

Analysing Binary Outcome Data from a Crossover Design Study using the SAS® System.

Abstract:

The crossover design is frequently used in clinical trials due to the various advantages gained from each patient receiving more than one treatment and hence acting as their own control. Statistical methods are available for analysing binary outcome data from these trials. There are three common methods of analysing such data collected from a 2-period, 2-treatment crossover design study – McNemar's test, the Mainland-Gart test and Prescott's test. These three tests have their own advantages and disadvantages that will be considered along with the methods for performing these tests using the SAS® System. McNemar's test and the Mainland-Gart test can be performed using PROC FREQ in the SAS® System. However, the main focus will be on Prescott's test that takes into consideration those patients with the same binary outcome in both periods of the study. This test cannot be performed using a SAS procedure directly but by using an algorithm, code can be written to calculate 1-sided or 2-sided p-values. This algorithm, along with the code required to calculate the mean and variance of the test statistic allowing a confidence interval to be constructed, will be discussed and illustrated with a case study.

Introduction:

This paper looks at the three common methods of analysing binary outcome data from a 2-period, 2-treatment crossover design study. The crossover design is often used in clinical trials with one of its advantages being that a patient can receive more than one treatment or dosing regimen and hence act as their own control in the study. The simplest crossover design is the 2-period, 2-treatment (or 2x2) crossover design where one group of patients receive Drug A followed by Drug B while the other group receive Drug B followed by Drug A. The methods in this paper are discussed in the context of the 2x2 design.

While there are advantages in using a crossover design it is only suitable for chronic conditions otherwise the patient could be cured of the condition in the first period of the study. It should also be ensured that an adequate "wash-out" period is planned between the first and second periods to avoid a residual effect of the treatment from the first period influencing the results in the second period.

Binary outcome data analysed in a clinical trial may either be data collected during the trial that can only have a binary response or continuous data dichotomised to create a binary outcome. For example the question "Which treatment did you prefer?" would have a binary response (Drug A or Drug B). However blood pressure may be collected during the trial and recorded as a continuous variable but the endpoint to be analysed could be the binary outcome of whether or not there was a 10% or greater reduction from baseline.

If the patient had a positive outcome on Drug A but a negative outcome on Drug B they could be said to “Prefer Drug A”. Similarly the patient could be said to “Prefer Drug B” if a positive outcome was recorded on Drug B but a negative outcome recorded on Drug A. The idea of “preference” to a drug is frequently used in the literature on crossover trials but can apply to any type of binary outcome measurement and is not solely used for assessment of treatment preference. It is also possible to expand on this if the binary outcome is collected at the end of each period by the creation of a “No Preference” category. If a particular binary outcome is measured at the end of each period and a patient records the same response under Drug A and Drug B then they can be said to have no preference to either treatment since the same result was observed regardless of treatment.

Binary outcomes in the 2x2 design crossover study can be analysed by three common methods. These are McNemar’s test, the Mainland-Gart test and Prescott’s test. The tests have their various advantages that will be discussed along with the methods for performing the test within the SAS® System.

Methods:

McNemar’s test considers the basic question of whether more patients prefer Drug A or prefer Drug B. McNemar’s test does not take into account the order in which patients receive the two drugs nor does it consider patients who record the same response on both treatments. The test conditions on the total number of patients who prefer Drug A and calculates the probability of this number of patients or more being observed under the null hypothesis that the treatments are equal.

Mainland-Gart extends this test by taking into consideration the order in which the patients receive the two drugs. The test does not take into account those patients recording the same response on both treatments. The Mainland-Gart test is based on the number of patients who prefer each drug within each sequence group. This creates a 2x2 contingency table and the association between period in which the preferred drug is received and sequence can be analysed.

Prescott’s test is the only one of these three tests that takes into account patients in the “No Preference” group. It also considers the order in which the two treatments were received. The test is based on the resulting 2x3 contingency table and the probability of observing this or a more extreme result.

Figure 1 is taken from Stephen Senn’s “Cross-over Trials in Clinical Research”.

Figure 1

Sequence	Prefer First Period	No Preference	Prefer Second Period	Total for all Preferences
AB	9	3	0	12
BA	1	5	6	12
Total for both Sequences	10	8	6	24

The table is based on sequence and whether the patient prefers the treatment in Period 1 or Period 2 or has no preference.

McNemar's test considers the number of patients who prefer each treatment regardless of the order the treatments were received in and ignoring those patients who do not express a preference. The data from Figure 1 can be presented as those with and without a response on Drug A and those with and without a response on Drug B as shown in Figure 2:

Figure 2

Drug B	Drug A	Response	No Response
Response		4	1
No Response		15	4

The 15 patients who respond on Drug A but not on Drug B are the patients in Sequence AB who prefer treatment in the first period (nine patients) and the patients in Sequence BA who prefer treatment in the second period (six patients). The one patient who responds on Drug B but not on Drug A is the patient in Sequence BA who prefers treatment in the first period. The number of patients across both sequences with "No Preference" (eight patients) can be split into those with a response in both periods and those without a response in either period.

McNemar's test does not consider the eight patients with the same result on each drug but to calculate the p-value using SAS two variables (for Drug A and Drug B) each with two levels (response and no response) are required. McNemar's test can then be performed in PROC FREQ as shown below:

```
PROC FREQ DATA=indata;
  TABLES druga*drugb / AGREE;
RUN;
```

The Mainland-Gart test takes into account the order in which the treatments are received but ignores the patients who do not express a preference leading to the 2x2 contingency table shown in Figure 3:

Figure 3

Sequence	Prefer First Period	Prefer Second Period	Total for both Preferences
AB	9	0	9
BA	1	6	7
Total for both Sequences	10	6	16

The number of patients in Sequence AB who prefer Drug A is shown here as the number of patients who prefer treatment received in the first period (nine patients) and the number who prefer Drug B as the number who prefer treatment in the second period (no patients). The number of patients in Sequence BA preferring each treatment can also be determined from the table in a similar way. The reason for presenting the data by period rather than drug is that the Mainland-Gart test looks for an association between sequence and period. This test for association can be performed using a Fisher's Exact test or a large sample approximation such as the Chi-Squared test, both of which can be performed using PROC FREQ as the example code shows:

```
PROC FREQ DATA=indata;
  TABLES seqgrp*period / CHISQ MEASURES EXACT;
  EXACT OR;
RUN;
```

In situations where many patients record the same result on each treatment using either McNemar's test or the Mainland-Gart test will lead to the exclusion of these patients from the analysis. This could mean an analysis being performed on only a small proportion of patients in the study and possibly too few patients to draw valid conclusions from the results obtained. In this situation it would be advantageous to use Prescott's test to analyse the data instead.

Prescott's test makes use of all the data recorded in Figure 1 and the analysis is based on the probability of observing this contingency table or a more extreme one. When analysing continuous non-parametric crossover data the Wilcoxon Rank Sum test is used and this test assesses treatment effect by testing the period differences between the two sequence groups. In a similar way Prescott's test looks at the period differences within each sequence group to test treatment effect. The period differences are represented by the difference in number of patients preferring the treatment received in the first and second periods within each sequence. Figure 4 shows the data from Figure 1 presented again, including the notation used in Stephen Senn's "Cross-over Trials in Clinical Research":

Figure 4

Sequence	Prefer First Period	No Preference	Prefer Second Period	Total for all Preferences
AB	9 (d)	3 (e)	0 (f)	12 (r ₁)
BA	1 (g)	5 (h)	6 (k)	12 (r ₂)
Total for both Sequences	10 (s ₁)	8 (s ₂)	6 (s ₃)	24 (N)

The difference between periods in Sequence AB can be represented by (d-f) and the difference between periods in Sequence BA can be represented by (g-k). Therefore (d-f) – (g-k) can be said to represent the preferences for one treatment over the other. Since the row and column totals are known if (d-f) is known then (g-k) can be determined. The contingency tables that are equal to or more extreme than the observed table can be determined by considering all contingency tables where (d-f) equals or exceeds the observed (d-f). The p-value is then determined based on the probability of observing this set of contingency tables.

The probability of observing a contingency table, assuming that the null hypothesis is true, is calculated using combinations for each of the three preference groups in that table. The number of ways of selecting r from n is denoted nC_r and is calculated as:

$${}^nC_r = \frac{n!}{r! (n-r)!}$$

In SAS this can be calculated using the COMB function where nC_r can be obtained from COMB(n,r).

For any contingency table the combination for each preference category can be calculated using the column totals s₁, s₂ and s₃ and the numbers in Sequence AB (d, e and f). So ${}^{s_1}C_d$ represents the number of ways in which d can be chosen from s₁, ${}^{s_2}C_e$ represents the number of ways in which e can be chosen from s₂ and ${}^{s_3}C_f$ represents the number of ways in which f can be chosen from s₃. These three values are multiplied together to calculate the probability of observing that table, assuming that the null hypothesis is true.

This is repeated for each possible table then all the probabilities are summed together and divided by ${}^NC_{r1}$, the number of ways of selecting the number of patients in Sequence AB from the total number of patients, to give a one-sided p-value. A two-sided p-value is obtained by multiplying the one-sided p-value by two.

The expected value and variance of the test statistic (d-f) can be calculated using the cell values from the observed table in the following equations:

$$E(d-f) = r_1(s_1-s_3) / N$$

$$Var(d-f) = r_1r_2 \{ (s_1+s_3) - (s_1-s_3)^2/N \} / \{ N(N-1) \}$$

These values could be used if required to construct a confidence interval for the test statistic (d-f).

Although Prescott's test has advantages over the other methods of analysing binary outcome data discussed here the test cannot be performed directly from a SAS procedure. Instead a macro needs to be used to perform the algorithm that creates all of the contingency tables where (d-f) is equal to or more extreme than the observed (d-f). Using the COMB function in SAS the various probabilities can then be calculated and used to derive the p-value.

The macro requires an input dataset with three variables, one variable representing sequence, one variable representing preference and one variable representing the count of patients with each response.

An example of the dataset structure is shown in Figure 5:

Figure 5

Sequence	Preference	Response
1	1	X
1	0	X
1	2	X
2	1	X
2	0	X
2	2	X

Sequence will have two levels representing Sequence AB (1) and Sequence BA (2). Preference will have three levels representing preference for treatment received in first period (1), preference for treatment received in second period (2) and no preference (0). Response holds the counts of patients.

Using the earlier notation the first value in the dataset represents the cell marked d in the contingency table. The second value represents cell e in the contingency table and so on until the sixth value representing k. From this observed contingency table the row totals (r_1 and r_2), the column totals (s_1 , s_2 and s_3) and the total number of patients in the two sequence groups (N) can be calculated. These remain fixed throughout the calculations and the observed (d-f) value is also retained in a separate variable.

Starting with d and f values of zero (the lowest possible cell value) the macro uses the fixed row and column totals to derive e, g, h and k. This 'contingency table' is output to become a row in a dataset but only if it meets the definition of a valid table (all cells are non-negative) and if (d-f) is equal to or more extreme than the observed (d-f). The value of f is then incremented by one and the process is repeated. Once the loop has completed with the maximum value of f (which will be the minimum of the third column total s_3 or the first row total r_1) then f is reset to zero and d is incremented by one. This process is repeated until a dataset is formed whose rows represent the valid

contingency tables where (d-f) is equal to or more extreme than the observed (d-f).

The probabilities of observing each table are calculated in variables holding the value of ${}^{s1}C_d$, ${}^{s2}C_e$ and ${}^{s3}C_f$ obtained from the function COMB(n,r). This function is also used to derive a variable holding the value ${}^NC_{r1}$. The product of the probabilities is calculated for each table then the products are summed over all tables and the total divided by ${}^NC_{r1}$. This value is the one-sided p-value for the difference between Drug A and Drug B.

A two-sided p-value is also produced by the macro along with the expected value and variance of the test statistic (d-f) obtained from the formulae above. An option will allow the valid contingency tables used in the calculation to be printed for verification purposes.

Worked Example:

The following data were collected from migraine sufferers in a 2x2 design crossover study of an active compound versus a placebo. The endpoint was a 50% or greater reduction in the number of migraine attacks over a 6-week treatment period compared to the baseline number of attacks for that period. The “No Preference” group was defined as those who had a 50% or greater reduction in both periods or those who did not have a 50% or greater reduction in either period. The results are shown for the two sequence groups in Figure 6 along with the notation used in the macro.

Figure 6

Sequence	Prefer First Period	No Preference	Prefer Second Period	Total for all Preferences
Active – Placebo	4 (d)	21 (e)	1 (f)	26 (r ₁)
Placebo - Active	6 (g)	20 (h)	1 (k)	27 (r ₂)
Total for both Sequences	10 (s ₁)	41 (s ₂)	2 (s ₃)	53 (N)

The test statistic (d-f) for the observed table is 3 and the p-value for the treatment effect in reducing number of migraine attacks is based on the probability of observing this contingency table or a more extreme table. Therefore contingency tables meeting this criteria will have all cells populated with non-negative values and a value of (d-f) of 3 or more.

Using the Prescott test macro all possible contingency tables are examined starting with d=0 and f=0. From the fixed row and column totals the cell values e=26, g=10, h=15 and k=2 are obtained. Since this (d-f)=0 so does not exceed the observed (d-f) this table is not added to the dataset. The process continues with d=0 and f=1 but again (d-f) does not exceed the observed. The first case where (d-f) exceeds the observed occurs when d=3 and f=0 leading to a (d-f) value of 3. The remaining cell values are e=23, g=7, h=18

and $k=2$. The process continues until all such contingency tables are stored in the dataset.

The probability of observing each table is calculated in the macro using the $\text{COMB}(n,r)$ function to calculate each combination. The probability of observing the table where $d=3$, $e=23$, $f=0$, $g=7$, $h=18$ and $k=2$ is calculated as the product of $\text{COMB}(10,3)$, $\text{COMB}(41,23)$ and $\text{COMB}(2,0)$. The product of combinations is calculated for each row of the dataset then all such probabilities are summed and the total is divided by ${}^N\text{C}_{r1}$ calculated by $\text{COMB}(53,26)$. The one-sided p-value is therefore 0.8048. Using the formula the macro calculates the expected value for the test statistic (d-f) to be 3.92 and the variance to be 2.749 which can be used to calculate a confidence interval for the test statistic (d-f)=3 if required.

Conclusion:

In conclusion there are three common methods for analysing binary outcome data from a 2-period, 2-treatment crossover trial but only Prescott's test makes full use of all data collected by including those patients who record the same result on both treatments. By using McNemar's test or the Mainland-Gart test, power to detect a difference in treatments can be lost by ignoring the data from these patients. Although the disadvantage of Prescott's test is that it cannot be performed directly using a SAS procedure, an algorithm can be written in SAS that will produce all the valid contingency tables required to calculate a p-value for the treatment effect. This p-value is based on the probabilities of observing contingency tables with a test statistic (d-f) equal to or more extreme than the value observed. In addition the expected value and variance of this test statistic can be calculated allowing a confidence interval to be presented.

References:

- Senn, S.J. (1993) *Cross-over Trials in Clinical Research*, John Wiley & Sons, New York.
- Prescott, R.J. (1981) *The comparison of success rates in cross-over trials in the presence of an order effect*, Applied Statistics, Vol. 30, 9-15.
- Stokes, M.E., Davis, C.S., Koch, G.G. (1995) *Categorical Data Analysis Using the SAS System*, Cary, NC: SAS Institute.