Bioarchaeology of the Near East, 5:41–44 (2011)

Technical Note: An R script for Smith's Mean Measure of Divergence

Arkadiusz Sołtysiak
Department of Bioarchaeology,
Institute of Archaeology, University of Warsaw,
ul. Krakowskie Przedmieście 26/28, 00-927 Warszawa, Poland
email: a.soltysiak@uw.edu.pl

Abstract: *The present paper introduces a script written in R language that generates a matrix of Smith's Measure of Divergence (MMD) for a set of non-metric trait frequencies. The script uses either the Anscombe or Freeman & Tukey transformation to obtain theta values, and a choice of the Freeman & Tukey or Grewal correction for small sample size to yield an MMD matrix. A matrix of standard deviations, using Sjøvold's formula, and statistical significance for each value are also calculated.*

Key words: R language; distance matrix; non-metric traits

Dental and skeletal non-metric trait data are frequently recorded as a part of standard osteological protocol (cf. Buikstra & Ubelaker 1992). These data may then be used in research on phenetic affinities among past and present human populations (Scott 2008). Two common distance measures are used for this form of research: 1) Mahalanobis D^2 based on a tetrachoric correlation matrix, and 2) C.A.B. Smith's Mean Measure of Divergence (MMD) (Grewal 1962; Smith 1972). The former measure requires a complete dataset, i.e. all traits must be observed for all individuals; the latter needs only summary data, i.e., occurence proportions in compared subsets for each trait. Thus, the MMD may be used when skeletons or dentitions are incomplete (Irish 2010). Considering that the state of preservation of human remains in the Near East is generally poor, the MMD would be the more practical distance measure for regional research.

Unfortunately, the MMD is not incorporated into any generally available statistical software; however, it can be implemented using R, a free software environment for statistical computing (R Development Core Team 2011). An R script for the MMD is provided in the Supplementary File available at the website of the *Bioarchaeology of the Near East* (www.anthropology.uw.edu.pl). The instructions for using this program are provided below.

Prior to any analysis, the R package should be downloaded from www.r-project.org and installed on the user's computer. The non-metric data should then be prepared using any spreadsheet software that can produce a Comma Separated Values (CSV) file. Two data sheets should be created by the user and then saved as two separate CSV files: 1) the total numbers of individuals scored for a given trait (i.e., whether present or absent), and 2) the amount present for each trait. In both files, the first row must include the names of each non-metric trait, and the first column should list the names of samples to be compared. Data in the second file may

Received 24 January 2012; accepted 12 February 2012; published online 16 February 2012 on www.anthropology.uw.edu.pl

be presented either as proportions, with a range of 0-1, or as percents, with a range of 0-100. No missing data are allowed.

Some user modifications to the script are possible, concerning four options. First, the Anscombe transformation is the default option under Section B (see script at the above link) as currently written:

```
## Anscombe transformation
theta <- function(n,p) { asin((n/(n+3/4))*(1-2*p)) }
## Freeman & Tukey transformation
#theta <- function(n,p) { 0.5*(asin(1-(2*p*n/(n+1)))+asin(1-(2*((p*n)+1)/ (n+1)))) }
```

If the Freeman & Tukey transformation is preferred, the # sign will need to be moved:

```
## Anscombe transformation
#theta <- function(n,p) { asin((n/(n+3/4))*(1-2*p)) }
## Freeman & Tukey transformation
theta <- function(n,p) { 0.5*(asin(1-(2*p*n/(n+1)))+asin(1-(2*((p*n)+1)/ (n+1)))) }
```

Second, there are three options for the correction of small sample size: Freeman & Tukey, Grewal, and the uncorrected formula (per Harris & Sjøvold 2003). Freeman & Tukey is the default. If one of the other options is preferred, the # sign will, as above, need to be moved accordingly in Section B. Specifically, it should be placed at the beginning of the line below the label "Freeman & Tukey correction" and removed from the beginning of the line below the label for desired option.

Third, it is assumed that the second user-created data file (see above) contains percents of occurrence for each non-metric trait (i.e., 0-100); if proportions are used instead (0-1), a similar substitution of the # signs must be made twice in Sections C and D, for lines below labels "percent frequencies" and "proportions."

The correction for the sample size produces negative values when a sample is compared to itself. It is possible to turn them automatically to 0 in the MMD matrix by removing the # sign in the last line of the Section D.

After the two user data files have been created and the script modified as needed, the latter file should be opened in the R environment using *File / Open script.* Next, the script should be launched via *Edit / Run all* (once the script window is active). The instructions for accessing files should be followed, and it may take some time for the process to finish. The script lines that have been executed will be displayed in red, assuming no colour preferences were changed by the user. If file selection was cancelled by user, the script would be executed anyway, with inappropriate data.

As the default, the script generates four matrices displayed in blue in the R console. First, a diagonal matrix is provided that contains data useful in trait selection. Two values are listed after each trait name; one is the mean MMD for this single trait, and the other is the proportion of positive MMD values for the trait. A negative MMD value indicates that the sample size may be too small and/or a given trait does not differentiate between samples; as such, it may be removed from the dataset as desired (see Harris & Sjøvold 2003). For clearer inter-population differentiation, only traits with a positive first value and the highest possible second value should be selected for the analysis. This numeric list is valid only if a correction for small sample size was used (an uncorrected matrix will have no negative distance values).

Three consecutive square matrices follow in the output (see **Figure 1**): (1) the MMD matrix, (2) the standard deviations counted using Sjøvold's formula (1973), and (3) the statistical significance for each MMD value, estimated using two-sided Z-scores counted as MMD/SD. Such an estimation of the statistical significance is based on the assumption that the MMD

follows normal distribution, which is not necessarily true, so the interpretation of this figure must be cautious (see also discussion in Harris & Sjøvold 2003 on the statistical significance of the MMD). Each matrix may be saved to a CSV file, by removing the # sign before the write.csv command in Section G; in addition, sample file names should be replaced by the proper names with full paths on the user's computer.

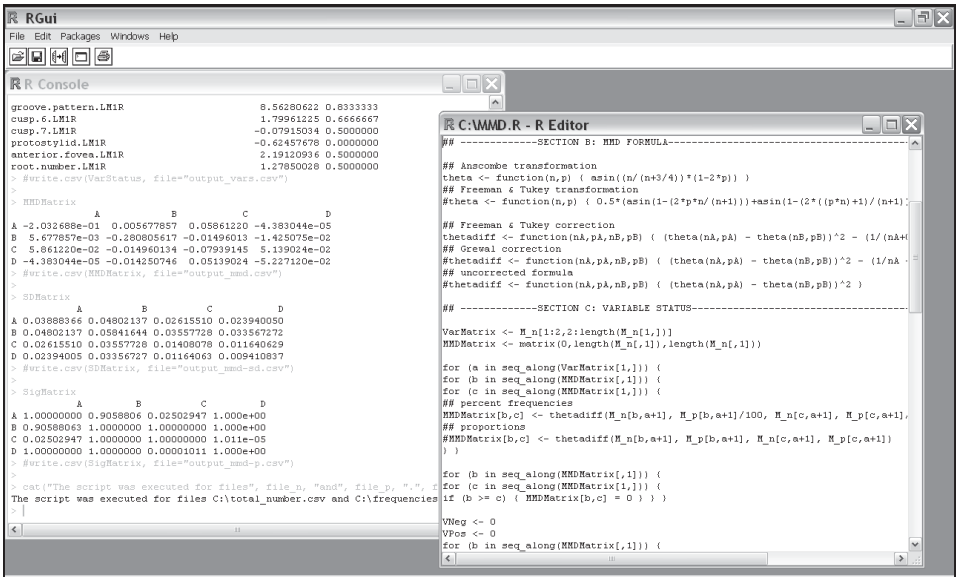

Figure 1. A sample output produced by the script in the R environment.

Acknowledgements

I am most grateful to Joel D. Irish (University of Alaska) for detailed comments concerning both the script and the instruction for users, to Edward F. Harris (University of Tennessee) for several general suggestions and to Piotr Jaskulski (Macrologic SA) for help in enhancing interactivity of the script.

References

Buikstra J.E., Ubelaker D.H., ed. (1994), *Standards of data collection from human skeletal remains*, "Arkansas Archaeological Survey Research Series" 44, Fayetteville.

Grewal M.S. (1962), *The rate of genetic divergence of sublines in the C57BL strain of mice*, Genetical Research 3:226-237.

Harris E.F., Sjøvold T. (2003), *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.

R Development Core Team (2011), *R: A language and environment for statistical computing*, Vienna: R Foundation for Statistical Computing, http://www.R-project.org/.

Scott G.R. (2008), *Dental morphology* [in:] "Biological anthropology of the human skeleton", 2nd edition, M.A. Katzenberg, R.S. Sauders (eds.), Hoboken: John Wiley & Sons, pp. 265-298.

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

Smith C.A.B. (1972), *Coefficients of biological distance. By T.S. Constandse-Westerman*, Annals of Human Genetics 36:241-245 [book review].