Randomization Inference and Bias of Standard Errors

Gary L. GADBURY

A nonparametric statistics course typically includes material regarding "distribution free" randomization based inference. However, the accuracy of reported p values and confidence intervals often relies on an unverifiable assumption of unit(subject)-treatment additivity. This assumption is not always explicitly stated in texts and, when the assumption does not hold, the implications on inference are seldom discussed. The focus of this article is the bias of standard errors of estimated mean treatment effects in the presence of nonadditivity. This bias is characterized and interpreted for a usual estimator of standard error in three common experimental designs: a two-sample completely randomized design, a matched-pairs design, and a balanced, two-period cross-over design. Even in the presence of nonadditivity, useful conservative estimates of a mean treatment effect can be obtained. This is illustrated using some previously published data.

KEY WORDS: Additivity; Experiments; Nonparametric; Permutation; Potential response.

1. INTRODUCTION

A typical nonparametric statistics course may include material on randomization-based inference. This type of inference involves permutation of observed responses under a specified null hypothesis. For instance, if the null hypothesis stated that there are no differences among treatments, then an observed response of a subject is fixed regardless of what treatment the subject would have received. More formally, suppose Y_i^j is a potential response (Rubin 1974) of the ith subject to the jth treatment, $i = 1, \ldots, N$ and $j = 1, \ldots, T$. Suppose the ith subject receives treatment t and that $Y_i^t = y_i$ is observed. The null hypothesis of no treatment effect would imply that $Y_i^j = y_i$ for all $j = 1, \ldots, T$. This is a condition of subject-treatment additivity referred to by Cox (1992) and Hinkelmann and Kempthorne (1994) and referred to as a constant effect by Holland (1986). Under this "additive model," p values are exact with the proviso that permutations of observed responses for all possible treatment assignment outcomes are computationally feasible. Otherwise, Monte Carlo techniques or statistical models can be used to approximate the randomization distribution. Rubin (1991) referred to this type of test as a randomization test of a sharp null hypothesis.

Students in my nonparametrics course, despite having been previously exposed to randomization tests, are somewhat surprised by the restrictive additivity assumption and puzzled over its practicality. Discussions often ensue and a need for further exposition of the subject becomes apparent. Results in this article convey added information regarding the associated implications on randomization inference when the additivity assumption is relaxed. Computing the randomization distribution of the estimator may not be possible when nonadditivity is present, but a single estimate and an estimated standard error can be computed. In such cases there remain two potential advantages of pursuing randomization based estimates of a treatment effect. First, inference results apply to the fixed set of subjects in the study so a random sample assumption from a larger population is not necessary. Rosenbaum (1995) made use of this fact with sensitivity analysis to analyze observational data from a fixed set of subjects in a study, and Lachin (1988) noted that a larger population model can only be invoked in clinical trials as an untestable assumption. Second, in the three designs considered here, estimated standard errors are positively biased and, hence, they tend to be conservative. So inferences that use the estimated standard error may also be conservative. The presence of bias in estimated standard errors was observed by Neyman (1935), but the concept seems notably absent from the current teaching literature on nonparametric inference, though it sometimes appears in material regarding design of experiments (e.g., Hinkelmann and Kempthorne 1994).

In the next section, an estimator for standard error and its bias are given in the context of a two-sample completely randomized design. The bias cannot be estimated because it involves a nonidentifiable subject-treatment interaction term discussed by Gadbury and Iyer (2000). In the following two sections the same is done for a matched-pairs design and a two-period balanced cross-over design. Then, an illustrative example is shown and some conclusions given. In all of the designs, two treatments are being compared, and the average treatment effect is the parameter of interest. Throughout it is assumed that there is no interference between subjects (Cox 1958), a condition generalized by Rubin's (1980) stable unit-treatment value assumption (SUTVA). These assumptions state that a particular subject's response to a treatment t will be the same regardless of what treatments other subjects receive and regardless of whether there may be different versions of treatment t (such as two or more different manufacturers of the same drug). If SUTVA is violated, such as might occur when adjacent plots in an agricultural experiment receive different treatments, randomization-based inference becomes considerably more complicated.

2. A TWO-SAMPLE COMPLETELY RANDOMIZED DESIGN

A test treatment t is being compared with a control treatment c. For notational clarity, suppose that if a subject receives treatment t the response would be X, and if a subject receives treatment c, the response would be Y. Potential responses for N subjects

Gary L. Gadbury is Assistant Professor, Department of Mathematics and Statistics, University of Missouri–Rolla, Rolla, MO 65409 (E-mail: gadburyg@umr.edu). The author thanks Hari Iyer for helpful advice on this research and the editor, an associate editor, and reviewer for helpful clarifications to the article's content.

are given by,

$$(X, Y) = \begin{pmatrix} X_1 & Y_1 \\ \vdots & \vdots \\ X_N & Y_N \end{pmatrix}. \quad (1)$$

Even though only one of the two responses can be observed at a given instant on any given subject, conceptually the "true" treatment effect for the ith subject is defined as $D_i = X_i - Y_i$. The vector of true treatment effects (length N) is denoted by $D = X - Y$. The quantity $\overline{D} = \overline{X} - \overline{Y}$ is the parameter of interest, where $\overline{X} = (1/N)\sum_{i=1}^{N} X_i$ and $\overline{Y} = (1/N)\sum_{i=1}^{N} Y_i$. Assume that $N = 2n$ and n subjects will be randomly selected to receive treatment t with the other n subjects receiving treatment c. Results can be easily generalized to the case of unequal numbers in each group. There are $k = \begin{pmatrix} 2n \\ n \end{pmatrix}$ possible treatment assignments, each occurring with probability $1/k$. The estimated treatment effect is

$$\overline{d} = \frac{1}{n}\sum_{i=1}^{2n} X_i\delta_i - \frac{1}{n}\sum_{i=1}^{2n} Y_i(1-\delta_i) = \overline{x} - \overline{y},$$

where $\delta_i = 1$ indicates subject i received treatment t. That is, the δ_i are the random components in $\overline{d}$. There are k possible values for $\overline{d}$ in the randomization distribution, and it can be shown that $E(\overline{d}) = \overline{D}$ where the expectation, $E(\cdot)$, is with respect to this distribution. The common pooled estimator for the variance of $\overline{d}$ is,

$$\widehat{\text{var}}(\overline{d}) = \frac{2}{n}\frac{\sum_{i=1}^{2n}(X_i^2\delta_i - \overline{x}^2) + \sum_{i=1}^{2n}\{Y_i^2(1-\delta_i) - \overline{y}^2\}}{2n-2}.$$

The bias of this estimator is given by,

$$\begin{aligned} \text{bias} &= E[\widehat{\text{var}}(\overline{d})] - \text{var}(\overline{d}) \\ &= \frac{1}{2n-1}\text{var}(X-Y), \end{aligned}$$

where $\text{var}(X-Y) = \text{var}(D) = (1/2n)\sum_{i=1}^{2n}(D_i - \overline{D})^2$. This was a fact observed by Neyman (1935), and the derivation can be a useful exercise for students. The important points are:

- Bias is always greater than or equal to zero.
- Under an additive model ($D =$ a constant vector), bias is zero.
- For a fixed value of $\text{var}(D)$, the bias decreases as the study size increases.

3. MATCHED-PAIRS DESIGN

The population is still of size $2n$ but subjects have now been matched into n pairs prior to random treatment assignment. There are $k = 2^n$ possible treatment assignment outcomes, each occurring with probability $1/k$. The potential responses are given by

$$\begin{pmatrix} X_1 - \epsilon_1 & Y_1 - \eta_1 \\ X_1 + \epsilon_1 & Y_1 + \eta_1 \\ - - - - & - - - - \\ \vdots & \vdots \\ - - - - & - - - - \\ X_n - \epsilon_n & Y_n - \eta_n \\ X_n + \epsilon_n & Y_n + \eta_n \end{pmatrix}, \quad (2)$$

where the parameters ϵ_i and η_i, $i = 1, \ldots, n$ represent a lack of homogeneity within the ith pair, and X_i, Y_i represent the average responses to the treatment and control, respectively, within the ith pair. With this notation, the treatment effects for the two units in the ith pair are $D_{i1} = X_i - Y_i - (\epsilon_i - \eta_i)$ and $D_{i2} = X_i - Y_i + (\epsilon_i - \eta_i)$. The parameter of interest is $\overline{D} = (1/2n)\sum_{i=1}^{n}\sum_{j=1}^{2} D_{ij} = (1/n)\sum_{i=1}^{n}(X_i - Y_i) = \overline{X} - \overline{Y}$. Let

$$T_i = \begin{cases} 1 & \text{if} \quad \begin{pmatrix} T_{i1} \\ T_{i2} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ 0 & \text{if} \quad \begin{pmatrix} T_{i1} \\ T_{i2} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \end{cases}$$

be an indicator random variable representing the treatment assignment for the ith pair where $P(T_i = 1) = 1/2$ for all $i = 1, \ldots, n$, and $T_{ij} = 1$ implies that subject j ($j = 1, 2$) in pair i receives treatment t. Write the observed treatment effect for the ith pair as

$$d_i = [X_i - Y_i - (\epsilon_i + \eta_i)]T_i + [X_i - Y_i + (\epsilon_i + \eta_i)](1 - T_i).$$

Then, $\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$ is the observed average treatment effect. Again, $E(\overline{d}) = \overline{D}$, where $E(\cdot)$ is with respect to the matched-pairs randomization distribution. It can be shown that the true variance of $\overline{d}$ is

$$\text{var}(\overline{d}) = \frac{1}{n^2}\sum_{i=1}^{n}(\epsilon_i + \eta_i)^2$$

so $\text{var}(\overline{d})$ depends on the effectiveness of the matching criteria (also noted by Copas 1973). A common estimator of the variance of $\overline{d}$ is

$$\widehat{\text{var}}(\overline{d}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(d_i - \overline{d})^2.$$

The bias of this estimated variance is

$$\text{bias} = \frac{1}{n-1}\text{var}(X-Y),$$

where $\text{var}(X-Y) = (1/n)\{\sum_{i=1}^{n}(X_i - Y_i - \overline{D})^2\}$. This indicates the following for matched pairs designs,

- The estimator for standard error is positively biased.
- It will be unbiased if and only if the mean treatment effect for each pair ($X_i - Y_i$) is the same for every pair (for every $i = 1, \ldots, n$). That is, $X - Y$ is a constant vector.
- For a finite value of $\text{var}(X-Y)$, the bias of the standard error will become smaller as the number of pairs in the study increases.
- A final note is that if $\epsilon_i = \eta_i = 0$ for all i (an untestable condition), then $\overline{d} = \overline{D}$.

4. A TWO PERIOD CROSS-OVER DESIGN

The potential responses of $2n$ subjects have the form shown below.

Subject	Time period 1		Time period 2	
1	$X_1 - t_1$	$Y_1 - \tau_1$	$X_1 + t_1$	$Y_1 + \tau_1$
⋮	⋮	⋮	⋮	⋮
2n	$X_{2n} - t_{2n}$	$Y_{2n} - \tau_{2n}$	$X_{2n} + t_{2n}$	$Y_{2n} + \tau_{2n}$

A true individual treatment effect for the ith subject is defined to be the average of the two treatment effects over the two time periods. That is,

$$D_i = \frac{1}{2}[(X_i - t_i) - (Y_i - \tau_i) + (X_i + t_i) - (Y_i + \tau_i)] = (X_i - Y_i),$$

where t_i and τ_i, $i = 1, \ldots, 2n$, are individual time effect parameters. The parameter of interest is then $\overline{D} = \overline{X} - \overline{Y}$. It is assumed that there are no carry-over effects from one period to the next.

Let T_i be a treatment indicator variable. If $T_i = 1$, then subject i receives the test treatment in time period 1 and the control treatment in time period 2. If $T_i = 0$, then subject i receives the two treatments in reverse order. Suppose that n of the $2n$ subjects are assigned to each treatment sequence, so $P(T_i = 1) = 1/2$ for each $i = 1, \ldots, 2n$. The observed treatment effect for subject i can be written as,

$$d_i = [(X_i - t_i) - (Y_i + \tau_i)]T_i + [(X_i + t_i) - (Y_i - \tau_i)](1 - T_i).$$

An estimator for the mean treatment effect is $\overline{d} = (1/2n)\sum_{i=1}^{2n} d_i$, and $E(\overline{d}) = \overline{D}$, where $E(\cdot)$ is with respect to the randomization distribution. The estimator $\overline{d}$ is the one used by Grizzle (1965). The variance of $\overline{d}$ with respect to the randomization distribution is

$$\begin{aligned}\text{var}(\overline{d}) &= \frac{1}{2n-1}\text{var}(t + \tau) \\ &= \frac{1}{(2n-1)2n}\sum_i \{(t_i + \tau_i) - (\overline{t} + \overline{\tau})\}^2,\end{aligned}$$

where $\overline{t} = (1/2n)\sum(t_i)$ and $\overline{\tau} = (1/2n)\sum(\tau_i)$. Note that if $t_i = -\tau_i$ for all $i = 1, \ldots, 2n$, one observes the true treatment effect for every subject regardless of the treatment assignment outcome, and $\text{var}(\overline{d}) = 0$. Grizzle (1965, 1974) also produced an estimator for the variance of $\overline{d}$. The bias of this estimator with respect to the randomization distribution is,

$$\begin{aligned}\text{bias} &= E(\widehat{\text{var}}(\overline{d}) - \text{var}(\overline{d})) \\ &= \frac{1}{(2n-1)}\text{var}(X - Y),\end{aligned}$$

where $\text{var}(X - Y) = \frac{1}{2n}\sum_i \{(X_i - Y_i) - (\overline{X} - \overline{Y})\}^2$. This result indicates the following:

- The bias is always greater than or equal to zero.
- For any study size, the vector of length $2n$, $X - Y$, must be a constant vector to ensure that the bias will be zero.
- For fixed $\text{var}(X - Y)$, the bias decreases as n increases.
- An untestable condition, $t + \tau =$ a constant, implies that $\overline{d} = \overline{D}$.

One could approach this problem from a different perspective. Suppose d_1 is a vector (length $2n$) of "potential" treatment effects of subjects assigned to the first treatment sequence (i.e., $T_i = 1$ for all i), and d_2 is a vector (length $2n$) of "potential" treatment effects of subjects assigned to the second treatment sequence ($T_i = 0$ for all i). Then results for the completely randomized design discussed in Section 2 could be used with $d_1 = X$ and $d_2 = Y$. Instead of potential treatment "responses" for each treatment as in Section 2, there are potential treatment effects for each treatment "sequence." Additivity in this sense would imply that $d_2 = d_1 + C$, where $C = 2(t + \tau)$ and where $(t + \tau)$ is a constant vector of length $2n$.

5. AN ILLUSTRATIVE EXAMPLE

The data were found in Senn (1993) and resulted from a two-period cross-over design concerning the effect of two drugs on peak expiratory ow (PEF) for asthma patients. The two treatments are 200 μg salbutamol, a well established bronchodilator (the control treatment), and 12 μg formoterol, a more recently developed bronchodilator (the test treatment). The treatment responses are PEF in liters per minute measured 8 hours after treatment, and it is assumed there are no carry over effects from the first to the second period. Table 1 shows observations for 12 subjects with 6 subjects in each treatment sequence, reported in the format of the potential response framework (the lower case x, y are used to denote observed vectors).

Table 1. Example Data

Time period 1		Time period 2	
$x - t$	$y - \tau$	$x + t$	$y + \tau$
310			260
370			300
410			390
250			210
380			250
330			365
	370	385	
	310	400	
	380	410	
	290	320	
	260	340	
	90	220	

The point estimate of $\overline{D}$ is equal to $\overline{d} = 54.17$. This is an estimate of the average increase in PEF due to the test treatment, formoterol, versus the control treatment, salbutamol, for the 12 subjects in the study. The estimated standard error is 14.40, using the estimator derived by Grizzle (1965, 1974) that assumes no carry-over effects. There are 924 possible values for $\overline{d}$ corresponding to the number of ways to assign 6 of 12 subjects to a treatment sequence. Note that the observed mean treatment effect, 54.17, is 3.76 estimated standard deviations away from zero. Furthermore, the estimate of standard error should be conservative. Equating point estimates with population quantities, and applying Chebyshev's inequality, one could estimate that "at least" 93% (859) of the 924 possible treatment assignments would have produced a positive value for $\overline{d}$. This is compelling evidence (with minimal assumptions) that the new treatment would have produced a higher average PEF value than the standard treatment, if each subject in the study could have been in

both sequences at the same time. That is, there is evidence that $\overline{X} > \overline{Y}$. Using a normal distribution model for treatment effects, the resulting p value would have been < 0.001 for a one-sided alternative hypothesis. Whether or not one believes that these results apply to a larger population would depend on one's judgment that the responses for the 12 subjects represent responses from individuals in the larger population.

6. CONCLUSION

This article presented some results that should benefit students' understanding of nonparametric randomization inference. In particular, the role of additivity in common randomization based tests of a sharp null hypothesis is clarified. Discussions in the literature concerning nonadditivity often focus on testing for "consequences" of nonadditivity and, if appropriate, trying to transform the data so that this consequence is not detected. A test for equality of variances among two treatment groups in a completely randomized design is one such test. Rank-based tests, such as Mann–Whitney, can be interpreted as randomization tests after first transforming the data to the integers.

In the presence of nonadditivity, a mean treatment effect and estimated standard error can still provide a nonparametric assessment of a treatment effect. Conclusions regarding the mean treatment effect may be conservative since estimates of standard error tend to overestimate the true value.

The biases that were reported herein cannot be estimated from observable data. Estimating the bias requires further assumptions about the data since the bias term depends on how the treatment effect varies from subject to subject (i.e., on the degree of subject-treatment interaction). A subject-treatment interaction term is a nonidentifiable quantity, but bounds for it can sometimes be estimated that may help an investigator to better understand how a treatment is affecting individuals in a study or population (Gadbury and Iyer 2000). If the degree of subject-treatment interaction is large, then the average treatment effect may not be a meaningful measure to use unless the source of this interaction (e.g., an unobserved covariate) can be identified.

[Received March 2001. Revised June 2001.]

REFERENCES

Copas, J. B. (1973), "Randomization Models for the Matched and Unmatched 2×2 Tables," *Biometrika*, 60, 467–476.

Cox, D. R. (1958), *Planning of Experiments*, New York: Wiley.

——— (1992), "Causality, Some Statistical Aspects," *Journal of the Royal Statistical Society*, Series A, 155, 291–301.

Gadbury, G. L., and Iyer, H. K. (2000), "Unit-Treatment Interaction and its Practical Consequences," *Biometrics*, 56, 882–885.

Grizzle, J. E. (1965), "The Two-Period Change-Over Design and its use in Clinical Trials," *Biometrics*, 21, 467–480.

——— (1974), "Corrections to 'The Two-Period Change-Over Design and its Use in Clinical Trials' (Grizzle, 1965)", *Biometrics*, 30, 727.

Hinkelmann, K., and Kempthorne, O. (1994), *Design and Analysis of Experiments* (vol. 1), New York: Wiley.

Holland, P. W. (1986), "Statistics and Causal Inference" (with discussion), *Journal of the American Statistical Association*, 81, 945–970.

Lachin, J. M. (1988), "Statistical Properties of Randomization in Clinical Trials," *Controlled Clinical Trials*, 9, 289–311.

Neyman, J. (1935), "Statistical Problems in Agricultural Experimentation" (with discussion), *Supplement to the Journal of the Royal Statistical Society*, Series B, 2, 107–180.

Rosenbaum, P. R. (1995), *Observational Studies*, New York: Springer-Verlag.

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

——— (1991), "Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism," *Biometrics*, 47, 1213–1234.

Rubin, D. R. (1980), Comment on Basu's "Randomization Analysis of Experimental Data" (1980), *Journal of the American Statistical Association*, 75, 591–593.

Senn, S. (1993), *Cross-Over Trials in Clinical Research*, West Sussex: Wiley.