Geometric Concerns Pertaining to Applications of Statistical Tests in the Atmospheric Sciences

PAUL W. MIELKE, JR.

Departments of Statistics and Atmospheric Science, Colorado State University, Fort Collins, CO 80523

(Manuscript received 13 July 1984, in final form 16 January 1985)

ABSTRACT

This paper is concerned with the application of well-known statistical methods (e.g. matched-pairs *t*-test, two-sample *t*-test, one-way analysis of variance and significance test of Pearson's correlation coefficient) in the atmospheric sciences. This concern results from the fact that these statistical methods are based on a complex nonintuitive geometry which does not correspond with the perceived Euclidean geometry of the data intended to be analyzed. The real and artificial examples of this paper demonstrate how these commonly used statistical methods yield conclusions which may contradict rational interpretations by investigators. The geometric problem underlying these well-known statistical methods is their dependence on a peculiar distance measure defined between all pairs of measurements (this distance measure does not satisfy the triangle inequality condition of metric spaces, e.g., the familiar Euclidean space). Alternative statistical methods are suggested which overcome this geometric problem.

1. Introduction

A primary goal of numerical modeling in the atmospheric sciences is to describe physical phenomena in a realistic manner. Similarly, it is tacitly assumed by most atmospheric science investigators that commonly used statistical tests provide measures of differences among data sets in a meaningful manner. Unfortunately, many commonly used statistical methods (e.g. matched-pairs *t*-test, two-sample *t*-test, one-way analysis of variance, and significance test of Pearson's correlation coefficient) depend on a very perplexing geometry.

This paper is intended to 1) identify the geometric problem underlying these well known statistical methods and 2) suggest alternative statistical methods which circumvent this geometric problem. Since both commonly used and alternative statistical methods considered here are embedded in a broader class of statistical procedures, descriptions of both geometric concerns and this broader class are given in Section 2. The effect of these geometric concerns on statistically based conclusions is illustrated in Section 3 with both real and artificial examples. Specifically, the real example of Section 3 involves precipitation data associated with a weather modification experiment.

Incidentally, the statistical methods considered in this paper are termed data dependent permutation procedures. The variability of these procedures is governed by the assumption that any one of the possible permutations of the actual data in question will occur with an equal chance. A test's ability to detect different alternatives depends on the test statistic's structure. In contrast, the variability of a parametric test is governed by both the test statistic's structure and the distribution (e.g. normal, lognormal, gamma, kappa, beta or Weibull) assumed to represent the data in question. The alternative hypothesis of a parametric test is usually specified by changes attributed to either a location or scale parameter. Since a simple analytic description of neither the distribution nor the alternative associated with the complex differential treatment effects on data of a meteorological experiment is seldom if ever possible, applications of parametric tests in the atmospheric sciences must be severely scrutinized.

2. Preliminaries and geometric concerns

The broader class of statistical procedures which includes both the commonly used and alternative statistical techniques is termed multiresponse permutation procedures (MRPP). While a more general description of MRPP is given elsewhere (Mielke, 1984), the following description involving univariate (single response) data satisfies the present purpose.

Let $\Omega = \{\omega_1, \ldots, \omega_N\}$ denote a finite population of N objects, let x_I designate a response measurement associated with object ω_I ($I = 1, \ldots, N$), and let $S_1, \ldots, S_g$ be an exhaustive partitioning of the N objects comprising Ω into g disjoint groups. Also, let $\Delta_{I,J}$ be a symmetric distance measure based on the response measurements associated with objects ω_I and ω_J. The MRPP statistic is given by

$$\delta = \sum_{i=1}^{g} C_i \xi_i,$$

where

$$\xi_i = \frac{2}{n_i(n_i - 1)} \sum_{I<J} \Delta_{I,J}\Psi_i(\omega_I)\Psi_i(\omega_J)$$

is the average distance measure value for all distinct pairs of objects in group S_i ($i = 1, \ldots, g$), $n_i \geqslant 2$ is the number of objects classified *a priori* (e.g., according to g treatments) in group S_i ($i = 1, \ldots, g$), $N = \sum_{i=1}^{g} n_i$, $g \geqslant 2$, $\sum_{I<J}$ is the sum over all I and J such that $1 \leqslant I < J \leqslant N$, $C_i > 0$ ($i = 1, \ldots, g$), $\sum_{i=1}^{g} C_i = 1$, and $\Psi_i(\omega_I)$ is 1 if ω_I belongs to S_i and 0 otherwise.

The null hypothesis for MRPP assigns an equal probability to each of the

$$M = N!/(\prod_{i=1}^{g} n_i!)$$

distinct allocations of the N objects to the g groups. The collection of all δ values associated with these M equally likely allocations is the permutation distribution of δ under the null hypothesis.

If $C_i = n_i/N$ (i.e. the simple proportion of objects in group S_i) for $i = 1, \ldots, g$ and $\Delta_{I,J} = |x_I - x_J|$ (i.e. ordinary Euclidean distance), then a small value of δ indicates a concentration of response measurements within the g groups. The purpose of the example given in Section 2 of Mielke *et al.* (1981a) is to provide a very simple description of the concept underlying MRPP. The P-value associated with an observed value of δ (say δ_0) is the probability under the null hypothesis given by $P(\delta \leqslant \delta_0)$. While an efficient algorithm to calculate the exact P-value exists (Berry, 1982), this approach is unreasonable when M is large (e.g., $M > 100\,000$). Thus a P-value approximation based on the exact mean, variance and skewness of δ under the null hypothesis (μ_δ, σ_δ^2 and γ_δ respectively) is used in applications involving large values of M (Mielke, 1984). Under the null hypothesis, the distribution of δ usually involves substantial negative skewness for small, moderate or large values of M (Brockwell *et al.*, 1982; Mielke, 1978; Mielke *et al.*, 1976; Robinson, 1983). To compensate for the negative skewness, the distribution of the standardized test statistic given by

$$T = (\delta - \mu_\delta)/\sigma_\delta$$

is approximated by the Pearson type III distribution having mean 0, variance 1 and skewness $\gamma = \gamma_\delta$ (Mielke, 1984; Mielke *et al.*, 1981a).

The choice of the symmetric distance measure ($\Delta_{I,J}$) determines the *analysis space* of MRPP. For example, consider the symmetric distance measures given by

$$\Delta_{I,J} = |x_I - x_J|^v$$

where $v > 0$. Because $\Delta_{I,J}$ is a Euclidean distance when $v = 1$, the corresponding analysis space of MRPP is an ordinary Euclidean space. If $v = 2$, $x_1 = 4$, $x_2 = 6$ and $x_3 = 7$, then $\Delta_{1,2} = 4$, $\Delta_{1,3} = 9$, $\Delta_{2,3} = 1$ and $\Delta_{1,2} + \Delta_{2,3} < \Delta_{1,3}$ (i.e. the triangle inequality condition $\Delta_{1,2} + \Delta_{2,3} \geqslant \Delta_{1,3}$ of a metric space is not satisfied). Thus the analysis space corresponding to $v = 2$ is a complex nonmetric space. The collection of observed response measurements ($x_1, \ldots, x_N$) upon which any comparisons are made is the *data space.* Since the data space (i.e., the visualized collection of response measurements in question) is perceived as a Euclidean space, the analysis space of MRPP and the data space are congruent only when $v = 1$ (Mielke and Berry, 1983). It should be noted that a data space may involve either observed or transformed response measurements since either set of data is visualized in a Euclidean space (naturally an observed or transformed collection of values will differ in their visualized appearance). If the analysis space of a statistical technique and the data space are congruent, then the *congruence principle* is satisfied. If the congruence principle is not satisfied, then there exists no basis to expect agreement between visual comparisons based on displayed data and the analytic comparisons of a statistical method. The geometric concerns of this paper involve statistical methods which do not satisfy the congruence principle. For example, the permutation version of one-way analysis of variance (the two-sided two-sample t-test when $g = 2$) is a special case of MRPP when $C_i = (n_i - 1)/(N - g)$ and $v = 2$ (Mielke *et al.*, 1982). Thus the permutation version of one-way analysis of variance does not satisfy the congruence principle. Similar geometric concerns involving statistical methods such as the matched-pairs t-test and the significance test of Pearson's correlation coefficient are considered elsewhere, along with alternative statistical methods which satisfy the congruence principle (Mielke, 1984). The examples of Section 3 demonstrate that major differences in conclusions may occur between those statistical methods which do and those which do not satisfy the congruence principle.

3. Examples and discussion

The two examples in this section are intended to demonstrate that substantial differences may be obtained in the conclusions of statistical methods which do and do not satisfy the congruence principle. The first example involves real data of a weather modification experiment. The second example involving artificial data is given to further clarify why contradictory results are achieved by two statistical methods in the first example.

The first example is based on actual data of the Climax I and II wintertime orographic cloud seeding experiments (Mielke *et al.*, 1971, 1981b). Let

$$D = \text{TGM} - \text{CM}$$

denote the difference between target and control values for each experimental unit. A complete description of the target group mean (TGM) and the control mean (designated CM in this paper) are given in Subsection 3c and the Appendix of Mielke *et al.* (1981b). Table 1 contains the 109 nonseeded and 108 seeded values of D for the combined Climax I and II experiments when the estimated 500 mb temperature over Climax is greater than or equal to −20°C. The version of MRPP characterized by $g = 2$, $v = 2$ and also $C_i = (n_i - 1)/(N - 2)$ is the permutation version of the two-sided two-sample t test and does not satisfy the congruence principle. The nonseeded versus seeded comparison P-value using this squared Euclidean distance ($v = 2$) test is 0.086 for the Table 1 values. In contrast, the version of MRPP characterized by $g = 2$, $v = 1$ and $C_i = n_i/N$ is a permutation test which does satisfy the congruence principle. The nonseeded versus seeded comparison P-value using this Euclidean distance ($v = 1$) test is 0.026 for the Table 1 values.

Thus a major difference in the two P-values occurs for the Table 1 values. One obvious feature involving the values of Table 1 is that there exists a very large value of D (0.709) among the nonseeded values. This large value of D is not a questionable value since it is well within the range of natural variability (i.e., infrequent occurrences of such values during a long sequence of events are expected). While this large value situation occurred in conjunction with a weather modification experiment, similar situations are not uncommon in other types of experiments in the atmospheric sciences and many other disciplines as well. While a permutation test satisfying the congruence principle is certainly affected by a large value, a test such as the permutation version of the two-sample t-test which does not satisfy the congruence principle is often overwhelmed by a single value (even when relatively large sample sizes are involved). Thus the routine practice of selecting a well-known test to analyze experimental results in advance of an experiment could easily doom the experiment for the wrong reason (i.e., the choice of a commonly used test which may be overwhelmed by a very few values). This concern did not occur with the Climax I and II experiments simply because of an early decision by the investigators to use rank tests (Mielke *et al.*, 1971, 1981b). Since the largest value is transformed into the largest rank order value, the effect of such a value is diminished in the most commonly used rank tests.

TABLE 1. Ordered values of D = TGM − CM for 109 nonseeded and 108 seeded experimental units of the combined Climax I and II experiments when estimated 500 mb temperatures are greater than or equal to −20°C.

Nonseeded cases				Seeded cases			
−0.343	−0.019	0.000	0.056	−0.208	0.000	0.000	0.118
−0.282	−0.019	0.000	0.057	−0.112	0.000	0.001	0.125
−0.196	−0.019	0.000	0.064	−0.098	0.000	0.002	0.128
−0.156	−0.015	0.000	0.076	−0.087	0.000	0.008	0.147
−0.139	−0.014	0.000	0.084	−0.080	0.000	0.013	0.168
−0.128	−0.011	0.000	0.095	−0.079	0.000	0.015	0.169
−0.113	−0.010	0.000	0.097	−0.074	0.000	0.021	0.170
−0.111	−0.010	0.000	0.110	−0.070	0.000	0.024	0.174
−0.108	−0.009	0.000	0.113	−0.066	0.000	0.025	0.183
−0.095	−0.008	0.000	0.126	−0.050	0.000	0.027	0.208
−0.084	−0.007	0.000	0.149	−0.049	0.000	0.033	0.213
−0.068	−0.007	0.000	0.157	−0.029	0.000	0.034	0.226
−0.066	−0.005	0.000	0.182	−0.026	0.000	0.036	0.234
−0.065	−0.003	0.000	0.186	−0.026	0.000	0.038	0.251
−0.059	−0.003	0.000	0.224	−0.026	0.000	0.043	0.251
−0.058	−0.001	0.000	0.284	−0.019	0.000	0.047	0.263
−0.051	−0.001	0.000	0.376	−0.014	0.000	0.050	0.336
−0.049	−0.001	0.000	0.427	−0.013	0.000	0.055	0.351
−0.048	0.000	0.012	0.709	−0.013	0.000	0.056	
−0.043	0.000	0.013		−0.009	0.000	0.061	
−0.039	0.000	0.021		−0.008	0.000	0.065	
−0.034	0.000	0.022		−0.008	0.000	0.065	
−0.034	0.000	0.033		−0.004	0.000	0.065	
−0.030	0.000	0.034		−0.001	0.000	0.069	
−0.030	0.000	0.035		0.000	0.000	0.070	
−0.029	0.000	0.039		0.000	0.000	0.077	
−0.025	0.000	0.041		0.000	0.000	0.077	
−0.025	0.000	0.045		0.000	0.000	0.082	
−0.024	0.000	0.046		0.000	0.000	0.108	
−0.023	0.000	0.053		0.000	0.000	0.112	

Since it is impossible to obtain the actual observed values from the rank order values, a tremendous amount of information may be lost by this transformation. However, if a test statistic satisfies the congruence principle, then the need to transform the actual observed values into rank order values may be moot. The next example involves artificial data and demonstrates that the overwhelming influence of a single value is dominated by the choice of v and has little to do with the present choices of C_i. The reason for using $C_i = (n_i - 1)/(N - g)$ is simply because the derivation of the one-way analysis of variance statistic depends on estimates of unknown parameters (a vacuous reason in the context of permutation tests).

The second example is based on the four artificial data sets (A, B, C and D) presented in Table 2. Each of the four data sets in Table 2 involves $g = 2$, $n_1 = n_2 = 15$ and $N = 30$. Data set A employs two groups of values, S_1 and S_2, where most of the larger values among the 30 combined values belong to S_2. Data set B(C) is identical to data set A except that one value among the 15 values of $S_1(S_2)$ is shifted. Data set D is identical to data set A except that two values are shifted, one value from S_1 and the other value from S_2. The three P-values associated with each data set in Table 3 are the P-value of MRPP with $v = 1$, the P-value of MRPP with $v = 2$, and the P-value of the two-sided two-sample t-test (i.e., a parametric test based on the normal distribution). Since $C_i = \frac{1}{2}$ for either n_i/N or $(n_i - 1)/(N - 2)$, the choice of C_i is eliminated as an issue in this example. The three P-values associated with data set A are essentially the same (i.e. all are very small). While all P-values associated with data sets B, C and D are larger, the P-values of MRPP with $v = 1$ are roughly two to three orders of magnitude less than the other two P-values. Since the P-value of MRPP with $v = 2$ (permutation version of the two-sided two-sample t-test) and the P-value of the two-sided two-sample t-test are about the same for data sets B, C and D, the enormous P-value differences are dominated by the geometric issue (i.e., the well known robustness issue concerning differences between the second and third P-values plays a trivial role). As previously indicated, a few values being relatively very large and/or very small are often anticipated with meteorological data sets (a typical example is the data of Table 1). The results of Table 3 show that the use of statistical methods which do not satisfy the congruence principle will often lead to erroneous statistical conclusions.

TABLE 2. Frequencies of S_1 and S_2 values for data sets A, B, C and D where $g = 2$, $N = 30$ and $n_1 = n_2 = 15$.

	Data set							
	A		B		C		D	
Value	S_1	S_2	S_1	S_2	S_1	S_2	S_1	S_2
16.3	0	0	0	0	0	1	0	1
18.5	1	0	1	0	1	0	1	0
18.6	4	0	4	0	4	0	4	0
18.7	6	1	5	1	6	1	5	1
18.8	3	3	3	3	3	2	3	2
18.9	1	4	1	4	1	4	1	4
19.0	0	5	0	5	0	5	0	5
19.1	0	2	0	2	0	2	0	2
19.7	0	0	1	0	0	0	1	0

TABLE 3. Two-sided P-value comparisons based on MRPP with $v = 1$, MRPP with $v = 2$, and the two-sample t-test for data sets A, B, C and D.

Data set	MRPP		Two-sample t-test
	$v = 1$	$v = 2$	
A	1.8×10^{-5}	2.8×10^{-5}	3.1×10^{-6}
B	2.1×10^{-4}	0.029	0.042
C	1.0×10^{-4}	0.95	0.71
D	9.2×10^{-4}	1.00	1.00

Acknowledgments. This work has been supported by the National Science Foundation, under Grant ATM-84-07543. The author appreciated constructive criticisms of an earlier version of this paper by the reviewers.

REFERENCES

Berry, K. J., 1982: Algorithm AS 179: Enumeration of all permutations of multi-sets with fixed repetition numbers. *Appl. Statist.*, **31**, 169–173.

Brockwell, P. J., P. W. Mielke and J. Robinson, 1982: On non-normal invariance principles for multi-response permutation procedures. *Austral. J. Statist.*, **24**, 33–41.

Mielke, P. W., 1978: Clarification and appropriate inferences for Mantel and Valand's nonparametric multivariate analysis technique. *Biometrics*, **34**, 277–282.

——, 1984: Meteorological applications of permutation techniques based on distance functions. *Handbook of Statistics, Vol. 4: Nonparametric Methods*, P. R. Krishnaiah and P. K. Sen, Eds., North-Holland, Amsterdam, 813–830.

——, and K. J. Berry, 1983: Asymptotic clarifications, generalizations, and concerns regarding an extended class of matched pairs tests based on powers of ranks. *Psychometrika*, **48**, 483–485.

——, L. O. Grant and C. F. Chappell, 1971: An independent replication of the Climax wintertime orographic cloud seeding experiment. *J. Appl. Meteor.*, **10**, 1198–1212; Corrigendum: **15**, 801.

——, K. J. Berry and E. S. Johnson, 1976: Multi-response permutation procedures for *a priori* classifications. *Commun. Statist. Theor. Meth.*, **A5**, 1409–1424.

——, —— and G. W. Brier, 1981a: Applications of multi-response permutation procedures for examining seasonal changes in monthly mean sea-level pressure patterns. *Mon. Wea. Rev.*, **109**, 120–126.

——, G. W. Brier, L. O. Grant, G. J. Mulvey and P. N. Rosenzweig, 1981b: A statistical reanalysis of the replicated Climax I and II wintertime orographic cloud seeding experiments. *J. Appl. Meteor.*, **20**, 643–659.

——, K. J. Berry and J. G. Medina, 1982: Climax I and II: Distortion resistant residual analyses. *J. Appl. Meteor.*, **21**, 788–792.

Robinson, J., 1983: Approximations to some test statistics for permutation tests in a completely randomized design. *Austral. J. Statist.*, **25**, 358–369.