The use and interpretation of the Friedman test in the analysis of ordinal-scale data in repeated measures designs

MICHAEL R. SHELDON and MICHAEL J. FILLYAW University of New England, Biddeford, Maine, USA
W. DOUGLAS THOMPSON University of Southern Maine, Portland, Maine, USA

ABSTRACT *The purpose of this paper is to review the use and interpretation of the Friedman two-way analysis of variance by ranks test for ordinal-level data in repeated measurement designs. Physical therapists frequently make three or more repeated measurements of the same individual to compare different treatments, or to assess the effect of a single treatment over time. When the measurements are ordinal-scaled, such as some ratings of functional status and muscle strength, statistical significance may be determined by the Friedman test. We illustrate the use of the Friedman test and a* post hoc *multiple comparison test with data from 27 subjects whose performance on a lifting task was rated on three occasions by use of an ordinal scale. We discuss the interpretation of ordinal-level data and recommend that therapists understand the limitations a measurement scale imposes on the inferences that can be made from these tests.*

Key words: Friedman test, ordinal-level data, repeated measurement design, scale of measurement.

INTRODUCTION

Physical therapists routinely compare measurements of attributes, behaviors, or aspects of performance of their patients or research subjects. When the measurements are interval- or ratio-scaled, the *F* statistic from an analysis of variance with a repeated measures model is appropriate to test the significance of the differences in group means. However, phyiscal therapists are often interested in attributes or behaviors of patients which are not interval- or ratio-scaled. Coordination (Schmitz, 1994), spasticity (Ashworth, 1964; Bohannon & Smith, 1987), muscle strength (Medical Research Council, 1943), angina and dyspnea (American College of Sports Medicine, 1995), standing balance (Shumway-Cook & Horak, 1986), or

functional status (Jette, 1980; Keith et al., 1987) are measured by ranking characteristics by use of distinct, ordinal scales. Although the ranks of the scales have the property of order and can be used to indicate whether the individual has more or less function, angina pain or muscle strength, the differences between ranks may not be equal. To the extent that the differences between the ranks of the scale are not equal, means and variances are in error (Siegel, 1956) and nonparametric procedures may be used for testing hypotheses.

Nonparametric tests, based on ranks, are alternatives to parametric statistics for testing hypotheses about relationships and differences for variables measured on an ordinal scale. Whereas many clinicians are probably familiar with the Wilcoxon matched-pairs signed-ranks test (Wilcoxon, 1945) for analyzing the difference between two correlated ordinal-level measurements, fewer may know the Friedman test (Friedman, 1937, 1940) for analyzing three or more repeated measurements of ordinal data. Because clinicians and researchers frequently make more than two repeated measurements on the same individual, physical therapists should be familiar with the Friedman test. The purpose of this paper is to review the Friedman test for analyzing three or more measures made on the same subjects from a single factor, repeated measures design, and to discuss the interpretation of the analysis of ordinal-scaled data. An example from physical therapy is used to illustrate the use of the Friedman test.

THE FRIEDMAN TEST

Friedman (1937, 1940) developed a procedure called the 'method of ranks' to test hypotheses related to ordinal-scaled data. (The Friedman test can also be used in place of the *F* test for repeated measures with interval- or ratio-level data that do not meet the assumptions of normality and homogeneity of variance and covariance (Ferguson, 1976).) Subsequently, the test has come to be known as the Friedman two-way analysis of variance by ranks (Ferguson, 1976; Portney & Watkins, 1993). The designation 'two-way' refers to the fact that subjects and treatments are considered separate independent variables in the analysis. The Friedman test is readily calculated by hand for small data sets and is available on several statistical software packages.

Procedure

Data from Sheldon (1994) were used to illustrate the use of the Friedman test. In this study, 27 elementary school students were taught a technique for lifting an empty crate from the floor. The lifting performance of each student was ranked on an ordinal scale on three occasions: before instruction on lifting (Pretest); two days after instruction (Post 1); and 6–7 weeks after instruction (Post 2). Students were expected to lift the crate keeping a straight back, flexing at the hips and knees, keeping the crate within 15 cm of the body and within the base of support, and without twisting. A student's score was determined by the number of components

observed during the lift. Therefore, scores ranged from 0 to 5. The students' scores are shown in Table 1 as a set of k=3 measurements (columns) for N=27 subjects (rows). Each student's raw scores are ranked (the average of the ranks is used when the raw scores are tied (Friedman, 1937)) and the ranked scores in each column are summed (R_j).

TABLE 1: Scores of lifting performance measured on three occasions

	Pretest		*Post 1*		*Post 2*	
Subject	*raw score**	*(ranked score)*	*raw score*	*(ranked score)*	*raw score*	*(ranked score)*
1	3	(1.5)	5	(3.0)	3	(1.5)
2	5	(2.0)	5	(2.0)	5	(2.0)
3	1	(1.0)	5	(2.5)	5	(2.5)
4	4	(1.0)	5	(2.5)	5	(2.5)
5	3	(1.0)	5	(3.0)	4	(2.0)
6	3	(1.5)	5	(2.0)	3	(1.5)
7	3	(1.0)	5	(3.0)	4	(2.0)
8	4	(1.0)	5	(2.5)	5	(2.5)
9	5	(2.0)	5	(2.0)	5	(2.0)
10	4	(2.0)	4	(2.0)	4	(2.0)
11	4	(1.0)	5	(2.5)	5	(2.5)
12	3	(1.0)	5	(2.5)	5	(2.5)
13	2	(1.0)	5	(2.5)	5	(2.5)
14	3	(1.0)	5	(2.5)	5	(2.5)
15	4	(2.0)	5	(3.0)	3	(1.0)
16	0	(1.0)	4	(2.0)	5	(3.0)
17	2	(1.0)	4	(2.0)	5	(3.0)
18	2	(1.0)	5	(2.5)	5	(2.5)
19	3	(1.0)	5	(3.0)	4	(2.0)
20	3	(1.0)	4	(2.5)	4	(2.5)
21	3	(1.0)	5	(3.0)	4	(2.0)
22	4	(1.0)	5	(2.5)	5	(2.5)
23	4	(1.5)	5	(3.0)	4	(1.5)
24	4	(1.0)	5	(2.5)	5	(2.5)
25	4	(1.0)	5	(2.5)	5	(2.5)
26	4	(1.0)	5	(2.5)	5	(2.5)
27	5	(2.5)	5	(2.5)	4	(1.0)
∑ ranks (Rj)		34.0**		69.0***		59.0†

*raw score (0 = lowest; 5 = highest). **$R_{Pretest}$; *** $R_{Post\ 1}$; †$R_{Post\ 2}$.

Hypotheses

Under the null hyposthesis (H_o:) the independent variable, instruction in lifting, is assumed to have no effect on the dependent variable, score on the performance of the lift; the three sets of scores come from the same population. Expressed symbolically, H_o: $R_{Pretest} = R_{Post\ 1} = R_{Post\ 2}$. If this is true, the sums of the ranks, R_j, will be

similar. Alternatively, if instruction changes lifting performance, the R_js will be different. Thus, the alternate hypothesis (H_a:) is that at least one set of scores is not from the same population.

Computation

The computational formula for the Friedman test (Friedman, 1937) is:

$$\chi^2_r = \frac{12}{Nk(k+1)} \sum_{j=1}^{k} R^2_j - 3N(k+1)$$

where k is the number of ranked observations or measurements (columns), N is the number of subjects (rows), and Rj is the sum of the ranked scores in each column. (Lehmann & D'Abrera (1975) provide an alternative formula to use when there are many tied ranks.) The numbers 12 and 3 are constants, not dependent on the number of subjects or experimental conditions. The test statistic χ^2_r is distributed according to the usual χ^2 distribution with k–1 degrees of freedom when the rankings are random, i.e. when the independent variable is having no effect on the dependent variable. As *N* and *k* increase, the approximation to the χ^2 distribution improves (Lehmann & D'Abrera, 1975). For small values of *N* and *k*, related computational formulae (Bradley, 1968; Ferguson, 1976; Daniel, 1990) and exact tables are available (Kendall & Smith, 1939; Bradley, 1968; Daniel, 1990).

Values from Table 1 are substituted into the formula to caluculate the test statistic χ^2_r:

$$\chi^2_r = \frac{12}{27\,(3)\,(4)} [(34)^2 + (69)^2 + (59)^2] - 324$$

$$\chi^2_r = 348.07 - 324$$

$$\chi^2_r = 24.07$$

The null hypothesis is rejected when χ^2_r is greater than the critical value of χ^2 for the chosen alpha level. In this example, the calculated χ^2_r of 24.07 is greater than 5.991, the critical value of χ^2 with two degrees of freedom for $p \leq 0.05$. Thus, we reject H_o: and conclude that the students' lifting performances were significantly different on the three days.

***Post hoc* analysis**

When the null hypothesis is rejected, nonparametric *post hoc* tests are available for determining where significant differences exist. One multiple comparison procedure suggested by Daniel (1990) determines significant differences by the formula:

$$\left| R_j - R_j' \right| \geq z \sqrt{\frac{Nk\,(k+1)}{6}}$$

where R_j and R_j' are the sums of ranks being compared (for example, $R_{Pretest}$ and

$R_{Post\ 1}$), N is the number of subjects, k is the number of measurements on each subject, and z is the z score from the standard normal curve corresponding to $\alpha/k\ (k–1)$. Based on an experiment-wise error rate of $\alpha = 0.05$, and $k = 3$, the value of z corresponding to $\alpha/k\ (k–1) = 0.05/3(2) = 0.008$ is 2.41.

From Table 1, $R_{Pretest} = 34.0$, $R_{Post\ 1} = 69.0$, and $R_{Post\ 2} = 59.0$. Inserting these values into the left side of the equation determines differences between the tests:

$$| R_{Pretest} - R_{Post\ 1} | = | 34.0 - 69.0 | = 35.0$$
$$| R_{Pretest} - R_{Post\ 2} | = | 34.0 - 59.0 | = 25.0$$
$$| R_{Post\ 1} - R_{Post\ 2} | = | 59.0 - 69.0 | = 10.0$$

By inserting the appropriate values into the right-hand side of the formula:

$$z \sqrt{\frac{Nk\,(k+1)}{6}}$$
$$= 2.41 \sqrt{\frac{27\,(3)\,(4)}{6}}$$
$$= 17.7$$

we see that any difference between sums of ranks greater than 17.7 is statistically significant. Because 35.0 and 25.0 are greater than 17.7, we note that significant differences exist between the lifting performances on the Pretest and Post 1 and between the lifting performances on the Pretest and Post 2. However, the difference in performance between Post 1 and Post 2 is not statistically significant. The results suggest that performance in lifting is improved by instruction in lifting technique. A significant improvement is apparent two days after instruction which is not abated within seven weeks.

ESTIMATING TREATMENT EFFECTS AND CLINICAL INTERPRETATION

In addition to statistical significance, researchers are also interested in estimating the magnitude of treatment effects. With large samples it is possible to detect statistically significant differences that may be of little practical interest. For parametric procedures, researchers estimate the value of a population parameter based on statistics from the sample. For example, the sample mean is the point estimate of the population mean and the confidence interval is an interval estimate that asserts, with a specified degree of confidence, the upper and lower values of the interval believed to include a population parameter. From these estimates, the researcher can decide what magnitude of treatment effects are statistically excluded on the basis of the empirical results (Thompson, 1987). With ordinal-scaled data, however, the validity of the mean and variance is questionable and the size of the treatment effect cannot be estimated by these statistics.

The median and the mode are appropriate point estimates for ordinal-scaled data, but because the numbers assigned to the ranks are somewhat arbitrary, it is difficult to interpret the size of treatment effects based on changes in these statistics. In the data from Sheldon (1994), the differences between scores on the scale of lifting performance are unknown and should not be assumed equal. If the scores on the scale had been 0, 10, 20, 30, 40 and 50, the median and mode would have changed from 30 at the Pretest to 50 at Post 1 and Post 2, instead of changing from 3 to 5, as occurred with the 0–5 scale. Although the change in performance appears much larger, by use of the 0–50 scale rather than the 0–5 scale, the effect of instruction is the same. Recall that the median is the middle score in a rank-ordered distribution and indicates the position, not amount, in a distribution; neither the median nor the mode are affected by the quantitative value of the scores (Portney & Watkins, 1993). One can only say that more components of the lift were observed after instruction in lifting. If, in the judgement of the clinician, the increase in scores denotes that lifting performance was more effective, efficient, or safer, for example, then the conclusion that the improvement was clinically important as well as statistically significant would be justified. With ordinal-scaled data, the onus is more on the clinician to interpret the importance of the treatment effects.

SCALES OF MEASUREMENT AND THE CHOICE OF TEST STATISTIC

There has been considerable discussion in the literature about whether it is necessary to consider the scale of measurement when choosing a statistic for analyzing ordinal data (Lord, 1953; Nunnally, 1967; Gaito, 1980; Townsend & Ashby, 1984; Barbeito & Simpson, 1991). Those who say that scale of measurement should not influence the choce of statistical tests (Nunnally, 1967; Gaito, 1980; Barbeito & Simpson, 1991) argue that hypothesis testing is concerned only with differences in numbers and that because 'the numbers don't remember where they came from' (Lord, 1953) ordinal data can be analyzed by parametric statistics. It is true that the probabilities associated with significance testing may be little affected whether ordinal data are analyzed by parametric or nonparametric procedures. When k=3, as in the example, the asymptotic efficiency of the Friedman test relative to the F test from a two-way analysis of variance with one observation per cell is 0.716 and increases to 0.955 for $k = \infty$ (Lehmann & D'Abrera, 1975; Ferguson, 1976).

However, statistical inference is concerned with more than the decision to reject the null hypothesis. Ultimately, inferences drawn from the results of statistical tests do depend on the measurement scale of the data. For example, if lifting performance was considered interval-ratio level data and analyzed by parametric statistics, and the mean increased from 2 to 4 in Clinic A and from 2 to 3 in Clinic B, Clinic A would have improved 50% more than Clinic B. If Clinic A used the 0–50 scale and the mean increased from 20 to 40 this would be a 10-fold greater improvement than an increase in the mean from 2 to 4 in Clinic B. In either case, one should not infer that Clinic A is more effective. In both examples, the intervals on the scales are assumed to be equal; in the second example, the scales differ by a magnitude of 10.

Thus, the differences between clinics could be attributed to these factors and not to superior treatment provided by Clinic A. Clinic A might receive more referrals or financial support for its programs to reduce low back injury by improving lifting mechanics, because of a misinference stemming from the parametric analysis of ordinal-level measurements. Merbitz et al. (1989) cautioned against basing decisions about the effectiveness of rehabilitation and allocation of resources on ordinal data which have been inappropriately analyzed.

Although it has been suggested that the problems of misapplication of mathematical operations and statistical inference would be lessened if ordinal-level scales were replaced by ratio-scaled measurements (Merbitz et al., 1989), clinicians will continue to use ordinal-level scales. Furthermore, replacing functional assessment scales with ratio-level measurements may not adequately capture the variation in patient functioning required to study outcome (Silverstein et al., 1989). Thus, whilst we agree that ratio-level measurements should be encouraged where appropriate, a prudent solution is for researchers to use the nonparametric statistical techniques appropriate for their research design when analyzing ordinal data, and to understand the limitations measurement scales impose on the inferences that can be drawn. In doing so, therapists will be in accord with the standards for measurement in physical therapy for tertiary purveyors of tests (teachers) and test users who 'must understand the different levels of measurement (i.e. nominal, ordinal, interval and ratio) and the mathematical operations that are appropriate for each level' (Task Force on Standards for Tests and Measurments in Physical Therapy, 1991).

CONCLUSION

The Friedman two-way analysis of variance by ranks test and the multiple comparisons procedure described in this paper are appropriate statistical tests to analyze ordinal-level data from a repeated measurement experimental design. Therapists need to understand the appropriate interpretation of ordinal-scale data as it relates to statistical significance and clinical importance.

REFERENCES

American College of Sports Medicine. Guidelines for Exercise Testing and Prescription (fifth edition). Baltimore: Williams & Wilkins, 1995.

Ashworth B. Preliminary trial of carisoprodol in multiple sclerosis. Practioner 1964; 192: 540.

Barbetio R, Simpson TL. Should level of measurement considerations affect the choice of statistic? Optometry and Vision Science 1991; 68: 236–242.

Bohannon R, Smith M. Interrater reliability of a modified Ashworth scale of muscle spasticity. Physical Therapy 1987; 67: 206–207.

Bradley JV. Distribution-free Statistical Tests. Englewood Cliffs, NJ: Prentice-Hall, 1968.

Daniel WW. Applied Nonparametric Statistics (second edition). Boston, MA: PWS–Kent Publishing Co., 1990.

Ferguson GA. Statistical Analysis in Psychology and Education (fourth edition). New York: McGraw-Hill, 1976.

Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association 1937; 32: 675–701.

Friedman M. A comparison of alternative tests of significance for the problem of *m* rankings. Annals of Mathematical Statistics 1940; 11; 86–92.
Gaito J. Measurement scales and statistics. Resurgence of an old misconception. Psychological Bulletin 1980; 87: 564–567.
Jette AM. Functional capacity evaluation: an emprical approach. Archives of Physical Medicine and Rehabilitation 1980; 61: 85–89.
Keith RA, Granger CV, Hamilton BB, Sherwin FS. The functional independence measure: a new tool for rehabilitation. In: Eisenberg MG, Grzesiak CS (eds). Advances in Clinical Rehabilitation (first edition). New York: Springer-Verlag, 1987, pp. 6–18.
Kendall MG, Smith BB. The problem of *m* rankings. Annals of Mathematical Statistics 1939; 10: 275–287.
Lehmann EL, D'Abrera HJM. Nonparametric Statistical Methods Based on Ranks. Oakland: Holden-Day Inc., 1975.
Lord FM. On the statistical treatment of football numbers. American Psychologist 1953; 8: 750–751.
Medical Research Council. Aids to the Investigation of Peripheral Nerve Injuries. London: HMSO, 1943, pp. 11–46.
Merbitz C, Morris J, Grip JC. Ordinal scales and foundations of misinference. Archives of Physical Medicine and Rehabilitation 1989; 70: 308–312.
Nunnally JC. Psychometric Theory. New York: McGraw-Hill, 1967.
Portney G, Watkins MP. Foundations of Clinical Research Applications to Practice. Norwalk: Appleton & Lange, 1993.
Schmitz TJ. Coordination assessment. In: O'Sullivan SB, Schmitz TJ (eds). Physical Rehabilitation Assessment and Treatment (third edition). Philadelphia: FA Davis, 1994.
Sheldon MR. Lifting instruction to children in an elementary school. Journal of Orthopaedic and Sports Physical Therapy 1994; 19: 105–110.
Shumway-Cook A, Horak F. Assessing the influence of sensory interaction on balance: suggestions from the field. Physical Therapy 1986; 66: 1548–1550.
Siegel S. Nonparametric Statistics for Behavioral Sciences. New York: McGraw-Hill, 1956.
Silverstein B, Kilgore K, Fisher W. Don't abandon FAS. Letter to the Editor. Archives of Physical Medicine and Rehabilitation 1989; 70: 864–865.
Task Force on Standards for Tests and Measurements in Physical Therapy. Standards for tests and measurements in physical therapy practice. Physical Therapy 1991; 71: 609.
Thompson WD. Statistical criteria in the interpretation of epidemiologic data. American Journal of Public Health 1987; 77: 191–194.
Townsend JT, Ashby FG. Measurement scales and statistics: the misconception misconceived. Psychological Bulletin 1984: 96: 394–401.
Wilcoxon F. Individual comparisons by ranking methods. Biometrics 1945; 1: 80–83.

Address correspondence to Michael R. Sheldon, MS, PT, Assistant Professor, Department of Physical Therapy, University of New England, 11 Hills Beach Road, Biddeford, Maine 04005, USA. E-mail: MSheldon@Mailbox.UNE.EDU