A comparison of the Exact Kruskal-Wallis Distribution to Asymptotic Approximations for All Sample Sizes Up to 105

J. Patrick Meyer
University of Virginia

Michael A. Seaman
University of South Carolina

September 6, 2011

An earlier version of this paper was presented at the annual meeting of the American Educational Research Association, March 2008.

Abstract

We generated exact probability distributions for sample sizes up to 35 in each of three groups ($N \leq 105$) and up to 10 in each of four groups ($N \leq 40$). We provided a portion of these exact probability tables and compared the exact distributions to the chi-square, gamma, and beta approximations. The beta approximation was best in terms of the root mean squared error. At specific significance levels either the gamma or beta approximation was best. These results suggest that the most common approximation, the chi-square approximation, is not a good choice, though for larger total sample sizes and equal numbers in each group, any of these three approximations are reasonable. For sample sizes up to 105, we can now provide exact tables that negate the use of an approximation.

A comparison of the Exact Kruskal-Wallis Distribution to Asymptotic Approximations for All Sample Sizes Up to 105

Kruskal and Wallis's (1952) rank-based test of location equality for three or more groups may be among the most useful of available hypothesis testing procedures for behavioral and social science research. It is also a relatively popular method. A recent search on APA PsycNet for "Kruskal-Wallis" returned 268 results whereas the search term "hierarchical linear models" returned only 233 results, after limiting the search to peer-reviewed journal articles published between 2000 and 2011. The popularity of the Kruskal-Wallis test may be attributed to its usefulness in a variety of disciplines such as education, psychology, and medicine. It is also suitable for, and has been applied to, a wide array of topics such as the validity of educational or psychological measures (Armstrong, MacDonald, & Stillo, 2010; Jang, Chern, & Lin, 2009; Rajasagaram, Taylor, Braitberg, Pearsell, & Capp, 2009; Tarshis & Huffman, 2007; Yin & Shavelson, 2008), teacher characteristics (Finson, Pedersen, & Thoms, 2006; Gömleksiz & Bulut, 2007), child development (Belanger & Desrochers, 2001), and adolescent behavior and learning disabilities (Plata & Trusty, 2005; Plata, Trusty, & Glasgow, 2005).

Most applications of the Kruskal-Wallis test use a large-sample approximation instead of the exact distribution. Indeed, none of the articles previously cited report use of an exact test and almost all of them report use of the chi-square approximation even though other approximations exist. Statistical packages such as SPSS and R default to the chi-square approximation and only offer the exact distribution for very small sample sizes. Unfortunately, very little is known about the veracity of the chi-square and other approximations for total sample sizes beyond about fifteen participants. This limitation is largely due to the lack of exact probability tables for

moderate to large sample sizes. To overcome this limitation, we recently extended the exact probability tables to a total sample size of 105 participants as explained below. The purposes of this paper are to (a) share a portion of our exact probability tables[1], and (b) examine three large sample approximations with respect to the exact distribution.

The Kruskal-Wallis Test

Parametric methods, along with the requirement for a stronger set of assumptions, continue to dominate the research landscape despite convincing studies that call into question the wisdom of making such assumptions (Micceri, 1989). Replacing original scores with ranks does not inherently lead to lower power, as one might suppose, but rather can result in a power increase at best and a slight power loss, at worst. This has been verified using both Pitman and finite efficiency indices (Hettmansperger, 1984).

A second criticism of nonparametric procedures, in general, and rank-based procedures, in particular, is that critical values or p-values are either difficult to compute or that tables of critical values are limited. Unlike the power myth, this criticism has some basis in reality. For example, when Kruskal and Wallis (1952) introduced their test, they provided exact probability tables for samples with five or less in each of three groups. Obviously such tables have limited value, though the way out of this predicament is to derive approximations that can be used when the table size is exceeded. Kruskal and Wallis derived three such approximations based on the chi-square, incomplete-gamma, and incomplete-beta distributions.

Most research on the Kruskal-Wallis statistic (*H*) has focused on comparing its performance to one-way analysis of variance (e.g. Boehnke, 1984; Harwell, Rubinstein, Hayes, & Olds, 1992) or evaluating its assumptions (e.g.Vargha & Delaney, 1998). Little has been done

[1] The complete tables for significance levels of .1, .05, and .01 span over 200 pages. They are available upon request from the authors. Complete exact distributions are also available but they require over 1 terabyte of disk storage.

to study the efficacy of these approximations yet such study should be paramount because other types of studies, including those referenced above, rely on these approximations. In short, if the approximations for the percentiles of *H* are problematic, studies that use approximate instead of exact cumulative probabilities for *H* are suspect.

It is really no surprise that there is a paucity of research on the large-sample approximations. Exact probability tables of the Kruskal-Wallis statistic have slowly progressed over the years. Kruskal and Wallis (1952) provided exact probability tables when they proposed their test, but their tables were limited to $N \leq 15$ where *N* is the total sample size. Iman, Quade, and Alexander (1975) published more extensive exact probability tables for *H*, but even for their tables, none of the sample sizes exceeded eight for any of the groups. More recently, exact probabilities have been computed for samples as large as $N = 45$ (Spurrier, 2003) and $N = 60$ (Meyer & Seaman, 2006).

Commercial software for computing exact probabilities for nonparametric statistics is more limited than the existing probability tables. SPSS Exact Tests does not provide exact probabilities for *H* for sample sizes larger than 15 (Mehta & Patel, 2010) and we were unable to obtain exact cumulative probabilities for *H* from StatXact 4.0 (Cytel, 2000) with samples as small as 30 participants. The statistical package R (R Development Core Team, 2011) includes a Kruskal-Wallis test in its base package but this test uses the chi-square approximation. Documentation for the R add-on package muStat (Wittkowski & Song, 2010) suggests it can provide exact probabilities for the Kruskal-Wallis test but this assertion is only true when there are no more than two groups. As such, muStat does not compute exact probabilities in cases where the Kruskal-Wallis test is most useful; it does not compute exact probabilities when there are more than two groups. Our comprehensive review of commercially available software did not

find any that provided exact critical or p values for even moderately sized samples. Indeed, even when software documentation claims to provide such values, the software resorts to Monte Carlo or theoretical distribution approximations for all but the smallest sample size arrangements. This is understandable given the resource-intensive nature of computing exact values, as we describe below.

Exact probability values are necessary to study the veracity of approximations, yet approximations are only necessary when exact probability values do not exist. As the exact probability tables cover an increasing larger sample size, the chi-square, gamma, and beta approximations can be studied more rigorously to support inferences about their usefulness in settings that involve even larger sample sizes. Our purpose in this paper is to provide critical values for *H* with even larger samples and then to compare these with approximate values. At the time of this writing, we have created the exact probability distributions for sample sizes up to 35 participants in each of three groups ($N = 105$) and 10 participants in each of four groups ($N = 40$). We created the tables for all possible configurations of unequal and equal sample sizes.

The H Statistic and Related Quantities

Consider *k* independent samples from distributions with CDFs of $F(x-\theta_1), F(x-\theta_2), \ldots, F(x-\theta_k)$, where θ_i is a location parameter for population *i*. We wish to know if there are differences in location among the *k* populations, so we can test the null hypothesis

$$H_0 : \theta_1 = \theta_2 = \ldots = \theta_k \tag{1}$$

against the alternative

$$H_1 : \theta_i \neq \theta_j \text{ for at least one } i \neq j. \tag{2}$$

This location parameter is general and merely denotes a shift of the otherwise common distribution functions. In practice, a test of the above null hypothesis is usually considered a test of median (or mean) equality and is therefore similar to the one-way analysis of variance (ANOVA) as a test of means. Unlike the ANOVA test, which requires a specific form of distribution identity, namely normal distributions, the Kruskal-Wallis test merely assumes continuous populations that might differ in location, rather than shape.

Kruskal and Wallis (1952) derived a test of the above hypothesis using the *H* statistic which can be viewed as the nonparametric analog of the *F* statistic in this one-way design. Given $i = 1,\ldots,k$ independent random samples, each with n_i observations, all $\sum_{i=1}^{k} n_i = N$ observations are ranked together from lowest to highest. The Kruskal-Wallis *H* statistic is based on the sum of ranks for each sample, R_i, and is given by,

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1). \tag{3}$$

The expected value, variance, and maximum value of *H* are

$$\mu = k - 1, \tag{4}$$

$$\sigma^2 = 2(k-1) - \frac{2[3k^2 - 6k + N(2k^2 - 6k + 1)]}{5N(N+1)} - \frac{6}{5} \sum_{i=1}^{k} \frac{1}{n_i}, \text{ and} \tag{5}$$

$$\eta = \frac{N^3 - \sum_{i=1}^{k} n_i^3}{N(N+1)}, \tag{6}$$

respectively (Kruskal & Wallis, 1952; Kruskal, 1952).

To conduct an α-level test of the null hypothesis of location equality, *H* can be compared to the $100(1-\alpha)$ percentile of *H* so that the H_0 is rejected if the observed value of *H* equals or

exceeds this percentile. As explained below, finding the correct percentile of *H* is not a trivial matter so that tables of these values are needed. In the absence of such tables, an approximation can be used. Kruskal and Wallis (1952) proposed three such approximations that we describe and critique later in this paper.

Computation of Exact Probabilities

Computational intensity is a primary drawback of exact procedures, even with the availability of faster processors. Consider, for example, the three-condition setting with $n_1 = n_2 = n_3 = 20$ (this notation indicates three groups with 20 participants in each group). The computation of *H* requires ranking all 60 observations, from 1 to 60. A straight-forward method for computing a critical value for the test of the location hypothesis using *H* as the test statistic would be to calculate *H* for all 5.8×10^{26} permutations of the 60 ranks into the three conditions. We would also need to store at least the largest values of *H* so that we could then determine the critical value based on determining the cut-off for some percentage (e.g. 5%) of these largest values.

To overcome this limitation of exact procedures, Iman et al. (1975) noted that for $i = 1,\ldots,k$ groups with fixed sample sizes, the distribution of *H* depends only on the rank sums $r_1,\ldots,r_k$. Therefore, the distribution of *H* may be obtained from the distribution of rank sums; there is no need to permute all observations.

For *k* independent groups, the exact probability of the rank sums $r_1,\ldots,r_k$ is given by

$$P(R_1 = r_1,\ldots,R_k = r_k) = \frac{(r_1,\ldots,r_k)W(n_1,\ldots,n_k)}{N!/(n_1!\cdots n_k!)}, \tag{7}$$

where $(r_1,\ldots,r_k)W(n_1,\ldots,n_k)$ denotes the number of ways to obtain the rank sums $r_1,\ldots,r_k$ from the sample sizes $n_1,\ldots,n_k$. A significance level for the Kruskal-Wallis statistic may be obtained

by summing Equation 7 for all rank sums that result in a value of *H* that is greater than or equal to the observed value, *h*.

Iman et al. (1975) described a recursive algorithm for obtaining the frequencies $(r_1,\ldots,r_k)W(n_1,\ldots,n_k)$ that depends on the largest observation's group number. For $k=3$, the recursion is given by,

$$\begin{aligned}\left(\mathrm{r}_1,\mathrm{r}_2,\mathrm{r}_3\right)\mathrm{W}\left(\mathrm{n}_1,\mathrm{n}_2,\mathrm{n}_3\right) &= \left(\mathrm{r}_1-\mathrm{N},\mathrm{r}_2\ ,\mathrm{r}_3\right)\mathrm{W}\left(\mathrm{n}_1-1,\mathrm{n}_2\ ,\mathrm{n}_3\right)\\ &+\left(\mathrm{r}_1,\mathrm{r}_2-\mathrm{N},\mathrm{r}_3\right)\mathrm{W}(\mathrm{n}_1,\mathrm{n}_2-1,\mathrm{n}_3)\\ &+\left(\mathrm{r}_1,\mathrm{r}_2,\mathrm{r}_3-\mathrm{N}\right)\mathrm{W}(\mathrm{n}_1,\mathrm{n}_2,\mathrm{n}_3-1)\end{aligned} \tag{8}$$

This algorithm is relatively fast, but it requires a large amount of storage. Frequencies for all possible permutations of the rank sums, (i.e. the $(r_1,\ldots,r_k)W(n_1,\ldots,n_k)$) must be stored on a computer's hard drive to implement the recursion.

We implemented the Iman et al. (1975) algorithm for computing the exact probability distribution in Java standard edition 1.6. Java has the advantage of offering BigInteger and BigDecimal classes. The BigInteger class has no limit on the size of the integer, and the BigDecimal class can compute decimals to an arbitrary level of precision. This is important because it is necessary to precisely track the number of permutations, even if you do not have to actually compute these permutations. In order to avoid overflow from repeating decimals, some level of precision must be specified for the BigDecimal class. We used the IEEE 754R Decimal128 format, which permits a precision of 34 significant decimal places. By comparison, the integer primitive type only allows integers as large as 2.47×10^9 and the double primitive type only allows for a precision of 15 significant decimal places.

Expanded tables of *H* critical values and exact p-values for equal sample sizes are listed in Table 1 for $k = 3$ groups and in Table 2 for $k = 4$ groups[2].

Approximations to the Exact H CDF

Three approximations described by Kruskal and Wallis (1952) are special cases of the gamma distribution. The most common approximation is based on a $\chi^2(k-1)$ distribution, and the proof of this approximation is provided by Kruskal (1952). Approximations based on the incomplete-Gamma(α,β) and incomplete-Beta(α, β) distributions were described by Kruskal and Wallis, and Wallace (Wallace, 1959). The latter two approximations are achieved by matching the moments of the distribution and require more calculations than the chi-square approximation. The parameters of the gamma approximation are $\alpha = \mu^2 / \sigma^2$, and $\beta = \sigma^2 / \mu$, where μ and σ^2 are defined by Equations 4 and 5. The beta approximation is for H / η, rather than H, and its parameters are

$$\alpha = \mu\left[\frac{\mu(\eta-\mu)-\sigma^2}{\eta\sigma^2}\right]$$

$$\beta = \alpha\left(\frac{\eta-\mu}{\mu}\right),$$

which requires Equations 4 through 6 (Kruskal & Wallis, 1952). Probabilities for the chi-square, gamma, and beta approximations were obtained from the JSci v0.93 Java library (JSci e-group, 2004).

Of these three approximations, the chi-square is the most common and it is implemented in numerous statistical packages. However, limited studies have shown the gamma and beta approximations to be better approximations to the exact distribution in most circumstances

[2] The complete tables are prohibitively large to include in this manuscript, but are available on request from either of the authors.

(Kruskal & Wallis, 1952; Spurrier, 2003; Wallace, 1959). The likely reason for this disparity is the computational burden of implementing the gamma and beta approximations relative to the chi-square approximation that traditionally existed. Exact probability tables for increasingly larger sample sizes negate the use of any approximation, but when they are needed, all approximations can be effortlessly implemented on modern computers.

Existing studies evaluating asymptotic approximations are limited to sample sizes smaller than $N = 45$ (Kruskal & Wallis, 1952; Spurrier, 2003; Wallace, 1959) and only evaluate the approximation at specific significance levels (e.g. .1, .05, and .01) rather than the entire distribution (Spurrier; Wallace). The study described below improved on these limitations by involving much larger sample sizes and the entire exact distribution.

Method

Methods for Studying the Proposed Approximations

Although exact tables negate the need to use any approximation, existing tables are still limited to small and moderate sample sizes, especially when the number of conditions exceeds three. Approximate percentiles of *H* are still needed for larger sample sizes. An additional advantage to creating more extensive tables, aside from the obvious benefits of obtaining exact distributions, is that these can be used to verify the value of a proposed approximation as well as to provide a standard by which to compare approximations to one another.

We conducted a study of the three approximations proposed by Kruskal and Wallis (1952) for sample sizes up to $n_1 = n_2 = n_3 = 35$ and $n_1 = n_2 = n_3 = n_4 = 10$. We studied all possible sample size configurations (i.e. equal and unequal sample sizes) up to 35 participants in one or more groups for three groups and up to 10 participants in one or more groups for four groups. To assess the veracity of an approximation, we used two primary methods. First, the

similarity of the approximate and exact CDFs was evaluated with the root mean squared error (RMSE). The RMSE provides a single value for each sample size configuration studied[3]. We defined the RMSE as

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{j=1}^{m}[A(H_j) - E(H_j)]^2}, \tag{9}$$

where $A(\cdot)$ is the approximate cumulative probability, and $E(\cdot)$ is the exact cumulative probability of H_j. The summation was over all *m* observed values of *H* in the exact distribution. Each sample size configuration yields a distinct CDF for *H*. The RMSE is an index of discrepancy between the entire exact and approximate CDF for each sample size configuration.

The second method we used to assess the approximations was to assess the Type I error rate of each approximation at specific probability values. To do this, we obtained the $p = 1 - \alpha$ quantile of the approximate CDF, $q = A^{-1}(p)$. We then obtained the exact p-value of this quantile, $E(q)$, by selecting the p-value for the smallest value of *H* from the exact distribution that was greater than or equal to $A^{-1}(p)$. Finally, we computed a Type I error rate discrepancy measure as $E(q) - \alpha$. For example, if the 95th percentile on the chi-square distribution is actually the 90th percentile on the exact *H* distribution, this yields a discrepancy score of .05. In our scheme, positive discrepancy scores indicate liberal Type I error rates while negative discrepancy scores are associated with conservative Type I error rates.

We plotted both the RMSE and the error rate discrepancy scores as a function of total sample size and the inequality of group sample sizes. The first type of plot is important for understanding if it is reasonable to make inferences once we get beyond the sample sizes

[3] The exact distribution is discrete. Therefore, we chose the RMSE rather than the mean integrated squared error to compare the asymptotic and exact distributions.

accommodated by the exact tables. In the second type of plot, we used the standard deviation of sample sizes as an indicator of the inequality of sample sizes. Large standard deviations represent groups with very different sample sizes, whereas small standard deviations represent groups with similar sample sizes. A standard deviation of zero represents the case where sample sizes are equal among groups. Note that our use of the standard deviation is not meant to imply that the sample sizes are random variables. It merely refers to the amount of inequality among group sample sizes. With the sample size standard deviation plots, we could determine if the RMSE and error rate discrepancies were affected by various degrees of group sample size inequalities.

We should note that we did not evaluate whether or not the asymptotic approximations were close to the nominal significance level. Our reasoning is that, for a given sample size configuration, the chi-square, gamma, and beta approximations are approximations of the exact Kruskal-Wallis distribution. They are not approximations of the nominal significance level. Therefore, the standard for comparison was the exact distribution, not the nominal significance level.

Results

Root Mean Squared Error

We computed the weighted average and standard deviation of RMSE values for each approximation, where the weight is the number of unique H values in the exact distribution for the particular combination of group sample sizes. As shown in Table 3, the RMSE values are smallest for the beta approximation and largest for the chi-square approximation. The beta and chi-square approximation RMSE values are also the least and most variable, respectively. RMSE values for the gamma approximation are consistently between the chi-square and beta approximations.

Figure 1 illustrates weighted kernel regression estimates of the RMSE values for each distribution as a function of total sample size, where the weight is the number of unique *H* values in the exact distribution for the particular combination of group sample sizes. These plots show that all three approximations are very good for larger sample sizes, but that the RMSE approaches zero fastest for the beta approximation followed by the gamma approximation. Differences in RMSE for the approximations are increased when the number of groups is increased from three to four, but the order of results is the same. RMSE goes to zero fastest for the beta approximation, followed by the gamma and chi-square approximations.

Figure 2 is a plot of RMSE values plotted as a function of the inequality of group sample sizes (i.e. the standard deviation of the sample sizes). As the inequality among sample sizes increases, the beta, gamma, and chi-square approximations become less accurate approximations of the exact distribution. However, the beta approximation is affected least by the sample size inequality followed by the gamma approximation. The chi-square approximation appears to be sensitive to differences in group sample size. Figure 2 again suggests that the beta approximation is the best, followed by the gamma and then the chi-square approximations.

Error Rate Discrepancy Scores

Table 4 lists the mean error rate discrepancy scores. Not surprisingly, given the results for the RMSE, mean discrepancy scores for the beta approximation are closer to zero than for the other two approximations. Moreover, the discrepancy scores are more variable for the chi-square approximation than for either of the other two approximations. A bit more surprisingly, the chi-square approximation outperforms the gamma approximation with three groups and a significance level of .1.

Figure 3 shows weighted kernel regression estimates of error rate discrepancy scores as a function of total sample size. This graph shows that all approximations improve as total sample

size increases. In addition, the chi-square approximation is consistently conservative and sometimes extremely so. Beta and gamma approximations can be either liberal or conservative. For sample sizes less than 8, the beta approximation is notably liberal. However, this effect is largely due to a handful of cases. If the total sample size is less than eight and the sample size configuration involves at least one group with a sample size of one, a high error rate discrepancy is observed for the beta approximation. Otherwise, the beta approximation error rate discrepancy is closer to zero than either the gamma or chi-square approximations.

Figure 4 reinforces the results observed thus far. Beta and gamma approximations are far less sensitive to sample size inequalities than the chi-square approximation. Indeed, the chi-square approximation error rate becomes increasingly discrepant from the exact distribution as group sample sizes become more unequal.

Discussion

We generated exact Kruskal-Wallis distributions for three- and four-groups. We then compared the exact CDFs to three approximations. In most cases the beta approximation provides percentiles that are close to the exact values, with exceptions occurring when the total sample size is less than eight. The fact that the beta approximation error discrepancy scores show it to be the closest approximation to the exact is consistent with the finding that this approximation also yielded the smallest RMSE. The RMSE is calculated across the entire exact distribution, unlike the discrepancy scores which we only calculated in the region of commonly used critical values. In those few cases where the beta approximation did not perform well, the gamma approximation is the most accurate approximation.

The three-group results were exaggerated in the four-group conditions and we are suspicious that this trend might continue for larger numbers of groups; a suspicion that we will

study in future research. Fortunately, all of the approximations are very good when the total sample size is large and when the sample sizes are equal or near equal across conditions.

The reason we believe these results to be fortunate is because for smaller sample sizes ($N \leq 105$) we can provide exact tables, which negates the need for approximations. What we are more concerned about is what the performance of the approximations in these smaller-sample conditions might imply about conditions for which there are still no exact tables. Our findings suggest that in such conditions the beta approximation is best, at least for three- and four-group conditions. The ubiquitous chi-square approximation was never the most accurate approximation to the exact distribution. Our results are consistent with those reported by others (Kruskal & Wallis, 1952; Spurrier, 2003; Wallace, 1959). In light of these results and the availability of modern computers, we discourage the use of the chi-square approximation and suggest that the beta approximation should be the new standard for larger study conditions that exceed the limits of our expanded exact critical value tables.

References

Armstrong, S. A., MacDonald, J. H., & Stillo, S. (2010). School counselors and principals: Different perceptions of relationships, leadership, and training. *Journal of School Counseling, 8*(15), 27pp.

Belanger, N. D., & Derochers, S. (2001). Can 6-month-old infants process causality in different types of causal events? *British journal of Developmental Psychology*, *19*(1), 11-21.

Boehnke, K. (1984). F- and H- test assumptions revisited. *Educational and Psychological Measurement, 44*, 609-617.

Cytel. (2000). *StatXact 4 for windows user manual* Cytel Software Corporation.

Finson, K. D., Pederson, J., & Thomas, J. (2006). Comparing science teaching styles to students' perceptions of scientists. *School Science and Mathematics*, *106*(1), 8-15.

Gömleksiz, M. N., & Bulut, İ. (2007). An evaluation of the effectiveness of the new primary school mathematics curriculum in practice. *Educational Sciences: Theory and Practice*, *7*(1), 81-94.

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing monte carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational and Behavioral Statistics, 4*, 315-339.

Hettmansperger, T. P. (1984). *Statistical inference based on ranks*. New York: Wiley.

Iman, R. L., Quade, D., & Alexander, D. A. (1975). Exact probability levels for the Kruskal-Wallis test. *Selected Tables in Mathematical Statistics,* pp. 329.

Jang, Y., Chern, J. -S, & Lin, K. -C (2009). Validity of the Loewenstein occupational therapy cognitive assessment in people with intellectual disabilities. American Journal of Occupational Therapy, *63*, 414-422.

JSci e-group. (2004). *JSci v0.93*. Retrieved from http://jsci.sourceforge.net.

Kruskal, W. H. (1952). A nonparametric test for the several sample problem. *Annals of Mathematical Statistics, 23*, 525-540.

Kruskal, W. H., & Wallis, A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 47*, 583-621.

Mehta, C. R., & Patel, N. R. (2010). *IBM SPSS Exact Tests*. Somers, NY: SPSS.

Meyer, J. P., & Seaman, M. A. (2006, April). *Expanded tables of critical values for the Kruskal-Wallis H statistic*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Micceri, T. (1989). The unicorn, the normal distribution, and other improbable creatures. *Psychological Bulletin, 105*, 156-166.

Plata, M., Trusty, J., & Glasgow, D. (2005). Adolescents with learning disabilities: Are they allowed to participate in activities? *Journal of Education Research*, *98*, 136-143.

Plata, M., & Trusty, J. (2005). Effect of socioeconomic status on general and at-risk high school boys' willingness to accept same-sex peers with LD. *Adolescence*, *40* (157), 47-66.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical computing, Vienna, Austria. ISBN: 3-900051-07-0, URL http://www.R-project.org.

Rajasagaram, U., Taylor, D. M., Braitberg, G., Pearsell, J. P. & Capp, B. A. (2009). Pediatric pain assessment: Differences between triage nurse, child and parent. *Journal of Pediatrics and Child Health*, *45*, 199-203.

Spurrier, J. D. (2003). On the null distribution of the Kruskal-Wallis statistic. *Journal of Nonparametric Statistics, 15*(6), 685-691.

Tarshis, T. P., & Huffman, L. C. (2007). Psychometric properties of the peer interaction in primary school (PIPS) questionnaire. *Journal of Developmental and Behavioral Pediatrics*, *28*, 125-132.

Vargha, A. s., & Delaney, H. (1998). The Kuskal-Wallis test and stochastic homogeneity. *Journal of Educational and Behavioral Statistics, 23*, 170-192.

Wallace, D. L. (1959). Simplified beta-approximations to the Kruskal-Wallis *H* test. *Journal of the American Statistical Association, 54*, 225-230.

Wittkowski, K., M., & Song, T. (2010). Package 'muStat' [statistical software]. Retrieved from http://cran.r-project.org/web/packages/muStat/index.html.

Yin, Y., & Shavelson, R. J. (2008). Application of generalizability theory to concept map assessment research. *Applied Measurement in Education*, 21, 273-291.

Table 1

Critical Values (CV) and Exact p-values at Significance Levels of .10, .05, and .01 for three groups and up to 35 participants in each group

	.10		.05		.01	
Sample Sizes	CV	p-value	CV	p-value	CV	p-value
5, 5, 5	4.560000	0.099520	5.780000	0.048777	8.000000	0.009459
6, 6, 6	4.538012	0.099849	5.719298	0.049438	8.222222	0.009942
7, 7, 7	4.593692	0.099327	5.818182	0.049108	8.378479	0.009924
8, 8, 8	4.595000	0.099331	5.805000	0.049733	8.465000	0.00906
9, 9, 9	4.582011	0.099584	5.844797	0.049946	8.564374	0.009982
10, 10, 10	4.583226	0.099717	5.855484	0.049897	8.640000	0.009957
11, 11, 11	4.587263	0.099340	5.847351	0.049970	8.670880	0.009959
12, 12, 12	4.582583	0.099758	5.875375	0.049969	8.726727	0.009988
13, 13, 13	4.583432	0.099826	5.880473	0.049987	8.756213	0.009986
14, 14, 14	4.581870	0.099902	5.87485	0.049968	8.793545	0.009998
15, 15, 15	4.592077	0.099967	5.906087	0.049957	8.814686	0.009997
16, 16, 16	4.592474	0.099986	5.906250	0.049930	8.852679	0.009990
17, 17, 17	4.587171	0.099853	5.908970	0.049989	8.870375	0.009984
18, 18, 18	4.591246	0.099993	5.911560	0.049989	8.890685	0.009993
19, 19, 19	4.586493	0.099985	5.919190	0.049939	8.905531	0.009998
20, 20, 20	4.588852	0.099921	5.920328	0.049991	8.924262	0.009998
21, 21, 21	4.594388	0.099927	5.928855	0.049996	8.935658	0.009993
22, 22, 22	4.593314	0.099934	5.928950	0.049983	8.954484	0.009997
23, 23, 23	4.590224	0.099966	5.928382	0.049998	8.960302	0.009998
24, 24, 24	4.591514	0.099969	5.932839	0.049982	8.971081	0.009997
25, 25, 25	4.590989	0.099998	5.933305	0.049998	8.980716	0.010000
26, 26, 26	4.593064	0.099986	5.937982	0.049985	8.992285	0.009998
27, 27, 27	4.594199	0.099992	5.938372	0.049994	8.997691	0.009998
28, 28, 28	4.593998	0.099997	5.943217	0.049997	9.006603	0.009999
29, 29, 29	4.594855	0.099968	5.942709	0.049999	9.012972	0.009994
30, 30, 30	4.594579	0.099989	5.944908	0.049997	9.020952	0.009998
31, 31, 31	4.595302	0.099989	5.945225	0.049998	9.025262	0.010000
32, 32, 32	4.596730	0.099980	5.947890	0.049996	9.032941	0.009997
33, 33, 33	4.595702	0.099981	5.948650	0.049998	9.038090	0.010000
34, 34, 34	4.596499	0.099991	5.950482	0.049994	9.042295	0.010000
35, 35, 35	4.595025	0.099999	5.951914	0.049998	9.047547	0.009998

Note: Critical values are the smallest value of *H* in the cumulative distribution for which the p-value is less than or equal to the nominal significance level.

Table 2

Critical Values (CV) and Exact p-values at Significance Levels of .10, .05, and .01 for four groups and up to 10 participants in each group

	.10		.05		.01	
Sample Sizes	CV	p-value	CV	p-value	CV	p-value
2, 2, 2, 2	5.666667	0.076190	6.166667	0.038095	6.666667	0.009524
3, 3, 3, 3	5.974359	0.099843	6.897436	0.045720	8.435897	0.009199
4, 4, 4, 4	6.088235	0.099001	7.235294	0.049217	9.286765	0.009990
5, 5, 5, 5	6.097143	0.099365	7.377143	0.049506	9.800000	0.009942
6, 6, 6, 6	6.120000	0.099862	7.440000	0.049909	10.100000	0.009999
7, 7, 7, 7	6.141450	0.099667	7.492611	0.049839	10.292048	0.009984
8, 8, 8, 8	6.161932	0.099732	7.542614	0.049887	10.434659	0.009997
9, 9, 9, 9	6.161161	0.099976	7.570571	0.049979	10.539540	0.009986
10, 10, 10,10	6.172683	0.099988	7.598049	0.049972	10.622927	0.009994

Note: Critical values are the smallest value of *H* in the cumulative distribution for which the p-value is less than or equal to the nominal significance level.

Table 3
Descriptive Statistics of RMSE for All Sample Size Configurations

Groups	Distribution	$\overline{x}$	$\hat{\sigma}$
Three	Chi-square	0.0014	0.0019
	Gamma	0.0011	0.0010
	Beta	0.0005	0.0007
Four	Chi-square	0.0087	0.0060
	Gamma	0.0051	0.0023
	Beta	0.0015	0.0013

Note: Statistics weighted by the number of *H* in each exact distribution.

Table 4
Mean Error Rate Discrepancy for All Sample Size Configurations (Standard Deviation in Parentheses)

		Significance Level		
Groups	Distribution	.10	.05	.01
Three	Chi-square	-0.0010 (0.0019)	-0.0020 (0.0017)	-0.0015 (0.0007)
	Gamma	0.0019 (0.0011)	0.0007 (0.0005)	-0.0004 (0.0003)
	Beta	-0.0008 (0.0009)	-0.0003 (0.0004)	0.0002 (0.0002)
Four	Chi-square	-0.0081 (0.0062)	-0.0101 (0.0047)	-0.0054 (0.0013)
	Gamma	0.0050 (0.0018)	0.0008 (0.0009)	-0.0020 (0.0008)
	Beta	-0.0014 (0.0015)	-0.0002 (0.0008)	0.0006 (0.0006)

Note: Statistics weighted by the number of *H* in each exact distribution.

Figure Captions

Figure 1. Weighted Kernel Regression Estimate of RMSE by Total Sample Size

Figure 2. Weighted Kernel Regression Estimate RMSE by Sample Size Standard Deviation

Figure 3. Weighted Kernel Regression Estimate Error Rate Discrepancy by Total Sample Size

Figure 4. Weighted Kernel Regression Estimate Error Rate Discrepancy by Sample Size Standard Deviation

Figure 1. Weighted Kernel Regression Estimate of RMSE by Total Sample Size

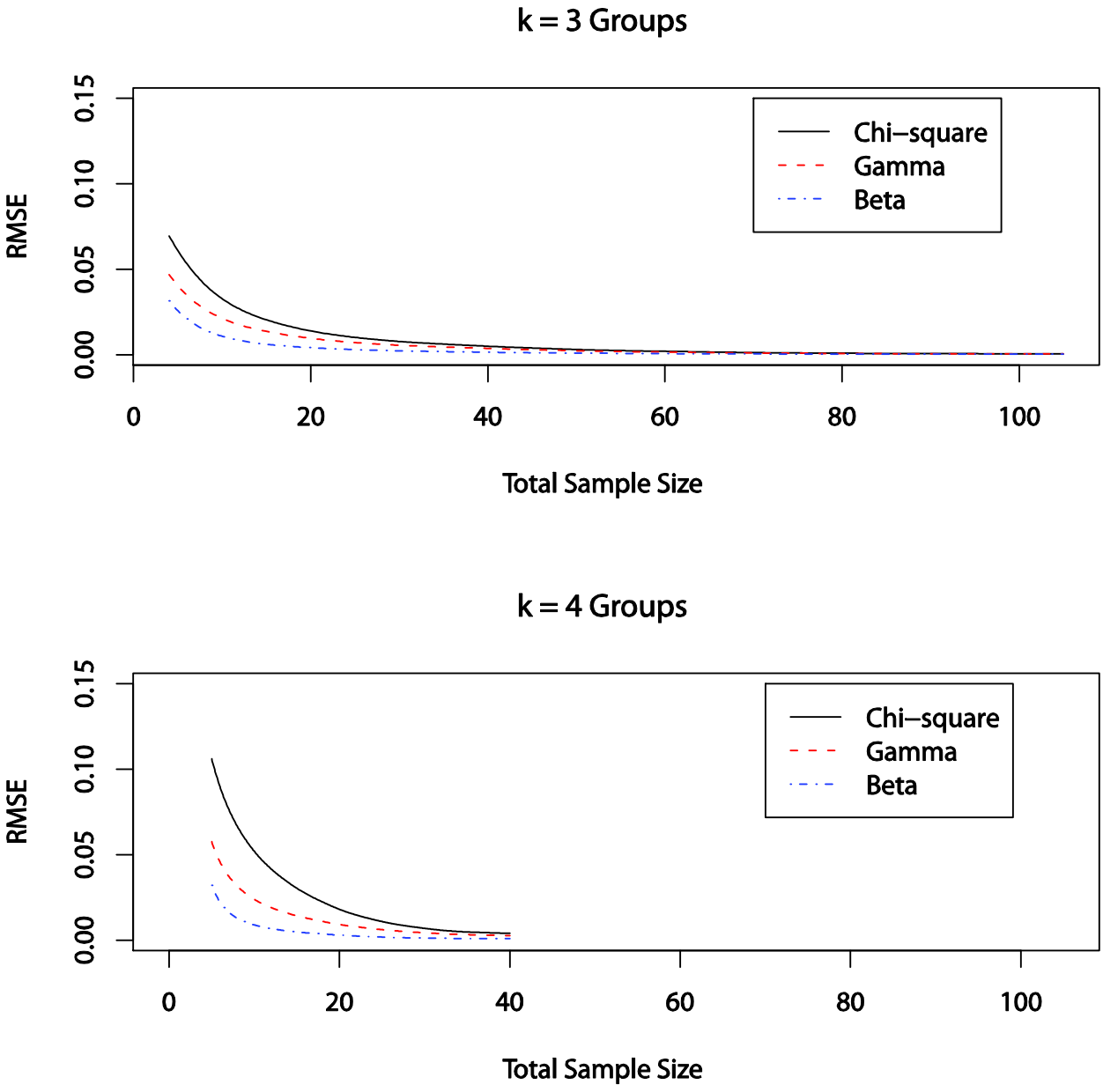

Figure 2. Weighted Kernel Regression Estimate of RMSE by Sample Size Standard Deviation

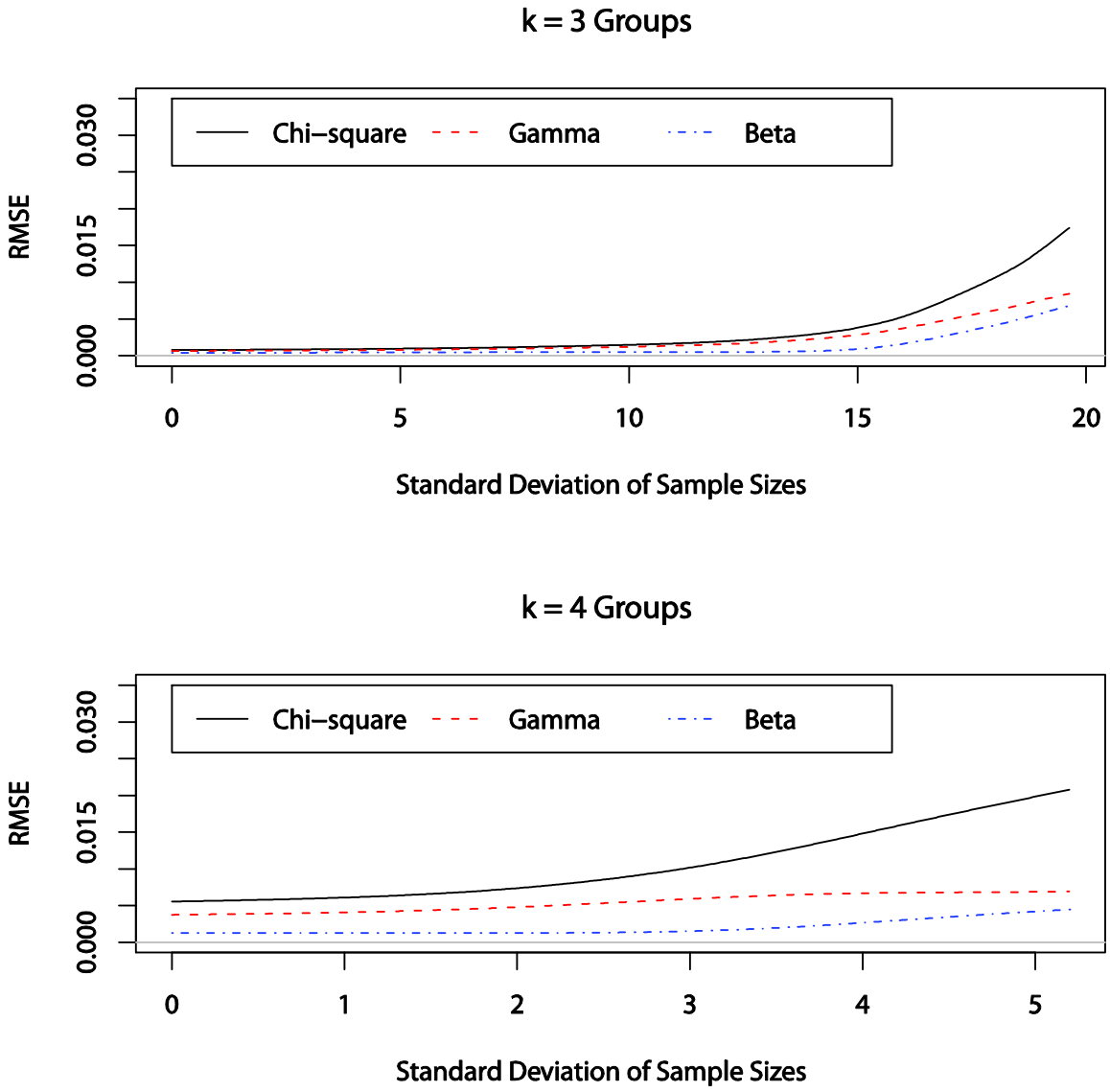

Figure 3. Weighted Kernel Regression Estimate of Error Rate Discrepancy by Total Sample Size

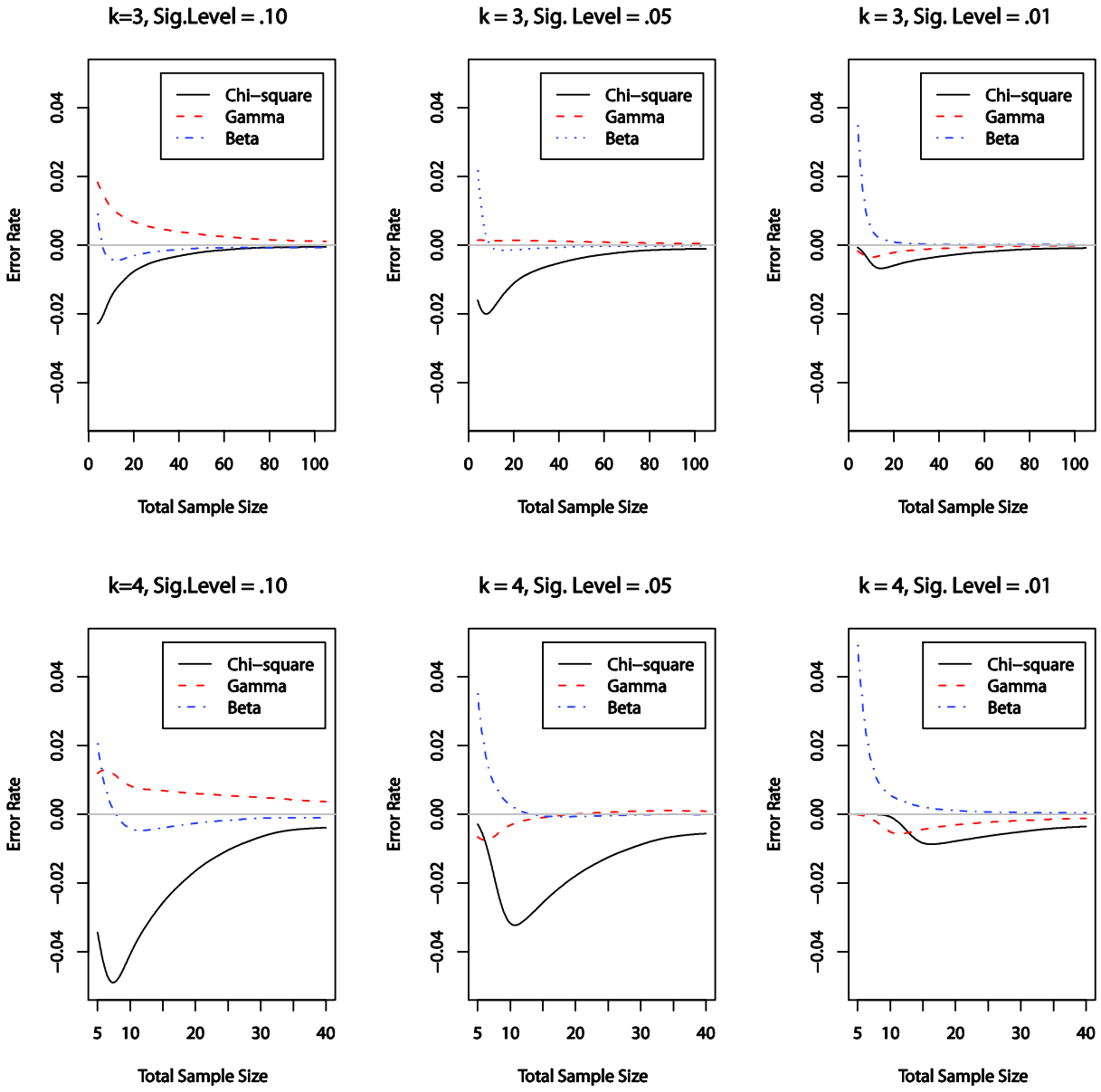

Figure 4. Weighted Kernel Regression Estimate of Error Rate Discrepancy by Sample Size Standard Deviation