2τ

Effect size for Period 2: $2\tau+2\lambda$

Average Period effect size (Our overall effect size) over the two orderings: $2\tau+\lambda$

Expected difference for Order AB: $\pi_2 - \pi_1 + 2\tau + \lambda$

Expected difference for Order BA: $\pi_1 - \pi_2 + 2\tau + \lambda$

Average Expected difference over orders: $2\tau+\lambda$ =Average period effect size

Treatment by Period Interaction, Difference between effect sizes, Period 2 –Period 1: 2λ

Treatment by Order Interaction: B–A difference, order AB minus order BA: $2(\pi_2 - \pi_1)$

“Pure Carryover effect”: Mean difference, Period 2–Period 1: $\pi_2 - \pi_1$

Note that if there is no pure carryover effect, there is no treatment by order interaction. That is the mean B–A treatment difference for order BA is the same as for order AB.

One-Sample Approach

Consider a single observation from this study. We will flip a balanced coin for the ordering and define the outcome per Table II. Only one of the columns will be observed, but without loss of generality, let us assume the unobserved X_{ij} exist.

The observed paired difference (B–A) is

$$Y = T(X_{21} - X_{11}) + (1 - T)(X_{12} - X_{22}) \quad (3)$$

where $T=1$ if the order is AB and $T=0$ if the order is BA.

T is independent of the X_{ij} , and takes values 1 or 0 with probability 0.5 each.

From Table I,

$$E(Y) = 0.5[\pi_2 - \pi_1 + 2\tau + \lambda] + 0.5[\pi_1 - \pi_2 + 2\tau + \lambda] = 2\tau + \lambda \quad (4)$$

Each Y is unconditionally unbiased for the arithmetic mean of the treatment differences B –A for the two orderings.

To calculate the variance of Y , note that $T(1-T)=0$, and T is independent of the X 's

$$\begin{aligned}
 E(Y^2) &= E(T^2)E[(X_{21} - X_{11})^2] + E[(1 - T)^2]E[(X_{12} - X_{22})^2] \\
 &= 0.5\text{Var}(X_{21} - X_{11}) + 0.5\text{Var}(X_{12} - X_{22}) + 0.5(\pi_2 - \pi_1 + 2\tau + \lambda)^2 + 0.5(\pi_1 - \pi_2 + 2\tau + \lambda)^2
 \end{aligned}$$

(5)

$$E(Y)^2 = (2\tau + \lambda)^2 \quad (6)$$

Upon taking the difference,

$$\text{Var}(Y) = E(Y^2) - E(Y)^2 = 0.5\text{Var}(X_{21} - X_{11}) + 0.5\text{Var}(X_{12} - X_{22}) + (\pi_2 - \pi_1)^2 \quad (7)$$

As noted in the Introduction, if we had a sample of N independent realizations of Y , the sample mean of these, $\bar{Y}$ is unbiased for the “mean effect size”, $2\tau + \lambda$, and the variance is

$$\text{Var}(\bar{Y}) = \text{Var}(Y)/N \quad (8)$$

Two-Sample Approach

This basic approach was suggested by Hills and Armitage [3].

Let us look at the following Period 2 –Period 1 half differences for the two orderings:

$$Y_{AB} = .5(X_{21} - X_{11}) \text{ and } Y_{BA} = .5(X_{22} - X_{12})$$

From the last row of Table I:

$$E(Y_{AB}) = .5(\pi_2 - \pi_1) + \tau + 0.5\lambda \text{ and } E(Y_{BA}) = .5(\pi_2 - \pi_1) - \tau - 0.5\lambda$$

$$E(Y_{AB}) - E(Y_{BA}) = 2\tau + \lambda. \{\text{same as (4)}\}$$

If we had N_1 and N_2 observations from the order AB and BA respectfully, then

$$E(\bar{Y}_{AB}) - E(\bar{Y}_{BA}) = 2\tau + \lambda \text{ and}$$

$$\text{Var}((\bar{Y}_{AB} - \bar{Y}_{BA})) = [\text{Var}(X_{21} - X_{11})/\{4N_1\} + \text{Var}(X_{12} - X_{22})/\{4N_2\}]. \quad (9)$$

Specifically, when $N_1 = N_2 = 0.5N$

$$\text{Var}((\bar{Y}_{AB} - \bar{Y}_{BA})) = 0.5[\text{Var}(X_{21} - X_{11})/N] + 0.5\text{Var}(X_{12} - X_{22})/N]. \quad (10)$$

Note that the two-sample method is unbiased for the same parameter as the one-sample method, but per (7) and (8), its variance is smaller unless the carryover effect is exactly zero (i.e. there is no treatment by order interaction).

The Relative efficiency of the two-sample approach to the one-sample approach is

$$\text{RE} = \text{Var}(\bar{Y}) / \text{Var}((\bar{Y}_{AB} - \bar{Y}_{BA})) = 1 + K \quad (11)$$

$$\text{where } K = (\pi_2 - \pi_1)^2 / \{0.5[\text{Var}(X_{21} - X_{11})] + 0.5\text{Var}(X_{12} - X_{22})\}. \quad (12)$$

Relative efficiency represents the approximate sample size ratio the less efficient to the more efficient needs to get the same precision as the more efficient.

Note that for the one-sample method, if there is a real carryover effect, forced balancing of the sample sizes in each order will eliminate conditional bias due to carryover, but it does not eliminate the impact on the increased variance over and above the two-sample method. In addition, careful inspection of Y in (3) reveals that it is a random mixture of orders AB and BA, with discrete mixing parameter T . We are therefore certain that unless there is no carryover effect, Y does not have a normal distribution. This would not be a problem for large sample applications, however. The two-sample approach does not have this limitation.

If the actual sample sizes are unequal, in the presence of a true carryover effect, the two-sample method but not the one-sample method would be conditionally unbiased, given the sample sizes.

In conclusion, the two-sample approach is superior to the one-sample approach, which should be avoided.

3. PLAUSIBLE DEPENDENT VARIABLE CHOICES FOR RANDOMIZED PHARMACEUTICAL CROSSOVER STUDIES

In pharmaceutical randomized crossover studies, with short term endpoints, there is often a baseline taken for each subject (often after a washout if on chronic medication) followed by one treatment, then a washout, followed by the second treatment. The order of treatments is randomized. The question we shall resolve for this type of crossover study is how to develop robust definitions of treatment effects for both (a) the use of post-tests only and (b) the use of changes from the start of the period to the end of the period (delta of delta). Our Table III, which is a slightly modified version of Table II of Jones and Kenward [2] will clarify the

issues, recalling from the previous section that the two-sample method produces unbiased estimates of the arithmetic mean of the “effect sizes” in Periods 1 and 2.

There are seven independent parameters: μ , τ , θ , λ , π_j $j=1,2,3$. Note that for randomized studies, since the subjects are drawn from a single population and since no treatment has been given at baseline, $E_{01} = E_{02}$. The parameters π_j represent the difference between the arithmetic mean of the two expectations in period j less the common mean, μ at baseline. These are the carryover effects. Formally, $\pi_j = \{(E_{j1} + E_{j2})/2\} - E_{01}$, $j=1,2,3$.

The parameter 2τ represents the pure treatment effect after the first period ($B-A$).

The parameter 2θ represents the interaction between treatment and washout. At baseline, the true mean difference between the treatments is zero. After washout, the true mean difference (after B minus after A) is 2θ .

The parameter 2λ is the interaction for effect sizes between treatment and period. The difference in means ($B-A$) after the first treatment is $(E_{12} - E_{11})=2\tau$, but after the second treatment is $(E_{31} - E_{32})=(2\tau+2\lambda)$. Hence the difference in treatment effects, $B-A$ for period 2– period 1 is 2λ .

A robust definition of the overall treatment effect is the arithmetic mean of the treatment effects in periods 1 and 2, either the average treatment effect in the post-test only analysis $2\tau + \lambda$ or the average change from the start of the period $2\tau + \lambda + \theta$ in the delta of delta analysis.

If the washout in a pharmacological intervention can on average get the subject back into their baseline condition, a common capability for short acting drugs, then the two dependent variable choices (post-test differences and delta of delta) will estimate the same target parameter. If that is the case, the choice of methods should depend upon the relative variances of the two methods. If that is not the case, then it is unclear as to which method is better.

Caution: Although this formulation of endpoints leads to virtually assumption-free data analysis, the study science may dictate the use of a different outcome. Section 6 will briefly consider this issue. But these endpoints are valid in many applications, and hence the exercises in Section 4 are of practical importance.

4. VARIANCE CONSIDERATIONS IN CROSSOVER STUDIES

In this section, we shall consider the variances for the difference in post-tests vs. the difference in changes from the start to the end of each period. We shall assume that the washout is full, namely $\theta=0$. (See Table 2). Note, however, that a non-zero pure carryover effect can exist, without altering the mean or variance of the outcome statistics for either approach.

Difference in Post-tests (Delta):

$$U_{ik}=0.5[(X_{3jk} - X_{1jk})] \quad (13)$$

Difference in Differences (Delta of Delta):

$$V_{jk}=0.5[(X_{3jk} - X_{2jk}) - (X_{1jk} - X_{0jk})] \quad (14)$$

where X_{ijk} is the k th-subject's measurement under order j ($j=1$ for AB and $j=2$ for BA), and $i=0$ (Baseline), $i=1$ (After period 1), $i=2$ (After washout), and $i=3$ (After Period 2).

We present two problems below, where we compare the variances of the delta vs. delta-of-delta variances. These assumptions can be useful in study design in certain situations, but the analysis we propose in the second paragraph of Section 6 does not require these assumptions.

4.1 Problem 1

Assuming per (15) below, that we have a “compound symmetric covariance” structure, which endpoint has lower variance? While this structure seems reasonable in our numerical example of Section 5, this assumption is mainly for exercise purposes, as it does not allow the correlation between measurements further apart in time to be lower than those closer together in time.

$$\text{Var}(X_{ijk})=\sigma^2 \quad \text{Cov}(X_{ijk}, X_{mjk})=\rho \sigma^2 \quad (m \neq i). \quad (15)$$

It follows from (2) that $\text{Var}(V_{jk})=(1-\rho)\sigma^2$ and $\text{Var}(U_{jk})=0.5[(1-\rho)\sigma^2]$

Hence, the “Relative efficiency” of U_{jk} to V_{jk} defined as $\text{Var}(V_{jk})/\text{Var}(U_{jk})=2$, making U_{jk} the method of choice. The use of V_{jk} doubles the variance compared with the use of U_{jk} .

Kenward and Roger [4] derived this relative efficiency for compound symmetric covariance structure, using more advanced mathematical methods.

4.2 Problem 2

Assuming per (16) below, we have an autoregressive covariance structure, which endpoint has lower variance? This exercise does allow higher correlations between measurements closer together in time to be lower than those further apart, but a geometric progression may or may not be reasonable in practice. However, Chi [5] has written an excellent article on using this type of correlation in crossover studies. But also see a discussion by Senn [6].

$$\text{Var}(X_{ijk})=\sigma^2 \quad \text{Cov}(X_{ijk}, X_{mjk})=\gamma^{|i-m|}\sigma^2, 0<\gamma<1. \quad (16)$$

Again using (1),

$$\text{Var}(V_{jk}) = \sigma^2[4 - 2\gamma - 4\gamma^2 + 2\gamma^3]/4 = \sigma^2[2 - \gamma][2 - 2\gamma^2]/4 = \text{Var}(U_{jk})[2 - \gamma].$$

Hence, the Relative efficiency of U_{jk} to V_{jk} , $\text{Var}(V_{jk})/\text{Var}(U_{jk}) = 2 - \gamma > 1$ when $0 < \gamma < 1$.

Note that only as γ approaches 1, will the relative efficiency of V_{jk} approach 1. But in that case, all of the correlations approach 1, making this problem virtually deterministic.

5. NUMERICAL EXAMPLE

In order to illustrate the methods, we present summary data from a real trial, whose patient level data has not been released to the public as yet. We therefore must present them as generic data. This is a randomized crossover trial with a washout, and the outcomes are provided in Tables IV and V.

Imagine that we used these data to design a future study.

The estimated relative efficiency of $U:V = \text{Var}(V)/\text{Var}(U) = [\text{SD}(V)/\text{SD}(U)]^2$

$$= 1.42 \quad \text{A then B}$$

$$= 2.64 \quad \text{B then A}$$

We estimate that a delta of delta analysis would need to inflate the sample size by 42% (164%) compared to that of a delta analysis, to get the same precision as a simple difference of the end of period results for orders AB(BA) respectively.

This paragraph can be skipped by those who have not yet studied statistical inference

Using the U_{jk} , {Difference in post-tests per (13)} with $N=64$ per order, the point estimate = -0.94 and standard error 1.64 , whereas using the V_{jk} , {Delta of Delta per (14)} with $N=64$ per order, the point estimate = 0.63 and standard error 2.18 . The overall estimated relative efficiency, end of period data only: delta of delta is $(2.18/1.64)^2 = 1.77$. In planning a replication of this study, a study of 64 per order using the end of period difference only would be estimated to have the same power of a study of 113 per order using delta of delta.

Inspection of Table 5 suggests that compound symmetry is a plausible approximation, and the final relative efficiency estimate of 1.77 is not far removed from the 2.00 suggested by compound symmetry.

6. ADVANCED DISCUSSION OF CROSSOVER DESIGNS

Although the intent of this article is to provide an interesting practical example for assessing means and variances of linear combinations of random variables, we provide a brief discussion of the topic of two-period crossover designs for interested instructors or advanced students.

First, let us begin by considering the two-sample method as discussed in Sections 2–4. If you use a Satterthwaite corrected t-test [7], the only assumption you are making is that you have a sufficient sample, such that the sample means are approximately normally distributed. It should be noted that Senn [8] and Shuster [9] describe the robustness to lack of normality, as long as outliers (values that are far removed from the central cloud of data) are lacking. But you are estimating a parameter, which may or may not be ideal for the end user. This method estimates the arithmetic mean of the effect sizes in periods 1 and 2, or equivalently the arithmetic mean of the treatment differences for order AB and order BA. The resulting estimates are unbiased for the above parameter, even when there exists treatment by period interaction or carryover effects. It is common, however, to seek to estimate the effect size in period 1. This parameter is attractive because it has no “baggage” from the second period, and is the same parameter estimated for a single stage randomized design. But unlike the arithmetic mean parameter, you can only do this completely cleanly if you ignore period 2. To use the data from period 2, you have no choice but to make certain unverifiable assumptions. We discuss the assumption issue further, below. Fleiss [10] provides an excellent summary of some of the difficulties in the methods that were published up to that time. In addition, as we state below, some crossover studies do not fit the mode we laid out in sections 2–4, and clearly need alternate methodology.

There are two excellent books on crossover designs. Senn [1] considers the one- and two-sample approaches in Chapter 3, but looks at other important topics as well: transformations, binary data (yes/no), and ordinal data (nonparametric: qualitative ratings whose numeric values only identify order but otherwise have no intrinsic importance) in Chapter 4; studies with more than two treatments in Chapter 5, and issues surrounding dropouts and missing data in Chapter 9.6. Jones and Kenward [2] considers the repeated measures aspects of crossover designs using mixed linear models for analysis in Chapter 5, analyzes discrete data in Chapter 6, discusses a Bayes approach in Chapter 2, and reviews the use of covariates in Chapter 2. Note that as explained in Kenward and Roger [4] and Senn [11], the use of period baselines as covariates is far from straight forward, unless used in an appropriate time-ordered linear mixed models.

The careful use of baseline covariates in such mixed models can be more efficient than our two-sample method described herein. However, these methods make distributional assumptions over and above those made in our context.

A popular method for crossover studies was advocated by Grizzle [12], where the analysis is conducted in two stages on the end of period results. In the first stage, a test for interaction between treatment and period is conducted (i.e. $H_0: \lambda=0$ in Table III). If the hypothesis is rejected, then in Stage 2, the analysis is done strictly on the first period. If the hypothesis is not rejected, then in Stage 2, the analysis will be done using the U_{jk} , per (13). Freeman [13] and Senn [11] reported that the Grizzle [12] method, despite 597 citations in the Web of Science as of 5/10/17, has major bias. As described in Shuster [14], there are two major problems with this strategy. First, the impact of the diagnostic test for treatment by period interaction is unknown when it comes to significance testing or point and interval estimation for the treatment effect. The error properties generally reported presume that the diagnostic test(s) produced the correct inference as to whether the assumptions were correct. The error

properties need to be evaluated from before the branching of the diagnostic test. In other words, does the resulting 95% confidence interval from before the diagnostic test produce 95% coverage? Second, a diagnostic test is simply a test of hypothesis. Failure to reject simply says the result is inconclusive. It cannot be used to prove no treatment by period interaction exists, just because there is no statistically significant observed interaction.

The use of methods such as Linear mixed models is very tempting because of the potential gain in efficiency. Here too, diagnostic tests for the parallel slope assumption do not work as a two-stage procedure analogous to Grizzle [12]. However, Zou et. al [15] proposed a mixed model approach that does control the study error from before the branching in a diagnostic test that has treatment by baseline covariate interaction in the model.

Two cautions: (i) Our proposed methods apply mainly to short term trials and to trials with a reasonably good washout. A referee correctly points out that the delta of delta method can have an advantage when this is not the case. Note that if for example, in a drug vs. placebo trial, the drug effect is permanent in a large fraction of potential patients. If you just use the post-tests, the placebo period following drug will see no difference in delta for these patients, whereas delta of delta will see a big difference in delta of delta, thanks to the big change in period 1. Of course, if we knew this in advance, we might prefer a single stage design for our primary question. Alternatively, a very long term washout period could resolve the issue, but in practice, that could lead to significant loss of subjects. (ii) When you conduct long term crossover studies, even where the much-feared dropout is not considered a major issue, it is advisable to collect the endpoint serially in each stage, rather than simply at the beginning and the end. In such a context, linear mixed models need to be employed, as the gain in efficiency over the methods of Sections 2–4, which discard a large portion of the data, is critical to a complete analysis. Further, mixed models do a better job of handling missing data.

In conclusion, the methods presented in Sections 2–4 are attractive for short-term studies in pharmaceutical companies who want to make their analyses as close as possible to being assumption-free. As long as their sample sizes are reasonably large, these methods accomplish this when the data are complete. When data are missing or subjects drop out, all methods have issues. It is beyond the intended scope of this article to address this question. A recent article on this subject by Ho et. al [16] suggests that a commonly used assumption “Missing at Random” may not work well in practice, and suggest methods that work well when there is no carryover effect. Note that missing values have a larger negative impact for our methods than for mixed models. Senn [11] made two relevant wishes in his essay, which the present paper unequivocally supports: (A) The use of the two-stage method should be discontinued; and (B) the use of “CROS”, which matches up exactly to the use of U_{jk} from (13) should be the method of choice, (at least for those with reluctance to using covariates or mixed models-author caveat added). The wish (A), after over 20 years still seems not to have been sufficiently heeded with 14, 6, 10, 19, and 13 Web of Science citations to Grizzle [10] in the years 2012, 2013, 2014, 2015, and 2016 respectively.

7. STUDENT EXERCISES

(A) The baseline and period 1 form a legitimate classical randomized single stage study. Should the baseline period result be subtracted from the end of period result? Under compound symmetry, the answer is yes, if and only if the correlation coefficient for repeated measures on the same subject exceeds 0.50. In practice, this is nearly always the case, and as seen in Table V, it can be considerably higher. Ironically, the conclusion we made about not using the baseline values for the crossover study, does not apply to the completely randomized (one period) study. (B) If we again presume compound symmetry, and just use the end of period 1 minus baseline, but ignore period 2 data, we will exactly double variance compared to the methods we derived for equation (10), irrespective of the underlying correlation coefficient. In terms of patients, thanks to the second period, the sample size requirement for the crossover study becomes half that of the completely randomized study. The number of patients can be halved, but the number of measurements needed at two for one is the same for both designs.

Acknowledgments

This work was supported in part by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR001427. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health. Special thanks go to Dr. Stephen Senn, who provided important references and suggestions.

References

1. Senn, SJ. Cross-over Trials in Clinical Research. 2. Chichester: Wiley; 2002.
2. Jones, B., Kenward, MG. Design and Analysis of Cross-over Trials. 3. CRC Press; 2014.
3. Hills M, Armitage P. The two-period cross-over clinical trial. *Br J Clin Pharmacol*. 1979; 8:7–20. [PubMed: 552299]
4. Kenward MG, Roger JH. The use of baseline covariates in crossover studies. *Biostatistics*. 2010; 11(1):1–17. [PubMed: 19915170]
5. Chi EM. (1992). Analysis of cross-over trials when within-subject errors follow an AR(1) process. *Biometrical Journal*. 1992; 34(3):359–365.
6. Senn SJ. (1993). A comment on the analysis of cross-over trials when within-subject errors follow an AR(1) process. *Biometrical Journal*. 1993; 36:125–128.
7. Satterthwaite FW. An approximate distribution of estimates of variance components. *Biometrics Bulletin*. 1946; 2:110–114. [PubMed: 20287815]
8. Senn SJ. A century of t-tests. *Significance*. 2008; 5(1):37–39.
9. Shuster JJ. Student t-tests for potentially abnormal data. *Stat Med*. 2009 Jul 20; 28(16):2170–84. DOI: 10.1002/sim.3581 [PubMed: 19326398]
10. Fleiss JL. A critique of recent research on the two-treatment cross-over design. *Controlled Clinical Trials*. 1989; 10:237–243. [PubMed: 2676339]
11. Senn SJ. The AB/BA crossover: past, present and future? *Statistical Methods in Medical Research*. 1994; 3(4):303–324. [PubMed: 7889225]
12. Grizzle JE. The Two-Period Change-Over Design and Its Use in Clinical Trials. *Biometrics*. 1965; 21(2):467–480. Corrected *Biometrics* 1974; **30**: 727–727. [PubMed: 14338679]
13. Freeman PR. The performance of the two-stage analysis of two-treatment, two-period crossover trials. *Stat Med*. 1989 Dec; 8(12):1421–32. [PubMed: 2616932]
14. Shuster JJ. Diagnostics for assumptions in moderate to large simple clinical trials: do they really help? *Stat Med*. 2005 Aug 30; 24(16):2431–8. [PubMed: 15977289]

15. Zou B, Jin B, Koch GG, Zhou H, Borst SE, Menon S, Shuster JJ. On model selections for repeated measurement data in clinical studies. *Stat Med*. 2015 May 10; 34(10):1621–33. Epub 2015 Jan 23. DOI: 10.1002/sim.6414 [PubMed: 25645442]
16. Ho WK, Matthews JN, Henderson R, Farewell D, Rodgers LR. Dropouts in the AB/BA crossover design. *Stat Med*. 2012 Jul 20; 31(16):1675–87. Epub 2012 Feb 23. DOI: 10.1002/sim.4497 [PubMed: 22362410]

Table I

Target Population Means of Response in Each Group (Treatment Order) by Time Point

Period	Treatment Order=AB	Treatment Order=BA	Average
After first treatment	$E_{11} = \pi_1 - \tau$	$E_{12} = \pi_1 + \tau$	π_1
After second treatment	$E_{21} = \pi_2 + \tau + \lambda$	$E_{22} = \pi_2 - \tau - \lambda$	π_2
Difference (B-A)	$\pi_2 - \pi_1 + 2\tau + \lambda$	$\pi_1 - \pi_2 + 2\tau + \lambda$	$2\tau + \lambda$
Difference Period 2-Period 1	$\pi_2 - \pi_1 + 2\tau + \lambda$	$\pi_2 - \pi_1 - 2\tau - \lambda$	$\pi_2 - \pi_1$

Table II

Hypothetical Results

Period	Treatment Order=AB T=1	Treatment Order=BA T=0
After first treatment	X_{11}	X_{12}
After second treatment	X_{21}	X_{22}
B-A difference	$X_{21} - X_{11}$	$X_{12} - X_{22}$

Table III

Mathematical Expectations of Response in Each Group (Treatment Order) by Time Point

Period	Treatment Order=AB	Treatment Order=BA	Average
0:Baseline	$E_{01} = \mu$	$E_{02} = \mu$	μ
1:After first treatment	$E_{11} = \mu + \pi_1 - \tau$	$E_{12} = \mu + \pi_1 + \tau$	$\mu + \pi_1$
2:After washout	$E_{21} = \mu + \pi_2 - \theta$	$E_{22} = \mu + \pi_2 + \theta$	$\mu + \pi_2$
3:After second treatment	$E_{31} = \mu + \pi_3 + \tau + \lambda$	$E_{32} = \mu + \pi_3 - \tau - \lambda$	$\mu + \pi_3$
B-A diff after treatment	$E_{31} - E_{11} = \pi_3 - \pi_1 + 2\tau + \lambda$	$E_{12} - E_{32} = \pi_1 - \pi_3 + 2\tau + \lambda$	$2\tau + \lambda$
Difference of DifferencesB-A (#3-#2)-(#1-#0) Order AB (#1-#0)-(#3-#2) Order BA	$(E_{31} - E_{21}) - (E_{11} - E_{01}) = (\pi_3 - \pi_2 - \pi_1) + 2\tau + \lambda + \theta$	$(E_{12} - E_{02}) - (E_{32} - E_{22}) = (\pi_1 + \pi_2 - \pi_3) + 2\tau + \lambda + \theta$	$2\tau + \lambda + \theta$

Table IV

Estimated Means and Standard Deviations

	Order AB j=1 N=64	Order BA j=2 N=64
Baseline X_{0jk}	60.47(28.70)	59.06(29.37)
After first RX X_{1jk}	63.28(28.46)	61.72(24.34)
After Washout X_{2jk}	60.47(27.63)	62.19(23.47)
After second RX X_{3jk}	62.22(27.30)	64.53(23.70)
Delta= $U_{ik}=.5[(X_{3jk} - X_{1jk})]$	0.47(11.18)	1.41(6.93)
DelDel= $V_{ik}=.5[(X_{3jk} - X_{2jk}) - [(X_{1jk} - X_{0jk})]]$	0.47(13.33)	-0.16(11.27)

Table: V

Estimated Correlation Coefficients

Order	Variable\With	X_{01k}	X_{11k}	X_{21k}
		X_{02k}	X_{12k}	X_{22k}
Order AB	X_{11k}	0.743		
Order BA	X_{12k}	0.748		
Order AB	X_{21k}	0.762	0.729	
Order BA	X_{22k}	0.673	0.830	
Order AB	X_{31k}	0.729	0.679	0.824
Order BA	X_{32k}	0.675	0.834	0.821

Upper (Lower) lines are correlate with Upper (Lower) lines, respectively.