

A Monte Carlo Evaluation of the Rank Transformation for Cross-Over Trials

George W. Pasdirtz, Hazleton Laboratories

Introduction

At Hazleton Laboratories, we have been investigating the use of a simple rank transformation (RT) to convert complicated, multivariate experimental designs, such as the cross-over trial (Ware, 1985; Jones and Kenward, 1989), into simpler univariate models. Our objective has been to determine whether hypothesis tests conducted within the distribution-free RT framework can be used in order to sidestep the assumptions of independently and identically distributed (iid) observations, homogeneity of variance, or multivariate normality. The availability of simple, flexible, and powerful distribution-free techniques would greatly increase the quality and timeliness of applied statistical analysis.

To evaluate the RT procedure, a limited number of Monte Carlo simulations were conducted using SAS*. In each simulation, parameters known to cause bias and error in the cross-over model were systematically varied. Results were then compared with or without two types of rank transformations: the RT-1 transformation, which ranks all continuous variables without regard for levels of treatment or control variables, and the *aligned rank* transformation (an RT-3 procedure), which eliminates nontreatment-related nuisance parameters prior to ranking.

The RT procedure was popularized by Conover and Iman (1981, 1982) and is widely used when normal-theory assumptions fail. The transformation is easy to construct with PROC RANK and can be used with any SAS/STAT procedure. The RT procedure offers the hope of extending distribution-free techniques beyond the restricted choices available in NPAR1WAY.

Recent Monte Carlo studies (Blair, Sawilowsky, and Higgins, 1987; Sawilowsky, Blair, and Higgins, 1989), however, have suggested that the RT-1 transformation might not be appropriate for the two-way Analysis of Variance (ANOVA). Because the cross-over model can be expressed as a two-way or higher-way ANOVA, there may be reasons for concern. Although Puri and Sen (1985) have suggested that the RT-3 procedure might be used instead, Conover and Iman (1981) have noted that some power might be lost in the process.

Cross-Over Model

In a multivariate framework, *within-subject* or *time-varying* covariates are the characteristics of the patient (respondent) that vary over time. For example, aging, behavioral status, and environmental variables all might

change over the duration of a study. On the other hand, *between-subject* covariates take a single value for the entire period of observation. Gender and race are examples of common classification variables that remain constant.

If different treatments are applied to the same patient over time, the design is defined as a cross-over trial. When the within-subject effect is simply elapsed time, a repeated measures analysis in PROC GLM is still appropriate. If the within-subjects design includes more than just elapsed time, a univariate regression with correlated errors must be used. One technique for dealing with correlated errors is the mixed-model ANOVA where subject effects are nested within sequence (Milliken and Johnson, 1984, Chapter 32). Both the repeated-measures and the mixed-model ANOVA assume compound symmetry (equal correlations and equal variances) among all within-subject covariates. The assumption is very restrictive.

Basically, the *cross-over trial* is a repeated measures design in which subjects receive different treatments during the different time periods. The advantages of a cross-over trial are that fewer patients may be needed and each patient can act as their own control. However, residual carry-over effects between periods may become confounded with treatment effects.

The *carry-over effect* is the impact of the previous treatment on the current period response. Carry-over effects can result from treatments that have not been entirely eliminated, or physiological and psychological changes in the patients over time.

A *sequence* is created taking permutations of the period and treatment levels. In a four-period, two-treatment cross-over, for example, the four permutations of treatment and period are:

$$\left\{ \begin{array}{l} \text{ABAB} \\ \text{BABA} \\ \text{ABBA} \\ \text{BAAB} \end{array} \right\}$$

When all permutations are present, the design is *balanced* and specialized multivariate techniques are available (Ware, 1985). If all the respondents are measured at all time points, the design is *balanced on time*. If all subjects and all groups are measured at all time points, the design is *completely balanced*.

Whether balanced or not, measurements related to different treatments are obtained for each subject and treatments are

compared *within subjects*. The goal is to remove, from the treatment and period comparisons, any component related to consistent individual differences. Carry-over effects can be minimized by wash-out periods during which no treatment is given. However, to determine whether the wash-out has been complete, its effect must be estimated. Treatment-by-period interactions can also exist if responsiveness to treatment changes over time. For example, treatments applied at the end of a clinical study might be more effective because the patients are in better health.

The cross-over model can be written symbolically as:

$$Y_{ijk} = \mu + s_k + \pi_j + \tau_{d(i,j)} + \lambda_{d(i,j-1)} + e_{ijk}$$

μ = a general mean

s_k = the effect of subject k in group i

π_j = the effect of period j

$\tau_{d(i,j)}$ = the direct effect of treatment

$\lambda_{d(i,j-1)}$ = the effect of carry-over from treatment $j-1$

$$\lambda_{d(i,0)} = 0$$

e_{ijk} = random error

The size of the carry-over effect is seldom of direct interest because the treatment sequences have been artificially constructed purely for the sake of experimentation. Likewise, subject effects are estimated only to control individual differences and not because there is any population parameter of interest. In most cases, the important question involves the direct effect of treatment. The other parameters, however, must be identified as sources of error and eliminated.

Assumptions

It would be tempting to apply Ordinary Least Squares (OLS) or the mixed-model ANOVA to the estimation of parameters in a cross-over model. Although the model looks deceptively simple, the assumptions are strong. Repeated measurements, subject effects, and non-normal clinical data often make OLS inappropriate.

- **Model specification.** For classification variables, the group average rather than the median or mode is assumed to be the correct measure of central tendency. For continuous effects, the relationship must be linear. The model must also include variables that take time effects into consideration.
- **Measurement and sampling of the X variables.** For the cross-over model, group assignment is always fixed under sampling. Some covariates, however, can be *random* (i.e., not under

experimental control). Some direct effects can also be random (e.g., subjects). Random-effects models can be estimated using PROC VARCOMP with some restrictions (e.g., only categorical covariates can be included). More general approaches can be coded in PROC IML (Laird and Ware, 1982; Chinchilli and Elswick, 1989).

- **Homogeneity of variance.** This assumption can be violated in two ways. First, the experimental groups might not respond uniformly to treatment. For example, patients in high-dose groups might respond more variably to treatment than those in low-dose groups. Second, variability might change over time as patients respond to treatment. A test developed by Levene (1960) and generalized by Draper and Hunter (1969) can be constructed using the absolute values of the regression residuals (i.e., the difference between predicted and actual values) to test for homogeneity of variance.
- **Correlated errors.** When the same subject is repeatedly observed for a specific indicator, a correlation pattern could be introduced. For example, body weight tends to fluctuate slowly around a set point. The Durbin-Watson statistic reported in both PROC REG and PROC AUTOREG can be used to determine serial correlation in models without lagged variables. When lagged independent variables exist, an alternative test can be constructed using the OLS residuals (Durbin, 1970).
- **Normality.** The familiar bell-shaped curve might not always be applicable. For example, the distribution of clinical chemistry data tends to be skewed. Or, the sample might actually have been drawn from two normal distributions with different parameters. A test developed by Shapiro and Wilk (1965) can be applied to the residuals using PROC UNIVARIATE.

To meet some of these assumptions, a repeated measures analysis in PROC GLM can be attempted.

$$Y_{[i,j,k]}, Y_{[i,j+1,k]}, \dots, Y_{[i,n,k]} = \mu_k + \tau_k + e_{ijk}$$

However, the estimating equation for the multivariate setup contains no carry-over, subject, or period parameters, and would be acceptable only when these effects are known to be zero. When cross-over parameters must be identified and eliminated, or when other assumptions fail, some transformation will be necessary to conduct hypothesis testing.

Transformations

There are three commonly used approaches to transformation: the search for an appropriate arithmetic function, a rank transformation, or a Generalized Least Squares (GLS) estimator.

- **Arithmetic transformations.** Draper and Hunter (1969) have suggested using a series of transformations (log, square root, arcsin, etc.) until either the normal-theory assumptions are met or the procedure fails. In practice, the search can be time consuming and can fail either because no transformation is possible, or because of outliers (extreme values) in the data set (Thakur, Trutter, and Korte, 1983). Even if a transformation can be found, it might not have a scientifically acceptable explanation.

- **Rank transformations (RT).** The RT approach has become popular based on the work of Conover and Iman (Conover, 1980; Conover and Iman, 1981; Iman, 1988; Iman and Conover, 1989). Carrying out an RT procedure involves replacing the original observations by the ranks and then calculating the usual parametric statistics.

- **The GLS model.** Although often not presented as such, Aitken's Generalized Least Squares (GLS) model can be viewed as a transformation.

$$P'y = P'Xb + P'e$$

P = an $n \times n$ transformation matrix

y = an $n \times 1$ vector of dependent variables

X = an $n \times m$ matrix of independent variables

b = an $m \times 1$ vector of unknown coefficients

e = an $n \times 1$ vector of unknown errors

$$\hat{b} = (X'PPX)^{-1}(X'Py)$$

For hypothesis testing, the transformation matrix must be chosen to create iid errors. One approach is to construct the transformations from the OLS residuals.

$$e = y - X\hat{b}$$

Functions of the residuals can be computed either directly or through iteration using an Expectation-Maximization (E-M) or a Maximum Likelihood (ML) algorithm (Ware, 1985).

The autoregressive (AR) model can be viewed as a GLS transformation. The subject-by-period interaction is modelled partly as a response to initial conditions, which may be changing as a result of the study, and partly as a function of remaining, unexplained trends in the residuals. The AR model can be used to remove the effects of aging, development, and fatigue, etc. The symbolic representation of the AR model is:

$$Y_{ijk} = \mu + \rho_2 y_{i,j-1,k} + s_k + \pi_j + \tau_{d(i,j)} + \lambda_{d(i,j-1)} + e_{ijk}$$

$$e_{ijk} = \rho_1 e_{i,j-1,k} + \omega_{ijk}$$

ω_{ijk} = iid random variable, mean zero

ρ_1 = first - order autoregression parameter

ρ_2 = autoregression on the previous response

The first-order autoregression equation is used to remove trend components from the residuals. Note that the regression coefficients in the AR model are interpreted as conditional on the respondent's initial position, as opposed to the standard carry-over model where the coefficients are interpreted unconditionally (Ware, 1985). When using PROC AUTOREG with lagged dependent variables, an instrumental variable regression must first be used to remove error components from the lagged values (Johnston, 1972: 319).

The AR model can be used to determine optimum treatment regimes for cross-over trials (Jones and Kenward, 1989: Chapter 7). When positive first-order autocorrelation is observed, treatments should be changed frequently. When negative first-order autocorrelation is observed, treatments should be changed infrequently.

Error and Bias in Estimators

Typically, RT procedures have been evaluated in terms of their impact on statistical decision making i.e., conservativeness of p -values, using up to 5,000 Monte Carlo replications. A similar question, which requires far fewer simulations (Efron, 1967), involves the statistical error of the estimator.

Bias refers to the deviation of the expected value of an estimator from the true value.

$$\text{Bias}(S) = E[S] - \mu$$

Mean squared error (MSE) is the expected value of the square of the deviation of the estimator from the true value.

$$\text{MSE}(S) = \sigma^2 + [\text{Bias}(S)]^2$$

Relative efficiency is the ratio of MSEs between two similar methods. Estimators should have a low bias and a minimum MSE.

Finally, a *consistent* estimator concentrates on the population value as sample size increases.

$$S \text{ is consistent if } \text{MSE}(S) \rightarrow 0$$

$$\text{as } n \rightarrow \infty$$

It has been shown analytically that autocorrelation and lagged dependent variables can produce large biases in

parameter estimates (Griliches, 1961). When all the fixed effects are zero, the cross-over model can be simplified through substitution.

$$Y_{ijk} = (\rho_1 + \rho_2)y_{i,j-1,k} - (\rho_1\rho_2)y_{i,j-2,k} + \omega_{ijk}$$

The second term indicates the presence of an *omitted variable* (Griliches, 1961) that, depending on the signs and magnitudes of the autocorrelation parameters, would introduce bias.

$$\text{Bias} \rightarrow (1 - \rho_2)$$

$$\text{as } \rho_1 \rightarrow 1$$

As the autocorrelation parameter approaches unity, bias approaches a fixed value. Autocorrelated disturbances without lagged y-values do not produce biased estimators, even in small samples, since the multiplicative terms cancel. Lagged y-values with random disturbances will give OLS estimators that are inconsistent and biased in finite samples.

It might seem unusual to be concerned about expected values for ranked data because the estimated rank is not a population parameter and would not seem worth estimating. Iman and Conover (1989), however, show how to map rank estimates back into sample values through linear interpolation. The *monotone regression* procedure allows a population estimate to be implied.

Monte Carlo Simulation

Large-scale simulation studies (Blair, Sawilowsky, and Higgins, 1987; Sawilowsky, Blair, and Higgins, 1989) have demonstrated some problems with the RT procedure. The presence of significant interactions and main effects in the two-way ANOVA model can seriously inflate Type I error rates when using ranks. This result should not be surprising because the RT is a nonlinear transformation and the interaction effects are sensitive to nonlinearity.

The simulations conducted by Blair, Sawilowsky, and Higgins were based on the following equations and parameter values:

$$y_{ijk} = v + a_i + b_j + (ab)_{ij} + e_{ijk}$$

$$a_1 = c$$

$$a_2 = -c$$

$$c = \begin{cases} 0.25\sigma \\ 0.50\sigma \\ 0.75\sigma \\ 1.00\sigma \end{cases}$$

Deviates were sampled from normal, uniform, t ($df = 3$), exponential, and mixed normal distributions. As the sample size approached moderate sizes (the maximum used being 50), the probability of making a Type I error approached unity when using ranks.

Puri and Sen (1985) suggest that the problems can be alleviated by classifying all but one of the main effects as *nuisance parameters* before ranking.

$$R[y_{ijk} - v - b_j - (ab)_{ij}] = a_i + e_{ijk}^*$$

For the standard cross-over model, the equation would be

$$R[Y_{ijk} - \mu - s_{ik} - \pi_j - \lambda_{d[i,j-1]}] = \tau_{d(i,j)} + e_{ijk}^*$$

with an added lagged dependent variable for the AR version.

$$R[Y_{ijk} - \mu - \rho_2 y_{i,j-1,k} - s_{ik} - \pi_j - \lambda_{d[i,j-1]}] = \tau_{d(i,j)} + e_{ijk}^*$$

The *aligned-rank transformation* (RT-3) is a three-step process.

- Estimate the nuisance parameters.
- Rank the resulting residuals.
- Estimate the one-way treatment effect.

If there are multiple parameters of interest, a Bonferroni correction should be used to adjust the probabilities (Miller, 1981).

The bias and MSE of the RT-1 vs. the RT-3 procedures can be computed by drawing a random sample, with replacement, using the probability distribution functions built into the SAS data step. The process is repeated many times to develop the *sampling distribution* of the estimators from which the "population" mean and variance can be computed.

The actual simulation was conducted using a modified SAS macro written by Carson (1985). The original macro was written to implement a computer-intensive technique called the bootstrap (Efron, 1979, 1982) that can be used to produce *synthetic* population estimates from actual data sets. The bootstrap also produces measures of bias, error, and relative efficiency. The macro was modified to generate Monte Carlo estimates but has the virtue of also being usable on actual data.

The *population* parameter values for the current simulation were chosen based on the Blair, Sawilowsky, and Higgins simulations. The values were limited to conditions where problems were observed. Parameters were chosen to create high-bias and low-bias conditions. Bias and MSE for OLS, AR, RT-1, and RT-3 cross-over models were then compared in small and moderate samples.

The high-bias conditions simulated significant autocorrelation, significant cross-over and treatment effects, and significant nonnormality. (The mixed normal distribution was formed by sampling with probability 0.95, from a normal distribution with mean 0, and standard deviation 80, and with probability 0.05 from a normal distribution with mean shifted 10 standard deviation units, and a standard deviation equal to ten times 80.) The remaining parameters were taken to simulate area under the plasma time curve (AUC) data from a bioequivalence study.

$$\text{high - bias} = \left\{ \begin{array}{l} \rho_2 = (-0.1) \\ \rho_1 = (-0.8) \\ \mu = 687.55 \\ s_e = 0 \\ \pi = 0.0\sigma \\ \tau = 1.0\sigma \\ \lambda = 1.0\sigma \\ \sigma = 80 \\ e = U(p < 0.95)N(0, \sigma) \Leftrightarrow \\ \quad U(p \geq 0.05)N(10\sigma, 10\sigma) \\ y_0 = 350 + N(0, 200) \\ N = 60 \end{array} \right.$$

$$\text{low - bias} = \left\{ \begin{array}{l} \rho_2 = (-0.1) \\ \rho_1 = 0.0 \\ \mu = 687.55 \\ s_e = 0 \\ \pi = 0.0\sigma \\ \tau = 1.0\sigma \\ \lambda = 0.0\sigma \\ \sigma = 80 \\ e = N(0, \sigma) \\ y_0 = 350 + N(0, 200) \\ N = 60 \end{array} \right.$$

For the low-bias simulations, all autocorrelation parameters and non-treatment related parameters were set to zero. Values of 60 and 36 were chosen to represent moderate and small sample sizes. For all simulations, 1000 Monte Carlo replications were conducted.

Results and Summary

The "population" estimates for treatment effect and their standard deviations are presented in Appendix I.

- OLS or GLS estimators provided acceptable location estimates. A possible exception was when using the AR model with small sample sizes and high bias. Scale estimates (standard errors) were badly inflated under highly biased conditions in all parametric models.
- When using the RT techniques, the location parameters were not stable in the face of bias, but the scale parameters were very stable. There seems to be little difference between the RT-1 and RT-3 procedures.

As a result of the limited simulations conducted for this paper, the RT procedure detected significant treatment effects when they existed even in the face of bias, while OLS and GLS techniques did not. For example, the constructed *t*-values for OLS and GLS estimators under high-bias would not be significant when in fact the simulation was designed to generate a significant treatment effect.

The results appear to support Conover's (1981) position, at least for the cross-over trial without higher-order interactions. The failure to confirm the Blair, Sawilowsky, and Higgins findings might be due to their focus on the *F*-statistic and Type I error rates rather than on estimators, standard errors, and bias.

An interesting extension of this work would be to add the computer-intensive bootstrap estimates (Efron, 1967, 1982; Carson, 1985) to compare against the simpler RT procedures.

References

- Carson, R. T. (1985), "SAS® Macros for Bootstrapping and Cross-Validating Regression Equations," *SUGI 10 Proceedings*, 1064-1069.
- Chinchilli, V. M. and R. K. Elswick (1989), "Multivariate Models for the Analysis of Crossover Experiments," *SUGI 14 Proceedings*, 1267-1271.
- Conover, W. J. (1980), *Practical Nonparametric Statistics*, New York: Wiley.
- Conover, W. J. and R. L. Iman (1981), "Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics," *The American Statistician*, 35, 124-133.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B* 39, 1-38.

Draper, N. R. and W. G. Hunter (1969), "Transformations: Some Examples Revisited," *Technometrics*, 11, 23-40.

Durbin, J. (1970), "Testing for Serial Correlation in Least-Squares Regression When Some of the Regressors are Lagged Dependent Variables," *Econometrica*, 38, 410-421.

Efron, B. (1967), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1-26.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*, Philadelphia: SIAM.

Griliches, Z. (1961), "A Note on Serial Correlation Bias in Estimates of Distributed Lags," *Econometrica*, 29, 65-73.

Iman, R. L. (1988), "The Analysis of Complete Blocks Using Methods Based on Ranks," *SUGI 13 Proceedings*, 970-978.

Iman, R. L. and W. J. Conover (1989), "Monotone Regression Utilizing Ranks," *SUGI 14 Proceedings*, 1310-1311.

Johnston, J. (1972), *Econometric Methods*, Second Edition, New York: McGraw-Hill.

Jones, B. and M. G. Kenward (1989), *Design and Analysis of Cross-Over Trials*, London: Chapman and Hall.

Laird, N. M. and J. H. Ware (1982), "Random-Effects Models for Longitudinal Data," *Biometrics*, 38, 963-974.

Levene, H. (1960), "Robust Tests for Equality of Variances," in *Contributions to Probability and Statistics*, (eds.) I. Olkin et al., Ch. 25, Stanford, CA: Stanford University Press, 278-292.

Miller, R. G. (1981), *Simultaneous Statistical Inference*, Second Edition, New York: Springer-Verlag.

Milliken, G. A. and D. E. Johnson (1984), *Analysis of Messy Data*, New York: Van Nostrand Reinhold.

Puri, M. L. and P. K. Sen (1985), *Nonparametric Methods in General Linear Models*, New York: John Wiley.

Shapiro, S. S. and M. B. Wilk (1965), "An analysis of variance test for normality (complete samples)," *Biometrika*, 52, 591-611.

Thakur, A. K., J. Trutter, and D. Korte (1983), "Classical Parametric (P) vs. Nonparametric (NP) Significance Testing in Toxicity Studies," *The Toxicologist*, 3.

Ware, J. H., (1985), "Linear Models for the Analysis of Longitudinal Studies," *The American Statistician*, 39, 95-101.

*SAS® is a registered trademark of the SAS Institute, Inc., Cary, NC, USA.

Author

George W. Pasdirtz, PhD
Hazleton Laboratories
3301 Kinsman Boulevard
Madison, WI 53704

Appendix 1. Monte Carlo Results.

Low-Bias	N=36		N=60	
	Parameter	Std. Error	Parameter	Std. Error
RT-3	9.334	2.813	15.216	3.630
RT-1	9.380	3.270	15.689	3.987
AR***	81.261	28.951	79.810	21.956
OLS**	79.678	25.785	80.725	20.284
RT-1*	7.873	1.035	13.043	1.376
High-Bias	N=36		N=60	
	Parameter	Std. Error	Parameter	Std. Error
RT-3	7.127	2.657	12.153	3.437
RT-1	7.594	2.507	12.551	3.107
AR	73.488	107.307	79.221	51.770
OLS**	78.983	70.441	79.231	52.486
OLS	1.236e+6	5.887e+5	1.232e+6	4.533e+5
*** Equivalent to lagged OLS with no autocorrelation			* Treatment contrast [-1,1]	
** Autoregression on previous response set to zero				