AN EVALUATION OF THE POWER OF THE K-SAMPLE BELL-DOKSUM STATISTIC

Amy Ro, Merck, Sharp, & Dohme Research Laboratories
Edward P. Barry, Schering Plough Corporation

1. Introduction

In the mid 1960's, Bell and Doksum (1) introduced a series of statistics based on random normal scores. For the k-sample problem, it involves replacing the original samples with the corresponding normal order statistics. This statistic has good asymptotic relative efficiency (A.R.E.) equal to 1 as compared with the F-test when the F-test is appropriate, and greater than 1 when the normality assumption is not satisfied. However, the Bell-Doksum test has a problem of being dependent on the random sample chosen from the standard normal distribution. The objectives of this paper are: 1) to examine the power of the Bell-Doksum statistic and its comparison with two classical k-sample tests, the F and Kruskal-Wallis statistics, under equal sample sizes and variances with an underlying distribution of normal, uniform or exponential; 2) to examine the influence of unequal sample sizes or variances on the power of the Bell-Doksum statistic; and 3) to examine the problem of the inherent variability for the Bell-Doksum statistic from different random normal samples chosen.

2. Calculating the Statistics

For testing $H_0$: the means for the $k$ populations are equal versus $H_1$: at least one of the populations has a larger mean than at least one of the other populations, the k-sample Bell-Doksum statistic is calculated as follows: pool all samples and rank the combined observations. Replace the $j$th value of the ranked observations by the $j$th order statistic of a random sample of the same size obtained from the standard normal distribution. Let $n_i$ be the sample size for group $i$, and $Z$ be the group mean and grand mean of the normal order statistics, respectively. The statistic

$$BD = \sum_{i=1}^{k} \frac{1}{n_i^2} (Z_i - Z)^2$$

is distributed as a Chi-square variable with $k-1$ degrees of freedom.

To represent the spread of the means, a metric-free measure from Cohen (3),

$$f = \frac{\sqrt{\frac{1}{k} \sum_{i=1}^{k} (u_i - \bar{u})^2}}{\bar{\sigma}}$$

is adopted in this paper. Here, $p_i$ is the proportion of total samples in group $i$, $u_i$ represents the population mean for the $i$th group and $\bar{u}$ represents the grand mean over the $k$ populations. The symbol $\sigma$ denotes the common population standard deviation. Cohen referred to the measure, $f$, as the standard deviation of the standardized means. $f$ assumes the value zero when $H_0$ is true and a large value when the differences among the $k$ means are large.

The alternative hypothesis can take on many patterns. In this paper, only the case where the $k$ means are equally spaced over a maximum distance of $d$ is considered. As described by Cook and Larntz (5), the power for the other patterns (given equal sample sizes among the $k$ groups) can be easily derived from this pattern.

3. Power of the Bell-Doksum Statistic with Equal Sample Sizes and Variances

Simulation studies were done for $k=2$ and $k=3$ with equal sample sizes and variances under normal, uniform and exponential distributions. The comparative power of the Bell-Doksum, $F$ and Kruskal-Wallis statistics for $k=2$ were observed to be very similar to that of $k=3$ regardless of the sample sizes. Since $k=3$ represents a more general case of a k-sample problem, only the results of $k=3$ are presented in this paper. The similarity between the results for $k=2$ and $k=3$ strongly suggested that the relative order of the power of the Bell-Doksum, $F$ and Kruskal-Wallis statistics might be preserved in cases where $k>3$.

One thousand sets of samples for each selected sample size (from hereon, sample size refers to the number of observations per group) were generated using SAS's random number generator. The separation among the 3 groups was achieved by adding the shifts to the generated observations for each group. The Bell-Doksum, $F$, and Kruskal-Wallis statistics were calculated for each sample and compared to the critical values ($\alpha=0.05$) of the appropriate distributions. The proportion of times that $H_0$ was rejected was taken as the simulated power of the corresponding statistic. As described before, the Bell-Doksum statistic depends on the particular random sample of normal values for normal deviations chosen. To stabilize the variability, seven Bell-Doksum statistics were calculated for each set of observations and the median was used to determine the rejection or acceptance of $H_0$. (Note: this is a modification of, rather than the actual Bell-Doksum statistic.)

Sample sizes of $5, 10, 15, 25$ and $50$ per group were considered. Three underlying distributions were examined: the normal, uniform and exponential. The results of the simulation studies on sample sizes of $5, 15, 25$ and $50$ are displayed in Figures 1-12. Illustration 1 contains the macro used to calculate the Bell-Doksum statistic. It allows the user to generate the statistic many times and count the number of times the statistic exceeds the
critical value of the Chi-square distribution with k-1 degrees of freedom. The user can specify the number of Bell-Doksum statistics desired and the (a-level) for the rejection of H0.

When the underlying distribution was normal and the sample size per group was smaller than 50, the F-test was superior to the Bell-Doksum and Kruskal-Wallis tests. In cases where the group size was 10 or smaller, when the Kruskal-Wallis statistic had greater power than the Bell-Doksum test. At sample size of 25, the power of the Bell-Doksum statistic was very similar to that of the Kruskal-Wallis test. They were both slightly smaller in power than the F-test. From Figure 5, it can be observed that for sample size of 50, the asymptotic relative efficiency is approximately 1 for both the Bell-Doksum and Kruskal-Wallis statistics when compared with the F-test.

In situations where the underlying distribution was uniform, the F-test was still better than the Bell-Doksum statistic given that the sample size was 10 or smaller. However, when the sample size increased to 25, the Bell-Doksum statistic became superior in power to the F and Kruskal-Wallis tests. The advantage of the Bell-Doksum statistic was even more evident when the sample size went up to 50. This suggested that when the underlying distribution is short-tailed, the Bell-Doksum statistic might be more powerful than both the F and Kruskal-Wallis tests as long as the sample size is moderately large.

For the exponential distribution, the Kruskal-Wallis statistic was more powerful than the F-test which was superior to the Bell-Doksum statistic with sample size of 5. At sample size of 10, the Bell-Doksum statistic was better than the F-test but was still less powerful than the Kruskal-Wallis statistic. At sample size of 15, the Bell-Doksum and Kruskal-Wallis tests were nearly the same and were both much greater in power than the F-test. Similar to situations where the underlying distribution was uniform, the Bell-Doksum statistic was superior in power to both the Kruskal-Wallis and F-tests when the sample size increased to 25.

4. Power of the Bell-Doksum Statistic with Unequal Sample Sizes or Variances

Barlow (2) compared the power of the Bell-Doksum test to that of the Mann-Whitney statistic when the underlying distributions for the two independent samples were different. In this section, the influence on the power of the Bell-Doksum statistic of unequal samples sizes or variances with k=3 and normal underlying distribution is considered. Figure 13 contains the power curves (of the Bell-Doksum, F and Kruskal-Wallis tests) of which the sample size of the smallest group differed from the largest group by a ratio of 1:3. The local number of observations was 75 with approximately 25 in the middle group. Comparing to curves where the sample sizes were all equal to 25, the power of the Bell-Doksum statistic was almost unchanged and its relative order to the F and Kruskal-Wallis statistics remained the same. This was true for cases where the sample size for the largest group was 5, 10 and up to 15 times the smallest group (results for sample size ratio of 1:15 are displayed in Figure 14).

Figures 15-16 display the power curves for the 3 statistics when the ratios between the smallest and largest variances were 1:3 and 1:15, respectively. The cases where the variance ratios were 1:10 and 1:15 were also studied. Sample size of 25 per group was used in each case. The power of the F and Kruskal-Wallis tests were both nearly unchanged for each of these ratios in comparison with the power curves of equal sample sizes (25 per group) and equal variances. However, the power of the Bell-Doksum statistic decreased with increasing variance ratio. It decreased by a small amount of 1 to 3 when the variance ratio was 1:3. The loss in power increased to 3 to 10% when the variance ratio was 1:15.

From these results, it can be stated that the power of the Bell-Doksum statistic, similar to that of the F and Kruskal-Wallis tests, changes very little even if the sample sizes differ by a ratio of 1:15 between the smallest and largest groups. However, unlike the F and Kruskal-Wallis tests, the power of the Bell-Doksum statistic decreases with increasing variance ratio.

5. Additional Variability of the Bell-Doksum Statistic

The greatest drawback of the Bell-Doksum statistic is the possibility of generating test values that lead to different conclusions. An example taken from Conover (4) is used here to demonstrate this problem. Four methods of growing corn and the results are summarized in Illustration 2. With the particular sample of normal deviates chosen, the Bell-Doksum statistic was 56.36. It was rejected at the 0.05 a-level. The Kruskal-Wallis test statistic was 25.46. Conover stated that this suggests that the Bell-Doksum test may have slightly more power than the Kruskal-Wallis test, at least for this set of data." However, when 1000 sets of random normal deviates were generated and mapped into this set of data, 60% of them produced a smaller Bell-Doksum test value than the Kruskal-Wallis statistic. This suggested an opposite interpretation of the comparative power of the two tests involved.

Conover further pointed out that "there has apparently been no investigation into the amount of variation that several different sets of random normal deviates would introduce into the Bell-Doksum test statistics." Based on two examples worked out in his book, he stated "judging from the close agreement of the Bell-Doksum tests statistics' values with the rank test statistics' values in the preceding two examples, we may suspect that the unwanted
variation is small when the number of observations is reasonably large. To examine the amount of variation that would be introduced by different random normal samples and its relationship with sample size, additional simulation studies were performed. The first one was to generate 25 sets of samples (equal sample sizes and variances and means equally spaced) under the normal distribution and calculate 10 Bell-Doksum statistics for each set of observations. The within-set variance of the Bell-Doksum statistic was calculated based on these 250 test values. Figure 17 plots the within-set variance versus different values of $f$, the measure of separation among the $k$ means. Contrary to Conover's suggestion, the variance of the Bell-Doksum statistic increased with increasing sample size. It also increased with $f$ and began to stabilize when $f$ reached the value of approximately 2. However, it should be noted that with large sample size or large $f$, the Bell-Doksum test value is much greater than the critical value for the rejection of $H_0$. This means that the Bell-Doksum statistics (that can be generated for the same set of data) fall only at the extreme tail of the appropriate Chi-square distribution and have a very small probability of causing opposite decision on the rejection of the hypothesis. This can be confirmed by the results displayed in Figure 18. In this figure, the percentage of times out of 200 that the 40th and 60th percentiles of 25 Bell-Doksum statistics gave contradictory rejection decisions are plotted against different values of $f$. Disagreement between the 40th and 60th percentiles indicated that approximately half of the Bell-Doksum test values infer the rejection of $H_0$ and the other half infer the opposite decision, which is the worst possible case. It is clear that the percentage of times disagreement arose decreased with increasing sample size.

6. Summary and Discussion

From these simulation studies, it can be concluded that the power of the Bell-Doksum statistic is good relative to the $F$ and Kruskal-Wallis tests with moderate sample size and normal underlying distribution. It is a better alternative than the $F$ and Kruskal-Wallis tests when the underlying distribution is uniform or exponential and the sample size per group is 25 or larger. Unequal sample sizes, at least up to a ratio of 1:15, has only a small effect on the power of the Bell-Doksum statistic as long as the average sample size per group is 25 or larger. However, the power of the Bell-Doksum statistic decreases with increasing difference in variance among the $k$ groups. The decrease is less than 3% when the variances differ by a ratio of 1:3 and is 3 to 10% when the variance ratio increases to 1:15. Although the Bell-Doksum statistic is not unique to a given set of data, the variation among the test values generated from different samples of normal deviates is small relative to the critical values for the rejection of $H_0$.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Joe Shih, Ms. Joanne Perry and Ms. Maryonna Kersten for their valuable suggestions and assistance with this paper.

References


POWER CURVES for the B-D, F, and K-W STATISTICS, EQUAL VARIANCES (normal data, maximum sample size ratio between groups = 1:5)

FIGURE 14

POWER CURVES for the B-D, F, and K-W STATISTICS, SAMPLE SIZE 25 (normal data, maximum variance ratio between groups = 1:5)

FIGURE 15

POWER CURVES for the B-D, F, and K-W STATISTICS, SAMPLE SIZE 100 (normal data, maximum variance ratio between groups = 1:5)

FIGURE 16

WITHIN-SET VARIANCE OF BELL-DOKSI STATISTIC (calculated from 25 data sets with 10 B-D statistics each)

FIGURE 17

PERCENT OF TIMES CONTRADICTORY REJECTION DECISIONS OCCURRED BETWEEN THE 40TH AND 80TH PERCENTILES OF 25 B-D STATISTICS

FIGURE 18