Journal of Data Science 12(2014), 415-436

Power of a Rank-Based Test for Differences Between Treatment Distributions in a Randomized Complete Block Design

Roy St. Laurent[1*], Philip Turk[2]

[1]*Northern Arizona University, Flagstaff, AZ*
[2]*Colorado State University, Fort Collins, CO*

Abstract: Friedman's test is a rank-based procedure that can be used to test for differences among *t* treatment distributions in a randomized complete block design. It is well-known that the test has reasonably good power under location-shift alternatives to the null hypothesis of no difference in the *t* treatment distributions. However the power of Friedman's test when the alternative hypothesis consists of a non-location difference in treatment distributions can be poor. We develop the properties of an alternative rank-based test that has greater power than Friedman's test in a variety of such circumstances. The test is based on the joint distribution of the *t*! possible permutations of the treatment ranks within a block (assuming no ties). We show when our proposed test will have greater power than Friedman's test, and provide results from extensive numerical work comparing the power of the two tests under various configurations for the underlying treatment distributions.

Key words: Friedman's test, goodness-of-fit, non-location shift, nonparametric test, power

1. Introduction

In this paper we develop the properties of a nonparametric test for differences among *t* treatment distributions in a randomized complete block design (RCB) with *b* blocks. The test statistic is based on the joint distribution of the *t*! possible orderings of the treatment ranks within a block, and is similar to the nonparametric test proposed by Friedman (1937) in its use of these ranks. Like Friedman's test, the assumptions necessary for the proposed test statistic to have a known, easily computed null distribution are less stringent than those required for the usual analysis of variance *F*-test. In particular, one need not assume a specific parametric family for the underlying treatment distributions.

Here we develop the test statistic X^2 by consideration of several inter-related null hypotheses, and show that the proposed test has better power than Friedman's test for detecting differences in treatment distributions under a variety of conditions.

* Corresponding author

2. Hypotheses

For each of t treatments, let X_{ij} denote the response to the j-th treatment in the i-th block, $j = 1,\ldots,t$ and $i = 1,\ldots,b.$ Let $\mathbf{X_i} = (X_{i1}, \ldots, X_{it})$ denote the t-vector of responses in the i-th block, assumed to have continuous joint distribution $\mathbf{F_i}(\mathbf{x})$. Denote the marginal distribution of X_{ij} by $F_{ij}(x)$.

Consider the null hypothesis that the t components of the vector of responses have identical marginal distributions:

$$FH_0: F_{i1}(x) = F_{i2}(x) = \cdots = F_{it}(x) \text{ for } i = 1,\ldots,b; \tag{1}$$

that is, within each block, observations from different treatment groups have the same distribution function. We call FH_0 the "Friedman null hypothesis."

In each block, rank the t responses from 1 to t (smallest to largest), and let R_{ij} denote the rank assigned to the j-th treatment response in the i-th block. Since the cumulative distribution function of each response is assumed to be continuous, it follows that the probability of a tie in rank between two or more treatments in a block is zero.

Following Quade (1984), we assume that blocks are independent and that all blocks have the same joint distribution of treatment ranks. If the only effect of blocks on the response is additive, then this condition will be met.

Letting $\bar{R}_{\bullet j}$ denote the average rank of the j-th treatment across the b blocks, Friedman's test statistic $Q = 12b/[t(t+1)]\sum_{j=1}^{t}[\bar{R}_{\bullet j} - \frac{1}{2}(t+1)]^2$ sums the squared deviations of the observed treatment average rank from the common expected value of the j-th treatment average rank under the assumption that all the treatment distributions are identical.

For small values of t and b, the exact distribution of Q has been tabled (see, for example, Friedman 1937, or Lehmann, 1975) or may be constructed with the aid of software (van de Wiel, 2004). Friedman (1937) showed that as $b \to \infty$, Q converges in distribution to $\chi^2_{(t-1)}$, a chi-square random variable with $(t-1)$ degrees of freedom. Iman & Davenport (1980) propose the general rule that an asymptotic approximation to the distribution of Q should not be used when $t = 3,$ and for $t > 3$ it should be used only when $t + b > 9.$

A null hypothesis different from the Friedman null given in (1) is:

$$H_0^R: E(R_{i1}) = E(R_{i2}) = \cdots = E(R_{it}), \text{ for } i = 1,\ldots,b; \tag{2}$$

that is, the expected rank of the j-th treatment is the same for $j = 1,\ldots,t$ (or equivalently $E(R_{ij}) = (t+1)/2$ for $j = 1,\ldots,t$). As Friedman's Q sums the squared deviations of the

observed average treatment ranks from their expectation under H_0^R in (2) above, *Q* is a direct test of this hypothesis: large values of *Q* support the complement to (2) that the expected values of the average treatment ranks are not the same. It is common practice for Friedman's *Q* to be applied in situations where interest is in testing equivalency of response means across treatments (St. Laurent & Turk, 2013). Even so, it is likely that in many applications, practitioners are unaware of the relationship between *Q* and the hypothesis in (2).

When H_0^R is not true, it can be shown that FH_0 also is not true (see Appendix 1). Thus large values of Friedman's *Q* can be considered evidence contradicting FH_0. However if the hypothesis in (2) is true, it is not necessarily the case that FH_0 is true. So failing to reject H_0^R via Friedman's test statistic *Q*, means only that there is insufficient evidence to support that there are differences in the expected treatment ranks. But that still allows for the possibility that the treatment distributions are not identical. It is in this sense *Q* is a direct test of (2) and an indirect test of (1). It is partially for this reason that we look for a more general, alternate approach to testing the hypothesis FH_0.

3. Alternate Test

3.1 Justification

For $i = 1, \ldots, b$ let $\mathbf{R_i} = (R_{i1}, R_{i2}, \ldots, R_{it})$ denote the random vector of ranks for the *t* treatments in the *i*-th block. For each *i*, the multivariate probability distribution of $\mathbf{R_i}$ has support on the set of all possible permutations of the vector of ranks $(1,2,\ldots,t)$. There are $s = t!$ such permutations, which we denote by $\pi_1,\ldots,\pi_s$. For each *i*, the probability that $\mathbf{R_i}$ is equal to π_k for any $k = 1,\ldots,s,$ is completely determined by the joint cumulative distribution function $\mathbf{F_i}(\mathbf{x})$ for the treatment response vector $\mathbf{X_i}$ For block i, let $p_{i,\pi_k} = P(\mathbf{R}_i = \pi_k),$ the probability that the vector of *t* treatment ranks in the *i*-th block matches the ordering of ranks in the *k*-th permutation. As we have assumed that the probability distribution of the ranks is identical across blocks, we write $p_{\pi_k} = P(\mathbf{R}_i = \pi_k)$ and let $\mathbf{p} = (p_{\pi_1},\ldots,p_{\pi_s})$ be the *s*-vector of probabilities for the $s = t!$ possible permutations, i.e., $\mathbf{p}$ is the same across all blocks (assumption IIa of Quade, 1984). Since $\sum_{k=1}^{s} P_{\pi_k} = 1,$ the set of all such vectors $\mathbf{p}$ is restricted to the standard $(s-1)$-simplex.

For example, when $t = 3$, the $s = 3! = 6$ permutations may be written $\pi_1 = (1,2,3),$ $\pi_2 = (1,3,2),$ $\pi_3 = (2,1,3),$ $\pi_4 = (3,1,2),$ $\pi_5 = (2,3,1),$ and $\pi_6 = (3,2,1).$ Given the joint cdf $\mathbf{F}_i(\mathbf{x})$ of the observations in the *i*-th block we can calculate p_{π_k}, e.g.,

$$p_{312} = p_{\pi_4} = P(\mathbf{R}_i = \pi_4) = P(R_{i1} = 3, R_{i2} = 1, R_{i3} = 2) = P(X_{i1} > X_{i3} > X_{i2}).$$

If the elements of $\mathbf{X}_i$ are exchangeable then each rank ordering of the treatments is equally likely and $\mathbf{p} = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$.

Appendix 3 includes several general results concerning the structure of the elements of **p** when the joint distribution of the random vector X exhibits some form of symmetry.

Now consider the hypothesis:

$$H_0^{\mathbf{p}}: p_{\pi_1} = p_{\pi_2} = \cdots = p_{\pi_s} = 1/s \tag{3}$$

meaning that within a block each permutation of the ranks is equally likely, and hence $p_{\pi_k} = 1/s$ for all permutations π_k. In general the three hypotheses (1), (2) and (3) are not equivalent. When the elements of the vector $\boldsymbol{X}_i$ are exchangeable, then FH_0 implies $H_0^{\mathbf{p}}$ which in turn implies H_0^R. However H_0^R does not imply $H_0^{\mathbf{p}}$, nor does $H_0^{\mathbf{p}}$ imply FH_0. In this sense, $H_0^{\mathbf{p}}$ is "closer" to FH_0 than is H_0^R. See Appendix 1 for a proof of this relationship.

Because of the relationships amongst these hypotheses, one might conjecture that a direct test of $H_0^{\mathbf{p}}$ would have greater power to detect departures from FH_0 than a direct test of H_0^R. Quade (1984) notes that the most general alternative to FH_0 that may be tested based on the rank vectors $\mathbf{R_1}, \ldots, \mathbf{R_b}$ is the complement of $H_0^{\mathbf{p}}$. In the remainder of this paper, we develop a direct test of $H_0^{\mathbf{p}}$, determine the properties of the test, and compare it to Friedman's test.

3.2 The Test Statistic

The random vector $\mathbf{R}_i$ has support on the $s = t!$ permutations $\pi_1, \ldots, \pi_s$. For $k = 1, \ldots, s$, let $Y_{ik} = \begin{cases} 1, & \mathbf{R}_i = \pi_k \\ 0, & \mathbf{R}_i \neq \pi_k \end{cases}$. Then the s-vector $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{is})$ is distributed $Mult_s(n = 1, \mathbf{p})$, i.e., $\mathbf{Y}_1, \ldots, \mathbf{Y}_b$ are independent, identically distributed s-dimensional multinomial random vectors with number of trials $n = 1$ and vector of cell probabilities $\mathbf{p} = (p_{\pi_1}, \ldots, p_{\pi_s})$. Therefore $\mathbf{M} = \sum_{i=1}^{b} \mathbf{Y}_i \sim Mult_s(n = b, \mathbf{p})$, with corresponding probability function $P(\mathbf{M} = \mathbf{m} \mid \mathbf{p}) = b! \prod_{k=1}^{s} \frac{p_{\pi_k}^{m_k}}{m_k!}$, where $\mathbf{m} = (m_1, \ldots, m_s)$ is the s-vector of observed counts, and $\sum_{k=1}^{s} m_k = b$.

When FH_0 is true then $H_0^{\mathbf{p}}$ is true and $\mathbf{p} = \mathbf{p}_0 \equiv \frac{1}{s}(1, \ldots, 1)$, in which case $E(\boldsymbol{M}) = b\mathbf{p_0}$ This suggests using a goodness-of-fit statistic for the s-dimensional multinomial distribution of

M as a method of indirectly testing FH_0, but getting us "closer" to Friedman's hypothesis than the test statistic *Q* proposed by Friedman.

One possibility is the chi-square goodness-of-fit statistic $X^2 = \sum_{k=1}^{s} \frac{(O_k - E_k)^2}{E_k}$, where O_k and E_k are, respectively, the observed and expected counts in the *k* th cell. In our application this simplifies to

$$X^2 = \sum_{k=1}^{s} \frac{(m_k - b/s)^2}{b/s} \tag{4}$$

since under H_0^{P}, $E_k = bp_{\pi_k} = b/s$.

Wormleighton (1959) develops the asymptotic properties of a "hierarchy of tests of permutation symmetry" of a *t*-variate distribution (as extensions of the familiar sign test for $t = 2$). The author notes that the test statistics *Q* and X^2 can be thought of as being at the two extremes of this hierarchy, with *Q* being a test of low order symmetry and X^2 being a test of high order symmetry. Wormleighton did not explore small sample properties of the X^2 test, nor did he consider its power for alternatives to H_0^{P}. Wormleighton's work has subsequently received scant attention in the literature. Quade (1984) briefly mentions X^2 including its asymptotic null distribution, but does not focus on the small sample properties of the test.

Rayner & Best (2001, ch. 6) consider the relative merits of Page's test, Anderson's test, *Q* and X^2 in testing for differences in treatment distributions in a randomized complete block design, and the relationships between these tests. They note that under the assumption of no difference in treatment distributions (FH_0), each of these test statistics is asymptotically distributed chi-square with degrees of freedom 1, $(t-1)$, $(t-1)^2$ and $(t!-1)$ respectively. In choosing which test to apply, Rayner & Best (1990) suggest that "…better tests were those whose degrees of freedom matched the dimensions of the alternative hypothesis."

Based on standard results concerning the chi-square goodness-of-fit test in multinomial sampling, when H_0^{P} is true: $E(X^2) = s-1$; $Var(X^2) = 2(s-1)(b-1)/b$; and for fixed *s*, as $b \to \infty$, the statistic X^2 converges in distribution to $\chi^2_{(s-1)}$, a chi-square random variable with $(s-1)$ degrees of freedom (Pearson, 1900).

In many if not most applications, we can expect that *b* will be small, possibly quite small relative to the corresponding rules of thumb for suggested use of the asymptotic chi-square reference distribution. By one such rule-of-thumb (Koehler & Larntz, 1980), under the uniform null hypothesis, each expected cell count b/s should be greater than $\sqrt{10/s}$, or equivalently, $b > \sqrt{10s} = \sqrt{10t!}$, provided that $b \geq 10$ For $t = 3$ this suggests an experiment with at least 10 blocks; for $t = 4$, at least 16 blocks; and for $t = 6$, at least 85 blocks. While this requirement on the number of blocks is not unrealistically large for $t = 3$ or 4 treatments, nonetheless in practice

it is useful to consider the exact, small sample distribution of these test statistics, via complete enumeration or simulation, particularly when *b* is small and *t* is greater than 4.

3.3 Enumeration of the Exact Sampling Distribution of the Test Statistic

The small sample properties, including the exact distribution, of both X^2 and *Q* depend entirely on *t*, *b* and the multinomial vector of probabilities **p**.

Starting from (4), the goodness of fit statistic simplifies to $X^2 = \frac{s}{b}\sum_{k=1}^{s} m_k^2 - b = \frac{s}{b}\mathbf{m}'\mathbf{m} - b,$ a quadratic function of **m**. Similarly *Q* may be expressed as a quadratic function of **m** and hence the distributions of *Q* and X^2 are both completely determined by the distribution of **m** (see Appendix 2). Taking advantage of this relationship, the authors have written a script, Joint.R, in the statistical software package R (version 2.15.2), to compute by complete enumeration the exact joint distribution of *Q* and X^2, as well as each of their marginal distributions, provided user input of *t*, *b*, and **p**. When $t = 3$, Joint.R can be used to construct the null distributions of these statistics under FH_0, and non-null distributions for any specific alternative to FH_0 (provided that the joint distribution of treatment ranks is the same in each block), for any value of *b*. The alternative of interest is determined by specification of the non-constant vector **p**. Joint.R may also be used when $t = 4$ or 5 to calculate the exact distributions of these statistics, but due to the computationally intensive nature of the calculations, realistically, only for very small values of *b*, e.g., $b < 6$. Joint.R is available upon request from the authors.

4. Power Comparisons

Rayner & Best (2001, pp. 97-100) report a simulation study comparing the power of four tests including *Q* and X^2 to detect a location shift between treatments in a complete block design with normal errors. They used a randomized test approach to ensure that each test had size $\alpha = 0.05$. For $t = 3$ and 4 treatments, $b = 5$ and 10, and two patterns of location-shift for each treatment, 10,000 simulations were run. Their results show that Friedman's test has greater power than X^2 for detecting location shift between treatment distributions. They note similar results were obtained with uniform and double exponential error distributions. They did not consider non-location differences between treatments in their study.

In what follows, we compare the power of X^2 to the power of Friedman's *Q* for plausible location and non-location alternatives to identical treatment distributions based on the exact (small-sample) distribution of the test statistics under both the null and various alternative hypotheses under consideration. We also include the RCB analysis of variance *F*-test for differences in treatment means in our comparisons as a benchmark, as it has certain well-understood optimality properties for detecting location shifts when the treatment distributions are

normal. Note that the *F*-test requires measurements on a continuous scale, while both the *Q* and X^2 tests require only the relative rankings of the observations in each block.

4.1 Design of Study

We looked at $t = 3,\ 4$ and 6 treatments for each of b = 5, 10, 20 and 40 blocks. To compare the power of the *Q*, X^2 and F tests to detect differences in treatment distributions, we chose examples in which treatment distributions differ in location (median or mean) only, in scale only, or in both location and scale. We considered both symmetric and skew treatment distributions and assumed additive block effects, which, without loss of generality were taken to be zero.

The scenarios used in this study are listed in Table 1. The notation is as follows: if the random variable X has a Student's t distribution with 2 degrees of freedom, $X \sim t_2$, then the distribution of $Y = \theta X + \mu$ is denoted $Y \sim t_2(\mu, \theta)$, a Student's t distribution with 2 degrees of freedom, shifted to have median μ and scale θ. Exponential distributions were median-centered and parameterized using the rate equal to $mean^{-1}$, i.e., let $X \sim Exp(\lambda)$ denote an exponential random variable with mean λ^{-1}, and hence median $\lambda^{-1}\ln(2)$. Then $X - \lambda^{-1}\ln(2)$ is a "median-centered exponential" random variable, which we denote $ExpMC(\lambda)$.

The superscript '*' denotes additional distributions added to each block with increasing t. For example, scenario 3 with $t = 3$ involves three independent treatment distributions per block, one each with responses distributed $t_2(0,1)$, $t_2(0,\theta)$, and $t_2(1,1)$; while for $t > 3$ this scenario involves t independent treatment distributions, $(t-2)$ of which were distributed $t_2(0,1)$, and one each of $t_2(0,\theta)$ and $t_2(1,1)$.

Table 1: Scenarios used in power study.

Scenario		Location Departure	Scale Departure	Parameter Values	Null	Non-Null
1.	$t_2(0,1)^*$ and $t_2(\mu,1)$	Varying	None	$\mu = 0, 1, 2, 3, 4$	FH_0	$\left(H_0^R\right)^c$
2.	$N(0,1^2)^*$ and $N(0,\sigma^2)$	None	Varying	$\sigma = \frac{1}{100}, \frac{1}{5}, \frac{1}{3}, \frac{3}{4}, 1, 2, 5, 20, 100$	FH_0	$H_0^R \cap \left(H_0^{\mathbf{p}}\right)^c$
3.	$t_2(0,1)^*$, $t_2(0,\theta)$, and $t_2(1,1)$	Fixed	Varying	$\theta = \frac{1}{100}, \frac{1}{5}, \frac{1}{3}, \frac{3}{4}, \frac{3}{2}, 3, 10, 50, 500$	--	$\left(H_0^R\right)^c$
4.	$N(0,1^2)^*$, $N(\mu,1^2)$ and $N(0,20^2)$	Varying	Fixed	$\mu = 0, 1, 2, 5, 10, 15$	$H_0^R \cap \left(H_0^{\mathbf{p}}\right)^c$	$\left(H_0^R\right)^c$
5.	$ExpMC(1)^*$ and $ExpMC(1/5)+\theta$	Varying	Fixed	$\theta = 0, \pm\frac{1}{2}, \pm 1, \pm 2, \pm 3, \pm 4$	--	$\left(H_0^R\right)^c$

In scenario 1 when $\mu = 0$ then FH_0 is true, while for all other values of μ, FH_0, $H_0^{\mathbf{p}}$, and H_0^R are all false as not only are the treatment distributions not identical, but for each $\mu \neq 0$ the corresponding vector **p** is non-constant, and the expected values of the treatment ranks are not $(t+1)/2$. Similarly, in scenario 2 when $\sigma = 1$, FH_0 is true. However for $\sigma \neq 1$, while H_0^R is still true, $H_0^{\mathbf{p}}$ is not. In scenarios 3 and 5, all values of the parameter θ yield treatment rank distributions where all three hypotheses are false. In scenario 4, $\mu = 0$ is a special case of scenario 2 where H_0^R is true but $H_0^{\mathbf{p}}$ is not, and when $\mu \neq 0$, all three of the hypotheses are false.

4.2 Type I Error Rates

Because of the discrete nature of the exact null distributions of *Q* and X^2, for any fixed nominal significance level $0 < \alpha < 1$ it is generally not possible to find critical values for both *Q* and X^2 that yield tests of size precisely equal to α. This is especially problematic when *b* is small. However it is difficult to compare the power of two tests that are not of the same size. For this reason, rather than fix α at 0.01, 0.05 or some other value, for each combination of *t* and *b*, we used the exact or estimated small-sample null distribution of each test statistic to find critical values that would result in comparable and reasonable size tests.

For $t = 3$, when $H_0^{\mathbf{p}}$ is true, the vector $\mathbf{p} = \frac{1}{6}(1,\ldots,1)$. For each *b*, the exact null distributions of *Q* and X^2 were constructed under $H_0^{\mathbf{p}}$, using the R script Joint.R. Using these exact distributions for each value of *b*, critical values for *Q* and X^2 were chosen to yield proximate nominal Type I error rates, α_Q and α_{X^2}.

For $t = 4$ and $b = 5,$ the R script was used in the same fashion as described above. However for $t = 4$ and each of $b = 10,\ 20$ and 40, an applet by Van de Wiel (2004) was used to compute the exact null distribution of *Q*, while the null distribution of X^2 was estimated via simulation $(n = 2,000,000).$ As in the case of $t = 3,$ using these distributions for each value of *b*, critical values for *Q* and X^2 were selected to yield proximate nominal Type I error rates (exact or estimated).

In the case of $t = 6,$ for all values of *b* and for both *Q* and X^2, the null distributions were estimated via simulation $(n = 2,000,000),$ and critical values for *Q* and X^2 were chosen to yield proximate, estimated Type I error rates. In those cases in which the null distribution was estimated via simulation, a conservative bound on the simulation standard error is $\sqrt{0.10(0.90)}/\sqrt{2,000,000} \approx 0.0002.$

For each value of *t* and *b*, the arithmetic average of the established Type I error rates for *Q* and X^2 was used to fix the value of the nominal Type I error rate α_F for the analysis of variance *F*-test, and the corresponding critical value was obtained from the *F* distribution with numerator and denominator degrees of freedom $(t-1)$ and $(t-1)(b-1)$ respectively. Note that when treatment distributions are normal each with common variance (scenario 2 when $\sigma = 1$), the ANOVA *F*-test will have a Type I error rate α_F. Otherwise the size of the *F*-test could be substantially different from the nominal value α_F.

The exact and estimated nominal Type I error rates obtained from this process are summarized in Table 2 together with the corresponding critical values c_*. Values for α_Q (c_Q) and α_{X^2} (c_{X^2}) that were obtained from a simulation-based estimate of the distribution of a test statistic are indicated in the table by a bold font.

Table 2: Nominal Type I error rates and corresponding critical values for Friedman's *Q*, and RCB ANOVA *F*.

t	b	$\alpha_Q\ (c_Q)$	$\alpha_{X^2}\ (c_{X^2})$	$\alpha_F\ (c_F)$
3	5	0.0239 (7.6)	0.0201 (15.4)	0.0220 (6.39)
3	10	0.0179 (7.8)	0.0172 (14)	0.0176 (5.10)
3	20	0.0374 (6.4)	0.0379 (11.8)	0.0377 (3.58)
3	40	0.0402 (6.45)	0.0410 (11.6)	0.0406 (3.34)
4	5	0.0167 (9.24)	0.0163 (47.8)	0.0165 (5.12)
4	10	0.0374 (8.16)	**0.0372 (38)**	0.0373 (3.25)
4	20	0.0370 (8.34)	**0.0374 (37.6)**	0.0372 (3.02)
4	40	0.0391 (8.31)	**0.0390 (36.8)**	0.0391 (2.88)
6	5	**0.0143 (12.6571)**	**0.0138 (1003)**	0.0141 (3.79)
6	10	**0.0611 (10.2857)**	**0.0606 (854)**	0.0609 (2.30)
6	20	**0.0261 (12.4571)**	**0.0260 (844)**	0.0261 (2.68)
6	40	**0.0254 (12.6714)**	**0.0254 (824)**	0.0254 (2.62)

4.3 Calculation of Power

The exact distribution of both *Q* and X^2, and hence the power of these tests, depends upon the treatments only through the vector **p** of multinomial probabilities associated with each possible ordering of the treatment ranks under the alternative. In turn, for each scenario under consideration **p** depends on a parameter θ. With this in mind, it is sometimes convenient to write $1-\beta_Q(\mathbf{p}(\theta))$ and $1-\beta_{X^2}(\mathbf{p}(\theta))$ as the power of the respective tests for a given vector $\mathbf{p}(\theta)$.

To determine the power of *Q* and X^2 for each alternative to $H_0^{\mathbf{p}}$, first the distribution of each test statistic was either calculated exactly using Joint.R (in most instances for $t=3$) or simulated. Exact calculation ($t=3$) requires specification of the 6×1 vector **p** for each value of the parameter involved $(\mu, \theta, \text{ or } \sigma)$ in each of the five scenarios for treatment distributions when $H_0^{\mathbf{p}}$ is not true (see Table 1). In these instances the elements of **p** were calculated via exact or numerical integration using MAPLE 15.0.1 © (MapleSoft, Waterloo Maple, Waterloo, Ontario). The resulting vectors **p** for $t=3$ are available from http://www.stat.colostate.edu/~pturk/AuxMaterial.pdf.

For a few cases where $t=3$ and $b=40$, and for all cases when $t=4$ or 6 for all values of *b*, simulation-based estimates of the non-null distribution were obtained $(n=100,000)$.

In all cases, whether the non-null distribution of *Q* was calculated exactly or estimated via simulation, the exact or estimated power $1-\beta_Q(\mathbf{p}(\theta))$ was taken to be the probability under the specific alternative (indexed by a parameter θ), that *Q* exceeds the critical value c_Q. The power

$1-\beta_{X^2}(\mathbf{p}(\theta))$ was obtained in an analogous fashion for X^2 In those cases where the power was based on a simulation-based estimate from the non-null distribution, a conservative bound on the simulation standard error in estimating the power is $0.5/\sqrt{100,000} \approx 0.0016$.

For the RCB ANOVA *F*-test, in all cases, the non-null distribution of the test statistic was estimated via simulation $(n=10,000)$, and the power $1-\beta_F(\theta)$ was estimated. Thus a conservative bound on the simulation standard error in estimating $1-\beta_F(\theta)$ by $1-\hat{\beta}_F(\theta)$ is $0.5/\sqrt{10,000}=0.005$.

5. Results

For the sake of brevity, graphs of the results for $t=4$ are not given here, but are discussed. In addition, tables of results on which Figures 1 through 5 are based are not included. Both are available from http://www.stat.colostate.edu/~pturk/AuxMaterial.pdf. The graphs in Figures 1 through 5 give linear interpolated power curves for each of the test statistics *Q*, X^2, and *F* when $t=3$ and 6 for each $b=5$, 10, 20 and 40, for each of the five scenarios discussed below.

5.1 Location Shift

When departure from identical treatment distributions is due solely to a location shift in one of the treatment distributions not only is FH_0 not true, but in addition H_0^R is not true – the expected rank of the *j*-th treatment differs from $(t+1)/2$ for some $j=1,2,\ldots,t.$

Scenario 1 allows comparison of the power of *Q*, X^2, and *F* to detect a location shift in one of the *t* treatment distributions, using a location-shifted Student's *t* distribution with 2 degrees of freedom, as described in section 4.1. Graphs of the results for the power calculations for a mean shift in one treatment by $\mu=0$, 1, 2, 3, or 4 are given in Figure 1. Due to the symmetry of the problem, for each test, location shifts of $\mu=\pm\mu_0$ will give the same power, hence we consider only non-negative values of μ.

While the method of construction of these tests insures that for each value of *t* and *b*, the achieved Type I error rate (when $\mu=0$) for *Q* and X^2 matches the nominal values in Table 2, this is not the case for the *F* test. The estimated Type I error rate for the *F* test is consistently low, ranging from 55% of its nominal value when $t=3$ and $b=5$ (0.0122 versus 0.0220), to 74% of its nominal value when $b=20$ for $t=3$, 4 or 6 (e.g., for $t=3$: 0.0279 versus 0.0377). This is perhaps not surprising given that the t_2 distribution used here is heavy-tailed and the *F* test is designed to detect location differences between normal treatment distributions.

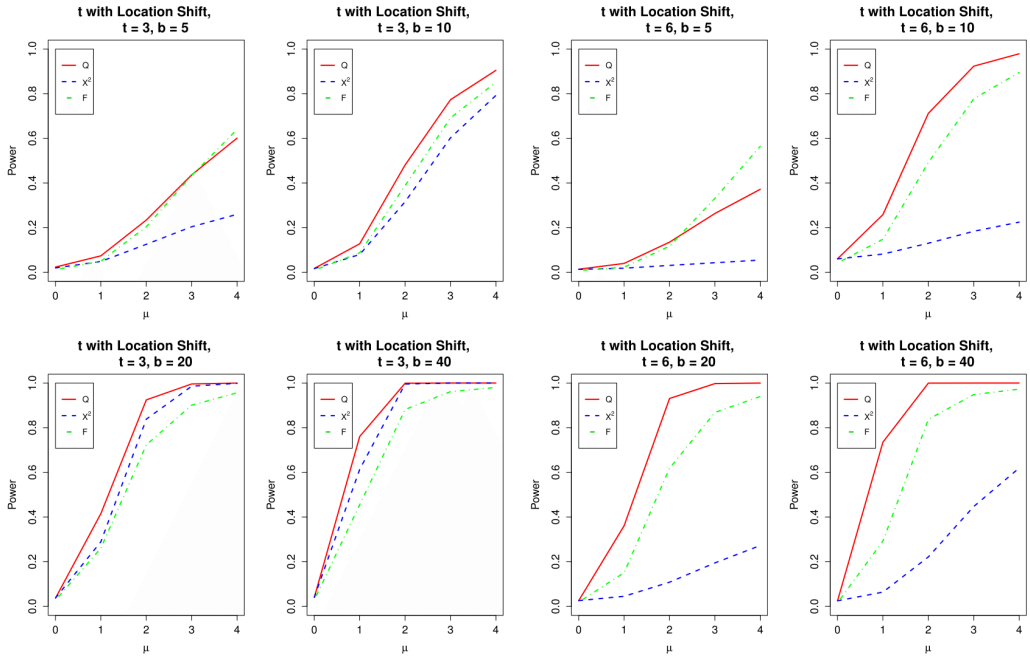

Figure 1: Estimated power curves for the *F*, *Q* and X^2 tests in Scenario 1.

With the exception of $b = 5$, these graphs show that *Q* has greater power for detecting a location shift under this scenario than either X^2 or *F*. For $t = 3$, the power of X^2 becomes competitive with that of *Q* for $b = 20$ and 40, particularly for larger values of μ. However for $t = 4$ (not shown here), this relationship is less evident, and for $t = 6$, the power of the X^2 test falls far short of *Q*.

Surprisingly, when $b = 5$ the *F* test is as powerful ($t = 3, 4$) or more powerful ($t = 6$) than *Q* in detecting a location shift in treatment distributions, even though the *F* test is conservative here.

Results (not shown here) were also obtained using a normal location family in place of the t_2 location family, for the same values of *t* and *b*. As one might expect, the *F* test had the greatest power to detect a location shift, followed by Friedman's *Q*, and then X^2 The relative power of X^2 to *Q* in the normal location family was very similar to the results discussed above.

From these examples, we conclude that the power of the X^2 goodness of fit test to detect location differences among treatment distributions, while better than *F* for heavy-tailed distributions when $t = 3$ and $b = 20$ or 40, does not do as well as *Q*. Excepting very small sample sizes ($b = 5$), *Q* outperforms the *F* test for detecting location differences in heavy-tailed distributions (including other examples examined but not reported here), but less well for light-tailed distributions. With respect to the relative behavior of the *F* and *Q* tests, this is consistent

with the results of O'Gorman (2001), though his results do not include symmetric distributions as heavy tailed as the t_2 – and he did not include X^2 in his study.

5.2 Scale Shift

Scenario 2 allows comparison of the power of *Q*, X^2 and *F* to detect a scale shift in one of the *t* treatment distributions, using a normal family, as described in section 4.1. The graphs in Figure 2 give the results based on power calculations for a shift in scale by $\sigma = \frac{1}{100}, \frac{1}{5}, \frac{1}{3}, \frac{3}{4}, 1, 2, 5, 20, 100$ – encompassing both scale compression and scale expansion. When $\sigma \neq 1$, both FH_0 and $H_0^{\mathbf{p}}$ are not true, however H_0^R is true – the expected rank of the *j*-th treatment is the same for all $j = 1, \ldots, t$. Consequently, this is a circumstance in which we expect that *Q* will have little power to detect departure from $H_0^{\mathbf{p}}$ and where X^2 will do well.

If one knew to expect that any potential differences among the treatments would be due to differences in scale, then one of several common tests might be used to look for treatment differences, including Hartley's test or the Brown-Forsythe test. Here, we imagine the practitioner using *Q*, X^2 and *F* to look for differences among treatments not knowing what the nature of the difference might be. In addition, the results here will be helpful in understanding the ability of the three tests under consideration to detect departures (from FH_0) that involve both location and scale shifts.

As one would expect, the estimated size of the *F* test under this scenario is within simulation error of its nominal value for all *t* and *b*.

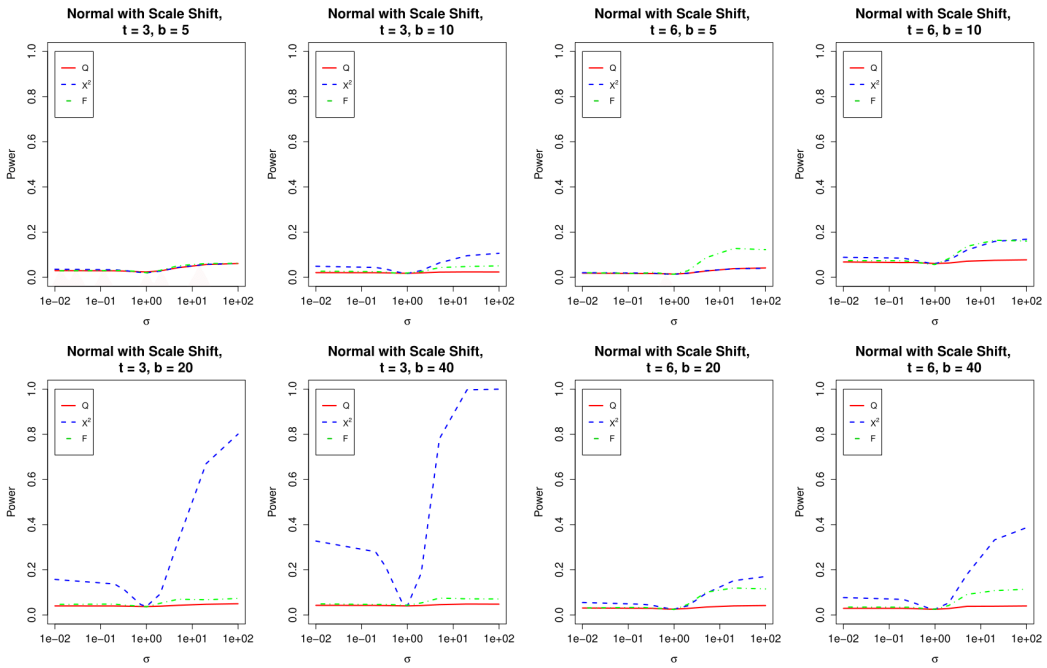

Figure 2: Estimated power curves for the *F*, *Q* and X^2 tests in Scenario 2.

None of the three tests has great power to detect a small-to-moderate shift in scale among the treatments. Friedman's *Q* does uniformly poorly having virtually no power to detect any shift in the scale of a treatment no matter how large, for $t = 3$, 4 or 6 and $b = 5$, 10, 20 or 40. The *F* test is worse except for $t = 3$, $b = 5$. The X^2 test is clearly superior for $b = 10$, 20 and 40 when $t = 3$, 4 or 6. For fixed *t*, it does better as *b* increases. For $t = 3$, X^2 is really the only serious choice for detecting scale shifts.

There is a form of invariance at work here: for $t = 3$ in a symmetric scale-shift family *f* with scale parameter $\theta > 0$, as θ increases the resulting multinomial probability vector $\mathbf{p} = \mathbf{p}_f(\theta)$ traces a line segment in the standard $(s-1)$-simplex – here $s-1=5$. The line segment traced is invariant to choice of *f*, though the parameterization of it (by θ) depends on *f*. Since the exact distributions of Q and X^2 depend only on $\mathbf{p}$, for a given $\mathbf{p}^*$ on the line segment and two such scale families (f,θ) and (g,ϕ), one can find a value $\theta = \theta^*$ and a value $\phi = \phi^*$ for which $\mathbf{P}_f(\theta^*) = \mathbf{P}_g(\phi^*) = \mathbf{P}$. It follows that for any choice of critical value, the power of *Q* to detect a scale shift θ^* in scale family (f,θ) is the same as the power of *Q* to detect a scale shift ϕ^* in scale family (g,ϕ), and similarly for X^2 A similar result holds for symmetric, scale-shift families when $t = 4$. The justification of this result for $t = 3$ is given in Appendix 3.

Consequently, the results for Q and X^2 for $t = 3$ in Figure 2 for the normal scale-shift family would be identical for any symmetric, scale-shift family, up to a change in the scale of the horizontal axis.

For $t = 3,$ similar results were obtained in an additional scale example when sampling from a skew normal distribution (Azzalini, 1985), with skew parameter $\alpha = 5$. Whether the skew normal treatment distributions are centered to have common mean, or to have common median, the power curves for the three tests (associated with a scale shift in one of the three treatment distributions) are nearly identical to those displayed in Figure 2. We conclude that (at least) in the presence of slight skew, the power to detect shifts in scale is well-described by the series of graphs in Figure 2.

5.3 Location and Scale Shift

In practice, as discussed by St. Laurent & Turk (2013), many misapplications of Friedman's test occur when practitioners are aware of heterogeneity in scale of treatments, and wish to assess whether there is evidence for differences in medians (or means) between treatments, irrespective of scale differences. In scenarios 3, 4 and 5 we investigate the power of Q, X^2, and F to detect combinations of location and scale shift among the treatment distributions.

Scenario 3 compares the power of each test to detect a difference among the t treatment distributions when $(t-2)$ treatment distributions are distributed $t_2(0,1)$, one treatment is mean shifted, $t_2(1,1)$, and the remaining treatment is scale shifted, $t_2(0,\theta)$. We considered $\theta = \frac{1}{100}, \frac{1}{5}, \frac{1}{3}, \frac{3}{4}, \frac{3}{2}, 3, 10, 50, 500$ – encompassing both scale compression and expansion. The results of the power calculations are graphed in Figure 3. Note that there is no value of θ in this scenario for which any of the stated null hypotheses in (1), (2) or (3) is true; that is, for all values of θ, H_0^R is not true.

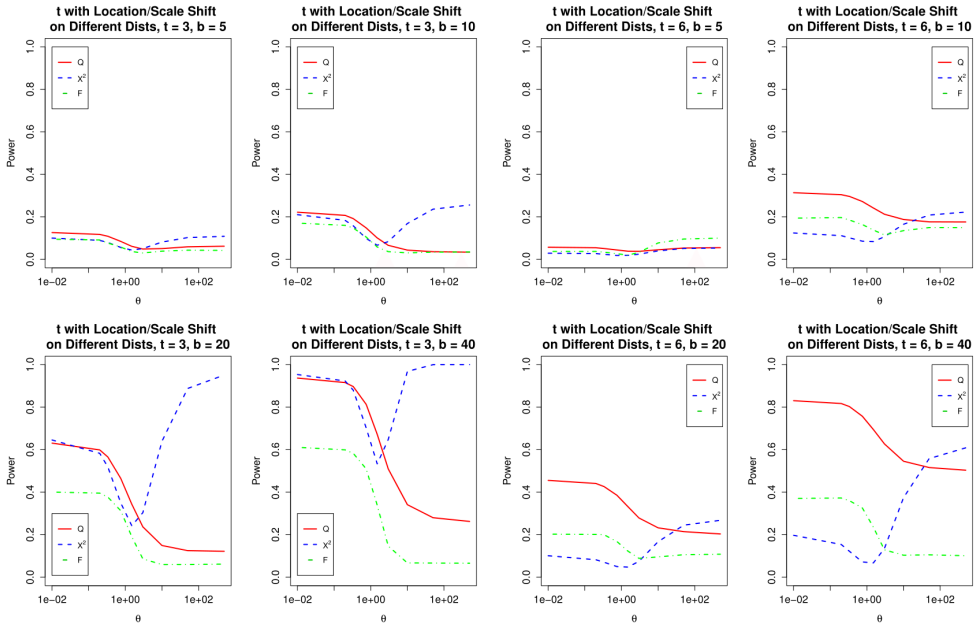

Figure 3: Estimated power curves for the *F*, *Q* and X^2 tests in Scenario 3.

The power curves for *Q* and *F* have a similar shape: each test generally has more power for detecting treatment differences involving scale compression and location shift than for detecting scale expansion and location shift. A reasonable explanation for this phenomenon is that in the presence of a location shift when θ is small, the differences in treatment distributions is dominated by the location shift in the $(t-1)$-st treatment, and both of these tests have reasonably good power for detecting location shifts (section 5.1). However, when θ is large, the differences in treatment distributions is dominated by the scale shift in the *t*-th treatment, and neither test is very good at detecting scale shift (section 5.2).

The behavior of the power curve for X^2 can be understood in a similar fashion: X^2 has good to moderate power to detect a location shift (section 5.1), and it also can have reasonable power to detect scale expansion (section 5.2). When θ is small, and the location shift is the dominant difference between the treatment distributions, the power of X^2 is smaller relative to *Q* and *F*, but still greater than its power when $\theta = 1$. However, when θ is moderate to large in value, the power of X^2 is greater, in fact exceeding that of *Q* and *F* – particularly for $t = 3$ (and for $t = 4$ not shown here).

Also note that excepting for $b = 5$, the power of *Q* exceeds the power of *F* for a fixed scale shift θ, for $t = 3$ and 6, and *Q* is substantially better for larger values of *b*. We attribute this to the relative advantage *Q* has over *F* when dealing with heavy-tailed distributions like Student's *t* with 2 degrees of freedom, for which the variance does not exist.

In Scenario 4 we compare the power of each test to detect a difference among t treatment distributions when $(t-2)$ are distributed $N(0,1^2)$, one treatment is scale shifted, $N(0,20^2)$, and the remaining treatment is mean shifted, $N(\mu,1^2)$. We considered $\mu = 0, 1, 2, 5, 10, 15$. As in Scenario 1, the power of each test here is invariant to the sign of the location shift, hence we consider only non-negative values of μ. Graphs of the results appear in Figure 4. Similar to Scenario 3, there is no value of μ here for which FH_0 is true. However when $\mu = 0$ then H_0^R is true, though $H_0^{\mathbf{p}}$ is not.

This is precisely the situation noted by Friedman (1937, second paragraph of footnote 4 on page 678) as requiring "further analysis." Implicitly, it seems that he recognized there might be difficulties in detecting differences in location with his test in the presence of non-constant variance across treatments.

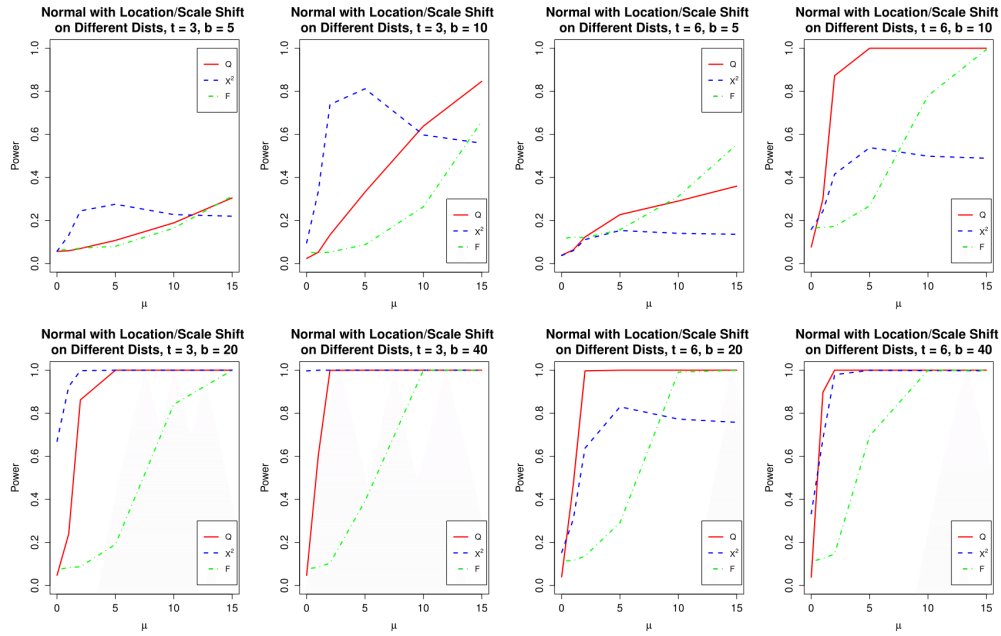

Figure 4: Estimated power curves for the *F*, *Q* and X^2 tests in Scenario 4.

With the exception of $t = 6$ and $b = 5$, the *F* test is not a serious competitor for detecting location shift among the treatments in the presence of a fixed scale shift in one treatment distribution. The behavior of *Q* is reasonably consistent across *t* as *b* increases. When $\mu = 0$ and the only difference among the treatment distributions is a scale shift, it has relatively poor power to detect that shift, however its power increases as μ increases, consistent with the results seen in scenario 1. For $t = 3$ (and for $t = 4$, available in the supplemental materials), and small *b*, X^2 has greater power than *Q* for detecting small or moderate location shift in the presence of

a scale shift, and greater or equivalent power for all location shifts for larger *b*. In essence, X^2 "already" has power to detect a difference among treatments due to scale shift when $\mu = 0$ (scenario 2) and this power increases as μ increases – at least up to a point – when $b = 5$ and 10, and monotonically for larger *b*. For $t = 6$, X^2 is not competitive with *Q* excepting when *b* is large $(b = 40)$. Generally for larger values of *b*, in the presence of a scale shift, X^2 is to be preferred to *Q* on the basis of its greater power to detect small location shifts and equivalent power to detect large location shifts.

In Scenario 5 we consider sampling from a distribution with fixed non-zero skew, where the treatments vary in scale, and location. As described in section 4.1, scenario 5 compares the power of each test to detect a difference among *t* treatment distributions when $(t-1)$ are distributed median-centered exponential, $ExpMC(1)$, and the remaining treatment is both scale and location shifted, $ExpMC(1/5)+\theta$. We considered $\theta = 0, \pm\frac{1}{2}, \pm 1, \pm 2, \pm 3, \pm 4$. Graphs of the results are provided in Figure 5. As in Scenario 3, H_0^R is not true for all choices of θ.

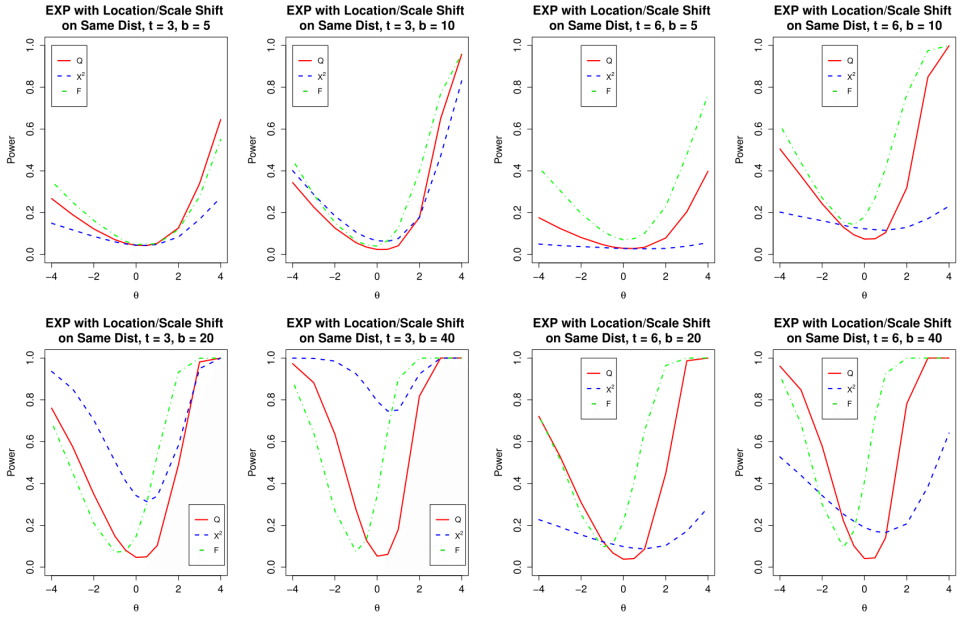

Figure 5: Estimated power curves for the *F*, *Q* and X^2 tests in Scenario 5.

For fixed *b*, the power curves for *Q* and *F* behave similarly for both $t = 3$ and 6 (and for $t = 4$, available in the supplemental materials). It is apparent that X^2 is generally not competitive with *Q* and *F* for $t = 6$, but for both $t = 3$ and 4, it is often superior when the location shift is opposite in direction to the skew, i.e., $\theta < 0$. There is no clear winner here; as the test with the greatest power depends on the values of *t*, *b*, and θ.

6. Discussion & Conclusions

In searching for differences among treatment distributions in a randomized complete block design, Friedman's *Q* provides a direct test of the hypothesis of no difference in the average treatment ranks, H_0^R, and hence an indirect test of FH_0. When interest is in establishing evidence for differences in treatments that may not be due to just a location shift, the X^2 goodness-of-fit test casts a wider net insofar as it provides a direct test of $H_0^{\mathbf{p}}$ – an hypothesis "closer" to the null hypothesis FH_0, and as a consequence has sensitivity to alternatives to FH_0 that are undetectable by Friedman's test. And as there are no rank-based tests that will detect a departure from FH_0 that is not also a departure from $H_0^{\mathbf{p}}$, one cannot find a rank-based test that can detect a larger class of alternatives.

While Friedman's test clearly has superior power for detecting shifts in location between treatment distributions, the power of the goodness-of-fit test X^2 is greater for detecting shifts in scale as well as combinations of location and scale shifts among treatments. The X^2 test does better when the number of treatments is small (e.g., $t = 3$ or 4) – likely a more realistic situation in practice. For example, in Scenario 4 for $t = 3$ or 4, in the presence of scale shift in one treatment, X^2 does much better than its competitors in detecting small to moderate location shift.

The X^2 test does tend to have less power for large *t*, at least over the range of values of *b* we studied here. Wormleighton (1959) intimates that the sensitivity of X^2 to a wider class of alternatives likely comes with it the need for larger sample sizes to detect those alternatives. Quade (1984) makes a similar conjecture. As a practical matter, we suspect that in the case of small sample sizes, this is in part due to the sparsity of the **m** vector, and resultant coarseness of the small sample distribution of X^2. For *t* treatments and *b* blocks, under the null hypothesis $H_0^{\mathbf{p}}$, it can be shown that the probability that a given element of **m** is non-empty is $1-(1-\frac{1}{t!})^b$. Consequently, the number of blocks needed for this probability to exceed $\frac{1}{2}$, say, quickly becomes prohibitively large: for $t = 3$, at least 4 blocks are needed; for $t = 4$, at least 17 blocks, while for $t = 6$, at least 499 blocks are required. Again, suggesting that X^2 would be best suited for use when $t = 3$ or 4.

Particularly surprising was the competitive, and sometimes modestly better, performance of the analysis of variance *F* test in detecting certain location and scale shifts (e.g., scenario 5) between treatments when the number of blocks is small, $b = 5$, for $t = 3$, 4 and 6.

Stochastically Ordered Alternatives. Tamhane and Dunlop (2000, p. 584), and Lehmann (1975, p. 262) state the alternative to the Friedman null hypothesis (1), as a stochastic ordering

among two or more of the distribution functions $F_{i1},\ldots,F_{it}$ in block *i* for all blocks. With respect to the current study: each of the alternatives considered in Scenario 1 is stochastically ordered; in Scenario 5 only the alternatives for $\theta = +3$ and $+4$ are stochastically ordered; and in the remaining scenarios none of the alternatives are stochastically ordered. Developing a better understanding of the small-sample behavior of Friedman's test and the X^2 test when the alternatives under consideration do, or do not, exhibit stochastic ordering is the subject of future work.

Aligned Ranks Test. The aligned ranks test proposed by Hodges & Lehmann (1962) and further developed by Sen (1968) is an alternative nonparametric test applicable in the randomized complete block setting that has good power relative to Friedman's test for detecting location shifts (O'Gorman, 2001). Comparing the X^2 test to the aligned ranks test in location and non-location shift settings is an avenue for future study.

In this study we considered only alternative distributions for the treatments that differed additively between blocks. An interesting question beyond the scope of this study would be to consider alternatives that incorporated variable treatment effects.

Alternative Approaches. Within the context of a rank-based methodology, we have focused on one approach to testing FH_0 via a test of $H_0^{\mathbf{p}}$ in the multinomial setting using X^2. Here one could also consider the class of power-divergence statistics for evaluating departures from $H_0^{\mathbf{p}}$ based on ranks. This class includes X^2 and the likelihood ratio test statistic G^2 as special cases (Cressie & Read 1984, Read & Cressie 1988).

Appendices

Appendices are included in the supplementary material available from <http://www.stat.colostate.edu/_pturk/AuxMaterial.pdf>.

References

[1] Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* **12**, 171-178.

[2] Cressie, N. and Read, T.R.C. (1984). Multinomial Goodness-of-Fit Tests. *Journal of the Royal Statistical Society*, Ser. B **46**, 440-464.

[3] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* **32**, 675-701.

[4] Hodges, J.L. and Lehmann, E.L. (1962). Rank methods for combination of independent experiments in the analysis of variance. *Annals of Mathematical Statistics* **33**, 482-497.

[5] Iman, R. L. and Davenport, J. M. (1980). Approximations of the critical region of the Friedman statistic. *Communication in Statistics, Part A - Theory and Methods* **9**, 571-595.

[6] Koehler, K.J. and Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association* **75**, 336-344.

[7] Lehmann, E.L. (1975). Nonparametrics: Statistical Methods Based on Ranks, San Francisco: Holden-Day, Inc.

[8] O'Gorman, T. (2001). A comparison of the F-test, Friedman's test, and several aligned rank tests for the analysis of randomized complete blocks. *Journal of Agricultural, Biological, and Environmental Statistics* **6**, 367-378.

[9] Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine* **50**, 157-175.

[10] Quade, D. (1984). "Nonparametric Methods in Two-Way Layouts," *Handbook of Statistics*, Vol. 4, 185-228, (P.R. Krishnaiah and P.K. Sen, Eds.) Elsevier Science.

[11] Rayner, J.C.W. and Best, D.J. (1990). A comparison of some rank tests used in taste testing. *Journal of the Royal Society of New Zealand* **20**(3), 269-272.

[12] Rayner, J.C.W. and Best, D.J. (2001). A Contingency Table Approach to Nonparametric Testing, Boca Raton: Chapman & Hall / CRC.

[13] Read, T.R.C. and Cressie, N.A.C. (1988). Goodness-of-Fit Statistics for Discrete Multivariate Data, New York: Springer-Verlag.

[14] St. Laurent, R. and Turk, P. (2013). The effects of misconceptions on the properties of Friedman's test. *Communications in Statistics – Simulation and Computation* **42**(7), 1596-1615.

[15] Sen, P.K. (1968). On a class of aligned rank tests in two-way layouts. *Annals of Mathematical Statistics* **39**, 1115-1124.

[16] Tamhane, A.C. and Dunlop, D.D. (2000). Statistics and Data Analysis: From Elementary to Intermediate. Upper Saddle River, NJ: Prentice-Hall.

[17] Van de Wiel, M. A. (2004). Exact null distributions of quadratic distribution-free statistics for two-way classification. *Journal of Statistical Planning and Inference* **120**, 29-40.

[18] Wormleighton, R. (1959). Tests of permutation symmetry. *Annals of Mathematical Statistics* **30**(4), 1005-1017.

Received July 8, 2013; accepted December 25, 2013.

Roy St. Laurent
Department of Mathematics & Statistics, PO Box 5717,
Northern Arizona University, Flagstaff, AZ 86011-5717, USA,
Roy.St.Laurent@nau.edu

Philip Turk
Department of Statistics
200 Statistics Building
Colorado State University
Fort Collins, CO 80523
pturk@rams.colostate.edu