Calculation of Smith's Mean Measure of Divergence for Intergroup Comparisons Using Nonmetric Data

Edward F. Harris[1]* and Torstein Sjøvold[2]

[1]*College of Dentistry, University of Tennessee, Memphis*
[2]*Stockholm University, Stockholm, Sweden*

ABSTRACT: The Mean Measure of Divergence (MMD) is a formula that converts a battery of trait frequencies into a numerical value such that the more dissimilar two samples are, the greater the value. This measure of phenetic distance was developed by the statistician Cedric A. B. Smith and has become popular among dental anthropologists and osteologists for estimating the dissimilarity among groups in order to help reconstruct populations' movements and structure over time and space. The purpose of the present study is to present the correct formulae and procedures for the MMD given that (1) numerous errors have entered into the literature concerning the formulae themselves, (2) improvements have been described that should be incorporated, and (3) various misunderstandings and misinterpretations have developed that need clarification. *Dental Anthropology* 2004;17(3):83-93.

Dental anthropologists commonly use morphological data to estimate the degree of dissimilarity among samples—so-called biological distance or phenetic distance. It is supposed that the greater the dissimilarity between two samples, the less the genetic contact between the groups due to separation by time and/or space.

An issue of some interest is how, statistically, to quantify the degree of dissimilarity among groups in an objective manner. Despite the numerous methods suggested in the literature (reviewed in Constandse-Westermann, 1972), dental anthropologists have focused almost exclusively on the use of Cedric A. B. Smith's mean measure of divergence (MMD). Our experience is, however, that there are misunderstandings about the MMD. There seems to be no commercially available computer program to calculate MMD, which would promote consistency, and the purported formula for MMD (if reported at all) differs among authors (including the repeated publication of statistical errors). The purpose of this note is to clarify the calculation of MMD in a simple, intentionally nontechnical manner.

Overview

Constandse-Westermann (1972) and, in particular, Sjøvold (1973, 1977) provide detailed descriptions of the development and use of the MMD. In brief, the British statistician Cedric A. B. Smith devised this statistic for M. S. Grewal (1962) who used it to estimate the biological divergences that had developed across generations in sublines of the common C57BL strain of laboratory mice. Grewal calculated trait frequencies for 27 cranioskeletal bony variants such as the occurrence of foramina, accessory sutures, and bony processes (traits primarily described by Grüneberg, 1950 and by Deol, 1955). It was supposed that the sublines diverged with time due to latent heterozygosity in the inbred line but, primarily, from the accumulation of mutations distinct to each subline—which is why the term *divergence* is used for this phenetic measure rather than *distance*. The MMD subsequently was popularized in anthropological circles by A. C. Berry and R. J. Berry, notably in their nonmetic skeletal comparisons among human groups (*e.g.*, Berry and Berry, 1967; Berry, 1968; Berry, 1974, 1976; and elsewhere). This pair of authors promoted the use of "minor skeletal variants" as epigenetic features that, from their analyses, had a genetic basis but were essentially independent of age and sex and size of the individual.

These minor skeletodental variants, such as the presence of nutrient foramina and accessory molar cusps, can each be viewed as a dichotomous feature, so summary of a sample is easily expressed as a trait frequency—and Smith's MMD provides a method of estimating the phenetic distances among samples arrayed through space and/or time.

Smith's original formula as described by Grewal (1962) is

*Correspondence to: Edward Harris, Department of Orthodontics, College of Dentistry, University of Tennessee, Memphis, Tennessee 38163 USA.
E-mail: eharris@utmem.edu

$$MMD = \frac{\sum_{k=1}^{r} \left(\theta_{ik} - \theta_{jk}\right)^2}{r} - \left(\frac{1}{n_i} + \frac{1}{n_j}\right) \quad \text{[Eq. 1]}$$

That is, the difference between samples *i* and *j* for the frequencies of trait *k* is calculated and then this difference is squared so that positive and negative differences do not cancel one another. The sum of the differences is divided by r, the number of traits used in the equation, in order to generate the "mean" or average difference between samples *i* and *j*. The correction term $1/n_i+1/n_j$ then is subtracted from this average to correct for sampling fluctuations. Grewal (1962:229-230) actually described the MMD in the text of his paper rather than presenting Eq. 1, which led to misinterpretations by other researchers.

It follows from this equation that the "size" of a MMD depends on the battery of traits used, and MMDs generated from different sets of variables are not comparable, even for the same pair of groups. These conditions hold for all measures of "biological distance" (Sokal and Sneath, 1963; Constandse-Westermann, 1972; Reyment, 1991). While it is not our purpose to critique the merits of the MMD, one noteworthy issue is that it does not account for intertrait correlations, which commonly is viewed as a shortcoming. Intertrait associations ("correlated traits") inflate the MMD because correlated traits share some of the same informational content, and this shared (redundant) information increases with the strength of the correlation. For example, the occurrence of incisor lingual shoveling (Hrdlička, 1920, 1922) is strongly intercorrelated on the maxillary central and lateral incisors (and between homologous teeth in the two quadrants), so including trait frequencies of shoveling on both UI1 and UI2 carries a lot of statistically redundant information. Studies have disclosed that nonmetric intertrait associations are more common than expected by chance (*e.g.*, Corruccini, 1974; Scott, 1977, 1978, 1979). On the other hand, Constandse-Westermann (1972) points out that, within an analysis, the same suite of traits is used for all of the pairwise comparisons so that, insofar as intertrait associations are a species-wide phenomenon, the effect of statistical redundancies can be viewed as a constant across the study.

Statistically significant intertrait correlations may also occur by chance. At the conventional alpha level of 0.05, one expects to make a Type I error (*i.e.*, reject a true null hypothesis) 5% of the time. Suppose that a battery of 30 morphological traits is scored (Table 1). One would expect that 21 of the matrix of 435 pairwise correlations would be statistically significant due to chance alone. An associated issue is that the ability to detect statistically significant differences depends on the available sample size (degrees of freedom) available (*e.g.*, Fisher and Van Belle, 1993). Biologically real but weak correlations generally cannot be detected with small sample sizes. Statistical textbooks deal with the subject in much more detail, but guidelines for detecting biologically real intertrait correlations are (1) comparable correlations should appear in the analyses of multiple samples and (2) correlations found in larger samples, where effects of sampling fluctuations are dampened, generally are more reliable. Weak associations, particularly with the sample sizes normally encountered in anthropological studies, will not seriously distort MMD results.

Frequency transformations

The MMD was devised to deal with percentages of dichotomous data (also termed nonmetric or, occasionally, discontinuous traits). This is in contrast to quantitative (interval and ratio scale) data where more common statistical methods can be employed, such as Pearson's (1926) virtually-defunct coefficient of racial likeness, Penrose's formulae (1953) for distance, size and shape (where distance = size + shape), and the current gold-standard, Mahalanobis' D^2 (Mahalanobis, 1936).

Qualitative data, like the frequency of the Dryopithecus Y-5 pattern on a lower molar (Hellman, 1928), generally are converted to percentages, commonly termed trait frequencies. Such data either are scored as dichotomous traits or a "cut-point" is decided upon along an ordinal grading scheme to create dichotomous traits. Formally, the sample frequency of a trait can be expressed as *p* (and the frequency of absence as *q*) such that $p + q = 1$ and $p = 1 - q$. This simply relates to the binomial distribution. The sample variance of this distribution is pq/n (*e.g.*, Sokal and Rohlf, 1995: 419), where *p* and *q* are the frequencies of trait presence and trait absence, respectively, and n is the sample size. For a given sample size, the sample variance is tied to the frequencies of *p* and *q*. The degree of distortion (that is, the changing value of the variance through the frequency distribution from zero to one) increases as the sample size decreases (Fig. 1). This nonlinear association between the variance and the trait frequency

TABLE 1. Number of statistically significant pairwise associations expected in variously-sized batteries of traits

Number of Traits	Number of Correlations	Number Expected from Chance
5	10	1
10	45	2
20	190	9
30	435	21
40	780	39
50	1,225	61

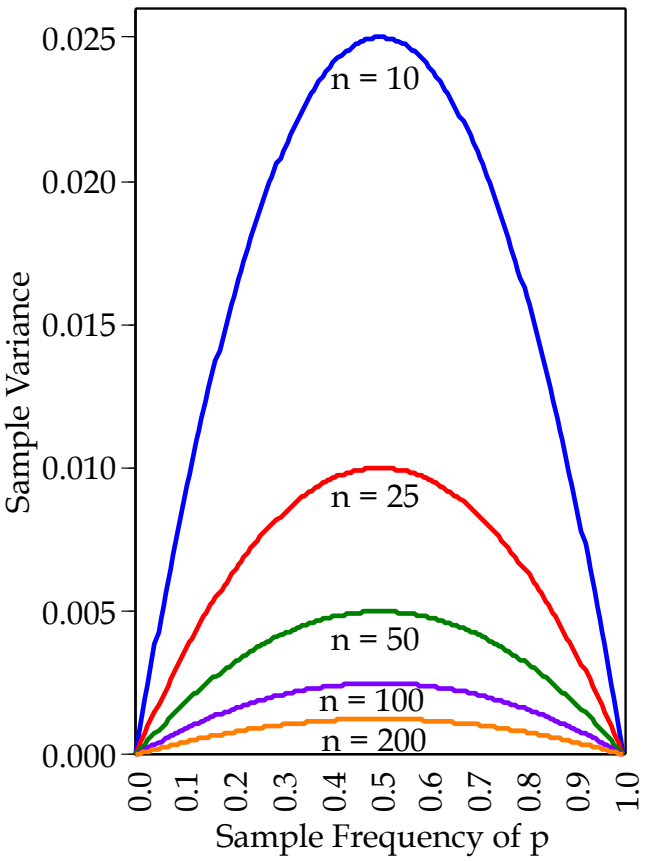

Fig. 1. Examples of how the variance of the trait frequency p changes depending on sample size. In all cases, sample variance is at its maximum when $p = q = 0.5$, but the range of values diminishes as sample size increases.

is obvious in the range of most anthropological samples—either of living or skeletal specimens.

An appropriate transformation of the percentages diminishes this association between a trait frequency and its variance, making the variance more stable. Historically, the transformation can be as simple as $\sin^{-1}\sqrt{p}$ (Fisher, 1958), but other choices work much better. The arcsine (or "arcsin" or "inverse sine") transformation is a trigonometric function, generally coded as $\sin^{-1}$, and it can be expressed either in degrees or radians. (The arcsine function $\sin^{-1}(x)$ is in no way related to $1/\sin(x)$ as might be guessed.)

Transforming trait frequencies introduces an issue that has not been appreciated universally when calculating MMDs. If a researcher uses a familiar transformation—such as $\sin^{-1}\sqrt{p}$ (Fisher, 1958)—with the units in degrees, then the sampling variance of this value is $820.7/n$ (Constandse-Westermann, 1972:118; Sjøvold, 1973:208). Historically, this value was cumbersome when manually calculating the MMD. Instead, the convention has been to express the transformation in radians rather than percentages, but, as Smith (1972: 242-244) illustrates, the results are mathematically identical. Radians are a trigonometric device that simplify many calculations. Several deterministic equalities between degrees and radians can be noted, such as

$$\pi \text{ radians} = 180°$$

$$2\pi \text{ radians} = 360°$$

$$1 \text{ radian} = \frac{180°}{\pi} \approx \frac{180°}{3.14159} \approx 57°17.75'$$

For present purposes, radians are desirable because the transformed frequencies of $\sin^{-1}\sqrt{p}$ have the simple-to-compute variance of *about* $1/4n$, where n is the sample size. The point needs to be emphasized that radians rather than degrees are to be used unless one also incorporates the appropriate variance correction into the MMD equation.

Grewal's (1962) transformation of p is $\sin^{-1}(1-2p)$, and its variance is 4 times as large as that for Fisher's transformation, namely $1/n$ (because 4 times $1/4n = 1/n$), when both are expressed in the same units, either degrees or radians.

Green and Suchey (1976) compared some published frequency transformations and concluded that the formula suggested by Freeman and Tukey (1950) did a decidedly better job of stabilizing the variance than Grewal's $\sin^{-1}(1-2p)$ transformation. The Freeman-Tukey transformation is

$$\theta = \frac{1}{2}\arcsin\left(1-\frac{2m}{n+1}\right)+\frac{1}{2}\arcsin\left(1-2\left(\frac{m+1}{n+1}\right)\right) \qquad \text{[Eq. 2]}$$

where m is the number of occurrences of the trait in the sample and n is the number of scorable specimens in the sample so the trait frequency is $p = m/n$. θ is computed for the kth trait in sample i and likewise for sample j, then these two values are entered into Eq. 1. This means that the raw counts (m and n) are needed to calculate the MMD, not the trait frequencies. Graphs of these three arcsine transformations of the trait frequency are shown in Figure 2.

In practice, there is very little improvement with the Freeman-Tukey transformation compared to another transformation proposed by Anscombe (1948), namely

$$\theta = \sin^{-1}\left(1-2\left(\frac{m+\frac{3}{8}}{n+\frac{3}{4}}\right)\right) \qquad \text{[Eq. 3]}$$

Indeed, according to the graphical comparisons in Green and Suchey (1976:63), Anscombe's transformation is slightly better than the Freeman-Tukey formula at asymptotically stabilizing sampling variance. Both

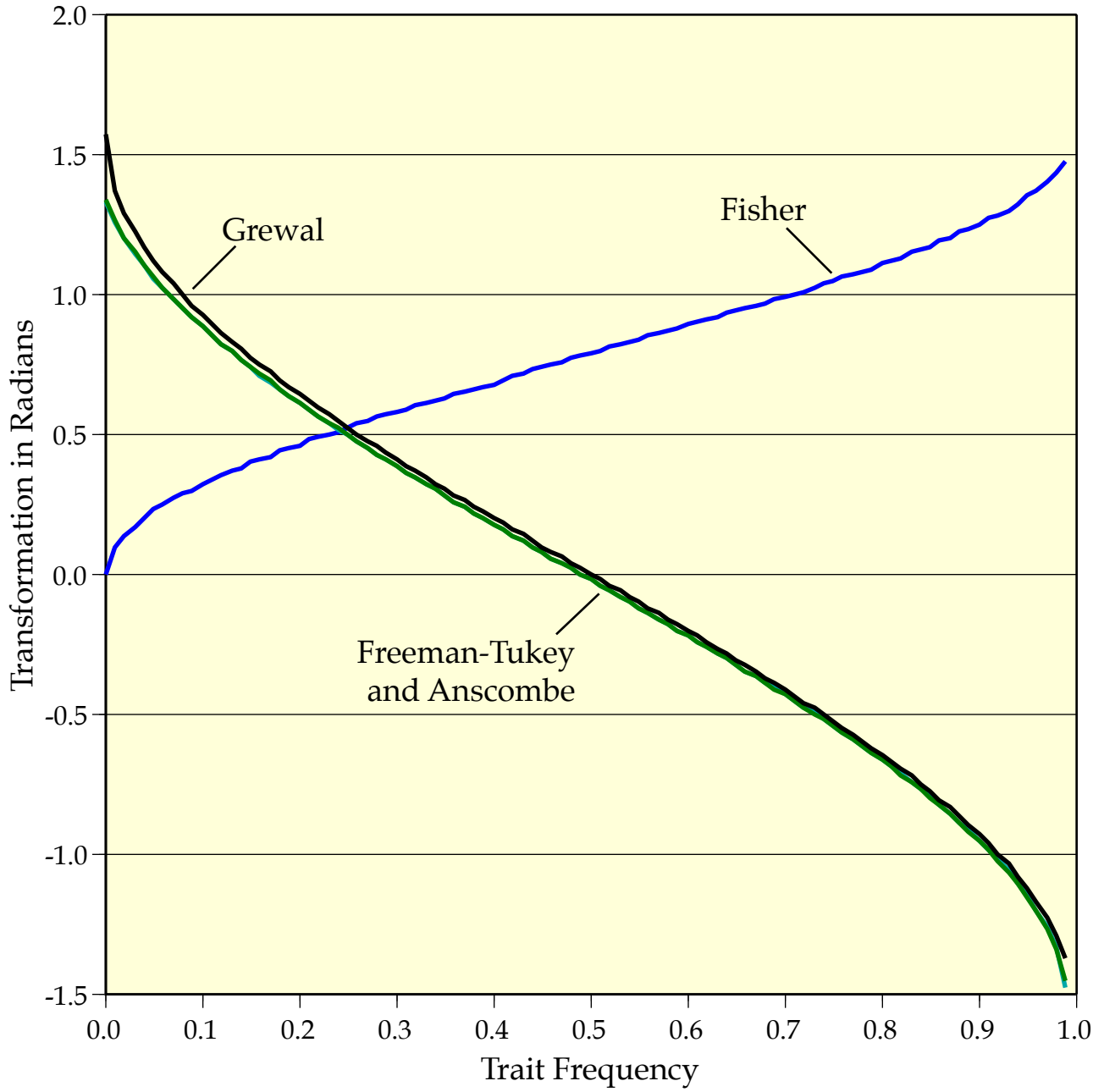

Fig. 2. Graphs of the arcsine transformations of trait frequency p discussed here, namely (1) Fisher's $\sin^{-1}\sqrt{p}$, (2) Grewal's $\sin^{-1}(1\text{-}2p)$, (3) the Freeman-Tukey transformation shown in Eq. 2, and (4) the Anscombe's transformation shown in Eq. 3. Sjøvold (1977) has shown that these latter two transformations are mathematically very similar—which is why they are superimposed here throughout their ranges.

are clear-cut improvements over Grewal's transformation in terms of stabilizing the variances of the binomial variable. We suggest that Anscombe transformation is preferrable for a couple of reasons. Historically, Rao (1952) recommended Anscombe's transformation when sample sizes are moderately large. This transformation also has the advantage that it can be rewritten as a single arcsine:

$$\sin^{-1}\left(\frac{n}{n+\frac{3}{4}}\right)(1\text{-}2p) \qquad \text{[Eq. 4]}$$

or, equivalently,

$$\sin^{-1}\left(\frac{1}{1+\left(\frac{3}{4}n\right)}\right)(1\text{-}2p) \qquad \text{[Eq. 5]}$$

The Freeman-Tukey transformation is quite complicated by comparison. Moreover, Anscombe's formula can be extended to multistate traits (in contrast to dichotomous traits)—though we do not discuss that option in this paper—and this is not true of the Freeman-Tukey formula.

Adjusting for variances

Smith's MMD originally was published without explicit directions (Grewal, 1962), then ambiguously by Berry and Berry (1967:370), and then incorrectly by Berry (1968:115). These shortcomings created a rocky start for the MMD, generating errors that occasionally reappear. Constandse-Westermann (1972: 119) was the first to explicitly publicize this formula:

$$\text{MMD} = \frac{\sum_{k=1}^{r}\left(\theta_{ik}-\theta_{jk}\right)^2 - \left(\frac{1}{n_{ik}}+\frac{1}{n_{jk}}\right)}{r} \qquad \text{[Eq. 6]}$$

though knowing what the equation should be makes the description by Grewal (1962:229-230) clear. Notice that in Eq. 6 the correction term applies to each variable, not just the summary value as indicated in Eq. 1. The quantity$(1/n_{ik}+1/n_{jk})$ is subtracted from the squared difference of trait frequencies to adjust for the mathematical properties of the squared differences between the theta values (θ) that overestimate the divergence between the corresponding populations. That is, $1/n_{ik}+1/n_{jk}$ is the variance of the two angular values. These theoretical and observed distributions coincide more closely as n increases. n_{ik} and n_{jk} are the sample sizes for the *k*th trait so that, depending on how fragmentary the dental or skeletal data are, the usable (scorable) sample sizes will vary from trait to trait.

Notice too that the correction term in Eq. 6 has the subscript *k* that was absent in Eq. 1. Equation 1 assumes that the data are complete, so sample sizes are identical across the whole suite of traits. This commonly is not the case because of damaged skeletal elements or attrition, caries, or loss of teeth. If there are missing data, sample size needs to be subscripted so it can vary by trait.

Green and Suchey (1976) and Green *et al*. (1979) note that this conventional correction formula overestimates the true variance and that, instead, the correction term (attributable to Freeman and Tukey, 1950) should be

$$\frac{1}{n_{ik}+\frac{1}{2}}+\frac{1}{n_{jk}+\frac{1}{2}} \qquad \text{[Eq. 7]}$$

Square-root transformation

If one reviews the various publications using the MMD, it will be seen that a square-root transformation crept into the formula with time. For example, A. C. Berry (1974:348) reports the formula to be:

$$\text{MMD}=\sqrt{\frac{\sum_{k=1}^{r}\left(\theta_{ik}-\theta_{jk}\right)^2 - \left(\frac{1}{n_{ik}}+\frac{1}{n_{jk}}\right)}{r}} \qquad \text{[Eq. 8]}$$

This square-root modification is due to R. J. Berry (1969), and we suggest a couple of reasons for this addition. The square-root modification may be supposed to be an improvement towards the goal of "triangular equality" among the MMDs. Given three groups, say A, B and C, the squared distance between two groups (say A and B) could be greater than the sum of the squared distances between the other two pairs, so AB > (BC + AC). This actually is not true. The actual effect of the square-root transformation is to change the reference space from Cartesian space to a sphere, which creates mathematical problems (Sjøvold, 1977). Even though the square root modification (Eq. 8) is commonly encountered, it stems from a misunderstanding, and we strongly suggest that it not be used.

Alternatively, the square-root modification may have been perceived as a "correction" for estimating a squared divergence, so taking the square root would estimate the unsquared (linear) divergence. Analogously, other researchers have used the square-root of Mahalanobis' D^2, supposing that D is a more relevant measure of intergroup distance than D^2. The modification is unwarranted, though, because the MMD (Eq. 6) is an unbiased estimated of the squared divergence between the populations from which the samples were drawn, but $\sqrt{\text{MMD}}$ is *not* an unbiased estimated of the unsquared divergence (Sjøvold, 1977: 46).

Consider too that the MMD commonly is less than 1.0, so $\sqrt{\text{MMD}}$ will be *larger* than MMD. Artificially increasing MMD by using the square-root transformations makes the test of significance (discussed below) inappropriate because the $\sqrt{\text{MMD}}$ are inflated values, so it is (falsely) harder to achieve statistical significance if it is not understood that the MMD and not the $\sqrt{\text{MMD}}$ needs to be tested.

Sample size

A tangential issue is how to score fragmentary data, particularly dental traits that typically occur bilaterally (*e.g.*, Turner *et al.*, 1991). Incompletely preserved skeletodental data, where the left and right occurrences of a trait cannot always be determined, is a common problem in archeological samples, but the same issue arises with living specimens when the dentition is compromised by caries, attrition, dental restorations, extractions and other causes of tooth loss. Green *et al*. (1979) reviewed the options for scoring incomplete data, concluding that the least biased method is to consider both left and right sides and calculate the trait frequency as the number of times the feature occurs on either side divided by the number of scorable sides. This maximizes the amount of usable information without artificially inflating sample sizes by using sides instead of individuals as the unit of study. It does assume that there is no systematic side preference in trait frequencies, which seems to be the case in the main.

A related issue of sample size becomes obvious from inspection of Eq. 6. If the sample size for a trait is small in one or both samples being compared, then the adjustment factor can be as large or larger than the phenetic difference that is measured as $(\theta_{ik}-\theta_{jk})^2$. This leads to a MMD that is zero or negative, but not because of the similarity in trait frequencies but because of small sample sizes. That is, the adjustment—which is wholly a function of sample sizes—can readily overwhelm the biological measure of difference $(\theta_{ik}-\theta_{jk})^2$, so MMD may well be "controlled" by inadequate sample sizes when dealing with samples in the range typically encountered in anthropological collections.

This artifactual effect of diminutive sample sizes can easily pervade an analysis for several reasons. One, the MMD almost invariably has been applied within a species, so the range of trait frequencies (and, thereby, differences between groups) is not great. Berry and Berry (*e.g.*, 1967) argued that discrete skeletodental traits exhibit considerable differences in frequencies among groups, but this has not been substantiated in the dental anthropological literature (*e.g.*, Lasker and Lee, 1957; Scott and Turner, 1997). Bigger between-group differences in trait frequencies obviously can "offset" the reductionist effect of small sample sizes. Two, sample sizes generally are comparable for the whole suite of traits in a sample; there is little chance of small sample sizes for some traits being offset by substantially larger samples of other traits. Three, when sample sizes are small vis-à-vis the phenetic difference $(\theta_{ik}-\theta_{jk})^2$, the adjustment produces a negative distance for that trait, but it seems that researchers have simply averaged this negative value into the MMD. In fact, a negative value for a trait has no biological meaning; it is wholly an artifact of the frequencies being too similar and/or the samples sizes being too small.

Table 2. Representative sample sizes and associated correction term[1]

Sample size	Correction term
10	0.040
15	0.018
20	0.010
25	0.006
30	0.004
40	0.003
50	0.002
75	$7x10^{-4}$
100	$4x10^{-4}$

[1]Sample size is the scorable number of individuals per group and assumes $n_i = n_j$.

Negative distances

Consider the largest possible difference between a pair of trait frequencies. Suppose, hypothetically,

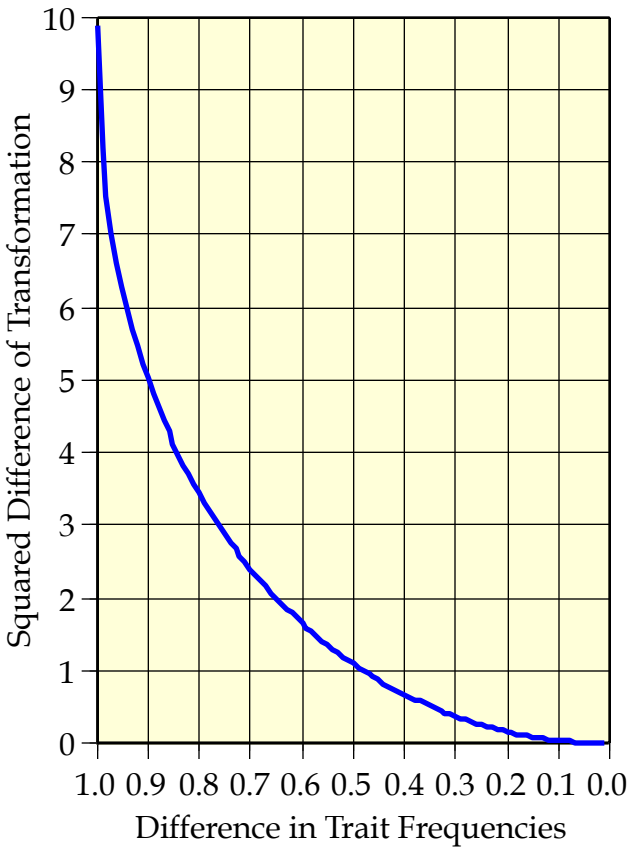

Fig. 3. Graph of the correspondence between the difference in trait frequencies in a pair of samples and the squared difference $(\theta_{ik}-\theta_{jk})^2$ using Grewal's transformation.

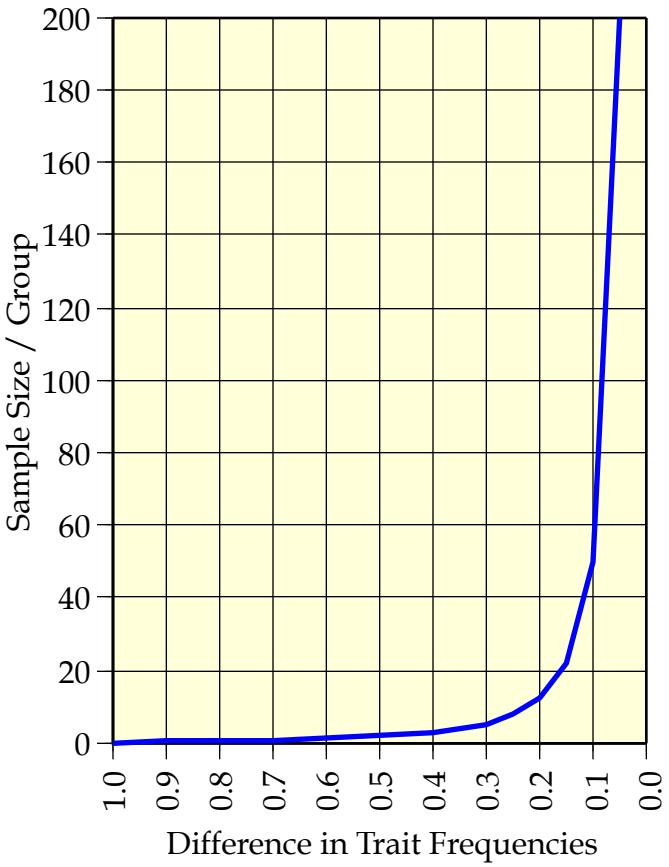

Fig. 4. Graph showing where the difference in trait frequencies (X axis) equals the correction term (Y axis) that is a function of sample sizes.

that a trait like the three-rooted mandibular first molar (Tratman, 1938) is virtually fixed at 99.99% in Group *i* but is quite rare, 0.01%, in Group *j*. This squared difference $(\theta_{ik}-\theta_{jk})^2$ using Grewal's arcsine transformation is 9.62. All other between-group comparisons other than this extreme will be less than 9.62. Obviously, too, as trait frequencies approach each another in two samples—as occurs when groups are genetically and phenotypically more similar—the smaller the $(\theta_{ik}-\theta_{jk})^2$ difference will be and the greater the relative influence of the correction term.

We can look at some simple examples to guage the influence of the correction term (Table 2). The relationship is linear. When sample sizes are less than about 20 (per sample, assuming $n_i = n_j$), the term is fairly large, in excess of 0.10. If sample sizes are 50, the term is 0.002, and if sample sizes are 100, the term is just 0.0004.

These values can be compared to those generated by the squared differences of the transformed frequencies $(\theta_{ik}-\theta_{jk})^2$ as shown in Figure 3. There is a negative hyperbolic relationship here. As examples, when the difference in trait frequency is 0.85, the contribution to the MMD will be 4; when the difference is 0.65, the contribution will be 2; and when the difference is 0.48 the difference will be 1.

Figure 4 graphs these two opposing values, namely the squared difference in trait frequencies $(\theta_{ik}-\theta_{jk})^2$ on the X axis and the sample size (per group) at which this difference is nullified by the correction factor. We see that sample size (per group) can be less than 20 and there will still be a positive contribution to the MMD so long as trait frequencies differ by at least 15 percentage points. If the difference in frequencies is just 10 points, then sample sizes less than 40 will generate a negative MD for that trait. If the difference is just 5 percentage points, sample sizes need to be at least 200 per group. This graph should provide some helpful guidelines when the researcher is deciding which skeletodental traits possess enough intergroup variation to generate meaningful MMDs.

One can see that the potential magnitude of an MMD is limited; the lower limit is zero and the upper limit is less than about 9.6. This upper limit assumes that the sample sizes of the two groups are very large (so the correction factor is effectively nil) and that the trait frequencies between groups are as different as possible for all traits considered. In practice, actual values for the MMD will be far smaller than this. Because the obtained MMD values are small (generally below 0.50), some researchers have multiplied them by 100 or 1,000 for presentation, and this has led to misunderstanding when the research report was not adequately scrutinized by subsequent investigators.

Test of significance

Two groups can have a nonzero MMD simply due to chance deviations because we are dealing with finite *samples* of specimens, not statistical populations. This might make a test of statistical significance useful. The smaller a group's sample sizes, the more the MMD can differ from zero due to sampling fluctuations that do not represent a "true" biological difference.

C. A. B. Smith developed a test of statistical significance for the MMD based on its variance, though, like the distance formula itself, several early publications contain errors. Constandse-Westermann (1972:120) lists the correct formulation of the variance of MMD:

$$4\frac{\sum_{k=1}^{r}\frac{\left(\frac{1}{n_{ik}}+\frac{1}{n_{jk}}\right)}{r}\left[\left(\theta_{ik}-\theta_{jk}\right)^2-\left(\frac{1}{n_{ik}}+\frac{1}{n_{jk}}\right)\right]}{r^2} \qquad \text{[Eq. 9]}$$

To be clear, the standard deviation of this variance is the square root of Eq. 9, namely

$$\sqrt{4\frac{\sum_{k=1}^{r}\frac{\left(\frac{1}{n_{ik}}+\frac{1}{n_{jk}}\right)}{r}\left[\left(\theta_{ik}-\theta_{jk}\right)^2-\left(\frac{1}{n_{ik}}+\frac{1}{n_{jk}}\right)\right]}{r^2}} \qquad \text{[Eq. 10]}$$

Sjøvold (1973:210; 1977:30; also see Green and Suchey, 1976:67) notes that, under the null hypothesis, the variance simplifies to

$$\frac{2}{r^2}\sum_{k=1}^{r}\left(\frac{1}{n_{ik}}+\frac{1}{n_{jk}}\right)^2 \qquad \text{[Eq. 11]}$$

so the square root of Eq. 11 is the standard deviation of MMD

$$\sqrt{\frac{2}{r^2}\sum_{k=1}^{r}\left(\frac{1}{n_{ik}}+\frac{1}{n_{jk}}\right)^2} \qquad \text{[Eq. 12]}$$

Standard statistical theory indicates that two samples will differ significantly at alpha = 0.05 when their means differ by at least 1.96 their standard deviation. This value of 1.96 rounds to 2, which is where the statements come from (*e.g.*, Sjøvold, 1973:216; Green and Suchey, 1976:67) that the null hypothesis of "no difference" can be rejected when the MMD is more than twice its standard deviation (Eq. 12). This rule of thumb is, however, a rough estimate, particularly if the usable sample sizes vary much among the traits used.

There are, however, at least three considerations that detract from the value of testing the statistical significance of MMD: One, the meaning of a "significant" difference is quite vague biologically. This relates to group selection; if two samples are sufficiently different on the basis of geography, anthropology (*i.e.*, race, language, and culture), or distance, then they already characterize separate populations, and no test is required. If, as occurs too frequently in the anthropological literature, samples differ in time, then of course they constitute samples of different populations because a biological population (Mayr, 1963:136) is a

> community of potentially interbreeding individuals at a given locality. All members of a local population share in a single gene pool, and such a population may be defined also as a group of individuals so situated that any two of them have equal probability of mating with each other and producing offspring….

This is where the oxymoron of a "skeletal population" is seen to be absurd (Cadien *et al.*, 1974).

It might be countered that the aim is to see whether two samples are so similar that they can be considered to be drawn fro the same statistical population. Smith (1972:243) notes that, "Alas, this seems to confuse the ideas and uses of a 'distance' and a 'test of significance'. Also, it is usually a nonsensical question, for two distinct populations are distinct, and are not in any reasonable way samples from a single population." Moreover, there are more appropriate and more efficient statistical methods for testing the differences in trait frequencies than the averaged result given by the MMD (see, *e.g.*, Fleiss, 1981; Sokal and Rohlf, 1995).

Two, "The crucial point in every problem concerning biological divergence or distance—and in fact for the study of biological distance in general as well—is the choice of variables of a given set to use" (Sjøvold, 1977:31). The issue here is that the size of the MMD between pairs of samples can be increased or diminished simply by varying the traits used. This issue has been reviewed in depth in books on numerical taxonomy because trait selection—which traits and how many traits—is so central to the results obtained (*e.g.*, Sokal and Sneath, 1963; Reyment, 1991). The issue revolves on two considerations (see Sjøvold, 1977:31), one is whether the chosen trait frequencies are sufficiently different among the groups while still being representative of the groups and, two, whether intergroup divergence is diminished or accentuated by the traits selected in the prior consideration. Those familiar with population differences in dental trait frequencies (reviewed in Turner *et al.*, 1991; Hillson, 1996; Scott and Turner, 1997) will appreciate that different traits discriminate between different groups; important discriminators for one comparison are noncontributory in other comparisons. The "best" discriminators depend wholly on the groups being compared. Put simply, the quantitative results from the MMD (and other distance statistics) are prone to researchers' biases in trait selection. A test of statistical significance is, then, of little practical use.

The researcher needs to be aware of the influence of trait selection and be prepared to defend the suite of traits used for an analysis. The simple inclusion of "lots" of dental traits actually is counterproductive because most do not differ sufficiently among groups or, like the paramolar tubercle of Bolk (Dahlberg, 1945) or the Uto-Aztecan premolar (Morris *et al.*, 1978), occur too infrequently to contribute numerically to a MMD. Sjøvold (1973:211) also makes the point that "dummy" variables are not to be used; these are traits that are fixed across all of the samples studied (either always present or always absent).

Sjøvold recommends the use of Bartlett's adjustment (Bartlett, 1936) when the trait frequency is fixed in a given sample: If the trait does not occur in a sample ($p = 0$) then it should be replaced by $p = 1/4n$. If the trait always occurs ($p = 1$) then it should be replaced by $p = 1 - (1/4n)$. Green and Suchey (1976) also promote the use of Bartlett's adjustment to help correct for ex-

treme trait frequencies.

Trait selection

The MMD necessitates some care in trait selection in order to preserve its statistical properties. This can present a perceptual conflict with the goals of numerical taxonomy. On the one hand, long-held goals in numerical taxonomy are *repeatability* and *objectivity*. A matrix of MMDs should not depend on the traits selected; instead, a goal is that different researchers, using different sets of traits should arrive at a comparable set of intergroup relationships. An obvious and attractive way of seeking this goal is to use *many* variables, without selection, so the resulting MMDs will constitute a broad, comprehensive consensus of how the groups are related phenetically. Sokal and Sneath put forth the seldom-achieved suggestion that, "At least sixty [traits] seem desirable, and less than forty should never be used" (1963:51). The idea is that *many* traits will more-thoroughly sample the battery of available or possible traits, thus diminishing the influence of any one or a few traits, and similarly will guard against biases in trait selection, thus making the phenetic distances more *objective*.

The statistical problem with this approach is that some—perhaps several or, even, most—traits will be nondiscriminatory among the groups. As seen from Eq. 6, when there is little or no difference in trait frequency among the groups, the contribution of that trait to the MMD will not be zero. Instead, because of the correction factor, the trait's contribution will be negative, which has no biological meaning. And, obviously, intergroup differences in trait frequencies need to be larger to be contributory when sample sizes are smaller.

An obvious solution to the accumulation of negative values in the calculation of an MMD would simply be to set the negative values to zero on a trait-by-trait basis. This, however, creates another problem, so it is not recommended. When MMDs are calculated as in Eq. 6, they are *unbiased* estimates of the underlying population differences. This feature is lost—and with it several statistical properties—if negative contributions are set to zero. If negative values are set to zero, the MMDs will over-estimate the population differences. Instead, we recommend the following two-step approach:

One, *a priori* a scientist should propose to use as large a battery of traits as feasible, thereby seeking the goals of repeatability and objectivity set forth by Sokal and Sneath (1963). This initial list needs to be made explicit in the publication; it may well supply important information for other researchers following up with later studies. However, these proposed traits need to be tested to see which ones contain contributory information, which we define as a trait showing a statistically significant difference between at least one pair of the groups being evaluated. These intergroup differences can be evaluated by any of a number of statistical tests appropriate for rates and proportions (*e.g.*, Fleiss, 1981; Siegel and Castellan, 1988).

This winnowing process (1) removes those traits that will generate negatives values across all pairs of groups during calculation of the MMDs, but (2) does not bias the MMDs' estimates. Again, we contend that it is important to provide the full list of traits (and their trait frequencies) prior to the omission of noncontributory traits.

As an optional third step, those MMDs that are negative can be set to zero, both conceptually and practically, if subsequent use is to be made of them (such as input for cluster analysis or phenograms or other graphical representations). Indeed, it is permissible to set all MMDs that are less than twice their standard deviations to zero since, statistically, these estimates of the underlying population differences are nonsignificant. Such values are simply within the range of random sampling fluctuations, so their expected values are zero.

The error of "standardization"

Sofaer and colleagues (1986) introduced quite a different approach to calculating the MMD that they term "standardized MMD." They developed their method to try and resolve a serious shortcoming of their data, namely: What if you want to develop a matrix of MMDs for a set of samples, but you did not score the same suite of morphological traits for all of the groups? Sofaer's solution was creative, but wrong.

In concept, one suite of traits ought to produce roughly the same phenetic relationships as another (*e.g.*, Sokal and Sneath, 1963). If enough traits are used, and all of them possess the same inter-group relationships, and each trait produces the same magnitude of intergroup "distances," then this would be approximately true. In actuality, of course, different sets of traits seldom produce comparable phenetic results.

Sofaer's solution was to use MMDs generated between pairs of groups—where different groups were represented by different traits and different numbers of traits. The authors then "standardized" the MMDs by dividing each MMD by its standard deviation (using a formula similar to Eq. 12). This was claimed to be analogous to the conventional z-score standardization,

$$z = \frac{(X - \mu)}{\sigma} \qquad \text{[Eq. 13]}$$

(*e.g.*, Sokal and Rohlf, 1995:101-111) but the analogy quickly breaks down.

Recall that standardizing a normally distributed sample yields z-scores with a mean of zero and a stan-

dard deviation of one (often termed "unit variance" since $\sigma^2 = 1 = \sigma$). Such a distribution occasionally is coded as N(0, 1). This standardization cannot be properly applied to a series of MMDs—unless all of the MMDs are zero (so $\mu = 0$), in which case the operation is pointless. The whole purpose of calculating MMDs among groups is that the groups differ according to some set of trait frequencies. More precisely, one supposes that the *populations* from which the samples are drawn possess meaningfully different trait frequencies; indeed, the degree of phenetic distance (MMD) is expected to differ among groups on a pair-by-pair basis—some groups being more similar and others more different than others for a given set of traits. For a given set of comparisons, some, most, or all of the MMDs will be different from zero. Regardless of particulars, the MMDs will not be zero, nor will the mean of the MMDs be zero.

Moreover, the standard deviation (Eq. 12) is going to suffer from random variations in sample size from trait to trait. Given that (1) most anthropological samples are modest in size, (2) they are samples of convenience (so sample size seldom can be controlled), (3) sample sizes differ among traits, sometimes dramatically, due to unscorable specimens, and (4) trait frequencies seldom vary much among groups, especially after sampling fluctuations are accounted for, "standardization" of MMDs effectively is an exercise in introducing random errors of unknown magnitude that differ in unknown but differing ways from comparison to comparison depending on sample sizes and other random errors, also of unknown and differing magnitudes.

There is, in fact, no analogy between the conventional z-score and Sofaer's treatment of the MMDs. With a set of MMDs, the population mean is not zero, and there is a different standard deviation for every MMD (Eq. 12). Since these standard deviations are primarily tied to the sample sizes of the traits available in the study, "standardization" as described by Sofaer *et al*. divides each MMD by a different and biologically meaningless value. We obviously see no merit—and several problems—with this attempt at "standardization."

Problems with "standardization" seem obvious to us, but the method was applied uncritically by Sutter and Mertz (2004)—evidently with the passive assent of the reviewers as well. What strikes us as particularly unfortunate is that (1) the proper source of the "standardization" method (*i.e.*, Sofaer *et al*., 1986) does not even appear in the literature cited and (2) the method is wrongly-attributed (on their page 136) to Sjøvold (1973), who decidedly did not mention or advocate any such approach. This error is yet another example of where hasty scholarship has created impediments to the correct calculation of MMDs.

Summary

The purpose of this note is to publicize the correct calculation of Cedric A. B. Smith's MMD. This can be summarized in four steps: (1) Eq. 6 is the correct formula for the MMD as devised by Smith and modified by Berry (1969); (2) Smith's arcsine transformation of trait frequencies should be replaced by Anscombe's transformation (Eq. 3) and expressed in radians, not degrees; (3) the sampling correction in Eq. 6 should be replaced by the more accurate term in Eq. 7; and (4) the preliminary battery of traits should be tested univariately for among-group differences and those traits without statistically significant differences in frequencies across all samples should be omitted. Additionally, Bartlett's adjustment should be applied when the sample trait frequency is fixed at 0 or 1. Statistical significance between a pair of populations occurs when the MMD exceeds twice its standard deviation (Eq. 12). The lack of statistical significance does not mean that the samples can be supposed to derive from the same population, but that it is not possible to distinguish the populations they come from by means of the data and/or the sample sizes available.

LITERATURE CITED

Anscombe FJ. 1948. The transformation of Poisson, binomial and negative-binomial data. Biometrika 35:246-254.

Bartlett MS. 1936. The square root transformation in the analysis of variance. J Roy Stat Soc suppl 3:68-78.

Berry AC. 1974. The use of non-metrical variations of the cranium in the study of Scandinavian population movements. Am J Phys Anthropol 40:345-358.

Berry AC. 1976. The anthropological value of minor variants of the dental crown. Am J Phys Anthropol 45:257-268.

Berry AC, Berry RJ. 1967. Epigenetic variation in the human cranium. J Anat 101:361-379.

Berry RJ. 1968. The biology of non-metrical variation in mice and men. In: Brothwell DR, editor. The skeletal biology of earlier human populations. London: Pergamon Press, p 103-133.

Berry RJ. 1969. History in the evolution of *Apodemus sylvaticus* (Mammalia) at one edge of its range. J Zool London 159:311-328.

Cadien JD, Harris EF, Jones WP, Mandarino LJ. 1974. Biological lineages, skeletal populations, and microevolution. Yrbk Phys Anthropol 18:194-201.

Constandse-Westermann TS. 1972. Coefficients of biological distance. The Netherlands: Oosterhout N. B.

Corruccini RS. 1974. An examination of the meaning of cranial discrete traits for human skeletal biological studies. Am J Phys Anthropol 40:425-445.

Dahlberg AA. 1945. The paramolar tubercle (Bolk). Am J Phys Anthropol 3:97-103.
Deol MS. 1955. Genetical studies on the skeleton of the mouse. XIV. Minor variants of the skull. J Genet 53: 498-514.
Fisher LD, Van Belle G. 1993. Biostatistics: a methodology for the health sciences. New York: John Wiley and Sons, Inc.
Fisher RA. 1958. Statistical methods for research workers, 13th ed. London: Oliver and Boyd.
Fleiss JL. 1981. Statistical methods for rates and proportions, 2nd ed. New York: John Wiley & Sons.
Freeman MF, Tukey JW. 1950. Transformations related to the angular and square root. Ann Math Stat 21: 607-611.
Green RF, Suchey JM. 1976. The use of inverse sine transformations in the analysis of non-metric cranial data. Am J Phys Anthropol 45:61-68.
Green RF, Suchey JM, Gokhale DV. 1979. The statistical treatment of correlated bilateral traits in the analysis of cranial material. Am J Phys Anthropol 50:629-634.
Grewal MS. 1962. The rate of genetic divergence in the C57BL strain of mice. Genet Res 3:226-237.
Grüneberg H. 1950. Genetical studies on the skeleton of the mouse. I. Minor variants of the vertebral column. J Genet 50:112-141.
Hellman M. 1928. Racial characters in human dentition. Proc Amer Phil Soc 67:157-174.
Hillson S. 1996. Dental anthropology. Cambridge: Cambridge University Press.
Hrdlička A. 1920. Shovel-shaped teeth. Am J Phys Anthropol 3:429-465.
Hrdlička A. 1922. Further studies of tooth morphology. Am J Phys Anthropol 4:141-176.
Lasker GW, Lee MMC. 1957. Racial traits in the human teeth. J Forensic Sci 2:401-419.
Mahalanobis PC. 1936. On the generalized distance in statistics. Proc Nat Inst Sci India 2:49-55.
Mayr E. 1963. Animal species and evolution. Cambridge: Belknap Press.
Morris DH, Dahlberg AA, Glasstone-Hughes S. 1978. The Uto-Aztecan premolars—the anthropology of a dental trait. In: Butler PM, Joysey KA, editors. Development, function, and evolution of the teeth. London: Academic Press, p 69-79.
Pearson K. 1926. On the coefficient of racial likeness. Biometrika 18:105-117.
Penrose LS. 1953. Distance, size and shape. Ann Eugenics 18:337-343.
Rao CR. 1952. Advanced statistical methods in biometric research. New York: John Wiley & Sons.
Reyment RA. 1991. Multidimensional palaeobiology. Oxford: Pergamon Press.
Scott GR. 1977. Lingual tubercles and the maxillary incisor-canine field. J Dent Res 56:1192.
Scott GR. 1978. The relationship between Carabelli's trait and the protostylid. J Dent Res 57:570.
Scott GR. 1979. Association between the hypocone and Carabelli's trait of the maxillary molars. J Dent Res 58:1403.
Scott GR, Turner CG II. 1997. The anthropology of modern human teeth: dental morphology and its variation in recent human populations. Cambridge: Cambridge University Press.
Siegel S, Castellan NJ. 1988. Nonparametric statistics for the behavioral sciences, 2nd ed. New York: McGraw-Hill, Inc.
Sjøvold T. 1973. The occurrence of minor non-metrical variants in the skeleton and their quantitative treatment for population comparisons. Homo 24: 204-233.
Sjøvold T. 1977. Non-metrical divergence between skeletal populations. Ossa 4:suppl. 1.
Smith CAB. 1972. Coefficients of biological distance. Ann Hum Genet 36:241-245.
Sofaer JA, Smith P, Kaye E. 1986. Affinities between contemporary and skeletal Jewish and non-Jewish groups based on tooth morphology. Am J Phys Anthropol 70:265-275.
Sokal RR, Rohlf FJ. 1995. Biometry: the principles and practice of statistics in biological research, 3rd ed. San Francisco: WH Freeman and Company.
Sokal RR, Sneath PHA. 1963. Principles of numerical taxonomy. San Francisco: WH Freeman and Company.
Sutter RC, Mertz L. 2004. Nonmetric cranial trait variation and prehistoric biocultural change in the Azapa Valley, Chile. Am J Phys Anthropol 123:130-145.
Tratman EK. 1938. Three-rooted lower molars in man and their racial distribution. Br Dent J 64:264-274.
Turner CG II, Nichol CR, Scott GR. 1991. Scoring procedures for key morphological traits of the permanent dentition: the Arizona State University dental anthropology system. In: Kelley MA, Larsen CS, editors. Advances in dental morphology. New York: Wiley-Liss, p 13-32.

Editor's note:
Copies of the publications by T. Sjøvold (1973, 1977) are available by contacting the author:

Prof. Torstein Sjøvold
Osteology Unit
Wallenberg Laboratory
Stockholm University
SE-106 91 Stockholm, Sweden

Torstein.Sjovold@ofl.su.se