Bioarchaeology of the Near East, 10:69–73 (2016)

Technical Note: A GNU Octave function for Smith's Mean Measure of Divergence

Andreas Bertsatos*, Maria-Eleni Chovalopoulou
Department of Animal and Human Physiology,
National and Kapodistrian University of Athens,
Panepistimioupolis 157 72, Athens, Greece
email: abertsatos@biol.uoa.gr (corresponding author)

Abstract: *The present paper introduces a function written in GNU Octave language for calculating Smith's Mean Measure of Divergence (MMD). The function uses by default the Freeman and Tukey angular transformation but also provides an optional choice for Anscombe's transformation. The Freeman and Tukey correction term is used in both cases. The function calculates the distance and variance matrices of MMD for an arbitrary number of groups (populations) of samples and trait frequencies. The generated matrices are returned in the Phylip format for direct use with tree construction tools.*

Key words: GNU Octave language; Smith's MMD; distance matrix; non-metric traits

Cedric A.B. Smith's Mean Measure of Divergence (MMD) is a dissimilarity measure of the phenetic affinities between samples. Its relaxed restriction for complete datasets has made MMD popular among physical anthropologists studying dental and skeletal non-metric traits (Irish 2010). The lack of commercially available software to calculate MMD along with the various mathematical misinterpretations and variants of the formula has led to inconsistencies among authors including the repeated publication of statistical errors (Harris & Sjøvold 2004). Sołtysiak (2011) introduced a script written in R language in an effort to promote consistency for calculating MMD. However, the need of user modifications in the script along with the need of Comma Separated Value (CSV) files for data input and output limit its practicality to R experts.

GNU Octave (Eaton et al. 2015) is a free software primarily intended for numerical computations featuring a language that is mostly compatible with MATLAB. GNU Octave software is freely available at www.gnu.org/software/octave/ under the GNU General Public License. The need for a standard function for calculating MMD, which can be used as standalone in the GNU Octave/MATLAB environments or directly called in other worker's programs, has led to the development of the *smithsMMD.m* function, which is available in the Supplementary File available on *Bioarchaeology of the Near East*'s website (www.anthropology.uw.edu.pl). This paper presents the underlying statistics of the function along with certain examples of its use.

Received 15 November 2016; accepted 14 January 2017; published online 14 January 2017 on www.anthropology.uw.edu.pl

The function requires at least two arguments, namely *frequency table* and *population*. The *frequency table* is a three-dimensional array with each corresponding dimension as follows: freq_table(population, trait, frequency), where population and trait are scalars and frequency is a 1×2 matrix containing the number of trait occurrences and the number of individuals examined for any given trait and population. The *population* argument is a string cell containing the names corresponding to each population included in the analysis and should be consistent with the first dimension of the *frequency table* argument. In the present example we consider the frequencies of five traits from four populations, and the respective array can be constructed by issuing the following commands.

```
>> freq_table = [7 13 12 104 28;3 2 3 34 10;21 7 5 33 3;15 21 4 58 34];
>> freq_table(:,:,2) = [168 211 193 198 218;86 85 84 86 86;83 86 83 85
    75;87 105 101 92 74];
```

The *population* string cell can as easily be constructed as follows.

```
>> population = {'Group A','Group B','Group C','Group D'};
```

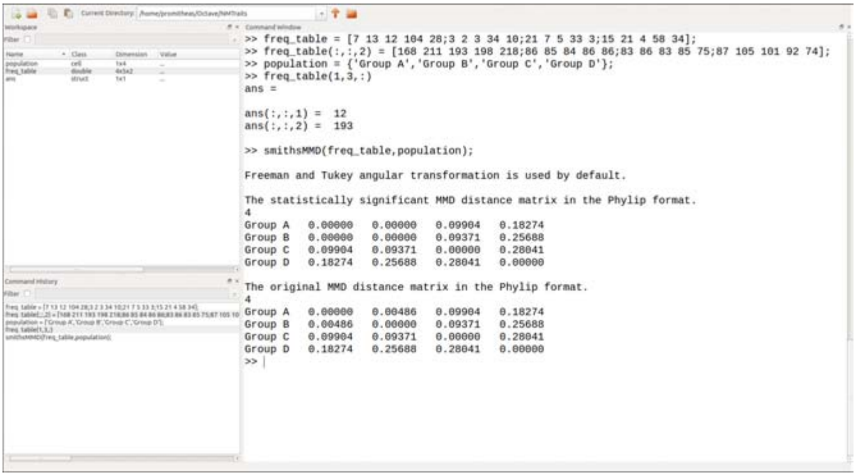
In the given example, the frequency of the third trait of the first population would be 12 occurrences in 193 observations.

```
>> freq_table(1,3,:)
ans =

ans(:,:,1) =  12
ans(:,:,2) =  193
```

Using the two arguments described above, the function will calculate MMD across all populations and print two matrices in the Phylip format as shown in **Figure 1** by issuing the command:

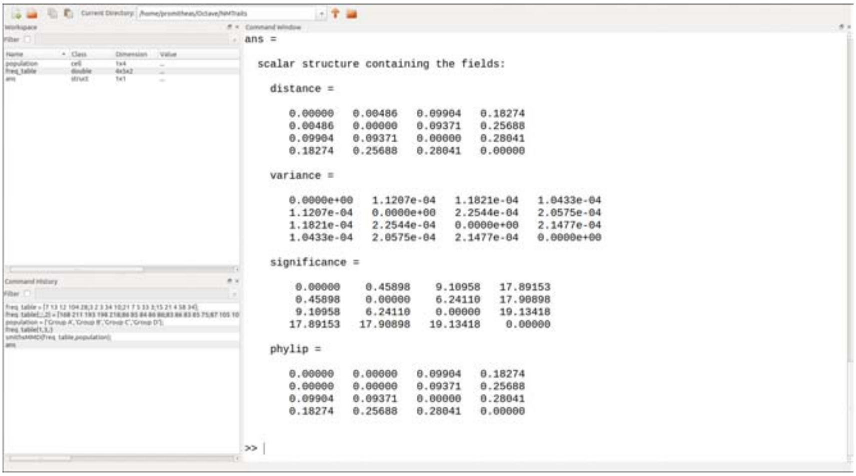
```
>> smithsMMD(freq_table,population);
```

The first matrix displays only the MMD distances that are considered statistically significant based on when the hypothesis of equality of proportions holds, that is, when the populations are not divergent, the MMD may be regarded as significant at a significance level of approximately 2.5% when larger than twice the standard deviation (Sjøvold 1977). All other distance values are set to zero. The second matrix displays the MMD distances as initially calculated. Since the distance matrix produced by MMD is likely to be subsequently used as input for cluster analysis, the Phylip format was chosen as an appropriate display format for the distance matrix. However, the function returns these two matrices along with the variance matrix and the significance matrix in a single structure variable. Note that the significance matrix displays the ratio $MMD\sqrt{varMMD}$ instead of the p-value as a significance measure for each respective MMD distance. The matrices can be displayed by simply typing *ans* in



```
>> freq_table = [7 13 12 104 28;3 2 3 34 10;21 7 5 33 3;15 21 4 58 34];
>> freq_table(:,:,2) = [168 211 193 198 218;86 85 84 86 86;83 86 83 85 75;87 105 101 92 74];
>> population = {'Group A','Group B','Group C','Group D'};
>> freq_table(1,3,:)
ans =

ans(:,:,1) =  12
ans(:,:,2) =  193

>> smithsMMD(freq_table,population);

Freeman and Tukey angular transformation is used by default.

The statistically significant MMD distance matrix in the Phylip format.
4
Group A   0.00000   0.00000   0.09904   0.18274
Group B   0.00000   0.00000   0.09371   0.25688
Group C   0.09904   0.09371   0.00000   0.28041
Group D   0.18274   0.25688   0.28041   0.00000

The original MMD distance matrix in the Phylip format.
4
Group A   0.00000   0.00486   0.09904   0.18274
Group B   0.00486   0.00000   0.09371   0.25688
Group C   0.09904   0.09371   0.00000   0.28041
Group D   0.18274   0.25688   0.28041   0.00000
>> |
```

Figure 1. Creating input arguments and calling *smithsMMD.m* function.



```
ans =

  scalar structure containing the fields:

    distance =

       0.00000   0.00486   0.09904   0.18274
       0.00486   0.00000   0.09371   0.25688
       0.09904   0.09371   0.00000   0.28041
       0.18274   0.25688   0.28041   0.00000

    variance =

       0.0000e+00   1.1207e-04   1.1821e-04   1.0433e-04
       1.1207e-04   0.0000e+00   2.2544e-04   2.0575e-04
       1.1821e-04   2.2544e-04   0.0000e+00   2.1477e-04
       1.0433e-04   2.0575e-04   2.1477e-04   0.0000e+00

    significance =

        0.00000    0.45898    9.10958   17.89153
        0.45898    0.00000    6.24110   17.90898
        9.10958    6.24110    0.00000   19.13418
       17.89153   17.90898   19.13418    0.00000

    phylip =

       0.00000   0.00000   0.09904   0.18274
       0.00000   0.00000   0.09371   0.25688
       0.09904   0.09371   0.00000   0.28041
       0.18274   0.25688   0.28041   0.00000


>> |
```

Figure 2. Displaying all matrices produced by the *smithsMMD.m* function.

GNU Octave's command prompt as illustrated in **Figure 2**. Alternatively, the output of the function can be stored in a user defined structure variable and subsequently each matrix can be accessed independently as shown in the example below.

```
>> MMD = smithsMMD(freq_table,population);
>> MMD.distance
ans =

   0.00000    0.00486    0.09904    0.18274
   0.00486    0.00000    0.09371    0.25688
```

```
    0.09904    0.09371    0.00000    0.28041
    0.18274    0.25688    0.28041    0.00000
```

The *smithsMMD.m* function uses the formula with the Freeman and Tukey correction term as it performs better at extreme trait frequencies (Green & Suchey 1976). By default, θ values are calculated according to the Freeman and Tukey (1950) angular transformation. However, the function also allows the input of a third argument, which optionally defines the use of the Anscombe (1948) angular transformation for calculating the θ values.

The angular transformation can be user defined as a character string. Calling the function as follows:

```
>> smithsMMD(freq_table,population,'Anscombe');
```

will include Anscombe's transformation in the MMD formula, whereas

```
>> smithsMMD(freq_table,population,'Freeman and Tukey');
```

will result the default operation such as with two input arguments. Note that if the third argument is misspelled, the function will return an error message. If no argument for angular transformation is defined by the user, Freeman and Tukey is used by default.

The calculated distances from the *smithsMMD.m* function were also compared against the results of Sołtysiak's R script, which are presented in **Table 1**, and were found similar. The minute differences observed between the two distance matricesare caused due to the input of the R script as percentage frequencies, which have been calculated with 0.01 precision. Nevertheless, both algorithms produce the same output of distance matrix provided the same input of trait frequencies.

Table 1. Distance results of the R script based on the same input data.

	Group A	Group B	Group C	Group D
Group A	0.000000000	0.004893326	0.09900205	0.18276822
Group B	0.004893326	0.000000000	0.09370111	0.25691473
Group C	0.099002050	0.093701108	0.00000000	0.28043944
Group D	0.182768218	0.256914734	0.28043944	0.00000000

Sołtysiak's R script provides a standard deviation matrix instead of the variance matrix for the calculated distances and also reports the respective p-values instead of the $MMD\sqrt{varMMD}$ ratios. However, reporting the statistical significance with the p-values implies that all the single MDs, and consequently the MMD, are related to the chi-squared distribution, which is true only if the number of observations for each trait in a population is constant. Since in most cases the sample sizes across traits vary to some extent, the MMD may be regarded as approximately normally distributed as long as single variances are not allowed to be too large compared with the sum of single variances and as many traits and observations as possible are used (Sjøvold 1973). Therefore, it is preferred to use Sjøvold's rule-of-thumb to consider the MMD significant when larger than twice its standard deviation. Nevertheless, both the *smithsMMD.m* function and Sołtysiak's R script produce similar and consistent results that can be directly compared to each other for all practical purposes.

Acknowledgements

The authors are grateful to Torstein Sjøvold for his comments on the statistics of MMD and the underlying angular transformations and to Arkadiusz Sołtysiak for providing the R script's results for the example dataset along with his suggestions on improving this technical note. We are also thankful to Przemysław Biecek and Efthymia Nikita for their constructive reviews.

References

Anscombe F.J. (1948), *The transformation of Poisson, binomial and negative-binomial data*, Biometrika 35:246-254.

Eaton J.W., Bateman D., Hauberg S., Rik Wehbring R. (2015), *GNU Octave version 4.0.0 manual: A high-level interactive language for numerical computations*, available online.

Freeman M.F., Tukey J.W. (1950), *Transformations related to the angular and square root*, The Annals of Mathematical Statistics 21:607-611.

Green R.F., Suchey J.M. (1976), *The use of inverse sine transformations in the analysis of non-metric cranial data*, American Journal of Physical Anthropology 45:61-68.

Harris E.F., Sjøvold T. (2004), *Calculation of Smith's Mean Measure of Divergence for intergroup comparisons using nonmetric data*, Dental Anthropology 17(3):83-93.

Irish D. (2010), *The Mean Measure of Divergence: Its utility in model-free and model-bound analyses relative to the Mahalanobis D 2 distance for nonmetric traits*, American Journal of Human Biology 22:378-395.

Sjøvold T. (1977), *Non-metrical divergence between skeletal populations*, OSSA. International Journal of Skeletal Research 4(Suppl. 1):1-133.

Sjøvold T. (1973), *The occurrence of minor non-metrical variants in the skeleton and their quantitative treatment for population comparisons*, HOMO – Journal of Comparative Human Biology 24:204-233.

Sołtysiak A. (2011), *Technical Note: An R script for Smith's Mean Measure of Divergence*, Bioarchaeology of the Near East 5:41–44.