



Comparing two mass size distributions

by F. Lombard*, and G.J. Lyman†

Synopsis

We consider in this paper the use of a modified version of Hotelling's statistic in the analysis of particle size distributions. The statistic can be adversely affected by the presence of outliers among the data. We propose a competitor to the statistic that is based on ranks, and hence is less sensitive to outlier effects. The results of a Monte Carlo study suggest that the rank test is highly competitive with the Hotelling test in its ability to detect differences between two mass size distributions. The calculation of the rank statistic is explained in detail and its application is illustrated on two sets of data.

Keywords

mass size distributions, bias testing, rank statistic.

Introduction

Lyman, *et al.* (2010) reported in this Journal the results of research, both applied and theoretical, on the analysis of particle size distributions. The Hotelling T^2 statistic that was proposed can be adversely affected by the presence of outliers among the data. The paper promised further research into an improved methodology. In this paper we try to fulfil the promise. We propose a robust test procedure based on ranks as a competitor to the T^2 statistic. In a typical sizing procedure, a mass M kg of raw material is sorted into $p + 1$ size intervals (in millimetres)

$$(0, c_1], (c_1, c_2], \dots, (c_p, c_{p+1}]$$

and the mass of material, M_1, \dots, M_{p+1} , in each of the size intervals recorded. Thus $M = M_1 + \dots + M_{p+1}$. Often the value of M that is needed to achieve a target precision in sizing is suggested by an industrial standard (e.g. an ISO document) and is supposed to remain fixed over successive sieve analyses. In practice it is hardly possible to keep M constant, and deviations of a greater or lesser magnitude from the prescribed value occur as a matter of course. Consequently, the results of a sieve analysis are invariably reported as a vector of proportions $\mathbf{x} = (x_1, \dots, x_{p+1})$, where

$$x_k = M_k / \sum_{i=1}^{p+1} M_i \geq 0 \quad [1]$$

so that

$$x_1 + \dots + x_{p+1} = 1 \quad [2]$$

As a consequence, the covariance matrix of the multivariate data set that results upon making $n > d$ independent observations on the random vector \mathbf{x} is singular.

In order to develop a statistical method to analyse such data, one thinks of the observation $\mathbf{x} = (x_1, \dots, x_{p+1})$ as a corrupted realization of a 'true' underlying size distribution $\mathbf{q} = (q_0, \dots, q_{p+1})$, $\sum q_k = 1$. The principal sources of corruption are mechanical sampling error and laboratory analysis error. Here q_k denotes the true, but unknowable, proportion of the total mass of a large (conceptually infinite) amount of material that falls in the k th size interval. In this paper we consider tests of the hypothesis of equality of two such underlying size distributions. In the first of the two applications to be considered, two samples were obtained from the same batch of material, one by each sampling method, each sample having been sized using a set of $p = 10$ sieves. This pairwise collection and sizing of samples was repeated on $n = 28$ independent batches of material. Thus, we have 28 pairs

$$(\mathbf{x}_i, \mathbf{y}_i) = ((x_{1i}, \dots, x_{p+1,i}), (y_{1i}, \dots, y_{p+1,i})) \quad [3]$$

of observed sizings. The question is whether or not it is reasonable to assume that the distributions \mathbf{q} underlying the \mathbf{x} results are the same as the distributions \mathbf{q}^* that underlie the \mathbf{y} results. We emphasize that the underlying size distributions typically vary from sample to sample because of the timewise variability in the supply of material being sampled. Thus we do not assume that the observation pairs $(\mathbf{x}_i,$

* Centre for Business Mathematics and Informatics, North-West University, Potchefstroom, South Africa.

† Materials Sampling and Consulting Pty Ltd, Southport, Australia.

© The Southern African Institute of Mining and Metallurgy, 2012. ISSN 2225-6253. Paper received Jul. 2011; revised paper received Nov. 2011.

Comparing two mass size distributions

y_i , $i = 1, \dots, 28$ are realizations of a fixed pair $(\mathbf{q}, \mathbf{q}^*)$ of underlying distributions. As such, an analysis should be based on the 28 sets of differences

$$\mathbf{x}_i - \mathbf{y}_i = (x_{1i} - y_{1i}, \dots, x_{p+1,i} - y_{p+1,i}), i = 1, \dots, 28 \quad [4]$$

Figure 1 shows for this data set boxplots of the 28 differences within each of the 10 size classes. Comparing the medians with the zero line, the overall visual impression is that the two size distributions differ somewhat. The sampler seems to be producing less coarse material, and consequently more of the finer material, than the stopped belt method. On the other hand, the Hotelling T^2 gives a non-significant result (p -value = 0.124), indicating, contrary to what we see in Figure 1, that the observed differences are nothing out of the ordinary. However, we also see that the data abound in outliers (the + signs in Figure 1) and that any real differences between the size distributions may have been masked by the large amount of variability in the data.

The main objective of the present paper is to demonstrate how outlier effects associated with the use of the T^2 statistic can be overcome. We will show that replacing the differences (Equation [4]) by appropriate rank scores takes care of the outlier problem without sacrificing much, if anything, in the ability to detect true differences (statistical power). A secondary objective of the paper is to show that the rank score test is, in fact, often more adept than Hotelling's T^2 at detecting real differences between commonly encountered size distributions.

The paper is structured as follows. We first review briefly the calculation and properties of the T^2 statistic in its application to sizing data. We then describe the rank score

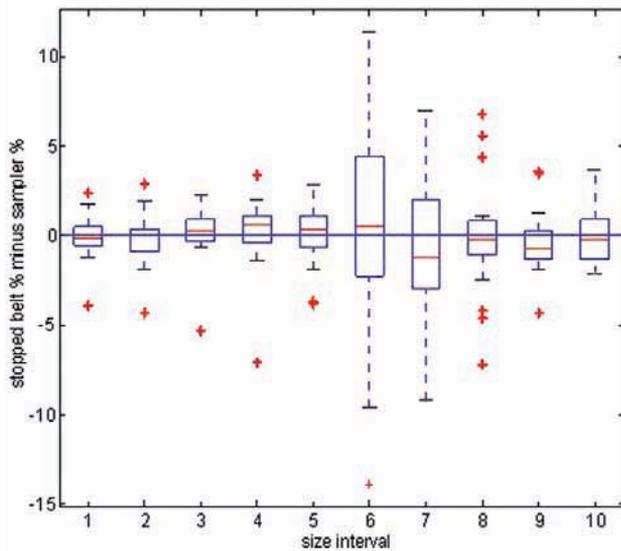


Figure 1—Boxplots of 28 size fraction (percentage) differences in each of 10 size classes. The centre line of the box indicates the median of the data values, while the upper and lower ends of a box are the 25th and 75th percentiles. The vertical dashed line extending from the upper (lower) end of a box marks the largest (smallest) data value that lies within 1.5 IQR of the upper (lower) end of the box (IQR=interquartile range, the spread between the 25th and 75th percentiles). Points outside these latter ranges are deemed to be outliers and are plotted as crosses

version of the statistic, illustrate its calculation on a small set of artificial data, and apply it to two real data sets. Following this, we use a Monte Carlo simulation method to compare the abilities of the tests to detect substantive differences between underlying size distributions. The results of the Monte Carlo study suggest that the proposed rank test is highly competitive with the Hotelling T^2 test. Finally, we make some concluding remarks and summarize our main results.

Hotelling's T^2 statistic

Set

$$\delta_{ki} = x_{ki} - y_{ki} \quad [5]$$

and form the data matrix of differences

$$\mathbf{D} = \begin{bmatrix} \delta_{11} & \delta_{12} & \dots & \delta_{1n} \\ \delta_{21} & \delta_{22} & \dots & \delta_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \delta_{p+1,1} & \delta_{p+1,2} & \dots & \delta_{p+1,n} \end{bmatrix} = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_{p+1} \end{bmatrix}$$

and the (column) vector of means

$$\bar{\delta} = (\bar{\delta}_1, \dots, \bar{\delta}_p)'$$

where $\bar{\delta}_k$ is the average of the n elements in row k of \mathbf{D} . Lyman *et al.* (2010) proposed a procedure based on Hotelling's T^2 statistic to test the statistical significance of the observed differences δ_{ki} by deleting any one of the rows of \mathbf{D} , say the last one δ_{p+1} , and computing the T^2 statistic on the remaining rows $\delta_1, \dots, \delta_p$. The T^2 statistic is

$$T^2 = \bar{\delta}' \mathbf{C}^{-1} \bar{\delta} \mathbf{x} (n-p) / p \quad [6]$$

where

$$\mathbf{C} = n^{-1} \sum_{k=1}^p (\delta_k - \bar{\delta}_k)' (\delta_k - \bar{\delta}_k) \quad [7]$$

denotes the covariance matrix. The reason for eliminating one row of \mathbf{D} from consideration is that the constraint Equation [2], which also applies to the y -data, together with Equation [5] implies that

$$\sum_{k=1}^{p+1} \delta_{k,j} = \sum_{k=1}^{p+1} x_{k,j} - \sum_{k=1}^{p+1} y_{k,j} = 0 \quad [8]$$

so that the covariance matrix \mathbf{C} is singular. This precludes calculation of T^2 on the full data matrix \mathbf{D} . It is important to note that the numerical value of T^2 computed in the manner suggested by Lyman *et al.* does not depend upon which one of the $p+1$ rows is eliminated from the data matrix. This fact is intuitively rather obvious because any p of the $p+1$ rows contain the same information as does the full set of $p+1$ rows—the missing row can be reconstructed exactly by applying the constraint (Equation [8]). Thus, the outcome of the statistical test is uniquely determined, no matter which one of the rows is eliminated from consideration.

The following equally important fact will also be used in this paper. Namely, that for any data matrix, not necessarily constrained as in Equation [2], and row δ_l the numerical value of T^2 computed on the p nonzero row difference $\delta_k - \delta_l$, $k \neq l$ does not depend upon l . For example, the numerical values of T^2 computed on the vectors

Comparing two mass size distributions

$$\delta_2 - \delta_1, \dots, \delta_{p+1} - \delta_1 \text{ and } \delta_1 - \delta_{p+1}, \dots, \delta_p - \delta_{p+1}$$

will be exactly the same. This follows from the fact that the numerical value of T^2 does not depend upon the ordering of rows in the data matrix \mathbf{D} .

A rank test

A version of the T^2 statistic that is robust to outliers is obtained upon replacing the differences δ_{ik} in row k of the data matrix by a rank score which is more or less immune to outlier effects. Towards this, notice that each δ_{ik} can be expressed as the product of its absolute value and its sign (+1 or -1),

$$\delta_{ki} = |\delta_{ki}| \times \text{sign of } \delta_{ki}. \quad [9]$$

In order to negate outlier effects one replaces $|\delta_{ki}|$ by its rank among the absolute values in row k ,

$$\text{rank of } |\delta_{ki}| = \text{number of } |\delta|$$

$$\text{in row } k \text{ that are } \leq |\delta_{ki}|.$$

Then, for example, the two largest absolute values in the row receive the ranks $n - 1$ and n , no matter how large they are compared to each other and to the other absolute values in the row. Appending the sign of δ_{ki} to its rank gives the rank score

$$s_{ki} = \text{rank of } |\delta_{ki}| \times \text{sign of } \delta_{ki}. \quad [10]$$

Comparing Equations [9] and [10] we see that robustness is effected by the replacement of the numerical value of $|\delta_{ik}|$, which may be an outlier relative to the other $|\delta|$ values in the row, with its rank, which is not an outlier relative to the other ranks in the row. In this way, each row $\delta_k = (\delta_{k1}, \dots, \delta_{kn})$ of the data matrix \mathbf{D} is replaced by a row of rank scores $s_k = (s_{k1}, \dots, s_{kn})$ to form a new data matrix

$$\mathbf{S} = \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \vdots \\ \mathbf{s}_{p+1} \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1n} \\ s_{21} & s_{22} & \dots & s_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p+1,1} & s_{p+1,2} & \dots & s_{p+1,n} \end{bmatrix}$$

The proposed rank statistic, denoted by T_w^2 , is obtained by choosing any one of the rows, say s_{p+1} , as a 'reference row' and calculating T^2 in Equation [6] on the reduced rank score matrix

$$\tilde{\mathbf{S}} = \begin{bmatrix} \mathbf{s}_1 - \mathbf{s}_{p+1} \\ \mathbf{s}_2 - \mathbf{s}_{p+1} \\ \vdots \\ \mathbf{s}_p - \mathbf{s}_{p+1} \end{bmatrix} = \begin{bmatrix} s_{11} - s_{p+1,1} & s_{12} - s_{p+1,2} & \dots & s_{1n} - s_{p+1,n} \\ s_{21} - s_{p+1,1} & s_{22} - s_{p+1,2} & \dots & s_{2n} - s_{p+1,n} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} - s_{p+1,1} & s_{p2} - s_{p+1,2} & \dots & s_{pn} - s_{p+1,n} \end{bmatrix}$$

The subscript W in T_w^2 serves to indicate that the rank scores are, in fact, those upon which the well-known Wilcoxon symmetry test is based. Notice that, in contrast to the δ_{ki} , the rank scores generally are not subject to the constraint $\sum_{k=1}^{p+1} s_{ki} = 0$. Hence, it would be possible in principle to compute T_w^2 on the full ranks score matrix \mathbf{S} . This is not advisable, though, because near-singularity of the corresponding covariance matrix is a frequent occurrence and gives rise to unnecessarily variable T_w^2 values. Eliminating one of the rank score row vectors from the calculation is also not advisable.

This is because, in the absence of the constraint $\sum_{k=1}^{p+1} s_{ki} = 0$, the value of T_w^2 will then depend upon which score vector is eliminated. On the other hand, as was pointed out in the Introduction, basing the calculation on the reduced rank score matrix $\tilde{\mathbf{S}}$ ensures a unique value of T_w^2 .

Numerical example

We illustrate the calculation of the rank scores on the following small set of artificial data:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \end{bmatrix} = \begin{bmatrix} .09 & .27 & .14 & .29 & .22 \\ .20 & .35 & .51 & .05 & .12 \\ .71 & .38 & .35 & .66 & .66 \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \end{bmatrix} = \begin{bmatrix} .14 & .33 & .10 & .26 & .20 \\ .21 & .27 & .40 & .07 & .17 \\ .65 & .40 & .50 & .67 & .63 \end{bmatrix}$$

Here, $p = 2$ and $n = 5$, that is, we have five paired sets of observations on each of three size fractions. The matrix of differences is

$$\begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix} = \begin{bmatrix} -.05 & -.06 & .04 & .03 & .02 \\ -.01 & .08 & .11 & -.02 & -.05 \\ .06 & -.02 & -.15 & -.01 & .03 \end{bmatrix}$$

From this we derive the matrix of signs

$$\begin{bmatrix} \text{sign}(\delta_1) \\ \text{sign}(\delta_2) \\ \text{sign}(\delta_3) \end{bmatrix} = \begin{bmatrix} -1 & -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 & 1 \end{bmatrix} \quad [11]$$

the matrix of absolute values

$$\begin{bmatrix} |\delta_1| \\ |\delta_2| \\ |\delta_3| \end{bmatrix} = \begin{bmatrix} .05 & .06 & .04 & .03 & 0.02 \\ 0.01 & .08 & .11 & .02 & .05 \\ .06 & .02 & .15 & .01 & .03 \end{bmatrix}$$

and the matrix of ranks

$$\begin{bmatrix} \text{ranks of } |\delta_1| \\ \text{ranks of } |\delta_2| \\ \text{ranks of } |\delta_3| \end{bmatrix} = \begin{bmatrix} 4 & 5 & 3 & 2 & 1 \\ 1 & 4 & 5 & 2 & 3 \\ 4 & 2 & 5 & 1 & 3 \end{bmatrix} \quad [12]$$

The rank score vectors are now found by multiplying corresponding elements of the matrices in Equations [11] and [12],

$$\begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \mathbf{s}_3 \end{bmatrix} = \begin{bmatrix} -4 & -5 & 3 & 2 & 1 \\ -1 & 4 & 5 & -2 & -3 \\ 4 & -2 & -5 & -1 & 3 \end{bmatrix}$$

and the Hotelling statistic is computed on the matrix

$$\begin{bmatrix} \mathbf{s}_1 - \mathbf{s}_3 \\ \mathbf{s}_2 - \mathbf{s}_3 \end{bmatrix} = \begin{bmatrix} -8 & -3 & 8 & 3 & -2 \\ -5 & 6 & 10 & -1 & -6 \end{bmatrix}$$

or on the matrix

$$\begin{bmatrix} \mathbf{s}_1 - \mathbf{s}_2 \\ \mathbf{s}_3 - \mathbf{s}_2 \end{bmatrix} = \begin{bmatrix} -3 & -9 & -2 & -1 & 4 \\ 5 & -6 & -10 & 1 & 6 \end{bmatrix}$$

Comparing two mass size distributions

or on the matrix

$$\begin{bmatrix} \mathbf{s}_2 - \mathbf{s}_1 \\ \mathbf{s}_3 - \mathbf{s}_1 \end{bmatrix} = \begin{bmatrix} 3 & 9 & 2 & 1 & -4 \\ 8 & 3 & -8 & -3 & 2 \end{bmatrix}$$

The value of T_w^2 is 0.915 in all three cases.

A Matlab (Mathworks Inc., 2007) program that does the required calculations for data sets of realistic sizes is available from either of the authors upon request.

Application to sizing data

For the data set exhibited in Figure 1 the observed value of the T_w^2 statistic is 2.980 with a p -value of 0.022 found from the F distribution with 9 and 19 degrees of freedom. (The p -value is the probability that the T_w^2 statistic exceeds the value 2.98 given that it has the indicated F distribution.) This p -value of 0.022 is quite different from the p -value of 0.124 produced by the T^2 statistic on this data. Thus, the rank test suggests strongly that the differences seen in Figure 1 are, in fact, real and not merely due to chance. Given this conclusion, an obvious question that arises is 'where do the differences mainly occur'? It was pointed out by Lyman *et al.* that some care should be exercised when attempting to answer this question. This is because a change in the mass fraction reported by any given size class must of necessity be accompanied by reported changes in one or more of the remaining size classes: the total of the reported fractions must be 1.00 under all circumstances. A glance at Figure 1 suggests that the reported proportions of coarser material (size fractions 7–10) have decreased while the proportions of finer material (size fractions 1–6) have increased correspondingly. This becomes clearer when we look at Figure 2, derived from Figure 1, which shows a plot of the medians of the 28 differences within each size class. Notwithstanding that the differences are relatively small, the pattern is clear.

Next, we turn to the data analyzed by Lyman *et al.* (2010). This involved a comparison of sizings from samples of iron ore obtained with two different cross-belt samplers, a Vezin-type belt-end sampler, and stopped belt sampling. A total of 14 pairs of sizings into 8 intervals from each procedure were available for analysis. Belt sampling was carried out by highly experienced staff and sample analysis (sizing) by a well-accredited laboratory that was closely supervised. The reproducibility of the sizing protocol was established prior to processing the test samples. Figure 3 (which is a reproduction of part of the top left hand one in Figure 3 of Lyman *et al.* (2010), except that the size fractions are now arranged left to right from finest to coarsest, shows boxplots of the 14 differences in each of the 8 size intervals for the belt-end sampler and the stopped belt samples. Student t -tests comparing each of the size fractions indicated a significant difference at only the finest size fraction. (We point out here that in Table 1 of Lyman *et al.* (2010) the numbers 0.160, 0.25, 2.68, and 0.64 in the last row of the belt-end block should be replaced by 0.887, 0.162, 14.6, and 5.48 respectively.) The Hotelling T^2 also indicated an overall significant difference (p -value = 0.05). Since a bias from cross-stream (Vezin) sampling was somewhat unexpected, the question was put whether outliers (indicated by a + in

Figure 3) could have been the cause of the significance of the values of the t and T^2 statistics. The answer seems to be that the observed differences are real—the rank test also gives a highly significant result for these data (p -value = 0.003). Some confirmation that the highly significant T^2 and T_w^2 values are at least in part due to the observed differences in the finest size fraction follows upon analysing the subcompositions $(x_{2i}, \dots, x_{8i}) / (1 - x_{1i})$ and $(y_{2i}, \dots, y_{8i}) / (1 - y_{1i})$ consisting of the size fractions in intervals 1 through 7 only. Then neither T^2 nor T_w^2 is significant (p -values of 0.54 and 0.31 respectively).

Technical note

The F -distribution with p and $n - p$ degrees of freedom attributed to the T^2 statistic and which is used to determine significance levels, requires a formal assumption that the matrix of differences originates from an underlying multivariate normal distribution. This assumption is hardly likely to be satisfied for sizing data, particularly since the

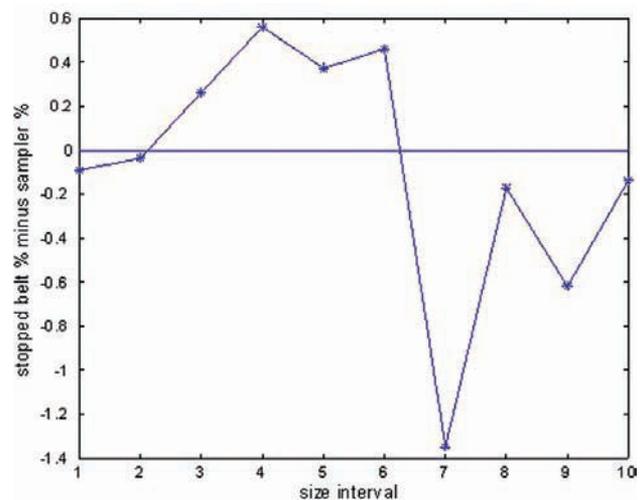


Figure 2—Medians of 28 pairs of size fraction differences (%) in 10 size classes (data from Figure 1)

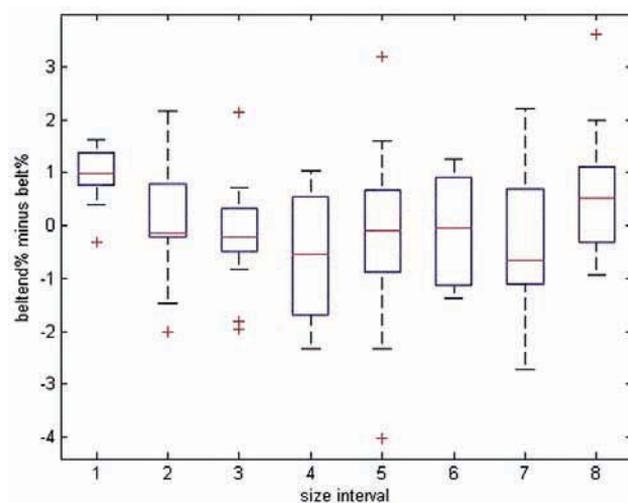


Figure 3—Boxplots of 14 belt-end and stopped belt size fraction differences in each of 8 size classes (from Lyman *et al.*, 2010)

Comparing two mass size distributions

entries in the vectors δ_k must all lie between -1 and $+1$. Nonetheless, extensive simulation results, some of which are reported in the next section, indicate that the F -distribution can be safely used. This can be explained partially by noting that the T^2 statistic depends on the raw data only through the mean vector $\bar{\delta}$ and the covariance matrix C , both of which are averages computed from the differences; see Equations [5] and [7]. Now, the central limit theorem guarantees approximate normality of averages, especially for difference data such as $\delta_k - \delta_l$ that are constrained to lie in a finite interval and that do not exhibit an excessively skewed distribution. Thus, it is perhaps not so surprising that the F -distribution is applicable to the data that we are dealing with, even when only relatively small amounts of such data are available.

Monte Carlo simulations

In this section of the paper we present the results of some Monte Carlo simulations with a view to (i) demonstrating the applicability of the F distribution when determining significance levels and (ii) comparing the statistical power of the rank test with that of the Hotelling test. The power of a statistical test (or a test statistic) is defined as the probability that the test will succeed in identifying a difference that is indeed present. A test with a power of 1 (or 100%) 'never fails'.

One approach towards these goals is to generate artificial data sets based on models that attempt to simulate the mechanisms that lead to an observed size distribution, such as particle breakage and particle sampling error (Brown and Wohletz (1995); Dacey and Krumbein (1979); Gy (1982); Lyman (1986)). However, since we have a number of real data sets at our disposal, we prefer to base our simulations directly on these.

In order to judge the applicability of the F distribution we require pairs of observed sizings that are generated from the same underlying size distribution and which differ only in respect of uncertainties added in the sampling and sieving operations. Given a data set consisting of n pairs of observed sizings (columns) (x_i, y_i) , $i = 1, \dots, n$, data sets with the required structure can be generated by interchanging at random the roles of x_i and y_i in the data matrix. That is, we flip an unbiased coin and if it shows heads we replace the pair (x_i, y_i) by (y_i, x_i) in the data matrix, otherwise leaving it unchanged. This means that we are sampling at random from a large collection of 2^n pairs of data sets of which any pair of columns is generated from identical underlying size distributions (the size distributions may vary from column pair to column pair). In the numerical example given earlier ($n = 5$), suppose five coin flips resulted in the outcome H, T, T, H, T . Then the corresponding randomized data matrix would be

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} .14 & .27 & .14 & .26 & .22 \\ .21 & .35 & .51 & .07 & .12 \\ .65 & .38 & .35 & .67 & .66 \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} .09 & .33 & .10 & .29 & .20 \\ .20 & .27 & .40 & .05 & .17 \\ .71 & .40 & .50 & .66 & .63 \end{bmatrix}$$

and the Hotelling and rank statistics would be computed on this data matrix in the manner indicated in the numerical example given earlier. This generation of randomized data sets is repeated a large number of times, say N times. We then compute the observed fractions of these N times in which each of the two statistics exceeded the upper $100(1 - \alpha)\%$ percentile c_α of the F distribution with p and $n - p$ degrees of freedom. If the F distribution is applicable, the fractions should be close to α . Table I shows the results obtained when applying this method to the data in Figure 1 ($n = 28$) and Figure 3 ($n = 14$) using in each instance $N = 50\,000$ trials and $\alpha = 0.10, 0.05$, and 0.01 . Clearly, the empirical exceedance probabilities of the rank test are quite close to the nominal values in all cases while the Hotelling statistic seems to undershoot the mark slightly.

To gauge the effect of outliers on the tests we again use the data from Figures 1 and 3, but now in a different manner. We replace in each interval those observed size fraction differences δ_{ki} that were identified as outliers by pseudo observations constructed as follows. Denote by iqr the interquartile range of the data in row k and denote by m their median. A Gaussian distribution of differences without outliers in row k would have a mean equal to m and a standard deviation equal to $0.69 \times iqr$. Notice that 0.69 is the ratio between the 68th (median + one standard deviation) and 75th percentiles of a Gaussian distribution. Each outlier in row k is now replaced by an observation from a Gaussian distribution which has mean m and standard deviation $\phi \times 0.69 \times iqr$. Here, ϕ is a variance inflation factor. The data set thus reconstructed has exactly the same configuration of median differences as the original. When $\phi = 1$, all outliers have essentially been removed. If ϕ is increased, more outliers are artificially introduced.

We illustrate the procedure using the data in the eighth size interval in Figure 1. There are five outliers, namely $-0.046, -0.042, 0.044, 0.056$, and 0.068 . The median and iqr of the differences in row 8 are -0.002 and 0.019 respectively. A Gaussian distribution with this median and iqr would have a mean of -0.002 and a standard deviation of $0.69 \times 0.019 = 0.013$. Accordingly, with $\phi = 1$ we generate five such Gaussian random numbers and insert them into column 8 in place of the original five outliers. If $\phi = 2$, we replace the outliers by five random numbers from a Gaussian distribution with mean -0.002 and standard deviation $2 \times 0.013 = 0.026$, etc. Figure 4 shows boxplots of the original data and those that resulted from one application of the above procedure at $\phi = 1$ (the pseudo observations were -0.022 ,

Table I

Observed and nominal significance levels when using the F distribution as approximation to the exact distributions of the T^2 and T_W^2 statistics

		$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
Figure 1 data	T^2	.007	.042	.088
$n = 28$	T_W^2	.010	.050	.101
Figure 3 data	T^2	.008	.042	.093
$n = 14$	T_W^2	.012	.049	.104

Comparing two mass size distributions

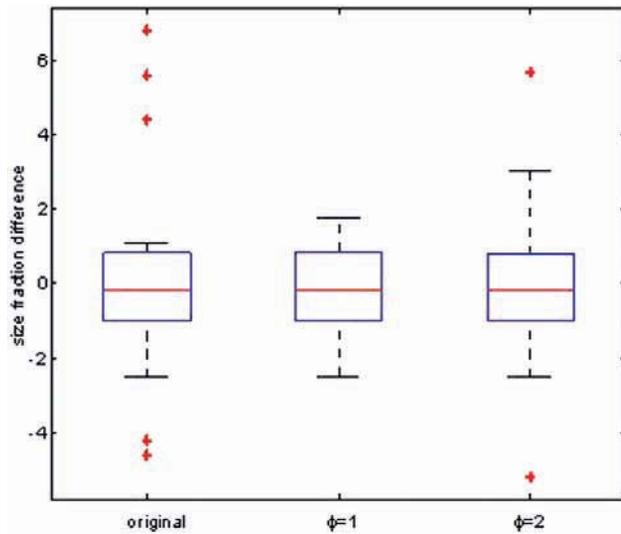


Figure 4—Demonstration of the removal of outliers from the original data and replacing them by Gaussian random numbers with standard deviation proportional to ϕ

–0.011, 0.010, 0.017, 0.018) and $\phi = 2$ (the pseudo observations were –0.050, –0.022, 0.008, 0.031, 0.057). We see that there are now no outliers at $\phi = 1$ and just two at $\phi = 2$.

As ϕ increases, so does the number of artificially generated outliers. Of course, after adding the perturbed values, we subtract from each column of differences its average in order that the necessary constraint $\sum_{k=1}^{p+1} \delta_{ki} = 0$ be satisfied for the new data. This process of replacing the original outliers is repeated N times for each of a range of values of ϕ . Then we find the fractions of the N times in

which each of the statistics exceeds c_α . These fractions are our estimates of the statistical powers of the tests. (Needless to say, this will be a fruitless exercise if neither of the tests gave a significant result on the original data set.) Table II shows the estimated powers of the tests when this methodology is applied to the data from Figures 1 and 3 with $\alpha = 0.05$ and $N = 10\,000$. Figure 5 shows for each value of ϕ in the first line of Table II (data from Figure 1) a boxplot of one randomly generated set of pseudo differences. Figure 6 shows the same in respect of the ϕ values in the fourth line of the table (data from Figure 3). A general conclusion from these simulation results is that the T^2 and T_W^2 statistics have similar powers when the numbers and sizes of outliers are not extensive, but that the former statistic loses power much more rapidly than the latter as the numbers and sizes of outliers increase. Overall, it would seem that the rank test is to be preferred because of its greater power robustness.

Concluding remarks

The data used in this study to illustrate the virtues of the ranked score version of the Hotelling procedure by

Table II

Statistical power of T^2 and T_W^2 tests at various noise levels

Figure 1 data $n = 28$	ϕ	1.0	1.5	2.0	3.0
	T^2	.89	.69	.52	.32
Figure 3 data $n = 14$	ϕ	1.0	3.0	4.0	5.0
	T^2	.89	.71	.51	.37
	T_W^2	.84	.75	.60	.48

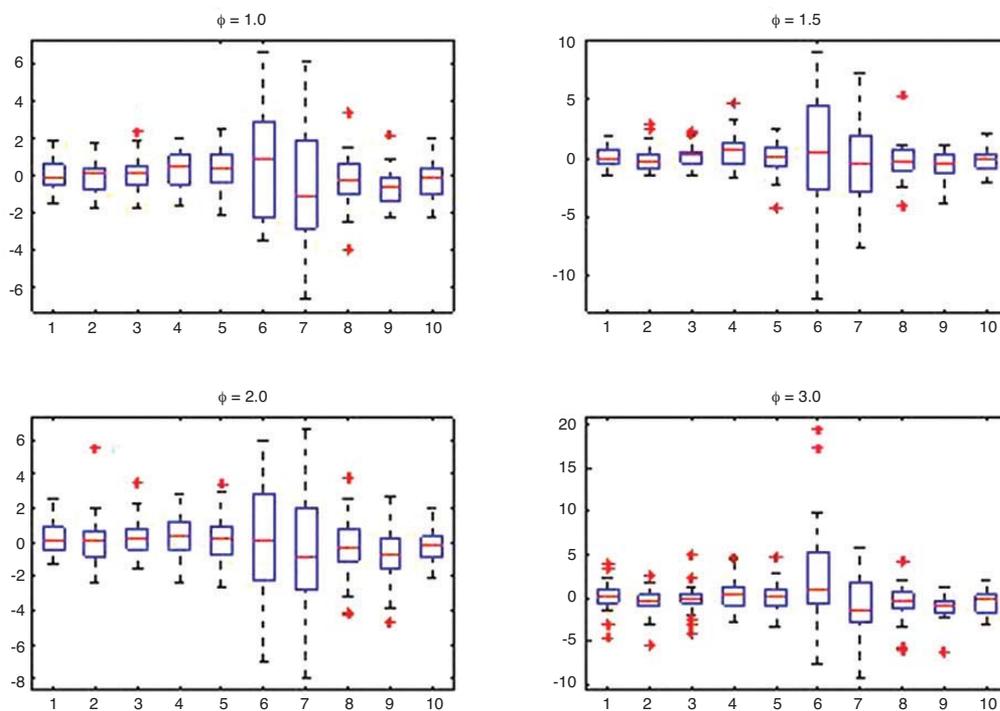


Figure 5—Typical distributions of size fraction differences for pseudo data constructed from Figure 1 data

Comparing two mass size distributions

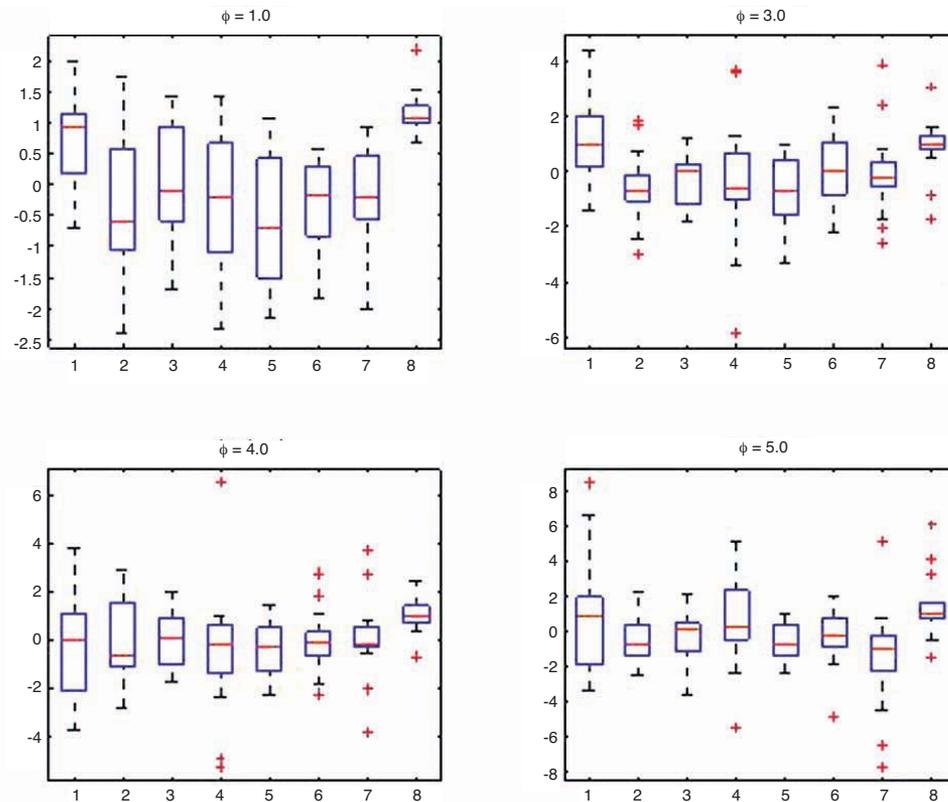


Figure 6—Typical distributions of size fraction differences for pseudo data constructed from the data in Figure 3

comparison to those of the conventional Hotelling test are real, and reflect the realities of bias testing using size distribution data. This contribution to the tools for making bias tests is motivated by the need to ensure that such tests are evaluated using the most powerful statistical tools.

However, it is critical in bias testing that the magnitude of the detectable bias be economically relevant. That is, the testing protocol, which includes sample masses collected, precision of the particle sizing process, the number of paired samples collected, and the statistical analysis procedure must be adequate to detect a difference between the test samples and the reference samples that is economically significant. Too often, a bias test is carried out using the guidelines provided by a standard. The accuracy of the work carried out is not specifically assessed and so the extent of the detectable bias is not quantified. A bias test which is not qualified by the extent of the detectable bias is no test at all, and may give comfort where none is to be had. Bias tests must be designed to detect a certain level of bias that has been based on an economic evaluation of the critical level of bias to be detected for all analytes of interest. Consideration of the size-by-size analyses of the material sampled for all economically important analytes, coupled with a critical consideration of the contractual tolerances on analytes, can be used to arrive at the critical level of bias to be detected in a test.

The authors continue to work towards a workable methodology for accurate assessment of the detectable bias over the full sizing distribution.

Summary

In this paper we propose a robust procedure based on ranks

to test for sizing bias in a mechanical sampling system. The data consists of observed mass size distributions and our results are applicable to such data in general. The proposed rank test has been applied to two data sets and its advantages over Hotelling's T^2 statistic has been illustrated in a Monte Carlo simulation study using a very realistic method of injecting outliers into the data set. In particular, and in contrast to the T^2 statistic, the power of the rank test is not unduly affected by the presence of outliers. Our simulation results indicate that use of the F -distribution to compute p -values is permissible, even if relatively few sizing pairs are available. The testing carried out in this paper clears the way for the ranked Hotelling test to become a reliable standard tool in size distribution comparison, which comparison is a best tool for bias testing of sampling systems.

References

- BROWN, W.K. and WOHLTZ, K.H. 1995. Derivation of the Weibull distribution based on physical principles and its connection to the Rosin-Rammler and lognormal distributions. *Journal of Applied Physics*, vol. 78, no. 4. pp. 2758–2763.
- DACEY, M.F. and KRUMBEIN, W.C. 1979. Models of breakage and selection for particle size distributions. *Mathematical Geology*, vol. 11. pp. 193–222.
- GY, P.M. 1982. *Sampling of Particulate Materials—Theory and Practice*, Elsevier, New York.
- LYMAN, G.J., NEL, M., LOMBARD, F., STEINHAUS, R., and BARTLETT, H. 2010. Bias testing of cross-belt samplers. *Journal of the Southern African Institute of Mining and Metallurgy*, vol. 110, no. 6. pp. 289–298.
- LYMAN, G.J. 1986. Application of Gy's sampling theory to coal: A simplified explanation and illustration of some basic aspects. *International Journal of Mineral Processing*, vol. 17. pp. 1–22.
- Mathworks Inc. 2007. *Matlab, Version 7.5 (release 2007b)*. ◆