Evaluation of the Power of Re-randomization Tests, with Application to Weather Modification Experiments.

By

K. RUBEN GABRIEL

and

CHIN-FEI HSU*

June, 1981

TECHNICAL REPORT 81/11

Department of Statistics and Division of Biostatistics

University of Rochester, Rochester, NY 14627 USA

Research supported in part by NSF grant ATM 79-05007 to Illinois State Water Survey, F. A. Huff and S. A. Changnon, principal investigators, and by NSF grant ATM 79-05536 to the University of Rochester, K. R. Gabriel and W. J, Hall, principal investigators.

*Illinois State Water Survey

Evaluation of the Power of Re-randomization Tests, with Application to Weather Modification Experiments.*

by

K. Ruben Gabriel and Chin-Fei Hsu
University of Rochester Illinois State Water Survey

ABSTRACT

Computations of power of re-randomization tests by exact methods are known to be computationally exhorbitant. We introduce a much cheaper "naive" method of evaluating power and show - both by simulation and analytically for a special case - that it very mildly overestimates true power. We further derive a normal approximation to re-randomization distributions of linear statistics and illustrate its closeness to the true distributions. This yields analytical formulas for calculating power approximately at negligible computational cost. The proposed methods therefore eliminate the previously prohibitive cost of calculating power for re-randomization tests and make it practical to evaluate the sample sizes needed for a randomized experiment to be analyzed by re-randomization. (See also Gabriel and Hall, 1981). Its application to weather modification experiments is illustrated.

*This is an extended version of a paper presented at the Third WMO Scientific Conference on Weather Modification at Clermont-Ferrand, July 21-25, 1980.

1. RE-RANDOMIZATION TESTS.

Re-randomization (permutation) methods are receiving increasing attention. These methods compare the evaluation of data from a randomized experiment with analogous evaluations of the same data calculated as though other experimental plans had been chosen. The probabilities for significance testing can then be based entirely on the chance selection of the experimental plan, and need not involve any assumptions about the stochastic nature of the data (Cox and Kempthorne, 1963). This robustness to specification is of paramount importance in meteorological experiments, where units of experimentation have to be used as they occur. Tukey, Brillinger and Jones (1978) recently came out strongly in favor of using re-randomization tests for confirmatory analyses of weather modification experiments. This was further emphasized by Gabriel (1979, 1981) who discussed some advantages of re-randomization tests over classical parametric tests. Unless there is a major breakthrough in understanding weather processes, so that better stochastic models become available, re-randomization tests are likely to be used more frequently in evaluating weather modification. Examples of such usage go back to Adderley (1961), and include several more recent papers (Bradley and Scott, 1980; Dennis et al, 1975; Elliot and Brown, 1971; Gabriel and Feder, 1969; Smith et al 1977). Similar considerations are relevant in many other areas of study.

One application of re-randomization tests is to experiments in which one randomly selects m out of N units to be treated, and leaves the N-m unselected units as controls. After imposing the treatment, its effect is evaluated by comparing some responses on the m treated units with the corresponding responses on the N-m controls. Mean responses may be compared,

or mean responses adjusted for some covariate observed on all N units prior to the treatment; treatment-control differences or ratios may be used; any one of a variety of statistics could be chosen - all would produce valid re-randomization tests. The typical statistic depends both on the experimental plan, i.e., selection of units for treatment, which we denote by E, and on the observed response vector, which we denote by $\underline{z}_E = (z_{E,1},\ldots,z_{E,N})$ The statistic can thus be written $s_E(\underline{z}_E)$

Re-randomizing mimics the experiment by making new random selections of m out of the N experimental units and then analyzing $\underline{z}_E$ as though these m units had been treated and those N-m had been used as controls. For such a re-randomization e, one calculates $s_e(\underline{z}_E)$ in exactly the same way as that for the experimental randomization E, except that one substitutes the re-randomization selection e of m units for the experimental selection E; the responses $\underline{z}_E$ remain the same.

The purpose of calculating statistics for re-randomizations e is to see how much they vary from one another. That allows one to gauge the experimental statistic $s_E(\underline{z}_E)$ against the random variation of the $s_e(\underline{z}_E)$'s.

The statistical test of a treatment effect is carried out by running repeated re-randomizations and finding how many of them have statistics $s_e(\underline{z}_E)$ as large as, or larger than, the experimental statistic $s_E(\underline{z}_E)$ The proportion of such re-randomizations is the P-value. If it is less than a, the effect is said to be significant at level a.

This argument does not depend on either the form of the statistics s( ) or the distribution of the responses $\underline{z}$. That is why re-randomization inference is robust and flexible both in allowing a wide choice of statistics and in being distribution-free, even to the extent that it does not require the $\underline{z}$'s to be random variables. All it does require is that

the experimental selection E have been picked at random from an ensemble $\mathcal{E}$ of K selections, be they all $\binom{N}{m}$ possible selections or any well defined subset of these selections. In actually applying a re-randomization test the statistic $s_e(\underline{z}_E)$ is computed for each $e \varepsilon \mathcal{E}$, i.e., for each of the K re-randomizations, and these K statistics are ordered as

$$s_{(1)}(\underline{z}_E) < s_{(2)}(\underline{z}_E) \leq \ldots \leq s_{(K)}(\underline{z}_E). \tag{1}$$

The a-level significance test is then to reject the null hypothesis if

$$s_E(\underline{z}_E) \geq s_{(K(1-\alpha))}(\underline{z}_E). \tag{2}$$

(Note that we assume α chosen so that Ka is an integer. For further discussion of such inferences see Gabriel and Hall, 1981).

2. POWER AND ITS NAIVE APPROXIMATION

To evaluate the power of this test at an alternative hypothesis, we must consider how many of the K possible experiments based on selections from ensemble E would have rejected the original hypothesis. The power 1-6 is the proportion of experiments in which this happens, i.e.,

$$1-\beta = \sum_{E \varepsilon \mathcal{E}} I\{s_E(\underline{z}_E) > s_{(K(1-\alpha))}(\underline{z}_E)\}/K, \tag{3}$$

where I{.} is one or zero according to whether its argument is true or false.

This exact evaluation of the power is seen to require the calculation of K statistics $s_e(\underline{z}_E)$, $e \varepsilon \mathcal{E}$, <u>for each E</u>; K^2 such statistics altogether. Unless the ensemble E is quite restricted this will be a prohibitively large amount of computation (Kempthorne and Doerfler, 1969). However, asymptotic methods are often useful (Davis, 1979).

Another method has sometimes been used to evaluate power by a naive analogy to parametric calculations. One first obtains the "null distribution" of the $s_e(\underline{\zeta})$, $e \varepsilon \mathcal{E}$, where $\underline{\zeta}$ are the potential values assumed to occur if no treatment effect existed. One then uses their upper α point $s_{(K(1-\alpha))}(\underline{\zeta})$ as a critical value. The "alternative" distribution of the treated

DISPLAY 1: A Simple Example of Exact and Naive Power Evaluation

---

Units	u	=	1	2	3	4
Potential Data	ζ_u	=	9	7	3	4
Experiments	$E\begin{cases} e_{1,u} \\ e_{2,u} \\ e_{3,u} \\ e_{4,u} \\ e_{5,u} \\ e_{6,u} \end{cases}$	=	t	t	c	c
		=	t	c	t	c
		=	t	c	c	t
		=	c	t	t	c
		=	c	t	c	t
		=	c	c	t	t

LEGEND: t = seeded
c = unseeded

Statistic $s_{e_i}(\underline{z})$ = (sum of z_u's with $e_{i,u}$ = t) - (sum of z_u's with $e_{i,u}$ = c)
= "seeded sum" - "unseeded sum"

Treatment Effect Doubling, i.e., $z_{E,u} = \begin{cases} \zeta_u & \text{if } E_u = c \\ 2\zeta_u & \text{if } E_u = t \end{cases}$

No Treatment Values[*] of $s_e(\underline{\zeta})$ = 9, 1, 3, -3, -1, -9, ; Max 9

Actual Values

							Max	1/6-level Significant[***]
expt $\underline{e}_1$ - values of $s_e(\underline{z}_{e_1})$ =	25,[**]	3,	5,	-5,	-3,	-25 ;	25 :	Yes
expt $\underline{e}_2$ - values of $s_e(\underline{z}_{e_2})$ =	15,	13,[**]	9,	-9,	-13,	-15 ;	15 :	No
expt $\underline{e}_3$ - values of $s_e(\underline{z}_{e_3})$ =	14,	6,	16,[**]	-16,	-6,	-14, ;	16 :	Yes
expt $\underline{e}_4$ - values of $s_e(\underline{z}_{e_4})$ =	13,	-3,	-7,	7,[**]	3,	-13 ;	13 :	No
expt $\underline{e}_5$ - values of $s_e(\underline{z}_{e_5})$ =	12,	-10,	0,	0,	10,[**]	-12 ;	12 :	No
expt $\underline{e}_6$ - values of $s_e(\underline{z}_{e_6})$ =	2,	0,	4,	-4,	0,	-2[**];	4 :	No

---

[*] The six $s_e(\underline{\zeta})$ values for $\forall e \in E$.

[**] Underlined are actual statistics $s_E(\underline{z}_E)$, for expt E - other $s_e(\underline{z}_E)$, $e \neq E$, are re-randomization statistics.

[***] 1/6 significant if $s_E(\underline{z}_E) = \max_e s_e(\underline{z}_E)$

EXACT POWER: Proportion of significant experiments, i.e., $1-\beta = 2/6$

NAIVE EVALUATION: Proportion of experiments with $s_E(\underline{z}_E) \geq \max_e s_e(\underline{\zeta}) = 9$, i.e., $1-b = 4/6$
(9 = the upper 1/6 point of the no treatment values)

responses' statistics $s_E(\underline{z}_E)$, $E \varepsilon \mathcal{E}$, is obtained next and yields the "naive" evaluation of power

$$1\text{-}b = \sum_{E \varepsilon \mathcal{E}} I \{s_E(\underline{z}_E) > s_{(K(1-\alpha))}(\underline{\zeta})\}/K \tag{4}$$

[9, Section 3.2].

This method is much cheaper computationally. It requires calculation of only 2K statistics s, i.e., a (K/2)-fold saving. However, it will not in general coincide with the exact evaluation (3). Treatment effects are likely to increase the variability of the $s_e(\underline{z}_E)$'s as compared to that of the $s_e(\underline{\zeta})$'s and so result in

$$s_{(K(1-\alpha))}(\underline{z}_E) > s_{(K(1-\alpha))}(\underline{\zeta}) \tag{5}$$

Hence the naive evaluation 1-b is likely to be in excess of true power 1-8.

A highly simplified example of power calculations by both methods is set out in Display 1. The naive calculation requires only the six no treatment values $s_e(\underline{\zeta})$ and the six actual values $s_E(\underline{z}_E)$, whereas the exact calculation requires all 36 values $s_e(\underline{z}_E)$, ~~$\forall$~~ e, ~~$\forall$~~E. This simple example also illustrates the power excess of the naive evaluation

$$1\text{-}b > 1\text{-}\beta. \tag{6}$$

Examples with larger numbers usually show much smaller excesses.

The cost of exact evaluations of power (3) may be prohibitive, especially if it is needed for several alternatives, several sample sizes, several significance levels and several statistics. The naive evaluation (4) may be more feasible in terms of cost but it is not clear whether it can be trusted. It would be important to know if and when the naive method is reliable in giving approximately correct evaluations of power.

An idea of the relative costs of the two methods may be obtained from a numbe of examples in which multiple regression tests based on

K=100 re-randomizations of 35 observations used approximately 200 CPU seconds of CDC CYBER 175 computer time for exact evaluation of power whereas the naive method used only 5 CPU seconds. The 40-fold savings was close to the expected of 50-fold savings.

3. MONTE CARLO STUDIES UNDER A MULTIPLICATIVE MODEL

Some examples were run with a view to studying the relation between naive and exact evaluations of power in realistic situations in the context of weather modification experimentation. Data consisted of monthly (May to September) and seasonal-average rainfalls for 10 counties in west-central Kansas, observed for 35 summers (1936-1970) (Hsu, 1979b).

For a practical study of power of re-randomization tests, the following experiment was mimicked. Two of the 10 Kansas counties were designated as "seeding targets", i.e., "target 1" and target 2", and the other 8 counties, which surround those two, were designated as controls, i.e., "control 1" to "control 8". A simulated experiment consisted of a random choice E of five of the summers 1936-70 to be "seeded", and the other 30 summers allocated to be "unseeded". The "seeding effect" was postulated to be multiplicative, so for each simulated experiment the "target" rainfalls were multiplied by $(1+\tau)$ in each "seeded" year; other rainfalls, in the "controls" and during "unseeded" summers, were left unaltered.

The present study used a restricted reference set of only 100 "experimental" selections of 5 out of the 35 summers. This set was randomly sampled at the beginning of these Monte Carlo studies and used for all subsequent re-randomizations.

The simulated "effects" applied to the "experiments" were τ = 0.1, 0.2, 0.3 and 0.4. The null hypothesis thus was H_o: τ=0, and the one-sided

DISPLAY 2: Power of the Double Ratio, evaluated exactly 1-6, and naively, 1-b, at the 5% and 10% Significance Levels.

		Target 1				Target 2				Target 3			
		Significance Level				Significance Level				Significance Level			
		5%		10%		5%		10%		5%		10%	
Month	T	1-b	1-6	1-b	1-6	1-b	1-6	1-b	1-6	1-b	1-6	1-b	1-6
May	0.1	.16	.11	.35	.31	.14	.10	.33	.22	.14	.13	.28	.24
	0.2	.39	.24	.59	.51	.33	.22	.58	.49	.35	.35	.57	.52
	0.3	.62	.41	.80	.73	.58	.51	.75	.69	.63	.59	.79	.75
	0.4	.80	.65	.91	.86	.74	.66	.86	.80	.82	.81	.92	.92
June	0.1	.21	.17	.31	.36	.16	.08	.18	.15	.23	.20	.29	.30
	0.2	.38	.25	.45	.30	.24	.17	.27	.26	.54	.49	.68	.69
	0.3	.50	.47	.70	.62	.33	.29	.37	.45	.82	.78	.85	.84
	0.4	.74	.66	.81	.78	.52	.43	.59	.65	.96	.92	.98	.97
July	0.1	.14	.11	.24	.22	.13	.13	.21	.20	.25	.17	.33	.30
	0.2	.25	.21	.55	.45	.25	.24	.27	.35	.48	.39	.57	.58
	0.3	.55	.46	.69	.66	.40	.38	.49	.48	.72	.66	.77	.77
	0.4	.69	.62	.86	.83	.52	.51	.64	.66	.87	.81	.90	.89
August	0.1	.17	.15	.23	.19	.11	.09	.22	.18	.15	.11	.26	.23
	0.2	.30	.25	.40	.38	.22	.17	.42	.36	.33	.25	.55	.49
	0.3	.49	.43	.59	.57	.42	.32	.58	.53	.61	.51	.75	.69
	0.4	.69	.59	.81	.80	.59	.52	.71	.65	.78	.77	.90	.91
September	0.1	.09	.08	.30	.23	.17	.13	.23	.23	.29	.25	.37	.34
	0.2	.24	.13	.50	.43	.34	.28	.43	.38	.60	.51	.70	.66
	0.3	.46	.41	.75	.73	.50	.43	.57	.55	.78	.75	.82	.81
	0.4	.69	.70	.87	.82	.64	.59	.71	.71	.86	.86	.90	.90
Season	0.1	.23	.18	.48	.44	.17	.15	.45	.37	.56	.42	.67	.62
	0.2	.73	.67	.90	.83	.60	.61	.80	.74	.90	.87	.92	.90
	0.3	.95	..95	.98	.98	.84	.85	.92	.91	.98	.98	.98	.99
	0.4	.99	.99	1.00	1.00	.94	.98	.99	.99	1.00	1.00	1.00	1.00

alternatives H_i: $\tau=i/10$, i=l,2,3,4. Several statistics were calculated for each simulated and "treated" "experiment", and each was tested at levels $\alpha = 0.10$ and $\alpha = 0.05$. The statistics used were the double ratio (DR) (Gabriel Feder,1969; Davis, 1979), two variants of multiple regression (MR) and five sum of rank power tests (SRP) (Hsu et al, 1981; Hsu, 1979b). Powers were evaluated by both formulas (3) and (4).

Display 2 illustrates the results for the Double Ratio statistic. It is evident that naive 1-b is usually a little higher than exact 1-6.

To summarize these power calculations, the exact power 1-B was plotted against its naive approximation 1-b. A curve of the form

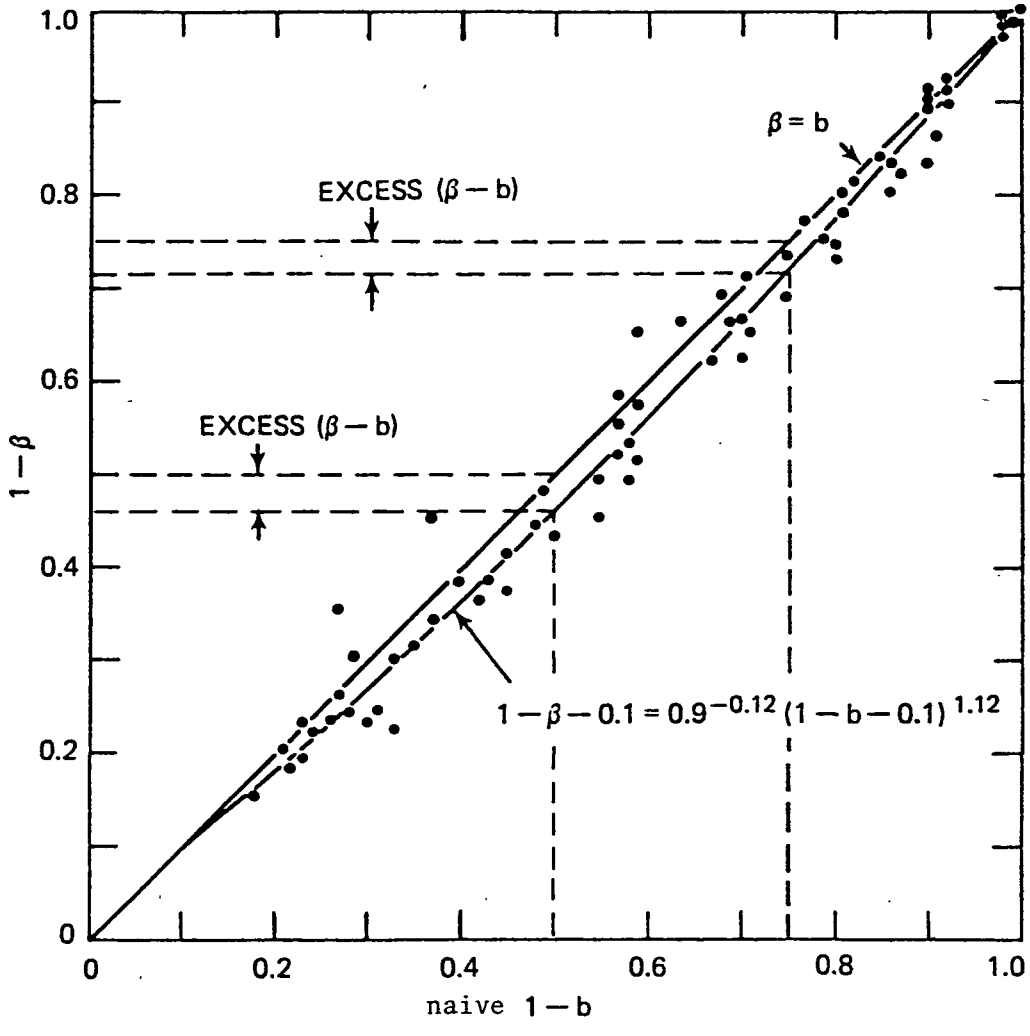
$$1-\beta = \alpha + (1-b-\alpha)^{1+r}/(1-\alpha)^{r} \qquad (7)$$

was found to provide a good fit of the regression of 1-β onto 1-b. Display 3 illustrates the scatter and fit for the DR statistic at level $\alpha = 0.10$; the 72 points represent all pairs of (1-b, 1-β) values for $\alpha=0.10$ in Display 2. The curve—fitted by linear regression of log(l-β-α) onto log(l-β-α)-- shows that 1-b is usually slightly above 1-B, with an excess of no more than 0.04.

Some confirmation of the reliability of these Monte Carlo estimates is obtained from a comparison with approximations obtained by Petrondas(1981) for the Double Ratio statistic. For the entire season, power calculations for the target average were as in Display 4. These generally support the present Monte Carlo estimates.

Similar tabulations and scatters were obtained for $\alpha = 0.10$ and $\alpha = 0.05$ for each one of the statistics concerned. The log-log regressions for all of them had coefficients of determination above 0.98. The results are summarized in Display 5 which shows the parameter r which measures

.DISPLAY 3: Powers of double ratio evaluated exactly and naively at 10% level.



DISPLAY 4. Three Estimates of the Power of Tests Comparing Average Target Area Total Season Precipitation for 5 "Seeded" Years with 30 "Unseeded" Ones

Level α	0.05				0.10			
"Effect" T	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
Simulations (Display 2)								
1-B	.42	.87	.98	1.00	.62	.90	.99	1.00
Naive 1-b	.56	.90	.98	1.00	.67	.92	.98	1.00
Petrondas's Formulas (1981)								
1-S	.39	.80	.97	1.00	.54	.90	.99	1.00
1-b	.41	.85	.99	1.00	.55	.92	1.00	1.00
Linear Approximations								
1-β(10)	.40	.89	1.00	1.00	.56	.95	1.00	1.00
1-b (12)	.41	.88	1.00	1.00	.56	.94	1.00	1.00

the closeness of naive 1-b to 1-β. The Display also shows the average excess of 1-b over 1-β

$$1\text{-}b\text{-}\alpha\text{-}(1\text{-}b\text{-}\alpha)^{1+r}/(1\text{-}\alpha)^{r} = (1\text{-}b\text{-}\alpha)\{1\text{-}(1\text{-}b\text{-}\alpha)^{r}/(1\text{-}\alpha)^{r}\} \qquad (8)$$

for two values of 1-b.

Thus, for example, a naively evaluated 1-b = 0.75 would, on the average, exceed exact power 1-8 by $(0.75\text{-}\alpha)\{1\text{-}(0.75\text{-}\alpha)^{r}/(1\text{-}\alpha)^{r}\}$. Most of the excesses of naive over exact power are below .05. Evidently the naive 1-b calculation usually exceeds the true power 1-8 by less than that. This suggests that if the naive method is used in power studies it will not bias the conclusions unduly, especially if some caution is exercised.

It is interesting to note that the excess of 1-b over 1-β is generally less at 1-b = 0.75 than at 1-b = 0.50, at α = 0.10 than at α = 0.05; less with ranks and normalized t statistics than with ratios and regressions. Presumably ranks are less sensitive to unusual randomizations than actual values. This needs further study.

4. APPROXIMATE FORMULAS FOR POWER

Power of tests under re-randomization may also be obtained through approximating their re-randomization distribution (see also Gabriel & Hall, 1981)

$$s_e(z_E) = \sum_{u=1}^{N} (d_{e,u}\text{-}\bar{d}_e) z_{E,u}/v_e, \qquad (9)$$

where $d_{e,1}\ldots,d_{e,N}$ - with mean $\bar{d}_e$ and standard deviation v_e - are treatment dosages assigned by experimental plan e. In the present context, the dosages are either $d_{e,u} = 1$ - if unit u is allocated to treatment - or $d_{e,u} = 0$ - if unit u is allocated to control. $N\bar{d}_e$ is therefore the number of treated units and $(N\text{-}1)v_e^2 = N\bar{d}_e(1\text{-}\bar{d}_e)$

Approximations are developed in the Appendix to this paper. This development is based on the following assumptions: (i) The effect of treatment on unit u is proportional to the dosage $d_{e,u}$; (ii) this effect

DISPLAY 5: Estimated r and excesses for 1-b = 0.50 and 0.75 at a = 0.10 and 0.05 levels

	Level α	r	Average Excess at l-b=.50	Average Excess at l-b=.75
Double Ratio	.10	.119	.037	.025
	.05	.192	.060	.040
Multiple Regression				
T	.10	.062	.020	.013
	.05	.051	.017	.011
D	.10	.147	.045	.030
	.05	.261	.080	.054
Sum of Rank Power				
A1	.10	.070	.022	.015
	.05	.200	.062	.041
A2	.10	.066	.021	.014
	.05	.102	.033	.021
A3	.10	.048	.015	.010
	.05	.046	.015	.010
C2	.10	.072	.023	.015
	.05	.052	.017	.011
C3	.10	.065	.021	.014
	.05	.070	.023	.015

T = Normalized t-Statistic

D = Average of Difference

$Ar = \Sigma(R_i)^r$; $Cr = \Sigma sign(D_i)|D_i|^r$

Summation is over the seeded sample; R. is the rank of the i-th seeded target-control ratio; $D_i = R_i(N+1)/2$, with N the total number of observations; sign(d) = 1,0 or -1 according to whether d is >0, 0 or <0.

is additive to the potential value that is/would be observed in the absence of treatment. Moreover, (iii) the reference set E is constrained to plans such that $v_e^2 = v_E^2$, say, for all $e \varepsilon E$. In the present context of a treatment-control experiment, constraints (iii) mean that all plans in the reference set must have the same number m of units to be treated.

The re-randomization distribution of $s_e(z_E)$, as defined in (9), is approximated by a Pearson Type II distribution (Appendix, (A.11), (A.12)). The moments, in standardized form (A.13), are seen to approach those of the Normal rapidly. Normal distribution theory is therefore applicable in approximating power at level a for proportional effect 6 as in (A.32) by

$$1-\beta_\alpha(\delta) = 1-\Phi\left(\frac{z_{1-\alpha}(1+\Delta^2)^{1/2}-\Delta(N-1)^{1/2}}{1-z_{1-\alpha}\Delta(N-1)^{-1/2}(1+\Delta^2)^{-1/2}}\right), \tag{10}$$

where $\Phi(\ )$ is the Normal probability integral, and

$$\Delta = \delta v_E / v_\zeta, \tag{11}$$

v_ζ being the standard deviation of the potential values. Naive power is similarly found to be - (A.35) -

$$1-b_\alpha(\delta) = 1 - \Phi(z_{1-\alpha}-\Delta(N-1)^{1/2}). \tag{12}$$

The difference between the Normal approximations of the naive evaluation and the exact power - (10) and (12) - becomes, for large N, approximately

$$[1-b_\alpha(\delta)]-[1-\beta_\alpha(\delta)] \approx \Phi(z_{1-\alpha}(1+\Delta^2)^{1/2}-\Delta(N-1)^{1/2})-\Phi(z_{1-\alpha}-\Delta(N-1)^{1/2}) \tag{13}$$

By Taylor's remainder theorem, that becomes

$$[1-b_\alpha(\delta)]-[1-\beta_\alpha(\delta)] \approx \frac{1}{\sqrt{2\pi}}((1+\Delta^2)^{1/2}-1)\exp[-1/2\{z_{1-\alpha}-\Delta(N-1)^{1/2} + \theta z_{1-\alpha}((1+\Delta^2)^{1/2}-1)\}^2], \tag{14}$$

for some $\theta \varepsilon (0,1)$.

DISPLAY 6: Moments of Permutation Distributions of $\sqrt{N-1}\ \cos(\underline{c}_e, \underline{z}_E - \bar{z}_E \underline{1})$, $e \in E$

Design*	Data**	Reference Set	No. of Replications***	Average Values Of					
				μ_1	μ_2	μ_3	μ_4	μ_5	μ_6
$\binom{35}{5}$	Normal	Sample of 100	13	-.08	.94	.03	2.42	0.47	9.51
		(Theoretical)		(0)	(1)	(0)	(2.83)	(0)	(12.68)
$\binom{35}{5}$	Uniform	Sample of 100	11	0	.97	.13	2.56	-.02	10.37
		(Theoretical)		(0)	(1)	(0)	(2.83)	(0)	(12.68)
$\binom{16}{8}$	Normal	Complete (12870)	1	0		0	2.86	0	12.98
		(Theoretical)		(0)	(1)	(0)	(2.65)	0	(10.45)
$\binom{12}{6}$	Normal	Complete (924)	2	0		0	2.59	0	9.65
		(Theoretical)		(0)	(1)	(0)	(2.54)	(0)	(9.31)
$\binom{12}{4}$	Normal	Complete (495)	1	0		-.14	2.53	-1.10	9.31
		(Theoretical)		(0)	(1)	(0)	(2.54)	(0)	(9.31)
$\binom{12}{3}$	Normal	Complete (220)	2	0		-.215	2.55	-1.70	9.72
		(Theoretical)		(0)	(1)	(0)	(2.54)	(0)	(9.31)
$\binom{12}{2}$	Normal	Complete (66)	2	0		-.33	2.48	-2.35	9.54
		(Theoretical)		(0)	(1)	(0)	(2.54)	(0)	(9.31)

* Design $\binom{N}{m}$ is one of taking m units - out of a total of N - to be treated.

** The data vector $\underline{z}$ was a sample from one of these distributions.

*** Each replication is a calculation on a different data set and/or reference set.

Evidently, the naive power evaluation will tend to exceed the exact power somewhat, but for large N the difference approaches zero. Moreover, for large A, the convergence to zero is faster.

In order to justify the use of these approximations, we checked them for a number of theoretical and empirical re-randomization distribution.

Display 6 shows the first six moments of some re-randomization distributions and compares them with the approximations (A.13). Apparently the distributions are not far from Normal even when experiments are as small as the ones shown, i.e., N = 35, 16 or even 12.

More important than the closeness of the moments is that of the approximate calculations of power by formulas (10) and (12). Display 7 illustrates such approximations and compares them with powers obtained exactly from restricted reference sets (For methods of computation see Gabriel and Hall, 1981). It is evident that approximation (10) usually overestimates power slightly and that naive approximation (12) is a little higher yet. For designs of 16 or more units we find that when true power is about 0.75, (10) would produce estimates averaging about 0.78 and naive (12) about 0.81. When the true power is larger than 0.80, the maximum discrepancy is 0.02, except in one case where it is 0.04. For skew designs of 12 the approximations are not so close - (10) as well as (12) overestimate true power more appreciably.

It would seem that for all but very small experiments both the approximation (10) of exact power and approximation (12) of the naive evaluation come reasonably close to true power - only few percentage points above it. This is particularly remarkable since the reference sets used

DISPLAY 7: Evaluations of Power by Exact and Approximate Methods

Design*	Data	Reference Set	a		Powers			
$\binom{35}{5}$	Normal	3 samples of 100	.05	(δ) l-β(exact) 1-β(10) l-b(12)	(.9) .58 .69 .71	(1.0) .68 .77 .79	(1.3) .90 .93 .94	(1.55) .97 .98 .98
			.10	(δ) l-β(exact) 1-β(10) l-b(12)	(.7) .62 .66 .67	(.9) .77 .82 .82	(1.1) .91 .92 .92	(1.3) .97 .97 .97
$\binom{35}{5}$	Uniform	3 samples of 100	.05	(δ) - 1-β(exact) 1-β(10) l-b(12)	(2.56) .53 .56 .59	(3.20) .70 .73 .75	(3.50) .76 .80 .82	(4.46) .94 .94 .95
			.10	(δ) l-β(exact) 1-β(10) l-b(12)	(1.00) .31 .28 .29	(1.74) .46 .48 .49	(2.50) .66 .70 .71	(3.50) .89 .90 .90
$\binom{16}{8}$	Normal	2 samples of 100	.05	(δ) 1-β(exact) 1-β(10) l-b(A2)	(4.36) .51 .56 .62	(5.35) .70 .72 .77	(6.00) .78 .82 .85	(7.00) .94 .92 .93
			.10	(δ) 1-β(exact) 1-β(10) l-b(12)	(3.00) .50 .50 .52	(4.00) .66 .67 .70	(4.50) .77 .75 .77	(6.00) .92 .92 .92
$\binom{12}{6}$	Normal	2 samples of 100	.05	(δ) l-β(exact) 1-β(10) l-b(12)	(.75) .49 .50 .60	(1.00) .73 .75 .81	(1.25) .91 .91 .94	
			.10	(δ) l-β(exact) 1-β(10) l-b(12)	(.55) .49 .50 .55	(.75) .66 .70 .73	(1.00) .90 .89 .90	

* See Display 6.

DISPLAY 7 (Continued)

Design	Data	Reference Set	α		Powers			
$\binom{12}{4}$	Normal	2 samples of 100	.05	(δ) 1-β(exact) 1-β(10) l-b(12)	(.65) .42 .37 .46	(1.00) .70 .70 .77	(1.25) .86 .88 .91	(1.50) .96 .97 .97
			.10	(δ) 1-β (exact) 1-β l-b(12)	(.55) .46 .47 .52	(.75) .66 .66 .70	(1.00) .86 .85 .87	(1.20) .95 .94 .94
$\binom{12}{3}$	Normal	Complete	.05	(δ) 1-β(exact) 1-β(10) 1-b	(1.00) .54 .52 .61	(1.15) .65 .63 .72	(1.35) .84 .77 .83	(1.70) .90 .93 .95
$\binom{12}{2}$	Normal	Complete	.05	(δ) 1-β(exact) 1-β(10) l-b(12)	(1.00) .52 .41 .51	(1.50) .67 .73 .80	(1.75) .82 .84 .90	(2.00) .97 .94 .95

are by no means complete - in fact, they are quite small. It suggests high robustness of the form of the re-randomization distribution to the choice of reference set (See also Iglewicz, Ascher and Begg, 1981).

Finally, we have applied power approximations (10) and (12) to the Kansas seasonal precipitation data described above. To compare with previous calculations we have now assumed an additive effect of seeding, of the order of $\beta = \tau x$(Average seasonal target precipitation), where T is the proportional effect above. Also, for these approximations we have considered a regression-type (or shift) statistic of the "target"-"control" precipitation onto an 0 ("unseeded") and 1 ("seeded") dosage. Formulas (10) and (12) thus apply. (The parameters were estimated from Kansas data as follows: Average "target" precipitation = 2.5 inches per average month. Standard Deviation of "target-control" precipitation = 0.36 = $v_{\underline{\zeta}}$. Standard Deviation of dosages, $v_E = \sqrt{(5x30)/35} = 0.35$)

These power calculations are shown in Display 4. They are not very different from those for the Double Ratio statistic obtained under a multiplicative effect model.

The comparative calculations presented here confirm that our Normal approximations to power are reasonably close even for quite small experiments. They further justify our earlier conclusion that the naive evaluation method produces a slight overestimate of true power. And they also indicate that power is relatively robust over a variety of methods of calculation and approximation and "reasonable" choice of statistics.

5. CONCLUSION

The exhorbitant computational cost of exact evaluation of power for re-randomization tests can be avoided by using the naive method with the understanding that with small experiments it may produce figures that are a few percent above true power. The proportional computational savings are of the order of K/2-fold, where K is the number of plans in the reference set.

The cost of computing power can be further reduced drastically by using an approximation developed in the Appendix. This approximation is strictly speaking valid only for regression-type or shift statistics and complete reference sets but seems to work well also for other sets.

The practical importance of evaluating power is in planning an experiment: Deciding on the adequacy of a design and/or choosing an adequate sample size. Such calculations are of necessity approximate and rough since the true alternative cannot possibly be known prior to the experiment. We therefore conclude that the power approximations developed here should in most cases serve quite satisfactorily in planning experiments.

ACKNOWLEDGEMENT

This research was supported in part by National Science Foundation grants ATM79-05007 and ATM79-05536. The opinions and conclusions expressed herein are those of the authors and do not necessarily reflect the views of NSF. Demetrios Petrondas's help with calculations is gratefully acknowledged.

REFERENCES

ADDERLEY, E.E. (1961). Non-parametric methods of analysis applied to large-scale cloud seeding experiments. J. Meteor., 18, 692-694.

BRADLEY, R.A. and SCOTT, E. (1980). Perspectives from a weather modification experiment. Commun. Statist. - Theor. Meth. (To appear.)

COX, D.F. and KEMPTHORNE, 0. (1963). Randomization tests for comparing growth curves. Biometrics, 19, 307-317.

DAVIS, A.W. (1979). On certain ratio statistics in weather modification experiments. Technometrics, 21, 283-289, and 22, 136.

DEMPSTER, A.P. (1969). Elements of Continuous Multivariate Analysis. Reading, Mass: Addison-Wesley.

DENNIS, A.S., MILLER, J.R., and SCHWALLER, R.L. (1975). Evaluation by Monte Carlo tests of effects of cloud seeding on growing season rainfall in North Dakota. J. Appl. Meteor., 14, 959-969.

ELLIOT, R.D, and BROWN, K.J. (1971). The Santa Barbara II Project -- Downwind effect. Preprints, Internat. Conf. on Wea. Mod., Canberra, Australia, Amer. Meteor. Soc, Sept. 6-11, 179-184.

GABRIEL, K. RUBEN (1979). Some statistical issues in weather experimentation. Comm. Statist. - Theor. Meth., A8 (10), 975-1015.

GABRIEL, K. RUBEN (1981). On the roles of physicists and statisticians in weather modification experimentation. Bull. Amer. Meteor. Soc. 62, (1)62-69.

GABRIEL, K.R. and FEDER, Paul (1969). On the distribution of statistics suitable for evaluating rainfall stimulation experiments. Technometrics, 11, 149-160.

GABRIEL, K.R. and HALL, W.J. (1981). Re-randomization inference on regression and shift effects: Computationally feasible methods. University of Rochester, Statistics Technical Report 81/7.

HSU, C.-F. (1979a). Two methods of computing statistical powers with application to weather modification. Proc. Statist. Comp. Section, American Statistical Association, Washington, D.C., 243-246.

HSU, C.-F. (1979b). Monte Carlo studies of statistical evaluation techniques for weather modification. Preprints, Seventh Conf. on Inadv. and Planned Wea. Mod., Banff, Canada, Amer. Meteor. Soc., J3-J4.

HSU, C.-F., GABRIEL, K.R., and CHANGNON, S. A. (1981). Statistical Techniques and Key Issues for the Evaluation of Operational Weather Modification. J. Weather Modification, 13, 195-199.

IGLEWICZ, B., ASCHER, S. and BEGG, C. (1981). Treatment Allocation in Sequential Clinical Trials. Submitted for publication.

KEMPTHORNE, 0. and DOERFLER, T.E. (1969). The behavior of some significance tests under experimental randomization. Biometrika, 56, 231-248.

PETRONDAS, DEMETRIOS (1981). Personal Communication.

SMITH, E.J., VEITCH, L.G., SHAW, D.E., and MILLER, A.J. (1977). A Cloud Seeding Experiment in Tasmania, 1964-1970. Report CP 183, CSIRO, Australia, 115 pp.

TUKEY, J.W., BRILLINGER, D.R., and JONES, L.V. (1978). The Management of Weather Resources. Volume II: The role of statistics. Weather Modification Advisory Board, Statistical Task Force. Washington, D.C.: U.S. Government Printing Office.

APPENDIX: ON AN APPROXIMATION TO THE RE-RANDOMIZATION DISTRIBUTION OF LINEAR STATISTICS..

Consider experiments performed on a batch of N units with potential value ζ_u for unit u(=l,...,N) - that is the response that would occur on that unit if no treatment were applied. The experimental plan, denoted E, is chosen randomly from a reference set E of plans, and assigns treatment dosage $d_{E,u}$ to unit u. The vector of dosages can be written $\underline{d}_E$ and the reference set E as a collection $\{\underline{d}_e \mid e\varepsilon E\}$ of such dosage vectors.

We make the following assumptions about the form of the treatment effects. (i) It is proportional to the dosage and (ii) it is additive to the potential value. Thus, the observed response on unit u is assumed to be

$$z_{E,u} = \zeta_u + \delta\, d_{E,u}\,, \tag{A.la}$$

for "effect" coefficient 6. Writing $\underline{\zeta}$ and $\underline{z}_E$ for the vectors of potential values and responses, this becomes

$$\underline{z}_E = \underline{\zeta} + \delta \underline{d}_E. \tag{A.lb}$$

For an analysis of the randomized experiment E under model (A.l) we use regression-type statistics relating observed responses to dosage. Thus, we define, for each eεE,

$$\bar{d}_e = N^{-1}\, \underline{1}'\underline{d}_e, \tag{A.2}$$

where 1 is a vector of N ones,

$$v_e^2 = (N-1)^{-1}\, \| \underline{d}_e - \bar{d}_e \underline{1} \|^2 \tag{A.3}$$

and

$$\underline{c}_e = v_e^{-1}(\underline{d}_e - \bar{d}_e \underline{1}) \quad .$$

The statistics $s_e(\underline{z})$ become

$$\underline{c}_e'\underline{z} = \Sigma_{u=1}^{N}(d_{e,u}-\bar{d}_e)z_u/\sqrt{\{\Sigma_{u=1}^{N}(d_{e,u}-\bar{d}_e)^2/(N-1)\}}. \qquad (A.5)$$

These statistics are normalized in so far as

$$\underline{1}'\underline{c}_e = \Sigma_{u=1}^{N} c_{e,u} = 0 \qquad (A.6)$$

and

$$\| \underline{c}_e \| = \sqrt{\Sigma_{u=1}^{N} c_{e,u}^2} = \sqrt{(N-1)}, \qquad (A.7)$$

for each $e\varepsilon E$.

To get a handle on the distribution of these statistics, it is useful to think of them geometrically. The coefficients $\underline{c}_e$ are in contrast space and lie on the surface of the (N-l)-dimensional origin-centered hypersphere of radius $\sqrt{N-1}$ orthogonal to $\underline{1}$. For observations $\underline{z}_E$ with mean $\bar{z}_E$, the statistics (A.5) become

$$\underline{c}_e'\underline{z}_E = \underline{c}_e'(\underline{z}_E-\bar{z}_E\underline{1}) = (N-1)v_{\underline{z}_E}\cos(\underline{c}_e,\underline{z}_E-\bar{z}_E\underline{1}), \qquad (A.8)$$

where

$$v_{\underline{z}_E}^2 = \| \underline{z}_E-\bar{z}_E\underline{1}\|^2/(N-1) \qquad (A.9)$$

and

$$\cos(\underline{c}_e,\underline{z}_E-\bar{z}_E\underline{1}) = \mathrm{corr}(\underline{d}_E,\underline{z}_E). \qquad (A.10)$$

For any given $\underline{z}_E$, the statistics $\underline{c}_e'\underline{z}_E$ are thus seen to be proportional to the co-ordinate $\sqrt{N-1}\cos(\underline{c}_e,\underline{z}_E-\bar{z}_E\underline{1})$ of $\underline{c}_e$ in the direction of $\underline{z}_E-\bar{z}_E\underline{1}$.

Re-randomization inference on the effect 6 is carried out by relating the observed statistic $\underline{c}_E'\underline{z}_E$ to the distribution of the $\underline{c}_e'\underline{z}_E$'s, $e\varepsilon E$. This distribution will be denoted $(\underline{c}_e'\underline{z}_E \mid e\varepsilon E)$. Its general form is not available, for any E and any $\underline{z}_E$, but an approximation may be obtained by analogy with a spherically symmetric distribution of $\underline{c}_e$'s, i.e., by assuming the $\underline{c}_e$'s to be uniformly distributed on the hypersphere of radius $\sqrt{N-1}$ that is origin-centered and orthogonal to $\underline{1}$

If this assumption of spherical symmetry were true, it would follow that, for any $\underline{z}_E$, the distribution of $\cos^2(\underline{c}_e, \underline{z}_E - \bar{z}_E\underline{1})$ would be Beta(1/2, N/2-1) ([6] Theorem 12.2.3). It follows after some reduction, that $\cos(\underline{c}_E, \underline{z}_E - \bar{z}_E\underline{1})$ has a Pearson Type II distribution with density

$$f(x) = (1-x^2)^{(N-4)/2}/\text{Beta}(1/2, N/2-1), \quad -1 \le X \le +1 \qquad \text{(A.11)}$$

and moments

$$Ex^{2n-1} = 0 \qquad n=1,2,\ldots \qquad \text{(A.12.1)}$$

and

$$Ex^{2n} = \frac{(2n-1)(2n-3)\ \ldots\ 3\cdot 1}{(N+2n-3)(N+2n-5)\ldots(N+1)(N-1)} \qquad n=1,2,\ldots \ . \qquad \text{(A.12.2)}$$

From that, it further follows that $u = \sqrt{N-1}\cos(\underline{c}_e, \underline{z}_E - \bar{z}_E\underline{1})$, the co-ordinate of $\underline{c}_e$ in the direction of $\underline{z}_E - \bar{z}_E\underline{1}$, has moments, for n = 1,2,...,

$$Eu^{2n-1} = 0 \qquad \text{(A.13.1)}$$

and

$$Eu^{2n} = (2n-1)(2n-3)\ldots 3.1\ (N-1)^n/(N+2n-3)(N+2n-5)\ldots(N-1). \qquad \text{(A.13.2)}$$

As N increases, these moments rapidly approach those of the standard Normal. Hence, one may use a Normal approximation to the re-randomization distribution of $\underline{c}_e'\underline{z}_E$. We write this

$$(\underline{c}_e'\underline{z}_E \mid e \in E) \sim N(0, (N-1)v_{z_E}^2). \qquad \text{(A.14)}$$

Now, observations $\underline{z}_E$ of experiment E yield statistics

$$\underline{c}_e'\underline{z}_E = \underline{c}_e'\underline{\zeta} + \delta\underline{c}_e'\underline{d}_E, \qquad e \in E \qquad \text{(A.15)}$$

and, in particular

$$\underline{c}_E'\underline{z}_E = \underline{c}_E'\underline{\zeta} + \delta(N-1)v_E \ . \qquad \text{(A.16)}$$

Arguing as above,

$$(\underline{c}_E'\underline{\zeta} \mid E \in E) \sim N(0, (N-1)v_{\underline{\zeta}}^2) \qquad \text{(A.17)}$$

and thus

$$(\underline{c}_E'\underline{z}_E \mid E \in \mathcal{E}) \sim N(\delta(N-1)v_E,\ (N-1)v_{\underline{\zeta}}^2) \tag{A.18}$$

For testing H_o: $\delta = 0$ against H_+: $\delta > 0$, one may use the upper α point of re-randomization distribution (A.14). This may be written as $(\underline{c}'\underline{z}_E)_{1-\alpha}$ and is, approximately

$$(\underline{c}'\underline{z}_E)_{1-\alpha} = z_{1-\alpha} v_{\underline{z}_E} \sqrt{N-1} \ . \tag{A.19}$$

The rejection region of an approximately α-level test thus becomes

$$\underline{c}_E'\underline{z}_E \geq z_{1-\alpha} v_{\underline{z}_E} \sqrt{N-1} \ . \tag{A.20}$$

To evaluate the power of that test for a given $\delta(>0)$ one may use (A.16) to rewrite that rejection region as

$$\underline{c}_E'\underline{\zeta} \geq z_{1-\alpha}\ v_{\underline{z}_E} \sqrt{N-1} \ - \delta(N-1)v_E \ . \tag{A.21}$$

Now

$$(N-1)v_{\underline{z}_E}^2 = (N-1)v_{\underline{\zeta}}^2 + (N-1)\delta^2 v_E^2 + 2\delta v_E\ \underline{c}_E'\underline{\zeta} \quad , \tag{A.22}$$

since (A16) can, using (A.4), be rewritten as

$$\underline{z}_E - \bar{z}_E\underline{1} = (\underline{\zeta} - \bar{\zeta}\underline{1}) + \delta v_E \underline{c}_E . \tag{A.23}$$

Rearranging (A.22) and introducing

$$\Delta = \delta v_E / v_{\underline{\zeta}} \tag{A.24}$$

yields

$$(N-1)v_{\underline{z}_E}^2 = (N-1)v_{\underline{\zeta}}^2(1+\Delta^2)\ \{1 + 2\underline{c}_E'\underline{\zeta}\ \Delta/(N-1)v_{\underline{\zeta}}(1+\Delta^2)\} \ . \tag{A.25}$$

Approximating the square root of the right hand side by using the first two terms of the Taylor expansion of the last term, yields

$$\sqrt{N-1}\, v_{z_E} \approx \sqrt{N-1}\, v_{\underline{\zeta}}(1+\Delta^2)^{1/2}(1+\underline{c}_E'\underline{\zeta}\Delta/(N-1)v_{\underline{\zeta}}(1+\Delta^2)), \tag{A.26}$$

that is

$$\sqrt{N-1}\, v_{z_E} \approx \sqrt{N-1}\, v_{\underline{\zeta}}(1+\Delta^2)^{1/2} + \underline{c}_E'\underline{\zeta}\, \Delta\{(N-1)(1+\Delta^2)\}^{-1/2} . \tag{A.27}$$

Using this, the rejection region (A.21) becomes approximately

$$\underline{c}_E'\underline{\zeta}(1-z_{1-\alpha}\Delta\{(N-1)(1+\Delta^2)\}^{-1/2}) \geq z_{1-\alpha}\sqrt{N-1}\, v_{\underline{\zeta}}(1+\Delta^2)^{1/2} - \delta(N-1)v_E , \tag{A.28}$$

that is

$$\frac{\underline{c}_E'\underline{\zeta}}{v_{\underline{\zeta}}\sqrt{N-1}} \geq \frac{z_{1-\alpha}(1+\Delta^2)^{1/2} - \Delta\sqrt{N-1}}{1 - z_{1-\alpha}\Delta/\sqrt{N-1}\sqrt{(1+\Delta^2)}} . \tag{A.29}$$

The right-hand side of (A.29) depends on E through Δ of (A.24). If the reference set E is such that for some v_E

$$v_E = v_E \qquad \forall\ E \in \mathcal{E} , \tag{A.30}$$

then

$$\Delta = \delta v_E / v_{\underline{\zeta}} , \tag{A.31}$$

independently of E and the right hand side of (A.29) is constant. Since the left hand side of (A.29) is, by virtue of (A.17) a standard Normal variable, the probability of (A.29), i.e., the power of the α-level test at true effect coefficient δ, becomes

$$1 - \beta_\alpha(\delta;\mathcal{E},\underline{\zeta}) = 1 - \Phi\left(\frac{z_{1-\alpha}\sqrt{(1+\Delta^2)} - \Delta\sqrt{(N-1)}}{1-z_{1-\alpha}\Delta/\sqrt{(N-1)}\sqrt{(1+\Delta^2)}}\right) \tag{A.32}$$

with the dependence on δ, E and ζ being expressed by Δ of (A.31).

For fixed δ, and hence fixed Δ, the argument of Φ will go to $-\infty$ as N becomes large and thus the power will tend to one. On the other hand, if one considers a sequence of δ's decreasing as N increases in

such a manner that

$$\rho = \Delta\sqrt{N-1} \tag{A.33}$$

remains constant, the argument of Φ will approach $z_{1-\alpha}-\rho$ as N increases. Thus, for such a sequence of effects 6

$$1 - \beta_\alpha(\delta;\underline{E},\underline{\zeta}) \to 1 - \Phi(z_{1-\alpha}-\rho). \tag{A.34}$$

The naive approximation to power is calculated by comparing the "alternative distribution" $(\underline{c}'_E\underline{z}_E \mid E\varepsilon\underline{E})$ with the "null distribution $(\underline{c}_E\underline{\zeta} \mid E\varepsilon\underline{E})$. Using the Normal approximations to these distributions ((A.17) and (A.18), respectively) and notation (A.31) the naive power is seen to be

$$1 - b_\alpha(\delta;\underline{E},\underline{\zeta}) = 1 - \Phi(z_{1-\alpha}-\Delta\sqrt{N-1}) \tag{A.35}$$

i.e., to be equal to the limit (A.34) of the exact power for large N and alternatives with constant $\rho = \Delta\sqrt{N-1}$

For finite N, and $a \leq 0.5$, it is readily seen that the argument of Φ in (A.35) is less than or equal to that in (A.32). Hence

$$1 - b_\alpha \geq 1 - \beta_\alpha \tag{A.36}$$

confirming that the naive power calculation is likely to exceed the exact evaluation of power.