Effect of Antipsychotic Withdrawal on Extrapyramidal Symptoms: Statistical Methods for Analyzing Single-Sample Repeated-Measures Data

Stephan Arndt, Ph.D., Charles S. Davis, Ph.D., Del D. Miller, Pharm.D., M.D., and Nancy C. Andreasen, M.D., Ph.D.

Using symptom severity ratings of extrapyramidal side effects (EPS) during a 4-week antipsychotic washout period, we illustrate particular problems associated with repeated measures of symptom severity and demonstrate four analysis methods. The often suggested analysis of variance and multivariate analysis of variance found no mean change in weekly Simpson Angus scores over the 4-week washout despite the fact that 43% had clinically significant EPS prior to drug discontinuation. On the other hand, the Friedman Analysis of Ranks and Cochran-Mantel-Haenszel (CMH) statistics found significant change over the washout period. These two less well-known techniques place fewer restrictions on the data, can be more sensitive to patterns of change, and may be more appropriate for psychiatric data. The CMH method is particularly attractive since it does not require complete data on all subjects as do many other techniques. This minimizes the number of cases lost to missing data and increases the generality of the results.
[Neuropsychopharmacology 8:67–75, 1993]

KEY WORDS: *Antipsychotic withdrawal; Extrapyramidal symptoms; Statistical methods; Repeated measures*

Researchers in psychiatry often study the processes involved with changes in clinical status. Since the most natural means of studying changes in symptom status or other states (e.g., blood levels) is to follow subjects over time, repeated-measures studies are widely used. Reviewing 343 articles in four psychiatry journals, Ekstrom et al. (1990) found that about one in five used repeated-measures data.

In addition to being a straightforward method to assess change, repeated-measurement designs offer several advantages over other methods (Baltes and Nesselroade 1979). In spite of these advantages, repeated-measures studies frequently encounter particular problems (Magnusson and Bergman 1990). For instance, the dependence among successive observations made on the same subject complicates the data analysis. Ekstrom et al. (1990) recently expressed concern over the relative appropriateness of two statistical techniques frequently used for psychiatric repeated-measures data: the repeated-measures analysis of variance (ANOVA) and the multivariate analogue (MANOVA). Among those articles that provided sufficient description to ascertain the method used (n = 41), Ekstrom et al. (1990) found that about one-half (21) of the analyses of repeated-measures data were done with either ANOVA or MANOVA. Both the MANOVA and ANOVA share a variety of wide-reaching and restrictive assumptions. Although we will discuss several assumptions in the context of symptom severity ratings, many of the problems and all of the alternatives we present apply to a

From the Division of Biostatistics, Department of Preventive Medicine and Environmental Health, and Department of Psychiatry, University of Iowa, Iowa City, Iowa.

Address reprint requests to: Dr. Stephan Arndt, MH-CRC JPP, University of Iowa Hospitals and Clinics, 200 Hawkins Drive, Iowa City, Iowa 52242-1057.

Received November 13, 1991; revised February 6, 1992; accepted February 11, 1992.

Published by Elsevier Science Publishing Co., Inc.
655 Avenue of the Americas, New York, NY 10010

0893-133X/93/$5.00

wide range of data. Many of these assumptions are often inappropriate for the data of interest to psychiatric researchers.

Symptom severity measures illustrate a common problem, that response variables often are not normally distributed, nor would we expect them to be, based on the nature of symptom severity or the population of interest. However, normality is a requirement of most parametric statistical techniques including ANOVA and MANOVA. Many people may be asymptomatic or very nearly so, either because the disease process is not present, has not involved the relevant system, or is in remission. The measurement instrument or rating scale of severity should aptly mark these individuals as such, by assigning them a zero value. On the other hand, subjects affected with symptoms may exhibit a wide variety of symptom severity scores, ranging to the scale's limit. Thus, we may not even anticipate that these data are normally or even symmetrically (i.e., balanced around a center point) distributed.

A good example of the aforementioned is ratings of drug-induced parkinsonism (DIP) which is manifested as tremor, rigidity, and akinesia individually or in combination; drooling, festinating gait, oily skin, dysarthria, and dysphagia may accompany the symptoms. Incidence estimates of DIP vary widely, ranging from 2.2% to 56% of persons receiving antipsychotic drugs (Ayd 1961; Korczyn and Goldberg 1976; Sheppard and Merlis 1967; Tarsy 1983). Much of the variation in the reported percentages may be explained by differences in the antipsychotic medication prescribed, length of treatment, dose of antipsychotic, individual characteristics such a sex and age, and definitions of extrapyramidal side effects (EPS). Generally, these reactions are more common with high-potency than low-potency antipsychotics and tend to occur most often in the elderly, with females being twice as likely to develop them (Korczyn and Goldberg 1976; Sheppard and Merlis 1967; Man 1973). Obviously, in any group of patients treated with antipsychotic drugs, there will be patients with few or no symptoms of DIP; when a rating scale such as the Simpson Angus (SA) scale (Simpson and Angus 1970) for EPS is used, they will have a score of less than 3. Those patients who develop DIP will have a wide range of severity with SA scores ranging from 3 to 40. A change in antipsychotic drug dose may be associated with a change in symptom severity, but there is a large amount of variability in the amount of change. Upon discontinuation of the antipsychotic medication, DIP symptoms generally decline and often resolve in 7 to 10 days. However, it may take several weeks to months for complete resolution, depending upon the drug, the dose, and the patient. During this time there can be large fluctuations in symptom severity.

There is a belief and some evidence that ANOVA and other parametric techniques are robust to violations of the normal distribution assumption. One might take this to imply that violations of the assumptions are of little concern. However, there is scant empirical evidence that this capacity to withstand incorrect assumptions extends to the case of repeated measurements. Furthermore, Micceri (1989) recently questioned the validity of the studies supporting the notion that parametric analyses are robust. Most of these studies used computer simulations based on specific nonnormal distributions; however, the nonnormal distributions chosen in the computer simulation studies do not necessarily characterize the kinds of data seen in psychiatric and behavioral research. Although the effects of violating the normality assumption for repeated-measures data are largely unknown, the consequences of departures from the covariance assumptions of repeated-measures ANOVA have been studied extensively. When the covariance assumptions are violated, the ANOVA F test of the repeated-measures factor will tend to be too liberal, producing significant results too often (Box 1954a; Hearne et al. 1983). Tests of specific contrasts using general error terms are even more unstable (Boik 1981) and can be either positively or negatively biased.

Measures of symptom severity may violate another psychometric assumption of the commonly used parametric techniques: that the errors of measurement are independent and uncorrelated with the true status of the subject. For instance, there may be a high degree of interrater agreement on subjects who do not have symptoms of DIP but less agreement may exist for cases with moderate or high SA ratings. If the errors of measurement are not constant across the scale, the error is "dependent on the scale." This violates a basic assumption of many parametric statistical techniques including ANOVA and MANOVA that the errors are independently and identically distributed.

Severity of symptoms may pose further problems for the analysis of repeated measures. The stability of a symptom's severity over time may relate to the level of symptom severity. Measurement issues aside, a subject without symptoms of DIP during antipsychotic treatment has a high probability of being asymptomatic during a drug washout, relative to the likelihood of a person with symptoms of DIP maintaining that same level of severity throughout the drug-free period. Subjects with less severe symptoms would show more symptom stability than people with severe symptoms. Higher variability between Time 1 and Time 2 for subjects with more severe symptoms produces the "fan-tailed" type of relationship shown in Figure 1. This type of relationship, another example of variance dependent on the scale, also violates the assumption of homogeneity of variance found in many parametric techniques, including both ANOVA and MANOVA. Al-

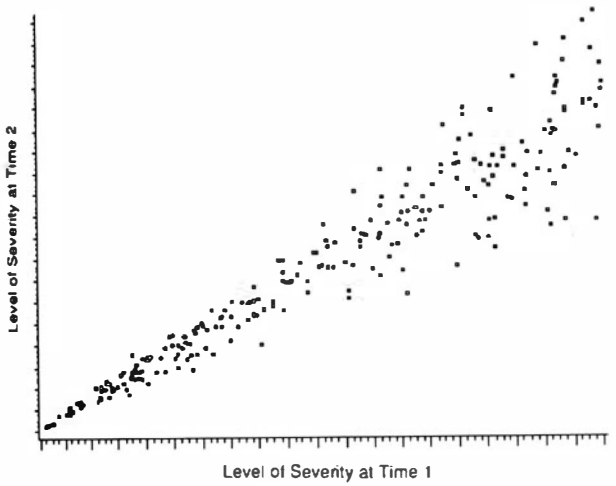

Figure 1 Hypothetical fan-tailed distribution of Time 1 and Time 2 symptom severity.

though this may be seen as a problem in the scale, it most probably reflects, in large part, true manifestations and variability of the symptoms. When the usual assumptions of parametric techniques are not compatible with the reality of symptom severity data, alternatives to both ANOVA and MANOVA become more attractive. Techniques with less restrictive assumptions about the symptom severity measurements may be more appropriate than the common parametric methods. We will discuss two such methods which require only that the measurement or rating scale can order people or times from the least to the worst symptom severity.

In this paper, we demonstrate and discuss four methods of analyzing repeated measures from a single sample. Table 1 depicts the general layout of this design. Two of the analysis approaches (ANOVA, MANOVA) are the most commonly used in psychiatric research (Ekstrom et al. 1990). Another method (the Friedman Analysis of Ranks) requires fewer assumptions. We also show a general methodology based on Cochran-Mantel-Haenszel statistics (CMH) applicable to longitudinal data (Agresti 1990; Mantel and Haenszel 1963; Mantel 1963). The CMH approach is particularly attractive since it does not require complete data on all subjects as do the three other techniques.

Table 1. General Layout of a One-Sample Repeated-Measures Study

	Time Point					
Subject	**1**	**2**	. . .	**j**	. . .	**T**
1	y_{11}	y_{12}		y_{1j}		y_{1T}
⋮						
i	y_{i1}	y_{i2}		y_{ij}		y_{iT}
⋮						
N	y_{N1}	y_{N2}		y_{Nj}		y_{NT}

The data we analyze are fairly typical of repeated-measures designs in psychiatric research. Ratings of the SA scale (Simpson and Angus 1970) of DIP severity were measured weekly over a 4-week antipsychotic medication washout period during which change in symptom severity was expected. Missing data is present in 21% of the subjects followed over a 5-week period ($T = 5$). The general null hypothesis is that the distribution of the ratings is the same at each of the 5 weeks.

SUBJECTS AND METHODS

Subjects. Forty-three patients who underwent a 4-week antipsychotic medication washout as part of a protocol for the University of Iowa Mental Health Clinical Research Center participated in this study. All patients met DSM III-R criteria for schizophrenia and had received treatment with antipsychotic drug prior to the study. Individuals who had received depot antipsychotics within the previous 6 months or had coexisting medical problems were excluded.

Procedures. After an initial observation of 2 to 3 days, antipsychotic medications were tapered and discontinued over a 2- to 4-day period (mean 3.2 days) depending on the dose of antipsychotic on which the patient was originally maintained (i.e., patients on higher doses underwent a longer tapering period).

Clinical Assessments. Trained research nurses made five weekly assessments on the SA scale for EPS. This scale rates 10 aspects (e.g., elbow rigidity, arm dropping) from 0 to 4 (normal to extremely symptomatic) and yields a score ranging from 0 to 40. Based on the work by Simpson and Angus (1970), a rating of 3 or greater was defined as clinically significant EPS. The first observation was a baseline rating, prior to discontinuation of antipsychotic medication, with three subsequent measures taken during the washout period. Missing data were present in nine patients due to a variety of reasons. For instance, one patient contracted chicken pox and returned home for 1 week.

RESULTS

Of the 43 participants, 30 were males and 13 were females with a mean age of 32.7 ± 9.7 years (range 22 to 56 years). Sixteen patients had been taking haloperidol (mean dose 29.6 mg/day, range 5 to 85 mg/day), eleven had been taking thiothixene (mean dose 33.6 mg/day, range 5 to 80 mg/day), five had been taking fluphenazine (mean dose 24.4 mg/day, range 2 to 60 mg/day), three had been taking trifluoperazine (mean

dose 23.0 mg/day, range 4 to 50 mg/day), two had been taking chlorpromazine (mean dose 200 mg/day), two received molidone (mean dose 50.0 mg/day, range 25 to 75 mg/day), one had been taking lozapine (dose 100 mg/day), and three patients were taking combinations of two different antipsychotics (i.e., one thioridazine and fluphenazine, one thioridazine and trifluoperazine, and one chlorpromazine and trifluoperazine).

Weekly means, medians, 75th percentiles, standard deviations, the mean difference, and paired *t*-tests comparing each week with the previous week are given in Table 2. The medians, 75th percentiles, and means indicate a general trend toward lower ratings in the following weeks. As might be expected, these data are not normally distributed. Since most patients do not exhibit DIP symptoms and some do, most ratings are very low, between 0 and 2. The distribution tapers off quickly but has a long tail. The relatively large standard deviations are produced by the skew in the data, occasional SA ratings between 10 and 24; thus, the data do not approximate a normal distribution and, in fact, have the expected shape as described in the introduction of this paper.

The pattern of the stability of scores was also of interest. We suggested that subjects with the lowest levels of symptom severity could be more stable from one measurement time to another compared to subjects displaying moderate or higher levels. Indeed, 12 of the 13 people (92.3%) rated as having no symptoms at Week 1 were so rated again at Week 2. People with initial ratings larger than zero seldom had the same rating at Week 2; only 6 of 29 people (21%) received the same score. We took the intrasubject standard deviation over time as an index of a subject's symptom stability, large numbers reflecting more week-to-week variation. That index correlated with the mean severity (Spearman $r = 0.91$, $p < 0.0001$) showing a dependence between stability and level of severity. Subjects' range of scores correlated similarly. Figure 2 shows the variability as each person's range of ratings ordered by their overall mean severity for the period. Higher symptom severity predicted greater variability over the period.

The Parametric Techniques: ANOVA and MANOVA

We will ignore the nonnormality and other violations of assumptions for the moment and proceed with the most often used tests for time-related differences, a repeated-measures ANOVA and MANOVA. Both analyses require complete data for all subjects and so use data for only 34 of our 43 patients.

The null hypothesis for the ANOVA is that the means at each measurement period are the same, that is, H_0: $\mu_1 = \mu_2 \ldots = \mu_T$. In addition to assuming that the data are normally distributed, repeated-measures ANOVA requires assumptions concerning the correlation structure of the repeated measures (Huynh and Feldt 1970). These assumptions are not likely to be satisfied when the measures are taken over time. A sufficient but not necessary condition to fulfill this assumption requires equal variances and covariances across the time periods. As a result of the equal variance-covariance condition, the correlations of the SA ratings between any two time periods should be the same. For instance, Week 1 SA ratings should correlate with Week 2 ratings about as much as Week 1 with Weeks 4 or 5. This is called the compound-symmetry, sphericity, or homogeneity of variance-covariance assumption and is frequently incorrect for longitudinal data (Ekstrom et al. 1990; Poor 1973; McCall and Appelbaum 1973).

Table 2. Weekly Simpson Angus Means, Standard Deviations, Mean Difference from the Preceding Week and Associated Paired *t*-Value

	Weeks				
	1	2	3	4	5
Complete data only (n = 34):					
75th percentile value	6.00	5.00	4.00	3.00	2.00
Median	2.00	1.00	0.50	1.00	1.00
Mean	3.32	2.76	2.59	2.50	1.94
Standard deviation	3.59	4.42	4.25	4.65	3.94
Difference from last week		−0.559	−0.176	−0.090	−0.559
t-Value (df = 33)		−1.05	−0.21	−0.17	−0.68
All Data					
n	42	42	42	39	36
75th percentile value	4.00	4.00	3.00	3.00	2.50
Median	1.50	0.50	0.00	1.00	1.00
Mean	2.90	2.36	2.21	2.41	2.06
Standard deviation	3.41	4.10	3.94	4.53	3.88
Difference from last week		−0.548	−0.146	0.128	−0.629
t-Value		−1.25	−0.21	0.25	−0.79
df		40	39	37	33

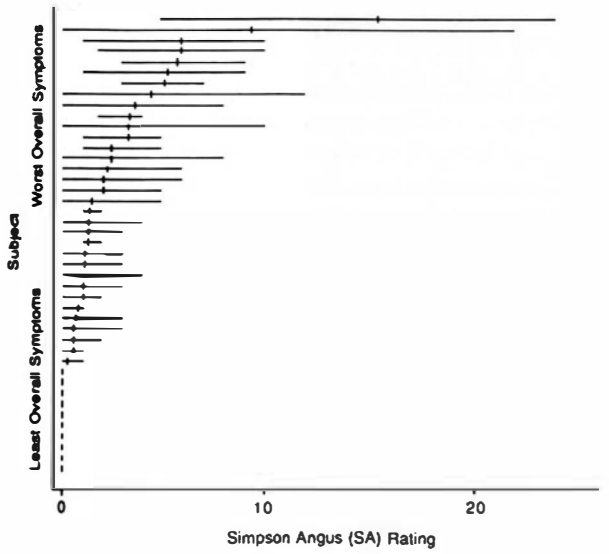

Figure 2 Individuals ordered from worst to least mean Simpson Angus (SA) rating. Horizontal bar is the individual's range of ratings over the 5 weekly ratings.

Violations of this assumption do affect the significance levels of the F statistics (Box 1954a; 1954b). Although there are procedures for correcting the ANOVA results, they usually result in an overly conservative test (Huynh and Feldt 1976; Geisser and Greenhouse 1985; Wallenstein and Fleiss 1981).

The parametric alternative, the MANOVA analogue of ANOVA (or the special case of Hotelling's T^2) treats the data only slightly differently (Finn 1974; Bock 1975; Milliken and Johnson 1984). Essentially, ANOVA questions whether the means at each time are different from one another. The MANOVA questions if any mean difference between repeated-measurement periods is zero. Differences between times are constructed and considered as the variates of interest. The test for a time effect is whether the multivariates have a mean zero (H_0; $\mu = 0$).

Because of the multivariate nature, no assumptions are made about equal covariances or correlations among the variates. This has led some authors to recommend MANOVA over ANOVA. However, the relaxation of the equal variance-covariance assumption has costs (Bock 1975; Milliken and Johnson 1984; Lavori 1990). The MANOVA assumes that the data come from a multinormal distribution and is frequently restricted by requiring more subjects than time points. The MANOVA also has less power than ANOVA when the ANOVA assumptions are correct.

For this sample, the F test for mean differences from the ANOVA was not significant, ($F[4, 132] = 0.91$, $p > 0.38$). Although Spearman correlations would be more appropriate to describe this set of data, the Pearson correlations shown in Table 3 are relevant to the assumptions of the ANOVA. These correlations show

Table 3. Pearson Correlations and p-Values in Parentheses of Simpson Angus Ratings for 5 Weeks ($n = 34$)

Week	2	3	4	5
1	0.716	0.321	0.471	0.089
	(0.001)	(0.064)	(0.004)	(0.616)
2		0.346	0.678	0.067
		(0.045)	(0.001)	(0.707)
3			0.776	0.791
			(0.001)	(0.001)
4				0.391
				(0.022)

a marked deviation ($p < 0.0001$) from the compound-symmetry condition using a likelihood ratio test (Mauchly 1940; Rogan et al. 1979). Thus, the data likely violate an assumption necessary for the ANOVA. The MANOVA result, multivariate $F(4, 30) = 1.31$ ($p > 0.28$) for the SA ratings, was only somewhat larger than the ANOVA result. In practice, both the MANOVA and ANOVA frequently provide the same conclusion. Using the parametric techniques there is no evidence for significant changes occurring over time. Although the parametric analyses suggest no changes over time, the nonparametric analyses provide quite different results.

Nonparametric Method 1: The Friedman Two-Way Analysis of Ranks

The Friedman statistic may be a more appropriate analytic method for this data since no normality assumption is necessary and our data are distinctively nonnormal. The Friedman approach makes inferences about the relative intraindividual ranks of the time-period ratings rather than about their absolute magnitude. Intraindividual change is accounted for by ranking each subject's scores from the lowest to the highest severity period, yielding T ranks ranging from 1 to T. The null hypothesis is that there is no consistent buildup of higher or lower ranks at any measurement period. A tendency for the highest ranks (e.g., subjects' worst symptom severity) to appear in the 1st week would lead to a rejection of the null hypothesis. Formulas for computing the Friedman analysis and exact probabilities for small samples are given by Siegel (1956) and Siegel and Castellan (1988). Computationally, a standard repeated-measures ANOVA calculated on the within-subject ranks will also produce F statistics and excellent small sample p-value approximations (Conover and Iman 1976). Using this technique is preferable to the standard χ^2 when the number of times is larger than the number of subjects (Iman and Davenport 1980).

The average ranks for each of the five weekly ratings appear in Table 4 as do the frequencies of each rank. This table and the significance tests use the 34 subjects with complete data. The Friedman analysis produces

Table 4. Percentages of Each Week's Rank in Symptom Severity for Subjects with Complete Data

	Week				
Rank	1	2	3	4	5
1 (least symptoms)	3	12	12	0	12
1.5–2	21	18	24	29	32
2.5–3	21	32	32	38	41
3.5–4	18	15	18	21	9
4.5–5 (most symptoms)	38	24	15	12	6
Mean rank	3.6	3.0	2.9	3.0	2.5

Table 5. Layout of One Subject's $T \times L$ Contingency Table for the CMH Statistics Whose Scores Were 5, 2, 1, 3, and 2 for the 5 Weeks, Respectively

	Simpson Angus Scale Severity Rating				
Week	1	2	3	4	5=L
1	0	0	0	0	1
2	0	1	0	0	0
3	1	0	0	0	0
4	0	0	1	0	0
T=5	0	1	0	0	0

an $F(4, 132) = 2.61$, $p < 0.038$. Using the $\chi^2(4)$ approximation we find a value of 9.96, $p < 0.041$. Thus, we can conclude that there is a pattern to the relative rankings during the antipsychotic medication washout period. Judging from Table 4, the worst week (ranks 4.5 to 5) tended to occur at the beginning of the washout period and tapered off. Many subjects (38%) were experiencing their worst week at Week 1 and only 6% experienced Week 5 as their worst.

Nonparametric Method 2: Cochran-Mantel-Haenszel Statistics

Although Friedman's test does not require that the response variable is continuous or normally distributed, it only uses complete cases. We now describe a method based on the use of CMH statistics (Agresti 1990; Landis et al. 1988; Mantel and Haenszel 1963; Mantel 1963) for analyzing one-sample repeated measures. Unlike other techniques, the CMH method can readily include subjects with incomplete data. Of course, the reasons that data are missing need to be irrelevant to the study (e.g., random or happenstance). To be more precise, any missing data must be missing completely at random (Rubin 1976). Individual subjects are strata for this use of the CMH analysis. The use of CMH statistics is quite common in epidemiology; however, the application to repeated-measures data is less well known.

The general framework is as follows. Let N, T, and L denote the number of subjects, the number of time periods, and the number of possible levels of the response variable, respectively. The resulting data are summarized in N $T \times L$ contingency tables, one table for each of the N subjects. For instance, one subject in our sample received the SA ratings of 5, 2, 1, 3, and 2 for the five time periods, respectively. The contingency table for this individual depicts "1's" for row (week) 1 and column (score) 5, row (week) 2 and column (score) 2, and so on. Zeros fill the remainder of the table (Table 5). In this example, L, the number of possible outcomes on the rating scale, is relatively small; however, there is no restriction on the number of possible outcomes in the variable. Thus, the CMH methodology described in this paper is equally applicable to continuous outcome variables.

The CMH statistics are summary test statistics for the independence of the ranked ratings (rows) from time (columns). Statistical inference is based on the multiple hypergeometric distribution; the row and column margin totals in each table are fixed. The only assumptions are independence between subjects and an order to the ratings (e.g., worst to least severe). Rejection of the null hypothesis indicates that the ranks tend to change as a function of the time period. Two versions of the test are commonly available: 1) a test sensitive to a monotonic correlation between the scores and time periods analogous to a Spearman correlation, and 2) a test for the time period differences in mean ranks analogous to a Friedman Analysis of Ranks. Both test statistics are easily obtained using the FREQ procedure of SAS (1990).

For comparison with the previous analyses we first used the sample of subjects with complete data (n = 34). The CMH test for differences in mean ranks yields a value of 9.955 ($df = 4$, $p < 0.041$). This is numerically equivalent to the χ^2 approximation found with the Friedman analysis. There is also a substantial monotonic relation (CMH statistic = 7.726, $df = 1$, $p < 0.005$) between the SA ratings and time. This supports our interpretation of Tables 2 and 4 that generally subjects' worst symptoms occur early in the period and then decline in severity.

The fact that the CMH approach can accommodate subjects for whom we do not have complete data is a major advantage. The CMH statistics do not require that all weeks (columns) be complete in the $T \times L$ contingency tables. In all of the previous analyses, only 34 cases with data present on all five occasions could be used. This gives 170 (i.e., 34 × 5) datapoints. If we include our nine subjects with missing data, the yield totals 201 valid observations, an 18% increase in the number of subjects. Based on all 43 subjects, the CMH statistic for differences increased to 12.153 ($df = 4$, $p < 0.016$) and for the monotonic correlation to 8.831 ($df = 1$, $p < 0.003$).

DISCUSSION

We have presented four different techniques for analyzing single-sample repeated-measures data, all of which are readily available in common statistical software (e.g., SAS 1990). The ANOVA, and to a lesser extent the MANOVA, are common procedures for analyzing repeated measures. Application of the Friedman analysis and CMH statistics to this situation is, however, relatively novel. We did not apply the various analyses in order to locate a significant result, rather we wished to demonstrate that the various methods each test subtly different hypotheses and each has its own benefits and deficits. It is oversimplifying to suggest that any one technique should be the standard method; however, the CMH or Friedman approach may often be more appropriate for analyzing repeated measures in psychiatric research. This is particularly true when the data do not meet the basic assumptions required of parametric analyses.

We found that 43% of the patients receiving a variety of antipsychotic medications had clinically significant DIP. Eighty-eight percent of our patients were receiving high-potency antipsychotics alone, which makes this rate consistent with previous reports (Ayd 1961; Korczyn and Goldberg 1976; Sheppard and Merlis 1967; Tarsy 1983). Although these symptoms tend to wax and wane over time, they generally resolve with discontinuation of antipsychotic medications. We assumed that since there was a substantial number of patients with DIP at baseline, the mean severity ratings would change significantly during a 4-week washout period. We were surprised that there was no mean change in weekly SA ratings over the 4-week washout period when analyzed by ANOVA and MANOVA. However, the repeated-measure ANOVA and MANOVA had gross violations of their assumptions occurring in the present dataset. Despite the restrictive assumptions, we included them for two reasons: 1) due to their popularity and recent suggestions to use MANOVA made by Ekstrom, et al. (1990) and others (Poor 1973; McCall and Appelbaum 1973; Keselman and Rogan 1980); and 2) to demonstrate that they can often be an inappropriate choice for psychiatric data. Paired *t*-tests are a special case for both of these methods so that some of the issues raised in this discussion apply to this simple test statistic as well.

These parametric techniques assume normal distributions for the dependent variable. Nonnormal distributions are fairly typical of symptom severity ratings in general. As noted in the discussion of Figure 1, many subjects had very low ratings, with 57% presenting asymptomatic or nearly so. Possibly, for those individuals displaying DIP, the scores are more variable and thus some appear as "outliers," but only in the context of a hypothetical normal distribution. The effect of this kind of observed distribution is that the mean moves away from the area of highest density, becoming less characteristic. Rather than question the validity of extreme scores to make the data more normal, it may be more appropriate to suspect the meaningfulness of calculating arithmetic means of ratings. The same suspicion about the mean's usefulness can be applied to the tests for differences among the means using ANOVA, MANOVA, or *t*-tests.

The simple ANOVA further made particular assumptions about the variances and the covariances, and hence correlations, among all of the measurement periods (i.e., the compound-symmetry condition). This latter assumption is often violated as in the present dataset since measures taken close together in time tend to be more highly correlated than measures that are more distal in time. More generally, this is true when measures are more or less spatially distant as in electroencephalograph leads placed closer or further apart on the scalp. Some journals have established editorial policy (e.g., *Psychophysiology*, Jennings 1987) requiring authors reporting repeated-measures data to account for these possible violations. This concern, however, is expressed over one particular aspect of the time dependency of subjects' scores. We speculated that repeated measures of symptom severity often violate other basic assumptions of both ANOVA and MANOVA. At least with the present data, the stability of ratings over time was dependent on the level of severity. For our sample data, subjects with more severe symptoms were less stable, as we expected. Discussing whether MANOVA is more appropriate than ANOVA is probably a less critical question than the more basic choice between the parametric linear models and the nonparametric alternatives.

For the present data, the effect of going from parametric to nonparametric techniques was dramatic. Although the weekly mean differences were nonsignificant using the ANOVA and MANOVA, there was substantial significance with the Friedman and CMH statistics. The dataset we used is fairly representative of symptom severity measurements. Thus it is clear that for this particular set of data, the nonparametric tests were more sensitive to changes occurring over time.

Choosing between the two nonparametric methods depends on the flexibility required when dealing with something other than a single-sample design with complete data. A Friedman-like analysis (Conover and Iman 1976, 1981; Iman 1974; Brunner 1991) can be performed when there are between-subjects factors, at present there is no similar extension of the CMH approach. This design is frequently of interest, as when case control groups are compared over time; however, like the ANOVA or MANOVA, present extensions of the Friedman analysis require complete data on all subjects.

On the other hand, the CMH method's ability to

accept cases with incomplete data is a powerful advantage. Losing an entire subject because of a missing datapoint is a high price to pay and may be unacceptable for any number of reasons. For instance, if costly or invasive procedures are involved, it is vital to maintain the data already obtained. A subgroup that consists of only "perfect completers" may be very atypical of the group as a whole, substantially reducing the generalizability of the results. Statistical power is also severely compromised. Providing that the data is not missing due to any factor related to the study, the CMH approach effectively minimizes the information loss and maintains generalizability by retaining as many subjects as possible. To our knowledge, repeated-measures studies in psychiatry have not exploited this aspect of the CMH methodology.

For univariate outcome variables that follow specific nonnormal distributions, such as binomial, Poisson, and gamma, generalized linear models are useful. Extensions to repeated measures have also been recently developed (Liang and Zeger 1986). However, these extensions focus on the marginal relationship between a response variable and one or more covariates, rather than on trends over time. In addition, it is still necessary to assume a particular parametric form for the univariate distribution of the response at each timepoint.

A simple but widely-used alternative that was not considered in this paper is response-feature analysis (Crowder and Hand 1990). The basic idea is to replace the repeated measures for each individual subject by a single-summary statistic indicating their trajectory or changes. Thus, a multivariate analysis is reduced to a univariate analysis. Response-feature analysis is useful not only in one-sample situations, but extends easily to between-subject factors. For these data, the within-person association of score with time could be summarized by a single measure of association such as a regression slope or the Spearman rank correlation coefficient. The resulting measures could then be analyzed for a nonzero mean or differences between groups. Although this principle is very similar to the use of the CMH correlation statistic, the CMH mean score statistic offers the advantage of being able to detect nonmonotonic relationships.

ACKNOWLEDGMENTS

This study was funded in part by National Institute of Mental Health MH46011, Grant #5 P50 MH43271, MH 40856, and MH31593.

REFERENCES

Agresti A (1990): Categorical Data Analysis. New York, John Wiley & Sons

Ayd FJ (1961): A survey of drug-induced extrapyramidal reactions. JAMA 175:1054–1060

Baltes PB, Nesselroade JR (1979): History and rationale of longitudinal research. In Nesselroade JR, Baltes PB (eds) Longitudinal Research in the Study of Behavior and Development. New York, Academic Press, pp 1–39

Bock RD (1975): Multivariate Statistical Methods in Behavioral Research. New York, McGraw-Hill

Boik RJ (1981): A priori tests in repeated measures designs: effects of nonsphericity. Psychometrika 46:241–255

Box GEP (1954a): Some theorems on quadratic forms applied to the study of analysis of variance problems. II: Effects of inequality of variance and of correlation between errors in the two-way classification. Ann Math Stat 25:484–498

Box GEP (1954b): Some theorems on quadratic forms applied to the study of analysis of variance problems. I: Effects of inequality of variance in the one-way classification. Ann Math Stat 25:290–302

Brunner E (1991): A nonparametric estimator of the shift effect for repeated observations. Biometrics 47:1149–1153

Crowder MJ, Hand DJ (1990): Analysis of Repeated Measures. London, Chapman & Hall

Conover WJ, Iman RL (1976): On some alternative procedures using ranks for the analysis of experimental designs. Commun Stat (A) 5:1348–1368

Conover WJ, Iman RL (1981): Rank transformations as a bridge between parametric and nonparametric statistics. Am Stat 35:124–133

Ekstrom D, Quade D, Golden RN (1990): Statistical analysis of repeated measures in psychiatric research. Arch Gen Psychiatry 47:770–772

Finn JD (1974): A General Model for Multivariate Analysis. New York, Holt, Rinehart and Winston

Geisser S, Greenhouse SW (1958): An extension of Box's results on the use of the F distribution in multivariate analysis. Ann Math Stat 29:885–891

Hearne EM, Clark GM, Hatch JP (1983): A test for serial correlation in univariate repeated-measures analysis. Biometrics 39:237–243

Huynh H, Feldt LS (1970): Conditions under which mean squares ratios in repeated measures designs have exact F-distributions. J Am Stat Assoc 65:1582–1589

Huynh H, Feldt LS (1976): Estimation of the Box correction for degrees of freedom from sample data in the randomized block and split plot designs. J Educ Stat 1:69–82

Iman RL (1974): A power study of a rank transform for the two way classification model when interactions may be present. Can J Stat 2:227–239

Iman RL, Davenport JM (1980): Approximations of the critical region of the Friedman statistic. Commun Stat (A) 9:571–595

Jennings JR (1987): Editorial policy on analyses of variance with repeated measures. Psychophysiology 24:474–475

Keselman HJ, Rogan JC (1980): Repeated measures F tests and psychophysiological research: Controlling the number of false positives. Psychophysiology 17:499–503

Korczyn A, Goldberg GJ (1976): Extrapyramidal effects of neuroleptics. J Neurol Neurosurg Psychiatry 39:866–869

Landis JR, Miller ME, Davis CS, Koch GG (1988): Some general methods for the analysis of categorical data in longitudinal studies. Stat Med 7:109–137

Lavori P (1990): ANOVA, MANOVA, my black hen: Comments on repeated measures. Arch Gen Psychiatry 47:755–778

Liang KY, Zeger SL (1986): Longitudinal data analysis using generalised linear models. Biometrika 73:13–22

Magnusson D, Bergman L (1990): Data Quality in Longitudinal Research. New York, Cambridge University Press

Man PL (1973): Long-term effects of haloperidol. Dis Nerv Syst 34:113–118

Mantel N (1963): Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. J Am Stat Assoc 58:690–700

Mantel N, Haenszel W (1963): Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 22:719–748

Mauchly JW (1940): Significance test for sphericity of a normal n-variate distribution. Ann Math Stat 11:204–209

McCall RB, Appelbaum MI (1973): Bias in the analysis of repeated-measures designs: Some alternative approaches. Child Dev 44:401–415

Micceri T (1989): The unicorn, the normal curve, and other improbable creatures. Psychol Bull 105:156–166

Milliken GA, Johnson DE (1984): Analysis of Messy Data. Volume 1: Designed Experiments. Belmont, CA, Lifetime Learning

Poor DDS (1973): Analysis of variance for repeated measures designs: Two approaches. Psychol Bull 80:204–209

Rogan JC, Keselman HJ, Mendoza JL (1979): Analysis of repeated measurements. Br J Math Stat Psychol 32: 269–286

Rubin DB (1976): Inference and missing data. Biometrika 63:81–92

SAS Institute Inc (1990): The FREQ procedure. In SAS Procedures Guide, Version 6 Third Edition. Cary, NC: SAS Institute

Sheppard C, Merlis S (1967): Drug-induced extrapyramidal symptoms: Their incidence and treatment. Am J Psychiatry 123:886–889

Siegel S (1956): Nonparametric Statistics for the Behavioral Sciences. New York, McGraw-Hill

Siegel S, Castellan NJ (1988): Nonparametric Statistics for the Behavioral Sciences, ed 2. New York, McGraw-Hill

Simpson GM, Angus JS (1970): A rating scale for extrapyramidal side effects. Acta Psychiatr Scand 212:11–19

Tarsy D (1983): Neuroleptic-induced extrapyramidal reactions: classification, description, and diagnosis. Clin Neuropharmacol 6:9–26

Wallenstein S, Fleiss JP (1981): Repeated measurements analysis of variance when the correlations have a certain pattern. Psychometrika 44:229–233