CLUSTERING ESTIMATES OF EFFECT MAGNITUDE FROM INDEPENDENT STUDIES

BY

LARRY V. HEDGES and INGRAM OLKIN

TECHNICAL REPORT NO. 173
APRIL 1981

PREPARED UNDER THE AUSPICES OF
NATIONAL SCIENCE FOUNDATION GRANT
MCS 78-07736

Ingram Olkin, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
CLUSTERING ESTIMATES OF EFFECT MAGNITUDE
FROM INDEPENDENT STUDIES

by

Larry V. Hedges and Ingram Olkin

TECHNICAL REPORT NO. 173
April 1981

PREPARED UNDER THE AUSPICES OF
NATIONAL SCIENCE FOUNDATION GRANT
MCS78-07736

Ingram Olkin, Project Director

1This research was supported in part by the Spencer
Foundation and by the National Science Foundation.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
CLUSTERING ESTIMATES OF EFFECT MAGNITUDE
FROM INDEPENDENT STUDIES

1. **Introduction**

There has recently been an increasing interest in the estimation of the magnitude of effects in educational and psychological research. An indicator of this interest is the fact that at least one journal now requests that authors report a measure of effect size along with their results. Part of the motivation for the interest has been the growing realization that the results of statistical significance tests alone do not adequately summarize the results of a study (see e.g., Bakan, 1966 or Morrison and Henkel, 1970). A second impetus for the current interest in measures of effect magnitude comes from the research synthesis movement. Judgments of research reviewers have frequently depended on counts of the number of studies that yielded statistically significant results (Light and Smith, 1971). Such vote-counting methods are shown to have undesirable statistical properties (e.g., vanishing power) in most research situations by Hedges and Olkin (1980). Glass (1976) advocates the use of quantitative indices of effect magnitude to synthesize the results of a series of studies. His procedure involves the estimation of effect magnitude for each study as either a correlation coefficient or a standardized mean difference. The estimates from each study are then averaged to obtain an overall estimate of effect magnitude. The implicit assumption in this procedure is that the estimates of effect magnitude from each study share the same population effect magnitude, i.e., that the estimates estimate the same thing.

Some statistical theory for estimating standardized mean differences is developed by Hedges (1980, 1981). The development of sound statistical procedures for estimating effect size, including tests for homogeneity of effect sizes across a series of experiments soon led to cases of obvious heterogeneity
of effect sizes among studies that purported to be replications. For example, Hedges and Gage (1980) found that a series of studies on open education yielded heterogeneous effects sizes on all nineteen variables considered. Humphreys (1979) has commented on similar findings in a different context. For concreteness let $\theta_1, \ldots, \theta_k$ be the indices of true (population) relationship between two variables. In this paper we are concerned with $\theta$'s that are either correlation coefficients or standardized mean differences. If the true relationship between variables is not the same for all studies in a collection, i.e., if $\theta_1 = \theta_2 = \ldots = \theta_k$ is false, then we may desire statistical procedures to accomplish any of several goals. One goal is to partition the heterogeneous populations into homogeneous groups of clusters based on sample data such that the $\theta$'s are close within clusters but differ between clusters. A second goal is suggested by the statistical literature on ranking and selection (Gibbons, Olkin, and Sobel, 1977), where the issue is finding statistical procedures for using sample data to rank populations so that $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_k$ or to select the best population (the one with the largest $\theta_1$). A third goal for statistical procedures is to identify the deviant $\theta_1$ when at most one $\theta_i$ is known to differ from the others. This last model is known as a slippage-model (Karlin and Truex, 1960).

This paper examines only the first problem posed above, that of clustering populations. It is an exposition of two clustering procedures for grouping estimates of effect magnitude (correlation coefficients or standardized mean differences) into homogeneous classes. The two procedures decompose a series of correlations or standardized mean differences into disjoint or overlapping classes respectively. In each case statistical significance levels are provided for the clusters. Both procedures are based on clustering theory for standard normal random variables. This theory is used to cluster
correlations and standardized mean differences by first transforming the estimators to be approximately standard normal variates.

2. Notation and Model

Consider a series of \( k \) studies each of which may be considered a replication of the others. We assume that each of the studies in the series produces an estimate of effect magnitude, either a correlation coefficient or a standardized mean difference.

2.1 Correlation Coefficient

Suppose the \( i \)th study measures two variables \( X_i \) and \( Y_i \) that have a bivariate normal distribution; let \( r_i \) and \( \rho_i \) denote the sample and population correlation coefficients between \( X_i \) and \( Y_i \) obtained in the \( i \)th study. Let \( n_i \) denote the sample size of the \( i \)th study. It will usually be the case that the variables \( X_i, i = 1, \ldots, k \) "measure the same characteristic," by some definition, i.e., are linearly equatable or have a single common factor and no unique factors, and the same will be true for the \( Y_i, i = 1, \ldots, k \). Note that this is a conceptual, rather than statistical condition.

2.1.1 Variance Stabilizing Transformation of \( r \)

Since \( r_i \) is a sample correlation coefficient with sample size \( n_i \), the well known Fisher z-transformation,

\[
(2.1) \quad z(r_i) = \frac{1}{2} \log[(1 + r_i)/(1 - r_i)] = \tanh^{-1} r_i
\]

stabilizes the variance of \( r_i \). In particular
\[ \sqrt{n_i - 3} \left[ z(r_i) - z(p_i) \right] \]

has an approximate standard normal distribution.

2.2 Standardized Mean Difference

Suppose the \( i \)th study uses an experimental (E) group and control (C) group with equal sizes \( n_i \). In the \( i \)th study there is an outcome variable \( w_i \) that is assumed to be normally distributed within the two groups with means \( \mu_i^E \) and \( \mu_i^C \), respectively, and common variance \( \sigma_i^2 \). The sample means of the experimental and control groups are denoted by \( \bar{w}_i^E \), \( \bar{w}_i^C \) and the pooled within group variance is denoted by \( s_i^2 \). The effect size for the \( i \)th experiment is the population standardized mean difference

\[ \delta_i = (\mu_i^E - \mu_i^C)/\sigma_i. \]

An estimate of the population effect size \( \delta_i \) is the standardized mean difference in the \( i \)th sample, namely

\[ (2.2) \quad \hat{\delta}_i = (\bar{w}_i^E - \bar{w}_i^C)/s_i. \]

We note in passing that Hedges (1981) has shown that \( \hat{\delta}_i \) is a consistent estimator of \( \delta_i \) with bias of less than 2 per cent when \( n_i > 20 \). As in the case of correlations, the variables \( w_i, i = 1, \ldots, k \) usually measure the same thing by some definition such as linear equatability.

2.2.1 A Variance Stabilizing Transformation for \( \hat{\delta}_i \)

The estimator \( \hat{\delta}_i \) has an approximate normal distribution with mean \( \delta_i \) and variance \( (2 + \delta_i^2/4)/n_i \) (Hedges, 1981). The variance stabilizing transformation \( h \) is obtained by solving the differential equation

\[ dh = \frac{d\delta}{\sqrt{\delta^2 + 8}} \]
and leads to the transformation

\[ h(x) = \sinh^{-1}\left(\frac{x}{2\sqrt{2}}\right) = \log\left(\frac{x + \sqrt{x^2 + 8}}{2\sqrt{2}}\right). \]

Therefore

\[ 2\sqrt{n} \left( h(\hat{\delta}_1) - h(\delta_1) \right) \]

is distributed approximately as a standard normal variable.

3. Clustering Theory for Standard Normal Random Variables

Suppose that \( U_1, \ldots, U_k \) is a sample from a standard normal distribution and \( U(1), \ldots, U(k) \) are the \( k \) values arranged in ascending order of magnitude, i.e., \( U(1) \leq U(2) \leq \cdots \leq U(k) \). These ordered values (order statistics) define \( k-1 \) "gaps" of lengths \( U(t+1) - U(t), \ t = 1, \ldots, k-1, \) between the values. Tukey (1949) showed that the expected number of gaps that exceed a prespecified length \( d \) is

\[ p = k \int_{-\infty}^{\infty} \left[ [\phi(x) + 1 - \phi(x + d)]^{k-1} - [\phi(x)]^{k-1} \right] d\phi(x), \]

where \( \phi(x) \) is the standard normal cumulative distribution function.

In the above we assume that all the \( k \) variates have a distribution with the same mean. If we declare that a gap greater than \( d \) is significantly large, we will be incorrect 100\( \alpha \) per cent of the time when all the \( k \) variates have the same mean. When all of the variates do not have the same mean, some of the gaps will be expected to be larger. This provides the basis for a significance
test for clusterings based on gaps. That is, if the \( k \) variates are divided into two clusters at a gap of length \( d \), the clustering is significant at the 100\( \alpha \) per cent level if the probability of a gap as large as \( d \) is less than 100\( \alpha \) when all variates have the same distribution. Table 1 provides a tabulation of the minimum significant gap length for various significance levels \( \alpha \).

3.1 Disjoint Clustering

The procedure for forming disjoint clusters from \( k \) variates consists in successively splitting at each of the gaps. The first step is to partition the \( k \) variates into two groups at the largest gap. The next division occurs at the second largest gap and divides one of the two groups obtained in the first step. The process can continue until there are \( k \) "groups," i.e., all of the variates are in a group by themselves. Alternatively, we may stop the clustering process when the gaps no longer exceed the 100(1-\( \alpha \)) percent critical value for some significance level \( \alpha \).

An Example. The process is illustrated by an example with \( k = 5 \) and the observations arranged in ascending order:

\[
U(1) = 1.7, \quad U(2) = 3.4, \quad U(3) = 6.9, \quad U(4) = 8.5, \quad U(5) = 10.3
\]

Then the gaps are

\[
U(2) - U(1) = 1.7, \quad U(3) - U(2) = 3.5, \quad U(4) - U(3) = 1.6, \quad U(5) - U(4) = 1.8
\]

The first step is to step separate the five variates into two groups at the largest gap between \( U(2) \) and \( U(3) \) to form the groups \( \{ U(1), U(2) \} \) and \( \{ U(3), U(4), U(5) \} \). The next step is to separate the variates into groups at the next largest gap between \( U(4) \) and \( U(5) \) to yield the groups \( \{ U(1), U(2) \}, \)
<table>
<thead>
<tr>
<th>( k )</th>
<th>.40</th>
<th>.30</th>
<th>.20</th>
<th>.10</th>
<th>.05</th>
<th>.01</th>
<th>.005</th>
<th>.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1.37</td>
<td>1.56</td>
<td>1.80</td>
<td>2.17</td>
<td>2.48</td>
<td>3.19</td>
<td>3.46</td>
<td>4.01</td>
</tr>
<tr>
<td>4</td>
<td>1.34</td>
<td>1.49</td>
<td>1.69</td>
<td>2.00</td>
<td>2.28</td>
<td>2.90</td>
<td>3.14</td>
<td>3.64</td>
</tr>
<tr>
<td>5</td>
<td>1.28</td>
<td>1.41</td>
<td>1.58</td>
<td>1.87</td>
<td>2.13</td>
<td>2.70</td>
<td>2.93</td>
<td>3.42</td>
</tr>
<tr>
<td>6</td>
<td>1.22</td>
<td>1.34</td>
<td>1.50</td>
<td>1.77</td>
<td>2.02</td>
<td>2.56</td>
<td>2.79</td>
<td>3.26</td>
</tr>
<tr>
<td>7</td>
<td>1.17</td>
<td>1.28</td>
<td>1.43</td>
<td>1.69</td>
<td>1.93</td>
<td>2.47</td>
<td>2.68</td>
<td>3.15</td>
</tr>
<tr>
<td>8</td>
<td>1.13</td>
<td>1.23</td>
<td>1.38</td>
<td>1.62</td>
<td>1.85</td>
<td>2.39</td>
<td>2.60</td>
<td>3.06</td>
</tr>
<tr>
<td>9</td>
<td>1.09</td>
<td>1.19</td>
<td>1.33</td>
<td>1.57</td>
<td>1.81</td>
<td>2.32</td>
<td>2.53</td>
<td>2.99</td>
</tr>
<tr>
<td>10</td>
<td>1.05</td>
<td>1.15</td>
<td>1.29</td>
<td>1.53</td>
<td>1.76</td>
<td>2.28</td>
<td>2.48</td>
<td>2.94</td>
</tr>
<tr>
<td>11</td>
<td>1.03</td>
<td>1.12</td>
<td>1.26</td>
<td>1.49</td>
<td>1.72</td>
<td>2.23</td>
<td>2.44</td>
<td>2.89</td>
</tr>
<tr>
<td>12</td>
<td>1.00</td>
<td>1.09</td>
<td>1.23</td>
<td>1.46</td>
<td>1.69</td>
<td>2.20</td>
<td>2.40</td>
<td>2.85</td>
</tr>
<tr>
<td>13</td>
<td>0.98</td>
<td>1.07</td>
<td>1.20</td>
<td>1.43</td>
<td>1.66</td>
<td>2.16</td>
<td>2.35</td>
<td>2.77</td>
</tr>
<tr>
<td>14</td>
<td>0.96</td>
<td>1.05</td>
<td>1.18</td>
<td>1.41</td>
<td>1.64</td>
<td>2.13</td>
<td>2.32</td>
<td>2.73</td>
</tr>
<tr>
<td>15</td>
<td>0.94</td>
<td>1.03</td>
<td>1.16</td>
<td>1.39</td>
<td>1.61</td>
<td>2.10</td>
<td>2.29</td>
<td>2.70</td>
</tr>
<tr>
<td>20</td>
<td>0.86</td>
<td>0.94</td>
<td>1.06</td>
<td>1.26</td>
<td>1.46</td>
<td>1.88</td>
<td>2.05</td>
<td>2.46</td>
</tr>
<tr>
<td>50</td>
<td>0.57</td>
<td>0.64</td>
<td>0.74</td>
<td>0.93</td>
<td>1.14</td>
<td>1.60</td>
<td>1.78</td>
<td>2.20</td>
</tr>
</tbody>
</table>
\{U(3), U(4)\} and \{U(5)\}. The third step separates \(U(1)\) from \(U(2)\) to yield the groups \(\{U(1)\}, \{U(2)\}, \{U(3), U(4)\}, \{U(5)\}\). The final step yields groups \(\{U(1)\}, \{U(2)\}, \{U(3)\}, \{U(4)\}, \{U(5)\}\). This process is illustrated in Figure 1.

\[
\begin{align*}
\text{Step 4} & \quad (d = 1.6, \ p < .20) \\
\text{Step 3} & \quad (d = 1.7, \ p < .20) \\
\text{Step 2} & \quad (d = 1.8, \ p < .20) \\
\text{Step 1} & \quad (d = 3.5, \ p < .05)
\end{align*}
\]

\[
\begin{align*}
U(1) = 1.7 & \quad U(2) = 3.4 & \quad U(3) = 6.9 & \quad U(4) = 8.5 & \quad U(5) = 10.3
\end{align*}
\]

Figure 1. Cluster map of hierarchical decomposition of five variates into disjoint clusters.

Examination of Table 1 shows that with \(k = 5\) and \(\alpha = .05\), the minimum significant gap length is \(d = 2.13\), so only the first step clustering is significant at the \(\alpha = .05\) level.

3.2 Clustering into Overlapping Clusters

A different clustering scheme can be used to divide the studies into overlapping clusters. The procedure is analogous to the Bonferroni multiple comparison test described by Dunn (1961) for testing the difference between means in the analysis of variance. The procedure used to form these clusters at a desired significance level \(\alpha\) depends on the computed 100(1-\(\alpha\))\% critical value \(c(k)\) of the range of \(k\) observations from a standard normal distribution, where \(\alpha = [1-(1-\alpha)^{k-1}]\). Table 2 is the compilation of critical values \(c(k)\) for various \(\alpha\) levels.

The procedure starts with the \(k\) variates arranged in ascending order; a comparison \(U(k) - U(1)\) is made between largest of the \(k\) variates \(U(k)\)
TABLE 2

Critical Values, c(k), for Bonferroni Multiple Comparisons of the Range of Standard Normal Variates for Overall Level of Significance $\alpha$

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>.40</th>
<th>.30</th>
<th>.20</th>
<th>.10</th>
<th>.05</th>
<th>.01</th>
<th>.005</th>
<th>.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.19</td>
<td>1.47</td>
<td>1.81</td>
<td>2.33</td>
<td>2.77</td>
<td>3.64</td>
<td>3.97</td>
<td>4.65</td>
</tr>
<tr>
<td>3</td>
<td>2.33</td>
<td>2.57</td>
<td>2.87</td>
<td>3.30</td>
<td>3.68</td>
<td>4.42</td>
<td>4.71</td>
<td>5.32</td>
</tr>
<tr>
<td>4</td>
<td>2.95</td>
<td>3.17</td>
<td>3.43</td>
<td>3.83</td>
<td>4.17</td>
<td>4.85</td>
<td>5.11</td>
<td>5.73</td>
</tr>
<tr>
<td>5</td>
<td>3.37</td>
<td>3.57</td>
<td>3.82</td>
<td>4.18</td>
<td>4.50</td>
<td>5.15</td>
<td>5.39</td>
<td>5.95</td>
</tr>
<tr>
<td>6</td>
<td>3.68</td>
<td>3.86</td>
<td>4.10</td>
<td>4.44</td>
<td>4.75</td>
<td>5.38</td>
<td>5.62</td>
<td>6.15</td>
</tr>
<tr>
<td>7</td>
<td>3.92</td>
<td>4.10</td>
<td>4.32</td>
<td>4.65</td>
<td>4.95</td>
<td>5.55</td>
<td>5.80</td>
<td>6.30</td>
</tr>
<tr>
<td>8</td>
<td>4.12</td>
<td>4.29</td>
<td>4.50</td>
<td>4.83</td>
<td>5.11</td>
<td>5.71</td>
<td>5.93</td>
<td>6.45</td>
</tr>
<tr>
<td>9</td>
<td>4.28</td>
<td>4.45</td>
<td>4.66</td>
<td>4.97</td>
<td>5.25</td>
<td>5.82</td>
<td>6.05</td>
<td>6.54</td>
</tr>
<tr>
<td>10</td>
<td>4.43</td>
<td>4.59</td>
<td>4.79</td>
<td>5.10</td>
<td>5.37</td>
<td>5.94</td>
<td>6.16</td>
<td>6.65</td>
</tr>
<tr>
<td>11</td>
<td>4.55</td>
<td>4.71</td>
<td>4.91</td>
<td>5.21</td>
<td>5.48</td>
<td>6.04</td>
<td>6.26</td>
<td>6.74</td>
</tr>
<tr>
<td>12</td>
<td>4.67</td>
<td>4.82</td>
<td>5.02</td>
<td>5.31</td>
<td>5.57</td>
<td>6.12</td>
<td>6.34</td>
<td>6.82</td>
</tr>
<tr>
<td>13</td>
<td>4.77</td>
<td>4.92</td>
<td>5.11</td>
<td>5.40</td>
<td>5.66</td>
<td>6.20</td>
<td>6.41</td>
<td>6.89</td>
</tr>
<tr>
<td>14</td>
<td>4.86</td>
<td>5.01</td>
<td>5.20</td>
<td>5.48</td>
<td>5.74</td>
<td>6.27</td>
<td>6.48</td>
<td>6.95</td>
</tr>
<tr>
<td>15</td>
<td>4.94</td>
<td>5.09</td>
<td>5.27</td>
<td>5.55</td>
<td>5.80</td>
<td>6.34</td>
<td>6.55</td>
<td>7.01</td>
</tr>
<tr>
<td>20</td>
<td>5.28</td>
<td>5.42</td>
<td>5.59</td>
<td>5.86</td>
<td>6.10</td>
<td>6.59</td>
<td>6.81</td>
<td>7.26</td>
</tr>
<tr>
<td>50</td>
<td>6.22</td>
<td>6.33</td>
<td>6.48</td>
<td>6.72</td>
<td>6.93</td>
<td>7.38</td>
<td>7.55</td>
<td>7.97</td>
</tr>
</tbody>
</table>
and the smallest $U(1)$. If this difference is larger than the critical value $c(k)$, test the difference $U(k) - U(2)$ between $U(k)$ and $U(2)$ using the critical value $c(k-1)$. If this is significant, test $U(k) - U(3)$ using the critical value $c(k-2)$. Continue testing $U(k)$ against successively larger variates until one of the differences is less than the appropriate critical value. The procedure is then repeated using $U(k-1)$ instead of $U(k)$, that is test whether the difference $U(k-1) - U(1)$ is significant using the critical value $c(k-1)$. Proceed in this way until you reach the stage of testing $U(2)$ versus $U(1)$.

An Example. Suppose that $k = 5$,

\[
U(1) = 1.7, \ U(2) = 3.4, \ U(3) = 6.9, \ U(4) = 8.5, \ U(5) = 10.3,
\]

and we use the 1% level of significance. In the first step we calculate $U(5) - U(1) = 8.6$. Since this is larger than $c(5) = 5.15$, calculate $U(5) - U(2) = 6.9$. Comparison of 6.9 with $c(4) = 4.85$ implies that we must compute $U(5) - U(3) = 4.4$, which is smaller than $c(3) = 4.42$.

Therefore $U(5)$ is significantly different from $U(1)$ and $U(2)$ but not $U(3)$ and $U(4)$. Next compute $U(4) - U(1) = 6.8$, which is larger than $c(4) = 4.85$. Thus we must compute $U(4) - U(2) = 5.1$ which is larger than $c(3) = 4.42$, so we compute $U(4) - U(3) = 1.6 < c(2) = 3.64$. This implies that $U(4)$ is significantly different from $U(1)$ and $U(2)$, but not from $U(3)$. The next step is to compute $U(3) - U(1) = 5.2 > c(3) = 4.42$ and $U(3) - U(2) = 3.5 < c(2) = 3.64$. Thus we conclude that $U(3)$ differs significantly from $U(1)$ but not from $U(2)$. Finally $U(2) - U(1) = 1.7 < c(2) = 3.64$, so $U(2)$ does not differ significantly from $U(1)$. 

10
These results are summarized in the following diagram, by listing the data and underscoring the variates that do not differ significantly.

\[
\begin{align*}
1.7 & \quad 3.4 & \quad 6.9 & \quad 8.5 & \quad 10.3 \\
\end{align*}
\]

Note that in this case, there are three overlapping clusters, \( \{U_{(1)}, U_{(2)}\} \), \( \{U_{(2)}, U_{(3)}\} \), and \( \{U_{(3)}, U_{(4)}, U_{(5)}\} \).

4. **Clustering Correlation Coefficients**

The clustering procedure described above can be applied to a series of \( k \) correlation coefficients by using the Fisher z-transformation (2.1). Define the weighted average \( \bar{z} \) via

\[
\bar{z} = \frac{\sum_{i=1}^{k} (n_i - 3)z(r_i)}{\sum_{i=1}^{k} (n_i - 3)}.
\]

A test for overall homogeneity of the \( \rho_i \), i.e., a test of

\[
H_0: \rho_1 = \rho_2 = \cdots = \rho_k,
\]

is provided by the statistic

\[
q_k = \frac{1}{\sum_{i=1}^{k} (n_i - 3)} \left[ z(r_i) - \bar{z} \right]^2.
\]

Under \( H_0 \), the statistic \( q_k \) has a chi-square distribution with \((k-1)\) degrees of freedom. This statistic can be calculated before any clustering of correlations is attempted. If the obtained value of the statistic \( q_k \)
is greater than the critical value of the chi-square distribution with 
(k-1) degrees of freedom, then the correlations are considered as samples 
from different populations. The clustering procedures are then used to 
determine populations for which the correlations may be considered to be 
homogeneous.

To carry out this procedure we first transform the data. If the 
sample sizes are equal, then let \( U'_i = \sqrt{n-3} \, z(r_{i}) \) for \( i = 1, \ldots, k \) in which 
the \( U'_i \) are independently distributed with a common mean and unit 
variance. To remove the effect of the mean, we can use \( U_i = \sqrt{n-3} \, (z(r_{i}) - \bar{z}) \) 
in which case the \( U_i \) have means zero and unit variance. Although the 
\( U_i \) are correlated, for large samples we treat these as independent.

If the sample sizes \( n_1, \ldots, n_k \) differ, we recommend replacing each 
\( n_i \) by the square-mean-root 
\[
n' = \left( \frac{1}{k} \, \sqrt{n_i - 3/k} \right)^2.
\]
Thus the clustering procedure is actually applied to the variates \( U_i = \sqrt{n_i' \, (z(r_{i}) - \bar{z})} \), \( i = 1, \ldots, k \), which have approximately standard normal distributions 
under \( H_0 \).

4.1 An Example for Clustering Correlations

Burt (1955) presented data consisting of correlations of IQ scores for 
identical twins reared together, identical twins reared apart, dizigotic 
twins reared together, siblings reared apart and unrelated children reared 
together. Although the data have since been questioned, it is interesting 
to apply our clustering methods to these correlations. Table 3 is a presentation
TABLE 3

Data for the Example of Clustering of Correlation Coefficients

<table>
<thead>
<tr>
<th>Correlation</th>
<th>$n_i$</th>
<th>$r_i$</th>
<th>$z(r_i)$</th>
<th>$\sqrt{n_i} (z(r_i) - \bar{z})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identical Twins Reared Together</td>
<td>83</td>
<td>.925</td>
<td>1.623</td>
<td>11.066</td>
</tr>
<tr>
<td>Identical Twins Reared Apart</td>
<td>21</td>
<td>.876</td>
<td>1.358</td>
<td>8.185</td>
</tr>
<tr>
<td>Dizigotic Reared Together</td>
<td>173</td>
<td>.551</td>
<td>.620</td>
<td>.163</td>
</tr>
<tr>
<td>Siblings Reared Apart</td>
<td>131</td>
<td>.517</td>
<td>.572</td>
<td>- .359</td>
</tr>
<tr>
<td>Unrelated Children Reared Together</td>
<td>287</td>
<td>.269</td>
<td>.276</td>
<td>-3.576</td>
</tr>
</tbody>
</table>

Note: These data were obtained from Burt, 1955.
of the raw data, and the z-transformed values for the correlations. The value of $\bar{z}$, the weighted average correlation, is $\bar{z} = .605$. An overall chi-square test for equality of correlations yields a highly significant value.

To carry out a clustering we first adjust for different sample sizes by computing the square-mean-root to obtain the value $n' = 118.16$. The values of the variates $U_i = n' (z(r_i) - \bar{z})$ are also given in Table 3. The gaps are

$$U_5 - U_4 = 2.881, \quad U_4 - U_3 = 8.022, \quad U_3 - U_2 = .521, \quad U_2 - U_1 = 3.218.$$  

Comparison of these gaps with the critical value 2.70 for $k = 5$, $\alpha = .01$ given in Table 1 suggests that every gap is significant except the gap between $U_3$ and $U_2$. Therefore the clusters obtained by this disjoint clustering method at the $\alpha = .01$ critical level are $\{U_1\}, \{U_2, U_3\}, \{U_4\}$, and $\{U_5\}$. Note that the only cluster larger than one contains the correlations for dizigotic twins reared together and siblings reared apart.

The overlapping clustering method yields slightly different clusters at the $\alpha = .01$ significance level. We see that $U_5 - U_3 = 10.903 > c(3) = 4.42$, but $U_5 - U_4 = 2.881 < c(2) = 3.64$. Therefore $\{U_5, U_4\}$ forms one cluster. Similarly $U_3 - U_1 = 3.739 < c(3) = 4.42$ but $U_3 - U_2 = .521 < c(2) = 3.64$. Therefore $\{U_1, U_2, U_3\}$ form a second cluster. Finally note that $U_4 - U_3 = 8.022 > c(2) = 3.64$ and $U_2 - U_1 = 3.218 < c(2) = 3.64$, so the first two clusters are the only clusters. We see that for these data, the overlapping clustering procedure has yielded nonoverlapping groups. One cluster consists of correlations between
identical twins either reared together \( U_{(1)} \) or reared apart \( U_{(2)} \).
The other cluster consists of correlations for pairs of individuals who are not nonidentical twins.

5. Clustering Standardized Mean Differences

The clustering procedures described for unit normal variates can also be applied to the problem of clustering standardized mean differences. Suppose \( \hat{\delta}_1, \ldots, \hat{\delta}_k \) are \( k \) standardized mean differences defined by (2.2), and \( \bar{h} \) is the weighted average of the transformed \( \hat{\delta}_i \)'s.

\[
\bar{h} = \frac{1}{\sum n_i h(\hat{\delta}_i)/2n_i},
\]

where \( h(x) \) is given by (2.3). A test for the homogeneity of the population standardized mean differences \( \delta_i \) is a test of the hypothesis

\[
H_0: \delta_1 = \delta_2 = \cdots = \delta_k
\]

versus the alternative that at least one of the \( \delta_i \) differs from the others.

The test statistic

\[
Q_k = \frac{1}{\sum n_i [h(\hat{\delta}_i) - \bar{h}]^2}
\]

has the chi-square distribution with \( (k-1) \) degrees of freedom under the null hypothesis \( H_0 \). Therefore if the statistic \( Q_k \) is greater than the critical value of the chi-square distribution with \( (k-1) \) degrees of freedom, \( H_0 \) is rejected and we cannot consider the \( \hat{\delta}_i, i = 1, \ldots, k \) as
samples from a population with a common \( \delta \). The clustering procedures for unit normal variates can be applied to see which \( \hat{\delta}_1 \) do not share a common population effect size.

To carry out the procedure transform the \( \hat{\delta}_i \) to \( 2\sqrt{n_i} \left[ h(\hat{\delta}_i) - \bar{h} \right] \), each of which has an approximate standard normal distribution under \( H_0 \). If the \( n_i \) are not equal, replace each \( n_i \) by

\[
n'' = \frac{1}{k} \sum_{i=1}^{k} \sqrt{n_i}^2,
\]

to yield

\[
U_i = 2\sqrt{n''} \left[ h(\hat{\delta}_i) - \bar{h} \right], \quad i = 1, \ldots, k.
\]

We can now carry out clustering procedures using these \( U \)-values.

5.1 An Example of Clustering Standardized Mean Differences

The techniques described in this paper were applied to data from nine studies on the effects of open instruction on the independence and self-reliance of students (Hedges and Gage, 1980). Table 4 is a tabulation of sample standardized mean differences \( \hat{\delta}_i \), sample sizes \( n_i \), and transformed values \( h(\hat{\delta}_i) \) for each of the studies. The weighted average \( \bar{h} \) of the transformed effect size estimates is \( \bar{h} = .055 \). The homogeneity test statistic \( Q_k = 58.93 \) is compared to the .05 critical value of the chi-square distribution with 8 degrees of freedom to yield a significant result, namely, the hypothesis of homogeneity of effect sizes is rejected.

The sample sizes \( n_i \) are not all equal so we compute the square-mean-root \( n'' = 29.266 \). The normalized variates \( U_{(i)} \) are also given in Table 4. The gaps are
### TABLE 4
Data for the Example of Clustering Standardized Mean Differences

<table>
<thead>
<tr>
<th>Study</th>
<th>$n_i$</th>
<th>$\hat{\delta}_i$</th>
<th>$h(\hat{\delta}_i)$</th>
<th>$2\sqrt{n_i} (h(\hat{\delta}_i) - \bar{h})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22</td>
<td>2.016</td>
<td>.663</td>
<td>6.578</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>.699</td>
<td>.234</td>
<td>1.937</td>
</tr>
<tr>
<td>3</td>
<td>32</td>
<td>.643</td>
<td>.225</td>
<td>1.839</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>.091</td>
<td>.032</td>
<td>-0.249</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>-.058</td>
<td>-.020</td>
<td>-0.811</td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td>-.079</td>
<td>-.028</td>
<td>-0.900</td>
</tr>
<tr>
<td>7</td>
<td>30</td>
<td>-.235</td>
<td>-.083</td>
<td>-1.493</td>
</tr>
<tr>
<td>8</td>
<td>30</td>
<td>-.494</td>
<td>-.174</td>
<td>-2.478</td>
</tr>
<tr>
<td>9</td>
<td>30</td>
<td>-.587</td>
<td>-.206</td>
<td>-2.824</td>
</tr>
</tbody>
</table>

Note: These data were obtained from Hedges and Gage, 1980.
\[ U(9) - U(8) = 4.641 \quad U(8) - U(7) = .098, \quad U(7) - U(6) = 2.088, \]
\[ U(6) - U(5) = .562, \quad U(5) - U(4) = .089, \quad U(4) - U(3) = .593 \]
\[ U(3) - U(2) = .985, \quad U(2) - U(1) = .346. \]

Comparison of these gaps with 1.81, the \( \alpha = .05 \) critical value of the gap for \( k = 9 \), we see that only the gaps between \( U(9) \) and \( U(8) \) and that between \( U(7) \) and \( U(6) \) are significant. Thus the gap procedure yields the clusters

\[ \{U(9)\}, \{U(8), U(7)\}, \text{ and } \{U(6), U(5), U(4), U(3), U(2), U(1)\}. \]

The overlapping clustering procedure yields different clusters at the \( \alpha = .05 \) significance levels. Comparing \( U(9) \) with \( U(8) \) we see that
\[ U(9) - U(8) = 4.641 > c(2) = 2.77, \quad \text{so } U(9) \text{ is different from all the other } U(1). \]
Comparing \( U(8) \) with \( U(2) \) and \( U(1) \) we see that
\[ U(8) - U(1) = 4.761 < c(8) = 5.11, \quad \text{so } U(8) \text{ is not different from } U(1). \]
Finally we note that \( U(7) - U(1) = 4.663 < c(9) = 4.95, \quad \text{so } U(1), \ldots, U(7) \text{ are not significantly different. No other differences are significant. The clusters resulting can be represented by the following diagram}

\[
\begin{array}{cccccccc}
U(1) & U(2) & U(3) & U(4) & U(5) & U(6) & U(7) & U(8) & U(9) \\
\end{array}
\]

where variates that are underscored do not differ significantly.
6. **The Effect of Unequal Sample Sizes**

Both of the clustering procedures depend on a transformation of the correlation coefficients or effect size estimates from each study into standard normal variates. When the sample sizes from the studies are unequal we recommend using the square mean root sample sizes $n'$ and $n''$ to create the variates $U_i$. If some of the studies have very different sample sizes, then some of the $U_i$ will have variances much smaller or larger than one. An obvious practical question is whether differences among the variances of the $U_i$ have a large effect on the statistical significance levels associated with the clustering procedures. This is a very difficult question in general, but some insight can be obtained by investigating the situation when there is a single discrepant observation. This is also a practically important situation since it is often the case that most studies of a phenomenon have similar sample sizes, and only a few studies are very much larger or much smaller than the rest.

6.1 **The Effect of Unequal Sample Sizes on the Disjoint Clustering Procedure**

Suppose that $U_2',...,U_k'$ are standard normal variates, and that $U_1$ is a normal variate with variance $\sigma^2$, i.e., $U_1 \sim N(0,\sigma^2)$, where $\sigma^2$ is not necessarily equal to one. Let $U(1), U(2),...,U(k)$ be the ordered values of the $U_i$, $i = 1,...,k$. The order statistics $U(i)$ define $(k-1)$ gaps:

\[
U(1) - U(2), U(2) - U(3),...,U(k-1) - U(k).
\]

A direct argument shows that the expected number of gaps that exceed a pre-specified length $d$ is
\[ p^*(d, k, \sigma) = \frac{1}{\sigma} \int_{-\infty}^{\infty} \left\{ [\Phi(y) + 1 - \Phi(y+d)]^k - [\Phi(y)]^k \right\} d\Phi(y) \]

\[ + (k-1) \int_{-\infty}^{\infty} \left\{ [\Phi(y) + 1 - \Phi(y+d)]^{k-2} [\Phi(y/\sigma) + 1 - \Phi((y+d)/\sigma)] - [\Phi(y)]^{k-2} [\Phi(y/\sigma)] \right\} d\Phi(y), \]

where \( \Phi(x) \) is the standard normal distribution function. Note that this expression reduces to the expression for \( p \) given previously for \( \sigma = 1 \).

We can study the effect of the value of \( \sigma \) on the proportion \( p^* \) of gaps expected to exceed \( d \) by evaluating the integrals numerically. By evaluating \( p^*(d, k, \sigma) \) for various values of the \( \sigma \) at the critical values \( d_\alpha \) given in Table 1, we can study the effect of \( \sigma \) on the true probability that at least one of \( (k-1) \) gaps exceeds \( d \). Table 5 is a presentation of values of \( p^*(d_\alpha, k, \sigma) \) for \( k = 3, 5, 10, 20, \sigma = .50, .75, 1.25, 1.50, 1.75, \) and 
\( d_{.10}, d_{.05}, d_{.01}, \) and \( d_{.001} \), where \( d_\alpha \) is the \( \alpha \) critical value of the gap given in Table 1. We see that \( p^* \) is sensitive to \( \sigma \), and that \( p^* \) is more sensitive to changes in \( \sigma \) in small samples than in large samples. Furthermore, for \( k \geq 5 \), the values of \( p^* \) are reasonably close to their nominal levels if \( .75 \leq \sigma \leq 1.25 \). There is also a trend for \( p^* \) to be less sensitive to large values of \( \sigma \), \( (\sigma > 1) \) than to small values of \( \sigma \), \( (\sigma < 1) \).

Since the variances of the transformed estimators of \( \rho \) and \( \delta \) are inversely proportional to \( n \), these results suggest that the significance levels of the disjoint clustering procedure will be reasonably accurate when \( k \geq 5 \) and the most deviant sample sizes are between 80% and 133% of the others.
6.2 The Effect of Unequal Sample Sizes on the Overlapping Clustering Procedure

Suppose that \( U_2, \ldots, U_k \) are standard normal variates and that \( U_i \sim N(0, \sigma^2) \). If \( U_{(1)}, U_{(2)}, \ldots, U_{(k)} \) are the ordered values of the \( U_i, i = 1, \ldots, k \) then the range is \( W = U_{(k)} - U_{(1)} \). A direct argument yields the cumulative distribution function \( F(w) \) for the range:

\[
F(w) = \frac{1}{\sigma} \int_{-\infty}^{\infty} [\Phi(x+\omega) - \Phi(x)]^{k-1} d\Phi(x/\sigma)
\]

\[
+ (n-1) \int_{-\infty}^{\infty} \left\{ [\Phi(x+\omega) - \Phi(x)]^{k-2} \left[ \Phi((x+\omega)/\sigma) - \Phi(x/\sigma) \right] \right\} d\Phi(x),
\]

where \( \Phi(x) \) is the standard normal distribution function. Note that this expression reduces to the usual expression for the cumulative distribution function for the range of standard normal variates when \( \sigma = 1 \) (see e.g., Harter, 1961).

We can study the effect of \( \sigma \) on tests of the significance of the range by evaluating the function \( F(w) \) numerically. Table 6 is a tabulation of the actual significance level of the range for various values of \( k \) and \( \sigma \), and various nominal significance levels \( \alpha \) (i.e., significance levels assuming \( \sigma = 1 \)). We see from Table 6 that the significance levels of the overlapping clustering procedure, which depend on the distribution of the range, are quite sensitive to \( \sigma \). Even for fairly large \( k \), it may be necessary to use a low nominal significance level, such as \( \alpha = .01 \), to assure a true significance level of .05.
TABLE 5

Exact Probabilities that at Least One Gap Among k Normal Order Statistics Exceeds the Critical Value for the Gap at Significance Level $\alpha$ When ($k-1$) of the Variates Have Variance 1 and One Variate Has Variance $\sigma^2$

<table>
<thead>
<tr>
<th>$k$</th>
<th>$\sigma$</th>
<th>$\alpha = .10$</th>
<th>$\alpha = .05$</th>
<th>$\alpha = .01$</th>
<th>$\alpha = .001$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.50</td>
<td>0.053</td>
<td>0.024</td>
<td>0.003</td>
<td>0.0002</td>
</tr>
<tr>
<td>3</td>
<td>0.75</td>
<td>0.070</td>
<td>0.035</td>
<td>0.005</td>
<td>0.0004</td>
</tr>
<tr>
<td>3</td>
<td>1.25</td>
<td>0.141</td>
<td>0.082</td>
<td>0.020</td>
<td>0.0030</td>
</tr>
<tr>
<td>3</td>
<td>1.50</td>
<td>0.190</td>
<td>0.120</td>
<td>0.038</td>
<td>0.0085</td>
</tr>
<tr>
<td>3</td>
<td>1.75</td>
<td>0.242</td>
<td>0.163</td>
<td>0.063</td>
<td>0.0193</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>0.076</td>
<td>0.038</td>
<td>0.007</td>
<td>0.0007</td>
</tr>
<tr>
<td>5</td>
<td>0.75</td>
<td>0.083</td>
<td>0.041</td>
<td>0.008</td>
<td>0.0007</td>
</tr>
<tr>
<td>5</td>
<td>1.25</td>
<td>0.130</td>
<td>0.071</td>
<td>0.017</td>
<td>0.0025</td>
</tr>
<tr>
<td>5</td>
<td>1.50</td>
<td>0.171</td>
<td>0.102</td>
<td>0.032</td>
<td>0.0074</td>
</tr>
<tr>
<td>5</td>
<td>1.75</td>
<td>0.218</td>
<td>0.141</td>
<td>0.055</td>
<td>0.0169</td>
</tr>
<tr>
<td>10</td>
<td>0.50</td>
<td>0.098</td>
<td>0.050</td>
<td>0.010</td>
<td>0.0009</td>
</tr>
<tr>
<td>10</td>
<td>0.75</td>
<td>0.094</td>
<td>0.048</td>
<td>0.009</td>
<td>0.0008</td>
</tr>
<tr>
<td>10</td>
<td>1.25</td>
<td>0.117</td>
<td>0.062</td>
<td>0.014</td>
<td>0.0016</td>
</tr>
<tr>
<td>10</td>
<td>1.50</td>
<td>0.147</td>
<td>0.084</td>
<td>0.024</td>
<td>0.0045</td>
</tr>
<tr>
<td>10</td>
<td>1.75</td>
<td>0.182</td>
<td>0.112</td>
<td>0.040</td>
<td>0.0106</td>
</tr>
<tr>
<td>20</td>
<td>0.50</td>
<td>0.104</td>
<td>0.051</td>
<td>0.010</td>
<td>0.0010</td>
</tr>
<tr>
<td>20</td>
<td>0.75</td>
<td>0.101</td>
<td>0.049</td>
<td>0.010</td>
<td>0.0010</td>
</tr>
<tr>
<td>20</td>
<td>1.25</td>
<td>0.109</td>
<td>0.055</td>
<td>0.012</td>
<td>0.0016</td>
</tr>
<tr>
<td>20</td>
<td>1.50</td>
<td>0.126</td>
<td>0.067</td>
<td>0.019</td>
<td>0.0041</td>
</tr>
<tr>
<td>20</td>
<td>1.75</td>
<td>0.148</td>
<td>0.085</td>
<td>0.030</td>
<td>0.0099</td>
</tr>
<tr>
<td>k</td>
<td>$\sigma$</td>
<td>$\alpha = .10$</td>
<td>$\alpha = .05$</td>
<td>$\alpha = .01$</td>
<td>$\alpha = .001$</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>3</td>
<td>0.50</td>
<td>0.047</td>
<td>0.021</td>
<td>0.004</td>
<td>0.0004</td>
</tr>
<tr>
<td>3</td>
<td>0.75</td>
<td>0.064</td>
<td>0.029</td>
<td>0.005</td>
<td>0.0004</td>
</tr>
<tr>
<td>3</td>
<td>1.25</td>
<td>0.158</td>
<td>0.089</td>
<td>0.024</td>
<td>0.0037</td>
</tr>
<tr>
<td>3</td>
<td>1.50</td>
<td>0.230</td>
<td>0.146</td>
<td>0.052</td>
<td>0.0118</td>
</tr>
<tr>
<td>3</td>
<td>1.75</td>
<td>0.309</td>
<td>0.214</td>
<td>0.094</td>
<td>0.0288</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>0.066</td>
<td>0.032</td>
<td>0.006</td>
<td>0.0006</td>
</tr>
<tr>
<td>5</td>
<td>0.75</td>
<td>0.073</td>
<td>0.035</td>
<td>0.007</td>
<td>0.0006</td>
</tr>
<tr>
<td>5</td>
<td>1.25</td>
<td>0.161</td>
<td>0.090</td>
<td>0.024</td>
<td>0.0036</td>
</tr>
<tr>
<td>5</td>
<td>1.50</td>
<td>0.254</td>
<td>0.163</td>
<td>0.058</td>
<td>0.0135</td>
</tr>
<tr>
<td>5</td>
<td>1.75</td>
<td>0.363</td>
<td>0.259</td>
<td>0.117</td>
<td>0.0377</td>
</tr>
<tr>
<td>10</td>
<td>0.50</td>
<td>0.082</td>
<td>0.042</td>
<td>0.010</td>
<td>0.0042</td>
</tr>
<tr>
<td>10</td>
<td>0.75</td>
<td>0.084</td>
<td>0.043</td>
<td>0.010</td>
<td>0.0042</td>
</tr>
<tr>
<td>10</td>
<td>1.25</td>
<td>0.156</td>
<td>0.089</td>
<td>0.027</td>
<td>0.0109</td>
</tr>
<tr>
<td>10</td>
<td>1.50</td>
<td>0.272</td>
<td>0.181</td>
<td>0.080</td>
<td>0.0449</td>
</tr>
<tr>
<td>10</td>
<td>1.75</td>
<td>0.427</td>
<td>0.322</td>
<td>0.187</td>
<td>0.1231</td>
</tr>
</tbody>
</table>
6. **References**


